Protective rare variants of rare diseases are inherently undetectable. My previous work looking at the possible correlations between suggested this was the case, as only slight negative correlations are possible while perfect correlation of (1) is always a possibility. Here, however, I prove this statement rigorously with an information-theoretic argument. I show that a Mendelian variant provides a strong amount of evidence even with low prevalences over a neutral mutation. However, even a perfectly protective variant provides no evidence over the null hypothesis even with prevalences up to 10%.

# Methods

We assume that the prevalence of the disease and the variant are both , which amounts to fixing both marginals of the joint distribution of disease and variant. The contingency table for a neutral allele is

|  |  |  |  |
| --- | --- | --- | --- |
| Variant\Disease | - | + | Marginals |
| - |  |  |  |
| + |  |  |  |
| Marginals |  |  | 1 |

While that of a Mendelian disease is

|  |  |  |  |
| --- | --- | --- | --- |
| Variant\Disease | - | + | Marginals |
| - |  |  |  |
| + |  |  |  |
| Marginals |  |  | 1 |

And a perfectly protective disease is

|  |  |  |  |
| --- | --- | --- | --- |
| Variant\Disease | - | + | Marginals |
| - |  |  |  |
| + |  |  |  |
| Marginals |  |  | 1 |

Detecting whether a variant is protective/causal amounts to obtaining evidence away from independence. Since we have an observational study here, both are response variables. In this case, the rigorous information provided by a single sample is the KL divergence from the independent distribution. The KL divergence is defined as

Which will always be non-negative. We can compute the divergences for a Mendelian disease an a perfectly protective disease as

And

These may not look very different but

While

So this curve has positive slope as for a Mendelian variant (in fact slope goes to infinity) but has slope 0 for a protective variant. To translate, even for the rarest of variants, we’re still obtaining information but we hardly extract information from even moderately common variant/disease combinations. This is supported by a visualization of the information zoomed in below.

A graph with a line

Description automatically generated

# Second intuition

The way we can tell the difference between the two distributions (independent vs alternative) is by shifts of probability mass. By assumption, the prevalence of the disease is . A perfectly protective mutation shifts mass mainly from people having the disease and variant to having the variant but no disease, which corresponds to moving mass from a category un as well as shifting people who have no variant and no disease to people who have no variant and have the disease.

# Further ramifications – Hypothesis testing

# Conclusions

Protective variants of rare disease exist, we’ll just never know without sample sizes greater than the number of atoms in the universe.