# Proposal review:
# The Prevalence of Recessive Traits in Indian Subpopulations

## proposed by:
## Jake Shapiro and Neal Wadhwa

I like this proposal and think that your ideas have the potential to make a real difference in the field. The significance and scope of the work are obvious. The innovation is clear although specific discussion of where your paper will diverge from Reich would be useful in the approach section.

My only complaint would be that you will need to flesh out the mathematical analysis suggested as being the "bulk of your paper" in the approach section of the proposal.

If as was suggested in the verbal review, analysis will require clustering analysis of binary vectors, then you will probably need to refine your approach substantially. While you are probably correct to assume that it will not be possible to naively cluster sparse binary vectors by comparison of their individual elements but it seems to me highly unlikely that the clusters derived from projecting all variation on to one dimension will produce a meaningful result.

This is mind, it seems to me that there are several other heuristic methods you could apply. In order of increasing abstraction, a few suggest themselves; perhaps some would work for your application in some flavor.

1. Suppose your binary vectors are of length $n$. If as you suggest, clustering the binary vectors in $n$ dimensions will likely fail, why not, in the fashion of "guilt by association" replace the binary vectors $\mathbf{b}$ by:

$$\begin{aligned}\mathbf{x} =& P\mathbf{b}\\ P_{ij} =& \frac{(\#\text{appearances of } r_i, r_j \text{ together})^2}{(\#\text{appearances of } r_i) \times (\#\text{ appearances of } r_j)}\end{aligned}$$

   and cluster $\mathbf{x}$'s?

2. Depending on the sparsity of your binary vectors, $P_{ij}$ may still be relatively sparse. Since the previous approach basically suggests transforming $\mathbf{b}$ by an affinity matrix, you could imagine trying the same fundamental approach using different affinity matrices. Since from any distance measure you can construct a corresponding affinity matrix

(and vice-versa) it is also possible to tackle the problem by deriving a distance between elements of **b**.

One concievable way to compute a distance measure between different recessive traits in a sparse matrix is based on degrees of seperation. Thus you could compute an adjacency matrix:

$$A_{ij} \equiv \begin{cases} 1 \text{ if } P_{i,j} > 0 \\ 0 \text{ if } P_{i,j} = 0 \end{cases}$$

and define a distance matrix:

$$D_{ij} \equiv \min_{k>0} \left\{ k | \left( \left( A^k \right)_{ij} > 0 \right) \right\}$$

and cluster using this distance measure.

3. Alternately, considering that your method suggested amounts to projecting **b** on to one dimension, you could tweak your method by projecting **b** to a subspace with dimension $d$:

$$n > d > 1 \tag{1}$$

You could derive this projection from a clustering of recessive traits themselves or by other means.

Otherwise, keep up the good work. Seems like a good project.