Identify IBD regions with rare variants

Eric Zhang Lu

August 13, 2016

1 Implementation

The genome is divided into windows (w, may be overlapped each other), for the genotypes (G_1,G_2) of rare variants(minor allele frequency$< 0.1\%$) in w, we have

$$\begin{aligned} P(G_1, G_2|IBD, D) &= \Sigma_{h_1,h_2,h_3} P(G_1|h_1, h_2)P(G_2|h_2, h_3)P(h_1, h_2, h_3|D) \\ P(G_1, G_2|\overline{IBD}, D) &= \Sigma_{h_1,h_2,h_3,h_4} P(G_1|h_1, h_2)P(G_2|h_3, h_4)P(h_1, h_2, h_3, h_4|D) \end{aligned} \tag{1}$$

where $h_i \in \{0, 1\}$. If we consider h_i follows Bernoulli distribution ($B(n, \theta)$) and is independent from each other, then

$$\begin{aligned} P(h_1, h_2, h_3|D) &= \int_0^1 P(h_1, h_2, h_3|\theta)P(\theta|D)d\theta \\ &= \int_0^1 P(h_1|\theta)P(h_2|\theta)P(h_3|\theta)P(\theta|D)d\theta \\ &= \int_0^1 \theta^{h_1}(1-\theta)^{1-h_1}\theta^{h_2}(1-\theta)^{1-h_2}\theta^{h_3}(1-\theta)^{1-h_3}P(\theta|D)d\theta \end{aligned} \tag{2}$$

Assuming $\theta|D$ follows Beta distribution($\beta(\alpha_{update}, \beta_{update})$) $\alpha_{update} = \alpha_{prior} + 2N_{training} + 2N_{test} - t$, $\beta_{update} = \beta_{prior} + t$, where t is the number of individuals hit by the variants in test set, $N_{training}$ and N_{test} are the number of individuals in training and test set. α and β are shortened form of α_{update} and β_{update} in the following text.

$$P(\theta|D) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{3}$$

Let's say $S = h_1 + h_2 + h_3$, Eq. 2 can be rewritten as

$$\begin{aligned} P(h_1, h_2, h_3|D) &= \int_0^1 \theta^{h_1+h_2+h_3}(1-\theta)^{3-h_1-h_2-h_3}\theta^{\alpha-1}(1-\theta)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}d\theta \\ &= \int_0^1 \theta^{S+\alpha-1}(1-\theta)^{2+\beta-S}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{S+\alpha-1}(1-\theta)^{2-S+\beta}d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{(2-S+\beta)!}{(S+\alpha)(S+\alpha+1)...(\alpha+\beta+2)} \end{aligned} \tag{4}$$

2 Beta distribution without sequencing error

Because $P(h_1, h_2, h_3|D)$ has included sequencing error in estimating θ, we assume k_1 and k_2 hits are observed for errors from ref− >alt and alt− >ref,respectively:

$$\Sigma_{k_1=0}^{\beta}\Sigma_{k_2=0}^{\alpha}P_{\epsilon,\beta}(k_1)P_{\epsilon,\alpha}(k_2)Beta(\alpha-k_2+k_1, \beta-k_1+k_2) =$$
$$\Sigma_{k_1=0}^{\beta}\Sigma_{k_2=0}^{\alpha}\binom{\beta}{k_1}\binom{\alpha}{k_2}\epsilon^{k_1+k_2}(1-\epsilon)^{\alpha+\beta-k_1-k_2}\frac{\theta^{\alpha+k_1-k_2-1}(1-\theta)^{\beta+k_2-k_1-1}\Gamma(\alpha+\beta)}{\Gamma(\beta+k_2-k_1)\Gamma(\alpha+k_1-k_2)} \tag{5}$$

3 The likelihood ration test for the shared rare variants in each window

This section provides a approach to find the IBD regions by evaluating the shared rare variants in each window. Assuming for each window we have k rare variants and $Q = 1 - 2p^2(1-p)^2 - (1-p)^4$ to denote two individuals shared a rare variant by random chance. $S = 1$ if the two individuals share the rare variant, otherwise $S = 0$. In IBD region, if two individuals are observed to share rare variants,there are half chance to be appeared on the IBD haplotype (H_2), see Figure 1.

Then we have:

$$\begin{aligned} \frac{P(D|IBD)}{P(D|\overline{IBD})} &= \prod_{i=1}^{k}\frac{P(d_i|IBD)}{P(d_i|\overline{IBD})} \\ &= \prod_{i=1}^{k}(\frac{1+Q}{2Q}S + \frac{1}{2}(1-S)) \\ &= \prod_{i=1}^{k}\frac{S+Q}{2Q} \end{aligned} \tag{6}$$

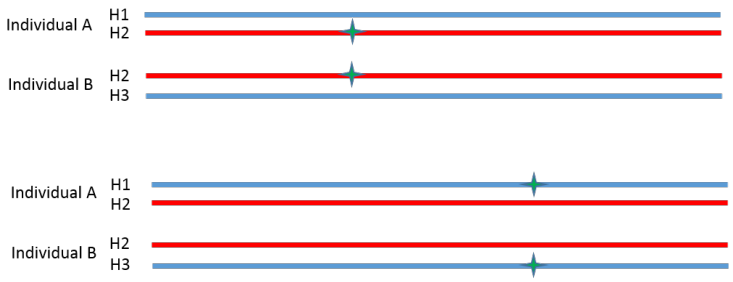
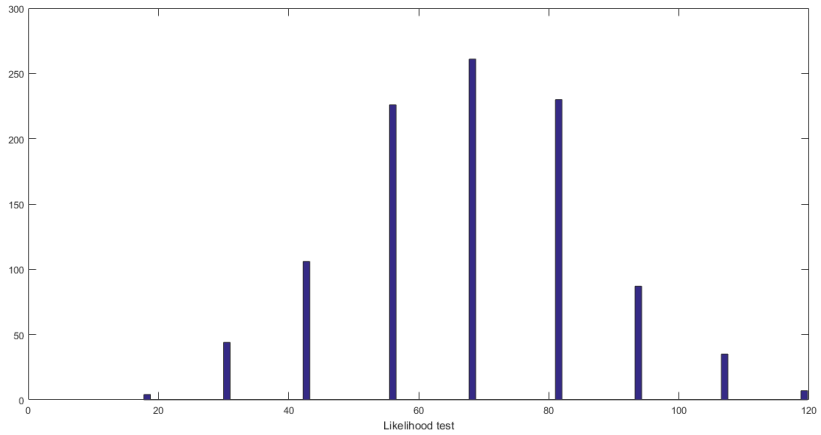


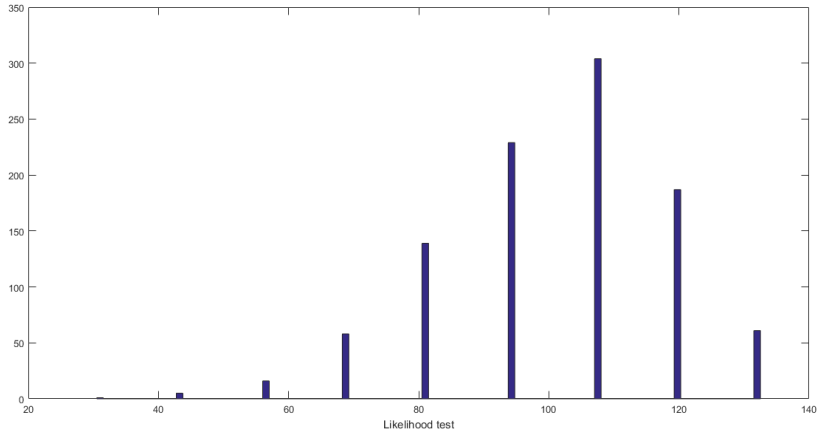
Figure 1: The two scenarios for sharing rare variants in IBD region.

3.1 Simulation result

We performed a simulation to compare the proposed likelihood ratio test. For the non-IBD region, $\frac{1}{2}$ probability is simulated for the rare variants shared by individual pair. For IBD region, three haplotypes are simulated and rare mutations hit each haplotype by half chance. We simulated 1,000 individual pairs and the minor allele frequency is fixed as 0.01.



(a) Random



(b) IBD

Figure 2: The likelihood ratio test comparing random and IBD regions.