Supplement S2: Statistical Details

Environmental DNA provides quantitative estimates of a threatened salmon species

qPCR statistical methods

Following general practice with qPCR, for each of $q = 1, 2, ..., 5$ qPCR plates, we ran a dilution series of known concentrations of Chinook salmon genomic DNA (D, between 10^{-1} and 10^{-7} μg μL^{-1}) to estimate the concentration of Chinook salmon DNA in field collections. We observed the estimated PCR cycle, C_{iq}, at which the flourescence of replicate-PCR reaction q became detectable above background levels for the ith replicate PCR reaction. Note that the observed PCR cycle is continuous, not discrete (i.e., C can take on fractional values). This number then allowed us to estimate regression coefficients β_0 and β_1 that describe the relationship between known DNA concentration and counts,

$$C_{iq} \sim Normal(\beta_{0q} + \beta_{1q}[\log_{10} D_{iq}], \sigma^2) \tag{1}$$

For some concentrations of the standard and some field samples, no PCR amplification was observed. As ignoring these instances lack of amplification can lead to bias in model estimates (Lahoz-Monfort, Guillera-Arroita, & Tingley, 2016), we included this information in our model by adding a presence-absence component. We use logistic regression to model the probability of occurrence, θ_{iq},

$$\begin{aligned} &Y_{iq} \sim Bernoulli(\theta_{iq}) \\ &logit(\theta_{iq}) = \phi_{0q} + \phi_{1q}[\log_{10} D_{iq}] \end{aligned} \tag{2}$$

where ϕ_0 and ϕ_1 are regression coefficients and $Y_{iq} = 1$ if PCR amplification was observed and $Y_{iq} = 0$ otherwise. For field collected samples, the concentration of Chinook salmon DNA was unknown and the primary object of interest. We define γ_{st} as the true $\log_{10}$ DNA concentration to be estimated at each site, s, in month, t. Let δ_{bst} represent the deviation for each replicate bottle, b, collected at each site-month combination so that the $\log_{10}$ DNA concentration is

$$\begin{aligned} &\log_{10} D_{bst} = \gamma_{st} + \delta_{bst} \\ &\delta_{bst} \sim Normal(0, \tau^2) \end{aligned} \tag{3}$$

Here, τ^2 represents the among-bottle variance in DNA concentration at a given site-month combination. In parallel to the likelihood for the dilution standard, we can connect the estimated DNA concentration to the observed PCR counts,

$$C_{ibstq} \sim Normal(\beta_{0q} + \beta_{1q}[\log_{10} D_{bst}], \sigma^2 + \omega^2) \tag{4}$$

and ω^2 is the additional variance in counts attributable to laboratory processing of the field samples beyond the variance contributed by the standards. We can write a model for the presence-absence component of the field samples as well

$$\begin{aligned} &Y_{ibstq} \sim Bernoulli(\theta_{bstq}) \\ &logit(\theta_{bstq}) = \phi_{0q} + \phi_{1q}[\log_{10} D_{bst}] \end{aligned} \tag{5}$$

For this statistical model, we are primarily interested in the estimates of the DNA concentration at each site-month combination, γ_{st}. However, we are also interested in the parameters that reveal the causes of variation in DNA concentration. Two processes contribute to the variability in observed qPCR counts: σ^2 (variance due to uncertainty in the standard curve) and ω^2 (within-bottle variance; differences among PCR replicates from identical field samples). We expect these quantities to be small relative to τ^2 (among-bottle, within-site variance). As σ^2 and ω^2 are in units of PCR counts and the other measures of variation are in units of $\log_{10}$ DNA concentration, we used the rules of random variables to convert σ^2 and ω^2 to units of $\log_{10}$ DNA concentration to enable comparison (see Supplement S2). Together, the variability attributable to PCR standards, PCR replicates, and among-bottle replicates can be combined to describe the total within-site variability in DNA concentration at each site-month combination (denoted "PCR + Bottles"; see Supplement S2). This within-site variability is comparable to the within-site variation of beach seine sampling (see below). We also calculated two derived measures of variation to understand variation at the scale of Skagit Bay. We calculated the standard deviation in γ_{st} among months at a each site ("Month"" variation) and standard deviation in γ_{st} among sites in each month ("Site" variation).

Calculation of variability in terms of DNA concentration

To compare the variability among different factors involved in the sampling and processing of eDNA samples, we need to convert all of the factors into shared units. Specifically, we need to convert σ^2 and ω^2 from units of PCR replicates to units of $\log_{10}$ DNA concentration. Since

$$C_{ibstq} \sim Normal(\beta_{0q} + \beta_{1q}[\log_{10} D_{bst}], \sigma^2 + \omega^2)$$

(eq. 4), we can rearrange the equation in terms of $\log_{10} D$ (supressing subscripts for brevity)

$$\log_{10} D \sim Normal\Big(\frac{C - \beta_0}{\beta_1}, \frac{\sigma^2}{\beta_1^2} + \frac{\omega^2}{\beta_1^2}\Big)$$

which means that the slope of the regression line, β_1, modifies the estimated variance among PCR standards (σ^2) and among PCR samples (ω^2) when they are translated into DNA concentration units.

Combining multiple terms to determine the variability within sites ("PCR + Bottles")

There are three components that contribute to variability within a site: variance arising from the use of PCR standards, variance among PCR replicates of field samples, and variance among replicate bottles collected at a site. The previous section provides the first two of these terms in units of $\log_{10}$ DNA concentration. We can use the MCMC draws from the joint posterior distribution to calculate the variability attributable to all three processes combined. For each MCMC draw we added the variance in DNA concentration among bottles within each site-month combination to the variance form the PCR components derived in the previous section, $\frac{\sigma^2}{\beta_1^2} + \frac{\omega^2}{\beta_1^2}$, to come up with a total within-site variance. We then calculated the average and 90% credible intervals for the variance for each site-month combination (grey points under "PCR+Bottles" in Fig. 2) as well as a grand average and confidence interval across all site-month combinations (black point under "PCR+Bottles" in Fig. 2).

Calculating probabilistic limit of detection

We use the estimated parameters defining the logistic regression parameters, ϕ_{0q} and ϕ_{1q}, for the five qPCR plates ($q = 1, .., 5$) to calculate the environmental DNA concentration at which our assay could detect Chinook salmon (Eqs. 2 and 5). We used the MCMC draws from the posterior distribution to calculate the DNA concentration at a defined probability of detection, θ,

$$\log_{10} D = \frac{logit(\theta)}{\phi_{1q}} - \frac{\phi_{0q}}{\phi_{1q}}$$

We evaluated this equation for $\theta = \{0.05, 0.25, 0.50\}$ for each plate and for the average across all five plates.

Across all five plates, we estimate a 5% of detecting a DNA concentration of $\log_{10} D = -5.79(0.99)$ [Mean(SD)]. A 25% probability of detection corresponds to $\log_{10} D = -5.15(0.46)$ and a 50% probability of detection is $\log_{10} D = -4.77(0.17)$ (see also Fig. S.2).

Statistical Methods for Beach-Seine Sampling

Similar to qPCR sampling, we created a statistical model for all beach seine samples collected. As beach seines directly observed Chinook salmon, we model the observed count of Chinook salmon in beach seine j where $j = 1, 2$ at site s in month t as

$$Z_{jst} \sim NegBinom(\log \psi_{st}, \nu) \tag{6}$$

where $\log \psi_{st}$ represents the natural logarithm of mean abundance at each site-month combination, and ν controls the amount of over dispersion. We estimate a single, shared overdispersion parameter for all site-month combinations. See Supplement S2 for information about prior distributions (Table S2.3).

As with the qPCR analysis, we are interested primarily in the estimated abundance at each site $\log \psi_{st}$ but we are also interested in the components of variation for beach seine sampling. For beach seines we can calculate the variance of catches at a particular site and month from the attributes of the Negative Binomial distribution: $Var[Z_{st}] = \psi_{st} + \frac{\psi_{st}^2}{\nu}$. This variance describes the variability in catches among beach seines conducted at a single site-month. However, for the Negative Binomial distribution, the mean and variance are not independent: the variance increases as a function of the square of the mean. This differs from the independence of the mean and variance in the normal distribution used for the qPCR analysis. We calculated the average standard deviation ($SD[Z_{st}] = \sqrt{Var[Z_{st}]}$) in seine catches across all site-month combinations to quantify within-site variability. In parallel to the qPCR results, we compare this values to the estimated standard deviation in ψ_{st} among months at a each site ("month"" variation) and estimated standard deviation in ψ_{st} among sites in each month ("site" variation).

Estimation

We estimated both the statistical models for qPCR and beach seine in Stan (Gelman et al. 2015; Carpenter et al. 2017) as implemented in the R environment [rstan, v.2.16.2; R Core Team, 2018, Stan Development Team 2018). Stan is a language that implements a Hamiltonian Markov Chain Monte Carlo sampler for Bayesian statisical models (Carpenter et al. 2017). For both the beach seine and qPCR analyses we used 5 parallel chains with diffuse starting locations and examined Gelman-Rubin diagnostics to esure adequate mixing among chains, and thereby an adequate search of parameter-space. To improve MCMC chain mixing, we included an offset in the model, replacing $\log_{10} D$ with $\log_{10} D - 4.5$ in all above equations. The inclusion of the offset does not meaningfully affect the interpretation of our results. We used diffuse prior distributions for all parameters; we provide all analytical code and data in the online supplement and data in (DRYAD ONLINE ARCHIVE). For all results we use MCMC samples from the posterior distribution to derive credible intervals for individual parameters and derived quantities.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. doi:/href%7Bhttps://doi.org/10.18637/jss.v076.i01%7D%7B10.18637/jss.v076.i01%7D

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization, *40*(5), 530–543.

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Stan Development Team. (2018). RStan: The r interface to stan, r package version 2.16.2. *Http://Mc-Stan.org/.*

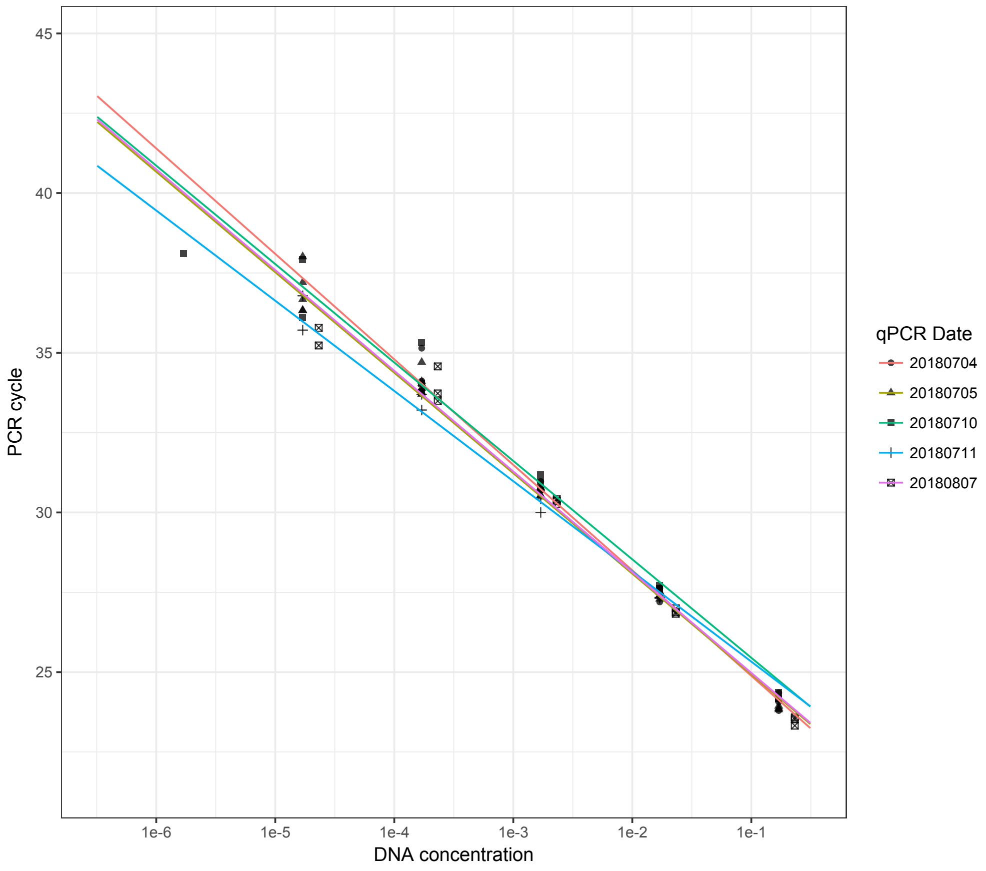

Figure S.1: Estimated relationships between DNA concentration and PCR cycle at which the flourescence of replicate-PCR reaction became detectable above background levels. Each line corresponds the mean linear regression estimate for each PCR plate

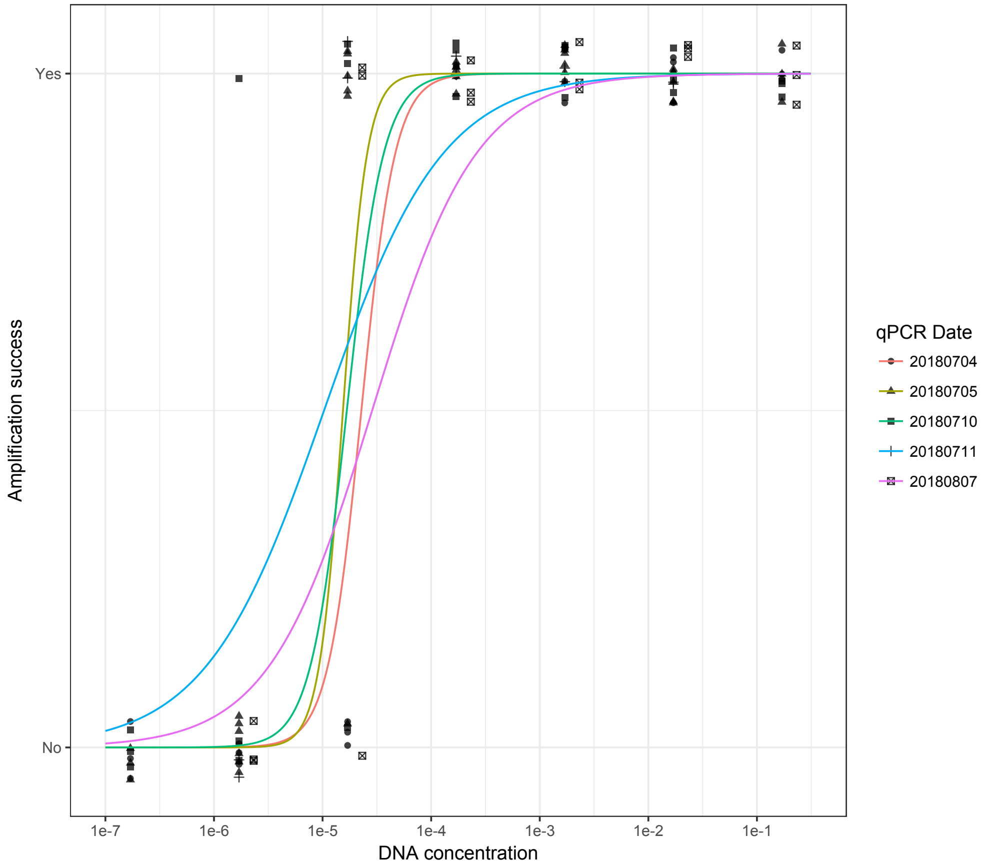

Figure S.2: Estimated relationships between DNA concentration and the probability of observing PCR amplification. Each line corresponds the mean logistic regression estimate for each PCR plate

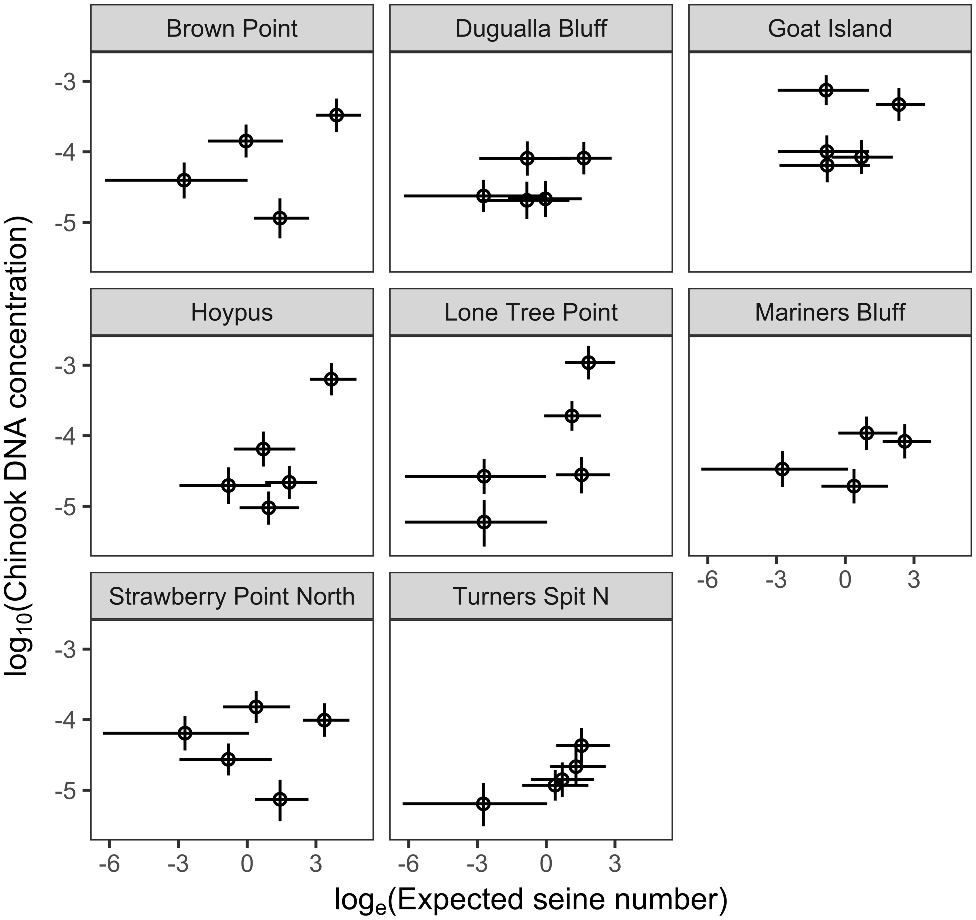

Figure S.3: Pairwise correlation at each site between the estimated expected catch from beach seines and log_{10} DNA concentration. This is the data contained in Fig. 3 separated by site.

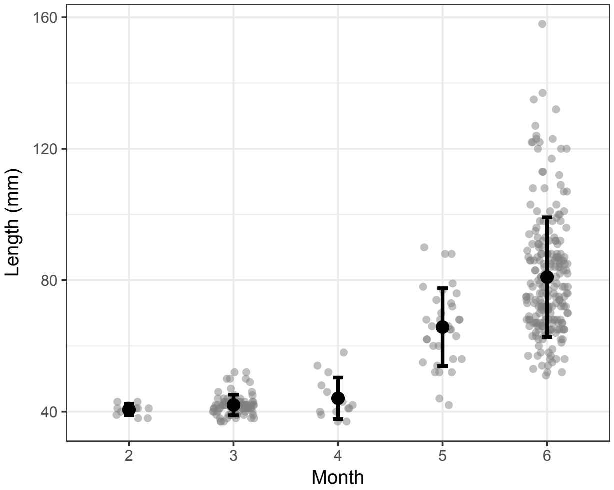

Figure S.4: Observed lengths (fork length) of Chinook salmon by month in 2018 as collected by beach seine. Grey points show individual fish, black points and error bars show monthly mean and one standard deviation.