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

Wei Liu[1], Sander Pajusalu[2,3], Nicole Lake[2,4], Geyu Zhou[1], Bradley Williams[5], Laura Rufibach[5], Monkol Lek[2*]

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, Tartu University Hospital, Tartu, Estonia
4. Murdoch Children's Research Institute, Royal Children's Hospital, Melbourne, VIC, Australia
5. The Jain Foundation, Seattle, WA, USA

[*] To whom correspondence should be addressed:
Monkol Lek
monkol.lek@yale.edu

Key words: Disease prevalence, Muscle diseases, Limb-girdle muscular dystrophy, Autosomal recessive disease

Abstract

As a fundamental measure in epidemiology, the prevalence of limb-girdle muscular dystrophies (LGMDs) can provide researchers and clinicians a better understanding of this disease. The LGMDs are a group of related diseases that present with a similar pattern of muscle weakness and have been divided into diverse subtypes based on the causal gene identified. However, even with the gene-level mechanisms known, it is still hard to get a reliable and generalizable prevalence estimator for each sub-type due to the limited number of epidemiology data and the low incidence of LGMDs. Taking advantage of recently published whole exome sequencing (WES) data and whole genome sequencing (WGS) data, we used a Bayesian method to develop a credible disease prevalence estimator. This method was applied in nine recessive LGMD subtypes (LGMD2) and the estimated disease prevalence determined by this method were comparable to published epidemiology studies. Along with each prevalence estimator, a confidence interval was calculated, making our estimators more robust to inclusions of future data.

Introduction

The limb girdle muscular dystrophies (LGMD) are a heterogeneous group of diseases, causing pelvic and shoulder girdle muscle weakness and wasting. There are currently 32[1] characterized subtypes with a diverse range of clinical phenotypes, such as differences in age of onset and progression and specific muscle wasting patterns including 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. The associated LGMD genes have cellular functions including glycosylation, muscle membrane integrity, maintenance and repair, which are a diverse range of mechanisms that when disrupted all result in muscle damage and degeneration. There currently does not exist any effective treatment for any LGMD sub-type, however promising gene therapy clinical trials are set to commence in 2019 for a subset of LGMD2 sub-types and thus improved estimates of disease prevalence is critical in planning and prioritizing these trials.

Historically, the prevalence of rare diseases has largely been estimated based on surveys and registries, which are highly biased, unreliable and varies between populations. In health care systems this is easier to monitor for childhood diseases, while with late onset and slowly progressing diseases such as LGMD, this is more difficult to collect [2]. The current published prevalence of LGMD subtypes can vary greatly between countries and even regions within countries. The genetic factors contributing to these regional differences include small sample size, founder mutations and consanguinity rates. For example, the c.191dupA variant in exon 5 of ANO5 (LGMD2L) [3] and the c.826C>A variant in FKRP (LGMD2I) [4] are known founder mutations observed broadly in Europeans; a retrotransposon insertion in FKTN (LGMD2M) is common in the Japanese [5], while the recently discovered TTN stop-gain variant (c. 107635C>T) is largely confined to the Serbian population [6]. Consanguineous marriages, which are still common in some regions, lead to increased incidences of recessive diseases like LGMD2 in those regions [7]. In addition, the resources and training available to regional healthcare system can also largely contribute to this regional variability. Besides, approximately 50% of LGMD patients remain undiagnosed. Therefore, increasing the genetic diagnosis rate is an ongoing challenge, regardless of region.

Using aggregated variants discovered in large human exome and genome research studies as population references has greatly aided in the filtering and interpretation of variants discovered in known disease genes and has provided contributing evidence for novel disease gene discoveries. In addition, this data can be used to identify genes and protein-coding regions depleted of variation and due to this regional constraint are likely to impact on human health and disease [8, 9]. Lastly, these population references have been used to study known disease mutations in the general population.

As variants aggregated from genomic datasets grow, these population reference databases have been more widely used for prevalence estimates. However, there have been two main challenges with estimating prevalence using allele frequencies from previous population reference databases. Firstly, the sample sizes have been insufficient to robustly estimate allele frequencies associated with rare diseases for which the majority of pathogenic variants are not

observed at all. In addition, many databases were inadequate to estimate prevalence in non-European populations. Although, Bayesian methods have been developed and applied on larger databases [10], they currently do not incorporate the various functional variants and subsequent allele frequencies that contribute to rare genetic diseases. In our method, the up-to-date population reference panel with the largest sample size is used, which alleviate the failure in estimating allele frequencies for extremely rare variants. By categorizing variants based on their consequences, our method can obtain more specific allele frequencies for variants since allele frequencies are affected by variant consequences due to natural selection.

In this study, we aim to use these publicly available population references to obtain a more reliable disease prevalence estimator for recessive LGMD (LGMD2). In contrast to directly taking one single observation as allele frequency estimates in either epidemiology studies or previous studies using population reference panel, we developed a Bayesian framework to gain more robust prevalence estimates with a confidence interval, which makes it more stable even with new reference panel data used. Utilizing population reference panel, ExAC/gnomAD to simultaneously re-estimate allele frequencies for various variant annotations via a Bayesian method and then estimate disease prevalence assuming Hardy-Weinberg equilibrium (see **Methods**), our approach can provide not only population-level disease prevalence estimation along with estimators in specific subpopulation like non-Finnish Europeans, but also a generalizable and repeatable estimation compared with previous epidemiology results, which is adapted to regional differences. Utilizing this method, our estimates for LGMD2 subtypes are consistent with published prevalence estimates along with a confidence interval, which makes our results robust to up to date data inclusion. As an additional evaluation of robustness, we also applied the method to another public genetic database (BRAVO) and found that prevalence estimates for 6 out of 9 LGMD2 subtypes using BRAVO were within the 95% prevalence confidence intervals estimated using gnomAD. Lastly, we have used well studied monogenic diseases with low incidences, Tay Sachs disease, sickle cell anemia and cystic fibrosis, as validation of our prevalence estimates. 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.

Results

Prevalence estimates in LGMD2 subtypes are comparable to published ones

The recessive LGMD (LGMD2) are autosomal recessive diseases, which can be caused by mutations in multiple genes. We applied our Bayesian method to nine subtypes of LGMD2 from 2A to 2L. After identifying potentially disease-causing variants based on clinical variant pathogenicity annotation databases and ACMG guidelines (see **Methods**), we used allele frequencies provided in ExAC/gnomAD to get a posterior allele frequency (the modified allele frequencies adjusted by observed allele frequencies and assumptions based on prior information, i.e. mean allele frequencies of all variants with the same consequence) for each variant. We assume the penetrance of compound heterozygotes variants is the same as for homozygotes. Disease prevalence was estimated for each gene, along with 95% confidence intervals.

Compared with published prevalence in previous epidemiology studies, our Bayesian method showed consistent results (see **Table 1**), especially for LGMD2A (published: 9.4 per million; Bayesian method: 8.39 per million, units are the same for estimators below), LGMD2E (published: 0.7; Bayesian method: 0.8) and LGMD2I (published: 4.3; Bayesian method: 4.5). For other subtypes with prevalence published, our method showed higher estimated prevalence, including LGMD2B (published: 1.3; Bayesian method: 7.4), LGMD2D (published: 0.7; Bayesian method: 3.3) and LGMD2L (published: 2.6; Bayesian method: 18). The higher prevalence estimates may be caused by the larger number of pathogenic variants used in our method compared to the number of published pathogenic ones in literature (see Table 4), like the number of used pathogenic variants here is 109 for 2L while the published number is 29, leading to the inflated prevalence estimate (18) compared to the published prevalence (2.6). Noticeably, the subtype LGMD2C shows smaller prevalence estimator (0.12) than the published one (1.3). For the other two subtypes, LGMD2F and LGMD2G, there is no prevalence published in epidemiology studies so far. These differences can be partly explained by the under diagnosis of these late-onset or slowly progressive LGMD subtypes [11, 12]. Genetics differences across regions would also lead to the discrepancies between our results and published estimators, since most epidemiology studies have been conducted in small regions while the databases we are using here include individuals with diverse genetic backgrounds.

Next, we applied our method to another genetics database, BRAVO, to estimate prevalence for the same 9 LGMD2 subtypes. When applied to a different database, our method provided more robust results compared with direct prevalence estimation using genetic data. Prevalence for 6 out of 9 subtypes estimated in BRAVO fell in the 95% confidence intervals (CIs) estimated in gnomAD. The other three subtype (2A, 2D and 2I) had estimated prevalence close to the lower bounds of the corresponding 95% CIs (see **Table 2**). The differences of applying the same method (either our Bayesian method or the direct way, see **Method**) in two different databases are much larger than that of using two methods in the same dataset, indicating the larger influence the database used has versus the method used. The large differences of results in different databases can be partly explained by the sampling biases and limited sample size in each genetics dataset.

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

The number of pathogenic variants reported in mutation databases for LGMD are quite limited due to their low incidences of diagnosis and systematic reporting is currently not widely adopted by clinical and academic laboratories. The above prevalence estimates are limited to reported pathogenic variants found in ExAC/gnomAD and do not account for other unreported pathogenic variants that may be in ExAC/gnomAD. When we included all rare missense variants (AF < 0.05%), not surprisingly the prevalence estimates increased dramatically compared to the results indicated above (Table 3). When using all observed rare missense variants, not surprisingly the inflated prevalence estimates were directly correlated to the size of the disease gene. Take 2A (CAPN3) and 2B (DYSF) for example, *DYSF* gene (233 kb) is much longer than *CAPN3* (64.2 kb) and the fold change of prevalence estimates for 2B including all rare missense variants versus only annotated pathogenic ones is also larger (168 versus 16 for 2A).

This analysis assumes all rare missense variants are pathogenic, which is likely not the case. We then applied the Combined Annotation Dependent Depletion (CADD) method to classify the pathogenicity of rare missense variants and decided on cut-off scores of 20 and 30, where the higher the score the more likely a variant will be pathogenic. Including variants with a cut-off of 20, decreases the prevalence estimate seen with the use of all rare variants, but is still larger than published values, while a cut-off of 30 produces closer estimates. For example, with LGMD2I, the estimated prevalence using the 30 cut-off is 6.70 per million, similar to the published 4.3 per million. These results show that improved pathogenicity predictions methods are required before potentially pathogenic variants can be included in disease prevalence estimates.

Comparison with epidemiology results in population stratified analysis

The majority of epidemiology studies have been conducted in small regions, leading to varying results across publications. LGMD2A serves as an example where the estimates vary greatly in two regions of Italy (6.1 and 16.48 per million). The gnomAD/ExAC data sets contain both global and population estimates allowing for the estimates based on populations: European (EUR), Finnish (FIN) and non-Finnish European (NFE).

After applying population stratification, estimated prevalence is more comparable with previously published results (see supplementary Table 1). For LGMD2A, the prevalence was estimated at 9.37 per million (95% CI: 7.13-11.82 per million) in NFE population, close to the published value of 9.4 per million in northeastern Italy [13]. However, after population stratification, some subtypes have a larger difference than published values. For LGMD2L, compared with the estimator (17.63 per million) in a mixed population, the estimator (27.33 per million) in NFE is even higher than the published prevalence (2.6 per million) in northern England [14]. The much higher results could be caused by the elevated allele frequency of a founder mutation, NM_213599.2:c.191dupA [3], in the NFE population (0.21%) of gnomAD database compared with the allele frequency in the mixed population (0.11%) of gnomAD database. In addition, other reported pathogenic variants also have higher allele frequencies in

the NFE population such as c.692G>T variant in ANO5 (0.11% and 0.09%). This suggests that varied genetic backgrounds can lead to population differences of disease prevalence estimates, which can be shown in results from both epidemiology studies and genetic databases.

Estimating prevalence in well studied diseases

To further confirm the reliability of our results, we also applied our estimator method to three non-neuromuscular diseases, Tay-Sachs disease, sickle cell anemia and cystic fibrosis, in their corresponding well-studied populations. Similar to what we did in the LGMD2s, pathogenic variants for each disease were firstly extracted from gnomAD (see **Methods**). For each disease-causing gene the allele frequencies for each variant was then updated via the Bayesian method integrating prior information from ExAC.

Tay-Sachs disease is a neurological disorder caused by recessive mutations in the *HEXA* gene. The disease is prevalent in the Ashkenazi Jewish population due to the founder mutation, a four base pair insertion. Using the Ashkenazi Jewish stratified allele frequencies and applying our method, we estimate a prevalence of 197 per million, along with a 95% CI of 137 to 265. Since only allele frequencies were taken into consideration without considering selection pressure, the prevalence obtained by our method is closer to prevalence at birth instead of population prevalence for diseases affecting life span like Tay-Sachs disease (and also Cystic fibrosis). The published 286 per million birth incidence result is close to the upper bound of what we estimated here.

Sickle cell anemia is a recessive blood disorder caused by mutations in *HBB*. Due to the c.365A>T missense mutation in HBB, the disease is rampant in Africa and among people of African ancestry the published prevalence is 2740 per million (one out of every 365) African-Americans [15]. Our estimated result in African or African-Americans is 3490 per million (95% CI: 3140 - 3853 per million), which is slightly higher than published result. The inflation of our estimates compared with published ones can be partly explained by our strong assumption that compound heterozygotes have the same penetrance with homozygotes of pathogenic variant.

Cystic fibrosis is a recessive disorder caused by mutations in *CFTR* and the published mean prevalence in European populations is 74 per million across 27 European Union countries [16]. Applying our method to NFE population, the estimated prevalence is 365 per million (95% CI: 338 - 393 per million). After adjusting the published prevalence 74 per million using the calibrating formula - incidence = 0.00019 + 0.00016 × prevalence per 10,000 [16], the incidence of 308 per million was obtained, which is closed to our estimated 365 per million.

Overall, prevalence estimated for these three diseases by our method are all close to well-adopted results published in previous studies, especially in specific populations, indicating that the results of our method are reliable across multiple autosomal recessive diseases.

Discussion

Through applying a Bayesian method to large publicly available genetic databases, we determined robust prevalence estimations for LGMD2 subtypes, which are consistent with previous reports. By applying our method of calculating prevalence to another genetics database, BRAVO, the robustness of the results was confirmed since most prevalence estimated in BRAVO were within our estimated confidence intervals using gnomAD. For further evaluation, we estimated prevalence for three non-muscular diseases using the method and the pattern was consistent, although not the same, with published results especially when the estimated population was specified.

By incorporating variant annotation databases and genetic databases, we obtained more repeatable and reliable diseases prevalence for LGMD2 subtypes, compared with epidemiology research. Our method builds and extends upon a previous prevalence estimation method [10] using genetic data in two improvements. We estimated LGMD2s prevalence by simultaneously considering more than one mutation and obtained results using much larger databases, which mitigates under estimation of disease prevalence. We also considered mutation consequences when updating allele frequency for each variant. Utilization of the largest genetic databases available also made our estimation closer to observation, since databases with insufficient sample data would lead to increased absence of rare pathogenic variants.

Like other disease prevalence estimating methods based on genetic data, our estimation has strong assumptions that may affect our results. First, we assumed pathogenic variants observed in compound heterozygous and homozygous states have the same penetrance (i.e. 100%), which may lead to an inflated prevalence. For example, the c.191dupA founder mutation in ANO5 is observed as homozygous 1 time in gnomAD suggesting lower penetrance and/or a much milder muscle phenotype associated with being homozygous for this variant. Second, pathogenic variants are assumed to be independent of each other and therefore, does not account for rare variants that occur on the same haplotype (i.e. linkage disequilibrium). Third, 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. We also assume that all pathogenic variants for a subtype have been identified in the database we used here, even for rare variants and unreported variants, which may lead to an under estimate of disease prevalence. Lastly, we assumed that only compound heterozygous or homozygous mutations can lead to LGMD2 diseases. However, several heterozygous mutations can also cause dominant phenotypes of LGMD2 subtypes, like a 21-bp deletion in CAPN3 [1].

In contrast to prevalence published in epidemiology studies, our results based on allele frequencies obtained from population reference databases are not impacted by public policy or health system specific to countries/regions. However, our results are also affected by different genetic backgrounds across regions (LGMD2L: 17.63 per million in global population and 27.33 per million in non-Finnish European population). More than that, varied sequence coverages in various regions in the genetic database used would also affect the identification of causal variants. Besides, although the sample size of the database used here is the largest up till now,

some rare pathogenic variants would still be missing due to the insufficient sample size, which further leads to under estimation of prevalence, especially for rarer subtypes. The under-estimated prevalence of LGMD2C (0.12 per million compared with 1.3 per million) may be caused by the failure in identifying pathogenic variants in the used database (gnomAD).

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 in the up-to-date largest population reference panel (gnomAD and ExAC), this method can obtain more refined allele frequencies for rare pathogenic variants and include more un-registered potentially pathogenic variants to achieve a stable and justified disease prevalence estimate along with a confidence interval.

Previous prevalence published in LGMD2 are usually from epidemiology research, which are inevitably underestimated since LGMD is a late-onset disease and is often under diagnosed. By applying a Bayesian method in genetic database, reliable disease prevalence have been estimated for LGMD2 by our method from genetics perspective.

Methods

Identifications of pathogenic variants

For each disease-causing gene, the identified variants were downloaded from the gnomAD database [8], which includes both exonic and flanking intronic variants. The Emory Genetics Laboratory (EGL) and ClinVar database [17] were used to annotate known pathogenic variants. We filtered all gnomAD variants with an allele frequency greater than 0.05%, while retaining variants that had been annotated as pathogenic or likely pathogenic by EGL or ClinVar.

Following the rules of defining pathogenic variants with strong evidence defined in ACMG guidelines [17, 18] we classify certain types of remaining variants as pathogenic (Table 5). For the other types of variants, as long as they are annotated as pathogenic in EGL or ClinVar database, they would be 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 [19] cut-offs to include more variants (mostly missense ones) as potentially pathogenic variants. We applied two CADD score cut-offs at 20 and 30 to include variants with predicted top 1% and 0.1% deleteriousness. 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

Problem setting and prior assumptions

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

Using the Bayesian method, we introduced a prior for q_i ,

$q_i \sim Beta(v_{c:i\in c}, w_{c:i\in c})$

$v_{c:i\in c}$ and $w_{c:i\in c}$ denote the prior parameters for variants belonging to the category c, which are estimated using empirical Bayes methods based on all variants data provided in the ExAC database [9]. We grouped all variants into seven main categories: frame shift, splice acceptor, splice donor, stop gained, missense, UTR (including 3' and 5' UTR) and other variants. Variants of different function annotations usually tend to have varied allele frequencies [20]. By categorizing variants based on their functional consequences, we can easily get better estimates of their allele frequencies which would be otherwise averaged among all variants, leading to inflated allele frequency estimates for variants with more deleterious function consequences. The allele frequencies for variants of a function annotation are assumed to follow the same prior distribution across all genes.

Posterior distribution of allele frequencies

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

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

Based on the equation above, we can infer that the posterior distribution of q_i is a beta distribution: $Beta(x_i + v_{c:\in c}, 2n_i - x_i + w_{c:\in c})$.

Posterior estimation of disease prevalence

For monogenic rare diseases the disease prevalence would be

$$D = [1 - \prod_i (1 - q_i)]^2$$

Here, we assume that effect for each pathogenic variants in the disease-causing gene is the same and occurrences of all pathogenic variants located in the same gene are independent of each other. In practice, q_i's that we are using here are quite small, therefore, we can use $D \approx (\sum_i q_i)^2$ to approximate the disease prevalence.

From the equation listed above, the approximation of prevalence involves the sum of multiple variables from different beta distributions, since q_i follows a beta distribution. Empirically, we use a normal distribution to approximate the distribution of the sum of beta variables .

$$t = \sum_i q_i$$

$$t|x, n \sim N(\mu, \sigma^2)$$

$$\mu = \sum_{i=1}^{k} \frac{x_i + v_{c:i\in c}}{2n_i + v_{c:i\in c} + w_{c:i\in c}}$$

$$\sigma^2 = \sum_{i=1}^{k} \frac{(x_i + v_{c:i\in c})(2n_i - x_i + w_{c:i\in c})}{(2n_i + v_{c:i\in c} + w_{c:i\in c})^2(2n_i + v_{c:i\in c} + w_{c:i\in c} + 1)}$$

Therefore, the posterior of the prevalence would follows a chi-square distribution with the degree of freedom being 1 and a non-central parameter. Using $\hat{D}$ denote the approximation term $(\sum_i q_i)^2$, we can get

$$\frac{\hat{D}}{\sigma^2} \sim \chi_i^2(\lambda), \lambda = \frac{\mu^2}{\sigma^2}$$

After getting the posterior distribution of the approximated disease prevalence, it is easy for us the get the prevalence estimator and its confidence intervals. 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 $F^{-1}(\frac{\alpha}{2}) \times \sigma^2$, where $F(\cdot)$ is the cumulative

distribution function for the chi-square distribution. Similarly, the upper bound for the estimator would be $F^{-1}(1-\frac{\alpha}{2}) \times \sigma^2$. We are using $\alpha = 0.05$ here to get the 95% confidence interval for our prevalence estimators.

Direct estimation of disease prevalence in genetic databases

For comparison, we also estimated disease prevalence by using the formula of disease prevalence and observed allele frequencies of pathogenic variants in genetic databases. More specifically, the disease prevalence can be estimated by :

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

where AC_i is the allele count for the variant i and AN_i is the corresponding allele number in the position. For a disease or a subtype, considering all identified pathogenic variants in the disease-causing gene, while 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 now available on https://github.com/vivid-/prevalence_estimation.

Used URLs:

gnomAD: http://gnomad.broadinstitute.org/downloads;
EGL genetics database: http://www.egl-eurofins.com/emvclass/emvclass.php;
ClinVar database (the version used here is 20180429, may be updated now):
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/

Tables and Figures

Table 1:

Subtype /gene	Bayesian estimator (per million)+ 95% CI	Direct estimator (per million)	Estimator published (per million)
2A/CAPN3	8.388(6.748,10.16)	8.330	9.4 [21]
2B/DYSF	7.457 (5.717, 9.363)	7.387	1.3 [14]
2C/SGCG	0.1169 (0.05125, 0.2008)	0.1133	1.3 [14]
2D/SGCA	3.361 (2.619, 4.169)	3.341	0.7 [21]
2E/SGCB	0.7967 (0.4203, 1.256)	0.7800	0.7 [21]
2F/SGCD	0.06594 (0.01283, 0.1457)	0.06095	Not available
2G/TCAP	0.03968 (0.02062, 0.06300)	0.03884	Not available
2I/FKRP	4.516 (3.207, 5.984)	4.475	4.3 [21]
2L/ANO5	17.63 (15.21, 20.18)	17.53	2.6 [14]

The table shows the comparison results between prevalence of LGMD2 subtypes estimated by our method ("Bayesian estimator") and by using the allele frequencies provided in gnomAD directly ("Direct estimator")

Table 2:

Subtype/gene	Direct in BRAVO	Direct in gnomAD	Difference (ratio)	Bayesian in BRAVO	Bayesian in gnomAD (95% CI)	Difference (ratio)
2A/ CAPN3	6.238	8.330	0.3352	6.279	8.388 (6.748, 10.16)	0.3349
2B/ DYSF	5.925	7.387	0.2467	5.966	7.457 (5.717, 9.363)	0.2500
2C/ SGCG	0.09156	0.1133	0.2379	0.09412	0.1169 (0.05125, 0.2008)	0.2423
2D/ SGCA	1.962	3.341	0.7026	1.977	3.361 (2.619, 4.169)	0.7002
2E/ SGCB	0.4909	0.7800	0.5891	0.4971	0.7967 (0.4203, 1.256)	0.6026
2F/ SGCD	0.03355	0.06095	0.8168	0.03506	0.06594 (0.01283, 0.1457)	0.8807
2G/ TCAP	0.02055	0.03884	0.8901	0.02172	0.03968 (0.02062, 0.06300)	0.8266
2I/FKRP	3.123	4.475	0.4330	3.141	4.516 (3.207, 5.984)	0.4378
2L/ ANO5	15.30	17.53	0.1456	15.39	17.63 (15.21, 20.18)	0.1457

This table lists the result of applying the direct method and our bayesian method to estimate prevalence of 9 subtypes of LGMD2. The units for the 2nd, 3rd, 5th and 6th column are all 'per million'. The 4th column, 'Difference (ratio)' column, denotes the differences between estimated prevalence by the direct calculation in gnomAD and that in BRAVO. Similarly, the 7th column also denotes differences using our bayesian method in gnomAD from that in BRAVO.

Table 3:

Subtype/Gene	Original (per million)	All rare missense (per million)	CADD 20 cut-off (per million)	CADD 30 cut-off (per million)	Published (per million)
2A/ CAPN3	8.388 (6.748, 10.16)	138.549 (125.498, 152,072)	99.137 (90.576, 107.981)	21.831 (18.762, 25.069)	9.4
2B/ DYSF	7.457 (5.717, 9.363)	1256.221 (1190.066,1323.690)	619.670 (586.771, 653.227)	105.177 (95.703, 114.977)	1.3
2C/ SGCG	0.1169 (0.05125, 0.2008)	19.610 (16.790, 22.588)	7.832 (6.647, 9.087)	0.907 (0.646, 1.200)	1.3
2D/ SGCA	3.361 (2.619, 4.169)	43.157 (38.017, 48.535)	17.816 (15.464, 20.290)	6.083 (4.968, 7.278)	0.7
2E/ SGCB	0.7967 (0.4203, 1.256)	27.886 (23.243, 32.834)	7.967 (6.392, 9.667)	1.111 (0.649, 1.660)	0.7
2F/ SGCD	0.06594 (0.01283, 0.1457)	10.088 (8.218, 12.096)	4.439 (3.580, 5.364)	0.299 (0.154, 0.477)	Not available
2G/ TCAP	0.03968 (0.02062, 0.06300)	4.715 (3.751, 5.758)	4.167 (3.484, 4.895)	0.255 (0.174, 0.348)	Not available
2I/ FKRP	4.516 (3.207, 5.984)	65.449 (56.434, 74.948)	32.290 (27.390, 37.481)	6.700 (5.038, 8.531)	4.3
2L/ ANO5	17.63 (15.21, 20.18)	133.717 (121.052, 146.842)	65.670 (59.523, 72.038)	24.825 (21.761, 28.036)	2.6

The result table shows disease prevalence estimated by our method when including computationally predicted pathogenic variants filtered by CADD score cut-offs. Numbers listed in brackets are the corresponding 95% confidence intervals.

Table 4:

Subtype/Gene	Gene length (Kb)	# used pathogenic variants in our method	# published pathogenic variants [22]
2A/ CAPN3	64.2	149	289
2B/ DYSF	233	225	213
2C/ SGCG	144	34	17
2D/ SGCA	11.7	51	63
2E/ SGCB	17.8	35	31
2F/ SGCD	897	26	8
2G/ TCAP	2.37	20	3
2I/ FKRP	30.9	49	95
2L/ ANO5	90.2	109	29

The table indicates numbers of pathogenic variants used for estimating prevalence in our method, compared with numbers of published pathogenic variants for each LGMD2 subtype.

Table 5:

Variant type	Pathogenicity
Frame shifting	Pathogenic when AF<0.05%
Splicing donor	Pathogenic when AF<0.05%
Splicing acceptor	Pathogenic when AF<0.05%
Nonsense	Pathogenic when AF<0.05%
Missense	Benign unless annotated as pathogenic at either ClinVar or EGL
Other	Benign unless annotated as pathogenic at either ClinVar or EGL

The table shows the criteria that we used to annotate variants as pathogenic or benign for later calculation of disease prevalence.

References

1. Vissing, J., *Limb girdle muscular dystrophies: classification, clinical spectrum and emerging therapies.* Current opinion in neurology, 2016. **29**(5): p. 635-641.
2. Mazzucato, M., et al., *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 Journal of Rare Diseases, 2014. **9**(1): p. 37.
3. Hicks, D., et al., *A founder mutation in Anoctamin 5 is a major cause of limb girdle muscular dystrophy.* Brain, 2011. **134**(1): p. 171-182.
4. Frosk, P., 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.* Human Mutation, 2004. **25**(1): p. 38-44.
5. Kobayashi, K., et al., *An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy.* Nature, 1998. **394**: p. 388.
6. Perić, S., et al., *A novel recessive TTN founder variant is a common cause of distal myopathy in the Serbian population.* European Journal Of Human Genetics, 2017. **25**: p. 572.
7. Pantoja-Melendez, C.A., et al., *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**(1): p. e0170280.
8. Samocha, K.E., et al., *Regional missense constraint improves variant deleteriousness prediction.* bioRxiv, 2017.
9. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.* Nature, 2016. **536**: p. 285.
10. Schrodi, S.J., et al., *Prevalence estimation for monogenic autosomal recessive diseases using population-based genetic data.* Human Genetics, 2015. **134**(6): p. 659-669.
11. Angelini, C., W. Grisold, and V. Nigro, *Diagnosis by protein analysis of dysferlinopathy in two patients mistaken as polymyositis.* Acta myologica : myopathies and cardiomyopathies : official journal of the Mediterranean Society of Myology, 2011. **30**(3): p. 185-187.
12. Sarkozy, A., et al., *Muscle MRI findings in limb girdle muscular dystrophy type 2L.* Neuromuscular Disorders, 2012. **22**: p. S122-S129.
13. Fanin, M., et al., *The frequency of limb girdle muscular dystrophy 2A in northeastern Italy.* Neuromuscular Disorders, 2005. **15**(3): p. 218-224.
14. Norwood, F.L.M., et al., *Prevalence of genetic muscle disease in Northern England: in-depth analysis of a muscle clinic population.* Brain, 2009. **132**(11): p. 3175-3186.
15. Hassell, K.L., *Population Estimates of Sickle Cell Disease in the U.S.* American Journal of Preventive Medicine, 2010. **38**(4, Supplement): p. S512-S521.
16. Farrell, P.M., *The prevalence of cystic fibrosis in the European Union.* Journal of Cystic Fibrosis, 2008. **7**(5): p. 450-453.
17. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Research, 2014. **42**(D1): p. D980-D985.
18. Richards, 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.* Genetics In Medicine, 2015. **17**: p. 405.
19. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nature genetics, 2014. **46**(3): p. 310-315.
20. MacArthur, D.G., et al., *A systematic survey of loss-of-function variants in human protein-coding genes.* Science (New York, N.Y.), 2012. **335**(6070): p. 823-828.

21. Thompson, R. and V. Straub, *Limb-girdle muscular dystrophies — international collaborations for translational research.* Nature Reviews Neurology, 2016. **12**: p. 294.
22. Di Fruscio, G., et al., *Are all the previously reported genetic variants in limb girdle muscular dystrophy genes pathogenic?* European journal of human genetics : EJHG, 2016. **24**(1): p. 73-77.