Statistical Inference From Stem Cell Barcoding Data Using Adaptive Approximate Bayesian Computation

Siyi Chen
Rice University
Katherine Y. King
Baylor College of Medicine
Marek Kimmel (kimmel@rice.edu)
Rice University

Research Article

Keywords: HSC, barcodes, approximate Bayesian computation, species problem, Dirichlet-multinomial

Posted Date: February 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-187743/v1

RESEARCH

Statistical inference from stem cell barcoding data using adaptive approximate Bayesian computation

Siyi Chen[1], Katherine Y. King[2] and Marek Kimmel[1*]

*Correspondence: kimmel@rice.edu
[1] Department of Statistics, Rice University, 6100 Main St, 77005 Houston, USA
Full list of author information is available at the end of the article

Abstract

Background: Barcodes that can be supplied to cells by transduction of a library of unique DNA sequences allow identification of heterogeneity in cell populations and lineage tracing applications. Estimation of the number of hematopoietic stem cell (HSC) clones is important since it also allows to approximate the number of hematopoietic stem cells from which the circulating blood cells descend. This problem is similar to the species problem, well-known to ecologists. However, an additional "degree of freedom" exists, since different HSC generally give rise to clones with different growth rates. This adds credibility to sampling models based on different versions of Dirichlet-multinomial distributions.
Results: We developed a truncated population approximate Bayesian computation (ABC) algorithm which is derived from sequential Monte Carlo ABC (SMC-ABC) and applied the method to the symmetric Dirichlet-multinomial model proposed by Zhang et al. (2005) and asymmetric Dirichlet-multinomial model we proposed. Methodology was tested using simulated and real-life data.
Conclusions: Results suggest that flexibility of the asymmetric Dirichlet-multinomial helps to obtain insight into heterogeneity of proliferating cell systems such as HSC. Estimates based on experimantal data approach the correct count of murine HSC.

Keywords: HSC; barcodes; approximate Bayesian computation; species problem; Dirichlet-multinomial

Background

In recent years, next generation sequencing (NGS) has enabled new methods to study cells at the single-cell level, including single-cell DNA and RNA sequencing. Cells can be also traced using NGS paired with DNA barcodes in cells that are either endogenously present or exogenously supplied. Here, we are concerned with barcodes, which can be supplied to cells either by transduction of a library of unique DNA sequences, thereby allowing identification of heterogeneity in cell populations and lineage tracing, or by use of an inducible transposon which allows achieving a DNA barcode without removing cells from their natural environment. In this latter case the DNA sequence flanking a random site of transposon integration provides the barcode.

The use of barcoding systems for single cell studies stimulated interest in the development of statistical tools to analyze the data produced by these systems. The diversity of clones that can be studied is limited by the range of barcodes that can be generated and subsequently detected. We study here the exogenous barcoding

system as published by Sun et al. (2014) and specifically undertake estimation of the number of the hematopoietic stem cells (HSC) based on distribution of barcodes in peripheral blood cell, which are HSC descendants. Our findings can be applied to other barcoding experiments and moreover, since the problem is similar to the one with which the classical "problem of species" is concerned, to ecological applications to estimate the true species diversity based on species frequencies in a field study.

As an informal introduction, let us consider an example taken from Sun et al. (2014) (Fig. 1). We would like to estimate the number of existing HSC based on experimental single-cell data. Among the transposon tags (Tn) detected in granulocytes analyzed in a single plate, some of the tags were found in single granulocytes, some of them appeared in two cells, and fewer tags were observed three times or more. In the experiment , 36 weeks after barcoding all cells (including all HSC) in the experimental animal, tag 187 (numbering is arbitrary) was observed three times, tags 188, 189, and 190 were observed twice, the rest of the tags were only observed once. The 36 weeks were allowed to complete the replacement of all the blood cells by descendants of the barcoded HSC.

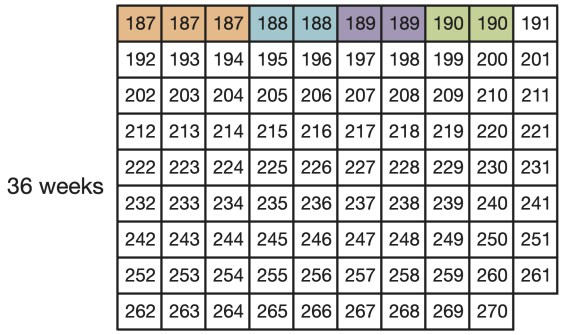

Figure 1 An example of Tn tags from barcoding results from Sun et al. (2014): In a plate of wells holding individual granulocytes, tag 187 (numbering is arbitrary) is observed three times, tags 188, 189, and 190 are observed twice, the rest of the tags are only observed once. The standard plate has 96 wells, but in some of them (7 in this case) tags were not identified.

Suppose that we detect w unique Tn barcodes among n granulocytes being descendants of K barcoded HSCs, and n_i is the number of granulocytes with tag i, while f_j is the number of different Tn tags present in j cells. Then n_i and f_j satisfy the following constraints:

$$\sum_{i=1}^{K} n_i = \sum_{i'=1}^{w} n'_{i'} = n, \tag{1}$$

$$\sum_{j=1}^{w} j f_j = n, \quad \sum_{j=1}^{w} f_j = w, \tag{2}$$

where $n'_{i'}$-s are the non-zero n_i-s. We will only use the latter, without causing confusion.

An important question is the development of an appropriate sampling model for observations relying on a sample of descendants of the HSC we try to count. If each

of the K HSC were producing a clone of large size, the sample of n descendants would constitute a multinomially distributed vector

$$(n_1, \ldots, n_K) \sim Mult(n; K^{-1}, \ldots, K^{-1}). \tag{3}$$

If $n \ll K$, then most of the entries of the vector $(n_1, \ldots, n_K)$ are 0-s. However, an additional "degree of freedom" exists, since different HSC generally give rise to clones with different growth rates. In such situation, the vector may be conditionally multinomial

$$(n_1, \ldots, n_K)|(\theta_1, \ldots, \theta_K) \sim Mult(n; \theta_1, \ldots, \theta_K), \tag{4}$$

with random coefficient vector $\theta = (\theta_1, \ldots, \theta_K)$. Because of the duality between multinomial and Dirichlet distributions, it seems natural to employ this latter for the coefficients. In this paper, we examine models involving symmetric and asymmetric Dirichlet distributions, and their use to obtain the posterior distribution of the number of barcoded clones K, given data, under adaptive priors.

Approximate Bayesian Computation (ABC) is a class of methods that apply for complex systems where likelihood function is intractable or impractical due to computational cost. The term ABC was first introduced by Beaumont et al. (2002). In ABC, the likelihood is approximated by a large number of simulations based on the acceptance-rejection. The algorithm first specifies a threshold value ϵ which indicates the required agreement level between observed and simulated data, and a prior distribution $\pi(\theta)$. We draw parameter θ from the prior distribution and generate simulated data y given θ according to the model $f(y|\theta)$, and determine the distance $d(y, y_{obs})$ between simulated and observed data. The algorithm accepts the parameter θ if $d(y, y_{obs}) < \epsilon$. If the data is high dimensional we may reduce the dimension by using summary statistics $s(y)$ rather than the vector y. The summary statistics $s(y)$ should be sufficient for θ however frequently we cannot identify a sufficient statistics. Use of a non-sufficient statistic introduces more 'approximation' to the result.

In this paper we present an adaptive truncated-population ABC method based on sequential Monte Carlo (SMC) ABC introduced by Sisson et al. (2007) and Toni et al. (2008). We estimate parameters of the asymmetric Dirichlet-multinomial model for the barcoding problem by ABC and a proposed adaptive ABC method. We also evaluate the performance of the adaptive method.

Methods

An adaptive truncated-population approximate Bayesian computation algorithm

Sisson et al. (2007) and Toni et al. (2008) proposed similar approaches of sequential ABC methods. We propose an adaptive truncated-population ABC algorithm based on their adaptive methods:

1 Following Toni et al. (2008), we define a series of non-increasing tolerance levels $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_t \geq \cdots \geq \epsilon_T$, where $t = 1, 2, \cdots, T$ indicates successive iterations.

2 The first iteration $t = 1$ is the pilot run. In the pilot run, for $i = 1, \cdots, N$ the algorithm generates the posterior distribution $\pi_{\epsilon_1}(\theta_i|y)$ using initial prior distribution $p(\theta_i)$ until the corresponding distance satisfies the acceptance criterion $dist(y, y_{obs}) < \epsilon_1$, with ϵ_1 being the first tolerance threshold. In this way we accept N estimated parameter values and we set an improper importance weight $w_1 = Dir(1, 1, \cdots, 1)$ with dimension N for each accepted parameter in the first iteration.

3 For iteration $t = 2, \cdots T$, we truncate the posterior distribution $\pi_{\epsilon_{t-1}}$ obtained in iteration $t-1$ to the interval $(\theta_{min}, \theta_{max})$, where θ_{min} and θ_{max} may be chosen as for example the α and $1-\alpha$ quantiles of $\pi_{\epsilon_{t-1}}$. In this way we can make sure new parameters will be sampled from a more 'concentrated' range thus expedite the algorithm. The algorithm also perturbs the parameter values by using kernel function K_t (e.g. a random walk distribution kernel, a uniform kernel or a Gaussian kernel) and obtains parameters from the kernel function. The truncated distribution under perturbation kernel K_t is $K_t(\theta_t|\theta_{t-1})$ and each parameter value is assigned with normalized weights w_t and kernel distribution function to smooth the distribution as well as favor parameters that are closer to the true posterior. We propose that for each iteration t the improper importance weight w_t, which is a Dirichlet random weight with size N defined as $w_t = Dir(1, 1, \cdots, 1)$ for each accepted parameter.

4 For each iteration t similarly we draw θ from truncated and perturbed parameters values with w_t until we accept N parameter values according to the corresponding ϵ_t chosen for the corresponding iteration.

Simulation study

An example from Sisson et al. (2007) and Beaumont et al. (2009) involves a Gaussian mixture model with two components with known variances and mixture weights. The unknown parameter is the common mean θ. This results in the "true" pdf

$$f(y|\theta) = 0.5\mathcal{N}(\theta, 1) + 0.5\mathcal{N}(\theta, 0.01). \tag{5}$$

A uniform is assumed $\theta \sim \mathcal{U}(-10, 10)$. The "true" (meaning an analytically derived one) posterior, given single observed data point $y_{obs} = 0$ is

$$\pi(\theta|y_{obs}) = 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(0, 0.01). \tag{6}$$

We employ the distance metric $dist(y_{obs}, y_{sim}) = (y_{obs} - y_{sim})^2$ and we run the adaptive truncated-population ABC strategy to find target $\pi(\theta|y)$, $N = 1500$ parameters values are accepted in each iteration. For iteration t the algorithm uses a uniform random walk kernel with its variance equals twice the empirical variance of the parameter values accepted in the previous iteration $t - 1$. Figure 2 depicts six iterations of the adaptive truncated-population ABC algorithm given the Gaussian mixture model. 'Observations' are sampled from the 'true' posterior (black lines in Fig. 2). The algorithm performs satisfactorily on this 'toy' example.

Symmetric Dirichlet-multinomial model for count data

As Zhang et al. (1997) and Boender and Rinnooy Kan (1987) suggested, in this section we apply a generalized multinomial model for estimation of the size of DNA barcoding population, with its sample value denoted as K. Let Y_i be the number of barcodes that are observed i-times in DNA sequencing results, when a sample n is taken from the population. Assume that the number of vector count $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_K)$ follows a multinomial distribution with probability $(\theta_1, \theta_2, \cdots, \theta_K)$:

$$\boldsymbol{Y}|(\theta_1, \theta_2, \cdots, \theta_K) \sim Mult(n; \theta_1, \theta_2, \cdots, \theta_K). \tag{7}$$

Dirichlet distribution can be applied to biological partition problems due to its exchangeability. Pitman (1996) considers species sampling problem by applying a Dirichlet random measure in his book. For DNA barcoding data, assume that the sampling process is independent and consider a symmetric Dirichlet distribution as the prior distribution for relative frequencies $(\theta_1, \theta_2, \cdots, \theta_K)$ with parameter α :

$$p(\theta|K, \alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K}(\prod_i^K \theta_i)^{\alpha-1}. \tag{8}$$

Assume the prior distribution for K is uniform. Then according to Bissiri et al. (2013) the posterior distribution of K given w and n is as follows:

$$p(K|w, n) = \frac{K(K-1)\cdots(K-w+1)\Gamma(K\alpha)/\Gamma(K\alpha+n)}{\sum_{l=w}^K l(l-1)\cdots(l-w+1)\Gamma(l\alpha)/\Gamma(l\alpha+n)}\mathbf{1}_{K\geq w}, \tag{9}$$

or

$$p(K|w, n) \propto \frac{K!}{(K-w)!}\frac{\Gamma(K\alpha)}{\Gamma(K\alpha+n)}, \tag{10}$$

which indicates that w is sufficient statistics for K if α is known. The posterior mode, which is also the Maximum A Posteriori (MAP) estimator, can be derived from this formula. Since in practical data analysis we do not know the value of α, the MAP estimator of K is thus not possible to compute.

The barcoding sampling problem can be treated as the joint estimation of both parameters K and α. Zhang et al. (2005) implemented MCMC algorithm to realize the estimation of both parameters, however the posterior of K relies heavily on its prior distribution. In order to minimize the effect of how we choose the prior, we apply the adaptive ABC approach to analyze the barcoding sampling data.

Asymmetric Dirichlet-multinomial model for count data

In the barcoding experiment, a biologically feasible assumption is likely that some of the barcoded HSC divide and differentiate abundantly whereas other remain dormant. In order to account for heterogeneity and variation within different clusters in barcodes, we then consider an asymmetric Dirichlet-multinomial model of cell probabilities $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_K)$ with positive real parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K)$,

which has the probability density function:

$$p(\boldsymbol{\theta}|K,\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_i^{\alpha_i-1}, \tag{11}$$

We can also rewrite this formula as for each barcode, the cell frequencies $\boldsymbol{\theta}$ follow an asymmetric Dirichlet distribution with parameters $\boldsymbol{q}\gamma = \boldsymbol{\alpha}$:

$$p(\boldsymbol{\theta}|K,\boldsymbol{q},\gamma) = \frac{1}{Z(\boldsymbol{q}\gamma)}\prod_{i=1}^{K}\theta_i^{q_i\gamma-1}, \tag{12}$$

where $Z(\boldsymbol{q}\gamma)$ is the normalizing constant in the Dirichlet distribution, and γ serves as a concentration parameter for $\boldsymbol{q}$, which can reflect the abundance level within the clusters of barcodes.

Now we specify the hierarchical Bayesian model for barcodes modeling as follows: first we take the model for the counts of observed species $\boldsymbol{y}$ to be multinomial,

$$\boldsymbol{y}|\boldsymbol{\theta}, n \sim Mult(n;\boldsymbol{\theta}), \tag{13}$$

then in the second stage,

$$\boldsymbol{\theta}|K,\boldsymbol{q},\gamma \sim Dir(\boldsymbol{q}\gamma). \tag{14}$$

We also assume that

$$\boldsymbol{q} \sim Dir(\alpha), \tag{15}$$

in order to make a complete Bayesian analysis. Without any prior knowledge we assume that α and γ are independent and from uniform distributions. The resulting hierarchical Bayesian model for multinomial count data is:

$$p(K,\boldsymbol{q},n,\boldsymbol{\theta},\alpha|\boldsymbol{y}) \propto p(\boldsymbol{q}|\boldsymbol{\theta},n)p(\boldsymbol{\theta}|K,\boldsymbol{q},\gamma)p(K)p(\boldsymbol{q}|\alpha)p(\gamma)p(\alpha). \tag{16}$$

Adaptive truncated-population ABC for Dirichlet-multinomial model

Assume that K, α (and γ) are drawn from uniform distributions, then according to approximate Bayesian computation algorithm:

1. Compute the summary statistics of observed dataset s_{obs}, and choose $\boldsymbol{s} = (w, f_1, f_2)$ as summary statistics. Draw parameters from prior distributions of α, K (and γ).
2. Simulate the Dirichlet-multinomial distribution and compute summary statistics s_{sim}.
3. For each iteration, accept parameter values if weighted L_2 distance: $dist(s_{sim}, s_{obs}) \leq \epsilon$, then accept parameters.

Choice of summary statistics and stopping rule

The Dirichlet-multinomial model depends on w as well as vector counts $\boldsymbol{f} = (f_1, f_2, \cdots, f_n)$. Considering that the nonparametric estimator from Chao (1984) is based on (w, f_1, f_2), we also choose $\boldsymbol{s} = (w, f_1, f_2)$ as the summary statistics.

Much research is carried out on how to determine the stopping criterion in an adaptive ABC algorithm (Simola et al. 2020). In this paper we apply the following simple criterion: starting from the fourth iteration, if a certain α quantile $\boldsymbol{\theta}_{\alpha t} = (\theta_{\alpha t1}, \theta_{\alpha t2}, \cdots)$ for example when $\alpha = 10\%$ of the posterior distribution in iteration t satisfies $0.9 \leq \frac{\theta_{\alpha ti}}{\theta_{\alpha(t-1)i}} \leq 1.1$ for each component i in the parameter vector, then the algorithm is stopped at iteration t and we assume that the posterior is stable.

Results

In order to address the poor-concentration issue caused by informative priors, an adaptive Bayesian method is applied in the following section. The adaptive ABC method is intended to improve the problems resulting from choosing uniform priors. A general summary of results is found in Table 1.

	true K	estimated K	true α	estimated α	symmetric/asymmetric model/data
Simulation I	2000	1967 (1708, 2571)	1	0.87 (0.55, 1.64)	symmetric model on symmetric data
Simulation II	8000	7817 (7424, 8360)	4	4.38 (2.55, 5.91)	symmetric model on symmetric data
Simulation III	2000	1982 (1718, 2333)	n/a	2.07 (1.35, 3.88)	asymmetric model on symmetric data
Simulation IV	8000	8545 (7960, 9019)	n/a	15.11 (8.40, 18.63)	asymmetric model on symmetric data
Simulation V	10000	9412 (8533, 10198)	2	2.34 (1.80, 3.16)	asymmetric model on asymmetric data
Simulation VI	6000	5860 (5582, 6088)	4	4.43 (3.36, 5.36)	asymmetric model on asymmetric data

Table 1 Summary of results of estimated K from six simulation studies. The true α value is not listed in Simulations 3 and 4, since the interpretation of parameter α in the symmetric model is different from that in the asymmetric model. Let us note that the crucial parameter K is correctly estimated also in these simulations.

Simulation study I

We simulate the species count data under symmetric Dirichlet-multinomial model with true $K = 2000$ and $\alpha = 1$. We also set initial uniform prior $\mathcal{U}(0.01, 8)$ for α and $\mathcal{U}(100, 6000)$ for K. 4 iterations are performed and results are presented in Figure 3(a). The final posterior gives $\widehat{K_{MAP}} = 1967$, with 95% credible interval $(1708, 2571)$ and $\widehat{\alpha_{MAP}} = 0.87$, with 95% credible interval $(0.55, 1.64)$. We also present the joint density plot of K and α (see Figure 3 (b)). Interestingly, the joint credible region for K and α is hyperbolic ("banana") in shape, indicating trade-off between the estimates of K and α. This effect might be influenced by sampling from uniform priors.

Simulation study II

We simulate barcoding data from symmetric Dirichlet-multinomial model with true $K = 8000$ and $\alpha = 4$. The uniform prior for the first iteration is $\mathcal{U}(1, 10)$ for α, and $\mathcal{U}(2000, 12000)$ for K. 5 iterations are performed and results are presented in Figure 4(a). The final posterior gives $\widehat{K_{MAP}} = 7817$ with 95% credible interval $(7424, 8360)$ and $\widehat{\alpha_{MAP}} = 4.4$, with 95% credible interval $(2.6, 5.9)$. We also present the joint density plot of K and α (see Figure 4(b)) and we also notice the hyperbolic ("banana") in shape, indicating trade-off between K and α.

Simulation study III

In order to check the sensitivity of the adaptive-prior ABC algorithm to model mis-specification, we also apply the algorithm involving the asymmetric Dirichlet-multinomial model to the barcoding experiment data simulated from symmetric Dirichlet-multinomial simulation with true $K = 2000$. 4 iterations are performed. The asymmetric Dirichlet-multinomial model gives estimate K is 1982 with credible interval $(1718, 2333)$, which covers the true value. Posterior distributions from the last iteration are presented in Figure 5(a). We also present the joint density plots of K and α and γ in Figure 5(b).

Simulation study IV

We also apply the asymmetric Dirichlet-multinomial model to the simulated barcoding data from symmetric Dirichlet-multinomial simulation with true $K = 8000$. 6 iterations are performed. The asymmetric Dirichlet-multinomial model gives estimate K of 8545 with credible interval $(7960, 9019)$. Posterior distributions from the last iteration with kernel densities presented in dash lines in Figure 6(a). We also present the joint density plot of K and α and γ in Figure 6(b).

Simulation study V

We also apply the asymmetric Dirichlet-multinomial model to the simulated barcoding experiment data from asymmetric Dirichlet-multinomial simulation with true $K = 10000, \gamma = 3000$ and $\alpha = 2$. 6 iterations are performed. The estimated K is 9412 with credible interval $(8533, 10198)$. The estimated γ is 2935 with credible interval $(2822, 3091)$. The estimated α is 2.3 with credible interval $(1.8, 3.2)$. Posterior distributions from the last iteration are presented in Figure 7(a). We also present the joint density plot of K and α and γ in Figure 7(b).

Simulation study VI

We apply the asymmetric Dirichlet-multinomial model to another simulated barcoding data from asymmetric Dirichlet-multinomial simulation with true $K = 6000, \gamma = 2000$ and $\alpha = 4$. 5 iterations are performed. The estimated K is 5860 with credible interval $(5582, 6088)$. The estimated γ is 2036 with credible interval $(1973, 2088)$. The estimated α is 4.43 with credible interval $(3.36, 5.36)$. Posterior distributions from the last iteration are presented in Figure 8(a). We also present the joint density plot of K and α and γ in Figure 8(b).

Application to barcoding data

We focus on the estimation of size population in DNA barcoding data, based on sequencing experimental results. We consider granulocyte barcoding experimental results presented in Sun et al. (2014). The barcoding data has a total of 290 single granulocytes, displaying 270 unique barcoding tags. 254 of them are singletons, 14 tags are doublets, one tag is a triplet (the data are from Sun et al, 2014). 6 iterations are performed. The adaptive ABC method gives estimate $\widehat{K_{MAP}} = 9850$ with credible interval $(9672, 10029)$, and it is close to the standard accepted value which is around 10^4. Histograms of the final posteriors for K, γ and α are depicted in Figure 9(a). We also present the joint density plot of K and α and γ in Figure 9(b).

Conclusions and Discussions

In this article we consider statistical problems related to the use of new technologies in hematology, specifically of cell barcoding. This technique is used primarily to label a high percentage of cells in the organism with unique labels, being short DNA sequences integrated randomly in cell genome. As mentioned, at the time of barcoding, the identification is unique, however with time progeny of barcoded cells that proliferated, inherit the label from their parents. At the time when the sample is ascertained, some cells will have identical barcodes. The number of these different clones, reflect the number of labeled cells that proliferated. However, when cells are sampled, it is possible to characterize only a limited count n of them, usually of the order of several hundred. The number K of originally labeled cells of interest, such as for example the hematopoietic stem cells (HSC) is of the order or 10^4 or higher. The typical problem a biologist wishes to deal with (see Background and Fisher et al. (1943)) is to estimate K given the sample of size n and more specifically given the count w and frequency f of the clones represented in the sample. This problem is similar to the well-established species-problem in theoretical population biology (see Zhang and Stern (2009) and Zhang (2007)) .

A number of approaches to estimation of K have been proposed, including both frequentist (Rocchetti et al. (2011), Good and Toulmin (1956), and Chao (1984)) such as the Chao's lower bound, and Bayesian (Efron and Thisted (1976) and Zhang and Stern (2005)). In this paper, we designed and employed an adaptive Approximate Bayesian Computation algorithm based on symmetric Dirichlet-multinomial and asymmetric Dirichlet-multinomial models. The multistage algorithm, employs truncation of the intermediate posterior distributions, followed by kernel smoothing and re-weighting, to counteract the effects of non-informative priors. We used simulations, to demonstrate the convergence of the the algorithm under different scenarios, scaled to be similar to the typical experimental data. In addition, we applied the algorithm to previously published HSC barcoding data, showing its consistency with previous results.

We notice that the asymmetric Dirichlet-multinomial model can be seen as the generalized form of symmetric Dirichlet-multinomial model that addresses the heterogeneity issue and the asymmetric Dirichlet-multinomial is applicable for population size estimation of simulated data from symmetric Dirichlet-multinomial model. The use of Dirichlet distribution with flexibly adjustable parameter α is a generalization of previous models, which used $\alpha = 1$ (Boender and Kan, 1987) and $\alpha \to \infty$ (Sun et al., 2014). It is interesting to notice that the α-value estimated from HSC data fluctuates between 3 and 4, which indicates heterogeneous clone size, however less so than if the multinomial parameters were sampled from uniform distribution on a simplex as it was proposed in Boender and Kan (1987).

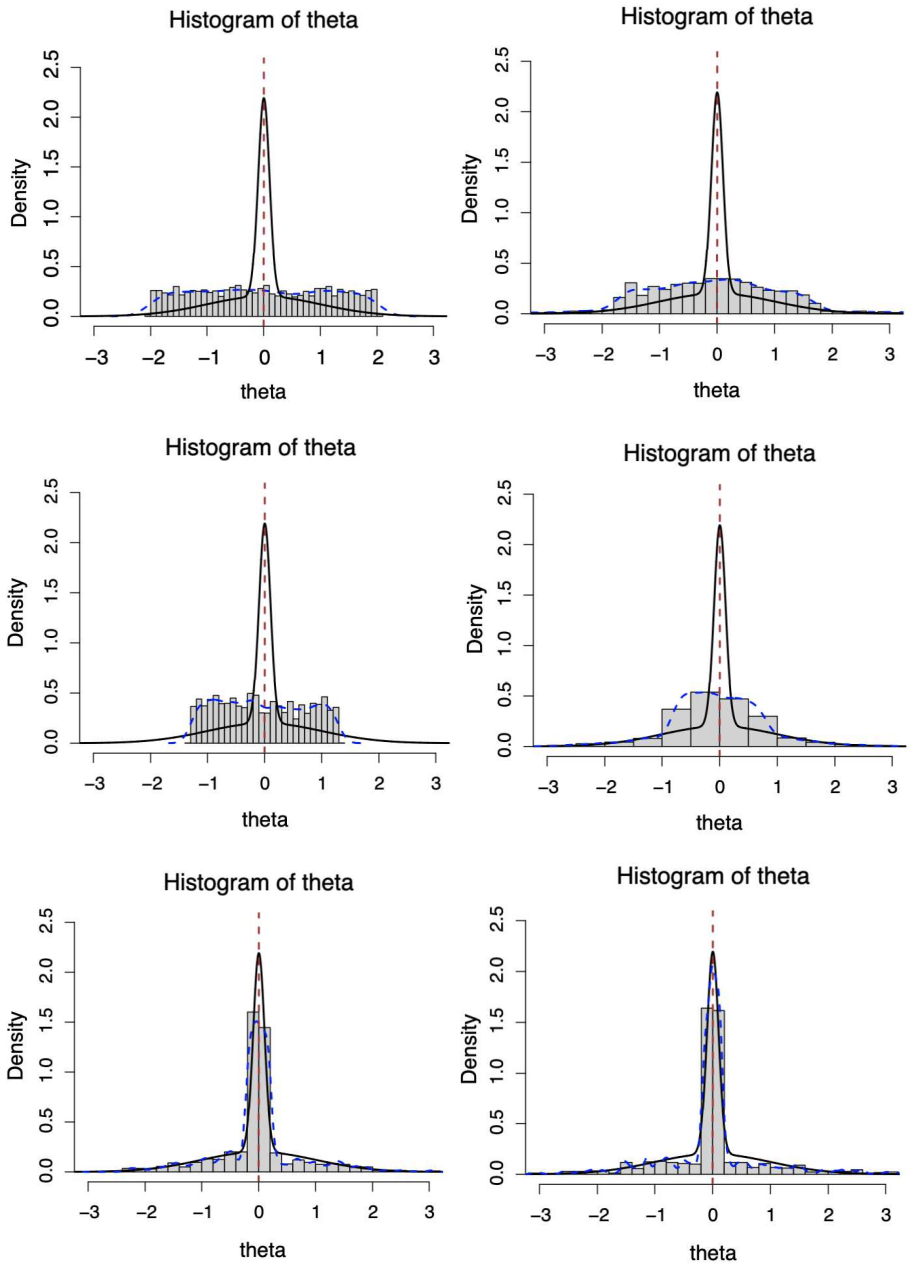

Figure 2 Histograms of six iterations of the adaptive truncated-population ABC given the Gaussian mixture model. Black lines represent true posterior densities, dashed blue lines are kernel density estimates from simulations.

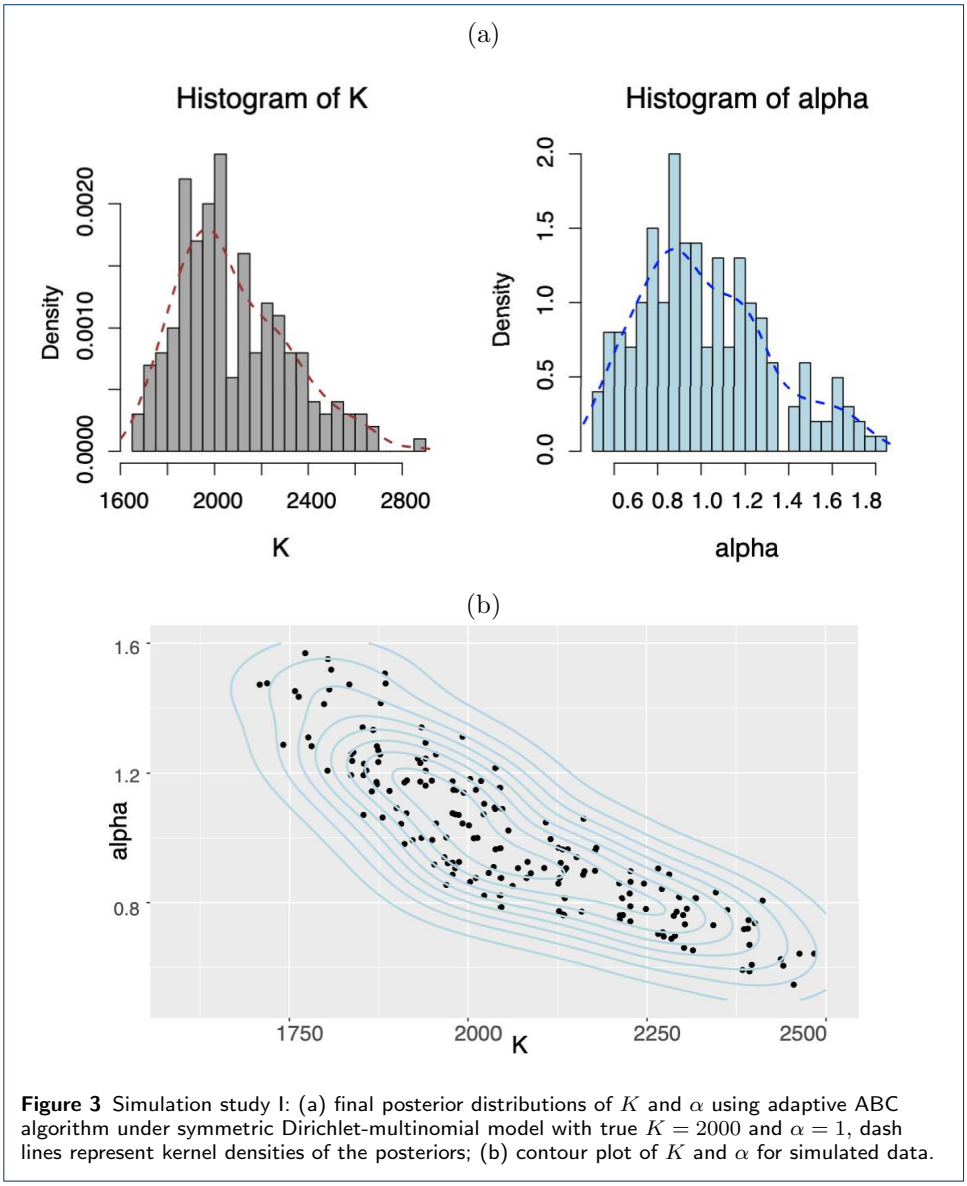

Figure 3 Simulation study I: (a) final posterior distributions of K and α using adaptive ABC algorithm under symmetric Dirichlet-multinomial model with true $K = 2000$ and $\alpha = 1$, dash lines represent kernel densities of the posteriors; (b) contour plot of K and α for simulated data.

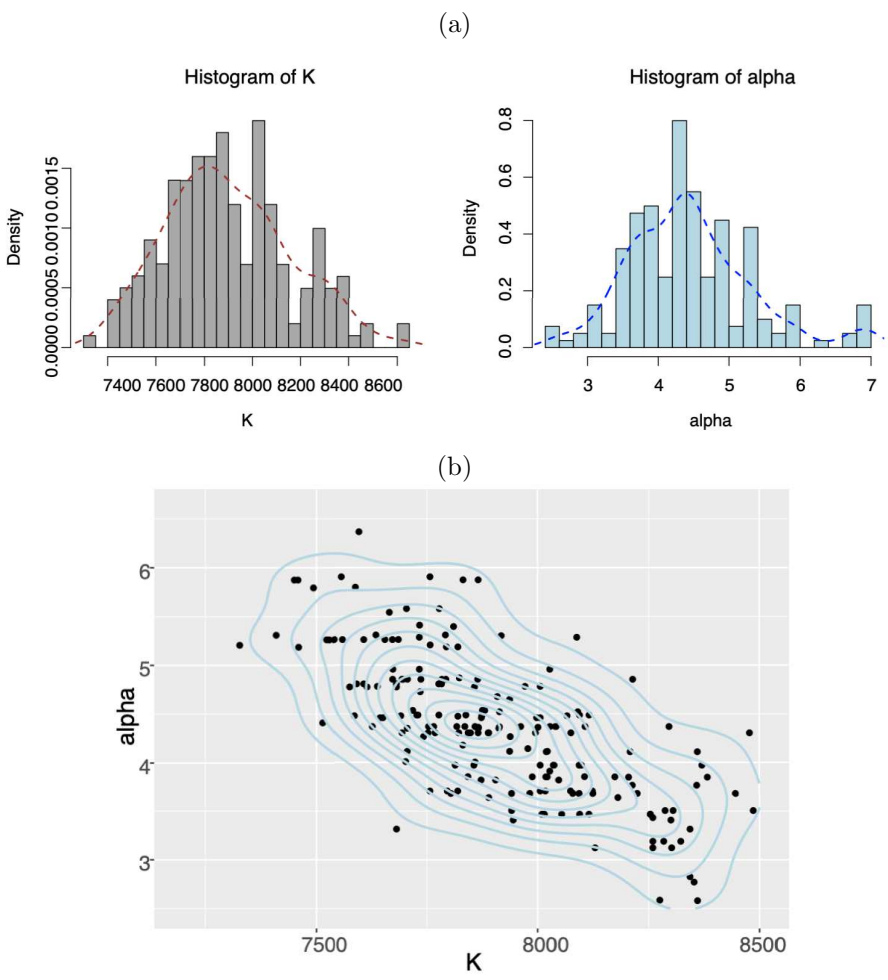

Figure 4 Simulation study II: (a) final posterior distributions of K and α using adaptive ABC algorithm with dash lines representing the kernel densities. The simulated data is from symmetric Dirichlet-multinomial model with true $K = 8000$ and $\alpha = 4$; (b) contour plot of K and α for simulated data.

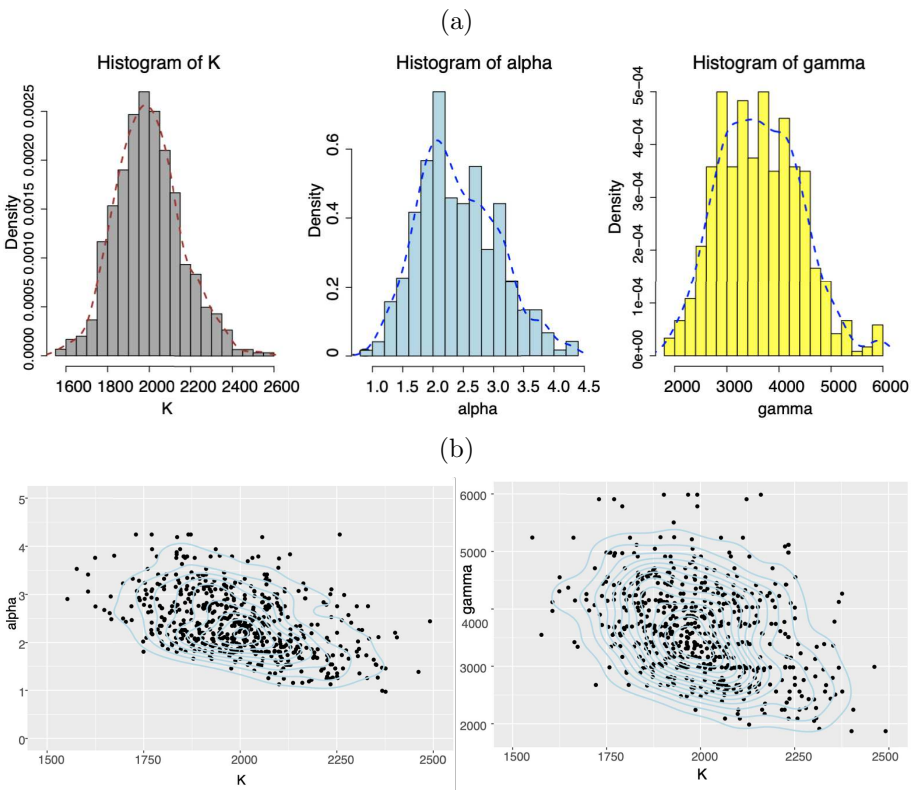

Figure 5 Simulation study III: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model with true $K = 2000$ with dash lines representing posterior kernel densities; (b) contour plots of K and α and γ for simulated data.

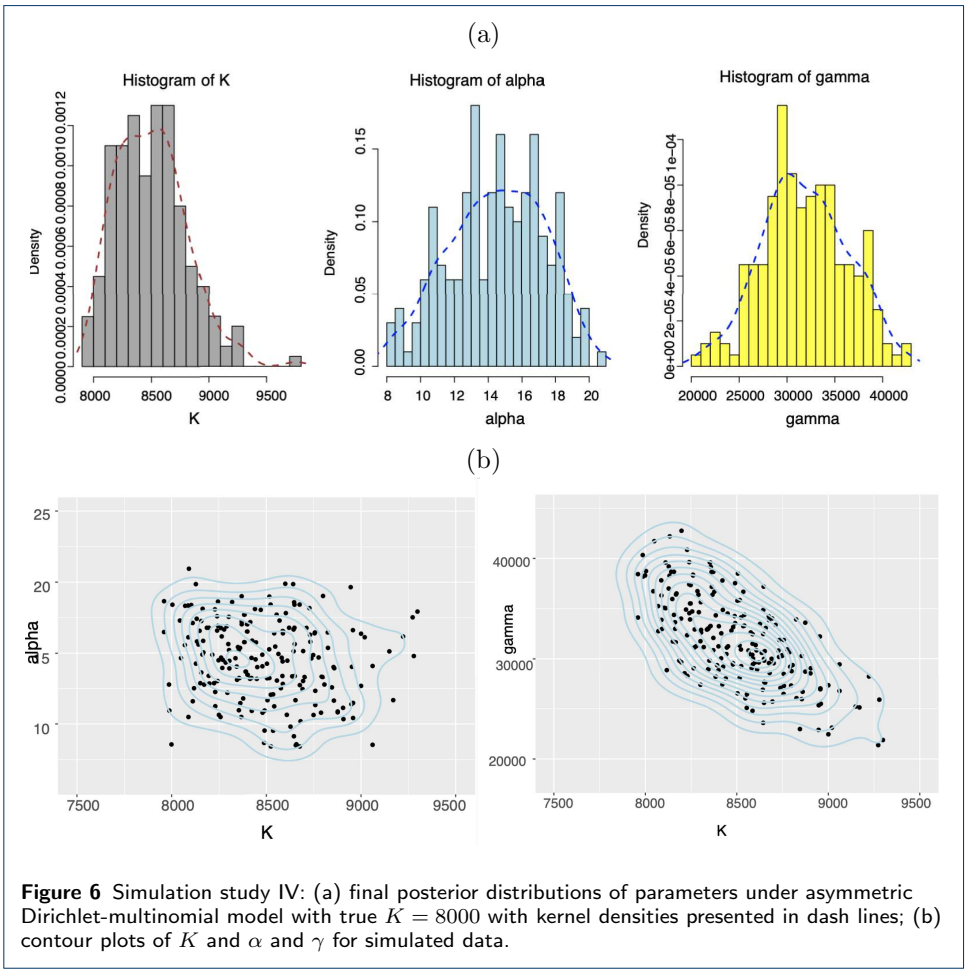

Figure 6 Simulation study IV: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model with true $K = 8000$ with kernel densities presented in dash lines; (b) contour plots of K and α and γ for simulated data.

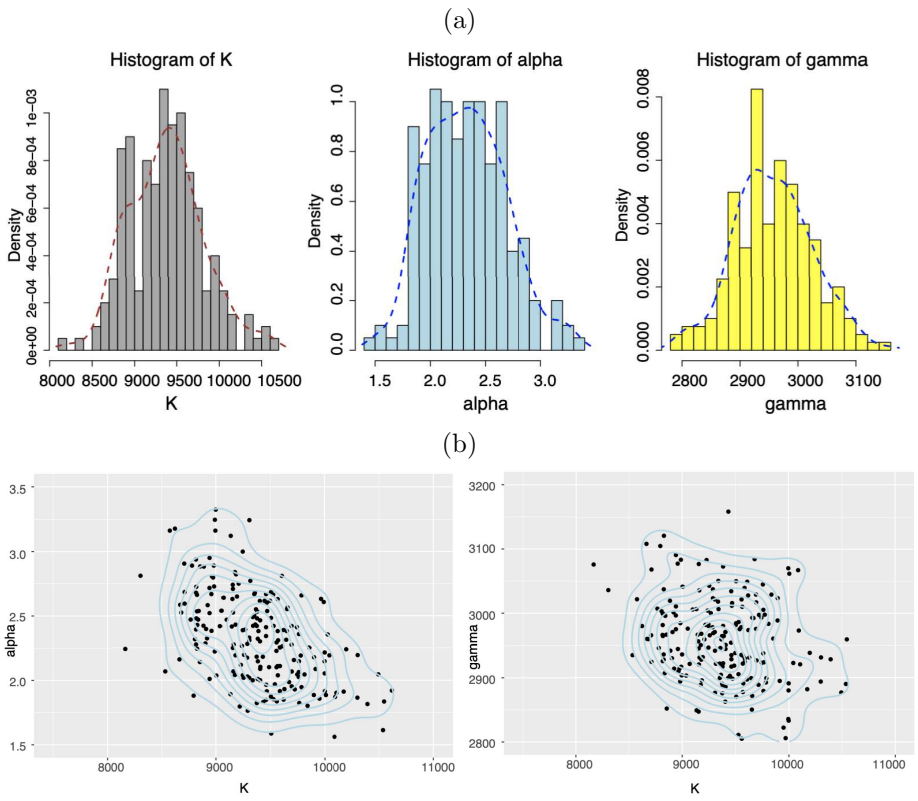

Figure 7 Simulation study V: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model with true $K = 10000, \gamma = 3000$ and $\alpha = 2$. Posterior kernel densities are presented in dash lines; (b) contour plots of K and α and γ for simulated data.

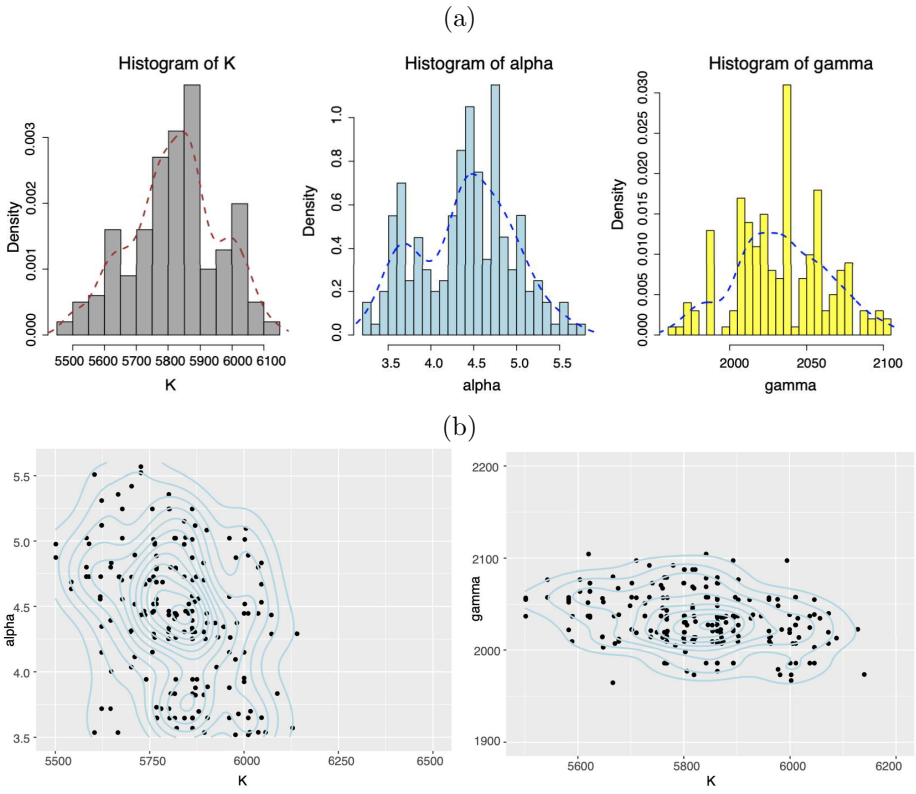

Figure 8 Simulation study VI: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model. True parameters values are $K = 6000, \gamma = 2000$ and $\alpha = 4$. Posterior kernel densities are presented in dash lines; (b) contour plots of K and α and γ for simulated data.

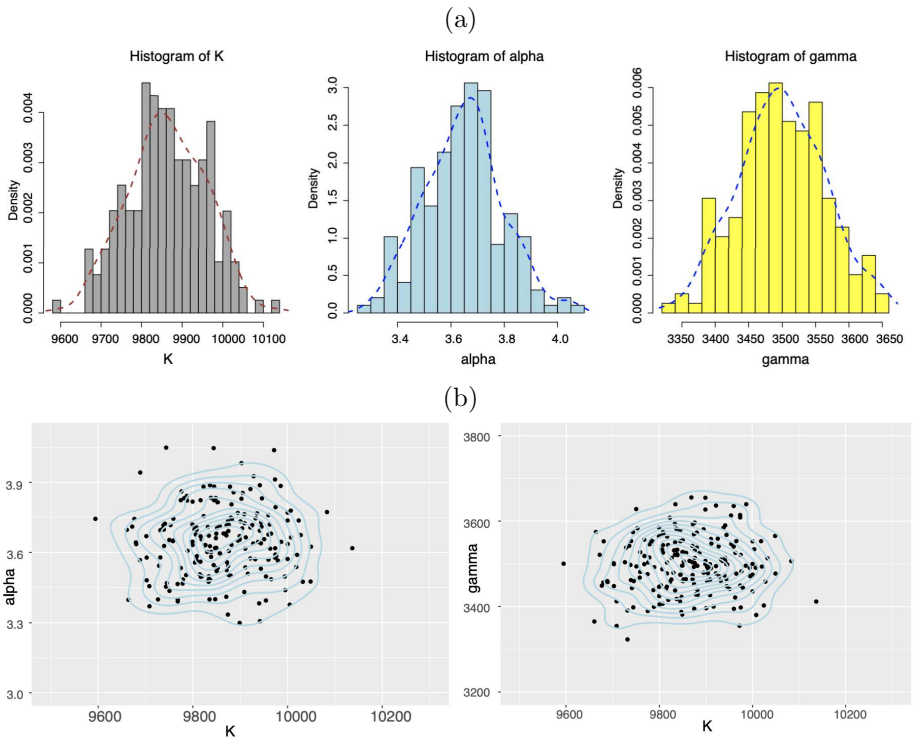

Figure 9 Real data application: (a) Bayesian estimates of real barcoding data parameters using asymmetric Dirichlet-multinomial model. Posterior kernel densities are presented in dash lines; (b) contour plots of K and α and γ for barcoding data.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
All estimation and simulation codes are available on request.

Competing interests
The authors declare that they have no competing interests.

Funding
SC, KYK and MK were supported in part by NIH grants (R01 HL136333, R01 HL134880) to KYK

Author's contributions
SC, KYK and MK contributed to the development and implementation of the mathematical result. SC conceived and designed the analysis, performed the data analysis, analyzed the data and produced the numerical results and plots using R. SC also wrote the paper, MK and KYK revised the paper for submission. SC, KYK and MK read and approved the final manuscript.

Acknowledgements
Not applicable .

Author details
[1] Department of Statistics, Rice University, 6100 Main St, 77005 Houston, USA. [2] Department of Pediatrics, Section of Infectious Diseases, Baylor College of Medicine , 1 Baylor Plaza, 77030 Houston, USA.

References

1. Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate bayesian computation in population genetics. Genetics **162**(4), 2025–2035 (2002)
2. Bissiri, P.G., Ongaro, A., Walker, S.G.: Species sampling models: consistency for the number of species. Biometrika **100**(3), 771–777 (2013). doi:10.1093/biomet/ast006. https://academic.oup.com/biomet/article-pdf/100/3/771/855078/ast006.pdf
3. Boender, C.G.E., Rinnooy Kan, A.H.G.: A multinomial Bayesian approach to the estimation of population and vocabulary size. Biometrika **74**(4), 849–856 (1987). doi:10.1093/biomet/74.4.849. https://academic.oup.com/biomet/article-pdf/74/4/849/786430/74-4-849.pdf
4. Bunge, J., Fitzpatrick, M.: Estimating the number of species: A review. Journal of the American Statistical Association **88**(421), 364–373 (1993). doi:10.1080/01621459.1993.10594330. https://doi.org/10.1080/01621459.1993.10594330
5. Bunge, J., Willis, A., Walsh, F.: Estimating the number of species in microbial diversity studies. Annual Review of Statistics and Its Application **1**(1), 427–445 (2014). doi:10.1146/annurev-statistics-022513-115654. https://doi.org/10.1146/annurev-statistics-022513-115654
6. Chao, A.: Nonparametric estimation of the number of classes in a population. Scandinavian Journal of Statistics **11**(4), 265–270 (1984)
7. Chao, A.: Estimating the population size for capture-recapture data with unequal catchability. Biometrics **43**(4), 783–791 (1987)
8. Chao, A.: An extension of chao's estimator of population size based on the first three capture frequency counts. Computational Statistics Data Analysis **55**(7), 2302–2311 (2011)
9. Del Moral, P., Doucet, A., Jasra, A.: Sequential monte carlo samplers. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(3), 411–436 (2006). doi:10.1111/j.1467-9868.2006.00553.x. https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2006.00553.x
10. Efron, B., Thisted, R.: Estimating the number of unsen species: How many words did shakespeare know? Biometrika **63**(3), 435–447 (1976)
11. Fisher, R.A., Corbet, A.S., Williams, C.B.: The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology **12**(1), 42–58 (1943)
12. Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**(430), 773–795 (1995). doi:10.1080/01621459.1995.10476572. https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476572
13. Marjoram, P.: Approximation bayesian computation. OA genetics **1**(3) (2013)
14. Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain monte carlo without likelihoods. Proceedings of the National Academy of Sciences **100**(26), 15324–15328 (2003). doi:10.1073/pnas.0306899100. https://www.pnas.org/content/100/26/15324.full.pdf
15. Ongaro, A.: Size-biased sampling and discrete nonparametric bayesian inference. Journal of Statistical Planning and Inference **128**(1), 123–148 (2005). doi:10.1016/j.jspi.2003.10.005
16. Pitman, J.: Combinatorial stochastic processes (1996)
17. Scott A. Sisson, M.B. Yanan Fan: Handbook of Approximate Bayesian Computation. New York: Chapman and Hall/CRC, ??? (2019)
18. Sisson, S.A., Fan, Y., Tanaka, M.M.: Sequential monte carlo without likelihoods. Proceedings of the National Academy of Sciences **104**(6), 1760–1765 (2007)
19. Sun, *et al.*: Clonal dynamics of native haematopoiesis. Nature **514**, 322–327 (2014)
20. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H.: Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface **6**(31),

187–202 (2009). doi:10.1098/rsif.2008.0172.
https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2008.0172

21. Zhang, H., Stern, H.: Investigation of a generalized multinomial model for species data. Journal of Statistical Computation and Simulation **75**(5), 347–362 (2005). doi:10.1080/0094965042000191631.
https://doi.org/10.1080/0094965042000191631

Figures

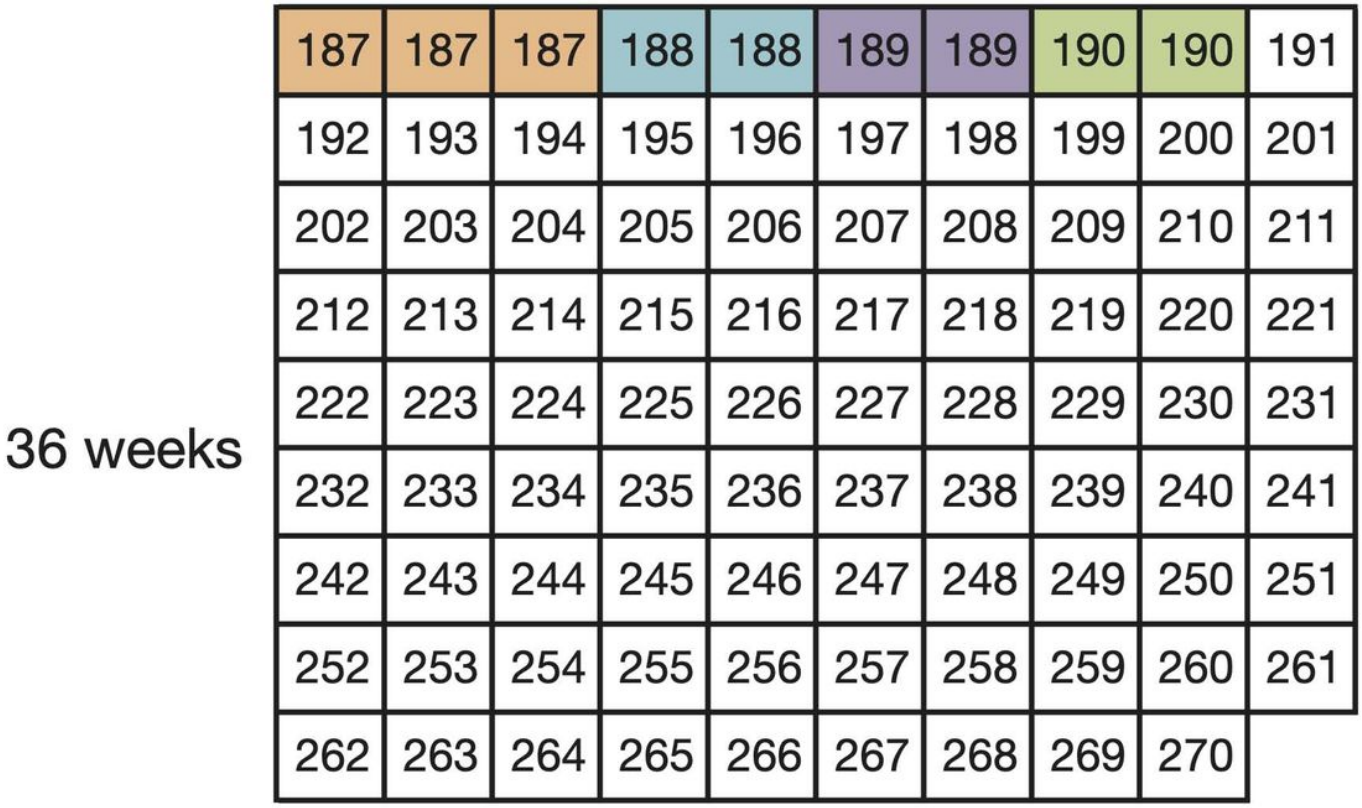

Figure 1

An example of Tn tags from barcoding results from Sun et al. (2014): In a plate of wells holding individual granulocytes, tag 187 (numbering is arbitrary) is observed three times, tags 188, 189, and 190 are observed twice, the rest of the tags are only observed once. The standard plate has 96 wells, but in some of them (7 in this case) tags were not identified.

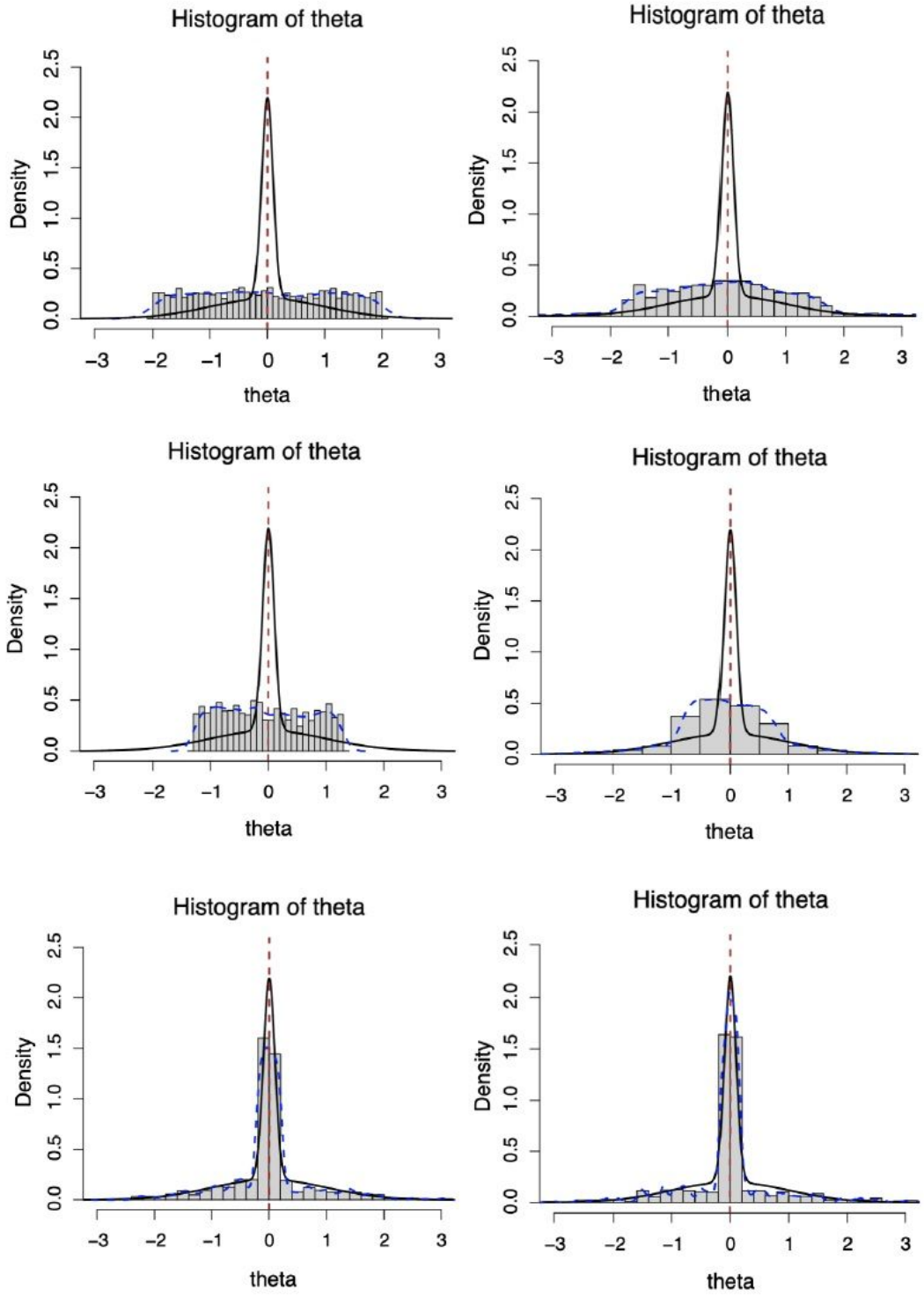

Figure 2

Histograms of six iterations of the adaptive truncated-population ABC given the Gaussian mixture model. Black lines represent true posterior densities, dashed blue lines are kernel density estimates from simulations.

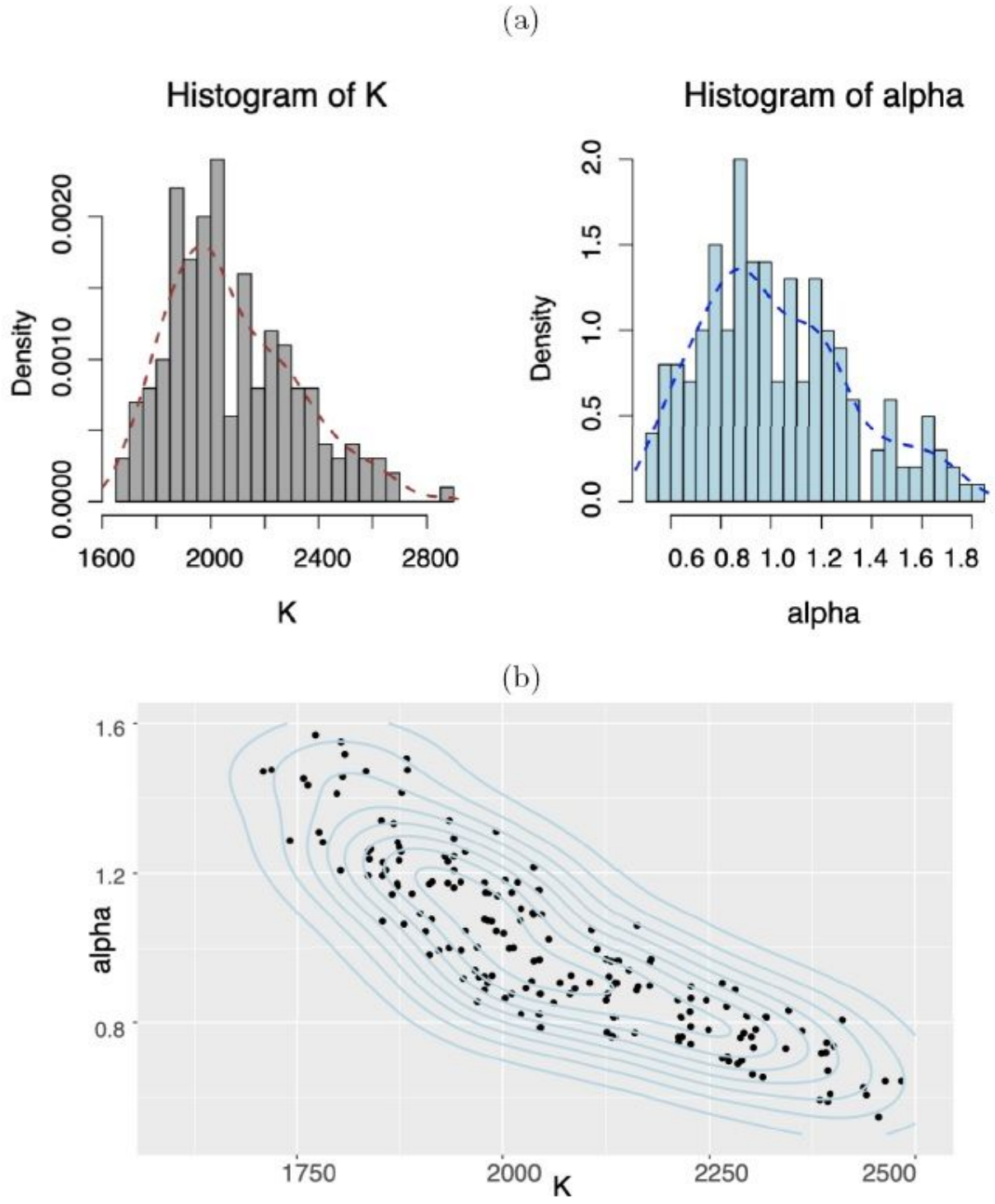

Figure 3

Simulation study I: (a) final posterior distributions of K and α using adaptive ABC algorithm under symmetric Dirichlet-multinomial model with true K = 2000 and α = 1; dash lines represent kernel densities of the posteriors; (b) contour plot of K and α for simulated data.

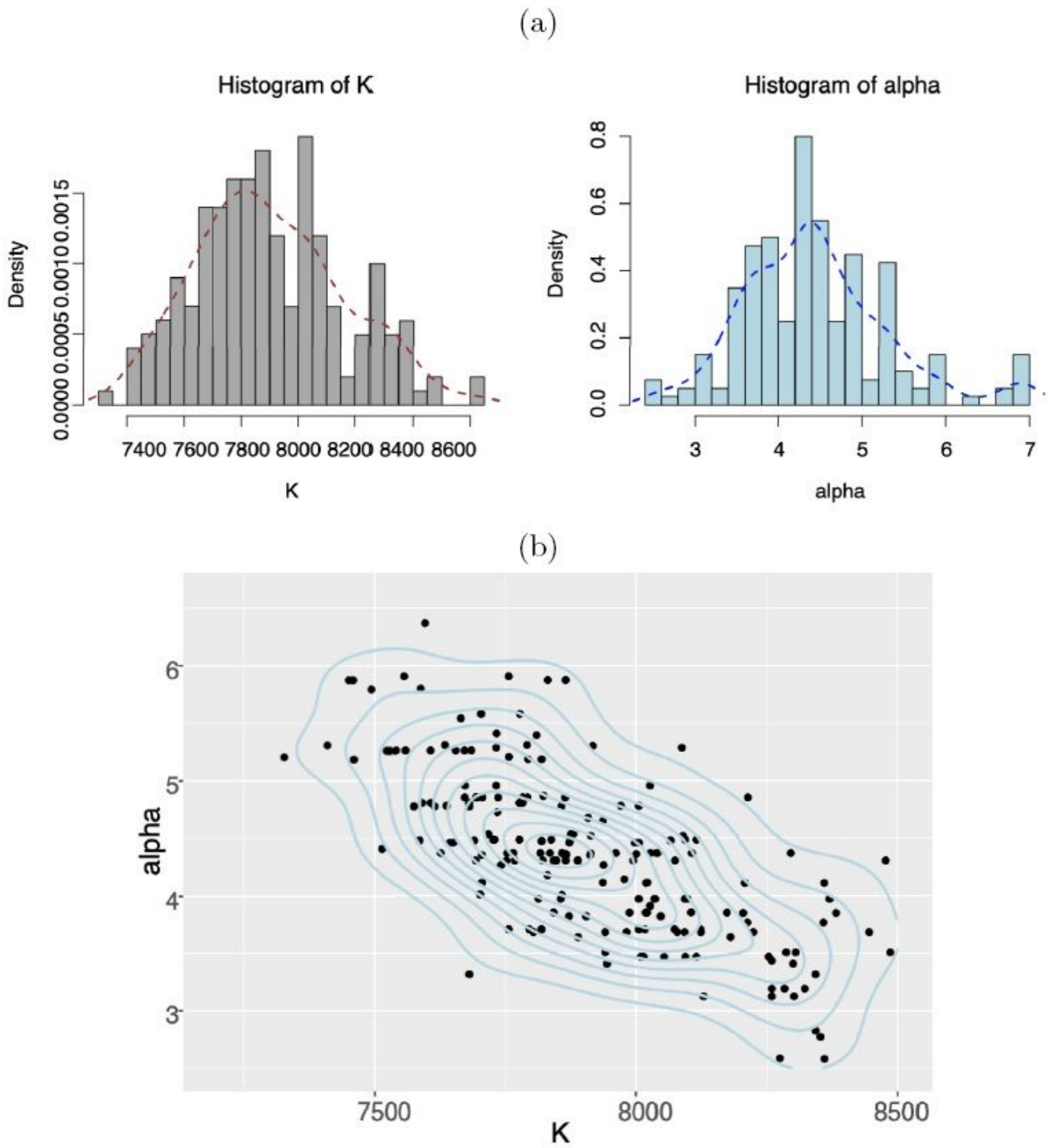

Figure 4

Simulation study II: (a) final posterior distributions of K and α using adaptive ABC algorithm with dash lines representing the kernel densities. The simulated data is from symmetric Dirichlet-multinomial model with true K = 8000 and α = 4; (b) contour plot of K and α for simulated data.

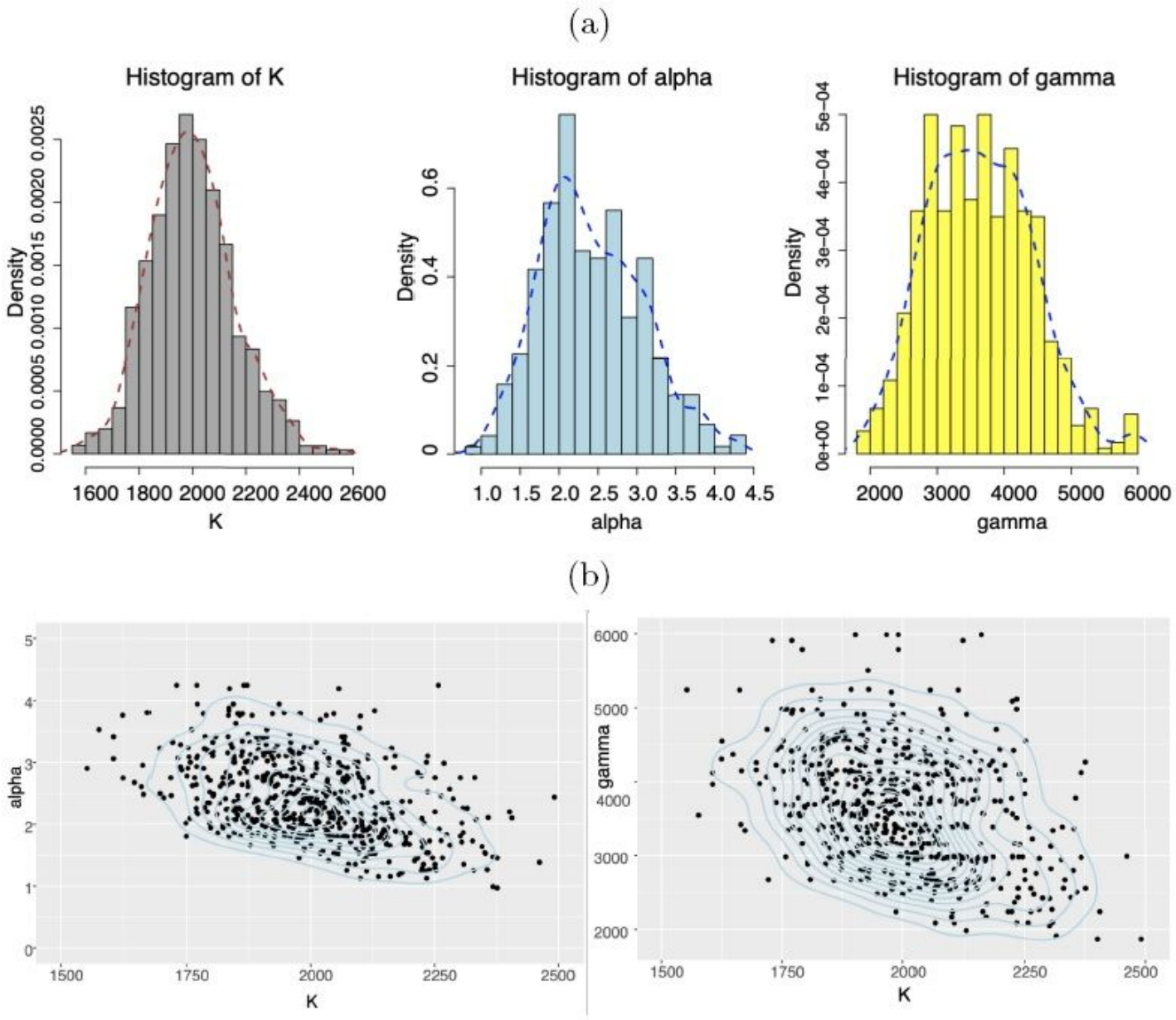

Figure 5

Simulation study III: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model with true K = 2000 with dash lines representing posterior kernel densities; (b) contour plots of K and α and γ for simulated data.

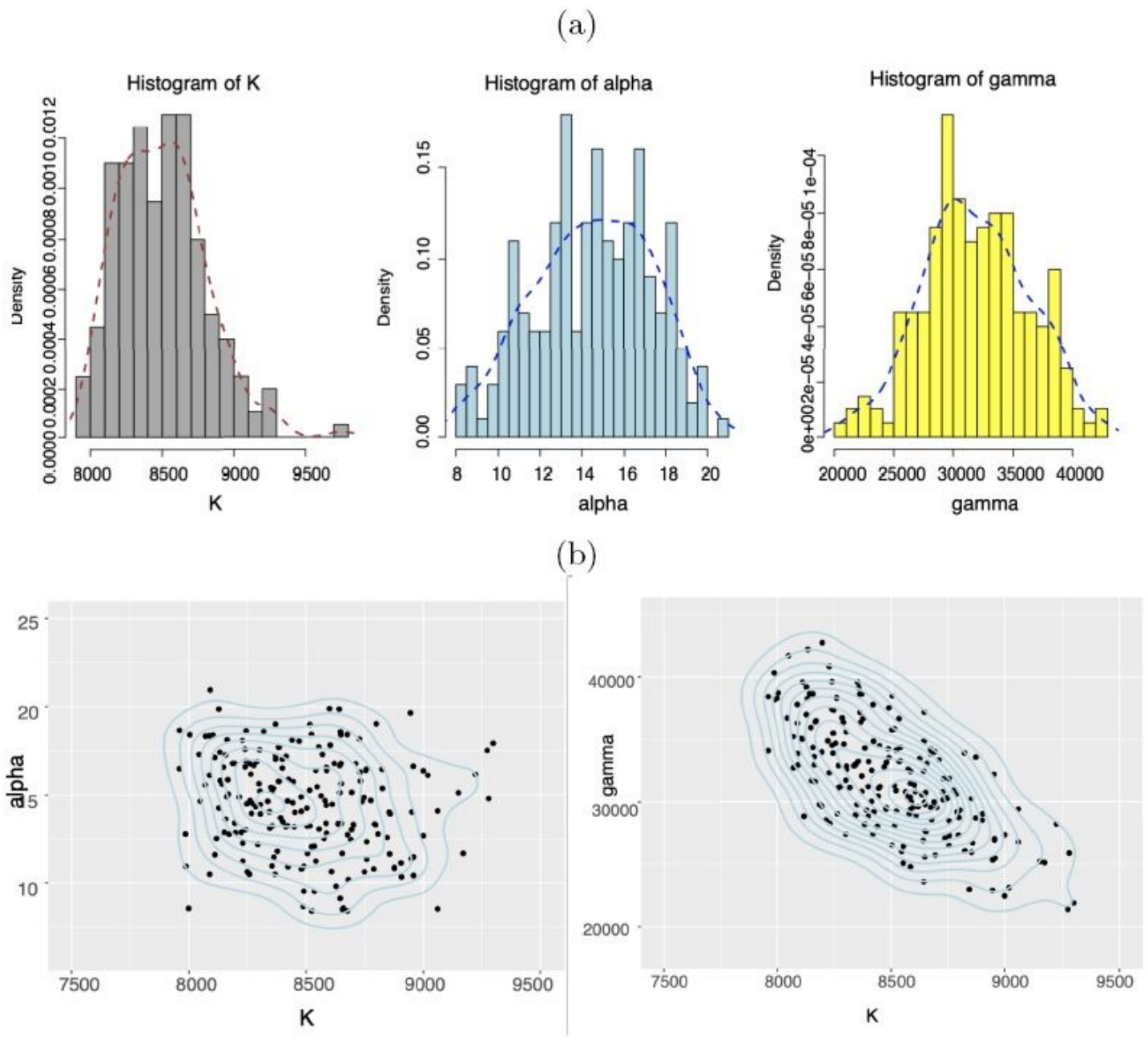

Figure 6

Simulation study IV: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model with true K = 8000 with kernel densities presented in dash lines; (b) contour plots of K and α and γ for simulated data.

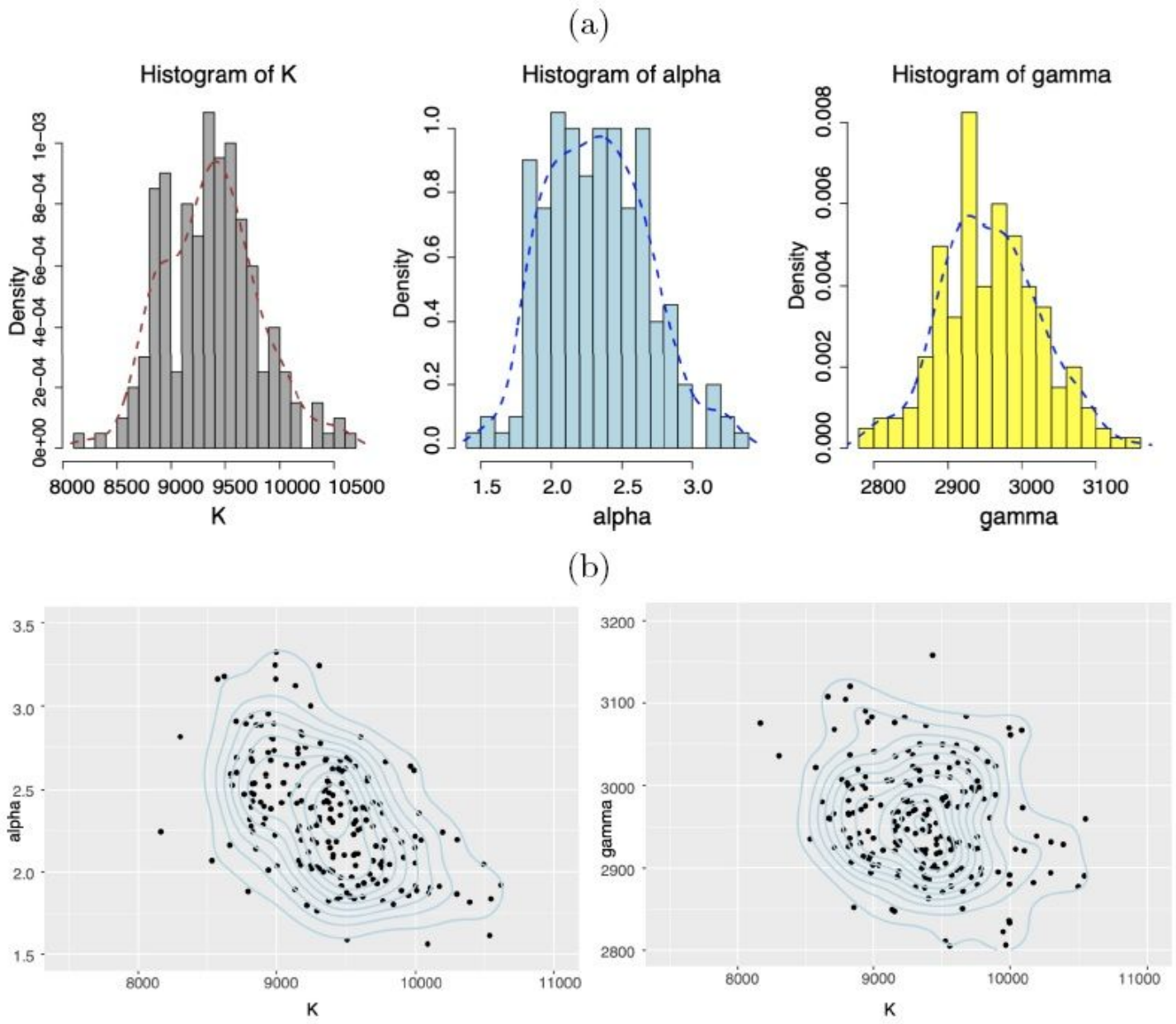

Figure 7

Simulation study V: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model with true K = 10000; γ = 3000 and α = 2: Posterior kernel densities are presented in dash lines; (b) contour plots of K and α and γ for simulated data.

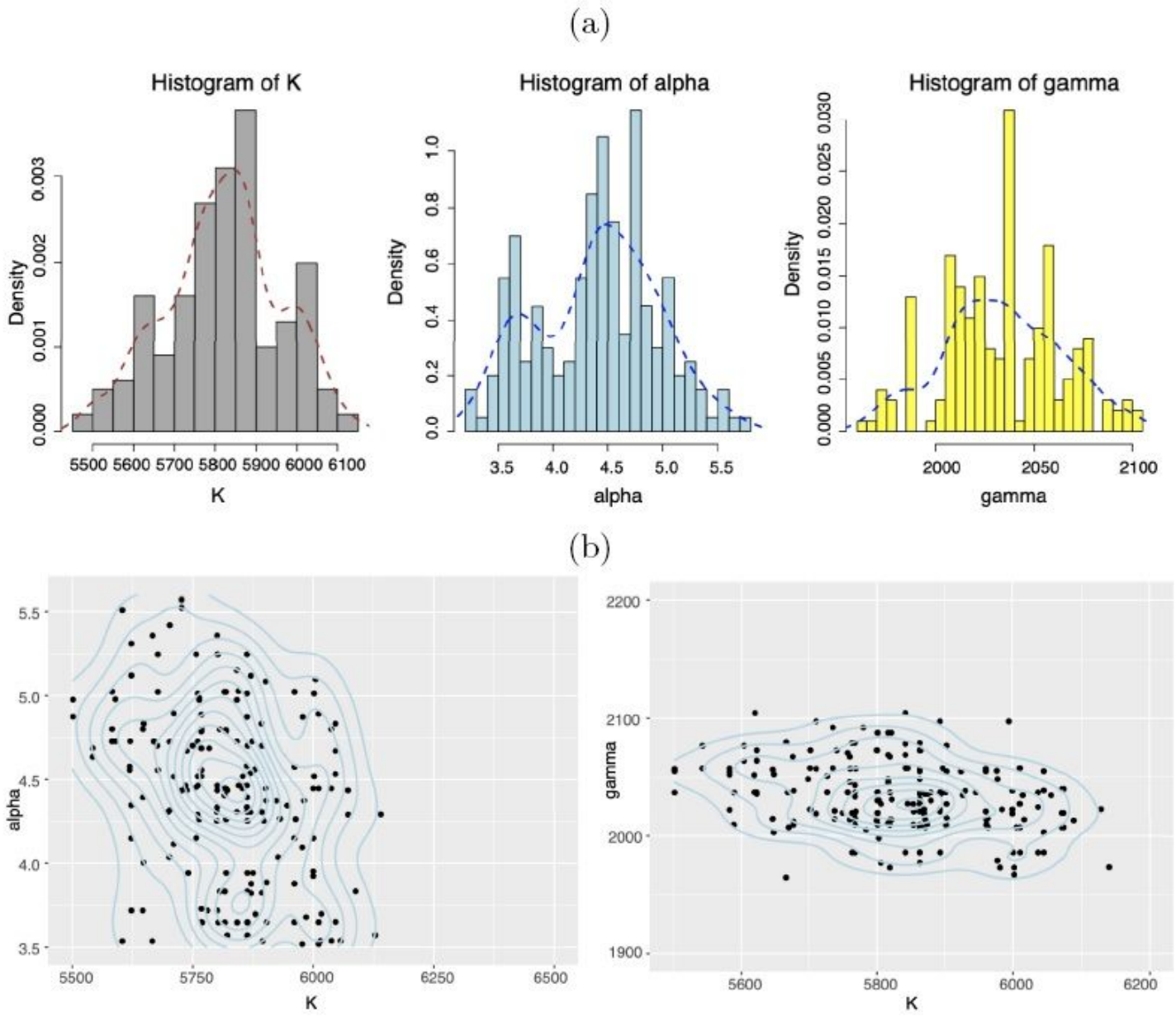

Figure 8

Simulation study VI: (a) final posterior distributions of parameters under asymmetric Dirichlet-multinomial model. True parameters values are K = 6000; γ = 2000 and α = 4: Posterior kernel densities are presented in dash lines; (b) contour plots of K and α and γ for simulated data.

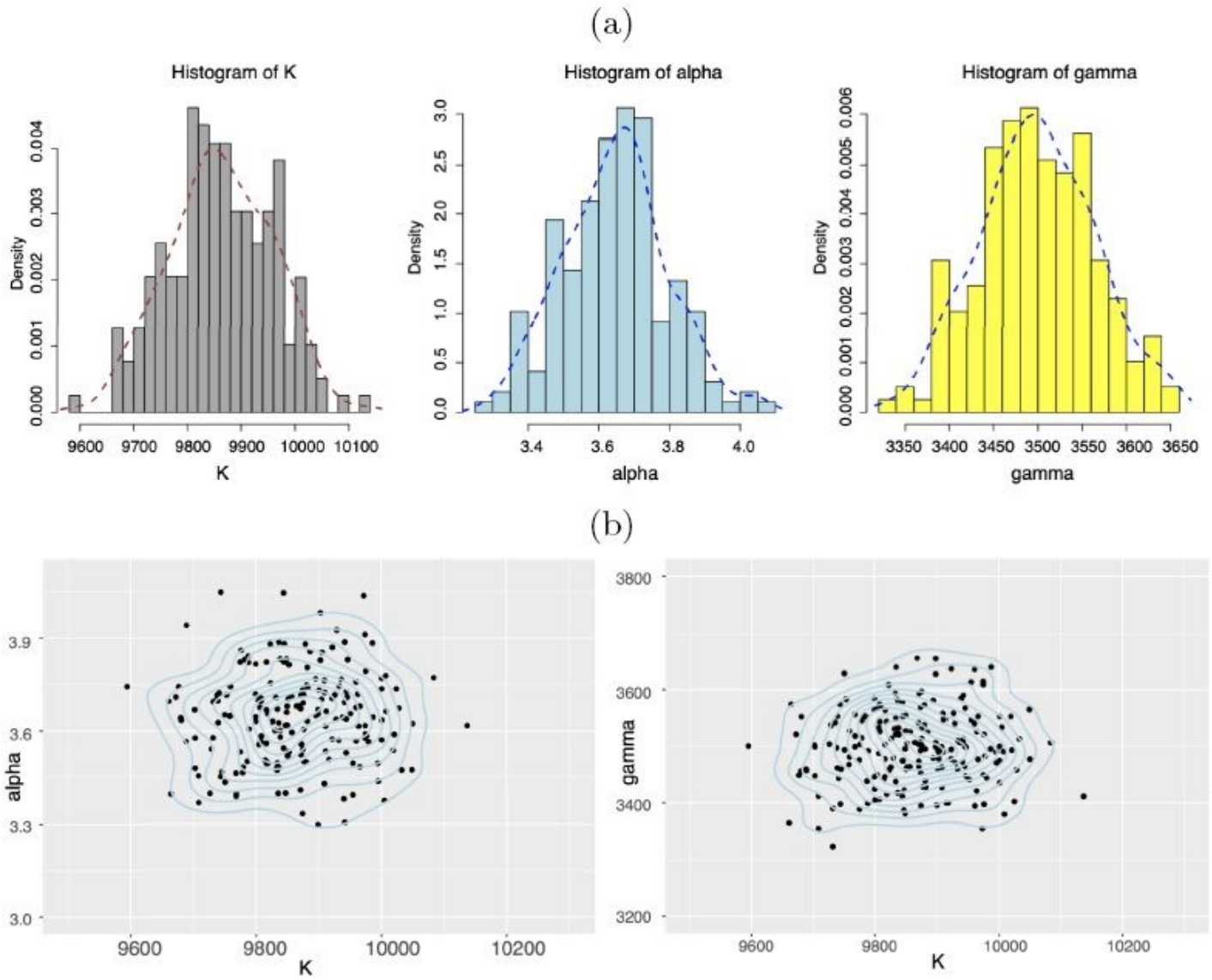

Figure 9

Real data application: (a) Bayesian estimates of real barcoding data parameters using asymmetric Dirichlet-multinomial model. Posterior kernel densities are presented in dash lines; (b) contour plots of K and α and γ for barcoding data.