The Pennsylvania State University

The Graduate School

Eberly College of Science

EVOLUTION OF NATURAL KILLER CELL RECEPTOR GENES

A Thesis in

Biology

by

Li Hao

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2006

The thesis of Li Hao was reviewed and approved* by the following:

Masatoshi Nei
Evan Pugh Professor of Biology
Thesis Advisor
Chair of Committee

Zhi-Chun Lai
Associate Professor of Biology

Wojciech Makalowski
Associate Professor of Biology

Robert F. Paulson
Associate Professor of Veterinary and Biomedical Sciences

Douglas R. Cavener
Professor of Biology
Head of the Department of Biology

*Signatures are on file in the Graduate School

ABSTRACT

Natural killer (NK) cells play important roles in the early immune response against tumor and virus-infected cells. Their activity is controlled by the interaction of the receptors on the NK-cell surface with major histocompatibility complex (MHC) class I molecules. Genes encoding families of NK receptors are organized into two chromosomal complexes: the leukocyte receptor complex (LRC) and the natural killer gene complex (NKC). To understand the diversity and evolution of these different gene families of NK receptors, the following three related studies were conducted. (1) Belonging to the immunoglobulin-like superfamily, the family of killer cell immunoglobulin-like receptors (KIR) is tandemly clustered in the LRC. I examined the evolutionary relationships of primate *KIR* genes and showed that the *KIR* gene family has experienced a rapid expansion in primate species. I also identified the positive selection at the MHC-binding sites (MBS) of *KIR* genes. Estimates of divergence times between duplicate genes suggested that *KIR* genes and their ligand, *MHC* class I genes, coevolved in primates. (2) The *KLRA* (previously known as *Ly49*) gene family in rodents belongs to the C-type lectin-like superfamily but performs the same function as that of *KIR* in primates. They are tandemly clustered in the NKC. I characterized the entire *KLRA* gene cluster in rats and identified a total of 33 genes, which is approximately twice as large as the *KLRA* repertoire in mice. The phylogenetic analysis of rodent *KLRA* genes suggested that this gene family expanded very quickly in rodents and the majority of these genes were generated after divergence of mice and rats. The similarity of the evolutionary pattern between primate *KIR* and rodent *KLRA* genes suggests that their rapid evolution is

caused by the functional divergence of the genes and the interaction with *MHC* class I molecules. (3) In addition to the *KLRA* gene family, the NKC contains a large number of other C-type lectin-like receptor genes which are expressed on NK and other immune-related cells. I characterized all the C-type lectin-like NKC genes and their organization from four major orders of placental mammals, primates (human), rodents (mouse and rat), carnivores (dog), and artiodactyls (cattle). Phylogenetic analysis of these genes indicated that the genes within the NKC are highly heterogeneous with respect to the rate of birth-and-death evolution within and between different mammalian species, but the NKC is also remarkably conserved in the gene organization and persistence of orthologous genes. Searching for putative NKC sequences in opossum and chicken genomes suggested that the expansion of the NKC gene families might have occurred before the radiation of placental mammals but after the divergence between birds and mammals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This graduate thesis marks not only a great achievement of my life, but also an end to a very special seven years of my life. At times it was difficult and it felt as if I would never finish. However, the support and encouragement of the many people in my life allowed me to reach my goal. I would like to take this opportunity to individually thank everyone involved.

Most of all I would express my deepest gratitude to my research advisor, Dr. Masatoshi Nei, for creating the opportunity for me to work in his lab, for giving me room to explore and learn, and most importantly for his continuous guidance, patience, and encouragement during these years. A very special thanks goes out to Dr. Jan Klein. His invaluable support and generous help, especially in my last project, made my dissertation possible. I am also very grateful to Dr. Robert Paulson, who was my advisor in the first year and continued to be my thesis committee. I want to thank him for his understanding and support throughout the period of my graduate program. My thanks also go to the other members of my committee, Drs. Zhi-chun Lai and Wojciech Makalowski for their insightful advice and warm-hearted encouragement.

I also want to thank many former members of the Dr. Nei's lab for their great help and discussion. They are Jongmin Nam, Yoshihito Niimura, Nikolas Nikolaidis, Helen Piontkivska, Kazuharu Misawa, Yoshiyuki Suzuki, Takeshi Itoh, Galena Glazko.

Last, but not the least, I would like to thank my wonderful family: my parents, Yangao Hao and Meiju Zhang, my younger sister, Jing Hao, my parents-in-laws, Pingxiong Wei and Furong Yu, my son, David, and in particular, my dear husband, Xin

Wei. Without their never ending love, encouragement, and support, I would not have finished this thesis.

CHAPTER 1

INTRODUCTION

Vertebrates are armed with two immune systems: innate and adaptive. The adaptive (acquired) immune system recognizes specific foreign pathogens and provides a life-long immunity to them after an initial attack (Hoffmann et al. 1999; Goldsby, Kindt, and Osborne 2000). The adaptive immune system is confined to jawed vertebrates. The three most important classes of molecules for adaptive immunity are the major histocompatibility complex (MHC) molecules, the immunoglobulins (Igs), and the T cell receptors (TCRs). The last two classes of proteins display a tremendous amount of diversity and can recognize a large number of different foreign antigens, which are presented by the highly polymorphic MHC molecules. Evolution of these three types of molecules has been studied extensively (e.g, Kasahara 2000; Ota, Sitnikova, and Nei 2000; Su and Nei 2001). In jawless vertebrates, a different type of immunoreceptors has been identified in the lamprey (Alder et al. 2005). It contains leucine-rich repeats (LRRs) rather than immunoglobulin domains as do the MHC, Ig, and TCR of jawed vertebrates. The remarkable diversity of the LRR-based receptors suggests that they function as lymphocyte antigen receptors like the Igs and TCRs.

The evolution of the innate immune system is not well understood. The innate immune system, as the first line of protection, provides broad and rather nonspecific host defenses, which can limit the attack of pathogens very rapidly. The system consists of highly diverse defense mechanisms, including physical barriers such as skin, inflammation molecules in the blood, complement proteins, and cells that engulf foreign antigens in the body. Plants and invertebrates do not have the adaptive immune system and exclusively rely on the innate immunity to protect themselves from the

attack of different pathogens. Some molecules involved in the innate immune system are highly conserved and are shared between vertebrates and invertebrates. For example, Toll molecules which, as the pattern recognition receptors (PRRs), recognize certain conserved pathogen-associated molecular patterns (PAMPs) and play a central role in the defense against different kinds of pathogens (Beutler 2004). Vertebrates have their unique innate defense mechanisms. One of them is natural killer (NK) cells that are present only in jawed vertebrates.

NK cells are developmentally related to T-cells and represent a population of large cytotoxic lymphocytes containing large granules in the cytoplasm. They are essential for the early response against tumor, virus-infected, and transplanted cells (Cerwenka and Lanier 2001). The cytotoxic activity of NK cells is controlled by the interaction of their cell surface receptors with MHC or MHC-related molecules. NK cells have also been shown to be capable of producing a variety of cytokines, which are involved in the regulation of the adaptive immune system. A large number of different receptors have been identified on the surface of NK cells (Kelley, Walter, and Trowsdale 2005). On the basis of their structure, the NK cell receptors (NKRs) can be classified into two groups: the immunoglobulin-like receptor superfamily (IgSF) and the calcium-independent C-type lectin-like receptor superfamily (CLSF). The genes encoding the IgSF and CLSF-type NKRs are located in different genomic regions. In humans, for example, the IgSF genes are clustered in a well-defined gene complex called the leukocyte receptor complex (LRC) on chromosome 19q13.4, whereas the CLSF genes are clustered in a genomic region called the natural killer gene complex

(NKC) on chromosome 12p13. Two representative NK cell receptors from the above two groups are killer-cell immunoglobulin-like receptors (KIRs) and killer-cell lectin-like receptors (KLRAs, also called Ly49), respectively. Although encoded by different gene superfamilies, they have essentially the same function and are considered as functional equivalents (Barten et al. 2001). Interestingly, KIRs are used primarily in primates, and KLRAs are used in rodents. Therefore, KIR and KLRA receptors represent an interesting case of functional convergence at the molecular level. How these different NK receptor gene families evolved in primates and rodents is unclear.

The *KIR* and *KLRA* gene families are highly diversified in primates and rodents, respectively. The *KIR* member genes in humans are tandemly clustered in the LRC, and the mouse *KLRA* genes are again tandemly clustered in the NKC. One of the most interesting features shared by the primate *KIR* and rodent *KLRA* gene families is their high degree of gene diversity. This diversity of *KIR* and *KLRA* genes is caused by multiple genes, multiple alleles, multiple haplotypes, and multiple splicing sites (e.g.,Wilson et al. 2000; Toneva et al. 2001; Makrigiannis et al. 2002; Wilhelm, Gagnier, and Mager 2002). For example, in the human population, there is a large number of *KIR* haplotypes and hundreds of different (diploid) *KIR* genotypes. The number of *KIR* genes varies considerably with haplotype. For *KLRA* genes, there is also substantial variation in gene number and gene content among different mouse strains. The expansion of genes in the *KIR* and *KLRA* families is caused primarily by repeated gene duplication. The rate of duplication of these genes seems to be

exceptional high among mammalian gene families. However, no extensive study of the evolution of these genes has been conducted.

The main purpose of my research is to examine the intraspecific and interspecific diversity of these two gene families and to study how rapidly they expanded during the evolution of primates and rodents, respectively. I have also been interested in studying the similarity and dissimilarity of the evolution of *KIR* genes and *KLRA* genes. Since these two families of NKRs are functionally equivalent, this comparison might shed some light on the relationship between function and evolution,.

In addition to KLRAs, many other C-type lectin-like receptors are located in the NKC, including KLRC, KLRD, CLEC2D, CD69, OLR1, and KLRB. Many of them are expressed on NK cells and are known as NKRs (Trowsdale et al. 2001). For example, KLRC (previously called NKG2A, C, E, and F) receptors form heterodimers with KLRD (previously known as CD94) and thus monitor the expression level of MHC class I molecules of the cells by recognizing the non-classical MHC class I molecule, HLA-E (O'Callaghan 2000). Some of the other lectin-like receptors, such as CLEC4, CD69, and OLR1, have been found to be expressed on a broader range of immune cells, including dendritic cells, although their function remains to be clarified. The comparison of NKC genes between different species suggests that the genomic structure of the NKC region might be conserved through the entire mammalian evolution, while the number of member genes of some of the gene families is highly variable among different species. As the genome sequences of more vertebrates are becoming available, it is possible to

examine the diversity and evolution of the C-type lectin-like genes in the NKC more extensively.

For these purposes, I have conducted three related studies to gain insights into the evolution of the diverse NKR genes in vertebrates. These three studies are as follows: (1) the expansion of *KIR* genes in primates and their coevolution with *MHC* Class I genes, (2) the genomic organization of rat *KLRA* gene cluster and the evolutionary analysis of *KLRA* genes in rodents, and (3) the evolutionary relationships of C-type lectin-like genes in the NKC and their evolutionary dynamics. These three sets of studies are discussed in detail in the following chapters.

CHAPTER 2

RAPID EXPANSION OF KILLER CELL IMMUNOGLOBULIN-LIKE RECEPTOR GENES IN PRIMATES AND THEIR COEVOLUTION WITH MHC CLASS I GENES

SUMMARY

The gene family of killer cell immunoglobulin-like receptors (KIRs) in primates provides the first line of defense against virus infection and tumor transformation. Interacting with MHC class I molecules, KIRs can regulate the cytotoxic activity of natural killer (NK) cells and distinguish the tumor and virus infected cells from normal body cells. Phylogenetic analysis and comparison of domain structures identified three major groups of *KIR* genes (group I, II, and III genes). These groups of *KIR* genes, generated by a series of gene duplications, have acquired different MHC-binding specificity. Inference of ancestral *KIR* sequences suggested that the functional divergence of group I genes from group II genes occurred by positive selection at the MHC-binding sites after duplication. Our evolutionary study has shown that group I genes diverged from group II genes about 17 million years ago (Mya) apparently after separation of hominoids from Old World (OW) monkeys. Around the same time, gene duplication generating the class I *MHC-C* locus appears to have occurred. These findings suggest that *KIR* and *MHC* class I genes have coevolved as an interacting system. The *KIR* gene family has experienced a rapid expansion in primate species. The rate of expansion of this gene family seems to be one of the highest among all hominoid gene families. The *KIR* gene family is also subject to birth-and-death evolution.

INTRODUCTION

Natural killer (NK) cells are a critical component of the innate immune system. They are essential for the early immune response against tumor and virus-infected cells. The cytotoxic activity of NK cells is regulated through the interaction of the receptors on the NK-cell surface with major histocompatibility complex (MHC) class I molecules. It has been shown that a large number of different receptors are expressed on the surface of NK cells (see review by Kelley, Walter, and Trowsdale 2005). On the basis of the protein structure, the NK cell receptors can be classified into two groups: the immunoglobulin-like receptor superfamily (IgSF) and the calcium independent C-type lectin-like receptor superfamily. The family of killer cell immunoglobulin-like receptors (KIR) belongs to the IgSF family, and in humans the member genes of the family are located tandemly in a genomic region called the leukocyte receptor complex (LRC) on chromosome 19q13.4. Many other genes belonging to the IgSF family are also located in the LRC.

Interestingly, primates and rodents use different NK cell receptors (KIRs for primates and KLRAs for rodents) to regulate the activity of NK cells, although they have a similar signaling pathway (Kelley, Walter, and Trowsdale 2005). The *KIR* gene family is composed of many member genes, and these member genes are highly diversified in hominoids and Old World (OW) monkeys (e.g., Khakoo et al. 2000; Hershberger et al. 2001; e.g., Guethlein et al. 2002). In cattle, multiple *KIR* genes and a single *KLRA* gene have been identified (McQueen et al. 2002; Storset et al. 2003).

The *KIR* genes appear to be absent in rodents except one or two presumably nonfunctional genes (Hoelsbrekken et al. 2003).

There is a high degree of diversity in primate *KIR* genes, which is generated by multiple genes, multiple alleles, multiple haplotypes, multiple splicing sites, and multiple recombinations (e.g., Toneva et al. 2001; Rajalingam, Parham, and Abi-Rached 2004). The expansion of gene members is caused primarily by repeated gene duplication (Martin et al. 2004). Human and chimpanzee *KIR* genes are quite different from each other (Khakoo et al. 2000). Examination of *KIR* genes in other primates also showed a pattern of species-specific diversification (Hershberger et al. 2001; Guethlein et al. 2002). The rapid evolution of high diversity of primate *KIR* genes seems to be exceptional if we consider their recent origin of duplicate genes.

Parham (1997) proposed a 'catch-up' model of evolution, in which *KIR* genes evolve rapidly because of the pressure for catching up with the evolutionary changes of their ligands, highly polymorphic MHC class I molecules. Hughes (2002) studied the evolution of human *KIR* genes using Ig-like domains as the evolutionary units and identified positive Darwinian selection in one of the Ig-like domains. However, it remains unclear how *KIR* and *MHC* genes have coevolved in the presence of their interaction. It is also interesting to investigate whether the positive selection plays any role in the coevolution.

The purpose of this study is three-fold: (1) investigation of the coevolution between *KIR* and *MHC* genes, (2) examination of possible involvement of positive selection on the evolution of human *KIR* genes, and (3) estimation of the divergence

time of different lineages of primate *KIR* genes to determine the rate of expansion of *KIR* genes by gene duplication. Because we are primarily interested in the long-term evolution of the *KIR* gene family, we used genes rather than individual Ig-like domains as the units of evolution.

MATERIALS AND METHODS

Background information

In the human population, there are a large number of haplotypes and hundreds of different (diploid) genotypes (e.g., Trowsdale et al. 2001; Hsu et al. 2002). The number of *KIR* genes varies considerably with haplotype. Eight major haplotypes are presented in figure 2.1. The first haplotype (H1) with a single activating receptor gene (*2DS4*) has the highest frequency (~50%) in the Caucasian population. The other haplotypes have at least two activating receptor genes (represented by 'S'; see below) (Hsu et al. 2002). Two pairs of *KIR* sequences, *3DL1*/*3DS1* and *2DL2*/*2DL3*, are likely to represent different alleles, because they are always located on the same genomic regions.

In this study the standard nomenclature of *KIR* genes was used (Marsh et al. 2003), and it is related to their organization of Ig-like domains (figure 2.2). Each KIR molecule contains the extracellular, transmembrane, and cytoplasmic regions. There are two or three Ig-like domains in the extracellular region. Each *KIR* gene is denoted by four alphanumerics. The first two letters (2D or 3D) represent the number of Ig-like domains of the molecule. In some KIR molecules, the cytoplasmic region contains one or two immunoreceptor tyrosine inhibitory motifs (ITIM). The KIRs with ITIM are inhibitory receptors, and they have the ability to inhibit the cellular activity when they bind to ligands. By contrast, the activating receptors do not contain ITIM. Instead, they recruit an adaptor protein through a charged amino acid in the transmembrane region and contribute to the activation of NK cells. 'L' denotes

long cytoplasmic tail and is most likely to represent an inhibitory receptor with ITIM, whereas 'S' represents an activating receptor without ITIM. The letter 'P' represents a pseudogene. Primate KIRs are divided into four subgroups (3DL, 3DS, 2DL, and 2DS) according to the number of Ig-like domains and the presence or absence of ITIM (figure 2.2). The last digit of the nomenclature designates the order of identification of the gene in each subgroup. The three Ig-like domains in subgroups 3DL and 3DS are designated as D0, D1, and D2, starting from the N-terminal (figure 2.2). The KIRs in subgroups 2DL and 2DS contain domains D1 and D2, except *2DL4* and *2DL5* which have domains D0 and D2 instead.

Phylogenetic analysis

All the sequence alignments were obtained by using computer program CLUSTAL X (Thompson et al. 1997). Additional modifications were done by visual inspection. Phylogenetic analysis was conducted by using the computer program MEGA2 (Kumar et al. 2001). We constructed phylogenetic trees for human genes and for primate genes separately by the neighbor-joining (NJ) method (Saitou and Nei 1987) after elimination of all alignment gaps (complete deletion option). We used p-distance (proportion of uncorrected nucleotide differences) to construct phylogenetic trees because p-distance has a smaller variance than other distances and often gives a better resolution of the topology (Nei and Kumar 2000). In the estimation of divergence times, however, we used the Jukes-Cantor, Kimura's 2-parameter, and Tamura-Nei distances to take care of multiple substitutions and the transition/transversion and GC content biases. However, the results obtained by these

three distance measures were virtually identical because the extent of sequence divergences was small and there were not much transition/transversion and GC content biases. We therefore present only the results obtained by Kimura's distance in this study. To examine the reliability of topologies generated by the NJ method, we also constructed MP trees using PAUP* 4.0 (Swofford 1998). For the dataset of human *KIR* genes, the branch-and-bound search was used with 1000 bootstrap replications. For the primate *KIR* genes, we used the standard stepwise addition and TBR search with 500 bootstrap replications. However, since the major pattern of the topology of the parsimony tree was essentially the same as that of the NJ tree, we presented only the NJ tree in this study.

The amino acid and nucleotide sequences of all ancestral nodes of the phylogenetic tree were inferred from the present-day sequences using the programs ANCESTOR and ANC-GENE (Zhang and Nei 1997; http://mep.bio.psu.edu/).

When we studied the coevolution of *KIR* and class I *MHC* genes in humans, we used 15 human *KIR* sequences excluding pseudogenes *2DP1* and *3DP1*. Two pairs of potential *KIR* alleles (*3DL1*/*3DS1* and *2DL2*/*2DL3*) were included in this study, because they are functionally different and we wanted to include every sequence to identify the potential positive selection and functional coevolution between *KIR* and *MHC* genes.

In the study of the rate of gene family expansion, we used 47 *KIR* sequences from humans (10), common chimpanzees (6), pygmy chimpanzees (3), gorillas (7), orangutans (6), and macaques (15) after the following data filtering processes. (a) We

excluded the *KIR* genes (*2DL4* and *2DL5*) that lacked the D1 domain because inclusion of these genes decreased the number of informative nucleotide sites in our phylogenetic analysis (complete-deletion option). (b) The distinction between loci and alleles was not always clear for non-human primate *KIR* sequences. Because we are interested in the estimation of gene expansion rate, the allelic sequence should not be included. We used only one of the potentially allelic sequences, which had a pairwise p-distance of less than 0.01. This criterion is conservative because the average nucleotide difference between different alleles in humans is about 0.003 (Li and Sadler 1991). For human *KIR* sequences, *2DL3* and *3DS1* were excluded because they are the alleles of *2DL2* and *3DL1*, respectively.

In our phylogenetic analysis, we included cattle *KIR* genes and recently identified rodent *KIR*-like sequences. The names of the sequences in the primate dataset and their GenBank accession numbers are given in table 2.1. The letter 'H' in the name of some macaque *KIR* genes represents a hybrid molecule, in which the extracellular region is more similar to those of subgroup 3DL genes, whereas the cytoplasmic region has only one ITIM that is similar to that of *2DL4*.

RESULTS

Evolutionary relationships of *KIR* and *MHC* class I genes

The NJ tree for 15 human *KIR* sequences is presented in figure 2.3. Since this NJ tree is based on only the D2 domain, transmembrane, and cytoplasmic regions because of the complete deletion option used, we constructed another NJ tree using the nucleotide sequences of domain D1 only (figure 2.4). Genes *2DL4*, *2DL5A*, and *2DL5B* were not used in this case because they lacked the D1 domain. This tree is consisted of the same sequence groups as those of the tree in figure 2.3. According to these trees, there are three distinct groups of human *KIR* genes. This observation is consistent with previous results (Khakoo et al. 2000; Guethlein et al. 2002), though our trees were rooted unlike the previous trees and the correlation between the phylogenetic grouping and the ligand binding specificity was clearly shown. Group I is the most recently diverged group with the largest number of genes, while group III is the most ancient. Eight *KIR* sequences are clustered into group I with a high bootstrap value and all of them share a pseudo-exon encoding a 'silent' D0 domain which is absent in the mature mRNA due to alternative splicing (Vilches, Pando, and Parham 2000). This supports that group I *KIR* genes are monophyletic.

The phylogenetic groups of human *KIR* genes are well correlated with the domain organization and ligand-binding specificity of their encoded proteins except for gene *3DL3*. The KIRs encoded by group II genes share the organization of three Ig-like domains (D0, D1, and D2). Both group I and group III KIRs have the structure of two Ig-like domains, but with different combinations (figure 2.3).

Furthermore, evolutionarily closely related KIRs tend to share similar MHC-binding properties. Six group I KIRs have been identified to bind to MHC-C alleles. In group II genes, 3DL1 and 3DS1 bind to MHC-B alleles, and 3DL2 to MHC-A alleles. As for group III genes, 2DL4 is known to bind to non-classical MHC-G molecules (see review by Boyington and Sun 2002). Interestingly, the ligand-binding specificity of chimpanzee KIRs is similar to that of human KIRs in each group (Khakoo et al. 2000). This suggests that these different *KIR* gene groups have acquired different ligand-binding specificities before the divergence of the two species. In addition, the evolution of new MHC-binding specificity for *KIR* genes is correlated with the evolution of *MHC* class I genes in primates (Piontkivska and Nei 2003). Therefore, we investigated the role of MHC ligands on the diversification of different groups of *KIR* genes and inferred the potential MHC-binding specificity of the ancestral *KIR* sequences.

Positive selection during the early expansion of group I human *KIR* genes

The MHC class I molecules are the only ligands of KIRs so far identified. The interaction of group I and II KIRs with MHC molecules requires mainly D1 and D2 domains of KIR proteins (Boyington and Sun 2002). The comparison of pairwise nonsynonmous (d_N) and synonymous (d_S) nucleotide substitutions for the individual Ig-like domain of human and other primate *KIR* genes (data not shown) has suggested that the d_N /d_S ratio is higher than 1 in the domain D1. Furthermore, Hughes (2002) suggested that the natural selection apparently operates at the MHC-binding region of D1 domain. These results are consistent with the observation

that the less conserved D1 domain is primarily responsible for the MHC locus or allotype binding specificity of KIRs (Gumperz et al. 1997; Winter and Long 1997). We therefore focused on domain D1 in the following study and studied the potential role of positive selection on the functional diversification of different groups of *KIR* genes.

The phylogenetic analysis of domain D1 suggested that group I *KIR* genes with MHC-C binding specificity diverged from group II *KIR* genes with MHC-A or B binding specificity after gene duplication occurred at node **a** of the tree in figure 2.4. We therefore inferred the ancestral nucleotide sequences for each interior node of this tree and estimated the numbers of nonsynonymous (a_N) and synonymous (a_S) substitutions per sequence per branch. There are 12 nonsynonymous substitutions but only 1 synonymous substitution on the branch **a-b**. The Fisher's exact test that compares the ratio of a_N/a_S with the expected ratio give substantial statistical support (P = 0.07) for the involvement of positive selection during the functional diversification of group I *KIR* genes (branch **a-b**).

Nonrandom amino acid substitutions for branch a-b

To examine what kind of amino acid changes occurred between nodes **a** and **b** in figure 2.4, we compared the ancestral amino acid sequences of domain D1 (figure 2.5). It appears that the amino acid changes from the ancestral sequences **a** to **b** are not random. Of a total of 9 amino acid substitutions between them, 4 are located at the MHC-binding sites (MBS) (labeled by asterisks in figure 2.5). The MBS of human group I and II KIRs is known to include 18 amino acid sites, 7 of which are

located in domain D1 and the remainder in D2 (Boyington and Sun 2002). There are 100 amino acids in the D1 domain and 7 of them are at the MBS. If we assume that the amino acid substitutions occur randomly along the sequence, the probability of 4 out of 9 total substitutions occurring at the MBS can be computed by using the hypergeometric distribution, and it is only 0.00096. This result supports the idea that an excessive number of amino acid substitution occurred at the MBS. A closer examination showed that all of the four amino acid substitutions between the sequences **a** and **b** at the MBS were polarity changes. For example, the hydrophobic isoleucine (I) in sequence **a** changed to the hydrophilic lysine (K) in **b** (labeled by the arrow sign in figure 2.5). These changes of polarity at the ligand-binding sites might be responsible for a change of the ligand-binding specificity.

Furthermore, to infer the potential ligand-binding specificity of ancestral sequences at nodes **a** and **b**, we compared their D1 domains with those of the present-day group I and II *KIR* sequences (figure 2.5). The ancestral sequence **a** has a high degree of sequence similarity with group II *KIR* genes, while **b** is similar to group I *KIR* genes. Therefore, it is very likely that the ancestral sequence **a** might have recognized MHC-A or B-related molecules similar to that of group II KIRs. After the gene duplication at node **a**, one gene appears to have retained the old binding specificity, and the other (sequence **b**) evolved the new MHC-C binding specificity of the present-day group I KIRs. Further support for this argument comes from the extended comparison of the MBS in domain D1 among different primate *KIR* genes (figure 2.6). The amino acid compositions of group I and II KIRs at the

MBS are similar to those of the ancestral sequences **b** and **a**, respectively. The functional studies of chimpanzee KIRs showed that the KIR molecule encoded by a group II gene *3DL1* recognizes MHC-A and B, while two group I *KIR* genes, *2DL6* and *3DL4*, have the MHC-C binding specificity (Khakoo et al. 2000). Overall, these results suggest that the amino acid substitutions in the MBS of the D1 domain are related to the functional divergence from group II to group I *KIR* genes.

As mentioned earlier, the D1 domain is important in determining the MHC allotype specificity. For example, the change of a single amino acid at the MBS (labeled by the arrow sign in figure 2.5) switches the binding specificities from one allotype of MHC-C to another (Winter and Long 1997). The human KIR molecules, 2DL2 and 2DL3, have lysine (K) at this site and they bind to the MHC-C allotypes (Cw1, 3, 7, and 8) which have asparagine (Asn) at amino acid position 80 (Asn80) of the α1 domain. However, 2DL1 has methionine (M) at this site and binds to the other MHC-C allotypes (Cw2, 4, 5, 6, and 15) with lysine at position 80 (Lys80). Therefore, we speculate that the ancestral sequence **b** might have recognized the specific MHC-C allotypes (Asn80) as in the case of 2DL2 and 2DL3.

Rapid differentiation of *KIR* genes by repeated gene duplication

One of the most distinctive features of evolution of the *KIR* gene family is the rapid increase of member genes by gene duplication. To study the rate of gene family expansion, we constructed a NJ tree of primate *KIR* genes using p-distance (figure 2.6). The cattle and rodent *KIR* genes which form distinct clusters from primate *KIR* genes were also included. According to the tree, both group I and group II *KIR* genes

are shared by all hominoid species used (humans, chimpanzees, gorillas, and orangutans). In each of these two groups, the orangutan *KIR* genes form a cluster separate from human and chimpanzee genes. Even between humans and chimpanzees, there are only two pairs of *KIR* genes (*2DS4* and *3DL1*) that might be considered as the real orthologs between these two species. All macaque *KIR* genes form a clade separate from hominoid group I and II *KIR* genes. Therefore, the *KIR* gene family in primates was diversified very rapidly and many gene duplication events have occurred in different primate species.

To examine the expansion rate of this gene family, we constructed a linearized tree (figure 2.7, Takezaki, Rzhetsky, and Nei 1995) and estimated the divergence times of duplicate genes using the human and orangutan divergence (13 million years ago, Mya) as the calibration point (Glazko and Nei 2003). To test the rate constancy among sequences, we used the branch length test as implemented in the LINTREE program (http://mep.bio.psu.edu/). This test indicated that there is no *KIR* sequence evolving significantly faster or lower than the average rate ($P < 0.01$). The molecular time scale in figure 2.7 shows that group I and group II *KIR* genes diverged about 17 Mya (16.6 ± 1.4, node B), which is after the divergence between OW monkeys and hominoids (23 Mya) and before the divergence of the human-chimpanzee lineage from the orangutan lineage (13 Mya). This explains why no macaque *KIR* genes are clustered with these two groups. Although the group III *KIR* genes (*2DL4* and *2DL5*) were not included in the analysis because of the lack of the D1 domain, they have been identified in both hominoids and OW monkeys (e.g., Khakoo et al. 2000;

Hershberger et al. 2001) and diverged from group I and group II hominoid *KIR* genes as a basal lineage when the macaque-specific *KIR* group diverged (data not shown). Therefore, we speculate that the expansion of primate *KIR* genes started about at least 25 Mya (25.4 ± 3.0 at node A in figure 2.7).

There are a total of 12 group I and II *KIR* genes excluding potential alleles in human genome but including pseudogenes *2DP1* and *3DP1* (figure 2.1). Because a gene family with *n* members should have been generated by *n–1* gene duplication events, the rate of gene expansion for the *KIR* gene family in the human lineage is roughly estimated to be 0.65 (=11/17) per million years (My). For the most diversified group (I), the gene expansion rate is even higher and becomes 0.69 (=9/13) per My under the assumption that humans and orangutans diverged 13 Mya. We have done similar computations for other primate species. The results have shown that in all primate species considered here the rate of expansion of *KIR* genes is very high (0.3~0.6, table 2.2). Particularly the macaque *KIR* gene family showed a rate as high as 1.1 duplications per My. It is known that *MHC* class I and class II genes also have experienced rapid gene duplications. The results obtained by a similar approach indicate that the rate of *MHC* class I gene expansion are 3 times lower than those of human *KIR* genes (table 2.2, Takahashi, Rooney, and Nei 2000; Piontkivska and Nei 2003). There is only one reported gene family that shows a rate of gene expansion as high as that for *KIR* genes. It is the *morpheus* gene family whose function is unknown (Johnson et al. 2001). In this gene family, gene duplication occurred 13 times in the human lineage after separation from orangutans. So, the rate becomes one duplication

per one My. The *KLRA* gene family in rodents, which is the functional equivalent of primate *KIR* genes, has an expansion rate as high as that of *KIR* genes in primates (Hao and Nei 2004).

Coevolution between *KIR* and *MHC* class I genes

The evolutionary pattern of primate *KIR* genes appears to coincide with that of classical *MHC* class I genes within primates. The group I *KIR* genes diverged from group II *KIR* genes soon after the gene duplication that generated the *MHC-C* locus 21-28 Mya (Piontkivska and Nei 2003). The orthologs of human *MHC-C* genes could be identified in other hominoids but not in OW monkeys and New World (NW) monkeys. Furthermore, although macaques have true orthologs of hominoid *MHC-A* and *B* genes, they are distantly related to each other (Adams and Parham 2001). These observations explain why no group I *KIR* genes are found in macaques and all macaque *KIR* genes form a species-specific cluster different from group I and II *KIR* genes (figure 2.6). Orangutan *KIR* genes form a separate cluster different from human and chimpanzee *KIR* genes in both group I and II. This is consistent with the observation that the classical *MHC* loci in orangutans are distantly related to the orthologs in chimpanzees and humans. It has been shown that the presence of the *MHC-C* is polymorphic in orangutan populations. In light of the positive selection at MBS of *KIR* genes shown earlier, the correlation of evolutionary pattern between *KIR* genes and *MHC* class I genes seems to suggest a coevolutionary relationship between them during the evolution of primates.

DISCUSSION

As mentioned earlier, the genetic variation of the *KIR* gene family is manifested primarily as haplotype diversity generated by tandem duplication, block duplication, nucleotide substitution, gene translocation, recombination, and gene loss, etc. Without having more information about the *KIR* haplotype structure in other primates, it is difficult to evaluate the contribution of each of these factors. However, our rough estimates of the divergence times of different *KIR* lineages provide us with an idea about how quickly the *KIR* gene family expanded and how their evolution correlated with the evolution of their ligands during the primate evolution.

We have emphasized the increase of functional genes in current species of hominoids and OW monkeys. However, it is likely that their ancestral species had many *KIR* genes and many previously functional genes have been lost or become pseudogenes. Indeed, the *KIR* gene family has several pseudogenes in humans, chimpanzees, and macaques. Therefore, this gene family appears to be subject to birth-and-death evolution, like the other multigene families such as the *MHC* and immunoglobulin gene families (Nei, Gu, and Sitnikova 1997). In our study, however, it was difficult to study the death rate of genes, because we could not estimate the number of genes in the genome of ancestral organisms.

The biological function of a haplotype must depend on the number of inhibitory and activating genes, allelic sequences, and splicing properties, etc. We can measure the extent of haplotype diversity (H) in a population by a quantity similar to gene diversity (or heterozygosity) of a gene (Nei 1987, pp.178). In the present case, it is

given by $H = n(1 - \sum_{i=1}^{N} X_i^2)/(n-1)$, where *n*, *N*, and *X* are the total number of genomes examined, the number of different haplotypes observed, and the frequency of the i-th haplotype, respectively. In a Caucasian population studied by Hsu et al. (2002), we have estimated *H* to be 0.81. This value is much higher than the gene diversity (0.135) observed by protein electrophoresis in human populations (Nei 1987, pp.192). However, this haplotype diversity must be minimum because only a small number of individuals have been studied so far.

It is interesting to note that the genetic variability of MHC class I molecules is generated primarily by allelic variation rather than by haplotype diversity. If *KIR* genes and *MHC* alleles coevolved to cope with antigen evolution as suggested by Parham (1997) and by the results presented here, why is haplotype diversity important in *KIR* genes? If the amino acid sequences of the MHC-binding regions of KIR coevolve with the corresponding regions of MHC molecules, there is no need for a high level of haplotype diversity. Probably, one reason for the occurrence of many haplotypes is that both inhibitory and activating genes are required in this system, and the two types of genes apparently must be relatively closely related to each other. If this is the case, unequal interlocus recombination may occur frequently, and this will increase the number of different haplotypes. At the present time, however, the real mechanism of the interaction between KIRs and MHC molecules is not well understood, and it would be difficult to know the reason for generating many different haplotypes.

In rodents a structurally different gene family called the *KLRA* gene family

plays essentially the same function as that of *KIRs* in primates as mentioned. As in the case of *KIR* genes, the *KLRA* gene family in rodents has experienced a rapid expansion of member genes by repeated gene duplication (Wilhelm, Gagnier, and Mager 2002; Hao and Nei 2004). It would be interesting to know how these two entirely different gene families have come to play the same biological function.

Table 2.1 Primate *KIR* gene sequences used

Humans (hum)	*2DS4* (AF258807)	*3DL5* (AF334620)
3DL1 (NM_013289)	Pygmy chimpanzees (bono)	*3DL6* (AF334621)
3DL2 (AJ276125)	*3DLb* (AF266733)	*3DL7* (AF334622)
3DL3 (AF352324)	*3DL4* (AF266731)	*3DL8* (AF334623)
3DS1 (NM_014514)	*3DSa* (AF266735)	*3DL9* (AF334624)
2DL1 (L41267)	Gorillas (gori)	*3DL10* (AF334625)
2DL2 (NM_014219)	*3DL7* (AY122869)	*3DL11* (AF334626)
2DL3 (L41268)	*2DLb* (AY122871)	*3DL16* (AF361084)
2DL4 (NM_002255)	*2DLc* (AY122872)	*3DL18* (AF361086)
2DL5A (AL133414)	*2DLd* (AY122873)	*3DH1* (AF334648)
2DL5B (AF217486)[a]	*2DLe* (AY122874)	*3DH2* (AF334649)
2DS1 (X89892)	*2DL6* (AY122870)	*3DH3* (AF334650)
2DS2 (U24079)	*2DSa* (AY122875)	*3DH4* (AF334651)
2DS3 (NM_012313)	Orangutans (oran)	Cattle (cattle)
2DS4 (U24077)	*3DLA* (AF470365)	*3DL1* (AF490402)[d]
2DS5 (NM_014513)	*3DLE* (AF470370)[c]	*3DL2* (AY075103)[e]
2DP1 (AL133414)[a]	*2DLA* (AF470358)	*3DS1* (AF490401)
Common chimpanzees (chimp)	*2DLB* (AF470359)	Mice (mice)
3DL1 (AF258798)[b]	*2DSA* (AF470360)	*3DL1* (NM_177748)
3DL4 (AF258800)	*2DSC* (AF470362)[c]	*2DL1* (AY152727)
3DL5 (AF258801)	Macaques (maca)	Rats (rat)
3DL6 (AF258802)	*3DL1* (AF334616)	*3DL1* (AF527797)
3DS2 (AF258803)	*3DL2* (AF334617)	

[a] The gene sequences were extracted from the genomic contig.

[b] The gene *3DL1* was denoted as *3DL1/2* in Khakoo et al. (2000)'s analysis.

[c] In Guethlein et al. (2002)'s paper, the gene *3DLE* was *3DLE1* and *2DSC* was *2DSC1*.

[d] Cattle *3DL1* and *3DS1* were obtained from Storset et al. (2003).

[e] Cattle *3DL2* is equal to the *3DL1* identified by McQueen et al. (2002).

Table 2.2 Rates of gene expansion for the group I and II hominoid *KIR* lineage, the macaque-specific *KIR* lineage, and other gene families

Gene Family	Evolutionary Time (My)	Duplication Events	Expansion Rate *per* My	Reference
Class II *MHC*	50	4	0.08	Takahashi et al. (2000)
Class I *MHC*	56	9	0.18	Adams and Parham (2001); Piontkivska and Nei (2003)
Group I and II human *KIR*	17	11	0.65	Our studies
Group I common chimpanzee *KIR*	13	4*	0.31	Our studies
Group I gorilla *KIR*	13	6*	0.46	Our studies
group I and II orangutan *KIR*	17	5*	0.29	Our studies
Macaque *KIR*	13	14*	1.08	Our studies
Morpheus	13	13	1.00	Johnson et al. (2001)

* The number of *KIR* genes might not be fully identified.

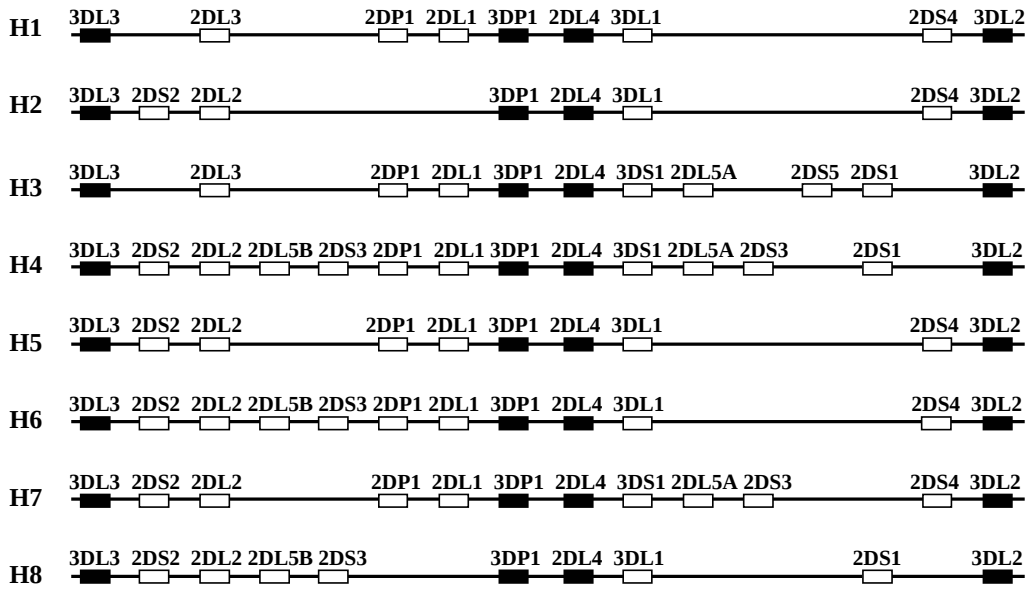


Figure 2.1. Genomic organizations of human *KIR* genes in 8 representative haplotypes (H1 ~ H8) (adapted from Hsu et al. 2002). Each haplotype has 8~14 *KIR* genes in a ~200 kb region. The black box represents the framework locus which is present on all the examined haplotypes. The open box represents the *KIR* gene which varies with haplotype. Only gene order is shown, and the distance between genes is not to scale. Each haplotype has occurred at least twice in the population studies (Hsu et al. 2002).

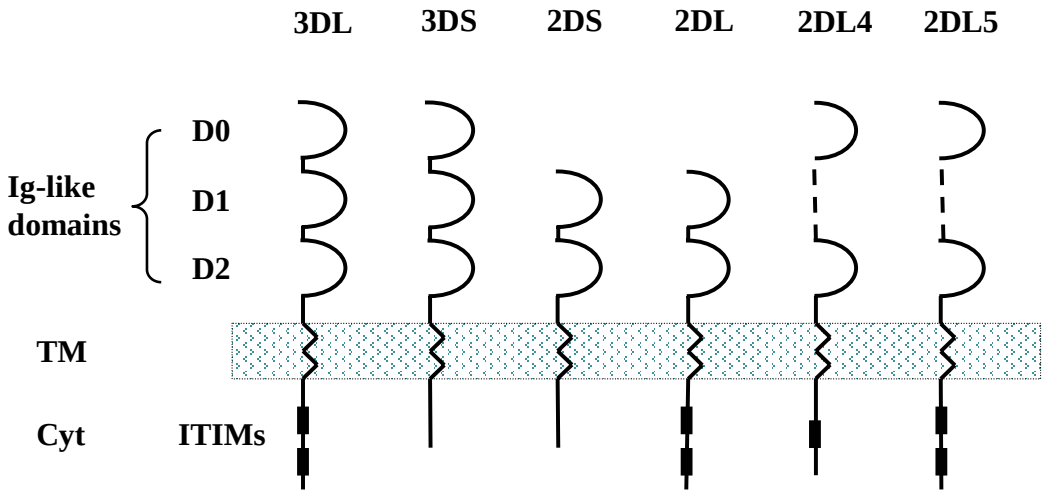


Figure 2.2. Domain organizations of KIRs. Most KIRs can be classified into 4 subgroups according to the domain organization. One recently identified cattle KIR molecule (2DS1) has an unusual domain organization (D0 and D1) (Storset et al. 2003), which was not shown here. TM, transmembrane region. Cyt, cytoplasmic region. ITIM, the immunoreceptor tyrosine inhibitory motif.

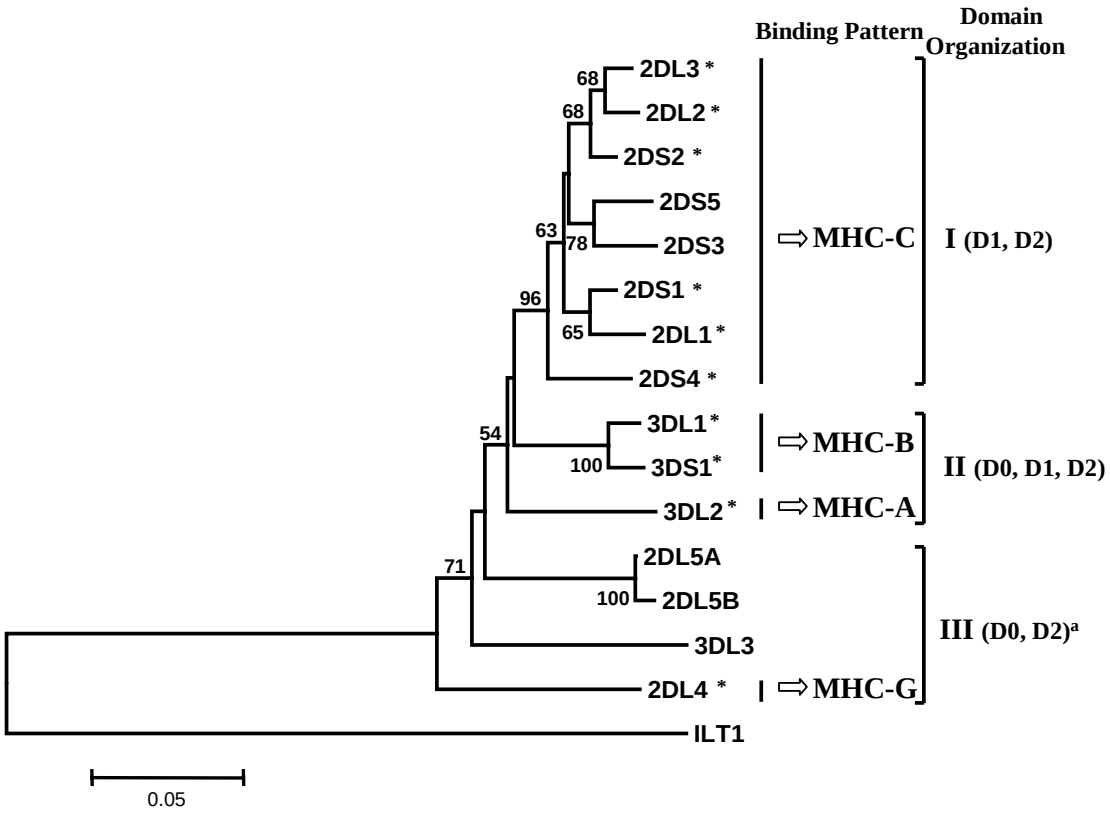


Figure 2.3. Neighbor-joining tree of 15 human *KIR* nucleotide sequences. The p-distance was used. The bootstrap value based on 1000 replications is shown for each interior branch wherever it is greater than 50%. A human IgSF gene closely related to *KIR* genes, called *ILT1* (immunoglobulin-like transcript), was used as the outgroup. Its Genbank accession number is U82275. The second Ig-like domain of this gene has been excluded in the alignment, since the gene has four Ig-like domains on the extracellular region and the first, third, and forth domains are most similar to D0, D1, and D2 domains of KIRs, respectively (data not shown). The asterisk "*" represents the KIR protein with known MHC-binding specificities. The inferred MHC-binding property for each group is shown on the right hand side of the tree. [a]*3DL3* has the structure of 3 Ig-like domains, which is different from *2DL4* and *2DL5*.

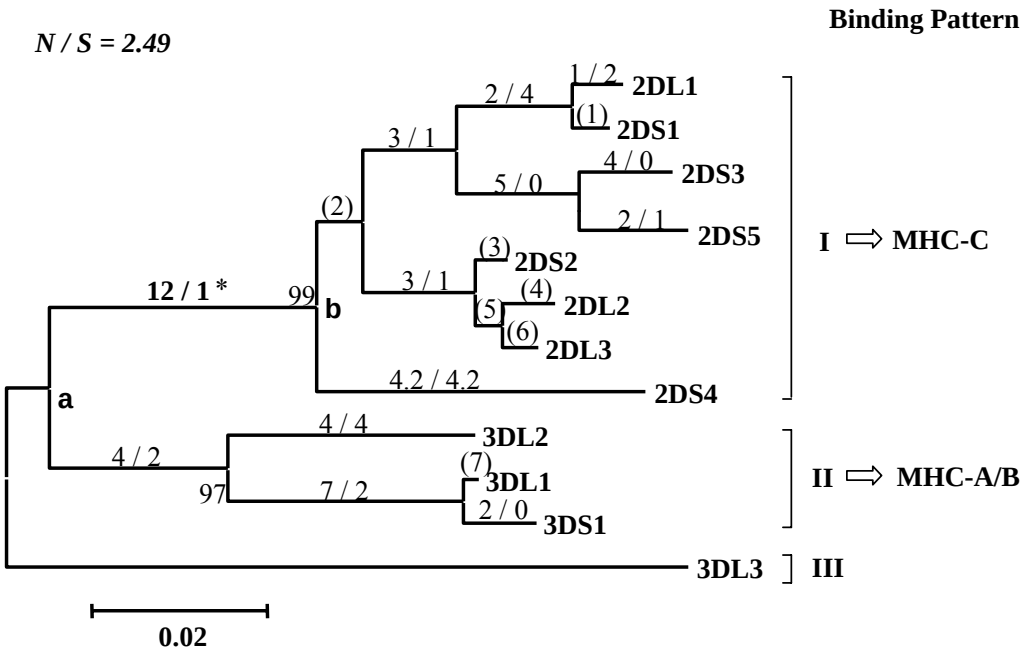


Figure 2.4. Neighbor-joining tree of nucleotide sequences of the D1 domain from 12 human KIRs. The topology of this tree was obtained by using p-distance, and the same topology was obtained by using JC distance. The pseudogenes *3DP1* and *2DP1* were not included because of the nonsynonymous/synonymous substitution rate test. The number of nucleotides of D1 domain is 300. The gene duplication occurred at node **a**, from which 2 different groups (I and II) of *KIR* genes are diverged. The numbers of nonsynonymous (a_N) and synonymous (a_S) substitutions per sequence per branch are presented as a_N /a_S above the branches. The *N*/*S* ratio for all the *KIR* sequences including ancestral sequences is given above the tree. (1) (2/2). (2) (1.5/0.5). (3) (1/0). (4) (1/2). (5) (0/2). (6) (1/1). (7) (1/0). The statistical significance of the difference between observed ratio a_N /a_S and expected ratio *N*/*S* was determined by Fisher's exact test which showed that the difference is significant at the 7% level (*).

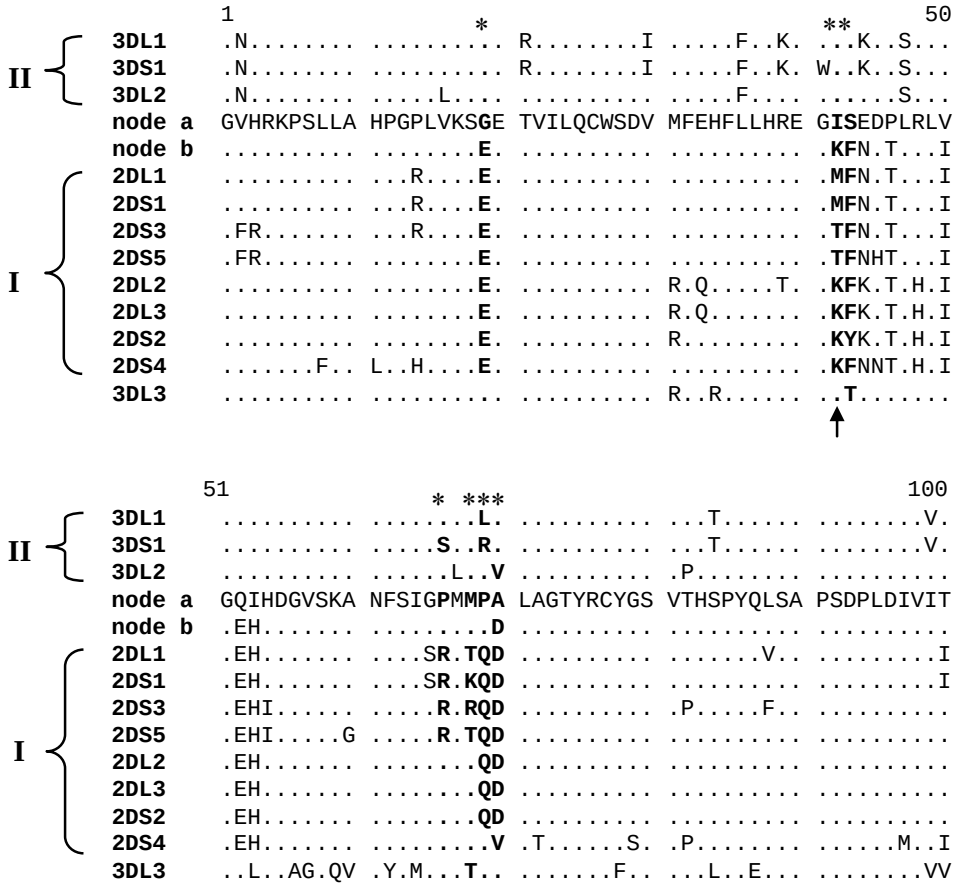


Figure 2.5. Comparison of the D1 domain between ancestral *KIR* sequences (node **a** and **b** in figure 2.4) and present-day group I and II human *KIR* sequences. Amino acids are presented by single-letter codes and dots represent the same amino acids as those of the sequence at node **a**. The MHC-binding sites (MBS) are shown in bold letters and also labeled with asterisks. The arrow sign represents the key amino acid which determines MHC-C allotype binding specificities.

Figure 2.6. Neighbor-joining tree of nucleotide sequences of 47 primate, 3 cattle, and 3 rodent *KIR* genes. We used both JC distance and p-distance and both trees had the same topology. The tree constructed by p-distance was shown here. Only extracellular regions were used and the number of nucleotide sites is 587. The three groups of *KIR* genes in primates are shown with vertical lines. The bootstrap value based on 1000 replications is shown for each interior branch wherever it is greater than 50%. The amino acids of MHC-binding sites (MBS) in the D1 domain (labeled with asterisks on the top) are shown for each primate *KIR* gene except for the pseudogene *2DP1*. The 4 sites which are different between two inferred ancestral sequences (node **a** and **b** in figure 2.4) are shown in bold letters. The human *ILT1* gene was used as the outgroup.

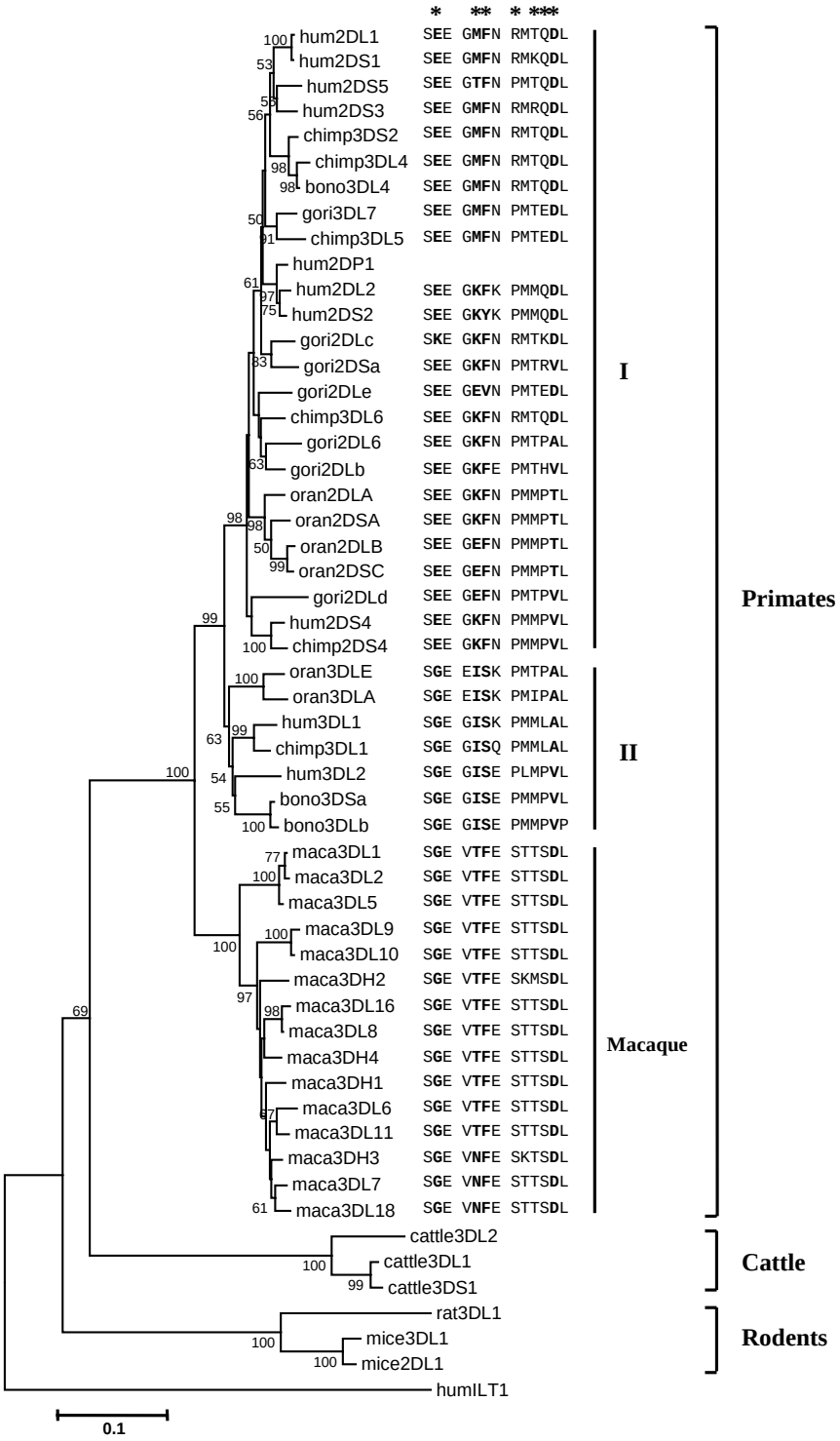
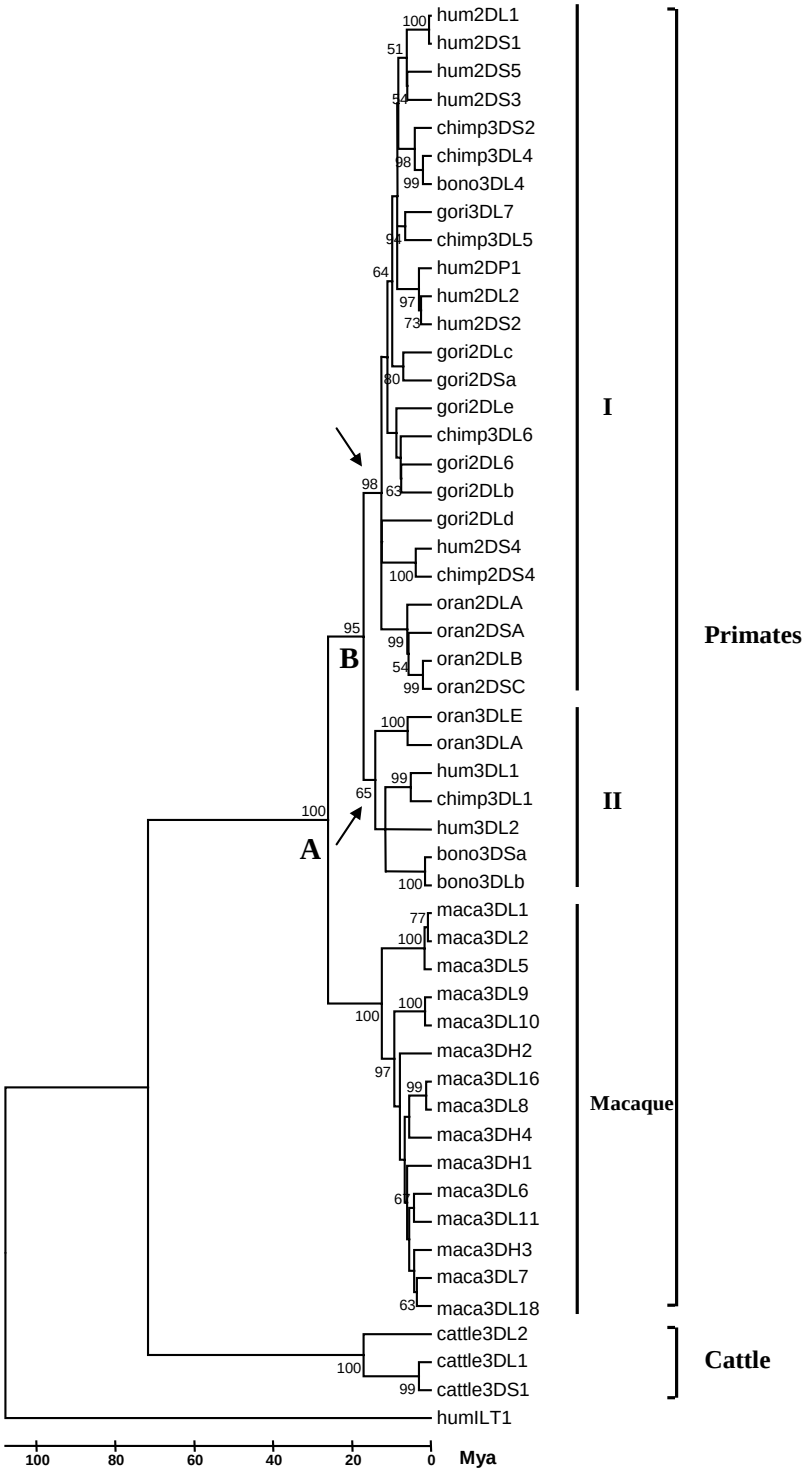

* ** * ***
hum2DL1 SEE GMFN RMTQDL
hum2DS1 SEE GMFN RMKQDL
hum2DS5 SEE GTFN PMTQDL
hum2DS3 SEE GMFN RMRQDL
chimp3DS2 SEE GMFN RMTQDL
chimp3DL4 SEE GMFN RMTQDL
bono3DL4 SEE GMFN RMTQDL
gori3DL7 SEE GMFN PMTEDL
chimp3DL5 SEE GMFN PMTEDL
hum2DP1
hum2DL2 SEE GKFK PMMQDL
hum2DS2 SEE GKYK PMMQDL
gori2DLc SKE GKFN RMTKDL
gori2DSa SEE GKFN PMTRVL
gori2DLe SEE GEVN PMTEDL
chimp3DL6 SEE GKFN RMTQDL
gori2DL6 SEE GKFN PMTPAL
gori2DLb SEE GKFE PMTHVL
oran2DLA SEE GKFN PMMPTL
oran2DSA SEE GKFN PMMPTL
oran2DLB SEE GEFN PMMPTL
oran2DSC SEE GEFN PMMPTL
gori2DLd SEE GEFN PMTPVL
hum2DS4 SEE GKFN PMMPVL
chimp2DS4 SEE GKFN PMMPVL
oran3DLE SGE EISK PMTPAL
oran3DLA SGE EISK PMIPAL
hum3DL1 SGE GISK PMMLAL
chimp3DL1 SGE GISQ PMMLAL
hum3DL2 SGE GISE PLMPVL
bono3DSa SGE GISE PMMPVL
bono3DLb SGE GISE PMMPVP
maca3DL1 SGE VTFE STTSDL
maca3DL2 SGE VTFE STTSDL
maca3DL5 SGE VTFE STTSDL
maca3DL9 SGE VTFE STTSDL
maca3DL10 SGE VTFE STTSDL
maca3DH2 SGE VTFE SKMSDL
maca3DL16 SGE VTFE STTSDL
maca3DL8 SGE VTFE STTSDL
maca3DH4 SGE VTFE STTSDL
maca3DH1 SGE VTFE STTSDL
maca3DL6 SGE VTFE STTSDL
maca3DL11 SGE VTFE STTSDL
maca3DH3 SGE VNFE SKTSDL
maca3DL7 SGE VNFE STTSDL
maca3DL18 SGE VNFE STTSDL
I
II
Macaque
Primates
cattle3DL2
cattle3DL1
cattle3DS1
Cattle
rat3DL1
mice3DL1
mice2DL1
Rodents
humILT1
0.1

Figure 2.7. Linearized tree of the primate *KIR* genes obtained by using Kimura's distance. Rodent *KIR*-like genes were excluded because the function of these genes is unknown. The time scale was calibrated with the divergence time between humans and orangutans (13 Mya, labeled with arrow signs). The nodes of which the divergence times we are interested in are labeled as capital letters in bold.


hum2DL1
hum2DS1
hum2DS5
hum2DS3
chimp3DS2
chimp3DL4
bono3DL4
gori3DL7
chimp3DL5
hum2DP1
hum2DL2
hum2DS2
gori2DLc
gori2DSa
gori2DLe
chimp3DL6
gori2DL6
gori2DLb
gori2DLd
hum2DS4
chimp2DS4
oran2DLA
oran2DSA
oran2DLB
oran2DSC
oran3DLE
oran3DLA
hum3DL1
chimp3DL1
hum3DL2
bono3DSa
bono3DLb
maca3DL1
maca3DL2
maca3DL5
maca3DL9
maca3DL10
maca3DH2
maca3DL16
maca3DL8
maca3DH4
maca3DH1
maca3DL6
maca3DL11
maca3DH3
maca3DL7
maca3DL18
cattle3DL2
cattle3DL1
cattle3DS1
humILT1
A
B
I
II
Macaque
Primates
Cattle
100
80
60
40
20
0
Mya

CHAPTER 3

GENOMIC ORGANIZATION AND EVOLUTIONARY ANALYSIS OF KLRA GENES ENCODING THE RODENT NATURAL KILLER CELL RECEPTORS: RAPID EVOLUTION BY REPEATED GENE DUPLICATION

SUMMARY

KLRA (also called *Ly49*) genes regulate the cytotoxic activity of natural killer (NK) cells in rodents and provide important protection against virus-infected or tumor cells. About 15 *KLRA* genes have been identified in mice, but only a few genes have been reported to date in rats. Here we studied all *KLRA* genes in the entire rat genome sequence and identified 17 putative functional and 16 putative nonfunctional genes together with their genomic locations in a 1.8 Mb region of chromosome 4. Phylogenetic analysis of these genes indicated that the *KLRA* gene family expanded rapidly in recent years, and this expansion was mediated by both tandem and genomic block duplication. The joint phylogenetic analysis of mouse and rat genes suggested that the most recent common ancestor of the two species had at least several *KLRA* genes but that the majority of current duplicate genes were generated after divergence of the two species. In both species *KLRA* genes are apparently subject to birth-and-death evolution, but the birth and death rates of *KLRA* genes are higher in rats than in mice. The rate of gene expansion in the *KLRA* gene family in rats is one of the highest among all mammalian multigene families so far studied. The biochemical function of *KLRA* genes is essentially the same as that of *KIR* genes in primates, but the molecular structures of the two groups of NK cell receptors are very different. A hypothesis was presented to explain the origin of the differential use of *KLRA* and *KIR* genes in rodents and primates.

INTRODUCTION

Natural killer (NK) cells are an important component of the innate immune system in mammals and are capable of discriminating virus-infected or tumor cells from healthy cells. This discrimination is accomplished by a large number of natural killer cell receptors that are expressed on the surface of NK cells (Trowsdale et al. 2001). Two of those NK cell receptors are killer-cell immunoglobulin-like receptors (KIR) and C-type lectin-like receptors (KLRA). The KIR and KLRA receptors have very different molecular structures and are encoded by different gene families. Yet, they have the same function and regulate the cytotoxic activity of NK cells through binding to major histocompatibility complex (MHC) class I molecules or MHC-related molecules (Colonna and Samaridis 1995; Brennan et al. 1996; Dorfman and Raulet 1996; Idris et al. 1999; Brown et al. 2001). Interestingly, KIRs are used primarily in primates, whereas KLRA are used in rodents. Therefore, KIR and KLRA receptors represent a rare case of functional convergence at the molecular level.

The diversity and evolution of the *KIR* gene family have been studied extensively in several primate species, and we now know that the intraspecific variation of KIRs is generated by duplicate genes, haplotype variation, allelic polymorphism, and alternative splicing (Kwon et al. 2000; Martin et al. 2000; Wilson et al. 2000; Gardiner et al. 2001). However, the primary factor of evolution of the *KIR* gene family is the rapid expansion of member genes by repeated gene duplication (Hershberger et al. 2001; Guethlein et al. 2002; Hao and Nei 2005). A similar mechanism has been suggested about the evolution of the *KLRA* gene family

in mice. The genetic diversity of mouse *KLRA* genes has been studied extensively with different strains such as C57BL/6 and 129/J (Brown et al. 1997; Depatie et al. 2000; Makrigiannis et al. 2002; Wilhelm, Gagnier, and Mager 2002). There is substantial variation in gene number and gene content among different strains. Rats are also likely to have many *KLRA* genes, but only 6 genes have been reported to date (Dissen et al. 1996; Naper et al. 2002aa; 2002bb). Since the genetic diversity of this gene family has been studied only in mice, it is difficult to know how quickly *KLRA* genes were generated in rodents.

Fortunately, a nearly complete rat genome sequence has recently become available (NCBI rat build 2, www.ncbi.nlm.nih.gov/genome/guide/rat). We therefore decided to study the number of duplicate genes in the *KLRA* family in rats and their evolutionary relationships with mouse genes. We have also investigated the similarity and dissimilarity of the evolution of *KIR* genes and *KLRA* genes. The results obtained will be presented below.

MATERIALS AND METHODS

Background information about *KLRA* genes

KLRA genes are located at a well-defined genomic region called the natural killer gene complex (NKC) together with some other C-type lectin-like genes such as the *KLRD1* and *KLRC* genes (figure 3.1A). The *KLRA* gene family is located on chromosome 12 in humans, chromosome 6 in mice, and chromosome 4 in rats (Brown et al. 1997; Renedo et al. 2000). However, the *KLRA* gene family is much larger in rodents than in primates and other mammals. Only one *KLRA* gene has been identified in humans and baboons (Mager et al. 2001), cattle (McQueen et al. 2002), and pigs, cats, and dogs (Gagnier, Wilhelm, and Mager 2003). About 5~6 *KLRA* genes have been identified in horses (Takahashi et al. 2004). *KLRA* genes are also called *Ly49* genes in some of the current literature, but in this study we use the standard terminology (NCBI Genome Annotation, http://www.ncbi.nlm.nih.gov/Genomes/index.html).

NK cell receptors can be categorized into inhibitory receptors and activating receptors according to the presence or absence of the immunoreceptor tyrosine-based inhibitory motif (ITIM) in the cytoplasmic region (Mason et al. 1996; 1997). When KLRA proteins bind to the ligand, the inhibitory receptors will inhibit the cytotoxic activity of NK cells through the ITIM. By contrast, the activating receptors will recruit an adaptor (DAP12) through a charged arginine (R) in the transmembrane domain, and this adaptor stimulates the cytotoxic activity of NK cells. The ITIM can

easily be recognized by the sequence signature "I/VxYxxV/L" in the cytoplasmic region, where "x" represents any of the 20 amino acids (Mason et al. 1997).

Identification and sequence analysis of rat *KLRA* genes

A draft assembly of the rat genome sequence (RGSC v3.1) covering more than 90% of the entire genome is available from NCBI. The sequences were obtained from 2 female rats (BN/SsNHsd) of a highly inbred line (Rat Genome Sequencing Project Consortium 2004). This means that the *KLRA* cluster identified here represents a single haplotype. Using the known rat *KLRA* gene (Ly49.9) as the query sequence (Dissen et al. 1996), we first located the *KLRA* gene cluster on the supercontig (NW_043770, updated by NW_047696 recently) of rat chromosome 4 with the computer program BLAST (Zhang et al. 1998b, ftp://ftp.ncbi.nih.gov/blast/). To identify the complete set of *KLRA* genes, a relaxed search was performed with a relatively high E value (0.01). The number of BLAST hits suggested that about thirty different *KLRA* genes are located in a 1.8 Mb region of NKC. After extracting each potential gene segment, we used PipMaker (Schwartz et al. 2000, http://bio.cse.psu.edu/pipmaker/) to identify the exon boundaries for each gene comparing the extracted genomic sequences with the known cDNA (Ly49.9).

The genes identified in this way were named according to the order of the genes on the chromosome starting from the telomeric end of the *KLRA* gene cluster (figure 3.1B). The list of the genes is available in table 3.1, and their sequences can be obtained from Nei's lab database (http://mep.bio.psu.edu/databases/Ly49). The amino acid sequence for each rat *KLRA* gene was deduced from the nucleotide

sequences of exons by the nucleotide-translation program, ExPASy (http://us.expasy.org/tools/dna.html). The repetitive elements in the *KLRA* gene cluster were identified by the computer program RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker.html).

Phylogenetic analysis

All the sequence alignments were generated by using the computer program CLUSTALX (Thompson et al. 1997). Additional modifications were done by visual inspection. Phylogenetic analysis was conducted by using the neighbor-joining method (Saitou and Nei 1987) implemented in the computer program MEGA2 (Kumar et al. 2001). The bootstrap values were computed with 1000 resamplings. We used Jukes-Cantor and Kimura distances for the nucleotide sequences of protein coding regions (Nei and Kumar 2000), but since these two distances generated virtually the same trees, we present only the trees obtained by Jukes-Cantor distances. (Note that the gene sequences used were closely related.) In addition, we used the parsimony method as implemented in PAUP* with 1000 bootstrap resamplings (TBR search) (Swofford 1998). In tree construction, all alignment gaps were eliminated (complete-deletion option). *KLRC1*, a C-type lectin-like gene, which is closely related to *KLRA* genes, was used as the outgroup.

The nucleotide sequences for mouse *KLRA* genes (except *KLRA_B*) were obtained from the Mouse Genome Resources (http://www.ncbi.nlm.nig.gov/genome/guide/mouse). The GenBank accession number for mouse *KLRA_B* is AK017140. Two primate *KLRA* genes were also

included in the analysis: Baboons (*Papio hamadryas*, AY028399) and humans (*Homo sapiens*, NM_006611). The six known rat *KLRA* genes identified in different strains were obtained from GenBank with the following accession numbers. Strain F344: Ly49.9, U56863; Ly49.12, U56822; Ly49.19, U56823; Ly49.29, U56824. Strain PVG: Ly49i2, NM_152848; Ly49s3, NM_153726. The GenBank asscession number for the rat *KLRC1* gene is AF021350.

RESULTS

Genomic organization of the gene families in the NKC region of rats

The gene families identified in the NKC regions of humans, mice, and rats are presented in figure 3.1A. Although the number of genes in rats is limited owing to the incomplete gene annotation, we observe a high level of conservation of NKC gene families with respect to the gene order and gene content among the three species, as was previously observed (Trowsdale et al. 2001). The conserved genes *KLRG1* and *CSDA* defined the boundary of the *KLRA* gene cluster.

Our analysis of the genomic sequence of the rat *KLRA* gene cluster revealed a total of 33 full-length or nearly full-length genes located within a ~1.8 Mb region (figure 3.1B). In addition, there were several fragmentary genes containing one or a few exons, but these genes will not be considered in this study. The number of rat *KLRA* genes is about twice as large as that of mice, which is about 15 (figure 3.1B). All rat genes are orientated with the same transcriptional direction as in the case of mouse genes. Of the 33 rat genes, 5 appear to be the same as those previously identified in different rat strains (table 3.1). One of the previously identified genes, *Ly49s3*, did not match to any of our 33 genes probably because different strains were used. We did not use this gene in the following study.

Sequence and structure analysis of rat *KLRA* genes

The KLRA protein is a homodimer and consists of the extracellular lectin-like domain, stem, transmembrane, and cytoplasmic regions (figure 3.2). The coding region of rat *KLRA* genes is composed of 6 protein-coding exons and 5 introns as in

the case of mouse genes (Takei, Brennan, and Mager 1997). Exons 1, 2, and 3 encode the cytoplasmic, transmembrane, and stem regions, respectively. The last three exons encode the extracellular lectin-like domain, which is responsible for the interaction with ligands. The lengths of 5 introns vary with gene, though the last intron (between exons 5 and 6) is generally longest. *KLRA-1* has the longest introns and the entire gene region extends over 30 kb. Excluding 5 incomplete genes (*KLRA-19, 21, 24, 29*, and *33*), in which one or two exons are deleted, we found a total of 28 full-length genes. We assumed that the genes containing at least one stop codon in the first five exons are pseudogenes and the remaining genes are functional. Using this criterion, we found 17 full-length putative functional genes and 11 putative pseudogenes (table 3.1). Therefore, the proportion of pseudogenes was 39% (11/28) among the full-length genes and 48% (16/33) among all genes including the incomplete ones. Note that there are some gaps in the supercontig which we used to identify rat *KLRA* cluster. Hence, our estimates of the proportion of pseudogenes could be overestimated.

The mouse strain C57BL/6 is known to have a total of 15 full-length or nearly full-length *KLRA* genes, of which only 4 genes (*X, K, N*, and *M*) are probably non-functional (McQueen et al. 1999; Kane, Silver, and Hazes 2001; Wilhelm, Gagnier, and Mager 2002). Therefore, the proportion of pseudogenes is about 27%. This mouse strain has another fragmentary gene (L) containing one exon (Wilhelm, Gagnier, and Mager 2002), but this gene will not be considered here. Using primers specific to two exons and one intron of known *KLRA* genes, Makrigiannis et al. (2002)

identified 19 genes or gene segments in mouse strain 129/J, but their functionality is unknown. We therefore do not consider these genes in this study.

We then looked for the sequence signature of ITIM in all putative translated rat *KLRA* sequences. The results showed that 10 genes are putative inhibitory receptors with ITIM (table 3.1). Three genes (*KLRA-4*, *6*, and *18*) had no ITIM and are probably activating receptors because of the presence of arginine in the transmembrane region. Interestingly, the remaining four genes (*KLRA-23*, *26*, *27*, and *28*) contained both the ITIM and transmembrane arginine. The function of these genes is unknown at present.

Evolutionary relationships of rat *KLRA* genes and genomic block duplication

To study the evolutionary relationships of rat *KLRA* genes, we constructed a phylogenetic tree of 28 full-length genes using nucleotide sequences (figure 3.3). For the outgroup, we used primate *KLRA* genes. According to this tree, there are six different clades of *KLRA* genes with high bootstrap values (except for clade II). If we exclude the single gene clade VI (gene 1), the first splitting of phylogenetic groups occurred between clade V and other clades though the bootstrap value is not very high. The splitting order of the remaining clades, I ~ IV, is not clearly resolved. Essentially the same results were obtained by parsimony analysis. All of the five groups (I ~ V) include one or a few pseudogenes. According to the tree, some pseudogenes such as genes 5ψ and 31ψ were generated very recently, while some others (e.g., genes 9ψ and 13ψ) might have lost their function a long time ago.

Because the genes were named according to their genomic location, we can easily see the relationships between the phylogenetic clades and genomic locations of the genes. If the number of *KLRA* genes increased mainly by tandem duplication, we would expect that the genes which are physically clustered in the genome form a monophyletic group. Figure 3.3 shows that all five genes (2~6) in clade III are closely related. This suggests that clade III genes were generated by tandem duplication.

However, other gene clusters do not necessarily show closely related phylogenetic relationships. A close examination of gene arrangement has led us to identify a genomic block duplication including four *KLRA* genes. Figure 3.4 shows the comparison of the two genomic blocks that were apparently generated by block duplication. The upper part of each row shows the genomic block starting with gene 23 and ending with gene 26, and the lower part represents the genomic block from gene 28 to gene 31. The genomic structures of the two blocks are virtually identical except for insertion or deletion of some repetitive elements in the non-coding regions. The nucleotide identity between the coding regions was 95% between genes 23 and 28, 99% between genes 24 and 29, 98% between genes 25 and 30, and 99% between genes 26 and 31. These high identity values suggest that the block duplication occurred very recently.

A crude estimate of the time of occurrence of this duplication can be estimated by computing the Kimura distances between genomic blocks 1 and 2 and between the orthologous pair of mouse and rat genes (mouse-B and rat-1; see below). Since the

divergence between the mouse and rat lineages (T_{MR}) has been estimated to have occurred approximately 33 Mya (Nei, Xu, and Glazko 2001), one can estimate the time of block duplication (T_D) by $(d_{12}/d_{MR}) \times 33$ Mya, where d_{12} is the distance estimate between genomic blocks 1 and 2 and d_{MR} is the distance between the orthologous mouse and rat genes. When we used the concatenate sequence of the coding regions of the four *KLRA* genes from each of genomic blocks 1 and 2, we obtained $d_{12} = 0.025 \pm 0.003$. For d_{MR}, we obtained 0.204 ± 0.017. Therefore, the genomic block duplication appears to have occurred about 4.0 ± 0.1 Mya.

The above examples of tandem duplication and genomic block duplication are probably not isolated cases, and it is possible that the other regions experienced these processes several times but we cannot see the trace of the occurrence anymore. Of course, it is also possible that some gene duplication has occurred by gene transposition. However, one thing is clear from the genomic maps of the *KLRA* clusters in mice and rats. That is, this gene family did not experience inverted gene duplication changing transcriptional direction, unlike many genes in other gene families such as the immunogloblin and olfactory receptor gene families (Matsuda et al. 1993; Niimura and Nei 2003).

Evolutionary relationships of rat and mouse *KLRA* genes

To understand the evolutionary relationships of mouse and rat *KLRA* genes, we constructed a phylogenetic tree for the genes from mice, rats, baboons, and humans (figure 3.5). In this study, we used the mouse genes from strain C57BL/6 and rat genes from strain BN/SsNHsd/MCW. The root of the tree was determined by using

the rat *KLRC1* gene, which is located next to rodent *KLRA* genes (see figure 3.1A). From this tree, we could identify 9 groups of rodent genes, which are supported by relatively high bootstrap values except for the single gene mouse clade MIII and rat clade II. Five of them (I-V) were rat-specific and are identical to those previously identified in figure 3.3. Group VI contains one mouse (gene B) and one rat (gene 1) genes, and this group is supported by a bootstrap value of 100%. Both of these genes are located at the telomeric end of the *KLRA* gene cluster (figure 3.1B) and appear to be an orthologous pair. However, the branching order of the remaining eight phylogenetic clades is unclear, because the relevant bootstrap values are low. There are three mouse-specific clades (MI, MII, and MIII), and they are intermingled with rat specific clades. This suggests that the most recent common ancestor (MRCA) of mice and rats already had several *KLRA* genes. However, the species-specific grouping suggests that the main rodent *KLRA* gene repertoires expanded rapidly after mice-rats divergence.

Gene conversion or recombination events might be involved in the evolution of *KLRA* genes. For example, conducting phylogenetic analysis of approximately 1.4 kb portions of intron 1 (between the first noncoding exon and exon 1) and intron 6 (between exons 5 and 6 in our definition) of the 14 mouse genes, Wilhelm et al. (2002) showed that the major groups of mouse genes were the same as those in figure 3.5 when the intron 6 region was used but that a somewhat different grouping of genes was obtained when the intron 1 region was used. From this and other observations, they proposed that gene conversion or recombination occurred in the

region of intron 3 (between exon 2 and 3 in our definition). However, since intron regions are often more susceptible to insertion and deletion than exon regions, we examined this hypothesis by constructing phylogenetic trees of rodent *KLRA* genes for exons 1-2 and exons 3-6, separately (figure 3.6). The major gene groups in the phylogenetic tree of extracellular region (figure 3.6B) are exactly same as those of figure 3.5 using full-length coding region. However, in the phylogenetic tree of cytoplasmic and transmembrane region (figure 3.6A), most major gene groups (except rat group I, VI, and mouse singleton gene group MIII) were divided into two subgroups. Interestingly, the two subgroups (i.e. rat group III and mouse group I, II) are often functionally different, and one subgroup includes mainly inhibitory receptors and the other represents activating receptors or pseudogenes. The *KLRA* genes with similar functionality (either inhibitory or activating) tend to cluster together. These results suggest that some degree of recombination might have occurred between extracellular and transmembrane regions in the evolution of rodent *KLRA* genes to promote the functional switch between the pair of inhibitory and activating receptors which share a highly similar lectin-binding domain, as suggested by Arase and Lanier (2004).

Rates of gene expansion by duplication

Figure 3.5 shows that the number of *KLRA* genes in mice and rats expanded rapidly by gene duplication after divergence of the two species. To have a rough idea about how quickly the gene expansion occurred, we estimated the time of occurrence when the *KLRA* repertoire has expanded from a common ancestor. For this purpose, we

used the linearized tree method (Takezaki, Rzhetsky, and Nei 1995). Using the branch length test of the LINTREE program (http://mep.bio.psu.edu), we first tested the molecular clock hypothesis and found that the mouse gene Kψ evolved faster than the average rate at the 1% significance level. However, elimination of this gene did not make much difference in the time estimates of branching points. We therefore constructed a linearized tree under the assumption of molecular clock using the entire set of genes except for the outgroup sequence. The linearized tree using the extracellular region of *KLRA* genes thus obtained is presented in figure 3.7. The timescale for this tree was obtained under the assumption that the putative orthologous mouse and rat *KLRA* gene lineages (indicated by arrows in figure 3.7) diverged about 33 Mya. The rate of nucleotide substitution was estimated to be approximately 3.1×10^{-9} per site per year.

Although the time estimates obtained from this tree are very crude, it appears that the mouse group MI, MII, and MIII genes were generated by repeated gene duplication from a sister gene (labeled with M) of the mouse-B (also rat-1) gene. This sister gene apparently existed about 47 Mya, and during the last 47 My the number of genes expanded from 1 to 14. Therefore, the rate of gene expansion by duplication is (14-1)/47 = 0.28 per gene per My. Figure 3.7 suggests that the rat group I, II, III, IV, and V genes expanded from a sister gene (labeled with R) of the rat-1 (also mouse-B) to 27 duplicate genes. There are five additional incomplete genes (genes 19, 21, 24, 29, and 33) that were not included in figure 3.7 because one or two exons were missing. These genes belong to group IV or V genes (data not shown). Therefore, if

we include these genes, there must have been 31 new genes generated during the last 57 My. This gives an expansion rate of 31/57 = 0.54 per gene per My. This rate is about two times higher than the mouse rate. We performed a similar analysis on the linearized tree of cytoplasmic and transmembrane region and the expansion rates were not significantly different (data not shown).

Table 3.2 shows the rates of gene expansion for some other gene families that are known to have experienced rapid gene duplication. It shows that human *KIR* genes expanded nearly at the same rate as that of the rat *KLRA* genes. *MHC* class I and class II genes are also known to have been subject to rapid gene duplication. Table 3.2, however, shows that the rates of gene expansion for these gene families in primates are about one third of the rate of rat *KLRA* genes. Of course, the estimates of *MHC* rates are very crude because they are based on only a few duplicate genes. The hominoid gene family *Morpheus*, whose function is unknown, has been acclaimed as the fastest evolving gene family so far identified (Johnson et al. 2001). Table 3.2 indicates that the rate of gene expansion for this gene family is nearly twice as high as the rat *KLRA* rate.

Lynch and Conery (2003) estimated the rates of gene duplication for genome-wide genes considering recently duplicated genes (less than 2 My old). Their estimates for animal and plant model organisms were 0.001 to 0.03 per gene per My. These estimates are supposed to include duplicate genes that may soon become pseudogenes. In our computation, however, we considered a relatively longer evolutionary time and may not have included pseudogenes which have already been

deleted from the genome. Therefore, other things being equal, our estimates are expected to be smaller than Lynch and Conery's duplicate rates. Yet, the *KLRA* or *KIR* rates are much higher than the Lynch and Conery rates. This suggests that the *KLRA* and *KIR* genes have duplicated much faster than the genome-wide genes. However, note that our rates are not the same as the mathematical rate of increase of genes by gene duplication (see Nei 1969). We did not use this rate here, because the actual evolutionary pattern is quite different from the mathematical model.

DISCUSSION

We have seen that the BN rat genome contains a total of 33 *KLRA* genes, of which 17 are putative functional genes and 16 are putative pseudogenes. The genome of mouse strain C57BL/6 is known to have 15 *KLRA* genes, of which 11 are functional and four are nonfunctional. These results suggest that the *KLRA* gene family has been subject to birth-and-death evolution and the birth and death rates of genes are higher in rats than in mice. These estimates are very crude, but if this is true, why should rats have a higher rate? One possible answer to this question is the difference in functional constraints that may exist between mice and rats. In the case of the *MADS-box* gene family, which controls the development of flowers and other characters in plants, the type-II gene subfamily is subject to stronger functional constraints than the type-I gene subfamily and shows a slower rate of birth-and-death evolution (Nam et al. 2004). To examine whether a similar mechanism is involved in the evolution of *KLRA* genes, we examined the rates of synonymous and nonsynonymous nucleotide substitution in mice and rats separately using Zhang et al.'s (1998a) method. Contrary to our expectation, however, the rates of synonymous and nonsynonymous substitution were virtually the same for both mice and rats (figure 3.8). This finding also suggests that *KLRA* genes have evolved essentially in a neutral fashion at the nucleotide level. It is then possible that the higher rate of birth-and-death evolution in rats is merely due to the stochastic errors of the birth-and-death process of genes. Note that our analysis of rat *KLRA* gene cluster is based on a single haplotype (rat strain BN/SsNHsd/MCW). It is possible that rat *KLRA* genes have multiple

haplotypes as in the case of mouse *KLRA* genes. It will be interesting to explore the *KLRA* gene cluster in other rat strains and compare its haplotype diversity between mice and rats.

The *Morpheus* gene family, which has the highest gene duplication rate in table 3.2, is known to have a rate of nonsynonymous substitution (r_N) significantly higher than the rate of synonymous substitution (r_S), and this was taken as evidence that the rate of gene duplication has been accelerated by natural selection (Johnson et al. 2001). This is a reasonable explanation. However, this explanation may not be the whole story, because the gene duplication rate can be enhanced substantially without a significant increase of r_N relative to r_S as in the case of rat *KLRA* genes. Note that the gene duplication rate is sometimes influenced substantially by genomic block duplication (e.g., Schable and Zachau 1993; Su and Nei 2001).

As mentioned earlier, the *KIR* genes in primates are structurally very different from the *KLRA* genes in rodents, but they have the same function. The number of member genes in the *KIR* family in primates is roughly similar to that of the *KLRA* family in mice, and the primate *KIR* genes have also evolved very rapidly by repeated gene duplication (Hao and Nei 2005). The similarity of the evolutionary patterns of *KIR* and *KLRA* genes suggests that their rapid evolution is caused by the function of the genes, that is, the interaction with *MHC* class I molecules, rather than by the protein structure encoded by the genes. It is interesting to note that the number of functional genes in the *KIR* gene family in humans and chimpanzees is about 10 though it depends on haplotypes (Hsu et al. 2002), and this number is only slightly

lower than that of the rodent *KLRA* gene family. This finding reminds us that the number of highly polymorphic *MHC* class I loci is only a few (often one to four) though this gene family contains a large number of genes including pseudogenes (Klein and Figueroa 1986; Anzai et al. 2003). It is possible that there is some kind of upper limit for the number of functional *KIR* or *KLRA* genes caused by interaction with MHC molecules.

The origins of *KIR* genes in primates and *KLRA* genes in rodents are quite mysterious. The molecular structures of KIR and Ly49 receptor molecules are so different that it is difficult to know how they originated. However, it is possible to speculate how these different molecules have come to be used in primates and rodents separately. Actually, although *KLRA* genes are concentrated primarily in rodents, the human and baboon genomes contain at least one *KLRA* gene, whether they are functional or not (figure 3.1A). Similarly, the mouse genome is known to have at least two *KIR*-like genes (Hoelsbrekken et al. 2003; Welch, Kasahara, and Spain 2003). Furthermore, multiple *KIR* genes and a single *KLRA* gene have been identified in cattle (McQueen et al. 2002; Storset et al. 2003). It is therefore quite likely that the common ancestor of primates and rodents used both KIR and Ly49 receptors. This view is supported by the fact that *KIR* related genes such as *ILT* genes in humans and *PIR* in mice and *KLRA* related genes such as *KLRC* genes are used in both primates and rodents (Trowsdale et al. 2001). After the primate and rodent lineages diverged, however, *KIR* genes apparently happened to be predominant in the primate lineage, and *KLRA* genes in the rodent. This differential use of NK cell

receptors may have happened by chance or for some adaptive reasons. It is also likely that once one type of NK cell receptors became predominant in an evolutionary lineage, it was probably more efficient to produce them exclusively and the other type of receptors therefore gradually ceased to be used.

At the present time, of course, this is merely a hypothesis. However, this hypothesis can be tested by examining the presence or absence of *KIR* or *KLRA* genes in other orders of placental mammals, marsupials, or even birds and reptiles and studying whether they are expressed as NK cell receptors.

Table 3.1. Rat *KLRA* genes and their predicted characteristics

Gene	Previous Gene	Group	Inhibitory/ Activating	Note
KLRA-1	-	VI	I	Orthologous to mouse *KLRA-B*
KLRA-2	-	III	I	
KLRA-3	*Ly49.9*	III	I	
KLRA-4	*Ly49.29*	III	A	
KLRA-5	-	III	-	Insertion, premature stop codon in exon 3
KLRA-6	-	III	A	
KLRA-7	-	II	I	
KLRA-8	-	I	-	Deletion, premature stop codon in exon 1
KLRA-9	Ly49i2	V	-	Mutation, premature stop codon in exon 4 (CAG→TAG)
KLRA-10	Ly49.19	I	-	Mutation, premature stop codon in exon 1 (TCA→TGA)
KLRA-11	-	V	I	
KLRA-12	-	V	-	Mutation, premature stop codon in exon 5 (AAA→TAA)
KLRA-13	-	II	-	Insertion of 10 nucleotides, premature stop codon in exon 1
KLRA-14	-	I	-	Deletion, premature stop codon in exon 2
KLRA-15	-	V	I	
KLRA-16	-	V	-	Mutation, premature stop codon in exon 5 (CGA→TGA)
KLRA-17	-	V	-	Mutation, premature stop codon in exon 5 (TTG→TAA)
KLRA-18	Ly49.12	I	A	
KLRA-19	-	V	-	Deletion of exon 6
KLRA-20	-	V	-	Mutation, premature stop codon in exon 4 (TAT→TAA)
KLRA-21	-	V	-	Deletion of exon 4
KLRA-22	-	IV	I	
KLRA-23	-	IV	I/A	

KLRA-24	-	IV	-	Deletion of exons 5 and 6
KLRA-25	-	IV	I	
KLRA-26	-	IV	I/A	
KLRA-27	-	IV	I/A	
KLRA-28	-	IV	I/A	
KLRA-29	-	IV	-	Deletion of exons 5 and 6
KLRA-30	-	IV	I	
KLRA-31	-	IV	-	Mutation, premature stop codon in exon 4 (TGG→TAG)
KLRA-32	-	IV	I	
KLRA-33	-	IV	-	Deletion of exon 1

Table 3.2. Rates of gene expansion by duplication for different gene families

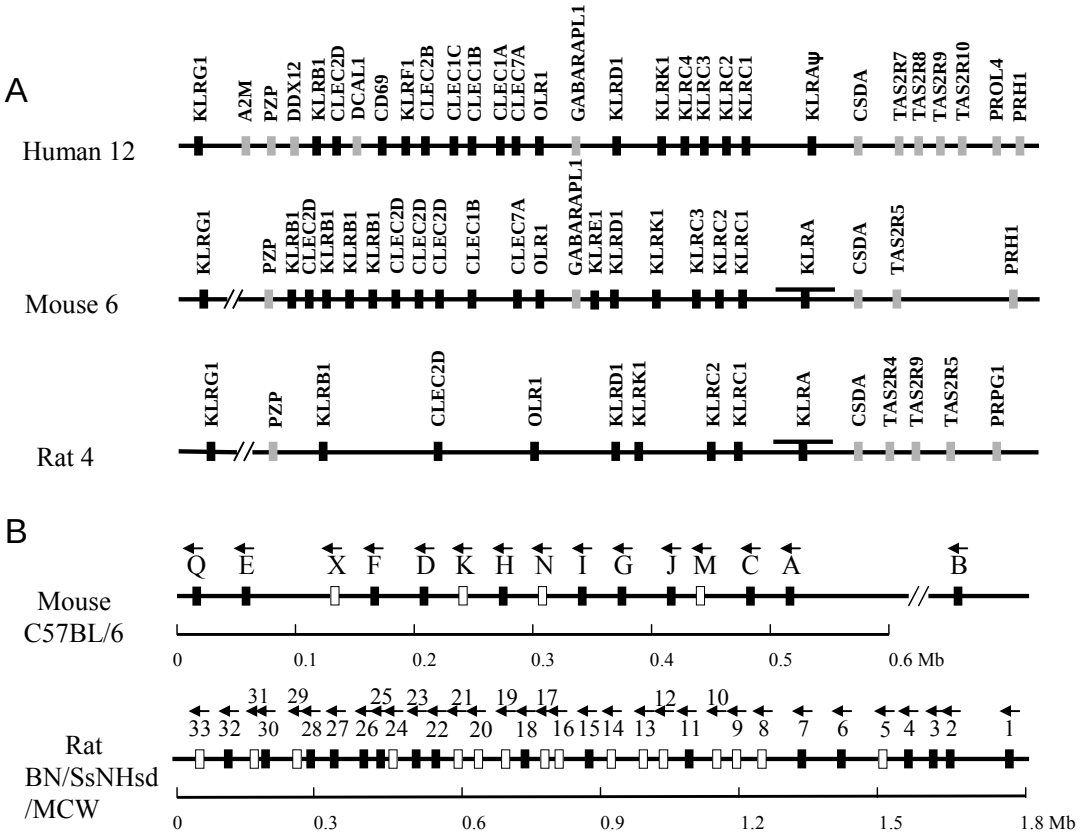
Gene Family	Evoluti-onary Time (MYA)	No. of Genes Increased/ Ancestor	Expansion Rate per MY	Data Used
Class II *MHC* (human *DRB*)	50	4	0.08	Satta et al. (1996); Takahashi et al. (2000)
Class I *MHC* (human)	56	9	0.18	Adams and Parham (2001); Piontkivska and Nei (2003)
Human *KIR*	21	11[a]	0.52	Hao and Nei (2005)
Mouse *KLRA*	47	13	0.28	This study
Rat *KLRA*	57	31[b]	0.54	This study
Morpheus (human)	13	13	1.00	Johnson et al. (2001)
Genome-wide genes			0.001-0.03 [c]	Lynch and Conery (2003)

[a] This number was obtained by considering only KIR domain D1.

[b] Five incomplete genes (19, 21, 24, 29, and 33) are not included in the tree of figure 3.6. However, since genes 19 and 21 belong to group V, and the other three genes belong to group IV, these five genes were added.

[c] Gene duplication rates were obtained from *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis eleganes*, *and Arabidopsis thaliana*. Strictly speaking, these rates are not the same as the gene expansion rate (see text).

Figure 3.1. Genomic organizations of the NKC and the *KLRA* gene families. (A) Comparison of the NKC gene families from humans, mice, and rats. Black boxes represent the C-type lectin-like genes, whereas gray boxes stand for other genes. Only the gene order is shown here, and the distances between genes are not to scale. This information is based on the gene annotations of the UCSC genome browser (http://genome.ucsc.edu/). The annotations for human, mouse, and rat genes are from the freeze of July, October, and June 2003, respectively. (B) The *KLRA* gene clusters of mice and rats. The gene locations are drawn roughly to scale on each chromosome. The arrow sign shows the transcription direction. The rat *KLRA* genes are named according to their genomic positions, starting from "*KLRA-1*" of the telomeric end to "*KLRA-33*" of the centromeric end. Solid boxes represent putative functional genes, and open boxes pseudogenes. The fragmentary gene L in mice is not included here because it has only one exon (Wilhelm, Gagnier, and Mager 2002).


A
Human 12
KLRG1
A2M
PZP
DDX12
KLRB1
CLEC2D
DCAL1
CD69
KLRF1
CLEC2B
CLEC1C
CLEC1B
CLEC1A
CLEC7A
OLR1
GABARAPL1
KLRD1
KLRK1
KLRC4
KLRC3
KLRC2
KLRC1
KLRAψ
CSDA
TAS2R7
TAS2R8
TAS2R9
TAS2R10
PROL4
PRH1
Mouse 6
KLRG1
PZP
KLRB1
CLEC2D
KLRB1
KLRB1
KLRB1
CLEC2D
CLEC2D
CLEC2D
CLEC1B
CLEC7A
OLR1
GABARAPL1
KLRE1
KLRD1
KLRK1
KLRC3
KLRC2
KLRC1
KLRA
CSDA
TAS2R5
PRH1
Rat 4
KLRG1
PZP
KLRB1
CLEC2D
OLR1
KLRD1
KLRK1
KLRC2
KLRC1
KLRA
CSDA
TAS2R4
TAS2R9
TAS2R5
PRPG1
B
Mouse
C57BL/6
Q E X F D K H N I G J M C A B
0 0.1 0.2 0.3 0.4 0.5 0.6 Mb
Rat
BN/SsNHsd
/MCW
31 29 25 23 21 19 17 12 10
33 32 30 28 27 26 24 22 20 18 16 15 14 13 11 9 8 7 6 5 4 3 2 1
0 0.3 0.6 0.9 1.2 1.5 1.8 Mb

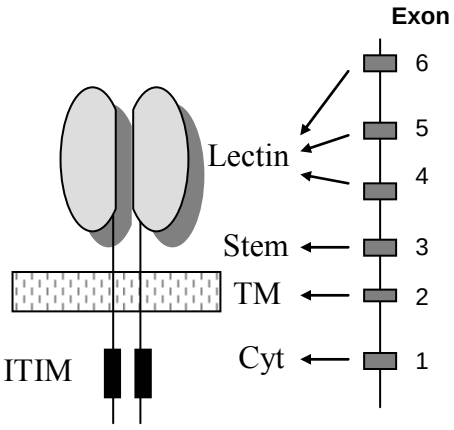


Figure 3.2. The relationships between the exon-intron organization and protein domains in *KLRA* genes. Cyt, cytoplasmic domain. TM, transmembrane domain. Stem, stem domain. Lectin, C-type lectin-like domain. Adapted from Takei et al. (1997). Note that the exon numbering here is different from the one commonly used for mouse *KLRA* genes because the 5' non-coding exon (exon 1 in mouse *KLRA* genes) was not included in this organization. Here we only showed inhibitory receptors as an example. The activating receptors have the same exon-intron organization as inhibitory receptors.

Figure 3.3. Phylogenetic tree of 28 full-length rat *KLRA* genes (790 nucleotides used). The tree was constructed by the NJ method using Jukes-Cantor distances. The bootstrap values based on 1000 replications are shown above the branches (only values higher than 50% shown). Essentially the same tree topology was obtained by using parsimony methods. Each group is shown by a bracket, labeled with a Roman numeral. Letter "ψ" represents potential pseudogene.

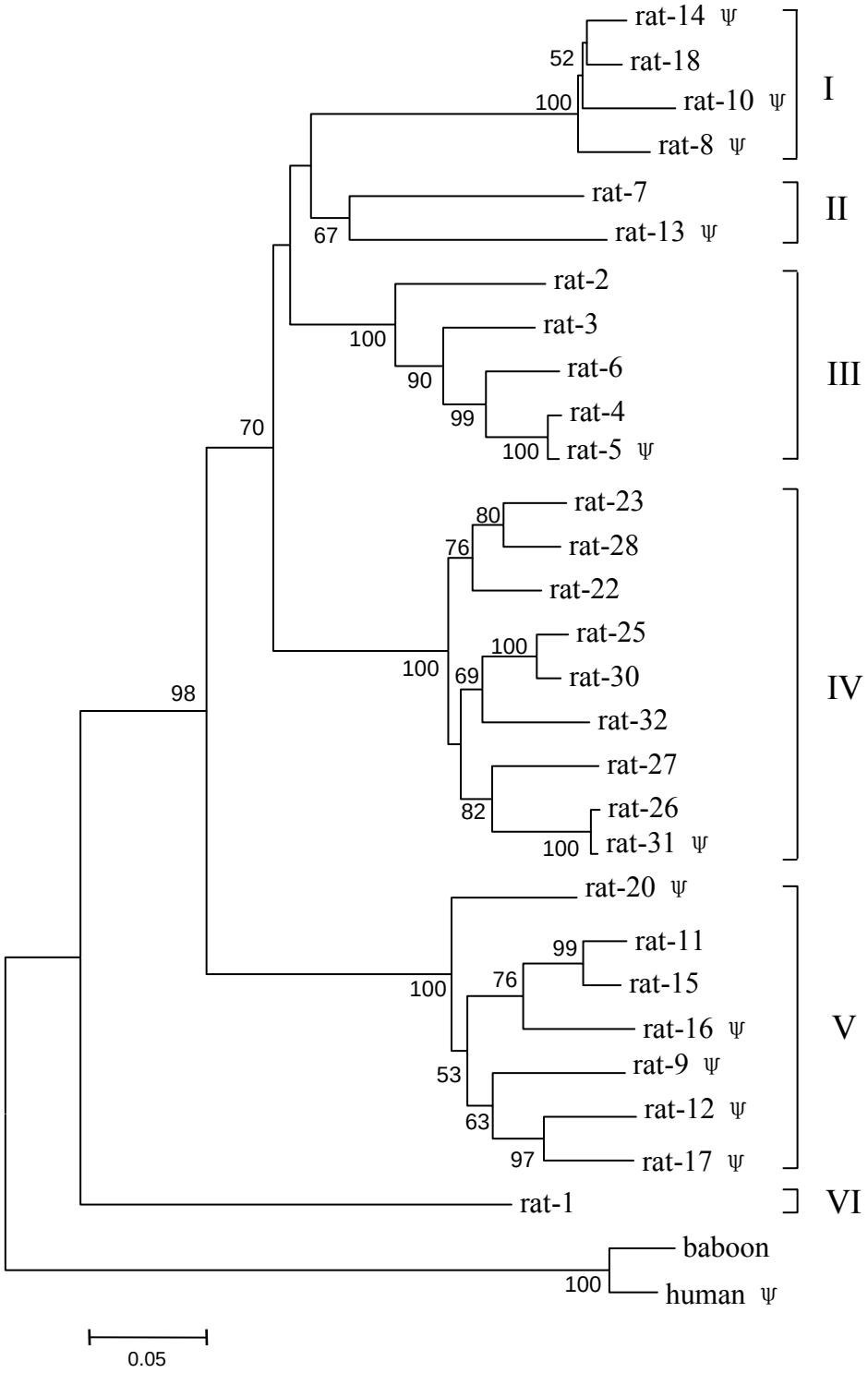

rat-14 ψ
rat-18
rat-10 ψ
rat-8 ψ
rat-7
rat-13 ψ
rat-2
rat-3
rat-6
rat-4
rat-5 ψ
rat-23
rat-28
rat-22
rat-25
rat-30
rat-32
rat-27
rat-26
rat-31 ψ
rat-20 ψ
rat-11
rat-15
rat-16 ψ
rat-9 ψ
rat-12 ψ
rat-17 ψ
rat-1
baboon
human ψ
52
100
67
100
90
99
100
70
80
76
100
100
69
98
82
100
99
76
100
53
63
97
I
II
III
IV
V
VI
0.05

Figure 3.4. Comparison of the genomic organization of *KLRA* genes and repetitive elements between two recently duplicated genomic blocks. The upper sequence represents the first genomic block, which extends from the 5'-end of the gene *KLRA-23* to the 3'-end of the gene *KLRA-26*. The lower sequence represents the second genomic block, which extends from the 5'-end of *KLRA-28* to the 3'-end of *KLRA-31*. Each sequence is approximately 140 kb long. The exons of each gene and the repetitive elements were drawn roughly to scale on the genomic sequence. The LINE elements include two major subfamilies L1 and Lx, with light grey color. LTR's are long terminal repeat retrotransposons including ERV and MaLR etc. The SINE elements include primarily ID and B1 subfamilies. The simple repetitive elements represent various microsatellite loci.

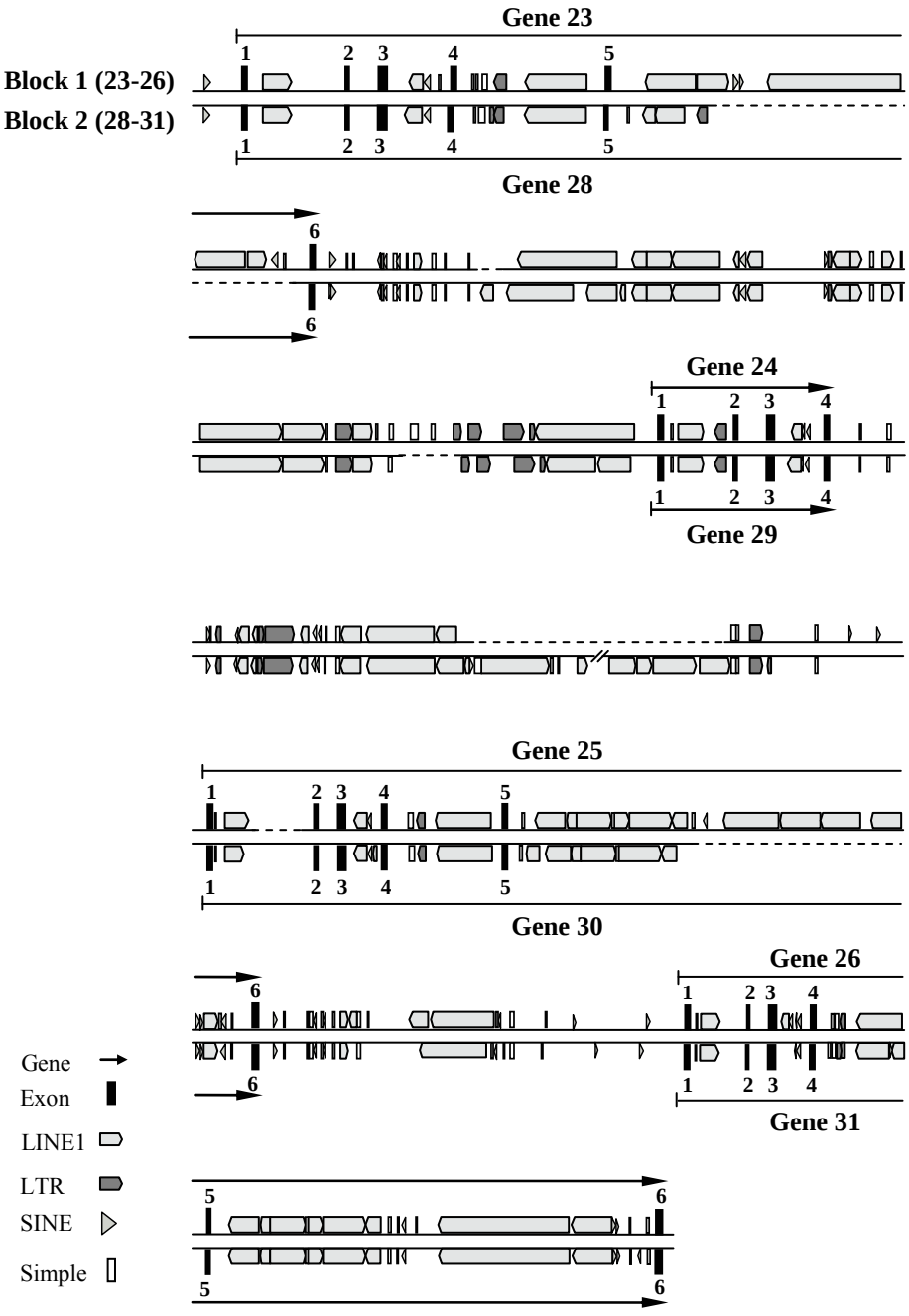

Gene 23
Block 1 (23-26)
Block 2 (28-31)
Gene 28
Gene 24
Gene 29
Gene 25
Gene 30
Gene 26
Gene 31
Gene
Exon
LINE1
LTR
SINE
Simple

Figure 3.5. Phylogenetic tree of rodent and other mammalian *KLRA* genes. This tree was constructed by the NJ method using Jukes-Cantor distance for the coding regions of 15 mouse, 28 rat, 2 primate *KLRA* genes, and the rat *KLRC1* gene, which was used as an outgroup (714 nucleotides used). Each group is shown by a bracket. The Roman numbers of rat *KLRA* gene groups are the same as that of figure 3.3. Each mouse *KLRA* group is labeled with a Roman numeral followed by the letter "M". The letter "ψ" represents potential pseudogene. The parsimony tree was virtually the same as the NJ tree if we disregard the branching patterns with low bootstrap values. Therefore, it is not presented here.

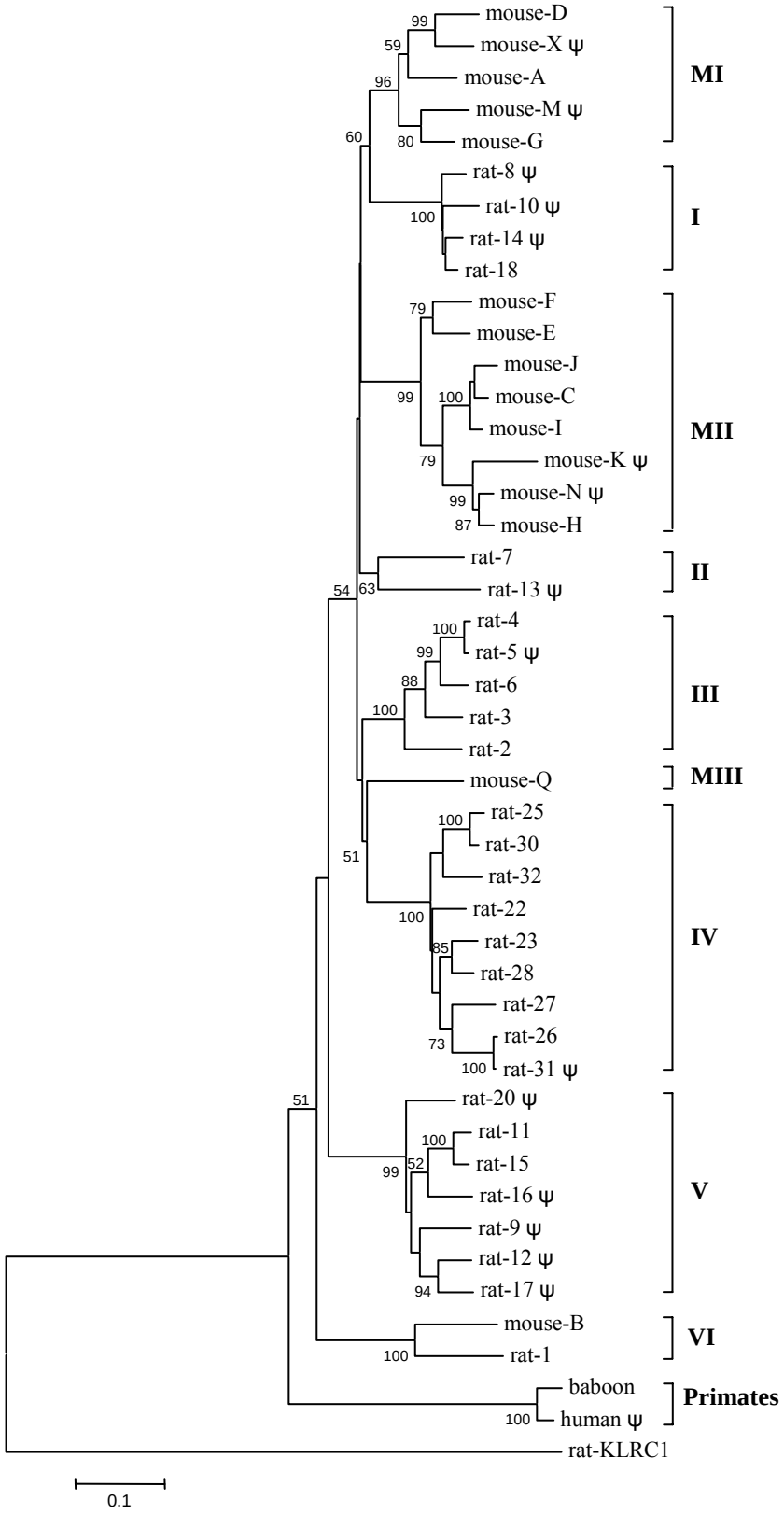

mouse-D
mouse-X ψ
mouse-A
mouse-M ψ
mouse-G
MI
rat-8 ψ
rat-10 ψ
rat-14 ψ
rat-18
I
mouse-F
mouse-E
mouse-J
mouse-C
mouse-I
mouse-K ψ
mouse-N ψ
mouse-H
MII
rat-7
rat-13 ψ
II
rat-4
rat-5 ψ
rat-6
rat-3
rat-2
III
mouse-Q
MIII
rat-25
rat-30
rat-32
rat-22
rat-23
rat-28
rat-27
rat-26
rat-31 ψ
IV
rat-20 ψ
rat-11
rat-15
rat-16 ψ
rat-9 ψ
rat-12 ψ
rat-17 ψ
V
mouse-B
rat-1
VI
baboon
human ψ
Primates
rat-KLRC1
0.1

Figure 3.6. Phylogenetic trees of rodent and primate *KLRA* genes using different region. The trees were constructed using p-distances. The bootstrap values based on 1000 replications are shown above the branches (only values higher than 50% shown). (A) Tree based on exons 1-2. 187 nucleotides used. “I” in the parenthesis represents inhibitory receptors. “A” represents activating receptors. “?” represents either pseudogene or functionally unknown. (B) Tree based on exons 3-6. 527 nucleotides used.

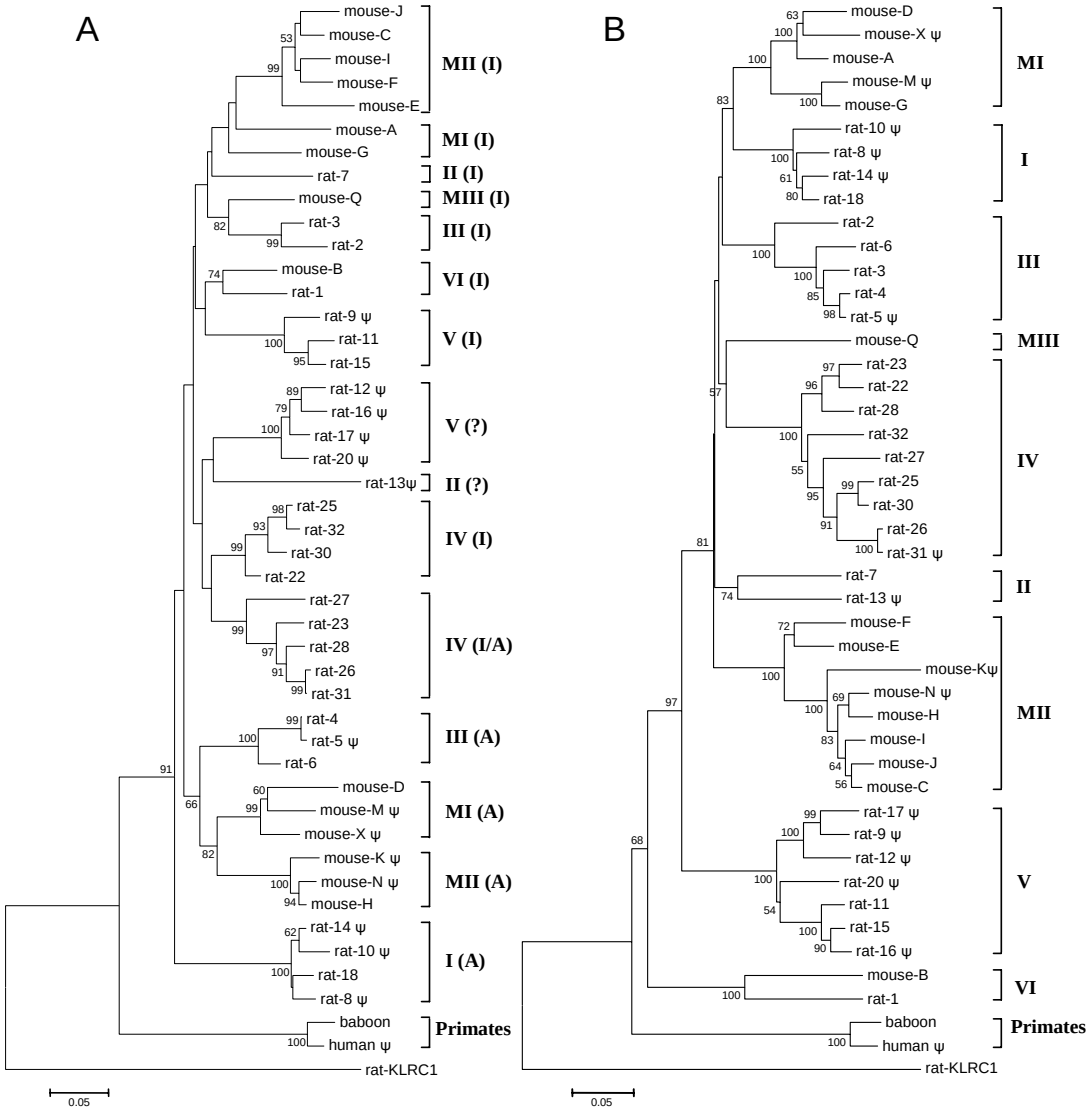
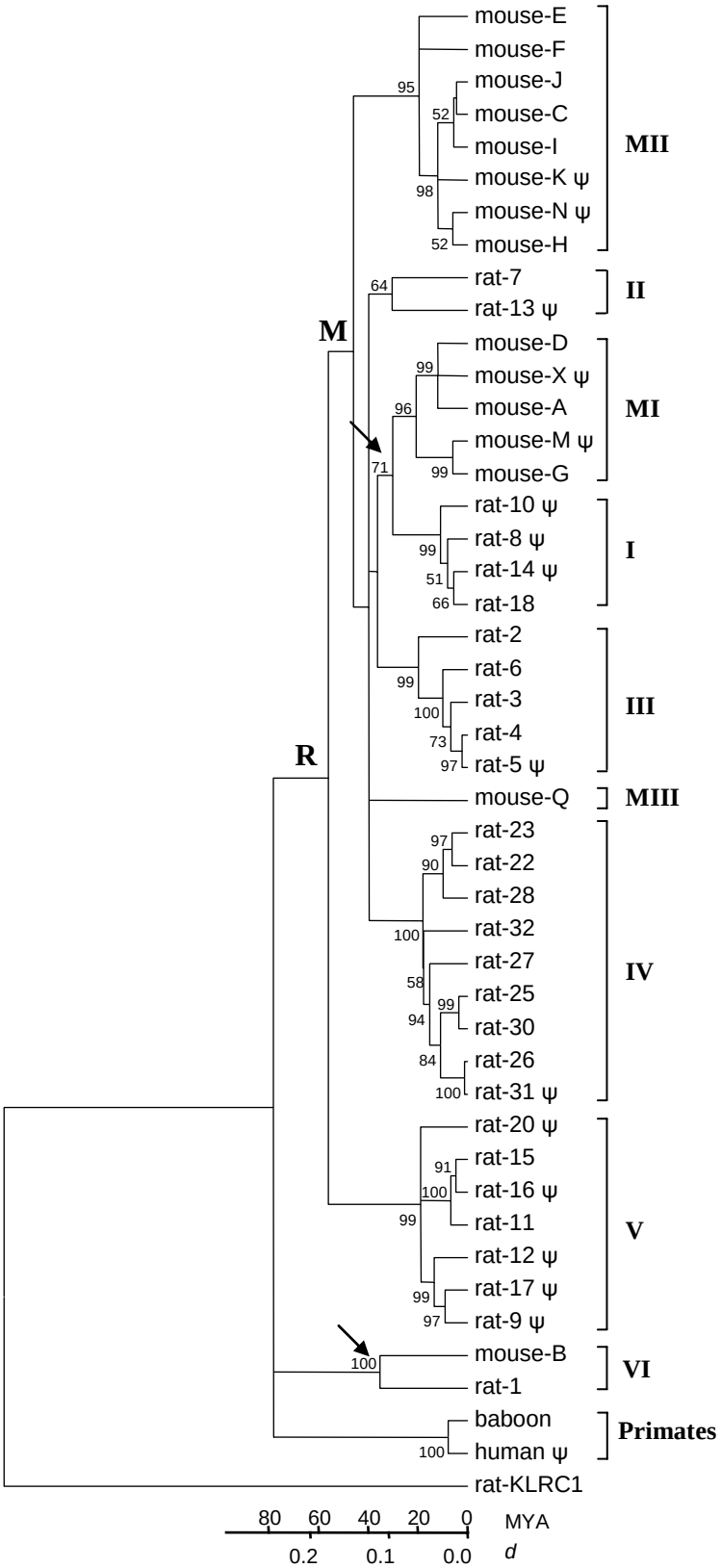

A
mouse-J
mouse-C
mouse-I
mouse-F
mouse-E
MII (I)
mouse-A
mouse-G
MI (I)
rat-7
II (I)
mouse-Q
MIII (I)
rat-3
rat-2
III (I)
mouse-B
rat-1
VI (I)
rat-9 ψ
rat-11
rat-15
V (I)
rat-12 ψ
rat-16 ψ
rat-17 ψ
rat-20 ψ
V (?)
rat-13ψ
II (?)
rat-25
rat-32
rat-30
rat-22
IV (I)
rat-27
rat-23
rat-28
rat-26
rat-31
IV (I/A)
rat-4
rat-5 ψ
rat-6
III (A)
mouse-D
mouse-M ψ
mouse-X ψ
MI (A)
mouse-K ψ
mouse-N ψ
mouse-H
MII (A)
rat-14 ψ
rat-10 ψ
rat-18
rat-8 ψ
I (A)
baboon
human ψ
Primates
rat-KLRC1
0.05
B
mouse-D
mouse-X ψ
mouse-A
mouse-M ψ
mouse-G
MI
rat-10 ψ
rat-8 ψ
rat-14 ψ
rat-18
I
rat-2
rat-6
rat-3
rat-4
rat-5 ψ
III
mouse-Q
MIII
rat-23
rat-22
rat-28
rat-32
rat-27
rat-25
rat-30
rat-26
rat-31 ψ
IV
rat-7
rat-13 ψ
II
mouse-F
mouse-E
mouse-Kψ
mouse-N ψ
mouse-H
mouse-I
mouse-J
mouse-C
MII
rat-17 ψ
rat-9 ψ
rat-12 ψ
rat-20 ψ
rat-11
rat-15
rat-16 ψ
V
mouse-B
rat-1
VI
baboon
human ψ
Primates
rat-KLRC1
0.05

Figure 3.7. Linearized tree of *KLRA* genes from rodent and primate species based on exons 3-6. The timescale was calibrated with 33 Mya under the assumption that the mouse and rat lineages diverged about 33 Mya (arrow signs).


mouse-E
mouse-F
mouse-J
mouse-C
mouse-I
mouse-K ψ
mouse-N ψ
mouse-H
MII
rat-7
rat-13 ψ
II
M
mouse-D
mouse-X ψ
mouse-A
mouse-M ψ
mouse-G
MI
rat-10 ψ
rat-8 ψ
rat-14 ψ
rat-18
I
rat-2
rat-6
rat-3
rat-4
rat-5 ψ
III
R
mouse-Q
MIII
rat-23
rat-22
rat-28
rat-32
rat-27
rat-25
rat-30
rat-26
rat-31 ψ
IV
rat-20 ψ
rat-15
rat-16 ψ
rat-11
rat-12 ψ
rat-17 ψ
rat-9 ψ
V
mouse-B
rat-1
VI
baboon
human ψ
Primates
rat-KLRC1
80 60 40 20 0 MYA
0.2 0.1 0.0 d

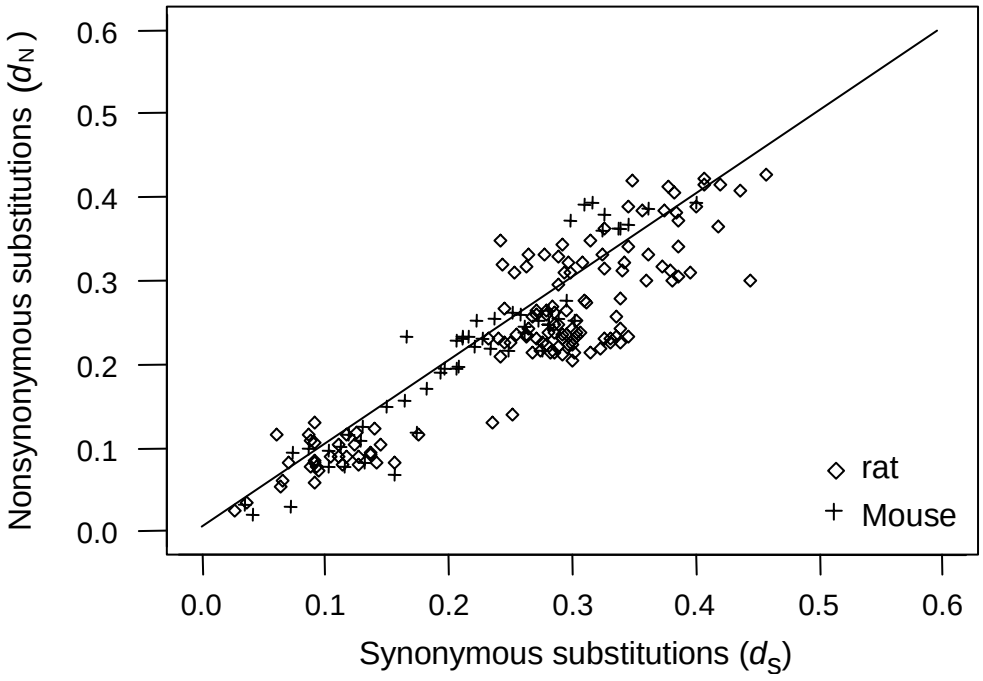


Figure 3.8. Relationships between the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) for mouse (n=11) and rat (n=17) putative functional Ly49 genes.

CHAPTER 4

HETEROGENEOUS BUT CONSERVED NATURAL KILLER RECEPTOR GENE COMPLEXES IN FOUR MAJOR ORDERS OF MAMMALS

SUMMARY

The natural killer (NK) receptor gene complex (NKC) encodes a large number of C-type lectin-like receptors which are expressed on NK and other immune-related cells. These receptors play an important role in regulating NK-cell cytolytic activity protecting cells against virus infection and tumorigenesis. To understand evolutionary history of the NKC, we characterized the C-type lectin-like NKC genes and their organization from four major orders of placental mammals, primates (human), rodents (mouse and rat), carnivores (dog), and artiodactyls (cattle), and then conducted phylogenetic analysis of these genes. The results indicate that the NKC of placental mammals is highly heterogeneous in terms of the gene content and rates of birth-and-death of different gene lineages, but the NKC is also remarkably conserved in its gene organization and persistence of orthologous gene lineages. Among the 28 identified NKC gene lineages, four, *KLRA1, KLRB1, CLEC2D,* and *CLEC4A/B*, have rapidly expanded in rodents only. The high birth-and-death rate of these four gene families might be due to functional differentiation driven by positive selection. Identification of putative NKC sequences in opossum and chicken genomes implies that the expansion of the NKC gene families might have occurred before the radiation of placental mammals but after the divergence of birds from mammals.

INTRODUCTION

Natural killer (NK) cells are a group of lymphocytes which have intrinsic cytolytic activity against certain virus-infected and tumor cells. They are crucial in innate immunity as demonstrated by their ability to kill target cells without prior exposure to pathogens (Seaman 2000). The ligands of the NK cell receptors (NKR) are mostly major histocompatibility complex (MHC) class I molecules. Most NKRs belong to two large gene families, the immunoglobulin superfamily (IgSF) and the C-type lectin superfamily (CLSF) (Trowsdale et al. 2001; Yokoyama and Plougastel 2003; Kelley, Walter, and Trowsdale 2005), clustered at different genomic region. The CLSF proteins are characterized by the possession of at least one C-type lectin-like domain (CTLD). The CLSF proteins are classified into 14 groups designated I – XIV, and each group is distinguished from other groups primarily by the type of additional domains its members share (Drickamer and Fadden 2002). The sequences of the CTLDs from the different groups are alignable, whereas the sequences outside of the CTLD are generally not.

In the human species, genes encoding the CTLD-bearing NKRs are clustered in a single region, the natural killer receptor gene complex (NKC) on chromosome 12p13 (Yokoyama and Plougastel 2003; Kelley, Walter, and Trowsdale 2005). The C-type lectin genes in the NKC belong to groups II and V of the 14 CLSF groups. While many group V proteins are known to function as NKRs, the remaining group V and all group II proteins are not expressed in NK cells and appear to have different functions (Yokoyama and Plougastel 2003). The original function of the CTLD

appears to have been binding of carbohydrates in the presence of Ca^{2+} ions. This function has been retained by most of the CTLDs in the 14 groups, but the CTLDs of the group V CLSFs have lost it and have become involved in protein-protein interactions instead (Drickamer and Fadden 2002).

Previous studies (Yokoyama and Plougastel 2003; Kelley, Walter, and Trowsdale 2005) described the conserved genomic structure of NKC in human, mouse, and rat. The gene content of the NKC from other species, the extent of variation in gene numbers among species, and the long-term evolutionary history of the NKC remain largely unknown. To address these issues, we searched six mammalian and one avian genomes, specifically, human, mouse, rat, dog, cattle, opossum, and chicken, for either the entire NKC genomic segment or the putative NKC sequences. By this large-scale, multi-species comparison, we have been able to trace the important evolutionary changes the NKC has undergone during the last 300 million years of its existence.

MATERIALS AND METHODS

Characterization of C-type lectin-like NKC sequences

We searched human (*Homo sapiens*), mouse (*Mus musculus*, strain C57BL/6J), and rat (*Rattus norvegicus*, strain BN/SsNHsd/Mcwi) genome sequences from ENSEMBL (www.ensembl.org), assembly versions v27.35a, v27.33c.1, and v28.3e.1, respectively; dog (*Canis familiaris*, breed boxer) and cattle (*Bos taurus*, breed Hereford) genome sequences deposited in NCBI, build number 2.1; as well as opossum (*Monodelphis domestica*) and chicken (*Gallus gallus*, inbred line UCD 001) sequences deposited in UCSC (http://hgdownload.cse.ucsc.edu/downloads.html), October and February 2004. In the case of human, mouse, rat, dog, and cattle genomes, we first identified roughly the location of the NKC using the computer program TBLASTN (Zhang et al. 1998b, ftp://ftp.ncbi.nih.gov/blast/), using the known mouse NKC protein sequences as queries. We then retrieved the entire sequence of the NKC-bearing chromosome and used it as a database to perform homology search for the C-type lectin-like NKC sequences using the TBLASTN program. By using high expected E-values (E=10), we ensured that all C-type lectin-like NKC sequences would be retrieved. For the human and mouse NKC analysis, the queries were all annotated CTLD (INTERPRO domain ID: IPR001304)-containing protein sequences from the NKC region retrieved from ENSEMBL by using the data mining tool, BioMart (http://www.ensembl.org/Multi/martview). For the poorly annotated genomes of rat, dog, and cattle, we used as queries not only the ENSEMBL-annotated C-type

lectin-like sequences in NKC, but also all the human and mouse sequences identified as described above. The sequence fragments which contain only part or none of CTLD were not considered.

Because of low sequence similarity outside of the CTLD in the opossum and chicken genomes, we limited our search to the CTLD of homologous NKC sequence only. The CTLD sequences of the representative NKC genes from human, mouse, rat, dog, and cattle genomes identified above were used as queries to search the opossum and chicken genomes. We also searched one of the largest chicken expressed sequence tag (EST) databases, BBSRC chickEST (http://www.chick.umist.ac.uk/), for sequences homologous to the CTLD of mammalian NKC genes.

Phylogenetic analysis and nomenclature

We constructed phylogenetic tree from protein sequences aligned by using the MAFFT program with the option of E-INS-I, which exploits an iterative refinement method incorporating local pairwise alignment information (http://timpani.genome.ad.jp/%7Emafft/server/). We used the neighbor-joining (NJ) method (Saitou and Nei 1987) in the computer program MEGA3 (Kumar, Tamura, and Nei 2004), with the pairwise deletion option. The proportional amino acid differences (*p*-distances) were used because they are known to give higher resolution in topology when the number of sequences is large. The tree was evaluated by 1,000 bootstrap resamplings. In the case of *CLEC2D* genes, which contain several pseudogenes in rodents, we constructed a NJ tree of the nucleotide sequences with Jukes-Cantor (JC) distance. The nucleotide sequences of *CLEC2D* genes were

retrieved by using the computer program Spidey (http://www.ncbi.nlm.nih.gov/spidey/index.html). The nucleotide sequences of all ancestral nodes of the phylogenetic tree were inferred from the present-day putatively functionally *CLEC2D* sequences using the program ANC-GENE (Zhang and Nei 1997).

For the gene names, we used the HUGO gene nomenclature (http://www.gene.ucl.ac.uk/nomenclature/) and MGNC (http://www.informatics.jax.org/mgihome/nomen/) with minor modification taking into account their phylogenetic relationship. Except two gene names, *CD69* and *OLR1*, all the remaining NKC genes were named starting with either KLR or CLEC. The forth and fifth letters (or numbers) refer to different orthologous genes or gene groups. We define an orthologous gene (or gene group) as a group of sequences from different species, all derived from a single most recent common ancestor.

The NKC described in some previous studies does not contain the *CLEC4* cluster of genes, which includes *CLEC4A/B/C*, *CLEC4D*, and *CLEC4E* etc. However, in this study, we included this cluster into the NKC because 1) its encoded proteins share the same domain organization with other NKC proteins; 2) in rodents this cluster is located between *KLRG1* and *KLRB1* of the NKC; 3) some of the genes in this cluster are known to be expressed on NK cells; 4) the genes encode inhibitory and activating immunoreceptors, as do several NKC genes.

RESULTS

To gain an insight into evolution of the NKC, in the first phase of the study, we searched the genomic databases of five mammalian species for C-type lectin-like NKC genes. The five species represent four orders of placental mammals: Primates (human), Rodentia (mouse and rat), Carnivora (dog), and Artiodactyla (cattle). We also determined the position of the individual genes in the genomes and classified them in terms of their homology to genes identified previously by other investigators. The results of this analysis are summarized in figure 4.1 and 4.2. Figure 4.1 compares the organization of the NKCs in the five species and figure 4.2 shows the phylogenetic relationships among the identified sequences. For additional information about the individual genes, e.g., alternative gene names and original references, see table 4.1.

Overall Organization of the NKC

In four of the five species, the NKC occupies a single chromosomal region (12p13 in human, 6 in mouse, 4 in rat, and 27 in dog). In cattle, however, it is divided between two chromosomes, 1 and 5. In all the five species, the NKC contains genes of two of the 14 CLSF groups – II and V. In the three non-rodent mammals, these two groups occupy separate genomic segments, which are adjacent to each other in human and dog, but on separate chromosomes in cattle (figure 4.1). In the two rodent species, one of the group V genes (*KLRG1*) has been transposed to a location at the other end of the group II segment. The transposition must have, therefore, occurred before the divergence of mice from rats, but presumably after the divergence of rodents from the

other three orders. The genes in the group II segment are arranged in the same order in the five species, except that the segment appears to be inverted in the dog. The possibility that the inversion is an artifact of an erroneous contig assembly could not be excluded, however. The group II segment does not seem to contain any proven or implied NK cell receptor genes in any of the five species.

The group V segment contains NK cell receptor genes (designated *KLR*, killer cell lectin-like receptor), as well as other genes encoding the CTLD (designated *CLEC*, C-type lectin), whose function is largely unknown, except that their products do not appear to act as NKRs. One of the *CLEC* genes (*CLEC2D*, previously called *Ocil* or *Clr*) apparently codes for the ligand of one of the *KLR* genes (*KLRB1*, see refs Iizuka et al. 2003; Carlyle et al. 2004). The *KLR* genes cluster at one end of the group V segment, except *KLRB1* and *KLRF* which are intercalated among the *CLEC2* genes (figure 4.1). Although the individual genes appear to have expanded and contracted by duplications and deletions during the evolution of the five species, the overall arrangement of the genes or gene groups has remained remarkably stable. This observation suggests that the expansions have occurred largely by tandem duplications and that the duplicated genes have remained clustered without subsequent rearrangements. The general order of the genes/gene clusters in the group V segment of the five species is:
KLRG1-KLRB1-CLEC2D-CD69-CLEC15A-CLEC1C-CLEC1D-CLEC1B-CLEC9A-CLEC1A-CLEC7A-OLR1-KLRD1-KLRK1-KLRC-KLRA1, with some minor variations. Exceptions to this order are found mainly in cattle, in which some of the

orthologs seemingly missing in the cattle group V segment are present in the unassigned genomic contigs. It is therefore possible that when the genome is completely assembled, the genes will find their way back into the NKC. The orphan group V genes of cattle include *KLRB1*, *KLRF1*, *KLRF2*, *CLEC1D*, *CLEC1B*, *CLEC9A*, *KLRH1*, and *KLRA1*. They are clearly orthologs to their name-sakes in the other species (figure 4.2), but their position in the genome is uncertain at this stage.

Phylogenetic relationships

We collected 172 C-type lectin-like NKC sequences of the five mammalian species and used their corresponding protein sequences of CTLD to construct a phylogenetic tree by the NJ method (figure 4.2). In the collection, we included pseudogenes with complete CTLD-encoding part, undisrupted by any stop codons. The alignment is published in supporting information as a text file and the full-length amino acid sequences of those putative functional NKC sequences can be found at our database (http://www.bio.psu.edu/People/Faculty/Nei/Lab/databases.html). The tree served two main purposes: to identify orthologous sequences in the five species and to examine changes during evolution of the different lineages of orthologous genes. Generally, the identification of orthologs posed no great difficulty since their clustering on the tree was supported by high bootstrap values. The only confounding situations were those in which the gene apparently duplicated once or more times in a given species. In such cases, identification of a single gene in that species as an ortholog was not possible. Examples are the expansions of the *CLEC4A/B/C*, *CLEC2D*, *KLRA1*, and *KLRB1* genes in the two rodent species. Some of these

duplications took place before and others after the divergence of mice and rats. For the purpose of the present discussion, we treat each such cluster as an orthologous group of the genes in other species.

On the tree in figure 4.2, we identified 28 lineages of orthologous genes, including 11 *KLR*s, 15 *CLEC*s, 1 *OLR1* and 1 *CD69*. In addition, there were five singleton genes, of which four were from cattle and were mostly unmapped, while one, *CLEC16*, was found in the dog and seemed to be related to the *KLRA* genes. For simplicity, we didn't take these 5 genes into account in the following discussion. Of the 28 gene lineages, 15 had an ortholog in each of the five species examined. The absence of an ortholog in one or more species in the remaining 13 lineages could be explained either by the incompleteness of the genomic data in the databases (especially in the dog and cattle genomes) or gene deletions in one species or in the ancestors of some of the species. Two of the 28 gene lineages (*CLEC15A* and *KLRF2*) were identified in this study. The *CLEC15A* gene was found in the mouse, rat, dog, and cattle, while the human ortholog may have been lost. The *KLRF2* gene was identified in the human and cattle, while the human sequence appears to be a pseudogene; the gene has apparently been lost in the two rodent species and has not been found in the dog genome thus far. The remaining 26 human and mouse genes were either identified previously or annotated by genome project (table 4.1). Most of the orthologous genes in the dog and cattle had not been described previously.

The determination of phylogenetic relationships among the different orthologous lineages has proved to be more difficult than the identification of the

orthologs themselves, especially among the *KLR* genes. Most if not all of the lineages apparently diverged before the divergence of the four mammalian orders sampled, but presumably at different times, since some of the lineages appear to be related more closely to one another than others. Among the *CLEC* genes, two well-supported monophyletic clades of orthologous lineages can be recognized. They are the CLEC4 clade composed of the *CLEC4A*-*E* and *CLEC4N* genes, and the CLEC2 clade containing the *CLEC2A*, *CLEC2B*, *CLEC2D*, and *CD69* genes. On the tree, the *CLEC15A* gene also groups with the CLEC2 clade, although the bootstrap support for the grouping is low in this case. Other monophyletic clades, admittedly not well-supported, are the CLEC1 clade composed of the *CLEC1A*-*D* genes and possibly also the *CLEC7A* and *OLR1* genes; and the KLRB/F clade containing the *KLRB1*, *KLRF1*, and *KLRF2* genes. The CLEC4 clade encompasses all the group II CLSF genes of the NKC; all the other clades contain group V CLSF genes. The groupings into clades are reflected in the nomenclature of these genes based on the standardized HUGO nomenclature with minor modification (table 4.1). For instance, *CLEC12A* gene has been renamed to *CLEC1C*, since it is more closely related to CLEC1 than to any other genes. Similarly, *CLEC6A* gene in human has been renamed to *CLEC4N*.

Interestingly, the members of each clade tend to cluster not only on the tree (figure 4.2) but also on the genomic map of the NKC (figure 4.1). As pointed out earlier, the physical clustering suggests that the members arose by tandem duplications and that most of them remained in place after the duplication over tens

of millions of years. This hypothesis is supported also by the observation that the transcriptional orientation of the orthologous genes in the different species has largely been conserved (figure 4.1).

Rapid expansion and contraction of rodent NKC genes

The lengths of the NKCs in human, dog, and cattle are comparable (2.8 Mb, 2.4 Mb, and 3.3 Mb, respectively), as are also the gene numbers (29, 22, and 32, respectively). However, the mouse and rat NKCs encompassing 8.7 and 10.3 Mb, respectively, are 2.7 and 3.6 times longer than the human NKC, respectively. Corresponding to this length differences are differences in the number of genes: the human NKC (29 genes) contains only about one half of the genes present in the mouse NKC (57 genes) and slightly more than one third of the genes present in the rat (75 genes). These observations suggest that an expansion of genes occurred in the mouse/rat lineage compared to the lineage leading to the three other species, and in the rat lineage in comparison to the mouse lineage.

This suggestion is borne out by the phylogenetic analysis. The tree in figure 4.2 indicated that several genes in the complex have duplicated repeatedly in the mouse and rat lineages. Gene duplications in other species (e.g., *CLEC1B* in human and *KLRI* in cattle) had occurred to much lesser extent. Rodent gene expansions occurred in both the *KLR* and *CLEC*-type genes and affected some genes (or some parts of the NKC) more than others. The *KLRA* gene underwent by far the most extensive expansions in the two rodent species. To keep the NKC maps and the tree within acceptable limits, we have limited the presentations of the mouse and rat *KLRA*

cluster to just a few genes in figures 4.1 and 4.2. In reality, however, the *KLRA* cluster of the mouse contains 15 genes and that of the rat 33 genes (Wilhelm, Gagnier, and Mager 2002; Hao and Nei 2004; Nylenna et al. 2005). In the CLEC category, the two gene lineages most subjected to duplications are *CLEC4A/B/C* and *CLEC2D*. In the former, there are five and four genes in the mouse and rat, respectively. In the latter, there are eight and eleven genes in the mouse and rat, respectively, of which four and two, respectively, might be nonfunctional. A phylogenetic tree of the mammalian *CLEC2D* genes including pseudogenes, constructed from nucleotide sequences, is shown in figure 4.3. The tree suggests the existence of two ancestral CLEC2D lineages in the mammals tested, CLEC2Da and CLEC2Db. The CLEC2Db lineage might have become extinct in the two rodent species and the cattle. All rodent *CLEC2D* genes are derived from the CLEC2Da lineage and they all form a single monophyletic clade supported by a high bootstrap value, suggesting that they started to diverge after the divergence of rodents from other mammals. Because of this phylogenetic relationship, we use a different nomenclature for the CLEC2D genes in rodents than the mouse genomic nomenclature committee (MGNC; see figure 4.1 legend for details). From these observations, it seems that most prone to amplifications are some of the genes at or near the *KLRA* and *CLEC4* ends of the NKC. This observed polarity may be accidental. It may not be coincidental, however, that one of the rapidly expanding genes (*CLEC2D*) has been identified recently (Iizuka et al. 2003; Carlyle et al. 2004) as encoding the ligand of one of the KLR genes in the NKC (*KLRB1*). Interestingly, *KLRB1* and *CLEC2D* genes are

intermingled within the NKC. It is tempting to speculate that the expansions might be related to the coevolution of the ligand with its receptor. The fact that the orthologies of the expanded genes in rodents are difficult to ascertain suggest that the sequences might have diverged very quickly after the duplications. On the other hand, however, the tree in figure 4.3 suggests that some of the duplications took place before the divergence of the mouse and rat species more than 30 Mya (Nei, Xu, and Glazko 2001). Analysis of the presence and absence of genes in the different species suggests that corresponding to the gains there have also been gene losses in rodents (data not shown). Overall, rodent NKCs evolve more rapidly and are less stable than NKCs of non-rodents because of the high birth-and-death rate.

Positive selection during the early expansion of the *CLEC2D* genes in rodents

On the phylogenetic tree in figure 4.3, the branch A leading to the clade of the expanded rodent *CLEC2D* genes is rather long, suggesting the possibility that positive selection might have been involved in the expansion. To test this possibility, we inferred the ancestral CLEC2D sequences from the functional CLEC2D sequences (pseudogenes excluded) and then estimated the number of nonsynonymous (a_N) and synonymous (a_S) substitutions per sequence per branch. The results show that, after speciation at node a, there were 13 nonsynonymous substitutions but no synonymous substitution on branch A (figure 4.3). The Fisher's exact test that compares the ratio of a_N/a_S with the expected ratio (*N/S*) gives statistically significant support (P=0.04) for the positive selection hypothesis. During the early divergence of the rodent *CLEC2Da* genes (branch A, B, and C), there are a total of 47

nonsynonymous substitutions and only four synonymous substitutions. For comparison, in the lineage leading to the cattle *CLEC2Da* genes, there were 17 nonsynonymous and 35 synonymous and substitutions. These observations suggest that a functional differentiation driven by the positive selection occurred at the early stage of the divergence of rodent *CLEC2Da* genes. At this stage, however, what kinds of functional change have occurred is unclear.

Putative NKC Sequences of Opossum and Chicken

The demonstration that the 28 orthologous NKC gene lineages diverged from one another before the divergence of the four placental mammal orders raises the possibility that the divergence might have occurred much deeper in the evolutionary past. To explore this possibility, we extended the search for NKC genes to the opossum, a representative of marsupial mammals, which diverged from the placental mammals approximately 170 Mya, and to the domestic fowl ("chicken"), a bird representative, which diverged from mammals approximately 310 Mya (Kumar and Hedges 1998). The genomes of these two species are currently being sequenced and, though the data thus far available are incomplete, they nevertheless afford us an insight into the long-term evolution of the NKC genes.

The search of the currently chromosome-wise unassembled opossum genome yielded eight sequences homologous to the NKC genes of placental mammals. We used their translated amino acid sequences, together with the chicken sequences (see below) and selected sequences of placental mammals, to construct a phylogenetic tree by the NJ method (figure 4.4). The criteria for the selection of the placental mammal

sequences were full coverage of the previously identified 28 gene lineages (figure 4.2) and representation of one rodent (mostly mouse) and one non-rodent (human in most case) genes for each lineage. The observed good correspondence between the topologies of the trees in figures 4.2 and 4.4 indicates that the selection procedure did not bias the sample. On the tree in figure 4.4, four of the eight opossum sequences ally themselves with the CLEC4 cluster and so presumably represent CLSF group II genes; the remaining four sequences are apparently group V genes. One of the 4 opossum group II sequences, *CLEC4E*, shows orthologous relationship to the placental mammalian *CLEC4E* genes. This observation indicates that the origin of the *CLEC4E* gene lineage, and hence the divergence of *CLEC4E* from the other CLEC4 gene lineages, occurred prior to the divergence of marsupial and placental mammals. The remaining three opossum group II sequences are in an outgroup position to the entire CLEC4 cluster, together with a singleton cattle sequence whose chromosomal location is undetermined. The existence of this outgroup cluster suggests that it might represent a new ancestral *CLEC4* gene lineage or even some other group II CLSF genes. The four group V opossum CLSF genes appear to be orthologs of the *CLEC1A*, *CLEC1B*, *KLRK1*, and *CD69* genes in placental mammals. In each case, the opossum sequence is in an outgroup position to the placental mammal members of the gene lineage and the clade is supported by high bootstrap value. Overall, these observations suggest that at least five NKC genes are present in the opossum and that they originated prior to the divergence of marsupial and placental mammals.

Searching the chicken genome, we identified ten NKC-like sequences, and by searching a chicken EST database, we have found nine additional sequences. Two of these sequences were reported previously (Rogers et al. 2005). From the phylogenetic tree (figure 4.4), only one group II CLSF sequence, which comes from the EST database, has been identified in the chicken, whereas the remaining 18 chicken sequences apparently represent CLSF group V genes. These two groups CLSF genes must have therefore diverged from each other before the divergence of mammals and birds. The two previously reported chicken NKC sequences are on chromosome 16, linked to the chicken MHC (Rogers et al. 2005). The genomic sequence corresponding to the first reported sequence, "*B-lec*", has only the first two exons of the CTLD-encoding part on chromosome 16, whereas the third exon is located on an unmapped contig. Yet, the EST database contains a full-length CTLD sequence corresponding to *B-lec*. These discrepancies are probably caused by genome assembly errors, although the possibility of haplotype polymorphism has not been excluded. On the tree, another reported sequence, "*B-NK*", forms an outgroup to the KLRB/F cluster, pushing its origin to the time prior to bird-mammal split. The remaining 17 sequences are affiliated with the CLEC2 cluster of genes in placental mammals, one forming an outgroup to the pair of *CLEC2A* and *CLEC2D* lineages, and the other 16 chicken sequences forming a monophyletic clade in an outgroup position to the group of *CLEC2A*, *CLEC2B*, and *CLEC2D* genes of placental mammals.

DISCUSSION

Our data indicate that an NKC similar to the human and mouse complexes exists also in three other species of placental mammals, rat, dog, and cattle (figure 4.1). Because these five species represent four very different orders of placental mammals, it is probably safe to predict that all placental mammals have an NKC, although not always as a single region on one chromosome: some species may have it split, like the cattle, into two parts residing on different chromosomes. The incompleteness of the characterization of the opossum and chicken genomes, compared to the five placental mammals, precludes us to make any firm statements about the organization of the NKC genes in non-placental mammals and birds, at this stage. For convenience, we refer to the homologs in these species as putative NKC sequences, keeping in mind that they may not be clustered together in a manner similar to that of the five species of placental mammals.

In terms of the type of genes it contains and their evolution, the NKC of the placental mammals is a heterogeneous chromosomal region and to discuss its evolution, it is expedient to distinguish five levels of its heterogeneity. At the first level, the complex contains two classes of genes: those that encode CTLDs and hence belong to the CLSF, and those that do not. The two classes are unrelated to each other evolutionarily. The non-CLSF genes are present in the NKCs of all five examined species of placental mammals, dividing the NKC into different sections (separated by double slash symbols in figure 4.1). The position of some of these non-CLSF genes [e.g, the gene *GABARAPL1* (gamma-aminobutyric acid (GABA) receptor-associated

protein-like 1) located between *OLR1* and *KLRE1*] are conserved in the NKCs of the five species of placental mammals, indicating that the genes were presumably present in the NKC in the MRCA of the five species of placental mammals.

At the second level, the NKC genes fall into two out of the 14 groups of CLSF genes, namely groups II and V. We have demonstrated that genes belonging to these two groups are present not only in placental mammals, but also in marsupials and in birds. The two groups must have therefore separated from each other before the separation of mammals and birds. How far back in evolution the separation of the two groups occurred is uncertain. Group II genes have been described in teleost fishes, as have been genes belonging to several other CLSF groups (Soanes et al. 2004; Zelensky and Gready 2004). The presence of group V genes in teleosts is, however, controversial. Some authors have failed to identify them in genomes of those fishes that have been sequenced (Zelensky and Gready 2004), but others (Sato et al. 2003; Kikuno et al. 2004) have described whole clusters of such genes in certain teleostean orders. A gene with weak phylogenetic affinity to human group V genes has even been reported to be present in the genome of a protochordate (Khalturin et al. 2003). In all the placental mammals tested, the group II and V genes of the NKC reside in separate genomic segments, either adjacent to each other on the same chromosome (human, mouse, rat, and dog), or on different chromosomes (cattle). Orphan genes belonging to these two groups may also be scattered over other genomic regions. Of the two different arrangements of group II and V NKC genes, the linked one seem to be more ancient since it is more parsimonious than the

alternative. Presumably, the separation of the two segments occurred in the evolutionary lineage leading to the artiodactyls. The divergence of group II and V genes presumably took place by tandem duplication, followed by a series of intra-group duplications. Although some of the duplicated genes may have been transposed to other chromosomes, the bulk of them has remained clustered together, the group II duplicates in one segment and the group V genes in an adjacent segment.

The third level of NKC gene heterogeneity is reflected in the distinction of the KLR from the CLEC genes. The KLR genes are unified by their expression in the NK cells and their function. The expression of an NKC gene in NK cells alone is apparently not sufficient to make it a KLR gene, since there are members of the CLSFs that are expressed in NK cells, yet do not function as KLR genes. Some of the CLEC genes apparently belong to this category (e.g., CLEC2D genes; ref. Iizuka et al. 2003); several other CLEC genes, however, are expressed in different cell types but not in NK cells (Yamanaka et al. 1998; Flornes et al. 2004). The CLEC genes lack a positive unifying feature that would distinguish them from the KLR genes: they are defined negatively as NKC genes whose products do not function as NK cell receptors. They will probably turn out to be a heterogeneous group with different genes specialized to different functions.

The fourth level of NKC gene heterogeneity is manifested in the existence of gene families – groups of genes belonging to different, but closely related orthologous lineages. In addition to their close phylogenetic relatedness, members of the same family are general also clustered on the genetic map of the NKC. These two

characteristics identify families in both the group II and group V segments, as well as in KLR and CLEC genes. The correspondence between the phylogenetic and physical clustering of genes within a family can be explained by the evolutionary history of the NKC genes, specifically by their origin by tandem duplication and the retention of their positions following the duplication events.

The fifth level of NKC gene heterogeneity is the differentiation into the individual genes, most of which have orthologs in the different species of placental mammals. At this level, too, there is a considerable degree of evolutionary conservation manifested in the existence of the 28 orthologous lineages presumably retained throughout the evolution of the placental mammals. The divergence of the 28 NKC gene lineages apparently predates the radiation of placental mammals. How far back the divergence has occurred is not clear. The three NKC gene lineages identified in the chicken imply that the NKC might have started to expand after the divergence of birds from mammals.

The heterogeneity at all the five levels reveals a surprising and seemingly paradoxical feature of NKC evolution. Much of the heterogeneity implies genomic instability, yet the complex also displays remarkable degree of conservation. Behind much of the instability is the birth-and-death process, in which new genes are created by repeated gene duplications and some duplicate genes can stay in the genome for a long time, whereas others may become pseudogenes and are eventually lost (Nei, Gu, and Sitnikova 1997). Evidence for the process is apparent everywhere within the complex and in all the species (figures 4.1 and 4.2), but nowhere is its effect as

striking as in the four orthologous gene groups in the two rodent species: *KLRA1*, *KLRB1*, *CLEC2D* and *CLEC4A/B/C*. The *KLRA* gene group shows an expansion, that occupies almost half of the entire NKC in the two rodents. By contrast, there is only one *KLRA* gene in each of the other placental mammals studied thus far (Gagnier, Wilhelm, and Mager 2003), except horse only (Takahashi et al. 2004). Similarly, the *CLEC2D* lineage is much expanded in rodents, and as are the *KLRB1* genes (figure 4.3), which code for the receptor of the *CLEC2D*-encoded proteins. Why the two rodents are so strongly affected by the birth-and-death process is unclear. Behind the increased rate of gene birth and gene loss could be factors peculiar to all rodents or to the individual NKC genes. If the former were the case, there would have to be an across the board increase in the rate of gene duplication and loss in a variety of genes over the entire rodent genome. While there seem to be some increase in the frequency of gene duplication/deletion in some other rodent gene families (Vincek et al. 1987; Grus et al. 2005), its magnitude is insufficient to explain the values observed in the case of the NKC. Hence, the rodent NKC-specific factors might also play important roles in the expansion. An indication that this might be so is the observation of an increased rate of nonsynonymous substitutions in the lineage leading to the rapidly duplicated rodent *CLEC2D* genes (figure 4.3). The increase is indicative of positive selection, which might be driven by the need of these genes to coevolve with the expanding rodent *KLRB1* genes encoding their receptors. For many of the other NKC-encoded KLRs, the ligands are encoded in the MHC genes (Kabat et al. 2002; Natarajan et al. 2002), which are well known for their instability in rodents. In some

rodent species (e.g., the mole-rat), the MHC class I genes have expanded to an estimated 100 copies (Vincek et al. 1987), while in others (e.g., Syrian hamsters), they have been reduced to a few genes only (Darden and Streilein 1984). Another evidence supporting the existence of rodent NKC-specific factors is the observation that *KIR* genes, which belong to the IgSF superfamily and perform function analogous to the *KLRA* in rodents, are expanded into a multi-gene cluster in primates, cattle, and other non-rodent mammals, while only one or two *KIR*-like genes are present in the mouse and rat (McQueen et al. 2002; Hoelsbrekken et al. 2003; Kelley, Walter, and Trowsdale 2005). Although the genomic instability manifested by the birth-and-death process may seem contradictory to the observed conservation of gene organization and of orthologous lineages, in fact the two features might have the same basis: the coevolution of the receptors with their ligands. If so, one would expect the ligands to evolve by a similar interplay of destabilizing and stabilizing processes. Indeed, there are some tantalizing hints that this might be the case.

Table 4.1. Characteristics of genes and gene groups comprising the NKC of placental mammals

Genes / Gene groups	Other Gene Symbols	Number of genes				Signaling	References
		mouse / rat	human	dog	cattle		
KLRA1	*Ly49*	15 / 33	1(p) [a]	1	1 [b]	I / A [c]	Wong et al. (1991)
KLRB1	*NKR-P1*	6 /4	1	1	1 [b]	I / A	Yokoyama et al. (1991)
KLRC	*NKG2A,C, E,F*	3 / 3	4	0 [d]	2	I / A	Yabe et al. (1993)
KLRD1	*CD94*	1	1	1(p)	1	-	Chang et al. (1995)
KLRE1	*NKG2I*	1 / 1	0	0	1	-	Wilhelm and Mager (2003)
KLRF1	*CLEC5C*	0 / 0	1	1	1 [b]	I	Roda-Navarro et al. (2000)
KLRF2 [e]	-	0 / 0	1	0	1 [b]	-	this study
KLRG	*MAFA*	1 / 2	1	1	1	I [f]	Butcher et al. (1998)
KLRH1	-	1 / 1(p)	0	1(p)	1 [b]	I	Naper et al. (2002c)
KLRI	-	2 / 2	0	0	2	I / A	Saether et al. (2005)
KLRK1	*NKG2D*	1 / 1	1	1	1	A	Yabe et al. (1993)
CLEC1A	*CLEC1*	1 / 1	1	1	1	-	Colonna et al. (2000)
CLEC1B	*CLEC2*	1 / 1	2	1	1 [c]	I	Colonna et al. (2000)
CLEC1C [e]	*CLEC12A, KLRL,CLL1*	1 / 0	1	1	1	I	Han et al. (2004)
CLEC1D [e]	*Macrophage antigen h*	1 / 1	1(p)	1	1 [b]	I	annotation
CLEC2A		0 / 0	1	0	1 [b]	-	annotation
CLEC2B	*CLECSF2, AICL*	0 / 0	1	1	0	-	Hamann et al. (1997)
CD69	-	1 / 1	1	1	1	-	Ziegler et al. (1993)
CLEC2D [e]	*OCIL, CLR*	8 / 11	2	1	4 [b]	-	Plougastel et al. (2001)

CLEC4A/B/C	*DCIR / DCAR*	5 / 4	2	1(p)	1	I / A	Bates et al. (1999)
CLEC4A1	*DCIR4*	1 / 1	0	0	0	I	Flornes et al. (2004)
CLEC4D	*CLECSF8, MCL*	1 / 1	1	1	1	-	Balch et al. (1998)
CLEC4E	*CLECSF9, MINCLE*	1 / 1	1	1	1	A	Balch et al. (2002)
CLEC4C [e]	*CLEC6A, CLECSF10, DECTIN-2*	1 / 1(p)	1(p)	0	1(p)	A	Ariizumi et al. (2000a)
CLEC7A	*CLECSF12, DECTIN-1*	1 / 1	1	1	1	I	Ariizumi et al. (2000b)
OLR1	*LOX-1, CLEC8A*	1 / 1	1	1	1	-	Yamanaka et al. (1998)
CLEC9A	-	1 / 1	1	1	1	I	annotation
CLEC15A	-	1(p) / 1	0	1	0	-	this study

[a] (p), putative or reported pseudogene

[b] The gene has not been mapped into chromosome

[c] I, inhibitory, A, activating receptors. The inhibitory receptors have an immunoreceptor tyrosine-based inhibitory motif (ITIM) in their cytoplasmic region. The activating receptors carry a positively charged residue in their transmembrane region which recruits an adaptor molecule containing an immunoreceptor tyrosine-based activating motif (ITAM).

[d] Dog NKC contains a few small fragments which might be relics of *KLRC* genes

[e] Gene names assigned in this study

[f] Rat *KLRG2* has no transmembrane domain and cytoplasmic tail

Figure 4.1**.** Genomic structure of natural killer receptor gene complex (NKC) in human, mouse, rat, dog, and cattle. Each pointed triangle represents a gene. Note that the distance between genes is not to scale. Green triangles represent KLR genes, red triangles CLEC-type genes, blue triangles *CD69* and *OLR1* genes, and orange triangles *CLEC15A* and *CLEC16* genes, which are identified in this study. The pseudogenes are labeled with 'p' at the end of the gene names. The light red line indicates orthologous relationships between genes of different species. Red bars indicate orthologous gene groups. The name of each gene in the rodent CLEC4A/B cluster is not shown because of space limitation. The asterisks (*) indicate deviation from standardized nomenclature of human and mouse genes. See Table 1 in the supporting information for details. *N*, number of genes; *Chr*, chromosome; *Mb*, length of NKC in megabases; *Un*, unassigned genomic contigs; //, segments containing non-CLSF genes in NKC.

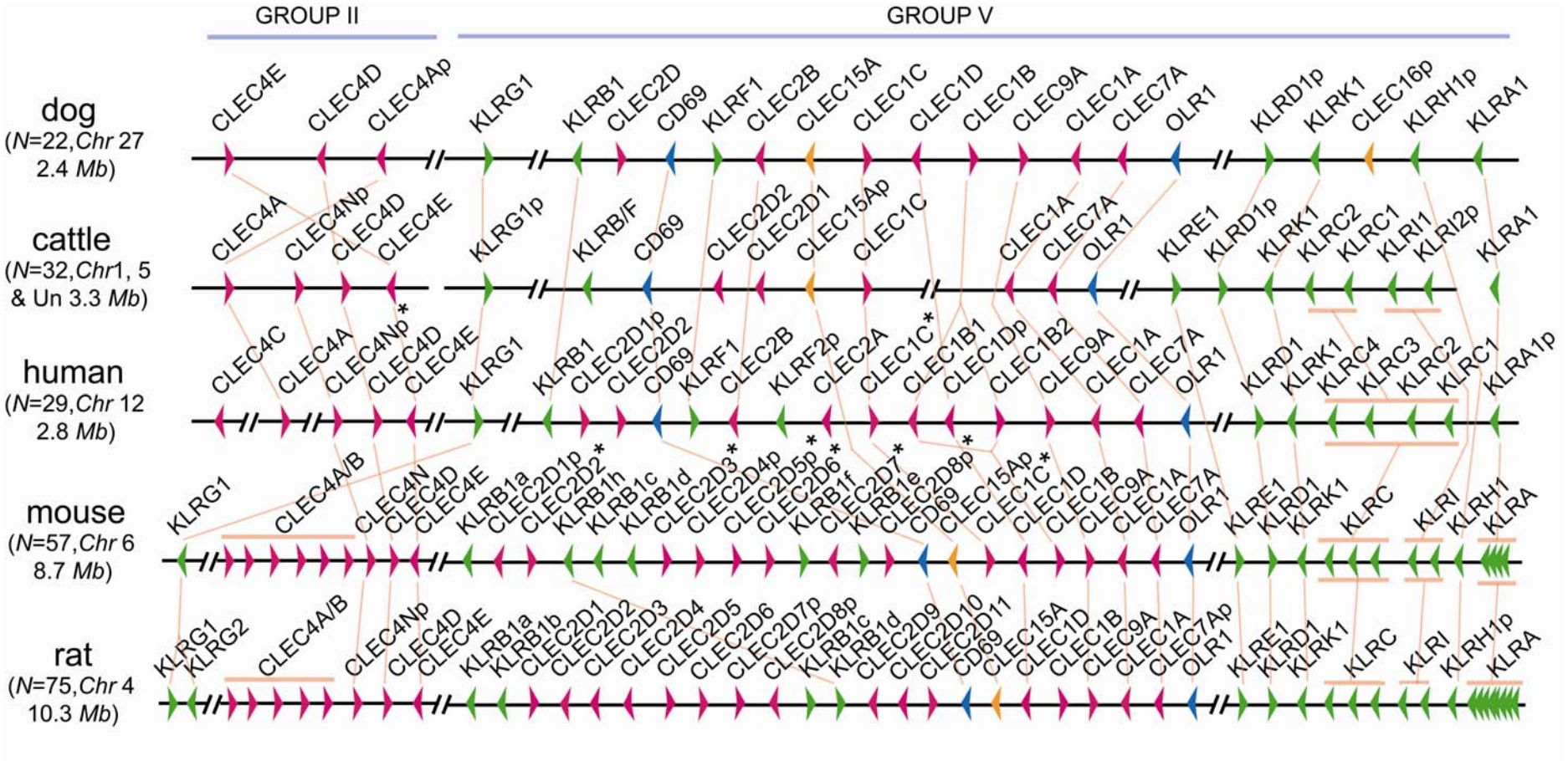
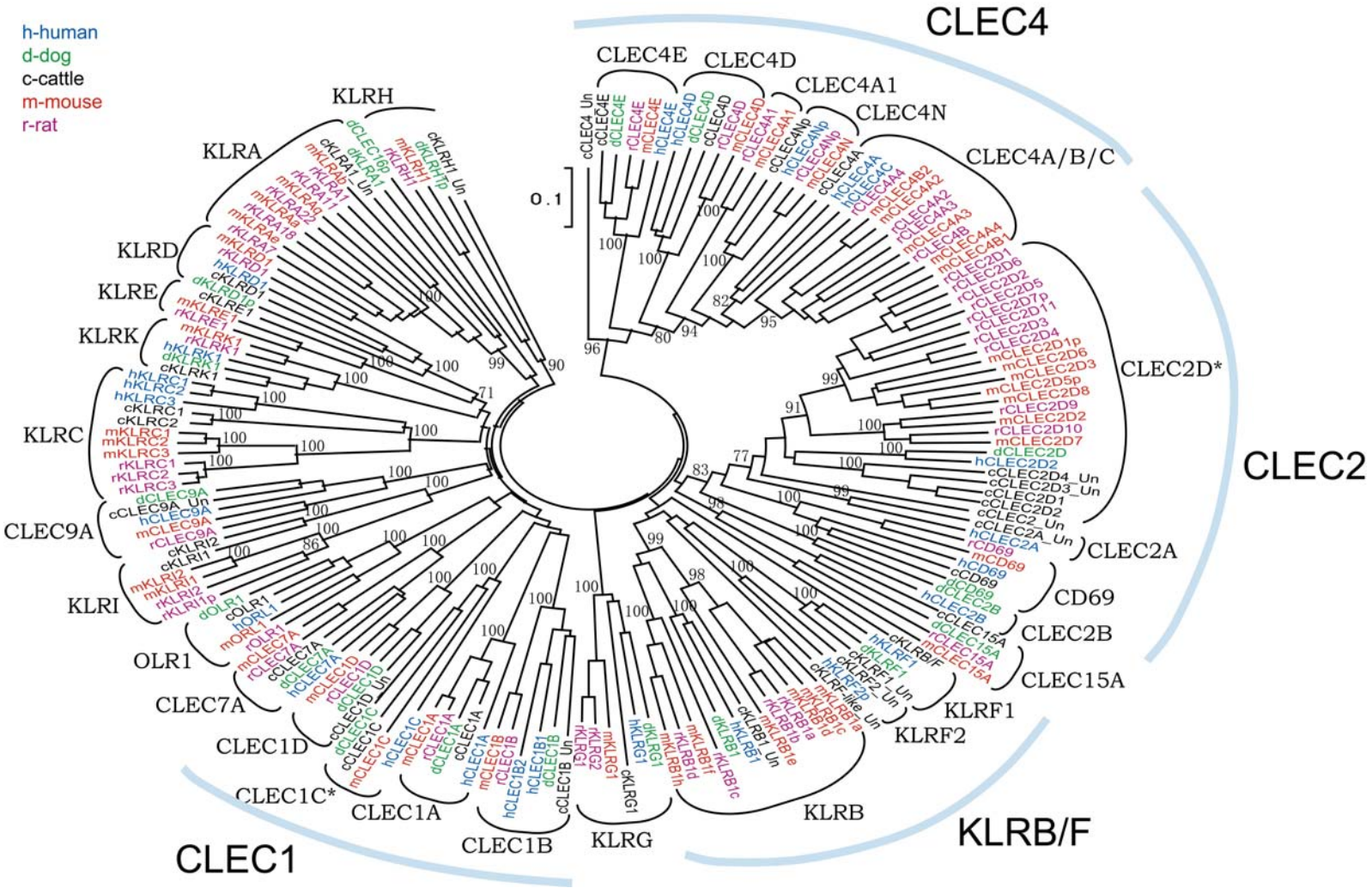

GROUP II
GROUP V
dog
(N=22,Chr 27
2.4 Mb)
CLEC4E
CLEC4D
CLEC4Ap
KLRG1
KLRB1
CLEC2D
CD69
KLRF1
CLEC2B
CLEC15A
CLEC1C
CLEC1D
CLEC1B
CLEC9A
CLEC1A
CLEC7A
OLR1
KLRD1p
KLRK1
CLEC16p
KLRH1p
KLRA1
cattle
(N=32,Chr1, 5
& Un 3.3 Mb)
CLEC4A
CLEC4Np
CLEC4D
CLEC4E
KLRG1p
KLRB/F
CD69
CLEC2D2
CLEC2D1
CLEC15Ap
CLEC1C
CLEC1A
CLEC7A
OLR1
KLRE1
KLRD1p
KLRK1
KLRC2
KLRC1
KLRI1
KLRI2p
KLRA1
human
(N=29,Chr 12
2.8 Mb)
CLEC4C
CLEC4A
CLEC4Np*
CLEC4D
CLEC4E
KLRG1
KLRB1
CLEC2D1p
CLEC2D2
CD69
KLRF1
CLEC2B
KLRF2p
CLEC2A
CLEC1C*
CLEC1B1
CLEC1Dp
CLEC1B2
CLEC9A
CLEC1A
CLEC7A
OLR1
KLRD1
KLRK1
KLRC4
KLRC3
KLRC2
KLRC1
KLRA1p
mouse
(N=57,Chr 6
8.7 Mb)
KLRG1
CLEC4A/B
CLEC4N
CLEC4D
CLEC4E
KLRB1a
CLEC2D1p
CLEC2D2*
KLRB1h
KLRB1c
KLRB1d
CLEC2D3*
CLEC2D4p
CLEC2D5p*
CLEC2D6*
KLRB1f
CLEC2D7*
KLRB1e
CLEC2D8p*
CD69
CLEC15Ap
CLEC1C*
CLEC1D
CLEC1B
CLEC9A
CLEC1A
CLEC7A
OLR1
KLRE1
KLRD1
KLRK1
KLRC
KLRI
KLRH1
KLRA
rat
(N=75,Chr 4
10.3 Mb)
KLRG1
KLRG2
CLEC4A/B
CLEC4Np
CLEC4D
CLEC4E
KLRB1a
KLRB1b
CLEC2D1
CLEC2D2
CLEC2D3
CLEC2D4
CLEC2D5
CLEC2D6
CLEC2D7p
CLEC2D8p
KLRB1c
KLRB1d
CLEC2D9
CLEC2D10
CLEC2D11
CD69
CLEC15A
CLEC1D
CLEC1B
CLEC9A
CLEC1A
CLEC7Ap
OLR1
KLRE1
KLRD1
KLRK1
KLRC
KLRI
KLRH1p
KLRA

Figure 4.2**.** Neighbor-joining tree of CTLD sequences from 172 NKC proteins identified in five placental mammals. Color code: human, blue; mouse, red; rat, purple; cattle, black; dog, green. The topology of this tree was obtained by using p-distance, the option of pairwise deletion was used. Bootstrap values are shown for orthologous groups. The asterisks (*) indicate deviation from standardized nomenclature of human and mouse genes. See Table 1 for details.


h-human
d-dog
c-cattle
m-mouse
r-rat
CLEC4
CLEC4E
CLEC4D
CLEC4A1
CLEC4N
CLEC4A/B/C
CLEC2D*
CLEC2A
CD69
CLEC2B
CLEC15A
KLRF1
KLRF2
KLRB
KLRG
CLEC1B
CLEC1A
CLEC1C*
CLEC1D
CLEC7A
OLR1
KLRI
CLEC9A
KLRC
KLRK
KLRE
KLRD
KLRA
KLRH
CLEC2
KLRB/F
CLEC1
0.1

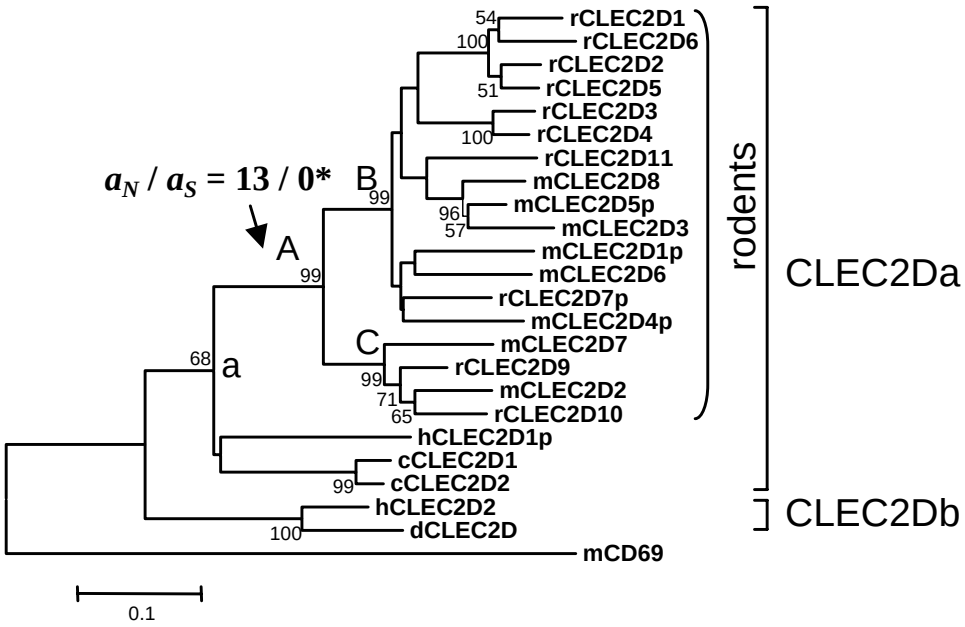
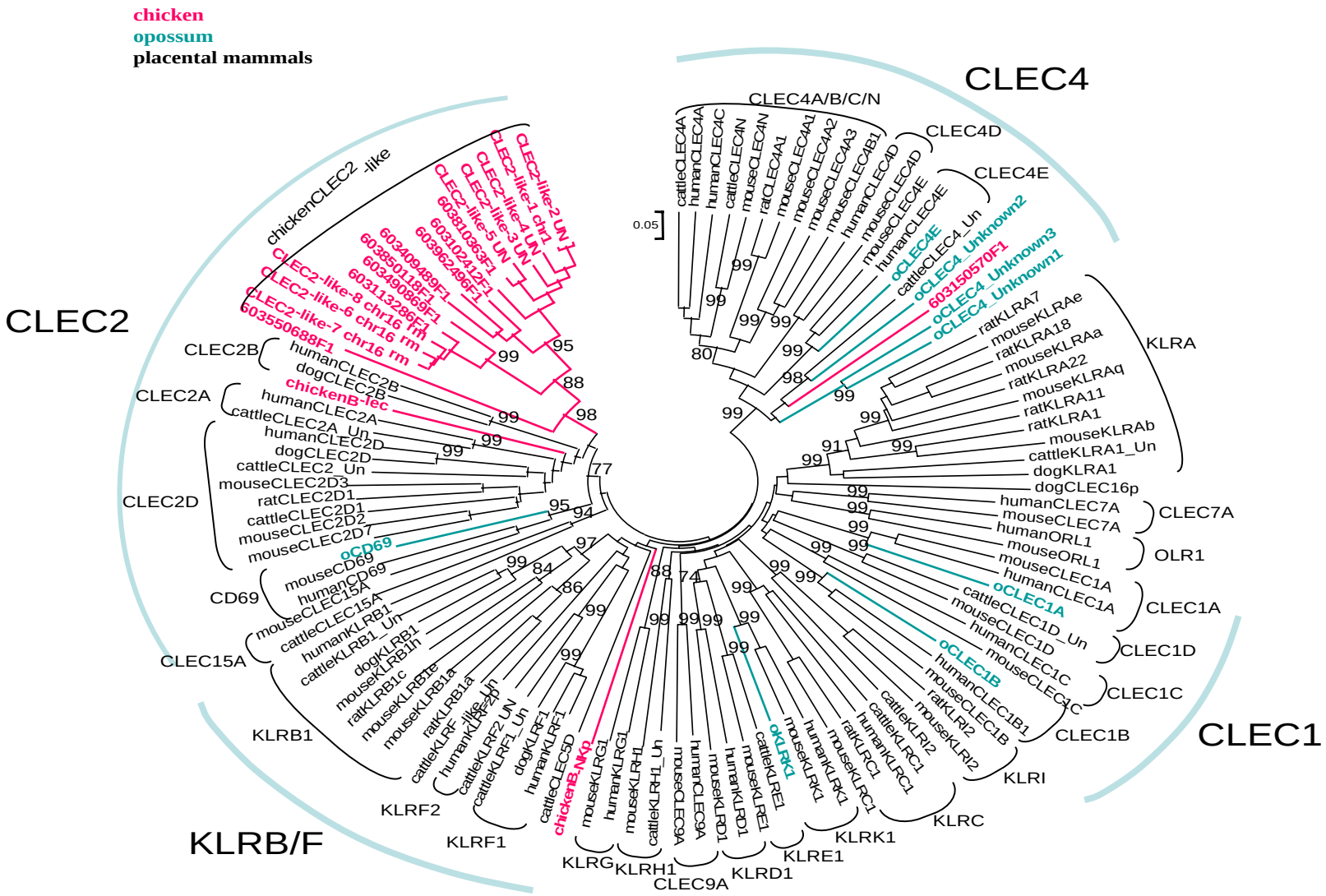


Figure 4.3. Neighbor-joining tree of *CLEC2D* genes from five placental mammals: c, cattle; d, dog; h, human; m, mouse; r, rat. Jukes-Cantor distances were used. Bootstrap values higher than 50% are shown above the branches. The numbers of nonsynonymous (a_N) and synonymous (a_S) substitutions per sequence per branch are shown for branch A. The statistical significance (at the 5% level) of the difference between observed ratio a_N / a_S and expected ratio *N/S* determined by the Fisher's exact test is indicated by the asterisk (*).

Figure 4.4**.** Neighbor-joining tree of CTLD sequences from 89 NKC proteins, including 8 opossum and 19 chicken sequences. Color code: opossum, green; chicken, red; placental mammals, black. The topology of this tree was obtained by using p-distance, the option of pairwise deletion was used. Bootstrap values are shown for orthologous groups.


chicken
opossum
placental mammals
CLEC4
CLEC4A/B/C/N
CLEC4D
CLEC4E
0.05
cattleCLEC4A
humanCLEC4A
humanCLEC4C
cattleCLEC4N
mouseCLEC4N
ratCLEC4A1
mouseCLEC4A1
mouseCLEC4A2
mouseCLEC4A3
mouseCLEC4B1
humanCLEC4D
mouseCLEC4D
humanCLEC4E
mouseCLEC4E
cattleCLEC4_Un
oCLEC4E
oCLEC4_Unknown2
oCLEC4_Unknown3
603150570F1
oCLEC4_Unknown1
CLEC2
chickenCLEC2 -like
CLEC2-like-2 UN
CLEC2-like-1 chr1
CLEC2-like-4 UN
CLEC2-like-3 UN
CLEC2-like-5 UN
603810363F1
603102412F1
603962496F1
603409489F1
603850118F1
603490869F1
603132866F1
CLEC2-like-8 chr16 rm
CLEC2-like-6 chr16 rm
CLEC2-like-7 chr16 rm
603550688F1
CLEC2B
humanCLEC2B
dogCLEC2B
CLEC2A
chickenB-lec
humanCLEC2A
cattleCLEC2A_Un
CLEC2D
humanCLEC2D
dogCLEC2D
cattleCLEC2_Un
mouseCLEC2D3
ratCLEC2D1
cattleCLEC2D1
mouseCLEC2D2
mouseCLEC2D7
oCD69
CD69
mouseCD69
humanCD69
CLEC15A
mouseCLEC15A
cattleCLEC15A
humanKLRB1
cattleKLRB1_Un
dogKLRB1
mouseKLRB1e
ratKLRB1c
mouseKLRB1a
ratKLRB1a
KLRB1
cattleKLRF_like-Un
humanKLRF2
cattleKLRF2_UN
cattleKLRF1_Un
dogKLRF1
humanKLRF1
KLRF2
cattleCLEC5D
KLRF1
chickenB-NKp
mouseKLRG1
humanKLRG1
KLRG
mouseKLRH1
cattleKLRH1_Un
KLRH1
mouseCLEC9A
humanCLEC9A
CLEC9A
mouseKLRD1
humanKLRD1
KLRD1
mouseKLRE1
cattleKLRE1
oKLRK1
KLRE1
mouseKLRK1
humanKLRK1
KLRK1
mouseKLRC1
humanKLRC1
cattleKLRC1
KLRC
mouseKLRI2
ratKLRI2
cattleKLRI2
KLRI
CLEC1
KLRB/F
KLRA
ratKLRA7
mouseKLRAe
ratKLRA18
mouseKLRAa
ratKLRA22
mouseKLRAq
ratKLRA11
ratKLRA1
mouseKLRAb
cattleKLRA1_Un
dogKLRA1
dogCLEC16p
CLEC7A
humanCLEC7A
mouseCLEC7A
OLR1
humanORL1
mouseORL1
CLEC1A
mouseCLEC1A
humanCLEC1A
oCLEC1A
CLEC1D
cattleCLEC1D_Un
mouseCLEC1D
CLEC1C
humanCLEC1C
mouseCLEC1C
CLEC1B
oCLEC1B
humanCLEC1B1
mouseCLEC1B
ratCLEC1B

BIBLIOGRAPHY

Adams, E. J., and P. Parham. 2001. Species-specific evolution of MHC class I genes in the higher primates. Immunol Rev **183**:41-64.

Alder, M. N., I. B. Rogozin, L. M. Iyer, G. V. Glazko, M. D. Cooper, and Z. Pancer. 2005. Diversity and function of adaptive immune receptors in a jawless vertebrate. Science **310**:1970-1973.

Anzai, T., T. Shiina, N. Kimura, K. Yanagiya, S. Kohara, A. Shigenari, T. Yamagata, J. K. Kulski, T. K. Naruse, Y. Fujimori, Y. Fukuzumi, M. Yamazaki, H. Tashiro, C. Iwamoto, Y. Umehara, T. Imanishi, A. Meyer, K. Ikeo, T. Gojobori, S. Bahram, and H. Inoko. 2003. Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. Proc Natl Acad Sci U S A **100**:7708-7713.

Arase, H., and L. L. Lanier. 2004. Specific recognition of virus-infected cells by paired NK receptors. Rev Med Virol **14**:83-93.

Ariizumi, K., G. L. Shen, S. Shikano, R. Ritter, 3rd, P. Zukas, D. Edelbaum, A. Morita, and A. Takashima. 2000a. Cloning of a second dendritic cell-associated C-type lectin (dectin-2) and its alternatively spliced isoforms. J Biol Chem **275**:11957-11963.

Ariizumi, K., G. L. Shen, S. Shikano, S. Xu, R. Ritter, 3rd, T. Kumamoto, D. Edelbaum, A. Morita, P. R. Bergstresser, and A. Takashima. 2000b. Identification of a novel, dendritic cell-associated molecule, dectin-1, by subtractive cDNA cloning. J Biol Chem **275**:20157-20167.

Balch, S. G., D. R. Greaves, S. Gordon, and A. J. McKnight. 2002. Organization of the mouse macrophage C-type lectin (Mcl) gene and identification of a subgroup of related lectin molecules. Eur J Immunogenet **29**:61-64.

Balch, S. G., A. J. McKnight, M. F. Seldin, and S. Gordon. 1998. Cloning of a novel C-type lectin expressed by murine macrophages. J Biol Chem **273**:18656-18664.

Barten, R., M. Torkar, A. Haude, J. Trowsdale, and M. J. Wilson. 2001. Divergent and convergent evolution of NK-cell receptors. Trends Immunol **22**:52-57.

Bates, E. E., N. Fournier, E. Garcia, J. Valladeau, I. Durand, J. J. Pin, S. M. Zurawski, S. Patel, J. S. Abrams, S. Lebecque, P. Garrone, and S. Saeland. 1999. APCs express DCIR, a novel C-type lectin surface receptor containing an immunoreceptor tyrosine-based inhibitory motif. J Immunol **163**:1973-1983.

Beutler, B. 2004. Innate immunity: an overview. Mol Immunol **40**:845-859.

Boyington, J. C., and P. D. Sun. 2002. A structural perspective on MHC class I recognition by killer cell immunoglobulin-like receptors. Mol Immunol **38**:1007-1021.

Brennan, J., G. Mahon, D. L. Mager, W. A. Jefferies, and F. Takei. 1996. Recognition of class I major histocompatibility complex molecules by Ly-49: specificities and domain interactions. J Exp Med **183**:1553-1559.

Brown, M. G., A. O. Dokun, J. W. Heusel, H. R. Smith, D. L. Beckman, E. A. Blattenberger, C. E. Dubbelde, L. R. Stone, A. A. Scalzo, and W. M. Yokoyama. 2001. Vital involvement of a natural killer cell activation receptor in resistance to viral infection. Science **292**:934-937.

Brown, M. G., S. Fulmek, K. Matsumoto, R. Cho, P. A. Lyons, E. R. Levy, A. A. Scalzo, and W. M. Yokoyama. 1997. A 2-Mb YAC contig and physical map of the natural killer gene complex on mouse chromosome 6. Genomics **42**:16-25.

Butcher, S., K. L. Arney, and G. P. Cook. 1998. MAFA-L, an ITIM-containing

receptor encoded by the human NK cell gene complex and expressed by basophils and NK cells. Eur J Immunol **28**:3755-3762.

Carlyle, J. R., A. M. Jamieson, S. Gasser, C. S. Clingan, H. Arase, and D. H. Raulet. 2004. Missing self-recognition of Ocil/Clr-b by inhibitory NKR-P1 natural killer cell receptors. Proc Natl Acad Sci U S A **101**:3527-3532.

Cerwenka, A., and L. L. Lanier. 2001. Natural killer cells, viruses and cancer. Nat Rev Immunol **1**:41-49.

Chang, C., A. Rodriguez, M. Carretero, M. Lopez-Botet, J. H. Phillips, and L. L. Lanier. 1995. Molecular characterization of human CD94: a type II membrane glycoprotein related to the C-type lectin superfamily. Eur J Immunol **25**:2433-2437.

Colonna, M., and J. Samaridis. 1995. Cloning of immunoglobulin-superfamily members associated with HLA-C and HLA-B recognition by human natural killer cells. Science **268**:405-408.

Colonna, M., J. Samaridis, and L. Angman. 2000. Molecular characterization of two novel C-type lectin-like receptors, one of which is selectively expressed in human dendritic cells. Eur J Immunol **30**:697-704.

Darden, A. G., and J. W. Streilein. 1984. Syrian hamsters express two monomorphic class I major histocompatibility complex molecules. Immunogenetics **20**:603-622.

Depatie, C., S. H. Lee, A. Stafford, P. Avner, A. Belouchi, P. Gros, and S. M. Vidal. 2000. Sequence-ready BAC contig, physical, and transcriptional map of a 2-Mb region overlapping the mouse chromosome 6 host-resistance locus Cmv1. Genomics **66**:161-174.

Dissen, E., J. C. Ryan, W. E. Seaman, and S. Fossum. 1996. An autosomal dominant

locus, Nka, mapping to the Ly-49 region of a rat natural killer (NK) gene complex, controls NK cell lysis of allogeneic lymphocytes. J Exp Med **183**:2197-2207.

Dorfman, J. R., and D. H. Raulet. 1996. Major histocompatibility complex genes determine natural killer cell tolerance. Eur J Immunol **26**:151-155.

Drickamer, K., and A. J. Fadden. 2002. Genomic analysis of C-type lectins. Biochem Soc Symp:59-72.

Flornes, L. M., Y. T. Bryceson, A. Spurkland, J. C. Lorentzen, E. Dissen, and S. Fossum. 2004. Identification of lectin-like receptors expressed by antigen presenting cells and neutrophils and their mapping to a novel gene complex. Immunogenetics **56**:506-517.

Flornes, L. M., Y. T. Bryceson, A. Spurkland, J. C. Lorentzen, E. Dissen, and S. Fossum. 2004. Identification of lectin-like receptors expressed by antigen presenting cells and neutrophils and their mapping to a novel gene complex. Immunogenetics **56**:506-517.

Gagnier, L., B. T. Wilhelm, and D. L. Mager. 2003. Ly49 genes in non-rodent mammals. Immunogenetics **55**:109-115.

Gardiner, C. M., L. A. Guethlein, H. G. Shilling, M. Pando, W. H. Carr, R. Rajalingam, C. Vilches, and P. Parham. 2001. Different NK cell surface phenotypes defined by the DX9 antibody are due to KIR3DL1 gene polymorphism. J Immunol **166**:2992-3001.

Glazko, G. V., and M. Nei. 2003. Estimation of divergence times for major lineages of primate species. Mol Biol Evol **20**:424-434.

Goldsby, R. A., T. J. Kindt, and B. A. Osborne. 2000. Immunology.

Grus, W. E., P. Shi, Y. P. Zhang, and J. Zhang. 2005. Dramatic variation of the

vomeronasal pheromone receptor gene repertoire among five orders of placental and marsupial mammals. Proc Natl Acad Sci U S A **102**:5767-5772.

Guethlein, L. A., L. R. Flodin, E. J. Adams, and P. Parham. 2002. NK cell receptors of the orangutan (Pongo pygmaeus): a pivotal species for tracking the coevolution of killer cell Ig-like receptors with MHC-C. J Immunol **169**:220-229.

Gumperz, J. E., L. D. Barber, N. M. Valiante, L. Percival, J. H. Phillips, L. L. Lanier, and P. Parham. 1997. Conserved and variable residues within the Bw4 motif of HLA-B make separable contributions to recognition by the NKB1 killer cell-inhibitory receptor. J Immunol **158**:5237-5241.

Hamann, J., K. T. Montgomery, S. Lau, R. Kucherlapati, and R. A. van Lier. 1997. AICL: a new activation-induced antigen encoded by the human NK gene complex. Immunogenetics **45**:295-300.

Han, Y., M. Zhang, N. Li, T. Chen, Y. Zhang, T. Wan, and X. Cao. 2004. KLRL1, a novel killer cell lectinlike receptor, inhibits natural killer cell cytotoxicity. Blood **104**:2858-2866.

Hao, L., and M. Nei. 2004. Genomic organization and evolutionary analysis of Ly49 genes encoding the rodent natural killer cell receptors: rapid evolution by repeated gene duplication. Immunogenetics **56**:343-354.

Hao, L., and M. Nei. 2005. Rapid expansion of killer cell immunoglobulin-like receptor genes in primates and their coevolution with MHC Class I genes. Gene **347**:149-159.

Hershberger, K. L., R. Shyam, A. Miura, and N. L. Letvin. 2001. Diversity of the killer cell Ig-like receptors of rhesus monkeys. J Immunol **166**:4380-4390.

Hoelsbrekken, S. E., O. Nylenna, P. C. Saether, I. O. Slettedal, J. C. Ryan, S. Fossum,

and E. Dissen. 2003. Cutting edge: molecular cloning of a killer cell Ig-like receptor in the mouse and rat. J Immunol **170**:2259-2263.

Hoffmann, J. A., F. C. Kafatos, C. A. Janeway, and R. A. Ezekowitz. 1999. Phylogenetic perspectives in innate immunity. Science **284**:1313-1318.

Hsu, K. C., S. Chida, D. E. Geraghty, and B. Dupont. 2002. The killer cell immunoglobulin-like receptor (KIR) genomic region: gene-order, haplotypes and allelic polymorphism. Immunol Rev **190**:40-52.

Hughes, A. L. 2002. Evolution of the human killer cell inhibitory receptor family. Mol Phylogenet Evol **25**:330-340.

Idris, A. H., H. R. Smith, L. H. Mason, J. R. Ortaldo, A. A. Scalzo, and W. M. Yokoyama. 1999. The natural killer gene complex genetic locus Chok encodes Ly-49D, a target recognition receptor that activates natural killing. Proc Natl Acad Sci U S A **96**:6330-6335.

Iizuka, K., O. V. Naidenko, B. F. Plougastel, D. H. Fremont, and W. M. Yokoyama. 2003. Genetically linked C-type lectin-related ligands for the NKRP1 family of natural killer cell receptors. Nat Immunol **4**:801-807.

Johnson, M. E., L. Viggiano, J. A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi, and E. E. Eichler. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature **413**:514-519.

Kabat, J., F. Borrego, A. Brooks, and J. E. Coligan. 2002. Role that each NKG2A immunoreceptor tyrosine-based inhibitory motif plays in mediating the human CD94/NKG2A inhibitory signal. J Immunol **169**:1948-1958.

Kane, K. P., E. T. Silver, and B. Hazes. 2001. Specificity and function of activating Ly-49 receptors. Immunol Rev **181**:104-114.

Kasahara, M. 2000. Genome paralogy: a new perspective on the organization and

origin of the major histocompatibility complex. Curr Top Microbiol Immunol **248**:53-66.

Kelley, J., L. Walter, and J. Trowsdale. 2005. Comparative genomics of natural killer cell receptor gene clusters. PLoS Genetics **1**:129-139.

Kelley, J., L. Walter, and J. Trowsdale. 2005. Comparative genomics of natural killer cell receptor gene clusters. PLoS Genet **1**:129-139.

Khakoo, S. I., R. Rajalingam, B. P. Shum, K. Weidenbach, L. Flodin, D. G. Muir, F. Canavez, S. L. Cooper, N. M. Valiante, L. L. Lanier, and P. Parham. 2000. Rapid evolution of NK cell receptor systems demonstrated by comparison of chimpanzees and humans. Immunity **12**:687-698.

Khalturin, K., M. Becker, B. Rinkevich, and T. C. Bosch. 2003. Urochordates and the origin of natural killer cells: identification of a CD94/NKR-P1-related receptor in blood cells of Botryllus. Proc Natl Acad Sci U S A **100**:622-627.

Kikuno, R., A. Sato, W. E. Mayer, S. Shintani, T. Aoki, and J. Klein. 2004. Clustering of C-type lectin natural killer receptor-like loci in the bony fish Oreochromis niloticus. Scand J Immunol **59**:133-142.

Klein, J., and F. Figueroa. 1986. Evolution of the major histocompatibility complex. Crit Rev Immunol **6**:295-386.

Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917-920.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform **5**:150-163.

Kwon, D., Y. J. Chwae, I. H. Choi, J. H. Park, S. J. Kim, and J. Kim. 2000. Diversity of the p70 killer cell inhibitory receptor (KIR3DL) family members in a single individual. Mol Cells **10**:54-60.

Li, W. H., and L. A. Sadler. 1991. Low nucleotide diversity in man. Genetics **129**:513-523.

Lynch, M., and J. S. Conery. 2003. The evolutionary demography of duplicate genes. J Struct Funct Genomics **3**:35-44.

Mager, D. L., K. L. McQueen, V. Wee, and J. D. Freeman. 2001. Evolution of natural killer cell receptors: coexistence of functional Ly49 and KIR genes in baboons. Curr Biol **11**:626-630.

Makrigiannis, A. P., A. T. Pau, P. L. Schwartzberg, D. W. McVicar, T. W. Beck, and S. K. Anderson. 2002. A BAC contig map of the Ly49 gene cluster in 129 mice reveals extensive differences in gene content relative to C57BL/6 mice. Genomics **79**:437-444.

Marsh, S. G., P. Parham, B. Dupont, D. E. Geraghty, J. Trowsdale, D. Middleton, C. Vilches, M. Carrington, C. Witt, L. A. Guethlein, H. Shilling, C. A. Garcia, K. C. Hsu, and H. Wain. 2003. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. Tissue Antigens **62**:79-86.

Martin, A. M., E. M. Freitas, C. S. Witt, and F. T. Christiansen. 2000. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. Immunogenetics **51**:268-280.

Martin, A. M., J. K. Kulski, S. Gaudieri, C. S. Witt, E. M. Freitas, J. Trowsdale, and F. T. Christiansen. 2004. Comparative genomic analysis, diversity and evolution of two KIR haplotypes A and B. Gene **335**:121-131.

Mason, L. H., S. K. Anderson, W. M. Yokoyama, H. R. Smith, R. Winkler-Pickett,

and J. R. Ortaldo. 1996. The Ly-49D receptor activates murine natural killer cells. J Exp Med **184**:2119-2128.

Mason, L. H., P. Gosselin, S. K. Anderson, W. E. Fogler, J. R. Ortaldo, and D. W. McVicar. 1997. Differential tyrosine phosphorylation of inhibitory versus activating Ly-49 receptor proteins and their recruitment of SHP-1 phosphatase. J Immunol **159**:4187-4196.

Matsuda, F., E. K. Shin, H. Nagaoka, R. Matsumura, M. Haino, Y. Fukita, S. Taka-ishi, T. Imai, J. H. Riley, R. Anand, and et al. 1993. Structure and physical map of 64 variable segments in the 3'0.8-megabase region of the human immunoglobulin heavy-chain locus. Nat Genet **3**:88-94.

McQueen, K. L., S. Lohwasser, F. Takei, and D. L. Mager. 1999. Expression analysis of new Ly49 genes: most transcripts of Ly49j lack the transmembrane domain. Immunogenetics **49**:685-691.

McQueen, K. L., B. T. Wilhelm, K. D. Harden, and D. L. Mager. 2002. Evolution of NK receptors: a single Ly49 and multiple KIR genes in the cow. Eur J Immunol **32**:810-817.

Nam, J., J. Kim, S. Lee, G. An, H. Ma, and M. Nei. 2004. Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. Proc Natl Acad Sci U S A **101**:1910-1915.

Naper, C., S. Hayashi, E. Joly, G. W. Butcher, B. Rolstad, J. T. Vaage, and J. C. Ryan. 2002a. Ly49i2 is an inhibitory rat natural killer cell receptor for an MHC class Ia molecule (RT1-A1c). Eur J Immunol **32**:2031-2036.

Naper, C., S. Hayashi, L. Kveberg, E. C. Niemi, L. L. Lanier, J. T. Vaage, and J. C. Ryan. 2002b. Ly-49s3 is a promiscuous activating rat NK cell receptor for nonclassical MHC class I-encoded target ligands. J Immunol **169**:22-30.

Naper, C., S. Hayashi, G. Lovik, L. Kveberg, E. C. Niemi, B. Rolstad, E. Dissen, J. C. Ryan, and J. T. Vaage. 2002c. Characterization of a novel killer cell lectin-like receptor (KLRH1) expressed by alloreactive rat NK cells. J Immunol **168**:5147-5154.

Natarajan, K., N. Dimasi, J. Wang, D. H. Margulies, and R. A. Mariuzza. 2002. MHC class I recognition by Ly49 natural killer cell receptors. Mol Immunol **38**:1023-1027.

Nei, M. 1969. Gene duplication and nucleotide substitution in evolution. Nature **221**:40-42.

Nei, M. 1987. Molecular evolutionary genetics.

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

Nei, M., and S. Kumar. 2000. Molecular evolution and phylogenetics.

Nei, M., P. Xu, and G. Glazko. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc Natl Acad Sci U S A **98**:2497-2502.

Niimura, Y., and M. Nei. 2003. Evolution of olfactory receptor genes in the human genome. Proc Natl Acad Sci U S A **100**:12235-12240.

Nylenna, O., C. Naper, J. T. Vaage, P. Y. Woon, D. Gauguier, E. Dissen, J. C. Ryan, and S. Fossum. 2005. The genes and gene organization of the Ly49 region of the rat natural killer cell gene complex. Eur J Immunol **35**:261-272.

O'Callaghan, C. A. 2000. Natural killer cell surveillance of intracellular antigen processing pathways mediated by recognition of HLA-E and Qa-1b by CD94/NKG2 receptors. Microbes Infect **2**:371-380.

Ota, T., T. Sitnikova, and M. Nei. 2000. Evolution of vertebrate immunoglobulin variable gene segments. Curr Top Microbiol Immunol **248**:221-245.

Parham, P. 1997. Events in the adaptation of natural killer cell receptors to MHC class I polymorphisms. Res Immunol **148**:190-194.

Piontkivska, H., and M. Nei. 2003. Birth-and-death evolution in primate MHC class I genes: divergence time estimates. Mol Biol Evol **20**:601-609.

Plougastel, B., C. Dubbelde, and W. M. Yokoyama. 2001. Cloning of Clr, a new family of lectin-like genes localized between mouse Nkrp1a and Cd69. Immunogenetics **53**:209-214.

Rajalingam, R., P. Parham, and L. Abi-Rached. 2004. Domain shuffling has been the main mechanism forming new hominoid killer cell Ig-like receptors. J Immunol **172**:356-369.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**:493-521.

Renedo, M., I. Arce, K. Montgomery, P. Roda-Navarro, E. Lee, R. Kucherlapati, and E. Fernandez-Ruiz. 2000. A sequence-ready physical map of the region containing the human natural killer gene complex on chromosome 12p12.3-p13.2. Genomics **65**:129-136.

Roda-Navarro, P., I. Arce, M. Renedo, K. Montgomery, R. Kucherlapati, and E. Fernandez-Ruiz. 2000. Human KLRF1, a novel member of the killer cell lectin-like receptor gene family: molecular characterization, genomic structure, physical mapping to the NK gene complex and expression analysis. Eur J Immunol **30**:568-576.

Rogers, S. L., T. W. Gobel, B. C. Viertlboeck, S. Milne, S. Beck, and J. Kaufman. 2005. Characterization of the chicken C-type lectin-like receptors B-NK and

B-lec suggests that the NK complex and the MHC share a common ancestral region. J Immunol **174**:3475-3483.

Saether, P. C., I. H. Westgaard, L. M. Flornes, S. E. Hoelsbrekken, J. C. Ryan, S. Fossum, and E. Dissen. 2005. Molecular cloning of KLRI1 and KLRI2, a novel pair of lectin-like natural killer-cell receptors with opposing signalling motifs. Immunogenetics **56**:833-839.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetics trees. Molecular Biology and Evolution **4**:406-425.

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

Sato, A., W. E. Mayer, P. Overath, and J. Klein. 2003. Genes encoding putative natural killer cell C-type lectin receptors in teleostean fishes. Proc Natl Acad Sci U S A **100**:7779-7784.

Satta, Y., W. E. Mayer, and J. Klein. 1996. Evolutionary relationship of HLA-DRB genes inferred from intron sequences. J Mol Evol **42**:648-657.

Schable, K. F., and H. G. Zachau. 1993. The variable genes of the human immunoglobulin kappa locus. Biol Chem Hoppe Seyler **374**:1001-1022.

Schwartz, S., Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker--a web server for aligning two genomic DNA sequences. Genome Res **10**:577-586.

Seaman, W. E. 2000. Natural killer cells and natural killer T cells. Arthritis Rheum **43**:1204-1217.

Soanes, K. H., K. Figuereido, R. C. Richards, N. R. Mattatall, and K. V. Ewart. 2004. Sequence and expression of C-type lectin receptors in Atlantic salmon (Salmo

salar). Immunogenetics **56**:572-584.

Storset, A. K., I. O. Slettedal, J. L. Williams, A. Law, and E. Dissen. 2003. Natural killer cell receptors in cattle: a bovine killer cell immunoglobulin-like receptor multigene family contains members with divergent signaling motifs. Eur J Immunol **33**:980-990.

Su, C., and M. Nei. 2001. Evolutionary dynamics of the T-cell receptor VB gene family as inferred from the human and mouse genomic sequences. Mol Biol Evol **18**:503-513.

Swofford, D. L. 1998. PAUP. Phylogenetic analysis using parsimony.

Takahashi, K., A. P. Rooney, and M. Nei. 2000. Origins and divergence times of mammalian class II MHC gene clusters. J Hered **91**:198-204.

Takahashi, T., M. Yawata, T. Raudsepp, T. L. Lear, B. P. Chowdhary, D. F. Antczak, and M. Kasahara. 2004. Natural killer cell receptors in the horse: evidence for the existence of multiple transcribed LY49 genes. Eur J Immunol **34**:773-784.

Takei, F., J. Brennan, and D. L. Mager. 1997. The Ly-49 family: genes, proteins and recognition of class I MHC. Immunol Rev **155**:67-77.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol **12**:823-833.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res **25**:4876-4882.

Toneva, M., V. Lepage, G. Lafay, N. Dulphy, M. Busson, S. Lester, A. Vu-Trien, A. Michaylova, E. Naumova, J. McCluskey, and D. Charron. 2001. Genomic diversity of natural killer cell receptor genes in three populations. Tissue

Antigens **57**:358-362.

Trowsdale, J., R. Barten, A. Haude, C. A. Stewart, S. Beck, and M. J. Wilson. 2001. The genomic context of natural killer receptor extended gene families. Immunological Reviews **181**:20-38.

Trowsdale, J., R. Barten, A. Haude, C. A. Stewart, S. Beck, and M. J. Wilson. 2001. The genomic context of natural killer receptor extended gene families. Immunol Rev **181**:20-38.

Vilches, C., M. J. Pando, and P. Parham. 2000. Genes encoding human killer-cell Ig-like receptors with D1 and D2 extracellular domains all contain untranslated pseudoexons encoding a third Ig-like domain. Immunogenetics **51**:639-646.

Vincek, V., D. Nizetic, M. Golubic, F. Figueroa, E. Nevo, and J. Klein. 1987. Evolutionary expansion of Mhc class I loci in the mole-rat, Spalax ehrenbergi. Mol Biol Evol **4**:483-491.

Welch, A. Y., M. Kasahara, and L. M. Spain. 2003. Identification of the mouse killer immunoglobulin-like receptor-like (Kirl) gene family mapping to chromosome X. Immunogenetics **54**:782-790.

Wilhelm, B. T., L. Gagnier, and D. L. Mager. 2002. Sequence analysis of the ly49 cluster in C57BL/6 mice: a rapidly evolving multigene family in the immune system. Genomics **80**:646-661.

Wilhelm, B. T., and D. L. Mager. 2003. Identification of a new murine lectin-like gene in close proximity to CD94. Immunogenetics **55**:53-56.

Wilson, M. J., M. Torkar, A. Haude, S. Milne, T. Jones, D. Sheer, S. Beck, and J. Trowsdale. 2000. Plasticity in the organization and sequences of human KIR/ILT gene families. Proc Natl Acad Sci U S A **97**:4778-4783.

Winter, C. C., and E. O. Long. 1997. A single amino acid in the p58 killer cell inhibitory receptor controls the ability of natural killer cells to discriminate between the two groups of HLA-C allotypes. J Immunol **158**:4026-4028.

Wong, S., J. D. Freeman, C. Kelleher, D. Mager, and F. Takei. 1991. Ly-49 multigene family. New members of a superfamily of type II membrane proteins with lectin-like domains. J Immunol **147**:1417-1423.

Yabe, T., C. McSherry, F. H. Bach, P. Fisch, R. P. Schall, P. M. Sondel, and J. P. Houchins. 1993. A multigene family on human chromosome 12 encodes natural killer-cell lectins. Immunogenetics **37**:455-460.

Yamanaka, S., X. Y. Zhang, K. Miura, S. Kim, and H. Iwao. 1998. The human gene encoding the lectin-type oxidized LDL receptor (OLR1) is a novel member of the natural killer gene complex with a unique expression profile. Genomics **54**:191-199.

Yokoyama, W. M., and B. F. M. Plougastel. 2003. Immune functions encoded by the natural killer gene complex. Nature Reviews Immunology **3**:304-316.

Yokoyama, W. M., J. C. Ryan, J. J. Hunter, H. R. Smith, M. Stark, and W. E. Seaman. 1991. cDNA cloning of mouse NKR-P1 and genetic linkage with LY-49. Identification of a natural killer cell gene complex on mouse chromosome 6. J Immunol **147**:3229-3236.

Zelensky, A. N., and J. E. Gready. 2004. C-type lectin-like domains in Fugu rubripes. BMC Genomics **5**:51.

Zhang, J., and M. Nei. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol **44 Suppl 1**:S139-146.

Zhang, J., H. F. Rosenberg, and M. Nei. 1998a. Positive Darwinian selection after

gene duplication in primate ribonuclease genes. Proc Natl Acad Sci U S A **95**:3708-3713.

Zhang, Z., A. A. Schaffer, W. Miller, T. L. Madden, D. J. Lipman, E. V. Koonin, and S. F. Altschul. 1998b. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res **26**:3986-3990.

Ziegler, S. F., F. Ramsdell, K. A. Hjerrild, R. J. Armitage, K. H. Grabstein, K. B. Hennen, T. Farrah, W. C. Fanslow, E. M. Shevach, and M. R. Alderson. 1993. Molecular characterization of the early activation antigen CD69: a type II membrane glycoprotein related to a family of natural killer cell activation antigens. Eur J Immunol **23**:1643-1648.

VITA

Li Hao was born in Le Shan, China, on January 13, 1975. She received a degree of Bachelor of Sciences in Genetics and Genetic Engineering from Fudan University in August 1996 and a degree of Master of Sciences in Genetics and Cell Biology from Beijing University in July 1999. In August 1999, she entered the Ph.D. program of Cell and Development Biology in the Huck Institute of Life Sciences at the Pennsylvania State University, and transferred to the program in Biology in 2001. Li Hao is a student member of the Society for Molecular Biology and Evolution.

Recent Publications

- **Hao, L.**, Klein, J., and M. Nei (2006) Heterogeneous but conserved natural killer receptor gene complexes in four major orders of mammals. *Proc. Natl. Acad. Sci. U. S. A.* 103:3192-3197

- Wei, X., **Hao, L.**, Ni, S., Liu, Q., Xu, J., and P. H. Correll (2005) Altered exon usage in the juxtamembrane domain of mouse and human RON regulates receptor activity and signaling specificity. *J Biol. Chem.* 280:40241-40251

- **Hao, L.** and M. Nei (2005) Rapid expansion of killer cell immunoglobulin-like receptor genes in primates and their coevolution with MHC Class I genes. *Gene*. 347:149-159.

- **Hao, L.** and M. Nei (2004) Genomic organization and evolutionary analysis of *Ly49* genes encoding the rodent natural killer cell receptors: rapid evolution by repeated gene duplication. *Immunogenetics* 56:343-354.

- Huang, J., **Hao, L.,** Liu, S., Li, L., Zhang, W. X., and Z. H. Dai (2002) Phylogenetic position of Chinese endemic *Drosophila curviceps* species subgroup in the *Drosophila immigrans* group. *Acta Genetica Sinica.* 29(5):417-423. In Chinese

- **Hao, L.**, Gu, Z. L., and Z. H. Dai (2000) The frequency distribution and establishment of fruit fly strain of segregation distorter in *Drosophila melanogaster* in China. *Acta Genetica Sinica.* 27(4):298-303. In Chinese

- **Hao, L.**, Wu, C-I, and Z. H. Dai (1999) The Segregation Distorter in *Drosophila melanogaster*. *Hereditas* 21(4):57-62. In Chinese