Chapter 1

# General Introduction

## General introduction

The study of genomes and genetic defects that underlie complex diseases is an active research area that raises much public interest. The complete DNA sequence of the human genome was published in 2003 and this milestone led to many expectations. With the knowledge of the human genetic sequence it is anticipated that we will gain better understanding of the molecular mechanisms behind development and disease and that we will be able to advance medicine by discovering new targets for drugs to combat today's incurable diseases like many types of cancer. Since the sequencing of the human genome, numerous other genome projects were started and today we have sequence information and map data of over 1000 organisms (NCBI database). However, the function of approximately half of all genes in human and other vertebrate species is still unknown. Therefore, the functional annotation of genome sequences is now the major challenge. Functional annotation of the human genome can be supported by loss-of-function (knock-out or knock-down) studies in model organisms and by the analysis of mutations (polymorphisms) that associate with human disorders.

This thesis is focused on the investigation of the *Mpzl3*, a novel gene with unknown function that was identified through mapping of the rough coat mutation in mice. The *Mpzl3* gene is a member of the myelin protein zero family that consists of proteins containing immunoglobulin domains and with suggested roles related to immune function and cell adhesion. Functional studies of the *Mpzl3* gene in two model organisms, mouse and zebrafish, and investigations of the human orthologue by *in silico* techniques, suggested that the product of this gene plays a role in the immune system and is a potential candidate gene for immune-related hereditary hair loss diseases in human.

## The rough coat mutation

The rough coat (*rc*) is a spontaneous recessive mutation in the inbred C57BL/6J mouse strain. The mutation was first observed at the Jackson laboratory in 1966 (Dickie, 1966). Homozygous *rc* mice (*rc*/*rc*) are indistinguishable from there littermates in the first two weeks, but from that time they start to develop a complex phenotype with several characteristic features. The most prominent peculiarity that the *rc*/*rc* mice develop is a cyclic and progressive hair loss starting from the first telogen phase in the hair cycle. In addition, they frequently develop spontaneous ulcerated wounds in the ventral region of the neck (Cao et al., 2007; Hayashi et al., 2004; Racz et al., 2009) (Figure 1). Histopathological examination of the homozygous mutant mice showed reduced amounts of extracellular matrix components, enlarged sebaceous glands in the skin and follicular atrophy in the ulcerated areas (Hayashi et al., 2004). The melanocyte pigments in the hair follicles change in color from black to light brown. Locally, extensive granulated tissue formations were noted with neu-

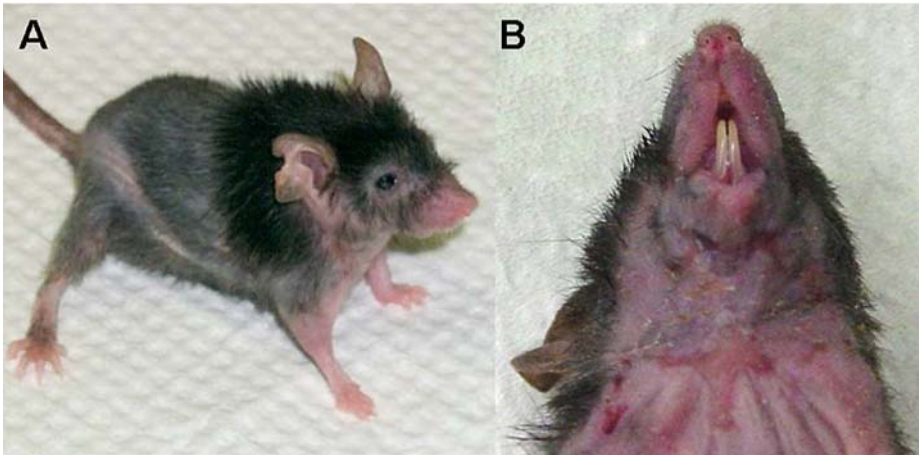


**Figure 1** The "rough coat" phenotype in the C57BL/6J mouse strain. **(A)** Hair loss phenotype of an adult *rc/rc* mouse, **(B)** Ulcerated skin wounds in the neck of rc/*rc* mouse

trophilic, mastocytic and lymphoplasmacytic dermatitis (Hayashi et al., 2004). In addition to the diversified skin phenotypes, the adult *rc/rc* mice also show growth retardation, abnormal bone structure and lower body weight than controls, with differences increasing progressively with age (Hayashi et al., 2004). In the femur of the *rc/rc* mice, lack of $Ca^{2+}$-derived basophilic material was observed, which might correlate with the elevated level of calcium in the blood (Hayashi et al., 2004). Histopathological study of different organs pointed to several immune system malfunctions. For example, in the hepatic sinusoids numerous erythrocytes, lymphocytes, polymorphonuclear leukocytes and enlarged Kupffer cells were detected, and in the trabeculae of the spleen, numerous macrophages were observed loaded with hemosiderin granules (Hayashi et al., 2004). The heart is also affected by this mutation. The cardiac muscle fibers are disoriented, and multifocal myocardial degeneration is also observed. In addition, the pups from homozygotes parents have lower survival rates compared with the litter from heterozygous parents. This phenomenon is probably due to abnormal maternality behavior of the *rc*/rc female mice (Hayashi et al., 2004).

By positional cloning the mutation in the rough coat mice was mapped to a 246-kb interval on chromosome 9 (Cao et al., 2007; Chapter 2). A missense mutation in this area was identified within a novel open reading frame, predicted to encode a protein with a conserved immunoglobulin-like V-type domain. The predicted protein showed strong homology to myelin protein zero (MPZ) and myelin protein zero-like 2 (MPZL2, also called epithelial V-like antigen) and was therefore named MPZL3 (myelin protein zero-like 3) (Cao et al., 2007; Chapter 2). The mutation in the *rc/rc* mice occurred at a highly conserved arginine residue within the conserved immunoglobulin domain, thus likely altering the MPZL3 protein function.

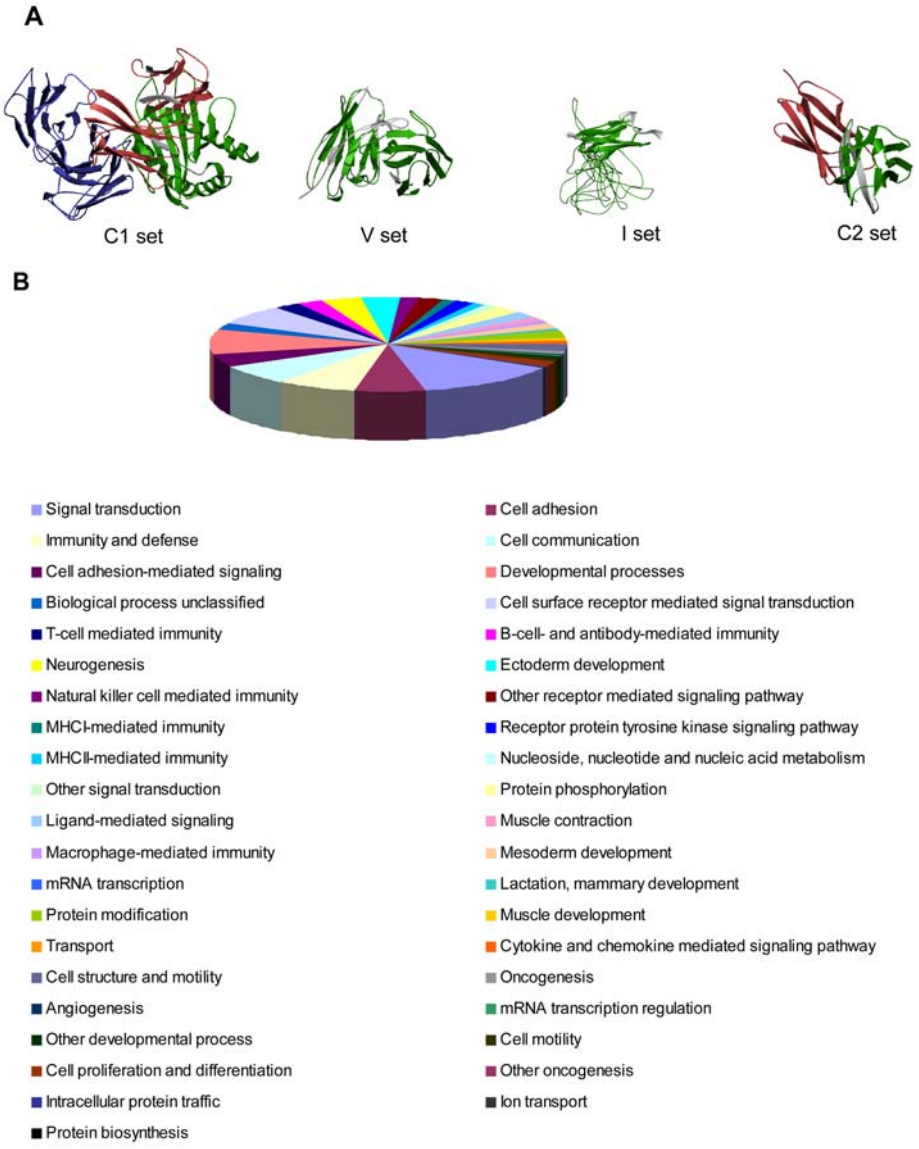


**Figure 2.** Structure and functions of IgSF proteins. **(A)** Structural models of proteins belonging to the different Ig sets. *C1-set*: complex of the human MHC class I glycoprotein hla-a2 and the T-cell coreceptor cd8 (PDB entry 1akj); *V-set*: fv fragment of mouse monoclonal antibody d1.3 (balb/c, igg1, k) variant chain l glu81->asp and chain H LEU312->VAL (PDB entry 1a7n) ; *I set*: central domain of cardiac myosin binding protein C (PDB entry 1gxe); *C2 set*: structures of an HIV and MHC binding fragment from human CD4 as refined in two crystal lattices (PDB entry 1cdh). Structures were imported from the Pfam database. **(B)** Gene ontology (GO) classification of the Ig domain containing proteins. The dataset was downloaded using BioMart service based on the Ensembl 56 database,

*Homo sapiens* genes dataset. GO classification was performed using Panther (Protein ANalysis THrough Evolutionary Relationships).

## The Immunoglobulin domain superfamily

Immunoglobulin domain containing proteins belong to one of the largest and most diverse protein superfamilies based on domain classification. The immunoglobulin domain appeared more than 500 million years ago and the structure can be recognized already in primitive organisms like members of the poriferan phylum, where it is found in cell-surface-receptor proteins (Buljan and Bateman, 2009). The domain structure was first identified as a sequence similarity of approximately 100 amino acids, repeated in antibody proteins. The number of proteins in the immunoglobulin superfamily (IgSF) extended with the evolution from invertebrates to vertebrates, in parallel with the evolution of the adaptive immune system (Barclay, 2003). Later it was discovered that immunoglobulin-like sequences are not unique only for antibody proteins, but can be found in an wide variety of proteins that plays diverse roles, for example in the immune system, and in neural and muscular processes (Buljan and Bateman, 2009).

Immunoglobulin superfamily domains contain relatively few highly conserved residues and their amino acid sequences are highly diverged, therefore it is possible that most algorithms underestimate their frequency (Barclay, 2003). The basic structure of the domain was identified more than 30 years ago (Davies et al., 1971; Poljak et al., 1973). The IgSF domains comprise a β-sandwich structure with seven or more antiparallel strands in two sheets (Barclay, 2003; Hutchinson and Thornton, 1993; Richardson, 1981) (Figure 2A). The average length of a protein domain is approximately 120 amino acids (Buljan and Bateman, 2009). Based on the amino acid sequence, two different types of immunoglobulin (Ig) domains can be distinguished. The first is the constant domain, or C-domain (IgC), which is shorter (seven strands) and contains several characteristics invariable residues, and the second is the variable-domain or V-domain (IgV), raging from eight, to ten strands. When more proteins were sequenced and shown to be Ig-related, the structures of these proteins were reminiscent of the V-domains and were termed V-set. This term does not indicate sequence variability as in the antibodies, just sequence similarity to the overall V-set sequences (Barclay, 2003). Later many IgSF domains were found with sequence patterns more similar to V-domains but more similar in size to C-domains. These were called C2-set and the original C-domains were named C1-set (Barclay, 2003). Another important set variation that is distinguished among the Ig domains is called the Intermediate or I-set domain. The I-set domain has sequence features of the V-set, but also has some structural features that were previously found only in the constant domains (Harpaz and Chothia, 1994). The different Ig sets have varying numbers of strands in each of the β-sheets that form the sandwich (Smith and Xue,

Table 1. The Myelin P0 protein fingerprint containing proteins in the human genome

| Ensembl Gene ID | Chromosome Name | Gene Start (bp) | Gene End (bp) | Strand | PRINTS ID | Associated Gene Name |
|---|---|---|---|---|---|---|
| ENSG00000158887 | 1 | 161274525 | 161279762 | -1 | PR00213 | MPZ |
| ENSG00000197965 | 1 | 167690429 | 167761156 | 1 | PR00213 | MPZL1 |
| ENSG00000144847 | 3 | 118619404 | 118864915 | -1 | PR00213 | IGSF11 |
| ENSG00000182985 | 11 | 115044346 | 115375112 | -1 | PR00213 | CADM1 |
| ENSG00000149575 | 11 | 118036187 | 118047241 | -1 | PR00213 | SCN2B |
| ENSG00000160593 | 11 | 118064442 | 118095809 | -1 | PR00213 | AMICA1 |
| ENSG00000160588 | 11 | 118097409 | 118123035 | -1 | PR00213 | MPZL3 |
| ENSG00000149573 | 11 | 118124137 | 118135009 | -1 | PR00213 | MPZL2 |
| ENSG00000166257 | 11 | 123499897 | 123525315 | -1 | PR00213 | SCN3B |
| ENSG00000079385 | 19 | 43009500 | 43032661 | -1 | PR00213 | CEACAM1 |
| ENSG00000105767 | 19 | 44126522 | 44143991 | -1 | PR00213 | CADM4 |
| ENSG00000105711 | 19 | 35521534 | 35531352 | 1 | PR00213 | SCN1B |
| ENSG00000167633 | 19 | 55327923 | 55378662 | 1 | PR00213 | KIR3DL1 |
| ENSG00000243772 | 19 | 55235964 | 55279336 | 1 | PR00213 | KIR2DL3 |
| ENSG00000198910 | X | 153126969 | 153174677 | -1 | PR00213 | L1CAM |
| ENSG00000101842 | X | 107288201 | 107322327 | 1 | PR00213 | VSIG1 |

The list was harvested from the Ensembl release 56 Sept. 2009 database using BioMart service.

1997). Ig domains are often characterized by a conserved disulphide bond that links the two beta sheets, however this boundary is not essential for the structure of the domain, as an active antibody and many other immunoglobulin domain-contain proteins lacking this disulphide have been reported (Barclay, 2003).

The functions of the IgSF proteins vary as much as their sequences (Figure 2B). First it was suggested that the proteins with Ig domains play a role in the immune system, however later more and more proteins were discovered with different attributes. Generally the function of the Ig-like domains is to allow binding reactions, however the specific interactions mediated by these domains vary widely (Fraser et al., 2006). They can bind small molecules, hormones, or large protein complexes, like muscle proteins through homo- or heterophilic interactions (Halaby and Mornon, 1998). The binding sites on these domains are located on the surfaces of the sheets or in the loops that connect the strands (Fraser et al., 2006). IgSF domains are mostly found in association with other IgSF domains on membrane proteins. However, they can also be associated with other domain types such as fibronectin type III, C-type lectin complement control protein, cytokine receptor domains and rarely EGF domains (Barclay, 2003; Korhonen et al., 1992; Letunic et al., 2002). Investigations of the evolution of Ig domains point also to an association with kinase domains in poriferans, which might suggest that the ancestral function of the Ig is also related to signaling (Buljan and Bateman, 2009).

## Proteins with a Myelin Protein Zero Domain

The Myelin Protein Zero Family is a small protein family in the Protein

Table 2. GO term classification of the human Myelin P0 domain containing proteins

| Biological Process | # Genes | Genes |
|---|---|---|
| Cell adhesion | 7 | MPZ, IGSF11, MPZL1, L1CAM, CEACAM1, MPZL2, VSIG1 |
| Signal transduction | 5 | KIR3DL1, KIR2DL3, AMICA1, L1CAM, CEACAM1 |
| Cell communication | 5 | KIR3DL1, KIR2DL3, AMICA1, L1CAM, CEACAM1 |
| Cation transport | 3 | SCN1B, SCN2B, SCN3B |
| Ion transport | 3 | SCN1B, SCN2B, SCN3B |
| Transport | 3 | SCN1B, SCN2B, SCN3B |
| Cell adhesion-mediated signaling | 3 | AMICA1, L1CAM, CEACAM1 |
| Developmental processes | 2 | L1CAM, CEACAM1 |
| Synaptic transmission | 2 | SCN1B, SCN3B |
| Neuronal activities | 2 | SCN1B, SCN3B |
| Cell structure | 2 | MPZ, MPZL1 |
| Cell structure and motility | 2 | MPZ, MPZL1 |
| Ligand-mediated signaling | 2 | KIR3DL1, KIR2DL3 |
| Cell surface receptor mediated signal transduction | 1 | KIR3DL1 |
| Ectoderm development | 1 | L1CAM |
| Neurogenesis | 1 | L1CAM |
| Other developmental process | 1 | CEACAM1 |
| Natural killer cell mediated immunity | 1 | KIR3DL1 |
| Immunity and defense | 1 | KIR3DL1 |

GO classification was performed using the Panther (Protein ANalysis THrough Evolutionary Relationships) server. The number of Myelin Protein o domain related genes that are associated to the GO term is indicated, together with the gene symbols. Data are based on the Ensembl release 56 Sept 2009 database.

Knowledgebase (UniProtKB) database. The common features are that all members of this family have an Ig V-set domain and possess at least three elements out of the characteristic six-element fingerprint of the Myelin Po protein (S-[KR]-S-x-K-[AG]-x-[SA]-E-K-K-[STA]-K.) (Entry: PS00568). Currently, four proteins have been assigned to this family: Myelin Protein Zero (MPZ), Myelin protein Zero Like 1 (MPZL1), Myelin protein Zero Like 2 (MPZL2), and Myelin protein Zero Like 3 (MPZL3). However, when we investigated the Ensembl release 56 Sept. 2009 *Homo sapiens* database we found altogether 16 proteins that have the Myelin Po protein fingerprint (Table 1). Investigation of the chromosomal localization of the family shows that one gene is located on the chromosome 3, two genes are on chromosome 1 and on the X chromosome, and the rest of the genes are spread over chromosomes 11 and 19 (Table 1). The predicted functions of the members of the family, based on the Panther (Protein ANalysis THrough Evolutionary Relationships) classification system, are mainly linked to cell adhesion , but several of these proteins are also linked to cell signaling, immune system processes, and ectoderm development (Table 2.). The classification of proteins according to GO terms allow for fast analyses based on experts knowledge of the biological data (Stevens et al., 2000), but such automated *in silico* analysis obviously has limitations, since not all published data are reflected in GO term annotations. For example, MPZL2 (EVA1) is annotated as an adhesion mol-

ecule, but it was reported that MPZL2 also plays a role in T cell and thymus development (DeMonte et al., 2007). Similarly, AMICA1 is associated with the general GO term signal transduction, but deeper investigation shows that this signaling function might affect the transmigration of leukocytes through epithelial and endothelial tissue (Moog-Lutz et al., 2003), suggesting that this myelin Po protein might play a role in the immune system too.

## Outline of the thesis

The goal of this thesis is the functional characterization of a novel member of the myelin Po protein family, *Mpzl3*, which is mutated in mice with the rough coat phenotype. To gain understanding of the molecular mechanisms behind the complex rough coat phenotype, the defects caused by the mutation and the expression pattern of the affected *Mpzl3* gene were studied in detail. In addition, knock-down and overexpression studies of the zebrafish homolog of *Mpzl3* were performed.

**Chapter 1** gives a general introduction about the rough coat mutation, a short description of the phenotype, and introduces the protein family of the mutated gene.

**Chapter 2** describes the microsatellite marker based mapping strategy used to identify the rough coat mutation. It shows that the mutation is mapped to a conserved residue in the Ig domain of the *Mpzl3* gene. Furthermore, it describes the expression pattern of the *Mpzl3* gene in mouse tissues and the localization of the MPZL3 protein in skin sections.

**Chapter 3** presents the results of *in silico* and histopathological analyses of the *MPZL3* gene in human and points out that this gene is a potential candidate for immune-related hereditary hair loss.

**Chapter 4** describes the functional analysis of the Mpzl3 protein in zebrafish by means of overexpression and morpholino knock-down experiments in embryos. The knock-down effect was further investigated at the transcriptome level by microarray analysis.

**Chapter 5** summarizes and discusses the work of this study.