Genetics inMedicine | ARTICLE

Estimating prevalence for limb-girdle muscular dystrophy based on public sequencing databases

Wei Liu, BSc[1], Sander Pajusalu, MD, PhD[2,3,4], Nicole J. Lake, MSc, PhD[2,5], Geyu Zhou, BSc[1], Nilah Ioannidis, MPhil, PhD[6,7], Plavi Mittal, PhD[6,8], Nicholas E. Johnson, Msc, MD[9], Conrad C. Weihl, MD, PhD[10], Bradley A. Williams, PhD[6], Douglas E. Albrecht, PhD[6], Laura E. Rufibach, PhD[6] and Monkol Lek, BE, PhD[2]

Purpose: Limb-girdle muscular dystrophies (LGMD) are a genetically heterogeneous category of autosomal inherited muscle diseases. Many genes causing LGMD have been identified, and clinical trials are beginning for treatment of some genetic subtypes. However, even with the gene-level mechanisms known, it is still difficult to get a robust and generalizable prevalence estimation for each subtype due to the limited amount of epidemiology data and the low incidence of LGMDs.

Methods: Taking advantage of recently published exome and genome sequencing data from the general population, we used a Bayesian method to develop a robust disease prevalence estimator.

Results: This method was applied to nine recessive LGMD subtypes. The estimated disease prevalence calculated by this method was largely comparable with published estimates from epidemiological studies; however, it highlighted instances of possible underdiagnosis for LGMD2B and 2L.

Conclusion: The increasing size of aggregated population variant databases will allow for robust and reproducible prevalence estimates of recessive disease, which is critical for the strategic design and prioritization of clinical trials.

Genetics in Medicine (2019) 21:2512–2520; https://doi.org/10.1038/s41436-019-0544-8

Keywords: limb-girdle muscular dystrophy; rare disease; disease prevalence

INTRODUCTION

The limb-girdle muscular dystrophies (LGMDs) are a heterogeneous group of diseases, causing pelvic and shoulder girdle muscle weakness and wasting. There are currently 32 characterized subtypes[1] with a diverse range of clinical phenotypes, which show variability in age of onset, rate of progression, specific muscle wasting patterns, and involvement of respiratory and cardiac muscles. The subtypes are broadly categorized by their pattern of inheritance as either dominant (LGMD1A-I) or recessive (LGMD2A-X), with the majority being recessive, and can harbor either loss-of-function or missense pathogenic variants. The proteins encoded by LGMD disease genes have cellular functions including glycosylation and muscle membrane integrity, maintenance, and repair, which are a diverse range of mechanisms that when disrupted all result in muscle damage and degeneration.

Currently, an effective treatment does not exist for any LGMD subtype; however, promising gene therapy clinical trials have commenced for LGMD2E and additional subtypes are set to commence in 2019–2020.[2] Disease prevalence information is critical to the planning and prioritization of these clinical trials. Historically, the prevalence of rare diseases has largely been estimated from epidemiological surveys and patient registries.[3–5] However, it can be difficult to achieve an accurate and meaningful prevalence estimate for rare genetic disorders through these traditional approaches. Many patients with rare disease experience a delayed or incorrect diagnosis, which can be more pronounced for late onset, slowly progressing diseases such as LGMD,[6] leading to underestimation of prevalence. Differences in the diagnostic criteria used between studies, as well as changes to these over time, can make it difficult to directly compare estimates across studies. The specific population studies can also bias the

[1]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA; [2]Department of Genetics, Yale School of Medicine, New Haven, CT, USA; [3]Department of Clinical Genetics, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia; [4]Department of Clinical Genetics, United Laboratories, Tartu University Hospital, Tartu, Estonia; [5]Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne, Australia; [6]Jain Foundation, Seattle, WA, USA; [7]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA; [8]In-Depth Genomics, Bellevue, WA, USA; [9]Department of Neurology, Virginia Commonwealth University, Richmond, VA, USA; [10]Department of Neurology, Washington University School of Medicine, St. Louis,, MO, USA. Correspondence: Monkol Lek (monkol.lek@yale.edu)
These authors Contributed equally: Sander Pajusalu, Nicole J. Lake

Submitted 7 February 2019; accepted: 2 May 2019
Published online: 20 May 2019

prevalence estimate, and indeed the current published prevalence of LGMD subtypes can vary greatly between countries and even regions within countries.[7] The factors contributing to these regional differences include small sample size, foundervariants, and consanguinity rates—all of which can lead to increased incidences of LGMD in those populations.[8–10] In addition, the resources and training available to each health-care system can contribute to regional variability. Improved methods for quantifying the prevalence of rare genetic disorders such as LGMD are thus needed.

Using variants identified by large human exome and genome research studies as population references has greatly aided the filtering and interpretation of variants found in individuals with rare disease, and the study of known disease variants in the general population.[11] The growth of these population genetic databases has enabled allele frequency data to be more widely used for estimating disease prevalence. However, there have been two main challenges with using allele frequencies from population reference databases to estimate prevalence. Firstly, the sample sizes can be insufficient to robustly estimate allele frequencies associated with rare diseases for which the majority of pathogenic variants are observed rarely in the general population. In addition, many databases have been inadequate for the estimation of disease prevalence in non-European populations. Although Bayesian methods for estimating disease prevalence have been developed and applied to allele frequency from large databases,[12] they currently do not incorporate separate prior distributions for each functional annotation (e.g., nonsense, missense, etc.).

In this study, we used publicly available population references to obtain a more robust disease prevalence estimation for recessive LGMD (LGMD2). Previous epidemiology studies (Table 1) and approaches using population reference panels have been biased and would vary a lot across different reference databases when using allele frequencies based on one single observation. Although overlapping variants from reference databases have similar allele frequencies for common variants (>0.5%), they may differ greatly at lower allele frequencies (<0.5%). In fact, 69% of European singletons within the Exome Sequencing Project (ESP) are not observed in the much larger ExAC data set.[13] To overcome this bias, we introduced a Bayesian method here to re-estimate allele frequencies, taking advantage of prior knowledge in the overall distributions of allele frequencies for different functional annotations (e.g., missense, frameshift, etc.). We developed a Bayesian framework to gain robust prevalence estimates with a confidence interval. By utilizing population reference panels from ExAC and gnomAD, we simultaneously re-estimated allele frequencies for various functional annotations via a Bayesian method and then estimated disease prevalence assuming Hardy–Weinberg equilibrium. Overall, we provide a generalizable and robust framework to estimate disease prevalence for LGMD2 subtypes that can be easily adapted for other autosomal recessive diseases.

MATERIALS AND METHODS

Identification of pathogenic variants

For each disease gene, variants were downloaded from the gnomAD database. The Emory Genetics Laboratory (EGL) and ClinVar databases[30] were used to annotate known pathogenic variants. Retrieved variants were first filtered based on their allele frequencies (AF). Only variants whose minor allele frequencies are less than 0.05% in the gnomAD database were kept, unless they have been annotated as pathogenic or pathogenic/likely pathogenic in either of these two databases (EGL and ClinVar). Using the American College of Medical Genetics and Genomics (ACMG) guidelines for defining pathogenic variants,[11,30] we classified loss-of-function type variants as pathogenic (e.g., frameshift, stop gain, splicing donor, splicing acceptor) whether or not they were listed as pathogenic in the EGL or ClinVar databases. For the other types of variants, as long as they were annotated as pathogenic in either the EGL or ClinVar database, they were classified as pathogenic.

The above analysis is limited to known pathogenic variants and loss-of-function variants. We used the Combined Annotation Dependent Depletion (CADD) score[31] cutoffs to include more variants as potentially pathogenic. We applied two CADD Phred-scaled score cutoffs at 20 and 30. For further comparison, we also included all rare (AF < 0.05%) missense variants to get the upper bound of estimated disease prevalence.

Bayesian estimation of allele frequencies and disease prevalence

The development of the disease prevalence estimator builds upon a previously published method and is detailed below.

Problem setting and prior assumptions

For a single variant, we would assume the observed allele count of the variant follows a binomial distribution $Binomial(q_i, 2n_i)$, where n_i is the number of individuals having genotypes genotyped at this position in the database and q_i is the true allele frequency for this variant.

Since the conditional distribution of the observed allele count for a variant conditioned on the allele frequency q_i is a binomial distribution, we introduced a conjugate prior of q_i, $q_i \sim Beta(v_{c:i\in c}, w_{c:i\in c})$, where $v_{c:i\in c}$ and $w_{c:i\in c}$ denote the prior parameters for variants belonging to the category c, which are estimated using method of moments based on all variant data provided in the ExAC database.[13] We grouped all variants into eight categories: frameshift, splice acceptor, splice donor, stop gained, missense, untranslated region (UTR) (including 3′ and 5′ UTR), other exonic, and other variants. The allele frequencies for variants of a functional annotation are assumed to follow the same prior distribution across all genes.

In an additional analysis exploring possibly more informative priors, the CADD score was incorporated in the prior as score ranges in four groups: <5, 5–30, >30, and those without a score. In combination with the eight functional categories (mentioned above), we created a total of 32 categories with

Table 1 Estimated prevalence (per million) in nine LGMD2s and published epidemiology estimates.

Subtype/gene	Population	Bayesian estimator[a]	Direct estimator	Published estimates
2A/*CAPN3*	AFR	27.0 (16.6, 39.1)	26.5	9.47 in northeastern Italy[17]
	ASJ	0.02 (2.6e-5, 0.09)	0.01	6.0 in northern England[18]
	EAS	13.6 (4.7, 25.6)	13.0	4300 in a Mexican village[10]
	EUR	7.0 (5.4, 8.8)	7.0	576 in a province of Spain[32]
	FIN	0.50 (0.12,1.1)	0.44	48 on Reunion Island[33]
	NFE	9.4 (7.1, 11.8)	9.3	
	All	8.4 (6.8, 10.2)	8.3	
2B/*DYSF*	AFR	34.2 (22.1, 48.2)	16.0	1.3 in northern England[6]
	ASJ	0.06 (1.5e-4, 0.24)	0.04	
	EAS	14.7 (3.5, 31.2)	11.2	
	EUR	4.4 (2.8, 6.2)	3.3	
	FIN	0.2 (0.03, 0.6)	0.1	
	NFE	6.0 (3.7, 8.5)	4.6	
	All	7.5 (5.7, 9.4)	7.4	
2C/*SGCG*	AFR	0.4 (0.05, 1.0)	0.4	1.3 in northern England[18]
	ASJ	0.07 (1.7e-4, 0.3)	0.05	1.72 in northeastern Italy[34]
	EAS	0.02 (5e-05, 0.08)	0.01	70 in Tunisia[35]
	EUR	0.13 (0.04, 0.20)	0.12	48.8 in Moroccan population[36]
	FIN	0.01 (3.2e-05, 0.05)	0.01	1.8 in Japan[37]
	NFE	0.17 (0.05, 0.3)	0.16	
	All	0.12 (0.05, 0.20)	0.11	
2D/*SGCA*	AFR	18.3 (10.1, 28.0)	18.0	0.7 in northeastern Italy[17]
	ASJ	3.5 (1.0, 7.2)	3.3	3.02 in northeastern Italy[34]
	EAS	0.3 (0.02, 0.7)	0.2	
	EUR	3.9 (3.0, 5.0)	3.9	
	FIN	5.9 (3.2, 9.1)	5.7	
	NFE	3.6 (2.6, 4.7)	3.5	
	All	3.4 (2.6, 4.2)	3.3	
2E/*SGCB*	AFR	2.3 (0.4, 5.2)	2.1	0.7 in northeastern Italy[17]
	ASJ	0.6 (0.03, 1.5)	0.5	0.86 in northeastern Italy[35]
	EAS	0.14 (0.002, 0.4)	0.11	
	EUR	0.6 (0.3, 1)	0.6	
	FIN	0.07 (0.002, 0.2)	0.06	
	NFE	0.8 (0.4, 1.3)	0.8	
	All	0.80 (0.42, 1.26)	0.78	
2F/*SGCD*	AFR	0.2 (0.002, 0.5)	0.12	Not available
	ASJ	0 (0, 0)	0	
	EAS	0.9 (0.001, 3.7)	0.5	
	EUR	0.06 (0.004, 0.15)	0.05	
	FIN	0.6 (0.002, 2)	0.4	
	NFE	0.02 (0.004, 0.04)	0.02	
	All	0.07 (0.01, 0.15)	0.06	
2G/*TCAP*	AFR	0.05 (4e-04, 0.2)	0.04	Not available
	ASJ	0 (0, 0)	0	
	EAS	1.2 (0.4, 2.3)	1.1	
	EUR	0.02 (0.004, 0.03)	0.01	
	FIN	0.0001 (1.2e-07, 6e-4)	0	
	NFE	0.02 (0.006, 0.05)	0.02	
	All	0.040 (0.020, 0.063)	0.039	
2I/*FKRP*	AFR	1.5 (0.2, 3.4)	1.4	4.3 in northeastern Italy[17]
	ASJ	0.03 (4.3e-05, 0.1)	0.02	
	EAS	4.2 (1.5, 7.9)	4.1	
	EUR	8.4 (5.7, 11.4)	8.3	

Table 1 continued

Subtype/gene	Population	Bayesian estimator[a]	Direct estimator	Published estimates
	FIN	7.7 (1.8, 16.3)	7.1	
	NFE	8.5 (5.7, 11.7)	8.4	
	All	4.52 (3.20, 6.00)	4.48	
2L/*ANO5*	AFR	4.8 (2.2, 8)	4.6	2.7 in northern England[8]
	ASJ	3.0 (0.8, 6.3)	2.9	20 in Finland[38]
	EAS	0.5 (0.09, 1.1)	0.5	10 in Denmark[39]
	EUR	28.5 (24, 33.2)	28.3	
	FIN	34.4 (35.8, 25.2)	35.3	
	NFE	27.3 (22.4, 32.5)	27.1	
	All	17.6 (15.2, 20.2)	17.5	

The table shows the comparison results between prevalence of LGMD2 subtypes estimated by our method ("Bayesian estimator"), by using the allele frequencies provided in gnomAD directly ("Direct estimator") and from epidemiology studies. It also shows population stratification estimates of LGMD2s prevalence.
AFR African/African American, *All* for mixed population in gnomAD, *ASJ* Ashkenazi Jewish, *CI* confidence interval, *EAS* East Asian, *EUR* European, *FIN* Finnish, *NFE* non-Finnish European.
[a]95% confidence intervals.

allele frequency priors, then calculated similarly across all genes.

We use method of moments to estimate two hyperparameters $v_{c:i\in c}$, $w_{c:i\in c}$ in the beta prior for allele frequency q_i. More specifically, we get these two parameters by solving the following linear system of equations:

$$\begin{cases} \frac{v_{c:i\in c}}{w_{c:i\in c}} & = \frac{\sum_{j=1}^{p}\frac{x_i}{2n_i}\mathbb{1}\{j\in c\}}{\sum_{j=1}^{p}\mathbb{1}\{j\in c\}} \\ \widehat{\mu_c} & = \frac{\sum_{j=1}^{p}\frac{x_i}{2n_i}\mathbb{1}\{j\in c\}}{\sum_{j=1}^{p}\mathbb{1}\{j\in c\}} \\ \frac{v_{c:i\in c}w_{c:i\in c}}{(v_{c:i\in c}+w_{c:i\in c})^2(v_{c:i\in c}+w_{c:i\in c}+1)} & = \frac{\sum_{j=1}^{p}\left(\frac{x_i}{2n_i}\mathbb{1}\{j\in c\}-\widehat{\mu_c}\right)^2}{\sum_{j=1}^{p}\mathbb{1}\{j\in c\}} \end{cases}$$

where p is the total number of variants in the reference panel here, including both pathogenic and nonpathogenic ones. $1\{j \in c\}$ is the indicator function indicating whether the variant j belongs to the category c or not. If it belongs to the category, then the function would give a value of 1; otherwise it would give 0.

Posterior distribution of allele frequencies

The posterior distribution of the allele frequency q_i given the observed allele counts x_i and prior assumption on the allele frequency would be

$$\pi(q_i|x_i,2n_i)=\frac{\pi(x_i,2n_i|q_i)\pi(q_i)}{\int_0^1\pi(x_i,2n_i|q_i')\pi(q_i')dq_i'}$$

$$\pi(q_i|x_i,2n_i)=\frac{\binom{2n_i}{x_i}B^{-1}(v_{c:i\in c},w_{c:i\in c})q_i^{x_i+v_{c:i\in c}-1}(1-q_i)^{2n_i-x_i+w_{c:i\in c}-1}}{\int_0^1 B^{-1}(v_{c:i\in c},w_{c:i\in c})(q_i')^{x_i+v_{c:i\in c}-1}(1-q_i')^{2n_i-x_i+w_{c:i\in c}-1}dq_i'},$$

where $B^{-1}(v_{c:i\in c},w_{c:i\in c})$ is the inverse of the beta function $B(v_{c:i\in c},w_{c:i\in c})$, which makes the total probability of beta distribution $Beta(v_{c:i\in c},w_{c:i\in c})$ be 1. Based on the equation above, we can infer that the posterior distribution of q_i is a beta distribution: $Beta(x_i+v_{c:i\in c},2n_i-x_i+w_{c:i\in c})$. For pathogenic variants (from EGL and ClinVar) unseen in the population reference panel, we would take x_i being 0 and n_i being the corresponding sample size in the mixed population or the specific subpopulation.

Posterior estimation of disease prevalence

For monogenic rare diseases the disease prevalence would be $D=[1-\prod_i(1-q_i)]^2$. This is the probability of both copies of the disease gene having at least one pathogenic variant. We can use $D\approx(\sum_i q_i)^2$ to approximate the disease prevalence, which indicates that the appoximated posterior of the prevalence is a chi-square distribution with one degree of freedom. Using $\hat{D}$ to denote the approximation term for disease prevalence $(\sum_{q_i})^2$, we can get $\frac{\hat{D}}{\sigma^2}\sim\chi_1^2(\lambda)$ and $\lambda=\frac{\mu^2}{\sigma^2}$. We are using the expectation $(\lambda+1)\sigma^2$ of the distribution as the prevalence estimator here. The lower bound of the estimator with the confidence $1-\alpha$ would be $F^{-1}\left(\frac{\alpha}{2}\right)\times\sigma^2$, where $F(.)$ is the cumulative distribution function for the chi-square distribution, similarly for the upper bound. We are using $\alpha=0.05$ here to get the 95% confidence interval. Detailed derivation of equations can be found in Supplementary Methods.

Direct estimation of disease prevalence in genetic databases

For comparison, we also estimated disease prevalence by using the observed allele frequency of a pathogenic variant in genetic databases as the direct estimator for q_i (without beta prior). More specifically, the disease prevalence can be estimated by

$$D_{direct}=\left[1-\prod_i\left(1-\frac{AC_i}{AN_i}\right)\right]^2$$

where AC_i is the allele count for the variant i and AN_i is the corresponding allele number in the position. As above, for a given disease or a subtype, the product is taken over all

identified pathogenic variants in the disease gene, where *i* is the index of those identified pathogenic variants.

The scripts for estimating recessive disease prevalence based on our Bayesian framework and also direct calculation are available at https://github.com/leklab/prevalence_estimation.

RESULTS

Prevalence estimates in LGMD2 subtypes are comparable with published values

The recessive LGMDs (LGMD2) are autosomal recessive diseases that can be caused by pathogenic variants in at least 24 genes.[1] We applied our Bayesian method to nine subtypes of LGMD2 from 2A to 2L (Table 1). The gnomAD data set was used to identify putative and reported pathogenic variants in each disease gene. The disease prevalence estimates calculated by our Bayesian method were generally consistent with published prevalence estimates from epidemiological studies (Table 1), in particular for LGMD2A, LGMD2E, and LGMD2I. For other subtypes, our method produced a higher estimated prevalence, including LGMD2B, LGMD2D, and LGMD2L. These differences can be partly explained by the underdiagnosis of these late-onset or slowly progressive LGMD subtypes.[14,15] In contrast, our disease prevalence estimation for subtype LGMD2C (0.12 per million) was notably lower than the lowest published value (1.3 per million). Genetic differences across regions would also contribute to discrepancies between our results and published estimators, since most epidemiology studies have been conducted in small regions, while the databases we used include individuals with diverse genetic backgrounds. Lastly, no comparison could be made for LGMD2F and LGMD2G because there are no published prevalence estimates.

Next, we applied our method to another genetic database, BRAVO, to estimate prevalence for the same nine LGMD2 subtypes. When applied to a different database, our method provided more robust results compared with direct prevalence estimation (see "Materials and Methods") using genetic data. Prevalence estimates for six of nine subtypes estimated in BRAVO fell in the 95% confidence intervals (CIs) estimated from the gnomAD data. The other three subtypes (2A, 2D, and 2I) had an estimated prevalence close to the lower bounds of the corresponding 95% CI (Table 2). Applying the same method (either our Bayesian method or the direct way, see "Materials and Methods") in two different databases yields much larger differences than applying two different methods in the same data set, indicating the database used is the greater influence, as opposed to the method. The large differences in results from different databases can be partly explained by the sampling biases and limited sample size in each genetic data set.

Including rare missense variants currently not reported as pathogenic increases prevalence estimates

The above prevalence estimates are limited to reported pathogenic and rare loss-of-function variants found in gnomAD and do not account for other unreported missense pathogenic variants that may be in gnomAD. When we included all rare missense variants (AF < 0.05%), not surprisingly the prevalence estimates increased dramatically (Table 3) compared with the results indicated above. This increased prevalence was proportional to the coding length of the gene, as larger genes will accumulate more rare variants by random chance.

This analysis assumes all rare missense variants are pathogenic, which is likely not the case. We then applied the CADD[16] method to classify the pathogenicity of rare missense variants. The CADD Phred-scaled cutoff scores of 20 and 30 were used to define pathogenicity, which respectively represent the top 1% and 0.1% of most deleterious substitutions predicted by the CADD method, i.e., the higher the score, the more likely a variant will be pathogenic. The published prevalence estimates still fell outside of the 95% CI calculated when missense variants with a cutoff score of 20 were included, while the more stringent cutoff score of 30 produced closer estimates (see Table 3). For example, with LGMD2E, the estimated prevalence using a cutoff score of 30 is 1.1 per million, similar to the published 0.7 or 0.86 per million, and is within the 95% CI (0.4 to 1.3) estimated when only considering rare loss-of-function variants and variants annotated as pathogenic in ClinVar or EGL. These results show that improved pathogenicity prediction methods are required to improve disease prevalence estimates.

Comparison with epidemiological results in population stratified analysis

The majority of epidemiological studies estimating disease prevalence have been conducted in small regions, leading to varying results across publications. LGMD2A serves as an example, where the estimates vary greatly in two small regions

Table 2 Estimated disease prevalence (per million) in gnomAD and BRAVO for nine LGMD2s.

Subtype/gene	Direct in BRAVO	Direct in gnomAD	Bayesian in BRAVO	Bayesian in gnomAD (95% CI)
2A/*CAPN3*	6.2	8.3	6.3	8.4 (6.8, 10.2)
2B/*DYSF*	5.9	7.4	6.0	7.5 (5.7, 9.4)
2C/*SGCG*	0.09	0.1	0.09	0.1 (0.05, 0.2)
2D/*SGCA*	2.0	3.3	2.0	3.4 (2.6, 4.2)
2E/*SGCB*	0.5	0.8	0.5	0.8 (0.4, 1.3)
2F/*SGCD*	0.03	0.06	0.04	0.07 (0.01, 0.15)
2G/*TCAP*	0.02	0.04	0.02	0.04 (0.02, 0.06)
2I/*FKRP*	3.1	4.5	3.1	4.5 (3.2, 6.0)
2L/*ANO5*	15.3	17.5	15.4	17.6 (15.2, 20.2)

This table lists the results of applying the direct method and our Bayesian method to estimate prevalence of nine subtypes of LGMD2.
CI confidence interval.

Table 3 Prevalence estimated (per million) including more predicted pathogenic variants.

Subtype/gene	Pathogenic variants[a,b]	All rare missense	CADD 20 cutoff	CADD 30 cutoff	Exon length (Kb)	Ratio of CADD cutoffs[c]
2A/*CAPN3*	8.4 (6.8, 10.2)	138 (125, 153)	99 (91, 108)	22 (19, 25)	29.2	2.6
2B/*DYSF*	7.5 (5.7, 9.4)	1260 (1190, 1320)	620 (590, 650)	105 (96, 115)	80.4	14
2C/*SGCG*	0.1 (0.05, 0.2)	19.6 (16.8, 22.6)	7.8 (6.6, 9.1)	0.9 (0.6, 1.2)	1.6	9
2D/*SGCA*	3.4 (2.6, 4.2)	43 (38, 49)	18 (15, 20)	6.1 (5.0, 7.3)	9.0	1.8
2E/*SGCB*	0.8 (0.4, 1.3)	28 (23, 33)	8.0 (6.4, 9.7)	1.1 (0.6, 1.7)	5.7	1.4
2F/*SGCD*	0.07 (0.01, 0.15)	10 (8, 12)	4.4 (3.6, 5.4)	0.3 (0.2, 0.5)	13.5	4.3
2G/*TCAP*	0.04 (0.02, 0.06)	4.7 (3.8, 5.8)	4.2 (3.5, 4.9)	0.26 (0.17, 0.35)	2.7	6.5
2I/*FKRP*	4.5 (3.2, 6.0)	65 (56,75)	32 (27, 37)	6.7 (5.0, 8.5)	19.7	1.5
2L/*ANO5*	17.6 (15.2, 20.2)	133 (121, 147)	66 (60, 72)	25 (22, 28)	14.0	1.4

The table shows disease prevalence estimated by our Bayesian method when including computationally predicted pathogenic variants filtered by CADD score cutoffs. Numbers listed in parentheses are the corresponding 95% confidence intervals.

[a]Including both annotated pathogenic variants and loss-of-function variants.

[b]The column is the same as the fifth column in Table 2.

[c]The ratio of estimates including CADD 30 cutoff variants over estimates in the second column.

of Italy (6.1 and 16.5 per million).[17] Due to the majority of the published estimates being from European populations, we limited our analysis to the subpopulations of European (EUR), Finnish (FIN), and non-Finnish European (NFE) here; results of additional subpopulations are shown in Table 1.

After applying population stratification, estimated prevalence is more comparable with previously published results (see Table 1). For LGMD2A, the prevalence was estimated at 9.4 per million (95% CI: 7.1–11.8 per million) in the NFE population, matching the published value of 9.4 per million in northeastern Italy. However, after population stratification, the prevalence estimations for some subtypes diverged further from the published values. For LGMD2L, compared with the estimator (17.6 per million) in a mixed population, the estimator (27.3 per million) in the NFE population is even higher than the published prevalence (2.6 per million) in northern England.[18] The much higher result could be caused by the elevated allele frequency of a founder variant, *ANO5* NM_213599.2:c.191dupA,[8] in the NFE population (0.21%) compared with the allele frequency in the mixed population (0.11%) in gnomAD. For subtypes only common in certain populations, the stratification can provide a more precise prevalence estimate. For example, the prevalence of 2G is estimated in East Asians (EAS) to be about 1.2 per million, while it is less than 0.05 per million in other populations (Table 1). This result suggests that varied genetic backgrounds can lead to population differences in disease prevalence estimates, which can be shown in results from both epidemiological studies and genetic databases.

Inclusion of CADD in the prior and unseen pathogenic variants

We next performed two additional analyses specifically facilitated by using a Bayesian framework. CADD scores were used to categorize variants in combination with functional categories for updating allele frequencies of pathogenic variants (see "Materials and Methods"). After categorizing variants into smaller specific groups, disease prevalence was re-estimated for LGMD2 subtypes (Supplementary Table 2). Compared with results in Table 1, prevalence estimates were overall very similar with only small changes in subpopulations.

There are a number of reported pathogenic variants that were not observed in gnomAD (Supplementary Table 1) due to sampling and being ultrarare variants. Using a Bayesian framework these variants can be included in the prevalence estimates resulting in slightly higher estimates in the subpopulations (Supplementary Table 3). Furthermore, this can provide a nonzero estimate and confidence intervals in instances where no pathogenic variants are observed in the subpopulation (e.g., LGMD2G prevalence in the Ashkenazi Jewish subpopulation).

Estimating prevalence in well-studied diseases

To further confirm the reliability of our results, we also applied our method to three non-neuromuscular diseases;

sickle cell disease,[19,20] cystic fibrosis,[21] and Tay–Sachs disease,[22,23] and estimated their prevalence in the subpopulation where they were sourced. Known pathogenic and putative loss-of-function variants for the corresponding disease genes *HBB*, *CFTR*, and *HEXA* were extracted from gnomAD (see "Materials and Methods") and our Bayesian method was used to calculate the posterior allele frequency distributions and an estimate of disease prevalence. The published estimates were within the confidence intervals for Tay–Sachs. For sickle cell disease, differences can be explained by HBB alleles associated with other β-hemoglobinopathies such as β-thalassemia.[20] In addition, prevalence adjustment[21] was required for cystic fibrosis as an early-onset life-shortening disease (Supplementary Material). Overall, taking this extra information into consideration, our prevalence estimates for these three diseases are similar to published figures, indicating that our method is robust across multiple autosomal recessive diseases.

DISCUSSION

Through the application of a Bayesian method to large publicly available genetic databases, we have determined robust prevalence estimations for LGMD2 subtypes that are consistent with published figures from epidemiological studies. By applying our method of calculating prevalence to another genetic database, BRAVO, the robustness of the method was confirmed since most prevalence estimates from BRAVO were within our estimated confidence intervals using gnomAD. For further evaluation, we estimated prevalence for three nonmuscular diseases using the method and generated similar values to published results.

Building upon a previous Bayesian prevalence estimation method,[12] we estimated LGMD2 prevalence by simultaneously considering more than one variant using much larger databases, which mitigates underestimation of disease prevalence. We also considered functional annotation when updating allele frequency for each variant. Utilization of the largest genetic databases available also made our estimation more robust, since databases with insufficient sample size would lead to increased absence of rare pathogenic variants. Although we have extended the previous Bayesian method, similar challenges still remain. First, there is an assumption of Hardy–Weinberg equilibrium, which can deviate when using aggregated population data. Specifically, the gnomAD data set contains consanguineous populations inferred by their higher inbreeding coefficients and also population stratification due to aggregating subpopulations into large continental groups.[13] Also, the classification and reporting of rare variants as pathogenic has been challenging despite established guidelines.[11] Although promising, our results, presented in Table **3**, suggest that further improvement in computational pathogenicity prediction methods for rare missense variants is required and overall the false positives are still too high for these to be used for prevalence estimates.[24,25] Lastly, pathogenic variants are assumed to be independent of each other and therefore this method does not account for rare variants that occur on the same haplotype (i.e., linkage disequilibrium).

In addition, there are assumptions that may affect our results in the context of LGMD prevalence estimates. First, we assumed pathogenic variants observed in compound heterozygous and homozygous states have the same severity, which may result in differences compared with published values. For example, the c.191dupA founder variant in *ANO5* is observed as homozygous in one individual in gnomAD, suggesting later onset and/or a much milder muscle phenotype associated with being homozygous for this variant. Second, the analysis is limited to single-nucleotide variants (SNVs) and small insertions and deletions. Large duplications and deletions account for some of the pathogenic variants discovered in neuromuscular disease genes with some having higher frequencies due to founder effects such as the exon 55 deletion in NEB[26] associated with autosomal recessive nemaline myopathy. Furthermore, we assume that all pathogenic variants for a subtype have been identified in the database we used here, which is likely not true (Supplementary Table 1), and may lead to an underestimate of disease prevalence. Conversely, the current analysis does not take into account the situation where multiple disorders are caused by variants in the same genes. For example, in the case of *FKRP*, the prevalence estimate includes both LGMD and Walker–Warburg syndrome variants,[27] and thus can overestimate the LGMD prevalence. As variant databases become more comprehensive, this information can be accurately extracted to mitigate this overestimation. Lastly, we have only estimated prevalence in this study for recessive LGMD2 disorders where compound heterozygous or homozygous variants cause disease. Recently, several heterozygous variants in genes associated with LGMD2 subtypes have been identified that can act dominantly, such as a 21-bp deletion in *CAPN3*.[1] The method we developed here is limited, however, for the estimation of dominant LGMD prevalence since dominant variants are expected to be largely absent from population databases ExAC and gnomAD, while any present may be further complicated by reduced penetrance. Taking the general and LGMD-specific assumptions together explains some of the discrepancies between published epidemiology reports and the results presented in our study.

In contrast to published prevalence estimates from epidemiology studies, our results based on allele frequencies obtained from population reference databases are not impacted by public policy and are not health system–specific to countries or regions. However, our results are also affected by different genetic backgrounds across regions (LGMD2L: 17.63 per million in the global population and 27.33 per million in the non-Finnish European population). Additionally, differences in sample sizes of various subpopulations in the genetic database used would also affect the identification of causal variants. Although the sample size of the database used here is the largest available, some rare pathogenic variants are likely to still be missing due to an insufficient sample size, which further leads to underestimation of

prevalence, especially for rarer subtypes. The underestimated prevalence of LGMD2C (0.12 per million compared with 1.3 per million) may be caused in particular by the absence of various pathogenic variants in the database used. The Bayesian framework allowed for reported pathogenic variants unseen in gnomAD to be included in the prevalence estimates; however, it did not result in much difference except in certain subpopulations (Supplementary Table 3). Future work may include estimating allele frequencies for the absent pathogenic variants by incorporating the UnseenEst method, which was successfully applied to estimate unseen variants in ExAC.[28]

Overall, our method provides a generalizable and robust framework to estimate disease prevalence for recessive forms of LGMD and can be adapted to estimate prevalence for other recessive diseases. By utilizing a Bayesian framework on data from the largest population reference panels (gnomAD and ExAC), this method can obtain more refined allele frequencies for rare pathogenic variants and include additional pathogenic variants from other disease databases to achieve improved disease prevalence estimates. This includes a framework for estimating the allele frequency priors, where functional annotation and CADD score groupings were used as an example. Future work will involve exploring more informative priors to improve estimates. Lastly, we have made our scripts and data available (see "Materials and Methods"), which can be easily adapted to other recessive disease genes of interest to calculate reproducible and robust estimates.

Published prevalence estimates for recessive LGMD are generally from epidemiological research studies, which are vulnerable to inaccuracies associated with delays in diagnosis or misdiagnosis, variation in diagnostic criteria used, and biases introduced by the specific population sampled.[29] By applying a Bayesian method to a genetic database, our method provides robust disease prevalence estimates for recessive LGMD from the genetics perspective.

URLS

gnomAD: http://gnomad.broadinstitute.org/downloads
Emory Genetics Laboratory database: http://www.egl-eurofins.com/emvclass/emvclass.php
ClinVar database (the version used here is 20180429): ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/
ExAC database: ftp://ftp.broadinstitute.org/pub/ExAC_release/release1/manuscript_data/
BRAVO database: https://bravo.sph.umich.edu/freeze5/hg38/

SUPPLEMENTARY INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-019-0544-8) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

N.J.L. is the recipient of a National Health and Medical Research Council (NHMRC) CJ Martin Early Career Fellowship and an American Australian Association scholarship. S.P. was supported by the Estonian Research Council grant (PUTJD827).

DISCLOSURE

M.L. has received consultant fees from Sarepta and L.E.K. Consulting. The other authors declare no conflicts of interest.

REFERENCES

1. Vissing J. Limb girdle muscular dystrophies: classification, clinical spectrum and emerging therapies. Curr Opin Neurol. 2016;29:635–641.
2. Nallamilli BRR, Chakravorty S, Kesari A, et al. Genetic landscape and novel disease mechanisms from a large LGMD cohort of 4656 patients. Ann Clin Transl Neurol. 2018;5:1574–1587.
3. Stence A, Westra S, Mathews KD, et al. Limb-girdle muscular dystrophy in the United States. J Neuropathol Exp Neurol. 2006;65:995–1003.
4. Magri F, Nigro V, Angelini C, et al. The Italian Limb Girdle Muscular Dystrophy Registry: relative frequency, clinical features, and differential diagnosis. Muscle Nerve. 2017;55:55–68.
5. Emery AE. Population frequencies of inherited neuromuscular diseases—a world survey. Neuromuscul Disord. 1991;1:19–29.
6. Mazzucato M, Visonà Dalla Pozza L, Manea S, Minichiello C, Facchin P. A population-based registry as a source of health indicators for rare diseases: the ten-year experience of the Veneto Region's rare diseases registry. Orphanet J Rare Dis. 2014;9:37.
7. Topaloglu H. Epidemiology of muscular dystrophies in the Mediterranean area. Acta Myol. 2013;32:138–141.
8. Hicks D, Sarkozy A, Muelas N, et al. A founder mutation in Anoctamin 5 is a major cause of limb girdle muscular dystrophy. Brain. 2011;134:171–182.
9. Frosk P, Greenberg CR, Tennese AAP, et al. The most common mutation in FKRP causing limb girdle muscular dystrophy type 2I (LGMD2I) may have occurred only once and is present in Hutterites and other populations. Hum Mutat. 2004;25:38–44.
10. Pantoja-Melendez CA, Miranda-Duarte A, Roque-Ramirez B, Zenteno JC. Epidemiological and molecular characterization of a Mexican population isolate with high prevalence of limb-girdle muscular dystrophy type 2A due to a novel calpain-3 mutation. PLoS ONE. 2017;12:e0170280.
11. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405.
12. Schrodi SJ, DeBarber A, He M, et al. Prevalence estimation for monogenic autosomal recessive diseases using population-based genetic data. Hum Genet. 2015;134:659–669.
13. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285.
14. Angelini C, Grisold W, Nigro V. Diagnosis by protein analysis of dysferlinopathy in two patients mistaken as polymyositis. Acta Myol. 2011;30:185–187.
15. Sarkozy A, Deschauer M, Carlier R-Y, et al. Muscle MRI findings in limb girdle muscular dystrophy type 2L. Neuromuscul Disord. 2012;22:S122–S129.
16. Rentzsch P, Kircher M, Witten D, Cooper GM, Shendure J. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2018;47:D886–D894.
17. Fanin M, Nascimbeni AC, Fulizio L, Angelini C. The frequency of limb girdle muscular dystrophy 2A in northeastern Italy. Neuromuscul Disord. 2005;15:218–224.
18. Norwood FLM, Harling C, Chinnery PF, Eagle M, Bushby K, Straub V. Prevalence of genetic muscle disease in Northern England: in-depth analysis of a muscle clinic population. Brain. 2009;132:3175–3186.
19. Hassell KL. Population estimates of sickle cell disease in the U.S. Am J Prev Med. 2010;38:S512–S521.
20. Thein SL. The molecular basis of β-thalassemia. Cold Spring Harb Perspect Med. 2013;3:a011700–a011700.
21. Farrell PM. The prevalence of cystic fibrosis in the European Union. J Cyst Fibros. 2008;7:450–453.

22. Rivas MA, Avila BE, Koskela J, et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. PLoS Genet. 2018;14:e1007329–e1007329.
23. Rozenberg R, Pereira LdV. The frequency of Tay–Sachs disease causing mutations in the Brazilian Jewish population justifies a carrier screening program. Sao Paulo Med J. 2001;119:146–149.
24. Ernst C, Hahnen E, Engel C, et al. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. BMC Med Genomics. 2018;11:35.
25. Li J, Zhao T, Zhang Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 2018;6:7793–7804.
26. Lehtokari V-L, Greenleaf RS, DeChene ET, et al. The exon 55 deletion in the nebulin gene—one single founder mutation with world-wide occurrence. Neuromuscul Disord. 2009;19:179–181.
27. Manzini MC, Gleason D, Chang BS, et al. Ethnically diverse causes of Walker–Warburg syndrome (WWS): FCMD mutations are a more common cause of WWS outside of the Middle East. Hum Mutat. 2008;28:E231–41.
28. Zou J, Valiant G, Valiant P, et al. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. Nat Commun. 2016;7:13293.
29. Auvin S, Irwin J, Abi-Aad P, et al. The problem of rarity: estimation of prevalence in rare disease. Value Health. 2018;21:501–507.
30. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42:D980–D985.
31. 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.
32. Urtasun M, Sáenz A, Roudaut C, et al. Limb-girdle muscular dystrophy in Guipúzcoa (Basque Country, Spain). Brain. 1998;121 pt 9:1735–1747.
33. Tomé FMS, Collin H, Fardeau M, et al. Juvenile limb-girdle muscular dystrophy: clinical, histopathological and genetic data from a small community living in the Reunion Island. Brain. 1996;119:295–308.
34. Fanin M, Duggan DJ, Mostacciuolo ML, et al. Genetic epidemiology of muscular dystrophies resulting from sarcoglycan gene mutations. J Med Genet. 1997;34:973.
35. Ben Hamida M, Fardeau M, Attia N. Severe childhood muscular dystrophy affecting both sexes and frequent in Tunisia. Muscle Nerve. 1983;6:469–480.
36. El Kerch F, Ratbi I, Sbiti A, Laarabi F-Z, Barkat A, Sefiani A. Carrier frequency of the c.525delT mutation in the SGCG gene and estimated prevalence of limb girdle muscular dystrophy type 2C among the Moroccan population. Genet Test Mol Biomarkers. 2014;18:253–256.
37. Okizuka Y, Takeshima Y, Itoh K, et al. Low incidence of limb-girdle muscular dystrophy type 2C revealed by a mutation study in Japanese patients clinically diagnosed with DMD. BMC Med Genet. 2010;11:49.
38. Penttilä S, Palmio J, Suominen T, et al. Eight new mutations and the expanding phenotype variability in muscular dystrophy caused by ANO5. Neurology. 2012;78:897.
39. Witting N, Duno M, Petri H, et al. Anoctamin 5 muscular dystrophy in Denmark: prevalence, genotypes, phenotypes, cardiac findings, and muscle protein expression. J Neurol. 2013;260:2084–2093.