freqpcr: estimation of population allele frequency using qPCR ΔΔCq measures from bulk samples

Short running title: 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

Telephone: +81-547-45-4419

ORCID ID:

0000-0001-9834-9857 (Masaaki Sudo)

0000-0002-2246-3431 (Masahiro Osakabe)

Manuscript type: Resource article: Molecular and Statistical Advances

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 numerous individuals, and subsequently handle each sample, makes quantification of rare mutations, including pesticide resistance genes at the early stages of resistance development, challenging. Meanwhile, pooling DNA from multiple individuals as a "bulk sample" may reduce handling costs. The qPCR output for a bulk sample, however, contains uncertainty owing to variations in DNA yields from each individual, in addition to measurement errors. In this study, we developed a statistical model for the interval estimation of allele frequency using ΔΔCq-based qPCR analyses of multiple bulk samples collected 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, while their confidence intervals suggest that collecting and pooling additional samples from individuals 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 specific alleles in populations is a key technique not only in population genetics and molecular ecology, but also in agricultural and regulatory sciences (Falconer, 1960; Kim et al., 2011; Yamamura & Hino, 2007). In applied entomology, field monitoring has been performed to detect resistance genes of arthropod pests to pesticides and genetically modified insecticidal plants, such as *Bt* crops (Andow & 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 associated with the treatment of living organisms have certain inherent drawbacks. Specifically, they are often labor-intensive and time-consuming. Although the resistance level can be directly measured using bioassays that detect the mortality of tested individuals, additional information, including the dominance of the resistance gene, is required to estimate 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), rapid advancements have recently been made in molecular diagnostics (Donnelly et al., 2016; Samayoa, et al., 2015; Toda et al., 2017). To quantify resistance-associated point mutations at the population scale, the most fundamental molecular technique is an individual-based polymerase chain reaction (PCR) analysis (Toda et al., 2017). Moreover, quantitative PCR (qPCR), based on real-time PCR, is also used for the point mutation of allele frequencies (Germer et al., 2000). 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, collecting, processing, and analyzing multiple DNA samples may not feasible, particularly when dealing with numerous samples from multiple sites, or when estimation

of a rare (<1%) mutation frequency is required for a given population, as is often the case in the early phase of resistance development.

Although rearing living insects is no longer necessary, the field of molecular diagnostics is still lacking a metaphorical silver bullet capable of reducing the required time and cost associated with 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 address some of these challenges. In fact, Osakabe *et al*. (2017) and Maeoka *et al*. (2020) developed diagnostic methods for detecting resistance to the acaricide, etoxazole, in the two-spotted spider mite, *Tetranychus urticae* Koch (Acari: Tetranychidae), which is conferred by an amino acid substitution in chitin synthase 1 (*CHS1*; I1017F) (Van Leeuwen et al., 2010). 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, known as the ΔΔCq method (Livak and Schmittgen, 2001). In the etoxazole-R diagnosis by Osakabe et al. (2017), glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) was used as the housekeeping gene.

In this study, we propose a statistical method for obtaining the interval estimate of allele frequency using ΔΔCq-based qPCR analyses forr multiple bulk samples taken from a population. We first introduced the random error structure to approximate the relative abundance 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 impacted the Cq measurements through ΔΔCq-based qPCR analysis. Finally, we combined the models and developed a maximum likelihood estimation 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). In fact, it has been suggested that body weights are distributed lognormally in many non-social insect species (Gouws et al., 2011).

In this study, we adopted a gamma, rather than lognormal, distribution 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. The two distributions are considered interchangeable (Wiens, 1999; Kundu and Manglick, 2005). Second, the sum and proportion of independent gamma distributions have closed forms under certain conditions. Using Eq. 1, let X ($X \geq 0$) be the DNA yield per single locus per individual:

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

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, while Appendix S1 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 $n - m$ carried S, the susceptible allele. When we capture insects from a wild population, the size of n is

obvious, however m is usually unknown (Figure 1A). Assuming random sampling from an infinite population with the R allele at the frequency p, m follows a binomial distribution (Eq. 2):

$$\mathrm{Bin}(m|n,p) = \frac{n!}{m!\,(n-m)!}p^m(1-p)^{n-m}. \qquad \textit{Eq. 2}$$

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 (Figure 1B).

$$X_\mathrm{R} \sim \mathrm{Ga}(mk, \theta),$$
$$X_\mathrm{S} \sim \mathrm{Ga}\big((n-m)k, \theta\big). \qquad \textit{Eq. 3}$$

When X_R and X_S independently follow gamma distributions with the same scale parameter, the observed allele frequency $Y_\mathrm{R} = X_\mathrm{R}/(X_\mathrm{S} + X_\mathrm{R})$ follows a beta distribution with the shape parameters mk and $(n-m)k$: Eq.

$$\mathrm{Beta}(Y_\mathrm{R}|mk, (n-m)k) = \frac{{Y_\mathrm{R}}^{mk-1}(1-Y_\mathrm{R})^{(n-m)k-1}}{\mathrm{B}(mk, (n-m)k)}, \qquad \textit{Eq. 4}$$

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).

Relative quantification of DNA by real-time PCR

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

The ΔΔCq (quantification cycle) method (Livak, 1997) is the most common method for relative quantification using qPCR, in which the quantities of complementary cDNA libraries are compared between samples to determine the relative expression levels of the genes of interest. Osakabe *et al.* (2017) expanded this concept and proposed the "RED-ΔΔCq method" (RED, restriction enzyme digestion), a derivative method that can measure the allele frequency from a single sample solution, to diagnose the regional prevalence of an acaricide-resistant point mutation in a *T. urticae* population.

In the RED-ΔΔCq method, the control was prepared as 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, however, was digested with restriction endonucleases prior to qPCR analysis (Figure 1A). 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 yet, positive variable giving the residual rate). Consequently, we obtained the template amount $X_{\mathrm{R}} + zX_{\mathrm{S}}$ at the target locus after digestion. To calibrate the template DNA amounts, the samples before and after digestion were also amplified using the primer set for a housekeeping gene as an internal reference.

Taken together, the single bulk sample results in a quartet of Cq measurements differentiating at the target loci (resistance-associated and housekeeping genes) × restriction enzyme digestion (undigested and digested). We can then formulate the allele frequencies by letting X^{HW} and X^{TW} represent the total amounts of template DNA at the housekeeping (H) and target (T) loci, respectively, included in the sample without digestion, the state denoted by W (Figure 1C).

$$\begin{aligned} X^{\mathrm{HW}} &= X_{\mathrm{R}} + X_{\mathrm{S}}, \\ X^{\mathrm{TW}} &= \delta_{\mathrm{T}}(X_{\mathrm{R}} + X_{\mathrm{S}}). \end{aligned} \qquad \textit{Eq. 5}$$

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

156
$$\begin{aligned} X^{\mathrm{HD}} &= \delta_{\mathrm{B}}(X_{\mathrm{R}} + X_{\mathrm{S}}), \\ X^{\mathrm{TD}} &= \delta_{\mathrm{B}}\delta_{\mathrm{T}}(X_{\mathrm{R}} + zX_{\mathrm{S}}). \end{aligned}$$

157 *Eq. 6*

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

160 As a result of qPCR, the Cq quartet, τ^{HW}, τ^{TW}, τ^{HD}, and τ^{TD} were obtained as:

161
$$\begin{aligned} \tau^{\mathrm{HW}} &= \frac{\ln(X_\Theta) - \ln(X_{\mathrm{R}} + X_{\mathrm{S}})}{\ln(1+\eta)} + \varepsilon_{\mathrm{c}}, \\ \tau^{\mathrm{TW}} &= \frac{\ln(X_\Theta) - \ln\delta_{\mathrm{T}} - \ln(X_{\mathrm{R}} + X_{\mathrm{S}})}{\ln(1+\eta)} + \varepsilon_{\mathrm{c}}, \end{aligned}$$

162
$$\begin{aligned} \tau^{\mathrm{HD}} &= \frac{\ln(X_\Theta) - \ln\delta_{\mathrm{B}} - \ln(X_{\mathrm{R}} + X_{\mathrm{S}})}{\ln(1+\eta)} + \varepsilon_{\mathrm{c}}, \\ \tau^{\mathrm{TD}} &= \frac{\ln(X_\Theta) - \ln\delta_{\mathrm{B}} - \ln\delta_{\mathrm{T}} - \ln(X_{\mathrm{R}} + zX_{\mathrm{S}})}{\ln(1+\eta)} + \varepsilon_{\mathrm{c}}. \end{aligned}$$

163 *Eq. 7*

164 Here, $1 + \eta$ ($\eta > 0$) and X_Θ denote the amplification efficiency per PCR cycle and its threshold,
165 respectively. According to Livak and Schmittgen (2001), we assume an ideal amplification, where X_Θ
166 is set within the early exponential amplification phase. The actual Cq data contain measurement errors
167 in addition to uncertainty due to experimental operations, such as sample dispensation or PCR
168 amplification. We express these using the common error term $\varepsilon_{\mathrm{c}} \sim \mathrm{N}(0, \sigma_{\mathrm{c}}^2)$, following the normal
169 distribution of mean = 0 and variance = σ_{c}^2 in the scale of raw Cq values. The validity of this error
170 structure is verified later.

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

$$\Delta\Delta\tau = \Delta\tau^{\mathrm{D}} - \Delta\tau^{\mathrm{W}} = -\frac{\ln\left(\frac{X_{\mathrm{R}} + zX_{\mathrm{S}}}{X_{\mathrm{R}} + X_{\mathrm{S}}}\right)}{\ln(1+\eta)} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, 4\sigma_{\mathrm{c}}^2) \cdot \qquad \textit{Eq. 8}$$

From Eq. 8, the expected value of $(X_{\mathrm{R}} + zX_{\mathrm{S}})/(X_{\mathrm{R}} + X_{\mathrm{S}})$ is calculated as $(1+\eta)^{-\Delta\Delta\tau}$. The coefficients δ_{B} and δ_{T} in Eq. 5 and Eq. 6 vanished by subtracting the Cq values and ΔCq values, respectively.

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 $(X_{\mathrm{R}} + zX_{\mathrm{S}})/(X_{\mathrm{R}} + X_{\mathrm{S}}) = \hat{Y}_{\mathrm{R}} + z\left(1 - \hat{Y}_{\mathrm{R}}\right)$ 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 a larger error exists 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. 7, 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 δ_{T} size of 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 were to be 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 T and H loci, respectively. Unlike the RED-ΔΔCq method, the control sample was not taken from the test sample solution, but rather was prepared as a DNA solution containing 100% R, hereafter denoted as U (= pUre R line) (Figure 1C).

X^{HU} and X^{TU} then denote the template DNA quantities:

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

Eq. 9

Though the definition of δ_{T} is the same as Eq. 5, the quantity is denoted by X'_{R} instead of $X_{\mathrm{S}} + X_{\mathrm{R}}$ as 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:

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

Eq. 10

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 R-specific primer set. As the primer set for the housekeeping gene was nonspecific, X^{HV} was 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. 8.

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. 7, the Cq values (denoted as τ_h^{HW}, τ_h^{TW}, τ_h^{HD}, and τ_h^{TD} for each bulk sample) are determined not only by the DNA quantities, denoted as $X_{h,\mathrm{R}}$ and $X_{h,\mathrm{S}}$, 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.

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:

$$\Delta\tau_h^{\mathrm{W}} \sim \mathrm{N}\left(-\frac{\ln\delta_{\mathrm{T}}}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right),$$

$$\Delta\tau_h^{\mathrm{D}} \sim \mathrm{N}\left(-\frac{\ln\delta_{\mathrm{T}} + \ln\left(\frac{X_{h,\mathrm{R}} + zX_{h,\mathrm{S}}}{X_{h,\mathrm{R}} + X_{h,\mathrm{S}}}\right)}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right). \qquad \textit{Eq. 11}$$

Although Eq. 11 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. 11, $\Delta\tau_h^{\mathrm{W}}$ 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. 7. Moreover, cancelation of δ_{B} also ensures that we can apply the model of Eq. 11 to the general ΔΔCq method of Eq. 9 and Eq. 10.

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 *h*th 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 *h*th 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. 12

In Eq. 12, ψ_{G} or ψ_{B} denotes the probability of obtaining $\Delta\tau_h^{\mathrm{D}}$ under the template DNA quantities of $X_{h,\mathrm{R}|m_h} = r$ and $X_{h,\mathrm{S}|m_h} = s$ if we formularize the two quantities by gamma distribution, or if we formularize their mixing ratio by the single beta distribution, respectively. 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} under the whole region of $X_{h,\mathrm{R}|m_h} = r$ and $X_{h,\mathrm{S}|m_h} = s$ for the interval $\{D: 0 \le r < \infty, 0 \le s < \infty\}$.

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

Eq. 13

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 $\ln[(r+zs)/(r+s)]$ in Eq. 13.

Since the computational burden for the double integration is large, we simplified the likelihood model with the beta distribution. By introducing $y = r/(r+s)$, the probability of obtaining $\Delta\tau_h^{\mathrm{D}}$ is replaced with ψ_{B} defined as follows:

$$\psi_{\mathrm{B}} = \int_0^1 \mathrm{N}\left(-\frac{\ln\delta_{\mathrm{T}}+\ln\big(z+y(1-z)\big)}{\ln(1+\eta)}, 2\sigma_{\mathrm{c}}^2\right)\mathrm{Beta}(y|m_h k,(n_h-m_h)k)dy. \qquad \textit{Eq. 14}$$

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 default is freqpcr(..., beta = TRUE), where the beta distribution model of Eq. 14 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 if we have conducted a (RED-)ΔΔCq analysis on multiple bulk samples. However, the size of z, the residue rate of the S allele, 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. 11, 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. 7 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. The "knownqpcr_unpaired" function was developed to handle incomplete data (i.e., the observations of τ^{HW}, τ^{TW}, τ^{HD}, and τ^{TD} have different data lengths). If the four Cq measures are available for all samples, then "knownqpcr" is used.

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 etoxazole-R allele frequency in *T. urticae* and the corresponding $2^{-\Delta\Delta Cq}$ measures (the approximate point estimate of the frequency), Osakabe *et al.* (2017) demonstrated linearity using a sample series of 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 $(1+\eta)^{-\Delta\Delta Cq}$.

Materials and 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 ten 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 of δ_B, δ_T, σ_c, z, and η was conducted with the "knownqpcr_unpaired" function of the freqpcr package (version 0.3.2). 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. 7, 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.

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 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) of the freqpcr package version 0.3.1 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 (p, k, δ_T, and σ_c) unknown. Further, the easiest way to estimate p derived from Eq. 8, we averaged the observed ΔΔCq values for *N* bulk samples and transformed them as $\hat{p} = (1 + \eta)\hat{}(-\overline{\Delta\Delta\tau})$.

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 $\hat{Y}_R$ (Figure 2). 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) (Figure 2A). 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 (Figure 2B). 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. 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

406 estimated confidence interval was accurate. The estimation success for the Cq measurement error, σ_c,
407 was 69.6% and 97.6% in the beta-distribution model with unknown k and $k = 1$, respectively. The
408 relative quantity of the target gene, δ_T, was 68.1% and 96.1%, respectively. However, the estimated
409 success of k was 59.9% with the beta distribution model, showing a lower performance than the other
410 parameters, implying that the likelihood is insensitive to the size of k. Conversely, the estimation of p
411 is robust to the size of k, as we show later in this section.

412 The estimation success of freqpcr() largely depended on the total sample size (Nn corresponding to
413 the facet 'ntotal' in the figures), as well as the level of p (Figure S1 and S2 for the beta and gamma
414 models, with all parameters unknown). In each parameter region, the quantity $\mathrm{Bin}(0|Nn, p)$ generally
415 gives the probability that the whole sample contains no R individuals. When Nn is larger enough,
416 $Nn > 3/p$ is approximately the requirement for the total sample size to contain at least one R
417 individual with 95% confidence, called the "rule of three" (Eypasch et al., 1995). The gray
418 backgrounds in the facets of Figures 3–4 and S1–S7 signify the regions where the total sample sizes
419 are smaller than the thresholds (e.g., 60 haploid individuals are required when $p = 0.05$). As shown in
420 Figures S1 and S2, the parameter estimation often failed when Nn did not meet the rule of three. Once
421 we exclude the parameter regions of $Nn \leq 3/p$, the estimation success rate of p with freqpcr(..., beta
422 = TRUE) improved to 84.3% and 99.9% with all parameters unknown and assuming $k = 1$,
423 respectively.

424 As for the estimation accuracy of p, the freqpcr() function assuming beta distribution provides an
425 unbiased estimator. Figures 3 and S3 show the estimated sizes of p using the beta model with all
426 parameters unknown and assuming $k = 1$, respectively. Both settings demonstrated that the estimator
427 converged to the true R frequency; the upper/lower bounds of the estimated 95% confidence intervals
428 (yellow/blue boxes in each plot) became narrower as we increased the total sample sizes (Nn) or
429 included more bulk DNA samples (N). Fixing the size of the gamma shape parameter to $k = 1$scarcely
430 affected the point estimates and intervals of p, as long as $Nn > 3/p$ is satisfied (Figure S3). However,

431 if every individual was analyzed separately, the interval estimation was only possible when *k* was
432 fixed (see the regions of "sample division = ntotal" cases in Figure 3).

433 When we used the gamma distribution model, the interval estimation of *p* was also possible and
434 unbiased (Figure S4). However, when we defined the point estimator of *p* as a simple average, that is,
435 $\hat{p} = (1+\eta)\hat{}(-\overline{\Delta\Delta\tau})$, it was strongly underestimated as the samples were more divided (N/Nn was
436 large) (Figure 4). The upper limit of 95% CI often violated 1, suggesting that the "simple average of
437 ΔΔCq" ± 1.96 SE is inadequate for the interval estimation based on the RED-ΔΔCq method.

438 Although the freqpcr() function with the gamma and beta distributions both showed an unbiased
439 estimation of *p*, the gamma model was disadvantageous regarding calculation time and the number of
440 iterations before convergence. The time varied largely in the model settings and sample sizes (Figures
441 S5–S7). Among the settings we tried, beta model with fixed *k* was the fastest and converged within a
442 few seconds in most parameter regions (median and 75 percentile: 0.32 and 0.69 s: Figure S6). It was
443 three and >10 times faster than the beta (0.91 and 2.4 seconds: Figure S5) and gamma (3.0 and 15 s:
444 Figure S7) model, respectively with all parameters unknown. The calculation time increased as the
445 dataset size increased - *Nn* and the sample was more divided (larger N/Nn) in the beta distribution
446 model, because the marginal likelihood was calculated for each bulk sample (Figures S5 and S6).
447 Conversely, the gamma distribution model (Figure S7) requires increased calculation time as the size
448 of each bulk sample becomes larger (larger n_h). This was considered because the combination of
449 $\mathrm{Bin}(m_h|n_h, p)$ exploded when n_h was large.

450 Furthermore, the estimation accuracy of the shape parameter, *k*, it was underestimated as the real
451 size of the parameter increased (e.g., $k = 27$) when the gamma distribution model was applied (Figure
452 S8B). Since the iterative fitting of the parameter in freqpcr() always starts internally from $k = 1$ (this
453 was determined due to the calculation stability), this bias suggests the likelihood function of ψ_G (Eq.
454 13), with little information on the size of *k* compared with *p*. Then, *k* tends to stay at its initial value,
455 suggesting that the gamma model is not suitable for the simultaneous estimation of *p* and *k*. Unlike the
456 gamma version, the fitting of *k* with freqpcr(beta = TRUE) was satisfactory when we divided the total

457 samples into more bulk samples (larger N/Nn), although the initial value dependence was still
458 observed, especially when p or N was small (Figure S8A). This may be because the estimation of k via
459 $\mathrm{Beta}(m_h k, (n_h - m_h)k)$ in Eq. 14 is comparable with measuring the overdispersion of $Y_{h,\mathrm{R}|m_h}$, which
460 is only possible when multiple bulk samples contain both R and S alleles.

461 Discussion

462 In the present study, we developed a statistical model to estimate the population allele frequency based
463 on qPCR across multiple bulk samples to address the issues facing the conventional point estimator for
464 allele frequency which averages the observed ΔΔCq values $\hat{p} = (1+\eta)\hat{}(-\overline{\Delta\Delta\tau})$. This conventional
465 method sometimes exceeds 1 when the frequency of the target allele is close to 1. Furthermore, when
466 one tries to quantify the rare mutant allele in a population, most bulk samples contain only the wild
467 type allele. The conventional $\hat{p}$ is vulnerable to many zero samples, which makes the frequency
468 estimation more difficult when p is small. To circumvent these problems, our interval estimation
469 explicitly models the number of individuals contained in each bulk sample (the binomial assumption)
470 as well as the individual DNA yields (the gamma assumption), thereby obtaining the interval estimate
471 over the entire range $0 < p < 1$.

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

480 The combination of mass trapping and bulk qPCR analysis solves the latter challenge by collecting
481 more individuals and pooling them. This can result in higher precision with less work than individual
482 PCR. For instance, we sampled 16 individuals from the population with an allele frequency of $p = 0.05$
483 and analyzed two individuals once in the numerical experiment (Figure 3: facet of ntotal = 16, sample
484 division = 8). The lower and upper limits of the 95% confidence interval p were estimated to be
485 0.0087 and 0.34, respectively, using freqpcr(..., beta = TRUE) (as the medians of the 1,000
486 independent trials). We also simulated the case of ntotal = 64 and sample division = 4 (i.e., analyzed
487 16 individuals together) and found the upper and lower limits to be 0.015 and 0.15, respectively. Thus,
488 we improved the precision of the interval estimate with half the handling effort.

489 Also, in non-quantitative PCR, sample pooling is considered as a tool for the detection of rare
490 (c)DNA in the population with practical labor requirements, and has been used as high throughput pre-
491 screening system for many samples e.g. in clinical examinations (Taylor et al., 2010; Yelin et al.,
492 2020). However, in some fields, such as plant quarantine, it is important to guarantee that a product is
493 not contaminated with pests or unapproved genetically modified seeds at a certain consumer risk. As
494 the assumed frequency range is low ($p \approx 0.001$), frequency estimation is not realistic (3,000 seeds are
495 needed to meet the "rule of three" when $p = 0.001$) and is not required for the current inspection
496 routine. Thus, group testing based on non-quantitative PCR has been conducted in these fields
497 (Yamamura et al., 2019). Yamamura and Hino (2007) proposed a procedure to estimate the upper limit
498 of the population allele frequency, in which they used the proportion of bulk samples detected as
499 "positive."

500 Overall, there has been a gap in methodology between the frequency estimation based on the
501 individual PCR and the non- or semi-quantitative PCR based on the non-quantitative bulk PCR.
502 Although it provides the highest estimation precision following binomial distribution, the former is
503 only available at a higher p; it becomes labor-intensive once we try to quantify rare alleles. The latter
504 can be applied to a lower range of p, but the precision is generally low or even non-quantitative.
505 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. 11 can also be applied for other qPCR protocols based on ΔΔCq. 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. 10 and X^{TU} in Eq. 9, 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 3), 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 S3), 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 negates 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 each bulk sample before the estimation with freqpcr(),

532 although the estimated size of σ_c changes from its original definition in Eq. 7. However, it is trivial if
533 the number of technical replicates is unified between bulk samples.

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

541 Acknowledgments

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

546 References

547 Andow, D. A., & Alstad, D. N. (1998). F2 screen for rare resistance alleles. *Journal of Economic*
548 *Entomology*, 91(3), 572–578.

549 Chen, H., Rangasamy, M., Tan, S. Y., Wang, H., & Siegfried, B. D. (2010). Evaluation of five
550 methods for total DNA extraction from western corn rootworm beetles. *PLoS One*, 5(8), e11963.

551 Donnelly, M. J., Isaacs, A. T., & Weetman, D. (2016). Identification, validation, and application of
552 molecular diagnostics for insecticide resistance in malaria vectors. *Trends in Parasitology*, 32(3),
553 197–206.

554 Eypasch, E., Lefering, R., Kum, C. K., & Troidl, H. (1995). Probability of adverse events that have not
555 yet occurred: A statistical reminder. *BMJ*, 311(7005), 619–620.

556 Falconer, D. S. (1960). *Introduction to quantitative genetics*. Oliver And Boyd; Edinburgh; London.

ffrench-Constant, R. H. (2013). The molecular genetics of insecticide resistance. *Genetics*, 194(4), 807–815.

Germer, S., Holland, M. J., & Higuchi, R. (2000). High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Research*, 10(2), 258–266.

Gould, F., Anderson, A., Jones, A., Sumerford, D., Heckel, D. G., Lopez, J., Micinski, S., Leonard, R., & Laster, M. (1997). Initial frequency of alleles for resistance to *Bacillus thuringiensis* toxins in field populations of *Heliothis virescens*. *Proceedings of the National Academy of Sciences*, 94(8), 3519–3523.

Gouws, E. J., Gaston, K. J., & Chown, S. L. (2011). Intraspecific body size frequency distributions of insects. *PLoS One*, 6(3), e16606.

Hothorn, T., Zeileis, A., Farebrother (pan.f), R. W., Cummins (pan.f), C., Millo, G., & Mitchell, D. (2019). lmtest: Testing Linear Regression Models (0.9-37) [Computer software]. https://CRAN.R-project.org/package=lmtest

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., & Witte, D. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, 12(1), 231.

Knapp, M. (2016). Relative importance of sex, pre-starvation body mass and structural body size in the determination of exceptional starvation resistance of *Anchomenus dorsalis* (Coleoptera: Carabidae). *PloS One*, 11(3), e0151459.

Kundu, D., & Manglick, A. (2005). Discriminating between the log-normal and gamma distributions. *Journal of the Applied Statistical Sciences*, 14, 175–187.

Kwon, D. H., Yoon, K. S., Strycharz, J. P., Clark, J. M., & Lee, S. H. (2008). Determination of permethrin resistance allele frequency of human head louse populations by quantitative sequencing. *Journal of Medical Entomology*, 45(5), 912–920.

Li, G., Reisig, D., Miao, J., Gould, F., Huang, F., & Feng, H. (2016). Frequency of Cry1F non-recessive resistance alleles in North Carolina field populations of *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *PloS One*, 11(4), e0154492.

Livak, K. J. (1997). Comparative Ct method. In *ABI Prism 7700 sequence detection system User Bulletin #2*, P/N 4303859 (Vol. 2, pp. 11–15). Applied Biosystems. http://tools.thermofisher.com/content/sfs/manuals/cms_040980.pdf

Livak, Kenneth J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2- ΔΔCT method. *Methods*, 25(4), 402–408.

Maeoka, A., Yuan, L., Itoh, Y., Saito, C., Doi, M., Imamura, T., Yamaguchi, T., Imura, T., & Osakabe, M. (2020). Diagnostic prediction of acaricide resistance gene frequency using quantitative real-time PCR with resistance allele-specific primers in the two-spotted spider mite *Tetranychus urticae* population (Acari: Tetranychidae). *Applied Entomology and Zoology*, 55, 329–335. https://doi.org/10.1007/s13355-020-00686-7

May, R. M. (1976). Models for single populations. In *Theoretical ecology* (pp. 5–29). Blackwell.

Narasimhan, B., Koller, M., Johnson, S. G., Hahn, T., Bouvier, A., Kiêu, K., & Gaure, S. (2019). cubature: Adaptive Multivariate Integration over Hypercubes (2.0.4) [Computer software]. https://CRAN.R-project.org/package=cubature

Osakabe, M., Imamura, T., Nakano, R., Kamikawa, S., Tadatsu, M., Kunimoto, Y., & Doi, M. (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.

R Core Team. (2019). R version 3.6.1. https://www.r-project.org/

Rakovski, C., Weisenberger, D. J., Marjoram, P., Laird, P. W., & Siegmund, K. D. (2011). Modeling measurement error in tumor characterization studies. *BMC Bioinformatics*, 12(1), 284.

Samayoa, L. F., Malvar, R. A., Olukolu, B. A., Holland, J. B., & Butrón, A. (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(1), 35.

Snoeck, S., Kurlovs, A. H., Bajda, S., Feyereisen, R., Greenhalgh, R., Villacis-Perez, E., Kosterlitz, O., Dermauw, W., Clark, R. M., & Van Leeuwen, T. (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., Inukai, K., Kitabayashi, S., Kuwazaki, S., & Jouraku, A. (2017). Molecular evaluation of diamide resistance in diamondback moth (Lepidoptera: Yponomeutidae) populations using quantitative sequencing. *Applied Entomology and Zoology*, 52(2), 353–357.

Sudo, M., Takahashi, D., Andow, D. A., Suzuki, Y., & Yamanaka, T. (2018). Optimal management strategy of insecticide resistance under various insect life histories: Heterogeneous timing of selection and interpatch dispersal. *Evolutionary Applications*, 11(2), 271–283.

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

Sugimoto, N., Takahashi, A., Ihara, R., Itoh, Y., Jouraku, A., Van Leeuwen, T., & Osakabe, M. (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., Patin, A. L., Dennehy, T. J., Liu, Y.-B., Carriere, Y., Sims, M. A., & Antilla, L. (2000). Frequency of resistance to *Bacillus thuringiensis* in field populations of pink bollworm. *Proceedings of the National Academy of Sciences*, 97(24), 12980–12984.

Takahashi, D., Yamanaka, T., Sudo, M., & Andow, D. A. (2017). Is a larger refuge always better? Dispersal and dose in pesticide resistance evolution. *Evolution*, 71(6), 1494–1503.

630 Taylor, S. M., Juliano, J. J., Trottman, P. A., Griffin, J. B., Landis, S. H., Kitsa, P., Tshefu, A. K., &
631 Meshnick, S. R. (2010). High-throughput pooling and real-time PCR-based strategy for malaria
632 detection. *Journal of Clinical Microbiology*, 48(2), 512–519.

633 Toda, S., Hirata, K., Yamamoto, A., & Matsuura, A. (2017). Molecular diagnostics of the R81T
634 mutation on the D-loop region of the β1 subunit of the nicotinic acetylcholine receptor gene
635 conferring resistance to neonicotinoids in the cotton aphid, *Aphis gossypii* (Hemiptera: Aphididae).
636 *Applied Entomology and Zoology*, 52(1), 147–151.

637 Uesugi, R., Hinomoto, N., & Goto, C. (2016). Estimated time frame for successful PCR analysis of
638 diamondback moths, *Plutella xylostella* (Lepidoptera: Plutellidae), collected from sticky traps in
639 field conditions. *Applied Entomology and Zoology*, 51(3), 505–510.

640 Van Leeuwen, T., Demaeght, P., Osborne, E. J., Dermauw, W., Gohlke, S., Nauen, R., Grbić, M.,
641 Tirry, L., Merzendorfer, H., & Clark, R. M. (2012). Population bulk segregant mapping uncovers
642 resistance mutations and the mode of action of a chitin synthesis inhibitor in arthropods.
643 *Proceedings of the National Academy of Sciences*, 109(12), 4407–4412.

644 Van Leeuwen, T., Vanholme, B., Van Pottelberge, S., Van Nieuwenhuyse, P., Nauen, R., Tirry, L., &
645 Denholm, I. (2008). Mitochondrial heteroplasmy and the evolution of insecticide resistance: Non-
646 Mendelian inheritance in action. *Proceedings of the National Academy of Sciences*, 105(16), 5980–
647 5985.

648 Van Leeuwen, T., Vontas, J., Tsagkarakou, A., Dermauw, W., & Tirry, L. (2010). Acaricide resistance
649 mechanisms in the two-spotted spider mite *Tetranychus urticae* and other important Acari: A
650 review. *Insect Biochemistry and Molecular Biology*, 40(8), 563–572.

651 Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., & Speleman, F.
652 (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of
653 multiple internal control genes. *Genome Biology*, 3(7), research0034-1.

654 Wiens, B. L. (1999). When log-normal and gamma models give different results: A case study. *The*
655 *American Statistician*, 53(2), 89–93.

656 Yamamura, K., & Hino, A. (2007). Estimation of the proportion of defective units by using group
657 testing under the existence of a threshold of detection. *Communications in Statistics—Simulation*
658 *and Computation*, 36(5), 949–957.

659 Yamamura, K., Mano, J., & Shibaike, H. (2019). Optimal definition of the limit of detection (LOD) in
660 detecting genetically modified grains from heterogeneous grain lots. *Quality Technology &*
661 *Quantitative Management*, 16(1), 36–53.

662 Yelin, I., Aharony, N., Shaer-Tamar, E., Argoetti, A., Messer, E., Berenbaum, D., Shafran, E., Kuzli,
663 A., Gandali, N., & Hashimshony, T. (2020). Evaluation of COVID-19 RT-qPCR test in multi-
664 sample pools. *Clinical Infectious Diseases*, 71, 2073–2078.

665

666

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

RED-ΔΔCq dataset from Osakabe et al. (2017).

ESM 2

R source code for Experiment 1 (Figure 2), including a brief guide to the "freqpcr" package.

ESM 3

R source code for the numerical simulation (Experiment 2) and the codes for Figures 3 and after.

Appendix S1

Formularization in the case of diploidy.

Author contributions

M.S. designed research, made statistical models and R package, and analyzed data. M.O. conducted laboratory work. Both authors wrote the final manuscript.

Tables

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_{Θ}	The termination threshold of the amplification in the real-time PCR process	$X_{\Theta} > 0$	Fixed 1 in 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

Figures

Figure 1 Scheme of population allele frequency estimation based on qPCR analyses. A: Insect sampling and subsequent qPCR analysis using the restriction enzyme digestion (RED)-ΔΔCq method. B: Probability distributions of the DNA amounts of the resistant (R) and susceptible (S) alleles and their ratio in the bulk sample. C: Template DNA involved in the RED-ΔΔCq analysis and a general ΔΔCq analysis using an R-specific primer set. In either method, the frequency of X_R in a test sample is quantified as $X_R + zX_S$ ($\cong X_R$) measured on the target gene, divided by $X_R + X_S$ measured on a housekeeping gene in the sample. As the copy numbers may differ between genes, the relative content δ_T is also quantified using a control sample.

Figure 2 Relationship between the allele frequency in the sample and A: the RED-ΔΔCq measures, B: the observed frequency calculated as $(1+\eta)\text{^}(-\Delta\Delta Cq)$, showing the results of etoxazole resistance in the two-spotted spider mites. The lines are not the regression on the actual Cq measurement (shown as points), but the theoretical relationship between true frequency of the R allele and the quantity defined as A: $-\ln\big(z + Y_R(1-z)\big)/\ln(1+\eta)$ or B: $z + Y_R(1-z)$, where $Y_R = X_R/(X_R + X_S)$. Parameters are $z = 0.00156$ and $\eta = 0.971$.

Figure 3 Estimation accuracy of the population 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 corresponds to *N*, or the "ntrap" parameter, the extent to which the collected individuals (ntotal) were divided to the bulk samples. 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 4 Estimation accuracy of the population allele frequency by simple averaging of ΔΔCq measures. The frequency was underestimated than its true value (horizontal broken line in each facet) as the samples were more divided. The Cq dataset was derived from the numerical experiment of "beta distribution, all parameters unknown."

freqpcr: estimation of population allele frequency using qPCR ΔΔCq measures from bulk samples

Masaaki Sudo, Masahiro Osakabe

Figures

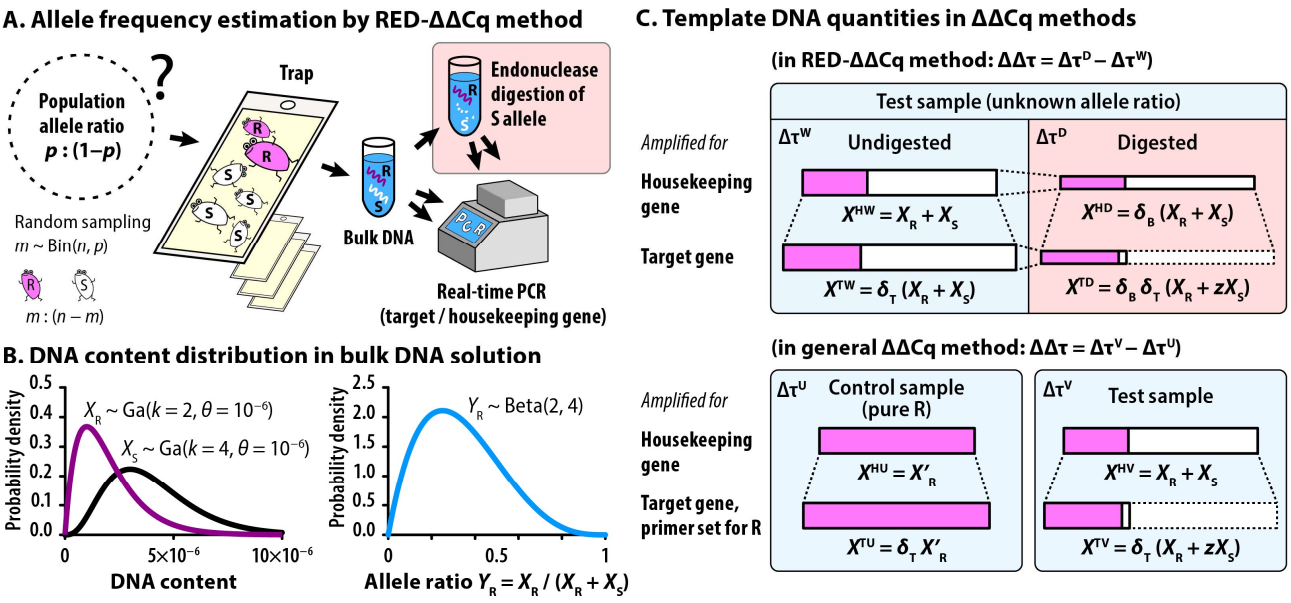

Figure 1 Scheme of population allele frequency estimation based on qPCR analyses. A: Insect sampling and subsequent qPCR analysis using the restriction enzyme digestion (RED)-ΔΔCq method. B: Probability distributions of the DNA amounts of the resistant (R) and susceptible (S) alleles and their ratio in the bulk sample. C: Template DNA involved in the RED-ΔΔCq analysis and a general ΔΔCq analysis using an R-specific primer set. In either method, the frequency of X_R in a test sample is quantified as $X_R + zX_S$ ($\cong X_R$) measured on the target gene, divided by $X_R + X_S$ measured on a housekeeping gene in the sample. As the copy numbers may differ between genes, the relative content δ_T is also quantified using a control sample.

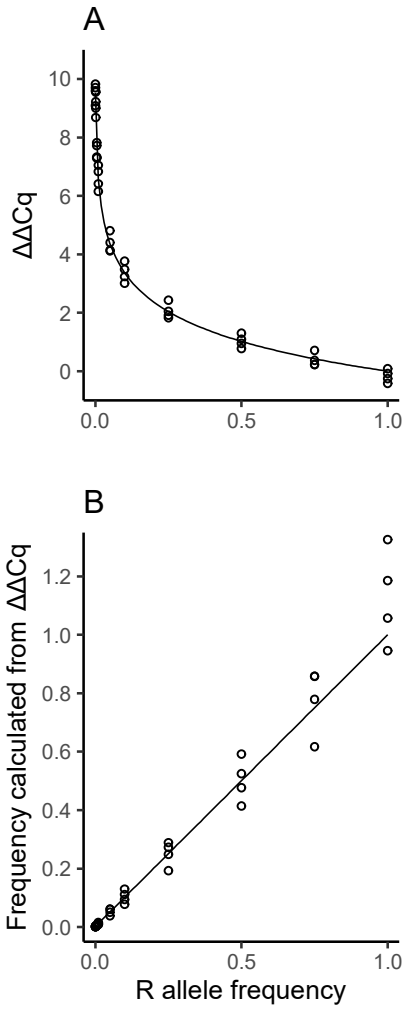

Figure 2 Relationship between the allele frequency in the sample and A: the RED-ΔΔCq measures, B: the observed frequency calculated as $(1+\eta)\text{^}(-\Delta\Delta\mathrm{Cq})$, showing the results of etoxazole resistance in the two-spotted spider mites. The lines are not the regression on the actual Cq measurement (shown as points), but the theoretical relationship between true frequency of the R allele and the quantity defined as A: $-\ln\big(z+Y_{\mathrm{R}}(1-z)\big)/\ln(1+\eta)$ or B: $z+Y_{\mathrm{R}}(1-z)$, where $Y_{\mathrm{R}}=X_{\mathrm{R}}/(X_{\mathrm{R}}+X_{\mathrm{S}})$. Parameters are $z=0.00156$ and $\eta=0.971$.

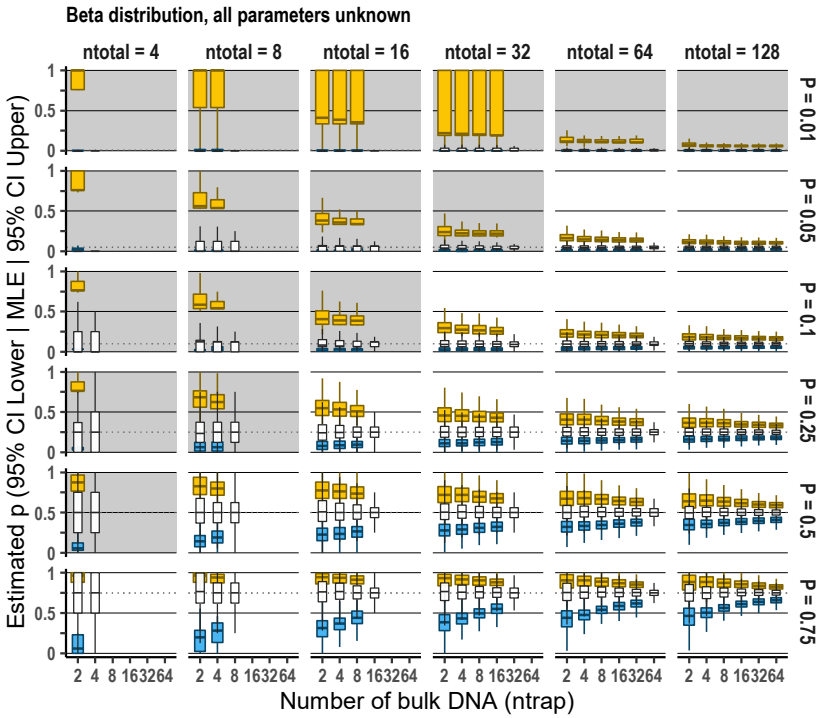

Figure 3 Estimation accuracy of the population allele frequency, *p*, with freqpcr() when the beta distribution was assumed, and all estimable parameters (P, K, targetScale, and sdMeasure) were set as unknown. The result of numerical experiments based on 1,000 dummy datasets per parameter region. The x-axes corresponds to *N*, or the "ntrap" parameter, the extent to which the collected individuals (ntotal) were divided to the bulk samples. 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*.

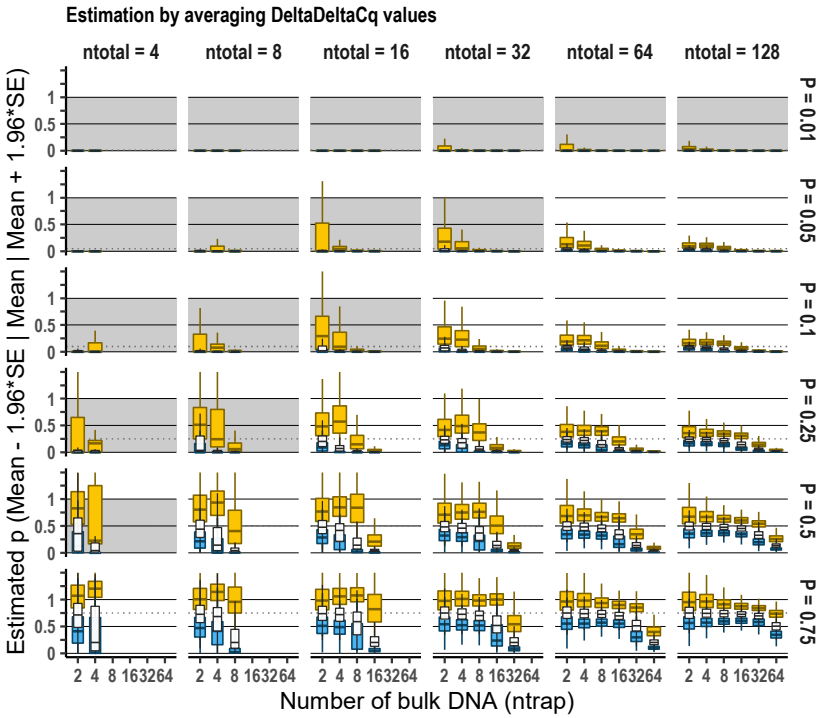

Figure 4 Estimation accuracy of the population allele frequency by simple averaging of ΔΔCq measures. The frequency was underestimated than its true value (horizontal broken line in each facet) as the samples were more divided. The Cq dataset was derived from the numerical experiment of "beta distribution, all parameters unknown."

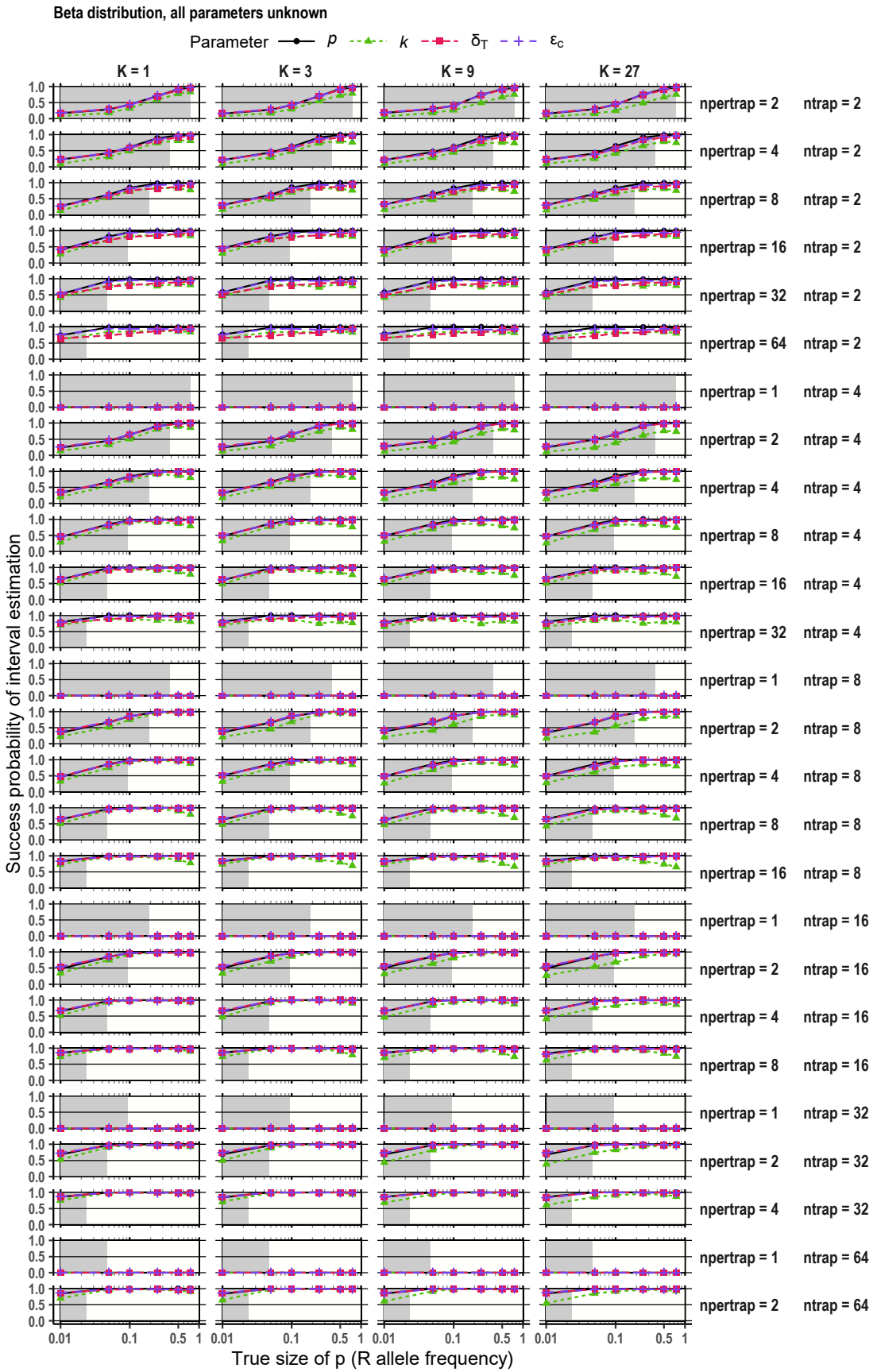

Figure S1 Probability of estimation success with freqpcr(). The beta distribution was assumed, and all estimable parameters (P, K, targetScale, and sdMeasure) were set as unknown. The shaded boxes in the background show the frequency ranges where the total sample sizes (ntotal) are smaller than $3/p$.

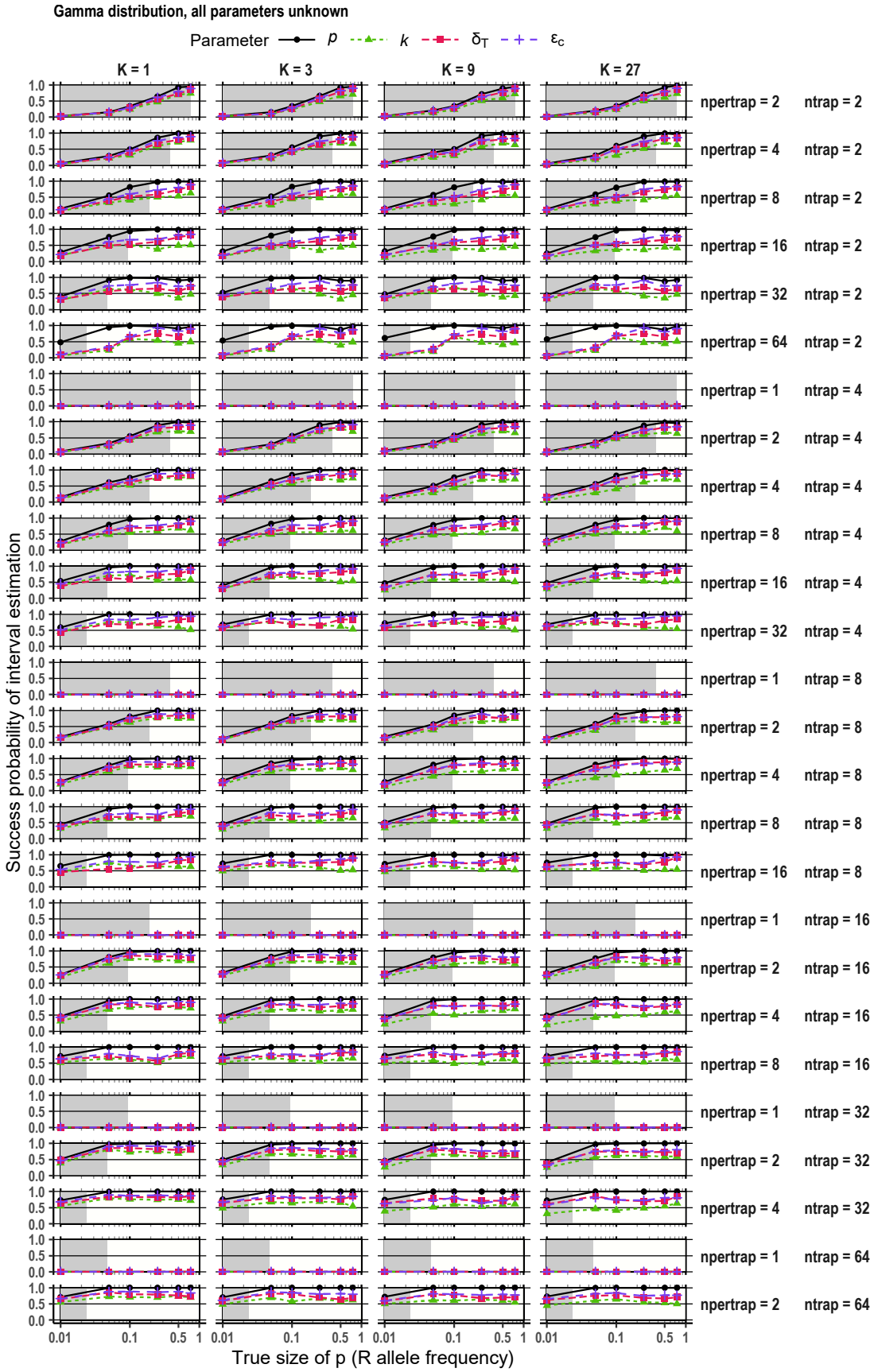

Figure S2 Probability of estimation success with freqpcr(). The gamma distributions were assumed, and all estimable parameters were set as unknown. The function often failed to calculate the CIs for *k* when npertrap (individuals in each bulk sample) were larger.

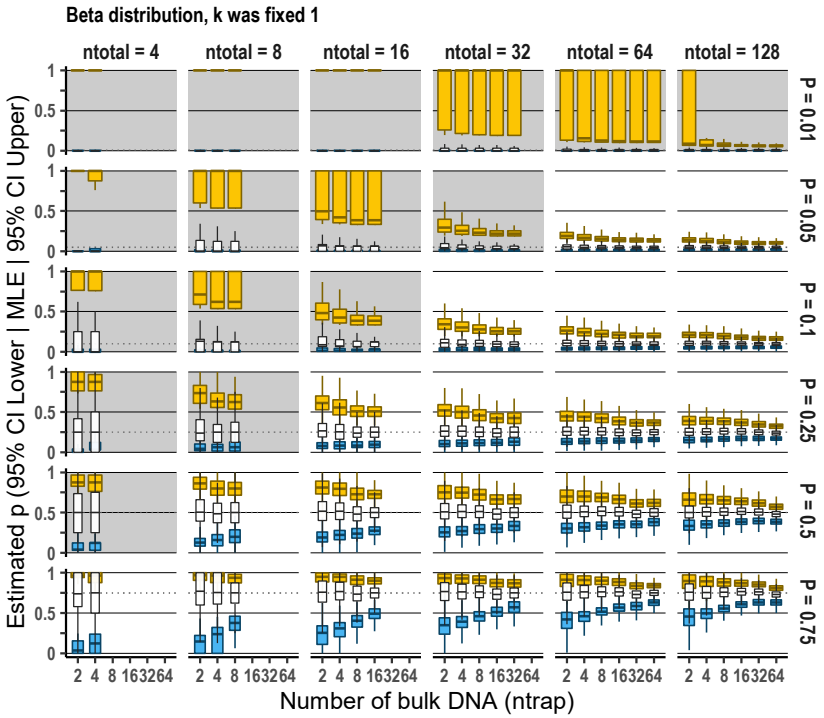

Figure S3 Estimation accuracy of the population allele frequency, *p*, with freqpcr() when the beta distribution was assumed, considering K = 1.

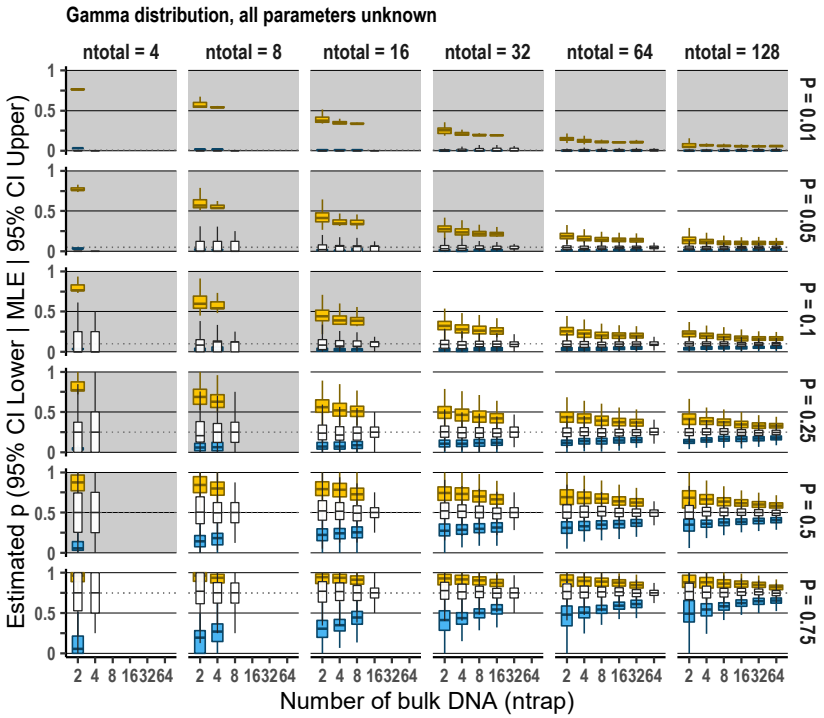

Figure S4 Estimation accuracy of *p* with freqpcr() when gamma distributions were assumed and all estimable parameters were set as unknown.

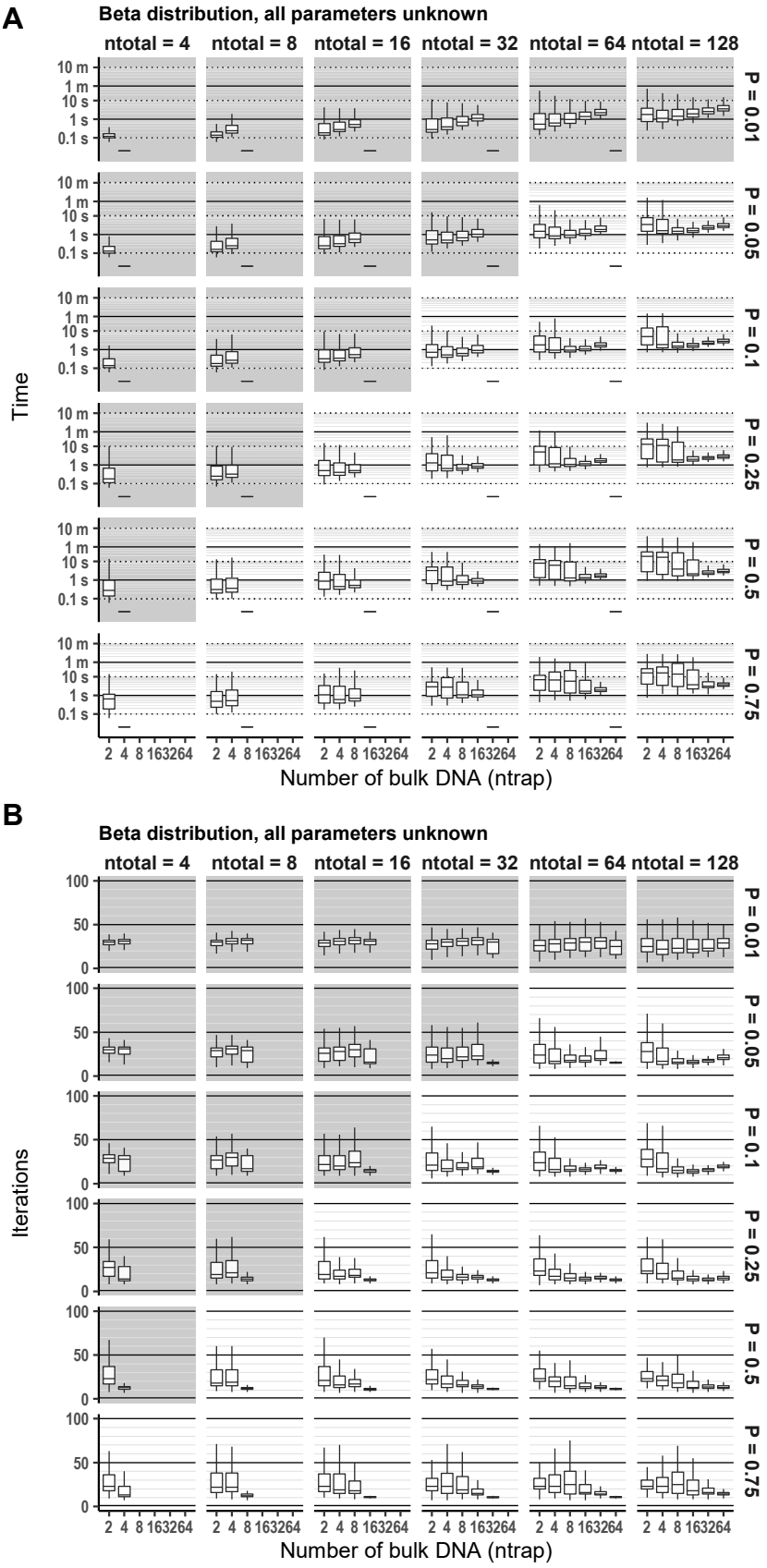

Figure S5 Calculation time (A) and number of iterations (B) until the freqpcr() function converges. The beta distribution was assumed, and all estimable parameters were set as unknown.

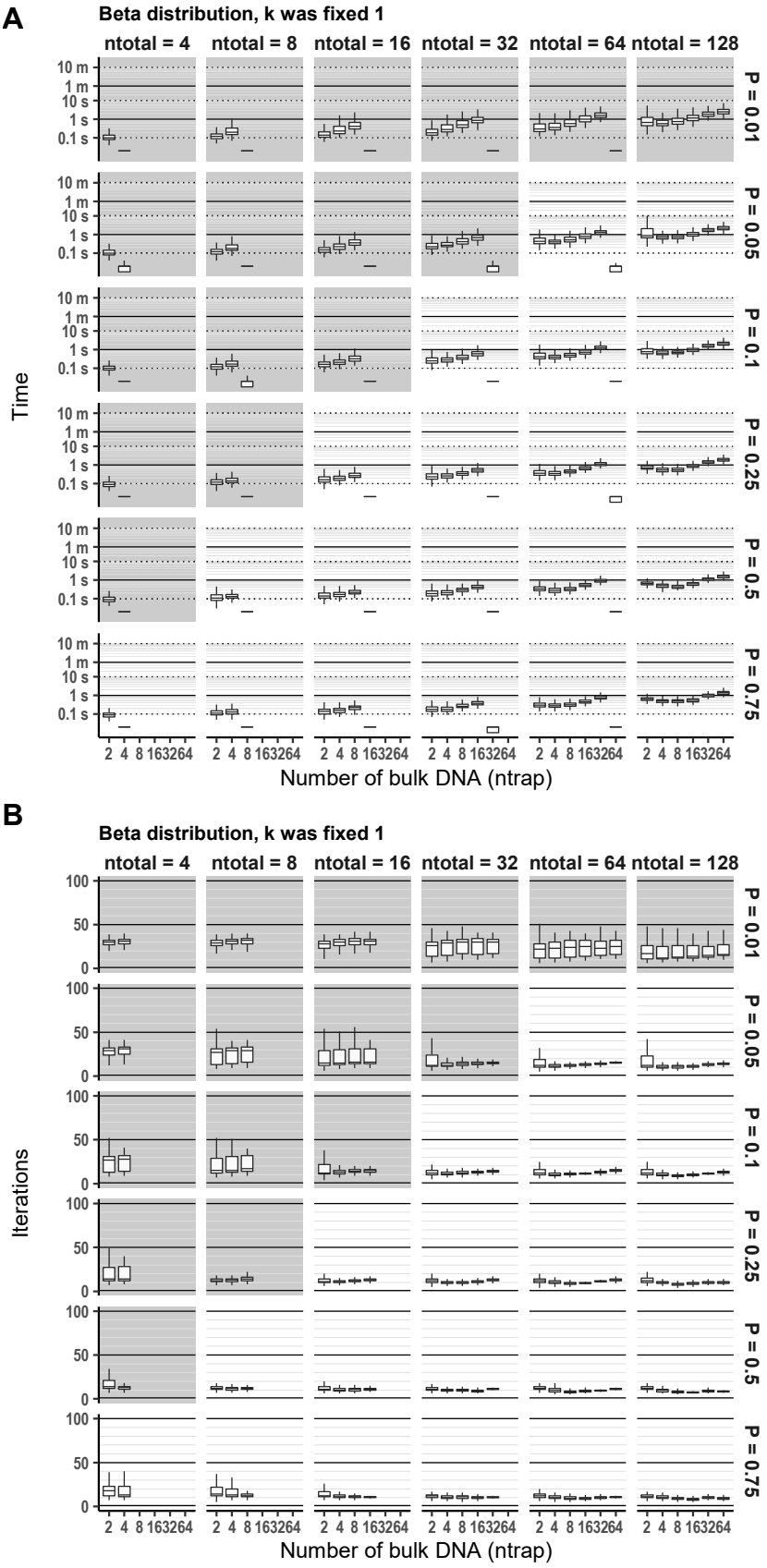

Figure S6 Calculation time (A) and number of iterations (B) until the freqpcr() function converges. The beta distribution was assumed, fixing the gamma shape parameter K = 1.

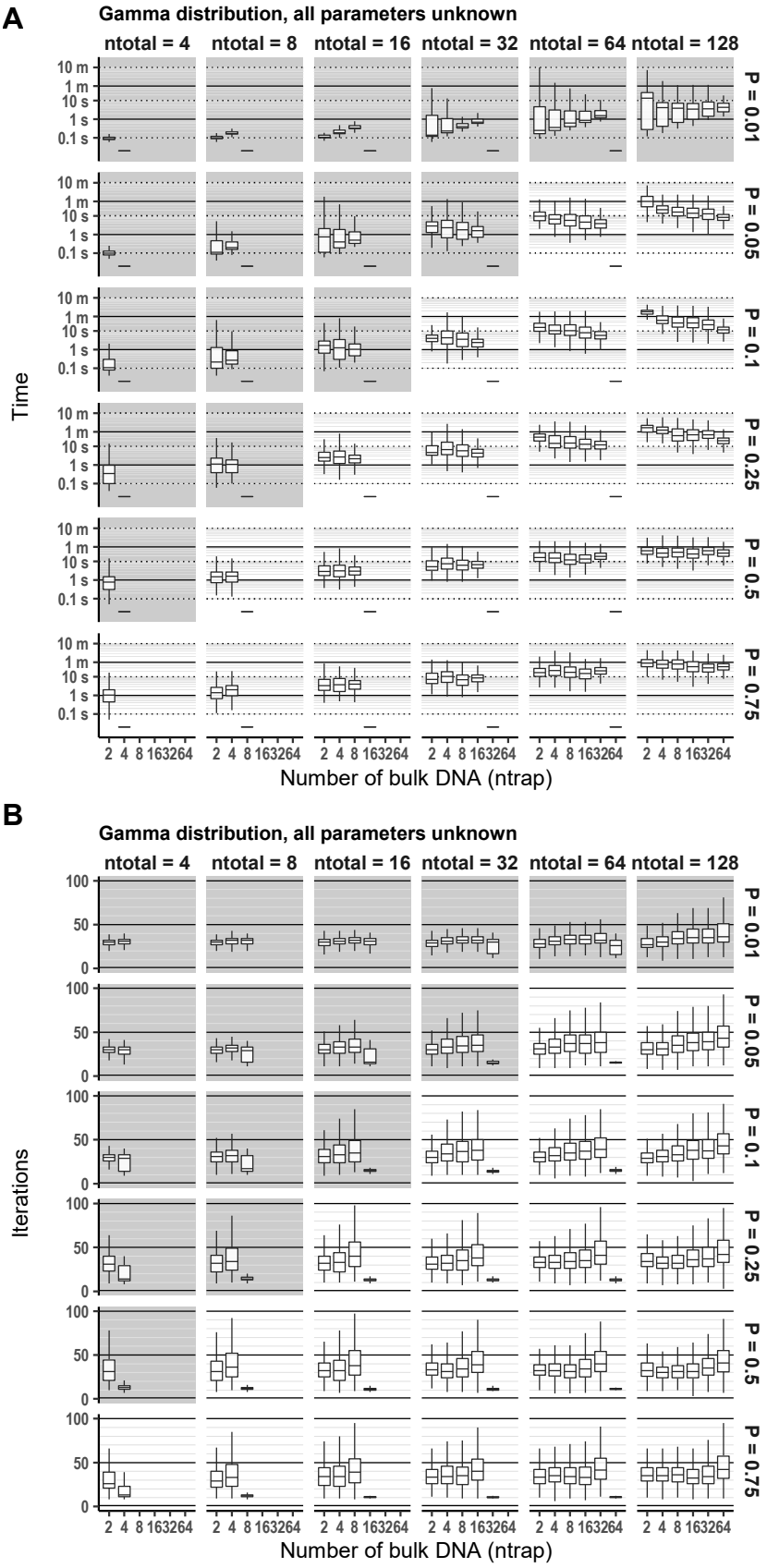

Figure S7 Calculation time (A) and number of iterations (B) until the freqpcr() function converges, assuming gamma distributions. All estimable parameters were set as unknown.

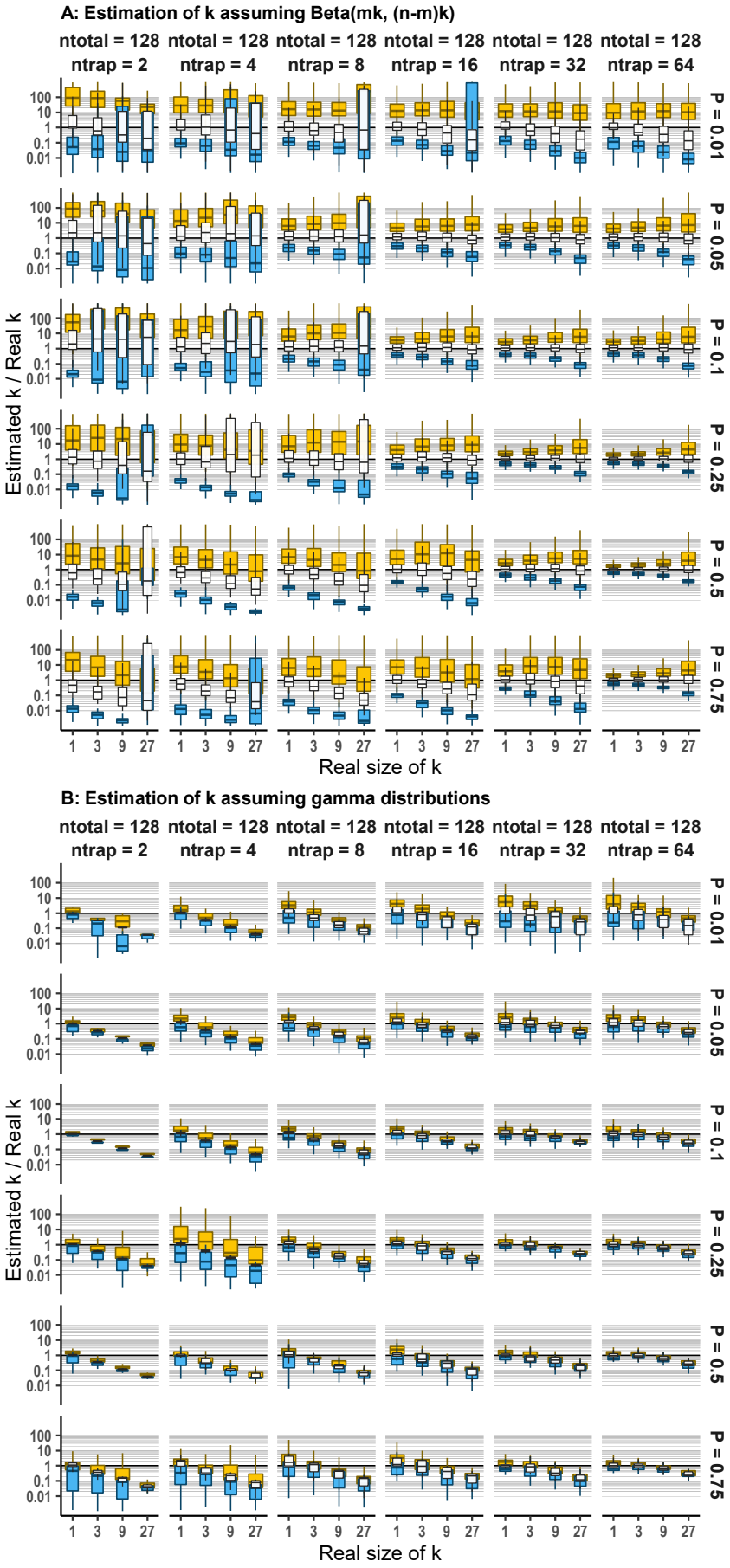

Figure S8 Estimation accuracy of *k* (the gamma shape parameter) in the simulation, showing the maximum likelihood estimate by freqpcr() divided by the actual parameter size.