1. Posterior over conditional frequencies in large populations

We start of by considering the empirical fraction of individuals with a given insomnia symptom (Onset, Maintenance or Terminal) that have alleles a and b. We denote these conditional empirical frequencies by F_a and F_b. Since these empirical frequencies are estimated from a finite sample, our goal is to infer the distribution over the frequencies f_a and f_b that we would have got if the population was arbitrarily large. We denote this distribution by $p(f_a, f_b|F_a, F_b, I)$. Here I is any set of information available for inference, including, for instance, N_a and N_b, the number of individuals in the sample with alleles a and b respectively in the data.

This distribution of interest can be written using the Bayes rule as

$$p(f_a, f_b|F_a, F_b, I) \propto p(F_a, F_b|f_a, f_b, I)p(f_a, f_b|I) \tag{1}$$

The first term on the right hand side of Eq. 1 can be easily estimated as

$$p(F_a, F_b|f_a, f_b, I) = \prod_{x=a,b} \binom{N_x}{N_x F_x} f_x^{N_x F_x}(1 - f_x)^{N_x(1-F_x)} \tag{2}$$

The second term in the right hand side of Eq. 1 reflects our prior belief about the actual large population limit frequencies. In analyzing the data in this paper we have considered two very different priors. The first prior is a uniform prior which assigns one to every possible outcome, namely

$$p(f_a, f_b|I) = 1 \tag{3}$$

The second prior is a mixture hierarchical prior which is represented as

$$p(f_a, f_b|I) = \int_0^\infty du \int_0^\infty dv p(f_a, f_b|u, v, I)p(u, v|I) \tag{4}$$

This is a mixture of probabilities $p(f_a, f_b|u, v, I)$ with parameters u and v weighted by a distribution over these parameters $p(u, v|I)$. For the probability distribution $p(f_a, f_b|u, v, I)$ we choose the beta distributions,

$$p(f_a, f_b|u, v, I) = \beta(u, v|I)\beta(u, v|I) \tag{5}$$

The choice of the beta distribution in this case is justified by the fact that the beta distribution is the conjugate prior to the binomial distribution in Eq. 2 making the calculation of the posterior in Eq. 1 straightforward.

The weighting term in the right hand side of Eq. 4 is chosen to be a gamma distribution as

$$p(u, v|I) = \gamma(u + v|1, 1000) = \frac{1}{1000} \exp\left\{\frac{u + v}{1000}\right\} \tag{6}$$

The weighting term Eq. 6 in Eq. 4 signals our lack of knowledge about how much data we need to win over the prior: the parameters u and v in the beta distributions in Eq. 5 determine such amount of data. We do not want to commit to a specific values for these parameters and rather consider a broad range for them. This is codified in the gamma distribution in Eq. 6 which gives a weight only to positive values of these parameters and is scale invariant for them in the log scale. Although we have used 1000 as the parameters in the exponential in Eq. 6 to concentrate the distribution over f_a and f_b along the diagonal, other choices of this parameter, e.g. 1, 100 or 1000000 yield quantitatively similar results.

With the priors in Eqs. 3 and 4 we can now calculate the posterior distributions in Eq. 1. The expression for the uniform prior can be calculated analytically and would be a product of two beta functions with updated parameters, while for the case of the hierarchical prior in Eq. 4, one obtains an integral form which should be calculated using Monte Carlo.

2. Measure of Significance

Once we have calculated the posterior in Eq. 1, we can determine pairs of alleles which show significant difference in their large population frequency differences f_a and f_b. As the main measure of allele frequency difference we consider $E[|f_a - f_b|]$, namely the expectation of the absolute value of the difference between the allele frequencies. This expected value can either be directly calculated from the samples of the distribution $p(f_a - f_b)$, or from first estimating the mean and standard deviation of this distribution and calculating the expected value of the difference using a Gaussian approximation to $p(f_a - f_b)$.

Although the expected value of the differences is the main quantity according to which we have sorted the genes in this paper, we note that other measures of significant difference, e.g. sorting the genes according the lowest 10-quantile of $p(f_a - f_b)$, yield quantitatively the same results.