Mutation

2

Chapter Summary

Mutations, the fundamental sources of evolution, are described in detail. They include nucleotide substitutions, insertions/deletions, recombinations, gene conversions, gene duplications, and genome duplications. Mutation rate estimates and methods to estimate mutation rates are also discussed.

2.1 Classification of Mutations

2.1.1 What Is Mutation?

Any change of nucleotide sequences in one genome can be considered as "mutation" in the broad sense. In the traditional view, mutations occur when genes are transmitted from parent to child. If we focus on diploids, miosis may be the only chance for mutations. However, DNA replications also occur during mitosis and DNA damage can happen at any time without DNA replication. Therefore, indication of time unit, per generation, per year, or per replication, is important when we discuss mutation rate differences.

The classic unit of mutation was gene, because gene was defined as a unit for particular phenotype or particular function. Now we know that DNA is the material basis of inheritance, and any modification of nucleotide sequences should be considered as "mutation." In the early days of molecular genetics, the term "point mutation" was often used. Change of one nucleotide may correspond to point mutation. However, this includes nucleotide substitution and deletion or insertion of one nucleotide. A change involving more than one nucleotide may also be considered as point mutation if they are contiguous. Because of this uncertainty, we should not use this term anymore. Mutations should be classified by their structural characteristics. Table 2.1 shows a list of mutations. Types of DNA polymorphisms caused by various types of mutations are also listed in Table 2.1, and they will be discussed in Chap. 4.

N. Saitou, *Introduction to Evolutionary Genomics*, Computational Biology 17,
DOI 10.1007/978-1-4471-5304-7_2, © Springer-Verlag London 2013

Table 2.1 List of mutation types

Type	Polymorphism
Nucleotide substitution	SNP (single nucleotide polymorphism)
Insertion and deletion	Indel
Gene conversion (allelic)	SNP-like
Gene conversion (paralogous)	SNP-like
Repeat number change	STR (short tandem repeat) or microsatellite
Single crossover	Recombination
Unequal crossover	CNV (copy number variation)
Double crossover	SNP-like
Inversion	Inversion polymorphism
Chromosomal translocation	Translocation polymorphism
Repeat insertion	Insertion polymorphism
Genome duplication	Genome number polymorphism

2.1.2 Temporal Unit of Mutation

The classic unit of time to measure mutation is one generation: between parents and children, for mutations were believed to occur only at meiosis in germ line. However, somatic mutation, or mutation in somatic cells, does occur. Acquired immune system of vertebrates is known to increase antibody amino acid sequence variation by incorporating somatic mutations [1]. Because mitosis may be involved in creating somatic mutations, the number of cell division in germ line is another important factor for mutation. Haldane ([2]; cited in [3]) already suggested in 1947 that the mutation rate might be much higher in males than in females because the number of germ-cell divisions per generation is much higher in the male germ line than in the female germ line. In fact, the long-term mutation rate for Y chromosomal DNA in mammals is clearly higher than that for autosomes and X chromosome [3, 4]. This difference is easy to be understood if we consider a huge number of sperms and a small number of eggs produced in one generation. Y chromosomes always pass through sperms, while autosomes pass through either sperm or egg with equal probability. An X chromosome has 1/3 and 2/3 probabilities for passing sperm and egg, respectively.

Mutations may not be restricted to cell divisions. They may occur at any time, for any damage to DNA molecules is always the starting point for a mutation. In any case, mechanisms of mutation are not yet understood in detail, and this is a future problem.

2.1.3 Mutations Affecting Small Regions of DNA Sequences

When only a small portion of the DNA sequence is modified, say, one to a few nucleotides, this may be called minute mutation. They are nucleotide substitutions,

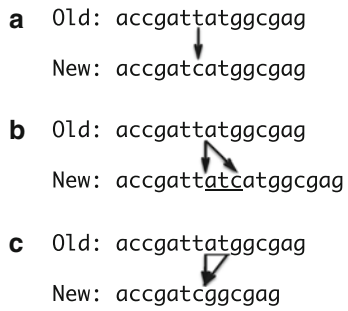

Fig. 2.1 Minute mutations. (**a**) Nucleotide substitution. (**b**) Short insertion. (**c**) Short deletion

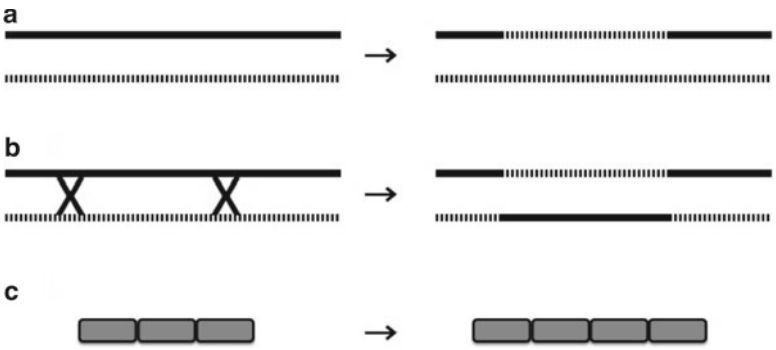

Fig. 2.2 Mini-scale mutations. (**a**) Allelic gene conversion. (**b**) Double crossover. (**c**) STR number change

short insertions, and short deletions. Figure 2.1 shows a schematic view of these minute mutations. Nucleotide substitutions will be discussed in detail at Sect. 2.2 and insertions and deletions at Sect. 2.3.

Mutations affecting somewhat larger regions of DNA sequences may be called mini-scale mutations (Fig. 2.2). Allelic gene conversion, double crossover, and short tandem repeat (STR) number change are included in this category.

2.1.4 Mutations Affecting Large Regions of DNA Sequences

The physical order of nucleotide sequence is modified in recombination, paralogous gene conversion, and inversion (see Fig. 2.3). Chromosomal level changes of DNA sequences are classified into inversion, translocation, and fusion as shown in Fig. 2.4. The human chromosomes have been well studied, and more detailed description will be given in Chap. 10. The largest type of mutation is genome duplication or polyploidization.

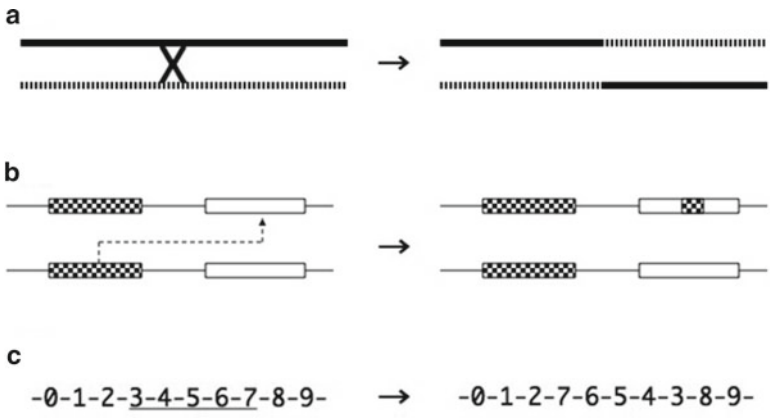

Fig. 2.3 Mutations affecting the physical order of nucleotide sequences. (**a**) Single crossover. (**b**) Paralogous gene conversion. (**c**) Inversion

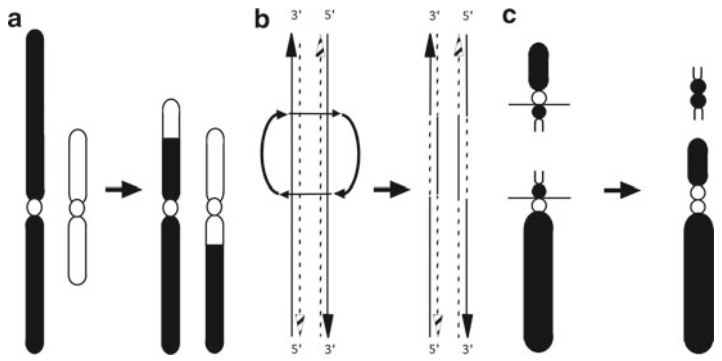

Fig. 2.4 Mutations affecting the large area of one chromosome (Based on [58])

2.2 Nucleotide Substitutions

2.2.1 Basic Characteristics of Nucleotide Substitutions

Nucleotide substitutions are mutual interchanges of four kinds of nucleotides or bases. Figure 2.5 shows all possible 12 nucleotide substitutions in DNA sequences. If a substitution is between chemically similar bases (see Chap. 1), i.e., between purines (adenine and guanine) or pyrimidines (cytosine and thymine), it is called transition. If a substitution is between a purine and a pyrimidine, it is called transversion.

It was predicted that transition should occur in higher frequency than transversion, because transitions have four possible intermediate mispair states, while transversions have only two such states. Figure 2.6 shows six possible intermediate mispairings. If we start from adenine–thymine-type normal base pairing, transition (A–T to G–C)

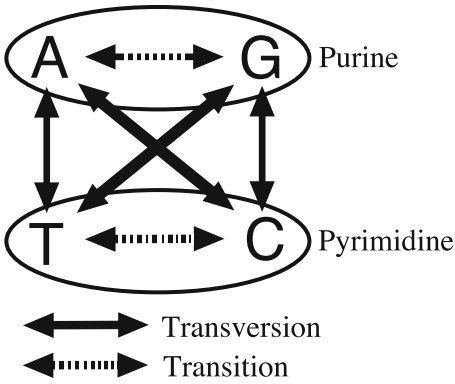

Fig. 2.5 Pattern of nucleotide substitution

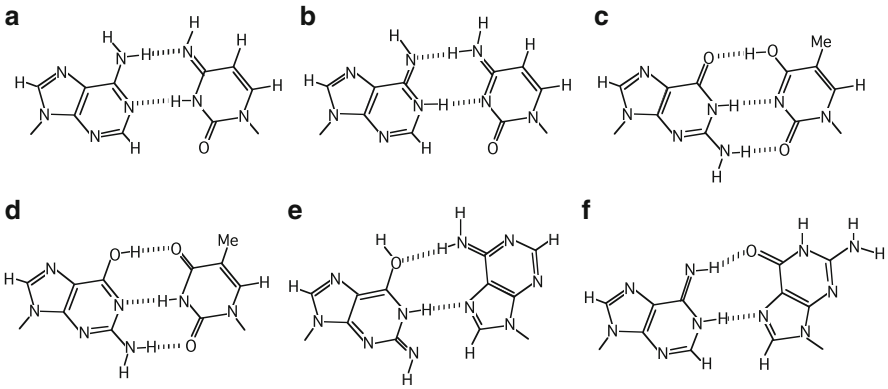

Fig. 2.6 Six possible intermediate mispairings (Based on Ref. [5]). (**a**) Adenine and imino-cytosine. (**b**) Imino-adenine and cytosine. (**c**) Guanine and enol-thymine. (**d**) Enol-guanine and thymine. (**e**) Imino-adenine and syn-adenine. (**f**) Imino-adenine and syn-guanine (based on [5])

can occur through (a) to (d) intermediate mispairings, while transversion (A–T to T–A) can occur only through (e) or (f) intermediate mispairings [5]. This is the basis of higher transitions than transversions.

Absolute rates of mutations for 12 kinds of directions are not easy to estimate, because we need to directly compare parental and offspring genomes, and the rate of fresh or de novo mutations in eukaryotes is usually quite low. Instead, we can compare evolutionarily closely related sequences. Relative mutation rates of six pairs of bases (A<==>G, C<==>T, A<==>T, A<==>C, G<==>T, and G<==>C) can be estimated by comparing many numbers of SNPs (single nucleotide polymorphisms) in one species. However, we need the closely related out-group species when the directionality of mutation comes in. Figure 2.7 shows how to estimate direction of mutation in this way.

The pattern of nucleotide substitutions in the human genome was estimated using the scheme of Fig. 2.7. More than 30,000 human SNP data determined for

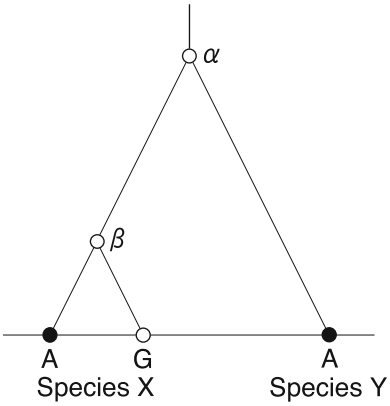

Fig. 2.7 Estimation of mutation direction in species X by using orthologous sequence data for species Y. Species X is polymorphic at some ne nucleotide site, and two nucleotides (A and G) coexist, while the corresponding nucleotide position for species Y is A. Using parsimony principle, ancestral nodes α and β are estimated to be both A. Therefore, the mutation direction is A to G.

Table 2.2 Pattern of nucleotide substitutions in the human genome (From [6])

	A	T	C	G
A	–	2.9	3.6	14.0
T	2.8	–	15.1	3.5
C	4.4	20.3	–	4.5
G	19.6	4.5	4.9	–

Unit: %

chromosome 21 were used, and the direction of mutation was inferred using the chimpanzee chromosome 22 as the out-group sequence [6]. We standardized these frequencies using Gojobori et al.'s (1982; [7]) method, so the sum becomes 100 %. The result is shown in Table 2.2. First of all, transitions shown in top-right to down-left diagonal are much more frequent than transversions. Among transitions, G=>A and C=>T are more frequent than their reverse directions (A=>G and T=>C). This corresponds to the fact that the mammalian genomes, here represented by the human genome, have about 40 % G+C proportion, or are A+T rich.

Transitions are known to be quite high in animal mitochondrial DNA. Kawai and Saitou (unpublished) analyzed complete mitochondrial DNA genome sequences of 7,264 human individuals and observed 4,939 substitutions with inferred direction among 2,179 fourfold degenerate synonymous sites. This result shows that transitions are about 30 times higher than transversions. Among transitions, C=>T changes are less than the other three directions, and C=>A changes are most abundant among transversions.

Evolutionary rates of nucleotide substitution are expected to be equal to the mutation rate of nucleotide substitution types, if the genomic region in question is evolving in purely neutral fashion (see Chap. 4). This characteristic is used for estimate mutation rates of various organisms, as shown in Sect. 2.6.3.

Table 2.3 Examples of various types of substitutions in the protein coding region

Synonymous substitution
All three possible substitutions at third position of a codon
GCT (Ala) → GCA (Ala), GCC (Ala), GCG (Ala)
Transitional type substitution at third position of a codon
AAT (Asn) → AAC (Asn)
Substitution at first position of a codon
CGA (Arg) → AGA (Arg)
Nonsynonymous substitution
All three possible substitutions at second position of a codon
GCT (Ala) → GAT (Asp), GGT (Gly), GTT (Val)
Transvertional type substitutions at third position of a codon
AAT (Asn) → AAA (Lys), AAG (Lys)
Substitution at first position of a codon
CGT (Arg) → AGT (Ser), GGT (Gly), TGT (Cys)
Nonsense substitution
Substitution at first position of a codon
CAA (Gln) → TAA (stop)
Substitution at second position of a codon
TCG (Ser) → TAG (stop)
Substitution at third position of a codon
TGG (Trp) → TGA (stop)
Stop codon to amino acid codon substitution
Substitution at first position of a codon
TAA (stop) → GAA (Glu)
Substitution at second position of a codon
TAG (stop) → TTG (Leu)
Substitution at third position of a codon
TGA (stop) → TGC (Cys)

2.2.2 Nucleotide Substitutions in Protein Coding Regions

When a nucleotide substitution occurs in a protein coding region, it may be either synonymous, nonsynonymous, or nonsense substitution; see the standard genetic code table shown in Table 1.1. Synonymous substitution may also be called silent substitutions in protein coding regions, and nonsynonymous substitution that seems to be used by Gojobori (1983; [8]) for the first time may also be called missense mutation or amino acid replacing mutation.

Amino acid sequence will not be changed when a synonymous substitution occurs, while amino acid will be changed when a nonsynonymous substitution happens. A nonsense substitution will change an amino acid codon to stop codon and will shorten the amino acid sequence. Change from stop codon to amino acid codon will elongate proteins. Table 2.3 shows examples of these four types of nucleotide substitutions.

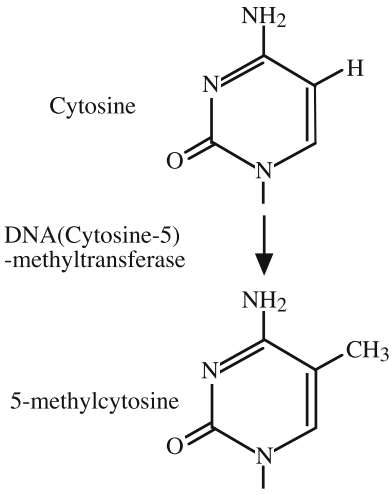

Fig. 2.8 Methylation of cytosine

2.2.3 Methylation Creates "Fifth" Nucleotide

DNA methyltransferase may recognize 5′-cytocine-guanine-3′ (CpG) and modify cytosine to 5-methylcytocine in mammalian genomes (Fig. 2.8). Because the CpG dinucleotide is complementary to itself, this double-strand DNA is methylated in both strands. This methylation is known to be related to imprinting and is an epigenetic phenomenon. Interestingly, CpG islands, regions with high CpG density just upstream of protein coding genes, are rarely methylated.

2.3 Insertion and Deletion

2.3.1 Basic Characteristics of Insertions and Deletions

The length of DNA does not change with nucleotide substitutions, while it is well known that genome sizes vary from organism to organism. It is thus clear that there exist mutations changing length of DNA. They are generically called "insertion" or "deletion" when the DNA length increases or decreases, respectively. When mutational directions are not known, combinations of insertions and deletions may be called gaps or indels. When the gap length is only one, this gap or indel polymorphism may be included as a special case of SNP (single nucleotide polymorphism). In real nucleotide sequence data analysis, insertions and deletions are detected only after multiple alignment of homologous sequences. The relationship of insertion and deletion with sequence alignment techniques will be discussed in Chap. 14.

A special class of insertions and deletions is repeat number changes. If repeat unit length is very short (less than 10 nucleotides), it is called STRs (short tandem repeats) or microsatellites. In contrast, "minisatellites" or VNTRs (variable number

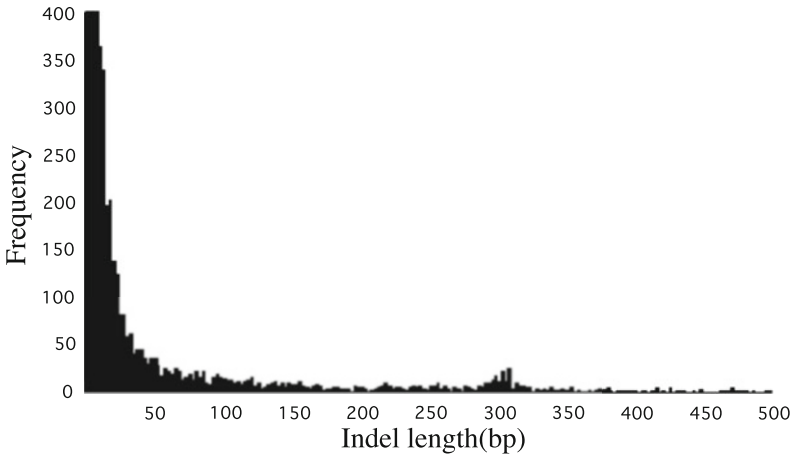

Fig. 2.9 Length distribution of indels in the human/chimpanzee lineages (From The International Chimpanzee Chromosome 22 Consortium 2004; [6])

of tandem repeats) have typically repeat unit lengths of 10–100 nucleotides. Because of their importance on DNA polymorphism studies, let us divide insertions and deletions into two types, unique sequence and repeat sequence, and we discuss them independently.

2.3.2 Insertions and Deletions of Unique Sequences

The length distribution of indels in the human or the chimpanzee lineages is shown in Fig. 2.9. Single nucleotide changes are most frequent, and the frequency quickly drops as the length becomes longer. It should be noted that there is a small peak around 300 bp. This is mostly due to Alu sequence insertions. Figure 2.10 shows data similar to those of Fig. 2.9 with mutational directions. Directions (either insertion or deletion) for each indel position were estimated by checking situations in gorilla and orangutan genomes [6]. Interestingly, the human genome experienced more insertions than the chimpanzee genome, especially for Alu sequences, as shown in Fig. 2.11. In contrast, the length distribution patterns for deletions do not differ significantly between human and chimpanzee genomes.

Minute length gaps or indels can be studied by examining multiple aligned orthologous (see Chap. 3) nucleotide sequences of closely related species. If they are located in the genomic regions under purely neutral evolution, their evolutionary rates can be considered as the mutation rates (see Chap. 4 for this rationale). Saitou and Ueda (1994; [9]) estimated mutation rates of insertions or deletions, for the first time, using primate species noncoding genomic regions, and they found molecular clocks (rough constancy of the evolutionary rates) both in mitochondrial and nuclear DNAs (see Fig. 2.12). The rate (approximately 2.0/kb/Myr) for mitochondrial DNA

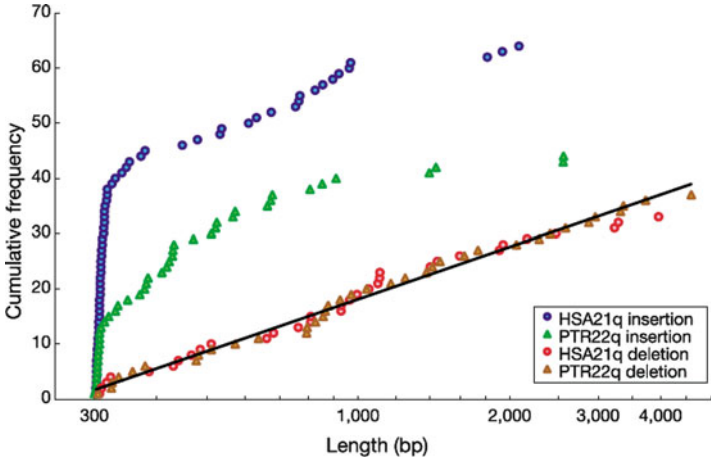

Fig. 2.10 Length distribution (more than 300 bp; cumulative) of insertions and deletions in the human and chimpanzee lineages (From The International Chimpanzee Chromosome 22 Consortium 2004; [6])

was found to be much higher than that (approximately 0.2/kb/Myr) for nuclear DNA. Because the rate of nucleotide substitutions in nuclear genome of primates is approximately 1×10^{-9}/site/year, the rate of insertions and deletions is about 1/5 of substitution. Ophir and Graur (1997; [10]) compared hundreds of functional genes and their processed pseudogenes in human and mouse nuclear genomes and found that deletions are more than two times more frequent than insertions. They also estimated that the rate of insertions is 1/100 of that of nucleotide substitutions. When chimpanzee chromosome 22, corresponding to human chromosome 21, was sequenced in 2004 [6], insertions and deletions were carefully analyzed. A total of 68,000 indels were found from the human–chimpanzee chromosomal alignment. More than 99 % of them are shorter than 300 bp. The human chromosome 21 long arm is 33.1 Mb, and chimpanzee chromosome 22 long arm is 32.8 Mb. If we take their average, 33.0 Mb, as compared length, and if we assume that the human and chimpanzee divergence time is 6 million years, then the overall rate of insertions and deletions in human and chimpanzee lineages becomes 0.38×10^{-9}/site/year (= 68,000/30Mb/6MY). This estimate is about two times higher than that obtained by Saitou and Ueda [9] using much smaller sequence data of primates. See Sect. 2.6.2 for more discussion.

2.3.3 Insertions and Deletions of Repeat Sequences

There are many studies on mutation mechanism of STRs or microsatellites. DNA slippage is commonly accepted to be the major mutation mechanism of microsatellites [60]. Factors affecting STR slippage include repeat number, locus length, motif

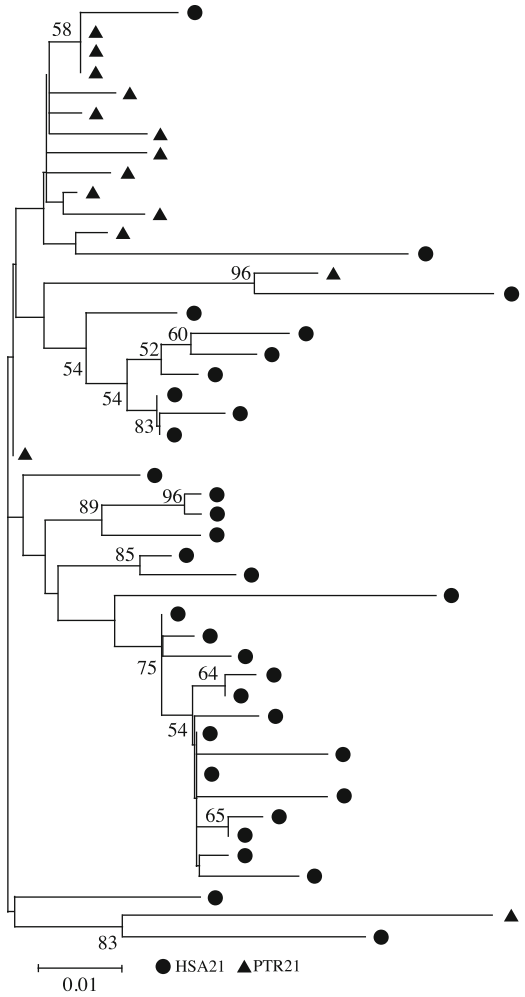

Fig. 2.11 A phylogenetic tree of human-specific and chimpanzee-specific Alu sequences (From The International Chimpanzee Chromosome 22 Consortium 2004; [6])

size, motif structure, type, and chromosomal location [61]. Among all those factors, repeat number is the strongest factor positively correlated to mutation rates of STRs [62].

Perfection status of STRs is one of the remaining factors which significantly affects STR mutation rates (e.g., [63]). Four types of status are often recognized: (1) perfect, STR purely composed of only one kind of motif; (2) imperfect, STR having one base pair which does not match the repetitive sequence; (3) interrupted, STR with short sequences within the repetitive sequences; and (4) composite (also called compound), STR with two distinctive, consecutive repetitive sequences linked. However, this classification is somehow idealistic that many of the microsatellites could not fall into any of the four categories in real cases. Algorithms searching STRs such as Sputnik, RepeatMasker, and Tandem Repeat Finder

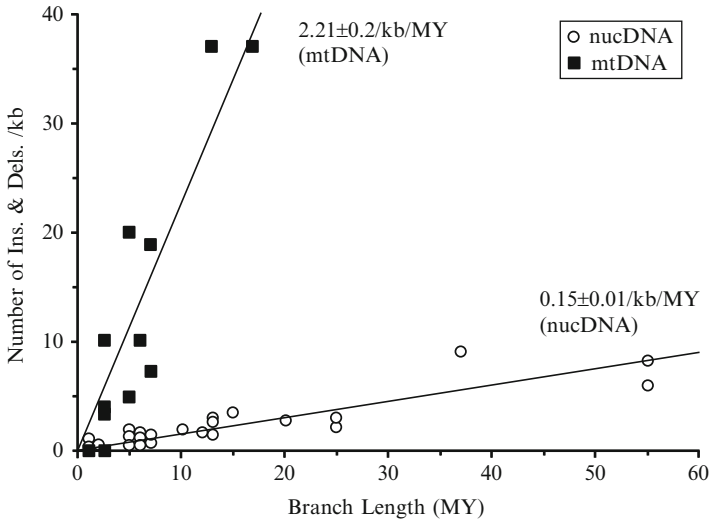

Fig. 2.12 The constancy of the evolutionary rates of insertions and deletions (From Saitou and Ueda 2004; [9])

(TRF) use different parameters, with major difference when dealing with mismatches (interruptions) and resulting nonuniform datasets [64]. Interruptions inside STRs are known to have stabilization effect, where mutation rate is greatly lowered. However, direct measure on mutation rates and comparison between imperfect microsatellites and perfect microsatellites were only recently analyzed [65].

Ngai and Saitou (2012; [66]) redefined STRs into four groups: perfect, imperfect, perfect compound, and imperfect compound. A perfect STR is defined as locus with a perfect repetitive run with its own motif type, abbreviated as the locus motif (LM) (Fig. 2.13a). An imperfect STR is defined as locus with a repetitive run that contains interruptions. Each interruption is from 1 to 10 bp long (Fig. 2.13b). When interruption is more than 10 bp, it is considered as a perfect locus. Besides perfect and imperfect, a locus could also be either compound or noncompound. A compound microsatellite is defined as locus which contains repetitive sequence composed of a non-locus motif, abbreviated as non-LM, where the repeat number passes the threshold value (say 3 repeats) and is within 10-bp flanking region of the locus (Fig. 2.13c, d).

Recently, direct sequence comparison of STR repeat numbers was conducted for parent–offspring pairs [67]. They examined ~2,500 STR loci for ~85,000 Icelanders and found more than 2,000 de novo repeat change-type mutations. Father-to-offspring mutations were three times higher than those for mother-to-offspring, and the mutation rate doubled as the father's age changed from 20 to 58, while no age effect was observed for mother. Mutation rate estimates per locus per generation for dinucleotide and trinucleotide STRs were 2.7×10^{-4} and 10.0×10^{-4}, respectively.

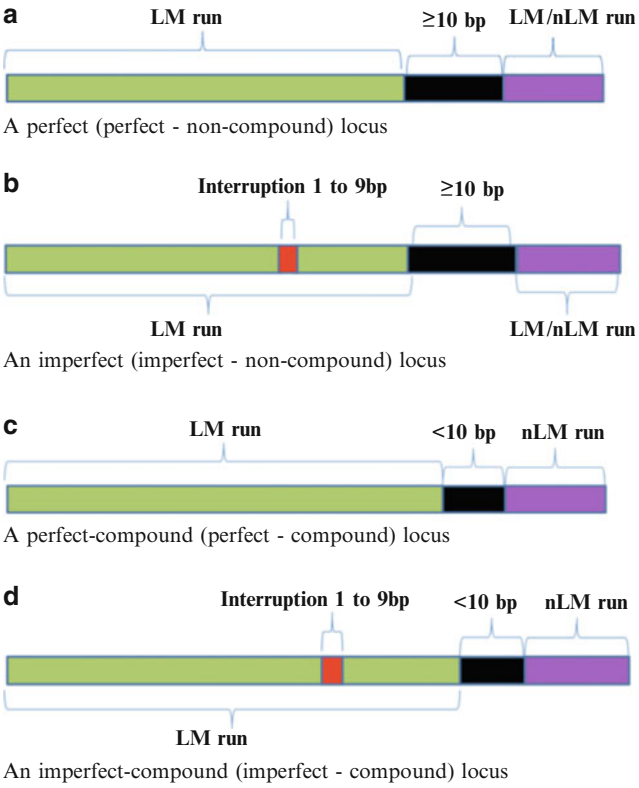

Fig. 2.13 Four types of STR loci (From Ngai and Saitou 2012; [66])

2.4 Recombination and Gene Conversion

Recombination was discovered by Thomas Hunt Morgan and his colleagues in the early twentieth century. The concept of "gene conversion" was first proposed by Winkler in 1930 [11], but it was not fully accepted for a long time, until studies on fungi clearly showed conversion events [12, 13]. Holliday (1964; [14]) proposed the "Holliday structure" model (Fig. 2.14) to connect gene conversion, or nonreciprocal transfer of DNA fragment, and recombination.

The general definition of recombination is reconnection of different nucleotide molecules. There are two types of recombination: homologous and nonhomologous. Homologous recombination usually occurs through crossing-over during meiosis, as already discussed using Figs. 2.2b and 2.3a. We restrict our discussion in this section only to eukaryotes, for "recombination" in prokaryotes are quite different.

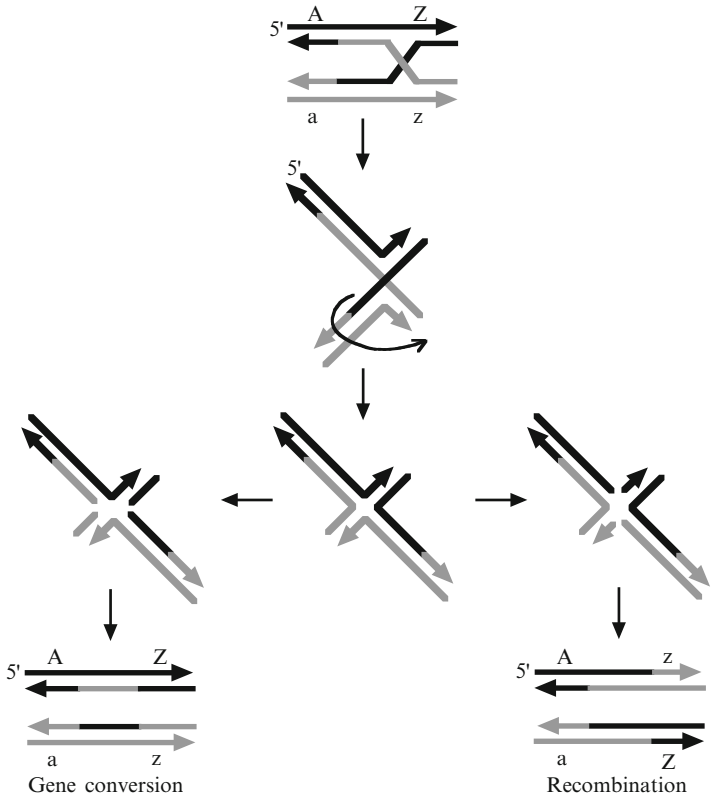

Fig. 2.14 Holliday structure model

2.4.1 Nature of Gene Conversion

Early studies on gene conversion were mostly restricted to fungal genetics. As molecular evolutionary studies of multigene family started, unexpected similarity of tandemly arrayed rRNA genes was found [15]. This phenomenon was termed "concerted evolution," and gene conversion or unequal crossing-over was proposed to explain this characteristic of some multigene families (e.g., [16]). New statistical methods were developed to detect gene conversion between homologous sequences [17, 18]. Program GENECONV developed by Sawyer [19] became the standard tool for analyzing gene conversions. We now know that conversion can occur in any genomic region irrespective of genes (DNA regions having function) or nongenic regions (e.g., [20]). However, "gene conversion" as technical jargon is currently widely accepted, and I follow this nomenclature. Gene conversion can be classified into two types: intragenic or between alleles and intergenic or between duplicated genes.

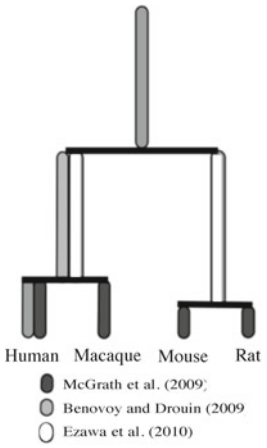

Fig. 2.15 Branches of phylogenetic trees where duplication events were used in three studies. Branch length proportions in terms of evolutionary time are after Ezawa et al. (2010; Ref. [27])

2.4.2 Gene Conversion on Duplog Pairs

After human genome sequencing was "completed" [21], genomes of mouse [22], rat [23], rhesus macaque [24], and other mammalian species were determined. Because genome sequences of these four species (human, macaque, mouse, and rat) are in good condition than those of other mammalian species, genome-wide analyses of gene conversion among duplogs (duplicated genes or DNA regions in one genome) were confined to these species. Boney and Drouin (2009; [25]) studied human duplogs, while Macgrath et al. (2009; [26]) studied young duplogs after speciation of human–macaque and mouse–rat, and Ezawa et al. (2010; [27]) used duplogs before these two speciations (see Fig. 2.15).

A number of duplog pairs used in three studies are 55,050, 3,996, and 1,121 for [25, 26], and [27], respectively. The total number (55,050) of duplog pairs used by [25] was much larger than the number (27,350) of known protein coding gene sequences they used. It is possible that many of these duplog pairs were counted more than once. The human genome has gene families with many copy numbers. For example, the number of functional olfactory receptor genes was estimated to be 388 [28]. Multiple countings of these large-sized multigene families could be a reason to reach such large duplog pairs. Independent duplog sets are preferable for statistical tests, and Ezawa et al. [27] carefully eliminated multiple countings, which were included in their previous study [29]. It is not clear whether McGrath et al. [26] also excluded double counting from their "Methods" section, but they focused on young duplogs (see Fig. 2.1), and two or more duplications in these relatively short evolutionary times for one gene may not be frequent.

It is interesting to compare frequency of gene duplication between primate and rodent lineages. A total of 549 and 363 duplog pairs were extracted from human and rhesus macaque genomes, respectively, in [27], while 1,913 and 1,171 pairs were found from mouse and rat genomes, respectively. Ezawa et al. [27] found 430 and

691 duplog pairs from primate and rodent lineages, respectively. Because primates and rodents started to diverge about 80–95 million years ago [27], we can compare the total number of duplications for these two lineages with the same evolutionary time: 886 (=[549 + 363]/2 + 430) for primate lineage and 2,233 (=[1,913 + 1,171]/2 + 691) for rodent lineage. It seems that the rodent lineage has more than two times higher rate of gene duplication. However, it is possible that gene duplication is more proportional to the number of nucleotide substitutions rather than evolutionary time. If we use neutral nucleotide divergence data shown in Fig. 2.2b of [27], duplication events can be normalized as 6,329 (=886/0.14) for primates and 8,270 (=2,233/0.27) for rodents per one nucleotide substitution per site. These two values are not much different. This suggests that the rate of gene duplication in mammals is more or less proportional to nucleotide substitutions. Because tandem gene duplication is usually assumed to start from unequal crossing-over that happens in meiosis, it is possible that nucleotide substitutions also happen during meiosis. In any case, it is clear that we need to have ample knowledge on mammalian duplogs if we are interested in intergenic gene conversion in mammalian genomes.

When Winkler [11] proposed gene conversion in 1930, it was a deviation from the Mendelian ratio. Later, detailed observations on baker's yeast and Neurospora [12, 13] established gene conversion, and Holliday's [14] model transformed gene conversion from phenomenon to mechanism. Nowadays several enzymes are known to be involved in DNA strand exchanges [30]. Abundant genome sequence data and their computational analyses again turned gene conversion or more flatly homogenization of homologous sequences from mechanism to phenomenon. We should be careful of any prejudice to a particular phenomenon when we try to interpret them with certain mechanism. One phenomenon, such as homologous sequence homogenization, may occur not only via gene conversion but with some other mechanisms, including one unknown to us at this moment. It is obvious that we should grasp molecular mechanism of gene conversion, including enzymatic machineries.

2.5 Gene Duplication

2.5.1 Classification of Duplication Events

Duplication of DNA fragment can happen in any region of chromosomes, but historically duplicated regions containing protein coding genes were the focus of research on duplication. Therefore, when we mention "gene duplication," nongenic regions may also be included. Under this broad meaning, gene duplication can be classified into four general categories: (1) tandem duplication, (2) RNA-mediated duplication, (3) drift duplication [31], and (4) genome duplication.

Tandem duplication results in two homologous genes in close proximity with each other in the same chromosome via unequal crossing-over (see Fig. 2.16a), while RNA-mediated duplication can create duplicate copies, complementary to original RNA molecules, far from the original gene with the help of reverse transcriptase

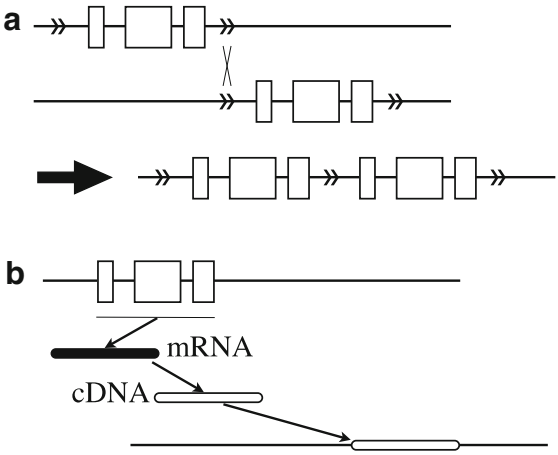

Fig. 2.16 Mechanisms of gene duplications. (**a**) Unequal crossing over. (**b**) Retrotransposition

(see Fig. 2.16b). Retrotransposons including SINEs (short interspersed elements) and LINEs (long interspersed elements) are in this category. Alu and L1 sequences are representatives of SINEs and LINEs in the human genome, respectively. This type of duplication is also characterized by intronless sequence, when mature mRNAs, formed after introns are spliced, are reverse transcribed. Intronless copies are also called "processed genes" because they went through processing called "splicing." Protein coding genes require appropriate enhancers and promoters to be transcribed. Therefore, these processed genes are most probably not transcribed and functionless. Most probably, they are "dead on arrival," that is, they become nonfunctional immediately after duplication. Because of this nature, these duplicates are often called "processed pseudogenes." It should be noted that immature mRNA before splicing may have the possibility of reverse transcription.

2.5.2 Drift Duplication

Ezawa et al. (2011; [31]) recently proposed a new category of gene duplication and named it "drift" duplication. Its physical distance distribution appears to peak around a few hundred Kb for vertebrates and a few dozen Kb for invertebrates, which is in between those of tandem duplication (short range) and retrotransposition (long range, i.e., mostly unlinked). Drift duplications are almost randomly oriented, with the frequency ratio of head-to-tail:tail-to-tail:head-to-head $\approx$ 2:1:1, as opposed to tandem duplications due to unequal crossing-overs, which are mostly head-to-tail. A drift duplication can also create multi-exon duplogs, as opposed to retrotransposition, whose products are mostly intronless. Retrotransposition is also drifting in a sense; however, it always passes through the RNA stage. This is the clear difference from drift duplication. With this name, "drift," Ezawa et al. [31] also implied that even some interchromosomal duplications may be attributed from drift duplication,

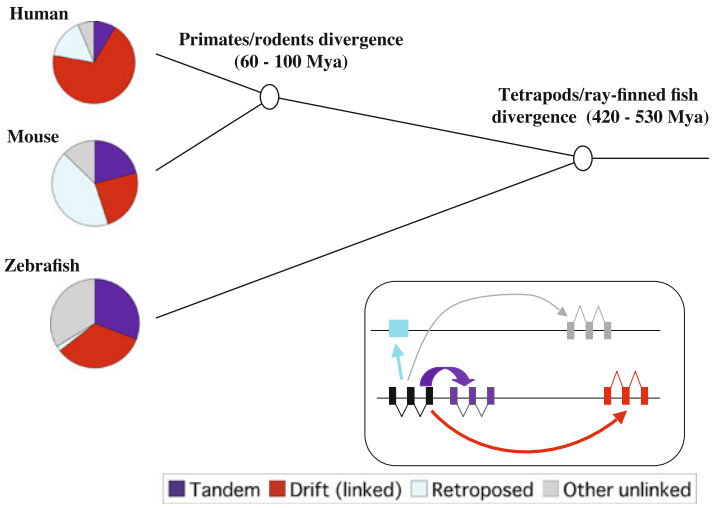

Fig. 2.17 Proportions of duplication types in vertebrate genomes (From Ezawa et al. 2011; [31])

though RNA-mediated duplications may be more frequent among interchromosomal duplications. DNA molecules for drift duplications are usually much larger than those for RNA-mediated duplications and may not be able to move to different chromosomes easily. This conjecture should be examined in future studies. Figure 2.17 shows proportions of duplication types in vertebrate genomes.

2.5.3 Genome Duplication

The last type of duplog generation is genome duplication. It is called "polyploidization" in plants, and this was probably why the word "genome" was coined by Hans Winkler (1920; [32]), a botanist. Genome duplication has never been demonstrated in prokaryotes (Bacteria and Archaea) and is rare in eukaryotes except for plants and some vertebrates (see Chap. 8). Genome duplication is important when we consider the origin of vertebrates, but after two-round genome duplications, the mammalian lineage has not experienced further genome duplication.

2.6 Estimation of Mutation Rates

2.6.1 Direct and Indirect Method

Estimation of mutation rates is quite important, for the mutation is the ultimate source of evolution. The natural way to estimate the mutation rate is to compare nucleotide sequences between parents and offsprings. This is called direct method. However, it was difficult for a long time to directly compare a large number of

nucleotide sequences. Therefore, examination of various phenotypes was used to estimate the mutation rate. Even now, this kind of study is conducted, e.g., Watanabe et al. (2009; [33]).

Because of the difficulty of direct estimates of mutation rates from comparison of large nucleotide sequences, various methods were used to indirectly estimate mutation rates, considering the balance among mutation, selection, and random genetic drift (see Chap. 3 of Nei (1987; [34]) for review). However, a more straightforward way is to compare neutrally evolving nucleotide sequences of relatively closely related species, as we will see in Sect. 2.6.3.

2.6.2 Direct Method

Various types of mutations were described in this chapter, and accumulation of these mutations through a long time period is the source of evolution. Therefore, the rate of mutation per generation, or mutation rate, is closely related to the rate of evolution. Classically, the temporal unit of mutation rate is one generation. This is because sudden changes of phenotypes between parent and children were fundamental for mutations in the classic sense. Estimation of mutation rates through comparison of parents and offsprings is called "direct" method. In this case, a temporal unit of the mutation rate is one generation.

The rate of mutation for human was estimated for the first time by using data on achondroplasia, an autosomal dominant Mendelian trait (OMIM #100800). This inherited disease is caused by a mutation that occurred in the gene coding for the fibroblast growth factor receptor-3 (FGFR3) at the short arm of human chromosome 4. The most frequent type is nonsynonymous substitution at the 380th codon, changing glycine to arginine. Haldane (1949; [35]) used data collected in Copenhagen, the capital of Denmark; 10 babies were achondroplastic out of 94,075 newborn babies. Two of them inherited the mutant gene from parents, for one parent was also achondroplastic. Therefore, the number of fresh or de novo mutations was eight. The mutation rate is thus estimated to be 4.3×10^{-5} ($= 8/[94{,}075 \times 2]$) per generation per gene. An alternative estimate, 1.4×10^{-5}, based on 7 mutants out of 242,257 births [36], is about 1/3 of the estimate originally obtained by Haldane [35].

As already mentioned in this chapter, there are various units of mutation rates for both temporal and spacial situations. Temporal units include one generation, one meiosis, one cell division, and 1 year, while typical spacial units are one gene, one nucleotide, 1 kb, or the whole genome. Because one cell division corresponds to one generation for unicellular asexual organisms such as prokaryotes, one cell division may be a good universal temporal unit for the mutation rate. However, naturally occurring radiations such as cosmic ray or background radiation as well as chemical mutagens may affect an organism at any time. Therefore, a physical time, such as 1 year, may also be a universal temporal unit of the mutation rate. Because some types of mutations may occur only during meiosis, one meiosis may be suitable for this kind of mutations. In the case of achondroplasia, the physical unit of the mutation rate was one gene that consists of up to more than one million nucleotides.

In the 1960s, the use of protein electrophoresis, particularly using starch gel (Smithies, 1995; [37]), became popular for detecting protein variations for many organisms. Because the amino acid sequence information is closely related to the nucleotide sequence (see Chap. 1), this technique should give much better estimate of mutation rates at the nucleotide sequence level. A mass scale study for estimating the mutation rate was conducted using protein electrophoresis for Japanese individuals who were exposed to various degrees of radiation from atomic bomb explosions at Hiroshima and Nagasaki on August 6 and 9, 1945, respectively. The Atomic Bomb Casualty Commission (ABCC) was created by the US government, and many human individuals were examined. One of the final results, published in 1986 led by James V. Neel [38], reported 3 mutations out of 539,170 gene transmissions from parent to offspring. The mutation rate was then estimated to be 1×10^{-8} per nucleotide site per generation. This value is about half of the estimate obtained by comparison of nucleotide sequences shown in the next paragraph.

DNA sequencing (see Chap. 11) became popular in the 1980s, but a huge effort of nucleotide sequencing was necessary to estimate the mutation rates by directly comparing parents and offspring. Therefore, this type of studies was started to be published in this century. Kondrashov (2002; [39]) assembled nucleotide sequence data for 20 Mendelian genetic disease-causing genes and estimated the human nuclear mutation rate to be 1.78×10^{-8} per nucleotide site per generation. The majority (1.7×10^{-8}) of them were nucleotide substitutions, and insertion-type mutations were 1/3 of deletion-type mutations. If we consider one generation of modern-day human as 30 years, the rate of substitution type mutation per nucleotide site per year becomes 0.56×10^{-9}. The mutation rate of insertions and deletions was estimated to be 8×10^{-10} per nucleotide site per generation, and it corresponds to 3×10^{-11}/site/year. This is 13 times lower compared to that (3.8×10^{-10}/site/year) obtained from the comparison of human and chimpanzee genomic sequences [6]. Probably the total amount of compared nucleotide sequences was too small to obtain a reliable estimate.

Recently, thanks to the so-called second-generation sequencer (see Chap. 11), genome-wide comparison of parents and offspring was made, and the mutation rate was estimated to be ~1.1×10^{-8} per nucleotide site per generation [40, 41]. Because one generation is about 30 years in human, the rate becomes ~0.4×10^{-9} per nucleotide site per year. During the final process of editing this book, two additional papers [58, 59] were published on this matter, and both showed 1.2×10^{-8} per nucleotide site per generation, which is slightly higher than the estimate obtained by two previous studies [40, 41].

So far, we discussed mutation rates of human using different kinds of data. Let us move to other animal species. More than 40-Mb nucleotide sequences were determined using the PCR-direct sequencing method, after accumulating mutations for hundreds of generations in *Caenorhabditis elegans* [42]. A total of 30 mutations (13 substitution type, 13 insertion type, and 4 deletion type) were discovered. The net mutation rate was estimated to be 2.1×10^{-8}/site/generation. This is similar to the value estimated for human, but insertion-type mutations are more than three times higher than deletion types. This is a good contrast to the estimate based on the

indirect method, where deletions are preponderant. This suggests that deletion-type mutations may be somewhat more advantageous than insertion-type mutations so as to keep the genome size smaller. A larger-scale study using second-generation sequencers (both 454 and Solexa) found 391 substitutional-type mutations out of 584-Mb sequences, resulting in the mutation rate to be 2.7×10^{-9}/site/generation [43]. If we assume that the average number of germ-line cell division per generation is 8.5 in this species, the mutation rate becomes 3.2×10^{-10}/site/cell division. This value is more or less similar to those for *S. cerevisiae* [44], *Drosophila melanogaster* [45], and human [43].

As for *Drosophila melanogaster*, a total of 37 fresh mutations were found through examination of 20-Mb sequences by combining DHPLC (denaturing high-performance liquid chromatography) and nucleotide sequencing [45]. The net mutation rate was estimated to be 8.4×10^{-9}/site/generation. Using Illumina sequencing technology, Keightley et al. (2009; [46]) sequenced three *Drosophila melanogaster* lines which accumulated mutations after 262 generations. They obtained the mutation rate to be 3.5×10^{-9}/site/generation, based on 174 de novo mutations out of 72-Mb sequences.

It is not clear how many generations may pass for 1 year in wild conditions for *Drosophila melanogaster*, but probably ten generations may correspond to 1 year (Dr. Takano-Shimizu Toshiyuki, personal communication). If so, the mutation rate becomes 8.4×10^{-8}/site/year. This is more than 100 times higher than that for human.

2.6.3 Indirect Methods

According to the neutral theory, the evolutionary rate (λ) is equal to the mutation rate in the genome region under pure neutral evolution (see Chap. 4). If we apply this idea, we can estimate the mutation rate by estimating the long-term evolutionary rate, under the simple equation:

$$D[i,j] = D[j,i] \tag{2.1}$$

where D is the evolutionary distance and T is the divergence time between the two lineages (see Fig. 2.18). This procedure is called the indirect method. For example, the substitutional difference (D) between human and chimpanzee is 1.23 % [47]. If we assume that the divergence time (T) of these two species is 6 million years, $\lambda = D/[2T] = 1 \times 10^{-9}$/site/year. Because application of the direct method usually takes a large amount of resources, this indirect method has been widely used from the advent of the molecular evolutionary studies. For example, Wu and Li (1985; [48]) showed that rodents seem to have higher mutation rate than primates.

Some cautions should be taken for this method. First of all, some genomic regions may not be under pure neutral evolution, but under purifying selection, as in the case of conserved noncoding sequences (e.g., Takahashi and Saitou, 2012; [71]). In this case, the mutation rate may be underestimated. Another problem is that the estimate is the long-term average of mutation rates for the two lineages. Generation

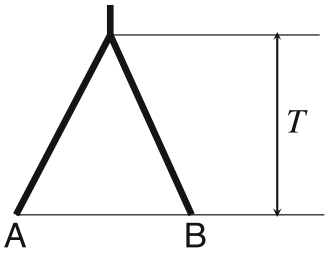

Fig. 2.18 A schematic phylogenetic tree for two lineages, A and B

times greatly vary from species to species. We are seeing only the average pattern of long-term evolution. This is a clear difference from results of the direct method, where a snapshot of the current populations is obtained.

2.6.4 Mutation Rates of Bacteria, Organella, and Viruses

Mutation rate estimates in prokaryotes can differ more than ten times between those using the direct method and those using the indirect method. Assuming the divergence time between *E. coli* and *Salmonella* to be about 100 million years ago, the long-term mean evolutionary rate was estimated to be 4.5×10^{-9}/site/year [49]. This was based on the number (0.9) of synonymous substitution per site, and it can be equated to be the mutation rate. On the other hand, the direct estimate based on lacZ revertants was ~5×10^{-10}/site/generation [49]. Because the number of generations per year for *E. coli* is at least 100, this mutation rate becomes ~5×10^{-8}/site/year.

Organella such as mitochondria and chloroplasts in eukaryotes have their own DNA and replicate them independently from nuclear DNA. Mutation rates of nuclear DNA and organella DNA are thus usually different. The mutation rate of animal mitochondria is more than ten times higher than that of nuclear DNA. However, among-lineage rate heterogeneity is considerable. There are also reports on time dependency. According to Burridge et al. (2008; [68]), pedigree-based mutation rate is the fastest (e.g., 5.1×10^{-7}/site/year) in human, followed by estimates based on 10,000-year-old ancient DNA (3.4–4.4×10^{-7}/site/year), and the slowest rate was obtained from phylogenetic estimates derived from Neogene primate divergences (0.5–2.4×10^{-7}/site/year). In birds, Millar et al. (2008; [69]) used Adélie penguins (*Pygoscelis adeliae*) whose ancient DNA samples were available. Direct mother–offspring mutation rate estimate and indirect estimate using 37,000-year-old ancient DNA samples gave 5.5×10^{-7}/site/year and 8.6×10^{-7}/site/year, respectively. Because these two rates are more or less the same, time dependency does not seem to exist in birds. In fish, however, Burridge et al. (2008; [68]) reported that the mtDNA substitution rates of galaxiid fishes from calibration points younger than 200 kyr (2–11×10^{-8}/site/year) were faster than those (0.8–5×10^{-8}/site/year) based on older calibration points, indicating the existence of time dependency.

In contrast, the mutation rate of plant mitochondria is about 1/10 of that of nuclear DNA. Interestingly, the mutation rate of plant chloroplast DNA is intermediate between those for mitochondrial DNA and nuclear DNA. The reason for such differences is not known.

The so-called RNA viruses have RNA genomes, and some of them have their own RNA replicase. Their replication error rates or mutation rates can be quite high than those of DNA genome organisms. Hanada et al. (2004; [50]) estimated rates of synonymous substitutions for 49 RNA viruses and showed that many of their rates were ~10^{-3}/site/year, though the whole range was heterogeneous, from 1×10^{-7}/site/year to 1×10^{-2}/site/year. Because the error rate of RNA replication is ~10^{-5} per site per replication, variation of replication cycles per unit of time seems to contribute to heterogeneity of the mutation rate per site per year in RNA viruses [50].

2.7 Mutation Affecting Phenotypes

2.7.1 What Is Phenotype?

Various conditions of organisms are collectively called "phenotypes." Phenotypes include macroscopic characters such as seed surface form (round or wrinkled) studied by Mendel (1866; [51]), human height, amino acid sequence of proteins, or even DNA sequences themselves. Genetics has been the science of connecting genotypes and phenotypes. Genotypes are straightforward and objective; one genotype corresponds to a specific nucleotide sequence. In contrast, phenotypes are products of human recognition. It is true that the operational definition of one phenotype is always possible, such as human head length. It is, however, not clear what kind of biological significance exists in head length variation. We should be careful about the subjective nature of phenotypes. It may be ideal to only consider phenotypes which has one-to-one correspondence with one genotype.

2.7.2 Mutations in Protein Coding Region

Mutations are changes in nucleotide sequences, but they may change amino acid sequences if they happen in protein coding regions. We already discussed three types of nucleotide substitutions occurring in protein coding regions at Sect. 2.2.2. Even if amino acid is changed, some changes may not affect protein function. In this case, this mutant may be selectively neutral, as we will see in Chap. 4. When the amino acid change occurred in regions important for protein function, often such changes will reduce protein function. One example is the nonsynonymous mutation at ABCC11 gene in the human genome (see Fig. 2.19). Amino acid change from glycine to arginine essentially nullified transporter function of this protein (Yoshiura et al. 2006; [52]).

Another classic example is the emergence of sickle-cell anemia that is resistant to malaria [53]. Hemoglobin is composed of heme (porphyrin) and globin (protein),

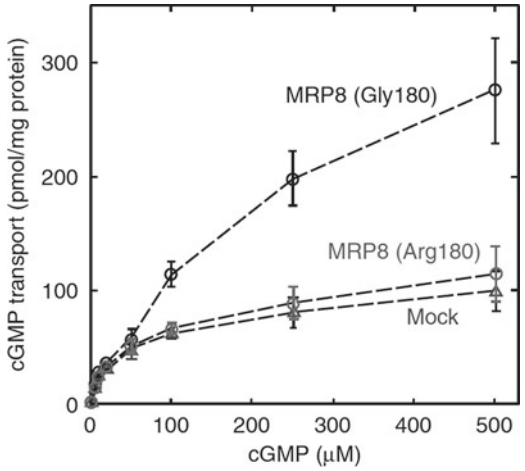

Fig. 2.19 Functional differences of two proteins coded in ABCC11 gene (From Yoshiura et al. 2006; [52])

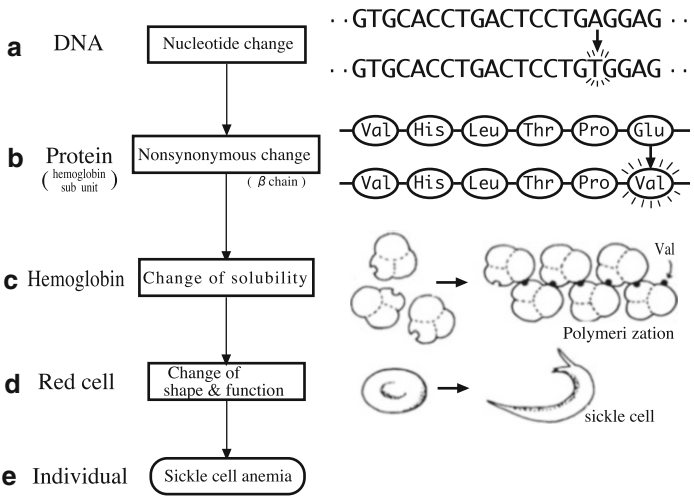

Fig. 2.20 Sickle cell anemia caused by one nonsynonymous substitution (From Saitou 2007; [70])

and its main function in animals is to transport oxygen to the entire body. Figure 2.20a shows normal hemoglobin A gene sequence and mutated hemoglobin S sequence. A to T nonsynonymous change caused Glu to Val change at position 6; see Fig. 2.20b. This position is not a part of the heme pocket where several amino acids are anchoring heme, but is located in globin surface. This change created a slight hump in the protein surface and can be connected to a hollow; see Fig. 2.20c. This produces linear globin polymer, and polymerization continues until all mutant proteins are connected. These polymers will form fibers and eventually cross the red cell body.

```
A allele   aaggatgtcctcgtggtgacccctggctggctggctcccattgtctgggaggg
O1 allele  aaggatgtcctcgtggt-acccctggctggctggctcccattgtctgggaggg
A protein  LysAspValLeuValValThrProTrpLeuAlaProIleValTrpGluGlyThr
O protein  LysAspValLeuValVal ProLeuGlyTrpLeuProLeuSerGlyArgAla
```

Fig. 2.21 O-1 allele of the ABO blood group gene sequence (Based on Ref. [54])

This is why normally round red cell turns to sickle shape; see Fig. 2.20d. Sickle cells do not easily go through capillary vessels, and anemia will be the final result for the human individual. This is a good example that even a single nucleotide change can affect the whole individual.

If insertions or deletions occur, they will shift protein coding frames unless their numbers are in multiplication of 3. These mutations are thus called "frameshift" mutations, and most of them will no longer produce functional proteins. One such example from the ABO blood group O allele [54] is shown in Fig. 2.21. It should be noted that this example is rather unique, for the O-1 allele that cannot produce the functional enzyme (A or B) has high frequency in human populations. If one protein is indispensable for organism, such frameshift-type mutations will be quickly eliminated from populations, as we will see in Chap. 5.

Another loss-of-function-type mutation is insertion of transposons. Mendel (1866; [51]) studied seven characters of pea (*Pisum sativum*), and one of them is seed shape. Round type was dominant to wrinkled type. After more than 100 years of his study, a British group discovered that a 0.8-kb-long transposon insertion caused functional protein gene to be nonfunctional [55]. This gene encodes a starch-branching enzyme, and nonfunctional gene somehow causes pea skin to be wrinkled.

2.7.3 Mutations in Noncoding Region

Protein expressions are controlled in various ways. One of them is transcription control, and there are two types; trans and cis. DNA sequences responsible for cis control is often called "cis-regulatory element." Some of these functions were discovered by their loss-of-function-type mutations that affected phenotypes (e.g., [56]). One classic example of temporal control of gene expression is lactose tolerance. Mammalian babies, by definition, drink mother's milk as their main food source, and lactose is abundant in milk. They express enzyme called lactase, and this cuts lactose into glucose and galactose. After the lactation period, gene expression of lactase is stopped. In some human individuals, however, lactase expression continues to adulthood, called lactose tolerance. Through association studies, it was found that single nucleotide polymorphism is located at one intron of the adjacent gene to lactose gene, LCT, is the origin of the lifetime expression of lactase gene [57].

References

1. Kato, L., Stanlie, A., Begum, N. A., Kobayashi, M., Aida, M., & Honjo, T. (2012). An evolutionary view of the mechanism for immune and genome diversity. *Journal of Immunology, 188*, 3559–3566.
2. Haldane, J. B. S. (1947). The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Annals of Eugenics 13*, 262–271. (cited by Ref. [3])
3. Crow, J. F. (1997). The high spontaneous mutation rate: Is it a health risk? *Proceedings of the National Academy of Sciences of the United States of America, 94*, 8380–8386.
4. Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K., & Yasunaga, T. (1987). Male-driven molecular evolution: A model and nucleotide sequence analysis. *Cold Spring Harbor Symposia on Quantitative Biology, 52*, 863–867.
5. Topal, M. D., & Fresco, J. R. (1976). Complementary base pairing and the origin of substitution mutations. *Nature, 263*, 285–289.
6. The International Chimpanzee Chromosome 22 Consortium. (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature, 429*, 382–388.
7. Gojobori, T., Li, W.-H., & Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of Molecular Evolution, 18*, 360–369.
8. Gojobori, T. (1983). Codon substitution in evolution and the "saturation" of synonymous changes. *Genetics, 105*, 1011–1027.
9. Saitou, N., & Ueda, S. (1994). Evolutionary rate of insertions and deletions in non-coding nucleotide sequences of primates. *Molecular Biology and Evolution, 11*, 504–512.
10. Ophir, R., & Graur, D. (1997). Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene, 205*, 191–202.
11. Winkler, H. (1930). *Die Konversion der Gene*. Jena: Verlag von Gustav Fischer (written in German).
12. Lindegren, C. C. (1953). Gene conversion in *Saccharomyces*. *Journal of Genetics, 51*, 625–637.
13. Michell, L. B. (1955). Aberrant recombination of pyridoxine mutants of Neurospora. *Proceedings of the National Academy of Sciences of the United States of America, 41*, 215–220.
14. Holliday, R. A. (1964). Mechanism for gene conversion in fungi. *Genetic Research Cambridge, 5*, 282–304.
15. Brown, D. D., Wensink, P. C., & Jordan, E. (1972). A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: The evolution of tandem genes. *Journal of Molecular Biology, 63*, 57–73.
16. Eickbush, T. H., & Eickbush, D. G. (2007). Finely orchestrated movements: Evolution of ribosomal RNA genes. *Genetics, 175*, 477–485.
17. Stephens, C. (1985). Statistical methods of DNA sequence analysis: Detection of intragenic recombination or gene conversion. *Molecular Biology and Evolution, 2*, 539–556.
18. Sawyer, S. A. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution, 6*, 526–538.
19. Sawyer, S. A. (1999). *GENECONV: A computer package for the statistical detection of gene conversion*. Available at http://www.math.wustl.edu/~sawyer
20. Kawamura, S., Saitou, N., & Ueda, S. (1992). Concerted evolution of the primate immunoglobulin a-gene through gene conversion. *Journal of Biological Chemistry, 267*, 7359–7367.
21. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*, 860–921.
22. Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature, 420*, 520–562.
23. Rat Genome Sequencing Project Consortium. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature, 428*, 493–521.
24. Rhesus Macaque Genome Sequencing and Analysis Consortium. (2007). Evolutionary and biological insights from the rhesus macaque genome. *Science, 316*, 222–234.
25. Benovoy, D., & Drouin, G. (2009). Ectopic gene conversions in the human genome. *Genomics, 93*, 27–32.

26. McGrath, C. L., Casola, C., & Hahn, M. W. (2009). Minimal effect of ectopic gene conversion among recent duplicates in four mammalian genomes. *Genetics, 182*, 615–622.
27. Ezawa, K., Ikeo, K., Gojobori, T., & Saitou, N. (2010). Evolutionary pattern of gene homogenization between primate-specific paralogs after human and macaque speciation using the 4-2-4 method. *Molecular Biology and Evolution, 27*, 2152–2171.
28. Nei, M., Niimura, Y., & Nozawa, M. (2008). The evolution of animal chemosensory receptor gene repertoires: Roles of chance and necessity. *Nature Reviews Genetics, 9*, 951–963.
29. Ezawa, K., OOta, S., & Saitou, N. (2006). Genome-wide search of gene conversions in duplicated genes of mouse and rat. *Molecular Biology and Evolution, 23*, 927–940.
30. Liu, Y., & West, S. C. (2004). Happy Hollidays: 40th anniversary of the Holliday junction. *Nature Reviews Molecular Cell Biology, 5*, 937–944.
31. Ezawa, K., Ikeo, K., Gojobori, T., & Saitou, N. (2011). Evolutionary patterns of recently emerged animal duplogs. *Genome Biology and Evolution, 3*, 1119–1135.
32. Winkler, H. (1920). *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche*. Jena: Fischer (written in German).
33. Watanabe, Y., et al. (2009). Molecular spectrum of spontaneous de novo mutations in male and female germline cells of *Drosophila melanogaster*. *Genetics, 181*, 1035–1043.
34. Nei, M. (1987). *Molecular evolutionary genetics*. New York: Columbia University Press.
35. Haldane, J. B. S. (1949). The rate of mutation of human genes. *Hereditas, 35*, 267–273.
36. Gardner, R. J. (1977). A new estimate of the achondroplasia mutation rate. *Clinical Genetics, 11*, 31–38.
37. Smithies, O. (1995). Early days of gel electrophoresis. *Genetics, 139*, 1–4.
38. Neel, J. V., Satoh, C., Goriki, K., Fujita, M., Takahashi, N., Asakawa, J., & Hazama, R. (1986). The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides. *Proceedings of the National Academy of Sciences of the United States of America, 83*, 389–393.
39. Kondrashov, A. S. (2002). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human Mutation, 21*, 12–27.
40. Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., & Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science, 328*, 636–639.
41. Conrad, D. F., et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics, 43*, 712–715.
42. Denver, D. R., Morris, K., Lynch, M., & Thomas, W. K. (2004). High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature, 430*, 679–682.
43. Denver, D. R., Dolan, P. C., Wilhelm, L. J., Sung, W., Lucas-Lledó, J. I., Howe, D. K., Lewis, S. C., Okamoto, K., Thomas, W. K., Lynch, M., & Baer, C. F. (2009). A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 16310–16314.
44. Lynch, M., et al. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 9272–9277.
45. Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B., & Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature, 445*, 82–85.
46. Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., & Blaxter, M. L. (2009). Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research, 19*, 1195–1201.
47. Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T. D., Itoh, T., Tsai, S.-F., Park, H.-S., Yaspo, M.-L., Lehrach, H., Chen, Z., Fu, G., Saitou, N., Osoegawa, K., de Jong, P. J., Suto, Y., Hattori, M., & Sakaki, Y. (2002). Construction and analysis of a human-chimpanzee comparative clone map. *Science, 295*, 131–134.
48. Wu, C.-I., & Li, W.-H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences of the United States of America, 82*, 1741–1745.

49. Ochman, H. (2003). Neutral mutations and neutral substitutions in bacterial genomes. *Molecular Biology and Evolution, 20*, 2091–2096.
50. Hanada, K., Suzuki, Y., & Gojobori, T. (2004). A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. *Molecular Biology and Evolution, 21*, 1074–1080.
51. Mendel, G. (1866). Versuche uber Pflanzenhybriden (written in German). *Verhandlungen des Naturforschenden Verenines, Abhandlungen, Brunn, 4*, 3–47.
52. Yoshiura, K., et al. (2006). A SNP in the ABCC11 gene is the determinant of human earwax type. *Nature Genetics, 38*, 324–330.
53. Branden, C., & Tooze, J. (1991). *Introduction to protein structure* (p. 40). New York: Garland.
54. Yamamoto, F., Clausen, H., White, T., Marken, J., & Hakomori, S. (1990). Molecular genetic basis of the histo-blood group ABO system. *Nature, 345*, 229–233.
55. Bhattacharyya, M. K., Smith, A. M., Ellis, T. H., Hedley, C., & Martin, C. (1990). The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell, 60*, 115–122.
56. Wray, G. A. (2007). The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics, 8*, 206–216.
57. Enattah, N. S., et al. (2002). Identification of a variant associated with adult-type hypolactasia. *Nature Genetics, 30*, 233–237.

Additional Citations Not Ordered According to Text Locations

58. Kong, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature, 488*, 471–475.
59. Campbel, C. D. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics, 44*, 1277–1283.
60. Ellegren, H. (2004). Microsatellites: Simple sequence with complex evolution. *Genetics, 5*, 435–445.
61. Bhargava, A., & Fuentes, F. F. (2010). Mutational dynamics of microsatellites. *Molecular Biotechnology, 44*(3), 250–266.
62. Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research, 18*, 30–38.
63. Oliveira, E. J., Padua, J. G., Zucchi, M. I., Vencovsky, R., & Vieira, M. L. (2006). Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology, 29*(2), 294–307.
64. Leclercq, S., Rivals, E., & Jarne, P. (2007). Detecting microsatellites within genomes: Significant variation among algorithms. *BMC Bioinformatics, 8*, 125.
65. Boyer, J. C., Hawk, J. D., Stefanovic, L., & Farber, R. A. (2008). Sequence-dependent effect of interruptions on microsatellite mutation rate in mismatch repair-deficient human cells. *Mutation Research, 640*, 89–96.
66. Ngai, M. Y., & Saitou, N. (2012). The effect of perfection status on mutation rates of microsatellites in primates (Unpublished).
67. Sun, J. X., et al. (2012). A direct characterization of human mutation based on microsatellites. *Nature Genetics, 44*, 1161–1165.
68. Burridge, C. P., et al. (2008). Geological dates and molecular rates: Fish DNA sheds light on time dependency. *Molecular Biology and Evolution, 25*, 624–633.
69. Millar, C. D., et al. (2008). Mutation and evolutionary rates in Adélie Penguins from the Antarctic. *PLoS Genetics, 4*, e1000209.
70. Saitou, N. (2007). *Genomu Shinkagaku Nyumon (written in Japanese, meaning 'Introduction to evolutionary genomics')*. Tokyo: Kyoritsu Shuppan.
71. Takahashi, M., & Saitou, N. (2012). Identification and characterization of lineage-specific highly conserved noncoding sequences in mammalian genomes. *Genome Biology and Evolution, 4*, 641–657.

http://www.springer.com/978-1-4471-5303-0

Introduction to Evolutionary Genomics
Saitou, N.
2013, XXIII, 461 p. 227 illus., Hardcover
ISBN: 978-1-4471-5303-0