Published in final edited form as:

Fam Cancer. 2017 October ; 16(4): 611–620. doi:10.1007/s10689-017-9989-6.

Power of Pedigree Likelihood Analysis in Extended Pedigrees to Classify Rare Variants of Uncertain Significance in Cancer Risk Genes

Elisabeth A. Rosenthal¹, John Michael O. Ranola², and Brian H. Shirts²

¹Division of Medical Genetics, School of Medicine, University of Washington, Seattle, WA, USA

²Department of Laboratory Medicine, University of Washington Medical Center, Seattle, WA, USA

Abstract

Rare and private variants of uncertain significance (VUS) are routinely identified in clinical panel, exome, and genome sequencing. We investigated the power of single family co-segregation analysis to aid classification of VUS. We simulated thousands of pedigrees using demographics in China and the United States, segregating benign and pathogenic variants. Genotypes and phenotypes were simulated using penetrance models for Lynch syndrome and breast/ovarian cancer. We calculated LOD scores adjusted for proband ascertainment (LODadj), to determine power to yield quantitative evidence for, or against, pathogenicity of the VUS. Power to classify VUS was higher for Chinese than United States pedigrees. The number of affected individuals explained the most variation in LODadj (21–38%). The distance to the furthest affected relative from the proband explained 1–7% of the variation for the benign VUS and Lynch associated cancers. Minimum age of onset explained 5–13% of the variation in families with pathogenic breast/ovarian cancer variants. Random removal of 50% of the phenotype/genotype data reduced power and the variation in LODadj was best explained by the distance to the furthest affected relative followed by the number of affected individuals and minimum age of onset when the founder was only 2 generations from the proband. Power to classify benign variants was ~2× power to classify pathogenic variants. Affecteds-only analysis resulted in virtually no power to correctly classify benign variants and reduced power to classify pathogenic variants. These results can be used to guide recruitment efforts to classify rare and private VUS.

Corresponding author: Elisabeth A. Rosenthal, University of Washington Medical Center, 1705 NE Pacific St, Box 357720, Seattle, WA 98195, erosen@uw.edu.

Web Resources

CoSeg R package:

CRAN – <https://cran.r-project.org/web/packages/CoSeg/index.html>

R-Forge – <https://r-forge.r-project.org/projects/coseg/>

Online cosegregation analysis tools:

<http://analyze.myvariant.org>

<https://www.msbi.nl/cosegregation/default.aspx>

Keywords

Variant classification; LOD score; private VUS; Lynch syndrome; breast cancer; ovarian cancer

Introduction

Genetic testing for diagnosis and risk assessment is becoming standard of care for a wide variety of inherited diseases. Next generation sequencing panels and exome/genome sequencing often identify extremely rare variants of uncertain significance (VUS) in known pathogenic genes [1]. These rare VUS may be private to a single family, precluding the ability to assess sufficient independent samples for case-control association testing [2]. As the number of patients receiving clinical or research sequencing, and the number of genes sequenced per patient, increases, it is expected that the number of extremely rare VUS will dramatically increase [3, 4]. Challenges in assessing pathogenicity of these variants will undoubtedly continue for decades.

The American College of Medical Genetics and Genomics (ACMG) and Association of Molecular Pathology (AMP) proposed recommendations for clinical variant classification [5]. Although these recommendations are nuanced, in general, a variant has strong or very strong evidence of pathogenicity if it is rare among all populations and any of the following are true: it is a null variant in a gene known to have other pathogenic null variants, there are multiple *in silico* and *in vitro* pieces of evidence supporting pathogenicity, or the prevalence of the variant is significantly higher in affected individuals than in unaffected individuals. Accordingly, rare and family-specific VUS, which have arisen in the last several hundred years, will have, at most, weak evidence for pathogenicity as they will be absent in population databases and are unlikely to have extensive functional data. Beyond frequency, the only ACMG-AMP specified data likely to be available for private VUS is from *in-silico* computational algorithms or *in vitro* studies of protein levels and activity.

Family studies are a source of independent information potentially available for all types of VUS that can aid their classification. Simple methods of counting affected heterozygotes, or the meioses connecting them, have been proposed, and may be ideal for completely penetrant traits [6]. However, meiosis-counting methods do not fully account for incomplete penetrance, variable expressivity, variable age of onset, or phenocopies. Despite these challenges, meiosis-counting methods are commonly used and often misused [1]. Alternatively, as has been proposed by the International Society for Gastrointestinal Hereditary Tumors (InSiGHT), family-based co-segregation analysis methods can be applied in the situation of rare VUS in genes known to be associated with risk phenotypes under a dominant mode of inheritance, adjusting for age of onset, sex, and the population based prevalence, among other risk factors [7–9]. Furthermore, the quantitative result from co-segregation can be validated and combined with other quantitative data, creating robust evidence to classify VUS [5, 9, 10].

Most clinical laboratories that offer genetic sequencing also offer family studies to at least a subset of families with a VUS [11]. However, policies and practices vary between clinical laboratories and segregation studies may be preferentially offered where initial family

structure suggests the variant could be pathogenic [11, 12]. Yet, few laboratories have reported on success rates for family studies. This may be because the timeline for classification is typically many years, and may be indefinite, particularly for VUS that are private to one family [2, 13, 14]. Furthermore, the types of relatives and family structures needed to yield strong classification evidence is not well defined. Previous research on the power of family segregation evaluated uniform, generic families, without taking into consideration population demographics and family structure for precise co-segregation analysis [2, 15, 16]. Power of typical linkage analysis, where the goal is to identify a polymorphic region of the genome in linkage with disease, increases with the number of informative meioses (i.e., parental alleles can be definitively assigned to an offspring)[17, 18]. However, little is known about how family structure influences variant classification when the gene is *a priori* known to be causal for the phenotype and the variant is extremely rare.

Here we report the results of modeling population demographics and family structure to determine the potential power of linkage analysis in single extended families to generate strong evidence for classification of VUS using Lynch syndrome [MIM: 120435] and breast/ovarian cancer [MIM: 604370 and 612555] penetrance models. Lynch syndrome and breast/ovarian cancer are among the most common cancers in the United States and China, making them ideal candidates [19, 20]. We focused on the most common genes underlying these disorders (*MLH1*[MIM: 120436]/*MSH2*[MIM: 609309]; *BRCA1*[MIM: 120436] and *BRCA2*[MIM: 600185]) as they are frequently the first genes sequenced in the clinic. In addition, we evaluated the kinds of relatives that can increase power, so that efforts at recruitment of relatives can lead to higher gains in power.

Materials and Methods

Simulated Pedigrees

We generated pedigrees designed to mimic historical demographic data from China (CHI) and the United States (USA) using the R package CoSeg (see Web Resources). We used reports from census and other population records to inform changes in family structure through the 20th century and into the early 21st century, including the US Census Bureau, gapminder.org, the World Bank, and academic texts [21–26]. Simulation models used estimated distributions from these sources for age at marriage, fertility, and mortality (Supplemental Table 1).

Simulated families were seeded with a proband whose age is sampled from a skewed Normal distribution (mean (μ) = 51.5 years, standard deviation (σ) = 10, skew (ξ) = 0.8), in order to mirror the age distribution of individuals seeking hereditary cancer testing at the University of Washington. We created realistically aged founder ancestors 2, 3, or 4 generations back from this proband creating 1000 each of small, medium and large families, respectively.

Marriage age was modeled using a positively skewed distribution around the reported median with $\sigma = 2.5$ and $\xi = 1.5$ for each decade between 1900 and 2000, and one estimated median for any marriage in the 19th century. We simulated a spouse, descendants, and death

year of the seed individual based on the demographic data specific to that individual's nationality, birth year, and sex. Simulated death ages were generated from data on average survival of 20-year-olds with a slight negative skew ($\sigma = 1/6$, $\xi = 0.8$). Simulated number of offspring followed a Poisson distribution with mean (λ) set at fertility minus child mortality. Probability of female offspring and probability of a child inheriting the VUS from a heterozygous parent were both set at 0.5. Functions for pedigree simulation and default parameters are documented and publicly available in the CoSeg package.

Simulated Genotypes and Phenotypes

After each pedigree was simulated, genotypes and phenotypes were simulated. Genotypes were simulated from the founder ancestor down the pedigree, while ensuring the proband inherited the VUS from the founder. We assumed all variants would be extremely rare so that simulated families could trace inheritance of a single VUS through their family. Given each genotype, phenotypes at all individuals were simulated using sex and age specific penetrance, assuming a dominant mode of inheritance and published penetrance models [27, 28]. Phenotypes included Lynch syndrome associated cancers (colon, endometrial or minor Lynch tumor defined as grouped small bowel, stomach, ovary, and urinary tract cancers) due to *MLH1* or *MSH2* and breast or ovarian cancer due to either *BRCA1* or *BRCA2*. As the penetrance model for *BRCA1* differs from that of *BRCA2*, separate simulations were performed for each of these genes. Phenotypes for each of these cancers were also simulated assuming the variant was benign, using SEER data for population-based penetrance of disease [20]. Finally, to ensure that the pedigree size was not too large for the analysis method (i.e., $N < 600$), low information individuals were pruned from the pedigree. Low information individuals were defined to be any individual who was neither a first-degree relative of any heterozygous or affected individual, nor an individual who connected heterozygous or affected relatives in the pedigree, as they are unlikely be heterozygous for the VUS. Typically, low information individuals are not ascertained in the clinic or research setting, as they do not have the disease and genotypes of relatives usually make it clear that they do not have the variant of interest, (e.g., if the variant is known to be maternal, paternal relatives are not genotyped). In this way pruned simulated pedigrees are more similar to expected pedigrees.

Simulated Missing Data

In order to assess the effect of missing data, we simulated 8000 each small and large USA families for pathogenic *MLH1/MSH2*, pathogenic *BRCA2*, and the respective benign variants. After the pedigrees were simulated, genotype and phenotype data were removed at random for each individual, with probability 0.5. As before, low information individuals were pruned from the pedigrees.

Statistical Analysis

We calculated a LOD score (LODadj), adjusted for the prior information that the proband is heterozygous for the VUS at a known pathogenic gene [29]. Briefly, the alternative hypothesis is that the VUS is in complete linkage disequilibrium ($LD=1$) with a truly pathogenic variant segregating perfectly with the phenotype (recombination rate (θ) = 0). The null hypothesis for this test is that the gene-specific VUS is not associated with

increased risk of cancer ($LD=0$, $\theta = 1/2$). We used FASTLINK to calculate $LOD_{adj} = LOD_3 + LOD_2 - (LOD_1 + LOD_4)$, where LOD_i is calculated under one of the following conditions:

LOD_1 : $LD=0$, $\theta = 1/2$ and all available genotypes are known.

LOD_2 : $LD=0$, $\theta = 1/2$ and only the proband genotype is known.

LOD_3 : $LD=1$, $\theta = 0$ and all available genotypes are known.

LOD_4 : $LD=1$, $\theta = 0$ and only the proband genotype is known.

Unless otherwise indicated, we assumed a VUS allele frequency of 0.001, as this is the overall frequency of pathogenic alleles in *BRCA1* and is higher than the overall frequency of pathogenic alleles in *BRCA2* or *MLH1/MSH2*. Furthermore, this higher allele frequency allows for the possibility that the VUS has entered the pedigree from someone other than the founder. We analyzed the data using an age and sex based step-wise approximation of the penetrance functions used to simulate the associated phenotype data when the VUS was pathogenic (Supplemental Tables 2, 3 and 4). The approximation breaks the continuous penetrance function into a step-wise function of cumulative risk for ten year periods. For analysis of the benign VUS simulated data, we used the Lynch-specific penetrance matrix for the Lynch-associated phenotypes, and the *BRCA1*-specific penetrance for the breast/ovarian cancer families. Unless otherwise indicated, genotypes are known for live individuals, only.

Power and type I error—To calculate power, we compared LOD_{adj} to cutoffs for variant classification from the international agency for research on cancer (IARC) guidelines[30]: Pathogenic if $LOD_{adj} \geq 2$, Likely Pathogenic if $LOD_{adj} \geq 1.3$, Likely Benign if $LOD_{adj} \leq -1.3$ and Benign if $LOD_{adj} \leq -2$. It is important to note that according to ACMG guidelines there are currently no cutoffs that define the relative strength of evidence from segregation, and segregation alone cannot prove that a VUS is pathogenic. However, segregation scores above reasonable thresholds can be used as moderate or strong evidence when combined with *in vitro*, functional, or evolutionary conservation evidence to establish pathogenicity of a VUS[5].

Linear Model Analysis—We used linear models to evaluate the effect and significance of multiple pedigree descriptors on LOD_{adj} for the large pedigrees. We chose to focus on the large pedigrees because they vary the most in information content and can give the most precise estimates of effect. As these pedigrees are unlikely to reflect those seen in the clinic, we also evaluated the pedigrees that had missing data in order to determine the effect of missingness on the significance of these descriptors. The descriptors of interest included those that can be calculated prior to genotyping individuals in the pedigree: number of affected individuals (N_{aff}), number of unaffected individuals (N_{unaff}), minimum age of onset (MAO), and degree of relatedness between the proband and the furthest affected relative (FAR). N_{aff} , N_{unaff} , and MAO were natural-log transformed. Since these descriptors are correlated, we performed step-wise analysis and evaluated the significance of interaction terms to determine the best model to explain the most variation in LOD_{adj} . We included the interaction terms if they were statistically significant at the 5% level in addition to the

individual terms remaining significant in the presence of the interaction, also at the 5% level. Additionally, in order to estimate the value of genotyping unaffected individuals, we estimated the percent variation in LODadj explained by the natural log transformed number of unaffected individuals heterozygous for the VUS (N_{UH}), separately from the other variables, as it can only be calculated after genotyping.

Results

Simulated Pedigrees

As expected, pedigree size increased with the number of generations, with the maximum size reaching into the mid-thousands (Supplemental Figures 1 and 2, and Table 1). Pruning away the low information individuals substantially reduced the pedigree sizes. Pruned pedigree sizes are larger under the CHI demographic model than under the USA demographic model due to the larger number of individuals in the CHI demography in earlier generations (Supplemental Figure 3). Pruned pedigree sizes are larger for the Lynch syndrome model than for the breast/ovarian cancer model, due to low frequency of affected males heterozygous for the VUS in the *BRCA1/2* pedigrees. Pedigrees with missing data were similarly sized to those with complete data, even after pruning. Additionally, the percent of pedigrees that had only one affected individual (i.e., the proband) decreases with increased number of generations (Tables 2 and 3).

The percent of affected individuals that are heterozygous for the VUS decreases as pedigree size increases (Table 2 and 3). For the pedigrees segregating a pathogenic variant, this is due to larger pedigrees having a higher overall probability of observing an affected individual whose underlying cause is unrelated to the VUS. In simulations where the variant is causal, >46% and >51% of the affected individuals (not including the proband) were heterozygous for the VUS, for Lynch syndrome and breast/ovarian cancer, respectively. In simulations where the VUS is benign, <20% of the affected individuals, not including the proband, are heterozygous for the VUS, for both Lynch syndrome and breast/ovarian cancer.

Power and type I error

Overall, the pedigrees simulated with a pathogenic VUS from the CHI demographic have higher LODadj than the pedigrees from the USA demographic (Supplemental Figure 4). This is likely due to the larger pedigree sizes from the CHI demographic. The power to detect moderate or strong evidence for pathogenicity is highest for *BRCA1* followed by *MLH1/MSH2*, and then *BRCA2* (Figure 1, Supplemental Figure 4, Supplemental Tables 5 and 6). The power and type I error are not affected by analysis with a smaller allele frequency ($q=10^{-6}$, data not shown) or doubling or halving the population based penetrance (data not shown). Larger pedigrees had more power as the proportion of pedigrees with only one affected individual (i.e., the proband) was smallest in this category. The presence of missing data affects the power, reducing by it by ~1/10 for the small families and ~1/3 for the large families (Supplemental Tables 5 and 6). This reduction is due to the even higher percentage of families with a single affected individual.

Overall, the power to assign a truly benign VUS as benign was $\sim 2\times$ higher than the power to assign a truly pathogenic VUS as pathogenic (Supplemental Figure 4, Supplemental Tables 5 and 6). For example, the power for a pathogenic *MLH1/MSH2* variant was 25% whereas the power for a benign *MLH1/MSH2* variant was 53% for the large CHI families. For the families with missing data, the power to correctly classify a benign variant was $\sim 6\times$ higher. None of the families segregating a benign variant yielded evidence supporting variant pathogenicity (type I error for benign variants) regardless of pedigree sizes and types (Supplemental Figure 4, Supplemental Tables 5 and 6). However, some families with simulated pathogenic variants (0–4.2%, depending on trait, pedigree size, and missingness) had $\text{LOD}_{\text{adj}} < -2$ (type I error for pathogenic variants); this “false negative” rate increased with pedigree size.

Linear Model Analysis

Step-wise linear model analysis indicated that N_{aff} explained the most variation (21–38%) in LOD_{adj} for all traits, generally (Table 4). N_{aff} is positively associated with LOD_{adj} when the VUS is pathogenic and is negatively associated with LOD_{adj} when the VUS is benign. Moreover, our simulations indicate that at least 5, 3, and 4 (and typically many more) affected individuals are necessary to achieve a $\text{LOD}_{\text{adj}} \geq 2$ for a VUS pathogenic for *MLH1/MSH2*, *BRCA1* and *BRCA2*, respectively. The addition of FAR to the model explained slightly more variation in LOD_{adj} for both SEER traits (1–5%) and Lynch syndrome (7%). FAR is negatively associated with LOD_{adj} for both SEER traits (benign VUS) and is negatively associated with LOD_{adj} when the VUS is pathogenic for Lynch, only after adjusting for N_{aff} . For both *BRCA1* and *BRCA2* the addition of MAO to the model explained slightly more variation (5–13%). MAO is negatively associated with LOD_{adj} when the VUS is pathogenic for either *BRCA1* or *BRCA2*. The best model included the interaction terms, except for breast/ovarian cancer with a benign variant in the USA demographic. Results were similar for the large USA pedigrees with missing data, for all variants, as well as for the small USA pedigrees with missing data segregating a benign variant.

In contrast, FAR explained the slightly more variation on LOD_{adj} than N_{aff} , for the small USA pedigrees with missing data, segregating a pathogenic variant. When considering all the pedigrees, FAR explained 14.7% and 15.5% of the variation in LOD_{adj} for *MLH1/MSH2* and *BRCA2* pathogenic variants, respectively. When considering the pedigrees with more than one affected individual, FAR explained 10.2% and 10.4% of LOD_{adj} , respectively. Furthermore, MAO and N_{aff} additionally explain ~ 2 and 3% more of the variation, for the *BRCA2* pathogenic variant (including interaction term) and the *BRCA2* benign variant, respectively.

Genotyping unaffected individuals was crucial for correctly identifying benign variants. For families with a benign VUS, N_{UH} explained more of the variance in LOD_{adj} (60%) than the corresponding best model using variables that are not dependent on genotype (37%) (Tables 4 and 5). Furthermore, affecteds-only linkage analysis resulted in much lower power (0.1% and 0.4% for Lynch syndrome and breast/ovarian cancers, respectively) to classify a benign variant correctly (Figure 2).

Discussion

We have shown how different parameters related to demographics of the United States and China influence power for classification of rare VUS in established cancer risk genes using family studies. Marriage age has increased and family size has decreased in both Chinese and American populations since the middle of the 20th century. However, these changes have been more dramatic in the Chinese population. Family size is related to power, as expected, and currently families from the Chinese demographic have higher power than families from the United States demographic. This is likely due to the larger family size in the 1950s and 1960s in China, resulting in overall larger simulated pedigrees, with more affected individuals, and therefore more information. This could suggest that extending pedigrees to have the most distant possible founder heterozygous for the VUS would be advantageous. However, as pedigree size increases, the possibility of phenocopies or other risk alleles related to the same trait also increases, complicating the analysis. Moreover, the more distantly related an affected relative is to the proband, the more likely, that he/she will not be heterozygous for the VUS.

Rather than focusing on recruiting distant relatives, initial recruitment efforts could concentrate on finding as many affected relatives as possible, regardless of the strength of their genetic relationship with the proband. We have shown that the power of single extended pedigrees to classify family specific VUS mostly depends on the number of affected individuals. This is most likely due to the high correlation between the number of affected individuals and the number of informative matings (i.e., where an informative meiosis is possible). This high correlation is explained as follows: 1. The only possible informative matings for these pedigrees involve a common homozygote with a heterozygote, as the VUS is extremely rare and 2. Most affected individuals are expected to be heterozygotes as the number of phenocopies are low. Therefore, most affected individuals represent at least one informative mating (the affected individual's parents) and possibly a second informative mating (the affected individual's mate). In contrast, results from the small families with missing data indicate that recruiting distantly related affected individuals may be crucial to the power. This is likely due to the lower number of affected individuals expected in this scenario.

In general, the degree of relationship to the farthest affected relative and the minimum age of onset in the pedigree, influences power to a lesser extent than the total number of affected individuals. For the pathogenic Lynch variants and both cancer type benign variants, the degree of relationship to the farthest affected relative influences power. This might indicate that efforts at finding distantly related affected individuals could be fruitful in the case when current family structure does not have power to classify a variant for a trait that has similar penetrance among both sexes. However, for the *BRCA1/2* pathogenic variants, the minimum age of onset contributed more to the power. Lower ages of onset may also increase the number of informative matings in the pedigree. Interestingly, *BRCA1* was first localized through the use of pedigrees with low ages onset of breast cancer [31]. This may be due to the fact that very few males have cancers due to *BRCA1/2*, so that only about half of heterozygotes have a non-negligible probability of developing cancer, potentially reducing the availability of distantly related affected heterozygotes. Furthermore, the penetrance for

breast/ovarian cancer in females is higher than the penetrance for Lynch syndrome, indicating that the underlying penetrance does not explain why minimum age of onset explains so much of the variation in power. This might indicate that recruitment efforts and screening should focus on younger individuals when the trait under consideration has much higher penetrance among one sex.

Although the rationale for segregation analysis is to demonstrate that a variant is pathogenic, most VUS, absent any clinical or *in vivo* data on tumor samples, are expected to be benign [12, 13, 32]. In fact, the ability to classify variants as benign may be the most powerful use of co-segregation analysis [33]. We have shown that family analysis has more power to classify a benign variant than a pathogenic variant, when all live individuals are genotyped, regardless of phenotype. The number of unaffected heterozygotes explains the largest portion of variation in LODadj in this situation. Moreover, when only affected individuals are analyzed, the LOD score approach fails to correctly classify benign variants correctly >99% of the time. This error is likely to be exacerbated when meiosis counting methods are used, possibly resulting in false evidence of pathogenicity [6]. Moreover, if strong quantitative evidence that a VUS is benign is reported to ClinVar, this could alleviate future emotional and financial costs when similar private variants are found in other families. Therefore, in addition to genotyping affected individuals, unaffected individuals who have a non-zero probability of being heterozygous for the VUS should also be genotyped. For example, if the proband's unaffected great-aunt is genotyped and does not have the VUS, then pursuing her offspring is not useful. However, if the unaffected great-aunt has the VUS, then she and her offspring, if genotyped, impart useful information.

Identifying and recruiting both affected relatives and potentially heterozygous unaffected relatives requires substantial time and effort. Although only a small percent of the pedigrees with pathogenic variants contained moderate or strong evidence for pathogenicity, the quantitative results can be combined with other data for variant classification [9]. There is almost always quantitative *in silico* data that could be combined with quantitative segregation evidence, even if neither is independently conclusive. In some cases, *in vitro* assays may be available to test the effect of the VUS. These assays can provide independent evidence for or against pathogenicity; however, these data should be combined with caution, particularly with discrepant results. If the quantitative results from co-segregation analysis are in conflict with *in vitro* data, this may indicate that the *in vitro* assay does not capture the underlying cancer biology and perhaps further, or different, *in vitro* analysis is warranted. [34, 35].

Our results indicating which individuals are likely to contribute valuable information are for the large pedigrees as they are likely to provide the most accurate measure of effect size and significance for pedigree descriptors on the resulting LODadj. However, these pedigrees are much larger than those usually tested in the clinic setting. It is possible that the number of affected individuals has a smaller effect size in smaller families. However, we believe the trend is the same: the larger the number of affected individuals, the higher the power. Realistically, many individuals will not be genotyped and second hand knowledge of relatives' phenotypes may not be accurate in the clinic setting, or missing, resulting in less informative pedigrees. In this case, it may be prudent to ascertain distantly related affected

individuals. Similarly, for small families with a trait in which penetrance is virtually zero for one sex, reducing the potential number of affected individuals, ascertaining distantly related affected individuals may have a larger effect on power. Another limitation of our study is that we used known age- and sex-based penetrances, which is unlikely to be the case for all variants. However, for many known pathogenic genes, there are multiple estimates of sex- and age-based penetrances available in the literature [27, 28, 36]. Furthermore, when most genotypes and phenotypes are known, small to moderate changes in the penetrance matrix had virtually no effect on the LOD score in our analysis.

We have presented simulation-based evidence to guide recruitment efforts for individuals considering family studies for VUS classification. Currently, there is little quantitative evidence from family segregation studies reported in ClinVar. We have shown that data from a single family can produce useful information, particularly for benign variants in large families, which could be added to ClinVar. Furthermore, new resources have made family studies more accessible to patients and genetics providers [11] and new statistical tools are available for quantitative family co-segregation analysis (see Web Resources). We hope that these tools will encourage increased reporting of quantitative results from co-segregation analysis in families and that future ACMG-AMP classification guidelines will incorporate them.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank David Goldgar for sharing example FASTLINK input assuming complete linkage disequilibrium and Elizabeth Blue for her thoughtful comments. This work was supported by the Damon Runyon Cancer Research Foundation (DRR-33-15), by the NHGRI (R21HG008513), and by development funds from the Fred Hutch/University of Washington Cancer Consortium (NCI 5P30 CA015704-39).

References

1. Amendola LM, Jarvik GP, Leo MC, et al. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am J Hum Genet.* 2016; doi: 10.1016/j.ajhg.2016.03.024
2. Shirts BH, Jacobson A, Jarvik GP, Browning BL. Large numbers of individuals are required to classify and define risk for rare variants in known cancer risk genes. *Genet Med.* 2013; 19(10): 187.doi: 10.1038/gim.2013.187
3. Consortium EA. ExAC Browser. Vol. 9. Cambridge, MA: 2015 Dec. 2015
4. Lek M, Karczewski K, Minikel E, et al. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv.* 2015; doi: 10.1101/030338
5. 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; 5(10):30.
6. Jarvik GP, Browning BL. Consideration of Cosegregation in the Pathogenicity Classification of Genomic Variants. *Am J Hum Genet.* 2016; 98(6):1077–81. DOI: 10.1016/j.ajhg.2016.04.003 [PubMed: 27236918]
7. Mohammadi L, Vreeswijk MP, Oldenburg R, et al. A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example. *BMC Cancer.* 2009; 9(211)doi: 10.1186/1471-2407-9-211

8. Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS. Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. *Hum Mutat.* 2008; 29(11):1265–72. DOI: 10.1002/humu.20897 [PubMed: 18951437]
9. Bayrak-Toydemir P, McDonald J, Mao R, et al. Likelihood ratios to assess genetic evidence for clinical significance of uncertain variants: hereditary hemorrhagic telangiectasia as a model. *Exp Mol Pathol.* 2008; 85(1):45–9. DOI: 10.1016/j.yexmp.2008.03.006 [PubMed: 18495117]
10. Lindor NM, Guidugli L, Wang X, et al. A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). *Hum Mutat.* 2012; 33(1):8–21. DOI: 10.1002/humu.21627 [PubMed: 21990134]
11. Garrett LT, Hickman N, Jacobson A, et al. Family Studies for Classification of Variants of Uncertain Classification: Current Laboratory Clinical Practice and a New Web-Based Educational Tool. *J Genet Couns.* 2016; doi: 10.1007/s10897-016-9993-2
12. Eggington JM, Bowles KR, Moyes K, et al. A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. *Clin Genet.* 2013; 8(10):12315.
13. Murray ML, Cerrato F, Bennett RL, Jarvik GP. Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical decisions. *Genet Med.* 2011; 13(12):998–1005. DOI: 10.1097/GIM.0b013e318226fc15 [PubMed: 21811163]
14. Rosenthal EA, Ranchalis J, Crosslin DR, et al. Joint linkage and association analysis with exome sequence data implicates SLC25A40 in hypertriglyceridemia. *Am J Hum Genet.* 2013; 93(6): 1035–45. DOI: 10.1016/j.ajhg.2013.10.019 [PubMed: 24268658]
15. Wong FL, Rotter JI. Sample-size calculations in segregation analysis. *Am J Hum Genet.* 1984; 36(6):1279–97. [PubMed: 6517052]
16. Chen WM, Abecasis GR. Estimating the power of variance component linkage analysis in large pedigrees. *Genet Epidemiol.* 2006; 30(6):471–84. DOI: 10.1002/gepi.20160 [PubMed: 16685720]
17. Boehnke M, Omoto KH, Arduino JM. Selecting pedigrees for linkage analysis of a quantitative trait: the expected number of informative meioses. *Am J Hum Genet.* 1990; 46(3):581–6. [PubMed: 2309706]
18. Ott, J. *Analysis of Human Genetic Linkage*. 3rd. Baltimore: Johns Hopkins University Press; 1999.
19. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. *CA Cancer J Clin.* 2016; 66(2):115–32. DOI: 10.3322/caac.21338 [PubMed: 26808342]
20. Howlader, N., Noone, AM., Krapcho, M., et al. *SEER Cancer Statistics Review*. Vol. 25. Bethesda, MD: National Cancer Institute; 2013 Oct. SEER Cancer Statistics Review, 1975–2013.
21. Gapminder World. Stockholm, Sweden: The Gapminder Foundation; 2008 May 11.
22. Fertility rate, total (births per woman) Data. Washington, DC, USA: The World Bank; 2015 May 11.
23. Lee, JZ., Campbell, CD. *Fate and Fortune in Rural China: Social Organization and Population Behavior in Liaoning 1774–1873*. Cambridge, UK: Cambridge University Press; 2007.
24. Lee, JZ., Wang, F. *One Quarter of Humanity: Malthusian Mythology and Chinese Realities, 1700–2000*. Cambridge, MA: Harvard University Press; 2001.
25. Sun, Y. *General Equilibrium Analysis and Simulation of Chinese Social Security System in An Overlapping Generation Model*. University of California; Santa Barbara: 2007.
26. Zeng Y, Vaupel JW, Yashin AI. Marriage and fertility in China: a graphical analysis. *Population and Development Review.* 1985; 11(4):721–36.
27. Wang F, Fang Q, Ge Z, Yu N, Xu S, Fan X. Common BRCA1 and BRCA2 mutations in breast cancer families: a meta-analysis from systematic review. *Mol Biol Rep.* 2012; 39(3):2109–18. DOI: 10.1007/s11033-011-0958-0 [PubMed: 21643751]
28. Quehenberger F, Vasen HF, van Houtwelingen HC. Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. *J Med Genet.* 2005; 42(6):491–6. DOI: 10.1136/jmg.2004.024299 [PubMed: 15937084]
29. Thompson D, Easton DF, Goldgar DE. A full-likelihood method for the evaluation of causality of sequence variants from family data. *Am J Hum Genet.* 2003; 73(3):652–5. Epub 2003 Jul 29. [PubMed: 12900794]

30. Plon SE, Eccles DM, Easton D, et al. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat.* 2008; 29(11):1282–91. DOI: 10.1002/humu.20880 [PubMed: 18951446]
31. Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science.* 1990; 250(4988):1684–9. [PubMed: 2270482]
32. Morris B, Hughes E, Rosenthal E, Gutin A, Bowles KR. Classification of genetic variants in genes associated with Lynch syndrome using a clinical history weighting algorithm. *BMC Genet.* 2016; 17(1):99.doi: 10.1186/s12863-016-0407-0 [PubMed: 27363726]
33. Tricarico R, Kasela M, Mareni C, et al. Assessment of the InSiGHT Interpretation Criteria for the Clinical Classification of 24 MLH1 and MSH2 Gene Variants. *Hum Mutat.* 2016; doi: 10.1002/humu.23117
34. Starita LM, Young DL, Islam M, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics.* 2015; 200(2):413–22. DOI: 10.1534/genetics.115.175802 [PubMed: 25823446]
35. de la Hoya M, Soukarieh O, Lopez-Perolio I, et al. Combined genetic and splicing analysis of BRCA1 c.[594-2A>C; 641A>G] highlights the relevance of naturally occurring in-frame transcripts for developing disease gene variant classification algorithms. *Hum Mol Genet.* 2016; doi: 10.1093/hmg/ddw094
36. Marabelli M, Cheng SC, Parmigiani G. Penetrance of ATM Gene Mutations in Breast Cancer: A Meta-Analysis of Different Measures of Risk. *Genet Epidemiol.* 2016; 40(5):425–31. DOI: 10.1002/gepi.21971 [PubMed: 27112364]

Figure 1. Power and type I error to classify the VUS from linkage analysis adjusting for the proband for Lynch syndrome pedigrees (A, B) and breast/ovarian cancer pedigrees (C, D) from the Chinese (A, C) and USA (B, D) demographics

Thickness of lines increases with pedigree size (Small, Medium, Large). Black lines = pathogenic variant. Gray lines = benign variant. In C and D, the solid lines are results from *BRCA1* and the dashed lines are results from *BRCA2*.

Figure 2. Scatterplots comparing LODadj for full data analysis and the affecteds-only analysis in the USA families

A) Lynch pathogenic, B) Lynch benign, C) BRCA1 pathogenic, D) breast and ovarian cancer benign, and E) BRCA2 pathogenic. The dashed line represents the $x=y$ line. Solid black line indicates the best fit line. Gray lines indicate cutoffs for classifying a VUS as Benign (-2).

Table 1

90th percentiles for number of individuals remaining in pruned pedigrees, over all pedigree sizes. CHI=Chinese demographic, USA=United States of America demographics, LS=Lynch syndrome with pathogenic variant. CE=Lynch syndrome with benign variant. B1=breast/ovarian cancer with pathogenic BRCA1 variant. B2=breast/ovarian cancer with pathogenic BRCA2 variant. OB=breast/ovarian cancer with benign variant.

	LS	CE	B1	B2	OB
CHI	188	189	169	165	165
USA	84	86	77	76	78

Table 2

Statistics for Number of Affecteds per simulated Lynch syndrome pedigree. CHI=Chinese demographics. USA=United States of America demographics. LS=Lynch syndrome with pathogenic variant. CE=Suspected lynch syndrome with benign variant. S=small (founder 2 generations from proband). M=medium (founder 3 generations from proband), L=large (founder 4 generations from proband). S50/L50 = Small/large Pedigrees with 50% missing data. %proband only = percent of pedigrees with only a single affected individual. %affected-heterozygotes = percent of affected individuals who are heterozygous for the VUS, not including the proband.

Simulation	Median	Mean	%proband only	%affected heterozygotes
CHI-LS-S	2	2.9	24	67
CHI-LS-M	5	6.3	8	58
CHI-LS-L	10	13.4	2	46
CHI-CE-S	1	1.8	55	15
CHI-CE-M	3	3.4	24	10
CHI-CE-L	6	8.7	8	6
USA-LS-S	2	2.4	29	72
USA-LS-M	3	4.0	13	65
USA-LS-L	6	7.8	3	55
USA-CE-S	1	1.5	65	17
USA-CE-M	2	2.3	38	14
USA-CE-L	3	4.5	16	10
USA-LS-S50	2	1.8	50	73
USA-LS-L50	4	4.5	12	56
USA-CE-S50	1	1.3	76	20
USA-CE-L50	2	2.7	32	9

Table 3

Statistics for Number of Affecteds per simulated breast/ovarian cancer pedigree. CHI=Chinese demographic. USA=United States of America demographics. B1=breast/ovarian cancer with pathogenic BRCA1 variant. B2=breast/ovarian cancer with pathogenic BRCA2 variant. OB= Suspected breast/ovarian cancer risk with benign variant. S=small (founder 2 generations from proband). M=medium (founder 3 generations from proband), L=large (founder 4 generations from proband). S50/L50 = Small/large Pedigrees with 50% missing data. %proband only = percent of pedigrees with only a single affected individual. %affected-heterozygotes = percent of affected individuals who are heterozygous for the VUS, not including the proband.

Family	Med	Mean	%proband only	%affected heterozygotes
CHI-B1-S	2	2.5	33	75
CHI-B1-M	4	4.8	13	65
CHI-B1-L	7	9.8	3	55
CHI-B2-S	2	2.3	38	73
CHI-B2-M	3	4.1	17	62
CHI-B2-L	7	9.2	5	51
CHI-OB-S	1	1.5	68	11
CHI-OB-M	2	2.3	38	10
CHI-OB-L	4	5.1	15	7
USA-B1-S	2	1.9	46	78
USA-B1-M	2	2.9	21	71
USA-B1-L	4	5.5	8	61
USA-B2-S	2	1.8	45	75
USA-B2-M	2	2.8	23	69
USA-B2-L	4	5.1	9	60
USA-OB-S	1	1.3	78	17
USA-OB-M	1	1.7	54	14
USA-OB-L	2	2.8	27	11
USA-B2-S50	1	1.5	63	78
USA-B2-L50	3	3.0	24	60
USA-OB-S50	1	1.2	87	22
USA-OB-L50	2	1.9	49	10

Table 4

Best linear model fit for each phenotype, by demographic simulation for the large pedigrees. The response variable is LODadj. CHI=Chinese demographic, USA=United States of America demographics, LS=Lynch syndrome with pathogenic variant. CE=Lynch syndrome with benign variant. B1=breast/ovarian cancer with pathogenic BRCA1 variant. B2=breast/ovarian cancer with pathogenic BRCA2 variant. OB=breast/ovarian cancer with benign variant. N_{aff} =Natural log transformed number of affected), FAR=degree of relationship between proband and his/her farthest affected relative, MAO= Natural log transformed minimum age of onset.

Simulation	Best Model	%variation explained (%variation explained by $\log(N_{\text{aff}})$ alone)
CHI LS	$N_{\text{aff}} + \text{FAR} + N_{\text{aff}} * \text{FAR}$	45 (38)
USA LS	$N_{\text{aff}} + \text{FAR} + N_{\text{aff}} * \text{FAR}$	38 (31)
CHI B1	$N_{\text{aff}} + \text{MAO} + N_{\text{aff}} * \text{MAO}$	29 (23)
USA B1	$N_{\text{aff}} + \text{MAO} + N_{\text{aff}} * \text{MAO}$	36 (23)
CHI B2	$N_{\text{aff}} + \text{MAO} + N_{\text{aff}} * \text{MAO}$	35 (26)
USA B2	$N_{\text{aff}} + \text{MAO} + N_{\text{aff}} * \text{MAO}$	26 (21)
CHI CE	$N_{\text{aff}} + \text{FAR} + N_{\text{aff}} * \text{FAR}$	37 (32)
USA CE	$N_{\text{aff}} + \text{FAR} + N_{\text{aff}} * \text{FAR}$	25 (23)
CHI OB	$N_{\text{aff}} + \text{FAR} + N_{\text{aff}} * \text{FAR}$	33 (28)
USA OB	$N_{\text{aff}} + \text{FAR}$	17 (16)

Table 5

Results of linear model of $\text{LOD}_{\text{adj}} \sim N_{\text{UH}}$, for benign VUS. N_{UH} = Natural log transformed number of unaffected heterozygotes. CHI=Chinese demographic, USA=United States of America demographics. CE=Lynch syndrome with benign variant. OB=breast/ovarian cancer with benign variant. Each model is statistically significant, with $p < 2e-16$.

Phenotype	Beta	%Variation explained
CHI CE	-3	70
USA CE	-2	65
CHI OB	-3	60
USA OB	-2	60