Leveraging large datasets accumulated through population carrier screening to inform variant classification

David Tran, Samuel Cox, K. Eerik Kaseniit, Elizabeth Collins, Matthew Meredith, Erik Zmuda

Myriad Women's Health

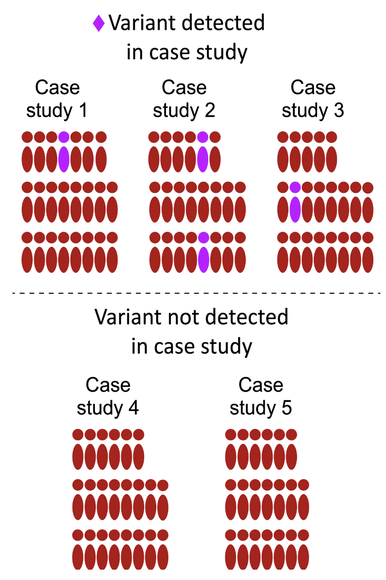
BACKGROUND

- Sequencing-based genetic testing requires real-time interpretation of variants encountered in patients.
- We have performed expanded carrier screening of recessive Mendelian disease genes for >1,000,000 samples, detecting >1,400,000 distinct variant alleles.
- Our variant database includes a large number of curated case studies garnered from the literature along with allele frequencies from our sample cohort.
- Here we demonstrate how the database enables accurate variant interpretation and improves variant classification methodologies.

Figure 1. Method for collating database to improve variant pathogenicity assessment

A. Since manual curation typically reviews studies that report the variant of interest, we accounted for disease studies that failed to detect the variant of interest (assuming the variant was detectable by the study methodology).

B. To improve the accuracy of variant/disease association tests, we calculated an estimate of the "total disease alleles" per gene/disease pair using all disease studies curated in our database with ≥20 cases.



METHODS

- We collated data from more than 40,000 manual reviews of literature as well as database evidence according to variant.
- We filtered for studies with ≥20 cases with a valid PubMed ID, resulting in 2,672 studies encompassing 267 genes (Figure 1).
- We focused on 154 primarily single-gene diseases to estimate total disease allele counts based on our filtering criteria.
- To identify and remove potential double-counting of disease alleles, we assessed whether publications reported on the same or overlapping cohorts.
- As an additional quality control measure, we compared total disease allele counts from this study with the disease-causing mutation alleles (DMs) in the HGMD database.
- ***Study Caveats:*** Studies were excluded that were too small and/or did not have a PubMed standard identifier. Counts will therefore be an underestimate. Secondly, some diseases are poorly represented in the literature. Lastly, patient ethnicity is often not reported, making it difficult to derive ethnicity-specific allele counts.

RESULTS

- Our estimates ranged from 32 disease alleles for *HYSL1*-associated Hydrolethalus syndrome to 83,436 alleles for *MEFV*-associated familial Mediterranean fever.
- 26% (426/1,639) of case studies with overlapping authors were confirmed to have shared study cohorts, which our curation SOP guards against.
- Genes with totals falling below the HGMD entry total were deemed to fail quality control (Figure 2).

Figure 2. Quality control step: Comparison of affected case allele counts (blue) and HGMD disease-causing mutation (DMs) class variant counts (orange)

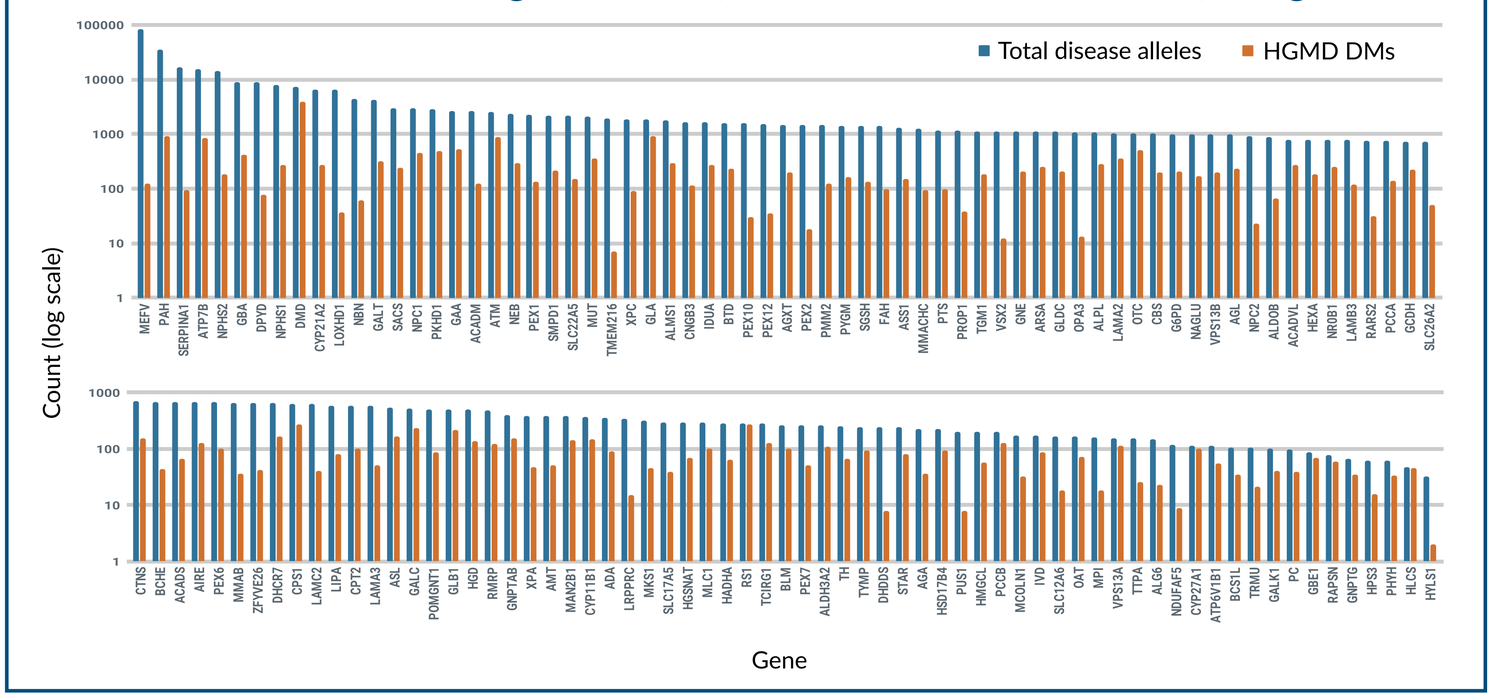


Figure 3. NM_000053.3(ATP7B):c.2972C>T(p.Thr991Met) case example

A. 9 variant disease alleles / 1,290 disease alleles vs 317 variant alleles / 265,732 gnomAD alleles — p=0.00004 — **SIGNIFICANT**

B. 9 variant disease alleles / **15,791** disease alleles vs 317 variant alleles / 265,732 gnomAD alleles — p=0.0216 — **NOT SIGNIFICANT**

C.

			# Co-Occuring	F	M
5 Alleles 3 Females 2 Males	KNOWN_DELETERIOUS	NM_000053.3(*ATP7B*):c.3207C>A	1	0	1
	KNOWN_DELETERIOUS	NM_000053.3(*ATP7B*):c.2305A>G	1	1	0
	LIKELY_DELETERIOUS	NM_000053.3(*ATP7B*):c.1568T>A(L523*)	1	0	1
	LIKELY_DELETERIOUS	NM_000053.3(*ATP7B*):c.2332C>A	1	1	0
	LIKELY_DELETERIOUS	NM_000053.3(*ATP7B*):c.2731-2A>G	1	1	0

- The *ATP7B* T991M variant is reported in 9/1,290 disease alleles in case studies compared to 317 variant alleles in the gnomAD database (population=ALL) (Figure 3A).
- We estimate that at least 15,791 Wilson's disease alleles have been reported in the literature (across multiple ethnicities) (Figure 3B).
- 2×2 Fisher exact test p-values differ in significance depending on which disease allele count is used.
  - The 15,791 *ATP7B* disease alleles estimate bring the p-value above the significance threshold of 0.0001.
- In support of this analysis, T991M co-occurs with 5 deleterious *ATP7B* variants in presumably healthy individuals undergoing carrier screening (Figure 3C).

CONCLUSION

- Through the course of population-scale expanded carrier screening, we have developed a database resource of literature cases, allele frequencies, and variant co-occurrence.
- Our database is used to supplement variant classification protocols and improve assessment of variant pathogenicity.

Presented at ASHG on October 16, 2019.