Neutral Evolution

5

Chapter Summary

Neutral evolution is the default process of genomic changes. This is because our world is finite, and the randomness, indispensable for neutral evolution, is important when we consider the history of a finite world. The random nature of DNA propagation is discussed using branching process, coalescent process, Markov process, and diffusion process. Expected evolutionary patterns under neutrality are then discussed on fixation probability, rate of evolution, and amount of DNA variation kept in population. We then discuss various features of neutral evolution starting from evolutionary rates, synonymous and nonsynonymous substitutions, junk DNA, and pseudogenes.

5.1 Neutral Evolution as the Default Process of Genomic Changes

It is now established that the majority of mutations fixed during evolution are selectively neutral, as amply demonstrated by Kimura [1] and by Nei [2]. Reports of many genome sequencing projects routinely mention neutral evolution in the twenty-first century, e.g., mouse genome paper in 2002 [3] and chicken genome in 2004 [4]. We thus discuss neutral evolution as one of the basic processes of genome evolution in this chapter.

Neutral evolution is characterized by the egalitarian nature of the propagation of selectively neutral mutants. For example, let us consider a bacterial plaque that is clonally formed. All cells in one plaque are homogeneous or have the identical genome sequences, if there are no mutations during the formation of that plaque. Because of identicalness in genome sequences, there will be no difference of genetic components for this plaque. Let us assume that six cells (A–F) at time 0 are

N. Saitou, *Introduction to Evolutionary Genomics*, Computational Biology 17,
https://doi.org/10.1007/978-3-319-92642-1_5

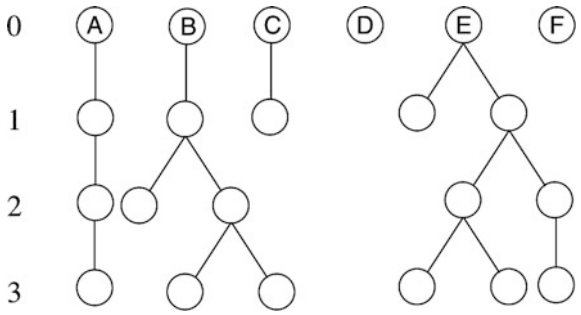

Fig. 5.1 Cell division history of six cells A–F during the three time units

in this clonal plaque (Fig. 5.1). Their descendant cells at time 3 also have the same genome sequences if there were no mutations, though the numbers of offspring cells at that time vary from 0 (C, D, and F) to 3 (cell E). This variation is attributed to nongenetic factors, such as heterogeneous distribution of nutrients. However, the most significant and fundamental factor is randomness, as we will see in Sect. 5.2.2 on branching process.

Mutation is the ultimate source of diversity of organisms. If a mutation occurring in some gene modifies gene function, there is a possibility of heterogeneity in terms of number of offsprings. This is the start of natural selection that will be discussed in Chap. 6. If genome sequences of six cells in Fig. 5.1 are heterogeneous, cell E might have some DNA sequences that have a higher ability of offspring cell production than those of cells D and F. However, some mutations may not change gene function. Although mutants and parental (wild) types are somewhat different in terms of DNA sequences, they are equal in terms of offspring propagation. This is the egalitarian characteristics of the selectively neutral mutants. If all members of evolutionary units, such as DNA molecules, cells, individuals, or populations, are identical in terms of their potential for producing their copies (offsprings), the frequency change of these types is dominated by random events. Randomness is the fundamental factor for neutral evolution.

5.1.1 Our World Is Finite

Randomness also comes in when abiotic phenomena are involved in organismal evolution. Earthquakes, volcanic eruptions, continental drifts, meteorite hits, and many other geological and astronomical events are not the outcome of biotic evolution, and they can be considered to be stochastic from organismal point of view.

Before the proposal of the neutral theory of evolution in 1968 by Kimura [5], randomness was not considered as the basic process of evolution, though some limited importance of the random genetic drift caused by finite population size was known mainly due to Wright [6]. Systematic pressure, particularly natural selection, was believed to play the major role in evolution. This view is applicable if the

population size, or the number of individuals in one population, is effectively infinite. However, the earth is finite, and the number of individuals is always finite. Even this whole universe is finite. This finiteness is the basis of the random nature of neutral evolution as we will see in later sections of this chapter.

5.1.2 Unit of Evolution

Nucleotide sequences contain genetic information, and one gene is often treated as a unit of evolution in many molecular evolutionary studies. A cell is the basic building block for all organisms except for viruses. It is thus natural to consider cell as a unit of evolution. One cell is equivalent to one individual in single-cell organisms. In multicellular organisms, by definition, one individual is composed of many cells, and a single cell is no longer a unit of evolution. However, if we consider only germ-line cells and ignore somatic cells, we can still discuss cell lineages as the mainstream of multicellular organisms as in the case of single-cell organisms. Alternatively, clonal cells of one single-cell organism can be considered to be one individual. Cellular slime mold cells form a single body with many cells, or each cell may stay independently, depending on the environmental conditions [7]. We therefore should be careful to define cell or individual.

Organisms are usually living together, and multiple individuals form one "population." We humans are sexually reproducing, and it seems obvious for us to consider one mating group. In classic population genetics theory, this reproduction unit is called "Mendelian population," after Gregor Johann Mendel, father of genetics. From an individual point of view, the largest Mendelian population is the species this individual belongs to. Asexually reproducing organisms are not necessary to form a population, and multiple individuals observed in proximity, which are often recognized as one population, may be just an outcome of past life history of the organism, and each individual may reproduce clonally. Therefore, there are many clonal lineages in one "population," and each clonal lineage may also be called "population." Gene exchanges also occur in asexually reproducing organisms, including bacteria. Therefore, by extending species concept, bacterial cells with similar phylogenetic relationship are called species. Population or species is also defined for viruses, where each virus particle is assumed as one "individual."

However, we have to be careful to define individuals and populations. One tree, such as cherry tree, is usually considered to be one individual, for it starts from one seed. Unlike most animal organisms, trees or many plant species can use part of their body to start new "individual." This asexual reproduction prompted plant population biologist John L. Harper to create terms genet (genetic individual) and ramet (physiological individual) [8]. We should thus be careful about the number of "individuals" especially for asexually reproducing organisms.

5.2 How to Describe Random Nature of DNA Propagation

We discuss the four major processes to mathematically describe the random characteristics of DNA transmission. The first two, branching process and coalescent process, are considering the genealogical relationship of gene copies, while the latter two, Markov process and diffusion process, treat temporal changes of allele frequencies.

5.2.1 Gene Genealogy Versus Allele Frequency Change

For organisms to evolve and diverge, we need changes or mutation. Supply of mutations to the continuous flow of self-replication of genetic materials (DNA or RNA) is fundamental for organismal evolution. This process is most faithfully described in phylogenetic relationship of genes. Because every organism is the product of eons of evolution, we are unable to grasp full characteristics of living beings without understanding the evolutionary history of genes and organisms. It is thus clear that the reconstruction of phylogeny of genes is essential not only for the study of evolution but also for biology in general. In another word, gene genealogy is the basic descriptor of evolution.

It should be emphasized that the genealogical relationship of genes is independent from the mutation process when mutations are selectively neutral. A gene genealogy is the direct product of DNA replication and always exists, while mutations may or may not happen within a certain time period in some specific DNA region. Therefore, even if many nucleotide sequences happened to be identical, there must be genealogical relationship for those sequences. However, it is impossible to reconstruct the genealogical relationship without mutational events. In this respect, search of mutational events from genes and their products is also important for reconstructing phylogenetic trees. Advancement of molecular biotechnology made it possible to routinely produce gene genealogies from many nucleotide sequences.

Figure 5.2 shows a schematic gene genealogy for ten genes. There are two types of genes that have small difference in their nucleotide sequences, depicted by open and filled circles. Both types are located in the same location in one particular chromosome of this organism. This location is called “locus” (plural form is “loci”), after a Latin word meaning place, and one type of nucleotide sequence is called “allele,” using a Greek word αλλο meaning “different.” Open circle allele, called allele A, is ancestral type, and filled circle allele, called allele M, emerged by a mutation shown as a star mark. The numbers of gene copies are 8 and 2 for alleles A and M, respectively. We thus define allele frequencies of these two alleles as 0.8 (=8/10) and 0.2 (=2/10), respectively. Allele frequency is sometimes called gene frequency. It should be noted that these frequencies are exact values if there are only ten genes in the population in question. If these ten genes were sampled from that population with many more genes, two values are sample allele frequency.

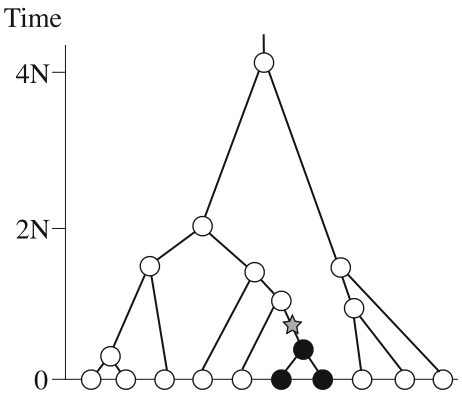

Fig. 5.2 Schematic gene genealogy for some locus of a population. Open circles and full circles designate two different alleles, and star is mutation. Timescale is in terms of generation, where N is the number of individuals. Autosomal locus of a diploid organism is assumed

Because all these ten genes are homologous at the same locus, they have the common ancestral gene. Alternatively, only descendants of that common ancestral gene are considered in the gene genealogy of Fig. 5.2. There are, however, many genes which did not contribute to the ten genes at the present time. If we consider these once existed genes, the population history may look like Fig. 5.3. In this figure, the gene genealogy starting from the filled circle gene at generation 1 is embedded with other genes which coexisted at each generation but became extinct in later generations. If we consider the whole population, it is clear that allele frequency changes continuously, and many genes shown in open circles did not contribute to the current generation. How can this allele frequency change occur? Natural selection does influence this change (see Chap. 6), but the more fundamental process is the random genetic drift. This occurs because a finite number of genes are more or less randomly sampled from the parental generation to produce the offspring generation. This simple stochastic process is the source of random fluctuation of allele frequencies through generations.

The random genetic drift can be described as follows. Let us focus on one particular diploid population with $N[t]$ individuals at generation t. We consider certain autosomal locus A, and the total number of genes on that locus at generation t is $2N[t]$. There are many alleles in locus A, but let us consider one particular allele A_i with n_i gene copies. By definition, allele frequency p_i for allele A_i is $n_i/2N[t]$. When one sperm or egg is formed via meiosis, one gene copy is included in that gamete from locus A. If males and females are assumed to have more or less the same allele frequency, the probability to have allele A_i in that gamete is p_i. This procedure is a Bernoulli trial, and the offspring generation at time $t + 1$ will be formed with $2N[t + 1]$ Bernoulli trials. Because all these trials are expected to be independent, we have the following binomial distribution to give the probability Prob[k] of having k copies among $2N[t + 1]$ genes in the offspring generation:

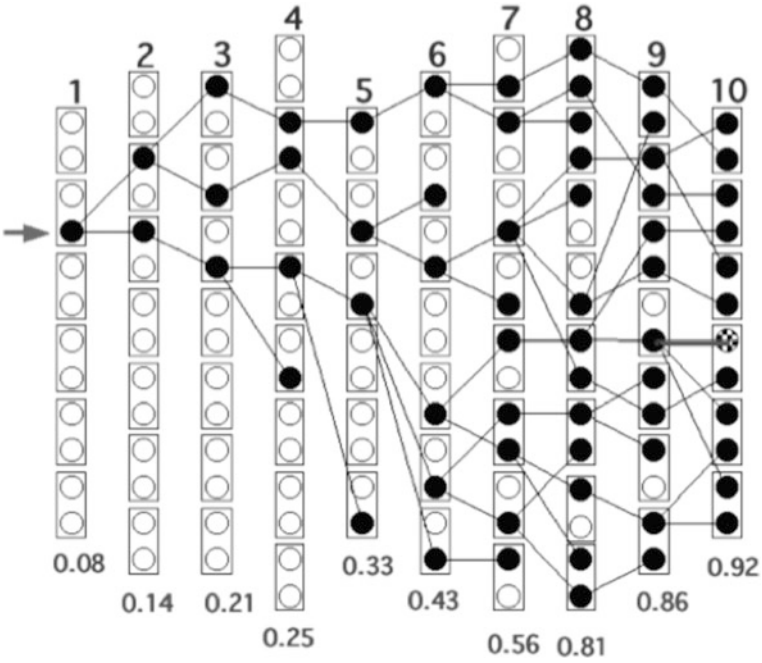

Fig. 5.3 Relationship between gene genealogy and allele frequency change (from [39])

$$\text{Prob}[k] = {}_{2N[t+1]}\text{C}_k p^k (1-p)^{2N[t+1]-k} \tag{5.1}$$

where ${}_x\text{C}_y$ ($=x!/[(x-y)!y!]$) is the possible combination to choose y out of x, and subscript i of p_i was dropped for simplicity. Continuation of this binomial distribution for many generations results in the random genetic drift of allele frequencies. When the number of individuals in that population, or population size, is quite large, this fluctuation is small because of "law of large numbers" in probability theory, yet the effect of random genetic drift will never disappear under finite population size. The random genetic drift was extensively studied by Sewall Wright and was sometimes called "Wrightian effect" in old literatures. Figure 5.4 shows examples of computer simulations of the random genetic drift under a set of very simple conditions: discrete generations, haploid, constant population size, no population structure, and no recombination. Population size (the total number of individuals or genes in one population) is 1000 in Fig. 5.4a and 10,000 in Fig. 5.4b. The initial allele frequency was set to be 0.2, and the temporal changes of up to 1000 generations are shown. In each case, five replications are shown. Clearly, as population size increases, fluctuation of allele frequencies decreases. This simplified situation is often called the Wright–Fisher model, honoring Sewall Wright and Ronald A. Fisher [9].

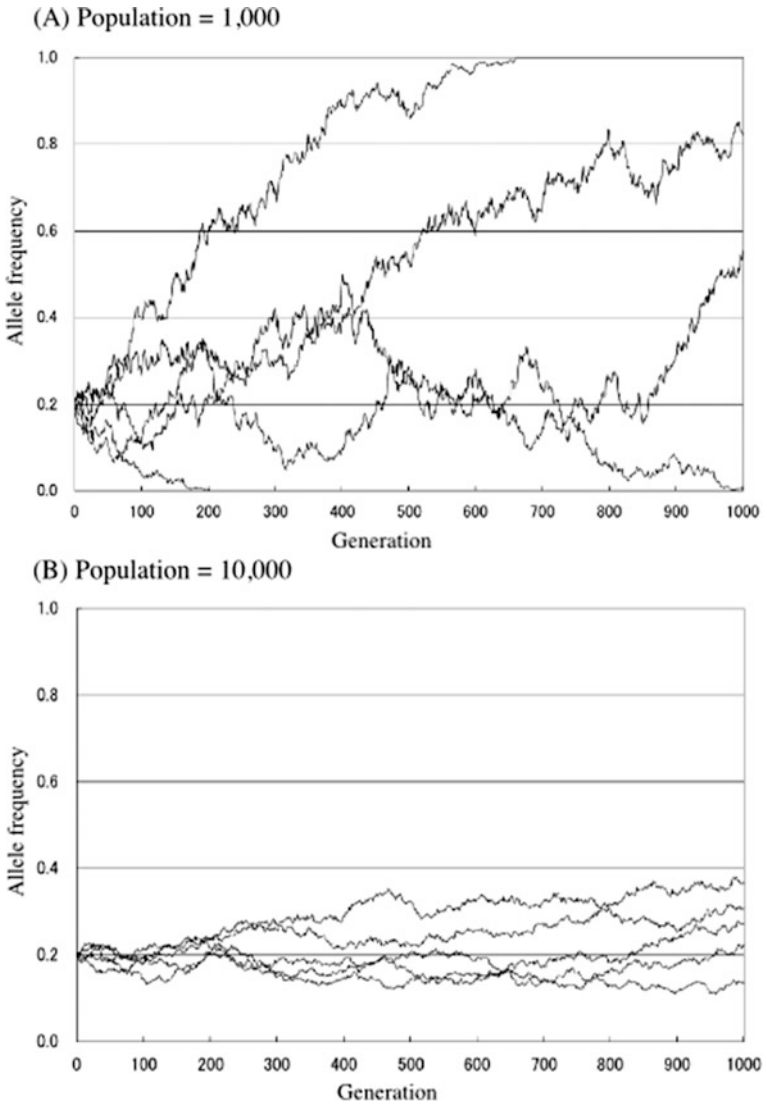

Fig. 5.4 Computer simulation of random genetic drift (from [39])

5.2.2 Branching Process

Francis Galton, a half-cousin of Charles Darwin, was interested in extinction probability of surnames. He was thus trying to compute the probability of surname extinction. He himself could not reach appropriate answer, so he asked some mathematicians. Eventually, he was satisfied with a solution given by H. W. Watson, who used generating function, and they published a joint paper in 1875 [10].

Because of this history, the mathematical model considered by them is sometimes called "Galton–Watson process," but is usually called "branching process" (see [11] for a detailed description of this process). It may be noted that surnames have been studied in human genetics (e.g., [12]) and in anthropology (e.g., [13]), for their transmissions often coincide with Y chromosome transmissions. It should be noted that Bienaymé ([14], cited in [11]) described this process mathematically much earlier than Watson and Galton [10] in French.

Fisher [15] applied this process to obtain the probability of mutants to be ultimately fixed or become extinct. Later in the 1940s, when physicists in the USA developed the atomic bomb, the branching process was used to analyze the behavior of neutron number changes (see [16]).

The distribution of transmission probability of gene copies from parents to offsprings is the basis of the branching process. The number of individuals in the population is usually not considered, for this process is mainly applied for the shallow genealogy of mutant gene copies within the large population. In a sense, the branching process is a finite small world in an infinite world.

A Poisson process is the default probability distribution for the gene copy transmission under random mating. Let us explain why the Poisson process comes in. We assume a simple reproduction process where one haploid individual can reproduce one offspring n times during its life span, and the probability, p, of reproduction is identical at each time unit (see Fig. 5.5). The probability Prob[k] of having k offspring during the n time units is given by the following binomial distribution:

$$\text{Prob}[k] = {}_n\text{C}_k p^k (1-p)^{n-k} \tag{5.2}$$

Equation 5.2 is equivalent to Eq. 5.1, though the meanings of parameters are quite different. The mean, m, of this binomial distribution is

$$m = np \tag{5.3}$$

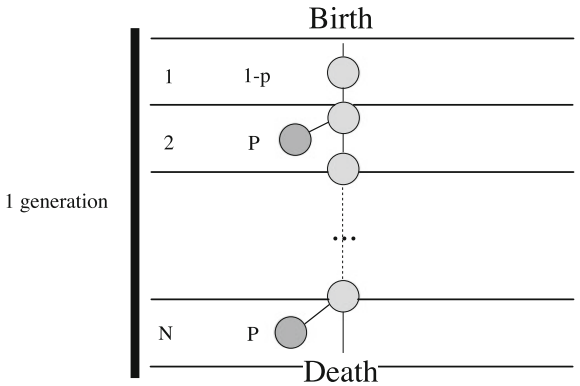

Fig. 5.5 A simple model of parent–offspring relationship

Table 5.1 Prob[k] values for various m values

Prob[k]			
k	m = 1.0	m = 1.5	m = 2.0
0	0.368	0.223	0.135
1	0.368	0.335	0.271
2	0.184	0.251	0.271
3	0.061	0.126	0.180
4	0.015	0.047	0.090
5	0.003	0.014	0.036

Let us increase n and decrease p while keeping m constant. The limit, $n = \infty$, gives

$$\text{Prob}[k] = m^k \mathrm{e}^{-m}/k! \tag{5.4}$$

where e (=2.718281828459…) is base of natural logarithm. Equation 5.4 is called Poisson distribution, after French mathematician Siméon Denis Poisson. When $m = 1$, the mutant gene is expected to keep its copy number, while $m > 1$ or $m < 1$ correspond to positive or negative selection situations (see Chap. 6). Table 5.1 shows Prob[k] values for various m values. It should be noted that Prob[0], or the probability of transmitting no offspring, is quite high. Even for $m = 2$, where the expected number of offspring is two times, Prob[0] is $\sim$0.135 even though the gene copy number explosion is expected to occur.

Fisher [15] showed that the mutant is destined to become extinct for $m \leq 1$. When $m = 1$, one may expect this is a stable situation and the mutant will continue to survive in the population. The population size is assumed to be infinite in the usual branching process, and this causes the mutant gene copy with $m = 1$ to become extinct. However, we live in finite environment, and the branching process under infinite population size is not appropriate when we consider the long-term evolution. When $m > 1$, the mutant is advantageous, and the probability of survival becomes positive, as we will see in Chap. 6. Readers interested in application of the branching process to fates of mutant genes should refer to Crow and Kimura [17].

Although the Poisson process is usually assumed in a random mating population, the real probability distribution of gene copy number may be different. In human study, pedigree data are used to estimate the gene transmission probability. A Kalahari San population (!Kung bushman) was reported to have a bimodal distribution of gene transmission, and the variance is larger than mean [18]. Interestingly, a Negrito population in the Luzon Island, who are also hunter-gatherers, had an approximate Poisson distribution with mean 1.05, as shown in Fig. 5.6 (based on Saitou et al. [19]). Figure 5.7 shows an example of the branching process with $m = 1$. A Monte Carlo method was used to generate this genealogy.

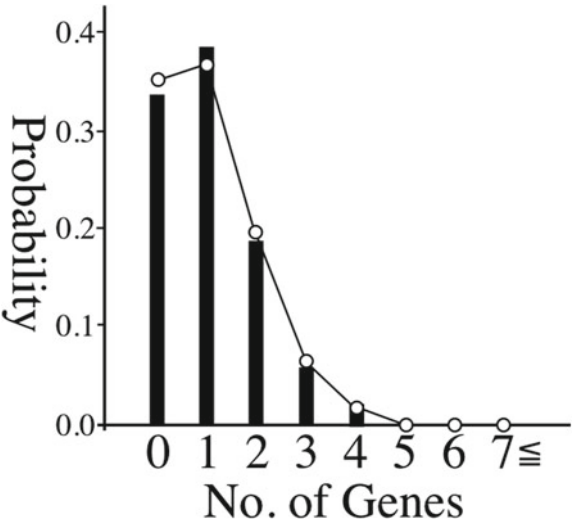

Fig. 5.6 Distribution of gene copy number transmission. Black bars represent observed numbers, and open circles represent expected numbers with a Poisson distribution of mean 1.05 (based on [19])

5.2.3 Coalescent Process

Mutant gene transmission follows the time arrow in the branching process. In another way, it is a forward process. However, as we saw, most of the gene lineages become extinct, and it is not easy to track the lineage which will eventually propagate in the population. Now let us consider a genealogy only for sampled genes. It is natural to look for their ancestral genes, finally going back to the single common ancestral gene. This is viewing a gene genealogy as a backward process. When two gene lineages are joined at their common ancestor, this event is called "coalescence" after Kingman [20]. It should be noted that Hudson [21] and Tajima [22] independently proposed essentially the same concept in 1983.

Let us consider Fig. 5.2 again. Left-most two gene copies coalesce first, followed by the coalescence of two mutant genes shown in filled circles. At this moment, there are eight lineages left, and one of them experienced mutation, shown with a star. After six more coalescent events, at around $2N$ generations ago, there are only two lineages. Then, it took another $\sim 2N$ generations to reach the final ninth coalescence. If there is no population structure in this organism, called "panmictic" situation, and if there is no change in population size (N), the time to reach the last common ancestral gene, or coalescent time, is expected to be approximately $4N$ generations ago, according to the coalescent theory of an autosomal locus for diploid organisms.

The simplest coalescent process is pure neutral evolution. Even if mutations accumulate, they do not affect survival of their offspring lineages. Because of this nature, gene genealogy and mutation accumulation can be considered separately. If natural selection, either negative (purifying) or positive, comes in for some mutant lineages, this independence between generation of gene genealogy and mutation accumulation no longer holds.

Another important assumption for the simplest coalescent process is the constant population size, N. In diploid organism, the number of gene copies for an autosomal locus is $2N$, while the number of gene copies for haploid organism locus is N. The former situation is assumed explicitly or implicitly in many literatures. However, the original lifestyle of organisms is haploid, and many organisms today are haploids. Therefore, we consider the situation in haploid organisms first. It should be noted that an approximate constant population size is more or less expected if we consider a long-term evolution. Otherwise, the species will become extinct or will have exponential growth. Though we, *Homo sapiens,* in fact experience population explosion, this is a rather rare situation among many species. In short-term evolution, population size is expected to fluctuate for any organism. Therefore, assumption of the constant population size is not realistic and is only for mathematical simplicity. We have to be careful about this sort of very simplistic assumptions inherent in many evolutionary theories. There are some more simplifications in the original coalescent theory: discrete generation and random mating. Random mating means that any gene copy is equal in terms of gene transmission to the next generation, and there is no subpopulation structure within the population of N individuals in question. These assumptions were also used for the Wright–Fisher model.

Let us first consider the coalescence of only two gene copies. What is the probability, Prob[2 → 1, 1], for two genes to coalesce to a single gene in one generation? If we pick up one of these two gene copies arbitrarily, this gene, say, G1, should have its parental gene, PG1, in the previous generation. Another gene, G2, also has its parental gene PG2. Because all genes are equal in terms of gene transmission probability under our assumption, all N genes, including PG1, can be PG2. We should remember Fig. 5.7, where multiple offsprings may be produced

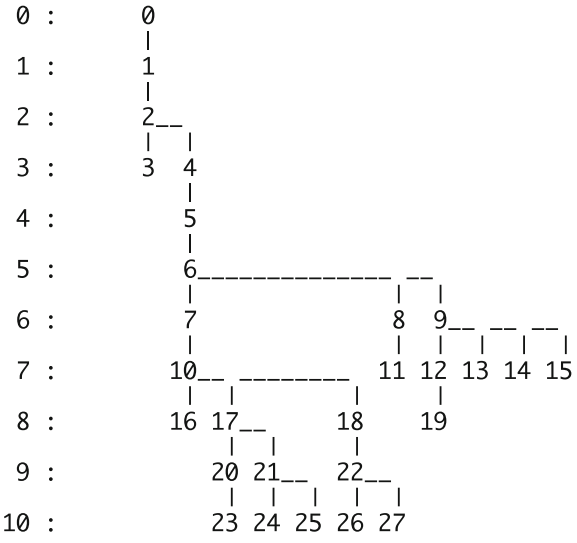

Fig. 5.7 Example of computer program output of branching process with a Poisson distribution of mean 1.0

from one individual during one generation, though this is a forward process. Therefore, having one offspring G1 does not affect the probability of having another offspring, for these reproductions are independent. It is then obvious that

$$\text{Prob}[2 \rightarrow 1, 1] = 1/N \tag{5.5}$$

The probability of the complementary event, i.e., no coalescence, can be written as Prob[2 → 2, 1] and

$$\text{Prob}[2 \rightarrow 2, 1] = 1-(1/N) \tag{5.6}$$

We now move to slightly more complicated situation. What is the probability, Prob[2 → 1, t], for two genes to coalesce exactly after t generations? The coalescent event must occur only after no coalescence through (t − 1) generations. Thus,

$$\text{Prob}[2 \rightarrow 1, t] = [1-(1/N)]^{t-1} \cdot [1/N] \tag{5.7}$$

When N is large, $[1 - (1/N)]^{t-1}$ can be approximated as exp[− t/N]. Then

$$\text{Prob}[2 \rightarrow 1, t] \sim \exp[-t/N]/N \tag{5.8}$$

We can obtain the mean, Mean[2 → 1, t], and the variance, Var[2 → 1, t], of the time, t, for coalescence, using this geometric distribution:

$$\text{Mean}[2 \rightarrow 1, t] = \Sigma_{t=1,\infty} t \cdot [1/N] \cdot [1-(1/N)]^{t-1} \tag{5.9}$$

After some transformations,

$$\text{Mean}[2 \rightarrow 1, t] = \text{N}. \tag{5.10}$$

The variance of this exponential distribution is

$$\text{Var}[2 \rightarrow 1, t] = \sum_{t=1,\infty} (t-N)^2 \cdot [1/N] \cdot [1-(1/N)]^{t-1} \tag{5.11}$$

It can be shown that

$$\text{Var}\ [2 \rightarrow 1, t] = N(N-1) \tag{5.12}$$

When $N \gg 1$, Var [2 → 1, t] ∼ N^2. Therefore, the standard deviation of t is ∼ N generations, same as its mean. When a diploid autosomal locus is assumed, mean and variance are $2N$ and $(2N)^2$, respectively. Let us now consider the coalescent process for n genes sampled from the population of N individuals. We assume $n \ll N$. The first step is the probability for two of n gene copies to coalesce during t generations. The probability of three gene copies to coalesce in one generation is $(1/N)^2$. If N is large, $(1/N)^2 \sim 0$, we can ignore the coalescence of

more than two genes in one generation and focus on the coalescence of the only pair of genes. Because there are ${}_nC_2$ [= $n(n-1)/2$] possible combinations to choose two out of n genes,

$$\text{Prob}[n \to n-1,\ 1] =_n C_2 \cdot [1/N] \tag{5.13}$$

We can thus generalize Eq. 5.7 to consider the probability that two genes among n genes sampled are coalesced in one generation as

$$\text{Prob}[n \to n-1, t] = [1-({}_nC_2/N)]^{t-1} \cdot [{}_nC_2/N] \tag{5.14}$$

The mean of t under this distribution is

$$\text{Mean}[n \to n-1, t] = N/{}_nC_2 = 2N/n(n-1) \tag{5.15}$$

We can then obtain the mean or expected time of coalescence from the current generation of n genes to single common ancestral gene by summing the means above:

$$\text{Mean}[n \to 1, t] = \sum_{i=2,n} 2N/i(i-1) \tag{5.16}$$

$$= 2N[1-(1/n)] \tag{5.17}$$

If n is large,

$$\text{Mean}[n \to 1, t] \sim 2N \tag{5.18}$$

When diploid autosomal genes are considered, this approximate mean becomes $4N$, and the variance of the coalescent time, when n is large, is given by Tajima [22]:

$$\text{Var}[n \to 1, t] \sim 16N^2\left(\pi^2/3-3\right) \tag{5.19}$$

If n is not much different from N, or almost exhaustive sampling was conducted, the possibility of coalescence of three or more gene copies together at one gene copy within one generation is no longer negligible, and Eq. 5.13 and later do not hold any more. We need to consider "exact" coalescence. The following explanation is after Fu [23]. If we consider a randomly mating population with constant size N, each gene copy at the present population was sampled from N gene copies of the previous generation with replacement. Therefore, if we choose one particular gene copy, say, copy ID 1, from the present population, the probability of its transmission from a specific gene copy of the previous generation is $1/N$. Then, the probability of gene copy ID 2 from the present population not sharing the same parental copy with copy ID 1 is $1 - [1/N]$. We then go to the next situation in which gene copy ID 3 from the present population shares the parental gene copy

with neither ID 1 nor ID 2. Its probability becomes $1 - [2/N]$. Applying a similar argument for IDs 4 to n ($n \leq N$), the probability, Prob$[n \rightarrow n, 1]$, that none of gene copies at the present generation shares the parental gene copy at the previous generation becomes

$$\text{Prob}[n \rightarrow n, 1] = \prod_{k=1,n-1} (1 - [k/N])^N \tag{5.20}$$

$$= N_{[n]}/N^n \tag{5.21}$$

$$N_{[n]} = N(N-1)(N-2)\ldots(N-n+1) \tag{5.22}$$

Therefore, the probability corresponding to Eq. 5.14 under the exact coalescent in which n gene copies at the present generation will coalesce to $n - 1$ ancestral gene copies at t generations ago becomes

$$\text{Prob}[n \rightarrow n-1, t] = \left[1-(N_{[n]}/N^n)\right] \cdot \left[N_{[n]}/N^n\right]^{t-1} \tag{5.23}$$

Generally speaking, the coalescent time for exact process is shorter than the approximation, or Kingman coalescence, first given by Kingman [20]. Figure 5.8 shows examples of gene genealogies of the same sample size under the exact coalescence and Kingman coalescence (reproduced from [23]).

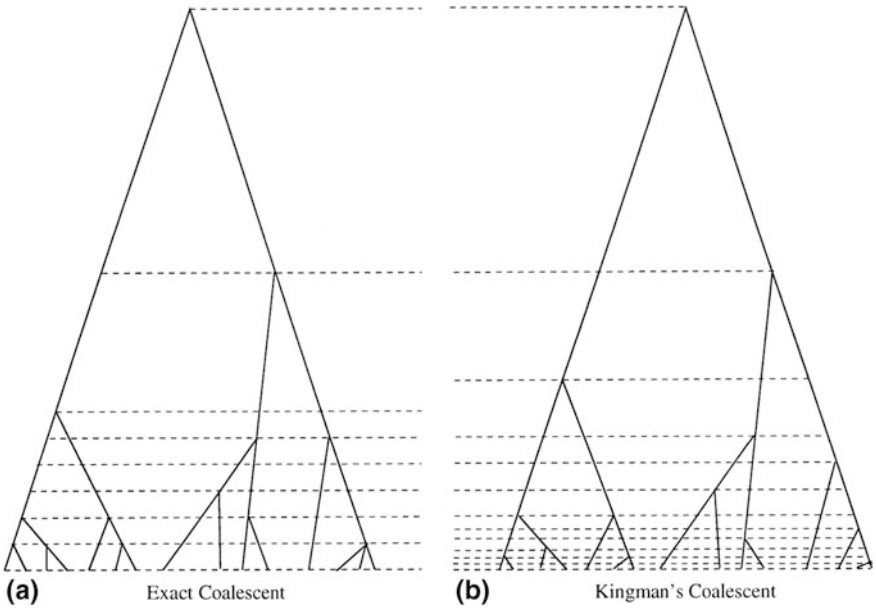

Fig. 5.8 Comparison of exact and Kingman coalescence (from [23])

Unlike the treatment of allele frequency changes to be discussed in later sections, the coalescent generation time is given in terms of the total number of genes in a population in the coalescent theory. Because of this, we can check the implicit assumption of the constant population size. For example, the total number of human population as of 2018 A.D. is over 7 billion. If we apply the coalescent theory under the constant population model, the expected number of generations for coalescence of an autosomal gene, $4N$, is 27 billion generations. If one generation is 20 years, the expected coalescent time becomes 540 billion years! This value is far greater than the start of this universe, i.e., Big Bang, approximately 14 billion years ago. This seemingly paradoxical situation simply comes from the population explosion, which violates the assumption of constant population size. To overcome this problem, the "effective population size" is often used. Modern human is estimated to have ca. 10,000 as the effective population size (e.g., [24]). There are two books on the coalescent theory [25, 26]. We will discuss various applications of the coalescent theory in Chap. 18.

5.2.4 Markov Process

We now move to the treatment of allele frequency changes. For simplicity, a constant population size (N) is assumed. We also consider haploid organism as before. Let us consider one particular allele A_i, and the number of A_i copies (hereafter called "gene copies") at generation t is denoted as i. Allele frequency for this allele at generation t is i/N. Then, the probability of having j gene copies among N genes in the next generation ($t + 1$) becomes

$$\mathrm{Prob}[i \to j; N] = {}_N\mathrm{C}_j [i/N]^j (1 - [i/N])^{N-j} \tag{5.24}$$

This is the transition probability of i to j gene copies from generation t to $t + 1$. For simplicity, let us denote $\mathrm{Prob}[i \to j]$ as $P_{i,j}$ ($0 \leq i, j \leq N$). Then, we can have the transition probability matrix $\mathbf{P}$ as:

$$\mathbf{P} = \begin{vmatrix} \mathrm{P}_{0,0} & \mathrm{P}_{1,0} & \mathrm{P}_{2,0} & \mathrm{P}_{3,0} & \mathrm{P}_{4,0} & \cdots & \mathrm{P}_{N-2,0} & \mathrm{P}_{N-1,0} & \mathrm{P}_{N,0} \\ \mathrm{P}_{0,1} & \mathrm{P}_{1,1} & \mathrm{P}_{2,1} & \mathrm{P}_{3,1} & \mathrm{P}_{4,1} & \cdots & \mathrm{P}_{N-2,1} & \mathrm{P}_{N-1,1} & \mathrm{P}_{N,1} \\ \mathrm{P}_{0,2} & \mathrm{P}_{1,2} & \mathrm{P}_{2,2} & \mathrm{P}_{3,2} & \mathrm{P}_{4,2} & \cdots & \mathrm{P}_{N-2,2} & \mathrm{P}_{N-1,2} & \mathrm{P}_{N,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \mathrm{P}_{0,N-2} & \mathrm{P}_{1,N-2} & \mathrm{P}_{2,N-2} & \mathrm{P}_{3,N-2} & \mathrm{P}_{4,N-2} & \cdots & \mathrm{P}_{N-2,N-2} & \mathrm{P}_{N-1,N-2} & \mathrm{P}_{N,N-2} \\ \mathrm{P}_{0,N-1} & \mathrm{P}_{1,N-1} & \mathrm{P}_{2,N-1} & \mathrm{P}_{3,N-1} & \mathrm{P}_{4,N-1} & \cdots & \mathrm{P}_{N-2,N-1} & \mathrm{P}_{N-1,N-1} & \mathrm{P}_{N,N-1} \\ \mathrm{P}_{0,N} & \mathrm{P}_{1,N} & \mathrm{P}_{2,N} & \mathrm{P}_{3,N} & \mathrm{P}_{4,N} & \cdots & \mathrm{P}_{N-2,N} & \mathrm{P}_{N-1,N} & \mathrm{P}_{N,N} \end{vmatrix} \tag{5.25}$$

We can derive the probability, $\mathrm{Prob}[i/N, t + 1]$, of having allele frequency i/N at generation $t + 1$, using this transition probability matrix and the probability at generation t as follows.

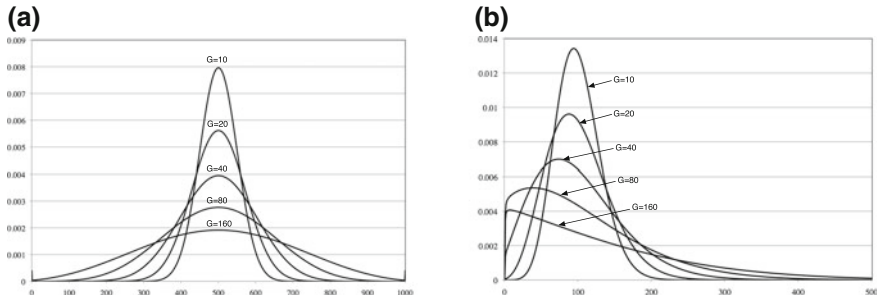

Fig. 5.9 Example of Markov process. **a** $N = 1000$, initial frequency = 0.5. **b** $N = 1000$, initial frequency = 0.1. G is generation

$$\text{Prob}[i/N, t+1] = \sum_{j=0,N} \text{Prob}[i/N, t] \cdot \text{P}_{j,i} \tag{5.26}$$

At the initial generation ($t = 0$), let us assume that there are k ($1 \leq k \leq N-1$) gene copies in the population. Then, Prob[k/N,0] = 1 and Prob[i/N,0] = 0 ($0 \leq i \leq N, i \neq k$). A stochastic process whose probability distribution is given this way is called a first-order Markov process.

Figure 5.9 shows some examples of allele frequency spectra using the Markov process for various combinations of N, k, and t. The Perl script for computing the Markov process is available upon request to the author. In the past, the Markov process was not extensively used, for it requires a large number of computations. Thanks to the great advancement of computational powers, we can now obtain allele frequency spectrum for relatively large number of populations. It may be interesting to apply this exact Markov process for various realistic situations in the future.

5.2.5 Diffusion Process

There are various mathematical models which can describe the evolutionary changes of allele frequency. The diffusion equation is the most widely used method. It can easily combine the stochastic effect, namely, random genetic drift, and deterministic effect such as natural selection random genetic drift alone is discussed in this section, and natural selection will be discussed in Chap. 6.

The starting point is the binomial distribution, the basic process for the random genetic drift (see Sect. 5.2.1). We assume that the population size is constant, and haploid organism is considered. The binomial distribution in Eq. 5.1 can be written as

$$\text{Prob}[k] = {}_N\text{C}_k p^k (1-p)^{N-k}. \tag{5.27}$$

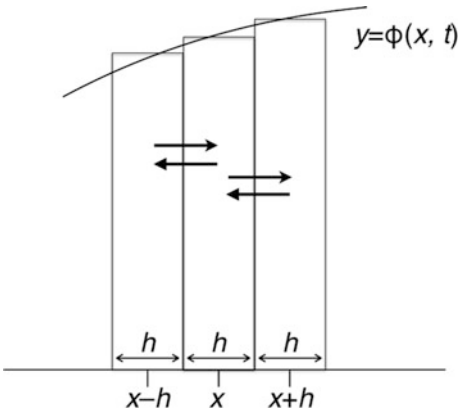

Fig. 5.10 Explanation of diffusion model (based on [27])

Let us note that the mean (m) and variance (v) of the gene copy numbers of this distribution are

$$m = Np, \tag{5.28}$$

$$v = Np(1-p). \tag{5.29}$$

This process can be approximated by a differential equation, a Kolmogorov forward (Fokker-Planck) equation for the random genetic drift:

$$\partial\varphi(p \to x;t)/\partial t = [1/4N][\partial^2/\delta x^2\{x(1-x)\phi(p \to x;t)\}]. \tag{5.30}$$

Figure 5.10 explains the basic concept of Eq. 5.30 on the change of allele frequency class, based on Kimura [27]. Let us consider a very small range of length h, and histograms of many rectangles approximate the probability density function $\partial(p \to x;t)$. Each rectangle has the width h and the height given by the value of $\partial(p \to x;t)$ at allele frequency x, at the middle of the rectangle unit. We also consider a very short time ∂t, so the change of allele frequency during that time period is restricted to at most to adjacent range, either left or right. If we take limits ($h \to$ zero and $\partial t \to$ zero), differential Eq. 5.30 is obtained.

The exact solution for this equation, for probability density distribution of allele frequency x at generation time t, starting from initial frequency p, is

$$\begin{aligned}\partial(p \to x;t) = &\sum_{i=1,\infty} p(1-p)i(i+1)(2i+1)\mathrm{F}(1-i, i+2, 2;p) \\ &\times \mathrm{F}(1-i, i+2, 2;x)\exp[-i(i+1)t/4N]\end{aligned} \tag{5.31}$$

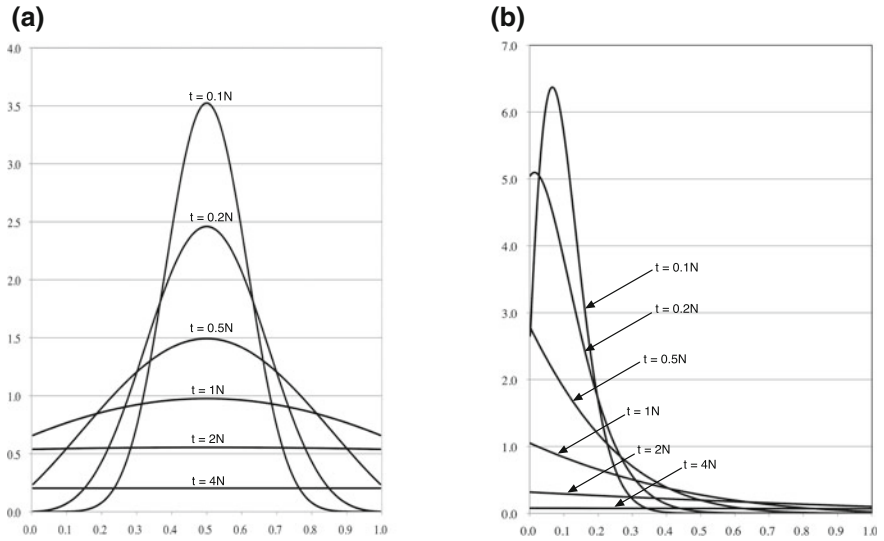

Fig. 5.11 Diffusion process. **a** Initial frequency = 0.5. **b** Initial frequency = 0.1

$F(a,b,c;z)$ in Eq. 5.31 is a hypergeometric function:

$$F(a,b,c;z) = \sum_{n=0,\infty} \{a_{[n]} \cdot b_{[n]} \cdot z^n\} / \{c_{[n]} \cdot n!\} \tag{5.32}$$

This solution was given by Kimura in 1955 [28]. Interested readers should refer to [17] and [27] for a detailed explanation of the diffusion process.

Figure 5.11 shows the probability density changes for various generations when the initial allele frequency is 0.5 and 0.1. The Perl script for computing the diffusion process is available upon request to the author. Initially, at time zero, all probability density is concentrated at the initial allele frequency. This is Dirac's delta function. As the random genetic drift starts to operate, allele frequency will start to diffuse. After long time, probability density becomes flat and low, and the majority of probabilities will be residing at either allele frequency of 0 or 1.

5.2.6 A More Realistic Process of Allele Frequency Change of Selectively Neutral Situation

In reality, the population size is not only finite but also not constant. Therefore, a more realistic process of frequency change of selectively neutral mutant alleles is as follows. Let us denote the total gene copy number of the population at generation t as $N[t]$ and the gene copy number of a selectively neutral allele A at generation t as $N_A[t]$. Then, the allele frequency at generation t, Freq_$A[t]$, becomes

$$\text{Freq\_}A[t] = N_A[t]/\text{N}[t] \tag{5.33}$$

We need to consider a finite population bounded by a finite maximum population size, or carrying capacity. Then, the population size fluctuation can be approximated by a Markov process with constant global population size or carrying capacity. The problem is that the carrying capacity itself will change depending on the change of environment. In the case of humans, the environment includes technological innovation. We need to redefine the transition probability matrix $P_{i,j}$, in which the population changes its size (number of individuals) from i to j. Because an extinct population cannot produce new population, $P_{0,j}$ $(0 < j \leq N\_\text{max})$ is zero. In contrast, $P_{N\_\text{max},j}$ $(0 \leq j \leq N\_\text{max})$ is not zero. Unfortunately, population genetics theory so far seems to be not considering this kind of more realistic dynamics of populations. It is left for future developments.

5.3 Expected Evolutionary Pattens Under Neutrality

We will discuss three categories when the pure neutral evolution is occurring: fixation probability, the evolutionary rate, and the amount of DNA variation. Because the majority of eukaryotic genome is evolving in this fashion, the understanding of the pure neutral evolutionary process is quite important for evolutionary genomics.

5.3.1 Fixation Probability

As stated at the beginning of this chapter, neutral evolution is characterized by the egalitarian nature of the propagation of mutants. Therefore, all genes at one generation have the same potential to leave offsprings. If one population is destined to continue for long time, eventually fixation of one gene will occur. Because any of N genes in the initial generation can become the common ancestor of later generations, the fixation probability of one gene in a population of N genes is $1/N$. In an autosomal locus of diploid organisms, the fixation probability becomes $1/2N$.

In reality, we do not know if one population in question at this time will continue to survive in later time. Therefore, the absolute fixation probability, Prob_fixation, of one gene should be

$$\text{Prob\_fixation} = \text{Prob\_existence} \cdot [1/N] \tag{5.34}$$

Prob_existence is the probability of the existence of that population for a certain long time. Unfortunately, we do not know this probability, and almost always it is implicitly assumed to be unity. Thus,

$$\text{Prob\_fixation} = 1/N \tag{5.35}$$

It should be noted that one mutant gene may never fix if the nucleotide sequence length of this gene is long and N is large. In this situation, new mutation(s) appear with a high frequency, and the original mutant gene is never fixed.

5.3.2 Rate of Evolution

If a gene fixation occurs in one population, there will be no change of allele frequency, though the gene genealogy will grow as time goes on. We definitely need mutations for evolution to proceed. If a mutation happens, the population of N genes with only one allele will again become polymorphic with a single copy mutant allele and $N - 1$ copies of the original allele. If all genes in later generations will become descendants of this mutant gene, now gene substitution is attained. Evolution of one gene or one locus can be seen as the accumulation of mutations. Therefore, the rate of gene substitution is equated as the rate of evolution.

Let us define the mutation rate per gene locus per generation as μ. Considering all N genes in this population, $N\mu$ mutants appear in every generation. During T generations, the total number of arising mutant genes becomes $N\mu T$, under the assumption of the constant population size. Because the fixation probability for each mutant gene is $1/N$, the total number of mutant genes fixed during T generation is $N\mu T\,[1/N] = \mu T$. The rate, λ, or speed of evolution in terms of continuous mutant fixation is thus

$$\lambda = \mu T/T = \mu \tag{5.36}$$

Equality of the evolutionary rate and the mutation rate was first shown by Kimura and Ohta [29] using the fixation probability. That explanation assumed a constant population size for a long time and may not be appropriate for a long-term evolution. There is also a possibility of nonfixation as we discussed in 5.3.1. We can relax this assumption in Eq. 5.36. Figure 5.12 shows a schematic gene genealogy for a single lineage during time T. The vertical axis represents the whole population, and the population size N can vary. Star symbols are mutations accumulated in this single lineage, and thin dotted lines represent increase of allele frequency for each mutant. The total number of mutations accumulated during time T is μT. Fixation of each mutant is not necessary. Therefore, the evolutionary rate, λ, of gene substitutions per generation should be $\mu T/T = \mu$, as shown in Eq. 5.36. This argument applies to any time irrespective of population size change. Even if speciation occurs, it does not affect this argument based on the single gene lineage. This is why we can consider the long-term evolution. Of course, any gene at the present population can be the starting point for the single lineage genealogy. This generality comes from the egalitarian nature of the selectively neutral mutant gene copies.

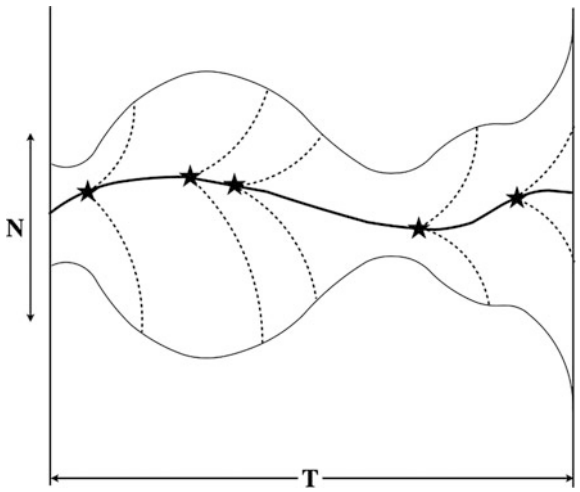

Fig. 5.12 Single lineage gene genealogy. N is population size, T is evolutionary time, and asterisks are mutations occurred on this gene genealogy. Solid and dotted lines denote population size changes and allele frequency changes, respectively

If the mutation rate, μ, does not change for long time and for diverse group of organisms, we can estimate the mutation rate by estimating the evolutionary rate in the neutrally evolving genomic region. This is the basis of the indirect method for estimating mutation rates discussed in Chap. 3.

5.3.3 Amount of DNA Variation Kept in Population

If we consider a relatively long DNA fragment, say composed of n nucleotides, as "locus," there are 4^n possible alleles in this locus. If we consider a 1-kb-long DNA fragment, $n = 1000$, and 4^{1000} is more than 10^{600}. Considering this enormous possibility of alleles for even a short DNA fragment, Kimura and Crow [30] proposed the infinite allele model. All new mutations are different with each other in this model. The phylogenetic relationship of alleles is not considered in the infinite allele model. Kimura [31] thus proposed the infinite site model where an infinitely long DNA sequence is considered. Now, new mutations appear by substituting one-nucleotide site, which was not changed before. In this sense, this model is similar to the infinite allele model, but now accumulation of nucleotide substitutions can be considered with the infinite site model. This means that a genealogical relationship of alleles is behind this model. In either case, the expected heterozygosity, H, under these two models is

$$H = 4N_e\mu/(1 + 4N_e\mu) \tag{5.37}$$

where N_e is "effective population size," and μ is mutation rate per locus per generation. The numerator of Eq. 5.37, $4N_e\mu$, which is often denoted as M or θ, should be identical with the nucleotide diversity, π, per nucleotide site [32].

5.4 DNA Polymorphism

When we compare gene copies of one locus in one organism, nucleotide sequences may be slightly different because various types of mutation may accumulate. In this case, this locus has genetic or DNA polymorphism. We have classified DNA polymorphisms according to the type of mutation (see Table 3.1 of Chap. 3). In classic evolutionary studies, "polymorphism" applies only to one species; however, the definition of species is often ambiguous, and there is no clear difference between within species genetic polymorphism and between species genetic differences. Therefore, when multiple closely related species are compared, nucleotide sites which have variations are sometimes called polymorphic.

Traditionally, one locus may be called polymorphic if the major allele frequency is equally or less than 0.99, while it is called monomorphic if the major allele frequency is more than 0.99. However, nowadays we often have sample size of more than 1000, and if some nucleotide sequences were found to be different from the major allele, this locus may be called polymorphic, even if the frequency of the major allele is more than 0.99.

Although there are no essential differences between haploid and diploid genomes in terms of the random genetic drift, patterns of genetic composition of alleles per locus, called "genotypes," are different. If there are two alleles, A_1 and A_2, at a locus, the possible genotypes are the same as alleles for haploids. In diploids, there are three genotypes, or possible combination of alleles: A_1A_1, A_1A_2, and A_2A_2. Genotypes with single type of allele are called "homozygotes," and those with two types of alleles are called "heterozygotes," after Greek words ὁμο and ἑτεροϛ meaning same and different, respectively. In general, if there are N types of alleles in one population, the possible types of homozygotes and heterozygotes are N and $N(N-1)/2$, respectively.

5.4.1 Single-Nucleotide Polymorphism (SNP)

DNA polymorphism observed at one nucleotide, the smallest unit of DNA molecule, is called "single-nucleotide polymorphism", or SNP. The majority of SNP is created via nucleotide substitution-type mutation, but sometimes one-nucleotide length insertion or deletion is also included as SNP. An SNP locus is usually biallelic. In nucleotide substitution-type SNPs, there are only two nucleotides in the population, for the mutation rate of nucleotide substitution is quite low. However, if we sample many individuals, such as for medical studies of humans, we may encounter SNP loci with three or four nucleotide alleles. There are gap or no-gap alleles for single-nucleotide indel SNPs.

SNPs observed in protein coding regions may be called cSNPs, and SNPs found in noncoding genomic region may be called gSNPs. There are synonymous and nonsynonymous cSNPs. See Chap. 12 for the SNP patterns in the human genome.

If we can estimate the ancestral SNP alleles, we can distinguish typical two alleles into ancestral and derived (mutated) alleles. If one allele has allele frequency lower than 0.5, it is called "minor" allele. Many databases for SNP are available, such as dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/).

5.4.2 Insertions and Deletions (indel)

Insertion and deletion (often abbreviated as indel)-type mutations create indel DNA polymorphisms. Broadly speaking, repeat number polymorphism and copy number polymorphism to be discussed later are also in this type; however, nonrepeat type indels are usually called as indel polymorphism. When the gap length is one, this indel polymorphism may be included in SNP, as mentioned above.

Insertions and deletions are detected as gaps in multiple alignments (see Chap. 15). Therefore, if nucleotide sequences are misaligned, we have incorrect indel information. Nucleotide sequences within the same species are expected to have quite high homology; however, if we are not aware of microinversions, misalignment will occur and often gaps are observed.

5.4.3 Repeat Number Polymorphism

When insertions or deletions occur within the repeat sequences, they are called repeat polymorphisms. Short repeat sequences of 1–5 nucleotides as unit are called "short tandem repeat" (STR) polymorphism or microsatellite DNA polymorphism. When the repeat length is longer, it is called "variable number of tandem repeat" (VNTR) polymorphism or minisatellite DNA polymorphism.

5.4.4 Copy Number Variation

If the repeat unit is much bigger, say at least a few kilobases, it is called "copy number variation" (CNV). A classic example is the Rh blood group D+/D- polymorphism. Many genes in the human genome were found to have CNV-type polymorphism [33, 34]. If CNV haplotype of more copy number is fixed in the population, the original gene is duplicated. Therefore, the frequent occurrence of CNVs suggests high frequency of gene duplications.

5.5 Mutation is the Major Player of Evolution

Mutations can be classified into deleterious, neutral, and advantageous ones according to their effects on organisms (see Chap. 6). Because the majority of the vertebrate genome is noncoding, mutations occurring in this region are selectively neutral unless they occur in evolutionarily conserved regions. If mutations occur in DNA regions where important genetic information such as protein amino acids and RNA sequences is coded, or in highly conserved noncoding regions, these mutations may become deleterious, and the mutant individual may have less possibility of transmitting that gene to the offsprings. In contrast, although in rare occasions, some DNA changes will cause that mutant individual to have more offsprings than those without mutant genes. This type of mutants is called "advantageous." In any case, when mutations occur, selectively neutral mutants are dominating. If we consider a long-term evolution, only a small fraction of these mutations will survive. Because deleterious mutations will soon disappear from the population (see Chap. 6), only neutral and advantageous mutants will survive in the population for a long time.

Because the fixation probability for advantageous mutants is higher than that for selectively neutral mutants, the proportion of advantageous mutations among the surviving mutations may be slightly higher than their proportion when they were produced. However, the majority of mutations surviving for long evolutionary time are selectively neutral. This is a clear difference from the prediction made by researchers who advocated the dominant power of natural selection in the 1960s and 1970s. As we will see, the fixation of selectively neutral mutations via stochastic effects is the main power of evolution, and the natural selection to choose advantageous mutations has only a limited contribution, although natural selection to eliminate deleterious mutations is quite effective to keep the current genetic entity. In short, natural selection is mostly conservative, and the chance effects, including the fixation of selectively neutral mutations, are really responsible for creative nature of evolution.

5.6 Evolutionary Rate Under the Neutral Evolution

We considered the fate of selectively neutral mutants in Sect. 5.3. In reality, there are deleterious and advantageous mutations. Because the fraction of advantageous mutations is expected to be much smaller than that of deleterious ones, we consider only neutral and deleterious mutations. Let us denote the fraction of neutral mutations as f. This fraction has the rate of evolution identical with the mutation rate μ. The remaining fraction, $1 - f$, is deleterious mutations, and all of them are assumed to be not fixed and do not contribute to gene substitution. Thus, the evolutionary rate λ becomes

$$\lambda = f \cdot \mu + (1-f) \cdot 0 = f\mu \tag{5.38}$$

The value of f varies from the genomic region to region, as we will see in this section. This simple relationship is under the assumption of Kimura [1], while Ohta [35] pointed out the importance of slightly deleterious mutations. In this case, the population size is involved in the mean evolutionary rate. We will discuss this problem in Chap. 6.

5.6.1 Molecular Clock

Human Rh blood group gene has paralogous genes as we saw in Chap. 4. Figure 5.13 shows the partial multiple alignment (see Chap. 15 for the procedure) of amino acid sequences for ten vertebrate RHCG proteins, products of a paralog of the Rh blood group gene. Amino acid sequence names are composed of UniProt accession number and genus. Only human amino acid sequence (Uniprot accession number = Q9UBD6) is fully written at the top, and the amino acids of remaining sequences are shown only when they are different from the corresponding human amino acid. If amino acid of nonhuman species RHCG protein is identical to the human counterpart, dot (.) is given. For example, rainbow trout RHCG amino acid sequence (Uniprot accession number = Q4VUZ1) has 56 amino acids different from those of human, among 200 amino acids. This proportion, p (0.28 = 56/200), can be used to estimate the number, d, of amino acid substitutions per amino acid site:

$$d = -\log_e(1-p) \tag{5.39}$$

This number is often called evolutionary distance, and d stands for "distance." See Chap. 16 for derivation of this equation. In any case, using this equation, d becomes 0.328 from p (=0.28). Evolutionary distances between human and the other ten vertebrate species are plotted in vertical axis of Fig. 5.14. The horizontal axis of this figure represents divergence time between human and corresponding species. Interestingly, evolutionary distances and divergence times are more or less proportional. This rough constancy of the evolutionary rate is often called

```
Q9UBD6 HUMAN VILIFQTAVNFILNLLKVKAMTTGAFGVTRIRNEQKERQNVQDLFMYISYHSAICSACSVAIALHKKALAVAMLMYAIVIILFVYLTFSRLHIING
Q5NVA3 PONAB ..........................................................................P........V...................
Q2@CR3 MACMU ........................Y..E............SA.....................V.........T......I.................
Q95JD3 RABIT .VML..ASA..L.HV.E...I.I......W.H.DHR...S.HN..I.M.NYA.....S...MV......G.......VIAV..........R..H.
Q9QXP0 MOUSE .M.....T.......IEA...........W.K.D..Q..S.H...I.S.F.A.LL..ST.TVIV.....G....T..VF.F..A......HR..H.
Q19KI0 PIG   .V..L..S........E............W.P.Y.....S.H.....V.H.A.........L..............SVI............R..H.
Q3BCQ4 CANLF .V..L..S........E............W.PGH.....S.H.....V.N.A.........L..............SVI.V..........RV.H.
Q2T9S6 BOVIN .V..L..SI.Y.....E..S..A.....AW.P.HL....STH.......N.A.........L...R...GLVLVLFS...V............VH.
Q6XL41 CHICK TV..L.....Y.....H.............FP....DKEG.H..Y....E.A......TM.FM.Q......S..T.SVI.VVY....I.K....H.
Q4VUZ1 ONCMY .V.VLG...EY.I...HAR..V..GY.ISWVP.H..RM.G.H.I..F.TD.GV.LCSTT...FSQ.TS..L.FIS..VL..MYIFIS.KT.K..HA
```

Fig. 5.13 Multiple alignment of ten vertebrate RHCG amino acid sequences. First six-character codes are UNIPROT accession numbers of this sequence, followed by species name and one-letter amino acid sequences. When the amino acid is identical with that of human at the top, it is designated as period (.)

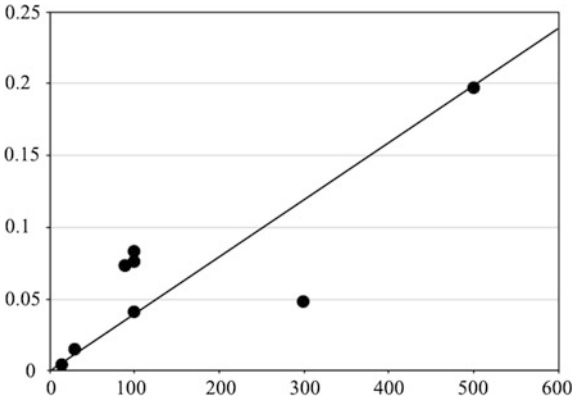

Fig. 5.14 Approximate linearity or molecular clock for vertebrate RHCG (based on data of Fig. 5.13)

"molecular clock" after Zuckerkandl and Poring [36]. It should be noted that evolutionary distances were obtained from molecular data determined in wet laboratories, while divergence times were obtained from paleontological studies.

Existence of the molecular clock is easily explained under the neutral theory. If the mutation rate (μ) and the fraction (f) of deleterious mutations are constant for long evolutionary time, the evolutionary rate λ ($=f\mu$) should be constant according to Eq. 5.38. In contrast, if the evolutionary rate is mainly determined by positive selection, not only mutation rate but also population size and selection coefficients of mutants affect the evolutionary rate, and the latter two are known to vary considerably. Therefore, the approximate constancy of the evolutionary rate is one evidence supporting the neutral theory of molecular evolution.

Even if we do not assume the constancy of the evolutionary rate, it is possible to consider the average rate of evolution by comparing two sequences. Figure 5.15 shows a schematic phylogenetic tree of two sequences, 1 and 2. They have the common ancestor T years ago, and the lineage specific evolutionary rates, $\lambda 1$ and $\lambda 2$, are given. Thus, the average rate, λ, of evolution between sequences 1 and 2 becomes

$$\lambda = (\lambda 1 + \lambda 2)/2 \tag{5.40}$$

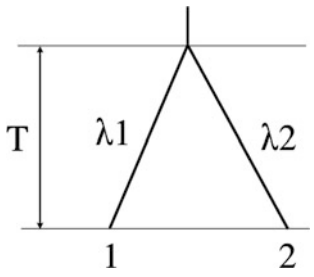

Fig. 5.15 Divergence of two lineages

Let us denote the evolutionary distance between sequences 1 and 2 as d. Then,

$$d = \lambda \cdot 2T \tag{5.41}$$

We can thus estimate the evolutionary rate:

$$\lambda = d/(2T) \tag{5.42}$$

If the constancy of the evolutionary rate approximately holds, we can estimate the divergence time:

$$T = d/(2\lambda) \tag{5.43}$$

This equation is often used because the divergence time of two sequences is usually unknown, while the molecular data such as amino acid sequences or DNA sequences can be easily determined.

5.6.2 Heterogeneous Evolutionary Rates Among Proteins

The fraction, f, of neutral mutations in Eq. 5.38 may vary in various situations. Let us first consider the heterogeneity among different proteins. Table 5.2 lists the rates of amino acid substitutions per amino acid site per year for 30 proteins taken from Table 1 of Dayhoff [37]. The evolutionary rates considerably vary from 0.00 to 3.3×10^{-9}/amino acid site/year. The highest amino acid substitution rate so far estimated is that (4.3×10^{-9}) for fibrinopeptide [38]. In contrast, histone proteins are the major basic protein family of nucleosome that binds DNA, an acid. The very low evolutionary rate for this protein family indicates that f, the fraction of neutral mutations, is quite low, and the majority of amino acid changing mutations are deleterious.

Fibrinopeptide is leftover of fibrinogen which was cut to fibrin and fibrinopeptide. The main function of blood coagulation is residing in fibrin, and the function of fibrinopeptide is just to keep fibrin not to become fibrous until it is detached from fibrin part. It is thus understandable that many amino acid substitutions on fibrinopeptide gene may be permissible; hence, its f became high.

Because of this relationship between f values and protein functions, it is routine to discuss the importance of one function in terms of its rate of amino acid substitutions. If the rate is slow, the protein may be called "quite important," and it may be "less important" if the rate is relatively high.

Table 5.2 Rates of amino acid substitutions (based on [37])

Protein	Rate ($\times 10^{-9}$)
Immunoglobulin kappa chain C region	3.7
Kappa casein	3.3
Immunoglobulin gamma chain C region	3.1
Complement C3a anaphylatoxin	2.7
Lactalbumin	2.7
Epidermal growth factor	2.6
Somatotropin	2.5
Pancreatic ribonuclease	2.1
Haptoglobin alpha chain	2.0
Serum albumin	1.9
Prolactin	1.7
Carbonic anhydrase	1.6
Globin alpha chain	1.2
Globin beta chain	1.2
Myoglobin	0.89
Trypsin	0.59
Endorphin	0.48
Insulin	0.44
Lactate dehydrogenase	0.34
Cytochrome c	0.22
Ferredoxin	0.19
Collagen	0.17
Alpha-crystallin B chain	0.15
Glucagon	0.12
Glutamate dehydrogenase	0.090
Histone H2B	0.090
Histone H2A	0.050
Histone H3	0.014
Histone H4	0.010
Ubiquitin	0.000

Unit per amino acid per year

5.6.3 Heterogeneous Evolutionary Rates Among Protein Parts

One protein has its specific 3D structure (see Chap. 2), and the functional part is often localized as "domains." Domains are often defined for many proteins because of their wide conservations (see Chap. 2). Therefore, it is natural for a domain part to have lower evolutionary rate than the remaining part of the protein. For example, Hox genes have highly conserved homeobox domain. If we compare amino acid sequences of human and mouse orthologous HoxA1–HoxA5 amino acid sequences, amino acid identities are certainly higher for the homeodomain region. Table 5.3

Table 5.3 Comparison of amino acid identity between homeodomain and the other regions of HoxA (from [39])

Amino acid similarity between human and mouse (%)		
Gene	Homeodomain region	Other regions
HoxA1	99.1	96.2
HoxA2	100	97.6
HoxA3	100	96.3
HoxA4	100	83.0
HoxA5	100	98.6

(taken from [39]) shows the proportions of amino acid similarity for this protein in two parts. As expected, the amino acid similarities of homeobox domains are quite high compared to those of the remaining parts.

5.6.4 Heterogeneous Evolutionary Rates Among Organisms

The evolutionary rate is proportional to f and μ. Therefore, if μ, the mutation rate differs among various lineages, molecular clock no longer holds. This is the case for the rodent lineage and other mammalian lineages, as first clearly shown by Wu and Li ([40, 41]). Hominoid and Old World monkeys diverged at $\sim$30 million years ago. Because human and rhesus macaque genomic distance is $\sim$0.06 [42], the average evolutionary rate in terms of nucleotide substitutions is, from Eq. 5.42, λ[primates] = 0.06/[2 $\times$ 30 million] = 1 $\times$ 10^{-9}/site/year. The genomic distance between mouse and rat in terms of fourfold degenerate synonymous sites (see Sect. 5.7.1) is $\sim$0.15 [43]. The divergence time between mouse and rat is not well known, so we use a range of 10–20 million years. Then, λ[rodents] = 0.15/[2 $\times$ {10–20} million] = 4–8 $\times$ 10^{-9}/site/year. Because mammalian genomes are mostly consisting of junk DNAs (see 5.7.2), genome-wide evolutionary rates are approximately mutation rate. It is clear that the mutation rate of rodents is several times higher than that for primates.

Compared to ordinary DNA genome organisms, genomes of RNA viruses such as influenza virus, SARS, and HIVs are RNA molecules, and their evolutionary rates are million times higher than those of DNA genome organisms (see Chap. 10).

If the value of f, fraction of neutral mutations, varies among lineages for a particular protein gene, the evolutionary rate obviously changes. In this case, molecular clock no longer holds, yet this variation naturally follows the pattern of neutral evolution. Although the molecular clock is often considered as the important characteristics of the neutral evolution, this comes from the simple relationship shown in Eq. 5.37. Therefore, if f and/or μ changes, the evolutionary rate should change, according to the neutral evolution. Figure 5.16 is the evolutionary history of rodent α crystallin [44]. The amino acid sequence of this protein is identical among mouse, rat, and hamster, and their sequence is identical with that of common ancestor or all rodents. In marked contrast to that situation, nine amino acid substitutions accumulated in the mole rat lineage during 40 million years. Mole rat eye

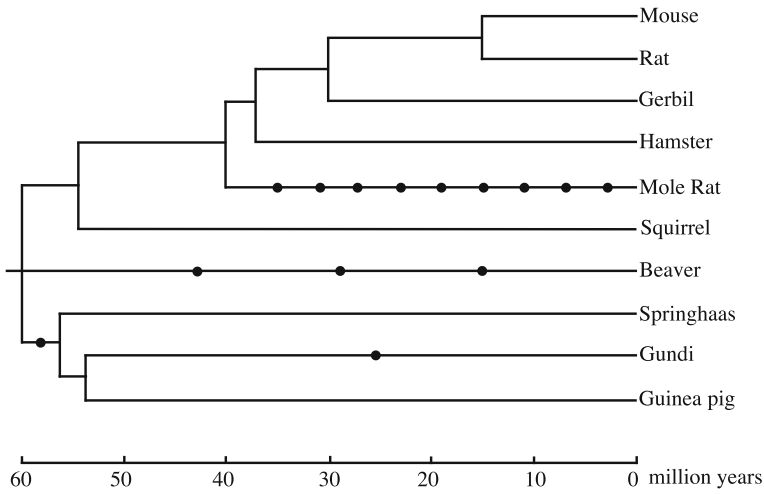

Fig. 5.16 Evolutionary history of rodent α crystallin (based on [44])

is diminished, and apparently, importance of α crystallin, the major lens protein, is reduced. It is natural to expect higher fraction (f) of selectively neutral mutations for mole rat than other rodents whose eyes are necessary for their existence.

5.6.5 Unit of Evolutionary Rate

We discussed the unit of mutation rate in Chap. 3. Because mutation is the main player of evolution, unit of the evolutionary rate is closely related to that discussion. While the generation time for many organisms is not known, divergence times of some organism groups such as vertebrates have been well documented thanks to paleontological studies. Thus, the rate of evolution is often obtained by Eq. 5.42, and the time unit is years, not generations.

5.7 Various Features of Neutral Evolution

We discuss the features of neutral evolution in terms of preponderance of synonymous substitutions to nonsynonymous ones, pure neutral evolution of junk DNA and pseudogenes, and neutral evolution at the macroscopic levels and at genomic levels.

5.7.1 Synonymous and Nonsynonymous Substitutions

If synonymous or nonsynonymous mutations (see Chap. 3) are fixed in populations, these are called synonymous and nonsynonymous substitutions, respectively. In some literatures, synonymous substitutions are called silent substitutions and nonsynonymous ones are amino acid-replacing substitutions.

If we consider the consequences of synonymous mutations, it is easy to expect that they are selectively neutral with original alleles because produced proteins are identical with each other. Nonsynonymous mutations may become deleterious because they may disrupt or reduce the protein function. As we saw in the evolution of fibrinopeptides, it is also possible that the effect of a nonsynonymous substitution may be very minor and essentially selectively neutral. It is therefore a good approximation that f (the fraction of neutral mutations) for synonymous mutations is 1, and the evolutionary rate is identical with mutation rate. As for nonsynonymous mutations, f is smaller than 1, and the evolutionary rate of nonsynonymous substitutions is expected to be smaller than that for synonymous substitutions. As we will see in Chap. 6, the evolutionary rate of nonsynonymous substitutions may become larger than the mutation rate when a special type of natural selection is operating, when any amino acid change is advantageous. In this case, the rate of nonsynonymous substitutions will be higher than that of synonymous substitutions. Figure 5.17 shows a schematic comparison of the rates of synonymous and nonsynonymous substitutions.

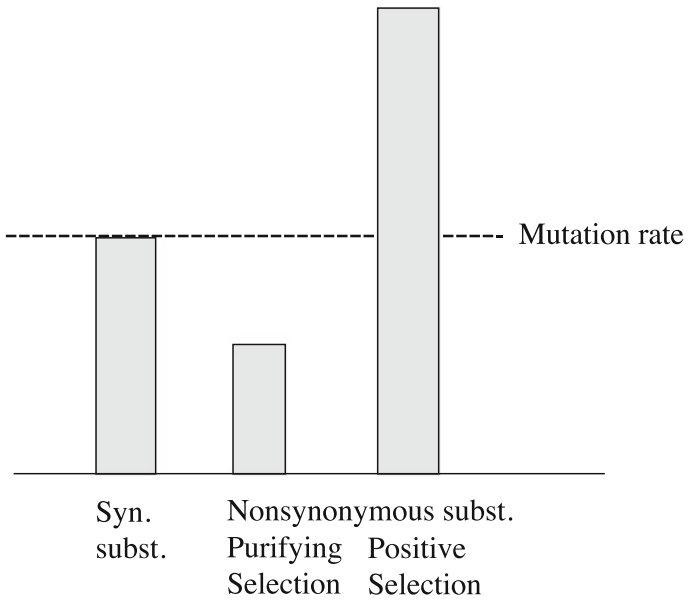

Fig. 5.17 A schematic comparison of synonymous and nonsynonymous substitutions. The evolutionary rate of synonymous substitutions is expected to be identical with the mutation rate, while that of nonsynonymous substitutions are lower when purifying selection operates. Only when positive selection operates, the evolutionary rate of nonsynonymous substitutions becomes higher than the mutation rate

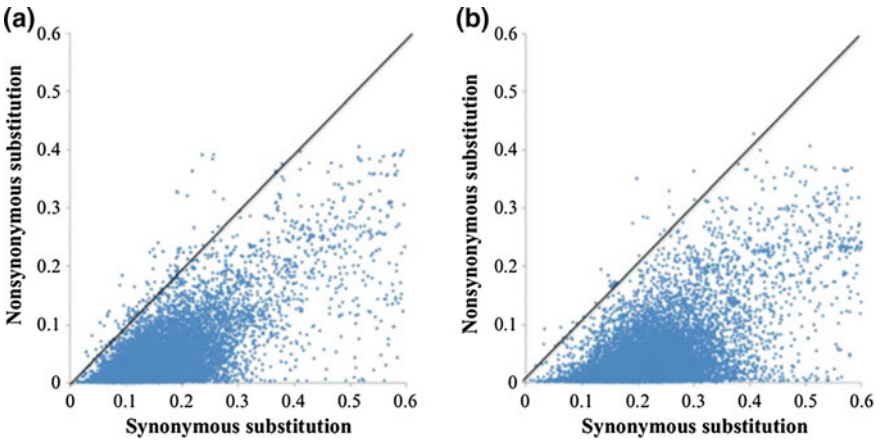

Fig. 5.18 Comparison of synonymous substitutions (horizontal axis) and nonsynonymous substitutions (vertical axis). **a** Comparison between mouse and rat. **b** Comparison between human and rhesus macaque

Because the number of synonymous substitutions per synonymous sites (*Ds*) and that of nonsynonymous substitutions per nonsynonymous sites (*Dn*) are simultaneously estimated for the orthologous proteins of different species (or different paralogous genes), comparison of *Ds* and *Dn* values is routinely conducted for many studies of genome comparison (see Chap. 16 for estimating methods). Figure 5.18a and b shows two examples for genome-wide comparisons: (a) between mouse and rat and (b) between human and rhesus macaque. In both cases, $Ds > Dn$ for the majority of protein coding genes.

It should be noted that the rate of synonymous substitutions may not be identical with the mutation rate, for biases of codon usages exist [45] and some unknown factors also exist [46]. We will discuss the consequences of these sorts of purifying selection on synonymous substitutions in Chap. 6.

5.7.2 Junk DNA

Susumu Ohno proclaimed the characteristics of mammalian genomes as "So much 'junk' DNA in our genome" in 1972 [47]. Junk DNA means functionless DNA. In fact, only 1.5% of the human genome is used for protein coding [48], and the rest are mostly junk. They are interspersed repeats (LINES and SINES), microsatellites, other intergenic regions, and introns (see Chap. 12). A small fraction of noncoding genomic regions are highly conserved (e.g., [49–53]), and they are expected to have some functions such as enhancers. Even some SINE is known to obtain an important function during the mammalian evolution (e.g., [54, 55]). It is still true, however, that the majority of noncoding genomic regions are functionless, and just junk DNAs. There are reports of transcriptions on many noncoding regions

(e.g., [56]). However, these results were obtained by problematic ChIP-chip techniques [57] and found to be artifact by checking ChIP-seq techniques [58]. Graur et al. [59] also condemned the ENCODE project statement [60] as an "evolution-free gospel."

Because the f value of Eq. 5.38 is 1 for junk DNA and for synonymous sites, their evolutionary rates are expected to be similar, if we ignore heterogeneity of mutation rates in one genome. In fact, the number (~ 0.15) of nucleotide substitutions per site in intergenic regions for mouse and rat genomes was shown to be quite similar to that of synonymous substitutions [43].

If we ignore a small portion of functional DNAs that are highly conserved among diverse organisms, the majority (more than 90%; see [61]) of mammalian or all vertebrate genomes are junk DNAs. Therefore, a genome-wide divergence of two species is a good approximation of the consequence of pure neutral evolution.

5.7.3 Pseudogenes

Pseudogenes are DNA sequences which are homologous to functional genes, but themselves are no longer functional. For example, if there are frameshift mutations and/or stop codons in a DNA sequence highly homologous to a known functional gene, it is called "pseudogene," for functional protein is expected to be not formed. Therefore, they are often products of gene duplications. Because of their nonfunctional nature, pseudogenes should be genuine members of junk DNAs.

There are four types of gene duplication (see Chap. 3). Among them, RNA-mediated duplication produces intronless sequences via reverse transcription of mRNAs. These cDNAs will be integrated into a DNA region unrelated to its place of origin, where a series of gene regulatory sequences exist. Therefore, such cDNA inserts are almost always "dead on arrival." We can see a clear enhancement of evolutionary rate for intronless (or processed) pseudogenes for the mouse p53 gene. The estimated numbers of nucleotide substitutions between *M. musculus* and *M. leggada* are 0.0157 and 0.0651 for functional genes and pseudogenes, respectively (data from [62]).

Nonfunctionalization can happen without gene duplication. Vitamins are molecules that exist in small quantity but essential for organisms, especially human, to survive. By definition, vitamins are not produced by the organism itself, and they should be taken in as a part of food. Their very existences are enigmatic, for these molecules are coming from other organisms which produce them. If vitamins are so important, why they are not produced by a certain species such as human? The neutral theory of evolution easily resolves this paradox. If vitamins are abundant in every day foods, even the mutants with no ability of producing a certain vitamin are selectively neutral compared to wild types with the ability to produce that vitamin through the existing enzymatic pathway.

Vitamin C, or ascorbic acid, is a good example. If appropriate intake of vitamin C is stopped for a long time, human will develop scurvy. King and Jukes [38] already predicted that the lack of ascorbic acid production could be explained by

assuming the neutral evolution. Not only human but all primates (except for prosimians), elephants, guinea pigs, and fruit bats lack the ability of producing ascorbic acid [63]. Medaka, a teleost fish, also does not produce ascorbic acid [64]. In fact, nonfuctionalization of L-glono-γ-lactone oxidase (enzyme number E. C.1.1.3.8) gene was confirmed by Nishikimi and his collaborators [65].

A more drastic situation of pseudogene formation without gene duplication is found in parasitic bacterial genomes. *Mycobacterium leprae*, a causative bacteria of leprosy, was found to have many pseudogenes in its genome [66]. This is because these bacteria are hiding deep in host body and receive many nutrients from host.

A gene function is often quite complex, and it is not easy to determine if a "pseudogene" is really nonfunctional. Even if protein is not produced, mRNA or even DNA sequences themselves may still have some function. Therefore, when we discuss the evolution of pseudogenes, it may be too simplistic to assume that f, fraction of neutral mutations, is 1 for a pseudogene. A "pseudogene" with some function is not surprising, for they were named so only because of sequence comparison.

5.7.4 Neutral Evolution at the Macroscopic Level

So far, we discussed the evolution of nucleotide or amino acid sequences and saw that the fixations of selectively neutral mutations are the major process of evolution. It is thus natural to expect that the evolution at the macroscopic or so-called phenotypic level is also following mostly neutral fashion. Unfortunately, this logically derived conjecture seems to be not kept by many evolutionary biologists. Ever since Charles Darwin, many biologists have been enchanted by seemingly powerful positive selection. They are biologists who study macroscopic morphology of organisms, who study animal behaviors, who study developmental process, and so on. As we will see in Chap. 6, we should be careful to discuss adaptation without clear demonstration at the molecular level.

It may be still optimistic to expect a rapid expansion of our knowledge on the genetic basis of developmental and behavioral traits in the near future. However, modern biology is proceeding to this direction, and I personally hope that the superficial dichotomy between molecules (genotypes) and phenotypes will disappear sooner or later. Evolutionary genomics is at the foundation of this edifice of modern biology. It should be added that Nei (2013, [67]) covers many interesting topics related to this problem.

5.8 Historical Developments of Population Structure Analysis Under Neutral Evolution

A simple random mating population with a constant population size is often assumed in many population genetics theories. However, one population will be divided if its population size increases or its environment becomes more heterogeneous. Population differentiation after population split naturally occurs. Thus, one panmictic random mating population is just illusion and unrealistic. We therefore discuss the historical developments of various theories regarding the population structure within one species.

5.8.1 Hardy–Weinberg Ratio

Let us assume that a diploid population has two alleles, A_0 and A_1, and their frequencies are p_0 and p_1, respectively. There can be three genotypes, A_0A_0, A_0A_1, and A_1A_1, and their frequencies can be approximated by using the binomial distribution $(p_0 + p_1)^2$ if male and female allele frequencies are more or less the same:

$$\text{Freq\_}A_0A_0 = p_0^2, \tag{5.44a}$$

$$\text{Freq\_}A_0A_1 = 2p_0p_1, \tag{5.44b}$$

$$\text{Freq\_}A_1A_1 = p_1^2. \tag{5.44c}$$

This simple relation is often called "Hardy–Weinberg ratio," after two persons who independently showed this relationship in 1908 [19, 20]. It is straightforward to extend the 2-allele case to more than two allele cases. In some old books on evolutionary genetics [e.g., 68], call this ratio as "equilibrium" as if thee ratios are important. However, this ratio is simply the outcome of random mating in a diploid sexually mating population, and we should consider this ratio as an approximation to obtain genotype frequencies from allele frequencies.

5.8.2 Wahlund Principle

Existence of population structure was first analyzed by Wahlund [69]. For simplicity, let us consider two populations A and B in one species. These two populations shared their common ancestor long time ago, and now allele frequencies on many loci are somewhat different between these two extant populations. Let us consider one particular locus with only two alleles 1 and 2. Allele frequencies in population A are A_1 and A_2 ($A_1 + A_2 = 1$) and those in population B are B_1 and B_2 ($B_1 + B_2 = 1$). Let us assume the Hardy–Weinberg ratio in each population. Genotype frequencies of population A thus become A_1^2, $2A_1A_2$, and A_2^2 for genotypes

11, 12, and 22, respectively, and those for population B are B_1^2, $2B_1B_2$, and B_2^2, respectively. Now we ignore the population label, and consider allele frequencies (T_1 and T_2) of the total population T (populations A and B are combined) under a simplified assumption of equal population sizes for populations A and B. We then have $T_1 = (A_1 + B_1)/2$ and $T_2 = (A_2 + B_2)/2$. Expected frequencies for genotypes 11, 12, and 22 become T_1^2, $2T_1T_2$, and T_2^2, respectively under the simple assumption of random mating in this total population assuming the Hardy–Weinberg ratio. However, real genotype frequencies are $(A_1^2 + B_1^2)/2$, $(2A_1A_2 + 2B_1B_2)/2$, and $(A_2^2 + B_2^2)/2$ for genotypes 11, 12, and 22, respectively. Let us compare the expected frequency (E_f) and real frequency (R_f) of heterozygote 12. Noting $A_2 = 1 - A_1$ and $B_2 = 1 - B_1$,

$$E_f = 2T_1T_2 = 2\{(A_1+B_1)/2\}\{(A_2+B_2)/2\} = (A_1+B_1)-(A_1+B_1)^2/2. \tag{5.45}$$

$$R_f = (2A_1A_2+2B_1B_2)/2 = (A_1(1-A_1)+B_1(1-B_1)) = (A_1+B_1)-\left(A_1^2+B_1^2\right). \tag{5.46}$$

$$E_f-R_f = \left(A_1^2+B_1^2\right)-(A_1+B_1)^2/2 = (A_1-B_1)^2/2. \tag{5.47}$$

Unless $A_1 = B_1$ (then $A_2 = B_2$), $E_f - R_f > 0$. Therefore, if one total population contains two differentiated subpopulations, the proportion of heterozygotes decreases from the situation of panmictic random mating of the whole population. This is called the Wahlund principle or the Wahlund effect. A more general case of many subpopulations is given by Crow and Kimura [17] using variance of allele frequencies. The extreme situation of population differentiation may be selfing or clonal subpopulations of plants. Each clonal population may be homozygous for different alleles, and there will be no heterozygotes.

5.8.3 F_{st} and G_{st}

The random nature of allele frequency changes caused by finite number of population size was not considered in the Wahlund principle. Wright [70] considered this random effect for K random mating subpopulations in the total population with a finite population size. Expanding this idea, Wright [71] proposed three kinds of fixation indices: F_{IS}, F_{IT}, and F_{ST}, where I, S, and T stand for individual, subpopulation, and total population. These were originally defined as correlations of genes [71]. Nei [72, 73] showed that these three fixation indices can be obtained from the means and covariances of allele frequencies by extending the Wahlund's principle. Nei [74] generalized F_{ST} by considering gene diversity and the mean of minimum genetic distances of Nei [75] and called it G_{ST}. Weir and Cockerham [76] also studied F_{ST} by extending Cockerham [77]. It should be noted that F_{ST}, which is now widely used as a measure of differentiation between two populations, was

originally considered an average of population differentiation into many subpopulations. Therefore, it is rather a misnomer to call the measure of the genetic distance between two populations as $F_{ST.}$

5.8.4 Genetic Distances Between Two Populations

When researchers got interested in differentiations of human populations in physical anthropology, a series of population distances were developed, such as coefficient of racial likeness [78] or Mahalanobis' D2 statistics [79]; see Nei [80] and Nei [2] for review. Later, similar distances based on allele frequencies were developed, such as those developed by Sanghvi [81] and Cavalli-Sforza and Edwards [82]. Nei [75] proposed three genetic distances between two populations: minimum, standard, and maximum. Arithmetic and geometric means of single-locus genetic distances are used in minimum and maximum genetic distances. These genetic distances are discussed in Chap. 18.

References

1. Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge: Cambridge University Press.
2. Nei, M. (1987). *Molecular evolutionary genetics*. New York: Columbia University Press.
3. Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature, 420,* 520–562.
4. International Chicken Genome Sequencing Consortium. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature, 432,* 695–716.
5. Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature, 217,* 624–626.
6. Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16,* 97–159.
7. Bonner, J. T. (2008). *The social amoebae: The biology of cellular slime molds*. Princeton: Princeton University Press.
8. Cook, R. E. (1979). Asexual reproduction: A further consideration. *American Naturalist, 113,* 769–772.
9. Ewens, W. J. (1979). *Mathematical population genetics*. Berlin: Springer.
10. Watson, H. W., & Galton, F. (1875). On the probability of the extinction of families. *Journal of Anthropological Institute of Great Britain and Ireland, 4,* 138–144.
11. Haccou, P., Jagers, P., & Vatutin, V. A. (2005). *Branching processes: Variation, growth, and extinction of populations*. Cambridge: Cambridge University Press.
12. Crow, J. F. (1989). The estimation of inbreeding from isonymy. *Human Biology, 61,* 935–948.
13. Saitou, N. (1983). An attempt to estimate the migration pattern in Japan by surname data (in Japanese). *Jinruigaku Zasshi, 91,* 309–322.
14. Bienaymé, I. J. (1845). De la loi de multiplication et de la durée des familles. *Société Philimatique de Paris Extraits, Series, 5*(10), 37–39.
15. Fisher, R. A. (1930). The distribution of gene ratios for rare mutations. *Proceedings of Royal Society of Edinburgh, 50,* 205–220.
16. Feller, W. (1968). Introduction to probability theory and its applications (3rd ed., Vol. 1). New York: Wiley.

17. Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. New York: Prentice-Hall.
18. Howell, N. (1979). *Demography of the Dobe !Kung*. New York: Academic Press.
19. Saitou, N., Shimizu, H., & Omoto, K. (1988). On the effect of the fluctuating population size on the age of a mutant gene. *Journal of the Anthropological Society of Nippon, 96,* 449–458.
20. Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability, 19A,* 27–43.
21. Hudson, R. R. (1983). Testing the constant rate neutral allele model with protein sequence data. *Evolution, 37,* 203–217.
22. Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics, 105,* 437–460.
23. Fu, Y.-X. (2006). Exact coalescent for the Wright-Fisher model. *Theoretical Population Biology, 69,* 385–394.
24. Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution, 10,* 2–22.
25. Hein, J., Schierup, M. H., & Wiuf, C. (2005). *Gene genealogies, variation, and evolution—A primer in coalescent theory*. Oxford: Oxford University Press.
26. Wakeley, J. (2008). *Coalescent theory: An introduction*. Greenwood Village: Roberts & Co.
27. Kimura, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability, 1,* 177–232.
28. Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proceedings of National Academy of Sciences USA, 41,* 144–150.
29. Kimura, M., & Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature, 229,* 467–469.
30. Kimura, M., & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics, 49,* 725–738.
31. Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics, 61,* 893–903.
32. Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research, 1,* 247–269.
33. Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., et al. (2004). Detection of large-scale variation in the human genome. *Nature Genetics, 36,* 949–951.
34. Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science, 305,* 525–528.
35. Ohta, T. (1973). Slightly deleterious mutant substitions in evolution. *Nature, 246,* 96–98.
36. Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In V. Bryson & H. J. Vogel (Eds.), *Evolving genes and proteins* (pp. 97–166). New York: Academic Press.
37. Dayhoff, M. O. (1978). Survey of new data and computer methods of analysis. In M. O. Dayhoff (Ed.), Atlas of protein sequence and structure, Vol. 2, Supplement 3, pp. 1–8. Washington, DC: National Biomedical Research Foundation.
38. King, J. L., & Jukes, T. H. (1969). Non-Darwinian evolution. *Science, 164,* 788–798.
39. Saitou, N. (2007). *Introduction to genome evolution studies (in Japanese)*. Tokyo: Kyoritsu Shuppan.
40. 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.
41. Li, W. H., & Wu, C. I. (1987). Rates of nucleotide substitution are evidently higher in rodents than in man. *Molecular Biology and Evolution, 4,* 74–82.
42. Rhesus Macaque Sequencing and Analysis Consortium. (2007). Evolutionary and biomedical insights from the rhesus macaque genome. *Science, 316,* 222–234.
43. Abe, K., Noguchi, H., Tagawa, K., Yuzuriha, M., Toyoda, A., Kojima, T., et al. (2004). Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution

of strain C57BL/6J, as defined by BAC end sequence-SNP analysis. *Genome Research, 14,* 2239–2247.
44. Hendriks, W., Leunissen, J., Nevo, E., Bloemendal, H., & de Jong, W. W. (1987). The lens protein alpha A-crystallin of the blind mole rat, *Spalax ehrenbergi*: Evolutionary change and functional constraints. *Proceedings of the National Academy of Sciences of the United States of America, 84,* 5320–5324.
45. Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution, 2,* 13–34.
46. Suzuki, R., & Saitou, N. (2011). Exploration for functional nucleotide sequence candidates within coding regions of mammalian genes. *DNA Research, 18,* 177–183.
47. Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symposium in Biology, 23,* 366–370.
48. International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431,* 931–945.
49. 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.
50. Matsunami, M., & Saitou, N. (2013). Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. *Genome Biology and Evolution, 5,* 140–150.
51. Babarinde, I. A., & Saitou, N. (2016). Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics. *Molecular Biology and Evolution, 33,* 1807–1817.
52. Hettiarachchi, N., & Saitou, N. (2016). GC content heterogeneity transition of conserved noncoding sequences occurred at the emergence of vertebrates. *Genome Biology and Evolution, 8,* 3377–3392.
53. Saber, M. M., & Saitou, N. (2017). Silencing effect of Hominoid highly conserved non-coding sequences on embryonic brain development. *Genome Biology and Evolution, 9,* 2122–2133.
54. Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., et al. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature, 441,* 87–90.
55. Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N., et al. (2008). Possible involvement of SINEs in mammalian-specific brain formation. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 4220–4225.
56. Johnson, J. M., Edwards, S., Shoemaker, D., & Schadt, E. E. (2005). Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics, 21,* 93–102.
57. Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature, 447,* 799–816.
58. van Bakel, H., Nislow, C., Blencowe, B. J., & Hughes, T. R. (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biology, 8,* e1000371.
59. Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., & Elhaik, E. (2013). On the immortality of television sets: function in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution, 5*(3), 578–590.
60. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489,* 57–74.
61. Babarinde, I. A., & Saitou, N. (2013). Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. *Genome Biology and Evolution, 5,* 2330–2343.
62. Ohtsuka, H., Oyanagi, M., Mafune, Y., Miyashita, N., Shiroishi, T., Moriwaki, K., et al. (1996). The presence/absence polymorphism and evolution of p53 pseudo- gene within the genus Mus. *Molecular Phylogenetics and Evolution, 5,* 548–556.

63. Lehninger, A. L. (1975). *Biochemistry*. New York: Worth Publishers.
64. Toyohara, H., Nakata, T., Touhata, K., Hashimoto, H., Kinoshita, M., Sakaguchi, M., et al. (1996). Transgenic expression of L-gulono-gamma-lactone oxidase in medaka (Oryzias latipes), a teleost fish that lacks this enzyme necessary for L-ascorbic acid biosynthesis. *Biochemical and Biophysical Research Communications, 223,* 650–653.
65. Nishikimi, M., Fukuyama, R., Minoshima, S., Shimizu, N., & Yagi, K. (1994). Cloning and chromosomal mapping of the human nonfunctional gene for L-gulono-gamma-lactone oxidase, the enzyme for L-ascorbic acid biosynthesis missing in man. *Journal of Biological Chemistry, 269,* 13685–13688.
66. Cole, S. T., & others. (2001). Massive gene decay in the leprosy bacillus. Nature, 409, 1007–1011.
67. Nei, M. (2013). *Mutation-driven evolution*. Oxford: Oxford University Press.
68. Dobzhansky, T. (1951). *Genetics and the origin of species, third edition, revised*. New York: Columbia University Press.
69. Wahlund, S. (1928). Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas, 11,* 65–106.
70. Wright, S. (1943). Isolation by distance. *Genetics, 28,* 114–138.
71. Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics, 15,* 323–354.
72. Nei, M. (1965). Variation and covariation of gene frequencies in subdividied populations. *Evolution, 19,* 256–258.
73. Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics, 41,* 225–233.
74. Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA, 70,* 3321–3323.
75. Nei, M. (1972). Genetic distance between populations. *American Naturalist, 106,* 283–292.
76. Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38,* 1358–1370.
77. Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution, 23,* 72–84.
78. Pearson, C. (1926). On the coefficient of racial likeness. *Biometrika, 18,* 337–343.
79. Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of National Institute of Science, India, 2,* 49–55.
80. Nei, M. (1977). Genetic distances (in Japanese). In E. Matsunaga & K. Omoto (Eds.), Anthropology, Vol. 10, Genetics, pp. 29–62. Tokyo: Yuzankaku Shuppan.
81. Sanghvi, L. D. (1953). Comparison of genetical and morphological methods for a study of biological differences. *American Journal of Physical Anthropology, 11,* 385–404.
82. Cavalli-Sforza, L. L., & Edwards, A. W. F. (1967). Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics, 19,* 233–257.