Adaptive Evolutionary Genetics

Dec 18th (Tue), 2nd-5th periods 10:25～18:35
Dec 19th (Wed), 2nd-5th periods 10:25～18:35
Dec 20th (Thu), 2nd-4th periods 10:25～16:40
Lecture Room, Bioscience BLDG.
Lecturer: Prof. Shoji Kawamura
# Adaptive Evolutionary Genetics

2018 A1/A2/W Term

Dec 18th (Tue), 2nd-5th periods 10:25～18:35  
Dec 19th (Wed), 2nd-5th periods 10:25～18:35  
Dec 20th (Thu), 2nd-4th periods 10:25～16:40  
Lecture Room, Bioscience BLDG.  
Lecturer: Prof. Shoji Kawamura

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec 18th</td>
<td></td>
</tr>
<tr>
<td>Chapter 1</td>
<td>1</td>
</tr>
<tr>
<td>Introduction to</td>
<td></td>
</tr>
<tr>
<td>Evolutionary</td>
<td></td>
</tr>
<tr>
<td>Genetics</td>
<td></td>
</tr>
<tr>
<td>Chapter 2</td>
<td>9</td>
</tr>
<tr>
<td>Population</td>
<td></td>
</tr>
<tr>
<td>Genetics I</td>
<td></td>
</tr>
<tr>
<td>(Genetic</td>
<td></td>
</tr>
<tr>
<td>Variation and</td>
<td></td>
</tr>
<tr>
<td>Natural</td>
<td></td>
</tr>
<tr>
<td>Selection</td>
<td></td>
</tr>
<tr>
<td>Chapter 3</td>
<td>19</td>
</tr>
<tr>
<td>Population</td>
<td></td>
</tr>
<tr>
<td>Genetics II</td>
<td></td>
</tr>
<tr>
<td>(Random Genetic</td>
<td></td>
</tr>
<tr>
<td>Drift)</td>
<td></td>
</tr>
<tr>
<td>Dec 19th</td>
<td></td>
</tr>
<tr>
<td>Chapter 4</td>
<td>48</td>
</tr>
<tr>
<td>Population</td>
<td></td>
</tr>
<tr>
<td>Genetics III</td>
<td></td>
</tr>
<tr>
<td>(Nucleotide</td>
<td></td>
</tr>
<tr>
<td>Sequence Data)</td>
<td></td>
</tr>
<tr>
<td>Chapter 5</td>
<td>89</td>
</tr>
<tr>
<td>Molecular</td>
<td></td>
</tr>
<tr>
<td>Evolution I</td>
<td></td>
</tr>
<tr>
<td>(Nucleotide</td>
<td></td>
</tr>
<tr>
<td>Substitution)</td>
<td></td>
</tr>
<tr>
<td>Dec 20th</td>
<td></td>
</tr>
<tr>
<td>Chapter 6</td>
<td>93</td>
</tr>
<tr>
<td>Molecular</td>
<td></td>
</tr>
<tr>
<td>Evolution II</td>
<td></td>
</tr>
<tr>
<td>(Phylogenetic</td>
<td></td>
</tr>
<tr>
<td>Tree)</td>
<td></td>
</tr>
<tr>
<td>Chapter 7</td>
<td>107</td>
</tr>
<tr>
<td>The Neutral</td>
<td></td>
</tr>
<tr>
<td>Theory</td>
<td></td>
</tr>
<tr>
<td>Chapter 8</td>
<td>113</td>
</tr>
<tr>
<td>Adaptive</td>
<td></td>
</tr>
<tr>
<td>Evolution</td>
<td></td>
</tr>
<tr>
<td>Chapter 9</td>
<td>132</td>
</tr>
<tr>
<td>Perspective of</td>
<td></td>
</tr>
<tr>
<td>Evolutionary</td>
<td></td>
</tr>
<tr>
<td>Genetics</td>
<td></td>
</tr>
</tbody>
</table>

**Recommended Reading**

- Molecular Evolutionary Genetics, M. Nei, Columbia University Press (1987)
Chapter 1 Introduction to Evolutionary Genetics

Objectives of evolutionary genetics and this lecture

Evolutionary genetics covers population genetics and molecular evolution, which describe, quantify and analyze genetic diversity within a species and genetic divergence between species, respectively. Evolutionary genetics concerns the mutations of DNA that are transmitted from generation to generation. The objective of evolutionary genetics is twofold: the first is to reveal the evolutionary mechanism by which variation within and between species is produced and the second is to reveal evolutionary history of organisms in terms of phylogenetic relationships and divergence times among them. In this lecture I shall give you an overview of its basic concepts and guide you how to detect adaptive genetic changes in evolution.

1.1. What is "evolution"?

Current organisms are descended from past organisms.

The characteristic of evolution is:

• Varying degrees of relatedness among current organisms through ancestral species at varying time points in the past
• Adaptation to diverse habitats

Charles Darwin "On the Origin of Species" 1859

The formal book title: "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life"

This is the first influential book which argued that current organisms are descended from past organisms with modifications and proposed the concept of natural selection as a mechanism to explain the adaptation of organisms to diverse habits.

Pre-evolutionary explanation of origin of organisms

Creation by god (Creationism)

In six days, 10,000 years ago, the universe, earth and organisms were all created by God

Still powerful in public conception and politics (new forms keep arising: Intelligent Design, etc.)

Is "evolution" common knowledge? truth? hypothesis? science?

Evolution is not a common knowledge yet. It cannot be regarded as a truth in a general sense because it cannot be witnessed. But evolution is not merely a hypothesis. Evolution is a "scientific truth" which has been built from testable theorems based on numerous observations and experiments. Thus, it requires a study to understand, which can be hampered by personal belief or religious faith.
1.2. Evidence of evolution

(1) Fossils: remnants of past organisms

- Existence of a world long before human records and existence of organisms in the distant past
- Change of organisms through stratum chronological order
- Discovery of intermediate morphs between extant organisms

(2) Varying levels of relatedness defined by shared characters (natural system of classification)

- Nested nature of taxonomy (species/genus/family/order/class/phylum/kingdom)

(3) Correlation of geographical distribution of organisms and their relatedness

- Close relatedness in the same continent under different physical conditions (e.g. climates)
- Distant relatedness between different continents under similar physical conditions
- Consistency of geographical patterns of relatedness with past climate or land level

(4) Consecutive patterns of morphology, physiology, ontogeny, behavior, etc.

- Similarity of early embryonic morphology among vertebrates
- Correspondence of limbic system of human brain and brain structure of other vertebrates

(5) Proportionality of genetic (DNA sequence) similarity and taxonomic relatedness

(6) Sharing of pseudogenes or "dead" genes between species

(7) Testability of basic components of evolution, e.g. mutation, natural selection and random genetic drift

(8) Replication of evolution in laboratory under limited conditions (e.g. in bacterial culture)

1.3. Overview (I) of evolutionary research

Evolution = Accumulation of genetic change through generations

Only DNA changes in germ cells are inherited by the next generation and contribute to evolution

Should be aware of conceptual distinction with non-inheritable changes

⇒ Source of enormous evolutionary changes in geological time scale and of tremendous biodiversity today

Early studies

(1) On evolutionary history (phylogenetic tree): Paleontology, comparative embryology, taxonomy

- Inter-species differences, long-term evolution
  - Problems: Incomplete fossil records, difficulty to complex morphological/physiological traits
  ⇒ Disagreement on phylogenetic details

(2) On evolutionary mechanism (population dynamics of mutant genes): Population genetics

- Intra-species genetic variation, short-term evolution (succession of short-term is long-term)
  - Conceptual framework: birth of a mutation → spread to population → fixation (100% freq.)
  - Mathematical formulation of natural selection as driving force of population dynamics of mutation
  - Theorization of random genetic drift manifesting importance of chance effect
Many fundamental concepts devised:
- Effective population size, habitat shift, life history strategy, kin selection, etc.
- Problems: scarcity of real data

1.4. Overview (II) of evolutionary research
Advent of molecular biology
(1) DNA and amino acid sequences can be examined directly, feasibly, massively
- Discovery of "molecular clock"
- Precision of phylogenetic relationships (contribution to "history" study in evolution)
(2) Discovery of tremendous amount of intra-species genetic variation (contribution to "mechanism" study)
- Refinement of population genetic theorem
- Conceptualization of neutral theory
- Theoretical ground of molecular clock
- 'History' and 'mechanism' studies integrated
(3) DNA and amino acid sequences should be considered into separate functional units (domains)
- Positions in the three-dimensional structure, hydrophilic/hydrophobic regions, catalytic domains, etc.
- Nucleotide position in a codon (1st, 2nd, or 3rd), pairing nucleotides in an RNA secondary structure,
  bias of codon usage and anti-codon tRNA repertoire, nucleotide positions relevant to mRNA
  stability/tturnover/processing, noncoding regions relevant to transcription regulation, etc.
(4) Non-universality of the genetic code
(5) Innovation of chromosome labeling and deeper understanding of chromosomal evolution and diversity
(6) Discovery of multigene families
(7) Discovery of pseudogenes or dead genes
(8) Discovery of transposons (mobile elements) and conceptualization of selfish genes and junk DNA

1.5. History of life (Panels 1-3)
1.5.1. Azoic Era (ca. 4.6–3 billion years ago: BYA)
Birth of Earth: ca. 4.6 BYA
The oldest fossil: ca. 3.5 BYA, bacteria-like microfossil
- The first life estimated at ca. 3.8 BYA: earth crust and ocean appeared before it
- No free oxygen in the atmosphere and no ozone layer: direct radiation of large amount of UV
  →Experimental field of pre-life chemical evolution
1.5.2. Archeozoic Era (ca. 3–1 BYA)
The oldest stromatolite at ca. 2.9 BYA (a laminated, rock-like structure built mainly by cyanobacteria)
- Increase of free oxygen due to photosynthesis by cyanobacteria
Atmospheric oxygen reached to stability at ca. 2 BYA: threatening to anaerobic organisms

From anaerobic or photosynthetic (cyanobacteria) to aerobic or amphimicrobe; mitochondria symbiosis

Origin of eukaryotes at ca. 2.2 BYA

Origin of multicellular organisms at ca. 1.8 BYA

Split to animals and plants at ca. 1 BYA; basic gene sets of multicellular animals established around this time

1.5.3. Proterozoic Era (ca. 1–0.6 BYA)

"Snowball" Earth (global glaciation) at ca. 0.8–0.6 BYA; evolution inconspicuous

1.5.4. Paleozoic Era (ca. 600–250 MYA)

Explosive increase of large multicellular organism fossils: ca. 600–500 MYA (Cambrian Period)

Ediacara biota (ca. 560 MYA)

Cambrian Explosion (ca. 540 MYA)

Most of the modern phyla first appeared

Continental shelves by continental breakup, global warming, oxygen increase, ozone layer

The sea as a big experiment station of evolution toward complex forms of organisms

Origin of vertebrates at ca. 500–400 MYA: Jawless fish (agnathans), suction feeding from sea bottom, armored, ancestor of cyclostomes (lampreys and hagfish)

Jawed fish (gnathans); ca. 400 MYA, revolution of feeding method, new food resources

Devonian Period (ca. 400–350 MYA)

Age of Fish (diverse fish emerged; bony, cartilaginous, placoderm, etc)

To the land; first by plants (whisk fern-like) at ca. 400 MYA, then by animals (wingless insects, amphibians) lobe-finned fish → amphibians

Carboniferous Period (ca. 350–280 MYA)

Forest of ferns and gymnosperms; reptiles (complete land life enabled by amnion)

1.5.5. Mesozoic Era (ca. 250–65 MYA)

Age of Dinosaurs

Ferns and gymnosperms (pines, ginkgo, cycad palms, etc); ammonites in the sea

Flowering angiosperms arose in latter Mesozoic, accompanied with insect evolution (mediating pollination; example of coevolution)

Origin of mammals: ca. 200 MYA, small and nocturnal?

Origin of birds: ca. 200 MYA [debate on significance of archaeopteryx; Nature 516, 18-19 (2014)]

1.5.6. Cenozoic Era (ca. 65 MYA ~ today)

Age of Mammals

Angiosperms (flowering plans) flourished; birds adaptive radiation

Age of Humans: Quarternary period (ca. 2 MYA ~ today)
Panel 1

Figure 2.1. Geological time and the early history of life. From Schopf et al. (1983). Reprinted by permission of Princeton University Press.

Figure 2.2. Divergence of the vertebrate groups according to geological and morphological evidence. Modified from McLaughlin and Dayhoff (1972).

Panel 2

1.6. Geological events important to evolution

(1) Continental drift (Panel 4)
The concept developed by Alfred Wegener in 1912, but his theory was neglected with lack of a mechanism
Rediscovered and accepted after 1950's and subsumed by the theory of plate tectonics

Breakup
Supercontinent Rodinia $\rightarrow$ Four continents in Cambrian Period (550 MYA), with continental shelves,
global warming, ozone layer

Join
Two continents (Laurasia and Gondwana) in Devonian Period (375 MYA)
(a) Age of Fish/ To the land (amphibians)

One supercontinent Pangea in Carboniferous ~ Jurassic Periods (300~200 MYA)
(b) ~ (e) amphibians $\rightarrow$ reptiles $\rightarrow$ dinosaurs: gymnosperm forests, insects, ammonites

Breakup
175~125 MYA (e) ~ (f)
India/ Antarctica/ Australia vs. Africa/ South-America
North-America vs. Europe
Latter dinosaur age, angiosperms, insect evolution

100~75 MYA (g)
India vs. Antarctica/ Australia
India drifted north and joined southern Asia (50 MYA) $\rightarrow$ Himalaya Mountains
South America vs. Africa
South America drifted and finally joined North America (4~5 MYA)

50 MYA (h)
Australia vs. Antarctica
Australia drifted north and now located south of Asia

(2) Glacial periods: drastic change of temperature, separation/ connection of lands due to sea level change

$\Rightarrow$ Strong impact to geographic distribution of organisms
    e.g. Primitive mammals in Australia and South America (marsupials and monotremes)
$\Rightarrow$ Utility to estimate divergence times between organisms
Panel 3

Panel 4

1.7. Extinction

Causes of "normal" extinction

(1) Competition among species

(2) Drastic environmental changes such as glacial epochs, volcanic eruptions, earthquakes, tsunamis

These occur often and cannot explain mass extinction.

Mass extinction (five major ones so far)

(1) 500 MYA (at the end of Cambrian Period): trilobites mostly disappeared

(2) 440 MYA (at the end of Ordovician Period): 57% of marine invertebrate genera disappeared

(3) 370 MYA (at the end of Devonian Period)

(4) 240 MYA (at the end of Paleozoic Era): the largest extinction; 96% of marine animals, trilobites gone

P-T (Permian-Triassic) extinction

(5) 65 MYA (at the end of Mesozoic Era): dinosaurs, many animals and plants, ammonites gone

**K-T (Cretaceous-Tertiary) [recently renamed as K-Pg (Cretaceous-Paleogene)] extinction**

Causes of mass extinction are not fully understood. Massive crustal deformations or climatic changes can be caused by mantle plumes triggered by continental drift. Mass extinctions sweep away major animals/plants and deform ecosystem overall. Thus, mass extinctions have a tremendous impact on evolution thereafter.

An extraterrestrial impact by an asteroid is regarded as a plausible explanation for K-Pg extinction

A layer of rock dated to this period is rich in iridium which is rare on Earth but rich in meteorites

A 180-km-wide crater in Mexico's Yucatan Peninsula (called Chicxulub), a 36-km one in Iowa, another in Russia, found and dated to this period.

⇒ A 10-km-diameter meteorite hit the Earth and choked the skies with debris that starved the Earth of the sun's energy, throwing a wrench in photosynthesis and sending destruction up and down the food chain.

⇒ Probability of asteroid hit of this size to the Earth is once a every 100 MY, according to a calculation by astronomers

However, a massive volcanism by mantle plumes associated with continental drift is also another hypothesis.

Chance and directionality in evolution

Evolution was realized only once, like history, and fraught with many chance events. If the K-Pg asteroid flew a little behind or in advance, it may have not hit the Earth. Then, there may have been no dinosaur extinction, no mammalian radiation, and no human evolution. Putting aside such an extreme example, chance or randomness is everywhere in major processes of evolution; normal extinctions, occurrence of mutations, spread of mutations in a population. On the other hand, natural selection appears to provide directionality to evolutionary changes. Evolution is a magnificent drama played by chance and occasional directionality.
Chapter 2 Population Genetics I (Genetic Variation and Natural Selection)

2.1. Genetic variation

Genetic variation in a population can reveal the characters of individuals, populations and species. It is relevant to personal medicine, individual identification during criminal investigations, and paternity identification in forensics. Above all, it is a central concern in evolutionary biology for deciphering population history, phylogenetic relationship, and mechanisms behind population and species differentiation.

2.1.1. Allele (or gene) frequency and genotype frequency

When there is a variation in a particular gene within a species, the gene is called polymorphic or under polymorphism. In this lecture, genetic variation within a species is called genetic diversity and that between species is called genetic divergence. Each variant gene type is called an allele, which occupies the same gene locus in the chromosome with other alleles. Frequency of an allele in a population is called allele frequency or gene frequency. In diploid organisms, an individual with two identical alleles on a gene locus is called a homozygote for that gene locus. An individual with two different alleles is called a heterozygote. The composition of the two alleles in an individual is called genotype. Owing to this combinatorial nature, change of genotype frequency between generations is more complex than change of allele frequency. Thus, allele frequency is primarily used in population genetic studies.

Alleles can be defined in various ways based on methodologies used and research purposes, e.g. reactivity difference to antibodies such as A, B, and O blood type alleles, electrophoresis mobility difference of a protein, length difference of DNA fragments cut by a restriction enzyme (restriction fragment length polymorphism: RFLP), length difference of PCR-amplified DNA fragments (amplified fragment length polymorphism: AFLP), and ultimately, the nucleotide sequence difference.

2.1.2. Number of alleles and heterozygosity

Basic statistical quantities are called summary statistics. The most fundamental summary statistics on genetic polymorphism are the number of alleles and the broadly-defined heterozygosity. While the original-meaning of heterozygosity is the heterozygote frequency in a population regarding a gene locus, the broadly-defined heterozygosity is generalized to mean the probability that two randomly chosen genes at a locus in a population are different alleles. Thus, the heterozygosity reflects the allele-frequency configuration of a population. It is also applicable to haploid organisms (e.g. male bees and ants) and haploid genes, such as Y-chromosomal genes and mitochondrial DNA. Heterozygosity, broadly-defined, is also called gene diversity.
2.1.3. Hardy-Weinberg principle

The Hardy-Weinberg principle is a simple principle to relate allele frequency with genotype frequency. We first begin with simple algebra:

When genotype frequencies at a generation are observed to be

\[ AA: \ P \quad Aa: \ Q \quad aa: \ R \quad (P + Q + R = 1) \]

allele frequencies at the generation are determined to be:

\[ A: \ p = \frac{2PN + QN}{2N} = P + \frac{Q}{2} \quad a: \ q = R + \frac{Q}{2} \quad (N \text{ is the number of individuals.}) \]

This is a simple algebra and not the Hardy-Weinberg principle. From these allele frequency data, genotype frequencies are estimated by the principle as the followings.

When frequencies of alleles \( A \) and \( a \) are \( p \) and \( q \), respectively (\( p + q = 1 \)), genotype frequencies are expected to be:

\[ AA: \ p^2 \quad Aa: \ 2pq \quad aa: \ q^2 \]

When the number of alleles is \( n \) and the frequency of allele \( A_i \) is \( p_i \) (\( \sum_{i=1}^{n} p_i = 1 \)), genotype frequencies are expected to be:

- Homozygous genotypes \( A_iA_i: \ = p_i^2 \)
- Heterozygous genotypes \( A_iA_j: \ = 2p_ip_j \)

The observed and expected genotype frequencies concur under the conditions described below. In other words, we estimate genotype frequencies in one generation after or before present from the present genotype data. Genotype frequencies are passed to generation to generation as constant values under certain conditions. This constancy is called Hardy-Weinberg equilibrium (HWE). HWE is conventionally tested by evaluating the difference between the observation and expectation of genotypes through \( \chi^2 \) test:

\[ \chi^2 = \sum \frac{(obs - exp)^2}{exp} \quad (2.1) \]

Be sure to use the numbers of individuals and not the frequency values for this calculation. Probability values are obtained using a \( \chi^2 \) probability graph (e.g. Figure 1.11 of Panel 5). Degrees of freedom (d.f.) is the number of data classes (here, three genotype classes, \( AA, Aa, aa \)) minus 1 minus the number of parameter classes estimated from data (here, one allele frequency class, \( p \)). Thus, d.f. here is \( 3 - 1 - 1 = 1 \). The other allele frequency \( q \) is \( 1 - p \), and given if \( p \) is estimated. Thus, \( q \) is not included in the parameter class.
2.1.4. Conditions to meet HWE

1. Diploid organism
2. Sexual reproduction
3. Non-overlapping generations
4. No difference of allele frequencies between sex
   <Natural selection>
5. No difference of survival between genotypes (no natural selection to survival)
6. No difference of mating success between genotypes (no natural selection on mating)
7. No difference of gamete production or fertility between alleles (no natural selection at gametes)
   <Random mating>
8. No selectivity for genotypes in mating (no sexual selection) (random mating regarding genotypes)
9. No population differentiation (random mating regarding populations)
   <Chance effect>
10. Population is large enough (no stochastic fluctuation of allele frequencies = no random genetic drift)
    <Novel variant>
Hardy-Weinberg principle is an extension of Mendel's law (which premises diploidy and sexual reproduction) of segregation (which states that an $Aa$ heterozygote produces an equal number of $A$-bearing and $a$-bearing gametes). Thus, conditions 1 and 2 are prerequisites. Condition 3 is a mathematical modeling for simplicity, which can provide a sufficient approximation to actual complex generation structures, and is also not to worry about. Condition 4 is met in many cases and is not cared about.

However, conditions 5 to 12 appear to be all unrealistic and problematic. Conditions 5~7 are regarding natural selection, 8 and 9 are on random mating, 10 is about chance effect, and 11 and 12 are regarding novel variants. Life history can be divided into (1) zygote to adult phase (to form gamete pool of population) and (2) gamete to zygote phase. Conditions 5, 9, 11, and 12 are involved in phase (1) and conditions 6, 7, 8, and 10 are involved in phase (2) (Panel 6).

It seems improbable that all these conditions are satisfied. Effectiveness of HWE would be therefore quite questionable. In reality, however, HWE is met in most cases. Why? This is because HWE, in fact, concerns only two consecutive generations for deviation of genotype frequency from expectation. In many situations, each condition has no enough power to cause a significant deviation of genotype frequency in only one generation.

However, condition 9 (population differentiation) is the only exception. For example, imagine one population has only allele $A$ and another population has only allele $a$. If we do not know such differentiation between populations and regard the two populations as one random-mating population, we will be surprised to find no $Aa$ heterozygous individuals. Such differentiated populations in allele frequency are called subpopulations. Even though mating is random within a subpopulation, there is a large deviation from random mating in the total population because there are no or few mating opportunities between subpopulations, resulting in significant deviation from HWE. If mating is limited to a close kin (inbreeding), inbred families are analogous to subpopulations. There is also a large deviation from random mating in the population consisting of inbred families, resulting in significant deviation from HWE. Population differentiation is achieved through many generations. While other HWE factors have an effectiveness of only one generation, the population differentiation reflects many generations and thus has a large effect on HWE. Hence, HWE is, in effect, to test if the population is one random-mating population.

If reproductive isolation occurs by geographical barrier, geographic distance, or other reasons, population
differentiation occurs in whole genome alike. If natural selection or sexual selection operates to a particular trait in a local population, population differentiation outstands in a gene(s) or genomic region(s) relevant to the trait (Panel 7). By searching for genome regions showing higher-than-average differentiation, researchers are actively finding genes relevant to local adaptations (such as to high altitude adaptation).

2.1.5. Source and fate of genetic variation

Source of genetic variation is mutation. Mutations include single nucleotide or amino acid change (point mutation), local gene duplication and deletion, recombination, gene conversion, transposition, etc. Genetic variation is also generated through migration (gene flow) from other population.

A new allele generated by mutation or migration may proliferate, diminish, or maintain its frequency in a population, which is governed by natural selection and random genetic drift. These two are the star players of evolutionary genetics. This chapter focuses on natural selection.

Panel 6

Panel 7

Cause and effect of population differentiation

Reproduction isolation (geographic barrier etc.)

Natural/sexual selection to a trait
2.2. Natural selection

Natural selection is a theorem to explain how all species show exquisite adaptation to their habit environment. Charles Darwin and Russell Wallace made the first publication of this idea as coauthors in 1858 in Linnean Society. But this was not paid much attention to.

Darwin published his 22-years long research since his South American voyage on the Beagle as a form of "abstract". It convinced people that evolution is a fact and that natural selection is its mechanism. Its influence was not just on biology but also on to social thoughts.

⇒ "On the Origin of Species"
- There is a variation among individuals under nature. (Fact)
- More offspring are produced than their living conditions allow. (Fact)
- Competition ("Struggle for Existence": not necessarily a contest) arises inevitably. (Logical fact)
- Individuals having any slight advantage over others would have the best chance of surviving and of procreating their kind. (Inference drawn)
- Favorable variations would be preserved by a "strong principle of inheritance" and, through competition in every generation, eventually extinguish any slight injurious variations. (Inference drawn)

This "preservation of favorable variations and the rejection of injurious variations" or this "principle by which each slight variation, if useful, is preserved" is termed Natural Selection by Darwin.

Through the action of natural selection in geological time scale, organisms become better adapted to their environment.

The same area can support more living beings as they increasingly diverge in structure, habits, and constitution. Therefore during the incessant struggle of all species to increase the number, the more diversified these descendants become, the better will be their chance of succeeding in the battle of life. Thus, the small differences among varieties of the same species will steadily increase till they come to the greater differences between species of the same genus, or even of distinct genera. Natural selection, on the other hand, leads to extinction of the less improved forms of life. In this way, the nature of affinities of all organisms is explained, which relates all organisms to each other in group subordinate to other groups, forming species, genera, families, orders, classes, and so on.

Afterwards, united with Mendel's law of inheritance, and by elucidation of DNA as the physical entity of gene, natural selection is endowed with a central role in modern evolutionary genetics.
2.3. Population genetics

Evolution can be regarded as a process of fixation of a variant allele in a population and as the repetition of the process. After a mutant gene (new allele) arises in a population, by what process does it spread to and fixed in the population? How do we formulate natural selection in the process? These are the questions of population genetics.

2.3.1. Fitness

To quantify natural selection, the concept of "fitness" was devised. Fitness is the number of offspring, surviving to maturity, that one individual can contribute to the next generation. When there are individuals with different genotype and if there is a difference in fitness between them, we state that natural selection is operating. Some people seem to treat natural selection as a kind of supernatural power, but, in reality it is just a situation where fitness is different between individuals.

In population genetics, fitness is defined as the average number of surviving offspring to maturity per individual. For organisms of non-overlapping generation like annual plants, fitness is directly this number (designated as $W$: called "selective value"). For overlapping-generation organisms like humans, it is the exponential change rate of number of individuals (designated as $m$: requiring a complex calculation using a function expressing survival rate and reproductive rate at every age). However, in general, $m$ can be approximated by the number of matured daughters per mother. An approximation $m = \ln W$ stands and the two cases can be treated interchangeably.

2.3.2. Positive and negative selection

When an allele is superior to other alleles in viability and fertility, and spreads to a population, the natural selection in such case is called positive selection, which is a popular and typical situation exemplifying Darwinian evolution. On the other hand, when an allele is deleterious to viability or fertility of the individuals carrying it, and is eventually eliminated from the population, such natural selection is called negative selection or purifying selection.

In principle, the term "positive" or "negative" represents the orientation of selection to spreading through or being eliminated from the population. But, the terms are used in many cases for a new mutant (or minority variant). In fact, a positive selection to a minor new mutant is also negative selection to the common standing variations. Negative selection to a new mutant is positive selection to the common types.

Well recognized examples of positive selection (to minority alleles) are actually rare. A few examples include the industrial melanism of moss and increment of pesticide-resistant genes. Many others are speculative. On
the other hand, examples of negative selection (to minority alleles) are countless, such as lethal mutations and hereditary disease mutations. Positive selection to minority alleles implies a change of organisms, while negative selection to minority alleles implies conservation of the current state. In this sense, much enthusiasm is devoted to find a positive selection to minority alleles to understand the diversity of life.

2.4. Change of allele frequency by natural selection (a deterministic model)

Let's calculate the process of frequency change of alleles in a population using fitness values of genotypes. In most cases, no absolute values but only relative values of fitness are required. Let's take one genotype as the standard with relative fitness 1, and give the others relative fitness values such as \((1+s)\) or \((1-s)\). We call \(s\) as selection coefficient. The hidden assumption here is the infinity of population size and no stochastic fluctuation of allele frequency. Thus, this is called a deterministic model.

2.4.1. Typical example 1 (Codominance selection)

Let \(A\) and \(a\) be a wild type allele and a new mutant allele at a locus, respectively. Assume fitness of genotypes \(AA, Aa, aa\) as 1, \(1+s\), \(1+2s\), respectively. Assume \(s\) is a small and positive value and \(s^2\) is indifferent to \(s\). \(q_t\) and \(p_t\) \((p_t + q_t = 1)\) are the allele frequencies of \(a\) and \(A\), respectively, at generation \(t\). Let's calculate time course change of \(q_t\).

**Genotype frequencies at generation \(t\)**

\[
\begin{array}{ccc}
AA & Aa & aa \\
p_t^2 & 2pq_t & q_t^2 \\
\end{array}
\]

(Hardy-Weinberg principle)

**Genotype frequencies at generation \(t+1\)**

\[
\begin{array}{ccc}
AA & Aa & aa \\
p_{t+1}^2 & \frac{2p_t q_t (1+s)}{w} & \frac{q_t^2 (1+2s)}{w} \\
\end{array}
\]

\(w\): average fitness in the population, \(w = p_t^2 + 2p_t q_t (1+s) + q_t^2 (1+2s)\)

\[
\Delta q_t = q_{t+1} - q_t \\
= \frac{q_t^2 (1+2s)}{w} + \frac{p_t q_t (1+s)}{w} - q_t \\
= \frac{sp_t q_t}{1+2s q_t} = sp_t q_t \quad \text{(for small \(s\))}
\]

(Refer to Panel 8 for more detail)

Transforming it to a differential equation \(\frac{dq_t}{dt} = sp_t q_t\), the solution is given by

\[
\ln \left( \frac{q_t}{p_t} \right) = \ln \left( \frac{q_0}{p_0} \right) + st \\
\Leftrightarrow \frac{q_t}{p_t} = \frac{q_0}{p_0} e^{st} \\
\Leftrightarrow q_t = \frac{q_0}{q_0 + p_0 e^{-st}}
\]
The logarithm of mutant-to-wild type ratio increases linearly through time, in other words, the mutant-to-wild type ratio increases exponentially through time. This equation is useful to have a rough idea on how quickly an advantageous mutation spreads through a population when the population is large enough.

For example, if $a$ is 0.1% more advantageous than $A$ (i.e. $s = 0.001$), it takes 13,813.5 generations for the mutant allele to increase from 0.1% to 99.9%. If one generation is one year, this is about 14,000 years. If the advantage is 10 times larger (1% more) (i.e. $s = 0.01$), it is about 1,400 years (Panel 9: Figure 2.1). This is very short in the geological time scale.

Panel 8

<table>
<thead>
<tr>
<th>Generation t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotype</td>
</tr>
<tr>
<td>$AA$</td>
</tr>
<tr>
<td>$Aa$</td>
</tr>
<tr>
<td>$aa$</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>$p_t^2$</td>
</tr>
<tr>
<td>$2p_t q_t$</td>
</tr>
<tr>
<td>$q_t^2$</td>
</tr>
<tr>
<td>$1$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No. indiv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Np_t^2$</td>
</tr>
<tr>
<td>$N2p_t q_t$</td>
</tr>
<tr>
<td>$Nq_t^2$</td>
</tr>
<tr>
<td>$N$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$</td>
</tr>
</tbody>
</table>

1 + $s$\[1 + 2s\]

Panel 9

\[\Delta q = q_{t+1} - q_t = \frac{q_t^2(1 + 2s)}{1 + 2s q_t} + \frac{p_t q_t (1 + s)}{1 + 2s q_t} - q_t = \ldots = \frac{s p_t q_t}{1 + 2s q_t} \approx s p_t q_t\]
Implication is: if a clearly advantageous mutant allele arises, old alleles could be replaced with the new advantageous allele in relatively a short while. By repeating this process, possible mutations are "tested" one after another at each locus, and the best ones are retained.

2.4.2. Typical example 2 (Overdominance selection)

When counteractive selections are operating, allele frequencies can be maintained at equilibrium. Such selection is generically called balancing selection. A typical example is the overdominance selection where heterozygous genotype has the highest fitness. Assume fitness of genotypes $AA$, $Aa$, $aa$ as 1, $1+s$, $1+t$, respectively ($s \geq 0$). Let allele frequencies of $A$ and $a$ be $p$ and $q$, respectively. Then, we obtain

$$\Delta q = \frac{pq(tq + s - 2sq)}{1 + 2spq + tq^2} \quad (2.2)$$

Transforming (2.2) to a differential equation and solving it, we obtain the time course change of $q$ for various initial frequency values of $a$ (Panel 9: Figure 2.2). The frequency of $a$ in the equilibrium can be obtained by assuming $\Delta q = 0$ in equation (2.2) and is given by

$$q = \frac{s}{2s - t}$$

If $t = 0$, the equilibrium frequency is 50%.

2.4.3. A generalization of fitness difference among genotypes

<table>
<thead>
<tr>
<th></th>
<th>$AA$</th>
<th>$Aa$</th>
<th>$aa$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1 + $s$</td>
<td>1 + $t$</td>
</tr>
</tbody>
</table>

When $t > 0$ and $s = ht$,

- $h > 1$  heterozygote superiority (overdominance)
- $h = 1$  $a$ is dominant ($A$ is recessive)
- $0 < h < 1$  $a$ is codominant
- $h = 0$  $a$ is recessive ($A$ is dominant)
- $h < 0$  heterozygote inferiority (disruptive selection)
- $t = 0$ and $s > 0$  overdominance
- $t = 0$ and $s < 0$  heterozygote inferiority
- $t = 0$ and $s = 0$  perfectly neutral
Chapter 3 Population Genetics II (Random Genetic Drift)

3.1. Random genetic drift (a stochastic model)

3.1.1. Finite population

The deterministic model we have seen is a simplified and idealistic model assuming infinitely large population size. It provides a useful approximation if population size is indeed large enough or the effect of natural selection is very large. However, in reality, population size is finite. The effect of the finiteness is not negligible. We need a stochastic model to understand how allele frequency changes through generations.

3.1.2. Allele frequency change by chance

Gametes are produced almost infinitely in a natural population of sexually reproducing organisms (gamete pool). However, only a fraction is sampled to reproduce the next generation. The allele frequency is not retained in the small sample. This fluctuation of allele frequency occurs by chance and is called random genetic drift or sampling effect. This was first noted by Sewall Wright (1931) and Ronald Aylmer Fisher (1930) independently. Thus, the stochastic model on random genetic drift is called Wright-Fisher model.

Panel 10 is simplistic explanatory figures of the sampling effect. Assume a diploidy population comprising $N$ individuals in every generation. The first gamete pool reflects the allele frequency of its producers. From the pool, $N$ male gametes and $N$ female gametes are randomly sampled and unite to form $N$ individuals of the next generation. Note that the allele frequency in the sample is varied from the pool by chance alone.
repeating this process, deviation is accumulated and allele frequency shifts greatly from the original state through generations.

Imagine you have one child. She/he has only one of your allele. Even if you have some more children, one of your alleles still may not be inherited to any of them. This is a familiar example of random genetic drift.

3.1.3. Probability function, mean, and variance of allele frequency

Panel 11 shows computer simulations of allele frequency change by random genetic drift. It monitors the fate of one allele \( A \). One trial continues until allele frequency becomes 1 (fixed) or 0 (lost). One trial corresponds to one line in the figures and corresponds to a fate of the allele \( A \) in one population. Fates differ among trials (populations).

Look at one population. Let the initial frequency of allele \( A \) be \( p_0 \). The probability of having \( i \) number of allele \( A \) in the sampled \( 2N \) gametes, \( P(\frac{i}{2N}) \), is equal to the product of the combination number of
choosing $i$ from $2N$, $\binom{2N}{i} = \frac{2N!}{i!(2N-i)!}$, the probability of choosing allele $A$ in $i$ times, $p_0^i$, and the probability of choosing non-$A$ alleles in all the rest, $(1 - p_0)^{2N-i}$. Thus, given by

$$P\left(\frac{i}{2N}\right) = \binom{2N}{i} p_0^i (1 - p_0)^{2N-i} \tag{3.1}$$

as a binomial distribution. "Binomial" means the repetition of alternative choice of $A$ or non-$A$.

When $N$ is large enough and $p_0$ is small enough, the binomial distribution can be approximated to the Poisson distribution, which is given by

$$P\left(\frac{i}{2N}\right) = \frac{m^i}{i!} e^{-m} \tag{3.1}^{'}$$

where $m$ is the expected value (mean) of $i$ and is $2Np_0$.

The equations (3.1) and (3.1)’ can be felt familiar with, when you think of some extreme cases such as the fate of a newly arisen mutation to be lost in as early as the next generation (Panel 12).

**Fate of a newly arisen mutation**

Panel 12

The initial number of mutant is 1, thus initial freq. is $p_0 = \frac{1}{2N}$

The mean number in the next generation is $2Np_0 = 1$

The probability of being lost in the next generation, $P\left(\frac{0}{2N}\right)$, is

if $N$ is not large, from (3-1),

if $N$ is large enough, from (3-1)’,

$$\left(\frac{1}{2}\right)^2 = 0.25 \quad \text{corresponds to one self-fertilizing flower}$$

$$N = 2 \quad \left(\frac{3}{4}\right)^2 \approx 0.316$$

$$\vdots$$

$$N = 10 \quad \left(\frac{19}{20}\right)^{20} \approx 0.358$$

$$\vdots$$

$$N = 100 \quad \left(\frac{199}{200}\right)^{200} = 0.367$$

$$N \rightarrow \infty \quad \frac{1}{e} = \frac{1}{2.718...} \approx 0.368$$

Even by self fertilization, it is lost by $\frac{1}{4}$ probability

In a very large sample, still lost by a little more than $\frac{1}{3}$ probability

The expected number of $A$ after one generation is $2Np_0$, so the expected frequency after one generation $E(p_1)$ is given by dividing it with $2N$. Thus,

$$E(p_1) = p_0$$

equal to the initial frequency. The variance of the number of $A$ after one generation is $2Np_0(1 - p_0)$, so the variance of frequency of $A$ after one generation $V(p_1)$ is given by dividing it with $(2N)^2$. This is because variance is the mean of the square of deviation from expectation. Thus,
\[ V(p_t) = \frac{p_0(1-p_0)}{2N} \]

which becomes larger for smaller population size.

By extension, the expected frequency of \(A\) after \(t\) generations \(E(p_t)\) is still equal to the initial frequency \(E(p_t) = p_0\).

However, variance becomes
\[
V(p_t) = p_0(1-p_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right]
\]

(3.2)

(For more detail, refer to Nei's "Molecular Evolutionary Genetics", p357~359)

When \(t = 0\), \(V(p) = 0\). When \(t \to \infty\), \(V(p_t) \to p_0(1-p_0)\)

When \(\frac{t}{2N} \ll 1\) (i.e. \(t \ll 2N\)),
\[
V(p_t) = p_0(1-p_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right] \approx p_0(1-p_0) \frac{t}{2N}
\]

(For a tip of approximation, refer to Panel 13.)

These formulas of \(V(p_t)\) indicate that deviation of allele frequency from the initial frequency increases linearly through time at the beginning. The deviation is amplified through time. The deviation magnitude is larger for smaller populations (Panel 11). In conclusion, allele frequency changes through generations by chance alone. This is a manifestation of random genetic drift.

A refresher of mathematics

Panel 13

Let the initial freq. be \(p_0 = \frac{1}{2N}\)

When \(N\) is large enough, according to formula (3-1)', \(P\left(\frac{i}{2N}\right) = \frac{1}{i!} e^{-m}\)

\[
\sum_{i=0}^{\infty} P\left(\frac{i}{2N}\right) = 1 = \frac{1}{e} \left(1 + 1 + \frac{1}{2} + \frac{1}{3!} + \frac{1}{4!} + \ldots\right) = e
\]

Generally, let the initial freq. be \(p_0\) and the mean be \(m\)

When \(N\) is large enough, \(P\left(\frac{i}{2N}\right) = \frac{m^i}{i!} e^{-m}\)

\[
\sum_{i=0}^{\infty} P\left(\frac{i}{2N}\right) = 1 = \frac{1}{e^m} \left(1 + \frac{m^2}{2} + \frac{m^3}{3!} + \frac{m^4}{4!} + \ldots\right) = e^m \text{ (Maclaurin's expansion)}
\]

When \(|x| \ll 1\)

\[e^x \approx 1 + x, \quad e^{-x} \approx 1 - x\]

Thus, \((1-x)^n \approx e^{-nx}\)

If \(|nx| \ll 1\)

\[e^{-nx} \approx 1 - nx \approx (1-x)^n\]

Very often used in this lecture
3.1.4. Diffusion model

How can we describe the time course of random genetic drift? As in Panel 14: Figure 2, let's assume that there are several subpopulations and that they all contain allele A with equal frequency at generation 0. Random genetic drift occurs in each subpopulation. After many generations, eventually some subpopulations are fixed with allele A and the others lose it. The proportion of the fixed subpopulations is expected to be equal to the initial frequency \( p_0 \). What is the intermediate stage? Panel 14: Figure 7.5 illustrates the distribution of subpopulations according to their allele frequency of allele A at every generation.

The basic process to derive this graph is the equation (3.1) which gives the probability to an allele frequency at a generation based on the allele frequency at one generation before. However, it is not practicable to trace allele frequency changes in all populations through all generations by applying (3.1). Instead, an approximation is applied which is analogous to heat diffusion through a metal rod, and is called the diffusion model (Panel 15). For more detail, refer to "A Primer of Population Genetics" (p 91-94).

The diffusion model itself may be too highly mathematical for non-specialists (like me!) to comprehend. But it is useful to note several important outcomes from the model. These include the probability of a mutant allele to be eventually fixed in a population (fixation probability) and the average time to fixation (given that it is eventually fixed) (conditional fixation time), and the probability of a mutant allele to be eventually
lost (extinction probability) and the average time to loss (given that it is eventually lost) (conditional extinction time). We shall come back to these topics in Chapter 5.

### 3.2. Population differentiation

Diffusion model may be too complex for non-specialists. In fact, the progress of random genetic drift can be evaluated in more a simple way from a viewpoint of population differentiation.

#### 3.2.1. Variance of allele frequency among populations as a measure of population differentiation

In Panel 11, the individual lines represent fate of individual populations regarding an allele (allele $A$). As random genetic drift proceeds, allele frequency becomes differed among populations by chance alone. Random genetic drift brings in population differentiation. Thus, we can regard the variance of allele frequency among populations as a progress indicator of random genetic drift and as a measure of population differentiation. The variance increases from the minimum of 0 (equal frequency in all populations) to the maximum of $p_0(1-p_0)$ (i.e. allele $A$ fixed in $p_0$ proportion of the populations and lost in $(1-p_0)$ proportion of the populations). Thus, let's indicate how close the variance is to the maximum value by giving

$$ V(p) = \frac{p_0(1-p_0)}{p_0(1-p_0)} = 1 - (1 - \frac{1}{2N}) \equiv F_{ST} \tag{3.3} $$

Wright defined $F_{ST}$ in this way as a measure of population differentiation. It ranges from 0 to 1.

#### 3.2.2. Heterozygosity as a measure of population differentiation

When looking at random genetic drift within a population, it can be regarded as a process of losing diversity in the population by fixation or loss of mutations. In other words, it is a process of lowering heterozygosity in a population. Then, let's try to transform the $F_{ST}$ formula (3.3) using heterozygosity.

As in statistics textbooks,

$$ V(p) = E(p^2) - \{E(p)\}^2 \tag{3.4} $$

Here, the expectation of $p^2$, $E(p^2)$, has a meaning of average homozygosity of allele $A$ among subpopulations based on HWE. The expectation of $p$, $E(p)$, is the average frequency of allele $A$ among subpopulations, which in turn can be regarded as the allele frequency of a uniform total population which is formed hypothetically by fusing all the subpopulations. Thus, $\{E(p)\}^2$ can be regarded as the homozygosity of allele $A$ in the total population based on HWE. Now, the familiar statistics formula (3.4) has a special meaning: as random genetic drift progresses, homozygosity of subpopulations becomes larger than homozygosity of total population. This is noted as Wahlund's principle.

The equation (3.4) holds for any allele. Imagine we describe equation (3.4) for every allele and add up the
left side through and right side through, that is
\[
\sum V(p) = \sum E(p^2) - \sum \{E(p)\}^2 \quad (3.5)
\]

Here, according to the equation (3.3), the left side of the equation (3.5), \( \sum V(p) \), can be described as
\[
\sum V(p) = F_{ST} \sum p_0 (1 - p_0) = F_{ST} \sum (p_0 - p_0^2) = F_{ST} (\sum p_0 - \sum p_0^2) = F_{ST} (1 - \sum p_0^2)
\]

Here, \( \sum p_0^2 \) means the homozygosity of the initial population based on HWE. Note the allele frequency of the total population does not change from that of the initial population (see page 22: \( E(p_t) = p_0 \)). Thus, \( \sum p_0^2 \) can be regarded as the homozygosity of the total population. Then, \( 1 - \sum p_0^2 \) means the heterozygosity of the total population, \( H_T \). In sum, the left side of the equation (3.5) is \( H_T F_{ST} \).

In the right side of the equation (3.5), \( \sum E(p^2) - \sum \{E(p)\}^2 \), \( \sum E(p^2) \) is the average homozygosity among subpopulations based on HWE. Using the average heterozygosity among subpopulations, \( \overline{H_s} \), it can be expressed as \( 1 - \overline{H_s} \). On the other hand, \( \sum \{E(p)\}^2 \) is the homozygosity of the total population based on HWE. It can be expressed as \( 1 - H_T \). In sum, the right side of the equation (3.5) is \( H_T - \overline{H_s} \).

Hence, the equation (3.5) can be rewritten as
\[
H_T F_{ST} = H_T - \overline{H_s} \quad (3.6)
\]

By transforming it, we obtain
\[
F_{ST} = \frac{H_T - \overline{H_s}}{H_T} = 1 - \frac{\overline{H_s}}{H_T} \quad (3.7)
\]

In "\( F_{ST} \)" "S" stands for Subpopulation and "T" stands for Total population. We see \( F_{ST} \) defined by allele frequency is equivalent with \( F_{ST} \) defined by heterozygosity. The latter definition (3.7) may be more popular.

From equation (3.6),
\[
\overline{H_s} = H_T (1 - F_{ST}) \quad (3.8)
\]
Further from equation (3.3)

\[ H_S = H_T \left( 1 - \frac{1}{2N} \right)^t \]  

(3.9)

Thus, we see the average heterozygosity of subpopulations decreases from the heterozygosity of total population (and initial population) by \( \frac{1}{2N} \) in every generation toward 0.

In conclusion, **progress of random genetic drift can be expressed using a degree of population differentiation** which can be quantified using variance of allele frequency (3.3) or heterozygosity (3.7).

### 3.3 Hierarchical population structure

By analogy from the equation (3.7), we compare the observed heterozygosity in a subpopulation, \( H_I \) ("I" stands for Individual), with the estimated heterozygosity based on HWE, \( H_S \).

\[ F_{IS} = \frac{H_S - H_I}{H_S} \]

If a subpopulation comprises inbred families, the subpopulation can be regarded as comprising further subpopulations (i.e. comprising hierarchical population structure). Within an inbred family, allele is biased to one kind and heterozygotes are in shortage (but, HWE holds within the family). Because different families are fixed with different alleles, the subpopulation as a whole still holds allelic variation. Thus, the estimated HWE-heterozygosity of the subpopulation is larger than the observed heterozugosity. In this way, \( F_{IS} \) provides a measure of population differentiation within a subpopulation and of deviation from HWE by inbreeding. Statistical significance of the deviation can be evaluated by permutation test where, in computer simulation, random mating is run many times under the given number of individuals and the given allele frequency and by checking if the observed \( F_{IS} \) is an outlier in the distribution of simulated \( F_{IS} \). Thus, \( F_{IS} \) is another way of HWE test than the equation (2.1) (chi-square test).

By averaging \( H_I \) and \( H_S \) among subpopulations, respectively, we redefine \( F_{IS} \) as

\[ F_{IS} = \frac{H_S - H_I}{H_S} = 1 - \frac{H_I}{H_S} \]  

(3.10)

Then, \( 1 - F_{IS} = \frac{H_I}{H_S} \)

From equation (3.7), \( 1 - F_{ST} = \frac{H_S}{H_T} \). Thus, \( (1 - F_{IS})(1 - F_{ST}) = \frac{H_I}{H_T} \) which we define as \( 1 - F_{IT} \). Hence,

\[ F_{IT} = \frac{H_T - H_I}{H_T} = 1 - \frac{H_I}{H_T} \]  

(3.11)
In equation (3.11), \( F_{IT} \) compares the estimated HWE-heterozygosity of the total population with the observed heterozygosity averaged among subpopulation. The deviation comes from combined effects of the dual population structure: the total population comprises subpopulations which comprise inbred families.

Inbred families and subpopulations are the same in that mating is limited to a certain range. In this sense, subpopulation is a form of inbreeding. Even if group members mate randomly, just because the mating range is limited, deviation from HWE accumulates in a larger scale and a hierarchical population structure emerges.

### 3.4. Fixation index \( F_t \)

As we have seen, random genetic drift is evident in small population. Small population means that mating is within the limited range and thus, in a sense, inbreeding. In this section, we reconsider the inbreeding from a viewpoint of gene genealogy back to an ancestry (coalescence) and redefine the random genetic drift by it. The concept revolutionizes the population genetics.

Let's assume a constant size population. If all members of the population transmit their diploid alleles by one copy each to the next generation, allele frequency does not change. Allele frequency changes because

![Diagram showing paths of ancestry of a set of alleles sampled in the present generation. The population is represented as having a constant size. The alleles in the original population are represented at the top. As generations progress forward in time (downward in the diagram), many alleles leave no descendants and therefore go extinct. Eventually one allele goes to fixation. Considering this process in reverse (bottom to top in the diagram), the sample observed in the present generation undergoes a series of coalescent events in which the \( k \) alleles present in the present generation had only \( k - 1 \) ancestral alleles. The coalescences continue backward in time until there is only one ancestral allele. The filled circles represent alleles present in previous generations that have left no descendant alleles in the present generation.](image)

someone does not transmit one or both alleles and someone else transmits two or more copies instead. In other words, there are two or more gene copies from one generation ago derived by DNA replication (i.e. brothers/sisters). **It is only by chance which one is transmitted twice or more, once, or no time (thus lost)** (Panel 16: Figure 3.14). **This is another view of random genetic drift.**

Two genes sampled from a population are traced back to a replication of a single DNA molecule (i.e. a common ancestral gene). This ancestral relation through a single molecule is called **coalescence**. When coalescent genes are identical in DNA sequence (i.e. no mutation has occurred), these two are called **"identical by descent (IBD)"**. The probability is called **fixation index** $F_t$ that two randomly chosen genes are coalescent by $t$ generations ago and IBD in a random-mating and kin-relation-unknown population. In this lecture, the probability is called **coalescence index** $C_t$ that two randomly chosen genes are coalescent by $t$ generations ago irrespective of having mutations in such populations. By ignoring mutations, $F_t$ is equivalent with $C_t$. When $F_t$ is 1, all genes in a current population are copies of single gene in an ancestor in $t$ generations, and the current population is fixed regarding the gene locus. Thus, the fixation index provides an average population-wide kinship, in other words, a degree of inbreeding of the population.

By the way, the IBD probability of two genes from two individuals with known kinship is called **coefficient of consanguinity**, **coefficient of kinship**, or **coefficient of co-ancestry**, and is different from the average population-wide kinship $F_t$. For example, the coefficient of consanguinity between full siblings is $\frac{1}{4}$. The probability that an IBD gene exists in a particular kin is called **coefficient/degree of relatedness/relationship** or simply **relatedness**. For example, the relatedness between full siblings is $\frac{1}{2}$. The IBD probability of the two alleles of an individual is called **inbreeding coefficient**. When meant in the context of the average population-wide kinship, inbreeding coefficient is equivalent with $F_t$ of the two parental genes. When meant in the context that the kinship between the parents is known, it is equivalent with the coefficient of consanguinity of the parental genes.

Let me return to the main topics. Let's express $F_t$ by a mathematical formula (Panel 17). Let's begin with a simpler case of no mutation. A constant size population ($2N$ genes) is assumed. Assume we pick up one gene from a population at random. This must be descended from a gene at generation 0. Then, pick up another gene at random. When the two coalesce by $t$ generations ago, it is either at one generation ago or in the remaining $t-1$ generations before it. Let's consider the coalescence at one generation ago. The second gene goes back to the only one parental gene of the first gene among $2N$ genes in the population. Thus, the
probability is \( \frac{1}{2N} \). Then, let's consider the coalescence in \( t - 1 \) generations before it. This is the probability

of no coalescence at one generation ago, \( 1 - \frac{1}{2N} \), multiplied by the probability of coalescence in \( t - 1 \) generations, i.e. \( F_{t-1} \). Thus, \( F_t \) is the sum of the two cases:

\[
F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1} \tag{3.13}
\]

Transforming it, we obtain a geometric progression function

\[
1 - F_t = (1 - \frac{1}{2N})(1 - F_{t-1}) \tag{3.14}
\]

By solving it, we obtain

\[
1 - F_t = (1 - \frac{1}{2N})^t (1 - F_0) \tag{3.15}
\]

Note we consider descents initiated by a DNA replication after the generation 0. Thus, by definition, \( F_0 = 0 \),

\[
F_t = 1 - (1 - \frac{1}{2N})^t \tag{3.15}
\]

Note \( F_t \) is independent from the initial allele frequency, and increases from 0 to 1 by progression of \( t \).

Panel 17

The probability of no coalescence in one generation is \( 1 - \frac{1}{2N} \) in every generation. Thus, the probability of

no coalescence throughout \( t \) generations is \( (1 - \frac{1}{2N})^t \). \( F_t \) is the probability of coalescence at some time in \( t \) generations, thus, should be given by subtracting the probability of no coalescence throughout \( t \) generations.
from 1. Indeed, the equation (3.15) is in such form.

Panel 18: Fibure 9 shows that $F_t$ increases as time goes by (i.e. population goes to uniform or inbred). The increment is rapid when population is small.

3.4.1. $F_t$, $F_{ST}$, and heterozygosity

Look at the equations (3.15) and (3.2). The $F_t$ defined by the coalescence theory appears in the equation for variance of allele frequency among subpopulations. By a step further, from equation (3.3), the $F_{ST}$ defined by the concept of population differentiation is equivalent with the $F_t$ defined by the coalescence theory. As time goes by, the degree how much populations differentiate is equivalent with the degree how much a population is becoming a kin group.

Then, let's replace $F_{ST}$ in equation (3.8) with $F_t$. Also, replace the total population heterozygosity, $H_T$, with the initial population heterozygosity, $H_0$, and rewrite the average subpopulation heterozygosity, $\overline{H_S}$, with a subpopulation heterozygosity at generation $t$, $H_t$. Then, we see the relation of heterozygosity to the population-wide kinship and the time course of $H_t$.

$$H_t = H_0 (1 - F_t) = H_0 (1 - \frac{1}{2N})^t$$  (3.16)

When $t \to \infty$, $F_t \to 1$ and $H_t \to 0$
Thinking about the meaning of $F_t$, the equation (3.16) is easy to understand (Panel 19). Genes at a fraction of $F_t$ in a current population are copies of an ancestral gene at generation 0 through DNA replications, and are identical (no mutation is assumed here). The heterozygosity in this fraction is 0. On the other hand, genes at a fraction of $1 - F_t$ at the current population are not descendant of single genes but descendant of different genes at generation 0. The allelic composition at generation 0 is expected to be retained in this $1 - F_t$ fraction, and the heterozygosity in this fraction is expected to be the same with the initial heterozygosity, $H_0$. Thus, the total current heterozygosity is given as the product of $H_0$ and $1 - F_t$.

Panel 19

Relation of heterozygosity to fixation index (no mutation)

In the $1 - F_t$ fraction of the population, other coalescences than that of the $F_t$ fraction would have occurred. But, averaging over many subpopulations, allelic composition in the $1 - F_t$ fraction can be regarded to be the same with the initial heterozygosity. Also, as will be explained in page 37, probability of having other coalescence events in a sample is negligible.

The equation (3.16) also indicates that heterozygosity decreases by $\frac{1}{2N}$ in every generation. Remember that the coalescent probability in one generation is $\frac{1}{2N}$. The decrease of diversity is due to the increase of copy by the effect of random genetic drift.

In fact, $F_t = F_{ST}$ holds only in an idealized situation where, as shown in Panel 11 or Panel 20: Figure 12.1A, subpopulations are all in equal size and derived from an initial population at the same time and are equally related to one another. In reality, this model almost never applies. Populations have some phylogenetic
(historical) relationships as illustrated in Panel 20: Figure 12.1B. Yet, \( \frac{H_T - H_S}{H_T} \) is useful as a measure of population differentiation and is used very often in actual researches.

Panel 20

"F" of \( F_{ST} \) was originally given by Wright from the relation to fixation index \( F_t \). Nei proposed to call \( \frac{H_T - H_S}{H_T} \) as \( GST \) for solely an indicator of population differentiation. Even if populations are related as in Panel 20: Figure 12.1B, the equation (3.12), \( (1 - F_{IS})(1 - F_{ST}) = 1 - F_{IT} \), still holds. Anyway, the equation (3.9) is useful to understand roughly how heterozygosity changes by random genetic drift.

\( F_{ST} \) (or \( GST \)) is often used as a measure of genetic distance between two populations. \( F_{ST} \) (or \( GST \)) is, however, a measure of entire population differentiation among multiple (ideally infinitely many) subpopulations, and is not appropriate for a measure of differentiation between only two populations. "Two" corresponds to doing the simulation in Panel 11 only twice. Due to stochastic fluctuation, allele composition in the total population would not reflect that in the initial population. Even the equation (3.2) would not hold.

For measures of genetic distance between two populations based on allele composition, methods have been developed such as Nei's genetic distance. Refer to

- Molecular Evolutionary Genetics, M. Nei, Columbia University Press (1987)

3.5. Effective population size

The genetic diversity and evolution concerns the number of individuals who participate in reproduction which is generally fewer than actual number of individuals in a population. The generalized number of
reproductive individuals is the **effective population size**, with which random mating is assumed allowing self-fertilization. This is indicated as \( N_e \) and formulated for several cases so that the decrease of genetic variation is \( \frac{1}{2N_e} \) in every generation (or variance of allele frequency change is \( \frac{p(1-p)}{2N_e} \) in every generation).

**Example 1) Fluctuation in population size**

By extension of equation (3.14), \( F_t \) is more precisely expressed in

\[
1 - F_t = (1 - \frac{1}{2N_{t-1}})(1 - F_{t-1}) = (1 - \frac{1}{2N_{t-1}})(1 - \frac{1}{2N_{t-2}})(1 - F_{t-2}) = (1 - \frac{1}{2N_{t-1}})(1 - \frac{1}{2N_{t-2}})...(1 - \frac{1}{2N_0})
\]

By equation (3.15), the left side can be replaced with \( (1 - \frac{1}{2N_e})' \). Thus,

\[
(1 - \frac{1}{2N_e})' = (1 - \frac{1}{2N_{t-1}})(1 - \frac{1}{2N_{t-2}})...(1 - \frac{1}{2N_0})
\]

By approximation,

the left side of the equation gives \( (1 - \frac{1}{2N_e})' \approx (1 - \frac{t}{2N_e}) \) (if \( \frac{t}{2N_e} \ll 1 \), or \( t \ll 2N_e \))

the right side gives \( (1 - \frac{1}{2N_{t-1}})(1 - \frac{1}{2N_{t-2}})...(1 - \frac{1}{2N_0}) \approx 1 - \sum_{i=0}^{t} \frac{1}{2N_i} \) Thus,

\[
\frac{1}{N_e} = \frac{1}{t} \left( \frac{1}{N_1} + \frac{1}{N_2} + \frac{1}{N_3} + ... + \frac{1}{N_t} \right)
\]

(3.17)

It is apparent that smaller \( N \) contributes more to \( N_e \). Even if \( N \) is large throughout most generations, only one severe drop of population size significantly influences \( N_e \). This is known as **bottleneck** effect.

**Example 2) When distinction of male and female is necessary (e.g. in the case of unequal sex ratio)**

We have assumed an idealized random mating system where no distinction of male and female is made and even self-fertilization is allowed. Consider the probability that two alleles in an individual are coalescent in the closest ancestor. Refer to Panel 21.

In the idealized case, the closest ancestor is in just one generation ago (parental generation). The probability is \( \frac{1}{2N_e} \) as explained in page 29.

In the case of distinction of male and female, the two alleles of an individual must come from different individuals (one male and one female). Let the number of reproductive males and females to be \( N_m \) and \( N_f \) respectively. Thus, the closest coalescent ancestor is in two generations ago (grandparental generation).
Focus on a parental gene pair (father gene and mother gene). We consider two cases where both came from the same grandfather gene or the same grandmother gene. The former: probability that one gene came from male is $\frac{1}{2}$, probability that the other gene came also from male is $\frac{1}{2}$, and probability that the two came from the same gene among $2N_m$ genes is $\frac{1}{2N_m}$. Thus, probability that a parental gene pair came from the same grandfather gene is the product of the three probabilities, $\frac{1}{8N_m}$. In the same way, probability in grandmother side is $\frac{1}{8N_f}$. Regarding the idealized and sexual distinction cases equivalent, we have

$$\frac{1}{2N_e} = \frac{1}{8N_m} + \frac{1}{8N_f},$$

thus,

$$N_e = \frac{4N_mN_f}{N_m + N_f} \quad (3.18)$$

Even if there are millions of males, if there is only one female, the population is equivalent with a population of only $N_e = 4$. It is quite common that males and females differ in reproductive individual number. The equation (3.18) is often used in the studies of sex difference of reproductive strategy, dispersal pattern, etc.

As we have seen, effective population size $N_e$ is generally much smaller than actual population size $N$. Hereafter, we shall use $N_e$ to describe population size.
3.6. Expected time of coalescence
How long does it take for \( n \) genes sampled from a population to go back in time and reach to the first coalescence in the past? By considering this question, we can estimate the time for a novel mutant gene to be fixed in the population by random genetic drift.

When the probability of an incidence is \( P \) in every trial, consider the waiting time for the first event to occur. Here, we consider the situation where the first coalescence occurred at \( t \) generations ago among \( n \) genes sampled from a population. How can we formulate the expectation of \( t \) and the probability distribution of \( t \), \( P(t) \). Note that \( P(t) \) is different from the probability of coalescence by \( t \) generation ago, \( C_t \) (or \( F_t \), if no mutation is assumed). The probability of coalescence at one generation ago is \( C_1 \). \( P(t) \) is the probability that, going back in time, no coalescence occurs during \( t - 1 \) generations and a coalescence occurs at the last generation. That is,
\[
P(t) = C_1 (1-C_1)^{t-1}
\]
(see Panel 22)

Thus, the expected time of the coalescence, \( \bar{t} \), is given by
\[
\bar{t} = \sum_{t=1}^{\infty} t C_1 (1-C_1)^{t-1}
\]

**Waiting time for first coalescence**

Panel 22

\[ 1 - C_1 = D \], defined, then,
\[
\bar{t} = C_1 (1 + 2D + 3D^2 + \cdots)
\]

\[ S_n = 1 + 2D + 3D^2 + \cdots + nD^{n-1} \], defined, then,
\[
DS_n = D + 2D^2 + \cdots + (n-1)D^{n-1} + nD^n \], thus, subtracting this from the above gives
\[
(1 - D)S_n = 1 + D + D^2 + \cdots + D^{n-1} - nD^n
\]
\[ X = 1 + D + D^2 + \cdots + D^{n-1}, \text{ defined, then,} \]

\[ DX = D + D^2 + \cdots + D^{n-1} + D^n, \text{ thus, subtracting this from the above gives} \]

\[ X = \frac{1 - D^n}{1 - D}, \text{ thus} \]

\[ (1 - D)S_n = \frac{1 - D^n}{1 - D} - nD^n, \text{ that is,} \]

\[ S_n = \frac{1 - D^n}{C_1^2} - \frac{nD^n}{C_1} \]

\[ \lim_{n \to \infty} nD^n = 0, \text{ and} \lim_{n \to \infty} D^n = 0, \text{ thus,} \lim_{n \to \infty} S_n = \frac{1}{C_1^2}, \text{ thus} \]

\[ t_i = \frac{1}{C_1} \quad (3.19) \]

It tells that the expected waiting time for the first coalescence is the reciprocal of the probability of coalescence in one generation.

When considering only two samples, the probability of coalescence in one generation \(C_1\) is

\[ C_1[n = 2] = \frac{1}{2N_e} \]

Thus, the expected waiting time for the first coalescence of two randomly chosen samples from a population, \(t_2\), is

\[ t_2 = 2N_e \text{ generations} \quad (3.20) \text{ (Panel 23)} \]

Then, let's consider a case of \(n\) samples (Panel 24). The probability \(C_1(n)\) that coalescence occurs among \(n\) samples in one generation is given by subtracting the probability of no coalescence in one generation from 1.

\[ C_1(n) = 1 - \left( \frac{2N_e - 1}{2N_e} \right) \left( \frac{2N_e - 2}{2N_e} \right) \left( \frac{2N_e - 3}{2N_e} \right) \cdots \frac{2N_e - (n - 1)}{2N_e} = 1 - \left( \frac{1}{2N_e} \right) \left( \frac{2}{2N_e} \right) \left( \frac{3}{2N_e} \right) \cdots \left( \frac{n - 1}{2N_e} \right) \]
When \( N_e \) is large enough and \( n \) is small enough relative to \( N_e \), an approximation gives

\[
C_i(n) \approx 1 - \left(1 - \frac{1}{2N_e}(1 + 2 + 3 + \cdots + n - 1)\right) = \frac{1}{2N_e} \cdot \frac{n(n-1)}{2} = \frac{n(n-1)}{4N_e}
\]

This is equivalent with the probability that two samples are randomly chosen from \( n \) samples and they coalesce in one generation. In other words, this is the probability that \( n \) genes reduces to \( n-1 \) by one coalescence event in one generation. It is possible that other coalescent pairs or three or more genes coalesce to one in one generation. However, when \( N_e \) is large enough and \( n \) is small enough relative to \( N_e \), these probabilities are negligible compared to the probability of coalescence of only two at a time.

Thus, the expected waiting time for the first coalescence among \( n \) samples, \( \bar{t}_n \), is given by equation (3.19)

\[
\bar{t}_n = \frac{4N_e}{n(n-1)} \text{ generations} \quad (3.21)
\]

**Panel 24**

**Coalescence waiting time among \( n \) samples: \( \bar{t}_n \)**

By extension, the expected total time that \( n \) genes take to coalesce eventually to one, i.e. the time to Most Recent Common Ancestor (\( T_{MRCA} \)), is the sum of the coalescence steps from \( n \) to \( n-1 \), to \( n-2 \), to \( \cdots \), to 2, and finally to 1 (Panel 25: Figure 3.15). Thus,

\[
T_{MRCA} = \sum_{i=2}^{n} \bar{t}_i = \sum_{i=2}^{n} \frac{4N_e}{i(i-1)} = 4N_e \sum_{i=2}^{n} \left(\frac{1}{i-1} - \frac{1}{i}\right) = 4N_e(1 - \frac{1}{n}) \text{ generations} \quad (3.22)
\]

As the sample size increases, \( T_{MRCA} \) approaches to \( 4N_e \). Hence, the expected waiting time for all genes in a population to coalesce to a single ancestral gene is \( 4N_e \) generations. This is the fixation time mentioned in page 23. Without using the complex diffusion model, the simple coalescence model solved it.
3.7. Random genetic drift and mutation

Thus far, we have ignored mutation for simplicity. All genes in a population are eventually descendant from replication of a single gene by random genetic drift. Thus, a population loses genetic diversity and becomes homogenized. However, in reality, mutations occur and supply genetic diversity to the population. The removal of variation by random genetic drift and the supply of variation by mutations go to a balance. Hence, without action of natural selection to maintain variation (balancing selection), genetic variation can persist in a population. By incorporating mutation to the equation of fixation index, we can formulate equilibrium between random genetic drift and mutation. Further, we can obtain expected values of heterozygosity and the number of alleles in population samples from the formulas.

3.7.1. Fixation index $F_t$ and mutation

Let mutation rate per generation per gene and effective population size be $u$ and $N_e$, respectively. Even mutations allowed, the coalescence index $C_t$ still follows the equation (3.15).

$$C_t = 1 - \left(1 - \frac{1}{2N_e}\right)^t$$ (3.23)

However, this equation does not hold for $F_t$ any longer. This is because coalescent genes could differ in DNA sequence due to mutation(s). Here, $F_t$ is the probability that two randomly chosen genes coalesce in $t$ generations and are exempted from mutation during the time. This is given by multiplying the probability of exempting mutation in one generation on either gene, $(1 - u)^2$, to each term of equation (3.13) (Panel 26).

$$F_t = \left[\frac{1}{2N_e} + \left(1 - \frac{1}{2N_e}\right)F_{t-1}\right](1 - u)^2$$ (3.24)

This is a recurrence formula and means that mutations are avoided throughout $t$ generations.
As shown in Panel 18: Figure 9, without mutation, coalescent copies rapidly increase in early phase and gradually slowdown toward the end. Somewhere in this process, the coalescence and mutation counterbalance each other. At the equilibrium, it is like the situation that coalescence supplies copies but mutation scribbles on them, and no more increase of copies occurs (see Panel 27).

At equilibrium, $F_t = F_{t-1} = F^{*}$. Let it expressed as $F^{*}$, then equation (3.24) is transformed as
\[ \hat{F} = \left[ \frac{1}{2N_e} + (1 - \frac{1}{2N_e}) \hat{F} \right](1 - u)^2 \]

\[ \hat{F} = \frac{(1-u)^2}{2N_e - (2N_e - 1)(1-u)^2} \approx \frac{1-2u}{4N_eu + 1} \approx \frac{1}{4N_eu + 1} \equiv \frac{1}{\Theta + 1} \] (3.25) (approximation by \( u \ll 1 \))

4\( N_eu \) is conventionally symbolized as \( \Theta \) and is a key parameter to describe genetic variation as explained later. \( \Theta \) reflects the number of mutant genes in a population in one generation and is called population mutation rate.

### 3.7.2. Heterozygosity at equilibrium

When considering the time scale for population differentiation or within-species genetic variation, mutation rate is small enough not to occur at a same nucleotide site twice or more. Thus, every mutation is safely assumed to create a novel allele. This is called infinite allele model. With mutations and random genetic drift, the genetic variation comes from (1) the \( 1 - C_t \) fraction of genes which are not descendent from a single ancestral gene and (2) the \( C_t - F_t \) fraction of genes which are descendent from a single ancestral gene but has mutations. Thus, by following the equation (3.16), heterozygosity at generation \( t \), \( H_t \), is given by

\[ H_t = H_0(1-C_t) + C_t - F_t \] (3.26)

As time goes by, \( C_t \) reaches to 1 according to equation (3.23). That is, \( C_t \) at equilibrium is 1. \( F_t \) at equilibrium is given by equation (3.25). Thus, the heterozygosity at equilibrium, \( \hat{H} \), is given as

\[ \hat{H} = 1 - F_t = \frac{4N_eu}{4N_eu + 1} = \frac{\Theta}{\Theta + 1} \] (3.27)

When \( N_eu \ll 1 \), that is, when \( u \ll \frac{1}{N_e} \), effect of mutation is much smaller than random genetic drift. Then,

\[ \hat{H} \to 0 \quad \text{This is the same as the case of no mutation assumed.} \]

When \( N_eu \gg 1 \), that is, when \( u \gg \frac{1}{N_e} \), effect of mutation is much larger than random genetic drift. Then,

\[ \hat{H} \to 1 \quad \text{All genes become different each other.} \]

Panel 28: Figure 12 illustrates equation (3.27). When \( 4N_eu \gg 1 \), very large genetic variation can be retained without action of balancing selection. In humans, the mutation rate per nucleotide site per year is inferred to be about \( 1 \times 10^{-9} \) based on the species divergence time estimated from fossil records and the neutral theory to be explained in Chapter 5. Thus, the mutation rate \( u \) per gene (~1000 bp) per generation (~25 years) is about \( 2.5 \times 10^{-5} \). The \( N_e \) of human is estimated to be about \( 1 \times 10^4 \) as explained in Chapter 4. Thus, \( 4N_eu \) per gene in humans is about 1, enabling human population to retain a large variation. At the genome level, the mutation rate per genome (~3 x 10^9 bp) is about 75 per generation and \( 4N_eu \) is about \( 3 \times 10^6 \). Thus,
herozygosity at genome level is practically 1. This is a ground for argument of the neutral theory explained in Chapter 7.

Panel 28

3.7.3. The number of alleles at equilibrium

At equilibrium of random genetic drift and mutation, we can also predict the number of alleles $k$ in $n$ samples randomly chosen from a population based on the $\Theta$ value. The expectation of $k$, $E(k)$, is given as

$$E(k) = 1 + \frac{\Theta}{\Theta + 1} + \frac{\Theta}{\Theta + 2} + \ldots + \frac{\Theta}{\Theta + n - 1}$$

Ewens' formula \hspace{1cm} (3.28)

When $\Theta \ll 1$, that is, when $Neu \ll 1$, effect of mutation is much smaller than random genetic drift, then, $E(k) \to 1$

When $\Theta \gg 1$, that is, when $Neu \gg 1$, effect of mutation is much larger than random genetic drift, then, $E(k) \to n$

3.7.4. Test of neutrality and population-size (demographic) constancy (Ewens-Watterson test)

At equilibrium between random genetic drift and mutation, the frequencies of alleles distribute so as to satisfy the heterozygosity under equation (3.27) and the number of alleles under equation (3.28). The distribution of allele frequency (frequency spectrum), that is, how many alleles are expected to be at low/high and middle frequencies, can be predicted (Panel 29)

The spectrum can be predicted from the probability that each of $k$ alleles in the sample takes $n_i$ number:

$$\text{Pr}\{n_1, n_2, \ldots, n_k, k\} = \frac{n!\Theta^k}{k!n_1n_2\cdots n_kS_n(\Theta)}$$

$S_n(\Theta) = \Theta(\Theta + 1)(\Theta + 2)\cdots(\Theta + n - 1)$

Ewens-Watterson formula \hspace{1cm} (3.30)
For example, when $k = 3, n = 10$ in the sample, probability of $n_1 = 2, n_2 = 3, n_3 = 5$ is
\[
\frac{10! \Theta^3}{3!2 \times 3 \times 5 \times \Theta(\Theta + 1)(\Theta + 2) \cdots (\Theta + 9)}
\]
If $\Theta = 1$, this is $\frac{1}{180}$.

This frequency spectrum is derived under the assumptions that population size is constant between generations and that alleles are not different in fitness effect (natural selection does not operate), i.e. allelic variation is selectively neutral. Hence, if the observed frequency pattern differs from the expectation, these assumptions are meant to be violated. The test of the deviation is called **Ewens-Watterson test**. The following two patterns of deviation are of particular importance (Panel 29).

Panel 29

(1) **Low and high frequency alleles are more abundant and middle frequency alleles are less abundant than expectation:** Every variant starts as a minority. Thus, this pattern is expected to occur when genetic variation is eliminated by bottleneck or founder effect and, since then, enough time has not passed for variants to increase to the middle level (equilibrium has not been attained). Novel variants are still minor and old alleles remain major. This is also expected to occur when population is expanding (novel alleles are all minor). When immigrants are minority from a population with distinct allele composition, this pattern is also expected. When a strong positive selective operates on a gene allele and drives it to fixation in relatively a short while, the gene and its surrounding are devoid of variation (selective sweep). Then, this pattern is expected to appear on the gene region.
When mutations in a gene are all slightly deleterious, variants are kept in low frequency by purifying selection. In this case, too, the pattern is expected to appear on the gene region.

(2) Low and high frequency alleles are less abundant and middle frequency alleles are more abundant than expectation: This pattern is expected to occur when population is shrinking (minor alleles are more likely to be lost). When immigrants are common from a population with distinct allele composition, this pattern is also expected. When balancing selection operates on a gene, this pattern is expected in and near the gene region.

3.8. Random genetic drift and migration

Genetic variation is supplied not only by mutation but also by migration. Because migration rate is much larger than mutation rate in general, effect of migration on genetic variation is much larger than mutation. Unlike mutation, migration suppresses population differentiation. By consideration of equilibrium between migration and random genetic drift, estimation of migration rate is possible.

3.8.1. Fixation index $F_t$ and migration

A variety of migration pattern is known. The simplest is the island model. Assuming equally differentiated populations as shown in Panel 20: Figure 12.1A, $m$ fraction of genes is randomly chosen from each population, pooled and mixed. Then, equal number of genes with the offer for pooling are returned to each population (Panel 30: Figure 6.18). Thus, allele frequency composition of the migrant genes is expected to be the same with that of total population.

Panel 30

The $m$ is a fraction of genes in a population which immigrated in one generation, and represent the migration rate per generation per gene. Let the effective population size be $N_e$. Under the assumption of
migrations, unlike mutations, the equation (3.23) does not hold for the coalescence index $C_t$, the probability of coalescence of two randomly chosen genes in a population in $t$ generations. This is because migrated genes are not replication copies from the parental generation in the population. By an analogy from the equation (3.24), $C_t$ is given by multiplying $(1 - m)^2$, the probability that either of two genes are not migrant.

$$C_t = \left[\frac{1}{2N_e} + (1 - \frac{1}{2N_e})C_{t-1}\right](1 - m)^2 \quad (3.31)$$

If ignoring mutations, $F_t$ is equivalent with $C_t$. Then,

$$F_t = \left[\frac{1}{2N_e} + (1 - \frac{1}{2N_e})F_{t-1}\right](1 - m)^2 \quad (3.32)$$

This is a recurrence formula and means that migrations are avoided throughout $t$ generations (Panel 31).

Panel 31

At the equilibrium, it is like the situation that coalescence supplies copies but the copies are replaced with migrants, and no more increase of copies occurs (see Panel 32).

At equilibrium, $C_t = C_{t-1}$. Let it expressed as $\hat{C}$, then as in equation (3.25), equation (3.31) is transformed as

$$\hat{C} \approx \frac{1}{4N_e m + 1} \quad (3.33)$$

(approximation by $m \ll 1$)

Under no mutation assumed, likewise,

$$\hat{F} \approx \frac{1}{4N_e m + 1} \quad (3.34)$$
3.8.2. Heterozygosity at equilibrium

Ignoring mutations, with migration and random genetic drift, the genetic variation comes from the $1 - C_t$ fraction of genes which are not descendent from a single ancestral gene. In island model, the allele frequency composition of migrants can be regarded as the same with that of total population (and initial population). Thus, the heterozygosity at generation $t$, $H_t$, is given as in equation (3.26). Here, $C_t = F_t$ (no mutation)

$H_t = H_0(1 - C_t)$

At equilibrium, heterozygosity $\hat{H}$ is given based on equation (3.33) as

$$\hat{H} = \frac{4N_em}{4N_e m + 1} H_0$$

(3.35)

When $N_e m \ll 1$, that is, when $m \ll \frac{1}{N_e}$, effect of migration is much smaller than random genetic drift. Then, $\hat{H} \to 0$  
This is the same as the case of no migration assumed.

When $N_e m \gg 1$, that is, when $m \gg \frac{1}{N_e}$, effect of migration is much larger than random genetic drift. Then, $\hat{H} \to H_0$  
This is the same as the case of no population differentiation.

When both mutations and migrations are assumed (Panel 33), $C_t$ retains the recurrent formula (3.31) and equilibrium (3.33). But, $F_t$ changes. The probability that a gene is not either migrant or mutant is

$$(1 - m)(1 - u) = 1 - (m + u) + mu \approx 1 - (m + u) \quad \text{ (approximation by } mu \ll m + u \ll 1)$$

Thus, the recurrent formula of $F_t$ is
\[ F_t = \left[ \frac{1}{2N_e} + (1 - \frac{1}{2N_e}) F_{t-1} \right] [1 - (m + u)]^2 \] (3.36)

At equilibrium,
\[ \hat{F} \approx \frac{1}{4N_e(m + u) + 1} \] (3.37)

Panel 33

In this case, the heterozygosity at generation \( t \), \( H_t \), is given as in equation (3.26).

\[ H_t = H_0 (1 - C_t) + C_t - F_t \]

At equilibrium, heterozygosity \( \hat{H} \) is given based on equation (3.33) and (3.37) as
\[ \hat{H} = \frac{4N_e m}{4N_e m + 1} H_0 + \frac{1}{4N_e m + 1} - \frac{1}{4N_e (m + u) + 1} \] (3.38)

However, generally, \( m \gg u \) holds and \( u \) can be negligible to \( m \). Practically,
\[ \hat{F} \approx \frac{1}{4N_e m + 1} \text{ and thus, } \hat{H} = \frac{4N_e m}{4N_e m + 1} H_0 \]

When \( N_e m \ll 1 \), that is, when \( m \ll \frac{1}{N_e} \), it is equivalent with the case of only mutations (see page 40).

When \( N_e m \gg 1 \), that is, when \( m \gg \frac{1}{N_e} \), as explained above, \( \hat{H} \rightarrow H_0 \). However, if effect of mutation is large and \( N_e u \gg 1 \), the heterozygosity of total population changes as \( H_0 \rightarrow 1 \). Then, \( \hat{H} \rightarrow 1 \).
3.8.3. Estimation of migration rate

As explained in pages 31-32, when populations diverged as in Panel 20: Figure 12.1A, $F_{ST}$, defined by $F_{ST} = \frac{H_T - \bar{H}_S}{H_T}$, is equivalent with the fixation index $F_r$. The island model assumes such pattern of population differentiation. Thus, at equilibrium between migration and random genetic drift, $F_{ST} = \hat{F}$ holds. Then, from the equation (3.34), $F_{ST} = -\frac{1}{4N_e m + 1}$ holds. From this, migration rate is given as

$$N_e m = \frac{1}{4} \left( \frac{1}{F_{ST}} - 1 \right)$$

(3.39)

$N_e m$ indicates the number of immigrant individuals at one generation. Hence, if we measure allele frequency from several populations and calculate $F_{ST}$, we can estimate $N_e m$. A number of assumptions are involved, but it offers a useful insight on migration rate.

3.8.4. Case of non-autosomal genes

The mitochondrial DNA is maternally inherited and is haploid. The coalescence probability of two randomly chosen DNA sequences in one generation is not $\frac{1}{2N_e}$ but is $\frac{1}{N_f}$ ($N_f$ is the effective female number).

Thus, in equations (3.32) and (3.34), by substituting $2N_e$ with $N_f$, and because of $F_{ST} = \hat{F}$, we have

$$F_{ST} = \frac{1}{2N_f m + 1}$$

then, following the equation (3.39), female migration rate is estimated as

$$N_f m = \frac{1}{2} \left( \frac{1}{F_{ST}} - 1 \right)$$

Likewise, in the case of Y-chromosomal genes, with the effective male number $N_m$, male migration rate is estimated as

$$N_m m = \frac{1}{2} \left( \frac{1}{F_{ST}} - 1 \right)$$

3.9. Cautions in handling real data

In calculations using real data, we need to take into consideration the number of populations, difference of the number of samples among populations, unbiased estimator compensation caused by sampling, etc. For details, refer to textbooks such as Nei and Kumar (2000) and Nei (1987).
Chapter 4 Population Genetics III (Nucleotide Sequence Data)

4.1. Genetic diversity at nucleotide level

Suppose we have two nucleotide sequences which differ at one site or at ten sites. If we regard a sequence as just a chunk, the two are only “different” irrespective of nucleotide difference level. Thus, the number of alleles and heterozygosity are not sufficient to describe the genetic variation at nucleotide sequence level.

4.1.1. The number of segregating sites and nucleotide diversity

Representative summary statistics to describe genetic variation at nucleotide sequence level are the number of segregating sites ($S$), or its quotient divided by the sequence length ($s$), and the nucleotide diversity ($\Pi$), or its quotient divided by the sequence length ($\pi$).

For example, suppose we have eight sequences of 90 nucleotides sampled from a population.

<table>
<thead>
<tr>
<th>Segregating site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>ggtactccgTttgctcagcaTaaacttgcccccaAtggactggaCtgtatggggagaactCgcacataatTagatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>ggtactccgTttgctcagcaTaaacttgcccccaAtggactggaCtgtatggggagaactCgcacataatTagatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>ggtactccgTttgctcagcaTaaacttgcccccaAtggactggaCtgtatggggagaactCgcacataatTagatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>ggtactccgTttgctcagcaCaaacttgcccccaTttggactggaActgatggggagaactCgcacataatCcgatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>ggtactccgTttgctcagcaCaaacttgcccccaTttggactggaActgatggggagaactCgcacataatCcgatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>ggtactccgCttgctcagcaCaaacttgcccccaCttggactggaCctgatggggagaactCgcacataatCcgatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>ggtactccgCttgctcagcaCaaacttgcccccaCttggactggaCctgatggggagaactCgcacataatCcgatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIII</td>
<td>ggtactccgCttgctcagcaCaaacttgcccccaCttggactggaCctgatggggagaactTggacataatCcgatctacgaGtcatcagctc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

configuration: $(3,5)\quad (3,5)\quad (2,3,3)\quad (2,6)\quad (1,7)\quad (3,5)\quad (1,7)$

The number of segregating sites is the number of polymorphic sites (the number of single nucleotide polymorphism: SNP, in more recent terminology). Here, $S = 7$, $s = 7 \div 90 = 0.0778$.

The nucleotide diversity is the average of pairwise nucleotide differences, $d_{ij}$, among all sequence pairs, $i$ and $j$, i.e. the mean pairwise difference $\Pi = \frac{1}{n} \sum_{i<j} d_{ij}$. Here, we have eight sequences and the number of pairwise combination is $sC_2 = \frac{8 \times 7}{2} = 28$. Each $d_{ij}$ is

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>III</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>IV</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Thus, $\Pi = \frac{5 \times 6 + 4 \times 13 + 3 \times 2 + 2 \times 1 + 1 \times 2}{28} = \frac{92}{28} = 3.3, \quad \pi = 3.3 \div 90 = 0.037$
The nucleotide diversity has another important aspect (Panel 34). Let's look at the sequences site-by-site. Let the number of different nucleotide kinds be $n_k$ and $n_l$. Then, the number of different nucleotide pair is $\sum_{k<l} n_k n_l$ at this site. For example, at the segregating site 1 in the previous page, we have five Ts and three Cs, and the number of different nucleotide pair is $3 \times 5 = 15$. Its quotient, $h_j = \frac{\sum_{k<l} n_k n_l}{\binom{n}{2}}$, which is divided by the number of pairwise combinations at the site, $\binom{n}{2}$ (here, $\binom{8}{2} = 28$), is the fraction of different nucleotide combinations among all combinations at this site. Summing up the fraction throughout the sites gives the mean pairwise differences, that is, $\Pi = \sum_{i=1}^{S} h_i$. Here,

$$\Pi = \frac{(3 \times 5) + (3 \times 5) + (2 \times 3 + 2 \times 3 + 3 \times 3) + (2 \times 6) + (1 \times 7) + (3 \times 5) + (1 \times 7)}{28} = \frac{92}{28}, \text{ the same as above.}$$

Panel 34

Let's transform $h_j = \frac{\sum_{k<l} n_k n_l}{\binom{n}{2}}$ further to understand it better.

$$h_j = \frac{\sum_{k<l} n_k n_l}{\binom{n}{2}} = \frac{1}{2} \sum_{k<l} 2n_k n_l = \frac{1}{2} \sum_{k<l} 2\left(\frac{n_k}{n} \times \frac{n_l}{n}\right)n^2 = \frac{n}{n-1} \sum_{k<l} 2\left(\frac{n_k}{n} \times \frac{n_l}{n}\right)$$
Here, $\sum_{k<l} 2\left(\frac{n_k}{n} \times \frac{n_l}{n}\right)$ is heterozygosity at a site, and $\frac{n}{n-1}$ corresponds to the unbiased estimator coefficient for a population from samples in statistics. Hence, $h_i$, at a site is the estimated heterozygosity at the site in the population. Therefore, summation of $h_i$ through all sites, $\Pi_i$, is the sum of estimated heterozygosity among sites in a population, and its quotient divided by sequence length, $\pi$, is the mean of estimated heterozygosity among sites in a population.

$$\Pi = \frac{\sum d_{ij}}{n \binom{c}{2}}$$
can also be related to heterozygosity at sequence level (Panel 35).

$$\Pi = \frac{\sum d_{ij}}{n \binom{c}{2}} = \frac{1}{2} \sum_{i<j} 2\left(\frac{1}{n} \times \frac{1}{n}\right) n^2 d_{ij} = \frac{n}{n-1} \sum_{k<l} 2\left(\frac{n_k}{n} \times \frac{n_l}{n}\right) d_{kl}$$

Panel 35

The frequency of a different sequence-type pair (heterozygosity) is multiplied by its nucleotide difference. Then, $\Pi$ is its estimated sum among all pairs in a population.

The nucleotide configuration at each site is expressed as $(n_k, n_l)$, where the numbers are arrayed from small to large (see the example in page 48). Heterozygosity at a site is high if its configuration gives a large
\[ \sum_{k<l} n_k n_l \]. Under the same number of segregating sites, a sequence set has a large \( \Pi \) if it contains many sites with high heterozygosity configuration, and has a small \( \Pi \) if it contains many sites with low heterozygosity configuration [the lowest configuration is \((1,n-1)\), called singleton].

4.1.2. Expectation of nucleotide diversity

\( \Pi \) is also an expected nucleotide difference between two randomly chosen sequences from a population. The expected coalescence time of two samples, \( \bar{t}_2 \), is \( 2N_e \) generations as shown in equation (3.20) (page 36). Thus, total time to consider mutations, \( \bar{T}_2 \), is \( \bar{T}_2 = 2 \times \bar{t}_2 = 4N_e \) generations. The expected number of mutations accumulated to the two sequences during this period is \( \Pi \) (Panel 36).

Panel 36

**Expectation of nucleotide diversity**

\[
E(\Pi) = \bar{T}_2 u = 4N_e u = \Theta
\]

The number of nucleotide site is almost infinite and the mutation rate per nucleotide site is very small. Thus, in the time scale of population differentiation or within-species genetic variation, every mutation can be assumed to occur at different site. This is called **infinite site model**. Also, the probability of having two or more mutations in the same DNA stretch at the same time is negligible. Hence, every mutation is assumed to create a new sequence (allele) (corresponding to the infinite allele model explained in page 40). In this situation, with the mutation rate \( u \) per generation per sequence, the expected number of mutations accumulated in the total time \( \bar{T}_2 \) is \( 4N_e u \) (Panel 36). Therefore, expectation of \( \Pi \) is given as

\[
E(\Pi) = 4N_e u = \Theta
\]

(4.1)

Let the mutation rate per generation per nucleotide be \( \mu \), the expectation of \( \pi \) is

\[
E(\pi) = 4N_e \mu = \Theta
\]

(4.2)

As mentioned in page 40, \( 4N_e u \) is a key parameter which ties together nucleotide-level diversity and gene-level diversity through equation (4.1) and equations (3.27) and (3.28), and is conventionally symbolized as \( \Theta \). \( 4N_e \mu \) is symbolized as \( \Theta \). Let the length of sequence be \( L \), and \( \Theta = L \Theta \).
4.13. Expectation of the number of segregating sites

Under the infinite site model, the number of segregating sites for \( n \) samples is equivalent with the number of mutations which occurred during the total coalescence time among the \( n \) samples. As in equation (3.21) of page 37, the expected waiting time for \( n \) samples to coalesce into \( n-1 \), \( \bar{t}_n \), is \( \frac{4N_e}{n(n-1)} \) generations (Panel 25: Figure 3.15). In page 37, the expected time, \( T_{\text{MRCA}} \), for \( n \) samples to eventually coalesce into 1 is given as \( \sum_{i=2}^{n} \bar{t}_n \). Here, we consider the total time, \( T \), the sum of all branch length in the coalescence pedigree (Panel 37). First, we consider the total time at each coalescence step. Let the total time for \( i \) genes to coalesce into \( i-1 \) be \( \bar{t}_i \). Then, \( \bar{t}_i = \bar{t}_{i-1} + 4N_e \frac{i}{i-1} \). Second, we sum up \( \bar{t}_i \) throughout to obtain the grand total time \( T \).

\[
T = \sum_{i=2}^{n} \bar{t}_i = 4N_e \sum_{i=2}^{n} \frac{1}{i-1} = 4N_e \sum_{i=1}^{n-1} \frac{1}{i}
\]

(4.3)

**Expectation of the number of segregating sites**

Panel 37

\[
E(S) = T \mu = 4N_e \mu \sum_{i=1}^{n-1} \frac{1}{i}
\]

Hence, the expectation of total number of mutations in this pedigree is obtained by multiplying the equation (4.3) and the mutation rate per generation per gene \( \mu \). This is the expectation of the number of segregating sites, \( S \). That is

\[
E(S) = uT = 4N_e \mu \sum_{i=1}^{n-1} \frac{1}{i} = \Theta \sum_{i=1}^{n-1} \frac{1}{i}
\]

(4.4)

The expectation of the proportion of segregating sites, \( s \), is given with the mutation rate per generation per nucleotide site, \( \mu \), as

\[
E(s) = \mu T = 4N_e \mu \sum_{i=1}^{n-1} \frac{1}{i} = \Theta \sum_{i=1}^{n-1} \frac{1}{i}
\]

(4.5)

52
4.2. Likelihood function of the number of segregating sites

We have seen that expected values of the nucleotide diversity and the number of segregating sites are governed by $\Theta$. In other words, the nucleotide diversity and the number of segregating sites are given stochastically under the parameter of $\Theta$. Or, these are given by likelihood functions with parameter $\Theta$. In this section, we consider the probability for each possible value of the number of segregating sites. By doing so, beyond simply knowing the expected value, we have an insight on what value is most likely and how more likely it is than other values.

In Chapter 3 page 35, we reached the probability function, $P(t)$, of the waiting time $t$ for the first coalescence to occur in $n$ samples (practically the probability of $t$ that $n$ samples take to coalesce into $n-1$). As in page 35, $P(t)$ is given as

$$P(t) = C_1(1-C_1)^{t-1}.$$ We rewrite it into a more manageable form. Because of $C_1 \ll 1$, an approximation $1-C_1 \approx e^{-C_1}$ holds (Panel 13). Thus, a handier exponential function is obtained.

$$P(t) = C_1 e^{-C_1(t-1)} = C_1 e^{C_1} e^{-C_1 t} \approx C_1 e^{-C_1 t} \quad \text{(because of } e^{C_1} \approx 1)$$

When the number of samples is two ($n = 2$)

Let's begin with the simplest case, $n = 2$ (Panel 38). As seen in page 36, when $n = 2$,

$$C_1[n = 2] = \frac{1}{2N_e}, \text{ thus, }$$

$$P(t_2) = \frac{1}{2N_e} e^{-\frac{1}{2N_e} t_2}$$

This kind of formula is called an exponential distribution with parameter $\frac{1}{2N_e}$, or an exponential distribution with expectation $2N_e$.

Then, we consider the probability distribution of the number of mutations during the waiting time $t_2$. The total time to consider mutation is the sum of two passes, from common ancestor to each gene, and is $2t_2$.

For simplification to avoid 2, we consider the total time $T_2$. The expectation of $T_2$ is $4N_e$. By an extension of the equation above, $P(T_2)$ is an exponential distribution with expectation $4N_e$.

$$P(T_2) = \frac{1}{4N_e} e^{-\frac{1}{4N_e} T_2} \quad (4.6)$$

The number of segregating site is equivalent with the number of mutations during this $T_2$ generations. Next, we consider the probability function of the number of mutations, $S_2$, during the time $T_2$. 

53
Panel 38

**Probability function of total time for 2 samples to coalesce**

\[
P(t_2) = \frac{1}{2N_e} e^{-\frac{1}{2N_e} t_2}
\]

\[
T_2 = 2t_2
\]

\[
P(T_2) = \frac{1}{4N_e} e^{-\frac{1}{4N_e} T_2}
\]

S2 mutations occur somewhere in T2 generations. The mutation rate u per gene per generation is small enough. We can assume that two or more mutations do not occur in one generation. Thus, we consider the probability that mutations occur in S2 generations out of T2 generations (i.e. a conditional probability of S2 under T2). This is a binominal distribution where a mutation of probability u occurs in S2 generations chosen from T2 generations.

\[
P(S_2|T_2) = C_{S_2} T_2^{S_2} (1-u)^{T_2-S_2}
\]

Because of u « 1 and T_2 » 1, \( P(S_2|T_2) \) can be regarded as a Poisson distribution with expectation of uT_2.

\[
P(S_2|T_2) = \frac{(uT_2)^{S_2} e^{-uT_2}}{S_2!}
\]

(4.7)

Hence, \( P(S) \), the likelihood function of S2, is given as

\[
P(S_2) = \int_0^\infty P(S_2|T_2)P(T_2)dT_2 = \int_0^\infty \frac{(uT_2)^{S_2} e^{-uT_2}}{S_2!} \frac{1}{4N_e} e^{-\frac{1}{4N_e} T_2} dT_2
\]

By solving this, we obtain

\[
P(S_2) = \frac{1}{1+\Theta} \left( \frac{\Theta}{1+\Theta} \right)^{S_2} \quad (\Theta = 4N_e u)
\]

(4.8) [Watterson (1975) Theor. Popul. Biol. 7:256-276]

This kind of formula is called a geometric distribution with parameter \( \frac{1}{1+\Theta} \). In more precisely, the equation (4.8) is the conditional probability function of S2 under \( \Theta \), and can be written as

\[
P(S_2|\Theta) = \frac{1}{1+\Theta} \left( \frac{\Theta}{1+\Theta} \right)^{S_2}
\]

(4.9)

The expectation of this probability distribution is \( \Theta \), the same as the expectation of \( \Pi \). Only in the case of \( n = 2 \), the equation (4.8) is the probability function of \( \Pi \) as well. Note that \( \Pi \) is the mean of pairwise nucleotide differences among \( n \) samples. When \( n \) is three or more, the probability function of \( \Pi \) differs from the
equation (4.8). The probability distribution of $\Pi$ is the distribution of the mean of multiple $S_2$ values sampled under the equation (4.8). Hence, by the central limit theorem, when $n$ is large enough, distribution of $\Pi$ approaches to the normal distribution with expectation $\Theta$.

When $S_2 = 0$, that is, when two sequences are identical, its probability is $\frac{1}{1 + \Theta}$ according to the equation (4.8). Note that this corresponds to $\hat{F}$ in the equation (3.25) in page 40, which is the probability that two randomly chosen genes are identical by descent, escaping mutations throughout. As mentioned in page 40, in humans, $\Theta$ of $\sim 1$ kb gene is about 1. Thus, the probability is about $1/2$ that two randomly chosen genes from human population are identical. $\Theta$ of human genome is about three million and the probability that two genomes are identical is almost 0.

When the number of samples is three or more ($n \geq 3$)

Panel 39

Probability function of total time at each coalesce

\[
\begin{align*}
T_2 &= 2t_2 & P(T_2) &= \frac{1}{4N_e} e^{-\frac{1}{4N_e}T_2} \\
T_3 &= 3t_3 & P(T_3) &= \frac{2}{4N_e} e^{-\frac{2}{4N_e}T_3} \\
T_4 &= 4t_4 & P(T_4) &= \frac{3}{4N_e} e^{-\frac{3}{4N_e}T_4} \\
& \vdots & & \vdots \\
T_i &= it_i & P(T_i) &= \frac{i-1}{4N_e} e^{-\frac{i-1}{4N_e}T_i}
\end{align*}
\]

When $n \geq 3$, as explained in page 37, the probability that $i$ genes coalesce into $i-1$ in one generation is $C_i(i) = \frac{i(i-1)}{4N_e}$. Then, the probability distribution $P(t_i)$ of waiting time $t_i$ that $i$ samples take to coalesce into $i-1$ follows an exponential distribution with expectation $\frac{4N_e}{i(i-1)}$. Thus,

\[
P(t_i) = \frac{i(i-1)}{4N_e} e^{-\frac{i(i-1)}{4N_e}}
\]

The total time $T_i$ to consider mutations is $it_i$ generations. As in the case of $n = 2$, the probability function
\( P(T_i) \) is an exponential distribution with expectation \( \frac{4N_e}{i-1} \). Thus,

\[
P(T_n) = \frac{i-1}{4N_e} e^{-\frac{i-1}{4N_e}T_i}
\] (Panel 39)

As in the case of \( n = 2 \), the probability distribution \( P(S_i | T_i) \) of the number of mutations \( S_i \) during the time \( T_i \) follows the Poisson distribution with expectation \( uT_i \). Then,

\[
P(S_i | T_i) = \frac{(uT_i)^S e^{-uT_i}}{S_i!}
\]

Thus, the probability distribution \( P(S_i) \) is

\[
P(S_i) = \int_0^\infty P(S_i | T_i) P(T_i) dT_i = \int_0^\infty \frac{(uT_i)^S e^{-uT_i}}{S_i!} \frac{i-1}{4N_e} e^{-\frac{i-1}{4N_e}T_i} dT_i
\]

By solving this, we obtain

\[
P(S_i) = \left\{ \frac{1}{1 + \frac{\Theta}{i-1}} \right\}^S \left\{ 1 - \frac{1}{1 + \frac{\Theta}{i-1}} \right\}^S_i
\] (4.10)


If manifesting it to be conditional to \( \Theta \),

\[
P(S_i | \Theta) = \left\{ \frac{1}{1 + \frac{\Theta}{i-1}} \right\}^S \left\{ 1 - \frac{1}{1 + \frac{\Theta}{i-1}} \right\}^S_i
\] (4.11)

The number of segregating sites, \( S \), corresponds to \( \sum_{i=2}^{n} S_i \). \( P(S | \Theta) \) is the synthesis of equation (4.11) throughout all coalescence steps \( i = 2 \sim n \) and is too complex to show here.

Panel 40 illustrates equations (4.9) and (4.11) by assigning values in them to visualize how the probability distributions of \( S_i \) look like. As an example, the case of \( \Theta = 5 \) is shown. This corresponds to a \( \sim 5 \) kb gene in humans. When the samples size \( i \) is smaller, the distribution of \( S \) becomes flatter. This trend is clearer with larger \( \Theta \). In the case of \( i = 2 \), i.e. the source of \( \Pi \), the distribution is flattest with the larger variance. As the sample size \( i \) increases, the distribution of \( S_i \) becomes narrower. This infers that, as explained later, in estimation of \( \Theta \), \( S \) is more reliable than \( \Pi \).

### 4.3. Estimation of basic parameters from sequence data

We have seen that the number of segregating sites and nucleotide diversity are given stochastically under parameter \( \Theta \). But, how can we know \( \Theta \) which cannot be directly measured?
4.3.1. The method of moments estimators for $\Theta$

Based on the equations (4.1) and (4.2) or the equations (4.4) and (4.5), we can estimate $\Theta$ simply from the observed $\Pi$ (or $\pi$) or $S$ (or $s$). That is,

$$\Theta_\Pi = \Pi \quad \text{(or} \quad \theta_s = \pi) \quad (4.12)$$

$$\Theta_s = S / \sum_{i=1}^{n-1} i \quad \text{(or} \quad \theta_s = s / \sum_{i=1}^{n-1} i) \quad (4.13)$$

These estimators correspond to the method of moments estimators. The relation of $\Pi$ to $\Theta$ was first noted by Tajima (1983 Genetics 105:437-460) and the estimator $\Theta_\Pi$ is also symbolized as $\Theta_T$ ($\theta_s$ as $\theta_T$).

The relation of $S$ to $\Theta$ was first noted by Watterson (1975 Theor. Popul. Biol. 7:256-276) and the estimator $\Theta_s$ is also symbolized as $\Theta_W$ ($\theta_s$ as $\theta_W$).

Their variances are

$$V(\Theta_\Pi) = \frac{n+1}{3(n-1)} \Theta_\Pi + \frac{2(n^2 + n + 3)}{9n(n-1)} \Theta_\Pi^2$$

$$V(\Theta_s) = \frac{1}{\sum_{i=1}^{n-1} \frac{1}{i}} \Theta_s + \frac{\sum_{i=1}^{n-1} \frac{1}{i^2}}{\left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^2} \Theta_s^2$$

$$V(\theta_\pi) = \frac{n+1}{3(n-1)L} \theta_\pi + \frac{2(n^2 + n + 3)}{9n(n-1)} \theta_\pi^2$$

$$V(\theta_s) = \frac{1}{\left(\sum_{i=1}^{n-1} \frac{1}{i}\right) L} \theta_s + \frac{\sum_{i=1}^{n-1} \frac{1}{i^2}}{\left(\sum_{i=1}^{n-1} \frac{1}{i}\right)^2} \theta_s^2$$

where $L$ is the length of sequence.
When \( n = 2 \), the coefficients of all terms are 1. As \( n \) increases, all coefficients become smaller and the variance become smaller, thus, reliability of estimate increases. The decrease of variance is larger for \( S \) (and \( s \)) than for \( \Pi \) (and \( \pi \)). Hence, \( S \) (and \( s \)) is more reliable than \( \Pi \) (and \( \pi \)) for estimating \( \Theta \) as in page 56.

4.3.2. The maximum likelihood estimate of \( \Theta \)

The expectation and variance are not sufficient as an estimate of \( \Theta \). Utilizing the idea of coalescence and applying the likelihood functions [such as equation (4.9)] or simulations to nucleotide sequence data, important parameters (such as \( \Theta \) and coalescence waiting time) can be estimated with confidence (or credible) interval under the maximum likelihood principle. Let the likelihood function \( P(S|\Theta) \) of \( S \) be an example. We fix \( S \) to an observed value and we change \( \Theta \) value to search for a \( \Theta \) which maximizes \( P(S|\Theta) \). The maximizing value is regarded as the maximum likelihood estimate. Use of likelihood function is one method and simulation is another for maximum likelihood estimation.

4.3.2.1. <Use of likelihood function for estimation of \( \Theta \)>

Panel 41A is an exemplary dataset with sample size \( n = 20 \) and sequence length \( L = 1000 \) bp (only segregating sites shown). The number of segregating sites is \( S = 8 \) and the nucleotide diversity is \( \Pi = 2.9 \). Thus, the method of moments estimates are \( \theta_s = 2.3 \times 10^{-3} \) and \( \theta_\pi = 2.9 \times 10^{-3} \). The panel 41B solid line illustrates the likelihood function \( P(S = 8|\Theta) \) with \( n = 20 \) derived based on the idea explained in pages 53-56. The maximum likelihood estimate coincides with the moments estimate from \( S \), \( 2.3 \times 10^{-3} \), which maximizes the likelihood function. The vertical axis is scaled by relative log likelihood with maximum being 0. The twofold of the difference of maximum log-likelihood from log-likelihood follows \( \chi^2 \) distribution. So, the difference within about 1.9 corresponds to about 95% confidence interval (Panel 41B horizontal line). Hence, the 95% confident interval of \( \Theta \) estimate is broad, ca. \( 1 \sim 6 \times 10^{-3} \). Once every 20 datasets the observation could go out of the range by chance alone. Therefore, the maximum likelihood estimation tells us that this dataset is not sufficient to reliably estimate the parameter.

4.3.2.2. < Use of simulation for estimation of \( \Theta \)>

The likelihood function of \( S \) is relatively simple when \( n = 2 \) [ \( P(S_2|\Theta) \) ]. But the function becomes complex for \( S \) with \( n \geq 3 \). The function of \( \Pi \) would be improbable to derive for non-experts. There are also other summary statistics to consider. To be realistic, we need to consider many conditions, such as migration, population differentiation, and population-size fluctuation. It is then not practical to devise likelihood functions for all of these. By using simulations based on the coalescence idea, we can circumvent the
difficulty and reach maximum likelihood estimates.

Panel 41

Panel 42 shows an outline of a simple case of simulation without population-size fluctuation, population differentiation, or migration.

(1) Suppose we have a dataset of \( n \) samples of a gene taken from a population. Let computer determine an order of coalescence among the \( n \) samples so that two genes coalesce to one at a time and \( n \) genes eventually coalesce to one ancestral gene (probabilities of coalescence of three or more genes at a time and of two or more coalescence events at a time are negligible as explained in page 37).

(2) Give a length (i.e. coalescence waiting time) at random to each interval between adjacent nodes under the probability function 

\[
P(t_i) = \frac{i(i-1)}{4N_e} e^{-\frac{i(i-1)}{4N_e}} \quad (i = 2, 3, \ldots, n).
\]

(See page 55)

(3) Calculate the total branch length of the tree (grand total time of coalescence) 

\[
T = \sum_{i=2}^{n} it_i.
\]

(4) Determine the number of mutations during \( T \) at random under the probability function 

\[
P(S|T) = \frac{(uT)^S e^{-uT}}{S!}.
\]

(See page 54)

(5) Assign these mutations to the branches at random proportionally to their length.
(6) Obtain a sequence dataset and calculate summary statistics of genetic variation (such as \( S \) and \( \Pi \)).

Panel 42

(1) Determine an order of coalescence among \( n \) samples

(2) Give a time length \( t_i \) to each coalescence

\[
P(t_i) = \frac{t_i(1-1)}{4N_e} e^{-t_i/2N_e}
\]

(3) Calculate the total branch length \( T \)

\[
T = \sum_{i=2}^{n} t_i
\]

(4) Determine the number of mutations \( S \)

\[
P(S|T) = \frac{(\mu T)^S e^{-\mu T}}{S!}
\]

(5) Assign the mutations to the branches proportionally

(6) Obtain a sequence dataset and calculate \( S, \Pi, \) etc.

In Panel 41B, small white boxes (□) show an example of such simulation result, with sample size \( n = 20 \) and sequence length \( L = 1000 \) bp. A value of \( \Theta \) is given (as a population size \( N_e \) and a mutation rate \( \mu \)) and simulation is repeated many times (200,000 times in Pane 41). Likelihood is defined as the proportion of trials where resulted \( S \) value coincided with the observation \( S = 8 \). Changing \( \Theta \) value by every \( 10^{-4} \), the \( \Theta \) value giving the highest proportion is defined as the maximum likelihood estimate. The simulated likelihood curve nicely fits the theoretically derived curve by likelihood function. This means that without knowing likelihood function we can reach the maximum likelihood estimate of \( \Theta \) (or \( \theta \)) by conducting a simulation. We only need to understand the coalescence simulation process (coalescence order, exponential distribution of waiting time, Poisson distribution of mutation number).

When nucleotide diversity \( \Pi \) is used to estimate \( \Theta \), the same procedure is followed. However, while \( S \) takes only nonnegative integers, \( \Pi \) has often a decimal fraction (the exact \( \Pi \) value in Panel 41 is 2.947368421\cdots). It is extremely improbable that a simulation gives the exact value of \( \Pi \). Then, the approximate likelihood method is used, where values within a certain range of deviation is accepted. The permissible range of deviation is denoted \( \delta \). \( \delta = 1\% \) is regarded strict and close to the true likelihood. As \( \delta \) increase, computation
time becomes shorter but reliability of estimation decreases. In Panel 41C, $\delta = 10\%$ fits as well as $\delta = 1\%$ and can be regarded accurate, but $\delta = 100\%$ is largely deviated from them. In this example, $\delta = 10\%$ was good but in other datasets, accuracy is not guaranteed.

In Panel 41C, the $\Pi$-based maximum likelihood estimate of $\Theta$ is $3.2 \times 10^{-3}$ and is slightly different from the method of moments estimate $2.9 \times 10^{-3}$. The 95% confidence interval is about $1 \sim 9 \times 10^{-3}$ and is broader than that in the $S$-based maximum likelihood estimation. As explained in page 58, estimation of $\Theta$ (or $\theta$) based on $\Pi$ has a larger variance than based on $S$, and is less reliable.

4.3.2.2.1. Rejection sampling method

Let me introduce a simpler method of maximum likelihood estimation, called rejection sampling, of $\Theta$ by coalescence simulation. When we have some information on probability distribution beforehand on a parameters of interest (prior probability distribution), this method is used. The prior distribution can be simply upper and lower limits, or more desirably some probability function. First, we choose one $\Theta$ value at random under the prior distribution. We run the simulation process (1) ~ (6) only once with the $\Theta$ value. If the resulted summary statistic of interest (e.g. $S$, $\Pi$) coincides with the observed one or within a certain range of deviation from it, we accept the $\Theta$ value. If not, we reject it. We repeat this process until the number of accepted values reaches to a certain amount (e.g. 10,000). This random sampling allows resampling the same value. Hence, the most often accepted value is defined as the maximum likelihood estimate. A credible interval (e.g. of 95%) is obtained from the location in its frequency distribution (terminology "credible" is used instead of "confidence" because Bayesian estimation is used).

4.3.2.3. Improvement of estimation accuracy of $\Theta$

As shown in Panel 41, sadly, a dataset with 20 sequences of 1000 bp is not sufficient to reliably estimate $\Theta$. How can we improve accuracy of estimation? Panel 43 illustrates the effect of sample size, sequence length, and the number of gene locus to reduce the variance of $\Theta$ estimation by coalescence simulation. Apparently, the number of study locus is most effective.

The rationale of the locus number effect is illustrated in Panel 44. Assume that we estimate $\Theta$ distribution based on, say, $S$ for each of $l$ loci. We then sample one $\Theta$ value from each locus and take average of them. We repeat this process many times. As in statistics textbook, the variance of mean is equivalent with the sum of variance among loci divided by $l^2$. If the variance is regarded same among loci, variance of $\Theta$ mean is one $l$th of one locus. Thus, if we study 10 loci and estimate $\Theta$ for each locus from $S$, taking average and using the average as the estimate of $\Theta$, the variance of it is reduced to 1/10 of a single gene. If we study 100 loci, the variance is 1/100. Reliability increases remarkably.
In practice (Panel 45), the following procedure is more efficient. An observed summary statistic (such as $S$) of one locus can be regarded as a result of one trial performed by nature under an unknown $\Theta$. When we study $l$ loci and examine their $S$ values, regarding them governed by a probability under the same $\Theta$, we can equate them with the results of nature's multiple trials. The variance of mean among them is reduced to $1/l$ of the variance in one trial. In a rejection resampling procedure, the effect of locus number can be incorporated by repeating the (1) – (6) process of pages 59-60 $l$ times for each $\Theta$ value and using the mean of a resulted summary statistic (such as $S$). The variance of the mean is reduced to $1/l$ of the variance in one locus. If the simulation mean by a $\Theta$ value is equal to or within a certain deviation from the observed mean in the $l$ loci, we accept the $\Theta$ value. By doing so, we can significantly improve the estimation of $\Theta$.
The mitochondrial DNA has often been used for studying population history of various organisms including humans. However, even if sequencing it in longer range, mitochondrial DNA is merely one locus due to lack of recombination in general. Argument based solely on it is rather dangerous. Microsatellite study is generally conducted with multiple loci and is far more meaningful. The whole-genome diversity study is getting feasible and would be the ultimate method for clarifying population history.

4.3.2.4. <Estimation of effective population size>

$\Theta$ (or $\theta$) is defined as $\Theta = 4N_e \mu$ ($\theta = 4N_e \mu$) and is comprised of the effective population size and the mutation rate. The mutation rate can be estimated from divergence time between species inferred from relevant fossil or paleogeographical information and their nucleotide divergence if the divergence is selectively neutral (neutral evolutionary rate = mutation rate, explained in Chapter 5). Recently, mutation rate can be directly measured by whole-genome sequencing of parents and an offspring. Thus, using these mutation rate information and the estimated $\theta$ value from genetic variation data, effective population size $N_e$ can be estimated.

In humans, nucleotide diversity $\pi$ is observed to be about 0.1%. If based on it, the method of moments estimate of $\theta = 4N_e \mu$ in humans is 0.1%. As mentioned in page 40, mutation rate per nucleotide site/year in humans is estimated to be about $1 \times 10^{-9}$. Assuming the generation time is 25 years, because $1 \times 10^{-9} = 4 \times N_e \times 1 \times 10^{-9} \times 25$, we reach $N_e = 1 \times 10^4$. Human effective population size is thus estimated to be roughly only
10,000. This is much smaller than the current census size of about 700 million, implying that humans experienced a severe bottleneck or our population had been historically very small (refer to page 33).

4.3.2.5. <Estimation of coalescence waiting time \(T_{MRCA}\) between two sequences>

In Chapter 3, we learned how to reach the expectation of coalescence waiting time. Two sequences sampled from a population can differ at 0, 1, 2, ... nucleotide sites. The expectation, \(\tilde{t}_2 = 2N_e\), of coalescence waiting time of two sequences is, in fact, a mean of these various cases. From equations (4.6), (4.7) and (4.8) in pages 53-54, we now know the likelihood functions \(P(T_2)\), \(P(S_2|T_2)\) and \(P(S_2)\). Using the Bayes theorem (refer to panel 46) and these functions and, we can reach the conditional probability function, \(P(T_2|S_2)\), of the total coalescence time \(T_2\) under \(S_2\), i.e. the likelihood function of \(T_2\) when data \(S_2\) is given.

\[
P(T_2|S_2) = \frac{P(S_2|T_2)P(T_2)}{P(S_2)} = \frac{T_2^S_2}{S_2!} \left( \frac{1 + \Theta}{4N_e} \right)^{S_2} e^{-\frac{1}{4N_e}T_2} \tag{4.14}
\]

\(T_{MRCA}\) is half of \(T_2\).

Panel 46

\[
P(T_2) = \frac{1}{4N_e} e^{-\frac{1}{4N_e}T_2}
\]

\[
P(S_2) = \frac{(uT_2)^{S_2}e^{-uT_2}}{S_2!} = \frac{1}{S_2!} \left( \frac{\Theta}{4N_e} T_2 \right)^{S_2} e^{-\frac{\Theta}{4N_e}T_2}
\]

\[
P(S_2|T_2) = \frac{1}{1 + \Theta} \left( \frac{\Theta}{1 + \Theta} \right)^{S_2}
\]

\[
P(T_2|S_2) = \frac{P(S_2|T_2)P(T_2)}{P(S_2)} = \frac{T_2^S_2}{S_2!} \left( \frac{1 + \Theta}{4N_e} \right)^{S_2+1} e^{-\frac{1}{4N_e}T_2}
\]

As an example, using the published genome sequences of a Japanese man (Fujimoto et al. 2010 Nat. Genet. 42:931-936), a Korean man (Kim et al. 2009 Nature 460:1011-1015), and an European man (Craig Venter) (Levy et al. 2007 PLoS Biol. 5:3254), let's derive the probability function of their Y-chromosome \(T_{MRCA}\) (Panel 47: courtesy of Dr. Shuhei Mano). We use the number of segregating sites (SNP) \(S_2\) and sequence length \(L\) from the genome data and the effective population size \(N_e\) and mutation rate \(\mu\) in humans as mentioned above. The Y chromosome is only in males and is haploid. In the equation (4.14), we replace \(4N_e\) with \(N_e\), and \(\Theta\) with \(N_e\mu\) \((=N_e\mu L)\). Then, the \(T_{MRCA}\) of the Japanese man's and Craig Venter's Y chromosomes is most likely
about 7,000 generations ago (175,000 years ago) and that of the Japanese and the Korean men's Y chromosomes is most likely about 2,000 generations ago (50,000 years ago). In Panel 47, "Prior" corresponds to the prior probability distribution equation (4.6), 

\[ P(T_2) = \frac{1}{N_e} e^{-\frac{T_2}{N_e}} \quad \text{\(T_{\text{MRCA}}\) is half of \(T_2\)}, \]

and is the probability distribution of coalescence waiting time when no information is given. By the sequence information given, the probability distribution is upgraded to the posterior probability with very sharp and high peak.

When two sequences are identical (i.e. \(S_2 = 0\)), the likelihood function (4.14) becomes

\[ P(T_2 | S_2 = 0) = \frac{1 + \Theta}{4N_e} e^{\frac{1 + \Theta}{4N_e} T_2} \quad \text{(4.15)} \]

that is, an exponential distribution with expectation \(\frac{4N_e}{1 + \Theta}\).

If two 1-kb sequences are identical, which are randomly chosen from human population, and given that \(\Theta\) of 1-kb gene in human is supposed to be 1, the expectation of their \(T_{\text{MRCA}}\) is \(N_e\) generations ago, that is, about 10,000 generations ago (250 thousand years ago). If 100 kb is identical, the expectation of their \(T_{\text{MRCA}}\) is about 5000 years ago. If 20 Mb is identical, the expectation of their \(T_{\text{MRCA}}\) is 1 generation ago, that is, the two could be siblings. These are only about expectation and don't consider variance, with simple assumptions. However, we learn that divergence time can be estimated even between identical sequences by applying population genetic theorem, which otherwise would be regarded 0 by a simple molecular clock sense.
4.3.2.6. <Parameter estimation under complex demographic conditions>
Under realistic demographic conditions such as with population size fluctuation, population differentiation, and migration, we can reach the maximum likelihood estimate of various parameter combinations (such as changing effective population size, time of the size change, divergence time of populations, and migration rate between subpopulations) which realizes various summary statistics of sequence data (such as the number of segregating sites, nucleotide diversity, the number of singleton sites, the number of haplotypes ("haplotype" in page 79), haplotype heterozygosity, and the most frequent haplotype frequency) (Panel 48).

Panel 48

In this complex case, the conceptual framework of coalescence still holds. The probability of coalescence of two genes in one generation $C_1$ is considered in three cases:
(A) A size-fluctuating population before population split
(B) Just after population split
(C) Two size-fluctuating populations after population split with migration each other

The two genes either
(1) belong to a same population, or
(2) belong to different populations each other

In any of these cases, for two sequences at generation $i$, $C_1$ concerns only population size one generation ago (at generation $i + 1$).
Let sample size be \( n \).

In the case (A), we only concern the situation (1). Then, \( C_1 \) in generation \( i \) is

\[
C_1 = \frac{n(n-1)}{4N_{A,i+1}},
\]

where \( N_{A,i+1} \) represents Ancestral effective population size in generation \( i+1 \).

Thus, taking the varying population size into consideration, the probability of coalescence at \( t \) generations ago is

\[
P(t) = \frac{n(n-1)}{4N_{A,t}} \prod_{i=1}^{t-1} \left[ 1 - \frac{n(n-1)}{4N_{A,i}} \right].
\]

In the case (B), we concern the size of population which is just before the split. Thus, we follow the case (A).

In the case (C), we concern both situations (1) and (2).

In the situation (1), we consider a case where the two genes belong to subpopulation 1 or a case where the two genes belong to subpopulation 2.

In the former case, both genes are non-immigrants which coalesce in subpopulation 1, or both genes are immigrants which coalesce in subpopulation 2. Thus,

\[
C_1 = (1 - m_{12,i})^2 \frac{n(n-1)}{4N_{I,i+1}} + m_{12,i}^2 \frac{n(n-1)}{4N_{2,i+1}},
\]

where \( m_{12,i} \) is migration rate in subpopulation 1 at generation \( i \) (probability that a gene is a migrant from subpopulation 2) and \( N_{I,i+1} \) is effective population size of subpopulation 1 at generation \( i+1 \), \( N_{2,i+1} \) is effective population size of subpopulation 2 at generation \( i+1 \).

In the latter case, similarly,

\[
C_1 = m_{21,i}^2 \frac{n(n-1)}{4N_{I,i+1}} + (1 - m_{21,i})^2 \frac{n(n-1)}{4N_{2,i+1}},
\]

where \( m_{21,i} \) is migration rate in subpopulation 2 at generation \( i \) (probability that a gene is a migrant from subpopulation 1).

In the situation (2), two genes coalesce in subpopulation 1 (one gene is an immigrant from 1 and another is native to subpopulation 1) or two genes coalesce in subpopulation 2 (one gene is an immigrant from 2 and another is native to subpopulation 2). Thus,

\[
C_1 = m_{21,i}(1 - m_{12,i}) \frac{n(n-1)}{4N_{I,i+1}} + m_{12,i}(1 - m_{21,i}) \frac{n(n-1)}{4N_{2,i+1}}.
\]

By following the case (A), \( P(t) \) is given.

We then incorporate probabilities into the steps (1) and (2) of page 59. We can incorporate the conditions when and how population size changes and when populations split into the simulation.

4.3.2.7. <Improvement of estimation accuracy of parameters under complex demographic conditions>

Under these complex models, we can improve parameter estimation by coalescence simulation. We examine
distribution of accepted parameter values which satisfy multiple summary statistics.

Panel 49 is an example where parameters considered are the present population size $N_p$, the timing of population reduction at $t_e$ generations ago (as a unit of $t_e$ times of $4N_p$ generations), ancestral population size $N_A$ which is expressed $k$ times of $N_p$. The summary statistics used are 6 kinds: the number of segregating site $S$, nucleotide diversity $\pi$, the number of singleton sites $s$, the number of haplotypes $h$, haplotype heterozygosity $H$, the most frequent haplotype frequency $F$. The initial setting is the sample number $n = 50$, the number of study gene locus $l = 50$, and sequence length 10 kb. Assumed parameter values are $N_p = 20,000$, $t_e = 0.1$, $k = 0.1$. Mutation rate is set as $\mu = 10^{-8}$. Then, 50 times simulations (because of 50 loci) are conducted to generate 50 virtual data sets. The mean of each summary statistics among 50 data sets is obtained.

Then, we look for parameter values that satisfy these summary statistics by coalescence simulation. By comparing the frequency distribution (posterior distribution) of the accepted parameter values with the initially set values, we evaluate the estimation accuracy.

Under the strictest condition requiring all six summary statistics to be satisfied (Panel 49B), the peak is high and narrow and well coincides with the assumed value in all three parameters, which means the estimation is very accurate. But calculation time is very long. In contrast, under the loosest conditions requiring only one parameter to be satisfied, calculation time is very short but accuracy is too bad. Then, in between, considering two or three summary statistics to be satisfied, some combinations are found accurate enough with less calculation time (Panel 49B). We can also examine the effect of increasing sample number and locus number (Panel 49C).

In summary, coalescence simulation enables us, in various complex demographic conditions, to estimate fundamental parameters such as changing population size, timing of population size change, timing of population split, and migration rate, with relatively a simple principle.

Panel 49

4.4. Tajima’s $D$: Relation between the number of segregating sites and nucleotide diversity

Under the equilibrium between random genetic drift and mutation, the number of alleles and the heterozygosity are correlated (pages 40-43). Likewise, the number of segregating sites and the nucleotide diversity are correlated. Tajima’s $D$ explains this relationship. A DNA segment with a new mutation would receive no more mutation in a short while before transmitted to the next generation. Thus, when the number of segregating site increases by one, the number of alleles increases by one: i.e. the number of segregating sites reflects the number of alleles. As explained earlier, the nucleotide diversity reflects the heterozygosity.
From equations (4.12) and (4.13), relation of the number of segregating sites and the nucleotide diversity is

\[ S / \sum_{i=1}^{n-1} \frac{1}{i} = \Pi \]  

(4.16)

\[ s / \sum_{i=1}^{n-1} \frac{1}{i} = \pi \]  

(4.17)

The nucleotide diversity \( \pi \) of the human population is about 0.1%, that is, there is about one nucleotide difference expected in every 1 kb between two randomly chosen genes. The value of \( \sum_{i=1}^{n-1} \frac{1}{i} \) is 2.83 (when \( n = 10 \)), 5.18 (when \( n = 100 \)), 7.48 (when \( n = 1,000 \)), 9.79 (when \( n = 10,000 \)). Thus, even if we collected 10,000 sequences from humans, we expect the number of segregating sites (the number of SNP) to be more or less 10 in 1 kb. If your data deviate from this by orders of magnitude, before considering some population genetic reasoning (such as selection or demographic effects), it is better to suspect some error or biased sampling. Because the human effective population size is about 10,000, estimation of SNP numbers in even larger samples need another theoretical framework.

Under the equilibrium between random genetic drift and mutation, nucleotide sequences from a population would satisfy both the number of segregating sites given by equation (4.2) and the nucleotide diversity (heterozygosity) given by equation (4.4). Thus, we can expect a distribution pattern of nucleotide configuration (nucleotide frequency spectrum) which tells us how many sites are low or high in heterozygosity.

### 4.4.1. Nucleotide frequency spectrum

Assume we have \( n \) sequences. At each segregating site, we can assume there are only two nucleotide kinds under the infinite site model. Let the number of one nucleotide to be \( i \) (then, the number of another is \( n - i \)) \((1 \leq i \leq n - 1)\). The expected number of segregating sites with \( i \) of one nucleotide kind is given as

\[ G_n(i) = \Theta \left( \frac{1}{i} + \frac{1}{n-i} \right) \]  

(Tajima 1989 Genetics 123:585-595)

For example, \( G_n(1) \) is the expected number of singleton sites. From equation (4.4), we have

\[ G_n(i) = S \left( \frac{1}{i} + \frac{1}{n-i} \right) / \sum_{j=1}^{n-1} \frac{1}{j} \]

For example, when \( n = 5 \), \( \sum_{j=1}^{n-1} \frac{1}{j} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{25}{12} \). Thus,

\[ G_5(1) = S \left( \frac{1}{1} + \frac{1}{4} \right) \frac{12}{25} = \frac{3}{5} S \], \quad G_5(2) = S \left( \frac{1}{2} + \frac{1}{3} \right) \frac{12}{25} = \frac{2}{5} S \], \quad G_5(1) + G_5(2) = S
$G_5(3)$ is equivalent with $G_5(2)$. $G_5(4)$ is equivalent with $G_5(1)$.

Panel 50: Figure 5 shows the expected nucleotide frequency spectrum $G_n(i)$ and the observed spectrum for human mitochondrial DNA sequence data ($n = 7, S = 45$). Low heterozygosity configuration sites appear to be more frequent than expected and high heterozygosity configuration sites appear less frequent.

This is reminiscent of Panel 29 explaining the interrelation between the number of alleles and heterozygosity. Difference would be that the number of segregating sites and nucleotide diversity is related by a simple formula (4.16) [or (4.17)]. As is the case of the number of alleles and heterozygosity, the equations (4.1) (4.2) (4.4) (4.5) hold when population size is constant through generations and nucleotide variations are selectively neutral. By testing if the equation (4.16) [or equation (4.17)] holds, we can evaluate if these conditions are met. This is Tajima's $D$ test.

4.4.2. Tajima's $D$ test

Tajima's $D$ test evaluate whether $\pi - s\sum_{i=1}^{n-1} \frac{1}{i}$ is positive, 0, or negative. Define $d = \pi - s\sum_{i=1}^{n-1} \frac{1}{i}$. Panel 50 right side is an image of Tajima's $D$ test. When $d > 0$ ($D > 0$ is used in the image: definition is given below), the nucleotide frequency spectrum is distorted so that high heterozygosity configuration sites are more frequent than the standard (constant and neutral) case. Conversely, when $d < 0$ ($D < 0$), low heterozygosity configuration sites are more frequent than the standard case.

The variance of $d$ is given by $\hat{V}(d) = c_1s + c_2s(s - \frac{1}{L})$, where $L$ is sequence length, $c_1 = \frac{b_1}{a_1} - \frac{1}{a_1^2}$, $c_2 = \frac{1}{a_1^2 + a_2}(b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2})$, $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$, $a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$, $b_1 = \frac{n+1}{3(n-1)}$, $b_2 = \frac{2(n^2 + n + 3)}{9n(n-1)}$, $n$ is the number of samples (Tajima 1989 Genetics 123:585-595). These variables appeared in page 57 as coefficients of terms in formulas to describe variance of $\Theta_{11} (\theta_\pi)$ and $\Theta_S (\theta_S)$. 
Tajima's $D$ is the quotient of $d$ divided by the estimated standard error of $d$, that is,

$$ \text{Tajima's } D = \frac{d}{\sqrt{V(d)}} $$

(4.18)

By applying a coalescence simulation (explained in pages 59-60), we can test whether a Tajima's $D$ value is statistically significantly positive or negative.

To conduct a coalescence simulation, we need to give a $\Theta$ (or $\theta$) value. The original article on Tajima's $D$ (Tajima 1989 Genetics 123:585-595) tested various $\Theta$ and $n$ values in the simulation and found that, irrespective of $\Theta$ values given, distribution of Tajima's $D$ value resembles the $\beta$ distribution of statistics. Tajima provided a table (Table 2 of the article) which shows significance level of Tajima's $D$ values under various $n$ values. Using this table, we can evaluate the significance of Tajima's $D$ in our data set. However, this method is not used today because the $\beta$ distribution is merely an approximation and not very accurate.

One simple method used today is

(1) Use only a single locus data. We regard $\Theta_w$ (i.e. $\sum_{i=1}^{n-1} \frac{1}{i}$), the method of moments estimate of $\Theta$ from the number of segregating site, as the true $\Theta$. We generate a virtual sequence data set by coalescence simulation under the $\Theta_w$. By repeating the simulation many times, we obtain a virtual distribution of Tajima's $D$. We evaluate if the observed Tajima's $D$ is within the 95% confidence interval of the distribution.

However, as explained before, the method of moments estimation is not very reliable even if based on $S$. In addition, estimation is based on only a single locus and has a large variance. Furthermore, without neutral reference data, we cannot know whether the result is due to selection or demographic effect. Thus, the following method should be taken.

(2) Use multiple loci data. First we obtain the distribution of, say, 10,000 accepted $\Theta$ values by following the rejection sampling procedure explained in page 62. Then, using each of 10,000 $\Theta$ values, 10,000 virtual sequence data sets are created. We thus obtain distribution of 10,000 Tajima's $D$ values. We evaluate if the observed Tajima's $D$ is within the 95% credible interval of the distribution.

If the purpose is to evaluate the constancy of demography (of population size), the study loci are desired to be from the regions where neutrality is highly expected, such as pseudogenes or non-genic regions. But, random choice of many regions from a genome is expected to generate a mostly neutral data set (the Neutral Theory to be explained in Chapter 7) and would also be acceptable.
Because of recombination, effect of natural selection (and sexual selection) to a gene is generally limited to
the gene and its nearby. In contrast, demographic events such as population size change and population
admixture affect entire genome regions similarly. Hence, to distinguish selection or demography, we need to
include reference regions in the analysis taken from supposedly neutral regions or many random choices of
genome regions.

If the purpose is to evaluate natural selection in a gene of interest, we first evaluate the observed Tajima’$D$ of
the neutral references in the theoretical standard distribution. If deviated, it is desirable to incorporate
population size change, population split, and migration into the simulation, and estimate these parameters to
satisfy the observed value (refer to section 4.3.2.6). Then, under the condition, if the gene of interest shows
deviation, we can interpret it due to natural selection.

4.4.3 Interpretation of Tajima's $D$ test
As in page 42-43 explaining the relation of the number of alleles and heterozygosy (Panel 29), Tajima's $D$
test results can be interpreted regarding demographic causes and natural selection causes.

In the case of Tajima’s $D > 0$

(1) Demographic causes
When population size reduces rapidly, low frequency alleles are easier to be lost. The number of alleles
reduces but heterozygosy is less affected. In other words, the low heterozygosy configuration sites tend to
lose its minor nucleotide kind and hence the number of segregating sites $S$ easily decreases, while high
heterozygosy configuration sites are not much affected and the nucleotide diversity $\pi$ does not decrease
much. If much time has not passed since this and a new equilibrium has not yet been attained, Tajima’s
$D > 0$ is observed. Another cause is an admixture of differentiated populations with comparable amounts
each other in nature or in sampling, which creates high heterozygosy configuration sites instantly.

(2) Natural selection causes
The higher frequency of high heterozygosy configuration sites implies operation of balancing selection.
There are several types of balancing selection: 1) overdominance which favors heterozygous genotypes (see
page 18), 2) negative frequency-dependent selection which favors minority genotypes (e.g. predator–prey
interaction or a form of sexual selection called disassortative mating system, that is, mating with a different
type from oneself in terms of the genetic trait in question), 3) niche divergence among genotypes, and 4)
mutual benefit of association, which predicts that individuals of each genotype benefit from being associated
with individuals of other genotypes in a polymorphic group.
In the case of Tajima’s $D < 0$

(1) Demographic causes

When population size expands rapidly, more alleles are born. Any allele starts as a minority. Thus, we have more minor alleles and lower heterozygosity than standard condition. In other words, at the time of mutation the site is always a singleton site, creating a low heterozygosity configuration site. Thus, we have lower $\pi$. If a population lost diversity by some reason and time has not passed enough, the same is expected. Examples are a bottleneck (a severe reduction of population size) and a founder effect (a small kin group starts a population). Another cause is to have a small number of immigrants from a differentiated population in nature or in sampling, which creates low heterozygosity configuration sites instantly.

(2) Natural selection causes

As explained in page 43, when a strong positive selective operates on an allele and drives it to fixation in relatively a short while, the gene and its surrounding are devoid of variation (selective sweep). This situation is similar with the start from the loss of diversity in (1). When mutations in a gene are all slightly deleterious, variants are kept in low frequency by purifying selection. In this case, too, low heterozygosity configuration sites increases. However, if mutations in a gene region are all too deleterious (e.g. dominant lethal), mutations are removed immediately. The situation is equivalent with low mutation rate, keeping both $\pi$ and $S$ at low value (called background selection) and not distorting their equilibrium. In this case, we do not expect Tajima’s $D < 0$, but Tajima’s $D \approx 0$.

Whether Tajima's $D$ is positive or negative, by examining if it is limited to a gene (genome) region or is pan-genomic, we can distinguish between demographic cause and selection cause.

4.4.4. Population-size effect on Tajima's $D$ viewed from coalescence branch length change

The effect of population size change on Tajima’s $D$ can be better understood if we consider changes of branch lengths in a coalescence tree (Panel 51). As in equation (3.21), the expected waiting time for the first coalescence is $\tilde{t}_i = \frac{4N_v}{i(i-1)}$. This means that, as time goes forward, branch lengths become shorter and shorter if population size is constant.
This shortening is relaxed if population size $N_e$ becomes larger as time goes by (terminal branch lengths are prolonged compared to the constant condition). Then, fraction of mutations near terminals (young mutations) increases. As young mutants are minority, they do not contribute to $\pi$ much while contributing to $S$ greatly. Hence, Tajima's $D$ becomes negative. If population size $N_e$ becomes larger as time goes by, the opposite occurs.

4.5. Analogues of Tajima's $D$

The population mutation rate $\theta$ can be estimated by various methods, not just by $S$ and $\pi$. Analogously to Tajima's $D$, by evaluating difference of estimates between different methods, we can reach an inference of natural selection and demographic changes.

4.5.1. Fu and Li's $G, F, D$  

Branches in a phylogenetic tree can be classified into external branches and internal branches (Panel 52). Let the number of mutations at external branches be $\eta_e$, the number of mutations at internal branches be $\eta_i$, and the number of total mutations be $\eta$. Then, their expectations are given as

\[ \eta_e = 4N_e \mu = \theta \]

\[ \eta = s = 4N_e \mu \sum_i \frac{1}{i} = \theta a_n \]

\[ \eta_i = \eta - \eta_e = \theta (a_n - 1) \]

(Fu and Li 1993 Genetics 133:693-709)
If we have outgroup sequence data, we can estimate \( \eta_e \) and \( \eta_i \).

Panel 52

As explained in Panel 51, if population expands, external branches are prolonged, thus, \( \eta_e \) becomes relatively larger. If population shrinks, the opposite occurs. From this viewpoint, tests are devised as:

- Fu and Li’s
  \[
  G = \frac{\eta_e - \eta_i}{\sqrt{V(g)}} \quad g \equiv \eta_e - \frac{\eta_i}{(a_n - 1)}
  \]
  \[
  Fu and Li’s \quad F = \frac{\pi - \eta_e}{\sqrt{V(f)}} \quad f \equiv \pi - \eta_e
  \]
  \[
  Fu and Li’s \quad D = \frac{n - a_n \eta_e}{\sqrt{V(d)}} \quad d \equiv n - a_n \eta_e
  \]

For the case we have no outgroup data, using the number of singleton sites, \( \eta_s \), and estimating \( \eta_e \) as

\[
\eta_e = \frac{n-1}{n} \eta_s
\]

the alternative tests are devised as

- Fu and Li’s
  \[
  F^* = \frac{\pi - n-1}{n} \eta_s \quad f^* \equiv \pi - \frac{n-1}{n} \eta_s
  \]
  \[
  Fu and Li’s \quad D^* = \frac{n-1}{n} \eta - a_n \eta_s \quad d^* \equiv \frac{n}{n-1} \eta - a_n \eta_s
  \]

4.5.2. Fay and Wu’s \( H \)

If we have outgroup sequence data, we can infer which nucleotide kind is ancestral and which is derived at each segregating site. Let the number of derived type nucleotide at each segregating site be \( i \) and let the number of segregating sites with \( i \) derived nucleotide kind be \( S_i \). Define

\[
\theta_H = 2 \frac{n}{n-1} \sum \left( \frac{i^2}{n} \right) S_i = 2 \sum \frac{i^2 S_i}{n(n-1)} \quad \text{(Fay and Wu 2000 Genetics 155:1405-1413)}
\]

Then, we reach

\[
\theta_H = 4N_e \mu = \theta
\]
That is, $\theta$ is estimated as twice of sum of unbiased estimator of homozygosity of derived nucleotide kind at each site.

Then, define

Fay and Wu’s $H \equiv \pi - \theta_H$

If a positive selection acts at a segregating site on the derived nucleotide kind and the allele type rapidly spreads into population, its linked haplotype in the neighboring genome region also spreads (selective sweep). Then, homozygosity of the region becomes high compared to other genome regions. The Fay and Wu's $H$ becomes a large negative value. The Fay and Wu's $H$ test is often used to detect selective sweep.

4.6. Effect of recombination on $\Pi$, $S$, and Tajima's $D$

We have ignored recombination so far. Recombination is as common as point nucleotide change. It would be useful to know how recombination affects the basic summary statistics $\Pi$, $S$, and Tajima's $D$.

Recombination has no effect on the expectations of $\Pi$ and $S$, and hence of Tajima's $D$. Imagine you have a sequence data set and you recombine them one another. The number of segregating site doesn't increase or decrease. As explained in page 50, the nucleotide diversity is the sum ($\Pi$) or the mean ($\pi$) of (estimated) heterozygosity of each site. Recombination does not change the configuration of each site and thus has no effect on heterozygosity at each site.

However, variances of them change. Recombination reduces variances of them. Distribution of observed values becomes narrow towards the expectations. Regarding the nucleotide diversity, smaller variance is because pairwise nucleotide differences become more similar among all pairs in a population. In the example in page 48, pairwise nucleotide differences range from 0 to 5. By recombination, you can imagine that all pairwise differences become close to the mean value 3.3. It's like recombination is shuffling sequences in a population.

Why does the variance of $S$ become small? Coalescence is helpful here again. Panel 53 is an explanatory example. Assume we sampled three sequences (alleles) (A, B, C) from a population. Assume recombination occurs to these sequences during going back to the past, combining the upper half of an allele to the lower half of the other allele in an individual. The genealogy of the upper half is depicted blue and of the lower half red. At black points, the two genealogies split. These points are the recombinant genes. The blue came from one parent and the red came from the other parent. Thus, the two go back through different paths. The blue lines and red lines eventually coalesce to different ancestors.
The coalescence time can be different between blue and red paths. The branching pattern can also be different: in blue A and B is closest, in red B and C are closest. These differences occur by change alone. The number of segregating sites among samples A, B and C is the sum of mutations which occurred in the blue path and in the red path. By recombination, we can regard this data set as an average of two independent trials of descent with mutations under the same $\theta$ conducted by nature. As explained in Panels 44 and 45, the variance of the mean of multiple trials is smaller than the variance of a single trial. This is why the variance of $S$ reduces by recombination.

To understand the effect of recombination on nucleotide diversity, coalescence is also helpful. In the blue genealogy nucleotide difference between A and B is smallest, while in the red genealogy nucleotide difference is smallest between B and C. In total, nucleotide differences between two of the three sequences are averaged. With more recombination events, this averaging effect becomes evident, resulting in similar differences among pairs.

Another explanation: The nucleotide diversity is the sum or average of heterozygosity among sites. High heterozygosity configuration sites have a mutation at older age in the coalescence tree. Low heterozygosity configuration sites have young mutations in the tree. Actual mutations do not always occur strictly in accord with the branch lengths in the tree by chance. Thus, the number of high sites and the number of low sites can fluctuate by chance. This is an explanation why the nucleotide diversity has a large variance. Recombination is like providing a multiple trial opportunity. Thus, averaging the number of high sites and the number of low sites among trials leads to a narrow range. Hence, the variance of nucleotide diversity becomes small by recombination.
Because both $\Pi$ and $S$ reduce their variance by recombination, Tajima's $D$, an evaluator of their difference, also reduces its variance.

4.7. Linkage disequilibrium

Nucleotide sequence data is a set of linked polymorphic sites. By evaluating a statistic called “linkage disequilibrium”, we can find regions where recombination is frequent (recombination hotspots) or is rare (linkage block). We can also detect a natural selection favoring a specific linkage of alleles between loci. We can also find a linkage marker for a disease allele of a gene. In this section, I shall overview the way of evaluation of linkage disequilibrium. For nucleotide sequence data, "locus" can be replaced with "polymorphic (SNP) site" and "allele" can be replaced with "allelic nucleotide".

4.7.1. What is linkage disequilibrium?

Consider two gene loci which are both polymorphic. A combination of alleles between the two loci in a gamete is called gametic phase. If the two loci are inherited independently each other, the frequency of the allele combination is expected to be the product of the frequencies of the two alleles. If this is met, the two gene loci are called to be in linkage equilibrium. If not, the two loci are called to be in linkage disequilibrium (LD). When the two loci are located in the same chromosome and when the recombination rate (explained later) between them is less than 0.5, these two gene loci are called linked genes. A gametic phase of linked genes is called a haplotype.

It would be easy to imagine that LD occurs between linked genes. LD is getting resolved by recombination through generations. If LD is not resolved after generations, it is implied that recombination is repressed for some reason (e.g. by inversion) or a natural selection is operating to maintain the specific haplotype.

However, LD could also occur between un-linked genes. Imagine the case of population differentiation where allelic composition differs between them by random genetic drift (in an extreme case, each population is fixed with a different allele). If we do not distinguish the two populations and sampled from them, we will see specific combinations appear with distorted frequencies (in the extreme case, only specific combinations appear). In this case, LD is not resolved unless gene flow occurs between the populations. Hence, LD is the population-level information, from which we can extract population genetic information such as on population structure and natural selection.

4.7.1.1. <Difference from linkage analysis>

LD analysis is often confused with the "linkage analysis". In the linkage analysis, through a test cross or a pedigree analysis, we examine whether inheritance patterns of alleles are associated between polymorphic
loci. We then learn whether they are linked or not. If linked, we calculate the recombination rate between them. We then estimate the physical distance between them. In the linkage analysis, we are not interested in phase itself (combination of alleles). In LD, we are interested in phase itself, interested in frequency of each allele and each phase in a population.

In principle, recombination rate is provided by the linkage analysis. Gametic phase information is also gained from test crosses or pedigree analysis, or determined directly by examining gametes. The haplotype of SNPs is determined only by DNA cloning and nucleotide sequencing or, recently, by single-molecule DNA sequencing. However, in LD analysis, it is not practical to gain information of gametic phase or haplotype by these authentic methods. LD analysis requires a lot of population samples. Physical distance between SNP sites can be too long to be easily cloned and sequenced (“Next Generation Sequencing” may break through this difficulty). Thus, gametic phase or haplotype is usually estimated from allele frequency data of a population using a specialized algorism, such as EM algorism or PHASE. The principle and practice of the estimation is complex and its explanation is omitted here.

The linkage analysis and LD analysis are used complementarily today. When LD is detected between two loci, the most probable cause is that the two are closely located. If we have a plenty of population samples and many polymorphic markers, we can find a marker of a gene of interest (e.g. disease-related genes) by LD analysis without performing painstaking and time-consuming crossing experiments or looking for rare pedigrees showing a disease trait. Today, genome-wide diversity analysis is becoming feasible. By regarding a genome region with high LD as one block, and by choosing a SNP in the region as a tag SNP representing the region, we can look for a tag SNP in whole genome which is associated with a phenotypic trait. Such Genome-Wide Association Study (GWAS) is very popular today.

4.7.1.2. <Measures of LD>

Let two alleles of locus A be $A_1$ and $A_2$ and two alleles of locus B be $B_1$ and $B_2$. When both loci are heterozygous in an individual, there are four possible combinations (four gametic phases): $A_1B_1$, $A_1B_2$, $A_2B_1$, $A_2B_2$. Let the gametic phase frequencies be $P_{11}$, $P_{12}$, $P_{21}$, $P_{22}$, respectively, in a population. Let the allele frequencies of $A_1$ and $A_2$ be $p_1$ and $p_2$, and the allele frequencies of $B_1$ and $B_2$ be $q_1$ and $q_2$, respectively, in the population.

If A and B are not linked, $P_{11} = p_1q_1$ is expected (Note we have a hidden assumption that simple size is large enough for random fluctuation of phase frequency not to occur) (see Panel 54)
A measure of LD, $D$, is defined as the deviation of observed $P_{11}$ from expectation $p_1q_1$.


text

Then, it follows as (see Panel 55 for explanation)

$$
\begin{align*}
P_{11} - p_1q_1 &= D \\
P_{12} - p_1q_2 &= -D \\
P_{21} - p_2q_1 &= -D \\
P_{22} - p_2q_2 &= D
\end{align*}
$$

(4.19)

Panel 55

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$B$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>18+D</td>
<td>28-D</td>
</tr>
</tbody>
</table>

Allele freq. given

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>42-D</td>
<td>28+D</td>
</tr>
</tbody>
</table>

Expectation of phase freq. derived from allele freq.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$B$</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>18+D</td>
<td>12-D</td>
</tr>
<tr>
<td>2</td>
<td>42-D</td>
<td>28+D</td>
</tr>
</tbody>
</table>

Phase $A_1B_1$ is observed more than expectation by $D$

The number of each allele is given and fixed

Compensate $A_1B_2$ and $A_2B_1$ numbers by minus $D$

Compensate $A_2B_2$ number by plus $D$
If an individual is heterozygous in both $A$ and $B$ loci, it has $A_1B_1$ and $A_2B_2$ phases or $A_1B_2$ and $A_2B_1$ phases. The incidence of the former is $P_{11}P_{22}$ and the incidence of the latter is $P_{12}P_{21}$. The difference of the two incidences is also $D$. See Panel 56.

$$P_{11}P_{22} - P_{12}P_{21} = (p_1q_1 + D)(p_2q_2 + D) - (p_1q_2 - D)(p_2q_1 - D)$$
$$= D(p_1q_1 + p_2q_2 + p_1q_2 + p_2q_1)$$
$$= D(p_1(q_1 + q_2) + p_2(q_1 + q_2))$$
$$= D(p_1 + p_2)$$
$$= D$$

$$P_{11}P_{22} - P_{12}P_{21} = D \quad (4.20)$$

If no LD, $P_{11}P_{22} = P_{12}P_{21} = p_1p_2q_1q_2$, and $D = 0$. In the extreme LD, either $P_{11}P_{22} = 0$ or $P_{12}P_{21} = 0$, and thus either $D = P_{11}P_{22}$ or $D = -P_{12}P_{21}$. In milder LD, $D$ is between the two values.

Because $P_{11}, P_{12}, P_{21}, P_{22}$ are all positive or 0, equation (4.19) gives

$$D \geq -p_1q_1$$
$$D \leq p_1q_2$$
$$D \leq p_2q_1$$
$$D \geq -p_2q_2$$

(4.21) Thus,

$$(\text{the larger of } -p_1q_1 \text{ and } -p_2q_2 = D_{\text{min}}) \leq D \leq (\text{the smaller of } p_1q_2 \text{ and } p_2q_1 = D_{\text{max}}) \quad (4.22)$$

Then, standardize $D$ by defining

$$D' \equiv \frac{D}{D_{\text{max}}} \quad \text{if } D > 0$$
$$D' \equiv \frac{D}{D_{\text{min}}} \quad \text{if } D < 0$$

(4.23) $0 \leq D' \leq 1$

$D'$ is one of representative measures of LD.

To evaluate if LD is significant or not, $\chi^2$ test is conducted to the difference between the number of observed phase kind and its expectation under no LD [$D = 0$ in equation (4.19)]. Let the sample number be
Degree of freedom is 1 [the number of data (phase kind) is 4, the number of one allele is calculated from the data for 2 loci (the number of the other allele is automatically given and is not counted in the df), thus, 4-1-1-1=1].

\[
\chi^2 = \frac{n^2(P_{11} - p_1q_1)^2}{np_1q_1} + \frac{n^2(P_{12} - p_1q_2)^2}{np_1q_2} + \frac{n^2(P_{21} - p_2q_1)^2}{np_2q_1} + \frac{n^2(P_{22} - p_2q_2)^2}{np_2q_2}
\]

\[
= \frac{nD^2}{p_1q_1} + \frac{nD^2}{p_1q_2} + \frac{nD^2}{p_2q_1} + \frac{nD^2}{p_2q_2} = \frac{p_2q_2 + p_2q_1 + p_1q_2 + p_1q_1}{p_1p_2q_1q_2} nD^2 = \frac{p_2(q_2 + q_1) + p_1(q_2 + q_1)}{p_1p_2q_1q_2} nD^2
\]

\[
= \frac{nD^2}{p_1p_2q_1q_2}
\]

\[
\frac{n(P_{11}P_{22} - P_{12}P_{21})^2}{p_1p_2q_1q_2}
\]

This is a regular \(2 \times 2\) independence test of allele frequency composition between two loci. As association becomes high, this becomes large.

By defining

\[
r^2 \equiv \frac{D^2}{p_1p_2q_1q_2}
\]

\(r^2\) is another representative measure of LD, which reflects association of allele frequency between two loci and statistical significance of LD.

From equation (4.22)

When \(D > 0\)

\[
0 \leq D^2 \leq \{\text{the smaller of } (p_1q_2)^2 \text{ and } (p_2q_1)^2\} \leq p_1p_2q_1q_2
\]

When \(D < 0\)

\[
0 \leq D^2 \leq \{\text{the smaller of } (p_1q_1)^2 \text{ and } (p_2q_2)^2\} \leq p_1p_2q_1q_2
\]

Hence

\[
0 \leq D^2 \leq p_1p_2q_1q_2
\]

Thus, from equations (4.25) and (4.26)

\[
0 \leq r^2 \leq 1
\]

Panel 57 is an exemplary explanation of LD calculation (\(D, D', r^2, \text{ and } \chi^2\)) for SNPs in a sequence data.
An example of LD calculation between SNP sites

<table>
<thead>
<tr>
<th>site 1 (A/G)</th>
<th>site 2 (T/C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
</tr>
</tbody>
</table>

**Site 1**

|--------|--------|--------|--------|--------|--------|

**Site 2**

<table>
<thead>
<tr>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

Focus on two polymorphic sites

Count 4 kinds of phase

Count alleles (nt kinds) at each site

**D** Cross phase frequencies diagonally and subtract each other

\[ D = \frac{3}{7} \times \frac{1}{7} - \frac{2}{7} \times \frac{1}{7} = \frac{1}{49} > 0 \]

Because D is positive, calculate \( D_{\text{max}} \)

(If D is negative, calculate \( D_{\text{min}} \))

**D_{\text{max}}** Cross allele frequencies nonparallel, choose the smaller

\[ D_{\text{max}} = \begin{cases} \text{the smaller of} & \frac{4}{7} \times \frac{2}{7} & \text{and} & \frac{5}{7} \times \frac{3}{7} \end{cases} = \frac{8}{49} \]

**\( D' \)**

\[ D' = \frac{D}{D_{\text{max}}} = \frac{1}{8} \]

\[ r^2 = \frac{D^2}{p_A p_B p_C p_C} = \frac{\frac{1}{49}}{\frac{120}{4 \times 3 \times 5 \times 2}} = \frac{1}{120} \]

\[ \chi^2 = n r^2 = \frac{7}{120} \quad d.f. = 1, \quad p \approx 0.8 \]

LD is not statistically significant

4.7.2. Difference of \( D' \) and \( r^2 \)

Panel 58

**Relation of allele frequency composition to \( r^2 \)**

\[ A-B \quad a \rightarrow B \quad a-B \quad b \rightarrow a \]

\[ a-b \text{ does not emerge without recombination between } a-B \text{ and } A-b \]

**A-b**

\[ D' = 1 \]

\[ a \text{ and } b \text{ rare, } r^2 \text{ small} \]

\[ a \text{ and/or } b \text{ common, } r^2 \text{ large} \]

\[ \text{no } A-B, \quad r^2 = 1 \]

\[ a-B \quad b \rightarrow A \quad b \rightarrow A \]

\[ A-B = \frac{(AB)(ab) - (Ab)(aB)}{AaBb} \]

\[ r^2 = \frac{(Ab)(aB)}{AaBb} \]

\[ = \frac{b^2 a^2}{AaBb} \]

\[ a = \frac{a}{A} \]

\[ = \frac{1}{81} \]

\[ \text{One of four phases is 0 } \Rightarrow D' = 1 \]

Its diagonal phase is also 0 \( \Rightarrow r^2 = 1 \)
Refer to Panel 58. Suppose we have gene loci A and B. At the beginning, both are monomorphic, having $A$ and $B$ alleles only, respectively. We have only one gametic phase $A - B$. First, at locus A, a mutant allele $a$ is born. We have two phases $A - B$ and $a - B$. Later, at locus B, a mutant allele $b$ is born to a $A - B$. Then, we have three phases $A - B$, $a - B$ and $A - b$. The fourth phase $a - b$ does not emerge unless recombination occurs between $a - B$ and $A - b$. In the absence of $a - b$, $D' = 1$. Thus, $D'$ is primarily an indicator that recombination occurred or not.

Under this maximum LD in terms of $D'$ (i.e. $D' = 1$), $r^2$ can be very small depending on allele frequencies of mutants $a$ and $b$. This is when both $a$ and $b$ are in low frequency. Because $a - B$ and $A - b$ are rare, they are unlikely to meet in an individual and recombine to produce $a - b$. As in equation (4.24), $r^2$ is an indicator of association of allele frequency composition between two loci and a measure of LD which reflects its statistical significance. In this case, absence of $a - b$ (i.e. LD) is likely to happen. In other words, statistical significance (i.e. $r^2$) of its absence is low. This is why $r^2$ is low.

If either $a - B$ or $A - b$, or both of them, are common, they are likely to meet in an individual and recombine to produce $a - b$. In these cases, absence of $a - b$ is unlikely to happen. In other words, statistical significance (i.e. $r^2$) of its absence is high. This is why $r^2$ is high.

Another explanation is this. $a$ is linked to $B$, and $b$ is linked to $A$. Thus, the frequencies of $a$ and $B$ are associated, and the frequencies of $b$ and $A$ are associated. However, when both $a$ and $b$ are rare, the $a$-bearing $B$ is few among all $B$, and the $b$-bearing $A$ is few among all $A$. In this case, the association is low. Thus, the indicator of association $r^2$ is low.

On the contrary, if $a$ increases, association between of $a$ and $B$ becomes large. If $b$ increases, association between of $b$ and $A$ becomes large. Hence, $r^2$ becomes large. If both $a$ and $b$ increase, $r^2$ becomes even larger. When $A - B$ happens to be lost, only $a - B$ and $A - b$ are left. Then, the association between $a$ and $B$ and that between of $b$ and $A$ become perfect (correlation coefficient is 1). That is, $r^2$ is 1. In these cases, the association is high. Thus, the indicator of association $r^2$ is high.

4.7.3. Time course of LD

The recombination rate is the proportion of gametic phases that are different from those inherited from its parents in germ cells in an individual. That is, the proportion of recombinant gametic phases produced in one generation. Let the recombination rate be $c$.

Suppose we have alleles $A_1$ and $A_2$ in locus A, and alleles $B_1$ and $B_2$ in locus B. Suppose an individual
inherited $A_1B_1$ from father and $A_2B_2$ from mother. If the two loci are not linked, the individual is expected to produce gametes with $A_1B_1, A_2B_2, A_1B_2,$ and $A_2B_1$ phases with equal portions. In this case, recombination rate is 0.5. This is the maximum of recombination rate. Thus, $0 \leq c \leq 0.5$.

When recombination rate is less than 0.5, the two loci are regarded as linked (located in the same chromosome and not as far away as recombination occurs in every meiosis).

Consider the probability that a randomly chosen gamete has the gametic phase $A_1B_1$. It either experienced a recombination or no recombination.

If the former, it is a recombination product between $A_1B_x$ and $A_xB_1$. $B_x$ can be $B_1$ or $B_2$. $A_x$ can be $A_1$ or $A_2$. The frequency of $A_1B_x$ is the same with the allele frequency of $A_1$ ($p_1$). The frequency of $A_xB_1$ is the same with the allele frequency of $B_1$ ($q_1$). Thus, the probability that $A_1B_1$ is a recombinant is $p_1q_1$ multiplied by the recombination rate $c$, i.e. $cp_1q_1$.

If the latter, $A_1B_1$ is in its parent and escaped from recombination. Its probability is given by the frequency of $A_1B_1$ ($P_{11}$) multiplied the probability of no recombination $1-c$, i.e. $(1-c)P_{11}$.

Let the probability that a gamete is $A_1B_1$ be $P_{11}$. Then,

$$P_{11} = cp_1q_1 + (1-c)P_{11}$$

By transforming it

$$P_{11} - p_1q_1 = (1-c)(P_{11} - p_1q_1)$$

From equation (4.8), $P_{11} - p_1q_1 = D$. Let $D$ at generation $t$ be $D_t$.

Then, we have a geometric progression function

$$D_t = D_{t-1}(1-c)$$

By solving it,

$$D_t = D_0(1-c)^t \quad (4.27)$$

It gives a time course of LD.

It tells that LD approaches to 0 through generations. As $c$ becomes smaller, its speed becomes slower (Panel 59: Figure 3.9). When $c$ is 0, LD is never resolved. Fastest is when $c$ is the maximum, 0.5, i.e., when two loci are not linked. LD is resolved by half in every generation. The equation (4.27) provides a baseline of LD in random mating population without natural selection.

Note, however, the explanation so far has ignored the effect of random genetic drift of allele frequency and
gamete phase frequency. A large population size is assumed, and only recombination is assumed to change phase frequency. In reality, even without recombination, allele frequency and phase frequency fluctuate, and LD values fluctuate through it. In small populations, the fluctuation becomes more evident.

Panel 59

4.7.4. Summary of various effects on LD

(1) The physical distance between loci (or SNPs). The shorter the distance is, the lower the recombination rate becomes, and the higher the LD becomes, in general. About 100 kb is a rough expectation to find a high LD.

(2) If there is a local increment of recombination rate (recombination hotspot) between loci, LD becomes smaller even if they are close.

(3) On the contrary, if recombination is repressed due to a chromosomal inversion or other reasons, LD occurs even between distant loci (well known in fruit fly).

(4) If a population consists of inbred families which differ in allele frequency composition (fixed with different alleles in extreme), LD occurs even between unlinked loci. Recombination is prohibited by reproductive barrier.

(5) If a population consists of subpopulations which differ in allele frequency composition (fixed with different alleles in extreme), LD occurs even between unlinked loci. This also happens if we sample differentiated populations as one population. Recombination is prohibited by reproductive isolation between subpopulations.

(6) If effective population size is small (the case includes bottleneck and founder effect), LD becomes higher than a large $N_e$ population. This is because smaller $N_e$ populations have shorter coalescence time and hence provide less time to resolve LD. The Out-of-Africa spread of human populations is a well-known example, where Eurasian populations (e.g. Europeans and Asians) have generally higher LD than larger $N_e$ African populations.
(7) When population expands, low heterozygosity configuration sites increase (Panel 51). This means $r^2$ value becomes small (Panel 58). On the contrary, when population shrinks, low heterozygosity configuration sites decrease, and $r^2$ value becomes large. See Panel 60.

Panel 60

(8) If natural selection favors a combination of alleles between two loci (SNPs), recombinants are eliminated and high LD is expected. If LD is low in other genome regions, this can be a supporting evidence of natural selection.

(9) If natural selection favors one allele at a locus (SNP) and selective sweep occurs to the genome region faster than recombination rate, high LD is expected. As in (8), if LD is low in other regions, this can be a supporting evidence of natural selection.

(10) Finally, when sample size is small, high LD can appear to arise even if LD does not occur in reality (Panel 54). This is an artifact and caution should be paid.

Nucleotide sequence data of a population is a rich source of SNPs and contains a wealth of LD information. By measuring LD to all pairs of SNP sites in a genome region, we can overview the distribution of LD blocks. This could provide us with many implications on natural selection, recombination hotspots, disease markers, etc. Panel 59: Figure 6 is an example of such LD distribution in a genome region.
Chapter 5 Molecular Evolution I (Nucleotide Substitution)

We have considered the genetic variation (diversity) within a species. Hereafter, we shift our focus on genetic divergence between species. On within-species diversity, we are interested in the process of fixation of an allele. On between-species divergence, we are interested in fixed mutations. Divergence of species is the difference of mutations which were fixed in each species. In this chapter, we consider how fast the genetic divergence is accumulated between species and what factors determine it.

5.1. Nucleotide substitution

In a nucleotide sequence, a mutation at a nucleotide site is spread through a population and fixed (the old nucleotide is substituted with the new one). At another site, an old nucleotide is substituted with another mutation, and so on. This repeating substitution process is called nucleotide substitution. The "substitution" is meant to be the fixation of mutations in a species. It is important to distinguish the substitution at population level from the mutation at individual level. From a viewpoint of genes, evolution is the repeating process of fixation of mutations (see Panel 61). Through understanding the process of nucleotide substitution, we can unite genetic diversity within a species and genetic divergence between species.

Panel 61 shows an image of successive nucleotide substitutions a, b and c and gradual change of nucleotide sequence. You may imagine that b arises after a is fixed and c arises after b is fixed. But as in Panel 62, b and c can arise before a is fixed. Note there are many other mutations that are eventually lost.
5.1.1. Fixation probability of mutations

To understand the process of nucleotide substitution, we first need to learn the fixation probability. As mentioned in page 23, this is the probability of a mutant allele to be eventually fixed in a population and is derived by the diffusion model. Its derivation requires a high expertise of mathematics and only typical examples are introduced here. Let \( u \) be the fixation probability hereafter.

(1) If the mutation is neutral

The fixation probability \( u \) is equal to the initial frequency of the mutant gene. In diploid organisms with population size \( N_e \), the mutant gene is only one when born. Then, initial frequency is \( \frac{1}{2N_e} \). Thus,

\[
u = \frac{1}{2N_e}
\]

This is intuitively understandable because all \( 2N_e \) genes have the same chance to be fixed.

(2) When the mutation is codominant (see page 16)

Let \( A \) be a wild type allele and \( a \) be a mutant advantageous allele. Assume fitness of genotypes \( AA, Aa, aa \) as 1, \( 1+s \), \( 1+2s \), respectively. In this case, the fixation probability of the mutant allele \( a \) is given as

\[
u = \frac{1 - e^{-4N_esp}}{1 - e^{-4N_es}} \quad (5.1)
\]

where \( p \) is the initial frequency of \( a \). Here, \( p = \frac{1}{2N_e} \), thus,

\[
u = \frac{1 - e^{-2s}}{1 - e^{-4N_es}}
\]
Special case 1: When $s$ is extremely close to 0 (i.e. when neutral)

Because $e^{-x} \approx 1 - x$ (if $x \ll 1$)

$$u = \frac{1 - e^{-2s}}{1 - e^{-4N_e s}} \approx \frac{1 - (1 - 2s)}{1 - (1 - 4N_e s)} = \frac{2s}{4N_e s} = \frac{1}{2N_e}$$

This is equivalent with (1).

Special case 2: When $s$ is small (but not extremely: at most a few %)

$e^{-2s} \approx 1 - 2s$ still holds but $4N_e s \ll 1$ does not hold, thus,

$$u = \frac{2s}{1 - e^{-4N_e s}}$$

When $N_e$ is large enough and $4N_e s \gg 1$, we have $e^{-4N_e s} \rightarrow 0$. Thus,

$$u = 2s \quad \text{neat!}$$

Implications: If $N_e$ is 50,000, the fixation probability of a neutral mutation is only 1/100,000. In the case of codominant selection, when advantage is 1% ($s = 0.01$), the fixation probability is 2%. This is much higher than the neutral case. However, still this advantageous mutant can be lost with 98% probability only by chance. In another view, disadvantageous mutations ($s$ is negative) can be fixed by chance if disadvantage is small and population is small (When $N = 1000$ and $s = -0.001$, $u = 0.004\%$). These remind us of the significance of chance effect on the fate of mutations.

Formulas of $u$ in various other situations are obtained [refer to Nei's "Molecular Evolutionary Genetics"].

By the presence or absence of natural selection, $u$ varies greatly. However, the fate of mutations in early phase is similar. The extinction probability in the first seven generations for a neutral mutation is 0.79, while that for a 1% advantageous mutation is 0.78. It tells us the significance of chance effect on the fate of mutations in early phase.

5.1.2. Fixation time of mutations

For mutations that are eventually fixed, the probability distribution of time from appearance to fixation (conditional fixation time) is derived by the diffusion model. The expectations are given as the followings.

(1) If the mutation is neutral

$4N_e$ generations (panel 61: Fig. 3.1)

Remember, this can also be reached by the coalescence theory (page 37).
(2) When the mutation is advantageous by $s$

$$\frac{2}{s} \ln 2N_e \text{ generations}$$

Implications: Suppose a mammalian species with generation time 2 years and effective population size 1 million. The expected fixation time of a neutral mutation is 8 million years, while that of a 1% advantageous mutation is only 5,800 years. Advantageous mutations are either lost in early phase or fixed quickly (Panel 61: Figure 2.6a). On the other hand, a neutral mutation changes its frequency very slowly in a large population and generate a large amount of transitional polymorphism.

5.1.3. The rate of nucleotide substitutions

The number of nucleotide substitutions per unit time is called the rate of nucleotide substitutions or, more generally, evolutionary rate. This is often symbolized as $\lambda$. Let the population size be $N_e$, the mutation rate per generation be $v$, and the fixation probability be $u$. There are $2N_e$ genes in a population. $2N_e v$ mutants emerge in every generation. Fraction of $u$ among the mutants is eventually fixed. Thus,

$$\lambda = 2N_e v u \quad (5.2)$$

Here, $v$ is given as per generation. Then, $\lambda$ is substitution rate per generation, $\frac{1}{\lambda}$ is the number of generations between two successive substitutions. If $v$ is given as per year, $\lambda$ is per year and $\frac{1}{\lambda}$ is in the number of years.

In the case of neutral mutations, $u = \frac{1}{2N_e}$, thus,

$$\lambda = v \quad (5.3) \quad \text{(Panel 61: Fig. 3.1)}$$

This simple equation between substitution rate and mutation rate in neutral case was first pointed out by Motoo Kimura (1968, Nature, 217:624-626). Implication is tremendous. Nucleotide substitution rate at species level is only dependent of mutation rate at individual level and irrelevant to population size.

In the case of advantageous mutations, if we take the example of codominant selection, $u = 2s$, thus,

$$\lambda = 4N_e sv \quad (5.4)$$

The substitution rate depends on population size, fitness, and mutation rate, much more complex than the neutral case.
Chapter 6 Molecular Evolution II (Phylogenetic Tree)

Molecular evolutionary studies provide a theoretical ground to explain why evolutionary changes of DNA sequences can occur proportionally to time (equation 5.3). In addition, the mutation rate per site is extremely low, the nucleotide change is always among three directions, and mutations occur on enormous amount of nucleotide sites at random in the genome. Thus, the probability is practically 0 that two sequences from different lineages converge to an identical sequence by chance. Hence, the similarity of DNA sequences reflects well the evolutionary relatedness and relative distance (relative divergence times) among organisms. In principle, classification and estimation of evolutionary relationship among organisms are carried out based on their any structural similarity. Then, the DNA sequence should be the best primary source for this purpose.

However, DNA does not tell us the absolute divergence times. Paleontological and paleogeographycal information is indispensable. DNA retrieval from ancient samples is possible today, but is severely limited. Fossil discovery and morphology is still a key in phylogenetic reconstruction.

Reconstruction of phylogenetic tree can be a research purpose. But, it also provides a framework along which evolution of any physiological function is interpreted. For example, regarding a character shared between species, we can distinguish whether it evolved independently in separate lineages or it occurred in a common ancestor of them. We can estimate when and where in a tree a physiological function arose and changed. We can apply a knowledge gained from a species to a different species based on their evolutionary relatedness. Truly, "nothing in biology makes sense except in the light of evolution" – Dobzhansky.

6.1. Estimation of evolutionary distance and divergence time

6.1.1. Alignment of sequences

The genetic divergence between species is also called evolutionary distance. To estimate evolutionary distance at the nucleotide or amino acid sequence level, we first need to line up the sequences by inserting gaps so that each site is arranged to be homologous among the sequences. This process is called sequence alignment. In practice, gaps are inserted so that evolutionary distances among them become minimal. If sequence differences are small, alignment can be easily carried out manually. However, divergent and long sequences are difficult to align manually, and algorithms are devised for this purpose (refer to Nei's "Molecular Evolutionary Genetics" for details). CLUSTAL W is one of popular alignment programs equipped in commercial DNA analysis software and free molecular phylogenetic software (e.g. MEGA: http://www.megasoftware.net/). Other popular examples include MUSCLE (equipped also in MEGA) and MAFFT. Panel 63 is an example of amino acid sequence alignment.
6.1.2. Evolutionary distance

During a long evolutionary time, one difference could correspond to two or more substitutions (Panel 64: Figure 1-5 lower row middle and right). Even seemingly no difference can be due to parallel substitutions (Panel 64: Figure 1-5 upper row middle) or a reverse substitution (Panel 64: Figure 1-5 upper row right). This assumption of multiple substitutions at a site is the difference from the case of within-species variation. The infinite site model is not applied here. The evolutionary distance is the number of nucleotide (or amino acid) substitutions, but not of differences. The nucleotide (amino acid) difference underestimates the evolutionary distance. When the difference is below ~5%, underestimation is not a problem. For larger differences, a correction needs to be conducted to estimate the number of substitutions from the differences.
Let $n$ be the sequence length, $n_d$ be the number of nucleotide (or amino acid) difference, $p (= n_d/n)$ be the proportion of the difference, and $d$ be the estimated number of the nucleotide substitutions per site. The evolutionary distance is $d$.

**Amino acid sequence:**

The simplest and popular correction formula is called Poisson correction

$$d = -\ln(1 - p)$$  \hspace{1cm} (6.1)

Let $\lambda$ be the evolutionary rate per site per year, $T$ be the divergence time (year unit). The evolutionary rate is equivalent with the probability of substitution at a site in one year. The total time to consider is $2T$ years. Thus, the expected evolutionary distance $d$ is given as

$$d = 2\lambda T$$

Let $r$ be the number of substitutions at one site. The probability distribution of $r$ during $2T$ years is equivalent with a binominal distribution of choosing $r$ years of substitutions from $2T$ years ($\lambda$ is so small that we regard two or more substitutions not to occur in one year). Thus,

$$P(r) = \frac{2T}{r!} C_r (\lambda)^r (1 - \lambda)^{2T-r}$$

Because of $\lambda \ll 1$ and $2T \gg 1$, $P(r)$ is approximated to a Poisson distribution with expectation $2\lambda T$. Then,

$$P(r) = \frac{(2\lambda T)^r e^{-2\lambda T}}{r!}$$  \hspace{1cm} \text{Because of} \hspace{0.5cm} 2\lambda T = d ,$$

$$P(r) = \frac{d^r e^{-d}}{r!}$$

Here, if we regard the probability of $r = 0$ as equivalent with the proportion of identical amino acid sites,

$$P(r = 0) = e^{-d} = 1 - p$$

Then, equation (6.1) is reached.

The standard error of $d$ is given as

$$\sqrt{\frac{p}{(1 - p)n}}$$

In this correction, we do not take into consideration the parallel and reverse substitutions (because of smaller probabilities). However, in the case that $p$ is larger than $\sim 0.3$, these effects are not negligible. An example of correction formula for such case is
\[ d = -\ln(1 - p - \frac{1}{3}p^2) \]

This is called the Kimura's empirical formula. Refer to Nei's "Molecular Evolutionary Genetics" for more details.

Nucleotide sequence:

The simplest and popular correction formula is called Jukes-Cantor correction.

\[ d = \frac{3}{4} \ln(1 - \frac{4}{3}p) \] (6.2)

The standard error is given as

\[ \sqrt{\frac{9p(1-p)}{(3-4p)^2n}} \]

By transforming the equation (6.2), we have

\[ p = \frac{3}{4} \left( 1 - e^{-\frac{4d}{3}} \right) \]

It tells that the maximum of \( p \) is 3/4 (Panel 65). There are only four nucleotide kinds. Two random sequences can match in 1/4 of sites by chance alone. It means that \( d \) cannot be estimated when the nucleotide difference is close to 75%. This is called saturation of nucleotide substitution.

Panel 65

Transitions are more often than transversions (see Panel 64: Tables 1.5, 1.6). Transitions are purine-to-purine (A \( \leftrightarrow \) G) and pyrimidine-to-pyrimidine (T \( \leftrightarrow \) C) nucleotide changes. Transversions are nucleotide changes between a purine and a pyrimidine. A correction formula considering this mutation bias is

\[ d = -\frac{1}{2} \ln[(1 - 2P - Q)\sqrt{1 - 2Q}] \] (Kimura's two-parameter method) (6.3) (SE equation omitted)

where \( P \) is the proportion of transition differences and \( Q \) is the proportion of transversion differences.
Many other estimation methods are devised which consider mutation bias, nucleotide content, etc., such as Tajima and Nei's formula. Refer to MEGA's Help menu or Nei's textbooks for more details.

The reliability of various correction methods can be tested using computer simulation where nucleotide changes are generated under some nucleotide content and transition/transversion ratio given. Except in the case of very biased nucleotide changes such as in mammalian mitochondrial DNA, estimation of $d$ is similar among methods when $d < 0.5$. When $d$ is small, the simple Jukes-Cantor correction is well applicable.

6.1.3. Estimation of evolutionary rate and divergence times

While giving us estimates of the evolutionary distance (the number of nucleotide or amino acid substitutions), sequence data cannot tell us the evolutionary rate and divergence times. Time standard can only come from paleontological (fossil) or geological (e.g. continental drift) information. Panel 66: Tables 2.1 and 2.2 are representative examples of such standards.

First, we estimate a $d$ value between sequences of a gene from standard organisms. Second, we divide $d$ by twice of their divergence time $2T$ inferred from paleontological or other information (calibration). This is the estimate of the evolutionary rate of the gene. The twice is because $d$ represents substitutions in two lineages after divergence.

$$\lambda = \frac{d}{2T}$$ (6.4)

The evolutionary rate per year is generally at the order of $10^{-9}$ in nuclear genes (Panel 67: Figures 16, 17).
Then, by using the \( \lambda \) value, we infer divergence times between organisms for which paleontological or other information is not available (see Panel 66: Figure 2.4 as an example).

\[
T = \frac{d}{2\lambda} \quad (6.5)
\]

How long would it take for neutral sequences (such as introns) to be too divergent to align each other? Assume the evolutionary rate of an intron to be \( \sim 4 \times 10^{-9} \) site/year. Assume also \( p \) is 0.4 between two sequences which we may feel a difficulty to align them. Then, the Jukes-Cantor estimate of \( d \) is about 0.6 according to the equation (6.2). Thus, according to the equation (6.5), the divergence time \( T \) in this case is \( 7.5 \times 10^7 \) years (75 million years) ago. Divergence time between human and mouse is about 100 million years ago. Thus, it is not surprising that human and mouse intron sequences are difficult to align.

6.2. Cautions in making a phylogenetic tree

6.2.1. Rooted tree and unrooted tree

A phylogenetic tree is a graph which consists of study objects (sequences) (called OTU: Operational
Taxonomic Unit), nodes and branches. Essentially, it is unrooted. Only when an outgroup reference is included, it becomes rooted (Panel 68: Figure 6). A root ("R" in Figure 6) is a node which corresponds to the common ancestor of the in-group. Rooted trees contain information of time flow. Unrooted trees provide only branch lengths (distances among nodes and OTUs).

Panel 68

<table>
<thead>
<tr>
<th>The number of OTU</th>
<th>The number of possible topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
</tr>
<tr>
<td>9</td>
<td>13,749</td>
</tr>
<tr>
<td>10</td>
<td>17,530</td>
</tr>
<tr>
<td>11</td>
<td>131,355</td>
</tr>
<tr>
<td>12</td>
<td>1,207,625</td>
</tr>
<tr>
<td>13</td>
<td>34,459,425</td>
</tr>
<tr>
<td>14</td>
<td>654,729,705</td>
</tr>
<tr>
<td>15</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>16</td>
<td>316,334,143,225</td>
</tr>
<tr>
<td>17</td>
<td>7,905,853,580,625</td>
</tr>
<tr>
<td>18</td>
<td>213,458,046,676,875</td>
</tr>
<tr>
<td>19</td>
<td>6,190,283,353,629,375</td>
</tr>
<tr>
<td>20</td>
<td>191,839,783,962,510,625</td>
</tr>
<tr>
<td>21</td>
<td>6,332,659,870,762,850,625</td>
</tr>
<tr>
<td>22</td>
<td>221,643,095,476,791,771,875</td>
</tr>
</tbody>
</table>

Modified from Satoh (1996), SHINKA Vol. 6, No. 3

Table 1 The number of possible topology of rooted trees

The number of possible tree topologies rapidly increases as the number of OTUs increases (Panel 68: Table 1). Choice of one correct topology among them is the purpose of phylogenetic tree construction.

The number of possible rooted tree topology for $n$ OTUs is $1 \times 3 \times 5 \times \cdots \times (2n - 3)$

In the case of unrooted trees, $n$ is replaced with $n-1$ (Panel 68: Figure 11.2).

6.2.2. Gene tree and species tree

Phylogenetic trees reconstructed from sequence information of genes are called gene trees. Gene trees are not necessarily matched with species tree which reflects species phylogeny. In species trees, nodes correspond to the time of reproductive isolation (speciation) between two species. In gene trees, until the time of species separation, the genes were present as alleles. Thus, divergence nodes of gene trees are generally older than those of species trees. If time interval between two speciation events is short, tree topology can differ between gene and species trees (Panel 69: Figure 11). In such case, it is necessary to study many gene loci,
and to choose the topology which is supported by more number of loci (Panel 70).

6.2.3. Statistical reliability

The nucleotide (amino acid) substitution is a stochastic event at each site. Thus, when sequence length to compare is short, statistical fluctuation becomes larger. Thus, it is necessary to evaluate its statistical reliability by using standard error or bootstrap test (to be explained in page 104). To increase reliability, longer sequences should be used.

6.2.4. Gene duplication and speciation

It is not unusual that both gene duplication and speciation are involved in a gene phylogeny (Panel 69: Figure 9). Genes separated by a speciation are called **orthologous**. Genes separated by a gene duplication are called **paralogous**. In a case such as in Panel 69: Figure 9 (a), if only Gene 1 is sampled from species A and only
Gene 4 is sampled from species B, it is possible that the time of gene duplication is mistaken as the time of speciation. Then, species divergence time is overestimated. Hence, it is important to be cautious about the orthology and paralogy among samples studied.

6.2.5. Gene conversion

When a stretch of DNA sequence in a gene is replaced by that in a paralogous gene, the phenomenon is called gene conversion. It can occur within a chromosome and between chromosomes. If a gene conversion occurs, the two sequences become identical in the converted region, thus lowering overall difference between the genes. This phenomenon is also called concerted evolution. If it happens, gene tree is erroneously estimated (Panel 71: Figure 10). To reach a correct gene phylogeny, it would be effective to draw gene trees separately for different regions of the gene. If one tree shows older origin of gene duplication than speciation and another tree shows independent duplications in each species, the former tree can be regarded to reflect the true phylogenetic relationship. The latter tree can be derived from the former by gene conversion. But, if the latter tree is true, the former tree cannot realize (unless horizontal gene transfer occurs between species).

Panel 71

6.2.6. Recombination within a gene

Panel 71: Figure 13 exemplifies a generation of new alleles C and D by recombination of A and B. In this situation, a network is the appropriate method to illustrate the relationship among the four.

6.3. Phylogenetic tree making methods

There are three methods to draw phylogenetic trees: distance matrix methods, maximum parsimony (MP) methods, and maximum likelihood (ML) methods. Distance matrix is a table which shows distances (such
as \( n_{ij}, p_{ij}, \) and \( d_{ij} \) between two sequences for all pairs (Panel 72: Table 11.1 is an example). Distance matrix methods use the evolutionary distance \( d \). MP methods choose a tree which requires minimum number of substitutions to explain the sequence differences. ML methods choose a tree which exhibits the highest probability to explain the sequence differences under a likelihood function on a substitution model.

A popular distance matrix method is the **neighbor-joining (NJ) method** (Panel 72: Figure 5 is an example). The principle of NJ method is to repeat a search for neighbors which minimize the total branch length. NJ method does not assume constancy of evolutionary rate. It is unrooted and requires an out group to obtain a rooted tree. **UPGMA method** is the simplest distance matrix method (Panel 73) (Panel 72: Figure 11.3 is an example). UPGMA assumes constant evolutionary rate. It assumes the most distant OUT to be the out group and gives a rooted tree.

Panel 72

### Table 11.1 Proportion of different nucleotides \( (p) \) (above the diagonal) and estimates of the number of nucleotide substitutions \( (d) \) per site and their standard errors (below the diagonal) obtained from nucleotide sequence data for five primate species. Data from Brown et al. (1982).

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Chimpanzee</th>
<th>Gorilla</th>
<th>Orangutan</th>
<th>Gibbon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>.088</td>
<td>.103</td>
<td>.160</td>
<td>.181</td>
<td></td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>.094 ± .011</td>
<td>.106</td>
<td>.170</td>
<td>.189</td>
<td></td>
</tr>
<tr>
<td>Gorilla</td>
<td>.111 ± .012</td>
<td>.115 ± .012</td>
<td>.166</td>
<td>.189</td>
<td></td>
</tr>
<tr>
<td>Orangutan</td>
<td>.180 ± .016</td>
<td>.194 ± .016</td>
<td>.188 ± .016</td>
<td>.188</td>
<td></td>
</tr>
<tr>
<td>Gibbon</td>
<td>.207 ± .017</td>
<td>.218 ± .017</td>
<td>.218 ± .017</td>
<td>.216 ± .017</td>
<td></td>
</tr>
</tbody>
</table>

Molecular Evolutionary Genetics, M. Nei, Columbia University Press (1987)

Figure 11.3. Phylogenetic tree reconstructed by UPGMA from the distance matrix in Table 11.1. The hatched bar represents one standard error on each side of the branching point.

Figure 5.—The phylogenetic tree for retinal (RH1, RH2, SWS1, SWS2, and LWS) and pineal gland-specific (P) pigments of pigeon, chicken, and American chameleon. The bootstrap supports are indicated next to branch nodes. Values after P indicate \( \lambda \) max values.

### 6.3.1. UPGMA method

Refer to Panel 73. Assume a distance matrix for four OTUs 1, 2, 3, and 4. We first connect the minimum distance pair. Assume the distance of 3-4 pair, \( d_{34} \), is minimal. Let 3 and 4 be clustered to make OTU (34). Assume its branch length to be half of \( d_{34} \). Rewrite the matrix with OTU (34). Look for the minimum. Assume \( d_{2(34)} \) is minimal. Let 2 and (34) be clustered. Its branch length is half of \( d_{2(34)} \). Lastly, let 1 and (234) be clustered. Branch length is half of \( d_{1(234)} \).
6.3.2. Maximum parsimony (MP)
Looking into an alignment, focus on only informative sites which have information on classifying OTUs. Exclude invariant sites. Exclude singleton sites, too. Panel 74: Figure 11.10 is an example. Dotted 90 sites are informative. These sites are classified according to polytypic pattern as in Panel 74:Table 11.2. 105 rooted trees are possible for 5 OTUs, but in this case only 4 topologies (Panel 74:Table 11.3) are considered because of taxonomic information on these five species. The polytypic patterns are classified into compatible topologies. Then, count the minimum number of substitutions in each topology. The least substitution tree (topology B) is chosen as the MP tree (Panel 74: Figure 11.11).

Panel 73

![Branch lengths in UPGMA diagram]

\[
d_{12} + d_{13} + d_{14} = \frac{d_{13} + d_{14}}{2}, \quad d_{23} = \frac{d_{23} + d_{24}}{2}, \quad d_{34} = \frac{d_{13} + d_{14}}{2}\]

Panel 74

![Nucleotide sequences of mitochondrial DNAs from humans and apes. Here, only polytypic sites are presented. In chimpanzees, gorillas, orangutans, and gibbons, the nucleotides that are identical with those of humans are not shown. Nucleotide sites with dots are informative sites. Data from Brown et al. (1982).]

![Table 11.2 Polytypic patterns and their frequencies in informative sites for determining the parsimonious tree for the human (H), chimpanzee (C), gorilla (G), orangutan (O), and gibbon (B). Singular mutations are listed separately.]

<table>
<thead>
<tr>
<th>Polytypic pattern</th>
<th>Frequency</th>
<th>Polytypic pattern</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Informative sites</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. (H)–(G)GB</td>
<td>10</td>
<td>i. (G)–(H)CG</td>
<td>8</td>
</tr>
<tr>
<td>b. (H)–(O)GB</td>
<td>5</td>
<td>j. (G)–(H)CG</td>
<td>29</td>
</tr>
<tr>
<td>c. (H)–(O)GB</td>
<td>2</td>
<td>k. (O)–(H)CG</td>
<td>1</td>
</tr>
<tr>
<td>d. (O)–(G)GB</td>
<td>4</td>
<td>l. (O)–(H)CG</td>
<td>2</td>
</tr>
<tr>
<td>e. (G)–(H)GB</td>
<td>10</td>
<td>m. (G)–(O)CG</td>
<td>1</td>
</tr>
<tr>
<td>f. (O)–(G)GB</td>
<td>2</td>
<td>n. (B)–(G)CG</td>
<td>2</td>
</tr>
<tr>
<td>g. (O)–(H)GB</td>
<td>4</td>
<td>o. (B)–(O)CG</td>
<td>1</td>
</tr>
<tr>
<td>h. (G)–(B)GB</td>
<td>7</td>
<td>Total</td>
<td>90</td>
</tr>
<tr>
<td>II. Singular mutations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p. Human</td>
<td>29</td>
<td>q. Orangutan</td>
<td>59</td>
</tr>
<tr>
<td>q. Chimpanzee</td>
<td>29</td>
<td>r. Gibbon</td>
<td>79</td>
</tr>
<tr>
<td>r. Gorilla</td>
<td>26</td>
<td>Total</td>
<td>210</td>
</tr>
</tbody>
</table>

![Table 11.3 Compatible polytypic patterns with the four topologies given in figure 11.11 and the minimum number of nucleotide substitutions required for each topology.]

<table>
<thead>
<tr>
<th>Topology</th>
<th>Compatible polytypic patterns</th>
<th>Number of compatible sites</th>
<th>Minimum number of substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. HCGOB</td>
<td>a, j, k</td>
<td>42</td>
<td>147</td>
</tr>
<tr>
<td>B. CGHOB</td>
<td>c, j, k, m, o</td>
<td>44</td>
<td>153</td>
</tr>
<tr>
<td>C. HGOB</td>
<td>b, j, k, l, n</td>
<td>44</td>
<td>148</td>
</tr>
<tr>
<td>D. CGHOB</td>
<td>d, c, m, o</td>
<td>16</td>
<td>175</td>
</tr>
</tbody>
</table>
Using the same dataset, UPGMA supports the topology A (Panel 72: Figure 11.3). So does NJ method. MP is not compatible with these methods. However, difference of topologies A and B is only 2. In the UGMA tree, nodes $a$ and $b$ are not separated with statistical significance. This example alerts that we should be cautious on interpretation of phylogenetic trees reconstructed.

### 6.4. Statistical reliability of phylogenetic trees

#### 6.4.1. Standard error method

Using the standard error (SE) of evolutionary distance $d$, we can evaluate whether a branch length between two nodes is significantly larger than 0. The SE values of UPGMA branch lengths are relatively easy to calculate. Calculation is much more complex in other tree-making methods. However, today, SE calculation is often implemented in software for molecular evolutionary analysis and non-experts can easily use it.

#### 6.4.2. Bootstrap method

Suppose we have $n$ sequences consisting of $m$ sites. Choose one site randomly $m$ times, allowing the same site twice or more, from the alignment to make a new sequence set. For this sequence set, construct a phylogenetic tree. Repeat sampling sequence sets and making trees (generally 100–1000 times). Count how many times the original clusters are restored. The restored proportion (%) is called the bootstrap probability. Generally, probabilities over 80% are regarded high. Bootstrap probabilities are indicated to each node in a phylogenetic tree (Panel 72: Figure 5).

### 6.5. Test evolutionary rate constancy (Relative rate test)

It is important to clarify whether or not evolutionary rate is same between two lineages after their separation when we choose an appropriate tree-making method, estimate divergence time, or infer natural selection. To evaluate the constancy of evolutionary rate, the relative rate test is conducted. As exemplified in Panel 75, suppose that species 1 and 2 separated at the node 0 and species 3 is an out group. The relative rate test evaluates whether distance 1—3 and distance 2—3 differ statistically significantly. Tajima's relative rate test (Tajima 1993 Genetics 135:599-607) is often used which is irrespective of the nucleotide and amino acid substitution models and whether or not the substitution rate varies with site (implemented in MEGA).
6.6. Comparison of tree-making methods

Let a computer simulate the sequence evolution along a model phylogeny under varied evolutionary rates and substitution patterns. For the sequence sets obtained, phylogenetic trees are reconstructed using various tree-making methods and are compared with the model tree. By doing so, we can compare the accuracy of tree-making among these methods.

When evolutionary rate varies greatly or multiple substitutions are often, the maximum parsimony (MP) tends to choose a wrong topology. Even when evolutionary rate is not constant, the neighbor-joining (NJ) and maximum likelihood (ML) frequently estimate a true tree. A drawback of ML is enormous amount of computation time. But, assumptions are explicit in ML and performance can be improved by correcting assumptions (models) to fit to observation. UPGMA is applicable in the case that evolutionary rate is constant.

In MP and ML, possible number of trees to examine is enormous if OTUs are many (over 2 million for 10 OTUs: Panel 68: Table 1). A regular computer may not be able process it. In this sense, NJ can be a practical method.

In the case we have so many multiple substitutions including parallel and reverse substitutions, evolutionary distance cannot be estimated correctly. In such case and in the case that evolutionary rate variation is too large, even NJ cannot estimate correctly. There is no almighty method. Even if we have high bootstrap values, we have to keep in mind that gene trees are estimates at any rate.

6.7. Summary of notes to obtain reliable phylogenetic trees

(1) Choice of sequence: Use as long sequences as possible. For comparison of closely related species, choose regions of high evolutionary rate [such as non-coding regions, synonymous sites (nucleotide sites where nucleotide changes do not alter amino acids), those in mitochondrial DNA]. For distant species, choose regions of low evolutionary rate [such as nonsynonymous sites (nucleotide sites where nucleotide changes alter amino acids), amino acid sequence, those in highly conserved genes, ribosomal RNA genes]. After substitutions are saturated, correct evolutionary distance cannot be estimated. Do not include sequences irrelevant to the research purpose.

(2) Choice of distance estimation method: If there is a bias in nucleotide content or transition/transversion ratio, we should choose a method which is designed to correct multiple substitutions under such conditions. If rate constancy is not guaranteed, we should use a method which does not assume rate constancy, such as NJ.

(3) Carry out a statistical test such as bootstrap.
(4) Choice of tree-making method: Apply two or more methods.
(5) Do not forget the cautions explained in pages 98-101 when interpreting the results.

6.8. Evolutionary genetics Software
Several free software of evolutionary genetic analyses are available through internet which perform phylogenetic tree construction or other analyses. Below are examples.

(1) MEGA (http://www.megasoftware.net/)
Developed by Koichiro Tamura (Tokyo Metropolitan University) and others. Several options of tree-making methods are available (such as NJ, ML, MP). Many analytical methods are supported such as Synonymous/nonsynonymous analysis, relative rate test, ancestral sequence estimation. Help menu is useful. Conveniently, corresponding page numbers of Nei and Kumar's "Molecular Evolution and Phylogenetics" (2000) are given for explanation. Updated relatively frequently and user-friendly overall.

(2) PHYLIP (http://evolution.genetics.washington.edu/phylip.html)
Developed by Joseph Felsenstein (University of Washington). Includes MP and NJ. Can perform bootstrap.

(3) PAML (http://abacus.gene.ucl.ac.uk/software/)
Developed by Ziheng Yang (University College London). Can use various programs of ML methods.

(4) DnaSP (http://www.ub.edu/dnasp/)
Population genetic tools available.

(5) Arlequin (http://cmpg.unibe.ch/software/arlequin3/)
Population genetic tools available.

(6) Phylogeny Programs (http://evolution.genetics.washington.edu/phylip/software.html)
Link collections on phylogenetic programs offered by Felsenstein.

When writing a paper, you should describe with references which method of distance estimation, tree-making, or statistical evaluation is used. Then, mention the software name. Example: "The number of nucleotide substitutions per site \((d)\) for two sequences was estimated by the method of Tamura and Nei (1993), and the phylogenetic tree was reconstructed by applying the neighbor-joining method (Saitou & Nei, 1987) to the \(d\) values using the MEGA program version X (Kumar et al., 2018). The reliability of the tree topology was evaluated by bootstrap analysis with 1,000 replications (Felsenstein, 1985)."
Chapter 7 The Neutral Theory

We have already used the neutral model in this lecture. The neutral model explains how genetic variation arises and persists within and between species by chance alone. By providing a limit of chance effect, the neutral model works as a null hypothesis to judge if natural selection operated or not. The neutral model already existed before the neutral theory was proposed. The neutral theory is a claim that majority of genetic evolution and variation is explained by the neutral model (by random genetic drift of mutations neutral to natural selection). This chapter introduces the neutral theory including its history.

After 1960's, genetic polymorphic data started to emerge through electrophoretic analysis of enzyme proteins. Amino acid sequence data also appeared. Subsequently, at nucleotide sequence level, within-species polymorphism and between-species divergence data explosively increased. These data showed:

(1) Inter-species sequence divergence reflects phylogenetic divergence between species very well (constancy of evolutionary rate: Panel 76).

Panel 76

---

(2) Evolutionary rate is slow in the regions of proteins and genes where functional constraint is strict (conservativeness of evolutionary rate: Panel 77).

Panel 77

Table 4.1. Evolutionary rates in terms of amino acid substitutions. These rates are based mostly on data from mammalian order. They are expressed per amino acid site per year taking $10^{-9}$ as the unit (‘pauling’)

<table>
<thead>
<tr>
<th>Proteins</th>
<th>$k_{aa} \times 10^9$/yr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibrinopeptides</td>
<td>8.3</td>
</tr>
<tr>
<td>Pancreatic ribonuclease</td>
<td>2.1</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>2.0</td>
</tr>
<tr>
<td>Hemoglobin α</td>
<td>1.2</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>0.89</td>
</tr>
<tr>
<td>Insulin</td>
<td>0.44</td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>0.3</td>
</tr>
<tr>
<td>Histone H4</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Figure 1-13. Amino acid substitution rate in proinsulin molecular evolution

(3) Within-species genetic variation is much larger than previously thought.
(4) Nucleotide substitution rate is estimated to be about 1 per year per genome which is unexpectedly high.

These are very difficult to explain by natural selection only. Then, Motoo Kimura proposed the **Neutral Theory** of Molecular Evolution (Nature, 217: 624-626, 1968).

7.1. Evolutionary rate constancy and Neutral Theory

Compare the equation (5.3) $\lambda = v$ and the equation (5.4) $\lambda = 4N_s v$. The neutral evolutionary rate is influenced by mutation rate only. On the other hand, in the case natural selection operates, evolutionary rate is affected not only by mutation rate but also by population size and selection strength. These should vary with gene function, habitat environment, and species. Hence, constancy is far realistic in the neutral case.

The evolutionary rate constancy is observed to be per year but not per generation. Thus, the **Neutral Theory** predicts from the equation (5.3) that

(A) The mutation rate would show constancy per year but not per generation.

(B) If per-year mutation rate differ between lineages, evolutionary rate also differ by the same extent.

Notes regarding (A): **Phenotypic mutations** (such as visible mutations and lethal mutations) are known to show constancy per generation. Phenotypic mutations involve not only simple nucleotide mutations but also other causes such as transposon insertions or epistatic interaction of multiple genes’ mutations. **It is not well**
understood why phenotypic mutations show constancy per generation, which is still a research topic.

On the other hand, per-year constancy of nucleotide mutations is easier to understand. The main cause of nucleotide mutations are errors of DNA replication in germ cells. The errors are expected to increase as cell divisions increase. The number of germ cell divisions is roughly proportional to generation time. Thus, the number of cell divisions (and number of mutations) is expected to be roughly the same per year between species of different generation time.

Notes regarding (B): If the nucleotide mutation rate correlates linearly with cell divisions of germ cells, evolutionary rate would differ between sex chromosomes and between sex chromosomes and autosomes. This is because male germ cells undergo through many more divisions than female germ cells. The Y chromosome would be fastest which goes through only males. The autosomes would be intermediate which go through males and females half and half. The X chromosome would be slowest which go through females in 2/3 and males in 1/3 (Panel 78).

Panel 78

Evolutionary rate difference among sex chromosomes and autosomes by male-driven evolution (in XY system)

\[ r = \frac{V_m}{V_f}, \]

\[
\begin{align*}
\sigma (XY) & \text{ mutation rate } V_m \quad \text{X chromosome evolutionary rate: } X = \frac{1}{3} V_m + \frac{2}{3} V_f \\
\varphi (XX) & \text{ mutation rate } V_f \quad \text{Y chromosome evolutionary rate: } Y = V_m \\
\frac{V_m}{V_f} & \quad \text{Autosome evolutionary rate: } A = \frac{1}{2} V_m + \frac{1}{2} V_f \\
\end{align*}
\]

\[
\begin{align*}
X &= \frac{1}{3} V_m + \frac{2}{3} V_f = 2 \left( \frac{V_m + 2 V_f}{V_m + V_f} \right) = \frac{2}{3} \left( \frac{r + 2}{r + 1} \right) \quad 1 \to \infty \implies 1 \to \frac{2}{3} \quad A > X \\
Y &= \frac{V_m}{V_f} = 2 \left( \frac{V_m}{V_m + V_f} \right) = \frac{2}{3} \left( \frac{r}{r + 1} \right) \quad 1 \to 2 \quad Y > A \\
\frac{V_m}{V_f} &= 3 \left( \frac{V_m}{V_m + 2 V_f} \right) = \frac{3}{3} \left( \frac{r}{r + 2} \right) \quad 1 \to 3 \quad Y > X \\
\end{align*}
\]

Let \( r \) be ratio of mutation rate of male to female. Let \( X, Y, \) and \( A \) be evolutionary rate of X chromosome, Y chromosome, and autosomes, respectively. Then, as in Panel 78, \( Y > A > X \) is predicted.

By comparing nucleotide sequences of an intron of ZFY gene and ZFX gene of primates and rodents, \( Y/X \) is calculated to 2.25 in primates and 1.42 in rodents. Thus, according to \( \frac{Y}{X} = 3 \left( \frac{r}{r+2} \right), r \) is estimated to be 6 in primates and 1.8 in rodents. In other genes, \( r \) is 3~6 in primates and about 2 in rodents. The observed male to
female ratio of germ cell divisions is 3~6 in humans and about 2 in mice and rats, being very concordant with \( r \) value.

The result implies that evolutionary rate is determined mainly by males. This phenomenon is called \textbf{male-driven evolution}. However, this XY-system result can also be explained by the difference of selective constraint on these chromosomes. Because X chromosome is haploid in males, deleterious mutations are directly exposed to natural selection in males. Thus, X chromosome genes are expected to evolve most slowly. On the other hand, Y chromosome is absent in females and its genes are expected to be not essential for survival. Deleterious mutations in Y chromosome would be less susceptible to natural selection.

Autosomes are in between X and Y. Thus, without assuming male-driven evolution, \( Y > A > X \) can be explained.

To test the male-driven hypothesis, ZW-system was examined, which is popular in birds and others. Males are ZZ and females are ZW. By analogy of XY system, the selective constraint hypothesis predicts \( W > A > Z \). On the other hand, the male-driven hypothesis predicts \( Z > A > W \) (Panel 79). A study on bird genes showed \( Z > A > W \). Then, male-driven evolution is supported. Refer to Graur and Li's "Fundamentals of Molecular Evolution" for more details.

By showing that the mutation rate determines the evolutionary rate, these studies support the claim of the neutral theory that majority of gene evolution occurs by neutral mutations.

\textbf{Evolutionary rate difference among sex chromosomes and autosomes by male-driven evolution (in ZW system)}

\[
\begin{align*}
\sigma^\theta &= \frac{V_m}{V_f} \\
\Psi &= \frac{V_m}{V_f} \\
\frac{Z}{A} &= \frac{2V_m + \frac{1}{3}V_f}{\frac{2}{3}V_m + \frac{1}{2}V_f} = \frac{2}{3} \left( \frac{2V_m + V_f}{V_m + V_f} \right) = \frac{2}{3} \left( \frac{2r + 1}{r + 1} \right) \quad \Rightarrow 1 \rightarrow \frac{4}{3} \quad Z > A \\
\frac{W}{A} &= \frac{V_f}{\frac{2}{3}V_m + \frac{1}{2}V_f} = \frac{V_f}{V_m + V_f} = 2 \left( \frac{V_f}{V_m + V_f} \right) = 2 \left( \frac{1}{r + 1} \right) \quad \Rightarrow 1 \rightarrow 0 \quad A > W \\
\frac{Z}{W} &= \frac{2V_m + \frac{1}{3}V_f}{\frac{2}{3}V_m + \frac{1}{2}V_f} = \frac{1}{3} \left( 2V_m + 1 \right) = \frac{1}{3} (2r + 1) \quad \Rightarrow 1 \rightarrow \infty \quad Z > W
\end{align*}
\]

\( Z > A > W \)
7.2. Conservativeness of evolutionary rate and Neutral Theory

If nucleotide substitutions are accumulation of advantageous mutations alone, functionally important
genes/regions are expected to have accumulated more mutations and the less important ones are not evolved
much. But, observation is opposite: functionally important genes/regions evolve slowly and less important
ones evolve fast.

Rewrite the equation (5.3) \( \lambda = v \) as

\[ \dot{\lambda} = f_0 v_T \]

(7.1)

where \( v_T \) is total mutation rate and \( f_0 \) is the fraction of neutral mutation. The fraction \( 1-f_0 \) is assumed to be
mostly deleterious and not to contribute to nucleotide substitution.

Functionally important regions are more likely to be damaged by mutations. Less mutations are expected to
be functionally equivalent (i.e. neutral) with the original nucleotide (or amino acid). That is, \( f_0 \) is smaller in
functionally important regions. When \( f_0 \) is 1, any mutation is equivalent with the original (i.e. not deleterious)
and evolutionary rate \( \lambda \) reaches to the maximum. The evolutionary rate has a upper limit given by total
mutation rate: \( \lambda \leq v_T \).

Generally in nonsynonymous sites, \( f_0 \) is small and evolutionary rate is slow. Functional constraint differs
among genes and evolutionary rate differs greatly accordingly (Panels 67, 77). On the other hand, in
synonymous sites, \( f_0 \) is larger and less variable among genes (Panel 67: Figure 16). Yet, in synonymous sites
\( f_0 \) is not 1 and some other constraints are operating (see (3) in page 3). So is not 1 in introns and noncoding
regions. Free from constraint are functionless pseudogenes (dead genes). Irrespective of synonymous or
nonsynonymous sites, evolutionary rate is very high and reflects total mutation rate in pseudogenes (Panel
67: Figure 17).

In equation (7.1), advantageous mutations are ignored because they are considered rare. The evolutionary
rate of advantageous mutations is, in fact, higher than the neutral maximum \( v_T \). The reason is explained in the
next chapter at Panel 80.

Notes: Neutral mutations are not mutations which occur in functionless regions (such as pseudogenes). But
are mutations of which the effect is equivalent with the original in function and fitness. Neutrality is
determined by the product of fitness effect of mutations and population size. If \( |N_c s| \ll 1 \), irrespective that
selection coefficient is not 0 and the mutation is advantageous or deleterious, the fixation probability is

\[ u = \frac{1 - e^{-2s}}{1 - e^{-4N_c s}} \approx \frac{1 - (1 - 2s)}{1 - (1 - 4N_c s)} = \frac{2s}{4N_c s} = \frac{1}{2N_c} \]
In this situation, fate of mutations is determined mostly by chance as in the case of neutral mutations.

\[ |N_e s| \ll 1 \Leftrightarrow |s| \ll \frac{1}{N_e} \]

Effect of selection is smaller than effect of random genetic drift.

\[ |N_e s| \gg 1 \Leftrightarrow |s| \gg \frac{1}{N_e} \]

Effect of selection is larger than effect of random genetic drift.

This idea is called “Nearly Neutral Theory” developed by Tomoko Ohta (Nature, 246: 96–98, 1973).

7.3. Within-species variation, genome-level nucleotide substitution rate and Neutral Theory

The genetic variation features (3), high within-species variation, and (4), high nucleotide substitution rate at the genome level, cannot be explained with positive selection alone by the following reasons. As explained earlier in page 15, a positive selection to mutations (minorities) is a negative selection to majority,

extinguishing major fraction of lower-fitness individuals from a population. To maintain population size, with the observed level of variation being all by positive selection, each individual needs to reproduce a large number of offspring. This is called cost of natural selection or substitution load.

As we have seen, nucleotide substitution rate in nuclear genes is in the order of $10^{-9}$ per site per year. Thus, at genome level with the order of $10^9$ nucleotides, the rate is several substitutions per year. Every year we have some substitutions somewhere in the genome. Suppose all of these are driven by positive selection, for one offspring to survive and reproduce, each parent needs to yield over 3 million offspring.

Early electrophoresis data of enzyme proteins revealed enormous genetic polymorphism in human and fruit fly populations: each individual is heterozygous in over 1000 loci. If this is explained by overdominance selection alone, each individual need to reproduce unrealistically large number of offspring. For more detail, refer to Kimura (1968) Nature 217:624-626 or Kimura's "The Neutral Theory of Molecular Evolution". In contrast, in the neutral model, the large amount of polymorphism can be realized without any load by the balance between mutation and random genetic drift alone. Refer to Panel 28: Figure 12 again.
Chapter 8 Adaptive Evolution

The Neutral Theory claims that the majority of mutations in the genome is neutral or deleterious and the majority of nucleotide substitutions is neutral. However, the theory does not rule out the presence of advantageous (adaptive) mutations and their contribution to evolution of life. Motoo Kimura, the founder of Neutral Theory, stressed this point in various occasions.

Adaptive mutations are supposed to have appeared and been tested one after another since early days of evolution, to create the fundamental machineries of life. Once the machineries are established, subsequent mutations are supposed to be mostly disturbing to the system, i.e. deleterious. In other words, as basic machineries got completed in early life evolution, the proportion of novel advantageous mutations is supposed to be decreased. The evolutionary conservativeness of genes with strict functional constraint is the reflection of negative selection to deleterious minority mutations. This is the other side of positive selection to the established majority.

However, besides the fundamental machineries, diverse adaptive evolution has been continuing. Though rare, but, once emerge, advantageous mutations have a much higher fixation probability than neutral mutations and contribute significantly to the adaptive evolution of life. As mentioned in page 15, the positive selection to minorities could transform life to adapt to changing environments. In this chapter, I shall overview methods to detect positive selection from sequence data. Hereafter, the positive selection is meant to be the positive selection to minorities.

8.1. Methods to detect positive selection
The detection of positive selection is basically the detection of deviation from the neutral model. To do it, we need to know the neutral mode. This lecture has been conducted to learn the neutral model. There are roughly three approaches to detect deviation from the neutral model: (1) focusing on within-species diversity, (2) focusing on between-species divergence, and (3) comparing diversity and divergence.

8.1.1. Methods focusing on within-species diversity
This is actually already explained in Chapters 3 and 4. Let me classify the methods to the followings.

8.1.1.1. <Population differentiation>
We compare $F_{ST}$ (also called $G_{ST}$) between a test gene and neutral reference regions (such as pseudogenes, introns, non-genic regions, etc.) (or random choice of multiple regions in the genome: expected to be neutral according to the Neutral Theory). If population differentiation of test gene is significantly larger than that of
the reference, a subpopulation is inferred to be under positive selection directing to fixation of an allele type. The lactase gene in humans is a good example. On the contrary, if population differentiation is significantly smaller than reference, a similar balancing selection is inferred in different populations at the gene.

8.1.2. <Summary statistics>

Compare various summary statistics [nucleotide diversity, the number of segregating sites, frequency spectra, linkage disequilibrium (LD), etc.] between a test gene and neutral references (or random choice of many genome regions).

Allele frequency spectrum: Ewens-Watterson test and others
Nucleotide frequency spectrum: Tajima's $D$, Fu and Li's $G$, $F$, $D$, Fay and Wu's $H$, and others

After estimating various demographic parameters by coalescence simulation using multiple neutral references, evaluate the deviation of various summary statistics from neutral prediction for a test gene.

Recent advance of genome data has enabled us to detect selective sweep and other characteristic regions through a genome scan for long LD block, high homozygosity block, etc. by using Extended Haplotype Homozygosity (EHH) test or other methods.

8.1.2. Methods focusing on inter-species divergence

The $d_N / d_S$ test and detection of convergent evolution are representative methods.

8.1.2.1. <$d_N / d_S$ test>

In this method, the evolutionary distance $d$ between species is calculated separately to the number of nonsynonymous substitutions per nonsynonymous site ($d_N$) and the number of synonymous substitutions per synonymous site ($d_S$). Then, we compare $d_N$ with $d_S$. We regard $d_S$ as reflecting the neutral evolutionary rate (total mutation rate) and $d_N$ as reflecting decelerated or accelerated evolutionary rate by natural selection. The following simple scheme is often used:

\[
\frac{d_N}{d_S} < 1 \rightarrow \text{negative selection, under functional constraint, implying the gene is functioning}
\]

\[
\frac{d_N}{d_S} = 1 \rightarrow \text{neutral, no functional constraint, implying the gene could be a pseudogene}
\]

\[
\frac{d_N}{d_S} > 1 \rightarrow \text{positive selection, implying the gene is acquiring (acquired) a new function}
\]

Why can $d_N$ exceed $d_S$ under positive selection? Didn’t equation (7.1), $\lambda = f_0 v_T$, tell us that the total mutation rate $v_T$ is the upper limit of evolutionary rate? This is due to the higher fixation probability of
adaptive mutations than of neutral mutations as explained in the followings (see also panel 80).

Consider one codon. Suppose population size is $N$. For simplicity, let's assume that the third site is synonymous for any nucleotide mutation and is free from functional constraint. Thus its evolutionary rate is supposed to be the total mutation rate $v_T$. The interval between substitutions is $1/v_T$ years during which one among $2N$ mutations are fixed ($2N v_T$ mutations every year for $1/v_T$ years) (Panel 80 uppermost: perfect neutrality).

At the second site, any mutation alters amino acid. Thus, the second site is under strict functional constraint. For simplicity, the first site is assumed to be nonsynonymous for any mutation as well. Because of amino acid change, many of mutations are deleterious. Hence, during $1/v_T$ years even one mutation cannot be fixed, i.e. evolutionary rate is decreased (Panel 80 middle: negative selection).

If advantageous mutations occur at the first or second site, their fixation probability is much larger than neutral mutations. Thus, during $1/v_T$ years more than one mutations can be fixed, i.e. evolutionary rate exceeds $v_T$ (Panel 80 bottom: positive selection).
As \( N \) increases, more advantageous mutations are expected to happen in \( 2N \) mutations and evolutionary rate increases. As for neutral mutations, irrespective of population size, one per \( 2N \) mutations is fixed in \( 1/vT \) years.

8.1.2.1.1. How to calculate \( d_N \) and \( d_S \)

Unweighted Pathway Method (Nei/Gojobori's method)

This is the simplest and easiest method. First, we need to determine how many synonymous sites and nonsynonymous sites are in a sequence. To do it, we assign nonsynonymous degree (synonymous degree) to each of three sites of a codon according to the codon table. Let's consider TTA (Leu) codon as an example. The nonsynonymous degree of the first site is \( 2/3 \), the second site is \( 1 \), and the third site is \( 2/3 \), in sum \( 7/3 \). Automatically, the synonymous degree of the first site is \( 1/3 \), the second site is \( 0 \), and the third site is \( 1/3 \), in sum \( 2/3 \). Total of two categories over 3 sites is of course 3. Do the same to all codons for each sequence. Sum of nonsynonymous degree through codons is the number of nonsynonymous sites of the sequence. So is the number of synonymous sites. If the number is different between sequences, we just take an average in Unweighted Pathway Method. The averaged numbers are the number nonsynonymous sites \( n_N \), and the number of synonymous sites \( n_S \) for a sequence pair \( (n_N + n_S = n; n \) is sequence length).

Next, we count the number of difference in a sequence pair codon by codon. Then, we determine the difference is nonsynonymous or synonymous. If there is only 1 difference in a codon, distinction is uniquely specified (e.g. for GTT (Val) and GTA (Val), one synonymous difference). If two sites differ between sequences, two ways of distinction are possible. For example, for TTT and GTA, two passes are:

Path 1  
TTT (Phe) ⇄ GTT (Val) ⇄ GTA (Val)

Path 2  
TTT (Phe) ⇄ TTA (Leu) ⇄ GTA (Val)

In path 1, we have 1 nonsynonymous change and 1 synonymous change. In pass 2, we have 2 nonsynonymous and no synonymous change. In Unweighted Pathway Method, we just take an average: 1.5 nonsynonymous differences and 0.5 synonymous difference. If there are 3 differences, 6 passes are possible, and we just take their average. Then, do the same for all codon pairs and sum up to gain the number of nonsynonymous difference \( n_{dN} \) and the number of synonymous difference \( n_{dS} \) for the sequence pair. Of course, \( n_{dN} + n_{dS} \) is equal to the number of nucleotide differences \( (n_d) \).

The number of nonsynonymous differences per nonsynonymous site \( (\rho_N) \) and the number of synonymous differences per synonymous \( (\rho_S) \) are given as:

\[
p_N = n_{dN}/n_N \quad p_S = n_{dS}/n_S
\]

\( d_N \) and \( d_S \) are obtained by multiple substitution correction such as by Jukes-Cantor method [equation (6.2)].
Several other methods to estimate $d_N$ and $d_S$ are developed including Weighted Pathway Methods (such as Modified Nei-Gojobori' method, Miyata-Yasunaga's method, Li's method). These are expected to give a better estimate. But, according to computer simulation, performance is not very different from Nei-Goobori's method. For more detail, refer to Nei and Kumar's "Molecular Evolution and Phylogenetics".

8.1.2.1.2. Statistical test of $d_N / d_S$

$d_N / d_S > 1$ and $d_N - d_S > 0$ are mathematically equivalent. $d_N$ and $d_S$ are independent estimators. Thus, the simplest test method of $d_N / d_S$ value is to evaluate whether $d_N - d_S$ is 0, positive or negative by Z test.

Software to calculate $d_N$ and $d_S$ and to conduct $d_N / d_S$ test is available through internet. See page.106.

The maximum likelihood method to evaluate $d_N / d_S$ values at every phylogenetic tree branch or at every codon (branch-site test) is often used (Yang and dos Reis 2011 Mol. Biol. Evol. 28:1217-1228). This method does not evaluate whether an observed $d_N / d_S$ value (designated as $\omega$) is significantly larger or smaller than 1. It sets up various null models on the distribution of $\omega$ values among nucleotide sites or among tree branches and evaluates whether the observed $\omega$ value distribution is deviated from the null models at a particular site or in a particular branch.

8.1.2.2. <Detection of convergent/parallel evolution>

"Convergent" in molecular evolution means that a substitution occurs to become a same nucleotide (amino acid) kind at a homologous site in two lineages independently. "Parallel" means that the original nucleotide (amino acid) is also the same kind. Thus, "parallel" is a special case of "convergent". Convergent evolution is possible at a site by chance in the inter-species evolutionary time scale (multiple substitutions: see "6.1.2. Evolutionary distance"). However, convergence in two or more closely located sites is very unlikely to happen by chance alone. Hence, if multiple convergent substitutions are detected in a region, it can support positive selection. For detail of statistical test, refer to Zhang and Kumar 1997, Mol Biol Evol 14: 527-536.

8.1.3. Methods comparing within-species diversity and inter-species divergence

Refer to Panel 81. Representative methods are Hudson-Kreitman-Aguade (HKA) test and McDonald-Kreitman (MK) test. We need sequence data from two species and population samples from both or one species. In HKA test, two unliked genome regions are examined (test region and reference region). In MK test, synonymous sites and nonsynonymous sites of a gene are examined.

When we align these sequences from two species, variable sites are classified into two states: (1) polymorphic (segregating) in one or both species and (2) fixed to different nucleotide kinds between two
species. The former is called **polymorphic sites and represents within-species diversity**, the latter is called fixed (or divergent) sites and represents inter-species divergence.

If neutral, the number of polymorphic (segregating) sites reflects the population mutation rate and the sample number \( E(s) = 4N_e\mu\sum_{i=1}^{n_i-1} \frac{1}{i} = \theta\sum_{i=1}^{n_i-1} \frac{1}{i} \) (see pages 52-57), while the number of fixed sites reflects the mutation rate (= evolutionary rate) and divergence time \( d = 2\mu T \) (see pages 92, 95). Under the same \( N_e, \mu, \) and \( T, \) if neutral, the ratio of the number of fixed sites to the number of polymorphic sites is expected to be the same between two regions. By setting one region as a neutral reference (synonymous sites in MK) and the other as a test (nonsynonymous sites in MK), we can test neutrality of the test region. \( N_e, \) differs among autosomes, sex chromosomes and mitochondrial DNA. \( \mu \) also differs among them and, strictly speaking, among regions within each. We need to consider these differences if necessary.

### Simplified HKA test

<table>
<thead>
<tr>
<th>Species A</th>
<th>Reference</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>- A - T - T - G - G - C -</td>
<td>- C - A - T - T - G - A -</td>
<td></td>
</tr>
<tr>
<td>- G - T - C - G - G - G -</td>
<td>- G - C - G - T - G - A -</td>
<td></td>
</tr>
<tr>
<td>- G - T - C - G - A - G -</td>
<td>- G - C - G - T - G - T -</td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Polymorphic</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species B</th>
<th>Reference</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>- G - T - C - A - A - G -</td>
<td>- G - C - G - T - G - T -</td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Polymorphic</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \frac{(1 \times 3 - 5 \times 3)^2}{6 \times 6 \times 8 \times 4} \times 12 = \frac{3}{2}
\]

\(d.f. = 1\)

\(0.2 < P < 0.3\)

No significant difference of Fixed/Polymorphic ratio between Reference and Test regions

In MK test, replace Reference with Synonymous, and Test with Nonsynonymous

A simple \(\chi^2\) test is conventionally conducted in an HKA test to evaluate whether the variable sites are "Fixed or Polymorphic" is independent from whether the variable sites are in "Reference or Test" regions. If the F/P ratio is not significantly different between reference and test regions, the independency and thus the neutrality is supported. More strictly, the number of polymorphic sites and fixed sites are estimated from the model, and deviation is evaluated of the observation from the model.
Likewise, a simple $\chi^2$ test is conventionally conducted in MK test to evaluate the independency between "Fixed or Polymorphic" and "Nonsynonymous or Synonymous" (Panel 82: Figure 3.7). Likewise again, more strictly, the number of polymorphic sites and fixed sites are estimated for nonsynonymous and synonymous sites from the model (Panel 83), and MK test is conducted by evaluating if the observation is deviated from the expectation or not (called Poisson random field method).

Panel 82

<table>
<thead>
<tr>
<th></th>
<th>$D. \text{ simulans}$ vs $D. \text{ yakuba}$</th>
<th>$D. \text{ melanogaster}$ vs $D. \text{ simulans}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$m = 6$</td>
<td>$n = 12$</td>
</tr>
<tr>
<td>Divergence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>$17$ (21.0)</td>
<td>$26$ (23.0)</td>
</tr>
<tr>
<td>Polymorphism</td>
<td>$29$ (25.0)</td>
<td>$21$ (23.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replacement</td>
<td>$6$ (6.1)</td>
<td>$36$ (36.2)</td>
</tr>
<tr>
<td></td>
<td>$0$ (0.9)</td>
<td>$2$ (1.8)</td>
</tr>
<tr>
<td>$X^2 (P)$</td>
<td>$8.6$ (0.0018)</td>
<td>$16.5$ (0.000026)</td>
</tr>
</tbody>
</table>

Panel 83

### Poisson Random Field approach (semidominant mutations)

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>1+s</td>
<td>1+2s</td>
</tr>
</tbody>
</table>

\[ S = 2N_{es} \]

<table>
<thead>
<tr>
<th></th>
<th>Polymorphism</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsynonymous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>$F(m) = \frac{1-x^n-(1-x)^n}{1-x} \frac{1-e^{-2Sx}}{2Sx} dx$</td>
<td>$G(m) = (1-x)^m - 1 - e^{-2Sx} \frac{2Sx}{2Sx} dx$</td>
</tr>
</tbody>
</table>

\[ Sawyer and Hartl, 1992 \]

 Courtesy of Dr. Naoki Osada, The National Institute of Genetics
HKA test is interpreted in the following way (Panel 84).

1) **When Fixed/Polymorphism ratio is larger in test than reference regions**

   Positive selection is suspected to operate on the test region and yield more fixations than neutral condition. Note there is a possibility that the reference region is, in fact, under a negative selection (due to RNA stability or some other functional constraints) and has fewer fixations than under truly neutral condition. Though even more unlikely, if ballancing selection operates on the reference region, it brings more polymorphic sites in the reference region than the neutral condition.

2) **When Fixed/Polymorphism ratio is smaller in test than reference regions**

   Ballancing selection is suspected to operate on the test region and maintain more polymorphism than neutral condition. Alternative interpretation is that negative selection operates on the test region and results in less fixation than neutral condition. Because functional genes are basically under negative selection, it is usually difficult to distinguish ballancing and negative selection with HKA test alone. Other methods, such as Tajima's $D$, is necessary to conduct together. Though unlikely to happen, note there is a possibility that positive selection operates on the reference region and yields more fixation than truly neutral condition.

We need to be careful and consider other methods when interpreting the result.
Interpretation of MK test is basically the same with HKA test. However, unlike HKA, because of close physical linkage between nonsynonymous and synonymous sites, the two variations are to behave similarly. For example, if selective sweep occurs due to a nonsynonymous mutation, nearby synonymous variations are also fixed together. In the case of balancing selection to a nonsynonymous variation, the nearby synonymous sites also have a long coalescence time and accumulate many mutations. In either case, unless selection-target nonsynonymous mutations are numerous, F/P ratio would not much differ between nonsynonymous and synonymous sites. Thus, generally, MK would be less sensitive than HKA to detect natural selection. See Panel 85.

Panel 85

**Problem of MK**
*(difference unclear between Syn and Nonsyn)*

![Diagram showing the problem of MK](image)

- **Synonymous**
  - Linked Syns extinguished together
  - Species A
  - Species B

- **Nonsynonymous**
  - An advantageous mutation S
  - Extinguished by S
  - Species A
  - Species B

- Fixed (inter-species divergence)
- Polymorphic (within-species diversity)
- Competing advantageous mutations
- Linked Syns maintained
8.2. Exemplary studies of positive selection detection

(1) L/M opsin genes of New World monkeys (Panel 86)

[Hiwatashi et al., 2010, Mol Biol Evol 27: 453–464]

New World monkeys, inhabiting Meso and South America, are highly polymorphic in color vision. This is primarily due to allelic variation of the X-linked single-locus L/M opsin gene. This study is aimed to test if the L/M opsin polymorphism is maintained by balancing selection by analyzing wild populations of capuchin monkeys and spider monkeys. First, distribution of \( \theta \) is estimated from the number of segregating sites and nucleotide diversity by rejection sampling method of coalescence simulation using multiple neutral reference loci (see pages 62). Then, by another coalescence simulation using the \( \theta \) distribution, the expected distributions of nucleotide diversity \( \pi \) and Tajima's \( D \) are obtained. The observations of \( \pi \) (Fig. 2) and Tajima's \( D \) (Fig. 4) of L/M opsin gene are significantly deviated from the expected distribution toward positive direction. This strongly supports that L/M opsin (color vision) polymorphism is maintained by balancing selection.

Panel 86

![Graphs](image_url)
MHC genes have so many alleles. The nucleotide difference among alleles is very high. Alleles are shared between species (cross-species polymorphism). To explain all of these characteristics including the observation of $d_S/d_S > 1$, balancing selection, in particular overdominance, is inferred to be operating. The larger the antigen-recognition repertoire becomes, the better the immune-function becomes, and the more the positive selection works toward this direction.
The huge repertoire of immunoglobulin (antibody) is created somatically by VDJ recombination and hyper mutation. In addition, V (variable) region genes accumulate many amino acid substitutions evolutionarily (in germ cells) as well. Within the V region, the hyper-variable "Complementarity-Determining Region (CDR)" binds directly to antigens and shows $d_N/d_S > 1$. In contrast, in other region of V, the Framework Region (FR) shows $d_S/d_S < 1$. This suggests that the positive selection is operating in CDR, as in MHC, to promote diversification.
Eosinophilic leukocyte RNase genes of primates (Panel 89)

[Zhang et al., 1998, Proc Natl Acad Sci USA 95: 3708-3713]

The Eosinophil Cationic Protein (ECP) and the Eosinophil-Derived Neurotoxin (EDN) are both RNase produced in the large granule of eosinophilic leukocytes. EDN has high RNase activity and is neurotoxic under non-physiological conditions. On the other hand, ECP has low RNase activity but is highly toxic to pathogenic bacteria and parasites. ECP was created from EDN at the common ancestor of humans and Old World monkeys. By inferring ancestral sequences and estimating $d_s/d_{\text{S}}$ at each branch of phylogenetic tree, significant $d_s/d_{\text{S}}>1$ is detected to ECP in the common ancestor of human and Old World monkeys. This is accompanied by large increase of arginine and is suspected to be relevant to acquisition of anti-pathogenic function.

Panel 89
(5) CS protein gene of malaria parasite (Plasmodium)

[Hughes, 1991, Genetics 127: 345-353]

The malaria pathogenic protozoa (*Plasmodium*) infect hosts at the sporozoite stage. The sporozoite cell-surface protein Circumsporozoit (CS) becomes an epitope presented by MHC to T cells. The epitope region of the CS protein of *Plasmodium falciparum* shows significant $d_N/d_S > 1$. This is thought to reflect positive selection to diversify parasite infectious ability against the immune system of the host.

(6) Neurotoxic conotoxin genes of poisonous snail (Panel 90)


The neurotoxic peptides conotoxins of *Conus*, a marine predatory snail, block diverse ion channels and neural receptors. A conotoxin is ~30 amino acids peptide (toxic region) excised from a precursor of 70–80 amino acids peptide. The precursor genes consist of multigene family and diverse conotoxins are produced. Among the multiple precursor genes of *Conus abbreviatus*, the toxic region shows significant $d_N/d_S > 1$ while the removed region (prepro region) does not.

Panel 90

![Fig. 2. Sliding window analysis of average $D_s$ and $D_e$ estimates for all toxin sequence comparisons. Codons 1–42 primarily include the prepro region; codons 43–63 only contain toxin sequences terminating before the stop codon; because the presumed cleavage site varies in position, window position 36–49 includes about a 1:1 ratio of prepro/toxin codons (see Table 1; gaps in the alignment are excluded from this codon-numbering scheme).](image-url)
Lysozymes are an enzyme which punctures bacteria. Most animals produce lysozymes. Generally, lysozymes are present in macrophages, tear, saliva, egg while of birds, mammalian milk. However, lysozyme is used as a digestive enzyme in three groups of animals [ruminant artiodactyls (e.g. cows), colobine primates, and Hoatzin (South American ground birds)]. They develop a specialized part of digestive tract, forestomach, where they keep bacteria and ferment leaves. In the next stomach, they digest bacteria using the lysozyme and take out fermented nutrition. The stomach lysozymes of the three animal groups independently acquire activity at low pH, resistance to protease, and expression in the stomach activity. Supporting this, convergent evolution at specific amino acid sites are frequently observed (Figure 7.8, TABLE 7.12).

In the phylogenetic tree, the primate lysozyme gene shows significant \( \frac{d_{N}}{d_{S}} > 1 \) at common ancestor of Colobinae. (FIG. 2). \( \frac{d_{N}}{d_{S}} > 1 \) is also detected in the common ancestor of hominoids. The meaning of this is not clear.
(8) Hybrid sterility gene OdsH gene of fruit fly
[Ting et al., 1998, Science 282: 1501-1504]

Hybrids of different species become sterile. In fruit fly, a hybrid-sterility gene OdsH evolves very fast even though they are a homeobox gene which generally evolves very slowly. This gene is expressed in testis in fruit fly, while it is expressed in neural tissues in mouse and nematode. In fruit fly, OdsH appears to regulate genes relevant to spermatogenesis, but remains to be proved. Inter-species comparison in fruit flies shows $d_S/d_S > 1$ and $d_S$ exceeds even intron evolutionary distance. OdsH in fruit fly appears to be related to speciation and reproductive isolation.

(9) Hemoglobin gene of Antarctic fish
[Bargelloni et al., 1998, Proc Natl Acad Sci USA 95: 8670-8675 and other works]

Notothenioids are a family of Antarctic bony fish and adapt to the cold temperature. Because of low temperature, the metabolic rate is low and the body fluid can hold high-concentration oxygen. Thus, these fish depend less on hemoglobin. In extreme, icefish lack hemoglobin. Generally vertebrates have multiple kinds of hemoglobin, but many Notothenioids have only one. Relaxation of functional constraint is expected to hemoglobin under such condition. Indeed, $d_S/d_S = 1$ is observed in 5 out of 6 species. However, in Gymnorchinaeutes which is closely related to icefish, significant $d_S/d_S > 1$ is observed. This implies that hemoglobin is remodeled in this species. But, its biological meaning remains unsolved.

8.3. Limitation and complementarity of detecting positive selection

8.3.1. $d_S/d_S$ test

In $d_S/d_S$ test, positive selection could not be detected unless many of nonsynonymous mutations are advantageous. The examples above are all such case. This holds for the situation where sequence difference itself is important, such as host defense, parasitism, reproductive isolation, and sexual selection, and the situation where many amino acid changes are required for functional change of a gene. However, in other situations, this would be rare. If a single amino acid change is sufficiently adaptive, other neutral mutations outnumber it and the $d_N/d_S$ value appears to become low. Even if advantageous mutations exist, $d_N/d_S$ value could be canceled out by negative selection in other sites. Also, if adaptive evolution occurred only during a certain period of time in the past (episodic evolution), it could be canceled out by neutral evolution and negative selection in other periods.

Note the hidden assumption of $d_S/d_S$ test: $d_S$ reflects neutral evolutionary difference. We should be careful about validity of this assumption. $d_S$ could be lowered by negative selection due to codon usage bias and other reasons. If $d_S$ is lowered by such reasons, $d_N / d_S > 1$ could happen even though $d_N$ is not high. Natural selection can work effectively in large $N_e$ species to lower $d_S$ and to result in $d_N / d_S > 1$. If sequence length
is short, $d_{S}/d_{S} > 1$ could also happen just due to a stochastic fluctuation.

8.3.2. Convergent evolution

Here again, if adaptation is reached by only a few convergent substitutions, statistical significance is difficult to attain. However, if such convergence happens independently in many lineages, positive selection could be suggested.

8.3.3. Intra-species polymorphism

Selective sweep could decay after $\sim 0.5N_e$ generations due to neutral mutations accumulated during it (Simonsen et al., 1995, Genetics 141:413-429, Properties of statistical tests of neutrality for DNA polymorphism data). Population sampling is generally not feasible. Sampling bias can also occur very often. We should also note that there are various assumptions (e.g. equilibrium between mutation and drift) and should be aware of the assumptions when interpreting the results. For more details, Refer to Nei and Kumar's "Molecular Evolution and Phylogenetics".

As shown, each method has a limitation. However, merit is large if we use them appropriately. In inter-species comparison, we focus on fixed mutations. Even studying one individual from each species, we can obtain a wealth of information. In within-species comparison, we witness ongoing evolution. Signal of natural selection is easier to appear in allele frequency spectrum, LD, and so on. Combined use of various methods is complementary to each other and is very effective. Recent accumulation of genome-wide inter-species divergence and within-species diversity data enable us to evaluate natural selection more comprehensively and more rigorously.

8.4. Meaning of detection of positive selection

Upon the limitation of methodology, not all natural selection is detectable. Also note that the detected positive selection is merely a statistical estimation as a deviation from the neutral model. In reality, biological meaning, or adaptive significance, of many of positive selections estimated is not clear. Detection itself is not the goal of the study. When we estimate positive selection, we have to test its biological meaning by experiments or observations.

Besides these evolutionary methods, we are able to identify amino acid sites (variations) which are important in gene function through such as random mutagenesis. However, evolutionary methods are of academic importance in predicting them and providing meaningful working hypotheses.
8.5. Distinction of adaptive evolution and functional evolution

Refer to Panel 92. Often confused are adaptive evolution and functional evolution. Adaptive evolution concerns whether functional changes in physiology, morphology and whatever contribute to fitness (survival and reproduction). If so, we express that the trait is adaptive to its environment. Adaptive evolution is always functional evolution. However, functional evolution does not concern if it is adaptive or not. Neutral evolution of function is possible. Even deleterious function can evolve.

Panel 92

Adaptive evolution ≠ Functional evolution

(more survival and reproduction)

Neutrality test

General applicability

Test by functional assay

Different assay systems from gene to gene

Two approaches are complementary and necessary

Functional evolution is tested by functional assays. Functional assays include in vitro experiments, cellular to individual physiological experiments, behavioral experiments, and field observation. In principle, adaptive evolution is tested by monitoring survival and reproduction through many generations. This is practically impossible to do, especially for long generation time organisms. However, by the development of evolutionary genetics, adaptive evolution can be tested as a form of neutrality test. Regarding functional assay, experimental and observational methodologies vary greatly among genes to be examined. In contrast, neutrality test can be applied to any gene. The two approaches are complementary and both are needed.

A functional variant can be fixed as neutral evolution, then, become indispensable machinery if habitat conditions change. Or, it can be replaced with another functional variant by neutral evolution. Such idea is not new. Darwin also mentioned the idea of neutral evolution in "Origin of Species" (Panel 93). Under a rich environment, especially in newly opened habitats without competitors, this is likely to happen. The explosive diversification of organisms in Cambrian Period with various weird forms of life could support the idea.
8.5.1. Cryptic variation and HSP 90


A molecular developmental study of fruit fly made an important suggestion that morphological evolution can happen rapidly as neutral evolution. The heat shock protein 90 (Hsp90) can rescue damaged proteins under stress including heat shock. Under no stress, Hsp90 works as a binding stabilizer of regulatory signaling proteins in cell proliferation and embryonic development. By this stabilizing role, even if signaling proteins have abnormality, Hsp90 masks the abnormality and controls the system to conceal the harmful effect. In the wild and in the lab, fruit flies were revealed to harbor so many "cryptic" variations which can cause morphological abnormality. If such cryptic variations are accumulated, even under the normal Hsp90 function, morphological abnormality arises. Under heat stress, many of Hsp90 are allocated to rescue damages. Then, normal role Hsp90 becomes shortage, and abnormality comes out more easily.

This implies that a morphology-altering mutation can spread to population as a cryptic variation by neutral evolution. Then, by change of environment, phenotype change can occur suddenly in the whole population.
Chapter 9 Perspective of Evolutionary Genetics

The goal of evolutionary genetics is to clarify the history and mechanism of evolution of life. Regarding history, our knowledge has been tremendously increased by studies of genes. Regarding mechanism, theorems of population genetics, molecular evolution and the Neutral Theory have provided us with a firm framework of our thinking. Toward the future, fully using this knowledge, for particulars of adaptive evolution and functional evolution, we are to clarify the physical entity of relevant mutations and their spreading process and force (selection, drift and demography). Some viewpoints are addressed below.

9.1. Restoration of ancestral sequences

Using the maximum parsimony principle, we can infer ancestral amino acid and nucleotide sequences from current sequences. By applying the maximum likelihood principle, we can also infer ancestral sequences and evaluate reliability of inference with posterior probability at each site for each ancestral sequence. Refer to Nei and Kumar's "Molecular Evolution and Phylogenetics" for details.

On the basis of the inference, we can create ancestral gene clones by applying the site-directed mutagenesis method to current gene clones. If functional assay is possible for the gene, we can test the functional change during evolution in the lab (Panels 91 and 94).

Panel 94

Ancestral sequences of the ultraviolet and violet opsins were inferred at the seven amino acid sites critical for absorption spectrum. The evolutionary spectral shift of the visual opsins was tested.
9.2. Test of predicted functional evolution from detected positive selection

By $d_N/d_S$ or other methods, if positive selection is inferred to genetic variations, these changes are supposed to have caused some functional evolution to the gene. By introducing these variations to the current gene, with a relevant assay system available, we can test the prediction. This is as if we exploit the experimental results brought by nature. This could be much more efficient than random mutagenesis. Knowledge on relevant physicochemical properties and structure of the proteins would be further helpful in designing functional assay. In addition, by combination with ancestral sequence inference, we can evaluate biological meaning of positive selection which operated during various periods of evolutionary time.

Panel 95

Shozo Yokoyama, at Syracuse University (currently at Emory University), found three convergent amino acid substitutions between human and a fish in red and green opsin genes. He suspected that these three amino acid differences cause the spectral difference between red and green opsins (Yokoyama and Yokoyama, 1990, Proc Natl Acad Sci USA 87:9315-9318). This prediction was later verified by site-directed mutagenesis experiment to the reconstructed opsin photopigments.

Pancreatic RNase gene in colobine primates (an exemplary study of adaptive evolution combined with functional assay) (Panel 96)


Colobines are folivorous Old World monkeys inhabiting Asia and Africa. They have a forestomach, keep symbiotic bacteria in the foregut to ferment leaves, break the bacteria in the next stomach by lysozyme (Panel 91), and recover nutrients. RNASE1 is a ribonuclease (RNase) which is secreted from pancreas and
transported to small intestine to degrade RNA and recycle nitrogen efficiently. Colobines have two RNASE1 genes by gene duplication: \( RNASE1 \) and \( RNASE1B \).

By inferring the ancestral sequence, authors showed that no nucleotide substitution occurred in the coding region in the colobine \( RNASE1 \). By contrast, \( RNASE1B \) carried 12 substitutions, 10 of which were non-synonymous. By relative rate test, they showed that evolutionary rate of \( RNASE1B \) was significantly accelerated by the nonsynonymous substitutions and concluded that it was under positive selection (Panel 96 Fig. 3).

Seven out of 9 amino acid substitutions in the mature peptide altered amino-acid charge, all of which increased the negative charge. The pH level in the small intestine of humans is 7.4-8.0, while that of colobines is 6-7 due to foregut fermentation and related digestive physiology. Authors expressed the recombinant RNASE1 and RNASE1B proteins and showed that the optimal pH of RNASE1 was 7.4 and that of RNASE1B was 6.3 (Panel 96 Fig. 4a), demonstrating the adaptive evolution of \( RNASE1B \) to the pH environment of colobine small intestine.

RNASE1 is found in many other tissues besides the pancreas and has enzyme activity in degrading double-stranded (ds) RNA. By recombinant protein assay, authors showed that colobine RNASE1 had high digestive activity to dsRNA like other primate ones, while RNASE1B drastically reduced the activity (Panel 96 Fig. 4b).
upper). By site-directed mutagenesis, they also showed that all the charge-altering amino-acid substitutions contributed to the reduction of the digestive activity to dsRNA (Panel 96 Fig. 4b lower).

Gene duplication enabled $RNASE1$ to maintain the digestive activity to dsRNA and $RNASE1B$ to be released from the functional constraint and adapt to the new digestive function required by folivores. This work is the textbook example of evolutionary genetics combined with functional assay treating important subjects in evolution, the gene duplication and natural selection.

Go out from lab
To learn the biological meaning of positive selection detected, functional assays in the laboratory are not sometimes sufficient. Then, many different approaches including field behavioral observation would be taken. This should take you to a new research frontier.

9.3. Evolution as a system
Genes do not work independently each other. Recently attention is directed to interaction among genes and their overall effect. The regulation of gene expression and development is a good example. Evolutionary study of the gene interactive system as a whole is wanted. Understanding and being able to apply theorems and methods of evolutionary genetics would be very important and useful there.

9.4. Population genomics
With technological innovation such as advent of Next Generation Sequencer, whole-genome sequencing is more and more feasible. This enables us to analyze huge amount of genetic variation data. Population genome data has a potential to solve previously unsolved problems of evolutionary history, such as migration history, gene flow, population fission/admixture, population size change, phylogenetic relationship, divergence time, and so on. For example, relative to the expected distribution of coalescence times from the constant population-size model, if estimated coalescence times to numerous genome regions are biased to some time-points, we can infer population-size reduction events at these periods. It also enables us to detect adaptive evolution with very high reliability by comparison among genome regions (by genome scan) of various measures of genetic variation (such as $F_{ST}$, $\Pi$, $S$, Tajima's $D$, and LD).

Examination of a short gene from one individual provides information of only two diploid alleles of ancestors no matter how long we trace back in time. However, whole genome data of one individual can provide us with information on enormous number of ancestral individuals because of separate chromosomes and recombination. For example, at 10 generations ago, your genome is supposed to be split into $2^{10}$ (~1000) persons in maximum. Claiming the ancestry to a particular historical person is not biologically meaningful.
9.5. Genome editing, stem-cell induction / differentiation, chromatin dynamics, epigenomics, and phenotypic evolution

Genome editing technologies, such as CRISPR/Cas9, enable us to introduce various genomic alterations to cultured cells and model organisms and to investigate their functional consequence at cellular to organismal levels. This could facilitate testing functional meanings of evolutionary changes. Application of the induced pluripotent stem (iPS) cell technology to non-model organisms and induction of desired cell / tissue / organ types could open an entirely new frontier to evolutionary studies by recapitulating gene expression and developmental patterns in them non-invasively, which has been otherwise inaccessible. Technologies to study topological features of chromatin, such as 3C, Hi-C and advanced imaging, have revealed interactive and dynamic association of genome regions and its importance in gene expression regulation and cellular functions. Epigenetic features, such as DNA methylation and histone methylation / acetylation impact regulation of gene expression, some features of which could even be inherited. Intra-species and inter-species variation of these chromatin and epigenomic features remains largely unknown. These features must also become important subjects of evolutionary studies. By integrating rapid progress of these technologies with wealth of evolutionary genetics knowledge, we could reach to understanding the mechanism of phenotypic plasticity, diversity and evolution.

9.6. Research foci of evolutionary genetics today and tomorrow

- Gene expression control system
- Developmental control genes
- Genes of which functional evolution can be tested: e.g. sensory genes (vision, olfaction, gustation,)
- Genome-wide phylogenetic analysis and within-species genetic diversity analysis
- Role of transposable elements in evolution
- RNA gene function and evolution
- Application of genome editing to cellular and model organisms
- Induction of pluripotent stem (iPS) cells and their differentiation for non-model organisms
- Comparative chromatin dynamics
- Epigenomic variations within and between species
- Contextual shift of gene function: e.g. emergence of neural genes before nervous system appeared
- Cross-disciplinary approach: e.g. collaboration of genetics, informatics and field ethology