

Lysoplex: An efficient toolkit to detect DNA sequence variations in the autophagy-lysosomal pathway

Giuseppina Di Fruscio,1,2, Angela Schulz,3,7, Rossella De Cegli,1 Marco Savarese,2 Margherita Mutarelli,1 Giancarlo Parenti,1,4 Sandro Banfi,1,2 Thomas Braulke,3 Vincenzo Nigro,1,2 and Andrea Ballabio1,4,5,6,*

1Telethon Institute of Genetics and Medicine (TIGEM); Naples, Italy; 2Dipartimento di Biochimica; Biosfica e Patologia Generale; Seconda Università degli Studi di Napoli; 3University Medical Center Hamburg-Eppendorf, Dept. Biochemistry; Children’s Hospital; Hamburg, Germany; 4Medical Genetics; Department of Translational Medicine; Federico II University; Naples, Italy; 5Department of Molecular and Human Genetics; Baylor College of Medicine; Houston, TX USA; 6Jan and Dan Duncan Neurological Research Institute; Texas Children’s Hospital; Houston, TX USA

*These authors equally contributed to this work.

Keywords: autophagy, genetic variants, lysosomal storage disorders, neuronal ceroid lipofuscinoses, next-generation sequencing

Abbreviations: ALP, autophagy-lysosomal pathway; LSDs, lysosomal storage disorders; NGS, next-generation sequencing; NCL, neuronal ceroid lipofuscinosis; WES, whole exome sequencing; WGS, whole genome sequencing.

The autophagy-lysosomal pathway (ALP) regulates cell homeostasis and plays a crucial role in human diseases, such as lysosomal storage disorders (LSDs) and common neurodegenerative diseases. Therefore, the identification of DNA sequence variations in genes involved in this pathway and their association with human diseases would have a significant impact on health. To this aim, we developed Lysoplex, a targeted next-generation sequencing (NGS) approach, which allowed us to obtain a uniform and accurate coding sequence coverage of a comprehensive set of 891 genes involved in lysosomal, endocytic, and autophagic pathways. Lysoplex was successfully validated on 14 different types of LSDs and then used to analyze 48 mutation-unknown patients with a clinical phenotype of neuronal ceroid lipofuscinosis (NCL), a genetically heterogeneous subtype of LSD. Lysoplex allowed us to identify pathogenic mutations in 67% of patients, most of whom had been unsuccessfully analyzed by several sequencing approaches. In addition, in 3 patients, we found potential disease-causing variants in novel NCL candidate genes. We then compared the variant detection power of Lysoplex with data derived from public whole exome sequencing (WES) efforts. On average, a 50% higher number of validated amino acid changes and truncating variations per gene were identified. Overall, we identified 61 truncating sequence variations and 488 missense variations with a high probability to cause loss of function in a total of 316 genes. Interestingly, some loss-of-function variations of genes involved in the ALP pathway were found in homozygosity in the normal population, suggesting that their role is not essential. Thus, Lysoplex provided a comprehensive catalog of sequence variants in ALP genes and allows the assessment of their relevance in cell biology as well as their contribution to human disease.

Introduction

The lysosome plays a central role in cellular clearance and energy metabolism by ensuring degradation and recycling of a variety of substrates. While extracellular substrates reach the lysosome via the endocytic and phagocytic pathways, intracellular material follows the autophagic pathway, mainly through the fusion of autophagosomes with lysosomes.1–4 In addition, recent studies have revealed that the lysosome is also involved in signaling pathways that play a crucial role in cell homeostasis and growth.1

The physiological importance of lysosome-mediated pathways is underlined by their involvement in human diseases. Lysosomal storage diseases are due to mutations in more than 50 genes encoding lysosomal soluble acidic hydrolases, integral membrane proteins, activators, transporters, or nonlysosomal proteins that are necessary for lysosomal function.5–7 The precise molecular diagnosis is very important as new therapies are becoming available for specific types of LSDs.8 Unfortunately, for the majority of these disorders, the diagnosis is difficult due to considerable clinical overlap and clinical variability. This is particularly true...
for neuronal ceroid lipofuscinosis, a genetically heterogeneous subtype of LSDs with a devastating phenotype that can be caused by mutations of at least 13 different genes. Mutations in the autophagy-lysosomal (ALP) pathway can also cause common neurodegenerative disorders, such as Parkinson and Alzheimer diseases. It is likely that mutations in the ALP pathway will be identified in an increasing number of diseases.

Next-generation sequencing is revolutionizing our view of life sciences by allowing the investigation of variations in DNA/RNA sequences at a genome scale. In the last few years, whole exome sequencing or whole genome sequencing (WGS) approaches have received unlimited consideration as universal tests for most Mendelian conditions, with the exceptions of those caused by complex structural variations. As a cheaper and more focused alternative to WGS, WES has been demonstrated to be an excellent and cost-effective solution with a consistent number of advantages and, above all, a higher throughput on coding sequences that favors the identification of novel disease genes. However, accumulating data indicate that WES may not be a totally reliable tool to assess sequence variations in a given gene. Efficiency of WES is undermined by nonuniform targeting and coverage that, for a significant fraction of the exome, is too low to allow thorough variant detection. In addition, false positive calls are frequently observed in segmental duplications and processed pseudogenes. This reduces WES sensitivity as well as specificity, leading to labor-intensive validation steps usually based on conventional PCR reactions and Sanger sequencing. For practical reasons, the validation steps are usually restricted to specific genes, missing the remaining variability of the exome. Finally, there is an increasing need to properly manage incidental findings detected by WES/WGS.

NGS-based targeted approaches, i.e. sequencing of selected genes and genomic regions of interest, have been developed as an alternative to WES/WGS. Using targeted approaches the ethical issues related to incidental findings are very limited. Our aim was to develop a robust sequencing platform to identify the sequence variations in the ALP pathway. A similar approach was used to detect sequence variations in mitochondrial pathways.

A potential obstacle to building a panel of bona fide ALP genes is represented by the fact that the identity of many key players is still unknown. Therefore, we used multiple bioinformatic tools to identify a comprehensive collection of ALP genes. The final, nonredundant, ALP list contains 891 genes, including 106 autophagy genes, 194 lysosomal genes, and 627 genes with a role in the endocytic pathway (Fig. 1).

Approximately 140 out of the total 891 genes are mutated in human Mendelian diseases. We used the HaloPlex enrichment protocol, which permitted accurate sequencing and high sequence coverage. This workflow was named Lysoplex.

**Results**

**Generation of the ALP gene list**

We selected the following sets of genes to generate the Lysoplex list: 1) a subset of genes encoding proteins with either an established role in autophagy or predicted by at least 2 of the following bioinformatics tools: AmiGO, Netview, UniProt, and MSigDB (see also Materials and Methods and Table S1); 2) a set of genes encoding proteins with lysosomal localization selected using a similar procedure involving both literature search and bioinformatic predictions; 3) all genes that are mutated in LSDs, including those coding for nonlysosomal proteins and the transcription factor TFEB, a master regulator of lysosomal biogenesis and autophagy, as well as 3 closely related paralogs (Mit/TFE members); 4) genes with a known/predicted role in the endocytic pathway; 5) genes involved in familial forms of Parkinson and Alzheimer diseases. As a result, the ALP list includes 891 genes (Table S1), for which all predicted exons and at least 10 nucleotides of flanking intronic sequences were included. The target region collectively spans 2.054 megabases of human DNA coding sequence (Fig. 1).

**Lysoplex design and validation with mutation-known samples**

To obtain a homogeneous and accurate coverage of the 891 ALP genes selected as indicated above, we designed capture oligonucleotide probes using the HaloPlex Target Enrichment System, which represents an excellent technology for complex targeting with high sensitivity and specificity. This procedure is based on restriction enzyme digestion of genomic DNA with multiple endonucleases and hybridization capture with probes that also work in the subsequent amplification step. We used this platform to sequence ~100 DNA samples. On average, sequence reads that mapped to the ALP gene target represented more than 92% of total reads (Fig. 2A), showing efficient and specific probe capture. This minimal “off-targeting” reduces the sequencing demand in comparison with other enrichment protocols and increases on-target coverage. Using a high-throughput setting of 32 libraries/lane, the average coverage depth was 40x for 95% and 100x for 80% of the ALP-target, which is significantly higher than the average depth we reached on the same genes with WES (Fig. 2B–D). This setting led to a high specificity in variant calls.
mutations in the Patient L44 turned out to be compound heterozygous for 2

diagnosis had not been previously made in the 2 patients in whom molecular
to identify the disease-causing mutations previously known mutations and to
procedure to the training set of 16 LSD
matic leukodystrophy (MLD), Pompe
disease, and Danon disease.

The application of the Lysoplex procedure to the training set of 16 LSD patients allowed us to detect all of the
identified the disease-causing mutations in the 2 patients in whom molecular
diagnosis had not been previously made (Table S2).

**Lysoplex application to molecular diagnosis of NCLs**

To test the efficacy of Lysoplex in the analysis of patients with
unknown/elusive molecular defects, we applied it to the search
for disease-causing mutations in patients with clinically diag-
osed NCLs, which are caused by mutations in at least 13 different
genes. We carried out Lysoplex on 48 NCL patients,
whose gene defect had not been previously detected (Table S3).

**Figure 3** describes in “exploded pies” the success rate of
Lysoplex to identify causative variants in the NCL patients. We
detected both pathogenic variants (homozygous or compound
heterozygous) in 29/48 cases. In 3 additional cases (Table 1; L33, L57 and L58), we only found a single heterozygous muta-
tion in known NCL genes (CLN3, CLN6), which were compatible
with the observed phenotype. We postulate that these
patients may harbor a second mutation in the same gene that is
not detectable, e.g., a deep intronic mutation or a large copy
number variation.

In total, pathogenic mutations were identified in 32 NCL patients in the following disease genes: CLN1 (n = 2), CLN2 (n = 3), CLN3 (n = 1), CLN5 (n = 1), CLN6 (n = 10), CLN7 (n = 8), CLN8 (n = 5), CLN11 (n = 1) and GLB1 (n = 1). Patient L44 turned out to be compound heterozygous for 2
mutations in the GLB1 gene, encoding β-galactosidase. Interest-
ingly, mutations in GLB1 are responsible for GM1-gangliosido-
sis, a neurodegenerative type of LSD whose phenotype partially
overlaps with that of CLN1. Overall, in all patients analyzed, the
molecular diagnosis was in agreement with the clinical diagnosis,
taking into account the phenotypic variability observed in NCLs. This is further exemplified by the fact that mutations in the
CLN6 gene can cause both late infantile as well as adult NCL
phenotypes (Table 1).

Of note, patient L42 had been previously proposed to repre-
represent a novel CLN locus, namely CLN9, based on a distinctive
fibroblast phenotype, gene expression pattern, ultrastructural
analysis of storage material, enzymatic activities, and sequencing
of NCL genes. However, we found that this patient carries a
homozygous CLN2 mutation.

Taken as a whole, the use of Lysoplex allowed us to identify
causative mutations in 67% of the NCL patients (Table 1) Eight-
teen patients had never received any test (“naïve” group), whereas
the remaining 30 individuals had been tested for some NCL
genes (“pre-screened”) but without positive findings. Interest-
ingly, one pre-screened family had previously been studied by
both linkage analysis and WES (Table 1; patients L37 and L38).
There was, however, no substantial difference in the success rate
between the pre-tested and the naïve group.

Interestingly, in addition to the causative mutations in a given
disease gene, we also identified in some patients additional muta-
tions in genes involved in other diseases (Table S4). This may
add further complexity to the genotype/phenotype correlation.

We then looked for recessive mutations in novel NCL candidate
genes, which could exert a pathogenic role in the 16 patients
in whom Lysoplex could not identify any mutations in known
disease genes. We focused our attention on loss-of-function var-
ts such as frame-shift, nonsense, splice defects, and potentially
damaging nonsynonymous amino acid substitutions in nondisease genes. We considered as novel candidate NCL genes those that showed these variants on both alleles in at least one NCL patient or those mutated in 2 unrelated patients. Following these selection criteria, we identified 3 potential candidate genes: \textit{STAB2} (stabilin 2), \textit{AGAP1} (ArfGAP with GTPase domain, ankyrin repeat and PH domain 1) and \textit{PLCG2} (phospholipase C, gamma 2 [phosphatidylinositol-specific]) (Table S5). How-

ever, functional analyses and in vivo studies in animal models are needed before we can conclude that these are causative mutations. Overall, we identified 37 different pathogenic mutations, including 21 novel mutations. They include 25 amino acid changes, 5 indels (insertions or deletions) including an insertion of 17 nucleotides, 5 nonsense mutations, and 2 splice site variants (Table 1).

\textbf{Sequence variation in ALP genes}

To assess the efficiency of Lysoplex in detecting sequence variations in ALP genes, we compared its performance to pre-

missense variations with a high probability to exert a damaging effect on the corresponding protein product. The above mutations were found in a total of 316 different genes (Table S6). Interestingly, some of the loss-of-function variations were also found in the normal population at a frequency higher than 1% (Table 2). Some of the genes (n = 6) carrying common loss-of-function variations in the normal population are involved in autophagy (\textit{ATG3}), endocytosis (\textit{ATP6V1G3}, \textit{CBLC}, \textit{SEC14L4}) and membrane trafficking (\textit{INPP5D}). These findings suggest that total loss-of-function in these genes does not appear to have major effects.

\section*{Discussion}

In the past few years, remarkable progress has been made in the characterization of the autophagic lysosomal pathway, discovering wide-ranging pathogenic roles in addition to vital homeostatic functions.\textsuperscript{1,49-51} Autophagy is involved in various existing public resources, primarily the Exome Variant Server (EVS). To avoid any bias due to sample selection, we first counted the number of synonymous variations/allele/kb in our patient cohort as compared with EVS data. We plotted each gene as a dot in a logarithmic scale where the X-axis represents the number of sequence variations/allele/kb in EVS and the Y-axis represents the same value obtained in the Lysoplex panel (Fig. 4A). This analysis revealed that the vast majority of genes show a much higher number of variations in the Lysoplex samples vs. EVS. These data were also confirmed for all other categories of sequence substitutions, including missense and loss-of-function mutations (Fig. 4B, C). This demonstrates that Lysoplex is much more sensitive than EVS in detecting sequence variability in ALP genes. We did not consider small insertions or deletions that grossly deviate from Hardy-Weinberg equilibrium, since it is well known that a high number of false positive variants are erroneously annotated in EVS and in all other databases.

Overall, we found 61 clear loss-of-function mutations (i.e., nonsense, indels leading to frameshifts and splice variants) and 488

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Success rate of Lysoplex in NCL patients. Schematic diagram summarizing the results of the Lysoplex procedure in the NCL patients analyzed. An overall detection rate of 67% was observed. In particular, mutations were identified in 8 NCL known genes. In one case, we found pathogenic mutations in the \textit{GLB1} gene (see text for details).}
\end{figure}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
\textbf{Gene} & \textbf{Frequency} & \textbf{Mutation Type} & \textbf{Allele/kb} & \textbf{Mutation Description} \\
\hline
\textit{STAB2} & 0.02 & Missense & 1 & S200D \\
\textit{AGAP1} & 0.01 & Nonsense & 1 & Q47X \\
\textit{PLCG2} & 0.005 & Splice site & 1 & c.237+5G>A \\
\hline
\end{tabular}
\caption{Summary of novel missense mutations identified in NCL patients.}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\textbf{Gene} & \textbf{Frequency} & \textbf{Mutation Type} & \textbf{Allele/kb} & \textbf{Mutation Description} \\
\hline
\textit{ATG3} & 0.001 & Frameshift & 1 & A157X \\
\textit{ATP6V1G3} & 0.003 & Indel & 1 & 5bp insertion \\
\textit{CBLC} & 0.002 & Nonsense & 1 & R47X \\
\textit{SEC14L4} & 0.004 & Missense & 1 & A28N \\
\textit{INPP5D} & 0.005 & Splice site & 1 & c.56+2T>G \\
\hline
\end{tabular}
\caption{Summary of novel loss-of-function mutations identified in NCL patients.}
\end{table}
| Patient ID | Nationality | Age of onset | First symptoms | Genomic coordinates allele 1 (hg19) | Genomic coordinates allele 2 (hg19) | Allele 1 | Allele 2 |
|------------|-------------|--------------|----------------|-----------------------------------|-----------------------------------|---------|---------|
| L137       | Indian      | late infantile | Epilepsy       | chr1:40558087                     | chr1:40558087                     | c.217G->T; p.Gly73Trp<sup>Y</sup> | c.217G->T; p.Gly73Trp<sup>Y</sup> |
| L144       | Indian      | infantile    | Epilepsy       | chr1:40546149                     | chr1:40546149                     | c.547G->T; p.Ala183Ser<sup>Y</sup> | c.547G->T; p.Ala183Ser<sup>Y</sup> |
| L42        | German      | juvenile     | Visual loss    | chr11:6638271                     | chr11:6638271                     | c.622C->T; p.Arg208X               | c.622C->T; p.Arg208X               |
| L150       | Lybian      | late infantile | Ataxia, epilepsy | chr11:6637287                     | chr11:6637287                     | c.1094G->A; p.Cys365Tyr            | c.1094G->A; p.Cys365Tyr            |
| L133       | Indian      | late infantile | Ataxia, epilepsy | chr11:6638858                     | chr11:6638858                     | c.379C->T; p.Arg127X               | c.379C->T; p.Arg127X               |
| L33        | German      | juvenile     | Dementia       | chr16:28489104                    | chr16:28489104                    | c.115T>C; p.Leu384Pro<sup>Y</sup> | n.f.                              |
| L128       | Egyptian    | juvenile     | Ataxia, epilepsy | chr1:1173,1174delAT; p.Tyr392X<sup>Y</sup> | chr1:1173,1174delAT; p.Tyr392X<sup>Y</sup> | c.1173,1174delAT; p.Tyr392X<sup>Y</sup> | c.1173,1174delAT; p.Tyr392X<sup>Y</sup> |
| L21        | Swiss       | late infantile | Ataxia         | chr15:68510922                    | chr15:68510922                    | c.150C->G; p.Tyr50X<sup>Y</sup>    | c.150C->G; p.Tyr50X<sup>Y</sup>    |
| L24        | Turkish     | early infantile | Dementia, epilepsy | chr15:68504188                    | chr15:68504188                    | c.311C->T; p.Ser104Phe             | c.311C->T; p.Ser104Phe             |
| L30        | German      | juvenile     | Dementia, epilepsy | chr15:68504126                    | chr15:68504126                    | c.373A->G; p.Ser125Gly<sup>Y</sup> | c.373A->G; p.Ser125Gly<sup>Y</sup> |
| L31        | German      | early infantile | Dementia, epilepsy | chr15:68501981                    | chr15:68501981                    | c.198+5C>T; p.Tyr66Phe<sup>Y</sup> | c.198+5C>T; p.Tyr66Phe<sup>Y</sup> |
| L39        | Turkish     | late infantile | Ataxia, epilepsy | chr15:68500518                    | chr15:68500518                    | c.586C->T; p.Pro299Leu             | c.586C->T; p.Pro299Leu             |
| L50        | German      | late infantile | Ataxia, epilepsy | chr15:68521918                    | chr15:68521918                    | c.5A>><G; p.Glu2Gly<sup>Y</sup> | c.5A>><G; p.Glu2Gly<sup>Y</sup> |
| L57        | Italian     | adult        | Visual loss, dementia | chr15:6850086                      | chr15:6850086                      | c.728C->T; p.Ala243Val<sup>Y</sup> | n.f.                              |
| L58        | Italian     | adult        | Visual loss, dementia | chr15:6850068                      | chr15:6850068                      | c.728C->T; p.Ala243Val<sup>Y</sup> | n.f.                              |
| L129       | German      | late infantile | Ataxia, epilepsy | chr15:6850058                      | chr15:6850058                      | c.502C->G; p.Leu168Val<sup>Y</sup> | c.502C->G; p.Leu168Val<sup>Y</sup> |
| L130       | Indian      | late infantile | Ataxia         | chr15:68506628                    | chr15:68506628                    | c.297G->T; p.Lys99Asn<sup>Y</sup> | c.297G->T; p.Lys99Asn<sup>Y</sup> |
| L20        | German      | juvenile     | Ataxia         | chr14:28849193                    | chr14:28849193                    | c.1439G->T; p.Gly480Val<sup>Y</sup> | c.1439G->T; p.Gly480Val<sup>Y</sup> |
| L22        | Swiss       | juvenile     | Ataxia         | chr14:28851955                    | chr14:28851955                    | c.881C->A; p.Tyr294Lys             | c.881C->A; p.Tyr294Lys             |
| L23        | Tamilian    | infantile    | Ataxia         | chr14:28849148                    | chr14:28849148                    | c.1394G->A; p.Arg465Gln            | c.1394G->A; p.Arg465Gln            |
| L40        | Turkish     | late infantile | Ataxia, epilepsy | chr14:28842794                    | chr14:28842794                    | c.1235C>T; p.Pro141Leu             | c.1235C>T; p.Pro141Leu             |
| L49        | German      | late infantile | Ataxia, epilepsy | chr14:28842927                    | chr14:28842927                    | dup17<sup>Y</sup>                 | dup17<sup>Y</sup>                 |
| L51        | Indian      | late infantile | Ataxia, epilepsy | chr14:28841922                    | chr14:28841922                    | c.1420C->T; p.Gln474X<sup>Y</sup> | c.1420C->T; p.Gln474X<sup>Y</sup> |
| L56        | Turkey      | late infantile | Ataxia, epilepsy | chr14:28841948                    | chr14:28841948                    | c.1394G->A; p.Arg465Gln<sup>Y</sup> | c.1394G->A; p.Arg465Gln<sup>Y</sup> |
| L140       | Indian      | late infantile | Ataxia, epilepsy | chr14:28863253                    | chr14:28863253                    | c.499delA; p.Ser167Val<sup>Y</sup> | c.499delA; p.Ser167Val<sup>Y</sup> |
| L37        | Turkish     | late infantile | Ataxia, epilepsy | chr8:1728661                      | chr8:1728661                      | c.789G->C; p.Trp263Cys             | c.789G->C; p.Trp263Cys             |
| L38        | Turkish     | late infantile | Ataxia, epilepsy | chr8:1728661                      | chr8:1728661                      | c.789G->C; p.Trp263Cys             | c.789G->C; p.Trp263Cys             |
| L60        | German      | late infantile | Ataxia, epilepsy | chr8:1728469                      | chr8:1728469                      | c.599delAT; p.Met200Val<sup>Y</sup> | c.599delAT; p.Met200Val<sup>Y</sup> |
| L139       | Indian      | late infantile | Ataxia, epilepsy | chr8:1719693                      | chr8:1719693                      | c.473A->G; p.Tyr158Cys             | c.473A->G; p.Tyr158Cys             |
| L141       | Dutch       | juvenile     | Dementia, epilepsy | chr8:1719719                      | chr8:1719719                      | c.595G->C; p.His199A<sup>Y</sup> | c.595G->C; p.His199A<sup>Y</sup> |
| L36        | Turkish     | late infantile | Ataxia, epilepsy | chr7:142427605                    | chr7:142427605                    | c.359C->A; p.Ser120Tyr<sup>Y</sup> | c.359C->A; p.Ser120Tyr<sup>Y</sup> |

*reported consanguinity  
*novel mutations  
n.f. = not found
conditions such as genetic diseases,\textsuperscript{52} metabolic disorders,\textsuperscript{53,54} neurodegenerative diseases,\textsuperscript{55} infections,\textsuperscript{56} and cancer.\textsuperscript{57,58}

While great attention has been given to the study of the mechanisms underlying the regulation and execution of autophagy in health and disease, the intrinsic genetic variation of the system in the human population has so far been ignored. Lysoplex was developed to map such genetic variation. When compared with public genome variant databases, the Lysoplex panel provided a significantly more accurate list of genetic variants that may affect the ALP pathways, in normal individuals and in patients.

When applied to the molecular diagnosis of LSDs, Lysoplex revealed a number of significant advantages over WES or WGS approaches. For example, the data load was approximately 1/10 and 1/100 compared to WES and WGS, respectively. This reduces costs and efforts, making Lysoplex applicable to a large number of samples. In addition, Lysoplex proved to be much more robust in terms of target reproducibility, specificity, and sensitivity, and its depth was such that there was no need for Sanger validation steps (i.e. to confirm variations or distinguish between homozygous and heterozygous status). The extent of off-targeting was minimal (8%) although slightly higher than that obtained with Motorplex (2%).\textsuperscript{23} However, Motorplex covers a target region that is 4-fold smaller than Lysoplex. It is also important to underline that Lysoplex is a very versatile tool as the list of genes to be analyzed can be easily updated as new relevant genes are identified.

In our study we detected all disease-causing mutations in a set of LSD patients. Lysoplex was successful in identifying the molecular defect in 67% of the NCL patients analyzed, which is a significant result when considering both the locus and allele heterogeneity of this class of diseases. Remarkably, patient L42 belongs to one of the 2 pairs of NCL siblings of German and Serbian descent, respectively, who were previously proposed to be affected by mutations in an undefined CLN9 locus.\textsuperscript{48} Lysoplex analysis allowed us to reclassify L42 as CLN2. Since the other sib pair (of Serbian origin) was reclassified earlier after homozygosity mapping as CLN5,\textsuperscript{59} no further evidence for an elusive CLN9 gene exists. Two others of the pre-tested patients, L37 and L38 (Table 1), had been previously analyzed by WES, but the pathogenic mutations were originally missed due to unclear family history (presence of one additional sibling with an uncertain diagnosis). Furthermore, in one of the patients originally diagnosed as affected by NCL, we identified missense mutations in both alleles of the GLBI gene, resulting in the molecular diagnosis of GM1 gangliosidosis. This disease is associated with accumulation of saposin and subunit c of the V-ATPase, and is characterized by common manifestation of ER and oxidative stress\textsuperscript{60–62} with a resulting phenotype that overlaps that observed in NCL. This suggests that some cases with a clinical diagnosis of NCL may indeed have GM1-gangliosidosis. There is another report in which a patient diagnosed with adult NCL was found to carry heterozygous mutations in SGSH defective in MPS IIIA (Sanfilippo syndrome A) and associated with lysosomal accumulation of heparin sulfate.\textsuperscript{63} Thus, Lysoplex appears to be instrumental for the differential diagnosis and correct classification between different types of LSDs. This result emphasizes the value of testing multiple genes for the same broad class of genetic disease (e.g., LSDs) with partially overlapping phenotypes.

Considering that ALP genes may be candidates for other genetic conditions, and that approximately 85% of them are not yet assigned to any disease, Lysoplex can effectively be used for the discovery of novel disease genes. Lysoplex application to a relatively small set of patients resulted in the identification of several homozygous or compound heterozygous mutations in putative novel disease genes for NCLs. These are: STAB2, that encodes a cell receptor that enables the scavenging and clearance of multiple ligands from the circulation salvaging their building block monomers by lysosomal degradation, as well as MAPK1/ERK2-MAPK3/ERK1 activated signaling,\textsuperscript{64,65} AGAP1 that encodes an endosome-associated, phosphoinositide-dependent ADP-ribosylation factor GTPase-activating protein,\textsuperscript{66} and PLCG2 that encodes a transmembrane signaling enzyme whose gain-of-function mutations have been associated with autoimmune and inflammatory diseases mediated by B cells.\textsuperscript{67} More data, such as the identification of additional families

### Table 2. List of common loss-of-function sequence variations (frequency > 1%) in ALP genes

| Gene symbol and RefSeq | Gene name | Biological function and biological process | Genomic coordinates (hg19) | Sequence variation | N Het samples | N Hom samples | Freq in EVS |
|------------------------|-----------|-------------------------------------------|---------------------------|-------------------|---------------|---------------|------------|
| ATG3 (NM_001278712)    | autophagy related 3 | Autophagic vacuole assembly | chr3:112253058 | c.921_922insT; p.Leu307fs | 9 | 2 | 0.601022 |
| ATP6V1G3 (NM_133326)   | ATPase, H\textsuperscript{+} transporting, lysosomal 13kDa, V1 subunit G3 | Transmembrane transport and endocytosis | chr1:198505831 | c.106C>T; p.Arg36X | 2 | 0 | 0.019689 |
| CBLC (NM_001130852)   | Cbl proto-oncogene C, E3 ubiquitin protein ligase | Cell surface receptor signaling pathway and endocytosis | chr19:45296846 | c.1115_1116insC; p.Asp372fs | 4 | 0 | 0.064297 |
| INPP5D (NM_001017915) | inositol polyphosphate-5-phosphatase, 145kDa | Membrane trafficking | chr2:234066969 | c.956C>G; p.Ser319X | 28 | 20 | 0.569864 |
| LPL (NM_000237)       | lipoprotein lipase | Protein binding | chr8:19819724 | c.1421C>G; p.Ser474X | 10 | 1 | 0.086345 |
| SEC14L4 (NM_001161368) | SEC14-like 4 (S. cerevisiae) | Golgi-to-plasma membrane transport | chr22:30891264 | c.400G>T; p.Glu134X | 1 | 0 | 0.032062 |
and/or functional assays, are needed to definitely prove the involvement of the above candidate genes in NCL pathogenesis.

Another application of Lysoplex may be in the field of modifier genes, complex disorders, and polygenic inheritance. It is well known that patients that share the same mutations may have different phenotypic spectrum. Thus, the effect of the primary molecular defects may be modified by the presence of additional and variable elements located in other genes that encode proteins involved in the same pathways. Remarkably, we have identified additional mutations in other disease genes in patients harboring pathogenic mutations in NCL genes (Table S4). This may represent disease modifiers that explain intra-familial phenotypic variability. However, to test their role as modifier genes, genotype-phenotype correlation studies are needed on a much larger number of cases.

There is increasing evidence that mutations in genes involved in the ALP pathway may predispose to common neurodegenerative diseases. Cohorts of patients affected by common neurodegenerative diseases, such as Parkinson and Alzheimer diseases, may be screened using Lysoplex to identify additional genes predisposing to such diseases. It may also be interesting to use Lysoplex for the screening of somatic mutations in ALP genes in tumor samples due to the known involvement of autophagy in cancer.

Figure 4. Comparison between the number of sequence variations found by Lysoplex and WES. Plotted data indicate the log of the number of variations/kb found in each gene by Lysoplex (Y-axis) and by public database (X-axis), if the sensitivity was identical, dots should tend to align. Dots above the line indicate a superior sensitivity of Lysoplex. Red dots indicate NCL genes and blue dots other disease genes. (A) synonymous variations; (B) missense variations; (C) loss-of-function variations.
Finally, Lysoplex is an exceptionally efficient tool to map amino acid sequence variations in proteins encoded by ALP genes. Such amino acid sequence variations may have functional relevance. Indeed, we have observed a high frequency of rare, nonconserved, amino acid changes as well as truncating variants. The observation of such protein variation may not be a total surprise, considering the predicted human genome variability, but it is remarkable considering the crucial roles of ALP genes in cell homeostasis. Much of this variability is overlooked when performing global WES or WGS approaches. In practice, WES or WGS data are usually filtered and the discordant variability is not dissected gene by gene according to function. Furthermore, when the coverage is low, the variability is always submerged by numerous sequencing errors and gaps that, in contrast, do not affect Lysoplex.

This catalog of functional variants may serve 2 main purposes. It provides a useful tool for basic researchers to study the biological relevance of protein variants in the ALP pathway. At the same time, clinical researchers can use this catalog to identify candidate causative or modifier genes for human diseases.

## Materials and Methods

### Patients

The training set of 16 genomic DNA samples belonged to patients affected by 14 different LSDs (Table S2). In all cases, the enzyme activities were measured in cultured fibroblasts, using artificial fluorogenic substrates, according to standard procedures. NCL patients received a diagnosis based on clinical evaluation, biochemical assays, and electron microscopy analysis. Some of the patients had already undergone some molecular tests (Table 1). All procedures on patients were approved by the Ethics Boards of the participating institutions and adhered to the tenets of the Declaration of Helsinki.

### Selection of Lysoplex genes

The complete gene list includes 891 members (Table S1): 98 previously described lysosomal genes, 29,9 genes colocalizing with the lysosomal marker LAMP2,48 the 4 MIT family members;41 all the known lysosomal and nonlysosomal protein-coding genes with a role in the different LSDs46,69 selected by using the Online Mendelian Inheritance in Man (OMIM).70,71

Lysoplex also includes 10 genes involved in hereditary forms of Parkinson disease and 4 involved in hereditary forms of Alzheimer disease. Moreover, bioinformatics prediction tools, including AmiGO,24 Netview,25 and UniProt26 were used. For lysosomal genes, in each database the complete human gene list was downloaded restricting the output to the “lysosome” term: the AmiGO list included 306 genes, the Netview list included 484 genes and the UniProt list included 305 genes. The 3 lists were compared in pairs: 83 genes predicted to be lysosomal by at least 2 out of 3 tools were considered good candidate genes and thus added to the final list. In Table S1 this subset of genes is listed and associated with the specific bioinformatics tools’ pair: 16 genes were shared by AmiGO and Netview lists, 63 genes were shared by AmiGO and UniProt lists, one gene was shared by Netview and UniProt lists and 3 genes were shared by all 3 lists.

As for the lysosomal genes, the selection of the 106 genes mainly playing a role in the autophagic pathway was performed by both literature analysis33,34 and bioinformatics tools. In detail, 73 genes were mapped to 12 autophagic subcategories on the GO hierarchy rooted in the “Autophagy” (GO:0006914) accession term: 2 more genes (SNX3 and CHPT1) were introduced because they were predicted as “autophagic vacuole” genes by Netview.

For the selection of genes with a known role in the endocytic pathway and trafficking, both literature analysis42,72–76 (n = 325) and the Molecular Signatures Database (MSigDB)27,77 were used. MSigDB, which is one of the most widely used repositories of well-annotated gene sets, was used to dissect lists of genes included in different “endocytosis” gene sets: we thus found 53 genes annotated by the AmiGO term (GO:0006897) “endocytosis”; 14 genes annotated by the GO term (GO:0030139) “endocytic vesicle”; 23 genes collected in the GO term (GO:0006898) “receptor mediated endocytosis”; 11 genes annotated by the GO term (GO:0030100) “regulation of endocytosis”; 178 genes included in the “endocytosis” term in KEGG78,79 and finally 23 genes included in the Reactome80 gene set term “transferrin endocytosis and recycling.”

Finally, the information about the specific cellular compartment, biological function and biological process associated with each Lysosplex gene was obtained by EntrezGene.81

### Targeted sequencing analysis

The libraries were sequenced using the HiSeq1000 system (Illumina inc., San Diego, CA, USA). The generated sequenc- es were analyzed using an in-house pipeline designed to automate the analysis workflow and composed of modules performing every step using the appropriate tools available to the scientific community or developed in-house.82 Paired sequencing reads were aligned to the reference genome (UCSC, hg19 build) using BWA83 and sorted with SAMtools84 and Picard (http://picard.sourceforge.net). Genome Analysis Toolkit (GATK)85 with parameters adapted to the Haloplex-generated sequences was then used to perform local realignment around insertions-deletions (indel) and single nucleotide variants (SNV) and small indel calling. The called SNV and indel variants produced with both platforms were annotated using ANNOVAR86 with: the relative position in genes using the RefSeq87 gene model, amino acid change, presence in dbSNP v137,88 frequency in NHLBI Exome Variant Server (http://evs.gs.washington.edu/ EVS), 1,000 genomes project,89 multiple cross-species conservation90–91 and prediction scores of damaging on protein activity.92–96 The annotated variants were then imported in the internal variation database, which stores all the variations found in internal sequencing projects. The database is then queried to generate the filtered list of variations and the internal database frequency in samples with unrelated phenotype is used as further annotation and filtering criterion. The alignments
at candidate positions were visually inspected using the Integrative Genomics Viewer (IGV). 97

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
We are grateful to Manuela Dionisi and Annalaura Torella and the TIGEM NGS Core for the NGS data generation and the TIGEM Bioinformatics core for support in sequencing data analysis. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also thank Nicola Brunetti-Pierri and Antonella de Matteis for help in ALP gene selection. We thank all patients and families as well as referring physicians for providing patient samples and data.

Supplemental Material
Supplemental material for this article can be accessed on the publisher’s website.

References
1. Settembre C, Fraldi A, Medina DL, Ballabio A. Signals from the lysosome: a control centre for cellular clearance and energy metabolism. Nat Rev Mol Cell Biol 2013; 14:283-96; PMID:23609508; http://dx.doi.org/10.1038/nrm3565
2. Mizushima N, Levine B, Cuervo AM. Chaperone-mediated autophagy and aging: a novel regulatory role of lipids related to cell death. Autophagy 2007; 3:38-9; PMID:17348364; http://dx.doi.org/10.4161/auto.3.4.2426
3. Kaushik S, Kiffin R, Cuervo AM. Chaperone-mediated autophagy and aging: a novel regulatory role of lipids related to cell death. Autophagy 2007; 3:38-9; PMID:17348364; http://dx.doi.org/10.4161/auto.3.4.2426

Acknowledgments
We are grateful to Manuela Dionisi and Annalaura Torella and the TIGEM NGS Core for the NGS data generation and the TIGEM Bioinformatics core for support in sequencing data analysis. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also thank Nicola Brunetti-Pierri and Antonella de Matteis for help in ALP gene selection. We thank all patients and families as well as referring physicians for providing patient samples and data.

Supplemental Material
Supplemental material for this article can be accessed on the publisher’s website.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
We are grateful to Manuela Dionisi and Annalaura Torella and the TIGEM NGS Core for the NGS data generation and the TIGEM Bioinformatics core for support in sequencing data analysis. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also thank Nicola Brunetti-Pierri and Antonella de Matteis for help in ALP gene selection. We thank all patients and families as well as referring physicians for providing patient samples and data.

References
1. Settembre C, Fraldi A, Medina DL, Ballabio A. Signals from the lysosome: a control centre for cellular clearance and energy metabolism. Nat Rev Mol Cell Biol 2013; 14:283-96; PMID:23609508; http://dx.doi.org/10.1038/nrm3565
2. Mizushima N, Levine B, Cuervo AM. Chaperone-mediated autophagy and aging: a novel regulatory role of lipids related to cell death. Autophagy 2007; 3:38-9; PMID:17348364; http://dx.doi.org/10.4161/auto.3.4.2426
3. Kaushik S, Kiffin R, Cuervo AM. Chaperone-mediated autophagy and aging: a novel regulatory role of lipids related to cell death. Autophagy 2007; 3:38-9; PMID:17348364; http://dx.doi.org/10.4161/auto.3.4.2426

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
We are grateful to Manuela Dionisi and Annalaura Torella and the TIGEM NGS Core for the NGS data generation and the TIGEM Bioinformatics core for support in sequencing data analysis. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also thank Nicola Brunetti-Pierri and Antonella de Matteis for help in ALP gene selection. We thank all patients and families as well as referring physicians for providing patient samples and data.
