Abstract

A huge amount of species living on the earth display extensively diversified phenotypes. During the long period of evolution from their common ancestors, characteristic shaping each species, i.e. "species-ness", has been acquired over time. One of the biggest issues in biology, since the age of Darwin, is to unveil the processes of acquiring species-ness in the course of evolution. Several approaches have been developed to infer the states of ancestral species by comparing morphological and ecological characteristics and/or the genomic contents of extant species. Despite that much evidence has been accumulated, the processes of the evolution of shaping species-specific characteristics are still limitedly understood. This would be because each inferred state at the ancestor only illustrates a point of time in the ancestral lineage, which would be hard to reconstruct a whole picture of evolutionary processes from ancestors to extant species. Today, it becomes easier to infer the ancestral genomes from those of extant species, and much more information can be extracted because of the rapid increase of number of species whose whole genome has been sequenced and rapid accumulation of knowledge on biological function of genes, which enable the inference of phenotype to genotype. If the whole pictures of ancestral genomes are provided with high resolution in timeline of evolution, the continuous process of evolution could be reconstructed. This can be most easily attained for the group of closely related species whose genomes are available and the genomes of common ancestors are almost wholly inferred with high accuracy.

My aim in this study is to uncover the process of acquisition of "species-ness", the specific characteristics of species, comparing the whole genomes of closely related

extant species as well as inference of ancestors. To achieve it, I designed an approach consisting of three integrative and sequential analyses based on the reconstruction of the genomes of ancestral species. The first is to distinguish different processes of genomic changes such as single point mutation, indels, and inversion to reveal the detail of structural evolution of genomes with finer precision. The second is to reconstruct the history of species based on refined genomic comparison attained by the first procedure. The third is to identify the species-specific genomic content that provides species to have each own species-ness. This can be carried out once the evolutionary history of species is firmly established by the second procedure.

In order to examine feasibility of above approaches, I conducted comparative genomic analyses of closely related species by selecting specific examples, some of which are currently controversial and much debated, as follows: (i) *Identification of ultramicro inversions within local alignments between closely related species*, (ii) *Reconstruction of the demographic histories of the human lineage using whole genome alignments*, and (iii) *Identification of the species-specific characteristics involved in the pathogenicity and adaptation to the host environments in Theileria parasites*.

(i) *Identification of ultramicro inversions within local alignments between closely related species*.

Inversion is one of the major mechanisms for generating genomic diversity in evolution. While the inversions of large size have been well investigated since the early 20th century, little is understood about the minimal size of inversions, which would have useful information for clarifying the minute structural changes of genomes. I developed an efficient method for identifying minimal-sized inversions that I call "Ultramicro

Inversions" of 5-125 bp buried in nucleotide alignments, and identified 3,330 ultramicro inversions within the human-chimpanzee genome alignments. Around 26% of the ultramicro inversions consisted of adenine (A) and thymine (T) only, and the ultramicro inversions were also frequently found in chromosome Y and regions close to transposable elements. These observations suggested that the ultramicro inversions are related to instability of the genomic structures. Ninety ultramicro inversions were found in gene regions, and 28 out of 90 were in the coding regions, indicating that some parts of the ultramicro inversions may contribute to gene evolution. At least 31% of the ultramicro inversions in the human-chimpanzee alignments were bounded by inverted repeats, suggesting that such ultramicro inversions involved in the chromosomal recombinations via DNA stem-loops. In addition, I identified ultramicro inversions in various lineages other than primates: 1,285, 40, and 62 ultramicro inversions in the fly, fungi, and rice genomes, respectively, and 20 on average in the genomes of four and two lineages of eubacteria and archaea, respectively. This observation indicates that ultramicro inversions are ubiquitous across the three domains of the living world. While frequencies of the ultramicro inversions were up to seven times different between the lineages, the mechanisms of ultramicro inversions seemed to be more various across the lineages. The fractions of AT-exclusive and stem-loop type ultramicro inversions were much more different across the lineages. Identification of ultramicro inversion hotspots *in silico* would be helpful for capturing the inversions in experiments and clarifying the mechanisms of minute genome structural evolution. Our inversion-identification method is also applicable in the fine-tuning of genome alignments by distinguishing ultramicro inversions from simple point mutations and indels.

(ii) *Reconstruction of the demographic histories of the human lineage using whole genome alignments*.

The demographic history of human would provide helpful information for identifying the evolutionary events that shaped the humanity but remains controversial even in the genomic era. In order to settle the controversies, I inferred the evolutionary history of human and great apes based on an estimation of the speciation times (T) and ancestral population sizes (N) in the lineage leading to human and great apes using the whole-genome alignments. A coalescence simulation determined the sizes of alignment blocks and intervals between them required to obtain recombination-free blocks with a high frequency. This simulation revealed that the size of the alignment block strongly affects the parameter inference, indicating that recombination is an important factor for achieving optimum parameter inference and that this simulation is helpful for the optimum data collection. From the whole genome alignments (1.9 giga-bases) of human (H), chimpanzee (C), gorilla (G), and orangutan, and the small-sized regions subject to the genomic changes by the other mechanisms than point mutations, such as CpG sites and ultramicro inversions, were excluded. 100-bp alignment blocks separated by ≥5-kb intervals were sampled from the alignments and subjected to estimate $\tau=\mu T$ and $\theta=4\mu gN$ using the MCMC method, where μ is the mutation rate and g is the generation time. Although the estimated τ_{HC} differed across chromosomes, τ_{HC} and τ_{HCG} were strongly correlated across chromosomes, indicating that variation in τ is subject to variation in μ across the lineages, rather than T, and thus, all chromosomes are likely to share a single speciation time. Subsequently, I estimated Ts of the human lineage from chimpanzee, gorilla, and orangutan to be 6.0-7.6, 7.6-9.7, and 15-19 MYA, respectively, assuming variable μ across lineages and chromosomes. These speciation times were consistent

with the fossil records. I conclude that the speciation times in our recombination-free analysis would be conclusive and the speciation between human and chimpanzee was a single event.

(iii) *Identification of the species-specific characteristics involved in the pathogenicity and adaptation to the host environments in Theileria parasites*.

Theileria is a tick-born apicomplexan group causing parasitosis in livestock. Some theilerias are parasitic to cattle, but the relationship between the theileria and cattle seem to have evolved specifically in each lineage. While *T. annulata* and *T. parva* (transforming theileria) induce abnormal proliferation of infected cells of lymphocyte or macrophage/monocyte lineages and are severely pathogenic, *T. orientalis* does not induce such transformation and shows moderate pathogenicity. Here, in order to clarify the process of acquiring the high pathogenicity and diverged systems infecting the hosts, I reconstructed the evolutionary history of theileria based on the comparative genomics of the almost whole genomes. While synteny across the chromosomes of the three theilerias was well conserved, subtelomeric regions were largely different: *T. orientalis* lacks the large tandemly arrayed subtelomere-encoded variable secreted protein-encoding gene family. Through the orthologue clustering, in addition, I found that duplication and deletion rates in the transforming theileria lineage were 1.66 and 1.95 times faster than those in the *T. orientalis* lineages, respectively. Expansion of particular gene families by gene duplication was found specifically in the two transforming theileria species. One of the most notable families is the TashAT/TpHN gene family, which is considered to be involved in transformation and abnormal proliferation of host leukocytes. The transforming theileria possessed around 20

TashAT/TpHN members, while only one member was identified in *T. orientalis*, and no homologues were found in a babesia and plasmodiums. I also found the gene families expanded specifically in *T. orientalis* lineages such as ABC transporters, implying species-specific strategies against host systems. Differences between the genome sequences of theileria species illustrated different tempo and mode of gene duplication and deletion between transforming theilerias and *T. orientalis*. It is implied, moreover, that such differences in evolutionary modes resulted in the novel abilities to transform and immortalize bovine leukocytes. The genomic changes between close relatives will provide insight into proteins and mechanisms that have evolved to induce and regulate this process.

In the above studies, I have examined and demonstrated effectiveness of the three steps of integrative and sequential approach for clarifying the evolutionary processes to attain "species-ness" at the genome level by reconstructing the genomes of ancestral species from closely related extant species. Even though the current approaches are based on the well-established fields of genomics, population genetics, and molecular phylogenetics, the integration of the approaches, as shown here, is innovative in the field of *in silico* genomic analysis and provides new insight on evolutionary biology. In the near future, comparative analysis of closely related species will be expanded for the genomes of species suitable to solve a particular biological issue. The integrative approach provided here would become one of de-facto standard for such analyses.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Supporting Data

Supporting data D2-1	Human-chimpanzee orthologous genome alignments.
Supporting data D2-2	List of ultramicro inversions in the human-chimpanzee alignments.
Supporting data D3-1	Whole and sampled genome alignments of human and three apes.
Supporting data D3-2	Genealogies and sequence alignments of the simulation data.
Supporting data D3-3	Estimated parameters for each chromosomal alignment set with 95% CI.
Supporting data D4-1	Gene families in piroplasms and plasmodium.

(Attached in DVD-ROM).

Chapter 1. General Introduction

A huge number of organisms live almost everywhere on the earth such as air, land surface, deep sea, and geological layers. For example, approximately 1.2 million and around 8.7 million species of eukaryotes are currently known and predicted to be present, respectively (Mora, et al. 2011). Based on a large amount of observations since the evolutionary study of Darwin's "On the Origin of Species" (Darwin 1859) to recent genomic studies (e.g. Carlton, et al. 2008; Clark, et al. 2007; Dujon, et al. 2004), it reached a conclusion that every extant species shares the common ancestor and that species have diverged from the ancestor and have been changing over time. In addition, the species display enormously diverse morphologic and ecological characters. Such varieties of phenotypes in organisms have been acquired through random mutations and adaptations to the environments, of which process we call evolution. The evolutionary processes consequently give every species its own "species-ness", which means characteristics representing specificity of species.

A whole picture of the species evolution can be illustrated by understanding changes of various characters of species continuously from past to present. However, two issues need to be solved to clarify the evolutionary process of shaping species specific characters. One is to reveal the evolutionary histories of organisms in continuous time. Despite a large amount of evidence on evolution, the processes of evolution are limitedly understood, since most of the evolutionary studies up to now only show sporadic pictures of ancestral states. Although fossils may show states of common ancestor of extant species, and the comparison of the extant species can illustrates the state of the ancestor, such analyses may be sparse in time. The other is to infer ancestral states of give characters at high accuracy; there is no way to see ancient species alive and obtain intact samples of ancient species. Although fossil records are

very helpful for inferring the states of ancestral species, they are scarce and their uses are limited: fossils are often found as limited parts of body, and records except bones, teeth, and skins are rarely available. Thus ancient state of organisms can be mainly inferred from the states of extant species. Nevertheless, comparison analyses of closely related species rarely face these problems. Since common ancestor of closely related species lived in near past, the changes of characters from common ancestors to the extant species can be inferred more continuously. In addition, the differences of closely related species are small enough so that nearly complete pictures of ancestors can be reconstructed, therefore changes of characters, some of which have generated "species-ness", can be easily identified for each evolutionary lineage. Comparison of the closely related species, therefore, is capable of illustrating whole pictures of extant and ancestral species nearly continuously in timeline like making an animation movie.

Reconstructing the evolutionary processes of organisms has been attempted with various approaches. Morphology has been long utilized as one of fundamental and direct keys for the purpose (Barton, et al. 2007; Futuyma 2005). Body shapes of ancestral species are inferred and compared with extant ones. Certain characteristics of extant and/or fossil species are shared on the ancestral node of their phylogenetic tree, called synapomorphies. Simultaneously, by a maximum parsimony framework using multiple characters, a phylogenetic tree of species can be inferred with inferring ancestral state of respective character at each node. Indeed this approach has solved a large number of issues on species phylogeny (e.g. whales are close to Artiodactyla (Fordyce and Barnes 1994; Novacek 1992)), which sometimes referred as fossil species (e.g. archaeopteryx belongs to a group of extinct dinosaurs rather than lineal ancestors of extant birds (Xu, et al. 2011; Zanno and Makovicky 2011)).

Ancient ecosystem is also important to infer ancestral states of species: it enables us to consider habitats and way of living of ancestors. Based on the knowledge on the distribution of extant species and geological events such as Plate tectonics and isolation of islands from continents, phylogeographical analyses can infer habitat changes of ancestral species (Futuyma 2005). Stratal organization in geological layers would be also informative. Similar structures to stromatolites generated by cyanobacteria were found in geological layers dated to 3.5 billion years ago (Allwood, et al. 2009).

However, it is often difficult to infer ancestral states by comparing morphological characters. For example, temporal holes in amniotes were assumed to be a derived character, and thus turtles, which lack the holes, could be primitive groups of amniotes or reptiles (Carroll 1988; Lee 1997). However, recent molecular phylogenomics clearly indicated that turtles belonged to a sister group of archosaurs (birds and crocodiles) and that squamates (snakes and lizards) had diverged earlier than the turtle-archosaur group (Iwabe, et al. 2005), indicating that temporal holes secondarily lost during the evolution of turtles. In another case, a morphological comparison had once divided larva, male, and female of a single species of whalefish into three different families (Johnson, et al. 2009). In addition, the quantities of morphological and ecological characteristics that we can obtain now are limited and somewhat arbitrary in that focusing on only visible ones.

Molecular phylogeny can be an alternative, though not mutually exclusive, approach to clarify the evolution based on the changes of molecules such as nucleotide and protein sequences and to infer the ancestral states of species. With using appropriate evolutionary models, we can infer nucleotide and/or amino acid sequences of genes at ancestral nodes and phylogenetic trees of genes. Reconstructing phylogenetic trees of genes and matching the evolutionary signatures of the gene sequences on the species

tree, we are able to infer the processes of gene evolution (Maddison 1997; Page and Charleston 1997). Using a set of appropriate genes, in addition, we can infer the species phylogeny itself (Delsuc, et al. 2005; Maddison 1997). However, molecular phylogenetic study of one-by-one genes may be hard to infer evolution of species because such analysis often has less quantity and sometimes can lead to wrong tree. Moreover, the choice of genes for the phylogenetic studies may still remain arbitrariness.

Massive data of genomes is capable of reconstructing the ancestral genomes with a quantity of information, which enable us to conduct various statistical tests with high credibility. Analysis of genomic sequences of different species can reveal the differences between the extant and ancestral genomes with reconstructing the genomes of ancestors. These genomic differences are the primary source of acquisition of phenotypic characteristics representing each species.

Compared to analyses of morphological and ecological characters, genomic analyses used to have several deficits; but they are rapidly improved. While morphological and ecological characters directly connect to phenotypes, large gaps have lied between genome information and phenotypes. However, biological functions of genomic regions have been elucidated rapidly. The ENCODE project uncovers the regions of human genome with some functions and reveals that many of them locates outside of the regions of conventional genes (Bernstein, et al. 2012). In addition, such nearly whole functional information of genomes is capable of identifying gene-gene interaction, which is uncovering the systematic outputs from the genomes next to phenotypes. Pathway databases such as KEGG and Reactome represent dynamics among genes and gene products *in vivo* (Joshi-Tope, et al. 2005; Ogata, et al. 1999).

The species whose genomes have been wholly sequenced are still few and sparse in the tree of life. Up to now, genomes have been completely sequences for 183 eukaryotes and 3,699 prokaryotes as in the Genomes OnLine Database (on Jan. 6, 2013) (Pagani, et al. 2012), representing only 0.015% of catalogued eukaryotes (Mora, et al. 2011). However, recent progress in DNA sequencing technology enables us to read the whole genome sequences with low price and high speed. Recently, genomes of species of interest such and model organisms, agricultural species, and pathogens, have been often analyzed together with their close relatives. They are referred to obtain the genomic characteristics involved in species-specific phenotypes (Carlton, et al. 2008; Liti, et al. 2009; Sakai, et al. 2011). In addition, the international genome projects targeting thousands of species in representing taxa have been launched: Genome 10k for 10,000 vertebrate species (Haussler, et al. 2009), and i5k for 5,000 insects (Consortium 2011). In 2012, the genome of Darwin finch was wholly sequenced as the first species for the Genome 10k project (Zhang, et al. 2012).

Due to accumulation of the genome sequences of various species and understandings of genotypes to phenotypes, we are now ready to infer the evolutionary processes of organisms based on the analyses of genomes by reconstructing ancestral genomes. Especially, comparative genomic analyses of closely related species can provide almost complete the genomic picture of ancestral species with high accuracy. This is because genomic sequences of closely related species can be aligned almost entirely and the ancestral sequences can be inferred at each and every site from the alignments. Moreover, common ancestors of closely related species are such recent that comparison of extant and inferred ancestral genomes can be conducted in high resolution in timeline. Comparison of genomes of closely related species can reveal the

degree of genetic variety of their ancestors (Dutheil, et al. 2009; Pamilo and Nei 1988; Pinho and Hey 2010), which can be used for the inference of the population structures and the processes of natural selections in common ancestors based on the theory of population genetics (Green, et al. 2010; Liti, et al. 2009). This approach based on the comparison of genomes of closely related species genomes is capable of inferring the changes of the genomes during evolution, which contributes to the understanding of phenotypic evolution of species.

My aim in this study is to clarify the processes of shaping "species-ness" of extant species at genome level. This can be achieved by comparing extant and ancestral species and connects them to phenotypes. For the purpose, three sequential approaches were applied. I first investigated detail mechanisms of structural changes observed in the evolution of genome. Subsequently, I reconstructed evolutionary history of species. Finally, genomic changes occurred at each evolutionary lineage are inferred and the characteristics shaping the "species-ness" are revealed. It is noted that the first procedure is required to the second one, and that the second procedure is required to the third one. The species phylogeny is inferred based on the phylogenetic or population genetic approaches. These approaches mainly use the mismatches of nucleotides observed in aligned genomes, which are considered to be generated by point mutations. Thus we should evaluate if every mismatches in an alignment represents a point mutation. In the other case small genomic structural changes enough have occurred and erroneously aligned in the alignments. Species-specific characteristic at the genome level can be accurately identified once the precise species phylogeny is obtained.

The aim could be achieved by integrating various methods established in diversified fields of genomics, population genetics, and molecular phylogeny.

Nonetheless, such integrative approach may promote innovations for *in silico* analysis for elucidating process of evolution at genome level. In addition the comparative analysis of genomes of closely related species will shed light on many aspects of genome and species evolution such and species specific characteristics which can be lead by adaptation specifically occurred in the lineages.

The outline of my study is as follows. The first section is the investigation on the mechanisms of genome evolution at minute scale. I identified a large number of minute structural changes, i.e., ultramicro inversions, within the genome alignments between human and chimpanzee (Hara and Imanishi 2011) and those of closely related species in the other lineages. The next is reevaluation of speciation history of the hominoids. Using the aligned nucleotides sequences of entire genomes of human and three great apes excluding the regions containing ultramicro inversions, I examined the speciation process of hominids and found no evidence of introgression between the ancestors of human and chimpanzee (Hara, et al. 2012). The third is the investigation on the origin of pathogenicity and parasitic mechanisms of theileria using newly sequenced genomes. Although the third procedure could be conducted with the human and its relatives genomes for consistency, functional analyses of human are still limited mainly due to the difficulties of molecular biological experiments. However it is noted that such difficulties can be solved by functional analyses at genome level such as ENCODE). Therefore, I used parasitic protists theileria as an example of analysis; several advantages are there to examine the process of shaping species-ness: small genome sizes, clear phylogenetic relationship, easier establishment of biological experiments, and clearly observed as lineage-specific characteristics such as pathogenicity and adaptations to hosts. Based on the comparison of the newly sequenced *Theileria*

orientalis genome with two known genomes of the transforming theileria (*T. annulata* and *T. parva*), I identified the candidate genes involved in the high pathogenicity of the transforming theileria in host leukocytes and those involved in the host-parasite interaction in *T. orientalis* (Hayashida, et al. 2012). Through these three approaches, I have shown that the integrative analyses of genomics, population genetics and molecular phylogenetics have successfully clarified mechanisms, history, and adaptation process of genome evolution and thus are capable of identifying the "species-ness" in organismal evolution.

References

Allwood AC, et al. 2009. Controls on development and diversity of Early Archean stromatolites. Proc Natl Acad Sci U S A 106: 9548-9555.

Barton NH, Briggs DEG, Eisen JA, Goldstein DB, Patel NH. 2007. Evolution. Cold Spring Harbor: Cold Spring Harbor Laboratory Pr.

Bernstein BE, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57-74.

Carlton JM, et al. 2008. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature 455: 757-763.

Carroll RL. 1988. Vertebrate paleontology and evolution. New York: Freeman.

Clark AG, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203-218.

i5k Insect and other Arthropod Genome Sequencing Initiative [Internet]. 2011. Available from: http://arthropodgenomes.org/wiki/i5K

Darwin C. 1859. On the Origin of Species. London: John Murray.

Delsuc F, Brinkmann H, Philippe H 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361-375.

Dujon B, et al. 2004. Genome evolution in yeasts. Nature 430: 35-44.

Dutheil JY, et al. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. Genetics 183: 259-274.

Fordyce RE, Barnes LG 1994. The Evolutionary History of Whales and Dolphins. Annual Review of Earth and Planetary Sciences 22: 419-455.

Futuyma D. 2005. Evolution. Sunderland: Sinauer.

Green RE, et al. 2010. A Draft Sequence of the Neandertal Genome. Science 328: 710-722.

Hara Y, Imanishi T 2011. Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. BMC Evol Biol 11: 308.

Hara Y, Imanishi T, Satta Y 2012. Reconstructing the Demographic History of the Human Lineage Using Whole-Genome Sequences from Human and Three Great Apes. Genome Biol Evol 4: 1133-1145.

Haussler D, et al. 2009. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. Journal of Heredity 100: 659-674.

Hayashida K, et al. 2012. Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of theileria-induced leukocyte transformation. MBio 3: e00204-12.

Iwabe N, et al. 2005. Sister group relationship of turtles to the bird-crocodilian clade revealed by nuclear DNA-coded proteins. Mol Biol Evol 22: 810-813.

Johnson GD, et al. 2009. Deep-sea mystery solved: astonishing larval transformations and extreme sexual dimorphism unite three fish families. Biol Lett 5: 235-239.

Joshi-Tope G, et al. 2005. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33: D428-432.

Lee MSY 1997. Pareiasaur phylogeny and the origin of turtles. Zoological Journal of the Linnean Society 120: 197-280.

Liti G, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458: 337-341.

Maddison WP 1997. Gene trees in species trees. Systematic Biology 46: 523-536.

Mora C, Tittensor DP, Adl S, Simpson AG, Worm B 2011. How many species are there

on Earth and in the ocean? PLoS Biol 9: e1001127.

Novacek MJ 1992. Mammalian Phylogeny - Shaking the Tree. Nature 356: 121-125.

Ogata H, et al. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34.

Pagani I, et al. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40: D571-579.

Page RDM, Charleston MA 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree species tree problem. Molecular Phylogenetics and Evolution 7: 231-240.

Pamilo P, Nei M 1988. Relationships between gene trees and species trees. Mol Biol Evol 5: 568-583.

Pinho C, Hey J 2010. Divergence with Gene Flow: Models and Data. Annual Review of Ecology, Evolution, and Systematics, Vol 41 41: 215-230.

Sakai H, et al. 2011. Distinct evolutionary patterns of Oryza glaberrima deciphered by genome sequencing and comparative analysis. Plant J 66: 796-805.

Xu X, You H, Du K, Han F 2011. An Archaeopteryx-like theropod from China and the origin of Avialae. Nature 475: 465-470.

Zanno LE, Makovicky PJ 2011. Herbivorous ecomorphology and specialization patterns in theropod dinosaur evolution. Proc Natl Acad Sci U S A 108: 232-237.

The genome of Darwin's Finch (Geospiza fortis). 2012. GigaScience. Available from http://dx.doi.org/10.5524/100040.

Chapter 2. Identification of Ultramicro Inversions within Local Alignments between Closely Related Species.

2.1. Summary

Inversion is one of the major mechanisms for generating genomic diversity in evolution. While the inversions of large size have been well investigated since the early 20th century, little is understood about the minimal size of inversions, which would have useful information for clarifying the minute structural changes of genomes. I developed an efficient method for identifying minimal-sized inversions that I call "Ultramicro Inversions" of 5-125 bp buried in nucleotide alignments, and identified 3,330 ultramicro inversions within the human-chimpanzee genome alignments. Around 26% of the ultramicro inversions consisted of adenine (A) and thymine (T) only, and the ultramicro inversions were also frequently found in chromosome Y and regions close to transposable elements. These observations suggested that the ultramicro inversions are related to instability of the genomic structures. Ninety ultramicro inversions were found in gene regions, and 28 out of 90 were in the coding regions, indicating that some parts of the ultramicro inversions may contribute to gene evolution. At least 31% of the ultramicro inversions in the human-chimpanzee alignments were bounded by inverted repeats, suggesting that such ultramicro inversions involved in the chromosomal recombinations via DNA stem-loops. In addition, I identified ultramicro inversions in various lineages other than primates: 1,285, 40, and 62 ultramicro inversions in the fly, fungi, and rice genomes, respectively, and 20 on average in the genomes of four and two lineages of eubacteria and archaea, respectively. This observation indicates that ultramicro inversions are ubiquitous across the three domains of the living world. While frequencies of the ultramicro inversions were up to seven times different between the lineages, the mechanisms of ultramicro inversions seemed to be more various across the

lineages. The fractions of AT-exclusive and stem-loop type ultramicro inversions were much more different across the lineages. Identification of ultramicro inversion hotspots *in silico* would be helpful for capturing the inversions in experiments and clarifying the mechanisms of minute genome structural evolution. Our inversion-identification method is also applicable in the fine-tuning of genome alignments by distinguishing ultramicro inversions from simple point mutations and indels.

2.2. Introduction

Chromosomal inversion, a type of genetic rearrangement involving the inversion of a chromosome segment, is one of the most important causes of genomic structural changes. Inversions have been identified as phylogenetic signatures since the first third of the twentieth century (Sturtevant 1921) and are thought to have affected speciation and phenotypic evolution (Kirkpatrick 2010; Kirkpatrick and Barton 2006). While large-size inversions (macroscopic inversions), detectable by microscopes and/or visible in genetic maps, were identified early on (Sturtevant 1921; Tan 1935), the recent abundance of genomic sequences and progress in sequence analysis has enabled the extensive detection of inversions of various sizes in genomes. In particular, comparative genomics between populations and between closely related species have revealed the occurrence of numerous inversions in genomes including relatively small-size inversions (Bansal, et al. 2007; Chaisson, et al. 2006; Feuk, et al. 2005). More than 1,500 inversions varying in length from 23 bp to 62 Mb occur in the human and chimpanzee genomes, suggesting that inversions are common mechanisms for differentiating genomes (Feuk, et al. 2005). Although several methods have been developed to identify these inversions, they focus only on the macroscopic or relatively small inversions which are inversions large enough to be detected as a single alignment or a string of local alignments (Feuk, et al. 2005).

Some inversions may be too small to be identified even as a local alignment block. Such "ultramicro inversions" are extremely difficult to detect using existing methods because they may be hidden within the local alignments of BLAST or other popular alignment softwares. These inversions are very short such that the alignment extends

beyond them allowing mismatches and gaps. In these cases, the ultramicro-inverted regions are treated as small arrays of mismatches and gaps within the local alignments (Fig. 2-1). The degree of overlooking the ultramicro inversions hidden within the local alignments would be high within the alignment between closely related genomes, because mismatches in short regions are negligibly small for highly similar alignments longer than ten kilo-bases.

These ultramicro inversions would be aligned with mismatches and gaps with higher density in an aligned region than would a random distribution of such differences in the whole alignment. Generally, mismatches and gaps within alignments account for nucleotide substitutions and insertions and deletions (indels), respectively (Fig. 2-2**A**, **B**). However, mismatches and gaps generated at inverted sites are a result of erroneous alignments (Fig. 2-2**C**). Whether or not differences in the alignments are caused by nucleotide substitutions and indels is apparently unclear. Thus, it is difficult to obtain information about ultramicro inversions from the local alignments themselves. Identifying ultramicro inversions within the alignments would be necessary for distinguishing the mismatches and gaps caused by nucleotide substitutions and indels from those caused by inversions.

The human genome is different from the chimpanzee genome, at 1.2% of sequence mismatches (2005) and 5% of sequence mismatches plus gaps (Britten 2002). Some of these differences are assumed to play important roles in the phenotypic evolution of the human lineage. Furthermore, macroscopic inversions are one of the major mechanisms of differentiating species (Kirkpatrick and Barton 2006). For example, pericentric inversion is one type of large genomic rearrangements which distinguishes the human karyotype from that of the chimpanzee (Kehrer-Sawatzki and

Cooper 2007; Nickerson and Nelson 1998), implying that such inversions are one of the important causes of speciation. Ultramicro inversions may also be spread across the human and chimpanzee genomes because the size distribution of the macroscopic and relatively small inversions decays as the size of the inversions increases (Feuk, et al. 2005). In addition, the differences in the human–chimpanzee alignments caused by inversions raise the average differences between the human and chimpanzee genomes. In order to examine the impact of ultramicro inversions on the genome alignment, I developed a method for identifying ultramicro inversions within the alignments between the human and chimpanzee genomes. I first generated 2.41 Gb of one-to-one (i.e., orthologous) alignments between the human and chimpanzee genomes using the G-compass pipeline (Fujii, et al. 2005; Kawahara, et al. 2009), and identified inversions in each local alignment. Subsequently, I examined the relationships of ultramicro inversions with the structural features of the human genome to see the molecular mechanisms of the inversions. Furthermore, I examined biologically functional segments to infer the effects of the inversions on the phenotypic evolution of the human lineage. Finally, I showed that the ultramicro inversions were ubiquitous across the living organism genomes. In this study, I improved the method of identification of ultramicro inversions developed by Hara and Imanishi (Hara and Imanishi 2011). Compared with the previous method, new one is more sensitive to minute size of ultramicro inversions (5 and 6 bp) and adenine and thymine-exclusive inversions with equivalent false positive rate.

2.3. Materials and Methods

Identification method for ultramicro inversions

Ultramicro inversions were detected within a local pairwise alignment by the following two procedures: identifying difference-rich regions and searching for inverted regions in these difference-rich regions. Fig. 2-2 illustrates the outline of these procedures. A region rich in mismatches and gaps was initially detected as a trio of the nearest mismatches and gap blocks which were more closely positioned on an alignment than expected (Fig. 2-2**A**). Each trio consisted of either three mismatches, two mismatches and one gap block, or one mismatch and two gap blocks that were located in different sequences of a pair. The trio was extracted by scanning the pairwise alignment. When a mismatch or gap block was found and the next two mismatches and/or gaps were located within a region of $n - 1$ consecutive sites, the conditional probability of the trio within n sites $P_{\mathrm{trio}}(n)$ ($n \geq 3$) was calculated by the equation given below,

$$P_{\mathrm{trio}}(n) = 1 - (1 - p_{\mathrm{d}})^{n-1} - (n-1)p_{\mathrm{d}}(1 - p_{\mathrm{d}})^{n-2}$$

where p_{d} represents the average number of mismatches and gap blocks per site.

During the detection process, some parts of the inversions were found to be aligned without mismatches, as follows:

ATGCCCGG----

----CCGGGCAT

The inversion of 8 bp included a palindrome in part and was aligned with the palindrome. I called this a partially palindromic inversion. In order to identify this kind of inversion, I search for the region where an identically aligned region was sandwiched

by two gap blocks inserted in different sequences of pair (Fig. 2-2**B**). The conditional probability of the two gap blocks within n sites $P_{\mathrm{duo}}(n)$ ($n \geq 2$) is given as follows:

$$P_{\mathrm{duo}}(n) = 1 - (1 - p_{\mathrm{g}})^{n-1}$$

where p_{g} is the average number of the gap blocks per site. I extracted such trios and duos where $P_{\mathrm{trio}}(n) < 0.01$ or $P_{\mathrm{duo}}(n) < 0.01$. These trios and duos were merged and then extended with 25 bp at both ends. The resultant regions were candidates and subject to subsequent analysis.

The candidates of the inversions were aligned by the Dynamic Programming assuming inversions (Chen, et al. 2004) (Fig. 2-2**C**). The alignment scores in Dynamic Programming were assumed as follows: match score at 10, mismatch penalty at 11, gap open penalty at 15, gap extension penalty at 5, and inversion penalty at 23. If the inversion was aligned with higher similarity than the corresponding forward alignment based on following the criteria stated below, I defined the inverted region as an ultramicro inversion —(i) For the region based on $P_{\mathrm{trio}}(n) < 0.01$, similarity of the inverted alignment should be >0.95, and the corresponding forward alignment should include the trio. (ii) For the region based on $P_{\mathrm{duo}}(n) < 0.05$, similarity of the inverted alignment should be 100%, and the corresponding forward alignment should include the pair of gaps at each end. Following these procedures, I excluded the some parts of AT-exclusive inversions which could be explained by other mechanisms than inversions. One of the inversions to be excluded consisted of mononucleotide repeats of A and T such as 5′-AAATTTTTTT-3′: the inversion 5′-AAAAAAATTT-3′ could be explained by with stretch and shrink of the repeats. The other consisted of staggered AT dinucleotide repeats such as 5′-ATATATATATA-3′: the inversion 5′-TATATATATAT-3′ could be

explained by insertion of deletion of A or T. The length of ultramicro inversions was defined as the length of the inverted segments determined by the blastn program.

Simulation

In order to evaluate the power of my identification method, simulations were performed using sequence alignment sets of random sequences, each consisting of 100,000 pairs of approximately 5,000 bp sequences and including an inversion in each pair. Each sequence pair was generated by the sequence evolution simulator Indelible (Fletcher and Yang 2009), allowing insertions and deletions (indels) from a random sequence of 5,000 bp, assuming the HKY sequence substitution model (Hasegawa, et al. 1985), the indel lengths distributed with the Lavalette distribution setting the decimal at 2 and the maximum indel length at 50, and the prior sequence conditions similar to the human–chimpanzee genome alignment: setting the base compositions of g_A, g_T, g_C, and g_G at 0.289, 0.304, 0.203, and 0.204, respectively, the transition/transversion ratio at 1.75, the average of sequence substitutions per site at 1.00, the indel/substitution ratio at 0.159, and the shape parameter for the gamma distribution and the number of categories for the discrete gamma approximation set at 0.65 and 5, respectively. A short region of 5 to 50 bp length in one sequence of each pair was inverted, and the pair was aligned with MAFFT (Katoh, et al. 2005). The inversion lengths were fixed in all 100,000 pairs of a sequence alignment set. I generated eleven sequence alignment sets with 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, and 50 bp inversions. The other sequence alignment groups were generated for assessing the intensity of AT-exclusive sequences. The sequence models and prior sequence parameters were equal to the simulation above except for the base compositions; (g_A, g_T, g_C, g_G) = (0.5, 0.5, 0, 0). For the AT-exclusive alignments, seven

alignment sets were generated, and 5, 10, 15, 20, 25, 30, and 50 bp inversions were included in the individual alignment sets. In these simulations, p_d and p_g were set at 0.0100 and 0.00150, respectively.

Identification of the ultramicro inversions within the human–chimpanzee alignments

To identify ultramicro inversions within the alignments between the human and chimpanzee genomes, one-to-one (i.e., orthologous) alignments were generated between the human and chimpanzee using the hg19 human genomic sequence and panTro2 chimpanzee genomic sequence from the UCSC genome browser (http://genome.ucsc.edu/). The alignments were constructed with the G-compass pipeline (Fujii, et al. 2005; Kawahara, et al. 2009) based on the lastz local alignments (Harris 2007) and its unique and non-redundant reciprocal best hits. The human-chimpanzee alignments were described in Supporting data D2-1. Applying the method above after setting p_d and p_g at 0.0136 and 0.00150 respectively, I obtained ultramicro inversions within the human–chimpanzee alignments.

The human–gorilla and human–orangutan one-to-one alignments were generated by the same procedures, using the gorGor3 gorilla and ponAbe2 orangutan genomic sequences from the UCSC genome browser (http://genome.ucsc.edu/). The human–chimpanzee alignments including the ultramicro inversions were grouped with the human–gorilla and human–orangutan alignments in which the human sequence overlapped the inversion segments in the human–chimpanzee alignments. In each group, the human, chimpanzee, and gorilla and/or orangutan sequences were multiply aligned by MAFFT (Katoh, et al. 2005).

The validation of ultramicro inversions based on phylogenic profiles was

performed using these multiple alignments. In the alignment sites of the inversions, if fewer than two mismatches or gaps were found between the human and outgroup sequences and three or more mismatches or gaps were found between the chimpanzee and outgroup sequences, I concluded that the chimpanzee sequence had been inverted. The human inverted sequences were also detected in the same way. If the phylogenetic profile of the inversion was inconsistent with the species phylogeny among human, chimpanzee, and gorilla, the inversion was verified with the incomplete lineage sorting.

In order to consider ultramicro inversions together with the genomic features, I used two kinds of genomic tracks available from the public database. Mapping information of exons and coding regions of human transcripts on the human genome were obtained from H-InvDB version 7.5 (http://hinv.jp/hinv/ahg-db/) (Imanishi, et al. 2004). Mapping information of *Alu* and *L1* was obtained from the chromOut repeat-masking annotation files on the human genome from the UCSC genome browser (http://genome.ucsc.edu/). To determine if ultramicro inversions preferentially occur in the neighborhood of transposable elements, I conducted a 1,000 times trial of the random distribution of the short segments on the human genome. Given that the size distribution of 3,300 short segments was identical to that of the ultramicro inversions within the local alignments between the human and chimpanzee genomes, these segments were randomly distributed on the human genome. Frequency distributions of every 100 bp of genomic distances between the segment and nearest transportable element were computed. If a boundary of the mobile element was included in the ultramicro inversion, the distance was set to zero. Frequencies of the short segments in every 100 bp were counted from the 1,000 times trial of the random distribution. The value of $p<0.001$ indicates no appearance of the short segment in the trial. I also applied

this procedure to determine if DNA stem-loops frequently included the ultramicro inversions. In this study, I extracted the ultramicro inversions within the loops with ≤10 bp spacer regions at each end and those within the stem-loops consisting of loops and ≥7 bp stems (as known as inverted repeats). It is empirically known that a pair of inverted repeats consisting of at least 7 bp nucleotides can make a stable stem-loop (Nag and Petes 1991). I also investigated the possibility that the inverted repeats were randomly distributed on both ends of the short segments instead of calculating the distance from the short segments to the transposable elements.

Ultramicro inversions in the alignment of closely related species or strains

Genome alignments of eukaryotes, fly (*Drosophila melanogaster* and *D. simulans*), budding yeasts (*Saccharomyces paradoxus* reference and A12 strains), rice (*Oryza sativa* and *O. glaberrina*), were obtained from AAA (Assembly/Alignment/Annotation) of 12 related Drosophila species (http://www.mmnt.net/db/0/0/ftp.biostat.wisc.edu/pub/cdewey/data/fly_alignments), Saccharomyces Genome Resequencing in Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html), and AfRicA DB in NIAS (http://green.dna.affrc.go.jp/Og/), respectively. Genome sequences of four eubacteria, *Neisseria meningitidis* (strains MC58 and Z2491), *E. coli* (strains K12 and O157), and *Helicobacter pylori* (strains J99 and 26695), and *Streptococcus pyogenes* (strains SSI-1 and MGAS8232), and two archaea, *Pyrobaculum* (*P. arsenaticum* and *P. oguniense*) and *Sulfolobus islandicus* (strains M.16.4 and Y.N.15.51) were obtained from NCBI Nucleotide (http://www.ncbi.nlm.nih.gov/nuccore). Genome alignments of the bacteria lineages were generated by the G-compass pipeline. Ultramicro inversions

were identified based on the same procedures to the human-chimpanzee alignments described above with the same parameters.

2.4. Results

Simulation

Firstly, I defined ultramicro inversions as the inverted regions buried within local alignments. With this definition, most of the "ultramicro" inversions are expected to be smaller than the "relatively small" inversions which are identified as a single alignment or a string of local alignments from the recent comparative genomics. Within the local alignments, the ultramicro inversions would be misaligned forwardly. I developed a method for identifying such ultramicro inversions hidden within regions of local pairwise-alignments rich in mismatches and gaps. In such regions, erroneously aligned ultramicro inversions would possess high density of mismatches and gaps. Assuming that the sequence differences are spread across the genome following a negative binominal distribution, I determined if these regions could be aligned inversely with greater similarity than the forward alignments (See Methods). A simulation was conducted in order to test the strength and accuracy of this algorithm using the Indelible program (Fletcher and Yang 2009) for evolving random sequences. Eleven sets of pairwise nucleotide alignments were generated allowing the creation of indels, each consisting of 100,000 pairs of 5,000 bp random nucleotide sequences, with parameters (e.g., differences and base composition) equivalent to the human–chimpanzee genome alignments. A short (5–50 bp) segment with fixed length was randomly chosen and inverted in one sequence of every pair. Re-aligned pairwise sequences were then subjected to the inversion identification.

Through the identification of inversions, the sensitivity of the algorithm was found to approximately range from 0.75 to 0.96 (Fig. 2-3**A**). Although ultramicro

inversions of ≥ 20 bp showed slightly higher sensitivity than the others, it is clearly indicated that this method is useful for identifying any sizes of inversions of ≥ 5bp buried in local alignments. In addition, only 4.4 false positives were found on average in a set of 100,000 pairs of sequences, indicating a very low false-positive rate. Only 21 false positives were expected in the human–chimpanzee alignments consisting of 2.4 Gb of alignment sites.

Identification of inversions between the human and chimpanzee genomes

I detected 3,330 ultramicro inversions hidden within the one-to-one alignments between the human and chimpanzee genomes. Interestingly, 871 inversion segments consisted of adenine and thymine exclusively (AT-exclusive inversions) (Table 2-1, Fig. 2-4**A**, and Supporting data D2-2). In addition, AT content of the inversions excluding the AT-exclusive segments was higher (66.5%) than the average AT content of the entire human–chimpanzee alignments (59.3%) (Fig. 2-4**A**), suggesting that ultramicro inversions preferentially occurred in AT-rich regions. The AT-exclusive regions possess considerably different conditions from the other regions to the extent of AT content and thus might show different power for the inversion identification from that assumed in the simulation in the previous subsection. In order to validate the strength and accuracy of the inversion identification methods for the AT-exclusive regions, I conducted a simulation under the same prior conditions except for different base compositions of the AT-exclusive regions (50% adenine and 50% thymine). Sensitivity for searching for the true ultramicro inversions in simulation in the AT-exclusive condition was less than that in the initial condition and, as well as the initial condition, increased with increasing inversion size from 0.503 for 5bp inversion to 0.868 for 50 bp inversion (Fig. 2-3**B**).

False positives were greater in the AT-exclusive conditions than in the initial simulation condition with less than 116 false positives on average in the AT-exclusive simulation set (Fig. 2-3**B**). Although approximately 560 false positives were expected in the 2.4 Gb AT-exclusive alignments, which was as large as the human-chimpanzee genome alignments, the number of false positives in the human-chimpanzee genome alignments may have been much lower than 560 since the AT blocks constitute small fractions of the genome. Blocks consisting of a series of at least 5 bp of adenines and thymines were summed at approximately 270 Mb in the human genome, in which all the 871 AT-exclusive inversions were included, indicating that less than 62 false positives of the AT-exclusive inversions could be expected in the human–chimpanzee alignments.

The size of ultramicro inversions between the human and chimpanzee genomes ranged from 5 to 125 bp (Fig. 2-4**B** and Supporting data D2-2), and the distribution of their lengths, which was classified into three characteristics, showed a peculiar shape. The distribution basically decayed in a fashion similar to the macroscopic and relatively small inversions between the human and chimpanzee genomes (Feuk, et al. 2005), suggesting that ultramicro inversions prefer small size. On the other hand, I found a small peak around 20 bp in the size distribution of the ultramicro inversions.

While the ultramicro inversions as well as the macroscopic and relatively small inversions were spread throughout the human genome (Feuk, et al. 2005), the density of inversions was significantly different on chromosome Y compared with that on the autosomes (Fig. 2-4**D**). Autosomes averaged 0.401 ± 0.176 AT-exclusive and 1.05 ± 0.210 guanine and cytosine-including (GC-including) inversions per Mb. However, chromosome Y possessed much more frequent GC-including inversions: 0.676 AT-exclusive inversions (p=0.119) and 1.94 GC-including inversions ($p < 1.00\times10^{-5}$)

per Mb. In contrast, the numbers of AT-exclusive and GC-including inversions on chromosome X (0.637 and 1.34 per Mb, respectively) were not significantly different from those on the autosomes (p = 0.180 and 0.168, respectively). In addition, the proportions of the inversion ratios between chromosome Y and autosomes (3.45 times) are larger than the proportion of the mutation rates (approximately 1.4 times (Hughes, et al. 2010)) between them. These observations suggest that the abundance of ultramicro inversions in chromosome Y is mainly subject to high diversity of the genomic structures specifically in chromosome Y (Hughes, et al. 2010) rather than male driven evolution. One possibility is that frequent intrachromosomal recombinations in chromosome Y (Hughes, et al. 2010) had been involved in frequent ultramicro inversions. Another is absence of the proofing by homologous recombination in Y chromosome (Li and Heyer 2008). Microdeletions (>100 bp), for example, are specifically frequently accumulated in chromosome Y and occasionally cause infertility in male (Foresta, et al. 2001; Onrat, et al. 2012; Pryor, et al. 1997).

Ultramicro inversions validated by phylogenetic profiles

By comparing the ultramicro inversions within the human–chimpanzee alignments with the orthologous sequences of the primate outgroups, the lineages in which the inversions occurred can be inferred (Fig. 2-5**A**). Generating multiple alignments of ultramicro inversions concatenating their neighbors of human, chimpanzee, gorilla, and/or orangutan as outgroups, the species possessing the inverted sequences were identified. In 2,613 ultramicro inversions out of 3,330, the lineages in which the sequences had inverted were definitely determined (Table 2-1). Seven hundred and thirteen and 1,757 inversions were identified specifically in the human and chimpanzee

sequences, respectively, suggesting that they had occurred specifically in the human and chimpanzee lineages after the separation between the two species. On the other hand, 143 inversions appeared to be inconsistent with the species phylogeny among human, chimpanzee, and gorilla, suggesting incomplete lineage sorting in the common ancestor of these three species: 78 ultramicro inversions shared between human and gorilla and 65 between chimpanzee and gorilla (Table 2-1). The former represented the gene phylogeny as ((Human, Gorilla), Chimpanzee) and the latter represented the gene phylogeny as (Human, (Chimpanzee, Gorilla)).

While my detection method for ultramicro inversions possessed a high degree of accuracy, it is noteworthy that these 2,459 ultramicro inversions were also supported by the phylogenetic profiles of the outgroups. Thus, I considered that these inversions were very plausible. Out of the rest of 871 ultramicro inversion, I could not obtained the strong support by phylogenetic profiles in 153 ultramicro inversions and the orthologous sequences of gorilla or orangutan in 718 ultramicro inversions. The fraction of the latter ultramicro inversions (22%) is larger than the loss-ratio of the size the whole genome alignments of human, chimpanzee, and gorilla (2.1 Gb) from those of human and chimp (2.4 Gb) (p=8.0×10^{-53}), still suggesting that the ultramicro inversion had occurred in the instable regions of genome structure such that the orthologous regions were not conserved in the outgroup species genomes.

Ultramicro inversions within genes

To examine the impact of ultramicro inversions on gene evolution in the human lineage, I searched for ultramicro inversions within those exons defined in H-InvDB (Imanishi, et al. 2004), and found a total of 90 inversions (Fig. 2-5**B**). More than half the

inversions were identified in the 3′ UTR region. Although 28 ultramicro inversions out of 90 were found in the coding regions, most of them (15 ultramicro inversions) were inferred to have occurred in the chimpanzee lineage specifically, and one was inferred to have occurred in the human lineages (Fig. 2-5**B**). The 17 genes of which ultramicro inversions were identified in the coding regions included several well-annotated ones such as tumor protein p73 (TP73), protein tyrosine phosphatase receptor type B (PTPRB), and NADPH oxidase organizer 1 (NOXO1). In 27 out of 28 inversions, biochemically different amino acids were observed between human and chimpanzee. These inversions ranges from five to 24 bp and affected the corresponding amino acid sequences from two to nine residues. In the remaining one, a stop codon was observed in the human sequence but not in the chimpanzee sequence: only four amino acids were extended in the chimpanzee sequence. These observations suggest that ultramicro inversions in coding regions have contributed to gene evolution mainly in the chimpanzee lineage.

2.5. Discussion

I developed a highly sensitive and distinctly specific method for identifying ultramicro inversions hidden within nucleotide alignments. This method could be very effective for sequences with average base compositions of the human and chimpanzee genomes as well as would work well for those with extremely biased base compositions such as 100% AT content (Fig. 2-3**B**) with extra filtering for simple repeats. Positive predictive values (number of true positives/(number of true positives + number of false positives)) are more than 0.9999 for the former case and 0.997 for the latter case, respectively. Although our previous method for identifying ultramicro inversions did not possess high sensitivity for 5 and 6 bp of ultramicro inversions, new method in this study overcame this weakness: sensitivities for 5 and 6 bp at 0.26 and 0.65 in previous method, respectively, those at 0.85 and 0.75, in the new method, respectively. Moreover, sensitivities for AT-exclusive inversions were also improved: 0.16 increase on average. It is noted that the new method shows positive predictive numbers equivalent to or slightly larger than that of previous one, 0.9998 for the average base compositions of the human and chimpanzee genomes and 0.993 for the100% AT content.

In addition to macroscopic and relatively small inversions, a large number of ultramicro inversions, ranging from five to 125 bp, were detected between the human and chimpanzee genomes using my method (Table 2-1). From this observation, I defined the size of ultramicro inversions equal to or less than 125 bp. Based on the simulations, at most approximately 82 false positives (2.5% of the total) were expected. On an average, 1.39 ultramicro inversions were found per Mb of the human–chimpanzee alignments. These inversions had been treated as mismatches and

gaps in the local alignment, suggesting that the identification of ultramicro inversions is one of the effective ways for fine-tuning the local alignments. However, I found only 0.0319% and 0.126% of mismatches and gaps in the whole human-chimpanzee genome alignments were derived from the ultramicro inversions, respectively. The nucleotide divergence between chimpanzees and humans before and after excluding the ultramicro inversions was estimated at 0.013276 and 0.013270, respectively, indicating that ultramicro inversions do not largely affect the nucleotide divergence between human and chimpanzee. Still, because of the relatively low sensitivity in detecting extra-short and palindrome-like inversions, the number of ultramicro inversions may be greater within the human–chimpanzee alignments. One of my most important findings was the large fraction of AT-exclusive ultramicro inversions (Fig 2-4A). My method included additional filtering of AT-exclusive inversions, which excluded inversions consisting of mono- or dinucleotide repeats of A and T. The simulation produced indicated a very high positive predictive rate. However, some of the AT-exclusive inversions may have been false positives because of the unknown aspects of genomic evolution. Filtering inversion candidates using the phylogenic profile would generate a highly specific subset of inversions (Chaisson, et al. 2006). By comparing inverted segments with the primate outgroup, 666 of the AT-exclusive inversions belonged to this specific subset (Table 2-1), still suggesting frequent AT-exclusive inversions.

Size distribution of the ultramicro inversions implies that more ultramicro inversions with smaller (<5 bp) size are buried in the local alignments. However, it would be very difficult to distinguish the real inversions <5bp form the stochastically-generated artifacts. Although it would be hard verify the candidates one-by-one, the whole pictures of the minute ultramicro inversions can be illustrated. If

such small inversions have occurred in the genome, specific arrays of four or less mismatches were preferably found in the alignments. All patterns of the mismatch arrays can be investigated if the lengths of the arrays are short. The inversions ranging from 23 to 125 bp could be any one of the ultramicro inversions hidden in a local alignment or small-size inversions identified as a single or a string of local alignment (Feuk, et al. 2005). Size distribution of the inversions roughly indicated that inversions less than 40 bp were preferably hidden in the local alignments between the human and chimpanzee genomes (Fig. 2-4**C**). The ultramicro inversions are also distinguished from the "relatively small" inversions that are detectable as a single alignment or a string of local alignments, in that the exact boundaries of ultramicro inversions can be identified easily within the local alignment. This may have a significant insight into the elucidation of the mechanism for the ultramicro inversions.

In this study, the human–chimpanzee alignments were generated by the G-compass pipeline (Fujii, et al. 2005). Although the G-compass pipeline is different from the UCSC axtNet alignment (Schwartz, et al. 2003) based on the definition of orthologous alignments, both methods initially generate local alignments with blastz (Schwartz, et al. 2003) or its successor lastz (Harris 2007). Thus an equivalent number of ultramicro inversions are likely to be obtained from the UCSC axtNet alignment. As expected, 3,036 ultramicro inversions were found in the human–chimpanzee alignments using UCSC axtNet alignment, suggesting that most of the ultramicro inversions are independent of the G-compass pipeline. Although I have not examined for ultramicro inversions within the genome alignment generated by local alignments other than blastz, differences in the ultramicro inversions between different alignment algorithms may be helpful in verifying the behavior of the alignment algorithms involving ultramicro

inversions either erroneously aligned or excluded from the local alignments.

Out of the 2,613 ultramicro inversions validated by phylogenetic profiling, 1,767 were found to have occurred specifically in the chimpanzee lineage, which were approximately twice more than those (713) in the human lineage (Table 2-1). Several studies have indicated that the sequence accuracy of the chimpanzee genome is poorer than that of the human genome (Hobolth, et al. 2011; Meader, et al. 2010) because of the lower coverage. This may be one of the causes of the abundance of ultramicro inversions in the chimpanzee lineage. However, the substitutions especially those in the chimpanzee lineage were at most 1.05 times more than those in the human lineage (Hobolth, et al.), indicating that a large number of ultramicro inversions in the chimpanzee lineage were unlikely to be the result of sequence errors. Higher level of false assembles of the sequence reads in the chimpanzee genome than the human's might be another explanation. It can be a cause for the false positives in the larger inversions as a single alignment or a string of local alignments (relatively small inversions) than ultramicro inversions. However, this may be also difficult to explain ultramicro inversions within a local alignment. Thus, the differences in inversion frequencies between humans and chimpanzees give an insight into the different histories of genomic structural changes between the two species. Furthermore, this observation ensures the abundance of ultramicro inversions in coding regions found specifically in the chimpanzee lineage. As shown in Fig. 2-5**A**, ultramicro inversions substitute more than one amino acid at a time into physicochemically different ones. The inversion in PTPRB genes in chimpanzee (Fig. 2-5**A**) had altered a string of two residues of glutamine and glycine into physicochemically different ones, (Q/H)1229P and G1230C, respectively. In contrast, the hydrophilic residue of (Q/H)1229 is conserved across

amniotes, and G1230 is conserved across tereosts and tetrapods. This implies that the ultramicro inversion had altered the function of the corresponding fibronectin type III domain. This implies that such ultramicro inversions played a role in drastic protein evolution in the chimpanzee lineage.

Although it has not been clear how ultramicro inversions have occurred, my findings of frequent ultramicro inversions in chromosome Y and the AT-exclusive regions suggests that ultramicro inversions are preferably located in those genomic regions that may relate to genomic instability. To examine the relationship between ultramicro inversions and genomic instability in detail, I compared the positions of ultramicro inversions with those of the genomic features involved in stability of the human genome. Ultramicro inversions significantly frequently overlapped on the boundaries of *L1* and *Alu* ($p < 0.001$) and were located in the region (<100 bp) closer to these transposable elements ($p < 0.001$), strongly suggesting that ultramicro inversions are associated with the genomic features generating instability as previously indicated by the macroscopic inversions (Lee, et al. 2008).

To elucidate the mechanisms of ultramicro inversions at the molecular level, I examined the genomic features near the inverted segments and found that large parts of ultramicro inversions were included within DNA stem-loop structures. This indicates that cruciform formation with DNA stem-loops mediated the ultramicro inversions (Kolb, et al. 2009). Inversions that possibly mediated by inverted repeats were found in various lineages and in various sizes (Carvalho, et al. 2011; Cui, et al. 2012; Furuta, et al. 2011; Mott and Symington 2011). Especially, such inversions in the chloroplast genomes in plants showed minute size at least 6 bp (Kim and Lee 2004) and occasionally consisted of the hotspots of inversion (Whitlock, et al. 2010). I found 1,041

ultramicro inversions (31% of total) in the human-chimpanzee genome alignments sandwiched by or including inverted repeats (Fig. 2-6**A**), which was significantly more frequent than expected ($p < 0.001$). This observation suggests that inversion via stem-loop structures following cruciform formation is one of the major mechanisms for generating ultramicro inversions. The ultramicro inversion in Fig. 2-6**B** is sandwiched between inverted repeats and was possibly generated via the cruciform formation (Fig. 2-6**C**). The inverted repeat next to the 3′ end of the ultramicro inversion segment is specifically found in the human sequence. This implies that inversion follows double-strand breaks, strand displacement by the invading 3′ end, *de novo* DNA synthesis, and concomitant DNA elongation (Kolb, et al. 2009). It is noted that three fourth of the ultramicro inversions may be explained other mechanisms. My observation indicated that ultramicro inversions are related to the genomic features involved in genomic instability, which is a characteristic similar to that of macroscopic inversions.

Up to here, I focused on the ultramicro inversions in the human and chimpanzee genomes. On the other hand, such ultramicro inversions would be found in different lineages. In order to examine the ubiquity of the ultramicro inversions across the living world, I searched for the ultramicro inversions between the genome alignments of closely related species or strains in various lineages. I generated the genome alignments between closely related species or strains of fly, budding yeasts, rice, four bacteria lineage, and two archaea lineages and identified ultramicro inversions buried in the alignments (Fig. 2-7**A**): from 13 ultramicro inversions in *Escherichia coli* strains to 1,285 in fly. Since the size of genomes and evolutionary distances between the species or strains varied across the lineages, I compared the numbers of the ultramicro inversions per 10,000 mutations across the lineages. These values were up to 6.8 times

different between the lineages (0.897 for rice to 6.06 for *Streptococcus pyogenes* strains). On the other hand, the ratio of the AT-exclusive inversions much varies across the lineages. Any AT-exclusive inversions were not found in some of the lineages while the variation of the fraction of At-exclusive ultramicro inversions seemed to be independent of the average AT-content of the genomes (Fig. 2-7**B**). In addition, I found the frequencies of the ultramicro inversions included within the DNA stem-loops also vary across the lineages (Fig. 2-7**C**). These observations suggested that the mechanisms of the ultramicro inversions are various across the lineages though ultramicro inversions are ubiquitous across the three domains of life.

I developed an effective method for identifying ultramicro inversions within pairwise alignments and found a large number of ultramicro inversions within the local alignments between the human and chimpanzee genomes. This is the first finding of an abundance of such inversions within the local alignments between closely related species. This observation suggests that a considerable number of ultramicro inversions could be found within the alignments between individuals from different populations. Furthermore, some of the adjacent SNPs may be ultramicro inversions as well as large inversions observed in HapMap data (Bansal, et al. 2007). Identification of ultramicro inversions within human populations may be helpful in elucidating how phenotypic characteristics have diversified during human evolution. These observations strongly indicate that my inversion-identification method was helpful for examining the impact of minute structural changes in genomes. This method is also applicable in fine-tuning of genome alignments by distinguishing ultramicro inversions from nucleotide substitutions and indels. This improvement of the alignments would be required for the comparative genomics between closely related species at high accuracy.

Fig. 2-1. Ultramicro inversions buried in a local alignment.

When one of the sequences of the mismatch and gap-rich region is altered in a complementary manner of nucleotides (within an orange rectangle in upper figure), the two sequences are identically aligned (lower figure).

```
S1 CATCTTCGGAGGAACGAGC---CGGAATTGC
   ||||| ||   |  ||  |   || ||||||
S2 CATCTCCG---GCTCGTTCCTCCGAAATTGC

S1 CATCTTCGGAGGAACGAGCCGGAATTGC
   ||||||||||||||||||||||||||||
S2 CATCTTCGGAGGAACGAGCCGGAATTGC
```

Fig. 2-1.

Fig. 2-2. Procedures of the identification method of ultramicro inversions within the alignments.

Candidates of the ultramicro inversions were extracted by the criteria (**A**) or (**B**), and such regions with the neighboring regions were globally aligned with Dynamic Programming assuming inversions (**C**). In (**C**), red and blue lines represent forward and inverted alignments, respectively.

A. Trios of the nearest mismatches and gap-blocks significantly closely located

```
CATCTTCGGAGGAACGAGC---CGGAATTGC
|||||  ||    |   ||   |    || ||||||
CATCTCCG---GCTCGTTCCTCCGAAATTGC
     TCGGAGGAACGAGCCGG
```

B. Pairs of the gap-blocks in different sequences significantly closely located
→assuming partial palindromes

```
GGATTCATGTGTGATCAC-----GAAAT
|||||      ||||||||      |||||
GGATT-----GTGATCACACATGGAAAT
      CATGTGTGATCAC
```

C

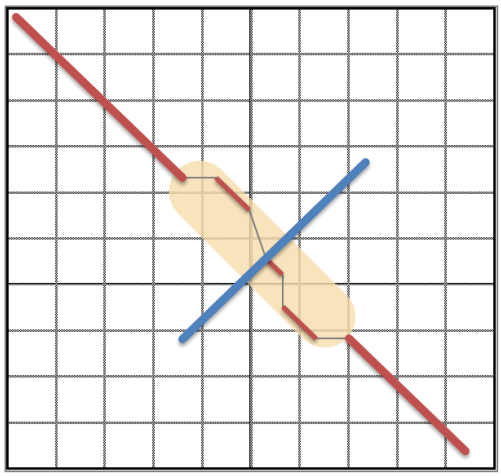
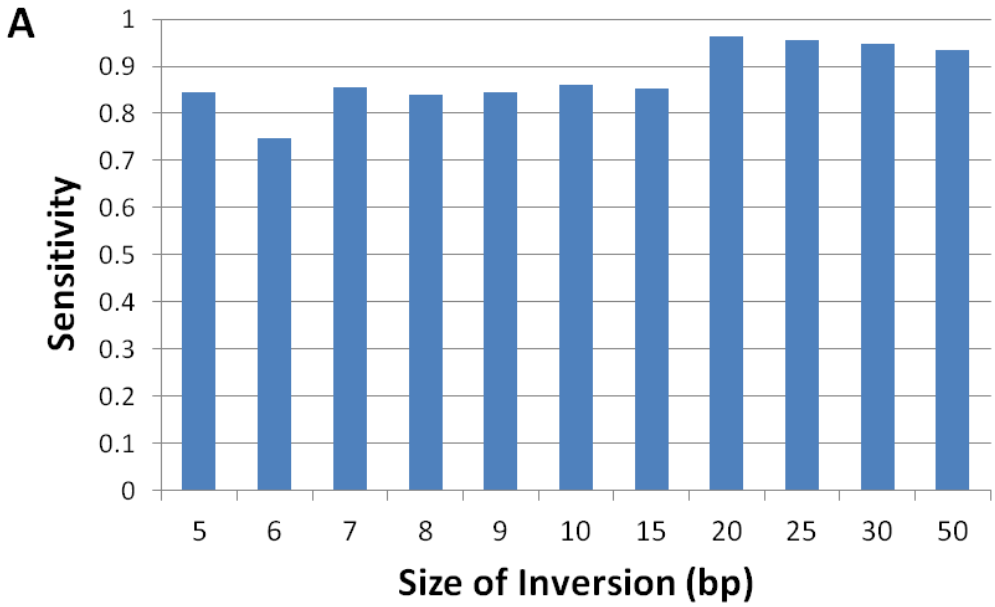


Fig. 2-2

Fig. 2-3. Results of the simulations of ultramicro inversion identification.

Sensitivities (blue lines) and numbers of false positives (red bars) for the simulations, assuming sequence parameters equivalent to the alignments between the human and chimpanzee genomes (**A**) and those of the AT-exclusive condition (**B**).


A
Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
6
7
8
9
10
15
20
25
30
50
Size of Inversion (bp)

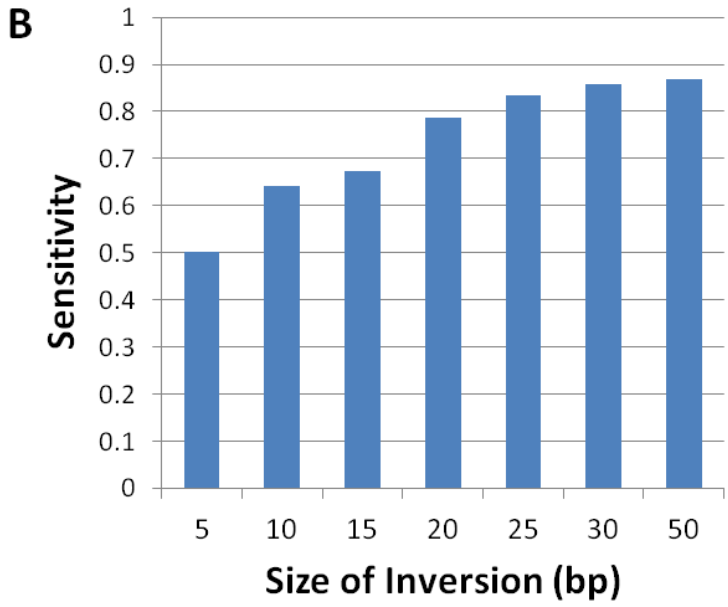

B
Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
50
Size of Inversion (bp)

Fig. 2-3

Fig. 2-4. AT content, size, and chromosomal distribution of ultramicro inversions. Distributions of ultramicro inversions over the ranges of AT content (**A**), sizes in nucleotides (**B** and **C**), and chromosomes (**D**). The red and blue bars represent the numbers of AT-exclusive and GC-including inversions. In (**A**), the average of AT content in the entire genome alignment between the human and chimpanzee genomes and that in GC-including inversions are also shown. In (**C**), the size distribution of the inversion by Feuk *et al*. (2005) is also shown.

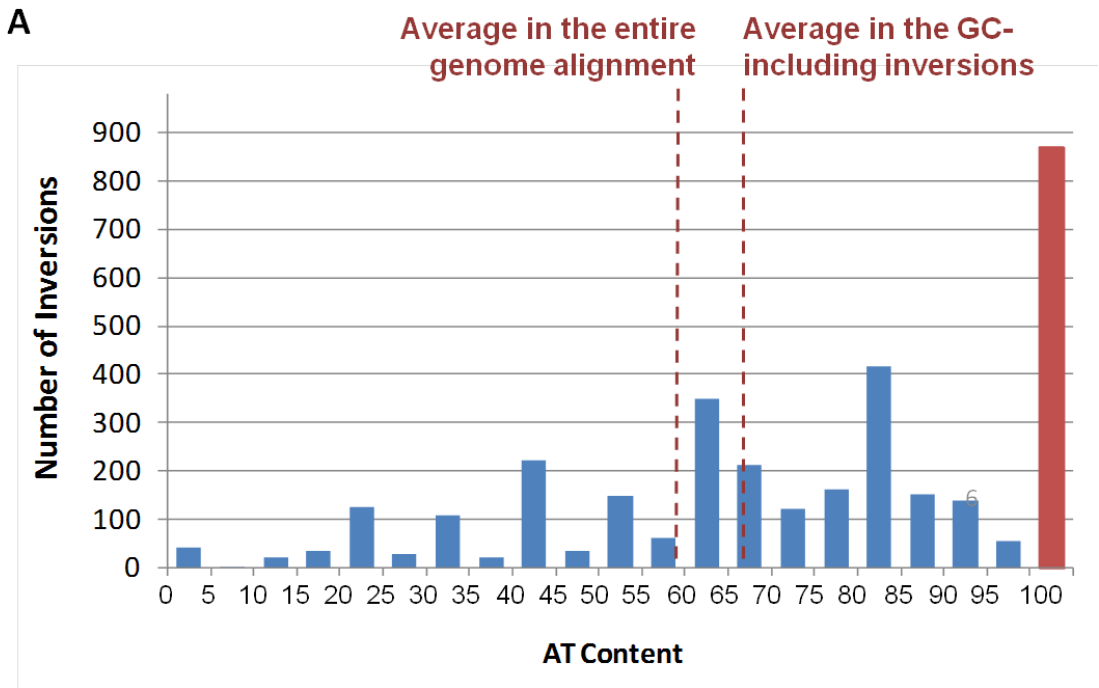
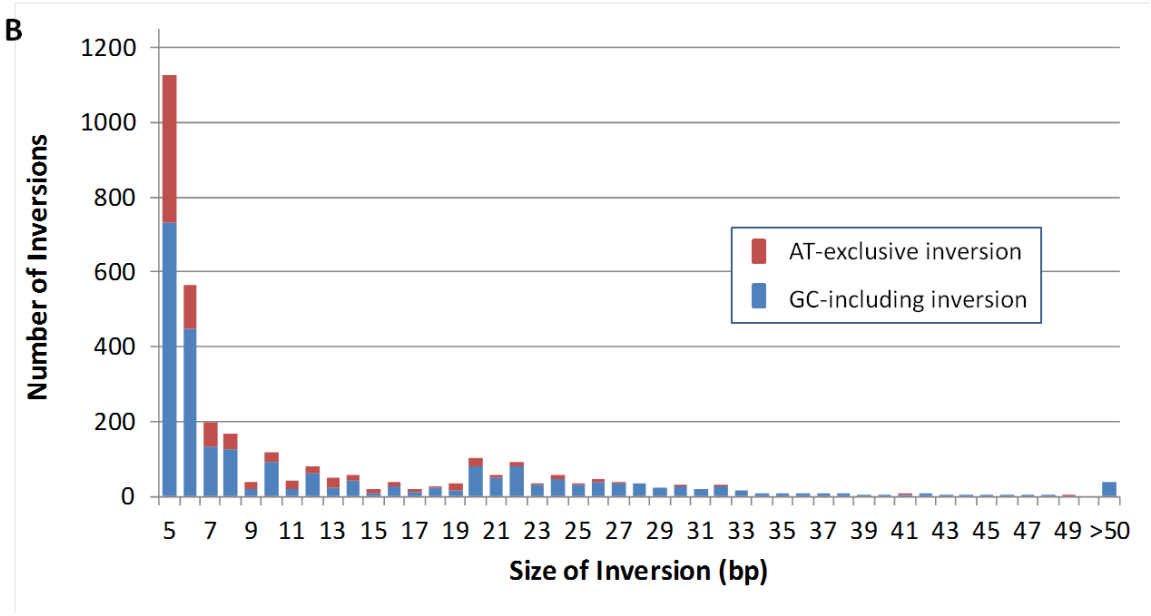

A
Average in the entire genome alignment
Average in the GC-including inversions
Number of Inversions
900
800
700
600
500
400
300
200
100
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
AT Content

Fig. 2-4


B
Number of Inversions
1200
1000
800
600
400
200
0
AT-exclusive inversion
GC-including inversion
5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 >50
Size of Inversion (bp)

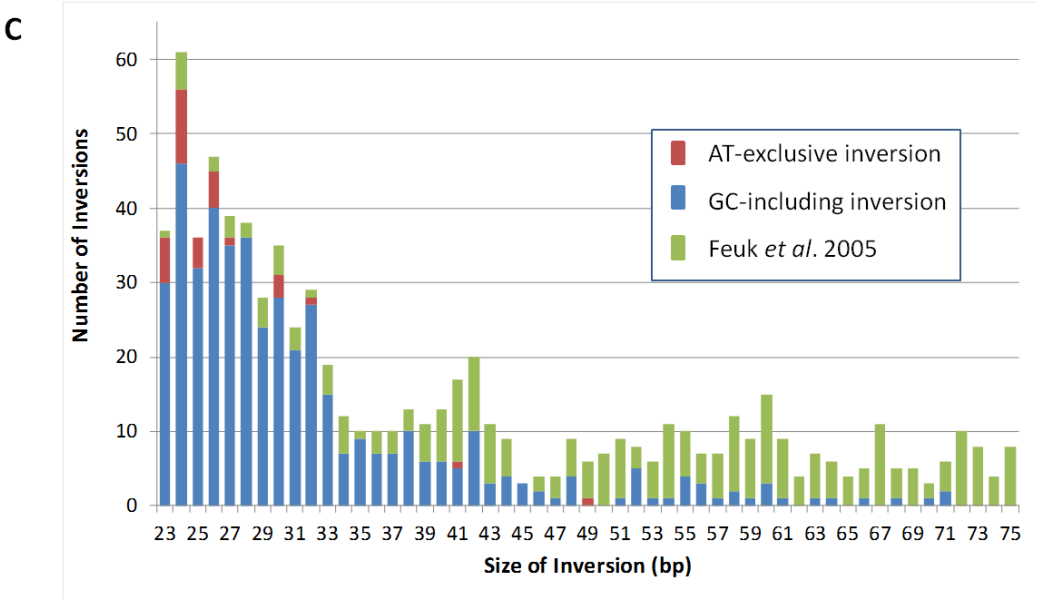

C
Number of Inversions
60
50
40
30
20
10
0
AT-exclusive inversion
GC-including inversion
Feuk *et al*. 2005
23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75
Size of Inversion (bp)

Fig. 2-4 (cont.)

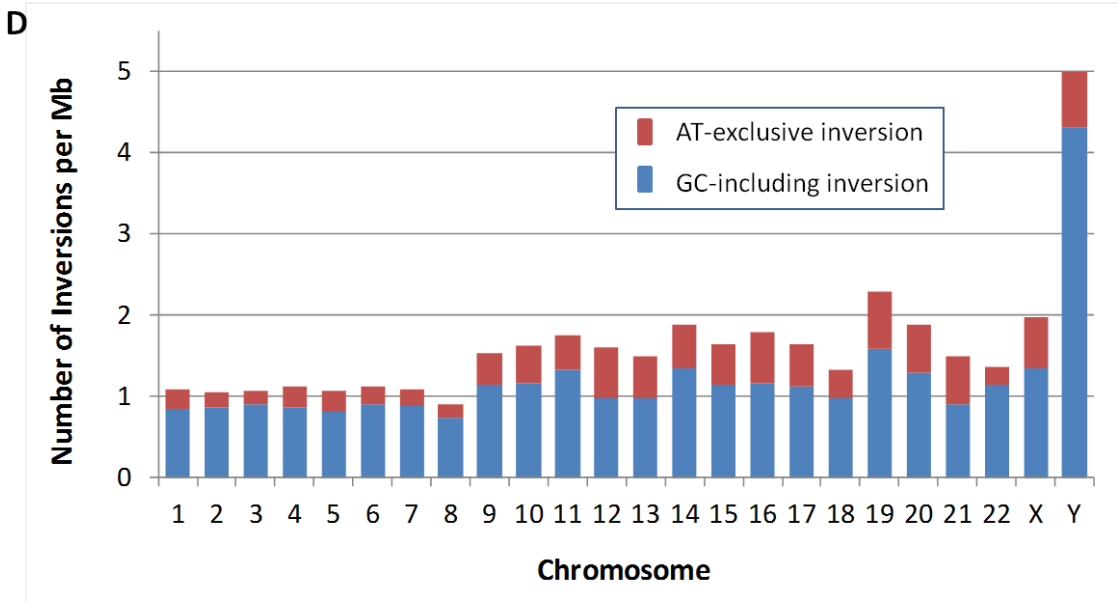

D
Number of Inversions per Mb
0
1
2
3
4
5
AT-exclusive inversion
GC-including inversion
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome

Fig. 2-4 (cont.)

Fig. 2-5. Ultramicro inversions found within genes.

(**A**) The multiple alignment around the ultramicro inversions specifically identified in the chimpanzee lineage in receptor-type tyrosine-protein phosphatase beta genes (PTPRB) and the genomic structures of a part of PTPRB transcripts in the human (HIT000321866 from H-InvDB) and chimpanzee genomes (XM_509219 from Refseq). This inversion is included in one of the Fibronectin type III domains in a tandem array in PTPRB protein. Blue characters indicate ultramicro inversions. Numbers within the boxes represent the exon numbers. The genomic regions with green and red backgrounds are subject to one-to-one alignment, and the red background corresponds to the multiple alignment. (**B**) Venn diagram of the ultramicro inversion frequencies in coding region sequences (CDS), 5′ UTR, and 3′ UTR. Numbers in parenthesis represent the ultramicro inversion frequencies that were inferred to have occurred specifically in the human and chimpanzee lineages, respectively. No ultramicro inversions in the genes showed the incomplete lineage sorting among human, chimpanzee, and gorilla.

A

```
              3'                                                  5'
Chimpanzee    gacaaggacctgacggagtggcggtttCCTTGtcttgttcctggaagg
Human         gacaaggacctgacggagtggcggtttcaaggccttgttcctggaagg
Gorilla       gacaaggacctgacggagtggcggtttcaaggccttgttcctggaagg
Orangutan     gacaaggacctgacggagtggcggtttcacggccttgttcctggaagg

              N                   C
Chimpanzee    DKDLTEWRFPCLVPGR
Human         DKDLTEWRFQGLVPGR
Gorilla       DKDLTEWRFQGLVPGR
Orangutan     DKDLTEWRFHGLVPGR
```

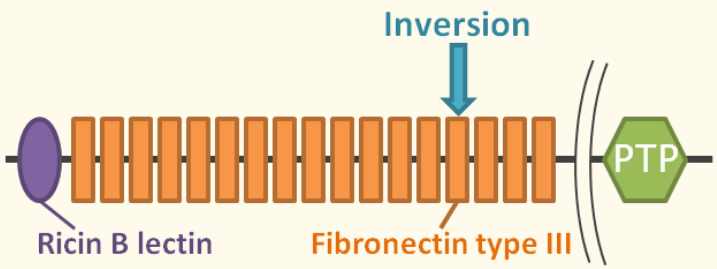


Fig. 2-5

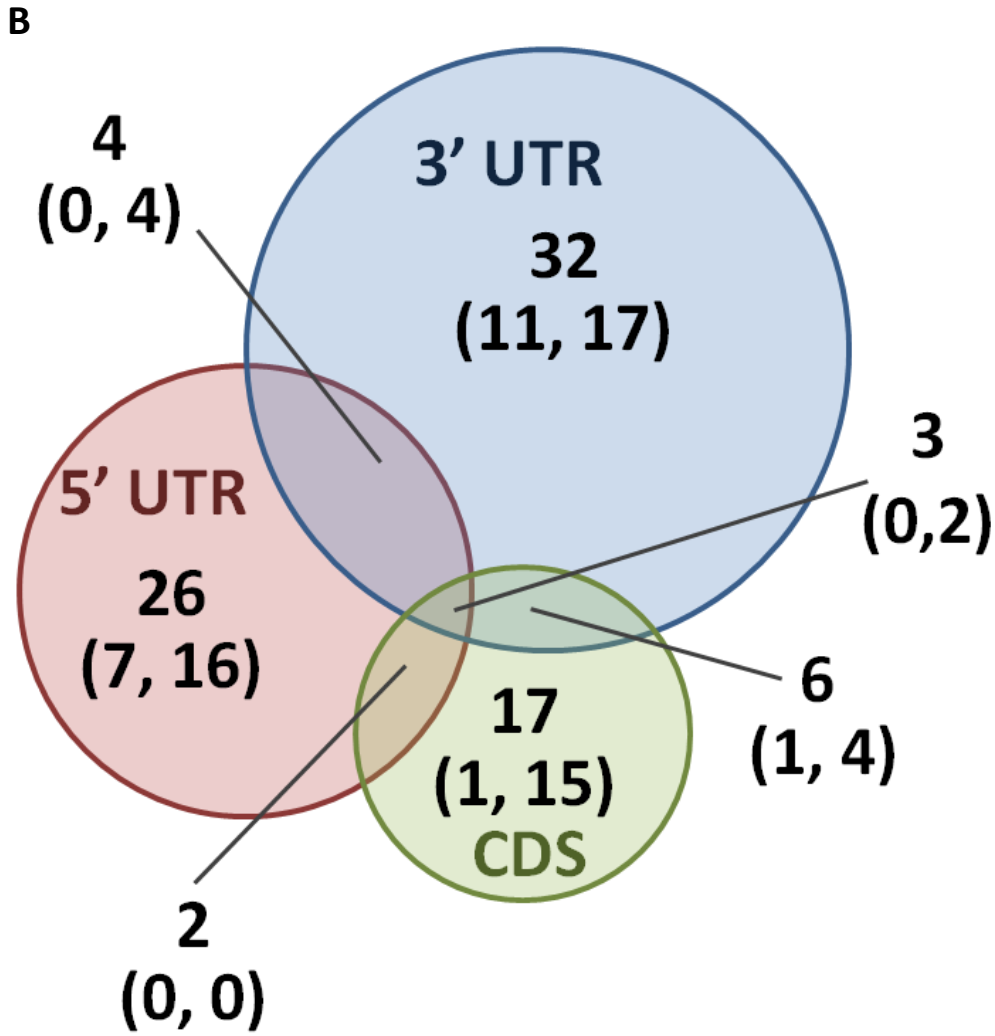


Fig. 2-5 (cont.)

Fig. 2-6. Candidates of cruciform-mediated ultramicro inversions.

(A) Ultramicro inversions possibly shaping the stem-loop structures. The left figure represents the ultramicro inversions (blue circles) within a loop and sandwiched by inverted repeats (green circles). In this case, ultramicro inversions located within the loop with ≤5 bp spacer regions at both ends were required. The right figure represents the ultramicro inversions including a stem or the inverted repeats. In both cases, stems of ≥7 bp were required. The lower figure represents the numbers of ultramicro inversions possessing these characteristics. (**B**) Nucleotide alignments of human chromosome (chromosome 4: 39332472-39332603) and its orthologous sequences to the chimpanzee and gorilla genomes including the ultramicro inversions sandwiched by inverse repeats. Characters in blue and green represent ultramicro inversions and inverted repeats, respectively. The inverted repeat at the 3′ end of ultramicro inversions may have been inserted by cruciform-formation following inversion. (**C**) One strand of cruciform-DNA inferred by the Mfold program (Zuker 2003). Characters highlighted in blue and green represent the ultramicro inversion and inverted repeats, respectively.

A

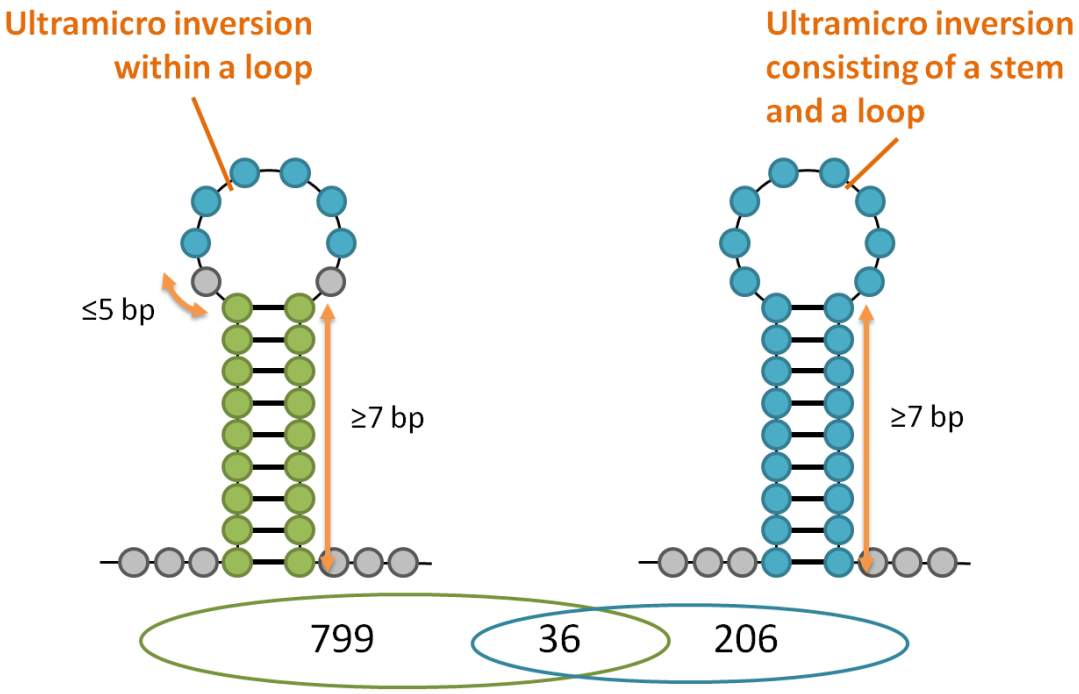

Ultramicro inversion within a loop
Ultramicro inversion consisting of a stem and a loop
≤5 bp
≥7 bp
≥7 bp
799
36
206

Fig. 2-6

B

```
Hg19, chr4: 39332472-39332603

            1                                                50
Human       tcaaccacgttacaaagagtcaacagactaggttcagtcattcATATATA
Chimpanzee  tcaaccacgttacaaagagtcaacagactaggttcagtcattcatatata
Gorilla     Tcaaccacgttacaaagagtcaacagactaggttcagtcattcatatat-

           51                                                98
            --TATAtACACACACACACACACATATATATATATtctccctacctaact
            tgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgttctccctacctaact
            -------gtgtgtgtgtgtgtgtgtgtgtgtgtgttctccctacctaact

           99                                 132
            actggttcttctctgcctctttgccagacgggaa
            actggttcttctctgcctctttgccagacaggaa
            actggttcttctctgcctctttgccagacaggaa
```

C

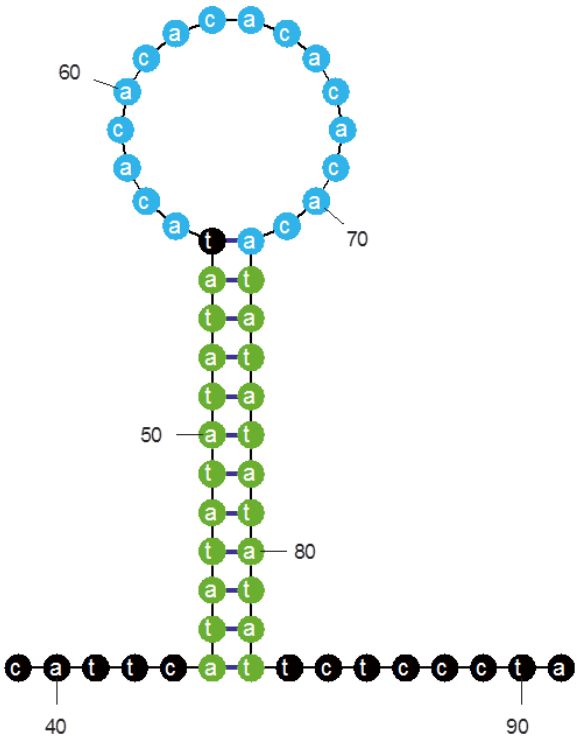
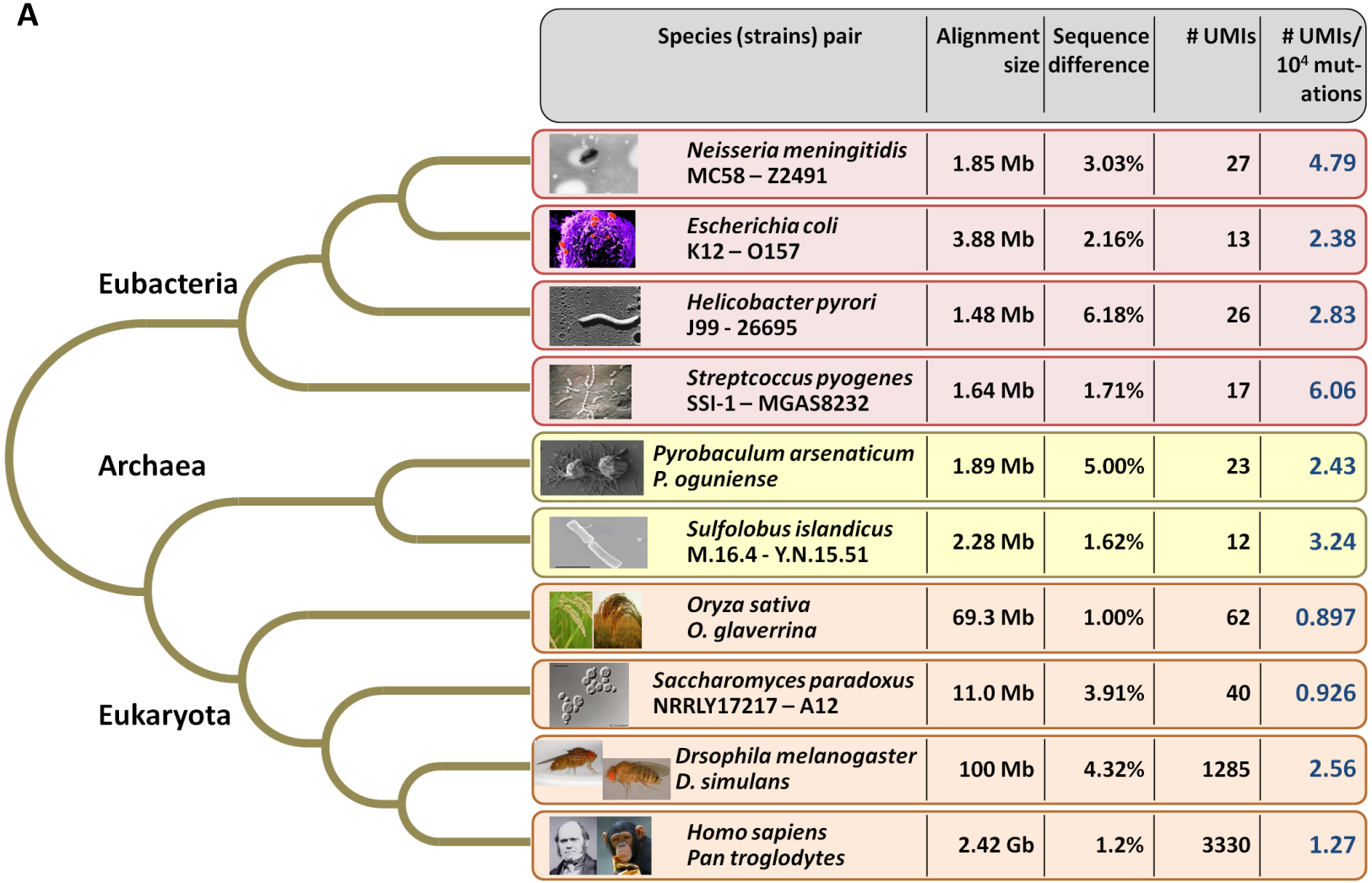


Fig. 2-6 (cont.)

Fig. 2-7. Ultramicro inversions ubiquitously found in the living world.

(**A**) The row representing each lineage in the table was colored according to the domains it belongs two. In the columns, average of genome size between the two species (strains), sequence difference in the whole genome alignment between them, number of ultramicro inversions (UMIs), and that per 10,000 mutations were shown. (**B**) Fractions of AT-exclusive ultramicro inversions and average AT-contents in the ultramicro inversions and the genomes in each lineage. **(C)** Fractions of the ultramicro inversions included within DNA stem-loops in each lineage.

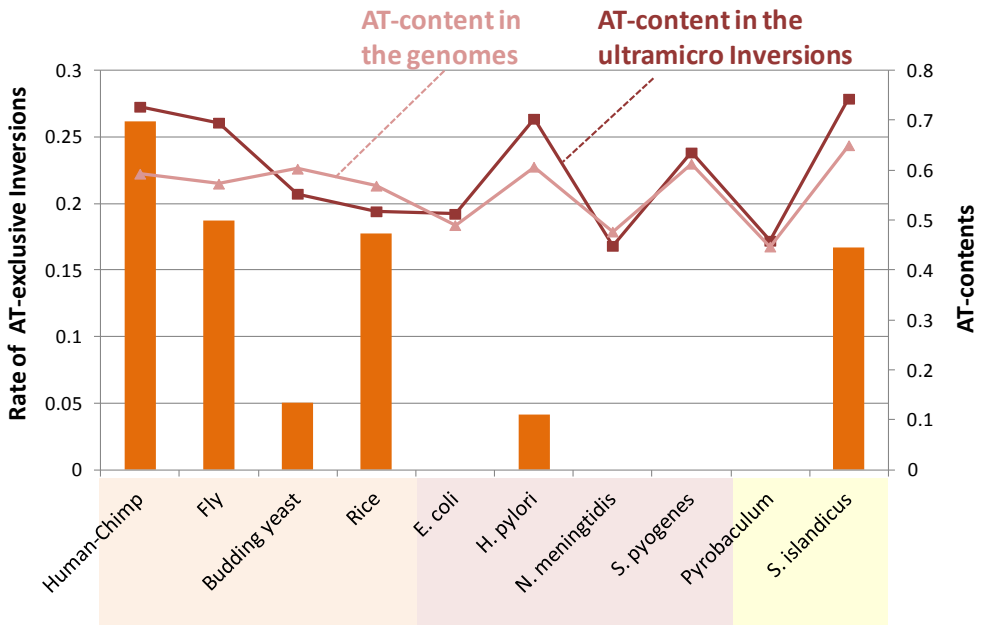
A



Species (strains) pair	Alignment size	Sequence difference	# UMIs	# UMIs/ 10^4 mut-ations
Neisseria meningitidis MC58 – Z2491	1.85 Mb	3.03%	27	4.79
Escherichia coli K12 – O157	3.88 Mb	2.16%	13	2.38
Helicobacter pyrori J99 - 26695	1.48 Mb	6.18%	26	2.83
Streptcoccus pyogenes SSI-1 – MGAS8232	1.64 Mb	1.71%	17	6.06
Pyrobaculum arsenaticum *P. oguniense*	1.89 Mb	5.00%	23	2.43
Sulfolobus islandicus M.16.4 - Y.N.15.51	2.28 Mb	1.62%	12	3.24
Oryza sativa *O. glaverrina*	69.3 Mb	1.00%	62	0.897
Saccharomyces paradoxus NRRLY17217 – A12	11.0 Mb	3.91%	40	0.926
Drsophila melanogaster *D. simulans*	100 Mb	4.32%	1285	2.56
Homo sapiens *Pan troglodytes*	2.42 Gb	1.2%	3330	1.27

Fig. 2-7

B



C

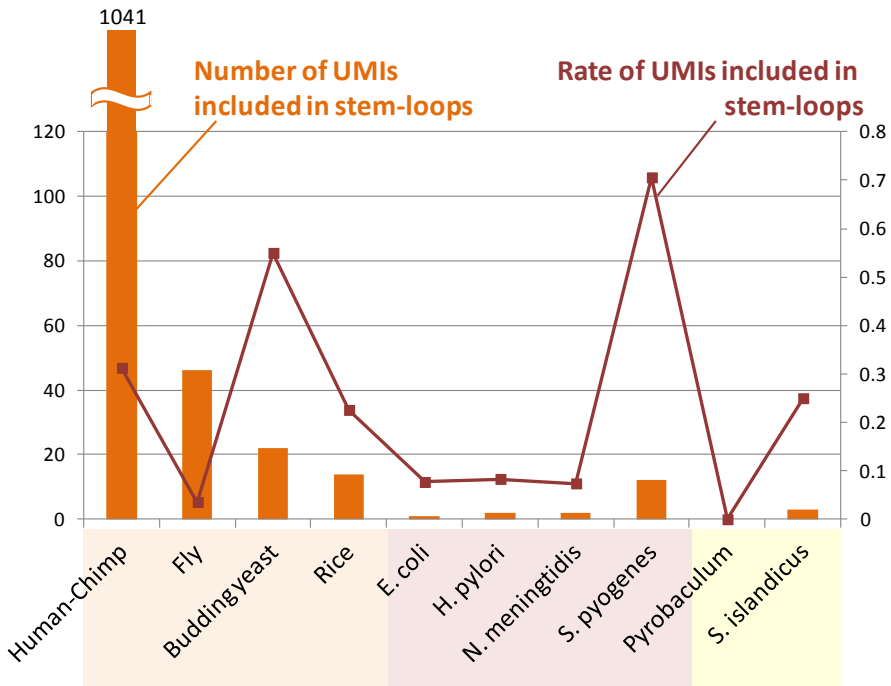


Fig. 2-7 (cont.)

Table 2-1. Overview of ultramicro inversions within alignments between the human and chimpanzee genomes.

Inversion classification and the lineage of inversion events	GC-including	AT-exclusive	Total
Ultramicro inversions	2459	871	3330
Inversions obtained with the support of phylogenetic profiles	1947	666	2613
Human lineage	587	126	713
Chimpanzee lineage	1268	489	1757
Human–Gorilla lineage*	48	30	78
Chimpanzee–Gorilla lineage*	44	21	65

*Subject to incomplete lineage sorting among human, chimpanzee, and gorilla.

Supporting data D2-1. Human-chimpanzee orthologous genome alignments.

Supporting data D2-2. List of ultramicro inversions in the human-chimpanzee alignments.

(Attached in DVD-ROM).

2.6. References

Bansal V, Bashir A, Bafna V 2007. Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Res 17: 219-230.

Britten RJ 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc Natl Acad Sci U S A 99: 13633-13635.

Carvalho CM, et al. 2011. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. Nat Genet 43: 1074-1081.

Chaisson MJ, Raphael BJ, Pevzner PA 2006. Microinversions in mammalian evolution. Proc Natl Acad Sci U S A 103: 19824-19829.

Chen ZZ, et al. 2004. A space-efficient algorithm for sequence alignment with inversions and reversals. Theoretical Computer Science 325: 361-372.

Consortium CSaA 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.

Cui L, Neoh HM, Iwamoto A, Hiramatsu K 2012. Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. Proc Natl Acad Sci U S A 109: E1647-1656.

Feuk L, et al. 2005. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet 1: e56.

Fletcher W, Yang Z 2009. INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26: 1879-1888.

Foresta C, Moro E, Ferlin A 2001. Y chromosome microdeletions and alterations of spermatogenesis. Endocr Rev 22: 226-239.

Fujii Y, et al. 2005. A web tool for comparative genomics: G-compass. Gene 364: 45-52.

Furuta Y, et al. 2011. Birth and death of genes linked to chromosomal inversion. Proc Natl Acad Sci U S A 108: 1501-1506.

Hara Y, Imanishi T 2011. Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. BMC Evol Biol 11: 308.

Harris R 2007. Improved pairwise alignment of genomic DNA. [Pennsylvania State University.

Hasegawa M, Kishino H, Yano T 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160-174.

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res 21: 349-356.

Hughes JF, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature 463: 536-539.

Imanishi T, et al. 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2: e162.

Katoh K, Kuma K, Toh H, Miyata T 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518.

Kawahara Y, et al. 2009. G-compass: a web-based comparative genome browser between human and other vertebrate genomes. Bioinformatics 25: 3321-3322.

Kehrer-Sawatzki H, Cooper DN 2007. Structural divergence between the human and chimpanzee genomes. Hum Genet 120: 759-778.

Kim KJ, Lee HL 2004. Complete chloroplast genome sequences from Korean ginseng (Panax schinseng Nees) and comparative analysis of sequence evolution among 17

vascular plants. DNA Res 11: 247-261.

Kirkpatrick M 2010. How and why chromosome inversions evolve. PLoS Biol 8.

Kirkpatrick M, Barton N 2006. Chromosome inversions, local adaptation and speciation. Genetics 173: 419-434.

Kolb J, et al. 2009. Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. Chromosome Res 17: 469-483.

Lee J, Han K, Meyer TJ, Kim HS, Batzer MA 2008. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. PLoS One 3: e4047.

Li X, Heyer WD 2008. Homologous recombination in DNA repair and DNA damage tolerance. Cell Res 18: 99-113.

Meader S, Hillier LW, Locke D, Ponting CP, Lunter G 2010. Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res 20: 675-684.

Mott C, Symington LS 2011. RAD51-independent inverted-repeat recombination by a strand-annealing mechanism. DNA Repair (Amst) 10: 408-415.

Nag DK, Petes TD 1991. Seven-base-pair inverted repeats in DNA form stable hairpins in vivo in Saccharomyces cerevisiae. Genetics 129: 669-673.

Nickerson E, Nelson DL 1998. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. Genomics 50: 368-372.

Onrat ST, Soylemez Z, Elmas M 2012. 46,XX, der(15),t(Y;15)(q12;p11) karyotype in an azoospermic male. Indian J Hum Genet 18: 241-245.

Pryor JL, et al. 1997. Microdeletions in the Y chromosome of infertile men. N Engl J Med 336: 534-539.

Schwartz S, et al. 2003. Human-mouse alignments with BLASTZ. Genome Res 13: 103-107.

Sturtevant AH 1921. A case of rearrangement of genes in drosophila. Proc Natl Acad Sci U S A. 7: 235-237.

Tan CC 1935. Salivary Gland Chromosomes in the Two Races of Drosophila Pseudoobscura. Genetics 20: 392-402.

Whitlock BA, Hale AM, Groff PA 2010. Intraspecific inversions pose a challenge for the trnH-psbA plant DNA barcode. PLoS One 5: e11533.

Zuker M 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406-3415.

Chapter 3. Reconstructing the Demographic History of the Human Lineage Using Whole-Genome Alignments.

3.1. Summary

The demographic history of human would provide helpful information for identifying the evolutionary events that shaped the humanity but remains controversial even in the genomic era. In order to settle the controversies, I inferred the evolutionary history of human and great apes based on an estimation of the speciation times (*T*) and ancestral population sizes (*N*) in the lineage leading to human and great apes using the whole-genome alignments. A coalescence simulation determined the sizes of alignment blocks and intervals between them required to obtain recombination-free blocks with a high frequency. This simulation revealed that the size of the alignment block strongly affects the parameter inference, indicating that recombination is an important factor for achieving optimum parameter inference and that this simulation is helpful for the optimum data collection. From the whole genome alignments (1.9 giga-bases) of human (H), chimpanzee (C), gorilla (G), and orangutan, and the small-sized regions subject to the genomic changes by the other mechanisms than point mutations, such as CpG sites and ultramicro inversions, were excluded. 100-bp alignment blocks separated by ≥5-kb intervals were sampled from the alignments and subjected to estimate $\tau=\mu T$ and $\theta=4\mu gN$ using the MCMC method, where μ is the mutation rate and g is the generation time. Although the estimated τ_{HC} differed across chromosomes, τ_{HC} and τ_{HCG} were strongly correlated across chromosomes, indicating that variation in τ is subject to variation in μ across the lineages, rather than *T*, and thus, all chromosomes are likely to share a single speciation time. Subsequently, I estimated *T*s of the human lineage from chimpanzee, gorilla, and orangutan to be 6.0-7.6, 7.6-9.7, and 15-19 MYA, respectively, assuming variable μ across lineages and chromosomes. These speciation times were consistent

with the fossil records. I conclude that the speciation times in our recombination-free analysis would be conclusive and the speciation between human and chimpanzee was a single event.

3.2. Introduction

Reconstructing the history of human evolution is helpful for elucidating the phenotypic characteristics that may have generated the nature of modern humans. In addition to fossil records (Harrison 2010), molecular characteristics are highly informative for reconstructing the evolutionary history of humans. In the early age of molecular evolutionary studies, immunoprecipitation (antigen-antibody interactions) and electrophoresis of peptides and DNA-DNA hybridization were used for estimation of the phylogenetic relationships among human and their relatives of great apes (King and Wilson 1975; Sarich and Wilson 1967; Sibley and Ahlquist 1984). These techniques have now been replaced by the *in silico* analysis based on nucleotide sequences. By comparing genome sequences among human and great apes, we can infer the phylogenetic relationships among these species and map their molecular and phenotypic signatures onto a phylogenetic tree. Characteristics associated with the lineage leading to modern humans are candidates for key factors in human phenotypic innovation.

The demographic history among closely related species is reconstructed using sets of orthologous nucleotide sequences in different genomic regions. If the divergence of sequences is determined by species divergence time alone, the extent of nucleotide sequence divergences between species can be the same for the entire genome. However, the nucleotide divergence varies among different regions. This variation partly reflects variation of segregation times due to different coalescence among regions caused by recombination as well as the stochastic variability of nucleotide substitutions. Takahata et al. pointed out that the parameters involved in a demographic history can be estimated using a single reference genome for each species (Takahata, et al. 1995). This

is because during the course of evolution a large number of recombination events have divided genomes into large numbers of small blocks, each of which represents a single genealogy. Assuming that coalescence occurred at random in each of such blocks in an ancestral population, the coalescence times would be geometrically distributed (Kingman 1982), leading to the simultaneous estimation of the parameters involving the speciation time $\tau=\mu T$ and the ancestral population size $\theta=4\mu gN$, where μ, T, g, and N represent the mutation rate per site per year, speciation time between species, generation time in years, and the effective population size of common ancestors, respectively (Takahata, et al. 1995). This estimation is usually conducted based on a maximum-likelihood approach (Takahata and Satta 1997; Takahata, et al. 1995; Yang 2002; Yang 2010), Bayesian (Burgess and Yang 2008; Hey 2010; Hey and Nielsen 2004; Rannala and Yang 2003; Yang 2002) or Hidden Markov Model frameworks (Dutheil, et al. 2009; Hobolth, et al. 2007). Each method has both advantages and disadvantages. The maximum-likelihood and Bayesian approaches are capable of addressing three or more species. However alignment data in both approaches must be sampled so that each block represents a single genealogy for estimating τ and θ precisely. The HMM approach uses an alignment of the entire genome and scans the alignments in small windows, while the approach can treat only three species. Some parts of the variation in nucleotide divergence may be subject to introgression after initial isolation (e.g. Pinho and Hey 2010; Wu and Ting 2004). The regions in which introgressions occurred may possess distinctly smaller nucleotide divergence than the genomic average.

Using the above-mentioned theoretical frameworks, the demographic history between human and chimpanzee has been inferred based on the limited numbers of randomly sampled genomic regions or protein-coding genes since the mid-1990s (Chen

and Li 2001; Takahata and Satta 1997; Takahata, et al. 1995; Yang 1997). Due to recent rapid progress in nucleotide sequencing techniques, the whole-genome sequences of not only human but also great apes have become available (Chimpanzee Sequencing and Analysis Consortium 2005; Locke, et al. 2011; Scally, et al. 2012). Thus it becomes possible to infer the demographic history of human and great apes using massive amounts of information.

Several attempts at estimating speciation times and ancestral population sizes have been conducted using relatively long sequences (>1 Mb), or even whole-genome sequences from human and great apes (Burgess and Yang 2008; Hobolth, et al. 2007; Hobolth, et al. 2011; Patterson, et al. 2006; Satta, et al. 2004; Scally, et al. 2012; Yang 2010). However the demographic history of human remains controversial. Most of these studies supported the occurrence of a simple speciation process between human and chimpanzee (allopatric speciation), which can be explained by a unique speciation time across the genomic regions. However, a few studies have indicated the existence of multiple speciation times across these genomes, implying that human and chimpanzee experienced a complex speciation history. A study performed by Patterson and colleagues resulted in the most debatable issue on the speciation between human and chimpanzee (Patterson, et al. 2006). This study estimated a significantly more recent speciation time based on X chromosome than that on the autosomes, concluding that this observed heterogeneity would be due to recent introgression between the human and chimpanzee ancestors after initial isolation and subsequent strong selection favoring X chromosome hybrids (Patterson, et al. 2006). Yang also showed multiple speciation times, even among the autosomes, using >5-Mb genomes of human, chimpanzee, and gorilla (Yang 2010). Osada and Wu estimated different divergence times of human and

chimpanzee between coding regions and intergenic regions, suggesting that genetic exchanges had occurred during the speciation history of human and chimpanzee (Osada and Wu 2005). On the other hand, a few studies using Patterson's data did not find the complex speciation (Innan and Watanabe 2006; Yamamichi, et al. 2011). In addition, both gorilla and orangutan genome consortiums estimated speciation times using nearly whole genomes of human, chimpanzee, gorilla, and orangutan. However they did not present a conclusion about the complex history of the human lineage (Locke, et al. 2011; Scally, et al. 2012).

To determine the evolutionary history of hominids comprehensively based on whole genomes, I inferred τ and θ using whole-genome alignments consisting of the most recent assemblies of the human, chimpanzee, gorilla, and orangutan genomes. This inference was conducted using Rannala's MCMC framework (Rannala and Yang 2003). The MCMC approach requires optimal sampling of alignments, each of which ideally represents a single genealogy, to obtain precise estimations of τ and θ. Thus, I simulated the evolution of nucleotide sequences under certain demographic models to search for the optimal conditions about sizes of alignment blocks and the lengths of intervals between them. Inference of the demographic histories of hominids was conducted using the optimal conditions from this simulation. In addition, to estimate speciation times and ancestral population sizes correctly, an evolutionary model including variability of evolutionary rates across lineages was required. Variation in mutation rates has been observed between Old World monkeys and hominoid lineages, and even within the hominoids (Elango, et al. 2006; Steiper and Seiffert 2012; Steiper and Young 2006). I assumed that the probability density function of the mutation rate on a branch was subject to that of the parental (adjacent older) branch. Through this analysis, I intend to

settle the controversy about human-chimpanzee speciation described above.

3.3. Materials and Methods

Generation of Whole-Genome Alignments

The human (hg19), chimpanzee (panTro3), gorilla (gorGor3), and orangutan (ponAbe2) genome sequences were obtained from the UCSC genome browser (http://genome.ucsc.edu/). Orthologous alignments among the four species were constructed based on two procedures as described below. Orthologous pairwise alignments between the human and each great ape sequences were generated with the G-compass pipeline (Fujii, et al. 2005; Kawahara, et al. 2009) based on lastz local alignments (Harris 2007) and its unique and non-redundant reciprocal best hits. Subsequently the human genomic regions which possessed the orthologous pairwise alignments to all the three great ape were extracted and multiply re-aligned with the corresponding sequences of the three apes with MAFFT (Katoh and Toh 2008).

Both ends (20 aligned sites) of each alignment were excluded due to the ambiguity of the alignments. The alignments were split into blocks of fixed lengths of 50, 100, and 200 bp. To obtain the alignments showing unambiguous orthology, I extracted the alignment blocks satisfying $d_{HC}<0.05$, $d_{HCO}<0.08$, and the null hypothesis of $d_H=d_C$ for a relative rate test (Tajima 1993), where $d_{HCO}=(d_{HO}+d_{CO})/2$ and d_H and d_C represent the evolutionary distances from the branching point between human and chimpanzee to their leaves, respectively. In the relative rate test, I calculated the exact p-values of binominal distributions because the observed values were >5 in most cases. 97.4% alignments out of the total satisfied these conditions. The alignments that did not include ultramicro inversions (Hara and Imanishi 2011) were chosen. Gapped sites and CpG dinucleotide sites were excluded from the alignment blocks, and the alignments in

which 80% or more of sites remained were subjected to the subsequent analyses. Finally, the 100 bp alignments blocks were extracted with ≥5 kb of the intervals. The whole and sampled alignments of human and three apes were attached in Supporting data D3-1.

MCMC Inference of Demographic History

To infer the demographic history parameters τ and θ, I applied the Rannala's MCMC framework (Burgess and Yang 2008; Rannala and Yang 2003) with an extension of the evolutionary model that assumes heterogeneous evolutionary rates among lineages. Under this condition, I assumed that the mutation rate for a branch was subject to that of its parental branch; thus, the mutation rate for a branch was log-normally distributed given the mutation rate on its parent branch (Yang 2006). The mean and standard deviation of the proportion of the mutation rate of the branch to that of its ancestor were calculated from the phylogenetic trees based on the orthologous genome alignments (1.03 Gb in total) consisting of human, chimpanzee, gorilla, orangutan, and macaque sequences. The multiple alignments of the orthologous regions among the five species were generated via the same procedure as used for the four species described above. The phylogenetic trees were inferred by RAxML (Stamatakis 2006). I assumed a log-normal prior distribution of $\mu_k/\mu_{anc(k)}$ with the sequence differences in the alignment, where $\mu_{anc(k)}$ is the mutation rate of the parent branch of k. In this analysis, $\mu_{HGCO}=\mu_O$, and the relative mutation rates were $r=\mu/\mu_H$. We could compute the relative ratios of the mutation rates between sister branches but not between a parent and a daughter. This is because we do not know the divergence times of the nodes separating the daughters in advance. Therefore, I assumed a prior distribution of $\mu_k/\mu_{anc(k)}=r_k/r_{k'}$, where k' is a sister of k, instead.

In this analysis, I assumed the existence of three evolutionary conditions on the heterogeneity of the mutation rates among lineages and among genomic regions: (i) the model assuming uniform mutation rates across lineages and across genomic regions (uniform model); (ii) the model assuming variations in mutation rates across lineages; (iii) the model assuming variations across lineages and across chromosome. In the method (iii), I applied the proportion of an average of total branch length of a block on every chromosome to that of whole genome as the parameter representing the variability of mutation rates across the chromosomes (Table 3-1).

The MCMC computation was conducted by a PC cluster consisting of 128 CPUs parallelizing the calculation of the joint log-likelihood of each locus in each step using the OpenMPI library. The extension of the evolutionary model and parallelization were performed via modification of the source code of MCMCcoal1.2a developed by Yang (http://abacus.gene.ucl.ac.uk/software/MCMCcoal.html) (Burgess and Yang 2008; Rannala and Yang 2003). After 100,000 burn-in steps, τ, θ, and the relative ratio of μ were sampled every 10 steps until accumulating 50,000 samples. Median and 2.5% and 97.5% CIs were calculated for each parameter.

To calculate speciation times and ancestral population sizes, I applied the number of *de novo* mutations per generation to the mutation rate per site per year, which included 1.17×10^{-8} *de novo* mutations per site per generation from a family trio of Hap Map CEU populations, 0.97 from a trio from the Hap Map YRI populations (Conrad, et al. 2011), 1.1×10^{-8} from a family quartet of Europeans (Roach, et al. 2010), and 1.28×10^{-8} from the *de novo* mutation database of monogenic disorders (Lynch 2010). To exclude the effect of the mutations at CpG dinucleotide sites, these values were multiplied by the ratios of non-CpG mutations among the total, which were 0.86, 0.89, 0.82, and 0.86 for

the respective studies. The first three were observed values based on the literatures, and the last was the average of the first three. I set the frequency of non-CpG dinucleotides in the whole human genome at 99% (Lander, et al. 2001; Saxonov, et al. 2006). In addition I assumed the average generation time to be 20 years based on those from chimpanzee (Teleki, et al. 1976), 19.1 years, and gorilla (Walsh, et al. 2008), 22 years, though the generation time of modern humans is longer than those (Matsumura and Forster 2008). From these conditions, the mutation rates per site per year were calculated as 0.508×10^{-9}, 0.436×10^{-9}, 0.456×10^{-9}, and 0.556×10^{-9}, respectively. In this analysis the maximum and minimum values among the four were used (Table 3-3).

Simulation

We applied MaCS software (Chen, et al. 2009) to simulate the demographic history among human and the three great apes at the mega-base level. I generated the demographic history of 10 Mb regions of the four species, setting $\mu=1\times10^{-9}$ per site per year, the recombination rate at 10 cM/Mb for hotspots and 1cM/Mb for the other regions, the average generation time at 20 years, T_{HC} at 300,000 generations, T_{HCG} at 400,000 generations, T_{HCGO} at 700,000 generations, and the population sizes at N_H=27,500, N_C=50,000, N_G=30,000, N_O=33,000, and N_{HC}= N_{HCG}=N_{HCGO}=60,000. Hotspots were distributed among 10%, 5%, and none of the regions at random, respectively. If adjacent regions were separated by recombinations but showed equal coalescence times, they are merged into a single genealogy. In the simulated region, blocks of a fixed length were set together with a fixed interval. The start site of the first block was randomly chosen within a length of the fixed interval from the end of the region. The blocks were subjected to examination of how many genealogies were

included in a block and how blocks shared a genealogy with adjacent ones. Based on the demographic history estimated with MaCS software (Chen, et al. 2009), random nucleotide sequences were evolved using Seq-Gen software (Rambaut and Grassly 1997). Alignments in blocks with fixed sizes (50 bp - 1 kb) and fixed intervals (500 bp - 5 kb) were extracted and subjected to estimation of τ and θ using Rannala's MCMC framework (Rannala and Yang 2003), assuming the uniform model (model (i) described above). The genealogy and sequence data generated by MaCS and Seq-gen were attached in Supporting data D3-2.

3.4. Results

Simulation of Coalescence

We simulated nucleotide sequences with MaCS software (Chen, et al. 2009) to obtain the optimal condition about the size of the alignment blocks and the length of the intervals between them. This procedure is intended to obtain recombination-free alignments with a high frequency. MaCS is much faster than the other available demographic simulation software and, thus, suitable for studies using mega-base pair or longer sequences (Chen, et al. 2009).

I assumed a 10-Mb region, in which recombination and coalescent events were generated according to a so far common feature of demographic history of human, chimpanzee, gorilla, and orangutan. The model included the following parameters: speciation times of humans from chimpanzee (T_{HC}), gorilla (T_{HCG}) and orangutan (T_{HCGO}) of 6, 8, and 14 million years ago (MYA), respectively; effective population sizes of the human (N_H), chimpanzee (N_C), gorilla (N_G), orangutan (N_O) lineages and the ancestral lineages (N_{HC}, N_{HCG}, and N_{HCGO}) of 27,500, 50,000, 30,000, 33,000, and 60,000, respectively. I considered a combination of two kinds of the recombination rates in a region. One represents an average recombination of 1 cM/Mb, equivalent to the average recombination rate across the human genome (Bouffard, et al. 1997; Nagaraja, et al. 1997; Pritchard and Przeworski 2001). The other represents a recombination rate of hotspots (Myers, et al. 2005), 10 cM/Mb. Ninety percents of a given region exhibited the former rate, while the remaining regions possessed the latter rate. The hotspots were randomly allocated across the region.

After constructing pieces of the genealogies in the 10-Mb region according to the

procedure described above, I allocated blocks to the region with a fixed size ranging from 50 bp to 1 kb and intervals with a fixed length ranging 500 bp to 5 kb. Under each combination of a block size and an interval length, I examined the number of genealogies in every alignment block and the number of alignment blocks sharing a genealogy with an adjacent block. The result showed that blocks with small sizes frequently present a single genealogy (Fig. 3-1**A**). While 87% of the 50-bp blocks with 5-kb intervals showed a single genealogy, only 17% of the 1-kb blocks with 5-kb intervals did. Interestingly, these values are more or less the same in different proportion of the two recombination rates in a region (Fig. 3-1**A**). In addition, I found that the longer the intervals between the blocks, the less frequent the blocks share a genealogy (Fig. 3-1**B**). 28% and 15% of blocks separated by 500-bp and 1-kb intervals shared a genealogy with the adjacent one, respectively. This was only true for 0.76% and 0.036% of blocks with 3-kb and 5-kb intervals, respectively.

Once genealogies were determined, the sequences of four species were simulated. If alignment blocks are set to be short with longer intervals, the blocks would be frequently allocated to a single different genealogy, leading to the precise estimation of speciation times and population sizes. I examined the impact of block sizes and the interval lengths on the accuracy of the estimation of τ and θ using the 10-Mb sequence alignments of the four species. After 20 replications of this procedure, I found that if alignment blocks were set to be short with longer intervals, speciation times and population sizes were estimated precisely (Fig. 3-1**C**-**H**). In most of the estimations of τ and θ with 50 and 100 bp blocks, the true values were included in the interquartile ranges, whereas in most of the τ and θ estimates based on 500 and 1 kb blocks, the true values were outside of the 95th percentiles. The variances in τ and θ were large in the

simulation of the short alignment blocks due to the low numbers of alignment sites. However, the variances in a real genome dataset, which would be nearly 200 times the size of the simulation dataset, would be negligibly small even if I use such short alignment blocks. These results indicate that blocks of 100 bp or less are preferable for estimations. It is noteworthy that the variances of θs are larger than those of τs under all conditions. I also found that the interval length between blocks was moderately influential in the estimation compared to the size of the blocks (irrespective of the proportions of the two recombination rates). The estimated τ and θ with more than 1-kb intervals appeared to be equivalent to the true values, whereas the estimates with 500-bp intervals can be inconsistent with the true values: the τ_{HCGO} and θ_{HCGO} differed from the true values (Fig. 3-1**C**-**H**).

Inference of τ and θ Using the Human and Great Apes Genomes

We inferred the τ and θ associated with the hominid demographic history using the human and three great apes genomes. I generated a total of 1.9 Gb of orthologous alignments using the human, chimpanzee, gorilla, and orangutan genomes. I inferred τ and θ via Rannala's MCMC framework (Rannala and Yang 2003), with modification of the heterogeneity of the mutation rates across the lineages (see Methods).

I used 50-bp alignment blocks together with 5-kb intervals to infer τ and θ but failed to compute realistic values: θ_{HC} was completely different from the values found in previous studies, and τ_{HC} was different between the four-species analysis and human-chimpanzee-orangutan analysis (Table 3-2). This may be because of failure of convergence in the Markov chain analysis (see Discussion). Therefore, I chose to use 100-bp blocks and 5-kb intervals instead.

To examine whether the estimated speciation times were unique between the autosomes and X chromosome, I inferred τ based on alignments of the autosomes and X chromosome separately (Table 3-1 and Supporting data D3-3). I estimated different τ_{HC} values between the autosomes and X chromosome: 0.00326 (95% CI: 0.00321 to 0.00331) for autosomes and 0.00277 (95% CI: 0.00263 to 0.00290) for the X chromosome. I also observed a similar difference in τ_{HC} between the autosomes and X chromosome when using the simplest model, which considers a uniform evolutionary rate across the lineage and across loci (Table 3-1). Furthermore, the estimated τ_{HC} for each chromosome varies, even across the autosomes (Table 3-1). The observed variability of τ across the autosomes appears to be consistent with Yang's estimation that was based on 5.2 Mb of alignments among human, chimpanzee, and gorilla genomes (Yang 2010).

Because τ is the product of the mutation rate, μ, and speciation time, T ($\tau=\mu T$), variation in τ across chromosomes can be explained by variability in T and/or μ. To determine which parameters affected the variation of τ, I plotted two τ values that reflect different species divergence time. The result showed that τ_{HC} and τ_{HCG} were strongly positively correlated across the chromosomes (R^2=0.822, p=2.58×10^{-9}) (Fig. 3-2**A**). To examine if this relationship was merely due to the fact that τ_{HC} and τ_{HCG} were simultaneously estimated from the same sequence data, I sampled alignment blocks that were not included in the original sample data and, using this new sample data (sample 2), estimated τ and θ (Supporting data D3-3). Interestingly, it was found that τ_{HC} and τ_{HCG} even from different sample data were strongly positively correlated (R^2=0.740, p=1.42×10^{-7} for Fig. 3-2**B** and R^2=0.774, p=3.24×10^{-8} for Fig. 3-2**C**). These results strongly suggested that the relationship between τ_{HC} and τ_{HCG} could not be explained by

the correlation of the data itself. This observation can be explained in two different ways. First, if τ_{HC} and τ_{HCG} are correlated, μ would vary across the chromosomes but T_{HC} and T_{HCG} would be constant. I further found that θ_{HC} and θ_{HCG} were significantly positively correlated across the chromosomes ($R^2 \geq 0.320$, $p \leq 0.00494$, Fig. 3-2**D-F**) though the correlation coefficients between θ_{HC} and θ_{HCG} were lower than those between τ_{HC} and τ_{HCG}. More complicated assumptions were likely to be required for explaining that the variability of the speciation time in each chromosome could be shared between the two temporally-independent speciation events. This finding also supports the idea that μ varied across the chromosomes. The second explanation is as follows: even under the constant μ among chromosomes, the same bias of T_{HC} and T_{HCG} in each chromosome, if any, could explain the variation of τ values. The latter explanation is rather unlikely. Thus, it is plausible that μ varied across the chromosomes, and that the speciation times of T_{HC} and T_{HCG} were unique across the chromosomes.

In addition to τ and θ, variation of mutation rates across lineages was simultaneously estimated by calculating the proportion of the mutation rate in a lineage to that in the human lineage (μ_H) through the MCMC procedure (Table 3-3 and Supporting data D3-3). Though both autosomes and X chromosome showed the slowdown during the course of the human evolution, the degree of the slowdown in X chromosome was higher than that in autosomes.

Estimation of Speciation Times and Ancestral Population Sizes

Assuming that μ varies across lineages and across chromosomes based on the result described above, I estimated τ and θ based on the blocks sampled from the whole genome alignments of human and three great apes. I collected 100 bp alignment blocks

separated by ≥5 kb intervals. Similar to the studies using whole or nearly whole-genome sequences (Hobolth, et al. 2011; Scally, et al. 2012; Yang 2010), the τ values estimated in my analysis were smaller than those found in previous studies involving smaller samples (Table 3-1) (Osada and Wu 2005; Takahata, et al. 1995; Yang 2002).

I estimated speciation times using the mutation rate from a recent estimation based on the number of *de novo* mutations per generation (Conrad, et al. 2011; Lynch 2010; Roach, et al. 2010), which was nearly half of the mutation rate previously estimated (Nachman and Crowell 2000). According to Nachman and Crowell (2000), the rate was calculated based on the $d=2\mu T+4\mu gN$, where d represents the nucleotide difference between the species at a local region. For the estimation of μ, they assumed T=5 MYA and N=10,000, both of which were smaller than those widely thought recently (Roach et al. 2010). This would lead to the estimation of large value of μ (Roach et al. 2010). Based on *de novo* mutations, I set the mutation rate, excluding CpG sites, in the human lineage (μ_H) to be from 0.44×10^{-9} to 0.56×10^{-9} per site per year, assuming the average generation time at 20 years (see Methods). Taking into account the variability of the mutation rates across the lineages (Table 3-3), the value of T_{HC} was calculated at 5.9-7.6 MYA, T_{HCG} at 7.6-9.7 MYA, and T_{HCGO} at 15-19 MYA (Table 3-4). It should be noted that these estimated speciation times were consistent with those from the fossil records (Carroll 2003, see Discussion).

Based on the θ values, I also estimated the effective population sizes in the ancestral lineages. The population sizes of N_{HC}, N_{HCG}, and N_{HCGO} were estimated to be 59,300-75,600, 51,400-66,000, and 159,000-203,000, respectively. In addition, the ancestral population sizes of the X chromosome were estimated to be 34,300-43,800 for $N_{HC(X)}$, 38,500-49,200 for $N_{HCG(X)}$, and 141,000-179,800 for $N_{HCGO(X)}$ (Table 3-4). Considering

the CI, all of the estimates of the ancestral population sizes for the X chromosome are roughly three-fourths of those for autosomes, as expected.

3.5. Discussion

To estimate accurate speciation times and ancestral population sizes using the MCMC framework, I conducted a systematic simulation about finding optimum method of sampling genomic regions that have single, independent genealogy. According to the simulation results, I inferred the demographic history of human and great apes based on whole-genome orthologous alignments of the human, chimpanzee, gorilla, and orangutan sequences. Finally, I obtained conclusive speciation times among human and great apes at the whole genome level and revealed that human-chimpanzee speciation was a single event.

In this study, I showed the importance of using recombination-free alignments to estimate precise τ and θ. Burgess and Yang stated that the length of loci has little influence on the estimation of τ and θ for alignments of hominids and Old World monkeys (Burgess and Yang 2008). However, my simulation targeting hominid alignments indicated that using alignment block of $\leq$100 bp was favorable for the estimation, but that block of $\geq$500 bp may result in poor estimations (Fig. 3-1). This discrepancy is clearly caused by the fact that the alignment blocks with recombination events were frequently observed in the $\geq$500 bp dataset. The reason for Burgess and Yang's observation is that they did not compare their estimated parameters to true values because the true values from real genome data are unknown (Burgess and Yang 2008). The greater the number of recombination events in alignment block is, the narrower the distribution of the evolutionary distance (*d*) of loci will centralize around the average. This leads to large τ and small θ estimates. Therefore, the appropriate size of loci should be used to infer τ and θ precisely. On the other hand, lengths of the

intervals between blocks ranging from 500 bp to 5 kb had moderate impact on the estimation of τ and θ (Fig. 3-1). Thus, it may be reasonable to choose a shorter interval between blocks and to collect a large number of blocks to reduce the variances of the parameters when the sequence information from genomes is limited. It is noteworthy that the appropriate size of alignment blocks and intervals does not rely on the proportion of the two different recombination rates: recombination hotspots and the other regions. In summary, my simulation is useful for obtaining the optimal conditions for the sampling of genome alignments. It should be, however, noted that making inferences under preferable conditions in a simulation is not always practical for obtaining inferences using actual data. I failed to estimate τ and θ when using 50-bp alignment blocks and, instead, estimated them with 100 bp alignment blocks. One of the reasons of this inexpedience can be the broadness of the likelihood surface contour with small block size. The simple two-species maximum-likelihood analysis based on the simulation data indicated that if the alignment blocks are small, the likelihood surface contour plot becomes broad (Hara, et al. 2012). In such a case, furthermore, the innermost area of the plot can become separated into two or more parts , which can lead to the convergence in the suboptimal condition. It is noted that τ and θ with the 100 bp blocks were comparable to those with the 200 bp blocks (Table 3-2), suggesting that the inference with the 100 bp blocks were not in the convergence in the suboptimal condition which might occur with 100 bp in the simulation dataset (Hara, et al. 2012).

I found that the variability of τ among the chromosomes can be explained solely by mutation rates, thus indicating a single speciation time between human and chimpanzee. This finding is inconsistent with Patterson's conclusion (Patterson, et al. 2006), but consistent with the results of follow-up analyses performed using Patterson *et al*.'s data

(Innan and Watanabe 2006; Yamamichi, et al. 2011). It is well known that the difference in mutation rates between autosomes and the X chromosome is due to the difference in the duration time in males, where most point mutations are generated in mammals. In contrast, the cause of the variation in mutation rates across autosomes remains unclear, though such variation is clearly observed between the human and chimpanzee genomes (Hodgkinson and Eyre-Walker 2011). I did not find statistically significant correlations between mutation rates and genomic characteristics such as GC contents, CpG proportions, chromosomal sizes, SNP densities, or recombination densities in the large genomic regions constituted by autosomes, implying that other mechanisms underlie the causes of chromosome specific mutation rates. On the other hand, the comparative genomics across the chromosomes in rodent genomes has revealed that large-scale genomic characteristics such as the degree of chromosomal rearrangements and replication time correlate to the variation in mutation rates across the chromosomes (Pink and Hurst 2010; Pink, et al. 2009). Thus, to clarify the causes of chromosome specific mutation rates in the human lineage, it would be required to examine the relationships between the mutation rates and the structural characteristics of chromosomes at large scale rather than the sequences themselves.

I then evaluated the possibility of complex speciation using a specific region such as coding sequences. Osada and Wu indicated that the coding regions in human-chimpanzee ancestors had experienced multiple genetic changes during the speciation history of these species (Osada and Wu 2005). This result should be carefully interpreted, because these authors used a full-length cDNA as a single locus. A full-length cDNA can be mapped in segments by exons in the genome, and thus, a full-length cDNA can have more than one genealogy. Therefore, I attempted to perform

speciation time estimations specifically using coding regions by selecting 100 bp blocks with intervals of ≥5 kb. In this analysis I used the well-annotated coding regions that were characterized as H-InvDB category I (Imanishi, et al. 2004). Using these alignments, I obtained τ_{HCG} and τ_{HC} values of 0.00249 and 0.00156, respectively (Table 3-1). Although both of these values are lower than those from the whole-genome analyses, these values are quite close to the regression line between τ_{HCG} and τ_{HC} (Fig. 3-2), suggesting that speciation time based on the coding regions is equivalent to that based on the whole genome analysis. From these 100-bp blocks of coding regions, I extracted four-fold degenerate (FFD) sites at the third codon positions, which are likely under neutral evolution. The values of τ obtained from these data were also plotted very close to that regression line, consistent with the whole coding sequence analysis. Although the estimation is rough, $N_{HC(FFD)}$ was only 1.1 times larger than the $N_{HC(WG)}$, where $N_{HC(FFD)}$ and $N_{HC(WG)}$ were the N_{HC} of FFD sites at the third codon positions and the whole genomes, respectively. Thus it is suggested that $N_{HC(CDS)}$ was overestimated due to non-neutral evolution in coding regions, where $N_{HC(CDS)}$ is N_{HC} of coding regions. It should be noted that my results did not completely reject the possibility of complex speciation processes between human and chimpanzee. It may be possible that introgressions occurred soon after the major speciation event, which may not be distinguishable from the stochastic variation of the distribution of coalescence times in the ancestral population.

The τ and θ values estimated in my analysis are slightly different from those reported by the gorilla genome consortium (Scally, et al. 2012) based on the most recent whole-genome analysis under the HMM framework. The main reason for this difference is the alignments used in the two studies. I chose unambiguously aligned regions, removing

the ends of the alignments, and excluded CpG sites from the alignments. In the human and chimpanzee genomes, a cytosine at a CpG dinucleotide site can mutate to thymine 15 times as frequently as that at other sites due to oxidative deamination of methylated cytosines at CpG sites (Elango, et al. 2008). Therefore, the CpG site removal was to approximate the mode of nucleotide substitutions in the alignment to the simple Jukes-Cantor model (Jukes and Cantor 1969), which both I and Rannala and Yang applied. Exclusion of CpG sites from aligned sites also decreases the evolutionary distances (*d*), which correspond to $\theta+2\tau$.

If the variation in τ and θ between my analysis and Scally *et al*. are explained based on the difference in the mutation rates excluding or including CpG sites, the θs and τs ratios would be constant between these two analyses. However, the τ_{HCG}/τ_{HC} ratio was different between the two analyses: $\tau_{HCG}/\tau_{HC}=1.3$ in my analysis and $\tau_{HCG}/\tau_{HC}=1.6$ in Scally *et al*. Thus the differences in τ can be explained by the other factors than CpG sites. One of such factors may be the correction of the heterogeneity of mutation rates across lineages. The analysis by Scally *et al.* corrected the variation in mutation rates after inference of τ and θ, simply by multiplying the proportion of μ to that of the human lineage (Scally, et al. 2012). In contrast, I inferred the coalescence time of each locus using a model with variable mutation rates across the lineages. In summary, CpG-sites were excluded for fitting the sequence data to the evolutionary model that I used, heterogeneity of mutation rates among lineages was assumed in each locus for considering the variation in mutation rates in hominids, and four species were applied for increasing the inner-node speciation. Thus I looked carefully at the condition for inferring the demographic history of human and the great apes precisely. However, these conditions do not seem to properly be dealt with in the analysis by Scally *et al*.

The speciation times observed in my analysis are also supported by fossil records of the Homininae (Fig. 3-3). *Nakalipithecus* and *Chororapithecus*, which lived 9.8-9.9 and 10-10.5 MYA, respectively, are morphologically related to extant hominines and suggest that the origin of hominines was in Africa (Kunimatsu, et al. 2007; Suwa, et al. 2007). The estimated speciation time T_{HCG} (7.6-9.7 MYA) suggests that *Nakalipithecus* and *Chororapithecus* may be related to the stem of the Homininae. The speciation time between human and chimpanzee was estimated to be 5.9-7.6 MYA in my analysis, which is consistent with both the most recent findings obtained using a genomic approach (Scally, et al. 2012) and the traditional view from fossil records (Carroll 2003). *Orrorin* (around 6 MYA) (Wood 2010) and *Sahelanthropus* (around 7 MYA) (Brunet, et al. 2005) both lived around the time of human-chimpanzee speciation. In contrast, *Ardipithecus* was considered to emerge after this speciation, with *Ar. kadabba* found 5.2-5.8 MYA (Haile-Selassie 2001) and *Ar. ramidus* 4.4 MYA (White, et al. 1994).

Based on my analysis, I propose a new approach for data collection from whole-genome alignments to infer demographic parameters. Simulation of coalescence is helpful for determining the appropriate size of alignment blocks and the interval length between the blocks. This approach can be applied for closely related species in various lineages. Although the HMM framework can cover entire genomic regions by scanning alignments with small windows, the MCMC framework developed by Rannala and Yang uses a fraction of whole-genome alignments to avoid the effect of recombination. At this time, however, only the MCMC methods are capable of addressing three or more species simultaneously. When the genomic sequences of several closely related species are available, increasing the number of estimated points (speciation times and ancestral population sizes) would lead to more accurate estimations. my method for inference of τ

and θ, assuming heterogeneity of mutation rates across lineages, may also be preferable when addressing multiple species with different mutation rates. Although this approach assuming heterogeneity of mutation rates across both lineages and blocks is still under development for practical use, it could contribute to reconstructing more precise demographic histories of related species, including hominids.

Fig. 3-1. Results in the simulations.

(**A**) Number of genealogies in a block under each of the block size conditions, setting the interval between the blocks at 5 kb. The frequency of hot spots was considered to cover 10% (upper graph), 5% (middle), and none (lower) of the genomes (see text). (**B**) Number of blocks sharing a genealogy with an adjacent block under each of the interval length conditions, setting the block size at 100 bp. These values are the average of the 1,000 replications of the coalescence simulation. (**C-H**) The estimated θs and τs from simulated sequences. Each boxplot consists of the averages of the 2.5^{th} percentile, lower quartile, median, upper quartile, and 97.5^{th} percentile from 20 replications, from bottom to top. A mark of X represents the median of each of the 20 replications. Dotted lines represent the true values. Under each condition, asterisks indicate that the true value is outside of the 95th percentile, and daggers indicate that the true value is smaller than or larger than all of the medians in the 20 replications.

A

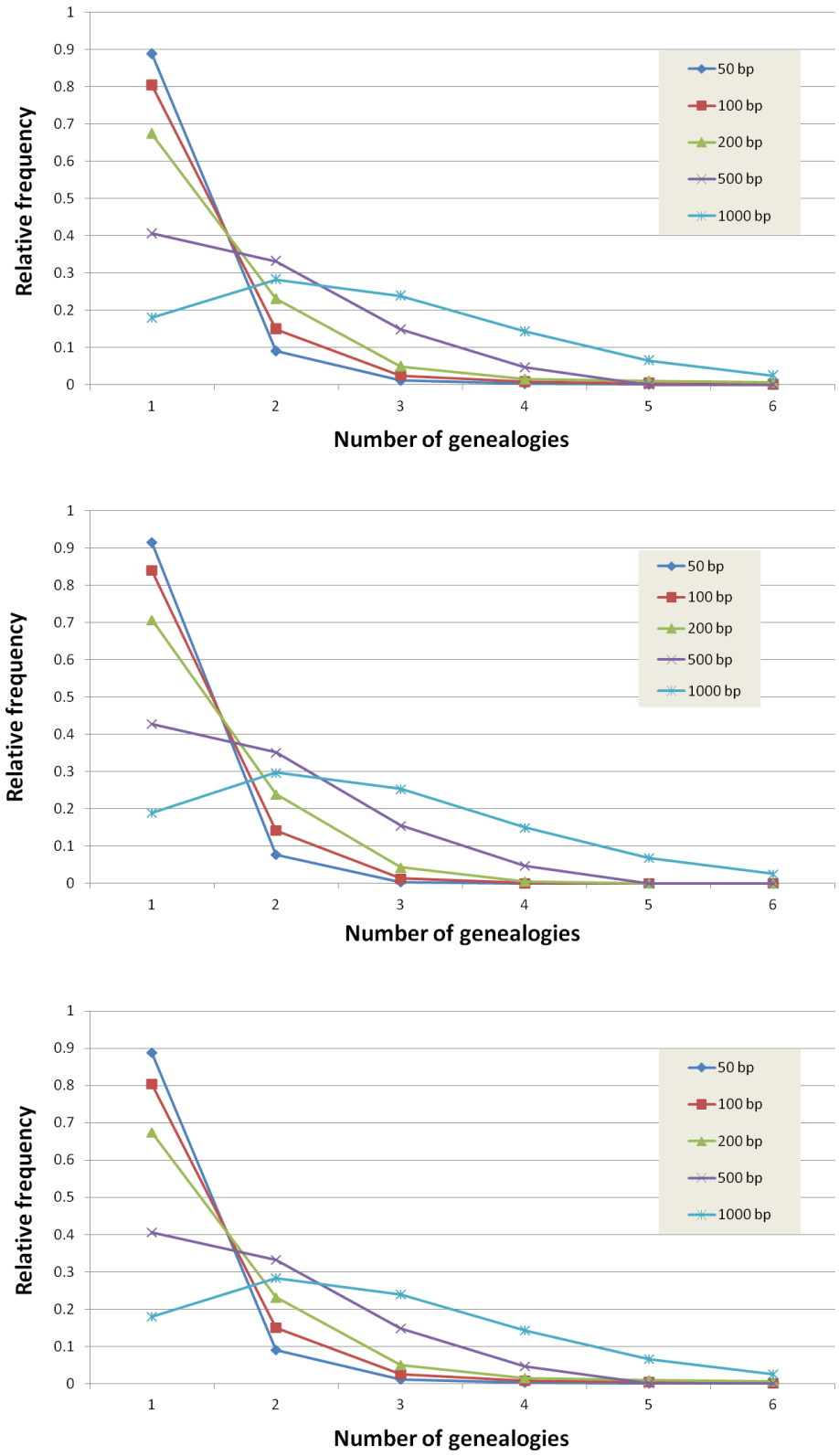

Relative frequency
Number of genealogies
50 bp
100 bp
200 bp
500 bp
1000 bp
Relative frequency
Number of genealogies
50 bp
100 bp
200 bp
500 bp
1000 bp
Relative frequency
Number of genealogies
50 bp
100 bp
200 bp
500 bp
1000 bp

Fig. 3-1

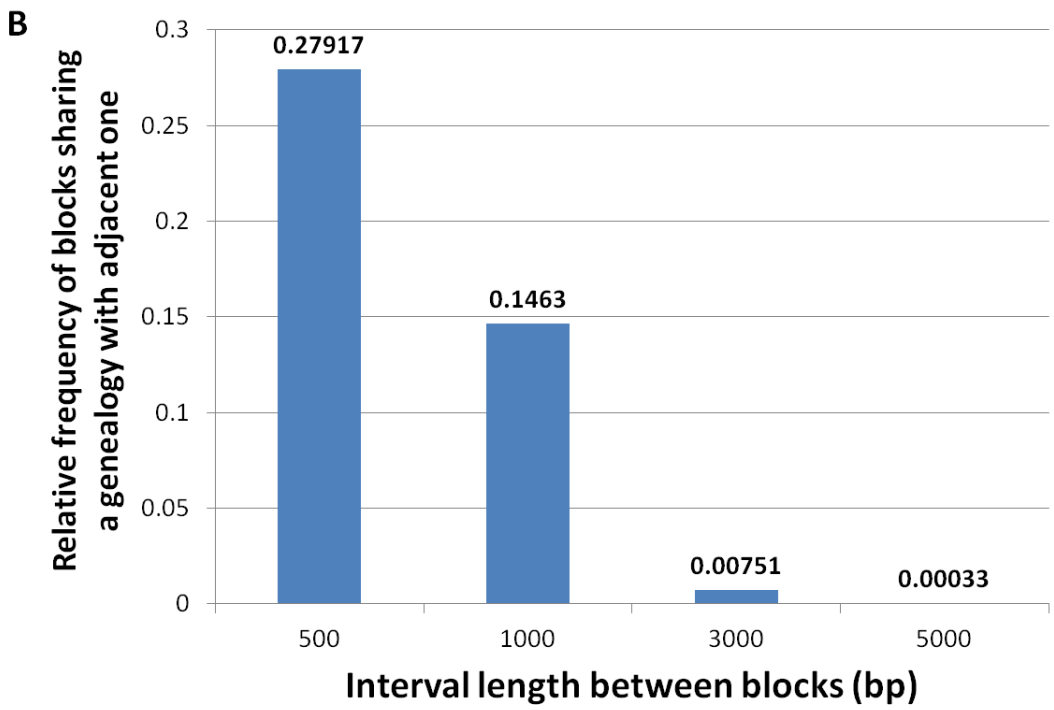

B
Relative frequency of blocks sharing a genealogy with adjacent one
0.3
0.25
0.2
0.15
0.1
0.05
0
0.27917
0.1463
0.00751
0.00033
500
1000
3000
5000
Interval length between blocks (bp)

Fig. 3-1 (cont.)

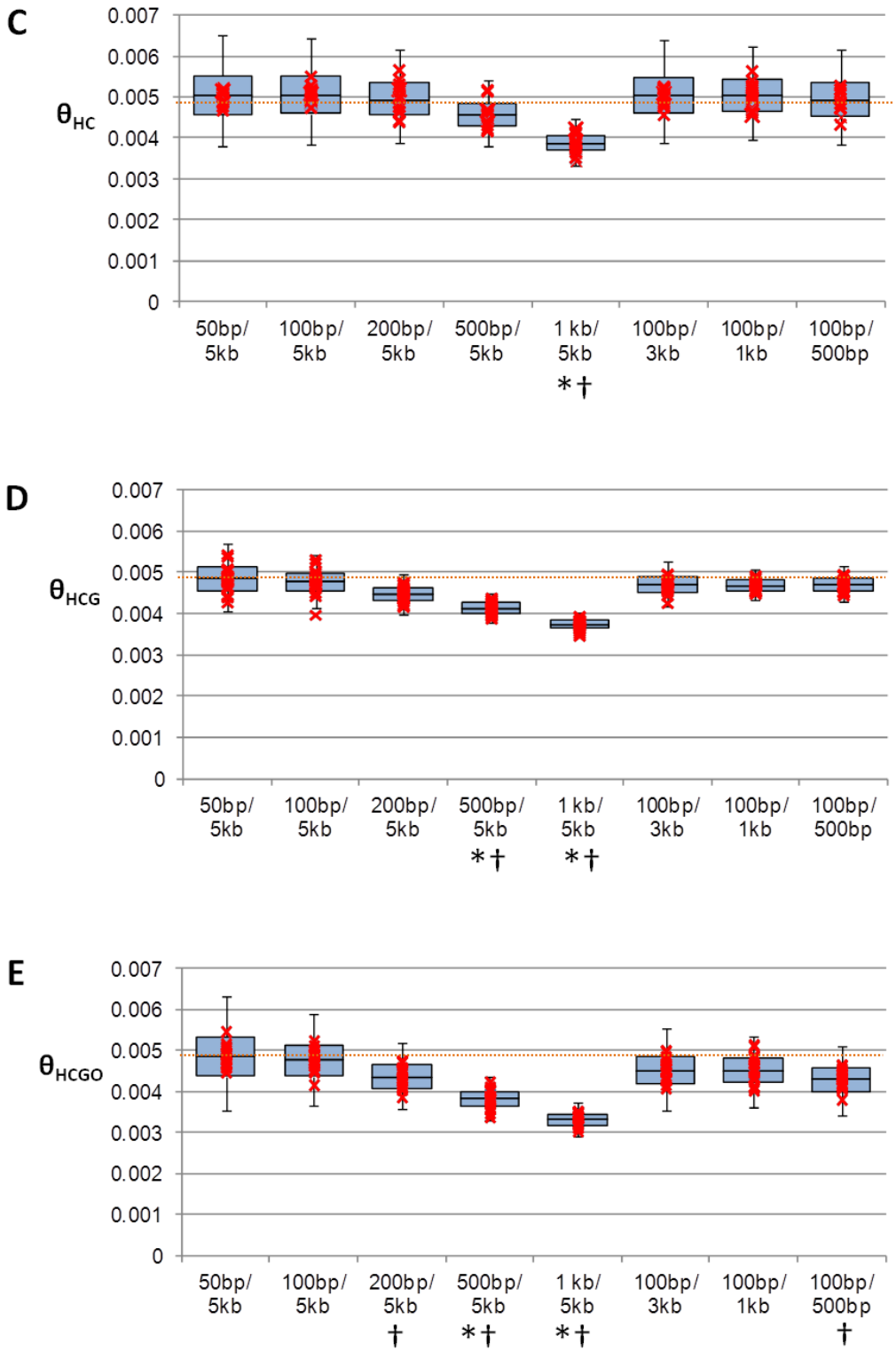

C
θHC
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
50bp/ 5kb
100bp/ 5kb
200bp/ 5kb
500bp/ 5kb
1 kb/ 5kb
100bp/ 3kb
100bp/ 1kb
100bp/ 500bp
*†
D
θHCG
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
50bp/ 5kb
100bp/ 5kb
200bp/ 5kb
500bp/ 5kb
1 kb/ 5kb
100bp/ 3kb
100bp/ 1kb
100bp/ 500bp
*†
*†
E
θHCGO
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
50bp/ 5kb
100bp/ 5kb
200bp/ 5kb
500bp/ 5kb
1 kb/ 5kb
100bp/ 3kb
100bp/ 1kb
100bp/ 500bp
†
*†
*†
†

Fig. 3-1 (cont.)

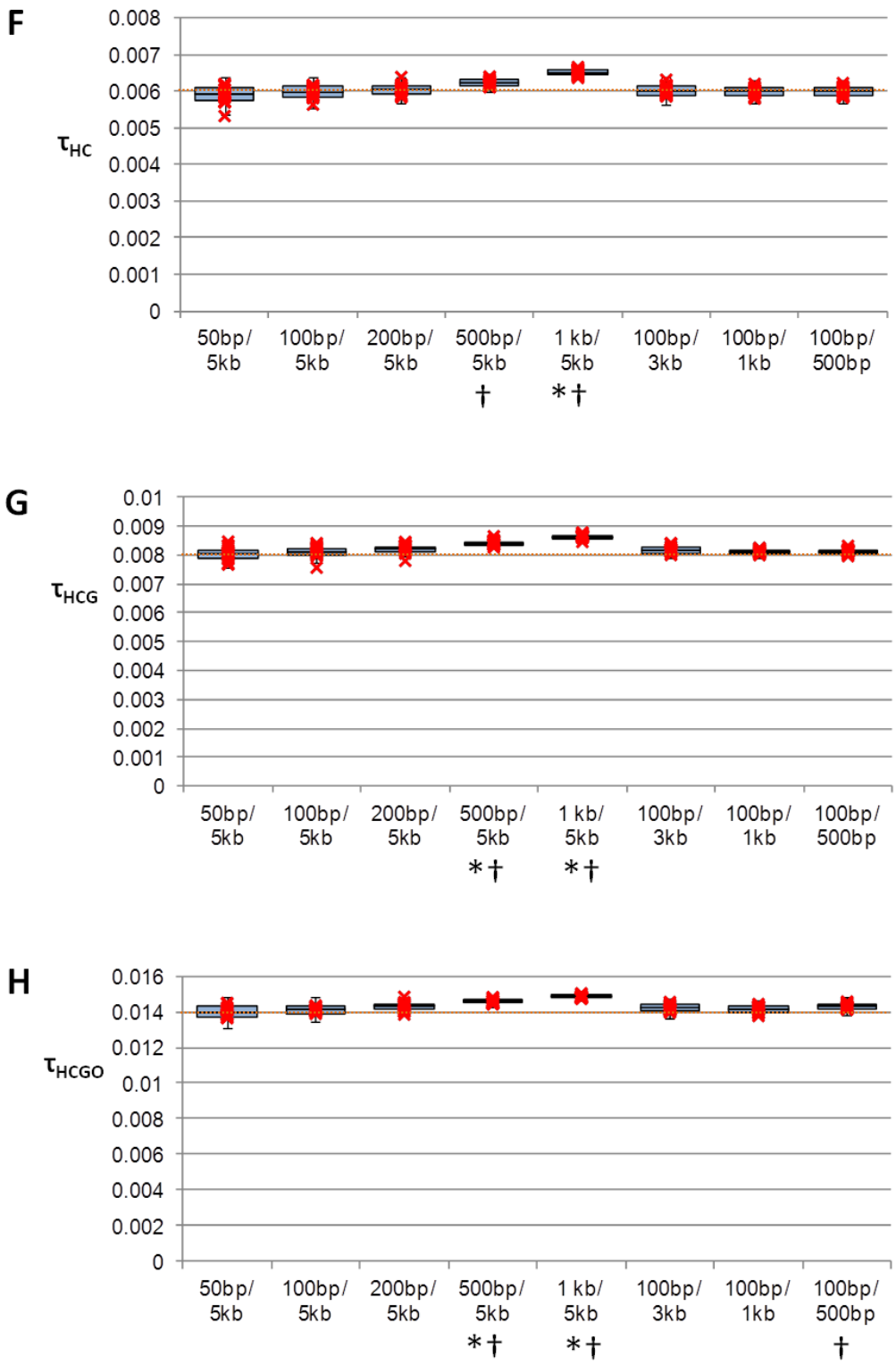

F
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
τHC
50bp/ 5kb
100bp/ 5kb
200bp/ 5kb
500bp/ 5kb
1 kb/ 5kb
100bp/ 3kb
100bp/ 1kb
100bp/ 500bp
†
*†
G
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
τHCG
50bp/ 5kb
100bp/ 5kb
200bp/ 5kb
500bp/ 5kb
1 kb/ 5kb
100bp/ 3kb
100bp/ 1kb
100bp/ 500bp
*†
*†
H
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
τHCGO
50bp/ 5kb
100bp/ 5kb
200bp/ 5kb
500bp/ 5kb
1 kb/ 5kb
100bp/ 3kb
100bp/ 1kb
100bp/ 500bp
*†
*†
†

Fig. 3-1 (cont.)

Fig. 3-2. Relationships between τ_{HC} and τ_{HCG} and between θ_{HC} and θ_{HCG} for each chromosome.

(**A-C**) Plots and a regression line between τ_{HC} and τ_{HCG} for each chromosome: τ_{HC} and τ_{HCG} from the original sample (**A**), τ_{HC} from the original sample and τ_{HCG} from the sample 2 (**B**), and τ_{HC} from the sample 2 and τ_{HCG} from the original sample (**C**). Diamonds represent autosomes, and an a cross, a triangle, or a square represents an X chromosome, coding region, or four-fold degenerate sites at the third codon positions, respectively. The regression line was calculated for autosomes and X chromosome and shown with its formula and the square of its correlation coefficient. **(D-F)** Plots and a regression line between θ_{HC} and θ_{HCG} for each chromosome: θ_{HC} and θ_{HCG} from the original sample (**D**), θ_{HC} from the original sample and θ_{HCG} from the sample 2 (**E**), and θ_{HC} from the sample 2 and θ_{HCG} from the original sample (**F**). *p*-values for correlation coefficients were calculated at 0.000792 for **D**, 0.00144 for **E**, and 0.00494 for **F**. The correlation coefficients between θ_{HC} and θ_{HCG} were lower than those between τ_{HC} and τ_{HCG}. This observation can be explained by the two inclusive causes: one is the large variation in θ estimates through the MCMC inference (Fig. 3-1), and the other is variation in *N* under non-neutral evolution in some parts of the genomes.

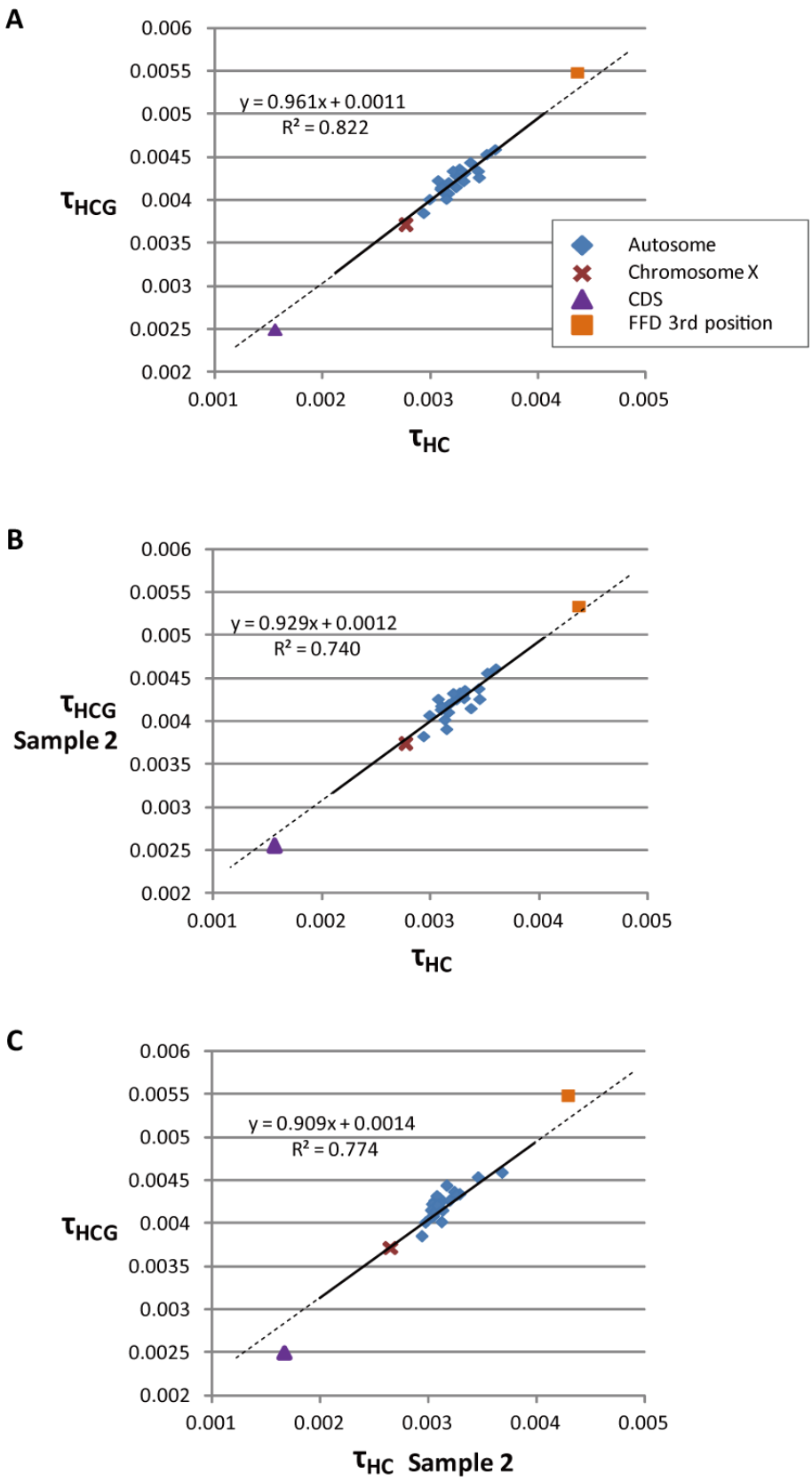

A
0.006
0.0055
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
τHCG
y = 0.961x + 0.0011
R² = 0.822
Autosome
Chromosome X
CDS
FFD 3rd position
0.001
0.002
0.003
0.004
0.005
τHC
B
0.006
0.0055
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
τHCG
Sample 2
y = 0.929x + 0.0012
R² = 0.740
0.001
0.002
0.003
0.004
0.005
τHC
C
0.006
0.0055
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
τHCG
y = 0.909x + 0.0014
R² = 0.774
0.001
0.002
0.003
0.004
0.005
τHC Sample 2

Fig. 3-2

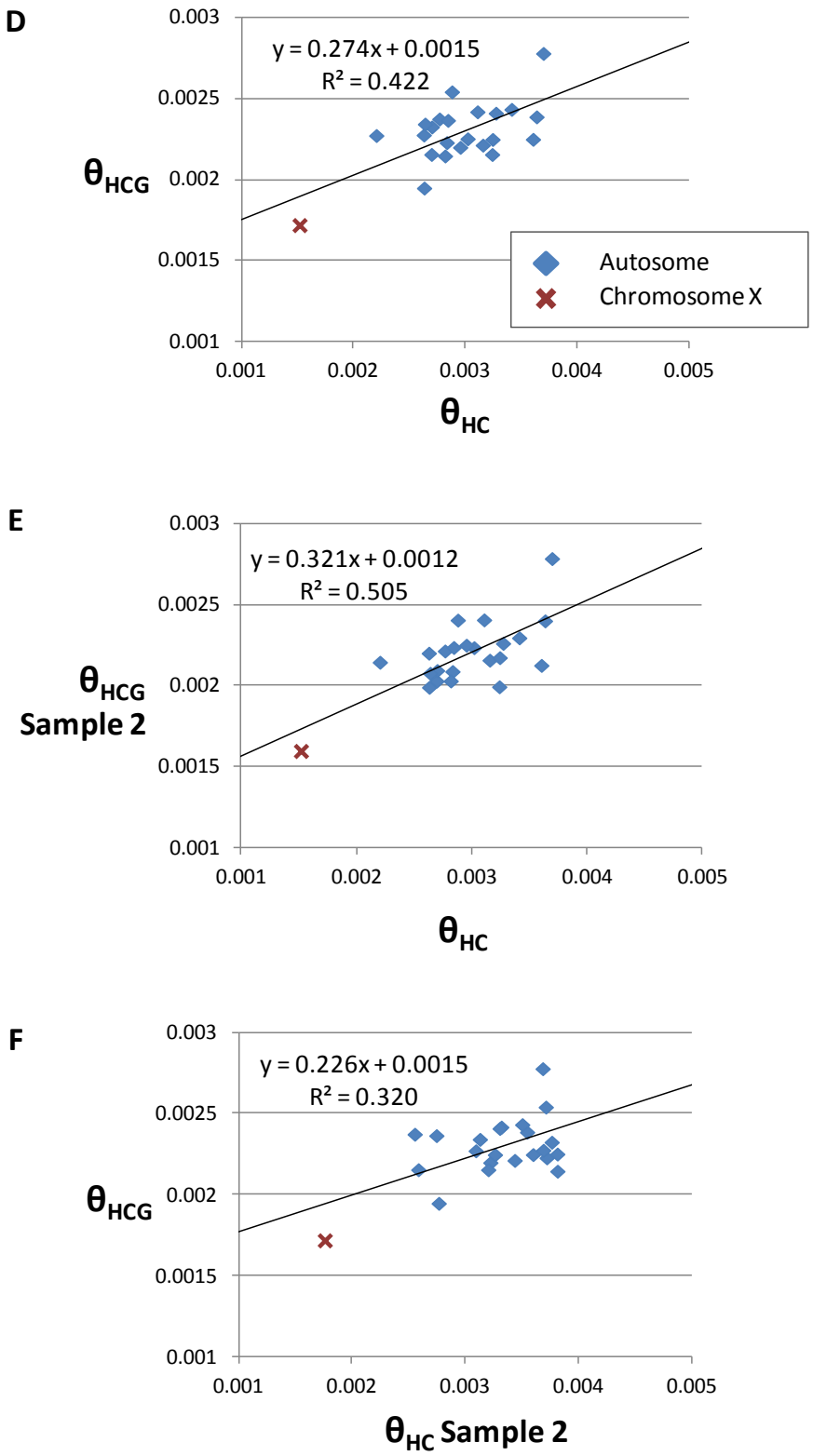

D
y = 0.274x + 0.0015
R² = 0.422
θHCG
θHC
Autosome
Chromosome X
E
y = 0.321x + 0.0012
R² = 0.505
θHCG Sample 2
θHC
F
y = 0.226x + 0.0015
R² = 0.320
θHCG
θHC Sample 2

Fig. 3-2 (cont.)

Fig. 3-3. Relationship between the estimated speciation times and the fossil records. The estimated inhabitation periods of fossil species were obtained from literatures (Brunet, et al. 2005; Gabunia, et al. 2001; Haile-Selassie 2001; Ishida, et al. 1999; Kunimatsu, et al. 2007; Sawada, et al. 1998; Suwa, et al. 2007; Wood 2010), for details see Discussion. Dotted lines represent the upper and lower bonds of the estimates of the speciation times (Table 3-4).

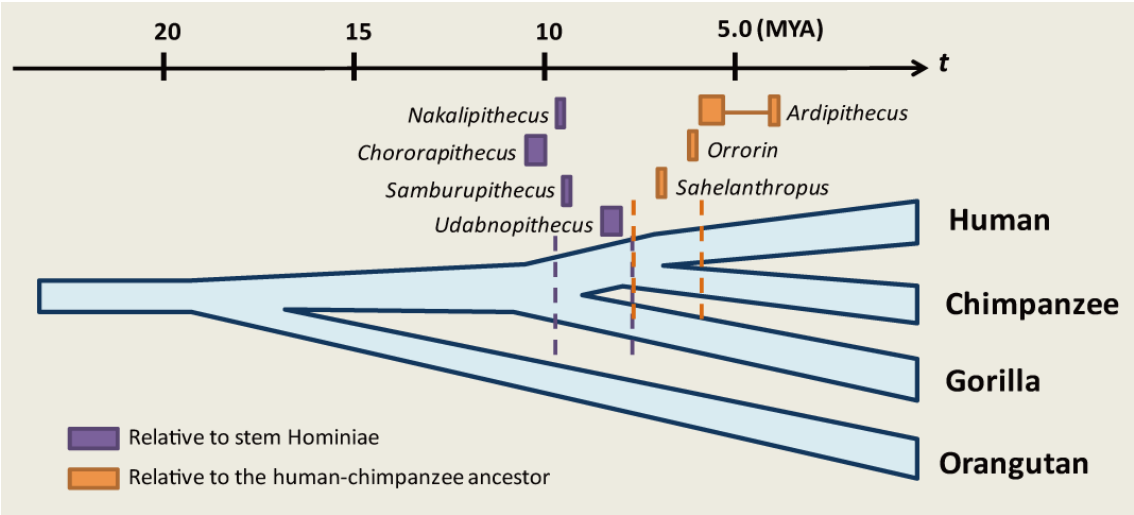

20
15
10
5.0 (MYA)
t
Nakalipithecus
Chororapithecus
Samburupithecus
Udabnopithecus
Ardipithecus
Orrorin
Sahelanthropus
Human
Chimpanzee
Gorilla
Orangutan
Relative to stem Hominiae
Relative to the human-chimpanzee ancestor

Fig. 3-3

Table 3-1. Estimated parameters for each chromosomal alignment set[†].

Regions[*1]	Alignment length (Mb)	τ_{HC}	τ_{HCG}	τ_{HCGO}	θ_{HC}	θ_{HCG}	θ_{HCGO}	Branch length[*2]	-lnL
Whole genome[‡]	40.8	0.00330	0.00423	0.00819	0.00264	0.00229	0.00709	0.0352	-66035045
Autosomes[‡]	38.7	0.00326	0.00423	0.00835	0.00286	0.00223	0.00659	0.0355	-62585914
Chr. 1	3.28	0.00313	0.00408	0.00778	0.00270	0.00216	0.00691	0.0335	-5289418
Chr. 2	3.49	0.00327	0.00426	0.00827	0.00284	0.00223	0.00671	0.0350	-5650076
Chr. 3	2.92	0.00324	0.00429	0.00837	0.00303	0.00225	0.00679	0.0352	-4736827
Chr. 4	2.79	0.00345	0.00434	0.00886	0.00288	0.00254	0.00676	0.0368	-4545603
Chr. 5	2.61	0.00327	0.00436	0.00855	0.00316	0.00221	0.00644	0.0354	-4243851
Chr. 6	2.43	0.00307	0.00423	0.00837	0.0036	0.00225	0.00629	0.0346	-3934109
Chr. 7	2.13	0.00345	0.00426	0.00826	0.00221	0.00227	0.00705	0.0352	-3458675
Chr. 8	2.09	0.00352	0.00453	0.00899	0.00311	0.00242	0.00650	0.0372	-3419007
Chr. 9	1.62	0.00332	0.00431	0.00786	0.00282	0.00215	0.00678	0.0340	-2620109
Chr. 10	1.86	0.00331	0.00422	0.00827	0.00263	0.00228	0.00723	0.0353	-3027929
Chr. 11	1.86	0.00324	0.00415	0.00827	0.00271	0.00233	0.00692	0.0348	-3022784
Chr. 12	1.93	0.00317	0.00408	0.00818	0.00264	0.00234	0.00671	0.0342	-3127323
Chr. 13	1.43	0.00321	0.00434	0.00878	0.00364	0.00239	0.00660	0.0361	-2336632
Chr. 14	1.29	0.00310	0.00413	0.00792	0.00296	0.00220	0.00721	0.0344	-2093555
Chr. 15	1.14	0.00317	0.00421	0.00786	0.00324	0.00216	0.00731	0.0343	-1850007
Chr. 16	1.07	0.00360	0.00459	0.00870	0.00277	0.00238	0.00730	0.0374	-1755416
Chr. 17	1.09	0.00294	0.00385	0.00726	0.00264	0.00195	0.00827	0.0330	-1752968
Chr. 18	1.11	0.00330	0.00433	0.00880	0.00342	0.00243	0.00616	0.0358	-1801757
Chr. 19	0.725	0.00315	0.00401	0.00751	0.00285	0.00237	0.00895	0.0350	-1173725

Chr. 20	0.878	0.00310	0.00415	0.00798	0.00325	0.00225	0.00706	0.0344	-1418404
Chr. 21	0.466	0.00337	0.00444	0.00910	0.00370	0.00278	0.00665	0.0376	-761073
Chr. 22	0.454	0.00299	0.00401	0.00796	0.00328	0.00241	0.00744	0.0348	-733646
Chr. X[*3]	2.07	0.00277	0.00371	0.00637	0.00153	0.00171	0.00627	0.0295	-3286104
Coding regions[‡]	2.37	0.00156	0.00249	0.00418	0.00367	0.00137	0.00552	0.0213	-3632995
FFD 3rd positions[*4‡]	0.351	0.00437	0.00548	0.0135	0.00401	0.00480	0.00708	0.05359	-598524

†, 95 % CI of each estimated parameter and the estimates based on the sample 2 were shown in Supporting data D3-3.
*1, The region with a double dagger were analyzed based on the method (iii), assuming heterogeneity of mutation rates across lineages and chromosomes, the others based on the method (ii), assuming heterogeneity of mutation rates across the lineages (see Methods)
*2, Average of sum of the branch lengths in each locus.
*3, $\theta=3\mu gN$ based on X chromosome
*4, Four-fold degenerate sites at third codon positions

Table 3-1 (cont.)

Table 3-2. Estimated parameters in different evolutionary models and different sequence collections.

Length of blocks	Species	Regions	Methods*	Alignment length (Mb)	τ_{HC}	τ_{HCG}	τ_{HCGO}	θ_{HC}	θ_{HCG}	θ_{HCGO}	lnL
100 bp	HCGO	Whole genome	(iii)	40.8	0.00330	0.00423	0.00819	0.00264	0.00229	0.00709	-66035045
	HCG		(iii)	44.2	0.00335	0.00428	-	0.00196	0.00342	-	-66254572
	HCO		(iii)	44.1	0.00340	-	0.00927	0.00210	-	0.00566	-69356941
	HGO		(iii)	44.1	-	0.00445	0.00935	-	0.00260	0.00544	-69687113
	HCGO		(i)	44.1	0.00288	0.00435	0.00873	0.00364	0.00268	0.00764	-66041311
	HCGO	Autosomes	(i)	38.7	0.00285	0.00436	0.00896	0.00384	0.00263	0.00704	-62592112
	HCGO	X	(i)	2.07	0.00250	0.00389	0.00722	0.00268	0.00300	0.00941	-3286253
200 bp	HCGO	Whole genome	(iii)	77.5	0.00336	0.00429	0.00862	0.00245	0.00214	0.00600	-125337308
50 bp	HCGO	Whole genome	(iii)	20.9	0.00371	0.00389	0.00772	0.000403	0.00267	0.00647	-33628306
	HCO		(iii)	22.6	0.00393	-	0.0109	0.000386	-	0.000532	-33678513

* (i): The uniform model assuming a uniform mutation rate, (iii): assuming heterogeneity of mutation rates across lineages and chromosomes (see Methods)

Table 3-3. Estimated relative ratios of the mutation rates to μ_H.

Relative ratio to μ_H	μ_H	μ_C	μ_G	μ_O	μ_{HC}	μ_{HCG}	μ_{HCGO}
Whole genome	1	1.004	1.034	1.091	1.005	1.025	1.091
X chromosome	1	0.9965	1.073	1.159	1.001	1.070	1.159

Table 3-4. Estimated speciation times and ancestral population sizes.

μ_H (/year·site)	T_{HC} (MYA)	T_{HCG} (MYA)	T_{HCGO} (MYA)	N_{HC}	N_{HCG}	N_{HCGO}	$N_{HC(X)}$	$N_{HCG(X)}$	$N_{HCGO(X)}$
0.436×10^{-9}	7.57	9.70	18.8	75,600	65,500	203,000	43,800	49,200	180,000
0.556×10^{-9}	5.94	7.61	14.7	59,300	51,400	159,000	34,300	38,500	141,000
1.00×10^{-9} *	*3.30*	*4.23*	*8.19*	*33,000*	*28,600*	*88,600*	*19,100*	*21,400*	*78,400*

*The value traditionally used. This value was not used for the conclusive estimation.

Supporting data D3-1. Whole and sampled genome alignments of human and three apes.

Supporting data D3-2. Genealogies and sequence alignments of the simulation data.

Supporting data D3-3. Estimated parameters for each chromosomal alignment set with 95% CI.

(Attached in DVD-ROM).

3.6. References

Bouffard GG, et al. 1997. A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. Genome Res 7: 673-692.

Brunet M, et al. 2005. New material of the earliest hominid from the Upper Miocene of Chad. Nature 434: 752-755.

Burgess R, Yang Z 2008. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol 25: 1979-1994.

Carroll SB 2003. Genetics and the making of Homo sapiens. Nature 422: 849-857.

Chen FC, Li WH 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68: 444-456.

Chen GK, Marjoram P, Wall JD 2009. Fast and flexible simulation of DNA sequence data. Genome Res 19: 136-142.

Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.

Conrad DF, et al. 2011. Variation in genome-wide mutation rates within and between human families. Nat Genet 43: 712-714.

Dutheil JY, et al. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. Genetics 183: 259-274.

Elango N, Kim SH, Vigoda E, Yi SV 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol 4: e1000015.

Elango N, Thomas JW, Yi SV 2006. Variable molecular clocks in hominoids. Proc Natl Acad Sci U S A 103: 1370-1375.

Fujii Y, et al. 2005. A web tool for comparative genomics: G-compass. Gene 364: 45-52.

Gabunia L, Gabashvili E, Vekua A, Lordkipanidze D. 2001. The late Miocene hominoid from Georgia. In: de Bonis L, Koufos G, Andrews P, editors. Hominoid Evolution and Climatic Change in Europe. Cambridge Cambridge University Press.

Haile-Selassie Y 2001. Late Miocene hominids from the Middle Awash, Ethiopia. Nature 412: 178-181.

Hara Y, Imanishi T 2011. Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. BMC Evol Biol 11: 308.

Hara Y, Imanishi T, Satta Y 2012. Reconstructing the Demographic History of the Human Lineage Using Whole-Genome Sequences from Human and Three Great Apes. Genome Biol Evol 4: 1133-45.

Harris RS 2007. Improved pairwise alignment of genomic DNA. [Ph.D. thesis]: Pennsylvania State University.

Harrison T 2010. Anthropology. Apes among the tangled branches of human origins. Science 327: 532-534.

Hey J 2010. Isolation with migration models for more than two populations. Mol Biol Evol 27: 905-920.

Hey J, Nielsen R 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. Genetics 167: 747-760.

Hobolth A, Christensen OF, Mailund T, Schierup MH 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent

hidden Markov model. PLoS Genet 3: e7.

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome Res 21: 349-356.

Hodgkinson A, Eyre-Walker A 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12: 756-766.

Imanishi T, et al. 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2: e162.

Innan H, Watanabe H 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. Mol Biol Evol 23: 1040-1047.

Ishida H, Kunimatsu Y, Nakatsukasa M, Nakano Y 1999. New Hominoid Genus from the Middle Miocene of Nachola, Kenya. Anthropolological Science 107: 189-191.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. . In: Munro HN, editor. Mammalian Protein Metabolism. New York: Academic Press. p. 21-132.

Katoh K, Toh H 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9: 286-298.

Kawahara Y, et al. 2009. G-compass: a web-based comparative genome browser between human and other vertebrate genomes. Bioinformatics 25: 3321-3322.

King MC, Wilson AC 1975. Evolution at 2 Levels in Humans and Chimpanzees. Science 188: 107-116.

Kingman JFC 1982. The coalescent. Stochastic Processes and their Applications 13: 235-248.

Kunimatsu Y, et al. 2007. A new Late Miocene great ape from Kenya and its implications for the origins of African great apes and humans. Proc Natl Acad Sci U

S A 104: 19220-19225.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Locke DP, et al. 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469: 529-533.

Lynch M 2010. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A 107: 961-968.

Matsumura S, Forster P 2008. Generation time and effective population size in Polar Eskimos. Proc Biol Sci 275: 1501-1508.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321-324.

Nachman MW, Crowell SL 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304.

Nagaraja R, et al. 1997. X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. Genome Res 7: 210-222.

Osada N, Wu CI 2005. Inferring the mode of speciation from genomic data: A study of the great apes. Genetics 169: 259-264.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441: 1103-1108.

Pinho C, Hey J 2010. Divergence with Gene Flow: Models and Data. Annual Review of Ecology, Evolution, and Systematics, Vol 41 41: 215-230.

Pink CJ, Hurst LD 2010. Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents. Mol Biol Evol 27: 1077-1086.

Pink CJ, et al. 2009. Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates. Genome Biol Evol 1: 13-22.

Pritchard JK, Przeworski M 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1-14.

Rambaut A, Grassly NC 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13: 235-238.

Rannala B, Yang Z 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.

Roach JC, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.

Sarich VM, Wilson AC 1967. Immunological Time Scale for Hominid Evolution. Science 158: 1200-&.

Satta Y, Hickerson M, Watanabe H, O'HUigin C, Klein J 2004. Ancestral population sizes and species divergence times in the primate lineage on the basis of intron and BAC end sequences. J Mol Evol 59: 478-487.

Sawada Y, et al. 1998. K-Ar ages of Miocene Hominoidea (Kenyapithecus and Samburupithecus) from Samburu Hills, Northern Kenya. Comptes Rendus De L Academie Des Sciences Serie Ii Fascicule a-Sciences De La Terre Et Des Planetes 326: 445-451.

Saxonov S, Berg P, Brutlag DL 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad

Sci U S A 103: 1412-1417.

Scally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-175.

Sibley CG, Ahlquist JE 1984. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J Mol Evol 20: 2-15.

Stamatakis A 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

Steiper ME, Seiffert ER 2012. Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. Proc Natl Acad Sci U S A.

Steiper ME, Young NM 2006. Primate molecular divergence dates. Mol Phylogenet Evol 41: 384-394.

Suwa G, Kono RT, Katoh S, Asfaw B, Beyene Y 2007. A new species of great ape from the late Miocene epoch in Ethiopia. Nature 448: 921-924.

Tajima F 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599-607.

Takahata N, Satta Y 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. Proc Natl Acad Sci U S A 94: 4811-4815.

Takahata N, Satta Y, Klein J 1995. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 48: 198-221.

Teleki G, Hunt EE, Pfifferling JH 1976. Demographic Observations (1963-1973) on Chimpanzees of Gombe-National-Park, Tanzania. Journal of Human Evolution 5: 559-598.

Walsh PD, et al. 2008. Gorilla gorilla. In. IUCN 2012. IUCN Red List of Threatened Species. Version 2012.1. <www.iucnredlist.org>. Downloaded on 19 August 2012.

White TD, Suwa G, Asfaw B 1994. Australopithecus Ramidus, a New Species of Early Hominid from Aramis, Ethiopia. Nature 371: 306-312.

Wood B 2010. Colloquium paper: reconstructing human evolution: achievements, challenges, and opportunities. Proc Natl Acad Sci U S A 107 Suppl 2: 8902-8909.

Wu CI, Ting CT 2004. Genes and speciation. Nat Rev Genet 5: 114-122.

Yamamichi M, Gojobori J, Innan H 2011. An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees. Mol Biol Evol 29: 145-156.

Yang Z. 2006. Computational Molecular Evolution. Oxford: Oxford University Press.

Yang Z 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 162: 1811-1823.

Yang Z 1997. On the estimation of ancestral population sizes of modern humans. Genet Res 69: 111-116.

Yang ZH 2010. A Likelihood Ratio Test of Speciation with Gene Flow Using Genomic Sequence Data. Genome Biology and Evolution 2: 200-211.

Chapter 4. Identification of the Species-specific Characteristics Involved in the Pathogenicity and Adaptation to the Host Environments in *Theileria* Parasites

4.1. Summary

Theileria is a tick-born apicomplexan group causing parasitosis in livestock. Some theilerias are parasitic to cattle, but the relationship between the theileria and cattle seem to have evolved specifically in each lineage. While *T. annulata* and *T. parva* (transforming theileria) induce abnormal proliferation of infected cells of lymphocyte or macrophage/monocyte lineages and are severely pathogenic, *T. orientalis* does not induce such transformation and shows moderate pathogenicity. Here, in order to clarify the process of acquiring the high pathogenicity and diverged systems infecting the hosts, I reconstructed the evolutionary history of theileria based on the comparative genomics of the almost whole genomes. While synteny across the chromosomes of the three theilerias was well conserved, subtelomeric regions were largely different: *T. orientalis* lacks the large tandemly arrayed subtelomere-encoded variable secreted protein-encoding gene family. Through the orthologue clustering, in addition, I found that duplication and deletion rates in the transforming theileria lineage were 1.66 and 1.95 times faster than those in the *T. orientalis* lineages, respectively. Expansion of particular gene families by gene duplication was found specifically in the two transforming theileria species. One of the most notable families is the TashAT/TpHN gene family, which is considered to be involved in transformation and abnormal proliferation of host leukocytes. The transforming theileria possessed around 20 TashAT/TpHN members, while only one member was identified in *T. orientalis*, and no homologues were found in a babesia and plasmodiums. I also found the gene families expanded specifically in *T. orientalis* lineages such as ABC transporters, implying species-specific strategies against host systems. Differences between the genome

sequences of theileria species illustrated different tempo and mode of gene duplication and deletion between transforming theilerias and *T. orientalis*. It is implied, moreover, that such differences in evolutionary modes resulted in the novel abilities to transform and immortalize bovine leukocytes. The genomic changes between close relatives will provide insight into proteins and mechanisms that have evolved to induce and regulate this process.

4.2. Introduction

Theileria is a tick-borne intracellular parasite belonging to phylum Apicomplexa and parasitic to mammals through blood sucking of ticks. The hosts of many of theilerias are domestic and wild ruminants, including cattle, Asian water buffalo, sheep, goat, and African buffalo, and others are parasitic to horse. Although infection by some theileria species is asymptomatic or persists as a chronic infection, *Theileria parva* and *T. annulata* can be highly pathogenic to cattle are so-called the "transforming theileria" due to their ability to transform and induce indefinite proliferation of infected host leukocytes (Brown 1990; Brown, et al. 1973; Hooshmand-Rad and Hawa 1973; Irvin, et al. 1975). Fatal East Coast fever caused by *T. parva* results in economic losses more than $200 million dollars per year in livestock industry in sub-Saharan Africa (Norval, et al. 1992). Although prompt development of vaccine and treatments for theileriosis would be required, molecular mechanisms of theileria for invasion to the hosts and initiation or regulation of the host cell transformation event have yet to be identified or fully validated (Dobbelaere 2009; Shiels, et al. 2006).

A comparative analysis of the transforming theilerias *T. parva* and *T. annulata* genomes was reported in 2005 (Gardner, et al. 2005; Pain, et al. 2005; Weir, et al. 2009). The genomic analysis revealed that the large numbers of the genes were shared between the two species. Moreover, it also found that some theileria genes that could be involved in the transformation process, while many of which were annotated hypothetical proteins of unknown function. One way in which the genes involved in transformation of leukocytes are uncovered would be to conduct the genome comparisons with the relatives without transforming abilities. Up to there, however, the closest relative that

the whole genome is available is *Babesia bovis*, which may be too far to be compared with the transforming theileria.

Theileria orientalis, an intra-erythrocytic parasite of cattle, is a member of the non-transforming group of genus *Theileria* that proliferate in the bovine host as an intra-erythrocytic form and can generate anemia and icterus but rarely cause fatal disease (Onuma, et al. 1998). *T. orientalis* is often classified into two major strains, the Chitose type and Ikeda type, which are distinguishable on the basis of diversity in the small subunit ribosomal RNA and major piroplasm surface protein (MPSP) gene sequences (Kubota, et al. 1996). The Ikeda type is limited to eastern Asian countries including Japan, Korea, the north-eastern part of China, and Australia (Kamau, et al. 2011) and it is present in areas where livestock succumb to severe clinical cases of theileriosis and serious production losses. In contrast, *T. orientalis* Chitose is found throughout the world and is usually associated with benign infection (Kakuda, et al. 1998; Kubota, et al. 1996). Although *T. orientalis* is likely to possess relatively moderate pathogenicity compared to the transforming theileria, epidemics of *T. orientalis* is still problematic for livestock industry (Islam, et al. 2011; Sugimoto 1997), and thus clarification of its parasitism at molecular level is required. In addition, *T. orientalis* is one of the closest relatives to transforming theileria, which would be very informative as the reference of transforming theileria. Thus, *T. orientalis* can be an important pathogen in its own right and many researchers have been looking forward to the derivation of the genomic sequence to provide an important resource for further studies.

In such circumstances, an international research community including me, which was lead by Prof. Chihiro Sugimoto, Hokkaido University, sequenced the whole

genome of *T. orientalis* Ikeda for a comparison with the genomes representing the transforming theileria species, *T. parva*/*T. annulata*. The main goals of our study were to provide supportive data on existing candidates and/or identify novel candidate genes that enable transformation of bovine leukocytes upon infection with *T. annulata* and *T. parva*, and to identify the genes in *T. orientalis* involved in its specific modes of host-parasite interactions. I was engaged in the comparative genomics and molecular phylogenetics in the study. This study revealed that the *T. orientalis* genome possessed slightly larger in size (9.0 Mb) and higher GC content (41.6%) than those of the transforming theileria species, while similar number of genes to them were annotated in the *T. orientalis* genome (Hayashida, et al. 2012). The InterPro domain of DUF529, known as the FAINT (Frequently Associated IN *Theileria*) domain, described later in detail, is found frequently in all theileria species sequenced to date (Hayashida, et al. 2012). In contrast, the PEST motif, associated with rapid degradation of (nuclear) proteins, was found to be encoded by several gene families in the genomes of the two transforming theileria species but was not identified in *T. orientalis* (Hayashida, et al. 2012).

To achieve our goals, comparative analysis of the repertoire of genes between closely related species is useful. Expanded gene families or gene deletion can contribute the acquisition of novel phenotypes (Flagel and Wendel 2009; Innan and Kondrashov 2010; Kuraku and Kuratani 2012; Nei, et al. 1997; Ohno 1970; Scannell and Wolfe 2008). Following the gene annotation by the TACT system (Yamasaki, et al. 2006), which is originally developed for the annotation of human transcripts, I generated orthologous groups of piroplasms consisting three theilerias and a babesia setting two plasmodiums as outgroups and investigated the molecular evolution of each gene

families. I first examined the tempo and mode of gene evolution within genus *Theileria*. I then focused on several family groups in the theileria lineage that showed evidence of marked expansion that could be associated with acquisition of the ability for the proliferation and transformation of the infected leukocytes and of the lineage-specific adaptation to the hosts.

4.3. Materials and Methods

Orthologue clustering

Orthologue groups consisted of *T. orientalis*, *T. annulata*, *T. parva*, *B. bovis*, *Plasmodium falciparum*, and *P. vivax* proteins: *T. orientalis* genes were from TOT-DB (http://totdb.czc.hokudai.ac.jp/) (Hayashida, et al. 2012), *T. annulata* genes were from GeneDB (http://old.genedb.org/genedb/annulata/), *T. parva* and *B. bovis* genes were from RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) except for proteins coded in the mitochondria in *T. parva*, the *T. parva* mitochondrial proteome was from UniProt (http://www.uniprot.org/), and *P. falciparum* and *P. vivax* genes were obtained from PlasmoDB (http://plasmodb.org/plasmo/). Orthologue groups were generated by OrthoMCL (Li, et al. 2003) based on sequence similarity, using an all-versus-all NCBI BLASTP search (Altschul, et al. 1997) with the default parameters of OrthoMCL except the similarity cut-off value, a bit-score <60. Because E-values from the BLASTP were applied for a similarity measure in OrthoMCL, I recomputed the exact E-values between closely related proteins if the E-value was approximated at 0.0. I integrated the orthologous groups assumed to be duplicated in the theileria lineage after separation from *Babesia* into a single group, using both automatic algorithms/software and manual integration as described below. Orthologue groups A and B were merged if any *Theileria-Theileria* gene pairs in which two genes belong to A and B, respectively, had larger bit-scores than any *Theileria-Babesia*/*Plasmodium* gene pairs within the single orthologue group A or B. Several orthologue groups were merged by manual curation based on sequence homology and genomic location if they generated tandem arrays on the chromosomes. I also merged non-clustered genes from the OrthoMCL clustering

into the orthologue groups with the same procedure. Finally, 3,502 orthologue groups were used for the following analyses: PiroF0100001–PiroF0100062 represent the merged orthologue groups, and PiroF0000001–PiroF0003675 represent the other orthologue groups. The orthologue clustering left 436,111, and 293 non-clustered genes in *T. orientalis*, *T. annulata*, and *T. parva*, respectively.

Molecular phylogenetic analysis

Amino acid sequences of each orthologue group were multiply aligned with the L-INS-i alignment strategy in MAFFT (Katoh, et al. 2005) and gap-rich sequences, such as truncated ones, were removed from the alignments with MaxAlign (Gouveia-Oliveira, et al. 2007). Ambiguously and/or poorly aligned sites were removed by Gblocks (Castresana 2000), and the remainders subject to phylogenetic analysis. Phylogenetic trees were inferred by Maximum Likelihood (ML) (Felsenstein 2004; Kishino, et al. 1990) with a heuristic ML tree search using RAxML (Stamatakis 2006), with the WAG-F model (Whelan and Goldman 2001). Heterogeneity of evolutionary rates among sites was modelled by a discrete gamma distribution, with optimization of gamma shape parameter alpha for each alignment set (Yang 1994). Bootstrap probability (Felsenstein 2004) was calculated for each tree node with 1,000 replications. d_S (number of synonymous substitutions per site) on each blanch of the gene tree was calculated based on yn00 methods (Yang and Nielsen 2000) using PAML software (http://abacus.gene.ucl.ac.uk/software/paml.html). The nucleotide alignments of the coding regions for estimation of d_S were generated based on the corresponding amino acids sequence alignments so that the nucleotide sequences were aligned at codon level.

4.4. Results

Generation of gene families

Expansion of gene families specific to different theileria species could offer a valuable insight into how these parasites have evolved and adapted to their different host environments, including acquisition of leukocyte transformation capability. To examine the expansion processes of gene families in the theileria lineages in detail, I first constructed gene families comprised of sequences representing the three theileria species, *B. bovis*, and two *Plasmodium* species (*Plasmodium falciparum* and *P. vivax*) based on the orthologue clustering framework of the OrthoMCL (Li, et al. 2003), as well as additional computational and manual curations. I assigned 3,419 orthologous groups out of 3,502 in which at least one theileria species was included. While 1,740 of these orthologous groups consisted of single-copy genes across all six species, 223 orthologous groups possessed theileria paralogues (Supporting data D4-1).

Tempo and mode of gene duplication and deletion

Gene duplication followed by selection is considered to be one of the main sources for phenotypic changes during evolution (Ohno 1970). In order to see how parasites utilize the gene duplication as the source for species specific adaptation, I investigated the variation of duplication tempo and mode among the apicomplex lineages using the orthologue groups. Phylogenetic tree of each group was inferred using the amino acid sequences (See Methods). Based on the topologies of these gene trees, the age distributions of duplications were plotted as d_S distribution of duplications (Fig. 4-1**A**, **B**): d_S, due to its neutrality, is considered to be well correlated to evolutionary time.

In every apicomplex species, numbers of duplication events monotonically decreased with the increase of d_S (Fig. 4-1**A**, **B**) as observed in some eukaryotes (Blanc and Wolfe 2004; Vandepoele, et al. 2004), suggesting that no polyploidizations and whole genome duplications in at least recent lineage. While the characteristics of the highest frequency at the lowest d_S and rapid dropping-off in the age distribution were shared across all six species, the numbers of duplications and the degree of decrease varied among the species (Fig. 4-1**A**) , consistent with the previous observation of many of eukaryotes (Lynch, et al. 2001; Vanneste, et al. 2012). In the theileria lineage, duplication frequencies in the lineages involved in transforming theilerias were remarkably larger than that in *T. orientalis* lineages when d_Ss were not large (d_S<0.5) (Fig. 4-1**B**). On the other hand, in *B. bovis*, numbers of duplication events dropped rapidly with increase of d_S, suggesting recent expansion followed by deletion in the babesia lineage.

One of the causes of such variation may be influence of the large gene families. Such gene families, consisting of more than 50 paralogues in a species, correspond to the large portion of the frequency distribution of duplication. In order to exclude the effect of the large gene families from the age distributions of the duplications, I plotted the d_S distribution of the duplications in the gene families with less than six paralogues in each species, and found that the age distribution of *T. orientalis* was not different from those of transforming theilerias ($p \geq 0.127$) (Fig. 4-1**C**). On the other hand, the age distributions in three theileria significantly differed from those in *B. bovis* ($p < 6.82 \times 10^{-4}$) and *P. vivax* ($p < 0.00126$). This observation suggests that gene duplication tempo in large gene family varies between *T. orientalis* and transforming theileria while that in small gene family are shared across them.

Age distribution of gene duplications can represents lineages-specific tempo and modes of gene duplications and deletions. In order to examine the duplication-deletion process during theileria evolution, I mapped numbers of duplication and deletion events upon the branches of the species tree (Fig. 4-2**A**). It was found that numbers of duplications and gene losses in the transforming theileria lineages were 1.66 and 1.95 times on average as large as those in the *T. orientalis* lineage. This suggested that more frequent duplication and deletion processes had occurred in transforming theileria lineages, which could be mainly composed by the member of large gene families. Fraction of the lineage specific duplication genes in *T. orientalis* was significantly less than those in transforming theilerias ($p<8.76E-04$, Fig. 4-2**A**), consistent with the heterogeneity of the duplication-and-loss between the two lineages. These observations suggest the hypothesis where *T. orientalis* have kept ancestral gene family sets of large size and the transforming theilerias newly expanded several gene families.

In order to clarify the models of gene expansion histories in the three theilerias, I investigated the gene expansion processes on the chromosomal positions. Genome comparisons of three theilerias revealed that gene components of the subtelomeric regions varied (Hayashida, et al. 2012). As well as the transforming theilerias (Gardner, et al. 2005; Hayashida, et al. 2012; Pain, et al. 2005), recently expanded genes in *T. orientalis* are frequently located in the genomic regions close to telomere (Fig. 4-2**B**). However, expanded genes in <0.1Mb telomeric position in the transforming theilerias are as 1.6 times as those in *T. orientalis*. In such regions, the genes including FAINT domains were frequently found (Hayashida, et al. 2012). In addition, bimodal peaks were observed specifically in transforming theileria. This peak distant from telomere is due to the large tandem arrays of Tash-AT/TpHN and/or Tpr/Tar genes in *T. annulata*

and *T. parva*. Size distribution of tandem arrays of duplicated genes showed that the large (≥15) arrays were specifically found only in *T. annulata* and *T. parva* amongst the apicomplexans (Fig. 4-2**C**). These observations suggest that large-size gene expansions in the transforming theilerias is resultant in not only enhance of expansion in subtelomeric regions and but generation of large tandem arrays of paralogues distant from telomeres.

Expansion of gene families in the genomes of transforming theileria species

The gene families specifically expanded in the transforming theileria lineage may have contributed to the phenotypic evolution involved in the transforming abilities of the host leukocytes. We identified 12 such gene families, each consisting of ≥4 members of *T. annulata* and *T. parva*, respectively, and less members of *T. orientalis* than transforming theileria (Table 4-1). Among them, three gene families showed a striking association with the genomes of the two transforming theileria species. Among the piroplasms specific gene families, the PiroF0100022 (Tar/Tpr family), PiroF0100037 (SVSP family), and PiroF0100038 (TashAT/TpHN family) are all significantly expanded within or unique to the genomes of the transforming theileria lineage, and are comprised of genes predicted to encode proteins possessing FAINT domains. The TashAT/TpHN family of *T. annulata* contains 17 tandemly arrayed genes, some of which have been shown to encode proteins that are translocated to the host nucleus, bind DNA and alter gene expression and protein profiles of transfected bovine cells (Swan, et al. 1999; Swan, et al. 2001). An orthologous cluster of 20 genes (TpHN) has also been identified in *T. parva* (Swan, et al. 1999). In sharp contrast, only a single TashAT/TpHN like gene, TOT0100571, was identified in the genome of *T. orientalis*.

Reciprocal best hits using BLASTP indicated that this *T. orientalis* gene was likely to be the orthologue of Tash-a (TA03110) and TP01_0621 in the transforming theileria species. Both of these genes are located at the 3' end of their respective clusters in the *T. annulata* and *T. parva* genomes (Hayashida, et al. 2012).

Phylogenetic analysis suggests that Tash-a and its orthologues represent ancestral members of the TashAT/TpHN clusters (Fig. 4-3**A**). I did not find any obvious TashAT/TpHN orthologues in *B. bovis* or two *Plasmodium* species genomes. Expression analysis of TashAT in *T. annulata* revealed definite differences of the expression profiles between Tash-a and the other members. Tash-a was upregulated in the piroplasm (erythrocytic) stage, while the others were upregulated in the macroschizont (lymphocytic) stage (Hayashida, et al. 2012). This observation was supported by indirect fluorescent antibody test (IFAT) using serum raised against a Tash-a fusion protein and co-localisation of Tash-a staining with a merozoite rhoptry antigen (Hayashida, et al. 2012). Thus it is proposed that Tash-a emerged and separated into two main sub-families in the common ancestors of theileria after separation from babesia, and that gene expansion and functional diversification of the macroschizont specific TashAT/TpHN clusters have then occurred as theileria species of the transforming lineage evolved, leading to the acquisition of transformation in the infected leukocytes.

Polypeptides encoded by the subtelomeric SVSP gene family (PiroF010037) are a major component of the predicted macroschizont secretome of *T. annulata* and *T. parva*, and a number of SVSPs have been predicted to translocate to the nucleus of the infected cell. Most SVSP genes are co-expressed in cultures of macroschizont-infected cells, and the SVSP family shows a high level of amino acid sequence diversity

(Schmuckli-Maurer, et al. 2009). Further work is needed to determine the function of SVSP proteins, whether they contribute directly to the transformation of the host cell or play a role in subverting the bovine immune response. Some of the SVSPs contain signal peptides detectable *in silico* (Emanuelsson, et al. 2007), suggesting secretion into the host cell cytoplasm. Though the expression patterns of *T. parva* SVSP proteins appear complicated and their involvement in phenotypic changes in the host leukocytes remains unclear, some SVSPs encode functional nuclear localization signals (NLS) in addition to a predicted signal sequence for secretion, suggesting that they may be transported to the host nucleus and modulate signalling pathways (Schmuckli-Maurer, et al. 2009). On the other hand, it is noted that SVSP loci were completely absent from the *T. orientalis* genomes. Thus, like the TashAT/TpHN clusters, SVSP gene expansion in *T. annulata*/*T. parva* appears to be associated with species of the transforming theileria lineage and may provide an, as yet, unknown function that promotes establishment or maintenance of the proliferating macroschizont-infected leukocyte.

In addition to the SVSP and TashAT/TpHN clusters, the Tar/Tpr (PiroF0100022) family of orthologous genes showed evidence of significant expansion in the transforming theileria lineages, as only five genes dispersed over the four chromosomes were detected in *T. orientalis*, compared with the 69 dispersed Tar genes in *T. annulata*. The function of the proteins encoded by Tar/Tpr genes is unknown. They lack a FAINT domain, and the presence of multiple transmembrane domains predicts a membrane location. Transcriptome studies indicate that copies of Tpr genes dispersed throughout the *T. parva* genome are expressed in the macroschizont stage (Baylis, et al. 1991), while those organised in a tandem array of 28 genes are expressed by the intra-erythrocytic piroplasm (Skilton, et al. 2000).

Expanded gene families in T. orientalis

Lineage-specific gene expansion may have shaped the evolution of parasite species-specific phenotypes, especially with regard to adaptation to host environmental niches (Gardner, et al. 2002). I identified 17 gene families displaying expansion in *T. orientalis* (Table 4-1). I therefore considered it of interest to gain insight into the predicted gene functions of these expanded *T. orientalis* genes, integrating molecular evolutionary and biological features from the sequences.

Of the groups expanded in *T. orientalis*, only one (PiroF0100018) was annotated to well-known InterPro entries; it can be classified as an ABC transporter family. Based on a BLASTP comparison, this group was most closely related to the PfCRT (*Plasmodium falciparum* chloroquine resistance transporter) and PfMDR1 (*Plasmodium falciparum* multidrug resistance gene) genes of the ABCC subfamily in *Plasmodium falciparum*, which are involved in conferring resistance to the antimalarial drugs chloroquine and mefloquine (Reed, et al. 2000; Valderramos and Fidock 2006). PiroF0100018 included 32 *T. orientalis* genes, nearly twice as many as the 18 genes and 13 genes identified for *T. annulata* and *T. parva* (Table 4-1). *T. orientalis* paralogues definitely created a monophyletic group in the phylogenetic tree (Fig. 4-3**B**) that the members dispersed throughout all four chromosomes. Unlike TashAT/TpHN and Tar/Tpr, they were located solely or in small tandem arrays of at most three paralogues. The *T. orientalis* paralogues are not always subtelomeric: 14 out of 32 genes were outside of the subtelomeric regions. Interestingly, the members of the same tandem array are not necessarily phylogenetically closest to each other, suggesting that duplication has not specifically occurred within the tandem array but has occurred

multiply in both subtelomeric tandem arrays and in genes far from the telomeres (Fig. 4-3**B**). Orthologue groups of PiroF0100018 were observed in the other theileria species, suggesting that the expansion of the ABC-transporter homologues had occurred multiple times following divergence from the common ancestor of theileria, but that expansion of one particular paralogous group was more pronounced in *T. orientalis*.

PiroF0000008, PiroF0000009, and PiroF0000023 were *T. orientalis*-specific groups composed by unknown conserved domains. Orthologues to PiroF0000009 were not found in any other species except for *T. uilenbergi*, a parasite found in sheep from China, suggesting that these gene families have evolved specifically in *T. orientalis* and closely related lineages. Interestingly, twelve of the genes within PiroF0000008 are located on three tandem arrays, one in chromosome 1 and two in chromosome 4 and all 13 predicted polypeptides encode a single transmembrane domain at the C-terminus which shows a high level of conservation relative to the rest of the amino acid sequences (Fig. 4-4**A**). This C-terminal region was well conserved even at nucleotide acid level (Fig. 4-4**B**), strongly suggesting that gene conversions had occurred multiple times between paralogues in the arrays and that the 5'-boundary of the gene conversion events had been determined definitely within the exon.

The evolution of the FAINT domain superfamily

As observed for *T. annulata* (Pain, et al. 2005) and *T. parva* (Gardner, et al. 2005), a large number of genes whose predicted polypeptides encode DUF529 domains (IPR007480 in InterPro), alternatively called FAINT domains, were found in *T. orientalis* (Table 4-1). Previous analysis revealed that ~900 copies of FAINT domains are present in the genomes of *T. annulata* and *T. parva* (Pain, et al. 2005). With our

pipelines for InterPro annotation, 686 FAINT copies were identified in 137 *T. orientalis* proteins, while 913 and 725 copies were identified in 126 *T. annulata* and 142 *T. parva* proteins, respectively. This suggests that FAINT domain-containing polypeptides (FAINT superfamily) is likely to have already expanded in the common ancestor of the three theileria species. In addition, orthologue clustering indicated that different FAINT families have been expanded in *T. orientalis* compared with *T. parva* and *T. annulata*. For example, the FAINT superfamilies of PiroF0001942 and PiroF0001943 are specifically expanded in *T. orientalis* (Table 4-1). In contrast, the PiroF0100056 orthologous group of SfiI-related genes showed greater expansion in *T. parva* and *T. annulata* (Table 4-1). A protein of the FAINT superfamily was also found in *T. equi* (Pain, et al. 2005), which has been considered to be an outlier species in the genus *Theileria* (Mehlhorn and Schein 1998). This indicates that FAINT domain polypeptides were present in early ancestral species of the *Theileria* genus and have subsequently been subjected to differential expansion or contraction pressures as the different species evolved.

Many of the FAINT superfamily members in *T. parva* and *T. annulata* are inferred to be secretory proteins (Shiels, et al. 2006). Out of 137 proteins of the FAINT superfamily identified for *T. orientalis*, signal peptides were found in 103 by SignalP (Emanuelsson, et al. 2007), indicating that members of the FAINT superfamily is significantly enriched in signal peptides ($p = 5.97\times10^{-55}$, Fisher's exact test). Thus, the differential expansion and diversification of the secretarial FAINT domain proteins could be associated with adaptation of different theileria species to preferential host niches that require specific host-parasite interactions.

4.5. Discussion

The comparative analysis of gene families based on molecular phylogenetics among theilerias, a babesia, and plasmodiums revealed that tempo and mode of gene duplication/deletion in *T. orientalis* in small gene families is equivalent to the other parasites. On the other hand, it was also revealed that *T. orientalis* has less large gene family expansion than the transforming theilerias. In addition, *T. orientalis* had the slowest tempo of gene duplications and deletions among the three theilerias in large families, suggesting that *T. orientalis* likely remains the gene family structures of ancestral theileria. Since the calculation of numbers of duplications and deletions in each lineage was based on the gene trees, inference of incorrect phylogenetic trees may fail to count the miscount of duplication and/or deletion events, which can be caused short and ambiguous inner branches and extremely heterogeneous evolutionary rates among branches. However, the gene gain-loss estimation program CAFE 2.1 (De Bie, et al. 2006), which is based on statistical frameworks and does not require topology information of the gene tree, indicated that tempo and mode of duplication and deletion events were similar to those based on the gene trees (1.88 and 2.05 times frequent duplications and deletions in transforming theileria), still supporting the hypothesis of the ancestral gene repertoire in *T. orientalis*.

Slow tempo and mode of gene duplication specifically observed in *T. orientalis* may be unusual across the parasitic protists: rapid gene expansions are likely to play important roles for adaptation to the hosts. One of the explanations of the success in less expansion and gene loss in the gene family structures may be the well-established parasite system. If parasites and hosts have get on with nearly benign relationships,

large changes in gene repertoires in the parasites followed by the changes in host-parasite relationships might not be necessarily adaptive during evolution. The presumed ancestors of theileria may have moderate pathogenicity and non-transformation activity for the host leukocytes, and so do *T. orientalis*. In addition, separation *T. equi* and the bovine-parasitic theilerias (*T. orientalis*, *T. annulata*, and *T. parva*) could correspond to that of their respective vertebrate hosts, Perissodactyla and Artiodactyla, suggesting higher specificity toward the hosts in ancestral theileria (Hara et al., in preparation). On the other hand, expansion and gene loss in large gene families have been accelerated in the transforming theileria, shaping unusual large tandem arrays, which could lead to the generation of high pathogenicity and resistance of the hosts against it.

It is noted that although the tempo and mode of gene family evolution were various among the theileria lineages, the numbers of genes were almost equivalent across them. Excluding the *de novo* genes, which were single copy and found in the specific lineage and might be partly misannotation, 3,567 genes in *T. orientalis*, 3,598 in *T. annulata*, and 3,659 in *T. parva*. Therefore, the concordant tempo of gene duplication and deletion may be subject to the genome sizes or gene numbers.

Unlike *T. orientalis*, transforming theilerias seem to have acquired the large gene families through recent gene expansion. Some of these gene families are considered to have one or more orthologues in the common ancestors of theileria, and others were newly born in the ancestor of transforming theileria following gene duplications. It is noted that a few of these gene families are located on large tandem arrays specifically found in transforming theileria. One of the remarkable families is TashAT/TpHN gene family as described in Results. While Tash-a genes were conserved among three

theileria species, the other TashAT/TpHN genes expanded specifically in the transforming theileria lineage. Gene expression of the TashAT/TpHN genes is considered to be associated with the proliferating macroschizont-infected leukocyte (Shiels B Fau - Kinnaird, et al. 1992). Analysis of the normalized dataset showed that, in general, TashAT/TpHN family expression is consistently down-regulated as the macroschizont undergoes differentiation to the merozoite and host cell proliferation subsides, as demonstrated previously for a number of individual family members (Swan, et al. 1999). In contrast, our *T. orientalis* genome sequencing team revealed that transcripts of Tash-a were significantly up-regulated during the differentiation process (Hayashida, et al. 2012), suggesting that the translation product of Tash-a possesses a distinct function from the other TashAT/TpHN members (macroschizont specific members): Tash-a seemed to be required through merozoite production. It is implied that the expression of the macroschizont specific TashAT/TpHN were related to the acquisition of the transformation of the hosts leukocytes.

Comparison of gene repertoires among three theileria species implied that *T. orientalis* has different strategies for adaptation to the host environments from transforming theileria. I found that a group of ABCC transporter family (PiroF0000018) closely related to the *P. falciparum* homologues involved in antimalarial drug resistance (Reed, et al. 2000; Valderramos and Fidock 2006) were specifically expanded in *T. orientalis*. The ABCC gene expansion seems to partly get free from a constraint where the ABCC genes in theileria were subject to the subtelomeric regions. 14 paralogues out of 32 *T. orientalis* ABCC genes were distant from the subtelomeric regions while only a tandem array consisting of three paralogues of transforming theileria located out of the subtelomeric regions. From the point of view of “neofunctionalization” (Ohno 1970), it

has been proposed that lineage-specific expansion of ABC transporter family genes is related to host adaptation (e.g., metabolizing toxins from host inner environments) (Dean and Annilo 2005). My results predict that parasites within the *T. orientalis* lineage have expanded their ABC transporter genes to successfully implement such an adaptive strategy (Fig. 4-5).

Another candidate related to parasite-host interaction phenotypes are the gene family PiroF0000008. Unlike the ABCC gene family, the members of PiroF0000008 were found only in *T. orientalis*, suggesting gene expansion of *de novo* genes in the *T. orientalis* lineage. A half of the members, most of which were located in one of the tandem arrays, possess very distinctive structures. These genes would generate proteins with long variable N-terminal external regions linked to an extremely highly conservative C-terminal region consisting of a transmembrane domain and a short cytoplasmic region, hypothesizing that each member has different ligands but share single intracellular signalling (Fig. 4-5). This structure would be analogous to that described for RIFIN protein families in plasmodium (Gardner, et al. 2002; Scherf, et al. 2008). The tandem arrays of RIFIN genes are thought to promote evasion of the immune response by generation of a repertoire of variant polypeptide sequences that are exposed externally on the infected erythrocyte surface. Whether *T. orientalis* proteins encoded within PiroF0000008 perform an analogous function requires experimental validation and cellular localization studies

It is noted that the largest gene family specifically expanded in *T. orientalis* is related to FAINT domains. As described in Results, FAINT superfamily may have emerged and expanded from the common ancestor of theileria and different FAINT families have expanded in each lineage. In addition, the proteins included in FAINT

domains are rich in secretome, indicating that FAINT-including proteins have played a major role for the secretomes through the theileria evolution. This implies that different repertoires of FAINT families in different theileria lineages have contributed to generate different parasitic systems including host-specificity. Investigation of the FAINT family in *T. equi* would be largely informative to see the contribution of FAINT families to phenotypic evolution of theileria.

T. orientalis is the first genome sequence of a non-transforming theileria species that occupies a phylogenetic position close to the transforming theileria, and thus provides an ideal opportunity to analyze unique features of the theileria lineage-specific parasitism from an evolutionary viewpoint. Comparative genomics of the non-transforming theileria species *T. orientalis* with the transforming theileria species *T. annulata* and *T. parva* highlighted lineage-specific evolutionary tempo and mode of the changes in gene repertoires. In addition, several lineage-specific gene family expansions were identified, which may have been coincident with development of the ability to transform host leukocytes and of the lineage-specific systems against host immunity. The study provides increased understanding of the evolution of transforming theileria species at the genomic level and has generated a database that will serve as the foundation for future studies on theileria pathobiology and parasite-host cell interaction. Comparative genomics between more closely related species or even strains will illustrate the more detailed gene evolution which enables to focus on the degree of natural selection. Prof. Sugimoto's team is now sequencing the whole genome of another strain *T. orientalis* Chitose. Interestingly, Chitose is infectious to cattle and buffalo while Ikeda is parasitic to cattle. Such comparative genomics would also uncover the characteristics involved in host range among them. Since the parasites with

wide host range can cause zoonosis, clarifying the genomic signature of the host range may be helpful for resolving public health threat by such parasites.

Fig. 4-1. d_S distribution of gene duplications.

(**A**) d_S distribution of gene duplications of all the paralogues in each species. (**B**) Details of (**A**) of d_S <1.5. (**C**) d_S distribution of gene duplications of the small gene families that number of the members in every species five or less.

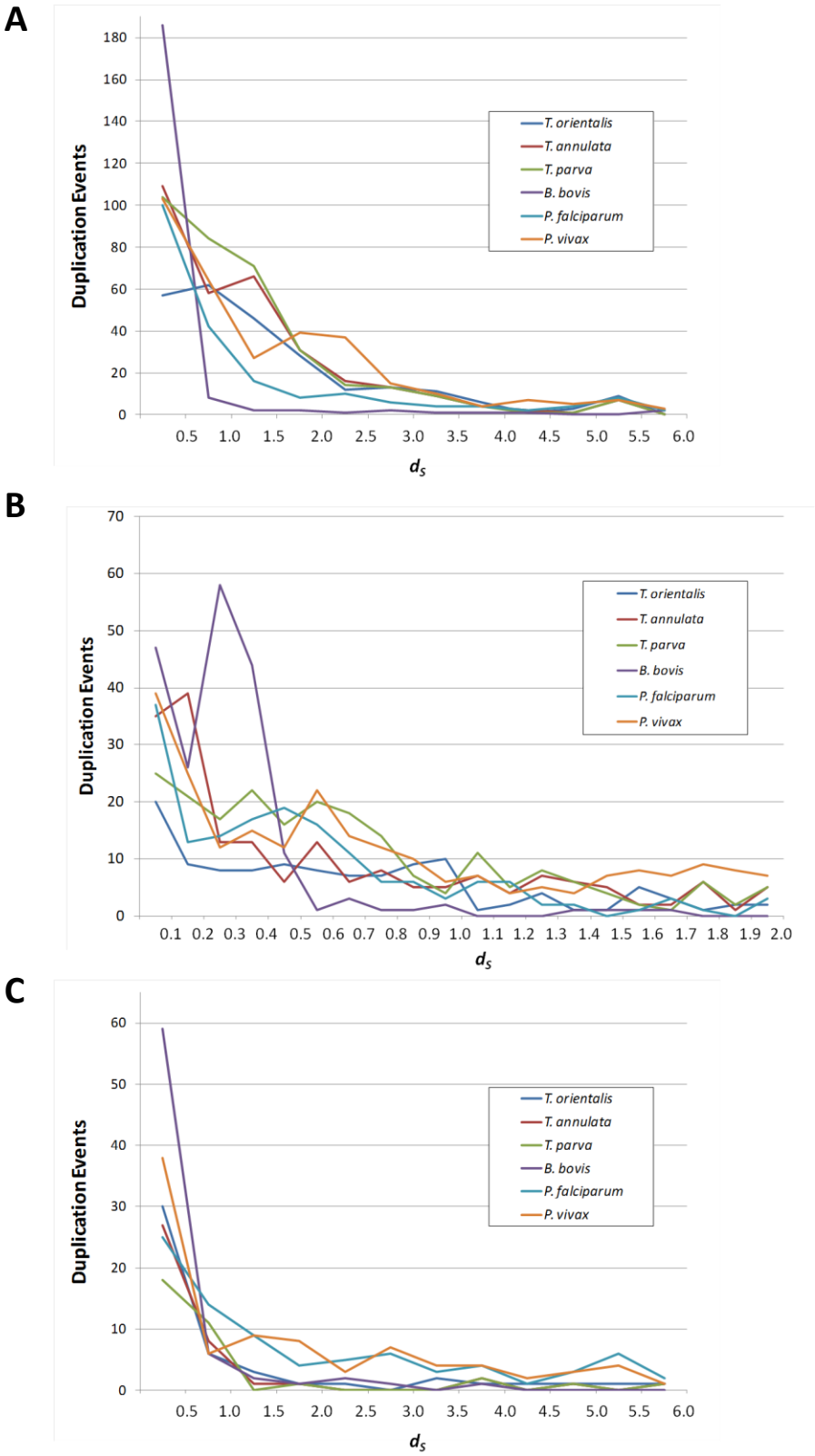

A
Duplication Events
180
160
140
120
100
80
60
40
20
0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
d_S
T. orientalis
T. annulata
T. parva
B. bovis
P. falciparum
P. vivax
B
Duplication Events
70
60
50
40
30
20
10
0
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0
d_S
T. orientalis
T. annulata
T. parva
B. bovis
P. falciparum
P. vivax
C
Duplication Events
60
50
40
30
20
10
0
0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0
d_S
T. orientalis
T. annulata
T. parva
B. bovis
P. falciparum
P. vivax

Fig. 4-1

Fig. 4-2. Evolution of gene families in each theileria lineage.

(**A**) Numbers of gene duplications and deletions in each theileria lineage. (**B**) Frequency distribution of the lineage-specific paralogues across the chromosomal regions from telomere. (**C**) Size distribution of tandem arrays of duplicated genes.

A

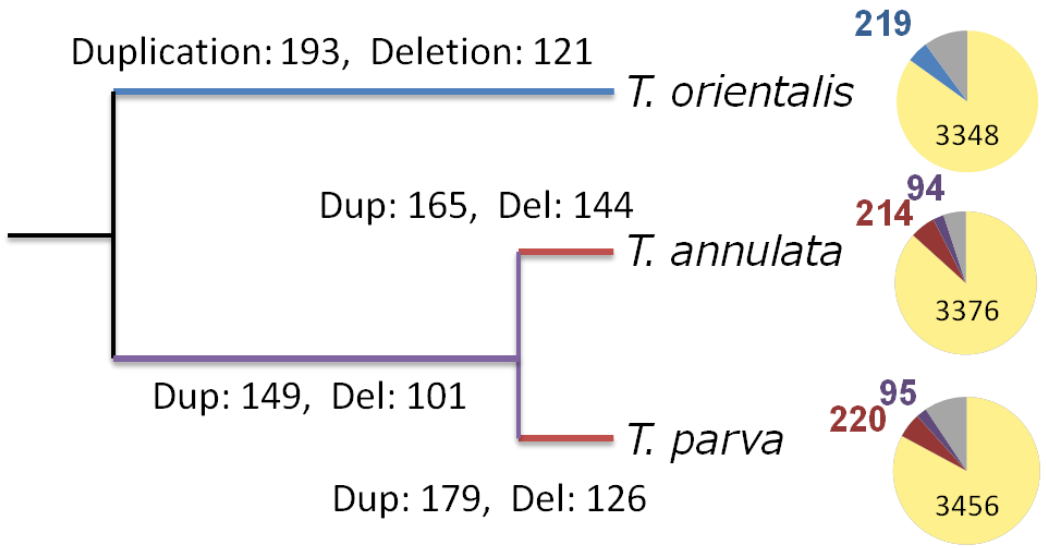

Duplication: 193, Deletion: 121
T. orientalis
219
3348
Dup: 165, Del: 144
T. annulata
214
94
3376
Dup: 149, Del: 101
T. parva
220
95
3456
Dup: 179, Del: 126

Fig. 4-2

B

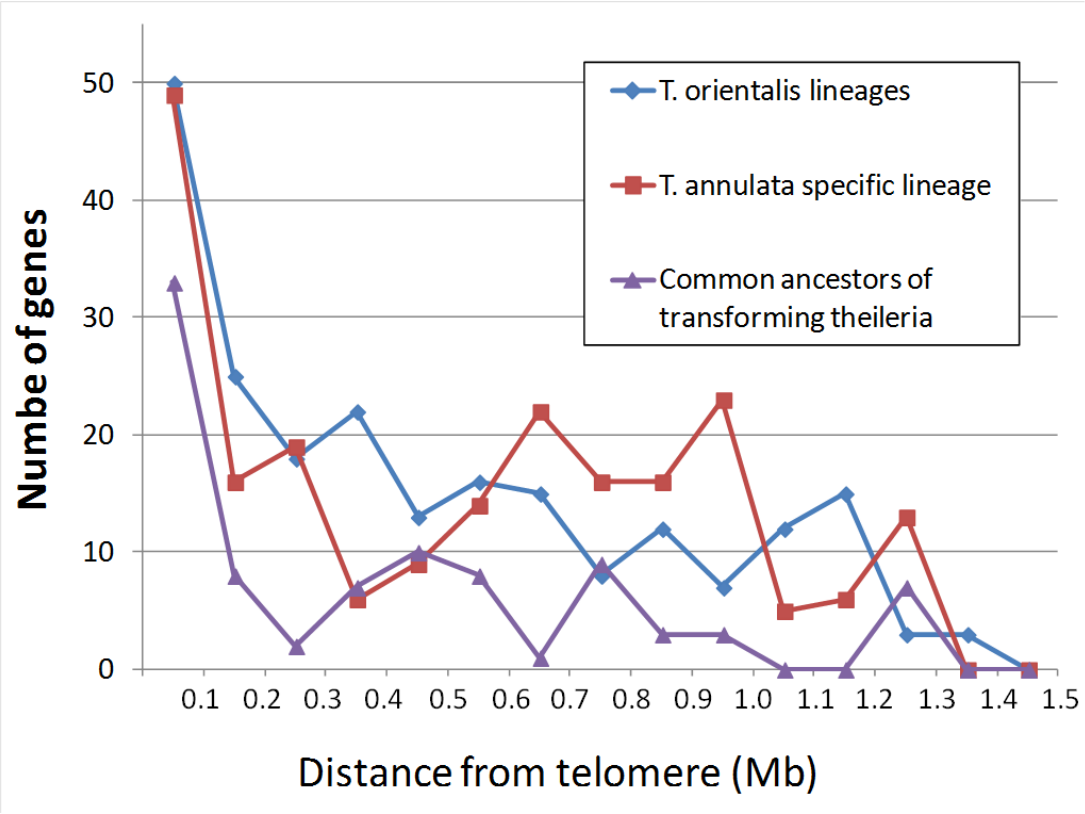

Numbe of genes
50
40
30
20
10
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Distance from telomere (Mb)
T. orientalis lineages
T. annulata specific lineage
Common ancestors of transforming theileria

Fig. 4-2 (cont.)

C

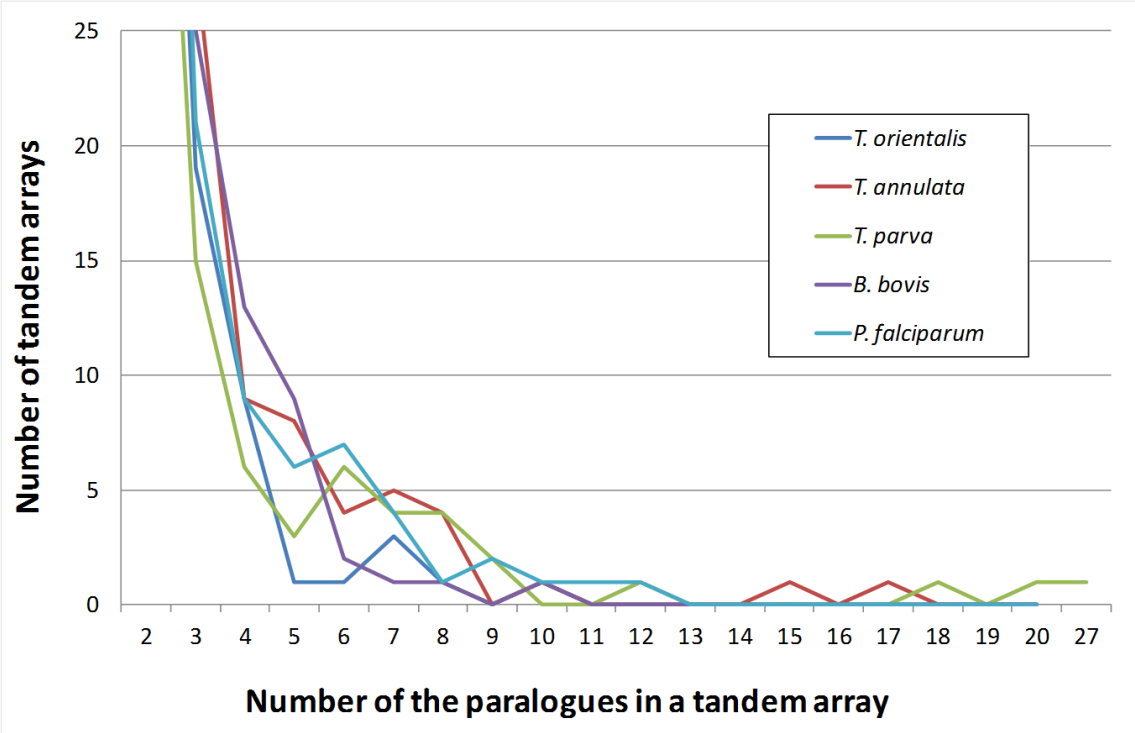

25
20
15
10
5
0
Number of tandem arrays
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
27
Number of the paralogues in a tandem array
T. orientalis
T. annulata
T. parva
B. bovis
P. falciparum

Fig. 4-2 (cont.)

Fig. 4-3. Phylogenetic relationships of TashAT/TpHN gene family (PiroF0100038) and ABC gene family (PiroF0000018).

(**A**) Phylogenetic tree of TashAT/TpHN gene family and (**B**) Phylogenetic tree of ABC gene family based on amino acid sequences. Proteins representative of *T. orientalis*, *T. annulata* and *T. parva* are indicated in red, blue and green, respectively. Bootstrap percentage (>60) values are shown at each node. In (**B**), the genes on subtelomeric regions or tandem arrays outside of the subtelomeric regions were represented by asterisk and cross, respectively.

A

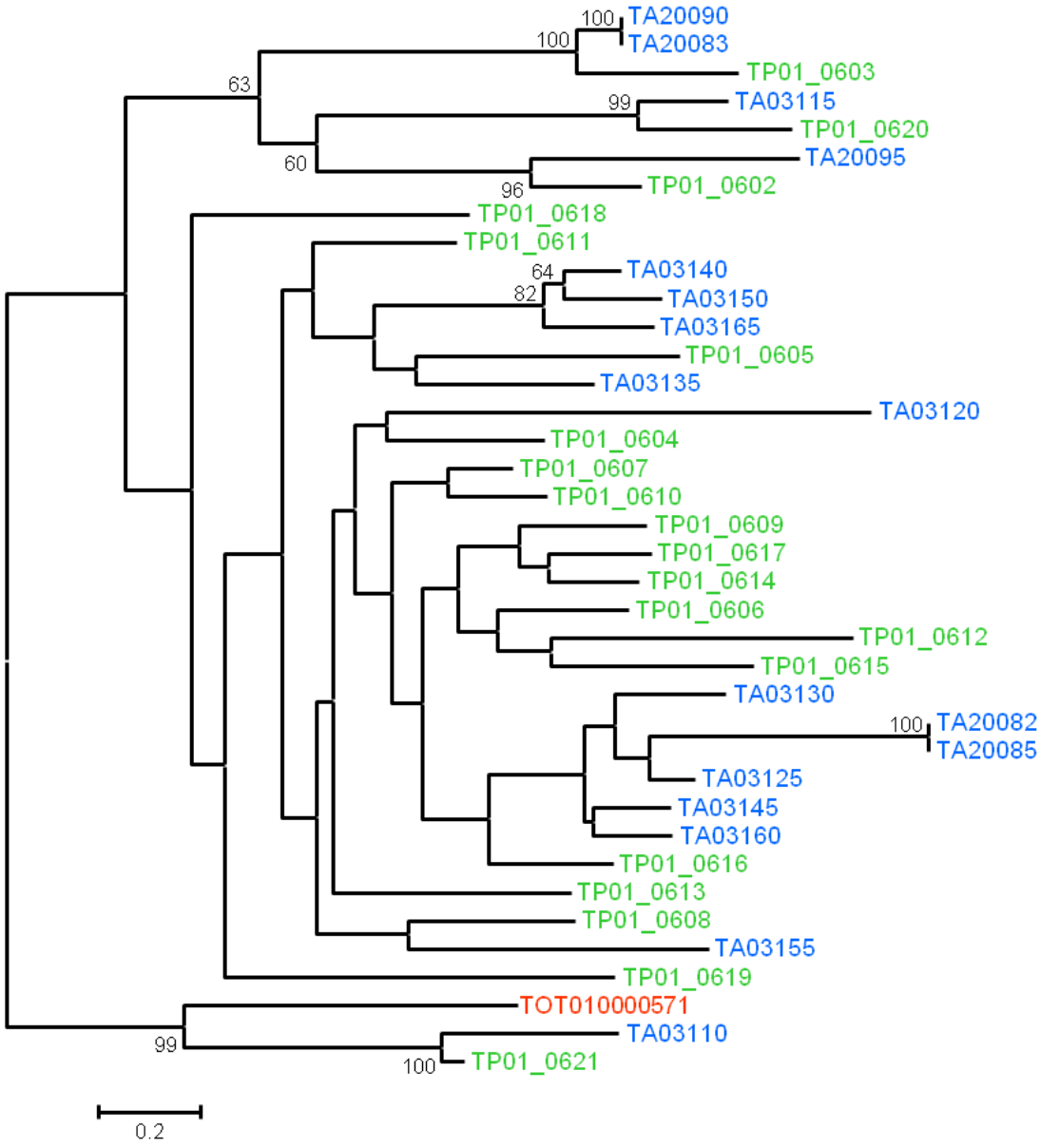

100
TA20090
100
TA20083
TP01_0603
63
99
TA03115
TP01_0620
60
TA20095
96
TP01_0602
TP01_0618
TP01_0611
64
TA03140
82
TA03150
TA03165
TP01_0605
TA03135
TA03120
TP01_0604
TP01_0607
TP01_0610
TP01_0609
TP01_0617
TP01_0614
TP01_0606
TP01_0612
TP01_0615
TA03130
100
TA20082
TA20085
TA03125
TA03145
TA03160
TP01_0616
TP01_0613
TP01_0608
TA03155
TP01_0619
TOT010000571
99
TA03110
100
TP01_0621
0.2

Fig. 4-3

B

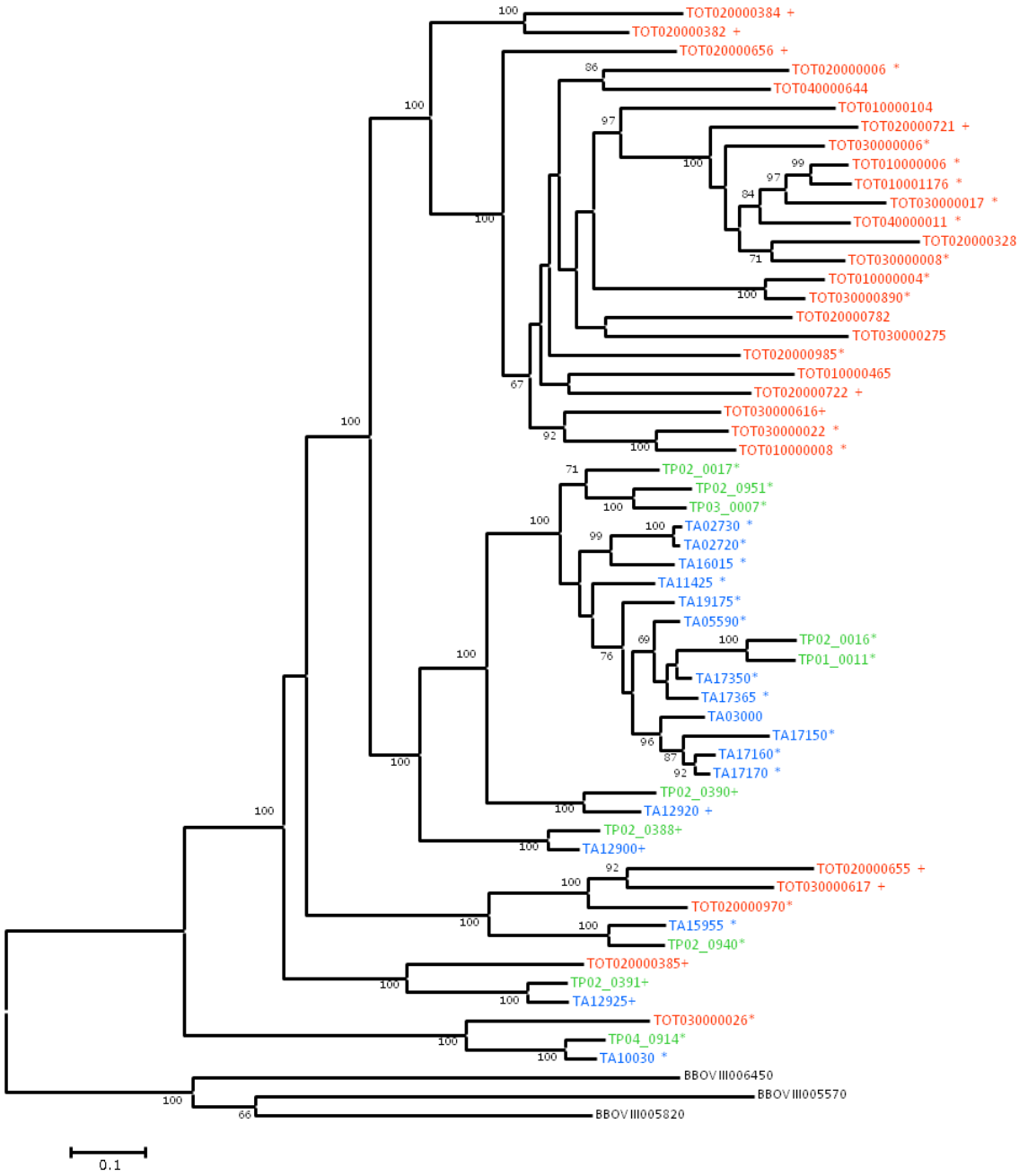

TOT020000384 +
TOT020000382 +
TOT020000656 +
TOT020000006 *
TOT040000644
TOT010000104
TOT020000721 +
TOT030000006*
TOT010000006 *
TOT010001176 *
TOT030000017 *
TOT040000011 *
TOT020000328
TOT030000008*
TOT010000004*
TOT030000890*
TOT020000782
TOT030000275
TOT020000985*
TOT010000465
TOT020000722 +
TOT030000616+
TOT030000022 *
TOT010000008 *
TP02_0017*
TP02_0951*
TP03_0007*
TA02730 *
TA02720*
TA16015 *
TA11425 *
TA19175*
TA05590*
TP02_0016*
TP01_0011*
TA17350*
TA17365 *
TA03000
TA17150*
TA17160*
TA17170 *
TP02_0390+
TA12920 +
TP02_0388+
TA12900+
TOT020000655 +
TOT030000617 +
TOT020000970*
TA15955 *
TP02_0940*
TOT020000385+
TP02_0391+
TA12925+
TOT030000026*
TP04_0914*
TA10030 *
BBOV III006450
BBOV III005570
BBOV III005820
0.1

Fig. 4-3 (cont.)

Fig. 4-4. Gene structure of PiroF0000008.

Amino acid sequence alignment (**A**) and nucleotide sequence alignment in the 36 aa C-terminal regions (**B**) are shown. Images of the alignment were generated by Jalview multiple alignment editor (http://www.jalview.org/). (**A**) The N-terminal sequences of TOT020000657 (65 aa) and TOT030000815 (21 aa) were excluded from the alignment. The C-terminal region and the putative transmembrane region were indicated by purple and black horizontal bars, respectively. The blown vertical lines indicate the members of the tandem arrays. (**B**) The nucleotide alignment of the region emphasized by the purple line in **A**. The more nucleotide sites conserved, the darker the bases were colored.

A

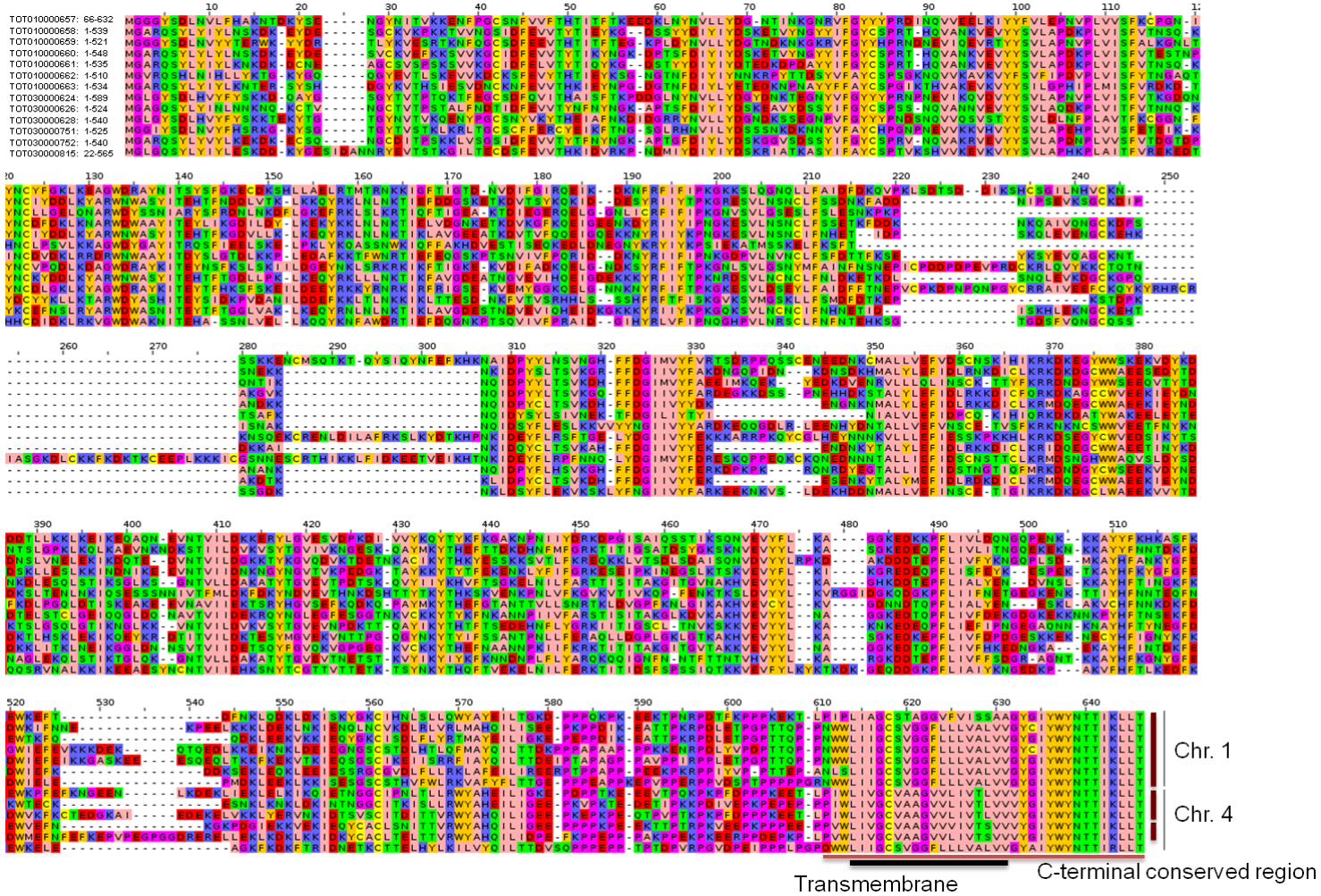


Fig. 4-4

B

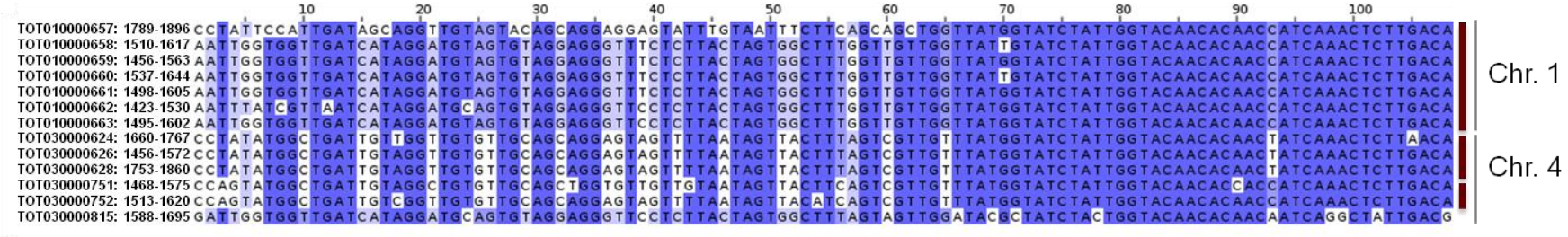

10 20 30 40 50 60 70 80 90 100
TOT010000657: 1789-1896 CCTATTCCATTGATAGCAGGTTGTAGTACAGCAGGAGGAGTATTTGTAATTTCTTCAGCAGCTGGTTATGGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT010000658: 1510-1617 AATTGGTGGTTGATCATAGGATGTAGTGTAGGAGGGTTTCTCTTACTAGTGGCTTTGGTTGTTGGTTATTGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT010000659: 1456-1563 AATTGGTGGTTGATCATAGGATGTAGTGTAGGAGGGTTTCTCTTACTAGTGGCTTTGGTTGTTGGTTATGGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT010000660: 1537-1644 AATTGGTGGTTGATCATAGGATGTAGTGTAGGAGGGTTTCTCTTACTAGTGGCTTTGGTTGTTGGTTATTGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT010000661: 1498-1605 AATTGGTGGTTGATCATAGGATGTAGTGTAGGAGGGTTTCTCTTACTAGTGGCTTTGGTTGTTGGTTATGGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT010000662: 1423-1530 AATTTATCGTTAATCATAGGATGCAGTGTAGGAGGGTTCCTCTTACTAGTGGCTTTGGTTGTTGGTTATGGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT010000663: 1495-1602 AATTGGTGGTTGATCATAGGATGTAGTGTAGGAGGGTTCCTCTTACTAGTGGCTTTGGTTGTTGGTTATGGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT030000624: 1660-1767 CCTATATGGCTGATTGTTGGTTGTGTTGCAGCAGGAGTAGTTTTAATAGTTACTTTAGTCGTTGTTTATGGTATCTATTGGTACAACACAACTATCAAACTCTTAACA
TOT030000626: 1456-1572 CCTATATGGCTGATTGTAGGTTGTGTTGCAGCAGGAGTAGTTTTAATAGTTACTTTAGTCGTTGTTTATGGTATCTATTGGTACAACACAACTATCAAACTCTTGACA
TOT030000628: 1753-1860 CCTATATGGCTGATTGTAGGTTGTGTTGCAGCAGGAGTAGTTTTAATAGTTACTTTAGTCGTTGTTTATGGTATCTATTGGTACAACACAACTATCAAACTCTTGACA
TOT030000751: 1468-1575 CCAGTATGGCTGATTGTAGGCTGTGTTGCAGCTGGTGTTGTTGTAATAGTTACTTCAGTCGTTGTTTATGGTATCTATTGGTACAACACCACCATCAAACTCTTGACA
TOT030000752: 1513-1620 CCAGTATGGCTGATTGTCGGTTGTGTTGCAGCAGGAGTAGTTTTAATAGTTACATCAGTCGTTGTTTATGGTATCTATTGGTACAACACAACCATCAAACTCTTGACA
TOT030000815: 1588-1695 GATTGGTGGTTGATCATAGGATGCAGTGTAGGAGGGTTCCTCTTACTAGTGGCTTTAGTAGTTGGATACGCTATCTACTGGTACAACACAACAATCAGGCTATTGACG
Chr. 1
Chr. 4

Fig. 4-4 (cont.)

Fig. 4-5. Putative functions of ABC transporter and PiroF0000008 in *T. orientalis*.

ABC transporters may be expressed on the surface on the parasite cells, as the malaria homologues do, to excrete the aversive substances (rounds in the figure) from the hosts. Each member of the ABC transporters may have different substances. A hypothesis is raised that the PiroF0000008 member is expressed on the surface of the parasite or host cells and that each member has different affinities against ligands (triangles) but shares single intracellular signalling (black arrowheads).

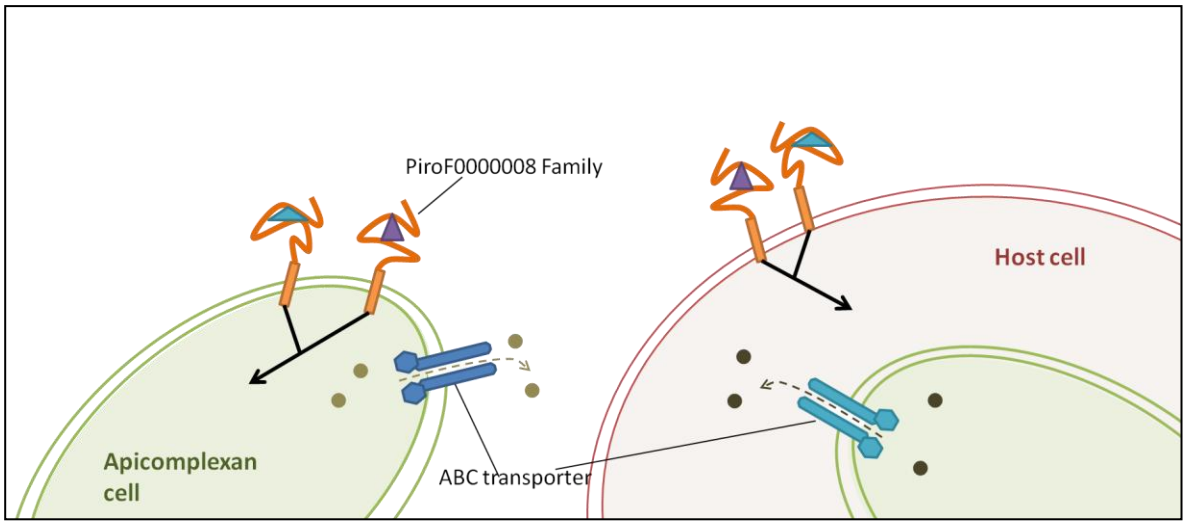

PiroF0000008 Family
Host cell
Apicomplexan cell
ABC transporter

Fig. 4-5

Table 4-1. Expanded gene families specifically in the T. orientalis and transforming Theileria lineages.

ID	#Spe-cies	#Seqs total	#Seqs TA	#Seqs TP	#Seqs TO	#Seqs BB	#Seqs PF	#Seqs PV	Definition
Transforming *Theileria* lineage expansion group									
PiroF0100022	4	122	69	44	5	4	0	0	Seven Residue (IPR011714) family.
PiroF0100056	6	189	64	97	25	1	1	1	Protein of unknown function DUF529 (IPR007480) family.
PiroF0100037	2	122	43	79	0	0	0	0	SVSP family
PiroF0100038	3	38	17	20	1	0	0	0	Tash PEST [IPR011695]; Tash1-like protein, putative Conserved hypothetical protein.
PiroF0100027	3	28	15	9	4	0	0	0	Piroplasms gene family/group candidate, PiroF0100027.
PiroF0000012	3	25	13	7	5	0	0	0	Piroplasms gene family/group candidate, PiroF0000012.
PiroF0100043	3	22	10	8	4	0	0	0	Protein of unknown function DUF529 (IPR007480) family.
PiroF0100003	6	22	8	7	4	1	1	1	Major facilitator superfamily MFS-1 (IPR011701) family.
PiroF0100024	4	21	7	8	4	2	0	0	HAD superfamily hydrolase-like, type 3 (IPR013200) family.
PiroF0100041	2	9	4	5	0	0	0	0	Protein of unknown function DUF529 (IPR007480) family.
PiroF0100024	4	21	7	8	4	2	0	0	HAD superfamily hydrolase-like, type 3 (IPR013200) family.
T. annulata expansion group									
PiroF0000038	2	11	10	1	0	0	0	0	Tash protein, PEST motif (IPR011695) family.
PiroF0000057	2	9	7	2	0	0	0	0	Piroplasms gene family/group candidate, PiroF0000057.
PiroF0000134	4	7	4	1	1	1	0	0	Seven Residue (IPR011714) family.
PiroF0000059	6	9	4	1	1	1	1	1	Histidine acid phosphatase (IPR000560) family.
T. parva expansion group									
PiroF0000026	2	14	2	12	0	0	0	0	Piroplasms gene family/group candidate, PiroF0000026.
PiroF0100029	3	11	2	8	1	0	0	0	Piroplasms gene family/group candidate, PiroF0100029.
PiroF0100039	2	9	3	6	0	0	0	0	Piroplasms gene family/group candidate, PiroF0100039.
PiroF0100042	2	6	2	4	0	0	0	0	Piroplasms gene family/group candidate, PiroF0100042.

ID	#Spe-cies	#Seqs total	#Seqs TA	#Seqs TP	#Seqs TO	#Seqs BB	#Seqs PF	#Seqs PV	Definition
T. orientalis expansion groups									
PiroF0100054	3	62	5	2	55	0	0	0	Protein of unknown function DUF529 (IPR007480) family.
PiroF0100018	5	67	18	13	32	3	0	1	ABC transporter-like (IPR003439) family.
PiroF0000009	1	25	0	0	25	0	0	0	Piroplasms gene family/group candidate, PiroF0000009.
PiroF0000008	1	25	0	0	25	0	0	0	Piroplasms gene family/group candidate, PiroF0000008.
PiroF0100023	4	42	9	7	24	2	0	0	Piroplasms gene family/group candidate, PiroF0100023.
PiroF0000030	1	12	0	0	12	0	0	0	Piroplasms gene family/group candidate, PiroF0000030.
PiroF0100046	6	16	3	3	7	1	1	1	Hly-III related (IPR004254) family.
PiroF0000037	2	11	0	4	7	0	0	0	Piroplasms gene family/group candidate, PiroF0000037.
PiroF0000032	3	12	4	2	6	0	0	0	Piroplasms gene family/group candidate, PiroF0000032.
PiroF0100004	6	14	3	3	5	1	1	1	General substrate transporter (IPR005828) family.
PiroF0001943	1	5	0	0	5	0	0	0	Protein of unknown function DUF529 (IPR007480) family.
PiroF0001942	1	5	0	0	5	0	0	0	Protein of unknown function DUF529 (IPR007480) family.
PiroF0001941	1	5	0	0	5	0	0	0	Piroplasms gene family/group candidate, PiroF0001941.
PiroF0100011	6	11	2	2	4	1	1	1	Uncharacterised protein family UPF0005 (IPR006214) family.
PiroF0002207	1	4	0	0	4	0	0	0	Protein of unknown function DUF529 (IPR007480) family.
PiroF0000199	3	6	1	1	4	0	0	0	Piroplasms gene family/group candidate, PiroF0000199.
PiroF0000103	4	8	2	1	4	1	0	0	Exon junction complex, Pym (IPR015362) family.

Table 4-1 (cont)

Supporting data D4-1. Gene families in piroplasms and plasmodium.

(Attached in DVD-ROM).

4.6. References

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Baylis HA, Sohal SK, Carrington M, Bishop RP, Allsopp BA 1991. An unusual repetitive gene family in Theileria parva which is stage-specifically transcribed. Mol Biochem Parasitol 49: 133-142.

Blanc G, Wolfe KH 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16: 1667-1678.

Brown CG 1990. Control of tropical theileriosis (Theileria annulata infection) of cattle. Parassitologia 32: 23-31.

Brown CG, Stagg DA, Purnell RE, Kanhai GK, Payne RC 1973. Letter: Infection and transformation of bovine lymphoid cells in vitro by infective particles of Theileria parva. Nature 245: 101-103.

Castresana J 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540-552.

De Bie T, Cristianini N, Demuth JP, Hahn MW 2006. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22: 1269-1271.

Dean M, Annilo T 2005. Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. Annu Rev Genomics Hum Genet 6: 123-142.

Dobbelaere DaB, M. . 2009. Theileria. In: Haas UESaA, editor. Intracellular Niches of Microbes: A Pathogens Guide Through the Host Cell . Weinheim, Germany. : Wiley-VCH Verlag GmbH & Co. KGaA.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953-971.

Felsenstein J. 2004. Inferring Phylogenies. Sunderland: Sinauer Associates, Inc.

Flagel LE, Wendel JF 2009. Gene duplication and evolutionary novelty in plants. New Phytol 183: 557-564.

Gardner MJ, et al. 2005. Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. Science 309: 134-137.

Gardner MJ, et al. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419: 498-511.

Gouveia-Oliveira R, Sackett PW, Pedersen AG 2007. MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics 8: 312.

Hayashida K, et al. 2012. Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of theileria-induced leukocyte transformation. MBio 3: e00204-12.

Hooshmand-Rad and Hawa NJ 1973. Malignant theileriosis of sheep and goats Trop Anim Health Pro 5: 97-102.

Innan H, Kondrashov F 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet 11: 97-108.

Irvin AD, Brown CG, Kanhai GK, Stagg DA 1975. Comparative growth of bovine lymphosarcoma cells and lymphoid cells infected with Theileria parva in athymic (nude) mice. Nature 255: 713-714.

Islam MK, Jabbar A, Campbell BE, Cantacessi C, Gasser RB 2011. Bovine theileriosis - An emerging problem in south-eastern Australia? Infection Genetics and Evolution 11: 2095-2097.

Kakuda T, et al. 1998. Phylogeny of benign Theileria species from cattle in Thailand, China and the U.S.A. based on the major piroplasm surface protein and small subunit ribosomal RNA genes. Int J Parasitol 28: 1261-1267.

Kamau J, et al. 2011. Emergence of new types of Theileria orientalis in Australian cattle and possible cause of theileriosis outbreaks. Parasit Vectors 4: 22.

Katoh K, Kuma K, Toh H, Miyata T 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518.

Kishino H, Miyata T, Hasegawa M 1990. Maximum-Likelihood Inference of Protein Phylogeny and the Origin of Chloroplasts. Journal of Molecular Evolution 31: 151-160.

Kubota S, Sugimoto C, Onuma M 1996. Population dynamics of Theileria sergenti in persistently infected cattle and vector ticks analysed by a polymerase chain reaction. Parasitology 112 (Pt 5): 437-442.

Kuraku S, Kuratani S 2012. Genome-wide detection of gene extinction in early mammalian evolution. Genome Biol Evol 3: 1449-1462.

Li L, Stoeckert CJ, Jr., Roos DS 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.

Lynch M, O'Hely M, Walsh B, Force A 2001. The probability of preservation of a newly arisen gene duplicate. Genetics 159: 1789-1804.

Mehlhorn H, Schein E 1998. Redescription of Babesia equi Laveran, 1901 as Theileria equi Mehlhorn, Schein 1998. Parasitol Res 84: 467-475.

Nei M, Gu X, Sitnikova T 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A 94: 7799-7806.

Norval RAI, Perry BD, Young AS. 1992. The Epidemiology of Theileriosis in Africa. London.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Onuma M, Kakuda T, Sugimoto C 1998. Theileria parasite infection in East Asia and control of the disease. Comp Immunol Microbiol Infect Dis 21: 165-177.

Pain A, et al. 2005. Genome of the host-cell transforming parasite Theileria annulata compared with T. parva. Science 309: 131-133.

Reed MB, Saliba KJ, Caruana SR, Kirk K, Cowman AF 2000. Pgh1 modulates sensitivity and resistance to multiple antimalarials in Plasmodium falciparum. Nature 403: 906-909.

Scannell DR, Wolfe KH 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome Res 18: 137-147.

Scherf A, Lopez-Rubio JJ, Riviere L 2008. Antigenic variation in Plasmodium falciparum. Annu Rev Microbiol 62: 445-470.

Schmuckli-Maurer J, et al. 2009. Expression analysis of the Theileria parva subtelomere-encoded variable secreted protein gene family. PLoS One 4: e4839.

Shiels B Fau - Kinnaird J, et al. 1992. Disruption of synchrony between parasite growth and host cell division is a determinant of differentiation to the merozoite in Theileria annulata. J Cell Sci. 101: 99-107.

Shiels B, et al. 2006. Alteration of host cell phenotype by Theileria annulata and Theileria parva: mining for manipulators in the parasite genomes. Int J Parasitol 36: 9-21.

Skilton RA, et al. 2000. A 32 kDa surface antigen of Theileria parva: characterization and immunization studies. Parasitology 120 (Pt 6): 553-564.

Stamatakis A 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

Sugimoto C 1997. Economic importance of theileriosis in Japan. Tropical Animal Health and Production 29: 49s-49s.

Swan DG, Phillips K, Tait A, Shiels BR 1999. Evidence for localisation of a Theileria parasite AT hook DNA-binding protein to the nucleus of immortalised bovine host cells. Mol Biochem Parasitol 101: 117-129.

Swan DG, et al. 2001. Characterisation of a cluster of genes encoding Theileria annulata AT hook DNA-binding proteins and evidence for localisation to the host cell nucleus. J Cell Sci 114: 2747-2754.

Valderramos SG, Fidock DA 2006. Transporters involved in resistance to antimalarial drugs. Trends Pharmacol Sci 27: 594-601.

Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A 101: 1638-1643.

Vanneste K, Van de Peer Y, Maere S 2012. Inference of Genome Duplications from Age Distributions Revisited. Mol Biol Evol.

Weir W, et al. 2009. Highly syntenic and yet divergent: a tale of two Theilerias. Infect Genet Evol 9: 453-461.

Whelan S, Goldman N 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691-699.

Yamasaki C, et al. 2006. TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB. Nucleic Acids Res 34: W345-349.

Yang Z 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39: 306-314.

Yang Z, Nielsen R 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32-43.

Chapter 5. General Conclusion

In this study, I first revealed a minute genomic structural change based on genome alignments of closely related species. I identified ultramicro inversions buried within the conventional human-chimpanzee alignments, which could be definitely distinguished from point mutations and indels. In addition, I found that ultramicro inversions are ubiquitous characters found in the organismal genomes. Using alignments with excluding the regions displaying the differences generated by molecular evolutionary mechanisms other than simple point mutations (e.g., CpG deamination and ultramicro inversions), I inferred the speciation process of human and great apes. Please note that the feasibility of method used was also examined by computer simulation, and it was shown that accurate speciation times and ancestral population sizes could be estimated based on an appropriate samplings of aligned genomic regions. The result clearly rejected the Patterson *et al*,'s hypothesis in which introgressions between the ancestors of human and chimpanzee were proposed (Patterson, et al. 2006). The speciation times estimated here could represent definitive ones in the hominoid evolutionary studies. The evolutionary history determined would be used for identifying characters, specifically emerged in the human lineage. Finally matching the events of gene evolution in the three theileria genomes with their species tree, I revealed different tempo and mode of gene duplication between *T. orientalis* and transforming theilerias. In addition, I found several gene families expanded in the genome of *T. orientalis* or transforming theileria, which may be involved in species specific parasitic characters and pathogenicity against the hosts. Expression of ,members of gene families expanded in transforming theileria would be the very first step to solve the molecular mechanisms of transformation of leukocyte and cancer-like pathogenicity in bovine hosts.

Even though I have selected different biological cases to examine the feasibility of approaches, my study certainly revealed that reconstructing the process of evolution from the closely related species could be successfully carried out by three steps of sequential analyses, i.e., (i) to obtain accurate genomic alignment with considering precise mechanism of genome evolution, (ii) to infer evolutionary histories of the species in fine precision, and (iii) to extrapolate phenotype or functional evolution based on the genome sequence comparison. As an example, I have successfully clarified the "species-ness" of theileria species. In accord, molecular biological experiments were carried out to examine the working hypothesis that gene expansions are linked to the phenotypic innovation of pathogenicity

As shown here, comparison of genomics will shed light on various aspects of biological problems, which was not possible by analyzing individual genes and much more deeper and detailed hypotheses will be proposed; e.g., differences in *cis*-regulatory regions, syntenies, three-dimensional structures of chromosomes which are related to spatial, temporal, and dosage regulation of gene expression. In near future, such working hypotheses will be replaced by the genome-centric functional resources such as ENCODE and KEGG. Knowledge on the function of respective genomic regions and functional interactions of the genes and their products is being accumulated at rapid rate, enabling us to clarify the "species-ness" in each evolutionary lineage.

Relationship between genotypes and phenotypes will be evaluated *in silico* with the help of biology understanding systems described above, which connects genotypes to phenotypes. Genomic analyses of closely related species by the approaches will resolve a lot of unanswered questions in evolutionary biology. One of the biggest and most interesting such issues is the one about the concept of "species". Even though I

have dealt with "species-ness" here, much ambiguity remains for the concepts of "species" (Hey 2001, 2006). It is certainly fascinating to consider why species evolve continuously while reproductive isolation is discontinuous. The key to tackle such problems is to analyze appropriate set of species by the integrative approaches. The species establishing allopolyploidy in plants (Rieseberg and Willis 2007) or fly species in which reproductive incompatibility was examined in details by genetic studies (Orr and Presgraves 2000) are certainly one of the best examples to which genome-centric analyses can be applied.

Takashi Miyata, a pioneer of molecular evolution, once said, "If testing a hypothesis in biology, suitable species for establishing a model must live on the earth and surely can be found because organisms are largely diversified in their living forms" (Miyata 2008). In near future, to solve primary biological questions, it will become much more important to identify the set of species than to obtain the genomic information, since the genome sequences of the species suitable for the analysis will be available from genomic resource project such as the genome 10K for vertebrates and i5k projects for insects and from *de novo* sequencing with low cost and high speed. Application of current approach to larger number of species will provide further insights because reconstructing ancestor genomes at multiple nodes can expose continuous course of species evolution with continuity. Ultimately, comparative genomics of individuals/populations on experimental biology may directly uncover the process shaping species-ness. Though the time span anyone can convey in an experiment is much shorter than the history of life and experimental environments provided may be different from those experienced in nature, experimental evolutionary studies can expose the process of evolution in real time (Blount, et al. 2008; Fry 2009; Izutsu, et al.

2012). My approach would be expected to be one of the efficient solutions for the comparative analysis of closely related species genomes in the era of non-model organisms and experimental biology.

References

Blount ZD, Borland CZ, Lenski RE 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc Natl Acad Sci U S A 105: 7899-7906.

Fry JD. 2009. Laboratory Experiments on Speciation. In: Theodore JG, Rose MR, editors. Experimental Evolution: Concepts, Methods, and Applications of Selection Experiments. Berkeley and Los Angeles, California: University of California Press.

Hey J 2001. The mind of the species problem. Trends Ecol Evol 16: 326-329.

Hey J 2006. On the failure of modern species concepts. Trends Ecol Evol 21: 447-450.

Izutsu M, et al. 2012. Genome features of "Dark-fly", a Drosophila line reared long-term in a dark environment. PLoS One 7: e33288.

Miyata T (宮田隆) 2008. DNA の情報の解読…生物物理から分子生物学・進化学へ. 物性研究 90: 505-521 (in Japanese).

Orr HA, Presgraves DC 2000. Speciation by postzygotic isolation: forces, genes and molecules. Bioessays 22: 1085-1094.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441: 1103-1108.

Rieseberg LH, Willis JH 2007. Plant speciation. Science 317: 910-914.

Acknowledgements

今西規先生（東海大医学部教授、前・産業技術総合研究所バイオメディシナル情報研究センター分子システム情報統合チーム長）には、研究における様々な機会を与えていただき、多大なご指導をいただきました。信頼をもって研究を任せてくださるとともに、研究成果の価値をより高められるように導いてくださいました。ここに深く感謝申し上げます。

颯田葉子先生（総合研究大学院大学先導科学研究科教授）には、共同研究において多大なご指導をいただいたのみならず、総研大にて学位を取得する機会を与えていただきました。所属チームの顧問である五條堀孝先生（国立遺伝学研究所教授）には、研究や進路に関する重要な岐路に、幾度にも貴重なご指導ならびに激励を賜りました。大田竜也先生（総合研究大学院大学先導科学研究科准教授）には、分野を跨ぐ本論文に統一性を持たせるための丁寧かつ示唆に富むご指導をいただきました。杉本千尋先生（北海道大学人獣共通感染症センター教授）には、タイレリア原虫論文を主論文の 1 つとして用いることを快諾していただき、共著者の承諾書の手配など学位取得に多くのご厚意をいただきました。以上の先生方にも深く感謝申し上げます。

チームの研究員、事務スタッフ、共同研究者を始めとする、これまでの研究生活における諸先輩方、同期、後輩の皆様方にも、研究活動の様々な場面においてお世話になりましたことを御礼申し上げます。特に、故渡辺純一博士、故伊藤真純博士には、優れた研究者とのネットワークに私を加えていただき、人生の先輩として色々相談させていただきました。直に御礼申し上げることが適わず残念な限りです。ここに哀悼と感謝の意を表します。

両親には心配と迷惑をかけてばかりでした。それでも絶えず信頼してくださり、暖かく見守ってくださったことは、研究を続けていくだけでなく、挫けず生きていくということにも大きな励みになりました。本当にありがとうございました。