Folkhälsan Institute of Genetics,

Neuroscience Center

and

Department of Medical Genetics,

University of Helsinki,

Finland

Molecular genetics of Cohen syndrome

Juha Kolehmainen

Academic Dissertation

To be publicly discussed with the permission of the Faculty of Medicine, University of Helsinki, in auditorium 2, Biomedicum Helsinki, on December 10th 2004, at 12 noon

Helsinki 2004

Supervised by:

Anna-Elina Lehesjoki MD, PhD
Professor and Research Director, Folkhälsan Institute of Genetics and
Neuroscience Center, University of Helsinki
Helsinki, Finland

Albert de la Chapelle, MD, PhD
Professor, Human Cancer Genetics Program,
Ohio State University,
Columbus, Ohio, U.S.A.

Reviewed by:

Marjo Kestilä PhD, Docent
Department of Molecular Medicine,
National Public Health Institute,
Helsinki, Finland

Pentti Tienari MD, PhD, Docent
Department of Neurology,
Helsinki University Central Hospital,
University of Helsinki,
Biomedicum Helsinki, Finland

Official opponent:

Han G. Brunner MD, PhD
Professor, Department of Human Genetics,
Radboud University Nijmegen Medical Centre,
Nijmegen, The Netherlands

ISBN 951-9170-91-X (paperback)
ISBN 952-10-2225-6 (PDF)
http://ethesis.helsinki.fi
Yliopistopaino
Helsinki 2004

To Kata

LIST OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The thesis is based on the following original articles, referred to in the text by the Roman numerals **I – IV**. Some additional unpublished data are presented.

I Kolehmainen J., Norio R., Kivitie-Kallio S., Tahvanainen E., de la Chapelle A., Lehesjoki A.E. (1997). Refined mapping of the Cohen syndrome gene by linkage disequilibrium. Eur. J. Hum. Genet. 5, 206-213.

II Kolehmainen J., Black G.C.M., Saarinen A., Chandler K., Clayton-Smith J., Träskelin A.L., Perveen R., Kivitie-Kallio S., Norio R., Warburg M., Fryns J-P., de la Chapelle A., Lehesjoki A.E. (2003). Cohen syndrome is caused by mutations in a novel gene, *COH1*, encoding a transmembrane protein with a presumed role in vesicle-mediated sorting and intracellular protein transport. Am. J. Hum. Genet. 72, 1359-1369.

III Falk M.J., Feiler H.S., Neilson D.E., Maxwell K., Lee J.V., Segall S.K., Robin N.H., Wilhelmsen K.C., Träskelin A.L., Kolehmainen J., Lehesjoki A.E., Wiznitzer M., Warman M.L. (2004). Cohen Syndrome in the Ohio Amish. Am. J. Med. Genet. 128A, 23-28.

IV Kolehmainen J*., Wilkinson R*., Lehesjoki A.E., Chandler K., Kivitie-Kallio S., Clayton-Smith J., Träskelin A.L., Waris L., Saarinen A., Khan J., Gross-Tsur V., Traboulsi E.I, Warburg M., Fryns J-P., Norio R., Black G.C.M., Manson F.D.C. (2004). Delineation of Cohen syndrome following a large-scale genotype-phenotype screen. Am. J. Hum. Genet. 75, 122-127.

*equal contribution

ABBREVIATIONS

AGU	aspartylglucosaminuria
ALMS	Alström syndrome
ALMS1	gene for Alström syndrome
APECED	autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy
AP3	adaptor-related protein complex 3
AP3B1, AP3B1	gene for adaptor-related protein complex β subunit, protein encoded by *AP3B1*
AS	Angelman syndrome
BAC	bacterial artificial chromosome
bp	base pair
BLAST	basic local alignment search tool
blastx	translated query homology search against protein database
BBS	Bardet-Biedl syndrome
cDNA	complementary deoxyribonucleic acid
CEPH	Centre dÉtudes du Polymorphisme Humain
chorein	protein for choreoacanthocytosis
CHAC	gene for choreoacanthocytosis
cM	centiMorgan (unit for one recombinational event in 100 meioses)
COH1, COH1	gene for Cohen syndrome, protein encoded by *COH1*
CNS	central nervous system
COP1	coatomer
COX6C	cytochrome c oxidase subunit VIc gene
CpG	dinucleotides CG linked by phosphate (p)
cR	centiRay
db	database
DGGE	denaturing gradient gel electrophoresis
DHPLC	denaturing high-performance liquid chromatography
DNA	deoxyribonucleic acid
DORFIN	gene for human double ring finger protein
EBI	European Bioinformatics Institute
ELA2	gene for elastase 2
ELK1	member of ets oncogene family
EMBL	European Molecular Biology Laboratory
ELN	elastin gene
ER	endoplasmic reticulum
EST	expressed sequence tag

ETS	*ETS* oncogene
ETS1P54	member of ets protein family
FASTA	fast sequence comparison algorithm
GC-rich	guanosine cytosine rich
GOA	gene ontology annotation
HPS2	Hermansky Pudlak syndrome type 2
IQ	intelligence quotient
kb	kilobase (unit for 1000 nucleotides)
LCR	ligation chain reaction
LD	linkage disequilibrium
LIMK1	gene for LIM domain kinase 1
lod score	logarithm of odds value
Mb	megabase pairs
MRI	magnetic resonance imagining
mRNA	messenger ribonucleic acid
MTN	multible tissue northern blot
NCBI	The National Center for Biotechnology Information
NIH	National Institutes of Health
NMD	nonsense-mediated mRNA decay
NRF2	nuclear respiratory factor 2 protein
OLA	oligonucleotide ligation assay
OMIM	Online Mendelian Inheritance in Man
OSR2	odd-skipped-related 2A gene
PAC	P1-derived artificial chromosome
PCR	polymerase chain reaction
POLR2K	polymerase (RNA) II (DNA directed) polypeptide K gene
PSI-BLAST	position-specific iterated BLAST
PTS2	peroxisomal targeting signal 2
PWS	Prader-Willi syndrome
q	long arm of chromosome
RFC2	gene for replication factor c, subunit 2
RH	radiation hybrid
RNA	ribonucleic acid
RP	retinitis pigmentosa
RT-PCR	reverse transcriptase polymerase chain reaction
SCOP	structural classification of proteins
SNP	single nucleotide polymorphism

SPAG1	gene for human sperm associated antigen 1
SSCP	single-stranded conformational polymorphism
SSRD	simple sequence repeats database
Start-p value	predicted probability for the CpG Island to locate over the transcription start site
STK3	gene for serine/threonine kinase 3
STS	sequence-tagged site
tblastn	protein query homology search against translated database
TF	transcription factor
3D-PSSM	three-dimensional position-specific scoring matrix
TIGR	The Institute for Genomic Research
TM	transmembrane
UCSC	University of California Santa Cruz
UTR	untranslated region
VEP	visual evoked potential
VNTR	variable number of tandem repeats
Vps	vacuolar protein sorting associated protein
Vps13, Vps13	gene for *S. cerevisiae* vacuolar protein sorting associated protein 13 (yeast homolog for human *COH1* gene), protein encoded by *Vps13*
VPS13C, VPS13D	human proteins belonging to VPS13 family
WBS	Williams-Beuren syndrome
YAC	yeast artificial chromosome

MEDICAL TERM GLOSSARY

Acanthocytosis a disorder characterized by abnormal red blood cells with multiple thorny projections or spicules

Alexithymia inability to identify own and others feelings and thus inability to communicate about them

Ataxia incoordination and unsteadiness due to the brain's failure to regulate the body's posture and regulate the strength and direction of limb movements

Cataract disease causing opacity in eye lens

Chorea ceaseless rapid complex body movements that look well coordinated and purposeful but are, in fact, involuntary

Chorioretinal dystrophy degeneration of choroideal and retinal layers that line the back of the eye

Choroidea vascular layer underlying retina that lines the back of the eye

Congenital malformation a physical defect in a newborn not defined to be either genetic or non-genetic by origin

Corpus callosum the area of the brain which connects two large brain halves

Craniofacial related to skull and face

Cyclic neutropenia cyclic low number of neutrophils varying in severity week to week, month to month, and possibly follows biorhythms

Dysmorphic feature a body characteristic that is abnormally formed

Granulocyte a type of white blood cell filled with microscopic granules

Granulocytopenia decrease in the number of granulocytes below normal values

Heterogeneous disorder inherited disorder that has variable inheritance pattern or can be caused by several genes

Hypogenitalism underdevelopment of the gonads

Hypotonia decreased tone of skeletal muscles

Intermittent neutropenia occasionally occurring low number of neutrophils

Joint laxity hyperextensibility of the joint

Kyphosis outward curvature of the spine, causing a humped back

Leukopenia decrease of the number of white blood cells below normal values

Lymphocytosis increase above normal values of lymphocytes

Mandible the bone of the lower jaw

Mental retardation limitations in mental functioning and in skills such as communicating, taking care of oneself, and social skills

Mental deficiency synonym for mental retardation

Microcephaly head circumference that is more than 2 standard deviations below the normal mean for age, sex, race, and gestation

Myopia nearsightedness, the ability to see close objects more clearly than distant objects

Neutrophil a subtype of white blood cell (specifically a form of granulocyte) filled with neutrally staining granules

Neutropenia decrease of the number of neutrophils below normal values

Nystagmus rapid rhythmic repetitious involuntary eye movements

Phenotype the appearance of an individual, which results from the interaction of the person's genetic makeup and his or her environment

Pigmentary retinopathy disease that causes accumulation of the pigment granules in retina

Philtrum the area from below the nose to the upper lip

Polydactyli increased number of digits

Pulmonary arterial stenosis narrowing of the pulmonary artery above pulmonic valve, which impedes the flow of blood from the right ventricle into the lungs

Retina light-sensitive nerve layer that lines the back of the eye

Retinitis pigmentosa any one of a large group of inherited disorders in which there are abnormalities of the photoreceptors (the rods and cones) in the retina, which leads to progressive visual loss

Retinochoroidal dystrophy synonym for chorioretinal dystrophy

Retinopathy any disease of the retina

Strabismus a condition in which the visual axes of the eyes are not parallel and the eyes appear to be looking in different directions

Supravalvular aortic stenosis narrowing of the aorta above aortic valve, which impedes the flow of blood from the left ventricle into the aorta and the arteries of the body

Synophrys eyebrows meet at midline

Tapering fingers narrow fingers

Triallelic inheritance inherited disorder in which mutations in three genes determine phenotype

The modifications for definitions at URL: http://www.medterms.com/script/main/hp.asp were used as a basis in the creation of this glossary for medical terms.

ABSTRACT

Cohen syndrome is an autosomal recessively inherited disorder with a broad spectrum of disease manifestations. Essential features for Cohen syndrome diagnosis include non-progressive psychomotor retardation, motor clumsiness and microcephaly, typical facial features, childhood hypotonia and hyperextensibility of the joints, ophthalmologic findings of retinochoroidal dystrophy and myopia in patients over five years of age, and granulocytopenia. As a result of published cases with a wide variety of clinical manifestations, a vivid debate over the diagnostic criteria of Cohen syndrome has been ongoing. Cohen syndrome is one of the diseases of the `Finnish disease heritage´. The incidence of Cohen syndrome is higher in the Finnish population—thirty-four patients with Cohen syndrome have been diagnosed in Finland, and over 100 Cohen syndrome case reports have been published worldwide. The mutation causing Cohen syndrome has been enriched in Finland, due to a demographic expansion of the Finnish population followed by restrictions of gene flow in genetic isolates, founder effects, genetic bottlenecks, and chance (genetic drift).

The main objectives of this study were to identify the gene underlying Cohen syndrome by a positional cloning approach, and to determine Cohen syndrome-associated mutations. Identification of the gene defect underlying Cohen syndrome further allowed determination of phenotype-genotype correlations and the definition of diagnostic criteria. Moreover, it laid the basis for *in silico*-based COH1 protein characterization. The present study was based on the assignment of the *COH1* gene to a 10 cM interval on chromosome 8q22.2-q22.3 by linkage analysis. The observation of linkage disequilibrium and conserved haplotypes in 75% of Finnish Cohen syndrome chromosomes allowed us to pinpoint the localization of the *COH1* gene, and limited the number of positional candidate genes subjected to mutation analysis. In a novel transcript, identified and assembled from the critical region, a two base pair deletion was identified in Finnish Cohen syndrome patients bearing the founder haplotype. Mutation analysis in Cohen syndrome patients revealed 31 additional *COH1* mutations. Lack of mutations in “Cohen-like” patients, in which the clinical features did not fulfill previously established diagnostic criteria, allowed molecular distinction between “true” Cohen syndrome and “Cohen-like” syndromes.

The full-length 14,093 bp *COH1* transcript was identified and assembled by *in silico*-based methods, and was verified by reverse transcriptase PCR (RT-PCR). The *COH1* gene is composed of at least 62 exons over ~864 kb of genomic DNA. Several alternatively spliced forms of *COH1* were observed. The 14,093 bp transcript is predicted to encode a 4,022 amino acid protein based on modelling with predicted transmembrane and other domains. Protein alignment against a domain family database indicated amino acid similarity with the *S. cerevisiae* Vps13 protein. This predicts that the COH1 protein has a function in the control of protein sorting.

The results presented in this thesis allow molecular confirmation of the clinical diagnosis of Cohen syndrome and confirm the previously established diagnostic criteria. Moreover, the results show that Cohen and "Cohen-like" syndromes are clinically and genetically distinct disorders. This work is the basis for further characterization of the COH1 protein and the molecular pathogenesis of Cohen syndrome.

INTRODUCTION

The human genome project began in 1990 with the aim to determine the entire 3,000 Mb human genome sequence. During this process the genome database information has grown exponentially, and the data submitted by the academic project has been freely available to the research community (Lander et al., 2001). Parallel to the academic genomic sequencing project, expressed sequence tagged (EST) databases, largely contributed by the commercial sequencing project of Celera (Venter et al., 2001), have evolved rapidly, and today contain over five million entries for sequence tagged sites (STSs) for human genes and 20 million sequences overall (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). This information has been utilized in compiling the 15,628 human full-length cDNAs reported in March, 2004 (http://mgc.nci.nih.gov/). This is about half of the expected total of 28,000-34,000 genes in humans (Crollius et al., 2002), a number derived from knowledge of other species' genomic sequence and gene sequence frequencies. However, the number of genes does not include functional units such as regulatory regions. Alongside these, the diversity of gene interactions and different expression patterns of the transcribed isoforms give versatility to protein function. The progress of the human genome project has increased database information of both the mapping elements in the genome as well as expressed sequences, and has offered tools for the positional mapping of genes as well as building blocks for gene discovery.

In Finland, concomitant with the human genome project, significant progress has been made in identifying the genes underlying disorders of the so-called Finnish disease heritage. The concept of the Finnish disease heritage covers a wide spectrum of inherited conditions occurring more frequently in Finland than elsewhere. In the majority of these the founding disease-causing mutation has been found only in Finland, but in some the founder mutation has originated elsewhere. For instance, in myoclonic epilepsy of Unverricht-Lundborg type (EPM1, Virtaneva et al., 1997) the founder mutation has been suggested to have been brought into Finland from North Africa (Moulard et al., 2002). On the other hand, Northern epilepsy (EPMR, Hirvasniemi et al., 1994) occurs exclusively in the Kainuu province in Finland and the disease-causing mutation has not been found elsewhere. The background for the positional cloning of Finnish disease heritage genes is built on the

extraordinary population structure and patterns of population movement during the early days of the inhabitation of Finland. The 36 Finnish disease heritage disorders can be divided into five subgroups, based on time of migration and geographic origin of the affected individuals (Norio, 2003a). Cohen syndrome belongs to the largest group, comprising about half of the Finnish disease heritage disorders, in which family origins are clustered in the area of late settlement (Norio, 2003a). Gene mutation enrichment in this group was initiated in the 1500s, when southern Savo farmers sought new cultivation land and populated the eastern, middle and northern parts of Finland (Norio, 2003a). The relatively small subisolates and low bi-directional gene flow between them provided conditions for the search for genes by linkage disequilibrium, which utilizes conservation of genomic regions around susceptibility loci.

To date, the disease gene for 29 Finnish disease heritage disorders have been identified, and the disease gene locus is known for an additional five diseases. We can now include the Cohen syndrome gene *COH1* in the growing group of Finnish disease heritage disorders in which the gene defect underlying the disease is described. The primary goals for this thesis work have been to identify the disease gene underlying Cohen syndrome, to set up methods for laboratory diagnosis, and to clarify the clinical definition of Cohen syndrome. The exceptional Finnish population structure has provided a firm ground for this endeavour.

REVIEW OF THE LITERATURE

1. Cohen syndrome

Cohen syndrome (OMIM#216550) is a developmental disorder inherited as an autosomal recessive trait. The first description of this multisystemic disease in 1973 introduced a syndrome with peculiar faces and multiple affected organs (Cohen et al., 1973). The phenotype was described in three affected individuals, one sibling pair and an unrelated patient, who all had hypotonia, obesity, a high nasal bridge, and prominent incisors as well as mental deficiency. Mottled pigmentation of the retina was also described. In 1978, Carey and Hall published four additional cases with a Cohen syndrome phenotype. The involvement of chorioretinal dystrophy and isolated granulocytopenia in Cohen syndrome was described in 1984 (Norio et al., 1984), based on observations in nine Finnish patients.

1.1. Cohen syndrome in Finland

The incidence of Cohen syndrome in Finland is one in 105,000 nationwide, and one in 60,000 when only the provinces with family histories of Cohen syndrome are considered (Norio, personal communication). This corresponds to the occurrence of approximately one affected newborn every two years. However, the number of new cases seems to be diminishing in Finland. This is probably due to migration from sparsely populated rural regions to densely populated communities. The geographical distribution of Cohen syndrome families covers practically the whole of Finland except the sparsely populated province of Lapland, but the highest prevalence is in the late settlement region including South Savo (Figure 1). To date, 34 Finnish patients have been clinically diagnosed with Cohen syndrome.

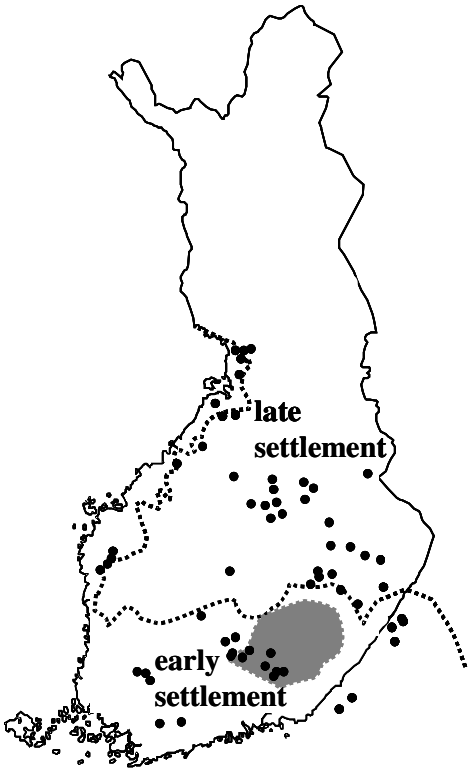


Figure 1. Geographical distribution of grandparental birthplaces of Cohen syndrome families in Finland. The area filled with gray color denotes the late settlement region in South Savo.

1.2. Clinical manifestation of Cohen syndrome in Finnish patients

Cohen syndrome is a clinical entity that has a complex multisystem involvement. In regards to diagnosis, the most important disease manifestations can be separated into four categories: affection of the central nervous system, dysmorphic bone development, retinal changes, and aberrance in leukocyte number. Both motor and mental developmental milestones are delayed and the intelligence quotient (IQ) varies from mild to severe mental deficiency (Kivitie-Kallio and Norio, 2001). The facial features include thick hair and eyebrows, flame-shaped lid-openings, prominent nose bridge, short philtrum and prominent and large upper central incisors (Norio et al., 1984). The faces of young Cohen syndrome patients have a charming general expression, whereas the facial features become coarser in

older patients. Granulocytopenia is present intermittently, with the granulocyte value at low or below normal values resulting in relative lymphocytosis. Cohen syndrome is a non-progressive disorder with the exception of retinal changes, which lead to a decrease in visual acuity and are usually present in patients from the age of five years, progressing finally to a severe visual defect (Norio et al., 1984). Progression of the eye manifestations follow a pattern similar to that in retinitis pigmentosa (RP), where the initial symptom is usually defective dark adaptation or "night blindness", followed by progressive constriction of visual fields *i.e.* "tunnel vision".

Based on analysis of 29 Finnish patients presumed to be genetically homogenous, Kivitie-Kallio and Norio (2001) determined the essential features of Cohen syndrome as non-progressive psychomotor retardation, motor clumsiness, microcephaly, typical facial features (high-arched or wave-shaped eyelids, short philtrum, thick hair, and low hairline), childhood hypotonia and hyperextensibility of the joints, retinochoroidal dystrophy and myopia, and periods of isolated granulocytopenia. Additional findings frequently observed (>50% of Finnish Cohen syndrome patients) include reduced fetal activity, neonatal feeding difficulties, delayed puberty, short stature, high and narrow palate, small or absent lobuli of ears, narrow hands and feet, wide gap between toes one and two, brisk tendon reflexes, high-pitched voice, kyphosis, and a cheerful disposition (Kivitie-Kallio and Norio, 2001).

1.3. Phenotype heterogeneity and intrafamiliar variation in Cohen syndrome

The clinical picture of Cohen syndrome has often been delineated. In many cases only some of the essential criteria are fulfilled (Balestrazzi et al., 1980; Goecke et al., 1982; Sack and Friedman, 1986; Massa et al., 1991), and few case reports depict patients who have a clinical picture consistent with Finnish Cohen syndrome patients (Carey and Hall, 1978, Fryns et al., 1996, Horn et al., 2000; Okamoto et al., 1998, Warburg et al., 1990). Of the approximately 100 patients described only 20 appear to have a disease phenotype similar with Finnish patients, in regards to the main diagnostic criteria (Kivitie-Kallio and Norio, 2001). Chandler et al. (2003) reported an additional 33 Cohen syndrome patients from 22 families of British, Arabic and Dutch origin. These patients represented a group with clinical features compatible with Finnish patients with the exception of three patients

who had normal leukocyte counts. The wide variation in the Cohen syndrome phenotype has been proposed to be due to either allelic or locus heterogeneity (Kondo et al., 1990; Kivitie-Kallio and Norio, 2001; Chandler and Clayton-Smith, 2002). Chandler et al. (2003) proposed modified diagnostic criteria, and suggested that diagnosis of Cohen syndrome should be based on the presence of at least two of the essential signs in a patient with learning difficulties/mental retardation: typical facial gestalt, pigmentary retinopathy, and neutropenia ($<2 \times 10^{-9}$).

Intrafamiliar variation in the Cohen syndrome phenotype has been reported on at least four occasions (North et al., 1985; Young and Moore, 1987; Carey and Hall, 1978; Horn et al., 2000). Kivitie-Kallio and Norio (2001) disputed the phenotype described in the first two of the above publications. The phenotype described in the other two is compatible with Finnish diagnostic criteria, and indicates phenotype variability in patients likely to be affected by the same mutation(s). Carey and Hall (1978) described four patients with Cohen syndrome, two of whom were sibs that differed in the presence of microcephaly and in facial habitus. Leukopenia, a sign considered to be essential in Cohen syndrome diagnosis (Kivitie-Kallio and Norio, 2001), was not evaluated in these patients. Whether they have Cohen syndrome is unproven, since no evidence of a *COH1* locus association has been shown. Horn et al. (2000) reported a multiple consanguineous kindred of Lebanese descent with intrafamilial variability in disease severity (Horn et al., 2000). The phenotype in these patients, two brothers and a cousin, co-segregated with the *COH1* locus in homozygosity mapping (Horn et al., 2000), and mutation analysis later confirmed the Cohen syndrome diagnosis (Hennies et al., 2004). These patients had moderate to severe mental retardation, microcephaly, short stature, and retinopathy. Neutropenia was absent. The presence of synophrys in these patients was exceptional in the Cohen syndrome facial gestalt, and the facial stigmata of these patients are proposed to be due to a different ethnogenic background (Horn et al., 2000).

1.4. Clinical investigations in Cohen syndrome

Due to the multitude of symptoms and clinical variability, Cohen syndrome has been suggested to be either a connective tissue disorder (Thomaidis et al., 1999) or a metabolic disorder (Okamoto et al., 1998). The components of the connective tissue disorder involve

an infrequently observed decreased left ventricular function arising in older patients, along with essential features like hypotonia, craniofacial dysmorphia and limb malformations (Kivitie-Kallio et al., 2001). Okamoto et al. (1998) proposed that the Cohen syndrome pathomechanism is associated with metabolic abnormality in three patients with essential signs and symptoms of Cohen syndrome and high levels of urinary hyaluronic acid. This sign was not present in Finnish patients, in whom the metabolic assay was negative. In addition, leukocyte morphology was normal and brain magnetic resonance imaging (MRI) was negative for any signs of lipid storage material accumulation within cells. Brain imaging in Cohen syndrome patients has not shown any gross pathological changes. The most significant observation in brain MRI has been a relatively enlarged corpus callosum (Kivitie-Kallio et al., 1998). This structure is made up of a substantial cluster of axonal fibers and works as a passage for nerve fibers between the two cerebral hemispheres. Recently, it has been reported that abnormal thinning of this part of the brain is associated with attention deficit syndrome (Pueyo et al., 2003) and alexithymia (Grabe et al., 2004), which is a manifestation of a deficit in emotional cognition. These observations link this part of the brain to the processing of emotions and support the proposed importance of this region in the development of the positive disposition of Cohen syndrome patients (Kivitie-Kallio et al., 1999).

1.5. Cohen syndrome differential diagnostics

Several developmental disorders have been often confused with Cohen syndrome. These are described in more detail below. In many of them multiple disease genes are involved, three of them belonging to continuous gene deletion syndromes (Williams-Beuren syndrome, Prader-Willi syndrome, Angelman syndrome), and one of them being genetically heterogeneous syndrome (Bardet-Biedl syndrome). The number of genes involved partly explains the phenotypic complexity of these disorders. Cohen syndrome and Alström syndrome are both monogenic disorders. Mirhosseini-Holmes-Walton syndrome (Mirhosseini et al., 1972) has been proposed to be an allelic variant of Cohen syndrome (Norio and Raitta, 1986).

1.5.1. Bardet-Biedl syndrome

Bardet-Biedl syndrome (BBS; OMIM#209900, Bardet, 1920; Biedl, 1922) is probably one of the most difficult disorders to distinguish from Cohen syndrome in differential diagnosis. Like Cohen syndrome, BBS patients have a manifestation of mental retardation, pigmentary retinopathy, and similar facial dysmorphic features. These two disorders differ in loss of central vision in adolescence, polydactyli, male hypogenitalism, kidney malformations, renal dysfunction, diabetes mellitus, facial characteristics, and normal intelligence in some BBS patients. Granulocytopenia, present almost in all Cohen syndrome patients, is absent in BBS. Facial dysmorphism is inconsistent in BBS, and the most outstanding feature is deep-set eyes. Similar facial features to Cohen syndrome are: microcephaly, thick hair, coarse eyebrows, downward slant of the eyelids, broad nasal bridge, short philtrum, and prominent incisors. BBS is known to be a heterogeneous disorder with at least eight genes underlying the disease (Mykytyn et al., 2002; Nishimura et al., 2001; Fan et al., 2004; Mykytyn et al., 2001; Li et al., 2004; Slavotinek et al., 2001; Katsanis et al., 2000; Badano et al., 2003; Ansley et al., 2003). Additionally, the inheritance pattern is contradictory to Cohen syndrome, as a triallelic inheritance mode has been proposed for BBS when an additional mutation in a second locus was observed in some BBS patients (Katsanis et al., 2001; Burghes et al. 2001). The gene mutation diagnosis in BBS is elaborate due to several causative genes, which are large in size, and probably many additional disease genes are yet to be determined.

1.5.2. Williams-Beuren syndrome

Williams-Beuren syndrome (WBS; OMIM#194050, Williams, 1961; Beuren, 1972; Grimm and Wesselhoeft, 1980) has a dominant inheritance pattern, and while it shares clinical features of mental deficiency, short stature and cataracts, the cardiovascular symptoms involving supravalvular aortic stenosis and multiple peripheral pulmonary arterial stenoses are not observed in Cohen syndrome. A characteristic “elfin” face is also distinctive in Williams-Beuren syndrome, including short palpebral fissures, a stellate pattern in the iris, medial eyebrow flare, a depressed nasal bridge with anteverted nares, and thick lips (Jones

and Smith, 1975). Genes known to be causative in Williams-Beuren syndrome include *ELN* (Ewart et al., 1993), *RFC2* (Peoples et al., 1996) and *LIMK1* (Tassabehji et al., 1996).

1.5.3. Prader-Willi syndrome and Angelman syndrome

Prader-Willi syndrome (PWS; OMIM#176270, Prader et al., 1956) is similar to Cohen syndrome in respect to mental retardation, growth retardation, newborn hypotonia (which is more profound in PWS), small hands and feet, tapering fingers and strabismus. Facial characteristics are narrow bifrontal diameter, upslanted almond-shaped eyes, full cheeks, and diminished mimic activity due to muscular hypotonia. Central obesity, infrequent in Cohen syndrome, is a major diagnostic criterium in PWS (Gunay-Aygun et al., 2001; Kivitie-Kallio and Norio, 2001). Ocular hypopigmentation is proposed to be a result of misrouting of optical fibers (Creel et al., 1986). In Cohen syndrome the optic disk, fundus as well as the retina around the pigment formation is pale, due to atrophy of the retina (Kivitie-Kallio et al., 2000). Abnormal visual evoked potential (VEP) and nystagmus have been observed in both PWS and Cohen syndrome (Roy et al., 1992; Kivitie-Kallio et al., 2000)

Angelman syndrome (AS; OMIM#105830, Angelman, 1965) resembles Cohen syndrome in the presence of motor and mental deficiency, in general more severe in AS, hypotonia, abnormal choroidal pigmentation, large mandible and open-mouth appearance (Bower and Jeavons, 1967). Choroidal pigment hypoplasia has also been reported in these patients. Infrequently seen in Cohen syndrome patients, epileptic seizures are often present in AS along with ataxia and an abnormal ‘happy puppet’ behavioral pattern (North et al., 1985, Thomaidis et al., 1999).

The genomic region containing genes responsible for PWS and AS overlap on chromosome 15q11-q13 (Magenis et al., 1990).

1.5.4. Alström syndrome

Alström syndrome (ALMS; OMIM#203800, Alström et al., 1959) involves dystrophic retinopathy and obesity. In contrast to Cohen syndrome, ALMS patients are not mentally retarded. In addition, the progress of retinal degeneration differs. Central vision is

exceptionally affected early on (Russell-Eggitt et al., 1998). Other features constantly seen in ALMS, but not in Cohen syndrome, involve deafness, diabetes mellitus, and abnormal lipid metabolism (Charles et al., 1990). The *ALMS1* gene on chromosome 2p13 is known to be causative (Collin et al., 2002, Hearn et al., 2002).

1.5.5. Mirhosseini-Holmes-Walton syndrome

Mirhosseini-Holmes-Walton syndrome (OMIM#268050, Mirhosseini et al., 1972, Mendez et al., 1985) clinically resembles Cohen syndrome, and whether these are clinically and genetically uniform entities has been disputed (Norio et al., 1986; Steinlein et al., 1991). These two disorders diverge considering the main clinical features only in respect to intermittent neutropenia, not reported in Mirhosseini-Holmes-Walton syndrome. The presence of mental retardation, ophthalmic changes with myopia, pigmentary retinal dystrophy and cataracts as well as typical craniofacial features, microcephaly, hypotonia, and hyperextensibility of joints in both Cohen and Mirhosseini-Holmes-Walton syndrome link these syndromes clinically.

2. Gene mapping and positional cloning

2.1. Approaches for gene mapping projects

The strategy for many molecular genetic research projects aiming at the identification of disease genes is to target the investigation to a specific, refined region in the genome. Identification of a disease gene on the basis of its location in the genome is called the positional cloning approach (Collins, 1992, 1995). In the positional candidate gene cloning approach determination of the gene localization is followed by analysis of a functionally relevant gene residing in the region. This method was first used in the identification of the *CFTR* gene underlying cystic fibrosis (Riordan et al., 1989). After completion of the total human genomic sequence positional cloning has been used almost without exception in gene hunting projects.

Earlier, functional cloning was a commonly used method. This approach is based on fundamental information about the basic biochemical defect without reference to chromosomal position. When the defective protein was known, knowledge of its amino acid sequence was utilized in the isolation of the disease gene. This approach was used, for example, to identify the HOGA disease-causing ornithine-δ-aminotransferase gene (Valle and Simell, 1983). Another approach has been candidate gene cloning, which solely focuses on a group of known genes which may be suspected, on the basis of their function, to have a role in the pathophysiology of the disease, without previous knowledge of the location of the sought-after gene in the genome.

The positional cloning approach consists of: 1) segregation analysis of the disease susceptibility locus by linkage-based methods; 2) linkage disequilibrium and haplotype analysis for refined disease gene locus determination; 3) physical mapping of the region; 4) identification of positional candidate genes in the sequence; 5) identification of the disease-associated mutation.

2.2. Linkage analysis

The first step in positional cloning is to genotype the affected families and search for the segregation of affection status with the disease gene locus, by studying the familial

transmission of marker alleles at consecutive polymorphic loci. This necessitates statistical methods to interpret genome-wide data. The descriptive unit for the strength of linkage is the *logarithm of odds i.e.* lod score value, which is based on an equation developed by Newton E. Morton (1955).

The lod score is the 10^{th} base logarithm for the likelihood ratio or odds ratio for the likelihood of linkage at a given recombination fraction (θ) between affection status and a marker locus to the likelihood of no linkage (Ott, 1985). In practice this ratio is computed for several values of recombination fraction. The frequency of one recombination event in 100 meioses equals a map distance of one centiMorgan (1 cM ≈ 0.01 θ) (Ott, 1991). This is 1 Mb on average in physical distance, but it varies between males and females and depends on chromosomal location. The estimate for linkage is the sum of lod scores at a given recombination fraction in single families. The lod score calculation is dependent on both the mode of transmission and penetrance of the disease phenotype. The estimation of linkage for a single genomic locus depends only on the last meiosis and gives a reliable, but usually also relatively gross localization for the affection locus, the most likely distance between the loci studied being the recombination fraction at which the lod score is highest. In theory the probability of two recombinations in a region of 1 Mb is, on average, one in 10,000, but depends on the true recombination frequency in a given region (Haldane, 1919). This figure holds true for one meiosis, but a linkage study utilizes information collected from siblings and several affected families. Lod scores ≥ 3 are considered significant since they indicate 1:1000 odds that the linkage did not occur by chance. Lod scores < -2 are generally considered as significant evidence against linkage (Morton, 1955, Ott, 1991).

2.3. Linkage disequilibrium and haplotype analysis

Linkage disequilibrium (LD) and haplotype analyses have been used frequently to refine the initial disease gene locus in positional cloning of disease genes (de la Chapelle and Wright, 1998; Peltonen et al., 1999). The concept of LD can be interpreted as conservation of a region of ancestral origin in the genome extending over polymorphic loci around the disease-causing locus. LD can be applied in a single consanguineous family with a recessive monogenic trait (Lander and Botstein, 1987) and in isolated populations with

small numbers of founders. When a gene defect originating from a founder is enriched in a population with low gene flow from the outside certain alleles are over-represented in affected when compared to unaffected individuals. The strength of LD is dependent on the age of the mutation and the frequency of the associated allele in a control population. The extent of LD decreases over time at a rate proportional to the recombination rate (Hästbacka et al., 1992; Lehesjoki et al., 1993; de la Chapelle, 1993). The age of the mutation can be estimated applying the Luria-Delbrück-based algorithm (Hästbacka et al., 1992). This method was further developed to calculate the distance between the affection locus and a polymorphic marker locus as a function of the proportion of disease-causing chromosomes descending from a common ancestor (Lehesjoki et al., 1993). The strength of the association is denoted by p_{excess}-value, which can be calculated using equation where the excess between a given allele frequency in disease-causing chromosomes and the frequency of the same allele in the general population is divided by the frequency of other alleles in the general population. In addition to manual linkage disequilibrium calculation, computer-based methods (DISLAMB for single locus and DISMULT for multiple loci LD calculation) have been developed (Terwilliger, 1995). The DISMULT program uses information from all marker loci simultaneously and has a built-in location parameter. The basic algorithm in both of these programs contain the parameter lambda (λ), which is equal to the proportion of increase of allele *i* in the disease chromosomes, relative to its population frequency (Terwilliger, 1995).

Haplotype analysis based on the concept of LD has been a method of choice in disease gene locus identification in many diseases more prevalent in Finland than elsewhere (de la Chapelle and Wright, 1998). An haplotype is a set of joined alleles in subsequent polymorphic loci in a given chromosome. Haplotype analysis is based on historical conservation of the genomic region around the disease gene in chromosomes sharing the common founder mutation. The length of the conserved haplotype is population-size and agedependent, and diminishes when recombinations or novel marker mutations occur in subsequent generations. In Finland, the time elapsed between mutation founding and the present is long enough for refined mapping of the disease gene by using the information of historical recombinations (de la Chapelle and Wright, 1998).

2.4. Polymorphic markers

Traditionally length polymorphisms (e.g. di-, tri-, and tetranucleotide repeats and VNTR markers) have been used in linkage, LD and haplotype analyses. In addition, the growing number of single nucleotide polymorphisms (SNPs) are nowadays being employed. SNP information is also utilized in loss-of-heterozygosity and haplotype-block analysis, and they can be studied as modifiers of the phenotype in genetic disorders. While length polymorphisms give comparatively higher analytical power, SNPs are in general more stable against *de novo* mutations (Ohashi and Tokunaga, 2003). The frequencies of mutation rates per generation for length polymorphisms is around 10^{-3}~10^{-4} on average compared to the considerably lower mutation rate for SNPs, approximated to be 10^{-8} or less (Drake et al., 1998). The benefit of SNPs is in higher resolution genetic maps. SNPs are estimated to occur every 357 bp (January 2004 release of NCBI SNPdb), and one might expect 9.1 million SNPs in the genome. In contrast, there are currently 944,592 known di-, tri-, and tetranucleotide repeats (SSRD; URL: http://www.ingenovis.com/ssr/).

2.5. Physical mapping

Physical mapping of the human genome had two objectives during the human genome project. Firstly, to create framework maps for sequencing projects, and secondly to locate ESTs, and to identify and position the full-length transcripts identified by EST contigs or by other *in silico*- (see Review of the Literature section 3.3.) and *in vitro*-based methods. Before the human genome sequence became available, physical maps were constructed with genomic libraries in which the human genome is fragmented in genomic clones containing human DNA inserts. Genomic cloning vectors designed for this purpose were yeast artificial chromosomes (YACs, Burke et al., 1987), bacterial artificial chromosomes (BACs, Shizuya et al., 1992), bacteriophage P1-derived chromosomes (PACs, Ioannou et al., 1994) and cosmids (Meyerowitz et al., 1980). The size of the insert depends on the vector, with the largest inserts of ~500 kb cloned in YACs, and the smallest ~48 kb in cosmids.

Another method developed for genomic mapping was radiation hybrid (RH) mapping (Goss and Harris, 1975; Cox et al., 1990: Walter et al., 1994). The RH method is

based on random fusion of irradiated human cells with hamster recipient cells after fragmentation of a donor genome by radiation (Goss and Harris, 1975). DNA from 80-100 independent hybrids is analyzed for the presence or absence of DNA markers of interest, and the mapping unit distances are calculated using a computer program designed to handle statistical data analysis of joined linkage groups (Boehnke et al., 1991, 1992; Lunetta and Boehnke, 1994; Slonim et al., 1997). The centiRay (cR) unit is equal to 280 kb in the Whitehead Institute GeneBridge RH panel (Gyapay et al., 1996) and 25 kb in the Stanford G3 panel (Stewart et al, 1997). The mapping unit order is computed by applying the minimized number of obligate chromosome breaks. The RH method has higher resolution than linkage mapping and is more robust than the cloning vector-based physical mapping approach (Bishop and Crockford, 1992). The advantage to linkage analyses is its ability to estimate a map position also for non-polymorphic markers.

2.6. Identification of coding sequences

Prior to the availability of annotated human genome sequence one had to rely on different laboratory methods to identify genes in genomic clones mapped to the region of interest. These include: identification of CpG islands (Gardiner-Garden and Frommer, 1987), cDNA direct selection (Lovett et al., 1991), and exon amplification (Duyk et al., 1990). To ensure the identification of as many genes as possible one usually had to apply many different methods simultaneously. Concominant with the progress of the human genome project a constantly growing number of ESTs, pinpointing the localization of coding sequences, were deposited in the databases. Now that the genomic location at the majority of human genes is available in published gene maps, the recognition of transcript units from the genomic region of interest by biocomputing has become possible. This has made gene identification easier and has replaced *in vitro*-based techniques for gene isolation. The first human gene map was published in June 1996 (Schuler et al., 1996). Mapping data was based on YAC - based contigs and RH maps, which were integrated into the framework human gene map containing 16,000 human genes and 1000 polymorphic genetic markers (Schuler et al., 1996). This preliminary map was followed by the human genome consortium release of 30,000 human genes in October 1998 (Deloukas et al., 1998).

Today, the entire human genomic sequence is available as a sequence contig, and the accurate position of transcript units can be determined from the mapping data present in an electronic form in sequence databases. To make it easier to interpret the mapping data, graphical interfaces for data mining have been developed in which the position of the mapping units and their relative distances to each other within a specific genomic region can be seen simultaneously. The latest assembly of human genome data is available from the University of California Santa Crutz (UCSC; URL: http://genome.ucsc.edu/cgi-bin/hgGateway), which is based on the National Center for Biotechnology Information (NCBI) Build 34 human reference sequence produced by the International Human Genome Sequencing Consortium (Lander et al., 2001). The UCSC genome browser also shows alignment of human sequence to chimpanzee, mouse, rat, and chicken as well as Fugu fish genomic sequence. Ensemble (http://www.ensembl.org/), a joint effort of EMBL-EBI and the Sanger Center, contains larger sets of genomic data, presently of 12 different species.

2.7. Mutation analysis

The final step in positional cloning is to identify disease-associated mutations in patient samples in genes identified from the region. Several methods exist with different sensitivities and costs. Methods used in mutation analysis include Southern (Southern, 1975) and Northern blot (Sambrook et al., 1989) analysis, single-strand conformation polymorphism analysis (SSCP; Orita et al., 1989), and denaturing gradient gel electrophoresis (DGGE; Fischer and Lerman, 1983). Any change detected by the above means has to be confirmed by sequencing to characterize the variation at the nucleotide level and for this reason sequencing is often used as a primary method. These methods are nowadays largely replaced by semi-automated mutation analyses, which are suitable for analysis of larger sample sets. Semi-automated techniques include heteroduplex analysis by denaturing high-performance liquid chromatography (DHPLC; Oefner and Underhill, 1995), and automated sequencing, which is performed by capillary electrophoresis (Karger, 1996). In addition, for large-scale diagnostic mutation analysis, minisequencing (Jalanko et al., 1992) and ligation-based methods e.g. oligonucleotide ligation assay (OLA; Alves and Carr, 1988; Landegren et al., 1988) and ligation chain reaction (LCR; Barany, 1991) have been developed.

3. Bioinformatics and gene identification tools

3.1. Strategy of human genome sequencing

Human genome sequencing was accomplished by a publicly funded project, primarily led by National Institutes of Health (NIH) and the U.S. Department of Energy, and the commercial Celera led project. The fundamental methodological difference between them was in the sequence assembly strategies. Celera used a whole-genome shotgun sequencing method (Venter et al., 1998) whereas the public consortium relied on a map-based approach. The public human genome project was carried out in three phases: 1*) A mapping phase*, when the first established genetic maps allowed the use of intermarker order and distances in physical map construction. The physical maps consist of clones of large genomic fragments arranged in contigs with overlapping marker loci. 2) *A sequencing phase*, which used automated sequencing of selected single clones covering the human genome in shotgun cloned genomic libraries, and *in silico*-assembly of the produced sequence. 3) *Utilization of obtained genomic sequence* data to gain knowledge about human sequence variation, gene identification, and elucidation of genomic organization by cross- and inter-species comparisons. These stages and the goals of this academically led project were reached over a twelve-year period (Collins et al., 2003), although analyses of the genomic sequence obtained and interpretation of the results are still continuing.

3.2. Tools to assemble sequence data in large sample sets

The assembly of provisional sequence from the library clones to genomic contigs demanded high biocomputing capacity. It also required the development of more efficient *in silico*-based programs for effective sequence quality analysis and alignment. Unix-based programs (http://www.phrap.org) for quality estimation (Phred; Ewing et al., 1998a; Ewing and Green, 1998b), sequence assembly (Phrap; Green, unpublished) and alignment (Consed; Gordon et al., 1998) were utilized to compose the single reads into sequence contigs. A quality assessment criterion for this was, depending on the sequencing center, eight- to ten-fold coverage of overlapping sequences.

3.3. Gene sequence identification

3.3.1. Sequence homology programs

The exponential growth of sequence information in databases has necessitated the development of more powerful computational methods to identify homologous sequence patterns. Sequence alignment has been used in genomic localization of a given sequence, in the search for transcript sequences, and for pattern similarity recognition of functional elements. The similarity search programs, which have evolved from simple algorithms for sequence alignment (FASTA; Lipman and Pearson, 1985), have resulted in increased calculation capacity. The development from the single-pass database-search method, basic local alignment search tool (BLAST; Altschul et al., 1990), to an iterated profile-based search method, PSI-BLAST (Altschul et al., 1997), which utilizes position-independent gap scores of Gapped Blast search, has permitted local blast searches with gapped alignments. This improvement has resulted to 10-100 times faster sequence alignment (Altschul and Koonin, 1998). While the Blast program similarity search is based on the length of continuous homology between the sequences, the Gapped Blast search also recognizes similarities that contain gaps in the middle of the homologous region. The cutting of the query sequence into smaller units in repeated similarity searches has enhanced sensitivity in similarity identification of sequences having intermittent segments of low homology.

3.3.2. Exon prediction algorithms

Several biocomputing tools to extract gene sequences from the entire genomic information have been developed. Prediction programs can be separated into those that utilize general models for gene structure and the regulatory elements in the genome (*ab initio* or *intrinsic* methods), and those that are based on cross- and intra-species conservation of protein coding sequences (*extrinsic* methods) (Korf et al., 2001). A third, integrated, approach is the homology-based method in which cross- or intra-species sequence comparisons are combined with structural information (*e.g.* Procrustes; Gelfand et al., 1996).

Signal detection and codon statistics based *intrinsic* methods utilize only the structural information of the genomic organization of the genes (Mathé et al., 2002). This

compositional and signal information, organized in training sets based on known genes, is used in the prediction of exons by intrinsic methods algorithms. The pattern recognition algorithms used by *intrinsic* methods are neural networks, discriminant analysis, and hidden Markow models (Murakami and Takagi, 1998). Homology search-based *extrinsic* methods compare the genomic sequence to known gene sequence at either the genomic, cDNA or protein level (Mathé et al., 2002). The basic assumption behind this method is that coding regions evolve slower than non-coding regions.

In silico exon prediction can only be suggestive, and all these methods have disadvantages. The exon prediction programs utilizing *intrinsic* approaches in exon discovery have a tendency to more reliably identify genes residing in GC-rich regions, when the preference for identification is of medium-size exons (length range between 70 and 200 nucleotides) and in internal exons, which do not contain start and stop signals for protein coding (Rogic et al., 2001). The weakness of *extrinsic* method is that genes without homologues in databases are missed and comparison of translated genomic sequence to protein sequence is sensitive to frameshift errors. Single programs have differences in accuracy, but the best prediction result can be obtained by combining the information from several programs (Murakami and Takagi, 1998).

3.3.3. CpG islands

Another approach to identify putative gene elements within genomic sequence is to search for regions having high, over 50%, C+G content *i.e.* CpG islands. In humans and mice, approximately 60% of all promoters co-localize with CpG islands devoid of methylation (Antequera, 2003). GC-rich regions usually represent upstream regulatory segments of genes, working possibly both in transcriptional and post-transcriptional regulation of gene expression, and are positioned either upstream or downstream from transcription factor (TF) binding sites (Gardiner-Garden and Frommer, 1987). Sometimes this regulatory element overlaps with the CpG island. Provisionally unmethylated CpG islands are detected in promoter regions of housekeeping and regulated genes (Bird, 1986, Larsen et al., 1992). The CpG island is methylation-free in somatic cells and is profusely associated with genes regularly activated (Ghazi et al., 1992). The exception for this rule is observed in some oncogenes (e.g. French et al., 2003; Strathdee et al., 2001). CpG methylation

results in silencing of the associated gene. Examples of computer programs developed to discover CpG island regions are CpG Island (Gardiner-Garden and Frommer, 1987), CpGPlot (Larsen et al., 1992) and accessory applications in the EMBOSS package (http://www.no.embnet.org/Programs/SAL/EMBOSS/). The CpG promoter program (Ioshikhes and Zhang, 2000) discriminates promoter-associated and non-associated CpG islands.

3.3.4. Expressed sequence tags (ESTs)

ESTs are usually partial sequences of cDNA clones representing small segments of expressed genes. Often they correspond to the 5'-coding or 3'-untranslated end of the gene. They are used mainly in gene discovery and physical mapping of genes.

The Institute of Genome Research (TIGR) was the first to start high-throughput cDNA library random sequencing in 1991 (Adams et al., 1991). Today, the gene indices (http://www.tigr.org/tdb/tgi/) contain over 3.7 million (835,000 of them human) unique EST sequences from 82 species. Another EST source is the NCBI EST database (http://www.ncbi.nlm.nih.gov/dbEST/index.html). The total number of ESTs collected in the NCBI databases to date is over 24 million (around six million of them human) and this figure grows rapidly. Because of the increasing number of EST sequences the Unigene collection of genes (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) was developed in NCBI in 1995 (Boguski and Schuler, 1995) (this database combines the ESTs released and creates clusters of overlapping gene sequences). The EST sequences collected in the Unigene database were converted to sequence tagged sites (STSs), which were used as a source for the release of the first gene map in 1996 (Schuler et al., 1996).

In addition to exploiting ESTs as tools to identify transcript units in the genome they have been used in many other applications as well. ESTs are utilized in the determination of expression profiles of genes (e.g. Gress et al., 1996; Khan et al., 1999). ESTs have also been useful in determination of alternatively spliced isoforms of transcripts, and for elucidation of their expression pattern in different libraries or tissues (Thanaraj et al., 2004; Pospisil et al., 2004). Functional annotation of ESTs has helped in determination of gene associations with metabolic and signaling pathways and gene ontology classification of transcripts (e.g. Whitfield et al., 2002; Lee et al., 1999). ESTs are also of

use in detection of single nucleotide polymorphisms (SNPs), which sometimes function as modifiers of the phenotype (Picoult-Newberg et al., 1999).

Although ESTs have greatly enhanced the discovery of novel genes they also have many disadvantages. Some EST databases contain cDNA sequences from cancer cell-line cDNA libraries, in which the transcript sequences can be highly reorganized and do not represent the intact transcript sequence. The accuracy of the EST sequences is dependent on the purity of the mRNA libraries. A small amount of genomic contamination can lead to cloning of the genomic insert instead of the cDNA fragment. The cDNA libraries are also susceptible to bacterial and viral contamination. The technique using poly-T probes to identify poly-A tails of transcripts is used for 'fishing' of putative transcript sequences and might lead to identification of poly-A regions of genome not associated with the expressed sequence. The mRNA libraries can also contain immature transcripts not yet processed to the mature form, containing intronic sequences.

3.4. Protein characteristics predicting programs

Nowadays bioinformatics is more and more concentrated on understanding functions and utilities at the molecular, cellular and organism levels (Kanehisa and Bork, 2003). For the prediction of protein function in cellular processes programs such as PSORTII (Nakai and Kanehisa, 1992) have been designed, which search for protein sorting signals and cell localization site-determining patterns in amino acid sequence. The vast amount of information available demands integration of protein data under one structured database, such as InterPro (Apweiler et al., 2000) and the larger ensemble in the Proteome Analysis database (Pruess et al., 2003), which combines information of protein families, domains, sites, and functions of complete genomes. The protein domain family database, ProDom (Gouzy et al., 1999), aligns proteins by conserved domain structures, and arranges the branching of protein sub-classes in a phylogenetic tree. The applications for ProDom include protein-protein interaction studies and structural genomics (Corpet et al., 2000). In the future, functional predictions will utilize the knowledge base of three-dimensional folding unit structure in functional domain structure identification. The structural classification of proteins (SCOP) database (Barton, 1994) compiles data from three-dimensional protein models according to folding patterns (Reedy and Bourne, 2003). The

three-dimensional position-specific scoring matrix (3D-PSSM) program is an application that utilizes protein fold profiles from the SCOP database to predict the folding pattern for a query protein by coupling 1D- and 3D-protein structures with protein secondary structure (Kelley et al., 2000). The constructed model facilitates the prediction of protein function. However, these computer-based predictions have restrictions in protein modeling and functional analysis. For instance, the huntingtin and orphan proteins are examples of proteins with novel functions not predictable by *in silico* -methods.

3.5. Comparative genomics

The development of cDNA sequence databases has allowed the integration of data for cross-species comparison of sequences, gene intron/exon identification and detection of multiple transcripts. Cross-species comparison of the regulatory regions for gene expression has helped to identify transcription factor binding sites, which should show high similarity in non-coding regions with generally low conservation between species. Liu et al. (2004) proposed an average identity of 69.5% for 127 human and mouse representative gene regulatory elements, with 81% of elements having over 50% similarity. This is considerably higher conservation than that of “background sequences” (Liu et al., 2004). To date, regulatory elements have been identified by applying comparative genomics for several human disease genes (Hansson et al., 2003; Zatyka et al., 2002; Touchman et al., 2001; Loots et al., 2000).

AIMS OF THE STUDY

I The primary aim was to identify and characterize the gene (*COH1*) responsible for Cohen syndrome by positional cloning, and to detect the mutations in Finnish and non-Finnish Cohen syndrome patients.

II The second aim was to correlate the clinical phenotype with the *COH1* gene mutations.

III The third aim was to establish uniform clinical criteria for Cohen syndrome and to facilitate the establishment of a gene test to support the clinical diagnosis.

IV The fourth aim was to predict the function and subcellular localization of the COH1 protein by *in silico*–based methods, and deduce the COH1 protein involvement in biological processes through the known function of the protein homologs.

SUBJECTS AND METHODS

1. Subjects

In total, 75 Cohen syndrome families, 21 of them Finnish (Figure 2), were included in the study. The majority of the patients, except patients in families 15-18, were involved in a nationwide clinical study of Cohen syndrome (Kivitie-Kallio and Norio, 2001) and were diagnosed by Doctor Satu Kivitie-Kallio at the Helsinki University Central Hospital and Professor Reijo Norio at the Family Federation of Finland. Patients in 31 Cohen syndrome families from the United Kingdom were diagnosed by clinical geneticists in a large clinical cohort study over a two year period (1999-2001) in St. Mary's Hospital, Manchester. Chandler et al. (2003) published the clinical phenotype of these patients. Additionally, we studied four Belgian families from Professor Jean-Pierre Fryns in Leuven University. The clinical data of these three patients have been published earlier (Fryns et al., 1996). Six of the eight unpublished Israeli families are from Doctor Varda Gross Tsur in Jerusalem Childrens Hospital. Two Danish families, one of them published (Warburg et al., 1990), were from Doctor Mette Warburg in Glostrup. Doctor Matthew L. Warman in Cleveland, Ohio, USA, sent the Amish family samples to us. Of the remaining eight families, four are from Germany, and one each from Holland, Austria, the USA and Syria. The set of control samples used in this study included 98 Généthon *CEPH* and 56 Finnish Red Cross samples, which were provided to us courtesy of Doctor Pertti Sistonen. The Cohen syndrome patients or their lawful custodians signed an informed consent form for participation in the research, and the study was reviewed and approved by the ethics review board of the Department of Medical Genetics at the University of Helsinki.

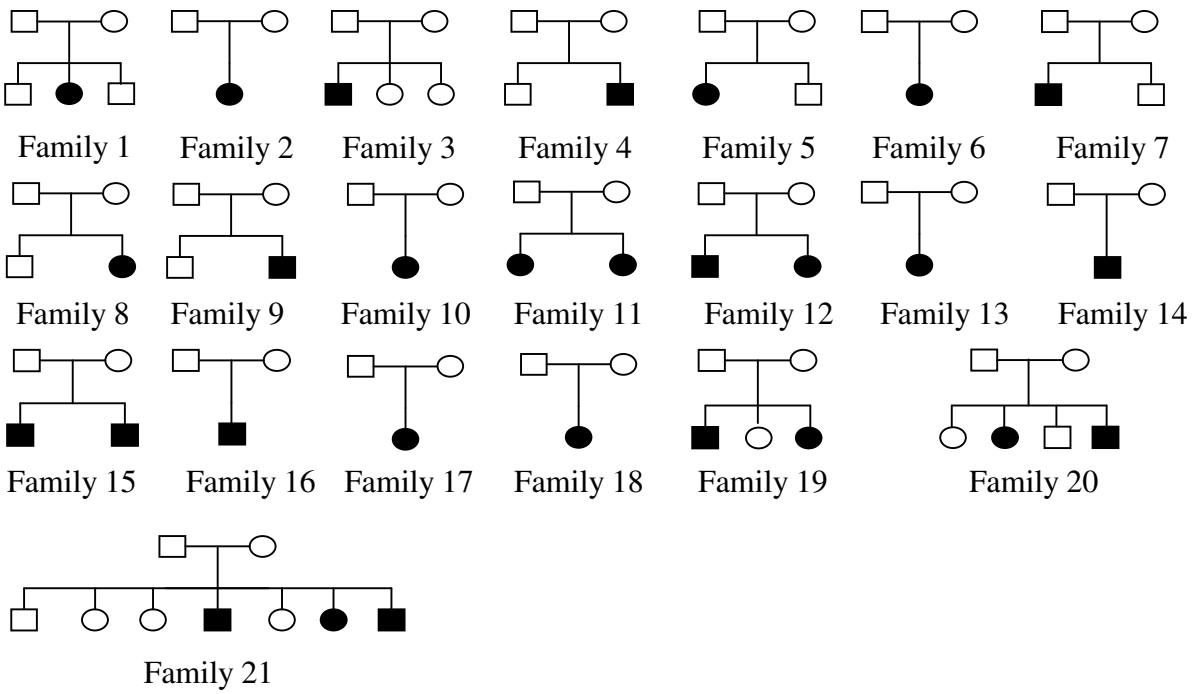


Figure 2. Pedigrees of 21 Finnish Cohen syndrome nuclear families included in the study.

2. Methods

Table 1. Laboratory methods

Material or method	Study	Reference
Polymerase chain reaction (PCR)	I, II, III, IV	Weber and May, 1989
RT-PCR	II, IV	Ohara et al., 1989
Haplotype analysis	I, II	Ramsay et al., 1993
Linkage disequilibrium (LD) analysis	I	Hästbacka et al., 1992; Lehesjoki et al., 1993
DNA extraction	I, II, III, IV	Sambrook et al., 1989
Mutation analysis	II, III, IV	See Review of the Literature section 2.7.
Linkage analysis	I	Botstein et al., 1980; Ott, 1991
Physical mapping	II, unpublished	Brown and Bird, 1986; Green and Olson, 1990
Direct sequencing	II, III, IV	Maxam and Gilbert, 1977; Sanger et al., 1977
Polyacrylamide gel electrophoresis (PAGE)	I, II	Maniatis et al., 1975
Silver staining of polyacrylamide gels	I, II	Bassam et al., 1991
Single strand conformational polymorphism (SSCP) gel electrophoresis	II, IV	Orita et al., 1989
Agarose gel electrophoresis	unpublished	Aaij and Borst, 1972
Restriction enzyme digestion analysis	II	Arber, 1974; Nathans and Smith, 1975
Northern blot analysis	II	Sambrook et al., 1989
Yeast artificial chromosomes (YACs)	unpublished	Burke et al., 1987
Bacterial artificial chromosomes (BACs)	II, unpublished	Shizuya et al., 1992
P1 clones (PACs)	unpublished	Ioannou et al., 1994
TA cloning	IV	Hu, 1993

Table 2. Programs and statistics used in computational analysis

Program	Study	Reference	URL
LINKAGE v.5.0 package	I, III	Lathrop et al., 1985	ftp://linkage.rockefeller.edu/softwar/linkage
Luria-Delbrück-based method	I	Hästbacka et al., 1992, Lehesjoki et. al., 1993,	
DISMULT v.2.1	I	Terwilliger J.D., 1995	ftp://linkage.cpmc.columbia.edu/software/
BLAST	II	Altschul et al., 1990	http://www.ncbi.nlm.nih.gov/BLAST/
Align	II	Pearson and Lipman, 1988	http://molbiol.soton.ac.uk/compute/align.html
Prosite	II	Hofmann et al., 1999	http://expasy.hcuge.ch/sprot/prosite.html
ProDom	II	Gouzy et al., 1999	http://protein.toulouse.inra.fr/prodom/current/html/home.php
Kyte-Doolittle hydrophobicity profile	II	Kyte and Doolittle, 1982	http://us.expasy.org/cgi-bin/protscale.pl
SOPMA	II	Geourjon and Deleage, 1995	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html
PSORT II	II	Nakai and Kanehisa 1992	http://psort.nibb.ac.jp
TMAP	II	Persson and Argos 1994	http://www.mbb.ki.se/tmap/
CAP3	II	Huang and Madan 1999	http://www.tigr.org/tdb/tgi/software/
ClustalW	II	Thompson et al., 1994	http://www.ebi.ac.uk/clustalw/
UCSC Human genome browser	II	Kent et al., 2002; Kent, 2002; Karolchik et al., 2003	http://genome.ucsc.edu/cgi-bin/hgNear
CpGProD	unpublished	Ponger and Mouchiroud, 2001	http://pbil.univ-lyon1.fr/software/cpgprod_query.html
MATCH (interconnected with TRANSFAC® database)	unpublished	Wingender et al., 2001	http://www.gene-regulation.com/pub/programs.html#match
SIM	unpublished	Huang and Miller, 1991	http://www.expasy.org/tools/sim-prot.html

RESULTS AND DISCUSSION

1. Fine-mapping of the *COH1* gene

1.1. Linkage, and linkage disequilibrium fine-mapping of the *COH1* locus (I)

We applied linkage analysis to further refine the *COH1* locus mapped by Tahvanainen et al. (1994). We were able to narrow the previous 10 cM *COH1* region bordered by markers *D8S270* and *D8S521* by only 0.4 cM, to a 9.6 cM region flanked by markers *D8S1699* and *D8S1714*. The allele sharing observed in Finnish *COH1* chromosomes prompted investigation of this region by LD and haplotype analysis. LD was investigated with eight polymorphic markers non-recombinant with *COH1*. Highly significant ($p<0.001$) LD was observed with markers *D8S559* and *D8S1762* (I, Table 1). Additionally, the frequencies of disease-associated alleles of marker loci *D8S506*, *D8S257*, and *D8S546* were significantly higher in Cohen syndrome patients ($0.001<p< 0.01$) than in control cohorts. A DISMULT multipoint LD analysis (Terwilliger, 1995) graph showed a one peak curve, and a highest lod score of 9.2 at map position 0.2 cM proximal to *D8S1808* (I, Figure 3). The 99% confidence interval for this position covered a 3.7 cM region, 1.2 cM telomeric of *D8S506* to 1.1 cM centromeric of *D8S1714*.

1.2. Initial haplotype analysis in Finnish Cohen syndrome patients (I and unpublished)

Haplotypes were constructed manually, and 25 of 29 disease chromosomes were assumed to represent a conserved major haplotype (I, Table 2). *COH1* gene localization by haplotype analysis to an interval flanked by markers *D8S1808* and *D8S546* was based on observation of a shared allele 5 at marker locus *D8S1762* (I, Table 2). This occurred in 87% (27/31) of the *COH1*-bearing chromosomes (I, Table 1) and in only 38% (9/24) of control chromosomes. In addition, the allele combination 2-5-7 in three consecutive marker loci, *D8S1808*, *D8S1762*, and *D8S546*, was seen in 24 disease chromosomes (I, Table 2) and in only two of 28 Finnish control chromosomes (data not shown). Based on the hypothesis that the majority of Cohen syndrome patients have one single founder mutation the haplotype data pinpointed the *D8S1808-D8S546* interval as the most likely region and the initial gene identification searches were focused within this area.

At the time of the first haplotype analysis, the order of markers and their physical map distances (Figure 3) were based on the Généthon genetic and Whitehead Institute physical YAC map (Dib et al., 1996). Subsequently, the sequence-based physical map (Lander et al., 2001; Venter et al., 2001; Figure 3) revealed that the original position of markers was incorrect. Moreover, subsequent studies with additional Cohen syndrome families and a denser marker map proved the initial localization to be incorrect, and allowed localization of *COH1* more proximally, in a region flanked by markers *D8S257* and *D8S559* (see Results and Discussion section 1.4.).

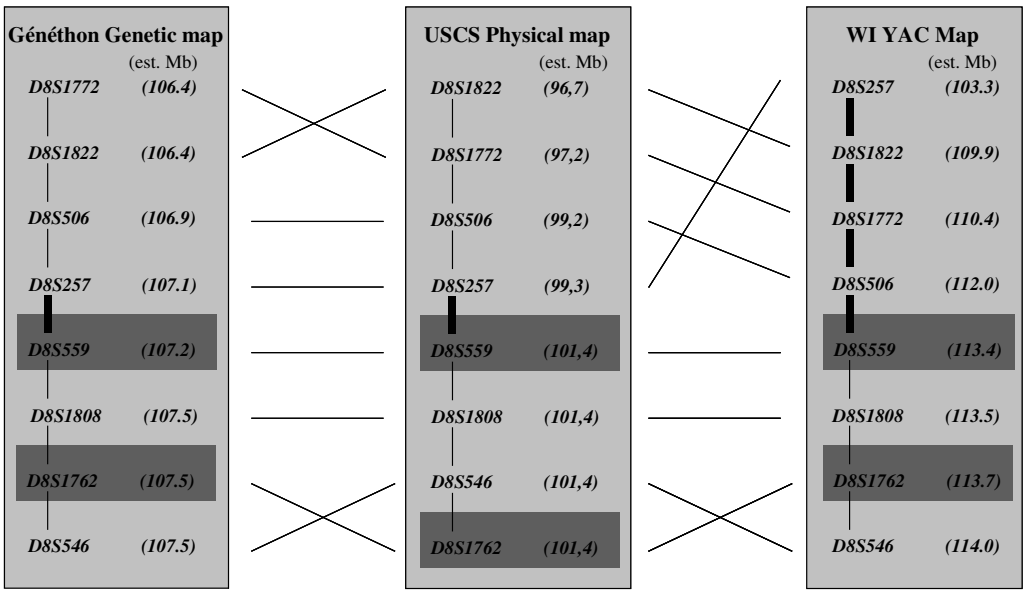


Figure 3. Relative order and intermarker distances of microsatellites around the *COH1* gene locus as presented in three different sources (Généthon genetic map, USCS physical map, and Whitehead Institute yeast artificial chromosome (YAC) map). The UCSC map is based on genomic sequence of the region. The interval in which the *COH1* gene was eventually identified (see Results and Discussion section 1.5.) is marked with a bold lines. Significant ($p<0.001$) linkage disequilibrium was observed with the two microsatellites indicated by a dark gray background color. The connecting lines between the maps indicate the relative positions of each marker.

1.3. Physical map of the initial *COH1* locus (II and unpublished data)

We first focused the physical mapping and gene identification efforts to the region between *D8S559* and *D8S1714*, covering a region of 3.8 cM and 700 kb, implicated by the initial haplotype mapping data. A low-resolution YAC contig was composed of CEPH mega-YACs (Cohen et al., 1993; Chumakov et al., 1995) onto which STSs from the region indicated by the haplotypes were mapped (Figure 4). Information available in public databases allowed the selection of YAC clones containing genomic markers residing within the region of interest. Eight of ten YACs were found to be non-chimeric and did not have large deletions. Our ultimate purpose was to build a gene map that would provide information that could be applied in the identification of candidate genes and in the longer-term goal of sequencing this region. For this the smaller and more stable BAC and PAC clones were ideal. Eight BAC and three PAC clones were found to reside within the region (Figure 4), and one of these, the 162-kb BAC clone 476J3 containing the *COH1* region refined by initial haplotypes around *D8S1762*, was purchased from Genome Systems, and eventually sequenced by us at the Ohio State University. Twelve ESTs were positioned in the physical map. Of these, four represent previously annotated genes: Homo sapiens polymerase (RNA) II (DNA directed) (*POLR2K*) represented by WI-8892, *SPAG1* by WI-7256, a pseudogene for Glyceraldehyde 3-phosphate dehydrogenase by WI-14243 and a potassium channel-like gene by WI-12835.

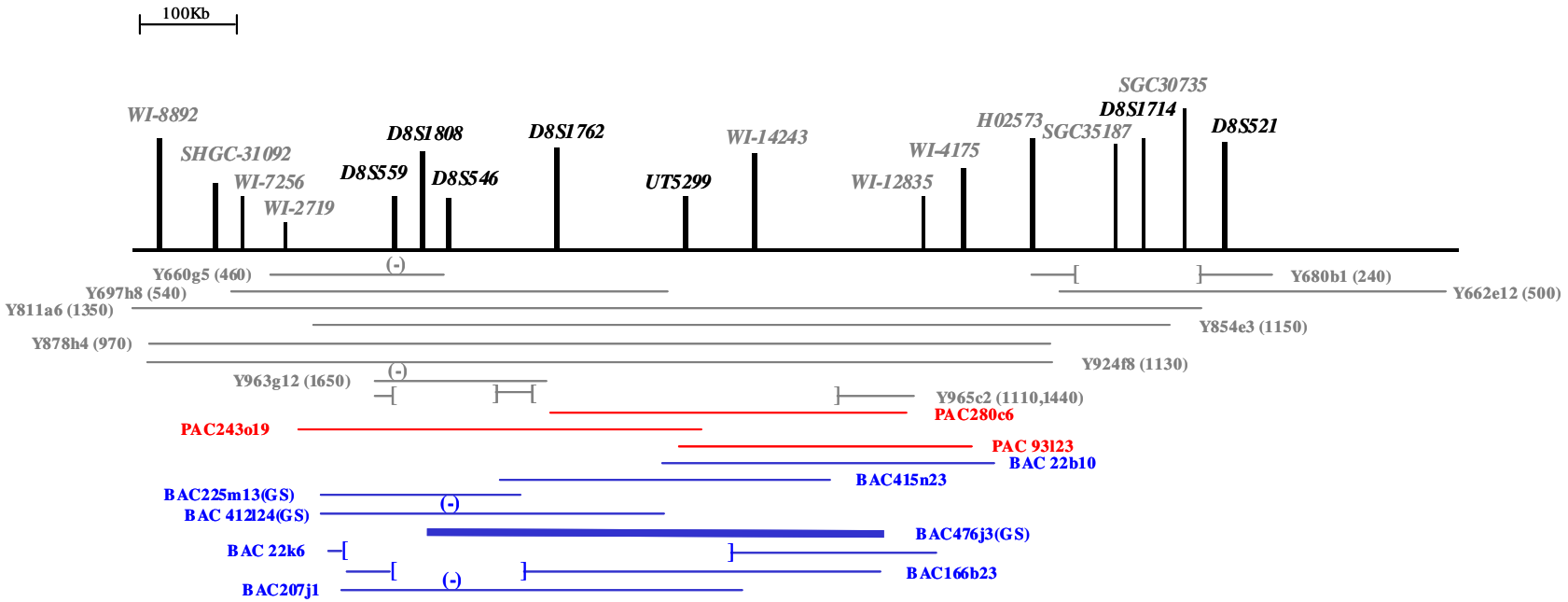


Figure 4. Physical map of the approximately 1.1 Mb genomic region flanked by *WI-8892* and *D8S521*, comprising ten YAC, eight BAC, and three PAC clones. The larger deletions in clones are shown in brackets, and missing single marker sites are in parentheses. The length scale is based on the physical distance between *WI-8892* and *D8S521* as indicated in the UCSC May 2004 assembly. A thick line indicates the BAC clone sequenced by us. Marker *D8S257*, defining the proximal boundary of the true *COH1* region (see Figure 5), resides outside this area ~ 1.8 Mb away, and *COH1* ~ 276 kb proximal to *WI-8892*. The physical map overlaps with the true *COH1* region (see Results and Discussion section 1.4.), which is flanked on the telomeric side by marker *D8S559*. Of the eight transcripts initially screened for mutations in Cohen syndrome patients, four resides within this physical map. Nucleotide changes are shown in table 4.

1.4. Extended haplotype analysis in Finnish Cohen syndrome patients (II, unpublished data)

As the extensive search for a gene defect within the region flanked by markers *D8S1808* and *D8S546* (I, Table 2) did not result in *COH1* gene identification, we extended the search to include the genes surrounding this region. Concomitantly, we typed an extended panel of Finnish patients with novel microsatellite and SNP markers. A rare amino acid-changing nucleotide substitution (c.704G>A; p.Ser235Asn denoted *DORFIN/PvuII*, Table 3) was found in one of the genes under study, the *DORFIN* gene (Niwa et al., 2001), and was present in 29 of 39 Finnish *COH1*-bearing chromosomes (Table 3), but not in 112 Finnish or 196 *CEPH* control chromosomes. The aforementioned 29 affected chromosomes included the majority (28/31) of initially presumed main *COH1* mutation carrying chromosomes and one affected chromosome in a newly diagnosed patient (No. 32 in Table 3). The rarity of the allele at the *DORFIN/PvuII* locus was additional proof for strong LD, which led us to search for and genotype novel single nucleotide polymorphisms (SNPs) more upstream from this marker site. The haplotypes constructed associated chromosome no. 32 with the major haplotype and excluded three chromosomes, which were previously assumed to carry main mutation. This directed our search for *COH1* proximal to *DORFIN* and helped to define a new, approximately 2 Mb, *COH1* region flanked by *D8S257* and *D8S559* (Table 3). We did an extended mutation search in the *DORFIN* gene by exon-by-exon sequencing, but this did not reveal additional changes in Cohen syndrome patients implying that *DORFIN* was unlikely to be the gene for Cohen syndrome. However, we found eight additional polymorphic sites predicting four amino acid changing variants and four silent changes (Table 4).

Table 3. Haplotypes in 39 chromosomes of 20 Finnish Cohen syndrome patients. The haplotypes indicate two possible locations for *COH1*. The initially assumed localization of *COH1* (chromosomes 1-31) is framed. Conserved founder haplotypes refining the true *COH1* critical region based on novel haplotype data and observed *COH1* mutations are shown in blue background color (chromosomes 1-28 and 32). These were observed in 76% of putatively unrelated Finnish Cohen syndrome chromosomes and were all associated with the founder mutation. Conserved haplotypes framed in chromosomes 33-35 share a second mutation. Markers D8S257 and CA-5, with CA-CEN1 located in *COH1* intron 33, flank the *COH1* gene.

	2.03 Mb							0.03 Mb		0.02 Mb						0.06 Mb				
		COH1																		
n:o	D8S257	CA-CEN1	CA-5	SNP-1	SNP-2	DORFIN/PvuII	SNP-6	D8S559	D8S1808	D8S546	SNP-3	SNP-4	D8S1762	SNP-7	SNP-8	SNP-9	SNP-10	SNP-11	SNP-12	CA-2	CA-3
1-18	1	3	1	1	2	1	2	4	2	7	1	2	5	1	1	1	2	1	1	4	3
19	N	3	1	1	2	1	2	4	2	7	1	2	5	1	1	1	2	1	1	4	3
20-25	2	3	1	1	2	1	2	4	2	7	1	2	5	1	1	1	2	1	1	4	3
26	N	3	1	1	2	1	2	4	2	7	1	2	5	1	1	1	2	1	1	4	2
27	1	3	1	1	2	1	2	4	2	7	1	2	5	1	1	1	2	1	1	2	3
28	1	3	1	1	2	1	2	4	2	7	1	2	5	1	1	1	1	1	1	4	3
29	3	6	2	2	2	2	2	6	2	7	2	1	5	1	1	1	2	1	1	4	2
30	4	2	1	2	1	2	1	6	2	7	2	1	5	1	1	1	1	1	1	4	2
31	N	5	2	2	2	2	2	6	3	1	2	1	5	2	2	2	1	2	2	2	2
32	1	3	1	1	2	1	2	6	3	2	2	1	6	2	2	2	1	2	2	2	2
33	2	4	1	2	2	2	2	6	3	2	2	2	1	2	2	2	1	1	1	4	2
34	2	3	2	2	2	2	2	6	3	2	2	2	1	2	2	2	2	2	1	3	2
35	2	3	2	2	2	2	2	6	3	7	2	2	1	1	2	2	2	2	1	3	2
36	5	1	2	2	2	2	2	6	3	2	2	1	6	2	2	2	1	2	2	2	2
37	N	3	1	1	1	2	1	6	2	7	2	2	3	2	2	2	2	1	1	2	5
38	6	5	1	1	1	2	1	6	2	2	2	2	6	2	2	2	1	2	1	4	2
39	3	N	1	1	N	2	2	7	3	1	2	2	3	2	2	2	2	1	1	2	2

1.5. Physical map of the true *COH1* locus (II and unpublished data)

After the genomic region for the *COH1* gene indicated by initial haplotype analysis was thoroughly examined for positional candidate genes and the *COH1* mutation, the novel haplotype data showed a new position for the *COH1* gene. At the time, a relatively good sequence-based physical map was available for this region, and a more exact map was being produced by the public project. The approximately 2 Mb critical interval for *COH1* defined by the extended haplotype analysis (Table 3) contained, in addition to the *COH1* gene, seven genes: six previously reported genes, and one partial transcript (Figure 5).

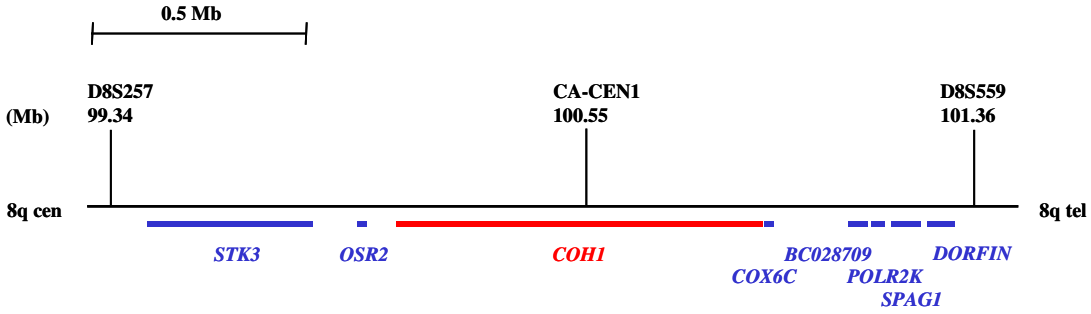


Figure 5. Location of six previously characterized genes (*STK3*, *OSR2*, *COX6C*, *POLR2K*, *SPAG1*, *DORFIN*), and partial mRNA of the gene encoding F-Box protein (*BC028709)* in an approximately 2.0 Mb genomic region flanked by *D8S257* and *D8S559*. These seven transcripts were screened for mutations in Cohen syndrome patients. These genes and the identified nucleotide changes are shown in Table 4. None of these genes were identified as *COH1*.

Table 4. Summary of the fifteen genes studied for mutations in Cohen syndrome patients before the identification of the *COH1* gene.

Gene (mRNA/protein ID) and map position*[1] (Mb)	**References**	**Transcript/protein variant**
DORFIN (101.2-101.3) (*AB029316*, *Q9NV58*)	Niwa et al., 2001; Niwa et al., submitted	**[c.704G>A; p.Ser235Asn]**;**[c.601T>G; p.Tyr201Asp]**; **[c.758A>C;p.His253Pro]**;**[c.1012G>C ; p.Asp338His]**; **[c.1091A>T; p.Gln364Leu]**;[c.105C>T;p.His35His]; [c.666G>A; p.Pro222Pro];[c.843C>T; p.Arg281Arg]; [1083A>C; p.Ile361Ile]
SPAG1 (101.1-101.2) (*AF311312*, *Q07617*)	Zhang et al., 1992; Lin et al., 2001; Lin et al., unpublished	[**c.ins1059_1060GAC; p.Lys353_Ser354insAsp**]; [c.2715G>A; p.Ser905Ser]
Putative F-Box gene (101.1) (*BC028709*)	Strausberg, unpublished	-
STK3 (99.4-99.8) (*U26424*, *Q13188*)	Creasy and Chernoff, 1995; Creasy, unpublished	-
OSR2 (99.9) (*AK074518*, *Q8N2R0*)	Isogai et al., unpublished; Isogai and Otsuki, unpublished	-
COX6C (100.8-100.9) (*BC000187*, *P09669*)	Strausberg et al., 2002	-
POLR2K (101.1) (*BC018157*, *P53803*)	Strausberg et al., 2002	-
HSPG1 (97.5-97.6) (*J04621*, *AAA52701*)	Marynen et al., 1989	**[c.91G>C; p.Ala31Pro]**;**[c.242C>T; p.Thr81Ile]**; **[c.391G>A; p.Ala131Thr]**;**[c.413G>C; p.Arg138Pro]**; **[c.416G>A; p.Arg139Lys]**;**[c.419G>A; p.Arg140Lys]**; [c.426G>A; p.Lys142Lys];**[c.463C>G; p.Leu155Val]**; **[c.799A>T; p.Thr267Ser]**;**[c.763G>A; p.Ala255Thr]**
KCNS2 (99.4) (*AB032970*, *Q9ULS6*)	Hirosawa et al., 1999; Strausberg et al., 2002	-
MATN2 (98.8-99.0) (*BC010444, Q96FT5*)	Deak et al., 1997; Strausberg et al., 2002	**[c.560C>T, p.Thr187Met]**;[c.816G>T, p.Ser272Ser]
OAZIN-1 (103.8) (*BC013420*, *O14977*)	Koguchi et al., 1997; Strausberg et al., 2002	[c.369G>A; p.Ala123Ala]
PRO1097 (99.4) (*AF119844*, *Q9P1A1*)	Zhang et al., unpublished	**[c.179T>C; p.Val60Ala]**
FLJ13955 (99.2-99.3) (*AK024017*, *Q9H841*)	Ota et al., 2004; Isogai and Otshuki, unpublished	[c. 924T>G; p.Thr308Thr]
FLJ36587 (101.5-101.6) (*AK093906*, *Q8N9S9*)	Ota et al., 2004; Suzugi et al., unpublished; Isogai and Yamamoto, unpublished	**[c.745G>T, p. Ala249Ser]**
Putative gene 1 (101.5) (*U79297*)	Andersson et al., 1996; Yu et al., 1997; Yu and Gibbs, unpublished	-

*[1] UCSC Genome Browser July 2003 Assembly (http://genome.ucsc.edu/)

Unpublished references are cited in GenBank.

Nonsynonymous substitutions are marked in bold.

2. The gene for Cohen syndrome (*COH1*)

2.1. Identification of the *COH1* gene (II)

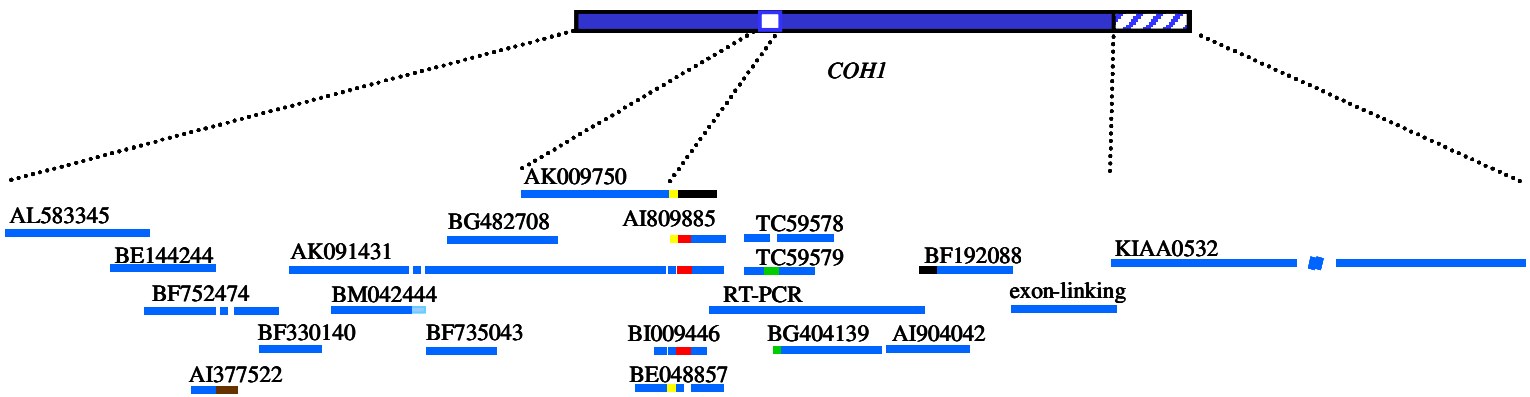
We did not observe *COH1*-associated mutations in genes residing within the novel *COH1* region for which sequences were publicly available in gene databases (Figure 5). Therefore we expanded the search for the disease gene to include murine homologs for human genes by performing blastx searches with the human genomic sequence from the *COH1* region. By this method a partial mouse protein BAB26477 (Genbank accession number *AK009750*) was identified. By comparing this sequence to human genomic sequences using the tblastn program, it was possible for us to predict intron-exon boundaries for the corresponding human gene. The exon sequences defined by the similarity search were used to design primers for two RT-PCR reactions in two Finnish Cohen syndrome patients. One of the two RT-PCR products from these patients showed a two-base pair deletion when sequenced. This mutation predicted a frameshift in the human homolog and was subsequently found to be present in all Cohen syndrome chromosomes bearing the founder haplotype, thus suggesting that this partial transcript sequence belonged to *COH1*. Further analysis revealed several Cohen syndrome-associated mutations in the newly assembled gene (Table 6). The full-length *COH1* transcript sequence was compiled from 18 ESTs, some of them overlapping, one large *KIAA0532* cDNA sequence, and *COH1*-associated *in silico*-predicted exons (Figure 6A). The partial transcripts found by BLAST searches and predicted exons were linked by RT-PCR. The *COH1* transcript of 14,092 nucleotides was found to span an 864 kb genomic region, comprising 62 exons in full-length form (Figure 6B). The transcription start site resides in the second exon and the full-length transcript predicts a 4,022 amino acid protein. The *COH1* transcript is predicted to have, based on ESTs and RT-PCR, at least eight alternative isoforms, three in-frame and five out-of-frame (II, Figure 3, IV). Of six alternatively spliced exons, two (exons 28 and 28b) were present either separately or together in *COH1* transcripts. When they occur separately they result in transcripts with substitution alternative splicing, and when occurring together they introduce length difference alternative splicing. The remaining four alternatively spliced exons resulted in transcripts with length difference alternative splicing. Velayos-Baeza et

al. (2004) suggested the occurrence of tandem duplication in the evolution of exons 28 and 28b (II, Figure 3), spliced by substitution alternative splicing.

2.2. *COH1* gene expression (II)

Multiple-tissue Northern (MTN) blots showed a wide expression pattern for *COH1* in human MTNI, MTNII, and fetal MTN blots (II, Figure 4) with two RT-PCR amplified probes. Three transcripts of ~2, ~5, and ~12-14 kb for *COH1* were observed. Of these, the largest showed significantly stronger signal in prostate, testis, ovary and colon than in other tissues. Interestingly, the brain expression was very low in a blot with adult tissues, while the fetal blot showed equally strong signal for all four tissues (brain, lung, liver, kidney) studied. Controversially, this pattern was not observed with a probe amplified from a cloned cDNA fragment (Velayos-Baeza et al., 2004). This discrepancy was suggested to be a result of nonspecific hybridization with RT-PCR amplified probes in the presence of highly represented RNAs as a similar pattern of bands was also observed with three other human *Vps*13 homologs by RT-PCR. Velayos-Baeza et al. (2004) showed wide expression of different *COH1* isoforms, with the brain showing the strongest signal for the full-length isoform (*AY223814*). This is not strongly expressed in other tissues, in which splice variant 1 (*AY223815*) is the main form. By RNA *in situ* hybridization, Mochida et al. (2004) showed wide expression of *COH1* in postnatal and adult mouse brain but lower expression in embryonic mouse brain, which is in contrast to our results in human MTN blots. Additionally, they showed wide expression of *COH1* in neurons of the central nervous system (CNS) while expression was absent in white matter glial cells (Mochida et al., 2004). Clearly more work is required in order to understand the transcription and developmental expression of *COH1*.

A



B

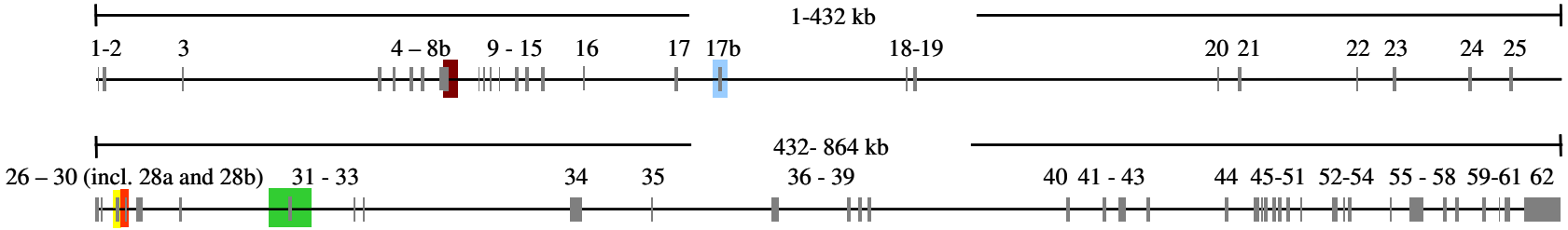


Figure 6. A) Complete *COH1* transcript sequence built by linking together partial transcripts residing within the *COH1* region and using *in silico*-based exon prediction algorithms. Alternative sequence is marked in different colors. Black indicates non-homologous regions in comparison to human sequence. **B)** Genomic structure of the Cohen syndrome gene *COH1* spanning an 864 kb region on chromosome 8q22.3. Five alternatively spliced exons are boxed and marked in transparent colors corresponding to those in Figure 6A. The relative length of exons and introns are to scale, but the introns are not to scale with exons.

3. *COH1* gene mutations

3.1. Overall characteristics of the *COH1* gene mutations (II, III, IV)

During this study 111 patients with a provisional clinical diagnosis of Cohen syndrome were analyzed for mutations in *COH1*. These patients were divided into three groups based on eight clinical criteria (Table 5). Seventy-two of the patients belonged to a "true" Cohen syndrome group fulfilling six or more previously established diagnostic criteria. Twenty-three patients belonged to Cohen-like group fulfilling less than six criteria. A third group of patients with incomplete clinical data consisted of 16 patients.

"True" Cohen syndrome patients were screened for mutations by sequencing except for 27 Finnish patients homozygous for the founder haplotype, who were analyzed by single-stranded conformational polymorphism (SSCP) after the founder mutation was observed in the cDNA of one of them by sequencing. The Cohen-like patients were analyzed by SSCP. Eight patients in the group of incomplete data were sequenced and the rest were studied by SSCP.

To date we have identified 32 different mutations (Table 6 and 7, Figure 7). Six are nonsense changes creating stop codons. Fifteen are small deletions and insertions leading to frameshifts and predicting early protein truncation. Four of the mutations are splice site changes. Of these, two affect splicing of *COH1* cDNA, predicting early protein truncation, while two have not been verified at the cDNA level. Two of the mutations are missense changes. Of the four large exonic deletions two are in-frame changes and two lead to predicted early protein truncation. In addition, two nucleotide changes, a one base pair insertion causing a frameshift and a missense change, were observed in the same allele inherited from both parents in an Amish patient and were shown to segregate in affected family members.

The distribution of mutations over the *COH1* gene (Figure 7) demonstrates that gene defects were seen only in exons present in the full-length form, and mutational hotspots were not observed. Two of the four splice site mutations were seen at the donor and acceptor sites of intron 37. Five of the mutations were seen in exon 34, one of the largest exons in the *COH1* gene. Specific correlations between the mutation site and clinical phenotype were not seen (see Results and Discussion section 3.4.).

In addition to the mutations described by us, others have published 22 additional mutations in *COH1* (Mochida et al., 2004; Hennies et al., 2004). Only one of these, a nonsense mutation (7051 C>T; R2351X) in a French patient (Mochida et al.,2004) was already observed by us. In total, 53 different mutations have been found in *COH1* to date, of which interestingly only five are missense changes. Most (85%) of the *COH1* mutations are protein truncating, probably causing non-functional protein or the transcript to be subjected to nonsense-mediated mRNA decay (NMD).

Table 5. Eight diagnostic criteria for separation of "true" Cohen syndrome patients and those with Cohen like-syndrome.

Diagnostic criteria
Developmental delay
Microcephaly
Typical facial dysmorphism
Obesity and slender extremities
Overly sociable behavior
Joint laxity
Myopia and/or retinal degeneration
Intermittent neutropenia

Table 6. Thirty-two different mutation types observed in *COH1*

Mutation type	**No.***
Nonsense mutations	6 (11)
Missense mutations	2 (3)
Small deletion/insertion/indel/duplication mutations	15 (7)
Large exonic deletion mutations	4 (0)
Splice site mutations	4 (1)
Double mutation (small insertion+missense)	1 (0)
Total number of different mutations	32 (22)

*Mutations published by others (Hennies et al., 2004; Mochida et al., 2004) are shown in parentheses.

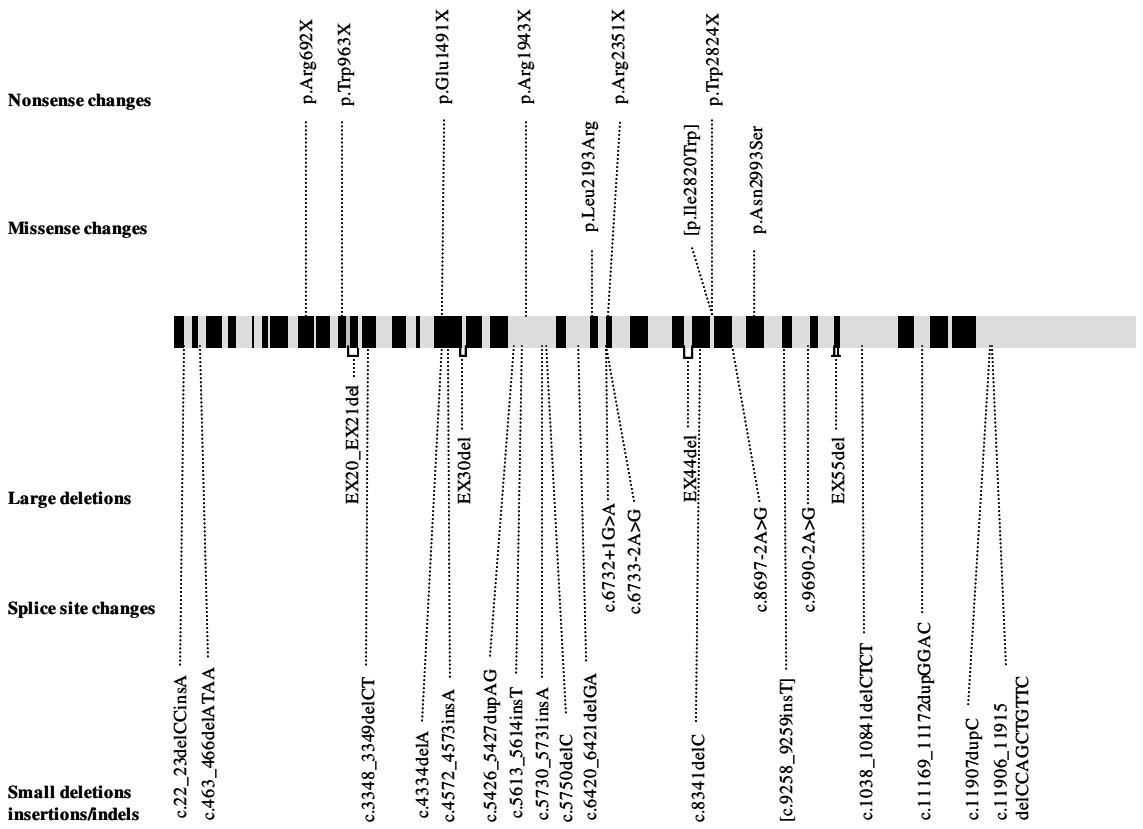


Figure 7. Distribution of mutations in the *COH1* gene cDNA grouped by mutation type. The two changes identified in the same allele in the Amish patient are shown in brackets.

3.2. *COH1* gene mutations in Finland (II, IV)

The Finnish founder mutation is a two nucleotide deletion [c.3348_3349delCT] resulting in a frameshift at codon 1117 and a stop codon after 8 codeshifted amino acids [p.Cys1117fsX8, Table 7]. This mutation is present in all Finnish patients except one. The founder mutation is homozygous in 27 Finnish patients, heterozygous in 12, and is present in 85% (66/78) of the affected chromosomes. The proportion is 73% (31/42) if only one affected sib per family is taken into account. The geographical distribution of the Finnish founder mutation covers a large part of Finland with the exception of Lapland. In addition to the founder mutation, we identified five different "minor" *COH1* mutations (see Table 7) in Finnish patients. This relatively high number of minor mutations is not frequently seen among the diseases of the Finnish disease heritage, as more than three Finnish minor

mutations have been seen only in six other disorders, *i.e.* AGU (4 minor mutations, founder mutation proportion in affected chromosomes 98%), APECED (7, 90%), Batten disease (4, 90%), cartilage-hair hypoplasia (4, 92%), congenital nephrosis (7, 78%), and retinoschisis (6, 70%). The minor mutations in *COH1* comprise one missense and one nonsense mutation, [c.6578T>G; p. Leu2193Arg], [c.5827C>T; p.Arg1943X], two frameshifts, including a small insertion [c.5730_5731insA; p.Ile1913fsX6] and deletion [c.10838_10841delCTCT; p.Leu3614fsX36], and one large in-frame deletion [EX20_21del; p.Gly942_Thr1027del], which is not determined on the genomic level and results in the skipping of two exons in the processed mRNA. Among the minor mutations, only the nonsense mutation [c.5827C>T] was found to be homozygous in one patient. This mutation was found in a family from a small parish in western Finland. In four Finnish patients, of whom two are siblings, only one disease allele was identified despite extensive screening of cDNA and/or genomic DNA. For two of these patients the cDNA was not available, and therefore gross genomic deletions and mutations in intronic or regulatory regions affecting splicing and/or transcript levels may have been missed with PCR-based sequencing of small fragments.

3.3. Consanguinity between Cohen syndrome parents (unpublished)

Consanguinity was verified in six Cohen syndrome sibships. Consanginity was shown in 12 parents carrying the main mutation [c.3348_3349delCT] (Figure 8, P1-3). In the largest pedigree (P2), five parents were descendants of one ancestral couple. As has been observed in other disorders of the Finnish disease heritage, the origin of the founder mutation is probably in south Savo, from where migration has brought it to eastern and northern parts of Finland. The parents of a patient (P4, Figure 8) homozygous for the minor mutation [c.5827C>T] were found to be remotely consanguineous. They and almost all of their ancestors were born in the same region in western Finland. The mother of another Cohen syndrome patient (P5, Figure 8) carried the same rare mutation. Most of the ancestors of this patient were from an entirely different region but one great-grandparent couple was born in the same region as family P4 (Figure 8). Consanguinity between these two ancestries could not be verified. This less common *COH1* founder mutation may have been brought to Finland from the west, or it has originated in western Finland. Due to the presence of the main founder mutation at a high

frequency, less common private mutations can manifest in the phenotype. Minor mutations are also probably present in other populations, but they are not seen because the prerequisite for phenotype expression, the second common founder mutation, is missing. Remote consanguinities between parents as shown here are typical in disorders of the Finnish disease heritage (Norio, 2003b)

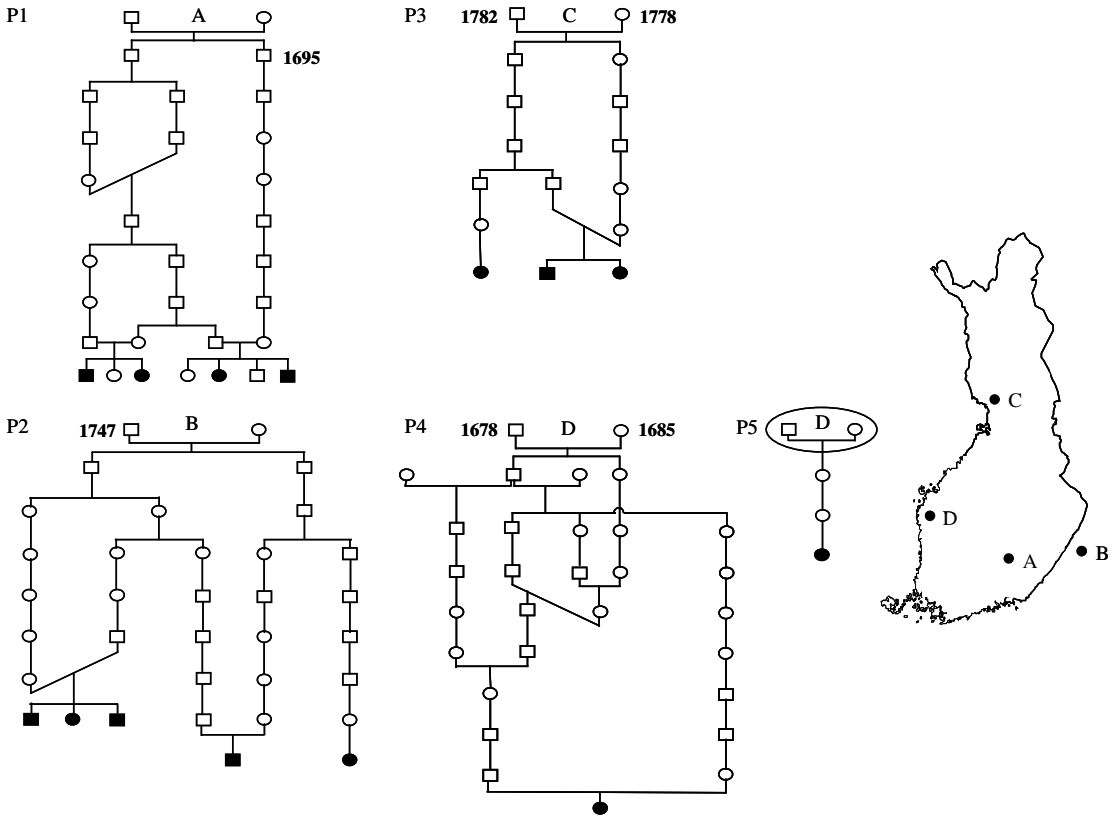


Figure 8. Pedigrees (P1-P5) showing consanguinities between Cohen syndrome parents. The letters A-D in the map of Finland show birth places of the common ancestors of the pedigrees. The Cohen syndrome parents not belonging to the consanguinities are omitted. All Cohen syndrome parents displayed in pedigrees 1-3 carry the main mutation [c.3348_3349delCT], which is homozygous in all patients with the exception of those in pedigree 3. The patients in the two branches of pedigree 3 share the main mutation [c.3348_3349delCT], and have their own private mutations (c.5730_5731insA and c.EX20_21del). The circled great-grandparents in pedigree 5 share the birth place of the common ancestor pair in pedigree 4, the parents shown in these families carry the same less common mutation.

3.4. Definition of Cohen syndrome (IV)

Mutation screening of the three groups totalling 111 patients revealed mutations only in 63 of 72 diagnosed as having Cohen syndrome and in two patients with incomplete data. This allowed us to determine the clinical features associated with molecularly defined Cohen syndrome.

In our study set all patients with an observed mutation in *COH1* had developmental delay (Table 7). Additional clinical features present in almost all patients with few exceptions were typical facial dysmorphism, myopia and/or retinal degeneration, and intermittent neutropenia. There was more variation between patients regarding the other four diagnostic criteria, and our results suggest that these are supportive, but not crucial for Cohen syndrome diagnosis. Thus developmental delay, typical facial dysmorphism, myopia and/or retinal degeneration and intermittent neutropenia can be considered essential features of Cohen syndrome.

Of the four diagnostic criteria important in Cohen syndrome facial features were typical in all patients except the Amish patients who had synophrys, not commonly seen in Cohen syndrome. However, other typical features were present in these patients including thick hair and long eyelashes, wave-shaped palpebral fissures, low nasal bridge, short philtrum and open mouth appearance. There is probably variation in Cohen syndrome facial characteristics due to the ethnic origin of the patients, as has been shown in other studies (Horn et al. 2000). Ophthalmological changes were present in all but three patients, of whom one was under five years of age at the time of the study, and two patients for whom data on ophthalmic changes or clinical confirmation of the ophthalmic changes was missing. Intermittent neutropenia was not reported in one patient and two patients had no data for the neutrophil count.

To date, *COH1* mutation analysis with patients having diagnosed Cohen syndrome and medical reports available has been performed in two studies in addition to ours. Mochida et al. (2004) reported patients of four families with *COH1* mutations with more phenotypic variability than the Cohen syndrome patients we presented. Granulocytopenia was not observed in patients from two families and was not assessed in one. However, blood counts were tested only once, which leaves open the possibility of an incorrect interpretation due to the intermittent nature of granulocytopenia (Kivitie-Kallio et al.,

1997). Ophthalmological studies were missing in two pedigrees, but in two families myopia and either retinal dystrophy or retinitis pigmentosa was reported (Mochida et al., 2004). Hennies et al. (2004) disputed the clinical criteria for Cohen syndrome presented in our study for cohorts having diverse origins, and postulated this to be due to allelic heterogeneity in Cohen syndrome. The only features present in all patients with mutations in *COH1* were developmental delay, early-onset myopia, joint laxity, and facial dysmorphism (Hennies et al., 2004). Their patients lacking retinopathy were young and mostly under five years of age, while retinal changes usually develop at a later age, but all of them had myopia. Granulocytopenia was only observed in about half of their patients despite repeated testing, which might indicate variability in Cohen syndrome concerning this feature.

Table 7. Clinical features used in study IV to differentiate "classical" Cohen syndrome patients from Cohen syndrome-like patients. The clinical phenotypes of 65 Cohen syndrome patients with 32 different mutations are shown.

Patient/ Origin	**Protein change**	**Location***	**Developmental delay**	**Microcephaly (under 2 SD)**	**Typical facial dysmorphism**	**Obesity and slender extremities**	**Overly sociable behavior**	**Joint laxity**	**Myopia and/or retinal degeneration**	**Intermittent neutropenia**
F1 / Finland	[p.Cys1117fsX8]+[p.Arg1943X]	EX23, EX34	+	+	+	+	+	+	+	+
F2 / Finland	[p.Arg1943X]+[p.Arg1943X]	EX34, EX34	+	+	+	+	+	+	+	+
F3 / Finland	[p.Cys1117fsX8]+[p.Ile1913fsX6]	EX23, EX34	+	+	+	+	+	+	+	+
F4 / Finland	[p.Cys1117fsX8]+[p.Leu3614fsX36]	EX23, EX56	+	+	+	-	-	+	+	+
F5 / Finland	[p.Cys1117fsX8]+[p.Gly942_Thr1027del]	EX23, EX20, EX21	+	+	+	+	+	+	+	+
F6 / Finland	[p.Cys1117fsX8]+[p.Gly942_Thr1027del]	EX23, EX20, EX21	+	+	+	+	+	+	+	+
F7 / Finland	[p.Cys1117fsX8]+[p.Leu3614fsX36]	EX23, EX56	+	+	+	-	-	+	+	+
F8 / Finland	[p.Cys1117fsX8]+[p.?]	EX23, ?	+	+	+	+	+	+	+	+
F9 / Finland	[p.Cys1117fsX8]+[p.?]	EX23, ?	+	+	+	+	+	+	+	+
F10 / Finland	[p.Cys1117fsX8]+[p.?]	EX23, ?	+	+	+	-	+	+	+	a

F11 / Finland	[p.Cys1117fsX8]+[p.?]	EX23, ?	+	+	+	-	+	+	+	a
F12 / Finland	[p.Cys1117fsX8]+[Leu2193Arg]	EX23, EX37	+	+	+	-	+	+	+	+
F13 / Finland	[p.Cys1117fsX8]+[Leu2193Arg]	EX23, EX37	+	+	+	-	+	+	+	+
F14-40 / Finland	[p.Cys1117fsX8]+[p.Cys1117fsX8]	EX23, EX23	+	+	+	b	+	+	c	+
UK1 / United Kingdom	[p.Ser3970fsX22]+[p.?]	EX61, ?	+	+	+	-	-	+	+	+
UK2 / United Kingdom	[p.Gln1445fsX7]+[p.?]	EX29, ?	+	+	+	-	+	+	+	+
UK3 / United Kingdom	[p.Val3340fsX9]+[p.?]	EX55, ?	+	-	+	-	+	+	+	+
UK4 / United Kingdom	[p.Leu2674_Gln2725del]+[p.?]	EX44, ?	+	+	+	-	+	+	+	+
UK5 / United Kingdom	[p.Asn156fsX4]+[p.Asn156fsX4]	EX4, EX4	+	+	+	+	+	+	+	+
UK6 / United Kingdom	[p.Glu2900fsX2]+[p.Glu2900fsX2]	IVS46, IVS46	+	-	+	+	+	+	+	+
UK7 / United Kingdom	[p.Glu2900fsX2]+[p.Glu2900fsX2]	IVS46, IVS46	+	+	+	+	-	+	+	+
UK8 / United Kingdom	[p.Trp2824X]+[p.?]	EX46, ?	+	+	+	+	-	+	+	-
UK9 / United Kingdom	[p.Trp963X]+[p.?]	EX19, ?	+	+	+	+	+	-	+	+
UK10 / United Kingdom	[Glu1491X]+[Glu1491X]	EX28, EX28	+	+	+	+	-	+	d	+
UK11 / United Kingdom	[Glu1491X]+[Glu1491X]	EX28, EX28	+	+	+	+	-	+	+	+
UK12 / United Kingdom	[Glu1525fs44X]+[Ala1570fs3X]	EX28, EX28	+	+	+	+	+	-	+	+
NL1 / Holland	[p.Ser1917fsX19]+[p.Arg3230fsX20]	EX33, IVS52	+	+	+	+	+	+	+	+
NL2 / Holland	[p.Ser1873fsX9]+[p.Arg3725fsX7]	EX33, EX57	+	+	+	+	+	+	+	+
D1 / Denmark	[p.Gln2140fsX27]+[p.Leu2781X]	EX36, EX45	+	+	+	+	+	-	+	+
D2 / Denmark	[p.Pro8fsX3]+[p.?]	EX2, ?	+	+	+	+	+	-	+	+
B1 / Belgium	[p.Asn2993Ser]+[p.Asn2993Ser]	EX49, EX49	+	+	+	+	+	+	+	+
B2 / Belgium	[p.Asn2993Ser]+[p.Asn2993Ser]	EX49, EX49	+	+	+	+	+	+	+	+
B3 / Belgium	[p.Arg692X]+[p.Trp1810fs20X]	EX15, EX34	+	+	+	+	+	+	+	+
B4 / Belgium	[p.Arg692X]+[p.Trp1810fs20X]	EX15, EX34	+	+	+	+	+	+	+	+
B5 / Belgium	[p.Arg2351X]+ [p.Arg2351X]	EX39, EX39	+	+	+	+	+	+	+	+
B6 / Belgium	[p.Arg2351X]+ [p.Arg2351X]	EX39, EX39	+	+	+	+	+	+	+	+
US1 / U.S.A.	[p.Leu3087fsX20+p.Ile2820Trp]+ [p. Leu3087fsX20+p.Ile2820Trp]	EX51, EX51	+	+	-	+	+	+	+	+
US2 / U.S.A.	[r.spl?;p.?]+[p.?]	IVS38, ?	+	+	+	a	+	a	+	a
I1 / Israel	[p.Pro3969fsX41]+[r.spl?;p.?]	EX62, IVS38	+	+	+	+	+	a	a	a

a) no data available

b) present in 17% of patients

c) retinal degeneration present in all but one patient under five years of age

d) no data

* prefix EX denotes exon and IVS intron

Double line frames and gray background color indicate the siblings.

4. Predicted characteristics of the COH1 protein (II and unpublished)

4.1. Complex structure of the COH1 protein

The hydrophobic and alpha helical amino acid sequence regions in COH1 indicate the protein has 10 transmembrane (TM) domains. This was supported by data from the Gene Ontology Annotation (GOA) database (Camon et al., 2004) available from Entrez Gene at NCBI. Similarity searches in the protein domain family database ProDom (Gouzy et al., 1999) revealed that the COH1 N- and C-terminal sequences are homologous to *S. cerevisiae* vacuolar protein sorting protein 13 (Vps13). In addition, the COH1 protein has a complex pattern of predicted functional motifs with signals for both endoplasmic reticulum (ER) retention (Teasdale and Jackson, 1996; aa 4018-4021) and peroxisomal targeting (McNew and Goodman, 1996; aa 263-271 and aa 3553-3561). Amino acid segments for a leucine zipper pattern (Landschulz et al., 1989; aa 30-51) and zinc metallopeptidase (Jongeneel et al., 1989; aa 3289-3298) were also detected. The participation of COH1 in vacuolar-targeted sorting of protein transfer was further supported when a possible vacuolar targeting motif (Stack et al., 1995) was found at amino acid positions 1518-1521.

The gene structure of *COH1* is similar to another human disease-causing gene, *CHAC* (Rampoldi et al., 2001; Ueno et al., 2001). The defect in *CHAC* leads to Choreoacanthosytosis (Critchley et al., 1968), a disorder with neuronal and haematopoietic involvement. The encoded chorein protein has moderate overall sequence homology to the COH1 protein. COH1 and chorein probably belong to a novel family of Vps13-like human proteins. The COH1, chorein and *S. cerevisiae* Vps13 all have conserved N- and C-terminal domain structure (II; Rampoldi et al., 2001; Ueno et al., 2001). Both *COH1* and *CHAC* encode large proteins, 4,022 and 3,174-amino acids respectively, and have on average small exons (62 and 72 exons). Both genes span a large genomic region (864 kb and 220 kb), which is four times larger for *COH1* (II; Rampoldi et al., 2001; Ueno et al., 2001).

4.2. ER retention signal in COH1 protein

The COH1 protein possesses a carboxy-terminal dibasic motif, a KKXX-like motif RKGF at the -5 position, in which one of the lysine residues is substituted with arginine. In this

motif a lysine or arginine residue is positioned at -3, -4, or -5 from the C-terminus, and the -4 lysine could be moved to the -5 position without losing function (Jackson et al., 1990, Shin et al., 1991). Although a conserved di-lysine motif is present in many ER proteins, some substitutions of lysine by arginine are permitted (Teasdale and Jackson, 1996). This dibasic motif is a typical signal for membrane proteins mediating sorting between the ER and the Golgi apparatus (Teasdale and Jackson, 1996). Di-lysine motifs bind complexes of cytosolic coat protein, COP1, and work in a secretory pathway to retrieve proteins from the Golgi to the ER (Teasdale and Jackson, 1996). The importance of the ER retention signal in Cohen syndrome and possible functional divergence of the COH1 protein is unclear. However, proteins interacting with the adaptor protein 3 (AP3) tetramer complex during protein sorting have ER retention function in addition to endosomal/lysosomal targeting. The cargo protein, neutrophil elastase, for the AP3 complex has an amino-terminus ER retention signal (XXRR), and one of the AP3 complex subunits, AP3B1, a vacuolar targeting motif. The AP3 complex is known to be involved in routing of proteins between *trans*-Golgi and lysosomes (Obermüller et al., 2002). Interestingly, a mutation in *AP3B1* encoding the β subunit of AP3, dissociates the tetramer leading to mislocalization of the cargo protein. Mutations in both the *ELA2* and *AP3B1* genes cause cyclic neutropenia in dogs (Ozsoylu, 2001; Dell'Angelica et al., 1998), and mutation of the *AP3B1* gene is causative in human Hermansky Pudlak syndrome type 2 (HPS2), which also involves retinal hypopigmentation. It has been suggested that mistargeting of neutrophil elastase, encoded by *ELA2*, also causes intermittent neutropenia in Cohen syndrome (Horwitz et al., 2004).

4.3. Rodent *COH1* orthologs

The predicted mouse homolog for the COH1 protein is comprised of 3,965 amino acids, and the partial transcript consists of 11,986 nucleotides. The gene has 62 exons, spanning ~558 kb has a translation starts from nucleotide 29 in exon two. Mouse *Coh1* has 85.5% overall mRNA similarity to human *COH1*. *In silico* protein analysis shows the signal sequences to be conserved in comparison to human COH1 protein, and the hydrophobicity profile predicts seven transmembrane domains.

The predicted 12,271 nucleotide rat *Coh1* spanning ~624 kb has a translation start site in exon two, 62 exons, and encodes a 3,995 amino acid protein. It has functional domains similar to the human ortholog. Exceptionally, the rat protein has a third peroxisomal targeting signal 2 (PTS2) signature domain at amino acid 2,465 and eight predicted transmembrane domains. Additionally, the first untranslated exon in mouse and rat included in ESTs BB596466 and CB691391, respectively, differs from the human *COH1* transcript. The similarity of the mouse and rat COH1 proteins to human COH1 protein isoform 1 is 87.5% and 86.7% respectively.

4.4. *COH1* promoter region (unpublished)

The CpGProD program (Ponger and Mouchiroud, 2001), designed to predict promoters associated with CpG islands, was used in the prediction of a potential transcription regulating region 5´-upstream of the *COH1* coding sequence. In humans a GC-rich promoter region was observed 1,310 nucleotides upstream (and 185 nucleotides downstream from the *COH1* transcription start codon), continuing 1,498 nucleotides downstream. The strength of prediction was moderate with a start-p value of 0.31. It is known that 50%-60% of human genes display a CpG island over the transcription start codon (Ponger and Mouchiroud, 2001). In the corresponding mouse and rat genome sequences, good start-p values of 0.57 and 0.53 for the GC-rich promoter region was predicted starting 1162 and 1179 nucleotides upstream from the putative *COH1* transcription starting codon.

Three conserved putative transcription factor (TF) binding sites (ELK1, ETS1P54, and NRF2) in human, mouse and rat were observed with the MATCH program (http://www.gene-regulation.com/pub/programs.html#match). These overlap in the genomic region upstream of the *COH1* transcription start codon (Figure 9). *ELK1* (Rao et al., 1989) belongs to the *ETS* gene family (Watson et al., 1988), of which the ETS1 family comprises proteins encoding nuclear phosphoproteins (Li et al., 2000). The ETS proteins are transcription factors that interact with purine rich promoter/enhancer region sequences of genes (Karim et al., 1990). Characteristically, proteins that have an ETS transcription factor do not possess typical transcription regulating “TATA” or “GAAT” elements (Mavrothalassitis et al., 1990; Jorcyk et al., 1991), which holds true also for the *COH1* gene

promoter region. In general, ETS proteins are associated with processes of cell growth control, embryological development and hematopoietic differentiation (Hromas and Klemsz, 1994; Seth et al., 1992; Crepieux et al., 1994).

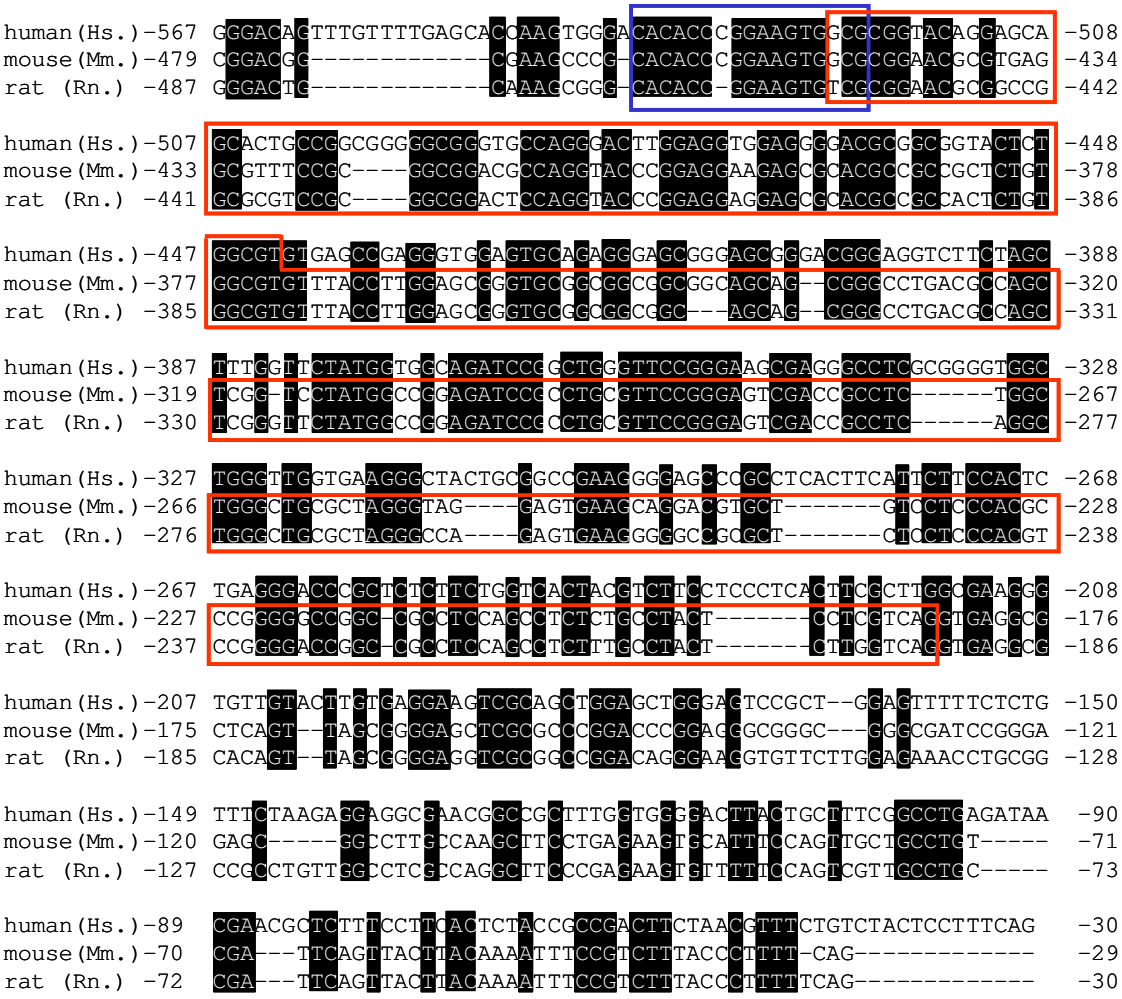


Figure 9. Predicted *COH1* promoter region and conserved transcription factor binding sites (boxed in blue) within the *COH1* 5´-upstream regulatory region in human, mouse and rat. The first non-coding exon is boxed in red.

5. COH1 function in respect of diseases involving trans-Golgi protein sorting

The role of COH1 in cellular events can only be hypothesized since experimental evidence for the protein function is not available. Interestingly, we observed homology of COH1 to *S. cerevisiae* Vps13. Vps proteins have several functions in secretory pathway protein sorting and vacuolar biogenesis in eukaryotes. The yeast vacuole and mammalian lysosome share many features and are thus believed to be functionally equivalent compartments, and derived from a branch of the secretory pathway. Vps proteins participate in several steps in the retrieval of proteins from the post-Golgi compartment. Vps-associated proteins in yeast are known to have a receptor function, and are involved in segregation and packaging of proteins into vesicular carriers. Vps proteins participate in vesicle formation and membrane fusion, are vesicle coat proteins, take part in targeting and recognition of the carriers, recycle receptors, and have a role in maintenance of vacuolar structure (Stack et al., 1995). Since COH1 has sequence homology to yeast Vps13 protein, we can assume the involvement of these proteins in similar processes in the cell. Four human proteins (COH1, chorein, VPS13C, and VPS13D) have been characterized to date that have N- and C-terminal homology to Vps13, and all except COH1 also have a third homologous region and a repeated 45 amino acid core element (Velayos-Baeza et al., 2004). These proteins form a new Vps13-like protein family (Ramboldi et al., 2001; Ueno et al., 2001; Velayos-Baeza et al., 2004). Of these proteins, COH1 and chorein, underlying choreoacanthocytosis, are known to be disease-causing if defective. Choreoacanthocytosis is characterized by progressive neurodegeneration and red cell acanthocytosis (Rubio et al., 1997). Interestingly, the disease phenotype in both Cohen syndrome and choreoacanthocytosis includes hematological aberration and neurologic abnormality.

COH1 predicted function as well as Cohen syndrome clinical features share some similarity to other diseases involved in protein sorting from the trans-Golgi network, in particular hypopigmentation disorders which also involve retinitis pigmentosa and neutropenia, essential features in Cohen syndrome. Many of the proteins underlying hypopigmentation disorders are known to participate in protein sorting between the trans-Golgi and lysosomes. Interestingly, there is growing evidence of the involvement of intracellular protein sorting and trafficking in disorders of the CNS. For example, disorders associated with congenital microcephaly and mental retardation are caused by mis-targeting

in apical sorting of proteins in neuronal progenitor cells (Sheen et al, 2004; Bond et al., 2003). However, the pathomechanism for aberrant brain development in these is probably different than that in Cohen syndrome, as based on the RNA *in situ* expression profile, the COH1 protein functions primarily in postmitotic cells and not in neuronal progenitor cells (Mochida et al., 2004).

CONCLUDING REMARKS AND FUTURE PROSPECTS

The clinical definition of Cohen syndrome and its diagnostic criteria has been disputed within the scientific community. Our results indicated Cohen syndrome to be a clinical entity, and excluded most features of what has been thought to represent phenotype heterogeneity. The results also emphasized the importance of careful ophthalmologic studies and repeated determination of the leukocyte count before diagnosing Cohen syndrome. The study also confirmed the age-dependency of retinal changes and the possibility of ethnic variation in facial features in Cohen syndrome. The observations described here will facilitate the diagnosis of Cohen syndrome and are of use in genetic counseling.

A wide variety of mutations were found in the *COH1* gene. Most of the mutations are protein truncating and the spectrum of mutations covers the whole gene. Due to the number of mutations and the large size of the gene, implementing a comprehensive diagnostic test is in practice impossible. However, the high proportion of affected chromosomes carrying a founder mutation in Finland provides a possibility for a diagnostic gene test. Diagnosis by minisequencing is indeed already available for this mutation. The currently available test is appropriate mainly for Finnish Cohen syndrome patients, since a world-wide founder origin has not been observed. The majority of observed mutations in *COH1* predict protein truncation with possible degradation already at the mRNA level, which might allow diagnostic approaches based on indications of reduced mRNA/protein levels.

Homology searches with the predicted COH1 protein sequence revealed that the *COH1* gene encodes a novel protein belonging to the group of vacuolar protein sorting-associated proteins. The significance of several other predicted functional domains needs experimental proof. Future studies of the COH1 protein might be based on the hypothesis that it has a function in vacuolar protein sorting.

Various steps in protein trafficking between the *trans*-Golgi network and lysosomes are affected in a group of hypopigmentation disorders. The involvement of a retinal defect as well as failure of granulocyte development connect these disorders and prompt the suggestion that a defect in COH1 affects granule structure-containing cell types. The

question of whether the protein-sorting machinery is also affected in Cohen syndrome should be unraveled in future protein studies.

The basic issues for future studies are the delineation of *COH1* alternative mRNAs and experimental clarification of the COH1 protein structure. Important objectives for future studies are also the determination of *COH1* expression regulation mechanisms, description of COH1 protein interactions, clarification of the subcellular localization of the COH1 protein, and determination of its involvement in developmental patterns. Putting this together, the clarification of COH1 protein function will help us to understand this disorder better, and might open alternatives for treatment. The study presented here serves as a basis for future research.

ACKNOWLEDGEMENTS

This work was done at the Folkhälsan Institute of Genetics, Department of Medical Genetics and Neuroscience Center, University of Helsinki, during the years 1996-2004. I am grateful to Albert de la Chapelle, Juha Kere, Pertti Aula, Leena Palotie, Anna-Elina Lehesjoki, Kristiina Aittomäki and Päivi Peltomäki, former and current Professors of the Department of Medical Genetics, and Per Henrik Groop, Research Director of the Folkhälsan Research Center, for their efforts to create an innovative atmosphere for high-class research.

I want to thank my supervisors Albert de la Chapelle and Anna-Elina Lehesjoki. Albert, thank you for putting me on the "payroll" back in 1995 at the Department of Medical Genetics, and offering the thesis project. I will remember with gratitude your support and inspiring discussions whenever we had a chance to discuss the project in person. Anna-Elina, I am thankful for many theoretical discussions about the current status of the ongoing study.

I want to thank Reijo Norio, who discovered and described Cohen syndrome in Finland, for good advice and comments on clinical and genealogical subjects. It can be safely said that without you the whole project surely wouldn't have been possible. Thank you for being a very emphatic and supportive person at times when the *COH1* 'genehunt' was not very feasible. I also want to thank Satu Kivitie-Kallio, who finished her thesis on Cohen syndrome, and studied a large set of Cohen syndrome patients. This has been very much a foundation of the present study.

I would like to direct my many thanks to our main collaborators, Kate Chandler, Graeme C. M. Black, and Forbes Manson at the Department of Clinical Genetics at St. Mary's Hospital in Manchester, who have been extremely helpful in studying a large set of British patients for *COH1* mutations, and who have been of crucial help in publishing two of the studies. I thank all the co-authors of this work for their personal contribution by giving valuable patient samples to our study.

I thank the reviewers of this thesis, Marjo Kestilä and Pentti Tienari for rational comments to improve the manuscript. I special thank to Jodie Painter for revising the English language in this thesis.

My warmest thanks to the following persons working with me in the Cohen project at different time points: Esa Tahvanainen, who introduced me to the Department of Medical Genetics, and made excellent work in the gross localization of the *COH1* gene, which offered a good premise for future studies. Ann-Liz Träskelin, for useful help in the lab, and also by giving a helping hand ('eye') in analysis of the piling up sequence information. Also, thank you for being a 'firm' organizer in the lab, and a good friend. Anne Saarinen, thank you for keeping all the records updated, and for your straightforward answers to my many questions concerning the on-going project. I am also grateful to Laura Waris, Mervi Kuronen, Nina Halla and Saara Tegelberg, who during different time periods were of precious help in the lab. A large part of this thesis work has involved sequencing. I would like to extend my warmest thanks to people who have contributed to the project at the Haartman Institute sequencing unit, Paula Kristo and Elvi Karila. I direct my sincere regards to Jinmin Miao, who was one of the very first persons I got acquainted within the laboratory at the time when I was a newcomer, many years ago. Thank you for assisting me and taking good care of the valuable Cohen syndrome cell lines. I thank Sinikka Lindh for collecting many/all of the Cohen syndrome family specimens during your trips to provinces of Finland. Warm thanks to Aila Riikonen and Minna Maunula who have ensured that the day-to-day businesses have gone smoothly.

I thank my "room mates" in Biomedicum, Liina Lonka, Kirsi Alakurtti, Tarja Salonen and Eija Siintola, for sharing their experiences with me and for making the atmosphere in the office from time to time very social, argumentative, and supportive. I would like to thank my colleague and good friend Peter Hackman, with whom I have had many precious discussions of both the items concerning work and turns of life in general. Recently, I have had the pleasure to get to know Bjarne Udd, who has made science an interesting topic during our lunch breaks. Ulla Lahtinen, thank you for the creative discussions about science and for giving me an excuse to visit VTT. All other persons working at FIG, thank you for your friendship and support, and for creating a memorable atmosphere in the lab.

During the years we have had many reasons to celebrate. A few names that particularly pop up are Kimmo Virtaneva, Kristiina Avela, Maria Aminoff, Tarja Joensuu, Maaret Ridanpää, Susanna Ranta, José Dieguez, Laura Huopaniemi, Elena D'Amato, Kati

Donner and Riikka Hämäläinen. It was nice to share the same office space with some of you and to share many moments to celebrate.

The Ulla Hjelt Fund of the Finnish Foundation for Pediatric Research, Finnish Cultural Foundation, the Finnish Medical Society Duodecim, the Maud Kuistila Memorial Foundation, the Centre of Excellence in Disease Genetics of the Academy of Finland, and the Folkhälsan Research Foundation have financially supported this work.

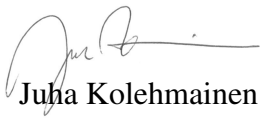
I have been blessed with a small, but good bunch of friends, with whom the friendships go back for many years. Thank you for many pleasant moments in off-work activities to "release fumes" on badminton and tennis courts.

I also thank the four-legged family members, dwarf rabbits Räpsy and Puppe (deceased) and lately dwarf pincher Tino, for giving much joy to my life.

To my mother I would like to present my deepest feelings and warm thanks for giving your full support to the different "projects" of my life. You have given me the prerequisites to a successful life with the encouraging and giving attitude of yours.

Finally, I would like to thank with love Katarina for all those good moments spent together now and in the future, as well as for your patience and tolerance toward my mood changes during the project. Your support, empathy, feet-on-the-ground attitude, and intelligence, in everyday and work related matters, have been an incredible help for me. I thank Niko for 'tough' *in silico* games over keyboards and exciting moments to catch the 'big fish' during summers.

Helsinki, December 2004



Juha Kolehmainen

REFERENCES

Aaij C, Borst P. (1972). The gel electrophoresis of DNA. Biochim Biophys Acta 269, 192-200.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno R, Kerlavage AR, McCombie WR, Venter JC. (1991). Complementary DNA sequencing: "expressed sequence tags" and the human genome project. Science, 252, 1651-1656.

Alström CH, Hallgren B, Nilsson LB, Asander H. (1959). Retinal degeneration combined with obesity, diabetes mellitus and neurogenous deafness: a specific syndrome (not hitherto described) distinct from the Laurence-Moon-Biedl syndrome. A clinical endocrinological and genetic examination based on a large pedigree. Acta Psychiat Neurol Scand 34 (suppl. 129), 1-35.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. J Mol Biol 215, 403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

Altschul SF, Koonin EV. (1998). Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. Trends Biochem Sci 23, 444-447.

Alves AM, Carr FJ. (1988). Dot blot detection of point mutations with adjacently hybridising synthetic oligonucleotide probes. Nucleic Acids Res 16, 8723.

Andersson B, Wentland MA, Ricafrente JY, Liu W, Gibbs RA. (1996). A "double adaptor" method for improved shotgun library construction. Anal Biochem 236, 107-113.

Angelman, H. (1965). 'Puppet children': a report of three cases. Dev Med Child Neurol 7, 681-688.

Ansley SJ, Badano JL, Blacque OE, Hill J, Hoskins BE, Leitch CC, Kim JC, Ross AJ, Eichers ER, Teslovich TM, Mah AK, Johnsen RC, Cavender JC, Lewis RA, Leroux MR, Beales PL, Katsanis

N. (2003). Basal body dysfunction is a likely cause of pleiotropic Bardet-Biedl syndrome. Nature 425, 628-633.

Antequera F. (2003) Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60, 1647-1658.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM; InterPro Consortium. (2000). InterPro—an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16, 1145-1150.

Arber W. (1974). DNA modification and restriction. Prog Nucleic Acid Res Mol Biol 14, 1-37.

Badano JL, Ansley SJ, Leitch CC, Lewis RA, Lupski JR, Katsanis N. (2003). Identification of a novel Bardet-Biedl syndrome protein, BBS7, that shares structural features with BBS1 and BBS2. Am J Hum Genet 72, 650-658.

Balestrazzi P, Corrini L, Villani G, Bolla MP, Casa F, Bernasconi S. (1980). The Cohen syndrome: clinical and endocrinological studies of two new cases. J Med Genet 17, 430-432.

Barany F. (1991). Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Natl Acad Sci USA 88, 189-193.

Bardet G. (1920). Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire). Thesis: Paris, 1920. Note: No. 479.

Barton GJ. (1994). Scop: structural classification of proteins. Trends Biochem Sci 19, 554-555.

Bassam BJ, Caetano-Anolles G, Gresshoff PM. (1991). Fast and sensitive silver staining of DNA in polyacrylamide gels. Anal Biochem 196, 80-83.

Beuren AJ. (1972). Supravalvular aortic stenosis: a complex syndrome with and without mental retardation. Natl Found March Dimes Birth Defects Orig Art Ser 8, 45-56.

Biedl A. (1922). Ein Geschwisterpaar mit adiposo-genitaler Dystrophie. Dtsch Med Wschr 48, 1630.

Bird AP. (1986). CpG-rich islands and the function of DNA methylation. Nature 321, 209-213.

Bishop DT, Crockford GP. (1992). Comparisons of radiation hybrid mapping and linkage mapping. Cytogenet Cell Genet 59, 93-95.

Boehnke M, Lange K, Cox DR. (1991). Statistical methods for multipoint radiation hybrid mapping. Am J Hum Genet 49, 1174-1188.

Boehnke M. (1992). Radiation hybrid mapping by minimization of the number of obligate chromosome breaks. Cytogenet Cell Genet 59, 96-98.

Boguski MS, Schuler GD. (1995). ESTablishing a human transcript map. Nat Genet 10, 369-371.

Bond J, Scott S, Hampshire DJ, Springell K, Corry P, Abramowicz MJ, Mochida GH, Hennekam RC, Maher ER, Fryns JP, Alswaid A, Jafri H, Rashid Y, Mubaidin A, Walsh CA, Roberts E, Woods CG. (2003). Protein-truncating mutations in ASPM cause variable reduction in brain size. Am J Hum Genet 73, 1170-1177.

Botstein D, White RL, Skolnick M, Davis RW. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32, 314-331.

Bower BD, Jeavons PM. (1967). The "happy puppet" syndrome. Arch Dis Child 223, 298-302.

Brown WR, Bird AP. (1986). Long-range restriction site mapping of mammalian genomic DNA. Nature 322, 477-481.

Burghes AHM, Vaessin HEF, de la Chapelle A. (2001). The land between mendelian and multifactorial inheritance. Science 293, 2213-2214.

Burke DT, Carle GF, Olson MV. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236, 806-812.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32, D262-D266.

Carey JC, Hall BD. (1978). Confirmation of the Cohen syndrome. J Pediatr 93, 239-244.

Chandler KE, Clayton-Smith J. (2002). Does a Jewish type of Cohen syndrome truly exist? Am J Med Genet 111, 453-454.

Chandler KE, Kidd A, Al-Gazali L, Kolehmainen J, Lehesjoki AE, Black GC, Clayton-Smith J. (2003). Diagnostic criteria, clinical characteristics, and natural history of Cohen syndrome. J Med Genet 40, 233-241.

Charles SJ, Moore AT, Yates JR, Green T, Clark P. (1990). Alstrom's syndrome: further evidence of autosomal recessive inheritance and endocrinological dysfunction. J Med Genet 27, 590-592.

Chumakov IM, Rigault P, Le Gall I, Bellanné-Chantelot C, Billault A, Guillou S, Soularue P, Guasconi G, Poullier E, Gros I, Belova M, Sambucy J-L, Susini L, Gervy P, Gilbert F, Beaufils S, Bui H, Massart C, De T and M-F, Dukasz F, Lecoulant S, Ougen P, Perrot V, Saumier M, Soravito C, Bahouayila R, Cohen-Akenine A, Barillot E, et al. (1995). A YAC contig map of the human genome. Nature 377, 175-297.

Cohen D, Chumakov I, Weissenbach J. (1993). A first-generation physical map of the human genome. Nature 366, 698-701.

Cohen MM, Jr, Hall BD, Smith DW, Graham CB, Lampert KJ. (1973). A new syndrome with hypotonia, obesity, mental deficiency, and facial, oral, ocular and limb anomalies. J Pediatr 83, 280-284.

Collin GB, Marshall JD, Ikeda A, So WV, Russell-Eggitt I, Maffei P, Beck S, Boerkoel CF, Sicolo N, Martin M, Nishina PM, Naggert JK. (2002). Mutations in *ALMS1* cause obesity, type 2 diabetes and neurosensory degeneration in Alstrom syndrome. Nat Genet 31, 74-78.

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

Collins FS. (1995). Positional cloning moves from perditional to traditional. Nat Genet 9, 347-350.

Collins FS, Morgan M, Patrinos A. (2003). The Human Genome Project: lessons from large-scale biology. Science 300, 286-290.

Corpet F, Servant F, Gouzy J, Kahn D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res 28, 267-269.

Cox DR, Burmeister M, Price ER, Kim S, Myers RM. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science 250, 245-250.

Creasy CL, Chernoff J. (1995). Cloning and characterization of a human protein kinase with homology to Ste20. J Biol Chem 270, 21695-21700.

Creel DJ, Bendel CM, Wiesner GL, Wirtschafter JD, Arthur DC, King RA. (1986). Abnormalities of the central visual pathways in Prader-Willi syndrome associated with hypopigmentation. N Engl J Med 314, 1606-1609.

Crepieux P, Coll J, Stehelin D. (1994). The Ets family of proteins: weak modulators of gene expression in quest for transcriptional partners. Crit Rev Oncog 5, 615-638.

Critchley EM, Clark DB, Wikler A. (1968). Acanthocytosis and neurological disorder without betalipoproteinemia. Arch Neurol 18, 134-140.

Crollius HR, Jaillon O, Bernot A, Pelletier E, Dasilva C, Bouneau L, Burge C, Yeh RF, Quetier F, Saurin W, Weissenbach J. (2002). Genome-wide comparisons between human and tetraodon. Ernst Schering Res Found Workshop 36, 11-29.

Deak F, Piecha D, Bachrati C, Paulsson M, Kiss I. (1997). Primary structure and expression of matrilin-2, the closest relative of cartilage matrix protein within the von Willebrand factor type A-like module superfamily. J Biol Chem 272, 9268-9274.

de la Chapelle A. (1993). Disease gene mapping in isolated human populations: the example of Finland. J Med Genet 10, 857-865.

de la Chapelle A, Wright FA. (1998). Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc Natl Acad Sci USA 95, 12416-12423.

Dell'Angelica EC, Klumperman J, Stoorvogel W, Bonifacino JS. (1998). Association of the AP-3 adaptor complex with clathrin. Science 280, 431-434.

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau MT, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJR, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Fox S, Gelling S, Green L, Harrison P et al. (1998). A physical map of 30,000 human genes. Science 282, 744-746.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, and Weissenbach J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380, 152 -154.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. (1998). Rates of spontaneous mutation. Genetics 148, 1667-1686.

Duyk GM, Kim SW, Myers RM, Cox DR. (1990). Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. Proc Natl Acad Sci USA 87, 8995-8999.

Ewart, AK, Morris CA, Atkinson D, Jin W, Sternes K, Spallone P, Stock AD, Leppert M, Keating, MT. (1993). Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. Nat Genet 5, 11-16.

Ewing B, Hillier L, Wendl MC, Green P. (1998a). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8, 175-185.

Ewing B, Green P. (1998b). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8, 186-194.

Fan Y, Esmail MA, Ansley SJ, Blacque OE, Boroevich K, Ross AJ, Moore SJ, Badano JL, May-Simera H, Compton DS, Green JS, Lewis RA, Van Haelst MM, Parfrey PS, Baillie DL, Beales PL, Katsanis N, Davidson WS, Leroux MR. (2004). Mutations in a member of the Ras superfamily of small GTP-binding proteins causes Bardet-Biedl syndrome. Nat Genet 36, 989-993.

Fischer SG, Lerman LS. (1983). DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. Proc Natl Acad Sci USA 80, 1579-1583.

French SW, Malone CS, Shen RR, Renard M, Henson SE, Miner MD, Wall R, Teitell MA. (2003). Sp1 transactivation of the TCL1 oncogene. J Biol Chem 278, 948-955.

Fryns JP, Legius E, Devriendt K, Meire F, Standaert L, Baten E, Van den Berghe H. (1996). Cohen syndrome: the clinical symptoms and stigmata at a young age. Clin Genet 49, 237-241.

Gardiner-Garden M, Frommer M. (1987). CpG islands in vertebrate genomes. J Mol Biol 196, 261-282.

Gelfand MS, Mironov AA, Pevzner PA. (1996). Gene recognition via spliced sequence alignment. Proc Natl Acad Sci USA 93, 9061-9066.

Geourjon C, Deleage G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 11, 681-684.

Ghazi H, Gonzales FA, Jones PA. (1992). Methylation of CpG-island-containing genes in human sperm, fetal and adult tissues. Gene 114, 203-210.

Goecke T, Majewski F, Kauther KD, Sterzel U. (1982). Mental retardation, hypotonia, obesity, ocular, facial, dental, and limb abnormalities (Cohen syndrome). Report of three patients. Eur J Pediatr 138, 338-340.

Gordon D, Abajian C, Green P. (1998). Consed: a graphical tool for sequence finishing. Genome Res 8, 195-202.

Goss SJ, Harris H. (1975). New method for mapping genes in human chromosomes. Nature 255, 680-684.

Gouzy J, Corpet F, Kahn D. (1999). Whole genome protein domain analysis using a new method for domain clustering. Comput Chem 23, 333-340.

Grabe HJ, Moller B, Willert C, Spitzer C, Rizos T, Freyberger HJ. (2004). Interhemispheric transfer in alexithymia: a transcallosal inhibition study. Psychother Psychosom 73, 117-123.

Green ED, Olson MV. (1990). Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proc Natl Acad Sci USA 87, 1213-1217.

Gress TM, Muller-Pillasch F, Geng M, Zimmerhackl F, Zehetner G, Friess H, Buchler M, Adler G, Lehrach H. (1996). A pancreatic cancer-specific expression profile. Oncogene 13, 1819-1830.

Grimm T, Wesselhoeft H. (1980). Zur Genetik des Williams-Beuren-Syndroms und der isolierten Form der supravalvulaeren Aortenstenose (Untersuchungen von 128 Familien). Z Kardiol 69, 168-172.

Gunay-Aygun M, Schwartz S, Heeger S, O'Riordan MA, Cassidy SB. (2001). The changing purpose of Prader-Willi syndrome clinical diagnostic criteria and proposed revised criteria. Pediatrics 108, E92.

Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, Muselet D, Prud'homme JF, Dib C, Auffray C, Morissette J, Weissenbach J, Goodfellow PN. (1996). A radiation hybrid map of the human genome. Hum Mol Genet 5, 339-346.

Haldane JBS (1919). The mapping function. J Genet 8, 299-309.

Hansson CM, Ali H, Bruder CE, Fransson I, Kluge S, Andersson B, Roe BA, Menzel U, Dumanski JP. (2003). Strong conservation of the human NF2 locus based on sequence comparison in five species. Mamm Genome 14, 526-536.

Hearn T, Renforth GL, Spalluto C, Hanley NA, Piper K, Brickwood S, White C, Connolly V, Taylor JF, Russell-Eggitt I, Bonneau D, Walker M, Wilson DI. (2002). Mutation of *ALMS1*, a large gene with a tandem repeat encoding 47 amino acids, causes Alstrom syndrome. Nat Genet 31, 79-83.

Hennies HC, Rauch A, Seifert W, Schumi C, Moser E, Al-Taji E, Tariverdian G, Chrzanowska KH, Krajewska-Walasek M, Rajab A, Giugliani R, Neumann TE, Eckl KM, Karbasiyan M, Reis A, Horn D. (2004). Allelic heterogeneity in the *COH1* gene explains clinical variability in Cohen syndrome. Am J Hum Genet 75, 138-145.

Hirosawa M, Nagase T, Ishikawa K, Kikuno R, Nomura N, Ohara O. (1999). Characterization of cDNA clones selected by the GeneMark analysis from size-fractionated cDNA libraries from human brain. DNA Res 6, 329-336.

Hirvasniemi A, Lang H, Lehesjoki AE, Leisti J. (1994). Northern epilepsy syndrome: an inherited childhood onset epilepsy with associated mental deterioration. J Med Genet 31, 177-182.

Hofmann K, Bucher P, Falquet L, Bairoch A. (1999). The PROSITE database, its status in 1999. Nucleic Acids Res 27, 215-219.

Horn D, Krebsova A, Kunze J, Reis A. (2000). Homozygosity mapping in a family with microcephaly, mental retardation, and short stature to a Cohen syndrome region on 8q21.3-8q22.1: redefining a clinical entity. Am J Med Genet 92, 285-292.

Horwitz M, Benson KF, Duan Z, Li FQ, Person RE. (2004). Hereditary neutropenia: dogs explain human neutrophil elastase mutations. Trends Mol Med 10, 163-170.

Hromas R, Klemsz M. (1994). The ETS oncogene family in development, proliferation and neoplasia. Int J Hematol 59, 257-265.

Hu G. (1993). DNA polymerase-catalyzed addition of nontemplated extra nucleotides to the 3' end of a DNA fragment. DNA Cell Biol 8, 763-770.

Huang X, Miller W. (1991). A time-efficient, linear-space local similarity algorithm. Adv Appl Math, 12, 337-357.

Huang X, Madan A. (1999). CAP3: A DNA sequence assembly program. Genome Res 9, 868-877.

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 3, 204-211.

Ioannou PA, Amemiya CT, Garnes J, Kroisel PM, Shizuya H, Chen C, Batzer MA, de Jong PJ. (1994). A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nat Genet 6, 84-89.

Ioshikhes IP, Zhang MQ. (2000). Large-scale human promoter mapping using CpG islands. Nat Genet 26, 61-63.

Jackson MR, Nilsson T, Peterson PA. (1990). Identification of a consensus motif for retention of transmembrane proteins in the endoplasmic reticulum. EMBO J 9, 3153-3162.

Jalanko A, Kere J, Savilahti E, Schwartz M, Syvanen AC, Ranki M, Soderlund H. (1992). Screening for defined cystic fibrosis mutations by solid-phase minisequencing. Clin Chem 38, 39-43.

Jones KL, Smith DW. (1975). The Williams elfin facies syndrome. A new perspective. J Pediatr 86, 718-723.

Jongeneel CV, Bouvier J, Bairoch A. (1989). A unique signature identifies a family of zinc-dependent metallopeptidases. FEBS Lett 242, 211-214.

Jorcyk CL, Watson DK, Mavrothalassitis GJ, Papas TS. (1991). The human ETS1 gene: genomic structure, promoter characterization and alternative splicing. Oncogene 6, 523-532.

Kanehisa M, Bork P. (2003). Bioinformatics in the post-sequence era. Nat Genet 33, 305-310.

Karger AE. (1996). Separation of DNA sequencing fragments using an automated capillary electrophoresis instrument. Electrophoresis 17, 144-151.

Karim FD, Urness LD, Thummel CS, Klemsz MJ, McKercher SR, Celada A, Van-Beveren C, Maki RA, Gunther CV, Nye JA. (1990). The ETS-domain: a new DNA-binding motif that recognizes a purine-rich core DNA sequence. Genes Dev 4, 1451-1453.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ; University of California Santa Cruz. (2003). The UCSC Genome Browser Database. Nucleic Acids Res 31, 51-54.

Katsanis N, Beales PL, Woods MO, Lewis RA, Green JS, Parfrey PS, Ansley SJ, Davidson WS, Lupski JR. (2000). Mutations in MKKS cause obesity, retinal dystrophy and renal malformations associated with Bardet-Biedl syndrome. Nat Genet 26, 67-70.

Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, Scambler PJ, Davidson WS, Beales PL, Lupski JR. (2001). Triallelic inheritance in Bardet-Biedl syndrome, a mendelian recessive disorder. Science 293, 2256-2259.

Kelley LA, MacCallum RM, Sternberg MJ. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299, 499-520.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002). The human genome browser at UCSC. Genome Res 12, 996-1006.

Kent WJ. (2002). BLAT--the BLAST-like alignment tool. Genome Res 12, 656-564.

Khan J, Saal LH, Bittner ML, Chen Y, Trent JM, Meltzer PS. (1999). Expression profiling in cancer using cDNA microarrays. Electrophoresis 20, 223-229.

Kivitie-Kallio S, Rajantie J, Juvonen E, Norio R. (1997). Granulocytopenia in Cohen syndrome. Br J Haematol 98, 308-311.

Kivitie-Kallio S, Autti T, Salonen O, Norio R. (1998). MRI of the brain in the Cohen syndrome: a relatively large corpus callosum in patients with mental retardation and microcephaly. Neuropediatrics 29, 298-301.

Kivitie-Kallio S, Larsen A, Kajasto K, Norio R. (1999). Neurological and psychological findings in patients with Cohen syndrome: a study of 18 patients aged 11 months to 57 years. Neuropediatrics 30, 181-189.

Kivitie-Kallio S, Summanen P, Raitta C, Norio R. (2000). Ophthalmologic findings in Cohen syndrome. A long-term follow-up. Ophthalmology 107, 1737-1745.

Kivitie-Kallio S, Norio R. (2001). Cohen syndrome: essential features, natural history, and heterogeneity. Am J Med Genet 102, 125-135.

Koguchi K, Kobayashi S, Hayashi T, Matsufuji S, Murakami Y, Hayashi S. (1997). Cloning and sequencing of a human cDNA encoding ornithine decarboxylase antizyme inhibitor. Biochim Biophys Acta 1353, 209-216.

Kondo I, Nagataki S, Miyagi N. (1990). The Cohen syndrome: does mottled retina separate a Finnish and a Jewish type? Am J Med Genet 37, 109-113.

Korf I, Flicek P, Duan D, Brent MR. (2001). Integrating genomic homology into gene structure prediction. Bioinformatics 17 Suppl 1, S140-148.

Kyte J, Doolittle RF. (1982). A simple method for displaying the hydropathic character of a protein. J Mol Biol 157, 105-132.

Landegren U, Kaiser R, Sanders J, Hood L. (1988). A ligase-mediated gene detection technique. Science 241, 1077-1080.

Lander ES, Botstein D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236, 1567–1570.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Landschulz WH, Johnson PF, McKnight SL. (1989). The DNA binding domain of the rat liver nuclear protein C/EBP is bipartite. Science 243, 1681-1688.

Larsen F, Gundersen G, Lopez R, Prydz H. (1992). CpG islands as gene markers in the human genome. Genomics 13, 1095-1107.

Lathrop GM, Lalouel JM, Julier C, Ott J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am J Hum Genet 37, 482-498.

Lee YH, Huang GM, Cameron RA, Graham G, Davidson EH, Hood L, Britten RJ. (1999). EST analysis of gene expression in early cleavage-stage sea urchin embryos. Development 126, 3857-3867.

Lehesjoki AE, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, de la Chapelle A. (1993). Localization of the *EPM1* gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum Mol Genet 8, 1229-1234.

Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, May-Simera H, Li H, Blacque OE, Li L, Leitch CC, Lewis RA, Green JS, Parfrey PS, Leroux MR, Davidson WS, Beales PL, Guay-Woodford LM, Yoder BK, Stormo GD, Katsanis N, Dutcher SK. (2004). Comparative genomics identifies a flagellar and basal body proteome that includes the *BBS5* human disease gene. Cell 117, 541-552.

Li R, Pei H, Watson DK. (2000). Regulation of Ets function by protein-protein interactions. Oncogene 19, 6514-6523.

Lin W, Zhou X, Zhang M, Li Y, Miao S, Wang L, Zong S, Koide SS. (2001). Expression and function of the HSD-3.8 gene encoding a testis-specific protein. Mol Hum Reprod 7, 811-818.

Lipman DJ, Pearson WR. (1985). Rapid and sensitive protein similarity searches. Science 227, 1435-1441.

Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S. (2004). Eukaryotic regulatory element conservation analysis and identification using comparative genomics. Genome Res 14, 451-458.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288, 136-140.

Lovett M, Kere J, Hinton LM. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. Proc Natl Acad Sci USA 88, 9628-9632.

Lunetta KL, Boehnke M. (1994). Multipoint radiation hybrid mapping: comparison of methods, sample size requirements, and optimal study characteristics. Genomics 21, 92-103.

Magenis RE, Toth-Fejel S, Allen LJ, Black M, Brown MG, Budden S, Cohen R, Friedman JM, Kalousek D, Zonana J, Lacy D, LaFranchi S, Lahr M, Macfarlane J, Williams CPS. (1990). Comparison of the 15q deletions in Prader-Willi and Angelman syndromes: specific regions, extent of deletions, parental origin, and clinical consequences. Am J Med Genet 3, 333-349.

Maniatis T, Jeffrey A, van de Sande H. (1975). Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. Biochemistry 14, 3787-3794.

Marynen P, Zhang J, Cassiman JJ, Van den Berghe H, David G. (1989). Partial primary structure of the 48- and 90-kilodalton core proteins of cell surface-associated heparan sulfate proteoglycans of lung fibroblasts. Prediction of an integral membrane domain and evidence for multiple distinct core proteins at the cell surface of human lung fibroblasts. J Biol Chem 264, 7017-7024.

Massa G, Dooms L, Vanderschueren-Lodeweyckx MJ. (1991). Growth hormone deficiency in a girl with the Cohen syndrome. J Med Genet 28, 48-50.

Mathé C, Sagot MF, Schiex T, Rouze P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30, 4103-4117.

Mavrothalassitis GJ, Watson DK, Papas TS. (1990). Molecular and functional characterization of the promoter of *ETS2*, the human c-ets-2 gene. Proc Natl Acad Sci USA 87, 1047-1051.

Maxam AM, Gilbert W. (1977). A new method for sequencing DNA. Proc Natl Acad Sci USA 74, 560-564.

McNew JA, Goodman JM. (1996). The targeting and assembly of peroxisomal proteins: some old rules do not apply. Trends Biochem Sci 21, 54-58.

Mendez HMM, Paskulin GA, Vallandro C. (1985). The syndrome of retinal pigmentary degeneration, microcephaly, and severe mental retardation (Mirhosseini-Holmes-Walton syndrome): report of two patients. Am J Med Genet 22, 223-228.

Meyerowitz EM, Guild GM, Prestidge LS, Hogness DS. (1980). A new high-capacity cosmid vector and its use. Gene 11, 271-282.

Mirhosseini SA, Holmes LB, Walton DS. (1972). Syndrome of pigmentary retinal degeneration, cataract, microcephaly, and severe mental retardation. J Med Genet 9, 193-196.

Mochida GH, Rajab A, Eyaid W, Lu A, Al-Nouri D, Kosaki K, Noruzinia M, Sarda P, Ishihara J, Bodell A, Apse K, Walsh CA. (2004). Broader geographical spectrum of Cohen syndrome due to COH1 mutations. J Med Genet 41, E87.

Morton NE. (1955). Sequential tests for the detection of linkage. Am J Hum Genet 7, 277-318.

Moulard B, Genton P, Grid D, Jeanpierre M, Ouazzani R, Mrabet A, Morris M, LeGuern E, Dravet C, Mauguiere F, Utermann B, Baldy-Moulinier M, Belaidi H, Bertran F, Biraben A, Ali Cherif A, Chkili T, Crespel A, Darcel F, Dulac O, Geny C, Humbert-Claude V, Kassiotis P, Buresi C,

Malafosse A. (2002). Haplotype study of West European and North African Unverricht-Lundborg chromosomes: evidence for a few founder mutations. Hum Genet 111, 255-262.

Murakami K, Takagi T. (1998). Gene recognition by combination of several gene-finding programs. Bioinformatics 14, 665-675.

Mykytyn K, Braun T, Carmi R, Haider NB, Searby CC, Shastri M, Beck G, Wright AF, Iannaccone A, Elbedour K, Riise R, Baldi A, Raas-Rothschild A, Gorman SW, Duhl DM, Jacobson SG, Casavant T, Stone EM, Sheffield VC. (2001). Identification of the gene that, when mutated, causes the human obesity syndrome BBS4. Nat Genet 28, 188-191.

Mykytyn K, Nishimura DY, Searby CC, Shastri M, Yen HJ, Beck JS, Braun T, Streb LM, Cornier AS, Cox GF, Fulton AB, Carmi R, Luleci G, Chandrasekharappa SC, Collins FS, Jacobson SG, Heckenlively JR, Weleber RG, Stone EM, Sheffield VC. (2002). Identification of the gene (BBS1) most commonly involved in Bardet-Biedl syndrome, a complex human obesity syndrome. Nat Genet 31, 435-438.

Nakai K, Kanehisa M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14, 897-911.

Nathans D, Smith HO. (1975). Restriction endonucleases in the analysis and restructuring of dna molecules. Annu Rev Biochem 44, 273-293.

Nishimura DY, Searby CC, Carmi R, Elbedour K, Van Maldergem L, Fulton AB, Lam BL, Powell BR, Swiderski RE, Bugge KE, Haider NB, Kwitek-Black AE, Ying L, Duhl DM, Gorman SW, Heon E, Iannaccone A, Bonneau D, Biesecker LG, Jacobson SG, Stone EM, Sheffield VC. (2001). Positional cloning of a novel gene on chromosome 16q causing Bardet-Biedl syndrome (BBS2). Hum Mol Genet 10, 865-874.

Niwa J, Ishigaki S, Doyu M, Suzuki T, Tanaka K, Sobue G. (2001). A novel centrosomal ring-finger protein, dorfin, mediates ubiquitin ligase activity. Biochem Biophys Res Commun 281, 706-713.

Norio R, Raitta C, Lindahl, E. (1984). Further delineation of the Cohen syndrome; report on chorioretinal dystrophy, leukopenia and consanguinity. Clin Genet 25, 1-14.

Norio R, Raitta C. (1986). Are the Mirhosseini-Holmes-Walton syndrome and the Cohen syndrome identical? Am J Med Genet 25, 397-398.

Norio, R. (2003a). Finnish Disease Heritage II: population prehistory and genetic roots of Finns. Hum Genet 112, 457–469.

Norio R. (2003b). Finnish Disease Heritage I: characteristics, causes, background. Hum Genet 112, 441-456.

North C, Patton MA, Baraitser M, Winter RM. (1985). The clinical features of the Cohen syndrome: further case reports. J Med Genet 2, 131-134.

Obermüller S, Kiecke C, von Figura K, Honing S. (2002). The tyrosine motifs of Lamp 1 and LAP determine their direct and indirect targetting to lysosomes. J Cell Sci 115, 185-194.

Oefner PJ, Underhill PA. (1995). Comparative DNA sequencing by denaturing high-performance liquid chromatography (DHPLC). Am J Hum Genet 57, A266.

Ohara O, Dorit RL, Gilbert W. (1989). One-sided polymerase chain reaction: the amplification of cDNA. Proc Natl Acad Sci U S A 86, 5673-5677.

Ohashi J, Tokunaga K. (2003). Power of genome-wide linkage disequilibrium testing by using microsatellite markers. J Hum Genet 48, 487-491.

Okamoto N, Hatsukawa Y, Arai H, Goto M. (1998). Cohen syndrome with high urinary excretion of hyaluronic acid. Am J Med Genet 76, 387-388.

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 86, 2766-2770.

Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, Tanaka T, Ishii S, Yamamoto JI, Saito K, Kawai Y, Isono Y, Nakamura Y, Nagahari K, Murakami K, Yasuda T, Iwayanagi T, Wagatsuma M et a.l. (2004). Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet 36, 40-45.

Ott J. (1985). Analysis of human genetic linkage. The Johns Hopkins University Press, Baltimore, Maryland.

Ott J. (1991). Analysis of human genetic linkage. Rev. ed. The Johns Hopkins University Press, Baltimore, Maryland.

Ozsoylu S. (2001). Neutrophil elastase gene mutations in cyclic neutropenia. Turk J Pediatr 43, 180.

Pearson WR, Lipman DJ. (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85, 2444-2448.

Peltonen L, Jalanko A, Varilo T. (1999). Molecular genetics of the Finnish disease heritage. Hum Mol Genet 8, 1913-1923.

Peoples R, Perez-Jurado L, Wang YK, Kaplan P, Francke U. (1996). The gene for replication factor C subunit 2 (RFC2) is within the 7q11.23 Williams syndrome deletion. Am J Hum Genet 58, 1370-1373.

Persson B, Argos P. (1994). Prediction of transmembrane segments in proteins utilising multiple sequence alignments. J Mol Biol 237, 182-192.

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M. (1999). Mining SNPs from EST databases. Genome Res 9, 167-174.

Ponger L, Mouchiroud D. (2001). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. Bioinformatics 4, 631-633.

Pospisil H, Herrmann A, Bortfeldt RH, Reich JG. (2004). EASED: Extended Alternatively Spliced EST Database. Nucleic Acids Res, 32, D70-74.

Prader A, Labhart A, Willi H. (1956). Ein Syndrom von Adipositas, Kleinwuchs, Kryptorchismus und Oligophrenie nach Myatonieartigem Zustand im Neugeborenenalter. Schweiz Med Wschr 86, 1260-1261.

Pruess M, Fleischmann W, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, Mittard V, Mulder N, Phan I, Servant F, Apweiler R. (2003). The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Res 31, 414-417.

Pueyo R, Maneru C, Junque C, Vendrell P, Pujol J, Mataro M, Estevez-Gonzalez A, Garcia-Sanchez C. (2003). Quantitative Signal Intensity Measures on Magnetic Resonance Imaging in Attention-Deficit Hyperactivity Disorder. Cogn Behav Neurol 16, 75-81.

Rampoldi L, Dobson-Stone C, Rubio JP, Danek A, Chalmers RM, Wood NW, Verellen C, Ferrer X, Malandrini A, Fabrizi GM, Brown R, Vance J, Pericak-Vance M, Rudolf G, Carre S, Alonso E, Manfredi M, Nemeth AH, Monaco AP. (2001). A conserved sorting-associated protein is mutant in chorea-acanthocytosis. Nat Genet 28, 119-220.

Ramsay M, Williamson R, Estivill X, Wainwright BJ, Ho MF, Halford S, Kere J, Savilahti E, de la Chapelle A, Schwartz M. (1993). Haplotype analysis to determine the position of a mutation among closely linked DNA markers. Hum Mol Genet 2, 1007-1014.

Rao VN, Huebner K, Isobe M, ar-Rushdi A, Croce CM, Reddy ES. (1989). elk, tissue-specific ets-related genes on chromosomes X and 14 near translocation breakpoints. Science 244, 66-70.

Reedy BV, Bourne PE. (2003). Protein structure evolution and the SCOP database. Methods Biochem Anal 44, 239-248.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS, Tsui LC. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245, 1066-1073.

Rogic S, Mackworth AK, Ouellette FB. (2001). Evaluation of gene-finding programs on mammalian sequences. Genome Res 11, 817-832.

Roy MS, Milot JA, Polomeno RC, Barsoum-Homsy M. (1992). Ocular findings and visual evoked potential response in the Prader-Willi syndrome. Can J Ophthalmol 27, 307-312.

Rubio JP, Danek A, Stone C, Chalmers R, Wood N, Verellen C, Ferrer X, Malandrini A, Fabrizi GM, Manfredi M, Vance J, Pericak-Vance M, Brown R, Rudolf G, Picard F, Alonso E, Brin M, Nemeth AH, Farrall M, Monaco AP. (1997). Chorea-acanthocytosis: genetic linkage to chromosome 9q21. Am J Hum Genet 61, 899-908.

Russell-Eggitt IM, Clayton PT, Coffey R, Kriss A, Taylor DS, Taylor JF. (1998). Alstrom syndrome. Report of 22 cases and literature review. Ophthalmology 105, 1274-1280.

Sack J, Friedman E. (1986). The Cohen syndrome in Israel. Isr J Med Sci 22, 766-770.

Sambrook J, Fritsch EF, Maniatis T. (1989). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, NY, Cold Spring Harbor Laboratory Press.

Sanger F, Nicklen S, Coulson AR. (1977). DNA sequencing with chain-terminating inhibitors. Biotechnology 24, 104-108.

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, RodriguezTome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJR, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C et al. (1996). A gene map of the human genome. Science 274, 540-546.

Seth A, Ascione R, Fisher RJ, Mavrothalassitis GJ, Bhat NK, Papas TS. (1992). The ets gene family. Cell Growth Differ 3, 327-334.

Sheen VL, Ganesh VS, Topcu M, Sebire G, Bodell A, Hill RS, Grant PE, Shugart YY, Imitola J, Khoury SJ, Guerrini R, Walsh CA. (2004). Mutations in ARFGEF2 implicate vesicle trafficking in neural progenitor proliferation and migration in the human cerebral cortex. Nat Genet 36, 69-76.

Shin J, Dunbrack RL Jr, Lee S, Strominger JL. (1991). Signals for retention of transmembrane proteins in the endoplasmic reticulum studied with CD4 truncation mutants. Proc Natl Acad Sci USA 88, 1918-1922.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci USA 89, 8794-8797.

Slavotinek AM, Stone EM, Mykytyn K, Heckenlively JR, Green JS, Heon E, Musarella MA, Parfrey PS, Sheffield VC, Biesecker LG. (2001). Mutations in MKKS cause Bardet-Biedl syndrome. Nat Genet 26, 15-16.

Slonim D, Kruglyak L, Stein L, Lander E. (1997). Building human genome maps with radiation hybrids. J Comput Biol 4, 487-504.

Southern EM. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98, 503-517.

Stack JH, Horazdovsky B, Emr SD. (1995). Receptor-mediated protein sorting to the vacuole in yeast: roles for a protein kinase, a lipid kinase and GTP-binding proteins. Annu Rev Cell Dev Biol 11, 1-33.

Steinlein O, Tariverdian G, Boll HU, Vogel F. (1991). Tapetoretinal degeneration in brothers with apparent Cohen syndrome: nosology with Mirhosseini-Holmes-Walton syndrome. Am J Med Genet 41, 196-200.

Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, Fang N, Hadley D, Harris M, Hussain S, Lee R, Maratukulam A, OConnor K, Perkins S, Piercy M, Qin F, Reif T, Sanders C, She XH, Sun WL, Tabar P, Voyticky S, Cowles S, Fan JB, Mader C, Quackenbush J, Myers RM, Cox DR. (1997). An STS-based radiation hybrid map of the human genome. Genome Res 7, 422-433.

Strathdee G, Appleton K, Illand M, Millan DW, Sargent J, Paul J, Brown R. (2001). Primary ovarian carcinomas display multiple methylator phenotypes involving known tumor suppressor genes. Am J Pathol 158, 1121-1127.

Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Buetow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF et al. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci U S A 99, 16899-16903.

Tahvanainen E, Norio R, Karila E, Ranta S, Weissenbach J, Sistonen P, de la Chapelle A. (1994). Cohen syndrome gene assigned to the long arm of chromosome 8 by linkage analysis. Nat Genet 7, 201-204.

Tassabehji M, Metcalfe K, Fergusson WD, Carette MJA, Dore JK, Donnai D, Read AP, Proschel C, Gutowski NJ, Mao X, Sheer D. (1996). LIM-kinase deleted in Williams syndrome. Nat Genet 13, 272-273.

Teasdale RD, Jackson MR. (1996). Signal-mediated sorting of membrane proteins between the endoplasmic reticulum and the golgi apparatus. Annu Rev Cell Dev Biol 12, 27-54.

Terwilliger, JD. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56, 777-787.

Thanaraj TA, Stamm S, Clark F, Riethoven JJ, Le Texier V, Muilu J. (2004). ASD: the Alternative Splicing Database. Nucleic Acids Res 32, D64-69.

Thomaidis L, Fryssira H, Katsarou E, Metaxotou C. (1999). Cohen syndrome: two new cases in siblings. Eur J Pediatr 158, 838-841.

Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673-4680.

Touchman JW, Dehejia A, Chiba-Falek O, Cabin DE, Schwartz JR, Orrison BM, Polymeropoulos MH, Nussbaum RL. (2001). Human and mouse alpha-synuclein genes: comparative genomic sequence analysis and identification of a novel gene regulatory element. Genome Res 11, 78-86.

Ueno S, Maruki Y, Nakamura M, Tomemori Y, Kamae K, Tanabe H, Yamashita Y, Matsuda S, Kaneko S, Sano A. (2001). The gene encoding a newly discovered protein, chorein, is mutated in chorea-acanthocytosis. Nat Genet 28, 121-122.

Valle D, Simell O. (1983). The hyperornithinemias. In: Stanbury JB, Wyngaarden JB, Fredrickson DS, Goldstein JL, Brown MS. The Metabolic Basis of Inherited Disease. New York: McGraw-Hill, 382-401.

Velayos-Baeza A, Vettori A, Copley RR, Dobson-Stone C, Monaco AP. (2004). Analysis of the human *VPS13* gene family. Genomics 84, 536-549.

Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M. (1998). Shotgun sequencing of the human genome. Science 280, 1540-1542.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XQH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang JH, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau C et al. (2001). The Sequence of the Human Genome. Science 291, 1304-1351.

Virtaneva K, D'Amato E, Miao J, Koskiniemi M, Norio R, Avanzini G, Franceschetti S, Michelucci R, Tassinari CA, Omer S, Pennacchio LA, Myers RM, Dieguez-Lucena JL, Krahe R, de la Chapelle A, Lehesjoki AE. (1997). Unstable minisatellite expansion causing recessively inherited myoclonus epilepsy, EPM1. Nat Genet 15, 393-396.

Walter MA, Spillett DJ, Thomas P, Weissenbach J, Goodfellow PN. (1994). A method for constructing radiation hybrid maps of whole genomes. Nat Genet 7, 22-28.

Warburg M, Pedersen SA, Horlyk H. (1990). The Cohen syndrome. Retinal lesions and granulocytopenia. Ophthalmic Paediatr Genet 11, 7-13.

Watson DK, McWilliams MJ, Lapis P, Lautenberger JA, Schweinfest CW, Papas TS. (1988). Mammalian ets-1 and ets-2 genes encode highly conserved proteins. Proc Natl Acad Sci USA 85, 7862-7866.

Weber JL, May PE. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44, 388-396.

Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB, Robinson GE. (2002). Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. Genome Res 12, 555-566.

Williams JC, Barratt-Boyes BG, Lowe JB. (1961). Supravalvular aortic stenosis. Circulation 24,1311-1318.

Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, Pruss M, Schacherer F, Thiele S, Urbach S. (2001). The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29, 281-283.

Young ID, Moore JR. (1987). Intrafamilial variation in Cohen syndrome. J Med Genet 24, 488-492.

Yu W, Andersson B, Worley KC, Muzny DM, Ding Y, Liu W, Ricafrente JY, Wentland MA, Lennon G, Gibbs RA. (1997). Large-scale concatenation cDNA sequencing. Genome Res 7, 353-358.

Zatyka M, Morrissey C, Kuzmin I, Lerman MI, Latif F, Richards FM, Maher ER. (2002). Genetic and functional analysis of the von Hippel-Lindau (VHL) tumour suppressor gene promoter. J Med Genet 39, 463-472.

Zhang ML, Wang LF, Miao SY, Koide SS. (1992). Isolation and sequencing of the cDNA encoding the 75-kD human sperm protein related to infertility. Chin Med J 105, 998-1003.