CHAPTER TWO

Properties of DNA Sequence Loci: Part I

Genomes harbor a plethora of different loci types that can be used as evolutionary markers in phylogenomic studies. In a research project that aims to elucidate the evolutionary history of particular gene families, there is little or no ambiguity about which locus or loci will be the target(s) for DNA sequence acquisition. However, in organismal phylogenomic studies the researcher must choose the types and numbers of loci to use—a decision that may, if poorly made, adversely affect the quality of phylogenomic inferences. Indeed, some phylogenomic analyses require DNA sequence datasets to meet certain evolutionary assumptions and therefore it is essential to carefully select the type and quantity of data to acquire.

Ideally, locus or loci choice will be based on a researcher's sound knowledge about the properties of each locus type with an eye on the anticipated analyses and associated assumptions. What are these properties of DNA sequence loci? The characteristics of a locus can be influenced or determined by a number of factors. For example, a locus may contain sites that are highly conserved in an evolutionary sense or they may not be conserved at all. Also mutation rates may be constant for all sites within one type of locus or vary among sites in another locus. Thus patterns and processes of mutation may not only vary between loci but also within them as well. Loci may also differ from each other depending on whether they exist as an entity in the nuclear genome or are found in an organellar genome such as the mitochondrial genome. As we will see in this chapter, loci in the former type of genome evolve quite differently from loci in the latter and thus these differences must be taken into account in phylogenomic studies.

In order to better understand these and other loci properties regardless whether a study is focused on molecular/genomic evolution or organismal evolution, a researcher should be knowledgeable about the characteristics of genomes (e.g., size and composition) as well as how DNA sequences evolve over time. In this chapter, we will begin by reviewing aspects of organismal genomes across the tree of life with a focus on eukaryotes followed by a discussion about key aspects concerning the molecular evolution of DNA sequences. Included in this discussion will be a brief review on the biology of so-called repetitive DNA, an important class of genomic DNA comprised of simple repeat DNA (e.g., microsatellites), transposable elements, and various types of pseudogenes including mitochondrial pseudogenes or "numts." This chapter is important because it will prepare us for Chapter 3, the section of this book that considers the many standard assumptions of phylogenomic analyses and describes the major classes of DNA sequence-based loci used in tree of life studies. Prepared with this knowledge, the researcher can make informed decisions about which loci will be most appropriate for a given study and provide important context for phylogenomic analyses of those data.

2.1 GENOMIC BACKGROUND

2.1.1 Genome Types and Sizes

The three main types of genomes include *prokaryotic genomes*, *eukaryotic nuclear genomes*, and *eukaryotic organellar genomes*. Organellar genomes consist of two types—*mitochondrial genomes* and *chloroplast genomes*. Mitochondrial and chloroplast genomes reside within their namesake organelles in the cytoplasm

and thus they are physically separated from the nuclear genome. Mitochondrial genomes are found in all different types of eukaryotes including protists, green alga, fungi, yeasts, animals, and land plants, but only photosynthetic eukaryotes also have chloroplast genomes (Brown 2007).

The genomes of prokaryotes and organellar genomes usually exist as single circular double-stranded DNA molecules though some prokaryotes have genomes subdivided into multiple linear chromosome-like molecules (Brown 2007). In contrast, all eukaryotes have their nuclear genomes apportioned among a number of linear chromosomes (Brown 2007). Interestingly, other similarities exist between prokaryotic and organellar genomes such as both having similar gene expression patterns and gene sequences (Brown 2007). The similarities between prokaryotic and eukaryotic organellar genomes may have an evolutionary basis according to the *endosymbiont theory*, which was developed by Lynn Margulis in the late 1960s. According to this theory, eukaryotic organellar genomes are descendants from free-living bacteria, which, long ago, not only lived inside primitive eukaryotic cells, but also had symbiotic relationships with their host cells (Sagan 1967; Margulis 1970).

Genome sizes are expressed in units of kilobases (kb), megabases (Mb), or gigabases (Gb), which are the equivalents to thousands, millions, and billions of bases, respectively. Prokaryotic genomes vary between 0.49 and 30 Mb (Table 2.1) with most being less than 5 Mb (Brown 2007). Organellar genomes range from about 6 kb to more than 11,000 kb (Table 2.2). A great amount of variation can be seen in the sizes of mitochondrial genomes. The protozoan that causes malaria (*Plasmodium falciparum*) has one of the smallest known mitochondrial genomes at 6 kb in size (Table 2.2). At the other end of the size spectrum, catchfly plants in the genus *Silene* have enormous mitochondrial genomes with one species attaining a size of 11,319 kb, which is >600 times larger than the human mitochondrial genome (Table 2.2)! The chloroplast genomes of photosynthesizing eukaryotes range in size from 120 kb for pea plants (*Pisum sativum*) to 195 kb in green alga (*Chlamydomonas reinhardtii*; Table 2.2).

From Table 2.2 it is obvious that the sizes of mitochondrial genomes are not closely related to organismal complexity, as some "simple" eukaryotes such as yeast and fungi have much larger mitochondrial genomes than "more complex" invertebrates or vertebrates but they are smaller than those found in flowering plants. Surprisingly, the mitochondrial genomes of *Chlamydomonas reinhardtii* (a green alga) and *Homo sapiens* are nearly the same size (Table 2.2). Even similar organisms such as the two protozoans in Table 2.2—*Plasmodium falciparum* and *Reclinomonas americana*—have mitochondrial genomes that differ from each by a factor of ten.

Eukaryotic nuclear genomes range from a minimum size of 12 Mb for the yeast (*Saccharomyces cerevisiae*) genome to the enormous genomes of *Fritillaria* lilies, some of which attain sizes up to 120,000–127,000 Mb (Table 2.3; Brown 2007; Ambrožová et al. 2010). To put this in perspective, the *Fritillaria* lily genomes are 40 times larger than the human genome and 10,000 times larger than the yeast genome! Despite the existence of NGS for more than a decade now, the complete sequences of these largest genomes have still not been fully sequenced. As of September 2015, the species with the largest complete genome sequence in Genbank is the White Spruce (*Picea glauca*), which has a genome size of 26.9 Gb (Table 2.3; Birol et al. 2013). The sizes of eukaryotic genomes are roughly correlated with organismal complexity, as, for example, yeast, fungi, and protists have the smallest genomes while vertebrates and plants have the largest genomes (Table 2.3; Brown 2007; Watson et al. 2014).

TABLE 2.1
Sizes of some prokaryotic genomes

	Size of genome (Mb)
Archaea	
Nanoarchaeum equitans	0.49
Methanococcus jannaschii	1.66
Archaeoglobus fulgidus	2.18
Bacteria	
Mycoplasma genitalium	0.58
Streptococcus pneumoniae	2.16
Vibrio cholerae El Tor N16961	4.03
Mycobacterium tuberculosis H37Rv	4.41
Escherichia coli K12	4.64
Yersinia pestis CO92	4.65
Pseudomonas aeruginosa PA01	6.26
Bacillus megaterium	30

SOURCE: Data from Table 8.3, Page 234, in Brown, T. A. 2007. *Genomes* 3. New York: Garland Science/Taylor & Francis. With permission.

TABLE 2.2
Sizes of mitochondrial and chloroplast genomes

	Type of organism	Genome size (kb)
	Mitochondrial genomes	
Plasmodium falciparum	Protozoan (malaria parasite)	6
Chlamydomonas reinhardtii	Green alga	16
Mus musculus	Vertebrate (mouse)	16
Homo sapiens	Vertebrate (human)	17
Metridium senile	Invertebrate (sea anenome)	17
Drosophila melanogaster	Invertebrate (fruit fly)	19
Chondrus crispus	Red alga	26
Aspergillus nidulans	Ascomycete fungus	33
Reclinomonas americana	Protozoa	69
Saccharomyces cerevisiae	Yeast	75
Suillus grisellus	Basidiomycete fungus	121
Viscum scurruloideum	Flowering plant (mistletoe)	66
Brassica oleracea	Flowering plant (cabbage)	160
Arabidopsis thaliana	Flowering plant (vetch)	367
Zea mays	Flowering plant (maize)	570
Cucumis melo	Flowering plant (melon)	2,500
Silene noctiflora	Flowering plant (catchfly)	6,728
Silene conica	Flowering plant (catchfly)	11,319
	Chloroplast genomes	
Pisum sativum	Flowering plant (pea)	120
Marchantia polymorpha	Liverwort	121
Oryza sativa	Flowering plant (rice)	136
Nicotiana tabacum	Flowering plant (tobacco)	156
Chlamydomonas reinhardtii	Green alga	195

SOURCE: Data from Table 8.5, Page 240, in Brown, T. A. 2007. *Genomes* 3. New York: Garland Science/Taylor & Francis, with permission, except for *Viscum scurruloideum*, *Silene noctiflora*, and *S. conica*, which were obtained from Figure 2 in Skippington, E. et al. 2015. *Proc Natl Acad Sci USA* 112:E3515–E3524.

2.1.2 Composition of Eukaryotic Organellar Genomes

As of October 2015, Genbank contained the complete sequences for over 6,000 mitochondrial and 700 chloroplast genomes. These data are providing us with a clearer picture about variation in the composition of these organellar genomes. There is also at least one peer-reviewed journal (i.e., *Mitochondrial DNA*) that is dedicated to publishing papers focused on the biology of mitochondrial genomes.

Tremendous variation in the composition of mitochondrial genomes exists among eukaryotes. The numbers of genes (RNA- and protein-coding) varies from a low of five in *Plasmodium falciparum* up to at least 92 genes in *Reclinomonas americana* (Table 2.4). Notice that these two protozoans bracket all other eukaryotes in terms of their gene numbers listed in Table 2.4. Thus, the number of genes found in mitochondrial genomes does not relate to the complexity of an organism. Mitochondrial genomes include genes that code for proteins involved in the respiratory complex, ribosomal RNAs, transfer RNAs, and a control region (Table 2.4; Randi 2000). In contrast to nuclear genomes, mitochondrial genomes generally contain few or no introns (Table 2.4). One of the two strands of the mitochondrial genome has a higher G + C

TABLE 2.3
Sizes of various eukaryotic genomes

Species	Type of organism	Genome size (Mb)
Human (*Homo sapiens*)	Vertebrate	3,260
Mouse (*Mus musculus*)	Vertebrate	2,804
Anolis lizard (*Anolis carolinensis*)	Vertebrate	1,799
Chicken (*Gallus gallus*)	Vertebrate	1,047
Zebrafinch (*Taeniopygia guttata*)	Vertebrate	1,232
Zebrafish (*Danio rerio*)	Vertebrate	1,412
Pufferfish (*Takifugu rubripes*)	Vertebrate	392
Fruit fly (*Drosophila melanogaster*)	Invertebrate	164
Nematode worm (*Caenorhabditis elegans*)	Invertebrate	100
Thale cress (*Arabidopsis thaliana*)	Land plant	127
Black cottonwood (*Populus trichocarpa*)	Land plant	417
Maize (*Zea mays*)	Land plant	2,068
White spruce (*Picea glauca*)	Land plant	26,936
Fritillary lily (*Fritillaria assyriaca*)	Land plant	120,000
Yeast (*Saccharomyces cerevisiae*)	Yeast	12
Leishmaniasis parasite (*Leishmania major*)	Protist	33
Malaria parasite (*Plasmodium falciparum*)	Protist	27

SOURCE: Data obtained from Genbank except for the Fritillary lily, which was obtained from Table 7.2, Page 208, Brown, T. A. 2007. *Genomes* 3. New York: Garland Science/Taylor & Francis.

TABLE 2.4
Composition of mitochondrial genomes

Feature	*Plasmodium falciparum*	*Chlamydomonas reinhardtii*	*Homo sapiens*	*Saccharomyces cerevisiae*	*Arabidopsis thaliana*	*Reclinomonas americana*
Total number of genes	5	12	37	35	52	92
			Types of genes			
Protein-coding genes	3	7	13	8	27	62
Respiratory complex	3	7	13	7	17	24
Ribosomal proteins	0	0	0	1	7	27
Transport proteins	0	0	0	0	3	6
RNA polymerase	0	0	0	0	0	4
Translation factor	0	0	0	0	0	1
Functional RNA genes	2	5	24	27	25	30
Ribosomal RNA genes	2	2	2	2	3	3
Transfer RNA genes	0	3	22	24	22	26
Other RNA genes	0	0	0	1	0	1
Number of introns	0	1	0	8	23	1
Genome size (kb)	6	16	17	75	367	69

SOURCE: Reproduced from Table 8.6, Page 241, in Brown, T. A. 2007. *Genomes* 3. New York: Garland Science/Taylor & Francis. With permission.

content than its complementary strand. Thus, the high G + C strand is referred to as the "heavy" or "H-strand," while the low G + C strand is called "light" or "L-strand" (Randi 2000).

The mitochondrial genomes of metazoans are relatively small (<21 kb) and thus many species in this group have had their mitochondrial genomes fully sequenced and annotated (see Boore 1999). With some exceptions (e.g., jellyfish), the composition of the metazoan mitochondrial genome is also remarkably stable. A typical metazoan mitochondrial genome is illustrated by the one found in the hummingbird *Chrysolampis mosquitus* shown in Figure 2.1. The gene composition in this mitochondrial genome includes: 13 protein-coding genes, 2 ribosomal RNA (rRNA) genes, and 22 transfer RNA (tRNA) genes (Figure 2.1). Additionally, metazoan mitochondrial genomes have a noncoding sequence called the *control region*. In vertebrates, this control region is called the "D-loop," while in invertebrates it is called the "AT-rich region" (Palumbi 1996; Randi 2000).

In addition to (usually) being circular, containing few genes and introns, there are other ways that mitochondrial genomes differ from their nuclear counterparts. While each cell only has one to several copies of a nuclear genome present in a nucleus depending on the ploidy level of the species, the same cells have an enormous number of mitochondrial genome copies. For example, human cells contain around 8,000 copies of the mitochondrial genome per cell (Pakendorf and Stoneking 2005). The number of mitochondrial genome copies per cell varies from 1,000 to 10,000 (Brown

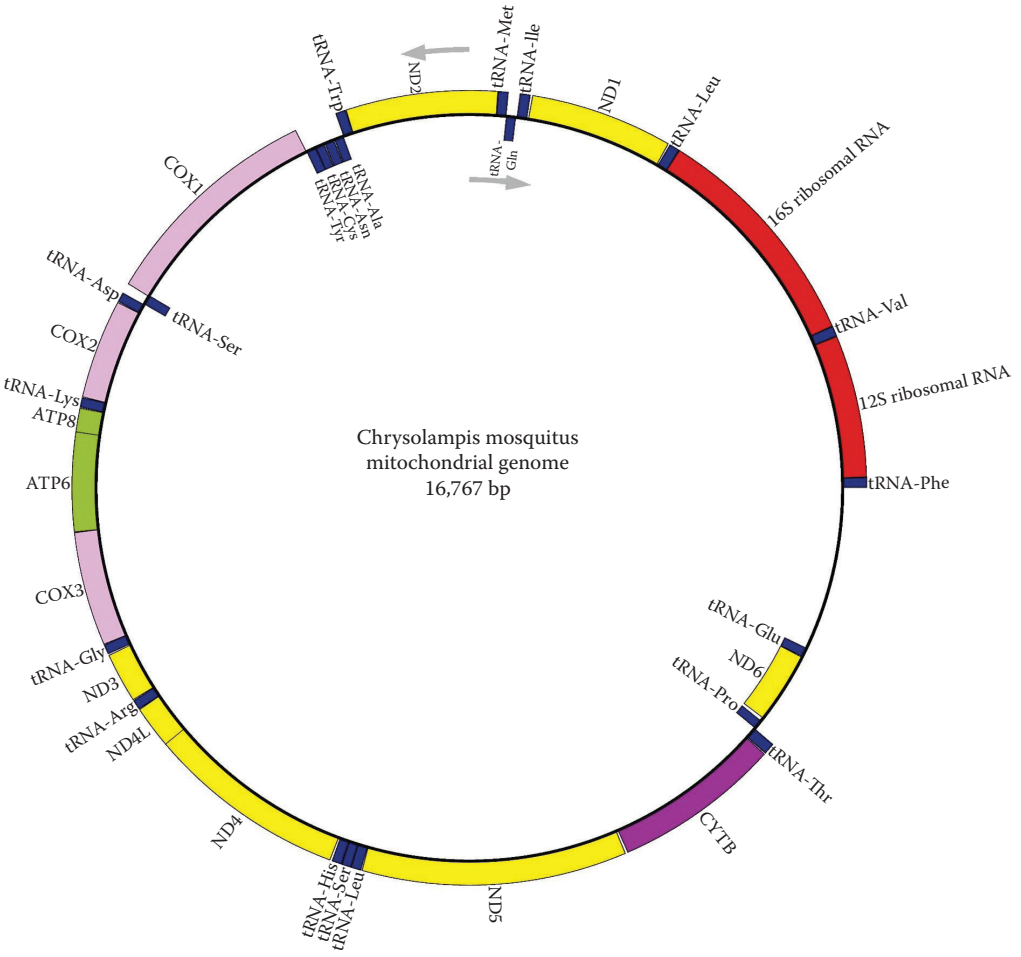

Figure 2.1. Complete mitochondrial genome for the hummingbird *Chrysolampis mosquitus* (Genbank # KJ619585; Souto, H. M. et al. 2014. *Mitochondrial DNA* 27:769–770) Figure was made using the software *OGdraw*. (From Lohse, M. et al. 2013. *Nucleic Acids Res*, doi: 10.1093/nar/gkt289. http://ogdraw.mpimp-golm.mpg.de/.)

2007). The mode of inheritance of mitochondrial genomes also differs from nuclear genomes. For most eukaryotes mitochondrial DNA is inherited strictly via the maternal line though some exceptions have been noted for marine and freshwater mussels that have biparental inheritance (Randi 2000; see review in Galtier et al. 2009). Owing to a high mutation rate, a cell can have a population of mitochondrial genomes that are not all identical in sequence. When more than two distinct types of mitochondrial genome sequences are present in a cell, this condition is called *heteroplasmy* (Randi 2000). However, only the copy present in highest abundance in a cell will be passed to the offspring (Randi 2000). Nuclear and mitochondrial genomes also differ regarding recombination. While recombination is a major feature of nuclear genomes, it is not yet clear if this evolutionary force plays any role in mitochondrial DNA evolution (Wilson et al. 1985; Pakendorf and Stoneking 2005; see review in Galtier et al. 2009). Evidently, though, the effects of any recombination appear to be negligible in mitochondrial genomes and thus researchers have assumed that these molecules are effectively passed from one generation to the next as a single nonrecombined "supergene" or linkage group (Wilson et al. 1985; Randi 2000). Excellent reviews of the properties of mitochondrial DNA can be found in Wilson et al. (1985), Moritz et al. (1987), Randi (2000), Ballard and Rand (2005), Pakendorf and Stoneking (2005), and Galtier et al. (2009).

In addition to not varying much in overall size (i.e., 100–200 kb), the composition of chloroplast genomes also appears to not vary much either (Brown 2007). Each chloroplast tends to show the same set of ~200 genes needed for photosynthesis, which includes proteins, ribosomal RNAs, tRNAs, and some introns (Brown 2007).

2.1.3 Composition of Eukaryotic Nuclear Genomes

Thanks to genome sequencing efforts over the past two decades, we not only have a better idea about the sizes of many nuclear genomes, but our knowledge of the genomic landscapes found across the tree of life has also dramatically improved. Many of these sequenced genomes have been well annotated and therefore we know far more now about the various elements and their representation in genomes. These elements include protein-coding genes and introns, regulatory regions for gene expression (e.g., promoters, enhancers, and noncoding RNAs), transfer and ribosomal RNAs, pseudogenes, various types of repetitive DNA, and intergenic DNA with presumably no function.

2.1.3.1 Gene Numbers and Densities among Nuclear Genomes

The total number of functional genes (i.e., RNA- and protein-coding sequences) per prokaryote genome ranges between 500 and 5,700 genes per genome (Brown 2007). In the human genome, the actual protein-coding genes (exons) occupy only about 48 Mb (1.5%) of the genome (from Figure 7.13 in Brown 2007). In addition to these functional genes, genomes also contain large numbers of gene-related DNA sequences that likely have little or no function such as introns, pseudogenes, and gene fragments. In the human genome, these gene-related elements are found in about 1,152 Mb (36%) of the genome (from Figure 7.13 in Brown 2007).

Within eukaryotes, the size and composition of genomes varies in different ways. As one might expect, both genome size and numbers of genes per genome increase with organismal complexity as the yeast, fruit fly, and human data show in Table 2.5. The numbers of introns also increases in this way. For example, the yeast genome has a total of 239 introns, whereas the human genome has 300,000 (Brown 2007). Similarly, the density of introns per gene is lowest in yeast, a little higher in the fruit fly genome, and highest in the human genome (Table 2.5). The percentage of genomes comprised of *genome-wide repeats*, a type of *repetitive DNA*, also increases with organismal complexity when low complexity eukaryotes are compared to high complexity eukaryotes (Table 2.5). However, when the gene count is examined in light of genome size, gene density shows the opposite trend with respect to organismal complexity. The yeast genome shows the highest gene density while the human genome the lowest (Table 2.5). This phenomenon can be explained by the huge amounts of noncoding DNA such as introns and especially repetitive DNA (Brown 2007).

2.1.3.2 Intergenic DNA

Located in the genomic spaces between RNA- and protein-coding genes is a type of DNA collectively called *intergenic DNA*. Intergenic DNA is a

TABLE 2.5
Genome statistics for three eukaryotes of varying complexity

	Yeast	Fruit fly	Human
Genome size (Mb)	12.1	180	3,200
Approximate number of genes	6,100	13,600	30,000–40,000
Gene density (average number of genes/Mb)	496	76	~11
Average number of introns/gene	0.04	3	9
Percentage of genome with interspersed repeats	3.4%	12%	44%

SOURCE: Data for yeast (*Saccharomyces cerevisiae*), fruit fly (*Drosophila melanogaster*), and human (*Homo sapiens*) are from Tables 7.3 and 7.4, Pages 210–211, Brown, T. A. 2007. *Genomes* 3. New York: Garland Science/Taylor & Francis. With permission.

heterogeneous class of DNA comprised of various types of repetitive DNA and nonrepetitive sequences. About 62% of the human genome is comprised of intergenic DNA most of which (46%) is repetitive DNA while the nonrepetitive fraction represents only about 16% of the genome (from Figure 7.13 in Brown 2007). These numbers may be modified in the future, as a study by De Koning et al. (2011) suggested that the current estimate of 46% repetitive DNA in the human genome might represent an underestimate owing to the possibility that current methods for finding repetitive DNA may lack sensitivity for detection. In other words, the accumulation of mutations over evolutionary time makes the identification of repetitive DNA such as transposable elements (see below) difficult or impossible. If true, then the amount of nonrepetitive DNA in the human genome would be even less than 16%.

The complete human genome sequence not only showed that genic regions comprise a miniscule fraction of the genome, but it also revealed that genes are dispersed throughout the genome much like oases in deserts (Venter et al. 2001). These *gene deserts*, which were defined as genomic regions of at least 500 kb in size that are devoid of genes, were found to occupy ~20% of the genome (Venter et al. 2001). Thus, a significant fraction of the intergenic DNA component of genomes is comprised of these gene deserts. In contrast to the parts of the genome containing RNA- and protein-coding sequences, gene deserts exhibit low G + C content, higher numbers of variable nucleotide sites, and decreased levels of sequence conservation (Ovcharenko et al. 2005).

Comparative genome analyses between species such as human versus mouse and human versus chicken show that intergenic DNA exhibits little sequence conservation suggesting that this genomic component is far less constrained by natural selection than the genic regions for maintaining essential biological functions. However, studies of these vast genomic landscapes are revealing the existence of biologically functional intergenic DNA and more such discoveries will undoubtedly be made in the future. It is also highly probable that vast amounts of intergenic DNA will always be considered nonfunctional. This is because careful consideration of the basic biology of certain genomic sequences (e.g., transposable elements and pseudogenes) leads one to the inescapable conclusion that most if not all of these sequences must be currently without a biological function (Graur et al. 2013). We will further explore the topic of functional versus nonfunctional DNA and its relevance to phylogenomics in Chapter 3.

Repetitive DNA—This class of DNA is comprised of two main groups: *tandemly repeated DNA* and *genome-wide repeats* (Brown 2007). Tandem repeats are found adjacent to each other or in clusters of repeats in close proximity to each other on the same chromosomes, whereas genome-wide repeats are dispersed randomly throughout genomes.

Different types of simple tandem repeats have been classified according to the overall length of the repeat sequence: (1) long stretches of *satellite* DNA, which span hundreds of kb and are largely confined to centromeric, telomeric, and subtelomeric regions; (2) *minisatellites*, which extend up to 20 kb along a chromosome and can occur throughout the genome; and (3) *microsatellites*, which are up to around 150 bp long and are also scattered across eukaryotic genomes (Scribner and Pearce 2000; Brown 2007). Most such repeats occur within intergenic regions and thus they are

generally free to mutate without any positive or negative consequence to the individual organism (Scribner and Pearce 2000). However, some microsatellites occur within protein-coding regions and can therefore lead to diseases (e.g., CAG repeats excessive in number within the *huntingtin* gene cause Huntington's Disease; Ashley and Warren 1995).

Genome-wide repeats, which are also called *transposons*, *transposable elements*, *mobile genetic elements*, and "*jumping genes*" consist of gene families with members dispersed throughout the genome. The process of one transposable element moving from one location in a genome to another is called *transposition*. Barbara McClintock discovered transposons in the 1940s during her research on maize genetics (McClintock 1950) and was subsequently awarded a Nobel Prize for this work (Ravindran 2012).

Transposable elements are largely inserted into random locations within genomes and thus they are common in intergenic regions; they can also be inserted within functional genes and other transposons. When transposons insert themselves into functionally important regions such as regulatory regions or in the reading frame of proteins they can alter gene expression patterns or severely disrupt gene function and cause diseases (Watson et al. 2014). Some transposable elements simply represent a single element that "jumps" around the genome to different chromosomal locations, whereas other types of transposons copy themselves and thereby proliferate in number within a genome.

Interestingly, much of the size variation among eukaryotic genomes is not due to the variable numbers of genes or other important functional elements of the genome, but is instead explained by varying amounts of genome-wide repeats (Brown 2007). For example, among vertebrates genome-wide repeats only account for about 3%–13% of the reptilian genome (including birds), whereas they represent 35%–50% of the mammalian genome (Janes et al. 2010). An even more dramatic range is observed in plants, as genome-wide repeat content in the carnivorous bladderwort plant (*Utricularia gibba*) genome is only 3% while for maize it approaches 85% (Lee and Kim 2014)!

Nonrepetitive intergenic DNA—When the genic and repetitive DNA fractions of the genome are accounted for, the remaining portion is referred to as *nonrepetitive intergenic DNA*. As previously mentioned, approximately ~16% (510 Mb) of the human genome is comprised of nonrepetitive intergenic DNA (Brown 2007). Like repetitive DNA, a major hallmark of nonrepetitive DNA is its overall lack of evolutionary conservatism. However, researchers are finding exceptional elements within these genomic regions that are evidently of vital importance to the organism. For example, not long after the human genome sequence was published, researchers discovered a whole new class of noncoding and nonrepetitive DNA that is biologically important. These "ultraconserved" elements, which have had their sequences essentially "frozen in time" (i.e., without modification) for hundreds of millions of years, apparently function as long-range regulatory elements for neighboring developmental genes (Bejerano et al. 2004; Katzman et al. 2007). Other interesting findings concerning the genomic structure and function involving noncoding/nonrepetitive DNA have been discovered as we will see in Chapter 3.

Gene deserts—Although gene deserts in general show low levels of sequence conservation, they exhibit variability in terms of their composition and evolution. This observation prompted Ovcharenko et al. (2005) to characterize two types of gene deserts, which are called *stable gene deserts* and *variable gene deserts*.

Stable gene deserts have >2% of their sequences conserved, are enriched for long-range regulatory elements, are depauperate in transposable elements, and usually exist as single syntenic blocks maintained since the time of divergence between mammalian and avian lineages (Ovcharenko et al. 2005). The preservation of these large syntenic blocks over evolutionary time is intriguing especially given the overall lack of sequence conservatism exhibited by these sequences. However, research findings from different studies are now painting a picture that suggests these syntenic blocks—which include the regulatory elements and flanking sequences—must have their structure maintained in order to preserve biological functions involving the regulation of developmental processes. First, these regulatory elements are highly conserved sequences maintained by purifying natural selection (Katzman et al. 2007). Secondly, duplications of these regions are apparently not tolerated (Bejerano et al. 2004). Thirdly, the regions flanking these regulatory elements are refractory to transposon insertions (Simons et al. 2006). The human genome contains ~3 million transposable elements, which corresponds to

one transposon every ~500 bp on average. Given this observation, it is surprising that 860 transposon-free regions (TFRs), which are defined as transposon free segments that are at least 10 kb long (longest is 81 kb), have been found (Simons et al. 2006). These TFRs are strongly associated with the distribution of highly conserved regulatory elements in stable gene deserts. Although some transposable elements are found in stable gene deserts, their syntenic positions have been maintained since the human and mouse lineages diverged from each other even though the actual transposons are lineage-specific (i.e., independently acquired; Simons et al. 2006).

In contrast, *variable gene deserts* have <2% of their sequences conserved are depauperate in regulatory elements, are enriched in transposable elements, and do not have syntenic blocks (Ovcharenko et al. 2005). Thus stable gene deserts appear to represent critically important genomic structures that are not able to tolerate transposon insertions and duplications, whereas variable gene deserts appear to have few if any selective constraints and thus might represent true genomic junkyards (Ovcharenko et al. 2005).

2.2 DNA SEQUENCE EVOLUTION

2.2.1 Patterns and Processes of Base Substitutions

Mutation occurs via different processes and at different levels in genomes: chromosomes can break into smaller pieces or they can fuse with each other, pieces of chromosomes can be inverted or translocated, mobile genetic elements mutate chromosomes by inserting themselves into random genomic locations, crossing over during meiosis alters DNA sequences, and, at the smallest scale, single base substitutions, insertions, and deletions change DNA sequences. It is essential for phylogenomics researchers to attain a solid understanding about patterns and processes of DNA mutations—particularly germline mutations—because this information is used to obtain improved estimates of gene tree topologies, evolutionary distances, and divergence times among sequences (Wakeley 1996; Yang and Yoder 1999; Graur and Li 2000).

There are two different ways DNA sequences can mutate at the level of a DNA site or short string of sites: (1) *site substitutions* or "point mutations" in which one nucleotide base is switched to another type and (2) *insertions or deletions* or "indels," which are caused when one or more consecutive nucleotide sites are added or deleted by mistake during the DNA replication process. Although both types of mutations represent potentially useful "information" in a sequence dataset, historically, the vast majority of molecular phylogenetic studies have analyzed sequence variation due to site substitutions. The primary reasons for this is that DNA sequences contain many more point mutations than indels and because it is more straightforward to model the base substitution process than modeling indel evolution.

There are a total of six distinct types of nucleotide substitutions, which are divided up into two classes: *transitions* and *transversions* (Figure 2.2). Two of these substitution types represent transitions, which occur when one purine changes into another purine (A ↔ G) or when a pyrimidine changes to another pyrimidine (C ↔ T), while the remaining four types of substitutions classed as transversions occur when a purine changes into a pyrimidine or vice versa (Figure 2.2).

2.2.1.1 Transition Bias

What are the patterns and processes of base substitution observed in DNA? This is a complicated subject with research still ongoing to better understand this aspect of molecular evolution. Nonetheless, we can still briefly explore this subject in order to obtain some understanding about the nature of base substitutions. First, let's consider the question: *Is DNA substitution a random process with respect to the directionality of change* (i.e., *each substitution*

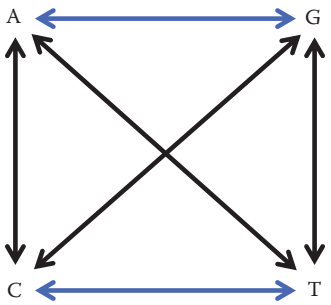

Figure 2.2. Diagram showing all six types of base substitutions. *Transitions* (blue arrows) are A ↔ G or C ↔ T changes, whereas *transversions* (black arrows) are A ↔ C, G ↔ T, A ↔ T, or G ↔ C substitutions.

is equally likely)? Before we can address this question, we need to define a relevant quantity, which is called *transition bias* or the transition:transversion rate ratio (Wakeley 1996). This important ratio is simply the estimated overall transition rate relative to the overall transversion rate and it is commonly abbreviated as *TI/TV rate ratio* (Wakeley 1996). If substitutions randomly occur, then we would expect to see a *TI/TV* rate ratio of 1:2 in real data (Graur and Li 2000). Although we do believe that DNA substitutions occur randomly with respect to the evolutionary fitness of the individual, an abundance of evidence and theory (as we will see later in this section and in the following Section 2.2.1.2) indicate that these mutations are not random when we are talking about the direction of change (Graur and Li 2000).

Brown et al. (1982) examined primate mitochondrial DNA and discovered that transitions vastly outnumbered transversions. These authors dismissed the hypothesis that natural selection was the sole explanatory factor to explain their observations because the *TI/TV* rate ratio bias was observed in tRNAs and in protein-coding genes each of which has different selection pressures. Instead, they attributed the high number of transitions to the mutation process itself. The "mutation process" Brown et al. (1982) referred to is the one that produces *spontaneous substitutions* or point mutations. Note that these substitutions only become mutations after they become fixed changes in the genome that are passed to the next generation (Graur and Li 2000). Substitutions can become so fixed if they either escape the cellular repair network or if they have no consequence to an individual's fitness (i.e., selectively neutral mutation). A later study by Tamura and Nei (1993) also found a preponderance of transitions relative to transversions in the mitochondrial control region in humans and chimpanzees, which represented a significant finding because control region sites are thought to experience little or no selection and hence the observed *TI/TV* ratio may reflect the rates of spontaneous substitutions in mitochondrial DNA (Graur and Li 2000).

High transition rates have also been found in nuclear DNA but the situation is more complicated and not fully understood. First, there is emerging evidence that rates of spontaneous substitution leading to transitions and transversions not only vary within genomes (Graur and Li 2000) but also among species (Yang and Yoder 1999; Keller et al. 2007). At the within genome level, research has shown that some bases are more mutable than others (Graur and Li 2000). For example, parts of the human genome that are rich in G + C bases (e.g., pseudogenes) display a *TI/TV* ratio of around five, which is 2.5-fold higher than the estimate observed for surrounding intergenic regions, which tend to be G + C poor regions (see Table 2.1 in Zhang and Gerstein 2003). In G + C rich sequences of mammalian genomes, a high percentage of CG dinucleotides (not to be confused with base pairs) or *CpG* sites ("CpG" represents—C—phosphate—G—on the same strand of DNA) experience a high rate of mutation into TpG dinucleotides, which occurs because the cytosines in CG dinucleotides are targets for methylation; methylated cytosines, in turn, induce C:G → T:A transitions (Petrov and Hartl 1999; Graur and Li 2000). Genomic regions that no longer contain many CpG sites such as "old" pseudogenes that have already undergone transition mutations at original CpG sites (Casane et al. 1997; Zhang et al. 2002) and intergenic DNA with no function (i.e., low G + C content; Graur and Li 2000) show *TI/TV* rate ratios, which likely reflect the rates of spontaneous substitutions due to DNA replication errors.

Studies to date suggest that the spontaneous substitution rate may not be fixed across species. For example, as was mentioned the genome-wide *TI/TV* rate ratio for humans was estimated to be about 2:1, a figure was that based on the analyses of CpG discounted pseudogenes (Zhang and Gerstein 2003). Costa et al. (2016) obtained a similar estimate for intergenic regions in the genomes of hominoids. However, in nonvertebrates the *TI/TV* rate ratio may vary. Petrov and Hartl (1999) observed a transition bias in nonfunctional parts of the *Drosophila* genome that is comparable to the mammalian ratio of two, whereas a later study on grasshoppers by Keller et al. (2007) observed a lower *TI/TV* rate ratio. These results for insect genomes suggest that the *TI/TV* rate ratio may vary across species more than is currently known.

Thus, spontaneous substitutions can originate during DNA replication or via nonreplicative processes (Graur and Li 2000; Smeds et al. 2016). Which process accounts for more germline point mutations? It has been believed that errors in DNA replication account for most spontaneous substitutions (Graur and Li 2000; Watson et al. 2014). Recent evidence from birds corroborates

this hypothesis, as Smeds et al. (2016) found that only ~13% of detected germline point mutations in an avian pedigree were explained by CpG site transitions.

2.2.1.2 Transition Bias and DNA Replication Errors

The extreme fidelity of DNA replication is one of the remarkable wonders of nature. Nonetheless, we know that single base errors do occasionally occur during replication, which results in commonly observed transition biases. *What causes these single base misincorporations and why are they generally nonrandom with respect to their directionality?* The molecular mechanism(s) that give rise to spontaneous substitutions have not been conclusively determined. However, the long-lived *rare tautomer hypothesis*, which was conceived by Watson and Crick (1953a), is currently the favored explanation for this mutation process (Graur and Li 2000; Harris et al. 2003; Wang et al. 2011; but see Echols and Goodman 1991). Although it remains to be seen if this hypothesis is correct, at the present time it nonetheless offers a simple and elegant explanation for why transition bias exists in genomes. Before we examine this hypothesis in detail, we will briefly review aspects of DNA replication fidelity.

The fidelity of DNA replication is largely the result of so-called "Watson–Crick geometry," which characterizes the geometrical configuration of the correct base pairings (i.e., when adenine base is paired with thymine or when guanine is paired with cytosine) during DNA synthesis (Watson and Crick 1953b; Kunkel and Bebenek 2000; Harris et al. 2003; Watson et al. 2014). The correct geometry is only achieved when a purine (adenine or guanine) is paired with a pyrimidine (cytosine or thymine)—thereby creating the correct spacing between the opposing bases for hydrogen bonding—and a sufficient number of hydrogen bonds are formed (i.e., two for the A:T pair and three for the G:C pair; Watson and Crick 1953b). However, the purine–pyrimidine pairing requirement alone is insufficient to explain the specificity of base pairing. Thus, it is likely the hydrogen bonding between the two opposed bases that best accounts for this phenomenon (Watson et al. 2014).

Watson and Crick (1953a) suggested that spontaneous base substitutions could arise when rare or "disfavored" minor tautomers of each natural base occasionally become incorporated into growing DNA chains during replication. The four bases that comprise DNA exist in two tautomeric states: the *major* (amino and keto) tautomeric states represent the normal or preferred base configurations, whereas the *minor* (imino and enol) states are rarely formed (Watson et al. 2014). These alternative tautomeric bases differ

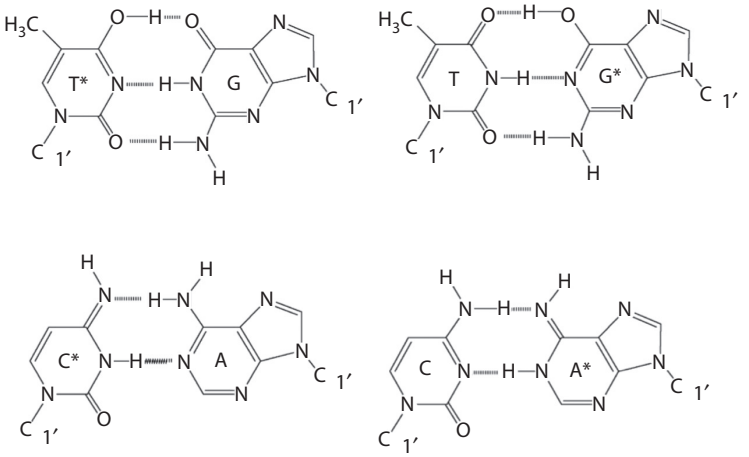

Figure 2.3. Hydrogen-bonding arrangements in pseudo-Watson–Crick base pairs. This figure shows how the minor tautomers (flagged by asterisks) of the natural bases form hydrogen bonds with complementary natural bases. Top: T*:G and T:G* pairs show the enol:keto and keto:enol pairings, respectively. Bottom: C*:A and C:A* pairs depict the imino:amino and amino:imino pairings, respectively. Hydrogen bonds are shown as hatched lines between the functional groups of opposing bases. (Reprinted from Harris, V. H. et al. 2003. *J Mol Biol* 326:1389–1401. With permission.)

from each other owing to changes in the attached functional groups, which in turn, alters their hydrogen-bonding properties (Figure 2.3). This means that the four minor tautomers will not form the usual G:C and A:T or Watson–Crick base pairs, but instead they will preferentially form G:T and C:A or "pseudo Watson–Crick" base pairs (Figure 2.3; Harris et al. 2003). Examination of the hydrogen-bonding capabilities in all possible base pairings involving the four minor base tautomers (not shown) makes clear why the pseudo Watson–Crick pairings shown in Figure 2.3 are most likely to occur during DNA replication. In other words, the pairings that satisfy the purine–pyrimidine pairing rule and maximum number of hydrogen bonds will be the stereochemically preferred pairings during DNA replication. Because the geometrical configurations of pseudo Watson–Crick pairings (Figure 2.3) mimic the traditional Watson–Crick pairings, DNA polymerase cannot discriminate the correct versus incorrect base pairs, respectively, thereby leading to occasional misincorporations during DNA replication (Harris et al. 2003). An important consequence of the rare tautomer hypothesis is that spontaneous mutations arising from incorporations of minor tautomers during DNA synthesis will result in the establishment of *transition*-type mutations in the genome.

2.2.1.3 Saturation of DNA Sites

If we were to assume that the rate of base substitution has remained constant through time, then we might naively expect that the total number of substitutions observed in two aligned homologous sequences will be linearly related to the amount of time since their most recent common ancestor or divergence time. This would be true for sequences with recent divergence times, but not for sequences with older divergence times (Upholt 1977; Brown et al. 1979, 1982). Why is this so? For sequences with older divergence times, the true total amount of evolutionary divergence will be underestimated because there will be some DNA sites that experienced multiple substitutions; hence in those situations counts of *observed* substitutions will be blind to previous unobserved substitutions (Upholt 1977). This phenomenon is known as *saturation* (Brown et al. 1982; Swofford et al. 1996; Arbogast et al. 2002) or *multiple hits* (Yang 2006) and it is of great importance to phylogenomics.

Let's take a closer look at how saturation occurs and how it can complicate efforts to estimate DNA sequence divergences. Consider a single 50 bp long gene sequence, which represents a portion of the open reading frame (ORF) for a protein-coding gene. If we sequence this locus in two different species and compare the sequences site by site, we can see that 15 of the sites are variable (Figure 2.4). Variable sites are also called *segregating* or *polymorphic sites*, whereas the sites showing no variation are labeled as *nonsegregating* or *monomorphic sites*. While looking at these two sequences in Figure 2.4, we would like to know how much evolutionary divergence has actually taken place between them. In other words, we would like to ask the question: *How many mutations have occurred between these sequences since the time of their most recent common ancestor?*

The simplest method for estimating this divergence between a pair of sequences involves calculating the *pairwise* or *p-distance*, which is expressed either as the number of substitutions/site or by percent sequence divergence (Swofford et al. 1996). The *p*-distance is equal to the total number of segregating sites divided by the total number of sites. From Figure 2.4, we can compute the *p*-distance as being 15/50 = 0.3 substitutions/site (or 30% sequence divergence). However, remember that this *p*-distance ignores any unobserved

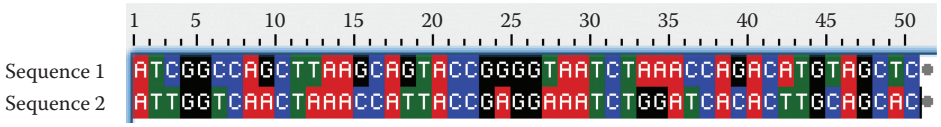

Figure 2.4. Alignment of two protein-coding DNA sequences spanning 50 sites obtained from two different species. Sequences are shown in a 5′ → 3′ orientation and site #1 at the 5′ end corresponds to the start of the reading frame (i.e., is 1st codon position). Fifteen of the sites are variable and 35 are invariable. The four bases are color-coded for easier visualization of sequence similarities and differences. Alignment was constructed using the sequence alignment software *Se-Al*. (From Rambaut, A. 2007. *Se-Al, version 2.0 a11*. Edinburgh: The University of Edinburgh.)

substitutions that could be hidden in these sequences. How bad can saturation effects be on estimates of genetic divergence?

Brown et al. (1979, 1982) provided empirical evidence for saturation effects from their studies of mitochondrial DNA evolution in primates (Figure 2.5). Notice that during the initial period of divergence—up to about 10 million years, the relationship is roughly linear (Figure 2.5). This initial phase of sequence divergence is characterized by having each new substitution occurring mostly at monomorphic sites, which generates new segregating sites. Thus, during this linear phase the estimated *p*-distance approximates the true number of mutations that occurred between two recently diverged sequences. However, beyond 10 million years (i.e., >20% sequence divergence), the rate of mutation accumulation begins to slow down (Figure 2.5). This is expected to occur because as time progresses there will be fewer and fewer available nonsegregating sites that can be subject to mutation for the first time and some polymorphic sites may revert to being monomorphic again. Consequently, the probability for multiple mutations within sites increases. By the time the sequences have exceeded 30% divergence, the curve in Figure 2.5 is flat, which implies that some factor (e.g., natural selection) is constraining the sequences from diverging further. If the sequences are completely unconstrained to evolve, then they can potentially continue to diverge from each other until they reach a maximum divergence of 75%, which means they are effectively random with respect to each other (Felsenstein 2004). Thus, Brown et al. (1979, 1982) results provided clear evidence that *p*-distances will underestimate the true distances for older divergences, which is in accord with expectation. The curve in Figure 2.5 is commonly referred to as a *saturation plot* and, as we will soon see, such plots vary depending on which sites are being considered.

Brown et al. (1982) observed another interesting aspect of DNA sequence evolution that concerns the effects of saturation on *TI*/*TV* rate ratios. These authors noticed that *TI*/*TV* estimates declined as the evolutionary distances between sequences increased (i.e., "time-dependency" of *TI*/*TV*). They explained this phenomenon by suggesting that, owing to multiple substitutions within sites, the number of observed transitions departs more and more from the true number with increased divergence times while the number of observed transversions remains close to the true number (i.e., do not reach point of saturation). Thus it is primarily an estimation problem caused by saturation and not due to any actual time-variable substitution rates.

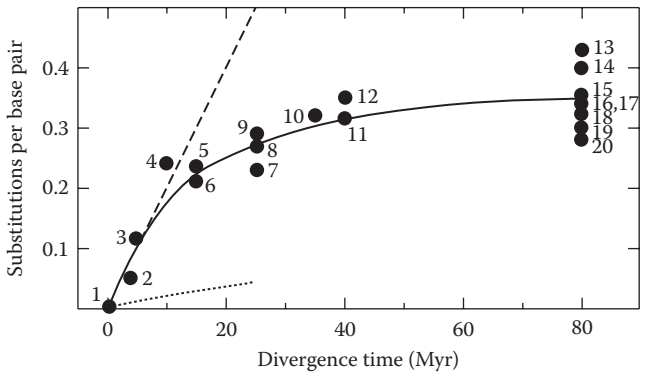

Figure 2.5. DNA sequence saturation illustrated by mammalian mtDNA divergences. This graph shows the relationship between mitochondrial *p*-distances (ordinate) generated for various mammal species versus the estimated divergence times in millions of years (Myr) between each pair. Divergence time estimates are based on fossil and protein data. The following *p*-distances are shown: (1) mean difference among humans (only intraspecific comparison on figure); (2) goat versus sheep; (3) human versus chimpanzee; (4) baboon versus rhesus; (5) guenon versus baboon; (6) guenon versus rhesus; (7) human versus guenon; (8) human versus rhesus; (9) human versus baboon; (10) rat versus mouse; (11) hamster versus mouse; (12) hamster versus rat; and (13–20), rodent versus primate species pairs. The broken line along the initial part of the curve represents the inferred mtDNA substitution rate. For comparative purposes, the substitution rate for single-copy nuclear DNA, which was obtained from an outside data source, is shown on the figure as a dotted line. (Reprinted from Brown, W. M., M. George, and A. C. Wilson. 1979. *Proc Natl Acad Sci USA* 76:1967–1971. With permission.)

2.2.1.4 *Among-Site Substitution Rate Variation*

Variation in nucleotide substitution rates *among* genomic sites represents another aspect of DNA sequence evolution that is of concern to phylogenomics. This is because ignorance of this phenomenon by researchers can lead to underestimates of the actual number of substitutions that had occurred (Golding 1983; Wakeley 1994; Swofford et al. 1996; Yang 1996). Yang (P. 370; 1996) described the phenomenon as follows: "*The existence of among-site rate variation means that most evolutionary changes occur at only a few sites, while many other sites never experience any substitutions.*"

An early empirical illustration of this comes from the study of Brown et al. (1982). When these authors partitioned the mitochondrial coding sequences by codon position and then made separate saturation plots for the 1st, 2nd, and 3rd codon positions, they observed three different saturation curves (Figure 2.6). The curve showing divergence among 3rd codon position bases exhibits the most rapid rate of divergence, 1st codon position bases show the second fastest rate, and the 2nd codon position shows the slowest rate of change (Figure 2.6). These curves make a lot of sense when you consider the evolutionary constraints against mutations in certain codon positions. We know that mutations at third codon positions are tolerated far more because they often do not result in amino acid replacements (because of the "wobble" phenomenon). Such mutations are referred to as *synonymous* or *silent substitutions*. If a mutation does cause an amino acid replacement, then it is called a *nonsynonymous* substitution. Nonsynonymous substitutions are much less tolerated because they can alter the protein structure in ways that result in less or no functionality and hence the mutation will likely be weeded out of the population via purifying natural selection.

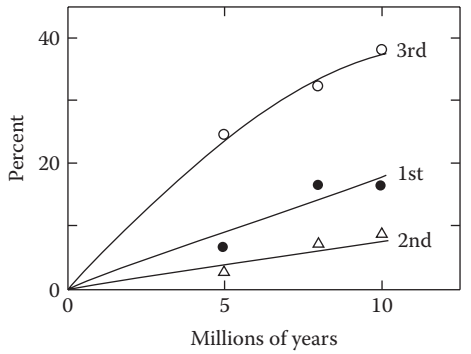

Figure 2.6. Codon bias in primate mtDNA protein-coding sequences. The graph shows observed percent sequence divergences (*p*-distances expressed as % values) for 1st, 2nd, and 3rd codon positions as a function of divergence time. Sequence divergences were calculated from an 896 bp sequence containing two coding regions (URF 4 and 5) from five primates and divergence times were obtained from an independent data source. (Reprinted from Brown, W. M. et al. 1982. *J Mol Evol* 18:225–239. With permission.)

The saturation plots for codon positions in Brown et al. (1982) provide a nice example of sites showing varying levels of *evolutionary conservatism*. Any sites in the genome that are under selection are expected to show such conservatism, while sites that are not maintained by natural election (directly or indirectly via linkage) are expected to show little among-site rate variation (Yang 1996). Evidence supporting the latter idea comes from the study of Costa et al. (2016) who found that nearly all of the 292 genealogically independent and presumably neutral loci in the hominoid genomes were best fit to substitution models having one substitution rate among sites.

Why is among-site rate variation a concern for us? It is huge concern for at least two reasons. First, distances that are estimated from sequences containing sites with varying substitution rates can yield badly underestimated distances (Golding 1983; Yang 1996). Secondly, *TI/TV* rate ratios—essential parameters in many nucleotide substitution models—can also be underestimated (Wakeley 1994, 1996). Moreover, the larger the actual distances, the larger will be the underestimates (Yang 1996). These adverse effects can have severe consequences for estimates of divergence times, *TI/TV* rate ratios, and gene tree inferences (Swofford et al. 1996; Yang 1996).

2.2.2 Tandemly Repeated DNA Sequences

This class of repetitive DNA consists of chromosomal elements that have undergone duplications via *unequal crossing over* and *DNA replication errors* (Brown 2007). Tandem repeats have the characteristic of being spatially proximal to each other along a chromosome. Repeat units can be as small as a single base or long segments of DNA.

Unequal crossing over can occur between two homologous chromosomes misaligned to each other during metaphase I of meiosis. The recombinant products of unequal crossing over include a chromosome containing a duplicated region

while the other homolog has this region deleted. The generation of new copies on a chromosome via unequal crossing over is one of several molecular processes that can give rise to *gene duplications*. When clusters of the same genes are produced via unequal crossing over, this results in the generation of a *gene family*.

There are several different consequences of such duplications when they involve functional genes. First, the new copies may simply contribute more of the same gene products, which may be deleterious, benign, or advantageous to the individual organism. Secondly, one or more of the new copies may evolve a novel function (Chang and Duda 2012). A third possibility is that the gene or one of its upstream regulatory elements might suffer a point mutation that renders the gene nonfunctional. Such genes that are inactivated by mutations are called *conventional pseudogenes* (Brown 2007). A conventional pseudogene can also result if the duplicated gene is missing part of its sequence (i.e., a gene fragment) as a result of unequal crossing over (Watson et al. 2014).

Another molecular mechanism by which tandem repeats can form is via "slippage" of DNA polymerase enzymes during replication (Levinson and Gutman 1987). Polymerase slippage is a phenomenon whereby the polymerase accidently inserts or deletes short stretches of DNA, which often correspond to the repeated motif already found in a sequence. For example, consider a stretch of sequence that has a string of 10 consecutive GC repeats (i.e., 5′-GCGCGCGCGCGCGCGCGC-3′ or "$(GC)_{10}$" repeat). Thus, after a DNA replication error, the above sequence could become a $(GC)_{11}$ sequence. There are many different repeats motifs found in genomes such as GC, AC, GT, CAG, AGCT, etc. Although DNA polymerase slippage may be the primary mechanism to initiate the formation of repeated nucleotides or motifs, once a short tandem repeat sequence is made, the conditions will exist for unequal crossing to dramatically increase the sizes of these repeat sequences (Levinson and Gutman 1987).

Microsatellite loci have the highest mutation rates for any genomic sequences, which is why such loci tend to have high allelic diversity in populations (Scribner and Pearce 2000). This attribute has thus made them the marker of choice for identifying individuals in criminal forensic studies and behavioral ecology studies (e.g., parentage studies). They are also excellent markers in some types of population genetic studies such as analyses of population structure (Waples and Gaggiotti 2006). However, microsatellites are poor markers for reconstructing gene trees because of significant problems with homoplasy; that is, it may be difficult to determine whether two identical allele sequences are related by common ancestry or through convergence (Scribner and Pearce 2000).

2.2.3 Transposable Elements

The transposition of DNA elements in genomes is a form of *site-specific* or *nonhomologous* recombination (Watson et al. 2014). There are several different transposons each with a different mechanism of transposition. Some transposons are *nonreplicative* meaning that the transposable element is enzymatically excised from its chromosomal location and then reinserted into a different genomic location. This form of transposition is called *cut and paste transposition* (Watson et al. 2014). Another form of transposition is *copy and paste transposition* or *replicative transposition* (Watson et al. 2014). Figure 2.7 shows examples of DNA transposons that use the "cut and paste" or "copy and paste" modes of transposition. Other types of transposons involve an mRNA intermediate stage to transpose. These *RNA transposons*, which are also called *retrotransposons* or *retroelements*, only use the replicative form of transposition. There are two main groups of RNA transposons: (1) *LTR retrotransposons* for "*long terminal repeat retrotransposons*" and (2) *non-LTR retrotransposons* (Watson et al. 2014).

Figure 2.8 shows an example of a DNA transposon, an LTR retrotransposon, and a non-LTR retrotransposon. Let's now take a look at the structure of these three elements. DNA transposons (Figure 2.8a) have several critical entities required for their transposition. First, DNA transposons contain the protein-coding gene for *transposase*, the enzyme that performs the transposition reaction. In addition to the *transposase* gene, the element has two terminal inverted repeats. These terminal repeats, which are also called recombination sites, are essential to transposition because they contain the recognition sequences for *transposase* (Figure 2.8a). The inverted repeats of the element will connect the element to the host chromosome once transposition is completed.

The structure of an LTR retroelement (Figure 2.8b) consists of two LTR sequences that flank a region coding for the enzymes *integrase* and *reverse*

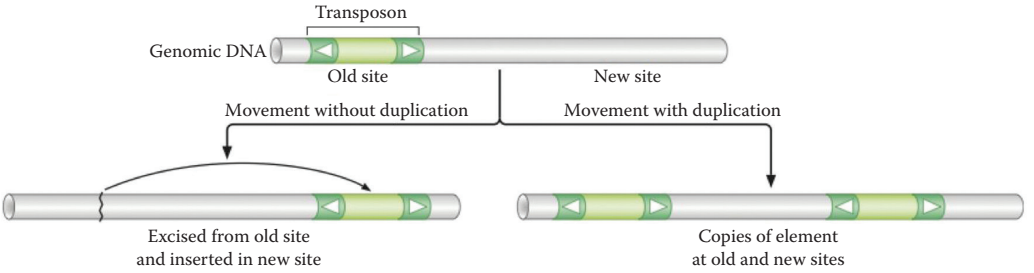

Figure 2.7. "Cut and paste" versus "copy and paste" transposition. Left side of figure shows cut and paste or nonreplicative transposition in which a transposon is excised from its host DNA and reinserted into the host genome at another location. Right side shows copy and paste or replicative transposition whereby a copy of a transposon is re-inserted into the host genome at a different location while the original copy remains in place. (Watson, James D.; Baker, Tania A.; Bell, Stephen P.; Gann, Alexander; Levine, Michael; Losick, Richard, *MOLECULAR BIOLOGY OF THE GENE*, 7th Ed., ©2014. Reprinted by permission of Pearson Education, Inc., New York, New York.)

transcriptase. The retrotransposon *integrase* enzyme functions much like *transposase*. Single inverted terminal repeats are found within each LTR and a promoter sequence for RNA polymerase is found within the upstream inverted terminal repeat (Figure 2.8b). LTR retrotransposons are sometimes called "virus-like retrotransposons" because they are likely derived from retroviruses (Watson et al. 2014). Thus, if a retrovirus invades a cell and successfully integrates its transposon genes into the host genome, then cellular RNA polymerases can produce many copies of mRNAs

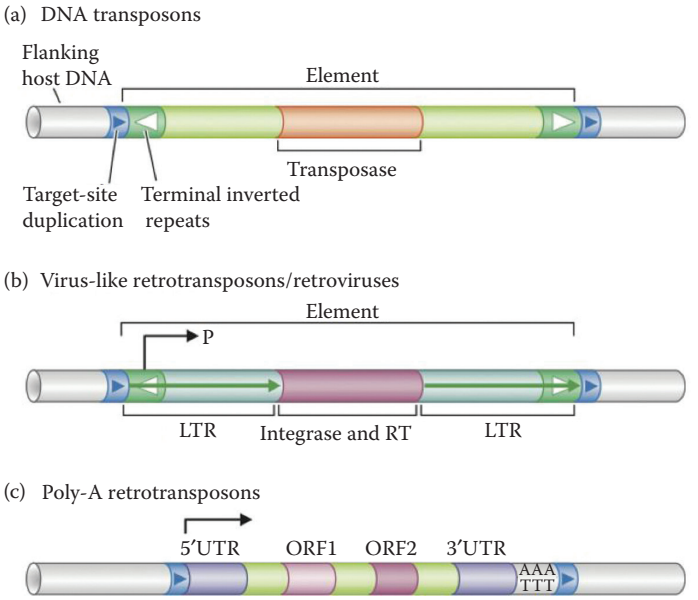

Figure 2.8. Structure of three different types of transposable elements. (a) DNA transposon elements include a *transposase* gene, which is flanked by terminal inverted repeat sequences (green and white arrows). The repeat sequences contain the recombination sites needed for transposition. (b) Virus-like retrotransposon elements contain coding regions for the enzymes *integrase* and *reverse transcriptase* and are flanked by two LTR sequences. (c) Poly-A retrotransposons or non-LTR retrotransposons elements contain coding sequences for an RNA-binding enzyme (ORF1) and an enzyme with both *reverse transcriptase* and *endonuclease* activities (ORF2). The 5′ and 3′ ends of the element are UTR sequences. (Watson, James D.; Baker, Tania A.; Bell, Stephen P.; Gann, Alexander; Levine, Michael; Losick, Richard, *MOLECULAR BIOLOGY OF THE GENE*, 7th Ed., ©2014. Reprinted by permission of Pearson Education, Inc., New York, New York.)

from the retroviral element, which are then converted into *cDNAs* or "copyDNA" before they are integrated into the host genome (Watson et al. 2014). Such host genome copies of a retrovirus are known as *endogenous retroviruses* or "ERVs" (Brown 2007). LTR retrotransposons are not only found in the genomes of plants, invertebrates, and vertebrates (Brown 2007), but this class of transposon accounts for much of the variation in genome sizes observed in plants (Ambrožová et al. 2010; Lee and Kim 2014).

Non-LTR retrotransposons (Figure 2.8c) do not have inverted terminal repeats, but they do have a coding region separated into two ORFs that code for the enzymes *reverse transcriptase* and an *integrase*-like endonuclease that performs the actual transposition. Also, fully functional LTR retrotransposons contain a promoter sequence upstream of the coding region (in a 5′ UTR or untranslated region). A 3′ UTR is located downstream of the two ORFs and the element has a poly-A string of bases on the nontemplate strand (Figure 2.8c). Although some of the details pertaining to the mechanism of transposition for LTR and non-LTR retrotransposons differ, both involve an mRNA intermediate that is produced by cellular RNA polymerases and both are reverse transcribed into a cDNA by *reverse transcriptase*. When transposition is completed, both cDNA elements reside in the host chromosome.

Non-LTR retrotransposons are sometimes called "poly-A retrotransposons" because they have a poly-A tail at the 3′ of the nontemplate DNA strand of the element (Watson et al. 2014). One family of non-LTR repeats, which are common in vertebrate genomes, consists of the *long interspersed nuclear elements* (LINEs; Watson et al. 2014). LINEs are a type of *autonomous* non-LTR retrotransposon because they encode proteins needed for reverse transcription and reintegration of LINE copies into the host genome (Watson et al. 2014). In contrast, *nonautonomous* non-LTR retrotransposons are unable to promote their own transposition because their elements only contain a promoter and poly-A tail (i.e., both ORFs are missing). An example of such a nonautonomous non-LTR retrotransposons are the short interspersed nuclear elements (SINEs), which are abundant in many vertebrate genomes. Nonetheless, SINEs can be transposed if they can borrow the key transposition proteins from LINE elements. Thus, if cellular RNA polymerases generate large numbers of mRNA transcripts from integrated LINEs, then non-LTR retroelements can, like the LTR retrotransposons, spread like wildfire in genomes. For example, ~34% of the human genome is comprised of non-LTR retrotransposons (Brown 2007). Another example is illustrated by CR1 LINE elements (for "Chicken Repeat 1"), which have approximately 200,000 copies inserted into the chicken genome (Kaiser et al. 2007). CR1 elements are common in birds and reptiles (Shedlock 2006; Shedlock et al. 2007; Janes et al. 2010; Kordis 2010).

Another interesting aspect about the biology of transposable elements is that there are generally no known mechanisms for maintaining their function indefinitely in genomes (Graur et al. 2013). Thus, mutations are apparently free to accumulate in transposable elements because such mutations generally do not affect the host organism. If a transposable element affects the host individual in a good or bad way with respect to the fitness of the host individual, it will be because of an unlikely insertion event into a sensitive part (e.g., regulatory region) of the host genome. But if the transposable element is inserted into an unimportant (i.e., nonfunctional) part of the genome, then the transposon can persist indefinitely without harming the host organism but it will thereafter be vulnerable to mutations, which would likely impair its ability to transpose itself again (Graur and Li 2000; Watson et al. 2014). Thus, each replicative transposon typically has a window of evolutionary time in which it can proliferate throughout genomes before mutations render each copy nonfunctional. Such nonfunctional transposons may persist in the genome as relics that only evolve via mutation and genetic drift.

There are a number of ways that transposable elements can be inactivated. DNA transposons can be disabled if point mutations (i.e., single base mutations) occur within the recombination sites (inverted terminal repeats) or *transposase* coding region. Likewise, point mutations in the recombination sites, promoter sequence, or coding region for the *integrase* and *reverse transcriptase* enzymes can render LTR retrotransposons nonfunctional. A functional non-LTR retrotransposon can similarly be inactivated through the occurrence of mutations in its promoter, either of its two ORFs, or in its poly-A tail. However, even if such point mutations are absent from the integrated elements, the transposons can effectively be rendered functionless due to the production

of defective retrotransposons. For example, when non-LTR retroelements become transcribed into RNA intermediates, subsequent reverse transcription of these mRNAS often yields double-stranded cDNA products that are truncated at their 5′ ends. These truncated elements are missing a segment of sequence that contains part of the promoter or promoter plus portions of coding regions and thus they are unable to be transcribed (Watson et al. 2014). These truncations apparently occur when the reverse transcriptase, which starts at the 3′ of the mRNA, discontinues synthesis before a complete cDNA product is made (i.e., has a complete promoter and both ORFs). The CR1 elements provide a good example of this truncated product phenomenon. The length of a functional CR1 element is 4.5 kb, yet most of the ~200,000 copies present in the chicken genome are shorter than 400 bp owing to their badly truncated 5′ ends (Kaiser et al. 2007). As a consequence of these truncations these defective elements are considered "dead-on-arrival" when they are inserted into the host genome (Haas et al. 1997; Petrov and Hartl 1999).

2.2.4 Processed Pseudogenes

Although *reverse transcriptase* enzymes encoded in LINEs are highly specific for their own retrotransposon mRNA transcripts, occasionally these enzymes accidentally synthesize cDNAs from cellular mRNAs (i.e., mRNAs transcribed from nontransposon genes in the host genome; Watson et al. 2014). Upon insertion into the genome, these *processed pseudogenes* are not functional because they lack a promoter, introns, and other sites depending on the severity of the 5′ end truncation (Graur and Li 2000; Brown 2007; Watson et al. 2014). Thus, following insertion, these genes evolve differently than their functional cousins, as they are free to accumulate mutations on any site without consequence to the organism (i.e., codon bias no longer exists). In other words, free from selective constraint all sites in a pseudogene are expected to have the same substitution rate.

2.2.5 Mitochondrial Pseudogenes ("Numts")

As early as the 1960s researchers obtained evidence that mitochondrial-like sequences were residing within eukaryotic nuclear genomes (Bensasson et al. 2001). However, the existence of such nuclear copies of mitochondrial DNA would not be confirmed until the 1980s when researchers obtained DNA sequence-based evidence from fungi (van den Boogaart et al. 1982; Farrelly and Butow 1983; Wright and Cummings 1983), invertebrates (Gellissen et al. 1983; Jacobs et al. 1983), and vertebrates (Hadler et al. 1983; Tsuzuki et al. 1983; Wakasugi et al. 1985). Ellis (1982) showed that organellar DNA transfer to the nuclear genome also occurred with chloroplast DNA. An even more dramatic finding was the discovery by Lopez et al. (1994) that a 7.9 kb long mitochondrial sequence had been inserted into the nuclear genome of domesticated cats. Since then a large number of studies have documented the existence of **nu**clear **mit**ochondrial **s**egments, called "numts" (Lopez et al. 1994), in various eukaryotic nuclear genomes (Bensasson et al. 2001). Such *numt* sequences range in size from tens of bases long up to the size of an entire mitochondrial genome (Zhang and Hewitt 1996; Bensasson et al. 2001; Henze and Martin 2001).

2.2.5.1 Numt Abundance in Eukaryotic Genomes

Preliminary surveys of genomic data from representative eukaryote species suggested that *numts* are widespread among species (see reviews in Zhang and Hewitt 1996; Bensasson et al. 2001; Richly and Leister 2004). A later more comprehensive analysis (Hazkani-Covo et al. 2010) of 85 fully sequenced eukaryotic genomes, which included 20 fungi, 11 protists, 7 plants, and 47 animals, corroborated these early findings. However, significant variation in the amount of *numts* per genome also exists among species. For example 8/85 examined genomes apparently did not contain *numts* while three other genomes had >500 kb of *numt* sequences (Hazkani-Covo et al. 2010). The highest *numt* content among animals and plants was found in the possum (*Monodelphis domestica*) and rice (*Oryza sativa Indica*) genomes, which had 2,000 kb and 800 kb of *numt* sequences, respectively (Hazkani-Covo et al. 2010). Fungi and protists, on the other hand, appear to have lower *numt* content than animals and plants (Hazkani-Covo et al. 2010). In general, there is a strong positive correlation between genome size and the total number of *numts* (Bensasson et al. 2001; Hazkani-Covo et al. 2010).

The total number of *numts* per genome reflects the actions of three processes. First, *numt* loci that

originated as a result of transfer of mitochondrial DNA to that genomic location can be considered *primary integrations*, whereas copies of *numts* that arose because of intragenomic duplication events may be referred to as *secondary integrations* (Tourmen et al. 2002). Because *numts* are not known to have a self-replicating mechanism (like true replicative transposons), they apparently increase their numbers using mechanisms that generate tandemly repeated or segmental duplications (Bensasson et al. 2003; Hazkani-Covo et al. 2010). As an example, the Lopez et al. (1994) 7.9 kb *numt* in the cat genome represents a single primary integration event, whereas its 38–76 tandemly repeated copies represent secondary integrations. Lastly, the deletion of genomic segments containing *numts* represents a third mechanism that can influence the total *numt* count (Hazkani-Covo et al. 2010).

2.2.5.2 Mechanisms of Primary Numt Integration

While it had been clear to researchers for a long time that transfer of mitochondrial DNA to the nuclear genome occurred in many eukaryotes, the mechanism(s) responsible for this transfer process were not immediately obvious. Two hypothetical mechanisms that could explain this phenomenon included: direct transfer of mitochondrial DNA into the nucleus where it could be integrated into the nuclear genome and, secondly, reverse transcription of mitochondrial mRNA into cDNA in the cytoplasm or nucleus prior to being integrated into the nuclear genome (Blanchard and Schmidt 1996; Bensasson et al. 2001). Early workers favored the latter hypothesis owing to two observations: that *numts* were frequently observed in close proximity to retroelements (Blanchard and Schmidt 1996); and secondly, some plant nuclear genomes contained *numts* that were evidently derived from edited mRNA transcripts (e.g., lack introns; Henze and Martin 2001). The study by Blanchard and Schmidt (1996) found some evidence showing that both mechanisms occur (Bensasson et al. 2001). However, Blanchard and Schmidt (1996) argued that the DNA-based transfer process better explained the existence of most *numts* across eukaryotes for the following reasons. First, they pointed out that the association of retroelements with *numts* could simply be coincidental because retroelements are ubiquitous in eukaryotic genomes. Secondly, they noted that many *numts* merely consisted of gene fragments or nontranscribed DNA, which likely did not originate from RNA. Lastly, they found that *numt* integration sites (i.e., the two junctions flanking each *numt*) were not consistent with transposon recombination sites, but instead resembled the short direct repeats associated with the process of *nonhomologous end-joining* (Blanchard and Schmidt 1996). Nonhomologous end-joining is the primary process that most, if not all, eukaryotic cells use to mend double-stranded chromosomal breaks (Watson et al. 2014). To repair a double-stranded break, the cell uses mitochondrial DNA fragments (and perhaps other available DNA) as "filler DNA" to rejoin two chromosomal pieces. Because DNA is inserted into a chromosome—in this case mitochondrial DNA, this is a mutagenic process (Watson et al. 2014). Thus, if *numts* are introduced into functional parts of the genome, then this process can lead to disease. However, this rarely occurs because the vast majority of *numts* are inserted into intergenic regions and thus they represent neutral mutations (Hazkani-Covo et al. 2010). Accordingly, the benefits of repairing double-stranded breaks using nonhomologous end-joining vastly outweigh the risks (Watson et al. 2014).

Another argument in favor of the direct DNA-transfer hypothesis comes from many independent discoveries of long segments of mitochondrial DNA—some of which include introns and nontranscribed regions—inside nuclear genomes (Henze and Martin 2001; Hazkani-Covo et al. 2010). For example, in addition to the 7.9 kb mitochondrial sequence found in the domestic cat genome (Lopez et al. 1994), a 12.5 kb *numt* was observed in the genomes of various species of *Panthera* (Kim et al. 2006; Antunes et al. 2007) and a 14.6 kb fragment—nearly the entire mitochondrial genome—was found in the human genome (Mourier et al. 2001). An even more startling example was the discovery of a 620 kb long segment of mitochondrial DNA—the *entire* mitochondrial genome plus some internally duplicated regions, which were found in the *Arabidopsis thaliana* nuclear genome (Lin et al. 1999; Henze and Martin 2001; Stupar et al. 2001). The RNA-intermediate hypothesis cannot account for these observations (Henze and Martin 2001). In later years, additional studies would instead support

the alternative hypothesis—that primary integrations of *numts* are the result of direct transfers of bulk mitochondrial DNA into the nucleus where they are incorporated into double-stranded breaks via nonhomologous end-joining (Ricchetti et al. 1999; Yu and Gabriel 1999; Bensasson et al. 2001; Hazkani-Covo and Covo 2008; Hazkani-Covo et al. 2010).

The finding that nonhomologous end-joining appears to be the primary mechanism for new *numt* integration also provides an explanation for why larger genomes tend to have larger numbers of *numts* than do smaller genomes. Larger genomes will have more opportunities for *numt* integration simply because in larger genomes there will be more chromosomal breaks in need of repair; in other words, the limiting factor for primary integration of *numts* is the number of required repairs of chromosomal breaks (Hazkani-Covo et al. 2010).

2.2.5.3 *Differences between Numts and Mitochondrial DNA*

The process of primary integration of *numts* into eukaryotic genomes appears to be continuous over time (Mourier et al. 2001; Tourmen et al. 2002; Bensasson et al. 2003; Hazkani-Covo et al. 2003; Antunes et al. 2007). Even within a genome, such as the human genome, *numts* show widely varying divergence times from their mitochondrial genome progenitors as well as among homologous secondary integration copies (Bensasson et al. 2003). Recently inserted *numts* exhibit little or no sequence divergences compared to their mitochondrial counterparts, whereas ancient *numts* are much less recognizable in genomes owing to the many base substitutions, indels, inversions, and deletions they have suffered (Jacobs and Grimes 1986; Tourmen et al. 2002). *In silico* searches of complete genome sequences for *numts* can produce varying estimates of the total number of *numts* depending on the type and stringency of computerized searches (Bensasson et al. 2003; Hazkani-Covo et al. 2003, 2010; Antunes et al. 2007). This not only suggests that estimates of the "total" number of *numts* per genome are likely underestimates (Tourmen et al. 2002; Antunes et al. 2007), but that *numts* are, unlike their mitochondrial counterparts, without selective constraint and thus are free to decay via mutation (Arctander 1995; Sorenson and Quinn 1998).

Several other lines of evidence exist suggesting that *numts* (at least in animals) are free from selective constraint and therefore evolve like pseudogenes. First, many *numts* are only gene fragments or noncoding mitochondrial DNA (Blanchard and Schmidt 1996) thus they likely do not have a function. Secondly, the genetic code for mitochondrial DNA is different from that for nuclear DNA, which implies that mRNAs generated from *numts* would yield defective protein products (Gellissen and Michaelis 1987; Perna and Kocher 1996). Lastly, *numts* often have indels, which creates frameshift mutations and premature stop codons—both of which would result in nonfunctional proteins (Jacobs and Grimes 1986; Smith et al. 1992; Arctander 1995; Bensasson et al. 2001).

Let's now take a closer look at how the mutational regimes differ between *numts* and their mitochondrial counterparts. Mitochondrial and chloroplast genomes undergo low levels of cytosine methylation at CpG sites but it is extensive in the nuclear genomes of higher plants (Huang et al. 2005 and references therein). In a study of the *Arabidopsis thaliana* and *Oryza sativa* genomes by Huang et al. (2005), they found that cytosines at CpG sites in *numts* were rapidly methylated and subsequently underwent C to T transitions (or G to A transitions on the opposite strand). Moreover, they noted that transitions at CpG sites were 5–10 times higher than rates at other *numt* sites and therefore they suggested that this represented a major force in the mutational decay of *numt* loci. It is unclear how extensive CpG methylation is in animals, but it has been found in some groups and thus similar elevated substitution rates at CpG sites are also observed (Petrov and Hartl 1999; Bensasson et al. 2001; Hazkani-Covo et al. 2010). A second difference concerns the overall substitution rates: because *numt* sites are embedded in nuclear DNA, their base substitution rates are expected to be 10-fold slower than for mitochondrial DNA sites (Brown et al. 1979). Also, the *TI*/*TV* rate ratio for *numt* sites is expected to conform to the level observed in nuclear DNA (i.e., *TI*/*TV* rate ratio ~ 2) rather than the level inferred for mitochondrial DNA (i.e., *TI*/*TV* rate ratio ~ 15; Zhang and Hewitt 1996; Graur and Li 2000). Thus, *numt* sites require more time to reach saturation than comparable mitochondrial sites (Zischler et al. 1998; Bensasson et al. 2001). A fourth difference is that while mitochondrial

genes are under selective constraints, particularly at certain codon positions (Brown et al. 1982), *numts* on the other hand are expected to have the same substitution rate across sites (Arctander 1995; Yang 1996; Zhang and Hewitt 1996; Bensasson et al. 2001, 2003). Lastly, in contrast to most mitochondrial sites, *numts* can experience insertions and deletions without the individual suffering adverse effects (Bensasson et al. 2001; Tourmen et al. 2002).

REFERENCES

Ambrožová, K., T. Mandáková, P. Bureš et al. 2010. Diverse retrotransposon families and an AT-rich satellite DNA revealed in giant genomes of *Fritillaria* lilies. *Ann Bot*. doi: 10.1093/aob/mcq235.

Antunes, A., J. Pontius, M. J. Ramos, S. J. O'Brien, and W. E. Johnson. 2007. Mitochondrial introgressions into the nuclear genome of the domestic cat. *J Hered* 98:414–420.

Arbogast, B. S., S. V. Edwards, J. Wakeley, P. Beerli, and J. B. Slowinski. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu Rev Ecol Syst* 33:707–740.

Arctander, P. 1995. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. *Proc R Soc Lond B: Biol Sci* 262:13–19.

Ashley Jr, C. T. and S. T. Warren. 1995. Trinucleotide repeat expansion and human disease. *Annu Rev Genet* 29:703–728.

Ballard, J. W. O. and D. M. Rand. 2005. The population biology of mitochondrial DNA and its phylogenetic implications. *Annu Rev Ecol Evol Syst* 36:621–642.

Bejerano, G., M. Pheasant, I. Makunin et al. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.

Bensasson, D., M. W. Feldman, and D. A. Petrov. 2003. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J Mol Evol* 57:343–354.

Bensasson, D., D. X. Zhang, D. L. Hartl, and G. M. Hewitt. 2001. Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends Ecol Evol* 16:314–321.

Birol, I., A. Raymond, S. D. Jackman et al. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*. doi: 10.1093/bioinformatics/btt178.

Blanchard, J. L. and G. W. Schmidt. 1996. Mitochondrial DNA migration events in yeast and humans: Integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Mol Biol Evol* 13:537–548.

Boore, J. L. 1999. Animal mitochondrial genomes. *Nucleic Acids Res* 27:1767–1780.

Brown, T. A. 2007. *Genomes* 3. New York: Garland Science/Taylor & Francis.

Brown, W. M., M. George, and A. C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci USA* 76:1967–1971.

Brown, W. M., E. M. Prager, A. Wang, and A. C. Wilson. 1982. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *J Mol Evol* 18:225–239.

Casane, D., S. Boissinot, B. J. Chang, L. C. Shimmin, and W.-H. Li. 1997. Mutation pattern variation among regions of the primate genome. *J Mol Evol* 45:216–226.

Chang, D. and T. F. Duda. 2012. Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Mol Biol Evol*. doi: 10.1093/molbev/mss068.

Costa, I. R., F. Prosdocimi, and W. B. Jennings. 2016. In silico phylogenomics using complete genomes: A case study on the evolution of hominoids. *Genome Res* 26:1257–1267.

De Koning, A. P., W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* 7:e1002384.

Echols, H. and M. F. Goodman. 1991. Fidelity mechanisms in DNA replication. *Annu Rev Biochem* 60:477–511.

Ellis, J. 1982. Promiscuous DNA–chloroplast genes inside plant mitochondria. *Nature* 299:678–679.

Farrelly, F. and R. A. Butow. 1983. Rearranged mitochondrial genes in the yeast nuclear genome. *Nature* 301:296–301.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sunderland: Sinauer.

Galtier, N., B. Nabholz, S. Glémin, and G. D. D. Hurst. 2009. Mitochondrial DNA as a marker of molecular diversity: A reappraisal. *Mol Ecol* 18:4541–4550.

Gellissen, G., J. Y. Bradfield, B. N. White, and G. R. Wyatt. 1983. Mitochondrial DNA sequences in the nuclear genome of a locust. *Nature* 301:631–634.

Gellissen, G. and G. Michaelis. 1987. Gene transfer. *Ann N Y Acad Sci* 503:391–401.

Golding, G. B. 1983. Estimates of DNA and protein sequence divergence: An examination of some assumptions. *Mol Biol Evol* 1:125–142.

Graur, D. and W.-H. Li. 2000. *Fundamentals of Molecular Evolution*, 2nd edition. Sunderland: Sinauer.

Graur, D., Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik. 2013. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5:578–590.

Haas, N. B., J. M. Grabowski, A. B. Sivitz, and J. B. Burch. 1997. Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. *Gene* 197:305–309.

Hadler, H. I., B. Dimitrijevic, and R. Mahalingam. 1983. Mitochondrial DNA and nuclear DNA from normal rat liver have a common sequence. *Proc Natl Acad Sci USA* 80:6495–6499.

Harris, V. H., C. L. Smith, W. J. Cummins et al. 2003. The effect of tautomeric constant on the specificity of nucleotide incorporation during DNA replication: Support for the rare tautomer hypothesis of substitution mutagenesis. *J Mol Biol* 326:1389–1401.

Hazkani-Covo, E. and S. Covo. 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet* 4:e1000237.

Hazkani-Covo, E., R. Sorek, and D. Graur. 2003. Evolutionary dynamics of large numts in the human genome: Rarity of independent insertions and abundance of post-insertion duplications. *J Mol Evol* 56:169–174.

Hazkani-Covo, E., R. M. Zeller, and W. Martin. 2010. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* 6:e1000834.

Henze, K. and W. Martin. 2001. How do mitochondrial genes get into the nucleus? *Trends Genet* 17:383–387.

Huang, C. Y., N. Grünheit, N. Ahmadinejad, J. N. Timmis, and W. Martin. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiol* 138:1723–1733.

Jacobs, H. T. and B. Grimes. 1986. Complete nucleotide sequences of the nuclear pseudogenes for cytochrome oxidase subunit I and the large mitochondrial ribosomal RNA in the sea urchin *Strongylocentrotus purpuratus*. *J Mol Biol* 187:509–527.

Jacobs, H. T., J. W. Posakony, J. W. Grula et al. 1983. Mitochondrial DNA sequences in the nuclear genome of *Strongylocentrotus purpuratus*. *J Mol Biol* 165:609–632.

Janes, D. E., C. L. Organ, M. K. Fujita, A. M. Shedlock, and S. V. Edwards. 2010. Genome evolution in Reptilia, the sister group of mammals. *Annu Rev Genomics Hum Genet* 11:239–264.

Kaiser, V. B., M. van Tuinen, and H. Ellegren. 2007. Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in galliform birds. *Mol Biol Evol* 24:338–347.

Katzman, S., A. D. Kern, G. Bejerano. 2007. Human genome ultraconserved elements are ultraselected. *Science* 317:915–915.

Keller, I., D. Bensasson, and R. A. Nichols. 2007. Transition–transversion bias is not universal: A counter example from grasshopper pseudogenes. *PLoS Genet* 3:e22.

Kim, J. H., A. Antunes, S. J. Luo et al. 2006. Evolutionary analysis of a large mtDNA translocation (numt) into the nuclear genome of the *Panthera* genus species. *Gene* 366:292–302.

Kordis, D. 2010. Transposable elements in reptilian and avian (sauropsida) genomes. *Cytogenet Genome Res* 127:94–111.

Kunkel, T. A. and K. Bebenek. 2000. DNA replication fidelity 1. *Annu Rev Biochem* 69:497–529.

Lee, S. I. and N. S. Kim. 2014. Transposable elements and genome size variations in plants. *Genomics Inform* 12:87–97.

Levinson, G. and G. A. Gutman. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203–221.

Lin, X., S. Kaul, S. Rounsley et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402:761–768.

Lohse, M., O. Drechsel, S. Kahlau, and R. Bock. 2013. Organellar Genome DRAW—A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res*. doi: 10.1093/nar/gkt289.

Lopez, J. V., N. Yuhki, R. Masuda, W. Modi, and S. J. O'Brien. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol* 39:174–190.

Margulis, L. 1970. *Origin of Eukaryotic Cells: Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth*. New Haven: Yale University Press.

McClintock, B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA* 36:344–355.

Moritz, C., T. E. Dowling, and W. M. Brown. 1987. Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. *Annu Rev Ecol Syst* 1987:269–292.

Mourier, T., A. J. Hansen, E. Willerslev, and P. Arctander. 2001. The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Mol Biol Evol* 18:1833–1837.

Ovcharenko, I., G. G. Loots, M. A. Nobrega, R. C. Hardison, W. Miller, and L. Stubbs. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome Res* 15:137–145.

Pakendorf, B. and M. Stoneking. 2005. Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet* 6:165–183.

Palumbi, S. R. 1996. Chapter 7. Nucleic acids II: The polymerase chain reaction. In *Molecular Systematics*, 2nd edition, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 205–247. Sunderland: Sinauer.

Perna, N. T. and T. D. Kocher. 1996. Mitochondrial DNA: Molecular fossils in the nucleus. *Curr Biol* 6:128–129.

Petrov, D. A. and D. L. Hartl. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc Natl Acad Sci USA* 96:1475–1479.

Rambaut, A. 2007. *Se-Al, version 2.0 a11*. Edinburgh: The University of Edinburgh.

Randi, E. 2000. Mitochondrial DNA. In *Molecular Methods in Ecology*, ed. A. J. Baker, 136–167. Oxford: Blackwell.

Ravindran, S. 2012. Barbara McClintock and the discovery of jumping genes. *Proc Natl Acad Sci USA* 109:20198–20199.

Ricchetti, M., C. Fairhead, and B. Dujon. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402:96–100.

Richly, E. and D. Leister. 2004. NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol* 21:1081–1084.

Sagan, L. 1967. On the origin of mitosing cells. *J Theor Biol* doi:10.1016/0022-5193(67)90079-3.

Scribner, K. T. and J. M. Pearce. 2000. Microsatellites: Evolutionary and methodological background and empirical applications at individual, population, and phylogenetic levels. In *Molecular Methods in Ecology*, ed. A. J. Baker, 235–273. Oxford: Blackwell.

Shedlock, A. M. 2006. Phylogenomic investigation of CR1 LINE diversity in reptiles. *Syst Biol* 55:902–911.

Shedlock, A. M., C. W. Botka, S. Zhao et al. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proc Natl Acad Sci USA* 104:2767–2772.

Simons, C., M. Pheasant, I. V. Makunin, and J. S. Mattick. 2006. Transposon-free regions in mammalian genomes. *Genome Res* 16:164–172.

Skippington, E., T. J. Barkman, D. W. Rice, and J. D. Palmer. 2015. Miniaturized mitogenome of the parasitic plant *Viscum scurruloideum* is extremely divergent and dynamic and has lost all nad genes. *Proc Natl Acad Sci USA* 112:E3515–E3524.

Smeds, L., A. Qvarnström, and H. Ellegren. 2016. Direct estimate of the rate of germline mutation in a bird. *Genome Res* 26:1211–1218.

Smith, M. F., W. K. Thomas, and J. L. Patton. 1992. Mitochondrial DNA-like sequence in the nuclear genome of an akodontine rodent. *Mol Biol Evol* 9:204–215.

Sorenson, M. D. and T. W. Quinn. 1998. Numts: A challenge for avian systematics and population biology. *Auk* 115:214–221.

Souto, H. M., P. A. Ruschi, C. Furtado, W. B. Jennings, and F. Prosdocimi. 2014. The complete mitochondrial genome of the ruby-topaz hummingbird *Chrysolampis mosquitus* through Illumina sequencing. *Mitochondrial DNA* 27:769–770.

Stupar, R. M., J. W. Lilly, C. D. Town et al. 2001. Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: Implication of potential sequencing errors caused by large-unit repeats. *Proc Natl Acad Sci USA* 98:5099–5103.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Chapter 11. Phylogenetic inference. In *Molecular Systematics*, 2nd edition, eds. D. M. Hillis, C. Moritz, and B. K. Mable, 407–514. Sunderland: Sinauer.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526.

Tourmen, Y., O. Baris, P. Dessen, C. Jacques, Y. Malthièry, and P. Reynier. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 80:71–77.

Tsuzuki, T., H. Nomiyama, C. Setoyama, S. Maeda, and K. Shimada. 1983. Presence of mitochondrial-DNA-like sequences in the human nuclear DNA. *Gene* 25:223–229.

Upholt, W. B. 1977. Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. *Nucleic Acids Res* 4:1257–1266.

van den Boogaart, P., J. Samallo, and E. Agsteribbe. 1982. Similar genes for a mitochondrial ATPase subunit in the nuclear and mitochondrial genomes of *Neurospora crassa*. *Nature* 298:187–189.

Venter, J. C., M. D. Adams, E. W. Myers et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.

Wakasugi, S., N. Hisayuki, F. Makoto, T. Teruhisa, and S. Kazunori. 1985. Insertion of a long KpnI family member within a mitochondrial-DNA-like sequence present in the human nuclear genome. *Gene* 36:281–288.

Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. *Mol Biol Evol* 11:436–442.

Wakeley, J. 1996. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *Trends Ecol Evol* 11:158–162.

Wang, W., H. W. Hellinga, and L. S. Beese. 2011. Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proc Natl Acad Sci USA* 108:17644–17648.

Waples, R. S. and O. Gaggiotti. 2006. INVITED REVIEW: What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol Ecol* 15:1419–1439.

Watson, J. D., T. A. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick. 2014. *Molecular Biology of the Gene*, 7th edition. New York: Pearson Education, Inc.

Watson, J. D. and F. H. C. Crick. 1953a. Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171:964–967.

Watson, J. D. and F. H. C. Crick. 1953b. Molecular structure of nucleic acids. *Nature* 171:737–738.

Wilson, A. C., R. L. Cann, S. M. Carr et al. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol J Linn Soc Lond* 26:375–400.

Wright, R. M. and D. J. Cummings. 1983. Integration of mitochondrial gene sequences within the nuclear genome during senescence in a fungus. *Nature* 302:86–88.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 11:367–372.

Yang, Z. 2006. *Computational Molecular Evolution* (Vol. 21). Oxford: Oxford University Press.

Yang, Z. and A. D. Yoder. 1999. Estimation of the transition/transversion rate bias and species sampling. *J Mol Evol* 48:274–283.

Yu, X. and A. Gabriel. 1999. Patching broken chromosomes with extranuclear cellular DNA. *Mol Cell* 4:873–881.

Zhang, D. X. and G. M. Hewitt. 1996. Nuclear integrations: Challenges for mitochondrial DNA markers. *Trends Ecol Evol* 11:247–251.

Zhang, Z. and M. Gerstein. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* 31:5338–5348.

Zhang, Z., P. Harrison, and M. Gerstein. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* 12:1466–1482.

Zischler, H., H. Geisert, and J. Castresana. 1998. A hominoid-specific nuclear insertion of the mitochondrial D-loop: Implications for reconstructing ancestral mitochondrial sequences. *Mol Biol Evol* 15:463–469.