Origin and evolution of sex determination systems in mammals

Yukako Katsura

DOCTOR OF PHILOSOPHY

Department of Evolutionary Studies of Biosystems

School of Advanced Sciences

The Graduate University for Advanced Studies

2011

Acknowledgements

Above all things, I am very grateful to my supervisor, Dr. Yoko Satta, for helpful advice, academic guidance, and encouragement throughout my study and for providing constructive and fruitful comments on this manuscript.

I would like to express many thanks to Dr. Naoyuki Takahata for many suggestions, comments, and encouragement on my research, and to Dr. Jenny Graves, who has conducted pioneering work on sex chromosome in a wide range of animals over many years, for providing me with the opportunity to visit to her laboratory and use their marsupial samples. I thank Ms. Hiroko Kondo, Drs. Atsushi Suenaga, Paul Waters, Veronica Murtagh, and Graves lab members for wonderful corroborations on my research and the staff of Kanazawa Zoo, Hamamatsu City Zoo, and Asa Zoological Parks for providing me with marsupial samples. I also thank Dr. Kateryna Makova and her lab members for sophisticated discussions and suggestions on my study.

I greatly appreciate the members of Department of Evolutionary Studies of Biosystems and Satta laboratory for their helpful advice throughout this study. I thank Drs. Mariko Hasegawa, Tatsuya Ota, and Asato Kuroiwa (Hokkaido University) of committees who provided many comments to my dissertation, and Dr. Koji Hirata of a supervisor of my research related to “science, technology, and society”. I also thank Ms. Kaori Kuno, Motoko Sumasu, Mr. Ken Nagata, and Drs. Hielim Kim, Takeshi Igawa,

Tasuku Nishioka, Yoshiki Yasukochi, Takahiro Yonezawa, and Hideyuki Tanabe for their technical advice and support, and Dr. Mineyo Iwase for constructive discussions and suggestions on my work.

In my research life, I have a lot of things to be thankful for, and good personal connection is the most important of them all. Dr. Ken-Ichiro Morohashi and his lab members provided me with consciousness as a scientist, and special techniques and knowledge about molecular biology, and sex determination. Drs. Hiromichi Morikawa, Atsushi Sakamoto, Misa Takahashi, Ikuo Miura, and Mr. Eisuke Matsuo encouraged me to be a researcher, and Megumi Fujita was always together having high motive.

Finally, my lovely family and trusted friends must also be thanked for their unflagging and vital support.

Abstract

Sex determination is essential to the reproductive success of an individual in sexually reproducing species, but the system of sex determination has evolved variously among organisms. In most sexually reproducing species, the sex of an individual is determined by combination of sex chromosomes. On the sex chromosome, different species have different types of a sex-determining gene, which is the most primary factor for the gonadal differentiation. I conducted evolutionary genetics research on the emergence and evolution of sex chromosomes and a primary sex-determination gene in mammals to better understand the evolution of sex determination systems. In Chapter 1, I provided a general introduction to this study and described the aims of this research.

In mammals, the sex-determining region Y (*SRY*) is a testis-determining gene on the Y chromosome. Chapter 2 described the molecular evolution of mammalian male-determining *SRY* genes. By comparing marsupial and eutherian *SRY* genes, I attempted to elucidate how *SRY* genes evolved and proposed a new scenario for explaining how the specialized function of male determination developed independently in marsupials and eutherians. The results revealed that the functional differentiation of the marsupial *SRY* differed from that of the eutherian. The lineage-specific changes that have been observed in the *SRY* and other sex determination-related genes (*SOX9* and *Ad4BP/SF-1*) implied that molecular coevolution of genes has occurred in the sex determination system of eutherians.

In Chapter 3, I proposed how therian (marsupial and eutherian) sex chromosomes became differentiated. Essentially, the X and Y chromosomes of these taxa originated from a pair of autosomes, with this differentiation of sex chromosomes being attributable to the suppression of recombination. Although a previous hypothesis proposed that X and Y differentiation in therian ancestors arose through a two-step process (called "evolutionary strata 1 and 2"; Lahn and Page, 1999), I posited that this differentiation arose only once and the entire sex chromosome differentiated simultaneously in the therian ancestor. However gene conversion in eutherians reduced the nucleotide divergence between some gametologs, which meant that they could subsequently be categorized as different strata. Based on these findings, I provided a new scenario to explain the differentiation of mammalian sex chromosomes by considering the effects of genomic rearrangements, such as a chromosomal inversion, on the sex chromosome.

Chapter 4 clarified the genome structure and gene family on sex chromosomes. I focused on intrachromosomal segmental duplications (ISDs) that produce tandem and/or inverted repeats (>50 kb) in neighboring regions; compared to other chromosomes, the X chromosomes of humans possess the highest number of these ISDs. Comparisons of mammalian sex chromosomes revealed that the pattern, number and/or size of ISDs on the chromosomes differed among the examined species (human, mouse, opossum, and platypus). In particular, the characteristics of these structures in the human and mouse X chromosome were shown to be considerably more complicated than those observed in the opossum and platypus. These findings implied that these

ISDs accumulated extensively in the X chromosomes of therian ancestors. I then discussed that the complexity of these structures on the eutherian X chromosome might be correlated with the evolution of multigene families, such as cancer testis antigen genes (CTAs).

In chapter 5, the molecular evolution and genome structure of the melanoma antigen gene (*MAGE*) family, which is one of CTAs and is located on the X chromosome, was examined in primate genomes. I proposed that human-specific palindromic sequences, including the *MAGE-A* genes, were conserved by negative selection. Since the *MAGE-A* genes encode epitopes of cancer cells, the binding capacity of the epitopes to highly divergent human leukocyte antigen (HLA) molecules was preserved. This finding was interesting because it could be used to better understand the significance of genomic structure on the X chromosome.

Chapter 6 provided general discussion about the results presented in Chapters 2 to 5, including sex determination systems in both mammalian and non-mammalian taxa. In Chapter 7, all of the chapters were summarized and, based on all of the findings presented, I provided a generalized description of evolution of sex determination systems, and described the biological significance of such a system having the apparently contradictory characteristics of evolutionary flexibility and stability. It is my hope that the various results and hypotheses presented here will be tested and examined further using a variety of molecular biology and evolutionary tools.

Table of contents

Chapter 1

General introduction

1.1 Sex determination mechanism

Sex determination (SD) is the process of development of reproductive organs and behavior. SD is a vital biological phenomenon in sexually reproducing species, but the mechanism varies among species. SD mechanisms are divided into two broad categories: genotypic sex determination (GSD) and environmental sex determination (ESD). In this introduction, these SD systems are reviewed.

1.1.1 Genotypic sex determination (GSD)

GSD systems occur in a large number of organisms that are distributed in three kingdoms of eukaryotes: animals, plants, and fungi. The great majority of GSD systems have a single segregating pair of chromosomes that mediate determination of the sexes, termed sex chromosomes. Among the GSD systems with differentiated sex chromosomes, female heterogametic systems are called ZZ/ZW or ZZ/ZO systems, and male heterogametic systems are XX/XY or XX/XO systems. In some species of mammals, fishes, and insects (e.g., Monotremes, *Stephanolepis cirrhifer*, *Gasterosteus*

aculeatus, Mantodae), multiple sex chromosomes, designated XnXn/XnYn, are seen (Kitano *et al*. 2009). Wrinkled Frogs (*Rana rugosa*), which are exceptional, have two different GSD systems, an XX/XY system in some populations and a ZZ/ZW system in others (Miura 2007). Some dioecious plants (e.g., *Silene alba*, *Fragaria*) and animals (e.g., some fish, and some frogs) have nascent sex chromosomes, which are not visibly differentiated (Liu *et al*. 2004, Ming *et al*. 2007; Almeida-Toledo *et al*. 2000, Peichel *et al*. 2004, Just *et al*. 2007). In general, the sex chromosome specific to the heterogametic sex is often degenerate and smaller than its partner sex chromosome, but some species have Y chromosomes that are larger than the X chromosomes (e.g. some Drosophila, polychaetes, frogs, turtles, and plants; Lewis and John 1963, Sato and Ikeda 1992, Solari and Pigozzi 1994, Matsunaga and Kawano 2001; Martinez *et al*. 2008).

There are two main types of molecular mechanisms that mediate chromosomally determined GSD. In one type of mechanism, primary sex-determination genes mediate GSD; this type is observed in several species that have a segregating sex chromosome. In humans (*Homo sapiens*), mice (*Mus musculus*), and medaka (*Oryzias latipes*), a primary male-determination gene is located on the Y chromosome. In contrast, African clawed frogs (*Xenopus laevis*) have a female-determination gene on the W chromosome. The second type of mechanism is dependent on relative gene or chromosome dose (dosage compensation). In fruit flies (*Drosophila melanogaster*), GSD is controlled by the ratio of X chromosomes to autosomes, rather than by the presence or absence of a Y chromosome (Bridges 1914). If this ratio is one, the individual develops as a female; if the ratio is two, the individual develops as a male.

In limited species, known as complementary sex determination (CDS) systems, sex is determined by ploidy. In Hymenoptera, males are often haploid and developed parthenogenetically from unfertilized eggs; whereas, females are diploid and developed from fertilized eggs (Beukeboom, Kamping and van de Zande 2007). In species with CSD, sex is genetically determined by a single *cds* locus with multiple alleles: individuals that are heterozygous at this locus develop into females; whereas, hemizygotes and homozygotes develop into haploid and diploid males, respectively (Whiting 1943; Beye *et al*. 2003).

1.1.2 Environmental sex determination (ESD)

Triggers for ESD systems are various environmental factors, such as temperature, body size, crowding, and stress. Although reptiles have sex chromosomes (Z and W chromosomes), their sex is often determined by ambient temperature during embryonic development; this phenomenon is often called temperature-dependent sex determination (TSD) (Quinn *et al*. 2007). The relationship between TSD and GSD systems was clarified by data from the Australian central bearded dragon lizard (*Pogona vitticeps*) (Quinn *et al*. 2007). Between 22 and 32°C, the sex ratios do not differ significantly from 1:1, which is consistent a ZZ/ZW GSD system (Quinn *et al*. 2007). However, between 34 and 37°C, there is an increasing bias toward female development, suggesting that temperature can override genotypic sex in some males (Quinn *et al*. 2007). The interaction between TSD and GSD systems in *P. vitticeps* is not common to other

reptiles such as lizards, turtles, and alligators (Ezaz *et al*. 2009; Bull 1980). The high diversity of SD mechanisms seen in reptiles (e.g. XY, XXY, ZW, ZZW, TSD, genetic-environment interactions) may be a remnant of the evolutionary lability of sex determination in reptiles (Ezaz *et al*. 2009).

ESD and TSD are also common in fishes. Moreover, environment-dependent sex reversal is often reported in teleost fish species. There are many different types of triggers for sex reversal; most are abiotic (e.g. temperature, pH, endocrine-disrupting chemicals, photoperiod, hypoxia), but a few biotic factors are also known to induce sex reversal (e.g. crowding, pathogens like Wolbachia, population size) (Stelkens and Wedekind 2010).

The water flea (*Daphnia magna*) can switch from parthenogenetic into sexual reproduction when environmental quality declines (Hebert 1978). Without the environmental stress, however, the analog of juvenile hormone or the high-expression of *doublesex* (*dsx*) *1* gene also can induce male production (Kato *et al*. 2011). ESD triggers are environmental. Furthermore, ESD is implemented by several internal factors, such as endocrine hormones and gene expression (Kato *et al*. 2011).

1.2 Mammals

Mammals represent a class of vertebrates (Mammalia) that is characterized several traits, including mammary glands, hair, and unique skeletal structures. The fossil records

reveals that mammals arose in or before the Early Jurassic, ~200 million yeas ago (MYA) (Luo 2007; Rowe, Macrini and Luo 2011). In the early Cenozoic, the adaptive radiation of mammals accelerated as new mammalian species occupied niches vacated due to the extinction of non-avian dinosaurs (Luo 2007). According to "Mammal Species of the World" (Wilson and Reeder 2005), 5416 extant mammalian species were known in 2005, and these species were distributed into 1,229 genera, 153 families, and 29 orders. Mammalian species have adapted to a variety of environments and climatic conditions worldwide, and this class contains substantial morphological diversity. The class is divided into two subclasses: the Prototheria (order of Monotremata; monotremes) and the Theria, which comprises two infraclasses—Metatheria (marsupials) and Eutheria (eutherians).

1.2.1 Monotremes

Monotremes have morphological characters of primitive mammals that are similar to characteristics of reptiles; monotremes lay eggs and have a single cloaca that is the opening duct for the intestinal, reproductive, and urinary tracts. There are five extant monotreme species that represent 3 genera, 2 families, and 1 order, and all five species live only in Australia. The oldest recorded monotreme fossil is from ~120 MYA (Rowe *et al*. 2007), and monotremes emerged in the early Cretaceous (Luo 2007). Molecular clock studies support the hypothesis that the monotreme and Theria clades diverged between 231 and 217 MYA(van Rheede *et al*. 2006).

1.2.2 Eutherians

Theria is viviparous and eutherians have placenta. More than 90% of extant mammals are eutherians, this infraclass includes 5032 known species (representing 1123 genera, 130 families, 21 orders), and those species exhibit ecomorphological diversity and have been radiated to the world since ~100 MYA.

The oldest recorded therian fossils are from ~167 MYA (Flynn *et al*. 1999), and molecular clock studies using extant therian genes indicated that the divergence between eutherians and marsupials occurred 148~190 MYA (Kumar and Hedges 1998; Woodburne, Rich and Springer 2003; van Rheede *et al*. 2006).

1.2.3 Marsupials

The newborn kangaroo is 0.003% of its mother's weight, but a mouse or human newborn is about 5% of its mother's weight (Tyndale-Biscoe 2005). Most female marsupials have pouches in which the fetus develops. In marsupials, 379 species (representing 103 genera, 21 families, and 7 orders) are known. More than 70% of marsupials live in Australia and/or on a nearby island; the remaining marsupial species live on the American continent. After 15th century, when Europeans colonized North and South America, a rapid extermination of marsupial and avian species occurred; often, hunting, ecocide, and introduction of larger invasive mammals are invoked to

explain this extinction.

Marsupials first emerged in South America and then spread into Australia from South America via Antarctica (Cox 1974). Molecular phylogenetic analyses suggest that all extant marsupials have a most recent common ancestor from the Late Cretaceous, ~76 MYA (Meredith, Westerman and Springer 2009). During this period, South America and Australia began to separate from Antarctica, according to geological studies (Archer and Kirsch 2006). The three continents separated completely in the Eocene, by sometime between 45 and 35 MYA (Archer and Kirsch 2006). After the geographical separation, marsupial radiations occurred independently in South America and Australia.

1.3 The aims of this study

In SD systems, primary SD factors seem to evolve quickly, but the genetic cascades controlling sexual differentiation seem to evolve more slowly. Primary SD factors are defined as a first operator for determining sex on the reproductive organ, including a sex-determination gene, or environmental factor, and induce differentiation of gonads or other sex-related traits. Some of primary SD factors are known a transcriptional factor, which regulates downstream genes or genetic cascades related to sexual differentiation. In Diptera, fruit flies (*Drosophila melanogaster*) determine femaleness via a chromosome ratio, but *Musca domestica* specify maleness based on the presence or

absence of the M gene, which resides on the Y chromosome (Shearman 2002). The different SD systems were established after these lineages diverged, 29 to 80 MYA (Wiegmann *et al*. 2003). The rapid evolution of SD systems has also occurred in the in vertebrate lineages. A male-determination gene *DMY* emerged 180-100 MYA in medaka (Matsuda *et al*. 2005), and *SRY* emerged 148-190 MYA in Theria (Wallis *et al*. 2007). In general, the master gene in a developmental system has been conserved in many lineages; the Paired box gene 6 (*Pax6*) that regulates eye development is a typical example of a conserved master gene (Kozmik 2008). Importantly, however, every known primary SD gene regulates sexual differentiation genes that encode a DM domain; DM domain-containing sexual differentiation proteins, such as dsx in invertebrates and dmrt in vertebrates, are essential for sexual development and sex-specific gonadogenesis. Sexual differentiation is regulated by sex-specific alternatively spliced isoforms of *dsx* in insects and by sex-dependent expression of *dmrt1* in vertebrates (Yoshimoto *et al*. 2010; Elllengren 2011). It is interested why and how the flexibility of primary SD factors and the stability of genetic cascades regulating sexual differentiation have evolved in each SD system. To understand that, I studied the origin and evolution of SD systems and sex chromosomes in mammals.

In chapter 2, I described the molecular evolution of *SRY*, a eutherian male-determination gene. By comparing marsupial and eutherian sequences, I attempted to reveal when and how those *SRY* genes have gained a function of male determination. In chapter 3, I investigated how therian sex chromosomes differentiated and claimed a

part of published theory regarding the evolution of sex chromosomes. In chapter 4, the evolution of repeat sequences and genes on the X chromosome was assessed. In chapter 5, the molecular evolution of a gene family on the X chromosome was investigated with a focus on primate genomes. In chapter 6, I overviewed the results presented in chapters 2-5, including the evolutionary flexibility of primary SD and stability of genetic cascades in non-mammalian organisms. In chapter 7, I provided an overarching summary of the fields of evolution of sex determination and sex chromosomes.

1.4 References

Archer, M. and Kirsch, J. A. W. (2006) 'The evolution and classification of marsupials. In: *Marsupials* (ed. P. J. Armati, C. R. Dickman and I. D. Hume).', pp. 1–21. New York, USA: Cambridge University Press Ltd.

Almeida-Toledo, L. F., Foresti, F., Daniel, M. F. and Toledo-Filho, S. A. (2000) 'Sex chromosome evolution in fish: the formation of the neo-Y chromosome in Eigenmannia (Gymnotiformes).', *Chromosoma* 109(3): 197-200.

Beukeboom, L. W., Kamping, A. and van de Zande, L. (2007) 'Sex determination in the haplodiploid wasp Nasonia vitripennis (Hymenoptera: Chalcidoidea): a critical consideration of models and evidence.', *Semin Cell Dev Biol* 18(3): 371-8.

Beye, M., Hasselmann, M., Fondrk, M. K., Page, R. E. and Omholt, S. W. (2003) 'The gene csd is the primary signal for sexual development in the honeybee and encodes an SR-type protein.', *Cell* 114(4): 419-29.

Bridges, C. B. (1914) 'DIRECT PROOF THROUGH NON-DISJUNCTION THAT THE SEX-LINKED GENES OF DROSOPHILA ARE BORNE BY THE X-CHROMOSOME.', *Science* 40(1020): 107-9.

Bull, J. J. (1980) 'Sex Determination in Reptiles.', *The Quarterly Review of Biology*. 55 (1): 3-21

Cox, C. B. (1974) 'Vertebrate palaeodistributional patterns and continental drift.', *Journal of Biogeography* 1 (2): 75-94

Ezaz, T., Sarre, S. D., O'Meally, D., Graves, J. A. and Georges, A. (2009) 'Sex chromosome evolution in lizards: independent origins and rapid transitions.', *Cytogenet Genome Res* 127(2-4): 249-60.

Flynn, J. J., Parrish, J. M., Rakotosamimanana, B., Simpson, W. F., Whatley, R. L. and Wyss, A. R. (1999) 'A Triassic Fauna from Madagascar, Including Early Dinosaurs.', *Science* 286(5440): 763-765.

Hebert, P. D. N. (1978) 'THE POPULATION BILOGY OF DAPHNIA (CRUSTACEA, DAPHNIDAE).', *Biological Reviews* 53 (3): 387–426

Just, W., Baumstark, A., Süss, A., Graphodatsky, A., Rens, W., Schäfer, N., Bakloushinskaya, I., Hameister, H. and Vogel, W. (2007) 'Ellobius lutescens: sex determination and sex chromosome.', *Sex Dev* 1(4): 211-21.

Kato, Y., Kobayashi, K., Watanabe, H. and Iguchi, T. (2011) 'Environmental sex determination in the branchiopod crustacean Daphnia magna: deep conservation of a Doublesex gene in the sex-determining pathway.', *PLoS Genet* 7(3): e1001345.

Kitano, J., Ross, J. A., Mori, S., Kume, M., Jones, F. C., Chan, Y. F., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M. et al. (2009) 'A role for a neo-sex chromosome in stickleback speciation.', *Nature* 461(7267): 1079-83.

Kozmik, Z. (2008) 'The role of Pax genes in eye evolution.', *Brain Res Bull* 75(2-4): 335-9.

Kumar, S. and Hedges, S. B. (1998) 'A molecular timescale for vertebrate evolution.', *Nature* 392(6679): 917-20.

Lewis, K. R. and John, B. (1963) 'Spontaneous interchange in Chorthippus brunneus.', *CHROMOSOMA* 14 (6): 618-637.

Liu, Z., Moore, P. H., Ma, H., Ackerman, C. M., Ragiba, M., Yu, Q., Pearl, H. M., Kim, M. S., Charlton, J. W., Stiles, J. I. et al. (2004) 'A primitive Y chromosome in papaya marks incipient sex chromosome evolution.', *Nature* 427(6972): 348-52.

Luo, Z. X. (2007) 'Transformation and diversification in early mammal evolution.', *Nature* 450(7172): 1011-9.

Martinez, P. A., Ezaz, T., Valenzuela, N., Georges, A. and Marshall Graves, J. A. (2008) 'An XX/XY heteromorphic sex chromosome system in the Australian chelid turtle Emydura macquarii: a new piece in the puzzle of sex chromosome evolution in turtles.', *Chromosome Res* 16(6): 815-25.

Matsuda, M. (2005) 'Sex determination in the teleost medaka, Oryzias latipes.', *Annu Rev Genet* 39: 293-307.

Matsunaga, S. and Kawano, S. (2001) 'Sex Determination by Sex Chromosomes in Dioecious Plants.', *Plant biol* (Stuttg) 3 (5): 481-488

Meredith, R. W., Westerman, M. and Springer, M. S. (2009) 'A phylogeny of Diprotodontia (Marsupialia) based on sequences for five nuclear genes.', *Mol Phylogenet Evol* 51(3): 554-71.

Ming, R., Wang, J., Moore, P. H. and Paterson, A. H. (2007) 'Sex chromosomes in flowering plants.', *Am J Bot* 94(2): 141-50.

Miura, I. (2007) 'An evolutionary witness: the frog rana rugosa underwent change of heterogametic sex from XY male to ZW female.', *Sex Dev* 1(6): 323-31.

Peichel, C. L., Ross, J. A., Matson, C. K., Dickson, M., Grimwood, J., Schmutz, J., Myers, R. M., Mori, S., Schluter, D. and Kingsley, D. M. (2004) 'The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome.', *Curr Biol* 14(16): 1416-24.

Quinn, A. E., Georges, A., Sarre, S. D., Guarino, F., Ezaz, T. and Graves, J. A. (2007) 'Temperature sex reversal implies sex gene dosage in a reptile.', *Science* 316(5823): 411.

Rowe, T. B., Macrini, T. E. and Luo, Z. X. (2011) 'Fossil evidence on origin of the mammalian brain.', *Science* 332(6032): 955-7.

Sato, M. and Ikeda. M (1992) 'Chromosomal complements of two forms of *Neanthes japonica* (Polychaeta, Nereididae) with evidence of male-heterogametic sex chromosomes.', *MARINE BIOLOGY* 112 (2): 299-307

Shearman, D. C. (2002) 'The evolution of sex determination systems in dipteran insects other than Drosophila.', *Genetica* 116(1): 25-43.

Solari, A. J. and Pigozzi, M. I. (1994) 'Fine structure of the XY body in the XY1Y2 trivalent of the bat Artibeus lituratus.', *Chromosome Res* 2(1): 53-8.

Stelkens, R. B. and Wedekind, C. (2010) 'Environmental sex reversal, Trojan sex genes, and sex ratio adjustment: conditions and population consequences.', *Mol Ecol* 19(4): 627-46.

Tyndale-Biscoe, H. (2005) *Life of Marsupials*, Collingwood, AUS: CSIRO PUBLISHING Ltd.

van Rheede, T., Bastiaans, T., Boone, D. N., Hedges, S. B., de Jong, W. W. and Madsen, O. (2006) 'The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians.', *Mol Biol Evol* 23(3): 587-97.

van Valen, L. (1973) 'A new evolutionary law.', *Evolutionary Theory* 1(1): 1-30.

Wallis, M. C., Waters, P. D., Delbridge, M. L., Kirby, P. J., Pask, A. J., Grützner, F., Rens, W., Ferguson-Smith, M. A. and Graves, J. A. (2007) 'Sex determination in platypus and echidna: autosomal location of SOX3 confirms the absence of SRY from monotremes.', *Chromosome Res* 15(8): 949-59.

Whiting, P. W. (1943) 'Multiple Alleles in Complementary Sex Determination of Habrobracon.', *Genetics* 28(5): 365-82.

Wiegmann, B. M., Yeates, D. K., Thorne, J. L. and Kishino, H. (2003) 'Time flies, a new molecular time-scale for brachyceran fly evolution without a clock.', *Syst Biol* 52(6): 745-56.

Woodburne, M. O., Rich, T. H. and Springer, M. S. (2003) 'The evolution of tribospheny and the antiquity of mammalian clades.', *Mol Phylogenet Evol* 28(2): 360-85.

Yoshimoto, S., Ikeda, N., Izutsu, Y., Shiba, T., Takamatsu, N. and Ito, M. (2010) 'Opposite roles of DMRT1 and its W-linked paralogue, DM-W, in sexual dimorphism of Xenopus laevis: implications of a ZZ/ZW-type sex-determining system.', *Development* 137(15): 2519-26.

Wilson, D. E. and Reeder, D. M. (2005) *Mammal Species of the World: A Taxonomic and Geographic Reference* (3rd ed). Baltimore, USA: The Johns Hopkins University Press Ltd.

Chapter 2

The evolution and origin of sex determination systems in Theria

2.1 Abstract

In eutherians, male gonads and maleness are determined genetically by the sex-determining region Y (*SRY*) gene on the Y chromosome. The *SRY* gene differentiated from the *SRY*-related HMG-box 3 (*SOX3*) gene before the divergence of marsupials and eutherians, but the involvement of marsupial *SRY* in male gonad development remains speculative. I examined the phylogenetic relationships among marsupial and eutherian homologs, and compared *SRY* sequences from several species. HMG domain, a DNA binding domain, in *SRY* is conserved among marsupials and eutherians, but the 3' and 5' regions adjacent to the HMG are not. In the domain, lineage-specific amino acids substitutions were found. Those amino acids substitutions might contribute functional differences between marsupial and eutherian *SRY*.

2.2 Introduction

Sex determination systems are important in reproduction. Genes involved in

gonadogenesis (e.g., *SOX9, SF-1, DMRT1, RSPO1*) are well conserved in the genetic cascades that regulate sexual differentiation in vertebrates; in contrast, upstream regulators in these cascades (i.e., primary sex determination genes) vary among vertebrates.

In eutherians, the sex-determining region Y (*SRY*) gene is the key factor in male determination (Sinclair *et al*. 1990). *SRY*, a transcription factor, contains a conserved DNA-binding high-mobility group (HMG) domain (78 amino acids). The complex formed by *SRY* and *Ad4BP/SF-1* proteins binds the testis enhancer region of *SOX9* directly and regulates *SOX9*, which drives testis formation (Sekido and Lovel-Badge 2008).

SRY has been identified in several orders of Australidelphia (Australian marsupials) (Fig. 2.1), but its functions in these taxa have not been fully examined (Foster *et al*. 1992). Although eutherian *SRY* is expressed in mainly testis and brain, wallaby *SRY* is expressed in a broad range of tissues including testis, brain, kidney, mesonephros among others (Harry *et al*. 1995). It is of interest to ask whether the marsupial *SRY* has a function in male determination.

The *SRY* gene is located on the Y chromosome and is differentiated from its allele *SOX3* on the proto-X chromosome (Wallis *et al*. 2007). If *SRY* emerged in the therian ancestor, Ameridaelphia (American marsupials) should also have *SRY* homologs. However, an *SRY* homolog has not been found in an American marsupial to date. Data from Katoh and Miyata (1999) indicates that therian *SRY* genes are monophyletic, but data from Nagai (2001) indicate that the phylogeny of eutherian and marsupial *SRY*

genes was not monophyletic. The orthology between eutherian and marsupial *SRY* is also debatable (Katoh and Miyata 1999, Nagai 2001).

To assess flexibility and stability within sex-determination systems, I first identified homologs of sex determination-related genes in phylogenetically diverged animals, including invertebrate and vertebrate. I then identified *SRY* homologs from Theria. Based on this phylogenetic analysis, which included an *SRY* homolog from an American marsupial (the opossum), *SRY* homologs form a monophyletic group. To determine when *SRY* gained a role in male sex determination, the process of functional differentiation in marsupial and eutherian *SRY* lineages was investigated using molecular evolutionary and computational analyses.

2.3 Materials and methods

2.3.1 Nucleotide sequences used in the analysis and homology search by blast program

Nucleotide sequence data and corresponding gene information were obtained from NCBI (http://www.ncbi.nlm.nih.gov/), Ensembl databases (release 62; http://uswest.ensembl.org/index.html), sea urchin genome from the Human Genome Sequencing Center at Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/project-species-o-Strongylocentrotus%20purpuratus.hgs

c?pageLocation=Strongylocentrotus%20purpuratus), Ghost Database (http://ghost.zool.kyoto-u.ac.jp/indexr1.html), *Branchiostome floridae* genome from EUKARYOTIC GENOMICS of DOE Joint Genome Institute (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html), and lamprey genome from UCSC Genome Bioinformatics (http://genome.ucsc.edu/cgi-bin/hgGateway?db=petMar1). The genomic or transcriptional sequences from 19 phylogenetically diverged animal species were analyzed and compared (Table 2.1 and Fig. 2.2).

First, to get insight on the divergence and conservation of sex determination systems, homologs of five genes (*SOX3*, *SOX9*, *Ad4BP/SF-1*, *DMRT1*, and *RSPO1*) involved in the genetic cascade regulating gonadogenesis and/or sex determination were identified in 19 animal species. Second, a BLAST search was carried out using the human or wallaby *SRY* genes as a query to identify *SRY* homologs in sequence data from eutherian and marsupial species. Nucleotide sequences of *SRY* from several marsupial species were also determined based on BLAST searches.

2.3.2 Animals samples used in the study

Tissue or hair samples from males of one American and 14 Australian marsupial species were used in this study. Liver and spleen samples were collected from four marsupials: opossums (*Monodelphis domestica*), swamp wallabies (*Wallabia bicolor*), eastern gray kangaroos (*Macropus Giganteus*), and koalas (*Phascolarctos cinereus*) at Kanazawa Zoo in Yokohama City, Japan. Hair samples were collected from four species: parma

wallabies (*Macropus parma*) and western grey kangaroos (*Macropus Fuliginosus*) at Hamamatsu City Zoo and sugar gliders (*Petaurus breviceps*) and brush-tailed rat kangaroos (*Bettongia penicillata*) at Asa Zoological Park in Hiroshima City, Japan. Genomic DNA was isolated from soft tissue samples using the DNeasy Blood & Tissue Kit (QIAGEN) and from hair samples using ISOHAIR (NIPPON GENE) and QIAamp DNA Micro Kit (QIAGEN). With genomic DNA purified from hair samples, whole genome amplification was performed using the REPLI-g Midi Kit (QIAGEN). Genomic DNA samples from seven species—tammar wallabies (*Macropus Eugenii*), striped-faced dunnarts (*Sminthopsis Macroura*), brushtail possums (*Trichosurus vulpecula*), tasmanian devils (*Sarcophilus harrisii*), fat tailed dunnarts (*Sminthopsis crassicaudata*), and eastern barred bandicoots (*Perameles gunnii*)—were a gift from Dr. Jenny Graves at Australian National University.

2.3.3 Polymerase chain reaction (PCR)

Genomic DNA (100 ng) was suspended in 50 μl of 1×Ex Taq PCR buffer, which contained 0.2 μM of each deoxyribonucleotide triphosphates (dNTP), 0.5 μM of the one pair of primers, and 1 unit of TaKaRa Ex Taq DNA polymerase (TaKaRa). The oligonucleotide primers used are shown in Table 2.2. PCR amplification included one cycle at 95°C for 30 seconds followed by 30-40 cycles of denaturing for 15 seconds at 95°C, annealing for 30-60 seconds at 50-60°C, and extension for 60 seconds at 72°C. A final extension was performed for 10 minutes at 72°C.

A different PCR method was used to amplify longer sequences, approximately 10 kb. Genomic DNA (1 μg) was suspended in 50 μl of 1×LA Taq PCR buffer, 0.4 μM of each dNTP Mixture, 0.5 μM of the two types of primers, and 1 unit of TaKaRa LA Taq DNA polymerase (TaKaRa). The oligonucleotide primers used are shown in Table 2.2. Primer sets were designed based on the evolutionarily conserved sequence. PCR conditions consisted of on cycle at 94°C for 2 minutes followed by 30 amplification cycles each consisting of denaturation for 10 seconds at 98°C and an annealing and extension step for 15 minutes at 68°C. A final extension was performed for 10 minutes at 72°C.

2.3.4 Subcloning and sequencing

PCR products were separated on 1% agarose gel and purified using the QIAquick Gel Extraction Kit (QIAGEN), or they were subcloned using the TOPO XL PCR cloning kit (Invitrogen). In the case of direct sequencing, PCR products were purified with ExoSAP-IT (United States Biochemical) for 30 minutes at 37°C followed by 15 minutes at 80°C. Those purified products were sequenced. The sequencing reactions were performed using the dideoxy chain-termination method (Sanger *et al*. 1977) using BigDye Terminator v1.1 or 3.1 Cycle Sequencing Kits (Applied Biosystems) and the sequencing reactions were analyzed on an Applied Biosystems 3130 genetic analyzer. The sequencing primers used are shown in Table 2.2.

2.3.5 Phylogenetic and data analyses

These nucleotide sequences were aligned using Clustal X (Thompson *et al*. 1997), and the results were also checked manually. Phylogenetic trees were constructed using all three methods available in the MEGA4.1 program (Tamura *et al*. 2007): neighbor-joining (NJ; Saitou and Nei 1987), minimum evolution (ME; Rzhetsky and Nei 1992), and maximum parsimony (MP; Sourdis and Nei 1988). The reliability of the trees was assessed by bootstrap re-sampling with 1000 replications. Phylogeny inference package version 3.68 (PHYLIP; Felsenstein 1989) and phylogenetic analysis by maximum likelihood (PAML; Yang 2007) were also used to construct a phlylogeny based on maximum likelihood (ML; Kishino and Hasegawa 1989).

In addition, to examine the nucleotide sequence similarity along the sequence between different species, window analysis implemented in mVISTA was conducted (Mayor *et al.* 2000). MatInspector (Quandt *et al.* 1995) was used to identify transcription factor binding sites (TFBS) of vertebrate in evolutionary conserved regions.

2.3.6 Analysis of binding affinity between DNA and proteins

The DNA-binding affinity of HMG domains from eutherian, marsupial, or mutant *SRY* proteins was investigated using molecular dynamics (MD) analysis and the molecular mechanics poisson-boltzmann surface area (MMPBSA) method, which is used to

calculate the free energy of proteins-DNA complexes in aqueous solutions (Gilson and Zhou 2007). The structure of the complex was modeled on a three-dimensional, nuclear magnetic resonance structure of human SRY HMG domain bound to a 14 nucleotide sequence (Murphy *et al.* 2001). The DNA-protein pairs used in this analysis are shown in Table 2.3. The mutant proteins were selected based on this study and a previous study. The *SRY* binding sequences from regulatory regions of the human anti-mullerian hormone (*Amh*) and marsupial *SRY* gene were used for this analysis (Table 2.3).

2.4 Results

2.4.1 Genes related to sex determination

To better understand the evolution of sex-determination systems, homologs of genes related to sex determination were investigated in 19 phylogenetically diverged animal species (Table 2.1 and Fig. 2.2). Every animal genome analyzed had homologs of essential genes for gonadogenesis and testis differentiation (*SOX3*, *SOX9*, *Ad4BP/SF-1,* and *DMRT1*) (Fig. 2.2). While only Jawed vertebrates had homologs of *RSPO1* essential for ovary differentiation (Kamata *et al*. 2004; Perma *et al*. 2006; Tomizuka *et al*. 2008), but invertebrate and jawless vertebrate (lampreys) did not (Fig. 2.2).

2.4.2 Identification of marsupial *SRY*

To assess the rate of evolution of primary sex-determination genes, I focused on the origin and evolution of *SRY* sequences. In particular, characterization of marsupial *SRY* homologs was important for understanding the origin and evolution of mammalian *SRY*. Then an *SRY* homolog from an American marsupial (AC239615; 61429-62068, 639 bp) was identified on an opossum BAC clone derived from the Y chromosome (Fig. 2.3). The sequence did not have any premature stop codons (Fig. 2.3) and thus could encode a protein-coding mRNA. The entire deduced amino acid sequence of opossum *SRY* had 63% and 67% similarity with the wallaby and stripe-faced dunnart *SRY* (S46279), respectively, and had 41% similarity with human *SRY* (NM003140). The *SRY* HMG domains (~246 bp) were highly conserved among marsupials and eutherians (Fig. 2.3), which the opossum HMG had 76%, 77%, and 64% similarity with the HMG of the stripe-faced dunnart, wallaby and humans, respectively. Additionally, *SRY* genes were identified empirically in 12 species, representing seven families, of Australian marsupials. Partial *SRY* sequences and sequences flanking *SRY* were isolated from seven species, representing five families: Macropodidae: swamp wallabies, eastern gray kangaroos, tammar wallabies; Phascolarctidae: koalas; Phalangeridae: brushtail possums; Dasyuridae: fat tailed dunnarts and Peramelidae: eastern barred bandicoots (Figs. 2.1 and 2.4). The *SRY* deduced amino acid sequences from these seven species had more than 70% similarity to those of other marsupial *SRY*. The putative *SRY* homologs in these seven species are shared gene order, with conserved synteny, as are the *SRY* genes in the opossum and tammar wallaby (unpublished date; tammar wallaby

nucleotide sequences of Y chromosome derived BAC were given by Dr. Jenny Graves and Dr. Paul Waters). Therefore, it can be concluded that the identified sequences were marsupial *SRY* homologs.

2.4.4 Regulatory regions of *SRY*

The expression patterns of marsupial and eutherian *SRY* transcripts differed. To identify the regulatory regions that are responsible for these differences, TFBSs in the 5' flanking region of marsupial *SRY* were investigated. The 5' regulatory region (~4 kb) of the wallaby *SRY* had about 600 TFBSs. There were three blocks of conserved sequence in the 5' flanking region of *SRY* genes from all marsupial species analyzed. Fig. 2.5 shows the regions that are conserved between wallaby and opossum Y chromosome, but the 5' regulatory regions of eutherian *SRY* homologs were not similar to those of marsupials. However, the 5' regions of marsupial *SRY* contain potential binding sites for some important TFs known to regulate gonadogenesis or male differentiation in eutherians. The 5' regulatory regions of some eutherian *SRY* homologs (i.e., human, bovine, goat, and pig homologs; but not the mouse gene) have conserved TFBSs (Ross *et al.* 2008). These conserved sequences include binding sites for *SP1*, homeobox genes (*Hox*), lim homeodomain factors (*Lhxf*), *Sox/Sry*, Brn POU domain factors (*Brnf*), hepatic nuclear factor1 (*Hnf1*), caudal related homeodomain protein (*Cdx*), and the CTCF and BORIS gene families (*Clox*).

2.4.3 Phylogenetic analyses of *SRY*

To understand the origin and evolution of the *SRY* gene, a phylogeny was constructed using amino acid sequences of *SRY* and the *SOXB1* family, including *SOX1-3* (Fig. 2.6). *SRY* sequences were gathered from annotated and non-annotated mammalian sequence data, and redundant or truncated sequences were removed from the analyses. The SRY cluster was monophyletic; this finding indicated that the marsupial and eutherian *SRY* homologs have a common origin. The topology of phylogeny was confirmed using four methods: NJ, ML, ME and MP. The data are inconclusive because the bootstrap values supporting the monophyly of *SOX3* are rather low.

In the *SRY* HMG domain, six marsupial-specific and 13 eutherian-specific amino acid substitutions were found (Figs. 2.6 and 2.7). Of the 13 eutherian-specific substitutions, three are evolutionary conserved (I55F, K59Q, and E68K), but in marsupials, six are specific and only one of them has substitution (V64K). By parsimony, four other substitutions could be speculated in the *SRY* of therian ancestor, although the substitutions are not conserved between *SRY* of marsupials and eutherians (Figs. 2.6 and 2.7).

2.4.4 Functional domain of *SRY*

Some conserved amino acids in HMG domains of eutherian *SRY* proteins (M9

Y69 F55 V5, H65, Y69 and Y72) are essential to the structure and function of these *SRY* proteins (Murphy *et al*. 2001; Assumpcao *et al*. 2002). However, the amino acids at two of these positions, M9M/I, Y69S/F, vary among marsupial SRY proteins (Fig. 2.7). The HMG domain of *SRY* binds in the minor groove of specific DNA sequences, resulting in substantial DNA bending (Murphy *et al*. 2001). An M9I change in the SRY HMG domain increases the angle of the bend in the DNA, and this more pronounced bend may prevent distally placed proteins from interacting with the transcription initiation complex (Assumpcao *et al*. 2002). The M9I mutation causes *SRY*-dependent 46XY sex reversal in humans (Assumpcao *et al*. 2002). M at position 9 is well conserved *SRY*/Sox1-3 HMG domains except that position was variable in marsupials (Fig. 2.7) (L9; Fig. 2.7). Amino acids V5, H65, Y69, and Y72 of SRY maintain the protein's structure by anchoring the N-terminal tail to the end of helix 3 and the beginning of the C-terminal tail. All but one of these amino acids are well conserved; Y at position 69 is replaced by S in the striped-faced dunnart and by F in the Macroposidae (kangaroo, wallaby and koala) respectively (Fig. 2.7). In addition, the eutherian-specific substitution, F55, is related to efficient packing interactions between the N-terminal strand, the N-terminal end of helix 1, and the inner face of helix. These eutherian specific substitutions may cause functional differentiation of marsupial and eutherian *SRY*.

To access the functional significance of amino acid differences in SRY homologs, DNA binding affinities of the several therian SRY HMG domains were investigated using MD analyses. The result indicated that the marsupial (wallaby,

opossum, and dunnart) HMG domains can bind DNA, but the binding affinity of these homologs seemed to be weaker than that of the human SRY HMG domain. Mutant versions of the human (M9I, F55I, Y69F, Q59K, and K68E) and wallaby (K64V) proteins were predicted to exhibit lower affinity for DNA than the wild-type proteins. The M9I mutant causes abnormal DNA bending (Murphy *et al*. 2001). The predicted binding affinities of Q59K, K68E, and K64V were much lower than that of M9I.

2.5 Discussion

2.5.1 The evolution of sex-determination systems

Male differentiation-related genes, *SOX3*, *SOX9*, *Ad4BP/SF-1,* and *DMRT1*, were conserved well in phylogenetically diverged animals (Fig. 2.2). All genes are transcriptional factors. *SOX3* is the ancestral gene of male determination *SRY*, but has possibility as a target of *SRY* (Graves 1998; Pask *et al*. 2000). *Ad4BP/SF-1* is essential for gonadogenesis and production of steroid hormones in both sexes. *SOX9* and *DMRT1* function in testis differentiation. Female differentiation-related genes *RSPO1* was found in only Jawed vertebrates. *RSPO1* is a non-secreted protein containing a thrombospondin type I motif, and might be a candidate of primary female-determination gene (Kamata *et al*. 2004; Perma *et al*. 2006), because RSPO1 enhances Wnt-4 signaling, which is essential for ovary differentiation (Tomizuka *et al*. 2008). Wilhelm

(2007) and Sekido and Lovell-Badge (2008) speculated that *RSPO1* might be suppressed by *SRY* in testes. *SRY* homologs were found in only Theria. These findings on *RSPO1* and *SRY* indicate that the primary genes in male or female determination are of more recent revolutionary origin or are more rapidly diversified than other, downstream genes in sex determination systems. The primary genes in male or female determination might be more recent revolutionary origin or more rapidly diversified than other, downstream genes in sex determination systems.

2.5.2 The origin of *SRY*

American and Australian marsupials diverged ~76 MYA (Meredith, Westerman, and Springer 2009), and I identified an *SRY* homolog in an American and several Australian marsupial. The present phylogenetic study of marsupial and eutherian *SRY* sequences strongly supported the hypothesis that *SRY* arose once in Theria (Foster *et al*. 1992; Wallis *et al*. 2007).

Phylogenetic relationships between *SRY* and *SOX3* have created controversy (Nagai 2001; Soullier *et al*. 1999; Katoh and Miyata 1999). Katoh and Miyata (1999) proposed a therian *SOX3* origin of *SRY* based on a heuristic approach and ML methods, but the phylogenetic topology they found was not reconstructed by Nagai (2001), by Soullier *et al*. (1999), or here in this study. The reasons for this inconsistency may be partly due to long branch attraction (LBA) because of rapid evolution of *SRY* on Y chromosomes and partly due to differences in the functional constraints on *SOX3* versus

those on *SRY*.

Following two different evolutionary scenarios (A and B) can be tested statistically. In hypothesis A, *SRY* originated in Theria (Fig. 2.8; Katoh and Miyata 1999); in hypothesis B, *SRY* originated in the ancestor of amniotes before the emergence of Theria (Fig. 2.8; this study, Nagai 2001 and Soullier *et al*. 1999). Based on the assumption that the evolutionary rate of branch is different between *SOX3* and *SRY*, hypothesis A and B were compared using the likelihood ratio test in PAML and aligned sequences of 81 amino acids. The log likelihood estimated for tree A (lnL = -720.393) was higher than that of tree B (lnL = -3.46±4.147), but no statistical difference was indicated by the P value of the KH normal test (pKH) or by the RELL bootstrap proportions (pRELL) (Kishino and Hasegawa 1989). The estimated log likelihood values were influenced by evolutionary models, such as substitution matrixes, but not statistically significant (JTT, WAG, or Dayhoff) or parameters. The analysis using a codon-based model of PAML and the nucleotide sequences (243 bp) also did not support one over the other. [Do you mean "a codon-based version of PAML and nucleotide sequences"? Please clarify.] The phylogenetic relationship between *SOX3* and *SRY* remains obscure, but the topology of tree B might be due to LBA of *SOX3* clusters.

Recently, a new hypothesis about the origin of eutherian *SRY* was proposed; Sato *et al*. (2009) suggested that *SRY* is a hybrid of *SOX3* and DiGeorge syndrome Critical Region gene 8 (*DGCR8*). In humans, exon 2 of *DGCR8* is highly similar to the 5' region of *SRY*. I investigated whether there was similarity between the

SRY and *DGCR8* genes within marsupials and other eutherians, as is the case in humans. But only the *SRY* and *DGCR8* gene pairs within primates were similar. The phylogeny of *SRY* and exon2 of *DGCR8* cannot be explained by a fusion between *DGCR8*-like sequence and *SRY* in primates. Currently, the possibility of convergent evolution of amino acids cannot be excluded.

2.5.3 Functional differentiation of *SRY*

To investigate the function of the amino acid substitution of HMG domain in marsupial and eutherian SRY, the DNA binding affinity of HMG was investigated using MD analyses. In the MD analysis, the HMG domains from marsupials (wallaby, opossum, and dunnart) did bind DNA, but the binding affinities of these HMG domains were lower than that of the HMG domain from humans. Mutant versions of the human HMG domain (M9I, F55I, Y69F, Q59K, and K68E) and of wallaby (K64V) were predicted to have lower affinity than the wild-type version. The M9I mutant results in protein malfunction because of abnormal DNA bending. The binding affinities of the Q59K, K68E and K64V mutants were predicted to be lower than that of M9I; these finding indicated that the lineage-specific substitutions were necessary for its DNA binding.

Functional differentiation of *SRY* homologs is considered to have occurred in two steps. The first step occurred in the common ancestor to marsupials and eutherians, as *SRY* was diverging from *SOX3*. The second step occurred as the marsupial and

eutherian lineages diverged; the marsupial and eutherian *SRY* differentiated independently. The rate of amino acid substitution in the ancestor (1.33×10^{-7} substitution per year) was significantly faster compared to the rate on each lineage leading to marsupial (5.94×10^{-8}) or eutherian *SRY* genes (1.73×10^{-7}), suggesting that substitution rate of *SRY* was higher than that of *SOX3* (Z test; $P<0.001$). In marsupials and eutherians, lineage-specific amino acid substitutions in *SRY* were found, and some of these substitutions were possibly important for the stability of DNA binding. Moreover, the DNA binding affinities of marsupial *SRY* proteins might be lower than those of eutherian *SRY* proteins.

A comparison of 5' regulatory regions of *SOX9* homologs supported the hypothesis that only eutherian genomes have the testis-specific enhancer that *SRY* and *Ad4BP/SF-1* can bind (Fig. 2.9); for example, the 5' regulatory regions of *SOX9* homologs in opossum and chicken genomes did not have the *SRY* and *Ad4BP/SF-1* TFBSs (Fig. 2.9). Therefore, eutherian-specific functional differentiation in the genetic cascade leading to gonadogenesis included the addition of an *SRY* binding enhancer to the *SOX9* gene, indicating that a novel functional relationship between *SRY* and *SOX9* emerged in eutherians.

2.5.4 Function of marsupial *SRY*

The mouse *SRY* gene is expressed at a critical stage of male determination in fetal testes

and in brains, and the *SRY* of other eutherians (humans, caws, sheep and pigs) is also expressed in testes and brains (Hacker *et al*. 1995, Lahr *et al*. 1995, Hanley *et al*. 2000, Mayer *et al*. 1998, Daneau *et al*. 1995, Payen *et al*. 1996, Parma, Pailhoux and Cotinot 1999). The 5' regulatory regions of *SRY* homologs are well conserved among many eutherians, excluding mice (Ross *et al*. 2008), but these eutherian regulatory regions are totally different from the regulatory regions of marsupial *SRY* homologs. The wallaby *SRY* is expressed in a broad range of tissues including, not only testis and brains, but also heart, mesonephros, kidney, lung, and others (Harry *et al*. 1995). Based on the expression pattern and the low DNA-binding affinity of marsupial *SRY* and the absence of a *SOX9* testis enhancer, the marsupial *SRY* gene might not contribute to testis determination during fetal development.

2.6 Conclusion

A comparison of marsupial and eutherian *SRY* homologs indicated that the process of functional differentiation in the *SRY* lineages might have occurred independently and differed in the marsupial and eutherian lineages. Lineage-specific changes in *SRY* and *SOX9* homologs were found, and these changes indicated that the primary-sex determination gene *SRY* and the sex-differentiation genes *SOX9* and *Ad4BP/SF-1* co-evolved in eutherians.

2.7 References

Assumpção, J. G., Benedetti, C. E., Maciel-Guerra, A. T., Guerra, G., Baptista, M. T., Scolfaro, M. R. and de Mello, M. P. (2002) 'Novel mutations affecting SRY DNA-binding activity: the HMG box N65H associated with 46,XY pure gonadal dysgenesis and the familial non-HMG box R30I associated with variable phenotypes.', *J Mol Med (Berl)* 80(12): 782-90.

Brennan, J. and Capel, B. (2004) 'One tissue, two fates: molecular genetic events that underlie testis versus ovary development.', *Nat Rev Genet* 5(7): 509-21.

Daneau, I., Houde, A., Ethier, J. F., Lussier, J. G. and Silversides, D. W. (1995) 'Bovine SRY gene locus: cloning and testicular expression.', *Biol Reprod* 52(3): 591-9.

Felsenstein, J. (1989) 'Mathematics vs. Evolution: Mathematical Evolutionary Theory.', *Science* 246(4932): 941-2.

Foster, J. W., Brennan, F. E., Hampikian, G. K., Goodfellow, P. N., Sinclair, A. H., Lovell-Badge, R., Selwood, L., Renfree, M. B., Cooper, D. W. and Graves, J. A. (1992) 'Evolution of sex determination and the Y chromosome: SRY-related sequences in marsupials.', *Nature* 359(6395): 531-3.

Gilson, M. K. and Zhou, H. X. (2007) 'Calculation of protein-ligand binding affinities.', *Annu Rev Biophys Biomol Struct* 36: 21-42.

Graves, J. A. (1998) 'Interactions between SRY and SOX genes in mammalian sex determination.', *Bioessays* 20(3): 264-9.

Hacker, A., Capel, B., Goodfellow, P. and Lovell-Badge, R. (1995) 'Expression of Sry, the mouse sex determining gene.', *Development* 121(6): 1603-14.

Hanley, N. A., Hagan, D. M., Clement-Jones, M., Ball, S. G., Strachan, T., Salas-Cortés, L., McElreavey, K., Lindsay, S., Robson, S., Bullen, P. *et al*. (2000) 'SRY, SOX9, and DAX1 expression patterns during human sex determination and gonadal development.', *Mech Dev* 91(1-2): 403-7.

Harry, J. L., Koopman, P., Brennan, F. E., Graves, J. A. and Renfree, M. B. (1995) 'Widespread expression of the testis-determining gene SRY in a marsupial.', *Nat Genet* 11(3): 347-9.

Kamata, T., Katsube, K., Michikawa, M., Yamada, M., Takada, S. and Mizusawa, H. (2004) 'R-spondin, a novel gene with thrombospondin type 1 domain, was expressed in the dorsal neural tube and affected in Wnts mutants.', *Biochim Biophys Acta* 1676(1): 51-62.

Katoh, K. and Miyata, T. (1999) 'A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian SOX-3 origin of the testis-determining gene SRY.', *FEBS Lett* 463(1-2): 129-32.

Kishino, H. and Hasegawa, M. (1989) 'Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea.', *J Mol Evol* 29(2): 170-9.

Lahr, G., Maxson, S. C., Mayer, A., Just, W., Pilgrim, C. and Reisert, I. (1995) 'Transcription of the Y chromosomal gene, Sry, in adult mouse brain.', *Brain Res Mol Brain Res* 33(1): 179-82.

Mayer, A., Lahr, G., Swaab, D. F., Pilgrim, C. and Reisert, I. (1998) 'The Y-chromosomal genes SRY and ZFY are transcribed in adult human brain.', *Neurogenetics* 1(4): 281-8.

Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S. and Dubchak, I. (2000) 'VISTA : visualizing global DNA sequence alignments of arbitrary length.', *Bioinformatics* 16(11): 1046-7.

Meredith, R. W., Westerman, M. and Springer, M. S. (2009) 'A phylogeny of Diprotodontia (Marsupialia) based on sequences for five nuclear genes.', *Mol Phylogenet Evol* 51(3): 554-71.

Murphy, E. C., Zhurkin, V. B., Louis, J. M., Cornilescu, G. and Clore, G. M. (2001) 'Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation.', *J Mol Biol* 312(3): 481-99.

Nagai, K. (2001) 'Molecular evolution of Sry and Sox gene.', *Gene* 270(1-2): 161-9.

Parma, P., Pailhoux, E. and Cotinot, C. (1999) 'Reverse transcription-polymerase chain reaction analysis of genes involved in gonadal differentiation in pigs.', *Biol Reprod* 61(3): 741-8.

Parma, P., Radi, O., Vidal, V., Chaboissier, M. C., Dellambra, E., Valentini, S., Guerra, L., Schedl, A. and Camerino, G. (2006) 'R-spondin1 is essential in sex determination, skin differentiation and malignancy.', *Nat Genet* 38(11): 1304-9.

Pask, A. J., Harry, J. L., Renfree, M. B. and Marshall Graves, J. A. (2000) 'Absence of SOX3 in the developing marsupial gonad is not consistent with a conserved role in mammalian sex determination.', *Genesis* 27(4): 145-52.

Payen, E., Pailhoux, E., Abou Merhi, R., Gianquinto, L., Kirszenbaum, M., Locatelli, A. and Cotinot, C. (1996) 'Characterization of ovine SRY transcript and developmental expression of genes involved in sexual differentiation.', *Int J Dev Biol* 40(3): 567-75.

Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) 'MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.', *Nucleic Acids Res* 23(23): 4878-84.

Ross, D. G., Bowles, J., Koopman, P. and Lehnert, S. (2008) 'New insights into SRY regulation through identification of 5' conserved sequences.', *BMC Mol Biol* 9: 85.

Rzhetsky, A. and Nei, M. (1992) 'Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference.', *J Mol Evol* 35(4): 367-75.

Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for reconstructing phylogenetic trees.', *Mol Biol Evol* 4(4): 406-25.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. and Smith, M. (1977) 'Nucleotide sequence of bacteriophage phi X174 DNA.', *Nature* 265(5596): 687-95.

Sato, Y., Shinka, T., Sakamoto, K., Ewis, A. A. and Nakahori, Y. (2010) 'The male-determining gene SRY is a hybrid of DGCR8 and SOX3, and is regulated by the transcription factor CP2.', *Mol Cell Biochem* 337(1-2): 267-75.

Sekido, R. and Lovell-Badge, R. (2008) 'Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer.', *Nature* 453(7197): 930-4.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., Foster, J. W., Frischauf, A. M., Lovell-Badge, R. and Goodfellow, P. N. (1990) 'A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif.', *Nature* 346(6281): 240-4.

Smith, C. A. and Sinclair, A. H. (2004) 'Sex determination: insights from the chicken.', *Bioessays* 26(2): 120-32.

Soullier, S., Jay, P., Poulat, F., Vanacker, J. M., Berta, P. and Laudet, V. (1999) 'Diversification pattern of the HMG and SOX family members during evolution.', *J Mol Evol* 48(5): 517-27.

Sourdis, J. and Nei, M. (1988) 'Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree.', *Mol Biol Evol* 5(3): 298-311.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) 'MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.', *Mol Biol Evol* 24(8): 1596-9.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997) 'The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.', *Nucleic Acids Res* 25(24): 4876-82.

Tomizuka, K., Horikoshi, K., Kitada, R., Sugawara, Y., Iba, Y., Kojima, A., Yoshitome, A., Yamawaki, K., Amagai, M., Inoue, A. *et al*. (2008) 'R-spondin1 plays an essential role in ovarian development through positively regulating Wnt-4 signaling.', *Hum Mol Genet* 17(9): 1278-91.

Wallis, M. C., Waters, P. D., Delbridge, M. L., Kirby, P. J., Pask, A. J., Grützner, F., Rens, W., Ferguson-Smith, M. A. and Graves, J. A. (2007) 'Sex determination in platypus and echidna: autosomal location of SOX3 confirms the absence of SRY from monotremes.', *Chromosome Res* 15(8): 949-59.

Wilhelm, D. (2007) 'R-spondin1--discovery of the long-missing, mammalian female-determining gene?', *Bioessays* 29(4): 314-8.

Yang, Z. (2007) 'PAML 4: phylogenetic analysis by maximum likelihood.', *Mol Biol Evol* 24(8): 1586-91.

2.8 Figure Legends

Figure 2.1 Schematic representation of therian diversification.

Gray triangles indicate animals in which an *SRY* ortholog has been identifies, *i.e*., tammar wallabies (*Macropus eugenii*; Order Diprotodontia; Family Macropodidae), stripe-faced dunnarts (*Sminthopsis macroura*: Order Dasyuromorphia), brushtail possums (*Trichosurus vulpecula*; Order Petauridae) and Northern brown bandicoot (*Isoodon macrourus*; Order Peramelemorphia) (Foster *et al*. 1992, Watson, Margan and Johnston 1998 and Eckery *et al*. 2002). The black triangle indicates the American marsupial in which an *SRY* ortholog was identified in this study. The time scale shows the divergence time of Theria. The last common ancestor of marsupials was ~76 MYA (Meredith, Westerman and Springer 2009). Didelphimorphia (opossums) diverged from Australidelphia ~73 MYA, the radiation of Australidelphia was ~63 MYA, and Diprotodontia diverged ~53 MYA (Meredith, Westerman and Springer 2009).

Figure 2.2 Sex determination-related genes in animals.

The left panel shows the cascade of sex determination in mice, a model of the eutherian sex-determination cascade. The right panel contains a list of sex determination-related gene homologs in animals.

Figure 2.3 An alignment of amino acid sequences from *SRY* genes.

Modo_Y (opossum *SRY*), Maeu_Y (wallaby *SRY*), Smmc_Y (stripe-faced dunnart *SRY*), Hosa_Y (human *SRY*). The blue bar represents the position of the HMG domain.

Figure 2.4 PCR products

The gene structure of *SRY* is based on one wallaby sequence inferred from a BAC derived from the Y chromosome. The gray bars represent the regions covered by the PCR products; the size and species-of-origin are indicated beside each bar.

Figure 2.5 Comparison of Y chromosomal regions surrounding *SRY* from opossums and wallabies.

Figure 2.6 Phylogeny of *SRY* genes

The tree was constructed using the NJ method and 58 amino acids. Inferred substitutions are shown on the tree. Four substitutions (M23L, R38Q, M66R, and K67E) were specific to the branch that *SRY* differentiated from *SOX3* in the ancestor of Theria. Twelve substitutions (G18D, K29Q, H31Q, A41Y, D42Q, L45M, D48E, I55F, D56E, K59Q, V64M, and E68K) were on the branch containing the eutherian *SRY* and

seven substitutions (D3S, G18S, D42S, A49N, R52Q, V64K, R75Q) were on the branch containing the marsupial *SRY*. The divergence time on the tree, 79 or 105 MYA coincides with the radiation of marsupials or eutherians from Figure 2.1, and 210-180 MYA is the divergence time of monotremes and Theria.

Figure 2.7 Comparison of HMG domain in *SRY* and *SOX* genes in eutherians and marsupials.

The amino acid sequence alignment of HMG domain is shown. The eutherian-specific substitutions are indicated in blue and marsupial-specific substitutions are indicated in green. The substitutions common to Theria are indicated in yellow. The dot at the top of alignment indicates an amino acid position that is critical to SRY function (magenta on the first line) or structure (multi-color on the second line).

Figure 2.8 The two hypothetical trees A and B used for the likelihood ratio test

Figure 2.9 The conservation of the Sox9 testis enhancers from humans, mice, opossums, and chickens.

The red boxes indicate *SRY* binding sites; the green boxes indicate *Ad4BP/SF-1* binding sites identified by Sekido and Lovell-Badge (2008).

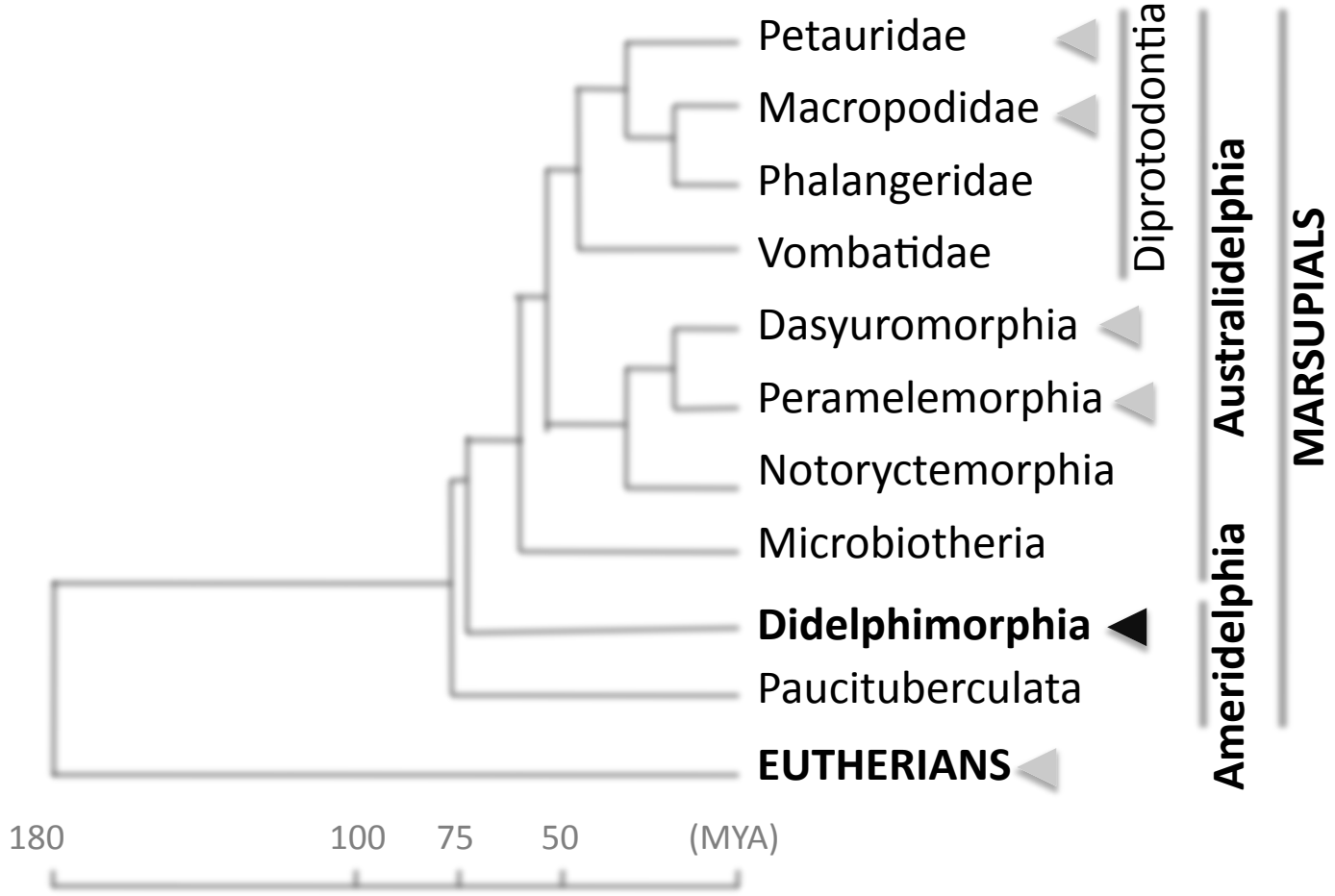
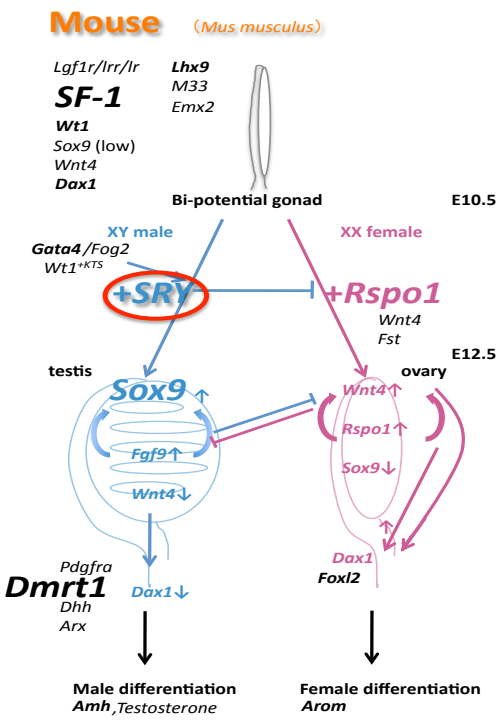


Figure 2.1
Schematic representation of therian diversification.



	SRY	Sox3	Sox9	SF-1	Dmrt1	Rspo1
Humans	◎ chr Y	◎ chr X	◎ chr 17	◎ chr 9	◎ chr 9	◎ chr 1
Chimpanzees	◎	◎	◎	◎	◎	◎
Orangutans	(-)	◎	◎	◎	◎	◎
Macques	◎	◎	◎	◎	◎ chr15	◎
Mouse	◎	◎	◎	◎ chr 2	◎ chr19	◎
Opossums	-	○ chr X	◎ chr2	◎	◎	◎
Wallabies	◎ chr Y	◎ chr X	○	◎	○	○
Platypuses	-	◎ chr 6	◎	◎ chr 4q	◎	◎ chr 16
Chickens	-	◎ chr 4	◎	◎ chr 17	◎	◎ chr 23
Alligators	-	◎	◎	◎	◎	◎
Turtles	-	◎	◎	◎	◎	◎
Lizards	-	○ scaffold 722	○ scaffold 336	○ scaffold 1639	○ scaffold 13	○ scaffold 28
Xenopus	-	◎	◎	◎	◎ chr1,2 DM-W chr3(w)	◎ Rspo2
Medaka	-	◎ sox2	◎ sox9a,b	◎ ftz-f1	◎ dmrt1, dmy	◎
Fugu	-	◎	◎	◎	◎ dmrt1-alpha	◎
Lampreys	-	○	◎ sox9, soxE1	○	○	-
Sea squirts	-	◎ soxB1	◎ soxE	◎ ftzf	◎	? Spo,1a, SCOspo
Lancelets	-	◎ soxB1	○ chrUn, ◎ soxE	○	◎ hypothetical	-
Sea urchins	-	◎ soxB1	◎ sox9	◎ ftz-f1	◎	? Spo, VSGP/Fspo, SCOspo
Fruit fries	(◎)	◎ soxNeuro	◎ sox100B chr3	◎ fushitarazu	◎ dsx, dmrt99B, 11E, 93B	? Fat-spo, Dfurin2
Nematodes	-	◎ sox-1,3	◎ sox-2	◎ nhr-25	◎ mab-3, dmd-4, 5	? Spo-1, unc-5

Figure 2.2
Sex determination-related genes in animals.

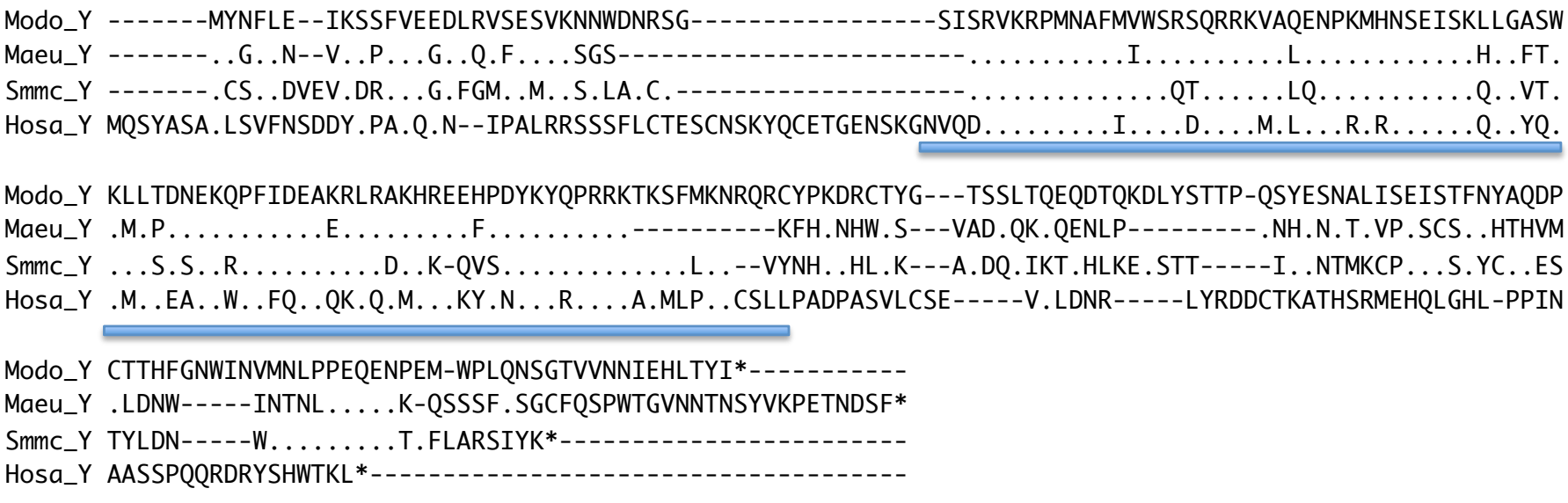


Figure 2.2
The alignment of amino acid sequences of *SRY* genes.

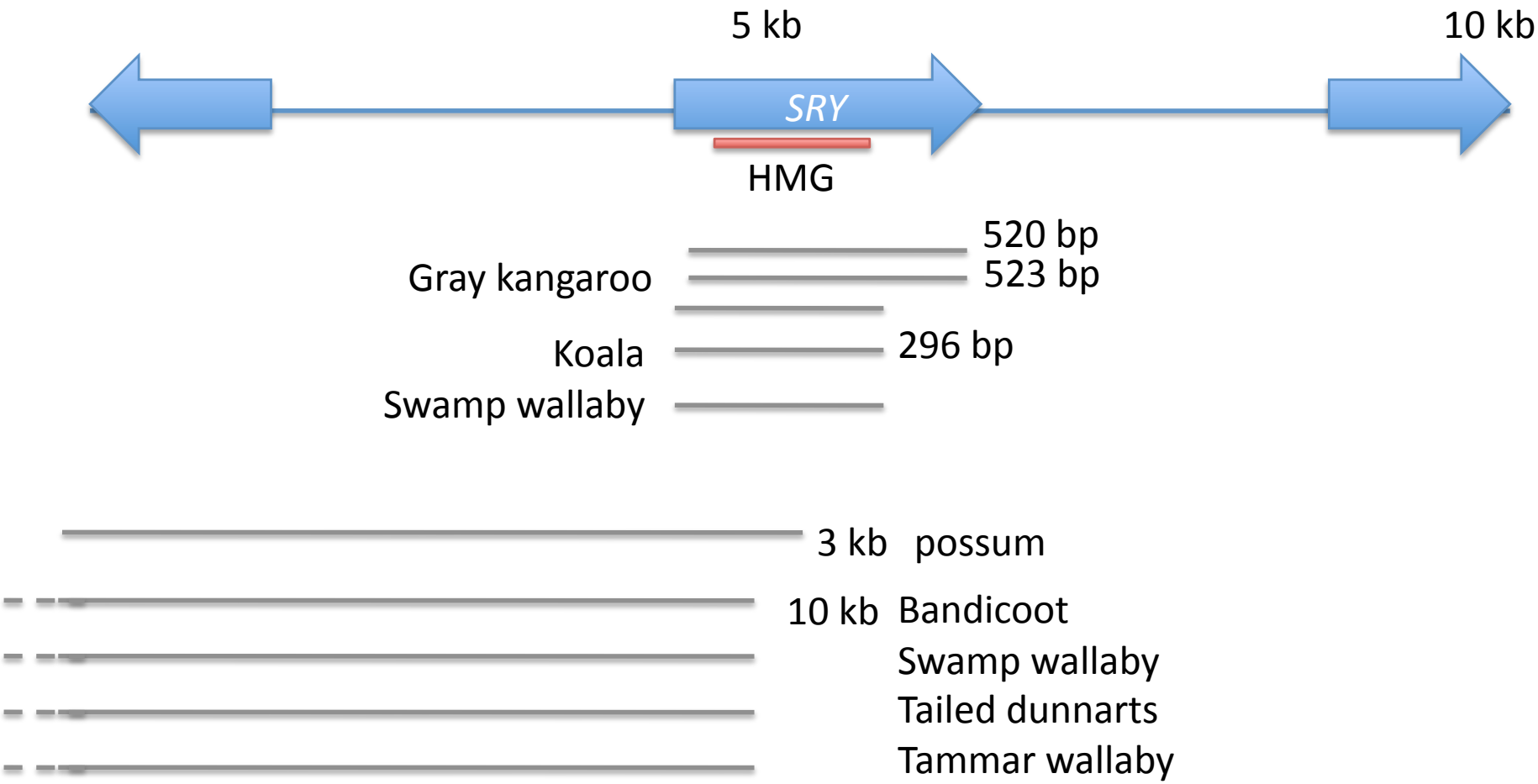


Figure 2.4
PCR products.

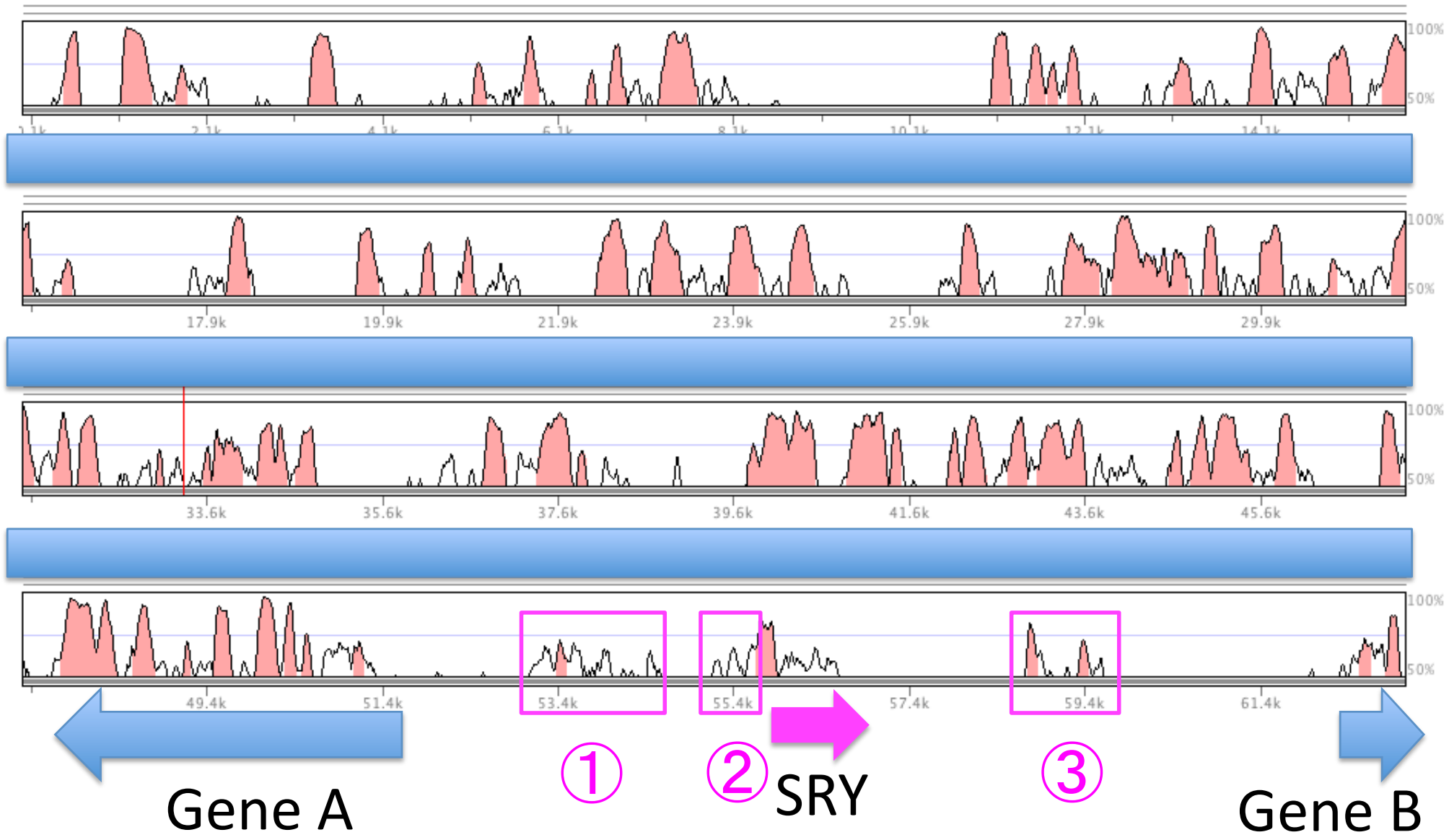


Figure 2.5

Comparison of Y chromosomal regions including *SRY* between opossums and wallabies.

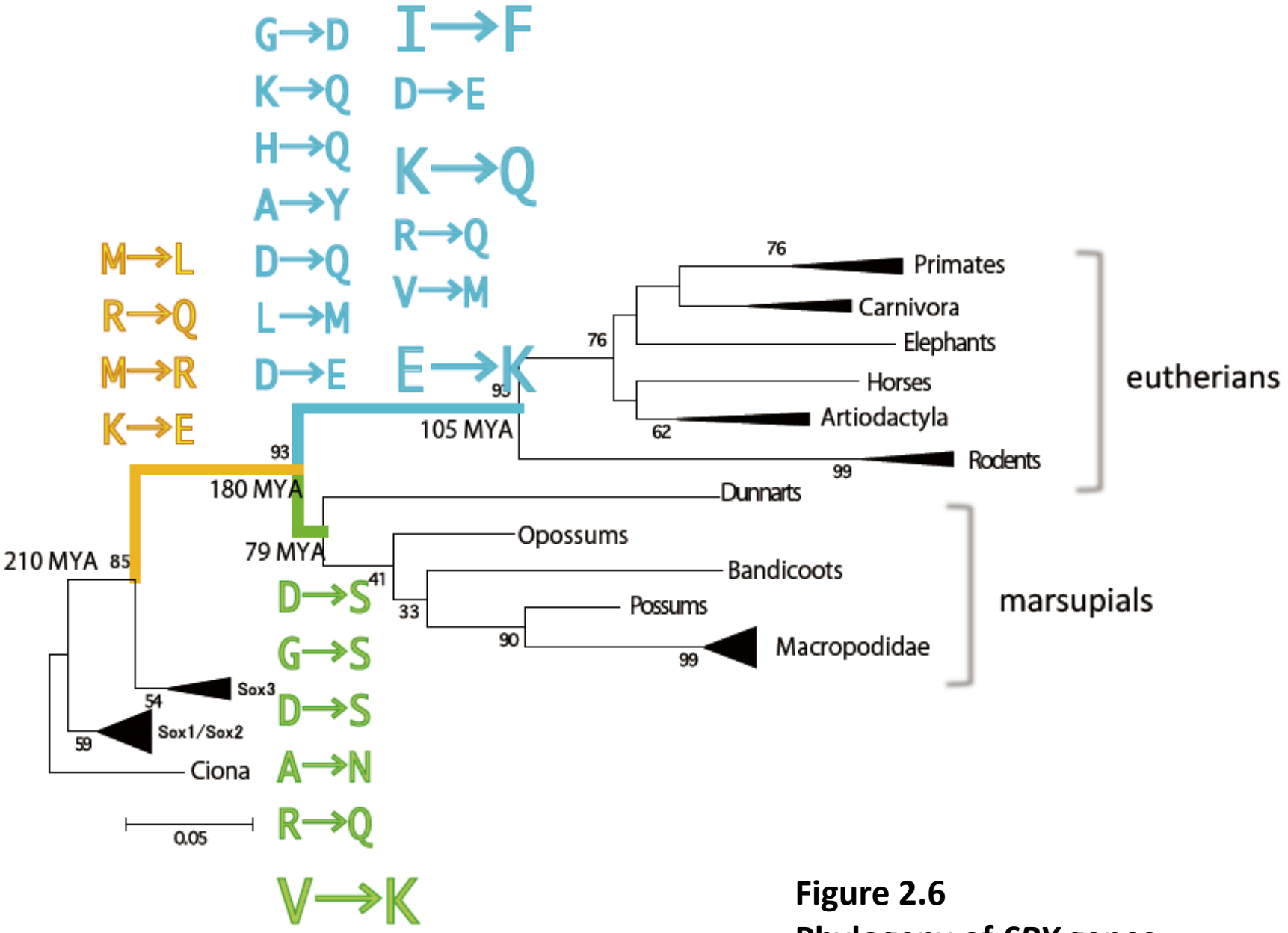


Figure 2.6
Phylogeny of *SRY* genes.

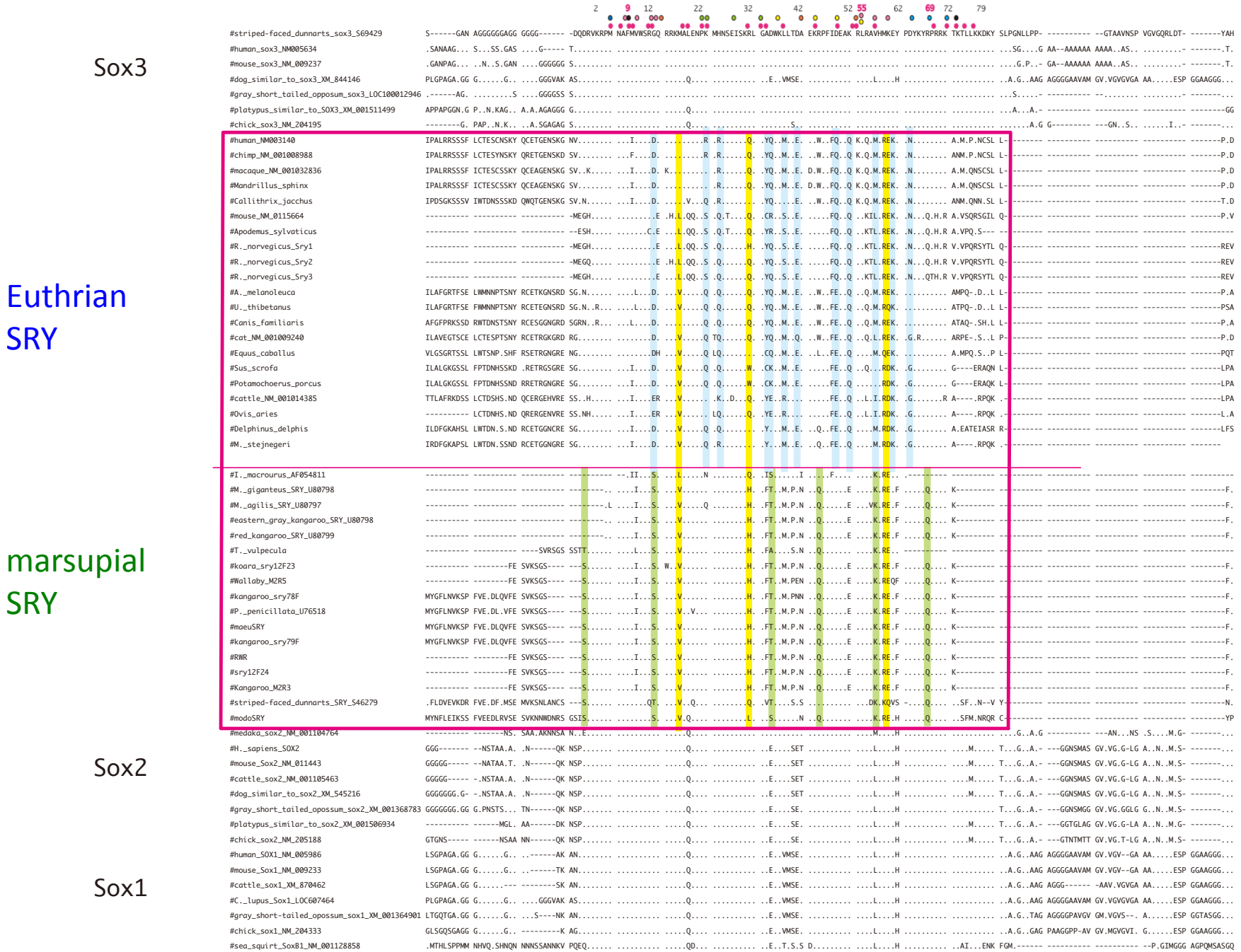


Figure 2.7

Comparison with HMG domain in *SRY* and *SOX* genes in eutherians and marsupials.

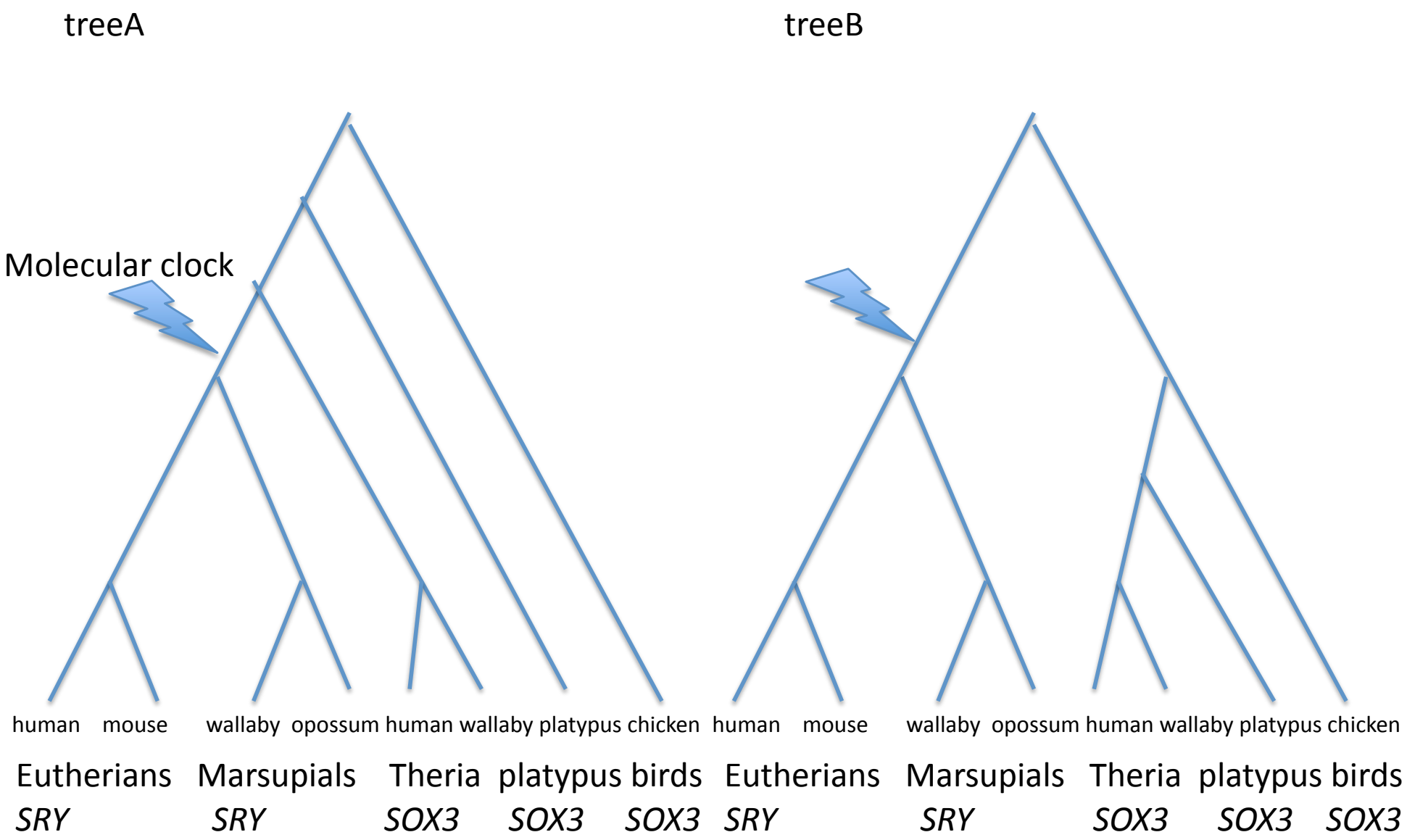


Figure 2.8
The topologies of tree A and B for likelihood ratio test

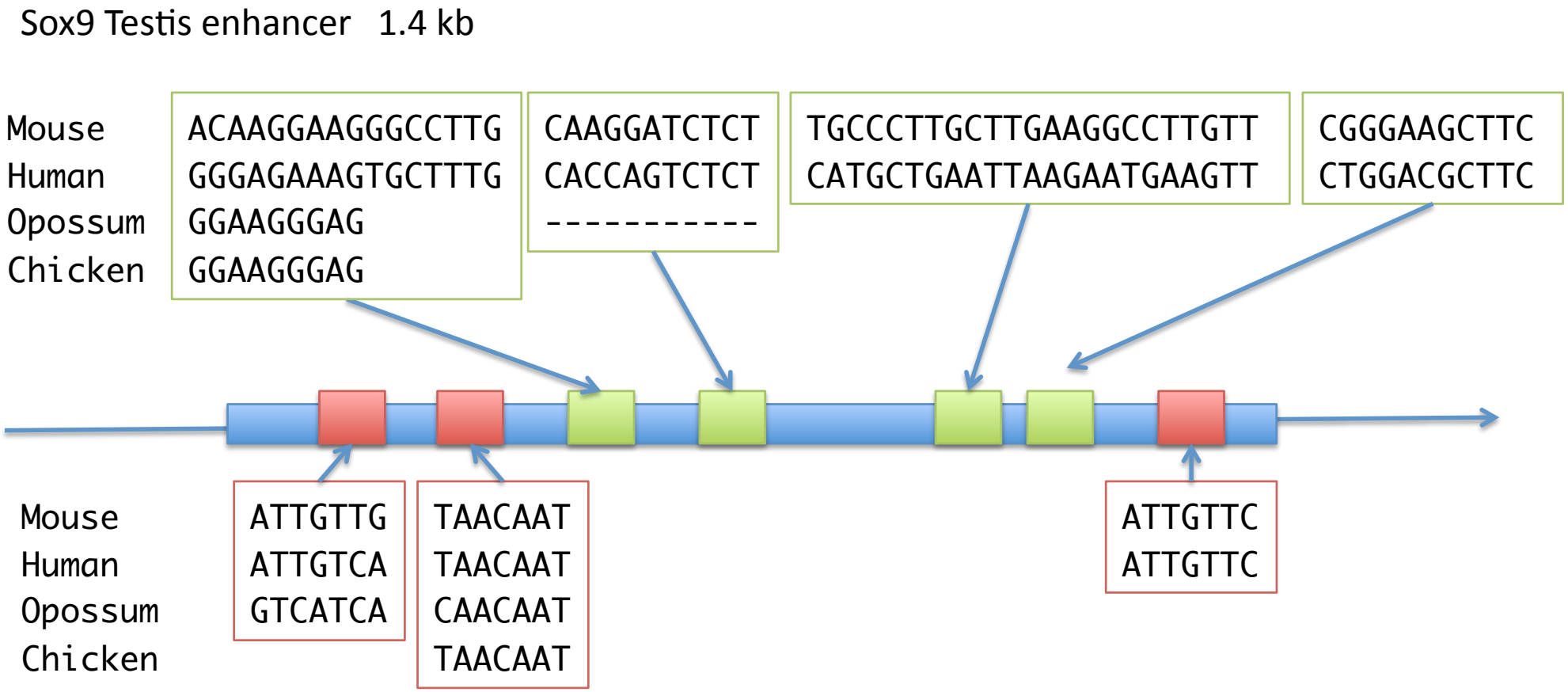


Figure 2.9
The conservation of Sox9 testis enhancer between humans, mice, opossums, and chickens.

Table 2.1
Accession numbers of nucleotide sequences used in chapter 2.

Species	**Gene**	**Accession No.**
striped-faced dunnart	*SOX3*	S69429
human	*SOX3*	NM005634
mouse	*SOX3*	NM_009237
dog	*SOX3*	XM_844146
opposum	*SOX3*	XM_001367288
platypus	*SOX3*	XM_001511499
chicken	*SOX3*	NM_204195
human	*SRY*	NM003140
chimpanzee	*SRY*	NM_001008988
macaque	*SRY*	NM_001032836
mandrillus sphinx	*SRY*	AF284330.2
marmoset	*SRY*	J527004
mouse	*SRY*	NM_011564
Apodemus sylvaticus	*SRY*	AB548702
Rattus norvegicus	*SRY1*	NM_012772
Rattus norvegicus	*SRY2*	AF275683
Rattus norvegicus	*SRY3*	AF275682
dog	*SRY*	AF107021.1
cat	*SRY*	NM_001009240
horse	*SRY*	NM_001081810.1
pig	*SRY*	GU991615
Potamochoerus porcus	*SRY*	FN186126.1
cattle	*SRY*	NM_001014385
Ovis aries	*SRY*	AY604733
Delphinus delphis	*SRY*	AB108522.2
Mesoplodon stejnegeri	*SRY*	AB108517
I. macrourus	*SRY*	AF054811
M. gugabteus	*SRY*	U80798
M. agilis	*SRY*	U80797
eastern gray kangaroo	*SRY*	U80798
T. vulpecula	*SRY*	U80799
P. penicillata	*SRY*	U76518
striped-faced dunnart	*SRY*	S46279
medaka	*SOX2*	NM_001104764
human	*SOX2*	NM_003106
mouse	*SOX2*	NM_011443
catle	*SOX2*	NM_001105463
dog	*SOX2*	XM_545216
opossum	*SOX2*	XM_001368783
platypus	*SOX2*	XM_001506934
chicken	*SOX2*	NM_205188
human	*SOX1*	NM_005986
mouse	*SOX1*	NM_009233
catle	*SOX1*	XM_870462
dog	*SOX1*	XM_844146
opossum	*SOX1*	XM_001364901
chicken	*SOX1*	NM_204333
sea squirt	*SOXB1*	NM01128858

Table 2.2
Primers used for PCR and sequencing.

Name	Primer sequence (5' → 3')	Specifity	Comments	Reference
MeSRY1_F	TTGAGTCCGTGAAAAGTGGGTC	SRY	PCR	Harry et al.1995
MeSRY1_R	TTGTGAATCTGCCACGCTTGTC	SRY	PCR	Harry et al.1995
MeSRY23_F	GCTATGTATGGCTTCTTGAATG	SRY	PCR	O' Neil et al. 1998
MeSRY2_R	CTGTCATTCGTTTCAGGTTTAAC	SRY	PCR	O' Neil et al. 1998
MeSRY3_R	AACTGTCATTCGTTTCAGGT	SRY	PCR	O' Neil et al, 1997
Modo_ATRY1	AT T TGT TGTGACTCT TGCCAT	ATRY	PCR	
Modo_ATRY1	GCCT T TACCT TCTGT TGCT T T	ATRY	PCR	
Modo_ATRX1	CAATAATGGATGAAAACAGCC	ATRX	PCR	
Modo_ATRX1	TGCCTGCT TCAAAAATCT TAC	ATRX	PCR	
M13_F	GTAAAACGACGGCCAG	M13	PCR	
M13_R	CAGGAAACAGCTATGAC	M13	PCR	
LA-F8	CTTCAGGGCTTTGCAGCACTTGAAGGAA	SRY genome	LA-PCR	
LA-R1	GACCATATCATAAAAGCATTCATAGGCCT	SRY genome	LA-PCR	
LA-F10	GTTCGAGACCTCAAGCCAAATGGAGC	SRY genome	LA-PCR	
LA-R2	CATAAAAGCATTCATAGGCCTCTTCAC	SRY genome	LA-PCR	
MeSwF1	GGAGCTAATATTCTGGTAAATGAGGAG	SRY genome	sequencing	
SmcrF1	GGATCTAATACTCTGGTAACTGAGGAG	SRY genome	sequencing	
BantiF1	GCTAATACTCTGGTAAATGAGGAG	SRY genome	sequencing	
MeSwBaF2	CGAGACCTCAAGCCAAATGGAGC	SRY genome	sequencing	
SmcrR1	GGTTATCTTTCCAACATTCAAAATGTTG	SRY genome	sequencing	
SwwaR1	GACCTCCCAACATTGGAAATG	SRY genome	sequencing	
SmcrR2	GGTTATCTTTCCAACATTC	SRY genome	sequencing	
MaeuR2	GTTAACCTCCCAACATTGG	SRY genome	sequencing	
MaeuR1	CACTTAAAGTTAACCTCCCAACATTGG	SRY genome	sequencing	
BantiR1	CTTAAAAGTTAACCTCCCAACACTGG	SRY genome	sequencing	
BantiR2	GTTAACCTCCCAACACTGG	SRY genome	sequencing	
SmcrF2	GTTCGGGACCTCAAGCCAAATGG	SRY genome	sequencing	
Posu102F1	GATTTGTAAATGCTTGAAG	SRY genome	sequencing	
Posu102F2	TTTGGAAATTTGTGGACAAC	SRY genome	sequencing	
Posu102F3	CTGTCTTAAAATTGGAGG	SRY genome	sequencing	
Posu102R1	CTTCAAGACATTTACAAATC	SRY genome	sequencing	
Posu102R2	CCATGAATCACTGTAATCTTTG	SRY genome	sequencing	
MSBF1	GACCACAGAAGAAAGCGTAAC	SRY genome	sequencing	
SF1	GACATTTATTAGACAAACTG	SRY genome	sequencing	
SF2	GAGACATTTATTAGACAAACTGC	SRY genome	sequencing	
MSBF2	CCAGATTCTGAAGGATTGAC	SRY genome	sequencing	
MSBF3	CTAAACTCAGTTATGAGAAAGG	SRY genome	sequencing	
MeSw3R1	GTAATTCCTCTACTTCATGTGGTCC	SRY genome	sequencing	
MeSw3R2	CCTCTACTTCATGTGGTCC	SRY genome	sequencing	
MeSw3R3	CTTTAATATAAGACTTATCCCTATC	SRY genome	sequencing	
SwBa3R1	CTTAAATATTGCTATCATTTTACCC	SRY genome	sequencing	
SwBa3R2	GTCCCTTCTGAATTTCCTAACT	SRY genome	sequencing	
Sm3R1	CATTGCCCACATCTGTAG	SRY genome	sequencing	
Sm3R2	CACTATTATCTTGTTATTCCAC	SRY genome	sequencing	
Sm3R3	GTCCCCTTCTAAATTTCCTG	SRY genome	sequencing	
Sm3R4	CACTTCTGGTTGATTCAATG	SRY genome	sequencing	
Swwa2R1	CCTCTTACGCAAACTCAAC	SRY genome	sequencing	
Banti2R1	CCTCTTACATAAACTCAAC	SRY genome	sequencing	
Maeu2R1	CCTCTTACACAAACTCAAC	SRY genome	sequencing	
Smcr2R1	CCACAAATTTCCAAATAAATTC	SRY genome	sequencing	
Smcr/102-F1	CAGTAAGGAGAGTATATAAG	SRY genome	sequencing	
MaBaSw-F1	CTGTCAGTGAGGAAACTATAG	SRY genome	sequencing	
P102-2R1	CCACAAATTTCCAAATAAATTC	SRY genome	sequencing	
P102-2R2	AACTTCAAGACATTTACAAATC	SRY genome	sequencing	
P102-2R3	CACAGACACACAAGAATTAT	SRY genome	sequencing	
P102-2R4	TTTATCATTAAATATCTTAAAATATG	SRY genome	sequencing	
P102-2F1	GAATTTATTTGGAAATTTGTGGAC	SRY genome	sequencing	
P102-2F2	TTTGGAGGGGTCAAGTTCCC	SRY genome	sequencing	
P102-2F3	TTTTAGTATGTTTTGGAGGG	SRY genome	sequencing	
MSB4F1	CTTAATAGTCCCATAGCATC	SRY genome	sequencing	
SB5R1	CACTGGGTTACTGTGGTC	SRY genome	sequencing	
MS5R1	CCTACTGACATTTGACAG	SRY genome	sequencing	
MS5R2	CGGGTATGGGTAGCAAG	SRY genome	sequencing	
Sc5R1	GTAAATTCTCATTCCAATAGCTG	SRY genome	sequencing	
Sc5R2	GAGTACTTTGCAGTTATGAC	SRY genome	sequencing	
Sw4F1	CTTCCTGTGTGTGCCAG	SRY genome	sequencing	
Sw5F1	CTTGTAGAGATAGCTTCCTC	SRY genome	sequencing	
MS4F1	GAGTATAGTTTCCCTTTGGC	SRY genome	sequencing	
Sc4F1	GGATATTATCCATCCCATG	SRY genome	sequencing	
BSS-102-F1	GTGAAGAGGCCTATGAATGCTTTTATGAT	SRY genome	sequencing	
BSS-102-F2	GTGAAGAGGCCTATGAATGC	SRY genome	sequencing	
L102Scr1	GTCCGTGAGAAGTGGATCAAGCAGTAC	SRY genome	sequencing	
L102Scr2	GTCCGTGAGAAGTGGATCAAGCAG	SRY genome	sequencing	
L102Scr3	CCGTGAGAAGTGGATCAAGC	SRY genome	sequencing	
LA102Scr4	CGTGAGAAGTGGATCAAGCAG	SRY genome	sequencing	
LBS1	GTGGATCAAGTAGAGTGAAGAGGCCTAT	SRY genome	sequencing	
LMBS1	CAAGTAGAGTGAAGAGGCCTATGAATGC	SRY genome	sequencing	
LMBS2	GTAGAGTGAAGAGGCCTATG	SRY genome	sequencing	
LMeu1	GTGGGTCAAGTAGAGTGAAGAGGCCTA	SRY genome	sequencing	

Table 2.3
The pair of proteins and DNA used for MD analysis.

#	SRY protein type	species	DNA binding sites derived from genes	species	comments
1	wild type	human	Amh	human	positive control
2	mutant F55I	human	Amh	human	essential function in humans
3	mutant Y69F	human	Amh	human	essential function in humans
4	wild type	wallaby	Amh	human	for specuration of marsupial SRY proteins
5	wild type	oppoaum	Amh	human	for specuration of marsupial SRY proteins
6	wild type	dunnart	Amh	human	for specuration of marsupial SRY proteins
7	mutant Q59K	human	Amh	human	conserved in eutherians
8	mutant K68E	human	Amh	human	conserved in eutherians
9	wild type	wallaby	SRY	wallaby	for specuration of marsupial SRY binding sequences
10	wild type	oppoaum	SRY	oppoaum	for specuration of marsupial SRY binding sequences
11	mutant K64V	wallaby	SRY	wallaby	conserved in marsupials
12	mutant M9I	human	Amh	human	negative control: sex reversal

Table 2.4
TFBS list.

BLOCK	Name	TF
1	V$LHXF	Lim homeodomain factors
1	V$HOXF	Paralog hox genes 1-8 from the four hox clusters A, B, C, D
1	V$BRNF	Brn POU domain factors
1	V$CART	Cart-1 (cartilage homeoprotein 1)
1	V$OCT1	Octamer binding protein
1	V$SATB	Special AT-rich sequence binding protein
1	V$EVI1	EVI1-myleoid transforming protein
1	V$MEF2	MEF2, myocyte-specific enhancer binding factor
1	V$FKHD	Fork head domain factors
1	V$CLOX	CLOX and CLOX homology (CDP) factors
1	V$CAAT	CCAAT binding factors
1	V$DMRT	DM domain-containing transcription factors
1	V$CREB	cAMP-responsive element binding proteins
1	V$PARF	PAR/bZIP family
1	V$STAT	Signal transducer and activator of transcription
1	V$BCL6	POZ domain zinc finger expressed in B-Cells
1	O$VTBP	Vertebrate TATA binding protein factor
1	V$HOMF	Homeodomain transcription factors
1	V$ETSF	Human and murine ETS1 factors
1	V$AP1R	MAF and AP1 related factors
1	V$CHRF	Cell cycle regulators: Cell cycle homology element
1	V$ARID	AT rich interactive domain factor
1	V$CEBP	Ccaat/Enhancer Binding Protein
1	V$SORY	SOX/SRY-sex/testis determinig and related HMG box factors
1	O$PTBP	Plant TATA binding protein factor
2	V$HOXF	Paralog hox genes 1-8 from the four hox clusters A, B, C, D
2	V$LHXF	Lim homeodomain factors
2	V$SORY	SOX/SRY-sex/testis determinig and related HMG box factors
2	V$BRNF	Brn POU domain factors
2	V$HNF1	Hepatic Nuclear Factor 1
2	V$CART	Cart-1 (cartilage homeoprotein 1)
2	V$CDXF	Vertebrate caudal related homeodomain protein
2	V$CAAT	CCAAT binding factors
2	V$CLOX	CTCF and BORIS gene family, transcriptional regulators with 11 highly conserved zinc finger domains
2	V$ZBPF	Zfx and Zfy - transcription factors implicated in mammalian sex determination
2	V$KLFS	Krueppel like transcription factors
2	V$SP1F	GC-Box factors SP1/GC
2	V$MAZF	Myc associated zinc fingers
2	V$SREB	Sterol regulatory element binding proteins
3	V$EGRF	EGR/nerve growth factor induced protein C & related factors
3	V$ABDB	Abdominal-B type homeodomain transcription factors
3	V$SNAP	snRNA-activating protein complex
3	V$HOXF	Paralog hox genes 1-8 from the four hox clusters A, B, C, D
3	V$HAND	Twist subfamily of class B bHLH transcription factors
3	V$NEUR	NeuroD, Beta2, HLH domain
3	V$RP58	RP58 (ZFP238) zinc finger protein
3	V$DMRT	DM domain-containing transcription factors
3	O$VTBP	Vertebrate TATA binding protein factor
3	V$BRN5	Brn-5 POU domain factors
3	V$OCT1	Octamer binding protein
3	V$ARID	AT rich interactive domain factor
3	V$BRNF	Brn POU domain factors
3	V$SATB	Special AT-rich sequence binding protein
3	V$FKHD	Fork head domain factors
3	V$MYT1	MYT1 C2HC zinc finger protein
3	V$CDXF	Vertebrate caudal related homeodomain protein
3	V$CART	Cart-1 (cartilage homeoprotein 1)
3	V$HBOX	Homeobox transcription factors
3	V$DLXF	Distal-less homeodomain transcription factors
3	V$NKXH	NKX homeodomain factors

Chapter 3

The differentiation of sex chromosomes in Theria

3.1 Abstract

Mammalian sex chromosomes originated from a pair of autosomes, and homologous genes on sex chromosomes (gametologs) differentiated as a result of reduction in recombination between proto-sex chromosomes. In eutherians, this differentiation took place in a stepwise fashion and generated "evolutionary strata" on the X chromosome. It was believed that strata 1 and 2 (corresponding to the first two steps, respectively) emerged in the ancestor of Theria (eutherians and marsupials). However, marsupial sex chromosomes have not been investigated for evidence of such strata. In this study, seven opossum orthologs of eutherian gametologs were identified. among them, five pairs of these gametologs (*SOX3/SRY, RBMX/Y, RPS4X/Y, HSFX/Y*, *XKRX/Y*) reside in strata 1 and two pairs (*SMCX/Y*, and *UBE1X/Y*) reside in strata 2 of the human X chromosome. However, phylogenetic analysis that estimated the divergence time of these gametologs (including *ATRX*, which had been found only in marsupials) revealed that they had differentiated simultaneously in the common ancestor to eutherians and marsupials. Evidence of gene conversion was observed at the 3' end of *SMCX/Y* and *UBE1X/Y* in eutherians, not in marsupials. Moreover, to know the extent of functional

constraint on gametologs, the ratios of nonsynonymous to synonymous substitutions on the branches leading to each gametolog was examined. Based on this analysis, *RBMY* and *SRY* were less functionally constrained than *RBMX*, *SOX3*, and orthologs in platypus and chicken. In contrast, *HSFY* was significantly more constrained than *HSFX*, and *HSFY* was very similar to the platypus ortholog. Based on our findings, we concluded that 1) at least eight genes differentiated simultaneously in the common ancestor to therians, but gene conversion in eutherians reduced the nucleotide divergence between some allelic gametologs, which resulted in the previous misclassification of these genes into stratum 2, and 2) some Y gametologs gained a new function (e.g., in sex determination or spermatogenesis), whereas *HSFY* might have maintained its ancestral function.

3.2 Introduction

Biological sex, female or male, is determined genetically by sex chromosomes in most organisms. Sex chromosomes have been found in diverse groups, mammals, birds, fishes, insects, plants, and fungi, and probably emerged independently multiple times in different lineages. In general, pairs of sex chromosomes are thought to have evolved from pairs of autosomes (proto-sex chromosomes) (Muller 1914; Ohno 1967).

In species with chromosomally determined sex, a primary sex-determination gene must be located on one of the sex chromosomes. Theoretically, it is thought that the

emergence of a primary sex-determination gene is accompanied by reduction of recombination (Ohno 1967; Nei 1969; B. Charlesworth and D. Charlesworth 2000; D. Charlesworth, B. Charlesworth, Marais 2005; Graves 2006 as a review). Recombination between sex chromosomes is reduced relatively compared with autosomal homologous recombination in a variety of organisms, including humans, mice, cats, chickens, dioecious plants, and smut fungi (Lahn and Page 1999; Sandstedt and Tucker 2004; Handley, Ceplitis and Ellegren 2004; Nicolas *et al*. 2005; Bergero *et al*. 2007; Pearks Wilkerson *et al*. 2008; Nam and Ellegren 2008; Votintseva and Filatov 2009).

Members of the class Mammalia, which comprises Monotremes, Metatheria (marsupials), and Eutheria (eutherians; placental mammals), have XY sex chromosomes (Painter 1923; Rens *et al*. 2004). In Monotremes, which diverged early in mammalian evolution, the origin of the sex chromosomes, is different from that in Theria (eutherians and marsupials) (Veyrunes *et al*. 2008). In the common ancestor of Theria, the XY chromosomes were derived from a pair of autosomes syntenic to chromosome 6 in the platypus (Wallis *et al*. 2007; Veyrunes *et al*. 2008).

The divergence of X or Y-linked genes was mediated by stepwise suppression of recombination (Lahn and Page, 1999). In the "evolutionary strata" hypothesis homologous X-Y gene pairs (gametologs) in humans are classified into four distinct categories based on the extent of synonymous nucleotide substitutions (K_S) between gametologs and the position on X chromosomes (Lahn and Page 1999; Skaletsky *et al*. 2003; Sandstedt and Tucker 2004). Lahn and Page (1999) estimated that stratum 1, which includes *SOX3/SRY,* differentiated before the divergence of Monotremes, and

stratum 2 differentiated before the therian divergence. Skaletsky *et al*. (2003) suggested that stratum 3 emerged before the eutherian radiation and stratum 4 formed after the divergence between prosimian and simian primates (Lahn and Page 1999; Skaletsky *et al*. 2003; Iwase *et al*. 2003). In Monotremes, *SRY* is absent and *SOX3* is located on an autosome (e.g. the platypus chromosome 6) (Wallis *et al*. 2007); therefore, gametologs in strata 1 and 2 may have diverged in the stem lineage of Theria. Examining marsupial gametologs is essential to understand what occurred at the early stages of the evolution of mammalian sex chromosomes, but they have not yet been investigated.

In this study, to examine the therian sex chromosomal differentiation, the opossum genome was searched for orthologs to human gametolog pairs. Particular attention was paid to the search for marsupial orthologs of the genes in strata 1 and 2. Phylogenetic relationships of gametologs in marsupials and eutherians analyses were investigated. Finally, we examined how functional divergence of X or Y gametologs occurred by assessing the extent of functional constraint.

3.3 Materials and Methods

3.3.1 Nucleotide sequences used in this study

Nucleotide sequence data and corresponding annotated gene information were obtained from NCBI (http://www.ncbi.nlm.nih.gov/) and Ensembl databases (release 62;

http://uswest.ensembl.org/index.html). The genomic or transcriptional sequences of six humans (*Homo sapiens*) and six other mammalian species were used (Table 3.1). The other six mammalian species were the gray short-tailed opossum (*Monodelphis domestica*) and tammar wallaby (*Macropus eugenii*), which represented the represent marsupials and mouse (*Mus musculus*), dog (*Canis familiaris*), cat (*Felis catus*), and cow (*Bos taurus*), which represented eutherians.

3.3.2 Gametologs

The following 32 pairs of human or New World monkey gametologs were compiled based on previous studies (Table 3.1). Of the 32 pairs, 19 *(SOX3/SRY, RBMX/Y, RPS4X/Y, SMCX/Y, UTX/Y, CASK/CASKP, DBX/Y, DFFRX/Y, ZFX/Y, EIF1AX/Y, TB4X/Y, AMELX/Y, KAL1/KALP, STS/STSP, PRKX/Y, ARSE/ARSEP, ARSD/ARSDP, GYG2/GYG2P*, and *UBE1X/Y*) were taken from Lahn and Page (1999). *UBE1Y* of *UBE1X/Y* was not found in humans, but it was found in New World monkeys, such as squirrel monkeys and marmosets; therefore, *UBE1X/Y* was included in this study. Another 10 pairs (*ARSF/ARSFP, ADLICAN/ADLICANP, NLGN4X/NLGN4Y, VCX/Y, TBL1X/TBL1Y, OA1/OA1P, APXL/APXLP, OFD1/OFD1P2, CXorf15/CYorf15A,B,* and *BCoR/BCoRP*) were from Skaletsky *et al*. (2003), and two gametolog pairs (*HSFX/Y*, *TSPX/Y*) were from Ross *et al*. (2005). Moreover, another pair (*XKRX/Y*), which was identified in Calenda *et al*. (2006) and Bhowmick, Satta and Takahata (2007), was also used. Homology searches for the gametologs were performed using the BLAST

program and genomic sequences from the seven mammalian species (including human sequences) as queries. Homologous sequences showing more than 70% similarity to a query sequence were used in subsequent analyses. To confirm the syntenic position and orthology of X gametologs, sequences adjacent to the homologs were examined. Due to difficulty in confirming orthology on the Y chromosome by synteny , sequences similar to Y gametologs were regarded as orthologs of the human genes. The gene nomenclature abbreviations used for the human gametologs was also used for putative homologs from opossums and other mammals.

3.3.3 Phylogenetic and molecular evolutionary analyses

The sequences recovered from the BLAST searches were aligned using ClustalX software (Thompson *et al*. 1997) and subsequent manual corrections were done. Using these alignments, the number of synonymous nucleotide substitutions per synonymous site (K_S) or the proportion of synonymous nucleotide differences per synonymous site (P_S) was calculated using the modified Nei-Gojobori method and an assumed transition/transversion bias of R=1 with the MEGA5.03 program (Tamura *et al.* 2007). The multiple-hit corrections were performed by means of the Jukes-Cantor model. The number of K_S was also estimated using the Li-Wu-Luo method (Li, Wu, Luo 1985), which uses the Kimura 2 parameter model for the correction, but the K_S value values of most pairs were similar to K_S values by Jukes-Cantor model. Phylogenetic trees were constructed with three methods available in the MEGA5.03 program (Tamura *et al.*

2007): the neighbor-joining (NJ; Saitou and Nei 1987), maximum likelihood (ML; Kishino and Hasegawa 1989), and maximum parsimony methods (MP; Sourdis and Nei 1988). The reliability of the trees was assessed using bootstrap resampling with 1000 replications.

3.3.4 Estimation of functional constraints on gametologs

The ratio of nonsynonymous nucleotide divergence (K_A) to K_S was used as an indicator of functional constraint. The K_A value, like the Ks value, was estimated using the modified Nei-Gojobori method with the Jukes-Cantor correction. To examine functional constraint on a branch leading to an X or Y gametolog, a tree of therian X and Y gametologs that included sequence from an outgroup (i.e. platypus or chicken) was constructed. The branch lengths on trees of K_A or K_S were estimated, and the K_A /K_S ratio for each branch was calculated. The probability of rejecting the null hypothesis of strict-neutrality (H_0: $K_A/K_S = 1$) was also calculated.

3.3.5 Detection of gene conversion

To detect gene conversion, the two-sample Runs Test (Takahata 1994) was used. Using methods developed by Takahata (1994), phylogenetically informative sites in the alignment of gametolog sequences from both eutherians and marsupials were examined. These alignments were assessed using the global test in GENECONV program version

1.81 (Sawyer 1999). The heterogeneity in nucleotide divergence along homologous sequences in these alignments was examined using the window analysis function (window size = 500 bp, no overlaps) in DnaSPv5(Librado and Rozas 2009).

3.3.6 Estimation of divergence time of gametologs

The oldest Theria fossil is reported from ~167 million years ago (MYA) (Flynn *et al*. 1999), but the divergence of eutherians and marsupials began 148–190 MYA based on the molecular clock approach and data from extant Theria (Kumar and Hedges 1998; Woodburne, Rich and Springer 2003; van Rheede *et al*. 2006); therefore, the timing of the initial divergence between eutherians and marsupials remains somewhat obscure. Consequently, both 148 and 190 million years were used for the estimate to calibrate a molecular clock.

When the opossum and human orthologs of autosomal genes are compared, the average of K_S is estimated to be 1.02 (0.76–1.44; Goodstadt *et al*. 2007). Therefore, the average synonymous substitution rate (μ) for autosomal genes was estimated as 2.68×10^{-9}–3.45×10^{-9}/site/year/lineage for the lineages leading to opossums and humans. Nucleotide sequences at silent sites (synonymous or non-coding sites) on X and Y chromosomes should evolve at different rates, because the mutation rate is higher in the male germ line than the female germ line (Miyata *et al.* 1987). Moreover, 2/3 of inherited X chromosomes are transmitted through a female, and only a third are transmitted through a male. Assuming a 1:1 sex ratio, mutation rate of X-linked genes

may be estimated as the sum of two-thirds of the female mutation rate and one-third of the male rate. In contrast, Y chromosome are only transmitted through the male germ line and mutations of Y-linked genes, therefore, occur at the male rate. The majority of previous studies indicate that the male mutation rate in most eutherian species is two times higher than the female rate (male mutation rate/female mutation rate (α) = ~2; Makova, Yang and Chiaromonte 2004; Lindblad-Toh *et al* 2005; Elango *et al*. 2009), but some estimates indicate that the male rate is six-times higher in some species (α = ~6; Makova and Li 2002; Taylor *et al*. 2005). Here α of 2 was used. If 2.68×10^{-9}–3.45×10^{-9} of nucleotide substitution rate (μ) for autosomal sequences was used, estimates of μ of X or Y chromosomal sequences were $2.38 \times 10^{-9} – 3.07 \times 10^{-9}$ or $3.57 \times 10^{-9} – 4.60 \times 10^{-9}$, respectively, because the autosomal substitution rate is the mean of the female and male substitution rates. Then substitution rate for gametologs is $5.95 \times 10^{-9} – 7.67 \times 10^{-9}$.

3.4 Results

3.4.1 Divergence of marsupial gametologs

Based on K_S values of human gametologs, previous studies have suggested that marsupial X chromosomes would have strata 1 and 2 (Lahn and Page 1999; Skaletsky *et al*. 2002). Here, 32 pair of human gametologs including the New World monkey

sequences were compiled from several studies (Lahn and Page 1999; Skaletsky *et al*. 2002; Ross *et al*. 2005; Calenda *et al* 2006; Bhowmick, Satta and Takahata 2007). Among the 32 pairs, five are in stratum 1, three are in stratum 2, 11 are in stratum 3, and 13 are in stratum 4. For simplicity and clarity, we did not consider the five-strata hypothesis proposed by Ross *et al*. (2005), in which stratum 4 was subdivided into strata 4 and 5. We searched for opossum orthologs of the 32 human gametologs and examined whether or not the same strata were observed on the opossum X chromosome. The opossum genome had orthologs of 30 gametolog pairs; seven of the 30 opossum orthologs were located on the sex chromosomes, 7 on chromosome 4, and 16 on chromosome 7 (Fig. 3.1). Opossum orthologs of two gametolog pairs in strata 2 and 4 of the human X chromosome (*TSPX/Y* and *VCX/Y*, respectively) were not found in the opossum genome. The 23 opossum genes that were found on opossum autosomes, chromosomes 4 and 7, were orthologous to human genes in strata 3 and 4 (Fig. 3.1). Of the seven sex-lined opossum orthologs, five genes (*SOX3, RBMX, RPS4X*, *HSFX* and *XKRX*) were homologous to human genes in stratum 1 and two (*SMCX* and *UBE1X*) were homologous to human genes stratum 2. Among the seven sex-linked opossum genes identified, six had Y gametologs, but one, *XKRX*, did not in the opossum genome.

The locations of the seven genes on the long arm of the opossum X chromosome differed from those of the human orthologs (Fig. 3.1). In particular, the opossum *UBE1X* and *SMCX* were located on the distal end of the long arm and were sandwiched between *SOX3* and *RPS4X* (Fig. 3.1), but the human *UBE1X* and *SMCX* genes are located at the proximal end of the short arm and the other five human genes

are on the long arm.

To elucidate the early stages of therian sex chromosomal evolution and differentiation, we focused on the gametologs that were found in both opossum and human X chromosomes (Table. 3.2). We estimated K_S values for 13 gametolog pairs (7 human pairs and 6 opossum pairs) and compared the values between marsupials and eutherians (Table. 3.2). The average K_S values for the eutherian gametolog pairs were 2.45 ± 0.31 for *HSFX/Y,* 1.15 ± 0.23 for *SOX3/SRY*, 0.88 ± 0.11 for *RBMX/Y,* 1.47 ± 0.13 for *XKRX/Y*, 1.20 ± 0.31 for *RPS4X/Y*, 0.70 ± 0.19 for *UBE1X/Y*, and 0.64 ± 0.12 for *SMCX/Y* (Table. 3.2). According to Lahn and Page (1999), the eutherian *SMCX* and *UBE1X* genes are in stratum 2, and the recombination rate between the X and Y alleles of these gene slowed after the alleles of the other five gametolog pairs had diverged due to reduced recombination rates. In opossums, however, K_S values for *SMCX/Y* and *UBE1X/Y* were 2.05 ± 0.042 and 1.47 ± 0.044, respectively, more than twice the corresponding eutherian values ($P < 0.001$, Table. 3.2). The P_S (Table 3.2) or d_S values calculated using the Li-Wu-Luo method also differed significantly between marsupials and eutherians ($P < 0.001$). For example, the d_S of *SMCX/Y* and *UBE1X/Y* were 3.45 ± 0.054 and 1.34 ± 0.042, respectively, in opossums and 0.83 ± 0.23 and 0.75 ± 0.14, respectively, in eutherians. However, K_S, P_S, and d_S values for the other gametolog pairs in opossum were not significantly different from those for the eutherian orthologs ($P > 0.05$). Furthermore, when the opossum *SMCX/Y* and *UBE1X/Y* were compared with the eutherian genes in the stratum 1, the K_S, P_S, and d_S values were not significantly different ($P > 0.05$). Based on these observations, the marsupial *SMCX/Y* and *UBE1X/Y*

gametolog pairs diverged at the same time when the eutherian gametolog pairs in stratum 1 diverged. Evidentially, the opossum gametolog pairs did not differentiate in two phases, as did their eutherian orthologs, and suppression of recombination probably occurred only once in marsupials.

3.4.2 Phylogenetic analyses of gametologs in Theria

To estimate the relative timing of differentiation of seven gametolog pairs, phylogenetic analyses of orthologs of these genes were performed using synonymous nucleotide substitutions (Fig. 3.2). Here, we tried to determine whether the alleles differentiated before or after the therian divergence. Phylogenies of four gametolog pairs (*HSFX/Y, SOX3/SRY, RBMX/Y*, and *XKRX/Y*) contained separate monophyletic clusters for X- and Y-linked genes (Fig. 3.2 *A*-*D*). This topology indicated that these gametolog pairs differentiated before the therian divergence. In contrast, the trees constructed using the other three gametolog pairs (*RPS4X/Y*, *SMCX/Y*, and *UBE1X/Y*) had topologies that differed from the trees described above (Fig. 3.2 *E*-*G*). The *RPS4X/Y* tree indicated that the marsupial gametologs were monophyletic, but the eutherian gametologs were paraphyletic (Fig. 3.2*E*). The clusters of marsupial X and Y genes were more closely related to eutherian X clusters than to Y clusters (Fig. 3.2*E*). By contrast, *SMCX/Y* and *UBE1X/Y* trees indicated that the eutherian gametologs formed monophyly, and the gametologs of eutherians and marsupials formed paraphyly (Fig. 3.2*FG*). The topology of seven phylogenies was confirmed using three different methods: NJ, ML, and MP.

The topologies of the trees based on all nucleotide substitutions (including both synonymous and nonsynonymous substitutions) were the same as those of trees based only on synonymous substitutions (data not shown).

To evaluate the topology of the gametolog phylogenies by different means, we examined the distribution and number of phylogenetically informative sites. For this purpose, we used only the second position of codons even for the comparisons of distantly related species; at this position, substitutions were unlikely to be saturated. *XKRX/Y* was excluded from this analysis because the Y homolog was not present in the opossum genome. For simplicity, four OTUs of gametolog phylogeny were used; these were X and Y sequences from both marsupials (opossums) and eutherians (humans or cats), which each OTU is represented by EX (eutherian X), EY (eutherian Y), MX (marsupial X) and MY (marsupial Y). A phylogenetically informative site supports one of three possible topologies (Fig. 3.3). One topology (topology A: Fig. 3.3*A*) is supported by the following partition: ((EX, MX), (EY, MY)); this type of partition of informative sites would indicate that gametologs differentiated before the therian divergence. In cases where gametologs had differentiated after the therian divergence, the partition is represented as ((EX, EY), (MX, MY)) (topology B: Fig. 3.3*B*). The partition represented as ((EX, MY), (MX, EY)) cannot be explained, but this topology could occur by chance (topology C: Fig. 3.3*C*). The number of phylogenetically informative sites for each topology is shown in Table 3.3.

Phylogenetically informative sites were analyzed for six gametolog pairs (*HSFX/Y, SOX3/SRY, RBMX/Y, RPS4X/Y*, *UBE1X/Y*, and *SMCX/Y*). The analysis of

informative sites in *SOX3/SRY* and *RBMX/Y* supported topology A (Fig. 3.3*A*), and indicated that these gametologs differentiated before the therian divergence (Table 3.3). These findings were consistent with the topology of the phylogenetic trees constructed based on analysis of nucleotide sequences (Figs. 3.2 *B* and *C*). The number of informative sites in *HSFX/Y* and *RPS4X/Y* was too small to support any model. The analysis of informative sites in *SMCX/Y* and *UBE1X/Y* supported topology B (Fig. 3.3*B*) and indicated that the gametologs differentiated after the therian divergence (Table 3.3). However, these results were inconsistent with the topology of the trees constructed based on analysis of nucleotide sequences. To explain this inconsistency, we invoked genetic exchange between X and Y chromosomes, which is also called gene conversion

3.4.3 Gene conversion between gametologs

We searched for bias in the distribution of informative sites. In only *UBE1X/Y*, the bias of distribution was found; for eutherian gametologs the topologies supported by the informative sites were different between the 5' (*SMCX/Ya*: 1st to 10th exons) and 3' (*SMCX/Yb*: 11th to last exons) ends of the gene (Table 3.3). The informative sites in *SMCX/Ya* indicated differentiation between X and Y before the therian divergence, but those in *SMCX/Yb* indicated differentiation occurred after the therian divergence (Table 3.3). In fact, the phylogenetic trees inferred for *SMCX/Ya* and *b* supported the above observation of informative sites (Fig. 3.4). Furthermore, the K_S of eutherian *SMCX/Yb* was lower than that of *SMCX/Ya* (Table 3.4). If the *SMCX/Y* gametologs differentiated

before the therian divergence, genetic exchange by gene conversion or recombination in *SMCX/Yb* must have occurred in the eutherian lineage. Here, gene conversion and recombination would be indistinguishable from each other, but the possibility of gene conversion between X- and Y-linked genes was investigated using statistical tests. A RUNS TEST showed that the distribution of phylogenetically informative sites was significantly different between *SMCX/Ya* and *SMCX/Yb* ($P<0.001$), and GENCONV revealed which segment of the eutherian *SMCX/Yb* region was subject to gene conversion.

In addition, we performed a comparison of nucleotide sequences for the whole region including introns using a window analysis to estimate the number of nucleotide differences per site. This analysis showed that the nucleotide differences varied greatly across the genes (Fig. 3.5). Along the human *SMCX/Y* genes, most regions showed evidence of substantial divergence (~0.6; Fig. 3.5*A*), but the sequence at the 3' end, which included *SMCX/Yb,* showed lower divergence (0.2–0.4; ~5 kb; Fig. 3.5*A*). Because *UBE1X/Y* was categorized as being in stratum 2 in only eutherians, mouse *UBE1X/Y* genomic sequences were also analyzed. The result showed that the nucleotide difference between mouse *UBE1X* and mouse *UBE1Y* was low (~0.4; ~2 kb) in the 3' region, as was the case for human *SMCX/Y* (Figs. 3.5*A* and *B*).

3.4.4 Comparison of functional constraint between genes

To compare the extent of functional constraint of gametologs in marsupials and

eutherians, functional constraint was examined by estimating the K_A /K_S ratio. First, K_A /K_S ratios were examined between gametologs in several species (humans, marmosets, mice, cats, cows, dogs, opossums, wallabies, and stripe-faced dunnarts) to assess the type of selection acting on gametologs. The analysis using most eutherian genes revealed that purifying selection (0.29 ± 0.15; P<0.05; Z = -2.00 – -16.11) was operating. However, purifying selection was not evident in mouse *SOX3/SRY* (1.92; P = 0.001; Z = 3.50) or cat *RPX4X/Y* (1.97; P = 0.022; Z = 2.33)", and the K_A /K_S ratios of these two gametologs may reflect relaxation of functional constraint or positive selection acting on X or Y gametologs.

To distinguish the two possibilities of relaxation of functional constraint or positive selection in the cat *RPX4X/Y* and mouse *SRY/SOX3*, we next studied the variation in K_A/K_S among the lineages leading to X and Y gametologs. The ratio at the branch leading to the mouse *SRY* was much higher (1.57; H_0: K_A/K_S=1, P<0.001; Z = 7.16) than those leading to human (0.13) or opossum *SRY* (0.14), and the estimate suggested the presence of positive selection in the mouse *SRY* lineage. This was consistent with the previous study demonstrating an elevated nonsynonymous substitution rate in the mouse *SRY* (Jansa, Lundrigan, Tucker 2003). The ratio at the branch leading to cat *RPX4Y* was also high (1.13 ± 0.024) compared with that leading to the human *RPX4Y* (0.13 ± 0.035). However, the ratio of cat *RPX4Y* was not significantly different from one (H_0: K_A/K_S=1, P>0.05), raising the possibility that the cat *RPX4Y* gene is becoming a pseudogene (pseudogenization).

In other eutherians, the ratios on the branch leading to the Y gametologs

(*UBE1Y, RPS4Y*, *RBMY* and *SRY*) were higher than those leading to the X gametologs (*UBE1X, RPS4X*, *RBMX*, and *SOX3*), suggesting a general tendency toward relaxation of functional constraint on the Y gametologs. In addition, compared with marsupial genes, the average K_A /K_S ratios of four eutherian genes (*UBE1X/Y, RPS4X/Y*, *RBMX/Y*, and *SOX3/SRY*) were relatively high (Table 3.5), even after excluding mouse *SOX3/SRY* and cat *RPX4X/Y*.

In marsupials, for three gametolog pairs (*SMCX/Y, RBMX/Y*, and *SOX3/SRY*), the ratios at the branch leading to Y gametologs were larger than those leading to X gametologs. In contrast, for the other three pairs (*HSFX/Y, UBE1X/Y*, and *RPS4X/Y*), the ratios for X gametologs were higher than those for Y gametologs (Table 3.5). In the opossum, the ratio at the branch leading to *SMCY* (0.43) was three times higher than that leading to *SMCX* (0.15), indicating that the functional constraint on the *SMCY* had relaxed. A similar tendency was observed for *RBMY* and *SRY* (Table 3.5). By contrast, the ratio at the branch leading to the marsupial *HSFX* (0.73) was three times higher than that leading to marsupial *HSFY* (0.22). The high K_A /K_S ratio of *HSFX* was also observed in the eutherian (*HSFX*: 0.57 ± 0.048, *HSFY*: 0.20 ± 0.13; Table 3.5). Compared with the ratio of the outgroup sequence, the ratio of marsupial and eutherian *HSFX* was also high, but that of the *HSFY* was approximately the same (Table 3.5). The functional constraint on therian *HSFX* was relaxed. Similarly, the relatively high ratio of marsupial *UBE1X* and *RPS4X* indicated that functional constraint was relaxed in the marsupial (Table 3.4). Compared with the corresponding outgroup, however, the ratios of *UBE1X* and *RPS4X* were relatively low, and that of *UBE1Y* and *RPS4Y* (Table 3.5)

was also low. Thus, the functional constraint on *UBE1X/Y* and *RPS4X/Y* may have become strong in the therian ancestor.

3.5 Discussion

3.5.1 Loss or pseudogenization of gametologs in therian evolution

The seven pairs of gametologs are not found in all species used in this study (Table 3.3). Each species might independently lose X-linked or Y-linked genes after speciation or divergence of taxa. The Y chromosome has degenerated rapidly and substantially reorganized (Hughes *et al*. 2010). For example, *UBE1Y* has been lost at least twice in primates, in the stem lineage of Catarrhini (hominoids and Old World monkeys) and in the lineage to marmosets after the radiation of New World monkeys (Mitchell *et al*. 1998). Additionally, *RPS4Y* has been found in primates, cats, and marsupials, but not in rodents, pigs, cows, and horses (Omoe and Endo 1996; Jegalian and Page 1998; Peaks Wilkerson *et al*. 2008). These examples show that independent loss of Y-linked genes occurred at least three times in eutherians. While *XKRY* is only present in primates, but *XKRX/Y* unlikely emerged in primates because of the relatively large estimates of K_S for *XKRX/Y* (Bhowmick, Satta and Takahata 2007). Considering that some Y-liked genes have been lost independently in different lineages, *XKRY* may also have been lost several times in mammals, but the absence of some Y-linked genes from current

genomic data may be due to incomplete Y chromosome sequences rather than evolutionary processes.

Most X-linked genes are well conserved in Theria (Murphy *et al*. 1999; Deakin *et al*. 2008). In some species, however, we could not find some X-linked genes. In mice and rats, an *HSFY*-like gene was found on an autosome, but *HSF* was not found on the X or Y chromosome in either species. Although *HSFX/Y* in mammals generally has a gene structure with introns, the autosomal copy in rodents had only one exon, this gene structure resembles processed genes. Moreover, these autosomal rodent sequences formed a monophyletic cluster with *HSFY* genes (Fig. 3.3A; Mumu Y). These findings may indicate that, in the ancestor of rodents, the *HSFY* retrotransposed to the autosome and the X- and Y-linked genes were subsequently lost. In addition, *TSPX/Y* are absent in marsupials. The estimated divergence time of human *TSPX* and *TSPY* was 178–138 MYA bases on analysis of synonymous substitutions (K_S = 1.06 ± 0.20; Table 3.6), and the estimated time was approximately the same as the therian divergence time (193–186 MYA). This dating, however, could not rule out either of two explanations for the emergence of *TSPX/Y*; these genes emerged first in the eutherian lineage or they were lost only in the marsupial lineage after having emerged in the therian ancestor.

The mechanism of gene loss could be deletion and/or accumulation of mutations. Indeed, the current study showed that cat *RPX4Y* may be in initial stage of pseudogenization. In addition, multiple copies of the human *HSFX/Y, XKRY, RBMY, RPS4Y,* and *TSPY* genes are present in the human genome and sequences comparisons indicated that functional constraints on extra copies are relaxed (data not shown). In fact,

pseudogenization was reported for a few copies of *XKRY, RBMY*, and *TSPY* (Bhowmick, Satta, and Takahata 2007).

3.5.2 Gene conversion between gametologs

Frequent gene conversion between X and Y chromosomes has been observed (Pecon Slattery, Sanner-Wachter and O'Brien 2000; Skaletsky *et al*. 2003; Rozen *et al*. 2003; Ross *et al*. 2005; Iwase *et al*. 2010). Our results indicated possible gene conversion at eutherian *SMCX/Y* and *UBE1X/Y* loci. We also suggested that the gene conversion might have occurred in the marsupial *RPS4X/Y*.

The large K_S value of *RPS4X/Y* indicated that X- and Y-linked genes diverged before the therian divergence. The phylogeny inferred from nucleotide sequences indicated that each therian *RPS4X* and *Y* gene did not form a separate monophyletic cluster, but rather the marsupial X/Y genes were more closely related to the eutherian X genes than to the eutherian Y genes (Fig. 3.3*E*), and the tree constructed using amino acid sequences had the same topology (data not shown). The branch lengths leading to the opossum and human *RPS4X* genes were approximately the same, but the branch length leading to *RPS4Y* was significantly shorter in opossums than in humans; the former was approximately one-sixth of the latter (the number of amino acid substitutions of opossum RPS4X: 8.25 ± 0.18, RPS4Y: 2.75 ± 0.10; human RPS4X: 4.25 ± 0.13, RPS4Y: 16.25 ± 0.25). These observations strongly suggested that the Y chromosomal sequence was converted to the X chromosomal sequences in the opossum.

In a similar way, the direction of gene conversion was determined as "from Y to X" for eutherian *SMCX/Y* (Figs. 3,3*F* and 3.4*B*).

The *K*s of human *SMCX/Ya* (the 5' end of gene) was 0.88 ± 0.050 and that of *SMCX/Yb* (the 3' end of gene) was 0.50 ± 0.020 (Table 3.4). The K_S of the entire gene was calculated as 0.59 ± 0.022, and this value was lower than the average K_S of genes in stratum 1 (1.44 ± 0.066; $P<0.001$). The K_S values estimated for mouse, cat, and dog *SMCX/Ya* and *b* were similar to the values for human (Table 3.4), suggesting that gene conversion at *SMCX/Yb* occurred in the eutherian ancestor. This observation was largely consistent with previous results of Sandstedt and Tucker (2004); they concluded that mouse *SMCX/Y* was in stratum 1, not stratum 2, based on evidence that the number of nucleotide differences between mouse *SMCX* and *SMCY* was low at the 3' end, as it is in humans. A partial reduction in nucleotide divergence between gametologs was observed in *UBE1X/Y* (Figs. 3.5*B*); because of this reduction, the K_S for the entire *UBE1X/Y* gene was low (Table 3.2, 3.4 and Figs. 3.5*A, B*), as was the case for *SMCX/Y*. Excluding the regions showing unusually low nucleotide divergence, *SMCX/Y* and *UBE1X/Y* showed approximately the same extent of divergence as that of the other five genes in stratum 1 (Table 3.2, 3.4 and Figs. 3.5*A, B*). If gene conversion between gametologs in eutherians occurred in *SMCX/Y* and *UBE1X/Y*, two independent reductions in recombination generating two strata, 1 and 2, as proposed by Lahn and Page (1999), are not required to explain the observed estimates.

3.5.3 Marsupial sex chromosome differentiation

Nucleotide and amino acid sequence divergence between gametologs in marsupials and eutherians support the hypothesis that the marsupial X chromosome did differentiate in a steps. In addition to the analysis of the seven genes described in the Results section, another gene, *ATRX*, was also examined. This analysis also supported the presence of single stratum in marsupials. Marsupial and eutherian *ATRX* are both located on the long arm of the X chromosome, but the Y gametolog was found only in marsupials (Pask, Renfree, and Graves 2000, Carvalho-Silva *et al*. 2004). K_S of the marsupial *ATRX/Y* was estimated as 1.03 ± 0.13 (Table. 3.6), this estimate was approximately same as those for the other six pairs of marsupial gametologs (Table. 1 and S2; P>0.05). In a phylogenetic tree of *ATRX/Y* homologs that was based on the synonymous differences, the marsupial *ATRX* did not form a cluster with the marsupial *ATRY*; instead, it clustered with eutherian *ATRX,* suggesting that the gametologs differentiated before the therian divergence (Fig. 3.6). In the eutherian ancestor, *ATRY* may have been lost from the Y chromosome (Pask, Renfree, and Graves 2000, Carvalho-Silva *et al*. 2004).

3.5.4 Functional constraints on eutherian and marsupial gametologs

In the comparison of K_A /K_S ratios, the extent of functional constraint on gametologs in marsupials was different from that in eutherians. In marsupials, the functional constraints on *SMCY, UBE1X* and *RPS4X* were relaxed; whereas, in eutherians, the

constraints on *UBE1Y, RPS4Y*, *RBMY*, and *SRY* were relaxed. These observations indicated that marsupial- or eutherian-specific functional differentiation occurred after the therian divergence.

In general, the individual K_A /K_S of branches leading to each X and Y gene implies that the extent of functional differentiation was different between X and Y chromosomes. The functional constraint on Y gametologs was relatively weaker than the constrains on X gametologs (Table 4, XY: 0.40 ± 0.033; X: 0.19 ± 0.095; Y: 0.21 ± 0.095 in marsupial *ATRX/Y*; XY: 0.88; X: 0.24; Y: 0.49 in the human *TSPX/Y*). Compared with the orthologs in the outgroups, the functional constraints on marsupial and eutherian *RBMY* and *SRY* were relaxed; these findings indicated that functional divergence of the Y genes from the X-linked alleles. However, *HSFX/Y* was a rare case; the functional constraints on the X gametologs had relaxed in both marsupials and eutherians. These observations may indicate that *HSFX* is in the initial stage of pseudogenization. In fact, the cow *HSFX* had premature stop codons and was annotated as a peseudogene (XR_084125). *HSF* belongs to a group of highly conserved regulators that play function as transcriptional activators of heat shock protein (HSP) family (Neuer *et al*. 2000, Tessari *et al*. 2004). The HSP family is expressed in response to stresses, such as elevated temperatures, and it plays an essential role in reproduction (Neuer *et al*. 2000). Human *HSFY* is expressed in testis and is predicted to have a function in spermatogenesis (Tessari *et al*. 2004), but the function of *HSFX* remains unknown. In all therian species examined in this study, the functional constraint on *HSFY* and ancestral orthologs were stronger than on *HSFX*. We hypothesized that the

function of *HSF* in the common ancestor of therians has been carried out by *HSFY* in therians. In short, the strong functional constraint on coding regions in *HSFY* might have been firmly established before therian divergence.

3.5.5 Differentiation of sex chromosome in Theria

Here, sex-chromosome differentiation during very early therein evolution was investigated (Fig. 3.7). At least eight gametolog pairs (*HSFX/Y, SOX3/SRY, RBMX/Y, XKRX/Y, RPS4X/Y, SMCX/Y, UBE1X/Y* and *ATRX/Y*) differentiated simultaneously in the stem lineage of Theria (Fig. 3.7). Repetitive sequences that could be used as cladistic markers were sought using RepeatMasker software (Smit 1996), but the gametologs did not have common or informative repetitive sequences. Instead, the divergence time of gametologs was estimated (see Methods for details). Except for three pairs of gametologs that were possibly subject to gene conversion (*RPS4X/Y, SMCX/Y, UBE1X/Y*) in specific lineages, the average of K_S of the eight genes was 1.33 ± 0.63; this *K*s value indicated that the differentiation of these gametologs occurred 224–173 MYA. This estimate of 224–173 MYA further indicated that X-Y differentiation occurred around or after monotremes diverged (231–217 MYA; van Rheede *et al*. 2005; Fig. 3.7).

The results of this study support the hypothesis that the recombination was suppressed along the entire proto-sex chromosome pair simultaneously in the therian ancestor. The gradual or stepwise suppression of recombination by genetic linkage of

male-specific genes on the Y chromosome has been proposed (Nei 1969; D. Charlesworth, B. Charlesworth, Marais 2005; Graves 2006). Here, it is proposed that the cause of the suppression of recombination between proto-sex chromosomes in the ancestral therian lineage might have been extensive rearrangement, a chromosomal inversion for example (Ohno 1967; Lahn and Page 1999). If, by chance, an autosome with an inversion also carried a sex-reversal mutation or an allele that conferred a sex-specific benefit, the corresponding autosome pair could have differentiated into sex chromosomes (van Doorn and Kirkpatick 2007). Our results indicated that, in the ancestral therian, *HSFY* may have specifically benefitted males.

3.6 Conclusion

In the present study, differentiation of sex chromosomes in the common ancestor of eutherians and marsupials was analyzed to understand the early process of sex chromosomal evolution. Contrary to the hypotheses proposed by Lahn and Page (1999), our data indicated that suppression of recombination occurred once in the therian ancestor. Moreover, we proposed that previously undetected gene conversion event confounded the findings of earlier studies. We concluded that eight gametolog pairs differentiated simultaneously in the therian ancestor. The initial event leading to sex-chromosome differentiation could have been a gross chromosomal rearrangement, such as an inversion, that lead to suppression of recombination between the proto-sex

chromosomes; subsequent functional diversification of a sex-determination gene and sex-differentiation genes then occurred on the X and/or Y chromosomes.

3.7 Gene nomenclature

Due to alterations in nomenclature, clarification is required for the following genes.

SMCX: aliases *KDM5C, JARID1C, DXS1272E,* and *XE169*

SMCY: aliases *KDM5D, JARID1D, HY, HYA,* and *KIAA0234*

UBE1X: aliases *UBA1, UBE1, A1S9T, A1S9, GXP1, POC20,* and *SBX*

UBE1Y: aliases *A1S9Y1, SBY, UBE2,* and *UBE1Y1*

3.8 Reference

Bergero, R., Forrest, A., Kamau, E. and Charlesworth, D. (2007) 'Evolutionary strata on the X chromosomes of the dioecious plant Silene latifolia: evidence from new sex-linked genes.', *Genetics* 175(4): 1945-54.

Bertrand, M., Huijbers, I., Chomez, P. and De Backer, O. (2004) 'Comparative expression analysis of the MAGED genes during embryogenesis and brain development.', *Dev Dyn* 230(2): 325-34.

Bhowmick, B., Satta, Y. and Takahata, N. (2007a) 'The origin and evolution of human

ampliconic gene families and ampliconic structure.', *Genome Res* 17(4): 441-50.

Bhowmick, B. K., Satta, Y. and Takahata, N. (2007b) 'The origin and evolution of human ampliconic gene families and ampliconic structure.', *Genome Res* 17(4): 441-50.

Bischof, J., Ekker, M. and Wevrick, R. (2003) 'A MAGE/NDN-like gene in zebrafish.', *Dev Dyn* 228(3): 475-9.

Bredenbeck, A., Losch, F. O., Sharav, T., Eichler-Mertens, M., Filter, M., Givehchi, A., Sterry, W., Wrede, P. and Walden, P. (2005) 'Identification of noncanonical melanoma-associated T cell epitopes for cancer immunotherapy', *J Immunol* 174(11): 6716-24.

Caballero, O. and Chen, Y. (2009) 'Cancer/testis (CT) antigens: potential targets for immunotherapy.', *Cancer Sci* 100(11): 2014-21.

Calenda, G., Peng, J., Redman, C. M., Sha, Q., Wu, X. and Lee, S. (2006) 'Identification of two new members, XPLAC and XTES, of the XK family.', *Gene* 370: 6-16.

Carvalho-Silva, D. R., O'Neill, R. J., Brown, J. D., Huynh, K., Waters, P. D., Pask, A. J., Delbridge, M. L. and Graves, J. A. (2004a) 'Molecular characterization and evolution of X and Y-borne ATRX homologues in American marsupials.', *Chromosome Res* 12(8): 795-804.

Carvalho-Silva, D. R., O'Neill, R. J., Brown, J. D., Huynh, K., Waters, P. D., Pask, A. J., Delbridge, M. L. and Graves, J. A. (2004b) 'Molecular characterization and evolution of X and Y-borne ATRX homologues in American marsupials.', *Chromosome Res* 12(8): 795-804.

Celis, E., Tsai, V., Crimi, C., DeMars, R., Wentworth, P., Chesnut, R., Grey, H., Sette, A. and Serra, H. (1994) 'Induction of anti-tumor cytotoxic T lymphocytes in normal humans using

primary cultures and synthetic peptide epitopes.', *Proc Natl Acad Sci U S A* 91(6): 2105-9.

Charlesworth, B. and Charlesworth, D. (2000) 'The degeneration of Y chromosomes.', *Philos Trans R Soc Lond B Biol Sci* 355(1403): 1563-72.

Charlesworth, D., Charlesworth, B. and Marais, G. (2005) 'Steps in the evolution of heteromorphic sex chromosomes.', *Heredity* 95(2): 118-28.

Chianese-Bullock, K. A., Pressley, J., Garbee, C., Hibbitts, S., Murphy, C., Yamshchikov, G., Petroni, G. R., Bissonette, E. A., Neese, P. Y., Grosh, W. W. *et al*. (2005) 'MAGE-A1-, MAGE-A10-, and gp100-derived peptides are immunogenic when combined with granulocyte-macrophage colony-stimulating factor and montanide ISA-51 adjuvant and administered as part of a multipeptide vaccine for melanoma', *J Immunol* 174(5): 3080-6.

Chomez, P., De Backer, O., Bertrand, M., De Plaen, E., Boon, T. and Lucas, S. (2001) 'An overview of the MAGE gene family with the identification of all human members of the family.', *Cancer Res* 61(14): 5544-51.

Deakin, J., Koina, E., Waters, P., Doherty, R., Patel, V., Delbridge, M., Dobson, B., Fong, J., Hu, Y., van den Hurk, C. *et al*. (2008a) 'Physical map of two tammar wallaby chromosomes: a strategy for mapping in non-model mammals.', *Chromosome Res* 16(8): 1159-75.

Deakin, J. E., Koina, E., Waters, P. D., Doherty, R., Patel, V. S., Delbridge, M. L., Dobson, B., Fong, J., Hu, Y., van den Hurk, C. *et al*. (2008b) 'Physical map of two tammar wallaby chromosomes: a strategy for mapping in non-model mammals.', *Chromosome Res* 16(8): 1159-75.

Delbridge, M., Patel, H., Waters, P., McMillan, D. and Marshall Graves, J. (2009) 'Does the

human X contain a third evolutionary block? Origin of genes on human Xp11 and Xq28.', *Genome Res* 19(8): 1350-60.

Delgado, C. L., Waters, P. D., Gilbert, C., Robinson, T. J. and Graves, J. A. (2009) 'Physical mapping of the elephant X chromosome: conservation of gene order over 105 million years.', *Chromosome Res* 17(7): 917-26.

Eckery, D. C., Lawrence, S. B., Juengel, J. L., Greenwood, P., McNatty, K. P. and Fidler, A. E. (2002) 'Gene expression of the tyrosine kinase receptor c-kit during ovarian development in the brushtail possum (Trichosurus vulpecula).', *Biol Reprod* 66(2): 346-53.

Elango, N., Lee, J., Peng, Z., Loh, Y. and Yi, S. (2009a) 'Evolutionary rate variation in Old World monkeys.', *Biol Lett* 5(3): 405-8.

Elango, N., Lee, J., Peng, Z., Loh, Y. H. and Yi, S. V. (2009b) 'Evolutionary rate variation in Old World monkeys.', *Biol Lett* 5(3): 405-8.

Escudier, B., Dorval, T., Chaput, N., André, F., Caby, M., Novault, S., Flament, C., Leboulaire, C., Borg, C., Amigorena, S. *et al*. (2005) 'Vaccination of metastatic melanoma patients with autologous dendritic cell (DC) derived-exosomes: results of thefirst phase I clinical trial.', *J Transl Med* 3(1): 10.

Foster, J. W., Brennan, F. E., Hampikian, G. K., Goodfellow, P. N., Sinclair, A. H., Lovell-Badge, R., Selwood, L., Renfree, M. B., Cooper, D. W. and Graves, J. A. (1992) 'Evolution of sex determination and the Y chromosome: SRY-related sequences in marsupials.', *Nature* 359(6395): 531-3.

Goodstadt, L., Heger, A., Webber, C. and Ponting, C. P. (2007) 'An analysis of the gene complement of a marsupial, Monodelphis domestica: evolution of lineage-specific genes

and giant chromosomes.', *Genome Res* 17(7): 969-81.

Gotoh, M., Takasu, H., Harada, K. and Yamaoka, T. (2002) 'Development of HLA-A2402/K(b) transgenic mice.', *Int J Cancer* 100(5): 565-70.

Gotter, A., Nimmakayalu, M., Jalali, G., Hacker, A., Vorstman, J., Conforto Duffy, D., Medne, L. and Emanuel, B. (2007) 'A palindrome-driven complex rearrangement of 22q11.2 and 8q24.1 elucidated using novel technologies.', *Genome Res* 17(4): 470-81.

Graff-Dubois, S., Faure, O., Gross, D., Alves, P., Scardino, A., Chouaib, S., Lemonnier, F. and Kosmatopoulos, K. (2002) 'Generation of CTL recognizing an HLA-A*0201-restricted epitope shared by MAGE-A1, -A2, -A3, -A4, -A6, -A10, and -A12 tumor antigens: implication in a broad-spectrum tumor immunotherapy.', *J Immunol* 169(1): 575-80.

Graves, J. A. (2006) 'Sex chromosome specialization and degeneration in mammals.', *Cell* 124(5): 901-14.

Groeper, C., Gambazzi, F., Zajac, P., Bubendorf, L., Adamina, M., Rosenthal, R., Zerkowski, H., Heberer, M. and Spagnoli, G. (2007) 'Cancer/testis antigen expression and specific cytotoxic T lymphocyte responses in non small cell lung cancer.', *Int J Cancer* 120(2): 337-43.

Grützner, F., Rens, W., Tsend-Ayush, E., El-Mogharbel, N., O'Brien, P. C., Jones, R. C., Ferguson-Smith, M. A. and Marshall Graves, J. A. (2004) 'In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes.', *Nature* 432(7019): 913-7.

Hacker, A., Capel, B., Goodfellow, P. and Lovell-Badge, R. (1995) 'Expression of Sry, the mouse sex determining gene.', *Development* 121(6): 1603-14.

Handley, L. J., Ceplitis, H. and Ellegren, H. (2004) 'Evolutionary strata on the chicken Z chromosome: implications for sex chromosome evolution.', *Genetics* 167(1): 367-76.

Hanley, N. A., Hagan, D. M., Clement-Jones, M., Ball, S. G., Strachan, T., Salas-Cortés, L., McElreavey, K., Lindsay, S., Robson, S., Bullen, P. *et al*. (2000) 'SRY, SOX9, and DAX1 expression patterns during human sex determination and gonadal development.', *Mech Dev* 91(1-2): 403-7.

Harry, J. L., Koopman, P., Brennan, F. E., Graves, J. A. and Renfree, M. B. (1995) 'Widespread expression of the testis-determining gene SRY in a marsupial.', *Nat Genet* 11(3): 347-9.

Hasegawa, M., Thorne, J. and Kishino, H. (2003) 'Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution.', *Genes Genet Syst* 78(4): 267-83.

Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A., Kel, O., Ignatieva, E., Ananko, E., Podkolodnaya, O., Kolpakov, F. *et al*. (1998) 'Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL.', *Nucleic Acids Res* 26(1): 362-7.

Hillig, R., Coulie, P., Stroobant, V., Saenger, W., Ziegler, A. and Hülsmeyer, M. (2001) 'High-resolution structure of HLA-A*0201 in complex with a tumour-specific antigenic peptide encoded by the MAGE-A4 gene.', *J Mol Biol* 310(5): 1167-76.

Hughes, J. F., Skaletsky, H., Pyntikova, T., Graves, T. A., van Daalen, S. K., Minx, P. J., Fulton, R. S., McGrath, S. D., Locke, D. P., Friedman, C. *et al*. (2010) 'Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.', *Nature* 463(7280): 536-9.

Iwase, M., Satta, Y., Hirai, H., Hirai, Y. and Takahata, N. (2010) 'Frequent gene conversion events between the X and Y homologous chromosomal regions in primates.', *BMC Evol*

Biol 10: 225.

JACOBS, P. A. and STRONG, J. A. (1959) 'A case of human intersexuality having a possible XXY sex-determining mechanism.', *Nature* 183(4657): 302-3.

Jegalian, K. and Page, D. C. (1998) 'A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated.', *Nature* 394(6695): 776-80.

Jansa, S. A., Lundrigan, B. L. and Tucker, P. K. (2003) 'Tests for positive selection on immune and reproductive genes in closely related species of the murine genus mus.', *J Mol Evol* 56(3): 294-307.

Katsura, Y. and Satta, Y. (2011) 'Evolutionary history of the cancer immunity antigen MAGE gene family.', *PLoS One* 6(6): e20365.

Kishino, H. and Hasegawa, M. (1989) 'Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea.', *J Mol Evol* 29(2): 170-9.

Kim, T. M., Hong, S. J. and Rhyu, M. G. (2004) 'Periodic explosive expansion of human retroelements associated with the evolution of the hominoid primate.', *J Korean Med Sci* 19(2): 177-85.

Kouprina, N., Mullokandov, M., Rogozin, I., Collins, N., Solomon, G., Otstot, J., Risinger, J., Koonin, E., Barrett, J. and Larionov, V. (2004) 'The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids.', *Proc Natl Acad Sci U S A* 101(9): 3077-82.

Kouprina, N., Noskov, V., Pavlicek, A., Collins, N., Schoppee Bortz, P., Ottolenghi, C., Loukinov, D., Goldsmith, P., Risinger, J., Kim, J. *et al*. (2007) 'Evolutionary

diversification of SPANX-N sperm protein gene structure and expression.', *PLoS One* 2(4): e359.

Kumar, S. and Hedges, S. B. (1998) 'A molecular timescale for vertebrate evolution.', *Nature* 392(6679): 917-20.

Kuroda-Kawaguchi, T., Skaletsky, H., Brown, L., Minx, P., Cordum, H., Waterston, R., Wilson, R., Silber, S., Oates, R., Rozen, S. *et al*. (2001) 'The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men.', *Nat Genet* 29(3): 279-86.

Lahn, B. T. and Page, D. C. (1999) 'Four evolutionary strata on the human X chromosome.', *Science* 286(5441): 964-7.

Lahr, G., Maxson, S. C., Mayer, A., Just, W., Pilgrim, C. and Reisert, I. (1995) 'Transcription of the Y chromosomal gene, Sry, in adult mouse brain.', *Brain Res Mol Brain Res* 33(1): 179-82.

Lange, J., Skaletsky, H., van Daalen, S., Embry, S., Korver, C., Brown, L., Oates, R., Silber, S., Repping, S. and Page, D. (2009) 'Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes.', *Cell* 138(5): 855-69.

Li, W. H., Wu, C. I. and Luo, C. C. (1985) 'A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes.', *Mol Biol Evol* 2(2): 150-74.

Lindblad-Toh, K. Wade, C. M. Mikkelsen, T. S. Karlsson, E. K. Jaffe, D. B. Kamal, M. Clamp, M. Chang, J. L. Kulbokas, E. J. Zody, M. C. *et al*. (2005) 'Genome sequence, comparative

analysis and haplotype structure of the domestic dog.', *Nature* 438(7069): 803-19.

Librado, P. and Rozas, J. (2009a) 'DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.', *Bioinformatics* 25(11): 1451-2.

Librado, P. and Rozas, J. (2009b) 'DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.', *Bioinformatics* 25(11): 1451-2.

Lucas, S., De Smet, C., Arden, K., Viars, C., Lethé, B., Lurquin, C. and Boon, T. (1998) 'Identification of a new MAGE gene with tumor-specific expression by representational difference analysis.', *Cancer Res* 58(4): 743-52.

Luescher, I., Romero, P., Kuznetsov, D., Rimoldi, D., Coulie, P., Cerottini, J. and Jongeneel, C. (1996) 'HLA photoaffinity labeling reveals overlapping binding of homologous melanoma-associated gene peptides by HLA-A1, HLA-A29, and HLA-B44.', *J Biol Chem* 271(21): 12463-71.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Røder, G., Justesen, S. *et al*. (2004) 'Definition of supertypes for HLA molecules using clustering of specificity matrices.', *Immunogenetics* 55(12): 797-810.

Lurquin, C., De Smet, C., Brasseur, F., Muscatelli, F., Martelange, V., De Plaen, E., Brasseur, R., Monaco, A. and Boon, T. (1997) 'Two members of the human MAGEB gene family located in Xp21.3 are expressed in tumors of various histological origins.', *Genomics* 46(3): 397-408.

López-Sánchez, N., González-Fernández, Z., Niinobe, M., Yoshikawa, K. and Frade, J. (2007) 'Single mage gene in the chicken genome encodes CMage, a protein with functional similarities to mammalian type II Mage proteins.', *Physiol Genomics* 30(2): 156-71.

Makova, K. D. and Li, W. H. (2002) 'Strong male-driven evolution of DNA sequences in humans and apes.', *Nature* 416(6881): 624-6.

Makova, K. D., Yang, S. and Chiaromonte, F. (2004) 'Insertions and deletions are male biased too: a whole-genome analysis in rodents.', *Genome Res* 14(4): 567-73.

Marturano, J., Longhi, R., Casorati, G. and Protti, M. (2008) 'MAGE-A3(161-175) contains an HLA-DRbeta4 restricted natural epitope poorly formed through indirect presentation by dendritic cells.', *Cancer Immunol Immunother* 57(2): 207-15.

Mayer, A., Lahr, G., Swaab, D. F., Pilgrim, C. and Reisert, I. (1998) 'The Y-chromosomal genes SRY and ZFY are transcribed in adult human brain.', *Neurogenetics* 1(4): 281-8.

Mayer, A., Mosler, G., Just, W., Pilgrim, C. and Reisert, I. (2000) 'Developmental profile of Sry transcripts in mouse brain.', *Neurogenetics* 3(1): 25-30.

Meredith, R. W., Westerman, M. and Springer, M. S. (2009) 'A phylogeny of Diprotodontia (Marsupialia) based on sequences for five nuclear genes.', *Mol Phylogenet Evol* 51(3): 554-71.

Mitchell, M. J., Wilcox, S. A., Watson, J. M., Lerner, J. L., Woods, D. R., Scheffler, J., Hearn, J. P., Bishop, C. E. and Graves, J. A. (1998) 'The origin and loss of the ubiquitin activating enzyme gene on the mammalian Y chromosome.', *Hum Mol Genet* 7(3): 429-34.

Miyahara, Y., Naota, H., Wang, L., Hiasa, A., Goto, M., Watanabe, M., Kitano, S., Okumura, S., Takemitsu, T., Yuta, A. *et al*. (2005) 'Determination of cellularly processed HLA-A2402-restricted novel CTL epitopes derived from two cancer germ line genes, MAGE-A4 and SAGE.', *Clin Cancer Res* 11(15): 5581-9.

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K. and Yasunaga, T. (1987) 'Male-driven

molecular evolution: a model and nucleotide sequence analysis.', *Cold Spring Harb Symp Quant Biol* 52: 863-7.

Muller, H. J. (1914) 'A FACTOR FOR THE FOURTH CHROMOSOME OF DROSOPHILA.', *Science* 39(1016): 906.

Murphy, W., Sun, S., Chen, Z., Pecon-Slattery, J. and O'Brien, S. (1999a) 'Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping.', *Genome Res* 9(12): 1223-30.

Murphy, W. J., Sun, S., Chen, Z. Q., Pecon-Slattery, J. and O'Brien, S. J. (1999b) 'Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping.', *Genome Res* 9(12): 1223-30.

Nam, K. and Ellegren, H. (2008) 'The chicken (Gallus gallus) Z chromosome contains at least three nonlinear evolutionary strata.', *Genetics* 180(2): 1131-6.

Nei, M. (1969) 'Heterozygous effects and frequency changes of lethal genes in populations.', *Genetics* 63(3): 669-80.

Nei, M. and Rooney, A. (2005) 'Concerted and birth-and-death evolution of multigene families.', *Annu Rev Genet* 39: 121-52.

Neuer, A., Spandorfer, S. D., Giraldo, P., Dieterle, S., Rosenwaks, Z. and Witkin, S. S. (2000) 'The role of heat shock proteins in reproduction.', *Hum Reprod Update* 6(2): 149-59.

Nicolas, M., Marais, G., Hykelova, V., Janousek, B., Laporte, V., Vyskot, B., Mouchiroud, D., Negrutiu, I., Charlesworth, D. and Monéger, F. (2005) 'A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants.', *PLoS Biol* 3(1): e4.

Nishimura, I., Sakoda, J. and Yoshikawa, K. (2008) 'Drosophila MAGE controls neural precursor proliferation in postembryonic neurogenesis.', *Neuroscience* 154(2): 572-81.

Novellino, L., Castelli, C. and Parmiani, G. (2005) 'A listing of human tumor antigens recognized by T cells: March 2004 update.', *Cancer Immunol Immunother* 54(3): 187-207.

Ohno, S. (1967) *Sex Chromosome and Sex Linked Genes*. New York: Springer-Verlag.

OHNO, S., KAPLAN, W. and KINOSITA, R. (1959) 'Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus.', *Exp Cell Res* 18: 415-8.

Ohtsubo, Y., Ikeda-Ohtsubo, W., Nagata, Y. and Tsuda, M. (2008) 'GenomeMatcher: a graphical user interface for DNA sequence comparison.', *BMC Bioinformatics* 9: 376.

Omoe, K. and Endo, A. (1996) 'Relationship between the monosomy X phenotype and Y-linked ribosomal protein S4 (Rps4) in several species of mammals: a molecular evolutionary analysis of Rps4 homologs.', *Genomics* 31(1): 44-50.

Painter, T. S. (1923) 'FURTHER OBSERVATIONS ON THE SEX CHROMOSOMES OF MAMMALS.', *Science* 58(1500): 247-8.

Pask, A., Renfree, M. B. and Marshall Graves, J. A. (2000) 'The human sex-reversing ATRX gene has a homologue on the marsupial Y chromosome, ATRY: implications for the evolution of mammalian sex determination.', *Proc Natl Acad Sci U S A* 97(24): 13198-202.

Pearks Wilkerson, A. J., Raudsepp, T., Graves, T., Albracht, D., Warren, W., Chowdhary, B. P., Skow, L. C. and Murphy, W. J. (2008) 'Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome.', *Genomics* 92(5): 329-38.

Pecon Slattery, J., Sanner-Wachter, L. and O'Brien, S. J. (2000) 'Novel gene conversion between X-Y homologues located in the nonrecombining region of the Y chromosome in

Felidae (Mammalia).', *Proc Natl Acad Sci U S A* 97(10): 5307-12.

Pold, M., Pold, A., Ma, H. J., Sjak-Shieb, N. N., Vescio, R. A. and Berensonb, J. R. (2000) 'Cloning of the first invertebrate MAGE paralogue: an epitope that activates T-cells in humans is highly conserved in evolution', *Dev Comp Immunol* 24(8): 719-31.

Põld, M., Põld, A., Ma, H., Sjak-Shieb, N., Vescio, R. and Berensonb, J. (2000) 'Cloning of the first invertebrate MAGE paralogue: an epitope that activates T-cells in humans is highly conserved in evolution.', *Dev Comp Immunol* 24(8): 719-31.

Rammensee, H., Falk, K. and Rötzschke, O. (1993) 'Peptides naturally presented by MHC class I molecules.', *Annu Rev Immunol* 11: 213-44.

Rens, W., Grützner, F., O'brien, P. C., Fairclough, H., Graves, J. A. and Ferguson-Smith, M. A. (2004) 'Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution.', *Proc Natl Acad Sci U S A* 101(46): 16257-61.

Ross, M. T. Grafham, D. V. Coffey, A. J. Scherer, S. McLay, K. Muzny, D. Platzer, M. Howell, G. R. Burrows, C. Bird, C. P. *et al*. (2005) 'The DNA sequence of the human X chromosome.', *Nature* 434(7031): 325-37.

Rozen, S., Skaletsky, H., Marszalek, J. D., Minx, P. J., Cordum, H. S., Waterston, R. H., Wilson, R. K. and Page, D. C. (2003) 'Abundant gene conversion between arms of palindromes in human and ape Y chromosomes.', *Nature* 423(6942): 873-6.

Saitou, N. and Nei, M. (1987a) 'The neighbor-joining method: a new method for reconstructing phylogenetic trees.', *Mol Biol Evol* 4(4): 406-25.

Saitou, N. and Nei, M. (1987b) 'The neighbor-joining method: a new method for reconstructing

phylogenetic trees.', *Mol Biol Evol* 4(4): 406-25.

Sandstedt, S. A. and Tucker, P. K. (2004) 'Evolutionary strata on the mouse X chromosome correspond to strata on the human X chromosome.', *Genome Res* 14(2): 267-72.

Sato, Y., Shinka, T., Sakamoto, K., Ewis, A. A. and Nakahori, Y. (2010) 'The male-determining gene SRY is a hybrid of DGCR8 and SOX3, and is regulated by the transcription factor CP2.', *Mol Cell Biochem* 337(1-2): 267-75.

Satta, Y., Hickerson, M., Watanabe, H., O'hUigin, C. and 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(4): 478-87.

Sawyer, S. (1989) 'Statistical tests for detecting gene conversion.', *Mol Biol Evol* 6(5): 526-38.

Sekido, R. and Lovell-Badge, R. (2008) 'Sex determination involves synergistic action of SRY and SF1 on a specific Sox9 enhancer.', *Nature* 453(7197): 930-4.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., Foster, J. W., Frischauf, A. M., Lovell-Badge, R. and Goodfellow, P. N. (1990) 'A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif.', *Nature* 346(6281): 240-4.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P., Cordum, H., Hillier, L., Brown, L., Repping, S., Pyntikova, T., Ali, J., Bieri, T. *et al*. (2003a) 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.', *Nature* 423(6942): 825-37.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T. *et al*. (2003b) 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.', *Nature* 423(6942):

825-37.

Sonnhammer, E. and Durbin, R. (1995) 'A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.', *Gene* 167(1-2): GC1-10.

Smit, A. F. (1996) 'The origin of interspersed repeats in the human genome.', *Curr Opin Genet Dev* 6(6): 743-8.

Sourdis, J. and Nei, M. (1988) 'Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree.', *Mol Biol Evol* 5(3): 298-311.

Sugihara, K., Sugiyama, D., Byrne, J., Wolf, D., Lowitz, K., Kobayashi, Y., Kabir-Salmani, M., Nadano, D., Aoki, D., Nozawa, S. *et al*. (2007) 'Trophoblast cell activation by trophinin ligation is implicated in human embryo implantation.', *Proc Natl Acad Sci U S A* 104(10): 3799-804.

Takahata, N. and Tajima F. (1991) 'Sampling Errors in Phylogeny.', *Mol Biol Evol* 8:494-502.

Takahata, N. (1994) 'Comments on the detection of reciprocal recombination or gene conversion.', *Immunogenetics* 39(2): 146-9.

Takahata, N., Ishii, K. and Matsuda, H. (1975) 'Effect of temporal fluctuation of selection coefficient on gene frequency in a population.', *Proc Natl Acad Sci U S A* 72(11): 4541-5.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007a) 'MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.', *Mol Biol Evol* 24(8): 1596-9.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007b) 'MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.', *Mol Biol Evol* 24(8): 1596-9.

Tatsumi, T., Kierstead, L., Ranieri, E., Gesualdo, L., Schena, F., Finke, J., Bukowski, R., Brusic, V., Sidney, J., Sette, A. *et al*. (2003) 'MAGE-6 encodes HLA-DRbeta1*0401-presented epitopes recognized by CD4+ T cells from patients with melanoma or renal cell carcinoma.', *Clin Cancer Res* 9(3): 947-54.

Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F. and Makova, K. D. (2006) 'Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison.', *Mol Biol Evol* 23(3): 565-73.

Tessari, A., Salata, E., Ferlin, A., Bartoloni, L., Slongo, M. L. and Foresta, C. (2004) 'Characterization of HSFY, a novel AZFb gene on the Y chromosome with a possible role in human spermatogenesis.', *Mol Hum Reprod* 10(4): 253-8.

Thompson, J., Gibson, T., Plewniak, F., Jeanmougin, F. and Higgins, D. (1997a) 'The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.', *Nucleic Acids Res* 25(24): 4876-82.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997b) 'The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.', *Nucleic Acids Res* 25(24): 4876-82.

Tsunoda, T. and Takagi, T. (1999) 'Estimating transcription factor bindability on DNA.', *Bioinformatics* 15(7-8): 622-30.

van Doorn, G. S. and Kirkpatrick, M. (2007) 'Turnover of sex chromosomes induced by sexual conflict.', *Nature* 449(7164): 909-12.

van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., De Plaen, E., Van den Eynde, B., Knuth, A. and Boon, T. (1991) 'A gene encoding an antigen recognized by cytolytic T

lymphocytes on a human melanoma.', *Science* 254(5038): 1643-7.

van Rheede, T., Bastiaans, T., Boone, D. N., Hedges, S. B., de Jong, W. W. and Madsen, O. (2006) 'The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians.', *Mol Biol Evol* 23(3): 587-97.

Veyrunes, F., Waters, P., Miethke, P., Rens, W., McMillan, D., Alsop, A., Grützner, F., Deakin, J., Whittington, C., Schatzkamer, K. *et al*. (2008) 'Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes.', *Genome Res* 18(6): 965-73.

Votintseva, A. A. and Filatov, D. A. (2009a) 'Evolutionary strata in a small mating-type-specific region of the smut fungus Microbotryum violaceum.', *Genetics* 182(4): 1391-6.

Votintseva, A. A. and Filatov, D. A. (2009b) 'Evolutionary strata in a small mating-type-specific region of the smut fungus Microbotryum violaceum.', *Genetics* 182(4): 1391-6.

Vujanovic, L., Mandic, M., Olson, W., Kirkwood, J. and Storkus, W. (2007) 'A mycoplasma peptide elicits heteroclitic CD4+ T cell responses against tumor antigen MAGE-A6.', *Clin Cancer Res* 13(22 Pt 1): 6796-806.

Wallis, M. C., Waters, P. D., Delbridge, M. L., Kirby, P. J., Pask, A. J., Grützner, F., Rens, W., Ferguson-Smith, M. A. and Graves, J. A. (2007) 'Sex determination in platypus and echidna: autosomal location of SOX3 confirms the absence of SRY from monotremes.', *Chromosome Res* 15(8): 949-59.

Wang, P., McCarrey, J., Yang, F. and Page, D. (2001) 'An abundance of X-linked genes expressed in spermatogonia.', *Nat Genet* 27(4): 422-6.

Wang, X., Cohen, W., Castelli, F., Almunia, C., Lethé, B., Pouvelle-Moratille, S., Munier, G., Charron, D., Ménez, A., Zarour, H. *et al*. (2007) 'Selective identification of HLA-DP4 binding T cell epitopes encoded by the MAGE-A gene family.', *Cancer Immunol Immunother* 56(6): 807-18.

Waters, P. D., Duffy, B., Frost, C. J., Delbridge, M. L. and Graves, J. A. (2001) 'The human Y chromosome derives largely from a single autosomal region added to the sex chromosomes 80-130 million years ago.', *Cytogenet Cell Genet* 92(1-2): 74-9.

Watson, C. M., Margan, S. H. and Johnston, P. G. (1998) 'Sex-chromosome elimination in the bandicoot Isoodon macrourus using Y-linked markers.', *Cytogenet Cell Genet* 81(1): 54-9.

Winandy, S., Wu, P. and Georgopoulos, K. (1995) 'A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma.', *Cell* 83(2): 289-99.

Yu, C., Meyer, D., Campbell, G., Larner, A., Carter-Su, C., Schwartz, J. and Jove, R. (1995) 'Enhanced DNA-binding activity of a Stat3-related protein in cells transformed by the Src oncoprotein.', *Science* 269(5220): 81-3.

Woodburne, M. O., Rich, T. H. and Springer, M. S. (2003) 'The evolution of tribospheny and the antiquity of mammalian clades.', *Mol Phylogenet Evol* 28(2): 360-85.

3.9 Figure Legends

Figure 3.1 The syntenic relationship between the human X chromosome and opossum chromosomes 7, 14, and X.

Human and opossum orthologs are connected by gray lines. In the human X chromosome, each stratum is indicated by a unique color (stratum 1, magenta; stratum 2, yellow; stratum 3, green; and stratum 4, blue). In the opossum chromosomes, regions homologous with strata 1 and 2 of the human X chromosome are indicated by magenta and yellow, respectively.

Figure 3.2 The phylogenetic relationships among seven gametologs.

These trees were based on the number of synonymous nucleotide differences per synonymous site (P_S). The bootstrap values indicated refer to branches only. A bootstrap value of more than 50% is shown. Sequences are listed in Table 3.1. The number of synonymous sites compared without gaps and the number of operation taxonomy units (OTUs) were as follows: (*A*) *HSFX/Y* (53 sites; 14 OTUs), (*B*) *SOX3/SRY* (101 sites; 16 OTUs), (*C*) *RBMX/Y* (107 sites; 17 OTUs), (*D*) *XKRX/Y* (70 sites; 15 OTUs), (*E*) *RPS4X/Y* (289 sites; 11 OTUs), (*F*) *SMCX/Y* (114 sites; 12 OTUs), (*G*) *UBE1X/Y* (140 sites; 11 OTUs). Platypus sequences were used as the outgroups, except in the cases of trees *B* and *D*. For trees *B* and *D*, chicken sequences were used as the outgroups. A vertical gray bar aside a tree showed a monophyletic cluster of X- or Y-linked genes. Bold lines in *E, F*, and *G* show marsupial or eutherian-specific clusters. OTU shows in bold were marsupial. The abbreviations used for species names are as follows: Bota (*Bos taraus*), Cafa (*Canis familiaris*), Caja (*Callithrix jacchus*), Eqca (*Equus caballus*), Feca (*Felis catus*), Gaga (*Gallus gallus*), Hosa (*Homo sapiens*), Loaf (*Loxodonta africana*), Maeu (*Macropus eugenii*), Modo (*Monodelphis domestica*),

Mumu (*Mus musculus*), Orna (*Ornithorhynchus anatinus*), and Smma (*Sminthopsis macroura*).

Figure 3.3 The three possible topologies for comparisons among four genes (eutherian and marsupial X/Y genes).

If gametologs differentiated before speciation, X- or Y-linked genes should form respective monophyletic clusters, as shown in (*A*). If gametologs differentiated after the speciation or lineage-specific recombination (gene conversion) between X and Y genes occurred, the two genes from a species should form a monophyletic cluster, as shown in (*B*). The remaining possibility is shown in (*C*), which cannot be explained by any simple evolutionary scenario. The abbreviations used in this figure are as follows: EX (a eutherian X gene); EY (a eutherian Y gene); MX (a marsupial X gene); MY (a marsupial Y gene).

Figure 3.4 The phylogenic relationships among *SMCX* and *SMCY* genes.

The tree was based on the number of synonymous nucleotide differences per synonymous site (P_S). Only the bootstrap value of more than 50% was indicated. The tree for the 5' portion of the gene (*SMCX/Ya*; 1st–10th exons) is shown in the left panel (A) and that of the 3' portion (*SMCX/Yb*; 11th-last exons) is shown in the right panel (B). The number of synonymous sites compared was 404 bp (*A*) or 972 bp (*B*) without gaps, and 11 OTUs were used. The vertical grays bar in (A) indicate monophyletic clusters of X- or Y-linked genes. Bold lines in (B) indicate a eutherian cluster of both X

and Y-linked genes. OTU shows in bold are marsupial. The abbreviations used for species names are the same as those used in Fig. 3.3.

Figure 3.5 Window analyses of nucleotide divergence of human *SMCX* and *SMCY* genes (A) and mouse *UBE1X* and *UBE1Y* genes (B).

The window size was 500 bp, and overlap between adjacent windows was not permitted. The ordinate represents the extent of nucleotide differences (*p*-distance), and the abscissa represents the position of the nucleotide (bp). Position 1 corresponds to the beginning of exon1.

Figure 3.6 The phylogenic relationship of *ATRX/Y* based on the P_S.

The bootstrap value is indicated for each branch. Only the bootstrap values of more than 50% are shown. The number of synonymous sites compared was 1473 without gaps, and nine OTUs were used. The vertical gray bars beside the tree indicate monophyletic clusters of X-linked genes. OUTs shown in bold are marsupial. The abbreviations for species names are the same as those used in Fig. 3.3.

Figure 3.7 Schematic diagram of sex chromosome evolution in Theria.

After the divergence of Theria, recombination was suppressed over a region containing least eight genes on the proto-XY chromosome. In the stem lineage of marsupials gene conversion occurred between *RPS4X* and *RPS4Y*, and in the stem lineage of eutherians, it occurred between *SMCX* and *SMCY* and between *UBE1X* and *UBE1Y*.

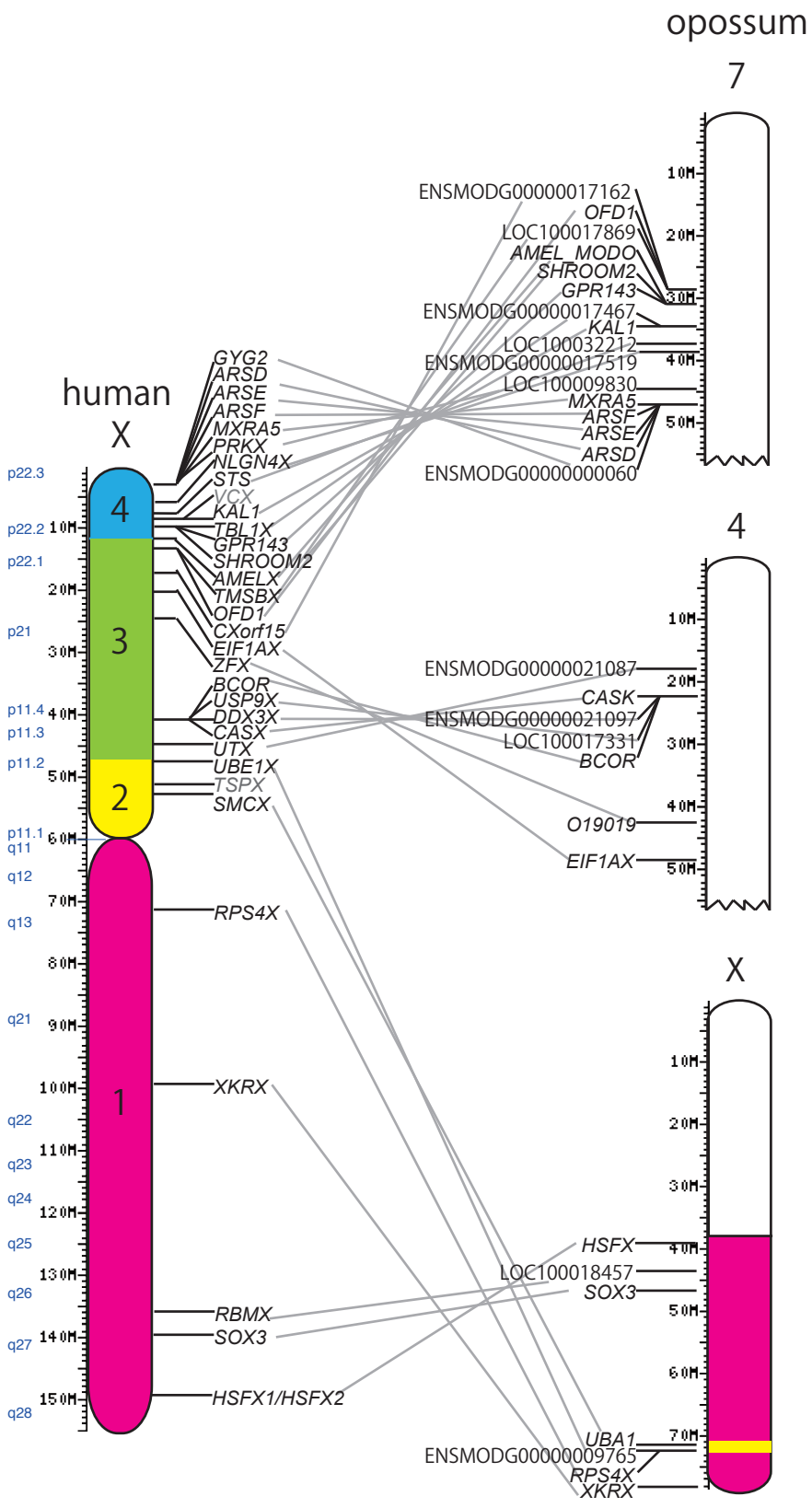


Figure 3.1

The syntenic relationship between the human X chromosome and the opossum 7, 4, and X chromosomes.

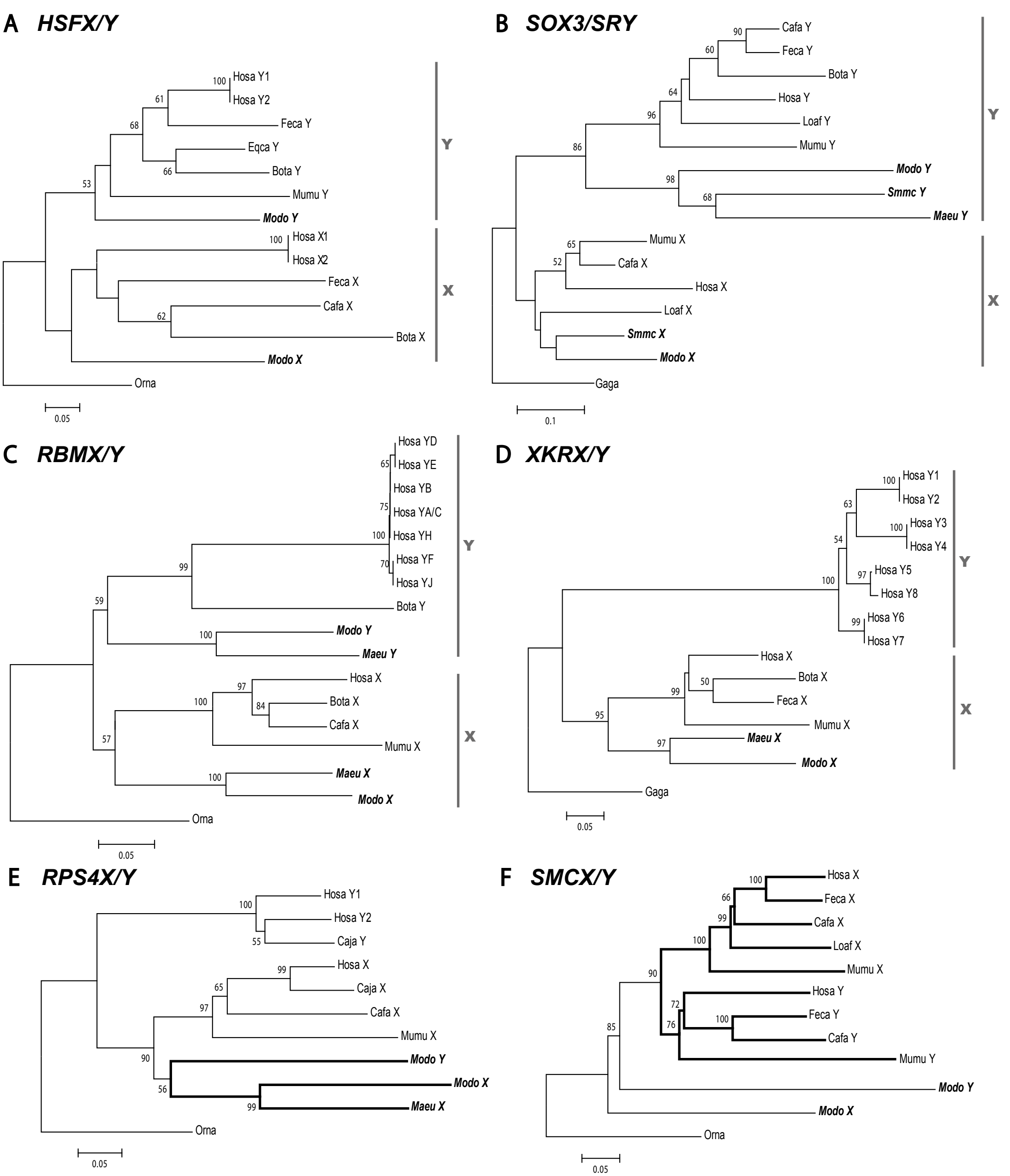


Figure 3.2
The phylogenetic relationships of seven gametologs.

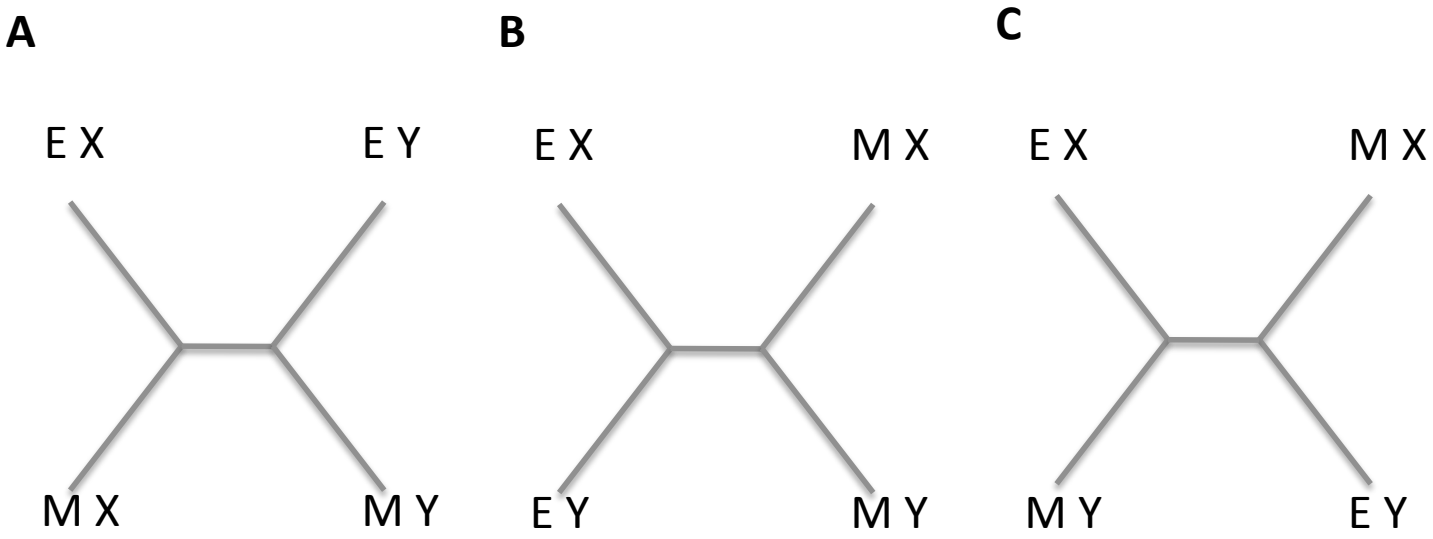


Figure 3.3
The three possible topologies among four genes (eutherian and marsupial X/Y genes).

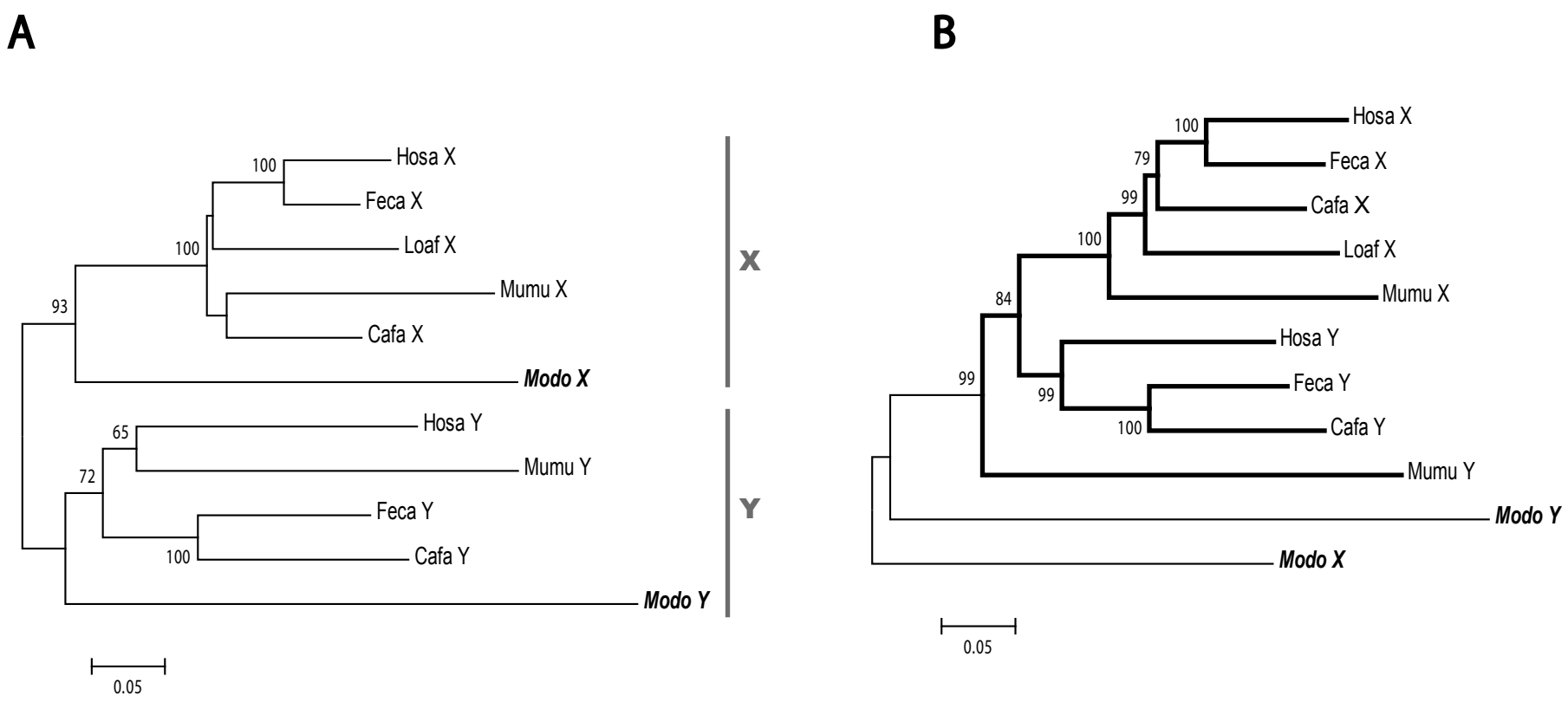
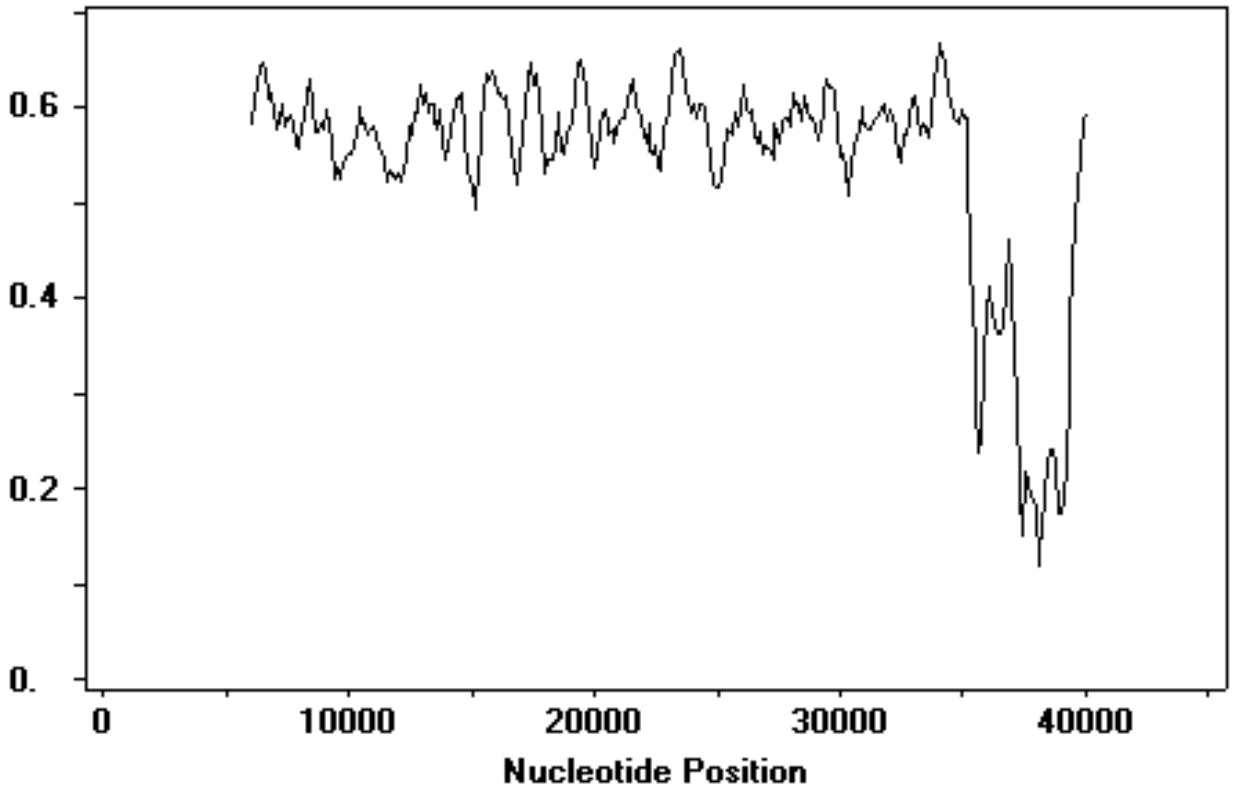


Figure 3.4
The phylogenic relationships of SMCX/Y.

***A* human SMCX/Y genes**



***B* mouse UBE1X/Y gene**

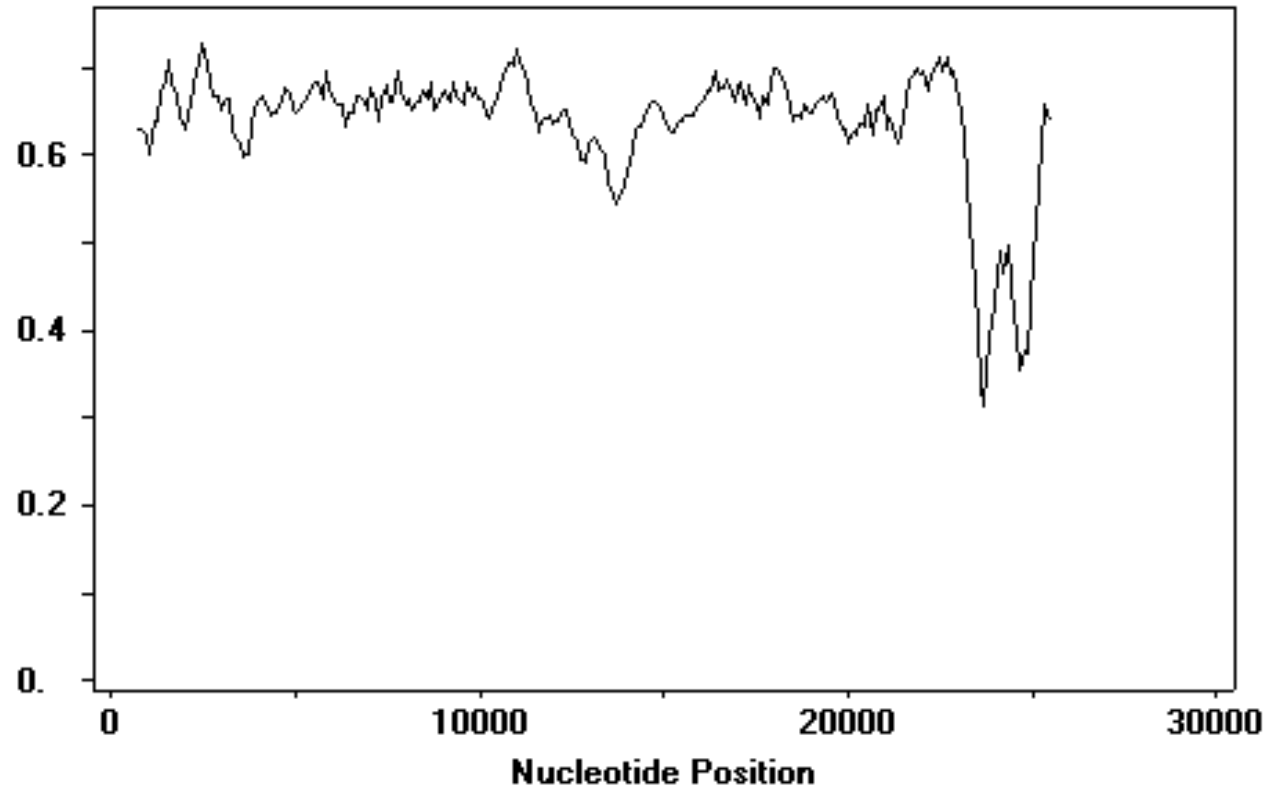


Figure 3.5
Window analyses of nucleotide divergence of human *SMCX/Y* (A) and that of mouse *UBE1X/Y* (B) genes.

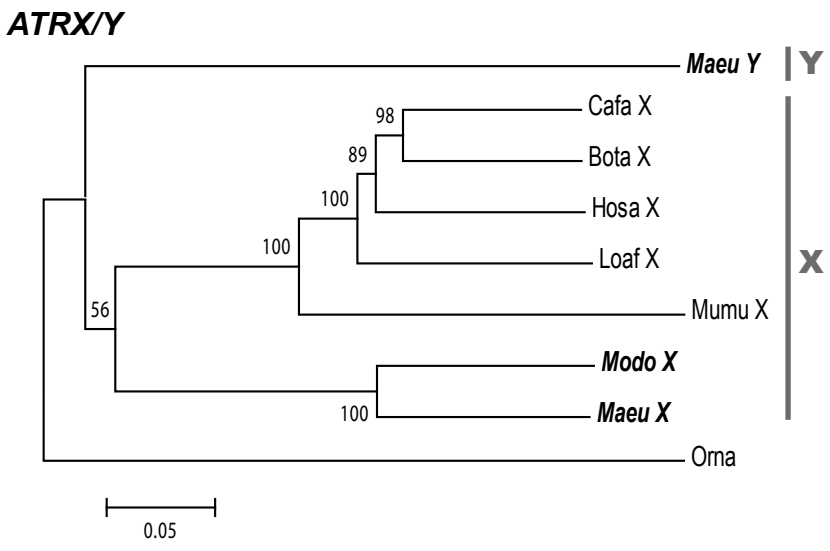


Figure 3.6
The phylogenic relationship of ATRX/Y.
The tree was based on the number of synonymous differences per site (p-distances).

Table 3.1
GENEBANK and Ensenble accession number of nucleotide sequences used in this study.

humans	chromosomal location	opossums	chromosomal location
GYG2 (ENSG00000056998)	X:2746863-2800859	ENSMODG00000000060 (ENSMODG00000000060)	7:47119939-47161669
ARSD (ENSG00000006756)	X:2822011-2847392	ARSD (ENSMODG00000000049)	7:47046839-47075941
ARSE (ENSG00000157399)	X:2852857-2882311	ARSE (ENSMODG00000000044)	7:46985689-47018873
ARSF (ENSG00000062096)	X:2959512-3030767	ARSF (ENSMODG00000025479)	7:46808022-46860449
MXRA5 (ENSG00000101825)	X:3226606-3264684	MXRA5 (ENSMODG00000000020)	7:46441638-46471702
PRKX (ENSG00000183943)	X:3522411-3631649	LOC100009830 XM_001362299.1 (ENSMODG00000000005)	7:45783562-45973283
NLGN4X (ENSG00000146938)	X:5758678-6146904	ENSMODG00000017519 (ENSMODG00000017519)	7:38885072-39215410
STS (ENSG00000101846)	X:7137497-7272851	LOC100032212 XM_001381242.1 (ENSMODG00000017503)	7:37024439-37125870
VCX (ENSG00000182583)	X:7810303-7812184	No homologues	–
KAL1 (ENSG00000011201)	X:8496915-8700227	KAL1 (ENSMODG00000017475)	7:34572132-34793913
TBL1X (ENSG00000101849)	X:9431335-9687780	ENSMODG00000017467 (ENSMODG00000017467)	7:33047290-33069728
GPR143 (ENSG00000101850)	X:9693386-9754337	GPR143 (ENSMODG00000017452)	7:32947332-33001610
SHROOM2 (ENSG00000146950)	X:9754496-9917483	SHROOM2 (ENSMODG00000017449)	7:32687003-32760832
AMELX (ENSG00000125363)	X:11311533-11318881	AMEL_MONDO (ENSMODG00000017370)	7:30712142-30715587
TMSB4X (ENSG00000205542)	X:12993226-12995346	LOC100017869 (ENSMODG00000017326)	7:28385814-28386554
OFD1 (ENSG00000046651)	X:13752832-13787480	OFD1 (ENSMODG00000017308)	7:27030190-27332110
CXorf15 (ENSG00000086712)	X:16804550-16862642	ENSMODG00000017162 (ENSMODG00000017162)	7:22857309-22901405
EIF1AX (ENSG00000173674)	X:20142636-20159962	EIF1AX (ENSMODG00000008104)	4:48149768-48307513
ZFX (ENSG00000005889)	X:24167290-24234206	O19019_MONDO (ENSMODG00000007512)	4:42457713-42507393
BCOR (ENSG00000183337)	X:39909068-40036582	BCOR (ENSMODG00000021102)	4:24014261-24049832
USP9X (ENSG00000124486)	X:40944888-41092185	LOC100017331 XM_001366565.1 (ENSMODG00000021098)	4:22456254-22624222
DDX3X (ENSG00000215301)	X:41192651-41223725	ENSMODG00000021097 (ENSMODG00000021097)	4:22331869-22343917
CASK (ENSG00000147044)	X:41374187-41782716	CASK (ENSMODG00000021095)	4:21956848-22098059
KDM6A (ENSG00000147050)	X:44732423-44971847	ENSMODG00000021087 (ENSMODG00000021087)	4:17894017-18145835
UBA1/UBE1X (ENSG00000130985)	X:47050260-47074527	UBA1/UBE1X (ENSMODG00000010677)	X:71766253-71772564
TSPYL2/TSPX (ENSG00000184205)	X:53111549-53117722	No homologues	–
KDM5C/SMCX (ENSG00000126012)	X:53221334-53254604	ENSMODG00000009765 (ENSMODG00000009765)	X:72253803-72268150
RPS4X (ENSG00000198034)	X:71475529-71497150	RPS4X (AF051136)	X:72,644,966-72,647,168
XKRX (ENSG00000182489)	X:100168431-100184422	XKRX (ENSMODG00000011758)	X:79079670-79091457
ATRX (ENSG00000085224)	X:76760359-77041719	ATRX (ENSMODG00000003920)	X:55899689-56064095
RBMX (ENSG00000147274)	X:135951351-135962884	LOC100018457(XM_001367159)	X:43,223,530-43,229,287
SOX3 (ENSG00000134595)	X:139585152-139587225	SOX3 (ENSMODG00000013932)	X:46390512-46391381
HSFX1 (ENSG00000171116)	X:148855726-148858525	LOC100025313/HSFX (NC_008809.1)	X:39012085-39014573
HSFX2 (ENSG00000171129)	X:148674172-148676974		

Table 3.1
GENEBANK and Ensenble accession number of nucleotide sequences used in this study.

Gene	Eutherians					
	humans (Hosa)	mice (Mumu)	cats (Feca)	dogs (Cafa)	cows (Bota)	marmosets (Caja)
UBE1X	NM_003334	NM_009457 NC_000086.6: 20235547-20260305	EU879978	ENSCAFG00000149	XM_001520965	-
UBE1Y	-	AF150963 NC_000087.6: 155156-180667	DQ329521	-	FJ959389	-
SMCX	NM_001146702 NC_000023.10: 53220503-53254604	AF127245.1	EU879976	NM_001048032	XM_002700121	-
SMCY	NM_004653 NC_000024.9: 21867301-21906825	AF127244.1	EU879977	NM_001113458/DQ156494.1	-	-
RPS4X	NM_001007	NM_009094	EU879986.1	XM_537399	NM_001035445	XM_002762988
RPS4Y	RPS4Y1 (XM_001510756) RPS4Y2 (ENSG00000157828)	-	-	-	-	FJ527003
XKRX	NT_086915: 238884-240414	NM_183319.2	ENSFCAG00000007003	ENSCAFG00000017579	XM_002699816.1	-
XKRY	XKRY1 (NT_011875: 6081676-6083256) XKRY2 (NT_011875: 6448145-6449725) XKRY3 (NT_011875: 6820619-6822204) XKRY4 (NT_011875: 7123594-7125179) XKRY5 (NT_011903: 1910506-1912104) XKRY6 (NT_011903: 2111515-2113110) XKRY7 (NT_011903: 3946439-3948034) XKRY8 (NT_011903: 4147409-4149006)	-	-	-	-	-
RBMX	NM_002139	NM_001166623	-	XM_861341	NM_001172039.1	-
RBMY	RBMY1A1/C (NM_005058) RBMYB (NM_001006121) RBMYD (NM_001006120) RBMYE (NM_001006118) RBMYF/J (NM_152585) RBMYH (NM_005404)	NM_011253	-	-	GU304599.1	-
SOX3	NM005634	NM_009237	-	XM_549298.2	-	-
SRY	NM003140	NM_0115664	NM_001009240	AF107021.1	NM_001014385	-
HSFX	HSFX1 (NM_016153) HSFX2 (NM_001164415)	-	ENSFCAG00000003179	XM_549326.2	XR_084125.1	-
HSFY	HSFY1 (NM_033108) HSFY2 (NM_153716)	NM_027661.2 (chr 1)	NM_001040123	-	gi|297469658:136-1389	-
ATRX	NM_138270.2	NM_009530.2	-	XM_538084.2	XM_002699982.1	-
ATRY	-	-	-	-	-	-
TSPX	NM_022117	NM_029836	ENSFCAG00000008997	gi|74007443:120-2258	gi|297469968:115-2256	-
TSPY	TSPY1 (NT_011878: 13242-16035) TSPY2 (NT_011878: 415676-416022) TSPY3 (NT_011896: 4905332-4908114) TSPY4 (NT_086998: 210118-212931) TSPY5 (NT_086998: 230451-233246) TSPY6 (NT_086998: 250730-253524) TSPY7 (NT_086998: 271075-273870) TSPY8 (NT_011878: 33554-36367) TSPY9 (NT_011878: 53837-56649) TSPY10 (NT_011878: 74167-76962) TSPY11 (NT_011878: 94349-97159) TSPY12 (NT_011878: 451744-454722) TSPY13 (NT_011878: 612690-615483) TSPY14 (NT_011875: 9730806-9733558)	NC_000087.6: 392207..395311	DQ329519.1	-	XM_001250467.2	-

Table 3.1
GENEBANK and Ensenble accession number of nucleotide sequences used in this study.

		Marsupials			Orthologous genes	
horses (Eqca)	elefants (Loaf)	opossums (Modo)	wallabies (Maeu)	stripe-faced dunnarts (Smma)	platypuses	chickens
–	–	XM_001363136	–	–	XM_001520965	XM420609
–	–	GQ253467	–	–		
–	ENSLAFG00000000303	ENSMODG00000009765	–	–	XM_001506932	–
–	–	XM_001364144	–	–		
–	–	AF051136	ENSMEUG00000007969	–	XM_001510756	NM_205108
–	–	AF051137	–	–		
–	–	ENSMODG00000011758	ENSMEUG00000010773	–	–	XM_001234325.1
–	–	–	–	–		
–	–	NM_001032987	AF034741.1	–	XM_001510739.1	EU477531.1
–	–	GU304607	U79565	–		
–	ENSLAFG00000030155	XM_001367288	–	S69429	–	NM_204195
NM_001081810	AF180946.1	AC239615; 61429-62068	Foster et al. 1992	S46279		
ENSECAG00000000093	–	GQ253469.1	–	–	XM_001512596.1	–
–	–	GQ253474	–	–		
–	ENSLAFG00000020570	AY445510	gi|46487452:58-7452	–	ENSOANG00000002389	–
–	–	GU304601	gi|71277006:378-5693	–		
–	–	–	–	–	–	–
–	–	–	–	–		

Table 3.2

The extent of nucleotide divergence per synonymous site (K_S) values ± standard error of seven gametologs in eutherians and marsupials.

	EUTHERIANS					MARSUPIALS	
	humans	mice	cats	dogs	cows	opossums	wallabies
UBE1X/Y	NY	0.77 ± 0.030 (0.48 ± 0.024)	0.57 ± 0.026 (0.40 ± 0.022)	NY	0.57 ± 0.067 (0.40 ± 0.056)	1.47 ± 0.044 (0.65 ± 0.029)	NX, NY
SMCX/Y	0.59 ± 0.22 (0.41 ± 0.018)	0.98 ± 0.029 (0.55 ± 0.021)	0.57 ± 0.021 (0.40 ± 0.018)	0.67 ± 0.023 (0.44 ± 0.019)	NY	2.05 ± 0.042 (0.70 ± 0.024)	NY
RPS4X/Y	0.98 ± 0.070 (0.55 ± 0.052)	NY	1.42 ± 0.11 (0.64 ± 0.073)	NY	NY	0.99 ± 0.070 (0.55 ± 0.052)	NY
XKRX/Y	1.47 ± 0.13 (0.64 ± 0.088)	NY	NY	NY	NY	NY	NY
RBMX/Y	0.82 ± 0.051 (0.50 ± 0.040)	1.01 ± 0.84 (0.56 ± 0.062)	NX, NY	NY	0.82 ± 0.058 (0.50 ± 0.045)	0.67 ± 0.046 (0.45 ± 0.037)	0.61 ± 0.044 (0.42 ± 0.037)
SOX3/SRY	1.39 ± 0.093 (0.63 ± 0.063)	0.93 ± 0.066 (0.53 ± 0.050)	NX	1.14 ± 0.085 (0.59 ± 0.061)	NX	2.30 ± 0.11 (0.72 ± 0.062)	NX
HSFX/Y	2.32 ± 0.14 (0.72 ± 0.076)	NX, NY	2.80 ± 0.12 (0.73 ± 0.060)	NY	2.22 ± 0.10 (0.71 ± 0.059)	0.82 ± 0.043 (0.50 ± 0.056)	NY

K_S values were estimated using the modified Nei-Gojobori method with corrections by Jukes-Cantor model. The standard error was calculated from the maximum variance (Takahata and Tajima 1991). Values in parentheses were estimates of P_S and its standard error. NY or NX means that the gametolog was not available on the Y or X chromosome respectively. In humans, the following genes possess multiple copies (number of copies); *HSFX* (2), *HSFY* (2), *RBMY* (7), *XKRY* (8), *RPS4Y* (2) and *TSPY* (14). In genes with multiple copies, the average value was estimated for all X-Y pairs.

Table 3.3

The phylogenetic informative sites at the second position of the codon.

	A	B	C
HSFX/Y	2	2	2
SOX3/SRY	14	1	0
RBMX/Y	3	0	3
RPS4X/Y	1	1	0
SMCX/Ya	11	3	0
SMCX/Yb	8	30	7
UBEX/Y	5	21	3

The number of sites to support each topology of Fig. 3.3 was shown. A, B or C means topology A, B or C in Fig. 3.3.

Table 3.4
The K_S and P_s of *SMCX/Ya* and *b*.

	SMCXYa	***SMCXYb***
human	0.88 ± 0.050 (0.52 ± 0.039)	0.50 ± 0.024 (0.37 ± 0.020)
mouse	1.30 ± 0.061 (0.62 ± 0.042)	0.89 ± 0.032 (0.52 ± 0.025)
dog	0.81 ± 0.048 (0.50 ± 0.038)	0.62 ± 0.026 (0.42 ± 0.022)
cat	0.69 ± 0.044 (0.45 ± 0.036)	0.52 ± 0.024 (0.38 ± 0.021)
opossum	2.37 ± 0.087 (0.72 ± 0.048)	1.95 ± 0.047 (0.69 ± 0.28)

In humans, mice, dogs, cats and opossums K_S values of *SMCX/Ya* and *b* were estimated using the same method used in table 3.2. The value in parentheses represents P_s.

Table 3.5

The synonymous/nonsynonymous ratio of seven gametologs in eutherians and marsupials.

	EUTHERIANS			MARSUPIALS			
	XY	X	Y	XY	X	Y	OUTGROUPS
UBE1X/Y	0.17 ± 0.035	0.074 ± 0.014	0.094 ± 0.021	0.069	0.062	0.007	0.34 ± 0.057
SMCX/Y	0.21 ± 0.027	0.089 ± 0.080	0.12 ± 0.060	0.58	0.15	0.43	0.095 ± 0.060
RPS4X/Y	0.055 ± 0.0055	0.0013 ± 0.006	0.054 ± 0.012	0.02	0.013	0.0068	0.033 ± 0.0057
XKRX/Y	0.57± 0.013	0.22	0.35	N.A.	N.A.	N.A.	0.36
RBMX/Y	0.34 ± 0.062	0.027 ± 0.012	0.31 ± 0.074	0.13 ± 0.099	0.033 ± 0.015	0.097 ± 0.015	0.014 ± 0.014
SOX3/SRY	0.67 ± 0.053	0.16 ± 0.041	0.50 ± 0.012	0.40 ± 0.031	0.16 ± 0.18	0.25 ± 0.18	0.14 ± 0.035
HSFX/Y	0.76 ± 0.082	0.57 ± 0.048	0.20 ± 0.13	0.95	0.73	0.22	0.27 ± 0.12

A column of XY indicates the K_A/K_S ratio in a comparison between gametologs. A column of X or Y indicates the ratio on the branch leading to X or Y gametologs. Outgroups indicate the ratio of outgoups in a phylogeny. The ratio of eutherians was calculated using the mouse, cat, cow *UBE1X/Y*; human, mouse, cat, dog *SMCX/Y*; human, marmoset *RPS4X/Y*; human *XKRX/Y*; human, mouse, cow *RBMX/Y*; human, dogs *SOX3/SRY*; human, cat, dog, cow *HSFX/Y*. The ratio of marsupials was calculated using the opossum *UBE1X/Y*, *SMCX/Y*, *RPS4X/Y* and *HSFX/Y*, opossum and wallaby *RBMX/Y*, opossum and stripe-faced dunnart *SOX3/SRY*. The ratio of outgroups was calculated using platypuses in *UBE1X/Y, SMCX/Y, RPS4X/Y, RBMX/Y* and *HSFX/Y* or chickens in *XKRX/Y* and *SOX3/SRY*.

Table 3.6
The K_S of marsupial *ATRX/Y* genes.

	EUTHERIANS					MARSUPIALS	
	humans	mice	cats	dogs	cows	opossums	wallabies
ATRX/Y	NY	NY	NX, NY	NY	NY	1.12 ± 0.082 (0.58 ± 0.059)	0.94 ± 0.027 (0.54 ± 0.020)

K_S values of genes were estimated using the same method used in table 3.3. The value in parentheses represents P_S.

Chapter 4

Comparison of intrachromosomal segmental duplications of sex chromosomes

4.1 Abstract

Segmental duplication is a powerful mechanism that alters chromosome and genome structure. We focused on intrachromosomal segmental duplications (ISDs), which produce tandem and/or inverted repeats (> 50 kb) in neighboring regions on the human chromosomes. Our surveys of the human genome revealed that the human X chromosome possesses the largest number of ISDs among the 24 chromosomes, including the Y chromosome. To understand the evolution of ISDs on sex chromosomes, I compared the regions with ISDs among platypuses, opossums, mice, and humans. This comparison revealed that the number and/or size of ISDs differed among species. In particular, the ISDs on human and mouse X chromosomes were more complex than those on the opossum X chromosome. In the opossum, regions of the X chromosome containing ISDs are gene-poor; in contrast, gene density and the number of different gene families are higher in ISDs-containing regions of human and mouse X chromosomes than those of the opossum X chromosome. Furthermore, the platypus X chromosomes had only one ISD, even though the platypus has five X chromosomes. Taken together, these observations indicated that ISDs accumulated on the X

chromosome in the therian ancestor. In the eutherian lineage, the amplification and complexity of ISDs on the X chromosome may have promoted the evolution of multigene families, such as the cancer testis antigens (*CTA*s).

4.2 Introduction

Intrachromosomal segmental duplications (ISDs) are generally relatively large duplications, ranging from 1 kb to >200 kb (The International Human Genome Sequencing Consortium 2001; Bailey *et al*. 2001), which duplicate segments often persist in close proximity to each other. Alteration of genomic structure (e.g., deletions, inversions, duplications, and insertions) often plays a fundamental role in genetic disease and gene evolution (Emanuel and Shaikh 2001 and reviewed in Samonte and Eichler 2002). Comparisons of the human, chimpanzee, and gorilla genomes demonstrate that genomic structure changed quickly and in species-specific ways (Newman *et al*. 2005; Venture *et al*. 2011).

ISDs are not evenly or randomly distributed throughout individual organism genomes. For example, the human and chimpanzee Y chromosome is rich in palindromes (Skaletsky *et al*. 2003, Kuroda-Kawaguchi *et al*. 2001, Bohwmick *et al*. 2007, Hughes *et al*. 2010). These Y chromosomal palindromes are much longer, ranging from 100 kb to ~3 Mb, than the few known palindromes on autosomes, which are less than 1 kb in length (Gotter *et al*. 2007). Genes in palindromes are often subject

to gene conversion, and this conversion plays an important role in maintaining homogeneity among the members of multigene families that reside in these regions (Bohwmick *et al*. 2007). However, rearrangements between palindromic structures may occasionally cause Y chromosome anomalies that can result in failure of spermatogenesis (Lange *et al*. 2009).

Mammalian X chromosomes are less diverse than mammalian Y chromosomes (Murphy *et al*. 1999). The genes within ISDs on X chromosomes are generally members of multigene families, and these genes are expressed mainly in testis cells, such as spermatogonia (Wang *et al.* 2001), as is the case for many genes within gene families on the Y chromosome. These testis-expressed and X-linked genes are testis microRNAs, testis-specific histone (H2A. Bbd), and cancer-testis antigens (*CTAs*) and these genes were subject to rapid evolution (Simpson *et al*. 2005; Guo *et al*. 2009; Caballero and Chen 2009; Ishibashi *et al*. 2010).

Here we counted the number of ISDs >50 kb on each human chromosome using dot-matrix analyses. The number of ISDs in the entire human genome is 310 and the number on the X chromosome is 41. Furthermore, the number on the X chromosome is larger than any of the autosomes and Y chromosomes. Multigene families, such as *CTAs*, were generally located within ISDs that included inverted repeats (IRs) (Warburton *et al*. 2004). The fast rate of evolution of *CTAs* may have resulted from rapid change of ISDs, but this causation is not well established because the relationship between the evolution of *CTAs* and the evolution of ISDs has been examined in only a few cases (Katsura and Satta 2011). Moreover, comparative

genomic analyses of X-linked ISDs have not been performed using data from phylogenetically diverged mammals.

CTAs are highly expressed in a wide range of cancer cells (i.e. in humans) essential to cancer immune systems, and potential targets for cancer immunotherapy (Caballero and Chen 2009). It is not fully understood how *CTAs* became distributed throughout the entire human genome. Of 136 *CTAs* listed (Almeida *et al*. 2009; CT database of July 2010), 36 are located on the X chromosome, and these X-linked *CTA*s are called CT-X antigens or *CT-X*. There are several *CTA* subfamilies: the G antigen (*GAGE*) family, the Sarcoma antigen (*SAGE*) family, and the melanoma antigen (*MAGE*) family among others. Some *CTAs* are of recent origin had a fast rate of evolution. *GAGE* genes diverged in the primate lineages (Gjerstorff and Ditzel 2008; Liu, Q. Zhu and N. Zhu 2008; Killen *et al*. 2011), and *MAGE* type I family is eutherian-specific (Katsura and Satta 2011). However the origin of most *CTAs* was not revealed.

In this chapter, to understand the apparent correlation between ISDs and *CT-X* multigene family, I assessed the distribution of ISDs and gene content of the ISDs on the X chromosomes of non-human mammals, such as platypuses, opossums, and mice. In addition, I investigated how many *CT-X* are located on the ISD and how the ISD and *CT-X* was formed in mammalian evolution.

4.3 Materials and methods

4.3.1 Sequences used

Human genome data (build 36 and 37.2), mouse genome data (build 37.2; X chromosome), opossum genome data (MonDom5; X chromosome), and platypus genome data (build 1.1; X1, 2, 3 and 5 chromosomes) and nucleotide sequence data and corresponding gene information were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). Among the five platypus X chromosomes, nucleotide sequences from the X4 chromosomes were not available yet, and X1, X2, X3 and X5 chromosomes were used in this study.

4.3.2 Identification of ISDs

I conducted a dot-matrix analysis using Dotter (Sonnhammer and Durbin 1995). The genomic sequence was divided into 1-Mb regions, and each region was compared with itself in the analyses.

4.3.3 Homology search and data analyses

I searched the CT database (http://www.cta.lncc.br/) to gather all of human *CTAs* and the peptide database in a Journal of the Academy of Cancer Immunology (http://www.cancerimmunity.org/peptidedatabase/Tcellepitopes.htm) for *CTA* peptide

sequences. To find orthologs of human *CTAs* in non-human mammals, I performed BLAST searches of the transcriptional and genomic databases, using human *CTAs* sequences as query and identified orthologs if the sequences have more than 70% similarity with human queries.

4.4 Results

4.4.1 ISDs on X chromosomes in four species

First, all ISDs (>50 kb) on the X chromosomes in humans, mice, and opossum and those in the platypus genome were identified. The distributions and sizes of these ISDs are shown in Tables 4.1-4. The numbers of X-linked ISDs differed substantially among species; there were 41 in human, 31 in mouse, and 19 in opossum (Tables 4.1-4). Surprisingly, only one X-linked ISDs was found in platypus. The regions exhibiting ISDs occupied 7.1% of the human, 14.7% of the mouse, 4.0% of the opossum, 0.5% of the platypus X chromosome. The mean lengths of the ISDs within each of three species (human, mouse, and opossum) ranged from 167 to 768 kb, and the medians in theses species ranged from 103 to 211 kb (Tables 4.1-3).

Here, the ISDs were categorized as palindromes (P), tandem repeats (T), short repeats (S), or as combinations, PT, PS, or TS. Repeats of relatively short regions (~100 bp) were categorized as S. The structures in the human and mouse were more

complicated than those in the opossum. In humans and mice, PT was the most frequent (42% in humans, 42% in mice) type of ISDs, P was the second most common type (37% in humans, 26% in mice), and T was less common (20% in humans, 16% in mice) and S was the least common (5% in humans, 13% in mice) (Table 4.1-2). In contrast, in opossums few ISDs (11%) were classified as P or T, and most were classified as S (68%) (Table 4.3).

The gene density (the number of genes per 10 kb) in the region exhibiting ISDs in the human (0.25) and mouse (0.33) was more than twice the average gene density on the entire X chromosome (0.11 in human or 0.12 in mouse) (Table 4.1-2). In the opossum, the average gene density in X-linked ISDs (0.04) was less than the average of the entire chromosome (0.07) (Table 4.3). Among the 239 genes in ISD regions of the human X chromosome, 130 (54%), 82 (34%), and 27 (11%) are protein-coding genes, pseudogenes, and non-coding RNA genes, respectively (Table 4.1). In mice, 453 genes were in X-linked ISDs; 224 of which were protein-coding (49%), 221 were pseudogenes (49%), and 8 were non-coding genes (2%) (Table 4.2). In the opossum, only 12 genes were located in X-linked ISDs; 6 of which were protein-coding genes (50%), 4 were pseudogenes (33%), and 2 were non-coding genes (17%) (Table 4.3). In the platypus, the gene density in the ISDs was relatively high (24%) compared to the rest of X chromosome (5%), and 4 protein-coding and 2 pseudogenes were found (Table 4.4).

4.3.2 Distribution of *CT-X* in human genomes

To assess the biological significance of ISDs on X chromosomes, the gene content in these X-linked ISDs was searched (Tables 4.1-3). In the human genome, most of genes present in regions of ISDs were *CT-X* genes (~70%). In the mouse genome, ~10% of the genes in ISDs were *CT-X* genes, but there is no *CT-X* in the opossum and platypus. In humans, the distribution of *CT-X* genes is shown in Fig. 4.1; of 35 *CT-X* genes, 16 (46%) were located in ISDs on the X chromosome.

4.3.3 Homolog search of *CT-X*

I also investigated the origin of the *CT-X* genes on the ISDs by comparing sequences of 16 human *CT-X* genes with genomes from chicken, platypus, opossum, cow, macaque, and chimpanzee. Ten *CT-X antigens* (*GAGE, CTAG, MAGE-C1, -C2, SPANX, CSAGE, CTAG, PHOXF2, CT47,* and *CT45*) were found on only macaque and on the chimpanzee X chromosomes, and six *CT-X* were found on only cow, macaque, and chimpanzee X chromosomes (*MAGE-A, XAGE, SSX, NXF2, Cxorf6, and CXorf48*).

4.5 Discussion

4.5.1 Evolution of genome structures on sex chromosomes

The ISDs on X chromosomes differed among three group of mammals; the X-linked ISDs are larger, more frequently observed, and more structurally complex than those in either marsupials or monotremes. These observations were consistent with results obtained by comparing the entire opossum and human genomes; the marsupial genome had fewer ISDs than the human genome (Mikkelsen *et al*. 2007). Moreover, the gene density in ISDs was substantially higher in the eutherians than in the marsupial.

Repetitive sequences are known to mediate genomic rearrangements and to be involved in the creation of duplications. However, in this analysis, the complexity (the number and size) of ISDs in a genome does not correlate with the number of repetitive sequences in that genome. For each species, the number of repetitive sequences was counted within the ISDs, and the repeat density within ISDs was compatible to the average repeat density along the entire X chromosome (Table 4.1-4). In mice, the density of repetitive sequences in the ISDs was more than twice of the average density estimated for the entire X chromosome. In other mammals, the density of repetitive sequences did not differ between ISDs and other region of the X chromosomes.

In eutherians, the number of ISDs on X chromosomes was larger in humans than in mice, but the ISDs in mice were much more gene-rich and longer in size than those in humans. In mice genes, such as *Xlr, Rhox* and *Xmr*, that are members of multigene families were located in ISDs (Table 4.2). In addition, pseudogenes were more frequent in mice than in humans. In mice, the large number of repetitive sequences in the ISDs might have actively enhanced the emergence and loss of multigene families.

4.5.2 Recent origin and rapid evolution of *CT-X antigens* and genomic structures

The gene content in the ISDs seemed to be species or lineage-specific, but some *CT-X* genes such as *SSX* and *MAGE* were found in ISDs in humans and mice. The mouse and human ISDs that contained orthologous genes might have emerged independently in these two species because the ISDs did not show synteny or similarity other than the *CT-X* orthologs. The phylogenetic analysis of the *MAGE* gene family revealed that *MAGE-A* have evolved in rodent lineages and primate lineages independently (see chapter 5; Katsura and Satta 2011).

The origin of human *CT-X* genes seemed, in general, to be recent, and most of those present in ISDs originated in the ancestor of primates or eutherians. After the divergence of marsupials and eutherians, moreover, complicated ISDs emerged in eutherians. The timing of amplification of *CT-X* genes was estimated to coincide with the timing of the accumulation of complicated ISDs.

The peptides derived from *CT-X* are often recognized by T-cell receptors as tumor antigens, and some *CT-X* play a role in spermatogonia. *CT-X* is expressed in highly proliferative cells. The duplicated copy of *CT-X* on the ISDs may compensate for high expression level. The number or complexity of ISDs and the number of *CT-X* might increase synergistically.

4.5.3 Mystery in the platypus X chromosome

Interestingly, the X chromosomes in platypus contain only one PT ISD (~250 kb). The origin of X chromosomes in monotremes is different from that in marsupials and eutherians; the marsupial and eutherian X chromosomes show synteny with chromosome 6 of platypus. Thus, platypus chromosome 6 was assessed, but no ISD was found on this chromosome too.

The platypus is unique and the platypus X chromosomes do not accumulate tandem or inverted repeats. The reason for the dearth of ISDs on platypus X chromosome is not known. If the platypus sex chromosome emerged recently there may have only been time for a small number of ISDs to accumulate on these chromosomes. The divergence time of platypus sex chromosomes has not been reliably estimated because X and Y gametologs have not been identified. Alternatively, the dearth of X-linked ISDs in platypus may be explained by the hypothesis that ISDs are deleterious in platypus. Given the overall dearth of ISDs in the platypus genome, this last hypothesis is plausible.

4.6 Conclusion and perspectives

The evolution of ISDs on sex chromosomes was rapid and species or lineage-specific. The dearth of X-linked ISDs in platypus is not consistent with the previous hypothesis that sex chromosomes accumulate ISDs. The genomes of many species must be

investigated to understand whether sex chromosomes accumulate ISDs in general or not. To address the question of why ISDs in eutherians were apparently more complicated than those in marsupials, the evolution of genes within the ISDs will be analyzed.

4.7 Reference

Almeida, L. G., Sakabe, N. J., deOliveira, A. R., Silva, M. C., Mundstein, A. S., Cohen, T., Chen, Y. T., Chua, R., Gurung, S., Gnjatic, S. *et al.* (2009) 'CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens.', *Nucleic Acids Res* 37(Database issue): D816-9.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. and Eichler, E. E. (2001) 'Segmental duplications: organization and impact within the current human genome project assembly.', *Genome Res* 11(6): 1005-17.

Bellott, D. W., Skaletsky, H., Pyntikova, T., Mardis, E. R., Graves, T., Kremitzki, C., Brown, L. G., Rozen, S., Warren, W. C., Wilson, R. K. *et al*. (2010) 'Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition.', *Nature* 466(7306): 612-6.

Bhowmick, B. K., Satta, Y. and Takahata, N. (2007) 'The origin and evolution of human ampliconic gene families and ampliconic structure.', *Genome Res* 17(4): 441-50.

Caballero, O. L. and Chen, Y. T. (2009) 'Cancer/testis (CT) antigens: potential targets for immunotherapy.', *Cancer Sci* 100(11): 2014-21.

Emanuel, B. S. and Shaikh, T. H. (2001) 'Segmental duplications: an 'expanding' role in genomic instability and disease.', *Nat Rev Genet* 2(10): 791-800.

Gjerstorff, M. F. and Ditzel, H. J. (2008) 'An overview of the GAGE cancer/testis antigen family with the inclusion of newly identified members.', *Tissue Antigens* 71(3): 187-92.

Gotter, A. L., Nimmakayalu, M. A., Jalali, G. R., Hacker, A. M., Vorstman, J., Conforto Duffy, D., Medne, L. and Emanuel, B. S. (2007) 'A palindrome-driven complex rearrangement of 22q11.2 and 8q24.1 elucidated using novel technologies.', *Genome Res* 17(4): 470-81.

Guo, X., Su, B., Zhou, Z. and Sha, J. (2009) 'Rapid evolution of mammalian X-linked testis microRNAs.', *BMC Genomics* 10: 97.

Hughes, J. F., Skaletsky, H., Pyntikova, T., Graves, T. A., van Daalen, S. K., Minx, P. J., Fulton, R. S., McGrath, S. D., Locke, D. P., Friedman, C. *et al*. (2010) 'Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content.', *Nature* 463(7280): 536-9.

Ishibashi, T., Li, A., Eirín-López, J. M., Zhao, M., Missiaen, K., Abbott, D. W., Meistrich, M., Hendzel, M. J. and Ausió, J. (2010a) 'H2A.Bbd: an X-chromosome-encoded histone involved in mammalian spermiogenesis.', *Nucleic Acids Res* 38(6): 1780-9.

Ishibashi, T., Li, A., Eirín-López, J. M., Zhao, M., Missiaen, K., Abbott, D. W., Meistrich, M., Hendzel, M. J. and Ausió, J. (2010b) 'H2A.Bbd: an X-chromosome-encoded histone involved in mammalian spermiogenesis.', *Nucleic Acids Res* 38(6): 1780-9.

Katsura, Y. and Satta, Y. (2011) 'Evolutionary history of the cancer immunity antigen MAGE gene family.', *PLoS One* 6(6): e20365.

Killen, M. W., Taylor, T. L., Stults, D. M., Jin, W., Wang, L. L., Moscow, J. A. and Pierce, A. J. (2011) 'Configuration and rearrangement of the human GAGE gene clusters.', *Am J Transl Res* 3(3): 234-42.

Kuroda-Kawaguchi, T., Skaletsky, H., Brown, L. G., Minx, P. J., Cordum, H. S., Waterston, R. H., Wilson, R. K., Silber, S., Oates, R., Rozen, S. *et al*. (2001) 'The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men.', *Nat Genet* 29(3): 279-86.

Lange, J., Skaletsky, H., van Daalen, S. K., Embry, S. L., Korver, C. M., Brown, L. G., Oates, R. D., Silber, S., Repping, S. and Page, D. C. (2009) 'Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes.', *Cell* 138(5): 855-69.

Liu, Y., Zhu, Q. and Zhu, N. (2008) 'Recent duplication and positive selection of the GAGE gene family.', *Genetica* 133(1): 31-5.

McPherson, J. D. Marra, M. Hillier, L. Waterston, R. H. Chinwalla, A. Wallis, J. Sekhon, M. Wylie, K. Mardis, E. R. Wilson, R. K. *et al*. (2001) 'A physical map of the human genome.', *Nature* 409(6822): 934-41.

Mikkelsen, T. S., Wakefield, M. J., Aken, B., Amemiya, C. T., Chang, J. L., Duke, S., Garber, M., Gentles, A. J., Goodstadt, L., Heger, A. *et al*. (2007) 'Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences.', *Nature* 447(7141): 167-77.

Murphy, W. J., Sun, S., Chen, Z. Q., Pecon-Slattery, J. and O'Brien, S. J. (1999) 'Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping.', *Genome Res* 9(12): 1223-30.

Newman, T. L., Tuzun, E., Morrison, V. A., Hayden, K. E., Ventura, M., McGrath, S. D., Rocchi, M. and Eichler, E. E. (2005) 'A genome-wide survey of structural variation between human and chimpanzee.', *Genome Res* 15(10): 1344-56.

Samonte, R. V. and Eichler, E. E. (2002) 'Segmental duplications and the evolution of the primate genome.', *Nat Rev Genet* 3(1): 65-72.

Simpson, A. J., Caballero, O. L., Jungbluth, A., Chen, Y. T. and Old, L. J. (2005) 'Cancer/testis antigens, gametogenesis and cancer.', *Nat Rev Cancer* 5(8): 615-25.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., Bieri, T. *et al*. (2003) 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.', *Nature* 423(6942): 825-37.

Sonnhammer, E. L. and Durbin, R. (1995) 'A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.', *Gene* 167(1-2): GC1-10.

Ventura, M., Catacchio, C. R., Alkan, C., Marques-Bonet, T., Sajjadian, S., Graves, T. A., Hormozdiari, F., Navarro, A., Malig, M., Baker, C. *et al*. (2011) 'Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee.', *Genome Res*.

Wang, P. J., McCarrey, J. R., Yang, F. and Page, D. C. (2001) 'An abundance of X-linked genes expressed in spermatogonia.', *Nat Genet* 27(4): 422-6.

Warburton, P. E., Giordano, J., Cheung, F., Gelfand, Y. and Benson, G. (2004) 'Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes.', *Genome Res* 14(10A): 1861-9.

4.8 Figure legend

Figure 4.1 The distribution of *CT-X* antigens on and the ISDs of the human X chromosome.

Yellow, red, blue, and green lines indicate palindrome and tandem repeats (PT), palindrome (P), tandem repeats (T), and short repeats (S), respectively. *CT-X* genes and the respective copy number are indicated beside the X chromosome.

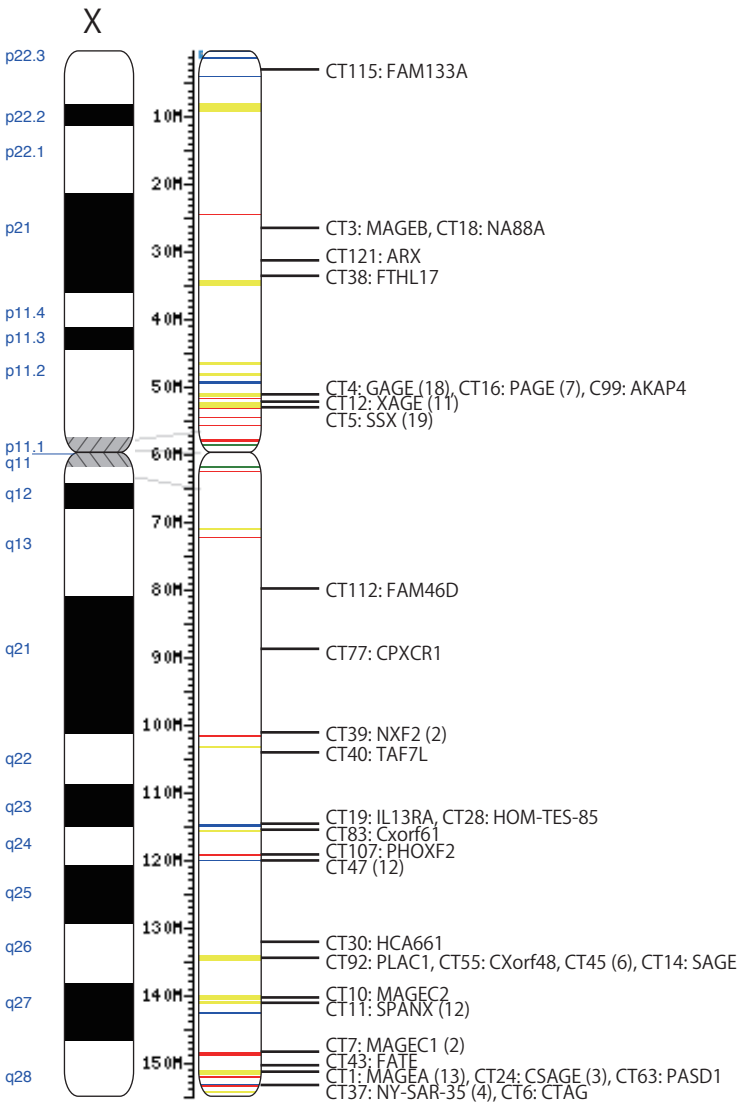


Figure 4.1

The distribution of CT-X antigens on and the ISDs of the human X chromosome.

Table 4.1
The list of genomic structures in the human X chromosomes.

#	chr	id	start	end	length	Gene number	Gene number/10kb	gene content	The number of pseudogene	The number of protein-coding gene	non-coding RNA	pseudogene number/gene number	protein-coding gene number/gene number	repetitive sequences number	repetitive sequences number/10kb	pattern
1	X	X.0	8921	106217	97297	0	0.00		0	0	0	N	N	38	3.91	T
2	X	X.0-1	955194	1129085	173892	1	0.06	pseudo	1	0	0	N	N	313	18.00	T
3	X	X.3	3745075	3865970	120896	10	0.83	tRNA isoleucine, miscRNA	1	7	2	0.04	0.50	247	20.43	T
4	X	X.7-8	7771039	8961066	1190028	11	0.09	VCX, PNPLA4, miscRNA, KAL, microRNA, pseudo, FAM9	3	6	2	0.05	1.00	1775	14.92	PT
5	X	X.24	24239876	24292277	52402	0	0.00		0	0	0	N	1.00	112	21.37	P
6	X	X.34	34054626	34875775	821150	4	0.05	FAM47, TMEM47, pseudo	2	2	0	0.02	1.00	1037	12.63	PT
7	X	X.46	46154438	46463666	309229	9	0.29	ZNF, pseudo, miscRNA, CHST7	5	1	3	0.04	0.50	516	16.69	PT
8	X	X.47-48	47749499	48190726	441228	19	0.43	ZNF, SSX, SPACA, pseudo,	11	8	0	0.08	0.53	825	18.70	PT
9	X	X.49	49046129	49344117	297989	19	0.64	SYP, GAGE, PPP1R3F, FOXP3, pseudo, CCDC	1	18	0	0.08	1.00	422	14.16	T
10	X	X.50	50759755	50940526	180772	2	0.11	pseudo (MAGEA10, 12)	2	0	0	0.01	N	168	9.29	PT
11	X	X.51a	51048876	51285717	236842	4	0.17	NUDT10, CXOrf67, NUDT11, pseudo	1	3	0	0.02	N	285	12.03	PT
12	X	X.51b	51427747	51485105	57359	1	0.17	pseudo	1	0	0	N	1.00	96	16.74	P
13	X	X.52a	52122417	52612860	490444	13	0.27	XAGE, peudo	6	7	0	0.05	0.44	601	12.25	PT
14	X	X.52b	52644752	52838211	193460	9	0.47	SSX, SPANXN5, pesudo	5	4	0	0.04	0.30	409	21.14	PT
15	X	X.52-53	52925114	53018335	93222	2	0.21	FAM156	0	2	0	0.01	1.00	149	15.98	P
16	X	X.54	54317361	54386350	68990	1	0.14	WNK3	0	1	0	N	N	164	23.77	P
17	X	X.55	55494033	55568872	74840	2	0.27	USP51, pseudo (MAGEE1)	1	1	0	0.01	1.00	98	13.09	P
18	X	X.57	57612387	57972819	360433	5	0.14	ZXD, peudo	3	2	0	0.02	1.00	295	8.18	P
19	X	X.58	58368471	58582379	213909	0	0.00		0	0	0	N	N	24	1.12	S
20	X	X.61	61598738	61832846	234109	0	0.00		0	0	0	N	N	13	0.56	S
21	X	X.62	62265195	62386540	121346	0	0.00		0	0	0	N	N	89	7.33	P
22	X	X.70-71	70790378	71009694	219317	8	0.36	OGT, ACRC, CXorf49. miscRNA, pseudo	1	5	2	0.03	0.20	325	14.82	PT
23	X	X.72	72132629	72223522	90894	3	0.33	PABPC1L2B, miscRNA	0	2	1	0.01	N	100	11.00	P
24	X	X.101	101339187	101631135	291949	6	0.21	TCEAL, BEX5, NXF2, pseudo	1	5	0	0.03	1.00	350	11.99	P
25	X	X.103	103061337	103242828	181492	7	0.39	RAB9B, H2B, TMSB15B	4	3	0	0.03	0.43	229	12.62	PT
26	X	X.114	114678625	114981540	302916	5	0.17	PLS3, pseudo	4	1	0	0.02	N	561	18.52	T
27	X	X.115	115493090	115722523	229434	2	0.09	SLC6A14, CXorf61	0	2	0	0.01	1.00	248	10.81	PT
28	X	X.119a	119056196	119216041	159846	3	0.19	miscRNA, NKAP, PHOXF2B	0	2	1	0.01	1.00	266	16.64	P
29	X	X.119b	119892964	119948552	55589	0	0.00		0	0	0	N	0.92	105	18.89	T
30	X	X.134a	134056724	134387311	330588	17	0.51	CXorf48, FAM127, miscRNA, pesudo	6	8	3	0.07	0.71	504	15.25	PT
31	X	X.134b	134554927	134856176	301250	10	0.33	SAGE1, CT45A1, DDX26B, miscRNA, pseudo	5	4	1	0.04	0.80	414	13.74	PT
32	X	X.139-140	139911071	140624559	713489	9	0.13	SPANX, LDOC1, miscRNA, pseudo	3	4	2	0.04	0.86	1066	14.94	PT
33	X	X.140-141	140790849	141124415	333567	3	0.09	MAGEC, pseudo	1	2	0	0.01	0.67	461	13.82	PT
34	X	X.142	142425113	142631948	206836	2	0.10	SPANXN3, pseudo	1	1	0	0.01	1.00	263	12.72	T
35	X	X.148	148421848	148866679	444832	14	0.31	miscRNA, HSFX, MAGEA, CXorf40A, TMEM185A, IDS, pseudo	4	8	2	0.06	0.67	519	11.67	P
36	X	X.151	151031433	151708023	676591	13	0.19	microRNA, MAGEA, GABR, miscRNA, pseudo	2	5	6	0.05	0.91	829	12.25	PT
37	X	X.152	152025037	152212579	187543	4	0.21	NSDHL, ZNF185, PNMA5	1	3	0	0.02	1.00	251	13.38	P
38	X	X.153a	153062716	153171946	109231	5	0.46	SSR4, PDZD4, LICAM, AVPR2, pseudo	1	4	0	0.02	0.67	165	15.11	T
39	X	X.153b	153218534	153276696	58163	5	0.86	microRNA, HCFC1, IRAK1, TMEM187	3	0	2	0.02	N	111	19.08	P
40	X	X.153c	153437292	153535994	98703	6	0.61	TEX28, TKTL1, OPN1MW, pseudo	2	4		0.03	0.71	188	19.05	P
41	X	X.154	154220219	154386384	166166	5	0.30		0	5	0	0.02	0.67	247	14.86	PT
		total			10988233	239			82	130	27	0.99	23.47	14878	573.46	
		avrg			268006	6	**0.25**		2	3	1	0.03	0.78	363	**13.99**	
		stdv			232928	5	0.22		2	3	1	0.02	0.25	342	5.00	
		median			206836	5	0.19		1	2	0	0.02	0.88	263	14.16	
		min			52402	0	0.00		0	0	0	0.01	0.20	13	0.56	
		max			1190028	19	0.86		11	18	6	0.08	1.00	1775	23.77	
	X	avrg			155000000	1669	**0.11**							211180	**13.62**	

Table 4.2
The list of genomic structures in the mouse X chromosomes.

#	chr	id	start	end	length	Gene number	Gene number/10kb	gene content	The number of pseudogene	The number of protein-coding gene	non-coding RNA	pseudogene number/gene number	protein-coding gene number/gene number	repetitive sequences number	repetitive sequences number/10kb	pattern
1	X	musX.3-5	3000000	4970000	1970000	38	0.19	predicted gene	24	14	0	0.63	0.37	1946	9.88	PT
2	X	musX.5	5600000	5780000	180000	4	0.22	Nudt, predicted	3	1	0	0.75	0.25	298	16.56	P
3	X	musX.7-8	7900000	8470000	570000	15	0.26	Ssxb, predicted	1	14	0	0.07	0.93	593	10.40	PT
4		musX.8	8530000	8660000	130000	18	1.38	Fthl17, predicted	12	6	0	0.67	0.33	145	11.15	PT
		musX.9	9150000	9230000	80000	0	0.00		0	0	0	N	N	28	3.50	S
5	X	musX.10	10500000	10730000	230000	1	0.04		1	0	0	1.00	N	330	14.35	PT
6	X	musX.23-33	23880000	33250000	9370000	91	0.10	sycp3, Xmr, Spindlin-like protein 2, pol protein, mgclh, gmcl1l	40	44	7	0.44	0.48	9359	9.99	PT
7	X	musX.34-35	34775000	35093000	318000	26	0.82	Rhox	4	22		0.15	0.85	600	18.87	PT
8	X	musX.35	35180000	35260000	80000	5	0.63	Rhox	1	4	0	0.20	0.80	124	15.50	P
9	X	musX.50	50888000	50999990	111990	6	0.54	Cxx1	2	4	0	0.33	0.67	206	18.39	P/S
10	X	musX.51	51080000	53550000	2470000	37	0.15	Xlr, Synaptonemal scomplex protein, slxl1	18	19	0	0.49	0.51	2912	11.79	T
11	X	musX.67	67622000	67,831,000	209000	13	0.62	pseudo (MAGEA4, HSFY2)	9	4	0	0.69	0.31	211	10.10	P
12	X	musX.70	70320000	70620000	300000	16	0.53	Xlr, DXBay18, F8a, Zfp275	14	2	0	0.88	0.13	371	12.37	PT
13	X	musX.71	72016000	72831000	815000	28	0.34	Xlr, Gab3, Dkc1, Mpp1, F8, Fundc2, Mtcp1, Brcc3, Vbp1, Rab39b	15	13	0	0.54	0.46	1424	17.47	PT
14	X	musX.75A	75119000	75248000	129000	2	0.16	Obp1f	1	1	0	0.50	0.50	155	12.02	PT
15	X	musX.75B	75430000	75500000	70000	2	0.29	Obp1b	0	2	0	N	1.00	73	10.43	P
16	X	musX.80	80690000	80767000	77000	2	0.26	Dmd	0	2	0	N	1.00	68	8.83	T
17	X	musX.83	83430000	83500000	70000	2	0.29	Nr0b1, MAGEB4	0	2	0	N	1.00	108	15.43	T
18	X	musX.88A	88289000	88429500	140500	0	0.00		0	0	0	N	N	192	13.67	P
19	X	musX.88B	88510000	88790000	280000	6	0.21	Mabeb2	3	3	0	0.50	0.50	401	14.32	T/S
20	X	musX.92	92129000	92182000	53000	0	0.00		0	0	0	N	N	228	43.02	S
21	X	musX.99	99900000	99999999	99999	7	0.70	Dmrtc1	2	5	0	0.29	0.71	149	14.90	T
22	X	musX.100A	100000000	100100000	100000	6	0.60	Dmrtc1	2	4	0	0.33	0.67	172	17.20	P
23	X	musX.100B	100180000	100300000	120000	4	0.33		4	0	0	1.00	N	169	14.08	P
24	X	musX.103	103420000	103650000	230000	5	0.22	Fnd3c	3	2	0	0.60	0.40	256	11.13	PT
25	X	musX.120	120150000	122700000	2550000	40	0.16	40S ribosomal protein S12, Srsx, Vmn2r121	13	27	0	0.33	0.68	4928	19.33	PT
26	X	musX.123	123290000	123340000	50000	2	0.40		1	1	0	0.50	0.50	693	138.60	S
27	X	musX.131	131650000	132100000	450000	13	0.29	Tceal6, Pramel	7	6	0	0.54	0.46	683	15.18	PT
28	X	musX.144	144200000	146550000	2350000	50	0.21	Ott, lysyl-tRNA, Luzp4 synthetase	38	12	0	0.76	0.24	4410	18.77	PT
29	X	musX.148	148850000	148950000	100000	2	0.20		0	2	0	N	1.00	125	12.50	P
30	X	musX.151	151250000	151550000	300000	8	0.27	Mega1	1	7	0	0.13	0.88	312	659.00	T
31	X	musX.166	166420000	167000000	580000	4	0.07	Mid1, miscRNA	2	1	1	0.50	0.25	659	11.36	S
		total			24583489	453			221	224	8	12.80	15.87	32328	1230.07	
		avrg			768234	14	**0.33**		7	7	0	0.51	0.59	1010	**38.44**	
		stdv			1733273	19	0.29		10	10	1	0.25	0.27	1935	115.52	
		median			194500	6	0.26		2	4	0	0.50	0.50	277	14.20	
		min			50000	0	0.00		0	0	0	0.07	0.13	28	3.50	
		max			9370000	91	1.38		40	44	7	1.00	1.00	9359	659.00	
	X	avrg			167000000	2025	**0.12**							264306	**15.83**	

Table 4.3
The list of genomic structures in the opossum X chromosomes.

#	chr	id	start	end	length	Gene number	Gene number/10kb	Gene content	The number of pseudogene	The number of protein-coding gene	non-coding RNA	pseudogene number/gene number	protein-coding gene number/gene number	repetitive sequences number	repetitive sequences number/10kb	pattern
1	X	modochX-3_2	6986339	7036339	50000	0	0.00		0	0	0	N	N	13	2.60	S
2	X	modochX-7_0	13905878	14005878	100000	3	0.30	pseudo, microRNA	1	0	2	0.08	N	87	8.70	S
3	X	modochX-8_0	14833434	14979026	145592	0	0.00		0	0	0	N	N	47	3.23	T
4	X	modochX-11_0	23315722	23575722	260000	0	0.00		0	0	0	N	N	260	10.00	P
5	X	modochX-11_1	23895722	24005722	110000	1	0.09	RNA binding protein	0	1	0	N	0.08	111	10.09	P
6	X	modochX-12_0	25011516	26011516	1000000	1	0.01	tax1-binding protein 1	0	1	0	N	0.08	983	9.83	PT
7	X	modochX-16_3	35307127	35357127	50000	0	0.00		0	0	0	N	N	33	6.60	S
8	X	modochX-17_6	42470296	42653091	182795	2	0.11	pseudo	2	0	0	0.17	N	292	15.97	T/S
9	X	modochX-18_1	44034091	44084091	50000	0	0.00		0	0	0	N	N	69	13.80	S
10	X	modochX-18_5	47734091	47814091	80000	0	0.00		0	0	0	N	N	24	3.00	S
11	X	modochX-18_7	50064091	50124091	60000	0	0.00		0	0	0	N	N	32	5.33	S
12	X	modochX-19_4	55222931	55325931	103000	0	0.00		0	0	0	N	N	105	10.19	S
13	X	modochX-21_0	58949209	59099209	150000	1	0.07	pseudo	1	0	0	0.08	N	95	6.33	T/S
14	X	modochX-21_5A	64054209	64125209	71000	0	0.00		0	0	0	N	N	26	3.66	S
15	X	modochX-21_5B	64679209	64929208	249999	0	0.00		0	0	0	N	N	651	26.04	S
16	X	modochX-21_6	64929209	65002959	73750	0	0.00		0	0	0	N	N	179	24.27	T
17	X	modochX-23_3	69517840	69672840	155000	4	0.26	CAPN6,serine/threonine-protein kinase PAK 3-like, zinc finger protein	0	4	0	N	0.33	335	21.61	S
18	X	modochX-25_0	72304275	72354275	50000	0	0.00		0	0	0	N	N	186	37.20	S
19	X	modochX-26_3	77112620	77338604	225984	0	0.00		0	0	0	N	N	616	27.26	S
		total			3167120	12			4	6	2	0.33	0.50	4144		
		avrg			166691	1	**0.04**		0	0	0	0.11	0.17	218	**12.933**	
		stdv			213120	1	0.09		1	1	0	0.05	0.14	263	9.884	
		median			103000	0	0.00		0	0	0	0.08	0.08	105	10.000	
		min			50000	0	0.00		0	0	0	0.08	0.08	13	2.600	
		max			1000000	4	0.30		2	4	2	0.17	0.33	983	37.200	
	X	avrg			79000000	532	**0.07**							130902	**16.57**	

Table 4.4
The list of genomic structures in the platypus X chromosomes.

#	chr	id	start	end	length	Gene number	Gene number/10kb	Gene content	The number of pseudogene	The number of protein-coding gene	pseudogene number/gene number	protein-coding gene number/gene number	repetitive sequences number	repetitive sequences number/10kb	pattern
1	X1	ch1-1_2	2350000	2600000	250000	6	0.24	xanthine dehydrogenase/oxidase/oxidoreductase, cysteine-rich secretory protein 3	2	4	0.33	0.67	430	17.20	PT
avrg	X1				46000000	299	**0.07**						86604	**18.83**	
avrg	X2				5,700,000	18	**0.03**						11660	**20.46**	
avrg	X3				6000000	29	**0.05**						11279	**18.80**	
avrg	X5				28000000	154	**0.06**						45903	**16.39**	
avrg	6				163000000	110	**0.01**						31388	**1.93**	
avrg	1				48000000	270	**0.06**						97859	**20.39**	
avrg	2				55000000	366	**0.07**						114506	**20.82**	
avrg	3				60000000	345	**0.06**						116498	**19.42**	
avrg	4				59000000	371	**0.06**						120359	**20.40**	
avrg	5				24600000	161	**0.07**						49509	**20.13**	
avrg	7				40000000	340	**0.09**						78378	**19.59**	
avrg	10				11200000	110	**0.10**						21479	**19.18**	
avrg	11				6800000	89	**0.13**						12130	**17.84**	
avrg	12				15900000	110	**0.07**						32385	**20.37**	
avrg	14				2700000	29	**0.11**						5141	**19.04**	
avrg	15				3800000	40	**0.11**						8267	**21.76**	
avrg	17				1400000	26	**0.19**						2224	**15.89**	
avrg	18				6600000	52	**0.08**						12629	**19.13**	
avrg	20				1820000	13	**0.07**						3769	**20.71**	
total	X				85700000	500	**0.06**						155446	**18.14**	

Chapter 5

Evolutionary History of the Cancer Immunity Antigen MAGE Gene Family

5.1 Abstract

The evolutionary mode of a multi-gene family can change over time, depending on the functional differentiation and local genomic environment of family members. In this study, we demonstrate that the evolution of the melanoma antigen (*MAGE*) gene family on the mammalian X chromosome was affected by both functional differentiation of duplicate genes and local genomic events, including palindrome formation. There are two gene types in the *MAGE* family; type I genes are of relatively recent origin, and they are expressed in cancer cell and encode epitopes that bind human leukocyte antigen (HLA). Type II genes are more ancient, and some are involved in apoptosis or cell proliferation. The evolutionary history of the *MAGE* gene family can be divided into four phases. In phase I, a single-copy ancestral *MAGE* gene was evolutionarily conserved; this phase lasted until the emergence of eutherian mammals. In phase II, a multi-gene family of 10 processed members was formed via RNA-mediated gene duplication (retrotransposition) of an ancestral gene, *MAGE-D*, and emergence of this family coincided with a transposition burst of long interspersed nuclear elements (*LINEs)* elements at the eutherian radiation. Phase III was characterized by DNA-mediated gene duplication. The formation of palindromes in the *MAGE-A* subfamily occurred in an ancestor of the Catarrhini. Phase IV was characterized by the decay of a palindrome in most non-human Catarrhini. Although the palindrome was truncated by

frequent deletions in apes and Old World monkeys, it was retained in humans. Here, we argue that this human-specific retention stems from negative selection acting on *MAGE-A* genes that encode cancer cells epitopes that bind to highly divergent HLA molecules.

5.2 Introduction

The evolution of any clustered multi-gene family is affected by functional divergence of duplicated member genes and the local structure of the genome (Nei, Gu, and Sitnikova 1997; Nei and Rooney 2005). Here, local structure of the genome refers to tandem or inverted repeats (IRs). Evolution of a gene family on IRs, in particular, can be complex because these families are particularly subject to homogenization by frequent gene conversion and structural change due to instability.

Warburton *et al*. (2004) found a preponderance of large, highly homologous IRs on the X and Y chromosomes; ~30% of IRs in the human genome are on the X and Y chromosomes. Many IRs on the X and Y contain genes expressed predominantly in the testis (Warburton *et al*. 2004). Warburton and his colleagues suggest that these IRs play an important role in human genome evolution. However, the precise role of IRs in evolution is still unclear. Therefore, in this study, we attempt to examine the tempo and mode of gene family evolution that are located in IRs. We focus on the melanoma antigen (*MAGE*) gene family because its members are located on a large (~100 kb) palindrome on the human X chromosome.

MAGE homologous sequences have been found in vertebrate (Bischof, Ekker, Wevrick 2003; López-Sánchez *et al*. 2007, van der Bruggen *et al*. 1991, Kirkin, Dzhandzhugazyan, and Zeuthen. 1998, Castelli *et al*. 2000, Chomez *et al*. 2001) and an invertebrate fruit fies (Põld Põld *et al.* 2000). In the human genome, this family is composed of 10 subfamilies, *MAGE-A, -B, -C, -D, -E*, *-F*, *-H*, *-L2*, *NDN*, *NDNL2*, and each subfamily comprises one to 15 genes (Chomez *et al*. 2001). In addition to the classification by subfamily,

MAGE genes are also classified as type I or type II based on their expression patterns and function. Type I and II comprise three (*MAGE-A*, *-B*, and*–C*) and seven (*MAGE-D, -E*, *-F*, *-H*, *-L2*, *NDN*, *NDNL2*) subfamilies, respectively. Type II genes are ubiquitously expressed in somatic cells, and some Type II genes are involved in apoptosis or cell proliferation (Bertrand *et al*. 2004). Type I genes, on the other hand, are expressed in highly proliferating cells such as tumors, placenta and germ line cells (van der Bruggen *et al*. 1991).

All type I *MAGE* genes are located on the X chromosome and encode tumor antigens that play a key role in cancer immunity. Peptides in the MAGE homology domain (MHD), which is 160–170 amino acids long, are recognized by human leukocyte antigen (HLA) class I molecules (van der Bruggen *et al*. 1991). When the antigen (MHD-peptide) on a tumor cell binds to a receptor on a killer T-cell, the T-cell attacks the tumor cell (van der Bruggen *et al*. 1991, Klein and Horejsí 1997). Although all type I *MAGEs* encode epitopes, *MAGE-A3* and *-A6* are highly expressed in tumor cells and encode the highest number of identified epitopes (van der Bruggen *et al*. 2002). *HLA* is exceptionally polymorphic in the human genome and different *HLA* alleles can bind different epitopes (Rammensee, Falk, and Rötzschke 1993; Lund *et al*. 2004). Each *MAGE* gene can encode several epitopes, and bind multiple HLA variants. For these reasons, it is of interest to trace the origin of the association between *HLA* and *MAGE* and to determine how the genetic diversity in the peptide-coding region referring to MHD has evolved and been maintained.

Many *MAGE* genes are reportedly mammalian-specific (Chomez *et al*. 2001). In addition, most *MAGE* genes have a single exon except for *MAGE-D* subfamily members which have 14 exons where an ORF is encoded between the second to 12th exon (Lucus, Brasseur and Boon 1999). Therefore, it has been thought that each subfamily was derived

from *MAGE-D* by RNA-mediated gene duplication (retrotransposition) (Chomez *et al*. 2001). Continuously, the members of subfamilies could amplify by the evolutionary process of retrotransposition and/or DNA-mediated gene duplication. Yet, the relationship between type I and type II genes has not been fully investigated, and the process of diversification of these genes remains unclear.

In this study, we investigate the evolutionary history of the *MAGE* gene family. First, we identified the most anciently diverged *MAGE* genes in vertebrate and invertebrate genomes. Second, we investigate how and when the ancestor of each subfamily emerged, and we focused on their mode of amplification. Third, we focused on the *MAGE-A* subfamily (one of the type I subfamilies) and demonstrated that the gene arrangement in this subfamily changed rapidly. Finally, we show that some human *MAGE-A* genes have been subject to negative selection that prevented homogenization by gene conversion and that maintained genetic variations among the MAGE-A amino acid sequences. We suggest that this selection is related to the maintenance of a variety of HLA binding sites in cancer cells.

5.3 Materials and Methods

5.3.1 Sequences used

Human (*Homo sapience*) nucleotide sequence data and corresponding gene information were obtained from the NCBI database (build 36.3; http://www.ncbi.nlm.nih.gov/). Syntenic or homologous genomic sequences from other primates and mammals, including opossums (*Monodelphis domestica*) and platypuses (*Ornithorhynchus anatinus*), were retrieved from the NCBI and Ensembl databases (http://uswest.ensembl.org/index.html). To find syntenic regions, homology search using human *MAGE* genes as queries were performed using the BLAST program to determine homologous regions in non-human primates and mammals.

5.3.2 Identification of genomic structures

Identification of IRs and tandem repeats was conducted using a dot-matrix approach (Sonnhammer and Durbin 1995). GenomeMatcher (Ohtsubo *et al*. 2008) was then used to obtain detailed information on nucleotide sequence similarity between duplicate units. A diagram drawn by this program depicts the extent of similarity between sequences using color codes, with red representing similarity greater than 95%, orange representing approximately 90%–95%, green representing approximately 85%–90%, and blue representing lower than 85%.

5.3.3 Phylogenetic and molecular evolutionary analyses

To study the phylogenetic relationships among *MAGE* family members, 158 coding sequences (CDSs) from human, chimpanzee (*Pan troglodytes*), macaque (*Macaca mulatta*), mouse (*Mus musculus*), cow (*Bos taurus*), dog (*Canis lupus*), opossum, platypus, and zebrafish (*Danio rerio*) genomes were retrieved from the NCBI database (Table 5.1). *MAGE* homologs were also sought in Ensembl database of sequences from the western African clawed frog (*Xenopus tropicalis*), lampreys (*Petromyzon marinus*), lancelets (*Branchiostoma floridae*), tunicates (*Ciona intestinalis*) and sea urchins (*Strongylocentrotus purpuratus*). In the searches for *MAGE* homologs, *MAGE-D* genes were used as a query because *MAGE-D* is thought to be most similar to the ancestral *MAGE* gene (Chomez *et al*. 2001). The retrieved sequences were also used in phylogenetic analyses.

In the human genome, there were 37 annotated *MAGE* genes on the X chromosome: 15 *MAGE-A*, 11 *MAGE-B*, three *MAGE-C*, five *MAGE-D*, two *MAGE-E*, and one *MAGE-H*. In addition, *MAGE-F* is located on chromosome 3, and *necdin-like 2* (*NDNL2*, also called *MAGE-G*), *MAGE-like 2* (*MAGE-L2*), and *necdin* (*NDN*) are on chromosome 15. In addition to the annotated genes, a homologous sequence (*psMAGEA-like: psMAGEAL,* NC_000023: 2765558..2770471) corresponding to the human *MAGE* pseudogene, *psMAGEA* (NC_000023: complementary 151952946..151957859), was identified. Gene abbreviations used in this study follow the standards used for human genes.

The sequences obtained were aligned using Clustal W software (Thompson *et al*. 1997) and subsequent manual corrections. Sequences of human *MAGE-H*, *-A5*, and mouse *-A9* were short, and they were discarded because inclusion of these sequences made meaningful sequence alignment shorter. The number of nucleotide differences per site

(*p*-distance) was then calculated using MEGA4 (Tamura *et al*. 2007), and a phylogeny was constructed with the neighbor-joining (NJ; Saitou and Nei 1987) method available in this software. Phylogenies were also constructed with Randomized A(x)ccelerated Maximum Likelihood (RAxML; Stamatakis *et al*. 2005) and Bayesian (Bayes) methods. The program used for the RAxML method was available on the internet at http://phylobench.vital-it.ch/raxml-bb/, and the program used for the Bayes method was MrBayes 3 (Ronquist *et al*. 2005). The alignments used here are available upon request from YS or YK. DnaSP v5 (Librado and Rozas 2009) was used for the window analysis of nucleotide divergence. RepeatMasker (Smit 1996) was used to screen sequences for interspersed repeats. For detection of gene conversion, a program, GENECONV (Sawyer, 1989) was used.

5.3.4 Transcription factor binding sites

Transcription factor binding sites (TFBSs) were examined using the TRANSFAC R4.3 database (Heinemeyer *et al.* 1998), which is available on the TFBIND website (http://tfbind.ims.u-tokyo.a.c.jp/; Tsunoda and Takagi 1999). To find a candidate TFBS, upstream sequences of target genes were aligned; highly conserved sequences were pursed as potential TFBSs. The sequences were checked for the presence of TFBSs in the database.

5.4 Results

5.4.1 Origin of the vertebrate and mammalian *MAGE* gene family

To identify *MAGE* orthologs in lampreys, lancelets, tunicates, and sea urchins, a BLAST search was performed using their genomic, cDNA and expressed sequence tag (EST) sequences as search substrate and human *MAGE-D* genes as queries. No *MAGE* homologs were indentified in lampreys or sea urchins, but hypothetical genes in both tunicates (XM_002119518) and lancelets (XM_002613563) showed 37% sequence similarity with the human *MAGE-D1*. The BLAST search indicated that an ancestral *MAGE* gene could have emerged before the divergence of Chordata.

The zebrafish genome possesses a single *MAGE* gene, *Necdin-like 2* (*DareNDNL2*; Bischof, Ekker, and Wevrick, 2003). *NDNL2* genes were found also in humans, mice and cows, but eutherian *NDNL2s* were processed genes and have a single exon, but *DareNDNL2* possessed ~11 exons. A phylogenetic tree based on predicted amino acid sequences showed that eutherian *NDNL2*s formed a cluster distinct from *DareNDNL2* (Figs. 5.1 and 5.2); eutherian *NDNL2s* were not one-to-one orthologs to *DareNDNL2*. *DareNDNL2* is "primary" ortholog to eutherian *MAGE* genes (Han and Hahn 2009). The topology of tree was supported by different three methods; NJ, RAxML and Bayes (data not shown).

The frog and chicken genomes each contained a single *MAGE* gene. In both cases, the synteny between the gene and *DareNDNL2* could not be determined because not all genes have been assigned to a chromosome in these species. However, given that phases at each exon and intron in the coding regions were well conserved (Table 5.2), the single *MAGE*

genes in the frog and chicken were likely to be one-to-one orthologous to *DareNDNL2*.

In the fish, frog, and chicken, a *MAGE* gene was single copy. Humans and mice, however, have multiple subfamilies of *MAGE* genes (Chomez *et al*. 2001). Thus, *MAGE* homologs was investigated in monotremes (platypus) and marsupials (opossum). A BLAST search of the platypus and opossum genomes using the human *MAGE-D1* as a query detected one and two *MAGE*-like (*MAGEL*) sequences, respectively. They were tentatively named *OrnaMAGEL* and *ModoMAGEL1*/*L2,* respectively. BLAST searches using other *MAGE* genes (e.g., *DareNDNL2*) as query also detected *OrnaMAGEL* and *ModoMAGEL1*/*L2*.

The opossum *ModoMAGEL1* and *ModoMAGEL2* genes were located on chromosomes X and 8, respectively. *ModoMAGEL1* contained 11 exons, and *ModoMAGEL2* contained only one exon; therefore, *ModoMAGEL2* was likely to be a processed gene derived from *ModoMAGEL1*. In fact, *ModoMAGEL1* and *ModoMAGEL2* formed a monophyletic cluster in a tree (Fig. 5.1). This cluster was reiterated in trees constructed using three different methods (NJ, RAx ML, and Bayes).

In platypuses the *OrnaMAGEL* gene was located on the contig Ultra 403, and it contained 10 exons. Although the number of exons differed between *OrnaMAGEL* and *ModoMAGEL1*, the phases and sizes of shared exons were well conserved (Table 5.2). Ultra 403 contained the ubiquitin ligase gene *HUWE1* (HECT, UBA and WWE domain containing 1), which was located ~600 kb upstream from *OrnaMAGEL*. An *in situ* hybridization study confirmed that, in the platypus, *HUWE1* is located on chromosome 6 (Delbridge *et al*. 2009); thus, it is likely that this contig is a part of chromosome 6. Platypus chromosome 6 is homologous to the autosomal ancestor of eutherian and marsupial X chromosomes (Delbridge *et al*. 2009). The region surrounding *OrnaMAGEL* on the contig showed a syntenic

relationship with the human Xp11 region. In the human genome, the corresponding position to *OrnaMAGEL* is occupied by *MAGE-D2* and -*D3* (Fig. 5.3). Human *MAGE-D2* and -*D3* possess 13 exons, and the phases and sizes of these exons were conserved with those of *OrnaMAGEL*, *ModoMAGEL1*, and the *MAGE* genes in the chicken, frog, and zebrafish genomes (Table 5.2).

5.4.2 Phylogeny of the mammalian *MAGE* gene family

A tree of human *MAGE* genes showed that the three type I subfamilies formed a monophyletic cluster that was separate from type II subfamilies (Fig. 5.4). This inference was supported by the evidence from five phylogenetically informative substitutions (D16Y, K23T, I62V, A113E, R156Q in the alignment of MHD within MAGE, Fig. 5.5). In addition, the *MAGE-D* subfamily also formed a monophyletic cluster. Based on the relatively low bootstrap probability at nodes of the subfamilies within type I or II, these type I or II subfamilies would have diverged from each other within short period of evolutionary time; however, the number of nucleotides used in this analysis was small (Figs. 5.2 and 5.4).

Since most *MAGE* genes, other than *MAGE-D g*enes, have a single exon for CDS, they are likely processed genes derived from transcripts of *MAGE-D* or other *MAGE-D* processed genes (Chomez *et al*. 2001; Artamonova and Gelfand 2004). Alternatively, another ancestral gene may have produced the *MAGE-D* clade and the single-exon genes. However, no such ancestor gene has been detected in any genome, suggesting that this scenario is highly unlikely.

To study how each gene family formed, representative nucleotide sequences of

subfamilies were compared with each other using dot-matrix analysis (Sonnhammer and Durbin 1995). If an entire coding region including flanking region was duplicated, the dot matrix analysis showed the similarity beyond the CDS. In contrast, if an ancestor of each subfamily was generated by retrotransposition, the analysis showed the similarity in the CDS only.

Except for comparisons between *MAGE-A* and *MAGE-C* sequences, analyses within and between the type I and II categories revealed similarities in CDS regions only. Comparison between *MAGE-A* and -*C* revealed similarities beyond the CDS. Therefore, the ancestral sequence of subfamilies was likely produced not by gene duplication, but by retrotransposition. *MAGE-A* and -*C* were exceptions. In total, eight retrotransposition events of *MAGE* sequences occurred in the ancestral genome, and each processed gene became the prototype of one subfamily. Following these retrotransposition events, DNA-mediated gene duplications took place resulting in independent amplification of each prototype and formation of each subfamily.

5.4.3 Gene duplication and palindrome formation

Notably, the phylogenetic clustering of *MAGE-A* genes differs from that of *MAGE-B* genes (Fig. 5.1). Each of 11 human *MAGE-B* genes formed a monophyletic cluster with their orthologs from other eutherians, whereas *MAGE-A* genes including 15 human genes, formed species- or taxon-specific clusters (Figs. 5.1 and 5.2). Moreover, three *MAGE-C* genes were likely primate-specific. Five *MAGE-E* and 11 *MAGE-D* genes also showed a clustering pattern (one-to-one orthologous correspondence) similar to that of *MAGE-B* (Fig. 5.1).

All 16 human *MAGE-A* genes were physically clustered into three blocks A, B, and C on the X chromosome (Fig. 5.5*A*). Blocks A and B contain five (*MAGE-A11*, *-A9*, *-A9B, -A8* and *psMAGEA7*) and ten (*MAGE-A4*, *-A5*, *-A10*, *-A6*, *-A2B*, *-A2*, *-A12*, *-A3*, *psMAGEA* and *psMAGEAL*) genes, respectively, whereas block C contains only a single gene (*MAGE-A1*) (Figs. 5.5*B* and *C*). Each of the three blocks contained a palindrome (Fig. 5.5*C*). In block B, most genes (six out of ten) are located on both arms of the palindrome (Fig. 5.5*C*); two nearly identical pairs of genes, *MAGE-A2*/*A2B* and *-A3*/*A6*, were located in symmetric positions on the arms (Figs. 5.5*B* and *C*), and *MAGE-A12* was located in the loop. The phylogenetic relationship among 16 *MAGE-A* genes, including *psMAGEAL* (Fig. 5.5*B*, see Materials and Methods), revealed that five genes in block B were monophyletic, whereas a pair of *psMAGEA*/*psMAGEAL* genes were distantly related to other *MAGE-A* genes. *MAGE-D* was used as the outgroup.

Human block B consisted of seven duplicate units. Each unit was 10–20 kb long and contained a *MAGE-A* and a chondrosarcoma associated gene (*CSAGE*) (Lin *et al*. 2002) (Fig. 5.6A). BLAST analysis of mammalian genomes also revealed the absence of *CSAGE* homologs in non-primate mammals. The palindrome in block B was not observed in non-primate genomes, such as the mouse, dog, or horse genome.

The block B was found in macaques (Fig. 5.6A). This block also contained seven duplicated units, but the form of the palindrome differed between the two species; unlike that observed in humans, a short stem and a large loop structure is expected in macaques (Fig. 5.6*B*). Further, the orthology of units between macaques and humans was curious given their positions. For convenience, we designated the seven duplicate units in block B as *h1* to *h7* in humans and *m1* to *m*7 in macaques (Fig. 5.5*A*) to help the examination of their phylogenetic

relationships (Fig. 5.5*B*). Units *h1*/*h7* each harbored *psMAGEAL* and *psMAGEA* genes, and were orthologous to *m1*/*m7*. Units *h3*/*h5* each harbored *MAGE-A2/A2B* genes and were orthologous to *m5* which harbored *MAGE-A2*; there is no partner to *m5* in macaques. Unit *h4* harbored *MAGE-A12* and was orthologous to *m3*, but unit *m3* did not possess a *MAGE* gene (Fig. 5.6A).

The relationships among *h2*/*h6*, *m*2, *m4*, and *m*6 were somewhat confusing. The *p*-distance between *h2* and *h6* was 0.7 ± 0.2%, and the *p*-distances among *m*2, *m4*, and *m*6 were much higher (12.1%). The distances based on pairwise comparisons of these duplication units between humans and macaques ranged from 8.3 ± 0.5 (%) to 17.7 ± 0.7 (%), and these distances were too large to indicate orthologous relationships. This phylogeny did not support the hypothesis that *m2*, *m4*, or *m6* was orthologous to *h2/h6* (Fig. 5.6C).

To identify orthologous relationships among these duplicated units, cladistic markers such as *SINE*s and *LINE*s were sought using RepeatMasker software (Fig. 5.7, Smit 1996). In general, the arrangements of *SINE*s, *LINE*s, *LTR*s, and short repeats (SRs) in block B were relatively similar between the human and macaque genomes, although there appeared to be a species-specific region. The species-specific region was ~40 kb long in humans, and it extended from the middle of *h2* to *h4*. By contrast, it was ~30 kb long in macaques, and it extended from the middle of *m2* to *m4*. Unlike the result that phylogeny and genetic distances showed (Figs. 5.6C and 5.7), the cladistic markers showed that *h2* and *m2*, which harbor human *MAGE-A6* and e macaque *MAGE3L*, respectively, are indeed orthologous to one another.

5.4.4 Human-specific palindrome and gene conversion

The dot-matrix analysis revealed that the palindrome in block B was apparent only in humans. Although there are sequencing gaps in chimpanzee and orangutan genome data, available sequences showed that the palindrome in block B was less evident in these two apes than in humans. Genes on palindromes may experience frequent gene conversion. Actually, the arms of one palindrome were almost identical based on a window analysis of 500 bp with a non-overlapping interval (Fig. 5.8). Furthermore, the program GENECONV also revealed evidence of gene conversion between arms for the majority of palindromes. However, in the middle of *h2* and of *h6*, there was significantly higher sequence divergence ($p = \sim 2\%$, $P < 0.001$) (Fig. 5.8). The highly diverged regions corresponds to a 673 bp region of the 5' ends of the *MAGE-A3* and *MAGE-A6* genes. To understand the biological significance of this region, we examined the distribution of of epitopes on MAGE proteins that bind HLA class I and II molecules (Fig. 5.9). Most type I MAGE proteins are expressed in tumor cells and have epitopes that bind HLA class I molecules, but some that are expressed in melanoma cells have epitopes that bind HLA class II molecules (van der Bruggen *et al*. 2002; Marsman *et al*. 2005; Crotez and Blum 2009). The highly variable 673 bp region of *MAGE-A3* and *MAGE-A6* specifically encodes peptides that are recognized by HLA molecules (Fig. 5.9). Among 13 amino acid changes between MAGE-A3 and -A6, 10 substitutions are concentrated in this epitope-coding region of these genes. Therefore, the human *MAGE-A3* and *MAGE-A6* genes each encode a protein that can bind to multiple HLA variants (Fig. 5.9).

5.5 Discussion

5.5.1 The ancient origin of *MAGE* genes (phase I)

The origin of *MAGE* genes was ancient; *MAGE* homologs were found in the tunicate and the lancelet genomes. Genes containing sequence that encode an MHD, which is found in MAGE proteins, have also been reported in insects (Põld *et al.* 2000; López-Sánchez *et al*. 2007). In fruit fries (*Drosophila melanogaster*), the gene (*DrmeMAGE*) plays a key role in neurogenesis (Nishimura, Sakoda, Yoshikawa 2008). The gene lacks an intron and could therefore be a processed gene. We searched for a *DrmeMAGE* copy with introns in the fly genome using FlyBase data (http://flybase.org/), but no candidate was found. We also carried out a TBLASTN search over the entire NCBI database. We found that the MHD encoded by *DrmeMAGE* had nearly 30% similarity with the MHDs encoded by vertebrate *MAGE* genes and that one of the epitope-coding regions in the human *MAGE-B16* (FLWGPRAKAE, Põld *et al.* 2000) is identical to the fly MHD. However, *DrmeMAGE* is not expressed in tumor cells and it does not have a function as immunity antigens in the fly. *MAGE* homologs were also found in the *Arabidopsis* genome. *Arabidopsis thaliana MAGE* shares 25% similarity with human *MAGE-A8*, although the function of MAGE in *Arabidopsis* is unknown. Since our study showed that the *MAGE-A* and -*B* subfamilies diverged in eutherians, apparent similarity of *MAGE-A* and -*B* with *MAGE* genes in insects and plants, respectively might result from convergence.

The conservation of phases in exons (Table 5.2) and synteny (Fig. 5.3) indicated that *OrnaMAGEL* is a ‘primary’ ortholog to *MAGE-D2* or -*D3* in humans. The ancestral

MAGE gene was probably most similar to extant *MAGE-D* and was also probably a single-copy gene with introns until the divergence between monotremes and therians. The ancestral *MAGE* gene (*MAGE-D*) was located on an autosome in the stem lineage of mammals; that pair of autosome then evolve into a pair of sex chromosomes, and consequently, the *MAGE-D* gene came X-linked in marsupials and eutherians. The *MAGE-D3* gene encodes trophinin (TRO), is expressed in the placenta, and affects embryo implantation (Sugihara *et al*. 2007); these finding indicate that *MAGE-D3* evolved its current function in eutherians.

The ancestral *MAGE-D* gene was on the proto-X chromosome in the platypus; therefore, the gene may have a gametolog on the extant Y chromosome. However, there are no *MAGE* homologs on the Y chromosomes of humans or other eutherians. The region syntenic to human Xp11 is located near the tip of the opossum X chromosome. However, in many eutherians the regions syntenic to human Xp11 are located near the centromere of the X chromosome. The ancestral region appears to have moved towards the centromere before the radiation of eutherians. This transposition on the X chromosome may have prevented pairing with the Y chromosome, leading to the loss of *MAGE* from the Y chromosome.

5.5.2 Formation of multi-gene families by retrotransposition (phase II)

In eutherians, the *MAGE* gene family comprises 10 subfamilies, and all of those subfamilies, except for the *MAGE-D* subfamily, comprise processed genes. Moreover, it is likely that the ancestral gene of eight subfamilies, *MAGE-A, -B, -E, -F, -H, -L2, NDN* and *NDNL2* subfamilies, were produced *via* retrotransposition events. Chomets *et al*. (2001) proposed that

the source for these eight retrotransposition events was a *MAGE-D* gene. We attempted to confirm this hypothesis using the extent of similarity among CDSs of *MAGE* genes, but the CDS sequences were too short to conclude the ancestry of the processed genes.

At least eight retrotranspositional events may have been necessary to produce ancestors of each of eight extant *MAGE* subfamilies early in eutherian evolution. The activation of reverse-transcriptase necessary for this transposition might have been provided by the activation of *LINE* elements at that time (Kim, Hong, and Rhyu 2004).

To be functional, any processed gene should gain promoter activity near the insertion site. *MAGE-A, -B*, and -*C* are all expressed in cancer cells and in the testis. Sequence similarity beyond the CDS shows that *MAGE-A* and *MAGE-C* were produced by DNA-mediated gene duplication. In addition, the tumor types where *MAGE-A* genes are expressed are similar to those where *MAGE-C* genes are expressed, but different from those where *MAGE-B* genes are expressed (Lurquin *et al.* 1997; Lucas *et al.* 1998; Caballero and Chen 2009). Based on the similarity between *MAGE-A* and *MAGE-C* gene expression, it is expected that the upstream region of *MAGE-A* and –*C* harbor similar TFB sequences and that these TFB were conserved in the upstream region after the gene duplication. In fact, the ~400 bp upstream region of the start codons of both *MAGE-A* and -*C* has potential TFBSs in common. Among several such TFBSs, transducers and activators of transcription factor STAT (TTCCCRKAA) and lymphoid transcription factor LYF (TTTGGGAGR) binding sites, which are known to act in cancer cells, are found (Yu *et al.* 1995; Winandy, Wu and Georgopoulos 1995).

5.5.3 Gene duplication and palindrome formation (phase III)

The high sequence similarity in the 5' flanking regions of *MAGE-A* genes, including possible regulatory, elements and the monophyly of the *MAGE-A* genes in the phylogeny (Figs. 5.1 and 5.4) indicate that *MAGE-A* subfamily members most likely originated from DNA-mediated gene duplication. Nucleotide divergences among members (ranging from 10 to 15%) show that most *MAGE-A* genes emerged in the stem lineage of Catarrhini or even earlier. For this reason, orthologs of *MAGE-A* genes might be present in New World monkeys as well. A database search for such orthologs revealed three sequences on contigs 7129, 6382, and 5036 in the common marmoset genome (*Callithrix jacchus, UCSC* WUSTL version Callithrix jacchus-2.0.2) with greater than 80% similarity to *MAGE-A2*/*A2B*, *A3*/*A6* and -*A12*. Moreover, three additional sequences on contig 880 and one sequence on contigs 1178 and 6382 also showed 76–79% similarity to several human *MAGE* genes. Thus, a total of eight *MAGE-A* homologs were detected in the common marmoset genome. Although the genomic locations of these homologs are not yet known, the duplication events that produced the *MAGE-A* genes probably took place in the stem lineage of simian primates.

It is worth noting that large palindromes on the Y chromosome also originated in the stem lineage of the Catarrhini or even earlier (Bohwmick, Satta, Takahata 2007). The eight palindromes on the human Y chromosome have seven gene families in them. Although nucleotide sequences in symmetrical positions on the palindromic arms are nearly identical, gene family members in asymmetric positions show nucleotide divergences ranging from 5.9 (± 1.0) to 13.9% (± 1.5). This range is similar to those observed between duplicated units in humans or macaques on the X chromosome. Therefore, these gene duplication events on the X and Y chromosome may have occurred simultaneously.

Unusual nucleotide substitutions between the human and macaque sequences on the palindromes require special attention. For example, the comparison between human *MAGE-A3*/*A6* and macaque -*A3*, -*3L*, and -*A3L* genes reveal unusual nucleotide substitutions. The phylogeny of the CDSs of these genes indicates that they diverged in the stem lineage of the Catarrhini (Fig. 5.10), yet synonymous nucleotide differences of human *MAGE-A3*/*A6* and macaque -*A3*, -*3L*, and -*A3L* are exceptionally high (p = 13.4 ± 2.2%, Table 5.3). The degree of functional constraint on newly duplicated genes may change, permitting frequent substitutions in CpG dinucleotides. And substitutions in CpG dinucleotides appear to have occurred in the present case as well. Among 315 codons in these *MAGE* genes, 45 codons contain CpG sites. If the latter codons are excluded, synonymous divergence decreases between human *A3* or *A6* and macaque *A3*, *3L*, or *A3L* to 7.8 ± 2.1% (Table 5.3, ranging from 6.4 ± 1.8% to 9.3 ± 2.4%), which is not significantly different from the average divergence between human and macaque orthologs for genes on the X chromosome (5.5 ± 0.3%) (Elango *et al*. 2009). These results confirm orthology among human *MAGE-A3*/-*A6* and macaque -*A3*, -*3L*, and -*A3L* genes. Importantly, the analysis of syntenic *LINE* and *SINE* insertions also clearly indicates one-to-one orthology between *MAGE-3L* in macaques and *MAGE-A6* in humans (Fig 5.10).

5.5.4 Human specificity in a palindrome (phase IV)

The overall sequence divergence among orthologous duplicate units in humans and macaques exceeds 10%. Since both humans and macaques have seven duplicate units, it is assumed that five pairs of duplicate units had formed a palindrome in the ancestral genome (Fig. 5.11).

Under this assumption, the present arrangement of duplicate units suggests species- or lineage-specific deletions in a loop region of the palindrome (Fig. 5.11).

Further examination of nucleotide divergence between the palindrome arms in humans shows the presence of a significantly diverged region in the middle of *MAGE-A3* and *-A6* (Fig. 5.8). Four synonymous substitutions have accumulated at only CpG sites between *MAGE-A3* and *-A6*, and 22 synonymous ones differentiate the human *MAGE-A6* from the macaque *MAGE-3L*. If these 22 substitutions accumulated the during 35 million years (myr) of divergence between the two species (Takahata 2001; Hasegawa, Thorne, and Kishino 2003; Satta *et al*. 2004), then the accumulation of four substitutions corresponds to 6.4 myr (35 myr × 4/22). This suggests that the divergence between *MAGE-A3* and *-A6* in humans occurred when the human and chimpanzee lineages diverged (7~6 MYA; Burnet *et al*. 2002). Although a one-to-one ortholog to human *MAGE-A6* has not been identified in the chimpanzee genome, chimpanzee *MAGE3* (a one-to-one ortholog to human *MAGE-A3*) apparently encodes a lower variety of epitopes than the human ortholog (Fig. 5.9). It is likely that the nucleotide differences between *MAGE-A3* and *-A6* have accumulated specifically in humans.

These findings lead to questions about the evolutionary forces generating and/or maintaining the diversity observed in human *MAGE-A3* and *-A6*. Two alternative explanations are 1) Darwinian selection elevating nonsynonymous substitutions or 2) negative selection against homogenization by gene conversion. Considering the role of *MAGE-A* proteins in cancer immunity (van der Bruggen *et al*. 1991), a diversity of epitopes might be advantageous, and *MAGE-A* might encode variable epitopes to maintain their ability to bind to HLA molecules, which are diverse. Alternatively, if gene conversion occurs in this

epitope-encoding region, as it does in other regions in arms, the diversity produced by point mutations is erased (Fig. 5.9). One way to prevent homogenization is to invoke negative selection against gene conversion, which may maintain diversity.

To distinguish these two possible causes, we examined relative substitution rates in *MAGE-A3* and -*A6* and their flanking region; we used the *MAGE-A2* sequence as a reference. If Darwinian selection operates in the epitope-coding region, then nucleotide divergence in the epitope-coding region should be higher than in the remaining non epitope-coding region. However, the nonsynonymous substitution rate in nonsynonymous changes between -*A2* vs. -*A3* and -*A2* vs. -*A6* was not elevated. The same result was obtained using different *MAGE-A* genes as reference. Thus, the divergence between *MAGE-A3* and -*A6* was not generated by an elevated nonsynonymous substitution rate. This conclusion was also supported by the comparison of nonsynonymous and synonymous substitutions between *MAGE-A3* and -*A6* ($d_N/d_S = 0.9$, $P < 0.001$). Since the highly diverged epitopes between -*A3* and -*A6* are manifest, negative (purifying) selection against homogenization by gene conversion is more likely. A similar effect of negative selection has been observed in immunoglobulin genes (Nei and Rooney 2005).

5.5.5 Co-evolution between HLA and MAGE epitopes

The operation of negative selection strongly argues for co-evolution between *HLA* and *MAGE-A3* and/or -*A6*. This negative selection of primate *MAGE-A* genes may be associated with rapid turnover of *HLA* class I loci in the primates (Sawai *et al*. 2004).

MAGE-A3 and *A6* encode seven different epitopes that bind HLA class I molecules,

and these epitopes can bind to HLA-A1, -A24, -A2, -B37, -B52, -B44, and -B35 molecule (Fig. 5.9). Curiously, in macaques, there are no corresponding allelic lineages producing these seven major histocompatibility complex (*MHC*: *HLA* homologs in macaques) molecules (data not shown). For this reason, the association between *MAGE* genes and the *MHC* in macaques may be different than the association between *MAGE* genes and *HLA* in humans.

The human-specific genetic diversification between *MAGE-A3* and *-A6* on the palindrome may have been associated with human evolution. Even after the divergence of the human and chimpanzee lineages, the ancestors of humans were arboreal. Subsequently, more recent ancestors left the forests and lived in savanna; even more recent ancestors lost their fur. The change in habitat likely resulted in direct exposure of the naked skin to strong ultra-violet (UV) light. Such exposure is known to increase the risk of tumors such as melanoma. As a means of protection against tumor progression, it is reasonable to imagine that MAGE-A3 and MAGE-A6 proteins with multiple HLA binding sites were favored by natural selection to enable appropriate HLA-mediated immunity.

5.6 Conclusion: Unique mode of evolution in the *MAGE* gene family

There are many gene families in the human genome, and they are generated by DNA-mediated gene duplication and/or retrotransposition. Well-known examples of DNA-mediate gene duplication include emergence of the ribosomal RNA (rRNAs; Fedoroff 1979; Eickbush and Eickbush 2007) and the alpha and beta hemoglobin gene families(Fritsch, Lawn, and Maniatis 1980, Czelusniak *et al*. 1982). In the case of the rRNA genes, the

requirement for a large amount of gene products leads to the multiplication and homogenization of duplicated units. In contrast, sequence divergence of members in the hemoglobin gene family depends on the requirement for physiological differentiation of these proteins. This kind of functional diversification in a multi-gene family is quite common. As discussed in this paper, the multiplication of *MAGE* genes is mediated by both retrotransposition and DNA-mediated gene duplication. Some members of the family have been homogenized by gene conversion, whereas others have been subjected to negative selection against homogenization. The evolution of human *MAGE* genes appears to be determined by the local genomic environment, such as *L1* activity and palindrome formation, and by differentiation in gene function, such as maintaining variation in protein structure to facilitate cancer immunity.

5.7 References

Artamonova, I. I. and Gelfand, M. S. (2004) 'Evolution of the exon-intron structure and alternative splicing of the MAGE-A family of Cancer/Testis Antigens.', *J Mol Evol* 59:620-631.

Bertrand, M., Huijbers, I., Chomez, P. and De Backer, O. (2004) 'Comparative expression analysis of the MAGED genes during embryogenesis and brain development.', *Dev Dyn* 230:325-334.

Bhowmick, B., Satta, Y. and Takahata, N. (2007) 'The origin and evolution of human ampliconic gene families and ampliconic structure.', *Genome Res* 17:441-450.

Bischof, J., Ekker, M. and Wevrick, R. (2003) 'A MAGE/NDN-like gene in zebrafish.', *Dev Dyn* 228:475-479.

Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A. *et al*. (2002) 'A new hominid from the Upper Miocene of Chad Central Africa.', *Nature* 418:145-151.

Caballero, O. and Chen, Y. (2009) 'Cancer/testis (CT) antigens: potential targets for immunotherapy.', *Cancer Sci* 100:2014-2021.

Castelli, C., Rivoltini, L., Andreola, G., Carrabba, M., Renkvist, N. *et al*. (2000) 'T-cell recognition of melanoma-associated antigens.', *J Cell Physiol* 182:323-331.

Chomez, P., Backer, O. D., Bertrand, M., De Plaen, E., Boon, T. and Lucas, S. (2001) 'An overview of the MAGE gene family with the identification of all human members of the family.', *Cancer Res* 61:5544-5551.

Crotzer, V. L. and Blum, J. S. (2009) 'Autophagy and its role in MHC-mediated antigen presentation.', *J Immunol* 182:3335-3341.

Czelusniak, J., Goodman, M., Hewett-Emmett, D., Weiss, M. L., Venta, P. J. *et al*. (1982) 'Phylogenetic origins and adaptive evolution of avian and mammalian hemoglobin genes.', *Nature* 298: 297-300.

Delbridge, M., Patel, H., Waters, P., McMillan, D. and Graves, J. M. (2009) 'Does the human X contain a third evolutionary block? Origin of genes on human Xp11 and Xq28.', *Genome Res* 19:1350-1360.

Eickbush, T. H. and Eickbush, D. G. (2007) 'Finely orchestrated movements: evolution of the ribosomal RNA genes.', *Genetics* 175: 477-485.

Elango, N., Lee, J., Peng, Z., Loh, Y. and Yi, S. (2009) 'Evolutionary rate variation in Old World monkeys.', *Biol Lett* 5:405-408.

Fedoroff, N. V. (1979) 'On spacers.', *Cell* 16: 697-710.

Fritsch, E. F., Lawn, R. M. and Maniatis, T. (1980) 'Molecular Cloning and Characterization of the Human /b-Like Globin Gene Cluster.', *Cell* 19: 959-972

Han, M. V. and Hahn, M. W. (2009) 'Identifying parent-daughter relationships among duplicated genes.', *Pac Symp Biocomput* 114-125.

Hasegawa, M., Thorne, J. and Kishino, H. (2003) 'Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution.', *Genes Genet Syst* 78:267-283.

Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E. *et al*. (1998) 'Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL.', *Nucleic Acids Res* 26:362-367.

Kim, T. M., Hong, S. J. and Rhyu, M. G. (2004) 'Periodic explosive expansion of human retroelements associated with the evolution of the hominoid primates.', *J Korean Med Sci* 19: 177-185.

Kirkin, A. F., Dzhandzhugazyan, K. and Zeuthen, J. (1998) 'The immunogenic properties of melanoma-associated antigens recognized by cytotoxic T lymphocytes.', *Exp Clin Immunogenet* 15:19-32.

Klein, J, and Horejsí, V. (1997) *Immunology*. Oxford: Blackwell Science Ltd.

Librado, P. and Rozas, J. (2009) 'DnaSP v5: a software for comprehensive analysis of DNA polymorphism data.', *Bioinformatics* 25:1451-1452.

Lin, C., Mak, S., Meitner, P. A., Wolf, J. M., Bluman, E. M. *et al*. (2002) 'Cancer/testis antigen CSAGE is currently expressed with MAGE in chondrosarcoma.', *Gene* 1-2:269-278.

López-Sánchez, N., González-Fernández, Z., Niinobe, M., Yoshikawa, K. and Frade, J. (2007) 'Single mage gene in the chicken genome encodes CMage, a protein with functional similarities to mammalian type II Mage proteins.', *Physiol Genomics* 30:156-171.

Lurquin, C., De Smet, C., Brasseur, F., Muscatelli, F., Martelange, V. *et al*. (1997) 'Two members of the human MAGEB gene family located in Xp21.3 are expressed in tumors of various histological origins.', *Genomics* 46:397-408.

Lucas, S., De Smet, C., Arden, K. C., Viars, C. S., Lethé, B. *et al*. (1998) 'Identification of a new MAGE gene with tumor-specific expression by representational difference analysis.', *Cancer Res* 58:743-752.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C. *et al*. (2004) 'Definition of supertypes for HLA molecules using clustering of specificity matrices.', *Immunogenetics* 55:797-810.

Marsman, M., Jordens, I., Griekspoor, A. and Neefjes, J. (2005) 'Chaperoning antigen presentation by MHC class II molecules and their role in oncogenesis.', *Adv Cancer Res* 93: 129-158.

Nei, M., Gu, X. and Sitnikova, T. (1997) 'Evolution by the birth-and-death process in multigene families of the vertebrate immune system.' *Proc Natl Acad Sci USA* 94:7799-7806.

Nei, M. and Rooney, A. 2005. 'Concerted and birth-and-death evolution of multigene families.', *Annu Rev Genet* 39:121-152.

Nishimura, I., Sakoda, J. and Yoshikawa, K. 2008. '*Drosophila* MAGE controls neural precursor proliferation in postembryonic neurogenesis.', *Neuroscience* 154:572-581.

Ohtsubo, Y., Ikeda-Ohtsubo, W., Nagata, Y. and Tsuda, M. (2008) 'GenomeMatcher: A

graphical user interface for DNA sequence comparison.', *BMC Bioinformatics* 9:376.

Põld, M., Põld, A., Ma, H. J., Sjak-Shie, N. N., Vescio, R. A. *et al*. (2000) 'Cloning of the first invertebrate MAGE paralogue: an epitope that activates T-cells in humans is highly conserved in evolution.', *Dev Comp Immunol* 24:719-731.

Rammensee, H., Falk, K. and Rötzschke, O. (1993) 'Peptides naturally presented by MHC class I molecules.', *Annu Rev Immunol* 11:213-244.

Ronquist, F., Huelsenbeck, J. P. and van der Mark, P. (2005) MrBayes 3.1. http://mrbayes.csit.fsu.edu/ondex.php.

Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for reconstructing phylogenetic trees.', *Mol Biol Evol* 4:406-425.

Satta, Y., Hickerson, M., Watanabe, H., O'hUigin, C. and 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.

Sawai, H., Kawamoto, Y., Takahata, N. and Satta, Y. (2004) 'Evolutionary relationships of major histocompatibility complex class I genes in simian primates.', *Genetics* 166:1897-1907.

Sawyer, S. (1989) 'Statistical tests for detecting gene conversion.', *Mol Biol Evol* 6:526-538.

Sonnhammer, E. and Durbin, R. (1995) 'A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.', *Gene* 167:GC1-10.

Stamatakis, A., Ludwig, T. and Meier, H. (2005) 'RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.', *Bioinformatics* 21: 456-463.

Smit, A. F. A. (1996) 'Origin of interspersed repeats in the human genome.', *Curr Opin Genet Devel* 6:743-749.

Sugihara, K., Sugiyama, D., Byrne, J., Wolf, D. P., Lowitz, K. P. *et al*. (2007). 'Trophoblast cell activation by trophinin ligation is implicated in human embryo implantation.', *Proc Natl Acad Sci USA* 104:3799-3804.

Takahata, N. (2001) In: *Humanity from African Naissance to Coming Millennia* (ed. Tobias, P. V., Raath, M. A., Moggi-Cecchi, J., Doyle, G. A.), pp. 299-305. Firenze: Firenze University Press and Johannesburg: Witwatersrand University Press Ltd.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) 'MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.', *Mol Biol Evol* 24:1596-1599.

Thompson, J., Gibson, T., Plewniak, F., Jeanmougin, F. and Higgins, D. (1997) 'The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.', *Nucleic Acids Res* 25:4876-4882.

Tsunoda, T. and Takagi, T. (1999) 'Estimating transcription factor bindability on DNA.', *Bioinformatics* 15:622-630.

van der Bruggen, P., Traversari, C., Chomez, P., Lurquin, C., Plaen, E. D. *et al*. (1991) 'A gene encoding an antigen recognized by cytolytic T lymphocytes on a human melanoma.', *Science* 254:1643-1647.

van der Bruggen, P., Zhang, Y., Chaux, P., Stroobant, V., Panichelli, C. *et al*. (2002) 'Tumor-specific shared antigenic peptides recognized by human T cells.', *Immunol Rev* 188:51-64.

Warburton, P. E., Giordano, J., Cheung, F., Gelfand, Y. and Beoson, G. (2004) 'Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes.', *Genome Res* 14: 1861-1869.

Winandy, S., Wu, P. and Georgopoulos, K. (1995) 'A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma.', *Cell* 83:289-299.

Yu, C., Meyer, D. J., Campbell, G. S., Larmer, A. C., Carter-Su, C. *et al*. (1995) 'Enhanced DNA-binding activity of a Stat3-related protein in cells transformed by the Src oncoprotein.', *Science* 269:81-83.

5.8 Figure Legends

Figure 5.1 Phylogeny of the *MAGE* gene family.

Coding sequences (CDSs) of 158 *MAGE* genes (see Table 1) were used. The CDS compared is 204 bp long without gaps. After alignment, all gaps were excluded for tree construction. Clusters of subfamilies are shown. The bootstrap values indicated refer to branches only. Fish *NDNL2* (*Dare NDNL2*) and mammal *NDNL2* are represented in blue.

The abbreviation for species names are as follows: Bota (*Bos taraus*), Capo (*Cavia porcellus*), Dare (*Danio rerio*), Gaga (*Gallus gallus*), Hosa (*Homo sapiens*), Mamu (*Macaca mulatta*), Modo (*Monodelphis domestica*), Mumu (*Mus musculus*), Orna (*Ornithorhynchus anatinus*), and Patr (*Pan troglodytes*).

Figure 5.2 Schematic representation of the MAGE gene family diversification history.

Each triangle indicates a subtree of the depicted subfamily. The bootstrap values indicated refer to branches only. Branch lengths are arbitrary, and they do not reflect evolutionary distances.

Figure 5. 3 Synteny between platypus contig Ultra 430 and human X chromosome Xp11.

Red bars indicate *MAGE* or *MAGEL* genes in the human or platypus, respectively. Black bars and gene names indicate syntenic genes between human and platypus. Blue bars and gene names indicate genes that are not syntenic. Other *MAGE-D* subfamily members, *MAGE-D1* and *MAGE-D4* are located at 51.6 M and 51.9 M on the human X chromosome, respectively.

Figure 5.4 Phylogeny of MHD in human *MAGE* genes.

The tree is based on the number of amino acid differences per site (*p*-distances). The number of sites compared represent 92 residues without gaps. The bootstrap values indicated refer to branches only. All sequences are listed in Table 5.1. *MAGE-E* has duplicated MHD and the duplication of MHD has occurred earlier than the emergence of typeI genes. *MAGEE1* (*MAGEE2*) and *MAGEE1_2* (*MAGEE2_2*) represent the MHD at the N and C termini of *MAGE-E1* (*MAGE-E2*), respectively.

Figure 5.5 Genomic structure and predicted palindromes in the 5 Mb region encoding the *MAGE-A* genes and a *MAGE-A* gene phylogeny.

(*A*) The 5 Mb region is divided into the three subregions, A, B, and C, which contain five, ten, and one *MAGE-A* gene(s), respectively. (*B*) The tree was constructed using the number of nucleotide differences (*p*-distances) in CDSs (1916 bp) among 16 genes. The bootstrap values indicated refer to branches only. Bootstrap values greater than 50% are shown. Operational taxonomic units (OTU) in magenta, green, and blue represent genes in subregions A, B, and C, respectively. (*C*) Three predicted palindromes are shown, one each in subregion A, B, and C.

In subregion B, most of *MAGE-A* genes occur on palindrome arms.

Figure 5.6 Genomic structures, phylogeny, and predicted palindrome in subregion B.

(*A*) The diagonal line in each panel from the left top to the right bottom indicates identity in self-comparison of human sequences (left panel) and macaque sequences (right panel). Gaps in the diagonal line in the right panel indicate data gaps in the macaque sequence. The colored boxes at the bottom of each panel indicate seven duplicate units. The same colored boxes within a species indicate that the units are more closely related to each other than to other units; similarly, units that are likely orthologs are indicated by shared coloring. (*B*) Predicted palindromes in subregion B were evident human (left) and macaque (right) sequences. The numbers beside the lines indicate each duplicate unit. (*C*) An NJ tree based on *p*-distances between duplicate units (2880 bp) is shown. The color-coding of OTU is the same as that in (*A*) and (*B*).

Figure 5. 7 Maps of cladistic markers in humans and macaques.

Green, light blue, red, dark blue, and purple triangles indicate interspersed elements (*LINEs* or *SINEs*), *LTRs*, DNA transposons (*DNA-TP*), and simple repeats (SR), respectively. Brackets under each line indicate duplicated units. Light pink arrows indicate palindrome structures. The light blue arrow indicates gaps in macaque sequence data. Letters a ~ l and a' ~ i' on the triangles indicate orthologous insertion elements in the human and macaque genomes. The light green bar indicates a human- or macaque-specific region and dotted lines indicate the boundary between species-specific and orthologous regions.

Figure 5. 8 Window analysis to assess nucleotide divergence between a pair of palindrome arms in the human genome.

Bars at the bottom of the figure indicate the locations of duplicate units and the *MAGE* genes therein. The ordinate represents nucleotide divergence (*d*), and the abscissa represents position (bp) relative to the center of the loop (position zero, indicated by a blue arrow).

Figure 5. 9 Alignments of MHD in the primate MAGE-A amino acid sequences. Based on references listed below, the MAGE-A epitopes for HLA alleles in humans are denoted by a square (magenta; human MHC class Ⅰ, light blue; human MHC class Ⅱ). HLA alleles that recognize each epitope are indicated directly below the corresponding amino acid sequence. Of the 13 amino acid substitutions between MAGE-A3 and -A6, 11 are in this region and are marked by stars; only two substitutions (P303L, A308V) were outside of this region. The 10 substitutions that contribute to produce epitopes which are recognized by different HLA alleles (E115K, D156L, L175V, T199A, L201F, V205I, K211R, D249H/D249Y, L279V/L279I, H298R) are indicated by green stars. The other substitution within this region (indicated by a blue star; F239L) on MAGE-A3 and -A6 does not contribute to produce epitopes which are recognized by different HLA alleles.

1. Pold, M., Pold, A., Ma, H. J., Sjak-Shieb, N. N., Vescio, R. A., *et al*. (2000) 'Cloning of the first invertebrate MAGE paralogue: an epitope that activates T-cells in humans is highly conserved in evolution.', *Dev Comp Immunol* 24: 719-731.2.
2. Bredenbeck, A., Losch, F. O., Sharav, T., Eichler-Mertens, M. Filter, M., *et al*. (2005) 'Identificatiof ncanical melama-associated T cell epitopes for cancer immutherapy.', *J Immul* 174: 6716-6724.3.

3. Chianese-Bullock, K. A., Pressley, J., Garbee, C., Hibbitts, S., Murphy, C., *et al*. (2005) 'MAGE-A1-, MAGE-A10-, and gp100-derived peptides are immugenic when combined with granulocyte-macrophage colony-stimulating factor and montanide ISA-51 adjuvant and administered as part of a multipeptide vaccine for melama.', *J Immul* 174: 3080-3086.4.
4. Celis, E., Tsai, V., Crimi, C., DeMars, R., Wentworth, P., *et al*. (1994) 'Inductiof anti-tumor cytotoxic T lymphocytes in rmal humans using primary cultures and synthetic peptide epitopes.', *Proc Natl Acad Sci U S A* 91: 2105-2109.5.
5. Hillig, R., Coulie, P., Stroobant, V., Saenger, W., Ziegler, A., *et al*. (2001) 'High-resolution structure of HLA-A*0201 in complex with a tumour-specific antigenic peptide encoded by the MAGE-A4 gene.', *J Mol Biol* 310: 1167-1176.6.
6. Martura, J,. Longhi, R., Casorati, G. and Protti, M. (2008) 'MAGE-A3(161-175) contains an HLA-DRbeta4 restricted natural epitope poorly formed through indirect presentation by dendritic cells.', *Cancer Immul Immuther* 57: 207-215.7.
7. Graff-Dubois, S., Faure, O., Gross, D., Alves, P., Scardi, A., *et al*. (2002) 'Generatiof CTL recognizing an HLA-A*0201-restricted epitope shared by MAGE-A1, -A2, -A3, -A4, -A6, -A10, and -A12 tumor antigens: implication in a broad-spectrum tumor immutherapy.', *J Immul* 169: 575-580.8.
8. Miyahara, Y., Naota, H., Wang, L., Hiasa, A., Goto, M., *et al*. (2005) 'Determinatiof cellularly processed HLA-A2402-restricted vel CTL epitopes derived from two cancer germ line genes, MAGE-A4 and SAGE.', *Clin Cancer Res* 11: 5581-5589.9.
9. Gotoh, M., Takasu, H., Harada, K. and Yamaoka, T. (2002) 'Development of HLA-A2402/K(b) transgenic mice.', *Int J Cancer* 100: 565-570.10.

10. Groeper, C., Gambazzi, F., Zajac, P., Bubendorf, L., Adamina, M., *et al*. (2007) 'Cancer/testis antigen expression and specific cytotoxic T lymphocyte responses in n small cell lung cancer.', *Int J Cancer* 120: 337-343.11.

11. Novelli, L., Castelli, C. and Parmiani, G. (2005) 'A listing of human tumor antigens recognized by T cells: March 2004 update.', *Cancer Immul Immuther* 54: 187-207.12.

12. Luescher, I., Romero, P., Kuznetsov, D., Rimoldi, D., Coulie, P., *et al*. (1996) 'HLA photoaffinity labeling reveals overlapping binding of homologous melama-associated gene peptides by HLA-A1, HLA-A29, and HLA-B44.', *J Biol Chem* 271: 12463-12471.13.

13. Escudier, B., Dorval, T., Chaput, N., André, F., Caby, M., *et al*. (2005) 'Vaccinatiof metastatic melama patients with autologous dendritic cell (DC) derived-exosomes: results of thefirst phase I clinical trial.', *J Transl Med* 3: 10.14.

14. Vujavic, L., Mandic, M., Olson, W., Kirkwood, J. and Storkus, W. (2007) 'A mycoplasma peptide elicits heteroclitic CD4+ T cell responses against tumor antigen MAGE-A6.', *Clin Cancer Res* 13: 6796-6806.15.

15. Tatsumi, T., Kierstead, L., Ranieri, E., Gesualdo, L., Schena, F., *et al*. (2003) 'MAGE-6 encodes HLA-DRbeta1*0401-presented epitopes recognized by CD4+ T cells from patients with melama or renal cell carcima.', *Clin Cancer Res* 9: 947-954.16.

16. Wang, X., Cohen, W., Castelli, F., Almunia, C., Lethé, B., *et al*. (2007) 'Selective identificatiof HLA-DP4 binding T cell epitopes encoded by the MAGE-A gene family.', *Cancer Immul Immuther* 56: 807-818.

Figure 5.10 The phylogeny of six *MAGE-A* genes from humans and macaques.

The NJ tree is based on synonymous differences among six *MAGE-A* CDSs; 314 sites were compared. The root was determined using *MAGE-A4* as an outgroup. The macaque *MAGE-A3/3L* pair was compared to the human *MAGE-A3*/*A6* pair.

Figure 5.11 Rearrangements in the subregion B.

Schematic diagrams of duplicate units containing *MAGE-A* genes in four extant species and their hypothetical common ancestor are shown. Each colored box indicates a different duplicate unit as in Fig. 5*A*. Gray bars indicate gaps in sequence data. Colored triangles indicate that a deletion occurred at each position independently. Triangles of the same color on the chimpanzee and human diagrams indicates that a deletion occurred in a common ancestor. An arrowhead in each rectangle represents the direction of the fragment.

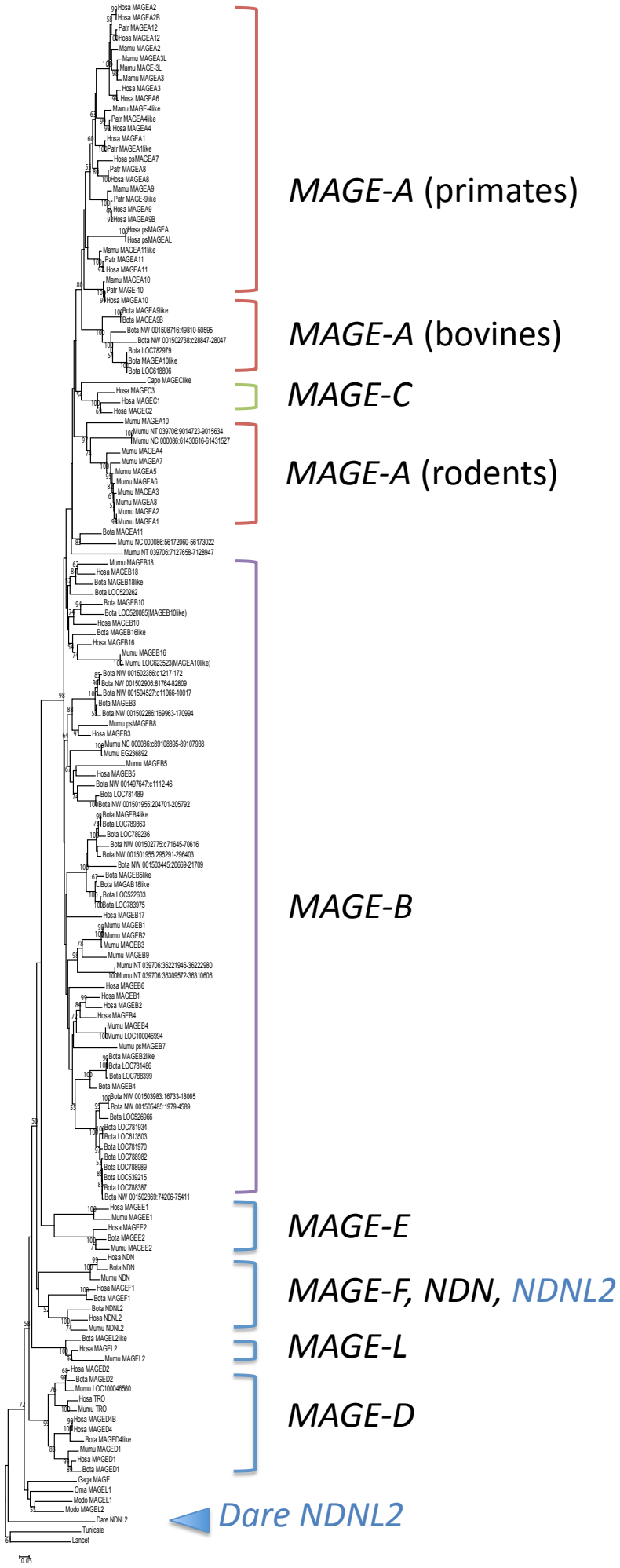


Figure 5.1 Phylogeny of the *MAGE* gene family.

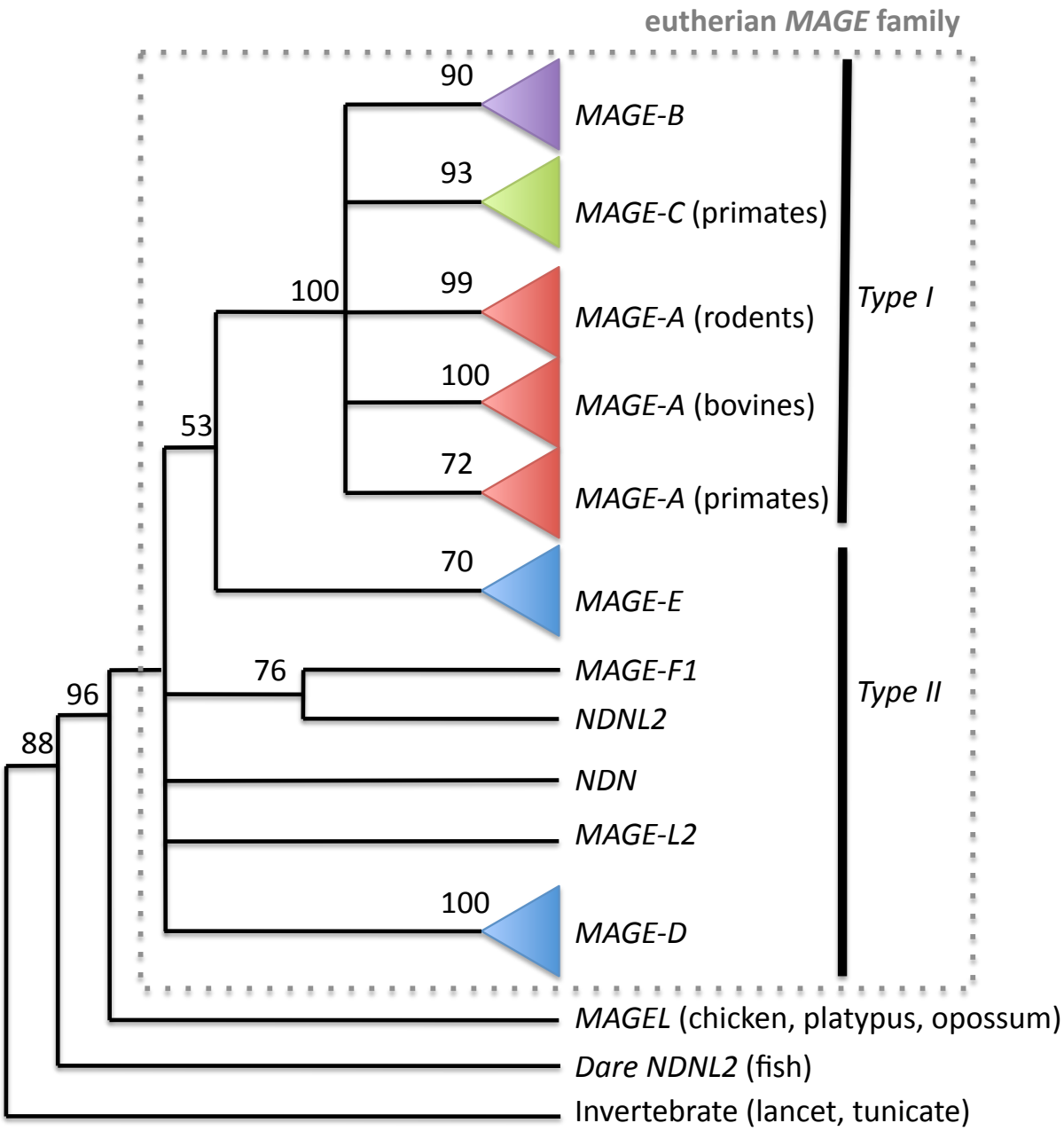


Figure 5.2
Schematic representation of the MAGE gene family diversification history.

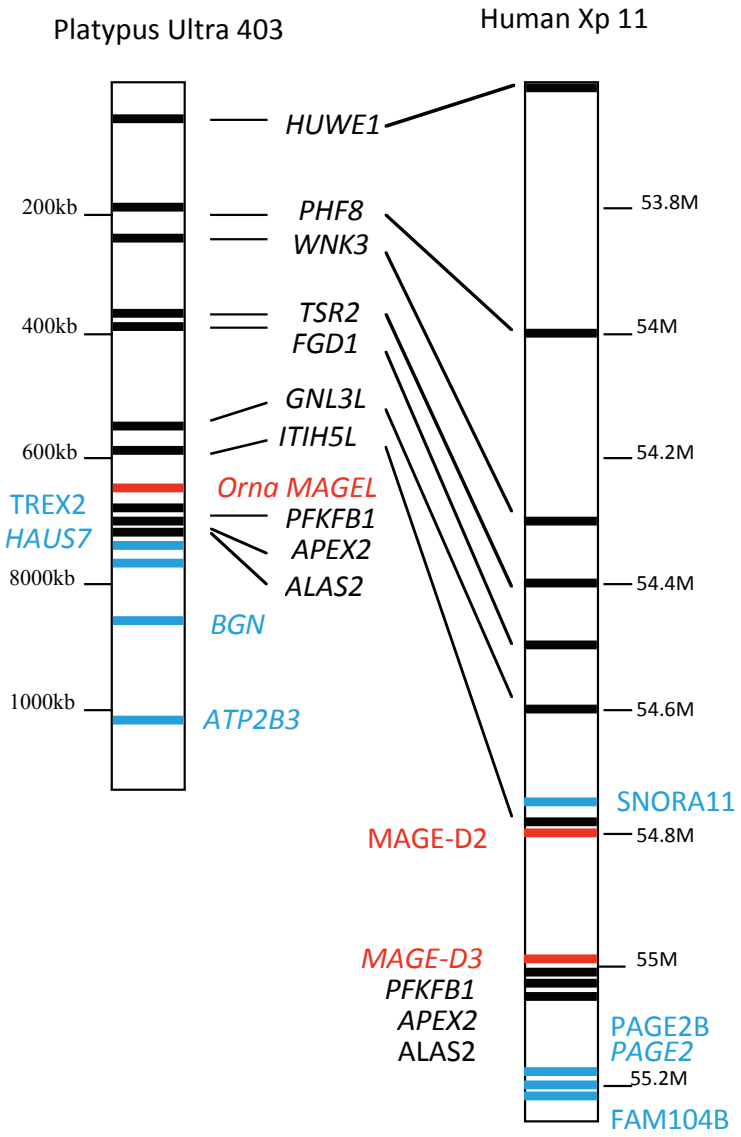


Figure 5.3
Synteny between platypus contig Ultra 430 and human X chromosome Xp11.

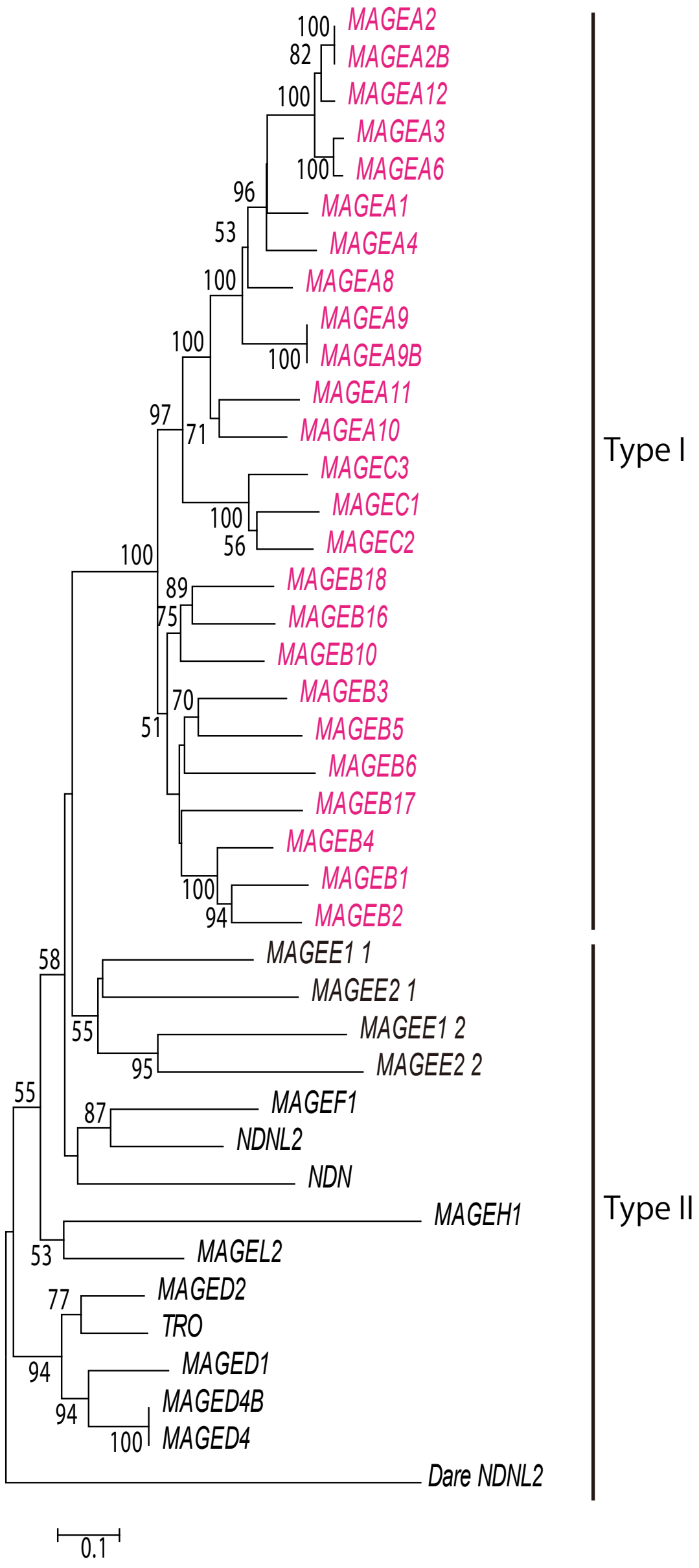


Figure 5.4
Phylogeny of MHD in human MAGE genes.

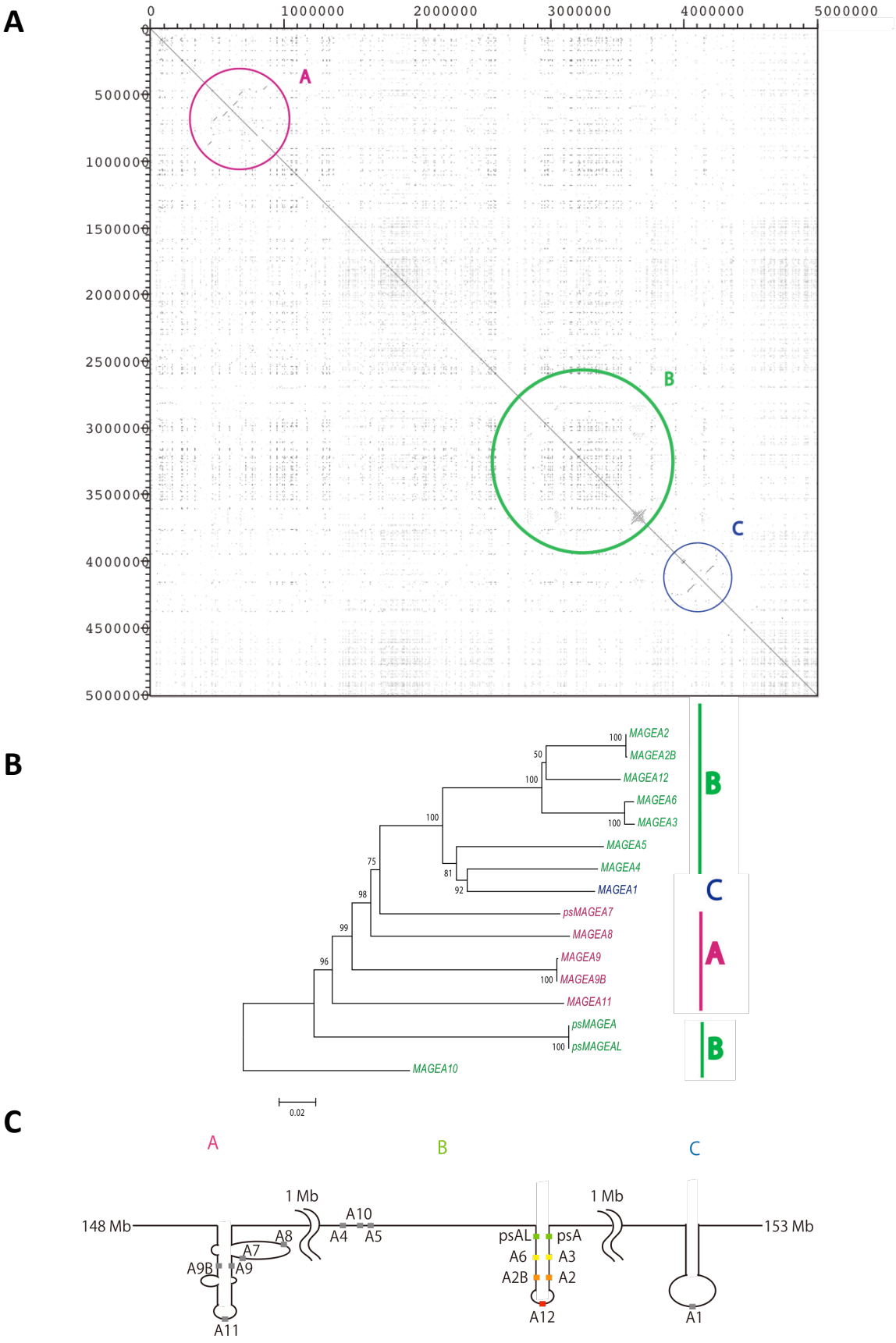


Figure 5.5

Genomic structure and palindrome prediction in the region (5 Mb) encoding *MAGE-A* genes, along with their phylogeny.

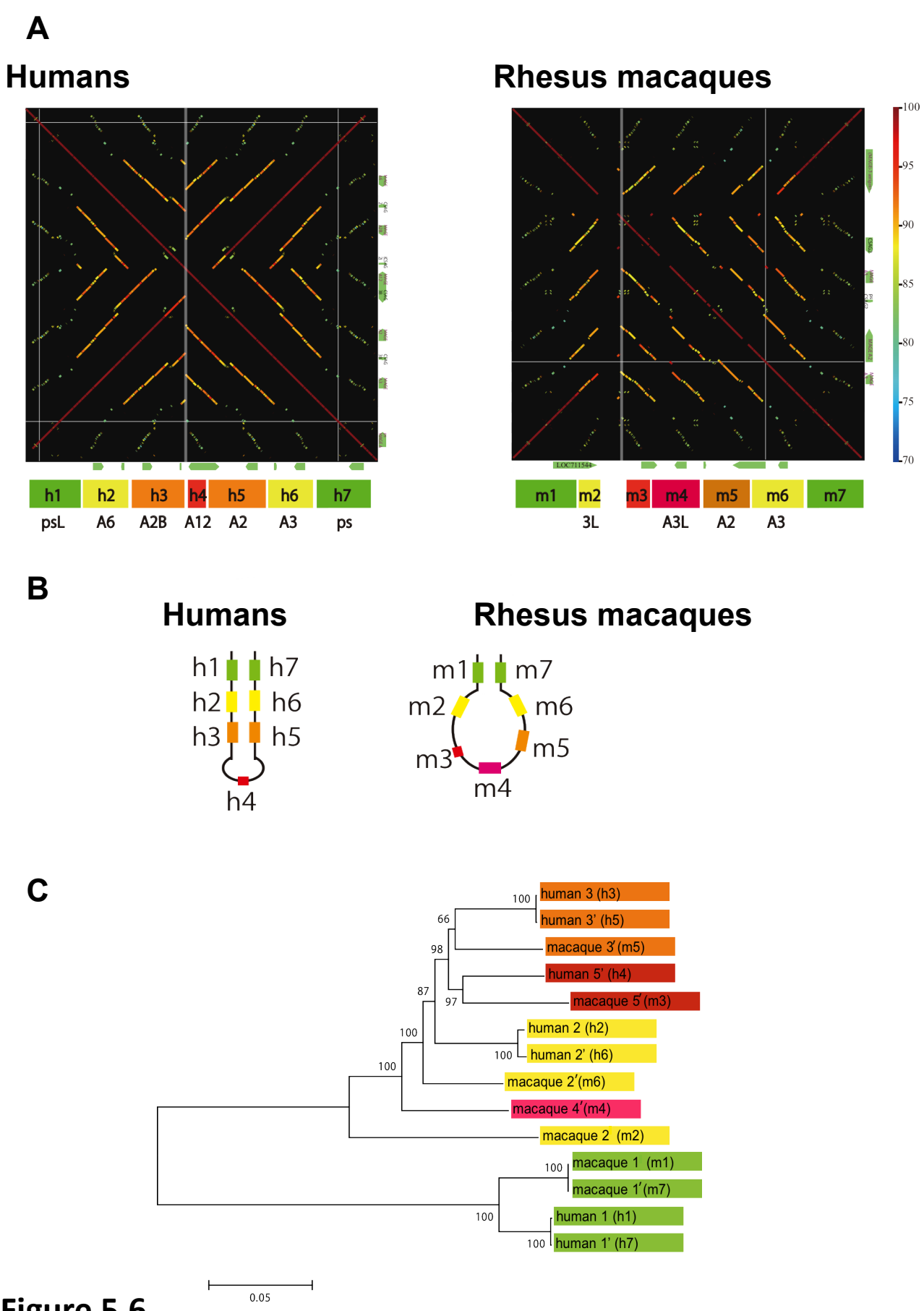


Figure 5.6

Genomic structures, phylogeny and predicted palindromes in subregion B.

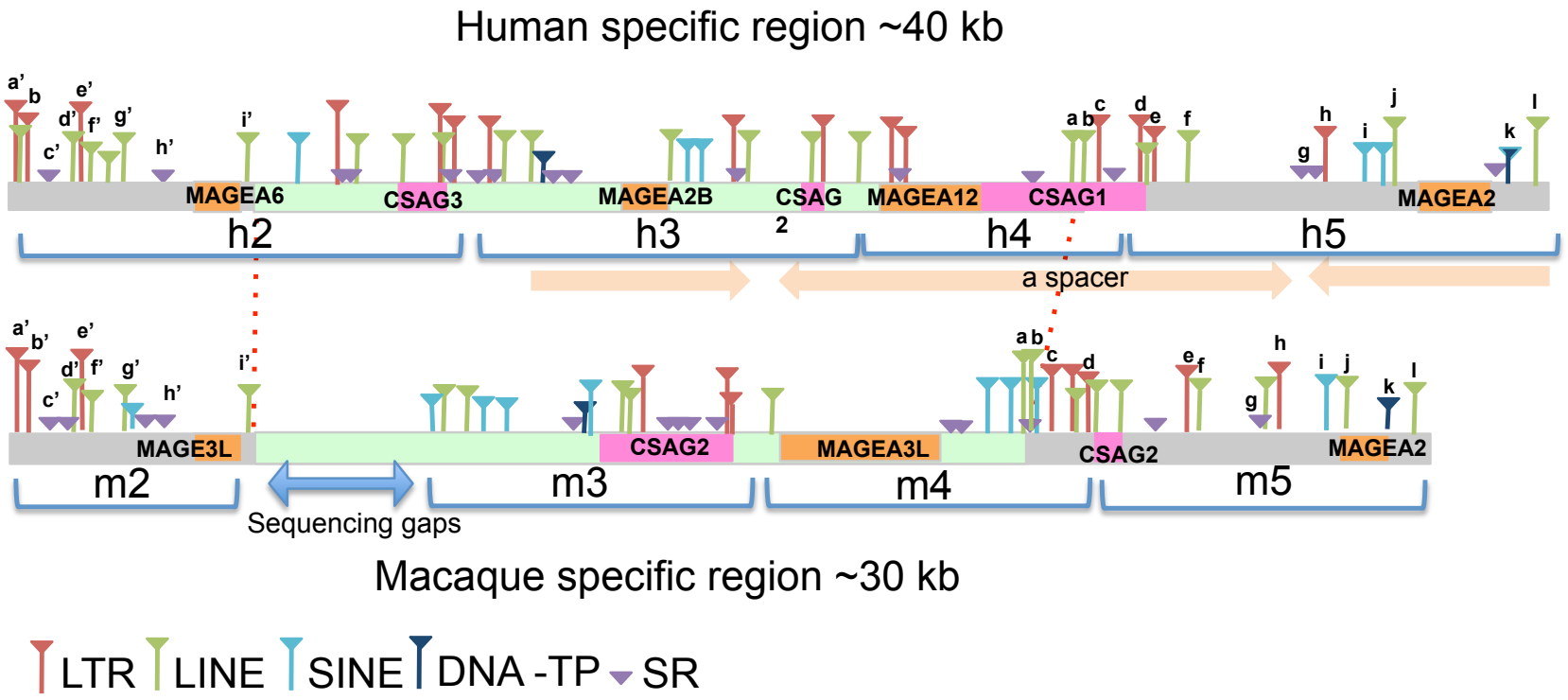


Figure 5.7
Maps of cladistic markers in humans and macaques.

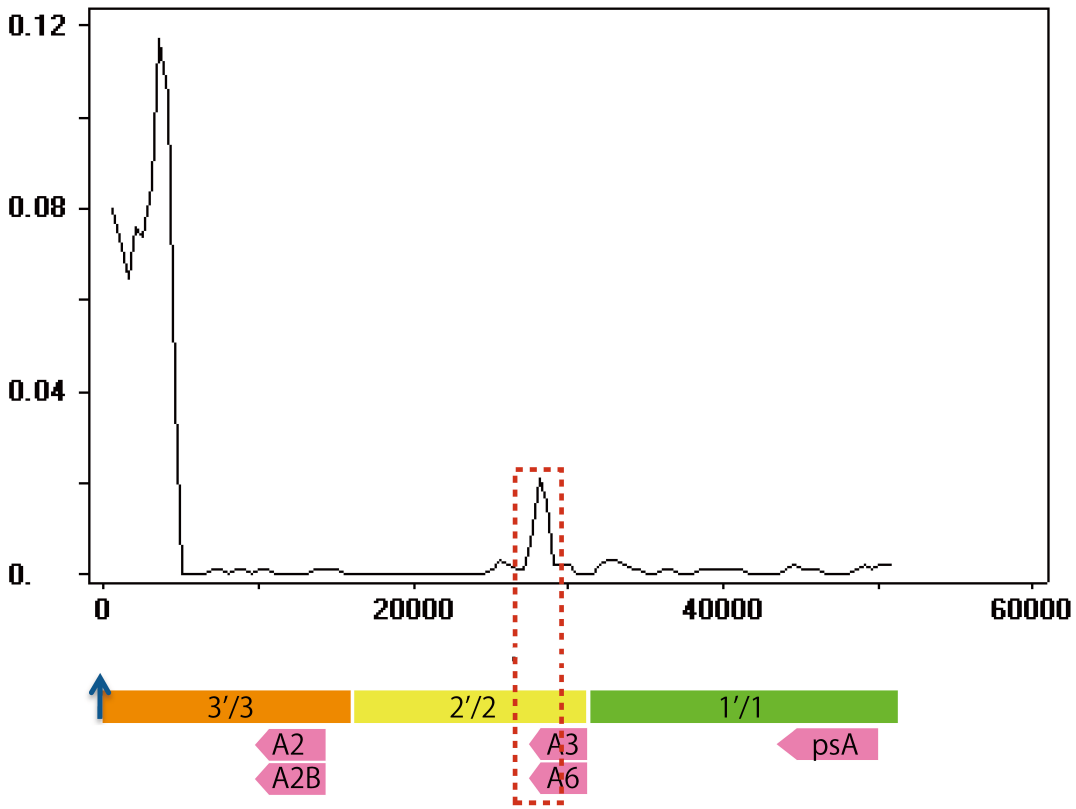


Figure 5.8
Window analysis of nucleotide divergence between a pair of palindrome arms in the human genome.

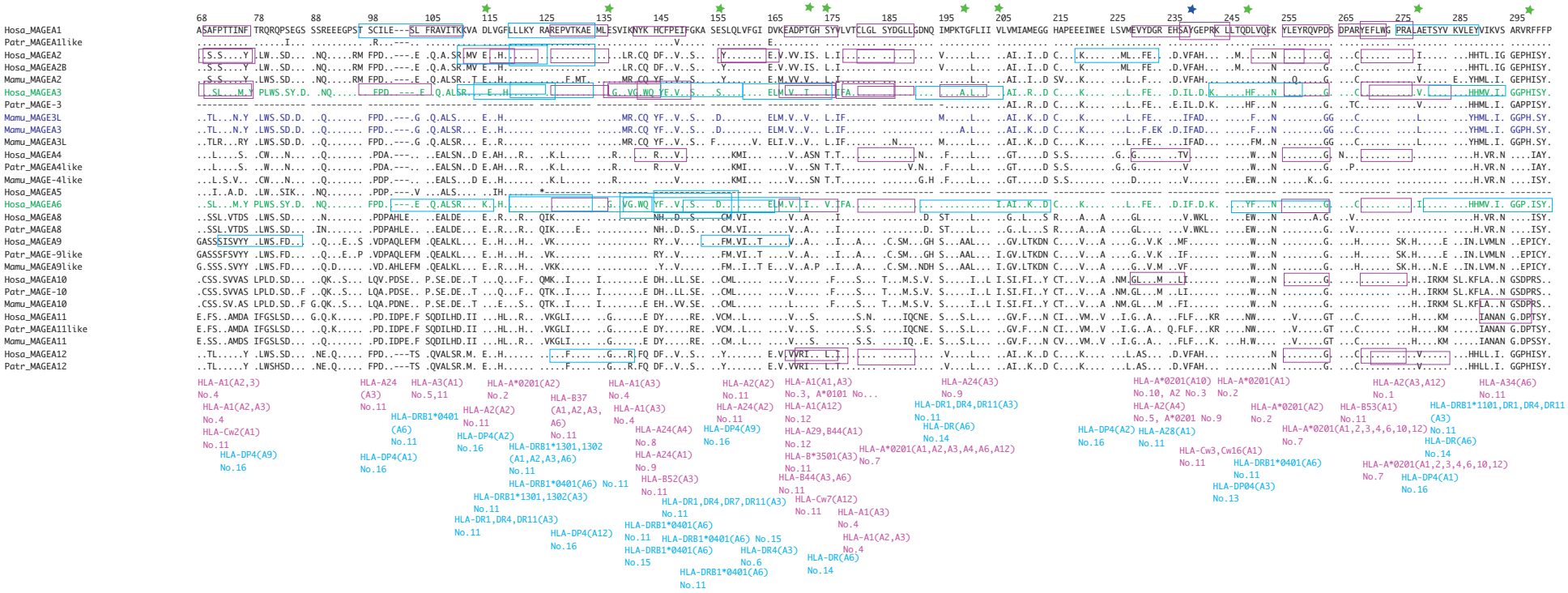


Figure 5.9

Alignments of primate MAGE-A amino acid sequences for an epitope

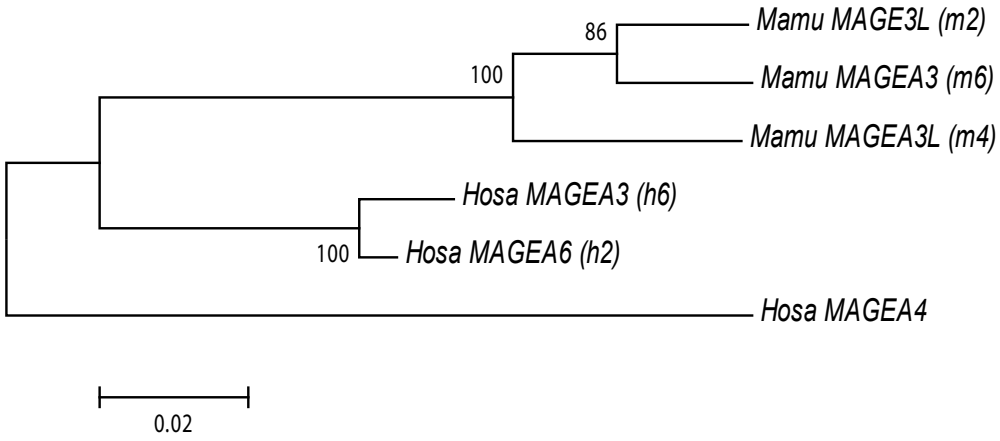


Figure 5.10
The phylogeny of six *MAGE-A* genes from humans and macaques.

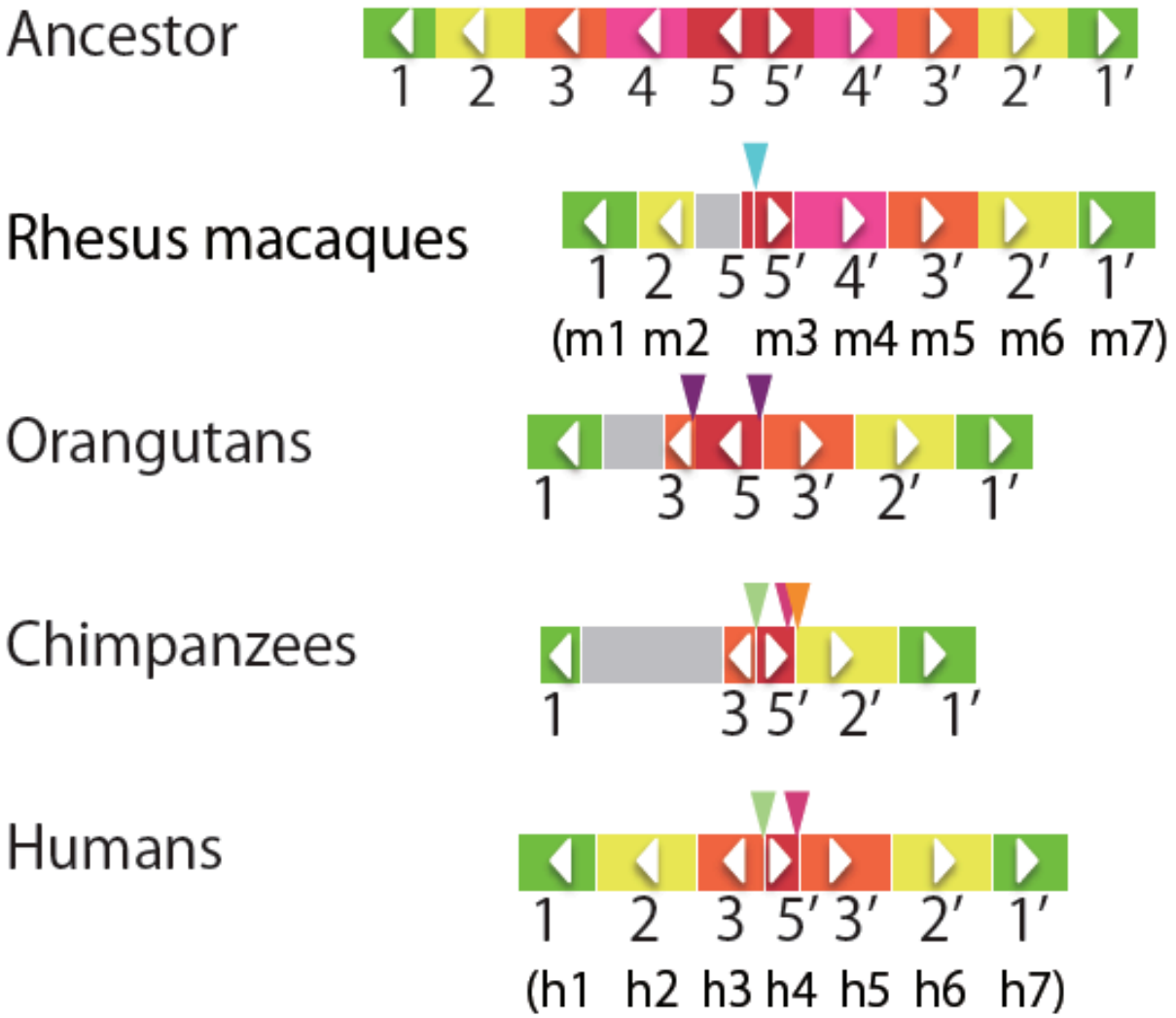


Figure 5.11
Rearrangements in the subregion B.

Table 5.1

Accession numbers of nulceotide sequences used in this study.

humans		chimpanzees		mice		bovines	
MAGEA1	NM_004988	MAGEA1like	XM_529226	MAGEA1	NM_020015	MAGEB5like	XM_001251181
MAGEA2	NM_005361	MAGE-3	XM_001136905	MAGEA2	NM_020016	MAGEA9like	XM_603753
MAGEA2B	NM_153488	MAGEA4like	XM_521309	MAGEA3	NM_020017	MAGEA9B	XM_603753
MAGEA3	NM_005362	MAGEA8	XM_529192	MAGEA4	BC104089	MAGEA10like	NC_007331
MAGEA4	NM_001011548	MAGE-9like	XM_529190	MAGEA5	NM_020018	MAGEA11	NM_001080732
MAGEA5	NM_021049	MAGE-10	XM_521312	MAGEA6	NM_020019	MAGEB16like	XM_586788
MAGEA6	NM_005363	MAGEA11like	XM_521299	MAGEA7	XM_001481307	MAGEB2like	XM_001789276
psMAGEA7	NG_001156	MAGEA12	XM_521314	MAGEA8	NM_020020	MAGEB3like	XM_586930
MAGEA8	NM_005364	**rhesus macaques**		MAGEA9	BC116353	MAGEB4like	XM_001256490
MAGEA9	NM_005365	MAGE-3L	XM_001100355	MAGEA10	NM_001085506	MAGEB10	XM_608078
MAGEA9B	NM_001080790	MAGEA3	XM_001094934	MAGEB1	NM_010759	MAGEB18like	XM_602824
MAGEA10	NM_001011543	MAGEA3L	XM_001094083	MAGEB2	NM_031171	MAGEE2	BT030739
MAGEA11	NM_005366	MAGE-4like	XM_001099496	MAGEB3	NM_008545	MAGEF1	NM_030801
MAGEA12	NM_005367	MAGEA9like	XM_001089793	MAGEB5	BC116773	MAGEH1	NM_001080728
psMAGEA	NC_000023	MAGEA10	XM_001099898	psMAGEB7	NM_001101595	NDN	BT020845
MAGEB1	NM_002363	MAGEA11	XM_001089907	psMAGEB8	NM_001101541	NDNL2	NM_001078080
MAGEB2	NM_002364	**opposums**		MAGEB9	XM_141933	MAGEL2like	XM_581873
MAGEB3	NM_002365	MAGEL1	XM_001373641	MAGEB16	XM_135953	MAGED1	NM_001046125
MAGEB4	NM_001033492	MAGEL2	ENSMODT00000021264	MAGEB18	NM_173783	MAGED2	NM_001075665
MAGEB5	XM_293407	**platypuses**		MAGEE1	NM_053201	MAGED4like	NM_001103311
MAGEB6	NM_173523	MAGEL1	XM_001510461	MAGEE2	BC138210		
MAGEB10	NM_182506	**zeblafishes**		MAGEH1	BC060080		
MAGEB16	XM_001099921	NDNL2	NM_198812	NDN	NM_010882		
MAGEB17	XM_001130425	**guiana pigs**		NDNL2	NM_030801		
MAGEB18	NM_173699	MAGEClike	ENSCPOT00000023751	MAGEL2	NM_019066		
MAGEC1	NM_005462			MAGEL2	BC054763		
MAGEC2	NG_015872			MAGED1	NM_019791		
MAGEC3	NM_138702			TRO	NM_001002272		
MAGEE1	NM_020932						
MAGEE2	NM_138703						
MAGEF1	NM_022149						
MAGEH1	NM_014061						
NDN	NM_002487						
NDNL2	NM_138704						
MAGEL2	NM_019066						
MAGED1	NM_001005333						
MAGED2	BC000304						
TRO	NM_001039705						
MAGED4	NM_001098800						
MAGED4B	NM_030801						

Table 5.2

Phases at exons in the MAGE coding sequence of zebrafishes, African clawed frogs, chickens and mammals.

Exon: [a]	1		2		3		4 (64)		5 (80)		6 (95)		7 (80)		8 (43)		9 (63)		10 (115)		11	
Phase:	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E
zebra fish	–	–	–	0	0	0	0	1	1	0	0	2	2	1	1	2	2	2	2	0	0	–
Frog					0	0	0	1	1	0	0	2	2	1	1	2	2	2	2	0	0	0
Chicken	–	–	–	0	0	0	0	1	1	0	0	2	2	1	1	2	2	2	2	0	0	–
Platypus					0	0	0	1	1	0	0	2	2	1	1	2	2	2	2	0	0	0
Opossum	0	0	0	0	0	0	0	1	1	0	0	2	2	1	1	2	2	2	2	0	0	0
human (*D2*)	–	0	0	0	0	0	0	1	1	0	0	2	2	1	1	2	2	2	2	0	0	–
human (*D3*)	–	0	0	0	0	0	0	1	1	0	0	2	2	1	1	2	2	2	2	0	0	–

a: Only protein coding exons are shown. Numbers in parentheses indicate the size of exons that are conserved from fishes to mammals. Exceptions are exon 6 in opossum and human *D3*; exon size is 98 bp and 92 bp, respectively. S: start, and E: end. Phase information for each species is ENSDART00000081038 for the zebra fish, ENSXETT00000047694 for the frog, DQ983362 for the chicken, NW_001794330 for the platypus, NW_001587054 for the opossum, ENST00000375068 for human *D2*, and ENST00000173898 for human *D3*. S: a phase for a codon at beginning of each exon. E: a phase for a codon at ending of each exon.

Chapter 6

General discussion

6.1 General discussion

In the studies presented in chapter 2-5, I attempted to figure out the evolutionary history of SD systems in mammals. In chapter 2, I showed the rapid evolution and functional differentiation of the primary sex-determination gene, *SRY*, in Theria. In chapter 3, I report on the evolutionary process that gave rise to therian sex chromosomes. In chapter 4 and 5, I present data indicated that the mammalian X-chromosome was subject to rapid evolutionary change, including events such as chromosomal rearrangement, accumulation of segmental duplications, and formation of and diversification within a gene family.

Here, I compare the SD systems in mammals with those in non-mammalian vertebrates. The evolution of primary sex-determination genes is remarkably dynamic in fishes and amphibians, as is the case with the mammalian *SRY* gene. In medaka (*Oryzias latipes*), *DMY* is a primary male-determination factor on the Y chromosome (Matsuda *et al*. 2002). In *Xenopus laevis*, *DM-W* is a primary female-determination factor on the W chromosome (Yoshimoto *et al*. 2008). *DMY* and *DM-W* were duplicated from *DMRT1* independently (Tanaka *et al*. 2007; Bewick, Anderson and Evans 2010).

Chapter 7

General Conclusion

In SD systems, both evolutionary flexibility and stability were observed. The flexibility in SD systems was evident in primary SD genes, which varied significantly among organisms; the biological significance of this variation might be environmental adaptation. The SD genes can maintain plasticity; for example, females or males can be determined by one SD gene because a dysfunction of or deleterious mutation in an SD gene induces sex reversal. Alternatively, the genes on sex chromosome are generally less stable than autosomal genes because of frequent pseudogenization, deletion, retrotransposition, or insertion of transposon, and, therefore, the primary SD gene might be also unstable. Other than the primary SD genes, components in SD systems (or cascades) are evolutionarily stable, this stability indicates that there are strong functional constraints on these systems; the consequence of these constraints may be reliable testicular and ovarian development and gamete production. If a gene at a downstream position in the SD cascade was disrupted by deletion or mutation, the individual would be unviable or infertile in many cases because genes in the cascade have an essential function in gonadogenesis or are pleiotropic.

The segmental duplication or genomic structure on sex chromosomes changed