UNIVERSITÀ DEGLI STUDI DI MILANO

Corso di Dottorato in Biologia Molecolare e Cellulare

XXIX Ciclo

Unraveling the molecular complexity of Parkinson's disease: from genetic risk factors to Mendelian causative genes

Letizia Straniero

PhD Thesis

Scientific tutor: **Dott.ssa Giulia Soldà**

Academic year: 2015-2016

SSD: BIO/11; BIO/13

Thesis performed at the Department of Medical Biotechnology and Translational Medicine and at the Humanitas Clinical and Research Center

Index

ABBREVIATIONS AND NOTES

AD: autosomal dominant

AR: autosomal recessive

ceRNAs: competing endogenous RNAs

CLEAR: coordinated lysosomal expression and regulation

CNS: central nervous system

ER: endoplasmic reticulum

FL: full length

GCase: beta-glucocerebrosidase

GD: Gaucher disease

GlcCer: glucosylceramide

GWAS: genome-wide association study

HGMD: human gene mutation database

HS: heparan sulfate

LNA: locked nucleic acid

lncRNAs: long non-coding RNAs

LOVD: leiden open variation database

miRNA: microRNA

MPS: mucopolysaccharidosis

MRE: miRNA recognition element

NGS: next-generation sequencing

NMD: nonsense-mediated mRNA decay

OR: odds ratio

PD: Parkinson's disease

PTC: premature stop codon

SNPs: single nucleotide polymorphisms

TFEB: transcription factor EB

UPR: unfolded protein response

WES: whole-exome sequencing

WGS: whole-genome sequencing

Note. In this thesis, materials and methods are reported in the result section, below each figure.

Part I

Abstract

Parkinson's disease (PD) is a neurodegenerative disorder characterized by the progressive death of dopaminergic neurons in the substantia nigra. PD is a complex disease caused by the combination of environmental factors and genetic components. In recent years, considerable progresses have been made in the identification of the genetic determinants of PD: several genes were shown to cause rare monogenic forms of the disease. Moreover, a larger number of predisposing genetic variants have been associated with sporadic PD by genome-wide association studies. Among them, GBA seems to be the main genetic risk factor for PD. However, despite these advancements, a large fraction of the expected PD heritability is still missing.

In this frame, the main aims of my PhD project were, on one hand, the study of the possible mechanisms of regulation of GBA, in particular the assessment of the role of GBAP1, its pseudogene, with the view to identify novel strategies to augment glucocerebrosidase activity and, on the other hand, the identification of novel genetic determinants for the disease by whole-exome sequencing (WES) of selected PD families.

Concerning GBA we demonstrated that mir-22-3p can modulate the levels of GBA transcripts down-regulating GCase expression and that GBAP1 might work as competing endogenous RNA acting as sponge and decreasing the miRNA mediated control on GBA. We then characterized the splicing pattern of the pseudogene and we showed that GBAP1 transcripts are down-regulated by the nonsense-mediated mRNA decay mechanism. On the basis of these results, we propose the existence of an RNA-based complex regulatory network involving GBA, GBAP1 and miR-22-3p.

Regarding the identification of novel PD genes, 24 PD families with a dominant or a recessive inheritance pattern were selected and whole-exome sequencing was performed on the proband of each family and on affected cousins or uncles, when available.

Six of these families resulted carriers of mutations in genes already associated with parkinsonism (4 GBA, 1 LRRK2, and 1 ATP7B). Moreover, we found 2 homozygous mutations in novel candidate genes in 2 consanguineous families: a splicing mutation in a gene involved in the unfolded protein response (DNAJC12) and a missense variant in a gene coding for a lysosomal enzyme (HGSNAT).

We characterized the splicing mutation and we demonstrated that this variation causes the skipping of the downstream exon and the introduction of the premature stop codon. Interestingly, we demonstrated that the silencing of the DNAJC12 gene increases the α-synuclein protein levels in SH-SY5Y neuroblastoma cells. Concerning the missense mutation, we measured the activity of the HGSNAT enzyme in the patient's fibroblasts, which resulted below the lower limit of the normal range. We hypothesize that the HGSNAT activity is not reduced enough to cause a lysosomal disorder, but may predispose to PD. Moreover, the pathogenic role of the two novel PD genes identified in my PhD project is strongly supported by the identification, by us and by our collaborators in Vancouver, of one additional unrelated PD family with mutations in the same gene for both HGSNAT and DNAJC12. Ongoing analyses will characterize the functional role of the novel candidate genes in PD pathogenesis.

1 | State of the Art

1.1 Parkinson's Disease

Parkinson's disease (PD) is a chronic and progressive neurodegenerative disease with a multifactorial etiology, resulting from the combination of different genetic determinants and multiple environmental factors.

The disease affects more than 1% of the population over 55 years of age and the percentage rises to 3% over the age of 75. The annual incidence (adjusted for age and gender) is 13.4 new cases per 100,000 individuals; in particular, the number of cases is higher in males (19:100,000) than in females (9.9:100,000) [Farlow et al., 2014]. The lifetime risk to develop the disease is close to 1.5% and the median age of onset is around age 60, with a life expectancy of 15 years after diagnosis [Lees et al., 2009]. In general, an onset before age 20 is very rare and is considered juvenile Parkinson's, before age 50 is called early-onset Parkinson's, while after age 60 is classified as late-onset and it is the most common form [Farlow et al., 2014]. It is estimated that 5% of patients manifest symptoms before age 40 [Schrag et al., 2006]. The symptoms characterizing the disease and on which the diagnosis is based, consist of: resting tremor, bradykinesia, rigidity and postural instability. Generally, in the early stages of the disease the symptoms appear asymmetrically and then affect the rest of the body. In late stages, the worsening of symptoms is associated with cognitive and psychiatric disorders, such as hallucinations, depression, sleep and speech disorders, memory loss and dementia.

1.1.1 Pathogenesis

PD is a disorder of the central nervous system (CNS) characterized by the progressive degeneration, and the consequent death by apoptosis, of the dopaminergic neurons localized in the substantia nigra, a region of the brain located between the midbrain and the diencephalon [Lees et al., 2009]. The first symptoms of the disease appear when the loss of dopaminergic neurons exceeds the 50%, a level beyond which compensation systems, which usually contribute to maintain normal mobility, are not enough to overcome the lack of dopamine (a neurotransmitter typically secreted by dopaminergic neurons). During disease progression, the neuronal degeneration does not remain confined to the substantia nigra, but also extends to other areas of the brain and other neuronal types, resulting in the appearance of secondary symptoms typical of the late stages of the disease.

It is widely accepted that the neurodegeneration occurs in response to a series of pathogenic events that could take place inside neurons themselves - such as mitochondria dysfunction, dysregulation of calcium and ROS (Reactive Oxygen Species) homeostasis, alterations in the protein quality control and degradation systems, lysosomal impairment, and alterations in the process of autophagy and apoptosis - or come from outside, with the transfer of the α-synuclein protein from an abnormal cell to a normal one, with a mechanism of action similar to the one described in prion diseases [Hirsch et al., 2013; Heman-Ackah et al., 2013; Antony et al., 2013].

In addition to the neuronal degeneration, another hallmark of PD is the presence of eosinophilic cytoplasmic aggregates of proteins, mainly consisting of α-synuclein, called Lewy bodies [Spillantini et al., 1999; Cooper et al., 2006; Credle et al., 2015; Hunn et al., 2015]. Lewy bodies are also found in other conditions generically defined as "synucleinopathies", and sometimes in Alzheimer's disease. Their role in PD pathogenesis has still to be defined: it is not clear whether they interfere with cell function or if they represent a protective response by which the neuron tries to confine and eliminate cytotoxic proteins [Minami et al., 2015].

1.1.2 Etiology

The etiology of PD is still unclear, although it is now widely accepted as a multifactorial disease, resulting from the interaction of environmental components and genetic factors. The role of genetic factors seems to be predominant in the early-onset forms, while the etiology of idiopathic late-onset PD still remains difficult to explain [Goldman, 2014]. In most cases the onset is sporadic and only in 5-10% of patients it is possible to trace a family history of the disease [Spatola and Wider, 2014]. The major risk factor is age, although it is not fully understood how the aging process is involved in the onset of the disease.

Epidemiological data indicate that some environmental and occupational factors such as lifestyle, exposure to chemicals, industrials, pesticides and metals (Al, Cu, Fe, Hg, Mn, Pb), the rural environment and professional activity (agricultural or mining work) may increase the risk to develop the disease. However, other studies are necessary to define in detail doses, exposure mode, and susceptible populations [Goldman, 2014]. Smoking and caffeine seem to be associated with a protective effect [Goldman, 2014], but there isn't a unanimous agreement on this subject and the debate is still open, with some authors suggesting the possibility that the apparent "neuroprotective" effect of smoking and caffeine observed in epidemiologic studies may be due to reverse causation [Ritz et al., 2014].

1.2 Genetics of PD

Until few years ago the role of environmental factors in the development of PD was considered as predominant, but accumulating evidence suggests that the genetic component may play a more important role than previously thought. Over the last twenty years, linkage analysis and positional cloning in large families with Mendelian forms of PD led to the identification of new genes involved in the pathogenesis of the disease, whereas association studies have identified an increasing number of loci and susceptibility genes responsible for sporadic forms of PD.

1.2.1 Linkage analyses

Various linkage studies have been performed on families with Mendelian or quasi-Mendelian transmission of PD to identify chromosomal regions associated with monogenic forms of the disease. Thanks to this approach, 14 loci have been identified, involved in both autosomal dominant (PARK

1/4, 3, 5, 8 and 17) and autosomal recessive forms of the disease (PARK 2, 6, 7, 9, 14 and 15). For 12 of these loci the causative gene was also identified [Lesage and Brice, 2009; Lin and Farrer 2014], as reported in Table 1.1. The number of genes involved in PD is recently increased considering the latest discoveries using the whole-exome sequencing approach [Zimprich A et al., 2011; Vilariño-Güell C et al., 2011; Edvardson et al., 2012; Krebs et al., 2013; Quadri et al., 2013; Vilariño-Güell C et al., 2014; Funayama et al., 2015; Lesage et al., 2016].

Concerning the autosomal dominant (AD) forms, the most important genes are SNCA and LRKK2. SNCA codes for α-synuclein, a protein predominantly expressed at the pre-synaptic membrane of CNS neurons that is considered the main component of Lewy bodies. Although the exact function of this protein is not yet entirely clear, it seems to have a role in synaptic plasticity, in vesicular trafficking, and in the inhibition of dopamine release. To date duplications/triplications of the SNCA locus and three missense mutations have been described (NM_000345.3 c.88G>C p.A30P; c.136G>A p.E46K; c.157G>A p.A53T); changes in α-synuclein expression or the presence of mutations that increase its tendency to form aggregates have a toxic effect on dopaminergic neurons [Hunn et al., 2015].

LRRK2 (leucine-rich repeat kinase 2) codes for a kinase expressed in different areas of the brain, including the substantia nigra. About 80 mutations in this gene have been reported, however, only seven can be considered pathogenic. Among these, the most important in terms of frequency are the NM_198578 c.4321C>T (p.R1441C) and the c.6055G>A (p.G2019S). It is thought that the protein is involved in intracellular signaling and vesicular trafficking, although its precise biological function has not been clarified yet [Martin et al., 2014; Spatola and Wider et al., 2014].

Concerning the autosomal recessive forms (AR), the main genes are: PRKN, PINK1, and DJ-1.

PRKN encode the parkin protein, an E3 ubiquitin ligase that conjugates ubiquitin to proteins targeted to be degraded by the proteasome. Parkin is recruited by mitochondria depolarization, where it induces the ubiquitination of the mitochondrial outer membrane proteins and promotes the removal of defective mitochondria by a specialized form of autophagy called mitophagy [Yin et al., 2013]. More than 100 mutations have been identified in the PRKN gene, most of which are large deletions or duplications of one or more exons; small deletions/insertions, nonsense and missense mutations were also described. Mutations in this gene are responsible for 50% of the recessive forms of early-onset PD [Lesage and Brice, 2012].

PINK1 (PTEN induced kinase 1) encodes a mitochondrial serine-threonine kinase that seems to protect neurons from mitochondrial dysfunction and oxidative stress-induced apoptosis. More than 50 mutations were discovered in this gene, most of which are point mutations, but deletions of the entire gene were also reported [Lesage and Brice, 2012].

Finally, the DJ-1 gene codes for a very abundant protein in the brain, and in particular in astrocytes. DJ-1, forming a complex with PINK1 and parkin, acts as sensor for oxidative stress, and it has a role in

the regulation of mitochondria morphology, as well as in the degradation of unfolded proteins [Wilhelmus et al., 2012]. Defects in this gene are quite rare, only 15 mutations were described [Lesage and Brice, 2012].

Table 1.1| Loci and genes associated with monogenic forms of PD

Locus	**Chromosome**	**Gene**	**Inheritance**	**Possible pathways**
PARK1/PARK4	4q21-q22	SNCA	AD	Synaptic function
PARK2	6q25.2-q27	PRKN	AR/sporadic	Mitochondrial function/mitophagy; ubiquitination
PARK3	2p13	uk	AD	-
PARK5	4p13	UCHL1	AD/sporadic	Ubiquitination
PARK6	1p35-p36	PINK1	AR/sporadic	Mitochondrial function/mitophagy
PARK7	1p36	DJ-1	AR	Mitochondrial function
PARK8	12q12	LRRK2	AD/sporadic	Synaptic function; autophagy/lysosomal degradation
PARK9	1p36	ATP13A2	AR	autophagy/lysosomal degradation
PARK11	2q36-37	uk	AD	-
PARK13	2p12	HTRA2	AD/sporadic	autophagy/lysosomal degradation
PARK14	22q13.1	PLA2G6	AR	Mitochondrial function
PARK15	22q12-q13	FBXO7	AR	Ubiquitination
PARK17	16q11.2	VPS35	AD	Autophagy/lysosomal degradation; endocytosis
PARK18	3q27.1	EIF4G1	AD	Translation
PARK19	1p31.3	DNAJC6	AR	Synaptic function; endocytosis
PARK20	21q22.11	SYNJ1	AR	Synaptic function; endocytosis
PARK21	3q22.1	DNAJC13	AD/sporadic	Synaptic function; endocytosis
PARK22	7p11.2	CHCHD2	AD	Mitochondrial function
PARK23	15q22.2	VPS13C	AR	Mitochondrial function

uk, unknown. Modified from Klein et al., 2012.

1.2.2 Association studies

Numerous case-control association studies, both on candidate genes and on the entire genome (Genome-Wide Association Study, GWAS), allowed the identification of genetic variants predisposing to PD in different populations.

Association studies were also instrumental to confirm the involvement of genes responsible for Mendelian forms of PD, like SNCA and LRRK2, in sporadic PD. In particular, regarding the SNCA gene, the association with sporadic PD of the dinucleotide microsatellite Rep1, located upstream of the transcription start site, has been repeatedly reported; moreover, haplotype analysis confirmed the association between different SNPs (Single Nucleotide Polymorphisms, SNPs) in the gene and the disease. Even meta-analysis studies support the association of SNCA with PD; in particular, the association signals are located in the promoter region and at the 3' end of the gene [Lesage and Brice, 2012]. These data were also confirmed in a large Italian cohort of 891 healthy subjects and 904 PD patients [Trotta et al., 2012]. The same study also verified the association between PD and MAPT (Microtubule-Associated Protein Tau), a gene encoding for tau, the microtubule-associated protein expressed in the adult CNS. SNPs in this gene have been recently associated with PD; in particular, the H1 haplotype, one of two common haplotypes for this locus in the Caucasian population, is a risk factor with an estimated odd ratio (OR) of 1.5 [Lesage and Brice, 2012]. Interestingly, mutations described in the MAPT gene, which usually cause tau hyperphosphorylation resulting in tau protein accumulation in neurons, are associated with various forms of dementia known as tauopathies [Kimura et al., 2014].

In addition, the most common and important risk factor for PD seems to be the GBA gene, whose association with PD was repeatedly validated [Sidransky et al., 2009; Do et al., 2011; IPDCG et al., 2011; Liu et al., 2011]. This gene codes for a lysosomal enzyme, the beta-glucocerebrosidase (GCase), whose severe deficiency causes the Gaucher disease (GD), one of the most frequent lysosomal storage disorders. Heterozygosity for mutations in this gene have been associated with a 5 times higher risk to develop PD [Lesage and Brice, 2012].

Besides SNCA, LRRK2, MAPT and GBA, now commonly considered the major susceptibility genes for PD, a large number of GWAS on different populations, followed by several meta-analyses, lead to the identification of new genetic determinants of PD listed in Table 1.2.

Table 1.2| Main GWAS results

Chromosome	Gene	OR	p value
1q21	GBA	5.43	$1.44 \cdot 10^{-14}$
1q21.1	SYT11	1.44	$1.18 \cdot 10^{-6}$
1q32	RAB7L1/PARK16	0.86	$1.52 \cdot 10^{-12}$
2q21.3	ACMSD	1.07	$3.16 \cdot 10^{-3}$
2q24.3	STK39	1.12	$1.64 \cdot 10^{-3}$
3q27	MCCC1/LAMP3	0.87	$6.92 \cdot 10^{-5}$
4p15	BST1	0.87	$3.94 \cdot 10^{-9}$
4p16	GAK	1.14	$7.46 \cdot 10^{-8}$
4q21.1	SCARB2	0.90	$7.60 \cdot 10^{-10}$
4q21.1	STBD1	0.91	$5.29 \cdot 10^{-5}$
6p21.3	HLA-DRB5	0.80	$5.50 \cdot 10^{-10}$
7p15	GPNMB	0.89	$3.86 \cdot 10^{-8}$
8p22	FGF20	0.88	$8.49 \cdot 10^{-7}$
12q24	CCDC62/HIP1R	1.13	$9.06 \cdot 10^{-7}$
16p11.2	STX1B	1.15	$8.21 \cdot 10^{-9}$
17p11.2	SREBF1	0.95	$5.60 \cdot 10^{-8}$
17q21.1	MAPT	0.80	$1.95 \cdot 10^{-16}$

Modified from Singleton et al., 2013.

1.3 GBA

The GBA gene encode the beta-glucocerebrosidase, a lysosomal enzyme involved in the metabolism of sphingolipids. There are two other isoforms of this enzyme: GBA2, which is localized at the plasma membrane, and GBA3, which is the cytosolic isoform [de Graaf et al., 2001; Boot et al., 2007].

The GBA gene is located on chromosome 1q22, it comprises 11 exons and spans 7.6kb (Figure 1.1). The gene is characterized by a proximal (P1) and a distal (P2) promoter: P1 contains a TATA box and different binding sites for transcription factors, while P2, located 2.6kb upstream of the translation start site, shows the typical features of an housekeeping gene promoter [Svobodová et al., 2011].

In the same genomic region, there is also a highly homologous pseudogene (GBAP1), located approximately 16kb downstream of the GBA gene, which shares the same exon-intron organization of the functional gene (Figure 1.1).

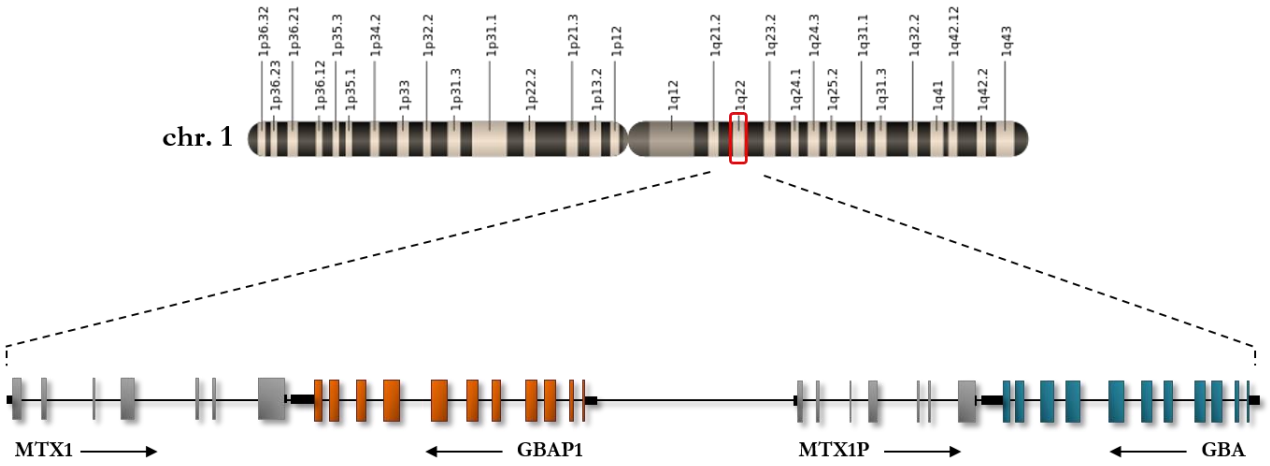


Figure 1.1 | Localization and structure of the GBA gene locus.
A schematic representation of chromosome 1 with the localization of the GBA and GBAP1 genes; introns are represented by black lines and exons by rectangles.

The glucocerebrosidase, belonging to the glycoside hydrolase family, is a protein mainly associated to the internal lysosomal membrane, even if small quantities of the enzyme are present on the extracellular membrane. GCase catalyzes the hydrolysis of glucosylceramide (GlcCer), a membrane glycosphingolipid, to ceramide and glucose.

The GCase is characterized by three structural domains. The domain I (residues 1-27 and 383-414) contains two disulfide bridges important for the correct protein folding and some specific glycosylated residues essential for catalytic activity. The domain II (residues 30-75 and 431-497) has a regulatory and structural function, and it is also important for the interaction with the saposin C, an enzyme activator also mediating the interaction with the lysosomal membrane. Finally, the domain III (residues 76-381 and 416-430) is constituted by a TIM barrel domain with catalytic function [Dvir et al., 2003].

Severe deficiencies of the enzyme lead to the accumulation of the substrate and are responsible for the multi-organ clinical manifestations of the Gaucher disease (GD), one of the most common genetic

lysosomal storage disorders. Although GD encompasses a continuum of clinical presentations from a perinatal lethal disorder to an asymptomatic form, traditionally, it has been divided in three major clinical types based upon the absence (type 1, non-neuronopathic, OMIM #230800) or presence of primary CNS involvement (Type 2, acute neuronopathic, OMIM #230900; and Type 3, subacute neuronopathic, OMIM #231000). Over 300 mutations have been identified in GBA and linked to GD so far [Hruska et al., 2008], and, interestingly, no strong correlation between the clinical phenotype, the genotype, and the residual enzyme activity was observed [Goker-Alpan et al., 2005; Lachmann et al., 2004].

In recent years, GD and PD have been connected on account of the clinical observation of parkinsonism and LB pathology in patients with GD [Lwin et al., 2004]. Compared with the general population, patients with GD type 1 have a 20-fold increased lifetime risk of developing parkinsonism [Westbroek et al., 2011], whereas individuals carrying heterozygous GBA mutations, have a five times greater risk of developing PD than non-carrier individuals [Sidransky et al., 2009]. Several studies confirmed that GBA mutations, in particular the two most common mutations N370S and L444P, are more frequent in PD patients than in healthy controls, demonstrating that genetic lesions in this gene are a common risk factor for the disease, especially in the familial form [International Parkinson's Disease Genomics Consortium, 2011].

Despite these efforts, to date, the mechanism underlying the relation between GBA mutations and the development of PD remains unclear and it is a cause of debate in the scientific community. In fact, there are reports in the literature supporting a gain-of-function effect of the mutated protein, as well as publications supporting a loss-of-function mechanism [Sidransky et al., 2012]. Moreover, not only a widespread deficiency of GCase activity has been demonstrated in the brains of PD patients carrying GBA mutations, but also PD patients without GBA mutations were shown to exhibit deficiency of GCase in substantia nigra [Gegg et al., 2012]. Importantly, recent data demonstrated that in neurons and in brains from this type of PD patients, the lysosomal accumulation of GlcCer directly influences the abnormal lysosomal storage of α-synuclein oligomers, thus resulting in a further inhibition of the GBA activity. These findings suggest, for the first time, that the bi-directional effect of GBA and α-synuclein accumulation forms a positive feedback loop that may lead to a self-propagating disease [Mazzulli et al., 2011].

1.4 MicroRNAs

MiRNAs are short (19-25 nucleotides) single-stranded RNAs found in plant and animals. They act as post-transcriptional regulators of gene expression by repressing target mRNAs translation and/or by inducing the degradation of the mRNA. Hundreds of miRNAs have been cloned and thousands more have been predicted bioinformatically, making them a major class of regulators [Berezikov et al., 2006]. Assisted by the RNA silencing machinery, each miRNA might inhibit the expression of hundreds of target mRNAs [Lim et al., 2005], whose recognition is based on imperfect complementary binding between miRNAs and their target sites, usually located within the 3' untranslated regions. In particular, a region of 7-8 nucleotides at the 5' end of the miRNA, called seed region, seems to be extremely important for target recognition [Nielsen et al., 2007]. MiRNAs have been implicated in a variety of biological processes, including embryonic development, cell differentiation, cell cycle regulation, and apoptosis [Esquela-Kerscher et al., 2006], as well as in pathological processes like cancer [Meltzer et al., 2005], Alzheimer disease and metabolic disorders [Krützfeldt et al., 2006]. The evidence of a dysregulated expression of specific miRNAs in different human diseases makes them both promising diagnostic markers and therapeutic agents or targets. In addition, miRNA-based drugs have already been developed and they have a great potential for treating cancer, psychiatric disorders and other human diseases [Srinivasan et al., 2013].

MiRNAs can be encoded as independent transcriptional units, or be part of polycistronic clusters. They are mainly transcribed by RNA polymerase II or III as primary transcripts (pri-miRNAs) containing domains that allow the folding into a hairpin structure. These pri-miRNAs are subsequently recognized and processed in the nucleus by the Drosha complex, which generates precursors of 70-100 nucleotides (pre-miRNA), that are subsequently exported in the cytoplasm by the exportin 5 in a Ran-GTP-dependent manner. In the cytoplasm, the Dicer complex recognizes the pre-miRNA and produces mature miRNA through a endonucleolytic cleavage that removes the loop of the hairpin, generating a double-strand RNA of 22 nucleotides with 3 'ends protruding, which is eventually recognized and assembled into the RISC complex (RNA-induced silencing complex). Usually, the functional miRNA, derived from one of the two strands, is selected and loaded on the Argonaut protein (Ago) of the complex, while the complementary strand, referred to as miRNA*, is generally degraded [He and Hannon, 2004]. However, more recent data have shown that miRNAs* are often present physiologically at relevant concentration and can be associated with an Ago protein inhibiting specific mRNA targets, as demonstrated both in cell culture and in transgenic animals [Okamura et al., 2008; Jazdzewski et al., 2009]. Considering that both strand could act as functional miRNA; the two miRNAs are usually called 3p or 5p depending on the strand of the hairpin from which they originate.

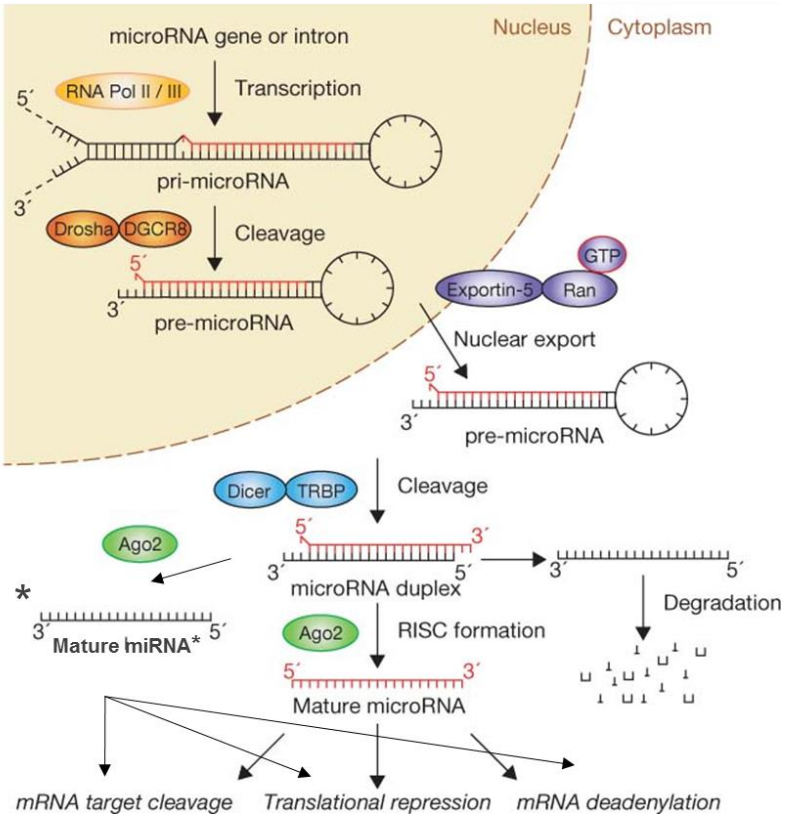


Figure 1.2|miRNA Biogenesis.
Modified from He and Hannon, 2004.

Several studies identified miRNAs that target genes associated with Mendelian or sporadic forms of PD. For example, two miRNAs, miR-7 and miR-153, were confirmed to target α-synuclein [Doxakis, 2010]. In particular, miR-7 is able to suppress α-synuclein-mediated cytotoxicity in neuronal cell models [Junn et al., 2009]. Moreover, studies on miRNA expression pattern in brains of PD patients showed a decrease in the levels of miR-34b and miR-34c. In PD cellular models, a miR-34b/c reduction was associated to a decrease of DJ-1 and Parkin levels, resulting in mitochondrial dysfunction and production of reactive oxygen species that lead to a compromised neuron viability [Miñones-Moyano et al., 2011]. Recently it was also demonstrated that these two microRNAs repress the expression of α-synuclein directly targeting its 3'UTR [Kabaria et al., 2015]. Finally, miR-205 was demonstrated to represses the expression of LRRK2. The observation that the levels of miR-205 in the frontal cortex was significantly lower in PD patients vs healthy controls suggests that this microRNA may play an important role in the modulation of LRRK2 expression also in the brains of sporadic PD patients [Cho et al., 2013].

Despite these interesting results concerning miRNAs that specifically target genes associated with familial forms of PD, the predominant role of miRNAs in PD pathogenesis might be their ability to regulate other disease-correlated genes. Therefore, it will be particularly important to understand the role of microRNAs in the regulation of networks of genes involved in the different pathways implicated in the pathogenesis of PD.

1.5 Competing endogenous RNAs

The advent of high-throughput techniques for the study of the transcriptome led to the discovery that more than 75% of the genome is transcribed, confirmed the existence of a large number of lncRNAs (long non-coding RNAs) and demonstrated that a number of pseudogenes are indeed actively transcribed. Recent data also indicate the increasing importance of non-protein coding genes, demonstrating the crucial role of lncRNAs in essential regulatory networks frequently dysregulated in pathological conditions [Milligan and Lipovich, 2015].

Among the many classes of non-coding RNAs with regulatory functions, an interesting new one is represented by competing endogenous RNAs (ceRNAs). These transcripts would act by competing with protein-coding mRNAs for a small pool of microRNA (miRNAs), thus modulating the miRNA-mediated expression regulation [Seitz, 2009; Poliseno et al., 2010].

Good candidates to play the role of ceRNA is the class of pseudogenes, since they have a high-sequence identity with the ancestral gene and, consequently, they could compete for the binding of the same miRNAs. Moreover, many among them are now known to be transcribed, even if they usually have lost the ability to generate a functional protein product [Poliseno et al., 2010]. Variations in the levels of pseudogene expression may result in changes in the levels of the transcripts regulated by the miRNAs targeting the pseudogene. In particular, high levels of ceRNA could bind a great amount of miRNA increasing the expression of the corresponding functional gene. By contrast, a low quantity of competitor increases the possibility for the shared miRNAs to interact with the protein coding gene, leading to the consequent reduction of its expression [Cazalla et al., 2010; Wang et al., 2010; Lee et al., 2010; Salmena et al., 2011].

Interestingly, the 5.5-kb-long GBAP1 pseudogene has maintained a 96% of sequence identity with the functional gene and shares the same exon-intron organization [Wafaei and Choy, 2005]. GBAP1 is the result of a tandem duplication that involves the GBA gene and the MTX1 gene, located immediately downstream of GBA (Figure 1.1). This genomic arrangement is evolutionary very recent, considering that it is present only in primates [Svobodová et al., 2011]. The main difference between GBA and GBAP1 consists in the presence of 3 additional Alu insertions in GBA introns, while a 55-bp deletion in exon 9 is a unique hallmark of GBAP1, and can be used to distinguish the two transcripts [Wafaei and Choy, 2005]. A large variety of predicted GBAP1 transcripts are annotated in the UCSC genome browser, however the majority of them is not experimentally validated (Figure 1.3).

Since it is known that GBAP1 is an expressed pseudogene, it could represent an endogenous competitor of the GBA mRNA, acting as a "sponge" for miRNAs targeting both GBA and GBAP1 transcripts.

Figure 1.3|GBAP1 annotated transcripts.
GBAP1 locus retrieved from UCSC Genome Browser. In blue the RefSeq genes, in light blue the UCSC Genes and in red the Ensembl Gene Predictions.

1.6 Next-generation sequencing and PD

The recent development of Next-Generation Sequencing (NGS) technologies allows the accurate determination of nearly all protein-coding sequence variants in an individual (the so-called whole-exome sequencing, WES) [Ng et al., 2009] or even the sequencing of the entire genome (whole-genome sequencing, WGS) [Foo et al., 2012]. While there has been a considerable debate in the scientific community whether to favour WES or WGS, a number of considerations still support the WES approach: i) WES costs at least five times less than WGS; ii) the size of WES data per patient are far less than WGS data, resulting in reduced processing time and less burden in terms of data storage; iii) the functional annotation of coding variations streamlines the prioritization of damaging variants; iv) the vast majority (85%) of mutations with large effects on disease-related traits is located within coding regions.

Indeed, the WES approach has been already effectively applied to a number of Mendelian disorders [Bamshad et al., 2011], proving that exome sequencing of a small number of related affected individuals is a powerful and efficient approach to discover novel disease genes. This strategy was also successfully applied to point out novel genes involved in autosomal dominant familial PD (such as VPS35 and DNAJC13) [Zimprich et al., 2011; Vilariño-Güell et al., 2011; Vilariño-Güell et al., 2014], whereas DNAJC6 and SYNJ1 were shown to cause autosomal recessive juvenile parkinsonism [Edvardson et al., 2012; Krebs et al., 2013; Quadri et al., 2013].

The identification of genes linked to familial PD has provided crucial insights into the molecular mechanisms of PD pathogenesis and identified possible targets for therapeutic intervention. In particular, the pathogenic mechanisms highlighted by the genes identified till now place the ubiquitin-proteasome system dysfunction, lysosomal system dysfunction, oxidative stress, and mitochondrial dysfunction at centre stage [Obeso et al., 2010]. Despite these advancements, the majority of the expected PD heritability is still missing; indeed, to date, the underlying cause has been identified in only ~40% of all familial PD cases. Undoubtedly, the discovery of novel PD−causative genes will dramatically further our knowledge on the cellular pathways that lead to neurodegeneration and may point to novel therapeutics targets for the treatment of PD.

2 | Aim of the project

Parkinson's disease is the second most common neurodegenerative disorder, affecting approximately 0.3% of the general population and 1% of people over the age of 75. Considering its prevalence in the population, implications of the costs for treatment (which is only symptomatic), diagnosis, and supporting regimens are a heavy burden for Public Health Institutions, making each improvement in the field desirable. PD is a complex disorder caused by the combination of so-far largely unidentified environmental factors and of predisposing susceptibility genetic components. In recent years, considerable progresses have been made in the identification of the genetic components of PD: several genes were shown to cause rare monogenic forms of PD, with autosomal-dominant (SNCA, LRRK2, VPS35) or autosomal-recessive (parkin, PINK1, DJ1) inheritance. Moreover, a large number of "common" predisposing genetic variants, each with a modest effect, have been associated with sporadic PD by genome-wide association studies (GWAS); among them, GBA seems to be the main genetic risk factor for the disease. These results strongly supported the idea that sporadic and familial PD share multiple genetic risks, and also pathogenic pathways. Despite these advancements, the majority of the expected PD heritability is still missing.

In this frame, my PhD project was mainly focused on better understanding disease pathogenesis through the identification of novel genetic determinants of PD and the in-depth study of GBA expression regulation as a necessary step to gain novel insights on the most important genetic component of the disease. I pursued these tasks by following two lines of research: one aimed at dissecting GBA gene expression regulation, focusing in particular on the role of GBAP1, a GBA pseudogene, in modulating GBA expression through an RNA-based post-transcriptional regulatory network, with the view to highlight novel strategies to augment GCase activity in PD patients. The second line of research was aimed at identifying novel genes responsible for familial PD, and potentially also implicated in sporadic PD. To this aim I have exploited a widespread approach, consisting on the whole-exome sequencing of selected PD families, followed by candidate genes prioritization and functional validation. The identification of novel genes responsible for rare Mendelian forms of PD may have a substantial impact in understanding disease pathogenesis by pointing to novel pathways involved in the disease and by identifying novel molecular targets, a crucial step in the design of innovative (molecular) mechanism-driven therapies.

3 | Results & conclusions: GBA regulation

3.1 Post-transcriptional regulation of GBA

The pathogenic mechanism underlying the association between mutations in the GBA gene and the development of PD remains to be clarified. Starting from the evidence that a widespread deficiency of GCase activity was demonstrated not only in PD patient brains with mutations in the GBA gene, but also in patients without mutations, we decided to evaluate the possible mechanisms that could alter GBA expression levels. In particular, we decided to start from the post-transcriptional regulation of this gene, searching for miRNAs targeting GBA.

3.1.1 MiRNA selection

We used different software to generate a list of miRNA candidates: we prioritized the ones predicted by at least five software and known to have a potential role in neurodegenerative diseases, or that were previously implicated in neuronal development. From this analysis were selected as possible candidates 3 miRNAs: miR-22-3p, miR-132 and, miR-212 (Table 3.1).

Table 3.1

miRNA (chromosomal position)*	**miRNA:GBA pairing****	**Brain expression*****
miR-22-3p (chr17:1,617,197-1,617,281)	3' ugucaagaagUUGACCGUCGAa 5' miR-22 || |||||||| 5' cagccaggaaAAAUGGCAGCUc 3' GBA	Cerebellum (643) frontal cortex (219)
miR-132 (chr17:1,953,202-1,953,302)	3' gcuGGUACCGACAUCUGACAAu 5' miR-132 ||| ||| |||||||| 5' ggcCCAAAACUGGAGACUGUUu 3' GBA	Cerebellum (203) frontal cortex (818)
miR-212 (chr17:1,953,565-1,953,674)	3' ccGGCACUGACCUCUGACAAu 5' miR-212 || |||||||||||| 5' gcCCAAAACUGGAGACUGUUu 3' GBA	Cerebellum (22) frontal cortex (70)

* According to UCSC genome browser (http://genome-euro.ucsc.edu/index.html) on Human Feb. 2009 (GRCh37/hg19) Assembly.
** The RNAhybrid software (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/) was used to visualize the miRNA: mRNA interactions.
*** According to deep-sequencing data available through miRBase (http://www.mirbase.org/); numbers in brackets refers to read counts per million of RNAseq experiments (annotation confidence: high for miR-22-3p and miR-132; not reported for miR-212).

MiR-22-3p is located on chromosome 17 p13.3 in an intron of the non-coding gene MIR22HG (host miR-22 gene). Recent studies have associated miR-22-3p with Huntington's disease and Alzheimer's disease, highlighting its protective role when over-expressed [Jovicic et al., 2013]. MiRNAs-132 and 212 belong to the same gene cluster on chromosome 17 p13.3, and they are characterized by seed regions with similar sequences; consequently, they may share common targets [Wanet et al., 2012]. Recent

studies have highlighted some important processes in which the two miRNAs are involved, especially in the brain, confirming their role in synaptic plasticity and neural development [Tognini et al., 2011]. Their expression is down-regulated in mouse models of PD; they also appear to be involved in psychiatric disorders, like Huntington's disease, Alzheimer's disease, and autism [Cogswell et al., 2008; Packer et al., 2008; Talebizadeh et al., 2008; Kim et al., 2007; Wanet et al., 2012].
The real-time RT-PCR method used for the quantitative analysis of selected miRNAs (based on a polyadenylation and a universal reverse transcription approach) did not allow us to reliably quantify miR-212 levels, at least in our cellular models (data not shown).

3.1.2 GBA and GBAP1 are targets of miR-22-3p

To verify if GBA and GBAP1 are target of miR-22-3p or miR132 we started evaluating the effect of microRNA over-expression in HeLa cells. To this end we cloned the precursors of both microRNAs into the expression plasmid for eukaryotic cells psiUx (kindly provided by prof. Bozzoni, Università di Roma "La Sapienza") [Denti et al., 2004]. After transfection experiments, the levels of endogenous GBA and GBAP1 expression were measured by real-time RT-PCR assays. The obtained results showed that miR-22-3p over-expression (50-fold increase) has a significant effect on the levels of GBA an GBAP1 transcripts, while the over-expression of miR-132 (180-fold increase) has no detectable effects. In particular, concerning miR-22-3p, the levels of GBA and GBAP1 are both reduced after 24 hours of transfection, respectively by 72% ($p=0.0003$) and 64% ($p=0.0002$) (Figure 3.2 A).

To confirm these results we set up cotransfection experiments using the vector expressing the miRNAs and a vector containing the GBA or GBAP1 3'UTRs cloned in the vector psiCHECK-2 (Promega), downstream of the renilla luciferase gene. The recombinant constructs were transfected into HeLa cells in the presence of psiUx empty vector (mock) or containing the sequence of pre-miRNAs. The results of transfection experiments confirmed that GBA and GBAP1 3'UTRs are responsive to miR-22-3p. On the contrary, and in accordance with previous results, no significant reduction was observed in the case of miR-132, further confirming that it has no effect on GBA or GBAP1 (Figure 3.2 B).

In order to validate the results obtained on miR-22-3p, these same experiments were repeated in a different cell line; in particular, we chose HEK293 cells for their low amount of endogenous miR-22-3p and the well detectable levels of GBA and GBAP1 (Figure 3.1). The obtained results confirmed that miR-22-3p over-expression (130-fold increase) leads to an about 50% reduction of both GBA and GBAP1 transcripts levels. To check whether modulation by miR-22-3p was detectable also at the protein level, the HEK293 cells, transfected for 48 hours with psiUx-miR-22-3p or with the empty plasmid, were used to measure the GCase enzymatic activity. All measurements were performed in collaboration with the laboratory of Dr. Massimo Aureli (Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano). The results of these experiments showed a

reduction (15% with a p-value of 0.0017) of GCase activity in miR-22-3p overexpressing cells (450-fold increase) compared to samples transfected with the empty plasmid (Figure 3.2 C).

We also replicated the luciferase experiments in HEK293 cells and, to evaluated if the region recognized by the miRNA in GBA and GBAP1 3'UTRs (miRNA recognition element, MRE) was the one predicted by bioinformatics analysis, we produced mutant constructs lacking the putative miR-22-3p binding site and then repeated the transfection experiments. The transcripts without the MRE were not responsive to miR-22-3p, suggesting that the observed down-regulation of GBA and GBAP1, as a result of miR-22-3p over-expression, was due to a direct miRNA-mRNA interaction (Figure 3.2 D).

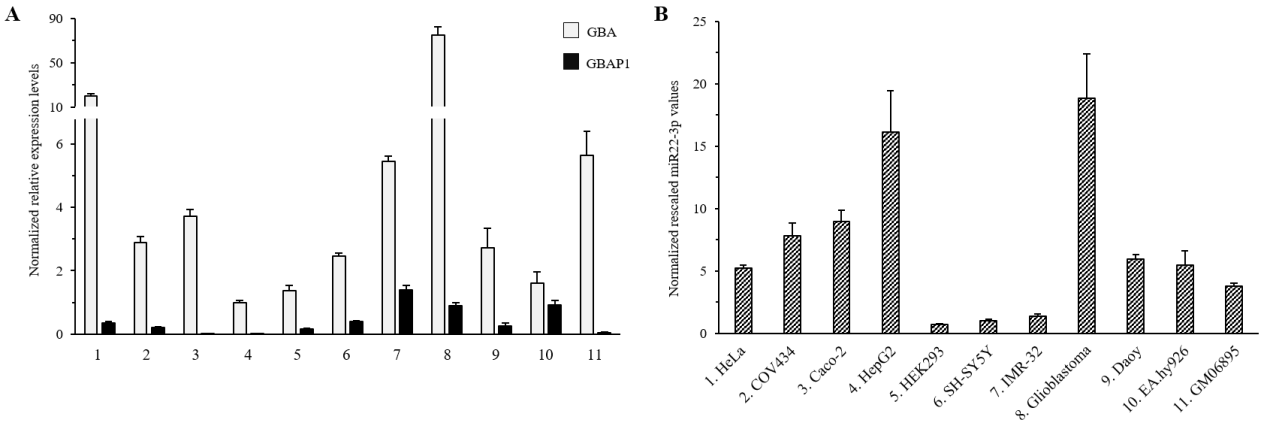


Figure 3.1|GBA, GBAP1 and miR-22-3p expression levels in 11 cell lines.

A. GBA and GBAP1 expression levels were measured by real time RT-PCR assays on RNA derived from 11 cell lines. HMBS transcripts was used as internal reference. **B.** MiR-22-3p expression levels were measured by real time RT-PCR assays (based on a polyadenylation and a universal reverse transcription approach) using the SYBR Green chemistry on RNA derived from 11 cell lines. U6 snRNA was used as internal reference. Results are presented as normalized rescaled values, setting as 1 the values of the SH-SY5Y line. Bars represent means + SD of three replicates.

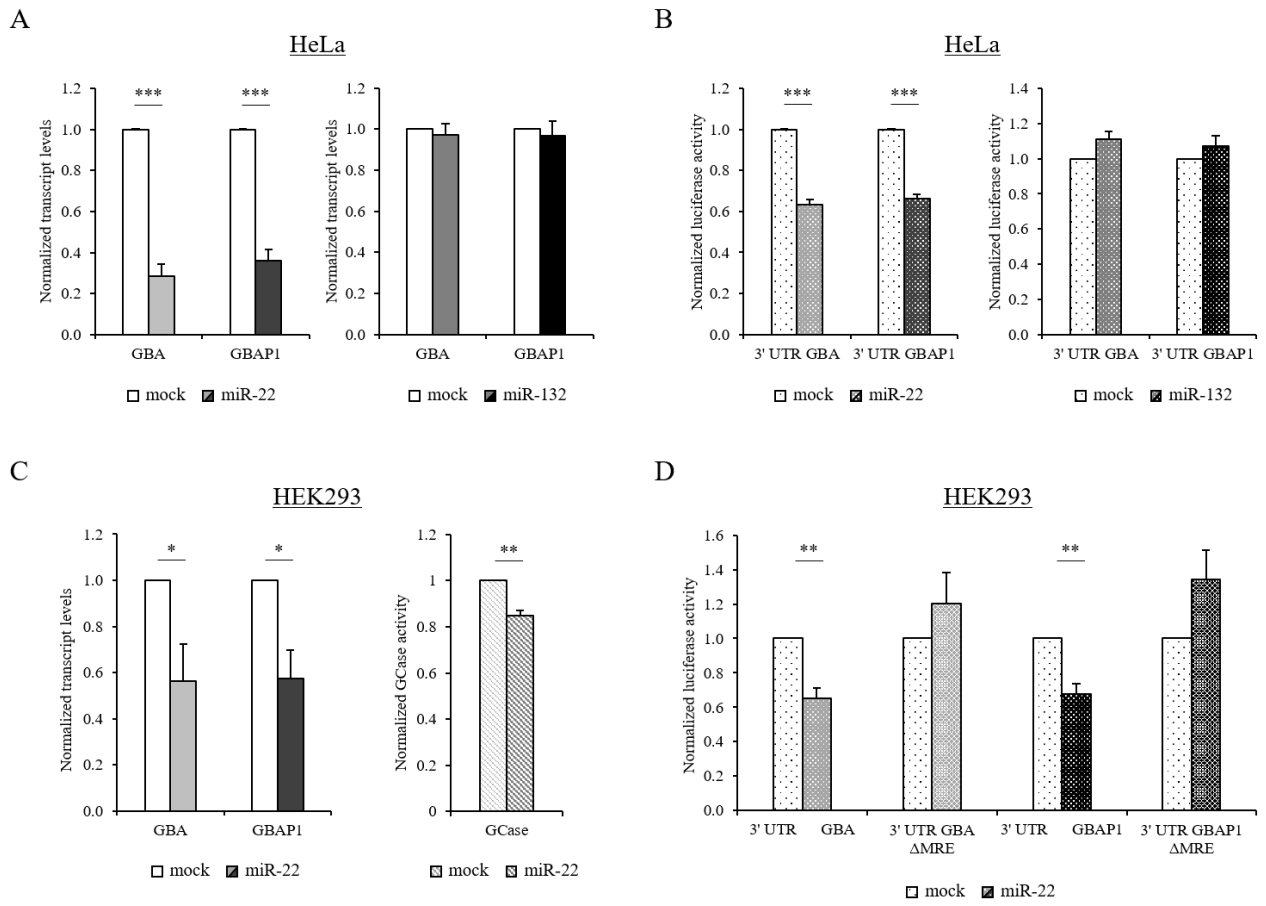


Figure 3.2| GBA and GBAP1 are targets of miR-22-3p.

A. HeLa cells were transfected with 3.5µg psiUX plasmid expressing either miR-22-3p or miR-132 precursors or the empty vector. 24h later, RNA was extracted and the GBA and GBAP1 levels were measured by real time RT-PCR. HMBS transcripts were used as internal reference. Expression levels are shown as normalized rescaled values, setting as 1 the value measured in cell transfected with an empty vector (psiUX, mock). **B.** The psiCHECK2 vectors, containing either the GBA or the GBAP1 3'UTR downstream of the luciferase reporter gene, were independently transfected in HeLa cells together with plasmids expressing either the pre-miR-22-3p or the pre-miR-132. After 48h, the luciferase activity was assayed in cell extracts. To normalize the renilla luciferase values, the firefly luciferase levels were used. Expression levels are shown as normalized rescaled values, setting as 1 the value measured in cell transfected with an empty vector (psiUX, mock). **C.** HEK293 cells were transfected with 875ng of psiUX plasmid expressing miR-22-3p, miR-132 precursors or the empty vector. Left panel, 24h later, RNA was extracted and the GBA and GBAP1 levels were measured by real time RT-PCR. HMBS transcripts were used as internal reference. Right panel, 48 hours after transfections, cells were collected for endogenous GCase activity measurements. In all cases, the results are shown as normalized rescaled values, setting as 1 the value measured in cell transfected with an empty vector (psiUX, mock). **D.** Luciferase reporter assays were repeated in HEK293 cells, by transfecting the psiCHECK2 vector coupled to the 3'UTR regions of GBA or GBAP1, with or without the putative miRNA recognition element (ΔMRE). Each of the four psiCHECK2 recombinant plasmid was cotransfected with the psiUX plasmid expressing miR-22-3p. 48 hours after transfection, cells were collected and lysates prepared to perform the reporter assays. To normalize the renilla luciferase values, the firefly luciferase levels were used. Expression levels are shown as normalized rescaled values, setting as 1 the value measured in cell transfected with an empty vector (psiUX, mock). In all panel, bars represent means + SEM of at least 3 independent experiments each performed at least in triplicate. The results were analyzed by unpaired t-test. *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$

3.1.3 GBAP1 acts as a GBA ceRNA

In order to evaluate the possible role of GBAP1 as a GBA ceRNA, we initially analyzed the expression levels of GBA, GBAP1 and miR-22-3p in a commercial panel of 20 human tissues and 24 different brain areas. The analysis was performed by real-time RT-PCR and the results showed that the pseudogene is expressed at well-detectable levels, although always lower than those of GBA (Figure 3.3), in all analyzed districts; the average ratio between the expression levels of the GBA/GBAP1 was equal to 70:1 (given that it oscillates between a minimum of 11:1 in the thymus and a maximum of 280:1 in the kidney). These results suggest that also the expression levels of the pseudogene may be differentially regulated between the different tissues.

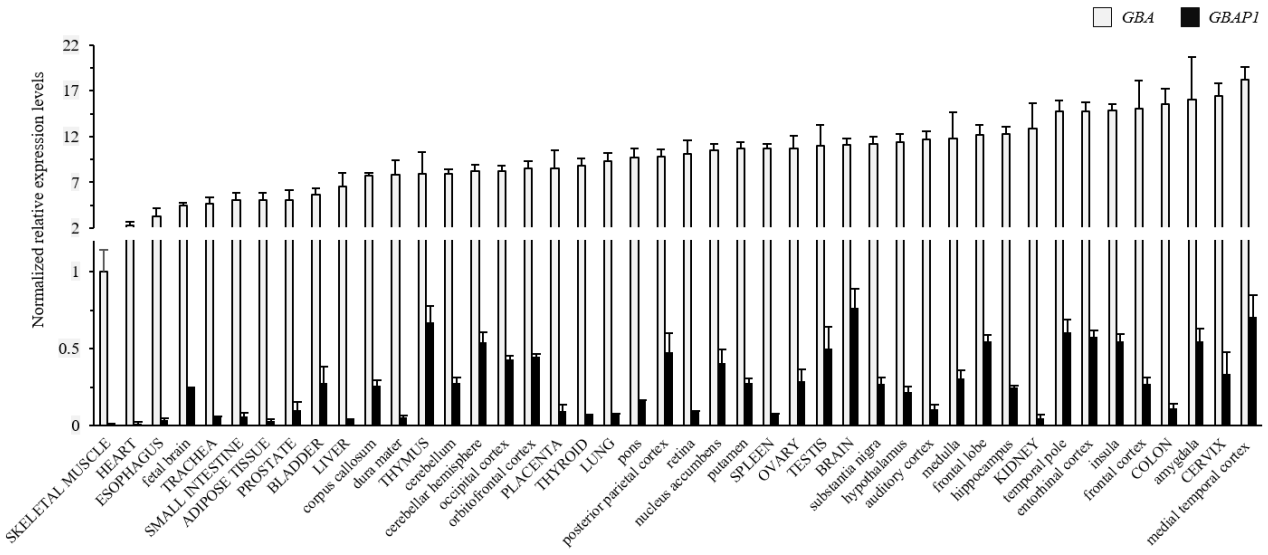


Figure 3.3| GBA and GBAP1 expression pattern.
GBA and GBAP1 expression levels were measured by real time RT-PCR assays using the SYBR Green chemistry on RNA derived from a commercial panel of 20 human tissues (upper-case letters) and of 24 brain areas (lower-case letters). HMBS transcripts were used as internal reference. Results are presented as normalized rescaled values, setting as 1 the GBA values of the SH-SY5Y line. Bars represent means + SD of three replicates.

After demonstrating that GBAP1 and GBA are targets of miR-22-3p, we focused our attention on the pseudogene, in order to assess its possible role as ceRNA in the regulation of GBA.

We performed transfection experiments over-expressing the 3'UTR of GBAP1 in HepG2 cells, a cell line expressing high levels of miR-22-3p and low levels of GBAP1 (Figure 3.1). We used as controls the empty vector and the MRE deleted 3' UTR. The endogenous levels of GBA were measured by real-time RT-PCR on total RNA extracted from cells 96 hours after transfection. The obtained results showed an increase in the expression levels of GBA transcripts (1.68 fold; $p=0.0016$) when the pseudogene full length 3'UTR was over-expressed (Figure 3.4 A). According to the ceRNA hypothesis, all transcripts that share the same miRNA response elements should be able to modulate each other [Salmena et al., 2011; Karreth and Pandolfi, 2013]. We therefore evaluated the effect of GBAP1 3'UTR over-expression on SP1 (Specificity Protein 1) and SIRT1 (Sirtuin1) transcripts, two validated targets of

miR-22-3p [Xu et al., 2011]. Moreover, we measured the levels of a transcript not predicted to be targeted by miR-22-3p: CELF1 (CUGBP, Elav-Like Family Member 1), a well-detectable ubiquitously-expressed mRNA, chosen for its long 3' UTR (6205bp) in order to take into account any nonspecific effects. As expected, the over-expression of the full length 3'UTR of GBAP1 induced a significant increase of about 1.7 ($p<0.015$) of SP1 and SIRT1 transcripts, while the expression levels of CELF1 did not significant change (Figure 3.4 A). These data support the hypothesis of a reciprocal regulation among transcripts sharing miR-22-3p as a post-transcriptional regulator.

We also tested the role of GBAP1 as a ceRNA at the protein level, measuring the enzymatic activity of GCase in the same transfection experiments. The enzymatic assay results showed a significant change in the GCase activity, which was increased by approximately 1.11 times in cells expressing the full length GBAP1 3'UTR compared to controls ($p=0.013$) (Figure 3.4 B). These results suggest that GBAP1 could be able to act as a molecular sponge for miR-22-3p and to modulate the levels of GBA.

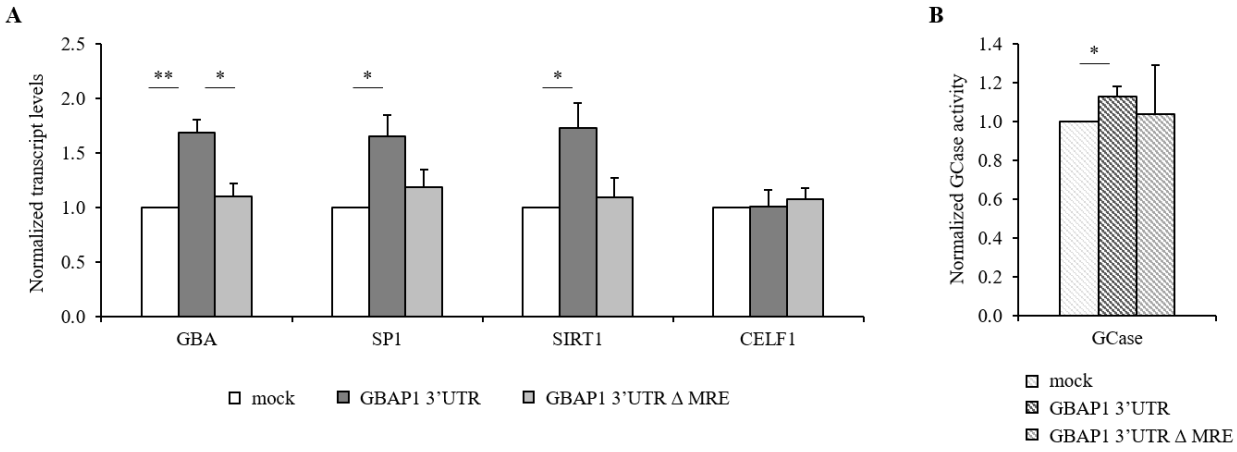


Figure 3.4| GBAP1 may act as a ceRNA for GBA and other miR-22-3p-3p targets.

A. HepG2 cells were transfected with 300ng of the psiCHECK2 plasmid containing the GBAP1 3'UTR with or without the MRE. 96h later, RNA was extracted and the levels of GBA, SP1, SIRT1, and CELF1 were measured by real time RT-PCR. HMBS transcripts were used as internal reference. Expression levels are shown as normalized rescaled values, setting as 1 the value measured in cells transfected with an empty vector (psiCHECK2, mock). **B.** The same overexpression experiments were performed to obtain, 96h after transfection, protein extracts for the measurements of the GCase activity. In all panels, bars represent means +SEM of three independent experiments, each performed at least in triplicate. The results were analyzed by unpaired t-test. *: $p<0.05$; **: $p<0.01$.

3.1.4 GBAP1 splicing pattern and NMD regulation

The information currently available in the literature and in databases on GBAP1 transcripts are rather fragmentary. However, it is known that some of the existing nucleotide differences between GBA and GBAP1 involve splicing sites, and is therefore likely that the mature transcripts of the pseudogene are slightly different from the GBA ones. For this reason, we decided to characterize the splicing profile of GBAP1 by RT-PCR on RNA extracted from HepG2 cells untreated and treated with cycloheximide, a known NMD (nonsense-mediated mRNA decay) pathway inhibitor. The study of GBAP1 splicing pattern is complicated by the high sequence identity between the pseudogene and its cognate protein coding gene; the main difference within the coding sequence is a region of 55 nucleotides deleted in the pseudogene exon 9. We designed RT-PCR assays specific for the GBA and the GBAP1 transcripts anchoring one of the two primers to the 55-bp in exon 9 or on the breakpoint of the deletion, respectively. The transcripts were initially analyzed using 3 separate assays (A, B, C for GBA and D, E, F for GBAP1) (Figure 3.5 A). Agarose gel electrophoresis of the RT-PCR products showed the amplification of a single major band for amplicons A, B, C in each condition, demonstrating that GBA exons are mainly present in all transcripts. Concerning GBAP1, amplicon F was characterized by a single band also when the cells were treated with cycloheximide, indicating the GBAP1 exons 9-11 are not subject to alternative splicing. By contrast, some of the exons upstream may be excluded from the mature mRNA (exon skipping), as shown by the presence of multiple RT-PCR products on the gel (amplicons D and E) (Figure 3.5 A). This remarkable transcript complexity was further characterized by nested or semi-nested RT-PCR assays, which led to the identification of the majority of alternative splicing events (both in-frame and out-of-frame) involving almost all exons (Figure 3.5 B). Some of them are already annotated as predicted transcripts in the UCSC Genome Browser (e.g. the intron 8a retention and the exon 4 skipping), while others are novel such as exon 6 skipping or the splicing events involving the 5' UTR.

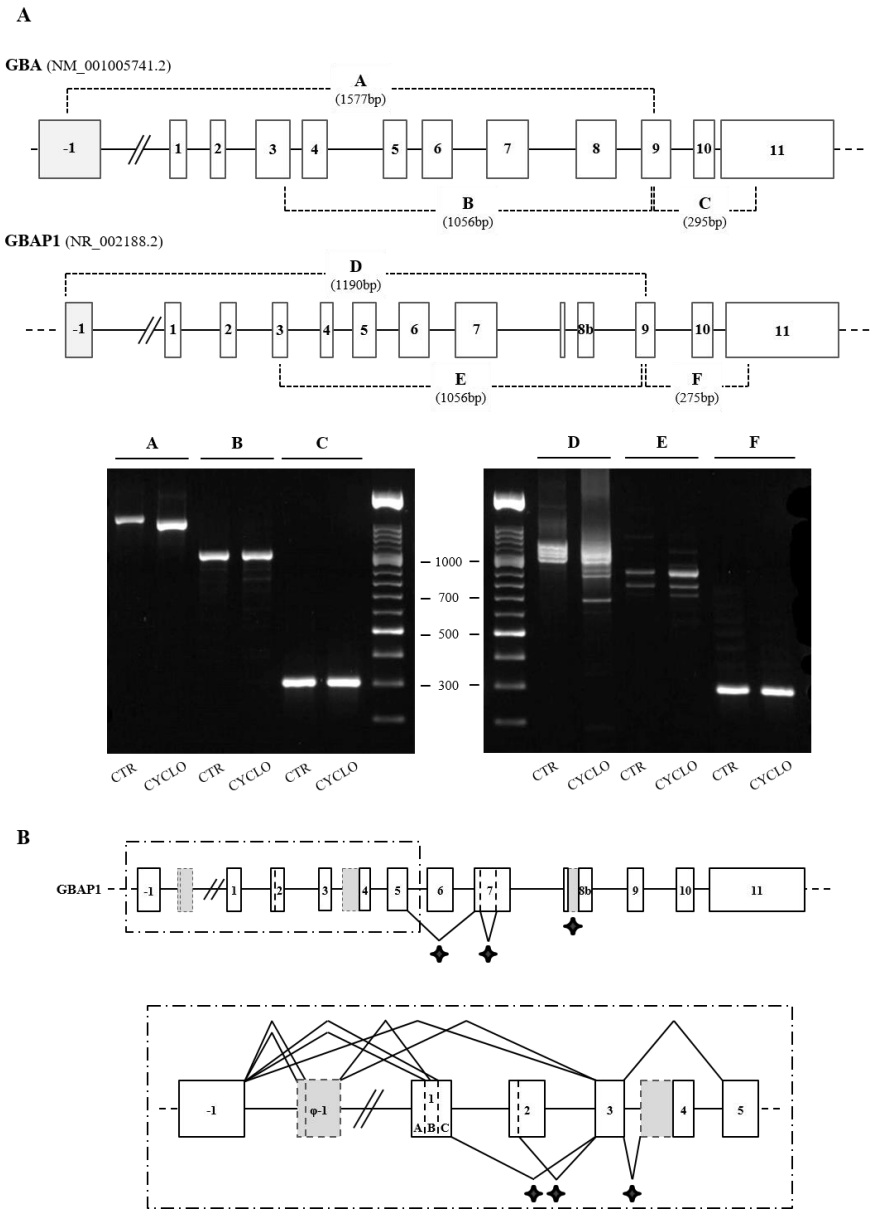


Figure 3.5|GBAP1 splicing profile.

A. In the upper part of the panel, a schematic representation of GBA and GBAP1 genes is reported. Exons are indicated by boxes, introns by line. The fragments amplified by RT-PCRs to analyze the GBA and GBAP1 splicing patterns are indicated by dashed lines and a letter. In the lower part of the panel, the electrophoretic analysis (agarose gels 2%) of RT-PCR amplicons is shown. RT-PCRs were performed on RNA extracted from HepG2 cells treated or untreated with the NMD inhibitor cycloheximide. On the top of each gel, letters indicate the relevant RT-PCR amplicons. **B.** In the upper part of the panel, a schematic representation of the whole gene is presented with broken lines pointing to the major splicing events characterizing the 3' portion of the transcript. Grey boxes indicate the presence of additional exons or extensions of annotated exons. In the lower part of the figure, a magnification of the 5' portion of the gene is reported, with all the identified alternative splicing events. Out-of-frame splicings are indicated in all cases by a star. All identified alternative splicing events were confirmed by Sanger sequencing of the relevant RT-PCR fragment.

Considering the high complexity of the GBAP1 splicing and given that the majority of pseudogene transcripts are characterized by the presence of premature stop codons (PTCs), we decided to check whether the nonsense-mediated mRNA decay could have a role in post-transcriptional regulation of GBAP1. Indeed, it is known that also non-coding transcripts containing PTC, as pseudogenes, can be subject to the NMD mechanism [Mitrovich and Anderson, 2005]. The possible degradation of GBAP1

transcripts due to NMD was investigated by treating HEK293 and HepG2 cells with cycloheximide. After 8 hours of treatment, we extracted total RNA and then performed real-time RT-PCR assays to measure the levels of GBAP1 and GBA transcripts using as reference the gene coding for connexins. The choice of using the connexin transcripts is due to the fact that the coding region of the corresponding genes is contained in a single exon, and therefore they are not subject to NMD degradation. The levels of two different transcripts of PRKCA gene (encoding for the protein kinase C alpha), were used as positive and negative control of the experiment [Paraboschi et al., 2014]. The results obtained from these experiments showed that there is an approximately 4-time increase ($p<0.045$) in the levels of GBAP1 following cycloheximide treatment compared to the untreated control in both cell lines (Figure 3.6). Even if GBA alternative transcripts were not identified in our experiments or reported in literature, it is interesting to observe that, after cycloheximide treatment, even the transcript levels of GBA undergo a modest increase (about 2-fold) compared to the untreated control (Figure 3.6). These data might indicate that the pseudogene GBAP1 is subject to the NMD degradation and suggest a possibly indirect regulation also of the GBA levels by this pathway.

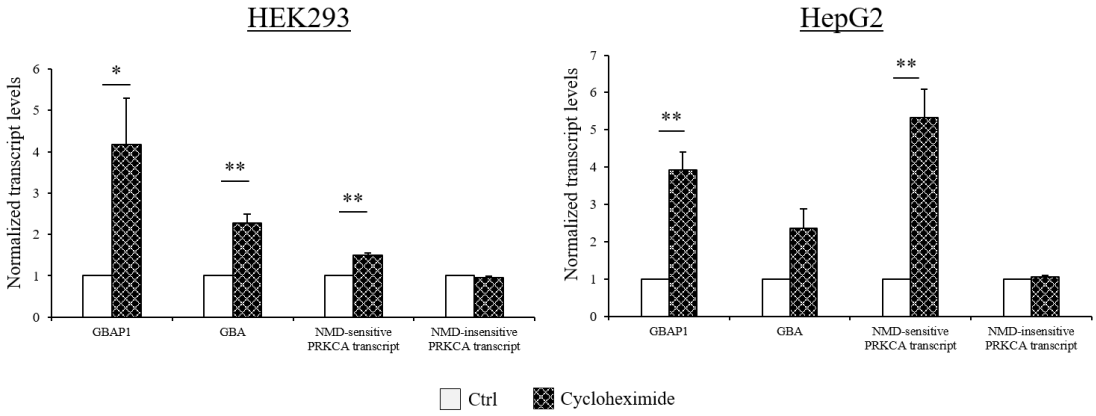


Figure 3.6|GBA and GBAP1 are regulated by the NMD mechanism.
HEK293 (left) and HepG2 (right) cells were treated with cycloheximide for 8h. After the treatment, RNA was extracted and the GBA and GBAP1 levels were measured by real time RT-PCR. Connexin 43 or 32 transcripts were used as internal reference. Expression levels are shown as normalized rescaled values, setting as 1 the value measured in untreated (Ctrl) cells. RT-PCRs performed on out-of-frame and in-frame PRKCA isoforms, known to be respectively sensitive and insensitive to the NMD inhibition were used as positive and negative controls. Bars represent means + SEM of at least 3 independent experiments, each performed at least in triplicate. The results were analyzed by unpaired t-test. *: $p<0.05$, **: $p<0.01$.

3.1.5 GBA, GBAP1, and miR-22-3p are expressed in iPS-derived neurons

To evaluate whether the proposed RNA-based regulatory network may be relevant in PD, we measured the levels of GBA, GBAP1 and miR-22-3p in a more disease-relevant cellular model. In particular, we evaluated their expression levels in iPS and iPS-derived dopaminergic neurons, obtained starting from fibroblasts of controls and of PD patients carriers of a GBA mutation (n=2 L444P, n=2 N370S), kindly provided by Dr. Michela Deleidi (German Center for Neurodegenerative Diseases, University of Tübingen), Dr. Alessio Di Fonzo (Department of Pathophysiology and Transplantation, University of Milan), and Prof. Rejko Kruger (Luxembourg Center for Systems Biomedicine, University of Luxembourg).

In order to assess whether GBA, GBAP1 and miR-22-3p are also coexpressed in dopaminergic neurons, and how their expression could change during neuronal development, we analyzed their transcript levels by real-time RT-PCR, using as internal reference the housekeeping gene HMBS and the snoRNA U43. The obtained results showed that the three transcripts are expressed in dopaminergic neurons, the diseased cells in PD. We also highlighted that GBA mRNA increases significantly during neuronal differentiation (8-fold $p=0.0002$) and that GBAP1 has a similar trend; by contrast, miR-22-3p decreases by almost 2.5 times in neurons ($p=0.018$) (Figure 3.7).

Next we compared the expression levels of GBA, GBAP1 and miR-22-3p in dopaminergic neurons of PD patients and healthy controls, to study possible differences. Concerning GBA, a clear difference is detectable between cases and controls: in PD patients' neurons the levels of GBA are less than half the ones measured in healthy controls ($p=0.0094$). GBAP1 levels seem to be stable, while miR-22-3p is sligthly increased in PD cases (Figure 3.7).

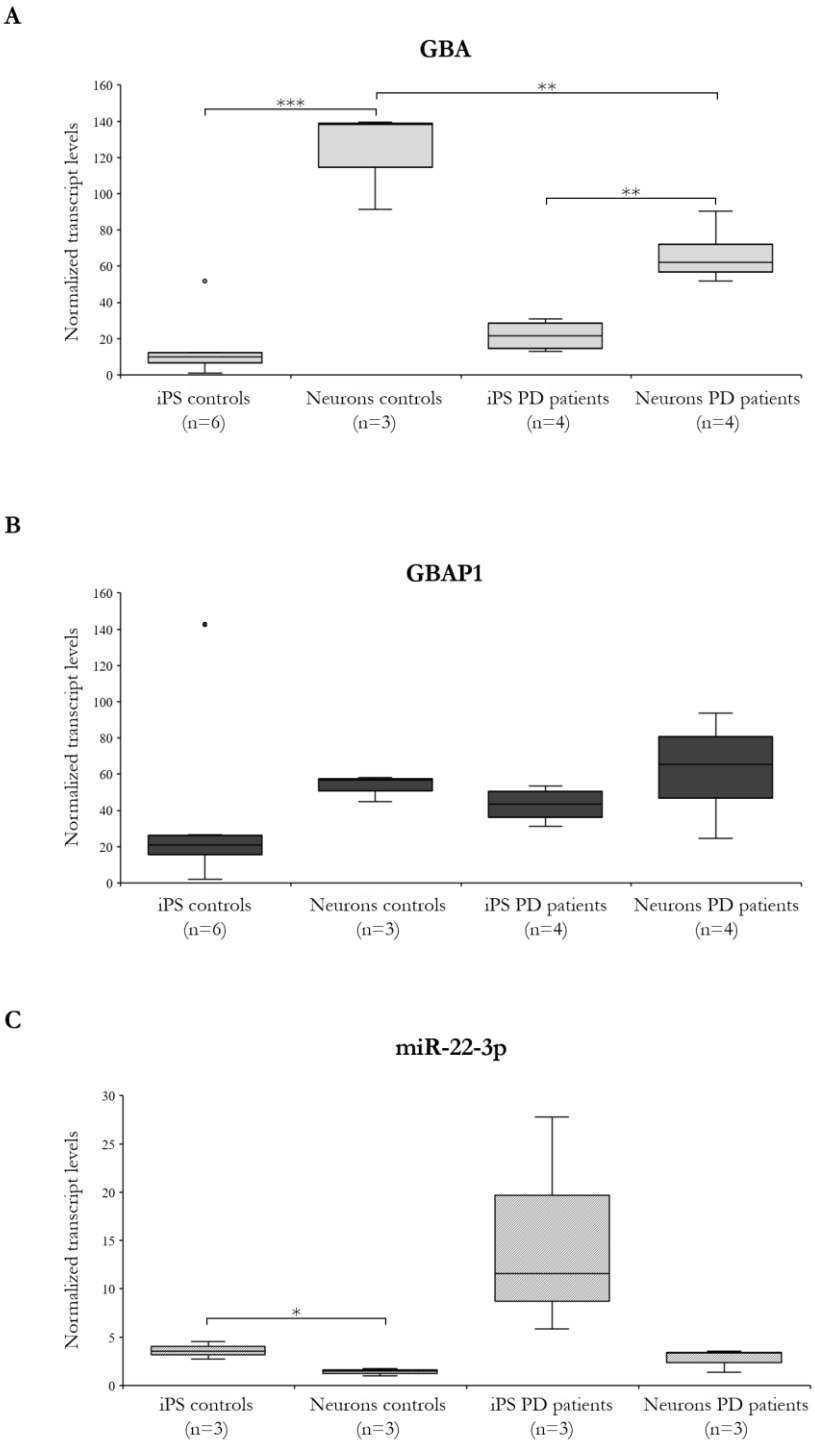


Figure 3.7| GBA, GBAP1, and miR-22-3p are expressed in iPS cells and iPSC-derived neurons.
GBA (panel A), GBAP1 (panel B), and miR-22-3p (panel C) expression levels were measured by real-time RT-PCR assays. Boxplots show expression levels according to the disease status; boxes define the interquartile range; the thick line represents to the median. Results are shown as normalized rescaled values. Significance level for differences between groups was calculated by a t-test, and shown only if significant. *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; n: number of analyzed subjects.

3.2 Transcriptional regulation of GBA and GBAP1

One of the main determinants that could explain the GBA expression variability is definitely the regulation at the transcriptional level. To understand the mechanisms that determine both the physiological levels of GBA and the differences described in the literature in the levels of GBA among PD patients and controls, it is therefore essential to analyze the two known promoters responsible for GBA expression. Furthermore, since GBAP1 might act as a non-coding RNA regulating GBA, also the study of the mechanisms that control the pseudogene expression is of great interest.

GBA is characterized by a promoter (P1), located immediately upstream of exon 1, which is not associated with CpG islands and contains a TATA box and different binding sites for transcription factors [Horowitz et al., 1989].

To date, in the UCSC Genome Browser, five alternative GBA transcripts are described, three of which appear to originate from an alternative promoter (P2), located 2.6kb upstream of the translation start site. All these transcripts contain an additional exon (exon -1), that does not alter the amino acid sequence of the enzyme glucocerebrosidase (Figure 3.8). Only one transcript seems to have an alternative start codon in exon 3. The distal promoter P2 has the characteristics of housekeeping promoters: it presents a CpG island and multiple binding sites for the Sp1 transcription factor, but no TATA or CAAT box [Svobodová et al., 2011].

According to the transcripts and the epigenetic markers annotated in the UCSC Genome Browser, even the pseudogene GBAP1 could present a proximal and a distal promoter.

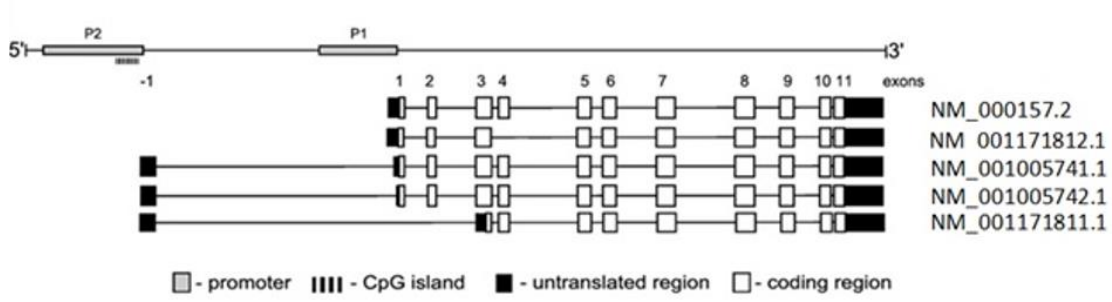


Figure 3.8|GBA alternative transcripts annotated in UCSC Genome Browser.

3.2.1 Characterization of GBA and GBAP1 promoters

To confirm the existence of the two GBAP1 promoters and to evaluate their functionality compared to GBA promoters, we cloned ~1kb of each promoter in the luciferase reporter vector pGL2 upstream of the firefly luciferase gene. These four constructs were then used to transfect different cell lines: HeLa, SH-SY5Y, HEK293, and HepG2.

Measurement of the luciferase activity showed, unexpectedly, that the distal promoter P2 is actually stronger than the promoter P1, for both GBA and GBAP1, in all transfected cell lines (Figure 3.9). In

particular, in HeLa and in HEK293, the GBA P2 promoter is ~12 times stronger than the corresponding promoter P1 (in Hela P2/P1=12.8, p=0.00017; in HEK293 P2/P1=11.7, p=0.03); in HepG2 P2 is 4.45 times stronger than P1 (p=0.02); while, in the neuronal line SH-SY5Y the difference is not significant. Also with regard to GBAP1, the distal promoter is stronger than the proximal one: in HeLa P2/P1=5.5 (p=0.00007), in SH-SY5Y P2/P1=2.4 (p=0.007), in HEK293 P2/P1=7 (p=0.002), and in HepG2 P2/P1=7.3 (p=0.007).

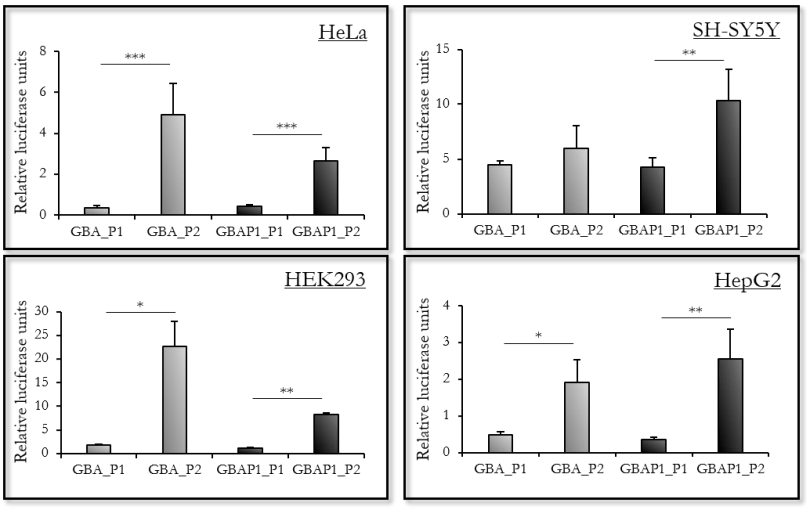


Figure 3.9| Characterization of the GBA and GBAP1 promoters.

In each transfection experiment, 1μg of pGL2 vectors containing the 4 different promoters were transfected in HeLa, SH-SY5Y, HEK293, and HepG2 cells. As controls, also the pGL2 basic, pGL2 promoter and pGL2 control plasmids were transfected. After 48h, the luciferase activity was assayed in cell extracts. To normalize luciferase values, each well was also transfected with 100ng of the pRL-TK vector, expressing the renilla luciferase gene. The reported normalized results were rescaled setting as 1 the values measured in the pGL2 promoter. Bars represent means + SEM of at least 3 independent experiments. The results were analyzed by unpaired t-test. *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$.

3.2.2 Analysis of GBA and GBAP1 promoters by 5'-serial deletions

In order to better characterize these four promoters, we produced serial 5'-deleted constructs of each promoter that were then transfected in SH-SY5Y cells, the most common neuronal cell line, to measure the possible differences in the luciferase expression levels. We created 16 plasmids: four starting from the full length (FL) P1 promoter of both GBA and GBAP1 (ΔA, ΔA+B, ΔA+B+C, ΔA+B+C+D) and three originated from the full-length P2 promoter of the gene and the pseudogene (ΔA, ΔA+B, ΔA+B+C) (Figure 3.10, left panels).

The obtained constructs were co-transfected in SH-SY5Y cells together with the pTK-RL vector, containing the renilla luciferase gene, used as reference and after 48h a luciferase reporter assay was performed. The strength of the GBA P1 promoter was significantly reduced in the construct ΔA+B+C+D, which contains only the last fragment of the P1 promoter: the luciferase activity was 4 times lower compared to the FL construct (p=0.0004) and 13 times compared to the ΔA+B+C vector (p=0.0072). Concerning the P1 promoter of GBAP1, there is a 6.5-fold decrease in luciferase activity as

a result of the deletion ΔA+B+C+D (p=0.0002) compared to the full length and a 4.6-fold decrease compared to the ΔA+B+C construct (p=0.0048). Conversely, for both the P2 promoters of GBA and GBAP1, the deletions did not change significantly the promoter strength (Figure 3.10).

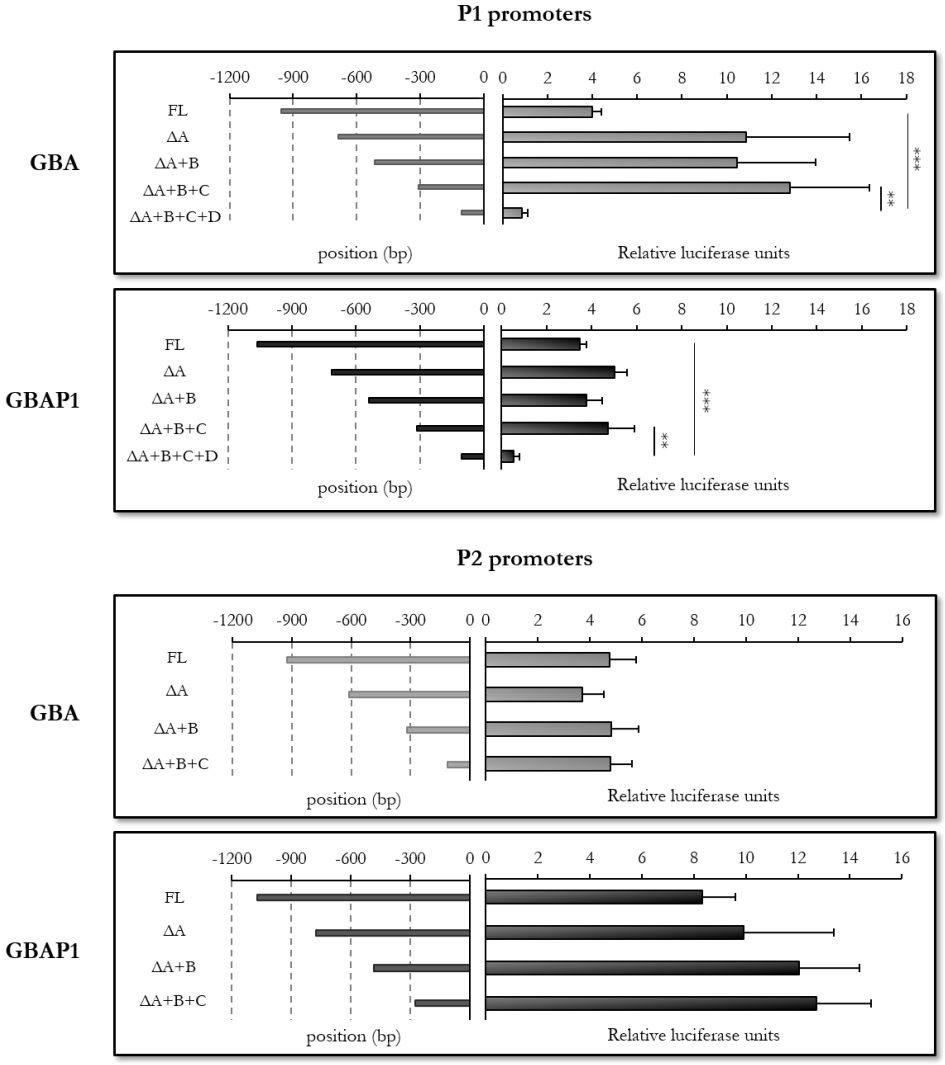


Figure 3.10 | 5'-serially-deleted constructs to analyze the GBA and GBAP1 promoters.
Upper panel, P1 promoters. Lower panel P2 promoters. In each transfection experiment, 1μg of pGL2 vectors containing the full-length and deleted promoters were transfected in SH-SY5Y cells. As controls, also the pGL2 basic, pGL2 promoter and pGL2 control plasmids were transfected. After 48h, the luciferase activity was assayed in cell extracts. To normalized luciferase values, each well was also transfected with 100ng of pRL-TK vector expressing the renilla luciferase gene. The reported normalized results were rescaled setting as 1 the values measured in the pGL2 promoter. Bars represent means + SEM of at least 3 independent experiment. The results were analyzed by unpaired t-test. **: p<0.01, ***: p<0.001.

3.2.3 P1 promoters of both GBA and GBAP1 are characterized by two CLEAR elements

The data obtained from transfection experiments of deleted constructs pointed to the last deletion, in the proximal promoter of the gene and the pseudogene as the most interesting one. Using the prediction software Genomatix Suite, we searched for possible transcription factor binding sites in this region. This analysis predicted the presence of particularly interesting sequences located at positions -118 and -87, i.e. two CLEAR (Coordinated Lysosomal Expression and Regulation) elements. These

sequences are recognized by TFEB (Transcription Factor EB), a transcription factor that regulates the expression of many lysosomal genes [Palmieri, 2011]. The presence of CLEAR elements in the promoter of the GBA gene and the hypothesis that these sequences are involved in TFEB-mediated GBA regulation has been previously described by Sardiello et al in 2009. In addition, the overexpression of TFEB is able to increase the GBA expression levels and to improve the GCase folding and transport to the lysosome [Song et al., 2013]. Interestingly we found that two predicted CLEAR elements are present also in the GBAP1 P1 promoter, in the same position.

To confirm the importance of the two CLEAR elements in the P1 promoters of GBA and GBAP1 we produced three additional constructs for each P1 promoters, corresponding to the full length promoter lacking one or both CLEAR elements (see Figure 3.11, left panels). These vectors were used to transfect the SH-SY5Y cell line. We observed that only the absence of both CLEAR elements completely abolishes the promoter activity.

We confirmed these results cotransfecting a commercial plasmid containing the TFEB cDNA with the CLEAR deleted constructs and showing that each CLEAR element is able to respond to TFEB with an increase in the luciferase activity comparable to that of the full length one (Figure 3.11). These results confirmed that this transcription factor is the main regulatory element of GBA expression and surprisingly demonstrated that also GBAP1 could be regulated by TFEB.

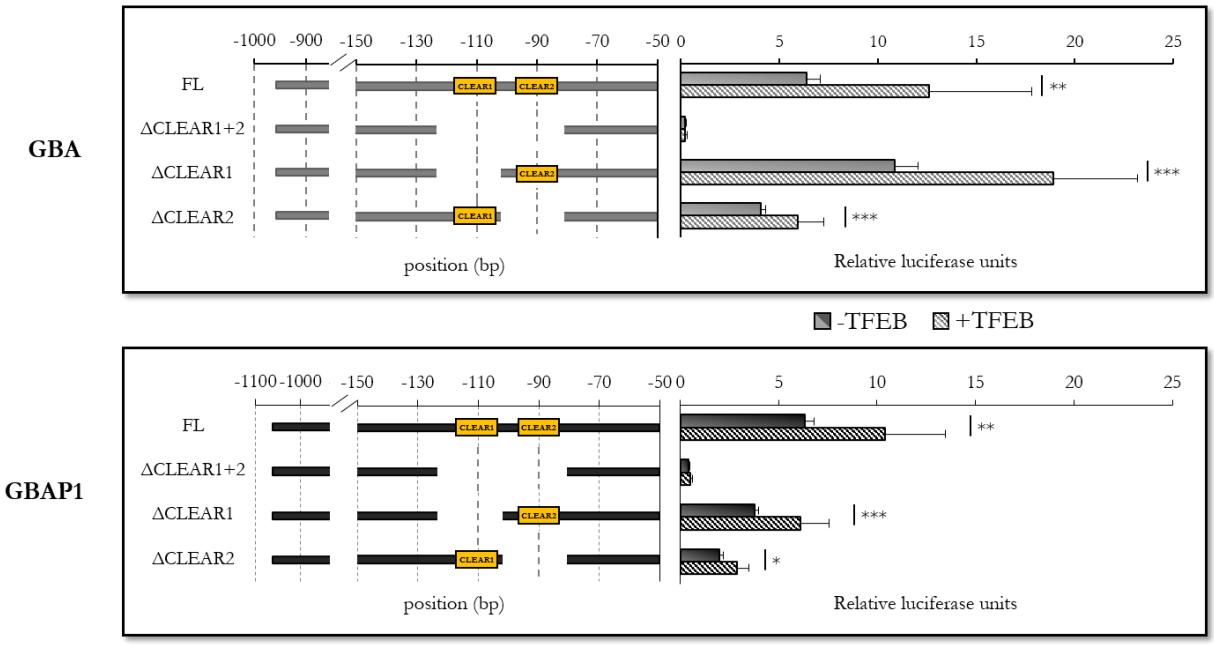


Figure 3.11|P1 promoters have two active CLEAR elements

In each transfection experiment, 1µg of pGL2 vectors containing the full length and deleted promoters were cotransfected in SH-SY5Y cells with 400ng of the pEGFP-TFEB plasmid (+TFEB) or the empty vector (-TFEB). As controls also the pGL2 basic, pGL2 promoter and pGL2 control plasmids were transfected. After 48h, the luciferase activity was assayed in cell extracts. To normalize luciferase values, each well was also transfected with 100ng of pRL-TK vector expressing the renilla luciferase gene. The reported normalized results were rescaled setting as 1 the values measured in the pGL2 promoter. Bars represent means + SEM of at least 3 independent experiment. The results were analyzed by unpaired t-test. *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$.

3.3 Conclusions and future perspectives

Parkinson's disease is a complex disorder with multifactorial etiology, which can result from the contribution of different genetic determinants and multiple environmental factors. GWAS and meta-analyzes have showed the importance of the GBA gene in the development of PD [Sidransky et al., 2009]. Numerous studies have confirmed that mutations in GBA are more frequent in patients with PD compared to healthy controls, showing that carriers of heterozygous mutations in this gene have a 5 times higher risk to develop PD than the general population [Sidransky et al., 2009; Asselta et al., 2014]. However, to date, the mechanism underlying the relation between GBA mutations and the development of PD remains unclear and it is a cause of debate in the scientific community. In fact, there are reports in the literature supporting a gain-of-function effect of the mutated protein, as well as publications sustaining a loss-of-function mechanism [Sidransky and Lopez 2012; Swan and Saunders-Pullman, 2013]. Moreover, a widespread deficiency of GCase activity has been demonstrated not only in the brains of PD patients carrying GBA mutations, but also PD patients without GBA mutations exhibit a deficiency of GCase in SN [Gegg et al., 2012]. These data suggest that a dysfunction of GCase activity or more in general of the lysosomal compartment, may participate in the pathogenesis of the disease even in the more frequent sporadic forms of PD.

This hypothesis stimulated the interests in the development of new therapeutic approaches aimed at augmenting the GCase activity [Schapira and Gegg, 2013]. It has already been showed that the overexpression of GBA, obtained by adeno-associated virus infection, is able to improve the cognitive deficit and to reduce the α-synuclein aggregates in a mouse model of synucleinopathy caused by a homozygous mutation in the GBA gene [Sardi et al., 2011; Sardi et al., 2013]. The study of the mechanisms that could regulate the levels of GCase is therefore an important prerequisite to better understand the involvement of this genetic risk factor in the disease, and to identify new therapeutic targets.

In this thesis, we describe a novel ceRNA-based network with a potential impact on Parkinson's disease involving GBA, its pseudogene GBAP1, and miR-22-3p (Figure 3.12).

Competing endogenous RNAs are a class of RNA regulators, recently described, that, working as molecular "sponges", are able to influence the expression levels of specific mRNAs by competing for a limited pool of miRNAs [Seitz, 2009; Poliseno et al., 2010]. In particular, two classes of lncRNAs are increasingly recognized as main ceRNA contributors: circular RNAs and pseudogene-derived transcripts [Thomson and Dinger, 2016]. Indeed, expressed pseudogenes, are considered ideal ceRNA candidates, since they have a high sequence identity with the corresponding ancestral gene and so they usually share the same miRNA binding sites [Tay et al., 2014; Thomson and Dinger, 2016]. To date, few pseudogenes have been experimentally demonstrated to act as ceRNAs, including: PTENP1 and

KRAS1P [Poliseno et al., 2010], OCT4-pg4 [Wang et al., 2013], BRAFP1 [Karreth et al., 2015], and CYP4Z2P [Zheng et al., 2015].

Concerning GBAP1, the GBA pseudogene, literature data are very limited and are primarily focused on the origin of this locus and the differences with the functionl protein-coding gene [Horowitz et al., 1989; Wafaei and Choy, 2005]. For this reason, we evaluated the expression levels of this pseudogene in different tissues and brain areas. The results obtained showed that the levels of GBAP1 are always lower than the GBA levels but also that this ratio is very variable among the different districts, suggesting that GBAP1 expression could be regulated.

To date there are no miRNAs reported in the literature to directly target GBA; only a miRNA mimic screening was described involving 875 miRNAs that evidenced three candidates (miR-127-5p, miR-16-5p, and miR-195-5p) with important consequences on the GCase activity. However, in all cases, the miRNA effect did not seem to be mediated by a direct interaction between the miRNA itself and GBA. In fact, the identified miRNAs acted either on the LIMP-2 receptor, which is involved in the trafficking of GCase from the endoplasmic reticulum to the lysosome, or on the expression levels of known modifiers of the GCase activity [Siebert et al., 2014].

In this thesis work, we showed that miR-22-3p is the first miRNA that could modulate GCase activity targeting GBA and moreover we suggest that this microRNA could also act on GBAP1, downregulating its transcript levels.

To date, few information on a possible involvement of these non-coding RNAs in Parkinson's disease is overtly present in the literature, with the reported dysregulation of miR-22* in Parkinson's disease patients actually referring to the 5p companion of "our" miR-22-3p [Margis et al., 2011]. Interestingly, a very recent paper suggested that miR-22 may exert a neuroprotective effect in Parkinson's disease, as it protects rat pheochromocytoma PC12 cells from 6-hydroxydopamine-induced injury, by modulating the levels of its target gene transient receptor potential melastatin 7 (TRPM7) [Yang et al., 2016]. However, it should be noted that Yang and colleagues in their paper linked miR-22 down-regulation with Parkinson's disease by mistakenly citing the work of Margis and collaborators [2011], which concerned miR-22-5p, as detailed above. Moreover, few studies reported a potential neuroprotective effect of miR-22-3p both in rat models of cerebral ischemia-reperfusion injury and in Huntington and Alzheimer's disease through a reduction in inflammation and apoptosis [Jovovic et al., 2013; Yu et al., 2015]. However, other studies suggested a pro-senescence role of miR-22 in endothelial progenitor cells, in cancer, and in the aging heart and brain [Li et al., 2011; Xu et al., 2011; Jazbutyte et al., 2013; Zheng and Xu, 2014].

The possible existence of a mutual regulation of the GBA and GBAP1 transcripts mediated by miR-22-3p was then investigated through overexpression experiments designed to increase the levels of the pseudogene 3'UTR. This increase is reflected in a significant increase of GBA transcript levels that

seems to be dependent on the presence of the identified miR-22-3p target site on GBAP1 3'UTR. The positive trend of the GCase activity corroborates our assumption. With this work, we therefore hypothesize the existence of a regulatory circuit among transcripts that share miR-22-3p as post-transcriptional regulator (SP1 and SIRT1) and, in particular, we assess a possible functional role of the pseudogene GBAP1 as ceRNA.

Since GBAP1 seems to act as a non-coding RNA with a regulatory significance for the GBA levels, it becomes important to understand the regulation of this transcripts and in particular, if there are physiological mechanisms that could increase the pseudogene levels.

With regard to the post-transcriptional regulation, it was supposed that also the PTC present in non-coding RNA, such as pseudogenes, can direct those molecules to degradation by the NMD pathway [Tay et al., 2014]. In particular, as regards GBAP1, some nucleotide differences between gene and pseudogene are located in the consensus sequence of splicing sites, suggesting the possible production of alternative transcripts potentially NMD target. We therefore decided to extensively analyze the GBAP1 splicing pattern. This analysis allowed the identification of a large variety of alternative transcripts, the majority of which cause the introduction of premature stop codons. The qPCR analysis of the GBAP1 levels in cells treated with NMD inhibitors suggest that this pathway is an important regulator of the GBAP1 expression. In absence of NMD regulation the levels increase up to 4 times compared with the basal levels. It is interesting to note that in these experiments it was also possible to observe an increase in the GBA transcripts levels. Since there are no literature data and we didn't observe for GBA any alternative splicing product potentially targeting the transcript to NMD, the increase of the GBA levels following cycloheximide (a known NMD inhibitor) treatment could represent a further confirmation of the reciprocal regulation of GBA and GBAP1 via microRNAs. In the future, it will be interesting to confirm this hypothesis blocking the NMD pathway in cells knockout for one of the proteins responsible for miRNA maturation, such as DICER.

Concerning the transcriptional regulation of GBAP1, literature data are essentially related to the GBA gene, for which two different promoters are known: a proximal one, placed immediately before the exon 1, and a distal one, 2.6kb upstream of the translation start site. Both promoters are characterized by the presence of several regulatory elements that may be responsible for a differential gene transcription in specific tissues and/or conditions [Svobodová et al., 2011]. According to the observation that GBAP1 predicted proximal and distal promoter sequences show a high level of identity with those of GBA and that epigenetic marks did not substantially differ between the gene and the pseudogene, as inferred from the UCSC Genome Browser ENCODE tracks (http://genome.ucsc.edu/; release Feb. 2009, GRCh37/hg19), also GBAP1 gene appears to present a proximal and distal promoter.

Luciferase experiments confirmed the existence of both GBAP1 predicted promoters and, surprisingly, they showed a transactivating activity comparable to the one of GBA promoters. These results are in agreement with the high level of sequence identity that GBA and GBAP1 share at the level of the promoter elements (described in Svobodová et al., 2011). On the other hand, this observation is in contrast with the evidence that the GBAP1 transcript levels are lower that the GBA ones in all the cell lines analyzed (Figure 3.1) indicating that epigenetic or post-transcriptional regulation mechanisms could have a main role in the pseudogene expression. We also demonstrated by 5'-serial-deletion analysis that the gene and the pseudogene share similar regulatory elements. In particular, we found two CLEAR elements also in the P1 promoter of GBAP1. We then confirmed that the transcription factor TFEB acts on these elements increasing the activity of both GBA and GBAP1 proximal promoters. Considering the similar strength of GBA and GBAP1 promoters it will be interesting in the future to evaluate the epigenetic regulation of these promoters in order to confirm if the low levels of GBAP1 transcripts are mainly due to the NMD mechanism or to a modifiable transcriptional regulation.

In conclusion, considering the results obtained in this thesis, it is possible to hypothesize the existence of an RNA-based complex regulatory network that involves GBA, GBAP1 and miR-22-3p (Figure 3.12). In this network, miR-22-3p (and probably other microRNAs) could modulate the levels of GBA transcripts down-regulating GCase activity. Moreover, probably in specific cells or developmental stages, upregulation of GBAP1, resulting from post-transcriptional or epigenetic regulatory mechanisms, might titrate miRNAs away from the GBA protein-coding transcripts, thus providing a physiologic ceRNA effect. This regulatory circuit has several regulatory points that might represent potential targets to modulate the GCase level.

The future goal of this project will be to replicate similar experiments in iPS-derived neurons in which we already demonstrated that GBA, GBAP1 and miR-22-3p are expressed, in order to evaluate if the modulation of this network could modify the mechanisms of neurodegeneration and for example the accumulation of α-synuclein one of the main PD hallmark.

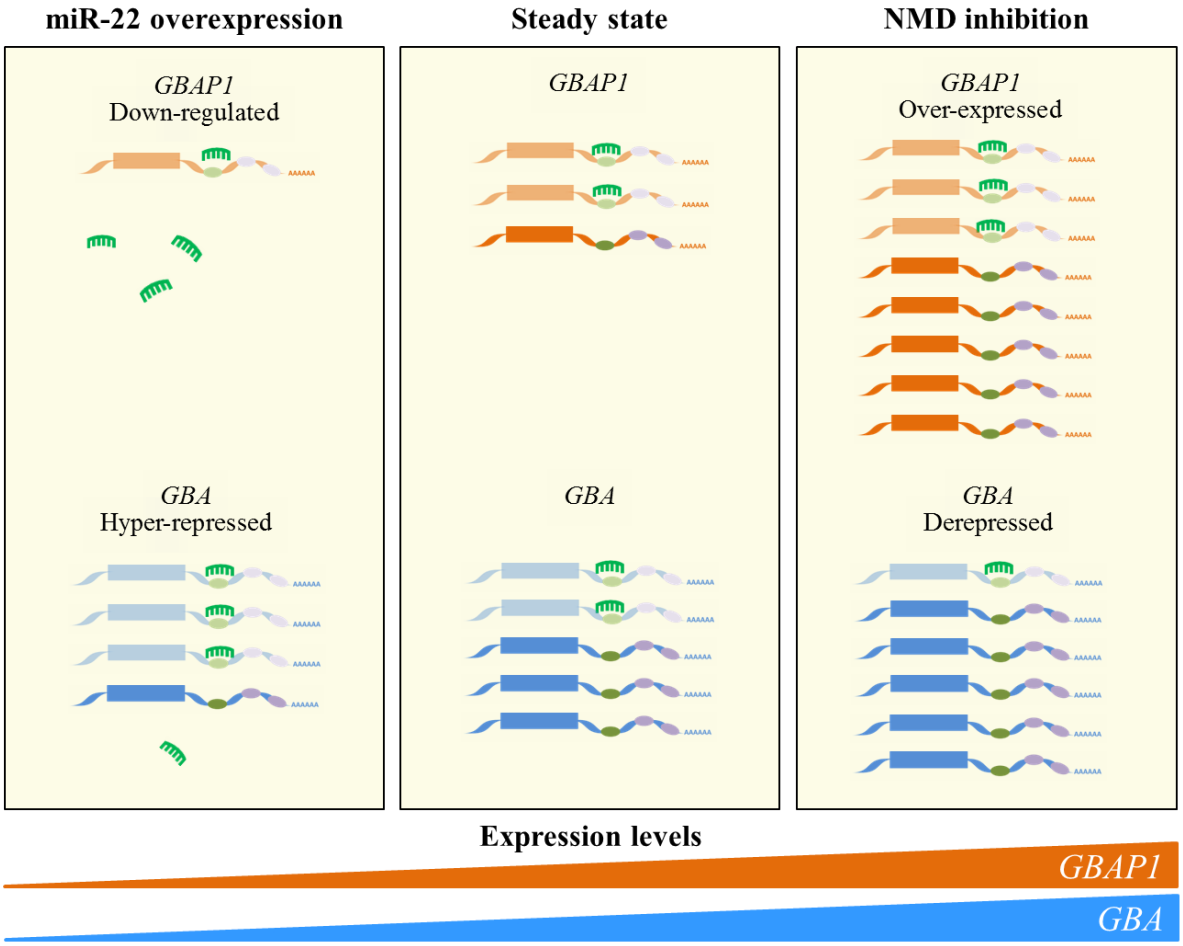


Figure 3.12| Schematic representation of the effect of modulating the GBA/GBAP1/miR-22-3p RNA-based network on endogenous GBAP1 and GBA levels.

Schematic representation of the ceRNA network involving GBA (blue transcripts) and GBAP1 (orange transcripts), harboring the same MRE sites (green and violet ovals). The green MRE sites bind to miR-22-3p (in green), whereas violet ones bind to other not-specified miRNAs. The experimental modulation of the proposed ceRNA network impacts on both coregulated transcripts. In particular, over-expression of miR-22-3p (left part of the figure) determines the down-regulation of both GBA and GBAP1 transcripts. Conversely, over-expression of GBAP1 (e.g., by inhibiting the NMD pathway, as experimentally verified in the present study; right part of the figure) will increase the cellular concentrations of miR-22-3p MREs, thus resulting in the de-repression of GBA. In the scheme, transcripts destined to degradation are colored in lighter shades.

4 | Results & conclusions: WES analysis

A total of 24 Italian families with dominant or recessive PD were selected after the screening for frequent PD causing mutations, in collaboration with the Parkinson Center of Milan.

Among them, 12 were consanguineous and therefore the disease in these families is expected to be the result of homozygous recessive mutations. Since only a small percentage (~1%) of potentially-deleterious homozygous variants within an individual are novel or rare, the number of possible causal variants can be substantially reduced. This is the reason why we decided to perform whole-exome sequencing (WES) only on the proband of each consanguineous family.

Concerning dominant families, one of the main difficulties of using exome sequencing approaches to identify Mendelian traits is the fact that each individual harbors several hundreds of novel heterozygous variants. As a result, the process of identifying the single causal change in a dominant disorder is dependent on methods of filtering out non-causal variants. In order to increase the probability of successfully identifying a causative gene, we have decided to perform WES on the probands and on the available affected cousins or uncles. This approach is currently used to generate shortlists of novel, potentially disease-causing variants for disease segregation/transmission studies within each multi-incident pedigree. This strategy already allowed the identification of a few novel PD genes, among them the two most robustly associated with PD are VPS35 and DNAJC13 [Zimprich et al., 2011; Vilariño-Güell et al., 2011; Vilariño-Güell et al., 2014].

For a first group of exome-targeted enrichment, library preparation and sequencing were performed at the UMass Deep Sequencing Core in collaboration with Prof. Landers (Department of Neurology, University of Massachusetts Medical School, USA). Target capture was performed by an oligonucleotide-based exome capture solution (SeqCap EZ Human Exome Library v2.0) from Nimblegen. The Nimblegen SeqCap EZ Exome Library consists of 2.1 million oligonucleotide probes used to capture ~300,000 exons derived from ~30,000 protein coding. The captured exonic DNA were then sequenced on an Illumina HiSeq2000 as 100 bp paired-end reads.

For the second half of samples, I directly performed the genomic capture, using the AmpliSeq Exome RDY kit from Thermo Fisher Scientific (a PCR-based target selection) and the sequencing as 200-bp single reads with the Ion Proton technology at the Centre of Applied Neurogenetics in Vancouver, during my stay in Prof. Farrer's laboratory. The AmpliSeq Exome RDY kit allows the production of ~294,000 amplicons starting from 12 pools of primers and spanning ~300,000 exons.

We obtained an average coverage of ~90x with a percentage of the covered target regions spanning from 75% (in the samples prepared with the Nimblegen kit) to 90% (in the samples prepared with the AmpliSeq kit). At this level of coverage, the sensitivity of detection of heterozygous variants approaches 100% and the estimated false heterozygote discovery rate is 6 per exome [Choi et al., 2009].

The resulted reads were aligned to the reference human genome (hg19) with the BWA (Burrows-Wheeler Aligner) software. The variant detection was accomplished using the GATK software package.

The obtained variants were then annotated and filtered at several levels using the Annovar program and the Galaxy data analysis web site. Non-synonymous changes were characterized by different software (SIFT, Polyphen) to bioinformatically predict whether the change is deleterious to the protein structure and/or function. Variants located within 10 base pairs of an exon/intron boundary were also classified as potential splicing mutations. Each variant was compared to several annotated SNP databases (dbSNP135, HapMap, 1000Genomes Project, Exome Aggregation Consortium-ExAC). Those variants that were previously documented to have a population frequency greater than 1% were eliminated given that they are unlikely to be the causal variant. We also checked the frequency in an in-house cohort of 3538 Italian control exomes to exclude population-specific variants. Of the remaining SNPs, 45% are predicted to be non-synonymous variants. In the families in which exome was performed on two affected individuals, we selected only the shared variants and we also tested the additional available family members for proper segregation; while for the consanguineous families, we considered only the homozygous variants.

At the end, we prioritized the filtered variants starting from known genes and then evaluated possible candidate genes according to the following criteria: i) interaction with PD causative genes, ii) involvement in pathways known to be connected with PD pathogenesis, and finally iii) expression in relevant tissues/cells. Finally, Sanger sequencing was used to confirm all selected variants and to study their segregation in the relevant families.

4.1 Identification of mutations in PD-related genes

The identification of PD patients carrying mutations in known PD genes is a priority to select families suitable for the identification of new genes involved in PD pathogenesis. After variant prioritization, we identified five families with a dominant inheritance pattern and one proband from a consanguineous family (Figure 4.1) bearing mutations in genes already associated with parkinsonism. These cases were considered genetically resolved and were then excluded from further analysis.

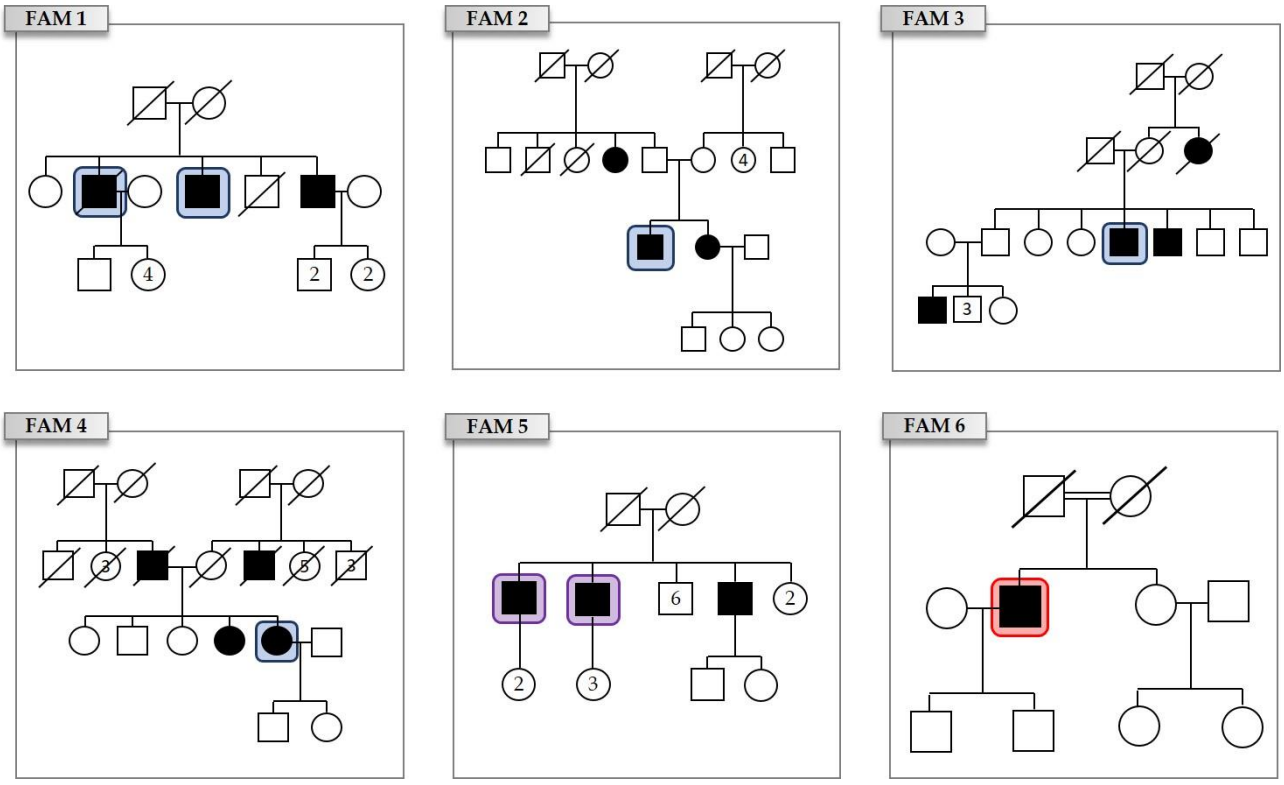


Figure 4.1|Pedigrees of families with mutations in known PD genes (FAM 1-6).
Filled symbols represent affected patients. Highlighted symbols indicate individuals analyzed by WES (in blue the GBA mutated, in violet the LRRK2 mutated, and in red the ATP7B mutated).

In particular, four families (FAM 1-4) were found to be carriers of heterozygous mutations in the GBA gene (Table 4.1), which, as discussed earlier, encode a lysosomal enzyme and is the main genetic risk factor for PD (see Introduction, paragraph 1.2). The variants identified in families FAM 1-3 are missense mutations predicted to be pathogenic by different software and absent in our in-house cohort of Italian controls (n=3538). Only one of them (p.R301C) is reported in dbSNP (with a frequency in ExAC lower than the 0.01%). The variant found in FAM 4 is a novel nonsense mutation (p.Q401*). All the three missense mutations affect amino acids located in the catalytic domain of the protein, which were previously found to be mutated in Gaucher disease patients [LOVD (Leiden Open Variation Database) v.2.0 Build 36].

Concerning FAM 5, a family with a dominant inheritance pattern, we sequenced the exome of two affected brothers (Figure 4.1) and we identified a novel heterozygous missense mutation in the LRRK2 gene: p.E193K (Table 4.1). This gene, as mentioned in paragraph 1.1, codes for a large multidomain protein with a kinase activity that seems to be involved in a number of cellular processes, including autophagy and vesicular trafficking, and whose association with PD is well-established [Martin et al., 2014]. The pathogenic role of this substitution, changing a negatively-charged residue into a positively-charged one, is suggested by its position within an exposed alpha-helix belonging to an Armadillo repeat of the N-terminal domain of the protein. The change of charge on the protein surface could

modify LRRK2 interactions with other proteins as well as promote its aggregation into inactive oligomers. Expression experiments to characterize the effect of the p.E193K mutation are in progress in collaboration with Dr. Piccoli (Center for Integrative Biology, University of Trento).

Finally, in the proband of consanguineous family FAM 6 (Figure 4.1), we found a homozygous missense mutation (p.I116T) in the ATP7B gene (Table 4.1). This gene encode the copper-transporting ATPase 2, a member of the P-type ATPase family, a group of proteins that transport metals through the plasma membrane. In particular, this protein is predominantly expressed in the liver and at a lower level in the kidney and in the brain, and it plays a central role in copper transport, especially in the efflux of hepatic copper into the bile. Mutations in this gene are associated to Wilson's disease, a recessive genetic disorder caused by copper accumulation in the liver and in the brain. Symptoms vary among and within families and include liver disease in 40% of the patients (recurrent jaundice, acute hepatitis or chronic liver disease), neurologic presentation in 40% (parkinsonism or rigid dystonia), and psychiatric disturbance in 20% (depression, anxiety disorders, and psychosis) [Patil et al., 2013]. The diagnosis is usually established starting for the evidence of one of these symptoms followed by the measurement of ceruloplasmin in the serum and copper in the serum and in urine.
Our patient at the time of the diagnosis (33 years) presented parkinsonism but no liver manifestations; the ceruloplasmin and copper evaluation was inconclusive, so he was diagnosed as having parkinsonism.
The identification of a novel mutation in the ATP7B gene in this family allowed us to re-define the diagnosis of this patient as an atypical Wilson disease and accordingly modify the therapeutic approach.

Table 4.1

				allelic frequency			**pathogenicity prediction**	
family	**gene**	**cDNA**	**AA change**	**Italian controls**	**ExAC**	**SNP_ID**	**Sift**	**Polyphen**
FAM 1	GBA	701G>A	G234E	/	/	/	D	D
FAM 2	GBA	901C>T	R301C	/	0.00004	rs374117599	D	P
FAM 3	GBA	1174C>T	R392W	/	/	/	D	D
FAM 4	GBA	1201C>T	Q401*	/	/	/	NA	NA
FAM 5	LRRK2	577G>A	E193K	/	/	/	D	D
FAM 6	ATP7B	347T>C	I116T	0.00027	0.00015	rs199773340	D	P

D: damaging, P: probable damaging, NA: not available.

4.2 FAM 7: a novel lysosomal gene involved in PD

We sequenced the exome of the only affected member of a consanguineous family (FAM 7, Figure 4.2 A). The patient is a 65 year-old man presenting an early onset PD (39 years) characterized by resting tremor and bradykinesia with L-DOPA response and no neuropsychiatric symptoms.

After NGS data analysis, the most promising candidate variant was a homozygous missense mutation in the HGSNAT gene (Heparan-Alpha-Glucosaminide N-Acetyltransferase), also known as TMEM76, located on chromosome 8. This variant is a transition from C to T at position 1,237 on the cDNA (NM_152419) that causes the amino acid substitution p.P413S. This missense variant is present in dbSNP138 (rs201346206) with a frequency lower than the 0.1% (frequency reported also in other mutation databases, ExAC and 1000genomes) and is predicted to be deleterious by different software (SIFT, Polyphen).

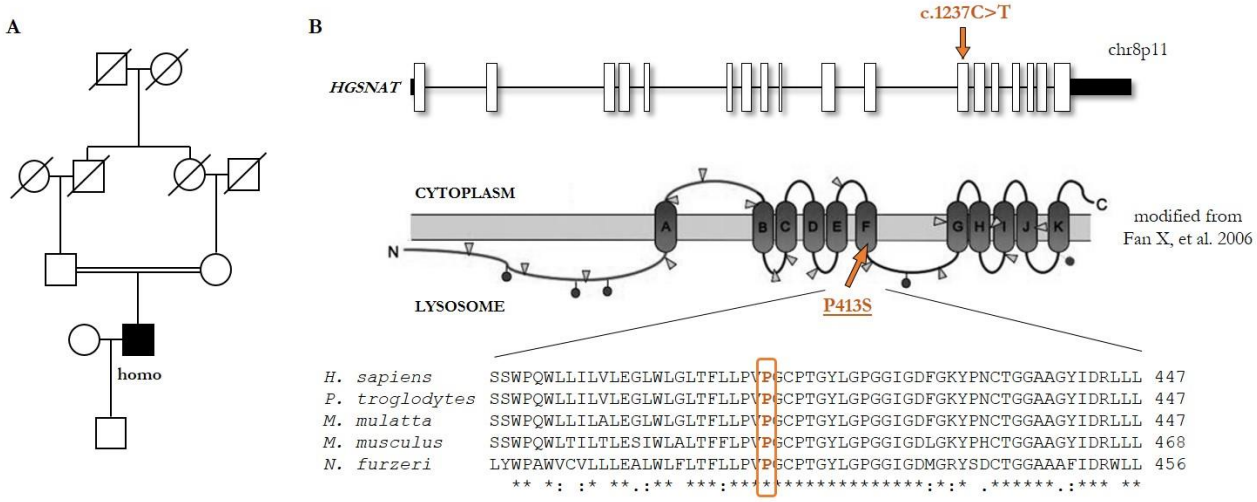


Figure 4.2|Identification of a homozygous mutation in the HGSNAT gene in family 7.
A. Pedigree of the family FAM 7; in black the affected subject. homo=homozygote **B.** Upper panel, the schematic representation of the exon-intron structure of the HGSNAT gene with the genomic localization; the arrow indicates the position of the mutation. Middle panel, schematic representation of the HGSNAT protein; the orange arrow specifies the position of the mutation in the protein; the circles represent putative Asn-linked oligosaccharide sites; the triangles indicate the exon-exon boundaries on the protein. Lower panel, multiple alignment of HGSNAT amino acid sequences from different species; proline 413 (in orange) is fully conserved in vertebrates.

The HGSNAT gene spans 62.4kb, is located at position 8p11.21 and it is composed by 18 exons (Figure 4.2 B). It encodes a lysosomal acetyltransferase: the heparan-alpha-glucosaminide N-acetyltransferase or N-acetyltransferase; this protein is involved in the heparan sulfate (HS) degradation. In particular, this enzyme catalyses the acetylation of the terminal alpha-glucosamine residues of HS converting it into a substrate for the subsequent steps of the degradation process [Fan et al., 2006].

The N-acetyltransferase is an integral membrane protein composed by 11 transmembrane domains and a large N-terminal lysosomal domain. It is also characterized by a 28 amino acid signal peptide that is not cleaved [Durand et al., 2010] and by five predicted sites for Asn-linked glycosylation [Fan et al., 2006] (Figure 4.2 B). Moreover, the protein is transcribed as a precursor and is then cleaved in the

lysosome into a α- and a β- chain, both required for protein activity; this cleavage is necessary for protein activation, probably making accessible the active site of the enzyme. It is also known that this enzyme works as a homo-oligomer presumably composed by six α- and six β-chains [Durand et al., 2010].

Mutations in the HGSNAT gene were classically associated with the mucopolysaccharidosis (MPS) type IIIC, also called Sanfilippo syndrome C (OMIM #252930), a lysosomal storage disorder [Fan et al., 2006]. It is a severe autosomal recessive disease with an infantile/early childhood onset (3-7 years). There are other three subtypes (A, B, D) of Sanfilippo syndrome, caused by mutations in genes coding for enzymes involved in the degradation of HS (SGSH, NAGLU, GNS) [Andrade et al., 2015].
The prevalence of MPS type III is estimated between 0.3 and 4.1 cases per 100,000 newborns with differences among geographical areas; the Sanfilippo syndrome C is the rarest with an incidence of 1:1,500,000. The four subtypes share the majority of the clinical features and the patients are usually classified only by the molecular defect. The main symptoms are progressive cognitive decline, behavioral problems, sleeping and speech disorder and, in severe forms, also hearing loss, and visceral manifestations, such as mild hepatomegaly, mild musculoskeletal abnormalities and deformities, mild coarse facies, and hypertrichosis [Andrade et al., 2015].
To date, 66 mutations in the HGSNAT gene were described, including: 37 missense mutations, 14 splicing variants, 11 small insertions/deletions, and 5 complex alleles [HGMD (Human Gene Mutation Database) release 2015.1] These mutations affect different protein domains but all lead to an absent, or extremely reduced, enzymatic activity.

Concerning our patient, the mutated proline is located in the 6th highly hydrophobic transmembrane domain and it corresponds to an amino acid residue very conserved during evolution, located in a highly conserved portion of the protein (Figure 4.2 B). The substitution of P413 with a serine, a smaller polar amino acid, could destabilize the transmembrane helix, leading to an incorrect protein folding. This assumption is also corroborated by the existence of another reported mutation in the same α-helix, W403C, described to cause protein misfolding, ER retention of the unfolded protein, and eventually causing a significant reduction of the protein activity [Feldhammer et al., 2009].
Therefore, we measured, in collaboration with the laboratory of Dr. Filocamo (Dipartimento di Neuroscienze, IRCCS Istituto G. Gaslini, Genova), the HGSNAT enzymatic activity in the patient's fibroblasts. The activity levels resulted 2.7 nmol/mg/17h, which is significantly lower than the normal range (22.1±10.2 nmol/mg/17h), confirming the deleterious effect of the mutation P413S on protein activity (Figure 4.3).
Moreover, we recently identified another PD patient carrying two heterozygous missense mutations in the HGSNAT gene (G565R and A615T). This patient presents a late-onset (53 years) PD with rigidity

and clumsiness of the right arm at the diagnosis. One year later DaTscan confirmed the diagnosis and the patient started the L-DOPA treatment with benefits. He does not present any neuropsychiatric symptoms until today. These two mutations are already annotated in dbSNP (rs148632988, rs112029032) with a frequency lower than 1% in the main mutation databases. We measured, also for this patient, the residual enzymatic activity in fibroblast, which resulted 6.7 nmol/mg/17h (Figure 4.3).

patient	genotype	Enzymatic activity* (nmol/mg/17h)
A	homo P413S	2.7
B	het G565R het A615T	6.7

*heparan-α-glucosaminide N-acetyltransferase
normal range= 22.1±10.2 nmol/mg/17h

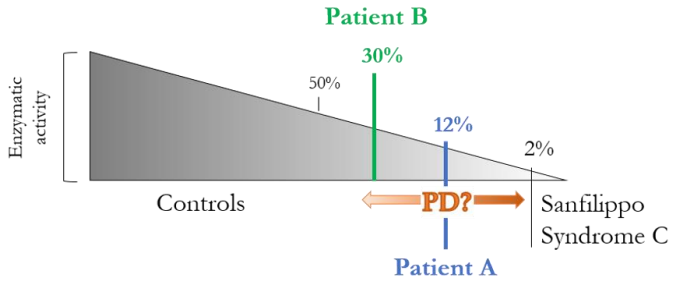


Figure 4.3|Enzymatic activity of HGSNAT mutants measured in vivo.
Left panel, the N- acetyltransferase enzymatic activity levels measured in patient fibroblasts using a fluorescence assay at the laboratory of Dr. Filocamo. Right panel, schematic representation of the residual activity in our patients and Sanfilippo syndrome C patients, setting as 100% the average activity of the controls.

The finding of two PD patients who do not present any symptoms characterizing the Sanfilippo C syndrome but are carrier of deleterious mutations in the HGSNAT gene, let us hypothesize an involvement of N-acetyltransferase deficiency in PD. In particular, the PD patients' residual enzymatic activity (12% and 30%, setting 100% the control average activity) is higher than the activity measured in Sanfilippo C patients (2%) but significantly lower compared to controls (Figure 4.3) and to individuals heterozygous for mutations causing MPS type III (data not shown). Our hypothesis is that an enzymatic activity not enough reduced to cause a lysosomal disorder may predispose to PD, as already clearly shown for GBA.

Supporting our hypothesis there are also literature evidences showing the presence of phosphorylated -α-synuclein across different brain regions in 3 cases of MPS IIIB [Hamano et al., 2008] and the accumulation of α-synuclein in cortical neurons of 2 cases of MPS IIIA [Winder-Rhodes et al., 2012]. Concerning genetic data, a study of 2,308 controls and 926 PD patients showed an association between Parkinson's disease and a common NAGLU haplotype (tag with the SNP rs2071046) with an OR of 1.3 [Winder-Rhodes et al., 2012].

Finally, these evidences and our results corroborate the link between MPS III and Parkinson's disease and more in general further stress the important role of lysosomal dysfunction in neurodegeneration and α-synuclein accumulation.

4.3 FAM 8: a novel DNAJ gene involved in PD

We performed exome sequencing on one individual (IV7) from the consanguineous family FAM 8 affected by early-onset PD (onset at 26 years old) (Figure 4.4). In particular, the patient shows bilateral tremor, slow progression and L-DOPA response with cognitive decline and neuropsychiatric symptoms.

The WES data analysis highlighted the presence of a homozygous splicing variant (c.79-2A>G, NM_021800) in the DNAJC12 gene. This variant was absent in the main mutation databases and was also not present in our cohort of 3538 Italian controls. The segregation in the family was confirmed by Sanger sequencing: the affected brother (IV5) resulted homozygous for the mutation while the two healthy siblings were one heterozygous (brother, IV2) and the other wild type (IV3). The other family members tested resulted carrier of this mutation (III2, III5, III13 and IV8) (Figure 4.4).

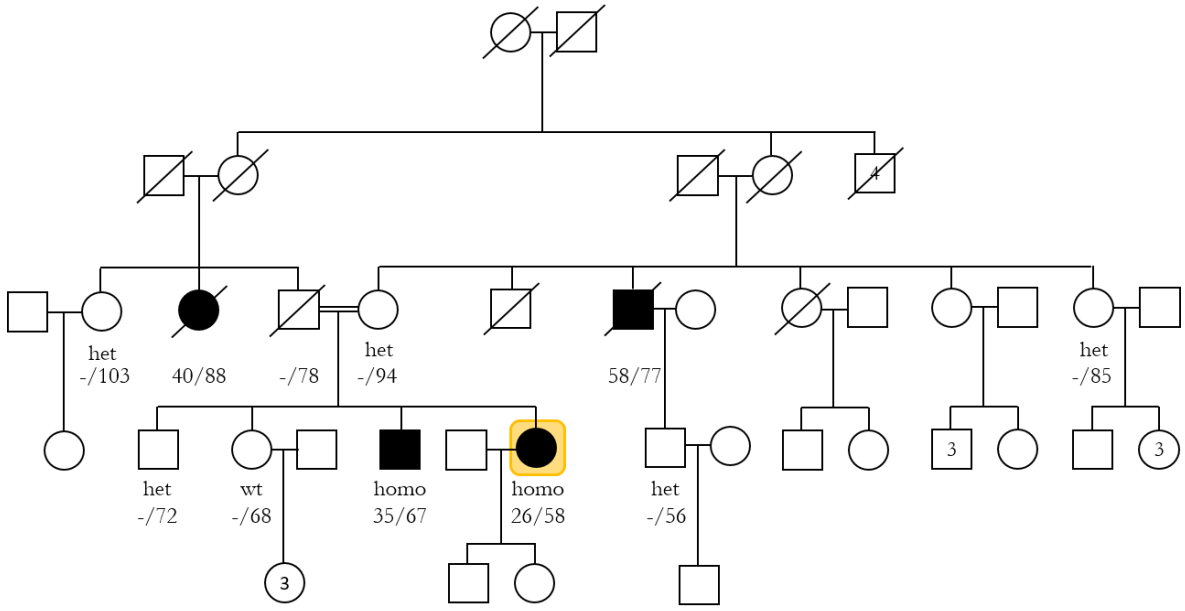


Figure 4.4|Pedigree of family FAM 8.
In black the affected subjects and highlighted by a yellow square the proband who was analysed by WES. The genotype for the c.79-2 A>G mutation, confirmed by Sanger sequencing, the age at the diagnosis/present age or age at death are reported, if available, under each family member. het=heterozygote, homo=homozygote, wt=wild type.

4.3.1 DNAJC12 gene and protein

The DNAJC12 gene spans 27kb, it is located in position 10q21.3 and comprises 5 exons. This gene encodes for a J domain-containing protein that belongs to the Hsp40/DNAJ protein family, a heterogeneous group of proteins that play the essential role of co-chaperone of the Hsp70 proteins [Qiu et al., 2006]. The Hsp40 proteins are characterized by the presence of a highly conserved J-domain, responsible for the interaction with the ATPase domain of Hsp70s and the stimulation of the ATP hydrolysis. The regions outside this domain are usually involved in the interaction with other cellular protein. These proteins are homodimeric proteins residing in various subcellular compartments as well as in the extracellular milieu [Kampinga and Craig, 2010]. Moreover, some of the DNAJ

proteins are present in specific tissues while others are ubiquitously expressed; the variety of expression and localization reflects the functional diversity of the Hsp70s [Kampinga and Craig, 2010].

The DNAJC12 protein is one of the lesser studied among members of the DNAJ family. It is known to contain a single J-domain (aa 14-79), mainly encoded by exon 1 and 2, while its specific function is still not clear. It was previoulsy reported that DNAJC12 expression is upregulated by the transcription factor AIbZIP in a prostate cancer cell line upon ER stress induction [Choi et al., 2014]. DNAJC12 seems to be localized at the cytoplasm level and to interact mainly with: Hsc70 (a chaperone ubiquitously and constitutively expressed), some mitochondrial proteins (such as the pyruvate carboxylase), and, under stress conditions, with Bip (which plays a central role in the ER-stress response) [Choi et al., 2014]. Moreover, two different DNAJC12 transcripts have been reported: a longer one (isoform "a") coding for a 198 amino acids protein and a shorter transcript (isoform "b"), generated by alternative polyadenylation, that encodes for a 109 aa polypeptide (Figure 4.5). Both isoforms contain the same J-domain but the functional role is still unknown.

We evaluated the mRNA expression profile of both isoforms in a commercial panel of 20 human tissues and 24 brain areas by real-time RT-PCR, using isoform-specific primers. The levels of isoform a are always higher the ones of isoform b. The two isoforms are predominantly expressed in the liver and testis among the analyzed tissues and in the cerebellar hemisphere among the brain areas. Isoform a also shows high levels in the brain while it is almost absent in the placenta and in the frontal cortex (Figure 4.5).

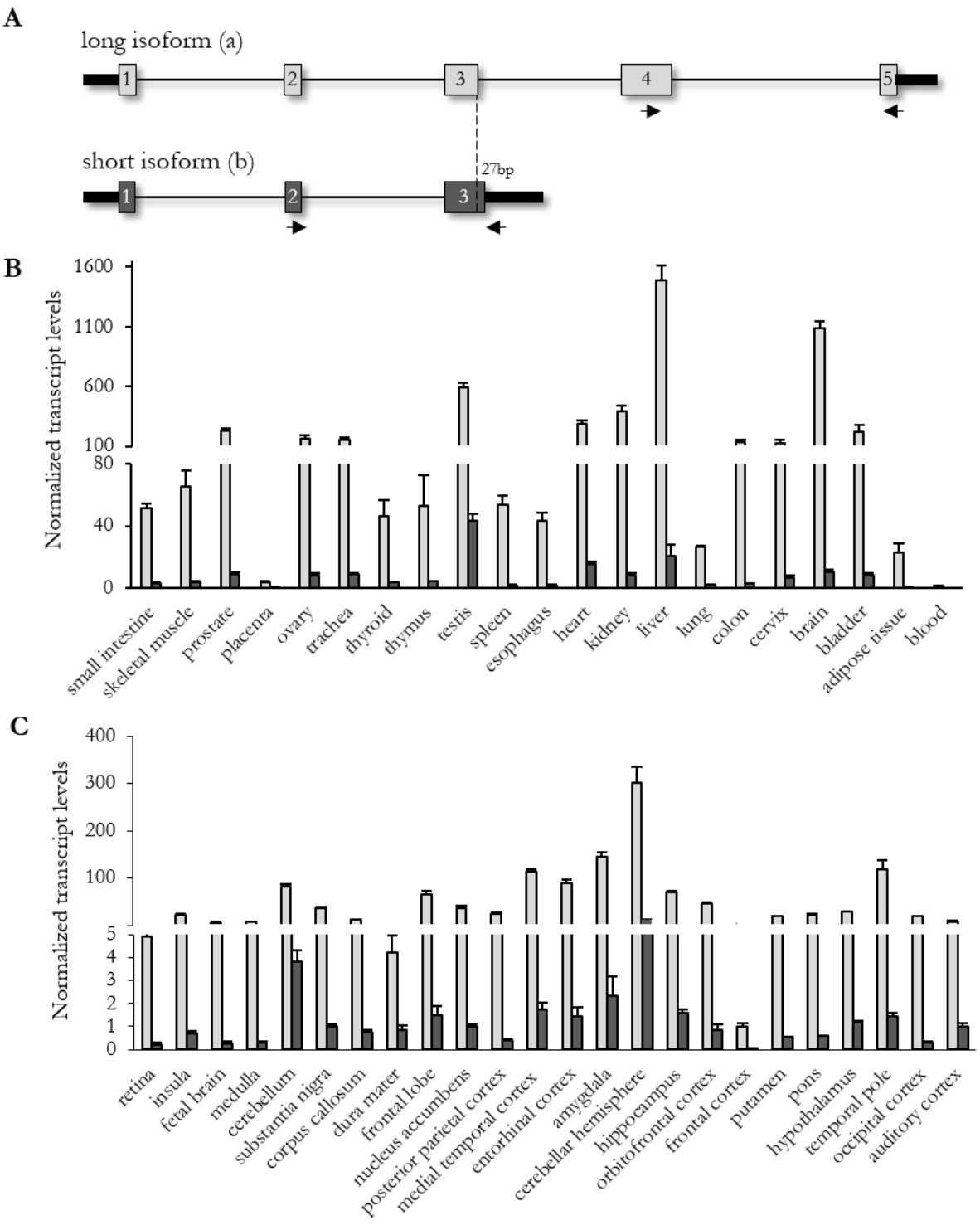


Figure 4.5|DNAJC12 isoforms expression pattern.

A. Schematic representation of the two DNAJC12 isoforms with the position of the primers (represented by arrows) used for the transcript levels measurement **B, C.** Isoform a and b expression levels were measured by real time RT-PCR assays using the SYBR Green chemistry on RNA derived from a commercial panel including 20 human tissues (B) and 24 brain areas (C). HMBS transcripts were used as internal reference.

4.3.2 Characterization of the splicing mutation c.79-2A>G

We evaluated the possible effect of this mutation c.72-2A>G on splicing in silico, using two different software: NNSPLICE 0.9 version and NetGene2. Both algorithms predicted the mutation to completely abolish exon 2 acceptor splice site. We confirmed these predictions evaluating by RT-PCR the possible aberrant splicing directly on the patient RNA extracted from total blood. In particular, agarose gel electrophoresis of the obtained PCR products showed the amplification of a lower molecular weight fragment (199bp) compared to the control (278bp). Sanger sequencing confirmed that the alternative product is due to complete skipping of exon 2. This aberrant splicing event leads to a frame-shift and consequently to the introduction of a premature stop after 13 alternative amino acids at the codon 39 of the mutated protein (Figure 4.6 A). Considering that out-of-frame transcripts are frequently degraded by the nonsense-mediated mRNA decay mechanism, we measured by real time RT-PCR the total levels of DNAJC12 transcripts in the patient and in control RNAs extracted from total blood. The results of this experiment showed that the transcript levels are significant reduced in the patient (Figure 4.6 B), even if the blood is not the perfect tissue to measure DNAJC12 transcripts considering its very low expression in this tissue.

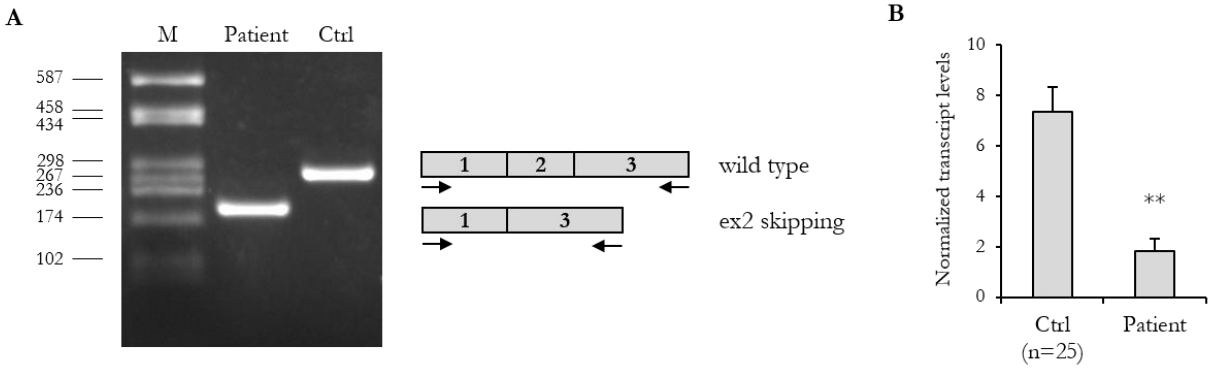


Figure 4.6| Characterization of the c.79-2 A>G mutation effect on splicing.
A. On the left, the agarose gel showing the results of the RT-PCR assay performed on the RNA extracted from total blood of the patient and of a control individual (Ctrl). On the right, the schematic representation of the amplicons obtained from the RT-PCR and sequenced by Sanger sequencing with the position of the primers used for the RT-PCR. **B.** Quantification of the total transcript levels of DNAJC12 in 25 controls and in the patient by real time RT-PCR using the HMBS transcripts as internal reference. The results were analyzed by unpaired t-test. **: $p<0.01$.

4.3.3 Effects of DNAJC12 silencing on α-synuclein

The possible link between DNAJC12 deficiency and PD was assessed performing silencing experiments in SH-SY5Y cells using an antisense oligonucleotide-based approach. In particular, we treated the cells with a custom GapmeR, an LNA (Locked Nucleic Acid) oligonucleotide, designed using the Exiqon proprietary design software and targeting both DNAJC12 isoforms. We transfected the SH-SY5Y cells with two doses (every 24h after cell seeding) of the GapmeR against DNAJC12 or the negative control GapmeR and after 24h we evaluated the silencing of DNAJC12 and the accumulation of α-synuclein by western blot. We obtained an average DNAJC12 protein

downregulation of the 40% compared to the negative control, using the GAPDH protein levels as an internal reference. The results of these experiments interestingly showed that the silencing of DNAJC12 causes a 2.2-fold increase in the level of monomeric α-synuclein (Figure 4.7).

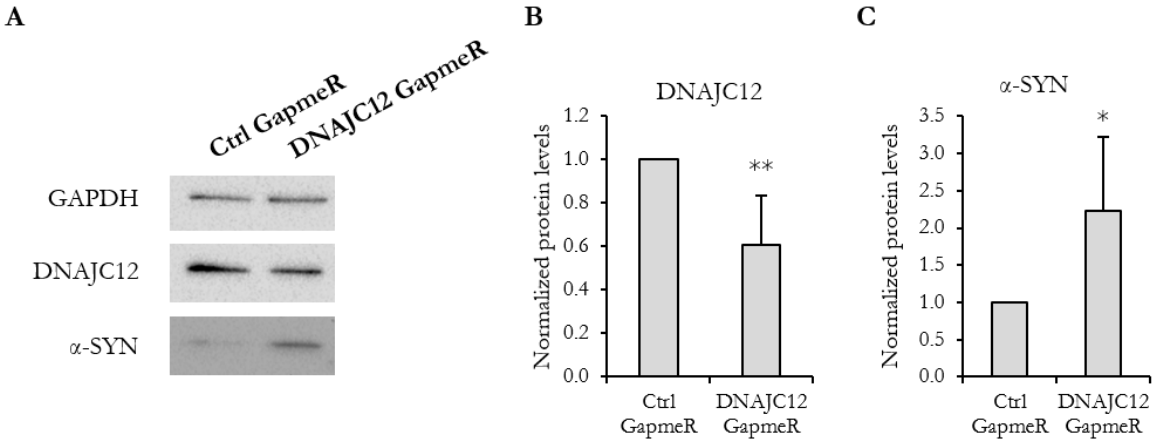


Figure 4.7| DNAJC12 silencing increases α-synuclein protein levels.

SH-SY5Y were transfected with two doses (every 24h) of 30nM of a GapmeR designed against DNAJC12 exon 2 or a Ctrl GapmeR. After 24h from the second dose, cells were harvested and sonicated for 10 seconds twice in ice. Next, 40μg of the total proteins extracted were loaded onto a 13% polyacrylamide gel; for Western Blot, primary antibodies used were: monoclonal antbody to α-synuclein (610787, BD Biosciences; diluted 1:1000), monoclonal antbody anti-DNAJC12 (ab167425, Abcam; diluted 1:5000), and polyclonal antibody to GAPDH (NB300-320, Novus Biologicals, diluted 1:5000); secondary peroxidase-conjugated antibodies were used in addition to SuperSignal West Dura Extended Duration Substrate revelation kit (Thermo Fisher Scientific). After scanning, the intensity of each band was estimated by densitometric quantification using ImageJ software. **A.** Western Blot representative of six independent experiments. **B.** Average DNAJC12 protein levels (B) and average α-synuclein protein levels (C) measured in SH-SY5Y cells treated with the DNAJC12 GapmeR compared to cells treated with the control GapmeR. GAPDH was used as internal reference. The results were analyzed by unpaired t-test. *: $p<0.05$, **: $p<0.01$.

4.3.4 Conclusions and future perspectives

Interestingly, the α-synuclein increase obtained by the silencing of DNAJC12 supports the possible role of this gene in the pathogenesis of Parkinson' disease. Moreover, this protein seems to be involved in the ER-stress response and to interact with proteins such as Hsc70 and Bip that play a central role in the protein homeostasis especially in the protein folding [Choi et al., 2014]. The ER stress response (or Unfolded Protein Response, UPR) is an evolutionary conserved cellular mechanism triggered by accumulation of misfolded proteins. Many genes involved in familial PD affect at different point the secretory pathway and usually their dysfunction leads to protein accumulation, in particular α-synuclein aggregation, and consequently to the activation of the UPR [Mercado et al., 2016]. Moreover, recent studies showed that α-synuclein aggregation by nutrient deprivation begins following the ER stress response [Jiang et al., 2014] and that silencing x-box binding protein 1 (XBP-1), a UPR transcription factor, triggers chronic ER-stress and dopaminergic neuron degeneration [Valdés et al., 2014]. Furthermore, the expression of an ER stress marker, glucose regulated protein (GRP78 or Bip), was increased in 1-methyl-4-phenylpyridinium (MPP^+)-treated neuronal cells. In addition, the expression of another ER-stress marker, C/EBP homologous protein (CHOP), was increased in a 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP)-induced Parkinson's disease mouse model [Tsujii S et al.,

2015]. Moreover, we measured both ER-stress markers (Bip and CHOP) in total blood of 26 PD patients and 25 controls using real time RT-PCR assays. We observed a significant increase of both markers: concerning Bip, the normalized mean levels in controls are 5.7 and reach 15.17 in PD patients (p=0.002), while the average levels of CHOP are 6.05 in controls and achieve 10.00 in cases (p=0.028) (Figure 4.8). These data corroborate the link between ER-stress and PD supporting the possible role of DNAJC12 in PD pathogenesis.

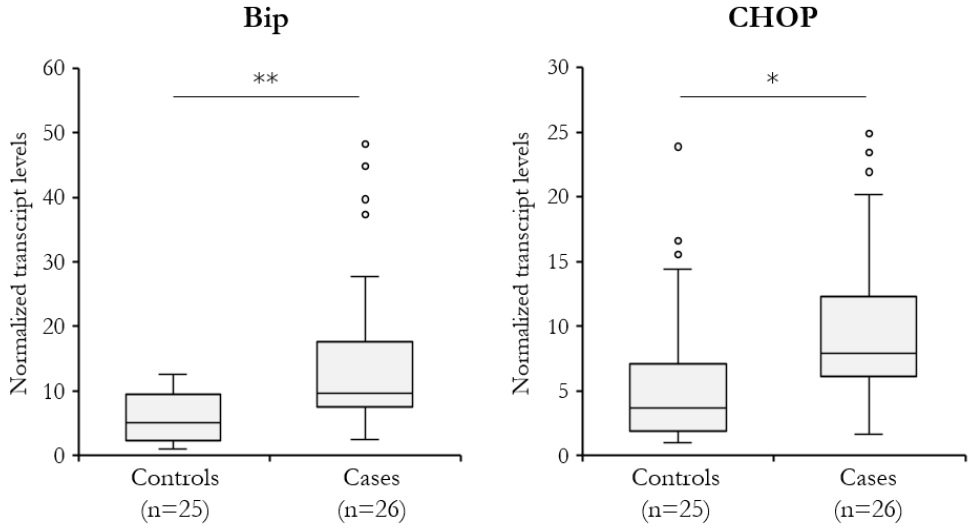


Figure 4.8|Expression levels of Bip and CHOP transcripts.
Quantification of the ER-stress markers, Bip (left panel) and CHOP (right panel), by real time RT-PCR on RNA extracted from total blood of PD patients and controls using SYBR Green chemistry. HMBS transcripts were used as internal reference. The results were analyzed by unpaired t-test. *: $p<0.05$, **: $p<0.01$.

Furthermore, DNAJC12 is not the first Hsp40 protein associated with PD; mutations in two other gene coding for DNAJ proteins were described in PD patients. In particular, mutations in the DNAJC6 gene were associated with autosomal recessive juvenile or early-onset PD [Edvardson et al., 2012] while mutations in the DNAJC13 gene were found in patients with late-onset PD [Vilariño-Güell et al., 2014]. Moreover, other DNAJ protein (DNAJB2, DNAJB6, DNAJC5, DNAJC19, and DNAJC29) were reported to be involved in other neurodegenerative disorder, like ataxia or motor neuropathy [Koutras and Braun, 2014]. All these proteins have very different functions in the cells but they are involved in different proteostasis events.

As further evidence of the causative role of the DNAJC12 gene, we also found, in collaboration with the laboratory of Prof. Farrer (Centre of Applied Neurogenetics, Vancouver, Canada), another early-onset PD patient presenting a homozygous nonsense mutation (K63*) in this same gene. This proband belongs to a Canadian family with no consanguinity reported; he presented a dopa-responsive, nonprogressive, juvenile (onset at 13 years) parkinsonism [Rajput et al., 1997].
Interestingly, the age of onset of PD fitted the predicted severity of the mutations, being 13 years for the nonsense mutation and 26 - 35 years for the splicing mutation, which can be expected to be

compatible with trace amounts, although undetectable by RT-PCR, of wild-type splicing.
We also performed the screening on 100 early-onset Italian PD patients by a PCR-based approach and NGS. In particular, we designed 11 amplicons spanning all the coding regions of the gene. We performed multiplexed PCR reactions on pools of 10 DNAs each and the obtained amplicons were ligated with adapters and unique indexes. The resulted DNA libraries were sequenced on the NextSeq500 Illumina sequencer. We didn't find any additional mutated subject but we are now extending this screening to other 100 early-onset and 400 familial PD patients. However, we have to consider that mutations in this gene could be very rare and so difficult to identify.

In conclusion, taking into account the literature data, our finding of two early-onset PD families presenting null homozygous mutations in the DNAJC12 gene and our results on α-synuclein accumulation upon DNAJC12 silencing, we propose DNAJC12 as a novel gene involved in the pathogenesis of Parkinson's disease. Additional functional analyses will need to be performed to establish the function of this protein in the cell and consequently the mechanism linking its deficiency to neurodegeneration. In particular, we plan to perform immunofluorescence experiments to assess the subcellular localization of DNAJC12 protein at basal state or under ER-stress condition (starvation, tunicamycin or DTT treatments) and to confirm its molecular interactors by immunoprecipitation experiments.
In addition, an interesting perspective may be the functional characterization of the role of this gene in PD pathogenesis using the patient cells. To this end, we plan to establish dopaminergic neurons derived from iPS cells generated starting from immortalized patient fibroblasts obtained from skin biopsy. This could be an interesting possibility to evaluate the pathways involved in the neurodegeneration process caused by mutations in the DNAJC12 gene in a disease relevant cellular model. iPSC-derived neurons have been successfully applied for the functional studies of mutations in PD-causing genes, such as SNCA, PINK1, Parkin and GBA [Oliveira et al., 2015, Seibler et al., 2011, Rakovic et al., 2015, Schöndorf et al., 2014]. iPS cells and differentiated dopaminergic neurons will be established through the already existing collaboration with Prof. Rejko Kruger laboratory (University of Luxembourg).
Finally, we are trying, in collaboration with Dr. Del Giacco (Department of Biosciences, University of Milan) to established an animal model to study DNAJC12 and its link to PD. In particular, we are planning to use a zebrafish model expressing human α-synuclein in dopaminergic neurons available in the laboratory of Dr. Del Giacco and knockdown the DNAJC12 gene by antisense oligonucleotide with a synthetic backbone called morpholino. The knockdown morphants will be further characterized for the presence of a phenotype or altered behavior. Moreover, cellular features using microscopy and altered gene expression by in situ hybridization or biochemical analyses will be performed to evaluate cellular phenotypes.

5 | References

Andrade F, Aldámiz-Echevarría L, Llarena M, Couce ML. Sanfilippo syndrome: Overall review. Pediatr Int. 2015 Jun;57(3):331-8.

Antony PM, Diederich NJ, Krüger R, Balling R. The hallmarks of Parkinson's disease. FEBS J. 2013 Dec;280(23):5981-93.

Asselta R, Rimoldi V, Siri C, Cilia R, Guella I, et al. Glucocerebrosidase mutations in primary parkinsonism. Parkinsonism Relat Disord. 2014 Nov;20(11):1215-20.

Bamshad MJ, Shendure JA, Valle D, Hamosh A, Lupski JR, et al. The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am J Med Genet A. 2012 Jul;158A(7):1523-5.

Berezikov E, van Tetering G, Verheul M, van de Belt J, van Laake L et al. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res. 2006 Oct;16(10):1289-98.

Boot RG, Verhoek M, Donker-Koopman W, Strijland A, van Marle J, et al. Identification of the non-lysosomal glucosylceramidase as beta-glucosidase 2. J Biol Chem. 2007 Jan 12;282(2):1305-12.

Cazalla D, Steitz JA. Down-regulation of a host microRNA by a viral noncoding RNA. Cold Spring Harb Symp Quant Biol. 2010;75:321-4.

Cho HJ, Liu G, Jin SM, Parisiadou L, Xie C et al. MicroRNA-205 regulates the expression of Parkinson's disease-related leucine-rich repeat kinase 2 protein. Hum Mol Genet. 2013 Feb 1;22(3):608-20.

Choi J, Djebbar S, Fournier A, Labrie C. The co-chaperone DNAJC12 binds to Hsc70 and is upregulated by endoplasmic reticulum stress. Cell Stress Chaperones. 2014 May;19(3):439-46.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009 Nov 10;106(45):19096-101.

Cogswell JP, Ward J, Taylor IA, Waters M, Shi Y, et al. Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways. J Alzheimers Dis. 2008 May;14(1):27-41.

Cooper AA, Gitler AD, Cashikar A, Haynes CM, Hill KJ, et al. Alpha-synuclein blocks ER-Golgi traffic and Rab1 rescues neuron loss in Parkinson's models. Science. 2006 Jul 21;313(5785):324-8.

Credle JJ, Forcelli PA, Delannoy M, Oaks AW, Permaul E, et al. α-Synuclein-mediated inhibition of ATF6 processing into COPII vesicles disrupts UPR signaling in Parkinson's disease. Neurobiol Dis. 2015 Feb 26 doi:10.1016/j.nbd.2015.02.005.

de Graaf M, van Veen IC, van der Meulen-Muileman IH, Gerritsen WR, Pinedo HM, et al. Cloning and characterization of human liver cytosolic beta-glycosidase. Biochem J. 2001 Jun 15;356(Pt 3):907-10.

Denti MA, Rosa A, Sthandier O, De Angelis FG, Bozzoni I. A new vector, based on the PolII promoter of the U1 snRNA gene, for the expression of siRNAs in mammalian cells. Mol Ther. 2004 Jul;10(1):191-9.

Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. 2011 Jun 7(6):e1002141.

Doxakis E. Post-transcriptional regulation of alpha-synuclein expression by mir-7 and mir-153. J Biol Chem. 2010 Apr 23;285(17):12726-34.

Durand S, Feldhammer M, Bonneil E, Thibault P, Pshezhetsky AV. Analysis of the biogenesis of heparan sulfate acetyl-CoA:alpha-glucosaminide N-acetyltransferase provides insights into the mechanism underlying its complete deficiency in mucopolysaccharidosis IIIC. J Biol Chem. 2010 Oct 8;285(41):31233-42.

Dvir H, Harel M, McCarthy AA, Toker L, Silman I, et al. X-ray structure of human acid-beta-glucosidase, the defective enzyme in Gaucher disease. EMBO Rep. 2003 Jul 4(7):704-9.

Edvardson S, Cinnamon Y, Ta-Shma A, Shaag A, Yim YI, et al. A deleterious mutation in DNAJC6 encoding the neuronal-specific clathrin-uncoating co-chaperone auxilin, is associated with juvenile parkinsonism. PLoS One. 2012;7(5):e36458.

Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer. 2006 Apr;6(4):259-69.

Fan X, Zhang H, Zhang S, Bagshaw RD, Tropak MB, et al. Identification of the gene encoding the enzyme deficient in mucopolysaccharidosis IIIC (Sanfilippo disease type C). Am J Hum Genet. 2006 Oct;79(4):738-44. Epub 2006 Aug 23.

Farlow J, Pankratz ND, Wojcieszek J, Foroud T. Parkinson Disease Overview. GeneReviews. 2004 May 25[updated 2014 Feb 27].

Feldhammer M, Durand S, Pshezhetsky AV. Protein misfolding as an underlying molecular defect in mucopolysaccharidosis III type C. PLoS One. 2009 Oct 13;4(10):e7434.

Foo JN, Liu JJ, Tan EK. Whole-genome and whole-exome sequencing in neurological diseases. Nat Rev Neurol. 2012 Sep;8(9):508-17.

Funayama M, Ohe K, Amo T, Furuya N, Yamaguchi J, et al. CHCHD2 mutations in autosomal dominant late-onset Parkinson's disease: a genome-wide linkage and sequencing study. Lancet Neurol. 2015 Mar;14(3):274-82.

Gegg ME, Burke D, Heales SJ, Cooper JM, Hardy J, et al. Glucocerebrosidase deficiency in substantia nigra of parkinson disease brains. Ann Neurol. 2012 Sep;72(3):455-63.

Goker-Alpan O, Hruska KS, Orvisky E, Kishnani PS, Stubblefield BK, et al. Divergent phenotypes in Gaucher disease implicate the role of modifiers. J Med Genet. 2005 Jun;42(6):e37.

Goldman SM. Environmental toxins and Parkinson's disease. Annu Rev Pharmacol Toxicol. 2014;54:141-64.

Hamano K, Hayashi M, Shioda K, Fukatsu R, Mizutani S. Mechanisms of neurodegeneration in mucopolysaccharidoses II and IIIB: analysis of human brain tissue. Acta Neuropathol. 2008 May;115(5):547-59.

He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nature Reviews Genetics. 2004 Jul 5(7):522-31.

Heman-Ackah SM, Hallegger M, Rao MS, Wood MJ. RISC in PD: the impact of microRNAs in Parkinson's disease cellular and molecular pathogenesis. Front Mol Neurosci. 2013 Nov 20;6:40.

Hirsch EC, Jenner P, Przedborski S. Pathogenesis of Parkinson's disease. Mov Disord. 2013 Jan 28(1):24-30.

Horowitz M, Wilder S, Horowitz Z, Reiner O, Gelbart T, et al. The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics. 1989 Jan 4(1):87-96.

Hruska KS, LaMarca ME, Scott CR, Sidransky E. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Hum Mutat. 2008 May 29(5):567-83.

Hunn BH, Cragg SJ, Bolam JP, Spillantini MG, Wade-Martins R. Impaired intracellular trafficking defines early Parkinson's disease. Trends Neurosci. 2015 Jan 29 doi:10.1016/j.tins.2014.12.009.

International Parkinson's Disease Genomics Consortium, Nalls MA, Plagnol V, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genomewide association studies. Lancet. 2011 Feb 19,377(9766):641-9.

International Parkinson's Disease Genomics Consortium, Wellcome Trust Case Control Consortium. A two-stage meta-analysis identifies several new loci for Parkinson's disease. PLoS Genet. 2011 Jun 7(6):e1002142.

Jazbutyte V, Fiedler J, Kneitz S, Galuppo P, Just A, et al. MicroRNA-22 increases senescence and activates cardiac fibroblasts in the aging heart. Age 2013; 35: 747-62.

Jazdzewski K, Liyanarachchi S, Swierniak M, Pachucki J, Ringel MD et al. Polymorphic mature microRNAs from passenger strand of pre-miR-146a contribute to thyroid cancer. Proc Natl Acad Sci U S A. 2009 Feb 3;106(5):1502-5.

Jiang P, Gan M, Lin WL, Yen SH. Nutrient deprivation induces α-synuclein aggregation through endoplasmic reticulum stress response and SREBP2 pathway. Front Aging Neurosci. 2014 Oct 8;6:268.

Jovicic A, Zaldivar Jolissaint JF, Moser R, Silva Santos Mde F, Luthi-Carter R. MicroRNA-22 (miR-22) overexpression is neuroprotective via general anti-apoptotic effects and may also target specific Huntington's disease-related mechanisms. PLoS One. 2013;8(1):e54222.

Junn E, Lee KW, Jeong BS, Chan TW, Im JY, Mouradian MM. Repression of alpha-synuclein expression and toxicity by microRNA-7. Proc Natl Acad Sci U S A. 2009 Aug 4;106(31):13052-7.

Kabaria S, Choi DC, Chaudhuri AD, Mouradian MM, Junn E. Inhibition of miR-34b and miR-34c enhances α-synuclein expression in Parkinson's disease. FEBS Lett. 2015 Jan 30;589(3):319-25.

Kampinga HH1, Craig EA. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. Nat Rev Mol Cell Biol. 2010 Aug;11(8):579-92.

Karreth FA, Pandolfi PP. ceRNA cross-talk in cancer: when ce-bling rivalries go awry. Cancer Discov. 2013 Oct 3(10):1113-21.

Karreth FA, Reschke M, Ruocco A, Ng C, Chapuy B, et al. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. Cell 2015;161:319-32.

Kim J, Inoue K, Ishii J, Vanti WB, Voronov SV, et al. A MicroRNA feedback circuit in midbrain dopamine neurons. Science. 2007 Aug 31;317(5842):1220-4.

Kimura T, Ishiguro K, Hisanaga S. Physiological and pathological phosphorylation of tau by Cdk5. Front Mol Neurosci. 2014 Jul 15;7:65.

Klein C, Westenberger A. Genetics of Parkinson's disease. Cold Spring Harb Perspect Med. 2012 Jan 2(1):a008888.

Koutras C, Braun JE. J protein mutations and resulting proteostasis collapse. Front Cell Neurosci. 2014 Jul 8;8:191

Krebs CE, Karkheiran S, Powell JC, Cao M, Makarov V, et al. The Sac1 domain of SYNJ1 identified mutated in a family with early-onset progressive Parkinsonism with generalized seizures. Hum Mutat. 2013 Sep;34(9):1200-7.

Krützfeldt J, Stoffel M. MicroRNAs: a new class of regulatory genes affecting metabolism. Cell metabolism. 2006 Jul 4(1):9-12.

Lachmann RH, Grant IR, Halsall D, Cox TM. Twin pairs showing discordance of phenotype in adult Gaucher's disease. QJM. 2004;97:199-204.

Lee DY, Jeyapalan Z, Fang L, Yang J, Zhang Y, et al. Expression of versican 3'-untranslated region modulates endogenous microRNA functions. PLoS One. 2010 Oct 25;5(10):e13599.

Lees AJ, Hardy J, Revesz T. Parkinson's disease. Lancet. 2009 Jun 13; 373(9680):2055-66.

Lesage S, Brice A. Parkinson's disease: from monogenic forms to genetic susceptibility factors. Hum Mol Genet. 2009 Apr; 15;18(R1):R48-59.

Lesage S, Brice A. Role of mendelian genes in "sporadic" Parkinson's disease. Parkinsonism Relat Disord. 2012 Jan 18 Suppl 1:S66-70.

Lesage S, Drouet V, Majounie E, Deramecourt V, Jacoupy M, et al. Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy. Am J Hum Genet. 2016 Mar 3;98(3):500-13.

Li N, Bates DJ, An J, Terry DA, Wang E. Up-regulation of key microRNAs, and inverse down-regulation of their predicted oxidative phosphorylation target genes, during aging in mouse brain. Neurobiol Aging 2011; 32: 944-55.

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 2005 Feb 17;433(7027):769-73.

Lin MK, Farrer MJ. Genetics and genomics of Parkinson's disease. Genome Med. 2014 Jun 30;6(6):48.

Liu X, Cheng R, Verbitsky M, Kisselev S, Browne A, et al. Genome-wide association study identifies candidate genes for Parkinson's disease in an Ashkenazi Jewish population. BMC Med Genet. 2011 Aug 3;12:104.

Lwin A, Orvisky E, Goker-Alpan O, LaMarca ME, Sidransky E. Glucocerebrosidase mutations in subjects with parkinsonism. Mol Genet Metab. 2004 Jan;81(1):70-3.

Margis R, Margis R, Rieder CR. Identification of blood microRNAs associated to Parkinson's disease. J Biotechnol 2011; 152: 96–101.

Martin I, Kim JW, Dawson VL, Dawson TM. LRRK2 pathobiology in Parkinson's disease. J Neurochem. 2014 Dec;131(5):554-65.

Mazzulli JR, Xu YH, Sun Y, Knight AL, McLean PJ, et al. Gaucher disease glucocerebrosidase and α-synuclein form a bidirectional pathogenic loop insynucleinopathies. Cell. 2011 Jul 8;146(1):37-52.

Meltzer PS. Cancer genomics: small RNAs with big impacts. Nature. 2005 Jun 9;435(7043):745-6.

Mercado G, Castillo V, Soto P, Sidhu A. ER stress and Parkinson's disease: Pathological inputs that converge into the secretory pathway. Brain Res. 2016 Oct 1;1648(Pt B):626-32.

Milligan MJ, Lipovich L. Pseudogene-derived lncRNAs: emerging regulators of gene expression. Front Genet. 2015 Feb 4;5:476. doi: 10.3389/fgene.2014.00476.

Minami A, Nakanishi A, Matsuda S, Kitagishi Y, Ogura Y. Function of α-synuclein and PINK1 in Lewy body dementia. Int J Mol Med. 2015 Jan;35(1):3-9.

Miñones-Moyano E, Porta S, Escaramís G, Rabionet R, Iraola S, et al. MicroRNA profiling of Parkinson's disease brains identifies early downregulation of miR-34b/c which modulate mitochondrial function. Hum Mol Genet. 2011 Aug 1;20(15):3067-78.

Mitrovich QM, Anderson P. mRNA surveillance of expressed pseudogenes in C. elegans. Curr Biol. 2005 May 24;15(10):963-7.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009 Sep 10;461(7261):272−6.

Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, et al. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA. 2007 Nov 13(11):1894-910.

Obeso JA, Rodriguez-Oroz MC, Goetz CG, Marin C, Kordower JH, et al. Missing pieces in the Parkinson's disease puzzle. Nat Med 16, 653−61 (2010).

Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, et al. The regulatory activity of microRNA* species has substantial influence on microRNA and 3'UTR evolution. Nat Struct Mol Biol. 2008 Apr 15(4):354-63.

Oliveira LM, Falomir-Lockhart LJ, Botelho MG, Lin KH, Wales P, et al. Elevated α-synuclein caused by SNCA gene triplication impairs neuronal differentiation and maturation in Parkinson's patient-derived induced pluripotent stem cells. Cell Death Dis. 2015 Nov 26;6:e1994.

Packer AN, Xing Y, Harper SQ, Jones L, Davidson BL. The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease. J Neurosci. 2008 Dec 31;28(53):14341-6.

Palmieri M, Impey S, Kang H, di Ronza A, Pelz C, Sardiello M, Ballabio A. Characterization of the CLEAR network reveals an integrated control of cellular clearance pathways. Hum Mol Genet. 2011 Oct 1;20(19):3852-66.

Paraboschi EM, Rimoldi V, Soldà G, Tabaglio T, Dall'Osso C, et al. Functional variations modulating PRKCA expression and alternative splicing predispose to multiple sclerosis. Hum Mol Genet. 2014 Jul 30.

Patil M, Sheth KA, Krishnamurthy AC, and Devarbhavi H. A Review and Current Perspective on Wilson Disease. J Clin Exp Hepatol. 2013 Dec; 3(4): 321–336.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, et al.A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010 Jun 24; 465(7301):1033-8.

Qiu XB, Shao YM, Miao S, Wang L. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. Cell Mol Life Sci. 2006 Nov;63(22):2560-70.

Quadri M, Fang M, Picillo M, Olgiati S, Breedveld GJ, et al. Mutation in the SYNJ1 gene associated with autosomal recessive, early-onset Parkinsonism. Hum Mutat. 2013 Sep;34(9):1208-15.

Rajput A, Kishore A, Snow B, Bolton CF, Rajput AH. Dopa-responsive, nonprogressive juvenile parkinsonism: report of a case. Mov Disord. 1997 May;12(3):453-6.

Rakovic A, Seibler P, Klein C. iPS models of Parkin and PINK1. Biochem Soc Trans. 2015 Apr;43(2):302-7.

Ritz B, Lee PC, Lassen CF, Arah OA. Parkinson disease and smoking revisited: ease of quitting is an early sign of the disease. Neurology. 2014 Oct 14;83(16):1396-402.

Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell. 2011 Aug 5;146(3):353-8.

Sardi SP, Clarke J, Kinnecom C, Tamsett TJ, Li L, et al. CNS expression of glucocerebrosidase corrects alpha-synuclein pathology and memory in a mouse model of Gaucher-related synucleinopathy. Proc Natl Acad Sci USA. 2011 Jul 19;108(29):12101-6.

Sardi SP, Clarke J, Viel C, Chan M, Tamsett TJ, et al. Augmenting CNS glucocerebrosidase activity as a therapeutic strategy for parkinsonism and other Gaucher-related synucleinopathies. Proc Natl Acad Sci USA. 2013 Feb.

Sardiello M, Palmieri M, di Ronza A, Medina DL, Valenza M, et al. A gene network regulating lysosomal biogenesis and function. Science. 2009 Jul 24;325(5939):473-7.

Schapira AH, Gegg ME. Glucocerebrosidase in the pathogenesis and treatment of Parkinson disease. Proc Natl Acad Sci U S A. 2013 Feb 26;110(9):3214-5.

Schöndorf DC, Aureli M, McAllister FE, Hindley C, Mayer F, et al. iPSC-derived neurons from GBA1-associated Parkinson's disease patients show autophagic defects and impaired calcium homeostasis. Nat Commun. 2014 Jun 6;5:4028.

Schrag A, Schott JM. Epidemiological, clinical, and genetic characteristics of early-onset parkinsonism. Lancet Neurol. 2006 Apr; 5(4):355-63.

Seibler P1, Graziotto J, Jeong H, Simunovic F, Klein C, Krainc D. Mitochondrial Parkin recruitment is impaired in neurons derived from mutant PINK1 induced pluripotent stem cells. J Neurosci. 2011 Apr 20;31(16):5970-6.

Seitz H. Redefining microRNA targets. Curr Biol. 2009 May 26;19(10):870-3.

Sidransky E, Lopez G. The link between the GBA gene and parkinsonism. Lancet Neurol. 2012 Nov 11(11):986-98.

Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. N Engl J Med. 2009 Oct 361(17):1651-61.

Siebert M, Westbroek W, Chen YC, Moaven N, Li Y, Velayati A, et al. Identification of miRNAs that modulate glucocerebrosidase activity in Gaucher disease cells. RNA Biol 2014;11:1291-300.

Singleton AB, Farrer MJ, Bonifati V. The genetics of Parkinson's disease: progress and therapeutic implications. Mov Disord. 2013 Jan 28(1):14-23.

Song W, Wang F, Savini M, Ake A, di Ronza A, et al. TFEB regulates lysosomal proteostasis. Hum Mol Genet. 2013 May 15;22(10):1994-2009.

Spatola M, Wider C. Genetics of Parkinson's disease: the yield. Parkinsonism. Relat Disord. 2014 Jan;20 Suppl 1:S35-8.

Spillantini MG. Parkinson's disease, dementia with Lewy bodies and multiple system atrophy are alpha-synucleinopathies. Parkinsonism Relat Disord. 1999 Dec 5(4):157-62.

Srinivasan S, Selvan ST, Archunan G, Gulyas B, Padmanabhan P. MicroRNAs -the Next Generation Therapeutic Targets in Human Diseases. Theranostics. 2013 Nov 29;3(12):930-942.

Svobodová E, Mrázová L, Lukšan O, Elstein D, Zimran A, et al. Glucocerebrosidase gene has an alternative upstream promoter, which has features and expression characteristic of housekeeping genes. Blood Cells Mol Dis. 2011 Mar 15;46(3):239-45.

Swan M, Saunders-Pullman R. The association between β-glucocerebrosidase mutations and parkinsonism. Curr Neurol Neurosci Rep. 2013 Aug 13(8):368.

Talebizadeh Z, Butler MG, Theodoro MF. Feasibility and relevance of examining lymphoblastoid cell lines to study role of microRNAs in autism. Autism Res. 2008 Aug;1(4):240-50.

Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014 Jan 16;505(7483):344-52.

Thomson DW, Dinger ME. Endogenous microRNA sponges: evidence and controversy. Nat Rev Genet 2016; 17: 272-83.

Tognini P, Putignano E, Coatti A, Pizzorusso T. Experience-dependent expression of miR-132 regulates ocular dominance plasticity. Nat Neurosci. 2011 Sep 4;14(10):1237-9.

Trotta L, Guella I, Soldà G, Sironi F, Tesei S, et al. SNCA and MAPT genes: Independent and joint effects in Parkinson disease in the Italian population. Parkinsonism Relat Disord. 2012 Mar;18(3):257-62.

Tsujii S, Ishisaka M, Shimazawa M, Hashizume T, Hara H. Zonisamide suppresses endoplasmic reticulum stress-induced neuronal cell damage in vitro and in vivo. Eur J Pharmacol. 2015 Jan 5;746:301-7.

Valdés P1, Mercado G, Vidal RL, Molina C, Parsons G, et al. Control of dopaminergic neuron survival by the unfolded protein response transcription factor XBP1. Proc Natl Acad Sci U S A. 2014 May 6;111(18):6804-9.

Vilariño-Güell C, Rajput A, Milnerwood AJ, Shah B, Szu-Tu C, et al. DNAJC13 mutations in Parkinson disease. Hum Mol Genet. 2014 Apr 1;23(7):1794-801.

Vilariño-Güell C, Wider C, Ross OA, Dachsel JC, Kachergus JM, et al. VPS35 mutations in Parkinson disease. Am J Hum Genet. 2011 Jul 15;89(1):162-7.

Wafaei JR, Choy FY. Glucocerebrosidase recombinant allele: molecular evolution of the glucocerebrosidase gene and pseudogene in primates. Blood Cells Mol Dis. 2005 Sep-Oct;35(2):277-85.

Wanet A, Tacheny A, Arnould T, Renard P. miR-212/132 expression and functions: within and beyond the neuronal compartment. Nucleic Acids Res. 2012 Jun;40(11):4742-53.

Wang J, Liu X, Wu H, Ni P, Gu Z, et al. CREB up-regulates long non-coding RNA, HULC expression through interaction withmicroRNA-372 in liver cancer. Nucleic Acids Res. 2010 Sep 38(16):5366-83.

Wang L, Guo ZY, Zhang R, Xin B, Chen R, et al. Pseudogene OCT4-pg4 functions as a natural micro RNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. Carcinogenesis 2013;34:1773-81.

Westbroek W, Gustafson AM, Sidransky E. Exploring the link between glucocerebrosidase mutations and parkinsonism. Trends Mol Med 2011;17:485-93.

Wilhelmus MM, Nijland PG, Drukarch B, de Vries HE, van Horssen J. Involvement and interplay of Parkin, PINK1, and DJ1 in neurodegenerative and neuroinflammatory disorders. Free Radic Biol Med. 2012 Aug 15;53(4):983-92.

Winder-Rhodes SE, Garcia-Reitböck P, Ban M, Evans JR, Jacques TS, et al. Genetic and pathological links between Parkinson's disease and the lysosomal disorder Sanfilippo syndrome. Mov Disord. 2012 Feb;27(2):312-5.

Xu D, Takeshita F, Hino Y, Fukunaga S, Kudo Y, et al. miR-22 represses cancer progression by inducing cellular senescence. J Cell Biol. 2011 Apr 18;193(2):409-24.

Yang CP, Zhang ZH, Zhang LH, Rui HC. Neuroprotective Role of MicroRNA-22 in a 6-Hydroxydopamine-Induced Cell Model of Parkinson's Disease via Regulation of Its Target Gene TRPM7. J Mol Neurosci. 2016 Dec;60(4):445-452.

Yin XM, Ding WX. The reciprocal roles of PARK2 and mitofusins in mitophagy and mitochondrial spheroid formation. Autophagy. 2013 Nov 1;9(11):1687-92.

Yu H, Wu M, Zhao P, Huang Y, Wang W, Yin W. Neuroprotective effects of viral overexpression of microRNA-22 in rat and cell models of cerebral ischemia-reperfusion injury. J Cell Biochem 2015;116:233-41.

Zheng L, Li X, Gu Y, Lv X, Xi T. The 3'UTR of the pseudogene CYP4Z2P promotes tumor angiogenesis in breast cancer by acting as a ceRNA for CYP4Z1. Breast Cancer Res Treat 2015;150:105-18.

Zheng Y, Xu Z. MicroRNA-22 induces endothelial progenitor cell senescence by targeting AKT3. Cell Physiol Biochem 2014;34:1547-55.

Zimprich A, Benet-Pagès A, Struhal W, Graf E, Eck SH, et al. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. Am J Hum Genet. 2011 Jul 15;89(1):168-75.

6 | Software & database links

BWA software: http://bio-bwa.sourceforge.net/

GATK software: http://www.broadinstitute.org/gatk

Annovar software: http://annovar.openbioinformatics.org/en/latest/

Galaxy software: https://usegalaxy.org/

SIFT software: http://sift.jcvi.org/

Polyphen-2 software: http://genetics.bwh.harvard.edu/pph2/

dbSNP database: http://www.ncbi.nlm.nih.gov/SNP

1000Genomes databases: http://www.internationalgenome.org/1000-genomes-browsers/

ExAC databases: http://exac.broadinstitute.org/

LOVD database: http://www.lovd.nl/3.0/home

HGMD database: http://www.hgmd.cf.ac.uk/ac/index.php

NNSPLICE software: http://www.fruitfly.org/seq_tools/splice.html

NetGene2: http://www.cbs.dtu.dk/services/NetGene2/

UCSC database: https://genome.ucsc.edu/

Part II

Content

Straniero L, Rimoldi V, Samarani M, Goldwurm S, Di Fonzo A, Krüger R, Deleidi M, Aureli M, Soldà G, Duga S, Asselta R. "The GBAP1 pseudogene acts as a ceRNA for the Parkinson-related gene GBA by sponging miR-22-3p". [Submitted to Scientific Reports]

The *GBAP1* pseudogene acts as a ceRNA for the Parkinson-related gene *GBA* by sponging miR-22-3p

Letizia Straniero,[1,2] Valeria Rimoldi,[2,3] Maura Samarani,[1] Stefano Goldwurm,[4] Alessio Di Fonzo,[5] Rejko Krüger,[6] Michela Deleidi,[7] Massimo Aureli,[1] Giulia Soldà,[2,3,*] Stefano Duga,[2,3] Rosanna Asselta[2,3]

[1] Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano, Milano, Italia
[2] Department of Biomedical Sciences, Humanitas University, Via Manzoni 113, 20089 Rozzano, Milan, Italy
[3] Humanitas Clinical and Research Center, Via Manzoni 56, 20089 Rozzano, Milan, Italy
[4] Parkinson Institute, ASST "Gaetano Pini-CTO", Milan, Italy
[5] IRCCS Foundation Ca' Granda Ospedale Maggiore Policlinico, Dino Ferrari Center, Neuroscience Section, Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
[6] Clinical and Experimental Neuroscience, Luxembourg Center for Systems Biomedicine (LCSB), University of Luxembourg and Centre Hospitalier de Luxembourg (CHL), Luxembourg
[7] German Centre for Neurodegenerative Diseases (DZNE) Tübingen within the Helmholtz Association, Tübingen, Germany; Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

Running title: An RNA-based circuit controlling *GBA* expression

***Corresponding author**:
Giulia Soldà
Department of Biomedical Sciences
Humanitas University
Via A. Manzoni 133 - 20089
Rozzano (Milano) Italy
Phone: +390282245214
Fax: +390282245290
E-mail: giulia.solda@hunimed.eu

Conflict of interest statement: The Authors declare no conflicts of interest.

ABSTRACT

Parkinson's disease is a neurodegenerative disorder characterized by the loss of dopaminergic neurons of the *substantia nigra*, which causes motor impairment and resting tremor. Currently, mutations in the *GBA* gene, encoding the glucocerebrosidase (GCase) lysosomal enzyme, represent the most frequent cause of Parkinson's disease. GCase deficiency was demonstrated in the brain of Parkinson's disease patients carrying *GBA* mutations, and, to a lesser extent, also in patients without *GBA* mutations, suggesting that dysregulated *GBA* expression may have a wider impact in disease pathogenesis. Since the increase of GCase activity in the brain has been proposed as a therapy to reduce alpha-synuclein accumulation and to revert symptoms, a detailed knowledge on the mechanisms controlling *GBA* expression might help design novel therapeutics.

Here, we explored the possible existence of a regulatory network involving *GBA*, microRNAs (miRNAs), and competing-endogenous RNAs (ceRNAs). ceRNAs are post-transcriptional regulators that titrate away miRNAs from their targets, thus upregulating mRNA expression. The highly-homologous and expressed *GBA* pseudogene (*GBAP1*) is particularly suited to act as a *GBA* ceRNA. To verify this hypothesis, we bioinformatically selected miRNAs potentially targeting both transcripts, and demonstrated that miR-22-3p directly targets both *GBA* and *GBAP1*, and significantly decreases their endogenous mRNA levels up to 70%. Over-expression experiments showed that the *GBAP1* 3'-untranslated region is able to "sponge" miR-22-3p, partially releasing *GBA* from the miR-22-3p control and thus increasing both *GBA* mRNA and GCase levels. The characterization of *GBAP1* splicing pattern identified multiple out-of-frame isoforms that are down-regulated by the nonsense-mediated mRNA decay (NMD) pathway, suggesting that *GBAP1* levels and, accordingly, its ceRNA effect, are significantly modulated by NMD. In an attempt to confirm these data in a more relevant cell model system, we measured miR-22-3p/*GBA*/*GBAP1* levels in induced pluripotent stem (iPS) cells derived from skin fibroblasts of Parkinson's disease patients with *GBA* mutations and controls. This analysis showed a significant up-regulation of *GBA* during the dopaminergic differentiation, which was nicely paralleled by anticorrelated miR-22-3p levels. Moreover, *GBA* levels in iPS-derived dopaminergic neurons were lower in Parkinson's disease patients compared to controls.

Altogether, our results describe the first miRNA controlling *GBA* levels and suggest that *GBAP1* can be considered a long non-coding RNA acting as a *GBA* ceRNA.

Key words: Parkinson's disease, *GBA*, pseudogene, competing endogenous RNA, induced pluripotent stem cells.

INTRODUCTION

Parkinson's disease is a neurodegenerative disorder affecting approximately 0.3% of the general population and more than 1% of people over 60 (De Lau and Breteler, 2006). It is characterized by depigmentation of the *substantia nigra* (SN) caused by the selective and progressive loss of dopaminergic (DA) neurons, and by the presence of intraneuronal inclusions known as Lewy bodies (LB) within the surviving neurons of the SN and other brain regions (Spillantini, 1999). Clinical features of Parkinson's disease include motor impairment, encompassing resting tremor, bradykinesia, postural instability, and rigidity, along with non-motor symptoms, such as autonomic, cognitive, and psychiatric problems (Fahn, 2003).

Parkinson's disease is the result of an interaction between multiple environmental factors and inherited genetic susceptibility (Klein and Schlossmacher, 2007). Linkage analyses, mainly performed in families with a pseudo-Mendelian transmission of the disease, identified at least 14 genes (Lesage and Brice, 2009; Lin and Farrer, 2014). The *SNCA* gene was the first to be discovered thanks to such kind of studies: it codes for α-synuclein, the principal component of LB; patients with duplications or triplications of this gene show a correlation between the quantity of expressed wild-type protein and the severity of the disease (Eriksen *et al.*, 2005). More recently, several genome-wide association studies and large-scale meta-analyses have been performed to identify risk factors for Parkinson's disease: these studies strongly stress the pivotal role of *SNCA*, *MAPT*, and *GBA* genes in disease susceptibility (Pankratz *et al.*, 2012; Nalls *et al.*, 2014 and references therein).

Until 2004 the *GBA* gene, coding for the enzyme glucocerebrosidase (GCase), was considered only responsible for the Gaucher's disease, one of the most common lysosomal storage diseases (Sidransky, 2012). GCase is mainly a lysosomal enzyme and only partly associated with the outer surface of the cell membrane (Aureli *et al.*, 2012). GCase catalyzes the hydrolysis of the membrane glucosylceramide (GlcCer) to ceramide and glucose, and its deficiency leads to the accumulation of the substrate, responsible for the multi-organ clinical manifestations of Gaucher's disease (De Fost *et al.*, 2003). Importantly, Gaucher's and Parkinson's diseases have been connected on account of the clinical observation of parkinsonism and LB pathology in patients with Gaucher's disease (Lwin *et al.*, 2004). Compared with the general population, patients with the milder form of Gaucher's disease (type 1) have a 20-fold increased lifetime risk of developing parkinsonism (Westbroek *et al.*, 2011), whereas individuals carrying heterozygous *GBA* mutations, have five times greater risk of developing Parkinson's disease than non-carrier individuals (Sidransky *et al.*, 2009). Several studies confirmed that *GBA* mutations, in particular the two most common ones (p.N370S and p.L444P), are more frequent in Parkinson's disease patients than in healthy controls, demonstrating that genetic lesions in this gene are a common risk factor for the disease (International Parkinson Disease Genomics Consortium, 2011; Asselta *et al.*, 2014). Recently, we proved the strong effect of GBA mutations also in PD progression and survival (Cilia *et al.*, 2016).

Despite many efforts, the mechanism underlying the relation between *GBA* mutations and the development of Parkinson's disease remains unclear. There are studies supporting a gain-of-function effect of the mutated protein (promoting α-synuclein aggregation), as well as others supporting a loss-of-function mechanism

(leading to substrate accumulation, and hence affecting α-synuclein processing and clearance) (Sidransky and Lopez, 2012). Importantly, not only a widespread deficiency of GCase activity has been demonstrated in the brains of Parkinson's disease patients carrying *GBA* mutations, but also Parkinson's disease patients without *GBA* mutations were shown to exhibit deficiency of GCase in SN as well as in blood (Gegg *et al.*, 2012; Alcalay *et al.*, 2015). Moreover, neurons and brains of Parkinson's disease patients showed accumulation of GlcCer that directly influences the abnormal lysosomal storage of α-synuclein oligomers, thus resulting in a further inhibition of the GCase activity. These findings suggested that the bi-directional effect of GlcCer and α-synuclein accumulation forms a positive feedback loop that may lead to a self-propagating disease (Mazzulli *et al.*, 2011). Recent data also linked GCase impairment to the cell-to-cell propagation of α-synuclein aggregates (Bae *et al.*, 2014). Based on the above-mentioned evidence, it is plausible that dysregulated *GBA* levels could represent a common feature in Parkinson's disease, whereas loss-of-function *GBA* mutations could constitute the specific trigger responsible for Parkinson's disease development in the *GBA*-associated disease.

Dysregulation of *GBA* expression may in theory be due to altered epigenetic, transcriptional, and/or post-transcriptional regulatory mechanisms. In particular, RNA-based networks, characterized by interactions between a specific mRNA, microRNAs (miRNAs), and competing endogenous RNAs (ceRNAs), are emerging as post-transcriptional regulators of gene expression (Tay *et al.*, 2014). Moreover, accumulating evidence points to deregulation of noncoding RNAs as an important and largely unexplored regulatory layer in human neurodegenerative disorders, like Parkinson's disease (Cooper *et al.*, 2009; Saugstad, 2010).

MiRNAs are ~20-nucleotide-long regulatory RNAs that act as post-transcriptional regulators of gene expression by repressing target mRNAs translation and/or by inducing mRNA degradation. About 2000 miRNAs have been experimentally validated in humans and many more have been predicted bioinformatically, making them a major class of regulators (Kozomara and Griffiths-Jones, 2014). Each miRNA might inhibit the expression of multiple target mRNAs, whose recognition is based on imperfect complementary binding between miRNAs and their target sites, usually located within the 3' untranslated region (3'UTR) (Lim *et al.*, 2005). Concerning ceRNAs, they were recently described as a novel category of regulatory RNAs: these transcripts compete with mRNAs for miRNAs, acting as molecular "sponges" and thus influencing mRNA levels (Tay *et al.*, 2014). Pseudogenes are the best ceRNA candidates, since they have a high-sequence identity with the ancestral gene (and, consequently, they could be targets of the same miRNAs), they can be transcribed, but usually they have lost the ability to generate a functional protein product (Poliseno *et al.*, 2010). Interestingly, a highly-homologous (96% sequence identity) expressed *GBA* pseudogene (*GBAP1*) is located 16 kb downstream of the functional gene (Horowitz *et al.*, 1989). *GBAP1* originated from a recent duplication event, being present only in primates (Martínez-Arias *et al.*, 2001).

With the aim to better understand *GBA* expression regulation at the post-transcriptional level, we explored the possible existence of a ceRNA-based network involving *GBA* and *GBAP1*. Here, we demonstrated that *GBAP1* may function as a ceRNA to regulate *GBA* expression by sponging miR-22-3p, thus enlightening a novel regulatory circuit that can play a role in the pathogenesis of Parkinson's disease

MATERIALS AND METHODS

Plasmid constructs

MiR-22-3p and miR-132 precursors were inserted into the psiUX expression vector (kindly provided by Prof. I. Bozzoni, Università di Roma La Sapienza, Rome, Italy). *GBA* and *GBAP1* 3'UTRs were directionally cloned downstream of the renilla luciferase gene in the psiCHECK2 reporter plasmid (Promega, Madison, USA). All constructs were produced by PCR amplifying the relevant genomic region from the DNA of a healthy subject using an appropriate PCR primer couple (Supplementary Table 1), and subsequently by cutting the amplified products with the proper restriction enzyme. Restricted products were ligated into the relevant plasmid.

The constructs carrying the *GBA* and *GBAP1* 3'UTR deleted of the miR-22-3p binding site (ΔMRE, miRNA recognition element) were obtained by site-directed mutagenesis, by means of the QuikChange kit (Agilent Technologies, Santa Clara, USA), following the manufacturer protocol.

A luciferase construct containing miR-22-3p antisense sequences (miR-22-3p sensor), kindly provided by Dr. Da-Zhi Wang (Children's Hospital Boston and Harvard Medical School), was used as a positive control (Huang *et al.*, 2013).

All plasmids were purified using the PureYield™ Plasmid Miniprep System kit (Promega) according to the manufacturer's instructions. All recombinant and mutagenized vectors were verified by conventional Sanger sequencing, as described (Asselta *et al.*, 2014).

Prediction of *GBA/GBAP1*-targeting miRNAs

Predictions were performed using publicly-available algorithms: microRNA.org (Betel *et al.*, 2008), MicroCosm Targets (Griffiths-Jones *et al.*, 2008), PITA (Kertesz *et al.*, 2007), as well as the miRWalk2 suite (Dweep and Gretz, 2015).

Cell cultures and transfection experiments

HEK293 and HeLa cells were cultured according to the standard procedures.

For miRNA over-expression experiments, cells were cotransfected using 3.5 μg (HeLa) or 875 ng (HEK293) of the psiUX plasmid expressing either miR-22-3p or miR-132 precursors.

For the miRNA-target interaction analysis, HEK293 cells were cotransfected using 300 ng of the psiUX plasmid expressing miR-22-3p together with 720 ng of the psiCHECK2 plasmid containing the relevant 3'UTR.

For the ceRNA-effect analysis, HEK293 cells were cotransfected using 300 ng of the psiUX plasmid expressing miR-22-3p together with 300 ng of the psiCHECK2 plasmid containing the *GBAP1* 3'UTR, whereas HepG2 cells were transfected only with 300 ng of the *GBAP1* 3'UTR.

In each experiment, an equal number of cells ($2.5*10^5$ for HeLa, $3*10^5$ for HEK293, $4*10^5$ for HepG2) were transfected with the Polyplus jetPRIME (EuroClone, Wetherby, UK) in 6-well plates, as described by the manufacturer. Depending on the measurement to be performed at the end of experiment, cells were collected 24, 48, 72, or 96 hours after transfection (detailed in the relevant figure legend), to obtain either total RNA, or cell lysates (see below).

RNA samples

Expression profiles of *GBA*, *GBAP1*, and miR-22-3p were determined using RNA from: a panel of 20 human tissues (First Choice total RNA; Ambion, Austin, USA), a panel of 24 human cerebral regions (Clontech Laboratories, Palo Alto, USA), 11 cell lines, induced pluripotent stem cells (iPSCs), and DA neurons differentiated from iPSCs (see below).

RNA from cell lines, iPSCs, DA neurons, as well as transfected cells was isolated using the Eurozol kit (Euroclone), according to the manufacturer's protocol. RNA concentration/quality was assessed using the Nanodrop ND-1000 (Thermo Fisher Scientific, Waltham, USA).

Semi-quantitative real-time RT-PCR

For the evaluation of expression levels of specific genes, random hexamers and the Superscript-III Reverse Transcriptase (Invitrogen, Carlsbad, USA) were used to perform first-strand cDNA synthesis starting from 1 µg of RNA extracted from cells, or RNA derived from a panel of human tissues. From a total of 20 µL of the reverse-transcription (RT) reaction, 1 µL was used as template for amplifications using the FastStart SYBR Green Master Mix (Roche, Basel, Switzerland) on a LightCycler 480 (Roche), following a touchdown thermal protocol. Expression levels were normalized using *HMBS* (hydroxymethylbilane synthase gene) and *ACTB* (β-actin) as housekeeping genes. To discriminate between the quasi-identical *GBA* and *GBAP1* genes, we took advantage of the 55-bp deletion in exon 9 characterizing *GBAP1* as well as of the few nucleotide differences between *GBA* and *GBAP1* spread along the two genes.

MiR-22-3p and miR-132 levels were measured by real-time RT-PCR by a poly(A) tailing and a universal reverse transcription approach, using the miRNA First Strand Synthesis kit (Agilent Technologies) and starting from 300 ng of total RNA, according to the manufacturer's instructions. RT-PCR reactions were performed using the universal reverse primer (Agilent Technologies) and miRNA-specific forward primers, as described (Soldà et al., 2012). U6 snRNA was used as housekeeping gene. Real-time reactions were performed as described above.

In all cases, real-time RT-PCR assays were performed at least in triplicate on a LightCycler 480, and expression levels were analyzed by the GeNorm software (Vandesompele *et al.*, 2002). Correlation between *GBA*/*GBAP1*/miR-22-3p expression profiles was calculated using the Pearson's correlation. Pearson's coefficients <-0.5 and >0.5 are considered as anti-correlation and positive correlation, respectively. P values <0.05 were considered as statistically significant.

Primer couples used in RT-PCR assays are listed in Supplementary Table 1.

Luciferase assays

For miRNA-target interaction assays, the activities of firefly/renilla luciferase were measured in lysates from transfected cells by using the Dual-Luciferase Reporter Assay System (Promega) and the Wallac 1420 VICTOR^3 V reader (PerkinElmer, Waltham, USA). The values of renilla luciferase were normalized against the corresponding values of firefly luciferase.

GCase enzymatic activity assays

Cells to be assayed were washed twice with phosphate buffered saline (PBS), harvested, and then lysed in water containing complete protease inhibitor cocktails (Roche). Total cell protein content was measured using the Micro BCA assay reagent (Pierce, Rockford, USA). Cells lysates were transferred to a 96-well microplate and assays were performed in triplicate. Cell-lysate associated GCase activity was analyzed using 4-methylumbelliferyl- β-D-glucopyranoside (MUB-Glc; Glycosynth, Warrington, UK), solubilized at a final concentration of 6 mM in McIlvaine Buffer (0.1 M Citrate/0.2 M Phosphate, pH 5.2) containing 0.1% Triton X-100. The reaction mixtures were incubated at 37°C under gentle shaking. The fluorescence was recorded after transferring 10 μL of the mixture in the microplate and adding 190 μL of 0.25 M glycine, pH 10.7. The fluorescence was detected by a Wallac 1420 VICTOR^3 V reader. Data were expressed as pmoles of converted substrate/mg cell proteins × hour.

***GBA* and *GBAP1* splicing pattern and sensitivity to the nonsense-mediated mRNA decay (NMD) pathway**

Analysis of *GBA*/*GBAP1* splicing patterns and susceptibility to NMD was undertaken in HepG2 and HEK293 cell lines. Cells were plated at a density of $4*10^5$ per 6-well dish and, after 72 hours, treated for 8 hours with cycloheximide (100 μg/mL; dissolved in dimethyl sulfoxide) or with the vehicle alone. After the treatment, cells were washed with PBS and total RNA extracted.

For the analysis of the splicing pattern, a set of gene-specific or pseudogene-specific RT-PCR assays (Supplementary table 1) was designed to catch the vast majority of possible alternative splicing events. RT-PCRs were performed as described above. The main amplified products, recovered from the agarose gel using the Wizard SV Gel and PCR Clean-Up System kit (Promega), were directly sequenced to confirm their identity.

Variations in the expression levels of *GBA*/*GBAP1* upon treatment were quantified by real-time RT-PCR assays using as reference an NMD-resistant transcript (*i.e.*, Connexin 43 or Connexin 32 mRNAs, whose coding sequences are all contained in a single exon, for HEK293 and HepG2, respectively). The NMD-sensitive and insensitive *PRKCA* transcripts were used as controls (Paraboschi *et al.*, 2014).

Fibroblast-derived iPSCs

IPSC lines derived from skin fibroblasts of six controls and four Parkinson's disease patients carrying heterozygous *GBA* mutations (p.L444P, n=2; p.N370S, n=2) were obtained as detailed in (Schondorf *et al.*, 2014) and following the protocol of Takahashi and colleagues (2007). These iPSCs were subjected to neuronal differentiation for 35 days *in vitro*, according to Kriks and collaborators' protocol (Kriks *et al.*, 2011).

This study has the approval of the local Ethics Committees and was performed according to the Declaration of Helsinki. Signed informed consent was obtained from all participants.

RESULTS

MiR-22-3p targets *GBA* and *GBAP1*

Since there is no information on miRNAs modulating *GBA* expression, we searched bioinformatically for miRNAs potentially targeting both *GBA* and its pseudogene. Predictions were performed using eight software; candidate miRNA selection was performed by prioritizing miRNAs: i) predicted by at least five algorithms; ii) containing at least 7-nt perfect seed match with *GBA* and *GBAP1* 3'UTRs; iii) known to be expressed in brain and previously implicated in neurodegenerative diseases. These filtering steps allowed the selection of three candidate miRNAs: miR-22-3p, miR-132, and miR-212. For functional validation, we prioritized miR-22-3p and miR-132, since they were expressed at higher level in both cerebellum and frontal cortex (Supplementary Table 2).

To verify that *GBA/GBAP1* can be targets of miR-22-3p and/or miR-132, we cloned both miRNA precursors in a suitable expression vector, and over-expressed them in HeLa cells for 24 hours. We obtained, on average, an over-expression of both miRNAs of ~100 fold respect to their endogenous basal levels. The results of these experiments showed that miR-22-3p over-expression can significantly reduce *GBA* and *GBAP1* endogenous mRNA levels (up to 72%; $P<0.0003$). Conversely, no *GBA* modulation was detected after miR-132 over-expression (Figure 1A).

To confirm these results, the 3'UTRs of *GBA* and *GBAP1* were cloned downstream of the luciferase gene in the psiCHECK2 vector. These UTRs differ for only 6 nucleotides, none of them mapping in the predicted binding sites for miR-22-3p and miR-132. We cotransfected in HeLa cells each of these reporter plasmids together with the vector expressing either the miR-22-3p or miR-132 precursor. The results of transfection experiments substantially confirmed previous observations, *i.e.* miR-22-3p was able to target both *GBA* and *GBAP1* UTRs (37% and 34% reduction, respectively; $P<0.0001$). Conversely, miR-132 did not affect the expression of the reporter gene (Figure 1B), and was hence not further investigated.

To better unravel the functional impact of miR-22-3p on the expression of *GBA/GBAP1*, we decided to study miR-22-3p/*GBA/GBAP1* expression profiles in 11 cell lines. Real-time RT-PCRs evidenced a ubiquitous expression of *GBA* in the analyzed lines, with highest levels present in HeLa and glioblastoma cells, and lowest levels in HepG2 cells. *GBAP1* was present in all cell lines, though at lower levels than *GBA* (from 186 to 1.8 times less) (Supplementary figure 1A). MiR-22-3p showed a nearly ubiquitous expression profile,

with highest levels in HepG2 and glioblastoma cells, and lowest levels in HEK293 (Supplementary figure 1B).

Based on these expression profiles, we decided to repeat miR-22-3p over-expression experiments in HEK293 cells. Results were comparable to those observed in the HeLa cell line, with *GBA* and *GBAP1* endogenous mRNA levels significantly decreased, after 24 hours, up to 44% ($P<0.05$; Figure 1C). We then confirmed the effects of miR-22-3p over-expression also at the protein level: measurements of endogenous GCase activity demonstrated that the enzyme was down-regulated of 15% after 48 hours of transfection ($P<0.002$; Figure 1C).

Finally, the specific interaction of miR-22-3p with *GBA* and *GBAP1* 3'UTRs was demonstrated by deleting the miR-22-3p putative miRNA responsive element (ΔMRE) in the reporter constructs containing the relevant UTR, and subsequently cotransfecting each mutagenized plasmid together with the miR-22-3p expressing one. In these experiments, a luciferase construct containing miR-22-3p antisense sequences (miR-22-3p sensor) was used as a positive control (Huang *et al.*, 2013). Our data showed that miR-22-3p responsiveness strictly depends on the presence of the predicted responsive element in the 3'UTR, since its deletion completely abolishes the miRNA-mediated regulation (Figure 1D). As expected, the level of luciferase activity in the miR-22-3p sensor control dramatically dropped (95% reduction).

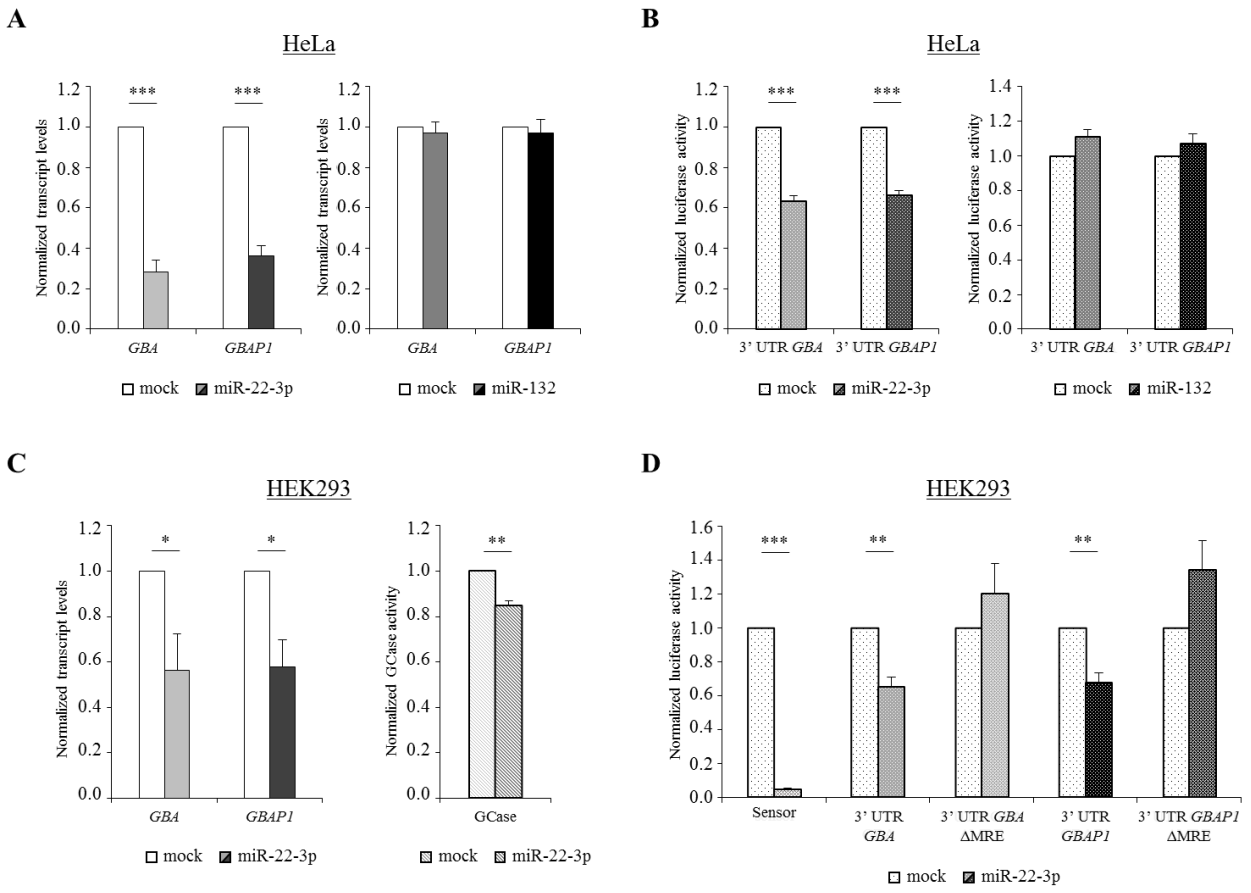


Figure 1: MiR-22-3p targets *GBA* and *GBAP1*.

A) The psiUX vectors, containing either the pre-miR-22-3p or the pre-miR-132 sequence, were independently transfected in HeLa cells; 24 hours after transfection, cells were collected and total RNA extracted. Endogenous expression levels of *GBA* and *GBAP1* were measured by semi-quantitative real-time RT-PCRs. In all cases, expression levels are shown as normalized rescaled values, setting as 1 the value measured in cell transfected with an empty vector (psiUX, mock).

B) The psiCHECK2 vectors, containing either the *GBA* or the *GBAP1* 3'UTR downstream of the luciferase reporter gene, were independently transfected in HeLa cells together with plasmids expressing either the pre-miR-22-3p or the pre-miR-132. 48 hours after transfection, cells were collected and protein lysates prepared to perform the reporter assays. The panels show the renilla luciferase activity normalized against the firefly luciferase activity, setting as 1 the value measured in cells cotransfected with an empty vector (psiCHECK2, no miRNA overexpression, mock).

C) Over-expression experiments of the psiUX vector, containing the pre-miR-22-3p sequence, were repeated in HEK293 cells. 24 or 48 hours after transfections, cells were collected for extracting total RNA (for endogenous *GBA* and *GBAP1* measurements by semi-quantitative real-time RT-PCRs; left) or for preparing protein lysates (for endogenous GCase activity measurements; right). In all cases, the value measured in cells cotransfected with an empty vector (psiUX, no miRNA over-expression, mock) was set as 1.

D) Luciferase reporter assays were repeated in HEK293 cells, by transfecting the psiCHECK2 vector coupled to the 3'UTR regions of *GBA* or *GBAP1*, with or without the putative miRNA recognition element (ΔMRE). Each of the four psiCHECK2 recombinant plasmid was cotransfected with the psiUX plasmid expressing miR-22-3p. 48 hours after transfection, cells were collected and lysates prepared to perform the reporter assays. The panels show the renilla luciferase activity normalized against the firefly luciferase activity, setting as 1 the value measured in cells cotransfected with an empty vector (psiCHECK2, no miRNA over-expression, mock). As positive control, a luciferase construct containing miR-22-3p antisense sequences (miR-22-3p sensor) (Huang *et al.*, 2013) was also transfected.

In all panels, bars represent means +SEM of three independent experiments, each performed at least in triplicate. Significance levels of t-tests are shown. *: P<0.05; **: P<0.01; ***: P<0.005.

GBAP1 acts as a ceRNA titrating miR-22-3p and up-regulating *GBA*

To assume that *GBAP1* can act as a *GBA* ceRNA by sponging miR-22-3p, we first verified the coexpression of the three transcripts in a broad range of samples (20 human tissues as well as 24 different cerebral regions). *GBA*/*GBAP1*/miR-22-3p were all ubiquitously expressed (Supplementary figure 2). In particular, *GBA* showed minimal expression in the skeletal muscle and the highest level in the medial temporal cortex (17 fold the skeletal muscle). *GBAP1* expression levels weakly correlated with those of *GBA* (Pearson's correlation coefficient of 0.53, P<0.0013), in agreement with the possible ceRNA role of *GBAP1*. Interestingly, *GBAP1* highest expression levels were registered in the brain, where the disproportion between the gene and pseudogene levels is one of the lowest (ratio 1:15). Concerning miR-22-3p, in general we observed the lowest levels of expression in brain regions, such as cerebellum, nucleus accumbens, medulla, and fetal brain. MiR-22-3p expression levels showed a tendency towards anticorrelation with the levels of *GBA*/*GBAP1* (Pearson's correlation coefficients -0.47 and -0.42, respectively; P<0.0043 and P<0.0094), in line with the demonstrated regulatory effect of miR-22-3p on *GBA*/*GBAP1*.

Such encouraging results prompted us to verify if altered levels of *GBAP1* could indeed modify the expression of *GBA*. First, as a proof of concept, we over-expressed both the 3'UTR of *GBAP1* and the miR-22-3p hairpin in HEK293 cells. In parallel, the over-expression experiment was conducted using as sponge the 3'UTR of *GBAP1* without the miR-22-3p responsive element. We showed that *GBAP1* 3'UTR over-expression causes a significant increase in the levels of endogenous *GBA* mRNA only in the presence of the miR-22-3p binding site (1.72 fold; P=0.019; Supplementary figure 3A). We also evaluated the ceRNA effect at the protein level, by measuring the GCase activity upon miR-22-3p and *GBAP1* 3'UTR over-expression. Our data confirmed that the *GBAP1* 3'UTR, containing the miR-22-3p binding site, causes a significant increase of GCase activity (1.11 fold; P=0.013) (Supplementary figure 3B).

Second, considering the high levels of miR-22-3p measured in HepG2 cells (16-fold the levels measured in HEK293; Supplementary figure 1), we over-expressed in this cell line the 3'UTR of *GBAP1* alone (with or without the miR-22-3p responsive element) and measured its effect on endogenous *GBA*. We observed a significant increase in the levels of endogenous *GBA* mRNA once again only in the presence of the miR-22-3p binding site (1.68 fold; P=0.0016; Figure 2A). The *GBAP1* ceRNA effect through miR-22-3p sponging was confirmed by measuring the expression levels of known miR-22-3p targets, *i.e.* the *SP1* and *SIRT1* genes (Xu *et al.*, 2011; Zhang et al., 2016), which both resulted upregulated of ~1.7 fold (P<0.015). Conversely, no up-regulation was observed for the *CELF1* transcript (Figure 2A), which does not contain any miR-22-3p responsive element. These results were corroborated by the measurements of GCase activity in HepG2 cells under the same experimental conditions (1.13 fold; P=0.049) (Figure 2B).

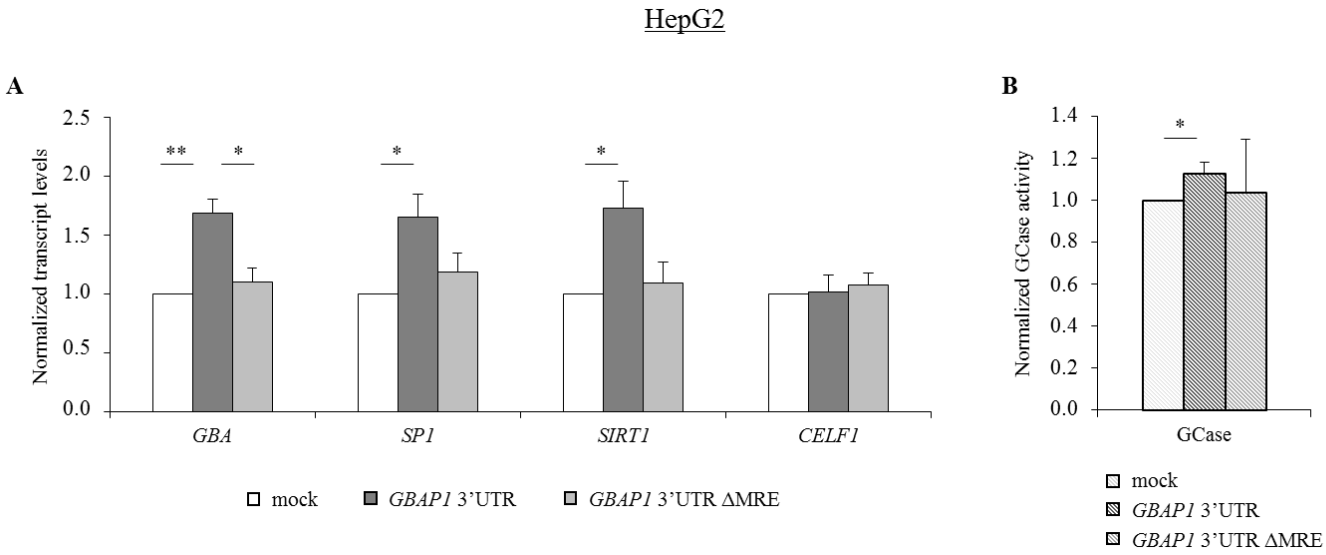


Figure 2: *GBAP1* acts as a ceRNA titrating miR-22-3p and up-regulating *GBA*.

A) The psiCHECK2 vectors, containing the *GBAP1* 3'UTR (with or without the miR-22-3p recognition element, ΔMRE) downstream of the luciferase reporter gene, were independently transfected in HepG2 cells. 96 hours after transfections, cells were collected for extracting total RNA for measurements by semi-quantitative real-time RT-PCRs of endogenous: i) *GBA*; ii) *SP1* (Sp1 Transcription Factor; known miR-22-3p target, positive control) (Xu *et al.*, 2011); iii) *SIRT1* (Sirtuin 1; known miR-22-3p target, positive control) (Zhang et al., 2016); and iv) *CELF1* (CUGBP, Elav-Like Family Member 1; negative control). The value measured in cells transfected with an empty vector (psiCHECK2, mock) was set as 1.

B) Over-expression experiments of psiCHECK2 vectors, containing the *GBAP1* 3'UTR (wild type or ΔMRE), were repeated to obtain protein lysates for endogenous GCase activity measurements. In this case, the GCase activity was measured 96 hours after transfections.

In all panels, bars represent means +SEM of three independent experiments, each performed at least in triplicate. Significance levels of t-tests are shown. *: P<0.05; **: P<0.01.

The *GBAP1* ceRNA effect could be modulated by the NMD pathway

To better unravel the reciprocal regulation of the couple *GBA*/*GBAP1*, we decided to comprehensively study *GBA* and *GBAP1* alternative splicing patterns and the possible regulation of expression of these two genes operated by NMD (Popp and Maquat, 2013; Mitrovich and Anderson, 2005).

To catch the vast majority of all possible splicing events, long-range RT-PCR assays were designed to completely cover both genes (Figure 3A). The specific amplification of either *GBA* or *GBAP1* in each assay was assured by anchoring one primer to exon 9, in correspondence of the pseudogene-specific 55-bp deletion. RT-PCR assays were performed on RNA extracted from HepG2 cells treated or not with the NMD inhibitor cycloheximide. This analysis allowed the identification of multiple alternatively-spliced isoforms for *GBAP1*; conversely, *GBA* did not show any detectable alternative isoform (Figure 3A). Notably, the heterogeneity of the splicing pattern of *GBAP1* increased after cycloheximide treatment, suggesting that multiple pseudogene splicing isoforms may be modulated by NMD. A tentative reconstruction of the main splicing variants of *GBAP1* was performed by a combination of isoform-specific semi-nested RT-PCRs and DNA sequencing, highlighting the presence of multiple transcripts containing a premature termination codon (Supplementary Figure 4).

The global effect of NMD degradation on *GBA* and *GBAP1* levels was also investigated by semi-quantitative real-time RT-PCR. This analysis showed a significant increase in the expression level of *GBAP1* in treated cells (4.18 and 3.92 fold in HEK293 and HepG2 cells, P=0.045 and P=0.0034, respectively), confirming that this pseudogene is down-regulated by NMD (Figure 3B). Also *GBA* transcripts were up-regulated upon NMD inhibition (2.28 and 2.35 fold in HEK293 and HepG2 cells), a rather unexpected result given the lack of out-of-frame *GBA* isoforms in our preliminary analysis. However, these results well fit the hypothesis that *GBAP1* levels may influence *GBA* expression through a ceRNA effect. As control, in-frame and out-of-frame *PRKCA* isoforms, known to be insensitive/sensitive to the NMD blockage (Paraboschi *et al.*, 2014), were also analyzed and gave the expected results (Figure 3B).

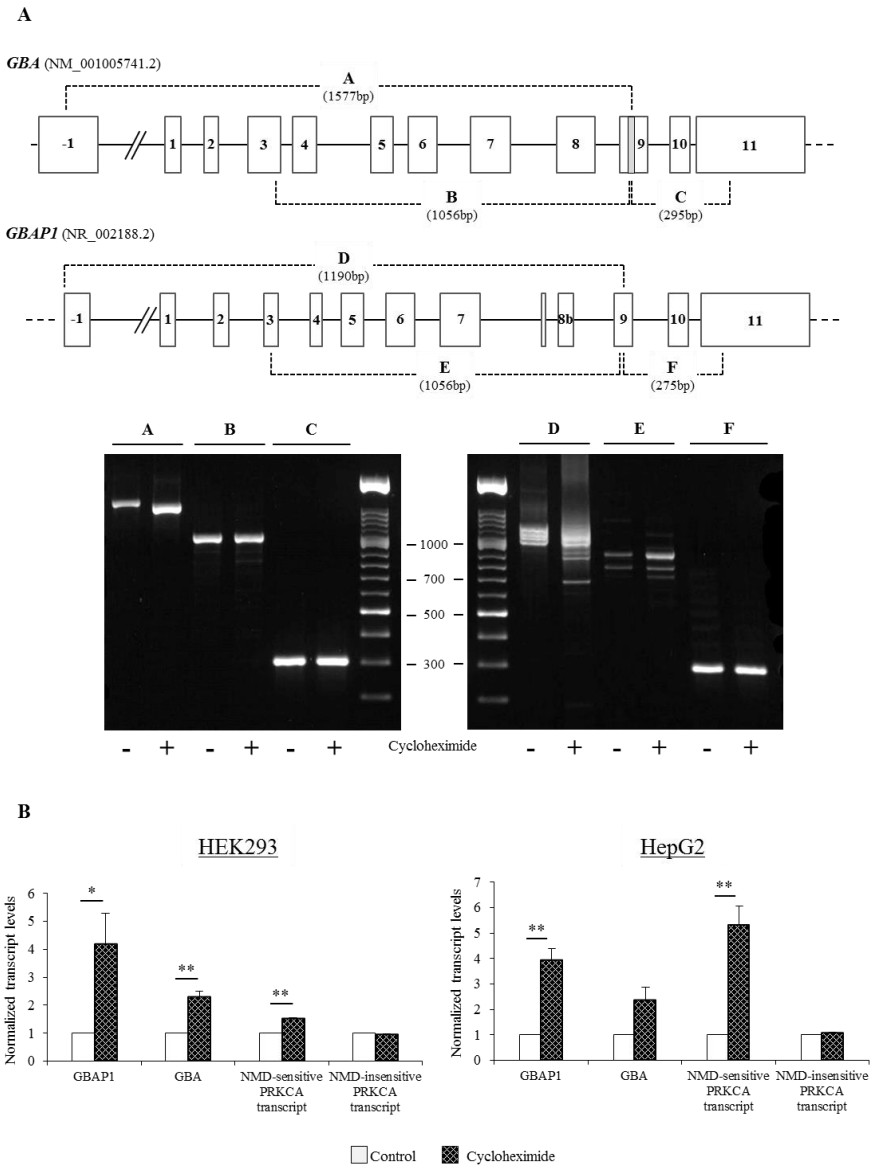


Figure 3: *GBAP1* codes for multiple alternatively-spliced isoforms and is modulated by NMD.
A) Analysis of *GBA* and *GBAP1* splicing patterns. In the upper part of the panel, a schematic representation of *GBA* (reference sequence: NM_001005741.2) and *GBAP1* (reference sequence: NR_002188.2) genes is reported. Exons are indicated by boxes, introns by line. The 55-bp-long sequence characterizing *GBA* exon 9 is specified by a grey rectangle. The scheme is approximately to scale. The overlapping fragments amplified by RT-PCRs to analyze the *GBA* and *GBAP1* splicing patterns are indicated by dashed lines and a letter. In the lower part of the panel, the electrophoretic analysis (agarose gels 2%) of RT-PCR amplicons is shown. RT-PCRs were performed on RNA extracted from HepG2 cells treated (+) or untreated (-) with the NMD inhibitor cycloheximide. On the top of each gel, letters indicate the relevant RT-PCR amplicons.
B) Demonstration of the NMD-mediated degradation of *GBAP1* transcripts. The two panel shows expression levels of *GBAP1* and *GBA* isoforms in HEK293 and HepG2 cells, untreated or treated for 8 hours with cycloheximide. Expression levels of endogenous *GBAP1*/*GBA* isoforms were measured by semi-quantitative real-time RT-PCRs (results are presented as normalized rescaled values, setting as 1 the value of the untreated samples). The expression level of the Connexin 43 or 32 transcripts, known to be insensitive to NMD, were used in the normalization step. RT-PCRs performed on out-of-frame and in-frame *PRKCA* isoforms, known to be respectively sensitive and insensitive to the NMD blockage (Paraboschi *et al.*, 2014), represent the positive and negative control. In all panels, bars represent means +SEM of three independent experiments, each performed at least in triplicate. Significance levels of t-tests are shown. *: P<0.05; **: P<0.01.

***GBA, GBP1,* and miR-22-3p are expressed in iPSC-derived neuronal cells**

To be relevant for the molecular pathogenesis of Parkinson's disease, the *GBA*/*GBAP1*/miR-22-3p network should work in tissues affected by the disease process, *e.g.* DA neurons. We hence verified the expression of *GBA*, *GBAP1*, and miR-22-3p in iPSCs and iPSC-derived neuronal cells (after 35 days of differentiation). Semi-quantitative real-time RT-PCR assays were performed on total RNA extracted from iPSCs/neurons derived from fibroblasts of six healthy controls and four Parkinson's disease patients (all carrying *GBA* mutations).

All the three players of the regulatory circuit were expressed both in iPSCs and iPSC-derived neurons, respectively. The process of differentiation towards neurons is accompanied by a significant up-regulation of *GBA* (8 fold in controls, P=0.0002; 3 fold in patients, P=0.0032) and by a parallel increase in expression levels of *GBAP1* (Figure 4A and B). In addition, we detected a significant down-regulation of the *GBA* transcript in Parkinson's disease patients respect to controls in DA neurons (0.54 fold, P=0.013). Finally, consistently with the observed up-regulation of *GBA*/*GBAP1* during neuronal differentiation, we detected lower expression levels of miR-22-3p in DA neurons respect to their precursors (0.39 fold in controls, P=0.018; 0.18 fold in patients) (Figure 4C).

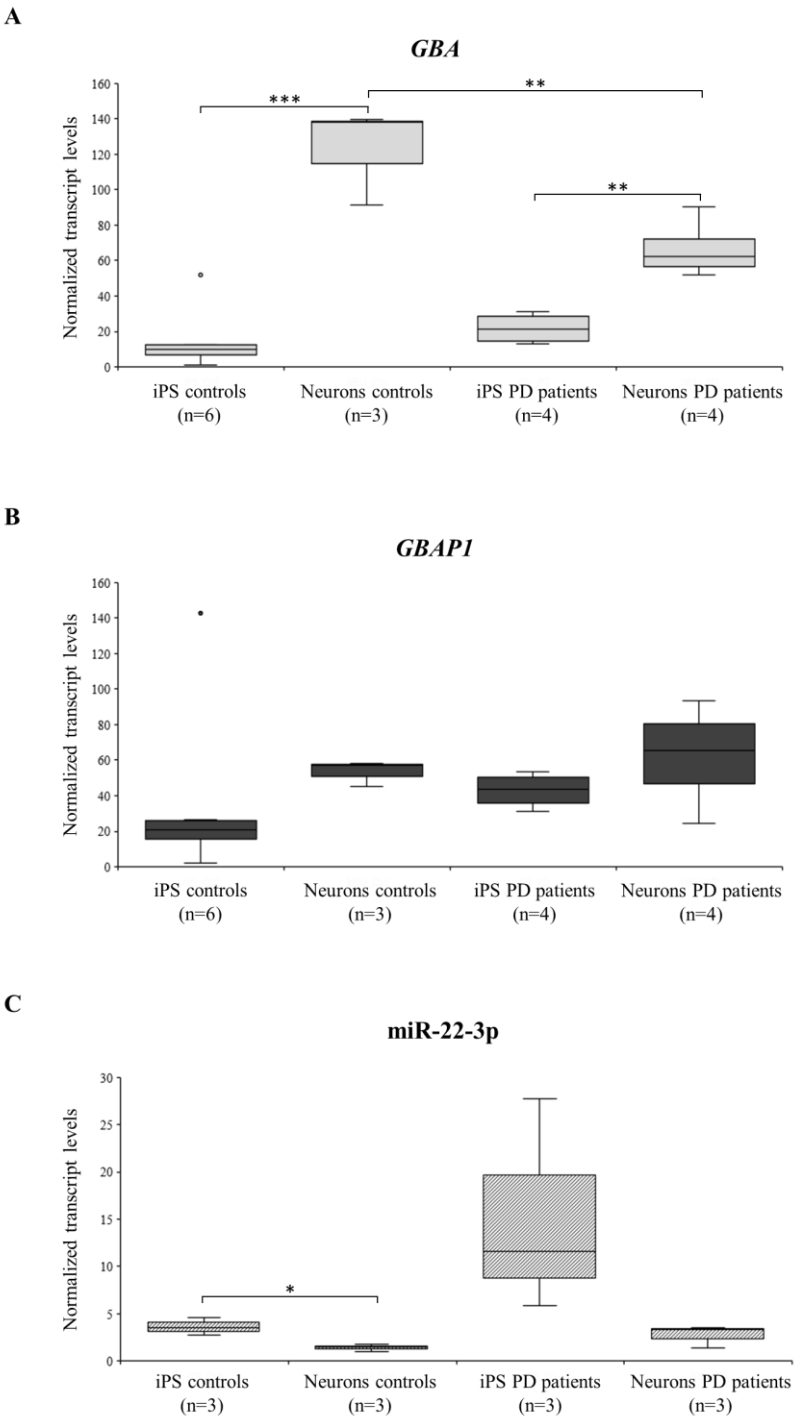


Figure 4: *GBA*, *GBP1*, and miR-22-3p are expressed in iPS cells and iPSC-derived neurons of Parkinson's disease patients and controls.

GBA (panel A), *GBAP1* (panel B), and miR-22-3p (panel C) expression levels were measured by semi-quantitative real-time RT-PCRs in up to six iPS and iPSC-derived neuronal cells of cases and controls. Boxplots show expression levels according to the disease status; boxes define the interquartile range; the thick line refers to the median. Results are presented as normalized rescaled values. Significance level for differences between groups was calculated by a t-test, and showed only if significant. *: P<0.05; **: P<0.01; ***: P<0.005.

DISCUSSION

Despite the substantial efforts, over the past few years, in understanding the role of long non-coding RNAs (lncRNAs) in health and disease, only some of them have been investigated for their biological function (Quek *et al.*, 2015). One promising –although debated- idea assigning to lncRNAs a generalized function is the "ceRNA hypothesis", based on the fact that specific RNAs can limit miRNA activity through sequestration, thus up-regulating the expression of miRNA target genes (Tay *et al.*, 2014). In particular, two classes of lncRNAs are increasingly recognized as main ceRNA contributors, *i.e.* circular RNAs and pseudogene-derived transcripts (Thomson and Dinger, 2016). Specifically concerning pseudogenes, to date there are more than 14,500 annotated pseudogenes (Gencode v24, August 2015 freeze, GRCh38; http://www.gencodegenes.org/stats/archive.html#a21), but only ~10% are actively transcribed (Sisu *et al.*, 2014), a pre-requisite for acting as ceRNAs. Indeed, transcribed pseudogenes, mostly deriving from duplication events, are considered optimal ceRNA candidates, as they share miRNA-binding sites with the ancestral genes (Tay *et al.*, 2014; Thomson and Dinger, 2016). A number of pseudogenes have been so-far experimentally demonstrated to act as ceRNAs, including: *PTENP1* and *KRAS1P* (Poliseno *et al.*, 2010), *OCT4-pg4* (Wang *et al.*, 2013), *BRAFP1* (Karreth *et al.*, 2015), and *CYP4Z2P* (Zheng *et al.*, 2015). In this study, we describe a novel ceRNA-based network - with a potential impact on Parkinson's disease - involving *GBA*, its pseudogene *GBAP1*, and miR-22-3p (Figure 5).

The molecular evolution, expression pattern, and mechanisms of transcriptional regulation of *GBA* have been previously investigated (Wafaei and Choy, 2005; Svobodová et al., 2011), mainly because of its direct link with the Gaucher's disease, the most frequently-encountered lipidosis as well as the most common inherited disorder in Ashkenazi Jews (Hruska et al., 2008). On the other hand, the few data available on *GBA* post-transcriptional regulation principally stem from a miRNA mimic screening aimed to identify miRNAs regulating the GCase activity in p.N370S homozygous Gaucher fibroblasts (Siebert et al., 2014). This screening involved 875 miRNAs and evidenced at least three candidates (miR-127-5p, miR-16-5p, and miR-195-5p), exhibiting a Z-score of at least +/- 2, with important consequences on the GCase activity. However, in all cases, the miRNA effect did not seem to be mediated by a direct interaction between the miRNA itself and *GBA*; rather, miRNAs acted either on the LIMP-2 receptor, which is involved in the trafficking of GCase from the endoplasmic reticulum to the lysosome, or on the expression levels of known modifiers of the GCase activity (Siebert et al., 2014). Hence, our work identifies miR-22-3p as the first miRNA directly targeting *GBA*. Interestingly, in the publicly-available dataset of Siebert and coworkers (2014) miR-22-3p mimic resulted to down-regulate GCase activity (Z-score=-1.5; suggestive P=0.066), according to our results.

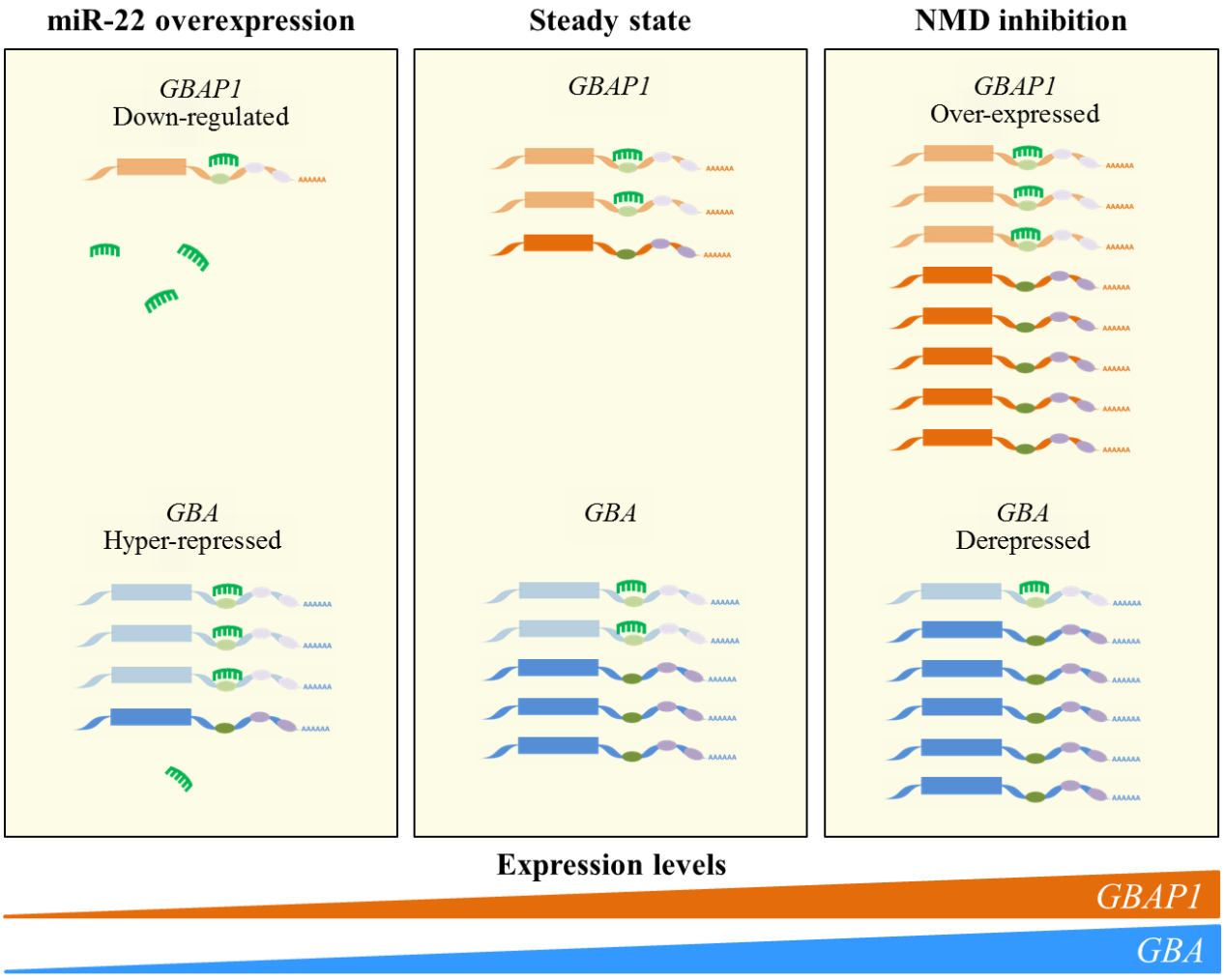


Figure 5: Schematic representation of the effect of modulating the *GBA*/*GBAP1*/miR-22-3p RNA-based network on endogenous *GBAP1* and *GBA* levels.

Schematic representation of the ceRNA network involving *GBA* (blue transcripts) and *GBAP1* (orange transcripts), harboring the same MRE sites (green and violet ovals). The green MRE sites bind to miR-22-3p (in green), whereas violet ones bind to other not-specified miRNAs. The experimental modulation of the proposed ceRNA network impacts on both coregulated transcripts. In particular, over-expression of miR-22-3p (left part of the figure) determines the down-regulation of both *GBA* and *GBAP1* transcripts. Conversely, over-expression of *GBAP1* (*e.g.*, by inhibiting the NMD pathway, as experimentally verified in the present study; right part of the figure) will increase the cellular concentrations of miR-22-3p MREs, thus resulting in the de-repression of *GBA*. In the scheme, transcripts destined to degradation are colored in lighter shades.

Concerning *GBAP1*, sparse information is available to date, and it is primarily focused on the evolution of the *GBAP1 locus*, as an example of a very recently acquired pseudogene (Horowitz *et al.*, 1989; Wafaei and Choy, 2005). We hence extensively studied *GBAP1* splicing pattern and expression profile, showing that it is subjected to multiple physiologic in-frame and out-of-frame splicing events and that it is broadly expressed, though often at low levels (Supplementary figure 2). More interestingly, we showed that *GBAP1* is targeted by NMD, which seems to be the main mechanism regulating its expression level: blocking NMD, the ratio between *GBA* and *GBAP1* substantially increased (on average from 1/100 to 1/68 in HepG2 cells, and from 1/9 to 1/4 in HEK293 cells). Of course, the *GBAP1* expression control exerted through NMD raises the question about the pseudogene translation, since RNAs should undergo a pioneer round of translation, associated with an inspection operated by the NMD machinery, before being degraded (Popp and Maquat, 2013). Interestingly, it has been recently reported that many lncRNAs, and even 5'UTRs, are translated, and *GBAP1* was identified among non-canonical human translated open reading frames (Ji *et al.*, 2015). The remarkable degradation of *GBAP1* operated by NMD not only may cause its low abundance, but also results in an increased degradation of bound miRNAs, possibly enhancing *GBAP1* efficiency as miRNA sponge, as suggested for other pseudogenes (Tay *et al.*, 2014).

The relevance of post-transcriptional regulation in determining the low *GBAP1* expression is also suggested by the observation that *GBAP1* proximal and distal promoters show high level of sequence identity with those of *GBA*, and are hence predicted to have similar transcriptional strength. For instance, the presence of two TATA boxes and two CAAT boxes in the proximal promoter of *GBAP1* exactly recapitulates the architecture of *in-cis* regulatory elements characterizing the *GBA* proximal promoter (Horowitz *et al.*, 1989). Moreover, epigenetic marks are not substantially different when comparing the gene and the pseudogene promoters, as inferred from the UCSC Genome Browser ENCODE tracks (http://genome.ucsc.edu/; release Feb. 2009, GRCh37/hg19). Indeed, our in-house preliminary data, obtained with reporter constructs, show that the activity of *GBAP1* promoters does reach the transcriptional levels of the corresponding *GBA* promoters (data not shown).

Our overexpression experiments in different cell lines clearly demonstrated that *GBAP1* 3'UTR, at supraphysiological concentrations, can modulate *GBA* mRNA levels through a miR-22-3p-mediated regulatory circuit. However, these results do not necessarily imply that this ceRNA-based regulation may also work in more physiological conditions. To confirm that *GBAP1* can act as a *GBA* ceRNA without overexpression, we exploited the predicted differential sensitivity to NMD of the gene and pseudogene transcripts (see Figure 3A). Cycloheximide treatment allowed us to increase the relative abundance of the endogenous *GBAP1* mRNA of around 4 times the basal level and was accompanied by a 2-fold increase in *GBA* transcripts, not directly attributable to NMD, and compatible with a ceRNA effect (Figure 3B). Hence, in specific cells or developmental stages, upregulation of *GBAP1*, resulting from post-transcriptional or epigenetic regulatory mechanisms, might titrate miRNAs away from the *GBA* protein-coding transcripts, thus providing a physiologic ceRNA effect.

The existence of an RNA-based network controlling *GBA* expression suggests the intriguing possibility that also miR-22-3p or *GBAP1* dysregulation could be associated with Parkinson's disease. To date, few information on a possible involvement of these non-coding RNAs in Parkinson's disease is overtly present in the literature, with the reported dysregulation of miR-22* in Parkinson's disease patients actually referring to the 5p companion of "our" miR-22-3p (Margis et al., 2011). Interestingly, a very recent paper suggested that miR-22 may exert a neuroprotective effect in Parkinson's disease, as it protects rat pheochromocytoma PC12 cells from 6-hydroxydopamine-induced injury, by modulating the levels of its target gene transient receptor potential melastatin 7 (*TRPM7*) (Yang et al., 2016). However, it should be noted that Yang and colleagues in their paper linked miR-22 down-regulation with Parkinson's disease by mistakenly citing the work of Margis and collaborators (2011), which concerned miR-22-5p, as detailed above.

To further explore the involvement of *GBA*/*GBAP1*/miR-22-3p in Parkinson's disease, we investigated their expression pattern in disease-relevant tissues using *in-silico* analyses of microarray datasets publicly available through the Gene Expression Omnibus repository (see Supplementary Materials and Methods), as well as *in-vivo* measurements performed on RNA extracted from iPSCs and iPSC-derived neuronal cells of Parkinson's disease cases and controls. In particular, we retrieved three microarray datasets evaluating differential gene expression in the SN of *post-mortem* brains, for a total of 51 cases and 42 controls (Supplementary Table 3). In the meta-analysis, we measured a significant down-regulation of both *GBA* and *GBAP1* transcripts in Parkinson's disease patients ($P<0.05$; Supplementary Figure 5). Notably, we observed the same significant down-regulation for *GBA* transcripts in iPS-derived DA neurons of Parkinson's disease patients; accordingly, miR-22-3p was slightly up-regulated in cases vs. controls (on average 1.96 fold, $P=0.13$; Figure 4).

A few studies have reported a potential neuroprotective effect of miR-22-3p both in rat models of cerebral ischemia-reperfusion injury and in Huntington and Alzheimer's disease through a reduction in inflammation and apoptosis (Jovovic et al., 2013; Yu et al., 2015). However, other studies suggested a pro-senescence role of miR-22 in endothelial progenitor cells, in cancer, and in the aging heart and brain (Li et al., 2011; Xu et al., 2011; Jazbutyte et al., 2013; Zheng and Xu, 2014). While the neuroprotective effects of miR-22 have suggested enhancing its expression as a potential therapeutic strategy for the treatment of neurodegenerative conditions, it may well be that miR-22 overexpression represents a pathophysiologic response to protect the cell from injury and stress also triggering other non-beneficial effects, like increased aging and reduced GCase activity.

In conclusion, we are aware of the fact that the connection of the here-presented RNA-based network and Parkinson's disease pathogenesis has not been formally proven. However, one can easily imagine a link between the down-regulation of the sister transcripts *GBA*/*GBAP1* - or, conversely, the up-regulation of miR-22-3p - and an aberrant α-synuclein metabolism, as already theorized (Sidransky and Lopez, 2012). A confirmed dysregulation of the *GBA*/*GBAP1*/miR-22-3p circuit in Parkinson's disease patients would suggest novel possible therapeutic strategies, based either on the direct control of the expression of the

miRNA/pseudogene, or on the modulation of the NMD pathway aimed at up-regulating *GBAP1* levels (Popp and Maquat, 2016).

Acknowledgments: Simone Digregorio, Francesca Balistreri, Nicole Tonsi, Chiara Baccin, and Giulia Rovaris are acknowledged for their invaluable work, assistance, and enthusiasm. Emanuele Frattini is specifically thanked for contributing to iPSCs and neurons preparation. Alba Bonetti, Francesca Natuzzi, Rosanna Morini, and all staff of Parkinson Institute for their effort to support the "Parkinson Institute Biobank" (http://www.parkinsonbiobank.com), member of the Telethon Network of Genetic Biobank (TELETHON Italy).

Funding: This study was supported by Fondazione Cariplo grant N°2015-1017, TELETHON Italy (project n. GTB12001), and "Fondazione Grigioni per il Morbo di Parkinson".

REFERENCES

Alcalay RN, Levy OA, Waters CC, Fahn S, Ford B, Kuo SH, Mazzoni P, Pauciulo MW, Nichols WC, Gan-Or Z, Rouleau GA, Chung WK, Wolf P, Oliva P, Keutzer J, Marder K, Zhang X. Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations. Brain 2015; 138: 2648-58.

Asselta R, Rimoldi V, Siri C, Cilia R, Guella I, Tesei S, Soldà G, Pezzoli G, Duga S, Goldwurm S. Glucocerebrosidase mutations in primary parkinsonism. Parkinsonism Relat Disord 2014; 20: 1215-20.

Aureli M, Bassi R, Loberto N, Regis S, Prinetti A, Chigorno V, Aerts JM, Boot RG, Filocamo M, Sonnino S. Cell surface associated glycohydrolases in normal and Gaucher disease fibroblasts. J Inherit Metab Dis 2012; 35: 1081-91.

Bae EJ, Yang NY, Song M, Lee CS, Lee JS, Jung BC, Lee HJ, Kim S, Masliah E, Sardi SP, Lee SJ. Glucocerebrosidase depletion enhances cell-to-cell transmission of α-synuclein. Nat Commun 2014; 5: 4755.

Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. Nucleic Acids Res 2008; 36: D149-53.

Cilia R, Tunesi S, Marotta G, Cereda E, Siri C, Tesei S, Zecchinelli AL, Canesi M, Mariani CB, Meucci N, Sacilotto G, Zini M, Barichella M, Magnani C, Duga S, Asselta R, Soldà G, Seresini A, Seia M, Pezzoli G, Goldwurm S. Survival and dementia in GBA-associated Parkinson's disease: The mutation matters. Ann Neurol 2016; 80: 662-73.

Cooper TA, Wan L, Dreyfuss G. RNA and disease [Review]. Cell 2009; 136: 777-93.

De Fost M, Aerts JM, Hollak CE. Gaucher disease: from fundamental research to effective therapeutic interventions [Review]. Neth J Med 2003; 61: 3-8.

De Lau LM, Breteler MM: Epidemiology of Parkinson's disease [Review]. Lancet Neurol 2006; 5: 525-35.

Dweep H, Gretz N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. Nat Methods 2015; 12: 697.

Eriksen JL, Przedborski S, Petrucelli L. Gene dosage and pathogenesis of Parkinson's disease [Review]. Trends Mol Med 2005; 11: 91-6.

Fahn S. Description of Parkinson's disease as a clinical syndrome [Review]. Ann N Y Acad Sci 2003; 991: 1-14

Gegg ME, Burke D, Heales SJ, Cooper JM, Hardy J, Wood NW, Schapira AH. Glucocerebrosidase deficiency in substantia nigra of parkinson disease brains. Ann Neurol 2012; 72: 455-63.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res 2008; 36: D154-8.

Horowitz M, Wilder S, Horowitz Z, Reiner O, Gelbart T, Beutler E. The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics 1989; 4: 87-96.

Hruska KS, LaMarca ME, Scott CR, Sidransky E. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA) [Review]. Hum Mutat 2008; 29: 567-83.

Huang ZP, Chen J, Seok HY, Zhang Z, Kataoka M, Hu X, Wang DZ. MicroRNA-22 regulates cardiac hypertrophy and remodeling in response to stress. Circ Res 2013; 112: 1234-43.

International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, Saad M, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet 2011; 377: 641-9.

Jazbutyte V, Fiedler J, Kneitz S, Galuppo P, Just A, Holzmann A, Bauersachs J, Thum T. MicroRNA-22 increases senescence and activates cardiac fibroblasts in the aging heart. Age 2013; 35: 747-62.

Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife 2015; 4: e08890.

Jovicic A, Zaldivar Jolissaint JF, Moser R, Silva Santos Mde F, Luthi-Carter R. MicroRNA-22 (miR-22) overexpression is neuroprotective via general anti-apoptotic effects and may also target specific Huntington's disease-related mechanisms. PLoS One 2013; 8: e54222.

Karreth FA, Reschke M, Ruocco A, Ng C, Chapuy B, Léopold V, Sjoberg M, Keane TM, Verma A, Ala U, Tay Y, Wu D, Seitzer N, Velasco-Herrera Mdel C, Bothmer A, Fung J, Langellotto F, Rodig SJ, Elemento O, Shipp MA, Adams DJ, Chiarle R, Pandolfi PP. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. Cell 2015; 161: 319-32.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007; 39: 1278-84.

Klein C, Schlossmacher MG. Parkinson disease, 10 years after its genetic revolution: multiple clues to a complex disorder [Review]. Neurology 2007; 69: 2093-104.

Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 2014; 42: D68-73.

Kriks S, Shim JW, Piao J, Ganat YM, Wakeman DR, Xie Z, Carrillo-Reid L, Auyeung G, Antonacci C, Buch A, Yang L, Beal MF, Surmeier DJ, Kordower JH, Tabar V, Studer L. Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. Nature 2011; 480: 547-51.

Lesage S, Brice A. Parkinson's disease: from monogenic forms to genetic susceptibility factors[Review]. Hum Mol Genet 2009; 18: R48-59.

Li N, Bates DJ, An J, Terry DA, Wang E. Up-regulation of key microRNAs, and inverse down-regulation of their predicted oxidative phosphorylation target genes, during aging in mouse brain. Neurobiol Aging 2011; 32: 944-55.

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 2005; 433: 769-73.

Lin MK, Farrer MJ. Genetics and genomics of Parkinson's disease [Review]. Genome Med 2014; 6: 48.

Lwin A, Orvisky E, Goker-Alpan O, LaMarca ME, Sidransky E. Glucocerebrosidase mutations in subjects with parkinsonism. Mol Genet Metab 2004; 81: 70-3.

Margis R, Margis R, Rieder CR. Identification of blood microRNAs associated to Parkinson's disease. J Biotechnol. 2011; 152: 96-101.

Martínez-Arias R, Calafell F, Mateu E, Comas D, Andrés A, Bertranpetit J. Sequence variability of a human pseudogene. Genome Res 2001; 11: 1071-85.

Mazzulli JR, Xu YH, Sun Y, Knight AL, McLean PJ, Caldwell GA, Sidransky E, Grabowski GA, Krainc D. Gaucher disease glucocerebrosidase and α-synuclein form a bidirectional pathogenic loop in synucleinopathies. Cell 2011; 146: 37-52.

Mitrovich QM, Anderson P. mRNA surveillance of expressed pseudogenes in C. elegans. Curr Biol 2005; 15: 963-7.

Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, DeStefano AL, Kara E, Bras J, Sharma M, Schulte C, Keller MF, Arepalli S, Letson C, Edsall C, Stefansson H, Liu X, Pliner H, Lee JH, Cheng R; International Parkinson's Disease Genomics Consortium (IPDGC); Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI); 23andMe; GenePD; NeuroGenetics Research Consortium (NGRC); Hussman Institute of Human Genomics (HIHG); Ashkenazi Jewish Dataset Investigator; Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE); North American Brain Expression Consortium (NABEC); United Kingdom Brain Expression Consortium (UKBEC); Greek Parkinson's Disease Consortium; Alzheimer Genetic Analysis Group, Ikram MA, Ioannidis JP, Hadjigeorgiou GM, Bis JC, Martinez M, Perlmutter JS, Goate A, Marder K, Fiske B, Sutherland M, Xiromerisiou G, Myers RH, Clark LN, Stefansson K, Hardy JA, Heutink P, Chen H, Wood NW, Houlden H, Payami H, Brice A, Scott WK, Gasser T, Bertram L, Eriksson N, Foroud T, Singleton AB. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat Genet 2014; 46: 989-93.

Pankratz N, Beecham GW, DeStefano AL, Dawson TM, Doheny KF, Factor SA, Hamza TH, Hung AY, Hyman BT, Ivinson AJ, Krainc D, Latourelle JC, Clark LN, Marder K, Martin ER, Mayeux R, Ross OA, Scherzer CR, Simon DK, Tanner C, Vance JM, Wszolek ZK, Zabetian CP, Myers RH, Payami H, Scott WK, Foroud T; PD GWAS Consortium. Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. Ann Neurol 2012; 71: 370-84.

Paraboschi EM, Rimoldi V, Soldà G, Tabaglio T, Dall'Osso C, Saba E, Vigliano M, Salviati A, Leone M, Benedetti MD, Fornasari D, Saarela J, De Jager PL, Patsopoulos NA, D'Alfonso S, Gemmati D, Duga S, Asselta R. Functional variations modulating PRKCA expression and alternative splicing predispose to multiple sclerosis. Hum Mol Genet 2014; 23: 6746-61.

Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature 2010; 465: 1033-8.

Popp MW, Maquat LE. Organizing principles of mammalian nonsense-mediated mRNA decay [Review]. Annu Rev Genet 2013; 47: 139-65.

Popp MW, Maquat LE. Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine [Review]. Cell 2016; 165: 1319-22.

Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 2015; 43: D168-73.

Saugstad JA. MicroRNAs as effectors of brain function with roles in ischemia and injury, neuroprotection, and neurodegeneration [Review]. J Cereb Blood Flow Metab 2010; 30: 1564-76.

Schöndorf DC, Aureli M, McAllister FE, Hindley CJ, Mayer F, Schmid B, Sardi SP, Valsecchi M, Hoffmann S, Schwarz LK, Hedrich U, Berg D, Shihabuddin LS, Hu J, Pruszak J, Gygi SP, Sonnino S, Gasser T, Deleidi M. iPSC-derived neurons from GBA1-associated Parkinson's disease patients show autophagic defects and impaired calcium homeostasis. Nat Commun 2014; 5: 4028.

Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, Bar-Shira A, Berg D, Bras J, Brice A, Chen CM, Clark LN, Condroyer C, De Marco EV, Dürr A, Eblan MJ, Fahn S, Farrer MJ, Fung HC, Gan-Or Z, Gasser T, Gershoni-Baruch R, Giladi N, Griffith A, Gurevich T, Januario C, Kropp P, Lang AE, Lee-Chen GJ, Lesage S, Marder K, Mata IF, Mirelman A, Mitsui J, Mizuta I, Nicoletti G, Oliveira C, Ottman R, Orr-Urtreger A, Pereira LV, Quattrone A, Rogaeva E, Rolfs A, Rosenbaum H, Rozenberg R, Samii A, Samaddar T, Schulte C, Sharma M, Singleton A, Spitz M, Tan EK, Tayebi N, Toda T, Troiano AR, Tsuji S, Wittstock M, Wolfsberg TG, Wu YR, Zabetian CP, Zhao Y, Ziegler SG. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. N Engl J Med 2009; 361: 1651-61.

Sidransky E. Gaucher disease: insights from a rare Mendelian disorder [Review]. Discov Med 2012; 14: 273-81.

Sidransky E, Lopez G. The link between the GBA gene and parkinsonism [Review]. Lancet Neurol 2012; 11: 986-98.

Siebert M, Westbroek W, Chen YC, Moaven N, Li Y, Velayati A, Saraiva-Pereira ML, Martin SE, Sidransky E. Identification of miRNAs that modulate glucocerebrosidase activity in Gaucher disease cells. RNA Biol 2014; 11: 1291-300.

Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang D, Rutenberg-Schoenberg M, Clark W, Diekhans M, Rozowsky J, Hubbard T, Harrow J, Gerstein MB. Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci USA 2014; 111: 13361-6.

Soldà G, Robusto M, Primignani P, Castorina P, Benzoni E, Cesarani A, Ambrosetti U, Asselta R, Duga S. A novel mutation within the MIR96 gene causes non-syndromic inherited hearing loss in an Italian family by altering pre-miRNA processing. Hum Mol Genet 2012; 21: 577-85.

Spillantini MG. Parkinson's disease, dementia with Lewy bodies and multiple system atrophy are alpha-synucleinopathies. Parkinsonism Relat Disord 1999; 5: 157-62.

Svobodová E, Mrázová L, Lukšan O, Elstein D, Zimran A, Stolnaya L, Minks J, Eberová J, Dvořáková L, Jirsa M, Hřebíček M. Glucocerebrosidase gene has an alternative upstream promoter, which has features and expression characteristic of housekeeping genes. Blood Cells Mol Dis 2011; 46: 239-45.

Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 2007; 131: 861-72.

Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition [Review]. Nature 2014; 505: 344-52.

Thomson DW, Dinger ME. Endogenous microRNA sponges: evidence and controversy. Nat Rev Genet 2016; 17: 272-83.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 2002; 3: RESEARCH0034.

Wafaei JR, Choy FY. Glucocerebrosidase recombinant allele: molecular evolution of the glucocerebrosidase gene and pseudogene in primates. Blood Cells Mol Dis 2005; 35: 277-85.

Wang L, Guo ZY, Zhang R, Xin B, Chen R, Zhao J, Wang T, Wen WH, Jia LT, Yao LB, Yang AG. Pseudogene OCT4-pg4 functions as a natural micro RNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. Carcinogenesis 2013; 34: 1773-81.

Westbroek W, Gustafson AM, Sidransky E. Exploring the link between glucocerebrosidase mutations and parkinsonism [Review]. Trends Mol Med 2011; 17: 485-93.

Xu D, Takeshita F, Hino Y, Fukunaga S, Kudo Y, Tamaki A, Matsunaga J, Takahashi RU, Takata T, Shimamoto A, Ochiya T, Tahara H. miR-22 represses cancer progression by inducing cellular senescence. J Cell Biol 2011; 193: 409-24.

Yang CP, Zhang ZH, Zhang LH, Rui HC. Neuroprotective Role of MicroRNA-22 in a 6-Hydroxydopamine-Induced Cell Model of Parkinson's Disease via Regulation of Its Target Gene TRPM7. J Mol Neurosci 2016 Sep 8. [Epub ahead of print]

Yu H, Wu M, Zhao P, Huang Y, Wang W, Yin W. Neuroprotective effects of viral overexpression of microRNA-22 in rat and cell models of cerebral ischemia-reperfusion injury. J Cell Biochem 2015; 116: 233-41.

Zhang S, Zhang D, Yi C, Wang Y, Wang H, Wang J. MicroRNA-22 functions as a tumor suppressor by targeting SIRT1 in renal cell carcinoma. Oncol Rep. 2016; 35: 559-67.

Zheng Y, Xu Z. MicroRNA-22 induces endothelial progenitor cell senescence by targeting AKT3. Cell Physiol Biochem 2014; 34: 1547-55.

Zheng L, Li X, Gu Y, Lv X, Xi T. The 3'UTR of the pseudogene CYP4Z2P promotes tumor angiogenesis in breast cancer by acting as a ceRNA for CYP4Z1. Breast Cancer Res Treat 2015; 150: 105-18.

SUPPLEMENTARY MATERIAL

The *GBAP1* pseudogene acts as a ceRNA for the Parkinson-related gene *GBA* by sponging miR-22-3p

Letizia Straniero,[1,2] Valeria Rimoldi,[2,3] Maura Samarani,[1] Stefano Goldwurm,[4] Alessio Di Fonzo,[5] Rejko Krüger,[6] Michela Deleidi,[7] Massimo Aureli,[1] Giulia Soldà,[2,3,*] Stefano Duga,[2,3] Rosanna Asselta[2,3]

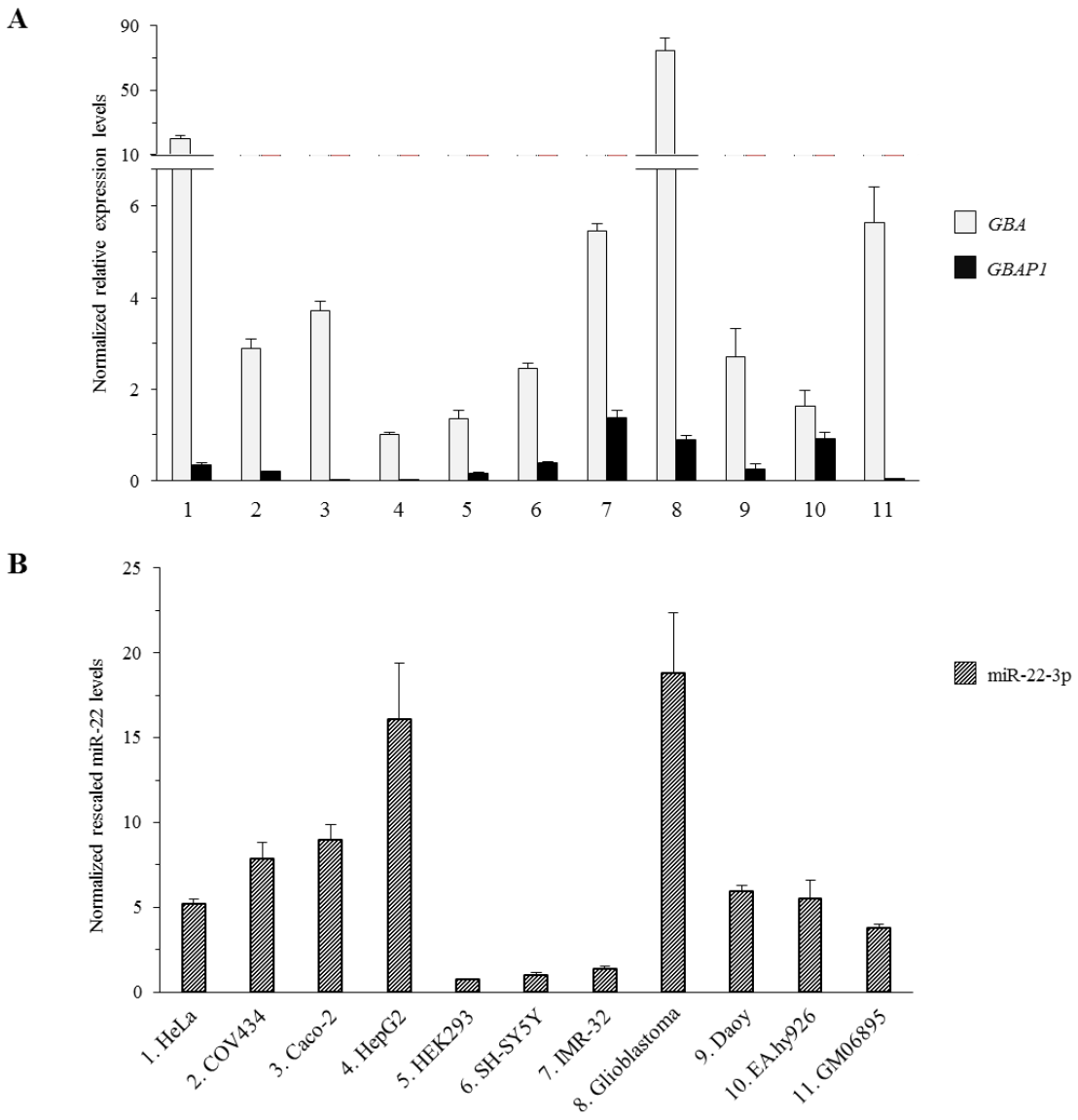
[1] Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano, Milano, Italia
[2] Department of Biomedical Sciences, Humanitas University, Via Manzoni 113, 20089 Rozzano, Milan, Italy
[3] Humanitas Clinical and Research Center, Via Manzoni 56, 20089 Rozzano, Milan, Italy
[4] Parkinson Institute, ASST "Gaetano Pini-CTO", Milan, Italy
[5] IRCCS Foundation Ca' Granda Ospedale Maggiore Policlinico, Dino Ferrari Center, Neuroscience Section, Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
[6] Clinical and Experimental Neuroscience, Luxembourg Center for Systems Biomedicine (LCSB), University of Luxembourg and Centre Hospitalier de Luxembourg (CHL), Luxembourg
[7] German Centre for Neurodegenerative Diseases (DZNE) Tübingen within the Helmholtz Association, Tübingen, Germany; Hertie Institute for Clinical Brain Research, University of Tübingen, Tübingen, Germany

List of contents:

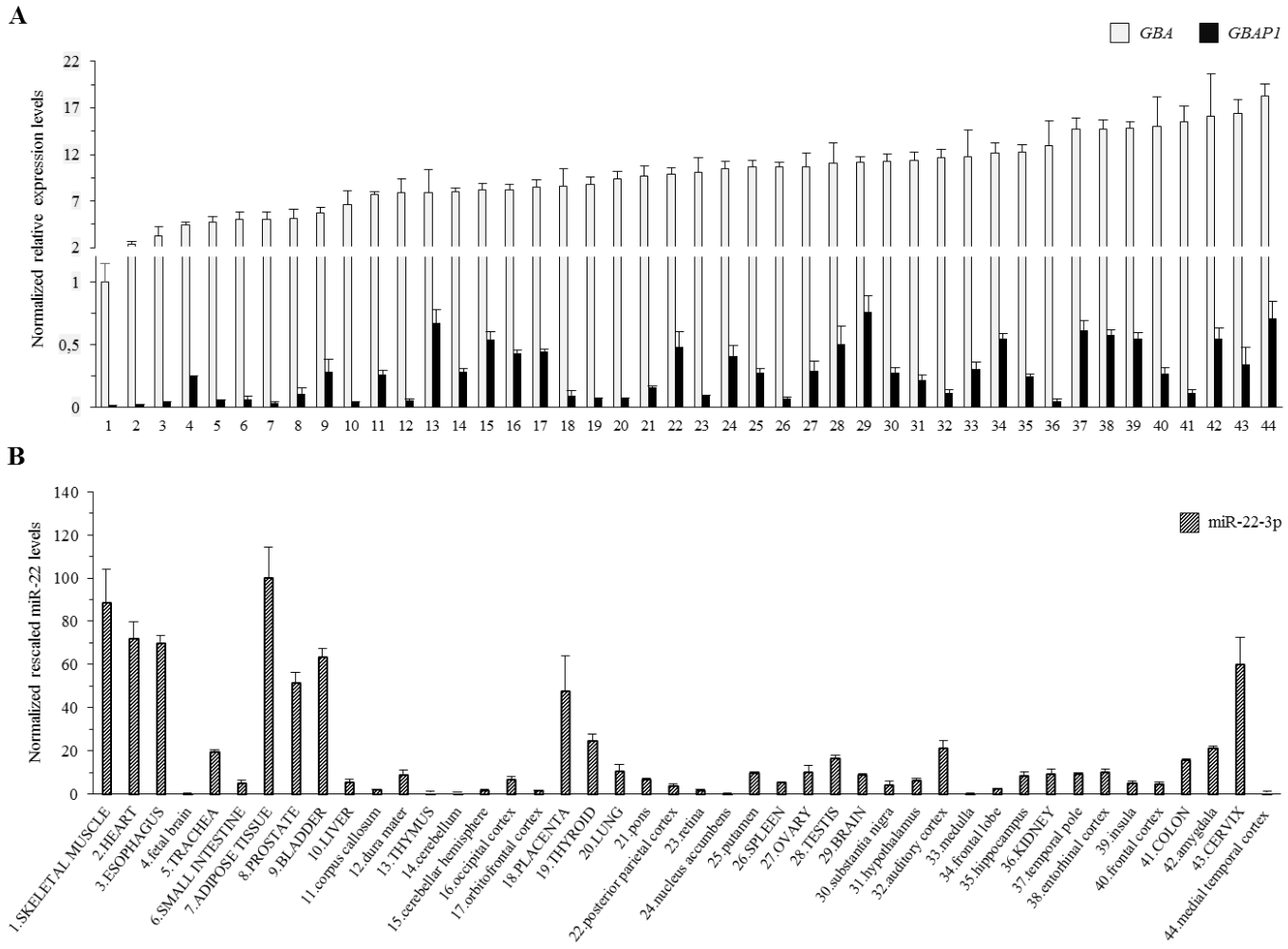
SUPPLEMENTARY MATERIALS AND METHODS

Microarray dataset retrieving and analysis

Microarray datasets were retrieved from the National Center for Biotechnology Information (NCBI) GEO database (http://www.ncbi.nlm.nih.gov/gds/). The database was searched using "Parkinson's disease", "*substantia nigra*", and "human" as keywords (accession date: June 14th, 2016). The manual review of the obtained entries allowed the identification of three datasets (GSE7621, GSE8397, and GSE20292), all characterized by expression data obtained from RNAs extracted from the *substantia nigra* of Parkinson's disease cases and healthy controls. Characteristics of selected datasets are listed in Supplementary Table 3. The three datasets were analyzed separately by comparing Parkinson's disease cases *vs* healthy controls, using the GEO2R web application (Barrett *et al.*, 2009; Barrett *et al.*, 2013). GEO2R, available at GEO repository, implements Bioconductor R packages (Gentleman *et al.*, 2004; R Core Team, 2013) and provides results as a list of genes ordered by significance (P value). We specifically searched this list for probe sets corresponding to *GBA* and *GBAP1*. The three identified probes (210589_s_at, 209093_s_at, and 216400_at; shared by the three datasets) were then individually blatted against the human genome, using the UCSC Genome Browser (hg19) to verify if they target correctly the corresponding gene. The 216400_at probe, which resulted to recognize intronic sequences, was excluded from further analysis. We hence performed a meta-analysis using the expression data associated with the two surviving probe sets and the Integrative Meta-analysis of Expression Data (INMEX) program (Xia *et al.*, 2013). Data were uploaded to INMEX, processed, annotated, checked, and meta-analyzed by using the "Combining P-values" option, based on the Fisher's method [-2*∑Log(p)] (Xia *et al.*, 2013). The threshold for significance was set to 0.05.



Supplementary figure 1: Expression profiles of *GBA*, *GBAP1*, and miR-22-3p in cell lines.
A) Expression levels of *GBA* and *GBAP1* were measured by semi-quantitative real-time RT-PCRs in a panel of 11 cell lines. *HMBS* and *ACTB* were used as housekeeping genes.
B) MiR-22-3p expression levels were determined in the same panel of cell lines using a poly(A) tailing and a universal reverse transcription approach, followed by real-time RT-PCRs. U6 was used as housekeeping gene.In all cases, expression levels are shown as normalized rescaled values, setting as 1 the value measured in HepG2 (pane A) and SH-SY5Y (panel B). Bars represent means +SD of three replicates.
Cell lines: HeLa (epithelial cervical carcinoma); COV434 (granulosa carcinoma); Caco-2 (epithelial colorectal adenocarcinoma); HepG2 (liver carcinoma); HEK293 (embryonic kidney); SH-SY5Y (neuroblastoma); IMR-32 (neuroblastoma); Glioblastoma; Daoy (medulloblastoma); EAhy296 (endothelium);- GM06895 (lymphoblastoid line).



Supplementary figure 2: Expression profiles of *GBA*, *GBAP1*, and miR-22-3p in 20 human tissues and 24 human cerebral districts.

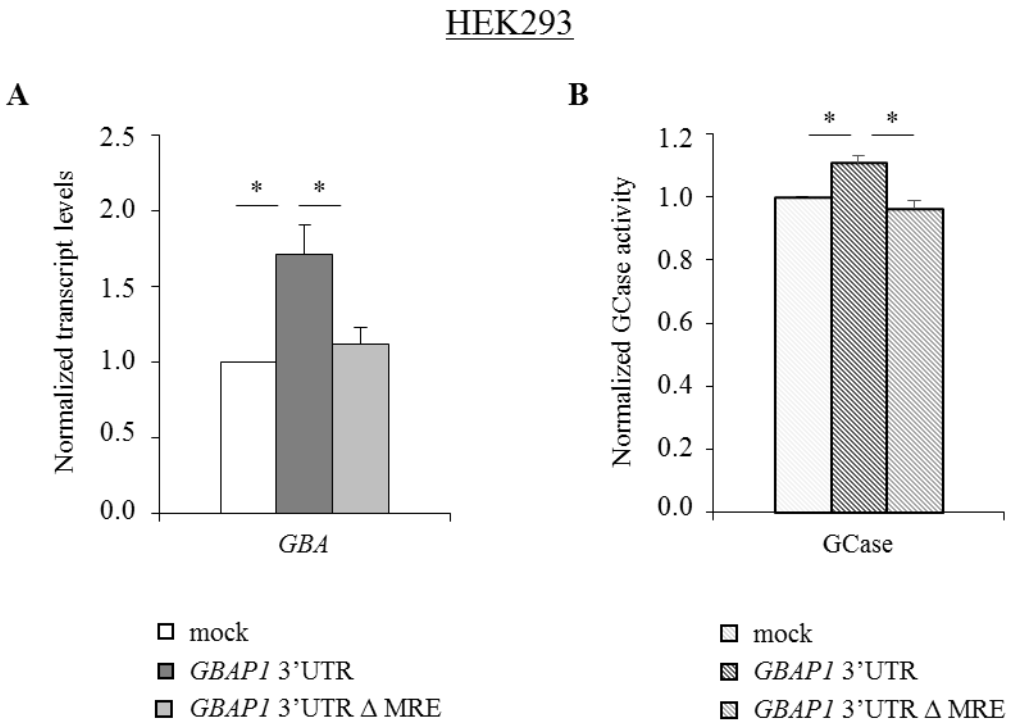
Expression levels of *GBA*, *GBAP1*, (**panel A**) and miR-22-3p (**panel B**) were measured by real-time RT-PCR assays on cDNAs derived from commercial panels of 20 human tissues (listed in upper-case letters) and 24 different RNAs from cerebral districts (in lower-case letters; for more details, see legend of Supplementary figure 1).

Each tissue sample consisted in a pool comprising RNA from at least 3 donors.

Concerning specifications for brain district samples:

- dura mater, parietal cortex posterior, auditory cortex, putamen, frontal cortex, entorhinal cortex, amygdala, hippocampus, orbital frontal cortex, temporal cortex medial, hypothalamus, and pons RNA samples all derive from a single person;
- fetal brain: this sample derives from normal brains of 59 spontaneously aborted male and female Caucasian fetuses (20-33 weeks);
- frontal lobe: pooled from 4 male and female Caucasian individuals (32-61 years old);
- nucleus accumbens: pooled from 6 male and female Caucasian individuals (23-56 years old);
- retina: pooled from 25 male and female Caucasian individuals (24-65 years old);
- insula: pooled from 15 male and female Caucasian individuals (20-68 years old);
- substantia nigra: pooled from 21 male and female Caucasian individuals (20-62 years old);
- medulla oblongata: pooled from 29 male and female Caucasian individuals (18-64 years old).

Results are presented as normalized rescaled values, setting as 1 the values corresponding to skeletal muscle (panel A), and thymus (panel B). Bars represent means +SD of three replicates.

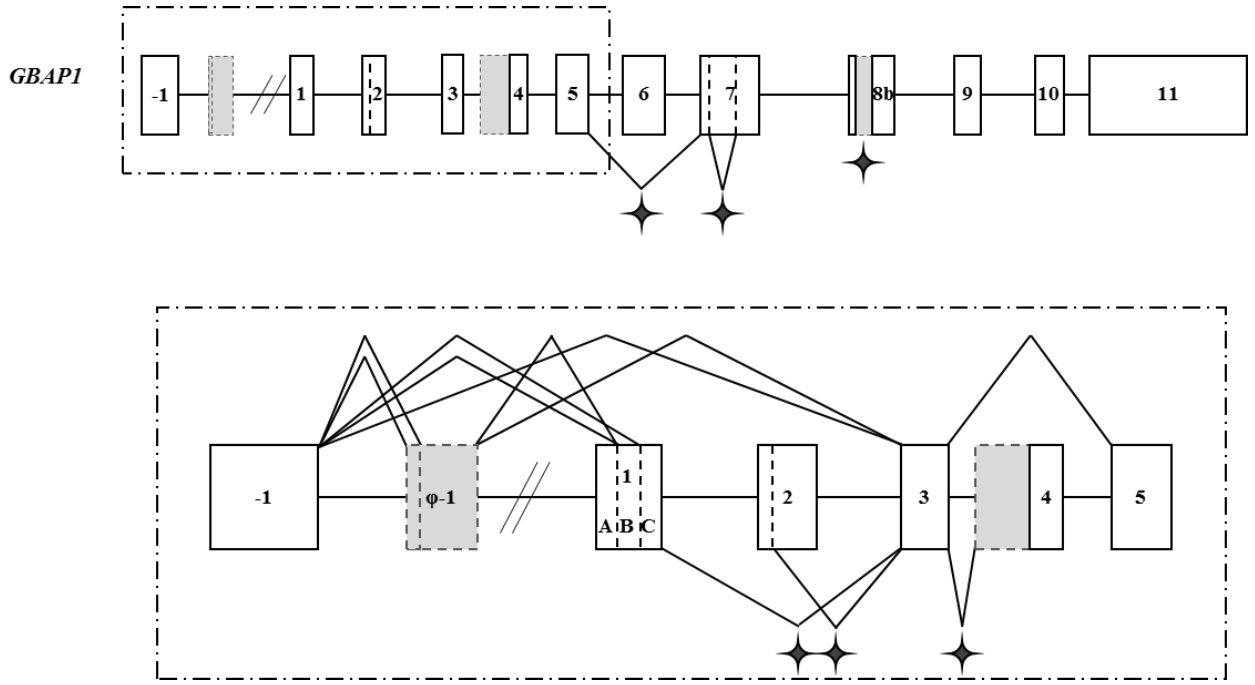


Supplementary figure 3: *GBAP1* acts as a ceRNA titrating miR-22-3p and up-regulating *GBA* in HEK293 cells.

A) The psiCHECK2 vectors, containing the 3'UTR region of *GBAP1* (with or without the miR-22-3p recognition element, ΔMRE) downstream of the luciferase reporter gene, were independently transfected in HEK293 cells together with the plasmid expressing the pre-miR-22-3p transcript. 24 hours after transfections, cells were collected for extracting total RNA for measurements of *GBA* endogenous levels by semi-quantitative real-time RT-PCRs. The value measured in cells transfected with an empty vector (psiCHECK2, mock) was set as 1.

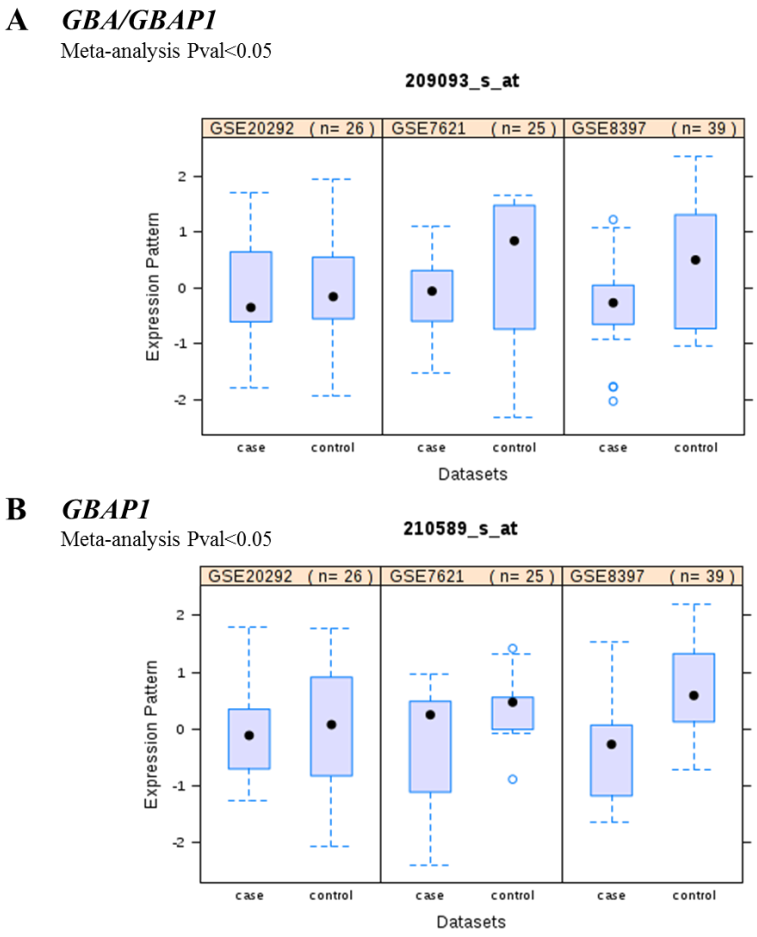
B) Over-expression experiments of psiCHECK2 vectors, containing the two different 3'UTR regions of *GBAP1*, upon miR-22-3p over-expression were repeated for preparing protein lysates for endogenous GCase activity measurements. In this case, the GCase activity was measured 48 hours after transfections.

In all panels, bars represent means +SEM of three independent experiments, each performed at least in triplicate. Significance levels of t-tests are shown. *: $P<0.05$.



Supplementary figure 4: *GBAP1* is characterized by multiple in-frame and out-of-frame splicing isoforms.

Alternative splicing of the *GBAP1* pre-mRNA was analyzed in HepG2 cells, treated with the NMD inhibitor cycloheximide, by RT-PCR using primer couples specific for the pseudogene. In the upper part of the figure, a schematic representation of the whole gene is presented, with white boxes representing annotated exons, lines indicating introns, and broken lines pointing to the major splicing events characterizing the 3' portion of the transcript. Grey boxes indicate the presence of additional exons or extensions of annotated exons. In the lower part of the figure, a magnification of the 5' portion of the gene is reported, with all the identified alternative splicing events. Out-of-frame splicings leading to the introduction of a PTC are indicated in all cases by a star. The precise mapping of all identified splicing events was confirmed by Sanger sequencing of the relevant RT-PCR fragment.



Supplementary figure 5: *GBA* and *GBAP1* appear dowregulated in the *substantia nigra* of Parkinson's disease patients.

The figure shows box-plots comparing the relative gene expression of *GBA*/*GBAP1* (probe set 209093_s_at, hybridizing against both genes; **panel A**) and *GBAP1* (probe set 210589_s_at, specific for *GBAP1*; **panel B**) in the *substantia nigra* of Parkinson's disease patients (case) and healthy controls (control). Both panels represent screen shots obtained from the INMEX program (Xia *et al.*, 2013), which rescales all expression levels in order to make data immediately comparable at visual inspection. For each box plot, the relevant dataset is indicated, together with the number of analyzed individuals. In the case of dataset GSE20292, three Parkinson's disease patients were excluded from the analysis, since they resulted outliers (abnormal low levels of both *GBA* and *GBAP1*). The meta-analysis P values for the two analyzed probe sets is indicated.

Supplementary table 1: Primers used for cloning, mutagenesis, as well as expression profiling experiments.

Primer	*Sequence (5'-3')* *	*Localization* **	*Application*
*GBA*_ex9_F *GBA*_ex10_R	ATTGGGTGCGTAACTTTGTC TCCAGGTCGTTCTTCTGACT	exon 9, chr1:155,205,549-155,205,568 exon 10, chr1:155,205,040-155,205,059	real-time RT-PCR assay to evaluate *GBA* expression
*GBAP1*_F *GBAP1*_R	GGACCGACTGGAACCCAT TCCAGGTCGTTCTTCTGACTG	exon 9, chr1:155,184,911-155,184,928 exon 10, chr1:155,184,413-155,184,433	real-time RT-PCR assay to evaluate *GBAP1* expression
*SP1*_F *SP1*_R	AGACAGTGAAGGAAGGGGCT GCGTTTCCCACAGTATGACC	exon 4/5, chr12:53,800,523-53,803,150 exon 5, chr12:53,803,280-53,803,299	real-time RT-PCR assay to evaluate *SP1* expression
*SIRT1*_F *SIRT1*_R	TGTTATTGGGTCTTCCCTCAA AAATGCAGATGAGGCAAAGG	exon 7, chr10:69,669,153-69,669,173 exon 8, chr10:69,672,275-69,672,294	real-time RT-PCR assay to evaluate *SIRT1* expression
*CELF1*_F *CELF1*_R	GAAGCCAGAAGGAAGGTCCA TCCCAAAGGGCATAAACATC	exon 10/11, chr11:47,494,779-47,496,902 exon 11, chr11:47,494,697-47,494,716	real-time RT-PCR assay to evaluate *CELF1* expression
*HMBS*_F *HMBS*_R	GTTCAGGAGTATTCGGGGAAACC TTCCTCAGGGTGCAGGATCTG	exon 8/9, chr11:118,960,963-118,962,132 exon 10, chr11:118,962,832-118,962,852	reference gene for real-time RT-PCR assays
CX32_F CX32_R	GCAGCAGCAGCCAGGTGTGG ATACTCGGCCAATGGCAGTA	exon 1, chrX:70,435,108-70,435,127 exon 2, chrX:70,443,608-70,443,627	reference gene for NMD experiments
CX43_F CX43_R	AAAGTACCAAACAGCAGCGG CTCCAGCAGTTGAGTAGGCT	exon 1, chr6:121,756,846-121,756,865 exon 2, chr6:121,768,038-121,768,057	reference gene for NMD experiments
*PRKCA*_ex3*_F *PRKCA*_4/5_R	TCCCCTGTATTGCTAGTCTGC CGCAGGTGTCACATTTCATC	exon 3*, chr17:64,550,643-64,550,663 exon 4/5, chr17:64,637,571- 64,641,506	positive control for NMD experiments
*PRKCA*_ex3/4F3 *PRKCA*_4/5_R	GGACCCGACACTGATGACC CGCAGGTGTCACATTTCATC	exon3/4 chr17:64,492,387- 64,637,476 exon 4/5, chr17:64,637,571- 64,641,506	negative control for NMD experiments
miR-22-3p_F1 Uni_R1	GCTGCCAGTTGAAGAACT CTCAGTCGCATAGCTTGAT	exon 3, chr17:1,617,210-1,617,227	real-time RT-PCR assay to evaluate miR-22-3p expression
miR-132-3p_F Uni_R1	AGTCTACAGCCATGGTCG CTCAGTCGCATAGCTTGAT	chr17:1,953,223-1,953,240	real-time RT-PCR assay to evaluate miR-132-3p expression
miR-22_*Kpn*I_F miR-22_*Xho*I_R	AGAGGTACCTTTCCCTTAGGAGCCTGT GGCCTCGAGCAGCCCATTTCTGTCACCTT	chr17:1,617,301-1,617,320 chr17:1,617,133-1,617,152	pre-miR-22 molecular cloning

miR-132_*Kpn*I_F miR-132_*Xho*I_R	AGAGGTACCCAGTCCCCGTCCCTCAG GGCCTCGAGCACGTGGGATCTTGACTCG	chr17:1,953,507-1,953,523 chr17:1,953,123-1,953,141	pre-miR-132 molecular cloning
UTR_*GBA-GBAP1_Sgf*I_F UTR_*GBA-GBAP1_Not*I_R	ACCGGCGATCGCTCACCTGGCTACTCCATTCA AAGGAAAAAGCGGCCGCCCACCCAGAATAAAGC CACT	exon 11, chr1:155,204,811-155,204,830 chr1:155,204,214-155,204,233	*GBA* and *GBAP1* 3'UTR molecular cloning
psiUx_F psiUx_R	GATCTTCCCCATCGGTGAT CGCTGATCGGAAGTGAGAAT	psiUx psiUx	recombinant bacterial colony screening
psiCHECK_2_F psiCHECK_2_R	AGGACGCTCCAGATGAAATG AGGACGCTCCAGATGAAATG	psiCHECK-2 psiCHECK-2	recombinant bacterial colony screening
3'UTR_*GBAP1*_ΔMRE_F 3'UTR_*GBAP1*_ΔMRE_R	TCCTATGGCACCAGCCAGGAAAAATCTTAAAGGA GAAAATGTTTGAGCCC GGGCTCAAACATTTTCTCCTTTAAGATTTTTCCTG GCTGGTGCCATAGGA		miR-22 MRE site-directed mutagenesis (deletion)
*GBAP1*_EX-1_F	GGGCTGCTTCTTGACTTCC	exon -1, chr1:155,197,281-155,197,299	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX-1/-φ1_F	TTCTCTTCGCCGACGGTT	exon -1/intron 1, chr1:155,197,168-155,194,848	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_-28-φ1_F	TGACAGGGCTTTCCCTATGT	intron 1, chr1:155,194,858-155,194,877	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_-φ1_F	CTGGGCTCAAAAGATCCTCA	intron 1, chr1:155,194,816-155,194,835	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX-1/1_F	CTCTTCGCCGACGTGGA	exon -1/1, chr1:155,197,168-155,188,730	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX-1/1B_F	CTCTTCGCCGACGAGACT	exon -1/1, chr1:155,197,168-155,188,711	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX-1/1C_F	TCTTCGCCGACGTGACC	exon -1/1, chr1:155,197,168-155,188,674	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX-1/3_F	CTCTTCGCCGACGGTGC	exon -1/1, chr1:155,197,168-155,187,837	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX2A_F	TCCCAAGCCTTCGGGTAG	exon 2, chr1:155,188,248-155,188,265	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX2_F	TTCGGGTAGGGTAAGCATCA	exon 2, chr1:155,188,237-155,188,256	RT-PCR assay to characterize *GBAP1* splicing pattern

*GBAP1*_EX2_R	CCACGACACTGCCTGAAGTA	exon 2, chr1:155,188,190-155,188,209	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX3_R	CCGGTGCAATTAGCCTGTAT	exon 3, chr1:155,187,760-155,187,779	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX3/5_F	TGCACCGGCACAGGAAT	exon 3/5, chr1:155,187,161-155,187,767	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX3/5_R	GATGTTATATCCGATTCCTGTGC	exon 3/5, chr1:155,187,148-155,187,766	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX4_F	TAAGATTTCGCCGCCTATCA	exon 4, chr1:155,187,360-155,187,379	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX4_R	GAGCCTGAGTCCGTAGCAGT	exon 4, chr1:155,187,333-155,187,352	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX5_R	ATAGGTGTAGGTGCGGATGG	exon 5, chr1:155,187,097-155,187,116	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX7_F	TGAGTGGATACCCCTTCCAG	exon 7, chr1:155,186,316-155,186,335	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_IVS8A_R	GGGAACAGGTGGTGTGTCTC	exon 8, chr1:155,185,466-155,185,485	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX8A/8B_R	ACCCACACAGGCCTTTAGC	exon 8A/8B, chr1:155,185,429-155,185,557	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_F	GGACCGACTGGAACCCAT	exon 9, chr1:155,184,911-155,184,928	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX9_R	ATGTCTACAATGATGGGTTCCAG	exon 9, chr1:155,184,899-155,184,921	RT-PCR assay to characterize *GBAP1* splicing pattern
*GBAP1*_EX11_R	TGAATGGAGTAGCCAGGTGA	exon 11, chr1:155,184,184-155,184,203	RT-PCR assay to characterize *GBAP1* splicing pattern

* Underlined sequences corresponds to nucleotides added at the primer end to introduce a site for restriction digestion in cloning experiments

** According to UCSC genome browser (http://genome-euro.ucsc.edu/index.html) on Human Feb. 2009 (GRCh37/hg19) Assembly.

Supplementary table 2: MiRNAs predicted as potential targets of *GBA* and *GBAP1*.

miRNA (chromosomal position)*	***miRNA:GBA pairing*** **	***Brain expression**** **	***Literature data on neurodegenerative diseases***
miR-22-3p (chr17:1,617,197-1,617,281)	3' ugucaagaagUUGACCGUCGAa 5' miR-22 || |||||||| 5' cagccaggaaAAAUGGCAGCUc 3' *GBA*	Cerebellum (643) frontal cortex (219)	AD: Cheng *et al.*, 2013 ALS: Parisi *et al.*, 2013 HD: Jovicic *et al.*, 2013; Lee *et al.*, 2011 PD: Margis *et al.*, 2011
miR-132 (chr17:1,953,202-1,953,302)	3' gcuGGUACCGACAUCUGACAAu 5' miR-132 ||| ||| |||||||| 5' ggcCCAAAACUGGAGACUGUUu 3' *GBA*	Cerebellum (203) frontal cortex (818)	AD: Smith *et al.*, 2015 ALS: Freischmidt *et al.*, 2013 HD: Lee *et al.*, 2011 PD: Alieva *et al.*, 2015
miR-212 (chr17:1,953,565-1,953,674)	3' ccGGCACUGACCUCUGACAAu 5' miR-212 || |||||||||||| 5' gcCCAAAACUGGAGACUGUUu 3' *GBA*	Cerebellum (22) frontal cortex (70)	AD: Smith *et al.*, 2015

Predictions were performed using microRNA.org, MicroCosm Targets, PITA, as well as the miRWalk2 suite (among the programs implemented in miRWalk2, we selected miRWalk2.0, RNA22, miRanda, miRDB, TargetScan, and PICTAR2). In all cases, default parameters were used.).
* According to UCSC genome browser (http://genome-euro.ucsc.edu/index.html) on Human Feb. 2009 (GRCh37/hg19) Assembly.
** The RNAhybrid software (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/) was used to visualize the miRNA:mRNA interactions.
*** According to deep-sequencing data available through miRBase (http://www.mirbase.org/); numbers in brackets refers to read counts per million of RNA-seq experiments (annotation confidence: high for miR-22 and miR-132; not reported for miR-212).
ALS: amyotrophic lateral sclerosis; AD: Alzheimer disease; HD: Huntington's disease; PD: Parkinson's disease.

Supplementary table 3: Characteristics of Parkinson's disease-related datasets included in the study.

Dataset (accession number)	*Array type*	*PD cases / controls*	*Origin*	*Age at death (years ± SD)*	*Notes on PD patients*	*Tissue*	*Reference*
GSE7621	Affymetrix Human Genome U133 Plus 2.0 Array	16 / 9	Caucasian	Mean age cases: 77.9 ± 13.1 Mean age controls: n.a.	All PD subjects had advanced disease with a mean H&Y stage of 4.5 ± 0.7.	SN tissue from post-mortem brains	Lesnick *et al.*, 2007 Papapetropoulos *et al.*, 2006
GSE8397	Affymetrix Human Genome U133A Array	24 / 15	n.a.	Mean age cases: 80 ± 5.7 Mean age controls: 70.6 ± 12.5	PD patients showed a mean disease duration of 13.4 ± 8.3	SN tissue from post-mortem brains (NS split into medial and lateral portions)	Moran *et al.*, 2006
GSE20292	Affymetrix Human Genome U133A Array	11 / 18	n.a.	Mean age cases: 76.7 ± 6.2 Mean age controls: 71.2 ± 11.1	-	SN tissue from post-mortem brains	Zhang *et al.*, 2005

Abbreviations: H&Y, Hoehn and Yahr scale; n.a., not available; PD, Parkinson's disease; SD, standard deviation; SN, *substantia nigra*.

SUPPLEMENTARY REFERENCES

Alieva AKh, Filatova EV, Karabanov AV, Illarioshkin SN, Limborska SA, Shadrina MI, Slominsky PA. miRNA expression is highly sensitive to a drug therapy in Parkinson's disease. Parkinsonism Relat Disord 2015; 21: 72–4.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 2009; 37(Database issue): D885–90.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 2013; 41(Database issue): D991–5.

Cheng XR, Cui XL, Zheng Y, Zhang GR, Li P, Huang H, Zhao YY, Bo XC, Wang SQ, Zhou WX, Zhang YX. Nodes and biological processes identified on the basis of network analysis in the brain of the senescence accelerated mice as an Alzheimer's disease animal model. Front Aging Neurosci 2013; 5: 65.

Freischmidt A, Müller K, Ludolph AC, Weishaupt JH. Systemic dysregulation of TDP-43 binding microRNAs in amyotrophic lateral sclerosis. Acta Neuropathol Commun 2013; 1: 42.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004; 5: R80.

Jovicic A, Zaldivar Jolissaint JF, Moser R, Silva Santos Mde F, Luthi-Carter R. MicroRNA-22 (miR-22) overexpression is neuroprotective via general anti-apoptotic effects and may also target specific Huntington's disease-related mechanisms. PLoS One 2013; 8: e54222.

Lee ST, Chu K, Im WS, Yoon HJ, Im JY, Park JE, Park KH, Jung KH, Lee SK, Kim M, Roh JK. Altered microRNA regulation in Huntington's disease models. Exp Neurol 2011; 227: 172–9.

Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, Henley JR, Rocca WA, Ahlskog JE, Maraganore DM. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet. 2007; 3: e98.

Margis R, Margis R, Rieder CR. Identification of blood microRNAs associated to Parkinson’s disease. J Biotechnol 2011; 152: 96–101.

Moran LB, Duke DC, Deprez M, Dexter DT et al. Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics 2006; 7: 1–11.

Papapetropoulos S, Ffrench-Mullen J, McCorquodale D, Qin Y, Pablo J, Mash DC. Multiregional gene expression profiling identifies MRPS6 as a possible candidate gene for Parkinson's disease. Gene Expr. 2006; 13: 205–15.

Parisi C, Arisi I, D'Ambrosi N, Storti AE, Brandi R, D'Onofrio M, Volonté C. Dysregulated microRNAs in amyotrophic lateral sclerosis microglia modulate genes linked to neuroinflammation. Cell Death Dis 2013;4:e959.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. URL http://www.R-project.org/.

Smith PY, Hernandez-Rapp J, Jolivette F, Lecours C, Bisht K, Goupil C, Dorval V, Parsi S, Morin F, Planel E, Bennett DA, Fernandez-Gomez FJ, Sergeant N, Buée L, Tremblay MÈ, Calon F, Hébert SS. miR-132/212 deficiency impairs tau metabolism and promotes pathological aggregation in vivo. Hum Mol Genet 2015; 24: 6721–35.

Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock RE. INMEX-a web-based tool for integrative meta-analysis of expression data. Nucleic Acids Res 2013; 41(Web Server issue): W63–70.

Zhang Y, James M, Middleton FA, Davis RL. Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am J Med Genet B Neuropsychiatr Genet 2005; 137B: 5–16.

Part III

During my PhD, I was also involved in side research projects mainly focused on the molecular dissection of mutations responsible for Mendelian disorders (i.e. cystic fibrosis and oculocutaneous albinism) and in particular, on the characterization of their effects on RNA metabolism.

Content

Straniero L, Rimoldi V, Soldà G, Mauri L, Manfredini E, Andreucci E, Bargiacchi S, Penco S, Gesu GP, Del Longo A, Piozzi E, Asselta R, Primignani P. Two novel splicing mutations in the SLC45A2 gene cause Oculocutaneous Albinism Type IV by unmasking cryptic splice sites. J Hum Genet. 2015 Sep;60(9):467-71.

Straniero L, Soldà G, Costantino L, Seia M, Melotti P, Colombo C, Asselta R, Duga S. Whole-gene CFTR sequencing combined with digital RT-PCR improves genetic diagnosis of cystic fibrosis. J Hum Genet. 2016 Aug 4.

Journal of Human Genetics (2015), 1–5

www.nature.com/jhg

ORIGINAL ARTICLE

Two novel splicing mutations in the *SLC45A2* gene cause Oculocutaneous Albinism Type IV by unmasking cryptic splice sites

Letizia Straniero[1], Valeria Rimoldi[2,3], Giulia Soldà[2,3], Lucia Mauri[4], Emanuela Manfredini[4], Elena Andreucci[5], Sara Bargiacchi[6], Silvana Penco[4], Giovanni P Gesu[4], Alessandra Del Longo[7], Elena Piozzi[7], Rosanna Asselta[2,3] and Paola Primignani[4]

Oculocutaneous albinism (OCA) is characterized by hypopigmentation of the skin, hair and eye, and by ophthalmologic abnormalities caused by a deficiency in melanin biosynthesis. OCA type IV (OCA4) is one of the four commonly recognized forms of albinism, and is determined by mutation in the *SLC45A2* gene. Here, we investigated the genetic basis of OCA4 in an Italian child. The mutational screening of the *SLC45A2* gene identified two novel potentially pathogenic splicing mutations: a synonymous transition (c.888G>A) involving the last nucleotide of exon 3 and a single-nucleotide insertion (c.1156+2dupT) within the consensus sequence of the donor splice site of intron 5. As computer-assisted analysis for mutant splice-site prediction was not conclusive, we investigated the effects on pre-mRNA splicing of these two variants by using an *in vitro* minigene approach. Production of mutant transcripts in HeLa cells demonstrated that both mutations cause the almost complete abolishment of the physiologic donor splice site, with the concomitant unmasking of cryptic donor splice sites. To our knowledge, this work represents the first in-depth molecular characterization of splicing defects in a OCA4 patient.
Journal of Human Genetics advance online publication, 28 May 2015; doi:10.1038/jhg.2015.56

INTRODUCTION

Albinism is a heterogeneous group of inherited genetic diseases, affecting all ethnic backgrounds with an overall prevalence of approximately 1/17 000 people.[1] Oculocutaneous albinism (OCA) is characterized by a general hypopigmentation of the skin, hair and eye, and by ophthalmologic abnormalities caused by a deficiency in melanin biosynthesis. Eye and optic system abnormalities include various degrees of congenital nystagmus, hypopigmentation of iris leading to iris translucency, reduced pigmentation of the retinal pigment epithelium, foveal hypoplasia, reduced visual acuity and refractive errors. Photophobia may be prominent. A characteristic feature is the misrouting of the optic nerves, consisting in an excessive crossing of the fibres in the optic chiasma, which can result in strabismus and reduced stereoscopic vision.[2] These clinical signs are common to all types of albinism and are probably related to melanin reduction during embryonic development and early postnatal life.[3]

OCA is an autosomal recessive inherited condition; the four principal OCA forms (OCA1, OCA2, OCA3 and OCA4) are diagnosed based on the presence of mutations in the *TYR*, *OCA2*, *TYRP1* and *SLC45A2* genes, respectively.[4–8] Recently, three additional *loci* involved in albinism (OCA5–7) have been identified. In particular, a novel OCA *locus* on chromosome 4q24 was found by linkage analysis in a large consanguineous Pakistani pedigree, but the causal gene (termed OCA5) has not been identified yet.[9] Conversely, two genes involved in melanosome maturation and melanocyte differentiation, namely *SLC24A5* and *C10orf11*,[10–12] were identified as responsible for OCA6 and OCA7.[13] However, few data about these recently described OCA forms are currently available. In addition, an X-linked form of ocular albinism (OA1), in which the phenotype is mainly restricted to eyes and optic system, is associated with mutations in the *GPR143* gene.[14]

OCA4 (OMIM *606202, #606574) is worldwide distributed with a prevalence of 1:100 000 in most populations,[7,15–19] but in Japan, it is the most common form, accounting for about 24% of OCA cases.[20] The causative gene is *SLC45A2* (solute carrier family 45, member 2; also known as *MATP* gene), located on chromosome 5p13.3 and coding for a membrane-associated transporter protein.[7] MATP is thought to be important for the proper processing and trafficking of melanosomal proteins Tyr, Tyrp1, Dct (Tyrp2), and seems to act as a proton-dependent transporter at the melanosomal membrane, showing a possible relationship with the P protein (OCA2).[21,22] Moreover,

[1]Dipartimento di Biotecnologie Mediche e Medicina Traslazionale, Università degli Studi di Milano, Milano, Italia; [2]Department of Biomedical Sciences, Humanitas University, Rozzano (Mi), Italy; [3]Humanitas Clinical and Research Center, Rozzano (Mi), Italy; [4]Medical Genetics Unit—Department of Laboratory Medicine, Niguarda Ca' Granda Hospital, Milan, Italy; [5]Medical Genetics Unit, Meyer Children's University Hospital, Florence, Italy; [6]Medical Genetics Unit, Department of Clinical and Experimental Biomedical Sciences 'Mario Serio', University of Florence, Florence, Italy and [7]Pediatric Ophthalmology Department, Niguarda Ca' Granda Hospital, Milan, Italy
Correspondence: Dr P Primignani, Medical Genetics Unit, Department of Laboratory Medicine, A.O. Niguarda Ca' Granda Hospital, Piazza Ospedale Maggiore 3, Milan 20162, Italy;
E-mail: paola.primignani@ospedaleniguarda.it
Received 5 March 2015; revised 30 March 2015; accepted 26 April 2015

recent data based on experiments of heterologous expression of MATP in *Saccharomyces cerevisiae* suggested that it also acts as a sucrose transporter.[23]

OCA4 clinical phenotype is characterized by a high degree of heterogeneity, ranging from profound hypopigmentation (similar to OCA1A) to near normal skin colour. Indeed, some individuals accumulate pigmentation with age, whereas others remain completely hypopigmented throughout their life.[13,20,24] To date, more than 80 mutations have been identified in the *SLC45A2* gene.[25]

In this study, we report the identification of two hitherto-unknown splice defects underlying OCA4 in one Italian proband. The functional consequences of both mutations were elucidated by *in vitro* expression experiments.

PATIENTS AND METHODS

Patient

The patient was initially recruited by the Genetic Unit and Ophthalmology Center of A.O.U. Meyer of Florence, Italy. Subsequently, the patient with his parents, came to Niguarda Ca' Granda Hospital to undergo a multidisciplinary diagnostic workup including ophthalmological, dermatological, ENT (ear, nose and throat) evaluation and genetic counselling.

This study was conducted according to the Declaration of Helsinki and to the Italian legislation on sensible data recording. A signed informed consent for the genetic analysis was obtained from the proband's parents.

Ophthalmologic evaluation

The patient underwent full ophthalmological and instrumental evaluations, including motor and sensory status, visual acuity, cycloplegic refraction, iris transillumination with slitlamp biomicroscope and macular translucency with indirect ophthalmoscope. The visual acuity assessment for distance and near was evaluated in monocular and binocular vision with appropriate tests (Pigassou figures, E test and Snellen charts).

The following instrumental examinations were performed: optical coherence tomography, visual evoked potential and electroretinogram. Spectral Domain OCT (Heildelberg, Germany) was used to evaluate the macular morphological characteristics, with particular reference to the presence/absence of foveal depression. Electrophysiological tests (Retimax, CSO, Scandicci, Italy) were performed to evaluate the optic nerve fiber decussation.

Genetic analysis

DNA was extracted from peripheral blood according to standard protocols.

The entire coding region, as well as the adjacent intronic sequences of *TYR* (NM_000372.4), *OCA2* (NM_000275.2) and *SLC45A2* (NM_016180.3) genes were PCR amplified from genomic DNA following standard amplification protocols. Amplified products were sequenced on both strands (primer sequences and experimental conditions available on request). Sequencing analysis was performed on an ABI Prism 3730 genetic analyzer (Life Technologies Corporation, Carlsbad, CA, USA) and the SeqScape software (Life Technologies Corporation) was used for mutation detection.

To screen for the presence of *TYR* and *OCA2* exons rearrangements (deletion/duplication), the MLPA (Multiplex Ligation-dependent Probe Amplification) kit SALSA P325 (MRC-Holland, Amsterdam, the Netherlands) was used following the manufacturer's instructions. The individual peak corresponding to each exon was identified based on the difference in migration relative to the size standards 500 LIZ™ (Life Technologies Corporation). The peak area of each fragment was compared with that of three control samples. Raw data were analyzed using the Coffalyser software v.9 (MRC-Holland, Amsterdam, the Netherlands). Each positive result was confirmed by two independent experiments and on two different DNA extractions.

In vitro analysis of splicing mutations

Computer-assisted analysis for splice-site prediction was accomplished using the NNSPLICE 0.9 (http://www.fruitfly.org/seq_tools/splice.html) and the Human Splicing Finder (HSF, http://www.umd.be/HSF/) programs.

For the functional characterization of the two newly identified mutations in *SLC45A2*, the relevant genomic DNA region was cloned in the hybrid alpha-globin-fibronectin minigene plasmid (pBS-KS_modified), in which the alternatively spliced extra-domain-B exon of fibronectin was removed to generate a site for the insertion of the exon under study.[26] In particular, for the analysis of the c.888G>A substitution, a 746-bp fragment of *SLC45A2* (introns 2 to 4) was PCR amplified from the patient's genomic DNA using the following primers: OCA4_ex3_*Nde*I_F 5′-ggaattccatatgAGCCCAAGATCACAGCAAGT-3′ and OCA4_ex3_*Nde*I_R 5′-ggaattccatatgGGACCAGAGGAAAATGCAGA-3′ and cloned into the pBS-KS_modified vector (lowercase letters indicate nucleotides added to the primers to introduce the *Nde*I restriction site).

Concerning the c.1156+2dupT variant, a 572-bp region of *SLC45A2* (including exon 5 with its flanking intronic sequences) was PCR amplified using the following primer pair OCA4_ex5_*Nde*I_F 5′-ggaattccatatgGAGAT CTGATGCAGCAAGCA-3′ and OCA4_ex5_*Nde*I_R 5′-ggaattccatatgGGAACC CACTGATTCCAAGA-3′, and cloned into the pBS-KS_modified vector.

The obtained wild-type (pBS-KS-OCA4_ex3_wt and pBS-KS-OCA4_ex5_wt) and mutant (pBS-KS-OCA4_ex3_mut and pBS-KS-OCA4_ex5_mut) plasmids were isolated by the PureYield Plasmid Miniprep System (Promega, Madison, WI, USA). The correct orientation of the two inserts, as well as the presence/absence of either the c.888G>A or the c.1156 +2dupT variant, were verified by direct DNA sequencing.

Recombinant plasmids were used for splicing assays, by transiently transfecting 4 μg of each minigene construct in HeLa cells (cultured according standard procedures). In each experiment, an equal number of cells (2.5×10^5) were transfected with the jetPRIME reagent (Euroclone, Wetherby, UK) in 6-well plates, as described by the manufacturer. Total RNA was isolated from the cells 24 h after transfection, using the Eurozol kit (Euroclone). Random hexamers and the ImProm-II Reverse Transcriptase (Promega) were used to perform first-strand cDNA synthesis starting from 500 ng of total RNA, according to the manufacturer's instructions. Of a total of 20 μl of the reverse-transcription (RT) reaction, 1 μl was used as template for amplifications, using primers annealing to the flanking α-globin/*FN1* exonic sequences (α2-3 and Bra2 primers, see Figure 1a). RT-PCRs were performed under standard conditions using the GoTaq DNA Polymerase (Promega) on a Mastercycler EPgradient (Eppendorf, Hamburg, Germany).

Quantitation of splicing isoforms by fluorescent RT-PCR

To obtain a relative quantitation of the different isoforms generated by minigene assays, competitive fluorescent RT-PCRs were performed on RNA from transfected cells. To this aim, the same oligonucleotide pair adopted for splicing assays was used, with the reverse Bra2 primer labelled with the 6-FAM fluorophore. Amplified fragments were separated by capillary electrophoresis on an ABI-3130XL Genetic Analyzer (Life Technologies Corporation) and quantitated by the GeneMapper v4.0 software (Life Technologies Corporation). The sum of all fluorescence peak areas in a single run was set equal to 100%, and the relative quantity of each transcript expressed as a fraction of the total.

RESULTS

Clinical features of the patient

The proband is a 3-year-old male of Italian origin. He has blonde platinum hair, pale skin and pale grey-blue iris. At 13 months, he had normal pupillary responses, horizontal nystagmus, astigmatism, no strabismus, iris transillumination (II-III grade of the Summers scale), foveal hypoplasia and albinotic fundus. Visual acuity could not be determined for low collaboration. No other apparent malformation was found at physical examination.

No consanguineous marriage was claimed by the parents.

The mutational screening of the proband revealed two novel putative splicing mutations in SLC45A2

The molecular screening of the patient was performed first by analyzing *TYR* and *OCA2* genes, which represent the most frequently mutated OCA genes in the Italian population.[27] Neither sequencing nor MLPA analysis of both genes revealed any point

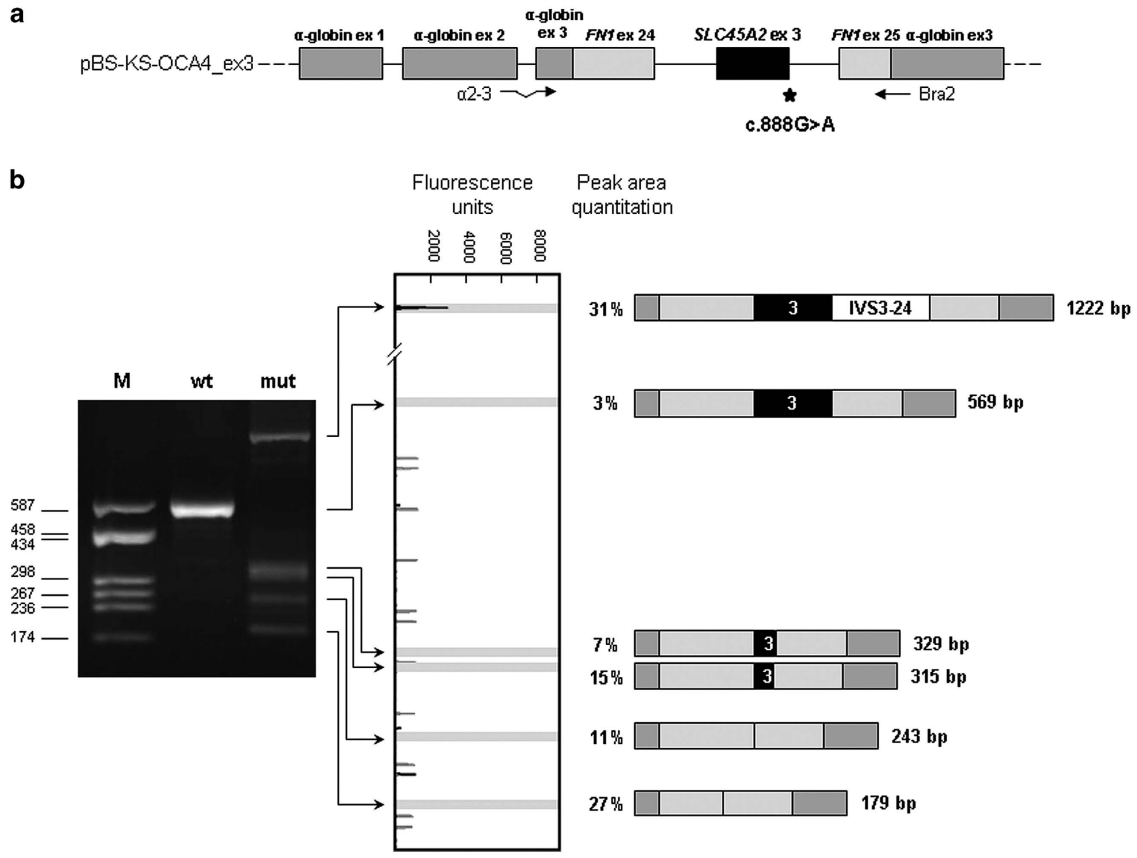


Figure 1 *In vitro* analysis of the effect of the c.888G>A (p.Q296Q) mutation on *SLC45A2* pre-mRNA splicing. (**a**) Schematic representation of the hybrid pBS-KS-OCA4_ex3 minigene: α-globin and fibronectin (*FN1*) exons are represented by gray boxes, whereas introns are represented by lines (not to scale). The *SLC45A2* exon 3 cloned, with its flanking intronic regions, into the *Nde*I site of the pBS-KS_modified vector is reported as a black box. The c.888G>A mutation, affecting the last nucleotide of exon 3, is indicated by a star. Primers used in RT-PCR experiments are indicated by arrows in correspondence of their position below the scheme of the minigene construct. (**b**) *Left panel*: RT-PCR products, obtained from RNA of HeLa cells transfected with the wild-type (wt) or the mutant (mut) minigene construct, separated on a 2% agarose gel, are shown. M: molecular weight marker (pUC9-*Hae*III). *Middle panel*: the GeneMapper window shows fluorescence peaks corresponding to the molecular species amplified by RT-PCR. Red peaks represent the size standard (ROX-500 HD); blue peaks (shaded in gray) correspond to the RT-PCR-labeled products, whose abundance is reported on the right of the panel (%). The *x* axis indicates fluorescence units. *Right panel*: schematic representation of the obtained RT-PCR products (verified by Sanger sequencing). The length of each fragment is indicated.

mutation or exon rearrangement. Subsequently, we analyzed the *SLC45A2* gene and identified two novel putative pathogenic variants. The first one, a c.888G>A transition, is a synonymous change (p.Q296Q) that involves the last nucleotide of exon 3, suggesting that it might affect splicing. The second one, a c.1156+2dupT, is a single-nucleotide insertion within the consensus sequence of the donor splice site of intron 5 (IVS5+2dupT). The proband's father was heterozygous for the c.888G>A variant, while the mother was heterozygous for the c.1156+2dupT insertion.

These two newly identified variants were absent in the cohort of 6503 European American and African American individuals, whose exome data are freely available through the Exome Variant Server.[28]

In vitro characterization of the two novel variants demonstrate their effect on splicing

The *in silico* analysis for the c.888G>A transition was performed by two different computer-assisted software for splice-site prediction: HSF detected the wild-type donor splice-site of exon 3 with a score of 85.90 and the mutated one with a score of 75.32. The NNSPLICE 0.9 software detected the wild-type donor splice-site with a score of 0.94, whereas, in the mutant sequence, it was no longer identified.

Given the unavailability of a suitable biological specimen to extract RNA, exon 3 of the *SLC45A2* gene with the surrounding intronic sequences was cloned, either in the wild-type or in the mutant version, into a pBS-KS_modified hybrid minigene vector (Figure 1a). The obtained constructs (pBS-KS-OCA4_ex3_wt and pBS-KS-OCA4_ex3_mut) were transiently transfected into HeLa cells and *SLC45A2* splicing products were analyzed by RT-PCR. All amplified fragments were then subjected to direct sequencing to characterize the aberrant splicing events. Indeed, transfection with the mutant vector gave rise to multiple aberrant products (Figure 1b). One PCR product was larger than the expected wild-type one (1222 vs 569 bp). In addition, four amplicons shorter than the wild type (329, 315, 243, and 179 bp long) were detected: (i) the first two caused by the retention of a shorter exon 3 because of the activation of two cryptic exonic donor splice sites, 86 and 72 bp downstream of the wild-type splice acceptor site, respectively; (ii) the third corresponded to the complete skipping of exon 3; and (iii) the last one resulted from the skipping of a part of fibronectin exon 24

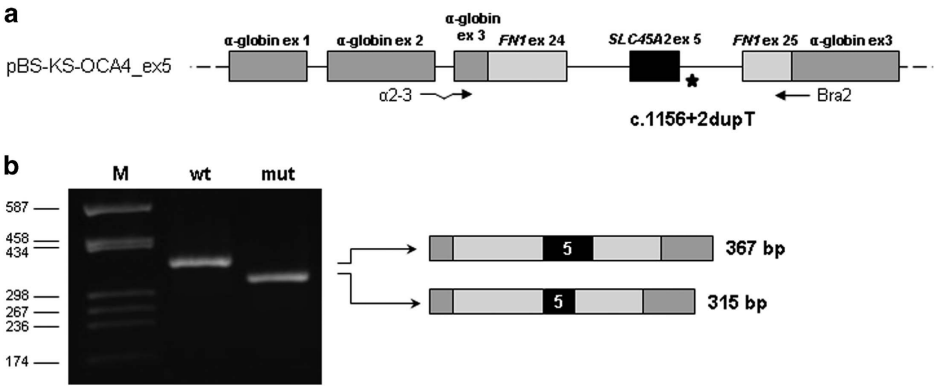


Figure 2 *In vitro* analysis of the effect of the c.1156+2dupT mutation on *SLC45A2* pre-mRNA splicing. (**a**) Schematic representation of the hybrid pBS-KS-OCA4_ex5 minigene: α-globin and fibronectin (*FN1*) exons are represented by gray boxes, whereas introns are represented by lines (not to scale). The *SLC45A2* exon 5 cloned, with its flanking intronic regions, into the *Nde*I site of the pBS-KS_modified vector is reported as a black box. The c.1156+2dupT insertion, affecting the donor splice site of exon 5, is indicated by a star. Primers used in RT-PCR experiments are indicated by arrows in correspondence of their position below the scheme of the minigene construct. (**b**) *Left panel*: RT-PCR products, obtained from RNA of HeLa cells transfected with the wild-type (wt) or the mutant (mut) minigene construct, separated on a 2% agarose gel, are shown. M: molecular weight marker (pUC9-*Hae*III). *Right panel*: schematic representation of the obtained RT-PCR products (verified by Sanger sequencing). The length of each fragment is indicated.

together with the entire *SLC45A2* exon 3, due to the activation of an exonic cryptic donor splice site (Figure 1b). While the skipping of the entire exon 3 or the inclusion of a shorter 72-bp-long exon 3 are predicted to cause a frameshift (leading to the introduction of a premature stop codon after 56 or 80 amino acids, respectively), the insertion of a 86-bp-long exon 3 is expected to determine the in-frame deletion of 80 amino acids. The relative quantitation by competitive fluorescent RT-PCR of all splicing isoforms generated by the minigene assay evidenced the presence of a residual amount (about 3%) of the wild-type transcript (Figure 1b).

The effect of the putative splicing mutation c.1156+2dupT was analyzed with the same approach. Computer-assisted splice-site analysis using HSF predicted a score reduction of the donor splice site (from 84.38 to 65.26) in the presence of the duplication. Moreover, a stronger cryptic donor splice site (located 72 bp downstream of the physiologic acceptor splice site of exon 5) was predicted. Instead, NNSPLICE 0.9 predicted the complete abolishment of the mutant donor splice-site compared with the wild type (whose score was 0.99).

Subsequently, appropriate minigene constructs were generated by cloning the relevant PCR-amplified genomic fragment (from intron 4 to 5) either in the wild-type or in the mutant version. Transfection with the mutant vector gave rise to a single splicing product, shorter than the expected wild-type one (315 vs 367 bp). Direct DNA sequencing confirmed that this aberrant product is due to the activation of the exonic cryptic donor splice site identified by the HSF *in silico* analysis. This aberrant splicing event would result in a frameshift followed by the introduction of a premature stop at codon 380 of the mutant protein, preceded by an abnormal sequence of 11 amino acids (p.V369Tfs*12) (Figure 2).

DISCUSSION

In this study, we analyzed the genetic defects underlying OCA4 in an Italian albino patient in whom we identified two novel splicing mutations, one of which is a synonymous transition (c.888G>A, p.Q269Q) involving the last nucleotide of exon 3 while the second one (c.1156+2dupT) lies in the consensus sequence of the donor splice site of *SLC45A2* intron 5 (IVS5+2dupT).

We applied both *in silico* prediction and *in vitro* functional validation to experimentally verify the possible effect of the newly identified mutations on *SLC45A2* pre-mRNA splicing. For both mutations, computer-assisted splice-site prediction analysis recognized a significant difference between the wild-type and the mutant sequences, suggesting that they might impact *SLC45A2* splicing. These predictions were tested by a well-validated *in vitro* hybrid minigene approach,[26,29,30] which we have recently applied to the study of two novel splicing mutations in the *OCA2* gene.[31]

Functional analysis demonstrated that both mutations inactivate the relevant physiologic donor splice site and, at the same time, may lead to the activation of cryptic sites. In particular, the c.888G>A mutation seems to induce multiple aberrant splicing events, ranging from intron retention, complete exon skipping and activation of cryptic donor splice sites. Noteworthy, this is the second exonic splicing mutation identified and characterized in a OCA gene. Instead, the main functional consequence of the c.1156+2dupT mutation on *SLC45A2* splicing is the unmasking of a cryptic exonic donor splice site, leading to the inclusion of a shorter out-of-frame exon 5. Although the *in vitro* characterization of splicing mutations might not completely mirror/recapitulate the actual splicing events occurring *in vivo*, it certainly provides a strong support to the pathogenic role of the identified variants. Indeed, minigene experiments have clearly evidenced a nearly complete loss of the wild-type transcript determined by the presence of either mutation (Figures 1 and 2). At the same time, most of the aberrant isoforms that might be produced *in vivo* as a consequence of the two identified mutations are predicted to cause the introduction of a premature termination codon, which might cause the transcript to be recognized and degraded by the nonsense-mediated mRNA decay mechanism.[32]

The activation of multiple cryptic splice sites within exon 3 observed in minigene studies prompted us to investigate the recognition of this exon in physiologic conditions. Interestingly, three *SLC45A2* mRNA isoforms skipping exon 3 are annotated in GenBank (accession numbers: BC003597 deriving from melanoma; CU678525 and CU678524 deriving from the ORFeome project) (Supplementary Figure 1), suggesting that these splicing events could also occur *in vivo*. In addition, *in silico* prediction with the HSF algorithm identifies a cryptic donor splice site within exon 3 having a score (80.73) comparable to that of the physiologic one (85.9). The use of this cryptic splice site would lead to the inclusion of a 86-bp-long in-

frame exon 3, corresponding to one of the aberrant isoforms detected in our minigene experiments (Figure 1b and Supplementary Figure 1), again suggesting its possible occurrence *in vivo* in physiologic conditions. Therefore, we attempted to characterize the splicing events involving exon 3 by RT-PCR assays performed on RNA extracted from hair follicles of a non-albino control. These experiments allowed the identification, beside the wild-type transcript, of an additional *SLC45A2* isoform. Sequencing revealed that such aberrant product corresponds to the inclusion of the 86-bp-long in-frame exon 3 (Supplementary Figure 1), confirming that the predicted cryptic donor splice site can be recognized even in the absence of the c.888G>A mutation. We cannot exclude that the skipping of the whole exon 3 could also occur physiologically, although its detection can be hampered by the concomitant effect of nonsense-mediated mRNA decay on this out-of-frame isoform. Indeed, evidence supporting nonsense-mediated mRNA decay as a mechanism controlling *SLC45A2* mRNA levels comes also from studies on the orthologous gene in birds.[33] Taken together, these observations seem to suggest that exon 3 might be 'leaky' in recognition during mRNA splicing process.

CONFLICT OF INTEREST

The authors declare no conflict of interest. The authors alone are responsible for the content and writing of the paper.

ACKNOWLEDGEMENTS

We thank Dr Emanuele Buratti (International Centre for Genetic Engineering and Biotechnology, Trieste, Italy) for kindly providing the hybrid pBS-KS vector. Laura Locatelli is also acknowledged for her technical assistance and enthusiasm. We would also like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010). We also thank 'Fondazione Paolina Brugnatelli' for its valuable support.

1 Grønskov, K., Ek, J. & Brondum-Nielsen, K. Oculocutaneous albinism. *Orphanet. J. Rare Dis.* **2**, 43–50 (2007)
2 Kirkwood, B. J. Albinism and implications with vision. *Insight* **34**, 13–16 (2009)
3 King, R. A., Hearing, V. J., Creel, D. J. & Oetting, W. S. Albinism. Scriver, C. R. *et al. The Metabolic and Molecular Basis of Inherited Disease*, 8th edn. Vol.II, 5587–5627 (McGraw-Hill, New York, NY, USA, 2001)
4 Tomita, Y., Takeda, A., Okinaga, S., Tagami, H. & Shibahara, S. Human oculocutaneous albinism caused by single base insertion in the tyrosinase gene. *Biochem. Biophys. Res. Commun.* **164**, 990–996 (1989)
5 Rinchik, E. M., Bultman, S. J., Horsthemke, B., Lee, S. T., Strunk, K. M., Spritz, R. A. *et al.* A gene for the mouse pink-eyed dilution locus and for human type II oculocutaneous albinism. *Nature* **361**, 72–76 (1993)
6 Boissy, R. E., Zhao, H., Oetting, W. S., Austin, L. M., Wildenberg, S. C., Boissy, Y. L. *et al.* Mutation in and lack of expression of tyrosinase-related protein-1 (TRP-1) in melanocytes from an individual with brown oculocutaneous albinism: a new subtype of albinism classified as "OCA3". *Am. J. Hum. Genet.* **58**, 1145–1156 (1996)
7 Newton, J. M., Cohen-Barak, O., Hagiwara, N., Gardner, J. M., Davisson, M. T., King, R. A. *et al.* Mutations in the human orthologue of the mouse underwithe gene (uw) underlie a new form of Oculocutaneous albinism, OCA4. *Am. J. Hum. Genet.* **69**, 981–988 (2001)
8 Gargiulo, A., Testa, F., Rossi, S., Di Iorio, V., Fecarotta, S., de Berardinis, T. *et al.* Molecular and clinical characterization of albinism in a large cohort of Italian patients. *Invest. Ophthalmol. Vis. Sci.* **52**, 1281–1289 (2011)
9 Kausar, T., Bhatti, M. A., Ali, M., Shaikh, R. S. & Ahmed, Z. M. OCA5, a novel locus for non-syndromic oculocutaneous albinism, maps to chromosome 4q24. *Clin. Genet.* **84**, 91–93 (2013)
10 Morice-Picard, F., Lasseaux, E., François, S., Simon, D., Rooryck, C., Bieth, E. *et al. SLC24A5* mutations are associated with non-syndromic oculocutaneous albinism. *J. Invest. Dermatol.* **134**, 568–571 (2014)
11 Wei, A. H., Zang, D. J., Zhang, Z., Liu, X. Z., He, X., Yang, L. *et al.* Exome sequencing identifies *SLC24A5* as a candidate gene for nonsyndromic oculocutaneous albinism. *J. Invest. Dermatol.* **133**, 1834–1840 (2013)
12 Grønskov, K., Dooley, C. M., Ostergaard, E., Kelsh, R. N., Hansen, L., Levesque, M. P. *et al.* Mutations in c10orf11, a melanocyte-differentiation gene, cause autosomal-recessive albinism. *Am. J. Hum. Genet.* **92**, 415–421 (2013)
13 Montoliu, L., Grønskov, K., Wei, A. H., Martínez-García, M., Fernández, A., Arveiler, B. *et al.* Increasing the complexity: new genes and new types of albinism. *Pigment Cell Melanoma Res.* **27**, 11–18 (2014)
14 Schiaffino, M. V., Bassi, M. T., Galli, L., Renieri, A., Bruttini, M., De Nigris, F. *et al.* Analysis of the OA1 gene reveals mutations in only one-third of patients with X-linked ocular albinism. *Hum. Mol. Genet.* **4**, 2319–2325 (1995)
15 Rundshagen, U., Zuhlke, C., Opitz, S., Schwinger, E. & Käsmann-Kellner, B. Mutations in the MATP gene in five German patients affected by oculocutaneous albinism type 4. *Hum. Mutat.* **23**, 106–110 (2004)
16 Sengupta, M., Chaki, M., Arti, N. & Ray, K. *SLC45A2* variations in Indian oculocutaneous albinism patients. *Mol. Vis.* **13**, 1406–1411 (2007)
17 Suzuki, T., Inagaki, K., Fukai, K., Obana, A., Lee, S. T. & Tomita, Y. A Korean case of oculocutaneous albinism type IV caused by a D157N mutation in the MATP gene. *Br. J. Dermatol.* **152**, 174–175 (2005)
18 Grønskov, K., Ek, J., Sand, A., Scheller, R., Bygum, A., Brixen, K. *et al.* Birth prevalence and mutation spectrum in Danish patients with autosomal recessive albinism. *Invest. Ophthalmol. Vis. Sci.* **50**, 1058–1064 (2009)
19 Konno, T., Abe, Y., Kawaguchi, M., Storm, K., Biervliet, M., Courtens, W. *et al.* Oculocutaneous albinism type IV: a boy of Moroccan descent with a novel mutation in *SLC45A2*. *Am. J. Med. Genet.* **149A**, 1773–1776 (2009)
20 Inagaki, K., Suzuki, T., Shimizu, H., Ishii, N., Umezawa, Y., Tada, J. *et al.* Oculocutaneous albinism type 4 is one of the most common types of albinism in Japan. *Am. J. Hum. Genet.* **74**, 466–471 (2004)
21 Costin, G. E., Valencia, J. C., Vieira, W. D., Lamoreux, M. L. & Hearing, V. J. Tyrosinase processing and intracellular trafficking is disrupted in mouse primary melanocytes carrying the underwhite (uw) mutation. A model for oculocutaneous albinism (OCA) type 4. *J. Cell Sci.* **116**, 3203–3212 (2003)
22 Cullinane, A. R., Vilboux, T., O'Brien, K., Curry, J. A., Maynard, D. M., Carlson-Donohoe, H. *et al.* Homozygosity mapping and whole-exome sequencing to detect *SLC45A2* and *G6PC3* mutations in a single patient with oculocutaneous albinism and neutropenia. *J. Invest. Dermatol.* **131**, 2017–2025 (2011)
23 Bartölke, R., Heinisch, J. J., Wieczorek, H. & Vitavska, O. Proton-associated sucrose transport of mammalian Solute Carrier Family 45: an analysis in Saccharomyces cerevisiae. *Biochem. J.* **464**, 193–201 (2014)
24 Suzuki, T. & Tomita, Y. Recent advances in genetic analyses of oculocutaneous albinism type 2 and 4. *J. Dermatol. Sci.* **51**, 1–9 (2008)
25 HGMD Professional http://biobase-international.com/hgmd/pro/all.php.
26 Baralle, M., Baralle, D., De Conti, L., Mattocks, C., Whittaker, J., Knezevich, A. *et al.* Identification of a mutation that perturbs NF1 agene splicing using genomic DNA samples and a minigene assay. *J. Med. Genet.* **40**, 220–222 (2003)
27 Mártinez-García, M. & Montoliu, L. Albinism in Europe. *J. Dermatol.* **40**, 39–24 (2013)
28 EVS, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL http://evs.gs.washington.edu/EVS/); accessed January 2015.
29 Costantino, L., Rusconi, D., Soldà, G., Seia, M., Paracchini, V., Porcaro, L. *et al.* Fine characterization of the recurrent c.1584+18672A>G deep-intronic mutation in the cystic fibrosis transmembrane conductance regulator gene. *Am. J. Respir. Cell Mol. Biol.* **48**, 619–625 (2013)
30 Paraboschi, E. M., Rimoldi, V., Soldà, G., Tabaglio, T., Dall'Osso, C., Saba, E. *et al.* Functional variations modulating PRKCA expression and alternative splicing predispose to multiple sclerosis. *Hum. Mol. Genet.* **23**, 6746–6761 (2014)
31 Rimoldi, V., Straniero, L., Asselta, R., Mauri, L., Manfredini, E., Penco, S. *et al.* Functional characterization of two novel splicing mutations in the OCA2 gene associated with oculocutaneous albinism type II. *Gene* **537**, 79–84 (2014)
32 Popp, M. W. & Maquat, L. E. Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu. Rev. Genet.* **47**, 139–165 (2013)
33 Gunnarsson, U., Hellström, A. R., Tixier-Boichard, M., Minvielle, F., Bed'hom, B., Ito, S. *et al.* Mutations in *SLC45A2* cause plumage color variation in chicken and Japanese quail. *Genetics* **175**, 867–877 (2007)

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Journal of Human Genetics (2016), 1–8

www.nature.com/jhg

ORIGINAL ARTICLE

Whole-gene *CFTR* sequencing combined with digital RT-PCR improves genetic diagnosis of cystic fibrosis

Letizia Straniero[1,7], Giulia Soldà[2,3,7], Lucy Costantino[4], Manuela Seia[4], Paola Melotti[5], Carla Colombo[6], Rosanna Asselta[2,3] and Stefano Duga[2,3]

Despite extensive screening, 1–5% of cystic fibrosis (CF) patients lack a definite molecular diagnosis. Next-generation sequencing (NGS) is making affordable genetic testing based on the identification of variants in extended genomic regions. In this frame, we analyzed 23 CF patients and one carrier by whole-gene *CFTR* resequencing: 4 were previously characterized and served as controls; 17 were cases lacking a complete diagnosis after a full conventional *CFTR* screening; 3 were consecutive subjects referring to our centers, not previously submitted to any screening. We also included in the custom NGS design the coding portions of the *SCNN1A*, *SCNN1B* and *SCNN1G* genes, encoding the subunits of the sodium channel ENaC, which were found to be mutated in CF-like patients. Besides 2 novel *SCNN1B* missense mutations, we identified 22 previously-known *CFTR* mutations, including 2 large deletions (whose breakpoints were precisely mapped), and novel deep-intronic variants, whose role on splicing was excluded by *ex-vivo* analyses. Finally, for 2 patients, compound heterozygotes for a *CFTR* mutation and the intron-9c.1210-34TG$_{[11-12]}$T$_5$ allele—known to be associated with decreased *CFTR* mRNA levels—the molecular diagnosis was implemented by measuring the residual level of wild-type transcript by digital reverse transcription polymerase chain reaction performed on RNA extracted from nasal brushing.
Journal of Human Genetics advance online publication, 4 August 2016; doi:10.1038/jhg.2016.101

INTRODUCTION

Cystic fibrosis (CF; OMIM#219700) is one of the most common autosomal recessive disorders with a median incidence in Europe of 1 in 3500.[1] The classical symptoms include chronic obstructive pulmonary disease, exocrine pancreatic insufficiency, and elevation of sodium and chloride concentrations in the sweat.[2] Moreover, most CF males present with infertility as a result of congenital bilateral absence of the vas deferens.[3] CF is caused by mutations in the CF transmembrane conductance regulator (*CFTR*; OMIM*602421) gene.[4,5] It encodes a protein that functions as a Cyclic adenosine monophosphate-activated chloride channel at the apical membrane of epithelial cells.[6] To date, over 1600 mutations have been described in the *CFTR* gene (http://www.genet.sickkids.on.ca/cftr, accessed 22 January, 2016, excluding sequence variations that are possibly polymorphisms and variants with an unknown functional meaning). About 14% of them are classified as splicing defects, but for ~40% of these, the specific effect on pre-mRNA splicing is unknown.[7] Despite extensive genetic screening, 1–5% of CF patients lack a definite molecular diagnosis, even after the sequencing of all exons and splice junctions.[8]

The genotype–phenotype relation in CF is very complex: some phenotypic features are mainly shaped by the *CFTR* genotype, whereas others are strongly influenced by modifying genetic and environmental factors. A prototypic example is the phenotypic variability associated with the (TG)m(T)n polymorphic locus at the 3′ end of *CFTR* intron 9 (formerly known as intron 8). This polymorphism determines the efficiency of exon-10 splicing, with alleles bearing less T repeats and more TG repeats associated with a lower rate of exon-10 inclusion.[9–11] The exon-10 skipped *CFTR* transcript encodes a misfolded and nonfunctional protein,[12] whose expression, in combination with a CF-causing mutation *in trans*, may be associated with monosymptomatic forms of CF, like congenital bilateral absence of the vas deferens or chronic airway diseases.[13] The T_5 allele is also known to modify the expressivity of the p.Arg117His mutation when *in cis*.[14]

To further complicate the CF genotype-phenotype picture, it has been suggested that mutations in genes coding for the multi-subunit amiloride-sensitive sodium channel ENaC (*SCNN1A*, *SCNN1B* and *SCNN1G*) could cause a CF-like phenotype.[15] These genes are expressed primarily on the apical membrane of the epithelial cells of kidney, lung and colon. Importantly, in the airways, the ENaC channel activity is inhibited by CFTR. Accordingly, in CF, the ENaC activity is elevated, thus resulting in a marked increase in sodium intake into epithelial cells.[16] The hypothesis is that an altered ENaC function, caused by mutations in one of these 3 genes, can ultimately mimic and/or contribute to the CF phenotype.[15]

[1]Department of Medical Biotechnology and Translational Medicine, University of Milan, Milan, Italy; [2]Department of Biomedical Sciences, Humanitas University, Milan, Italy; [3]Humanitas Clinical and Research Center, Milan, Italy; [4]Medical Genetics Laboratory, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy; [5]Cystic Fibrosis Center, Azienda Ospedaliera Universitaria Integrata di Verona, Verona, Italy and [6]Cystic Fibrosis Center of Milan, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
[7]These authors contributed equally to this work.
Correspondence: Professor S Duga, Department of Biomedical Sciences, Humanitas University, Via A Manzoni 113, Rozzano, 20089 Milan, Italy.
E-mail: stefano.duga@hunimed.eu or stefano.duga@humanitasresearch.it
Received 7 April 2016; revised 31 May 2016; accepted 7 July 2016

In recent years, the introduction of massively parallel next-generation sequencing (NGS) has allowed the simultaneous analysis of hundred thousand DNA fragments and thus dramatically changed our approach to the genetic analysis of inherited diseases.[17] The major advantage offered by NGS is the ability to produce an enormous volume of data in a time- and cost-efficient way. This, combined with the possibility to multiplex both the capture of target genomic regions and the sequencing, allows the simultaneous analysis of single genes or panel of selected genes in a large number of patients at unprecedentedly possible speed and costs. At present, commercial kits for amplicon resequencing of *CFTR* exons and NGS are available from different companies or have been in-house developed by research groups.[18] Target capture and sequencing protocols have also been developed to screen the whole *CFTR* 190-kb-long genomic region. However, to our knowledge, whole-gene sequencing was applied only as a proof-of-principle approach, to search for mutations in an already characterized cohort of CF patients.[19]

Here, we developed a whole-gene NGS protocol from one side to assess the efficacy/sensitivity of targeted resequencing and from the other to test the method for molecular screening of CF in a panel of Italian patients with either incomplete or no genetic diagnosis. In 2 cases, absolute mRNA molecule count by digital reverse-transcription (RT) PCR assays was used to precisely quantify the residual level of the *CFTR* wild-type transcript, hence completing the molecular diagnosis.

MATERIALS AND METHODS

This study was approved by local Ethical Committees and was performed according to the Declaration of Helsinki and to the Italian legislation on sensible data recording. Signed informed consent was obtained from all participants and from parents of subjects younger than 18 years.

NGS experiments and data analysis

NGS libraries were prepared starting from 1 μg of genomic DNA from each patient, using a Nimblegen SeqCap EZ Library custom design capture kit (Roche/Nimblegen, Madison, WI, USA). Equal amounts of 6 samples were pooled prior to capture. Libraries were sequenced as 75-bp paired-end reads on an HiSeq2000 (Illumina, San Diego, CA, USA), according to the manufacturer's protocols. Details on DNA extraction, target capture, library preparation, quality check and sequencing are reported in Supplementary Materials.

Pre-capture pooling, target-enrichment, as well as NGS were performed through an external service provider (Yale Genome Center; Yale University, USA).

Data were analyzed between November 2012 and March 2013 using an in-house developed pipeline (Supplementary Materials). Considering that indel and structural variant (SV) calling still represent a tricky step, all data were reanalyzed with the GeneSpring NGS software (Agilent Technologies, Palo Alto, CA, USA), implementing algorithms specifically developed for the identification of indels and SV (that is, 'SNP detection' and 'SV detection' options).

Ex-vivo splicing assays of selected deep-intronic variants

To generate hybrid minigene constructs, diverse *CFTR* regions were PCR amplified from the genomic DNA of patients and cloned into the α-globin-fibronectin hybrid plasmid (modified pBS-KS),[20] as previously described.[21] Splicing assays were performed by transiently transfecting 4 μg of each of the 8 recombinant vectors into HeLa cells (cultured according to standard procedures). In each experiment, an equal number of cells (2.5×10^5) were transfected with the FuGENE 6 reagent (Roche) in 6-well plates, as described by the manufacturer. Total RNA was isolated from cells 24 h after transfection, using the Eurozol kit (Euroclone, Wetherby, UK). Reverse transcription polymerase chain reactions (RT-PCRs) were carried out with primers mapping in the flanking fibronectin exonic regions of the plasmid. All primer sequences are available on request.

Absolute quantification of *CFTR* transcripts

RNA from nasal brushing was obtained from 2 CF patients and 9 healthy controls, and reverse transcribed as described in Supplementary Materials.

The level of wild-type *CFTR* transcripts was evaluated by digital RT-PCR. Reaction mixtures were prepared by combining 1 μl of cDNA with the QuantStudio 3D Digital PCR Master Mix and a custom TaqMan assay (Sigma-Aldrich, Milan, Italy), designed to amplify *CFTR* transcripts spanning exons 10-11. For the normalization step, a second TaqMan assay was designed to co-amplify, in the same tube, the hydroxymethylbilane synthase (HMBS) mRNA. The sequences of primers and probes used in digital RT-PCR assays are listed in Supplementary Table 1.

Digital RT-PCRs were performed on a QuantStudio 3D Digital PCR System (Thermo Fisher Scientific, Wilmington, DE, USA). Each reaction mixture was loaded onto a QuantStudio 3D Digital PCR Chip, which contains 20 000 wells, and cycled under standard conditions for 40 cycles. End-point fluorescence data were collected and analyzed using the QuantStudio 3D Digital PCR Instrument and the QuantStudio 3D AnalysisSuite according to the manufacturer's instructions. The precision of all copy counts showed with a standard error of lower than 5%.

RESULTS

Patients

Among individuals referred to the Cystic Fibrosis Centers of Milan and Verona in the last 5 years, 23 CF patients and one CF carrier (named CF-1 to CF-24, all of Caucasian origin) were selected for being submitted to NGS of the entire *CFTR* gene as well as of the coding regions and exon/intron boundaries of the *SCNN1A*, *SCNN1B* and *SCNN1G* genes (Supplementary Table 2).

Among selected patients, 17 had an incomplete genetic diagnosis, even after screening by conventional copy-number analysis and Sanger sequencing. In particular, for 5 of them, no mutations were identified, whereas 12 were heterozygous for only one genetic defect (2 were carrier of the F508del mutation). Finally, 3 patients (CF-22, 23 and 24) were not subjected to any previous molecular screening. In all cases, the inclusion criterion for entering the study was the availability of the DNA from both parents.

As controls, we selected 3 patients (CF-15, 17 and 18) showing a complete genetic diagnosis (all compound heterozygous for substitutions/small deletions), and one heterozygous carrier of a large deletion (CF-12) (Supplementary Table 3).

NGS sequencing statistics

For NGS analysis, we chose a custom-designed target-enrichment strategy based on the Nimblegen SeqCap EZ Library capture kit. Probes were designed on the basis of the genomic coordinates of the entire 189-kb-long *CFTR* gene (human genome release GRCh37/hg19, chr7:117,120,017-117,308,718), and of all exons and splicing junctions of the ENaC genes (*SCNN1A*, mapping on chr12p13.31, and *SCNN1B* and *SCNN1G* both located on chr16p12.2). Altogether, the target regions accounted for a total of 207 508 bp.

A pilot group of DNAs was analyzed by performing the enrichment step multiplexing six samples, and the sequencing step in a single HiSeq2000 lane. The overall coverage of the target region was >98%, with only about 3 kb of *CFTR* intronic repetitive sequences left uncovered, and a mean coverage of ~3500× (in all cases the coverage was >2600×) (Supplementary Table 4). As the coverage was extremely high, we scaled up the number of pooled samples in the subsequent experiments: multiplexing 6 samples in the capture phase and 18 in the sequencing step, we obtained a mean depth >1150× (in all cases the coverage was >800×), again with a 98% coverage of the target region (Supplementary Table 4). Unfortunately, one patient (CF-20) failed the sequencing step.

On average, each subject resulted carrier of 162 genetic variants, 24 of which having a minor allele frequency <1% in the general population.

Identification of genetic variants in *CFTR*

Variants were annotated by filtering them against an in-house developed database and classified in one of the following categories: (1) already known pathogenic mutations (retrieved from: the Cystic Fibrosis Mutation Database, http://www.genet.sickkids.on.ca/app, the Human Gene Mutation Database, HGMD, http://www.hgmd.cf.ac.uk/ac/index.php, and from the CF Centre of Milan database); (2) potentially pathogenic variants included in dbSNP135; (3) common (minor allele frequency, MAF>1%), likely not pathogenic, variants annotated in dbSNP135; (4) rare single nucleotide variations (SNVs) of unknown functional relevance; (5) novel variants. Variants annotated in group 1 were considered of immediate diagnostic relevance, whereas variations included in groups 2, 4 and 5 were selected for Sanger sequencing validation. Probably due to the obtained high coverage, all variants were confirmed (100% concordance rate).

Our pipeline allowed the identification of all pathogenic variants already detected by routine screening (18 alleles for a total of 13 different mutations), including the 21-kb deletion in the CF-12 control individual. In addition, we identified: (i) 3 point mutations that were already reported in the literature to be clinically associated with CF but were not identified by Sanger sequencing (CF-4, 10 and 11); (ii) one large deletion (CF-16, see further); (iii) 5 known mutations in the newly-screened CF patients (CF-22, 23 and 24); and (iv) 22 novel deep-intronic variants of unknown significance (Table 1; Supplementary Table 5). Finally, the poly-thymidine-guanine and poly-thymidine c.1210–34$TG_{[11–13]}T_{[5–9]}$ tracts were also inspected. To this aim, the variant caller GeneSpring SNP detection was used, and it correctly assigned different alleles to all individuals (as subsequently verified by Sanger sequencing). In particular, we found 2 patients carrying the T_5 pathogenic variant, thus possibly completing the diagnosis for both of them (see further; Table 1).

Overall, for 6 of the 11 patients with incomplete genetic diagnosis successfully analyzed, NGS allowed the detection of the second causative allele, corresponding to a diagnostic yield of 54.6%. Conversely, no clearly pathogenic mutation was found in the five patients negative for CFTR mutations after previous screenings.

Characterization of *CFTR* large deletions

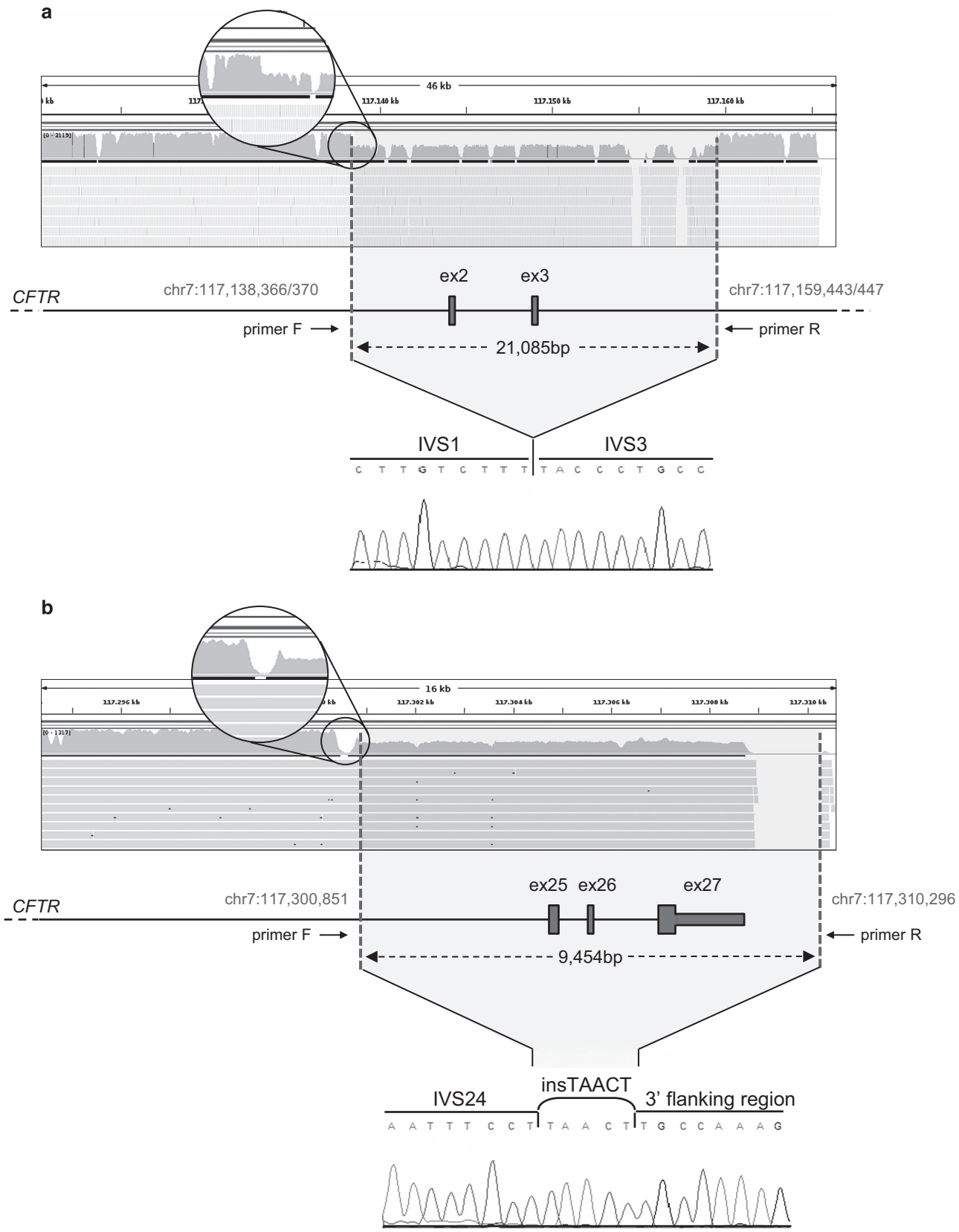
Using the specific algorithm 'SV detection' implemented in the GeneSpring software, we were able to identify 2 large deletions, both present in the heterozygous state (Table 1).

Table 1 List of mutations identified by NGS in the *CFTR* gene

Patient	*cDNA position*[a]	*Protein*[b]	*No. of missing CFTR mutations*	*Detection*[c]	*$TG_{[11–13]}T_{[5–9]}$ genotype*
CF-1	c.220C>T/c.3808G>A[d]	A74W/D1270N	1	Confirmed	$TG_{11}T_7$-$TG_{11}T_7$
CF-2	c.1585-1G>A	/	0	Confirmed	$TG_{10}T_7$-**$TG_{11}T_5$**
CF-3	—	—	2	—	$TG_{11}T_7$-$TG_{11}T_7$
CF-4	c.1650delA	G551VfsX18	0	This study	$TG_{11}T_7$-$TG_{10}T_7$
	c.1585-1G>A	/		Confirmed	
CF-5	c.2657+5G>A	/	0	Confirmed	$TG_{10}T_7$-**$TG_{12}T_5$**
CF-6	c.266A>G	Y89C	1	Confirmed	$TG_{11}T_7$-$TG_{10}T_7$
CF-7	—	—	2	—	$TG_{11}T_7$-$TG_{10}T_7$
CF-8	c.1521_1523delCTT	F508del	1	Confirmed	$TG_{10}T_9$-$TG_{10}T_9$
CF-9	c.1521_1523delCTT	F508del	1	Confirmed	$TG_{10}T_9$-$TG_{11}T_7$
CF-10	c.2657+5G>A	/	0	Confirmed	$TG_{11}T_7$-$TG_{10}T_7$
	c.1210-2A>C	/		This study	
CF-11	c.1040G>C	R347P	0	Confirmed	$TG_{11}T_7$-$TG_{11}T_7$
	c.4250delA	E1417EfsX15		This study	
CF-12	ex2-3del (21 kb)	/	0	Confirmed	$TG_{11}T_7$-$TG_{11}T_7$
CF-13	—	—	2	—	$TG_{10}T_9$-$TG_{11}T_7$
CF-14	—	—	2	—	$TG_{11}T_7$-$TG_{10}T_7$
CF-15	c.182T>C	L61P	0	Confirmed	$TG_{11}T_7$-$TG_{10}T_7$
	c.1002-1110_1113delTAAG	/		Confirmed	
CF-16	c.1013C>T	T338I	0	Confirmed	$TG_{11}T_7$-$TG_{10}T_7$
	ex25-27del (9.4 kb)	/		This study	
CF-17	c.1624G>T	G542X	0	Confirmed	$TG_{10}T_9$- $TG_{10}T_7$
	c.1002-1110_1113delTAAG	/		Confirmed	
CF-18	c.1624G>T	G542X	0	Confirmed	$TG_{10}T_9$- $TG_{11}T_7$
	c.349C>T	R117C		Confirmed	
CF-19	—	—	2	—	$TG_{11}T_7$-$TG_{10}T_7$
CF-21	c.1495C>G	P499A	1	Confirmed	$TG_{11}T_7$-$TG_{10}T_7$
CF-22	c.1521_1523delCTT	F508del	1	This study	$TG_{10}T_9$- $TG_{10}T_9$
CF-23	c.1624G>T	G542X	0	This study	$TG_{10}T_9$-$TG_{10}T_7$
	c.1002-1110_1113delTAAG	/		This study	
CF-24	c.14C>T	P5L	0	This study	$TG_{10}T_9$-$TG_{11}T_7$
	c.3909C>G	N1303K		This study	

The $TG_{11/12}T_5$ alleles are bolded. /, For splicing mutations, deep intronic mutations and large deletions, the effect at the protein level was not indicated.
[a]Nucleotide position is according to the NM_000492.3 sequence, starting from the first nucleotide of the translation start codon.
[b]When possible, mutations are named after their predicted effect at the protein level (according to the sequence deposited in GenBank under accession number NP_000483.3).
[c]'This study' indicates mutations already reported in the literature but newly detected in the analyzed patients by NGS (highlighted in gray).
[d]These 2 mutations are present *in cis*.

As expected, the first one was identified in individual CF-12, from one side confirming the already-available genetic diagnosis, and from the other underlining the ability of the NGS approach to detect such genetic lesions. The mutation consisted of a 21-kb-long deletion encompassing introns 1–3 of the *CFTR* gene (Figure 1a). The second structural variant was identified in patient CF-16: it is a



9.5-kb-long deletion involving the 3′ end of the *CFTR* gene (removing exons 25–27) (Figure 1b).

Inspection of the NGS read alignment also allowed an easy definition of the deletion breakpoints for both lesions. Breakpoints were confirmed by designing couples of primers mapping immediately upstream and downstream of the breakpoints predicted by the 'SV detection' software, and PCR amplifying the mutant allele from the patient's DNA. Concerning the 21-kb-long deletion (21 085 bp), Sanger sequence analysis of the PCR product evidenced that, while the wild-type sequence is characterized by the presence of a 4-bp-long CTTT sequence at both cleavage sites, in the deleted allele only one copy is retained. Hence, the precise location of the 5′ and 3′ breakpoints cannot be unambiguously determined, and may range between positions chr7:117,138,366-117,138,370 and chr7:117,159,443-117,159,447, respectively (Figure 1a). Concerning the 9.5-kb-long deletion (9454 bp), sequencing of the deleted allele evidenced the presence of five additional nucleotides (TAACT) between the breakpoints, as already described (Figure 1b).[22]

Ex-vivo characterization of deep-intronic variants

Among the 22 novel deep-intronic nucleotide substitutions, 4 variants (named V1–V4) were selected for further characterization because bioinformatically predicted to potentially interfere with normal splicing (Supplementary Table 6).

The molecular characterization of the putative V1–V4 mutations was tackled by using an *ex-vivo* approach, as RNA to be extracted from nasal epithelial cells of patients (or their relatives) was not available to the study. In particular, we generated recombinant minigene constructs (4 wild-type, 4 mutant; Supplementary Figure 1A) that were independently transfected in HeLa cells. Subsequent RT-PCR assays did not evidence any splicing alteration for transcripts originating from minigenes carrying the V1, V2 or V3 variants (Supplementary Figure 1B). Concerning the V4 variant, RT-PCR fragments amplified from cells transfected with the wild-type or the mutant construct were again identical, but in this case an additional band was visible for both samples. Sequencing analysis revealed that the additional band originates from the insertion of a pseudoexon of 91 nucleotides, corresponding to a fragment of *CFTR* intron 23 lying between positions IVS23+152 and IVS23+243. However, this unexpected pseudoexon activation is independent from the presence of the V4 variant, which was no further investigated.

Quantification of residual wild-type *CFTR* transcript by digital RT-PCR

Since the intron-9c.1210-34TG$_{11\text{-}12}$T$_5$ alleles are known to be associated with high rates of exon-10 skipping (and hence with low wild-type *CFTR* expression levels), we explored the possibility that the phenotype in patients 2 and 5 could indeed be explained by the presence of the T$_5$ allele in combination with the corresponding CF-causing mutation *in trans* (that is, the splicing mutations c.1585-1G>A and c.2657+5G>A, determining the out-of-frame skipping of exons 12 and 16, respectively).[23–25] To this aim, we set up digital RT-PCR assays for the absolute quantification of the wild-type *CFTR* mRNA, designing PCR primers to specifically amplify exon-10-containing transcripts (Figure 2a). Digital PCR[26,27] is an accurate technique allowing the absolute quantification of RNA transcripts by partitioning the PCR reaction mix over a large number of wells, so that each well contains a single copy or no copies of the target region. On the basis of the assumption of Poisson distribution of copies, the number of template copies originally present in the sample can be recalculated simply counting the number of wells in which amplification has successfully occurred.[28]

Digital RT-PCRs were performed on RNA extracted from nasal brushing of the 2 patients and of 9 healthy controls (none carrying the T$_5$ allele). Digital counts of the wild-type *CFTR* transcript resulted 58 and 76 molecules/μl for patient CF-2 and 5, respectively. These values correspond to 21 and 28% of the mean wild-type *CFTR* level measured in controls, in whom the number of molecules/μl ranged from 128 to 456 (Figure 2b).

Genetic variants in *SCNN1B*

Sequencing of *SCNN1A*, *SCNN1B* and *2SCNN1G* genes led to the identification of 2 missense variants in the heterozygous state, both located in the *SCNN1B* gene. In particular, in patient CF-3, the c.1313A>G transition (p.Glu438Gly) was identified, whereas in

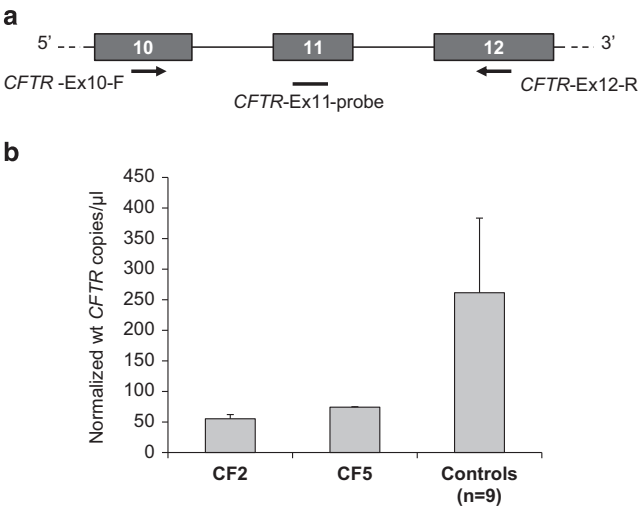


Figure 2 Evaluation of residual wild-type *CFTR* transcript using digital RT-PCR. Absolute expression levels were measured by digital RT-PCRs in patients CF-2 and CF-5 (that is, subjects carrying a null mutation and the T$_5$ allele in compound heterozygosity), as well as nine healthy controls (none being carrier of the T$_5$ allele). (**a**) Schematic representation of the *CFTR* gene between exons 10 and 12. Exons and introns are represented by boxes and lines, and are not drawn to scale. The positions of the primer couple used for the digital RT-PCR assay are shown by arrows, whereas the TaqMan probe is indicated by a horizontal line. (**b**) Histogram representing the results of digital RT-PCR. The number of copies of the target region is shown, normalized on the number of copies of the housekeeping *HMBS* mRNA (co-amplified with a second TaqMan assay). Bars represent the mean+the precision of the assay as calculated by the software.

Figure 1 Identification of the 21-kb-long and the 9.5-kb-long large deletions. (**a**, **b**) The main features characterizing, respectively, the 21-kb-long and the 9.5-kb-long deletions, both identified in the heterozygous state by NGS. For each deletion, a window of the IGV software showing the alignment of the reads in the region of the mutation is shown, together with a close-up view at the 5′ deletion breakpoint (highlighting the drop in the number of reads to about half). Below the IGV window, the relevant *CFTR* genomic region is displayed, with exons represented by boxes and introns by lines (drawn to scale). In this scheme, the position of primers used to PCR amplify the region across deletion breakpoints, as well as the breakpoint nucleotide positions along chromosome 7, are shown. In the lower part of the panel, the sequence electropherogram of the region corresponding to the deletion junction is shown. NGS, next-generation sequencing; PCR, polymerase chain reaction. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

patient CF-11 the c.818A>G substitution (p.Tyr273Cys) was disclosed. Both missense variants lead to an amino-acid change affecting a conserved residue (Supplementary Figure 2), and are found in European non-Finnish individuals from the Exome Aggregation Consortium (ExAC) database, both with a global frequency below 1:10 000 (http://exac.broadinstitute.org, last accessed 22 January 2016). The p.Tyr273Cys variant was classified as potentially damaging by seven of nine common programs used to predict the functional impact of identified variants (Supplementary Materials), whereas the p.Glu438Gly variant was predicted to be deleterious by 4 of 9 programs (Supplementary Table 7).

To understand the possible consequences of the newly-identified variants on the SCNN1B protein structure, the molecular model of the human protein was inspected (Supplementary Figure 2). The SCNN1B protein is characterized by the presence of 2 transmembrane domains and a wide extracellular loop (including ~70% of the whole amino-acid sequence). This loop, together with the extracellular domains of the other 2 subunits, forms a funnel that directs ions from the lumen into the pore of ENaC.[29] The p.Glu438Gly mutation affects the extracellular domain of the protein, in a region well exposed to the solvent: the amino-acid substitution is predicted to change a negatively-charged residue with a small apolar one, thus possibly modifying the electrostatic signature of this region. Also the p.Tyr273Cys mutation involves the SCNN1B extracellular domain: the substitution falls within a beta strand, close to a highly-conserved Cys residue (Cys272), which is known to participate to the maintenance of the overall structure of the channel. Hence, the addition of a second Cys in this context could perturb the bond network normally characterizing the region (Supplementary Figure 2).

DISCUSSION

In the era of the P4 medicine ('personalized', 'predictive', 'preventive' and 'participatory'), the possibility to have a complete genetic diagnosis is the fundamental prerequisite for directing therapeutic strategies. CF conforms to this paradigm, especially in the light of the many efforts made in the last few years to develop new therapies targeting specific *CFTR* mutations.[30] At present, commercial kits for amplicon resequencing of *CFTR* exons using multiplex amplification, pooling of barcoded samples and NGS are available from different companies. These solutions promise to completely replace the traditional multi-step genetic testing of CF, reducing the costs of analysis and increasing the success rate of the diagnosis. Here, we explored the application of whole-gene *CFTR* sequencing to CF genetic testing by screening a number of 'difficult' cases, that is, patients with an incomplete diagnosis even after a preliminary extensive molecular screening. Our approach was successful both in identifying all mutations (including a large deletion) already detected by routine screening, and in ameliorating the percentage of genetically-diagnosed patients (Table 1).

Indeed, our data together with previous literature[19] suggest that whole-gene resequencing can in principle be used as a 'one-step strategy' to the molecular diagnosis of CF, for a number of reasons. First, this strategy has the advantage of not suffering from allele-dropout complications, because the library preparation procedure is based on probes for the capture step rather than on PCR amplification.[31,32] This advantage is worth not only in comparison with conventional Sanger sequencing, but also with amplicons-based NGS methods. Second, our strategy should allow the identification of large intragenic deletions, duplications, and rearrangements, which may go undetected by exon-only resequencing strategies. The inspection of the NGS read alignment should also allow an immediate definition of the deletion breakpoints at the nucleotide level, a process often laborious and time consuming. Third, thanks to the possibility to multiplex samples, both at the capture and at the sequencing level, we observed an increased time efficacy of whole-gene sequencing over Sanger sequencing. In addition, the foreseeable cost reductions coupled with the possibility to further multiplex samples (we reached a mean depth of >1150) promise to make this choice even more affordable over traditional approaches. Last, even in the unfortunate event of not being able to complete a genetic diagnosis, the availability of the sequence of all *CFTR* intronic regions still represents a great source of information. For instance, we identified a total of 22 novel deep-intronic variants, 18 of which at the moment remain without a functional meaning (Supplementary Table 5; Supplementary Figure 1). It is conceivable that in the future the accumulation of data on intronic sequences and the introduction of high-throughput functional analyses will assign a pathogenic meaning to some of these genetic variants, further increasing the diagnostic yield.

Even after this extensive sequencing strategy, five probands did not show any mutation in *CFTR*, whereas six were still lacking the second mutation. Missing mutations could either reside in the few intronic regions not covered by our NGS experiment, or be one of the already identified deep-intronic variants that we did not functionally characterize. Alternatively, considering that indels and structural variants still represent a problem in the alignment procedures, it is possible that they were not properly called by the adopted software. Finally, we cannot rule out the possibility that some patients, despite their CF or CF-like appearance, are indeed phenocopies with no mutations in their *CFTR* gene.[33] Accordingly, no mutated allele was identified in the five analyzed patients negative after conventional CFTR genetic screening. Further studies by whole-exome sequencing might help a better phenotype classification and might potentially reveal novel gene/s related to the CF phenotype.

Interestingly, it has been reported that patients with a CF-like disorder, but having mutations in only one copy of their *CFTR* genes, were carrier of gain-of-function mutations in the ENaC-encoding subunits.[15,34,35] More recently, Ramos *et al.*,[36] also suggested an oligogenic nature for CF-like phenotypes, with the description of patients having no mutations in *CFTR* but multiple variants in *SCNN1A*, *SCNN1B SCNN1G*, and even in the *SERPINA1* gene. The inclusion of the ENaC subunit genes in our NGS design was aimed at taking into account also this variable. However, we only found 2 missense variants in the *SCNN1B* gene, in one case in a patient with a complete molecular diagnosis of CF (CF-11), in the other in a patient with no pathogenic mutations in *CFTR* (CF-3). In both cases, the functional impact of the relevant variant still remains to be clarified.

RNA analysis, even though not routinely undertaken in diagnostic laboratories, has proven to be an important step in both the genetic diagnosis of CF and in the definition of the pathogenic potential of point mutations in *CFTR*. For instance, RT-PCR studies on RNA extracted from nasal epithelial cells or lymphocytes have been instrumental to clarify the role of deep-intronic variants affecting splicing.[7,37–39] We used this approach to complete the genetic diagnosis of a couple of patients resulted compound heterozygotes for a 'clear' disease-causing mutation and the c.1210-34$TG_{[11-12]}T_5$ alleles (CF-2 and CF-5; Table 1). The use of digital RT-PCR, which is an extremely accurate and sensitive method for the absolute quantification of transcripts, allowed us to verify that both patients have an overall amount of *CFTR* mRNA below 30%. This value should indeed be considered as an overestimate in the case of the CF-5 patient. In fact, a proper evaluation of the level of wild-type *CFTR* transcripts should take into consideration all possible splicing events as well as the

stability of transcripts bearing premature stop codons. In this frame, patient CF-5, who carries a splicing mutations affecting exon 16, is expected to have a lower level of the wild-type *CFTR* mRNA than that measured by digital RT-PCR, because the assay was designed for the specific amplification of the region encompassing exons 10–12. Hence, our results are in line with previous studies, describing a wide range of full-length *CFTR* transcripts (4–20%) associated with mild CF phenotypes,[39–41] and hence reinforce the notion that the T_5 allele can contribute to the CF phenotype in some patients.

In summary, we propose a new diagnostic protocol to be applied to the genetic diagnosis of CF: a one-shot PCR-based test followed by full-gene NGS screening (Supplementary Figure 3). The preliminary PCR-based screening is aimed at the identification of the most common CF mutation, that is, F508del, which can be easily detected by using either a simple and economic high-resolution melting analysis or other relatively low-cost genotyping methods (fluorogenic 5′ nuclease PCR assays or fluorescent fragment analysis on a capillary electrophoresis sequencer). Should this preliminary analysis be negative, the subsequent NGS step will allow the most comprehensive search for point mutations and gross alterations. In the light of the possible improvements in the programs used for mutation detection and in the hope of further cost drops, this strategy promises to be the most straightforward, convenient and rapid diagnostic protocol for CF patients.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Italian Cystic Fibrosis Foundation, grant numbers FFC#6/2011 and FFC#5/2015 (to s.d.), and by Italian fiscal contribution '5 × 1000' 2011 devolved to Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico. We wish to thank Carlo Castellani (Cystic Fibrosis Centre, Ospedale Civile Maggiore, Verona, Italy), Fabiola Corti and Amalia Negri (Division of Pediatrics, Livorno Hospital, Italy), and Rita Padoan (Cystic Fibrosis Centre, Spedali Civili, Brescia) for providing additional clinical data on some patients.

Author contributions: LS and GS designed the experiments and analyzed NGS data. LS performed gene expression analysis by digital PCR, functional validation of intronic variants, and mapping of deletion breakpoints. GS, RA and SD were responsible for data interpretation and drafting the manuscript. CC, MS and PM were involved in the design of the work (patient selection), data interpretation and critical revision of the manuscript. All authors read and approved the final version submitted for publication.

1 Southern, K. W., Munck, A., Pollitt, R., Travert, G., Zanolla, L., Dankert-Roelse, J. *et al.* A survey of newborn screening for cystic fibrosis in Europe. *J. Cyst. Fibros.* **6**, 57–65 (2007).
2 Knowles, M. R. & Durie, P. R. What is cystic fibrosis? *N. Engl. J. Med.* **347**, 439–442 (2002).
3 Chillón, M., Casals, T., Mercier, B., Bassas, L., Lissens, W., Silber, S. *et al.* Mutations in the cystic fibrosis gene in patients with congenital absence of the vas deferens. *N. Engl. J. Med.* **332**, 1475–1480 (1995).
4 Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073 (1989).
5 Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
6 Collins, F. S. Cystic fibrosis: molecular biology and therapeutic implications. *Science* **256**, 774–779 (1992).
7 Faà, V., Incani, F., Meloni, A., Corda, D., Masala, M., Baffico, A. M. *et al.* Characterization of a disease-associated mutation affecting a putative splicing regulatory element in intron 6b of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. *J. Biol. Chem.* **284**, 30024–30031 (2009).
8 Sosnay, P. R., Siklosi, K. R., Van Goor, F., Kaniecki, K., Yu, H., Sharma, N. *et al.* Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat. Genet.* **45**, 1160–1167 (2013).
9 Cuppens, H., Lin, W., Jaspers, M., Costes, B., Teng, H., Vankeerberghen, A. *et al.* Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes. The polymorphic (Tg)m locus explains the partial penetrance of the T5 polymorphism as a disease mutation. *J. Clin. Invest.* **101**, 487–496 (1998).
10 Pagani, F., Buratti, E., Stuani, C., Romano, M., Zuccato, E., Niksic, M. *et al.* Splicing factors induce cystic fibrosis transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element. *J. Biol. Chem.* **275**, 21041–21047 (2000).
11 Buratti, E., Brindisi, A., Pagani, F. & Baralle, F. E. Nuclear factor TDP-43 binds to the polymorphic TG repeats in CFTR intron 8 and causes skipping of exon 9: a functional link with disease penetrance. *Am. J. Hum. Genet.* **74**, 1322–1325 (2004).
12 Strong, T. V., Wilkinson, D. J., Mansoura, M. K., Devor, D. C., Henze, K., Yang, Y. *et al.* Expression of an abundant alternatively spliced form of the cystic fibrosis transmembrane conductance regulator (CFTR) gene is not associated with a cAMP-activated chloride conductance. *Hum. Mol. Genet.* **2**, 225–230 (1993).
13 Schwarz, M., Gardner, A., Jenkins, L., Norbury, G., Renwick, P. & Robinson, D. *Testing Guidelines for Molecular Diagnosis of Cystic Fibrosis* (Clinical Molecular Genetics Society, UK, 2009).
14 Ferec, C. & Cutting, G. R. Assessing the disease-liability of mutations in CFTR. *Cold Spring Harb. Perspect. Med.* **2**, a009480 (2012).
15 Sheridan, M. B., Fong, P., Groman, J. D., Conrad, C., Flume, P., Diaz, R. *et al.* Mutations in the beta-subunit of the epithelial Na+ channel in patients with a cystic fibrosis-like syndrome. *Hum. Mol. Genet.* **14**, 3493–3498 (2005).
16 Wine, J. J. The genesis of cystic fibrosis lung disease. *J. Clin. Invest.* **103**, 309–312 (1999).
17 Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M. *et al.* Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
18 Trujillano, D., Weiss, M. E. R., Köster, J., Papachristos, E. B., Werber, M., Kandaswamy, K. K. *et al.* Validation of a semiconductor next-generation sequencing assay for the clinical genetic screening of CFTR. *Mol. Genet. Genomic. Med.* **3**, 396–403 (2015).
19 Trujillano, D., Ramos, M. D., González, J., Tornador, C., Sotillo, F., Escaramis, G. *et al.* Next generation diagnostics of cystic fibrosis and CFTR-related disorders by targeted multiplex high-coverage resequencing of CFTR. *J. Med. Genet.* **50**, 455–462 (2013).
20 Baralle, M., Baralle, D., De Conti, L., Mattocks, C., Whittaker, J., Knezevich, A. *et al.* Identification of a mutation that perturbs NF1 gene splicing using genomic DNA samples and a minigene assay. *J. Med. Genet.* **40**, 220–222 (2003).
21 Rimoldi, V., Straniero, L., Asselta, R., Mauri, L., Manfredini, E., Penco, S. *et al.* Functional characterization of two novel splicing mutations in the OCA2 gene associated with oculocutaneous albinism type II. *Gene* **537**, 79–84 (2014).
22 Taulan, M., Girardet, A., Guittard, C., Altieri, J. P., Templin, C., Beroud, C. *et al.* Large genomic rearrangements in the CFTR gene contribute to CBAVD. *BMC Med. Genet.* **8**, 22 (2007).
23 Kerem, B. S., Zielenski, J., Markiewicz, D., Bozon, D., Gazit, E., Yahav, J. *et al.* Identification of mutations in regions corresponding to the two putative nucleotide (ATP)-binding folds of the cystic fibrosis gene. *Proc. Natl Acad. Sci. USA* **87**, 8447–8451 (1990).
24 Highsmith, Jr W. E., Burch, L. H., Zhou, Z., Olsen, J. C., Strong, T. V., Smith, T. *et al.* Identification of a splice site mutation (2789 +5 G>A) associated with small amounts of normal CFTR mRNA and mild cystic fibrosis. *Hum. Mutat.* **9**, 332–338 (1997).
25 Sharma, N., Sosnay, P. R., Ramalho, A. S., Douville, C., Franca, A., Gottschalk, L. B. *et al.* Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. *Hum. Mutat.* **35**, 1249–1259 (2014).
26 Vogelstein, B. & Kinzler, K. W. Digital PCR. *Proc. Natl. Acad. Sci. USA* **96**, 9236–9241 (1999).
27 Huggett, J. F. & Whale, A. Digital PCR as a novel technology and its potential implications for molecular diagnostics. *Clin. Chem.* **59**, 1691–1693 (2013).
28 Nuzzo, F., Paraboschi, E. M., Straniero, L., Pavlova, A., Duga, S. & Castoldi, E. Identification of a novel large deletion in a patient with severe factor V deficiency using an in-house F5 MLPA assay. *Haemophilia* **21**, 140–147 (2015).
29 Edelheit, O., Hanukoglu, I., Dascal, N. & Hanukoglu, A. Identification of the roles of conserved charged residues in the extracellular domain of an epithelial sodium channel (ENaC) subunit by alanine mutagenesis. *Am. J. Physiol. Renal. Physiol.* **300**, F887–F897 (2011).
30 Corvol, H., Thompson, K. E., Tabary, O., le Rouzic, P. & Guillot, L. Translating the genetics of cystic fibrosis to personalized medicine. *Transl. Res.* **168**, 40–49 (2016).
31 Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
32 Blais, J., Lavoie, S. B., Giroux, S., Bussières, J., Lindsay, C., Dionne, J. *et al.* Risk of misdiagnosis due to allele dropout and false-positive PCR artifacts in molecular diagnostics: analysis of 30,769 genotypes. *J. Mol. Diagn.* **17**, 505–514 (2015).
33 Tsui, L. C. & Dorfman, R. The cystic fibrosis gene: a molecular genetic perspective. *Cold Spring Harb. Perspect. Med* **3**, a009472 (2013).
34 Azad, A. K., Rauh, R., Vermeulen, F., Jaspers, M., Korbmacher, J., Boissier, B. *et al.* Mutations in the amiloride-sensitive epithelial sodium channel in patients with cystic fibrosis-like disease. *Hum. Mutat.* **30**, 1093–1103 (2009).

35 Fajac, I., Viel, M., Gaitch, N., Hubert, D. & Bienvenu, T. Combination of ENaC and CFTR mutations may predispose to cystic fibrosis-like disease. *Eur. Respir. J.* **34**, 772–773 (2009).
36 Ramos, M. D., Trujillano, D., Olivar, R., Sotillo, F., Ossowski, S., Manzanares, J. *et al.* Extensive sequence analysis of CFTR, SCNN1A, SCNN1B, SCNN1G and SERPINA1 suggests an oligogenic basis for cystic fibrosis-like phenotypes. *Clin. Genet.* **86**, 91–95 (2014).
37 Chillón, M., Dörk, T., Casals, T., Giménez, J., Fonknechten, N., Will, K. *et al.* A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+1.6kbA–>G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. *Am. J. Hum. Genet.* **56**, 623–629 (1995).
38 Costantino, L., Claut, L., Paracchini, V., Coviello, D. A., Colombo, C., Porcaro, L. *et al.* A novel donor splice site characterized by CFTR mRNA analysis induces a new pseudo-exon in CF patients. *J. Cyst. Fibros.* **9**, 411–418 (2010).
39 Costantino, L., Rusconi, D., Soldà, G., Seia, M., Paracchini, V., Porcaro, L. *et al.* Fine characterization of the recurrent c.1584+18672 A>G deep-intronic mutation in the cystic fibrosis transmembrane conductance regulator gene. *Am. J. Respir. Cell. Mol. Biol.* **48**, 619–625 (2013).
40 Chu, C. S., Trapnell, B. C., Curristin, S., Cutting, G. R. & Crystal, R. G. Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nat. Genet.* **3**, 151–156 (1993).
41 Rave-Harel, N., Kerem, E., Nissim-Rafinia, M., Madjar, I., Goshen, R., Augarten, A. *et al.* The molecular basis of partial penetrance of splicing mutations in cystic fibrosis. *Am. J. Hum. Genet.* **60**, 87–94 (1997).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)