freqpcr: interval estimation of population allele frequency based on quantitative PCR ΔΔCq measures from bulk samples

Allele-frequency estimation based on ΔΔCq

Masaaki Sudo*[1], Masahiro Osakabe[2]

[1] Tea Pest Management Unit, Institute of Fruit Tree and Tea Science, NARO: Kanaya Tea Research Station, 2769, Shishidoi, Kanaya, Shimada, Shizuoka 428-8501, Japan

[2] Laboratory of Ecological Information, Graduate School of Agriculture, Kyoto University: Kyoto, Japan

* To whom correspondence: masaaki@sudori.info

ORCID ID:

0000-0001-9834-9857 (Masaaki Sudo)

Telephone: +81-547-45-4419

Manuscript types: article

Abstract

PCR techniques, both quantitative (qPCR) and non-quantitative, have been used to estimate allele frequency in a population. However, the labor required to sample more individuals and handle each sample makes it difficult to quantify rare mutations, such as pesticide-resistance genes at the early stages of resistance development. Pooling DNA from multiple individuals as a "bulk sample" may reduce handling costs. The output of qPCR on a bulk sample, however, contains uncertainty owing to variations in DNA yields from each individual, in addition to measurement error. In this study, we developed a statistical model for the interval estimation of allele frequency via ΔΔCq-based qPCR analyses of multiple bulk samples taken from a population. We assumed a gamma distribution as the individual DNA yield and developed an R package for parameter estimation, which was verified with real DNA samples from acaricide-resistant spider mites, as well as a numerical simulation. Our model resulted in unbiased point estimates of the allele frequency compared with simple averaging of the ΔΔCq values, and their confidence intervals suggested collecting more samples from individuals and pooling them may produce higher precision than individual PCR tests with moderate sample sizes.

Keywords: Real-time polymerase chain reaction, group testing, confidence interval, maximum likelihood estimation, R language

Introduction

Estimating the frequency of certain alleles in populations is one of the key techniques not only in population genetics and molecular ecology, but also in agricultural and regulatory sciences (Falconer 1960; Kim et al. 2011; Yamamura and Hino 2007). In applied entomology, field monitoring has been performed to detect resistance genes of arthropod pests to pesticides and genetically modified (GM) insecticidal plants, such as *Bt* crops (Andow and Alstad 1998; Sonoda et al. 2017).

Entomologists have traditionally estimated resistance allele frequencies via bioassays (Gould et al. 1997; Li et al. 2016; Tabashnik et al. 2000), in which insects directly collected from fields or their offspring reared in laboratories are exposed to chemical compounds of interest to obtain measurements, such as mortality rate. However, bioassays have drawbacks associated with the treatment of living organisms. It is usually labor-intensive and time-consuming. Although the resistance level can be directly measured using a bioassay as the mortality of tested individuals, additional information including the dominance of the resistance gene is required to estimate the allele frequency.

In accordance with the development of genome-wide association studies on resistance genes (ffrench-Constant 2013; Snoeck et al. 2019; Sugimoto et al. 2020), molecular diagnostics have rapidly developed in recent years (Donnelly et al. 2016; Samayoa et al. 2015; Toda et al. 2017). To quantify the resistance-associated point mutation at the population scale, the most fundamental molecular technique is an individual-based polymerase chain reaction (PCR) analysis (Toda et al. 2017). If the alleles are distributed randomly in the target population, a simple binomial assumption enables us to estimate the population allele frequency and its confidence interval. However, it may not be realistic to extract and analyze DNA individually, especially when dealing with many samples from multiple sites or when we need to estimate mutation frequency, which is rare in the population (below 1%), as is often the case in the early phase of resistance development.

Although rearing living insects is no longer needed, molecular diagnostics still require a silver bullet to reduce the time and cost of handling multiple samples while guaranteeing estimation precision and accuracy. The use of a “bulk sample” (i.e., pooling multiple individual samples and processing a single DNA extract), in coordination with statistical methods, such as group testing, may help. Quantitative PCR (qPCR), based on real-time PCR, is also used for the point estimation of allele frequency (Germer et al. 2000). Osakabe *et al*. (2017) and Maeoka *et al*. (2020) developed diagnostic methods for acaricide resistance in the two-spotted spider mite, *Tetranychus urticae* Koch (Acari: Tetranychidae), where they used a bulk sample to measure the frequency of the resistant point mutation in field mite populations. To calculate the point estimate, these studies compared the relative quantity of the resistance allele with an internal reference (housekeeping gene) in the sample, which is known as the ΔΔCq method (Livak and Schmittgen 2001).

In this study, we propose a statistical procedure to obtain the interval estimate of allele frequency using ΔΔCq-based qPCR analyses over multiple bulk samples taken from a population. We first introduced the random error structure to approximate the amounts of the two alleles (wild-type and mutant) and their ratios in the bulk DNA sample. Thereafter, we formulated how the relative amounts of the two alleles in a sample solution resulted in the Cq measurements through qPCR analysis. Finally, we combined the models for individual DNA yields and the model for ΔΔCq-based qPCR analysis. We developed a maximum likelihood estimation (MLE) procedure to estimate an allele frequency implemented using the R language. The package source is available on the Internet (https://github.com/sudoms/freqpcr).

Model

Approximation of allele quantities contained in a bulk DNA sample

When DNA is directly extracted from the whole body of a living organism, the DNA yield is roughly proportional to its body weight (Chen et al. 2010). For insects, the intra-population frequency distribution of body weight is often approximated using a unimodal and right-skewed continuous distribution, typically lognormal or gamma distribution (May 1976; Rakovski et al. 2011; Knapp 2016). A study suggested that body weights are distributed lognormally in many non-social insect species (Gouws et al. 2011).

In this study, we adopted a gamma distribution, instead of a lognormal, to approximate the DNA amount per individual organism for two reasons. First, it is difficult to distinguish which distribution a real population obeys when the sample size is small. They are considered interchangeable (Wiens 1999; Kundu and Manglick 2005). Second, the sum and proportion of independent gamma distributions have closed forms under certain conditions. Assuming, let X ($X \geq 0$) be the DNA yield per single locus per individual:

$$\mathrm{Ga}(X|k,\theta)=\frac{1}{\Gamma(k)}\left(\frac{1}{\theta}\right)^{k} X^{k-1}\exp\left(-\frac{X}{\theta}\right),$$

Eq. 1

where $\Gamma(\cdot)$ denotes the gamma function. The parameters k and θ ($k,\theta > 0$) are the shape and scale parameters of the gamma distribution, respectively. The mean is given by $k\theta$.

Using Eq. 1, let us consider the amounts of allelic DNA in the sample extracted from multiple individuals at once, hereafter referred to as "a bulk sample." Table 1 lists the variables and parameters of the model structure. For simplicity, we model the case of haploidy in the main text. Appendix A1 describes the approximated formulation for diploids. Now, we have n insects, of which m ($m = 0,1,,\ldots,n$) are the genotypes resistant to an insecticide (hereafter denoted by R). The rest $n-m$ had S, the susceptible allele. When we capture insects from a wild population, the size of n is obvious, but m is usually unknown. Assuming random sampling from an infinite population with the R allele at the frequency p, m follows a binomial distribution:

$$\mathrm{Bin}(m|n,p) = \frac{n!}{m!\,(n-m)!}p^{m}(1-p)^{n-m}.$$

Eq. 2

When the bulk sample contains at least one resistant individual, $X_{\mathrm{R}} = \sum_{i=1}^{m} X_i$ denotes the total R content. If there is no systematic error in the efficiency of DNA extraction between the genotypes, and if X_i, the individual DNA yield obeys the gamma distribution of Eq. 1, then X_{R} follows the gamma distribution with the shape parameter mk and scale parameter θ based on the reproductive property. Conversely, the amount of S allele is denoted by $X_{\mathrm{S}} = \sum_{i=m+1}^{n} X_i$, which follows the gamma distribution with $(n-m)k$ and θ.

$$X_{\mathrm{S}}\sim\mathrm{Ga}\big((n-m)k,\theta\big),$$

$$X_{\mathrm{R}}\sim\mathrm{Ga}(mk,\theta).$$

Eq. 3

When X_R and X_S independently follow gamma distributions with the same scale parameter, the observed allele frequency $Y_R = X_R/(X_S + X_R)$ follows a beta distribution with the shape parameters mk and $(n-m)k$:

$$\mathrm{Beta}(Y_R|mk,(n-m)k) = \frac{{Y_R}^{mk-1}(1-Y_R)^{(n-m)k-1}}{\mathrm{B}(mk,(n-m)k)},$$

Eq. 4

where $\mathrm{B}(\cdot)$ is a beta function. This error structure was originally developed to model allele frequencies measured via quantitative sequencing (Sudo et al. in press). In quantitative sequencing, unlike qPCR, we cannot directly observe the quantities of template DNA (X_R and X_S). Instead, the output from the Sanger sequencer is reflected as Y_R. Although Sudo *et al*. (in press) used Eq. 4 to approximate DNA yield distribution in dead insect bodies on a trap, that is, considering variations in body weight plus post-mortem DNA degradation, it is also applicable to DNA from fresh bodies.

Relative quantification of DNA by real-time PCR: ΔΔCq and RED-ΔΔCq methods

Relationship between the template DNA amount and qPCR measure

In real-time qPCR, the target molecule is amplified at a nearly constant efficiency until it exhausts nucleotides (dNTPs) to synthesize the new DNA strand. After amplification cycles with an appropriate primer set, the abundance of the initial template DNA was measured as Cq: quantification cycle (Bustin et al. 2009), also known as cycle threshold (Ct). According to Livak and Schmittgen (2001), we assume an ideal amplification, where the threshold X_Θ is set within the early exponential amplification phase:

$$X_\Theta = X_0 \times (1+\eta)^\tau.$$

Eq. 5

Here, X_0 and $1+\eta$ ($\eta > 0$) denote the initial amount of template DNA and its amplification efficiency, respectively. Standard PCR protocols are designed so that η obtain the range 80% to 120% i.e., doubling in each cycle. The size of Cq, τ, is then defined as:

$$\tau = \frac{\ln(X_\Theta) - \ln X_0}{\ln(1+\eta)}.$$

Eq. 6

Relative quantification of template DNA between experimental levels: ΔΔCq method

The ΔΔCq (Ct) method (Livak 1997) is the most common method for relative quantification using qPCR. In a typical scenario, an experiment is conducted at two levels (i.e., treated *versus* control [calibrator]) and complementary cDNA libraries are obtained reflecting different gene expression levels at a single target locus (hereafter abbreviated as "TG" or simply T). It is possible to directly compare the cDNA quantities measured by qPCR if a primer set is available to amplify the TG locus. However, there is no guarantee that the samples with different treatments have the same cDNA preparation efficiency.

Hence, an internal reference, which is dispensed in accordance with the sample in question, should be included in relative quantification, such as the ΔΔCq method. The corresponding primer set usually targets the locus of a housekeeping gene (hereafter abbreviated as "HK" or H), a gene that shows a constant expression level and is thus considered the same concentration between treatments. If the experiment had

two levels, we amplified at least four samples (two levels, two primer sets for TG and HK loci, ignoring technical replicates). ΔCq is then defined as the difference of the Cq values of "TG − HK" for each treatment level, which is equivalent to the abundance of target cDNA offset by housekeeping gene (= TG / HK) in each sample (Schefe et al. 2006). Finally, we obtained $\Delta\Delta\mathrm{Cq} = \Delta\mathrm{Cq}^{\mathrm{treated}} - \Delta\mathrm{Cq}^{\mathrm{control}}$ from the Cq measures. Derived from Eq. 6, $2^{-\Delta\Delta\mathrm{Cq}}$ gives the relative abundance of template DNA between the treatment levels (Livak and Schmittgen 2001; Pfaffl 2012) ($1 + \eta = 2$ was presupposed there).

Allele frequency estimation from a single bulk sample: RED-ΔΔCq method

The original ΔΔCq method compares the quantities of (c)DNA between samples to determine the relative expression levels of the genes of interest. Osakabe *et al.* (2017) expanded it and proposed the "RED-ΔΔCq method" (RED stands for restriction enzyme digestion), a derivative method that can measure the allele frequency from a single sample solution, to diagnose the regional resistance prevalence of the two-spotted spider mite, *Tetranychus urticae* Koch (Acari: Tetranychidae), to the acaricide etoxazole, which is conferred by an amino acid substitution in chitin synthase 1 (*CHS1*; I1017F) (Van Leeuwen et al. 2010).

The RED-ΔΔCq method also utilized ΔΔCq as a proxy for relative quantity, but the Cq measurements were all taken from a single bulk sample, which was collected from a population in which each individual possesses R or S. The calibrator was an intact sample containing total DNA ($= X_{\mathrm{R}} + X_{\mathrm{S}}$) on the target locus. The sample in question was the same DNA extract, but digested with restriction endonucleases prior to qPCR analysis. The restriction site is designed to recognize the S allele on the target locus so that the operation digests the major part of S (denoted by $1 - z$: z is a small, but positive variable giving the residual rate). Consequently, we obtained the template amount $X_{\mathrm{R}} + zX_{\mathrm{S}}$ at the target locus after digestion.

The samples before and after digestion were also amplified using the HK primer set as an internal reference. In the etoxazole-R diagnosis by Osakabe *et al.* (2017), glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) was used. Taken together, the single bulk sample results in a quartet of Cq measurements differentiating at the target loci (*CHS1* and *GAPDH*) × restriction enzyme digestion (undigested and digested).

We now formulate the allele frequencies. Let X^{HW} and X^{TW} represent the total amounts of the template DNA at the housekeeping (H) and target (T) loci included in the sample without digestion, the state denoted by W.

$$\begin{aligned} X^{\mathrm{HW}} &= X_{\mathrm{S}} + X_{\mathrm{R}}, \\ X^{\mathrm{TW}} &= \delta_{\mathrm{T}}(X_{\mathrm{S}} + X_{\mathrm{R}}). \end{aligned}$$

Eq. 7

The coefficient δ_{T} ($\delta_{\mathrm{T}} > 0$) provides the relative content of the target gene to the housekeeping gene in genomic DNA (the difference in the DNA extraction efficiencies is also included). After digestion (state D), X^{HD} and X^{TD} denote the DNA amounts at the H and T loci, respectively:

$$\begin{aligned} X^{\mathrm{HD}} &= \delta_{\mathrm{B}}(X_{\mathrm{S}} + X_{\mathrm{R}}), \\ X^{\mathrm{TD}} &= \delta_{\mathrm{B}}\delta_{\mathrm{T}}(zX_{\mathrm{S}} + X_{\mathrm{R}}). \end{aligned}$$

Eq. 8

The common coefficient δ_{B} ($\delta_{\mathrm{B}} > 0$) provides the rate of certain locus-independent changes in the quantities of template DNA accompanying the restriction enzyme treatment.

As a result of qPCR, the Cq quartet, τ^{HW}, τ^{TW}, τ^{HD}, and τ^{TD} were obtained. From Eq. 6,

$$\begin{aligned}\tau^{\mathrm{HW}} &= \frac{\ln(X_\Theta) - \ln(X_\mathrm{S} + X_\mathrm{R})}{\ln(1+\eta)} + \varepsilon_\mathrm{c},\\ \tau^{\mathrm{TW}} &= \frac{\ln(X_\Theta) - \ln\delta_\mathrm{T} - \ln(X_\mathrm{S} + X_\mathrm{R})}{\ln(1+\eta)} + \varepsilon_\mathrm{c},\\ \tau^{\mathrm{HD}} &= \frac{\ln(X_\Theta) - \ln\delta_\mathrm{B} - \ln(X_\mathrm{S} + X_\mathrm{R})}{\ln(1+\eta)} + \varepsilon_\mathrm{c},\\ \tau^{\mathrm{TD}} &= \frac{\ln(X_\Theta) - \ln\delta_\mathrm{B} - \ln\delta_\mathrm{T} - \ln(zX_\mathrm{S} + X_\mathrm{R})}{\ln(1+\eta)} + \varepsilon_\mathrm{c}.\end{aligned}$$

Eq. 9

The actual Cq data contain measurement errors in addition to uncertainty due to experimental operations, such as sample dispensation or PCR amplification. We express these using the common error term $\varepsilon_\mathrm{c} \sim \mathrm{N}(0, \sigma_\mathrm{c}^2)$, following the normal distribution of mean = 0 and variance = σ_c^2 in the scale of raw Cq values. The validity of this error structure is verified later.

The two ΔCq values were then defined as $\Delta\tau^{\mathrm{W}} = \tau^{\mathrm{TW}} - \tau^{\mathrm{HW}}$ and $\Delta\tau^{\mathrm{D}} = \tau^{\mathrm{TD}} - \tau^{\mathrm{HD}}$, respectively. Their ΔΔCq are:

$$\Delta\Delta\tau = \Delta\tau^{\mathrm{D}} - \Delta\tau^{\mathrm{W}} = -\frac{\ln\left(\frac{zX_\mathrm{S} + X_\mathrm{R}}{X_\mathrm{S} + X_\mathrm{R}}\right)}{\ln(1+\eta)} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, 4\sigma_\mathrm{c}^2).$$

Eq. 10

From Eq. 10, the expected value of $(zX_\mathrm{S} + X_\mathrm{R})/(X_\mathrm{S} + X_\mathrm{R})$ is calculated as $(1+\eta)^{-\Delta\Delta\tau}$.

The point estimate of the resistance allele frequency, $\hat{Y}_\mathrm{R}$, is defined as $X_\mathrm{R}/(X_\mathrm{R} + X_\mathrm{S})$ for each bulk sample. When z is much smaller than $\hat{Y}_\mathrm{R}$, the quantity $(zX_\mathrm{S} + X_\mathrm{R})/(X_\mathrm{S} + X_\mathrm{R}) = \hat{Y}_\mathrm{R} + z\left(1 - \hat{Y}_\mathrm{R}\right)$ itself can approximate the frequency, which will be the case with enough digestion time before qPCR. However, the use of the point estimate may introduce a problem in that the size of $\hat{Y}_\mathrm{R}$ often exceeds 1 when the R frequency is high and there is a larger error in the Cq measurement (also see the result of Experiment 2).

Although the value of $1 + \eta$ may vary on the primer sets, both target and housekeeping loci share the same amplification efficiency in Eq. 9. This is because practical PCR protocols were designed to be $1 + \eta \cong 2$. We can also approximately cancel the effect of heterogeneous amplification efficiencies by fitting the size of δ_T the sample sets with known allele ratios (Experiment 1).

Measurement of ΔΔCq using allele-specific primer sets

While the RED-ΔΔCq method enabled us to measure allele frequency from the bulk sample, enzyme availability is a prerequisite to digest the S-allele-specific restriction site at the target locus. A longer digestion period (3 h) was also required to quantify etoxazole resistance in the protocol by Osakabe *et al.* (2017).

Maeoka *et al*. (2020) demonstrated that a general ΔΔCq method without restriction enzyme treatment could be used for allele-frequency measurement if a specific primer set was designed to amplify only the R allele at the target locus. Similar to the RED-ΔΔCq method, DNA samples with unknown mixing ratios were dispensed and amplified using primer sets corresponding to TG and HK loci, respectively. Unlike the RED-ΔΔCq method, the control sample was not taken from the test sample solution, but was prepared as a DNA

solution containing 100% R, hereafter denoted as U (= pUre R line). Then, X^{HU} and X^{TU} denote the template DNA quantities ready for subsequent PCR amplification:

$$\begin{aligned} X^{\mathrm{HU}} &= X'_{\mathrm{R}}, \\ X^{\mathrm{TU}} &= \delta_{\mathrm{T}} X'_{\mathrm{R}}. \end{aligned}$$

Eq. 11

Though the definition of δ_{T} is the same as Eq. 7, the quantity is denoted by X'_{R} instead of $X_{\mathrm{S}} + X_{\mathrm{R}}$ because it no longer originates from the R portion of the test sample itself (i.e., not internal).

For the test sample (denoted as V), the template DNA quantities amplified at the housekeeping (X^{HV}) and target (X^{TV}) loci are expressed as follows:

$$\begin{aligned} X^{\mathrm{HV}} &= X_{\mathrm{S}} + X_{\mathrm{R}}, \\ X^{\mathrm{TV}} &= \delta_{\mathrm{T}}(zX_{\mathrm{S}} + X_{\mathrm{R}}). \end{aligned}$$

Eq. 12

In the PCR process of the modified ΔΔCq method, the small positive number z provides the template quantity of S, which is non-specifically amplified even with the specific primer set, which is designed to amplify only the R allele at the target locus. As the primer set for the housekeeping gene was non-specific, both X^{HU} and X^{HV} were fully amplified. Assuming that all four template DNAs are amplified with efficiency $1 + \eta$, we define the two ΔCq values as $\Delta\tau^{\mathrm{U}} = \tau^{\mathrm{TU}} - \tau^{\mathrm{HU}}$ and $\Delta\tau^{\mathrm{V}} = \tau^{\mathrm{TV}} - \tau^{\mathrm{HV}}$. Finally, their ΔΔCq values are $\Delta\Delta\tau = \Delta\tau^{\mathrm{V}} - \Delta\tau^{\mathrm{U}}$, which yields a formula identical to Eq. 10.

Interval estimation of allele frequency and experimental parameters based on qPCR over multiple bulk samples

Finally, we consider the likelihood model to obtain the interval estimate of the allele frequency based on the (RED-)ΔΔCq analysis over multiple bulk samples. Assume that the population has the R allele at the frequency p from which N bulk samples are taken. The hth sample ($h = 1,2,3,\ldots,N$) consists of n_h haploid individuals, of which m_h are resistant mutants. As shown in Eq. 9, each Cq value is determined not only by the DNA quantities, which are denoted as $X_{h,\mathrm{R}}$ and $X_{h,\mathrm{S}}$ for each sample, but also by parameters such as δ_{T} or σ_{c}^2 accompanying the experimental operation. We can simultaneously estimate these if we have multiple bulk samples, for which the likelihood function of obtaining the Cq values under the parameters is defined.

Although it was possible to define a joint likelihood for each Cq quartet, or we could define the likelihood of a single ΔΔCq value, we propose the joint likelihood for the two ΔCq values, $\Delta\tau_h^{\mathrm{W}} = \tau_h^{\mathrm{TW}} - \tau_h^{\mathrm{HW}}$ and $\Delta\tau_h^{\mathrm{D}} = \tau_h^{\mathrm{TD}} - \tau_h^{\mathrm{HD}}$, for the convenience of numerical calculation:

$$\begin{aligned} \Delta\tau_h^{\mathrm{W}} &\sim \mathrm{N}\left(-\frac{\ln\delta_{\mathrm{T}}}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right), \\ \Delta\tau_h^{\mathrm{D}} &\sim \mathrm{N}\left(-\frac{\ln\delta_{\mathrm{T}} + \ln\left(\frac{zX_{h,\mathrm{S}} + X_{h,\mathrm{R}}}{X_{h,\mathrm{S}} + X_{h,\mathrm{R}}}\right)}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right). \end{aligned}$$

Eq. 13

Although Eq. 13 is defined for the RED-ΔΔCq method, it is also applicable to the ΔΔCq method by Maeoka *et al*. (2020) by substituting $\Delta\tau_h^{\mathrm{W}}$ and $\Delta\tau_h^{\mathrm{D}}$ to $\Delta\tau_h^{\mathrm{U}} = \tau_h^{\mathrm{TU}} - \tau_h^{\mathrm{HU}}$ and $\Delta\tau_h^{\mathrm{V}} = \tau_h^{\mathrm{TV}} - \tau_h^{\mathrm{HV}}$, respectively.

Formulation of likelihood based on gamma or beta distribution

Using the relationship between m_h, n_h, and p in Eq. 2, we proceed to the likelihood function defined as the probability of observing the set of $\Delta\tau_h^{\mathrm{W}}$ and $\Delta\tau_h^{\mathrm{D}}$ under the given values of p, n_h, and other experimental parameters. In Eq. 13, $\Delta\tau_h^{\mathrm{W}}$ is not affected by the R : S ratio in the bulk sample; it is only affected by the experimental parameters, δ_{T}, η, and σ_{c}^2. In addition, by taking the differences, there is no need to estimate as X_Θ and δ_{B} appear in Eq. 9.

Conversely, we must consider the amount of DNA in the bulk sample to calculate the probability of obtaining $\Delta\tau_h^{\mathrm{D}}$. When the size of m_h is specified under the binomial assumption, the quantities of DNA in the hth bulk sample, $X_{h,\mathrm{R}|m_h}$ and $X_{h,\mathrm{S}|m_h}$, can independently take any positive values following the gamma distribution of Eq. 3, and their proportions $Y_{h,\mathrm{R}|m_h} = X_{h,\mathrm{R}|m_h}/\left(X_{h,\mathrm{R}|m_h} + X_{h,\mathrm{S}|m_h}\right)$ are $\mathrm{Beta}(m_h k, (n_h - m_h)k)$ as shown in Eq. 4. If the sample contains only S or R, then $X_{h,\mathrm{R}|m_h=0} = 0$ or $X_{h,\mathrm{S}|m_h=n_h} = 0$ is guaranteed.

The likelihood function for the observed ΔCq values on the hth bulk sample L_h is defined as follows:

$$L_h = P\left(\Delta\tau_h^{\mathrm{W}}|\delta_{\mathrm{T}}, \eta, \sigma_{\mathrm{c}}^2\right) \sum_{m_h=0}^{n_h} \left[\mathrm{Bin}(m_h|n_h, p) P\left(\Delta\tau_h^{\mathrm{D}}|m_h, \delta_{\mathrm{T}}, z, \eta, \sigma_{\mathrm{c}}^2\right)\right],$$

$$P\left(\Delta\tau_h^{\mathrm{D}}|m_h, \delta_{\mathrm{T}}, z, \eta, \sigma_{\mathrm{c}}^2\right) = \begin{cases} \mathrm{N}\left(-\dfrac{\ln(z\delta_{\mathrm{T}})}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right) & (m_h = 0) \\ \psi_{\mathrm{G}} \quad \text{or} \quad \psi_{\mathrm{B}} & (m_h = 1,2,\ldots,n_h - 1). \\ \mathrm{N}\left(-\dfrac{\ln\delta_{\mathrm{T}}}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right) & (m_h = n_h) \end{cases}$$

Eq. 14

We must consider not only the possible cases of m_h, but also the entire range of the DNA amounts. If we use the gamma distributions, for every case $m_h = 1,2,\ldots,n_h - 1$, we need to calculate the double integration for ψ_{G}, the probability of obtaining $\Delta\tau_h^{\mathrm{D}}$ under the whole region of $X_{h,\mathrm{R}|m_h} = r$ and $X_{h,\mathrm{S}|m_h} = s$ for the interval $\{D: 0 \le r < \infty, 0 \le s < \infty\}$.

$$\psi_{\mathrm{G}} = \iint_D \mathrm{N}\left(-\frac{\ln\delta_{\mathrm{T}} + \ln\left(\frac{zs+r}{s+r}\right)}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right) \mathrm{Ga}(r|m_h k, \theta)\mathrm{Ga}(\mathrm{s}|(n_h - m_h)k, \theta) dr ds.$$

Eq. 15

The common scale parameter of the gamma distributions, θ, is not identifiable from the data, although we can substitute arbitrary values $\theta = 1$ for it because it is canceled in $\Delta\tau_h^{\mathrm{D}}$ as a quotient.

Since the computational burden for the double integration is large, we simplified the likelihood model with the beta distribution. As shown in Eq. 4, the proportion $Y_{h,\mathrm{R}} = X_{h,\mathrm{R}}/\left(X_{h,\mathrm{R}} + X_{h,\mathrm{S}}\right)$ is as follows: $\mathrm{Beta}(m_h k, (n_h - m_h)k)$. Then, the probability of obtaining $\Delta\tau_h^{\mathrm{D}}$ is replaced with, ψ_{B} defined as follows:

$$\psi_{\mathrm{B}} = \int_0^1 \mathrm{N}\left(-\frac{\ln\delta_{\mathrm{T}} + \ln\left(z + y(1-z)\right)}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right) \mathrm{Beta}(y|m_h k, (n_h - m_h)k) dy.$$

Eq. 16

We provide an R function "freqpcr()" to estimate the parameters p, k, δ_{T}, and σ_{c} simultaneously when the set of Cq measurements (τ_h^{HW}, τ_h^{TW}, τ_h^{HD}, and τ_h^{TD}) and n_h are given for each of the N bulk samples. The package source is available at https://github.com/sudoms/freqpcr. The default is "beta = TRUE," where the beta distribution model of Eq. 16 was used instead of gamma. Regardless of the algorithms, the asymptotic confidence intervals are calculated using the inverse of the Hessian matrix evaluated at the last iteration. The functions nlm() of R and cubintegrate() in the R package "cubature" (Narasimhan et al. 2019) are used for the iterative optimization and the double integration, respectively.

Identification of auxiliary parameters using DNA samples with known allele-mixing ratios

The likelihood introduced above ensures that we can estimate the sizes of p and k together with other experimental parameters, δ_{T} and σ_{c}, if we have conducted a (RED-)ΔΔCq analysis on multiple bulk samples. However, the size of z is not identified and must be specified as a fixed parameter. The amplification efficiency, η, is estimated in theory over the iterative calculation of Eq. 13, but in fact, simultaneous estimation sometimes fails when η is set as unknown.

Therefore, the experimenter should identify the sizes of these auxiliary parameters. To estimate their plausible sizes, one can conduct (RED-)ΔΔCq analysis using DNA solutions with known allele ratios; for instance, DNA can be extracted from each of the pure breeding lines of S and R and mix the solutions at multiple ratios, or make a dilution series of R by S. As the ratio of X_{R} to X_{S} is strictly fixed, Eq. 9 is directly applicable to express the relationship between DNA quantities and the four Cq measurements. The R functions knownqpcr() and knownqpcr_unpaired() appearing in the package provide the maximum likelihood estimation for δ_{B}, δ_{T}, σ_{c}, z, and η. These values can be used as fixed parameters in the freqpcr () function.

Another objective of the analysis with known-ratio samples is to test the homoscedasticity of the qPCR data at the scale of Cq measures. Regarding the relationship between the R allele frequency and the corresponding $2^{-\Delta\Delta\mathrm{Cq}}$ measures (the approximate point estimate of the frequency), Osakabe *et al.* (2017) demonstrated linearity using a sample series of *T. urticae* DNA with multiple mixing ratios on CHS1 (I1017F). In the next section, we recycled the same data to compare whether the Cq measurements in the RED-ΔΔCq analysis obey the homoscedasticity in the scale of ΔΔCq or $(1+\eta)^{-\Delta\ \mathrm{Cq}}$.

Materials and laboratory methods

Experiment 1: estimation of auxiliary parameters and verification of homoscedasticity in Cq measurements based on mite DNA samples with known allele-mixing ratios

Experimental setup

In the experiment by Osakabe *et al*. (2017), the resistant mite strain (SoOm1-etoR strain) originated from a field population collected in Omaezaki City, Shizuoka, Japan (34.7°N, 138.1°E) in January 2012. The susceptible strain was obtained from Kyoyu Agri Co., Ltd. (Kanagawa, Japan) (Kyoyu-S strain). For each strain, two pairs of females and males were used separately. Each pair was allowed to mate and oviposit on a kidney bean leaf square (2 × 2 cm) for four days. The mites were then confirmed to be homozygous on the

CHS1 locus using sequence analysis. Genomic DNA extracted from the offspring of each pair was used for qPCR analysis. For each pair, the DNA extracts were prepared twice, each of which was a mixture from 50 adult females homogenized together, that is, four extracts (replicates) for each strain.

To verify the validity of the RED-ΔΔCq method, qPCR analysis was performed with heterogeneous DNA solutions with 10 mixing ratios of $X_R/(X_R + X_S)$ = {0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1}. The net DNA concentration of each mixed solution was adjusted to 1 ng μl^{-1}, from which 15 ng was dispensed into each of the two tubes. Only one was digested with the restriction enzymes before qPCR. For digestion, the samples were treated with a mixture of two enzymes, *MluC* I (10 units) and *Taq*$^{\alpha}$I (20 units; New England BioLabs, Ipswich, MA, USA), at 37 °C for 3 h, followed by incubation at 65 °C for 3 h. This is due to the polymorphism of the CHS1 loci; the 1017 codon of *T. urticae* displays ATT (Kyoyu-S strain) or TTT (SoOm1-etoR) sequences, whereas the upstream 1016 codon displays a synonymous TCG or TCA independent of the strains (Van Leeuwen et al. 2012). Therefore, we need to digest both TCGATT (underline shows the restriction site of *Taq*$^{\alpha}$I) and TCAATT (*MluC* I) to diminish the entire S allele.

qPCR analysis using the intercalator method was performed using the LightCycler Nano System (Roche Diagnostics, Basel, Switzerland) with SYBR Fast qPCR Mix (Takara, Kusatsu, Japan) as described previously (Osakabe et al. 2017). The primer sets were tu03*CHS1* (forward: 5'-GGCACTGCTTCATCCACAAG-3' and reverse: 5'-GTGTTCCCCAAGTAACAACGTTC-3') and tu25*GAPDH* (forward: 5'-GCACCAAGTGCTAAAGCATGGAG-3' and reverse: 5'-GAACTGGAACACGGAAAGCCATAC-3').

Statistical analysis

The maximum likelihood estimation of δ_B, δ_T, σ_c, z, and η was conducted with the "knownqpcr_unpaired" function of the freqpcr package. The raw Cq data are available as ESM 1 along with a step-by-step guide for statistical analyses (ESM 2). Due to the limitation of the handling capacity of the thermal cycler, qPCR analysis was not conducted on undigested samples of the nine mixing ratios other than $X_R/(X_R + X_S) = 1$ (i.e., pure R solution). Thus, in each replicate, Osakabe *et al*. (2017) used the observed $\Delta\tau^W$ value when the ratio = 1 for other ratios to calculate the conventional ΔΔCq indices. As we have shown in Eq. 9, this operation does not affect the point estimates of p, although the size of the Cq measurement error (σ_c) will be underestimated if we recycle the observed Cq value multiple times. The "knownqpcr_unpaired" function was developed to deal with such incomplete data (i.e., the observations of τ^{HW}, τ^{TW}, τ^{HD}, and τ^{TD} have different data lengths). If the four Cq measurements are available for all samples, then "knownqpcr" can be used.

Regarding the relationship between the true mixing ratio and the RED-ΔΔCq measures in the sample, the linearity was analyzed using a linear model via the function "lm" running on R version 3.6.1 (R Core Team 2019), where the response variables were put into the model at the scale of Cq or $(1 + \eta)^{-\Delta\Delta Cq}$. Based on the linear models, we tested heteroscedasticity using the Breusch-Pagan test via the bptest() function of the R library "lmtest" (Hothorn et al. 2019).

Experiment 2: evaluation of the simultaneous estimation method with randomly generated data

Since the experiment by Osakabe *et al.* (2017) used a sample series with strict mixing ratios, the effect of individual differences in DNA yield was not evaluated. Instead, we conducted a numerical experiment to verify the accuracy of the simultaneous parameter estimation under uncertainty in the individual DNA yield. The frequency of the R allele in the population, p, was set to 0.01, 0.05, 0.1, 0.25, 0.5, or 0.75.

For the sampling strategy, N bulk samples (the parameter 'ntrap' in the R source code), each comprising of n individuals (n was fixed among the samples: the parameter 'npertrap' in the code), were generated by random sampling from a wild population of a haploid organism. To assess how the estimation interval responds to the sample sizes, we evaluated the combination of N = {2, 4, 8, 16, 32, 64} and n = {4, 8, 16, 32, 64}, though the combinations with $Nn > 128$ were excluded (Nn corresponds to 'ntotal' in the code). The DNA quantities (X_R and X_S) contained in each bulk sample were generated as random numbers that followed the gamma distributions of Eq. 3. To cover a plausible variability range of the DNA yield, the gamma shape parameter was varied as k = {1, 3, 9, 27}. Depending on the size of k, the gamma scale parameter was set at $\theta = 1 \times 10^{-6}/k$ to fix the mean of the individual DNA yield to 1×10^{-6}. The termination threshold for qPCR X_Θ was fixed at 1.

We fixed the other parameters due to limitations of the computing resources. From the results of Experiment 1, δ_T = 1.2, δ_B = 0.24, z = 0.0016, and η = 0.97 were presupposed. As for the random errors in the PCR amplification process and/or the Cq measurement, σ_c = 0.2 was assumed regardless of the initial template quantity. For each of the 624 parameter regions, the dummy datasets comprising N bulk samples were generated 1,000 times independently with different random number seeds (i.e., 1,000 replicates), for which the parameter estimation with freqpcr(..., beta = TRUE) was run on the R 3.6.1 environment. The simulation code is available in ESM 3.

As we also implemented the gamma distribution model as freqpcr(..., beta = FALSE), a numerical experiment with the gamma model was also conducted for the first 250 replicates, and the estimation accuracy was compared between the two assumptions. Furthermore, we also fitted the function with the settings freqpcr(..., K = 1), that is, assuming the gamma shape parameter was fixed at 1 (a.k.a. exponential distribution), in addition to the default simulation with all parameters unknown. Further, the easiest way to estimate p derived from Eq. 10, we averaged the observed ΔΔCq values for N bulk samples and transformed them as $\hat{p} = (1 + \eta)\text{^}(-\overline{\Delta\Delta\tau})$.

Results

Estimation of auxiliary parameters and verification of homoscedasticity

Based on the Cq measures, the auxiliary parameters were estimated based on the RED-ΔΔCq analysis of the I1017F mutation of *T. urticae*. As for the initial quantity of template DNA (the parameter "meanDNA" on the R code; defined as X/X_Θ), the maximum-likelihood estimate was 1.256×10^{-6} (95% confidence interval: 7.722×10^{-7} to 2.041×10^{-6}). The relative quantity of the target gene to the housekeeping gene δ_T (targetScale) was estimated to be 1.170 (95% CI: 1.069–1.280). The locus-independent change rate in the template quantity accompanying the restriction enzyme treatment δ_B (baseChange) was 0.2361 (95% CI: 0.2040 to 0.2731). The measurement error in the scale of Cq σ_c (SD) was 0.2376 (95% CI: 0.2050 to 0.2755). The residue rate of the S allele after digestion z (zeroAmount) was 0.001564 (95% CI: 0.001197–0.002044). The efficiency of amplification per PCR cycle η (EPCR) was 0.9712 (95% CI: 0.9231–1.022).

In the RED-ΔΔCq analysis of the etoxazole resistance of *T. urticae*, the relationship between the true R allele frequency ($Y_R = X_R/(X_R + X_S)$ in the sample) and the corresponding Cq measures exhibited higher homoscedasticity in the scale of the measured ΔΔCq values rather than in $(1 + \eta)^{-\Delta\Delta Cq}$, the transformation to $\hat{Y}_R$ (Fig. 1). The linear regression of the ΔΔCq values on $-\ln[0.001564 \times (1 - Y_R) + Y_R]/\ln(1 + 0.971)$ showed high linearity (intercept = −0.07694, coefficient = 1.025, adjusted R^2 = 0.9936). The homoscedasticity of the coefficient of determination was not rejected at the 5% level of significance

(Breusch-Pagan test: BP = 3.1577, *df* = 1, $p = 0.07557$) (Fig. 1A). Conversely, the linear regression of $1.971^{-\Delta\Delta Cq}$ on $[0.001564 \times (1 - Y_R) + Y_R]$ showed a slightly lower linearity (intercept = −0.008625, coefficient = 1.092, adjusted $R^2 = 0.9709$). The Breusch-Pagan test was highly significant (BP = 13.978, *df* = 1, $p = 0.0001849$), rejecting homoscedasticity (Fig. 1B). These results suggest that it is easier to model the error structure of the RED-ΔΔCq method on the scale of Cq values (logarithm) rather than frequency (linear scale).

Evaluation of the simultaneous estimation method with randomly generated data

Among the 624 parameter regions of the numerical simulation with 1,000 replicates (250 for the gamma model), the total success rate of the interval estimation p using freqpcr(..., beta = TRUE) was 70.6% and 94.5% when all parameters were unknown, and when the gamma shape parameter was fixed as $k = 1$, respectively. The "success rate" here indicates the probability when the function returns certain values other than NA (i.e., the diagonal of the Hessian was not negative): no guarantee that the estimated confidence interval was accurate. The estimation success for the Cq measurement error, σ_c, was 69.6% and 97.6% in the beta-distribution model with unknown k and $k = 1$, respectively. The relative quantity of the target gene, δ_T, was 68.1% and 96.1%, respectively. However, the estimation success of k was 59.9% with the beta distribution model, showing a lower performance than the other parameters. This result implies that the likelihood is insensitive to the size of k. Conversely, the estimation of p is robust to the size of k, as we show later in this section.

The estimation success of freqpcr() largely depended on the total sample size (Nn corresponding to the facet 'ntotal' in the figures), as well as the level of p (Figure S1 and S2 for the beta and gamma models, with all parameters unknown). In each parameter region, the quantity $\mathrm{Bin}(0|Nn, p)$ generally gives the probability that the whole sample contains no R individuals. When Nn is larger enough, $Nn > 3/p$ is approximately the requirement for the total sample size to contain at least one R individual with 95% confidence, called the "rule of three" (Eypasch et al. 1995). The gray backgrounds in the facets of Figures 2–4 and S1–S6 signify the regions where the total sample sizes are smaller than the thresholds (e.g., 60 individuals are required when $p = 0.05$). As shown in Figure S1, the parameter estimation often failed when Nn did not meet the rule of three. Once we exclude the parameter regions of $Nn \leq 3/p$, the estimation success rate of p with freqpcr(..., beta = TRUE) improved to 84.3% and 99.9% with all parameters unknown and assuming $k = 1$, respectively.

As for the estimation accuracy of p, the freqpcr() function assuming beta distribution provides an unbiased estimator. Figures 2 and 3 show the estimated sizes of p using the beta model with all parameters unknown and assuming $k = 1$, respectively. Both settings demonstrated that the estimator converged to the true R frequency; the upper/lower bounds of the estimated 95% confidence intervals (yellow/blue boxes in each plot) became narrower as we increased the total sample sizes (Nn) or included more bulk DNA samples (N). Fixing the size of the gamma shape parameter to $k = 1$ scarcely affected the point estimates and intervals of p, as long as $Nn > 3/p$ is satisfied (Figure 3). However, if every individual was analyzed separately, the interval estimation was only possible when k was fixed (see the regions of "sample division = ntotal" cases in Figure 2).

When we used the gamma distribution model, the interval estimation of p was also possible and unbiased (Figure S3). However, when we defined the point estimator of p as a simple average, that is, $\hat{p} = (1 + \eta)\text{^}(-\overline{\Delta\Delta\tau})$, it was strongly underestimated as the samples were more divided (N/Nn was large) (Figure 4). The upper bound of 95% CI often violated 1, suggesting that the "simple average of ΔΔCq" ± 1.96 SE is inadequate for the interval estimation based on the RED-ΔΔCq method.

Although the freqpcr() function with the gamma and beta distributions both showed an unbiased estimation of p, the gamma model was disadvantageous regarding calculation time and the number of iterations before convergence. The time varied largely in the model settings and sample sizes (Figures S4–S6). Amongst the settings we tried, beta model with fixed k was the fastest; it converged within a few seconds in most parameter regions (median and 75 percentile: 0.32 and 0.69 s: Figure S6). It was three and >10 times faster than the beta (0.91 and 2.4 seconds: Figure S4) and gamma (3.0 and 15 s: Figure S5) model, respectively with all parameters unknown. The calculation time increased as the dataset size increased - Nn and the sample was more divided (larger N/Nn) in the beta distribution model, because the marginal likelihood was calculated for each bulk sample. Conversely, the gamma distribution model (Figure S5) requires increased calculation time as the size of each bulk sample becomes larger (larger n_h). This was considered because the combination of $\mathrm{Bin}(m_h|n_h,p)$ exploded when n_h was large.

Regarding the estimation accuracy of the shape parameter, k, it was underestimated as the real size of the parameter increased (e.g., $k = 27$) when we used the gamma distribution model (Figure S7B). Since the iterative fitting of the parameter in freqpcr() always starts internally from $k = 1$ (this was determined due to the calculation stability), this bias suggests the likelihood function of ψ_{G} (Eq. 15), with little information on the size of k compared with p. Then, k tends to stay at its initial value, suggesting that the gamma model is not suitable for the simultaneous estimation of p and k. Unlike the gamma version, the fitting of k with freqpcr(beta = TRUE) was satisfactory when we divided the total samples into more bulk samples (larger N/Nn), although the initial value dependence was still observed, especially when p was small (Figure S7A). This may be because the estimation of k via $\mathrm{Beta}(m_h k, (n_h - m_h)k)$ in Eq. 16 is comparable with measuring the overdispersion of $Y_{h,\mathrm{R}|m_h}$, which is only possible when multiple bulk samples contain both R and S alleles.

Discussion

In the present study, we developed a statistical model to estimate the population allele frequency based on qPCR across multiple bulk samples. There have been problems with the conventional point estimator of the allele frequency by averaging the observed ΔΔCq values $\hat{p} = (1 + \eta)\hat{}(-\overline{\Delta\Delta\tau})$. It sometimes exceeds 1 when the frequency of the target allele is close to 1. Furthermore, when one tries to quantify the mutant allele rare in the population, most bulk samples contains only the wild type. The conventional $\hat{p}$ is vulnerable to many zero samples, which makes the frequency estimation more difficult when p is small. To circumvent these problems, our interval estimation explicitly models the number of individuals contained in each bulk sample (the binomial assumption) as well as the individual DNA yields (the gamma assumption), thereby obtaining the interval estimate over the entire range $0 < p < 1$.

The explicit modeling of individuals also allows sample division to various degrees, which helps us to balance our sampling strategy on the cost-precision tradeoff. We can achieve higher precision (narrower confidence interval) by increasing the total sample size, $\sum_{h=1}^{N} n_h$ although it also increases the costs associated with sample collection and laboratory work, including library preparation and PCR analysis. Recent advances in molecular diagnosis have relieved sampling costs. We can now extract DNA from dead insect bodies obtained from sticky traps (Uesugi et al. 2016). Nevertheless, a larger sample size still imposes a larger handling cost if we analyze the collected individuals individually via non-quantitative PCR.

The combination of mass trapping and bulk qPCR analysis solves the latter by collecting more individuals and pooling them. This can result in higher precision with less work than individual PCR. For instance, we sampled 16 individuals from the population with an allele frequency of $p = 0.05$ and analyzed two

individuals at once in the numerical experiment (Figure 2: facet of ntotal = 16, sample division = 8). The lower and upper bounds of the 95% confidence interval p were estimated to be 0.0087 and 0.34, respectively, using freqpcr(..., beta = TRUE) (as the medians of the 1,000 independent trials). We also simulated the case of ntotal = 64 and sample division = 4 (i.e., analyzed 16 individuals together). The upper and lower bounds were 0.015 and 0.15, respectively. Thus, we improved the precision of the interval estimate with half the handling effort.

Also in non-quantitative PCR, sample pooling has been considered as a tool to detect (c)DNA rare in the population with practical labor, sometimes as a high throughput pre-screening of a number of samples e.g. in clinical examination (Taylor et al. 2010; Yelin et al. 2020). In some fields, such as plant quarantine, it is important to guarantee that a product is not contaminated with pests or unapproved GM seeds at a certain consumer risk. As the assumed frequency range is extremely low ($p \approx 0.001$), frequency estimation is not realistic (3,000 seeds are needed to meet the "rule of three" when $p = 0.001$) and is not required for the current inspection routine. Thus, group testing based on non-quantitative PCR has been conducted in these fields (Yamamura et al. 2019). Yamamura and Hino (2007) proposed a procedure to estimate the upper limit of the population allele frequency, in which they used the proportion of bulk samples detected as "positive."

Overall, there has been a gap in methodology between the frequency estimation based on the individual PCR and the non- or semi-quantitative PCR based on the non-quantitative bulk PCR. Although it provides the highest estimation precision following binomial distribution, the former is only available at a higher p; it becomes labor-intensive once we try to quantify rare alleles. The latter can be applied to a lower range of p, but the precision is generally low or even non-quantitative. Bridging the gap, our qPCR-based procedure offers an allele-frequency estimation in the mid-low range (p = 0.01 to 0.25), which is considered a critical range for decision making in some fields like pesticide resistance management (Takahashi et al. 2017; Sudo et al. 2018).

Although this study focused on resistance genes, the likelihood model in Eq. 13 can be used for other qPCR protocols based on this ΔΔCq method. If both the specific and nonspecific primer sets are available to amplify the mutant and "wild type + mutant" alleles at the target locus, they can be used for the test and control samples equivalent to X^{TV} in Eq. 12 and X^{TU} in Eq. 11, respectively. However, there is a caveat in determining which allele should be amplified with a specific primer set and which affects the estimation accuracy due to the intrinsic nature of $(1+\eta)^{-\Delta\Delta\tau}$. As shown, the 95% confidence intervals were broader when p = 0.75 than when p = 0.25 (Figure 2), the accuracy was not symmetric around 0.5, but more accurate when the frequency was low. That is, one should design a specific primer set to amplify the allele that would be rare in the population to improve the signal-to-noise ratio.

The maximum likelihood estimation with freqpcr() relies on the assumption that the quantities of the S and R alleles in each bulk sample independently follow gamma distribution and that their quotient is expressed using beta distribution. Fixing the size of the gamma shape parameter k further accelerated the optimization, which was owing to the robustness of p to the size of k. However, once the size of k was fixed much larger than the actual size of the gamma shape parameter (i.e., the individual DNA yield was regarded as almost a fixed value), the iterative optimization using the nlm() function sometimes returned an error. Therefore, one should start with a smaller shape parameter e.g., k = 1 (the exponential distribution: Figure 3), which is currently the default setting of the freqpcr package.

In qPCR applications for diagnostic use, ΔΔCq is often used with calibration. One of the popular methods is the involvement of technical replicates; each sample is dispensed and analyzed using qPCR multiple times, which cancels the Cq measurement error. The measurement error obeys a homoscedastic normal distribution in the Cq scale, as shown in Experiment 1. Thus, a simple solution is to average the Cq values measured for

every bulk sample before the estimation with freqpcr(), although the estimated size of σ_c changes from its original definition in Eq. 9. However, it is trivial if the number of technical replicates is unified between bulk samples.

Moreover, the comparison of Cq values is sometimes conducted on more than one internal reference because there is no guarantee that the expression level of a "housekeeping gene" is always constant (Vandesompele et al. 2002). Future updates of freqpcr() will handle multiple internal references. As long as qPCR is used to estimate population allele frequency, the use of statistical inferences on the bulk samples, as presented in this study, will continue to be a realistic option for regional allele monitoring and screening for practitioners, such as those in agricultural, food security, and public health sectors.

Acknowledgements

We appreciate Dr. Kohji Yamamura and Dr. Takehiko Yamanaka for earlier discussion on the gamma assumption of the individual DNA yield. The work was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics-based Technology for Agricultural Improvement): PRM05 to MO and PRM07 to MS.

Conflict of Interest

None declared.

Data Accessibility

The R package source is available at https://github.com/sudoms/freqpcr. The output data of the numerical experiment are available at https://figshare.com/collections/freqpcr/5258027. The source code for the figures, including the mite dataset from Osakabe *et al*. (2017), are available as electronic supplementary materials.

ESM 1

Verification of the RED-ΔΔCq method: raw dataset used by Osakabe et al. (2017).

ESM 2

R source code for Experiment 1 (Figure 1), including a brief guide to the "freqpcr" package.

ESM 3

R source code for the numerical simulation of freqpcr() (Experiment 2), and the codes for Figures 2 and after.

References

Andow, D. A., and D. N. Alstad. 1998. F2 screen for rare resistance alleles. Journal of Economic Entomology 91:572–578.

560 Bustin, S. A., V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, et al. 2009. The
561 MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments.
562 Clinical chemistry 55:611–622.

563 Chen, H., M. Rangasamy, S. Y. Tan, H. Wang, and B. D. Siegfried. 2010. Evaluation of five methods for
564 total DNA extraction from western corn rootworm beetles. PLoS one 5:e11963.

565 Donnelly, M. J., A. T. Isaacs, and D. Weetman. 2016. Identification, validation, and application of molecular
566 diagnostics for insecticide resistance in malaria vectors. Trends in parasitology 32:197–206.

567 Eypasch, E., R. Lefering, C. K. Kum, and H. Troidl. 1995. Probability of adverse events that have not yet
568 occurred: a statistical reminder. Bmj 311:619–620.

569 Falconer, D. S. 1960. Introduction to quantitative genetics. Oliver And Boyd; Edinburgh; London.

570 ffrench-Constant, R. H. 2013. The molecular genetics of insecticide resistance. Genetics 194:807–815.

571 Germer, S., M. J. Holland, and R. Higuchi. 2000. High-throughput SNP allele-frequency determination in
572 pooled DNA samples by kinetic PCR. Genome research 10:258–266.

573 Gould, F., A. Anderson, A. Jones, D. Sumerford, D. G. Heckel, J. Lopez, S. Micinski, et al. 1997. Initial
574 frequency of alleles for resistance to Bacillus thuringiensis toxins in field populations of Heliothis
575 virescens. Proceedings of the National Academy of Sciences 94:3519–3523.

576 Gouws, E. J., K. J. Gaston, and S. L. Chown. 2011. Intraspecific body size frequency distributions of insects.
577 PLoS One 6:e16606.

578 Hothorn, T., A. Zeileis, R. W. Farebrother, C. Cummins, G. Millo, and D. Mitchell. 2019. lmtest: Testing
579 Linear Regression Models. https://cran.r-project.org/package=lmtest

580 Kim, S. Y., K. E. Lohmueller, A. Albrechtsen, Y. Li, T. Korneliussen, G. Tian, N. Grarup, et al. 2011.
581 Estimation of allele frequency and association mapping using next-generation sequencing data. BMC
582 bioinformatics 12:231.

583 Knapp, M. 2016. Relative importance of sex, pre-starvation body mass and structural body size in the
584 determination of exceptional starvation resistance of Anchomenus dorsalis (Coleoptera: Carabidae). PloS
585 one 11:e0151459.

586 Kundu, D., and A. Manglick. 2005. Discriminating between the log-normal and gamma distributions. Journal
587 of the Applied Statistical Sciences 14:175–187.

588 Li, G., D. Reisig, J. Miao, F. Gould, F. Huang, and H. Feng. 2016. Frequency of Cry1F non-recessive
589 resistance alleles in North Carolina field populations of Spodoptera frugiperda (Lepidoptera: Noctuidae).
590 PloS one 11:e0154492.

591 Livak, K. J. 1997. Comparative Ct method. Pages 11–15 *in*ABI Prism 7700 sequence detection system User
592 Bulletin #2, P/N 4303859 (Vol. 2). Applied Biosystems.

593 Livak, K. J., and T. D. Schmittgen. 2001. Analysis of relative gene expression data using real-time
594 quantitative PCR and the 2- ΔΔCT method. methods 25:402–408.

595 Maeoka, A., L. Yuan, Y. Itoh, C. Saito, M. Doi, T. Imamura, T. Yamaguchi, et al. 2020. Diagnostic
596 prediction of acaricide resistance gene frequency using quantitative real-time PCR with resistance allele-
597 specific primers in the two-spotted spider mite Tetranychus urticae population (Acari: Tetranychidae).
598 Applied Entomology and Zoology 55:329–335.

May, R. M. 1976. Models for single populations. Pages 5–29 *in*Theoretical ecology. Blackwell, Oxford.

Narasimhan, B., M. Koller, S. G. Johnson, T. Hahn, A. Bouvier, K. Kiêu, and S. Gaure. 2019. cubature: Adaptive Multivariate Integration over Hypercubes. https://cran.r-project.org/package=cubature

Osakabe, M., T. Imamura, R. Nakano, S. Kamikawa, M. Tadatsu, Y. Kunimoto, and M. Doi. 2017. Combination of restriction endonuclease digestion with the ΔΔCt method in real-time PCR to monitor etoxazole resistance allele frequency in the two-spotted spider mite. Pesticide biochemistry and physiology 139:1–8.

Pfaffl, M. W. 2012. Quantification strategies in real-time polymerase chain reaction. Quantitative real-time PCR. Appl Microbiol 53–62.

R Core Team. 2019. R version 3.6.1.

Rakovski, C., D. J. Weisenberger, P. Marjoram, P. W. Laird, and K. D. Siegmund. 2011. Modeling measurement error in tumor characterization studies. BMC bioinformatics 12:284.

Samayoa, L. F., R. A. Malvar, B. A. Olukolu, J. B. Holland, and A. Butrón. 2015. Genome-wide association study reveals a set of genes associated with resistance to the Mediterranean corn borer (Sesamia nonagrioides L.) in a maize diversity panel. BMC plant biology 15:35.

Schefe, J. H., K. E. Lehmann, I. R. Buschmann, T. Unger, and H. Funke-Kaiser. 2006. Quantitative real-time RT-PCR data analysis: current concepts and the novel "gene expression's C T difference" formula. Journal of molecular medicine 84:901–910.

Snoeck, S., A. H. Kurlovs, S. Bajda, R. Feyereisen, R. Greenhalgh, E. Villacis-Perez, O. Kosterlitz, et al. 2019. High-resolution QTL mapping in Tetranychus urticae reveals acaricide-specific responses and common target-site resistance after selection by different METI-I acaricides. Insect biochemistry and molecular biology 110:19–33.

Sonoda, S., K. Inukai, S. Kitabayashi, S. Kuwazaki, and A. Jouraku. 2017. Molecular evaluation of diamide resistance in diamondback moth (Lepidoptera: Yponomeutidae) populations using quantitative sequencing. Applied entomology and zoology 52:353–357.

Sudo, M., D. Takahashi, D. A. Andow, Y. Suzuki, and T. Yamanaka. 2018. Optimal management strategy of insecticide resistance under various insect life histories: Heterogeneous timing of selection and interpatch dispersal. Evolutionary applications 11:271–283.

Sudo, M., K. Yamamura, S. Sonoda, and T. Yamanaka. in press. Estimating the proportion of resistance alleles from bulk Sanger sequencing, circumventing the variability of individual DNA. Journal of Pesticide Science.

Sugimoto, N., A. Takahashi, R. Ihara, Y. Itoh, A. Jouraku, T. Van Leeuwen, and M. Osakabe. 2020. QTL mapping using microsatellite linkage reveals target-site mutations associated with high levels of resistance against three mitochondrial complex II inhibitors in Tetranychus urticae. Insect Biochemistry and Molecular Biology 103410.

Tabashnik, B. E., A. L. Patin, T. J. Dennehy, Y.-B. Liu, Y. Carriere, M. A. Sims, and L. Antilla. 2000. Frequency of resistance to Bacillus thuringiensis in field populations of pink bollworm. Proceedings of the National Academy of Sciences 97:12980–12984.

Takahashi, D., T. Yamanaka, M. Sudo, and D. A. Andow. 2017. Is a larger refuge always better? Dispersal and dose in pesticide resistance evolution. Evolution 71:1494–1503.

639 Taylor, S. M., J. J. Juliano, P. A. Trottman, J. B. Griffin, S. H. Landis, P. Kitsa, A. K. Tshefu, et al. 2010.
640 High-throughput pooling and real-time PCR-based strategy for malaria detection. Journal of clinical
641 microbiology 48:512–519.

642 Toda, S., K. Hirata, A. Yamamoto, and A. Matsuura. 2017. Molecular diagnostics of the R81T mutation on
643 the D-loop region of the β1 subunit of the nicotinic acetylcholine receptor gene conferring resistance to
644 neonicotinoids in the cotton aphid, Aphis gossypii (Hemiptera: Aphididae). Applied entomology and
645 zoology 52:147–151.

646 Uesugi, R., N. Hinomoto, and C. Goto. 2016. Estimated time frame for successful PCR analysis of
647 diamondback moths, Plutella xylostella (Lepidoptera: Plutellidae), collected from sticky traps in field
648 conditions. Applied entomology and zoology 51:505–510.

649 Van Leeuwen, T., J. Vontas, A. Tsagkarakou, W. Dermauw, and L. Tirry. 2010. Acaricide resistance
650 mechanisms in the two-spotted spider mite Tetranychus urticae and other important Acari: a review. Insect
651 biochemistry and molecular biology 40:563–572.

652 Van Leeuwen, T., P. Demaeght, E. J. Osborne, W. Dermauw, S. Gohlke, R. Nauen, M. Grbić, et al. 2012.
653 Population bulk segregant mapping uncovers resistance mutations and the mode of action of a chitin
654 synthesis inhibitor in arthropods. Proceedings of the National Academy of Sciences 109:4407–4412.

655 Vandesompele, J., K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, and F. Speleman. 2002.
656 Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal
657 control genes. Genome biology 3:research0034–1.

658 Wiens, B. L. 1999. When log-normal and gamma models give different results: a case study. The American
659 Statistician 53:89–93.

660 Yamamura, K., and A. Hino. 2007. Estimation of the proportion of defective units by using group testing
661 under the existence of a threshold of detection. Communications in Statistics—Simulation and
662 Computation® 36:949–957.

663 Yamamura, K., J. Mano, and H. Shibaike. 2019. Optimal definition of the limit of detection (LOD) in
664 detecting genetically modified grains from heterogeneous grain lots. Quality Technology & Quantitative
665 Management 16:36–53.

666 Yelin, I., N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, et al. 2020.
667 Evaluation of COVID-19 RT-qPCR test in multi-sample pools. Clinical Infectious Diseases 71: 2073–
668 2078. doi:10.1093/cid/ciaa531
669

Appendix

Appendix A1: 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 ($m_1 = 0,1,,\ldots,n$) individuals of RR homozygotes, $n - m_1 - m_0$ RS heterozygotes, and m_0 ($m_0 = 0,1,,\ldots,n$) 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 $(1-p)^2$

$$\mathrm{Tri}(m_1, m_0 | n, p^2, (1-p)^2) = \frac{n!}{m_1!\, m_0!\, (n - m_1 - m_0)!} \cdot p^{2m_1} \cdot (1-p)^{2m_0} \cdot (2p - 2p^2)^{(n-m_1-m_0)}. \quad \textit{Eq. 17}$$

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. However, 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 $X_{*\in(\mathrm{S,R})|\mathrm{homo}}$ for the yield of S or R from the homozygotes, and $X_{*\in(\mathrm{S,R})|\mathrm{hetero}}$ 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 $X_{*|\mathrm{homo}}$ and $X_{*|\mathrm{hetero}}$:

$$X_\mathrm{R} = 2 \times X_{\mathrm{R}|\mathrm{homo}} + X_{\mathrm{R}|\mathrm{hetero}}, \quad X_\mathrm{S} = X_{\mathrm{S}|\mathrm{hetero}} + 2 \times X_{\mathrm{S}|\mathrm{homo}},$$

$$2 \times X_{\mathrm{R}|\mathrm{homo}} \sim \mathrm{Ga}(m_1 k, 2\theta), \quad X_{\mathrm{R}|\mathrm{hetero}} \sim \mathrm{Ga}\big((n - m_1 - m_0)k, \theta\big),$$
$$X_{\mathrm{S}|\mathrm{hetero}} = X_{\mathrm{R}|\mathrm{hetero}}, \qquad 2 \times X_{\mathrm{S}|\mathrm{homo}} \sim \mathrm{Ga}(m_0 k, 2\theta). \quad \textit{Eq. 18}$$

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. 18 gives the probability of obtaining the ΔCq measures in Eq. 13. However, a drawback arises from the constraint of the amounts of R and S possessed by heterozygotes. The applicability of the likelihood model (Eq. 15 or Eq. 16) depends largely on the independence of X_R and X_S. If we define the likelihood using Eq. 18 as it was, we must convolve the DNA amounts not on the two-dimensional parameter space spanned by X_R and X_S, but a three-dimensional space by $X_{\mathrm{R}|\mathrm{homo}}$, $X_{\mathrm{S}|\mathrm{hetero}} = X_{\mathrm{R}|\mathrm{hetero}}$, and $X_{\mathrm{S}|\mathrm{homo}}$, which would increase the calculation time by 1,000 to 10,000 times.

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

$$X_{\mathrm{R|homo}} \sim \mathrm{Ga}(2m_1k, \theta), \quad X_{\mathrm{R|hetero}} \sim \mathrm{Ga}\big((n - m_1 - m_0)k, \theta\big),$$
$$X_{\mathrm{S|hetero}} \sim \mathrm{Ga}\big((n - m_1 - m_0)k, \theta\big)\ i.i.d., \quad X_{\mathrm{S|homo}} \sim \mathrm{Ga}(2m_0k, \theta).$$

Eq. 19

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

$$X_{\mathrm{R}} \sim \mathrm{Ga}\big((n + m_1 - m_0)k, \theta\big), \quad X_{\mathrm{S}} \sim \mathrm{Ga}\big((n - m_1 + m_0)k, \theta\big).$$

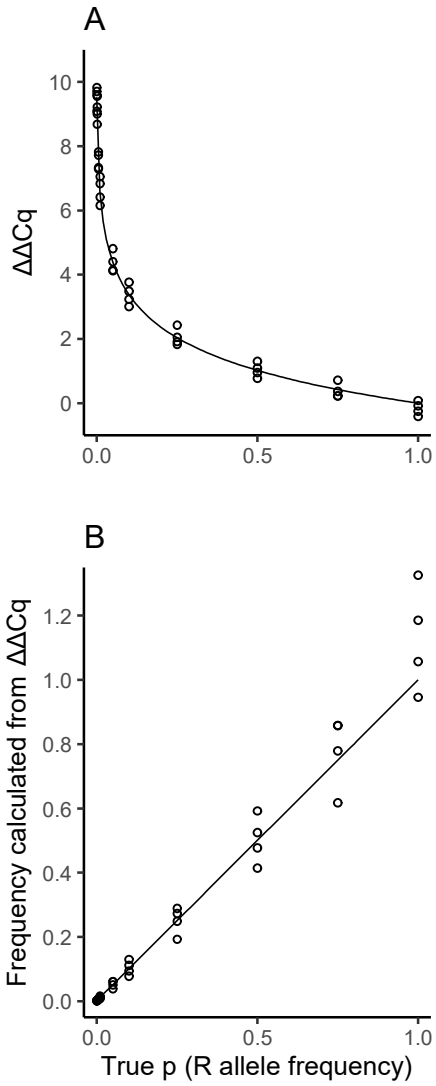
Eq. 20

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

Table and figure captions

Table 1. Description of variables and parameters

Symbol	Description	Range	Arguments in the R package
p	Frequency of the R (resistant) allele in a population	$0 \leq p \leq 1$	P
X_{S}, X_{R}	Amounts of DNA belonging to S (susceptible) or R alleles included in a bulk sample	$X_{\mathrm{S}} \geq 0, X_{\mathrm{R}} \geq 0$	—
Y_{R}	The observed frequency of R in the bulk sample, defined as $X_{\mathrm{R}}/(X_{\mathrm{R}} + X_{\mathrm{S}})$	$0 \leq Y_{\mathrm{R}} \leq 1$	—
k, θ	Shape and scale parameters of the gamma distribution $\mathrm{Ga}(k, \theta)$	$k > 0, \theta > 0$	K
N	Number of bulk samples taken from a population, each of which consists of n_h individuals ($h = 1,2,3,\ldots,N$)	$N \in \mathbb{N}$	ntrap
n, n_h	Number of individuals constituting the (hth) bulk sample	$n \in \mathbb{N}$	npertrap
m, m_h	Number of R individuals included in the (hth) bulk sample	$0 \leq m \in \mathbb{Z} \leq n$	m (as an internal variable)
qPCR-related variables and parameters			
η	Per-cycle efficiency in the PCR amplification (as $1 + \eta$)	$\eta > 0$	EPCR
X_0, X_Θ	Initial amount of template DNA and the termination threshold of the amplification in the real-time PCR process	$X_0 > 0, X_\Theta > 0$	X_Θ is fixed 1in the package
τ	Cq value: the number of PCR amplification cycles before termination	$\tau \in \mathbb{R}$	τ_h^{TW}: target0, τ_h^{TD}: target1, τ_h^{HW}: housek0, τ_h^{HD}: housek1
δ_{T}	Relative content of the target gene to the internal reference (housekeeping gene)	$\delta_{\mathrm{T}} > 0$	targetScale
δ_{B}	(In RED-ΔΔCq method) the locus-independent change rate of the template DNA quantity accompanying the restriction enzyme treatment.	$\delta_{\mathrm{T}} > 0$	baseChange
z	(In RED-ΔΔCq method) residual rate of restriction enzyme digestion, or (in general ΔΔCq analyses) portion of the off-target allele amplified in the PCR	$z > 0$	zeroAmount
ε_{c}	Cq measurement error (standard deviation)	$\varepsilon_{\mathrm{c}} > 0$	sdMeasure



716

717 Figure 1 The relationship between the allele frequency in the sample and A: the RED-ΔΔCq measures, B: the
718 observed frequency calculated as $(1+\eta)^{-\Delta\ Cq}$, showing the results of etoxazole resistance in the two-
719 spotted spider mites. The lines are not the regression on the actual Cq measurement (shown as points), but
720 the theoretical relationship between true frequency of the R allele and the quantity defined as A:
721 $-\ln\big(z+Y_{\mathrm{R}}(1-z)\big)/\ln(1+\eta)$ or B: $z+Y_{\mathrm{R}}(1-z)$, where $Y_{\mathrm{R}}=X_{\mathrm{R}}/(X_{\mathrm{R}}+X_{\mathrm{S}})$. Parameters are $z=$
722 0.00156 and $\eta=0.971$.

723

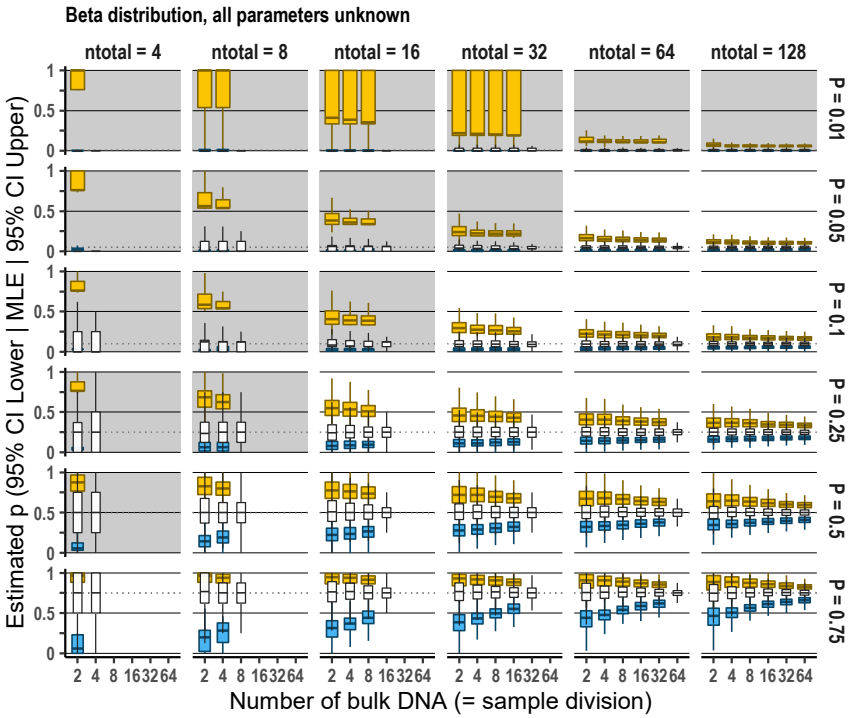


Figure 2 Estimation accuracy of the resistance allele frequency, *p*, with freqpcr() when the beta distribution was assumed, and all estimable parameters (P, K, targetScale, and sdMeasure) were set as unknown. The result of numerical experiments based on 1,000 dummy datasets per parameter region. The x-axes correspond to the parameter "ntrap." The three box plots (white thin, blue, and yellow wide) in each region show the maximum likelihood estimates (MLE), lower bound of the 95% CI, and the upper bound, respectively. In each boxplot, the horizontal line signifies the median of the simulations, hinges of the box show 25 and 75 percentiles, and the upper/lower whiskers correspond to the 1.5 × interquartile ranges. The shaded facets show that the total sample sizes (ntotal) are smaller than 3/*p*.

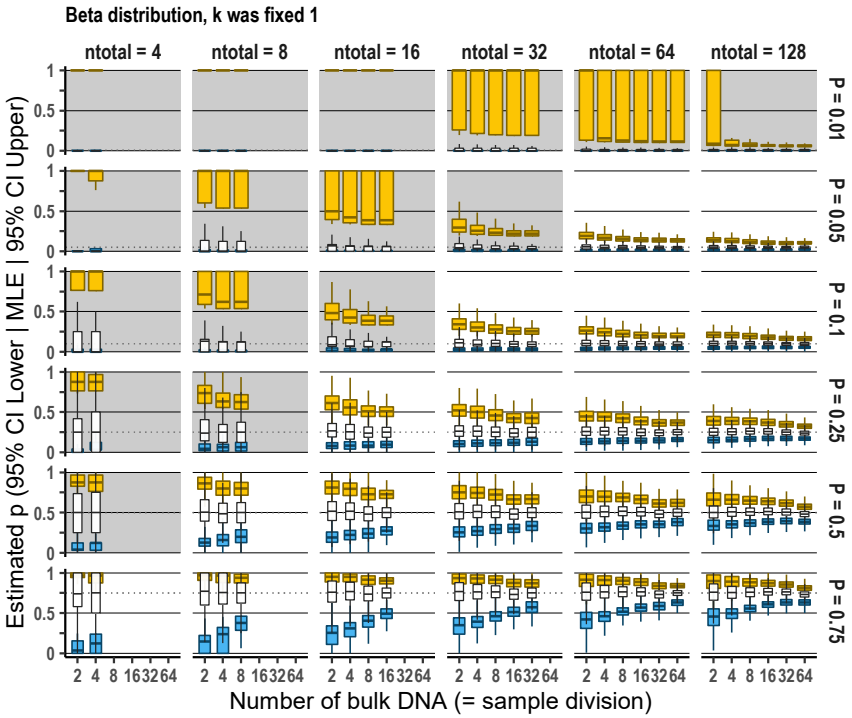


Figure 3 Estimation accuracy of the resistance allele frequency with freqpcr() when the beta distribution was assumed, fixing K = 1.

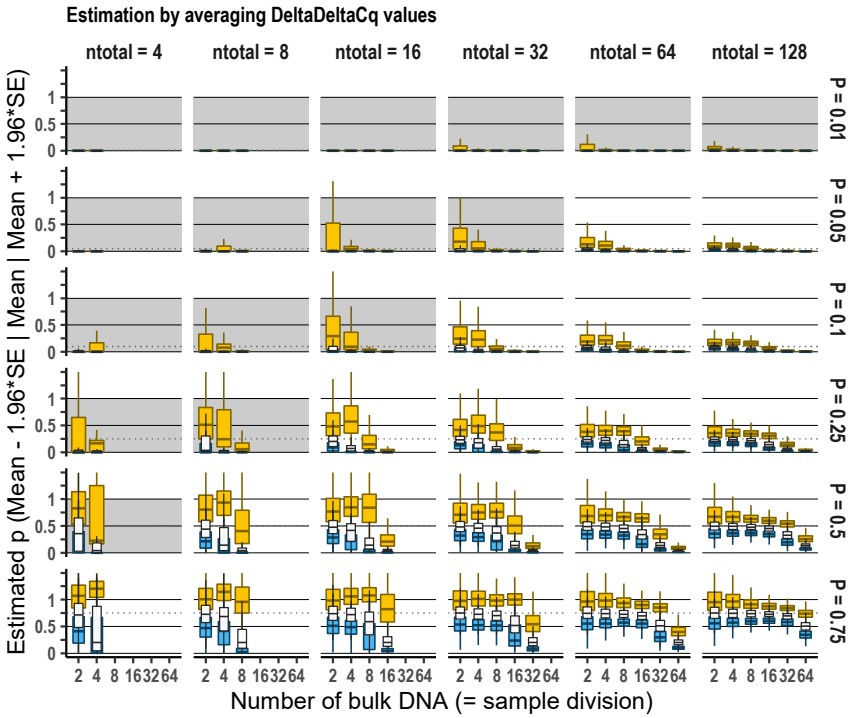


Figure 4 Estimation accuracy of the resistance allele frequency by simple averaging of ΔΔCq measures. The dummy dataset was derived from the numerical experiment of "beta distribution, all parameters unknown."

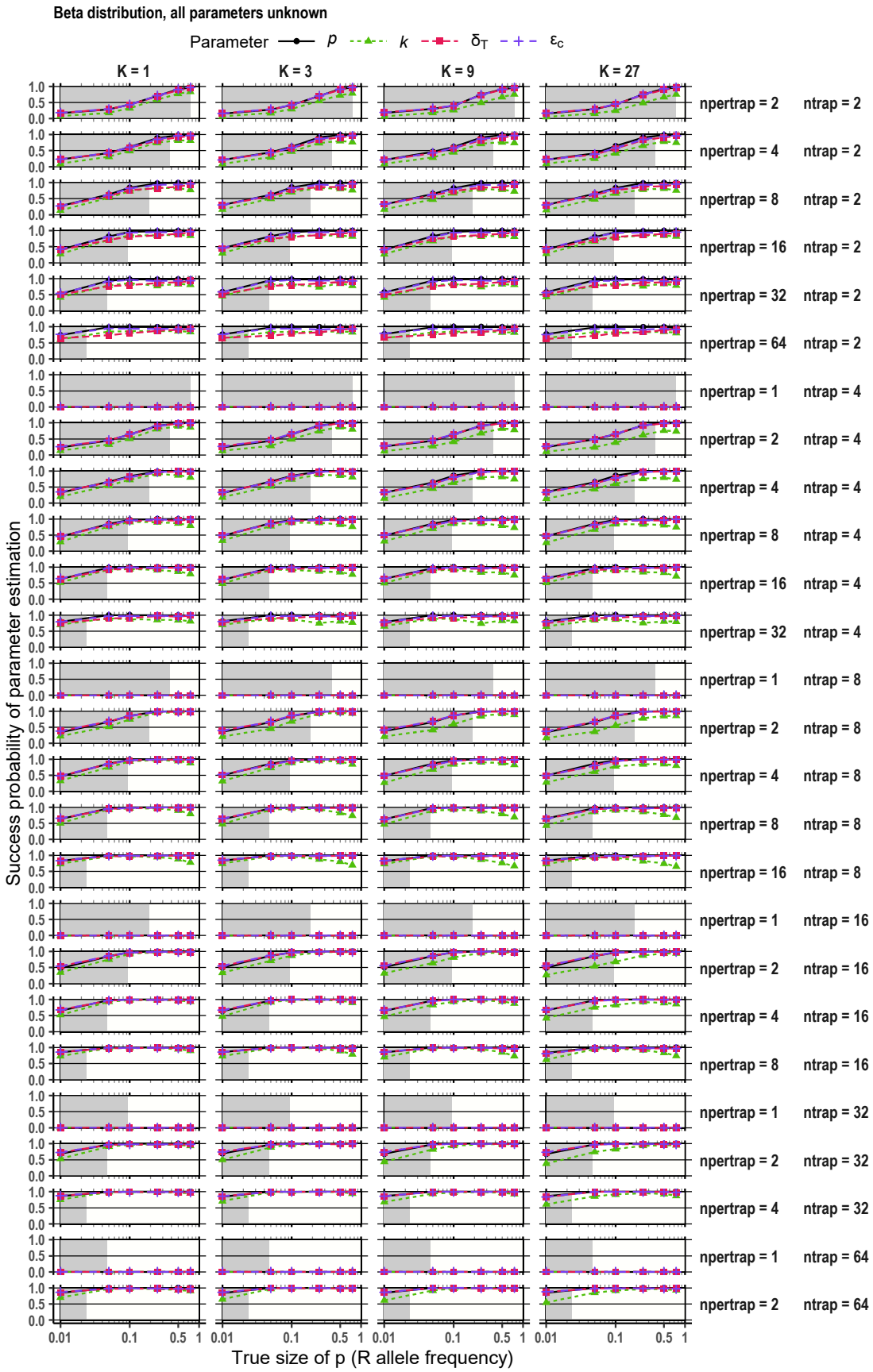


Figure S1 Probability of estimation success with freqpcr() in each parameter region. The beta distribution was assumed, and all estimable parameters (P, K, targetScale, and sdMeasure) were set as unknown. The shaded boxes in the background show the frequency ranges where the total sample sizes (ntotal) are smaller than $3/p$.

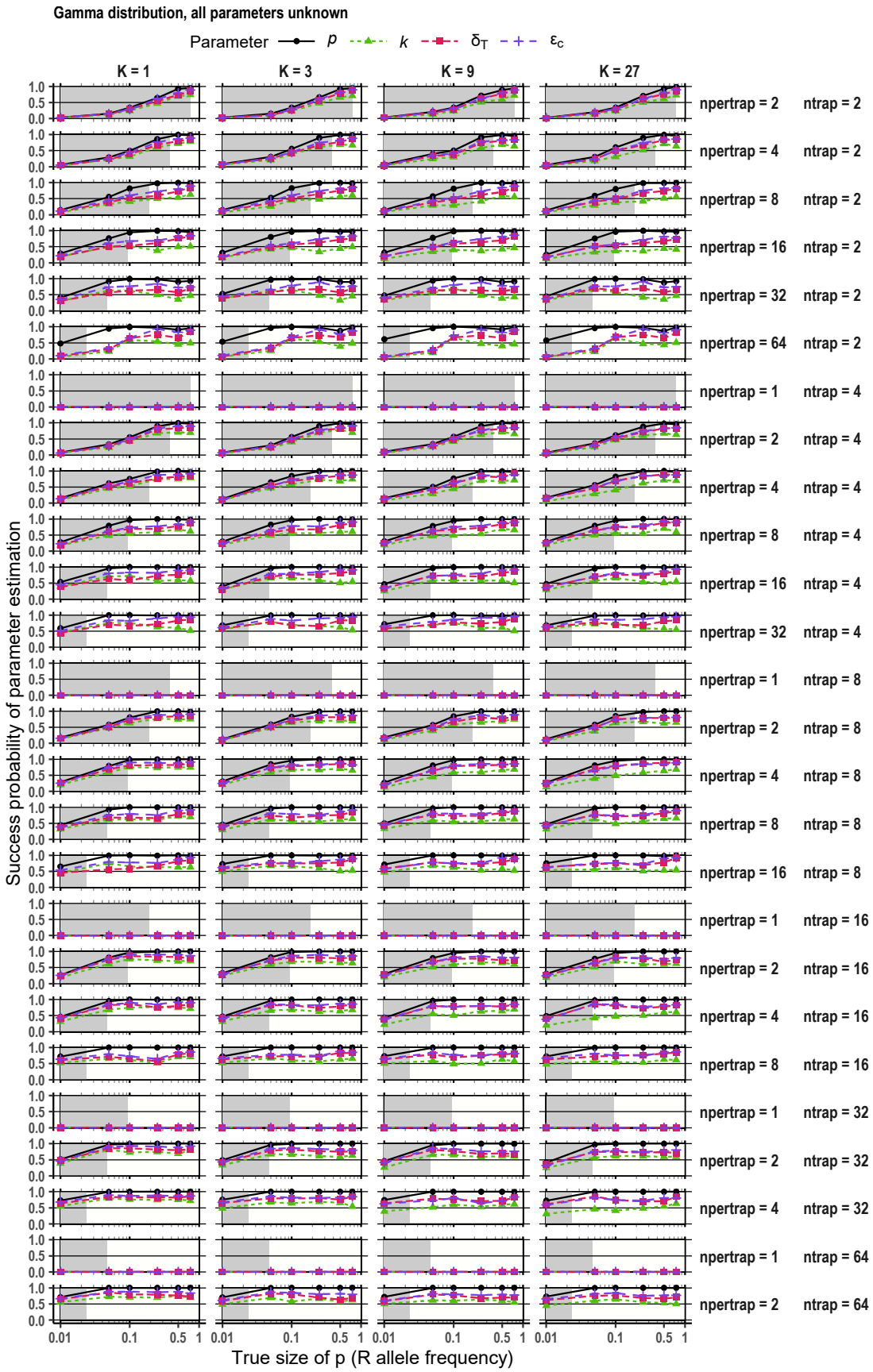


Figure S2 Probability of estimation success with freqpcr() in each parameter region. The gamma distributions were assumed, and all estimable parameters were set as unknown. The function often failed to calculate the CIs for *k* when npertrap (individuals in each bulk sample) were larger, possibly due to the accumulation of numerical calculation error.

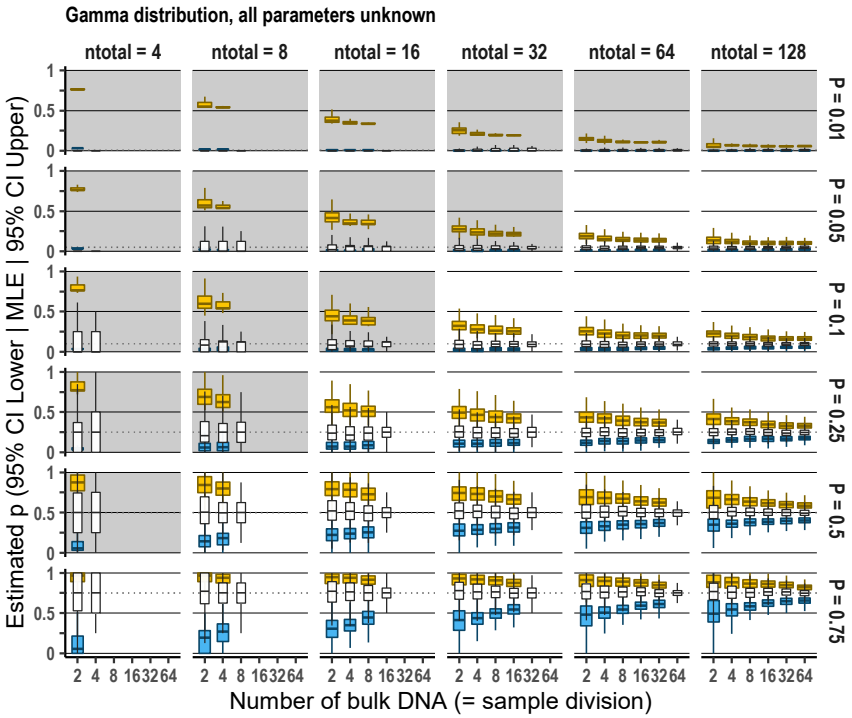
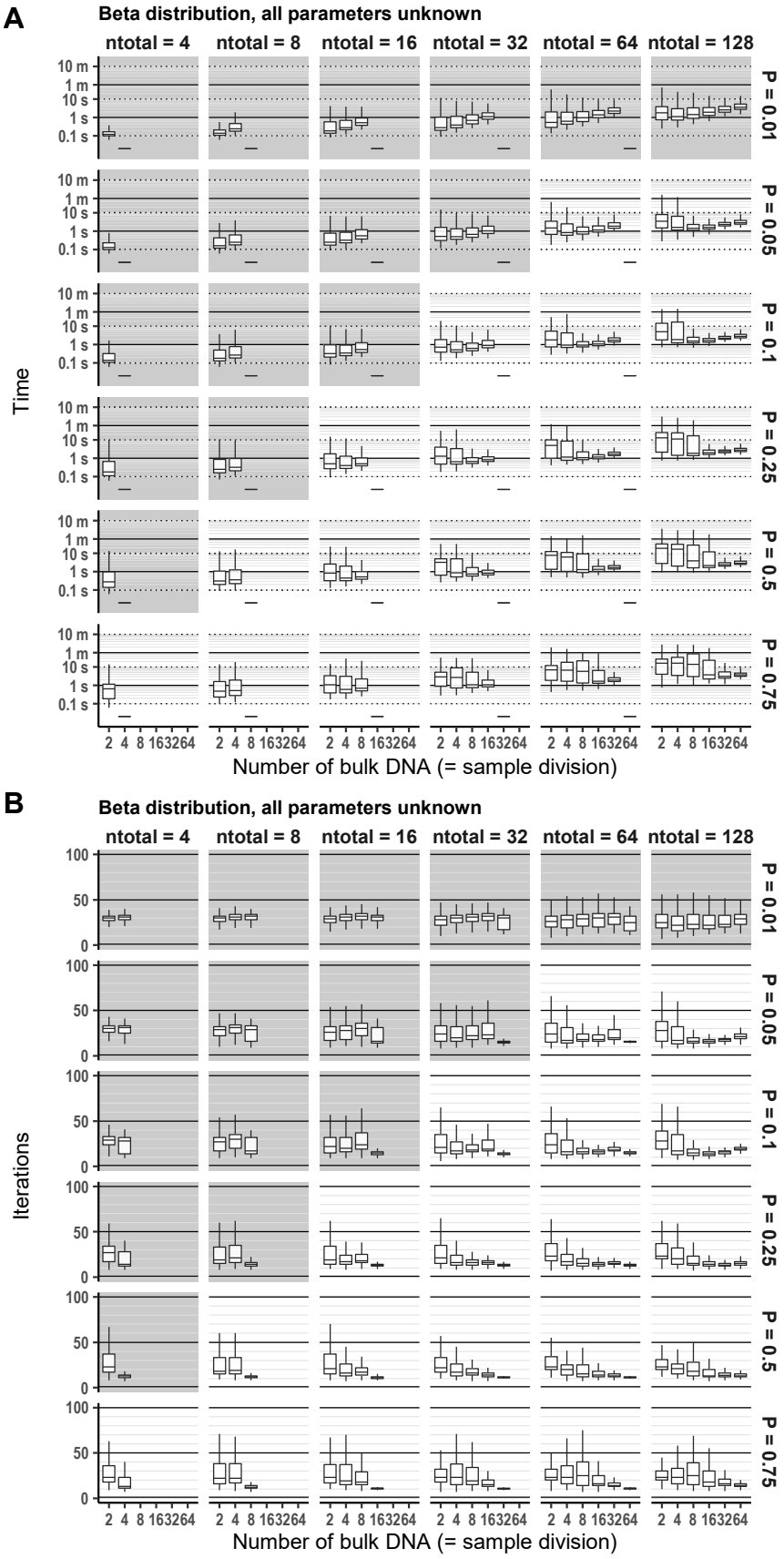


Figure S3 Estimation accuracy of p with freqpcr() when gamma distributions were assumed and all estimable parameters were set as unknown. The shaded facets show that the total sample sizes (ntotal) are smaller than $3/p$.



757

758 Figure S4 A: Calculation time and B: number of iterations until the freqpcr() function converges. The beta
759 distribution was assumed, and all estimable parameters were set as unknown.

760

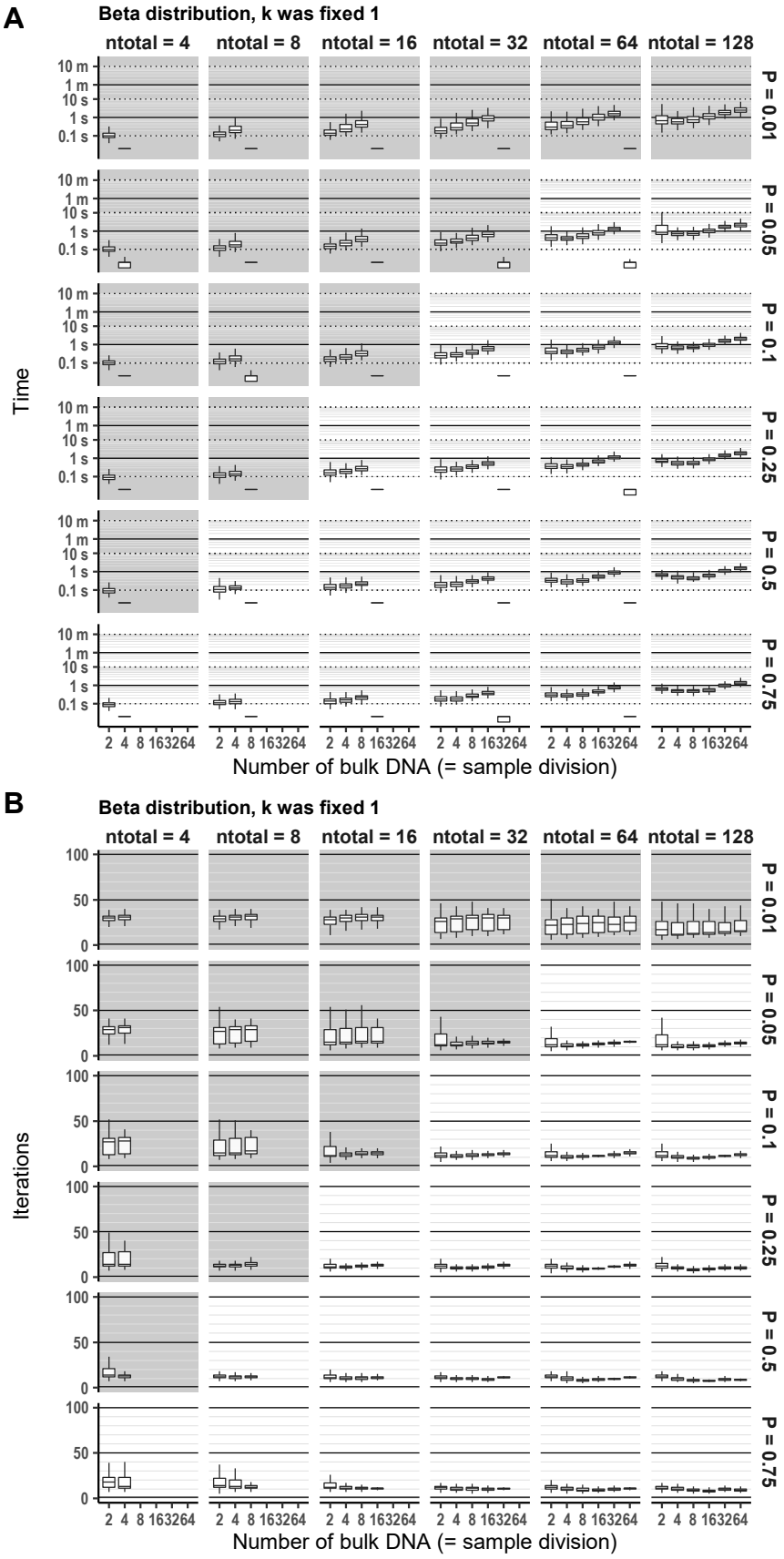
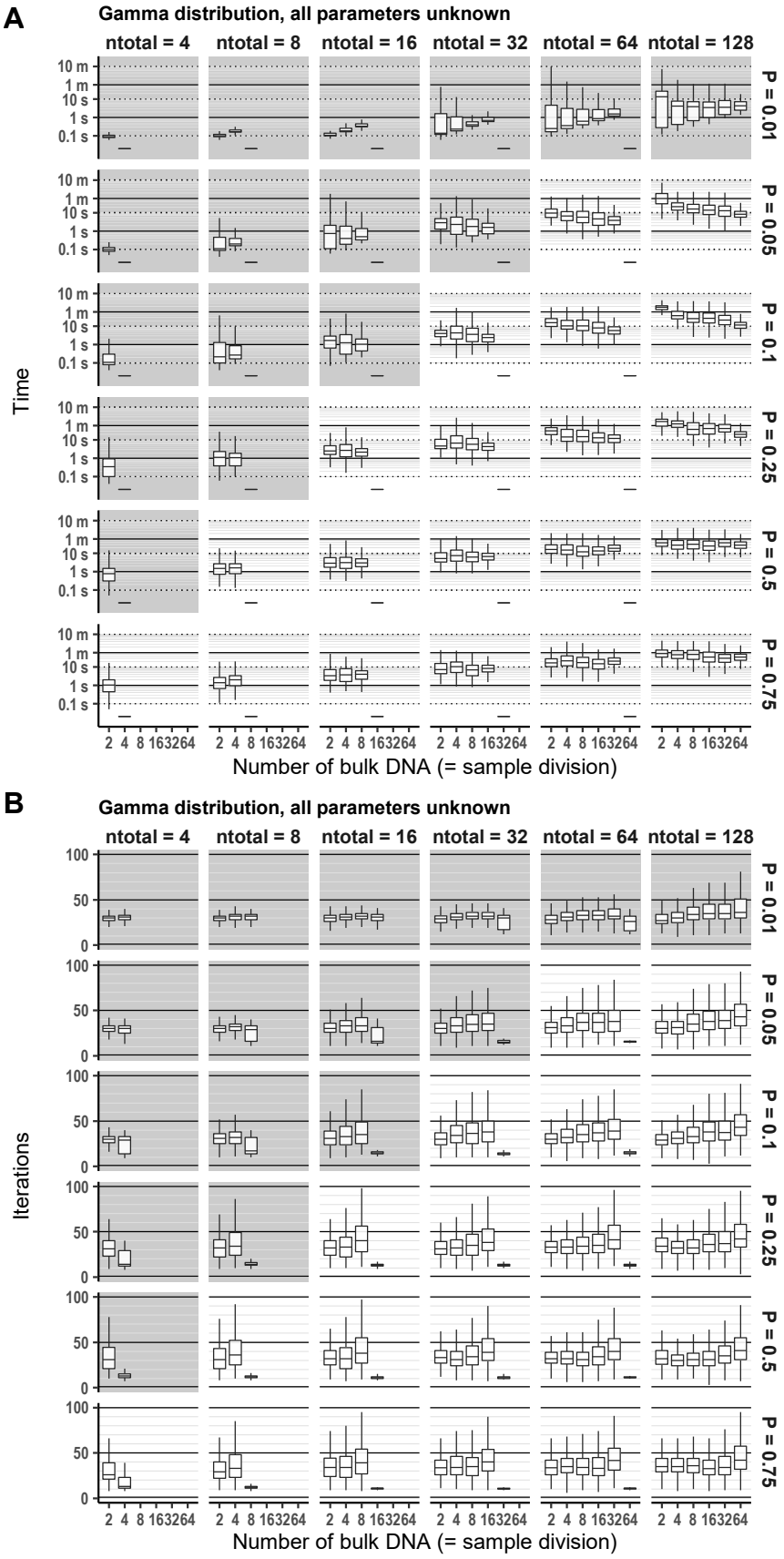


Figure S5 A: Calculation time and B: number of iterations until the freqpcr() function converges. The beta distribution was assumed, fixing the gamma shape parameter K = 1.



765

766 Figure S6 A: Calculation time and B: number of iterations until the freqpcr() function converges, assuming
767 gamma distributions. All estimable parameters were set as unknown.

768

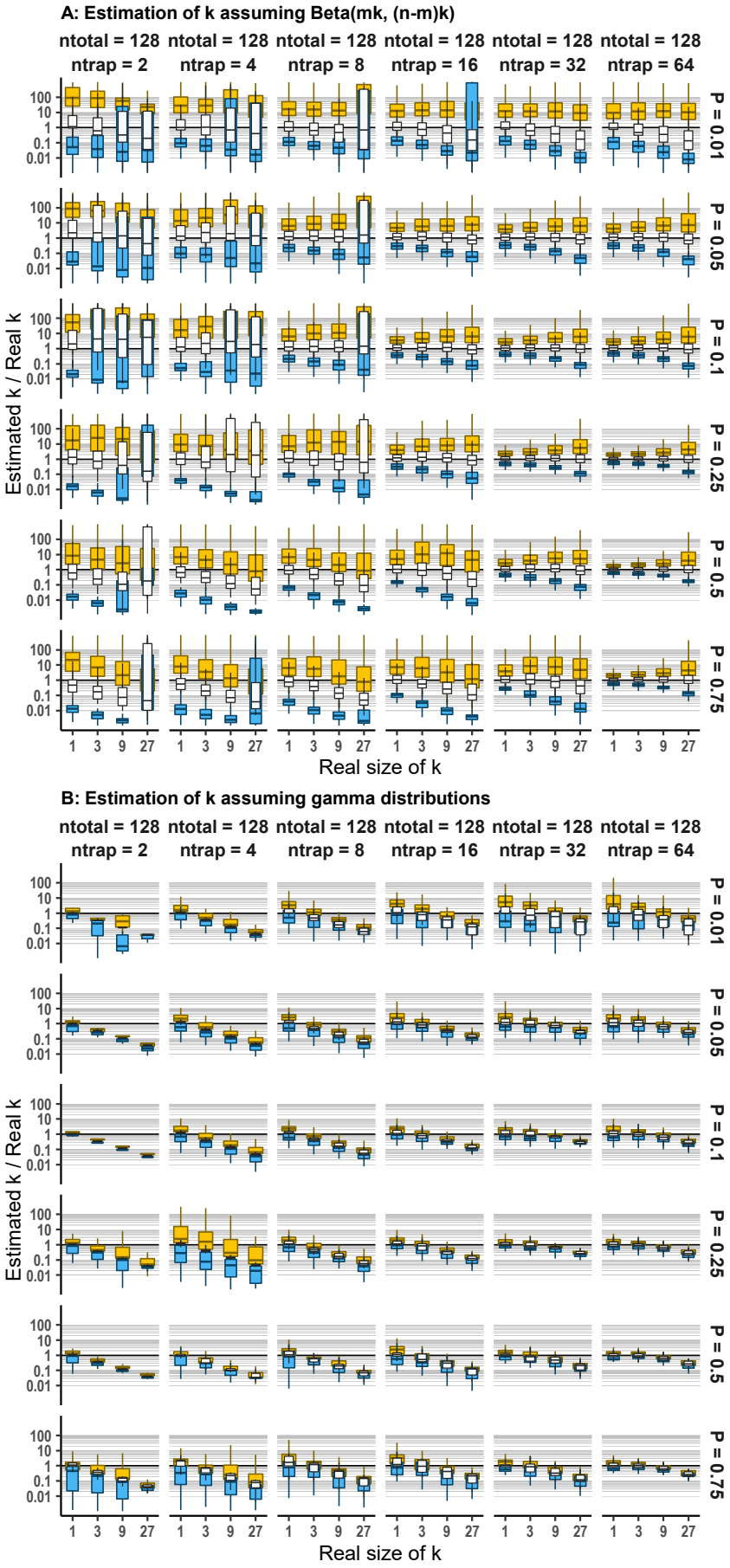


Figure S7 Estimation accuracy of *k* (the gamma shape parameter) in the simulation, showing the maximum likelihood estimate by freqpcr() divided by the actual parameter size.