Analysis of NCL Proteins from an Evolutionary Standpoint

Neda E. Muzaffar[1] and David A. Pearce[1,2,3,*]

[1]*Center for Neural Development and Disease,* [2]*Aab Institute of Biomedical Sciences, Department of Biochemistry and Biophysics,* [3]*Department of Neurology, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA*

Abstract: The Neuronal Ceroid Lipofuscinoses (NCLs) are the most common group of neurodegenerative disorders of childhood. While mutations in eight different genes have been shown to be responsible for these clinically distinct types of NCL, the NCLs share many clinical and pathological similarities. We have conducted an exhaustive Basic Local Alignment Search Tool (BLAST) analysis of the human protein sequences for each of the eight known NCL proteins- CLN1, CLN2, CLN3, CLN5, CLN6, CLN7, CLN8 and CLN10. The number of homologous species per CLN-protein identified by BLAST searches varies depending on the parameters set for the BLAST search. For example, a lower threshold is able to pull up more homologous sequences whereas a higher threshold decreases this number. Nevertheless, the clade confines are consistent despite this variation in BLAST searching parameters. Further phylogenetic analyses on the appearance of NCL proteins through evolution reveals a different time line for the appearance of the CLN-proteins. Moreover, divergence of each protein shows a different pattern, providing important clues on the evolving role of these proteins. We present and review in-depth bioinformatic analysis of the NCL proteins and classify the CLN-proteins into families based on their structures and evolutionary relationships, respectively. Based on these analyses, we have grouped the CLN-proteins into common clades indicating a common evolving pathway within the evolutionary tree of life. CLN2 is grouped in Eubacteria, CLN1 and CLN10 in Viridiplantae, CLN3 in Fungi/ Metazoa, CLN7 in Bilateria and CLN5, CLN6 and CLN8 in Euteleostomi.

INTRODUCTION

The neuronal ceroid lipofuscinoses (NCLs, also known as Batten disease) are the most common neurodegenerative disease occurring in childhood with an overall frequency of approximately 1:12,500 [1]. They are a group of neurodegenerative storage disorders inherited in an autosomal recessive manner and are characterized by the ubiquitous accumulation of an autofluorescent lipopigment in tissues of the affected individual [2]. Clinical features are characterized by impaired vision, progressive neurodegeneration resulting in varying degrees of seizures, ataxia progressive psychomotor deterioration and eventually premature death.

The different types of NCLs, which have been categorized on the basis of the age of onset of disease, are infantile NCL (INCL), late-infantile NCL (LINCL), juvenile NCL (JNCL or Battens) and adult NCL (ANCL or Kufs disease). Other variant forms are Finnish variant late-infantile NCL, variant late-infantile NCL, Turkish variant late-infantile NCL, Northern epilepsy (EPMR) and Congenital NCL (CNCL). These NCLs result from mutations in the CLN genetic loci *CLN1, CLN2, CLN3, CLN4, CLN5, CLN6, CLN7, CLN8* and *CLN10,* respectively [3].

To date, the only NCL proteins whose function has been elucidated are palmitoyl protein thioesterase (PPT1) encoded by *CLN1*; tripeptidyl peptidase 1 (TPP1) by *CLN2*; and cathepsin D by *CLN10*. CLN3, CLN5, CLN6, CLN7, and CLN8 while having been identified, have no definitive functional classification. The genes for *CLN4* and *CLN9* remain to be identified.

CLN1/PPT1 (INCL)

Disease Progression

Infantile NCL is the most severe form of NCL and accounts for 25% of all cases in the US [4]. Children with INCL start to show symptoms at around 6-12 months of age when signs of decreased head growth and psychomotor degeneration become apparent. They suffer from progressive vision loss which results in complete blindness by the age of 2. Within the following year most of the affected infants begin to lose higher brain functions and survive into their early to mid teenage years [5-7].

Disease Autofluorescent Storage Material

All tissues in the body show cellular accumulation of autofluorescent storage material characterized by a mix of saposins A and D. The saposin proteins are involved in the degradation of sphingolipids in lysosomes [8, 9].

CLN1 or *PPT1* Gene and PPT1 Protein

INCL results from mutations in the *CLN1* gene, localized to human chromosome 1p32 [10] which encodes the enzyme palmitoyl protein thioesterase 1 (PPT1) [11]. PPT1 is a solu-

*Address correspondence to this author at the Center for Neural Development and Disease, Department of Biochemistry and Biophysics, Box 645, University of Rochester, School of Medicine and Dentistry, Rochester, New York 14642, USA; Tel: (585) 273-1514; Fax: 585-506-1972; E-mail: david_pearce@urmc.rochester.edu

ble lipase that cleaves fatty acids from cysteine residues in proteins during lysosomal protein degradation [12-14]. Mutations in *CLN1* cause decreased enzyme activity that can be readily detected diagnostically [15-19]. There is a good correlation between the level of measurable enzyme activity and the severity of clinical phenotype [20].

PPT1 Structure

The 37kD, 306 amino acid enzyme contains motifs characteristic of other thioesterases i.e., a G-X-S-X-G in the N-terminal half of the protein and a G-D-H near the C-terminus [11]. X-ray crystallographic studies of the structure of PPT1 have provided insight into the molecular basis for phenotypes of INCL associated with known PPT1 mutations [21]. PPT1 is a globular protein with a classical alpha/beta hydrolase fold typical of lipases. There is a hydrophobic groove in the center of the enzyme which binds to the palmitate portion of the substrate that is to be depalmitoylated. The classical hydrolytic triad consists of serine(S)-115, aspartate (D)-233 and histidine (H)-289. As a result, the mutations that will affect the catalytic activity of PPT1 most severely are the ones that affect residues near the active site and in the hydrophobic core of the enzyme because they affect the folding of the enzyme. Therefore, these mutations correlate with the most severe phenotypes seen in INCL patients. Other mutations which affect the peptide binding pocket or the periphery of the enzyme will still allow residual enzyme activity and have been observed to result in the less severe, late-onset phenotype of INCLs [22]. Over 40 mutations have been described in *CLN1* including missense, nonsense, small deletions or insertions, and mutations affecting splice sites [23]. Mutations leading to decreased enzyme stability and inactivity cause infantile-onset INCL [20].

PPT1 Function

Palmitoylation is a lipid-modification which attaches a 16-carbon fatty acid chain *via* a labile thioester linkage to cysteine residues. The palmitoylated proteins eventually undergo depalmitoylation in the course of their degradation and disposal *via* the lysosome. PPT1, which is a soluble lipase, is involved in the depalmitoylation of these palmitoylated proteins and is targeted to the lysosomes to execute this function. The targeting of PPT1 to the lysosome occurs through a phosphate modification on mannose residues of asparagines-linked oligosaccharides *via* binding to the mannose 6-phosphate receptor [13, 24]. PPT1 was first purified from bovine brain (30,000-fold enrichment) on the basis of its ability to cleave palmitate from a model ^{3}H-palmitoylated substrate, H-Ras [12, 25]. In addition to this substrate, a number of other *S*-acylated proteins and peptides are substrates, including palmitoylated G_α proteins [25] and fatty acyl-CoAs.

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

The highest incidence of occurrence of INCL is in Finland where the carrier frequency estimated is 1:70 [11]. Over 90% of these Finnish patients are homozygous for a severe missense point mutation (R122W) leading to enzyme inactivity [23]. Other *CLN1* mutations that result in the severe disease phenotype are H39Q, G42E and V181M. However,

mutations such as T75P, G250V, D79G and Q177E result in a milder phenotype with later ages of onset of disease [17, 26, 27].

The severity of these disease causing point mutations can be correlated to their degree of conservation in a multiple sequence alignment of all the known homologs of PPT1 (Fig. **1**). HomoloGene detected 14 homologs for PPT1 which are present in *H. sapiens, P. troglodytes, C. familiaris, M. musculus, R. norvegicus, G. gallus, D. melanogaster, A. gambiae, C. elegans, S. pombe, M. grisea, N. crassa, A. thaliana* and *O. sativa* (Fig. **2**). The domain that is conserved in all of these sequences homologous to PPT1 is 279 amino acids in length which spans almost the entire length of the 306 amino acid PPT1 in *H. sapiens* (Fig. **1**). The multiple sequence alignment was generated using MultAlin which labels conserved mutations in black and grey. It was observed that the early onset associated severe mutations such as R122W, H39Q and V181M were highly conserved in all the homologs of PPT1.

CLN2/TPP1 (LINCL)

Disease Progression

Late Infantile NCL (LINCL) results from mutations in the *CLN2* gene. The symptoms manifest in children at 2 to 4 years of age with progressive neurodegeneration, complete loss of motor function, vision and speech eventually resulting in premature death by the ages 8 to 14. LINCL is associated with instances of generalized gangliosidosis, severe neuronal loss and gliosis of brain tissue and widespread accumulation of storage material in lysosomes [28-33].

Disease Autofluorescent Storage Material

The principal protein component of the storage bodies observed in patients with LINCL is the subunit c of the mitochondrial ATP synthase [34, 35].

CLN2 or *TPP1* Gene and TPP1 Protein

The *CLN2* gene contains 13 exons spanning a 6.65 kb region on chromosome 11p15 [36]. The gene was mapped by means of a proteomics approach that compared brain lysosomal protein contents of normal controls to patients afflicted with LINCL [36].

CLN2 encodes a lysosomal serine protease [37], tripeptidyl-peptidase I (TPP I) [38, 39], which is 563 amino acids in length and 46 kD in size [40]. TPP1 is a member of the family of serine carboxyl proteinases which play a crucial role in lysosomal protein degradation.

TPP1 Structure

Human TPP1 cDNA encodes a preproenzyme of 563 amino acid residues, which includes a 19 amino acid signal peptide cleaved off cotranslationally, a 176 amino acid prodomain removed during the maturation process, and a mature enzyme of 368 amino acid residues [37, 40-42].

TPP1 Function

TPP1 functions by sequentially removing tripeptides from the N-termini termini of small polypeptides that are sent to the lysosome to be degraded [43-45]. In addition, it

has a minor endoprotease activity [46]. When the aminopeptidase activity of TPP1 is disrupted as a result of several mutations [47] it results in the late infantile variant of NCL (LINCL). Subunit c of mitochondrial ATP synthase, the major protein component of the storage material, is likely to be a substrate of CLN2 [48].

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

Mutational analyses, inhibition studies and sequence similarities with other family members have identified Ser280, Glu77, and Asp81 as the catalytic residues in TPP1 [49]. These three residues are highly conserved in the multiple sequence alignment (Fig. **3**) of the 9 homologous TPP1 proteins present in *H. sapiens, P. troglodytes, C. familiaris, M. musculus, R. norvegicus, G. gallus, M. grisea, X. oryzae* and *S. solfataricus.*

To date, 52 *CLN2* mutations have been associated with LINCL [47] but the most widespread mutation is the R208X mutation which results in the premature termination of TPP1 synthesis resulting in no translational product [32, 50-52]. As with the essential catalytic residues, this residue is also highly conserved among the homologous species (Fig. **4**) as it has a critical role in the proper functioning of TPP1. The missense mutations that are found in patients with LINCL-

N286S, I287N, T353P and Q422H result in amino acid substitutions that induce major misfolding of the precursor peptidase. Consequently, post-translational processing and lysosomal targeting of tripeptidyl-peptidase I is affected. These amino acids, as would be expected, are also present in regions of the protein that are conserved highly by evolution as they are essential for the proper functioning of TPP1 [32].

CLN3 (JNCL)

Disease Progression

JNCL is the most common of the neurodegenerative disorders and is characterized clinically by vision loss at 3- 5 years of age, seizures of increasing severity followed by progressive psychomotor decline resulting in premature death in the late 20s to early 30s. Though this represents the typical disease course, the onset and severity of disease symptoms can vary depending on the causative genetic mutation as well as environmental influences [53, 54].

Disease Autofluorescent Storage Material

The AFSM accumulates in the lysosomes of tissues of the patients suffering from Batten disease. The main component of the storage material in these patients is the subunit c of the F_o region of the ATP synthase complex of the inner mitochondrial membrane [55].

(Fig. 1) contd....

```
          181                                                           240
H.sapiens1  NAGAYSKYVQ ERLVQAEYWH DIIKE-DVYR NHSIFLADIN QERGI-NESY K-KNIMAIKK
P.troglody  NAGAYSKVVQ ERLVQAEYWH DIIKE-DVYR NHSIFLADIN QERGI-NESY K-KNIMAIKK
C.familiar  NAGAYNKAIQ ERLVQAEYWH DIIKE-DTYR NHSIFLADIN QERGV-NESY K-RNIMAIKK
M.musculus  NAGAYSKIVQ ERLVQAQYWH DIIKE-SVYR NYSIFLADIN QERCV-NESY K-KNIMAIKK
R.norvegic  NAGAYSKVVQ ERLVQAQYWH DIIKE-EVYR NCSIFLADIN QERHI-NESY K-ENIMAIKK
G.gallus    DLGAYTQAVQ EHLVQAQYWH DPIKE-EDYR KKSIFLADIN NELFI-NKFY I-ENLQKIKK
D.melanoga  DNAAYAPEVQ KALVQATYWH DPIME-NKYR LGSTFLADIN NELFI-NKFY I-ENLQKIKK
A.gambiae   NYAAYTDLMQ NFLVQATYWH DPLNE-SKYR TSSTFLADIN NEQFI-NKTY V-KNLQKINK
C.elegans   DLGAYLPEVQ KRVVQAQYWH DPNQV-EEYR KRSIFLADIN NENNN-NPTY K-RNLLSIKN
S.pombe     GLGIWHSWIQ NHVVQAQYYR T-EKQYDKYL ENNKFLTHIN NEVLHDNYIR NIEKIKEIDN
M.grisea    G-NTFSQFVQ SRLVPAQYFR DPANNYDKYL ESSNFLADIN NERALKNQTY K-KNIAKLVN
N.crassa    G-NAWTQFVQ SRLVPAQYYR N-LDEYEQYL EHSNFLADIN NERMFKNRTY R-ENIAKIEN
A.thaliana  KAELYSDYMQ EHIAPSGELK IPTDI-AGYM EGCRFLPKIN NELPV-KNST YKERFSSIEN
O.sativa    KLELYSNYVQ AHIAPSGYLK IPTDM-TDYL KGCKFLPKIN NEIPSERNAT YKQRFSSIEN
Consensus   ...ay...vQ ..lvqaq%.. dp......Y. ..s.FLadiN #g....n ty ...nl..LEn
```

```
          241                                                           300
H.sapiens1  FVMVKFLNDS IVDPVDSEWF GFYRSGQAKE TIPLQCTSLY TDRLGLKDM DNAGQIVBLA
P.troglody  FVMVKFLNDS IVDPVDSEWF GFYRSGQAKE TIPLQCTSLY TDRLGLKEM DNAGQIVBLA
C.familiar  FVMVKFLNDS IVDPVDSEWF GFYRSGQAKE TIPLQCTALY TDRLGLKEM DNAGQIVBLA
M.musculus  FVMVKFLNDS IVDPVDSEWF GFYRSGQAKE TIPLQPSTLY TERLGLKKM DKACKIVBLA
R.norvegic  FVMVKFLNDS IVDPVDSEWF GFYRSGQAKE TIPLQCTTLY TERLGLKKM DKACKIVBLA
G.gallus    FVMVKFLNDT MVDPPISEWF GFYKSGQAKD TIPLKCTLLY KEDRLGLQQM DKACKIVBLG
D.melanoga  FVMVQFLNDT IVQPKE-QWF QYYTTGQNKV IQCFTESKVY QD--IGLDKM HRQCQIVBLG
A.gambiae   FVMVQFNNDS IVQPLQTQWF GYYKPGQDKE IQGLKCTNIY IDDRLGLKKV DDQNKIVBLE
C.elegans   LVMVKFNQDH MVVPKDGSWF GFYKDCDIDT ILPMNCTDLY KEDRIGLKKL HESCRIHBMD
S.pombe     IVAVSFERDD IVEPPYSTGF GWINE---TT GENIEMEDFV LYESLGLKDL VNQCKTETIS
M.grisea    FVMFAGDEDT TVIPRETAWF EEVNG---TE VIPLRARKLY TEDWIGIREL DRNCGIKLRT
N.crassa    LVLYMGDEDT TVVPKETAWF EEVNG---TE ITPLRARKMY SEDWIGLREL DRKCGLHBRT
A.thaliana  LVLIMPEHDT ILIPKETSWF GYYPDGSFKT ILPPQETKLY TEDWIGLRTL DEACKVKBVN
O.sativa    LVLIMPEDDA VLIPRETAWF GYYPDGAFSP VQPPQKIKLY TEDWIGLKAL EEACRVKBVS
Consensus   lV.v.F..D. iv.P..s.WF g.y...g.... ..p..et.ly e#.lGLk.$ d..G.l.f..
```

```
          301                                                           360
H.sapiens1  TE DHLQLSE EWFYAHIIPF IG
P.troglody  TE DHLQLSE EWGYAHIIPF IG
C.familiar  VE DHLQLSE EWFYAHIIPF IE
M.musculus  KE DHLQISK EWFTAHIIPF IK
R.norvegic  KE DHLQISK EWFTAHIIPF IK
G.gallus    VE DHIHFTE EWGYTHILPF LQ
D.melanoga  VE DHIAISK AWFIQNIVPL LLEK
A.gambiae   CE NHLQFTK EWGRENLFSY LK
C.elegans   VD DHLQIPR SVLVNDIIKK YFM
S.pombe     FP RHLQMRW GDFDALVLKY FKDEKEEKTE LEESTRPSNF LSTYFVSPLV SAIDGTVDYL
M.grisea    CP EHMQLS- ---DELLNKT FTEFFGPNNK KFGMGGQAGD LDVTREDEL
N.crassa    AP DHMQLT- ---DQLLNET FAEFFGPIKK AKKSAPRPLM LGGGDL
A.thaliana  VS NHLQISH TDMKKHIVPY LCD-----KS SSSSSTTMAV SGSSSSHQWL PSVGNIIMDL
O.sativa    VP EEGVPSD QF
Consensus   v.G.hlg.s. ..f...i... .......... .......... .......... .........
```

Fig. (1). Conserved Domains of human PPT1.

The conserved domain of human PPT1 is Palmitoyl protein thioesterase (pfam02089), indicated by the box shaded in grey that spans the multiple sequence alignment (MSA). It is 279 amino acids in length and aligns 99.6% with the human PPT1 protein from residue 28 to residue 305. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in gray. All other neutral residues are not highlighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.
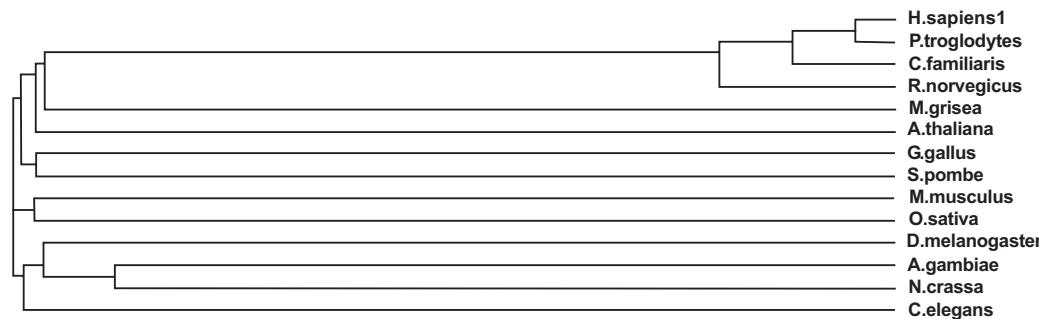
Fig. (2). Phylogenetic tree of PPT1 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 14 homologous sequences are proportional to the amount of inferred evolutionary change.

CLN3 Gene and CLN3 Protein

The *CLN3* gene contains 15 exons and spans a 15 kb region between chromosome 16p12.1- p11.2. The *CLN3* gene encodes the hydrophobic CLN3 protein which is 438 amino acids in length [56, 57].

CLN3 Structure

The CLN3 protein is predicted to contain 6 transmembrane domains [58] and has been shown to reside in the lysosomal and endosomal membrane of cells [59-61].

```
                 1                                                          60
H.sapiens2                                          MGLQACLLG LFALILSGKC SYSPEPDQRR
P.troglody                                          MGLQACLLG LFALILSGKC SYSPEPDQRR
C.familiar                                          MRLRTCLLG LLALCVAGKC SYSPEPDQQR
M.musculus                                          MGLQARLLG LLALVLAGKC TYNPEPDQRW
R.norvegic                                          MGLQARFLG LLALVIAGKC THSPEPDQRW
 M.grisea      MFAFWTLSQS SFHLLIHSAN PVRSTVGQGR YREVIAYAAL ATVLPLSASA ASNVVVESLP
 S.solfatar                                    MYR YIFLMSMLLI SIIPLVFAS- --NPNMYQNP
  X.oryzae                                                              MHTVLAGSER
  G.gallus
  Consensus    .......... .......... .......... .......... .......... ..........

                 61                                                        120
H.sapiens2     TLPPGWVSLC RADPEDELSL TFALRQQNVE RLSELVQAVS DPSSFQNGKY LTLENVADIV
P.troglody     TLPPGWVSLC RADPEDELSL TFALRQQNVE RLSELVQAVS DPSSFQNGKY LTLENVADIV
C.familiar     TLPPGWVSLC RVDSEDELSL TFALRQQNVE RLSELVQAVS DPSSFHNGKY LTLENVAEIV
M.musculus     MLPPGWVSLC RVDPEDELSL TFALRQRNIE RLSELVQAVS DPSSFQNGKY LTLEDVAEIV
R.norvegic     MLPPGWVSLC RVDPEDELSL TFALKQQNLD RLSELVQAVS DPSSFRNGKY LTLEDVAEIV
 M.grisea      AVPEGWEKIQ DADPDSKIRL KVALRQPNLA LFEREMYAVS SPLSARVGKH LKREEVRDIM
 S.solfatar    ITLKEFREIC TLNANEVIV TIFVPLKNID LLYYYASGAS NEASPLHKF LSPHEVQQIF
  X.oryzae     HTVPSILEAC KPPQDDRLRV LLAVRMPALN AAADALLQLQ PNTIPVA--- LTRAAFSERF
  G.gallus
  Consensus    ........g .....e.... ..a....nl. ..........s .p.sp.y.k. l....v..l.

                 121                                                       180
H.sapiens2     RESPLTLHTV QKWLLAACAQ KCHSVITQDF LTCWLSIRCA ELLIPGAEFH HYVGCPTETH
P.troglody     RESPLTLHTV QKWLLAACAR KCHSVITQDF LTCWLSIRCA ELLIPGAEFH HYVGCPTETH
C.familiar     RESPLTLHTV QKWLSAACAR KCHSVITQDF LTCWLSVRCA ELLISGAEFH HYVGCPTETH
M.musculus     QESPLTLLTV QKWLSAACAR KCDSVITQDF LTCWLSVRCA ELLISGAEFH HYVGCPTKTH
R.norvegic     QESPLTLRTV QKWLLAACAR DCHSVITQDF LTCWLSVRCA ELLIPGAEFH HYVGCPAKTH
 M.grisea      KDDPASTSTV MQWLENAGIP SQAIHEDSDW INIVTSVRNA SALTK-ADFG VYGHANSDLK
 S.solfatar    LPTE-EYNQI LNYVKSSCFQ VIFTASNSV IVIKGTVGQV EKYILGTKYA VYSNGCVTY
  X.oryzae     AASAADLQAV ADFAASYCLA VEHSHADSGN VILEGTVQQC EAAFH-VCLR EYVHCTS--R
  G.gallus
  Consensus    .p.......v .........g. .........s. ........v.g. e..l...... .y..g.s...

                 181                                                       240
H.sapiens2     VVRSPHEYQL EQALAPHVDF VGGLHRF-- -EPTSSLRQR PEPQVT---- -------GTV
P.troglody     VVRSPHEYQL EQALAPHVDF VGGLHRF-- -EPTSSLRQR PEPQVT---- -------GTV
C.familiar     VIRSLREYQL EKALAPHVDF VGGLHRF-- -EPTSSLRQR PEPQVS---- -------GTV
M.musculus     VIRSPHEYQL EQALAPHVDF VGGLHRF-- -EP-SSPRQR PEPQQV---- -------GTV
R.norvegic     IIRSPHEYQL EQALAPHVDF VAGLHRF-- -EPLSSPRQR PEPQGV---- -------GPV
 M.grisea      KIRALE-YSV EDNVRDHITM VAPIIRFGEI KPKRSIVKSV EDSKGSIQAA EVPTGQLDVA
 S.solfatar    YTNYGYEKIN AYVYSSNISA IFFAH----- KPSTLITEST IKSFQQ---- -----EINQT
  X.oryzae     YRGRTGEVSI EQALTGIVTA VLGLDA---- RPQAQTLPSV PDASASLPTA TPPPTPASPE
  G.gallus
  Consensus    .......p.. ..p....... v......p. ......s. .......... ..........

                 241                                                       300
H.sapiens2     -GLHLGVTES VIRKRYNLTS QDVGSGTSNN SQACAQFLEQ YFHDSDIAQF MRLFG----G
P.troglody     -GLHLGVTES VIRKRYNLTS QDVGSGTSNN SQACAQFLEQ YFHDSDIAQF MRLFG----G
C.familiar     -GLHLGVTES VIRKRYNLTA QDVGSGTTNN SQACAQFLEQ YFHASDIAEF MRLFG----G
M.musculus     -SLHLGVTES VIRQRYNLTA KDVGSGTTNN SQACAQFLEQ YFHNSDLTEF MRLFG----G
R.norvegic     -GLHLGVTES VIRQRYNLTA RDVGSGTTNN SQACAQFLEQ YFHNSDLTEF MRLFG----S
 M.grisea      -ACNKTITEE CLRALVQIS-G DYKATGSPKS RFGIAGYLEE YARYSEIEKF QEVYAPYTLG
 S.solfatar    -FPLEGYWPT VIQKVYNVTT E-----GENT TIGILDFYGD PYIVQQIAYF DKITGLPNPP
  X.oryzae     DGPIMQYTPP QLAQLYGFPE HD----GHGQ CIGIIVLGGG YAREQMAAYE AQL-RIPMPT
  G.gallus
  Consensus    .......tp. .l.....y.... .........gi .....y.....l.f ..........

                 301                                                       360
H.sapiens2     NFAHQASVAR VVGQQGRGRA GIEASLDVQY LMSAGANIST WVYSSPGR- -----------
P.troglody     NFAHQASVAR VVGQQGRGRA GIEASLDVQY LMSAGANIST WVYSSPGR- ----------
C.familiar     NFAHQASVAR VVGQQGRGRA GIEASLDVEY LMSAGANIST WVYSSPGR- ----------
M.musculus     NFTHQASVAR VVGKQGRGRA GIEASLDVEY LMSAGANIST WVYSSPGR- ----------
R.norvegic     NFAHQASVAR VVGKQGRGRA GIEASLDVEY LMSAGANIST WVYSSPGR- ----------
 M.grisea      SNFSVVTING GLNDQSSTAD SVEANLDMQY GVALGYNTNI VYYSTGGRGP LVPDLDQPNG
 S.solfatar    NFSVVPIGPY NPNLGIVTGW AGEISLDVEV AHIIAPKANI TLYIANPNIP L---------
  X.oryzae     LVDVLLPGAT NTVSHGTANA DVEAQMDIQI AGALVPGAKL VVLFAP---- ----------
  G.gallus
  Consensus    .......... .........  ....ea.ld......a....... ..y.... ..........

                 361                                                       420
H.sapiens2     -HEGQEPFIQ WIMLLSN--E S-ALEHVHTV SYGDDEDSLS SAYIQRVNTE LMKAAARGLT
P.troglody     -HEGQEPFIQ WIMLLSN--E S-ALEHVHTV SYGDDEDSLS SAYIQRVNTE LMKAAARGLT
C.familiar     -HEGQEPFIQ WIMLLSN--E S-ALEHVHTV SYGDDEDSLS SAYIQRVNTE FMKAAARGLT
M.musculus     -HEAQEPFIQ WIMLLSN--E S-SLEHVHTV SYGDDEDSLS SIYIQRVNTE FMKAAARGLT
R.norvegic     -HEAQEPFIQ WIMLLSN--E S-SLEHVHTV SYGDDEDSLS SVYIQRVNTE FMKAAARGLT
 M.grisea      NDSSNEPYID FLEYMKNLPD D-ELPQTLTT SYCENEQSVP REFAVKVCNM FGELGARGVS
 S.solfatar    --PAIIAYIT SQNKVDTLSQ SFSIPESLFS SLFNGP--LF YSCIILSDEY YALGSAEGIT
  X.oryzae     --NTDNGFIE AINAAIHNTE --HAEGIIVI SWCFTESQWT PQSRCAYDCA EQAAFLMGIT
  G.gallus     RHESQEPFIA WLLLLSMS-- --ALEWVHSV SYGDDEDSLS QAYILRVNVE ELKAAARGLT
  Consensus    .....ep%l. .l....l... ...LF...... Syg...e.sl. ....q.v... %...aaarG.t

                 421                                                       480
H.sapiens2     LLFASGDSGA C--CWSVSGR H--QERPTFP ASSPYVTIVG GTSFQEPELI TNEIVDYISG
P.troglody     LLFASGDSGA C--CWSVSGR H--QERPTFP ASSPYVTIVG GTSFQEPELI TNEIVDYISG
C.familiar     LLFASGDSGA C--CWSVSRR H--QERPSFP ASSPYVTIVG GTSFQNPERV TTEIVDYISG
M.musculus     LLFASGDTGA C--CWSVSGR H--KDRPSEP ASSPYVTIVG GTSFKNPELI TDEVVDYISG
R.norvegic     LLFASGDTGA C--CWSVSGR H--KDRPSFP ASSPYVTIVG GTSFKNPELV TNEVVDYISG
 M.grisea      IFFSSGDEGP CSACQSNDGK NTTRDDAFPI AACPYVTAVG GTRYVAE--- --TGAILAG
 S.solfatar    FIASGGDACG SGGYSNGPICT ------VGYP STSPFVTSVG GTTVYVQIPN GSYYQTIWSN
  X.oryzae     VCIAAGDDCA C---DGQPGL N-----VQFP ASSPFVLACG GTRLQVTADS ANCQAWANGC
  G.gallus     VLFASGDSGA C--CRRVPGG NH-TERPSFP ASSPYVTIVG GTSFKNPEMV TAEVTDYISG
  Consensus    .lfasGD.Ga .g.c......n.. .f.p.%p%P assP%Vt.vG GT....pf... ..e....sg
```

(Fig. 3) contd….

```
                   481                                                           540
H.sapiens2    GGF--SNVFP  RESYQEEAVT  KFLSSSPHLP  PSSYFNASGR  AYPDVAALSD  GY---WVVSN
P.troglody    GGF--SNVFP  RESYQEEAVT  KFLSSSPHLP  PSSYFNASGR  AYPDVAALSD  GY---WVVSN
C.familiar    GGF--SNVFP  RESYQEEAVT  QFLSSSPHLP  PSSYFNASGR  AYPDVAALSD  GY---WVVSN
M.musculus    GGF--SNVFP  RPPYQEEAVA  QFLSSSPHLP  PSSYFNASGR  AYPDVAALSD  GY---WVVSN
R.norvegic    GGF--SNVFP  QPSYQEEAVA  QFLSSSSHLP  PSSYFNASGR  AYPDVAALSD  GY---WVVSN
  M.grisea    GGF--SDIWP  QEWCKTGVE   KYLSILGDKW  KGLY-NPSGR  AYPDIAGQAY  SY---HVFDS
 S.solfatar   YGFVPNNVNY  GGSTGGVSII  EPKPWYQWGL  PTPSTYPNCK  LIPPISANAN  VPPGIYIVLP
   X.oryzae   GCE--GRFEA  REAWCKDLRL  TDAQQQSRQL  RM-------R  GVPLVAANAD  AQTGYYLSIN
   G.gallus   GGF--SNVFP  MEDYQVRAG
  Consensus   gGf..snvfp  .p..q..a..  ..........  .......gr  ..pd.aa.e.  ..y.......

                   541                                                           600
H.sapiens2    RVPIPWVSGT  SASTPVFGGI  ISLINEHRIL  SGRPPLGFLN  PRLYQQHCAG  LFDVTRCCHE
P.troglody    RVPIPWVSGT  SASTPVFGGI  ISLINEHRIL  SGRPPLGFLN  PRLYQQHCAG  LFDVTRCCHE
C.familiar    SVPIPWVSGT  SASTPVFGGI  ISLINEHRIL  SGRPPLGFLN  PRLYQQHCAG  LFDVTRCCHE
M.musculus    MVPIPWVSGT  SASTPVFGGI  ISLINEHRIL  NGRPPLGFLN  PRLYQQHCTG  LFDVTRCCHE
R.norvegic    MVPIPWVSGT  SASTPVFGGI  ISLINEHRIL  NGRPPLGFLN  PRLYQQHCAG  LFDVTHGCHE
  M.grisea    GKDV-LVAGT  SASSPSVAAI  FALLNDARLA  ANQPPMGWIN  EWLYEIGLKG  LTDITVCGSR
 S.solfatar   SNTTGITGGT  SEASPLTGV   LATPESYTHH  RIGLLNPILT  YMAENYYCKV  IEPITFCYNI
   X.oryzae   GQPA-VMGGT  SAAPLWAAL   LARI----YG  ANRMQPHFVL  PRLYQPCA-  FRDIVACDN-
   G.gallus
  Consensus   ........gt  sa..p..a..  .la.i.....  ..........  p.ly..g..  .dit.g..

                   601                                                           660
H.sapiens2    SCLDEEVEG-  --------QC  FCSGPGWDPV  TGWCTPNFPA  ILKTLLNP
P.troglody    SCLDEEVEG-  --------QC  FCSGPGWDPV  TGWCTPNFPA  ILKTLLNP
C.familiar    SCLDEEVEG-  --------QC  FCSGPGWDPV  TGWCTPNFPA  ILKALLKP
M.musculus    SCLNEEVEG-  --------QC  FCSGPGWDPV  TGWCTPNFPA  ILKTLLNP
R.norvegic    SCLNEEVEG-  --------QC  FCSGPGWDPV  TGWCTPNFPA  ILKTLLNP
  M.grisea    GCTGKSIYSG  LPAPKVPPAS  WNATEGWDPV  TGLCTPLFNK  ILELSTTGAQ  AQKIAN
 S.solfatar   PWVATYGYNL  VTGYGTINAG  YFEKILPTLN  LSKELNVIVS  VYNTSDETVS  PQQFYPGQRI
   X.oryzae   ---------  --------AG  FRACTGWDAN  TGLCAEDGAR  IAEVLRES
   G.gallus
  Consensus   ..........  ........ag  .....gwd..  tg.g.p...  l.....g...  ..........
```

Fig. (3). Conserved Domains of human TPP1.

TPP1 in humans contains the conserved Pro-kumamolisin activation domain (pfam09286), indicated by the box shaded in grey. Members of this family are found in various subtilase propeptides, and adopt a ferredoxin-like fold, with an alpha+beta sandwich. Cleavage of the domain results in activation of the peptide. The conserved domain sequence is 142 amino acids in length and aligns 100% with TPP1 protein sequence. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in gray. All other neutral residues are not highlighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.
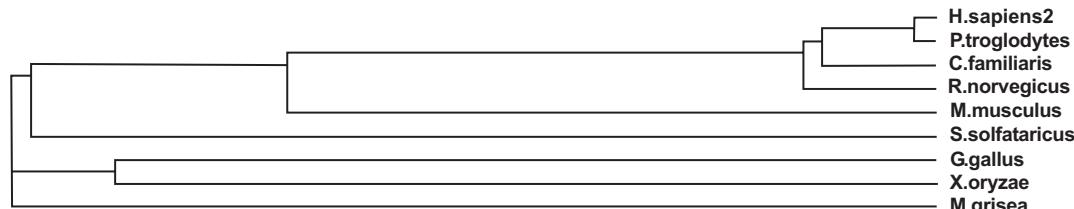
Fig. (4). Phylogenetic tree of TPP1 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 9 homologous sequences are proportional to the amount of inferred evolutionary change.

CLN3 Function

CLN3 has been implicated in controlling the acidic pH in lysosomes. Studies of fibroblasts from Batten disease patients show elevated lysosomal pH [62] and yeast lacking the CLN3 protein homolog, Btn1p, have abnormal vacuolar pH in the early phases of growth. This defect in Btn1p lacking yeast strains can be reversed by complementation with either the yeast wild-type *btn1* gene or human *CLN3* [63, 64].

In addition to its role in regulating lysosomal pH, CLN3 has also been suggested to function in maintenance of biophysical membrane properties [65], control of apoptosis [66, 67] and control of protein trafficking [68]. However, the role of CLN3 in the pathological mechanism leading up to Batten disease still remains unclear.

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

The mutation responsible for 85% of Batten disease chromosomes is a 1.02 kb deletion resulting in a frameshift mutation that generates a premature termination codon. The result of this mutation is a truncated protein 181 amino acids in length, consisting of the first 153 residues of the protein, followed by 28 novel amino acids before the stop codon [69]. Patients who inherit this mutation homozygously always manifest severe symptoms including blindness, epilepsy, dementia and premature death at approximately 24 years of age. Additionally, 31 other Batten disease mutations have been described [70] which can result in JNCL when inherited in a compound heterozygous manner. Several of these mutations described to result in JNCL cause a disruption or deletion of the highly conserved stretches of amino acids in CLN3, 184**WSSGTGGAGLLG**195, 291**VYFAE**295 and 330**VFASRSSL**337 [67]. These regions are highly conserved (Figs. **5** and **6**) in the CLN3 homologs- *H. sapiens, P. troglodytes, C. familiaris, M. musculus, R. norvegicus, C. elegans, D. melanogaster, A. gambiae, N. crassa, M. grisea, S. pombe, K. lactis, S. cerevisiae* and *E. gossypii.*

```
                1                                                          60
H.Sapiens3                                                     GFWILGLQNN FSYVVMLSAA
C.familiar  MGGCAGSRRR LLDSEEEETA PEPR-PPRSY HKGALWKNVM GFWILGLQNN FSYVVMLSAA
M.musculus  MGSSAGSWRR LEDSEREETD SEPQ-APRLD SRSVLWKNAV GFWILGLQNN FSYVVMLSAA
R.norvegic  MGSSESSWRR LEDSEREETD SEPQ-TPPLD NLSVLWKNAV GFWILGLQNN FSYVVMLSAA
D.melanoga  MTATKQQEAE AEHDPQVVVT PGDA-SRSSR QDRGLWRDLT SYWILGLQNN YGYVVMLTAA
A.gambiae              PSVVPLT ENDP-DQAPP KDKGLWRDLV AYWILGLQNN YGYVVMLSAA
C.elegans                      M VNWTNFRDIL AFWLLGLQNN YGYVVMLSAA
S.pombe                     MIKL RLTKDAKVGC QLIFGLNNN LLYVIILSAA
S.cerevisi                     MSDKSHQIYC YFWLFGLNNN VLYVVLLSAA
E.gossypii                     MDRRKLIFG KFWLFGLNNN VLYVVILAAA
K.lactis                     M ELDRDKKTFA YFWLFGLNNN ILYVVILSAA
M.grisea    MPGSTSASWA VYRS-RLVSV FRGVDSSVLI AFWLFGLINN VLYVVILSAA
N.crassa                     M NRSPSSSGLL P--PGLINN VLYVIILSAA
P.troglody  MGGCAGSRRR FSDSEGEETV PEPRLPLLDH HQGARWKNAV GFWLLGLQNN FSYVVMLSAA
Consensus   .......... .......... .......... .......... .fwl.GL.NN .YV!.LsAA
```

```
                61                                                         120
H.sapiens3               NQSHVDPGP TPIPHNSSSR FDCNSVSTA VLLADILPTL VIKLLAPLG-
C.familiar  HDILSHQRAS GNQSHVDPDP PPTAHNSSSR FDCNSVSTA VLLADILPTL ILKLLAPLG-
M.musculus  HDILKQEQAS GNQSHVEPGP TPTPHNSSSR FDCNSISTA VLLADILPTL VIKLLAPLG-
R.norvegic  HDILKQEQAS GNQSHVEPGP TPMPHNSSSR FDCNSISTA VLLADILPTL VIKLLAPLG-
D.melanoga  HDIIKQ---- -----FNPND ESEESSS--G RNCHLVSTG ILLADVLPSL FVKILMPF--
A.gambiae   HDIIKE---- -----LEGGD GHAKSASPLA RPCNKLSTG ILLADILEAL AVKSIASF--
C.elegans   EDILDDQKG- ---TNSSSTN SSTCETHLDS RHCQSISTG VLLADNLPAL VVQTTFPFF-
S.pombe     LDLVG----- ---------- ---ANVSKGV VLLSNIVPSL AGKISTSILH
S.cerevisi  VDIVG----- ---------- ---PTLPKSL VLLADIFPSL AIKLCSPFF-
E.gossypii  VDIVG----- ---------- ---PTLPKSL ILLADILPSF LVKLVAPFF-
K.lactis    SDIIG----- ---------- ---PSLPKSI VLLFDIMPSF LIKLSAPFF-
M.grisea    QDLVG----- ---------- ---TNVPKGV VLLADVLPSF LTKLVAPFF-
N.crassa    QDLVG----- ---------- ---NGIPKGV VLLADVMPSF LTKLVAPFF-
P.troglody  HDILSH---- ---------- KRTSGNQSHA VLLADILPTL VIKLLAPLG-
Consensus   .Di....... .......... .......a !LLaD!lPsl .Ki.ap...
```

```
                121                                                        180
H.sapiens3  LHLLPY--SP RVLVSGICA GSFVLVAF-- -SHSVGTS C GVVFASISSG LGEVTFILSLT
C.familiar  LHLLPY--SP RVLVSGICA GSFILVAF-- -SHSVGTS C GVLASISSG VGEVTFILSLT
M.musculus  LHLLPY--SP RVLVSGVCSA GSFVLVAF-- -QSVGLS C GVLGSISSG LGEVTFILSLT
R.norvegic  LHLLPY--SP RVLVSGVCSA GSFVLVAF-- -QSVGLS C GVLASISSG LGEVTFILSLT
D.melanoga  -FPFWV--NF RIALAVAFSA ACFLLVG-- -ANVEWMALL GVIITSASSG LGETTFIAYS
A.gambiae   -LPLAK--NV RIFLCVAAA ACFLTAF-- -ATVEWVLFV GVIATSFSSG LGBATFFAYA
C.elegans   MTRFPE--GF RAFLVFLLQ SSYFVVAY-- -SKNLAMSLA GVCMASLGQG IGEITFIAVA
S.pombe     VHKFKP--AK RIGFCVFMSI LCMQWLAW-- -SSSVPSKLL GVSLLAISSS FGEISTHLS
S.cerevisi  IDRIKY--SY RIWSLITMSC LCMFLVS--- -FKNLFVCLL GVSFASISSG FGEVTFIQLT
E.gossypii  IDRVEY--HY RIWSLIALSG FCMLLVA-- -GRLGVCIV GIVLASISSG VGEITFIQLT
K.lactis    VISIHY--DK RIPILILLSM LCIILVS--- -TRSLWLCP GIVLASLSSG FGEITFIQLT
M.grisea    IHRIPY--YA RILIFVALSS SCMLLVALTP AERGVAARLV GVVLASISSG GGELSETFGLT
N.crassa    IHRIPY--AT RILIFVTLSS ACMLLIALTP PSRSVAVKLI GVVLSSISSG GGELSFLGLT
P.troglody  LHLLPYRSGP RVLVSGICA GSFILVAF-- -SHSVGTS C EAC
Consensus   .h..py.... R!.......sa .g...l!af.. .s.sv....l gv..as.ssg .ge.tfl.lt
```

```
                181                                                        240
H.Sapiens3  AFYPRAVISW WSSGTGGAGL LCALSYLGLT Q-AG SPQQ LLSMLGIFAL ILASYFLLLT
C.familiar  AFYPRAVISW WSSGTGGAGL MCALSYLGLT Q-AG SPQH LLSMLGIFAL NIASYFFLLT
M.musculus  AFYPRAVISW WSSGTGGAGL LCSLSYLGLT Q-AG SPQH LLSMLGIVAL ILASYFLLLT
R.norvegic  AFYPRAVISW WSSGTGGAGL LCSLSYLGLT Q-AG SPQH LLSMLGIVAL NIASYFLLLT
D.melanoga  SRNKNVIST WSSGTGGAGV IGSLSYASTR S-LDFSPRDI MLIMLIFFAI FAFAFWLLLR
A.gambiae   TYNKNVIST WSSGTGGAGI AGSLSYTGLT G-LGTTPKTI ILIMLVVGGL RAAAFWLLLR
C.elegans   AHLIPETIAS WSSGTGGAGL IGSFSYAFIL Q-AG SPSNL LIVQLFIFVV FAGAYFFLLT
S.pombe     SRHSVSLPC WSSGTGLAGL FCASSYLVMT TWFNFSVRST LIISSFLLLF LIIMYFFVLP
S.cerevisi  HYKQISLNG WSSGTGGAGI ICGASYMFLT SIFKVPVKLL LLVFSLLLFA FFYFKLESN
E.gossypii  HYSHVALNG WSSGTGGAGL ACSFLMLLT SILKIPVSRS LLLFSILLIG FLYFTLQVE
K.lactis    HFGSKSLTG WSSGTGGAGL VGSFSYLLT TVFRLNIQLS LLLYAALLLI FLYYKI--N
M.grisea    HYGPMSLAA WGSTGGAGL AGAGLLVAMT QWFGTVKTS LLASALLLGV MVVSFFVIP
N.crassa    HYGHMSLAA WGSTGGAGL IGSLLLVMLT DWIG SVKTS LLASAFLLII MLVSFFLILP
P.troglody
Consensus   ...y....... wssgtggagl .gs.sy..lt ....l....t ll.....p.. .l.......l.
```

```
                241                                                        300
H.Sapiens3  SPEAQDPGG- --------EE EAESAARQP- ---------- ---------- ----LIRTEA
C.familiar  SPEPQDPGG- --------EE EAETSARQP- ---------- ---------- ----LIDSET
M.musculus  SPEPLDPGG- --------EN EAETAARQP- ---------- ---------- ----LIGTET
R.norvegic  SPEPLDPGG- --------QD EAETAARQP- ---------- ---------- ----LIGTET
D.melanoga  RPQVDILPV- --------TT VEST------ ---------- ---------- ----------
A.gambiae   HKDTDKPAD- --------IS EQEAGSEKP- ---------- ---------- ----E-EIDY
C.elegans   IPPTVYSPTI HPSTWIIPKN YDKDVFEAA- ---------- ---------- ----EARGDV
S.pombe     ESES------ ---TSPSINN NYTPIES--- ---------- ---------- -----IDLRA
S.cerevisi  DTNL------ ---TYQSLQQ IDEAEDDQLV ---------- ---------- ----PFPVAF
E.gossypii  RTAY------ ---EPLASGH FMEAEDNSGS ---------- ---------- ----VISLEA
K.lactis    DSIV------ ---SSQVYQS LDMPIDSSDG ---------- ---------- -----PIELLK
M.grisea    RAPLKEGQAA GVKAYSALPD RDLEDEDVEG -------MSA GDASASLLAP GPPGVASTAL
N.crassa    HGPLR---AS ARKTYEPIAN RDSFQEESEG AENNFDDIPT STASSSLAP GP-AVAATAY
P.troglody
Consensus   .......... .......... .......... .......... .......... ..........
```

```
                301                                                        360
H.Sapiens3  PESKPGS--- -------SS SLSLRERWTV FKGLW-YIV PLVVVYFASY FINQGLFELL
C.familiar  PESKPDS--- -------SS NLSLQERWTV FKGLW-YIV PLVLVYFASY FINQGLFELL
M.musculus  PESKPGS--- -------SW DLSLQERWTV FKGLW-YIL PLVLVYFASY FINQGLFELL
R.norvegic  PESRPGT--- -------SW DLSLQERWTV FKGLR-YIL PLVLVYFASY FINQGLFELL
D.melanoga  EVLISDE--- -------KP LVGFKEKLFY IKHFFK-WML PLCIVYFFSY FINQGLFELV
A.gambiae   NTLPEDE--- -------RP LENWNQRLRY IPSFI-YML PIILVYLLSY YINQGLFELV
C.elegans   LNTKKVP--- -------QR ELGPLERIKL IGPFY-L V PLATVYTASY MINQGLTELI
S.pombe     GHVSFN---- --------F VNSLKQTFIF MQPYLLSH F QFLVYFSY TINIGVAPTI
S.cerevisi  THTNASQS-- ------LYST RQHILQTVKR LRRVFPYVI PLTTVYLFLY LINQAVAPTI
E.gossypii  PRTTVDNARD CASRPRLAAL RERIEVTMRR LKVVVPTMI PLSTVYLFSY LINQGVSPTI
K.lactis    SNGWVD---- -------IVKR LSKVFPYMI PLSTVYLFSY LINQGVAPTI
M.grisea    HAVDAEGGGH GAAAARGAAA KPTLARNLHR AKSFPYTML PLLLVYVAEY TINQGVSPTI
N.crassa    SSHPTED--- -------AR KDSLAAKIQR SKSFFPYTL PLLLVYIASY TINQGVSPTI
P.troglody
Consensus   .......... .......... ....l..ym. pl.vy..ey .inqg.....
```

(Fig. 5) contd….

```
            361                                                                      420
H.Sapiens3  FF---WNT--  ------SLSH AQQYRWYQML  YQAGVFASRS  SLRCCRIRF-  -TWALALLQC
C.familiar  FF---RNT--  ------SLNH AQQYRWYQML  YQAGVFVSRS  SLHCCRIRF-  -TWVLALLQC
M.musculus  FF---RNT--  ------SLSH AQQYRWYQML  YQAGVFASRS  SLQCCRIRF-  -TWVLALLQC
R.norvegic  FF---RNT--  ------SLSH AQQYRWYQML  YQAGVFASRS  SLQCCRIRF-  -TWILALLQC
D.melanoga  YF---EDI--  ------ELDK DSQYRWLNVD  YQIGVFISRS  SVNLFQLDK-  -IWLMSIFQF
A.gambiae   YF---PGI--  ------WLSQ SAQYRWYQVI  YQIGVFISRS  SVNLIQFRQ-  -VWIMAVFQF
C.elegans   IFNCSQGL--  ------SLPL SSQYRWYQVL  YQLGVFISRS  SVKFFEMPLW  LIWCLPILQC
S.pombe     LFPPEKAG--  ------ESSF RDFYPTYQTV  YQIGVFLSRS  SISFFTVPY-  LRTLAITQF
S.cerevisi  LFPINGDERS  KSMPFFEHKY RDIYVTYGTL  YQLGVFISRS  FGHLMRMRS-  LYILAFIQG
E.gossypii  MFPIHEGYGT  SQL---EHKY RDIYVAYGTL  YQLGVFISRS  SGSWVRIRG-  LYLLSVLQF
K.lactis    LFPIDTTP--  ------EVKY RDIYVTYGTL  YQLGVFISRT  WGHKLPVKN-  LYLLSVLQL
M.grisea    LFPLEQSP--  ------ETEF RSFYPFYGFL  YQLGVFISRS  STPFYRVHN-  LYLPSVLQV
N.crassa    LFPLESSP--  ------EAEF RSFYPFYGFL  YQVGVFISRS  SIAFIRIHH-  LYLPSVLQV
P.troglody
Consensus   .f........  ... f... ...y..y..l  yq.gvfisrs  s........  ......lq.


            421                                                                      480
H.Sapiens3  LNLVFLLADV  WFGFLPSIYL  VELIILYEGL  LGGAAYVNTF  HNIALETSDE  HREEAMAATC
C.familiar  LNLAFLLVDV  WFSFLPSIYL  VELIILYEGL  LGGAAYVNTF  HNIALETSDQ  HREEAMAAAC
M.musculus  LNLALLLADV  CLNFLPSIYL  IEIIILYEGL  LGGAAYVNTF  HNIALETSDK  HREEAMEAAC
R.norvegic  LNLAFLLADV  CLTFLPSIYL  IEIVILYEGL  LGGAAYVNTF  HNIALETSDK  HREEAMEAAC
D.melanoga  VNVVYFLTEV  IWWYTPSIWI  VEAVLWEGL   LGGGAYVNTF  YRMSKEISPE  RQQEAMAMVV
A.gambiae   LNVVYFTFEA  VYYFTPSIWI  IEVLILWEGL  LGGGGYVNTF  YRIQTDVPAA  RREYAMMVTS
C.elegans   VNMIFFFFEA  VYWFTPTIII  IEVLILWEGL  LGGSAYVNTY  NKIHKKVNPD  VREYSLSAAS
S.pombe     IILLFTILQS  ALYLTSSYHF  VLFLIFVEGL  IGGTVYVNVY  -HSLQTTESS  QRELAISTVG
S.cerevisi  VNLCITVLQS  WFYVTHSPWA  VMILIFYEGF  LGGASYVNTF  LNILEQEDPD  ETEEAMGAVS
E.gossypii  LNLVILLIQS  WYYVIHSVWV  IMAIVLYEGL  LGGASYVNSF  LNILKDVPLA  EREEALGAVS
K.lactis    INLLITLSQS  YYYWTDSISW  IMVLIFYEGL  IGGSSYVNCF  MNILKDVDPN  EREEVLGSVS
M.grisea    ANLVLLTLHA  VLDFIPSVYL  VEAVIFWEGL  LGGAVYVNTF  AEIMENVPAA  DREESLGATS
N.crassa    ANLVLLTLHA  LLNFIPSVYI  VEVIIFWEGL  LGGCVYVNTF  AEIMEHVPAE  DREESLGATS
P.troglody
Consensus   .nl.......  .....ps...  .f..i..egl  lgg..yvntf  ..i.......  .ref...a.s


            481                                      520
H.Sapiens3  ISDTLGISTS  GLIALPIHDF  LCQLS
C.familiar  ISDTLGISTS  GLIALPIHDF  LCHLS
M.musculus  ISDTLGISTS  GVIALPIHDF  LCHLP
R.norvegic  ISDTLGISLS  GVIALPIHDF  LCHLP
D.melanoga  QSDSYGIALA  GFIAIPVHNA  ICGLPAAARS  LVW
A.gambiae   ISDSVGIALA  GIAAIPSHNA  ICDLPVPDRL  L
C.elegans   MGNSLGTNIA  AFISIPIYNW  VCTMKAPGR
S.pombe     SSDSSGIFLA  SLVSLFIEPS  LCHFQA-DRG  RDWCALT
S.cerevisi  IADSFGVFLA  ALIGLGIEPK  LCRHQI-ADD  RPWCRME
E.gossypii  ISDSSGITLV  AEIGILIEPV  LCSHQV-KTG  RPWCQLE
K.lactis    ISDSLGTLIA  AFIGIILEPS  LCKHQI-GTG  RPWCRME
M.grisea    VSDSGGICLA  SLIGMVIEVS  LCDYQV-RNG  RDWCRRIEAS
N.crassa    VSDSGGICVA  SFISMAIEIW  LCNWQV-DHG  RDWCRRIKAG
P.troglody
Consensus   .sds.gi.la  ..l...l...  lc........  ..w.......
```

Fig. (5). Conserved Domains of human CLN3.

The conserved domain is the human CLN3 protein indicated by the box in grey. A missense mutation of glutamic acid (E) to lysine (K) at position 295 in the human protein has been implicated in Juvenile neuronal ceroid lipofuscinosis (Batten disease). It is 409 amino acids in length and aligns 92% with the human CLN3 protein from residue 4 to residue 368. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in gray. All other neutral residues are not highlighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.

CLN5 (FINNISH VARIANT LINCL)

Disease Progression

The Finnish variant of late-infantile NCL (Finnish vLINCL or variant Jansky-Bielschowsky disease) is de-scribed in Finnish patients with onset at 2 to 7 years of age due to mutations in the *CLN5* gene. The children are afflicted with problems in concentration, motor coordination, mental retardation, visual failure, ataxia, myoclonus and epilepsy [71]. The age at death varies considerably from 14 to 36 years [72-74].
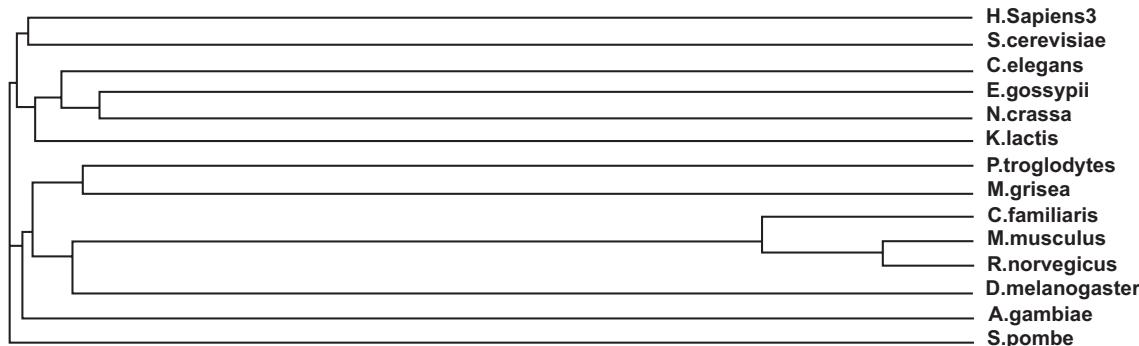
Fig. (6). Phylogenetic tree of CLN3 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 14 homologous sequences are proportional to the amount of inferred evolutionary change.

Disease Autofluorescent Storage Material

Finnish variant LINCL, like the classical LINCL and juvenile variants, contains storage material whose main component is the subunit c of hydrophobic protein mitochondrial ATP synthase. The storage material accumulates in the lysosomes of cells [75] of the patients afflicted by Finnish LINCL.

CLN5 Gene and CLN5 Protein

Finnish vLINCL is caused by defects in the *CLN5* gene present on chromosome 13q22 [76]. It consists of four exons that span 13kb of genomic DNA and encode a 407 amino acid protein, CLN5, which is 60 kD in size.

CLN5 Structure

The CLN5 protein is predicted to be a transmembrane protein [77]. However, co-immunoprecipitation experiments by Isosomppi *et al*. (2002) [78] have suggested that CLN5 may represent a soluble lysosomal glycoprotein which is glycosylated and targeted to lysosomes. Soluble and membrane bound forms of the CLN5 protein may exist due to the use of alternative initiation methionines [78, 79].

CLN5 Function

The function of CLN5 is currently unknown.

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

To date, four disease mutations have been described in *CLN5* [74, 77] of which three result in premature termination of the polypeptide chain. The most common mutation among Finnish *CLN5* patients is a 2 bp deletion resulting in Tyr392Stop. The other two mutations resulting in truncated polypeptides are G1517A and the SWE mutation INS(C) 1961. The fourth *CLN5* mutation, G2127A results in an amino acid substitution of Asp279Asn [78]. All of these sites are highly conserved throughout evolution in *C. familiaris, M. musculus, R. norvegicus* and *P. troglodytes* (Figs. **7** and **8**).

CLN6 (VARIANT LINCL)

Disease Progression

Variant LINCL patients display the same symptoms of the disease as classical LINCL patients but the onset is delayed and the course is milder [80].

Disease Autofluorescent Storage Material

Variant LINCL AFSM material is largely composed of storage bodies in the lysosome whose primary protein content is the subunit c of the mitochondrial ATP synthase [81].

CLN6 Gene and CLN6 Protein

CLN6 present on chromosome 15q21-23 [76] causes variant late-infantile disease (vLINCL) in families of Indian ancestry [36] and in descendants of Spanish settlers in Costa Rica [82]. The *CLN6* gene contains 7 exons that span approximately 23 kb of genomic DNA. A single 2.4kb mRNA is predicted to encode a 311 amino acid that is 30 kD in size, resides in the endoplasmic reticulum (ER), and has unknown function [83-86].

CLN6 Structure

CLN6 is predicted to have 7 transmembrane domains [84] and is predicted to be an ER resident membrane protein [85, 86] . It contains an N-terminal cytoplasmic domain, and a luminal C-terminus [35]. CLN6 contains no asparagine-linked glycosylation sites and can form dimers upon overexpression [85]. Like other NCL membrane proteins (CLN3, CLN5, and CLN8), CLN6 has no homology with known proteins or functional domains, but the sequence is highly conserved across vertebrate species [83].

CLN6 Function

The function of the CLN6 protein is unknown.

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

The majority of vLINCL mutations result in a frameshift or nonsense change, with the introduction of a premature stop codon. However, a Portuguese patient was homozygous for a 3-bp deletion in exon 4 (c.460_462delATC) of *CLN6* which is predicted to remove a single amino acid (I154del) within the predicted third hydrophilic loop of the protein. This residue is within a region of the protein that is highly conserved across at least five species (human, mouse, cow, pig and chicken), suggesting that it is likely to have an important role in the function of CLN6. In addition, a Costa Rican patient was homozygous for an exon 4 missense mutation (c.368G→A) that changes glycine to aspartic acid (G123D) within the predicted third transmembrane domain.

```
                        61                                                                120
            H.sapiens5                     MRRN LRLGPSSGAD AQGQGAPRPG ---LAAPRML LPPASQASRG
            C.familiar
              G.gallus  EEGKQRHPCG FTPFPCIPHV AEQREPRSAA RPSKGAPRTR PRPVGQPRAA GAGAGAALSR
            M.musculus
            R.norvegic  VSGAASGWSE ICVASLQLLP GGQRFFASRA RAVEQRRSSW ARTVEASRSR RSYAQAGARA
            R.norvegic      MGGGLE LLICKPAY-- DGPCKSQEVS GYVICKWWDL AALDLPSDKK DDLFPAGQIV
            P.troglody                     MRRN LRLLPSSGAD AQGQGAPRPG ---LAAPRML LPPASQASRG
             Consensus  .......... .......... ....ps.ga. ....gapr.. .....apr.. ...a.r.

                        121                                                               180
            H.sapiens5  SGSTGCSLMA QEVDTAQGAE MRCAGAARG RASWCWALAL LWLAVVPGWS RVSGIPS--R
            C.familiar          MA QAGSADPCVG GHWAAGPRC- -APWRWALAL LWLATAACG- -----PS--R
              G.gallus  PTSLGPAAAS HGRQRPRHGT AELGASRNAM SAAGRHQWLL ILLLAAACRP CLSGAPHAPQ
            M.musculus                  MLR GGPCGAHWRP ALALALLGLA TILCASPTSG
            R.norvegic  RLSRRRGGAH ARPTPGPGSR RDFRESRMPR VGPGGAHWRP ALALALLGLA AILCASPTSG
            R.norvegic  PTSACSLGIE -RPTPWPGSH RDFRES---- VGPGGAHWRQ ALALALLGLA AILCASPTSG
            P.troglody  SGSTGCSLMA QEVDTAQGAE MRCAGAARG RASWCWALAL LWLAAVQGWS RASGIPS--R
             Consensus  ..s.g...ma q......g.. .r.gag.a.. .a.w.walal lwlaa..g.. ..sg.ps..r

                        181                                                               240
            H.sapiens5  RHWPVPYKRF DFRPKPDPYC QAKYTFCPTG SPIPVMEGDD DIEVFRLQAP VWEFKYGDLL
            C.familiar  RQWPVPYKRF SFRPEPDPYC QAKYTFCPTG SPIPVMKGDD VIEVFRLQTP VWEFKYGNLL
              G.gallus  RRWPVPYRRF DFRPKSDPYC QARYTFCPTG SAIPVMKEED IIEVFRLQAP VWEFKYGDLL
            M.musculus  QRWPVPYKRF SFRPKTDPYC QAKYTFCPTG SPIPVMKDND VIEVLRLQAP IWEFKYGDLL
            R.norvegic  QRWPVPYKRF SFRPKTDPYC QAKYTFCPTG SPIPVMKDND VIEVLRLQAP VWEFKYGDLL
            R.norvegic  QRWPVPYKRF SFRPKTDPYC QAKYTFCPTG SPIPVMKDND VIEVLRLQAP VWEFKYGDLL
            P.troglody  RHWPVPYKRF DFRPKPDPYC QAKYTFCPTG SPIPVMEGND DIEVFRLQAP VWEFKYGDLL
             Consensus  r.WPVPYkRF dFRPkpDP%C QAKYTFCPTG SpIPVMkg#D .IEVFRLQaP !WEFKYG#LL

                        241                                                               300
            H.sapiens5  GHLKIMHDAI GERSTLTGKN YTMEWYELFQ LGNCTFPHLR EEMDAPFWCN QGAACFFEGI
            C.familiar  GHLKIMHDAI GEKSTLTGKN YTMEWYELFQ LGNCTFPHLR EEMNAPFWCN QGAACFFEGI
              G.gallus  GHLKIMHDAV GEKSSLTGKN YTMEWYELFQ LGNCTFPHLR PGMDAPFWCN QGAACFYEGI
            M.musculus  GHFKIMHDAV GERSTLTGKN YTIEWYELFQ LGNCTFPHLR DDKSAPFWCN QGAACFFEGI
            R.norvegic  GHFKIMHDAI GERSTLTGKN YTVEWYELFQ LGNCTFPHLR EEVNAPFWCN QGAACFFEGI
            R.norvegic  GHFKIMHDAI GERSTLTGKN YTIEWYELFQ LGNCTFPHLR EEVNAPFWCN QGAACFFEGI
            P.troglody  GHI-LLYLAP GR
             Consensus  GHlki$hdA. GE.s.ltgkn yt.ewyelfq lgnctfphlr p....apfwcn qgaacf.egi

                        301                                                               360
            H.sapiens5  DDVHWKENGT LVQVATISGN MFNQMAKWVK QDNETGIYYE TWNVKASPEK GAETWFDSYD
            C.familiar  DDIHWKENGT LVLVATISGN TFNQMAKWVK RDNETGIYYE TWTVQASPTK GAETWFESYD
              G.gallus  DDAHWKENGT IVLITKISGT MFNEMAKWVK YDNETGIYYE TWTVQASPDK KSVVWBSYE
            M.musculus  DDKHWKENGT ISSVVATISGN TFAKVAEWVK QDNETGIYYE TWTVRAGPGQ GAQTWBESYD
            R.norvegic  YDKHWKENGT ISLVATVSGN TFAKVAEWVK QDNETGIYYE TWTVRASPGK GAQTWBESYD
            R.norvegic  YDKHWKENGT ISLVATISGN TFAKVAEWVK QDNETGIYYE TWTVRASPGK GAQTWBESYD
            P.troglody
             Consensus  dd.hwkengt l.....isg. .fn..a.wvk .dnetigyye twtv.asp.k ....wf.sy.

                        361                                                               420
            H.sapiens5  CSKFVLRTFN KLAEFGAEEK NIETNYTRIF LYSGEPTYLG NETSVFGPTG NKTLGLAIKR
            C.familiar  CSKEVLRTYK KLAELGAEEK KIETNYTRIF LYSGEPTYLG NETSIFGPTG NKTLALAIKR
              G.gallus  CSKEILRTYQ KLADLGATEN KIQTNYTSII LFSGEPIYLG NETSIFGPTG NKTLALAIKK
            M.musculus  CSNEVLRTYK KLAEFGTEEK KIETNYTKIF LYSGEPIYLG NETSIFGPKG NKTLALAIKK
            R.norvegic  CSNEVLRTYE KLAEFGTDEK KIETNYTKIF LYSGEPIYLG NETSIFGPKG NKTLALAIKK
            R.norvegic  CSNEVLRTYE KLAEFGTEEK KIETNYMKIF LYSGEPIYLG NETSIFGPKG NKTLALAIKK
            P.troglody
             Consensus  cs.f.lrty. kla..g..f. ki.tnyt.i. l.sgep.ylg netsifgp.g nktla.ai..

                        421                                                               480
            H.sapiens5  FYYPFKPHLP TKEELLSLLQ IFDAVIVHKQ FYLFYNFEYW FLPMKFPF-- --IKITYEEI
            C.familiar  FYYPFKPHLS TKEELLSILQ IFDAVIIHRE FYLFYNFEYW FLPMKFPF-- --IKITYEEI
              G.gallus  FYYPFKPHKT VREEFVDLIK IIDRVILNHQ FYLFYNFEYW FLPMKFPY-- --LIVVYEEV
            M.musculus  FYGPFRGYLS TKDELMNFIK IFDTVIIHRQ FYLFYNFEYW FLPMKPPF-- --VKITYEET
            R.norvegic  FYGPFKRYSS TKDELLNFIK IFDTVIMHRE FYLFYNFEYW FLPMKPPF-- --VKITYEET
            R.norvegic  FYGEVKRYSS TKDELLNFIK IFDTVIMHRE FYLFYNFEYW FLPMKPPFPP SAAATLHPPA
            P.troglody
             Consensus  fy.pfky... ...f....l. i.d.vi.... fylfyn.eyw flpmk.p.   .k...yee.

                        481                                    518
            H.sapiens5  PLPIRNKTLS GL
            C.familiar  PLPKRNETLS GL
              G.gallus  PLPIGSKTSS GV
            M.musculus  PLPTRHTTFT DL
            R.norvegic  PLPTQHTTFT DL
            R.norvegic  SIFSRGCAYT GVFGHLHLLG RLVGSTVYVK TPGKKQKP
            P.troglody
             Consensus  plp....t.. ..
```

Fig. (7). Conserved Domains of human CLN5.

There are no known conserved domains in the human CLN5 protein. However, the MSA of CLN5 and its homologs show regions of high sequence similarity that may indicate potential unidentified domains. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in grey. All other neutral residues are not high-lighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.

Fig. (8). Phylogenetic tree of CLN5 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 7 homologous sequences are proportional to the amount of inferred evolutionary change.

The introduction of a charged amino acid is predicted to disrupt this domain [83]. Sharp *et al.* (2003) [87] identified 8 mutations in *CLN6* bringing the total number of mutations found in this disorder to 18, of which 10 mutations affected single amino acids. These mutations are conserved across the vertebrate species in *C. familiaris, R. norvegicus, M. musculus* and *P. troglodytes* as noted in the multiple sequence alignment (Figs. **9** and **10**).

CLN7/MFSD8 (TURKISH VLINCL)

Disease Progression

Children afflicted by the Turkish variant Late Infantile NCL show an onset of disease symptoms at a mean age of 5.1 years, ranging between 2 and 7 years. They suffer from the classical NCL disease symptoms including epileptic seizures, progressive psychomotor deterioration, visual failure,
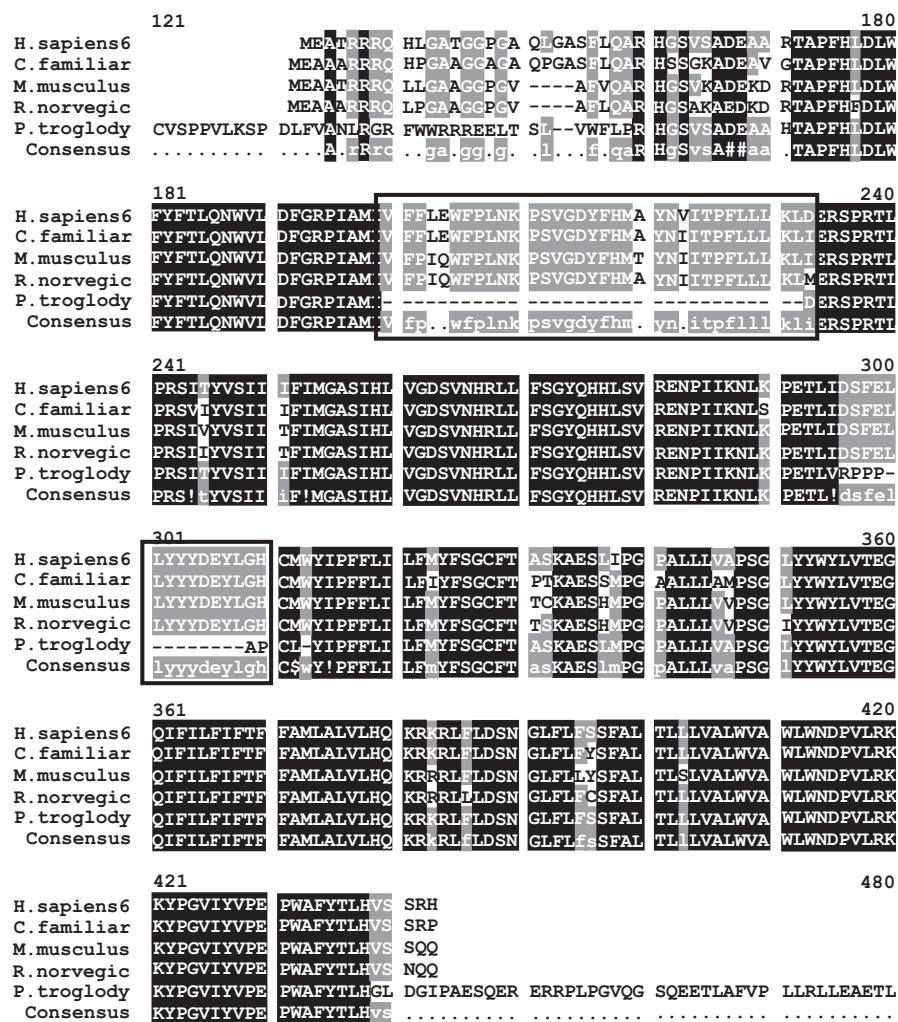
Fig. (9). Conserved Domains of human CLN6.

There are no known conserved domains in the human CLN6 protein. However, the MSA of CLN6 and its homologs show regions of high sequence similarity that may indicate potential unidentified domains. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in grey. All other neutral residues are not highlighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.
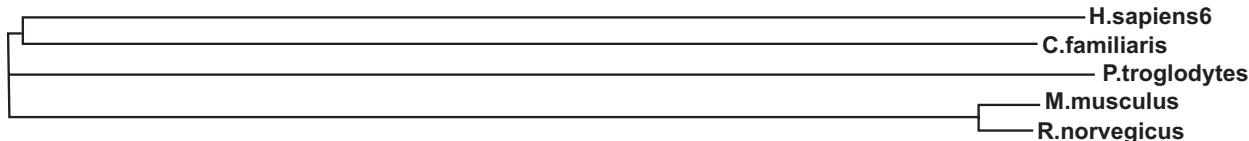
Fig. (10). Phylogenetic tree of CLN6 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 5 homologous sequences are proportional to the amount of inferred evolutionary change.

and premature death. However, the Turkish variant NCL patients suffer from seizures that are more severe than in patients that suffer from classical LINCL [88].

Disease Autofluorescent Storage Material

Biochemical characteristics of AFSM in Turkish variant LINCL patients have not been determined.

***CLN7* Gene and MFSD8 Protein**

The Turkish variant LINCL gene, *CLN7*, was first characterized by Wheeler *et al*. [89] and was believed to be allelic to previously characterized NCL genes- *CLN8* [90, 91] and *CLN6* [92]. However, Siintola *et al*. [93] mapped *CLN7* to a unique locus on chromosome 4q28.1-q28.2 using a genomewide scan with Single Nucleotide Polymorphism (SNP) markers and homozygosity mapping. *CLN7* or the *MFSD8* gene belongs to the major facilitator superfamily of transporter proteins and encodes MFSD8, a putative lysosomal transporter.

MFSD8 Structure

MFSD8 (Major Facilitator Superfamily Domain-containing protein-8) is predicted to be a 518 amino acid protein that localizes mainly to the lysosomal compartment. It is approximately 58 kD in size with 12 predicted transmembrane domains. MFSD8 is expressed ubiquitously with several alternative splice variants that were detected by Northern blot and Expressed Sequence Tag (EST) database analysis [93].

MFSD8 Function

The function of MFSD8 is unknown.

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

Analysis of the MFSD8 amino acid sequence using the Pfam (Protein FAMily) domain database revealed that MFSD8 contains a Major Facilitator Superfamily (MFS) domain and a sugar transporter domain between amino acid positions 42 - 477 and 72 - 147 respectively. A BLAST search of MFSD8 returned several homologs for MFSD8 in different species- *H. sapiens, P. troglodytes, M. musculus, R. norvegicus, G. gallus, D. melanogaster, D. rerio, A. gambiae* and *C. elegans,* suggesting that it is an evolutionarily conserved protein (Figs. **11** and **12**).

Of the six mutations that Siintola *et al*. (2007) [93] identified in the *MFSD8* gene two of them- GLY310ASP and GLY429ASP resulted in amino acid substitutions in exon 10 and 12 respectively, in the Turkish patients carrying the defective *MFSD8* gene. A third mutation, TYR298TER, was

identified in an Indian patient. This results from a transversion event in exon 10 of the *MFSD8* gene which gives rise to a truncated protein.

CLN8 (EPMR/NORTHERN EPILEPSY)

Disease Progression

Northern epilepsy, also known as progressive epilepsy with mental retardation (EPMR), is caused by a Finnish founder mutation in the *CLN8* gene. It has the most protracted course of all the NCLs and is characterized by the onset of generalized seizures between 5 and 10 years of age and subsequent progressive mental retardation. Visual problems are not severe, myoclonus does not exist and the clinical progression of the disease is much slower [71, 91].

Disease Autofluorescent Storage Material

Northern epilepsy, like Finnish vLINCL is pathologically characterized by intraneuronal cytoplasmic deposits of autofluorescent granules. Mitochondrial ATP synthase subunit c is the main stored protein in both disorders [71].

***CLN8* Gene and CLN8 Protein**

CLN8 encodes a ubiquitously expressed 286 amino acid transmembrane protein [94]. It has been suggested that CLN8 is an endoplasmic reticulum (ER) resident protein that recycles between the ER and ER–Golgi intermediate compartment (ERGIC) in non-neuronal cells and the ER in neuronal cells [95].

CLN8 Structure

CLN8 is a 33kD non-glycosylated transmembrane protein. It contains an ER-retrieval signal KKRP in the C-terminus (aa 283-286) [95]. Sequence homology links CLN8 to a large eukaryotic protein family of TLC-domain homologs (TRAM, Lag1, CLN8 homology domain; SMART accession number SM00724), (Fig. **13**) [96]. Members of this family have been shown to facilitate translocation of nascent polypeptide chains into the ER and export of glycosylphosphatidylinositol-anchored proteins out of the ER [35, 97-99].

CLN8 Function

The function of CLN8 protein is unknown.

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

Northern Epilepsy patients have been observed to carry some of the following mutations, ARG24GLY, TRP263-CYS, ARG204CYS, 1-BP DEL, 88G, ALA30PRO, and 1-BP DEL, 66G. All these residues are highly conserved in all vertebrate species that have homologs of CLN8-

```
           121                                                         180
H.sapiens7                 MA GLRNESEQCP LIGD-TPCSR EWDILETEEH
P.troglody                 MA GLRNESEQCP LIGD-TPCSR EWDILETEEH
M.musculus                 MA NLGSEARCP  LIGPGSPCSR EWSEIETQEH
R.norvegic LPSVSTAPAR KELRAREWAA RRLVSAVIMA NLGSEARCP LIGTGSPCSR EWGMIDTQEH
G.gallus   VGGSRRSRRS RRRGARKRPE LGMAAAEGRE PLLRAGECRP ERQQQQQCR- --DVVETEEH
D.rerio                     M PLEESGCNTP LIRG------ --EIISDGDT
D.melanoga             MEFAR RVSARFCTKR LPEDVDDC-- ----LETLEE
A.gambiaeS                                       DC-- ----LETAAE
C.elegans            MGVYPSESEI STPSKDSCIS LDSIRSPLHE
Consensus  .......... .......... ......se.c. ..l......g. .....et..e

           181                                                         240
H.sapiens7 YKSRWRSIRI LYLTMFLSSV GFSVVMSIW  PYLQKIDPTA DTSFLGWVIA SYSIGQMVAS
P.troglody YKSRWRSIRI LYLTMFLSSV GFSIVMSIW  PYLQKIDPTA DTSFLGWVIA SYSIGQMVAS
M.musculus YKSRWKSVRI LYLTMFLSSV GFSIVIMSIW PYLQKIDQTA DASFLGWVIA SYSIGQMVAS
R.norvegic YKSRWKSVRI LYLTMFLSSV GFSIVIMSIW PYLQKIDQTA DASFLGWVIA SYSIGQMVAS
G.gallus   YKSRWRSIWI MYLTMFLSSV GFSIVIMSIW PYLQKIDPTA DASFLGWIIA SYSIGQMVAS
D.rerio    DRSRWRSIRV MYTMFLSSV  GFTIVITSIW PYLKKIDESA DASFLGWVVA AYSIGQMVAS
D.melanoga YKQRWRSVRI IWFTMFYMAL GFSIILTGIW PYLNKLDPLA GKEFMGEIVA ANPIGQMILS
A.gambiaeS YTERWTSVRI IYYTMFTMSL GFSIILTGVW PYLDKLDFEA GKEFMGWIVG ANPVGQMILS
C.elegans  RSTEWRAMWI SIAMQFVVGV QISVYYMSVW PYISGLDKTA DMDFLGWVVA ACNMGCTITN
Consensus  y...rWre.r! .y.tmFl.Sv gfs!!.ms.W PYL.klDptA d..F$Gw!!a a...Ggm!.s

           241                                                         300
H.sapiens7 PIFGLWSNYR PR-KEPLIVS ILISVAANCL YAYLHIPASH NKYYMIVAFG IIGICAGNVA
P.troglody PIFGLWSNYR PR-KEALIVS ILISVAANCL YAYLHIPASH NKYYMIVAFG IIGICAGNVA
M.musculus PLFGLWSNYR PR-KEPLIVS ISLSVAANCL YAYVHVPAAH NKYYMIIAFG IVGICAGNVA
R.norvegic PLFGLWSNYR PR-KEPLIVS IFISVAANCL YAYVHVPAAH NKYYMITAFG IVGFCAGNVA
G.gallus   PLFGLWSNYR PR-REPLVVS TAISVAANCL YAYVHIPHSH NKYYMITARA LVGFCAGNVA
D.rerio    PFFGLWSNHR PR-REPLVCS IFINVSANIV YSYVYLPPSH NQIHMLLAFT FVGICAGNVA
D.melanoga PIFGWWQNKL GKIRIPLLLS LSLFTLASGI YSSLETRPDN VKYYMISSLF IIGVSSANIA
A.gambiaeS PLVGWWCNRL GSIRIPLLCS LALFSIASGI YACLETFPTD QKYWMLYSLF IIGVSSSSVA
C.elegans  PLYGLWNQKT LSCKWPTVTG FLIAAVGQVM YGAISEVEQY GKWYMLVARV VTGLCVIA
Consensus  PlfGlW.#.. ...r.Pl..S ..i...a..l Y....l.... kvvML.aR. l.G.g.gnvA

           301                                                         360
H.sapiens7 VVRSYTAGAT SLQERTSSMA NISMCCALGF ILGPVFQTCF TFLGEKGVTW DVIKIQINMY
P.troglody VVRSYTAGAT SLQERTSSMA NISMCCALGF ILGPVFQTCF TFLGEKGVTW DVIKIQINMY
M.musculus VVRSYTAGAT SLQERTNAMA NTSTCCALGF ILGPVFQTCF ALIGEKGVTW DIIKIQVNMY
R.norvegic VVRSYTAGAT SLQERTGAMA NTSTCCALGF ILGPVFQTCF ALIGEKGVAW DIIKIQINMY
G.gallus   VVRSYTAGAT SLTERTSAMA NTSACCAVGF ILGPVFQTCF TLIGEEGITW KSIHIQLNMY
D.rerio    VVRSYVAGAT SLKERTSAM  NM-ACCALGF ILGPALQAVL SFIGETGVDW DVIKIQVNMY
D.melanoga LCRSYLSAAT RISERTHAVA MVSLACVLGF IIGPTLQAAV TLPLGDQGHVW LWGKLHFNMY
A.gambiaeS VCRSYLSAAT KVKERTGAVS MVSLACTLGF IVGPVLQGAV TMFGEDGLPL VKNKIHINMY
C.elegans  ALRAYGTAS  TPKDRLKAIS YGTAGFVFCI SFGPAISFF  TPLGETGFNI WIVK--INMY
Consensus  v..RsY.a.At ..k#Rt.A.a ..s..g.lGf i.GP..qA.f t..G#.G... ..kl..NMY

           361                                                         420
H.sapiens7 TTPVLISAFL GILNIIILI- AILREHRVDD SGRQCKSINF EEASTDEA-Q VPQGNIEQVA
P.troglody TTPVLISAFL GILNIIILI- AILREHRVDD SGRQCKSINF EEASTDEP-Q VPQGNIDQVA
M.musculus TAPVLIAFL  GILNIIILI- EILREHRVDD LGRQCKSVNF QEENTDEP-Q IPEGNIEQVA
R.norvegic TAPVLIASFL GILNIIILI- EILREHRVDD LGRQCKSVNF PEENADEL-Q IPEGSIEQVA
G.gallus   TAPVLFGALL GIINIIITF- AIFREHRVDD MGRQCNSINF EREEGGVLDQ DAEGNIEHVA
D.rerio    TAPALIACF  GIINIIIVI- LVLREHSVDD HGNRIRPINY TPEERVEVPP EVEGDIEHIA
D.melanoga TTSGWINVLM SIGNFMMFIP GVFEEHKI-- ---AREVMV MQGGTSEKET WKGIEPSYLS
A.gambiaeS TTSGWINVLM GILNFCIFIP EVFKEKRI-- ---ALEAMK QQGMQSEKET WKSMEPIYLS
C.elegans  TSAAYMAFI  CVILAAVIMV FQELY----- --AGIIDKN SSDEKQCADY VVVPKFLVP
Consensus  Ta...l.a..g!ln.$.l.f...#.h...... .e........ .......c.. ....kd.v.

           421                                                         480
H.sapiens7 VVAINVLFFV TIFIFALFET IITPLTMDY  AWTQEQAVLY NGIILAALGV EAVVILGVK
P.troglody VVAINVLFFV AIFIFALFET IITPLTMDY  AWTQEQAVLY NGIILAALGV EAVVILGVK
M.musculus VVAINIVFFV VIFIFAVYET IILTPLTMDY AWTQEQAVLY DGILLVAFGV EAVLVIMGVK
R.norvegic VVAINIVFFV VIFIFAVYET IILTPLTMDY AWTQEQAVLY NGIILVAFGV EAVLVIMGVK
G.gallus   VVAVNFLFFV IFIFAVFET  IATPLTMDY  SWTREKAVEY NGIILVAFGV ESVIVMVVK
D.rerio    VFTSNVLFFI IIFIFAVFET ISTPISMDME AWTRKEAVYY NGIILAAIGF ESILVLVAK
D.melanoga AWTLIVSFFI LVFNVLIET  LGTSLTMDME AWSNDEALWY MGIMMTTAAT VCSLVTVLE
A.gambiaeS AWTLISAFFI LVFNVLIET  LATPLTMDME AWKAFALYMAW MAVGALASAM CLMLD
C.elegans  ALVCIYLYMI VSMATNIEV  MSTPLTTVLY DWKDSQAILY MGIIESITCI VSVLINFGTG
Consensus  a...i.l%f!...fif...Et..LpLtmd$%.awt..A.lYngii.....i.sv..f..l.

           481                                                         540
H.sapiens7 LLSKKIGEEA IILL-GGIIVV WVGFFILIPW GNQFPKIQWE DLHNNSIPNT TFGEIIGLW
P.troglody LLSKKIGEEA IILL-GGIIVV WVGFFILIPW GNQFPKIQWE DLHNNSIPNT TFGEIIGLW
M.musculus LLSKKIGEEA IILL-GGFVVV WVGFFILIPW GNQFPKIQWE DLHNSSTPNT TFGEIIGLW
R.norvegic LLSKKIGEEA IILL-GGFVVV WVGFFILIPW GNQFPKIQWE DLHNSSTPNT TFGEIIGLW
G.gallus   TLSKKTGEEA IIH-GGILII LICFFILIPW GKKLPNIQWQ EIKNNSVPQT TPSEMVMPFW
D.rerio    VVSARVGEEP ILL-GGILIL FVGFFMLIPW GNQYPKIQWT DINNNTIPPI QLAP------
D.melanoga PMCKIIALIY VLIWGGFSLM FLGRVLFVPW GPDEPKIAQP FNASW--NLS ESDPI-----
A.gambiaeS PLCKRFPEQQ VLIWGGFFLM VLCRAVYIEM SDTPEPKIALP ENATLAVDFD YSGPIPTA--
C.elegans  KTRIGKIKRIQILFGIGVF VAFHVVNYPW WFYSGPIHFL PPGSNTTIVG ----------
Consensus  ...k...#r..ili.gGl.....g.....Pw.g...pkl......n.......l.......

           541                                                         600
H.sapiens7 K--SPMEDDNERPTGCSIE QAWCLYTIVI HLAQFLTSAV LI-GLGYPVC NLMSYTLYSI
P.troglody K--SPMEDDNERPTGCSIE QAWCLYTIVI HLAQFLTSAV LI-GLGYPVC NLMSYTLYSI
M.musculus N--SSREDHSEQPTGCIE QTWCLYTIVI HLAQFLTSAV LI-GTGYPAC SVMSYTLYSI
R.norvegic N--SSREDHSEQPTGCIE QAWCLYTIVI HLAQFLTAAV LV-GIGYPAC SVMSYTLYSI
G.gallus   SFRAMQLSSNHSEPAGCIT QSWCLNTMI YLAQYISSDI LI-GLGYPVC NVMSYTLYSK
D.rerio    ----TPASNSLEPIGCSE QTWCLFTIVI HLAQYITSDI LI-GVGYPTC NVMSYTLYSK
D.melanoga ----------FIGCTT QKWCGELALTLTQFIGFAET-SIGYPIGVTLIQLIFSK
A.gambiaeS ---GTETLDGGIDAIGCPS QEWCKTTRGM TISQFLVGYAET-AIGYPIGVTLITLIFSK
C.elegans  ---------------GCLDS YSWCNTTRV PMATYFAFV FFGTGYFV ESPAAALYSE
Consensus  ...........GCp..g.wc.tg...laq%i......f..gig%P.....tl%st
```

(Fig. 11) contd….

Fig. (11). Conserved Domains of human CLN7.

The conserved domain of human CLN7 is Major Facilitator Superfamily. (pfam07690) indicated by the box shaded in gray that spans the multiple sequence alignment (MSA). It is 346 amino acids in length and aligns 82.9% with the human PPT1 protein from residue 42 to residue 354. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in grey. All other neutral residues are not highlighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.
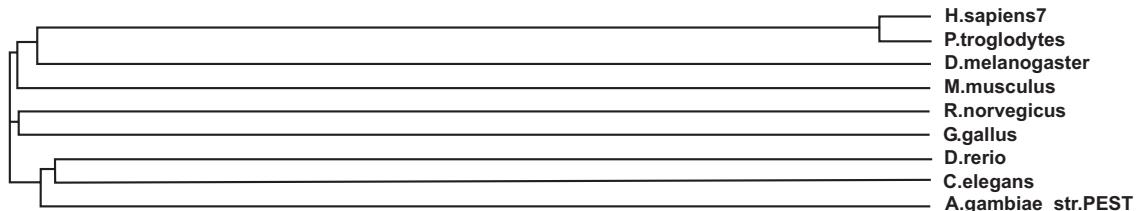

Fig. (12). Phylogenetic tree of CLN7 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 9 homologous sequences are proportional to the amount of inferred evolutionary change.

H. sapiens, P. troglodytes, C. familiaris, M. musculus, R. norvegicus, G. gallus and *Vibrio* sp. MED222 (Figs. **13** and **14**).

CLN10/CATHEPSIN D (CONGENITAL NCL)

Disease Progression

Congenital NCL (CNCL) is a rare congenital disorder that was first described in 1941 [100]. It is characterized by microencephaly, rigidity, seizures, and respiratory difficulties resulting in death usually within a few hours or weeks after birth. This disorder has been identified in about ten individuals who upon post-mortem examination reveal a small, firm brain with severe neuronal loss, gliosis, white matter lacking myelin and accumulation of storage material within cells in the brain and the reticuloendothelial system [100-105].

Disease Autofluorescent Storage Material

Sintolla *et al.* [105] showed that the storage material typical of CNCL affected individuals stained positively for the sphingolipid activator protein D which is also found only in INCL patients.

CLN 10/ CTSD Gene and Cathepsin D Protein

CTSD gene consists of 9 exons and is located on chromosome 11p15.5 [106-108]. It encodes a 412 amino acid pro-

tein cathepsin D (CTSD) which is a lysosomal aspartic protease that belongs to the pepsin family [109, 110].

Cathepsin D Structure

Mature cathepsin D is a two-chain, glycosylated, lysosomal aspartic protease [111].

It is classified in the A1 family of aspartyl proteinases. Cathepsin D consists of two polypeptides encoded by the *CTSD* gene. The *CTSD* gene first encodes a preproenzyme which undergoes several proteolytic processing steps resulting in a single-chain active polypeptide that is 43kD in size. This active single chain polypeptide is further processed resulting in a mature form consisting of two polypeptides [109] interlinked by disulphide bridges [112]. Both the polypeptides contain an aspartic acid residue essential for the enzymatic activity of the mature protein [109, 113]. The mature protease is 31kD in size.

Cathepsin D Function

Cathepsin D is a ubiquitously expressed lysosomal aspartic protease that belongs to the pepsin family [110]. There are several proteins described to function as substrates of CTSD *in vitro* but the *in vivo* substrates are still unknown [35]. Aspartyl proteinases consist of two domains each of which contains an aspartate residue. The residues come together and link to a water molecule at the active site where

```
                  1                                                          60
H.sapiens8        MNEA SDGGTSESIF DLDYAS GIE STLMVAGE VF YLGVE VVCE Q
P.troglody        MNEA SDGGTSESIF DLDYAS GIE STLMVAGE VF YLGVE VVCE Q
C.familiar        MTEM SDGGTSESIF DLDYTS KIE STLAVAGE VF YLGVE VVCE Q
M.musculus        MTEV SSHGLAESIF DLDYAS KIE STLAVAGE VF YLGVE VVCE Q
R.norvegic        MTEV SNHGVAESIF DLDYAS KIE STLAVAGE VF YLGVE VVCE Q
G.gallus          MNEA NSNQVECTIF DWDYIL EWE LTFIAAGE LI YLGVE LIS W
vibriosp.M   MIAEIGHFAM ILSLGLALLL SVLPLYCAAR NNTLIMNSAE PLSWGMFGFL AISFEILGYA
Consensus    .................m.p...s.....g.if#.dyl.w..P..t....agf...ylgvE.lch.

                  61                                                         120
H.sapiens8        L-SSSLNAT  RSIVAREKVI WDIAATRAVE GVQSTAAGLW ALL--GDPVLHADKARGQQN
P.troglody        L-SS SVNAPC RSIVART TEK GDIA TRAVE GVQSTAAGLW ALI--GDPVLHADKARGQQN
C.familiar        L-SSSLNAT  RSIVAREKVI WNIAATRAVE GVQSTAAGLW ALL--VDPVLHADKARGQQN
M.musculus        L-SSSLNAT  RSIAAKEKVI WNIAATRAVE GVQSTTAGLW ALL--GDPVLYADKALGQQN
R.norvegic        L-SSSLNAT  RSILAKEKVI WNIAATRAVE GIQSTAAGLW ALL--GDPVLYSNKALGQQN
G.gallus          L-SSRMST V RAIYKEKVI WNMAVTHGAI GIOSCVAGLW ALL--IDPVFHADKVYSOQK
vibriosp.M   FYTNDFTIC VASNSNSQLP WYYRIAVW CAHEGSLLIV VLIQACGWTIA VATFSRGMPQ
Consensus    l.ss....t.ral.a.ekvi.W...a.TraviG.qs..agLWaLL..gdpV..Adk.rgqq.

                  121                                                        180
H.sapiens8        WCWFHITT T GFFC--FENV AVEISNLIFE TFDLFLVIHH LFAFLGFLGC LVNLQAGHYL
P.troglody        WCWFHITT T GFFC--FENV AVEISNLAFE TFDLFLVIHH LFAFLGFLGC LVNLQAGHYL
C.familiar        WCWFHIAT T GFFF--FENV AVEISNVLFR TFDLFIAIHH LFAFLGFLGC VVNLGAGHYL
M.musculus        WCWFHITT T GFFF--FENV AVEISNIFFR TFDLFLVVHH LFAFLGFLGC AINLRAGHYL
R.norvegic        WCWFHITT T GFFL--FENA AVEISNIFFR TFDIFLVHH LFAFLGFLGC AVNLRAGHYL
G.gallus          WCWENCIIA A GFFL--IENV AVEVSNIVFE TFDVFIGGSS LTCEWWLGWS SNQVKSGHYL
vibriosp.M   ESVARVIAIM CLITVGFLLF IIVTSNPIIE TLPFEPVDGR DLNPILQDPGIIIHHPPMLYI
Consensus    wswf..I.a.Gff...fenv.a!h.SN.ffRTfd.Flv...l..fl.....l.....ghY$

                  181                                                        240
H.sapiens8        AHTTLLLEMS TPFTCVSWMI LKAGMSE-SL FWKLNQWLMI HMFHC--RMV LTYHMWWVCF
P.troglody        AHTTLLLEMS TPFTCVSWMI LKAGWSE-SL FWKLNQWLMI HMFHC--RMV LTYHMWWVCF
C.familiar        AHSTLLLEAS TPFTCISWMI LKAGWSE-SL FWKLNQWLMI HMFHC--RMV LTYHMWWVCF
M.musculus        AHTTLLLEMS TPFTCISWMI LKAGWSD-SL FWKANQWLMI HMFHC--RNI LTYHMWWVCF
R.norvegic        AHTTLLLEMS TPFTCVSWMI LKAGWSH-SL FWKVNOWLMI HMFHC--RNI LTYHMWWVCF
G.gallus          PLMCMLLEMS TPATCISWMI LKACCAN-TF FWKANQVMI HLFHC--RMI LTYHMWWVCI
vibriosp.M   GYVCFSVAFS FAIASLMSGR LDTAWARWSR EWTIAAWIFL TVGIALGSWW AYYEIGVGGW
Consensus    ...g.llemS tp.tc.swmI Lkagwa..s..fWk.nqWlmi.h.fhc..rm..ltYh$wWvc.

                  241                                                        300
H.sapiens8        WH  GLVSSL Y PHLTLFLV GIALLILIIN PYWTHKKTQQ LINPVDWNFA QPEAKS-RPE
P.troglody        WH  GLVSSL Y PHLTLFLV GIALLILIIN PYWTHKKTQQ LINPVDWNFA QPEAKS-RPE
C.familiar        WH  GLVSSL Y PHLALFLV GLGLLTLVIN PYWTHKKTQQ LINPVDWNFA QPAPRSSRPA
M.musculus        CH  ALTSSL H PHWALFLF GIALLTAVIN PYWTHKKTQQ LLHPVDWNFA QEEAGSRQE
R.norvegic        QH  ALASSL Y PHLALFLF GIALLVIIN PYWTHKKTQQ LINPVDWNFA QPEAKGDRQE
G.gallus          FN  SVVENL G LHFIVLFS GIFAFLILN PYWTYKKTQQ LISPTDWNF- --ENKATENG
vibriosp.M   WFNDPVENAS FIEWLAG--T AIMHSLAVTE KRGIFKAWTV LIAISAFSLS LLGTFLVRSG
Consensus    w.w.#..v...l..$phla....gl...t.!!#.pywT.kktqq.LL..p.dwnf....e.k..r.g

                  301                              360
H.sapiens8        GENGQLLRKK RP
P.troglody        GENGQLLRKK RP
C.familiar        GANGQVPQKK GQ
M.musculus        RNNGQVPRKK RL
R.norvegic        RNNGQVPRKK RL
G.gallus          KINGETSQKK RM
vibriosp.M   IVSVHAFAS DEARGMFILG FLVFVIGGSL LLFAVKGASV RVRGNFDLVS RENALLGNNI
Consensus    .ing....kk.rg........ .......... .......... .......... ..........
```

Fig. (13). Conserved Domains of human CLN8.

The conserved domain of human CLN8 shows a high degree of conservation with the TLC, TRAM, LAG1 and CLN8 homology domains (smart00724). These protein domains contain at least 5 transmembrane alpha-helices. Lag1p and Lac1p are essential for acyl-CoA-dependent ceramide synthesis, TRAM is a subunit of the translocon and the CLN8 gene is mutated in Northern epilepsy syndrome. The family may possess multiple functions such as lipid trafficking, metabolism, or sensing. Trh homologues possess additional homeobox domains. The conserved domain is indicated by the box shaded in gray that spans the multiple sequence alignment (MSA). It is 208 amino acids in length and aligns 95.7% with the human CLN8 protein from residue 67 to residue 258. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in grey. All other neutral residues are not highlighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.
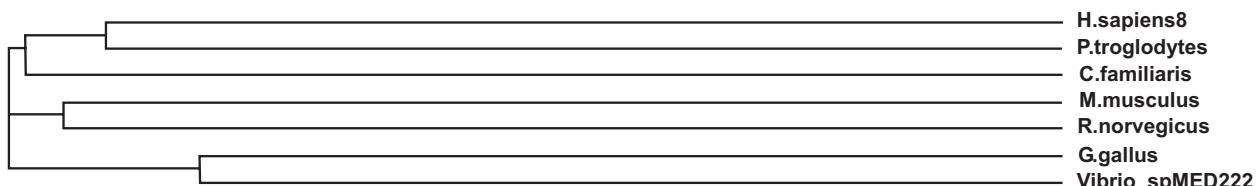
Fig. (14). Phylogenetic tree of CLN8 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 7 homologous sequences are proportional to the amount of inferred evolutionary change.

the substrate peptide bond is hydrolyzed. Hence, mutation of the aspartate residues which may be present at distant regions of the protein, results in elimination of enzymatic activity of cathepsin D without affecting its processing [114, 115].

Evolutionary Relationships- Alignment, Conserved Domains and Mutations

Cathepsin D is a conserved protein through evolution and has 15 homologous gene sequences present in *H. sapiens, C. familiaris, M. musculus, R. norvegicus, G. gallus, D. rerio, D. melanogaster, A. gambiae, C. elegans, S. cerevisiae, K. lactis, M. grisea, N. crassa, A. thaliana* and *O. sativa* (Figs. **15** and **16**). Though the aspartyl protease domain (pfam00026) is conserved in all of the homologs, the A1 propeptide

(pfam07966) is present only in humans and fruitflies. Cathepsin D mutant flies exhibit several features related to NCL pathology such as the progressive accumulation of AFSM in neurons and modest neurodegeneration [116]. It is interesting to see that A. thaliana and O. sativa have SapB_2 (pfam03489, Saposin-like type B, region 2) and SapB_1 (pfam05184, Saposin-like type B, region 1) conserved domains present in their homologous cathepsin D protein sequence as saposin D is a component of the ultrastructural AFSM that accumulates in CNCL patients. Sheep [117] and mice [118] deficient in functional cathepsin D protein also recapitulate CNCL phenotypes and are being exploited as model organisms to study the disease. Some of the important mutations occurring in patients that results in CNCL are PHE229ILE, TRP383CYS [119] and Y255X [105].

[Figure: Multiple sequence alignment of cathepsin D homologs from H. sapiens, C. familiaris, M. musculus, R. norvegicus, G. gallus, D. rerio, D. melanogaster, A. gambiae, C. elegans, S. cerevisiae, K. lactis, O. sativa, with consensus line; residue positions 1–360 across six alignment blocks.]

(Fig. 15) contd….

```
        361                                                                      420
H.sapien10  ITLKLGGKGY  KLSPEDYTLK  VSQA-GKTLC  LSGFMGMDIP  PPSGPL--WI  LGDVFIGRYY
C.familiar  VTLKLGGKLY  KLSSEDYTLK  VSQG-GKTIC  LSGFMGMDIP  PPGGPL--WI  LGDVFIGCYY
M.musculus  VYLKLGGKNY  ELHPDKYILK  VSQG-GKTIC  LSGFMGMDIP  PPSGPL--WI  LGDVFIGSYY
R.norvegic  ITEKLGGQNY  EIHPEKYILK  VSQA-GKTIC  LSGFMGMDIP  PPSGPL--WI  LGDVFIGCYY
G.gallus    VTLMLGGKPY  QLTGEQYVFK  VSAQ-GETIC  LSGFSGLDVP  PPGGPL--WI  LGDVFIGPYY
D.rerio     ISFSLGGKVY  SITGEQYILK  ESQG-GHDIC  LSGFMGLDIP  PPAGPL--WI  LGDVFIGQYY
D.melanoga  IKEVLGGKTF  EIEGKDYILR  VAOM-GKTIC  LSGFMGLDIP  PPNGPL--WI  LGDVFIGKYY
A.gambiae   ITFTLGGKQF  TIEGADYILR  VAOM-GKTIC  LSGFMGIDIP  PPNGPL--WI  LGDVFIGKYY
C.elegans   VSEVIGGQEF  SLKGEDYVLK  VSQG-GKTIC  LSGFMGIDLP  ERVGEL--WI  LGDVFIGRYY
S.cerevisi  LIENFNGYNF  TIGPYDYTLE  VSGS-----C  ISAITPMDFP  EPVGPL--AI  VGDAFLRKYY
K.lactis    LTLNFNGYNF  TITAYDYTLE  VSGS-----C  ISAFTPMDFP  EPVGPL--AI  IGDAFLRKYY
O.sativa    DQLRAETKPA  KVCSSVGLCT  FDGTHGVSAG  IRSVVDDEVG  KSSGPFSSAM  CNACETAVVW
Consensus   ..f.lggk.f  .l.g.dy.l.  vsq...g.tic lsgfmg.#.p  pp.Gpl..wi  lgdvfig.yy

        421                                                                      480
H.sapien10  TVFDRDNNRV  GFAEAARL
C.familiar  TVFDRDQNRV  GLAQATRL
M.musculus  TVFDRDNNRV  GFANAVVL
R.norvegic  TVFDREYNRV  GFAKAATL
G.gallus    TVFDRDNDSV  GFAKCV
D.rerio     TVFDRENNRV  GFAKAKSV
D.melanoga  TEFDMGNDRV  GFADAK
A.gambiae   TEFDMGNDRV  GFATAV
C.elegans   SVFDFDQNRV  GFAQAKTADG  RPVDPAPRPF  -RSVFDNESE  ESMEQDDE
S.cerevisi  SIYDLGNNAV  GLAKAI
K.lactis    SIYDIGHDAV  GLAKAA
O.sativa    MHTQLAQNQT  QDLVLQYID-  QLCDRLPSPM  GESSVDCSSL  ASMPDIAFTI  GGNKFVLKPE
Consensus   t.f#...n#rv gfa.a.....  ..........  ..........  ..........  ..........
```

Fig. (15). Conserved Domains of human CLN10.

The conserved domains of human CLN10 i.e., the Cathepsin D protein are A1 Propeptide (pfam07966) and the Eukaryotic aspartyl protease (pfam00026) indicated by the boxes shaded in gray that span the multiple sequence alignment (MSA) from the 5' to the 3' end respectively. The A1 Propeptide domain is 29 amino acids in length and aligns 96.6% with the human CLN10 protein from residue 22 to residue 49. The Eukaryotic aspartyl protease domain is 314 amino acids in length and aligns 100% with the human CLN10 protein from residue 78 to residue 409. The MSA generated by MultAlin highlights individual residues with high consensus value (above 90%) in black and those with low consensus value (above 50%) in grey. All other neutral residues are not highlighted. There are regions of high sequence homology that are not indicated by Multalin; these are shown by black boxes around the conserved regions.

DISCUSSION

Eight NCL genes have been identified- *CLN1*, *CLN2*, *CLN3*, *CLN5*, *CLN6*, *CLN7*, *CLN8* and *CLN10*. Their gene products- PPT1, TPP1, CLN3, CLN5, CLN6, MFSD8, CLN8 and cathepsin D, respectively, are highly conserved. The high degree of sequence similarity between the homologs implies a conserved functional role to each of the corresponding human NCL proteins. Consequently, several model organisms have been used to study the various clinical and pathological features of the NCL genes. However, despite extensive research the functional role of several NCL proteins remains elusive.

The NCLs are neurodegenerative lysosomal storage disorders that show pathological accumulation of autofluorescent storage material in the lysosomes of the tissues of the affected individuals. However, beyond the cursory grouping
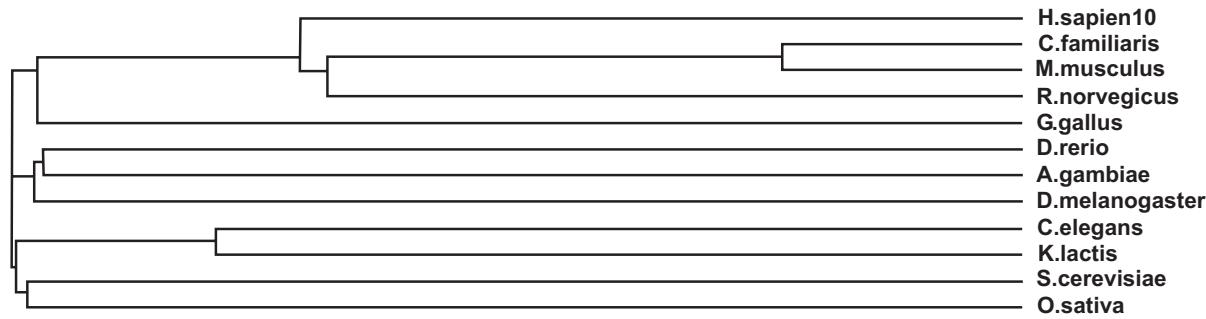
Fig. (16). Phylogenetic tree of CLN10 homologs.

The phylogram tree is generated using the program ClustalW. In the above figure the branch lengths of the 12 homologous sequences are proportional to the amount of inferred evolutionary change.

of these proteins resulting in a disease with similar clinical manifestations, there is not sufficient sequence or domain similarity between the NCL genes to assign a similar function to them or to group them in the same biological pathway.

We have therefore performed a comparative biological analysis of the NCL proteins in the context of phylogeny. A sound classification of gene family relationships is a prerequisite in understanding how genes evolved along with the proteins that they encode. A reliable gene phylogeny is a powerful tool with which the structure and function of uncharacterized proteins can be predicted and mechanisms by which new genes appeared and assumed characteristic functions can be inferred. Such a phylogenic reconstruction also allows a better understanding of how biochemical pathways were established and what the role of the homologous proteins was in their evolutionary ancestors. Therefore, co evolutionary relationships can be analyzed in a meaningful way to elucidate the dynamics among proteins and to understand the links between genomic change and how this change is manifested morphologically [120, 121].

The generation of multiple sequence alignments and corresponding phylogenetic trees of each of the NCL proteins is the first step in attempting to resolve the evolutionary origins

of these individual proteins. A look at when the NCL proteins first appeared in the tree of life allows a grouping of them based on their corresponding cladistic confines. CLN2 appears the earliest in the Bacterial clade in the organisms *Pseudomonas* sp. 101 and *Xanthomonas* sp. T-22. TPP1 has significant sequence similarities to two previously characterized bacterial pepstatin-insensitive carboxyl peptidases from *Xanthomonas* and *Pseudomonas*. CLN2 is synthesized as an inactive zymogen that is autocatalytically converted to an active serine protease at acidic pH in the lysosome [37]. BLAST searches were also able to identify a sequence with homology to CLN2 in *Sulfolobus solfataricus* which is present in the Archae clade (Figs. **3**, **4** and **17**).

Next, CLN1 and CLN10 appear in the eukaryotic clade, Viridiplantae. Their resulting NCL pathology shares the same predominant AFSM of saposins. CLN1 and CLN10 also encode soluble lysosomal enzymes whose functions are known. CLN1 and CLN10 protein sequences contain the conserved domains for palmitoyl protein thioesterase and eukaryotic aspartyl protease, respectively, which span the entire length of the protein. A common role for PPT1 and cathepsin D appears to be in breakdown of lysosomal substrates, either through lipid removal from protein or degradation of proteins, respectively.
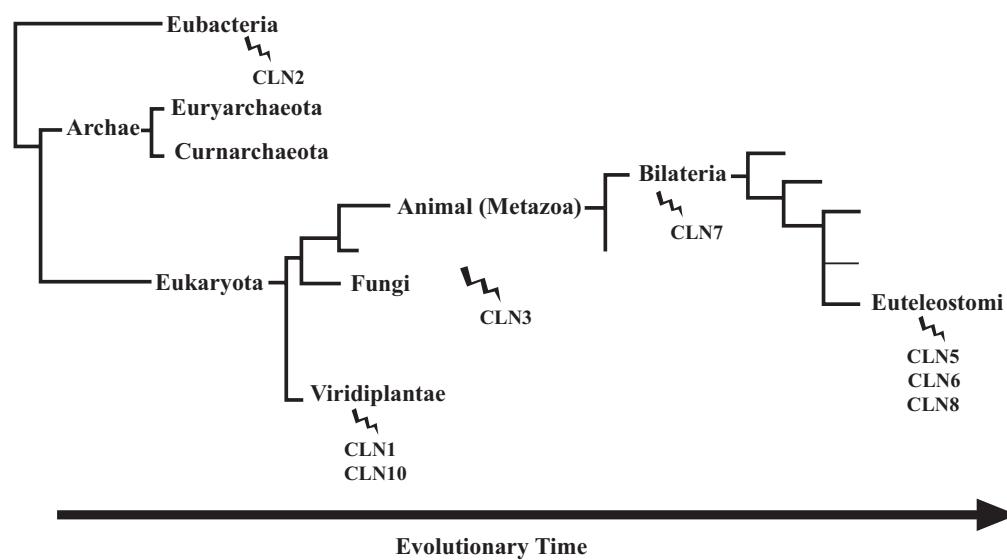
Fig. (17). Clade Confines of the various NCL proteins.

CLN2 is conserved in the Eubacteria clade. CLN1 and CLN10 are conserved in the Viridiplantae group. CLN3 appears in the Fungi and Metazoan group. CLN7 is conserved in Bilateria. CLN5, CLN6 and CLN8 are conserved in the Euteleostomi group.

CLN3 appears in the Fungal/ Metazoan clade. In LINCL and JNCL the primary component of the AFSM is the mitochondrial ATP synthase subunit C. The function of CLN3 is unknown though it is a very well conserved protein that is known to localize to the lysosomal membrane.

CLN7 or MFSD8 appears in the Bilateria clade of the metazoan organisms that include the nematode worms, *Caenorhabditis elegans*. *C. elegans* are the only homolog of CLN7 that also contain the Major Facilitator Superfamily domain found in humans. Lastly, CLN5, 6 and 8 appear in the more recent Euteliostomi clade in the organism *Danio rerio* also known as Zebra fish. In addition to its eukaryotic homologs identified by HomoloGene, BLAST searches identify a bacterial species protein in *Vibrio* sp. MED222 that shares sequence similarity with CLN8. CLN5 and CLN6 multiple sequence alignments indicate that there is a high degree of sequence conservation among the homologs however; no conserved domains have been identified. CLN8 contains the conserved domain TRAM which contains at least five transmembrane alpha helices. TRAM containing proteins may possess multiple functions such as lipid trafficking, metabolism, or sensing. However, the precise function of CLN8, like CLN5, CLN6 and CLN7 is unknown. CLN5 is the only soluble lysosomal enzyme among these more recently evolved proteins. The remaining are all transmembrane proteins occurring in various organelles in the endosomal/ lysosomal pathway. It is possible that CLN1 and CLN10 as well as CLN5, CLN6 and CLN8 that are subsequently confined to the same clade may have evolved along with common substrates or other proteins in related pathways. Thus, the classification of NCL proteins based on their phylogenetic relationships with each other provides an alternative way to group the NCLs. This grouping does however appear to overlap somewhat with the types of AFSM associated to the loss of the NCL-protein. Thus, continued studies on how the NCL proteins and proteins with biological relevance to NCL-associated pathways evolved may provide clues about the function of NCL-proteins.

REFERENCES

[1] Banerjee, P., Dasgupta, A., Siakotos, A. N., Dawson, G. Evidence for lipase abnormality: high levels of free and triacylglycerol forms of unsaturated fatty acids in neuronal ceroid-lipofuscinosis tissue. *Am. J. Med. Genet.* **1992**, *42*: 549-54.

[2] Zeman, W., Dyken, P. Neuronal ceroid-lipofuscinosis (Batten's disease): relationship to amaurotic family idiocy? *Pediatrics* **1969**, *44*: 570-83.

[3] Weimer, J. M., Kriscenski-Perry, E., Elshatory, Y., Pearce, D. A. The neuronal ceroid lipofuscinoses: mutations in different proteins result in similar disease. *Neuromol Med.* **2002**, *1*: 111-24.

[4] Hofmann, S. L., Das, A. K., Lu, J. Y., Soyombo, A. A. Positional candidate gene cloning of CLN1. *Adv. Genet.* **2001**, *45*: 69-92.

[5] Santavuori, P., Haltia, M., Rapola, J., Raitta, C. Infantile type of so-called neuronal ceroid-lipofuscinosis. 1. A clinical study of 15 patients. *J. Neurol. Sci.* **1973**, *18*: 257-67.

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

[7] Wisniewski, K. E. Neuronal ceroid-lipofuscinoses. Gene Reviews at http://www.genetests.org. **2005**.

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

[9] Tyynela, J., Baumann, M., Henseler, M., Sandhoff, K., Haltia, M. Sphingolipid activator proteins (SAPs) are stored together with glycosphingolipids in the infantile neuronal ceroid-lipofuscinosis (INCL). *Am. J. Med. Genet.* **1995**, *57*: 294-7.

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

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

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

[13] Verkruyse, L. A., Hofmann, S. L. Lysosomal targeting of palmitoyl-protein thioesterase. *J. Biol. Chem.* **1996**, *271*: 15831-6.

[14] Lu, J. Y., Verkruyse, L. A., Hofmann, S. L. 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. USA* **1996**, *93*: 10046-50.

[15] van Diggelen, O. P., Keulemans, J. L., Winchester, B., Hofman, I. L., Vanhanen, S. L., Santavuori, P., Voznyi, Y. V. A rapid fluorogenic palmitoyl-protein thioesterase assay: pre- and postnatal diagnosis of INCL. *Mol. Genet. Metab.* **1999**, *66*: 240-4.

[16] Voznyi, Y. V., Keulemans, J. L., Mancini, G. M., Catsman-Berrevoets, C. E., Young, E., Winchester, B., Kleijer, W. J., van Diggelen, O. P. A new simple enzyme assay for pre- and postnatal diagnosis of infantile neuronal ceroid lipofuscinosis (INCL) and its variants. *J. Med. Genet.* **1999**, *36*: 471-4.

[17] van Diggelen, O. P., Thobois, S., Tilikete, C., Zabot, M. T., Keulemans, J. L., van Bunderen, P. A., Taschner, P. E., Losekoot, M., Voznyi, Y. V. Adult neuronal ceroid lipofuscinosis with palmitoyl-protein thioesterase deficiency: first adult-onset patients of a childhood disease. *Ann. Neurol.* **2001**, *50*: 269-72.

[18] de Vries, B. B., Kleijer, W. J., Keulemans, J. L., Voznyi, Y. V., Franken, P. F., Eurlings, M. C., Galjaard, R. J., Losekoot, M., Catsman-Berrevoets, C. E., Breuning, M. H., Taschner, P. E., van Diggelen, O. P. First-trimester diagnosis of infantile neuronal ceroid lipofuscinosis (INCL) using PPT enzyme assay and CLN1 mutation analysis. *Prenat. Diagn.* **1999**, *19*: 559-62.

[19] Kohan, R., de Halac, I. N., Tapia Anzolini, V., Cismondi, A., Oller Ramirez, A. M., Paschini Capra, A., de Kremer, R. D. Palmitoyl Protein Thioesterase1 (PPT1) and Tripeptidyl Peptidase-I (TPP-I) are expressed in the human saliva. A reliable and non-invasive source for the diagnosis of infantile (CLN1) and late infantile (CLN2) neuronal ceroid lipofuscinoses. *Clin. Biochem.* **2005**, *38*: 492-4.

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

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

[22] Lu, J. Y., Hofmann, S. L. Thematic review series: lipid posttranslational modifications. Lysosomal metabolism of lipid-modified proteins. *J. Lipid Res.* **2006**, *47*: 1352-7.

[23] Williams, R. E., Aberg, L., Autti, T., Goebel, H. H., Kohlschutter, A., Lonnqvist, T. Diagnosis of the neuronal ceroid lipofuscinoses: an update. *Biochim. Biophys. Acta* **2006**, *1762*: 865-72.

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

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

[26] Stephenson, J. B., Greene, N. D., Leung, K. Y., Munroe, P. B., Mole, S. E., Gardiner, R. M., Taschner, P. E., O'Regan, M., Naismith, K., Crow, Y. J., Mitchison, H. M. The molecular basis of GROD-storing neuronal ceroid lipofuscinoses in Scotland. *Mol. Genet. Metab.* **1999**, *66*: 245-7.

[27] Wisniewski, K. E., Kida, E., Connell, F., Zhong, N. Neuronal ceroid lipofuscinoses: research update. *Neurol. Sci.* **2000**, *21*: S49-56.

[28] Lyon, G., Raymond, A. D., Kolodny, E. H. *Neurology of hereditary metabolic diseases of children*, McGraw-Hill: New York **1996**.

[29] Goebel, H. H., Mole, S. E., Lake, B. D. *The neuronal ceroid lipofuscinoses (Batten disease)*. IOS: Washington, DC **1999**.

[30] Kida, E., Golabek, A. A., Wisniewski, K. E. Cellular pathology and pathogenic aspects of neuronal ceroid lipofuscinoses. *Adv. Genet.* **2001**, *45*: 35-68.

[31] Wisniewski, K. E., Kida, E., Golabek, A. A., Kaczmarski, W., Connell, F., Zhong, N. Neuronal ceroid lipofuscinoses: classification and diagnosis. *Adv. Genet.* **2001**, *45*: 1-34.

[32] Steinfeld, R., Steinke, H. B., Isbrandt, D., Kohlschutter, A., Gartner, J. Mutations in classical late infantile neuronal ceroid lipofuscinosis disrupt transport of tripeptidyl-peptidase I to lysosomes. *Hum. Mol. Genet.* **2004**, *13*: 2483-91.

[33] Donahue, S., Zeman, W., Watanabe, I. In *Inborn Disorders of Sphingolipid Metabolism*; Perfamon Press: Oxford, **1967**, pp. 3-22.

[34] Palmer, D. N., Fearnley, I. M., Walker, J. E., Hall, N. A., Lake, B. D., Wolfe, L. S., Haltia, M., Martinus, R. D., Jolly, R. D. Mitochondrial ATP synthase subunit c storage in the ceroid-lipofuscinoses (Batten disease). *Am. J. Med. Genet.* **1992**, *42*: 561-7.

[35] Kyttala, A., Lahtinen, U., Braulke, T., Hofmann, S. L. Functional biology of the neuronal ceroid lipofuscinoses (NCL) proteins. *Biochim. Biophys. Acta* **2006**, *1762*: 920-33.

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

[37] Lin, L., Sohar, I., Lackland, H., Lobel, P. The human CLN2 protein/tripeptidyl-peptidase I is a serine protease that autoactivates at acidic pH. *J. Biol. Chem.* **2001**, *276*: 2249-55.

[38] Rawlings, N. D., Barrett, A. J. Tripeptidyl-peptidase I is apparently the CLN2 protein absent in classical late-infantile neuronal ceroid lipofuscinosis. *Biochim. Biophys. Acta* **1999**, *1429*: 496-500.

[39] Vines, D. J., Warburton, M. J. Classical late infantile neuronal ceroid lipofuscinosis fibroblasts are deficient in lysosomal tripeptidyl peptidase I. *FEBS Lett.* **1999**, *443*: 131-5.

[40] Liu, C. G., Sleat, D. E., Donnelly, R. J., Lobel, P. Structural organization and sequence of CLN2, the defective gene in classical late infantile neuronal ceroid lipofuscinosis. *Genomics* **1998**, *50*: 206-12.

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

[42] Golabek, A. A., Kida, E., Walus, M., Wujek, P., Mehta, P., Wisniewski, K. E. Biosynthesis, glycosylation, and enzymatic processing *in vivo* of human tripeptidyl-peptidase I. *J. Biol. Chem.* **2003**, *278*: 7135-45.

[43] McDonald, J. K., Hoisington, A. R., Eisenhauer, D. A. Partial purification and characterization of an ovarian tripeptidyl peptidase: a lysosomal exopeptidase that sequentially releases collagen-related (Gly-Pro-X) triplets. *Biochem. Biophys. Res. Commun.* **1985**, *126*: 63-71.

[44] Page, A. E., Fuller, K., Chambers, T. J., Warburton, M. J. Purification and characterization of a tripeptidyl peptidase I from human osteoclastomas: evidence for its role in bone resorption. *Arch. Biochem. Biophys.* **1993**, *306*: 354-9.

[45] Vines, D., Warburton, M. J. Purification and characterisation of a tripeptidyl aminopeptidase I from rat spleen. *Biochim. Biophys. Acta* **1998**, *1384*: 233-42.

[46] Ezaki, J., Takeda-Ezaki, M., Oda, K., Kominami, E. Characterization of endopeptidase activity of tripeptidyl peptidase-I/CLN2 protein which is deficient in classical late infantile neuronal ceroid lipofuscinosis. *Biochem. Biophys. Res. Commun.* **2000**, *268*: 904-8.

[47] Mole, S. Gene table: neuronal ceroid lipofuscinoses. *Eur. J. Paediatr. Neurol.* **2002**, *6*: 129-30.

[48] Ezaki, J., Tanida, I., Kanehagi, N., Kominami, E. A lysosomal proteinase, the late infantile neuronal ceroid lipofuscinosis gene (CLN2) product, is essential for degradation of a hydrophobic protein, the subunit c of ATP synthase. *J. Neurochem.* **1999**, *72*: 2573-82.

[49] Oyama, H., Fujisawa, T., Suzuki, T., Dunn, B. M., Wlodawer, A., Oda, K. Catalytic residues and substrate specificity of recombinant human tripeptidyl peptidase I (CLN2). *J. Biochem.* **2005**, *138*: 127-34.

[50] Bukina, A. M., Tsvetkova, I. V., Semiachkina, A. N., Il'ina, E. S. [Tripeptidyl peptidase 1 deficiency in neuronal ceroid lipofuscinosis. A novel mutation]. *Vopr. Med. Khim.* **2002**, *48*: 594-8.

[51] Kononov, A. I., Bukina, M. N. Luminescence excitation spectra reveal low-lying excited states in stacked adenine bases. *J. Biomol. Struct. Dyn.* **2002**, *20*: 465-71.

[52] Bukina, I. E., Baev, A. A. [Baker's cysts in patients with gonarthrosis in early stages according to arthrosonography and magnetic resonance tomography]. *Klin. Med. (Mosk)* **2002**, *80*: 27-30.

[53] Kwon, J. M., Rothberg, P. G., Leman, A. R., Weimer, J. M., Mink, J. W., Pearce, D. A. Novel CLN3 mutation predicted to cause complete loss of protein function does not modify the classical JNCL phenotype. *Neurosci. Lett.* **2005**, *387*: 111-4.

[54] Marshall, F. J., de Blieck, E. A., Mink, J. W., Dure, L., Adams, H., Messing, S., Rothberg, P. G., Levy, E., McDonough, T., DeYoung, J., Wang, M., Ramirez-Montealegre, D., Kwon, J. M., Pearce, D. A. A clinical rating scale for Batten disease: reliable and relevant for clinical trials. *Neurology* **2005**, *65*: 275-9.

[55] Kominami, E., Ezaki, J., Wolfe, L. S. New insight into lysosomal protein storage disease: delayed catabolism of ATP synthase subunit c in Batten disease. *Neurochem. Res.* **1995**, *20*: 1305-9.

[56] Mitchison, H. M., Munroe, P. B., O'Rawe, A. M., Taschner, P. E., de Vos, N., Kremmidiotis, G., Lensink, I., Munk, A. C., D'Arigo, K. L., Anderson, J. W., Lerner, T. J., Moyzis, R. K., Callen, D. F., Breuning, M. H., Doggett, N. A., Gardiner, R. M., Mole, S. E. Genomic structure and complete nucleotide sequence of the Batten disease gene, CLN3. *Genomics* **1997**, *40*: 346-50.

[57] Mitchison, H. M., Taschner, P. E., Kremmidiotis, G., Callen, D. F., Doggett, N. A., Lerner, T. J., Janes, R. B., Wallace, B. A., Munroe, P. B., O'Rawe, A. M., Gardiner, R. M., Mole, S. E. Structure of the CLN3 gene and predicted structure, location and function of CLN3 protein. *Neuropediatrics* **1997**, *28*: 12-4.

[58] Mao, Q., Foster, B. J., Xia, H., Davidson, B. L. Membrane topology of CLN3, the protein underlying Batten disease. *FEBS Lett.* **2003**, *541*: 40-6.

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

[60] Kida, E., Kaczmarski, W., Golabek, A. A., Kaczmarski, A., Michalewski, M., Wisniewski, K. E. Analysis of intracellular distribution and trafficking of the CLN3 protein in fusion with the green fluorescent protein *in vitro*. *Mol. Genet. Metab.* **1999**, *66*: 265-71.

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

[62] Holopainen, J. M., Saarikoski, J., Kinnunen, P. K., Jarvela, I. Elevated lysosomal pH in neuronal ceroid lipofuscinoses (NCLs). *Eur. J. Biochem.* **2001**, *268*: 5851-6.

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

[64] Chattopadhyay, S., Muzaffar, N. E., Sherman, F., Pearce, D. A. The yeast model for batten disease: mutations in BTN1, BTN2, and HSP30 alter pH homeostasis. *J. Bacteriol.* **2000**, *182*: 6418-23.

[65] Das, A. M., von Harlem, R., Feist, M., Lucke, T., Kohlschutter, A. Altered levels of high-energy phosphate compounds in fibroblasts from different forms of neuronal ceroid lipofuscinoses: further evidence for mitochondrial involvement. *Eur. J. Paediatr. Neurol.* **2001**, *5 Suppl A*: 143-6.

[66] Puranam, K. L., Guo, W. X., Qian, W. H., Nikbakht, K., Boustany, R. M. CLN3 defines a novel antiapoptotic pathway operative in neurodegeneration and mediated by ceramide. *Mol. Genet. Metab.* **1999**, *66*: 294-308.

[67] Persaud-Sawin, D. A., VanDongen, A., Boustany, R. M. Motifs within the CLN3 protein: modulation of cell growth rates and apoptosis. *Hum. Mol. Genet.* **2002**, *11*: 2129-42.

[68] Chattopadhyay, S., Roberts, P. M., Pearce, D. A. The yeast model for Batten disease: a role for Btn2p in the trafficking of the Golgi-associated vesicular targeting protein, Yif1p. *Biochem. Biophys. Res. Commun.* **2003**, *302*: 534-8.

[69]   Haskell, R. E., Carr, C. J., Pearce, D. A., Bennett, M. J., Davidson, B. L. Batten disease: evaluation of CLN3 mutations on protein localization and function. *Hum. Mol. Genet.* **2000**, *9*: 735-44.

[70]   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., Jarvela, I., Taschner, P. E. New mutations in the neuronal ceroid lipofuscinosis genes. *Eur. J. Paediatr. Neurol.* **2001**, *5 Suppl A*: 7-10.

[71]   Ranta, S., Savukoski, M., Santavuori, P., Haltia, M. Studies of homogenous populations: CLN5 and CLN8. *Adv. Genet.* **2001**, *45*: 123-40.

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

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

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

[75]   Tyynela, J., Suopanki, J., Santavuori, P., Baumann, M., Haltia, M. Variant late infantile neuronal ceroid-lipofuscinosis: pathology and biochemistry. *J. Neuropathol. Exp. Neurol.* **1997**, *56*: 369-75.

[76]   Mole, S., Gardiner, M. Molecular genetics of the neuronal ceroid lipofuscinoses. *Epilepsia* **1999**, *40*(Suppl 3): 29-32.

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

[78]   Isosomppi, J., Vesa, J., Jalanko, A., Peltonen, L. Lysosomal localization of the neuronal ceroid lipofuscinosis CLN5 protein. *Hum. Mol. Genet.* **2002**, *11*: 885-91.

[79]   Vesa, J., Chin, M. H., Oelgeschlager, K., Isosomppi, J., DellAngelica, E. C., Jalanko, A., Peltonen, L. Neuronal ceroid lipofuscinoses are connected at molecular level: interaction of CLN5 protein with CLN2 and CLN3. *Mol. Biol. Cell* **2002**, *13*: 2410-20.

[80]   Auger, K. J., Ajene, A., Lerner, T. Progress toward the cloning of CLN6, the gene underlying a variant LINCL. *Mol. Genet. Metab.* **1999**, *66*: 332-6.

[81]   Jolly, R. D., Shimada, A., Dopfmer, I., Slack, P. M., Birtles, M. J., Palmer, D. N. Ceroid-lipofuscinosis (Batten's disease): pathogenesis and sequential neuropathological changes in the ovine model. *Neuropathol. Appl. Neurobiol.* **1989**, *15*: 371-83.

[82]   Haines, J. L., Boustany, R. M., Alroy, J., Auger, K. J., Shook, K. S., Terwedow, H., Lerner, T. J. Chromosomal localization of two genes underlying late-infantile neuronal ceroid lipofuscinosis. *Neurogenetics* **1998**, *1*: 217-22.

[83]   Wheeler, R. B., Sharp, J. D., Schultz, R. A., Joslin, J. M., Williams, R. E., Mole, S. E. 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.* **2002**, *70*: 537-42.

[84]   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., MacDonald, M. E. Mutations in a novel CLN6-encoded transmembrane protein cause variant neuronal ceroid lipofuscinosis in man and mouse. *Am. J. Hum. Genet.* **2002**, *70*: 324-35.

[85]   Heine, C., Koch, B., Storch, S., Kohlschutter, A., Palmer, D. N., Braulke, T. Defective endoplasmic reticulum-resident membrane protein CLN6 affects lysosomal degradation of endocytosed aryl-sulfatase A. *J. Biol. Chem.* **2004**, *279*: 22347-52.

[86]   Mole, S. E., Michaux, G., Codlin, S., Wheeler, R. B., Sharp, J. D., Cutler, D. F. CLN6, which is associated with a lysosomal storage disease, is an endoplasmic reticulum protein. *Exp. Cell Res.* **2004**, *298*: 399-406.

[87]   Sharp, J. D., Wheeler, R. B., Parker, K. A., Gardiner, R. M., Williams, R. E., Mole, S. E. Spectrum of CLN6 mutations in variant late infantile neuronal ceroid lipofuscinosis. *Hum. Mutat.* **2003**, *22*: 35-42.

[88]   Topcu, M., Tan, H., Yalnizoglu, D., Usubutun, A., Saatci, I., Aynaci, M., Anlar, B., Topaloglu, H., Turanli, G., Kose, G., Aysun, S. Evaluation of 36 patients from Turkey with neuronal ceroid lipofuscinosis: clinical, neurophysiological, neuroradiological and histopathologic studies. *Turk J. Pediatr.* **2004**, *46*: 1-10.

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

[90]   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., Mole, S. E. Turkish variant late infantile neuronal ceroid lipofuscinosis (CLN7) may be allelic to CLN8. *Eur. J. Paediatr. Neurol.* **2001**, *5 Suppl A*: 21-7.

[91]   Ranta, S., Topcu, M., Tegelberg, S., Tan, H., Ustubutun, A., Saatci, I., Dufke, A., Enders, H., Pohl, K., Alembik, Y., Mitchell, W. A., Mole, S. E., Lehesjoki, A. E. Variant late infantile neuronal ceroid lipofuscinosis in a subset of Turkish patients is allelic to Northern epilepsy. *Hum. Mutat.* **2004**, *23*: 300-5.

[92]   Siintola, E., Topcu, M., Kohlschutter, A., Salonen, T., Joensuu, T., Anttonen, A. K., Lehesjoki, A. E. Two novel CLN6 mutations in variant late-infantile neuronal ceroid lipofuscinosis patients of Turkish origin. *Clin. Genet.* **2005**, *68*: 167-73.

[93]   Siintola, E., Topcu, M., Aula, N., Lohi, H., Minassian, B. A., Paterson, A. D., Liu, X. Q., Wilson, C., Lahtinen, U., Anttonen, A. K., Lehesjoki, A. E. The novel neuronal ceroid lipofuscinosis gene MFSD8 encodes a putative lysosomal transporter. *Am. J. Hum. Genet.* **2007**, *81*: 136-46.

[94]   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., Lehesjoki, A. E. The neuronal ceroid lipofuscinoses in human EPMR and mnd mutant mice are associated with mutations in CLN8. *Nat. Genet.* **1999**, *23*: 233-6.

[95]   Lonka, L., Aalto, A., Kopra, O., Kuronen, M., Kokaia, Z., Saarma, M., Lehesjoki, A. E. The neuronal ceroid lipofuscinosis Cln8 gene expression is developmentally regulated in mouse brain and up-regulated in the hippocampal kindling model of epilepsy. *BMC Neurosci.* **2005**, *6*: 27.

[96]   Winter, E., Ponting, C. P. TRAM, LAG1 and CLN8: members of a novel family of lipid-sensing domains? *Trends Biochem. Sci.* **2002**, *27*: 381-3.

[97]   Hegde, R. S., Voigt, S., Rapoport, T. A., Lingappa, V. R. TRAM regulates the exposure of nascent secretory proteins to the cytosol during translocation into the endoplasmic reticulum. *Cell* **1998**, *92*: 621-31.

[98]   Barz, W. P., Walter, P. Two endoplasmic reticulum (ER) membrane proteins that facilitate ER-to-Golgi transport of glycosyl-phosphatidylinositol-anchored proteins. *Mol. Biol. Cell* **1999**, *10*: 1043-59.

[99]   Heinrich, S. U., Mothes, W., Brunner, J., Rapoport, T. A. The Sec61p complex mediates the integration of a membrane protein by allowing lipid partitioning of the transmembrane domain. *Cell* **2000**, *102*: 233-44.

[100] Norman, R., Wood, N. Congenital form of amaurotic family idiocy. *J. Neur. Psychiat.* **1941**, *4*: 175-90.

[101] Brown, N. J., Corner, B. D., Dodgson, M. C. A second case in the same family of congenital familial cerebral lipoidosis resembling amaurotic family idiocy. *Arch. Dis. Child* **1954**, *29*: 48-54.

[102] Sandbank, U. Congenital amaurotic idiocy. *Pathol. Eur.* **1968**, *3*: 226-9.

[103] Humphreys, S., Lake, B. D., Scholtz, C. L. Congenital amaurotic idiocy--a pathological, histochemical, biochemical and ultrastructural study. *Neuropathol. Appl. Neurobiol.* **1985**, *11*: 475-84.

[104] Barohn, R. J., Dowd, D. C., Kagan-Hallet, K. S. Congenital ceroid-lipofuscinosis. *Pediatr. Neurol.* **1992**, *8*: 54-9.

[105] Siintola, E., Partanen, S., Stromme, P., Haapanen, A., Haltia, M., Maehlen, J., Lehesjoki, A. E., Tyynela, J. Cathepsin D deficiency underlies congenital human neuronal ceroid-lipofuscinosis. *Brain* **2006**, *129*: 1438-45.

[106] Qin, S., Nakai, H., Byters, M. G., Eddy, R. L., Haley, L. L., Henry, W. M., Wang, X., Watkins, P. C., Chirgwin, J. M., Shows, T. B. Mapping FSHB, CAT, and CTSD to specific sites on 11p. (Abstract). *Cytogenet. Cell Genet.* **1987**, *46*: 678 only.

[107] Henry, I., Puech, A., Antignac, C., Couillin, P., Jeanpierre, M., Ahnine, L., Barichard, F., Boehm, T., Augereau, P., Scrable, H., Rabbitts, T. H., Rochefort, H., Cavenee, W., Junien, C. Subregional mapping of BWS, CTSD, MYOD1, and a T-ALL breakpoint in 11p15. (Abstract). *Cytogenet. Cell Genet.* **1989**, *51*: 1013 only.

[108] Hasilik, A., von Figura, K., Grzeschik, K.-H. In *Cytogenet. Cell Genet.*, **1982**; Vol. *32*, pp. 284 only.

[109]   Rawlings, N. D., Barrett, A. J. Families of aspartic peptidases, and those of unknown catalytic mechanism. *Methods Enzymol.* **1995**, *248*: 105-20.

[110]   Press, E. M., Porter, R. R., Cebra, J. The isolation and properties of a proteolytic enzyme, cathepsin D, from bovine spleen. *Biochem. J.* **1960**, *74*: 501-14.

[111]   Gulnik, S., Baldwin, E. T., Tarasova, N., Erickson, J. Human liver cathepsin D. Purification, crystallization and preliminary X-ray diffraction analysis of a lysosomal enzyme. *J. Mol. Biol.* **1992**, *227*: 265-70.

[112]   Metcalf, P., Fusek, M. Two crystal structures for cathepsin D: the lysosomal targeting signal and active site. *EMBO J.* **1993**, *12*: 1293-302.

[113]   Faust, P. L., Kornfeld, S., Chirgwin, J. M. Cloning and sequence analysis of cDNA for human cathepsin D. *Proc. Natl. Acad. Sci. USA* **1985**, *82*: 4910-4.

[114]   Lin, X. L., Wong, R. N., Tang, J. Synthesis, purification, and active site mutagenesis of recombinant porcine pepsinogen. *J. Biol. Chem.* **1989**, *264*: 4482-9.

[115]   Wittlin, S., Rosel, J., Hofmann, F., Stover, D. R. Mechanisms and kinetics of procathepsin D activation. *Eur. J. Biochem.* **1999**, *265*: 384-93.

[116]   Myllykangas, L., Tyynela, J., Page-McCaw, A., Rubin, G. M., Haltia, M. J., Feany, M. B. Cathepsin D-deficient Drosophila recapitulate the key features of neuronal ceroid lipofuscinoses. *Neurobiol. Dis.* **2005**, *19*: 194-9.

[117]   Tyynela, J., Sohar, I., Sleat, D. E., Gin, R. M., Donnelly, R. J., Baumann, M., Haltia, M., Lobel, P. A mutation in the ovine cathepsin D gene causes a congenital lysosomal storage disease with profound neurodegeneration. *EMBO J.* **2000**, *19*: 2786-92.

[118]   Koike, M., Nakanishi, H., Saftig, P., Ezaki, J., Isahara, K., Ohsawa, Y., Schulz-Schaeffer, W., Watanabe, T., Waguri, S., Kametaka, S., Shibata, M., Yamamoto, K., Kominami, E., Peters, C., von Figura, K., Uchiyama, Y. Cathepsin D deficiency induces lysosomal storage with ceroid lipofuscin in mouse CNS neurons. *J. Neurosci.* **2000**, *20*: 6898-906.

[119]   Steinfeld, R., Reinhardt, K., Schreiber, K., Hillebrand, M., Kraetzner, R., Bruck, W., Saftig, P., Gartner, J. Cathepsin D deficiency is associated with a human neurodegenerative disorder. *Am. J. Hum. Genet.* **2006**, *78*: 988-98.

[120]   Koonin, E. V., Tatusov, R. L., Galperin, M. Y. Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **1998**, *8*: 355-63.

[121]   Koonin, E. V., Tatusov, R. L., Rudd, K. E. Protein sequence comparison at genome scale. *Methods Enzymol.* **1996**, *266*: 295-322.