SUPPLEMENTAL MATERIAL

Supplemental Methods

Further Explanation of the Primary Analysis Using Monte Carlo Permutation Testing

We were interested in comparing the number of distinct phenotypes in women and men and testing whether more phenotypes appear in women. But simple counting was confounded by the unequal sample sizes. Obviously, we should have expected to see more phenotypes in the sample of women simply because there are more women due to the 2:1 enrollment ratio in the VIRGO study. Therefore, we asked the question, “how can we tell if there's an effect of sex over and above any artifact caused by the differing sample sizes?”

A statistical method known as permutation testing can address this problem.

Suppose that AMI phenotypes were actually independent of sex. If that were true, there would be nothing special about splitting the participants by sex versus splitting them in any other way that results in groups of the same sizes. So following the standard logic of frequentist inference testing, if we knew the distribution of the difference in phenotype numbers under the null hypothesis of no sex differences (but with the unequal sample sizes in place), we could look at the difference we actually did observe and decide whether to reject the hypothesis that only the sample size difference, not sex, was at work in the observed data.

While it is not at all clear how one might derive a theoretical distribution under the null hypothesis, it is rather easy to empirically approximate that distribution using the observed data. We can randomly split the observed data into two groups, A and B, where the sizes of A and B equal the number of women and men in the VIRGO dataset, but membership in A or B is determined randomly

with no regard to sex, to see what the difference in the number of phenotypes is. We can do this again and again, a hundred thousand times, to empirically build up what our distribution would be if phenotypes were independent of sex (the test is called a permutation test because we randomly permute the group labels attached to the individual phenotypes). Technically, it is a Monte Carlo permutation test because there are far too many permutations to look at them all, so we take a random sample of them.

With the distribution of differences between Groups A and B under the null hypothesis in hand, we are able to look where the difference falls in that distribution when group A is actually women and group B is actually men. Just as we do for traditional test statistics such as *t* or *F*, we look at what proportion of the distribution under the null hypothesis has a value at least as extreme as what we observed and use that proportion as our *p* value. Since the statistic we're using here is two-tailed – in principle, there could be more phenotypes in women or in men – using the proportion directly leads to a one-tailed test, which is reasonable given our directional hypothesis that there are more phenotypes in women than in men. To perform a two-tailed test, we would divide the proportion by 2 to obtain *p* if our observed data were in the upper half of the distribution or subtract the proportion from 1 and then divide by 2 if our observed data were in the lower half. And although this is not tied to the statistical inference, the median of the distribution under the null hypothesis can be interpreted as a point estimate of how big the difference is that we would expect to see just because of the unequal sample sizes, giving us a rough sense of how much of the observed difference might be attributable to sex.

Although we have chosen to investigate the differences in the number of phenotypes appearing in women and men, this permutation framework can be used to examine any statistic that can be calculated from the data, so long as sex-independence is the proper null hypothesis. Another possible statistic to test is the Pearson chi-square, treating the space of phenotypes as the rows of a contingency table and the sexes (or groups A and B) as the columns (or vice versa). This test asks a different question

- whether the distribution of phenotypes is different in men and women - but retains the null hypothesis of sex-independent phenotypes.

The different sample sizes for women and men is not an obstacle here, but for this analysis, a permutation test is needed for a different reason. Because there are many cells with small (or even zero) frequencies, the test statistic, despite being calculated using the standard formula for a Pearson chi-square, will not have an asymptotic chi-square distribution. Instead of relying on the standard distribution, we can use our collection of permuted datasets to approximate the true distribution of this statistic under the null (i.e., calculate a Pearson chi-square for every permuted data set, using groups A and B rather than sex) and use that for significance testing. This procedure is essentially a Monte Carlo approximation to Fisher's exact test when the number of categories is too large for an exact test to feasibly be calculated.

However, there is a deeper issue to consider with this test. There are 1,024 possible phenotypes, many of which either do not appear in the data at all or only appear in one sex (or in one group for the permuted datasets). We have no way to tell which of these empty cells are structural (i.e., the phenotypes do not exist in the population, so their cells' occupancies will be zero in every sample) and which ones are random (i.e., the phenotypes are rare and we did not see them in this sample but might see them in a different sample). In practice, we use the observed frequency tables to calculate the chi-squares. This effectively treats phenotypes that appear in neither group as structural zeroes (they may not be – some of them certainly are not, unless the VIRGO sample managed to capture every phenotype present in the human race – but if we tried treat them as random, we would calculate an infinite chi-square because entire rows or columns of the contingency table would have zero counts). Also, this effectively treats the zeroes in only one group as random (which is probably fair for the permutations, although some of those zeroes may actually be structural in the observed data where the groups are sexes). In the end, a Monte Carlo chi-square would actually test sex independence only

among the subset of phenotypes that appear somewhere in the VIRGO dataset, leaving open the question of how well that result generalizes to phenotypes that may exist in the population but are too rare to be detected in this dataset.