Title:

Estimation of concentration ratio of indicator to pathogen-related gene in environmental water based on left-censored data

Short title:

Estimation of indicator/pathogen concentration ratio based on left-censored data

Author names and affiliations:

Tsuyoshi Kato[a], Ayano Kobayashi[b], Toshihiro Ito[b], Takayuki Miura[b], Satoshi Ishii[b], Satoshi Okabe[b] and Daisuke Sano[b]*

[a]Department of Computer Science, Graduate School of Engineering, Gunma University, Tenjinmachi 1-5-1, Kiryu, Gunma 376-8515, Japan

[b]Division of Environmental Engineering, Faculty of Engineering, Hokkaido University, North 13, West 8, Kita-ku, Sapporo, Hokkaido 060-8628, Japan

***Corresponding Author. Address:**

Division of Environmental Engineering, Faculty of Engineering, Hokkaido University, North 13, West 8, Kita-ku, Sapporo, Hokkaido 060-8628, Japan. Telephone/Fax: +81-11-706-7597. E-mail: dsano@eng.hokudai.ac.jp

Abstract

The indicator/pathogen ratio is frequently used in quantitative microbial risk assessment for calculating the pathogen concentration from the concentration of fecal indicator in water. However, the concentration of a pathogen in water is often below the value of the analytical quantification limit, which makes it difficult to calculate the indicator/pathogen concentration ratio. In the present study, we constructed a stochastic model for estimating the ratio between a fecal indicator and a pathogen based on left-censored data, which include a substantially high number of non-detects. River water samples were taken for 16 months at 6 points in a river watershed, and conventional fecal indicators (total coliforms and general *Escherichia coli*), genetic markers (*Bacteroides* spp.), and virulence genes (*eaeA* of enteropathogenic *E. coli* and *ciaB* of *Campylobacter jejuni*) were quantified. The quantification of general *E. coli* failed to predict the presence of the virulence gene from enteropathogenic *E. coli*, different from what happened with genetic markers (Total Bac and Human Bac). A Bayesian model that was adapted to left-censored data with a varying analytical quantification limit was applied to the quantitative data, and the posterior predictive distributions of the concentration ratio were predicted. When the sample size was 144, simulations conducted in this study suggested that 39 detects was enough to accurately estimate the distribution of the concentration ratio, when combined with a dataset with a positive rate higher than 99%. To evaluate the level of accuracy in the estimation, it is desirable to perform a simulation using an artificially generated left-censored dataset that have the identical number of non-detects with actual data.

Keywords: analytical quantification limit, *Bacteroides*, Bayesian estimation, pathogens, indicator microorganisms, left-censored data.

Introduction

Fecal indicator microorganisms, such as coliforms, fecal coliforms, and *Escherichia coli*, are used for determining the sanitary quality of surface, recreational, and shellfish growing waters (Scott et al., 2002; Setiyawan et al., 2014). The concentration of fecal indicator microorganisms is often used in quantitative microbial risk assessment (QMRA), for estimating the pathogen concentration based on the indicator/pathogen concentration ratio (Labite et al., 2010; Itoh, 2013). This indicator/pathogen concentration ratio is acquired based on quantitative data of a field investigation, in which a pathogen and a fecal indicator are simultaneously quantified from an identical sample (Machdar et al., 2013; Silverman et al., 2013; Lalancette et al., 2014). However, pathogen quantitative data commonly include a substantially high number of non-detects, in which the pathogen concentration falls below the quantification limit (Wu et al., 2011; Kato et al., 2013). This kind of dataset is called a left-censored dataset, which does not allow us to calculate the indicator/pathogen concentration ratio at each investigation event.

Substitution of the non-detect data with specific values such as the limit of quantification, the half value of quantification limit, or zero has been used as a classical approach for dealing with non-detects, but the substitution gives inaccurate estimation of distribution parameters when the distribution of concentration is predicted (Gilliom & Helsel, 1986; Helsel, 2006). Alternatively, the Bayesian approach adapted for left-censored data was proposed, and applied to actual left-censored datasets, such as pesticide residue concentrations in food (Paulo et al., 2005). To study the density of enteric viruses in wastewater, such as those in the genus of *Enterovirus, Heptovirus, Rotavirus* and *Norovirus* (Bosch et al., 2008), we previously applied the Bayesian model proposed by Paulo et al. (2005) with a slight modification, in which the occurrence of the real zero of virus density is not assumed (Kato et al., 2013). In this Kato et al (2013)' model, virus density is assumed to follow a lognormal distribution, which is one of the probabilistic distributions previously modeled for enteric virus density in water (Tanaka et al., 1998).

In this study, we employed the extended Kato et al (2013)' model to estimate the distribution of the indicator/pathogen concentration ratio, with a further modification of the entry of quantification limit value. The previous model (Kato et al., 2013) requires entering an identical quantification limit value within a dataset. However, the quantification limit values change over time due to changes in methods, protocols, and instrument precision even within a single laboratory (Helsel et al., 2006). Those facts motivated us to develop a new Bayesian model to analyze dataset with a varying quantification limit. Let us refer to the Kato et al (2013)' model as the common limit model hereinafter. Datasets of concentrations of pathogens and indicator microorganisms were acquired from a watershed, and posterior predictive distributions of these concentrations were estimated with the new Bayesian model for varied quantification limit values. In order to ensure the accuracy of the prediction, 100 paired datasets were artificially generated, in which the simulated data were assigned to detects and non-detects by setting values of the limit of quantification to obtain the number of detects that was identical to the actual data. Then, the new Bayesian model for varied quantification limit values was applied to the simulated and censored data. The estimated mean and standard deviation were compared with true values by calculating root mean square deviation (RMSD), and the influence of the sample size and positive rate value on the estimation accuracy of the posterior predictive distribution was discussed. Furthermore, a numerical procedure is employed to obtain the distribution of the concentration log-ratio between a fecal indicator and a pathogen by integrating the posterior predictive distribution of the fecal indicator concentration with that of pathogen concentration. The accuracy of the distribution estimation of the fold change was evaluated by Kullback-Leibler (KL) divergence.

Methods

Water samples and measurement of water quality parameter.

River water samples (10L of surface water) were collected from Toyohira River (Site 1), Kamogamo River (Site 2), Nopporo River (Site 3), Atsubetsu River (Site 4), and Motsukisamu River (Site 5) in Sapporo City, and Atsubetsu River (Site 6) in Ebetsu City, Hokkaido, Japan (latitude-longitude locations are listed in Table S1, supplementary data). River water samples were collected about twice a month from January 2012 to April 2013. Total sample number was 144. No major fecal contamination sources were located near Site 1. On the other hand, effluents from wastewater treatment plants were discharged in proximity to Sites 3 and 5. Wild waterfowl such as wild ducks were observed in Sites 2 and 5. Domestic stock farms were located near Sites 4 and 6, so contamination by animal feces is expected in these sites. Total coliforms and general *E. coli* were measured for each water sample according to the standard method using defined substrate (APHA et al., 2005).

Recovery of bacterial cells from water samples.

To monitor the DNA loss during the bacterial cell recovery and DNA extraction processes, *E. coli* MG1655 Δlac::kan was used as the sample process control (SPC) for genetic markers and pathogenic bacteria (Kobayashi et al., 2013b). One hundred microliters of *E. coli* MG1655 Δlac::kan were added in 5L of river water before the recovery of bacterial cells. Bacterial cells in 5L of river water were collected by pressure filtration with a 0.22-μm-pore-size polyethersulfone membrane filter (Millipore). Bacterial cells on the membrane filter were eluted by soaking in 30 ml of sterile PBS with gelatin buffer (NaH_2PO_4: 0.58g, Na_2HPO_4: 2.5g, NaCl: 8.5g and Gelatin: 0.1g per litter) and vigorously shaken by a vortex mixer (Ishii et al.,2014b). Suspended cells in the PBS with gelatin buffer were collected by centrifugation at 10,000 ×*g* for 15 min at 4˚C, and the pellet was re-suspended in 0.8 ml of distilled MilliQ water.

DNA extraction and quantitative PCR assays.

Total DNA was extracted from bacteria cell suspensions (200 μL) obtained from water samples by

using the PowerSoil DNA Isolation Kit (MoBio Laboratories, Carlsbad, CA, USA). The concentrations of total and human-specific genetic markers (Total Bac and Human Bac, respectively), and genes of pathogenic bacteria (*ciaB* for *Campylobacter jejuni* and *eaeA* for enteropathogenic *E. coli*) were quantified using quantitative PCR (qPCR) methods previously developed (Ishii et al., 2013; Okabe et al., 2007). *E. coli* MG1655 Δlac::kan gene was quantified using a qPCR method, which does not amplify indigenous bacterial genes (Kobayashi et al. 2013b). The virulence gene *ciaB* encodes a 73-kDa secreted protein (CiaB), which enhances the internalization of *C. jejuni* into epithelial cells (Konkel et al., 1999). The virulence gene *eaeA* encodes a 94-97 kDa outer membrane protein that mediates adherence of enteropathogenic *E. coli* to epithelial cells (Frankel et al., 1995). Levels of PCR inhibition were evaluated by the addition of internal amplification control (IAC) to the sample DNA prior to qPCR. We used Chicken-Bac plasmid (Kobayashi et al., 2013a) as the IAC in this study. Amplification efficiencies of IAC were calculated as quantitative values of IAC in environmental DNA samples divided by the quantitative value of IAC in a pure plasmid solution. All primers and probes used in this study are shown in Table S2 (Supplementary data).

The qPCR assays were performed using SYBR Green chemistry to quantify the Total Bac, *E. coli* MG1655 Δlac::kan, and Chicken-Bac (IAC). In SYBR Green assays, each PCR mixture (25μl) was composed of 1x SYBR Premix Ex Taq II (Takara Bio, Otsu, Japan), 1x ROX Reference Dye (Takara Bio, Otsu, Japan), 400 nM each of forward and reverse primers, and 2 μl of template DNA. TaqMan qPCR assay was also performed to quantify Human-Bac, in which each PCR mixture (25μl) was composed of 1x Premix Ex Taq (Takara Bio, Otsu, Japan), 200 nM each of forward and reverse primers, 200 nM of fluorogenic probe, and 2.0 μl of template DNA. PCR reactions using SYBR Premix Ex Taq II (Takara Bio, Otsu, Japan) and Premix Ex Taq (Takara Bio, Otsu, Japan) were performed in MicroAmp Optical 96-well reaction plates with Applied Biosystems 7500 Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). The reaction was carried out by

heating at 95˚C for 30 sec, followed by 40 cycles of denaturation at 95˚C for 5 sec and annealing and extension at 60˚C for 34 sec. Following the amplification step, melting curve analysis was performed for SYBR Green assays to confirm that no unexpected PCR products had been obtained.

The TaqMan qPCR assay was also applied to the quantification of *ciaB* of *C. jejuni* and *eaeA* of enteropathogenic *E. coli*, in which each PCR mixture (20 μl) was composed of 1x FastStartTaqMan Probe Master (Roche, Deutschland), 900 nM each of forward and reverse primers, 250 nM of fluorogenic probe, and 2.0 μl of template DNA.PCR reactions using FastStart TaqMan Probe Master (Roche, Deutschland) were performed in MicroAmp Optical 96-well reaction plates with the ABI PRISM 7000 sequence detection system (Applied Biosystems, Foster City, CA, USA). The reaction was carried out by heating at 95˚C for 10 min, followed by 40 cycles of denaturation at 95˚C for 15 sec and annealing at 60˚C for 1 min.

A DNA template to generate standard curves for the qPCR assays was prepared using recombinant pCR 2.1-TOPO vector plasmids inserted with target sequences as described previously (Ishii et al., 2013; Kobayashi et al., 2013a; Kobayashi et al., 2013b). The standard plasmids for the quantification of these targets were prepared using the TA cloning system. For the standard plasmid for the quantification of *E. coli* MG1655 Δlac::kan, pUC19 vector carrying the PCR amplicon generated from *E. coli* MG1655 Δlac::kan with the primer set Kan-res-F and DS-Kan-R was used. The ligated products were transformed into *E. coli* TOP10 competent cells (Invitrogen, Carlsbad, CA, USA). Plasmids were extracted and purified from *E. coli* cells using QIAprep Spin Miniprep Kit (QIAGEN, Hilden, Germany). The concentrations of plasmid DNA were adjusted from 10^{-1} to 10^{-8} ng per μL and used to generate standard curves. Standard curves were generated by linear regression analysis between threshold cycles (C_T) and the concentration of the plasmid DNA using Applied Biosystems 7500 Real-Time PCR System software version 2.0.4 (Applied Biosystems, Foster City, CA, USA). The quantification limit was defined as the lowest concentration of plasmid

DNA that was amplified within the linear range of the standard curve.

Statistical methods.

The normality of logarithmic concentrations of detected bacteria was determined by a chi-square goodness-of-fit test at a significance level of 0.01. A Pearson product-moment correlation coefficient at a significance level of 0.01 (two-tailed test) was calculated between datasets that were log-normally distributed. The p-values in the chi-square goodness-of-fit test and Pearson product-moment correlation coefficient were calculated by a chi-square distribution and t-distribution, respectively. All statistical analysis was done using the Microsoft Excel program version 2012 (Microsoft corporation, SSRI, Tokyo).

Estimation of the posterior predictive distribution of indicators and virulence genes and the concentration ratio of an indicator to a virulence gene.

In this study, the common limit model (Kato et al., 2013) was extended so that different quantification limits for different observed concentrations were expressed mathematically. Suppose we are trying to measure concentrations of a target organism n times, and each concentration $x_i (i = 1, \cdots, n)$ is non-negative and observed only when the concentration exceeds a quantification limit 10^{θ_i}. It is worthy to note that, although the common limit model can express only the case of $\theta_1 = \cdots = \theta_n$, the new model allows a different quantification limit for each sample. The dataset is denoted by n tuples $(x_1, y_1), \ldots, (x_n, y_n) \in \boldsymbol{R} \times \{1,0\}$, where y_i indicates whether the concentration is detected. If $x_i > 10^{\theta_i}$, $y_i = 1$ is given; $y_i = 0$ otherwise. The value of x_i is unknown if $y_i = 0$.

If n concentrations are assumed to follow according to the log-normal distribution with a mean parameter μ and precision parameter β, the probabilistic density of a detected concentration x_i is represented by the truncated log-normal (TLN)

distribution $\mathrm{TLN}(x;\mu,\beta^{-1},\theta) \coloneqq \frac{1}{Z(\mu,\beta^{-1},\theta)x}\exp\left(-\frac{\beta}{2}(\mu-\log_{10}x)^2\right)$ where $Z(\mu,\beta^{-1},\theta) \coloneqq \sqrt{2\pi}\ln(10)\cdot\left(1-\varphi\left(\sqrt{\beta}(\theta-\mu)\right)\right)\beta^{-1}$ and φ is the cumulative density function of the standard normal distribution. Therein, the notation := has been used to denote a definition. Note that every sample is assumed to be drawn from a probabilistic density function with an exactly equal parameter θ in the common limit model (Kato et al., 2013), whereas the new model permits different quantification limits for different samples leading to probabilistic density functions of different shapes.

Derived from the fact that the normal random variable drawn from $N(\mu,\beta^{-1})$ falls short of the quantification limit $\theta_i$with the probability$\sqrt{\beta}(\theta_i-\mu)$, it is possible to express the probability mass function of y_i as$\left(\varphi\left(\sqrt{\beta}(\theta-\mu)\right)\right)^{1-y_i}\left(1-\varphi\left(\sqrt{\beta}(\theta-\mu)\right)\right)^{y_i}$. Therein, the indicator variabley_i is treated as a Bernoulli variable where, in the Bernoulli trial, one side of the unfair coin, corresponding to $y_i=0$, appears with probability $\left(\varphi\left(\sqrt{\beta}(\theta-\mu)\right)\right)$ and the other side $y_i=1$ appears with probability $\left(1-\varphi\left(\sqrt{\beta}(\theta-\mu)\right)\right)$. Instead of the Bernoulli distribution, the common limit model uses the binomial distribution (e.g. Cohen (1959)), although it is impossible to employ the binomial distribution in the setting of the different quantification limits assumed in this work.

To infer the values of model parameters, Bayesian inference is adopted. So far many works have employed the maximum likelihood estimation. This estimation method is useful if a large sample is available, although, if not, the model parameters often over-fit to the sample. To avoid over-fitting, Bayesian inference is opted in this study. For Bayesian analysis, a definition of the likelihood function and the prior distribution is required.

Based on the modeling described above, the following likelihood function of two model parameters, μand β, is employed:

$p(X|\mu,\beta)=\prod_{i=1}^{n}\left(\varphi\left(\sqrt{\beta}(\theta_i-\mu)\right)\right)^{1-y_i}\left(\left(1-\varphi\left(\sqrt{\beta}(\theta_i-\mu)\right)\right)\mathrm{TLN}(x_i;\mu,\beta^{-1},\theta_i)\right)^{y_i}$. The prior distributions, $\mu\sim N(0,100)$ and $\beta\sim Gam(0.01,0.01)$, are employed, following the original common limit model (Kato et al., 2013), where $N(m,v)$ and $Gam(a,b)$, respectively, denote the normal distribution with mean m and variance v and the Gamma distribution with shape parameter a and rate parameter b.

Applying the Bayesian inference technique to the probabilistic model described above for the concentration datasets of indicators and pathogens, the predictive posterior distributions $p_{pred}(x|X_{ind})$ and $p_{pred}(x|X_{path})$, where X_{ind} and X_{path} are the datasets of indicators and pathogens, respectively, are estimated. Then, the concentration log-ratio can be obtained as the probabilistic distribution of the difference of the two random variables. The details are referred to in Kato et al. (2013).

The Bayesian algorithm is summarized as follows.

1. The posterior parameter distribution $p(\mu,\beta|X_{ind})$ of an indicators dataset, which is proportional to the product of the likelihood function $p(X_{ind}|\mu,\beta)$ and the prior parameter distribution $p(\mu,\beta)$, is computed.
2. The predictive posterior distribution $p_{pred}(x|X_{ind})$ of an unseen log-transformed concentration x is computed by integrating out the model parameters from the product of the posterior parameter distribution and the model distribution.
3. Similarly, the posterior parameter distribution $p(\mu,\beta|X_{path})$ is estimated from a pathogens dataset, and its predictive posterior distribution $p_{pred}(x|X_{path})$.
4. The probabilistic distribution of the concentration log-ratio is obtained from two distributions

$p_{pred}(x|X_{ind})$ and $p_{pred}(x|X_{path})$.

Accuracy evaluation of the extended Bayesian estimation.

To test the accuracy of Bayesian estimation, 100 left-censored datasets were generated for each indicator or pathogen-related gene, and each generated dataset was applied to the extended Bayesian model to estimate distributional parameters. The generation process of a left-censored dataset with 144 total samples including n_v detects is composed of two steps. In the first step, a dataset with 144 total samples, which is equal to that of the actual dataset (Table S3, supplementary data), is generated. The model parameters $(\hat{\mu}, \hat{\beta})$ estimated by the maximum a posteriori (MAP) from the posterior predictive distribution (Table S3, supplementary data) were regarded as true values, and used to generate 144 data points $x_1, \cdots, x_n$ from the log-normal $TLN(x; \hat{\mu}, \hat{\beta}^{-1}, -\infty)$. In the second step, a detection limit value was chosen randomly from the observed detection limit values in the corresponding actual dataset, and data points below the assigned detection limit value were erased to make the dataset left-censored. These two steps were repeated until the dataset included the target number of detects, n_v. This process was repeated 100 times, which gave 100 left-censored datasets, in which all datasets including the same number of detects, n_v. The generation of 100 left-censored datasets were conducted for each number of detects in the actual datasets (Table S3, supplementary data). Finally, the dataset was applied to the extended Bayesian model to estimate posterior distributions of distributional parameters, which were used to obtain the posterior predictive distributions of log-concentration of microbes, and log-concentration ratio between an indicator and a pathogen-related gene. The estimated distributions of log-concentration ratio based on the generated left-censored datasets were compared with true distributions by calculating KL divergence (Kato et al., 2013)

Results

Characterization of quantitative data.

The log-normality of quantitative data was first tested because we assumed a log-normal distribution of the concentration of microbes in the Bayesian estimation. All 144 samples were positive for total coliforms and Total Bac (Table S3, supplementary data). Normal probability plots of these microbes looked straight (Fig. S1(a) and S1(c)), but a chi-square test at the significance level of 1% showed that the logarithmic quantitative data of total coliforms were not normally distributed (p-value was 3.16×10^{-3}, Table S3). The normality of logarithmic concentration values of Total Bac was not rejected by the chi-square test at a significance level of 0.01, with a p-value of 0.03 (Table S3). Only one out of 144 samples was negative in the quantitative data of general *E. coli* and Human Bac (Table S3). The normal probability plot of the 143 logarithmic concentration values of general *E. coli* and Human Bac also looked straight (Fig. S1(b) and S1(d)), and the normality was not rejected by the chi-square test (p-values were 0.22 and 0.01 for general *E. coli* and Human Bac, respectively, Table S3). Compared to the high positive rate of these indicator microorganisms and genetic markers, virulence genes were detected at a lower frequency, with 39 and 28 positive samples for *eaeA* of enteropathogenic *E. coli* and *ciaB* of *C. jejuni*, respectively. The linearity of the normal probability plot larger than the quantification limit values is the necessary condition for the log-normality of whole dataset of these virulence gene concentrations. The normal probability plot of the logarithmic concentration values of *eaeA* looked straight, but that of *ciaB* did not (Fig. S1(e) and S1(f)). The chi-square test did not reject the log-normality of the observed values of *eaeA* with a p-value of 0.33, but rejected *ciaB* with a p-value of 1.16×10^{-4} (Table S3). This does not assure the log-normality of the whole dataset of *eaeA* concentration, but the further analyses were performed under the assumption that the quantified values of *eaeA* concentration are log-normally distributed.

Pearson's product-moment correlation coefficients between the virulence gene *eaeA* and three indicators were calculated (Table S4). The coefficient value between general *E. coli* and *eaeA* was

0.24 with a p-value of 0.17. This result means that the quantification of general *E. coli* fails to predict the presence of the virulence gene from enteropathogenic *E. coli* in the study area and period. Therefore, in the subsequent analyses, the quantitative data of general *E. coli* were not used. On the other hand, genetic markers (Total Bac and Human Bac) gave higher correlation coefficient values, which were 0.72 and 0.62 with p-values of 2.31×10^{-7} and 2.74×10^{-5}, respectively.

Predictive distribution of the concentration of genetic markers and *eaeA*.

Since significant correlations were detected between genetic markers (Total Bac and Human Bac) and the virulence gene *eaeA*, the posterior predictive distributions of the concentration of two genetic markers and *eaeA* were estimated individually. The posterior mean values of μ and log (σ), where $\sigma = \beta^{-1/2}$, of Total Bac were 3.45 and 0.03, respectively (Table 1). These values were identical with those calculated from raw data and estimated from the normality probability plot (Table S3). Since the genetic marker of Total Bac was detected in 100% (144/144) of the samples, the posterior mean and SD (Fig. 1(a)) and the predictive distribution of Total Bac concentration (Fig. 1(b)) were accurately estimated. In order to clarify the extent of accuracy of predictive distribution with 100% positive samples (144/144), 100 datasets were simulated using the posterior mean values of μ and log (σ). Quantile values and KL divergence are shown in Table 1. When there are 100 datasets of 144 quantified values, 100-times estimation of the posterior predictive distribution gave a KL divergence of 0.04 or less.

The genetic marker of Human Bac was detected 143 times out of 144 total samples. The posterior mean values of μ and log (σ) of Human Bac were 3.04 and 0.03, respectively (Table 2). These values were similar to those estimated from the normality probability plot (Table S3), which means that one non-detect out of 144 does not affect the accuracy of the estimation. The estimation of the posterior predictive distribution of Human Bac concentration (Fig. 1(c) and 1(d)) was also regarded as accurate, because the KL divergence of 100-times simulated datasets using the posterior mean

values of μ =3.04 and log (σ) = 0.03 was 0.04 or less (Table 2), which was the same accuracy level of 100% positive datasets (Table 1).

The proportion of positives for *eaeA* (27.1%) was significantly lower than those of Total Bac (100%) and Human Bac (99.3%). The virulence gene was detected 39 times out of 144 total samples (Table S3). The small number of positives resulted in a posterior distribution with low entropy and relatively large RMSDs, as shown in Fig.1(e) and Table 3. The posterior mean values of μ and log (σ) of *eaeA* were -1.14 and 0.01, respectively (Table 3), while those estimated from normality probability plot were -1.73 and 0.07, respectively (Table S3). In order to clarify the estimation accuracy, 100 simulated datasets with 27.1% positives were created using the posterior mean values of μ = -1.14 and log (σ) = 0.01. The maximum KL divergence was 0.12, while 75% of estimation gave KL divergence values less than 0.02 (Table 3). It is impossible to define how accurate is accurate in the estimation. However, the maximum KL divergence of 0.12 is lower than that obtained when 100% positives were obtained in the total sample number of 12 (Kato et al., 2013), which means that the posterior predictive distribution was estimated with relatively high accuracy by using 39 detects out of 144 total samples (Fig. 1(e) and 1(f)).

Distributions of the concentration ratio of genetic markers to *eaeA*.

The distributions of the concentration ratio between genetic markers and *eaeA* were depicted in Fig. 2. Since the posterior mean of μ for Total Bac (3.45, Table 1) was larger than that for Human Bac (3.04, Table 2), the distribution of Total Bac / *eaeA* (Fig. 2(a)) shifts to right compared to that of Human Bac / *eaeA* (Fig. 2(b)). These distributions were regarded to be very accurate, because the maximum values of KL divergence were 0.02 when 100 simulated datasets of genetic markers and *eaeA* were used to estimate the distributions of concentration ratio (Table 3). This level of KL divergence is comparable with that obtained for a pair of 100% positives with the total sample number of 48 (Kato et al., 2013). These results indicated that the combination of datasets with the

positive rates of 27.1% (39/144) and 99.3 (143/144) gave high estimation accuracy of the distribution of concentration ratio.

Discussion

The present study attempted to establish a computational procedure for inferring the concentration ratio between a fecal indicator and a pathogen based on left-censored datasets. Field investigation in a river watershed was conducted for 16 months, and quantitative datasets of conventional fecal indicators, genetic markers, and virulence genes of pathogenic bacteria were obtained. A Bayesian model that was adapted to a left-censored dataset with varying analytical quantification limit values was applied to the quantitative dataset. The posterior predictive distributions of the concentration ratio were predicted, and we found that 39 detects out of 144 total sample number was enough to accurately estimate the distribution of the concentration ratio, when combined with a dataset with a positive rate higher than 99%.

The most simple and convenient approach to investigate the ratio value must be the data accumulation of the fecal indicator/pathogen concentration ratio. However, this fecal indicator/pathogen concentration ratio is usually difficult to obtain, because the concentration of a pathogen is generally low, and the significant fraction is composed of non-detects (Kato et al., 2013), which makes it impossible to calculate the ratio value. The approach proposed in this study overcomes the difficulty in acquiring the concentration ratio between fecal indicators and pathogens in environmental water when left-censored datasets were obtained. The application of truncated probability distribution allows us to estimate the posterior predictive distribution of microbe concentrations based on left-censored data, which can be used for inferring the distribution of concentration ratio (Paulo et al., 2005).

Our analysis has three stages: The first stage checks the log-normality, the second stage confirms the correlation between parameters, and the third stage performs Bayesian analysis. However, one may think that a single framework containing all these steps is more elegant. However, when developing an algorithm for some analysis, a trade-off between elegance and reliability is usually faced in general. In this study, reliability was regarded as more important than simplicity. Unfortunately, few methods that check the normality only in a Bayesian framework are established, well-verified and accepted widely. Therefore, rather than putting all analysis steps in a single Bayesian framework, the chi-square goodness-of-fit test and the spearman's correlation test, which are well-established and widely accepted, are applied for checking the log-normality and the correlation between parameters before Bayesian analysis.

We employed a chi-square test at the significant level of 0.01 for checking the log-normality of quantitative datasets. However, when the dataset is left-censored, it is impossible to test the log-normality of a whole dataset, because only a part of a dataset (values of detected data) can be used in the test. In the other words, the acceptance of the null hypothesis in the chi-square test does not ensure that the log-normality assumption is really applicable. The normality check as the first stage of the proposed approach is just to exclude datasets that cannot be used in the framework. Under this setting, the significant level is subject to change, and in addition to it, the other tests such as Shapiro-Wilk test can be employed for the normality check. Investigators should scrutinize the nature of quantified data carefully in terms of the applicability of the assumption, and can select an appropriate test at an appropriate value of the significant level. In the chi-square test, we found that the null hypothesis was rejected for total coliforms and a virulence gene of *C. jejuni* (*ciaB*) (Table S3). We therefore excluded the total coliforms and *ciaB* in subsequent analyses. Although the chi-square test did not reject the log-normality of Total Bac and Human Bac, p-values for these genetic markers are 0.03 and 0.01, respectively (Table S3), which means that the log-normality of datasets of these genetic markers is rejected at a significant level of 0.05 in the chi-square test. Very

411 apparently one of the limitations of the proposed approach is the assumption of log-normality of
412 datasets. Since the occurrence of microorganisms in water is really episodic, the assumption of
413 stationarity in parameters of concentration distribution (e.g., mean and SD) is sometimes
414 inappropriate (Haas and Heller, 1988). It is worth preparing other Bayesian estimation algorithms
415 using a different probability distribution, such as gamma distribution and Weibull distribution
416 (Englehardt and Li, 2011). The comparison of estimation results between the log-normality-
417 assumed model and other distributions-assumed models may give us insights about the nature of
418 quantified data obtained in field investigations, which should be included in the further study.
419
420 The Bayesian estimation algorithm used in this study is available for a pair of parameters, even if
421 there is no correlation between them, which means that it is possible to obtain the distribution of
422 concentration ratio between a pathogen and an unrelated water quality parameter. Needless to say,
423 however, it is meaningless to investigate the concentration ratio between a pathogen and a water
424 quality parameter devoid of the correlation with pathogen occurrence. Thus, the correlation analysis
425 is essential as a component of the process for determining the concentration ratio between an
426 indicator and a pathogen. However, the correlation between an indicator and a pathogen is
427 extremely case-specific, as discussed for a long time (Wu et al., 2011). We detected the significant
428 correlations between genetic markers (Total Bac and Human Bac) and *eaeA* (Table S4), which does
429 not mean that these significant correlations can be observed in the other watersheds. That's why this
430 study is a case study presenting the estimation process of the distribution of concentration ratio
431 between an indicator and a pathogen. Although the result of correlation analysis is devoid of
432 generality, the proposed process must be available in the other settings, when the log-normality of
433 datasets is not rejected and a significant correlation is detected between pathogen and indicator
434 concentrations. If a dataset to be analyzed does not follow the log-normal distribution, another
435 Bayesian estimation algorithm using an appropriate probability distribution has to be prepared and
436 applied, as already discussed above.

Accuracy is always the most important issue in the estimation of the distribution of concentration ratio between a fecal indicator and a pathogen (Kato et al., 2013). Since fecal indicators are detected more easily than pathogens because of the relatively high concentration, the accuracy is usually dependent on the number of detects in the quantitative dataset of pathogens. In this study, 39 detects out of 144 total samples was enough to accurately estimate the distribution of concentration ratio when combined with datasets with a high proportion of positives (100% for Total Bac and 99.3% for Human Bac). In Bayesian analysis, the tolerable accuracy level depends on situations of scenarios of applications. It may be thus desirable to perform a simulation using artificially generated left-censored data that have the identical number of non-detects with actual data, for confirming how large KL divergence is obtained.

For accuracy evaluation of our Bayesian algorithm, we generated 100 datasets artificially. Here a justification of this procedure is described by answering two questions that why 100 datasets are needed and why artificial data is used. If only a single dataset was generated, the estimation result might be much better accidentally or might be much worse. To perform accurate assessment, repeat experiments are necessary. Furthermore, to use quartile values for assessment, the number of repeat experiments is chosen as 100 in this study. The reason why artificial data is used is that there is no way to know the true distribution of any real-world data. By using artificial data, the assessment can be done by comparison of the estimated distribution to the true distribution that has generated the artificial data.

In the present case study, only two virulence genes from pathogenic bacteria (*eaeA* of enteropathogenic *E. coli* and *ciaB* of *C. jejuni*) were investigated. However, a variety of virulence genes from multiple pathogens are present in water environments (Ishii et al., 2014b), which requires us to identify the most important target for analyzing the quantitative relationship with

indicators. It may be also necessary to employ appropriate virus markers (Kitajima et al., 2011; Fumian et al., 2013; Love et al., 2014), if pathogens of concern are viruses. Chemical markers are also available for indicating fecal contaminations (Black et al., 2007; Kuroda et al., 2012). Pathogens occur in water episodically and those posing infectious risks may vary from place to place and from season to season in terms of species and strains, which means that the quantification of multiple pathogens and indicators (Wong et al., 2013; Ishii et al., 2014a) has to be conducted at each study area. Accumulation of multiple quantitative data at each location is a basis for better understanding the quantitative relationship between indicators and pathogens.

Conclusions

A Bayesian model for estimating the fecal indicator/pathogen concentration ratio was constructed, in which a left-censored dataset with varying analytical quantification limit values was available. When the sample size was 144, numerical simulations concluded that 39 detects was enough to accurately estimate the distribution of concentration ratio when combined with datasets with a positive rate higher than 99%. To evaluate the level of accuracy in the estimation, it is desirable to perform a simulation using an artificially generated left-censored data that has the identical number of non-detects with actual data.

Acknowledgements

We thank Ms. Rie Nomachi and Reiko Hirano for their technical help. This work was supported by the Japan Society for the Promotion of Science Through Grant-in-Aid for Scientific Research (B) (26303011).

References

APHA, AWWA&WEF. 2005 *Standard method for the examination of water and wastewater*, 21stedition.American Public Health Association, Washington D.C.

Black, L. E., Brion, G. M. & Freitas, S. J. 2007 Multivariate logistic regression for predicting total culturable virus presence at the intake of a potable-water treatment plant: Novel application of the atypical coliform/total coliform ratio. *Appl. Environ. Microbiol.* **73**(12), 3965-3974.

Bosch, A., Guix, S., Sano, D. & Pinto, R. M. 2008. New tools for the study and direct surveillance of viral pathogens in water. *Food Environ. Virol.* **19**, 295-310.

Cohen, C. Jr. 1959, Simplified estimators for the normal distribution when samples are singly censored or truncated. *Technometrics*, **1**(3), 217-237.

Englehardt, J. D. & Li R. 2011The discrete Weibull distribution: An alternative for correlated counts with confirmation for microbial counts in water. *Risk Anal.* **31**(3), 370-381.

Frankel, G., Candy, D. C. A., Fabiani, E., Adu-Bobie, J., Gil, S., Novakova, M., Phillips, A. D.& Dougan, G. 1995 Molecular characterization of a carboxy-terminal eukaryotic-cell-binding domain of intimin from enteropathogenic *Escherichia coli*. *Infect. Immun.* **63**, 4323-4328.

Fumian, T. M., Vieiram C. B., Leite, J. P. G. & Miagostovich, M. P. 2013 Assessment of burden of virus agents in an urban sewage treatment plant in Rio de Janeiro, Brazil. *J. Water Health* **11**(1), 110-119.

Gilliom, R. J. & Helsel, D. R. 1986 Estimation of distributional parameters for censored trace level water quality data, 1. estimation techniques. *Water Resour. Res.* **22**(2), 135-146.

Haas, C. N. & Heller, B. 1988. Test of the validity of the Poisson assumption for analysis of most-probale-number results. *Appl. Environ. Microbiol.* **54**(12), 2996-3002.

Helsel, D. R. 2006 Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **65**, 2434-2439.

Ishii, S., Kitamura, G., Segawa, T., Kobayashi, A., Miura, T., Sano, D. & Okabe, S. 2014a Microfluidic quantitative PCR for simultaneous quantification of multiple viruses in

515 environmental water samples. *Appl. Environ. Microbiol.* **80**(24), 7505-7511.
516 Ishii, S., Nakamura, T., Ozawa, S., Kobayashi, A., Sano, D. & Okabe, S. 2014b Water quality
517 monitoring and risk assessment by simultaneous multipathogen quantification. *Environ. Sci.*
518 *Tech.***48**, 4744-4749.
519 Ishii, S., Segawa, T. & Okabe, S. 2013 Simultaneous quantification of multiple food and waterborne
520 pathogens by use of microfluidic quantitative PCR. *Appl. Environ. Microbiol.* **79**(9), 2891-2898.
521 Itoh, H. 2013. Effect of the ratio of illness to infection of *Campylobacter* on the uncertainty of
522 DALYs in drinking water. *J. Water Environ. Tech.* **11**(3), 209-224.
523 Kato, T., Miura, T., Okabe, S. & Sano, D. 2013 Bayesian modeling of enteric virus density in
524 wastewater using left-censored data. *Food Environ. Virol.* **5**(4), 185-193.
525 Kitajima, M., Haramoto, E., Phanuwan, C. & Katayama, H. 2011 Prevalence and genetic diversity
526 of Aichi viruses in wastewater and river water in Japan. *Appl. Environ. Microbiol.* **77**(6), 2184-
527 2187.
528 Kobayashi, A., Sano, D., Hatori, J., Ishii, S. & Okabe, S. 2013a Chicken- and duck-associated
529 *Bacteroides-Prevotella* genetic markers for detecting fecal contamination in environmental
530 water. *Appl. Microbiol. Biotech.* **97**, 7427-7437.
531 Kobayashi, A., Sano, D., Taniuchi, A., Ishii, S. & Okabe, S. 2013b Use of a genetically-engineered
532 *Escherichia coli* strain as a sample process control for quantification of the host-specific
533 bacterial genetic markers. *Appl. Microbiol. Biotech.* **97**, 9165-9173.
534 Konkel, M. E., Kim, B. J., Rivera-Amill, V. & Garvis, S. G. 1999 Bacterial secreted proteins are
535 required for the internalization of *Campylobacter jejuni* into cultured mammalian cells. *Mol.*
536 *Microbiol.* **32**, 691-701.
537 Kuroda, K., Murakami, M., Oguma, K., Marumatsu, Y., Takada, H. & Takizawa, S. 2012
538 Assessment of groundwater pollution in Tokyo using PPCPs as sewage markers. *Environ. Sci.*
539 *Tech.* **46**, 1455-1464.
540 Labite, H., Lunani, I., van der Steen, P., Vairavamoorthy, K., Drechsel, P. & Lens, P. 2010

Quantitative microbial risk analysis to evaluate health effects of interventions in the urban water system of Accra, Ghana. *J. Water Health* 8(3), 417-430.

Lalancett, C., Papineau, I., Payment, P., Dorner, S., Servais, P., Barbeau, B., Di Giovanni, G. D. and Prevost, M. 2014 Changes in *Escherichia coli* to *Cryptosporidium* ratios for various fecal pollution sources and drinking water intakes. *Water Res.* **55**, 150-161.

Love, D. C., Rodriguez, R. A., Gibbons, C. D., Griffith, J. F., Yu, Q., Stewart, J. R. & Sobsey, M. D. 2014 Human viruses and viral indicators in marine water at two recreational beaches in Southern California, USA. *J. Water Health* **12**(1), 136-150.

Machdar, E., van der Steen, N. P., Raschid-Sally, L. & Lens, P. N. L. 2013 Application of quantitative microbial risk assessment to analyze the public health risk from poor drinking water quality in a low income area in Accra, Ghana. *Sci. Total Environ.* **449**, 132-142.

Okabe, S., Okayama, N., Savichtcheva, O. & Ito, T. 2007 Quantification of host-specific *Bacteroides- Prevotella* 16S rRNA genetic markers for assessment of fecal pollution in freshwater. *Appl. Microbiol. Biotech.* **74**, 890-901.

Paulo, M. J., van der Voet, H., Jansen, M. J. W., ter Braak, C. J. F. & van Klaveren J. D. 2005 Risk assessment of dietary exposure to pesticides using a Bayesian method. *Pest Manag. Sci.* **61**, 759-766.

Scott, T. M., Rose, J. B., Jenkins, T. M., Farrah, S. R. & Lukasik, J. 2002 Microbial source tracking: Current methodology and future directions. *Appl. Environ. Microbiol.* **68**, 5796-5803.

Setiyawan, A. S., Yamada, T., Fajri, J. A. & Li, F. Characteristics of fecal indicators in channels of johkasou systems. *J. Water Environ. Tech.* **12**(6), 469-480.

Silverman, A. I., Akrong, M. O., Amoah, P., Drechsel, P. & Nelson, K. L. 2013 Quantification of human norovirus GII, human adenovirus, and fecal indicator organisms in wastewater used for irrigation in Accra, Ghana. *J. Water Health* **11**(3), 473-488.

Tanaka, H., Asano, T., Schroeder, E. D. & Tchobanoglous, G. 1998 Estimating the safety of wastewater reclamation and reuse enteric virus monitoring data. *Water Environ. Res.* **70**(1), 39-

567 51.
568 WHO. 2011 *Guidelines for Drinking Water Quality 4th Edition*.
569 http://www.who.int/water_sanitation_health/publications/2011/dwq_guidelines/en/ (access:
570 2013.6.18).
571 Wong, M. V. M., Hashsham, S. A., Gulari, E., Rouillard, J.-M., Aw, T. G. & Rose, J. B. 2013
572 Detection and characterization of human pathogenic viruses circulating in community
573 wastewater using multi target microarrays and polymerase chain reaction. *J. Water Health*
574 **11**(4), 659-670.
575 Wu, J., Long, S. C., Das, D. & Dorner, M. 2011 Are microbial indicators and pathogens correlated?
576 A statistical analysis of 40 years of research. *J. Water Health* **9**(2), 265-278.

577 **Table 1**

578

Table 1. Estimation accuracy of μ and log (σ), and the predictive distribution of Total Bac

	μ			log (σ)			Kullback-Leibler divergence
	Posterior mean	RMSD	Posterior SD	Posterior mean	RMSD	Posterior SD	
μ and log (σ) estimated by the Bayesian approach	3.45	-	-	0.03	-	-	-
Minimum	3.17	0.08	0.08	-0.02	0.03	0.03	0.00
25%tile	3.40	0.09	0.09	0.02	0.03	0.03	0.00
Median	3.45	0.11	0.09	0.03	0.03	0.03	0.00
75%tile	3.51	0.13	0.09	0.04	0.04	0.03	0.01
Maximum	3.74	0.30	0.10	0.08	0.06	0.03	0.04

579

580

581

582 **Table 2**

583

Table 2. Estimation accuracy of μ and log (σ), and the predictive distribution of Human Bac

	μ			log (σ)			Kullback-Leibler divergence
	Posterior mean	RMSD	Posterior SD	Posterior mean	RMSD	Posterior SD	
μ and log (σ) estimated by the Bayesian approach	3.04	-	-	0.03	-	-	-
Minimum	2.85	0.08	0.08	-0.04	0.03	0.03	0.00
25%tile	2.98	0.10	0.09	0.03	0.03	0.03	0.00
Median	3.07	0.12	0.09	0.04	0.03	0.03	0.00
75%tile	3.11	0.13	0.09	0.05	0.04	0.03	0.01
Maximum	3.29	0.27	0.11	0.11	0.08	0.03	0.04

584

585

586

587

588 **Table 3**

589

Table 3. Estimation accuracy of μ and log (σ), and the predictive distribution of *eaeA*

	mean			log (σ)			Kullback-Leibler divergence
	Posterior mean	RMSD	Posterior SD	Posterior mean	RMSD	Posterior SD	
μ and log (σ) estimated by the Bayesian approach	-1.14	-	-	0.01	-	-	-
Minimum	-1.43	0.14	0.12	-0.10	0.05	0.05	0.00
25%tile	-1.22	0.16	0.15	-0.02	0.06	0.05	0.00
Median	-1.17	0.17	0.16	0.02	0.07	0.05	0.01
75%tile	-1.12	0.20	0.18	0.05	0.08	0.06	0.02
Maximum	-1.01	0.37	0.23	0.16	0.16	0.06	0.12

590

591
592

593 **Table 4**
594

Table 4. Kullback-Leibler divergence of the distribution of concentration ratio between *eaeA* and genetic markers

	Total Bac *vs eaeA*	Human Bac *vs eaeA*
Minimum	0.00	0.00
25%tile	0.00	0.00
Median	0.00	0.00
75%tile	0.01	0.01
Maximum	0.02	0.02

595
596

FIGURE LEGENDS

Figure 1. Posterior distribution of μ and log (σ) and posterior predictive distribution of concentration in environmental water. (a) Posterior distribution of μ and log (σ) of Total Bac. (b) Posterior predictive distribution of Total Bac concentration. (c) Posterior distribution of μ and log (σ) of Human Bac. (d) Posterior predictive distribution of Human Bac concentration. (e) Posterior distribution of μ and log (σ) of *eaeA*, a virulence gene of enteropathogenic *Escherichia coli*. (f) Posterior predictive distribution of *eaeA* concentration. Red circles are the observed concentrations, and blue circles are the detection limits. The area of each red circle is proportional to the number of observations at the value. The area of each blue circle is proportional to the number of undetected concentrations with the detection limit at the position. Since there were two values of analytical quantification limit for *eaeA* (data not shown), there are two blue circles in panel (f).

Figure 2. Distribution of concentration ratio between a generic marker (Total Bac or Human Bac) and *eaeA*, a virulence gene of enteropathogenic *Escherichia coli*. (a) Total Bac *vs eaeA*. (b) Human Bac *vs eaeA.*

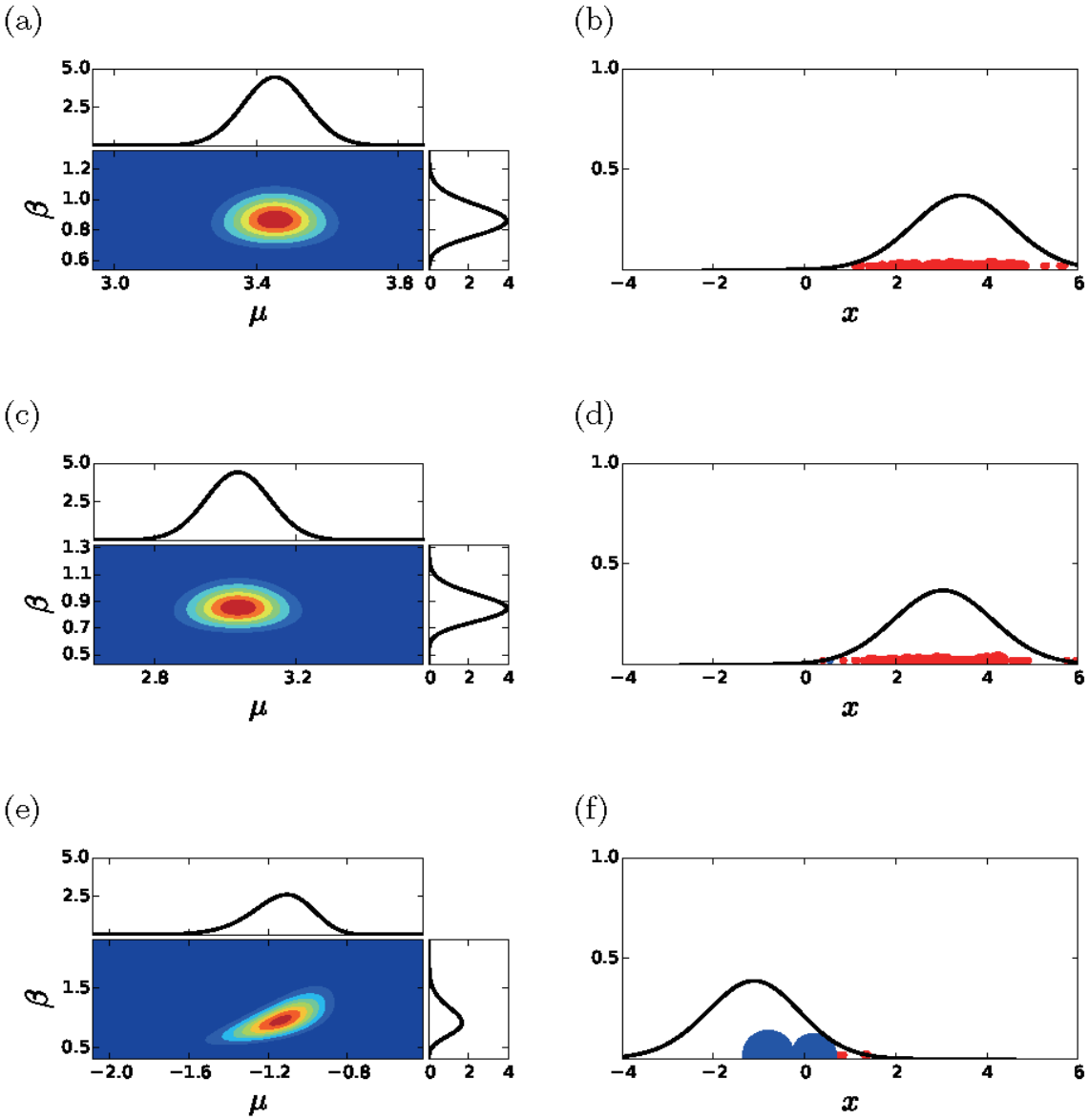

Figure 1. Posterior distribution of μ and log (σ) and posterior predictive distribution of concentration in environmental water. (a) Posterior distribution of μ and log (σ) of Total Bac. (b) Posterior predictive distribution of Total Bac concentration. (c) Posterior distribution of μ and log (σ) of Human Bac. (d) Posterior predictive distribution of Human Bac concentration. (e) Posterior distribution of μ and log (σ) of eaeA, a virulence gene of enteropathogenic Escherichia coli. (f) Posterior predictive distribution of eaeA concentration. Red circles are the observed concentrations, and blue circles are the detection limits. The area of each red circle is proportional to the number of observations at the value. The area of each blue circle is proportional to the number of undetected concentrations with the detection limit at the position. Since there were two values of analytical quantification limit for eaeA (data not shown), there are two blue circles in panel (f).

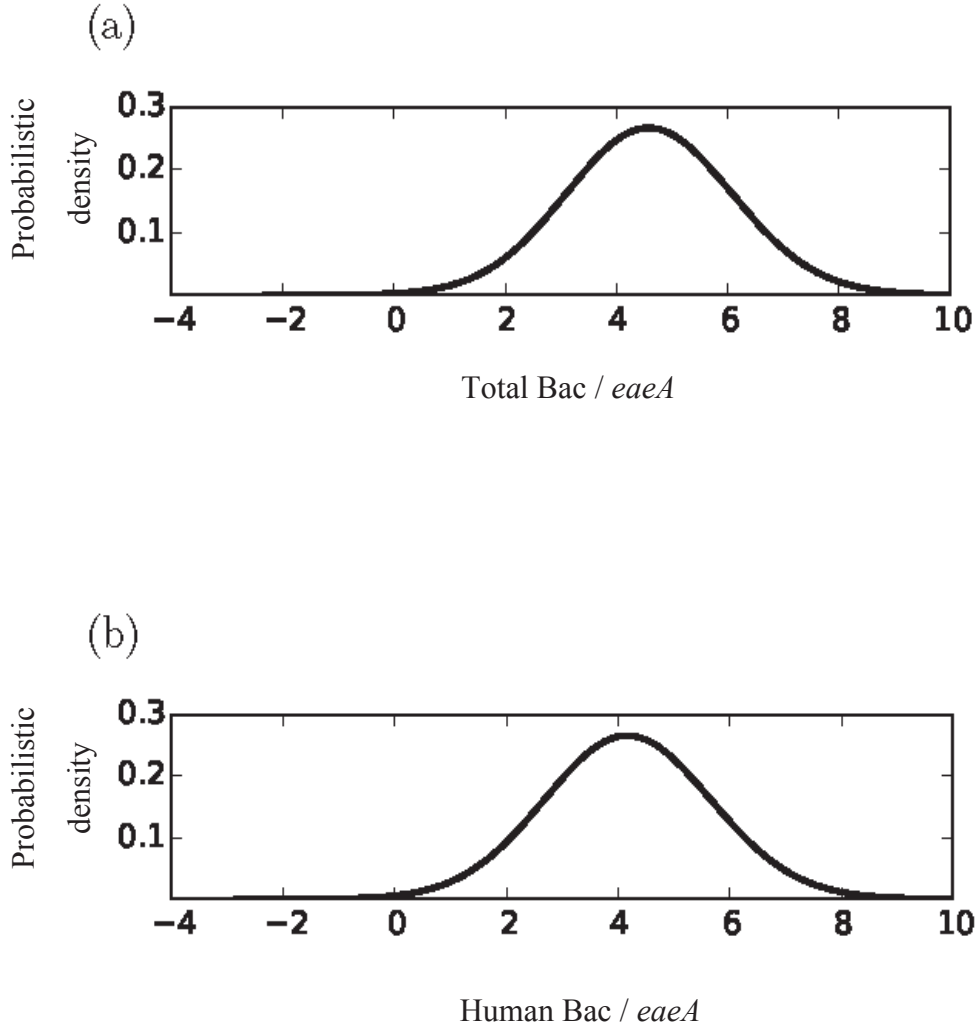

Figure 2. Distribution of concentration ratio between a generic marker (Total Bac or Human Bac) and *eaeA*, a virulence gene of enteropathogenic *Escherichia coli*. (a) Total Bac vs *eaeA*. (b) Human Bac vs *eaeA*.