Meta-analysis of genotype-phenotype associations in Alström syndrome.

Brais Bea-Mascato\textsuperscript{1,2}, Diana Valverde\textsuperscript{1,2}* 

\textsuperscript{1} CINBIO, Universidad de Vigo, 36310, Spain.
\textsuperscript{2} Grupo de Investigación en Enfermedades Raras y Medicina Pediátrica, Instituto de Investigación Sanitaria Galicia Sur (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain

*Correspondence: Diana Valverde

CINBIO Facultad de Biología, Universidad de Vigo, Campus As Lagoas-Marcosende s/n, 36310 Vigo, Spain
Tel +34 986 811 953
Email dianaval@uvigo.es

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract:

Introduction: Alström syndrome (ALMS, #203800) is an ultra-rare monogenic recessive disease. This syndrome is associated with mutations in the ALMS1 gene, which codes for a centrosome structural protein responsible for centrosome cohesion. The type of mutation associated with ALMS is mostly cLOF (97%) and they are mainly located in exons 8, 10 and 16 of the gene. Other studies in the literature have tried to establish a genotype-phenotype correlation in this syndrome with little success. The difficulty in recruiting a large cohort in rare diseases is the main barrier to conducting this type of study.

Methods: In this study we have collected all cases published to date for ALMS. We have created a database with those patients who had a genetic diagnosis and an individualised clinical history. Finally, we have attempted to establish a genotype-phenotype correlation using the truncation site of the patient’s longest allele as a grouping criterion.

Results: We collected a total of 357 patients of which 227 had complete clinical information, complete genetic diagnosis and meta information regarding sex and age. We have seen that there are 5 variants in high frequency with p.(Arg2722Ter) being the most common variant with 28 alleles. There are no gender differences in disease progression. Finally, truncating mutations in exon 10 seem to be correlated with a higher prevalence of liver disorders in patients with ALMS.

Conclusion: The location of the mutation in the ALMS1 gene does not have a major impact on the phenotype developed by the patient. Only mutations in exon 10 of the ALMS1 gene were associated with a higher prevalence of liver disease.

Keywords: ALMS, Alström syndrome, ciliopathy, genotype-phenotype, metabolic disease, obesity, meta-analysis, rare disease.
Introduction:

Alström syndrome (ALMS, #203800) is an ultrarare monogenic disease caused by mutations in the \textit{ALMS1} gene. It is an autosomal recessive disorder with an estimated incidence of 1-9 cases per 1,000,000 inhabitants and currently there are an estimated 1000 cases described worldwide (Orphanet; 03-05-2022).

Most of mutations associated with this disease are complete loss-of-function (cLOF) mutations that generate a stop codon either at the same site as the mutation or result in a frameshift alteration leading to a stop codon (Marshall et al., 2015). There are currently 388 loss-of-function variants reported in ClinVar and 253 in gnomAD. These pathogenic variants have a uniform distribution along the gene and are mainly located in coding regions. Whole exome sequencing (WEG) is a standard practice for the genetic testing of rare diseases, which implies that intronic regions are poorly studied, explaining why most pathogenic variants are detected in coding regions (Sawyer et al., 2016). Exons 8, 10 and 16 are considered mutation hotspots, but this seems to be due to their large size rather than a specific regulatory correlation. Exon 8 (6,108bp) for example covers 50\% of the total gene sequence (12,844bp).

ALMS presents a very heterogeneous phenotype in which symptoms can be aggregates into two main groups (Marshall et al., 2007a, 2011). A first group includes the presence of retinal dystrophy from early life, various metabolic disorders (obesity, hypertriglyceridemia or type 2 diabetes mellitus), hearing loss, liver and kidney dysfunction and cardiac disorders such as dilated cardiomyopathy (DCM) (Marshall et al., 2005, 2007b, 2007a). The second group of symptoms would include short stature, recurrent pulmonary infections, mental and cognitive impairment and various endocrine disorders affecting the thyroid and reproductive system (Marshall et al., 2007a). This second group would also include other very rare symptoms, such as alterations in the fingers, alopecia and spinal abnormalities (Marshall et al., 2007a).

ALMS is characterised by high intra- and inter-familial phenotypic heterogeneity (Marshall et al., 2005, 2007a, 2007b, 2011). This means that siblings with the same genotype may have different phenotypes, which complicates the establishment of a genotype-phenotype correlation (Hollander et al., 2017).
Several studies have attempted to establish a genotype-phenotype correlation without success using cohorts with 12-18 patients (Bond et al., 2005; Minton et al., 2006). Studies in larger cohorts, 58 patients, have shown that there is correlation between disease-causing variants in exon 16 and the presence of retinal degeneration before one year of age, the occurrence of urological dysfunction, DCM and diabetes (Marshall et al., 2007b). Moreover, this study found a significant correlation between disease-causing variants in exon 8 and the absent, mild or delayed kidney disease.

Premature termination codon (PTC) mutations are often classified as cLOF and are associated with the activation of nonsense-mediated mRNA decay (NMD), which results in the elimination of the expression of the mutated gene (Supek et al., 2021). However, it has been shown that NMD is not 100% efficient when nonsense mutations are located in the last exons of genes (Kerr et al., 2001; Lindeboom et al., 2016; Supek et al., 2021). In the case of ALMS, there are patients who continue to express the ALMS1 protein even when carrying two cLOF variants (Chen et al., 2017). In that study of 23 patients where the clinical manifestations were related to the expression or non-expression of the ALMS1 gene, it was observed that those patients with residual ALMS1 gene expression presented milder phenotypes (Chen et al., 2017).

Although there are several databases that list the mutations described in the ALMS1 gene (GnomAD, ClinVar), only one, LOVD, offers a register with 140 patients but the clinical data is absent or incomplete in many cases (Fokkema et al., 2011; Landrum et al., 2018; Karczewski et al., 2020). There is still no open patient registry that collects clinical and genetic information in a comprehensive way. In this study, we have reviewed 112 publications related to ALMS and compiled all available clinical and genetic information. We have collected a total of 357 patients, manually curated, on which various analyses have been performed to look for a genotype-phenotype correlation. Finally, we have made this dataset available to the scientific community in a single public access database (http://genesandfunctions.uvigo.es/alm-s-db/).
Methods:

Preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines:

This meta-analysis was performed following the PRISMA guidelines, as described in Niederlova et al. (Niederlova et al., 2019), for Bardet Biedl syndrome (BBS). In this case, the main PICO question our study sought to address was: Do ALMS patients have different phenotypes depending on where in the gene, the stop mutation occurs?.

Similar studies have been conducted in smaller cohorts grouping patients by major mutational hotspots in the ALMS1 gene. (Marshall et al., 2005, 2007b; Astuti et al., 2017).

Search strategy:

The PubMed and Google Scholar databases were searched in January 2022 using the following keyword combination: ("Alström syndrome" OR ALMS) AND ("genotype phenotype" OR “cohort” OR "case report"). The screening of the search results was carried out by BB-M by initially examining the title and abstract to determine whether it corresponded to the subject of the analysis. Some of the articles analysed in this study were not initially found by the database search but were detected by the recommendations and bibliography of other more relevant works. The search was conducted in English only, covering a period between the origin of the databases and the date of the search (January 2022).

Study selection:

All articles selected in a first search were carefully reviewed to meet the criteria for inclusion in the meta-analysis. After a thorough reading of the article, the study was included if the following characteristics were met: the cohort of the article included a patient with a diagnosis of ALMS, and the diagnosis had a genetic and clinical characterisation. Articles whose diagnosis of ALMS was based solely on phenotype were
discarded, (Charles et al., 1990; Holder et al., 1995; Russell-Eggitt et al., 1998; Koray et al., 2001; Worthley and Zeitz, 2001; Benso et al., 2002; Satman et al., 2002; Paisey et al., 2004; Hoffman et al., 2005; Hamamy et al., 2006; Koç et al., 2006; Gogi et al., 2007; Silan et al., 2013; Bronson et al., 2015; Boerwinkle et al., 2017; Davies et al., 2018), as well as those that simply presented or reported the patient's mutations without giving an complete (Lazar et al., 2015) or individualised clinical history (Patel et al., 2006; Marshall et al., 2007b, 2015; Redin et al., 2012; Kilpinen et al., 2017; Gao et al., 2019; Baig et al., 2020). From an initial selection of 108 studies, 31 (Millay et al., 1986; Charles et al., 1990; Holder et al., 1995; Russell-Eggitt et al., 1998; Koray et al., 2001; Worthley and Zeitz, 2001; Benso et al., 2002; Satman et al., 2002; Wu et al., 2003; Iannello et al., 2004; Paisey et al., 2004; Hoffman et al., 2005; Marshall et al., 2005, 2007b, 2015; Koç et al., 2006; Patel et al., 2006; Hamamy et al., 2006; Gogi et al., 2007; Hitz et al., 2008; Catrinoiu et al., 2009; Redin et al., 2012; Silan et al., 2013; Lazar et al., 2015; Bronson et al., 2015; Ahmad et al., 2016; Huang et al., 2016; Lindsey et al., 2017; Boerwinkle et al., 2017; Gao et al., 2019; Baig et al., 2020) were discarded and 76 (Titomanlio et al., 2004; Bond et al., 2005; Minton et al., 2006; Joy et al., 2007; Ö zgül et al., 2007; Malm et al., 2008; Aldahmesh et al., 2009; Khoo et al., 2009; Liu et al., 2009; Kocova et al., 2011; Wang et al., 2011, 2016a, 2016b, 2021; Aliferis et al., 2012; Jatti et al., 2012; Piñeiro-Gallego et al., 2012; Taokesen et al., 2012; Corbetti et al., 2013; Katagiri et al., 2013; Kuburović et al., 2013; Liang et al., 2013; Mahamid et al., 2013; Taşdemir et al., 2013; Kaya et al., 2014; Bahmad et al., 2014; Louw et al., 2014; Ozantürk et al., 2014; Sanyoura et al., 2014; Casey et al., 2014; Khan et al., 2015; Kim et al., 2015; Long et al., 2015; Sathy Priya et al., 2015; Chakroun et al., 2016; Laxer et al., 2016; Nikopoulos et al., 2016; Xu et al., 2016; Zmyslowska et al., 2016; Chen et al., 2017; Cruz-Aguilar et al., 2017; Das Bhowmik et al., 2017; Hollander et al., 2017; Astuti et al., 2017; Jinda et al., 2017; Bakar et al., 2017; Wicher et al., 2017; Broofferio et al., 2017; Yang et al., 2017; Castro-Sánchez et al., 2017; Dotan et al., 2018; Kilinc et al., 2018; Kvarnung et al., 2018; Maltese et al., 2018; Nasser et al., 2018; Sanchez-Navarro et al., 2018; Tsai et al., 2018; Waldman et al., 2018; Shurygina et al., 2019; Weiss et al., 2019; Gatticchi et al., 2020; Hirano et al., 2020; Hull et al., 2020; Kamal et al., 2020; Lombardo et al., 2020; Mauring et al., 2020; Rethanavelu et al., 2020; Spinelli et al., 2020; Torkamandi et al., 2020; Zhou et al., 2020;
were selected for information extraction and subsequent analyses.

Data extraction and curation:

For the extraction of information, three main groups were defined: metadata, genetic information, and clinical information.

The metadata refers to the: ID of each patient within the dataset, the ID of the family to which they belong (patients with the same family ID are siblings) and the reference from where the information of that patient has been extracted.

In the case of genetic information, whenever possible, the following should be extracted: the original allele reference, the nomenclature of the mutation for the cDNA and protein sequence, the exon/intron where the mutation is located and the genotype of the patient (homozygote or compound heterozygote). All reported mutations were named following the HGVS guidelines with the transcript NM_015120.4 and validated with the name checker of the mutalyzer software (Lefter et al., 2021). The back translator tool of the mutalyzer software was also used to determine the mutation nomenclature at the cDNA level when only the protein annotation of the mutation was provided in the article. In these cases, only nonsense and stop mutations could be determined, as the algorithm does not work with frameshift mutations.

With regard to the clinical information, sex, age and ethnicity were extracted (whenever possible), the diagnostic criteria named by Marshall et al., (Marshall et al., 2007a) were used to try to homogenise it: vision impairments (history of nystagmus in infancy/childhood, legal blindness, cone and rod dystrophy by ERG); obesity and/or insulin resistance and/or T2DM; history of DCM/CHF; hearing loss; hepatic disfunction; renal failure; pulmonary dysfunction; short stature; males: hypogonadism; females: irregular menses and/or hyperandrogenism; thyroid disorders; mental retarded; abnormal appearance of a finger, Intestinal problems, scoliosis/flat wide feet, epilepsy and alopecia (Table 1). Initially, the full information reported in the article was noted for each individual symptom and the age of onset if this was reported. This was then converted into a binary matrix to simplify and homogenise the subsequent analysis.
<table>
<thead>
<tr>
<th>Abbrev.</th>
<th>Complete name</th>
<th>Phenotype</th>
<th>Prevalence (%)</th>
<th>HPO terms added in each category</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI</td>
<td>Vision impairments</td>
<td>History of nystagmus in infancy/childhood, legal blindness, cone and rod dystrophy by ERG</td>
<td>96.54</td>
<td>HP:0000556 (Retinal dystrophy)</td>
</tr>
<tr>
<td>MT</td>
<td>Metabolic anomalies</td>
<td>Obesity and/or insulin resistance and/or Type 2I diabetes mellitus (T2DM)</td>
<td>85.88</td>
<td>HP:0001513 (Obesity)</td>
</tr>
<tr>
<td>HL</td>
<td>Hearing anomalies</td>
<td>Hearing loss</td>
<td>62.54</td>
<td>HP:0000365 (Hearing loss)</td>
</tr>
<tr>
<td>HRT</td>
<td>Heart anomalies</td>
<td>History of Dilated cardiomyopathy/congestive heart failure (DCM/CHF)</td>
<td>49.86</td>
<td>HP:0001644 (DCM)</td>
</tr>
<tr>
<td>LIV</td>
<td>Liver anomalies</td>
<td>Hepatic dysfunction</td>
<td>34.58</td>
<td>HP:0001410 (Decreased liver function)</td>
</tr>
<tr>
<td>REN</td>
<td>Renal anomalies</td>
<td>Renal failure</td>
<td>29.68</td>
<td>HP:0000077 (Renal anomaly)</td>
</tr>
<tr>
<td>MEND</td>
<td>Mental anomalies</td>
<td>Mental disability</td>
<td>22.48</td>
<td>HP:0001249 (Intellectual disability)</td>
</tr>
<tr>
<td>REP</td>
<td>Reproductive system anomalies</td>
<td>Males: hypogonadism; Females: irregular menses and/or hyperandrogenism</td>
<td>15.56</td>
<td>HP:0000026 (Male hypogonadism)</td>
</tr>
<tr>
<td>PUL</td>
<td>Pulmonary anomalies</td>
<td>Pulmonary dysfunction</td>
<td>14.99</td>
<td>HP:0002795 (Impaired pulmonary function)</td>
</tr>
<tr>
<td>SCO</td>
<td>Spine/feet anomalies</td>
<td>Scoliosis/flat wide feet</td>
<td>14.12</td>
<td>HP:0002650 (Scoliosis)</td>
</tr>
<tr>
<td>TYD</td>
<td>Thyroid metabolism anomalies</td>
<td>Thyroid disorders</td>
<td>11.53</td>
<td>HP:0000820 (Abnormality of the thyroid gland)</td>
</tr>
<tr>
<td>SHS</td>
<td>Stature anomalies</td>
<td>Short stature</td>
<td>7.78</td>
<td>HP:0004322 (Short stature)</td>
</tr>
<tr>
<td>ALO</td>
<td>Alopecia</td>
<td>Alopecia</td>
<td>4.61</td>
<td>HP:0001596 (Alopecia)</td>
</tr>
<tr>
<td>NER</td>
<td>Nervous system anomalies</td>
<td>Epilepsy</td>
<td>3.75</td>
<td>HP:0001250 (Seizure)</td>
</tr>
<tr>
<td>ABFING</td>
<td>Finger anomalies</td>
<td>Abnormal appearance of a finger</td>
<td>1.15</td>
<td>HP:0001155 (Abnormality of the hand)</td>
</tr>
<tr>
<td>INT</td>
<td>Intestine anomalies</td>
<td>Intestinal problems</td>
<td>1.15</td>
<td>HP:0011024 (Abnormality of the gastrointestinal tract)</td>
</tr>
</tbody>
</table>
Statistical frequency analysis:

To determine the existence of a genotype-phenotype correlation in our patient registry, we selected a sub-cohort of 227 patients who met the following criteria: I) complete genetic information (Detection of two mutated alleles with complete notation at the cDNA and protein level), II) complete clinical information (Sex, age and presence or absence of the 5 most prevalent clinical manifestations used in these patients: vision impairments, metabolic anomalies, hearing anomalies, heart anomalies and liver anomalies). The heterogeneity of the clinical notation associated with these 5 categories was summarised in a binary matrix (1=presence, 0=absence) which was used for frequency calculations to determine the prevalence of each symptom in the cohort and the different subgroups.

Following the methodology described by Niederlova et al., (Niederlova et al., 2019) a syndromic score between 0 and 1 was calculated. This syndromic score gives an estimate of how many of the 5 most prevalent clinical manifestations described above are present in each patient.

In this study the lack of race was not considered to exclude a patient from being included in the study cohort. Due to the high heterogeneity of this characteristic and the small cohort size, correlations with enough statistical power could not be established.

Patient ages were initially aggregated into 10-year age intervals: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59. After studying the evolution of the syndromic score in the different age intervals, it was found that the number of patients above 39 years was very small. This is consistent with the life expectancy of patients with ALMS, which is usually not more than 50 years. For this reason, a regrouping of the last 3 intervals in patients over 30 years of age was made for the subgroup analysis.

Following the methodology used by Marshall et al., (Marshall et al., 2007b) patients were genotypically grouped according to the different mutational hotspots: exon 8 (E8), exon 10 (E10) and exon 16 (E16). For compound heterozygous patients this aggregation was performed based on the patient's longest allele. Since ALMS is an autosomal recessive disease, it is expected that a single functional copy of the gene is sufficient to
prevent the appearance of the symptoms of the disease (Collin et al., 2002; Hearn, 2018). In this way, it would be expected that a protein truncated in exon 16-20 could have more possibilities of maintaining a certain functionality than if it is truncated in exon 5.

Approximately 90% of the mutations were found in the mutation hotspots (E8, E10, and E16), but to include the remaining % in the study, 3 intervals were defined: longest allele truncated before exon 9 (E8), longest allele truncated between exon 9 and 14 (E10), longest allele truncated after exon 14 (E16).

In cases where only two groups were compared, a Wilcoxon test was used. In trials where comparisons were made between several groups (more than 2), an overall p-value calculation was performed using the non-parametric Kruskal-Wallis test, followed by a comparison of means by groups of two using the Wilcoxon test with correction for FDR, with corrected p-values of less than 0.05 being considered significant.

For the analysis of the prevalence of symptoms in the different genetic groups, contingency tables were created showing the number of positive/negative cases within each patient group. Differences in the prevalence of phenotypes in the different patient groups were assessed using pairwise Fisher's exact test. Statistical significance of differences between individual groups was determined with the FDR correction for multiple comparisons.
Results:

Cohort description:

**Figure 1.** Cohort Description of 227 ALMS patients. (A) Counting of the different alleles in the cohort. Alleles with more than 2 copies in the cohort are represented. (B) Number of alleles per exon in the study cohort. (C) Patients grouped by their copy of the ALMS1 gene with the later truncation mutation. (D) Age composition in the subgroups with the latest truncation site in the ALMS1 gene in exon 8 (E8), exon 10 (E10) and exon 16 (E16).

A total of 357 patients were collected from a total of 76 scientific publications that passed the initial screening. Only patients with information on sex, age and a complete genetic characterisation (description of two non-functional alleles; methods) were used for further analysis. Only mutations in coding regions were considered in the study. This reduced the initial cohort to a total of 227 patients, 128 (56.38%) males and 99 females (43.61%). This study cohort contained a total of 176 variants where 168 were cLOF and 8 missense mutations (Figure S1). Most of them were private mutations, but 5 mutations were shown to have a high frequency: p.(Arg2722Ter) (28 alleles), p.(Gln3495Ter) (22 alleles), p.(Thr3592LysfsTer6) (14 alleles), p.(Glu3773TrpfsTer18) (11 alleles) and p.(Pro3911GlnfsTer16) (10 alleles) (Figure 1A). The number of alleles
described for these mutations shows certain inconsistencies with the gnomAD database and in some cases the variant does not even appear in gnomAD database (Table 2).

Table 2. Differences in the number of alleles reported in gnomAD and in our database for the most recurrent mutations.

<table>
<thead>
<tr>
<th>Pathogenic variant</th>
<th>ALMSDB</th>
<th>gnomAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>p.(Arg2722Ter)/p.(Arg2720Ter)</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>p.(Gln3495Ter)/p.(Gln3493Ter)</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>p.(Thr3592LysfsTer6)/p.(Thr3590LysfsTer6)</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>p.(Glu3773TrpfsTer18)/p.(Glu3771TrpfsTer18)</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>p.(Pro3911GlnfsTer16)</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Patients were initially grouped into age ranges by decades (from 0 to 59 years) but due to the low number of patients in the age ranges 40-49 and 50-59 years, they were added to the 30-39 years group and reclassified as over 30 years for further analysis.

Out of a total of 454 alleles, 207 (45.59%) were in exon 8, 107 (23.57%) in exon 16 and 81 (17.84%) in exon 10 (Figure 1B). Making these exons the main mutation hotspots in our cohort, consistent with the literature.

Due to the recessive nature of ALMS, we decided to define the characteristic allele of each patient according to the mutation furthest from the origin of the gene (Figure 1C). Many of the patients carrying mutations in exon 8 were compound heterozygotes with mutations in exon 16. After this regrouping we saw that 80 (35.34%) patients had the largest allele truncated in exon 8, 42 (18.50%) in exon 10 and 75 (33.03%) in exon 16.

In accordance with other similar study (Marshall et al., 2007b) we used the mutation hotspots (exons 8, 10 and 16) to subdivide the patients into genetic groups E8, E10 and E16 respectively. Due to the low frequency of mutations in other exons of the ALMS1 gene we decided to add patients with truncated alleles in adjacent exons to these mutation hotspots. Group E8 includes all patients with the longest truncated allele before exon 9. Group E10 includes patients with the longest truncated allele between exon 9 and exon 13. Finally group E16 included patients with the longest truncated allele from exon 14 to exon 20. As a result, groups E8, E10 and E16 contained 85, 45 and 97 patients respectively (Figure 1D).
Patients with mutations in exon 10 have a higher syndromic score than the other subgroups:

For the strategy of collecting phenotypic manifestations, 16 syndromic groups (methods) were initially defined. In the genotype-phenotype correlation analysis, a minimum prevalence threshold of 15% (n=33) was established in the cohort (Figure S2).

This reduced the initial phenotypic manifestations to 9 groups: visual impairments (VI, 97.80%), metabolic anomalies (MT, 85.02%), hearing anomalies (HL, 59.91%), heart anomalies (HRT, 49.34%), liver anomalies (LIV, 36.56%), renal anomalies (REN, 29.52%), mental anomalies (MEND, 24.67%), reproductive system anomalies (REP, 17.62%), pulmonary anomalies (RES, 19.38%).

Following the methodology developed by Niederlova et al., (Niederlova et al., 2019), the 5 main disease conditions (VI, MT, HL, HRT and LIV; methods) were used to create a discrete syndromic score ranging from 0 to 1 (Figure 2A). The mean of this syndromic score in the total cohort is around 0.7, with the most common values being 0.6 (3 of the 5 symptoms) and 0.8 (4 of the 5 symptoms) (Figure 2B). No significant differences in syndromic score were observed between sexes. However, the syndromic score increased significantly in the different age ranges of the patients (p-value = 3.1 e-10; Figure 2C-D). This is consistent with the progressive worsening that these patients undergo throughout their lives (Marshall et al., 2005, 2007a). Finally, we observed how the syndromic score was distributed in the different groups defined according to their largest transcript. Patients whose longest transcript was truncated around exon 10 (E10) had a higher syndromic score compared to the E8 (p-value = 0.017) and E16 (p-value = 0.016) groups (Figure 2E).
Figure 2. Study of the phenotypic impact on different subgroups of the study cohort. (A) Prevalence of the 5 main phenotypic groups in the cohort: visual impairments (VI), metabolic anomalies (MT), hearing anomalies (HL), heart anomalies (HRT) and liver
anomalies (LIV). (B) Distribution of syndromic scores calculated from the presence or absence of the 5 most relevant phenotypic groups. (C) Gender comparison of syndromic scores. (D) Comparison of the syndromic scores between age groups. (E) Comparison of the syndromic scores between subgroups created from the longest copy of the ALMS1 gene.

Patients with truncation variants around exon 10 have a higher prevalence of liver disease:

Figure 3 Correlation study between symptoms prevalence and the truncation site of the ALMS1 gene. (A) Evolution of the syndromic score among the subgroups by the longest
copy of the ALMS1 gene in the different age groups. (B) Prevalence of the different symptom clusters in the subgroups by the longest copy of the ALMS1 gene.

To determine whether the differences between the E8, E10 and E16 groups were due to an unequal composition between the different age ranges in each group, the syndromic score was studied by combining both parameters (age and genetic group) (Figure 3A). In the first decade of life, no significant differences in syndromic score were observed between the groups. Between the ages of 10 to 19 years, a higher syndromic score was observed in the E10 group compared to the E16 group (FDR = 0.061). From 20-29 years of age, a higher syndromic score was observed in the E10 group compared to the E8 and E16 groups (FDR = 0.009, in both cases). Finally, in patients over 30 years of age, no significant differences were observed between the genetic groups.

After concluding that the differences observed between the groups were not due to an unequal age composition between them, the prevalence of the 9 phenotypic manifestations in each of these groups was studied (Figure 3B). Patients within the E10 group were found to have a higher prevalence of liver disorders compared to patients in the E8 and E16 groups. For the remaining 8 syndromic manifestations no significant differences were found between the genetic groups.

Discussion:

Meta-analyses are a useful tool to address one of the main limitations when attempting to establish a genotype-phenotype correlation in a rare disease, the inability to recruit a large cohort to draw robust and statistically significant conclusions. Although meta-analyses mainly focus on the aggregation of several studies with well-defined and well-studied cohorts, this is often not possible in rare diseases. Alternatively, one can try to aggregate information obtained from case reports in the literature. However, the lack of uniform criteria in how authors describe their patients is one of the main biases of this approach. In this paper we adapt and apply the methodology described by Niederlova et al (Niederlova et al., 2019) from a polygenic disease such as BBS to a monogenic disease such as ALMS.
Despite the fact that ALMS is a monogenic disease (Hearn et al., 2002), to date several studies have already discussed the possible existence of several tissue-specific isoforms for the ALMS1 gene, which may have different regulatory roles that explain the high symptomatic heterogeneity (Collin et al., 2002, 2012; Hearn et al., 2002; Braune et al., 2017). Although this is an interesting approach, the isolated nature of the data published in the case reports does not allow validation it. In order to validate this hypothesis, the recruitment of a large cohort is necessary. Different cohorts have been described in the literature, the largest being Marshall et al. (Marshall et al., 2007b), followed by Ozantürk et al. (Ozantürk et al., 2014), the NIH clinical centre cohort (Brofferio et al., 2017; Waldman et al., 2018), Chen et al. (Chen et al., 2017) and Rethanavelu et al. (Rethanavelu et al., 2020) with 58, 44, 38, 23 and 21 ALMS patients, respectively. Significant genotype-phenotype correlations were only detected in Marshall et al (Marshall et al., 2007b), highlighting the importance of the sample size used in this type of study.

The causal mutations of ALMS are mainly of the cLOF type (generating a PTC) (Marshall et al., 2015). These types of mutations activate nonsense-mediated decay (NMD), preventing the translation of the sequence from mRNA to protein (Supek et al., 2021). However, certain situations have been described, such as PTC mutations in the last exons of a gene, in which NMD can be prevented or incompletely produced (Lindeboom et al., 2016; Supek et al., 2021). Chen et al., have detected ALMS1 gene expression in patients carrying two cLOF mutations leading to PTC (Chen et al., 2017). That study has also correlated residual ALMS1 gene expression with the development of milder phenotypes (Chen et al., 2017). Taking this into account, it could happen that cLOF mutations in the terminal exons of the ALMS1 gene do not activate NMD, generating a partially functional protein or a non-functional misfolded protein. Under this hypothesis, we grouped the patients in our study according to the longest transcript they had. Three different groups related to the main mutational hotspots of the ALMS1 gene, exons 8, 10 and 16, were defined. A similar approach has already been used on a cohort of 58 ALMS patients by Marshall et al. (Marshall et al., 2007b). Subsequently, a syndromic score was created by adding the 5 clinical manifestations more prevalent of ALMS. This methodology was adapted from Niederlova et al, (Niederlova et al., 2019).
The results showed that the syndromic score increases with age (Figure 2D), which was consistent with the gradual worsening that these patients suffer throughout their lives (Marshall et al., 2005, 2007a). This helped us to validate the effectiveness of the syndromic score in measuring the severity of the patient symptoms. When comparing the syndromic scores of the different genetic groups by age groups, it was observed that patients with truncation mutations in exon 10 (E10) evolve more unfavourably than the rest (Figure 3A). The fact that these differences are not appreciable after the age of 30 years could be due to a higher mortality in this group in the second and third decades of life. This supposed higher mortality after 30 years could be correlated with the greater prevalence of liver problems that patients with truncation mutations at exon 10 (E10) appear to have (Figure 3B). This correlation was not detected in previous genotype-phenotype analyses (Bond et al., 2005; Minton et al., 2006; Patel et al., 2006; Marshall et al., 2007b). Marshall et al., (Marshall et al., 2007b) described disease-causing variants in exon 16 as leading to urological dysfunction, DCM and diabetes. In our study, we did not look for correlations in urological dysfunction due to the lack of data reported in the literature, leading to a low prevalence of these types of symptoms. In the case of DCM and diabetes the prevalence was the same among the established genetic groups. Marshall et al., (Marshall et al., 2007b) also found a significant association between alterations in exon 8 and absent, mild or delayed kidney disease; findings that were not observed in our analysis.

The explanation for this event could be related to the residual expression of ALSM1 gene. Mutations in exon 10 could prevent NMD activation and give rise to a misfolded protein that causes greater hepatotoxicity. In the case of mutations in exon 16 or adjacent, perhaps the NMD is not activated either, but the generated protein, despite not being functional, could have a semi-normal conformation that prevents the formation of aggregates. This hypothesis could be easily tested if these protein isoforms can be simulated by homology, but unfortunately the protein structure of the ALMS1 gene remains unknown and cannot be simulated using the artificial intelligence models such as AlphaFold (van Breugel et al., 2022).

Finally, the ALMS1 sequence has already described the presence of a IncRNA, ALMS1-IT1, with a role in regulating proliferation in various types of cancers and
neuroinflammation and a pseudo gene, *ALMS1P1*, whose function is still unknown (Lu et al., 2021; Luan et al., 2021; Mei et al., 2021). Due to the large length of the *ALMS1* gene, the presence of more of these elements would not be uncommon and could explain why the localisation of the cLOF variant can lead to different phenotypes. These events could be of help understanding the regulation of *ALMS1* in the different tissues involved in the disease, and the great heterogeneity observed for the clinical symptoms.

**Conclusion:**

In conclusion, 5 highly prevalent mutations were detected in our cohort. Not all of these variants appear in public databases. There are no gender differences in the prevalence of the different symptoms of ALMS. Patients' symptoms worsen with age, and the site of the *ALMS1* gene mutation does not influence the worsening of symptoms in most cases. Only truncation of the ALMS1 protein at exon 10 level correlated with a higher prevalence of liver problems in these patients.

**Author Contributions:** BB-M, DV designed the study. BB-M selected, reviewed, collected and curated data from the scientific articles. BB-M designed and executed the analysis pipeline. BB-M created the publicly accessible online database. BB-M and DV drafted the article. All authors wrote the manuscript and provided approval for publication.

**Funding:** This work was funded by Instituto de Salud Carlos III de Madrid FIS project PI15/00049 and PI19/00332, Xunta de Galicia (Centro de Investigación de Galicia CINBIO 2019-2022) Ref. ED431G-2019/06, Consolidación e estructuración de unidades de investigación competitivas e outras accións de fomento (ED431C-2018/S4). Brais Bea-Mascato (FPU17/01567) was supported by graduate studentship awards (FPU predoctoral fellowship) from the Spanish Ministry of Education, Culture and Sports.

**Informed Consent Statement:** Not Apply.

**Data Availability Statement:** Data are available via [https://github.com/BreisOne/meta-analysis_ALMS](https://github.com/BreisOne/meta-analysis_ALMS) or [http://genesandfunctions.uvigo.es/alms-db/](http://genesandfunctions.uvigo.es/alms-db/).

**Code Availability Statement:** The code used in this study can be consulted on the GitHub repository [https://github.com/BreisOne/meta-analysis_ALMS](https://github.com/BreisOne/meta-analysis_ALMS).
References:


