The following are my own comments on some things that need changing in the thesis.

Throughout:

Standardize indicator function notation to $\mathcal{I}\{Y \in \mathcal{A}\}$, for example.

☒

Front Matter

p. viii Do the acknowledgments section. Use outline in notebook.

☐

Chapter 1. Introduction

p. 9 Whole locus Gibbs sampler and Heath and Thompson stuff]

☐

Chapter 2. Importance Sampling

p. 26 ET wants a figure here that describes how the sampling is done sequentially over alleles.

End Add an Extensions and Caveats section that covers:

1. the logistic approximation for computational efficiency, (sort of)
2. accounting for extinction probabilities
3. the problem of allele order in the sequential method

☐

Chapter 3. Pólya Urn Model

p. 59 Note that the likelihood derived from the WF model may tend to overestimate effective size (but this effect could be terribly slight).

☐

p. 61 Change the graph so it has the RB'd curve in there.

☐

Discussion and Extensions Add a brief discussion section in which the following topics are addressed

1. Fluctuating λ (mention models for detecting that or for seeing if it depends on the actual number of breeding individuals)
2. Census sizes not known without error.

☐

Chapter 4. Overlapping Year Classes

Beginning Include Nielsenetal2000 in the motivation (supportive breeding stuff). Also include the other Nielsenetal paper on archived fish scales and DNA.

☐

p. 64 Need some major clarification about why the model in which "population sizes don't change" is crucial.

☐

Somewhere Point out that sampling from juveniles to adults still assumes H-W equilibrium.

☐

p. 92 Come up with a different notation for $\dot{S}$

☐

p. 93 in 4.5, (25,60) is not less pointed than (60,125).

☐

Wrap-up Include a short concluding section that also notes that these methods would be very appropriate to salmon hatchery situations (fecundity is well known, census is well known, etc.) Also that once again, λ may vary over time. As in the previous chapter these methods could be extended to include model comparison (those in which λ changes and those in which it doesn't)—particularly appropriate for testing for changes in λ over time—good example is before and after the establishment of a supplemental breeding program, or coinciding with changes in hatchery management style.

☐

Chapter 6. Direct Modeling of Hybridization

Somewhere Note that the distinction between *a priori*, known, fixed differences between species and frequency differences b/t species can both be handled by the choice of prior.

☐

p. 137 Major tense overhaul required in writing about the four analyses for demonstration.

☐

p. 148 Add something about fewer problems with trapping states because the component specific parameters are not greatly influenced by sparsely-filled components.

☐

Bibliograply and Appendices

Appendix B This thing needs major cleaning up. Note in the beginning of it that the notation departs a bit from the conventions in the main text—roman lowercase now denotes a random vector so that the uppercase roman can be a matrix.

☐

The following are my paraphrases of Robin's comments on the dissertation.

Chapter 1. Introduction

p. 6 "populations, however" should be "populations; however," and that semi-colon should be used elsewhere before "however."

☐ **Check this out in the Chicago Manual of Style**

p. 8 Define latent variables, say that the abbreviation for Markov chain Monte Carlo is MCMC, and explain Monte Carlo before MCMC. Also, perhaps the section on Monte Carlo should precede the section on Monte Carlo in genetics.

☒ *Did all the above, except left the sections in the same order as they are now. Did some re-working so that there is a mini-definition of Monte Carlo before the MC in Genetics section, but point out that a more in-depth introduction to MC will appear later in the first chapter.*

p. 10 "For example X or θ" is not a sentence. "The two" means what?

☒ *Added a semi-colon. Deleted the sentence about the Bayesian framework.*

p. 12 Monte Carlo also useful for approximating distributions...

☒ *Included a short paragraph explaining how approximations of distributions are approximations of the probability that the rv falls in an interval (for many different intervals). Each of these probabilities may be expressed as an expectation.*

p. 16 When is it not the case that $\frac{h(x)}{h(x)} = 1$?

☒ *Changed this around a bit, and clarified it.*

Chapter 2. Importance Sampling

p. 21 The data are *random* genetic samples.

☒

p. 25 Missing "be."

☒

p. 27 "The ability to estimate N_e (add: *with adequate precision*) requires data from many loci.

☒

p. 27 Regarding the assumption that the alleles are in linkage disequilibrium at the beginning and remain in linkage equilibrium over time—clearly there will be LD from finite population size. Explain....

☐ **Explain how it is that that assumption is implicit in the formulation. Say that this will be violated in practice, but on expectation, the LD resulting from finite population size should not greatly change the result from the temporal method. However, in the discussion I should say that joint modeling of LD and temporal changes of frequencies over time might be a fruitful area of research.**

p. 34 Note that the sample size is four times the effective size—isn't that sort of odd, or at least it requires explanation.

☐ **Explain.**

p. 36 Note explicitly that the simulation is not a simulation study with many different replicates—it is a single simulated data set only.

☐ **Make this fact explicit. Discuss the fact that this is not a simulation study to determine the properties of the estimator—rather a set of simulated data to show that the computational method yields a likelihood curve.**

p. 37, Figure 2.3 Features described in the caption are not visible in the graph.

☐ **Include in caption that the features are almost impossible to see because the error is quite small.**

p. 38 Are the advantages of ML over F-statistics with respect to bias, precision, or both?

☐ **Cite Williamson and Slatkin and also Wang saying it is both (though primarily in the presence of low freq alleles. Also, though, greater flexibility in modelling changes in population size, *etc.***

p. 39 Should be able to compute Pollak's estimator while excluding the Pgm locus.

☐ **I did this—the new estimate via Pollak's method is close to the previous estimate via Pollak's method (268 instead of 250).**

p. 40 Needs a discussion of the underlying assumptions about N_e over time. And also more expansion upon the idea of λ and how it could change over time.

☐ **Include this discussion**

Chapter 3. Pólya Urn Model

In general What does the method estimate when λ fluctuates over time? Is it the harmonic mean λ, etc.?

☐ **This is really an interesting question that gets at the heart of the issue that when combining data from many intervals to estimate a single N_e or λ one must weight the information from different intervals in some way, either implicitly or explicitly. Include a discussion section to describe this sort of point.**

p. 41 The census size C is easy to count as the number of spawners in salmon populations, but what about other organisms? What gets counted in C?]

☐ **Review Nunney's paper about this and describe.**

p. 51 Family size varies. . . over time? across families? Be more explicit.

☐ **Across families. Also, this section needs substantial editorial assistance.**

p. 51 "It is possible to derive the inbreeding and variance effective sizes of (should be "in") the urn model." (How can you have te N_e of a model?)

☒

p. 52, Bottom Change "previous generation" to generation 1.

☒

p. 53 Define IBD.

☒

p 56 Allele fixation stuff. I must make clear that this is the loss of an allele from the population in a single generation.

☐

p. 57, Bottom "probably too low" should be "probably too high"

☒

p. 58 It is not clear why use of the Wright-Fisher model implies the assumption that the distribution of family sizes are bimodal.

☐ **Explain this more fully, and describe better what I mean...**

p. 59 Once again: fixation probabilities within what time frame? One generation? Many generations? *etc.*

☐

p. 59 "Likelihood derived from WF model may overestimate the effective size"... which effective size am I talking about there?

☐ **Dramatically clarify.**

p. 61 Note in caption and text that this figure is based on the data from Begon.

☐

Chapter 4. Overlapping Year Classes

p. 66 On determining the age of juveniles—can do with scales, but also, in some spp. all the juveniles will be of a single age class.

☐ **Point out the distinction between steelhead and some forms of chinook, *etc.***

p. 74 Juveniles and adults can be treated as independent binomial draws from the previous generation.

☐ **Read over Nei & Tajima and Waples and figure out if this would or would not apply to the present modeling scheme.**

p. 74 Determining age of the parent of a gene copy sampled amongst juveniles could be done via pedigree analysis.

☐ **Point this out. Cite "family-printing."**

p. 90 Allele counts in the years 1950 to 1953 were simulated via the urn model—describe more fully.

☐ **State explicitly how it follows the assumption in the prior**

p. 90 How do you know that 1954 to 1963 is long enough to simulated to allow the allele frequencies to settle into a correlated pattern?

☐ **Don't know for certain. Cite Robin's manuscript that 20 years was enough for F's between generations to settle down—10 years may have not been quite enough, but, the resulting allele counts probably represent a sample from the "stationary" distribution resulting from some series of census sizes back before 1949.**

p. 90 Juvenile survivorship figures—were those for stream or ocean type fish?

☐ **Check Healey. I think they were sort of aggregate measures. Note that the absolute values are not particularly important—the relative fecundities are the more important figures there.**

p. 90 Cite the data source as Beamsderfer *et al.*

☒

p. 92 λ may vary by age, but it is much more likely to vary by year.

☐ **Add a discussion, and discuss this, along with some other things.**

p. 95 The fact that the 90% credible interval overlaps .4 is nice, but not very convincing by itself.

☐ **Point out again that this is a simulated dataset for demonstrating that the computational method works.**

Chapter 5. Mixture and Admixture

p. 110, Figure 5.2 Correlation between α and ξ_P are only apparent for $\alpha < 1.5$.

☐ **Note this in the caption, and also refer to it in the discussion when talking about setting the upper limit on α.**

Chapter 6. Direct Modeling of Hybridization

p. 136 Make it more explicit at this point that the sample of juveniles was not a random sample—they were believed to be cutthroat.

☐

p. 141 "It is unlikely that any purebred individual will receive high posterior probability of being in a non-purebred genotype frequency category" (assuming you know what pure SH and CUTT look like).

☐ **Figure out what he is getting at here, and expand.**

Bibliograply and Appendices

p. 161 Thompson and Heath 1997 also says "in press." Straighten that out.

☐ **Will get straightened out when I figure out just which of that series of articles to cite.**

Following are some of Joe's comments on the dissertation:

Chapter 1. Introduction

p. 12, section 1.5.1 Was Monte Carlo actually used for development of the A-bomb during the war? My understanding is that it was used for development of the H-bomb after the war.

☐ **Check out the book, Stanislaw Ulam 1909–1984, Los Alamos, N.M. : Los Alamos National Laboratory, 1987, QA7 .S79 1987**

p. 12 line 5 from bottom I understand that Monte Carlo methods are most often used to approximate expectations, but they also serve to approximate variances.

☒ *Explain in the text that many things can be expressed as expectations, including variances and distributions,* etc.

p. 17 Again a quibble. Yes, to do what you describe, $h(x)$ should be easy to compute. In our MCMC method for inferring ML estimates of $4N\mu = \Theta$, we actually have a case where $h(x)$ is known only up to a constant. In that case instead of the integral g_n we only get it as $g_n/(\int h)$, but that is good enough to give us the likelihood ratios and permit ML inference.

☐ **Include another paragraph describing what happens when $h(x)$ is known only up to a normalizing constant. Check out Adrian's chapter in MCMC in Practice—I believe that has a lot about different variations on importance sampling. Or maybe Gelman et al's section on importance sampling. Also include a description of the method as applied in Joe's paper too.**

pp. 18–19 I am not sure after reading all this that I still know what Rao-Blackwellization is. I think it needs a bit more intuitive explanation. I couldn't winkle it out of the notation, though I am sure it is there.

☒ *I tried to give a preliminary, intuitive explanation in a new paragraph.*

Chapter 2. Importance Sampling

p. 22 Explain a little – the reader may not get it that the assumption is that there a great many more juveniles than adults, and hence that one can indeed (almost) sample with replacement from the adults.

p. 22 The uniform prior on $P_{N_e}(\mathbf{X}_0)$ is tossed in without much comment as to the reasonableness of doing this.

☐ **Recall that Ellen and Monty looked at the differences this made and found that it didn't make much difference if they used a beta prior. Primarily because, I believe, if you have a sample of 200 alleles at the first sample, that information completely overwhelms the weight in the prior. With very many alleles, then it might be better to use a unit-information prior—*i.e.*, the Jeffreys prior. But it should make very little difference, since it really will not change the posterior distribution of the latent allele counts in that first generation greatly.**

p. 27 The unbiased estimators of variances on this page, and the confidence interval estimates sound like they are in effect assuming that the $P_{N_e}(Y, X^{(i)})$ are i.i.d. Of course they aren't, but are these estimates somehow reliant on having large enough m that they may be regarded as i.i.d. Might say so if so.

☐ **In fact, the $P_{N_e}(Y, X^{(i)})$ are i.i.d.! This is an importance sampling exercise. What I don't mention, though I suppose I could, is that the same block of random numbers is used for each value of N_e so that, in fact, there is some correlation that way which makes the curve smoother—it doesn't improve the estimation of the absolute value of the likelihood, but it does improve the estimation of the relative likelihoods between values of N_e.**

pp. 31–34 The reflection stuff loses me. I think readers may have the same problem. It is too much, too densely, without any intuitive explanation.

☐ **Give an intuitive introduction to the section**

Chapter 3. Pólya Urn Model

p. 41, 7 lines from bottom "leads to unattractive computational situations". The reader will find this to whiz by too fast with no explanation. In fact, section 3.2 does explain it. Needs some signposting here that this will be covered in that section.

☐ **Signpost it!**

p. 47, line 12 When it is said that X follows the compound multinomial Dirichlet distribution, presumably that is true conditionally given the n_i, not marginally. This should be made clear.

☐ **The whole section could be cleaned and cleared up a little bit.**

p. 55 (3.16) N is described above it, but then in this equation C is used instead.

☐ **This was a typo that I should check to make sure I have already fixed.**

pp. 55–59 I must be missing something here. Surely with an initial frequency 0.025 of an allele, in all these models the fixation probability of that allele (long-term) must be the same for all copies of the gene and thus be exactly 0.025. Or am I not getting it? (In which case readers might not either).

☐ **Include, very explicitly, the fact that this is allele loss IN ONE GENERATION.**

p. 62 "significantly" (typo)

☐ **Better fix it.**

Chapter 4. Overlapping Year Classes