Supporting Information

Model Details

As described in the main text, we write our model in terms of the marginal distributions for each category,

$$F(Y_i|\boldsymbol{\mu}, N, \phi) = \begin{cases} p\left(Y_i = 0|\boldsymbol{\mu}, N, \phi\right) & = (1-\mu_i)^{N\phi} \\ p\left(Y_i = N|\boldsymbol{\mu}, N, \phi\right) & = \left(1-(1-\mu_i)^{N\phi}\right)\prod_{j\neq i}(1-\mu_j)^{N\phi} \\ p\left(Y_i = y_i|0 < y_i < N, \boldsymbol{\mu}, N, \phi\right) & = \dfrac{y^{\kappa_i N\phi-1}\left(N-y\right)^{(1-\kappa_i)N\phi-1}}{N^{N\phi-1}B\left(\kappa_i N\phi, \left(1-\kappa_i\right)N\phi\right)}[1-(1-\mu_i)^{N\phi}- \\ & \quad \left(1-(1-\mu_i)^{N\phi}\right)\prod_{j\neq i}(1-\mu_j)^{N\phi}] \\ where & \\ \quad \kappa_i = \dfrac{\mu_i}{\sum_i \mathbf{I_i}\mu_i} & \end{cases} \tag{1}$$

Here $\mathbf{I_i}$ is an indicator function that takes on a value of 1 if $0 < Y_i < N$ and 0 otherwise. We can conceptualize the model as a two step process in which each species is determined to be absent ($Y_i = 0$) or comprise the entire sample ($Y_i = N$). Then conditioned on the categories that have observations $0 < Y_i < N$, each category's marginal distribution is a generalized beta distribution stretched over the interval (0, N) and jointly across all the categories the observations follow a generalized Dirichlet distribution stretched over the same interval. The final line of the equation renormalizes the μ_i based on the categories that are between 0 and N and ensures that $\sum_i \kappa_i = 1$. Thus the renormalization links all I categories into a single model and ensures that the total sample size observed across categories is fixed (i.e. $N = \sum_i Y_i$). While the model describes the marginal density for a single category, this is actually a joint model for all categories (see below).

This model utilizes a non-standard parameterization of the generalized beta distribution, using the parameters μ, ϕ, and N rather than the standard α, β parameterization of the beta distribution.

The above parameterization of the beta distribution can be connected using the following equalities,

$$\alpha_i = \mu_i N\phi \tag{2}$$

$$\beta_i = (1-\mu_i)N\phi \tag{3}$$

and the effective sample size is $\alpha_i + \beta_i = N\phi$. Unlike other parameterizations of mixture distributions involving the beta distribution [zero-inflated or zero- and one-inflated; Ospina and Ferrari (2012); Joseph et al. (2016); Liu (2021)], this parameterization uses a low-rank parameterization that links the discrete outcomes (0, N) and the continuous component using only the parameter vector $\boldsymbol{\mu}$ and a single scaling term ϕ.

Connections to multivariate distributions

Presence-absence model and the multinomial distribution

We note a strong connection between the presence-absence component of our mixture model and the multinomial model. We are motivated by the multinomial pmf which can be expressed using gamma functions as

$$g(y_1, y_2, ..., y_I; N, \mu_1, \mu_2, ..., \mu_I) = \frac{\Gamma(\sum_i y_i + 1)}{\prod_i \Gamma(y_i + 1)} \prod_{i=1}^{I} \mu_i^{y_i} \tag{4}$$

As above, μ_i is the probability of occurrence for species i and N be the known sample size. However, we are only interested in the cases where one or more observations (y_i) are zero. The marginal probability of observing zero in category i, $y_i = 0$ is easy to calculate by recognizing when $y_i = 0$ the remaining categories must sum to N and the multinomial collapses to two categories,

$$\begin{aligned} p(y_i = 0|n, \mu_1, \mu_2, ...\mu_I) &= \frac{\Gamma(\sum_i y_i + 1)}{\prod_i \Gamma(y_i + 1)} \prod_{i=1}^{I} \mu_i^{y_i} & (5) \\ &= \frac{\Gamma(N+1)}{\Gamma(1)\Gamma(N+1)} \mu_i^0 (1 - \mu_i)^N & (6) \\ &= (1 - \mu_i)^N & (7) \end{aligned}$$

which has clear similarity to the marginal distribution presented above and in the main text. Simply substitute $N\phi$ for N; using the gamma function form of the multinomial, there is no requirement that the y_i take on integer values. The probability the entire sample size is in a single category in the multinomial distribution is $p\,(y_i = N|\mu_i, N) = \mu_i^N$, which differs from the probability in our model. Thus the model presented in the main text and eq. S1 has a marginal probability $p\,(Y_i = 0|\mu_i, N, \phi)$ comparable to the marginal probability of the multinomial model with equivalent μ_i and sample size $N\phi$ but a different marginal probability for all observations occurring in a single category, $p\,(Y_i = N|\mu_i, N, \phi)$ (main text and eq. S1). While we would prefer to use a multinomial-like model, in the following we illustrate the computational complexities that arise from trying to use a multinomial-like structure for the two discrete outcome ($Y_i = 0$ and $Y_i = N$) generally but show how this structure can be implemented directly when there are relatively few categories.

The issues with the multinomial model occur when trying to calculate the probability of all possible combinations of zero observations. We know that if $I - 1$ categories are 0, the remaining categories observations must sum to N. We can calculate the probability of all possible combinations of categories being present or absent using the probability of each category being absent. We let $\boldsymbol{j}$ index the J species that are absent $\boldsymbol{j} = \{j_1, j_2, ..., j_J\}$ so the probability of observing all J species absent conditioned on the parameters and samples size is the sum of the probabilities for the relevant categories,

$$p\,(\boldsymbol{y_j} = 0|\boldsymbol{\mu}, N, \phi) = \left(1 - \sum_{\boldsymbol{j}} \mu_{\boldsymbol{j}}\right)^{N\phi}. \tag{8}$$

This is the probability of observing all $\boldsymbol{j}$ species absent but includes all possible combinations of presence and absence of the remaining $I - J$ species. To calculate the probability of exactly the set of $\boldsymbol{j}$ being equal to zero and the remaining species being non-zero, we have to enumerate and substract the probabilities associates with one or more of the $I - J$ remaining categories being exactly zero. We define $\boldsymbol{k}$ as the vector of elements $\boldsymbol{k} = \{k_1, k_2, ..., k_{I-J}\}$ such that $\boldsymbol{j} \cap \boldsymbol{k} = \emptyset$. We also define an indicator matrix $\boldsymbol{M}$ where each entry contains either a zero or one and each row is of length $I - J$ and comprises a particular combination in which $y_k = 0$ ($\boldsymbol{M}_{.,k} = 1$) or $y_k > 0$ ($\boldsymbol{M}_{.,k} = 0$). Then $\boldsymbol{M}$ has $L = 2^{I-J} - 2$ rows (there are two possible combinations which are not included: the probability all categories are absent [which occurs with probability 0] and the probability all other categories are present [the quantity of interest]).

$$p\left(\boldsymbol{y}_{\boldsymbol{j}}=0,\boldsymbol{y}_{\boldsymbol{k}}>0|\boldsymbol{\mu},N,\phi\right)=\left(1-\sum_{j=1}^{J}\mu_j\right)^{N\phi}-\sum_{l=1}^{L}\left(1-\sum_{m=1}^{I-J}\boldsymbol{M}_{l,m}\mu_{k_m}\right)^{N\phi} \tag{9}$$

While explicitly calculating these probabilities is not difficult when there are only a few categories, as the number of categories grow the number of unique combinations of presence and absence rapidly expand. It is computationally more efficient to write the right hand side of the equation using inner products but there are still more computations and thus slower calculations than is desirable.

Note that when there is only one category, Y_k, that is non-zero (so $I-J=1$) the second term on the right hand side above is zero and

$$\begin{aligned}
p\left(\boldsymbol{Y}_{\boldsymbol{j}}=0,Y_k>0|\boldsymbol{\mu},N,\phi\right) &= \left(1-\sum_{j=1}^{J}\mu_j\right)^{N\phi} && (10)\\
&= \left(1-(1-\mu_k)\right)^{N\phi} && (11)\\
&= \mu_k^{N\phi} && (12)
\end{aligned}$$

which matches the probability $p\left(Y_k=N|\boldsymbol{\mu},N,\phi\right)$ in the multinomial model above as it should.

As a simple worked example, imagine a situation where there are five possible categories so $\boldsymbol{\mu}=\{\mu_1,\mu_2,\mu_3,\mu_4,\mu_5\}$ and are interested in calculating the probability of an observation in which $\boldsymbol{Y}=\{Y_1=0,Y_2>0,Y_3=0,Y_4>0,Y_5>0\}$. Then $\boldsymbol{j}=\{1,3\}$ and $\boldsymbol{k}=\{2,4,5\}$ and

$$\boldsymbol{M}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\\1&1&0\\1&0&1\\0&1&1\end{bmatrix} \tag{13}$$

In this case the are only six rows in $\boldsymbol{M}$ and the subsequent calculations are trivial. When there are many more categories, however, the computational burden in evaluating all possible combinations is prohibitive. In practice, we have found that explicitly evaluating all of these probabilities only is practicable when there are fewer than 20 categories. Specifically, the computations are exponential in the number of categories; it is $O(2^K)$. In the next section we detail some potential approaches to speed calculations. We note that our problem here is closely related to some classical probability problems (Holst 1986, O'Neill 2021) and this literature may provide additional algorithms to speed calculations further. At present, we use the form in the main text that treats the probability of occurrence as independent among categories and eliminates the need to explicity calculate the probabilities described in this section. This is an approximate solution made for computational reasons that provides reasonable performance.

Alternative methods for computing the probability of zeroes in a multinomial.

In addition to the approach described above, below we outline a potential, recursive approach to calculating the desired probability which is $O(KN^2)$ in computational complexity rather than the $O(2^K)$ described above. This approach would be akin to certain common calculations on hidden Markov chains (e.g., the Baum-Welch algorithm). It involves cycling over the different components for the multinomial random vector, and for each one accumulating a probablity vector that gives the probability of the sum of each component up to and including the current value.

Some new notation should help to express this. Note that, because we intend to implement this in C^{++} we will index the K components of $\boldsymbol{Y}$ from 0 to $K-1$ (because vectors are indexed starting with zero rather than one in C^{++}). First, let $\breve{\mu}_i$ denote the cell probability μ_i scaled so that $\sum_{j=i}^{K-1} = 1$. Thus, we have

$$\breve{\mu}_i = \mu_i \left(\sum_{j=i}^{K-1} \mu_j \right)^{-1} = \mu_i \left(\prod_{j=0}^{i-1} (1-\mu_j) \right)^{-1}. \tag{14}$$

Now, think about simulating the multinomial random vector in a sequential fashion, as a series of binomials, each one conditioned on the results of the preceding ones. Then, for $i = 0, 1, \ldots, K-1$, let $p(n,i)$ denote the probability, after values $y_0, \ldots, y_i$ have been realized/simulated, that their sum is n. In other words:

$$p(n,i) = P\left(\sum_{j=0}^{i} Y_j = n \right) \tag{15}$$

If we further allow ourselves to condition the outcome upon the occurrence that none of the $Y_i = 0$, then we also have as definitions

$$p(0,i) \equiv 0 \text{for} \quad i = 0, 1, \ldots, K-1 \tag{16}$$

$$p(n,i) \equiv 0 \text{if} \quad N - n < K - 1 - i \tag{17}$$

The last condition simply states that the sum of the first i components of the multinomial vector cannot be so great that there is no way for the remaining $K-i$ components to each have at least the value of 1. This can be rearranged to say that $p(n,i) \equiv 0$ whenever $n > N - K + i + 1$

For the initial case of $i = 0$ it should be clear that $p(n,0)$ is the probability that $Y_0 = n$, for $n > 0$ and $n \leq N - K + i + 1$:

$$p(n,0) = \frac{N!}{n!(N-n)!} \mu_0^n (1-\mu_0)^{N-n} \quad , \quad n = 1, \ldots, N-K+1. \tag{18}$$

For all remaining cases, with $i > 0$, the values of $p(n,i)$ can be found with the following recursion:

$$p(n,i) = \sum_{L_r \leq r \leq H_r} p(r, i-1) er \frac{(N-r)!}{(n-r)!(N-n)!} \breve{\mu}_i^{n-r} (1-\breve{\mu}_i)^{N-n} \quad , \quad L_n \leq n \leq H_n, \tag{19}$$

where L_n, H_n, L_r, and H_r are defined as follows. Since we will be cycling over values of n, we will start with the global definitions for L_n and H_n, and then we will define L_r and H_r, conditional on n.

$$\begin{aligned} L_n &= N, \text{ if } i = K-1; \quad i+1, \text{ otherwise} \\ H_n &= N - K + i + 1 \\ L_r &= i \\ H_r &= n - 1 \end{aligned}$$

We can rewrite that recursion as:

$$p(n,i) = \sum_{L_r \le r \le H_r} p(r, i-1)\mathcal{B}(r, n, \breve{\mu}_i) \quad , \quad L_n \le n \le H_n, \tag{20}$$

where,

$$\mathcal{B}(r, n, \breve{\mu}_i) = \frac{(N-r)!}{(n-r)!(N-n)!} \breve{\mu}_i^{n-r}(1-\breve{\mu}_i)^{N-n} \tag{21}$$

and the value of N is considered fixed (as it is). Note that $\mathcal{B}$ is simply the probability that a binomial random variable of size $N-r$ trials and success probability $\breve{\mu}_i$ has $n-r$ successes.

In conclusion, this shows a way to compute the probability that all the components of a multinomial random vector are greater than 0. It scales linearly in K, and quadratically in N. For large N, this approach may still be too computentially intensive to incorporate in an estimation model. But, this implementation makes it easy to calculate values to compare the approximation used in zoid to the exact value in a faster manner. Future work to speed computation may be gained by reclaiming a constant factor by computing the binomial probabilities recursively. Note that to be fully general to really large N and small μ, some sort of underflow protection would also likely be important.

Non-zero components and the scaled Dirichlet distribution

The previous section discusses the mixture components that specify the probability of zero and N observations. Here, we describe the connection between the continuous component of the mixture distribution as a scaled Dirichlet distribution. Specifically, the marginal distributions of the Dirichlet distribution are beta distributions so the marginal beta distribution for observations $0 < Y_i < N$ presented in the main text correspond to a scaled Dirichlet distribution.

Let $\boldsymbol{X}$ be drawn from a Dirichlet distribution with parameter vector $\boldsymbol{a} = (\alpha_1, \ldots, \alpha_I)$, and use $\alpha_\bullet$ to denote $\sum_{i=1}^{I} \alpha_i$. Then, for the categories in which the observations are not zero, the vector $\boldsymbol{Y}$ can be seen to have a scaled Dirichlet distribution. Let $\boldsymbol{Y} = N\boldsymbol{X}$. The inverse transformation is $\boldsymbol{X} = u(\boldsymbol{Y}) = \boldsymbol{Y}/N$, a transformation whose Jacobian is an $(I-1) \times (I-1)$ matrix with diagonal elements of $1/N$, and 0s elsewhere, such that its determinant is $N^{-(I-1)}$. It follows that the density of $\boldsymbol{Y}$ can be written as

$$p(\boldsymbol{Y}|\boldsymbol{a}, N) = \frac{\Gamma(\alpha_\bullet)N^{-(I-1)}}{\prod_{i=1}^{I}\Gamma(\alpha_i)} \prod_{i=1}^{I} (y_i/N)^{\alpha_i - 1} \quad , \quad 0 < y_i < N \text{ and } \sum y_i = N,$$

This leads to the marginals in Equation 1 with $\alpha_i = \kappa_i N \phi$.

Simulation Method Details

Given mixture proportions $\boldsymbol{\mu}$, sample size N, and overdispersion ϕ, we first determine whether categories have observations of 0 using independent Bernoulli draws, removing sets of simulated values in which all categories are zero. Mixture observations in which only one category is non-zero necessarily represent observations of N for the lone non-zero category. Then, for the categories that were non-zero, the proportion of N falling into each category is determined using the stick-breaking algorithm for generating Dirichlet observations from beta distributed marginal distributions [see Gelman et al. (2014); page 583] . Thus for all non-zero categories, we can generate a realized proportion p_i – which across all I categories follows a Dirichlet distribution – and the observed values in each non-zero category are then $y_i = p_i N$. For each simulation, this maintains a fixed $\sum_i y_i = N$ and the appropriate marginal distributions for each category. Additional overdispersion is incorporated into simulations by decreasing the expected sample size by some specified amount (e.g., 50%) and subsequently increasing variability in simulated observations. The R script for simulating proportional data is provided in the `zoid` package.

Available Bias Correction

Under the basic model specification, model predicted values for the probability of 0 and N observations are biased relative to realized frequencies of these observations from simulations. Because simulations reject observations in which all categories are 0, observations are normalized by the probability of all categories being 0 ($\prod_i(1-\mu_i^{N\phi})$) while model expectations are not. This results in simulated frequencies of 0 observations lower than naive model expectations and higher frequencies of N observations; this effect is most pronounced at low effective sample sizes (e.g., $N\phi<5$) (Fig. S1). At high sample sizes, probabilities of 0 or N observations become so small that the influence of this bias becomes negligible.

To address this situation-specific bias, we also provide a modified model, implemented in Stan, that normalizes probabilities of 0 and N observations based on the probability of all categories being 0. Specifically, we define category-specific values p_i equal to $(1-\mu_i)^{N\phi}$; these represent the uncorrected probabilities of 0 observations for each category. Then, corrected probabilities are expressed as follows,

$$F(Y_i|\mu_i,N,\phi)=\begin{cases} p\left(Y_i=0|\mu_i,N,\phi\right) & =\dfrac{p_i(1-\prod_{j\neq i}p_j)}{(1-\prod_i p_i)} \\ p\left(Y_i=N|\mu_i,N,\phi\right) & =\dfrac{(1-p_i)\prod_{j\neq i}p_j)}{(1-\prod_i p_i)} \\ p\left(Y_i=y|0<Y_i<N,\boldsymbol{\mu},N,\phi\right) & =\dfrac{y^{\kappa_i N\phi-1}\left(N-y\right)^{(1-\kappa_i)N\phi-1}}{N^{N\phi-1}B\left(\kappa_i N\phi,\left(1-\kappa_i\right)N\phi\right)}(1-\dfrac{p_i(1-\prod_{j\neq i}p_j)}{(1-\prod_i p_i)}- \\ & \dfrac{(1-p_i)\prod_{j\neq i}p_j)}{(1-\prod_i p_i)}) \\ & \\ where & \\ \qquad \kappa_i=\dfrac{\mu_i}{\sum_i \mathbf{I_i}\mu_i} & \end{cases} \tag{22}$$

We further demonstrate that this correction has negligible effect on model performance, compared to the basic model, over the ranges of sample sizes considered in our model evaluation (Fig. S2).

References

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. Bayesian data analysis. Third edition.

Holst, L. 1986. On birthday, collectors', occupancy and other classical urn problems. International Statistical Review 54:15–27.

Joseph, M. B., D. L. Preston, and P. T. J. Johnson. 2016. Integrating occupancy models and structural equation models to understand species occurrence. Ecology 97:765–775.

Liu, F. 2021. Zoib: Bayesian inference for beta regression and zero-or-one inflated beta regression.

Magnussen, E. 2011. Food and feeding habits of cod (gadus morhua) oon the faroe bank. ICES Journal of Marine Science 68:1909–1917.

O'Neill, B. 2021. The classical occupancy distribution: Computation and approximation. The American Statistician 75:364–375.

Ospina, R., and S. L. Ferrari. 2012. A general class of zero-or-one inflated beta regression models. Computation Statistics & Data Analysis 56:1609–1623.

Satterthwaite, W. H., J. Ciancio, E. Crandall, M. L. Palmer-Zwahlen, A. M. Grover, M. R. O'Farrell, E. C. Anderson, M. S. Mohr, and J. C. Garza. 2015. Stock composition and ocean spatial distribution from california recreational chinook salmon fisheries using genetic stock identification. Fisheries Research 170:166–178.

Tables

Table S 1: Prey types encountered and modeled from cod diet data obtained from (Magnussen 2011).

Prey ID #	Scientific Name
1	Hydrozoa
2	Anthozoa
3	Scleractinia
4	Polychaeta
5	Gastropoda
6	Bivalvia
7	*Loligo forbesi*
8	Crustacea spp
9	*Meganyctiphanes sp*
10	*Pandalus montagui*
11	*Lithodes maja*
12	*Munida sp*
13	*Galathea intermedia*
14	*Galathea sp*
15	*Galathea nexa*
16	*Hyas coarctatus*
17	*Hyas sp*
18	*Inachus leptochirus*
19	*Atelecyclus rotundatus*
20	*Liocarcinus holsatus*
21	*Liocarcinus turbeculatus*
22	*Ophiura sp*
23	*Amphiura sp*
24	Pisces spp
25	*Ammodytes sp*
26	*Ammodytes tobianus*
27	*Callionymus lyra*
28	Pleuronectidae
29	Stones

Figures

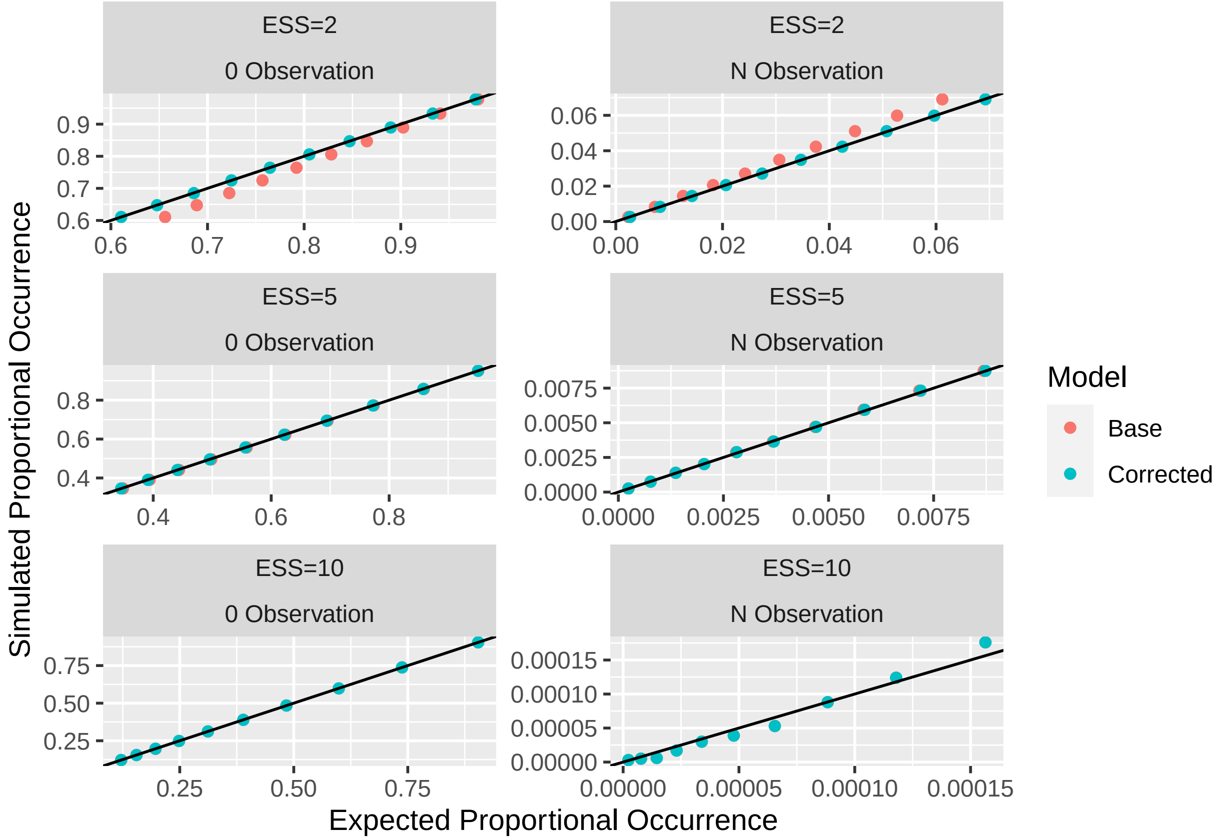

Figure S 1: Expected and simulated proportional occurrences of 0 and N observations with the base and bias corrected model structures. The black line represents a 1:1 relationship between simulated and expected values. Simulation inputs featured 1e6 iterations, 10 categories, effective sample sizes (ESS) varying from 2, 5, and 10, and category proportions equal to: (0.01,0.03,0.05,0.07,0.09,0.11,0.13,0.15,0.17,0.19).

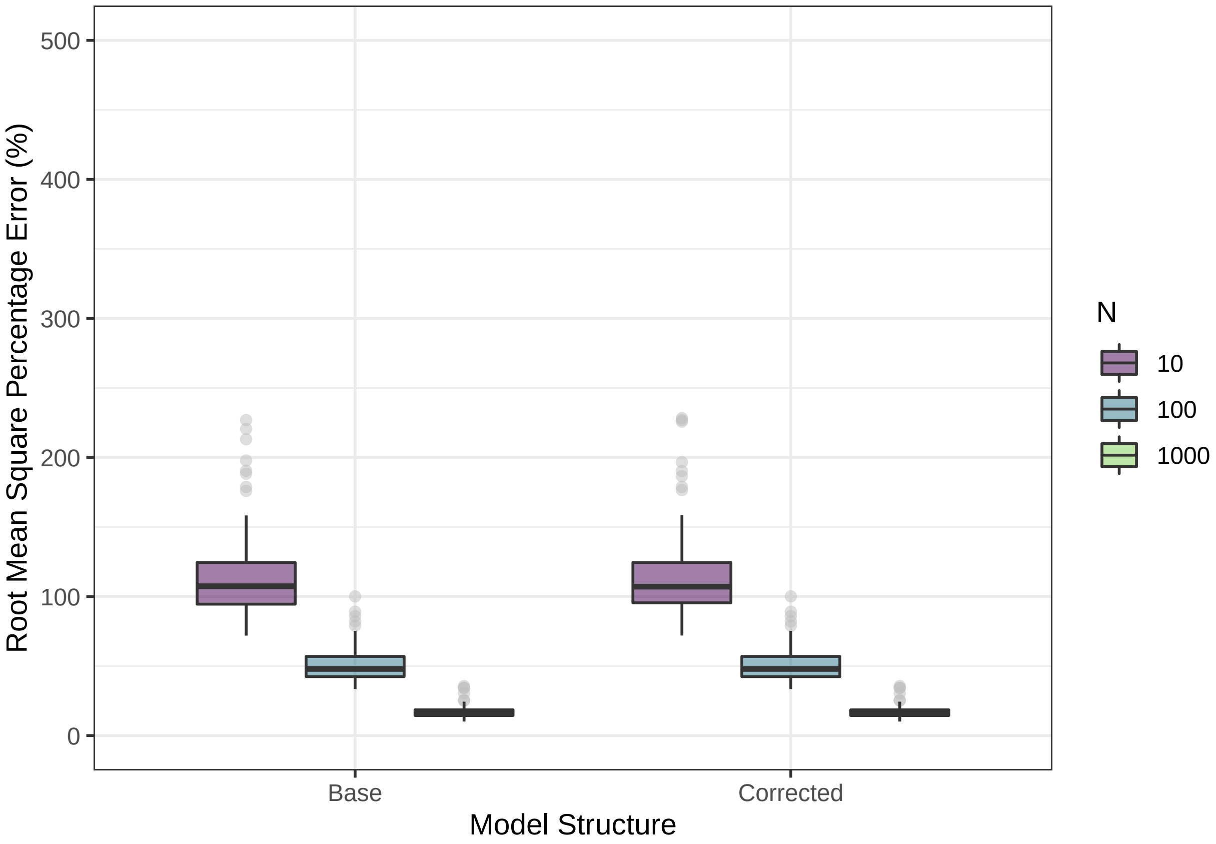

Figure S 2: Boxplots of root mean square percentage error (RMSPE) values calculated from simulated mixture proportions and MCMC iteration-specific predicted values for simulation scenario #1 and varying sample sizes for the base and bias-corrected model structures. Grey points represent outliers.

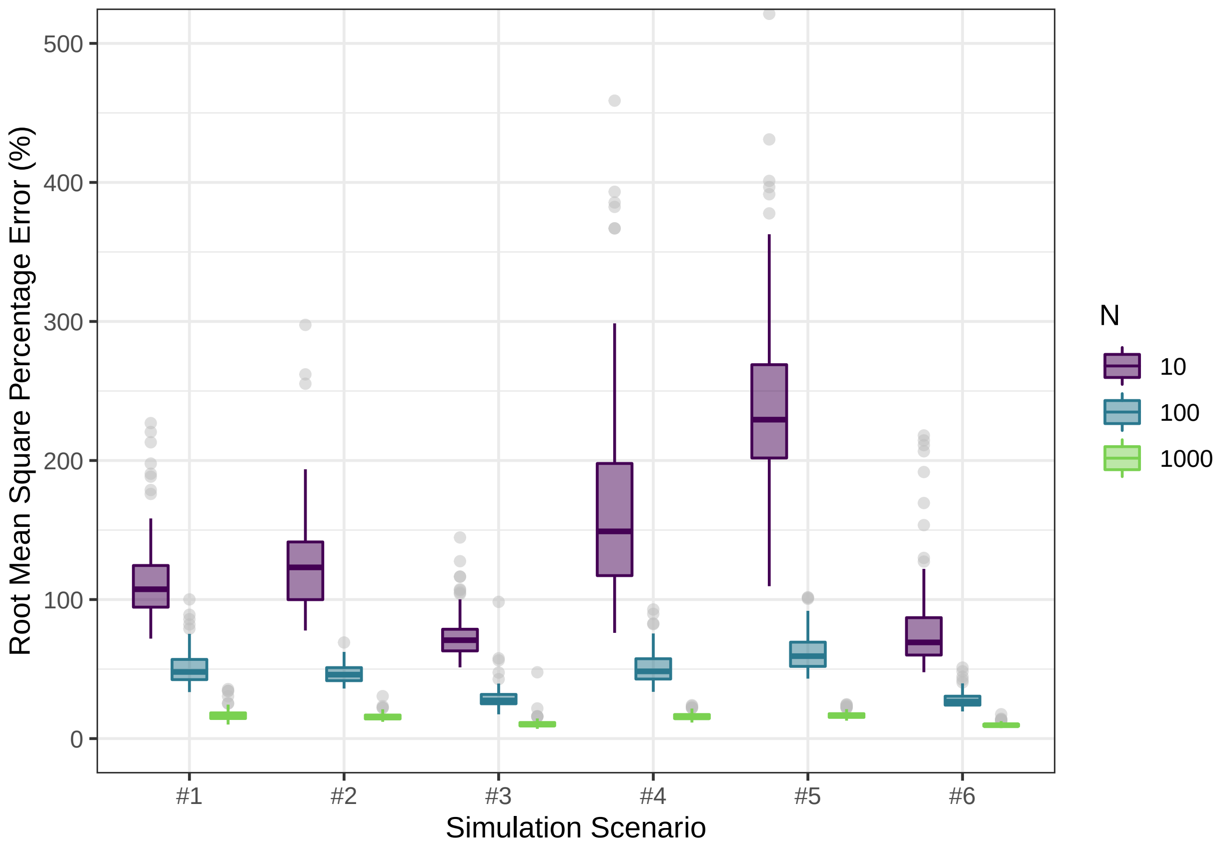

Figure S 3: Boxplots of root mean square percentage error (RMSPE) values calculated from simulated mixture proportions and MCMC iteration-specific predicted values for all simulation scenarios and sample sizes. As sample size increases, RMSPE asymptotically goes toward zero. Grey points represent outliers.

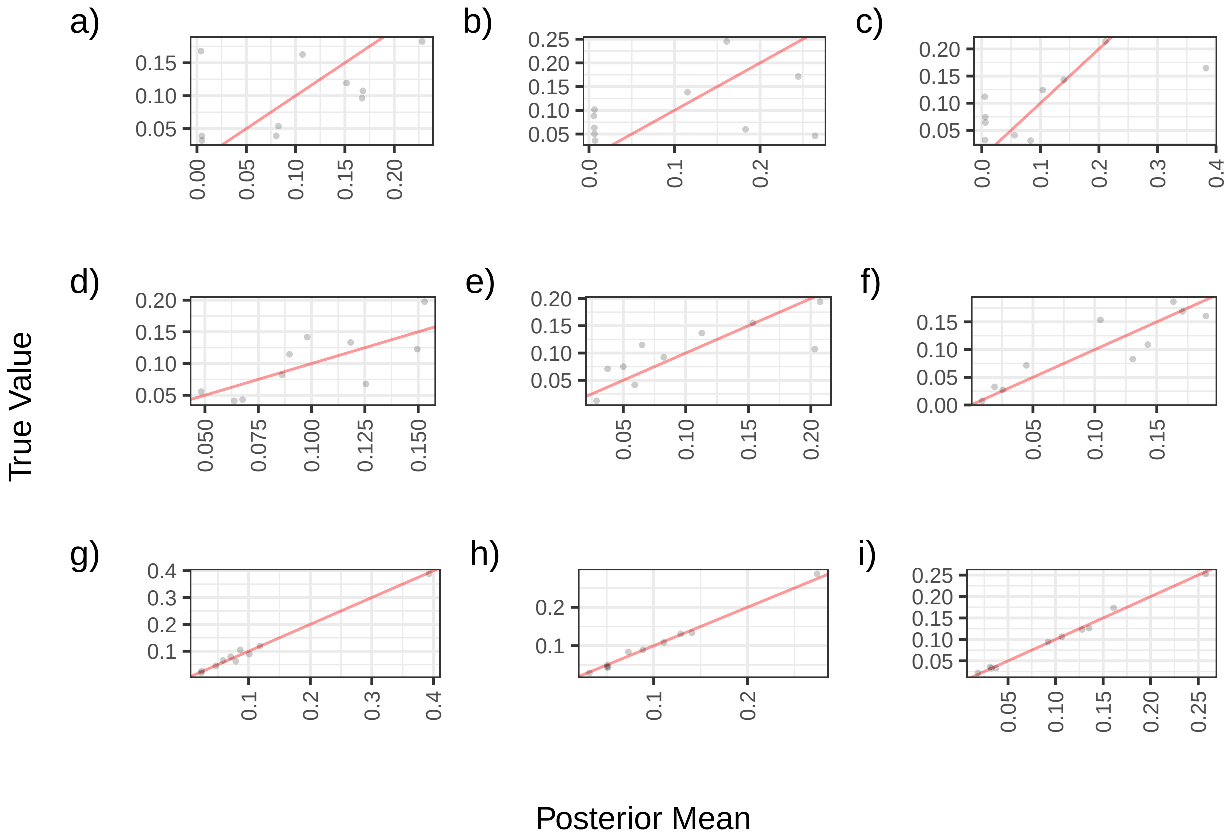

Figure S 4: Correlation between simulated and predicted values (posterior mean) for the first three simulated datasets of simulation scenario #1 with a sample size of 10 (a-c), 100 (d-f), and 1000 (g-i) per observation; a 1:1 reference line is shown in red.

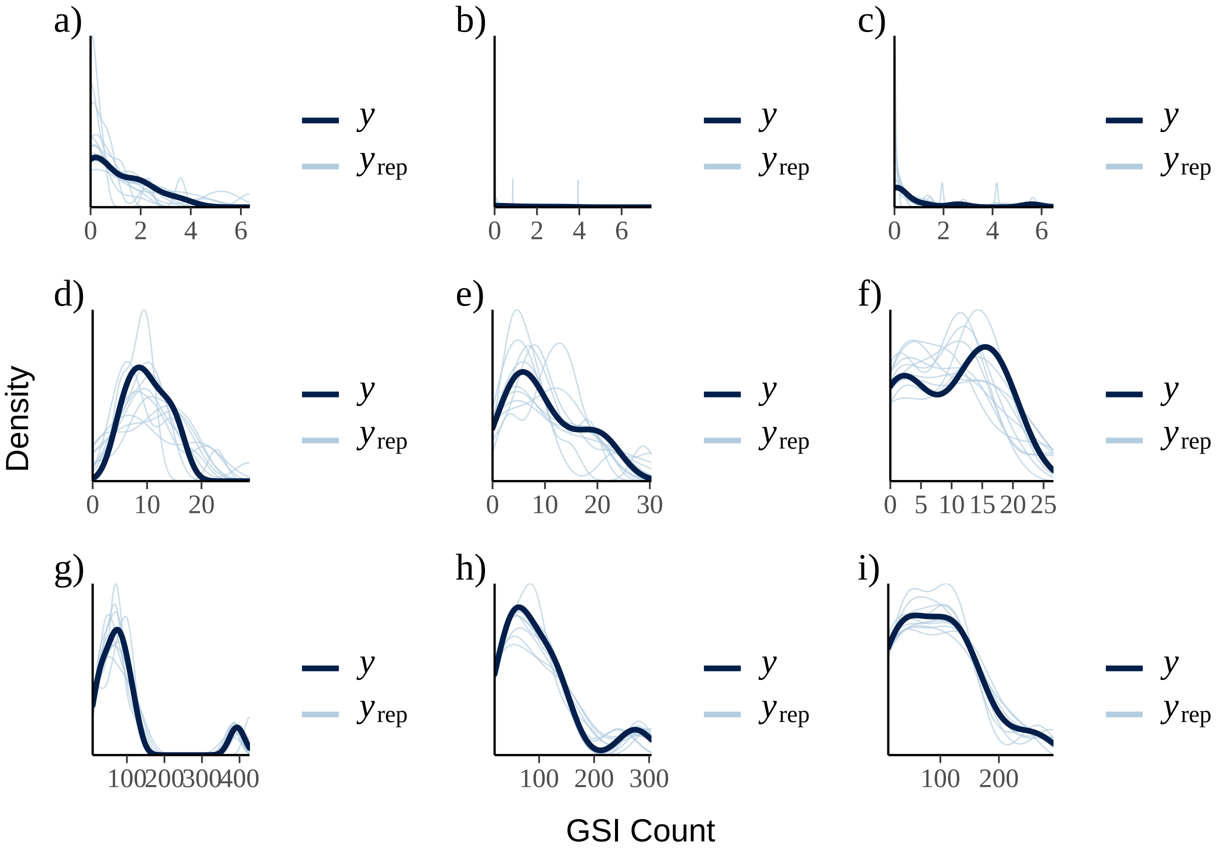

Figure S 5: Graphical posterior predictive checks comparing model fitting data and MCMC iteration-specific simulated values. We present results for the first 10 MCMC iterations from the first three datasets for simulation scenario #1 with a sample size per observation of 10 (a-c), 100 (d-f), and 1000 (g-i). y is the density of observed (simulated) data and yrep is the density of MCMC-simulated data.

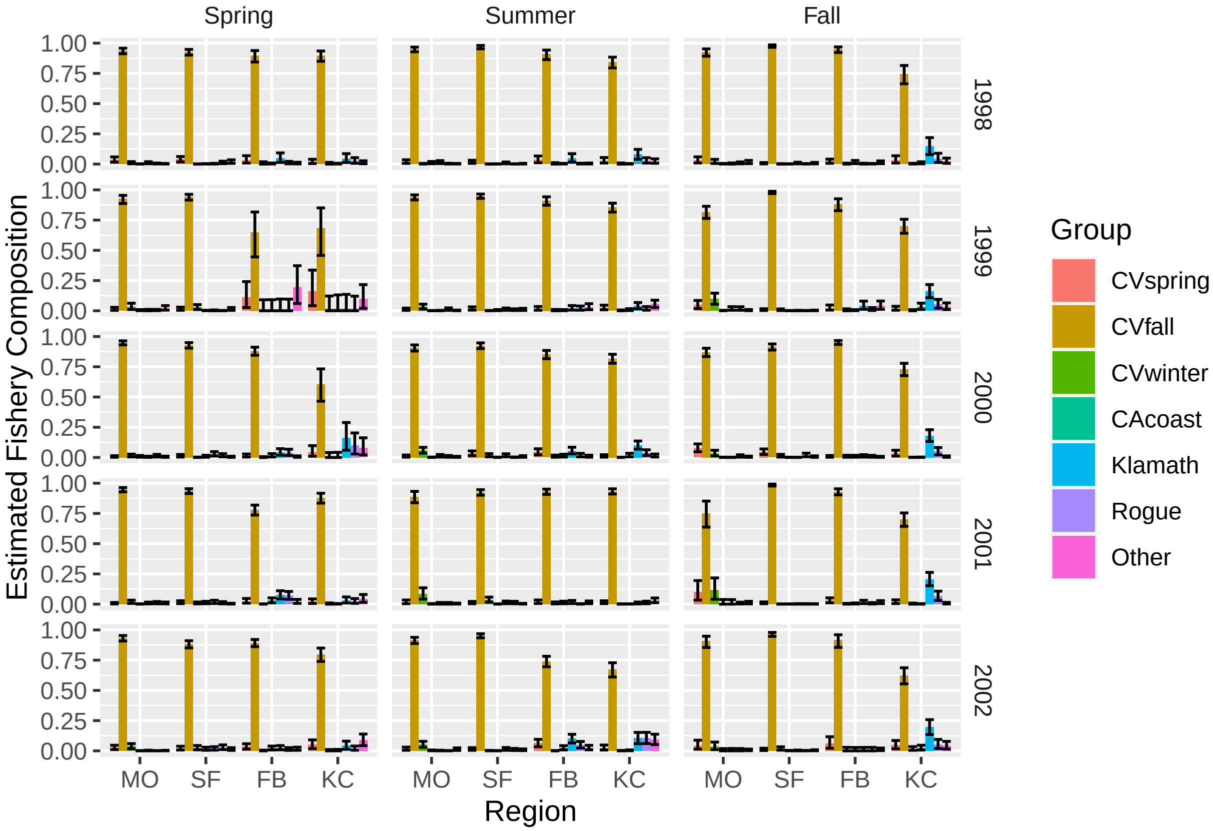

Figure S 6: Estimated fishery mixture proportions for each combination of region, season, and year from the mixed-stock Chinook salmon fishery in California's coastal waters. Error bars represent 95% credible intervals. Regions MO, SF, FB, and KC refer to the Monterey, San Francisco, Fort Bragg, and Klamath Management Zone areas, respectively (Satterthwaite et al. 2015).

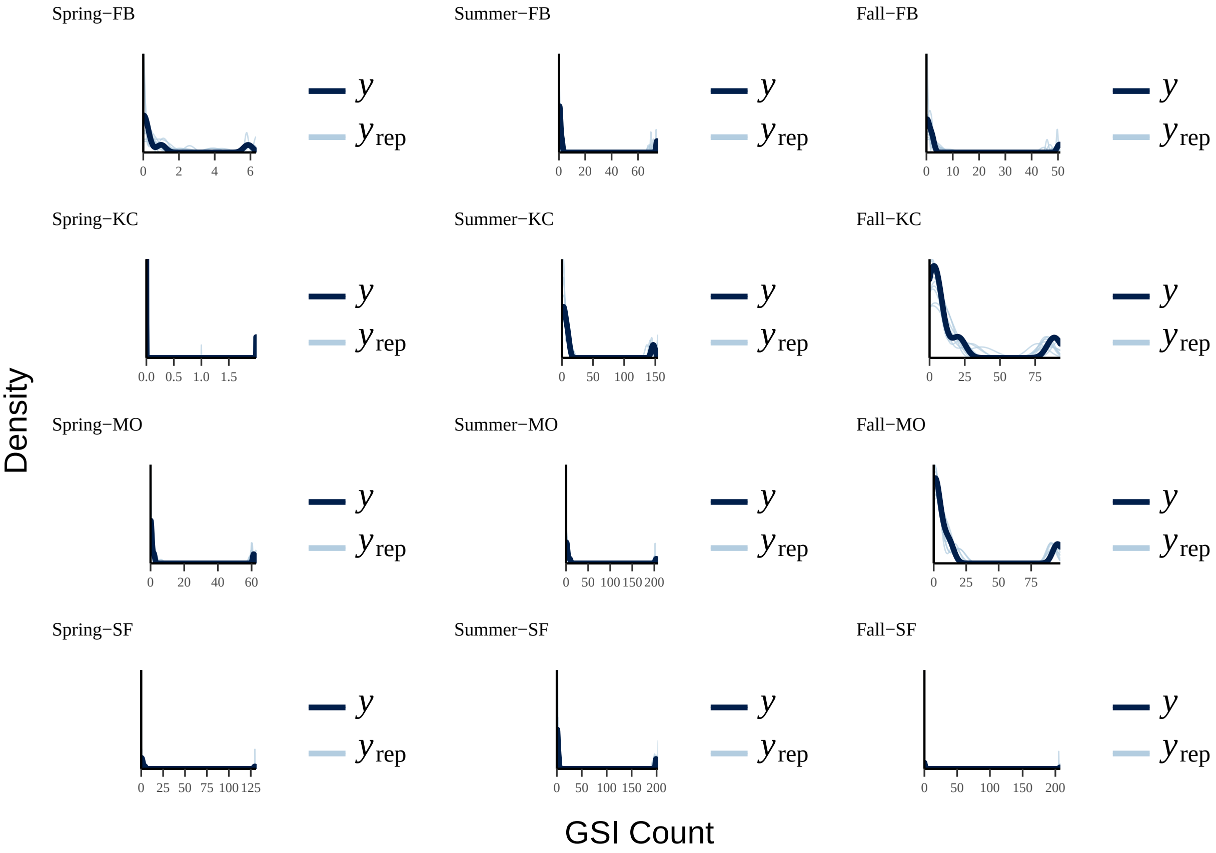

Figure S 7: Graphical posterior predictive checks comparing MCMC iteration-specific simulated values for the first 10 MCMC iterations to genetic assignment data from the California recreational fishery on Chinook salmon in 1999. Values y and yrep refer to the density of observed and model-simulated data.

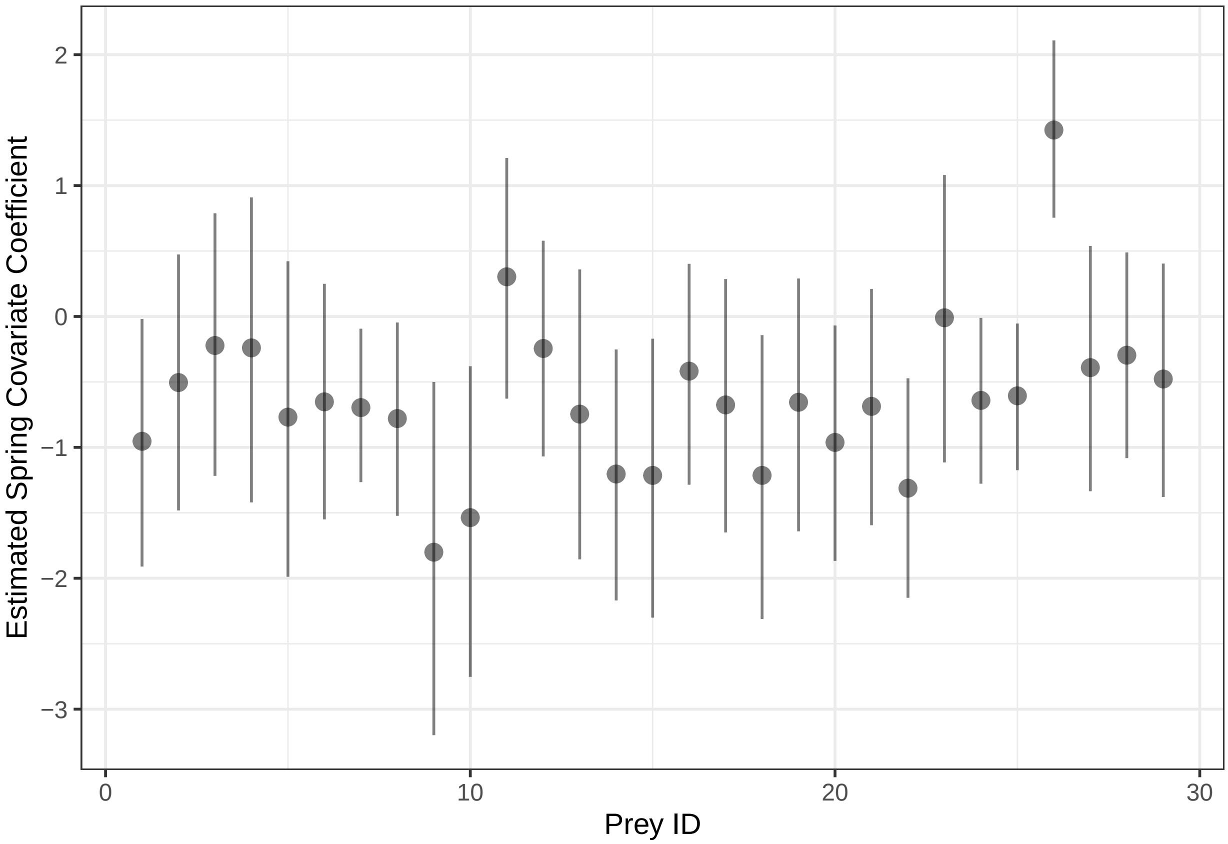

Figure S 8: Estimated spring coefficient values for cod diet types. Points represent mean values and lines represent 95% credible intervals. Prey ID values correspond to scientific names provided in Table S1.