# Long-read single molecule real-time (SMRT) sequencing of GBA1 locus in Gaucher disease national cohort from Argentina reveals high frequency of complex allele underlying severe skeletal phenotypes: Collaborative study from the Argentine Group for Diagnosis and Treatment of Gaucher Disease

Guillermo I. Drelichman [a,*], Nicolas Fernández Escobar [a], Barbara C. Soberon [a], Nora F. Basack [a], Joaquin Frabasil [b], Andrea B. Schenone [b], Gabriel Aguilar [c], Maria S. Larroudé [c], James R. Knight [d], Dejian Zhao [d], Jiapeng Ruan [e], Pramod K. Mistry [e], Argentine Group for Diagnosis and Treatment of Gaucher Disease[1]

[a] *Unidad de Hematología, Hospital de Niños "Ricardo Gutiérrez", Ciudad Autónoma de Buenos Aires, Argentina*
[b] *Laboratorio de Neuroquímica "Dr. N. A. Chamoles", Ciudad Autónoma de Buenos Aires, Argentina*
[c] *Centro de Diagnóstico Dr. Rossi, Ciudad Autónoma de Buenos Aires, Argentina*
[d] *Yale University Center for Genome Analysis, Yale School of Medicine, New Haven, CT, United States*
[e] *Department of Internal Medicine, Yale School of Medicine, New Haven, CT. United States*

## ARTICLE INFO

## ABSTRACT

Gaucher disease is renowned for extreme phenotypic diversity that does not show consistent genotype/phenotype correlations. In Argentina, a national collaborative group, Grupo Argentino de Diagnóstico y Tratamiento de la Enfermedad de Gaucher, GADTEG, delineated uniformly severe type 1 Gaucher disease manifestations presenting in childhood with large burden of irreversible skeletal disease. Here using Long-Read Single Molecule Real-Time (SMRT) Sequencing of GBA1 locus, we show that the Rec*Nci*I allele is highly prevalent and it is associated with severe skeletal manifestations with onset in childhood or in young adults. Additionally, we described novel GBA1 variants not previously described.

## 1. Introduction

Gaucher disease (GD) is a prototype lysosomal storage disease due to bi-allelic mutations in *GBA1,* which encodes lysosomal acid β-glucosidase (glucocerebrosidase, EC 3.2.1.45) [1]. Deficiency of acid β-glucosidase leads to a progressive accumulation of glucosylceramide (GlcCer) and glucosylsphingosine (GlcSph) in the lysosomes of myeloid cells, most prominently displayed by the macrophages [2,3]. Three broad phenotype categories have been classified based on the absence of (type 1, GD1: non-neuronopathic, OMIM # 230800), or the presence of, and severity of early onset neurodegenerative symptoms (type 2, GD2: acute neuronopathic form, OMIM # 230900; type 3, GD3: chronic neuronopathic forms, OMIM # 231000) [3,4]. In GD1, some patients develop neurodegeneration as adults manifesting as Parkinson's disease and Lewy Body Dementia [ 5].

*GBA1,* located on chromosome 1q21 of GRCh37/hg19 (now on 1q22

of the latest version GRCh38/hg38), is comprised of 11 exons and 11 introns spanning 7.6 Kb DNA. Notably, it is located in a highly gene dense region that harbors seven genes and two pseudogenes within only 85 Kb of DNA [6,7] (Fig. 1). There is a highly homologous pseudogene (*GBAP1*), 16 Kb downstream from the *GBA1* with an exon and intron organization spanning 5.7 kb, similar to *GBA1*. In fact, the assigned exons of the *GBAP1* share up to 98% sequence homology with the coding region of *GBA1* [6]. Notably, *GBAP1* harbors many mutations which, if present in the *GBA1*, causes Gaucher disease. The genomic organization of the *GBA* locus results in a propensity for gene conversion events, which underlie numerous disease mutations and complex alleles involving *GBA1* and *GBAP1*. These attributes of *GBA* present significant challenges for the accurate and comprehensive genotyping of patients in the clinics and in large cohorts studies.

Currently, more than 600 *GBA1* pathogenetic variants have been catalogued (HGMD professional 2020.4; CentoLSD™, https://www.cent ogene.com/centolsd.html) The most prevalent *GBA1* variants are N370S (c. 1226A > G; p.Asn409Ser), a founder mutation from Eastern Europe, and L444P (c.1448 T > C; p.Leu483Pro), a variant in *GBAP1* transferred to *GBA1* by gene conversion events that occurs recurrently in all populations of the world [7,8].

Delineation of the genotype/phenotype correlations have been the focus of many studies. N370S mutation is predictive of GD type 1, therefore is a neuro-protective variant for childhood-onset of neurodegenerative disease (GD2 or GD3). However, it does not protect from late-onset neurodegenerative diseases, Parkinson's disease and Lewy Body Dementia, seen in some patients with GD1. It accounts for more than 70% of pathogenetic variants in GD type 1 Ashkenazi Jewish patients, and ~ 30% of pathogenetic variants in European non-Jewish GD patients [9]. In contrast, homozygosity for the L444P mutation is strongly associated with GD3, and, when present in context of a complex allele, it can be associated with the most severe forms neuronopathic GD [8].

A wide spectrum of *GBA1* mutations have been reported, including missense/nonsense, indels, gross deletions/insertions, duplications, alternative splicing, promoter elements, regulatory RNAs, and complex recombinant alleles. GD complex alleles arise from the high homology and the physical proximity between *GBA1* and *GBAP1*, that enable reciprocal as well as nonreciprocal homologous recombination events [10]. To date, more than 20 *GBAP1* derived recombinant alleles have been reported, where recombination sites are variable spanning intron 2 to exon 11 of *GBA1*. The most frequently encountered complex alleles are Rec*Nci*I and RecDelta55 (c.1263-1317Del55) [11]. Recombination events underlying Rec*Nci*I mutation occur in the area from intron 9 to exon 10, involve the incorporation of a *GBAP1* segment that harbors three variants: L444P, A456P, and the silent change of V460V [12]. Therefore, targeted NGS sequencing or Sanger sequencing analysis of only L444P (as occurs in many diagnostic panels not using whole gene sequencing) can miss a complex allele, hindering optimal genetic care and confounds genotype/phenotype studies. Taken together, the complexity of the *GBA1* locus challenges and confound precise genotype assignment which can hamper accurate assessment of individual patients and large cohort studies.

A major cause of morbidity and disability in GD1 is complex skeletal disease, manifesting as chronic unrelenting bone pain, avascular osteonecrosis, complex lytic bone lesions and fragility fractures. In Europe and the US, bone disease occurs in GD patients with a frequency of 50-60% [13,14]. In contrast, in the Argentinian GD population, bone involvement is more frequent at diagnosis (71%) and remains predominant after long-term follow-up (69.8%), despite the Enzyme Replacement Therapy (ERT) [15]. To understand what types of mutation are associated with bone disease in GD, comprehensive *GBA1* analysis is necessary. Previous studies have examined prevalent individual pathogenetic variants and in general shown that "N370S/other allele variant (s)" is associated with more severe skeletal disease [16]. Hitherto,
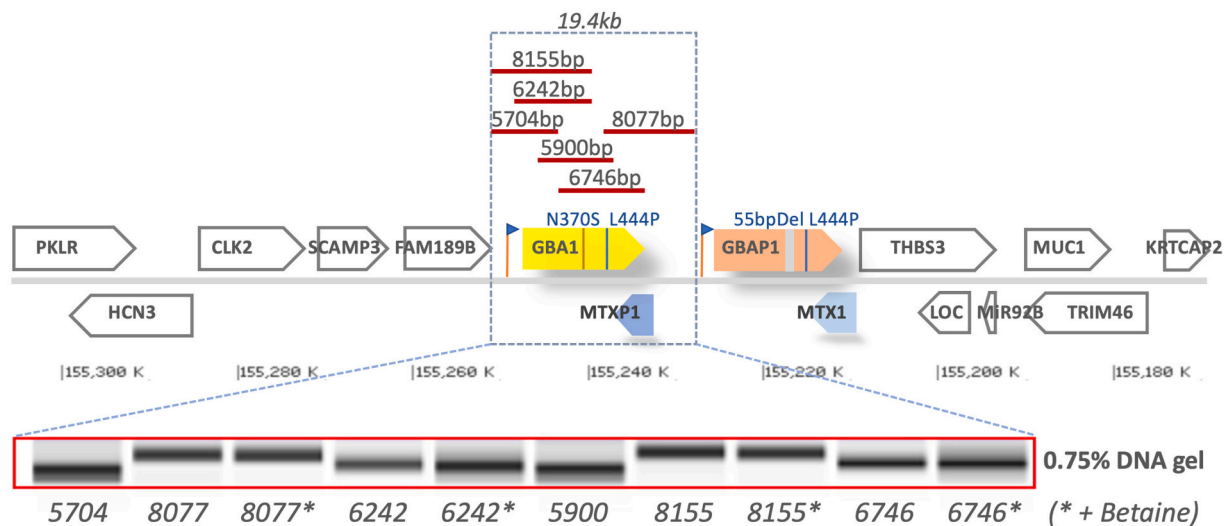


**Fig. 1.** **Location of GBA1 on Chromosome 1q21 with flanking genes, and the LR-PCR amplicons for SMRT sequencing.** The 133kb human *GBA1* loci (GRCh38. p13, Chr1q22; NC_000001.11) consist of 15 genes: **PKLR** (pyruvate kinase L/R, Chr1:155,289,293..155,301,438. Length:12,146nt); **HCN3** (hyperpolarization activated cyclic nucleotide gated potassium channel 3, complement Chr1:155,277,427..155,289,848. Length:12,422nt), **CLK2** (CDC like kinase 2, Chr1:155,262,868..155,273,504. Length:10,637nt), **SCAMP3** (secretory carrier membrane protein 3, Chr1:155,255,981..155,262,360. Length:6,380nt), **FAM189B** (family with sequence similarity 189 member B, Chr1:155,247,205..155,255,892. Length:8,688nt), **GBA1** (glucosylceramidase beta, Chr1:155,234,452..155,244,627. Length:10,176nt), **MTX1P1** (metaxin 1 pseudogene 1, complement Chr1:155,230,976..155,234,451. Length:3,476nt), **GBAP1** (glucosylceramidase beta pseudogene 1, Chr1:155,213,825..155,227,534. Length:13,710nt), **MTX1** (metaxin 1, complement Chr1:155,208,699..155,213,839. Length:5,141nt), **THBS3** (thrombospondin 3, Chr1:155,195,588..155,209,180. Length:13,593nt), **LOC (THBS3-AS1**/LOC105371450, complement Chr 155,196,035..155,200,571. Length:4,537nt), **MIR92B** (microRNA 92b, complement Chr1:155,195,177..155,195,272. Length:96nt), **TRIM46** (tripartite motif containing 46, complement Chr1:155,173,381..155,184,971. Length:11,591nt), **MUC1** (mucin 1, cell surface associated, Chr1:155,185,824..155,192,915. Length:7,092nt), **KRTCAP2** (keratinocyte associated protein 2, Chr1:155,169,408..155,173,304. Length:3,897nt). GBA1 pseudogene (GBAP1) is approximately 12 kb downstream of GBA1 gene. Red bar, length and location of the six long-range (LR) SMRT amplicons used in this study (primers in Table 1). Lower panel, purified LR-PCR amplicons on 0.75% agarose gel.

approaches to ascertainment of *GBA1* mutations in such studies have relied mostly on screening for common mutations and not full gene sequencing. Therefore, the nature of genotype/phenotype correlation with respect to bone disease in GD is not fully understood.

The Argentine Group for Diagnosis and Treatment of Gaucher Disease (Grupo Argentino de Diagnóstico y Tratamiento de la Enfermedad de Gaucher, GADTEG) was created in 2006. Our collaborative group is formed by ~70 physicians throughout Argentina, to monitor the phenotypic spectrum, natural history, and treatment outcomes in 300 patients across the country. We showed a strikingly prevalence of skeletal disease in our GD1 population [15]. Until 2017, only 30% of our patients underwent limited genotyping for common pathogenetic variants. Therefore, our cohort of GD patients offers a unique opportunity to understand the genetic contribution of *GBA1* variants to high burden of skeletal disease. Such studies are essential prelude for unravelling putative modifier genes. Moreover, knowledge of comprehensive genotype of in our patients promises to advance precision medicine for optimal management of Gaucher disease.

## 2. Materials and methods

### 2.1. Patients

A total of 192 patients provided informed consent to participate in the study. All patients had enzymatically confirmed diagnoses of Gaucher disease in peripheral blood leucocytes. Patients have been longitudinally followed and comprehensively evaluated for hematological, visceral and bone disease indicators as described previously [14]. Response to ERT (Enzyme Replacement Therapy) has also been carefully documented.

DNA samples were processed at Laboratorio "Dr. N. A. Chamoles" in Argentine and sequenced at Yale Center of Genome Analysis (YCGA). PacBio long-read Single Molecule Real-Time (SMRT) *GBA1* deep sequencing was developed using *GBA1* specific primers depicted in Fig. 1, covering from the 5'-UTR to 3'-URT of *GBA1* gene, avoiding amplification of *GBAP1*.

To fully genotype *GBA1*, a total of six specific LR-PCR amplicons (5.7 to 8.15 kb, spanning 19.4 kb) were designed as shown in Fig. 1. Primers (Table 1) were optimized to yield amplicons of similar amount to enhance sequencing efficiency and loading capacity for SMRT sequencing. LR-PCR fragments were amplified from 100 ng of genomic DNA, using 1× PrimeSTAR GXL polymerase (R050B, Takara Bio USA Inc.) on a 25 μl of PCR reaction volume with 200 nM of barcode tagged primers. Initial denaturation was performed for 8 min at 98°, followed by 30 cycles of 10 s at 98 °C, 15 s at 60 °C, and 10 min at 68 °C, respectively. Final extension was 10 min at 68 °C. In some cases, 5 ul of 5M Betaine solution (B0300, Sigma-Aldrich USA) was included to the 25 μl of LR-PCR reaction to increase PCR efficiency. Pooled PCR products were size selected, purified, and visually inspected on agarose gel. Equal amount of pooled amplicons were used to generate SMRT libraries and

**Table 2**
LR-PCR primers used to amplify *GBA1* gene for Sanger sequencing.

| Primer ID | Primer Sequence | Amplicon Size (bp) |
|---|---|---|
| NA2568F | CCATCCTCTGGGATTTAGGAGC | |
| NA2568R | GAAGTCAGGGTCCAAAGAAAGGG | 2568 |
| NB2664F | TGCATCCCTAAAAGCTTCGGCTA | |
| NB2664R | GGTGAGTACTGTTGGCGAGGG | 2664 |
| NC2470F | CTCAAGACCAATGGAGCGGT | |
| NC2470R | TCGACAAAGTTACGCACCCA | 2470 |
| C1600F | CTTCCTGCAAAGCAGACCTCA | |
| C1600R | TTGGGCCCAGCTTTCCTAGTC | 1600 |

sequenced on SMRT cells in PacBio RS II system [17] according to manufacturer's instructions. Briefly, the damages of the pooled amplicons were first repaired, followed by end-repair and A-tailing. Next, PacBio sequencing adaptors were ligated to the pooled amplicons and purified using AMPure PB beads. The final PacBio libraries were then annealed to the sequencing primer, bound to the polymerase, and loaded on the PacBio RS II for sequencing. SMRT phasing was then performed to resolve individual alleles. The genotype was validated in original patient DNA sample by Sanger sequencing.

For Sanger sequencing, four Sanger LR-PCR were designed (Table 2) covering the whole *GBA1* (Table 1). Sanger LR-PCR enzyme and conditions were the same as for PacBio LR-PCR, except the extension time was decreased to 3 min. For each patient, four gel-purified Sanger LR-PCR fragments were prepared and stored individually at -20 °C. Each variant detected by SMRT was verified at least two times from both 5'- and 3'- directions by LR-PCR based Sanger sequencing.

Genotyping results were evaluated for correlation with skeletal phenotype and compared our genotype distribution with that reported in the Gaucher International Registry data (International Collaborative Gaucher Group, ICGG, https://clinicaltrials.gov/NCT00358943).

### 2.2. Statistical methods

Qualitative variables are expressed as frequency and percentage, while quantitative variables are expressed as mean, median, minimum, and maximum. For univariate analysis, contingency tables were evaluated using chi-square test or Fisher's exact test, as appropriate. Yates' continuity correction was applied to 2 × 2 contingency tables. Logistic regression models were used for multivariate analysis. Alpha values of 0.05 were considered as statistically significant.

In the multivariate statistical analysis we included 4 variables: Rec*Nci*l allele, Rec*Nci*l/N370S genotype, Rec*Nci*l/other genotype and Argentine ancestors.

In addition, we took 2 models as a dependent variable:

· **MODEL I**: less severe bone manifestations during follow up, i.e., bone marrow infiltration and EFD.

**Table 1**
LR-PCR primers used to amplify the 1q21 *GBA1* region for PacBio SMRT.

| Primer ID | Primer Sequence (Universal tag + gene specific sequence)* | Amplicon Size (bp) |
|---|---|---|
| F5704/F8155 | <u>GGTAG</u>GCGCTCTGTGTGCAGCtCGGGGTTGGGATTCGCACT | |
| R5704 | <u>CCATC</u>TCATATGTAGTACTCTtGATGTCCAGGGGCTGGCAA | 5704 |
| R8155 | <u>CCATC</u>TCATATGTAGTACTCTtGATGTCCAGGGGCTGGCAA | 8155 |
| F6242 | <u>GGTAG</u>GCGCTCTGTGTGCAGCgGCCACACCATGGACAGCTT | |
| R6242 | <u>CCATC</u>TCATATGTAGTACTCTtTGGGTCCTCCTTCGGGGTT | 6242 |
| F5900 | <u>GGTAG</u>GCGCTCTGTGTGCAGCaGCAGATGTGTCCATTCTCCATGT | |
| R5900 | <u>CCATC</u>TCATATGTAGTACTCTtTGTCTCCATCCAGCGGGCA | 5900 |
| F6746 | <u>GGTAG</u>GCGCTCTGTGTGCAGCgGTCCACTTTCTTGGCCGGA | |
| R6746 | <u>CCATCT</u>CATATGTAGTACTCTaACCTATTGCTATGAAAAGGAGCAG | 6746 |
| F8077 | <u>GGTAG</u>GCGCTCTGTGTGCAGCgGACCGACTGGAACCTTGCC | |
| R8077 | <u>CCATC</u>TCATATGTAGTACTCTgCCAGCACACCCTTAGTGGG | 8077 |

* Each primer has 5 base padding sequence at 5'-end (underline) following with barcode sequence and *GBA1* gene specific sequence (starts from the lower-case nucleotide).

· **MODEL II**: severe bone manifestations during follow up, i.e., acute and/or chronic avascular necrosis and bone marrow infarcts).

We found, in both models (I and II), that RecNciI/N370S genotype and patient carrying heterozygote a RecNciI allele have a statistically significant association with model I (p = 0.017) and model II (p = 0.004) Argentine ancestors were not a significant variable (p = 0.059) in either of the two models.

## 3. Results

Of a total of 192 samples, 146 (76%) were successfully genotyped by SMRT sequencing. In 46 samples, full *GBA1* genotype could not be ascertained in the first round of PacBio SMRT analysis. As the key objective of our study was delineating the comprehensive *GBA1* genotype of our cohort, we removed these 46 samples from further analysis. Separately, we are conducting studies to optimize our sequencing strategies to overcome this limitation. All genotypes assigned by SMRT were confirmed by Sanger sequencing in the original gDNA sample.

### 3.1. Pathogenetic variants

The most frequent Gaucher disease mutation in the Argentine cohort is N370S, with 126 patients (86.3%) harboring at least one N370S allele. Notably, this frequency is similar to that reported in the ICGG (June 2014) [13] for Ashkenazi Jewish and non-Jewish European populations (Table 3).

The second most frequent allele was RecNciI, with 77 patients (52.7%) harboring this complex mutation (Table 4). Notably, this allelic frequency for RecNciI is the highest reported in the literature, including a previous small study from Argentina [18–23].

### 3.2. Genotypes

We retained the older mutation nomenclature to facilitate comparison with past studies, *i.e.*, N370S for p. Asp409Ser and L444P for p. Lys483Pro, respectively. The most frequent genotype in our population was N370S/RecNciI. These compound heterozygotes mutations occurred in 68 patients (46.6%). This is the highest frequency of compound heterozygote N370S/RecNciI genotype reported in the literature. In our study of Argentine GD population, it accounted for 46.6% of genotypes compared to only 1.9% in the International Gaucher Registry (p = 0.001) (Table 5). Interestingly, 14 patients (9.6%) were homozygous for N370S mutation. Equally prevalent genotype was N370S/L444P with 14 patients (9.6%) harboring this compound heterozygous genotype. This is similar to other regions worldwide. In 14 patients (9.6%), we found multiple rare or novel mutations (Table 6); in 10 patients, the rare mutation was in compound heterozygous state with N370S and in only one patient with L444P. One patient was homozygous for F411I mutation, and 6 patients were heterozygous for F411I.

### 3.3. Clinical characteristics of N370S/RecNcil genotype patients at diagnosis

The finding of the high prevalence of N370S/RecNciI genotype in our cohort offered an unprecedented opportunity to understand the phenotypic spectrum associated with this combination of mutations. First, all patients had type 1 GD, underscoring the notion that presence of at least one N370S mutation predicts GD1 and that it is neuroprotective against neuronopathic GD [24]. Patients with N370S/RecNciI genotype had severe disease indicated by onset of symptoms at mean age 8.9 years, and it was not until mean age 21.4 years that they were diagnosed with Gaucher disease and started ERT at mean age 28.1 years. These results reveal a wide gap in interval between onset of symptoms in pediatric age group and delayed diagnosis as adults and moreover significant delay from diagnosis to initiation of ERT for a population that is severely affected. Notably, patients with this genotype had severe visceral and bone disease at diagnosis and as well as during follow-up. The numbers of patients with other genotypes were too small to conduct meaningful comparisons.

Bone involvement is the most disabling and incapacitating complication of GD1. The major goal of treatment for Gaucher disease is to improve bone health by preventing irreversible complications such as avascular necrosis and fragility fractures as well as ameliorate chronic bone pain. In our cohort of N370S/RecNciI Gaucher disease patients, we found a higher prevalence of skeletal disease at diagnosis (71%) as well as during follow up on ERT (69.8%), compared to the ICGG Registry which shows bone disease in 59.8% of patients. Moreover, there was major diagnostic delay with onset of symptoms in childhood and diagnosis in adulthood. These findings underscore the large burden of irreversible bone complications at initiation of treatment, such as avascular osteonecrosis (Table 7).

#### 3.3.1. N370S/RecNcil genotype and bone manifestations

Comparing the presence of bone manifestations in our RecNciI/N370S genotype patients with those reported at the ICGG (June 2019), we observed that bone manifestations were significantly more frequent in our population not only at diagnosis (61% vs. 38%, p = 0.001), but also after long-term follow-up (median 154 months) (70% vs. 57%, p = 0.001).

#### 3.3.2. Univariate analysis

The presence of bone lesions at diagnosis and follow-up (mean duration of ERT 154 months) was compared in N370S/RecNciI and F411I/RecNciI patients (n = 77) and those with other genotypes (n = 69). The prevalence of bone lesions at diagnosis did not differ in the two genotype groups(p = 0.183). At follow-up, bone lesions were significantly more frequent among patients harboring RecNciI allele (p = 0.017). Additionally, follow-up on ERT, both genotypes (RecNciI/N370S and RecNciI/ F411I), were significantly associated with bone lesions (p = 0.017 and p = 0.001, respectively). Rates of splenectomy in different genotypes were similar, ~9% (data not shown).

**Table 3**

N370S allele frequency in Argentine compared with other regions worldwide (ICGG 2014) [13].

| Region | Argentine | Europe[1] | Japac[2] | Latin America[3] (Without Argentine) | Middle East[4] | North America[4] | TOTAL (Without Argentine) |
|---|---|---|---|---|---|---|---|
| Total genotyping | 146 | 971 | 135 | 497 | 697 | 1854 | 4156 |
| 1 N370S ALLELE (Source: ICGG 2014) | 86.3% | 74.8% | 4.4% | 79% | 84.6% | 83.8% | 79.2% |

[1]**Europe:** Albania, Austria, Balearic Islands, Belgium, Bulgaria, Czech Republic, Denmark, England, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Lithuania, Netherlands, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovenia, Spain, Switzerland, Turkey, and Ukraine. [2] **JAPAC:** China, Hong Kong, India, Japan, Korea, Malaysia, Philippines, Taiwan, and Thailand. [3]**Latin América:** Bolivia, Brasil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, México (n = 13; 1.4%), Panamá, Paraguay, Perú, Suriname, Uruguay, and Venezuela. [3]**North America:** Canada and United States. [4]**Middle East:** Egypt, Israel, Jordan, Kuwait, Oman, Saudi Arabia, and Unites Arab Emirates.

**Table 4**

Comparison of Rec*Nci*l allele frequency in Argentine and other regions.

| Country | Argentine [15] | Egypt [16] | Spain [17] | Argentine [18] | India [19] | Brazil [20] |
|---|---|---|---|---|---|---|
| Total genotyping | 146 | 26 | 193 | 31 | 22 | 58 |
| 1 Rec*Nci*l ALLELE | 52.7% | 13.4% | 0.7% | 21% | 7% | 15.5% |

**Table 5**

Genotype frequency in Argentine compared with other regions of the world (ICGG 2019)[12].

| Region | Argentina | Europe[1] | Japac[2] | Latin America[3] (without Argentina) | Middle East[4] | North America[4] | Oceania | Total (without Argentina) |
|---|---|---|---|---|---|---|---|---|
| Total genotyping | 146 | 1345 | 139 | 284 | 921 | 2146 | 42 | 4877 |
| N370S/Rec*Nci*I (Source: ICGG 2019) | 46.6% | 4.8% | 0% | 2.8% | 0.5% | 0.7% | 4.8% | 1.9% |
| N370S/L444P (Source: ICGG 2019) | 9.6% | 17.8% | 2.9% | 26.8% | 4.7% | 13.4% | 19% | 13.8% |

[1]**Europe:** Albania, Austria, Balearic Islands, Belgium, Bulgaria, Czech Republic, Denmark, England, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Lithuania, Netherlands, Norway, Polad, Portugal, Romania, Russia, Serbia, Slovenia, Spain, Switzerland, Turkey, and Ukraine. [2]**JAPAC:** China, Hong Kong, India, Japan, Korea, Malaysia, Philippines, Taiwan, and Thailand. [3]**Latin America:** Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Mexico (n = 13; 1.4%), Panama, Paraguay, Peru, Suriname, Uruguay, and Venezuela. [3]**North America:** Canada and United States. [4]**Midde East:** Egypt, Israel, Jordan, Kuwait, Oman, Saudi Arabia, and United Arab Emirates. [6] **Oceania:** Australia and New Zealand.

**Table 6**

Summary of *GBA1* genotypes in Argentine GD national cohort.

| Genotype (n = 146) | Frequency |
|---|---|
| Rec*Nci*l/N370S | 68 (46.6%) |
| N370S/N370S | 14 (9.6%) |
| New (see Table 10) | 14 (9.6%) |
| Rec*Nci*l/F411I | 9 (6.2%) |
| L444P/N370S | 9 (6.2%) |
| F411I/F411I | 6 (4.1%) |
| F411I/N370S | 4 (2.7%) |
| R285C/N370S | 2 (1.4%) |
| L444P/R496H | 2 (1.4%) |
| H255Q-D409H/F411I | 2 (1.4%) |
| F411I/R48W | 2 (1.4%) |
| G202R/N370S | 2 (1.4%) |
| G195W/N370S | 2 (1.4%) |
| H255Q-D409H/N370S | 1 (0.7%) |
| R120Q/N370S | 1 (0.7%) |
| V394L/N370S | 1 (0.7%) |
| R163X/R463C | 1 (0.7%) |
| I161N/N370S | 1 (0.7%) |
| L371V/L444P | 1 (0.7%) |
| I260T/N370S | 1 (0.7%) |
| F411I/E233X | 1 (0.7%) |
| F397S/N370S | 1 (0.7%) |
| F411I/Y135C | 1 (0.7%) |
| Total | 146 |

**Table 7**

Bone manifestations at diagnosis and follow-up in the Argentine population vs. world population (Source: ICGG 2014) [13].

| | Argentine Basal | ICGG Basal | Argentine follow-up | ICGG follow-up |
|---|---|---|---|---|
| Patients, N | 260 | 4625 | 266 | 4625 |
| Bone pain | 65.6% | 49.8% | 41% | 37.8% |
| Bone crisis | 35.6% | 13.6% | 14.5% | 2.6% |
| Infiltration | 91% | 87.3% | 89.5% | 80.8% |
| Erlenmeyer | 89.7% | 73.1% | 87.7% | 71.1% |
| Infarcts | 60% | 43.8% | 59.6% | 48.9% |
| Necrosis | 43.6% | 32% | 42.7% | 28% |
| % Total of bone disease | 71% | 60.8% | 69.8% | 58.8% |

### 3.3.3. Argentine ancestor analyses in patients harboring RecNciI allele

The preponderance of Rec*Nci*I allele in nearly 50% of our national cohort is highest of any cohort described to date. We examined the origin of patients using questionnaires designed by experts in population genetics based on the origins of the families to the birthplace of the third

**Table 8**

Association between Rec*Nci*I allele and Argentine ancestor.

| | Without association | With association | P value |
|---|---|---|---|
| One Argentine ancestor, N = 80 | 24 (30%) | 56 (70%) | <0.001 |
| Two Argentine ancestors, N = 50 | 11 (22%) | 39 (48.8%) | <0.001 |

**Table 9**

Rec*Nci*I/N370S genotype and Argentine ancestors.

| | Without association | With association | P value |
|---|---|---|---|
| One Argentine ancestor, N = 80 | 20 (29.4%) | 48 (70.6%) | 0.001 |
| Two Argentine ancestors, N = 50 | 17 (34%) | 33 (66%) | 0.001 |

generation. Of the respondents, approximately 70% (n = 56) of patients reported at least one Argentine ancestor and 49% (n = 39) had two ancestors from Argentine in one paternal and/or maternal lineage.

### 3.3.4. Multivariate analysis

Variables included were: 1) Rec*Nci*I allele, 2) N370S/Rec*Nci*I genotype, 3) Rec*Nci*I/other genotype, and 4) Argentine ancestors. Taking as a dependent variable solitary bone lesion in the follow-up (model I) and severe multiple lesions at follow-up (model II) we found, in both models, that N370S/Rec*Nci*I genotype and a presence of Rec*Nci*I allele was significantly associated with severe bone lesions at follow-up (p = 0.017) as well as with irreversible severe lesions (necrosis and acute and chronic infarcts) (p = 0.004). Argentine ancestors were not a significant variable (p = 0.059) in either of the two models.

Interestingly, we found a significant correlation between the presence of a Rec*Nci*I allele and the presence of, one Argentine ancestor (p < 0.001) or both Argentine ancestors (p < 0.001) (Table 8). Notably, there was also a significant correlation between N370S/Rec*Nci*I genotype and at least one Argentine ancestor (p = 0.001) or two Argentine ancestors (p = 0.001) (Table 9). There was no significant correlation between history of Argentine ancestors and bone lesions at diagnosis (p = 0.470) or at follow-up (p = 0.549).

### 3.3.5. Novel GD pathogenetic variants

Using our comprehensive genotyping strategy, we found rare GD

**Table 10**

Novel Pathogenic variants found in PacBio Argentina patients.

| Novel Var. | Numbers in GD | ArgGD ID. | PolyPhen Score | Position/Change | Clinical Significance | Genotype |
|---|---|---|---|---|---|---|
| L-17SfsX36 | 1 | #31 | | 1:155210462 (Del:AG > A) | Pathogenic | L-17SfsX36/N370S |
| R48GfsX4 (c.259DelC) | 1 | #186 | | 1:155209724 (Del:CG > C) | Pathogenic | R48GfsX4 (c.259DelC)/N370S |
| D218A | 1 | #178 | 0.892 | 1:155207361 (T > G) | Uncertain | L444P/D218A |
| P332L | 1 | #147 | 1 | 1:155206148 (G > A) | Uncertain | P332L/N370S |
| W348R | 1 | #10 | 0.863 | rs765182863/1:155206101 (A > T) | Pathogenic | L444P/W348R |
| L372P | 1 | #59 | 0.987 | 1:155205628 (A > G) | Pathogenic | F411I/L372P |
| Y313H + V375L G > C | 3 | #86, #153, #154 | 0.795 | rs398123528/1:155205620 (C > G) (rs74598136)/1:155205541 (G > C) | Pathogenic | N370S/Y313H + V375L G > C |
| P401R | 1 | #93 | 1 | | Pathogenic | P401R/N370S |
| S424R | 1 | #190 | 1 | 1:155205102 (G > C) | Uncertain | S424R/N370S |
| F426S | 1 | #58 | 1 | 1:155205097 (A > G) | Uncertain | F426S/N370S |
| S488IfsX38 | 1 | #165 | | 1:155204819 (Ins:G > GTAGC) | Pathogenic | S488IfsX38/N370S |
| L461P | 1 | #105 | 0.998 | 1:155204992 (A > G) | Uncertain | L461P/N370S |

pathogenetic variants in 14 patients (9.6%), which have not been previously reported. We assessed their damaging properties *in silico*, as shown in the Table 10. Phenotype of patients with these pathogenetic variants is severe, with early childhood presentations (mean age of diagnosis: 6.9 years) and almost 64% prevalence of advanced bone disease (Table 10).

## 4. Discussion

Comprehensive *GBA1* genotyping remains challenging [10,11,25,26], due to the vast multiplicity of disease mutations and highly homologous pseudogene in proximity, that harbors numerous disease mutations. Several variants normally present in the *GBA1P* sequence, when they occur in *GBA1*, generally cause severe forms of GD, *e.g.,* L444P, complex alleles with multiple pseudogene variants in tandem, D409H, and RecDelta55 deletion, etc. The occurrences of such disease mutations around the world have been attributed to the propensity of *GBA1* for gene conversion events. Several approaches have been developed to analyze the coding regions of *GBA1* by Sanger sequencing on amplicons generated using primers designed to exploit limited differences in the *GBA1* and *GBA1P* sequences. While the Sanger sequencing approach is adequate for most clinical applications, its short-read (from 500bp to 1000bp) and low-throughput limitations may hinder its application for refined genetic counselling and for study of large cohorts. Moreover, this standard approach does not allow for phasing of the pathogenic variants. Next generation sequencing (NGS) platforms [27], utilizing short read lengths (from 25 bp to 400 bp), are also inadequate for the precise identification of *GBA1* variants, especially for complex alleles and structural variants (SVs) [11,26].

Third-generation sequencing (TGS) technologies display the capability to produce read lengths higher than 10,000 bp (typically 5,000 bp to 20,000 bp), offering the advantage of identifying repetitive and complex genome regions. Superior to short-read sequencing and arrays, long-read sequencing technologies have now reached a level of accuracy and yield that allows their application to variant detection on a scale of tens to thousands of samples [28]. Therefore, TGS is a promising tool for identifying recombinant alleles, SVs and phasing bi-allelic GD mutations.

However, TGS using Nanopore [29] or PacBio [30] reportedly shows results with low accuracy reads. So far, there is only one such algorithm designed to exploit *GBA1* long-read data, using Nanopore technology studied on a relatively small cohort to compare with Sanger sequencing results [31].

A high proportion of the Rec*Nci*I complex allele has been noticed in an earlier Argentinian GD report [21]. Our results confirmed the superiority of SMRT to NGS in detecting the Rec*Nci*I allele, and to Sanger sequencing in scaled-up bi-allelic sequencing of the whole *GBA1* gene.

In the Argentinian GD cohort, bone disease is highly prevalent.

Advanced bone lesions are frequent both at diagnosis (71%) and at follow-up (69.8%), underscoring the irreversible nature of bone disease such as osteonecrosis. There was large gap between diagnosis in childhood and initiation of treatment as young adults which may contribute to this finding. Previous studies have demonstrated that the maximal impact of treatment in prevention of irreversible bone lesions occurs when ERT is initiated within 2 years of diagnosis compared to when there is a larger interval between diagnosis and starting ERT [32,33]. There are no published descriptions of a comprehensive genotype/ phenotype correlation regarding bone manifestations in GD patients, hence our cohort provides valuable insight.

We found that the high frequency of advanced bone lesions in Argentinian GD patients is correlated with an unprecedentedly high frequency of N370S/N370S genotype. Additional contributors to advanced bone disease in the Argentinian GD phenotype include prolonged intervals between childhood onset of disease, and long gaps between diagnosis and initiation of ERT in adulthood, which likely promotes irreversible bone lesions such as osteonecrosis and bone marrow fibrosis around focal collections of lipid-laden macrophages, 'Gaucheromas'. ERT targets tissue macrophages via the mannose receptor and exhibits very high uptake by the liver and spleen, hindering delivery to the bone marrow compartment [34,35].

In 2016, the GADTEG identified five unfavorable prognostic factors for advanced bone manifestations of Gaucher disease, viz a viz, history of splenectomy, diagnostic delay and long gap between diagnosis and initiation of ERT, poor compliance and suboptimal dose of ERT [15].

Our study found that the Argentinian GD population harbors the highest burden of Rec*Nci*I GBA1 mutation accounting for 52.7% of all disease alleles ever reported. This is higher than those reported in the ICGG Registry and in a smaller cohort study from Argentina (1.9% and 21%, respectively) [21].

The high prevalence of Rec*Nci*I allele found in our study in Argentina is the most accurate ascertainment to date, enabled by the SMRT strategy to decipher GBA locus. Previous studies reporting lower prevalence of this disease allele may have under-estimated the prevalence due to screening for only common variants.

There are several important findings in our study. First, our patient population harbors high prevalence of the Rec*Nci*I *GBA1* complex allele, which is associated with severe early childhood onset of GD1. Second, there is large gap between the onset of GD1 symptoms and diagnosis to initiation of ERT. Therefore, by the time of initiation of ERT, there is already high burden of irreversible bone manifestation, attenuating the impact of therapy in the skeletal compartment. Third, the high prevalence of the Rec*Nci*I mutation highlights that the gene conversion event (s) giving rise to this variant could disrupt another gene in the tightly packed GBA locus that could impact the skeletal phenotype. This topic merits further investigation for a potential candidate gene for severe skeletal manifestations. Finally, we found about 10% of disease variants

were not described before, hence contributing to the catalogue of *GBA1* mutations.

It is intriguing that we found a strong association of the N370S/Rec*Nci*I genotype with Argentinian ancestry. It seems likely that the Rec*Nci*I mutation is founder mutation in the Argentine GD patient population. Generally the Argentinian population is said to be comprised, in a greater proportion, by immigrants, while only a minor proportion with Argentinian ancestors. However, in 2012, Avena et al. [36], demonstrated, after analyzing national surveys and census, that 65% of Argentinians have European ancestors, 31% have Argentinian ancestors, and 4% have African. These frequencies vary according to geographical region, for example, the proportion of indigenous Americans is higher in the northwest (60%) and south (50%) [12,13]. Our analysis showed that 54.8% of GD patients have at least one Argentinian ancestor, and 30.5% have two ancestors in any paternal/maternal lineage.

Univariate and multivariate analyses showed a statistically significant correlation between the presence of the Rec*Nci*I allele and at least one Argentine ancestor ($p < 0.001$) or two Argentinian ancestors ($p < 0.001$). This demonstrates a strong relationship between the presence of Argentine ancestors and the Rec*Nci*I allele as well as the Rec*Nci*I/N370S genotype ($p = 0.001$). This high incidence of Argentinian ancestors should explain the high frequency of Rec*Nci*I allele.

A limitation of our study was that, of a total of 192 samples, 146 (76%) were successfully genotyped by SMRT sequencing but for 46 samples full genotypes could not be ascertained. However, having 146 patients comprehensively genotyped enabled a robust genotype/phenotype correlation study. We are currently optimizing and adapting our strategy to achieve successful genotyping in all samples.

## 5. Conclusions

Complete *GBA1* gene sequencing using long-read SMRT plus short-read validation sequencing provides a rapid scaled-up mutation analysis, with essential information that cannot be obtained by Sanger or NGS sequencing alone. In our population, a high frequency of the Rec*Nci*I allele (52.7%) and Rec*Nci*I/N370S genotype (46.6%) was detected, data not captured in the international registry. Using a questionnaire to assess familial origin, we show that Argentina GD patients have at least one Argentine ancestor (54.8%) or two Argentine ancestors (30.5%). Moreover, there was a statistically significant correlation between the presence of a Rec*Nci*I allele and one Argentine ancestor ($p < 0.001$) or two Argentine ancestors ($p < 0.001$). Our study indicates that the Rec*Nci*I allele and Rec*Nci*I/N370S genotype are significantly associated with severe bone manifestations at presentation and during follow-up on ERT ($p = 0.017$).

## References

[1] E. Beutler, G. Grabowski, The mucopolysaccharidoses, in: C. Scriver, A. Beaudet, W. Sly, D. Valle (Eds.), The Metabolic and Molecular Bases of Inherited Disease, McGraw Hill, New York, 2001, pp. 3635–3668.

[2] G.A. Grabowski, Gaucher disease and other storage disorders, Hematology Am Soc Hematol Educ Program. 2012 (2012) 13–18.

[3] G. Drelichman, Escobar N. Fernández, N. Basack, et al., Actualización del consenso argentino de enfermedad de gaucher: grupo argentino para el diagnóstico y tratamiento de la enfermedad de gaucher. *HEMATOLOGÍA* Volumen 19 Suplemento Enfermedad de Gaucher, 2015, pp. 4–51.

[4] V. Murugesan, W.L. Chuang, J. Liu, A. Lischuk, K. Kacena, H. Lin, G.M. Pastores, R. Yang, J. Keutzer, K. Zhang, P.K. Mistry, Glucosylsphingosine is a key biomarker of Gaucher disease, Am. J. Hematol. 11 (2016) 1082–1089.

[5] G. Bultron, K. Kacena, D. Pearson, M. Boxer, R. Yang, S. Sathe, G. Pastores, P. K. Mistry, The risk of Parkinson's disease in type 1 Gaucher disease, J. Inherit. Metab. Dis. 33 (2) (2010) 167–173.

[6] M. Horowitz, S. Wilder, Z. Horowitz, O. Reiner, T. Gelbart, E. Beutler, The human glucocerebrosidase gene and pseudogene: structure and evolution, Genomics. 4 (1) (1989) 87–96.

[7] S.L. Winfield, N. Tayebi, B.M. Martin, E.I. Ginns, E. Sidransky, Identification of three additional genes contiguous to the glucocerebrosidase locus on chromosome 1q21: implications for Gaucher disease, Genome Res. 7 (1997) 1020–1026.

[8] V. Koprivica, D.L. Stone, J.K. Park, M. Callahan, A. Frisch, I.J. Cohen, N. Tayebi, E. Sidransky, Analysis and classification of 304 mutant alleles in patients with type 1 and type 3 Gaucher disease, Am. J. Hum. Genet. 66 (6) (2000) 1777–1786.

[9] T. Vanier, R. Froissart, Diagnostic testing (enzyme and mutational analysis), in: G. A. Grabowski, 1st (Eds.), Advances in Gaucher Disease: Basic and Clinical Perspectives, Future Medicine, London, 2013, pp. 170–181.

[10] N. Tayebi, B.K. Stubblefield, J.K. Park, E. Orvisky, J.M. Walker, M.E. LaMarca, E. Sidransky, Reciprocal and nonreciprocal recombination at the glucocerebrosidase gene region: implications for complexity in Gaucher disease, Am. J. Hum. Genet. 72 (2003) 519–534.

[11] S. Zampieri, S. Cattarossi, B. Bembi, A. Dardis, GBA1 Analysis in Next-Generation Era: Pitfalls, Challenges, and Possible Solutions, J. Mol. Diagnost. 19 (5) (2017) 733–741.

[12] N. Eyal, S. Wilder, M. Horowitz, Prevalent and rare pathogenetic variants among Gaucher patients, Gene. 96 (1990) 277–283.

[13] International Collaborative Gaucher Group (ICGG) Gaucher Registry, Standard Report Population Report: Annual Report Based on 26 June, 2014.

[14] Data Analysis Request Report ICGG Gaucher Registry DAR Committee Date: 5/13/2019.

[15] G. Drelichman, Escobar N. Fernández, B. Soberón, G. Aguilar, M. Larroude, et al., Skeletal involvement in Gaucher disease: an observational multicenter study of prognostic factors in the Argentine Gaucher disease patients, Am. J. Hematol. 91 (2016) E448–E453.

[16] T.H. Taddei, K.A. Kacena, M. Yang, R. Yang, A. Malhotra, M. Boxer, K.A. Aleck, G. Rennert, G.M. Pastores, P.K. Mistry, The underrecognized progressive nature of N370S Gaucher disease and assessment of cancer risk in 403 patients, Am. J. Hematol. 84 (4) (2009) 208–214.

[17] J. Eid, A. Fehr, J. Gray, et al., Real-time DNA sequencing from single polymerase molecules, Science. 323 (2009) 133–138.

[18] G. Drelichman, Escobar N. Fernández, P.K. Mistry, et al., Correlación fenotípica /genotípica de la Prevalencia del genotipo N370S/Rec*Nci*I en una corte argentina de 197 pacientes con Enfermedad de Gaucher. *Poster durante el LATAM round table, 2018.* Buenos Aires Argentina.

[19] T. Saleem, M. Hassan, A. El-Abd Ahmed, A. Sayed, et al., Clinical and genetic assessment of pediatric patients with Gaucher disease in Upper Egypt, Egypt. J. Med. Human Genet. 18 (2017) 249.255.

[20] P. Alfonso, A. Cenarro, J. Pérez Calvo, P. Giraldo, et al., Mutation prevalence among 51 unrelated spanish patients with gaucher disease: identification of 11 novel pathogenetic variants, Blood Cells Mol. Dis. 27 (2001) 882–8912.

[21] B. Cormand, T.L. Harboe, L. Gort, C. Campoy, M. Blanco, N. Chamoles, A. Chabas, L. Vilageliu, D. Grinberg, Mutation analysis of Gaucher disease patients from Argentina high Prevelence of the Rec*Nci*I mutation, Am. J. Med. Genet. 80 (1998) 343–351.

[22] J. Sheth, R. Bhavsar, M. Mistri, D. Pancholi, A. Bavdekar, A. Dalal, P. Ranganath, K. M. Girisha, A. Shukla, S. Phadke, et al., Gaucher disease: single gene molecular characterization of one-hundred Indian patients reveals novel variants and the most prevalent mutation, BMC Med. Genet. 20 (2019) 31.

[23] S.P. Basgaluppa, V. Altmanna, F.P. Vairob, et al., Is there any difference in GBA1 allele frequencies depending on the region of Brazil. WORLD SymposiumTM, Mol. Genet. Metab. 126 (2019) S2–S6.

[24] G.A. Grabowski, E.H. Kolodny, N.J. Weinreb, B.E. Rosenbloom, A. Prakash-Cheng, P. Kaplan, J. Charrow, G.M. Pastores, P.K. Mistry, Gaucher disease: phenotypic and genetic variation, in: D.L. Valle, S. Antonarakis, A. Ballabio, A.L. Beaudet, G. A. Mitchell (Eds.), The Online Metabolic and Molecular Bases of Inherited Disease, McGraw Hill, 2019. Accessed October 30.

[25] D. Mandelke, R.J. Schmidt, A. Ankala, G.K. McDonald, et al., Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing, Genet. Med. 18 (2016) 1282–1289.

[26] E.G. Woo, N. Tayebi, E. Sidransky, Next-generation sequencing analysis of GBA1: the challenge of detecting complex recombinant alleles, Front. Genet. 12 (2021) 684067.

[27] S. Goodwin, J.D. McPherson, W.R. McCombie, Coming of age: ten years of next-generation sequencing technologies, Nat. Rev. Genet. 17 (6) (2016) 333–351.

[28] W. De Coster, M.H. Weissensteiner, F.J. Sedlazeck, Towards population-scale long-read sequencing, Nat. Rev. Genet. 22 (9) (2021) 572–587.

[29] R.R. Wick, L.M. Judd, K.E. Holt, Performance of neural network base calling tools for Oxford Nanopore sequencing, Genome Biol. 20 (1) (2019) 129.

[30] S. Ardui, A. Ameur, J.R. Vermeesch, M.S. Hestand, Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics, Nucleic Acids Res. 46 (5) (2018) 2159–2168.

[31] M. Leija-Salazar, F.J. Sedlazeck, M. Toffoli, S. Mullin, K. Mokretar, M. Athanasopoulou, A. Donald, R. Sharma, D. Hughes, A.H.V. Schapira, C. Proukakis, Evaluation of the detection of GBA missense mutations and other

variants using the Oxford Nanopore MinION, Mol. Genet. Genomic Med. 7 (3) (2019), e564.

[32] P.K. Mistry, P. Deegan, A. Vellodi, J.A. Cole, M. Yeh, N.J. Weinreb, Timing of initiation of enzyme replacement therapy after diagnosis of type 1 Gaucher disease: effect on incidence of avascular necrosis, Br. J. Haematol. 147 (4) (2009) 561–570.

[33] D. Hughes, P. Mikosch, N. Belmatoug, F. Carubbi, T. Cox, O. Goker-Alpan, A. Kindmark, P. Mistry, L. Poll, N. Weinreb, P. Deegan, Gaucher disease in bone: from pathophysiology to practice, J. Bone Miner. Res. 34 (2019) 996–1013.

[34] P.K. Mistry, E.P. Wraight, T.M. Cox, Therapeutic delivery of proteins to macrophages: implications for treatment of Gaucher's disease, Lancet. 348 (1996) 1555–1559.

[35] N.W. Barton, R.O. Brady, J.M. Dambrosia, A.M. Di Bisceglie, S.H. Doppelt, S. C. Hill, H.J. Mankin, G.J. Murray, R.I. Parker, C.E. Argoff, et al., Replacement therapy for inherited enzyme deficiency–macrophage-targeted glucocerebrosidase for Gaucher's disease, N. Engl. J. Med. 324 (21) (1991) 1464–1470.

[36] S. Avena, M. Via, E. Ziv, E.J. Pérez-Stable, C.R. Gignoux, C. Dejean, S. Huntsman, G. Torres-Mejía, J. Dutil, J.L. Matta, K. Beckman, E.G. Burchard, M.L. Parolin, A. Goicoechea, N. Acreche, M. Boquet, C. Ríos Part Mdel, V. Fernández, J. Rey, M. C. Stern, R.F. Carnese, L. Fejerman, Heterogeneity in genetic admixture across different regions of Argentina, PLoS One 7 (2012) e34695.