Identify IBD regions with rare variants

Eric Zhang Lu

August 8, 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:

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