J. Pestic. Sci. **46**(2), 160–167 (2021)
DOI: 10.1584/jpestics.D20-064

JOURNAL OF PESTICIDE SCIENCE

Regular Article

Estimating the proportion of resistance alleles from bulk Sanger sequencing, circumventing the variability of individual DNA

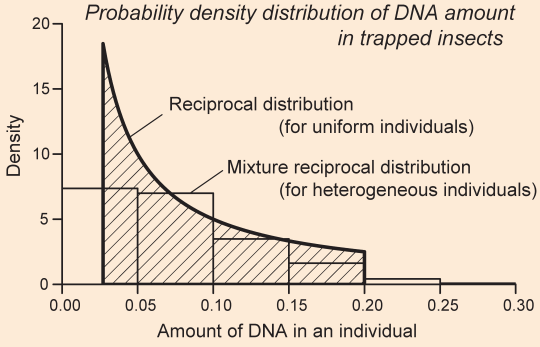
Masaaki Sudo,[1,†] Kohji Yamamura,[2,*,†] Shoji Sonoda[3] and Takehiko Yamanaka[2]

[1] *Institute of Fruit Tree and Tea Science, NARO, Kanaya Tea Research Station, 2769 Shishidoi, Kanaya, Shimada, Shizuoka 428–8501, Japan*
[2] *Institute for Agro-Environmental Sciences, NARO, 3–1–3 Kannondai, Tsukuba, Ibaraki 305–8604, Japan*
[3] *School of Agriculture, Utsunomiya University, Utsunomiya, Tochigi 321–8505, Japan*

(Received September 1, 2020; Accepted November 24, 2020)

Supplementary material

Specimens should be examined as much as possible to obtain a precise estimate of the proportion of resistance alleles in agricultural fields. Monitoring traps that use semiochemicals on sticky sheets are helpful in this regard. However, insects captured by such traps are ordinarily left in the field until collection. Owing to DNA degradation, the amount of DNA greatly varies among insects, causing serious problems in obtaining maximum likelihood estimates and confidence intervals of the proportion of the resistance alleles. We propose a statistical procedure that can circumvent this degradation issue. R scripts for the calculation are provided for readers. We also propose the utilization of a Sanger sequencer. We demonstrate these procedures using field samples of diamide-resistant strains of the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). The validity of the assumptions used in the statistical analysis is examined using the same data.



Keywords: DNA degradation, Sanger sequencer, confidence interval, maximum likelihood estimation, proportion of resistance, bulk sequencing.

Introduction

The evolution of resistance in pests against chemicals such as antibiotics or pesticides is one of the most serious threats we face in food production and human health. Continuous use of pesticides facilitates the evolution of resistance; hence, we should carefully regulate pesticide use to prevent the spread of resistance. Monitoring the emergence of resistance alleles is an essential part of establishing appropriate regulations. However, the intensive monitoring of resistance is practically very difficult in actual fields.

The preferred method for detecting resistance alleles has evolved from live bioassays to molecular-based methods. Entomologists traditionally estimated the proportion of resistance alleles by performing bioassays in which they exposed live insects to chemical compounds.[1–3] However, the preparation of sufficient live insects requires significant labor, especially if we want to avoid an inbreeding depression.[4] The development of molecular technology provided alternative methods to estimate the proportion of the resistance alleles.[5–9] Qualitative analysis using polymerase chain reaction (PCR) enabled us to detect the specified gene from dead insects. Therefore, living insects are not required if there is an appropriate primer for detection of the gene, though we would need as many dead specimens as possible for precise estimations.

There are two primary problems in handling dead insects when a precise estimate of the proportion of the resistance alleles is required: (1) sufficient insects should be collected in the field to enhance detectability and precision, and (2) these specimens should be processed efficiently by PCR to save time

* To whom correspondence should be addressed.
E-mail: yamamura@affrc.go.jp
† These authors contributed equally to this work.
Published online February 18, 2021

and costs. To solve the first problem, we can use several types of monitoring traps that have been developed for judging the necessity of control activities.[10–12] Several combinations of capturing methods (*e.g.*, sticky sheets or a water pan) and attractants (semiochemicals, ultraviolet lights, or attractive colors) are used in these traps, and sufficient insects can be collected with these systems.[13,14] To solve the second problem, we can use quantitative methods such as real-time PCR to handle many insects in a single load. Quantitative methods are usually based on the assumption that an individual has the same quantity of the target DNA. This assumption is frequently violated, especially when using monitoring traps to collect insects. Insects captured by monitoring traps are left in the field until we visit the field. Therefore, the amount of DNA measured from an insect varies greatly due to degradation of the DNA, depending on the exposure time of the insect. The confidence intervals of the proportion of resistance alleles will be biased if we do not consider the variability of the amount of DNA.

Our study has two purposes. First, we propose a statistical procedure that should be used when the amount of DNA in insects varies due to degradation in the field. The R function for the estimation is provided in the electronic appendix. Second, we propose utilization of the output from a Sanger sequencer to measure the proportion of resistance alleles in DNA. A Sanger sequencer measures the signal intensities of the dye fluorescence corresponding to each of the resistant (R) and the susceptive (S) genotypes. The existence of resistance alleles can be ascertained quickly and easily by a sequencer even if only a small amount of the resistance DNA is extracted from multiple insects. This procedure is called "bulk sequencing."[12,15] However, the signal from a sequencer is not designed for quantitative measurement, and we cannot directly quantify the relative abundance of the resistant DNA from the output of the sequencers. Therefore, a transformation method is required to estimate the actual proportion of resistance DNA from the output of a Sanger sequencer.

To visualize the second procedure, *i.e.*, the transformation method for a Sanger sequencer, we used a sampling record of diamide-resistant strains of the diamondback moth, *Plutella xylostella* Linnaeus (Lepidoptera: Plutellidae). The sample record (Excel spreadsheet) is provided in the electronic appendix. We also utilize the data to examine the validity of the assumptions adopted in the first procedure, *i.e.*, the statistical analysis to estimate the proportion of resistance alleles.

Materials and methods

1. Statistical methods

1.1. Estimation of the proportion of resistance alleles

Here we discuss how to estimate the proportion of resistance alleles from a given proportion of DNA observed from dead specimens with degraded DNA in the field. The probability distribution of the observed values has three components: (1) the distribution of the numbers of resistant alleles captured by the traps, (2) the distribution of the amount of DNA in an allele, and (3) the distribution of the proportion of resistance DNA for a given number of captured alleles. We will formulate these three distributions consecutively and then combine them to produce the full probability distribution, which can be utilized to estimate the proportion of resistance alleles in the field.

1.1.1. Number of resistant alleles in a trap

We first derive the distribution of the numbers of resistant alleles captured by the traps. Let p be the proportion of the resistant alleles in the field, m the number of resistant alleles in a sample obtained by a trap, and n the total number of the alleles in the sample. The number of susceptible alleles in a sample is given by $n-m$. We assume that the alleles are mixed well by random mating of the individuals. The distribution of the numbers of captured resistant alleles is given by a binomial distribution with the parameters n and p:

$$\text{Binomial}(m\,|\,n,p)=\frac{n!}{m!(n-m)!}p^m(1-p)^{n-m}. \qquad (1)$$

1.1.2. Amount of DNA in an allele

We use an approximation to describe the probability density function of the amount of measurable DNA per allele in a trapped individual. Let y be the amount of measurable DNA per allele, T the total duration of trapping, and t the time from trapping to collection of an individual. We can use a uniform distribution of the range from 0 to T for the time from trapping to the collection of individuals (t) if we ignore the decrease in attracting ability of a trap (*e.g.*, depletion of pheromone lures or saturation of sticky sheets) and if no information about the phenology of the insects is available. We can assume that individuals enter the trap at random in such situations. We denote the uniform distribution by $U(t|0,T)$. Trapped individuals soon die in the trap, and then DNA degradation (fragmentation) begins because of remnant endogenous DNase activity and ultraviolet solar radiation.[10] Exponential decay can generally be assumed for DNA as well as for other materials.[16,17] Therefore, the amount of measurable DNA after duration t is given by $a\exp(-\lambda t)$, where λ is the rate of decay and a is the initial amount of DNA. The duration of decay, t, follows the uniform distribution $U(t|0,T)$; hence, we obtain the probability density function of the amount of DNA by transforming $U(t|0,T)$ by the function $a\exp(-\lambda t)$. Therefore, the probability density function of the amount of DNA (denoted by y) for a given set of λ and T is given by the reciprocal distribution[18]

$$f(y\,|\,\lambda,T)=\frac{1}{\lambda T y},\ \ (a\exp(-\lambda T)\le y\le a). \qquad (2)$$

This equation indicates that the probability density of the amount of DNA (y) falls sharply to zero when y exceeds its maximum quantity a. A schematic example of the probability density is shown in the hatched area of Fig. 1.

The quantity of a in Eq. 2, which indicates the initial amount of DNA before the beginning of decay, will vary between individuals due to the variability in body size.[19] Hence, the actual probability density distribution of DNA will be given by a mixture distribution of Eq. 2, *i.e.*, a mixture reciprocal distribution,

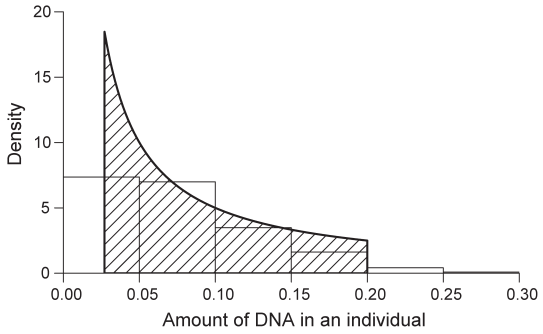


Fig. 1. Schematic illustration of the theoretical probability density of the amount of DNA in an individual insect captured by a monitoring trap. The hatched area indicates the reciprocal distribution given by Eq. 2 for $a=0.2$ and $\lambda T=2$. The histogram indicates a simulated example of the mixture reciprocal distribution where the initial amount of DNA (a) fluctuates by following a distribution with the arithmetic mean 0.2.

where the initial amount of DNA (a) varies between individuals. We do not know the distribution of the initial amount of DNA, but the mixture reciprocal distribution will be generally given by a distribution smoothly decreasing toward the right. An example of a mixture reciprocal distribution is shown in the histogram in Fig. 1. This example was generated in a simulated calculation by assuming that the initial amount of DNA follows a lognormal distribution, but the basic form of mixture reciprocal distribution will not largely depend on the underlying distributions of the initial amount of DNA.

We use another class of distribution, a gamma distribution, for approximately describing the mixture reciprocal distribution of the amount of DNA:

$$\mathrm{Gamma}(y\,|\,k,\theta)=\frac{1}{\Gamma(k)}\left(\frac{1}{\theta}\right)^{k} y^{k-1}\exp\left(-\frac{y}{\theta}\right),\ (0\le y), \quad (3)$$

where $\Gamma(\cdot)$ indicates a gamma function. The parameters k and θ are the shape and scale parameters of the gamma distribution, respectively. The mean and variance are given by $k\theta$ and $k\theta^2$, respectively. The quantity of $1/k$ is equivalent to the coefficient of variation (CV) for the distribution. The validity of this approximation will be shown using the actual field data in a later section. Note that if the trapping duration T is large compared to the mean life-time of DNA ($1/\lambda$), *i.e.*, if the quantity of λT is sufficiently large, Eq. 2 becomes identical to Eq. 3 with $k=1/(\lambda T)$ and $\theta=(\lambda T)^2$, even if no variability exists in the initial amount of DNA.

1.1.3. Proportion of resistance DNA for a given number of alleles

We next calculate the probability of obtaining a certain proportion of the resistance DNA for a given set of numbers of resistant alleles (m) and the total number of alleles in the sample (n). Let r be the observed proportion of resistance DNA. The amount of resistance DNA in a trap is given by the sum of the DNA of m alleles. The amount of DNA in an allele is approximately given as a gamma variable with the shape parameter k and the scale parameter θ, as discussed above (Eq. 3). We further assume that the distribution of the amount of DNA in an allele is approximately independent although it is not exactly true for diploid insects. Then, the total amount of resistance DNA follows a gamma distribution with the shape parameter km and the scale parameter θ because of the reproducible property of the gamma distribution. Similarly, the total amount of the susceptible DNA in a trap follows a gamma distribution with the shape parameter $k(n-m)$ and the scale parameter θ. Consequently, the distribution of the proportion of resistance DNA, which is denoted by r, is given by a beta distribution with the parameters km and $k(n-m)$ due to the relationship between a gamma distribution and a beta distribution.[20] We denote the distribution $\mathrm{Beta}(r|km, k(n-m))$ as follows:

$$\mathrm{Beta}(r\,|\,km,k(n-m))=\frac{\Gamma(kn)}{\Gamma(km)\Gamma(k(n-m))}r^{km-1}(1-r)^{k(n-m)-1}. \quad (4)$$

1.1.4. Proportion of resistance DNA in a trap

Finally, we calculate the probability of observing the proportion of resistance DNA (r) for a given total number of alleles in the sample (n). The number of resistant alleles (m) among the total alleles in the sample (n) follows a binomial distribution given by Eq. 1. The proportion of the resistance DNA for a given set of m and n follows a beta distribution given by Eq. 4. Hence, we can calculate the probability of observing the proportion (r) for a given number of the total alleles (n) by the convolution of the binomial distribution and the beta distribution for $m=1,2,\ldots,n-1$, if the proportion falls within a range of $0<r<1$. The proportion becomes $r=0$ if and only if $m=0$. The proportion becomes $r=1$ if and only if $m=n$. Thus, the probability $h(r|n,k,p)$, in which the obtained proportion of resistance DNA in a trap is r, is given by a set of discrete and continuous distributions:

$$h(r|n,k,p)=\begin{cases}\displaystyle\sum_{m=1}^{n-1}\mathrm{Beta}(r|km,k(n-m))\,\mathrm{Binomial}(m|n,p)\,\mathrm{d}r & \text{for } 0<r<1\\ \mathrm{Binomial}(0|n,p) & \text{for } r=0\\ \mathrm{Binomial}(n|n,p) & \text{for } r=1.\end{cases} \quad (5)$$

If we calculate the proportion of resistance (r) for each bulk sample, we can estimate the proportion of the resistant alleles (p) as well as the parameter k of the gamma distribution by finding a set of p and k that maximizes the sum of $\log_e(h(r|n,k,p))$ given by Eq. 5. A computer program for the estimation of parameter p and its approximate confidence intervals using the R language[21] is provided in Electronic Supplementary Material 1 (ESM1). The approximate confidence intervals are calculated using the inverse of the Hessian matrix evaluated at the last iteration of the program.

1.2. Example of estimation using the R function

The proportion of resistant individuals can be estimated by using the R function `Resist_est()` that is based on the theory described above. The R function is given in ESM1. All ESMs are available on the journal site and on the following site: ‹http://cse.naro.affrc.go.jp/yamamura/Resistance_estimation_from_bulk.html›. We test the function using data generated by simulations. We use the binomial distribution with the proportion of resistant individuals p=0.1 (Eq. 1) and the gamma distribution with the parameter k=1 (Eq. 3).

Let us first consider a case where the numbers of captured individuals in three traps are 8, 10, and 12 and the corresponding proportions of resistance DNA are 0, 0.018, and 0.005. The numbers of alleles are 2×(8, 10, 12) for a diploid species. The text file ESM1_R_function_resistance_estimation.txt, which contains the R function, should be placed in the working directory of R software. Then we can perform the estimation as follows: We first use the `source` function to enable the function. Then we create two vectors named `N` and `ObsP`; the `N` vector contains the number of individuals in each trap, and the `ObsP` vector contains the quantity of the observed proportion of resistance DNA. The vectors are passed to the R function `Resist_est()`.

```
source("ESM1_R_function_resistance
_estimation.txt")
N <- 2*c(8,10,12)
ObsP <- c(0,0.018,0.005)
(result <- Resist_est(N,ObsP))
```

The output is as follows:

```
     Estimates  Lower.95.CL  Upper.95.CL
P   0.04165607  0.009400105    0.1660438
K   0.47431017  0.051890908    4.3354443.
```

The maximum likelihood estimate of the proportion of resistance is $\hat{p}$=0.042. It should be noted that the estimate is larger than any of the three observations (0, 0.018, and 0.005). The 95% confidence interval is from 0.009 to 0.166. The estimate of the shape parameter of the gamma distribution is $\hat{k}$=0.474. If we know a reliable estimate of parameter k beforehand from other sources of data, we can use the k-value by adding an optional statement. For example, if we know beforehand that k=1.00, we can use the following script by adding option `K=1`, although the value of k seems not to largely influence the estimate of p in this case.

```
(result <- Resist_est(N,ObsP,K=1))
```

The output is as follows:

```
     Estimates  Lower 95%CL  Upper 95%CL
P   0.03597262  0.009014123    0.1327551.
```

The confidence interval of the proportion of resistance alleles may be biased if we use the classical method, which is based on the assumption of normal errors. For example, if we use the `t.test()` function of R by specifying

```
t.test(ObsP),
```

then the results are as follows:

```
95 percent confidence interval:
 -0.01541488  0.03074821
sample estimates:
 mean of x    0.007666667.
```

The lower limit of confidence intervals becomes negative (−0.015), and we cannot show the lower limit of the proportion of resistance.

We next use the data for a single trap, where the number of captured individuals is 8 and the observed proportion of resistance DNA is 0.052. The number of alleles is 16 for a diploid species. Hence, we use the following script:

```
N <- 2*c(8)
ObsP <- c(0.052)
(result <- Resist_est(N,ObsP)).
```

The output is as follows:

```
     Estimates  Lower 95%CL  Upper 95%CL
P   0.06249915  0.008728831    0.3354198
K  29.82938907  1.894776396  469.6028798.
```

The estimate of the proportion of resistance is $\hat{p}$=0.062. The 95% confidence interval is from 0.009 to 0.335. If we know beforehand that k=1.00, for example, we can add an option `K=1`. The output is as follows:

```
     Estimates  Lower 95%CL  Upper 95%CL
P   0.09228071   0.01337477    0.4325929.
```

The estimate of the proportion of resistance seems to be improved by using this option. Note that the `t.test()` function cannot yield an estimate of the confidence interval in this case because we have only one observation for the proportion of resistance DNA.

We next use the data for three traps, where the numbers of captured individuals are 4, 2, and 2 and the observed proportion of resistance DNA is 0, 0, and 0, respectively. The numbers of alleles are 8, 4, and 4 for diploid species. No resistance genes are detected in this case; hence, the function yields the simple binomial estimate as follows:

```
     Estimates  Lower 95%CL  Upper 95%CL
P            0            0    0.2059072.
```

The estimate of the proportion of resistance is $\hat{p}$=0, and the 95% confidence interval is from 0 to 0.206.

1.3. Calculation of the proportion of resistance DNA from the output of a Sanger sequencer

We next discuss the procedure to calculate the proportion of the resistance DNA (r) from the output of a Sanger sequencer. The DNA sequencer reports the peak height of the fluorescence detector corresponding to each base of the DNA. We denote the peak heights of the critical base of the resistant type (R) and the

susceptible type (S) as H_R and H_S, respectively. The sequencer is designed for qualitative analysis to determine the sequence of DNA; it is not designed for quantitative analysis. Consequently, the peak height does not necessarily indicate the amount of DNA. Therefore, considerable bias may arise if we directly correspond the peak height to the amount of DNA. A proper transformation of the peak height is required to obtain an appropriate quantity of the proportion of resistance DNA (r).

Let x_R and x_S be the DNA amount of the resistant and susceptible types, respectively. Our current purpose is to calculate the ratio, $r=x_R/(x_R+x_S)$, from the observed peak heights, H_R and H_S. In the procedure of qualitative detection of materials, the signal for detection should be clearly expressed as a binary form, such as 0 or 1. The signals should increase rapidly with increasing amount of material, while the signals should plateau if the amount of material is sufficiently large. If we want to utilize the qualitative output as a quantitative output, we should take account of the saturation characteristics of the signals. We use the following empirical form of the saturation curve, which is sufficiently flexible:

$$H_R = H_{Max}\{1-\exp[-(a_R x_R)^b]\}, \quad (6)$$

$$H_S = H_{Max}\{1-\exp[-(a_S x_S)^b]\}, \quad (7)$$

where H_{Max} is a constant that indicates the plateau of the strength of the signals; a_R and a_S indicate the sensitivity of the signals to the DNA amount of the resistant and susceptible types (x_R and x_S), respectively; and b determines the form of the saturation curve. This form of saturation-curve has frequently been used empirically. For example, Kono and Sugino[22] used it to describe the proportion of rice stems damaged by the rice stem borer, *Chilo suppressalis* (Walker) (Lepidoptera: Crambidae). Uhlig *et al.*[23] also used it for the detection probability in PCR assays.

Let $H_{A,R}$ and $H_{A,S}$ be the adjusted peak heights for the resistant and susceptible DNA, respectively, which are defined by the following equations:

$$H_{A,R} = -\log_e\left(1-\frac{H_R}{H_{Max}}\right), \quad (8)$$

$$H_{A,S} = -\log_e\left(1-\frac{H_S}{H_{Max}}\right). \quad (9)$$

By combining Eqs. 6–9, we obtain the following relation:

$$\log_e(H_{A,R})-\log_e(H_{A,S}) = b[\log_e(x_R)-\log_e(x_S)]+d, \quad (10)$$

where d is given by $d=b\log_e(a_R/a_S)$.

The true proportion of resistance DNA is defined by $r=x_R/(x_R+x_S)$. Hence, the formula inside the bracket on the right side of Eq. 10 is identical to logit(r), that is, $\log_e(r/(1-r))$. Similarly, we define the adjusted proportion of the resistant peak height by

$$r_H = H_{A,R}/(H_{A,R}+H_{A,S}). \quad (11)$$

Then the left side of Eq. 10 is identical to logit(r_H), that is, $\log_e(r_H/(1-r_H))$. Thus, Eq. 10 is simply written as

$$\mathrm{logit}(r_H) = b\,\mathrm{logit}(r)+d. \quad (12)$$

The intercept parameter d is defined by $d=b\log_e(a_R/a_S)$, where the parameters a_R and a_S indicate the sensitivity of signals to the amount of resistant and susceptible DNA types (x_R and x_S), respectively, as defined above. The sensitivity may fluctuate depending on the conditions of the samples and DNA-sequencing instruments. However, the ratio of sensitivity, a_R/a_S, should be nearly constant for each (detector) instrument; hence, we can assume that d is a constant in the following analysis.

A nonlinear least squares method is used to estimate the parameters b, d, and H_{Max} in Eq. 12. The peak height is influenced by various factors in a multiplicative manner. Consequently, the error concerning each factor also influences the peak height in a multiplicative manner. If we use a logarithmic scale, the error influences the logarithmic height in an additive manner. Therefore, the logarithmic height will follow a normal distribution with a common variance because of the central limit theorem. Consequently, it is appropriate to apply a nonlinear least squares method to logit(r_H) when estimating the parameters of Eq. 12.

A potential problem may arise in the process of estimation in that the peak height (H_R or H_S) may exceed the quantity of H_{Max} if the peak height is very high. The estimation procedure fails in such cases; hence, the peak height exceeding the quantity of H_{Max} should be replaced by $H_{Max}(1-\Delta)$ for the convenience of estimation, where Δ is a small quantity for adjustment. Therefore, we have four unknown parameters: b, d, H_{Max}, and Δ. We can calculate the proportion of the resistance DNA by using the estimated parameters as follows:

$$r = \frac{1}{1+\exp[-(\mathrm{logit}(r_H)-\hat{d})/\hat{b}]}, \quad (13)$$

where the peak height (H_R or H_S) is replaced by $\hat{H}_{Max}(1-\hat{\Delta})$ when the peak height is greater than $\hat{H}_{Max}$. An example of the estimation is provided in an Excel spread sheet in ESM2. The detailed procedure for calculating r is provided in ESM3.

2. *Laboratory methods*

We demonstrate the validity of our method by using the data for the diamondback moth, *P. xylostella*, which is a major pest of cruciferous vegetables (ESM2). A single nucleotide mutation of the ryanodine receptor gene (*RyR*) causes the amino acid mutation from glycine to glutamate at amino acid position 4946 (G4946E), which is responsible for resistance to diamide pesticides such as flubendiamide and chlorantraniliprole.[24,25] On September 6–26, 2015, one of the authors (Shoji Sonoda) collected adult males of *P. xylostella* in the cabbage field of Kagawa Prefectural Agricultural Experiment Station (in Ayagawa: 34°13′54″N, 133°56′08″E) (hereafter referred to as the Kagawa population) using sticky traps equipped with pheromone lures (Sumitomo Chemical Co., Ltd., Osaka, Japan). We extracted DNA individually from the collected insects, time of capture

varying from one to five days after installation of the traps. Genotyping for the G4946E mutation was conducted using the individually extracted DNA according to the method reported previously.[25] In all, 47 susceptible homozygotes (SS individuals having the nucleotide G) and 48 resistant homozygotes (RR having A) were obtained and used for the subsequent analysis.

PCR amplification of partial *RyR* was conducted using primers 5′-tgtaaaacgacggccagtagactggcgctaccaagtgt-3′ and 5′-cccgtatgcgtgacagact-3′. In the former primer, the M13-21 primer sequence was included in the 5′ end, as underlined. Quick Taq HS DyeMix (Toyobo Co., Ltd., Osaka, Japan) was used for the PCR amplification. The PCR conditions were 1 cycle of 2 min at 98°C, 32 cycles of 10 sec at 98°C, 15 sec at 58°C, and 15 sec at 68°C, finishing with the final extension of 68°C for 5 min. Amplified DNA fragments were sequenced directly using the M13–21 primer. The nucleotide sequencing was conducted using a dye terminator cycle sequencing kit (Applied Biosystems, Carlsbad, CA, USA) and a DNA sequencer (3130xl, Applied Biosystems). The peak heights of nucleotides corresponding to G4946E were measured from the sequence chromatogram using software (PowerPoint 2016; Microsoft Japan, Tokyo, Japan).

During the process of capillary electrophoresis conducted in a Sanger sequencer, the signal intensities of fluorescence detectors basically correspond to the amounts of the terminal bases, *i.e.*, adenine (A), cytosine (C), guanine (G), and thymine (T), in the hybrid DNA solution. Although the DNA solution contains up to four types of a single-base polymorphism (A, C, G, and T), we discuss only two genotypes, the resistant type R and the susceptible type S, since the frequencies of the other two were negligibly low.

Results

1. Estimation of the calibration parameters (*first experiment*)

In this experiment, we estimated the parameters of Eq. 13 to calculate the proportion of the resistance DNA (r). We generated artificial solution samples of DNA in which the proportions of resistance DNA were set *exactly* at predetermined quantities. We first created genuine DNA solutions for the RR and the SS types separately. We mixed a sufficient number of individuals to create the DNA solutions in which the heterogeneity in the amount of DNA among individuals disappeared. We mixed 47 and 48 individuals for the RR and SS types, respectively. An equal volume of DNA solution was used for each individual. We next created solution samples by dispensing the genuine solutions of RR and SS types in the following nine ratios: 1:15, 1:9, 1:5, 1:2, 1:1, 2:1, 5:1, 9:1, and 15:1; that is, we set $r(=x_R/(x_R+x_S))$ at 1/16, 1/10, 1/6, 1/3, 1/2, 2/3, 5/6, 9/10, and 15/16. Three replicate solution samples were prepared for each ratio.

We obtained the following estimates of parameters by using the nonlinear least squares for Eq. 12: $\hat{b}=1.28$, $\hat{d}=0.268$, $\hat{H}_{Max}=4.35$, and $\hat{\Delta}=2.71\times10^{-4}$. A sample Excel spreadsheet for the calculation is given in ESM2. The left panel of Fig. 2 indicates that the logit of the non-adjusted peak height, $\log_e(H_R)-\log_e(H_S)$, curvilinearly increases with increasing logit proportion of resistance DNA, and the fitted curve of second-order polynomial regression fits better than the linear regression. In contrast, the right panel of Fig. 2 indicates that the logit of the adjusted peak height, $\log_e(H_{A,R})-\log_e(H_{A,S})$, almost linearly increases with increasing logit proportion of resistance DNA, $\log_e(x_R)-\log_e(x_S)$, as was indicated by Eq. 10.

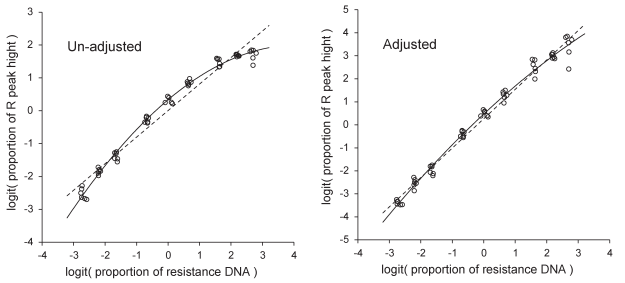


Fig. 2. Effect of calibration on the peak height from the sequencer. Left panel: the logit proportion of the unadjusted peak height of the resistance DNA, $\log_e(H_R)-\log_e(H_S)$. Right panel: the logit proportion of the adjusted peak height of the resistance DNA, $\log_e(H_{A,R})-\log_e(H_{A,S})$. The horizontal axis indicates the logit proportion of the resistance DNA, $\log_e(x_R)-\log_e(x_S)$. The dotted lines and solid lines indicate the regression curves for the first- and second-order polynomial regression for showing the effectiveness of calibration, respectively. To improve visibility, all of the points are jittered horizontally.

2. Evaluation of the validity of the assumption (*second experiment*)

We assumed a gamma distribution (Eq. 3) as an approximation of the mixture reciprocal distribution, which was derived from the assumption of an exponential decay of DNA. We conducted the second experiment to evaluate the validity of this assumption.

We used the same field samples as in the first experiment (ESM2). However, in the second experiment, the DNA mixing procedure was changed: we mixed the DNA solutions of a fixed number of RR individuals with those of a fixed number of SS individuals. The individuals were selected at random from the whole samples of the RR and the SS individuals (47 individuals for SS and 48 individuals for RR). Five combinations were used for the number of mixed individuals: RR:SS=1:15, 1:9, 1:5, 1:2, and 1:1; that is, the combination of (n,m) is (16,1), (10,1), (6,1), (3,1), and (2,1). In the case of $(n,m)=(10,1)$, for example, one DNA tube was selected at random from the 48 DNA tubes of the RR type, while nine DNA tubes were selected at random from the 47 DNA tubes of the SS type. The same amount of solution was used from these tubes to create a mixed solution. Twelve replicates were prepared for each of the five combinations.

We examined the histogram of the proportion of resistance DNA by calculating the proportion of resistance DNA using Eq. 13 with the parameters we estimated from the first experiment ($\hat{b}=1.28$, $\hat{d}=0.268$, $\hat{H}_{Max}=4.35$, and $\hat{\Delta}=2.71\times10^{-4}$). If the mean of r is not large, a beta distribution is given approximately by a gamma distribution,[26] where the shape parameter of the gamma

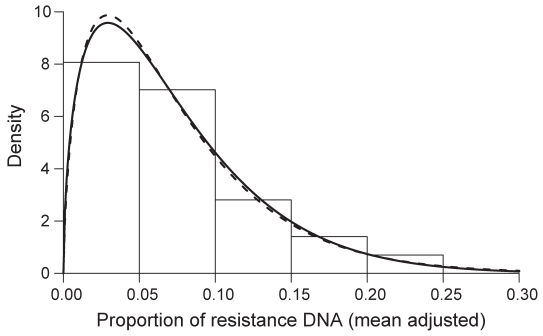


Fig. 3. Validity of the gamma distribution for describing the fluctuation in the proportion of resistance DNA. The histogram shows the empirical density distribution of the estimated proportion ($\hat{r}$) of the resistance DNA for *P. xylostella* in the mixed samples. See Yamamura *et al.*[27] for the calculation procedure. The solid curve indicates the beta distribution ($\hat{k}=1.57$), while the dotted curve indicates the gamma distribution that was fitted using the maximum likelihood method ($\hat{k}=1.66$).

distribution coincides with the "first parameter" of the beta distribution. We fixed the number of the RR individuals at $m=1$ in the second experiment; hence, the "first parameter" of the beta distribution of Eq. 4, that is km, was fixed at k in this case, *i.e.*, the shape parameter of the corresponding gamma distribution was fixed at k. The gamma distributions with a fixed shape parameter further reduce to an identical gamma distribution if we adjust the mean of the gamma distributions by changing the scale parameter of the distribution.[27] Therefore, we adjusted r so that the average of r becomes (1/16), which is the minimum quantity of the average of r in our experiment; we can adjust the average of r to any other quantity if we like. In the case of $(n,m)=(10,1)$, where the average of r is (1/10), for example, all r values were multiplied by (1/16)/(1/10) so that the average of r changes from (1/10) to (1/16). This procedure is useful in evaluating the validity of the underlying assumptions of the probability distribution in the field.[27]

The resultant histogram is shown in Fig. 3. Three observations from the experiment of $(n,m)=(2,1)$ were excluded in this histogram because the peak height was greater than the maximum quantity H_{Max} in these observations; such an observation may significantly distort the estimate of the histogram. The estimate of common k (and the 95% confidence interval) was $\hat{k}=1.57$ (1.09, 2.19) if we fitted a beta distribution and $\hat{k}=1.66$ (1.15, 2.32) if we fitted a gamma distribution. We judge the approximation by gamma distributions (Eq. 3) and the resultant beta distributions (Eq. 4) seems satisfactory based on this histogram.

Discussion

One of the difficulties of the quantitative method of genetic detection is the fluctuation of DNA amounts between samples. If the amount of DNA varies among individuals, we cannot precisely evaluate the proportion of the specified alleles in a quantitative analysis. Qualitative methods have an advantage in this respect. We can obtain the quantitative information, such as the proportion of resistance alleles, from the results of qualitative methods if we combine them with the methodology of group testing procedures.[28,29] In this paper, we explored another possibility for solving the problem of fluctuating DNA amounts. We can circumvent the problem if we know the form of the fluctuation in the amount of DNA. We considered the exponential decay of DNA. In this case, the resultant mixture reciprocal distribution is approximately described by a gamma distribution where the shape parameter k determines the fluctuation. The scale parameter θ determines the absolute quantity of DNA but does not influence the fluctuation of the proportion of specified alleles. Hence, only the estimation of k is important. The quantity of k may change depending on various conditions, such as trap types and climatic conditions. If we can predict the k value beforehand, we will be able to obtain a superior estimate of the proportion of resistance alleles from a smaller set of samples. Thus, the factors influencing the quantity of k should be examined in future studies.

We also proposed a procedure to obtain the proportion of resistance DNA from the output of a Sanger sequencer. A Sanger sequencer provides a qualitative method of detection, but it has semi-quantitative characteristics, *i.e.*, some quantitative information is included. We attempted to fully utilize the quantitative information. The saturation characteristics of the signals, which are given approximately by Eqs. 6 and 7, may be the same as those by other sequencers, but the parameters for calibration may change depending on the instrument. Hence, the parameters should be estimated for each independently.

Recently, environmental DNA has been frequently utilized as a fingerprint in examining the composition of ecological communities. Uchii *et al.*[30] used environmental DNA to examine the ratio between the Japanese native strain and a non-native strain of common carp (*Cyprinus carpio*). The exponential degradation of DNA after emission from living organisms is an important factor that influences analyses of environmental DNA as well.[31,32] An approximation by a gamma distribution, which we confirmed in this paper, will become a useful tool to improve the reliability of quantitative analyses in ecological communities.

Acknowledgements

The work was supported by a grant from the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics-based Technology for Agricultural Improvement, PRM01 to S.S. and PRM07 to T.Y. and M.S.). The authors declare that they have no conflicts of interest.

Electronic supplementary materials

The online version of this article contains supplementary material (EMS1-3), which is available at https://www.jstage.jst.go.jp/browse/jpestics/.

References

1) F. Gould, A. Anderson, A. Jones, D. Sumerford, D. G. Heckel, J. Lopez, S. Micinski, R. Leonard and M. Laster: Initial frequency of alleles for resistance to *Bacillus thuringiensis* toxins in field populations of *Heliothis virescens*. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3519–3523 (1997).
2) B. E. Tabashnik, A. L. Patin, T. J. Dennehy, Y.-B. Liu, Y. Carrière, M.

A. Sims and L. Antilla: Frequency of resistance to *Bacillus thuringiensis* in field populations of pink bollworm. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 12980–12984 (2000).
3) G. Li, D. Reisig, J. Miao, F. Gould, F. Huang and H. Feng: Frequency of Cry1F non-recessive resistance alleles in North Carolina field populations of *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *PLoS One* **11**, e0154492 (2016).
4) D. A. Andow and D. N. Alstad: F_2 Screen for rare resistance alleles. *J. Econ. Entomol.* **91**, 572–578 (1998).
5) M. Grbić, T. Van Leeuwen, R. M. Clark, S. Rombauts, P. Rouzé, V. Grbić, E. J. Osborne, W. Dermauw, P. C. Thi Ngoc, F. Ortego, P. Hernández-Crespo, I. Diaz, M. Martinez, M. Navajas, É. Sucena, S. Magalhães, L. Nagy, R. M. Pace, S. Djuranović, G. Smagghe, M. Iga, O. Christiaens, J. A. Veenstra, J. Ewer, R. M. Villalobos, J. L. Hutter, S. D. Hudson, M. Velez, S. V. Yi, J. Zeng, A. Pires-daSilva, F. Roch, M. Cazaux, M. Navarro, V. Zhurov, G. Acevedo, A. Bjelica, J. A. Fawcett, E. Bonnet, C. Martens, G. Baele, L. Wissler, A. Sanchez-Rodriguez, L. Tirry, C. Blais, K. Demeestere, S. R. Henz, T. R. Gregory, J. Mathieu, L. Verdon, L. Farinelli, J. Schmutz, E. Lindquist, R. Feyereisen and Y. Van de Peer: The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* **479**, 487–492 (2011).
6) R. ffrench-Constant: The molecular genetics of insecticide resistance. *Genetics* **194**, 807–815 (2013).
7) M. Donnelly, A. Isaacs and D. Weetman: Identification, validation, and application of molecular diagnostics for insecticide resistance in Malaria vectors. *Trends Parasitol.* **32**, 197–206 (2015).
8) L. F. Samayoa, R. A. Malvar, B. A. Olukolu, J. B. Holland and A. Butrón: 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 Biol.* **15**, 35 (2015).
9) S. Toda, K. Hirata, A. Yamamoto and A. Matsuura: Molecular diagnostics of the R81T mutation on the D-loop region of the β1 subunit of the nicotinic acetylcholine receptor gene conferring resistance to neonicotinoids in the cotton aphid, *Aphis gossypii* (Hemiptera: Aphididae). *Appl. Entomol. Zool.* **52**, 147–151 (2017).
10) R. Uesugi, N. Hinomoto and C. Goto: Estimated time frame for successful PCR analysis of diamondback moths, *Plutella xylostella* (Lepidoptera: Plutellidae), collected from sticky traps in field conditions. *Appl. Entomol. Zool.* **51**, 505–510 (2016).
11) Y. Itagaki and S. Sonoda: Seasonal proportion change of ryanodine receptor mutation (G4946E) in diamondback moth populations. *J. Pestic. Sci.* **42**, 116–118 (2017).
12) S. Sonoda, K. Inukai, S. Kitabayashi, S. Kuwazaki and A. Jouraku: Molecular evaluation of diamide resistance in diamondback moth (Lepidoptera: Yponomeutidae) populations using quantitative sequencing. *Appl. Entomol. Zool.* **52**, 353–357 (2017).
13) S. Foster and M. Harris: Behavioral manipulation methods for insect pest-management. *Annu. Rev. Entomol.* **42**, 123–146 (1997).
14) P. Witzgall, P. Kirsch and A. Cork: Sex pheromones and their impact on pest management. *J. Chem. Ecol.* **36**, 80–100 (2010).
15) D. H. Kwon, K. S. Yoon, J. P. Strycharz, J. M. Clark and S. H. Lee: Determination of permethrin resistance allele frequency of human head louse populations by quantitative sequencing. *J. Med. Entomol.* **45**, 912–920 (2008).
16) R. Lance, K. Klymus, C. Richter, X. Guan, H. Farrington, M. Carr, N. Thompson, D. Chapman and K. Baerwaldt: Experimental observations on the decay of environmental DNA from bighead and silver carps. *Manage. Biol. Invasions* **8**, 343–359 (2017).
17) S. Tsuji, M. Ushio, S. Sakurai, T. Minamoto and H. Yamanaka: Water temperature-dependent degradation of environmental DNA and its relation to bacterial abundance. *PLoS One* **12**, e0176608 (2017).
18) R. W. Hamming: "Numerical Methods for Scientists and Engineers," Dover, New York, 1973.
19) E. J. Gouws, K. J. Gaston and S. L. Chown: Intraspecific body size frequency distributions of insects. *PLoS One* **6**, e16606 (2011).
20) C. Minotani: "Handbook of Statistical Distributions," Asakura, Tokyo, 2003.
21) R Core Team: "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, 2019.
22) T. Kono and T. Sugino: On the estimation of the density of rice stems infested by the rice stem borer. *Jpn. J. Appl. Entomol. Zool.* **2**, 184–188 (1958).
23) S. Uhlig, K. Frost, B. Colson, K. Simon, D. Mäde, R. Reiting, P. Gowik and L. Grohmann: Validation of qualitative PCR methods on the basis of mathematical–statistical modelling of the probability of detection. *Accredit. Qual. Assur.* **20**, 75–83 (2015).
24) B. Troczka, C. T. Zimmer, J. Elias, C. Schorn, C. Bass, T. G. E. Davies, L. M. Field, M. S. Williamson, R. Slater and R. Nauen: Resistance to diamide insecticides in diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae) is associated with a mutation in the membrane-spanning domain of the ryanodine receptor. *Insect Biochem. Mol. Biol.* **42**, 873–880 (2012).
25) S. Sonoda and Y. Kataoka: Genotyping for the G4946E site of ryanodine receptor gene in *Plutella xylostella* (Lepidoptera: Yponomeutidae) considering gene duplication. *Appl. Entomol. Zool.* **51**, 195–204 (2016).
26) K. Yamamura and T. Sugimoto: Estimation of the pest prevention ability of the import plant quarantine in Japan. *Biometrics* **51**, 482–490 (1995).
27) K. Yamamura, S. Fujimura, T. Ota, T. Ishikawa, T. Saito, Y. Arai and T. Shinano: A statistical model for estimating the radiocesium transfer factor from soil to brown rice using the soil exchangeable potassium content. *J. Environ. Radioact.* **195**, 114–125 (2018).
28) K. Yamamura and A. Hino: Estimation of the proportion of defective units by using group testing under the existence of a threshold of detection. *Commun. Stat. Simul. Comput.* **36**, 949–957 (2007).
29) J. Mano, Y. Yanaka, Y. Ikezu, M. Onishi, S. Futo, Y. Minegishi, K. Ninomiya, Y. Yotsuyanagi, F. Spiegelhalter, H. Akiyama, R. Teshima, A. Hino, S. Naito, T. Koiwa, R. Takabatake, S. Furui and K. Kitta: Practicable group testing method to evaluate weight/weight GMO content in maize grains. *J. Agric. Food Chem.* **59**, 6856–6863 (2011).
30) K. Uchii, H. Doi and T. Minamoto: A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. *Mol. Ecol. Resour.* **16**, 415–422 (2016).
31) P. F. Thomsen, J. Kielgast, L. L. Iversen, P. R. Møller, M. Rasmussen and E. Willerslev: Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One* **7**, e41732 (2012).
32) K. M. Strickler, A. K. Fremier and C. S. Goldberg: Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biol. Conserv.* **183**, 85–92 (2015).