MOLECULAR AND CELL BIOLOGY OF INFANTILE (CLN1) AND VARIANT LATE INFANTILE (CLN5) NEURONAL CEROID LIPOFUSCINOSES

Juha Isosomppi

Department of Molecular Medicine,
National Public Health Institute, Helsinki, Finland
and
Department of Medical Genetics,
University of Helsinki
Finland

Academic Dissertation

To be publicly discussed with the permission of the Medical Faculty of the University of Helsinki in the large lecture hall of the Haartman Institute, Haartamaninkatu 3, Helsinki, on March 28th, 2003 at 12 o'clock noon.

Helsinki 2003

Supervised by

Professor Leena Peltonen-Palotie
National Public Health Institute
and University of Helsinki, Finland

Reviewed by

Professor Jorma Panula
Institute of Biomedicine/Anatomy
University of Helsinki
Finland

and

Docent Maija Wessman
Division of Genetics
Department of Biosciences
University of Helsinki, Finland

To be publicly discussed with

Professor Marja-Liisa Savontaus
Department of Medical Genetics
University of Turku, Finland

Publications of the National Public Health Institute
KTL A3/2003

Julkaisija – Utgivare – Publisher

Kansanterveyslaitos (KTL)
Mannerheimintie 166
00300 Helsinki
puh. vaihde 09-47441, telefax 09-47448408

Folkhälsoinstitutet
Mannerheimvägen 166
00300 Helsinki
tel. växel 09-47441, telefax 09-47448480

National Public Health Institute
Mannerheimintie 166
00300 Helsinki, Finland
phone +358-9-47441, telefax +358-9-47448408

ISBN 951-740-340-2
ISSN 0359-3584
ISBN 951-740-341-0 (pdf) http://ethesis.helsinki.fi
ISSN 1458-6290 (pdf)

Cosmoprint Oy, Helsinki 2003

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original articles, which are referred to in the text by their Roman numerals.

I. Laan, M., **Isosomppi, J**., Klockars, T., Peltonen, L., and Palotie, A. (1996) Utilization of FISH in positional cloning: an example on 13q22. *Genome Research* 6, 1002-1012

II. Klockars, T., **Isosomppi, J**., Laan, M., Kakko, N., Palotie, A., and Peltonen, L. (1997) The visual assignment of genes by fiber-fish: BTF3 protein homologue gene (BTF3) and a novel pseudogene of human RNA helicase A (DDX9P) on 13q22 *Genomics* 44, 355-357

III. **Isosomppi, J**., Vesa, J., Jalanko, A. and Peltonen, L. (2002) Lysosomal Localization of the Neuronal Ceroid Lipofuscinosis CLN5 Protein. Human Molecular Genetics 11 (8), 885-891.

IV. **Isosomppi, J**.*, Heinonen, O*., Hiltunen, J. O., Greene, N. D. E., Vesa, J., Uusitalo, A., Mitchison, H. M., Saarma, M., Jalanko, A., and Peltonen, L. (1999) Developmental expression of palmitoyl protein thioesterase in normal mice. *Brain Research Developmental Brain Research* 118, 1-11.

*These authors contributed equally to the respective article.

Publication I is found in the thesis of Maris Laan (1997) and publication II in theses by Tuomas Klockars (1998) and Maris Laan.

ABBREVIATIONS

ANCL	adult neuronal ceroid lipofuscinosis
AP	adaptor protein complex
BHK-21	Syrian golden hamster kidney cells
BLAST	basic local alignment tool
bp	base pair
cDNA	complementary deoxyribonucleic acid
CLN1	infantile neuronal ceroid lipofuscinosis locus
CLN2	classical late infantile neuronal ceroid lipofuscinosis locus
CLN3	juvenile neuronal ceroid lipofuscinosis locus
CLN4	adult neuronal ceroid lipofuscinosis locus
CLN5	variant late infantile neuronal ceroid lipofuscinosis locus, Finnish type
CLN6	variant late infantile neuronal ceroid lipofuscinosis locus
CLN7	variant late infantile neuronal ceroid lipofuscinosis locus, Turkish type
CLN8	Northern epilepsy locus
COS-1	African green monkey kidney cells
DNA	deoxyribonucleic acid
ER	endoplasmic reticulum
EST	expressed sequence tag
FISH	fluorescent *in situ* hybridization
GGA	Golgi localized, gamma-adaptin ear homologous, ADP-ribosylation factor binding protein
GROD	granular osmiophilic deposit
INCL	infantile neuronal ceroid lipofuscinosis
JNCL	juvenile neuronal ceroid lipofuscinosis
kb	kilobase
LINCL	late infantile neuronal ceroid lipofuscinosis
Mb	megabase pair(s)
Man-6-P	mannose 6-phosphate
M6R	mannose 6-phosphate receptor
mRNA	messenger ribonucleic acid
NCL	neuronal ceroid lipofuscinosis
ORF	open reading frame
PAC	P1 derived artificial chromosome
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PFGE	pulsed field gel electrophoresis
PPT1	palmitoyl protein thioesterase 1
RT	reverse transcription
SAP	sphingolipid activator protein
STS	sequence tagged site
TGN	trans Golgi network
UTR	untranslated region
vLINCL	variant form of late infantile neuronal ceroid lipofuscinosis
WT	wild-type
YAC	yeast artificial chromosome

ABSTRACT

Neuronal ceroid lipofuscinoses (NCL) are a group of common progressive recessively inherited neurodegenerative disorders of childhood. All types of NCL diseases cause progressive visual and mental decline, motor disturbances, epilepsy and behavioral changes, and lead to premature death. Prior to this study the first NCL gene was recently identified using the positional cloning approach. Mutations in the palmitoyl protein thioesterase (*PPT1*) were shown to result in the infantile form of NCL (INCL). At the same time, the positional cloning of the vLINCL disease gene (*CLN5*) was in progress. The position of the *CLN5* gene was assigned by linkage analysis to chromosomal region 13q21.1-q32 and physical mapping of the region was ongoing. Both of these diseases are especially enriched in the Finnish population.

In this thesis, fluorescent *in situ* hybridization on DNA fibers (fiber-FISH) was utilized in the physical mapping project of the critical CLN5 region. This visual mapping approach was essential in our efforts to produce a genomic clone contig over the CLN5 region. The fiber-FISH method not only enabled rapid confirmation of the order of genomic clones, but it also allowed the detection of overlaps between various clones. Thus, high-density mapping was possible without the tedious methods traditionally used in physical mapping approaches. In addition, a novel ultra-sensitive tyramide-based amplification system was used successfully to visualize short probes representing transcribed sequences in the critical region.

The physical map of the critical region facilitated the identification of the *CLN5* gene, and its expression was analyzed as a part of this thesis. The biosynthesis, post-translational processing and intracellular localization of the CLN5 protein was investigated in transiently transfected BHK-21 cells. Confocal immunofluorescence microscopy and immunoprecipitation analysis showed that wild type CLN5 is a lysosomally targeted 60-kDa glycoprotein, which is partially secreted into the culture medium. Secretion of the polypeptide into the culture medium would imply that CLN5 is a soluble lysosomal glycoprotein, not an integral transmembrane protein as predicted earlier. The most common naturally occurring CLN5 disease mutation represents a premature stop codon that leaves the 16 C-terminal amino acids of the protein untranslated. These polypeptides were not targeted to lysosomes, which would imply that the pathogenesis of

vLINCL might be associated with the defective lysosomal trafficking of the corresponding polypeptide.

In order to better understand the destruction of neurons in the central nervous system in the childhood forms of NCL-disorders, we characterized expression of the *PPT1* gene in developing mouse brain and embryo. Northern blot analysis, *in situ* hybridization and immunohistochemistry revealed gradual increase in expression of *PPT1* mRNA and protein during mouse development. A notable increase in *PPT1* mRNA expression was monitored during a developmental stage of the mouse brain when new synaptic contacts are extensively formed. In addition to that, a relatively high prevalence of PPT protein was observed in the neuronal extensions. Based on these findings, it was suggested that PPT1 might have a role for survival of neural networks, possibly associated with the development and maintenance of the synaptic machinery.

REVIEW OF THE LITERATURE

1. Neuronal ceroid lipofuscinoses

1.1 Introduction

1.1.1 Classification of neuronal ceroid lipofuscinoses

Neuronal ceroid lipofuscinoses (NCLs) are a group of neurodegenerative disorders that are linked by common clinical and pathological features (Goebel, 1995; Santavuori, 1988). The term NCL derives from the accumulation of ceroid- and lipofuscin-like storage cytosomes in various tissues (Zeman & Dyken, 1969). These diseases are characterized by progressive visual and mental decline, motor disturbances, epilepsy and behavioral changes, and ultimately they all lead to premature death. Traditionally, NCLs have been divided into four main types: infantile NCL (INCL; locus definition CLN1), classical late infantile NCL (LINCL; CLN2), juvenile NCL (JNCL; CLN3) and adult NCL (ANCL, CLN4). In addition to these four main types, several variant NCL subtypes have been defined (Table 1).

Table 1. Classification of NCL diseases

Locus	*Clinical type*	*Chromosomal location*	*Gene product*
CLN1	Infantile	1p32	Palmitoyl protein thioesterase
CLN2	Late infantile Classical	11p15	Pepstatin insensitive protease
CLN3	Juvenile	16p12	Membrane protein
CLN4	Adult, Kufs or Parrys disease	Not known	?
CLN5	Late infantile, Finnish variant	13q22	Lysosomal protein (III)
CLN6	Late infantile, Variant	15q21-q23	Membrane protein
CLN7	Late infantile, Turkish variant	8p32	?
CLN8	Northern epilepsy	8p32	Membrane protein
?	Congenital	?	?

(See chapter 1.1.2 for references)

In examinations with the electron microscope each classical NCL type has a characteristic ultrastructural appearance of storage material. The storage material forms granular osmiophilic deposits (GRODs) in the INCL, curvilinear pattern (CVP) in the LINCL and fingerprint profile (FPP) in JNCL (Santavuori, 1988). The ultrastructure of storage material in ANCL is mixed; being either a combination of FPP and CVP or GRODs (Berkovic *et al.*, 1988b; Martin *et al.*, 1987). Biochemical analyses have shown that the GROD bodies mostly consist of sphingolipid activator proteins A and D (Tyynelä *et al.*, 1993). In CVP and FPP inclusions the major accumulated material is mitochondrial ATP-synthase subunit c (Hall *et al.*, 1991).

1.1.2 Genetic and cell biological studies of NCL-disorders

To date, six NCLs with different gene locations have been recognized by linkage analysis: 1p32 for CLN1 (Järvelä *et al.*, 1991), 11p15 for CLN2 (Sharp *et al.*, 1997), 16q22 for CLN3 (Gardiner *et al.*, 1990), 13q22 for CLN5 (Savukoski *et al.*, 1994), 15q21-q23 for CLN6 (Sharp *et al.*, 1997) and 8p32 for CLN8 (Tahvanainen *et al.*, 1994). In 1999, a Turkish variant, LINCL (CLN7) was excluded from all known NCL loci, suggesting that it represents a novel genetic locus for LINCL (Wheeler *et al.*, 1999). Later on CLN7 was mapped onto the 8p32chromosomal region (Mitchell *et al.*, 2001), which is known to contain the *CLN8* gene responsible for the Finnish disease Northern epilepsy (EPMR) (Ranta *et al.*, 1999; Tahvanainen *et al.*, 1994). Thus, it is probable that CLN7 is allelic to CLN8. To date no locus has been identified for ANCL (Berkovic *et al.*, 1988a; Boehme *et al.*, 1971) or congenital NCL (Norman & Wood, 1941).

The defective genes behind the six human NCL diseases are known. Two of them encode soluble lysosomal enzymes. Palmitoyl protein thioesterase 1 (PPT1) is defective in CLN1 (Hellsten *et al.*, 1996; Verkruyse & Hofmann, 1996; Vesa *et al.*, 1995) and tripeptidyl peptidase (TPP1) in CLN2 (Sleat *et al.*, 1997). PPT1 removes palmitate groups from proteins *in vitro* (Camp & Hofmann, 1993; Camp *et al.*, 1994) and TPP1 cleaves tripeptides from the N-terminus of small peptides (Vines & Warburton, 1999). The *CLN3* gene was identified in 1995 (The International Batten Disease Consortium, 1995). This protein is an integral transmembrane protein, which may have a role in the regulation of the vacuolar pH (Pearce *et al.*, 1999). In addition to lysosomes (Järvelä *et al.*, 1998), several other intracellular localizations have been proposed for CLN3 (Katz *et al.*, 1997; Kremmidiotis *et al.*, 1999; Margraf *et al.*, 1999). In neurons the protein has been shown to be transported along the neuronal extensions and to be targeted to

neuronal synapses (Järvelä *et al.*, 1999; Luiro *et al.*, 2001). The *CLN8* gene encodes a membrane protein of unknown function (Ranta *et al.*, 1999). It has been shown to be transported between the ER and ER-Golgi intermediate compartment (Lonka *et al.*, 2000). The CLN6 gene was very recently cloned, and it is predicted to encode a novel transmembrane protein with unknown function (Gao *et al.*, 2002; Wheeler *et al.*, 2002). The sixth known NCL gene is *CLN5*. The predicted amino acid sequence of CLN5 shows no homology to previously reported proteins and its function remains to be determined (Savukoski *et al.*, 1998).

Currently there is no unifying hypothesis, which would explain the molecular and cellular basis of the NCLs. It is unclear, how mutations in different genes result in similar diseases. Based on the acid phosphatase activity and electron microscopic studies, the storage material in NCLs is associated with lysosomes (Rapola, 1993). The identification of mutations in lysosomal proteins (CLN1, CLN2, CLN3) also indicates that pathogenesis of the NCL disorders is somehow related to lysosomes.

NCL disorders have been comprehensively reviewed in several recent articles (Mitchison & Mole, 2001; Mole, 1998; Peltonen *et al.*, 2000; Weimer *et al.*, 2002) and special journal issues and books have been also published (The Neuronal Ceroid Lipofuscinosis, 1999; Proceedings of the 8th international congress on the neuronal ceroid lipofuscinoses, 2000). This thesis deals with CLN5 and CLN1 and they are described in more detail in the following sections.

1.2 Finnish variant late infantile neuronal ceroid lipofuscinosis (CLN5)

1.2.1 Clinical features

Finnish variant late infantile neuronal ceroid lipofuscinosis (vLINCL; MIM 256731) has its clinical onset at 2-7 years of age. The first symptom is motor clumsiness, followed by progressive visual failure, mental and motor deterioration and later by myoclonus and seizures. The ultrastructure of the storage material consists of curvilinear and fingerprint profiles. Subunit c of the mitochondrial ATP synthase is the major protein in vLINCL brain storage cytosomes. These cytosomes also contain minor amounts of sphingolipid activator proteins (SAPs) (Tyynelä *et al.*, 1997). The age at death varies from 14 to 36 years (Holmberg *et al.*, 2000; Santavuori *et al.*, 1991; Santavuori *et al.*, 1982). Cerebellar atrophy is

the most striking abnormality in brain imaging studies (Autti *et al.*, 1992) and in autopsy specimens (Tyynelä *et al.*, 1997).

1.2.2 *CLN5* gene and mutations

The *CLN5* gene was identified in 1998 using the positional cloning approach (Klockars *et al.*, 1996; Savukoski *et al.*, 1994; Savukoski *et al.*, 1998). The gene has four exons and it has an open reading frame (ORF) of 1380 bp. The predicted amino acid sequence of CLN5 shows no homology to previously reported proteins. Very little is known about the expression of the gene and the function of the protein is not known. Based on the results of Northern and dot blot hybridizations the gene is expressed in a wide variety of tissues (Savukoski *et al.*, 1998). *In situ* hybridization and immunohistochemical studies have demonstrated that *CLN5* is expressed at varying stages of corticogenesis in humans beginning at the early developmental stage and the expression level of *CLN5* increases during brain development (Heinonen *et al.*, 2000b).

To date, four disease mutations have been described, of which three result in premature termination of the polypeptide chain (Holmberg *et al.*, 2000; Savukoski *et al.*, 1998). The most common mutation among Finnish CLN5 patients is a two base pair deletion, del(AT)2467-2468 resulting in Tyr392Stop. Another disease mutation found among Finnish patients is G1517A leading to a very truncated polypeptide (Trp75Stop). The SWE mutation, ins(C)1961, was found in one Swedish and one Finnish CLN5 patient, both being compound heterozygotes for the mutation. The fourth CLN5 mutation, G2127A, was found in a Dutch family and results in an amino acid substitution of Asp279Asn. All the mutations seem to result in a similar clinical phenotype (Holmberg *et al.*, 2000).

1.3 Infantile neuronal ceroid lipofuscinosis (CLN1)

1.3.1 Clinical features

The most severe form of the NCLs is INCL. Early development of children with INCL is normal until the age of 8–14 months, when retardation of psychomotor development is first observed. All patients enter a terminal stage before the age of 3 and usually die between 6–15 years of age. The disorder leads to an extraordinary degree of brain atrophy. The cerebral cortex is almost completely

destroyed and the cerebellum is also extremely atrophic (Haltia *et al.*, 1973; Rapola, 1993; Santavuori *et al.*, 1974).

1.3.2 *PPT1* gene and mutations

The defective gene, *PPT1*, underlying INCL was isolated using a positional cloning strategy (Hellsten *et al.*, 1993; Järvelä *et al.*, 1991; Vesa *et al.*, 1995). The gene is composed of nine exons and it spans a 25 kb region in genomic DNA (Schriner *et al.*, 1996). To date, 39 disease causing mutations have been identified in PPT1 gene. All known mutations in the *PPT1* gene (and in other NCL-genes) are contained in the NCL mutation database (http://www.ucl.ac.uk/ncl/) (Mole *et al.*, 2001). Most of the PPT1 mutations cause severe an early onset INCL phenotype. However, certain mutations in PPT1 gene have been reported to produce phenotypes which are clinically indistinguishable from later onset NCLs; LINCL, JNCL and ANCL (Hofmann *et al.*, 1999; Mitchison *et al.*, 1998; Van Diggelen *et al.*, 2001).

1.3.3 PPT1 protein

PPT1 enzyme was originally purified from bovine brain cytosol (Camp & Hofmann, 1993). The function of PPT1 is to remove long-chain fatty acids (usually palmitate) from lipid-modified cysteine residues in fatty acylated proteins. Initially, lysosomal localization of PPT1 was considered unlikely, because of the neutral pH optimum of this enzyme (Camp *et al.*, 1994). In 1996, it was shown that PPT1 is one of the most abundant mannose-6-phosphorylated glycoproteins in the rat brain (Sleat *et al.*, 1996). As the mannose 6-phosphate modification is a hallmark of lysosomal enzyme trafficking (Kornfeld, 1990), PPT1 was suggested to be a lysosomal hydrolase (Sleat *et al.*, 1996). Later on, it was confirmed that PPT1 is targeted to lysosomes through the mannose 6-phosphate receptor pathway in transiently transfected COS-1 cells (Hellsten *et al.*, 1996; Verkruyse & Hofmann, 1996). Moreover, the lysosomal nature of the site of PPT1 function in the lymphoblastoid cells is clearly demonstrated - PPT1 has a role in the degradation of fatty-acylated proteins in the lysosomes (Lu *et al.*, 1996; Lu *et al.*, 2002). However, recent studies have suggested that in neurons PPT1 is localized in synaptosomes and synaptic vesicles rather than in the lysosomal compartment (Heinonen *et al.*, 2000a; Lehtovirta *et al.*, 2001). Neurons and their synapses are enriched in palmitoylated proteins, and due to its reversible nature, protein palmitoylation appears to have a crucial role in the functioning of the nervous system (Bizzozero *et al.*, 1994). Thus, it is speculated that in addition to

lysosomal protein degradation, PPT1 might also have a biological role outside lysosomes (Heinonen *et al.*, 2000a; Lehtovirta *et al.*, 2001; Suopanki *et al.*, 2002). Moreover, it has been shown that different substrates show different pH optima for PPT1, which further indicates a potential extralysosomal function for PPT1 (Cho *et al.*, 2000). However, the palmitate groups that modify proteins are normally found on the cytoplasmic face of the plasma membrane and based on our current knowledge, PPT1 is located on the luminal side of vesicles. Thus, it needs to be explained how PPT1 could act on cytoplasmic substrates.

The crystal structure of bovine PPT1 has been resolved (Bellizzi *et al.*, 2000). The model contains amino acids 28-306, which corresponds to the entire mature PPT1 polypeptide after cleavage of the 27-residue signal peptide. PPT1 has an α/β-hydrolase fold which is a characteristic structure of two previously determined thioesterases. The catalytic triad of PPT1 consists of serine 115, aspartic acid 233 and histidine 289. Most of the PPT1 mutations, which cause INCL and LINCL phenotypes are located close to the active site and palmitate binding pocket, or they disrupt the folding of the PPT1 protein. The mutations associated with later onset phenotype (JNCL) are located away from active site and are predicted to cause less dramatic changes to the structure of the PPT1. Some of the late onset mutations have been shown to retain some residual PPT activity, which further explains the milder phenotype of these patients (Das *et al.*, 2001; Hofmann *et al.*, 1999). The effects of different PPT1 mutations have also been studied in transient cell expression systems. While the wild type PPT1 is transported to lysosomes in nonneuronal cell lines, all the studied mutants are trapped in the ER and they do not show any detectable enzyme activity (Hellsten *et al.*, 1996; Salonen *et al.*, 2001; Vesa *et al.*, 1995). However, in infected mouse primary neuron cultures PPT1 polypeptides with severe mutations reside in the ER, whereas polypeptides with mild mutations migrate further in neurons (Salonen *et al.*, 2001).

Despite intense investigation of the PPT1 protein, its *in vivo* substrate is not known and pathogenesis of the INCL disorder remains to be resolved. Recently developed PPT1 knockout mouse model produce a characteristic NCL-like phenotype (Gupta *et al.*, 2001). Neurological abnormality is evident in 100% of PPT1-deficient mice by their eighth month and mice were deat before they were ten months old. Autofluorescent storage material, typical for INCL patients, was observed throughout the brains of PPT1 knockout mice. Thus, this mouse model provides a valuable tool to clarify the pathogenesis of the INCL disorder.

2. Targeting of lysosomal proteins

2.1 Soluble proteins

Lysosomes are acidic organelles in which endogenous and internalized macromolecules are degraded by lumenal hydrolases (Kornfeld & Mellman, 1989). The targeting of most of the soluble lysosomal hydrolases is dependent on the addition of mannose 6-phophate residues (Man-6-P) to their carbohydrates and recognition of this signal by receptors, which mediate the delivery of the proteins to lysosomes (Figure 1) (Kornfeld, 1990; Kornfeld & Mellman, 1989). The specificity of this Man-6-P pathway is determined by the Golgi-resident enzyme UDP-N-acetylglucosamine 1-phosphotransferase (phosphotransferase), which transfers N-acetylglucosamine-1-phosphate from UDP-N-acetylglycosamine to mannose residues of the high mannose-type oligosaccharide side chains of lysosomal enzymes. Phosphotransferase recognizes its substrates on the basis of a specific arrangement of lysine residues on the surface of lysosomal proteins (Cuozzo *et al.*, 1998; Tikkanen *et al.*, 1997). In a second reaction the N-acetylglucosamine is removed by another intra-Golgi enzyme (N-acetylglucosamine 1-phosphodiester α-N-acetylglucosaminidase) generating Man-6-P residue on the oligosaccharide side chains. In the late Golgi compartments, lysosomal enzymes bind to mannose 6-phophate receptors (MPRs). Two MPRs with overlapping functions have been identified to date. The first is a large (300 kDa) type I membrane glycoprotein that also binds insulin-like growth factor II. The second, a cation dependent MPR is a smaller type I transmembrane glycoprotein (45 kDa) (Le Borgne & Hoflack, 1998; Ludwig *et al.*, 1995). The receptor bound enzymes are packed into clathrin-coated transport vesicles that are targeted into the endosomal compartment. Collection of MPRs into clathrin-coated vesicles is directed by tyrosine- and dileucine-based motifs in their cytoplasmic domains. These motifs are recognized by the GGA (Golgi localized, gamma-adaptin ear homologous, ADP-ribosylation factor binding proteins) proteins (Doray *et al.*, 2002). The GGAs functions in the trans Golgi network (TGN) as adaptor proteins selecting cargo molecules for incorporation into AP-1 containing clathrin coated vesicles. In endosomes, the hydrolases dissociate from their receptors and subsequently reach lysosomes. The endosomal MPRs are recognized by a 47 kDa protein (TIP47), which facilitates collection of MPRs into transport vesicles destined to go back to the Golgi complex (Diaz & Pfeffer, 1998).

Although Man-6-P -dependent targeting is the most common pathway for the transport of soluble lysosomal enzymes, some of them are transported to lysosomes independently of the Man-6-P sorting signal (Glickman & Kornfeld, 1993). One of these targeting mechanisms involves membrane associatio, as is demonstrated for prosaposin, cathepsin D and β-glucosylceramidase (Rijnboutt *et al.*, 1991).

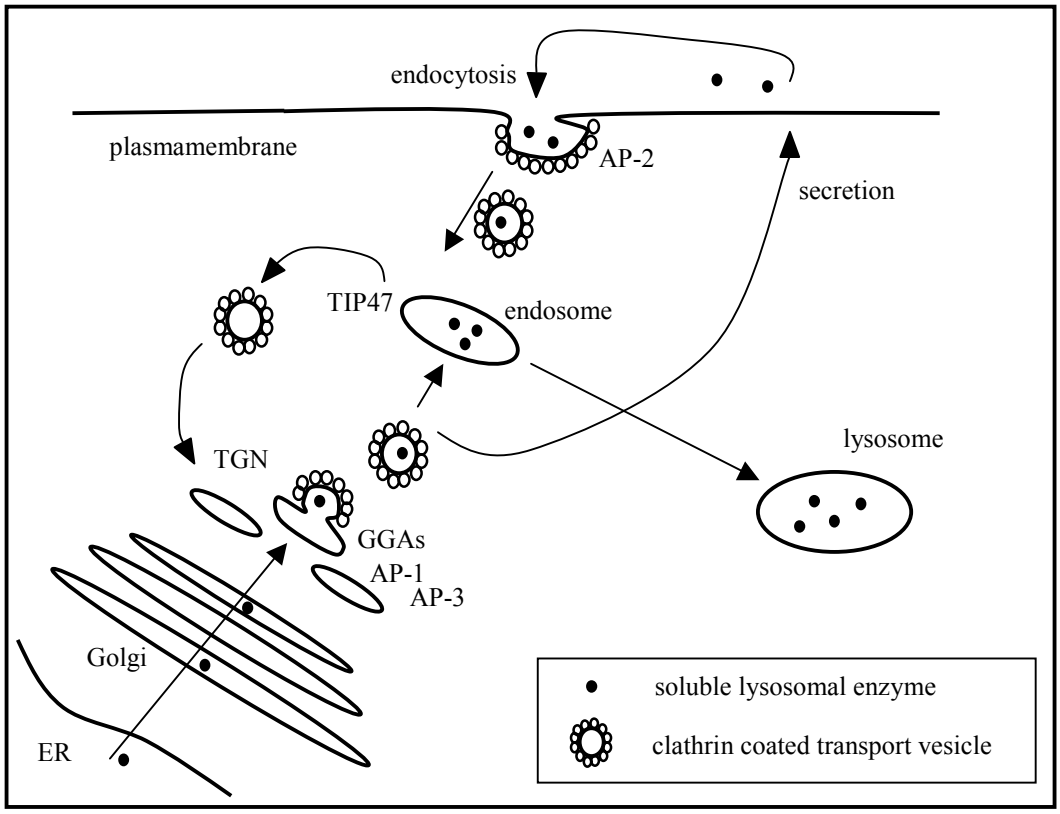

Figure 1. Targeting of lysosomal proteins (Modified from Rouille et al 2000). Soluble lysosomal enzymes are sorted into lysosomes at the trans-Golgi network (TGN) by the mannose 6-phosphate receptors (MPRs). Lysosomal hydrolases can also be secreted outside of the cell, and endocytosed back to the cell from the plasmamembrane. Also lysosomal transmembrane proteins are sorted to the lysosomes in the TGN. Membrane traffic requires the formation of clathrin coated transport intermediates. Different kinds of adaptor proteins (APs and GGAs) have important functions in the cargo inclusion into the transport vesicles. Recycling of the MPRs to the TGN from late endosomes requires TIP47 protein.

2.2 Membrane proteins

Lysosomal transmembrane proteins are also sorted to lysosomes in the TGN. Targeting of numerous lysosomal transmembrane proteins from the TGN (or from the cell surface) is mediated by tyrosine-based and/or dileucine-based sorting signals present in their cytoplasmic domains (Hunziker & Geuze, 1996; Kirchhausen, 1999). Most tyrosine-based signals conform to the consensus motifs YXXØ (Y is tyrosine, X is any amino acid, and Ø is an amino acid with a bulky hydrophobic side chain) or NPXY (N is asparagine and P is proline) (Kirchhausen, 1999; Marks *et al.*, 1997). While MPR traffic relies on the AP-1 adaptor complex, the proper targeting of many lysosomal membrane proteins requires AP-3 (Le Borgne *et al.*, 1998).

3. Fluorescence *in situ* hybridization in positional cloning

3.1 Positional cloning

Today, it is possible to isolate a disease-related gene simply on the basis of its position in the genome. No knowledge is needed about the biochemical background of the disease or how the gene functions. This technique is commonly referred to as positional cloning (Collins, 1992; Collins, 1995) (Figure 2). The first step towards isolation of a disease gene is collecting families where the trait of interest is segregating. Specific chromosomal localization of the disease gene is determined with linkage analysis (Ott & Bhat, 1999). Linkage analysis is based on polymorphic markers, which are used to detect variations between individuals. This allows separation of maternal and paternal chromosomes. Due to the recombination events in meioses only the markers that are close to the disease gene co-segregate with the disease phenotype. Usually it is possible to restrict the disease gene to a 5 Mbp region by linkage analysis (Collins, 1995). However, in isolated populations, like in Finland, critical chromosomal region can be narrowed down to less than 0.1 kb with linkage disequilibrium mapping and haplotype analysis (Hästbacka *et al.*, 1992; Hästbacka *et al.*, 1994; Peltonen *et al.*, 1995).

After the disease gene region is established, it is possible to utilize the fruits of the Human Genome Project (http://www.ornl.gov/hgmis/) (Lander *et al.*, 2001) and move to use genome browsers (http://genome.ucsc.edu/ (Kent *et al.*, 2002) or http://www.ensembl.org/ (Hubbard *et al.*, 2002)) to search for candidate disease genes from the restricted DNA region, and finally to identify specific disease causing mutations. However, before the first draft of the human genomic sequence was released, gene hunters were forced to construct physical maps over the critical disease gene regions for detailed sequence analyses of regional genes. Physical mapping means isolating and ordering of genomic clones along the disease gene region. An overlapping clone contig is necessary for a large scale sequencing in the critical chromosomal region. Physical mapping is still necessary on certain chromosomal regions, because of sequence annotation problems and still existing gaps in the sequence of the human genome. In the first draft of the human genome, around 90% of the gene-rich (euchromatin) portion of the genome was considered to be completed. This means that only 25 % of the whole genome was in its finished stage (Bailey *et al.*, 2001; Bork & Copley, 2001). Updated information about the progress of the sequencing project can be obtained from http://www.ornl.gov/hgmis/project/progress.html.

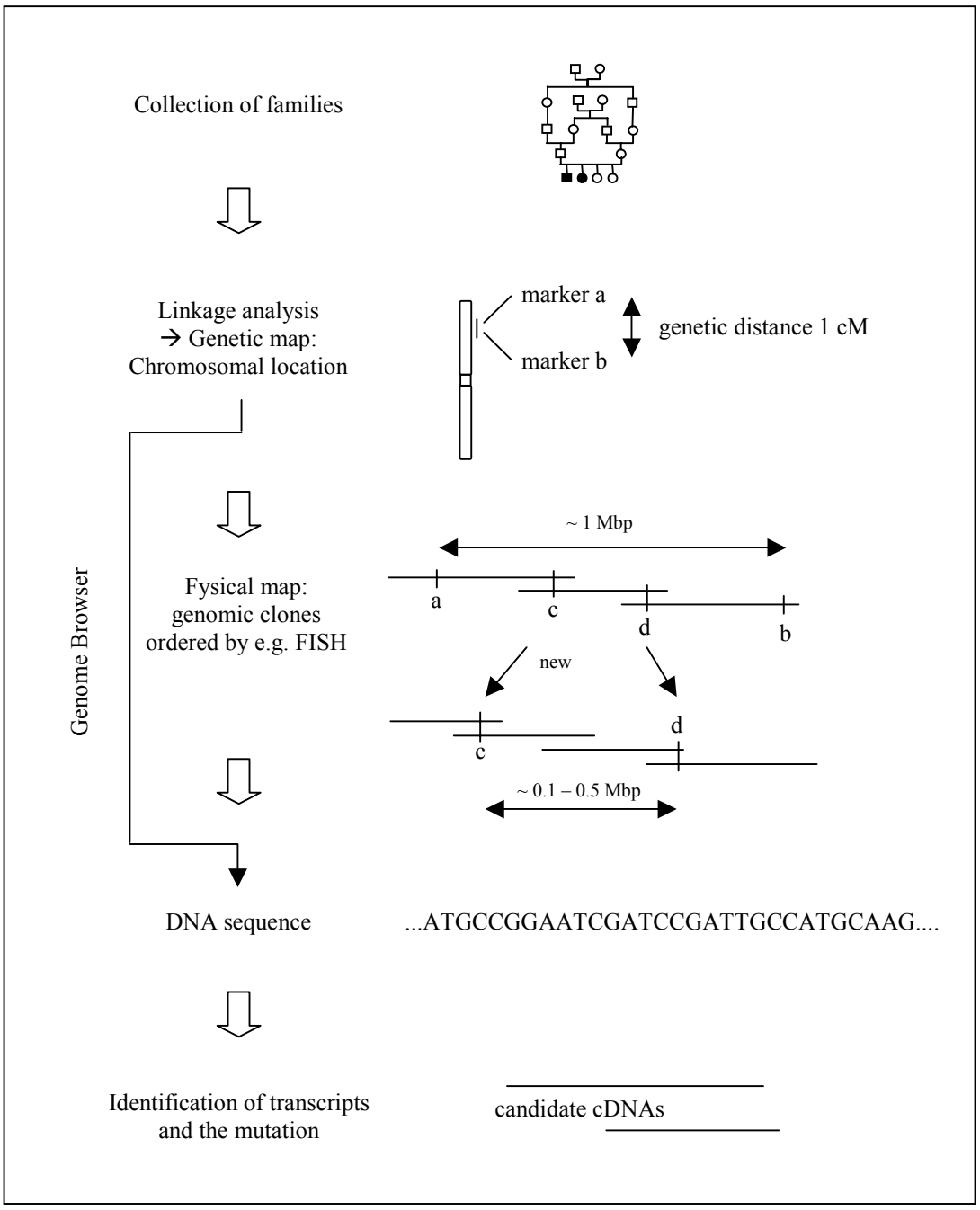

Figure 2. Schematic presentation of different stages of the positional cloning strategy.

3.2 Principle of FISH technique

Fluorescence *in situ* hybridization (FISH) is a technique that allows visualization of specific DNA targets on microscopic slides. The principle of the method is that labeled nucleotide sequences (probes) are hybridized directly to pieces of DNA or RNA with the complementary sequences in metaphase chromosomes, nuclei, tissues or in free chromatin (Figure 3) (Heng & Tsui, 1998; Trask, 1991). The technique involves labeling of the probe with a reporter molecule (e.g. biotin or digoxigenin), followed by hybridization of the labeled probe and target DNA and detection of hybridization with immunofluorescent reagents (directed directly or indirectly against the labeled probe). Finally, the hybridization signal is observed with a fluorescence microscope. The *in situ* hybridization technique was developed in 1969 (John *et al.*, 1969; Pardue & Gall, 1969), at the time when radioisotopes were the only available labels for nuclei acid probes. Tagging of the probes with different fluorescent colors (Pinkel *et al.*, 1986), in conjunction with improvements in fluorescence microscopy and computer based image analysis, has made the technique safe, fast, reliable and sensitive. This has allowed FISH to be applied to a broad spectrum of biological and clinical problems (Heng *et al.*, 1997; Lichter & Ward, 1990; Luke & Shepelsky, 1998). Some examples of them are listed in Table 2.

Table 2. Applications for FISH

Research	**Clinical**
Gene mapping	Clinical cytogenetics
Nuclear architecture	Prenatal diagnosis
Chromatin packaging	Cancer diagnostics
DNA replication	Infectious disease diagnostics
RNA processing	
Gene amplification	
Gene integration	
Chromatin elimination	
Tumour biology	

(For references see Lichter et al. 1990; Luke et al. 1998; Heng et al. 1997)

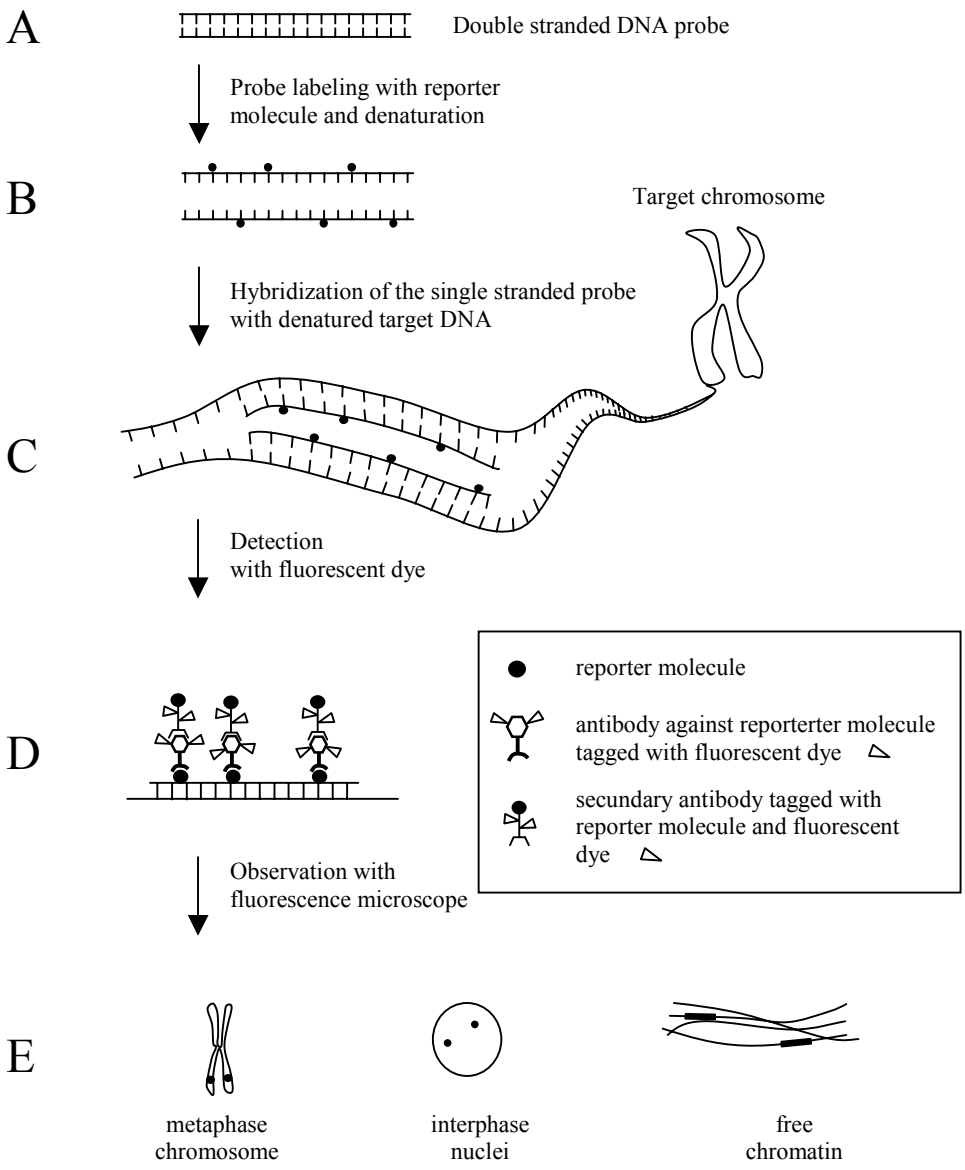

Figure 3. Principle of the FISH technique. A double stranded DNA probe is labeled with reporter molecule (A and B). The probe and target DNA are denatured and allowed to hybridize with each other (B and C). The reporter molecule is detected with fluorescently labeled antibodies (D). Observation of the hybridized sequences is done with fluorescence microscope. Different kinds of signals are seen at the site of the probe hybridization depending on which kind of target DNA is used in the hybridization reaction (E).

3.3 Different resolution FISH applications can be utilized in different stages of physical mapping

Visual mapping by FISH represents the most direct approach for the ordering and orientation of genomic clones. It can be utilized in different stages of a mapping project to speed up ordering of genomic clones (Heiskanen *et al.*, 1996b). Genomic clones for physical mapping can be obtained from genomic DNA libraries in which genomic DNA fragments are cloned into vector DNAs and maintained in yeast or bacterial hosts. Different kinds of vectors allow the incorporation of DNA inserts of various sizes: yeast artificial chromosomes (YACs) up to 2000 kilobase pairs (Burke *et al.*, 1987), bacterial artificial chromosomes (BACs) up to 300 kb (Shizuya *et al.*, 1992), P1 and P1 derived artificial chromosomes (PACs) 80 – 300 kb (Ioannou *et al.*, 1994; Sternberg, 1990) and cosmids 30 – 45 kb (Collins & Hohn, 1978). Ordering of the genomic clones can be initiated by searching for overlapping sequence tagged sites (STS) from the clones. STSs are unique DNA sequences that can be easily amplified with PCR and they can function as landmarks that define the position on the physical map (Olson *et al.*, 1989). Other commonly used methods in physical mapping are pulsed-field gel electrophoresis (PFGE) (Schwartz & Cantor, 1984) and radiation hybrid mapping (Cox *et al.*, 1990). All of these physical mapping techniques are very labour and time consuming. Moreover, they do not provide any information on the size of the overlap or gap between two clones.

Resolution of the FISH mapping depends on the condensation level of the target chromatin (Figure 4). In metaphase chromosomes, differentially labelled probes can be distinguished if they are separated by approximately 1 – 3 megabases (Mb) (Hopman *et al.*, 1986; Lichter *et al.*, 1990). Prometaphase chromosomes have been used for the ordering of DNA sequences in the 50 kb – 1 Mb range (Inazawa *et al.*, 1994; Lebo *et al.*, 1992), mechanically stretched chromosomes in the 0.1 Mb (Haaf & Ward, 1994b; Laan *et al.*, 1995) and interphase nuclei between 50 kb – 1 Mb range (Lawrence *et al.*, 1990; Trask *et al.*, 1993). The highest level of resolution (1 –500 kb) is reached when free chromatin fibers are used as a target for hybridization. Different kinds of techniques have been developed to release chromatin fibers from cells for the assembly of high-resolution physical maps (Heng & Tsui, 1998; Weier, 2001). First fiber-FISH protocols were introduced in 1992. Heng et al. described a chromatin releasing method, which was based on different drug and alkaline treatments (Heng *et al.*, 1992). In parallel with Heng, Wiegant et al. developed a method, which was based on highly extended DNA loops (halo-like structures) arranged around the nuclear matrix (Wiegant *et al.*,

1992). Subsequently, several other fiber-FISH techniques have been described, e.g. a direct visual DNA map (DIRVISH) (Parra & Windle, 1993; Windle *et al.*, 1995), extended chromatin fibers (ECF) (Haaf & Ward, 1994a), free DNA (Fidlerova *et al.*, 1994; Senger *et al.*, 1994), fiber-FISH (Heiskanen *et al.*, 1995; Heiskanen *et al.*, 1996a; Heiskanen *et al.*, 1994; Heiskanen *et al.*, 1996b) and quantitative DNA fiber mapping (Bensimon *et al.*, 1994; Weier *et al.*, 1995).

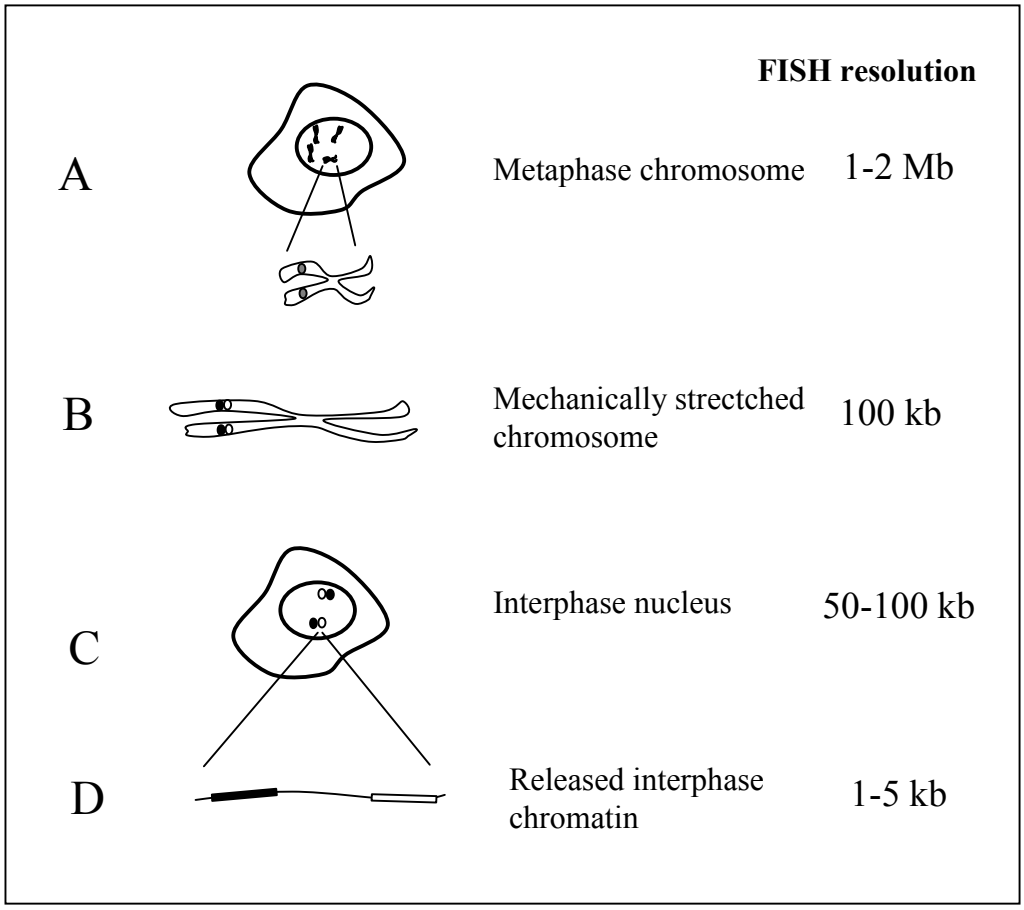

Figure 4. The resolution of FISH depends on the degree of packing of the target DNA. Signals of two probes located 500 kb apart from each other can not be resolved as two separate signals on metaphase chromosomes (A). Mechanical stretching of chromosomes by cytocentrifugation improves the resolution of metaphase FISH and orientation of the probes can be determined (B). In the interphase nucleus the level of chromatin condensation is low. The probes are seen as closely paired signals within the interphase nucleus (C). Free chromatin, which is released from the interphase nucleus, provides the highest resolution for FISH (D).

4. Computational characterization of identified disease genes

4.1 Introduction

The progress in sequencing projects of the human and other species has made computational sequence analysis of the gene a critical step that can provide clues to the molecular basis of pathogenesis and invaluable insights for further experimental analysis (Sreekumar *et al.*, 2001). There are many biological databases and computational sequence analysis tools available on the Internet. Links to different kind of tools and databases can be easily accessed through several molecular biology servers (Table 3). The Molecular Biology Database Collection (Baxevanis, 2002) is a good initial point to start to find useful tools and databases. It provides searchable summaries and updates for each of the databases and is freely available to everyone through the Nucleic Acid Research web site at http://nar.oupjournals.org.

Table 3. Links to some major molecular biology servers.

National Center for Biotechnological Information (NCBI)	http://www.ncbi.nlm.nih.gov/
European Bioinformatics Institute (EMBL-EBI)	http://www.ebi.ac.uk/services/
DNA Database of Japan	http://www.ddbj.nig.ac.jp/
UCSC Genome Bioinformatics	http://genome.ucsc.edu/
Expert Protein Analysis system (ExPASy)	http://www.expasy.ch/
GenomeNet	http://www.genome.ad.jp/

Although biologists are increasingly turning to web-based bioinformatics programs to analyse their molecules of interest, there are also certain limits to computational sequence analysis of which users should be aware. There are often plenty of options available even in simple programs and using default setting may end up with suboptimal results. Potentially interesting biological facts can be overlooked or on the contrary, computationally produced artefacts are easily approved when they seem to point to some really exciting biology. Thus, it is crucial to read the documentation associated with any bioinformatics program or database used, because it often includes useful tips as well as some of the avoidable pitfalls. (Bork, 2000; Claverie, 2000; Fuchs, 2000; Peri *et al.*, 2001; Wolfsberg *et al.*, 2002).

4.2 Nucleotide sequences

4.2.1 Structure of the gene

In order to understand the function of the identified disease gene properly, it is important to define the structure of the gene in detail. Characterization of the gene should include a description of the promoter region, determining of transcription and translation initiation sites, finding of open reading frame, defining of exon/intron boundaries and analysing the structure of the untranslated 3' end region as well. Alternatively spliced forms of the gene should also be identified, as it is suggested that alternative splicing is one of the most significant components of the functional complexity of the human genome (Modrek & Lee, 2002). Thus, the formation of mRNA is only the first step in a long sequence of events resulting in the synthesis of a protein (Figure 5) (Graves & Haystead, 2002).

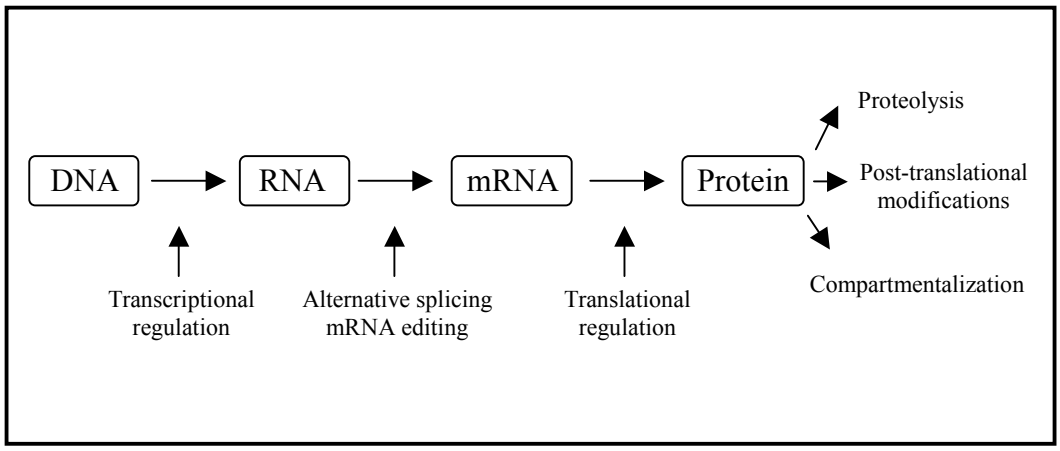

Figure 5. Mechanism by which a single gene can give rise to multiple gene products. Modified from Graves and Haystead (2002).

4.2.1.1 Elucidation of complete gene structures

A good way to initiate characterization of the gene structure is to use The Human Genome Browser (http://genome.ucsc.edu/; Kent *et al.*, 2002) or The Ensembl Genome Browser (http://www.ensembl.org/; Hubbard *et al.*, 2002). They are automatically annotated web tools, which can be used to display graphically any requested portion of the genome. These tools combine information collected from

a wide range of methods and sources (e.g. known genes, gene prediction programs, EST clusters). Nature Genetics has recently published a User's Guide to the Human Genome, which is a handbook for using these browsers (Wolfsberg *et al.*, 2002). The AceView system at the NCBI's web site is also worth visiting (http://www.ncbi.nih.gov/IEB/Research/Acembly/index.html; unpublished). It contains information on human genes, based upon the analysis of all the human mRNAs and ESTs available in Genbank. All of these three tools provide pre-computed analysis of sequences that a user can browse but not alter. However, if a more extensive characterization is needed (e.g. promoter predictions, searching for alternative polyadenylation sites or splice sites beyond known EST sequences), there are also tools for which users can choose their own sequences for analysis (Fortna & Gardiner, 2001). Web-based tools like NIX (http://www.hgmp.mrc.ac.uk/) or RUMMAGE (http://gen100.imb-jena.de/rummage/; Taudien *et al.*, 2000) are useful if a limited length of sequence need to be analyzed. Programs that users run on their own computers are advantageous, if large sequence patterns have to be analyzed (e.g. Genotator, http://www.fruitfly.org/~nomi/genotator/; Harris, 1997).

4.2.1.2 Promoter analysis

Understanding the regulation of the gene expression is an important aspect of understanding the gene function. The promoter is a primary component that controls gene expression. It can be defined as a region of DNA surrounding the transcription start site (TSS) that is able to direct transcription from the correct TSS (Fickett & Wasserman, 2000). The Eukaryotic Promoter Database (EPD, http://www.epd.isb-sib.ch/) is a collection of experimentally defined promoters, but unfortunately promoters have not been defined for most human genes (Praz *et al.*, 2002). Thus, reliable computational methods for recognition and characterization of promoters are needed. However, there are no clear signals for motifs that could be uniformly related to the control of transcription, and performance of many promoter prediction systems has been reported to be very poor (Fickett & Hatzigeorgiou, 1997; Reese *et al.*, 2000). One of the most current methods for the predicting of promoter regions is CONPRO (http://stl.bioinformatics.med.umich.edu/conpro/), which combines several previously developed methods for promoter identification. As a new feature, it utilizes information of EST and mRNA sequences to place potential promoter regions in a genomic sequence. In a test set of 120 promoters, the program detected promoters correctly for about 71% of the human genes with known mRNAs (Liu & States, 2002).

4.2.1.3 Initiation of translation

The identification of correct translation initiation codons is an important aspect of interpreting actual open reading frames of novel mRNA sequences. It has been traditionally suggested that eukaryotic ribosomes initiate translation almost exclusively at the 5' proximal AUG codon (Kozak, 1987; Kozak, 1995), although some exceptions to this rule have been reported (Kozak, 1996). However, in a recent study, it was reported that initiation of translation from upstream AUGs is quite common. This study suggested that leaky scanning and, reinitiation or internal initiation of translation have a much greater role than previously believed (Peri & Pandey, 2001). The program ATG_EVALUATOR can be used for computational prediction of start codons (http://www.itba.mi.cnr.it/webgene/; Rogozin *et al.*, 2001).

4.2.1.4 Utilization of EST sequences

The database of expressed sequence tags (dbEST, http://www.ncbi.nlm.nih.gov/dbEST/; Boguski *et al.*, 1993) is a division of GenBank (Benson *et al.*, 2002) that contains information on partial cDNA sequences from number of different organisms. New data are submitted to dbEST continuously, and in September 2002 there were almost five million human ESTs in this database. Although the stated purpose of most large-scale EST sequencing programs has been gene discovery, it has turned out that these sequence resources are invaluable both for gene prediction and for confirming models of gene structure (Lewis *et al.*, 2000). Besides that EST sequences are very useful for detecting of exon/intron boundaries, they are also extremely convenient for identifying alternate polyadenylation sites (Beaudoing *et al.*, 2000; Beaudoing & Gautheret, 2001; Gautheret *et al.*, 1998) and detecting alternatively spliced forms of the gene (Brett *et al.*, 2000; Brett *et al.*, 2002; Mironov *et al.*, 1999; Modrek & Lee, 2002; Modrek *et al.*, 2001; Xu *et al.*, 2002). Several gene prediction programs (Claverie, 1997) take advantage of EST sequence information, but there are also specialized databases and tools that can be utilized in structural analysis of the gene. For example, SpliceNest (http://splicenest.molgen.mpg.de/) is a web-based graphical tool to explore gene structure, which is based on clustered EST sequences (Krause *et al.*, 2002).

4.2.2 Similarity searches

Database searching with methods like BLAST (Altschul *et al.*, 1990) or FASTA (Pearson & Lipman, 1988) is probably the most familiar stage of sequence analysis for many scientists who have analysed a gene of interest. Database searches may provide information about the function of the gene, if the query sequence appears to be homologous to experimentally annotated gene(s) (Andrade *et al.*, 1999). Different kinds of BLAST programs and comprehensive information how to use them can be accessed through NCBIs web pages (http://www.ncbi.nlm.nih.gov/BLAST/). Similarly, FASTA programs and information how to use them can be found from EMBLs web pages (http://www.ebi.ac.uk/fasta33/).

4.3 Protein sequences

4.3.1 Function

The discovery of protein function directly from sequence has become a fundamental question as thousands of unknown proteins and increasing numbers of complete genomes are made available daily in the public domain (Rigoutsos *et al.*, 2002). In the human genome, it is estimated that there are approximately 30 000 – 40 000 genes in total and the number of annotated genes with unknown function is approximately 40 – 60% (Lander *et al.*, 2001; Venter *et al.*, 2001). Probably the most common way to learn more about the functions of protein molecules is to search for similarities between a query protein and proteins with known annotations in databases. It is possible by using similarity search algorithms like BLAST (Altschul *et al.*, 1990), available for example at http://www.ncbi.nlm.nih.gov/BLAST/, or FASTA (Pearson & Lipman, 1988), available for example at http://www.ebi.ac.uk/fasta33/. Ideally, a search output will show unequivocal similarity to a well-characterized protein over the full length of the query. However, the usual result is a list of partial matches to various unrelated protein families. Thus, identification of functions of multidomain proteins is often problematic. The solution to this problem is to use pattern databases (Table 4), which can be used to assign an unknown query sequence to a known protein family (Attwood, 2000).

Table 4. Some of the major pattern databases

PROSITE	http://www.expasy.ch/prosite/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
SMART	http://smart.embl-heidelberg.de/
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
BLOCKS	http://www.blocks.fhcrc.org/
TIGRFAMs	http://www.tigr.org/TIGRFAMs/

The functional assignments by homology usually involve identification of some specific molecular function of the protein, like enzymatic activity. However, a full description of "protein function" requires a broad range of attributes and features. It is essential to define the function of the protein in the cellular context, e.g., in which metabolic pathway a protein is working on and to define its interacting partners. Finally, we must understand how the protein functions in physiological

subsystems and together with environmental stimuli defines the phenotypic properties of the organism (Bork *et al.*, 1998; Eisenberg *et al.*, 2000).

New computational methods have been developed to place the proteins in their context of cellular function (Eisenberg *et al.*, 2000). These methods utilize information from the fully sequenced genomes of numerous organisms. The method of phylogenetic profiles is based on the assumption that functionally linked proteins evolve in a correlated fashion, and, therefore, they have homologs in the same subset of organisms (Pellegrini *et al.*, 1999). The domain fusion or Rosetta stone method looks for groups of proteins that are distinct in a given organism but appear as a single product in another organism. It is based on the assumption that if a composite protein is uniquely similar to two component proteins in another species, the component proteins are most likely to interact (Enright *et al.*, 1999). The gene neighbour method assumes that it is possible to predict functional coupling genes based on conservation of gene clusters between genomes. This method is most robust for prokaryotic genomes, where gene clusters are typically composed of functionally related genes (Overbeek *et al.*, 1999).

As the need for automated approaches for the functional assignment of proteins increases, new methods are published regularly. One of them is the dictionary-driven protein annotation approach (http://cbcsrv.watson.ibm.com/Tpa.html) (Rigoutsos *et al.*, 2002). It is based on similarity searches in the Bio-Dictionary, which is a collection of small amino acid sequences derived from public databases using the TEIRESIAS algorithm. This algorithm is designed for discovery of rigid patterns in biological sequences (Rigoutsos & Floratos, 1998). Another recent protein function prediction method is the ProtFun (http://www.cbs.dtu.dk/services/ProtFun/). This approach utilizes functional attributes, which are predictable from amino acid sequence – like post-translational modifications and protein sorting signals (Jensen *et al.*, 2002).

4.3.2 Structure of the protein – soluble or membranous?

There are several tools on the Internet, which allow computational analysis of virtually all aspects of protein structure – primary, secondary and even tertiary structure of the protein can be analyzed. The ExPASy server provides a variety of tools to perform these analyses (http://us.expasy.org/tools/). For example, secondary structure prediction of protein can be done quite accurately by using

the PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) server. It incorporates three recently developed methods for predicting structural information about a protein from its amino acid sequence alone (McGuffin *et al.*, 2000). The prediction of the tertiary structure of a protein is much more difficult. Although there have been a number of promising advances in predicting the structure from amino acid sequence alone, homology based modeling is still the most accurate method to make these predictions (Baker & Sali, 2001). One of the latest programs for modeling of the three-dimensional structures is ESyPred3D (http://www.fundp.ac.be/urbm/bioinfo/esypred/; Lambert *et al.*, 2002).

An important structural issue in analysis of novel protein sequences is to classify them into either soluble or membrane proteins. Several computational methods using different algorithms have been developed for prediction of transmembrane helices directly from amino acid sequences (Moller *et al.*, 2001). Many approaches rely on two basic rules: (I) transmembrane helices are short amino acid stretches with a high overall hydrophobicity, and (II) positively charged residues (arginine and lysine) are mainly found in the non-transmembrane parts of the protein on the cytoplasmic side determining the orientation of the protein in the membrane (von Heijne, 1996). Thus, identification of transmembrane segments is often based on hydrophobicity blots (Kyte & Doolittle, 1982) and on "the positive-inside rule" (von Heijne, 1992). However, these basic rules are easily blurred and correct prediction of the location and orientation of all transmembrane segments has proved to be a difficult problem. Thus, several methods have been developed to improve the accuracy of predictions (Sonnhammer *et al.*, 1998). In the recent evaluation of the performance of the currently best known and most widely used methods for the prediction of transmembrane regions, the best performing program was a hidden Markov model based on TMHMM, available at http://www.cbs.dtu.dk/services/TMHMM/ (Krogh *et al.*, 2001). Apart from its performed best in determining transmembrane regions, it was also especially good at reliably distinguishing between soluble and transmembrane proteins (Moller *et al.*, 2001). One of the common problems of transmembrane prediction programs is their tendency to interpret the hydrophobic parts of signal sequences and transit peptides as membrane spanning regions. Therefore, all predictions should be performed with the consultation of signal sequence prediction methods like SignalP 2.0 (Nielsen *et al.*, 1997).

4.3.3 Intracellular localization

The functional description of a protein very often indicates the cellular compartment where the protein is located. Thus, subcellular localization of a newly identified protein is a key attribute to define its function (Eisenhaber & Bork, 1998; Mott *et al.*, 2002). Currently, there are three conceptually different computational methods to predict the subcellular localization of the protein from its amino acid sequence (Emanuelsson & von Heijne, 2001; Mott *et al.*, 2002). The first category utilizes sorting signals, like signal peptides, membrane spanning segments, lipid anchors, nuclear import signals and different organelle targeting motifs. The second category of approaches is based on the observation that proteins from different compartments tend to differ in subtle ways in their overall amino acid composition. Thirdly, a phylogenetic profile can be used to assign query proteins to subcellular locations. This method is based on the finding that the phylogenetic profiles of proteins with the same cellular location are often similar (Marcotte *et al.*, 2000).

In the Internet, there are some programs available for the prediction of intracellular localization. PSORT (http://bioweb.pasteur.fr/seqanal/interfaces/psort2.html) program requests a full-length amino acid sequence, then it calculates values for various sorting features, e.g. different signal sequences and motifs, and displays some of the most probable localization for the protein (Nakai & Horton, 1999). TargetP (http://www.cbs.dtu.dk/services/TargetP/) is a neural network-based tool for location prediction. It utilizes N-terminal sequence information only, and discriminates between proteins destined for the mitochondrion, the chloroplast, the secretory pathway, and other localizations (Emanuelsson *et al.*, 2000).

An important analysis for a new protein sequence is to characterize the presence or absence of the N-terminal signal peptide (Nielsen *et al.*, 1997). Targeting of the protein to the secretory pathway, to mitochondria and to chloroplasts normally depends on an N-terminal presequence that can be recognized by receptors on the surface of the appropriate organelle. Currently, the most widely used method to predict secretory signal peptides is the neural network based SignalP predictor (http://www.cbs.dtu.dk/services/SignalP-2.0/) (Nielsen *et al.*, 1997). The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models. For the prediction of mitochondrial targeting peptides MitoProt

(http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter) can be utilized (Claros & Vincens, 1996).

4.3.4 Post-translational modifications

After synthesis proteins can be further processed to enhance their capabilities. Most proteins are cleaved or trimmed protelytically following translation. The initiation methionine is usually removed after protein synthesis and a cleavage of the N-terminal signal sequence is also a common proteolytical modification. In addition, a variety of different kinds of protein modifications are known. Some of them, like glycosylation, phosphorylation and lipidation, can play very important physiological roles and thus it is of special interest to predict these events directly from amino acid sequences (Nakai, 2001).

4.3.4.1 Glycosylation

Asparagine-linked (N-linked) glycosylation is often found to occur in secretory and membrane proteins. In the early secretory pathway, the N-linked glycans play a pivotal role in protein folding, oligomerization, quality control, sorting and transport. In the Golgi complex, the glycans acquire more complex structures and a new set of functions. It is known that the consensus sequence Asn-X-Ser/Thr is necessary, but not sufficient for N-glycosylation (Helenius & Aebi, 2001; Parodi, 2000). In O-glycosylation the glycan moiety is covalently linked to the hydroxyl group of serine or threonine residue. It influences a number of properties of proteins including proteolytic resistance, solubility, immunological properties and ligand binding. Certain rules and acceptor motifs have been proposed for O-glycosylation, but there are no definite rules, which distinguish O-glycosylated amino acids from non-glycosylated residues (Gupta *et al.*, 1999). Thus, neither N-linked nor O-linked glycosylation can be predicted solely on the consensus sequences. The NetNGlyc 1.0 server (http://www.cbs.dtu.dk/services/NetNGlyc/) can be used for predicting of N-glycosylation sites (Gupta et al. 2002, in preparation, see the web page), and NetOGlyc 2.0 for predicting of O-glycosylation sites (Hansen *et al.*, 1998). They are both based on an artificial neural network method.

AIMS OF THE PRESENT STUDY

Prior to this study, the *PPT1* gene responsible for the INCL (CLN1) disorder had been recently identified (Vesa *et al.*, 1995), and physical mapping of the critical chromosomal region of the vLINCL (CLN5) gene was initiated. This study was undertaken to further understand the pathogenesis of these disorders, which are enriched especially in the Finnish population.

The specific aims of this study were:

1. To utilize the fiber-FISH technique in the positional cloning of the *CLN5* gene (I, II).

2. To characterize intracellular processing and localization of the wild-type and mutant CLN5 protein in transiently transfected cell lines (III)

3. To characterize the expression of *PPT1* in developing mouse brain (IV).

MATERIALS AND METHODS

The materials and methods are described in more detail in the original publications (I – IV).

1. Visual mapping by fiber-FISH (I, II)

The order and orientation of the PAC and cosmid clones of the CLN5 region (isolation of clones is described by Klockars et al. 1996) were verified by FISH on extended DNA fibers as previously described in detail (Heiskanen *et al.*, 1996a; Heiskanen *et al.*, 1994). Briefly, clones were labeled by standard nick-translation protocol with either biotin-11-dUTP (Sigma Chemical, St. Louis, MO, U.S.A.) or digoxigenin-11-dUTP (Boehringer Mannheim, Germany). Target DNA fibers were prepared from lymphocytes embedded in agarose blocks containing about 5 μg human genomic DNA. A piece of agarose block was placed on a microscopic slide precoated with 0.15% gelatin and 0.2% Poly-L-Lysine. An agarose block was melted with 20 μl of deionized water in a microwave oven and the DNA was extended mechanically on a slide. Hybridization and detection of probes was performed using standard FISH protocols (Pinkel *et al.*, 1986). Biotinylated probes were detected using TRITC-conjugated avidin D and the signal was amplified by biotinylated goat anti-avidin D and another layer of avidin-TRITC (Vector, Burligname, CA). For digoxigenin labeled probes, mouse anti-digoxigenin antibodies (Boehringer Mannheim) and fluorescein conjugated sheep anti-mouse and donkey anti-sheep antibodies (Sigma Chemicals) were used. To prevent fading slides were mounted in antifade solution (Vectashield, Vector).

For positioning of the genes on the CLN5 region the high-sensitive tyramide-based detection was performed using the Tyramide Signal Amplification (TSA) kit (NEN-DuPont, Boston, MA, USA) by modifying the manufacturer's instructions and the protocol published by Raap et al. (Raap *et al.*, 1995). Briefly, the probes for the RNA Helicase A pseudogene and for the HUMBTFB were labeled by nick-translation with biotin-11-dUTP (Sigma Chemical, St. Louis, MO) and genomic PAC clone 76N15 with digoxigenin-11-dUTP (Boehringer Mannheim, Germany). The biotinylated probes were visualized with the TSA kit and Streptavidin-Texas Red (Vector, Burlingame, CA) antibodies. To visualize the digoxigenin-labeled probes simultaneously, the slides were co-incubated with mouse anti-digoxigenin and FITC-conjugated sheep anti-mouse antibodies

(Sigma Chemicals). The signal of digoxigenin labeled probe was amplified with an additional layer of FITC-conjugated donkey anti-sheep antibody (Sigma). For the detailed protocol see the original article (II).

Hybridization signals were analyzed using a digital multicolor image analysis system, based on a CCD (charge-coupled device) camera (Photometrics PXL, Photometrics Inc., Waterloo, Ontario, Canada) attached to a Power Mac 7100/Av workstation. The image acquisition, the distance measurements and the positions of the gene specific probes within the genomic clone probe were performed on a Macintosh system using the IP Lab software options. The distance measurements were based on the known sizes of the probes.

2. Transient cDNA expression (III)

2.1 The expression constructs

For the transient cell expression analyses the coding region of *CLN5* was amplified by RT-PCR from human fibroblast RNA and cloned into the mammalian expression vector pCMV5 (Andersson *et al.*, 1989). For *in vitro* translation analysis cDNA was subcloned into a pGEM3-vector (Promega, Madison, WI, U.S.A.). The FIN_M mutant cDNA construct was generated by the QuickChange site-directed *in vitro* mutagenesis kit with the WT *CLN5* cDNA in pCMV5 as a template.

2.2 Cell culture and transfection

For the transfection of the CLN5 cDNA constructs, BHK-21 cells (Syrian golden hamster kidney cells, CCL-10; ATCC, Manassas, VA, U.S.A.) were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum (FCS, Gibco BRL, NY, U.S.A) and antibiotics (penicillin and streptomycin, Sigma Chemical, St. Louis, MO, U.S.A.). One day before transfection, the cells were seeded on 9.6 cm^2 plates at a density of 1.5 x 10^5 cells per well. Transfection was performed with the FuGENE 6 transfection reagent (Roche Diagnostics, Indianapolis, IN) following the guidelines supplied by the manufacturer. Experiments were performed 48 h after transfection.

2.3 Antibodies

To obtain specific antibody against the CLN5 protein (5289), a keyhole limpet hemocyanin coupled synthetic peptide (CYETWNVKASPEKGAET) corresponding to amino acids 258 273 was purchased from Genosys Biotechnologies Ltd (Europe, London) and used to immunize rabbits. The resulting antiserum was IgG-purified using Protein A Sepharose CL-4B (Amersham Pharmacia Biotech). The medial Golgi-specific antibody CTR433 and the lysosome/late endosome-specific antibody lgp120 were kind gifts from Dr. Michel Bornens (Institute CURIE, Paris, France) and Dr. Jean Gruenberg (Department of Biochemistry, Geneva, Switzerland), respectively. Secondary antibodies, FITC- (excitation/emission maxima: 494/518 nm) or TRITC- (554/576 nm) conjugated anti-rabbit or anti-mouse antibodies, were purchased from Jackson's Immunoresearch Laboratories (Bar Harbor, ME).

2.4 Immunoprecipitation

To investigate biosynthesis and processing of the CLN5 protein, transfected BHK-21 cells were metabolically labeled by starving them in methionine- and cysteine-free medium (Lifetechnologies, Rockville, MD, U.S.A.) for 1 hour and thereafter labeling with 50 µCi/ml of both [^{35}S]methionine and [^{35}S]cysteine (Amersham, Buckinghamshire, UK) for 2 hours. After the pulse labeling, cells were harvested and lysed with RIPA-buffer (50 mM Tris pH 8.0, 150 mM NaCl, 1% IGEPAL, 0.5% deoxycholic acid, 0.1% SDS) supplemented with protease inhibitors (Complete^{TM}, Roche Diagnostics, Indianapolis, IN). Lysed cells and culture media samples were immunoprecipitated with the 5289 anti-CLN5 antibody and Pansorbin® cells (Calbiochem). Immunocomplexes were separated on 12% SDS-PAGE and visualized by fluorography (Amplify, Amersham, Buckinghamshire, England). EndoH and PNGase F digestions of immunocomplexes were performed as recommended by the manufacturer (New England BioLabs Inc., Beverly, MA, U.S.A.).

2.5 Immunofluorescent cell staining and confocal microscopy

To investigate intracellular localization of the CLN5 protein, transfected BHK-21 cells were grown on coverslips and fixed 48 h after transfection with 4% paraformaldehyde (PFA) for 20 min, permeabilized for 2 min with -20ºC

methanol and blocked with 0.5% bovine serum albumin (BSA) prior adding primary antibodies. Cells were washed with 0.5% BSA and incubated with secondary antibodies. After washing with phosphate buffered saline (PBS) and water, the cells were mounted in GelMount (Biomeda, Foster City, CA) and analyzed using Leica DMR confocal immunofluorescence microscopy equipped with the argon laser and TCS NT software.

2.6 *In vitro* translation

In vitro translation of WT CLN5 in pGEM3Z plasmid was performed using the TNT® T7 Quick Coupled Transcription/Translation System (Promega, Madison, WI). Aliquots of 0.5 µg of plasmid DNA and 10 µCi of Redivue L-[^{35}S] methionine (Amersham Pharmacia Biotech) were incubated for 90 min at 30ºC in a final volume of 25 µl. Control reactions were performed without plasmid DNA. Fixed (50% methanol, 10% acetic acid, 40% water, 30 min) and dried gels were exposed directly to Kodak BioMax MR film for visualization.

3. mRNA and protein expression in mouse tissues (IV)

3.1 Tissue specimens

The mouse strains used in the study were Hsd-NIHS (produced at the National Public Health Institute in Kuopio, Finland) and CD-1 (Institute of Child Health, London, UK). The samples were either immersion fixed in 4% PFA in 0.1% M phosphate buffer over night (for immunohistochemistry) or frozen in dry ice (for *in situ* hybridizations and northern blot). Embryos for whole-mount *in situ* hybridization were explanted from the uterus, rinsed in diethylpyrocarbonate (DEPC)-treated PBS and fixed in 4% PFA overnight. Animal care and handling were at all times consistent with the guidelines set out in the National Research Council's guide for laboratory animal care and use. The experiments were approved by the Laboratory Animal Committee of the National Public Health Institute.

3.2 Northern blot

For Northern blot analysis, mRNA was extracted from mouse brain at different developmental stages. Eight micrograms of mRNA was loaded per lane on 0.8% agarose formaldehyde gels and the separated mRNA was blotted onto nylon

membranes (Hybond N, Amersham, UK) according to standard protocols. A 32P-labeled PCR fragment of the mouse PPT cDNA (nucleotides 232 – 1002) was used as a probe. The filter was stripped and rehybridized with a ß-actin cDNA probe (Clontech). The values of the relative mRNA expression levels were obtained by densitometric scanning of autoradiograms using Bio Image densitometry (Millipore, Ann Arbor, MI, U.S.A.). The PPT signals were normalized against the ß-actin signal of each mRNA sample.

3.3 *In situ* hybridization

For expression analysis of the *PPT* mRNA, two fragments of the mouse PPT cDNA (nucleotides 232-1002 for the brain samples and 123-496 for the embryos) were cloned into a pBluescript II vector (Stratagene, CA, U.S.A.). Digoxigenin labeled sense and anti-sense probes were generated using a DIG RNA labeling kit in accordance with the manufacturer's instructions (Boehringer Mannheim). Non-radioactive *in situ* hybridization for brain samples (cryosections) were performed as described previously (Nolo *et al.*, 1996). For more details see the original publication (IV). Embryos for whole-mount *in situ* hybridization were processed as described previously (Wilkinson, 1992). Nicholas D. E. Green performed whole-mount *in situ* hybridizations in London.

3.4 Antibodies

Rabbit antiserum against PPT glutathione S-transferase (GST) fusion protein (Hellsten *et al.*, 1996) was used for immunohistochemistry. In addition, rabbit antiserum against a synthetic peptide corresponding to amino acids 103 – 119 (CPKLQQGYNAMGFSQGGQ) of human PPT protein was raised (code 336). The commercially available primary antibodies used in the study were monoclonal anti-myelin basic protein (Boehringer Mannheim, Germany), polyclonal anti-cow glial fibrillar acidic protein (DAKO, Glostrup, Denmark) and monoclonal anti-synapthophysin (DAKO).

3.5 Immunohistochemistry

Endogenous expression of PPT1 protein in mouse tissues was analyzed with immunohistochemical stainings. Tissue slides were deparafinized and nonspecific staining was reduced by incubating the sections in 5% H_2O_2 for 5 min followed

by incubation with normal goat serum (Vector). Immunostaining of the sections was performed as described in the original article (IV).

3.6 Microscopy

Bright-field images of *in situ* and immunohistochemical analyses were digitized using an Olympus AX70 Provis microscope (Olympus Optical, Tokyo, Japan) equipped with a photometrix SenSys CCD camera and Image ProPlus 3.0 software (Media Cybernetics, Silver Spring, MD, USA). For whole-mount *in situ* hybridization, photomicrographs were first taken using a Zeiss SV11 microscope (Zeiss, Jena, Germany) and then scanned using a UMAX Astra 1200S color scanner (UMAX Technologies, USA).

RESULTS AND DISCUSSION

1. Utilization of FISH in characterization of the CLN5 region (I and II)

1.1 Construction of the visual physical map over the CLN5 region (I)

The CLN5 locus was originally positioned by linkage analysis to the chromosomal region 13q21.1-q32 between markers D13S162 and D13S160 (Savukoski *et al.*, 1994). Based on the Genethon human genetic linkage map (Gyapay *et al.*, 1994), the genetic distance between these markers was estimated to be about 4 cM (corresponding roughly to 4 Mb physical distance) (Varilo *et al.*, 1996). Construction of the physical map over this region was initiated by isolating YAC-clones positive to either of these markers (Klockars *et al.*, 1996). As it is known that chimerism and other rearrangements limit the usefulness of YAC-clones in physical mapping (Green *et al.*, 1991), metaphase FISH analysis was utilized to confirm nonchimeric character of isolated YAC-clones. This analysis also refined the chromosomal region of the CLN5 to be 13q22. Also mechanically stretched chromosomes (Laan *et al.*, 1995) and interphase nuclei were used as hybridization targets in FISH analysis to confirm orders and relative distances of individual YAC clones on the region.

Finding of the two polymorphic markers, COLAC1 and AC224, allowed narrowing of the CLN5 region down to a single YAC clone (Klockars *et al.*, 1996). However, YAC clones are not suitable for large scale sequencing projects. They are often too large in size for shotgun sequencing and moreover, metaphase FISH analysis cannot completely exclude chimerism of these clones. Thus, construction of a physical map with shorter insert genomic clones was necessary. Screening of the PAC library with polymorphic markers and STSs (sequenced from the end of identified clones) resulted in isolating of several PAC clones. Overlaps between many of these clones were established with STS content mapping, but there were still gaps in the contig. Visual mapping with fiber-FISH (Heiskanen *et al.*, 1995; Heiskanen *et al.*, 1996a; Heiskanen *et al.*, 1994) improved essentially analysis of the PAC contig. It not only enabled rapid confirmation of clone orders obtained from STS-based mapping, but also made it possible to directly visualize overlaps and gap distances between various PAC clones. For example, detection of the overlap between PACs 76n15 and 186a15 was solely based on the fiber-FISH analysis (Figure 6). To estimate the physical distance of the critical CLN5 region, fiber-FISH was carried out using the cosmid clones positive with restricting markers COLAC1 and AC224, and the PAC

76n15. Based on the determined sizes of genomic clones (Klockars *et al.*, 1996), critical CLN5 region was estimated to be approximately 200 kb.

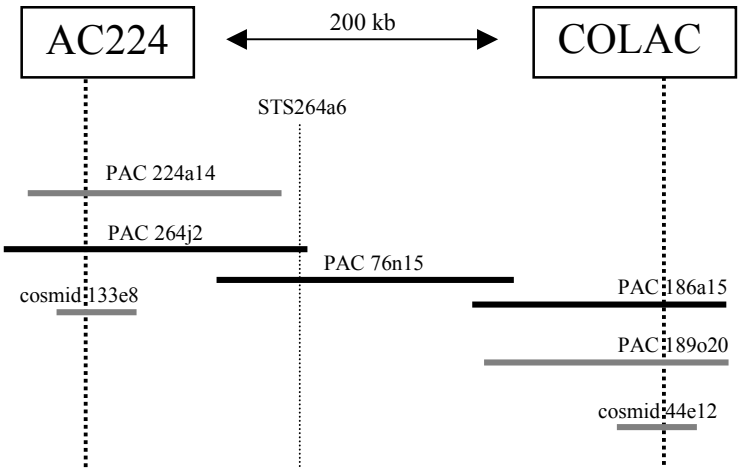

Figure 6. Schematic presentation of a complete physical map covering the CLN5 region. Positions of markers AC224 and COLAC1 restricting the critical CLN5 region are shown. Three PACs sequenced as a part of the Human Genome Project at the Whitehead Institute (Cambridge, MA, USA) are indicated in black. Klockars *et al.* (1996) and I

The continuous contig of genomic clones over the critical CLN5 region was composed with several overlapping PAC and cosmid clones (Figure 6). The fiber-FISH mapping replaced many traditional time-consuming mapping methods. This very same fiber-mapping method has been successfully utilized in several other positional cloning projects as well (Aaltonen *et al.*, 1997; Leppanen *et al.*, 1996; Nikali *et al.*, 1997; Paavola *et al.*, 1999). As the Human Genome Project has progressed rapidly, physical mapping is no more needed for most of the euchromatic regions of the human genome. However, physical map assembly in heterochromatic regions by conventional methods (e.g. by STS content mapping) has proved to be difficult, and different kinds of fiber-mapping techniques can be useful on these regions (Weier, 2001). Moreover, genome geography has been shown to be more complex than previously recognized. Especially, segmental duplicated regions, which may underlie a great amount of human phenotypic variation and disease, are marginalized within both private and public genome

assemblies (Bailey *et al.*, 2002; Bailey *et al.*, 2001; Eichler, 2001). Published results have demonstrated the power of the FISH methodology to clarify genomic structures on regions where duplications, deletions or inversions have occurred (Gervasini *et al.*, 2002; Giglio *et al.*, 2001).

1.2 Positioning of coding regions by high-sensitive tyramide-based detection method (II)

After the critical CLN5 region was covered by three PAC clones (see Figure 6), partial shotgun sequencing of these clones was performed in order to identify coding regions and to create new STSs. This search resulted in the identification of the BTF3 protein homologue gene (BTFBL1, accession No. NM_001208) from PAC 224a14, and a previously unknown pseudogene for RNA helicase A (Accession No. L13848) from PAC 264j2 (see Figure 6). We utilized the fiber-FISH method to precisely assign these genes on the physical clones. A PCR positive cosmid 50c4 was used as a probe for BTF3L1 and two 2 kb genomic PCR products as a probe for the RNA helicase pseudogene. Since the visualizing of short probes (few kb) may be difficult, we utilized the highly sensitive tyramide-based signal amplification method (Raap *et al.*, 1995) in combination with the conventional labeling technique (Pinkel *et al.*, 1986). In conventional FISH hybridization the signal is detected by immunoassays using fluoroschrome-conjugated antibodies. The sensitivity of tyramide detection is based on a precipitation reaction of the biotinylated tyramide by a horse-radish peroxidase enzyme over a biotin labeled probe. In the original report (Raap *et al.*, 1995), tyramide-based detection was used for a single-color FISH. In this study, this method was combined with the conventional labeling technique, which allowed us to use the tyramide-detection method in a two-color FISH analysis. As a result, the 4 kb PCR fragment was detected as a short and bright (red) hybridization signal inside the PAC 76n15 signal (green). The signal of the cosmid 50c4 was observed in a short distance from the end of the PAC 76n15. Based on the known size of cosmid 50c4 (32 kb), the distance between these two genes was determined to be 35 kb.

Later on, this same method was used successfully for detection of short DNA-fiber targets ranging from 0.3 to 1 kb (Horelli-Kuitunen *et al.*, 1999). Moreover, development of different tyramide conjugates has enabled the detection of multiple targets in different colours in the same hybridization reaction (Speel *et al.*, 1997; van Gijlswijk *et al.*, 1997). Although it is quite clear, that tyramide-detection increases sensitivity of the FISH, it is suggested that the limit of the

FISH sensitivity is determined by hybridization efficiency – not by the ability to generate sufficient signal from small probes (van de Rijke *et al.*, 2000).

2. Characterization of the CLN5 protein (III)

2.1 Cloning of the CLN5 cDNA

The CLN5 gene was identified by screening a fetal brain cDNA library with one of the PAC clones (76n15, see Figure 6) covering the critical CLN5 region (Savukoski *et al.*, 1998). This search resulted in identification of several cDNA clones (Klockars *et al.*, 1999) and one of them turned out to be the CLN5 gene. At the same time with cDNA library screenings the PAC clones were sequenced as a part of the Human Genome Project at the Whitehead Institute. The genomic sequence of the PAC 76n15 (accession number AC001226) and numerous overlapping EST sequences greatly facilitated assembling the structure of the *CLN5* gene (Figure 7) (Savukoski *et al.*, 1998).

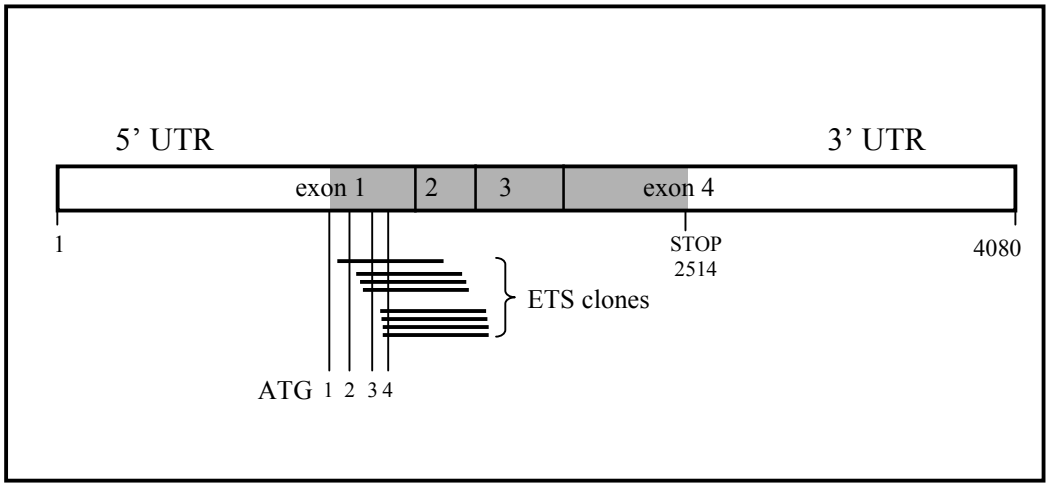

Figure 7. Structure of the CLN5 cDNA. An open reading frame (nucleotides 1293 – 2514) is shown in grey. The positions of the four in frame ATG codons are indicated (ATG 1 at 1293, ATG 2 at 1380, ATG 3 at 1440, ATG 4 at 1476). The most 5' end EST clones of the *CLN5* gene are shown. (Modified from Savukoski et al. 1998).

In order to characterize intracellular processing and localization of CLN5 protein in transiently transfected cell lines, we amplified the open reading frame region of *CLN5* (gray region in Figure 7) by RT-PCR from human fibroblast RNA and cloned it into a mammalian expression vector pCMV5. Although we were able to amplify CLN5 cDNA by RT-PCR, it was a very difficult task. Difficulties in RT-PCR may have been related to the high content of C and G nucleotides (CpG island) at the 5' end of the gene. However, an interesting feature of the *CLN5* is that it has an exceptionally long 5' end UTR region and there are four in frame initiation methionines in the 5' end of the gene. Later on, our expression analyses indicated that translation of *CLN5* is not necessarily initiated from the first in frame initiation codon, but potentially upstream AUGs are used (see 2.2.2 and 2.2.3). This may indicate that the originally reported *CLN5* cDNA may include the retained 5' end intron. It is suggested that frequent cloning of introns containing cDNAs is not accidental. Slow removal of the 5' end intron in particular might be a way to regulate the expression of mammalian genes (Kozak, 1996). The suggestion that the *CLN5* gene retains the 5' end intron is further supported by the fact that none of the EST sequences overlap with the first in frame AUG codon (see Figure 7). Thus, it is possible that also we have cloned a pre-mRNA form of *CLN5*, where the most 5' end intron is still intact. For expression analyses we used this "full-length" cDNA, which contains all of the four initiation methionines.

2.2 Expression analysis of the wild-type and mutant CLN5 proteins

2.2.1 Intracellular localization of WT and FIN_M CLN5 proteins

The subcellular localization of the CLN5 protein was studied in transiently transfected BHK-21 cells using the 5289 peptide antibody, organelle-specific antibodies and confocal microscopy. Wild-type CLN5 was seen in vesicular structures, which overlapped almost completely with the lysosomal/endosomal marker lgp120 (Figure 8A). The targeting of the most common vLINCL mutant (FIN_M), lacking the 16 carboxy-terminal amino acids, was evidently different from that of the wild type protein. FIN_M proteins showed co-localization with the medial Golgi marker CTR433 (Figure 8B). This indicates that the mutant CLN5 polypeptides are initially sufficiently correctly folded to pass the quality control of the ER but are not recognized as lysosomal proteins by the intracellular sorting system in the Golgi.

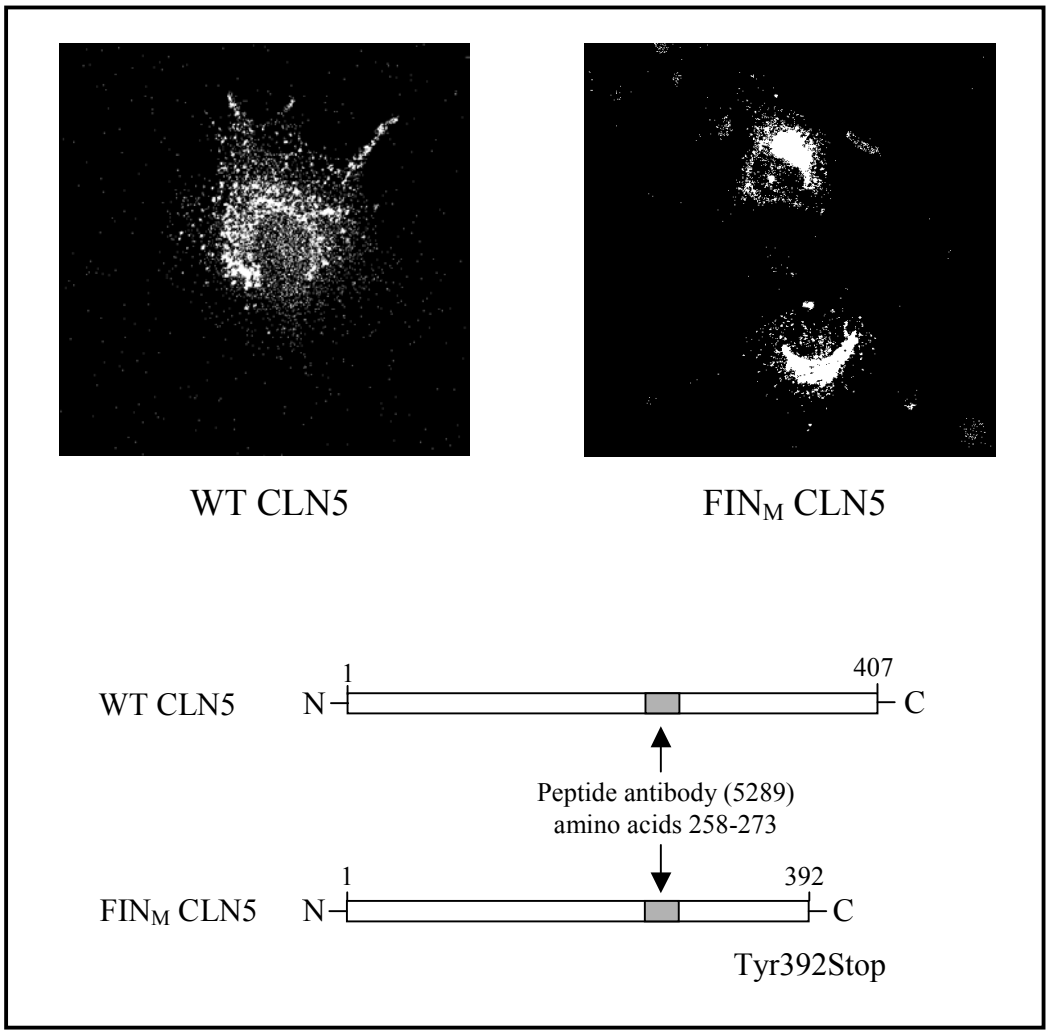

Figure 8. Subcellular localization of WT and mutant CLN5 proteins in transiently transfected BHK-21 cells. WT CLN5 proteins are transported to lysosomes and FIN_M polypeptides are seen within a Golgi region. The predicted consequence of FIN_M mutation on the CLN5 polypeptide and position of the 5289-peptide antibody recognition site are shown in the box above immunofluorescence figures.

Immunoprecipitation and computational sequence analyses of the CLN5 polypeptide implied that it is probably a soluble protein (see chapter 2.2.2). This raises the possibility that CLN5 could represent a soluble lysosomal enzyme targeted to lysosomes via the Man-6-P receptor –mediated pathway. However, further studies are needed to determine whether the high mannose-type

oligosaccharides (see 2.2.2) of the CLN5 protein are phosphorylated and what is the exact lysosomal targeting mechanism and the function of CLN5. Interestingly, FIN_M polypeptides were targeted to lysosomes in COS-1 cells (Vesa *et al.*, 2002). This would suggest that the lysosomal transport machinery of the CLN5 protein might vary between different cell types. As it is known that a deficiency of CLN5 affects mainly cells in the central nervous system, the intracellular localization of CLN5 should also be determined in neurons in order to determine the localization of both WT and mutated polypeptides in a more relevant cellular background.

2.2.2 Biosynthesis of WT and FIN_M CLN5 proteins

Immunoprecipitation analysis was used to monitor the biosynthesis and intracellular processing of the WT and FIN_M CLN5 proteins in transiently transfected BHK-21 cells. This analysis revealed that WT CLN5 is a 60 kDa and FIN_M a 52-kDa glycoprotein. After deglycosylation with PNGase F WT protein was collapsed to a single 38 kDa polypeptide and FIN_M protein to a single 35 kDa band. Interestingly, the theoretical molecular weight of the WT protein translated from the first initiation methionine is substantially more: 46.3 kDa. Based on this finding, it seemed possible that CLN5 protein is not translated from the first methionine, but probably one of the upstream methionines is used. This suggestion is further supported by the fact that in the corresponding mouse gene there is only one initiation methionine and its position is comparable to the position of the fourth AUG codon in the human CLN5 gene. Moreover, if the fourth initiation methionine is used, a SignalP program (Nielsen *et al.*, 1997) predicts that the CLN5 polypeptide undergoes proteolysis of the N-terminal signal peptide. The theoretical molecular weight of such a protein would be 36.6 kDa. It corresponds quite well with the observed 38-kDa molecular weight of the deglycosylated form of WT CLN5.

Interestingly, both WT and FIN_M CLN5 polypeptides were also secreted into the culture medium. This would imply that CLN5 is a soluble lysosomal glycoprotein, not an integral transmembrane protein as predicted earlier (Savukoski *et al.*, 1998). This prediction was based on the results of the Tmpred transmembrane prediction program and a Kyte-Doolittle hydrophobicity blot. However, these programs are very poor in their ability to distinguish between soluble and membranous proteins (Moller *et al.*, 2001). The best performing programs have been reported to be TMHMM (Krogh *et al.*, 2001) and SOSUI (Hirokawa *et al.*, 1998). Neither of them predicts any transmembrane regions for

CLN5. In conclusion, our results indicate that CLN5 is a soluble lysosomal protein. Two other previously characterized NCL gene products; CLN1 and CLN2 have been shown to be lysosomal enzymes. This should encourage further analyses of potential enzymatic functions of CLN5.

2.2.3 Utilization of alternative in frame translation initiation codons

Immunoprecipitation analysis with the 5289-peptide antibody indicated that the CLN5 polypeptide is not translated from the first AUG codon in BHK-21 cells. The structure of the mouse Cln5 gene further supported this idea and emphasized the biological significance of the fourth initiation methionine in the human *CLN5* gene (see chapter 2.2.2). To further analyze the possibility that alternative AUGs are used, we cloned WT *CLN5* into pGEM3 vector and performed *in vitro* translation analysis in the absence of microsomal membranes. The theoretical molecular weights of polypeptides translated from these alternative start codons are 46.3, 43.4, 41.5 and 40.3 kDa. An *in vitro* translation assay produced protein bands of 47, 44, 42 and 40 kDa, which indicates that upstream AUGs can potentially be used for the translation initiation of the *CLN5* gene. Later on, in another study, it was shown that alternative AUG codons of *CLN5* are used for translation initiation in transiently transfected COS-1 cells (Vesa *et al.*, 2002). In this study, expression of CLN5 was monitored using an antibody, which recognizes the N-terminal part of the protein (Figure 9).

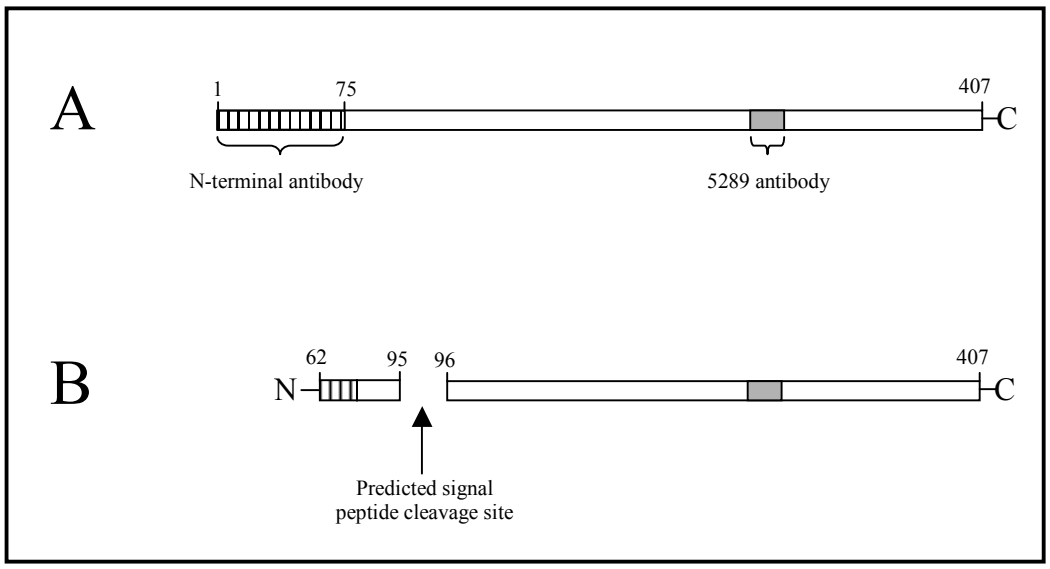

Figure 9. Schematic presentation of CLN5 proteins translated from the first (A) and fourth (B) initiation methionines. Recognition sites of two different antibodies and predicted signal peptide cleavage site are shown. N-terminal antibody is used in Vesa *et al.* (2002) and the 5289 antibody in III.

Interestingly, the N-terminal antibody produced rather different results than the 5289 peptide antibody, which was used in our original study (III). The N-terminal antibody recognized a 47 kDa form of CLN5, which was bound to membranes. Moreover, a great majority of this protein was shown to lack any N-linked glycans. Thus, it is possible that different forms of CLN5 polypeptides might exist in some cells or tissues. At least different forms of CLN5 are expressed *in vitro*, in transient cDNA expression systems. It would be of value to uncover what explains the different results obtained with the two different antibodies, even though the very same expression construct was used in both studies. It is quite clear that the N-terminal antibody is not able to detect CLN5 protein translated from the fourth initiation methionine, if the N-terminal signal sequence is cleaved off as predicted (Figure 9B). It is more difficult to understand, why the 5289 antibody is not able to recognize the 47 kDa polypeptide (Figure 9A). One likely explanation is that the 60 kDa protein is the major cellular form of CLN5 and only low levels of the 47 kDa form are expressed. Consequently, the 5289 antibody can identify the 60 kDa protein, but is not sensitive enough to detect 47 kDa polypeptides.

The most interesting finding obtained with the N-terminal antibody was the interaction of the unglycosylated form of CLN5 with two other NCL proteins,

CLN2 and CLN3 (Vesa *et al.*, 2002). This interaction was demonstrated by an *in vitro* binding assay and coimmunoprecipitation analysis. Previous studies with the yeast two-hybrid system had failed to observe any interactions between NCL proteins (Cottone *et al.*, 2001; Leung *et al.*, 2001; Zhong *et al.*, 2000). It would be interesting to monitor in which tissues and physiological conditions the interacting form of CLN5 is expressed.

3. Expression of PPT1 in developing mouse tissues (IV)

The amino acid sequence of the mouse PPT1 enzyme is highly conserved with the human and rat PPT1 (84.3 and 94.1% identity at the amino acid level) (Salonen *et al.*, 1998). The mouse gene was cloned in 1998 and its expression pattern in transiently transfected COS-1 cells was shown to be highly identical with the human PPT1 (Hellsten *et al.*, 1996; Salonen *et al.*, 1998). Thus, it is expected that mouse would be a good model organism to study INCL disease. Here the expression pattern of PPT1 is analyzed in the developing and adult mouse by Northern blot, *in situ* hybridization and immunohistochemistry. Also Western blot analysis was performed, but our antibodies were not able to detect any specific PPT1 bands in these experiments (data not shown).

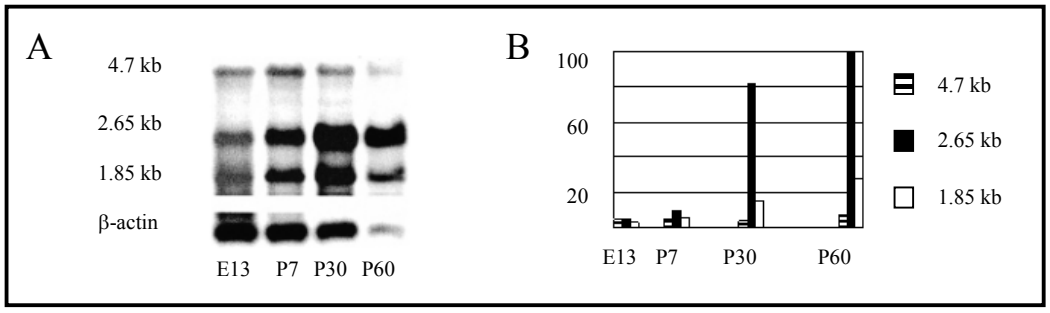

Figure 10. PPT mRNA expression during mouse brain development. (A) Northern Blot analysis. (B) Densitometric analysis of autoradiograph shown in A. E = embryonic day; P = postnatal day. (For more details, see IV).

3.1 Expression of PPT mRNA

To monitor the developmental expression of PPT1 mRNA, Northern analysis of the mouse brain was carried out from embryonal day 13 (E13) to postnatal day 60 (P60). Previously, mouse *PPT1* has been shown to be expressed as a major 2.65 kb and a smaller 1.85 kb transcript in the brain (Salonen *et al.*, 1998). In addition to these transcripts, an additional hybridization signal of 4.7 kb was detected in the present Northern analysis (Figure 10). The most prominent of the three was the 2.65 transcript. Its level of expression was considerably increased between P10 and P30. The increase in the 1.85 transcript paralleled that of the 2.65, but its level was consistently lower. The weak additional 4.7 kb transcript was expressed

at a constant level throughout development. Upregulation of PPT mRNA expression in developing brain has been demonstrated in other studies as well (Suopanki *et al.*, 1999; Zhang *et al.*, 1999).

A progressive increase in the level of *PPT1* mRNA expression was further observed using *in situ* hybridization. The whole-mount *in situ* hybridization analysis revealed that *PPT1* is already expressed at E8, when it was most apparent throughout the neuroepithelium in regions of both the closed neural tube and the open neural folds. From E9 onward, *PPT1* expression is widespread throughout all tissues of the embryo. However, by E11 some regionalization of the signal became apparent in the prospective mid- and hindbrain. This may reflect the more specific expression pattern detected at later developmental stages using *in situ* hybridization on brain sections. The most intense signal in the brain was detected in the neurons of the cerebral cortex in layer II and IV – V, hippocampal CA1 – CA3 pyramidal cells, granule cells of the dentate gyrus, and the hypothalamus (Figure 11). Regionalization of *PPT1* expression in the brain has also been reported in other studies. Especially, high expression of *PPT1* is observed in the hippocampus using brain region specific Northern blot (Zhang *et al.*, 1999) and immunohistochemistry (Suopanki *et al.*, 2002).

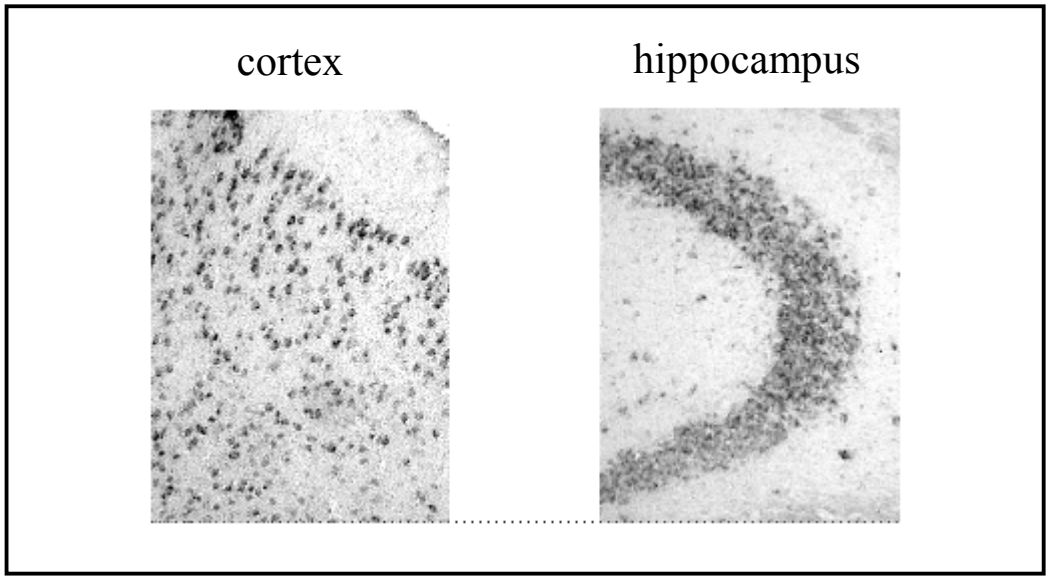

Figure 11. Expression of *PPT1* mRNA in mouse brain revealed by *In situ* hybridization analysis. Expression of PPT1 is seen in cerebral cortex and hippocampal pyramidal cells.

3.2 Expression of PPT1 protein

Immunohistochemical analysis of mouse brain sections with PPT1 specific antibodies revealed that the expression of PPT1 protein was spatiotemporally comparable to that of *PPT1* transcripts. A remarkable increase in immunoreactivity was detected between P5 and P10. PPT1 protein was detected in the cerebral cortex, cerebellum, and in the hippocampus. At P60, PPT1 immunostaining was especially strong in the hippocampus and subiculum. Immunopositive granules were distributed throughout the cell soma and neuronal extensions (Figure 12). These findings are in agreement with the *in vitro* finding from primary neuron cultures that PPT1 is targeted from the cell soma to the neuritis and nerve terminals (Heinonen *et al.*, 2000a).

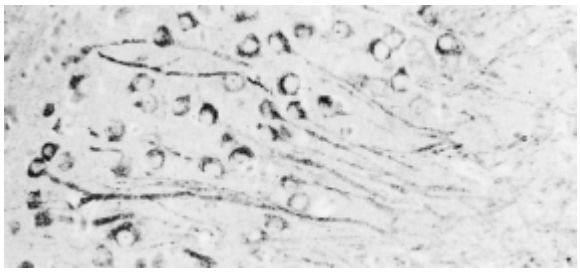

Figure 12. Expression of PPT1 protein in mouse brain revealed by immunohistochemistry. Strong immunoreactivity is evident in neuronal extensions in subiculum.

Expression of PPT1 protein and mRNA was shown to gradually increase with period, when new synaptic contacts are extensively formed in mouse brain (White *et al.*, 1997). Additionally, in embryos a notable increase in *PPT1* mRNA expression was monitored just before the early synaptogenic period (E11 – E14) (Vaughn, 1989). This is well in line with recent data from primary neuron cultures showing that expression of PPT is under developmental control and precedes expression of synaptic markers (Ahtiainen *et al.*, 2003). Also, in a study of human embryonic brains, the expression of PPT1 was shown to increase during brain development (Heinonen *et al.*, 2000b). Based on these findings, it was suggested that PPT1 might have a role for the survival of neural networks, possibly associated with the development and maintenance of the synaptic machinery. In a recent cell fractionation studies PPT1 has been associated with synapses, which further supports this hypothesis (Lehtovirta *et al.*, 2001; Suopanki *et al.*, 2002).

CONCLUDING REMARKS

The neuronal ceroid lipofuscinoses (NCL) are a group of inherited neurodegenerative disorders, frequent in childhood but also present in adulthood. In recent years, modern molecular genetic methods have allowed identification of genes responsible for these disorders. Currently, the defective genes behind the six human NCL diseases are known. They are either lysosomal enzymes or membrane proteins. Identification of mutations has allowed rapid prenatal and carrier diagnostics. However, the pathogenesis of these disorders is still very poorly understood. Biochemical analyses have shown that accumulating material in INCL mostly consists of sphingolipid activator proteins (SAPs) A and D. In other NCL types, subunit c of the mitochondrial ATP synthase complex is the major storage component. However, storage material accumulates in many cell types and it is not understood why neuronal cells of NCL patients are most affected. It remains elusive, whether different NCL proteins are working on the same metabolic cascade or possibly interacting among themselves. There are two naturally occurring mouse models for NCLs (CLN6 and CLN8), and knockout models have been published for CLN1 and CLN3. These animal models provide a valuable tool for further analysis of these disorders. Understanding of the molecular mechanism behind NCL disorders would greatly benefit our understanding of the general biology of the human central nervous system as well.

This study was undertaken to promote the understanding of the pathogenesis of vLINCL (CLN5) and INCL (CLN1) disorders, which are enriched especially in the Finnish population. During the positional cloning of the *CLN5*, we were able to show that the FISH strategy can replace many laborious traditional physical mapping methods. Visual mapping provided direct information on the location of genomic clones and the distances between them. The produced physical map allowed identification of the *CLN5* gene. In transient cell expression analysis we showed that CLN5 is a lysosomal protein. Further, our results indicated that CLN5 is a soluble protein, not an integral transmembrane protein as predicted previously. Defective intracellular transport of the mutant CLN5 suggested that defective lysosomal trafficking could be responsible for the molecular basis of vLINCL.

In order to understand the molecular pathogenesis of NCL disorders, a knowledge of spatiotemporal expression pattern of NCL proteins during development is of great importance. As the most drastic destruction of central nervous system is observed in INCL disorder, we analyzed expression pattern of PPT1 in

developing mouse tissues. We were able to demonstrate a gradual increase in PPT1 expression in developing mouse brain and embryo during a time when new synaptic contacts are being extensively formed. Moreover, the relatively high prevalence of PPT was seen in the neuritic shafts and nerve terminals. These findings indicate that PPT1 might have extra-lysosomal functions, possibly associated with the maintenance of the synaptic machinery.

ACKNOWLEDGEMENTS

This study was carried out at the Department of Molecular Medicine, National Public Health Institute in Helsinki. I wish to thank the head of the Institute, Professor Jussi Huttunen for providing excellent research facilities.

I am grateful to my supervisor Professor Leena Peltonen-Palotie. I have been privileged to work under her guidance. Working in her lab has been a challenging experience.

Professor Jorma Panula and Docent Maija Wessman are acknowledged for carefully reviewing my thesis.

Other members of the CLN5 group, Tuomas Klockars, Minna Savukoski and Ville Holmberg are warmly thanked for forming such a great team. Likewise, I would like to thank Anu Jalanko and Outi Kopra. Their support and guidance have been essential for this study to be completed.

All the other past and present members of the department are thanked for making the lab a good place to work. I would especially like to thank Jani Saarela, Nina Aula and Tarja Salonen for sharing the same writing room in Biomedicum. Kai Tenhunen deserves special thanks for e-mails (not only junk) and relaxing lunch breaks. Jouni Vesa is acknowledged for guidance in the practical laboratory work in the beginning of this study and for collaboration later on. Ismo Ulmanen and Irma Järvelä are thanked for discussions on a variety of matters and Sari Kivikko and Sari Mustala for expert secretarial work.

Professor Aarno Palotie's research group is greatly acknowledged for collaboration on FISH-studies. In addition to the FISH-collaboration, Maris Laan and Nina Horelli-Kuitunen are thanked for their friendship during these years.

Donald Smart is thanked for revising the English of this thesis.

I owe my warmest thanks to my family – Sari, Joona and Veera. Thank you for being there.

This research was financially supported by the Academy of Finland, the Rinnekoti Research Foundation (Espoo, Finland), The Foundation of Pediatric Research (Ulla Hjelt Fund) and Helsinki University Science Fund.

Helsinki, January 2003

Juha Isosomppi

REFERENCES

Aaltonen J., Horelli-Kuitunen N., Fan J. B., Bjorses P., Perheentupa J., Myers R., Palotie A., and Peltonen L. (1997). High-resolution physical and transcriptional mapping of the autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy locus on chromosome 21q22.3 by FISH. *Genome Res* **7:** 820-9.

Ahtiainen L., Van Diggelen O. P., Jalanko A., and Kopra O. (2003). Palmitoyl protein thioesterase 1 is targeted to the axons in neurons. *J Comp Neurol* **455:** 368-77.

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-10.

Andersson S., Davis D. L., Dahlback H., Jornvall H., and Russell D. W. (1989). Cloning, structure, and expression of the mitochondrial cytochrome P- 450 sterol 26-hydroxylase, a bile acid biosynthetic enzyme. *J Biol Chem* **264:** 8222-9.

Andrade M. A., Brown N. P., Leroy C., Hoersch S., de Daruvar A., Reich C., Franchini A., Tamames J., Valencia A., Ouzounis C., and Sander C. (1999). Automated genome sequence analysis and annotation. *Bioinformatics* **15:** 391-412.

Attwood T. K. (2000). The quest to deduce protein function from sequence: the role of pattern databases. *Int J Biochem Cell Biol* **32:** 139-55.

Autti T., Raininko R., Launes J., Nuutila A., and Santavuori P. (1992). Jansky-Bielschowsky variant disease: CT, MRI, and SPECT findings. *Pediatr Neurol* **8:** 121-6.

Bailey J. A., Gu Z., Clark R. A., Reinert K., Samonte R. V., Schwartz S., Adams M. D., Myers E. W., Li P. W., and Eichler E. E. (2002). Recent segmental duplications in the human genome. *Science* **297:** 1003-7.

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

Baker D., and Sali A. (2001). Protein structure prediction and structural genomics. *Science* **294:** 93-6.

Baxevanis A. D. (2002). The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Res* **30:** 1-12.

Beaudoing E., Freier S., Wyatt J. R., Claverie J. M., and Gautheret D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Res* **10:** 1001-10.

Beaudoing E., and Gautheret D. (2001). Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res* **11:** 1520-6.

Bellizzi J. J., 3rd, Widom J., Kemp C., Lu J. Y., Das A. K., Hofmann S. L., and Clardy J. (2000). The crystal structure of palmitoyl protein thioesterase 1 and the molecular basis of infantile neuronal ceroid lipofuscinosis. *Proc Natl Acad Sci U S A* **97:** 4573-8.

Bensimon A., Simon A., Chiffaudel A., Croquette V., Heslot F., and Bensimon D. (1994). Alignment and sensitive detection of DNA by a moving interface. *Science* **265:** 2096-8.

Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Rapp B. A., and Wheeler D. L. (2002). GenBank. *Nucleic Acids Res* **30:** 17-20.

Berkovic S. F., Andermann F., Andermann E., Carpenter S., and Wolfe L. (1988a). Kufs disease: clinical features and forms. *Am J Med Genet Suppl* **5:** 105-9.

Berkovic S. F., Carpenter S., Andermann F., Andermann E., and Wolfe L. S. (1988b). Kufs' disease: a critical reappraisal. *Brain* **111:** 27-62.

Bizzozero O. A., Tetzloff S. U., and Bharadwaj M. (1994). Overview: protein palmitoylation in the nervous system: current views and unsolved problems. *Neurochem Res* **19:** 923-33.

Boehme D. H., Cottrell J. C., Leonberg S. C., and Zeman W. (1971). A dominant form of neuronal ceroid-lipofuscinosis. *Brain* **94:** 745-60.

Boguski M. S., Lowe T. M., and Tolstoshev C. M. (1993). dbEST--database for "expressed sequence tags". *Nat Genet* **4:** 332-3.

Bork P. (2000). Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res* **10:** 398-400.

Bork P., and Copley R. (2001). The draft sequences. Filling in the gaps. *Nature* **409:** 818-20.

Bork P., Dandekar T., Diaz-Lazcoz Y., Eisenhaber F., Huynen M., and Yuan Y. (1998). Predicting function: from genes to genomes and back. *J Mol Biol* **283:** 707-25.

Brett D., Hanke J., Lehmann G., Haase S., Delbruck S., Krueger S., Reich J., and Bork P. (2000). EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett* **474:** 83-6.

Brett D., Pospisil H., Valcarcel J., Reich J., and Bork P. (2002). Alternative splicing and genome complexity. *Nat Genet* **30:** 29-30.

Burke D. T., Carle G. F., and Olson M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806-12.

Camp L. A., and Hofmann S. L. (1993). Purification and properties of a palmitoyl-protein thioesterase that cleaves palmitate from H-Ras. *J Biol Chem* **268:** 22566-74.

Camp L. A., Verkruyse L. A., Afendis S. J., Slaughter C. A., and Hofmann S. L. (1994). Molecular cloning and expression of palmitoyl-protein thioesterase. *J Biol Chem* **269:** 23212-9.

Cho S., Dawson P. E., and Dawson G. (2000). In vitro depalmitoylation of neurospecific peptides: implication for infantile neuronal ceroid lipofuscinosis. *J Neurosci Res* **59:** 32-8.

Claros M. G., and Vincens P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem* **241:** 779-86.

Claverie J. M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* **6:** 1735-44.

Claverie J. M. (2000). From bioinformatics to computational biology. *Genome Res* **10:** 1277-9.

Collins F. S. (1992). Positional cloning: let's not call it reverse anymore. *Nat Genet* **1:** 3-6.

Collins F. S. (1995). Positional cloning moves from perditional to traditional. *Nat Genet* **9:** 347-50.

Collins J., and Hohn B. (1978). Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc Natl Acad Sci U S A* **75:** 4242-6.

Cottone C. D., Chattopadhyay S., and Pearce D. A. (2001). Searching for interacting partners of CLN1, CLN2 and Btn1p with the two- hybrid system. *Europ J Paediatr Neurol* **5:** 95-8.

Cox D. R., Burmeister M., Price E. R., Kim S., and Myers R. M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250:** 245-50.

Cuozzo J. W., Tao K., Cygler M., Mort J. S., and Sahagian G. G. (1998). Lysine-based structure responsible for selective mannose phosphorylation of cathepsin D and cathepsin L defines a common structural motif for lysosomal enzyme targeting. *J Biol Chem* **273:** 21067-76.

Das A. K., Lu J. Y., and Hofmann S. L. (2001). Biochemical analysis of mutations in palmitoyl-protein thioesterase causing infantile and late-onset forms of neuronal ceroid lipofuscinosis. *Hum Mol Genet* **10:** 1431-9.

Diaz E., and Pfeffer S. R. (1998). TIP47: a cargo selection device for mannose 6-phosphate receptor trafficking. *Cell* **93:** 433-43.

Doray B., Ghosh P., Griffith J., Geuze H. J., and Kornfeld S. (2002). Cooperation of GGAs and AP-1 in packaging MPRs at the trans-Golgi network. *Science* **297:** 1700-3.

Eichler E. E. (2001). Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17:** 661-9.

Eisenberg D., Marcotte E. M., Xenarios I., and Yeates T. O. (2000). Protein function in the post-genomic era. *Nature* **405:** 823-6.

Eisenhaber F., and Bork P. (1998). Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol* **8:** 169-70.

Emanuelsson O., Nielsen H., Brunak S., and von Heijne G. (2000). Predicting subcellular localization of proteins based on their N- terminal amino acid sequence. *J Mol Biol* **300:** 1005-16.

Emanuelsson O., and von Heijne G. (2001). Prediction of organellar targeting signals. *Biochim Biophys Acta* **1541:** 114-9.

Enright A. J., Iliopoulos I., Kyrpides N. C., and Ouzounis C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402:** 86-90.

Fickett J. W., and Hatzigeorgiou A. G. (1997). Eukaryotic promoter recognition. *Genome Res* **7:** 861-78.

Fickett J. W., and Wasserman W. W. (2000). Discovery and modeling of transcriptional regulatory regions. *Curr Opin Biotechnol* **11:** 19-24.

Fidlerova H., Senger G., Kost M., Sanseau P., and Sheer D. (1994). Two simple procedures for releasing chromatin from routinely fixed cells for fluorescence in situ hybridization. *Cytogenet Cell Genet* **65:** 203-5.

Fortna A., and Gardiner K. (2001). Genomic sequence analysis tools: a user's guide. *Trends Genet* **17:** 158-64.

Fuchs R. (2000). Analyse this...or: intelligent help for the rest of us. *Bioinformatics* **16:** 491-3.

Gao H., Boustany R. M., Espinola J. A., Cotman S. L., Srinidhi L., Antonellis K. A., Gillis T., Qin X., Liu S., Donahue L. R., Bronson R. T., Faust J. R., Stout D., Haines J. L., Lerner T. J., and MacDonald M. E. (2002). Mutations in a Novel CLN6-Encoded Transmembrane Protein Cause Variant Neuronal Ceroid Lipofuscinosis in Man and Mouse. *Am J Hum Genet* **70:** 324-35.

Gardiner M., Sandford A., Deadman M., Poulton J., Cookson W., Reeders S., Jokiaho I., Peltonen L., Eiberg H., and Julier C. (1990). Batten disease (Spielmeyer-Vogt disease, juvenile onset neuronal ceroid- lipofuscinosis) gene (CLN3) maps to human chromosome 16. *Genomics* **8:** 387-90.

Gautheret D., Poirot O., Lopez F., Audic S., and Claverie J. M. (1998). Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res* **8:** 524-30.

Gervasini C., Bentivegna A., Venturin M., Corrado L., Larizza L., and Riva P. (2002). Tandem duplication of the NF1 gene detected by high-resolution FISH in the 17q11.2 region. *Hum Genet* **110:** 314-21.

Giglio S., Broman K. W., Matsumoto N., Calvari V., Gimelli G., Neumann T., Ohashi H., Voullaire L., Larizza D., Giorda R., Weber J. L., Ledbetter D. H., and Zuffardi O. (2001). Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet* **68:** 874-83.

Glickman J. N., and Kornfeld S. (1993). Mannose 6-phosphate-independent targeting of lysosomal enzymes in I- cell disease B lymphoblasts. *J Cell Biol* **123:** 99-108.

Goebel H. H. (1995). The neuronal ceroid-lipofuscinoses. *J Child Neurol* **10:** 424-37.

Graves P. R., and Haystead T. A. (2002). Molecular biologist's guide to proteomics. *Microbiol Mol Biol Rev* **66:** 39-63; table of contents.

Green E. D., Riethman H. C., Dutchik J. E., and Olson M. V. (1991). Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* **11:** 658-69.

Gupta P., Soyombo A. A., Atashband A., Wisniewski K. E., Shelton J. M., Richardson J. A., Hammer R. E., and Hofmann S. L. (2001). Disruption of PPT1 or PPT2 causes neuronal ceroid lipofuscinosis in knockout mice. *Proc Natl Acad Sci U S A* **98:** 13566-71.

Gupta R., Birch H., Rapacki K., Brunak S., and Hansen J. E. (1999). O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* **27:** 370-2.

Gyapay G., Morissette J., Vignal A., Dib C., Fizames C., Millasseau P., Marc S., Bernardi G., Lathrop M., and Weissenbach J. (1994). The 1993-94 Genethon human genetic linkage map. *Nat Genet* **7:** 246-339.

Haaf T., and Ward D. C. (1994a). High resolution ordering of YAC contigs using extended chromatin and chromosomes. *Hum Mol Genet* **3:** 629-33.

Haaf T., and Ward D. C. (1994b). Structural analysis of alpha-satellite DNA and centromere proteins using extended chromatin and chromosomes. *Hum Mol Genet* **3:** 697-709.

Hall N. A., Lake B. D., Dewji N. N., and Patrick A. D. (1991). Lysosomal storage of subunit c of mitochondrial ATP synthase in Batten's disease (ceroid-lipofuscinosis). *Biochem J* **275:** 269-72.

Haltia M., Rapola J., and Santavuori P. (1973). Infantile type of so-called neuronal ceroid-lipofuscinosis. Histological and electron microscopic studies. *Acta Neuropathol (Berl)* **26:** 157-70.

Hansen J. E., Lund O., Tolstrup N., Gooley A. A., Williams K. L., and Brunak S. (1998). NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj J* **15:** 115-30.

Harris N. L. (1997). Genotator: a workbench for sequence annotation. *Genome Res* **7:** 754-62.

Hästbacka J., de la Chapelle A., Kaitila I., Sistonen P., Weaver A., and Lander E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat Genet* **2:** 204-11.

Hästbacka J., de la Chapelle A., Mahtani M. M., Clines G., Reeve-Daly M. P., Daly M., Hamilton B. A., Kusumi K., Trivedi B., Weaver A., and et al. (1994). The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78:** 1073-87.

Heinonen O., Kyttala A., Lehmus E., Paunio T., Peltonen L., and Jalanko A. (2000a). Expression of palmitoyl protein thioesterase in neurons. *Mol Genet Metab* **69:** 123-9.

Heinonen O., Salonen T., Jalanko A., Peltonen L., and Copp A. (2000b). CLN-1 and CLN-5, genes for infantile and variant late infantile neuronal ceroid lipofuscinoses, are expressed in the embryonic human brain. *J Comp Neurol* **426:** 406-12.

Heiskanen M., Hellsten E., Kallioniemi O. P., Makela T. P., Alitalo K., Peltonen L., and Palotie A. (1995). Visual mapping by fiber-FISH. *Genomics* **30:** 31-6.

Heiskanen M., Kallioniemi O., and Palotie A. (1996a). Fiber-FISH: experiences and a refined protocol. *Genet Anal* **12:** 179-84.

Heiskanen M., Karhu R., Hellsten E., Peltonen L., Kallioniemi O. P., and Palotie A. (1994). High resolution mapping using fluorescence in situ hybridization to extended DNA fibers prepared from agarose-embedded cells. *Biotechniques* **17:** 928-9, 932-3.

Heiskanen M., Peltonen L., and Palotif A. (1996b). Visual mapping by high resolution FISH. *Trends Genet* **12:** 379-82.

Helenius A., and Aebi M. (2001). Intracellular functions of N-linked glycans. *Science* **291:** 2364-9.

Hellsten E., Vesa J., Olkkonen V. M., Jalanko A., and Peltonen L. (1996). Human palmitoyl protein thioesterase: evidence for lysosomal targeting of the enzyme and disturbed cellular routing in infantile neuronal ceroid lipofuscinosis. *Embo J* **15:** 5240-5.

Hellsten E., Vesa J., Speer M. C., Makela T. P., Järvelä I., Alitalo K., Ott J., and Peltonen L. (1993). Refined assignment of the infantile neuronal ceroid lipofuscinosis (INCL, CLN1) locus at 1p32: incorporation of linkage disequilibrium in multipoint analysis. *Genomics* **16:** 720-5.

Heng H. H., Spyropoulos B., and Moens P. B. (1997). FISH technology in chromosome and genome research. *Bioessays* **19:** 75-84.

Heng H. H., Squire J., and Tsui L. C. (1992). High-resolution mapping of mammalian genes by in situ hybridization to free chromatin. *Proc Natl Acad Sci U S A* **89:** 9509-13.

Heng H. H., and Tsui L. C. (1998). High resolution free chromatin/DNA fiber fluorescent in situ hybridization. *J Chromatogr A* **806:** 219-29.

Hirokawa T., Boon-Chieng S., and Mitaku S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14:** 378-9.

Hofmann S. L., Das A. K., Yi W., Lu J. Y., and Wisniewski K. E. (1999). Genotype-phenotype correlations in neuronal ceroid lipofuscinosis due to palmitoyl-protein thioesterase deficiency. *Mol Genet Metab* **66:** 234-9.

Holmberg V., Lauronen L., Autti T., Santavuori P., Savukoski M., Uvebrant P., Hofman I., Peltonen L., and Järvelä I. (2000). Phenotype-genotype correlation in eight patients with Finnish variant late infantile NCL (CLN5). *Neurology* **55:** 579-81.

Hopman A. H., Wiegant J., Raap A. K., Landegent J. E., van der Ploeg M., and van Duijn P. (1986). Bi-color detection of two target DNAs by non-radioactive in situ hybridization. *Histochemistry* **85:** 1-4.

Horelli-Kuitunen N., Aaltonen J., Yaspo M. L., Eeva M., Wessman M., Peltonen L., and Palotie A. (1999). Mapping ESTs by fiber-FISH. *Genome Res* **9:** 62-71.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyras E., Gilbert J., Hammond M., Huminiecki L., Kasprzyk A., Lehvaslaiho H., Lijnzaad P., Melsopp C., Mongin E., Pettett R., Pocock M., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I., and Clamp M. (2002). The Ensembl genome database project. *Nucleic Acids Res* **30:** 38-41.

Hunziker W., and Geuze H. J. (1996). Intracellular trafficking of lysosomal membrane proteins. *Bioessays* **18:** 379-89.

Inazawa J., Ariyama T., Tokino T., Tanigami A., Nakamura Y., and Abe T. (1994). High resolution ordering of DNA markers by multi-color fluorescent in situ hybridization of prophase chromosomes. *Cytogenet Cell Genet* **65:** 130-5.

Ioannou P. A., Amemiya C. T., Garnes J., Kroisel P. M., Shizuya H., Chen C., Batzer M. A., and de Jong P. J. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nat Genet* **6:** 84-9.

Järvelä I., Lehtovirta M., Tikkanen R., Kyttala A., and Jalanko A. (1999). Defective intracellular transport of CLN3 is the molecular basis of Batten disease (JNCL). *Hum Mol Genet* **8:** 1091-8.

Järvelä I., Sainio M., Rantamaki T., Olkkonen V. M., Carpen O., Peltonen L., and Jalanko A. (1998). Biosynthesis and intracellular targeting of the CLN3 protein defective in Batten disease. *Hum Mol Genet* **7:** 85-90.

Järvelä I., Schleutker J., Haataja L., Santavuori P., Puhakka L., Manninen T., Palotie A., Sandkuijl L. A., Renlund M., White R., and et al. (1991). Infantile form of neuronal ceroid lipofuscinosis (CLN1) maps to the short arm of chromosome 1. *Genomics* **9:** 170-3.

Jensen L. J., Gupta R., Blom N., Devos D., Tamames J., Kesmir C., Nielsen H., Staerfeldt H. H., Rapacki K., Workman C., Andersen C. A., Knudsen S., Krogh A., Valencia A., and Brunak S. (2002). Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol* **319:** 1257-65.

John H. A., Birnstiel M. L., and Jones K. W. (1969). RNA-DNA hybrids at the cytological level. *Nature* **223:** 582-7.

Katz M. L., Gao C. L., Prabhakaram M., Shibuya H., Liu P. C., and Johnson G. S. (1997). Immunochemical localization of the Batten disease (CLN3) protein in retina. *Invest Ophthalmol Vis Sci* **38:** 2375-86.

Kent W. J., Sugnet C. W., Furey T. S., Roskin K. M., Pringle T. H., Zahler A. M., and Haussler D. (2002). The human genome browser at UCSC. *Genome Res* **12:** 996-1006. [doi].

Kirchhausen T. (1999). Adaptors for clathrin-mediated traffic. *Annu Rev Cell Dev Biol* **15:** 705-32.

Klockars T., Holmberg V., Savukoski M., Lander E. S., and Peltonen L. (1999). Transcript identification on the CLN5 region on chromosome 13q22. *Hum Genet* **105:** 51-6.

Klockars T., Savukoski M., Isosomppi J., Laan M., Järvelä I., Petrukhin K., Palotie A., and Peltonen L. (1996). Efficient construction of a physical map by fiber-FISH of the CLN5 region: refined assignment and long-range contig covering the critical region on 13q22. *Genomics* **35:** 71-8.

Kornfeld S. (1990). Lysosomal enzyme targeting. *Biochem Soc Trans* **18:** 367-74.

Kornfeld S., and Mellman I. (1989). The biogenesis of lysosomes. *Annu Rev Cell Biol* **5:** 483-525.

Kozak M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15:** 8125-48.

Kozak M. (1995). Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. *Proc Natl Acad Sci U S A* **92:** 2662-6.

Kozak M. (1996). Interpreting cDNA sequences: some insights from studies on translation. *Mamm Genome* **7:** 563-74.

Krause A., Haas S. A., Coward E., and Vingron M. (2002). SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. *Nucleic Acids Res* **30:** 299-300.

Kremmidiotis G., Lensink I. L., Bilton R. L., Woollatt E., Chataway T. K., Sutherland G. R., and Callen D. F. (1999). The Batten disease gene product (CLN3p) is a Golgi integral membrane protein. *Hum Mol Genet* **8:** 523-31.

Krogh A., Larsson B., von Heijne G., and Sonnhammer E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305:** 567-80.

Kyte J., and Doolittle R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157:** 105-32.

Laan M., Kallioniemi O. P., Hellsten E., Alitalo K., Peltonen L., and Palotie A. (1995). Mechanically stretched chromosomes as targets for high-resolution FISH mapping. *Genome Res* **5:** 13-20.

Lambert C., Leonard N., De Bolle X., and Depiereux E. (2002). ESyPred3D: Prediction of proteins 3D structures. *Bioinformatics* **18:** 1250-6.

Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J. P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J. C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R. H., Wilson R. K., Hillier L. W., McPherson J. D., Marra M. A., Mardis E. R., Fulton L. A., Chinwalla A. T., Pepin K. H., Gish W. R., Chissoe S. L., Wendl M. C., Delehaunty K. D., Miner T. L., Delehaunty A., Kramer J. B., Cook L. L., Fulton R. S., Johnson D. L., Minx P. J., Clifton S. W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J. F., Olsen A., Lucas S.,

Elkin C., Uberbacher E., Frazier M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lawrence J. B., Singer R. H., and McNeil J. A. (1990). Interphase and metaphase resolution of different distances within the human dystrophin gene. *Science* **249:** 928-32.

Le Borgne R., Alconada A., Bauer U., and Hoflack B. (1998). The mammalian AP-3 adaptor-like complex mediates the intracellular transport of lysosomal membrane glycoproteins. *J Biol Chem* **273:** 29451-61.

Le Borgne R., and Hoflack B. (1998). Protein transport from the secretory to the endocytic pathway in mammalian cells. *Biochim Biophys Acta* **1404:** 195-209.

Lebo R. V., Lynch E. D., Bird T. D., Golbus M. S., Barker D. F., O'Connell P., and Chance P. F. (1992). Multicolor in situ hybridization and linkage analysis order Charcot- Marie-Tooth type I (CMTIA) gene-region markers. *Am J Hum Genet* **50:** 42-55.

Lehtovirta M., Kyttala A., Eskelinen E. L., Hess M., Heinonen O., and Jalanko A. (2001). Palmitoyl protein thioesterase (PPT) localizes into synaptosomes and synaptic vesicles in neurons: implications for infantile neuronal ceroid lipofuscinosis (INCL). *Hum Mol Genet* **10:** 69-75.

Leppanen P., Isosomppi J., Schleutker J., Aula P., and Peltonen L. (1996). A physical map of the 6q14-q15 region harboring the locus for the lysosomal membrane sialic acid transport defect. *Genomics* **37:** 62-7.

Leung K. Y., Greene N. D., Munroe P. B., and Mole S. E. (2001). Analysis of CLN3-protein interactions using the yeast two-hybrid system. *Europ J Paediatr Neurol* **5:** 89-93.

Lewis S., Ashburner M., and Reese M. G. (2000). Annotating eukaryote genomes. *Curr Opin Struct Biol* **10:** 349-54.

Lichter P., Tang C. J., Call K., Hermanson G., Evans G. A., Housman D., and Ward D. C. (1990). High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247:** 64-9.

Lichter P., and Ward D. C. (1990). Is non-isotopic in situ hybridization finally coming of age? *Nature* **345:** 93-4.

Liu R., and States D. J. (2002). Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res* **12:** 462-9.

Lonka L., Kyttala A., Ranta S., Jalanko A., and Lehesjoki A. E. (2000). The neuronal ceroid lipofuscinosis CLN8 membrane protein is a resident of the endoplasmic reticulum. *Hum Mol Genet* **9:** 1691-7.

Lu J. Y., Verkruyse L. A., and Hofmann S. L. (1996). Lipid thioesters derived from acylated proteins accumulate in infantile neuronal ceroid lipofuscinosis: correction of the defect in lymphoblasts by recombinant palmitoyl-protein thioesterase. *Proc Natl Acad Sci U S A* **93:** 10046-50.

Lu J.-Y., Verkruyse L. A., and Hofmann S. L. (2002). The effects of lysosomotropic agents on normal and INCL cells provide further evidence for the lysosomal nature of palmitoyl-protein thioesterase function. *Biochim Biophys Acta* **in press**.

Ludwig L., Le Borgne R., and Hoflack B. (1995). Roles for mannose-6-phosphate receptors in lysosomal enzyme sorting, IGF-II binding and clathrin-coat assembly. *Trends in Cell Biology* **5:** 202-206.

Luiro K., Kopra O., Lehtovirta M., and Jalanko A. (2001). CLN3 protein is targeted to neuronal synapses but excluded from synaptic vesicles: new clues to Batten disease. *Hum Mol Genet* **10:** 2123-2131.

Luke S., and Shepelsky M. (1998). FISH: recent advances and diagnostic aspects. *Cell Vis* **5:** 49-53.

Marcotte E. M., Xenarios I., van Der Bliek A. M., and Eisenberg D. (2000). Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A* **97:** 12115-20.

Margraf L. R., Boriack R. L., Routheut A. A., Cuppen I., Alhilali L., Bennett C. J., and Bennett M. J. (1999). Tissue expression and subcellular localization of CLN3, the Batten disease protein. *Mol Genet Metab* **66:** 283-9.

Marks M. S., Ohno H., Kirchhausen T., and Bonifacino J. S. (1997). Protein sorting by tyrosine-based signals: adapting to the Ys and wherefores. *Trends Cell Biol.* **7:** 124-128.

Martin J. J., Libert J., and Ceuterick C. (1987). Ultrastructure of brain and retina in Kufs' disease (adult type-ceroid- lipofuscinosis). *Clin Neuropathol* **6:** 231-5.

McGuffin L. J., Bryson K., and Jones D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16:** 404-5.

Mironov A. A., Fickett J. W., and Gelfand M. S. (1999). Frequent alternative splicing of human genes. *Genome Res* **9:** 1288-93.

Mitchell W. A., Wheeler R. B., Sharp J. D., Bate S. L., Gardiner R. M., Ranta U. S., Lonka L., Williams R. E., Lehesjoki A. E., and Mole S. E. (2001). Turkish variant late infantile neuronal ceroid lipofuscinosis (CLN7) may be allelic to CLN8. *Europ J Paediatr Neurol* **5:** 21-7.

Mitchison H. M., Hofmann S. L., Becerra C. H., Munroe P. B., Lake B. D., Crow Y. J., Stephenson J. B., Williams R. E., Hofman I. L., Taschner P. E., Martin J. J., Philippart M., Andermann E., Andermann F., Mole S. E., Gardiner R. M., and O'Rawe A. M. (1998). Mutations in the palmitoyl-protein thioesterase gene (PPT; CLN1) causing juvenile neuronal ceroid lipofuscinosis with granular osmiophilic deposits. *Hum Mol Genet* **7:** 291-7.

Mitchison H. M., and Mole S. E. (2001). Neurodegenerative disease: the neuronal ceroid lipofuscinoses (Batten disease). *Curr Opin Neurol* **14:** 795-803.

Modrek B., and Lee C. (2002). A genomic view of alternative splicing. *Nat Genet* **30:** 13-9.

Modrek B., Resch A., Grasso C., and Lee C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* **29:** 2850-9.

Mole S. E. (1998). Batten disease: four genes and still counting. *Neurobiol Dis* **5:** 287-303.

Mole S. E., Zhong N. A., Sarpong A., Logan W. P., Hofmann S., Yi W., Franken P. F., van Diggelen O. P., Breuning M. H., Moroziewicz D., Ju W., Salonen T., Holmberg V., Järvelä I., and Taschner P. E. (2001). New mutations in the neuronal ceroid lipofuscinosis genes. *Europ J Paediatr Neurol* **5:** 7-10.

Moller S., Croning M. D., and Apweiler R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17:** 646-53.

Mott R., Schultz J., Bork P., and Ponting C. P. (2002). Predicting protein cellular localization using a domain projection method. *Genome Res* **12:** 1168-74.

Nakai K. (2001). Review: prediction of in vivo fates of proteins in the era of genomics and proteomics. *J Struct Biol* **134:** 103-16.

Nakai K., and Horton P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24:** 34-6.

Nielsen H., Engelbrecht J., Brunak S., and von Heijne G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10:** 1-6.

Nikali K., Isosomppi J., Lonnqvist T., Mao J. I., Suomalainen A., and Peltonen L. (1997). Toward cloning of a novel ataxia gene: refined assignment and physical map of the IOSCA locus (SCA8) on 10q24. *Genomics* **39:** 185-91.

Nolo R., Kaksonen M., and Rauvala H. (1996). Developmentally regulated neurite outgrowth response from dorsal root ganglion neurons to heparin-binding growth-associated molecule (HB-GAM) and the expression of HB-GAM in the targets of the developing dorsal root ganglion neurites. *Eur J Neurosci* **8:** 1658-65.

Norman R., and Wood N. (1941). Congenital form of amaurotic family idocy. *J Neur Psychiat* **4:** 175-190.

Olson M., Hood L., Cantor C., and Botstein D. (1989). A common language for physical mapping of the human genome. *Science* **245:** 1434-5.

Ott J., and Bhat A. (1999). Linkage analysis in heterogeneous and complex traits. *Eur Child Adolesc Psychiatry* **8:** 43-6.

Overbeek R., Fonstein M., D'Souza M., Pusch G. D., and Maltsev N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96:** 2896-901.

Paavola P., Avela K., Horelli-Kuitunen N., Barlund M., Kallioniemi A., Idanheimo N., Kyttala M., de la Chapelle A., Palotie A., Lehesjoki A. E., and Peltonen L. (1999). High-resolution physical and genetic mapping of the critical region for Meckel syndrome and Mulibrey Nanism on chromosome 17q22-q23. *Genome Res* **9:** 267-76.

Pardue M. L., and Gall J. G. (1969). Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc Natl Acad Sci U S A* **64:** 600-4.

Parodi A. J. (2000). Protein glucosylation and its role in protein folding. *Annu Rev Biochem* **69:** 69-93.

Parra I., and Windle B. (1993). High resolution visual mapping of stretched DNA by fluorescent hybridization. *Nat Genet* **5:** 17-21.

Pearce D. A., Ferea T., Nosel S. A., Das B., and Sherman F. (1999). Action of BTN1, the yeast orthologue of the gene mutated in Batten disease. *Nat Genet* **22:** 55-8.

Pearson W. R., and Lipman D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85:** 2444-8.

Pellegrini M., Marcotte E. M., Thompson M. J., Eisenberg D., and Yeates T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96:** 4285-8.

Peltonen L., Pekkarinen P., and Aaltonen J. (1995). Messages from an isolate: lessons from the Finnish gene pool. *Biol Chem Hoppe Seyler* **376:** 697-704.

Peltonen L., Savukoski M., and Vesa J. (2000). Genetics of the neuronal ceroid lipofuscinoses. *Curr Opin Genet Dev* **10:** 299-305.

Peri S., Ibarrola N., Blagoev B., Mann M., and Pandey A. (2001). Common pitfalls in bioinformatics-based analyses: look before you leap. *Trends Genet* **17:** 541-5.

Peri S., and Pandey A. (2001). A reassessment of the translation initiation codon in vertebrates. *Trends Genet* **17:** 685-7.

Pinkel D., Straume T., and Gray J. W. (1986). Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proc Natl Acad Sci U S A* **83:** 2934-8.

Praz V., Perier R., Bonnard C., and Bucher P. (2002). The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res* **30:** 322-4.

Proceedings of the 8th International Congress on the neuronal ceroid lipofuscinosis (2000). Mitchison H. M. and Mole S. E. (eds.)

Raap A. K., van de Corput M. P., Vervenne R. A., van Gijlswijk R. P., Tanke H. J., and Wiegant J. (1995). Ultra-sensitive FISH using peroxidase-mediated deposition of biotin- or fluorochrome tyramides. *Hum Mol Genet* **4:** 529-34.

Ranta S., Zhang Y., Ross B., Lonka L., Takkunen E., Messer A., Sharp J., Wheeler R., Kusumi K., Mole S., Liu W., Soares M. B., Bonaldo M. F., Hirvasniemi A., de la Chapelle A., Gilliam T. C., and Lehesjoki A. E. (1999). The neuronal ceroid lipofuscinoses in human EPMR and mnd mutant mice are associated with mutations in CLN8. *Nat Genet* **23:** 233-6.

Rapola J. (1993). Neuronal ceroid-lipofuscinoses in childhood. *Perspect Pediatr Pathol* **17:** 7-44.

Reese M. G., Hartzell G., Harris N. L., Ohler U., Abril J. F., and Lewis S. E. (2000). Genome annotation assessment in Drosophila melanogaster. *Genome Res* **10:** 483-501.

Rigoutsos I., and Floratos A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14:** 55-67.

Rigoutsos I., Huynh T., Floratos A., Parida L., and Platt D. (2002). Dictionary-driven protein annotation. *Nucleic Acids Res* **30:** 3901-16.

Rijnboutt S., Kal A. J., Geuze H. J., Aerts H., and Strous G. J. (1991). Mannose 6-phosphate-independent targeting of cathepsin D to lysosomes in HepG2 cells. *J Biol Chem* **266:** 23586-92.

Rogozin I. B., Kochetov A. V., Kondrashov F. A., Koonin E. V., and Milanesi L. (2001). Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics* **17:** 890-900.

Salonen T., Heinonen-Kopra O., Vesa J., and Jalanko A. (2001). Neuronal trafficking of palmitoyl protein thioesterase provides an excellent model to study the effects of different mutations which cause infantile neuronal ceroid lipofuscinocis. *Mol Cell Neurosci* **18:** 131-40.

Salonen T., Hellsten E., Horelli-Kuitunen N., Peltonen L., and Jalanko A. (1998). Mouse palmitoyl protein thioesterase: gene structure and expression of cDNA. *Genome Res* **8:** 724-30.

Santavuori P. (1988). Neuronal ceroid-lipofuscinoses in childhood. *Brain Dev* **10:** 80-3.

Santavuori P., Haltia M., and Rapola J. (1974). Infantile type of so-called neuronal ceroid-lipofuscinosis. *Dev Med Child Neurol* **16:** 644-53.

Santavuori P., Rapola J., Nuutila A., Raininko R., Lappi M., Launes J., Herva R., and Sainio K. (1991). The spectrum of Jansky-Bielschowsky disease. *Neuropediatrics* **22:** 92-6.

Santavuori P., Rapola J., Sainio K., and Raitta C. (1982). A variant of Jansky-Bielschowsky disease. *Neuropediatrics* **13:** 135-41.

Savukoski M., Kestila M., Williams R., Järvelä I., Sharp J., Harris J., Santavuori P., Gardiner M., and Peltonen L. (1994). Defined chromosomal assignment of CLN5 demonstrates that at least four genetic loci are involved in the pathogenesis of human ceroid lipofuscinoses. *Am J Hum Genet* **55:** 695-701.

Savukoski M., Klockars T., Holmberg V., Santavuori P., Lander E. S., and Peltonen L. (1998). CLN5, a novel gene encoding a putative transmembrane protein mutated in Finnish variant late infantile neuronal ceroid lipofuscinosis. *Nat Genet* **19:** 286-8.

Schriner J. E., Yi W., and Hofmann S. L. (1996). cDNA and genomic cloning of human palmitoyl-protein thioesterase (PPT), the enzyme defective in infantile neuronal ceroid lipofuscinosis (Erratum appears in Genomics 1996 Dec 15;38(3):458). *Genomics* **34:** 317-22.

Schwartz D. C., and Cantor C. R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37:** 67-75.

Senger G., Jones T. A., Fidlerova H., Sanseau P., Trowsdale J., Duff M., and Sheer D. (1994). Released chromatin: linearized DNA for high resolution fluorescence in situ hybridization. *Hum Mol Genet* **3:** 1275-80.

Sharp J. D., Wheeler R. B., Lake B. D., Savukoski M., Järvelä I. E., Peltonen L., Gardiner R. M., and Williams R. E. (1997). Loci for classical and a variant late infantile neuronal ceroid lipofuscinosis map to chromosomes 11p15 and 15q21-23. *Hum Mol Genet* **6:** 591-5.

Shizuya H., Birren B., Kim U. J., Mancino V., Slepak T., Tachiiri Y., and Simon M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89:** 8794-7.

Sleat D. E., Donnelly R. J., Lackland H., Liu C. G., Sohar I., Pullarkat R. K., and Lobel P. (1997). Association of mutations in a lysosomal protein with classical late- infantile neuronal ceroid lipofuscinosis. *Science* **277:** 1802-5.

Sleat D. E., Sohar I., Lackland H., Majercak J., and Lobel P. (1996). Rat brain contains high levels of mannose-6-phosphorylated glycoproteins including lysosomal enzymes and palmitoyl-

protein thioesterase, an enzyme implicated in infantile neuronal lipofuscinosis. *J Biol Chem* **271:** 19191-8.

Sonnhammer E. L., von Heijne G., and Krogh A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6:** 175-82.

Speel E. J., Ramaekers F. C., and Hopman A. H. (1997). Sensitive multicolor fluorescence in situ hybridization using catalyzed reporter deposition (CARD) amplification. *J Histochem Cytochem* **45:** 1439-46.

Sreekumar K. R., Aravind L., and Koonin E. V. (2001). Computational analysis of human disease-associated genes and their protein products. *Curr Opin Genet Dev* **11:** 247-57.

Sternberg N. (1990). Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc Natl Acad Sci U S A* **87:** 103-7.

Suopanki J., Lintunen M., Lahtinen H., Haltia M., Panula P., Baumann M., and Tyynelä J. (2002). Status epilepticus induces changes in the expression and localization of endogenous palmitoyl-protein thioesterase 1. *Neurobiol Dis* **10:** 247.

Suopanki J., Tyynelä J., Baumann M., and Haltia M. (1999). Palmitoyl-protein thioesterase, an enzyme implicated in neurodegeneration, is localized in neurons and is developmentally regulated in rat brain. *Neurosci Lett* **265:** 53-6.

Tahvanainen E., Ranta S., Hirvasniemi A., Karila E., Leisti J., Sistonen P., Weissenbach J., Lehesjoki A. E., and de la Chapelle A. (1994). The gene for a recessively inherited human childhood progressive epilepsy with mental retardation maps to the distal short arm of chromosome 8. *Proc Natl Acad Sci U S A* **91:** 7267-70.

Taudien S., Rump A., Platzer M., Drescher B., Schattevoy R., Gloeckner G., Dette M., Baumgart C., Weber J., Menzel U., and Rosenthal A. (2000). RUMMAGE--a high-throughput sequence annotation system. *Trends Genet* **16:** 519-20.

The International Batten Disease Consortium (1995). Isolation of a novel gene underlying Batten disease, CLN3. *Cell* **82**: 949-57.

The Neuronal Ceroid Lipofuscinosis (Batten Disease) (1999). Goebel H. H., Mole S. E., Lake B. D. (eds.). IOS press, Amsterdam.

Tikkanen R., Peltola M., Oinonen C., Rouvinen J., and Peltonen L. (1997). Several cooperating binding sites mediate the interaction of a lysosomal enzyme with phosphotransferase. *Embo J* **16:** 6684-93.

Trask B. J. (1991). Fluorescence in situ hybridization: applications in cytogenetics and gene mapping. *Trends Genet* **7:** 149-54.

Trask B. J., Allen S., Massa H., Fertitta A., Sachs R., van den Engh G., and Wu M. (1993). Studies of metaphase and interphase chromosomes using fluorescence in situ hybridization. *Cold Spring Harb Symp Quant Biol* **58:** 767-75.

Tyynelä J., Palmer D. N., Baumann M., and Haltia M. (1993). Storage of saposins A and D in infantile neuronal ceroid-lipofuscinosis. *FEBS Lett* **330:** 8-12.

Tyynelä J., Suopanki J., Santavuori P., Baumann M., and Haltia M. (1997). Variant late infantile neuronal ceroid-lipofuscinosis: pathology and biochemistry. *J. Neuropathol. Exp. Neurol.* **56:** 369-375.

van de Rijke F. M., Florijn R. J., Tanke H. J., and Raap A. K. (2000). DNA fiber-FISH staining mechanism. *J Histochem Cytochem* **48:** 743-5.

Van Diggelen O. P., Keulemans J. L., Kleijer W. J., Thobois S., Tilikete C., and Voznyi Y. V. (2001). Pre- and postnatal enzyme analysis for infantile, late infantile and adult neuronal ceroid lipofuscinosis (CLN1 and CLN2). *Europ J Paediatr Neurol* **5:** 189-92.

van Gijlswijk R. P., Zijlmans H. J., Wiegant J., Bobrow M. N., Erickson T. J., Adler K. E., Tanke H. J., and Raap A. K. (1997). Fluorochrome-labeled tyramides: use in

immunocytochemistry and fluorescence in situ hybridization. *J Histochem Cytochem* **45:** 375-82.

Varilo T., Savukoski M., Norio R., Santavuori P., Peltonen L., and Järvelä I. (1996). The age of human mutation: genealogical and linkage disequilibrium analysis of the CLN5 mutation in the Finnish population. *Am J Hum Genet* **58:** 506-12.

Vaughn J. E. (1989). Fine structure of synaptogenesis in the vertebrate central nervous system. *Synapse* **3:** 255-85.

Weier H. U. (2001). DNA fiber mapping techniques for the assembly of high-resolution physical maps. *J Histochem Cytochem* **49:** 939-48.

Weier H. U., Wang M., Mullikin J. C., Zhu Y., Cheng J. F., Greulich K. M., Bensimon A., and Gray J. W. (1995). Quantitative DNA fiber mapping. *Hum Mol Genet* **4:** 1903-10.

Weimer J. M., Kriscenski-Perry E., Elshatory Y., and Pearce D. A. (2002). The Neuronal Ceroid Lipofuscinosis: Mutations in different proteins result in similar disease. *Neuromolecular Medicine* **1:** 111-124.

Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A., Holt R. A., Gocayne J. D., Amanatides P., Ballew R. M., Huson D. H., Wortman J. R., Zhang Q., Kodira C. D., Zheng X. H., Chen L., Skupski M., Subramanian G., Thomas P. D., Zhang J., Gabor Miklos G. L., Nelson C., Broder S., Clark A. G., Nadeau J., McKusick V. A., Zinder N., Levine A. J., Roberts R. J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A. E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T. J., Higgins M. E., Ji R. R., Ke Z., Ketchum K. A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G. V., Milshina N., Moore H. M., Naik A. K., Narayan V. A., Neelam B., Nusskern D., Rusch D. B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., et al. (2001). The sequence of the human genome. *Science* **291:** 1304-51.

Verkruyse L. A., and Hofmann S. L. (1996). Lysosomal targeting of palmitoyl-protein thioesterase. *J Biol Chem* **271:** 15831-6.

Vesa J., Chin M. H., Oelgeschlager K., Isosomppi J., DellAngelica E. C., Jalanko A., and Peltonen L. (2002). Neuronal Ceroid Lipofuscinoses Are Connected at Molecular Level: Interaction of CLN5 Protein with CLN2 and CLN3. *Mol Biol Cell* **13:** 2410-20.

Vesa J., Hellsten E., Verkruyse L. A., Camp L. A., Rapola J., Santavuori P., Hofmann S. L., and Peltonen L. (1995). Mutations in the palmitoyl protein thioesterase gene causing infantile neuronal ceroid lipofuscinosis. *Nature* **376:** 584-7.

Wheeler R. B., Sharp J. D., Mitchell W. A., Bate S. L., Williams R. E., Lake B. D., and Gardiner R. M. (1999). A new locus for variant late infantile neuronal ceroid lipofuscinosis-CLN7. *Mol Genet Metab* **66:** 337-8.

Wheeler R. B., Sharp J. D., Schultz R. A., Joslin J. M., Williams R. E., and Mole S. E. (2002). The Gene Mutated in Variant Late-Infantile Neuronal Ceroid Lipofuscinosis (CLN6) and in nclf Mutant Mice Encodes a Novel Predicted Transmembrane Protein. *Am J Hum Genet* **70:** 537-42.

White E. L., Weinfeld L., and Lev D. L. (1997). A survey of morphogenesis during the early postnatal period in PMBSF barrels of mouse SmI cortex with emphasis on barrel D4. *Somatosens Mot Res* **14:** 34-55.

Wiegant J., Kalle W., Mullenders L., Brookes S., Hoovers J. M., Dauwerse J. G., van Ommen G. J., and Raap A. K. (1992). High-resolution in situ hybridization using DNA halo preparations. *Hum Mol Genet* **1:** 587-91.

Wilkinson D. G. (1992). "In Situ Hybridization: A Practical Approach," IRL Press, Oxford, UK.
Windle B., Silvas E., and Parra I. (1995). High resolution microscopic mapping of DNA using multi-color fluorescent hybridization. *Electrophoresis* **16:** 273-8.
Vines D. J., and Warburton M. J. (1999). Classical late infantile neuronal ceroid lipofuscinosis fibroblasts are deficient in lysosomal tripeptidyl peptidase I. *FEBS Lett* **443:** 131-5.
Wolfsberg T. G., Wetterstrand K. A., Guyer M. S., Collins F. S., and Baxevanis A. D. (2002). A user's guide to the human genome. *Nat Genet* **32 Suppl:** 1-79.
von Heijne G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225:** 487-94.
von Heijne G. (1996). Principles of membrane protein assembly and structure. *Prog Biophys Mol Biol* **66:** 113-39.
Xu Q., Modrek B., and Lee C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* **30:** 3754-66.
Zeman W., and Dyken P. (1969). Neuronal ceroid-lipofuscinosis (Batten's disease): relationship to amaurotic family idiocy? *Pediatrics* **44:** 570-83.
Zhang Z., Mandal A. K., Wang N., Keck C. L., Zimonjic D. B., Popescu N. C., and Mukherjee A. B. (1999). Palmitoyl-protein thioesterase gene expression in the developing mouse brain and retina: implications for early loss of vision in infantile neuronal ceroid lipofuscinosis. *Gene* **231:** 203-11.
Zhong N. A., Moroziewicz D. N., Ju W., Wisniewski K. E., Jurkiewicz A., and Brown W. T. (2000). CLN-encoded proteins do not interact with each other. *Neurogenetics* **3:** 41-4.