

# Determining the prevalence of McArdle disease from gene frequency by analysis of next-generation sequencing data

Mauricio De Castro, MD<sup>1</sup>, Jennifer Johnston, PhD<sup>1</sup> and Leslie Biesecker, MD<sup>1</sup>

**Purpose:** McArdle disease is one of the most common glycogen storage disorders. Although the exact prevalence is not known, it has been estimated to be 1 in 100,000 patients in the United States. More than 100 mutations in *PYGM* have been associated with this disorder. McArdle disease has significant clinical variability: Some patients present with severe muscle pain and weakness; others have only mild, exercise-related symptoms.

**Methods:** Next-generation sequencing data allow estimation of disease prevalence with minimal ascertainment bias. We analyzed gene frequencies in two cohorts of patients based on exome sequencing results. We categorized variants into three groups: a curated set of published mutations, variants of uncertain significance, and likely benign variants.

**Results:** An initial estimate based on the frequency of six common mutations predicts a disease prevalence of 1/7,650 (95% confidence interval (CI) 1/5,362–1/11,108), which greatly deviates from published estimates. A second method using the two most common mutations predicts a prevalence of 1/42,355 (95% CI 1/24,536–1/76,310) in Caucasians.

**Conclusions:** These results suggest that the currently accepted prevalence of McArdle disease is an underestimate and that some of the currently considered pathogenic variants are likely benign.

*Genet Med* advance online publication 5 March 2015

**Key Words:** allele frequency; McArdle disease; next-generation sequencing; prevalence; *PYGM*

## INTRODUCTION

McArdle disease (glycogen storage disease type V) is an inherited disorder of glycogen metabolism that exclusively affects skeletal muscle. It was initially described in 1951 by British physician Brian McArdle, who described a patient with exercise intolerance who failed to produce lactate. Symptoms consist of rapid fatigue, myalgia, and cramping associated with exercise. There is clinical variability: Some patients have mild symptoms (fatigue or poor stamina) related to exercise,<sup>1</sup> whereas others have more pronounced proximal muscle weakness.<sup>2</sup> A fatal, rapidly progressive neonatal form with widespread muscle weakness has also been reported.<sup>3</sup> A classic finding in patients with the disease is the rapid improvement of symptoms with rest (the so-called second-wind phenomenon). In mildly to moderately affected patients the clinical diagnosis requires a high degree of suspicion, especially in older patients in whom the only symptom may be exercise intolerance. The diagnosis is confirmed with identification of biallelic pathogenic variants in the *PYGM* gene that encodes for the muscle phosphorylase protein, the only gene known to be associated with McArdle disease.<sup>4</sup> If the results are unclear, muscle biopsy with measurement of phosphorylase enzyme activity can be helpful. A recently described less invasive method involves the use of antibodies to determine the expression of *PYGM* in white blood cells.<sup>5</sup>

The prevalence of McArdle disease has been reported to be 1 in 100,000 in the United States,<sup>6</sup> at least 1 in 170,000 in Spain,<sup>7</sup> and

1 in 350,000 in the Netherlands.<sup>8</sup> In Spain and the Netherlands the calculations were based on the number of affected individuals from national McArdle disease registries. Because McArdle disease can cause mild symptoms, it is possible that an estimate of prevalence based on ascertainment by clinical presentation to a metabolic disease expert could severely underestimate the prevalence. Access to exome sequencing data allows us to estimate the prevalence of this disorder based on carrier frequency using the Hardy-Weinberg equilibrium, reducing the bias associated with clinical ascertainment.

## MATERIALS AND METHODS

We evaluated variant call data from the ClinSeq cohort ( $n = 951$ ) and the National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project (ESP) ( $n = 4,297$  European Americans (EAs) and 2,201 African Americans). The ClinSeq cohort is composed of 951 patients, predominantly of Caucasian descent, ascertained for their family history of cardiovascular disease; participants are otherwise healthy and were not selected for known muscular conditions or symptoms. The ESP cohort is composed of several groups of patients: Most have a personal or family history of cardiovascular or pulmonary disease, some are healthy controls, whereas others are affected with hyperlipidemia, cardiovascular disease, or other associated conditions. None of the cohorts were selected for primary muscle disease. We first analyzed variant calls for the *PYGM* gene in the ClinSeq database (materials and methods

<sup>1</sup>National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA. Correspondence: Mauricio De Castro ([Mauricio.decastro-pretelt@nih.gov](mailto:Mauricio.decastro-pretelt@nih.gov))

Submitted 19 November 2014; accepted 13 January 2015; advance online publication 5 March 2015. doi:10.1038/gim.2015.9

for the ClinSeq study are described elsewhere<sup>9</sup>); DNA isolation, library preparation, capture, sequencing, and alignment, as well as base calling were performed as described in previous reports.<sup>10</sup> *PYGM* variant analysis was performed in VarSifter version 1.6 (ref. 11). Variants were filtered for mutation type and population frequency.

Variants that met population frequency (minor allele frequency (MAF) <0.5% in ClinSeq and ESP) and quality filters were further classified by cross-referencing them with mutations in the Human Gene Mutation database. The pathogenicity of these variants was evaluated by reviewing publications with clinical, functional, and/or genetic data. To be considered pathogenic, a variant had to be reported in the literature in a patient with classical manifestations of the disease with compatible ancillary testing (e.g., characteristic muscle biopsy, absent muscle phosphorylase levels, or second-wind phenomenon on treadmill testing) and the identification of biallelic variants in *PYGM*. The phase of the variants had to be known and appropriate Mendelian segregation confirmed. For variants not described in the literature, further classification was limited to allele frequency in the general population and in silico model predictions: PolyPhen-2, SIFT,<sup>12</sup> and combined annotation-dependent depletion score.<sup>13</sup> Variants that did not meet our criteria for classification as pathogenic, were predicted to be deleterious by all four models, and had a MAF <0.5% were considered to be variants of uncertain significance. Variants with a MAF >0.5% or unpublished variants predicted to be benign by one or more in silico models were considered to be likely benign.

Statistical analysis for the 95% confidence intervals (CIs) was performed using the exact binomial method based on the beta distribution as described by Clopper and Pearson.<sup>14</sup> Variants p.Arg50\* and p.Gly205Ser in the ClinSeq cohort were verified by Sanger sequencing, but this was not possible for variants in the ESP cohort.

## RESULTS

The ClinSeq data were evaluated first. Two variants were excluded (p.Thr395Met and p.Arg414Gly) because they were above the frequency limit. We were left with 59/951 ClinSeq participants, who had among them 27 *PYGM* variants (Table 1). No participant had two minor alleles. Fifteen participants were heterozygous for one of six published mutations. Thirteen participants were heterozygous for 12 variants of uncertain significance, and 31 participants were heterozygous for 9 likely benign variants. We then evaluated the ESP data set for EAs for the mutations that we identified in ClinSeq. In the ESP EA data set, 105 participants were heterozygous for one of the six published mutations. Twenty-six participants were heterozygous for 6 of the 12 variants of uncertain significance, and 64 participants were heterozygous for 1 of the 9 likely benign variants.

To increase power, we combined our results with data from the ESP project, which yielded 5,248 exomes. Although there were no homozygotes for any of these variants in the NHLBI ESP, we could not exclude compound heterozygosity because that database does not provide these data.

A total of 27 variants among 59 individuals from the 951 participants in ClinSeq were considered. Six of these 27 variants have been claimed to be pathogenic in prior publications. These six variants, which were present in a total of 15 participants—for a MAF of 0.00789—predict a disease prevalence of 1/16,080 (95% CI 1/5,940–1/51,163). Because the CIs of this estimate were so large, we expanded our data set by analyzing the NHLBI ESP EA cohort, for a total of 5,248 individuals. Between the two data sets, there were a total of 120 participants with one of the six pathogenic variants, for a MAF of 0.0114, which predicts a prevalence of 1/7,650 (95% CI 1/5,362–1/11,108).

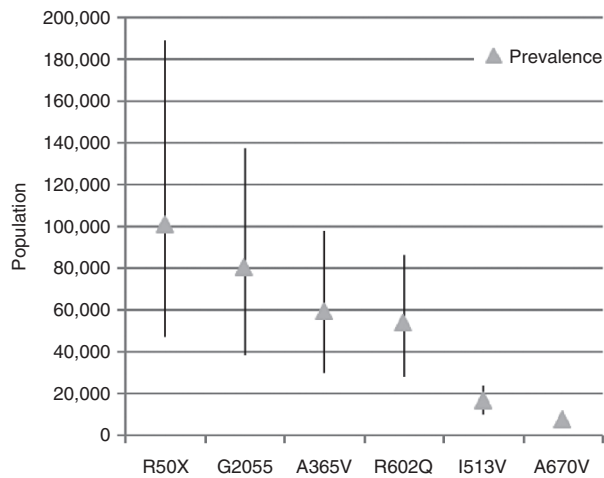
Given the discrepancy with published estimates, we critically evaluated the evidence supporting the pathogenicity of the variants and rank ordered them from most evidence to least evidence. The p.Arg50\* variant was the highest ranked because it is present in large numbers of affected individuals compared with controls and has been shown to undergo nonsense-mediated decay in muscle tissue from patients with McArdle disease.<sup>15</sup> We calculated the predicted disease prevalence based on that variant alone. In the combined ClinSeq and ESP EA data, the MAF for this variant was 0.00313, which predicts a disease prevalence of 1/101,166 (95% CI 1/51,349–1/213,345). We then took the variant with the next most strong evidence, p.Gly205Ser, and added the frequencies of that variant to those of p.Arg50\* and estimated the frequency of the disease. This variant is located in a critical region for tetramerization of the *PYGM* enzyme, and mutations in residue 205 have been shown to lead to misfolding of the protein in human cell lines.<sup>16</sup> The MAF of those two variants in the combined data set was 0.00352, which predicts a disease prevalence of 1/80,478 (95% CI 1/42,407–1/162,198). This series of calculations was continued for all six mutations, showing that the previous estimated prevalence of the disease is accounted for by only the p.Arg50\* variant and that the upper 95% CI of our calculations falls to about 1/100,000 when accounting for only three mutations (Figure 1). Indeed, by using all six of the published variants identified in ClinSeq, the predicted disease frequency is far more common than prior estimates. Although there are more than 100 reported *PYGM* mutations, we calculated a predicted disease prevalence of 1/7,650 (95% CI 1/5,362–1/11,108) using only 6 published mutations.

To provide yet another approach to these estimates, we calculated the prevalence by deriving the total fraction of all other pathogenic alleles using data from affected patients.<sup>17</sup> First, we tabulated the total mutation burden for the two most common mutations: p.Arg50\* and p.Gly205Ser. The former is the most common mutation in McArdle disease, with the actual prevalence of the mutation varying among populations. The estimated prevalence of p.Arg50\* among patients with McArdle disease in the United States is 63%.<sup>1,18</sup> p.Gly205Ser is the second most common mutation in Europe and the United States, comprising about 9% of pathogenic alleles. The combination of these two alleles should account for 72% of alleles for McArdle disease in EAs in the United States. The prediction using both allele frequencies, and assuming this accounted for 72% of

**Table 1** Variants evaluated in this study

Complementary DNA	AA change	Individuals with variant in ClinSeq (n)	Individuals with variant in EA ESP (n)	Total population of ClinSeq + ESP EA (n)	Individuals with variant in AA ESP (n)	Variant classification	Published in the literature	SIFT	PolyPhen2	CADD score
c.148C>T	p.R50*	6	27	33	2	Pathogenic	Yes	LOF	LOF	40
c.613G>A	p.G205S	1	3	4	0	Pathogenic	Yes	Damaging	Probably damaging	36
c.1094C>T	p.A365V	2	4	6	2	Pathogenic	Yes	Damaging	Probably damaging	23.6
c.1537A>G	p.I513V	1	35	36	3	Pathogenic	Yes	Tolerated	Benign	10.83
c.1805G>A	p.R602Q	1	0	1	0	Pathogenic	Yes	Damaging	Probably damaging	36
c.2009C>T	p.A670V	4	35	39	6	Pathogenic	Yes	Damaging	Probably damaging	35
c.100C>T	p.R34W	1	0	1	0	VOUS	No	Damaging	Probably damaging	35
c.209G>A	p.R70H	1	0	2	0	VOUS	No	Damaging	Probably damaging	35
c.482G>A	p.R161H	1	0	0	0	VOUS	No	Damaging	Probably damaging	34
c.832C>T	p.R278C	1	0	0	0	VOUS	No	Damaging	Probably damaging	22.1
c.848A>G	p.N283S	2	12	14	2	VOUS	No	Damaging	Probably damaging	24.2
c.1160G>A	p.R387H	1	0	1	0	VOUS	No	Damaging	Probably damaging	31
c.1558C>T	p.R520C	1	2	3	2	VOUS	No	Damaging	Probably damaging	22.8
c.1885G>T	p.D629Y	1	1	2	5	VOUS	No	Damaging	Probably damaging	24.89
c.2083G>A	p.G695R	1	0	1	0	VOUS	No	Damaging	Probably damaging	28.3
c.2446C>T	p.R816C	1	1	2	0	VOUS	No	Damaging	Probably damaging	20.9
c.2467C>T	p.R823W	1	1	2	0	VOUS	No	Damaging	Probably damaging	19.07
c.2500C>T	p.R834C	1	0	1	0	VOUS	No	Damaging	Probably damaging	20.4

AA, African American; CADD, combined annotation-dependent depletion; EA, European American; LOF, loss of function; VOUS, variant of uncertain significance.



**Figure 1 Ordinal mutation prevalence.** Prevalence estimates with 95% confidence intervals starting with the mutation with the most evidence for pathogenicity (p.Arg50\*) and subsequently adding published mutations in decreasing order of evidence for pathogenicity.

causative alleles, resulted in a prevalence of 1/42,355 (95% CI 1/24,536–1/76,310), which does not overlap with the currently estimated prevalence.

## DISCUSSION

These data suggest that McArdle disease is significantly more common among European-descended Americans than the currently accepted 1/100,000 prevalence, and we conclude that the disorder is at least twice as common, in the range of 1/50,000. There are two potential explanations: (i) McArdle disease is under diagnosed, and/or (ii) the penetrance of some of the variants in McArdle disease is overestimated. It is possible that some mutations in *PYGM* are not fully penetrant, thus overestimating the prevalence when calculating based on combined allele frequencies. We believe this is one of the strengths of the calculations that use only the two most common mutations (p.Arg50\* and p.Gly205Ser), which all evidence to date suggests are fully penetrant. That both methods predict a higher frequency supports our thesis. Expressivity should also be considered—if there is a wider range of expressivity than currently appreciated, there could be many patients who have a very mild form of this disease. This would be just as interesting and important—we suggest that there could be present in a patient a very mild form of McArdle disease, which is not diagnosed as such but has significant implications for exercise tolerance. A separate issue to consider is the possibility that many of the variants in McArdle disease are actually benign, which would erroneously increase the calculated prevalence (for instance, the variant p.Ile513Val seems to be just as common as p.Arg50\* in certain populations). We do not believe this to be valid because our higher prevalence is supported by the method of extrapolating from only two variants that are essentially certain to be pathogenic, which makes questions

of individual pathogenicity assessment of other variants irrelevant. Nearly all variants other than p.Arg50\* and p.Gly205Ser would have to be benign for the 95% CI of our estimates to overlap with the current prevalence estimate, which we think is an unreasonable hypothesis.

It is possible that some mutations in *PYGM* cause a very clinically mild phenotype of McArdle disease. This has been described for autosomal recessive metabolic disorders such as biotinidase deficiency,<sup>19</sup> pyruvate kinase deficiency,<sup>17</sup> and Gaucher disease, but not for McArdle disease. Because McArdle disease is a condition with high clinical variability, symptoms can go unrecognized for many years before being diagnosed. It is possible that many affected patients develop an aversion to anaerobic exercise that does not limit their life enough to seek a diagnosis, and because of this they are not included in current prevalence estimates.

There are some limitations to this approach. We assumed that McArdle disease is a monogenic condition and all variants can be accounted for by looking at *PYGM*. If locus heterogeneity were a possibility for McArdle disease, then the prevalence of mutations would be higher than those we are suggesting here. A second limitation is that we are not able to ascertain the phase of the variants for the NHLBI ESP data set. Given that our estimates of prevalence are much higher than the inverse of the NHLBI ESP data set, we think this is unlikely to be an issue.

Finally, it is important to point out the technical limitations of identifying variants from next-generation sequencing data. Appropriate depth of coverage, deep intronic mutations, mutations in the promoter region, and inability to detect large deletions or duplications would lead to underascertainment of pathogenic variants. However, such an error would again make our estimate conservative, and the disease would be more common than we predict.

The estimation of disease frequency based on patients who present to specialty clinics is biased toward those with typical, recognizable, and more severe presentations. We predict that as sequencing is applied more widely in the clinic and in larger research cohorts, undiagnosed individuals with biallelic mutations in *PYGM* will be identified. This approach of genome-driven ascertainment (as opposed to phenotype-driven ascertainment) mitigates the inherent ascertainment bias toward more severe presentations. Identifying patients by mutations and following that with clinical research will be important to elucidate the possible associated phenotype; this has been termed hypothesis-generating clinical research.<sup>20</sup> Such identifications will allow a better appreciation of the true spectrum of clinical phenotypes associated with variation in this gene. We predict that a substantial number of such identified individuals will have abnormal biochemistry and exercise tolerance and that the full delineation of this phenotype will become a component of predictive medicine.

## ACKNOWLEDGMENTS

We thank Neal Oden for his invaluable assistance with the statistical calculations.

## DISCLOSURE

L.B. is an unpaid consultant to the Illumina Corporation. He also receives royalties from Genentech and Amgen. The other authors declare no conflict of interest.

## REFERENCES

1. el-Schahawi M, Tsujino S, Shanske S, DiMauro S. Diagnosis of McArdle's disease by molecular genetic analysis of blood. *Neurology* 1996;47:579–580.
2. Wolfe GI, Baker NS, Haller RG, Burns DK, Barohn RJ. McArdle's disease presenting with asymmetric, late-onset arm weakness. *Muscle Nerve* 2000;23:641–645.
3. DiMauro S, Hartlage PL. Fatal infantile form of muscle phosphorylase deficiency. *Neurology* 1978;28:1124–1129.
4. García-Consuegra I, Rubio JC, Nogales-Gadea G, et al. Novel mutations in patients with McArdle disease by analysis of skeletal muscle mRNA. *J Med Genet* 2009;46:198–202.
5. de Luna N, Brull A, Lucia A, et al. PYGM expression analysis in white blood cells: a complementary tool for diagnosing McArdle disease? *Neuromuscul Disord* 2014;24:1079–1086.
6. Haller RG. Treatment of McArdle disease. *Arch Neurol* 2000;57:923–924.
7. Lucia A, Ruiz JR, Santalla A, et al. Genotypic and phenotypic features of McArdle disease: insights from the Spanish national registry. *J Neurol Neurosurg Psychiatry* 2012;83:322–328.
8. van Alfen N, de Bie HJ, Wevers RA, Arenas J, van Engelen BG. The prevalence and genetic characteristics of McArdle's disease in the Netherlands. *Neuromuscul Disord* 2002;12:718–783.
9. NISC Comparative Sequencing Program; Biesecker LG, Mullikin JC, Facio FM, et al. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res* 2009;19:1665–1674.
10. Johnston JJ, Rubinstein WS, Facio FM, et al. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *Am J Hum Genet* 2012;91:97–108.
11. Teer JK, Green ED, Mullikin JC, Biesecker LG. VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer. *Bioinformatics* 2012;28:599–600.
12. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;11:863–874.
13. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46:310–315.
14. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934;26:404–413.
15. Nogales-Gadea G, Rubio JC, Fernandez-Cadenas I, et al. Expression of the muscle glycogen phosphorylase gene in patients with McArdle disease: the role of nonsense-mediated mRNA decay. *Hum Mutat* 2008;29:277–283.
16. Birch KE, Quinlivan RM, Morris GE. Cell models for McArdle disease and aminoglycoside-induced read-through of a premature termination codon. *Neuromuscul Disord* 2013;23:43–51.
17. Beutler E, Gelbart T. Estimating the prevalence of pyruvate kinase deficiency from the gene frequency in the general white population. *Blood* 2000;95:3585–3588.
18. Tsujino S, Shanske S, DiMauro S. Molecular genetic heterogeneity of myophosphorylase deficiency (McArdle's disease). *N Engl J Med* 1993;329:241–245.
19. Wolf B, Norrgard K, Pomponio RJ, et al. Profound biotinidase deficiency in two asymptomatic adults. *Am J Med Genet* 1997;73:5–9.
20. Biesecker LG. Hypothesis-generating research and predictive medicine. *Genome Res* 2013;23:1051–1053.