

freqpcr: estimation of population allele frequency using qPCR $\Delta\Delta C_q$ measures from bulk samples

Masaaki Sudo, Masahiro Osakabe

Appendix

Appendix S1: Case of Diploidy

Although we considered sampling from haploid organisms, many insects and vertebrates are diploid. Let us consider that the population of a diploid insect species has the R allele frequency p , from which we collected n individuals. The bulk sample then consists of m_1 ($m_1 = 0, 1, \dots, n$) individuals of RR homozygotes, $n - m_1 - m_0$ RS heterozygotes, and m_0 ($m_0 = 0, 1, \dots, n$) SS homozygotes ($m_1 + m_0 \leq n$). The joint probability of obtaining $\{m_1, m_0\}$ obeys the trinomial distribution with probabilities p^2 and $(1 - p)^2$

$$\text{Tri}(m_1, m_0 | n, p^2, (1 - p)^2) = \frac{n!}{m_1! m_0! (n - m_1 - m_0)!} \cdot p^{2m_1} \cdot (1 - p)^{2m_0} \cdot (2p - 2p^2)^{(n - m_1 - m_0)}.$$

Eq. 15

The total R allele in the bulk sample comes from two R/R sets contained in the m_1 homozygotes and a single set of R from the $n - m_1 - m_0$ heterozygotes. Likewise, two S/S sets from m_0 homozygotes and a single S set from the $n - m_1 - m_0$ heterozygotes constitute the total S body. Note that the yields of R and S from these heterozygotes would be the same unless there is a genotype-dependent systematic error in the extraction efficiency.

Let us define the amount of DNA copies per genome: the random variable $X_{* \in \{S, R\} | \text{homo}}$ for the yield of S or R from the homozygotes, and $X_{* \in \{S, R\} | \text{hetero}}$ for S or R from the heterozygotes. As in the case of haploidy, X_R and X_S denote the allele contents in the bulk sample; they are the linear combinations of $X_{* | \text{homo}}$ and $X_{* | \text{hetero}}$:

$$\begin{aligned} X_R &= 2 \times X_{R | \text{homo}} + X_{R | \text{hetero}}, & X_S &= X_{S | \text{hetero}} + 2 \times X_{S | \text{homo}}, \\ 2 \times X_{R | \text{homo}} &\sim \text{Ga}(m_1 k, 2\theta), & X_{R | \text{hetero}} &\sim \text{Ga}((n - m_1 - m_0)k, \theta), \\ X_{S | \text{hetero}} &= X_{R | \text{hetero}}, & 2 \times X_{S | \text{homo}} &\sim \text{Ga}(m_0 k, 2\theta). \end{aligned}$$

Eq. 16

Parameter estimation

There are $n - i + 1$ cases from $m_0 = 0$ to $m_0 = n - i$ when the number of RR homozygotes is given by $m_1 = i$. The segregation ratio in the bulk sample has $\sum_{i=0}^n (n - i + 1)$ total combinations. For each combination of n , m_0 , and m_1 , Eq. 16 gives the probability of obtaining the ΔC_q measures in Eq. 11. However, a drawback arises from the constraint of the amounts of R and S possessed by heterozygotes. The applicability of the likelihood model (Eq. 13 or Eq. 14 in the main text) depends largely on the independence of X_R and X_S . If we define the likelihood using Eq. 16 as it was, we must convolve the DNA amounts not on

33 the two-dimensional parameter space spanned by X_R and X_S , but a three-dimensional space by $X_{R|homo}$,
 34 $X_{S|hetero} = X_{R|hetero}$, and $X_{S|homo}$, which would increase the calculation time.

35 Therefore, we removed the constraint and assumed that $X_{R|*}$ and $X_{S|*}$ were distributed independently and
 36 identically; that is, instead of the heterozygotes, we captured $n - m_1 - m_0$ individuals of haploid R and
 37 another $n - m_1 - m_0$ individuals of haploid S separately. Regarding homozygotes, we also assumed that we
 38 captured $2m_1$ R haploids and $2m_0$ S haploids instead of m_1 RR and m_0 SS, respectively. Then,

$$39 \quad \begin{aligned} X_{R|homo} &\sim \text{Ga}(2m_1 k, \theta), & X_{R|hetero} &\sim \text{Ga}((n - m_1 - m_0)k, \theta), \\ X_{S|hetero} &\sim \text{Ga}((n - m_1 - m_0)k, \theta) \text{ i. i. d.}, & X_{S|homo} &\sim \text{Ga}(2m_0 k, \theta). \end{aligned}$$

40 *Eq. 17*

41 Finally, we can approximate the DNA amounts of a diploid organism in the bulk sample by simply
 42 substituting Eq. 3 in the main text:

$$43 \quad X_R \sim \text{Ga}((n + m_1 - m_0)k, \theta), \quad X_S \sim \text{Ga}((n - m_1 + m_0)k, \theta).$$

44 *Eq. 18*

45 In addition, at probability $\text{Bin}(0|2n_h, p)$, all (hypothetically haploid) individuals become S or R; in that case,
 46 there is no need to convolve the DNA amounts.