Biochimica et Biophysica Acta 1832 (2013) 1831–1841

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbadis

Review

Bioinformatic perspectives in the neuronal ceroid lipofuscinoses[☆]

Stanislav Kmoch [a], Viktor Stránecký [a], Richard D. Emes [b], Hannah M. Mitchison [c,*]

[a] Institute for Inherited Metabolic Disorders, First Faculty of Medicine, Charles University in Prague, 120 00 Prague, Czech Republic
[b] School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, Leicestershire, UK
[c] Molecular Medicine Unit and Birth Defects Research Centre, Institute of Child Health, University College London, London, UK

ARTICLE INFO

Article history:
Received 7 November 2012
Received in revised form 16 December 2012
Accepted 19 December 2012
Available online 26 December 2012

Keywords:
Neuronal ceroid lipofuscinosis
Bioinformatics
Linkage analysis
Next generation sequencing
Topology

ABSTRACT

The neuronal ceroid lipofuscinoses (NCLs) are a group of rare genetic diseases characterised clinically by the progressive deterioration of mental, motor and visual functions and histopathologically by the intracellular accumulation of autofluorescent lipopigment – ceroid – in affected tissues. The NCLs are clinically and genetically heterogeneous and more than 14 genetically distinct NCL subtypes have been described to date (*CLN1–CLN14*) (Haltia and Goebel, 2012 [1]). In this review we will chronologically summarise work which has led over the years to identification of NCL genes, and outline the potential of novel genomic techniques and related bioinformatic approaches for further genetic dissection and diagnosis of NCLs. This article is part of a Special Issue entitled: The Neuronal Ceroid Lipofuscinoses or Batten Disease.

1. Introduction

The neuronal ceroid lipofuscinoses are a diverse group of inherited childhood lysosomal storage disorders, now known to represent at least 14 genetically distinct neurodegenerative diseases. The NCLs were relatively undefined until the advent of molecular genetic research, in contrast to the well-stratifed disease subtypes that are now recognised [1]. Computer-based and bioinformatic analyses have contributed to improved understanding of the NCLs, from linkage studies to predicting disease protein topology and mutation pathogenicity. This review will summarise the bioinformatic advances that have occurred in the NCL field over recent years, and look towards future contributions.

2. The '90s — an era of positional cloning

The NCLs until the early 1990s were a genetically uncharacterised grouping of clinical entities not amenable to functional or candidate gene cloning [2]. Genetic dissection of NCLs was initially facilitated by the discovery of polymorphic protein markers, restriction fragment length polymorphisms (RFLPs) and the concept of disease gene mapping through linkage analysis [3]. This approach later became more efficient with discovery of multi-allelic short tandem repeat polymorphisms (STRPs) [4], and subsequent developments in the availability of efficient PCR-based genotyping technologies, the construction of human genetic maps [5–7], development of computer programs for linkage analysis [8], construction of physical maps [9], the general evolution of positional cloning strategies [10] and most importantly through the availability of publically accessible human DNA sequence data from Human Genome Project [11].

This instrumental framework and the gradual build-up of impressive cohorts of patients and families opened up an avenue for deciphering the genetic and molecular basis of NCLs. The era of NCL genetics started with identification of linkage between juvenile NCL or Batten disease (now *CLN3*) and the polymorphic protein marker haptoglobin on chromosome 16q22 [12]. Further studies using RFLP markers confirmed linkage to chromosome 16 [13] and this was gradually refined, using STRP markers, leading to the genetic and physical localization of *CLN3* to chromosome 16p12 [14,15] which indicated a major founder effect underlying the disease by demonstration of a common haplotype of genetic markers on most of the patients' *CLN3* chromosomes [16,17]. Additional fine genetic mapping [18,19] and construction of physical maps spanning the *CLN3* locus [20,21] provided the framework for the final identification of *CLN3* gene encoding the 'battenin' protein, and delineation of the major founder effect mutation (1 kb deletion) [22]. In parallel to these developments, four major clinical subtypes of NCL had been recognised through clinical and genetic advances, and discovery of the *CLN3* locus on chromosome 16 allowed exclusion of this region in *CLN1* [23], *CLN2* [15,24] and *CLN5* [25].

[☆] This article is part of a Special Issue entitled: The Neuronal Ceroid Lipofuscinoses or Batten Disease.

[*] Corresponding author at: Molecular Medicine Unit and Birth Defects Research Centre, UCL Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK. Tel.: +44 207 905 2866; fax: +44 207 404 6191.
E-mail address: h.mitchison@ucl.ac.uk (H.M. Mitchison).

0925-4439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbadis.2012.12.010

CLN1 was initially mapped to chromosome 1p [26]. Additional fine mapping to chromosome 1p32 [27] and construction of the *CLN1* physical map [28] led to identification of disease causing mutations in the *PPT1* gene encoding palmitoyl-protein thioesterase 1 [29]. *CLN5* was initially mapped to chromosome 13q21.1–q32 [30]. The candidate region was further narrowed using linkage disequilibrium analysis [31] and physically localised using the fiber-FISH technique [32]. The *CLN5* gene was later identified using a broad spectrum of positional cloning strategies in 1998 [33], although its function in controlling the itinerary of lysosomal sorting receptors has been proposed only recently [34]. *CLN2* was initially excluded from the *CLN3*, *CLN1* and *CLN5* loci [35] and then localised, using homozygosity mapping, to chromosome 11p15 [36]. Identification of the CLN2 protein (tripeptidyl-peptidase 1, TPP1) and its gene was achieved not by positional cloning but by biochemical methods, using comparative proteomic analysis of mannose-6-phosphate-containing proteins in brain extracts from patients versus controls, with subsequent bioinformatic searching for cDNA clones encoding the N-terminal sequence of the corresponding protein found to be absent in patients' samples [37].

The *CLN6* gene was initially excluded from the *CLN3* locus on chromosome 16, and from *CLN5* on chromosome 13, and was then localised, using homozygosity mapping, to chromosome 15q [36]. Additional fine mapping together with construction of physical map and transcription maps [38,39] and in silico cloning approach [40] led to a definition of a set of positional candidate genes. Application of a systematic expression analysis and mutation screening finally revealed recurrent mutations in an uncharacterised protein FLJ20561 [41,42], encoding an endoplasmic reticulum membrane protein [43,44]. *CLN8* was originally assigned, based on neuropathological studies, to Northern epilepsy [45], when the gene was first mapped to chromosome 8p [46,47]. *CLN8* was later identified by positional cloning [48], and also found to encode an endoplasmic reticulum membrane protein [49]. Linkage analysis and homozygosity mapping in Turkish CLN7 patients suggested that in a subgroup of these patients the disease may be allelic to Northern epilepsy [50,51].

3. The era of genomics

The final phase in cloning of these initial NCL genes in the '00s including *CLN6* and *CLN8* demonstrated the enormous potential for gene identification using the publically accessible data from the Human Genome Project that developed over the nineties, alongside the increased capacity of DNA sequencing that was arising from associated technological developments. The positional cloning approach was also greatly facilitated by other evolving resources that included gene and tissue expression arrays [52], biological chips [53], and SNP genotyping arrays [54]. Furthermore, there was parallel expansion in discovery of SNP polymorphisms and establishment of the dbSNP database [55], construction of SNP-based genetic maps [56], combined linkage-physical maps [57]. In addition, methods were improved by the availability of computer programs capable of handling tens of thousands of genotypes for linkage analysis and haplotyping [58], and the development of integrated database resources and web-based genome browsers such Map Viewer [59], ENSEMBL [60] and the UCSC genome browser [61] that provided an opportunity for search, retrieval and analysis of genomic sequences and annotation data.

The discovery of *CLN7* and definition of *CLN9* further demonstrated the power of these technological and conceptual advances, including use of molecular arrays of genes and genetic markers. The CLN7 NCL subtype was firstly found to be a genetically heterogeneous disease subtype, since mutations in the *CLN8* and *CLN6* genes were identified in subsets of patients with similar features that led to them originally being classed together as ‘Turkish vLINCL’ [62]. Then the *CLN7* gene was mapped in families that did not carry *CLN6* or *CLN8* mutations, using genome-wide genotyping with Affymetrix Human GeneChip arrays, to a ~20 Mb genomic region on chromosome 4q28.1–q28.2. Positional and functional relevant candidate genes were then screened for mutations by genomic sequencing. This analysis revealed recurrent mutations in *MFSD8* encoding a lysosomal transmembrane protein [63]. The CLN9 disease subtype was excluded from other CLN forms after negative mutation analysis of all the known *CLN* genes, and due to the distinctive phenotype and gene expression profile of CLN9 deficient patient derived fibroblasts [64].

CLN10 was identified through recognising a phenotypic similarity between the phenotype of affected patients and that of cathepsin D (Ctsd)-deficient mice [65] and sheep [66]. Subsequent targeted sequence analysis of the human *CTSD* gene in patients with non-identified genetic causes of NCL revealed pathogenic inherited mutations [67,68]. The CLN4 terminology was designated early on for adult-onset NCL disease (Kufs) but underlying genetic heterogeneity was evident for this disease entity, since both autosomal recessive (form A) and autosomal dominant (form B, Parry-type) modes of inheritance were found among the patients and families. What is now known as the CLN4A category of disease, which is characterised by autosomal recessive progressive myoclonus epilepsy, was identified through linkage mapping and candidate gene sequencing which revealed in fact that most of the individuals carried mutations in *CLN6* [69].

4. Post-cloning bioinformatic approaches

The successful identification of the *CLN1*, *CLN2*, *CLN3*, *CLN5*, *CLN6*, *CLN7*, *CLN8* and *CLN10* genes and developments in routine DNA sequencing dramatically changed what we understand about NCL biology over the last 2 decades, and moreover right from the beginning greatly impacted in improving diagnostic procedures [70]. These advances have led to identification of more than 360 NCL-causing mutations and have revealed both phenotypic divergence and phenotypic convergence of individual genetic defects [71]. Huge biological and clinical advances have been made as a result, in dissecting correlations between the underlying genotype and resulting clinical phenotype for the NCL forms [71].

The interpretation of genotypes is made within the context of what is known about the encoded disease proteins' functional domains and putative biological roles. Thus, significant value has come from bioinformatic computational studies of NCL protein tertiary topology structure, derived from structure and function information. This information has been generated from available protein domain homologies, protein resolution structures, and comparative protein domain evolutionary conservation data. Of the first NCL genes to be isolated, bioinformatics searches on the CLN1/PPT1, CLN2/TPP1 and CLN10/Cathepsin D proteins identified them as lysosomal hydrolase enzymes expressed as proenzymes requiring cleavage of their signal peptide to mature to their enzymatic forms. All three were recognised entities prior to being connected to NCL disease, and crystal structure work has allowed excellent understanding of the effect of NCL-associated mutations in these proteins [72–74]. The CLN5 protein has been more controversial since it has no recognisable domains and exists as multiple protein isoforms not all of which have a defined signal peptide. However it is now widely recognised to be a soluble lysosomal glycoprotein [75], rather than as was initially reported being a transmembrane protein [33].

The other novel NCL proteins CLN3, CLN6, CLN7 and CLN8 were all identified to encode transmembrane proteins using various modelling prediction programs such as TMHMM to predict hydrophobic transmembrane (TM) helices [76]. CLN3 and CLN6 have no similarity to other known proteins and are typically modelled as simple multiple pass proteins localised to the endosome–lysosome system vesicular membranes or the endoplasmic reticulum membrane, respectively (Fig. 1). To determine their toplogy they have been subject to many complementary experimental studies using antisera and tags, and the consensus is that CLN3 is a six TM protein with cytoplasmic N- and