High Resolution Physical Map of the Distal Part of the Natural Killer Cell Gene Complex

Øyvind Nylenø

Department of Anatomy
Institute of Basic Medical Sciences, Medical Research Curriculum

October 2004
1 ACKNOWLEDGEMENT

This thesis is a result of my work at the Immunological laboratory, Dept. of Anatomy, Institute of Basic Medical Science, University of Oslo. I started in Sigbjørn Fossum's laboratory autumn 1998 and received student fellowship in 1999 from the Norwegian Research Council. I have later received fellowship from The Norwegian Cancer Society. Further grants have been given from the Odd Fellow medisinsk-Vitenskaplig forskningsfond.

I have been lucky to work in an inspiring, encouraging, well informed and friendly milieu made by present and past colleagues. This has, together with an interesting work, been catalytic on the hours in laboratory and made them fly. There are some people deserving to be thanked in specific:

My supervisor Sigbjørn Fossum for excellent guidance and support. He has given me freedom in planning and choosing the laboratory work. Always given me support and been helpful in an optimistic and enthusiastic manner even when the experiments fail and have to be redone. His knowledge of immunology and basal medicine is invaluable and has been indispensable.

My co-supervisor Erik Dissen for enormous valuable knowledge of conducting laboratory work in an accurate and precise manner. He has always been given support when I have asked for assistance and at the same time been quick with a joke. Erik and Sigbjørn make it a stimulating and instructive environment.

Marianne Kollerud and Wendi Jensen for good technical assistance and help in the laboratory.

Sigurd Erik Hoelsbrekken, Per Christian Sæter, Ingunn H. Westgaard and Line Flornes for help and many hours of good company in the lab.

My family for invaluable patience and support throughout the work.
2 INDEX

1 ACKNOWLEDGMENT ................................................................. p 2

2 INDEX ................................................................. p 4

3 ABBREVIATIONS ......................................................... p 5

4 INTRODUCTION ......................................................................... p 6

  4.1 Natural Killer Cells .............................................................. p 7
   1. Definition, morphology and distribution
   2. Function

  4.2 NK cells and their targets ...................................................... p 9
   1. Allogenic cells
   2. Cancer cells
   3. Infected cells
   4. NK cells and immunomodulation
   5. The cytotoxic mechanisms

  4.3 NK cell receptors ............................................................... p 14
   1. Killer cell lectin-like receptors
      a. NKR-P1
      b. NKG2 / CD94
      c. The Ly49 subfamily
   2. Killer cell Ig-like receptors .......................................................... p 17
      a. Killer cell Ig-like receptors (KIR)
      b. Leukocyte Ig-like receptors (LILR) and LAIR
      c. NKp46

5 METHODS & TECHNIQUES ..................................................... p 19

  5.1 Polymerase Chain Reaction
  5.2 Hybridization
  5.3 Gel electrophoresis
  5.4 TA-cloning
  5.5 Fingerprinting
  5.6 Chromosomal mapping and bioinformatics

6 RESULTS ..................................................................................... p 25

  6.1 Identification of Ly49 genes from the genomic sequences
  6.2 High resolution physical map
  6.3 Cluster size and gene density
  6.4 Sequence analysis
  6.5 Receptor groups and blocks
  6.6 Gene homogenization

7 DISCUSSION AND CONCLUSION ............................................. p 36

8 REFERANCE LIST ................................................................. p 38

3 Abbreviations

BAC – Bacterial Artificial Chromosome
BLAST – Basal Local Alignment Sequence Tool
BN rat – Brown Norwegian, rattus norvegicus
CD – Cluster of Differentiation / Cluster Designation
CLSF – C-type Lectin like Super Family
CMV – Cytomegalo Virus
CRD – Carbohydrate Recognition Domain
CTLD – C-Type Lectin like Domain
EBV – Ebstein Barr Virus
GE – Gele electrophoresis
GM – CSF Granulocyte and Macrophage Colony Stimulating Factor
HCV – Hepatitis C Virus
HSV – Herpses Simplex Virus
IFN-γ – Interferon gamma
Ig – Immunoglobulin
IL – Inter Leukin
ILT – Ig-like transcript
ITAM – immunoreceptor tyrosine-based activating motif
ITIM – immunoreceptor tyrosine-based inhibitory motif
KIR – Killer cell Ig (immunoglobuline) –like receptors
KLR – Killer cell Lectin like Receptors
LAIR – leukocyte-associated Ig-like receptor
LAK – Lymphokine Activated Killer cells
LGL – Large Grranulated Lymphocyte
LILR – Leucocyte Ig-like receptors
LRC – Leukocyte Regulatory gene Complex
MHC – Major Histocompatibility Complex
MICA/B – MHC class I chain related A / B
MIP 1 α/β – Macrophage Inflammatory Protein 1α and 1β
MIR – Monocyte / macrophage immunoglobuline-like receptor
NCBI – National Center for Biothechnology Information
NK 1 – NK cell secreting IL10, -12 and IFN-γ
NK 2 – NK cell secreting IL4, -5 and -13
NK cell – Natural Killer cell
Nka – Natural killer alloreactivity
NKC – Natural Killer cell gene Complex
NKLLR – Natural Killer cell Lektin-Like receptors
PAC – P1 derived Artificial Chromosome
PCR – Polymerase Chain Reaction
PLC – Phospholipase C
RANTES - regulated on activation normal T-cell-expressed and secreted
RT – Reverse Transcriptase
SH2 – Src Homology 2 domain
TNF-α – Tumor Necrosis Factor alpha
ULBP – UL16 binding protein
Almost six years ago I started to work at the Immunobiological Laboratory (IBL) as a Norwegian Research Council student fellow followed by a fellowship in the Norwegian Cancer Society. My project was to make a high resolution physical map of the rat Natural Killer cell gene complex (NKC). At that time the NKC were thought to span ~800 kb and contain in the order of 20 genes. The project was more demanding than anticipated and we now know that the rat NKC is about four times bigger, spanning ~ 3.3 Mb and containing 67 genes, all but one encoding related proteins, the exception is a GABA receptor. Here I describe the gene content of the distal third of this complex, the Ly49 region.

As part of my project I PCR-scanned a PAC library (in collaboration with The Wellcome Trust Centre for Human Genetics, University of Oxford) and identified PAC (P1 derived artificial chromosome) clones containing NKC genes, and arranged these into a tailing pattern to establish the physical chromosomal orders of the genes. The same pattern is seen in figure 7 using BAC clones instead of PAC clones. Specific PCR primers for the genes were already available from earlier work in the laboratory. The next step was to establish a high resolution physical map with regards on the nucleotide sequence and we began sequencing three of the PAC clones. At the time we got the sequences, the rat genome project had started and went much faster than foreseen. We therefore stopped sequencing more PAC clones and based our mapping on the BAC clone sequences and later the whole rat genome sequence released from the rat genome project (based on the BN rat strain). It should be emphasized that the rat genome sequence is based on work done by robots, automatic sequencers and computers, with little human intervention, and it is up to the scientific community to characterize the sequences deposited in the databanks. Identifying genes from these sequences is often referred to as “in silico cloning”. Before they can be published, however, in silico cloned genes have to be verified by in vitro cloning. We have so fare done both and the work here reported is therefore a combination of mapping of Ly49 genes.
already cDNA cloned in our laboratory, and in vitro cloning of new genes in silico predicted from the genome sequence.

The Ly49 receptor family makes up a cluster of genes in the distal part of the NKC and encode type 2 transmembrane receptors expressed by natural killer (NK) cells and a subset of T cells. The NKC is located on the rat chromosome 4q42, the mouse chromosome 6F3 and on the human chromosome 12p13.31. In the rat the Nka, autosomal dominant locus controlling NK-mediated alloreactivity, is mapped to the distal part of the Ly49 region. The receptors belong to the C-type lectin superfamily (CLSF), are involved in NK cell recognition of target cells and influence the NK cell activity. The Ly49 receptors come in two functional forms, activating and inhibitory, where both forms are expressed on the NK cell. Whether the NK cells shall recognize another cell as a target to be killed, or as not a target to be spared, depends on integration of the signals transmitted by the opposing receptors. In the mouse and the rat there are many different Ly49 receptors among the two types. Here I report the cloning of new Ly49 receptor genes in the rat, describe how they are organized in the genome and discuss the evolution of the rat Ly49 receptor repertoire. First I shall give some background information on NK cells and the C-type lectin-like receptors.

4.1 Natural Killer cells

4.1.1 Definition, morphology and distribution

In 1975 Kiessling and co-workers discovered a population of murine splenic lymphocytes that could kill malignant cells in vitro without prior sensitization. The lytic activity was shown to be non-MHC restricted and to be performed by a cell type unknown at the time. The cells were therefore called Natural Killer (NK) cells. In the microscope they looked like lymphocytes, but with scattered large granules, and were accordingly also referred to as “Large Granulated Lymphocytes” (LGL) as can be seen in the picture below, figure 1.
Compared with resting T and B cells, LGL are larger and with a higher cytoplasm/nucleus ratio. The nucleus is kidney shaped and the cytoplasm contains azurophilic granules. NK cells share many of the surface markers with T and B cells, e.g. CD2 on NK and T cells. In the human NK cells are defined by the expression of NKR-P1 and the lack of CD3.

NK cells belong to the lymphocyte lineage. In contrast to T cells, NK cells are short-lived with an approximate life span of 1-2 weeks. The NK cells are divided into two subgroups, NK1 and NK2, depending on cytokines secreted, see 2.2.4 page 11. NK cells are found in the blood where they constitute approximately 5-20% of the lymphocytes, in the liver (pit cells), the red pulp of the spleen (where they constitute approximately 5% of the lymphocytes), the alveolar capillaries in the lung and in the uterus. In contrast to T cells NK cells do not recirculate between blood and lymphoid organs.

4.1.2 Functions

Together with granulocytes and macrophages, NK cells constitute the effector cells in the innate immune system. Through cytokine release, like IFN-γ and TNF-α, they have influence on the adaptive immune system. NK cells can be activated and migrate to sites of infection, where they proliferate, release cytokines and kill cells harbouring pathogens. This process is initiated by activating receptors leading to expression of receptors like chemokine- and cytokine receptors which induce chemotaxis and activation of the NK cell. The NK cells recognize their targets by detecting abnormal protein expression patterns, like altered levels of MHC–I or related molecules, on the target cell surface. They are able to detect and eliminate some tumor cell types,
some types of infected cells and allogenic cells. Killing is mediated by two different mechanisms: 1) secretion of perforin and granzymes from intracellular granules and 2) expression of Fas ligand. The end point is cell membrane destruction with cellular lysis or induction of apoptosis respectively.

### 4.2 NK cells and their targets

#### 4.2.1 Allogenic cells

Cells that do not express the correct amount or type of MHC class I molecules are susceptible to NK cell killing. Host NK cells thus recognize and eliminate transplanted MHC incompatible allogeneic cells. The evolution of the alloreactivity recognition system may seem puzzling considering that natural transfer of allogeneic cells would be expected to occur only rarely. However, in mammals fetal semiallogeneic cells are in close contact with the maternal blood circulation (spiral arteries in the decidua) and may easily invade the mother if not kept in check. NK cells are thought to play a role in spontaneous abortion. Large colonies of NK cells are localized in the decidua and probably take part in the process of fetal rejection. Pregnancy is therefore one possible reason why this surveillance system has developed and sustains.

#### 4.2.2 Cancer cells

Because NK cells were first detected through their ability to kill tumor cells, they were postulated to play a role in the restriction of primary tumor formation and metastases development. At the present antigens presented on tumor cells are neither fully characterized nor understood. Tumor cell MHC class I and class I related molecules seemed to play a role in recognition and induction of tumor cell killing. A tumor cell presenting an altered pattern of self – MHC classes may shift the balance from not kill to kill due to integration of activating and inhibiting NK cell receptors. MICA and MICB (MHC class I chain related A and B) are molecules that resemble the MHC class I and are expressed by some epithelial tumors. In the human the NK cell receptor NKG2D recognize MICA and MICB, and binding of NKG2D to these ligands results in enhanced NK cell
activity and killing of the tumor cells. Additional receptors in the human that recognize tumor ligands and lead to enhanced killing are the NKp30, NKp44, NKp46 and CD16.

4.2.3 Infected cells

NK cells recognize and kill cells infected by cytomegalovirus (CMV), Epstein-Barr virus (EBV) and herpes simplex virus (HSV). In addition NK cells possess the property of chemotaxis, which result in cellular movement towards the site of infection. For detection of intracellular parasites mammals have developed a sophisticated report system whereby MHC class I molecules sample peptides resulting from partial degradation of proteins synthesized intracellulary and bring these to the cell surface for inspection by CD8+ T cells. If the T cells recognize foreign peptides on the surface of a cell, signifying intracellular parasites, the cell is killed. Some viruses are able to interfere with this report system. By downregulating MHC class I expression they evade detection by T cells and killing. In this context, the NK cells constitute a back-up system which detects and eliminates cells with low or no MHC class I expression.

To evade killing by NK cells some viruses, like the cytomegalovirus, hinder down regulation of particular inhibitory MHC I ligands and some even encode MHC class I protein analogues which are expressed and act as decoys on the surface of the infected cells. Another example on the race-of-arms between parasite and host is the infection with hepatitis C virus 1 (HCV-1), which leads to expression of HCV-E1, a ligand for the CD81 NK cell receptor. HCV-E1 bound to the CD81 receptor leads to blocked NK cell activation, reduced cytokine production, reduced cytotoxic granule release and inhibition of NK cells proliferation.

Detection of reduced levels of MHC class I on target cells depend on inhibitory NK cell receptors for these ligands. A significant lowered inhibitory signal via inhibitory receptors will lead to loss of inhibition and hence activation of the cell. However, the NK cells also have at its disposal activating receptors such as NKG2D, Ly49s3 and NKp46, which upon ligation induce cytotoxicity, proliferation and cytokine production. A human cell infected by CMV upregulates
membrane molecules like MICA, MICB, ULBP1/2/3 that are ligands for the activating NK cell receptor NKG2D leading to increased cytotoxicity and killing. Some viruses encode proteins that can interfere with the NK cell recognition by the activating receptors. An example is UL16, an early protein transcribed from the hCMV genome, which binds to and blocks the interaction between ULBP1/2 or MICB and NKG2D and thereby hinders killing of the cell.

4.2.4 NK cells and immunomodulation

Via activating receptors and cytokines the NK cell can recognize structures induced on infected or stressed cells leading to target cell lysis, NK cell proliferation and production of cytokines and chemokines such as INF-γ, TNFα, GM-CSF, MIP-1α/1β and RANTES. INF-γ produced by NK cells leads to a variety of effects, such as inhibition of virus replication, induction of nitric oxide synthase (iNOS) by macrophages and induction of MHC I expression in macrophages, thereby augmenting presentation of virus peptide to the cytotoxic T cells. As previously mentioned there are two different subsets of NK cells based on the production of different sets of cytokines: 1) NK1 cells secreting IL-10, IL-12 and INF-γ; and 2) NK2 cells secreting IL-4, IL-5 and IL-13.

4.2.5 The cytotoxic mechanisms

Two killing mechanisms are used by the NK cell 1) induction of cell lysis and 2) apoptosis.

1) Lysis of a target cell is mediated by the release of granzymes and perforin by exocytosis of cytotoxic granules within the NK cell. Granzymes belong to a family of serine proteases involved in the lytic activity, where for example granzyme B promotes DNA fragmentation and is involved in cell death. Perforin is a pore-forming protein which leads to death by making the cell membrane leaky, but it is not fully established whether the cell death is caused by membrane pore formation or endosome permeabilization.

2) Apoptosis, programmed cell death, is induced via binding of Fas ligand on the target cell to the Fas receptor on the NK cell.
NK cell killing can be understood as a shift in the balance between activation and inhibition, either loss of inhibitory receptors or a gain in activating receptors. Examples of activating receptors are Ly49s3 in the rat and NKG2-D, NKp30, NKp44 and NKp46 in the human, and inhibitory receptors Ly49i2 in the rat and NKRP-1a in the human.

When killing is initiated the NK cell becomes polarized where the cytolytic granules that are released via exocytosis upon ligation to a target cell are directed to the part of the membrane in contact with the target cell, and the receptors involved in apoptosis are organized on the cellular surface toward the target cell. This area of contact between target and “killer” is referred to as the immunologic synapse. Binding and activation of different receptors (immune receptors, costimulatory- and cytokine receptors) lead to phosphorylation of intracellular proteins and rearrangement of the cytoskeleton, which is a prerequisite for directing the cytotoxic granules towards the immunologic synapse. Conversely, when inhibitory signals dominate, the activating signal pathways are blocked by phosphatases, leading to dephosphorylation, dormant NK cell and no exocytosis of cytotoxic granules.
**Figure 2**
A cartoon showing some of the intracellular events when a target cell engages with an inhibitory NK cell receptor. The ITIM motif is phosphorylated by Scr kinases and activated, facilitating binding of SHP-1 and downstream dephosphorylation of intracellular proteins leading to NK cell inhibition.

**Figure 3**
A cartoon showing some of the intracellular events when a target cell engages with an activating NK cell receptor. An adaptor molecule (green “squash”) with ITAM motif is recruited when target cell binds to the NK cell receptor. This adaptor protein is linked to the activating Ly49 molecule by coupling a negatively charged amino acid in the adaptor protein to an arginine in the Ly49 molecule. The ITAM becomes phosphorylated and activated leading to Syk binding. Syk, a kinase, phosphorylate downstream molecules and leads to activation of the NK cell and then killing of the target cell.
4.3 NK CELL RECEPTORS

2.3.1 Killer cell lectin-like receptors (KLR)

Lectins are carbohydrate binding receptors containing a carbohydrate-recognition domain - CRD. In vertebrate lectins there are several different types of CRDs. The largest family are the C-type lectins (CTLs), so called because the saccharide binding is Ca\(^{2+}\) dependent. CTLs are involved in a variety of functions, such as cell adhesion (selectins), glycoprotein clearance and innate immunity (collectins). Fourteen different groups of CTLs have been described. CTL group V is primarily expressed by NK cells and referred to as killer cell lectin-like receptors - KLR. The KLR are atypical in that most of them seem to bind protein ligands instead of saccharides.

In the human, the rat and the mouse the genes encoding the KLRs are located in a chromosomal cluster called the Natural Killer cell gene complex (NKC). In the rat this complex is located on the distal part of the long arm on chromosome 4, 4q42. The proteins encoded by this gene complex are type II receptors, i.e. single pass transmembrane (TM) proteins with the amino terminus located intracellularly. The NKC in the rat contains 67 receptor genes, which can be divided into different subfamilies regarding protein structure. Some of these subfamilies contain multiple members and others only one gene (such as CD94 and KLRE). Three of the former are the NKR-P1, the NKG2 and the Ly49 subfamilies, all of which contain both activating and inhibitory receptors.

4.3.1a NKR-P1

The molecules constituting this subfamily are disulphide bonded homodimers expressed on both NK (CD3\(^{-}\)) and a subset of T cells referred to as NKT (CD3\(^{+}\)) cells. In the rodents the genes are located to the proximal part of the NKC. NKR-P1A and -C mediate activating signals, whereas NKR-P1B and -D are inhibitory. In the human only one gene, hNKR-P1A, has been detected \(^{19}\). It putatively mediates activating signals to the human NK cell.
4.3.1b NKG2/CD94

NKG2 members form disulphide bonded heterodimers together with CD94. The heterodimer is expressed by NK cells and a subset of T cells in the rat, mouse and human. CD94 is the invariant part of the receptor complex whereas there are three different NKG2 members in rodents, both activating and inhibitory molecules. The inhibitory NKG2A contains an ITIM in the intracellular part of the receptor. The ITIM motif is an intracellular protein sequence recruiting and activates the scr protein which upon activation leads to intracellular down stream inhibition of the NK cell, see figure 2 page 13. On the other hand the activating NKG2C and -E contain no ITIM motif but a charged amino acid (lysine) in the transmembrane domain which recruit and bind to DAP10. DAP 10 is a membrane bound activating adaptor protein housing an intracellular ITAM motif, which lead to activation of the NK cell, see figure 3 page 13.

4.3.1c The Ly49 subfamily

The Ly49 subfamily constitute the largest KLR family in the rat, and seem to form disulphide bonded homodimers expressed by NK cells and subsets of T cells. In the human Ly49L is the only Ly49 gene detected which is a truncated not expressed gene. In rodents there are multiple genes, 34 in the rat (see figure 5) and 23 in the mouse, named Ly49A to Ly49W. The so far known ligand for these receptors are MHC class I molecules and MHC class I homologues like MICA.

The inhibitory Ly49 receptor molecules have an ITIM motif in the cytoplasmatic region which the activating molecules lack. The activating molecules holds, whereas the inhibitory molecules lack, a transmembrane region harbouring a positively charged amino acid. This arginine is the link to activation via coupling to an activating adaptor molecule. Following the transmembrane region is a stalk separating the ligand binding domain from the cell membrane. The ligand binding region within the Ly49 family is the highly conserved C-type lectin-like domain (CTLD), shown in figure 4, consisting of two α-helices (α1 and α2), two anti-parallel β-sheets formed by
β-strands (β0, β1 and β5) and (β2, β2’, β3 and β4). There are four intrachain disulphide bonds, two of which are the characteristic invariant disulphides found in C-type lectins located in position C1 and C2, see figure 12. A third disulphide bond is linking strands β0 and β1. The fourth is unique to the Ly49 and links the N-terminus of strand β5 to the first turn of helix α1 (Cys167 – Cys253 in figure 4).

The intracellular part (N-terminus) of the inhibitory Ly49 receptors has an immunoreceptor tyrosine-based inhibitory motif (ITIM) close to the N-terminus. Upon receptor ligation the ITIM is phosphorylated, which leads to downstream inhibition of the NK cell by recruitment of SH-2 domain containing tyrosine phosphatases or by the lipid phosphatase inositol-5 phosphatase (SHIP)\(^{27,28}\). The activating receptors lacks the ITIM motif, has a shorter intracellular domain and has instead arginine, a positively charged amino acid, in the transmembrane domain which associates with DAP10, a membrane bound adaptor molecules containing immunoreceptor tyrosine-based activating motifs (ITAMs). Upon receptor ligation the ITAMs become phosphorylated and recruit downstream Syk or Zap 70 leading to cellular activation\(^{29}\), see figure 3. In addition to these two receptor types I have cloned three receptors that contain both an ITIM and an arginine in the transmembrane domain. How these apparently ‘bifunctional’ receptors work is not known and analyses are about to be started.
4.3.2 Killer cell Ig-like receptors

The Killer cell Immunoglobulin- (Ig-) like Receptor (KIR) family is characterized by molecules having one or several Ig-related domains. The receptors are encoded by genes located in the Leukocyte Regulatory gene Complex (LRC) on chromosome 1 in the rat. The LRC is constituted by the KIR family, the leukocyte Ig-like receptor (LILR) family, the leukocyte-associated Ig-like receptor (LAIR) family, GPVI, Fcα receptor (CD89) and NKp46. Some of these molecules are expressed on different cell types, whereas NKp46 is expressed on NK cells only. KIRs are expressed on NK cells and a subset of T-cells.

4.3.2a Killer cell Ig-like Receptors (KIR)

The KIRs are expressed as membrane bound monomer, modulating the NK cytotolytic response by either activation or inhibition. Subset of MHC class I molecules seem to be ligand for the KIRs. The KIR molecules are immunoglobuline (Ig) holding receptors and are given name depending on the number of extracellular Ig domains and the length of the intracellular domain. KIR2D members have two and KIR3D three Ig domains. L and S depict whether the KIR molecule has a long or a short intracellular domain, respectively. Only one KIR has been detected in the rat so far, whereas in the human there are 14 \(^{30,31}\). The long KIRs (L) tend to be inhibitory by containing ITIM motifs in the cytoplasmatic region, whereas the short ones (S) have a charged amino acid in the transmembrane region associating with DAP12, an activating adaptor protein containing an ITAM motif \(^{32}\). One interesting molecule is the putatively bifunctional KIR2DL4 that contains both a charged amino acid and an ITIM motive. At present its biological function is unknown.

4.3.2b Leukocyte Ig-like receptors (LILR) and leukocyte-associated Ig-like receptor (LAIR)

The LILR family, also known as the Ig-like transcripts (ILT), resembles the KIRs, FcαR and gp49 \(^{33}\). It also includes the leukocyte immunoglobulin-like receptor 1(LIR-1) which can bind UL18, a cytomegalovirus encoded protein which mediate inhibitory signaling via ITIMs \(^{34}\).
The last member is the monocyte/macrophage immunoglobulin-like receptor (MIR) which has an ITIM motifs, these receptors are expressed not only on NK cells but also on macrophages, monocytes and dendritic cells\textsuperscript{35}. The activating molecules associate with the FcεRIγ adaptor molecule which transmit the activating signal. The LAIRs are type 1 single transmembrane receptors with one Ig domains in the extracellular region and 2 ITIM motives intracellularly. The LAIR family consists of two inhibitory members, LAIR1 and LAIR2. LAIR1 is expressed both in the human and the mouse, whereas LAIR2 is at present detected in the human only. These receptors are expressed on different mononuclear cells, such as NK cells, B cells, T cells, monocytes, dendritic cells and thymocytes\textsuperscript{36}.

4.3.2c NKp46

Due to its expression in all NK cells and NK cells only, NKp46 is a NK cell marker. NKp46 has two Ig domains extracellularly and an arginine within the transmembrane domain, hence an activating receptor\textsuperscript{37}. In the rat NKp46 is associated with a heterodimer of CD3ζ and FcεRIγ ("Rat NKp46 activate natural killer cell cytotoxicity and is associated with a heterodimer of CD3ζ and FcεRIγ, Westgaard, IH – submitted manuscript").
5 METHODS & TECHNIQUES

5.1 The polymerase Chain Reaction (PCR)

PCR is a method for rapid amplification of a specific DNA segments. Briefly, the principle of the method is as follows, take also a look at figure 5: The DNA polymerase, which adds nucleotides to a new strand using a pre-existing strand as a template, cannot initiate the synthesis of the new strand from the start (i.e. cannot prime DNA synthesis), only continue to add nucleotides to the tail of partially synthesized strand. The sites from where synthesis occurs on the template are thereby determined by where the partially synthesized strands are found. During DNA replication, these ‘primers’ are made by an RNA polymerase, called primase. This process can be imitated in vitro, by making synthetic oligonucleotides that match specific sites of the template DNA, thereby directing DNA synthesis to these sites. A specific DNA sequence is amplified by making primers flanking the sequence, complementary to the two antiparallel strands in the DNA double helix.

Figure 5 – Principles of the PCR method
In a PCR reaction the following ingredients are added: 1) DNA template, 2) the two primers, 3) the four nucleotides as activate triphosphates (dNTPs, where N stands for A, C, G or T), 4) a suitable buffer and 5) heat stable DNA polymerase. The template DNA is first denatured, i.e. the two stands are separated by heating the PCR mix to 94°C. By cooling the mix down the DNA strands start to re-anneal, i.e. the complementary sequences bind to each other. Because the primers are in much higher concentrations than the template, the primers bind much more rapidly to the complimentary region in the template DNA than the original template strands. Immediately following primer binding, the DNA polymerase starts to synthesize a new strand from the 3’ end of the primer with the base sequence determined by the template strand (see figure 5). From the third round amplification is exponential, so that running 30 cycles results in \(2^{30}\) (> 268 million) sequence specific copies. The amplified products are separated by gel electrophoresis and visualized by staining with e.g. ethidium bromide, which intercalates with the DNA and fluoresce when exposed to UV light (see figure 6, page 23)).

5.2 Gel electrophoresis

To visualize and quantify small molecules they first ought to be separated which is done by using a matrix gel, normally made from agarose or polyacrylamide, and then staining the gel by different means. Gel electrophoresis (GE) takes advantage of the migration of charged molecules within an electric field, negatively and positively charged molecules migrating towards the opposite charged electrode. The gel matrix obstruct larger molecules more than it does small, and will separate the molecules on the basis of size; small molecules migrating faster than large. Hence molecules of same size and charge migrate and position equally in the gel. Agarose gels are used to separate large molecules, like PCR products whereas polyacrylamide gels are more used for smaller molecules such as proteins. After separating different molecules within a sample by using GE the molecules can be visualized by using a molecule specific staining method. Ethidium bromide is one example which stains nucleic acids by intercalating between the nucleic
acid molecules and makes it fluorescent when exposed to UV light. Other methods can be applied to visualize smaller molecules after they have been transferred to a membrane (blotting). Silver staining or Comassie brilliant blue dyes can show proteins, as can antibodies after blotting. Figure 6 shows an ethidiumbromide stained gel after gel electrophoresis has been run.

5.3 Hybridization

Hybridization is a method for detecting specific DNA sequences. The principle is complementary binding of a certain DNA sequence to a template. It is a method to see if a known sequence (query) is present in a sample of DNA. If the query is present within a DNA sample it will bind complementary to the template, where it is identical, and this binding can be visualized by placing a radioactive sensitive film over the washed and cleaned membrane. Since the template is non radioactive whereas the query is, the film will be labelled where the query binds to the template only. The sample is usually 1) genomic DNA, 2) cDNA, 3) cDNA cloned into a plasmid vector, or 4) genomic sequences cloned into e.g. BAC clones. Ad 1) If the genomic DNA is digested with restriction enzymes followed by gelelectrophoresis and blotting onto a membrane, the technique is referred to as Southern blotting. Ad 2) If the cDNA is blotted in arrays onto filters or glass plates, the technique is referred to as dot blots, macro arrays or micro arrays. Ad 3) The cloned cDNA is usually a cDNA library, where the goal is to clone homologous sequences. Ad 4) Here the purpose can be to identify BACs containing specific genes, e.g. with the aim of finding which genes are mapping together on the chromosomes.

A probe is needed in order to search for a specific sequence. The probe is a copy of the query and is labeled with a substance that can be used for detection, for instance a radioactive tag. The template is immobilised on a membrane, which is incubated with the labeled probe and then washed (the incubation and washing conditions are collectively referred to as ‘stringency’ - see below). The probe will bind to its complementary sequence, if present, and is then visualized. If for example a radiolabeled probe is used, the membrane is covered with a sensitive film and the binding (and position) of the probe detected. The stringency of hybridization can be altered by
using different baking temperature (increase of temperature leads to increase of stringency) and salt concentration during washing (decrease in salt concentration leads to increase in stringency).

5.4 TA cloning

TA cloning is a method to generate copies of a DNA sequence, for instance a PCR product, by using a prokaryote host such as E.Coli. The desired DNA sequence is ligated into a cloning vector using the mechanisms of A-T complementary binding. Taq polymerase attach adenosine residues to the flanking ends of the DNA sequence, these DNA ends can then bind to the complimentary thymine residues flanking the cloning sites in a cloning vector. Ligation of the DNA fragment into the vector is mediated by a DNA ligase. The vector-DNA-unit is transfected into a bacterial host by instabilising the bacterial membrane. This is done by shocking the cells either electrically or by heat. Within the vector sequence there are genes coding for antibiotic resistance which will lead to a selection of the bacterias containing the vector when it is grown in an antibiotic containing LB medium. The bacteria containing the vector are grown over night at 37°C on LB medium plates, agar plates containing antibiotics, giving rise to many colonies representing one bacterial clone. Vector specific PCR is run on different colonies from the LB plate to check which colony representing the starting point sequence. One selects several colonies showing the proper PCR results and these are grown in bottles with liquid LB medium added antibiotics, one bottle for each colony. In this way the starting point sequence has been cloned and multiplied. The next step is to separate the vector-DNA-unit from the bacterial solution, a process that can be done in different ways. One method is the use of Maxi-prep preparation kit.

5.5 Fingerprinting

Fingerprinting is a method to give a hint of the similarity between different DNA sequences by exposing the DNA sequences to a combination of restriction enzymes. Two identical DNA strands cut with the same restriction enzymes will show the identical pattern when compared on the gel after GE. Two partly identical DNA strands will show some similarity and some disparity
whereas totally different strands will not share any similar pattern. This is based on the feature that similar DNA sequences give rise to the same DNA fragments and gel electrophoresis pattern when cut with the same restriction enzymes.

By cutting a DNA sequence with a set of three or four endonucleases and running the resultant product on a gel the resultant product gives rise to a “DNA fingerprint”.

This can be used to indicate whether different BAC clones (Bacterial Artificial Chromosome, a vector capable to carry a DNA fragment up to 250 kbp) share the same DNA sequence and hence will be overlapping, see figure 6 below.

5.6 Making a PAC contig – presented in a tailing pattern

A PAC (P1-derived artificial chromosome) contig is a continuous row of overlapping PAC clones representing a segment of the genome. A PAC clone is a vector containing approximately 130,000 nucleotides from a genome. The PACs are identified by the use of PCR and is then organized into a pattern where different PACs containing the same sequence is overlapping. By running PCR on many different PACs one can establish a contig of overlapping PACs making up a tailing pattern. I did this on PACs sent from The Welcome Trust Centre for human genetics by using NKC marker specific PCR primers and built up a contig representing the NKC in the rat. The same can be done on BACs (Bacterial Artificial Chromosomes), but this time it is done electronically, see next chapter and figure 7.
5.7 Chromosomal mapping and bioinformatics

The sequencing of the rat genome, The Rat Genome Project, has been an international collaboration where the resultant, fully sequenced rat genom, generated from the BN strain, is available on the internet site “Ensembl” (http://www.ensembl.org/Rattus_norvegicus/). On this site the chromosomal localization of genes can be done by performing a pairwise blast with a query DNA sequence against the whole BN rat genom. An exact position of the query within the sequenced genome will be given. The position is given by chromosomal numbering and the exact base pair position relatively from the first chromosome base pair. A single search gives rise to several hits within the genome, some identical sequences and others just sharing spares similarity, to segregate the hits the resemblance between the query and the genome is indicated by percentage similarity. Because this site is under continuous upgrading the exact position within the chromosome vary depending on when the blast is performed. Another recurring problem is the gaps within the sequence. These gaps are of different size and one consequence of the gaps is that these might harbor sequences that are genes.

BAC clones containing pieces of the BN genome are shot-gun sequenced and build up to overlap and cover the whole BN strain genome. These clones are available via the “NCBI Blast 2 sequences site” and can be used to build up a contig. The contig will establish the consecutive order of the genes on basis of what genes are located within which BAC clone when there is a BAC clone overlap along the genome, see figure 7.
6 RESULTS

Six rat Ly49 genes (cDNA) had previously been published (the new names): Ly49-9 (i1), -12 (s2), -19 (p8) and -29 (s1, from the F344 rat strain) and Ly49i2 and Ly49s3 (from the PVG rat strain) (all from IBL in collaboration with Veterans Administration Medical Center and University of California, San Francisco). In addition, Christian Naper had cloned, but not yet published, another seven rat Ly49 genes, provisionally called -i5, -s5, -7 (i6), -19P (s7), -29P (s4), -30 (i8) and -35 (i4, from the PVG strain). In all 13 rat Ly49 genes had been cloned, four from the F344 and nine from the PVG strains. I have screened the released genomic sequences (both individual BAC clones and the assembled whole rat genome sequence) with these cDNA sequences in order 1) to identify their chromosomal positions, and 2) to find novel Ly49 genes. To the extent that novel genes have been found, I have then tried to cDNA clone these in vitro. In the following, I shall briefly describe my findings. The gene nomenclature is based upon the receptor function of the specific Ly49 molecule and when they were identified, giving the first identified activating/stimulatory molecule the name s1. The receptor function is denoted with either s – stimulatory, i – inhibitory or si – bifunctional. As will be shown, 21 new Ly49 genes were identified. Of these, I have cDNA cloned five new genes. Nine of the genes are predicted to be non-functional in the BN strain. In the following these findings will be presented in more detail, including analysis of the sequences with respect to predicted functional properties of the receptors and the evolution of this gene complex.

6.1 Identification of Ly49 genes from the genomic sequence

New Ly49 exons were identified by similarity search (blasting) BAC clones (paired blasts) and the whole genome sequence with previously sequenced Ly49 cDNA sequences, a process called in silico cloning. Exon-intron splice junctions were deduced by using information from the already existing Ly49 molecules, contiguous exons joined and the resulting sequences translated and compared with previously known proteins. Sequences containing all six coding exons giving complete open reading frames were defined as candidate functional genes. Sequences with
nonsense (stop codons) or frameshift mutations or lacking exons were considered likely to be pseudogenes. Primers were constructed from the putative 5’ and 3’ untranslated regions of the candidate functional genes and used for RT-PCR on mRNA isolated from DA or PVG LAK cells (lymphokine activated killer cells, i.e. NK cells cultured for ~10 days in the presence of interleukin-2). The PCR products were analyzed by gel electrophoresis. Bands of expected lengths were excised form the gels, amplified by TA cloning (see Methods & Techniques, page 22) and the inserts sequenced form both ends in Germany at the Medigenomix. In this way the following five new genes were cDNA clones (rat strains from which they were cloned in parenthesis): Ly49si2 (DA and PVG), Ly49si3 (PVG), Ly49si1 (DA), Ly49i3 (DA) and Ly49i7 (DA).

6.2 High resolution physical map

Before the whole rat genome sequence was released in year 2003, I made a high resolution map of the Ly49 genes as deduced from overlapping BAC clones arranged in a tiling pattern, as shown:

![Map of the Ly49 gene region as deduced from overlapping BAC clones](image)

The BACs sequenced by the rat genome consortium are from the BN strain. Their names are shown in the first column. The Ly49 genes are lined in order at the top. The horizontal yellow lines represent each BAC with its corresponding genes. Y (yes) means that the gene is present in that clone, N (no) that it is absent, G (gap) that there appears to be a gap in that particular clone (AC111574). i = predicted inhibitory receptor, s = stimulatory (activating) molecule, b= bifunctional molecule and ψ= pseudogene. †= pseudogen in the BN, functional in the PVG (s5).
After assembly of the whole rat genome sequence the various Ly49 sequences were blasted against this sequence and their consecutive positions and orientations identified, as shown in the figure below. The top line represents the whole natural killer cell gene complex, which is seen to span ~3.3 Mb. The Ly49 (KLRA) gene region is situated in the distal part of the NKC, is the largest cluster within the NKC and makes up slightly more than half of the gene complex in total.

Protein sequence comparison of all the Ly49 molecules lead to division of the gene complex into three major blocks containing multiple closely related genes, with three genes falling outside the blocks, Ly49i13, -i6 and i8 from proximal to distal location (see figure 8). There is almost absolute colinearity when comparing the physical map with the BAC contig.

![Figure 8](image_url)

**Figure 8**  
*High resolution map of Ly49 region of the NKC as deduced from the rat genome sequence (Ensembl release 19.3b.2 - February 2004)*

Arrows indicate gene orientation with length of the arrows corresponding to 20 kb. Vertical boxes represent individual exons with the width indicating exon length. The numbers below each box indicate positions of exon I for each gene to be added to the number to the left (e.g. Ly49i13 at position 167 Mb + 425 kb = 167 425 kb from the centromere). * exon 6 cannot be localized in the Ensembl genome and the internal distance between exon 5 and 6 is estimated to 9 kb. - - inhibitory, + + activating, +/- bifunctional, gf: gene fragment, p: pseudogene.

### 6.3 Cluster size and gene density

With the exceptions of Ly49s6, which seems to lack exon 3, encoding the membrane proximal stalk, and Ly49i8, which has an additional exon of 21 bp (corresponding to 7 amino acids)
between exon 1 and exon 2, all the other Ly49 receptors are encoded by 6 exons: exon 1 encoding the cytoplasmic part, exon 2 the transmembrane domain, exon 3 the membrane proximal stalk and exons 4 to 6 the lectin-like domain.

The Ly49 cluster spans 1,766 Mb from one end to the other. The average length of an open reading-frame Ly49 gene is 21.3 kb [range 11.1 kb (si1) to 32.2 kb (Ly49i8)]. The median distance between two consecutive genes (exon 1 to exon 1) is 51 kb (mean 73.8 kb). The central third of the complex has the highest gene density - 2.25 genes per 100 kb, whereas the corresponding figures are 1.77 and 1.79 for the proximal and the distal thirds, respectively.

All of the Ly49 genes are encoded by the minus strand (i.e. exon 6 is located proximal to the exon 1) and thus have the same orientation. For all the genes intron 5, separating exons 5 and 6, is by far the largest, usually > 5000 nucleotides and in some cases even >10,000 nucleotides separating the two exons.

---

**Figure 9 – Distribution of gene lengths**
The lengths of the Ly49 genes, including introns, as measured by the genomic distance between the start and stop codons. Mode $x \in [19,22]$ kb, median 20.9 kb and mean 21.3 kb.

**Figure 10 – Distribution of intergenic distances**
Distributions of distances between start codons of contiguous genes. Mode between 30 and 40 kb, median 48 kb and mean 51.7 kb.
6.4 Sequence analysis

The Ly49 molecules are very similar concerning nucleic acid sequences and protein structure. Here I will comment on some of the variant features among the members. Amino acid alignment of exons 1 and 2 for all the genes are shown below in fig.8. Inhibitory Ly49 receptors have an ITIM motif intracellularly, characterized by the amino acid sequence I/VxYxxV/L (I=isoleucine, V=valine, Y=tyrosine and x=any amino acid). Activating Ly49s have a positively charged amino acid (arginine - R) in the transmembrane domain, recruiting an activating adaptor molecule holding an ITAM motif by binding to a corresponding negatively charged amino acid in the transmembrane domain of the adaptor. Five of the receptors exhibit both features and are therefore referred to as bifunctionals. To what extent these are activating or inhibitory (or both) is presently unknown. Ly49i8 has the most deviant sequence, including an extra exon (exon 1b), encoding 7 amino acids, inserted between exons 1 and 2. In the genomic sequence this extra exon 1/2 lies 780 bp downstream of exon 1 and 655 bp upstream of exon 2.

Use figure 4, 11 and 12 to get a visual impression of the molecule. The lectin domain starts with an α-helix (α1), followed by a β-sheet, (β1), via a connecting strand. There is a cluster of acidic amino acids (aspartic acid - D or glutamic acid - E) at the end of exon 4 of all the receptors except Ly49s6, which lacks the five last amino acids. The conserved WIGL motif (aa210-213), follows just after the β2-strand. The first isoleucine and the second valine are in some molecules substituted with equivalent hydrophobic amino acids, valine and serine/phenylalanine respectively. The next β-strands, β3-5, are encoded by exon 6. The Ly49 lectin domain is characterized by two S-S bonds connecting the α1 helix to the β5 strand. Ly49i8 may have a third S-S bond here. Another striking novel observation is that eight of the predicted Ly49 receptors have a conserved cysteine residue at the end of exon 6, in position C279, which would be expected to face an extra cysteine residue in the beginning of exon 4, in position C142, found in these receptors only at base of the lectin domain, with the two probably forming an extra disulphide bridge (the positions marked with ▼1 in the alignment).
Figure 11 – Exon alignment
Alignment of peptide sequence encoded by exon 1 and 2 I/VxYxxV in exon 1 represents the ITIM motif (red: inhibitory molecules, violet: bifunctional). R in exon 2 (turquoise: activating molecules, violet: bifunctional) represents the arginine involved in the interaction with the DAP12 adaptor molecule. Predicted transmembrane region underlined.
**EXON 4**

\[ ▼^1 \] C G CYY D W C^1 C^2 L IDDEDEL EXON 5 W G
s5 RCVMHWFCHGIKCYYFIMDRITWHEKCTQCNYNLSFRKIDDEDEL KFLQDHIIDSDY WGL SYNNKEKWISWIDSPNL
s4 RCVMHWFCHGIKCYYFIMDRITWHEKCTQCNYNLSFRKIDDEDEL KFLQDHIIRSDY WGL SYNNKEKWISWIDSPNL
i3 (DA) RCDEMHWFCGIKCYYFIMDRITWHEKCTQCNYNLSFRKIDDEDEL KFLQDHIIRSDY WGL SYNNKEKWISWIDSPNL
s3 RCDEMHWFCGIKCYYFIMDRITWHEKCTQCNYNLSFRKIDDEDEL KFLQDHIIRSDY WGL SYNNKEKWISWIDSPNL
i2 RCDEMHWFCGIKCYYFIMDRITWHEKCTQCNYNLSFRKIDDEDEL KFLQDHIIRSDY WGL SYNNKEKWISWIDSPNL
s2 RCDEMHWFCGIKCYYFIMDRITWHEKCTQCNYNLSFRKIDDEDEL KFLQDHIIRSDY WGL SYNNKEKWISWIDSPNL

binding. Cn indicate disulphide binding within the lectin domain between the same numbers in exon 4 and 6. At the end of exon 6 there is a conserved

**EXON 6**

C^1 C^2 C^3 ▼^1
s5 DLACACPLKGTGICYFIMDSMTLHYDDCGRKHCICERGMDKIAPLCVSXESQSA V
s4 DLACACPLKGTGICYFIMDSMTLHYDDCGRKHCICERGMDKIAPLCVSXESQSA V
i3 (DA) DLACACPLKGTGICYFIMDSMTLHYDDCGRKHCICERGMDKIAPLCVSXESQSA V
s3 DLACACPLKGTGICYFIMDSMTLHYDDCGRKHCICERGMDKIAPLCVSXESQSA V
i2 DLACACPLKGTGICYFIMDSMTLHYDDCGRKHCICERGMDKIAPLCVSXESQSA V
s2 DLACACPLKGTGICYFIMDSMTLHYDDCGRKHCICERGMDKIAPLCVSXESQSA V
s1 SLNLKYNCNKGCMFSTTMRKLMNCAKCMNPFCICRKRFLDFPH

\[ ▼^1 \] C G CYY D W C^1 C^2 L IDDEDEL EXON 5 W G

\[ ▼^1 \] C G CYY D W C^1 C^2 L IDDEDEL EXON 5 W G

**Figure 12 – Exon alignment, exon 4 - 6**

Figure representing the exon 4 - 6. WxGx representing the WIGL motif, central part of the lectin domain. ▼^1 indicate possible site for a disulphide binding, C^1 indicate disulphide binding within the lectin domain between the same numbers in exon 4 and 6. At the end of exon 6 there is a conserved cluster of acidic amino acids. α and β represent the different helices. L represents the loops within the lectin domain.
6.5 Receptor groups and blocks

To the left is shown a dendrogram of the sequences, where the branching pattern illustrates the patterns of sequence similarities. The shorter is a horizontal branch at the end the higher is the similarity between the genes. Three major branches containing several genes are seen. In addition there are four smaller branches, three with only one gene (Ly49i13, i7 and i8) and one with two genes (Ly49s2 and Ly49p8) four smaller ones are seen. Interestingly, the branches coincide with the chromosomal gene order, which thereby can be defined as consisting of three larger blocks, denoted I, II and III in figure 13. Each block contains inhibitory as well as activating or bifunctional members. Block I lies most centromeric, block II in the central part of the complex and block III at distal end. To what extent Ly49i13 and –i6 can be assigned to a block has not yet been determined.

In contrast, Ly49i8 is clearly positioned outside the blocks, at the far distal end of the NKC.

When comparing mouse and rat Ly49 genes, the rat is seen to have more genes (25 genes with open reading frame plus pseudogenes) than the mouse (19 genes). Another striking feature is the lack of overlap on a combined dendrogram, see figure 14, with rat Ly49 genes clustering separately from the mouse Ly49 genes. There are some exceptions: mouse Ly49Q, which is more similar to the rat block I genes than to the other mouse genes, rat Ly49i6, which is more similar to the mouse than to the rat genes, and most strikingly, mouse Ly49B and rat Ly49i8, which are most similar to each other. The last two genes occupy the same position at the distal end of their...
respective gene complexes. These two genes therefore seem to represent true orthologs, i.e. they represent direct descendants from a single gene, with sequence differences resulting from speciation. For the other genes, the closer intraspecific sequence similarities either mean that the genes were duplicated after the split between mouse and rat ~40 million years ago, or that they somehow have been homogenized (i.e. made similar to each other after speciation), if so, most likely by gene conversion.

Figure 14 also shows Ly49 genes identified in the human (h), baboon (ph), cat (fc), dog (cf), swine (ss) and cow (bt). In all of these species only one Ly49 gene has been identified, and as can be seen from the figure, without clear resemblance to a specific rodent Ly49 gene. We cannot therefore identify with certainty which of the rodent Ly49 genes represent the most ancestral gene.

Figure 14 – Dendrogram of Ly49 molecules from different species

Dendrogram of all the Ly49 receptors from rat (r), mouse (m), dog (cf - canis familiaris), cat (fc - felis catus), baboon (ph - papio hamadryas), human (h), pig (ss - sus scrofa) and cow (bt - bos taurus). The Ly49 molecules make up families which to a lesser extent mix with each other. rLy49-7 is more closely related to the mouse molecules and the mLy49Q is more similar to the rat block I. The mLy49B and the rLy49-30 make up their own group separated from the other molecules both in the mouse and the rat.
6.6 Gene homogenization

Below are dendrograms of exon1 and exon 4 from all the rat receptors, figure 15 A & B. The following pattern emerges: Whereas inhibitory receptors tend to cluster separately from activating and from bifunctional for exon 1, the pattern is just the opposite for exon 4. Here smaller groups are formed, in most cases containing both inhibitory and activating members (or bifunctional instead of activating members). The significance of this observation is uncertain. Exon 1 (and 2) are involved in signal transduction, whereas exons 4 - 6 most probably are involved in ligand binding. The ligands are yet unknown, but as deduced from results in the mouse, probably are MHC class I molecules or MHC class I like molecules. It is possible that receptors showing high sequence similarity in the lectin domain share ligands. If so, one interpretation of this observation is that the grouping together of inhibitory and activating (or bifunctional) reflects a repertoire of receptors with opposite signaling functions directed at shared ligands. Why we should have such a system is presently unclear.
A question is whether this system is based on selective pressure conserving the ligand recognition parts of the receptors, or whether it reflects homogenization. In the latter case, one would expect that some of the exons from opposing receptors exhibit particularly high sequence similarities.

This seems indeed to be the case. Figure 16 below shows a dendrograms of all the exons of eight selected receptors, four inhibitory and four activating. The complete sequence identities for some of the exons bear evidence of relatively recent homogenization events.

**Figure 16 – Dendrogram showing homogenization**

Some molecules show complete exon identity seen in the figure as horizontal lines without vertical lines, e.g. exon 4 in Ly49-29 and Ly49-9. (Ly49-9 = Ly49i1, Ly49-29 = Ly49s1, Ly49-35 = Ly49i4 and Ly49-36 = Ly49s6)
7 DISCUSSION AND CONCLUSION

In this work I present a high resolution physical map of the Ly49 region of the rat NKC, based on the rat genomic sequence. From the rat genome sequence 34 Ly49 genes have been identified, of which 7 are predicted to encode activating receptors, 13 inhibitory receptors, 5 bifunctional receptors, and 9 pseudogenes. All the genes have the same orientation and are distributed fairly regularly along the chromosome, with an average distance of ~52 kb and an average gene length of ~21 kb. In addition to the new Ly49 cDNAs cloned here, novel findings are:

1) The presence of five apparently bifunctional Ly49 receptors, never previously described in other species (three of which I have cloned), as well as novel sequence features for some of the receptors, in particular extra cysteine residues that seem to make additional intrachain disulphide bonds.

2) The number of intrachain disulphide bonds in the lectin domains of Ly49 receptors is higher than in other CLSF members. Why these extra stabilizing bonds have been selected for is unknown.

3) The rat Ly49 gene region can be divided into three major blocks encoding related receptors. High sequence similarities, even complete sequence identities, indicate extensive occurrence of gene homogenization. This seems not to take place between genes in different blocks, only between genes within the same block, although the genes can be spaced far apart from each other within the block (e.g. si2 and si3 in block I). In order that duplicated genes shall develop into separate sequences with different functional properties, it is necessary that homogenization is escaped. The NKC must have developed by duplications of a single ancestral gene, whereafter separate single and multigene subfamilies were formed. The close sequence similarities forming the basis of multigene families may not necessarily represent recently duplicated genes, but genes that have not yet escaped homogenization. It is possible that the lack of homogenization observed between the three Ly49 blocks is an expression of incipient escape, so that with time the three blocks will drift apart into separate subfamilies, but inside each block retains their sequence similarities due to continued homogenization.
4) Within each block the receptors form smaller groups, showing high sequence similarities of the exons encoding the lectin-like domain. As a rule, most of the smaller group contain both inhibitory and activating (or bifunctional) receptors, suggesting that the system is designed so that the ligands (not yet defined, but probably MHC class I molecules) shall be recognized by both types of functionally opposed receptors.

5) Comparison with mouse sequences gives a picture of a highly dynamic genetic region in rapid change. This is supported by comparisons between different rat strains, where e.g the pseudogene s5† (see figure 7) in the BN strain probably is the equivalent of Ly49s5 in the in the PVG strains, where it seems to be functional (activating).

In conclusion I have cloned cDNA new rat Ly49 receptors, and by sequence and genome analysis detected several novel features about this large subfamily of C-type lectin receptors expressed by NK cells. The findings should contribute to understanding the evolution and shedding light on the function of this group of receptors.
8 REFERENCE LIST


APPENDIX

The genes and gene organization of the Ly49 region of the rat Natural killer cell gene complex

Øyvind Nylenna1,5, Christian Naper1,5, Jon T. Vaage2, Peng Y. Woon3, Dominique Gauguier3, Erik Dissen1, James C. Ryan4, Sigbjørn Fossum1

1Institute of Basic Medical Sciences, University of Oslo, 2Institute of Immunology (IMMI), Rikshospitalet University Hospital, Oslo, 3Wellcome Trust Centre for Human Genetics, University of Oxford, 4Department of Arthritis and Immunology, University of California 5Equal contributions

Brief title: The Ly49 gene region of the rat NKC
Key words: NK cells, gene complex, lectin-like receptors, phylogenesis

Corresponding author:
Sigbjørn Fossum, Department of Anatomy, University of Oslo, N-0317 Oslo, Norway.
Phone: +47 22 85 12 13 Fax: +47 22 85 12 78
e-mail: sigbjorn.fossum@basalmed.uio.no

Abbreviations: CLSF: C-type lectin superfamily, ITIM: immunoreceptor tyrosine-based inhibitory motif, KLR: killer cell lectin-like receptor, NKC: natural killer cell gene complex, TM: transmembrane

This study was supported by The Norwegian Cancer Society, The Norwegian Research Council, Bergljot and Sigurd Skaugen’s fund and The Wellcome Trust.
Abstract

We have previously published the cDNA sequences of six rat Ly49 genes and here report eleven new genes with full and three with incomplete open reading frames. Although obtained from different inbred rat strains (PVG, DA and F344) the cDNAs so far cloned seem to represent non-allelic genes as they match with different loci in the BN rat genome. The BN rat genome is predicted to contain 34 Ly49 loci distributed over ~1.75 Mb of the distal part of the rat NK cell gene complex. Some of the cloned genes appear to be mutated to non-function in the BN genome, which harbour additional genes with full open reading frames. Our analysis indicates that there may be at least 26 non-allelic functional Ly49 genes in the rat. Based on deduced amino acid sequences, thirteen of the encoded receptors are predicted to be inhibitory, eight to be activating, whereas five may be both (‘bifunctional’). Phylogenetic analysis bears evidence of a highly dynamic genetic region, where only the most distally localized Ly49 gene has a clear-cut mouse ortholog. In phylograms the majority of the rat genes cluster into three larger subgroups. The genes within each subgroup map together, defining three chromosomal regions which seem to have undergone recent expansions. When comparing only the exons encoding the lectin-like domains, the receptors form smaller subgroups, most containing at least one inhibitory and one activating or ‘bifunctional’ receptor. The close sequence similarities between exons from receptors with opposite signalling functions suggest recent homogenization, probably by gene conversion.
1 Introduction

Upon antigenic challenge NK cells respond by releasing cytokines and by killing target cells. In addition to eliminating infected and malignant cells, they recognize and kill semiallo- and fully allogeneic bone marrow derived cells, a feature of uncertain physiological relevance, although of clinical importance in bone marrow transplantation. To the extent target cell ligands are known, they have in many cases turned out to be MHC class I or related molecules. The first MHC class I binding NK cell receptor cloned was mouse Ly49A[1]. A plethora of MHC-class I binding NK cell receptors have by now been found [2]. They are of two different types: the killer cell lectin-like receptors (KLR) and the killer cell and leukocyte Ig-like receptors (KIR and LIR). Both types contain inhibitory as well as activating members [3-5], the former with intracellular domains containing immunoreceptor tyrosine-based inhibitory motifs (ITIMs), the latter as a rule with a positively charged amino acid in the transmembrane (TM) domain mediating association with adaptor proteins containing immunoreceptor tyrosine-based activating motifs (ITAMs). Following receptor ligation, tyrosine phosphorylated ITAMs recruit protein tyrosine kinases [6;7], triggering signaling pathways that activate NK cells. In contrast, tyrosine phosphorylated ITIMs recruit and activate the tyrosine phosphatase SHP-1 that counteracts the activating signals [8-10]. The KLR are encoded by the natural killer cell gene complex (NKC), localized on mouse chromosome 6F3 [11], rat chromosome 4q42 [12] and human chromosome 12p13.31 [13]. In the human there are many KIR/LIR loci [14], but only a single Ly49 gene, which is mutated to non-function [15]. In rodents the situation is reverse, with Ly49 by far the largest subfamily of NK cell receptors, and only a few KIR genes [16;17].
According to the current BN rat genome assembly (8 Feb 2004), the rat NKC, defined as the chromosomal region between *Nkrp1b* (=*Klrb1*) and *Ly49i8*, spans ~3.2 Mb and contains 67 KLR genes divided among five multigene families and twelve single genes (unpublished). This amounts to all KLR genes hitherto identified, except *Klrg1*, encoding the mast cell receptor MAFA-1 (KLRG1), which lies ~7 Mb proximal to the NKC. Furthermore, it contains no other types of genes, with the single exception of the GABA receptor associated protein-like gene, *Gabarapl1*, which lies in the middle part of the NKC, close to *Klre1* [18]. All KLR genes, including the Ly49 genes, encode type 2 single-pass transmembrane disulphide-linked homodimeric proteins with an N-terminal C-type lectin-like domain.

In the mouse, the functional gene *Cmv1*, conferring resistance to cytomegalovirus infection, was demonstrated to be associated with Ly49H [19], and *Chok*, involved in recognition of Chinese hamster ovary (CHO) cells, to be associated with Ly49D [20]. In the rat, we mapped the alloreactivity gene *Nka* to the distal part of the Ly49 gene region.

In the pursuit of the associated structural gene, our wanted to obtain an accurate inventory of all rat genes in this chromosomal region, preferably from the inbred PVG and DA strains defining *Nka* [12]. We here report the cDNA cloning of new Ly49 genes in addition to the six previously published. Moreover, on the background of the seemingly anomalous propensity of rat NK cells to be triggered by alloantigens [12;21-23], we used the recently released genome sequence of the Brown Norway (BN) rat[24] to compare the organization of the rat and the mouse Ly49 gene regions.

2 Results

2.1 Gene cloning
By *in silico* homology screening (BLAST) of rat BN BAC clones with the known rat and mouse Ly49 cDNAs as probes, 34 putative Ly49 loci were detected and hypothetical cDNA sequences constructed from predicted coding exons. Two different strategies were used for *in vitro* cDNA cloning: 1) cloning by RT-PCR on mRNA isolated from PVG or DA NK cells, with primers designed from the hypothetical genes identified *in silico*, yielding ten novel cDNAs. 2) *in vitro* homology screening of a PVG NK cell cDNA library with a mixture of previously cloned Ly49 genes as probe, yielding nine cDNAs. Eight of these were novel and one represented the PVG allele of the *Ly49s2* gene, previously cloned in the F344 strain [12]. Four of the cDNAs were obtained by both methods, giving a net result of 14 novel rat Ly49 cDNAs. Of these, 11 exhibited complete and three incomplete ORFs.

As a rule, the Ly49 genes have a long 5’-UTR intron 50-100 nucleotides (nt) upstream of the translation start codon, and many, but not all, have a short 3’-UTR intron immediately downstream of the stop codon (Fig.1). For the cDNAs cloned by RT-PCR, primers were therefore designed to lie close to the predicted translation start and after the predicted 3’-UTR intron. The larger part of the 5’ and 3’UTR sequences, including the putative first exon, is consequently not contained within these cDNAs, and the presence of a 5’-UTR unproven (Fig. 1, top). In order to simplify the presentation the first translated exon is here referred to as exon 2 for all the genes. A list of the cloned genes is given in table 1, along with the previously cloned and the deduced BN genes, in the order they occur in the BN genome in the centromeric to telomeric direction. The previously cloned *Ly49s3* (PVG) gene did not match any of the BN sequences, but was most similar to *Ly49s4*. 
2.2 Analysis of deduced amino acid sequences

The translated part of Ly49 genes consists of six exons. Exon 2 encodes the cytoplasmic domain, where inhibitory Ly49 receptors have an ITIM motif, characterized by the amino acid sequence VxYxxV/L (Fig. 2). This motif was found in 20 of the receptors (including the predicted BN sequences) (Table 1). Exon 3 encodes the transmembrane domain, where activating Ly49 receptors have an arginine (R) residue, found in 18 of the sequences. Five of the receptors exhibited both features. Exon 4 encodes the membrane proximal stalk, which probably function mainly as a spacer and in dimer formation. Exons 5 - 7 encode the ligand-binding lectin-like domain, characterized by evolutionary conserved amino acids with a typical succession of β-strands, α-helices and loops [25;26] (Fig. 2). Particularly conserved are the cysteines that form the two disulphide bridges between the α1-helix and the β5-strand (in the case of Ly49i8 possibly even three) and a single bridge between the beginning of the β3- to the end of the β4-strand. Like some other KLR, the Ly49 receptors have an additional disulphide-linked loop at the start of the lectin-like domain, formed by the β0- and the beginning of the β1-strand [27]. Eight of the predicted sequences may have a fifth intrachain disulphide bond, clamping the end of the β5 strand to the basis of the lectin-like domain just below the β0-loop. Ly49i8 exhibits the most deviant sequence, including 7 amino acids encoded by a short extra exon (exon 2b). In the rat genome assembly the extra exon is found approximately midway in the ~1350 nt sequence separating exons 2 and 3. A novel splice variant of Ly49s3 similarly contained a 39 bp/13aa sequence in this position (Fig. 2).

Inspection of the aligned sequences shows that the receptors fall into subgroups with shared substitutions compared with the consensus sequence. This is more clearly
visualized in Fig. 3, where most of the receptors are seen to cluster in three larger subgroups, each containing nearly equal numbers of inhibitory (i) and activating (s) members (in the case of group I, bifunctional receptors (si) substitute for activating). In addition, there is one group with two members, p8 and s2, and three with a single member only, all inhibitory (i13, i6 and i8). The dendrogram in Fig. 3 is based on the translated part of the genes only (amino acid sequences). The rat Ly49 genes can be divided into four major subgroups with respect to their 3’-ends (detailed in legend to Fig. 1). As indicated in Fig. 3, these coincide with the branching pattern of the dendrogram.

2.3 Physical map

Ly49 is by far the largest of the multigene subfamilies within the NKC, its 34 loci making up half of the KLR genes, spread across the distal ~1.75 Mb of this genetic region (Fig. 4a). The genes, all oriented the same transcriptional direction, are regularly spaced, with a median size of ~22 kb from exon 2 to exon 7 and median distance ~51 kb between the start codons of contiguous genes. For all the genes, intron 6 (separating the exons encoding the middle and last part of the lectin domain) is by far the largest, usually >5000 nt, in some cases even >10000 nt.

The order of loci deduced from a contig of overlapping BAC clones exhibited complete co-linearity with the physical map based on the genome assembly (Fig. 4b). However, the 325 kb BAC clone AC114442 lacked the Ly49si1 gene, expected to lie in the middle of its reported 279 kb contig, and contained only incomplete sequences for si3, i11 and p2. Furthermore, in the two BAC clones AC130159 and AC134637 flanking each side of AC114422, the hypothetical pseudogenes Ly49p2 and –p3 were positioned close together, as local duplicates. In the genome assembly, however, they were assigned
positions 190 kb apart. This region displays a repeated pattern with respect to deduced receptor function: si3 – i11 – p2 – si1 and si2 – i10 – p3 – si5, (Fig. 4), with si3 most closely related to si2, i11 to i10 and si1 to si5 (Fig. 3), and with p2 and p3 closest relatives when nucleotide sequences are compared (98 % sequence identity, not shown in Fig. 3), including sharing the same defect (missing exons 6 and 7 - table 1). The pattern is compatible with a chromosomal segment duplication involving four genes. Segmental duplication does not explain, however, the missing parts of AC114442 and the presence of both Ly49p2 and -p3 in AC130159 and in AC134637, which according to the genome assembly do not overlap (see Discussion).

The Ly49 gene region can be divided into three major subregions or blocks, each containing the genes belonging to the larger subfamilies described above (with s2, p5, p8 and p9 intermingled with the block II genes). The three single gene subfamilies fall outside the blocks, with Ly49i13 as the most proximal and Ly49i8 as the most distal gene, and Ly49i6 between blocks II and III. In block II sequence similarities between s4 - p7, i4 - i3, and p5 - p8 (Fig. 3) suggest segmental duplication involving three genes.

2.4 Allelic variation

Considerable allelic variation with respect to gene content and sequences has been reported for the mouse Ly49 gene region [28-31]. Although a high resolution map of this genetic region is yet only available in the rat for the BN strain, extensive allelic variation can be deduced by sequence alignments with cDNA data derived from other rat strains. Of the 21 cDNAs hitherto cloned from the PVG, DA or F344 strains, the majority were polymorphic at the DNA level when compared with their BN alleles (table 1). Seventeen seem to be functional in the strain from which it was cloned, as inferred from a complete
ORF and mRNA expression. Of these, six are predicted to be non-functional in the BN strain: one (Ly49s3) is missing from the BN genome and five have inactivating mutations (nonsense or frameshift – Ly49si3, -s5, -s4, -i2, -s7) (for -s5 confirmed by cDNA sequencing of the BN allele, for the others by the presence of the mutation in at least two independent genomic trace sequences, except for s4 where only a single trace sequence was found). According to the sequence of BAC clone AC125741 and the genome assembly, a seventh gene, Ly49i5, lacks the last coding exon. However, the untranslated 3’-UTR exon 8 is present in the clone AC119655, and a complete cDNA was obtained by RT-PCR on BN NK-cell mRNA (Dai, KZ – personal communication).

One of the previously cloned genes (Ly49p8) had multiple defects, found also in the deduced BN allele. Three cloned cDNAs were incomplete: Ly49s8(DA) Ly49s6(PVG) and Ly49i9(PVG) (Table 1). All three clones of Ly49s8(DA) shared the same defect, as did two independent clones of Ly49i9(PVG). RT-PCR of Ly49s6(PVG) gave only truncated products. The defects are therefore likely to be real. Both the BN and the PVG Ly49s6 are probably non-functional, but for different reasons (Table 1). As the causes of the splicing defects are unknown, faulty splicing due to cryptic splice sites remains a possibility also for the Ly49s8 and the Ly49i9 alleles. We have, however, chosen not to assign these genes pseudogene labels (see Discussion).

2.5 Phylogenetic analyses

As detailed above, whereas NK alloreactivity seems to be regulated primarily through inhibitory receptors in man and mouse, allospecific activating receptors seems to be equally important in the rat. With the mapping of the functional ‘alloreactivity’ gene Nka
to the distal part of the Ly49 region of the rat NKC [12], it would be of interest to compare this genetic region in the mouse and the rat. Furthermore, whereas rodents have many Ly49 genes, the species so far investigated in primates, carnivores (dog, cat) and artiodactyls (cow, pig) have only one [15;32], although perissodactyls (horse) have at least six [33]. Phylogenetic analyses of these mammalian orders showed that the primate Ly49 genes were more closely related to those from carnivores and artiodactyls, than those from the mouse [32], which contrasts with other genes, where molecular phylogenetics has placed primates closer to rodents [34;35]. The dendrogram in Fig. 5 shows that also the rat, together with the mouse genes, segregate from those of the other mammalian orders in a distinct clade. Equally striking is the lack of overlap between the mouse and the rat genes. Only rat Ly49i8 and mouse Ly49B form a clear-cut orthologous pair, with the most telomeric position in both species. Furthermore, the most centromeric mouse gene, Ly49Q, is more closely related to the rat genes (encoded by block I), than to other mouse genes, and vice versa for rat Ly49i6, which maps between blocks II and III. For all the other genes, the intraspecific similarities are greater than the interspecific, confirming the dynamics of this genetic region.

2.6 Recent duplications or gene homogenization?

The closer intraspecific sequence similarities for the majority of the genes could be due to duplications after the split between mouse and rat ~40 million years ago, or result from homogenization events, or both. Evidence for the latter mechanism emerges when the genes are compared exonwise (Fig. 6). When comparing exon 2, inhibitory receptors tend to cluster separately from activating and bifunctional. When comparing exon 5, however, smaller groups are formed, in most cases containing both inhibitory and activating
members (or bifunctional instead of activating members). Fig. 7 shows a closer view of eight selected receptors, four presumed activating and four inhibitory, exhibiting particularly distinct homogenization patterns. For the first two translated exons, the inhibitory receptors are more similar to each other than to any of the activating, in some case even exhibiting complete sequence identities. For the third exon (encoding the stalk) this pattern is broken, and for the last three translated exons (encoding the ligand-binding domain), activating and inhibitory receptors form closely related pairs, with complete sequence identities for some of the pairs. This pattern of ‘split homogenization’ could be explained by gene duplication followed by deletion of the chromosomal segment between e.g. introns 2 of neighbouring receptors of opposite signalling functions. However, this cannot explain the homogenization observed with other genes, such as Ly49s6 and Ly49i2, mapping far from each other. It should be noted, however, that the pattern is only seen between genes within one of the three major blocks.

3 Discussion

We here present evidence of 34 rat Ly49 loci in the BN genome, of which 15 seem to be pseudogenes in this strain. Some of these may be erroneously categorized as pseudogenes due to sequencing errors or incomplete sequences. However, the presence of multiple independent sequences from the genomic trace archive showing the same nonsense or frameshift mutations suggests that most are real, and for Ly49s5 the nonsense mutation has been confirmed by cDNA sequencing of the BN allele (Naper, manuscript in preparation). In the case of missing exons, the presence of multiple gaps in the BAC clone sequences and the genomic assembly make faulty categorization as pseudogenes more likely. This is evidently the case for Ly49i5, where the BAC clone used for sequencing (AC125741) exhibits a ~600 nt gap 8kb downstream of exon 6, which is the
expected region of this exon. Here, a complete cDNA was obtained by RT-PCR of BN NK cells (see note to table 1). The cDNA cloning of Ly49 genes from other inbred strains (PVG, DA, F344) furthermore reveals that many of the predicted pseudogenes in the BN rat probably are functional in the other strains. Ultimately, whether the genes are functional or pseudogenes must be determined at the protein level, which so far only has been done for Ly49i2 [36], -s3 [37], -i4, -i5 and -s5 (Naper, manuscript in preparation). By the less stringent criteria of complete ORFs and mRNA expression in the case of the cloned genes, or complete ORFs constructed from the genomic sequence in the case of the BN alleles, as many as 26 of the rat genes may be functional, a figure that may get higher when genes from other laboratory strains and wild rats are investigated. (Note that we have chosen not to label Ly49s6 a pseudogene, although the defect in the cloned PVG allele makes it likely to be non-functional.)

As for the number of genes, a caveat should be made for the apparent duplicated four-gene segment in block I, as it proved impossible to construct a consistent gene order for the three overlapping BAC clones AC130159, AC114442 and AC134637. Persistent heterozygosity could be an explanation, although the possibility seems remote on the background that the BN rat substrain selected as the sequencing target (BN/SsNHsd/Mcw) was subject to 13 additional generations of inbreeding to ensure homozygosity [24].

Of the 26 putative functional receptors, 13 are predicted to be inhibitory and 7 to be activating (the eighth activating receptor, Ly49s3, is not represented in the BN genome). There are in addition five predicted receptors with both ITIMs and an arginine in the TM domains. Such Ly49 variants have not previously been reported, and their signaling
properties are yet unknown. The clustering of inhibitory and activating receptors into
groups with closely related lectin-like domains (Figs. 6 and 7) (recently also reported by
Hao & Nei [38]), can be interpreted as evidence for functionally ‘opposing pairs’
recognizing the same ligand. As ‘bifunctional’ receptors substitute for activating in some
of these groups, a prediction would be that they at least transmit activating signals
(currently under investigation).

As for the genomic organization, the division of the genes into three major blocks
encoding related receptors is striking. To a certain extent it may reflect recent
duplications of individual genes, or as in the case of block I and possibly block II,
duplications of chromosomal segments bearing more than one gene. No doubt
expansions/contractions of the gene region through duplications/deletions as well as
homogenization by gene conversion contribute to rapid evolution of this gene region, as
reflected by the great interspecific sequence differences between the rat and the mouse
[32]. It is noteworthy that gene homogenization seems restricted to genes within the same
block, although they can be spaced far apart from each other within the block.
The close sequence similarities shared by members of the multigene families within the
NKC may not necessarily reflect recent duplications, but rather genes that have not yet
escaped homogenization. It is possible that the lack of homogenization observed between
the three Ly49 blocks is an expression of incipient escape, so that with time the three
blocks will drift apart into separate subfamilies, but inside each block retain their
sequence similarities due to continued homogenization.
The Ly49 gene region is flanked by Ly49i13 proximally and Ly49i8 distally, both falling outside the major blocks. The latter has the most deviant sequence of all the Ly49 genes. In particular is the last part of the lectin-domain different from other Ly49 genes, suggesting different ligand-binding properties, which may be linked to its chromosomal position outside the blocks, non-participation in gene homogenization and retention of sequence similarity with its murine ortholog, Ly49B. Identification of its ligand will clearly be highly informative. It should be added that the additional seven amino acids inserted between the cytoplasmic tail and the TM-region in Ly49i8, encoded by a separate exon (exon 2b), is not present in the orthologous mouse Ly49B receptor. It probably represents a splice variant rather than a splice artifact, as a splice variant with 39bp/13 aa, encoded by a similar exon 2b, also was observed for Ly49s3. Furthermore, also the rat Nkg2d (= Klrk1) gene contains a small (39 bp) intron in this position [39].

Another structural feature deserving comment is the high number of intrachain disulphide bonds in the lectin domains of Ly49 receptors. The CLSF receptors usually have two intrachain SS-bonds stabilizing the lectin domain. Like many other CLSF members, the Ly49 receptors have in addition a third at the base of the domain. Characteristic for the Ly49 family, however, is a second disulphide bond between the α1-helix and the β5-strand and many of the rat receptors are in addition predicted to possess a fifth SS-bond between the end of the β5-strand and the base of the lectin-like domain (Fig. 2). Soluble forms of the Ly49 receptor ligands, MHC class I, seem to abound in the body fluids [40]. In order to strip off bound soluble ligands, the Ly49 receptors may be subject to frequent acid rinses by recycling via early endosomes, which may be a reason why Ly49 receptors need the extra SS-bonds for added stability.
Comparisons with the cloned genes from the PVG, DA and F344 strains with the predicted genes from the BN strain bear evidence of extensive allelic polymorphism. The near complete segregation of the rat and mouse receptors in phylograms further underscores the rapid evolutionary change of this genetic region. Rapid evolution is also observed for two other immunoreceptor encoding genetic regions, the human leukocyte receptor complex (LRC), encoding KIR/LIR and related Ig-like receptors [41], and the MHC (as recently described for the rat in [42]). These three gene complexes have in common that they encode non-rearranging receptors. In order to keep pace with rapid microbial evolution, considerable plasticity seems required, with expansion/contraction through repeated unequal crossing-over events and gene conversion as central mechanisms. The importance of gene conversion in MHC evolution is debated, in particular whether it functions as an adaptive mutator under selection (the ‘strong’ model) [43]. In the case of the Ly49 region (and the LRC) we argue for a strong model. The Ly49 (and the KIR/LIR) receptors differ from the MHC in that they are signaling receptors with activating as well as inhibitory members. According to current orthodoxy in NK-cell target recognition, rooted in the strong position of the missing self-hypothesis [44], inhibitory Ly49 receptors are ascribed supremacy, with activating variants considered evolutionary afterthoughts. For example, with pairs of opposing Ly49 receptors recently described with near identical lectin domains and shared ligand specificities [45;46], the activating variants were proposed to have arisen from the inhibitory through gene conversion as a result of pathogen-driven selective pressure [47]. In the recently characterized bony fish NKC [48], however, the majority if not all cloned receptors lacked ITIMs and possessed an arginine in the TM region, predicting activating
function [49]. For functionally related receptors, like CD94/NKG2 and KIR/LIR and PIR, a recurrent finding is the existence of pairs of opposing receptors with shared ligand specificities. If, as argued above, the ‘bifunctional’ rat Ly49 receptors have activating properties, the rat possesses a near balanced repertoire of activating and inhibitory Ly49 receptors, forming smaller subgroups with opposing receptors with closely related lectin-like domains. On this background, a reasonably interpretation seems to be that the system is designed so that receptors of opposite signaling functions shall recognize the same ligands. Gene conversion then seems to represent a particularly suitable mechanism to ensure that the opposing receptors keep pace with each other in chasing the rapidly evolving ligands. Why we should have a system of pairs of opposing receptors is still a puzzle. It may e.g. make the system more robust against pathogen exploitation or allow for more accurate estimation of ligand densities. A corollary is that the activating Ly49 variants have a more central role than previously thought, compatible with observations made long ago in the rat that target allospecificities can trigger allorejection involving Ly49 receptors [12;21-23].

PVG NK cells recognize and rapidly eliminate allogeneic bone marrow derived target cells, discriminating between BN, AO and DA targets [50]. In contrast, although DA NK cells efficiently lyse tumor cell lines, they are inert towards allogeneic targets. This genetic difference was the basis for defining the alloreactivity gene, Nka, which was mapped to the distal part of the Ly49 gene region [12]. As the associated structural gene seemed to activate alloreactivity [12], most likely by encoding an activating receptor, it is noteworthy that the (purely) activating receptors all map to the distal part of the Ly49 region and exhibit particularly high allelic polymorphism. Of eight putative activating
receptors, only Ly49s1 and –s2 seem to be functional in both the strains investigated so far. Ly49s3, which is an important activating alloreceptor in the PVG strain, has no counterpart in the BN genome. Ly49s4, -s5, -s6 and -s7 have inactivating mutations in the BN strain, and the DA Ly49s8 is subject to faulty splicing. The activating receptors thus seem particularly vulnerable to inactivation by mutation, at least in inbred strains kept under laboratory conditions. The clustering of receptors with opposite signalling functions together in subsets with similar lectin domains (as argued above, suggesting shared ligands), probably constitutes the physical substrate underlying the capacity of PVG NK cells to discriminate between different MHC encoded allospecificities. An intriguing possibility is that the structural gene associated with Nka is not a single gene, but several of the genes encoding activating Ly49 receptors, each contributing to the general loss of alloreactivity observed in DA NK cells. A prediction would be that DA NK cells express few if any functional activating Ly49 receptors. We have previously shown that DA NK cells do indeed have very low expression levels of Ly49 genes [12], possibly reflecting a limited repertoire of expressed genes. On this background, it is noteworthy that the only activating DA Ly49 gene cloned seems to be non-functional.

In conclusion, our findings demonstrate a highly dynamic genetic region, containing a large number of closely related immunoreceptors, with the highest number of Ly49 genes in any species reported so far. However, rather than considering this an anomaly, the high gene content of the rat Ly49 gene region seems to offer a rich source for further explorations into the properties and evolution of a gene complex encoding non-rearranging immunoreceptors. The complex seems designed not only to meet the major
challenge of chasing rapidly evolving ligands, but to do so in step for pairs of receptors of opposing signalling functions.
4 Methods

4.1 Ly49 Homology screening

The cDNA library was from a subset of PVG NK cells (KLRH1+) and has previously been described, except that primary cDNA was ligated into the ZAP Express vector instead of into pMET7 [51]. Complexity of the primary library comprised more than 2x10^6 clones. 500,000 pfu of primary cDNA library were plated out on 160 mm NZY agar plates at a density of 50,000 pfu/plate. Replica nitrocellulose filters were lifted from each plate, denatured according to the manufacturer’s instructions (Stratagene), rinsed in 0.2M Tris-HCl (pH 7.5) and 2x SSC buffer, and blocked in hybridization solution (6x SSC, 5x Denhardt’s, 0.5% SDS, 20 µg/ml sonicated salmon sperm DNA, 50% Formamide). Hybridization was performed with a combination of [32P]-dCTP labelled cDNAs from rat Ly49i2, rat Ly49s1 (Ly49.29), rat Ly49s2 (Ly49.12) and mouse Ly49H overnight at 42°C. The filters were then washed twice with 2x SSC, 0.1% SDS and twice with 0.5 SSC, 0.1% SDS and films were developed after 3 days of exposure at 80°C. The plaque lifts were oriented and crosshybridizing plaques were purified to homogeneity. Clonal phage stocks were processed into PBK-CMV phagemids by in vivo excision according to the manufacturer’s instructions (Stratagene). Both strands of cloned cDNAs were sequenced in PBK-CMV by automated fluorescent sequencing at the UCSF Cancer Center molecular biology core facility. In this way eight new genes were obtained: Ly49s4, -s6, -s7, -i4, -i5, -i6, -i8 and -i9, plus the PVG allele of Ly49s2 previously cloned from the F344 strain [12]. In an alternative expression-based cloning strategy a ninth gene was also isolated, Ly49s5 (Naper, manuscript in preparation).

4.2 Identification of Ly49 genes from BAC clones and the genomic BN sequence
New Ly49 exons were identified by sequence similarity search (using the NCBI BLAST program) of BAC clones (listed in Fig. 4) and the whole genome sequence (www.ensembl.org/ Rattus norvegicus) against previously sequenced Ly49 cDNA sequences. Exon-intron splice junctions were deduced (following consensus rules for splice junctions and for intron phases of group V C-type lectins [39]), contiguous exons joined and the resulting sequences translated and compared with previously known proteins. Sequences containing all six coding exons giving complete open reading frames were defined as potentially functional genes. Sequences with nonsense or frameshift mutations or lacking exons were considered to be pseudogenes.

4.3 Cloning by RT-PCR

Primers were designed in the putative 5’ and 3’ untranslated regions (UTR) of the potentially functional genes and used for RT-PCR on mRNA isolated with Dynabeads mRNA Direct kit (Dynal Biotech) from DA or PVG LAK cells (lymphokine activated killer cells, i.e. NK cells cultured for ~10 days in the presence of interleukin-2). First strand cDNA synthesis was carried out with M-MLV reverse transcriptase RNase H (Promega) using 1 µg total RNA in a 20 µl reaction volume. The PCR products were analyzed by gel electrophoresis. Bands of expected lengths were excised from the gels, amplified by cloning (TOPO TA cloning kit, Invitrogen) and the inserts sequenced from both ends (MediGenomix, Germany), with internal sequencing primers generated to get full sequences in both directions. Ten new genes were detected: Ly49si1, -si2, si3, -i3, -i7, -s8, -s7, -s4, -i5, -i4, of which the latter four also were obtained by homology cloning, as described above. Three independent clones of the first six were fully sequenced in both directions.
4.4 Gene mapping and biocomputing

In silico analyses of rat, mouse and human genomic sequences were performed by pairwise BLAST against BAC clones and by sequence comparisons (BLAST) followed by visual inspection of the annotated rat genome (the analyses continually updated against the latest version, currently v. 21.3b.1, 8 February 2004). Biocomputing, including sequence analysis, was performed using software supplied by the Norwegian EMBNet node at the Biotechnology Centre in Oslo, Norway. The presented dendrograms were generated using the pileup program in the GCG package version 8.0. Alignments were also generated by using the program CLUSTAL X [52], and analyzed with the neighbor-joining method [53]. Bootstrap values were computed with 1000 iterations. As the two methods yielded the same major branches, only the figures from the GCG package are presented here.
Acknowledgements

The expert technical assistance of Ms Wendi Jensen and Marianne Laurtizen and the generous financial support by The Norwegian Cancer Society, The Norwegian Research Council, Bergljot and Sigurd Skaugen’s fund and The Wellcome Trust is gratefully acknowledged. Dominique Gauguier holds a Wellcome Senior Fellowship in Basic Biomedical Science (057733).
Figure legends

**Fig. 1**

Intron-exon structure of rat Ly49 genes. Boxes indicate exons, numbered as indicated below the boxes, with white showing translated and gray untranslated parts. Translated parts of the protein shown below the figure: cyt - cytoplasmic domain, TM - transmembrane part, stalk - membrane proximal stalk, pLD, mLD and dLD - proximal, middle and distal part of lectin domain. ATG - start codon. Stop - stop codon. The figures following the stop indicate length from stop codon to 5'-UTR intron and typical length of intron. The four 3'-UTR patterns shown are shared by multiple genes with gene name of typical member shown below type designation. Although positions and length may vary slightly, the sequences of the 5'-UTR introns are highly conserved within each group (generally > 90 % identity). The patterns concur with chromosomal positions (as divided into segments or blocks - see fig 4), as follows: Type 1 genes: all block I encoded genes (top). Type 2 genes: four genes within block II (s2, p5, p8 and p9). Type 3 genes: All the genes within block III. Type 4 genes: The remaining genes in block II and Ly49i8. The most centromeric gene, Ly49i13, has a longer intron, but conserved the flanking parts of the adjacent type I genes. Ly49i6, situated between blocks II and III, has an 227 nt intron 10 nt downstream of the stop codon. For the genes not cloned the introns were deduced from the rat genome sequence.

**Fig. 2**

Alignment of predicted amino acid sequences for all cloned rat receptors, except Ly49s5(PVG) (Naper, in preparation), and the two incomplete cDNAs Ly49s8(DA) and Ly49i9(PVG). The less informative sequence of the long stalk region has for practical
reasons not been included. Con - consensus sequence. TM region underlined. Predicted N-glycosidation sites boxed. Numbers on top of exons 5 - 7 indicate cysteine pairs involved in intrachain disulphide bonding. Bottom line: conserved secondary structures (α-helices, β-strands and loops) of the C-type lectin domain [25;26]. Exon 2: The residues forming the ITIM motif (VxYxxV/L) of predicted inhibitory receptors are shown in bold on the top line and in the consensus sequence. In the receptor sequences, amino acids replacing these residues are shown in bold. Exon 3: Similarly shown in bold in the sequences are the residues departing from the arginine (R) in the TM region, characteristic of activating receptors. The nomenclature is based on the principle introduced by Naper et al [36;37]. (Predicted) inhibitory receptors are called Ly49in (i - inhibitory, n - number assigned in the order the genes were first described); stimulatory receptors Ly49sn; ‘bifunctional’ Ly49sin; pseudogenes Ly49pn (numbered according to chromosomal position). The letters refers to the strain from which the gene was cloned: F - F344, P - PVG and D - DA. Accession numbers: Ly49i3(DA) - AY659932, Ly49i4(PVG) - AY649834, Ly49i5(PVG) - AY653729, Ly49i6(PVG) - AY651018, Ly49i7(DA) - AY659934, Ly49i8(PVG) - AY649835, Ly49i9(PVG) - AY649839, Ly49s2(PVG) - AY649838, Ly49s4(PVG) - AY649831, Ly49s6 (36) - AY649837, Ly49s7 - AY649833(PVG), Ly49si1(DA) - AY659935, Ly49si2(DA) - AY659933, Ly49s3(PVG) splice variant - AY747628, Ly49si3(PVG) - AY653730. The protein coding part of additional sequences predicted from the BN rat genome to encode functional genes (rLy49i9, i10, i11, i12, i13, s5, s8, si4 and si4) can be downloaded from http://www.med.uio.no/imb/anatomi/immunobiolab/fossum/index.html

Fig. 3
Phylogenetic tree (dendrogram) based on deduced amino acid sequences of presumed functional receptors (bootstrap values, 1000 iterations) in per cent shown for the main branches only). Three of the clusters contain multiple genes. In addition there is one with two genes (\textit{p8} and \textit{s2}) and three single gene families (\textit{i13}, \textit{i6} and \textit{i8}). The groups coincide with gene localization, defining three chromosomal blocks, denoted I, II and III (see physical map). Each block contains inhibitory (\textit{i}) and activating (\textit{s}) or bifunctional (\textit{si}) members. Following the submission of our manuscript, two papers were published on \textit{in silico} deduced Ly49 sequences from the rat genome \cite{38,54}, with predicted receptors named in consecutive order from the telomeric end, permitting gene identification \cite{38}. The predicted phylogenetic tree in \cite{38} is highly similar to the one presented here, with the reservation that \textit{i13} was not reported and \textit{p1} - \textit{p4} and \textit{i5} were omitted (\textit{i5} classified as a pseudogene), whereas the predicted pseudogenes \textit{p1} - \textit{p7} are omitted here.

\textbf{Fig. 4}

A. Physical map of the Ly49 region, with position of loci as deduced from the rat genome sequence (Ensembl release 19.3b.2 - February 2004), but with naming of the genes a composite of results derived from four different strains, with presumed functional genes in the PVG, DA or F344 given priority over apparent pseudogenes in the BN strain (cf. table 1). Arrows indicate transcriptional orientation with lengths corresponding to 20 kb. Vertical boxes represent individual exons. Numbers below boxes indicate genomic position of the first coding exon - exon 2 (to be added to the number on the left - e.g. Ly49i13 at position 167 Mb + 425 kb = 167 425 kb from the centromere). The related genes clustering in the three major groups in the phylogram (Fig. 3) are here shown to map together, defining three major blocks, named I, II and III, with \textit{Ly49i13}, \textit{-i6} and \textit{-i8}
falling outside the blocks. Deduced functional properties according to the presence or absence of ITIMs or charged amino acid (arginine) in the TM domains indicated by -: inhibitory, +: activating, +/-: bifunctional, p: pseudogene. Gene names as explained in Fig. 2, letter in parenthesis rat strain from which the gene was cloned: F- F344, P - PVG, D - DA, *(B) - not yet cloned, sequence in silico deduced from the BN rat genome sequence. B. Physical map of rat Ly49 genes as deduced from overlapping BAC clones. The line segments indicate the BAC clones. ● indicates the presence of the gene in the BAC clone, semicircle that only some of the exons are present.

Fig. 5

Dendrogram of all Ly49 molecules (amino acid sequence of the whole proteins) from rat (r), mouse (m), human (hs), dog (cf), cat (fc), baboon (ph), pig (ss), cow (bt) and horse (ec). Only mLy49B and rLy49i8 bear clear-cut evidence of orthologous relationship, underscored by the encoding genes mapping to the same localization at the distal end of the gene clusters in the two species. The majority of the rat and mouse Ly49 receptors cluster in separate branches (i.e. exhibit greater intraspecific than interspecific sequence similarities) indicating post-speciation subfamily expansions by gene duplications or gene homogenization by gene conversion. The exceptions are mLy49Q and rLy49i6. Bootstrap values for main branches only (in percent after 1000 iterations). The almost complete separation of the mouse and rat receptors, with the exceptions noted here, were also reported in [38;54].

Fig. 6
Dendrograms of exon 2 and exon 5 (deduced amino acid sequences) of the rat Ly49 genes. The peptide sequences encoded by exons 2 and 3 are involved in signal transduction, by exons 5 - 7 in ligand binding. Only bootstrap values > 80% shown.

**Fig. 7**

Dendrograms based on exonwise comparisons on the six translated exons of four selected inhibitory and four activating receptors.
Table 1. Overview of rat Ly49 receptors.

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Strain cloned</th>
<th>Identity to BN *</th>
<th>Signaling motif</th>
<th>Predicted function</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>i13</td>
<td></td>
<td></td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>p1</td>
<td></td>
<td></td>
<td>-</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>i12</td>
<td></td>
<td></td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Si3</td>
<td>PVG</td>
<td>99.4 %</td>
<td>ITIM/TM-R</td>
<td>bifunctional</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>i11</td>
<td></td>
<td></td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>p2</td>
<td></td>
<td></td>
<td>ITIM</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Si1</td>
<td>DA</td>
<td>99.7 %</td>
<td>ITIM/TM-R</td>
<td>bifunctional</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Si4</td>
<td></td>
<td></td>
<td>ITIM/TM-R</td>
<td>bifunctional</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Si2</td>
<td>DA</td>
<td>99.9 %</td>
<td>ITIM/TM-R</td>
<td>bifunctional</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>i10</td>
<td></td>
<td></td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>p3</td>
<td></td>
<td></td>
<td>ITIM</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Si5</td>
<td></td>
<td></td>
<td>ITIM/TM-R</td>
<td>bifunctional</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>i9</td>
<td>PVG</td>
<td>98.6 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>p4</td>
<td></td>
<td></td>
<td>TM-R</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>s5</td>
<td>PVG**</td>
<td>97.4 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>i5</td>
<td>PVG</td>
<td>99.7 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>s2</td>
<td>(.12) F344/ PVG</td>
<td>99.9 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>s6</td>
<td>PVG</td>
<td>99.8 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>s4</td>
<td>PVG</td>
<td>99.9 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>i4</td>
<td>PVG</td>
<td>96.2 %</td>
<td>TM-R</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>p5</td>
<td></td>
<td></td>
<td>TM-R</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>p6</td>
<td></td>
<td></td>
<td>TM-R</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>p7</td>
<td></td>
<td></td>
<td>TM-R</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>i3</td>
<td>DA</td>
<td>99.7 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>p8</td>
<td>(.19) F344</td>
<td>99.9 %</td>
<td>TM-R</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>i2</td>
<td>PVG</td>
<td>99.3 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>p9</td>
<td></td>
<td></td>
<td>TM-R</td>
<td>non-functional</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>i6</td>
<td>PVG</td>
<td>99.1 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>s8</td>
<td>DA</td>
<td>100 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>s7</td>
<td>PVG</td>
<td>99.4 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>s1</td>
<td>(.29) F344</td>
<td>99.5 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>i1</td>
<td>(.9) F344</td>
<td>100 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>i7</td>
<td>DA</td>
<td>100 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>i8</td>
<td>PVG</td>
<td>100 %</td>
<td>ITIM</td>
<td>inhibitory</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>s3</td>
<td>PVG</td>
<td>100 %</td>
<td>TM-R</td>
<td>activating</td>
<td></td>
</tr>
</tbody>
</table>

* Identities based on nucleotide sequences of coding parts, disregarding missing parts. **Ly49s5(PVG), obtained by expression cloning, has by functional studies been confirmed to encode a stimulatory receptor (Naper et al. In preparation). *** Full length clone with all six coding exons has been obtained by RT-PCR on mRNA from BN rat NK cells (Dai, KZ – personal communication).
Fig. 1
Fig. 5
Fig. 6
Fig. 7
References


