Determining the incidence of rare diseases

Matthew N. Bainbridge

Rady Children's Institute for Genomic Medicine, 1 Children's Way, San Diego, California, USA.

Abstract

Extremely rare diseases are increasingly recognized due to wide-spread, inexpensive genomic sequencing. Understanding the incidence of rare disease is important for appreciating its health impact, and allocating recourses for research. However, estimating incidence of rare disease is challenging because the individual contributory alleles are, themselves, extremely rare. We propose a new method to determine incidence of rare, severe, recessive disease in non-consanguineous populations that uses known allele frequencies, estimates the combined allele frequency of observed alleles and estimates the number of causative alleles that are thus far unobserved in a disease cohort. Experiments on simulated and real data show that this approach is a feasible method to estimate the incidence of rare disease in European populations but due to several limitations in our ability to assess the full spectrum of pathogenic mutations serves as a useful tool to provide a lower threshold on disease incidence.

Keywords

Rare disease; Genetics; Incidence; Simulation

1. Introduction

The radical decrease in cost of genome wide sequencing has led to a boom in newly discovered Mendelian diseases particularly in severe, rare, pediatric conditions. Currently there are over 7,000 rare diseases affecting 25-40 million Americans. Over 40 Billion dollars are spent annually on treatments for rare diseases (2018; Cannizzo et al. 2018). Determining the incidence of rare diseases is important for understanding how many affected individuals there might be in a population. This data can be used to help determine resource allocation or help estimate the difference between the number of individuals who are diagnosed versus those who are affected.

However, this estimation is often made challenging by the rarity of the disease and its pathogenic alleles. Further, ascertainment is imperfect, even in developed countries. Children may never be diagnosed or diagnosed only later in life, even if the disease is evident congenitally(Grier et al. 2018). And, although families will often come together to form support groups, there are typically no registries and diagnosed individuals may never come to the attention of the scientific community(Rode 2005; Valdez et al. 2016). For recessive disease, where we assume the vast majority of incidence is through inherited (as opposed to de novo) mutations, incidence of disease may be calculated if all pathogenic alleles are known and have known minor allele frequency (MAF). If the population is large enough and under random mating the frequency of homozygous or bialleleic (affected) individuals can be determined by the use of Hardy-Weinberg equilibrium (Weinberg 1909; Hardy 2003). Determining the MAF of all pathogenic alleles is difficult, typically a disease may have a small (1-5) number of common alleles (defined here as MAF $>1\mathrm{x}10^{-5}$) (the head of the distribution) and a larger number of very rare (MAF $<10^{-5}$) alleles (the tail) (Kobayashi et al. 2017) which may be so rare that they may be unobserved in large public databases.

Previous approaches to estimate monogenic recessive disease incidence have primarily aimed to estimate disease incidence using only head alleles (Schrodi et al. 2015). However, with very rare diseases extremely rare alleles can make a significant contribution to incidence. This means that many of the pathogenic alleles may be too rare to have an associated MAF from a public database, which typically have $\sim 10^5$ individuals, and many may be too rare to have ever been observed in an affected individual. Thus, there is a need to take a statistical approach to estimate the total pathogenic MAF for these rare diseases.

Here, we present a framework to estimate incidence of rare autosomal recessive diseases. This method relies on having a moderately sized (~50) cohort of individuals with known pathogenic alleles and MAF information from a large public database (Karczewski et al. 2017). We also rely on the fact that the MAF of each individual pathogenic allele need not be known, but only the combined MAF of all pathogenic alleles. Further, this framework makes multiple simplifying assumptions:

1. All pathogenic alleles segregate independently. This assumption is likely reasonable for very rare alleles (Browning and Thompson 2012)

2. All populations are out-bred and matings are random with respect to pathogenic alleles

3. All pathogenic alleles are fully penetrant independent of genetic context

4. The MAF in the public databases is an accurate representation of the true MAF for the population of the disease cohort

5. Affecteds in the cohort have been selected without bias

These conditions are critical for ensuring that alleles in the cohort are randomly drawn from the population and although these assumptions are almost certainly not true for most cases, they may be reasonable for the majority of alleles in a suitably rare disease.

Materials and Methods

Given a set of n pathogenic alleles (A), any particular allele (A_i) will have an MAF estimated from a public database of unaffected individuals (A_i^m) and a count in our disease cohort (A_i^c). We imagine three broad classes of allele, those which have both MAF and count information, those with no MAF but count information (i.e. extremely rare alleles), and those which may or may not have MAF information but have not been observed in our disease cohort ($A_i^c = 0$).

First we will estimate the MAF of all observed alleles ($A_i^c \neq 0$) without MAF information (functionally, $A_i^m = 0$). Among the observed alleles in our cohort (i.e. $A_i^c \neq 0$) with MAF information (i.e. $A_i^m \neq 0$) we compute m as the sum of the MAF, and c as the total count. The ratio m/c is a estimator of the required MAF to have one count in our cohort. Let $X = \sum_i A_i^c$ then the total MAF, M, of all observed alleles is given by M=(m/c)*X.

When the number of unobserved alleles is low, M is a reasonable approximation of the true MAF of all pathogenic alleles. However, when there are a very large number of rare alleles that are unobserved it is critical to approximate the total number of alleles in the population. To do this we make yet another assumption that all unobserved alleles have identical MAF and thus have an equal chance of having been sampled in our cohort. Although this assumption is almost certainly false, it may be approximately true for rare ($A_i^m < r$) alleles, where r is an arbitrarily picked or empirically determined threshold of rarity. Estimating the number of unobserved alleles can be accomplished by fitting our rare allele count to a conditional Poisson distribution $P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$. In this distribution, cases where $A_i^c = 0$ (unobserved) are censored. From this we can estimate λ (Cohen 1960) using a maximum likelihood estimator $\lambda / (1 - e^{\lambda}) = \underline{x}$. Using this estimator for λ it is possible to estimate the total number of unobserved, u, pathogenic alleles in the population as $u = O/(1 - e^{-\lambda}) - O$ where O is the number of observed alleles with $A_i^m < r$. To estimate the MAF for each allele we take the average MAF for all rare alleles as $R = \sum_{i \in J} A_i^m / \sum_{i \in J} A_i^c$ where $J = \{i \mid A_i^m < r\}$. The total contribution of unobserved alleles is given by U=R*u. And following, the total pathogenic MAF for all alleles is F=M+U.

Results

We use both simulated data sets and real world data to test our procedure. We use three sets of simulated data which consists of 1) predominantly common alleles, 2) predominantly rare alleles, 3) all alleles at the same frequency. Real world data is obtained from the TESS foundation for SLC13A5 deficiency, a disease with ~60 affected families worldwide.

Simulation

We developed a program that takes as input the true underlying MAF distribution for pathogenic alleles. Using these MAFs it will simulate a large public database (120,000

alleles) of individuals. By simulating the database we capture the uncertainty of estimating the MAF of very rare alleles from such a database. Further, we randomly generate 50 individuals who are biallelic or homozygous for a pathogenic allele. We then use our above described approach to estimate the total pathogenic allele frequency. We simulated 4 distinct pathogenic allele distributions 1) a distribution where there are 5 'common' alleles (MAF: $2x10^{-4}$ to $2x10^{-5}$, total of $3.9x10^{-4}$) and 20 rare alleles (MAF: $3x10^{-6}$) with a total MAF across all alleles of $4.7x10^{-4}$. 2) A distribution with 5 common alleles (as before) and 110 rare alleles (30 with an MAF of $5x10^{-6}$ and 80 with an MAF of $3x10^{-6}$) with a total MAF of $7.8x10^{-4}$ 3) A distribution with 43 alleles with identical frequencies ($1x10^{-5}$) with a total MAF of $4.3x10^{-4}$ and 4) A distribution of 143 alleles with identical frequencies $3x10^{-6}$ with a total MAF of $4.29x10^{-4}$. The first two distributions more accurately reflect allele distributions in reality, while the latter two represent a challenging distribution for our approach. For each distribution the simulation was run 100 times with predetermined random number generator seeds. A predetermined cut off of $r=1x10^{-5}$ was used to separate rare from common alleles. For each variable in our approach we were able to use the given MAF values to determine our error in the contributory MAF of all alleles observed in our cohort (M), the contributory MAF of alleles that have not been observed in our cohort, U, the number of estimate unobserved alleles, u, and the total estimated MAF, F. Table 1 shows the averaged results and standard deviation of 100 simulations.

First we note the estimated Observed Count MAF, M, very closely approximates the true observed MAF, with a maximum average error of ~2.6%. Calculating the MAF contribution of unobserved alleles, U, however, is more problematic with over estimation as high as 116%. This is driven, in part, by overestimating the number of missing alleles (u), by as much as 145%. This results in a large overall error in the contribution of the unobserved alleles to the total estimated MAF. This is likely driven by the small cohort size and was anticipated by Cohen(Cohen 1960). The overall effect is mitigated by the small contribution that the unobserved alleles make to F. Unexpectedly, the total MAF estimates using our approach for the even-distributions is more accurate than the head and tail distributions. This is likely driven by the large number of observed counts that are lost to the common alleles in the head, leaving fewer observations to accurately estimate the tail of the distribution.

To test whether increasing cohort size would improve our estimate for u, we reconducted the simulation for the two head and tail distributions with cohort sizes of 50, 100, and 200 each (Table 2). For the short-tail distribution we often observed every allele, and over-estimated the number of missing alleles. When the cohort size is 50 we would, on average, estimate there were 7.78 additional missing alleles. When the cohort size was doubled, we would only estimate 0.362 missing alleles, and this value drops to approximately 0 when using a 200 person cohort. Similar results were observed for the long tail distribution with the percent error in missing alleles dropping to just 1.5% with the largest cohort. As expected, the estimates for F, the total MAF of all pathogenic alleles, also becomes extremely accurate with a 200 person cohort and was estimated to be $4.73x10^{-4}$ ($1.86x10^{-5}$) and $7.86x10^{-4}$ ($3.18x10^{-5}$) for the short and long tail, respectively, a difference of ~1%.

It is notable that determination of pathogenicity of variants is imperfect. That is, an observed variant that is disease causing may not be recognized as such. To test uncertainty in

determining whether an allele is pathogenic we altered our model to randomly dismiss a portion of alleles that were only observed once. This effect will be most pronounced in the long tail cohort. Whereas, we normally see a ~5% over-estimate in MAF (Table 1) when we dismiss 25, 50, or 100% of alleles only observed once we observe under estimates of the true MAF by 12, 31 and 68%, respectively.

Additionally, we assume that unknown pathogenic alleles are a rare cause of disease, however, it may be possible that relatively common alleles may be unappreciated causes of disease, for example, non-coding variants or copy neutral structural variants. Such alleles are missing from the model in the model and can have a large impact on estimated incidence given by (T-m)2, where T is the true MAF of all pathogenic alleles and m is the MAF of the relatively common missing allele. In practice, if m=0.1T, then the disease incidence is underestimated by 19% and 35% if m=0.2T.

Real world data

Several assumptions in this framework are clearly not true and/or not possible to assess in practice. Perhaps most concerning is the biased nature of the cohort selection, which will, currently, heavily favor European ancestry and those living in more developed countries. Further, MAF estimates for Europeans are generally better than for other populations allowing for more accurate MAFs for rare alleles. This section, therefore seeks to provide practical advice for determining our estimator and provide an example of implementing in practice.

First, setting the value of r (what qualifies as a rare allele) is straightforward and should be on the same order as the public database (currently ~10^{-5}). However, r, may be empirically determinable based on observation in the cohort. Second, in the vast majority of cases, we will observe a bias in the ascertained alleles towards those of European ancestry. One may wish to exclude non-Europeans from the cohort, however, this may also adversely affect the estimator. In any case, using a MAF that is ethnicity specific will likely give more accurate results than using the overall (pan-ethnic) MAF for any particular allele. Ideally, this framework could be implemented for every ethnicity separately, but typically cohort size is limiting. Next, one must take care in counting the occurrence of each allele in the cohort. In particular, related individuals (typically siblings) should not be counted twice as these are not independent observations of the allele. Homozygous rare alleles are concerning for potential consanguinity, which also does not constitute independent observations of the allele. One may wish to count these alleles only once or leave the individual out entirely, especially if the individual is from a region that practices consanguineous marriage. However, many recessive diseases are frequently assessed in such regions and this may represent a substantial portion of the cohort for some diseases. Once the approach is established it should be possible to construct a table similar to Table 3.

Table 3 is composed of cohort data collected for SCL13A5 deficiency [MIM:608305] (the TESS cohort). Defects in this gene cause Epileptic encephalopathy, early infantile, 25, an autosomal recessive encephalopathy that can result in severe, early onset seizures, with intellectual disability and failure to thrive. The current registry contains information on 37 independent families collected from around the world but with a strong bias to those residing

in the United States and those with European ancestry. The minor allele frequency is derived from non-Finnish Europeans from the gNomad database (accessed 6/25/2019)

For this data setting $r=1 \times 10^{-5}$ seems reasonable as it separates the more commonly observed alleles from the tail of the distribution, although 9×10^{-6} could also be considered as to include allele #6 which has a relatively high count in our cohort. In the end using one value over the other makes only a very small difference to the final outcome (data not shown). The total measured MAF, m, is 4.53×10^{-4}, c is 49 and X is 74, yielding a total observed MAF, M, of 6.84×10^{-4}. Turning our attention to the tail of the distribution (allele numbers >5) we find the average MAF, R, to be 1.97×10^{-6} and the average count, $\underline{x}$, to be 1.77. This gives an estimated lambda (Cohen 1960) of 1.2828 and, from the Poisson distribution with $k=0$, we estimate 27.7% of alleles to be unobserved, yielding 8.44 expected unobserved alleles, u. The total MAF contribution of unobserved alleles, U, is therefore estimated to be 1.67×10^{-5}. Finally we determine the total pathogenic MAF for SLC13A5 deficiency to be the sum of U and M, $F=7.01 \times 10^{-4}$. The United States has approximately 3.8 million births per year, and thus this translates to 1.87 births per year. The SCL13A5 registry contains year and country of birth for each subject and thus can also be used to estimate the occurrence of SLC13A5 deficiency. Table 4 lists year of birth and 5 year running count for all children in the SLC13A5 registry born in the United states. From this table we can see that the registry contains approximately 0.6 to 0.8 births per year on average, with both 2018 and 2019 having 2 births. This 5 year running estimate is not significantly different from our estimator (p=0.18, Fisher's Exact).

Discussion

Determining the incidence of rare disease is a daunting task because of the rarity of the contributing alleles and small number of affected individuals. Our framework relies on a number of simplifying conditions in order to estimate the frequency of a rare disease. Although these assumptions are likely not true, the error they introduce into the estimator appears minimal from simulation results. Further, it simplifies calculating the estimator and can be easily done by hand. Despite this, there are several limitations to this approach; critically the ethnicities of the disease cohort should be known and accurate population allele frequencies needs to be known, something that is not possible for many ethnicities. Inbreeding is also a complicating factor and is especially poignant for recessive diseases which are frequently first identified in communities that practice consanguineous marriage. The ability to correctly identify pathogenic alleles is also critical, and can be daunting when contributory alleles are extremely rare and frequently only observed as singletons. Failure to recognize pathogenic mutations as pathogenic can lead to moderate to severe underestimates of incidence. Further, we assume that common disease causing alleles are known and unknown disease mutations are rare. If, however, this is not the case, then this approach will lead to a severe underestimate of the true incidence of disease. Such diseases may be recognized by low molecular diagnostic rates despite high clinical suspicion. Further, in such cases we would expect to see multiple incidences where only a single pathogenic allele can be identified and the second 'hit' is unrecognized. For such diseases, this approach will lead to an underestimate of disease incidence and caution must be taken when interpreting these results.

It is also notable that some diseases may be genetically heterogeneous, that is, disease causing variants in multiple genes may lead to a similar clinical phenotype (e.g. Noonan syndrome) in such cases the approach outlined here must be taken with each gene individually and the individual incidences summed up to get the total incidence. Some disease may also be clinically heterogeneous (e.g. STXBP1 deficiency), that is, pathogenic variants in the same gene may lead to different clinical presentations. Although this is an important factor, its impact is lessened with widespread genetic testing which provide a molecular diagnosis to supplement the clinical diagnosis. However, in some cases, the clinical manifestations may be so mild or non-specific, that no genetic testing is pursued and these patients will go unappreciated. Thus, an additional limitation of this approach is that it will only assess the incidence of disease for the well-described (typically severe) form of disease.

Currently there are no other methods to estimate the incidence of very rare disease. Existing methodologies rely on the majority of causative alleles having known frequencies (Schrodi et al. 2015) and these methods will underestimate incidence when many of the causative alleles are extremely rare. Underestimating incidence can significantly affect allocation of resources for affected individuals, their families and caregivers. When compared to an empirical derivation using the TESS cohort data, our estimator is approximately twice as high as the observed estimate, although not statistically significantly so. For several reasons the registry may be an underestimate of the real incidence of SCL13A5 deficiency. First, genetic sequencing may fail to ascertain some individuals and has only become common in the past 5 years. Further, the registry may fail to ascertain individuals who have a genetic diagnosis. Lastly, as affected children become older, they are less likely to receive a diagnosis. Thus it is feasible that our estimator for the incidence for SLC13A5 deficiency is accurate.

This framework provides a reasonable estimate of the true incidence of rare, recessive disease and can help rare disease organizations better understand the total disease burden in a population. This can be critical for helping set health policy, obtaining funding for disease research from government agencies, or generating interest from private corporations.

Acknowledgements

References

Browning SR, Thompson EA (2012) Detecting Rare Variant Associations by Identity-by-Descent Mapping in Case-Control Studies. Genetics 190:1521–1531. 10.1534/genetics.111.136937 [PubMed: 22267498]

Cannizzo S, Lorenzoni V, Palla I, et al. (2018) Rare diseases under different levels of economic analysis: current activities, challenges and perspectives. RMD Open 4:. 10.1136/rmdopen-2018-000794

Cohen AC Jr (1960) Estimating the Parameters of a Modified Poisson Distribution. J Am Stat Assoc 55:139–143. 10.1080/01621459.1960.10482054

Grier J, Hirano M, Karaa A, et al. (2018) Diagnostic odyssey of patients with mitochondrial disease: Results of a survey. Neurol Genet 4:e230 10.1212/NXG.0000000000000230 [PubMed: 29600276]

Hardy GH (2003) Mendelian proportions in a mixed population. 1908. Yale J Biol Med 76:79–80 [PubMed: 15369635]

Karczewski KJ, Weisburd B, Thomas B, et al. (2017) The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Res 45:D840–D845. 10.1093/nar/gkw971 [PubMed: 27899611]

Kobayashi Y, Yang S, Nykamp K, et al. (2017) Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med 9:. 10.1186/s13073-017-0403-7

Rode J (2005) Rare Diseases: Understanding this Public Health Priority. Rare Dis 14

Schrodi SJ, DeBarber A, He M, et al. (2015) Prevalence estimation for monogenic autosomal recessive diseases using population-based genetic data. Hum Genet 134:659–669. 10.1007/s00439-015-1551-8 [PubMed: 25893794]

Valdez R, Ouyang L, Bolen J (2016) Public Health and Rare Diseases: Oxymoron No More. Prev Chronic Dis 13:. 10.5888/pcd13.150491

Weinberg W (1909) Über Vererbungsgesetze beim Menschen. Z Für Indukt Abstamm-Vererbungslehre 2:276–330. 10.1007/BF01975801

(2018) Updated Study Analyzes Use and Cost of Orphan Drugs. In: NORD Natl. Organ. Rare Disord https://rarediseases.org/updated-study-analyzes-use-and-cost-of-orphan-drugs/. Accessed 19 Nov 2019

Table 1.

Estimates of Observed Count MAF, Missing MAF, and their errors from the true values as well as the error in the true estimate of the number of missing alleles and the final total estimated MAF.

Allele Distribution	Observed Count MAF, M	Error in M (%)	Missing MAF Estimate, U	Error in U (%)	Error in unobserved pathogenic alleles, u (%)	Total Estimated MAF, F	TrueMAF
Head and Short Tail	4.39E-4 (2.0E-5)	0.39 (4.1)	6.86E-5 (5.2E-5)	116 (206)	145 (241)	5.08E-4 (5.7E-5)	4.7E-4
Head and Long Tail	5.4E-4 (2.5E-5)	1.0 (3.9)	3.23E-2 (1.7E-4)	19 (58)	33.9 (59.9)	8.64E-4 (1.8E-4)	7.8E-4
Short, even	3.93E-4 (2.2E-5)	0.19 (4.2)	4.74E-5 (1.6E-5)	55 (120)	55.1 (120)	4.4E-4 (3.1E-5)	4.3E-4
Long, even	2.24E-4 (1.8E-5)	2.6 (5.3)	2.34E-4 (8.4E-5)	12.6 (48)	12.8 (49.8)	4.5E-4 (9.8E-5)	4.29E-4

Standard deviations given in parentheses

Table 2.

Average missing allele count and percent error in missing allele count for the short and long tail distributions, respectively.

Cohort Count	50	100	200
Short Tail missing allele count	7.78 (13.5)	0.346 (0.2)	0.1 (1)
Long Tail **%** error missing allele count	12.4 (39.3)	6.9 (12.5)	1.5 (5.6)

Standard deviations given in parentheses

Table 3.

European MAFs and observed cohort counts for SLC13A5 deficiency

Allele Number	Allele Count	MAF
1	19	2.73E-04
2	3	4.41E-05
3	10	3.10E-05
4	2	3.10E-05
5	1	3.10E-05
6	6	9.09E-06
7	2	8.94E-06
8	3	8.81E-06
9	2	8.80E-06
10	1	7.77E-06
11	2	0
12	2	0
13	2	0
14	1	0
15	1	0
16	2	0
17	1	0
18	1	0
19	1	0
20	3	0
21	1	0
22	1	0
23	1	0
24	1	0
25	2	0
26	2	0
27	1	0

Table 4.

Subject ID, year of birth and 5 year running count for individuals in the TESS cohort.

Subject ID	Year of Birth	5 Year running Count
59	1999	1
1	2003	2
5	2005	2
4	2006	3
2	2007	4
3	2008	4
58	2012	2
6	2016	2
7	2018	3
54	2018	3
43	2019	4
55	2019	4