FILE S1

MATERIALS AND METHODS

Estimation of population MAF in OWAS

There can be several ways to estimate population MAF. For example, MB estimates it from control individuals (MADSEN and BROWNING 2009). We choose to estimate population MAF in the following way. Denoting p_i as population MAF of variant i, we first assume that its true overall sample frequency is equal to observed overall sample frequency.

$$p_i^+ + p_i^- = \hat{p}_i^+ + \hat{p}_i^- \tag{17}$$

p_i^+ and p_i^- are defined in terms of p_i (Equations 3, 4), and we can rewrite (Equation 17) as

$$\frac{\gamma_i p_i}{(\gamma_i - 1)p_i + 1} + p_i = 2\hat{p}_i^{\pm} \quad \left(\text{where } \hat{p}_i^{\pm} = \frac{\hat{p}_i^+ + \hat{p}_i^-}{2} \right) \tag{18}$$

We can compute p_i in terms of γ_i and $\hat{p}_i^{\pm}$ by finding the root of (Equation 18).

$$p_i = \frac{b + \sqrt{b^2 + 8(\gamma_i - 1)\hat{p}_i^{\pm}}}{2(\gamma_i - 1)} \quad \text{where } b = 2\hat{p}_i^{\pm}(\gamma_i - 1) - (\gamma_i + 1) \tag{19}$$

Approximation of x_i, $\hat{\mu}_i$ and $\hat{\sigma}_i$ of MB

In this section, we show that x_i, $\hat{\mu}_i$ and $\hat{\sigma}_i$ of MB can be approximated as (Equation 12). First, MB calculates a weight of variant i ($\hat{w}_i$) as

$$\hat{w}_i = \sqrt{N \cdot q_i(1 - q_i)} \quad \text{where } q_i = \frac{m_i^U + 1}{2n_i^U + 2} \tag{20}$$

N is the total number of case and control individuals, m_i^U is the number of mutations for variant i in control individuals, and n_i^U is the number of control individuals.

MB then calculates the genetic score (γ_j) of each individual j.

$$\gamma_j = \sum_{i=1}^{M} \frac{I_{ij}}{\hat{w}_i}$$

where M is the number of variants, and I_{ij} is the number of mutations observed in individual j at variant i. MB ranks all individuals (both cases and controls) by their genetic scores and calculates the sum of the ranks of cases as its test statistic (x).

$$x = \sum_{j \in \text{cases}} \text{rank}(\gamma_j)$$

Madsen and Browning reports that x can also be computed using the sum of genetic scores instead of the sum of ranks, and the two methods have very similar power. Hence, we will compute x as the sum of genetic scores.

$$x = \sum_{j \in \text{cases}} \gamma_j$$

First, we observe that the sum of genetic scores of cases is equivalent to the sum of observed MAF of each variant in cases divided by the weight of the variant. In other words, we sum the number of mutations per variant instead of the number of mutations per individual.

$$\sum_{j \in \text{cases}} \gamma_j = \sum_{i=1}^{M} \frac{\mathcal{N}/2 \cdot \hat{p}_i^+}{\sqrt{\mathcal{N} q_i (1-q_i)}} \tag{21}$$

Assuming $q_i \approx \hat{p}_i^-$ since q_i is an estimate of MAF of variant i in controls, the statistic of variant i, x_i, in (Equation 21) is

$$x_i = \frac{\sqrt{\mathcal{N}}}{2} \frac{\hat{p}_i^+}{\sqrt{\hat{p}_i^- (1-\hat{p}_i^-)}} \tag{22}$$

Next, we derive the statistic of the null distribution denoted as x_i^*. First $\hat{p}_i^+$ and $\hat{p}_i^-$ have the following distribution under the null distribution.

$$\hat{p}_i^+ \sim \mathsf{N}\left(p_i, \frac{p_i(1-p_i)}{\mathcal{N}/2} \right) \tag{23}$$

$$\hat{p}_i^- \sim \mathsf{N}\left(p_i, \frac{p_i(1-p_i)}{\mathcal{N}/2} \right) \tag{24}$$

By multiplying $\hat{p}_i^+$ in (Equation 23) by $\frac{\sqrt{\mathcal{N}}}{2\sqrt{\hat{p}_i^- (1-\hat{p}_i^-)}}$ and assuming $\hat{p}_i^- \approx p_i$, we can derive x_i^* that is approximately equivalent to x_i in (Equation 22). x_i^* and its distribution are then

$$x_i^* = \frac{\sqrt{\mathcal{N}}}{2} \frac{\hat{p}_i^+}{\sqrt{\hat{p}_i^- (1-\hat{p}_i^-)}} \approx \frac{\sqrt{\mathcal{N}}}{2} \frac{\hat{p}_i^+}{\sqrt{p_i(1-p_i)}} \sim \mathsf{N}\left(\frac{\sqrt{\mathcal{N}}}{2} \frac{p_i}{p_i(1-p_i)}, \frac{1}{2} \right) \tag{25}$$

Thus, the mean ($\hat{\mu}_i$) of x_i^* is $\frac{\sqrt{\mathcal{N}}}{2} \frac{p_i}{p_i(1-p_i)}$, the standard deviation ($\hat{\sigma}_i$) is $\sqrt{1/2}$, and x_i is (Equation 22).

LITERATURE CITED

MADSEN, B. E. , and S. R. BROWNING, 2009 A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.