MISIDENTIFICATION ERROR IN NON-INVASIVE GENETIC MARK-RECAPTURE SAMPLING: CASE STUDY WITH THE CENTRAL GEORGIA BLACK BEAR POPULATION

by

JAMIE L. SKVARLA SANDERLIN

(Under the Direction of Nicole Lazar)

ABSTRACT

Advances within the fields of molecular and genetic biology have increased the ability to use genetic analyses in wildlife studies, especially with non-invasive sampling methods. The presence of genetic error, namely allelic dropout and false alleles, can bias demographic estimates from non-invasive studies. The balance of accepting certain levels of error versus genotyping additional loci has not been quantified in the literature. The main objective of this study was to develop a cost-effective method of estimating misidentification error from non-invasive sampling. Error could then be incorporated into black bear abundance estimates for a case study in the central Georgia population (CGP). Calibration samples of known individuals from tissue and hair samples collected from the CGP were used. Model verification was conducted with both simulated data and data from the case study. The objective function for optimal selection of a marker panel with and without genetic error was also described.

INDEX WORDS: misidentification error, non-invasive sampling, black bear, *Ursus*, optimal marker panel

MISIDENTIFICATION ERROR IN NON-INVASIVE GENETIC MARK-RECAPTURE SAMPLING: CASE STUDY WITH THE CENTRAL GEORGIA BLACK BEAR POPULATION

by

JAMIE L. SKVARLA SANDERLIN

B.S., Purdue University, 2002

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2009

MISIDENTIFICATION ERROR IN NON-INVASIVE GENETIC MARK-RECAPTURE SAMPLING: CASE STUDY WITH THE CENTRAL GEORGIA BLACK BEAR POPULATION

by

JAMIE L. SKVARLA SANDERLIN

Major Professor:	Nicole Lazar
Committee:	Michael J. Conroy
	Jaxk Reeves

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2009

ACKNOWLEDGEMENTS

I would first like to thank my major advisor, Dr. Nicole Lazar, for her many helpful comments with model aspects of this project and encouragement. I am also thankful for my committee members, Drs. Michael Conroy and Jaxk Reeves, for their unique insights to different components of the thesis project. Research support was provided jointly by a research grant from The International Bear Association and the Georgia Department of Natural Resources. I would like to thank the GA DNR for assistance with collection of blood, tissue, and hair samples from the known bears, since the calibration study would not be possible without these efforts. I am grateful to Drs. Joe Nairn and John Carroll for providing laboratory space and various laboratory equipment for all of the genetic analyses. Genetic lab assistance (especially with hair extractions!) is greatly appreciated from Soo hyung Eo, Lauren Wilson, Dr. Michael Conroy, and Jeremy Sanderlin. Brant Faircloth deserves a special thank you for initial training in the lab and for the many discussions regarding marker panel development and laboratory techniques, as well as other members of the Wildlife Genetics Lab. I would like to thank all my office mates and fellow graduate students for great discussions about anything from bear biology to modeling to other non-academic topics! On the computer simulation side, I am fully indebted to my husband, Jeremy, for all of his assistance with securing computers to run a lot of the models. I would also like to thank my family and friends for their words, cards, and notes of support and encouragement through the years in graduate school. Lastly, I could never fully express my gratitude towards my husband, especially for staying up to help during many a long night (and early morning!), for all of his words of encouragement, and for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

Animals that occur at low densities or have elusive behavior are difficult to sample for population inference, and often lead to low sample sizes with physical captures. Non-invasive sampling techniques, or techniques that do not require physical capture of animals, allow many populations of animal species to be monitored with greater detail. Advances within the fields of molecular and genetic biology have increased the ability to use genetic analyses in wildlife studies. Genetic samples (e.g., shed hairs, feathers, feces, shed skin), collected non-invasively in the field, are often small and contain degraded DNA. The ability to create multiple copies of DNA from these samples with PCR (polymerase chain reaction) has advanced non-invasive genetic sampling techniques profoundly (Waits 1999). Non-invasive genetic samples are currently utilized with many animal species globally, and in particular, bear species for problems of demographics (Taberlet et al. 1997, Boulanger et al. 2002, Bellemain et al. 2005), habitat relationships (Apps et al. 2004), dispersal and/or effectiveness of corridors (Dixon et al. 2006, Proctor et al. 2004), to name a few.

To create multiple copies of specific locations within the DNA of an organism with PCR, markers, or the start and end points of the segment of DNA of interest, must be selected. There are different types of markers that can be used in DNA analysis: mitochondrial DNA (mtDNA) markers, Y chromosome markers, and nuclear DNA markers. Nuclear DNA, located in the nucleus of mammalian cells, is inherited by both parents and can be used for maternal and

paternal evolutionary history, gene flow, genetic diversity and relatedness. The Y chromosome (*i.e.*, sex chromosome) is inherited from father to son and can be used for paternal evolutionary history, gene flow, and genetic diversity (Waits 1999).

In most capture-recapture studies with genetic sampling, nuclear DNA from microsatellite loci, also known as microsatellite markers, is used to infer identity (Waits 1999). Additionally, the Y chromosome may be used to determine the sex of an individual in capture-recapture studies (Waits 1999). Microsatellite markers consist of short tandem repeats of 2-9 bases, where bases can be adenine (A), guanine (G), cytosine (C), or thymine (T) (Hartl 2000).

Some studies list possible benefits of non-invasive genetic sampling, such as field methods that may be less expensive, and less harmful to the animal than physical captures and the mark, or genetic identity, is visible, read clearly, and permanent (Foran et al. 1997, Woods et al. 1999). These assumptions are adopted in many non-invasive studies, although I believe they warrant further investigation. There remains doubt in the clarity of genetic identity (Taberlet et al. 1999, Bonin et al. 2004) and cost-effectiveness of non-invasive techniques. Thus, the tradeoff between cost and information quality is unknown.

The presence of genetic error (e.g., allelic dropout or false alleles) is a key factor with accuracy measures for genetic non-invasive techniques. There is also an additional trade-off between the ability to determine individuals and cost (monetary and time) with the number of microsatellite loci used in genetic analyses. The balance of accepting certain levels of error versus genotyping additional loci is recognized as a challenge (Hoffman and Amos 2005). However, few approaches that quantify an optimal number of loci have been described in literature.

Genetic error

A series of microsatellite markers, or loci, are used in capture-recapture studies to infer the identity of individuals. At each locus, individuals may have a different number (*k*) of copies of the microsatellite repeat (Hartl 2000). PCR primers are used to amplify the microsatellite repeat region in individuals, that consist of alleles that differ in fragment size due to varied number of core repeats of factor *k* (Hartl 2000). Individuals that are homozygous produce one band on a gel, while heterozygous individuals yield two distinct bands. Species within a population can have multiple alleles of different lengths at a specific locus. This leads to several possible combinations of homozygotes and heterozygotes. An individual in a population may have a unique genotype (series of specific alleles at each locus) with respect to other individuals in the population. Twins (or other multiple births) are an exception to identical genotypes in a population. Siblings and parent-offspring pairs will have similar genotypes (more identical alleles at each locus than non-related individuals), or even identical, if an insufficient number of loci are used, labeled the 'shadow effect' (Mills et al. 2000).

In any genetic analysis, especially non-invasive analyses, there are two main types of errors that may occur at a specific locus with respect to alleles: allelic dropout and false alleles. In allelic dropout, one of the two alleles present may be lost during the analysis. Allelic dropout with a homozygous individual would not be problematic, since the individual would still be considered a homozygote at that locus. However, with a heterozygous individual, the individual would be classified as homozygous under allelic dropout, leading to an error in classification. A false allele is the addition of one allele to a locus during an analysis. True homozygous individuals with a false allele event would be considered heterozygous at the locus, and true heterozygous individuals would have three alleles present. In the case of a true heterozygous

individual with a false allele event, it is clearly evident that an error occurred since only two alleles are possible at a locus. However, errors are not always clear with a true homozygous individual. An accumulation of genetic errors may lead to misidentification of individuals within a population.

Sources of genetic error

Genetic errors can occur at various steps within a genetic study (e.g. sampling in the field, DNA extraction, molecular analysis, scoring of genotypes, data entry and analysis), and be a result of various sources (e.g., human error, technical error, biological processes) (Bonin et al. 2004). Technical error can include PCR amplification artifacts (Taberlet et al. 1999, Rodriguez et al. 2001), biochemical anomalies (Smith et al. 1995), electrophoresis (Fernando et al. 2001, Delmotte et al. 2001), temperature variation in the laboratory (Davison and Chiba 2001), and with the quality and type of DNA used (Goosens et al. 1998, Taberlet et al. 1999). Scoring of genotypes can be complicated by 'stutter bands' that are created by the slippage of *Taq* polymerase during PCR (Litt et al. 1993, Ginot et al. 1996).

Allelic dropout is often a result of amplification of a small amount of DNA and pipetting the DNA into a dilute sample, otherwise known as sampling stochasticity in the laboratory. This means only a fragment of the total sequence may be in that sample (Goossens et al. 1998, Taberlet et al. 1999, Woods et al. 1999). False alleles can occur with amplification artifacts from PCR with dinucleotide microsatellites (Goossens et al. 1998, Taberlet et al 1999, Woods et al. 1999) or contamination.

The amount of DNA in a sample, presumably, influences the likelihood of a genetic error occurring. In general, more DNA leads to less error, however, there are several

recommendations to the minimal amount needed. Although, laboratory error from human sources does not decrease with more DNA. The amount of DNA available is determined by the non-invasive source and size of the sample. Goossens et al. (1998) determined that using 10 hairs from an animal would have the lowest error rate (out of 1, 3, or 10 hairs), regardless of the hair extraction technique.

Efforts to reduce error

Efforts to reduce error in genetic studies fall into the categories of: 1) laboratory methods, 2) pilot studies (Taberlet and Luikart 1999), and 3) simulations (Taberlet et al. 1996, Petit and Valière 2006). Laboratory methods include: the multiple tubes approach where DNA samples are amplified independently several times (Taberlet et al. 1996), comparison of genotypes with samples from tissue or blood (Wasser et al. 1997, Ernest et al. 2000, Sloane et al. 2000, Parsons 2001, Fernando et al. 2003), reamplifying loci prone to error (Miller et al. 2002) and selecting samples based on some criteria of DNA quantity (Morin et al. 2001, Segelbacher 2002). Many of these methods are specific to the study organism and genetic project, and there is minimal generality among studies. Statistical methods that provide a general framework to assess genetic error allow comparisons to be made across projects and species, as the use of non-invasive genetic sampling increases among bear and other wildlife projects.

Therefore, the objectives in this study were to develop a cost-effective method of estimating misidentification error from non-invasive mark-recapture sampling, which ultimately could be incorporated into estimates of abundance or density in a Bayesian framework. Model verification was done using both simulated data and with the black bear (*Ursus americanus*) central Georgia population (CGP), as a case study. In addition, an algorithm for evaluating the

optimal (in terms of cost and accuracy) number of genetic markers for genetic analysis with the CGP, was evaluated with and without estimates of error rates.

Literature Cited

Apps, C. D., B. N. McLellan, J. G. Woods, and M. F. Proctor. 2004. Estimating grizzly bear distribution and abundance relative to habitat and human influence. Journal of Wildlife Management 68: 138-152.

Bellemain, E., J.E. Swenson, D. Tallmon, S. Brunberg, and P. Taberlet. 2005. Estimating population size of elusive animals with DNA from hunter-collected feces: four methods for brown bears. Conservation Biology 19: 150-161.

Bonin, A., E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. 2004. How to track and assess genotyping errors in population genetics studies. Molecular Ecology 13: 3261-3273.

Boulanger, J., G. C. White, B. N. McLellan, J. Woods, M. Proctor, and S. Himmer. 2002. A meta-analysis of grizzly bear DNA mark-recapture projects in British Columbia, Canada. Ursus 13: 137-152.

Davison, A., and S, Chiba. 2003. Laboratory temperature variation is a previously unrecognized source of genotyping error during capillary electrophoresis. Molecular Ecology Notes 3: 321-323.

Delmotte, F., N. Leterme, and J.C. Simon. 2001. Microsatellite allele sizing: difference between automated capillary electrophoresis and manual technique. Biotechniques 31: 810,814-816,818.

Dixon, J.D., M.O. Oli, M.C. Wooten, T.H. Eason, J.W. McCown, and D. Paetkau. 2006. Effectiveness of a regional corridor in connecting two Florida black bear populations. Conservation Biology 20: 155-162.

Ernest, H.B., M.C.T. Penedo, B.P. May, M. Syvanen, and W.M. Boyce. 2000. Molecular tracking of mountain lions in the Yosemite Valley region in California: genetic analysis using microsatellites and faecal DNA. Molecular Ecology 9: 433-441.

Fernando, P, B.J. Evans, J.C. Morales, and D.J. Melnick. 2001. Elecrophoresis artifacts- a previously unrecognized cause of error in microsatellite analysis. Molecular Ecology Notes 1: 325-328.

Fernando, P., T.N.C. Vidya, C. Rajapakse, A. Dangolla, and D.J. Melnick. 2003. Reliable noninvasive genotyping: fantasy or reality? Journal of Heredity 94: 115-123.

Foran, D.R., S.C. Minta, and K.S. Heinemeyer. 1997. DNA-based analysis of hair to identify species and individuals for population research and monitoring. Wildlife Society Bulletin 25: 840-847.

Ginot, F., I. Bordelais, S. Nguyen, and G. Gyapay. 1996. Correction of some genotyping errors in automated fluorescent microsatellite analysis by enzymatic removal of one base overhangs. Nucleic Acids Research 24: 540-541.

Goossens, B., L.P. Waits, and P. Taberlet. 1998. Plucked hair samples as a source of DNA: reliability of dinuccleotide microsatellite genotyping. Molecular Ecology 7: 1237-1241.

Hartl, D. L. 2000. A primer of population genetics. Sinauer Associates, Inc., Massachusetts.

Hoffman, J. I. and W. Amos. 2005. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology 14:599-612.

Litt, M., X. Hauge, and V. Sharma. 1993. Shadow bands seen when typing polymorphic dinucleotide repeats-some causes and cures. Biotechniques 15: 280.

Miller, C., P. Joyce, and L.P. Waits. 2002. Assessing allelic dropout and genotype reliability using maximum likelihood. Genetics 160: 357-366.

Mills, L. S., J. Citta, K. Lair, M. Schwartz, and D. Talmon. 2000. Estimating animal abundance using noninvasive DNA sampling: promises and pitfalls. Ecological Applications 10: 283-294.

Morin, P.A., K.E. Chambers, C. Boesch, and L. Vigilant. 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglogytes verus)*. Molecular Ecology 10: 1835-1844.

Parsons, K.M. 2001. Reliable microsatellite genotyping of dolphin DNA from faeces. Molecular Ecology 1: 341-344.

Petit, E. and N. Valière. 2006. Estimating population size with noninvasive capture-mark-recapture data. Conservation Biology 20: 1062-1073.

Proctor, M.F., B.N. McLellan, C. Strobeck, and R.M.R. Barclay. 2004. Gender-specific dispersal distances of grizzly bears estimated by genetic analysis. Canadian Journal of Zoology 82: 1108-1118.

Rodriguez, S., G. Visedo, and C. Zapata. 2001. Detection of errors in dinucleotide repeat typing by nondenaturing electrophoresis. Electrophoresis 22: 2656-2664.

Segelbacher, G. 2002. Noninvasive genetic analysis in birds: testing reliability of feather samples. Molecular Ecology Notes 2: 367-369.

Sloane, M.A., P. Sunnucks, D. Alpers, B. Beheregaray, and A.C. Taylor. 2000. Highly reliable genetic identification of individual northern hairy-nosed wombats from single remotely collected hairs: a feasible censusing method. Molecular Ecology 9: 123-124.

Smith, J.R., J.D. Carpten, M.J. Brownstein et al. 1995. Approach to genotyping errors caused by nontemplated nucleotide addition by *Taq* DNA-polymerase. Genome Research 5: 312-317.

Taberlet, P. ,S. Griffen, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L. P. Waits and J. Bouvet. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 24: 3189-3194.

Taberlet, P., J. Camarra, S. Griffin, E. Uhres, O. Hanotte, L.P. Waits, C. Dubois-Paganon, T. Burke, and J. Bouvet. 1997. Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Molecular Ecology 6: 869-876.

Taberlet, P. and G. Luikart. 1999. Noninvasive genetic sampling and individual identification. Biology Journal of the Linnean Society 68: 41-55.

Taberlet, P., L.P. Waits, G. Luikart. 1999. Noninvasive genetic sampling: look before you leap. Trends in Ecology and Evolution 14: 323-327.

Waits, L.P. 1999. Molecular genetic applications for bear research. Ursus 11: 253-260.

Wasser, S.K., C.S. Houston, G.M. Koehler, G.C. Cadd, and S.R. Fain. 1997. Techniques for application of faecal DNA methods to field studies of Ursids. Molecular Ecology 6: 1091-1097.

Woods, J.G., D. Paetkau, D. Lewis, B.N. McLellan, M. Proctor, and C. Strobeck. 1999. Genetic tagging of free-ranging black and brown bears. Wildlife Society Bulletin 27: 616-627.

CHAPTER 2

GENETIC MISIDENTIFICATION ERROR MODEL WITH SIMULATED CALIBRATION DATA[1]

[1] Sanderlin, J. S., N. Lazar, M.J. Conroy and J. Reeves. To be submitted to *Biometrics*.

Introduction

Animals that occur at low densities or have elusive behavior are difficult to sample for population inference, and often lead to low sample sizes with physical captures. Non-invasive sampling techniques, or techniques that do not require physical capture of animals, allow many populations of animal species to be monitored with greater detail. Advances within the fields of molecular and genetic biology have increased the ability to use genetic analyses in wildlife studies. Genetic samples (e.g., shed hairs, feathers, feces, shed skin), collected non-invasively in the field, are often small and contain degraded DNA. The ability to create multiple copies of DNA from these samples with PCR (polymerase chain reaction) has advanced non-invasive genetic sampling techniques profoundly (Waits 1999). Non-invasive genetic samples are currently utilized with many animal species globally for problems of demographics (Taberlet et al. 1997, Boulanger et al. 2002, Bellemain et al. 2005, Banks et al. 2003, Creel et al. 2003, Kendall et al. 2008, Prugh et al. 2005), habitat relationships (Apps et al. 2004), paternity and mating systems (Constable et al. 2001, Garnier et al. 2001, Oka and Takenaka 2001), and dispersal and effectiveness of corridors (Dixon et al. 2006, Proctor et al. 2004).

In any genetic analysis, especially non-invasive analyses, there are two common types of errors that occur at a specific locus with respect to alleles are allelic dropout and false alleles. Genetic errors can occur at various steps within a genetic study (e.g. sampling in the field, DNA extraction, molecular analysis, scoring of genotypes, data entry and analysis), and as a result of various sources (e.g., human error, technical error, biological processes) (Bonin et al. 2004). In allelic dropout, one or both of the two alleles may be lost during the analysis. Allelic dropout with a homozygous individual would not be problematic, since the individual would still be

considered a homozygote at that locus. However, with a heterozygous individual, the individual would be classified as homozygous under allelic dropout, leading to an error in classification. Allelic dropout is often a result from sampling stochasticity in the laboratory, amplification of small amounts of DNA and pipetting template DNA into a dilute sample (Goossens et al. 1998, Taberlet et al. 1999, Woods et al. 1999).

A false allele is the addition of one allele to a locus during the genetic analysis. True homozygous individuals with a false allele event would be classified as heterozygous at the locus, and true heterozygous individuals would have three alleles present. In the case of a true heterozygous individual with a false allele event, it is evident that an error occurred since only two alleles are possible at a locus; however, errors are not always clear with a true homozygous individual. False alleles can occur with PCR amplification artifacts from dinucleotide microsatellites (Goossens et al. 1998, Taberlet et al 1999, Rodriguez et al. 2001, and Woods et al. 1999) or contamination of samples. An accumulation of both types of genetic error may lead to misidentification of individuals within a population.

Mark-recapture models with genetic error

Recent studies have indicated population size estimates are sensitive to genetic error (Creel et al. 2003, Waits and Leberg 2000). As more non-invasive genetic studies are conducted, the reporting rate for genetic error has increased and even incorporated into parameter estimates (Figure 2.1, Appendix A). Reported error rates can be in the range of 0.001-0.92 per allele, locus, or study (Appendix B) and vary by sample type. The most common reported types of error include allelic dropout and false alleles with these studies. The method of quantifying genotyping errors, however, varies greatly between studies. Following notation from Broquet

and Petit (2004) in all formulas for genotyping errors, there are three common ways to calculate allelic dropout (ADO) and false alleles (FA), summarized in the ensuing paragraphs.

The unbiased estimate of allelic dropout (ADO_u) is essentially the number of errors detected in heterozygous genotypes divided by all heterozygous genotypes. Thus if p is the frequency of ADO at locus j:

$$\widehat{p}_j = \frac{D_j}{A_{het_j}}$$

(eqn 1)

This most often is reported as the ratio of observed ADO over L loci on the total number of heterozygous genotypes, or the weighted average of p_j for L loci (see Broquet and Petit 2004 for more details):

$$\widehat{p} = \frac{\sum_{j=1}^{L} D_j}{\sum_{j=1}^{L} A_{het_j}}$$

(eqn 2)

Biased estimates of ADO include: 1) ADO_1, which considers all PCR attempts (P_j), either successful or not or if individuals are homozygous or heterozygous, and 2) ADO_2, which considers all successful amplifications (A_j), regardless of homozygous or heterozygous. The estimator for ADO_1 is:

$$\widehat{p}_j = \frac{D_j}{P_j}$$

(eqn 3)

The estimator for ADO_2 is:

$$\widehat{p}_j = \frac{D_j}{A_j}$$

(eqn 4)

The unbiased estimate of FA (FA_u) is essentially the number of amplifications leading to the creation of one or more false alleles (F_j) at locus j, regardless if an individual is a true homozygote or heterozygote, divided by the total number of successful amplifications (A_j). The estimator is:

$$\widehat{f}_j = \frac{F_j}{A_j}$$

(eqn 5)

Again, over multiple loci L, the estimator is:

$$\widehat{f} = \frac{\sum_{j=1}^{L} F_j}{\sum_{j=1}^{L} A_j}$$

(eqn 6)

Biased estimates of FA include: 1) FA_1, the frequency of false alleles over all PCR attempts, regardless if successful or not, or 2) FA_2, only calculating FA with heterozygotes, assuming equal probabilities between homozygotes and heterozygotes. The estimator for FA_1 is:

$$\widehat{f}_j = \frac{F_j}{P_j}$$

(eqn 7)

Current methods of incorporating genetic error in non-invasive sampling mark-recapture models use maximum-likelihood methods (Lukacs and Burnham 2005, Kalinowski et al. 2006),

Bayesian methods (Wright et al. 2009, Petit and Valière 2006), and *ad hoc* approaches (Paetkau 2003, McKelvey and Schwartz 2004). However, a unified approach has not been adopted across genetic projects. Many approaches also require multiple PCR attempts to assess error (i.e., the multiple-tubes approach from Taberlet et al. 1996), which increases the cost per sample, with an undetermined amount of information gained. Therefore the cost per unit of information about genetic error ('efficiency') is a concern with genetic projects, particularly genetic projects with smaller budgets and non-invasive sampling projects.

We propose use of a calibration sample, which utilizes two different genetic sample types. Presumably, one type has a lower genetic error probability and is higher quality and the other type has a lower quality with a higher genetic error probability. Similar studies have used this type of genotype comparison with samples from tissue or blood to non-invasively collected samples (Wasser et al. 1997, Ernest et al. 2000, Sloane et al. 2000, Parsons 2001, Fernando et al. 2003), but have not incorporated a rigorous statistical technique of error estimation after the laboratory procedure. The genetic error calibration of non-invasive samples in a study utilizes a smaller subset of the samples by comparing genotypes from higher quality DNA types to lower quality DNA types. This project provides a general cost-effective approach for assessing genetic error in wildlife studies, with particular application to the central Georgia population (CGP) of black bears.

Estimation model

The possible events that occur in the misidentification error model closely follow the empirical penetrance model for pedigrees in Sobel et al. (2002). Empirical penetrance refers to the conditional probability of an observed genotype given the underlying true genotype. If

misidentification error is ignored through statistical models, then the observed outcomes either equal the true genotypes or they do not (i.e., binary). However, if error is included, there is some probability that the observed outcome is equal to the true genotype.

Here notation follows Hadfield et al. (2006), and originally in Wang (2004). We assume loci are independent in each individual, since we test for linkage disequilibrium during the selection of a marker panel for each genetic study. We compare a *n*x2 matrix for each locus, *l*, where *n* is the number of individuals genotyped, of true genotypes, **G,** to the *n*x2 matrix of observed genotypes, $\mathbf{G}^{(obs)}$, at locus, *l*. Since genotypes are comprised of two alleles per locus, notation must be specified for each allele at a locus. Specifically, the true genotypes can be specified as $\mathbf{G_{w,x}}$ with *w* and *x* representing the two possible alleles at one locus. The observed genotypes can be specified as $\mathbf{G}^{(obs)}_{\mathbf{u,v}}$ at one locus with *u* and *v* representing the possible observed alleles.

For simplicity, first consider the probability of observing one allele (A_u^{obs}) given the true allele (A_w) at one locus; there are three scenarios possible. The observed allele matches the true allele ($u = w$), the observed allele does not match the true allele ($u \neq w$), or the allele was not observed (*nd*, an allelic dropout event occurred and there is no data). We assume that two scenarios or processes can lead to the observed outcomes: a false allele event and allelic dropout. The probabilities associated with observing one allele given the true allele with *k* alleles possible at a given locus, e_1 the probability of one false allele event and e_2 the probability of one allele dropping out, are as follows:

$$\Pr(A_u^{obs} \mid A_w, e_1, e_2) = \begin{cases} (1-e_1(k-1))(1-e_2), & \text{where } u = w \\ e_1(k-1)(1-e_2), & \text{where } u \neq w \\ e_2, & \text{where } u = nd \end{cases} \qquad \text{(eqn. 8)}$$

Since our species has two independent alleles at a given locus, the probabilities of observing a two-locus genotype can be combined from the above probabilities. At any given locus, the true genotype can be either homozygous ($w = x$) or heterozygous ($w \neq x$). Given the true genotype is homozygous ($w = x$), there are three main categories for the observed genotype:

1) homozygous ($u = v$), with two possibilities

a) homozygous observed genotype is equal to the true genotype

($u = v = w$; $u = v, v = nd$; $v = w, u = nd$)

b) homozygous observed genotype is not equal to true genotype

($u = v \neq w$; $u \neq w$, $v = nd$; $v \neq w$, $u = nd$)

2) heterozygous ($u \neq v$), with two possibilities

a) heterozygous observed genotype contains one true allele

($u = w$, $u \neq v$; $v = w$, $u \neq v$)

b) heterozygous observed genotype contains no true alleles

($u \neq v \neq w$)

3) No observed genotype ($u = v = nd$)

The probability of observing a genotype given the true homozygous genotype with k alleles possible at a given locus (when k>2 per locus), where e_1 is the probability of one false allele event and e_2 the probability of one allele dropping out, follows a multinomial (1, p) distribution with the respective p probabilities for the different categories given below. See Appendix C for the special case of when k=2.

$$\Pr(\underline{G}_{u,v}^{obs} \mid \underline{G}_{w,x}, \underline{e}) = \begin{cases} (1-(k-1)e_1)^2(1-e_2)^2 + 2(1-(k-1)e_1)(1-e_2)e & \textit{where } u=v=w, \textit{ or } u=w, v=nd \\ 2e_1(k-1)(1-e_2)e_2 + (k-1)e_1^2(1-e_2)^2 & \textit{where } u=v\neq w, \textit{ or } u\neq w, v=nd \\ 2(1-(k-1)e_1)(k-1)e_1(1-e_2)^2 & \textit{where } u=w, u\neq v \textit{ or } v=w, u\neq v \\ [(k-1)^2-(k-1)]e_1^2(1-e_2)^2 & \textit{where } u\neq v\neq w \\ e_2^2 & \textit{where } u=v=nd \end{cases}$$

(eqn. 9)

Given that the true genotype is heterozygous ($w \neq x$), there are three possibilities for the observed genotype:

1) homozygous ($u = v$), with two possibilities

a) observed genotype is missing one allele

($u = v = w$) or ($u = v = x$)

b) observed genotype does not contain any true alleles

($u = v \neq w$) or ($u = v \neq x$)

2) heterozygous ($u \neq v$), with two possibilities

a) observed genotype is the true genotype

($u = w, v = x$)

b) observed genotype contains one true allele

($u = w, v \neq x$) or ($u = x, v \neq w$) or ($v = w, u \neq x$) or ($v = x, u \neq w$)

c) observed genotype contains no true alleles

($u \neq w, u \neq x, v \neq w, v \neq x$)

3) No observed genotype ($u = v = nd$)

Therefore, the probability of observing a genotype given the true heterozygous genotype when $k>2$ alleles per locus follows a multinomial $(1, p)$ distribution with the respective p probabilities for the different categories given below. The special case of $k=2$ is given in Appendix C.

$$\Pr(\underline{G}_{u,v}^{obs} \mid \underline{G}_{w,x}, \underline{e}) = \begin{cases} (1-(k-1)e_1)^2(1-e_2)^2 + e_1^2(1-e_2)^2 \\ 2(k-2)(1-(k-1)e_1)e_1(1-e_2)^2 + 2(k-2)e_1^2(1-e_2)^2 \\ (k-2)(k-3)e_1^2(1-e_2)^2 \\ 2(1-(k-1)e_1)e_2(1-e_2) + 2e_1e_2(1-e_2) + 2e_1(1-(k-1)e_1)(1-e_2)^2 \\ 2(k-2)e_1e_2(1-e_2) + (k-2)e_1^2(1-e_2)^2 \\ e_2^2 \end{cases}$$

$$\begin{array}{ll} where & u=w, v=x \quad or \quad u=x, v=w \\ where & u=w, v\neq x \quad or \quad u=x, v\neq x \quad or \quad v=w, u\neq x \quad or \quad v=x, u\neq w \\ where & u\neq w, u\neq x \quad or \quad v\neq w, v\neq x \\ where & u=v=w \quad or \quad u=v=x \quad or \quad u=w, v=nd \quad or \quad u=x, v=nd \\ where & u=v\neq w \quad or \quad u=v\neq x \quad or \quad u\neq w, v=nd \quad or \quad u\neq x, v=nd \\ where & u=v=nd \end{array}$$

(eqn 10)

The sample mean ($\bar{x}$) and sample variance (v) of both types of error calculated from previous studies (Appendix B) can be use as solutions for the parameters of a prior Beta distribution with the first two method-of-moments estimates. The two parameters (α, β) for the Beta distribution are approximated by the following: 1) $\alpha = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{v} - 1\right)$, and 2) $\beta = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{v} - 1\right)$. The prior distribution of e_1, or the probability of obtaining a false allele, would approximately follow a Beta (0.79,12.56) distribution, based on previous studies (n=10) (Figure 2.2, Figure 2.3). The prior distribution of e_2, or the probability of allelic dropout would approximately follow a Beta (0.55, 5.23) distribution, based on previous studies (n=27) (Figure 2.4). Figures were created with the base graphic package in program R (R Development Core Team 2008).

The true genotypes for each locus can be sampled conditional on the observed genotypes, allele frequencies (**a**) in the population, and the misidentification error rates (**e**):

$$\Pr(\underline{G} \mid \underline{G}^{obs}, \underline{a}, \underline{e}) \propto \Pr(\underline{G} \mid \underline{a}) \Pr(\underline{G}^{obs} \mid \underline{G}, \underline{e}) \qquad \text{(eqn 11)}$$

This step typically assumes that the population is in Hardy-Weinberg equilibrium (e.g., Hadfield et al. (2006)). However, this assumption can be relaxed. We assume that the observed genotypes and error probabilities are independent of the true state. In most cases, the true genotype distribution is unknown and must be included in the model. We are using a calibration sample that conditions on the tissue and blood samples as the true genotypes, and the observed genotypes as the hair samples. Therefore, we know the proportion of true genotypes and the allele distribution of our sample from the true genotype matrix. For the purposes of estimating genotyping error in our sample, error probabilities are the only parameters of interest in the calibration sample application.

To obtain the posterior distributions of the error probabilities given the observed hair samples and true genotypes and allele frequencies in the true blood and tissue samples, and using the likelihoods from equations 9 and 10, see below:

$$\Pr(e_1,e_2 \mid \underline{G}^{obs},\underline{G},\underline{a}) \propto \Pr(e_1)\Pr(e_2)\prod_{i=1}^{n_i}\prod_{l=1}^{n_l}\Pr(\underline{G}_{i,l}^{obs} \mid \underline{G}_{i,l\_},e_1,e_2) \qquad \text{(eqn 12)}$$

where n_i is the number of i individual samples, and n_l is the number of loci l, e_1 and e_2 are the probabilities of obtaining a false allele and allelic dropout, respectively.

The likelihood of the true genotype belonging to these categories is the likelihood from the error equations (above) multiplied by the probability of that genotype given allele frequencies, a, in the population. The true genotypes and allele frequencies are known and fixed under the calibration sample scenario. Essentially, this breaks down to:

$$[G \mid G^{\text{obs}}][G^{\text{obs}}] = [G^{\text{obs}} \mid G][G] \qquad \text{(eqn 13)}$$

$$[G \mid G^{\text{obs}}] \propto [G^{\text{obs}} \mid G][G] \qquad \text{(eqn 14)}$$

Or, more formally, including all parameters:

$$[G \mid G^{\text{obs}}, e_1, e_2, a] \propto [G^{\text{obs}} \mid G, e_1, e_2][G \mid a][a][e_1][e_2] \qquad \text{(eqn 15)}$$

If the likelihood of the true genotype in the observed sample is desired, simply multiply the known fixed quantities of [G|a] and [a] of the true genotype with the blood or tissue sample of individual i with the other likelihoods. For the purpose of this study, we are only concerned with estimating the posterior distributions of the error probabilities (eqn 12). The error probabilities can be combined into the full Bayesian model of abundance estimation of the central Georgia black bear population.

Error rates may differ by locus, which means some loci are more prone to errors than others. Error rates may differ by allele, specifically; studies indicate alleles with longer base pair length are more likely to contain error. Error rates may also differ by individual, due to poor sample quality, or groups of individuals, due to weather, time of year sample was collected, or length of time until DNA extraction. Further development of these models is warranted. We focus on the differences in error rates among loci, but not base pair length of alleles or individuals.

Simulated data with estimation model

Model verification is conducted in two steps with simulated data and using known individuals from the bear CGP. Simulated data, in the form of individual genotypes, were created using Python, version 2.5.2 (Python Software Foundation, http://python.org) programming language from known genotype frequencies at a specified amount of loci in a population. Following Hadfield et al. (2006), the true genotype of individual (i) at locus (l) can be sampled from a multinomial distribution with one trial, and $k_l(k_l+1)/2$ categories of

genotypes, where k is the number of alleles at a specific locus, and multinomial cell probabilities from a uniform Dirichlet distribution with $k_l(k_l+1)/2$ categories. This eliminates the restriction of Hardy-Weinberg equilibrium, which is often a poor assumption to make with natural populations. Samples were drawn from the given population of genotypes to simulate a random sample for use as the calibration sample.

After samples were drawn, they were assigned an observed genotype, based on the multinomial probabilities of observing a genotype given the true genotype and error rates (equations 9 and 10). For each iteration under every scenario, the error rates were selected from beta distributions based on previous genetic studies (See previous section). Table 2.1 describes the parameter levels (N, n, alleles per locus, number of loci) of simulated genetic data. For each scenario, 500 replicates of the simulation-MCMC estimation algorithm were processed with the Metropolis-Hastings algorithm implemented in PyMC 2.0 (Patil, A., Huard, D. and Fonnesbeck. C, http://pymc.googlecode.com) in Python, version 2.5.2 (Python Software Foundation, http://python.org) on 20 PCs and evaluated with the frequentist properties of Bayesian credible interval (BCI) percent coverage, BCI length, relative mean squared error (RRMSE), and relative bias (RBIAS) for the error parameters. Each data set replicate of the simulation-MCMC process includes posterior estimates based on 10,000 iterations with a burn-in period of 5,000 iterations and no thinning. Each replicate took 1 second (or 8.3 hours for 500 replications).

The 95% Bayesian credible interval (BCI) coverage was evaluated by summing the number of replications where the true parameters (error rates) were contained in the BCI and dividing this value by the total number of replications. The BCI interval is the distance between the lower and upper 95% credible interval values. Relative root mean-squared error (RRMSE) with r being the total number of replicates, i the individual replicate, $\hat{\theta}_i$, the estimated parameter

at replicate i, θ_i the true value of parameter at replicate i, and $\bar{\theta}$ the mean of the true parameter values over all replicates, was calculated as:

$$RRMSE = \frac{\sqrt{\frac{1}{r}\sum_{i}^{n}(\theta_i - \theta_i)^2}}{\bar{\theta}}$$

Relative bias (RBIAS), following the same notation as above, was calculated as:

$$RBIAS = \frac{1/r\sum_{i}^{n}(\hat{\theta}_i - \theta_i)}{\bar{\theta}}$$

Results

Percent coverage of Bayesian credible interval

The error parameters for false alleles (e_1) allelic dropout (e_2) did not have a consistent pattern with the BCI over all levels of true population size, calibration sample size, number of loci or number of alleles per locus (Table 2.2, Figure 2.5 and 2.6). The BCIs were high (0.92-0.96) and near the nominal frequentist value of 0.95.

Bayesian credible interval length

The error parameters for false alleles (e_1) allelic dropout (e_2) did not have a consistent pattern with the BCI length over all levels of true population size and the number of alleles per locus (Table 2.2, Figure 2.7 and 2.8). There was a decrease in the interval length with an increase in the calibration sample size and an increase in the number of loci for both error types.

Relative bias

The error parameters for false alleles (e_1) allelic dropout (e_2) did not have a consistent pattern of relative bias over all levels of true population size, number of loci and number of alleles per locus (Table 2.2, Figure 2.9 and 2.10). There was a decrease in bias with an increase in calibration sample size for both error types.

Relative root mean square *error*

The error parameters for false alleles (e_1) allelic dropout (e_2) did not have a consistent pattern with the RRMSE over all levels of true population size and number of alleles per locus (Table 2.2, Figure 2.11 and 2.12). There was a decrease in RRMSE, or increase in accuracy, with an increase in calibration sample size and an increase in number of loci for both error types.

Sample size

The true population size and calibration sample size were both increased to 1000 to evaluate the influence of sample size with the best (8 alleles per locus and 15 loci) and worst (3 alleles per locus and 5 loci) scenarios. The percent BCI coverage was the same as the original lower sample size values, but the interval length was smaller (Table 2). The relative bias and RRMSE were also lower (Table 2.2).

Sensitivity to priors

To evaluate the sensitive of priors, a Beta (1,1) distribution was selected for both error rates with the best (true population size of 500, calibration sample size of 75, 8 alleles per locus, and 15 loci) and worst (true population size of 100, calibration sample size of 25, 3 alleles per

locus, 5 loci) scenarios. This is a uniform prior, or a non-informative prior. The percent BCI coverage was slightly lower compared to the 81 simulation combinations, although the interval length was about the same (Table 2.2). Relative bias and RRMSE were slightly higher for the worst scenario and lower for the best scenario, compared with the 81 simulation combinations for both error rates.

Values of error rate

The values for the prior distributions of both error rates were low and might influence the frequentist measures of BCI percent coverage and length, relative bias, and RRMSE. Therefore, a Beta(15,15) distribution was selected with both error rates (or error rates of 0.5 with some variance) for the best (true population size of 500, calibration sample size of 75, 8 alleles per locus, and 15 loci) and worst (true population size of 100, calibration sample size of 25, 3 alleles per locus, and 5 loci) scenarios. The BCI percent coverage was about the same over all the 81 simulation combinations for both error types (Table 2.2). The relative bias and RRMSE were slightly smaller with this prior distribution for both error types and for both the best and worst scenarios. This indicates a slight bias with the use of small values, or a different prior distribution, for error rates (Table 2.2).

Discussion and conclusions

There have been several efforts to minimize the total error in genetic studies within the categories of: 1) laboratory methods, 2) pilot studies (Taberlet and Luikart 1999), and 3) simulations (Taberlet et al. 1996, Petit and Valière 2006). Laboratory methods include: the

multiple tubes approach where DNA samples are amplified independently several times (Taberlet et al. 1996), reamplifying loci prone to error (Miller et al. 2002) and selecting samples based on some criteria of DNA quantity (Morin et al. 2001, Segelbacher 2002). Many of these methods are specific to the study organism and genetic project, thus there is minimal generality among studies. Statistical methods that provide a general framework to assess genetic error allow comparisons to be made across projects and species, as the use of non-invasive genetic sampling increases among bear and wildlife projects. Our model provides a cost-efficient method of estimating genetic error. The error probability could then be incorporated into other models like abundance models, for example.

Percent coverage of Bayesian credible interval and length

The percent coverage for both error parameters was high and within the nominal rate of 0.95. There was a decrease in the interval length with an increase in the calibration sample size and an increase in the number of loci for both error types. This should be expected as sample size increases in any study.

Relative bias

Similarly, as sample size increased, there was a decrease in bias with an increase in calibration sample size for both error types. One would not expect a pattern with the other parameters.

Relative root mean square *error*

There was a decrease in relative root mean square error, or an increase in accuracy, with

an increase in calibration sample size and an increase in number of loci for both error types. This follows with an increase in sample size. The other parameters of true population size and number of alleles per locus should not affect accuracy, and would not be desired since this indicates a bias in the model.

Sample size

With and increase in sample size, the percent Bayesian credible interval coverage was the same as the original lower sample size values, but the interval length was smaller. This indicates a slight advantage of estimation with larger sample sizes. The relative bias and accuracy were also lower, which can be attributed to the larger sample size.

Sensitivity to priors

To determine if the model was sensitive to selection of a prior distribution, we choose an uninformative prior. With this prior the percent BCI coverage was slightly lower compared to the eighty-one levels of simulations, even though the interval length was about the same. This indicates a more informative distribution may be optimal. This could be a result of the low expected probabilities of error (near zero), compared to levels of error that would fall around fifty percent. Relative bias and accuracy were slightly higher for the worst scenario with the uninformative prior, but lower for the best scenario, compared to the other simulation levels for both error rates. In conclusion, the model is slightly affected by the choice of a prior distribution. However, the results indicate that the error estimates are still within acceptable limits.

Values of error rate

Since a low expected error rate with both allelic dropout and false alleles may bias results, a higher error rate was used in the simulation/estimation process. The BCI percent coverage was about the same as the simulation levels for both error types. The relative bias and accuracy were slightly smaller with this prior distribution for both error types and for both the best and worst scenarios. This indicates a slight bias with the use of small values, or a different prior distribution, for error rates.

In conclusion, the calibration sample study with the error estimation model is valid. The frequentist properties of percent coverage and interval length, relative bias, and root mean square error were adequate. Sample size, sensitivity to priors and different values of error rate influences the frequentist properties slightly, and should be considered when applying real data. It would probably be optimal to use more informative priors in a genetic study, as outlined in this chapter.

Literature Cited

Apps, C. D., B. N. McLellan, J. G. Woods, and M. F. Proctor. 2004. Estimating grizzly bear distribution and abundance relative to habitat and human influence. Journal of Wildlife Management 68: 138-152.

Banks, S.C., S.D. Hoyle, A. Horsup, P. Sunnucks and A.C. Taylor. 2003. Demographic monitoring of an entire species (the northern hairy-nosed wombat, *Lasiorhinus krefftii* by genetic analysis of non-invasively collected material. Animal Conservation 6: 101 107.)

Bellemain, E., J.E. Swenson, D. Tallmon, S. Brunberg, and P. Taberlet. 2005. Estimating - population size of elusive animals with DNA from hunter-collected feces: for brown bears. Conservation Biology 19: 150-161.

Bonin, A., E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. four methods 2004. How to track and assess genotyping errors in population genetics studies. Molecular Ecology 13: 3261-3273.

Boulanger, J., G. C. White, B. N. McLellan, J. Woods, M. Proctor, and S. Himmer. 2002. A meta-analysis of grizzly bear DNA mark-recapture projects in British Columbia, Canada. Ursus 13: 137-152.

Broquet, Thomas and Eric Petit. 2004. Quantifying genotyping errors in noninvasive population genetics. Molecular Ecology 13: 3601-3608.

Constable, J. L., M. V. Ashley, J. Goodall, and A. E. Pusey. 2001. Noninvasive paternity assignment in Gombe chimpanzees. Molecular Ecology 10: 1279-1300.

Creel, Scott, G. Spong, J.L. Sands et al. 2003. Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12: 2003-2009.

Dixon, J.D., M.O. Oli, M.C. Wooten, T.H. Eason, J.W. McCown, and D. Paetkau. 2006. Effectiveness of a regional corridor in connecting two Florida black bear populations. Conservation Biology 20: 155-162.

Ernest, H.B., M.C.T. Penedo, B.P. May, M. Syvanen, and W.M. Boyce. 2000. Molecular tracking of mountain lions in the Yosemite Valley region in California: genetic analysis using microsatellites and faecal DNA. Molecular Ecology 9: 433-441.

Fernando, P., T.N.C. Vidya, C. Rajapakse, A. Dangolla, and D.J. Melnick. 2003. Reliable noninvasive genotyping: fantasy or reality? Journal of Heredity 94: 115-123.

Garnier, J. N., M. W. Bruford, and B. Goossens. 2001. Mating system and reproductive skew in the black rhinoceros. Molecular Ecology 10: 2031-2041.

Goossens, B., L.P. Waits, and P. Taberlet. 1998. Plucked hair samples as a source of DNA: reliability of dinuccleotide microsatellite genotyping. Molecular Ecology 7: 1237-1241.

Hadfield, J.D., D.S. Richardson, and T. Burke. 2006. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. Molecular Ecology 15: 3715-3730.

Kalinowski, S.T., M.L. Taper, and S. Creel. 2006. Using DNA from non-invasive samples to identify individuals and census populations: and evidential approach tolerant of genotyping errors. Conservation Genetics 7: 319-329.

Kendall, K. C., J. B. Stetz, D. A. Roon, L. P. Waits, J. B. Boulanger, and D. Paetkau. 2008. Grizzly bear density in Glacier National Park, Montana. Journal of Wildlife Management 72: 1693-1705.

Lukacs, P. M. and K.P. Burnham. 2005. Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. Journal of Wildlife Management 68: 439-448.

McKelvey, Kevin and Michael Schwartz. 2004. Genetic errors associated with population estimation using non0invasive molecular tagging: problems and new solutions. Journal of Wildlife Management 68: 439-448.

Miller, C., P. Joyce, and L.P. Waits. 2002. Assessing allelic dropout and genotype reliability using maximum likelihood. Genetics 160: 357-366.

Morin, P.A., K.E. Chambers, C. Boesch, and L. Vigilant. 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglogytes verus)*. Molecular Ecology 10: 1835-1844.

Oka, T. and O. Takenaka. 2001. Wild Gibbons' parentage tested by non-invasive DNA sampling and PCR-amplified polymorphic microsatellites. Primates 42: 67-73.

Paetkau, D. 2003. An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology 12:1375-1387.

Parsons, K.M. 2001. Reliable microsatellite genotyping of dolphin DNA from faeces. Molecular Ecology 1: 341-344.

Petit, E. and N. Valière. 2006. Estimating population size with noninvasive capture-mark-recapture data. Conservation Biology 20: 1062-1073.

Proctor, M.F., B.N. McLellan, C. Strobeck, and R.M.R. Barclay. 2004. Gender-specific dispersal distances of grizzly bears estimated by genetic analysis. Canadian Journal of Zoology 82: 1108-1118.

Prugh, L. R., C. E. Ritland, S. M. Arthur, and C. J. Krebs. 2005. Monitoring coyote popoulation dynamics by genotyping faeces. Molecular Ecology 14: 1585-1596.

Rodriguez, S., G. Visedo, and C. Zapata. 2001. Detection of errors in dinucleotide repeat typing by nondenaturing electrophoresis. Electrophoresis 22: 2656-2664.

Segelbacher, G. 2002. Noninvasive genetic analysis in birds: testing reliability of feather samples. Molecular Ecology Notes 2: 367-369.

Sloane, M.A., P. Sunnucks, D. Alpers, B. Beheregaray, and A.C. Taylor. 2000. Highly reliable genetic identification of individual northern hairy-nosed wombats from single remotely collected hairs: a feasible censusing method. Molecular Ecology 9: 123-124.

Sobel, E., J. C. Papp, and K. Lange. 2002. Detection and integration of genotyping errors in statistical genetics. American Journal of Human Genetics 70: 496-508.

Taberlet, P. ,S. Griffen, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L. P. Waits and J. Bouvet. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 24: 3189-3194.

Taberlet, P., J. Camarra, S. Griffin, E. Uhres, O. Hanotte, L.P. Waits, C. Dubois-Paganon, T. Burke, and J. Bouvet. 1997. Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Molecular Ecology 6: 869-876.

Taberlet, P., L.P. Waits, G. Luikart. 1999. Noninvasive genetic sampling: look before you leap. Trends in Ecology and Evolution 14: 323-327.

Waits, L.P. 1999. Molecular genetic applications for bear research. Ursus 11: 253-260.

Waits, Juliann L. and Paul L. Leberg. 2000. Biases associated with population estimation using molecular tagging. Animal Conservation 3:191-199.

Wang, J. 2004. Sibship reconstruction from genetic data with typing errors. Genetics 166: 1963-1979.

Wasser, S.K., C.S. Houston, G.M. Koehler, G.C. Cadd, and S.R. Fain. 1997. Techniques for application of faecal DNA methods to field studies of Ursids. Molecular Ecology 6: 1091-1097.

Wright, J. A., R. J. Barker, M. R. Schofield, A. C. Frantz, A. E. Byrom, and D. M. Gleeson. 2009. Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. *In press*: Biometrics.

Woods, J.G., D. Paetkau, D. Lewis, B.N. McLellan, M. Proctor, and C. Strobeck. 1999. Genetic tagging of free-ranging black and brown bears. Wildlife Society Bulletin 27: 616-627.

Table 2.1. Simulation parameters, or the parameter levels, and a description of why the levels were selected.

Parameter	Levels	Description
True population size (N)	100, 250, 500	Low, medium, and high levels of true population sizes for bear populations
Calibration sample size (n)	25, 50, 75	Low, medium, and high numbers of samples used in a calibration sample. A laboratory will likely not sample more than 100 individuals in a pilot study, based on budget constraints.
Alleles per locus (k)	3, 5, 8	These numbers are the average number of alleles per locus (fixed average, but varied alleles at specific loci) that reflect low, medium, and high levels of diversity with bear populations.
Number of loci (l)	5, 10, 15	The number of loci used to assess genotypes were chosen to have low, medium, and high numbers, based on previous bear studies with goals ranging from estimating population size (less loci) to parentage assignments (more loci).
TOTAL COMBINATIONS	81	

Table 2.2. Summary of all simulations of all levels

True pop. size	Calib. sample size	alleles per loc	loci	true mean e_1	true mean e_2	e_1 % BCI	e_2 % BCI	e_1 RBIAS	e_2 RBIAS	e_1 RRMSE	e_2 RRMSE	e_1 BCI mean length	e_2 BCI mean length
100	25	3	5	0.058	0.093	0.946	0.954	0.019	-0.002	0.302	0.245	0.055	0.079
250	25	3	5	0.058	0.094	0.946	0.938	-0.016	0.019	0.279	0.233	0.054	0.079
500	25	3	5	0.056	0.097	0.944	0.932	-0.011	-0.008	0.277	0.244	0.053	0.078
100	50	3	5	0.055	0.097	0.932	0.936	0.013	-0.002	0.222	0.169	0.039	0.058
250	50	3	5	0.059	0.095	0.948	0.934	-0.003	0.006	0.198	0.174	0.040	0.057
500	50	3	5	0.062	0.085	0.930	0.942	-0.013	-0.007	0.196	0.185	0.040	0.055
100	75	3	5	0.059	0.093	0.948	0.940	-0.008	0.014	0.165	0.137	0.033	0.047
250	75	3	5	0.058	0.100	0.964	0.942	0.014	-0.003	0.164	0.123	0.033	0.047
500	75	3	5	0.061	0.096	0.940	0.934	-0.005	-0.010	0.167	0.143	0.033	0.046
100	25	5	5	0.061	0.088	0.944	0.938	-0.018	0.000	0.267	0.225	0.053	0.067
250	25	5	5	0.059	0.100	0.928	0.940	-0.008	0.001	0.291	0.230	0.053	0.069
500	25	5	5	0.058	0.095	0.938	0.938	0.014	-0.009	0.278	0.217	0.052	0.068
100	50	5	5	0.057	0.096	0.950	0.930	-0.001	-0.003	0.195	0.157	0.038	0.050
250	50	5	5	0.060	0.101	0.950	0.954	-0.009	0.004	0.183	0.143	0.040	0.052
500	50	5	5	0.057	0.095	0.924	0.938	-0.006	0.010	0.219	0.171	0.038	0.050
100	75	5	5	0.058	0.100	0.934	0.922	-0.004	-0.003	0.163	0.125	0.032	0.042
250	75	5	5	0.065	0.091	0.938	0.916	0.002	-0.002	0.156	0.138	0.033	0.040
500	75	5	5	0.059	0.097	0.930	0.950	-0.006	-0.002	0.158	0.117	0.032	0.041
100	25	8	5	0.055	0.092	0.946	0.958	0.022	-0.012	0.276	0.191	0.050	0.063
250	25	8	5	0.066	0.095	0.940	0.940	-0.003	-0.009	0.245	0.194	0.054	0.064
500	25	8	5	0.056	0.096	0.942	0.934	0.001	0.012	0.281	0.197	0.052	0.065
100	50	8	5	0.053	0.091	0.946	0.934	-0.001	0.004	0.182	0.151	0.035	0.045
250	50	8	5	0.062	0.098	0.926	0.942	-0.002	-0.003	0.191	0.137	0.038	0.046
500	50	8	5	0.061	0.106	0.942	0.944	0.007	-0.004	0.181	0.135	0.038	0.047
100	75	8	5	0.060	0.097	0.934	0.930	-0.005	-0.012	0.162	0.117	0.031	0.038
250	75	8	5	0.063	0.097	0.948	0.954	-0.004	-0.004	0.149	0.104	0.032	0.038

Table 2.2. (Continued) Summary of all simulations of all levels.

True pop. size	Calib. sample size	alleles per loc	loci	true mean e_1	true mean e_2	e_1 % BCI	e_2 % BCI	e_1 RBIAS	e_2 RBIAS	e_1 RRMSE	e_2 RRMSE	e_1 BCI mean length	e_2 BCI mean length
500	75	8	5	0.059	0.092	0.936	0.946	0.004	0.007	0.164	0.112	0.031	0.037
100	25	3	10	0.055	0.102	0.970	0.912	-0.002	0.002	0.188	0.176	0.039	0.058
250	25	3	10	0.058	0.102	0.944	0.918	0.017	-0.017	0.223	0.174	0.040	0.058
500	25	3	10	0.062	0.100	0.954	0.934	-0.004	0.014	0.189	0.173	0.041	0.058
100	50	3	10	0.060	0.092	0.938	0.938	-0.001	-0.002	0.133	0.124	0.028	0.041
250	50	3	10	0.059	0.095	0.936	0.942	-0.001	0.003	0.115	0.093	0.029	0.042
500	50	3	10	0.060	0.094	0.956	0.942	-0.002	-0.009	0.133	0.125	0.029	0.041
100	75	3	10	0.056	0.096	0.938	0.936	-0.002	-0.004	0.113	0.095	0.023	0.034
250	75	3	10	0.059	0.091	0.944	0.938	0.002	-0.005	0.113	0.105	0.023	0.034
500	75	3	10	0.060	0.095	0.938	0.936	0.004	-0.004	0.116	0.105	0.024	0.034
100	25	5	10	0.056	0.087	0.950	0.948	-0.007	0.004	0.195	0.160	0.037	0.048
250	25	5	10	0.059	0.088	0.932	0.934	-0.006	0.003	0.203	0.170	0.038	0.049
500	25	5	10	0.062	0.096	0.928	0.950	-0.010	-0.012	0.199	0.144	0.039	0.050
100	50	5	10	0.060	0.096	0.932	0.930	-0.004	0.004	0.138	0.112	0.027	0.035
250	50	5	10	0.058	0.087	0.942	0.940	-0.007	-0.009	0.139	0.112	0.027	0.035
500	50	5	10	0.058	0.093	0.938	0.940	0.002	0.002	0.144	0.113	0.028	0.035
100	75	5	10	0.064	0.093	0.930	0.956	-0.003	0.000	0.113	0.085	0.024	0.030
250	75	5	10	0.059	0.096	0.924	0.936	0.010	0.003	0.125	0.096	0.023	0.030
500	75	5	10	0.059	0.101	0.942	0.948	0.004	-0.001	0.115	0.076	0.023	0.030
100	25	8	10	0.059	0.084	0.946	0.950	-0.009	0.006	0.193	0.153	0.037	0.044
250	25	8	10	0.063	0.097	0.934	0.938	0.002	-0.004	0.188	0.142	0.039	0.046
500	25	8	10	0.061	0.089	0.936	0.932	-0.011	-0.002	0.175	0.157	0.038	0.045
100	50	8	10	0.060	0.099	0.950	0.940	-0.001	-0.001	0.128	0.097	0.027	0.033
250	50	8	10	0.060	0.095	0.942	0.940	0.000	0.002	0.128	0.105	0.027	0.033
500	50	8	10	0.057	0.096	0.930	0.934	-0.006	-0.006	0.141	0.100	0.027	0.033
100	75	8	10	0.059	0.093	0.956	0.940	-0.006	-0.004	0.112	0.084	0.022	0.026

Table 2.2. (Continued) Summary of all simulations of all levels.

True pop. size	Calib. sample size	alleles per loc	loci	true mean e_1	true mean e_2	e_1 % BCI	e_2 % BCI	e_1 RBIAS	e_2 RBIAS	e_1 RRMSE	e_2 RRMSE	e_1 BCI mean length	e_2 BCI mean length
250	75	8	10	0.062	0.087	0.956	0.942	-0.006	-0.002	0.098	0.083	0.022	0.025
500	75	8	10	0.059	0.102	0.958	0.944	0.000	0.003	0.105	0.079	0.022	0.028
100	25	3	15	0.059	0.088	0.952	0.946	0.005	0.000	0.167	0.154	0.033	0.046
250	25	3	15	0.062	0.087	0.950	0.944	-0.006	0.006	0.164	0.156	0.034	0.047
500	25	3	15	0.058	0.097	0.922	0.952	0.002	-0.007	0.176	0.137	0.033	0.048
100	50	3	15	0.057	0.090	0.958	0.944	0.003	0.001	0.119	0.107	0.023	0.033
250	50	3	15	0.058	0.103	0.934	0.950	0.008	-0.005	0.126	0.094	0.024	0.034
500	50	3	15	0.057	0.094	0.944	0.938	-0.006	0.005	0.125	0.105	0.023	0.033
100	75	3	15	0.056	0.091	0.952	0.944	-0.010	-0.006	0.091	0.090	0.019	0.028
250	75	3	15	0.057	0.094	0.950	0.924	-0.002	0.003	0.094	0.089	0.019	0.029
500	75	3	15	0.056	0.096	0.946	0.928	0.013	0.001	0.100	0.087	0.019	0.028
100	25	5	15	0.058	0.089	0.916	0.962	-0.002	0.009	0.175	0.124	0.031	0.040
250	25	5	15	0.059	0.096	0.956	0.950	-0.012	-0.002	0.165	0.111	0.032	0.041
500	25	5	15	0.052	0.097	0.954	0.950	0.000	0.005	0.160	0.121	0.030	0.041
100	50	5	15	0.062	0.091	0.942	0.926	-0.005	-0.005	0.112	0.091	0.023	0.029
250	50	5	15	0.062	0.090	0.932	0.948	-0.005	0.004	0.114	0.089	0.023	0.029
500	50	5	15	0.057	0.090	0.936	0.932	-0.007	-0.010	0.114	0.094	0.022	0.028
100	75	5	15	0.060	0.097	0.928	0.952	0.001	-0.001	0.094	0.073	0.019	0.024
250	75	5	15	0.062	0.099	0.934	0.940	0.001	-0.001	0.085	0.075	0.019	0.024
500	75	5	15	0.060	0.087	0.922	0.932	-0.008	-0.001	0.096	0.081	0.018	0.023
100	25	8	15	0.061	0.099	0.946	0.930	0.003	-0.014	0.152	0.116	0.032	0.039
250	25	8	15	0.061	0.106	0.942	0.940	-0.001	-0.021	0.155	0.111	0.032	0.039
500	25	8	15	0.059	0.095	0.942	0.928	0.004	-0.003	0.157	0.125	0.031	0.037
100	50	8	15	0.056	0.091	0.944	0.954	-0.003	0.001	0.111	0.078	0.021	0.027
250	50	8	15	0.056	0.100	0.946	0.944	0.002	0.007	0.111	0.079	0.021	0.027
500	50	8	15	0.060	0.095	0.946	0.958	-0.009	-0.001	0.112	0.085	0.022	0.027

Table 2.2. (Continued) Summary of all simulations of all levels.

True pop. size	Calib. sample size	alleles per loc	loci	true mean e_1	true mean e_2	e_1 % BCI	e_2 % BCI	e_1 RBIAS	e_2 RBIAS	e_1 RRMSE	e_2 RRMSE	e_1 BCI mean length	e_2 BCI mean length
100	75	8	15	0.062	0.097	0.944	0.924	0.007	0.004	0.086	0.072	0.018	0.022
250	75	8	15	0.058	0.101	0.932	0.932	0.004	0.000	0.091	0.066	0.018	0.023
500	75	8	15	0.059	0.102	0.930	0.940	0.008	-0.006	0.093	0.063	0.018	0.023
beta(15,15) simulation and estimation model													
100	25	3	5	0.503	0.498	0.968	0.940	-0.005	0.004	0.090	0.063	0.181	0.123
500	75	8	15	0.502	0.499	0.926	0.940	0.001	0.000	0.034	0.021	0.062	0.041
sample size													
1000	1000	3	5	0.062	0.089	0.940	0.932	0.002	0.000	0.044	0.043	0.010	0.013
1000	1000	8	15	0.060	0.097	0.930	0.946	0.000	0.000	0.025	0.019	0.005	0.006
beta(1,1) prior in estimation model													
100	25	3	5	0.061	0.094	0.924	0.910	0.088	0.074	0.318	0.266	0.060	0.084
500	75	8	15	0.059	0.093	0.926	0.914	0.006	0.009	0.100	0.069	0.018	0.022

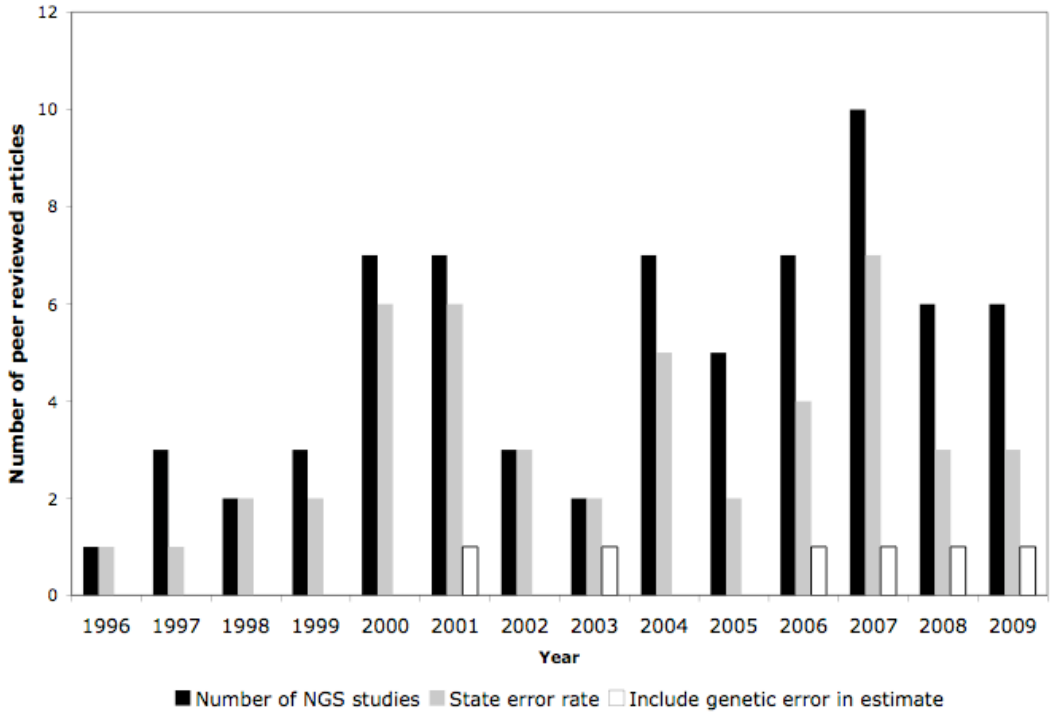


Figure 2.1. Number of peer-reviewed articles for non-invasive genetic studies that state the error rate for their study and/or include genetic error into parameter estimates.

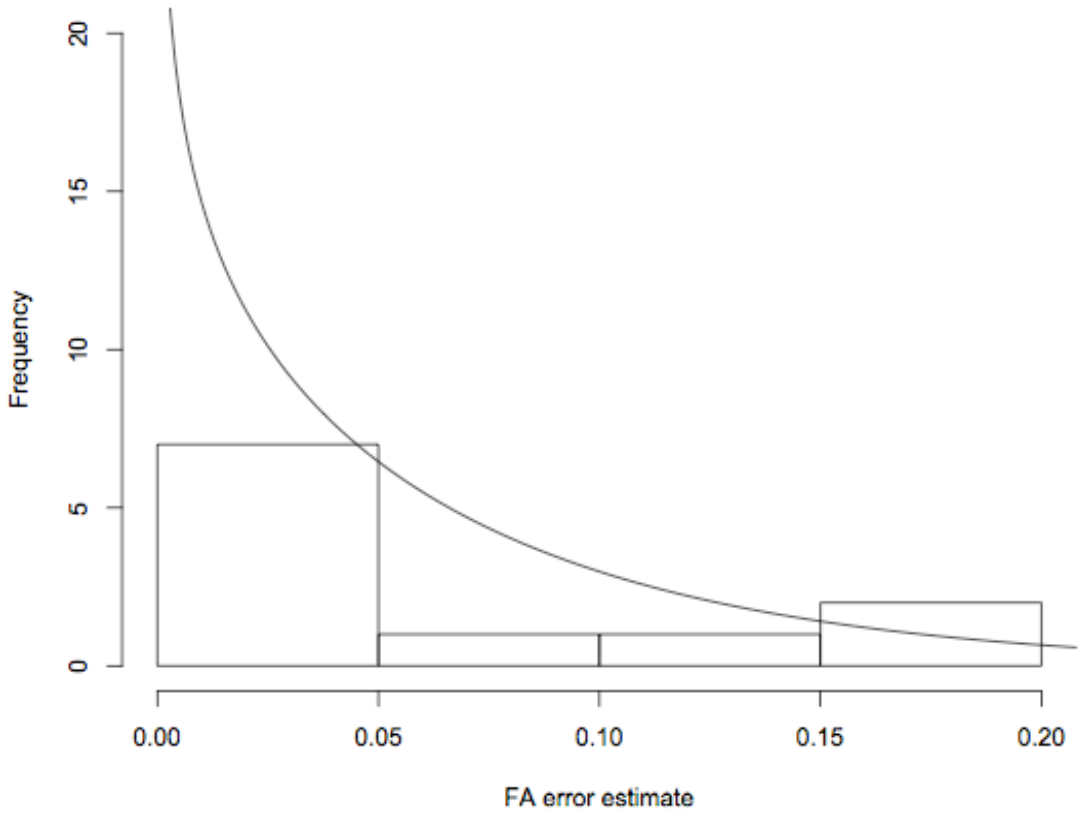


Figure 2.2. Prior distribution of e_1, or the probability of obtaining a false allele. The distribution approximately follows a Beta (0.79,12.56) distribution, based on previous studies (n=11).

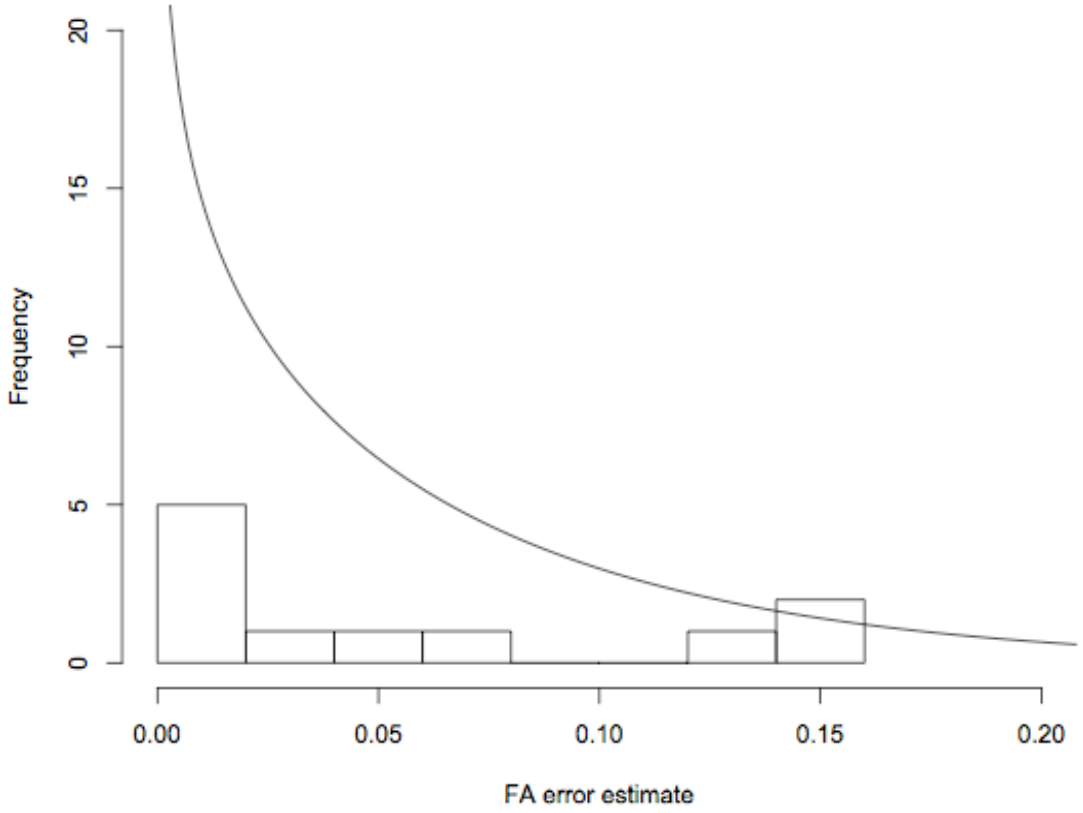


Figure 2.3. Prior distribution of e_1, or the probability of obtaining a false allele. The distribution approximately follows a Beta (0.79,12.56) distribution, based on previous studies (n=11) with a different number of bins with the histogram.

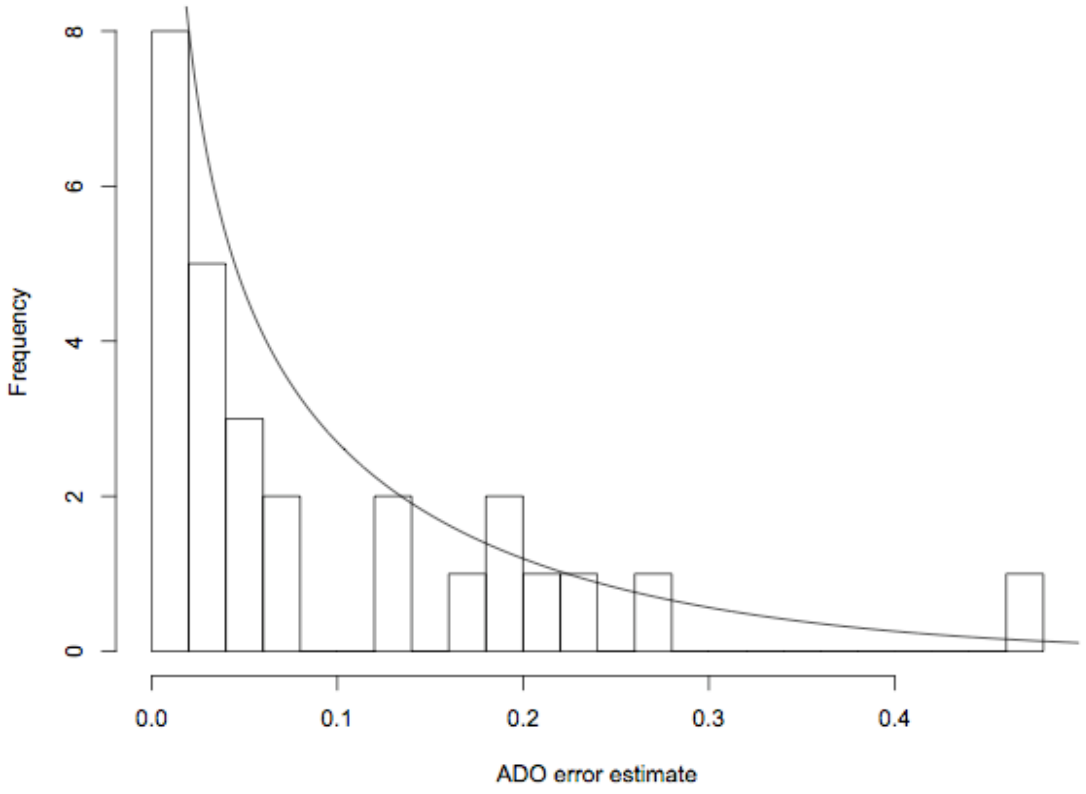


Figure 2.4. Prior distribution of e_2, or the probability of allelic dropout. The distribution approximately follows a Beta (0.55, 5.23) distribution, based on previous studies (n=27).

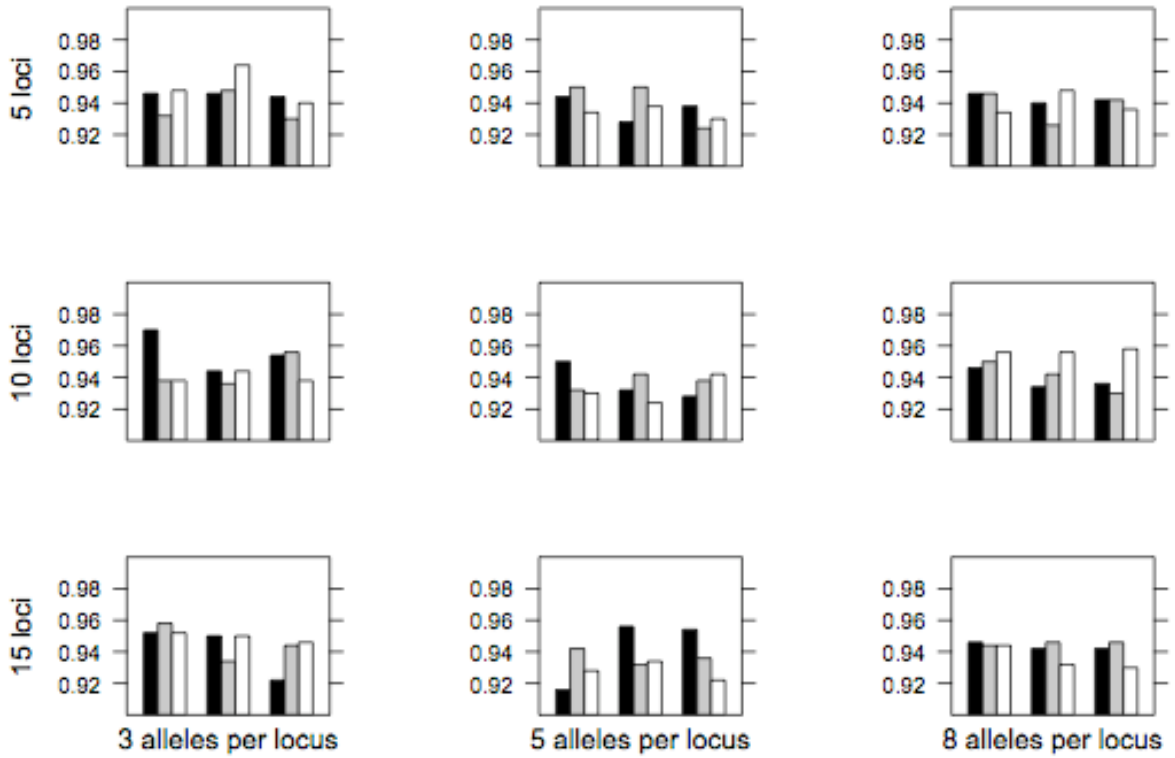


Figure 2.5. Percent BCI coverage for false allele (e_1) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (black bars), 50 (gray bars), and 75 (white bars) with true population sizes of 100, 250, and 500.

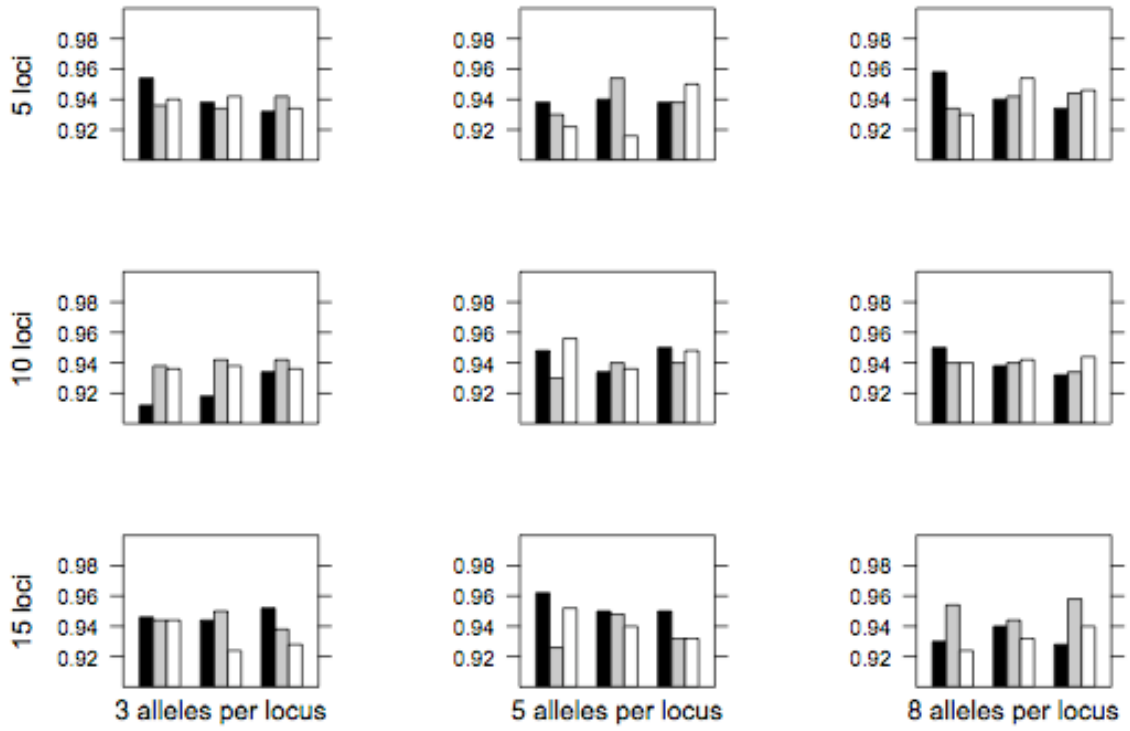


Figure 2.6. Percent BCI coverage for allelic dropout (e_2) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (black bars), 50 (gray bars), and 75 (white bars) with true population sizes of 100, 250, and 500.

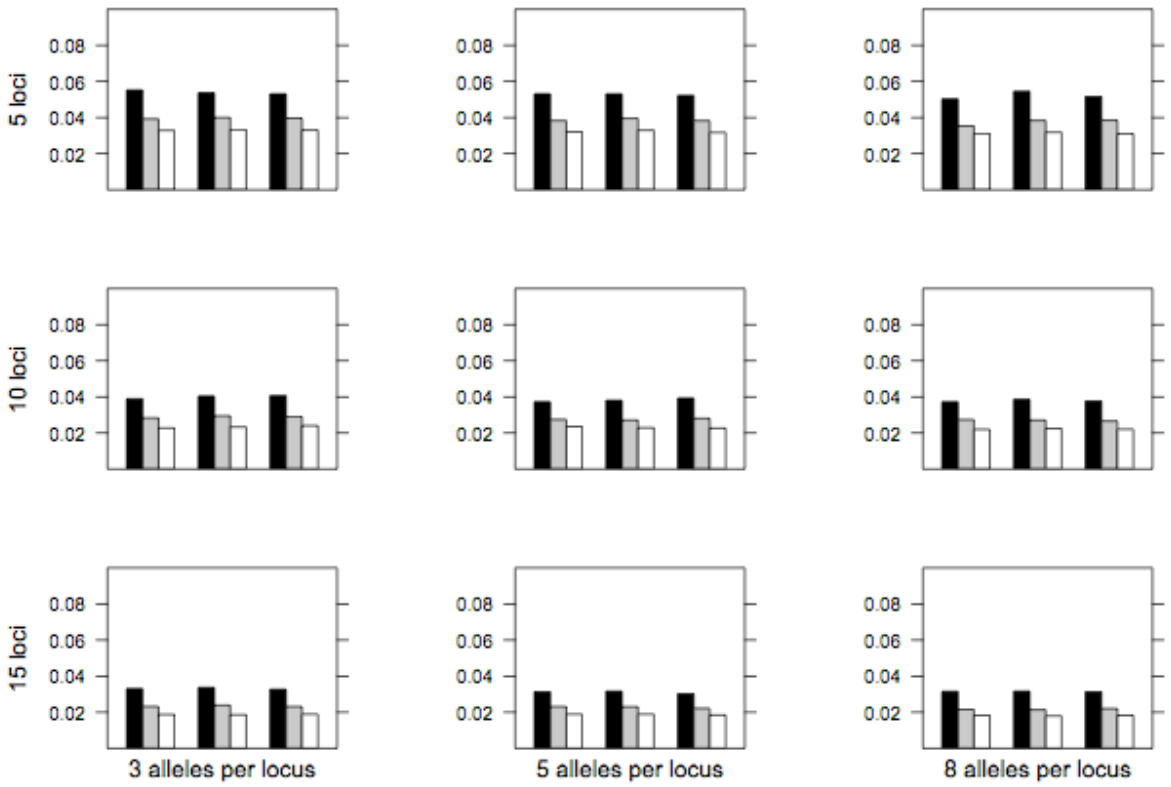


Figure 2.7. BCI length for false allele (e_1) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (black bars), 50 (gray bars), and 75 (white bars) with true population sizes of 100, 250, and 500.

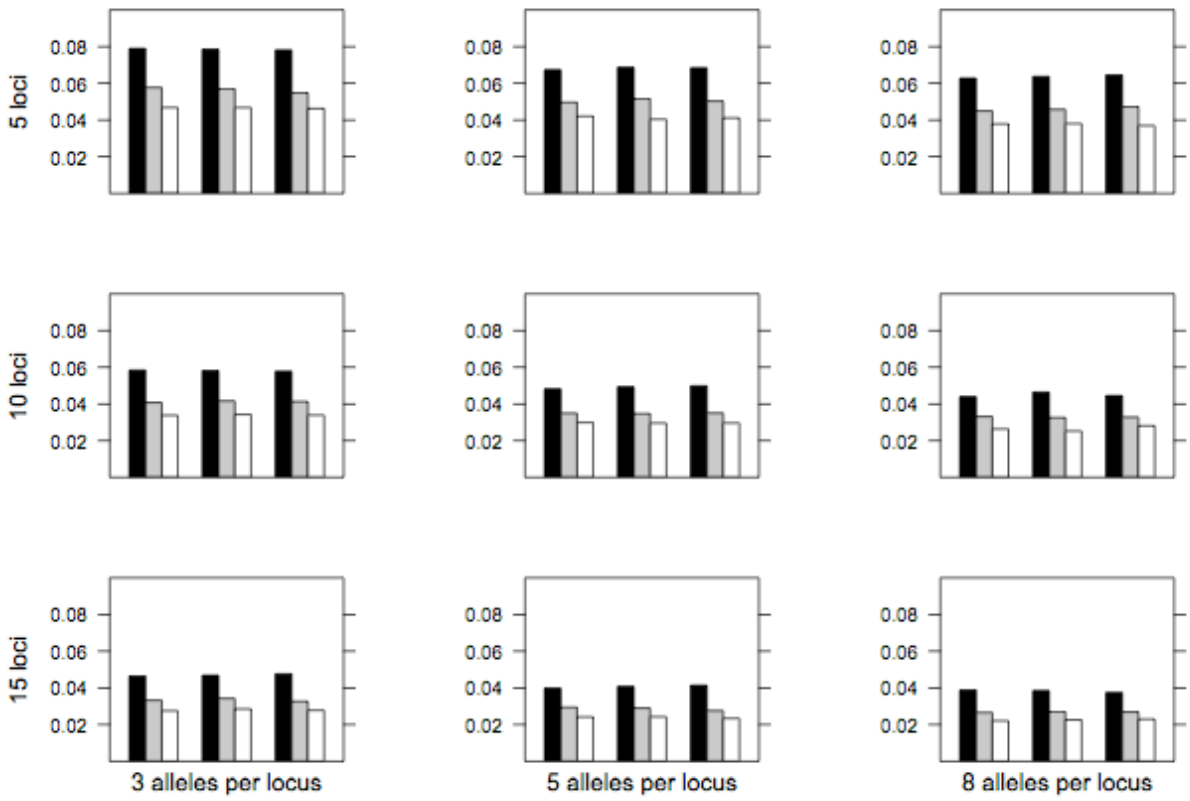


Figure 2.8. BCI length for allelic dropout (e_2) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (black bars), 50 (gray bars), and 75 (white bars) with true population sizes of 100, 250, and 500.

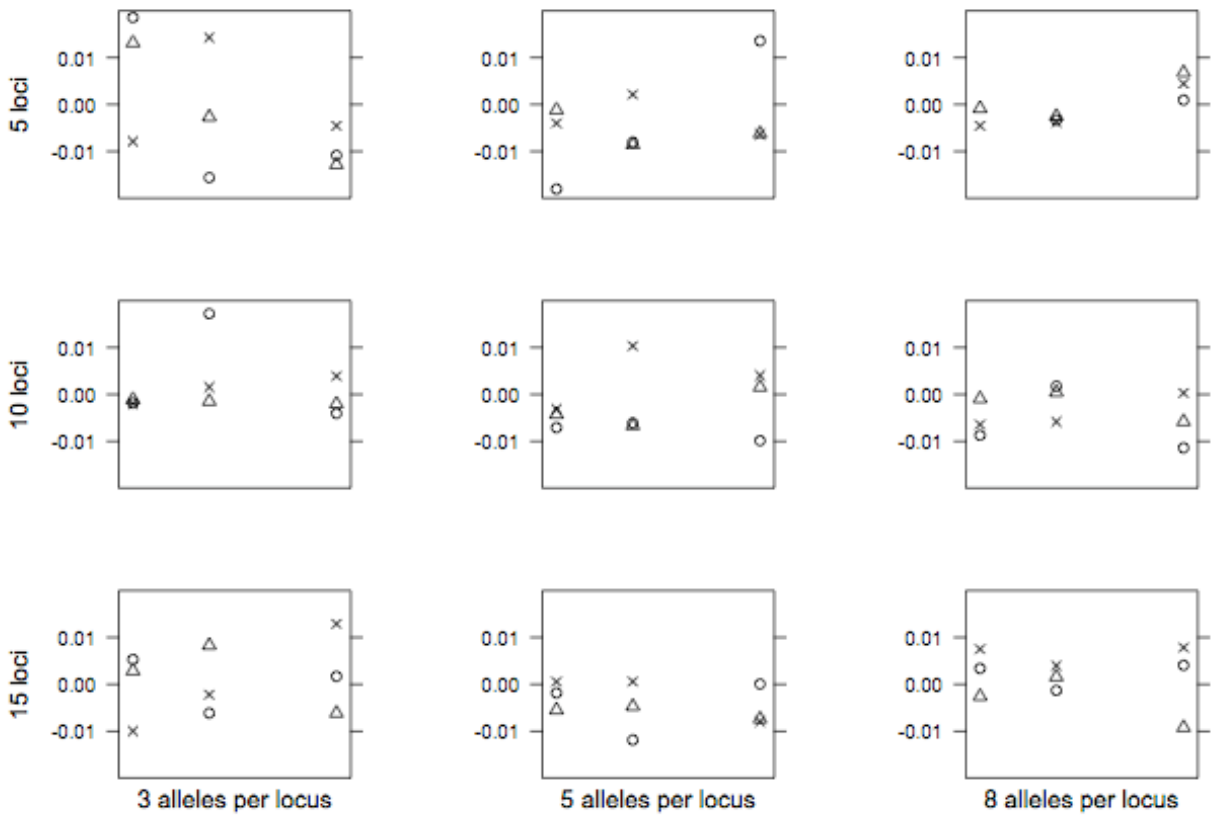


Figure 2.9. Relative bias for false allele (e_1) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (○), 50 (△), and 75 (×) with true population sizes of 100, 250, and 500.

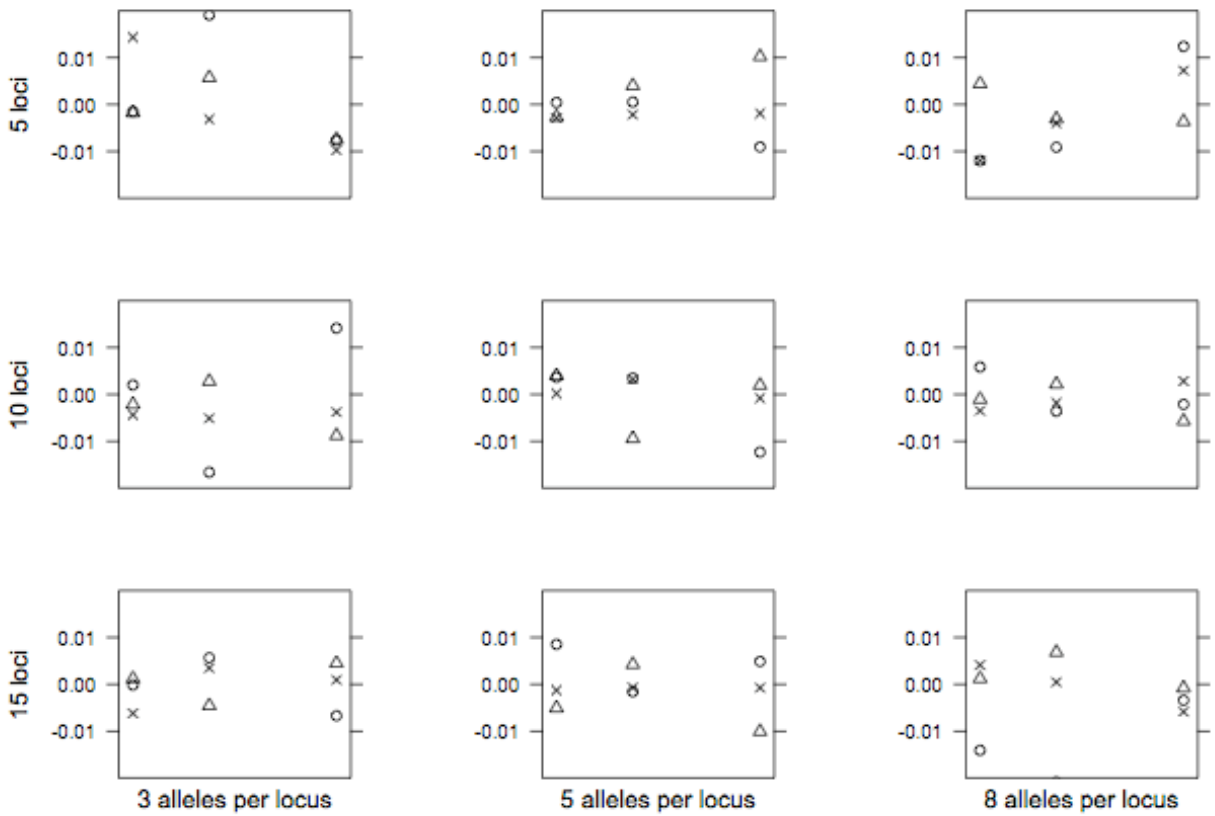


Figure 2.10. Relative bias for allelic dropout (e_2) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (○), 50 (△), and 75 (×) with true population sizes of 100, 250, and 500.

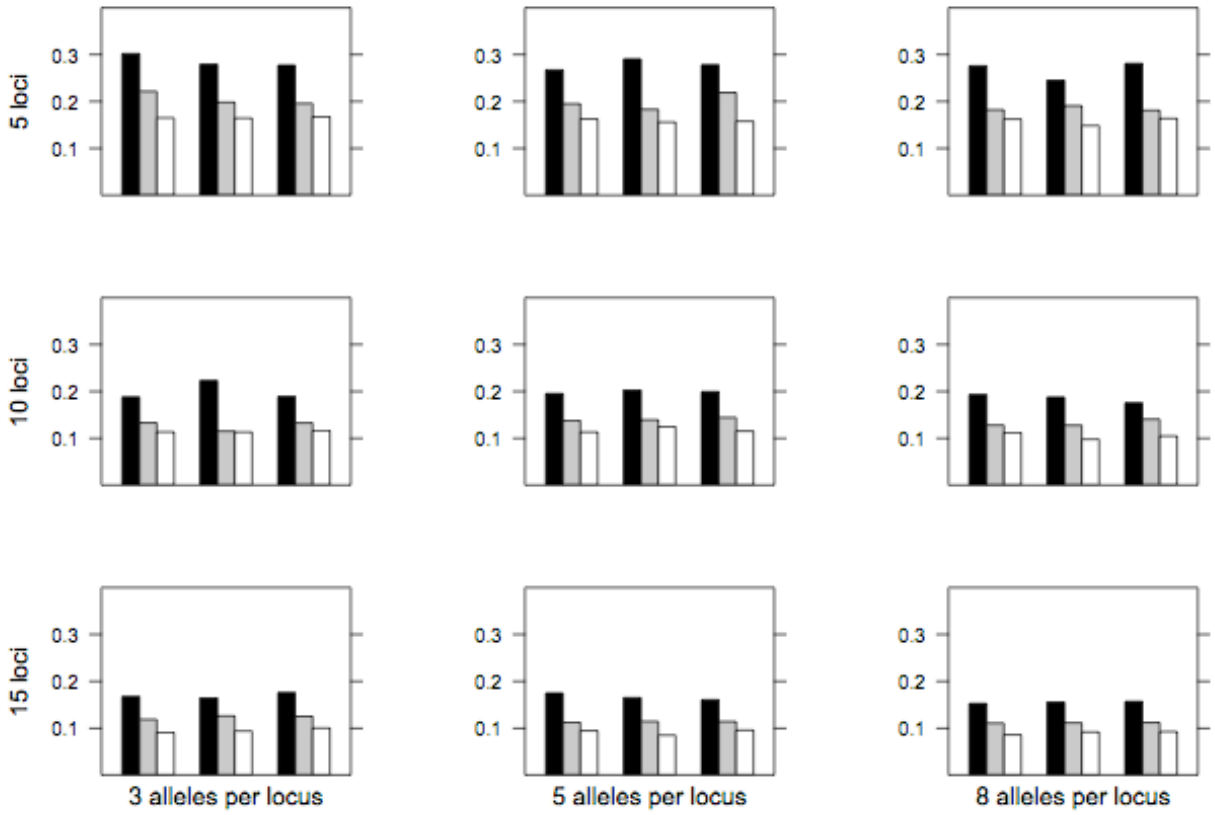


Figure 2.11. Relative RRMSE for false allele (e_1) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (black bars), 50 (gray bars), and 75 (white bars) with true population sizes of 100, 250, and 500.

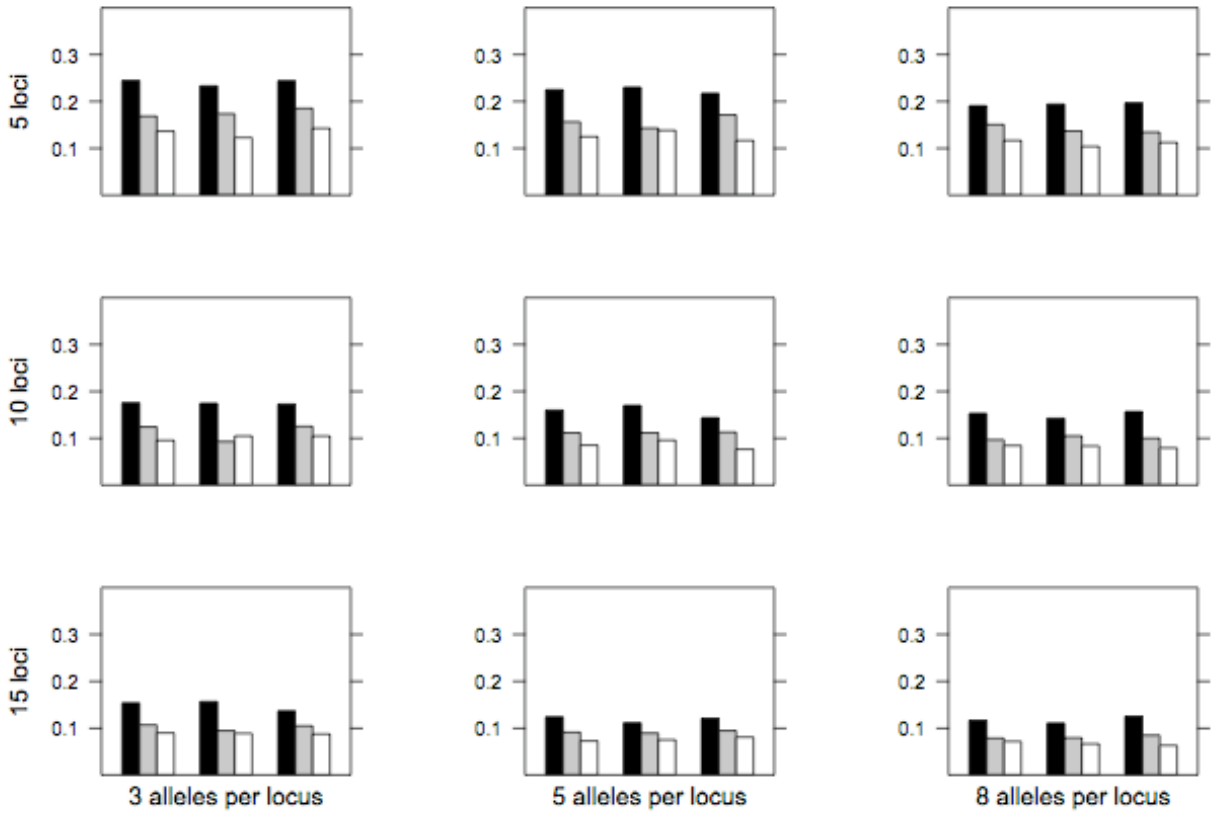


Figure 2.12. Relative RRMSE for allelic dropout (e_2) error for 81 simulation combinations of true population size, calibration sample size, alleles per locus, and number of loci. Each number of loci x alleles per locus panel includes calibration samples of 25 (black bars), 50 (gray bars), and 75 (white bars) with true population sizes of 100, 250, and 500.

CHAPTER 3

GENETIC MISIDENTIFICATION ERROR MODEL WITH BLACK BEAR POPULATION CASE STUDY[1]

[1] Sanderlin, J. S., N. Lazar, M. J. Conroy, and J. Reeves. To be submitted to *Biometrics*.

Introduction

The central Georgia population (CGP) of black bears (*Ursus americanus*) is considered to inhabit mostly forested land in and around 186 km^2 associated with the Ocmulgee River drainage system, and likely a core area of contiguous forest in the Oaky Woods and Ocmulgee Wildlife Management Areas (WMAs). The density of the CGP is estimated at 1 bear per 3.1 km^2, based on physical captures from a preliminary study (Grahl 1985). A current, more accurate estimate is needed to make informed management decisions. Various sampling techniques were applied to estimate density, including non-invasive genetic sampling. Barbed wire enclosures, or hair snares, designed to obtain hair samples from individual bears entering the devices (Woods et al. 1998) were placed on Oaky Woods and Ocmulgee Wildlife Management Areas (WMAs) and some private property from 2003 through 2006. Hair snares were sampled during summer and fall seasons. Hair snares were placed in a trapping web design (Anderson et al. 1983). Digital cameras were randomly placed at the same locations of hair snares and webs were monitored daily with radiotelemetry to detect marked bears in or near the webs. In addition, tissue and hair samples from known individuals from the CGP were collected to evaluate genetic error as a calibration sample with the genetic misidentification error model described in chapter 2. The main research objectives of the larger project were to estimate demographic parameters (e.g., survival and reproduction) and density to construct models for predicting population viability of the CGP of black bears.

Methods

Study area

The principal study areas for physical captures, hair snares, and camera traps are

Ocmulgee and Oaky Woods WMAs in Bleckley, Bibb, Houston, Pulaski, and Twigg Counties, located in central Georgia. The WMAs consist of a variety of habitat types (pine stands, bottomland hardwood, mixed forest, upland hardwood, black-belt prairie, clearcuts, thinned pine stands, and cypress-gum swamps).

Field Methods

Bears were captured and immobilized with a 2:1 mixture of ketamine hydrochloride (Ketaset) and xylazine hydrochloride (Rompun) at a dosage of 4.4 mg/kg of Ketaset and 2.2 mg/kg of Rompun, for estimated body weights by Georgia Department of Natural Resources personnel. Bears were captured in the study area (Figure 3.1) using Fremont foot trap snares (Freemont 1986) in each of 4 trapping seasons (Figures 3.2, 3.3, 3.4, 3.5) that extend from May through August each year (2003-2006). Culvert traps were used to trap nuisance bears and released on Oaky Woods WMA. An upper pre-molar tooth for age estimation by cementum annuli analysis (Willey 1974), blood samples, and hair follicles were collected from each captured bear. Sectioning, staining, and aging of teeth were conducted by Matson's Laboratories (Milltown, Montana). All bears were uniquely marked using a combination of collars, lip tattoos, and ear tags/streamers. Pertinent physiological data were recorded for each captured bear. Most bears were fitted with Advanced Telemetry Systems (Asanti, MN) radio transmitter collars (VHF, very high frequency) equipped with mortality signal sensors and 4 males were fitted with radio collars contained Global Positioning technology. All collars fitted to bears during the project were equipped with either a mechanical timer release (GPS) or a degradable release tab (VHF). The tissue and hair samples (n=85) were used in the calibration genetic analyses with known captured individuals (52M: 31F), plus samples collected by DNR

personnel of one road mortality (1 M) and one capture mortality (1 M).

Laboratory Methods

After field collection, hair samples were stored in silica desiccant then transferred to a -20°C freezer. Prior to extraction and after field collection, blood and tissue samples were also stored in a -20°C freezer. Extraction of DNA from Georgia tissue samples was done with the DNeasy Kit (QIAGEN) and with one captured bear blood sample using the GenomicPrep DNA isolation kit (GE Healthcare). DNA from hair samples were extracted with Chelex100 (10% solution) (Promega) and proteinase K (Phenix Research Products, QIAGEN) (modified from Boersen 2001). The root portion (1 cm.) from a maximum of 10 hairs per sample were cut and placed into 150 *µl* of Chelex 100 (10% solution) (Promega). The number of hairs and quality of sample were recorded. If the number of hairs was less than 10, the entire strand of hair was used in the sample. Low quality samples had little or no visible roots, and usually consisted of under-fur (thin) hairs. Medium quality samples were classified as samples with half of the hairs with roots visible and some guard hairs. High quality samples were classified by a majority of the hairs as guard hairs with most or all of the roots visible, and including visible skin cells. After roots were placed in the 10% solution Chelex 100 (Promega), 10 *µl* proteinase-K was added to assist with DNA digestion. The hair samples were incubated at 65 °C overnight (~8 hours). Samples were vortexed, and then boiled at 100°C for 10 minutes. After removal, the samples were centrifuged at 10,000-12,000 rpm for 3 minutes. The supernatant was pulled off and placed into a clean tube, and stored at -20°C until PCR analysis.

PCR amplifications were performed in 10 µL volumes using Bio-Rad MyCycler thermal cyclers for both tissue and hair samples with 16 tetranucletide loci (UA-BM3-P1F04, UA-BM4-P1H06, UA-BM4-P1H10, UA-BM4-P2B06, UA-BM4-P2C10, UA-BM3-P1A08, UA-BM4-

P2E11, UA-RM3-P2G10, UA-RM3-P2H03 , UA-BM3-P1D05, UA-BM4-P2A02, UA-BM4-P2A06, UA-BM4-P2B08, UA-BM3-P1B05, UA-RM3-P2H01, and BM4-P2C02, hereafter named Bear 10Y, Bear 12Y, Bear 13, Bear 17G, Bear 19Y, Bear 2, Bear 23, Bear 25, Bear 27B, Bear 30B, Bear 32, Bear 33B, Bear 35G, Bear 6, Bear 26, and Bear 36, respectively) previously described in Sanderlin et al. (2009). Loci Bear 10Y, Bear 12Y, Bear 17G, Bear19Y, Bear 27B, Bear 30B, Bear 33B, and Bear 35G were directly labeled primers with the dyes NED (Y), HEX (G), and FAM (B). Final concentrations for optimizing reactions with unlabelled primers were 10 mM Tris pH 8.4, 50 mM KCl, 0.5 μM "pigtailed" primer, 0.05 μM CAG or M13-reverse tagged primer (CAG or M13-reverse + primer), 0.45 μM dye labeled tag (HEX or FAM + CAG or M13-reverse), 1.5 mM $MgCl_2$, 0.5 mM dNTPs, 0.5 U Ampli*Taq* Gold DNA Polymerase (Applied Biosystems), and 50 ng DNA. Final concentrations for optimizing reactions with directly labeled primers were 10 mM Tris pH 8.4, 50 mM KCl, 0.5 μM upper directly labeled primer, 0.5 μM lower directly labeled primer, 1.5 mM $MgCl_2$, 0.5 mM dNTPs, 0.5 U Ampli*Taq* Gold DNA Polymerase (Applied Biosystems), and 50 ng DNA. We ran reactions using one touchdown thermal cycling program (Don et al. 1991), encompassing a 10.5 ^{o}C span of annealing temperatures (range: 60-49.5 ^{o}C).

For tissue samples, cycling parameters were: 21 cycles of 96 ^{o}C for 20 s; highest annealing temperature for 30 s minus 0.5 ^{o}C per annealing cycle; and 72 ^{o}C for 1 min 30 s followed by 14 cycles of 96 ^{o}C for 20 s; 50 ^{o}C, for 30 s; 72 ^{o}C for 1 min 30 s; and a final extension period of 10 min. at 72 ^{o}C. For hair samples, cycling parameters were: 20 cycles of 96 ^{o}C for 20 s; highest annealing temperature for 30 s minus 0.5 ^{o}C per annealing cycle; and 72 ^{o}C for 1 min 30 s followed by 30 cycles of 96 ^{o}C for 20 s; 50 ^{o}C, for 30 s; 72 ^{o}C for 1 min 30 s; and a final extension period of 10 min. at 72 ^{o}C. We checked PCR products for amplification and

sized fragments using a 3730*xl* DNA sequencer (Applied Biosystems) with GENESCAN Rox500 fluorescent size standard (PE Applied Biosystems). Results were analyzed using GENEMAPPER software (Applied Biosystems) using the local Southern size-calling method.

Statistical Methods

Tissue samples were classified as true genotypes for individuals in the estimation model and hair samples from known individuals were classified as observed genotypes in the previously described estimation model of chapter 2. Models were processed with the Metropolis-Hastings algorithm implemented in PyMC 2.0 (Patil, A., Huard, D. and Fonnesbeck. C, http://pymc.googlecode.com) in Python, version 2.5.2 (Python Software Foundation, http://python.org). For each statistical model, 100,000 iterations were used with 50,000 iterations used as a burn-in period. No thinning interval was used. Three chains were run with the top model based on model selection criteria (see below). CODA (Best et al. 1995) in program R (R Development Core Team 2008) was used to create figures and traces of the posterior distributions for allelic dropout (ADO) and false alleles (FA) for the most likely models in the candidate model sets.

Priors for parameters

Informative priors were selected, based on several previous studies (Appendix B, additional samples marked with *). The sample mean ($\bar{x}$) and sample variance (s^2) of both types of error were calculated from the previous studies. These values can be used to approximate a prior Beta distribution with the first two method-of-moments estimates (ADO: mean is 0.136, variance is 0.029, FA: mean is 0.029, variance is 0.002). The two parameters (α, β) for the Beta

distribution have solutions from the first 2 methods of moments of: 1) $\alpha = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$, and 2) $\beta = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$. The prior distribution of e_1, or the probability of obtaining a false allele, would approximately follow a Beta (0.49, 16.65) distribution, based on previous studies (n=46 estimates, some from the same studies) (Figure 3.6). The prior distribution of e_2, or the probability of allelic dropout would approximately follow a Beta (0.42, 2.66) distribution, based on previous studies (n=69 estimates, some from the same studies) (Figure 3.7). Figures were created with the base graphic package in program R (R Development Core Team 2008). Uninformative priors Beta (1,1) for both error probabilities were selected to identify if any convergence discrepancies existed with the most likely models from the candidate model sets.

Candidate models and model selection

Model selection criteria was used to assess a candidate model set and determine the most likely models, given data from the central Georgia Population. The log-likelihood, Aikaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and Deviance Information Criterion (DIC) were recorded. AIC (Akaike 1973) is defined as:

$$AIC = -2\log(\ell(\hat{\theta} \mid y)) + 2K,$$

where $\ell(\hat{\theta} \mid y)$ is the log-likelihood and K is the number of estimable parameters. BIC (Schwarz 1978) is defined as:

$$BIC = -2\log(\ell(\hat{\theta} \mid y)) + K * \log(n),$$

where $\ell(\hat{\theta} \mid y)$ is the likelihood , n is the sample size, and K is the number of estimable parameters.

DIC (Spiegelhalter et al. 2002) is defined as:

$$DIC = 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y),$$

where deviance is defined as -2 times the log-likelihood: $D(y,\theta) = -2\log p(y \mid \theta)$. The $D_{\hat{\theta}}(y)$ portion obtains the discrepancy of the point estimate with a summary of y the data and a point estimate for θ, like the mean of the posterior simulations. The estimated average discrepancy, $\hat{D}_{avg}(y)$, is the average discrepancy over the posterior distribution, or the range of possible parameter values. The Bayesian models were hierarchical, which indicates that DIC would be the best choice of the three model selection criteria. For comparison, the AIC and BIC model selection criteria were also calculated.

Four subsets of models were run, specifically: 1) 16 loci with informative priors including records that were classified as 'no data', 2) 16 loci with informative priors without including records that were classified as 'no data', 3) 9 loci with informative priors including records that were classified as 'no data', and 4) 9 loci with informative priors without including records that were classified as 'no data'. Within each subset of models, four models (constant ADO/FA among loci, constant ADO/ varied FA among loci, varied ADO among loci/ constant FA among loci, varied ADO and FA among loci) were run. In total, sixteen models were run with the error estimation model described in chapter 2.

Results

Data summary

Nine hair samples (6 M: 3 F) were classified as bad samples, since they had positive amplification with only a few loci and were removed from the error analysis. The total possible

individuals for the genetic error models were 76 (48 M: 28 F). There were 20 errors over 76 bears and 16 loci detected in the genetic analysis (Table 3.1). Three bears accounted for 50% of the errors detected (Bear 32: 2 errors, Bear CM1: 5 errors, Bear 37: 3 errors). Some tissue and hair samples had positive PCR amplification, but were censored from the analysis because there was too much product in the samples for positive allele sizing. The total possible bears will be less than 76 for some loci (Table 3.2).

Statistical analysis

Twenty candidate models were run in PyMC 2.0. The most likely model of the 16 loci with informative priors including records that were classified as 'no data' was variation across all loci for the probabilities of ADO and FA, with a DIC value of 298.61 (Table 3.3). Median values of the locus specific ADO probabilities have a range from 0.002 to 0.295 (Table 3.4, Figures 3.8. 3.9, 3.10, 3.11). Median values of the locus specific FA probabilities have a range from 0.001 to 0.033 (Table 3.5, Figures 3.12, 3.13, 3.14, 3.15). All three chains for all parameters mixed quickly for this model.

The most likely model of the 16 loci with informative priors without including records that were classified as 'no data' was a constant ADO probability and variation across all loci for the FA probability, with a DIC value of 94.01(Table 3.6). The median and 95% posterior density interval for the ADO probability is 0.011 (95% posterior density interval: 0.005-0.018) (Figure 3.16). Median values of the locus specific FA probabilities have a range from 0.001 to 0.032 (Table 3.7, Figures 3.17, 3.18, 3.19, 3.20). All three chains for all parameters mixed quickly for this model. For comparison, the medians and 95% posterior density intervals for the constant ADO and FA probability model of the 16 loci with informative priors without including records

that were classified as ‘no data’ are: 1) ADO probability 0.011 (95% posterior density interval: 0.005-0.019), and 2) FA probability is 0.005 (95% posterior density interval: 0.002-0.009) (Figure 3.21).

The most likely model of the 9 loci with informative priors including records that were classified as ‘no data’ was variation across all loci for the probabilities of ADO and FA, with a DIC value of 167.46 (Table 3.8). Median values of the locus specific ADO probabilities have a range from 0.002 to 0.211 (Table 3.9, Figures 3.22, 3.23). Median values of the locus specific FA probabilities have a range from 0.001 to 0.033 (Table 3.10, Figures 3.24, 3.25). All three chains mixed quickly for all parameters for this model.

The most likely model of the 9 loci with informative priors without including records that were classified as ‘no data’ was a constant ADO probability and variation across all loci for the FA probability, with a DIC value of 73.20 (Table 3.11). The median and 95% posterior density interval for the ADO probability is 0.010 (95% posterior density interval: 0.004-0.020) (Figure 3.26). Median values of the locus specific FA probabilities have a range from 0.001 to 0.032 (Table 3.12, Figures 3.27, 3.28). All three chains mixed quickly for all parameters for this model. For comparison, the medians and 95% posterior density intervals for the constant ADO and FA probability model of the 9 loci with informative priors without including records that were classified as ‘no data’ are: 1) ADO probability 0.010 (95% posterior density interval: 0.003-0.020), and 2) FA probability is 0.009 (95% posterior density interval: 0.004-0.016) (Figure 3.29).

Discussion and conclusions

When the records classified as ‘no data’ were included for both the 16 loci and 9 loci models, the fully parameterized model with locus specific ADO and FA probabilities was considered the most likely model in the candidate model set. ADO probabilities had a wider range, and were on average, higher than FA probabilities for most loci. This is consistent with other studies that estimate genetic error (Appendix B). Both types of genetic error had low probabilities of error compared to other studies (Appendix B). Typically, population estimates are not as biased with rates of error less than 1%.

When the records classified as ‘no data’ were not included for both the 16 loci and 9 loci models, the model with locus specific FA probabilities and a constant ADO probability was considered the most likely model in the candidate model set. The sample size with these candidate models was smaller. This may lead to selection of less explanatory models. Error probabilities for both types of error were similar to the models that included the ‘no data’ records, with ADO probability typically larger than FA probabilities. Approximately 50% of the loci without the ‘no data’ records had estimates for FA close to zero for both 16 and 9 loci model sets.

Models without ‘no data’ records were analyzed because it is impossible to separate the true allelic dropout events in the laboratory from general PCR artifacts that might cause a negative amplification of the whole sample or another laboratory anomaly. Therefore, the true allelic dropout probabilities are probably somewhere between the parameter estimates including ‘no data’ records and the parameter estimates not including ‘no data’ records. A future extension of the model parameterization could include a latent parameter that would capture the negative

amplification from laboratory anomalies from the true double allelic dropout events.

There was some evidence of bears having unequal probabilities of genetic error. Bears, such as CM1 and 37, had many genotype errors, compared to several bears that did not have any. Again, three individual bears contributed to 50% of the detected errors in the data set. The hair samples from these bears may be of lower quality than the other samples.

A potential model for future analyses would include grouping loci into categories according to the base pair length of the PCR products. Some studies indicate loci with longer base pair lengths may have a higher probability of error. This is one reason why shorter lengths of genetic markers for non-invasive studies are selected. A linear model with base pair size as a predictor may even be a better model than grouping loci into categories. Special consideration in selection of a marker panel is especially important for conservation and management studies based on genetic data.

Literature Cited

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267-281 *in* B. N. Petrov, and F. Csaki, (eds.) *Second International Symposium on Information Theory*. Akademiai Kiado, Budapest.

Anderson, D.R., K.P. Burnham, G.C. White, and D.L. Otis. 1983. Density estimation of small mammal populations using a trapping web and distance sampling methods. Ecology 64:674-680.

Best, N. G., M. K. Cowles, and S.K. Vines. 1995. CODA Manual version 0.30. MRC Biostatistics Unit, Cambridge, UK.

Boersen, M. R. 2001. Abundance and density of Louisiana black bears on the Tensas River National Wildlife Refuge. M. S. thesis, University of Tennessee, Knoxville.

Don. R.H., P. T. Cox, B. J. Wainwright, K. Baker, and J. S. Mattick. 1991. 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Research 19: 4008.

Fremont, A. D. and G. J. Steil. 1986. Foot snare live trap. Patent No. 4581843. USA.

Grahl, D.K., Jr. 1985. Preliminary investigation of Ocmulgee River drainage black bear population. Georgia Department of Natural Resources. Final Report, Fed Aid Proj. W 37-R, Study B-2, Atlanta, Georgia, USA. 15 pp.

R Development Core Team (2008). R: A language and environment for statistical computing, reference index version 2.7.2. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sanderlin, J. S., B.C. Faircloth, B. Shamblin, and M. J. Conroy. 2009. Tetranucleotide microsatellite loci from the black bear (*Ursus americanus*). Molecular Ecology Resources 9: 288-291.

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6: 461-464.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64: 1-34.

Willey, C. H. 1974. Aging black bears from first premolar tooth sections. Journal of Wildlife Management 38: 97-100.

Woods, J.G., D. Paetkau, D. Lewis, B.N. McLellan, M. Proctor, and C. Strobeck. 1999. Genetic tagging of free-ranging black and brown bears. Wildlife Society Bulletin 27: 616-627.

Table 3.1. Errors detected in the genetic analysis. The locus, bear name, and the actual base pair lengths of both alleles for the tissue samples and hair samples of each error detected are listed below.

Locus	Bear	Tissue allele 1	Tissue allele 2	Hair allele 1	Hair allele 2
10Y	RK1	294	298	294	294
12Y	32	256	256	248	256
12Y	33	260	260	252	260
17G	14	185	193	197	197
17G	CM1	193	193	185	185
23	CM1	256	271	271	271
25	32	382	393	382	382
26	31	356	361	356	356
27B	22	183	183	187	187
27B	30	187	187	183	187
27B	35	187	187	183	187
27B	82	191	191	187	191
30B	CM1	443	447	443	443
32	37	196	204	204	204
32	CM1	196	204	204	204
33B	37	272	286	272	272
33B	CM1	272	286	272	272
35G	5	224	224	216	224
35G	8	216	224	224	224
36	37	198	207	207	207

Table 3.2. Summary of data for genetic error analysis. True tissue samples are either homozygous or heterozygous. Observed hair samples are either true, with no alleles true, one allele true, or no data because the hair samples did not amplify.

TRUE		tissue samples			homozygous					heterozygous					
OBSERVED		hair samples			homozygous		heterozygous			heterozygous			homozygous		
Locus	Alleles	Total bears	Total hom	Total het	TRUE	no alleles true	one allele true	no alleles true	no data	TRUE	one allele true	no alleles true	one allele true	no alleles true	no data
bear10Y	7	70	35	35	35	0	0	0	0	34	0	0	1	0	0
bear12Y	4	68	31	37	29	0	2	0	0	36	0	0	0	0	1
bear13	3	70	62	8	60	0	0	0	2	8	0	0	0	0	0
bear17G	4	70	18	52	17	1	0	0	0	50	0	0	0	1	1
bear19Y	4	73	31	42	31	0	0	0	0	42	0	0	0	0	0
bear2	3	73	33	40	31	0	0	0	2	38	0	0	0	0	2
bear23	6	72	23	49	22	0	0	0	1	46	0	0	1	0	2
bear25	6	74	49	25	46	0	0	0	3	24	0	0	1	0	0
bear27B	4	70	56	14	48	1	3	0	4	12	0	0	0	0	2
bear30B	5	69	21	48	20	0	0	0	1	45	0	0	1	0	2
bear32	3	71	40	31	40	0	0	0	0	29	0	0	2	0	0
bear33B	5	74	27	47	26	0	0	0	1	44	0	0	2	0	1
bear35G	3	72	30	42	29	0	1	0	0	41	0	0	1	0	0
bear6	5	74	44	30	42	0	0	0	2	30	0	0	0	0	0
bear26	2	76	60	16	54	0	0	NA	6	11	1	NA	NA	NA	4
bear36	2	76	36	40	35	0	0	NA	1	36	1	NA	NA	NA	3

Table 3.3. Candidate model set of the 16 loci with informative priors including records that were classified as 'no data'. The log likelihood, BIC, AIC, and DIC values are listed. The low DIC model is considered the most likely model in this candidate model set.

Model	Log likelihood	BIC	AIC	DIC
ADO and FA varied across loci	-49.64	263.25	163.28	298.61
ADO varied across loci, FA constant	-84.14	255.39	202.28	311.74
ADO constant, FA varied across loci	-100.75	288.61	235.51	361.97
ADO and FA constant	-180.71	371.66	365.41	374.31

Table 3.4. Median values and 95% posterior density intervals of the locus specific ADO probabilities, under the model with 16 loci with informative priors including records that were classified as ‘no data’.

Locus (name)	Median	95% posterior density interval	
1 (Bear 10Y)	0.013	0.000	0.058
2 (Bear 12Y)	0.026	0.005	0.073
3 (Bear 13)	0.115	0.042	0.218
4 (Bear 17G)	0.036	0.011	0.082
5 (Bear 19Y)	0.002	0.000	0.026
6 (Bear 2)	0.088	0.043	0.154
7 (Bear 23)	0.067	0.030	0.124
8 (Bear 25)	0.105	0.048	0.183
9 (Bear 27B)	0.211	0.128	0.309
10 (Bear 30B)	0.069	0.031	0.126
11 (Bear 32)	0.029	0.002	0.087
12 (Bear 33B)	0.060	0.024	0.116
13 (Bear 35G)	0.006	0.000	0.041
14 (Bear 6)	0.057	0.020	0.123
15 (Bear 26)	0.295	0.206	0.393
16 (Bear 36)	0.100	0.050	0.170

Table 3.5. Median values and 95% posterior density intervals of the locus specific FA probabilities, under the model with 16 loci with informative priors including records that were classified as 'no data'.

Locus (name)	Median	95% posterior density interval	
1 (Bear 10Y)	0.002	0.000	0.017
2 (Bear 12Y)	0.014	0.003	0.042
3 (Bear 13)	0.002	0.000	0.017
4 (Bear 17G)	0.015	0.003	0.043
5 (Bear 19Y)	0.001	0.000	0.015
6 (Bear 2)	0.002	0.000	0.018
7 (Bear 23)	0.002	0.000	0.017
8 (Bear 25)	0.002	0.000	0.017
9 (Bear 27B)	0.033	0.011	0.075
10 (Bear 30B)	0.002	0.000	0.018
11 (Bear 32)	0.002	0.000	0.021
12 (Bear 33B)	0.002	0.000	0.018
13 (Bear 35G)	0.010	0.001	0.035
14 (Bear 6)	0.001	0.000	0.016
15 (Bear 26)	0.002	0.000	0.021
16 (Bear 36)	0.002	0.000	0.018

Table 3.6. Candidate model set of the 16 loci with informative priors without including records that were classified as 'no data'. The log likelihood, BIC, AIC, and DIC values are listed. The low DIC model is considered the most likely model in this candidate model set.

Model	Log likelihood	BIC	AIC	DIC
ADO constant, FA varied across loci	33.58	19.94	-33.17	94.01
ADO and FA varied across loci	57.39	49.18	-50.78	103.01
ADO and FA constant	-44.88	100.01	93.77	104.96
ADO varied across loci, FA constant	26.92	33.26	-19.84	115.53

Table 3.7. Median values and 95% posterior density intervals of the locus specific FA probabilities, under the model with 16 loci with informative priors without including records that were classified as ‘no data’.

Locus (name)	Median	95% posterior density interval	
1 (Bear 10Y)	0.002	0.000	0.017
2 (Bear 12Y)	0.015	0.003	0.042
3 (Bear 13)	0.001	0.000	0.016
4 (Bear 17G)	0.017	0.003	0.047
5 (Bear 19Y)	0.001	0.000	0.014
6 (Bear 2)	0.002	0.000	0.016
7 (Bear 23)	0.002	0.000	0.018
8 (Bear 25)	0.002	0.000	0.018
9 (Bear 27B)	0.032	0.011	0.068
10 (Bear 30B)	0.002	0.000	0.019
11 (Bear 32)	0.003	0.000	0.024
12 (Bear 33B)	0.002	0.000	0.020
13 (Bear 35G)	0.009	0.001	0.033
14 (Bear 6)	0.001	0.000	0.017
15 (Bear 26)	0.003	0.000	0.024
16 (Bear 36)	0.003	0.000	0.021

Table 3.8. Candidate model set of the 9 loci with informative priors including records that were classified as 'no data'. The log likelihood, BIC, AIC, and DIC values are listed. The low DIC model is considered the most likely model in this candidate model set.

Model	Log likelihood	BIC	AIC	DIC
ADO and FA varied across loci	-6.76	95.50	49.53	167.46
ADO varied across loci, FA constant	-20.06	85.66	60.12	175.78
ADO constant, FA varied across loci	-30.91	107.35	81.81	198.13
ADO and FA constant	-96.69	202.49	197.38	206.07

Table 3.9. Median values and 95% posterior density intervals of the locus specific ADO probabilities, under the model with 9 loci with informative priors including records that were classified as ‘no data’.

Locus (name)	Median	95% Posterior density interval	
1 (Bear 10Y)	0.014	0.000	0.058
2 (Bear 12Y)	0.025	0.005	0.074
3 (Bear 17G)	0.036	0.011	0.082
4 (Bear 19Y)	0.002	0.000	0.026
5 (Bear 27B)	0.211	0.128	0.308
6 (Bear30B)	0.069	0.031	0.127
7 (Bear 33B)	0.059	0.024	0.114
8 (Bear 35G)	0.009	0.000	0.046
9 (Bear 36)	0.099	0.050	0.168

Table 3.10. Median values and 95% posterior density intervals of the locus specific FA probabilities, under the model with 9 loci with informative priors including records that were classified as ‘no data’.

Locus (name)	Median	95% Posterior density interval	
1 (Bear 10Y)	0.002	0.000	0.019
2 (Bear 12Y)	0.014	0.003	0.042
3 (Bear 17G)	0.015	0.003	0.045
4 (Bear 19Y)	0.001	0.000	0.015
5 (Bear 27B)	0.033	0.011	0.075
6 (Bear30B)	0.001	0.000	0.017
7 (Bear 33B)	0.002	0.000	0.017
8 (Bear 35G)	0.009	0.001	0.034
9 (Bear 36)	0.002	0.000	0.017

Table 3.11. Candidate model set of the 9 loci with informative priors without including records that were classified as 'no data'. The log likelihood, BIC, AIC, and DIC values are listed. The low DIC model is considered the most likely model in this candidate model set.

Model	Log likelihood	BIC	AIC	DIC
ADO constant, FA varied across loci	32.90	-20.85	-45.79	73.20
ADO and FA varied across loci	48.90	-15.82	-61.79	79.51
ADO and FA constant	-32.35	73.81	68.70	79.55
ADO varied across loci, FA constant	29.84	-14.15	-39.69	85.32

Table 3.12. Median values and 95% posterior density intervals of the locus specific FA probabilities, under the model with 9 loci with informative priors without including records that were classified as ‘no data’.

Locus (name)	Median	95% posterior density interval	
1 (Bear 10Y)	0.002	0.000	0.018
2 (Bear 12Y)	0.014	0.003	0.042
3 (Bear 17G)	0.017	0.003	0.048
4 (Bear 19Y)	0.001	0.000	0.015
5 (Bear 27B)	0.032	0.010	0.071
6 (Bear30B)	0.002	0.000	0.019
7 (Bear 33B)	0.003	0.000	0.026
8 (Bear 35G)	0.008	0.001	0.033
9 (Bear 36)	0.002	0.000	0.022

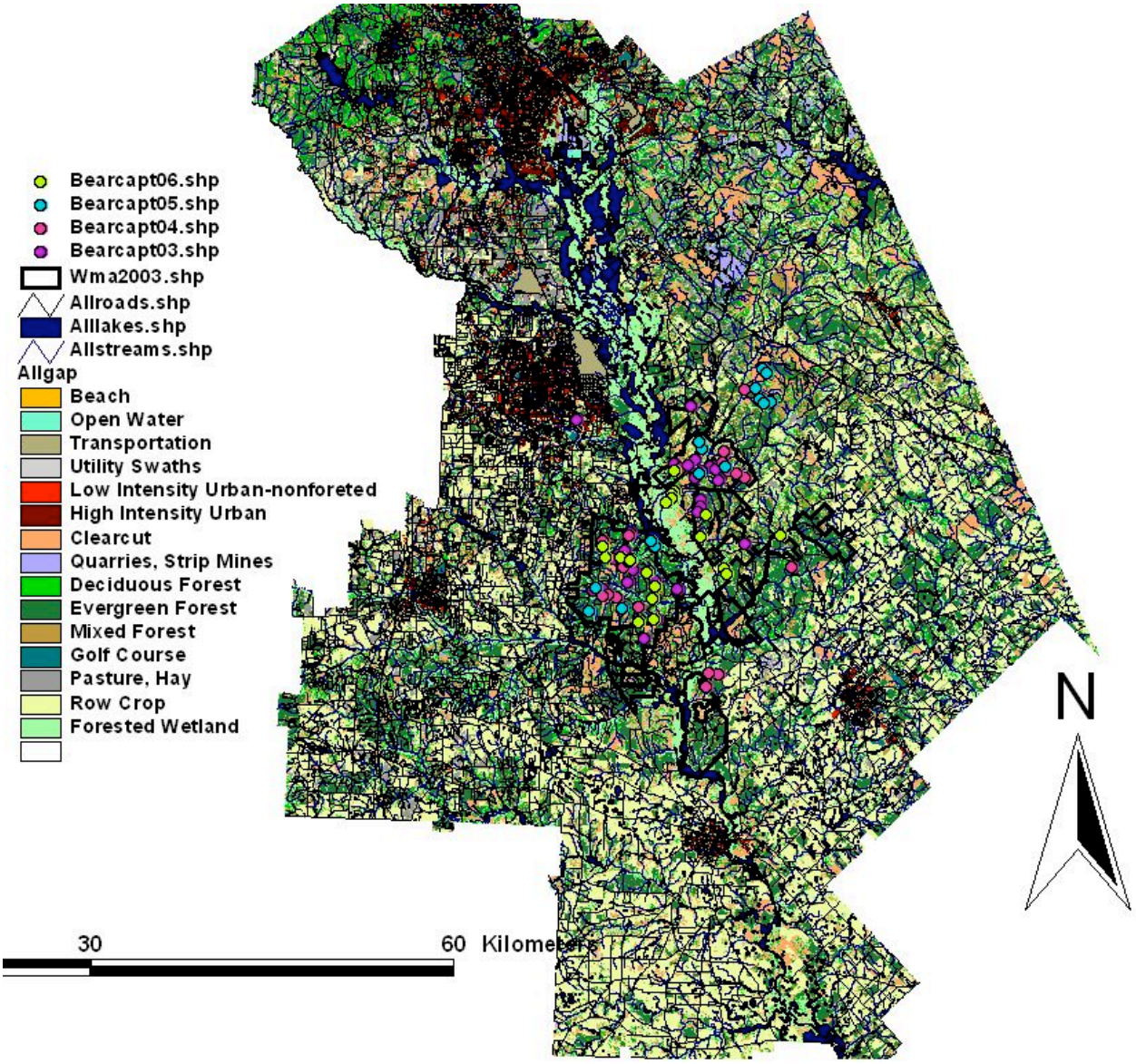


Figure 3.1. Capture coordinates for years 2003-2006 of initial and recaptured bears and WMA boundaries

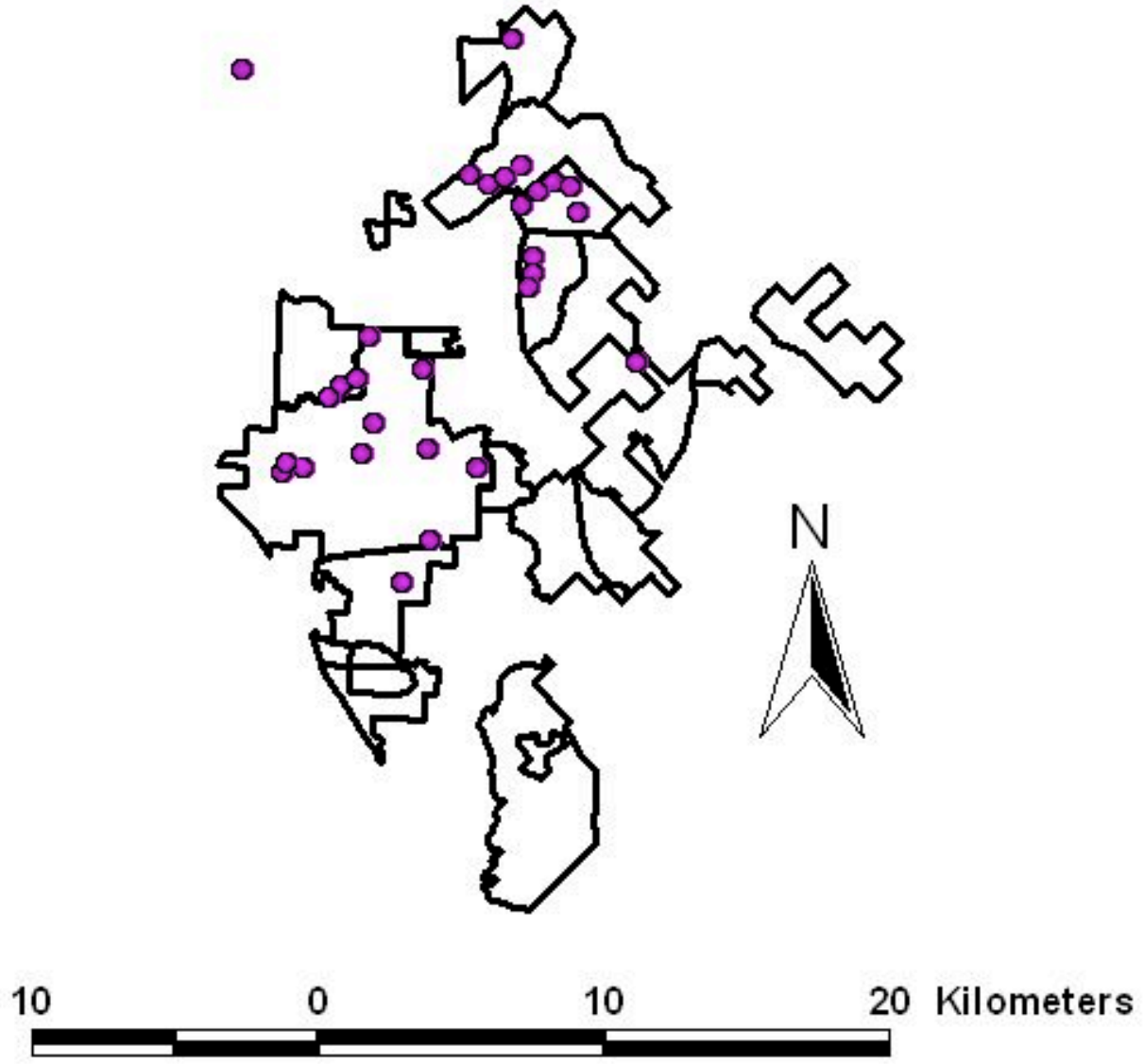


Figure 3.2. Initial and recapture coordinates for bears from 2003 and WMA boundaries

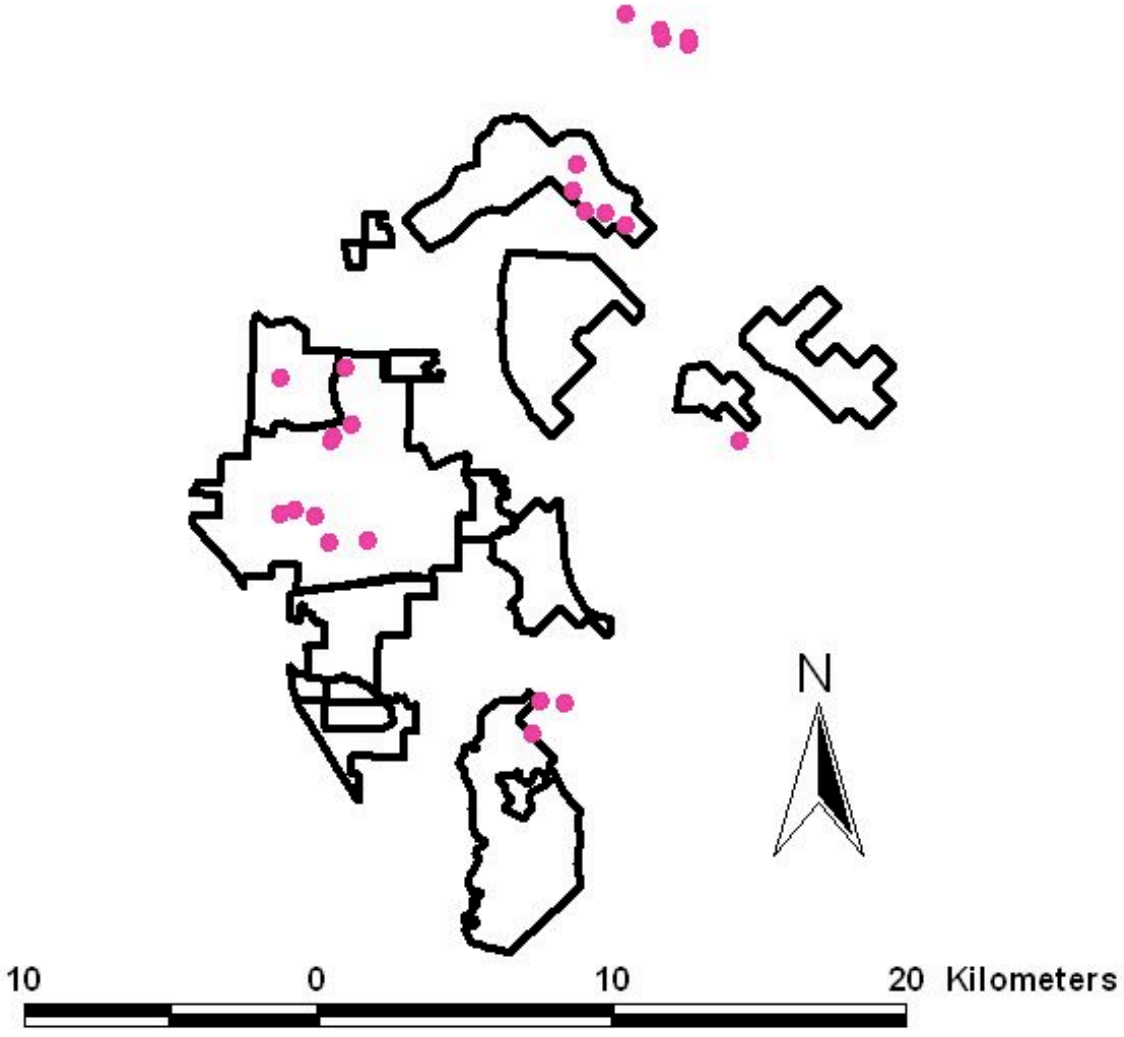


Figure 3.3. Initial and recapture coordinates for bears in 2004 and WMA boundaries

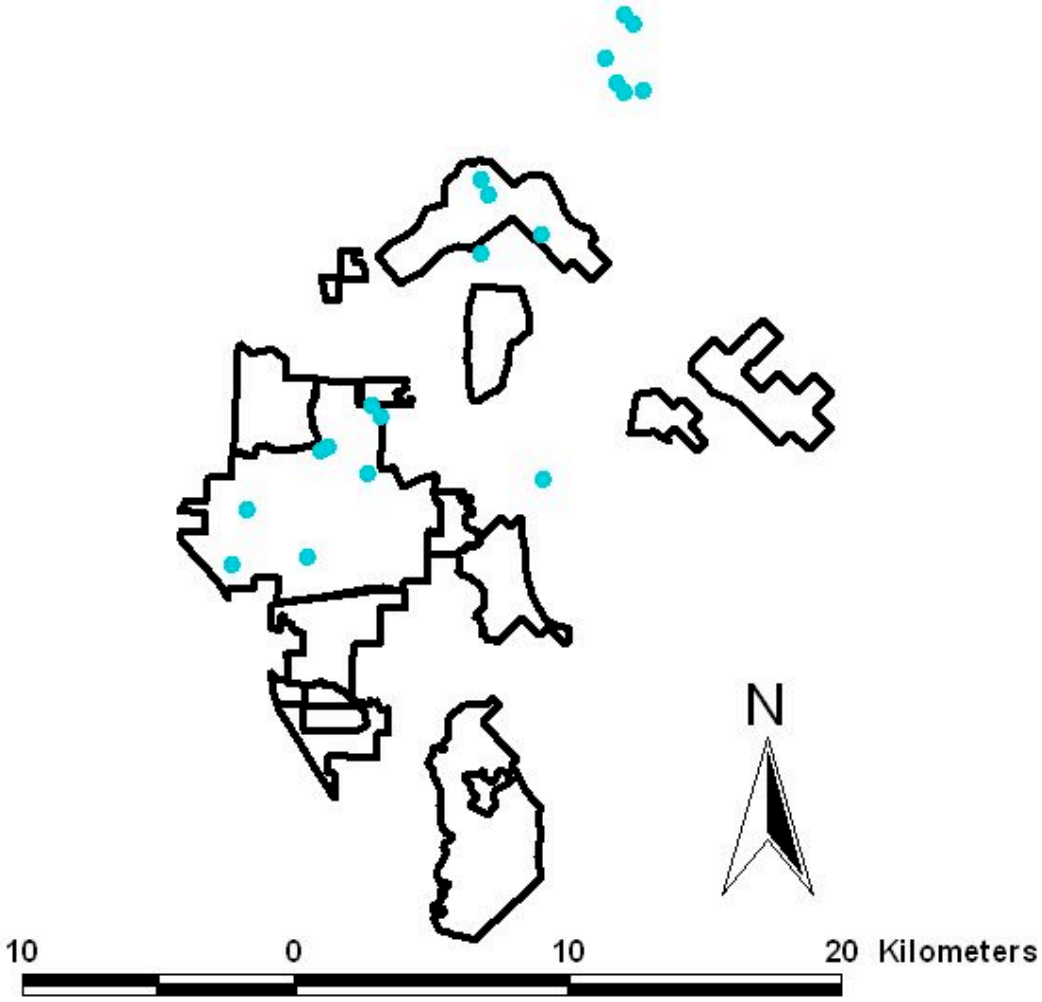


Figure 3.4. Initial and recapture coordinates for bears in 2005 and WMA boundaries

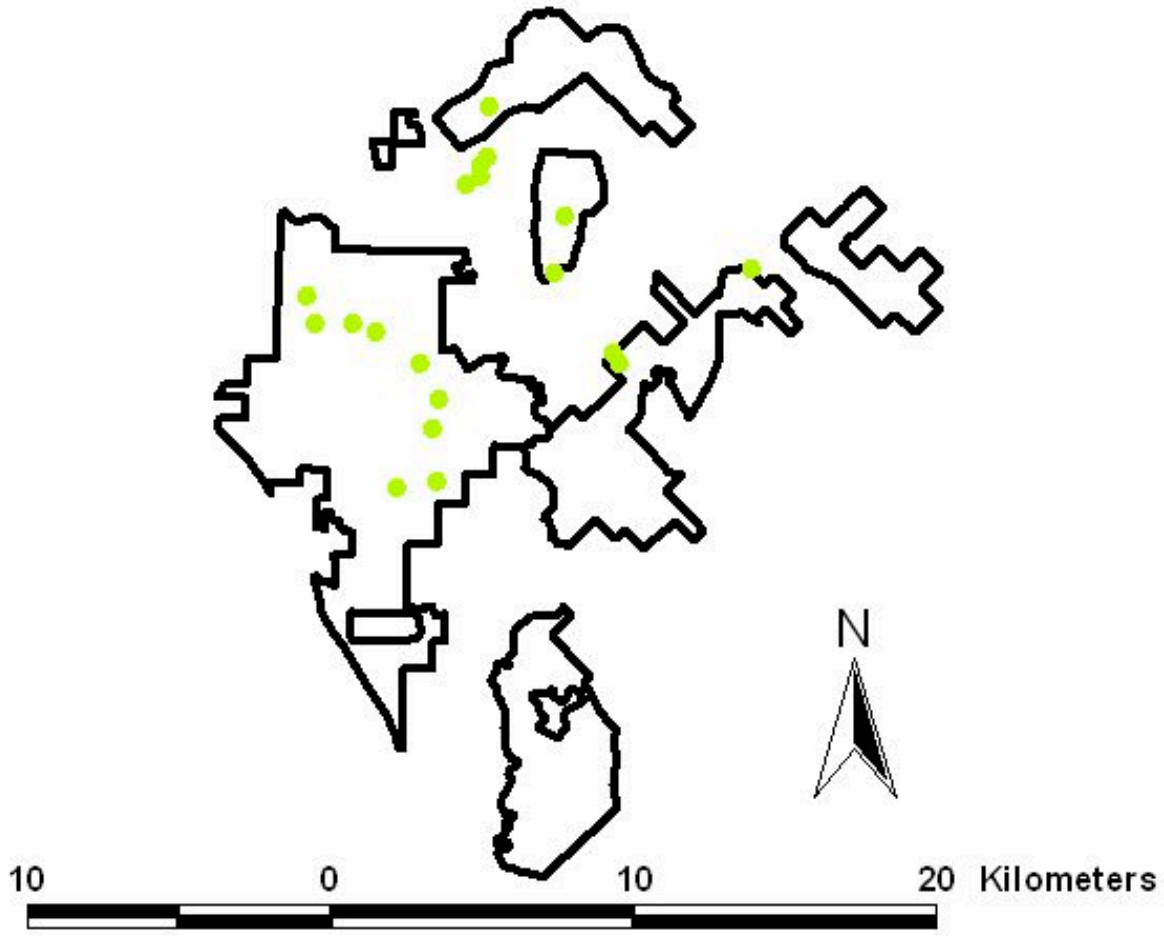


Figure 3.5. Initial and recapture coordinates for bears in 2006 and WMA boundaries

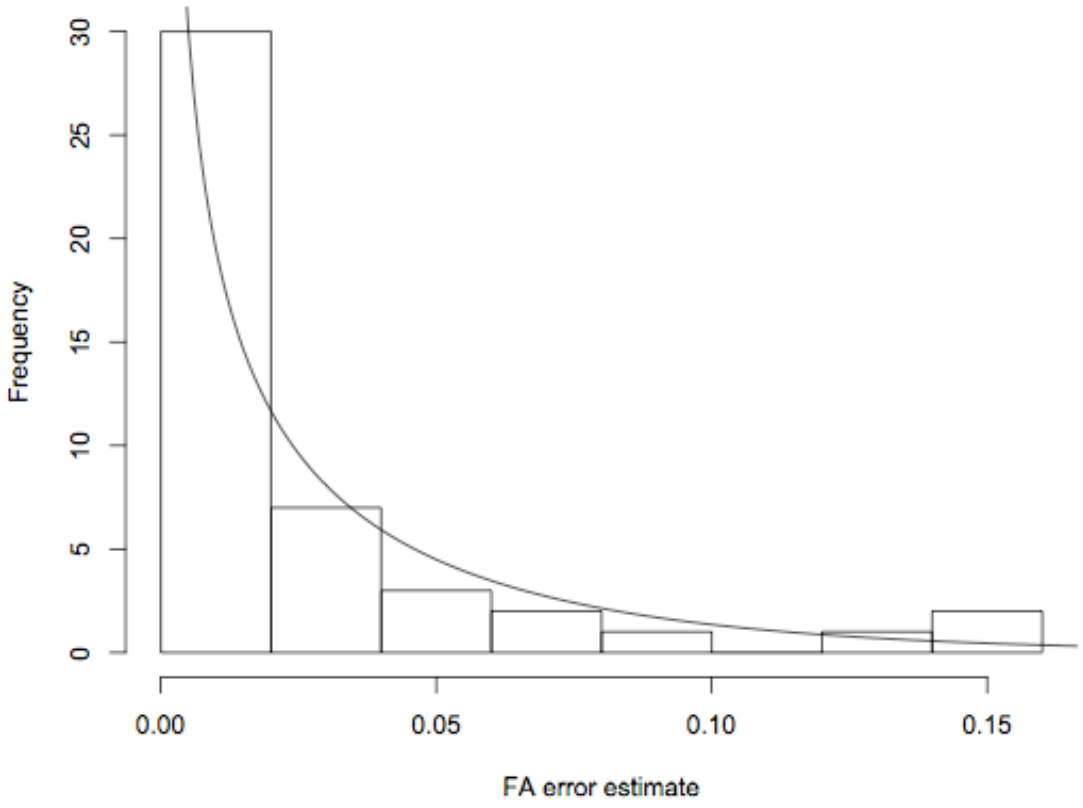


Figure 3.6. Histogram of the prior distribution of e_1, probability of obtaining a false allele, which approximately follows a Beta (0.49, 16.65) distribution (n=46 estimates, some from the same previous studies).

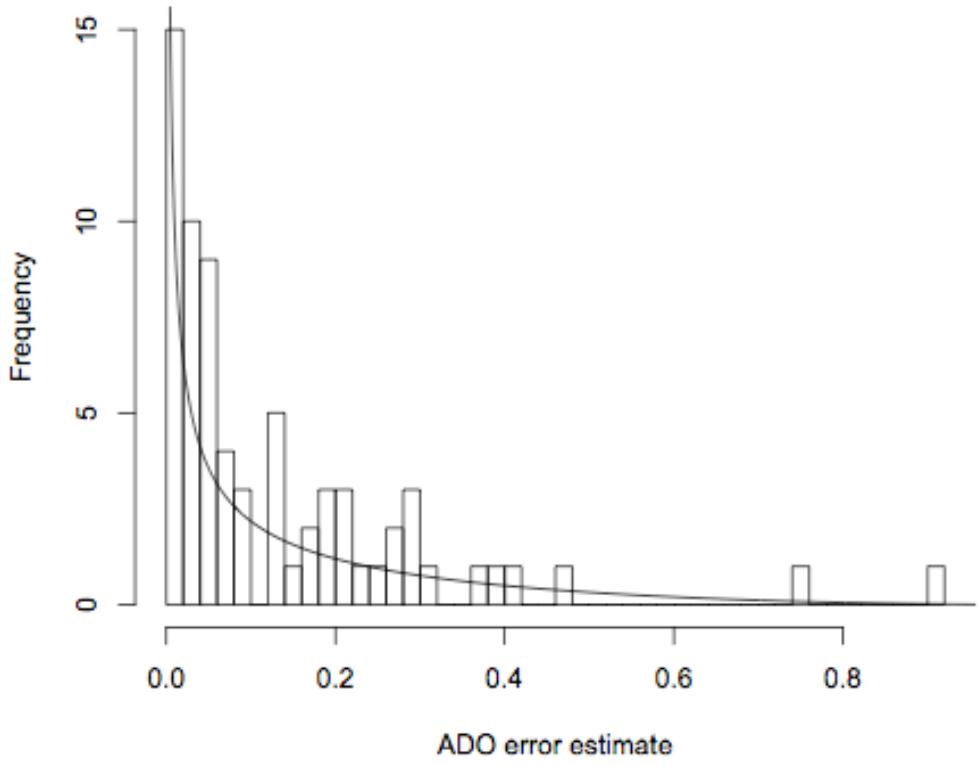


Figure 3.7. Histogram of the prior distribution of e_2, probability of allelic dropout, which approximately follows a Beta (0.42, 2.66) distribution (n=69 estimates, some from the same previous studies).

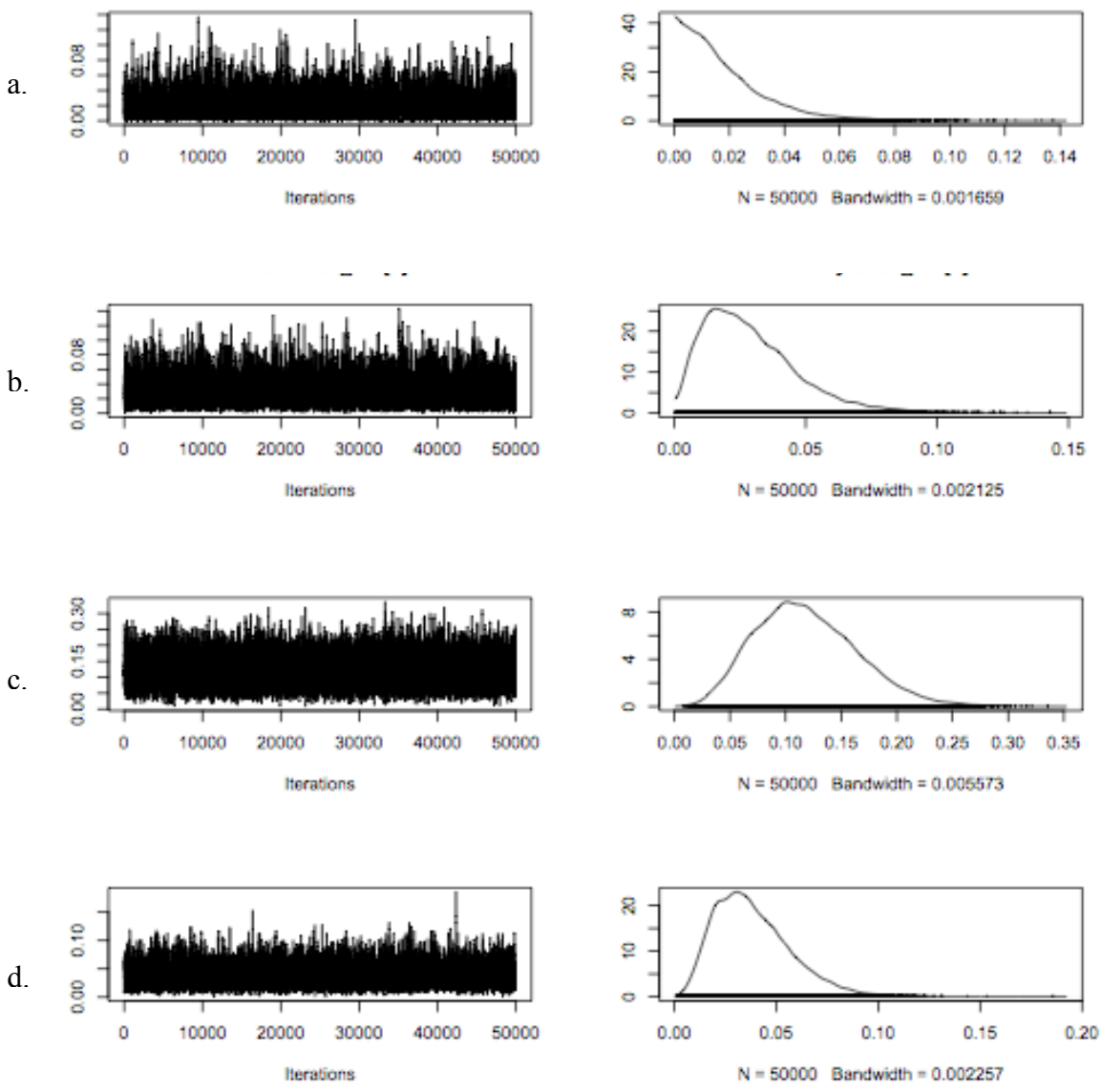


Figure 3.8. Posterior densities and traces of locus specific ADO probabilities for locus 1 (a), locus 2 (b), locus 3 (c), and locus 4 (d) under the model of 16 loci with informative priors including records that were classified as 'no data'.

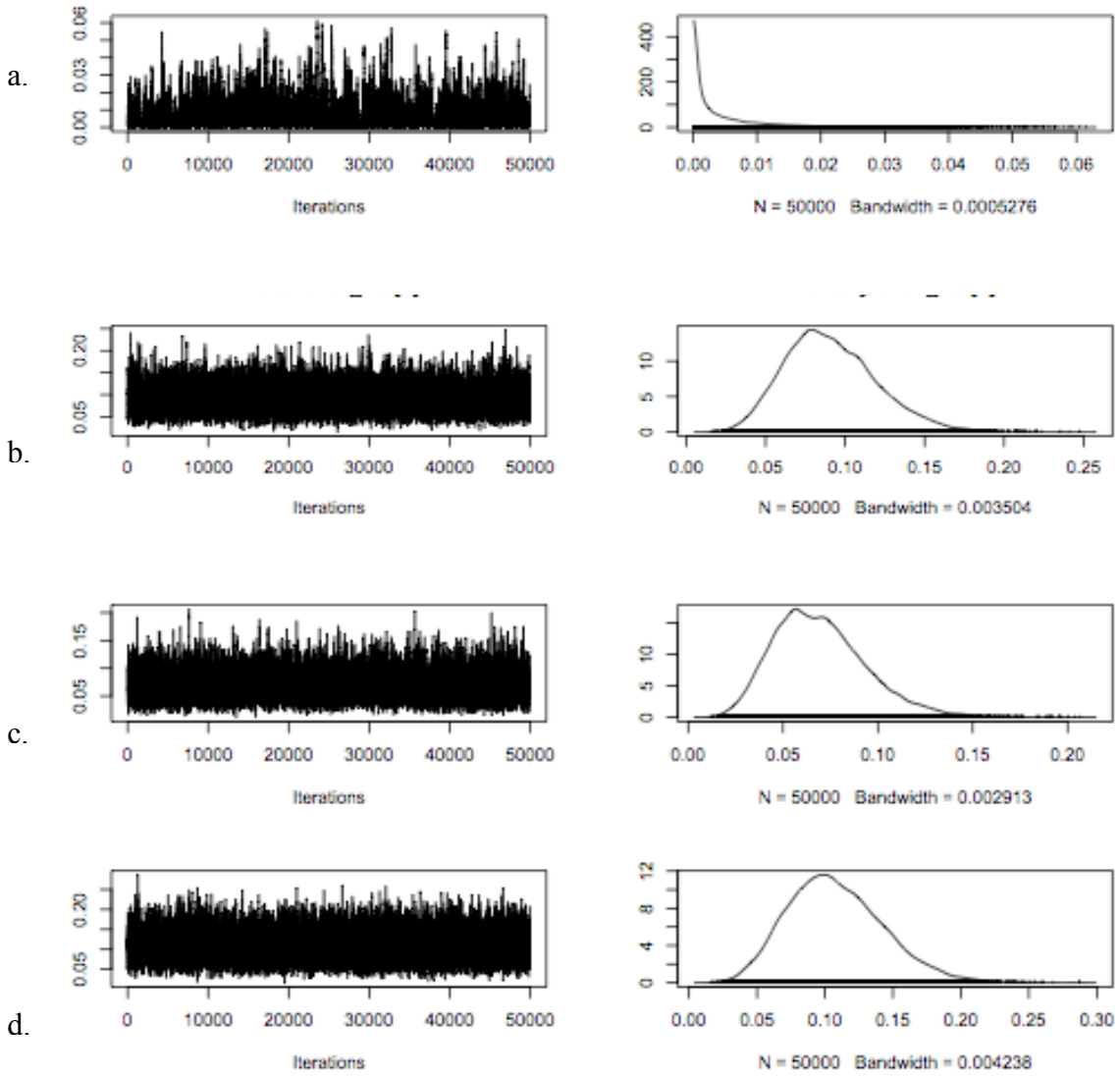


Figure 3.9. Posterior densities and traces of locus specific ADO probabilities for locus 5 (a), locus 6 (b), locus 7 (c), and locus 8 (d) under the model of 16 loci with informative priors including records that were classified as 'no data'.

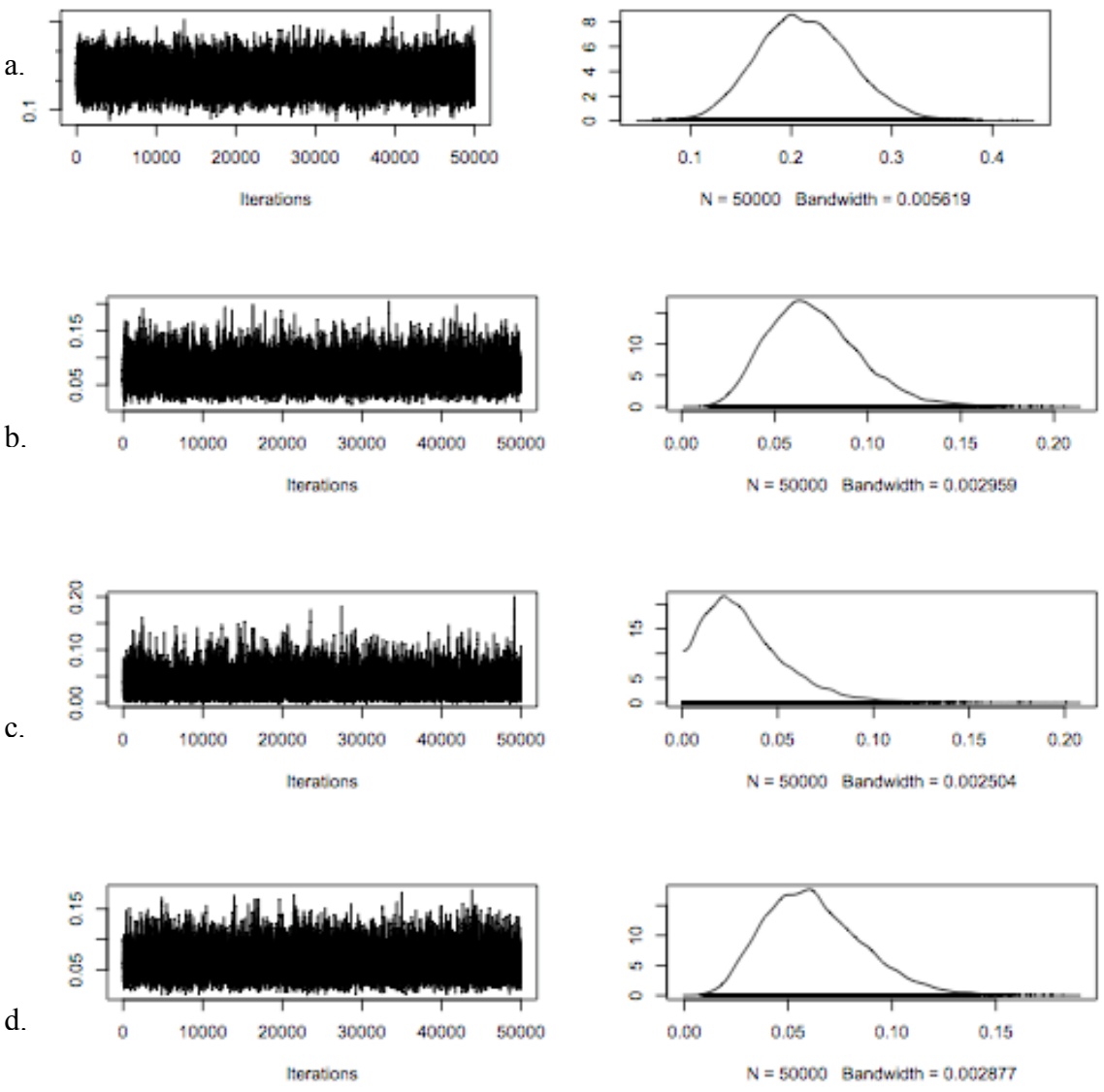


Figure 3.10. Posterior densities and traces of locus specific ADO probabilities for locus 9 (a), locus 10 (b), locus 11 (c), and locus 12 (d) under the model of 16 loci with informative priors including records that were classified as ‘no data’.

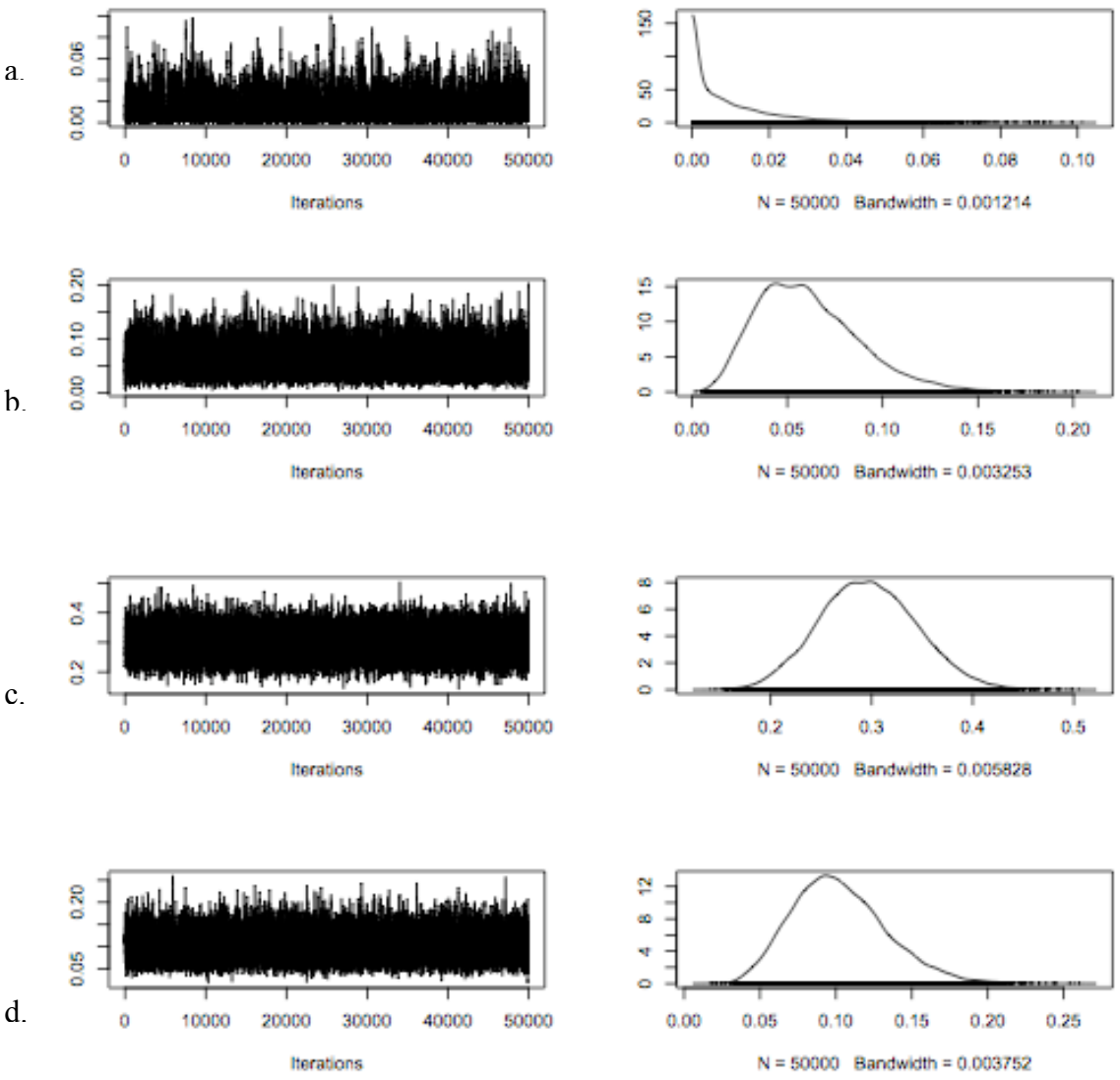


Figure 3.11. Posterior densities and traces of locus specific ADO probabilities for locus 13 (a), locus 14 (b), locus 15 (c), and locus 16 (d) under the model of 16 loci with informative priors including records that were classified as 'no data'.

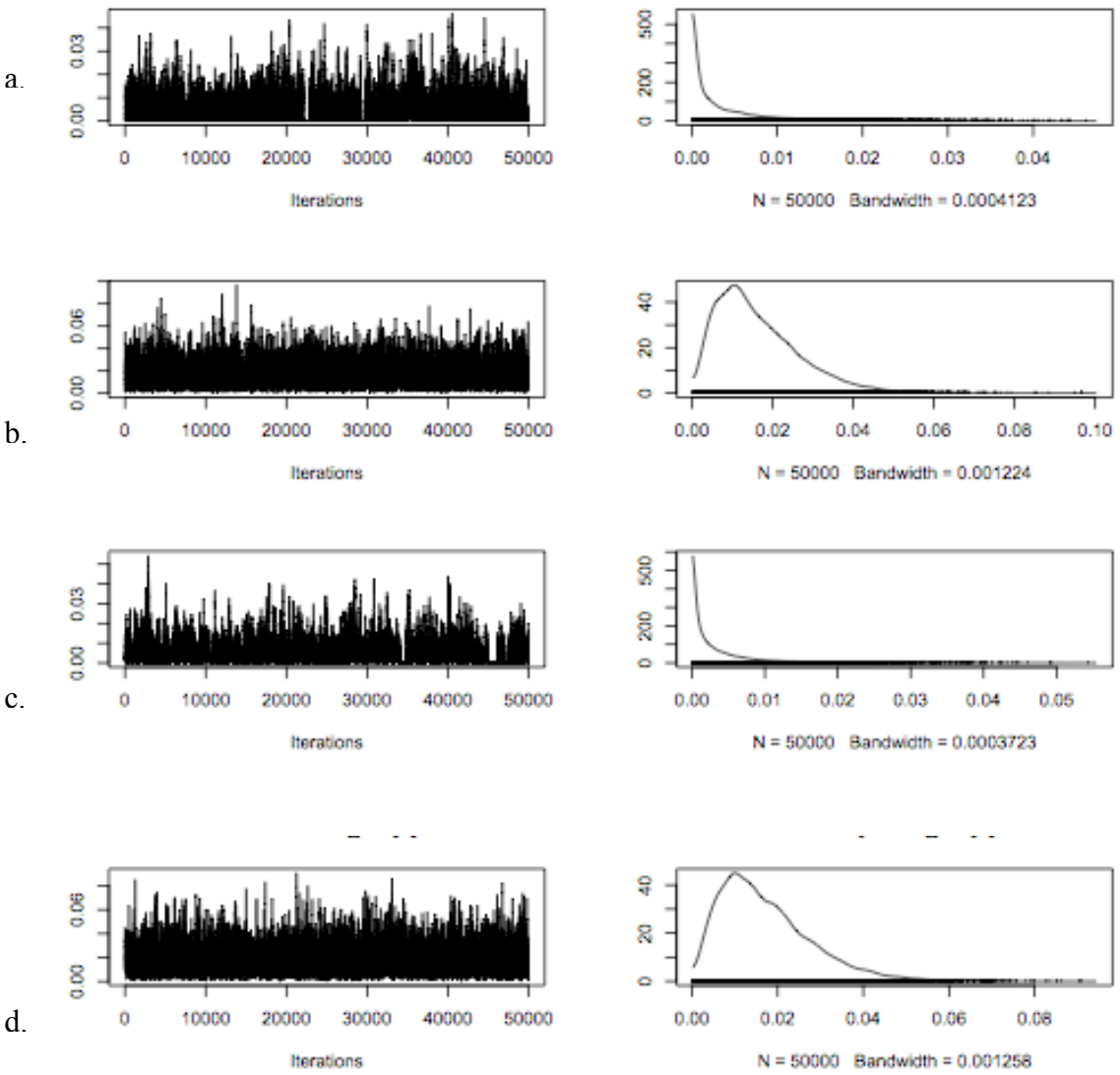


Figure 3.12. Posterior densities and traces of locus specific FA probabilities for locus 1 (a), locus 2 (b), locus 3 (c), and locus 4 (d) under the model of 16 loci with informative priors including records that were classified as 'no data'.

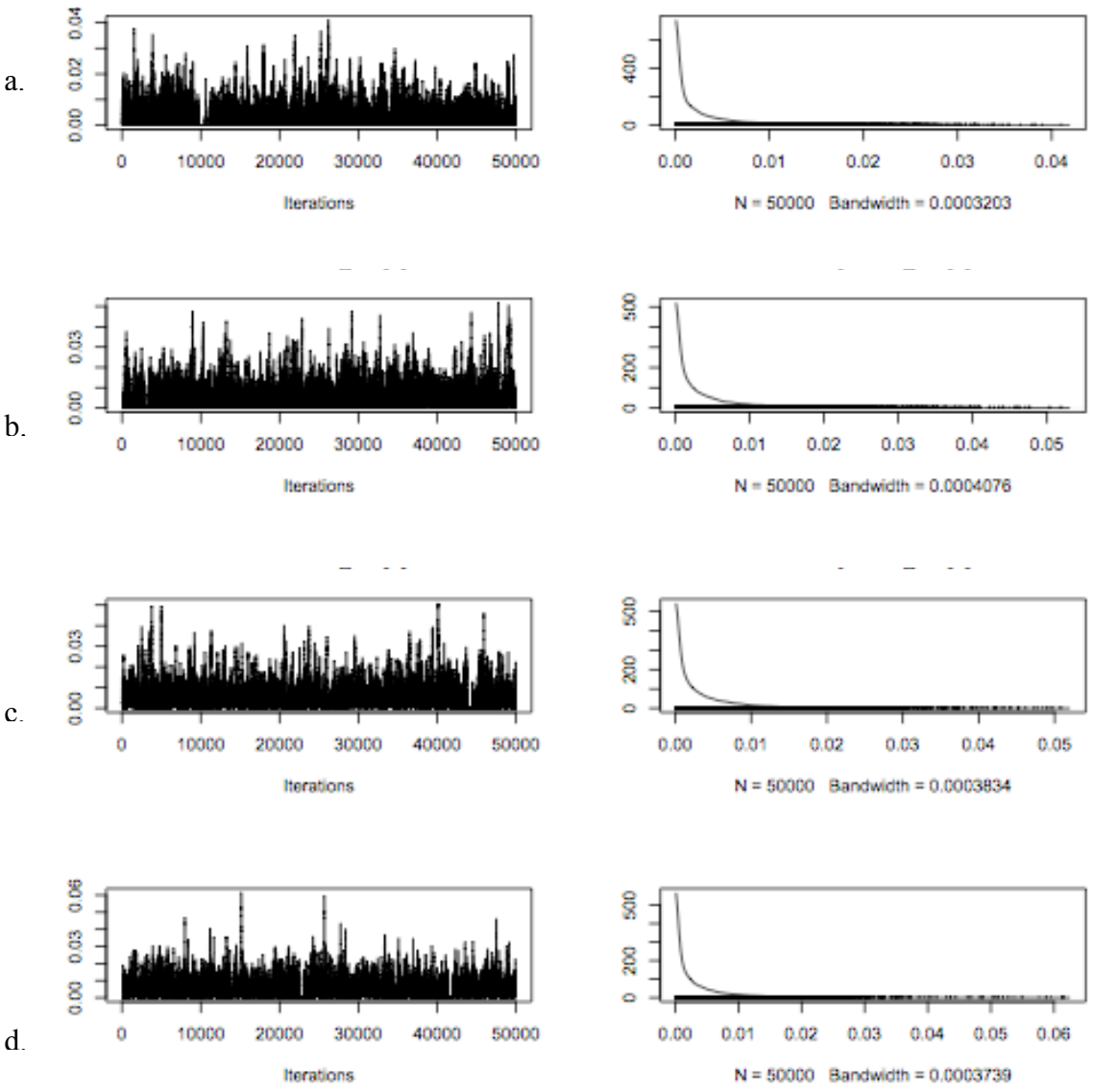
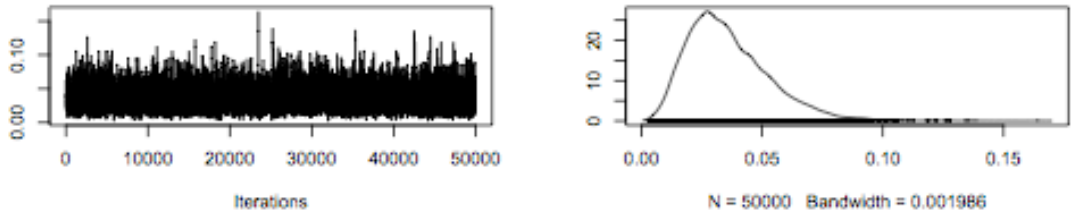
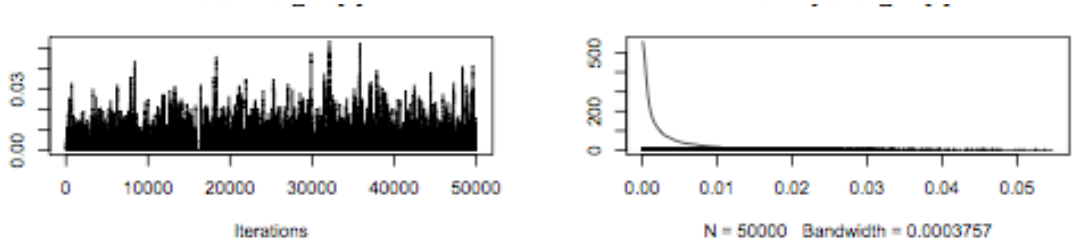


Figure 3.13. Posterior densities and traces of locus specific FA probabilities for locus 5 (a), locus 6 (b), locus 7 (c), and locus 8 (d) under the model of 16 loci with informative priors including records that were classified as 'no data'.

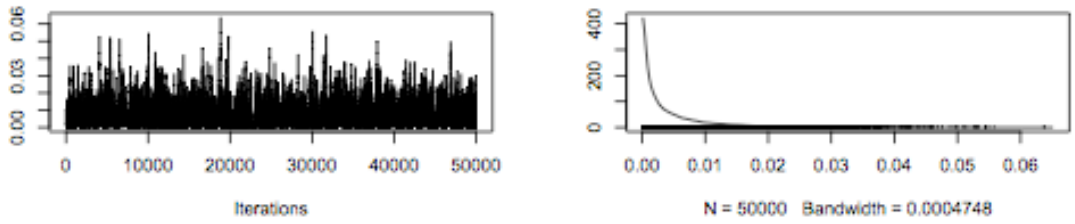
a.



b.



c.



d.

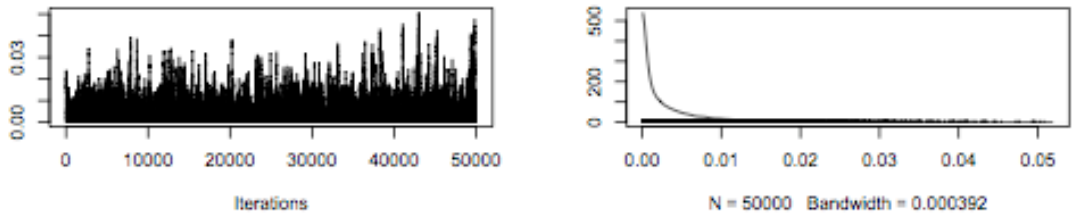


Figure 3.14. Posterior densities and traces of locus specific FA probabilities for locus 9 (a), locus 10 (b), locus 11 (c), and locus 12 (d) under the model of 16 loci with informative priors including records that were classified as ‘no data’.

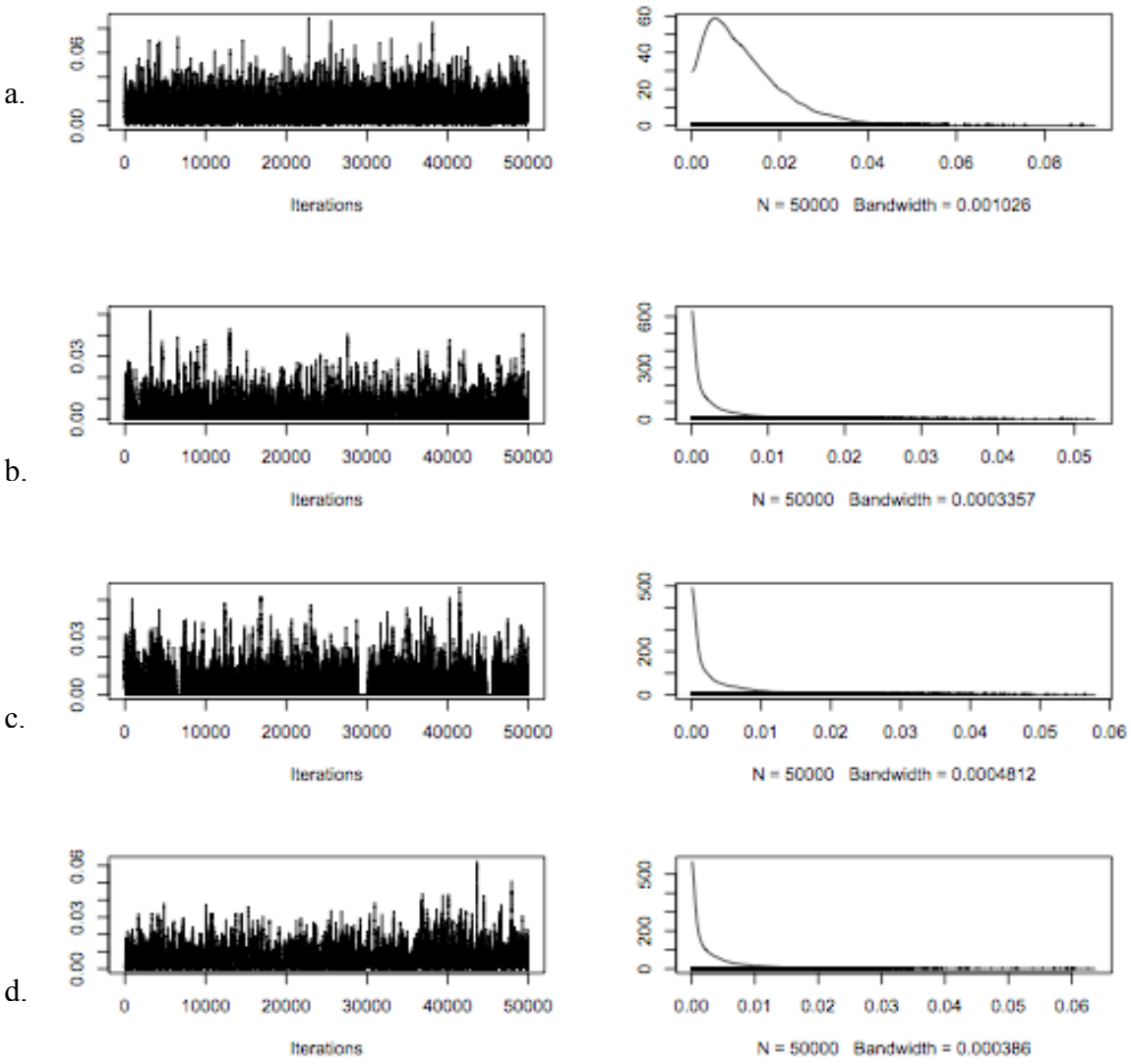


Figure 3.15. Posterior densities and traces of locus specific FA probabilities for locus 13 (a), locus 14 (b), locus 15 (c), and locus 16 (d) under the model of 16 loci with informative priors including records that were classified as 'no data'.

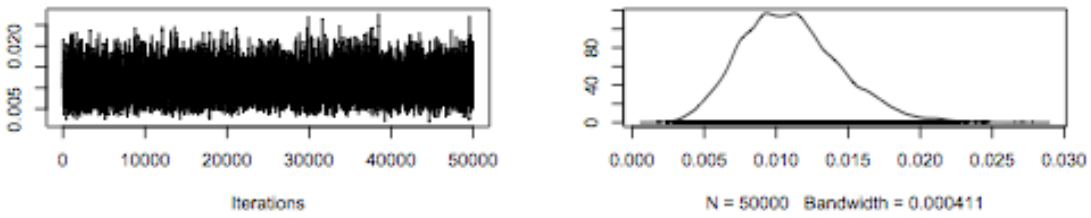


Figure 3.16. Posterior density and trace of the ADO probability over all loci under the model of 16 loci with informative priors without including records that were classified as ‘no data’.

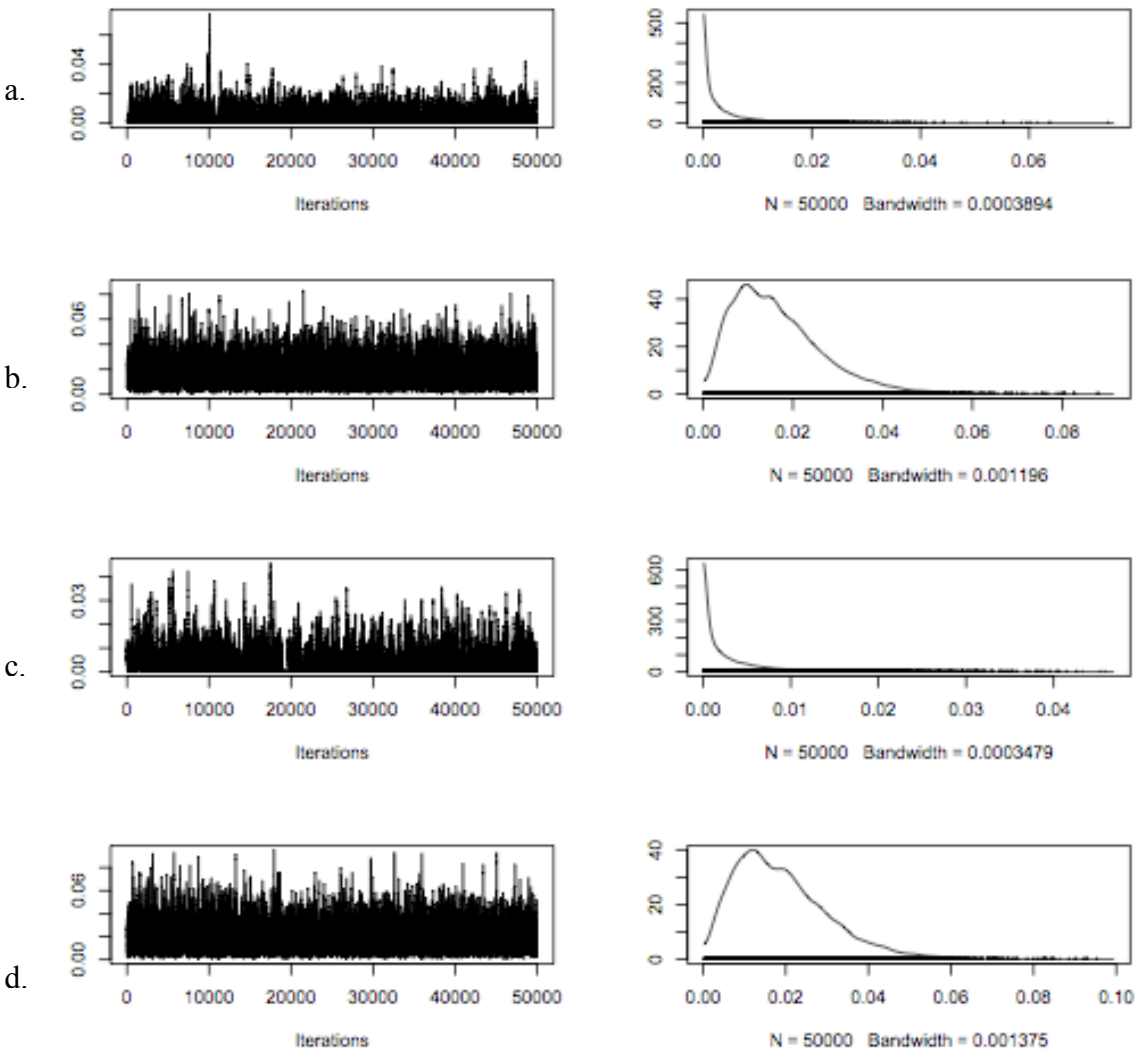


Figure 3.17. Posterior densities and traces of locus specific FA probabilities for locus 1 (a), locus 2 (b), locus 3 (c), and locus 4 (d) under the model of 16 loci with informative priors without including records that were classified as 'no data'.

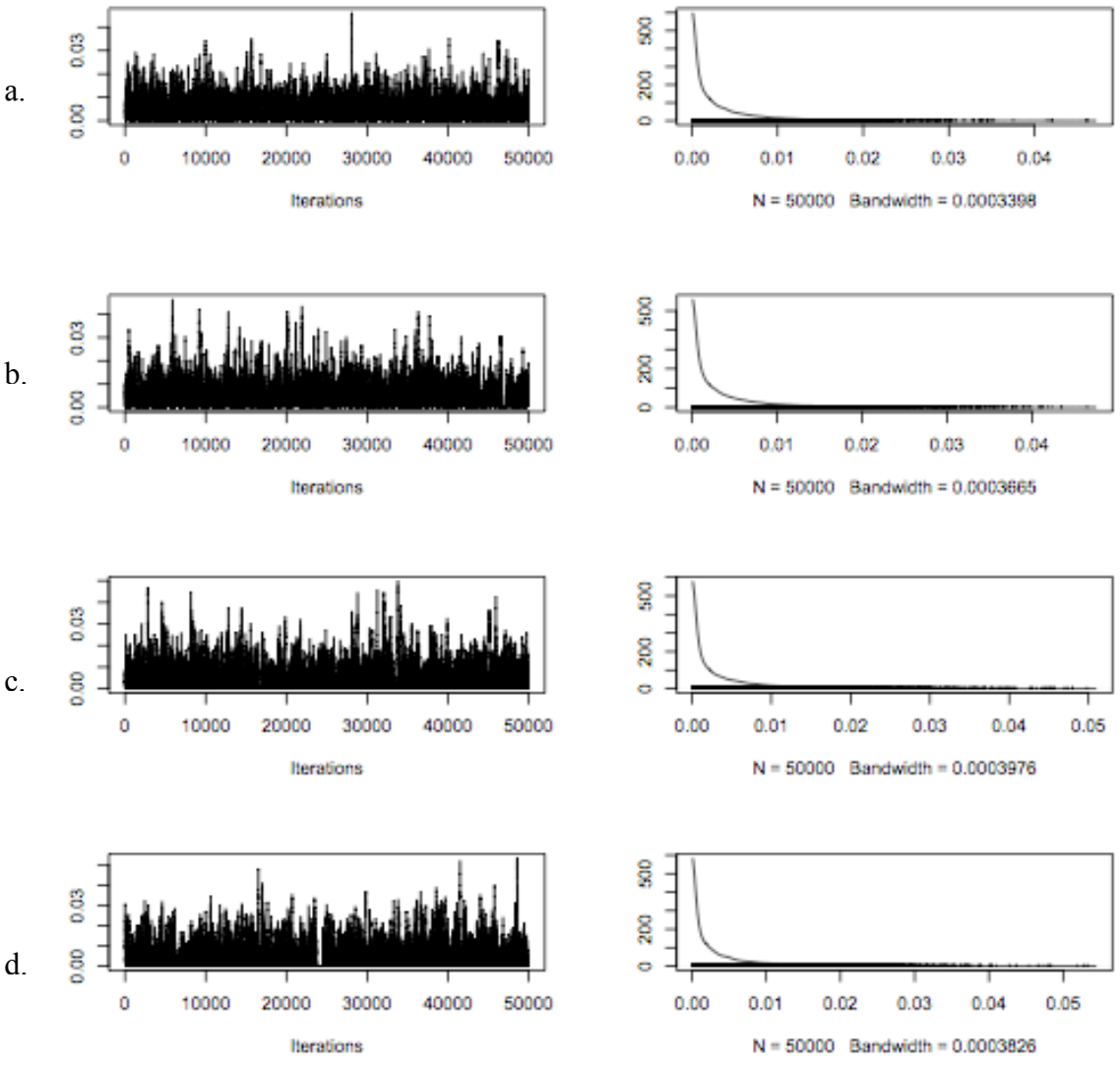


Figure 3.18. Posterior densities and traces of locus specific FA probabilities for locus 5 (a), locus 6 (b), locus 7 (c), and locus 8 (d) under the model of 16 loci with informative priors without including records that were classified as 'no data'.

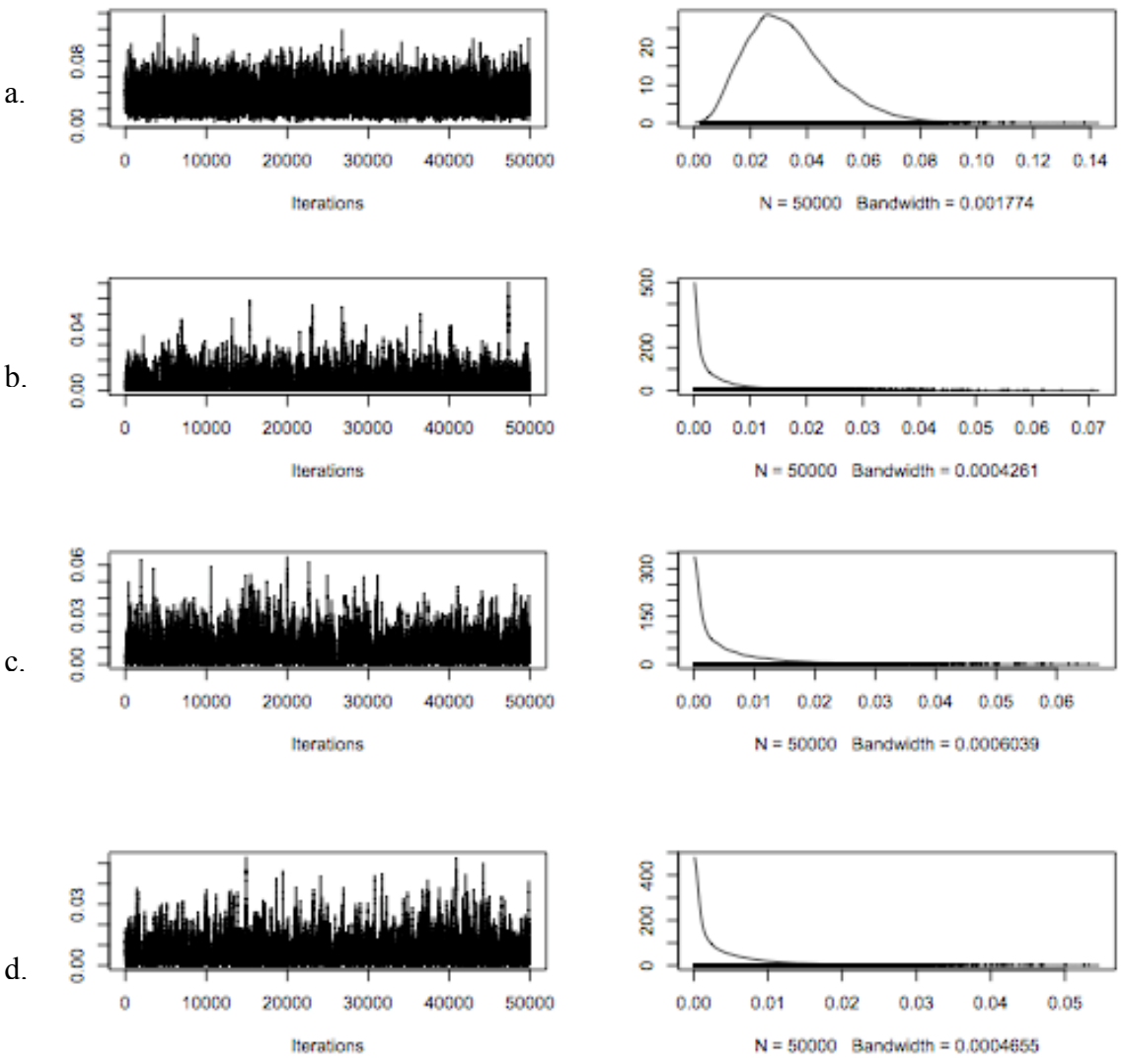
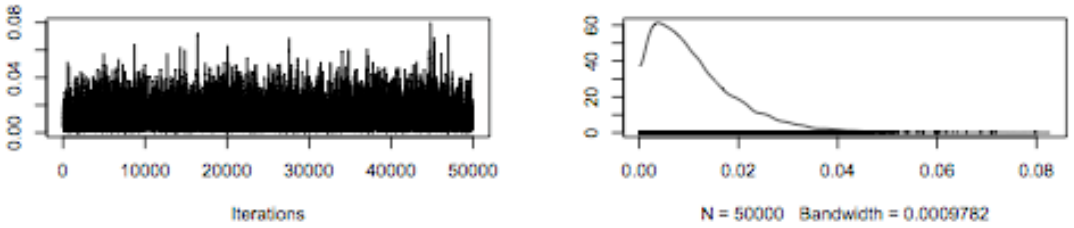
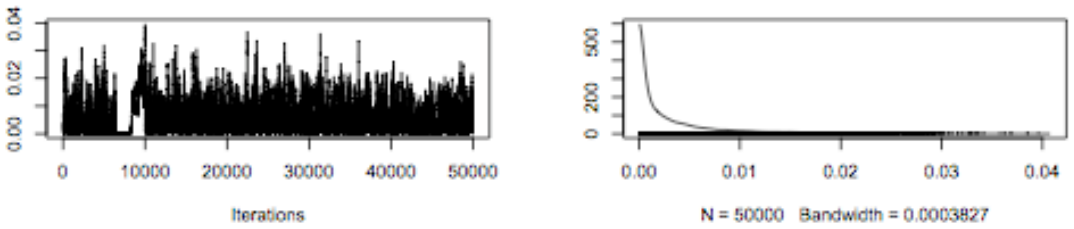


Figure 3.19. Posterior densities and traces of locus specific FA probabilities for locus 9 (a), locus 10 (b), locus 11 (c), and locus 12 (d) under the model of 16 loci with informative priors without including records that were classified as 'no data'.

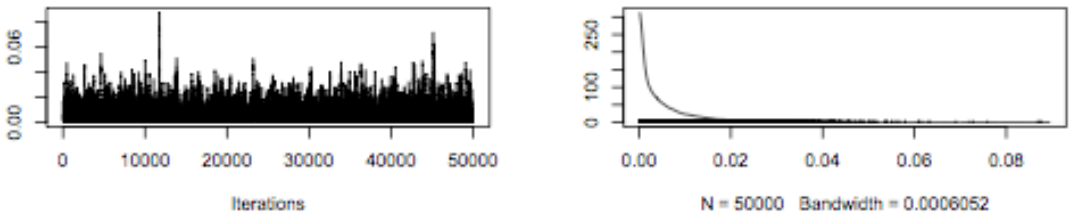
a.



b.



c.



d.

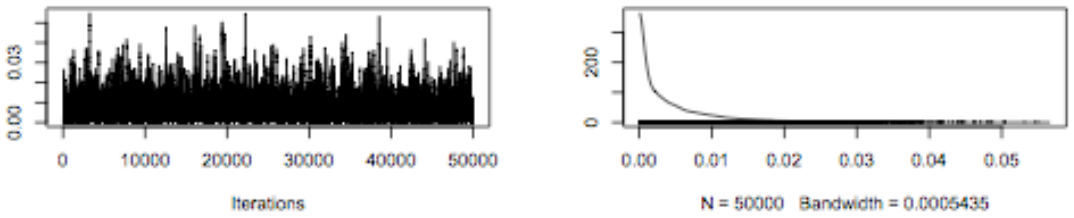
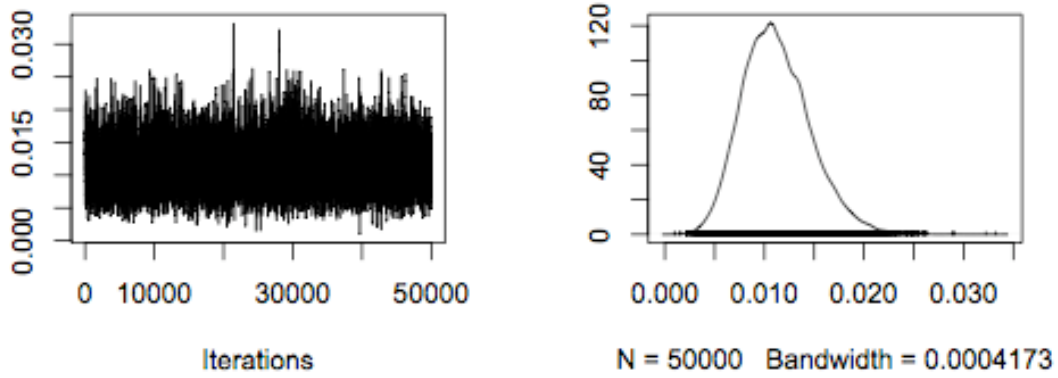


Figure 3.20. Posterior densities and traces of locus specific FA probabilities for locus 13 (a), locus 14 (b), locus 15 (c), and locus 16 (d) under the model of 16 loci with informative priors without including records that were classified as ‘no data’.

a.



b.

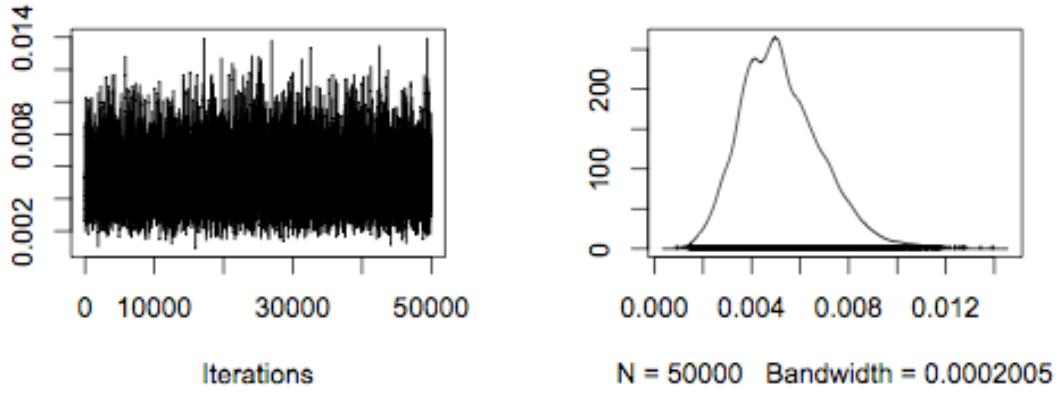
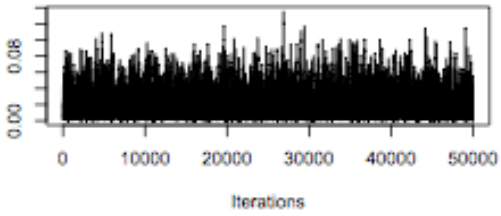
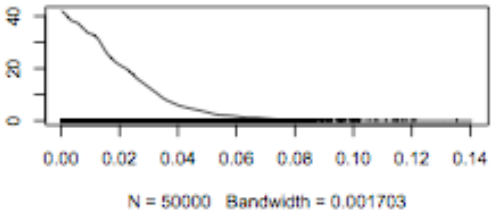


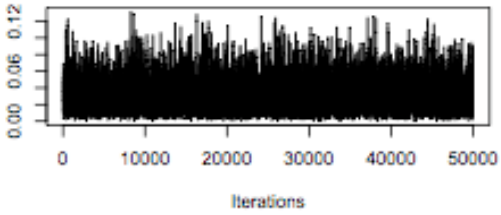
Figure 3.21. Posterior densities and traces of the ADO probability (a) and the FA probability (b) over all loci under the model of 16 loci with informative priors without including records that were classified as ‘no data’.

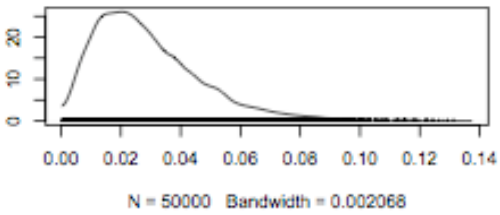
a.



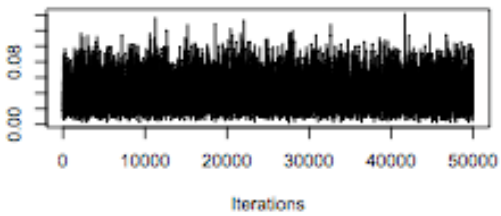


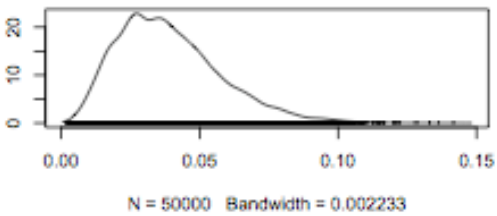
b.



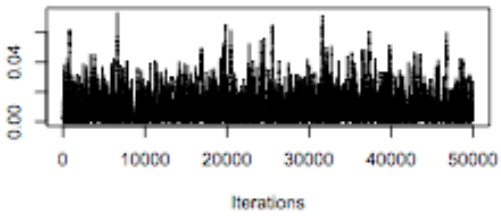


c.





d.



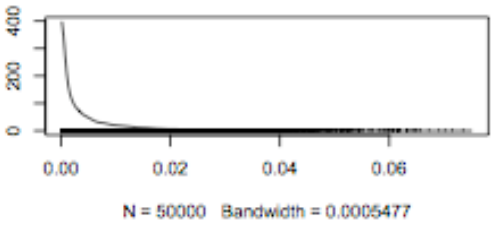


Figure 3.22. Posterior densities and traces of locus specific ADO probabilities for locus 1 (a), locus 2 (b), locus 3 (c), and locus 4 (d), under the model of 9 loci with informative priors including records that were classified as ‘no data’.

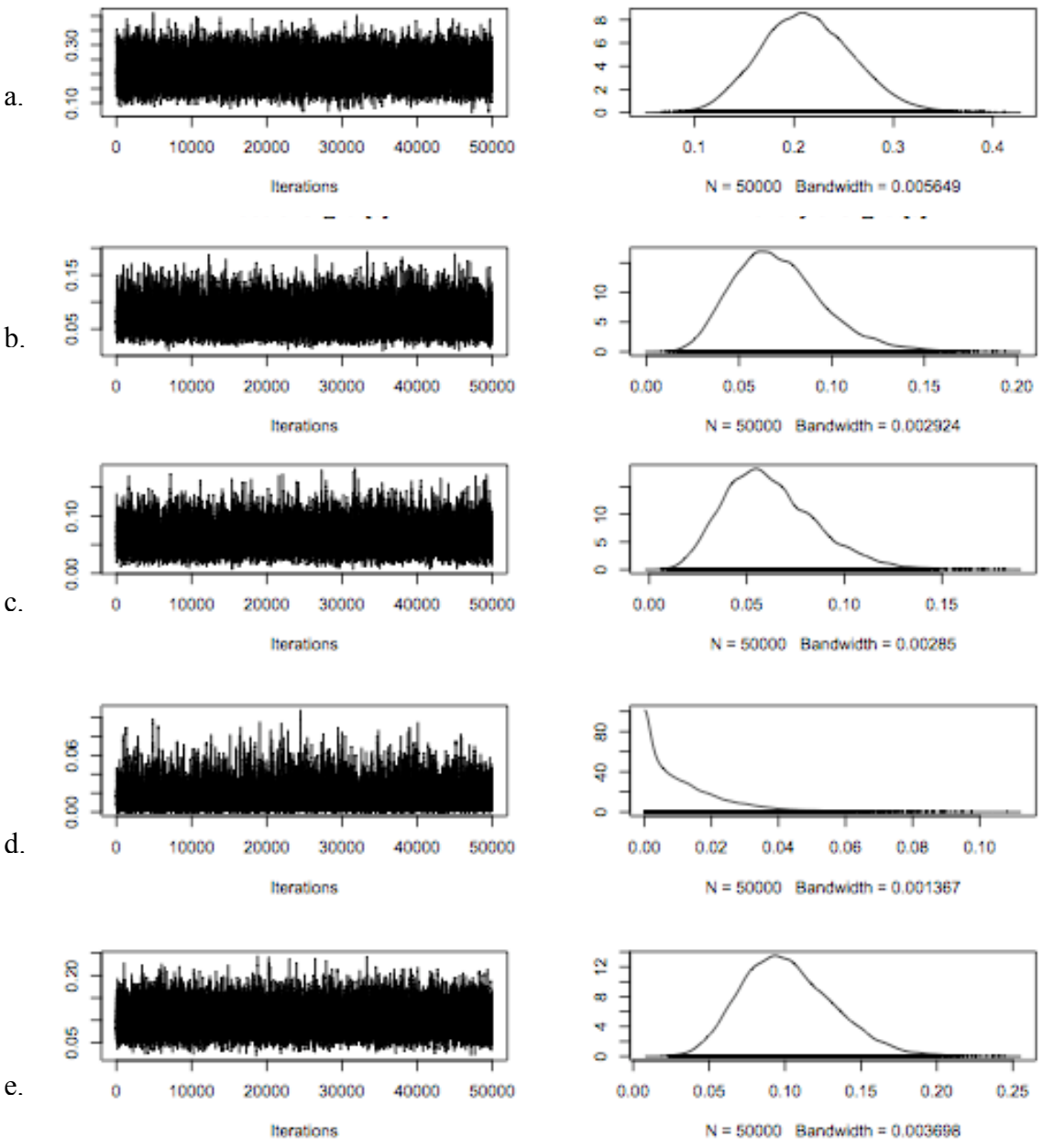


Figure 3.23. Posterior densities and traces of locus specific ADO probabilities for locus 5 (a), locus 6 (b), locus 7 (c), locus 8 (d), and locus 9 (e) under the model of 9 loci with informative priors including records that were classified as 'no data'.

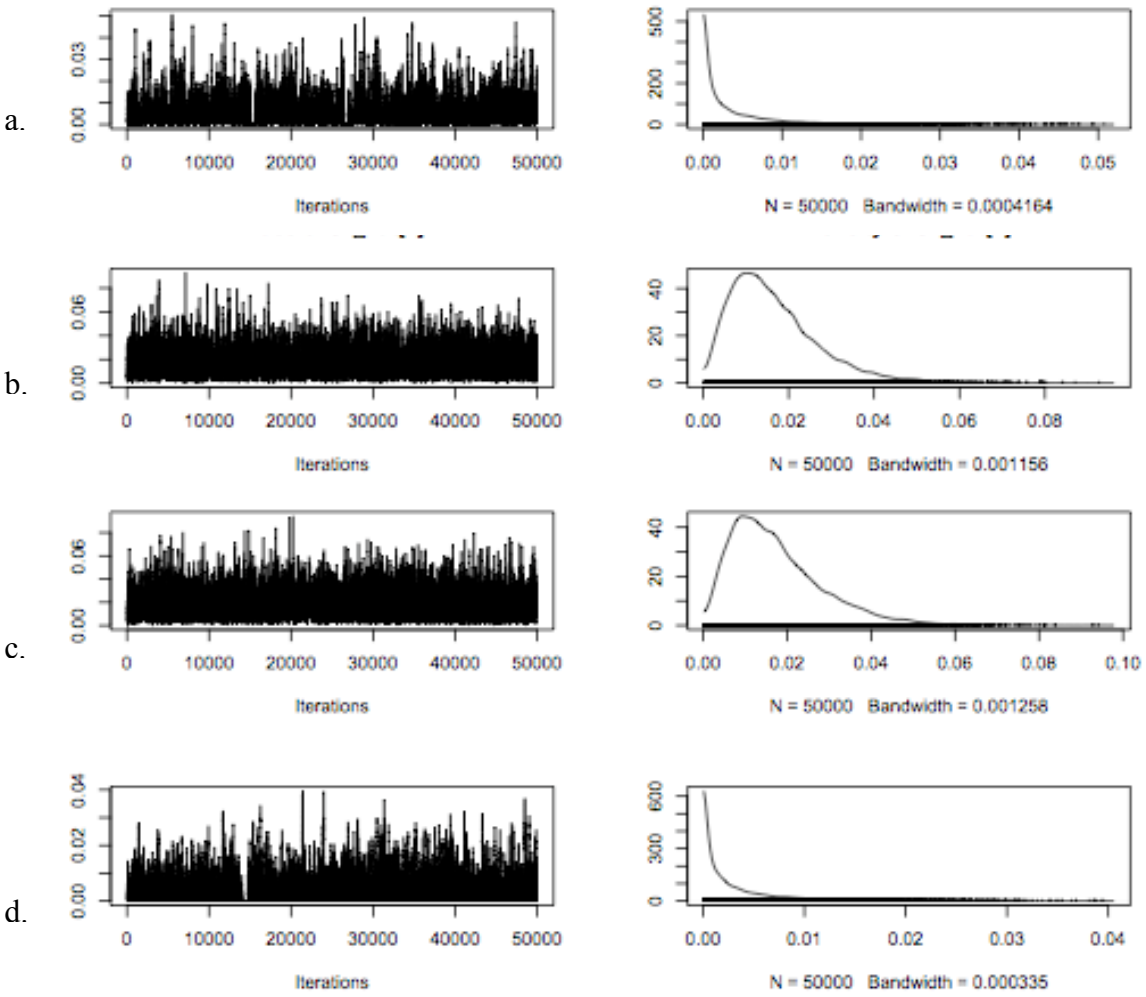


Figure 3.24. Posterior densities and traces of locus specific FA probabilities for locus 1 (a), locus 2 (b), locus 3 (c), and locus 4 (d) under the model of 9 loci with informative priors including records that were classified as 'no data'.

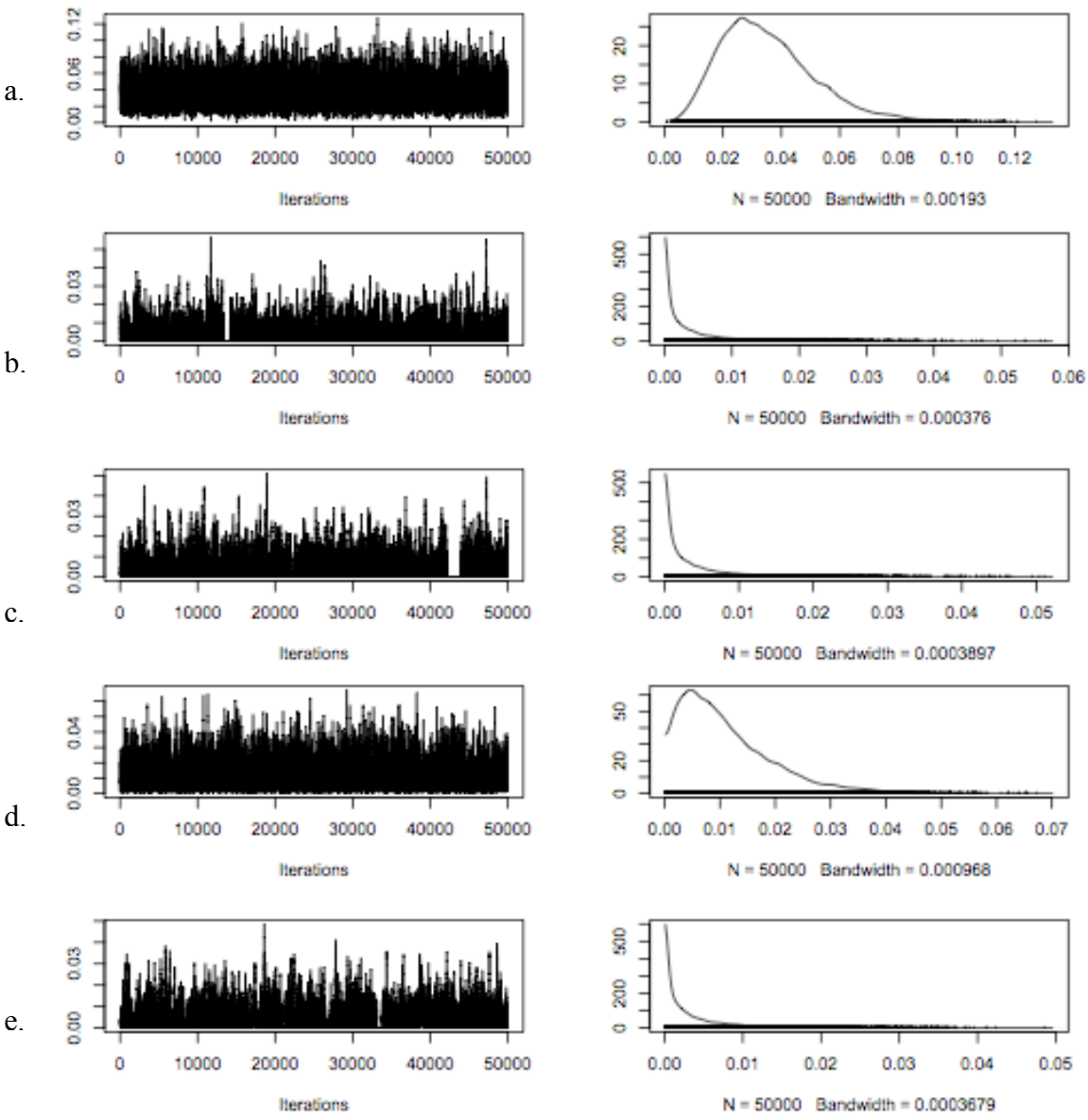


Figure 3.25. Posterior densities and traces of locus specific FA probabilities for locus 5 (a), locus 6 (b), locus 7 (c), locus 8 (d), and locus 9 (e) under the model of 9 loci with informative priors including records that were classified as 'no data'.

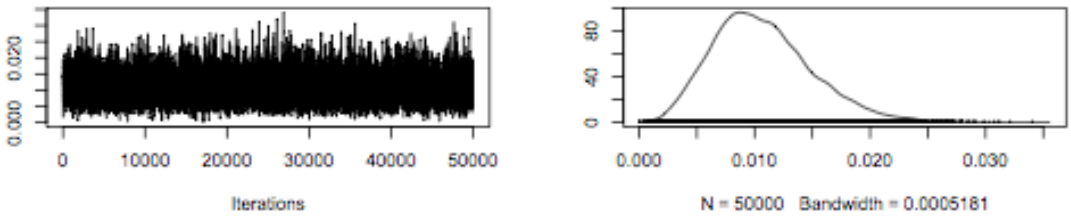


Figure 3.26. Posterior density and trace of the ADO probability over all loci under the model of 9 loci with informative priors without including records that were classified as ‘no data’.

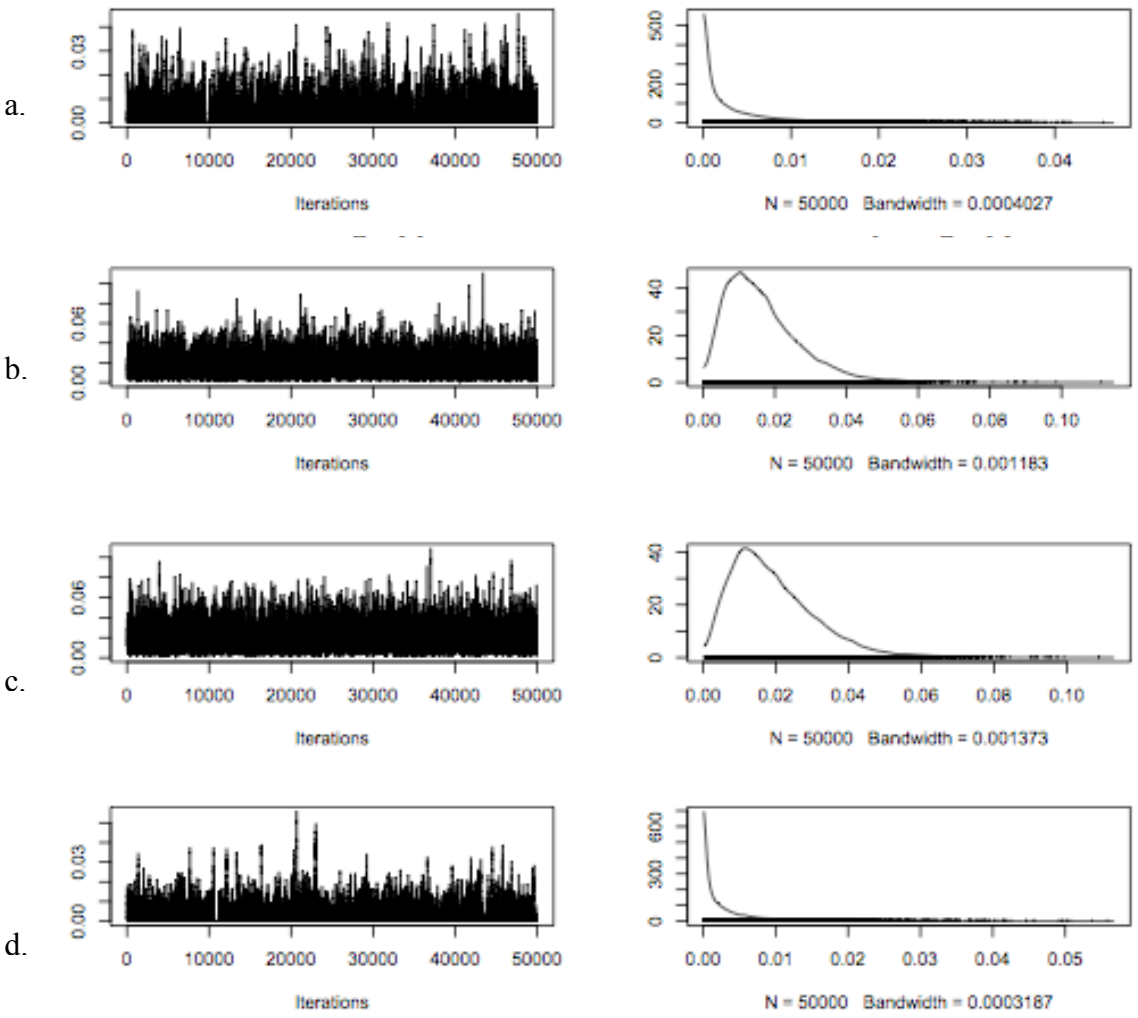


Figure 3.27. Posterior densities and traces of locus specific FA probabilities for locus 1 (a), locus 2 (b), locus 3 (c), and locus 4 (d) under the model of 9 loci with informative priors without including records that were classified as 'no data'.

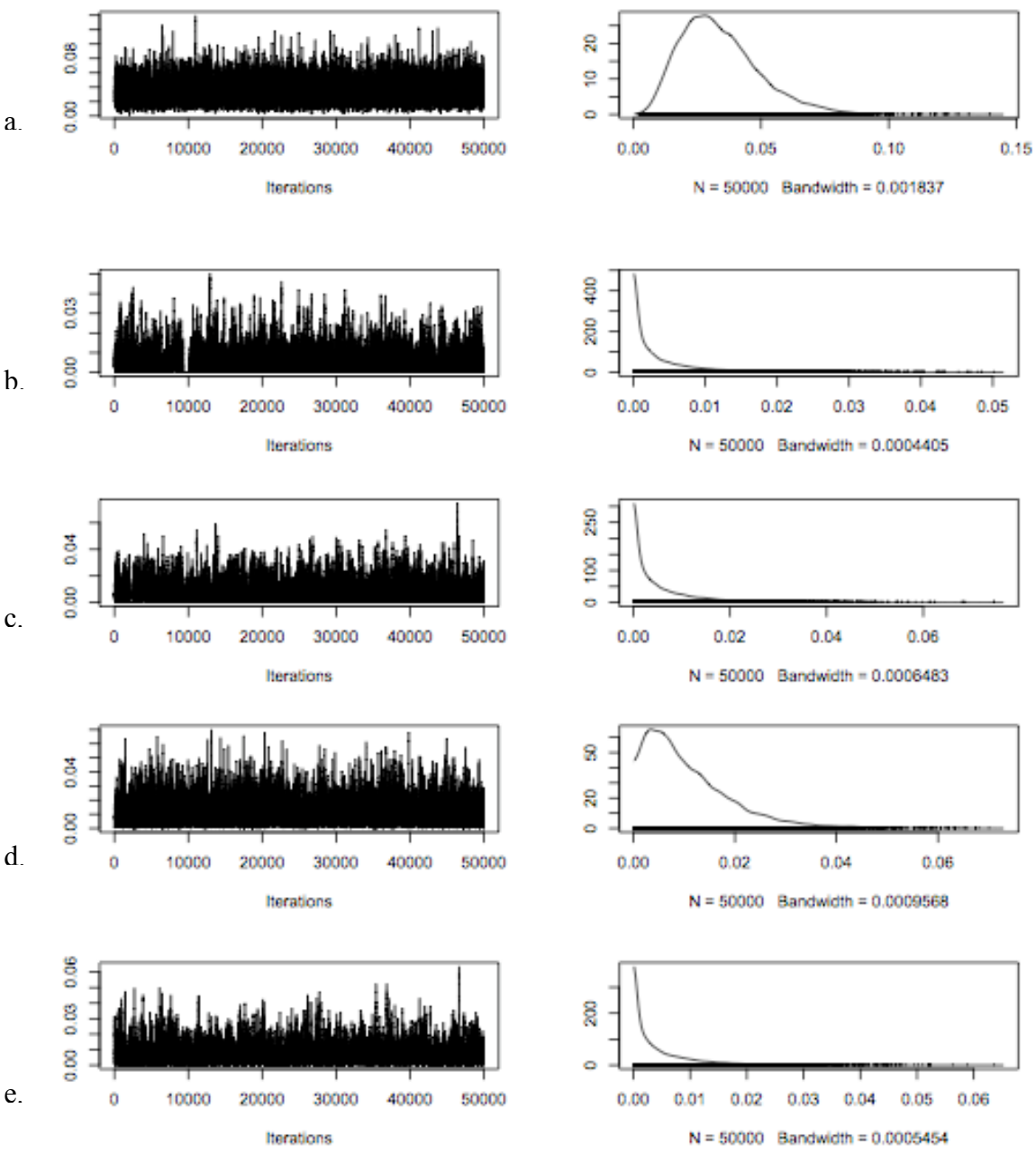
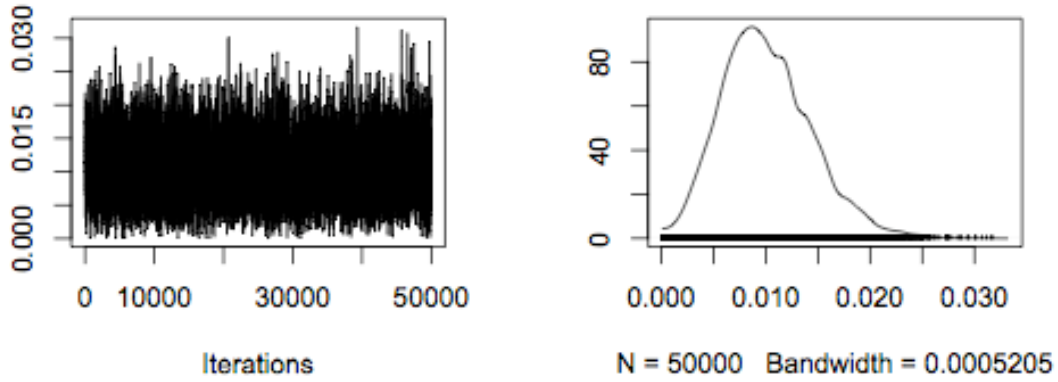


Figure 3.28. Posterior densities and traces of locus specific FA probabilities for locus 5 (a), locus 6 (b), locus 7 (c), locus 8 (d), and locus 9 (e) under the model of 9 loci with informative priors without including records that were classified as 'no data'.

a.



b.

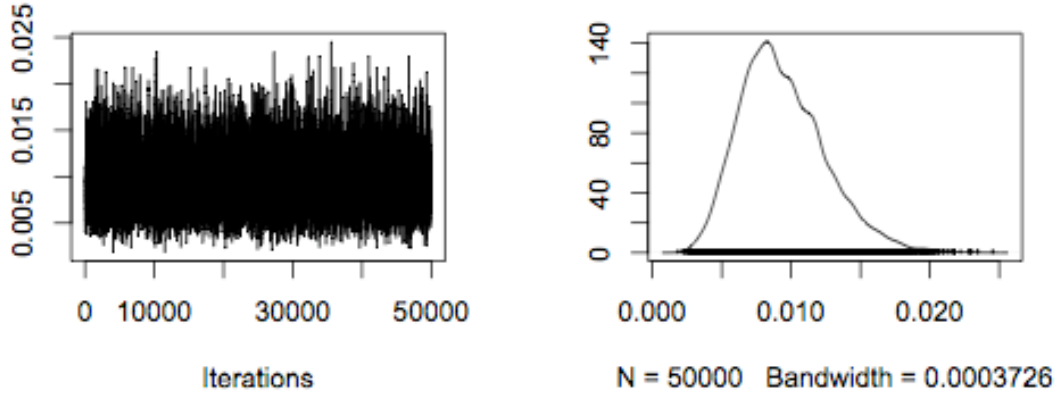


Figure 3.29. Posterior densities and traces of the ADO probability (a) and the FA probability (b) over all loci under the model of 9 loci with informative priors without including records that were classified as ‘no data’.

CHAPTER 4

COST-EFFICIENT SELECTION OF A MARKER PANEL IN NON-INVASIVE STUDIES[1]

[1] Sanderlin, J. S., N. Lazar, M. J. Conroy, and J. Reeves. To be submitted to *Molecular Ecology Resources* or *Conservation Genetics*.

Introduction

Non-invasive sampling techniques, or techniques that do not require physical capture of animals, provide more opportunities to sample and monitor wildlife populations, particularly populations of species that occur in low densities or have elusive behavior. Genetic samples (e.g., shed hairs, feathers, feces, shed skin), collected non-invasively in the field, are often small and contain degraded DNA. The ability to create multiple copies of DNA from these samples with PCR (polymerase chain reaction) has advanced non-invasive genetic sampling techniques (Waits 1999). Non-invasive genetic samples are currently utilized with many animal species globally to ask questions pertaining to demographics (Taberlet et al. 1997, Boulanger et al. 2002, Bellemain et al. 2005, Banks et al. 2003, Creel et al. 2003, Kendall et al. 2008, Prugh et al. 2005), habitat relationships (Apps et al. 2004), paternity and mating systems (Constable et al. 2001, Garnier et al. 2001, Oka and Takenaka 2001), and dispersal and effectiveness of corridors (Dixon et al. 2006, Proctor et al. 2004).

In any genetic analysis, especially non-invasive analyses, two common types of errors occur at a specific locus with respect to alleles, namely allelic dropout and false alleles. In allelic dropout, one or both of the two alleles may be lost during the analysis. Allelic dropout is often a result from sampling stochasticity in the laboratory, amplification of small amounts of DNA and pipetting template DNA into a dilute sample (Goossens et al. 1998, Taberlet et al. 1999, Woods et al. 1999). A false allele is the addition of one allele to a locus during an analysis. False alleles can occur with PCR amplification artifacts from dinucleotide microsatellites (Goossens et al. 1998, Taberlet et al. 1999, Rodriguez et al. 2001, and Woods et al. 1999) or contamination of samples. An accumulation of both types of genetic error may lead to misidentification of individuals within a population.

Genetic error and the number of loci

The number of microsatellite loci used to determine individuals varies. Without genetic error, more loci will theoretically increase the ability to determine individuals at a higher cost, in terms of time or monetary resources. The minimum number of markers should reduce the "shadow effect" (Mills et al. 2000), or when several individuals have the same genetic tag. This is often a result of using too few loci or loci with low heterozygosity. A reduction in the shadow effect is normally achieved through reducing the probability of identity (P_{ID}), or the probability that two randomly chosen individuals in a population will have identical genotypes (Paetkau and Strobeck 1994) using the following formula:

$$\sum_i p_i^4 + \sum_i \sum_{j>i} (2 p_i p_j)^2, \qquad \text{(eqn 1)}$$

where p_i and p_j are the frequencies of the ith and jth alleles. A more conservative estimate is P_{IDsib}, or the probability of identity among siblings (Evett and Weir 1998), summarized below:

$$0.25 + (0.5 \sum p_i^2) + [0.5(\sum p_i^2)^2] - (0.25 \sum p_i^4) \qquad \text{(eqn 2)}$$

where p_i is the frequency of the ith allele.

Non-invasive studies tend to be more sensitive to biases with the selection of a marker panel and the number of loci in the panel. Waits and Leberg (2000) performed several computer simulations of simulated population sizes, with individuals randomly assigned alleles from microsatellite loci similar to frequencies of a native wolf population. They varied sampling intensity and error rates of genotype assignment. They concluded that as the population size increases, more loci are needed to distinguish among related individuals, at the cost of increasing the number of genetic errors. Waits and Leberg (2000) proposed using fewer, but more informative loci to reduce errors in genotyping. Similarly, Creel et al. (2003) found overestimation of population size was greatest when the maximum number of loci were used in

the analysis, and suggest the minimum number of loci with a low probability of identity is optimal. However, Hoffman and Amos (2005) suggest more heterozygous loci with greater number of alleles tend to possess more stutter bands, leading to more errors. In addition, Hoffman and Amos (2005) found that mode product size was strongly correlated with the number of errors in their study.

Paetkau (2004) states the number of markers in a study will affect the "efficiency not efficacy" due to reanalysis of samples; therefore, an increase in markers decreases efficiency. Since the degree of relatedness among individuals in a population, population size, or degree of isolation is unknown in a population, it may be difficult to choose an optimal number of markers. Often, evaluation of the number of markers can only be completed in "retrospect" (Paetkau 2004). Although the history of a population is unknown at the onset of a study, a reasonable solution is to evaluate a population subsample in a pilot study for optimal selection of markers. Optimal selection would be based both on project goals and the amount of time and money available for the study. There are guidelines to reduce error in non-invasive genetic studies, but no formal techniques or algorithms for choosing a marker set.

Objectives and case study

The central Georgia population (CGP) of black bears (*Ursus americanus*) is considered to inhabit mostly forested land in and around 186 km^2 associated with the Ocmulgee River drainage system, and likely a core area of contiguous forest in the Oaky Woods and Ocmulgee Wildlife Management Areas (WMAs). A current, more accurate estimate is needed to make informed management decisions with this population. Various sampling techniques were applied to estimate density, some of which include non-invasive genetic sampling from DNA hair snares

(Woods et al. 1999). Tissue and hair samples from known individuals from the CGP were collected to evaluate genetic error as a calibration sample with the genetic misidentification error model described in chapter 2. The main research objectives of this larger project were to estimate demographic parameters (e.g., survival and reproduction) and density to construct models for predicting population viability of the CGP of black bears.

The black bear CGP serves as a case study for the general question of how to select a marker panel, given an excess amount of markers available for a study species. The algorithm was evaluated with bear tissue samples of known identity, and hair samples from the same individuals. These samples from known individuals were used to determine the marker panel for all subsequent hair samples from unknown individuals in the population used in the larger project. The overall objective was to optimize the number of markers included in the panel with the minimal probability of identity and minimum level of genetic error at a fixed cost.

Methods

Study area

The principal study areas for bear physical captures were Ocmulgee and Oaky Woods WMAs in Bleckley, Bibb, Houston, Pulaski, and Twigg Counties, located in central Georgia, USA. The WMAs consist of a variety of habitat types (pine stands, bottomland hardwood, mixed forest, upland hardwood, black-belt prairie, clearcuts, thinned pine stands, and cypress-gum swamps).

Field Methods

Bears were captured and immobilized with a 2:1 mixture of ketamine hydrochloride (Ketaset) and xylazine hydrochloride (Rompun) at a dosage of 4.4 mg/kg of Ketaset and 2.2 mg/kg of Rompun, for estimated body weights by Georgia Department of Natural Resources personnel. Bears were captured in the study area (Figure 4.1) using Fremont foot trap snares (Freemont 1986) in each of 4 trapping seasons, which extend from May through August each year (2003-2006). Culvert traps were used to trap nuisance bears and released on Oaky Woods WMA. An upper pre-molar tooth for age estimation by cementum annuli analysis (Willey 1974), blood samples, and hair follicles were collected from each captured bear. The tissue and hair samples (n=85) were used in the calibration genetic error and marker panel selection analyses with known captured individuals (52M: 31F), plus samples collected by DNR personnel of one road mortality (1 M) and one capture mortality (1 M).

Laboratory Methods

After field collection, hair samples were stored in silica desiccant then transferred to a -20°C freezer. Prior to extraction and after field collection, blood and tissue samples were also stored in a -20°C freezer. Extraction of DNA from Georgia tissue samples was done with the DNeasy Kit (QIAGEN) and with one captured bear blood sample using the GenomicPrep DNA isolation kit (GE Healthcare). DNA from hair samples were extracted with Chelex100 (10% solution) (Promega) and proteinase K (Phenix Research Products, QIAGEN) (modified from Boersen 2001). The root portion (1 cm.) from a maximum of 10 hairs per sample were cut and placed into 150 μl of Chelex 100 (10% solution) (Promega). The number of hairs and quality of sample was recorded. If the number of hairs was less than 10, the entire strand of hair was used

in the sample. Low quality samples had little or no visible roots, and usually consisted of under-fur (thin) hairs. Medium quality samples were classified as samples with half of the hairs with roots visible and some guard hairs. High quality samples were classified by a majority of the hairs as guard hairs with most or all of the roots visible, and including visible skin cells. After roots were placed in the 10% solution Chelex 100 (Promega), 10 *µl* proteinase-K was added to assist with DNA digestion. The hair samples were incubated at 65 °C overnight (~8 hours). Samples were vortexed, and then boiled at 100°C for 10 minutes. After removal, the samples were centrifuged at 10,000-12,000 rpm for 3 minutes. The supernatant was pulled off and placed into a clean tube, and stored at -20°C until PCR analysis.

PCR amplifications were performed in 10 µL volumes using Bio-Rad MyCycler thermal cyclers for both tissue and hair samples with 16 tetranucletide loci (UA-BM3-P1F04, UA-BM4-P1H06, UA-BM4-P1H10, UA-BM4-P2B06, UA-BM4-P2C10, UA-BM3-P1A08, UA-BM4-P2E11, UA-RM3-P2G10, UA-RM3-P2H03 , UA-BM3-P1D05, UA-BM4-P2A02, UA-BM4-P2A06, UA-BM4-P2B08, UA-BM3-P1B05, UA-RM3-P2H01, and BM4-P2C02, hereafter named Bear 10Y, Bear 12Y, Bear 13, Bear 17G, Bear 19Y, Bear 2, Bear 23, Bear 25, Bear 27B, Bear 30B, Bear 32, Bear 33B, Bear 35G, Bear 6, Bear 26, and Bear 36, respectively) previously described in Sanderlin et al. (2009). Loci Bear 10Y, Bear 12Y, Bear 17G, Bear19Y, Bear 27B, Bear 30B, Bear 33B, and Bear 35G were directly labeled primers with the dyes NED (Y), HEX (G), and FAM (B). Final concentrations for optimizing reactions with unlabelled primers were 10 mM Tris pH 8.4, 50 mM KCl, 0.5 µM "pigtailed" primer, 0.05 µM CAG or M13-reverse tagged primer (CAG or M13-reverse + primer), 0.45 µM dye labeled tag (HEX or FAM + CAG or M13-reverse), 1.5 mM $MgCl_2$, 0.5 mM dNTPs, 0.5 U Ampli*Taq* Gold DNA Polymerase (Applied Biosystems), and 50 ng DNA. Final concentrations for optimizing reactions with

directly labeled primers were 10 mM Tris pH 8.4, 50 mM KCl, 0.5 μM upper directly labeled primer, 0.5 μM lower directly labeled primer, 1.5 mM $MgCl_2$, 0.5 mM dNTPs, 0.5 U Ampli*Taq* Gold DNA Polymerase (Applied Biosystems), and 50 ng DNA. We ran reactions were using one touchdown thermal cycling program (Don et al. 1991), encompassing a 10.5 ^{o}C span of annealing temperatures (range: 60-49.5 ^{o}C).

For tissue samples, cycling parameters were: 21 cycles of 96 ^{o}C for 20 s; highest annealing temperature for 30 s minus 0.5 ^{o}C per annealing cycle; and 72 ^{o}C for 1 min 30 s followed by 14 cycles of 96 ^{o}C for 20 s; 50 ^{o}C, for 30 s; 72 ^{o}C for 1 min 30 s; and a final extension period of 10 min. at 72 ^{o}C. For hair samples, cycling parameters were: 20 cycles of 96 ^{o}C for 20 s; highest annealing temperature for 30 s minus 0.5 ^{o}C per annealing cycle; and 72 ^{o}C for 1 min 30 s followed by 30 cycles of 96 ^{o}C for 20 s; 50 ^{o}C, for 30 s; 72 ^{o}C for 1 min 30 s; and a final extension period of 10 min. at 72 ^{o}C. We checked PCR products for amplification and sized fragments using a 3730*xl* DNA sequencer (Applied Biosystems) with GENESCAN Rox500 fluorescent size standard (PE Applied Biosystems). Results were analyzed using GENEMAPPER software (Applied Biosystems) using the local Southern size-calling method.

Analytical Methods

The three main components of a model of the marker panel selection system include: decision variables and parameters, constraints or restrictions, and an objective function (Taha 1976). The decision variables (x_i, where i=1,…,16) are the identities of loci in a proposed marker panel (x is binary with '1' indicating the locus is in the marker panel and '0' indicating the locus is not in the marker panel). The sum of x_i is the number of loci in the proposed marker

panel. Parameters include the PID_{sib}, allelic dropout (ADO) probability, false allele (FA) probability, and cost per locus.

The cost function includes a fixed per sample cost for the first locus (C_1), and an additional cost (C_2) for each locus (L) in a marker panel after the first locus.

$$Total\ Cost = C_1 + C_2 * (L-1)$$

Allele frequencies of each locus (necessary for probability of identity calculations), observed and expected heterozygosities for each locus, and total exclusionary power were calculated using Cervus 2.0 (Marshall *et al.* 1998). A program to evaluate probability of identity and probability of identity among siblings was developed in Python, version 2.5.2 (Python Software Foundation, http://python.org). Estimates of genetic error from chapter 3 were selected from the most likely model with 16 loci with the inclusion of samples classified as 'no data'. The model with samples classified as 'no data' likely overestimates genetic error in the calibration sample. Therefore, conclusions from the marker panel selection will be conservative in regards to genetic error.

Restrictions to the system include a maximum probability of identity among siblings, maximum allowable mean overall estimates of both types of error, and a maximum number of loci based on a fixed cost for the genetic analysis. The objective function is used to find the optimum solution the model, which is obtained when the 'corresponding values of the decision variables yield the best value of the objective function while satisfying all the constraints' (Taha 1976). Therefore, the model can be summarized as follows.

$$\text{Minimize } x_0 = C_1 + C_2 * \sum_{i=1}^{L-1} x_i ,$$

and subject to:

$$\prod_{i=1}^{L} PID_{sib,i}, \text{ for all } x_i=1,\ i=1,\ldots L \le 0.01$$

$$\frac{\sum_{i=1}^{L} ADO_{median,i}}{\sum_{i=1}^{L} x_i}, \text{ for all } x_i=1,\ i=1,\ldots L \le 0.05$$

$$\frac{\sum_{i=1}^{L} FA_{median,i}}{\sum_{i=1}^{L} x_i}, \text{ for all } x_i=1,\ i=1,\ldots L \le 0.01$$

$$L \le 10,$$

where L is the total number of loci in the marker panel, and ADO_{median} and FA_{median} are the posterior median values of allelic dropout and false allele probabilities, respectively. A more conservative restriction on PID_{sib} could be a marker panel of loci with a value less than 0.004. A marker panel was selected with and without the genetic error restraints. Minimization was accomplished graphically, instead of analytically due to the non-linearity in some of the constraints.

After optimal marker panels were selected, GENEPOP 3.4 (Raymond, Rousset 1995) was used to test for Hardy-Weinberg equilibrium (HWE) and genotypic linkage disequilibrium (LD) with *a posteriori* sequential Bonferroni correction (Rice 1989) was conducted among loci in the marker panels. Linkage disequilibrium refers to when alleles at two distinctive loci occur in gametes more frequently than expected given the known allele frequencies and recombination fraction between the two loci. Evidence of linkage disequilibrium between pairs of loci, violates assumptions of independence among loci and is not optimal in a marker panel. Violations of HWE indicate possible non-random mating, selection, limited population size, random genetic drift, or mutations in the population (Hartl 2000).

Results

Data summary

Nine hair samples (6 M: 3 F) were classified as bad samples, since they positively amplified at less than half of the loci, and were removed from the error and amplification success analysis. The total possible individuals for the genetic error models were 76 (48 M: 28 F). There were 20 errors over 76 bears and 16 loci detected in the genetic analysis (Table 4.1). Three bears accounted for 50% of the errors detected (Bear 32: 2 errors, Bear CM1: 5 errors, Bear 37: 3 errors). Some tissue and hair samples had positive PCR amplification, but were censored from the analysis because there was too much product in the samples for positive allele sizing, so some loci with have less than 76 possible bear individuals (Table 4.2). For probability of identity calculations, only tissue samples from resident bears of central Georgia (n=84, 54M: 30 F) were used.

Probability of identity, genetic error, and amplification success

The average number of alleles per locus over all 16 loci was 3.4 (range 2-5) for the central Georgia populations (n=84 bears). The probability of identity among siblings (Evett, Weir 1998) over all 16 loci for the central Georgia population was 1.85×10^{-4} (Table 4.3). The probability of identity was 1.37×10^{-8}. Median values of the locus specific ADO probabilities have a range from 0.002 to 0.295 (Table 4.4). Median values of the locus specific FA probabilities have a range from 0.001 to 0.033 (Table 4.5).

Selection of marker panel, without including genetic error estimates

Given the original restrictions, the optimal number of loci is 7 (loci: Bear 12Y, Bear 17G, Bear 19Y, Bear 23, Bear 30B, Bear 33B, Bear 35G) without including genetic error estimates (Figure 4.2). With a more conservative restriction of 0.004 for PID_{sib}, the optimal number of loci is 9 (loci: Bear 12Y, Bear 17G, Bear 19Y, Bear 23, Bear 27B, Bear 30B, Bear 33B, Bear 35G, Bear 36) without including genetic error estimates (Figure 4.2). In the central Georgia population, no loci with the first panel with 7 loci, and only one locus, Bear 27B (p=0.000), with the second panel, deviated from HWE following *a posteriori* Bonferroni correction (Rice 1989). Significant LD between 2 loci (Bear 30B and Bear33B, p=0.0) was detected after Bonferroni correction in both marker panels.

Selection of marker panel, including genetic error estimates

Given the original restrictions, the optimal number of loci is 7 (loci: Bear 10Y, Bear 12Y, Bear 17G, Bear 19Y, Bear 23, Bear 33B, Bear 35G) including genetic error estimates (Figures 4.3, and 4.4). With a more conservative restriction of 0.004 for PID_{sib}, the optimal number of loci is 9 (loci: Bear 10Y, Bear 12Y, Bear 17G, Bear 19Y, Bear 23, Bear 30B, Bear 32, Bear 33B, Bear 35G) without including genetic error estimates (Figures 4.3, and 4.4). All potential marker panels had low mean FA probabilities; therefore, FA probability is not a limiting restriction. In the central Georgia population with both marker panels, no loci deviated from HWE following *a posteriori* Bonferroni correction (Rice 1989). No significant LD was detected after Bonferroni correction in the marker panel, but significant LD between 2 loci (Bear 30B and Bear33B, p=0.0) was detected after Bonferroni correction in the second panel with 9 loci.

Comparison of both methods of marker panel selection

Both methods of marker panel selection with and without genetic error selected the optimal number of loci as 7 and 9 for the less restrictive and more conservative restrictions on PID_{sib}. However, the specific loci in the marker panels differed. Locus Bear 10Y was selected when genetic error was included, but locus Bear 30B was selected when genetic error was not included with the less restrictive PID_{sib}. In this case, six out of the seven loci were identical in the optimal marker panel. Loci Bear 10Y and Bear 32 were selected when genetic error was included, but Bear 27B and Bear 36 were selected when genetic error was not included with the more restrictive PID_{sib}. The optimal marker panel differed for two out of the nine loci in this scenario.

Discussion and Conclusions

The techniques of this study provide formal procedures for choosing a marker set in a non-invasive study, with restrictions of cost, genetic error limitations, and the ability to distinguish among individuals. Optimal selection of a marker panel ultimately depends on both project goals and the amount of time and money available for the study. The optimization algorithm was evaluated using pilot study bear tissue samples of known identity, and hair samples from the same individuals. The pilot study serves a dual purpose of both marker panel selection, and to determine if a non-invasive study is feasible with the given amount of heterozygosity in the sampled population. Bears were captured over a wide range of the central Georgia habitat for black bears, so the sample can be considered representative of population allele frequencies.

The three main components of a model of the marker panel selection system were the decision variables and parameters, constraints or restrictions, and an objective function. Restrictions to the system are arguably subjective, since there are no formal guidelines for probability of identity or the level of genetic error acceptable in a study. Lukacs and Burnham (2005) suggest keeping levels of genetic error to less than 5%. Depending on research goals, these restrictions can be modified to be more stringent or less conservative.

Certainly, time costs and more stringent laboratory costs could also be included into the cost function. Modifications of the cost function would determine the maximum number of loci allowed in the marker panel. After all constraints are met, there might be several potential marker panels still in the solution space with the same number of loci, which translates to the same cost. Further restrictions would then need to be applied towards probability of identity and allowable genetic error. Project goals will influence these values and may weight one measure more heavily than the other.

Estimates of genetic error from chapter 3 from the most likely model with 16 loci with the inclusion of samples classified as 'no data' were included. The model including samples as 'no data' likely overestimates genetic error in the calibration sample. Therefore conclusions from the marker panel selection were conservative in regards to genetic error.

Contrary to Waits and Leberg (2000) and Creel et al. (2003), increases in the number of loci did not necessarily increase the average genetic error with the loci in this study. This has implications in marker panel selection. In general, the probability of identity, or the ability to distinguish among individuals, particularly siblings, decreased as more loci were included. The optimal number of loci did not differ with the inclusion or exclusion of genetic error estimates into the objective function. However, there were differences in the compositions of the optimal

marker panels. Since genetic error is a concern in non-invasive studies, a pilot study that includes estimates of genetic error will aid in selection of a marker panel designed to reduce genetic error. The black bear CGP serves as a case study for the general question of how to select a marker panel, given an excess amount of markers available for a study species.

Literature Cited

Apps, C. D., B. N. McLellan, J. G. Woods, and M. F. Proctor. 2004. Estimating grizzly bear distribution and abundance relative to habitat and human influence. Journal of Wildlife Management 68: 138-152.

Banks, S.C., S.D. Hoyle, A. Horsup, P. Sunnucks and A.C. Taylor. 2003. Demographic monitoring of an entire species (the northern hairy-nosed wombat, *Lasiorhinus krefftii*) by genetic analysis of non-invasively collected material. Animal Conservation 6: 101-107.

Bellemain, E., J.E. Swenson, D. Tallmon, S. Brunberg, and P. Taberlet. 2005. Estimating population size of elusive animals with DNA from hunter-collected feces: four methods for brown bears. Conservation Biology 19: 150-161.

Boersen, M. R. 2001. Abundance and density of Louisiana black bears on the Tensas River National Wildlife Refuge. M. S. thesis, University of Tennessee, Knoxville.

Boulanger, J., G. C. White, B. N. McLellan, J. Woods, M. Proctor, and S. Himmer. 2002. A meta-analysis of grizzly bear DNA mark-recapture projects in British Columbia, Canada. Ursus 13: 137-152.

Constable, J. L., M. V. Ashley, J. Goodall, and A. E. Pusey. 2001. Noninvasive paternity assignment in Gombe chimpanzees. Molecular Ecology 10: 1279-1300.

Creel, S., G. Spong, J.L. Sands et al. 2003. Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12: 2003-2009.

Dixon, J.D., M.O. Oli, M.C. Wooten, T.H. Eason, J.W. McCown, and D. Paetkau. 2006. Effectiveness of a regional corridor in connecting two Florida black bear populations. Conservation Biology 20: 155-162.

Don. R.H., P. T. Cox, B. J. Wainwright, K. Baker, and J. S. Mattick. 1991. 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Research 19: 4008.

Evett, IW, and B.S. Weir. 1998. Interpreting DNA Evidence: Statistical genetics for forensic scientists. Sinauer, Sunderland.

Fremont, A. D. and G. J. Steil. 1986. Foot snare live trap. Patent No. 4581843. USA.

Garnier, J. N., M. W. Bruford, and B. Goossens. 2001. Mating system and reproductive skew in the black rhinoceros. Molecular Ecology 10: 2031-2041.

Goossens, B., L.P. Waits, and P. Taberlet. 1998. Plucked hair samples as a source of DNA: reliability of dinuccleotide microsatellite genotyping. Molecular Ecology 7: 1237-1241.

Hartl, Daniel L. 2000. A primer of population genetics. Sinauer Associates, Inc., Massachusetts.

Hoffman, J. I. and W. Amos. 2005. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology 14:599-612.

Kendall, K. C., J. B. Stetz, D. A. Roon, L. P. Waits, J. B. Boulanger, and D. Paetkau. 2008. Grizzly bear density in Glacier National Park, Montana. Journal of Wildlife Management 72: 1693-1705.

Lukacs, P. M. and K.P. Burnham. 2005. Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. Journal of Wildlife Management 68: 439-448.

Marshall, T. C. , J. Slate, L. E. B. Kruuk, and J. M. Pemberton JM. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Molecular Ecology 7: 639–655.

Mills, L. S., J. Citta, K. Lair, M. Schwartz, and D. Talmon. 2000. Estimating animal abundance using noninvasive DNA sampling: promises and pitfalls. Ecological Applications 10: 283-294.

Oka, T. and O. Takenaka. 2001. Wild Gibbons' parentage tested by non-invasive DNA sampling and PCR-amplified polymorphic microsatellites. Primates 42: 67-73.

Paetkau, D. and C. Strobeck. 1994. Microsatellite analysis of genetic variation in black bear populations. Molecular Ecology 3: 489-495.

Paetkau, D. 2004. The optimal number of markers in genetic capture-mark-recapture studies. Journal of Wildlife Management 68: 449-452.

Proctor, M.F., B.N. McLellan, C. Strobeck, and R.M.R. Barclay. 2004. Gender-specific dispersal distances of grizzly bears estimated by genetic analysis. Canadian Journal of Zoology 82: 1108-1118.

Prugh, L. R., C. E. Ritland, S. M. Arthur, and C. J. Krebs. 2005. Monitoring coyote popoulation dynamics by genotyping faeces. Molecular Ecology 14: 1585-1596.

Raymond, M., and F. Rousset. 1995. GENEPOP (version 1.2.): population genetics software for exact tests and ecumenicism. Journal of Heredity 86: 248–249.

Rice, W. R.1989. Analyzing tables of statistical tests. Evolution 43: 223–225.

Rodriguez, S., G. Visedo, and C. Zapata. 2001. Detection of errors in dinucleotide repeat typing by nondenaturing electrophoresis. Electrophoresis 22: 2656-2664.

Sanderlin, J. S., B.C. Faircloth, B. Shamblin, and M. J. Conroy. 2009. Tetranucleotide microsatellite loci from the black bear (*Ursus americanus*). Molecular Ecology Resources 9: 288-291.

Taberlet, P., J. Camarra, S. Griffin, E. Uhres, O. Hanotte, L.P. Waits, C. Dubois-Paganon, T. Burke, and J. Bouvet. 1997. Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Molecular Ecology 6: 869-876.

Taberlet, P., L.P. Waits, G. Luikart. 1999. Noninvasive genetic sampling: look before you leap. Trends in Ecology and Evolution 14: 323-327.

Taha, H. A. 1976. Operations research, an introduction. 2nd edition. Macmillan Publishing Co., Inc, New York.

Waits, L.P. 1999. Molecular genetic applications for bear research. Ursus 11: 253-260.

Waits, J. L. and P. L. Leberg. 2000. Biases associated with population estimation using molecular tagging. Animal Conservation 3:191-199.

Willey, C. H. 1974. Aging black bears from first premolar tooth sections. Journal of Wildlife Management 38: 97-100.

Woods, J.G., D. Paetkau, D. Lewis, B.N. McLellan, M. Proctor, and C. Strobeck. 1999. Genetic tagging of free-ranging black and brown bears. Wildlife Society Bulletin 27: 616-627.

Table 4.1. Errors detected in the genetic analysis. The locus, bear name, and the actual base pair lengths of both alleles for the tissue samples and hair samples of each error detected are listed below.

Locus	Bear	Tissue allele 1	Tissue allele 2	Hair allele 1	Hair allele 2
10Y	RK1	294	298	294	294
12Y	32	256	256	248	256
12Y	33	260	260	252	260
17G	14	185	193	197	197
17G	CM1	193	193	185	185
23	CM1	256	271	271	271
25	32	382	393	382	382
26	31	356	361	356	356
27B	22	183	183	187	187
27B	30	187	187	183	187
27B	35	187	187	183	187
27B	82	191	191	187	191
30B	CM1	443	447	443	443
32	37	196	204	204	204
32	CM1	196	204	204	204
33B	37	272	286	272	272
33B	CM1	272	286	272	272
35G	5	224	224	216	224
35G	8	216	224	224	224
36	37	198	207	207	207

Table 4.2. Summary of data for genetic error analysis. True tissue samples are either homozygous or heterozygous. Observed hair samples are either true, with no alleles true, one allele true, or no data because the hair samples did not amplify.

TRUE		tissue samples			homozygous					heterozygous					
OBSERVED		hair samples			homozygous		heterozygous			heterozygous			homozygous		
Locus	Alleles	Total bears	Total hom	Total het	TRUE	no alleles true	one allele true	no alleles true	no data	TRUE	one allele true	no alleles true	one allele true	no alleles true	no data
bear10Y	7	70	35	35	35	0	0	0	0	34	0	0	1	0	0
bear12Y	4	68	31	37	29	0	2	0	0	36	0	0	0	0	1
bear13	3	70	62	8	60	0	0	0	2	8	0	0	0	0	0
bear17G	4	70	18	52	17	1	0	0	0	50	0	0	0	1	1
bear19Y	4	73	31	42	31	0	0	0	0	42	0	0	0	0	0
bear2	3	73	33	40	31	0	0	0	2	38	0	0	0	0	2
bear23	6	72	23	49	22	0	0	0	1	46	0	0	1	0	2
bear25	6	74	49	25	46	0	0	0	3	24	0	0	1	0	0
bear27B	4	70	56	14	48	1	3	0	4	12	0	0	0	0	2
bear30B	5	69	21	48	20	0	0	0	1	45	0	0	1	0	2
bear32	3	71	40	31	40	0	0	0	0	29	0	0	2	0	0
bear33B	5	74	27	47	26	0	0	0	1	44	0	0	2	0	1
bear35G	3	72	30	42	29	0	1	0	0	41	0	0	1	0	0
bear6	5	74	44	30	42	0	0	0	2	30	0	0	0	0	0
bear26	2	76	60	16	54	0	0	NA	6	11	1	NA	NA	NA	4
bear36	2	76	36	40	35	0	0	NA	1	36	1	NA	NA	NA	3

Table 4.3. Characteristics of 16 primer pairs of 84 bears from central Georgia.

Locus	alleles	number of hair samples amplified with positive amplification	total possible hair samples *	number of tissue samples with positive amplification	hair amplification success	Hobs	Hexp	Size (bp)	P_{HW}	PIDsib (per locus)	PID (per locus)
Bear10Y	5	70	70	84	1.000	0.488	0.454	238-298	0.854	0.612	0.349
Bear12Y	3	67	68	84	0.985	0.571	0.603	248-260	0.218	0.512	0.245
Bear13	3	68	70	84	0.971	0.107	0.124	213-221	0.288	0.881	0.773
Bear17G	4	70	72	81	0.972	0.716	0.672	185-201	0.735	0.462	0.182
Bear19Y	4	76	76	80	1.000	0.600	0.658	356-375	0.218	0.471	0.192
Bear2	2	69	74	81	0.932	0.556	0.488	173-177	0.256	0.603	0.383
Bear23	4	73	76	78	0.961	0.705	0.623	256-276	0.284	0.496	0.223
Bear25	5	71	74	84	0.959	0.333	0.380	382-397	0.237	0.671	0.424
Bear26	2	66	76	84	0.868	0.190	0.263	356-361	0.022	0.764	0.579
Bear27B	3	50	73	62	0.685	0.258	0.539	183-191	0.000	0.562	0.317
Bear30B	4	66	69	84	0.957	0.631	0.612	439-451	0.927	0.500	0.218
Bear32	2	71	71	84	1.000	0.429	0.439	196-204	1.000	0.635	0.413
Bear33B	4	74	76	82	0.974	0.634	0.631	272-302	0.786	0.487	0.203
Bear35G	3	64	73	81	0.877	0.556	0.527	220-228	0.270	0.559	0.282
Bear36	2	72	76	84	0.947	0.548	0.503	198-207	0.509	0.594	0.375
Bear6	4	73	76	81	0.961	0.370	0.340	239-253	0.595	0.697	0.463

Table 4.4. Median values and 95% posterior density intervals of the locus specific ADO probabilities, under the model with 16 loci with informative priors including records that were classified as 'no data'.

Locus (name)	Median	95% posterior density interval	
1 (Bear 10Y)	0.013	0.000	0.058
2 (Bear 12Y)	0.026	0.005	0.073
3 (Bear 13)	0.115	0.042	0.218
4 (Bear 17G)	0.036	0.011	0.082
5 (Bear 19Y)	0.002	0.000	0.026
6 (Bear 2)	0.088	0.043	0.154
7 (Bear 23)	0.067	0.030	0.124
8 (Bear 25)	0.105	0.048	0.183
9 (Bear 27B)	0.211	0.128	0.309
10 (Bear 30B)	0.069	0.031	0.126
11 (Bear 32)	0.029	0.002	0.087
12 (Bear 33B)	0.060	0.024	0.116
13 (Bear 35G)	0.006	0.000	0.041
14 (Bear 6)	0.057	0.020	0.123
15 (Bear 26)	0.295	0.206	0.393
16 (Bear 36)	0.100	0.050	0.170

Table 4.5. Median values and 95% posterior density intervals of the locus specific FA probabilities, under the model with 16 loci with informative priors including records that were classified as 'no data'.

Locus (name)	Median	95% posterior density interval	
1 (Bear 10Y)	0.002	0.000	0.017
2 (Bear 12Y)	0.014	0.003	0.042
3 (Bear 13)	0.002	0.000	0.017
4 (Bear 17G)	0.015	0.003	0.043
5 (Bear 19Y)	0.001	0.000	0.015
6 (Bear 2)	0.002	0.000	0.018
7 (Bear 23)	0.002	0.000	0.017
8 (Bear 25)	0.002	0.000	0.017
9 (Bear 27B)	0.033	0.011	0.075
10 (Bear 30B)	0.002	0.000	0.018
11 (Bear 32)	0.002	0.000	0.021
12 (Bear 33B)	0.002	0.000	0.018
13 (Bear 35G)	0.010	0.001	0.035
14 (Bear 6)	0.001	0.000	0.016
15 (Bear 26)	0.002	0.000	0.021
16 (Bear 36)	0.002	0.000	0.018

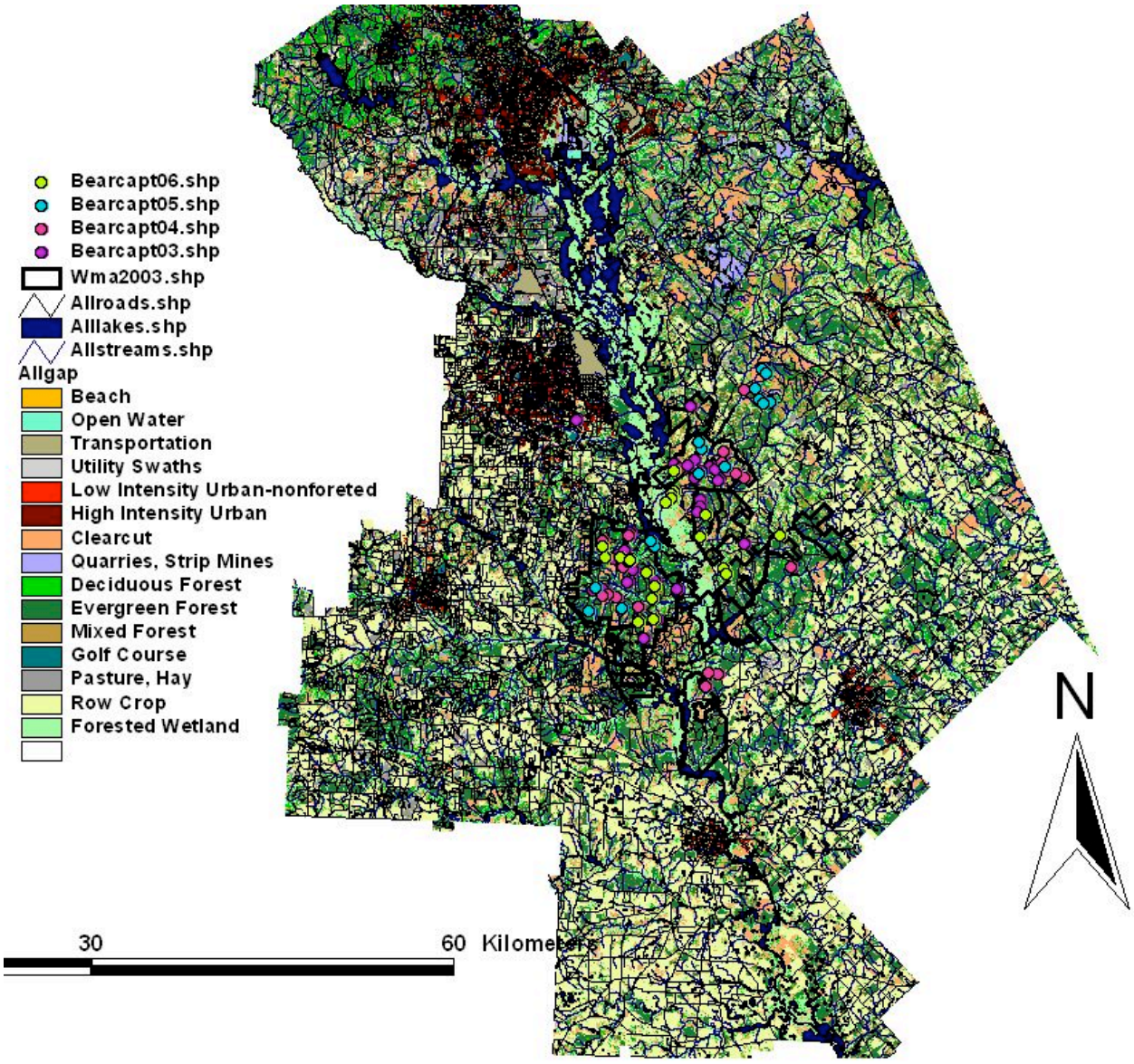


Figure 4.1. Capture coordinates for years 2003-2006 of initial and recaptured bears and WMA boundaries in central Georgia, USA.

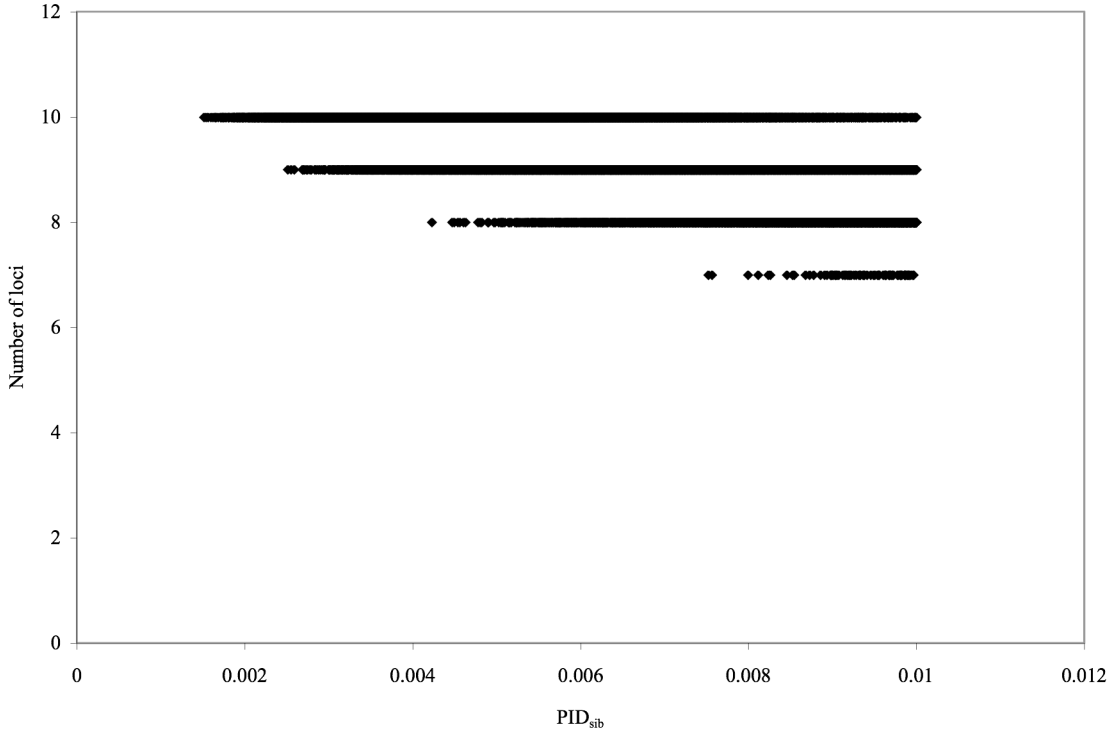


Figure 4.2. Graphical representation of all possible marker panel sets' overall PID_{sib} and number of loci in each panel without including genetic error estimates. The optimal number of loci with the original restrictions was 7. With the additional restriction of PID_{sib}<0.004, the optimal number of loci is 9 for the CGP pilot study.

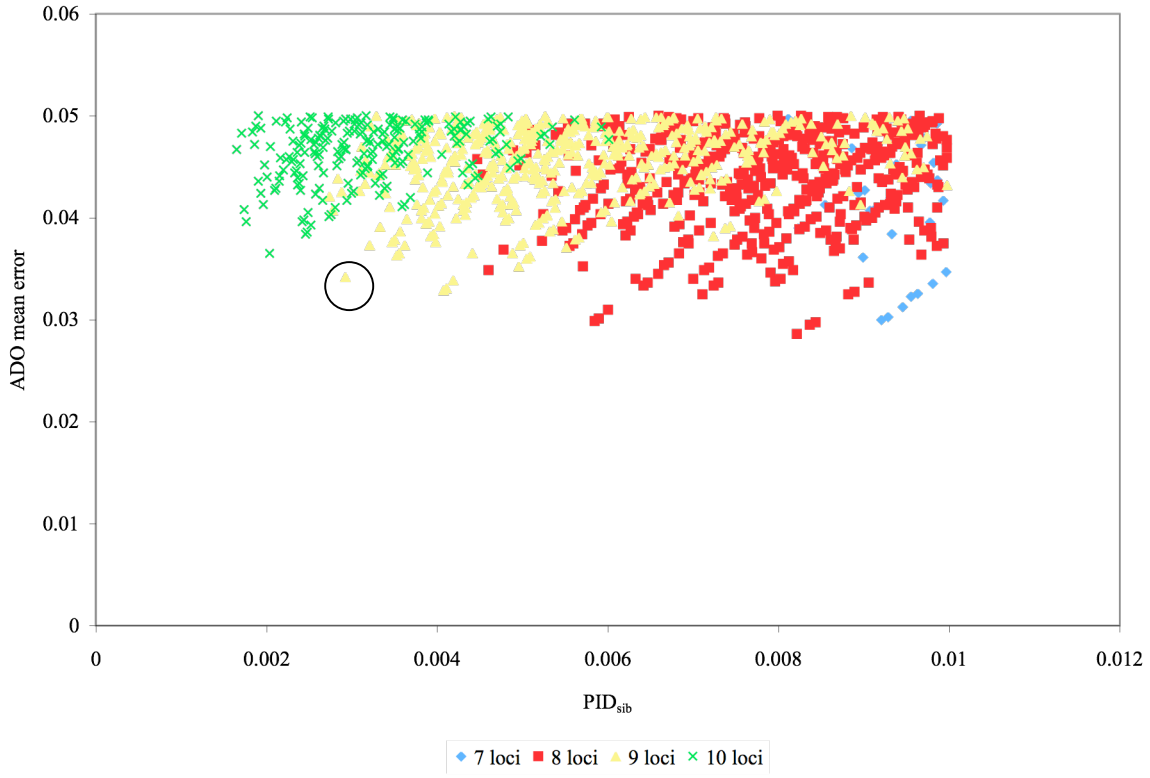


Figure 4.3. Graphical representation of all possible marker panel sets' overall PID_{sib} , mean ADO error, and number of loci in each panel including genetic error estimates. The optimal number of loci with the original restrictions was 7. With the additional restriction of PID_{sib}<0.004, the optimal number of loci is 9 for the CGP pilot study. The circled point would be the optimal marker panel set, with a minimum PID_{sib} and mean ADO error.

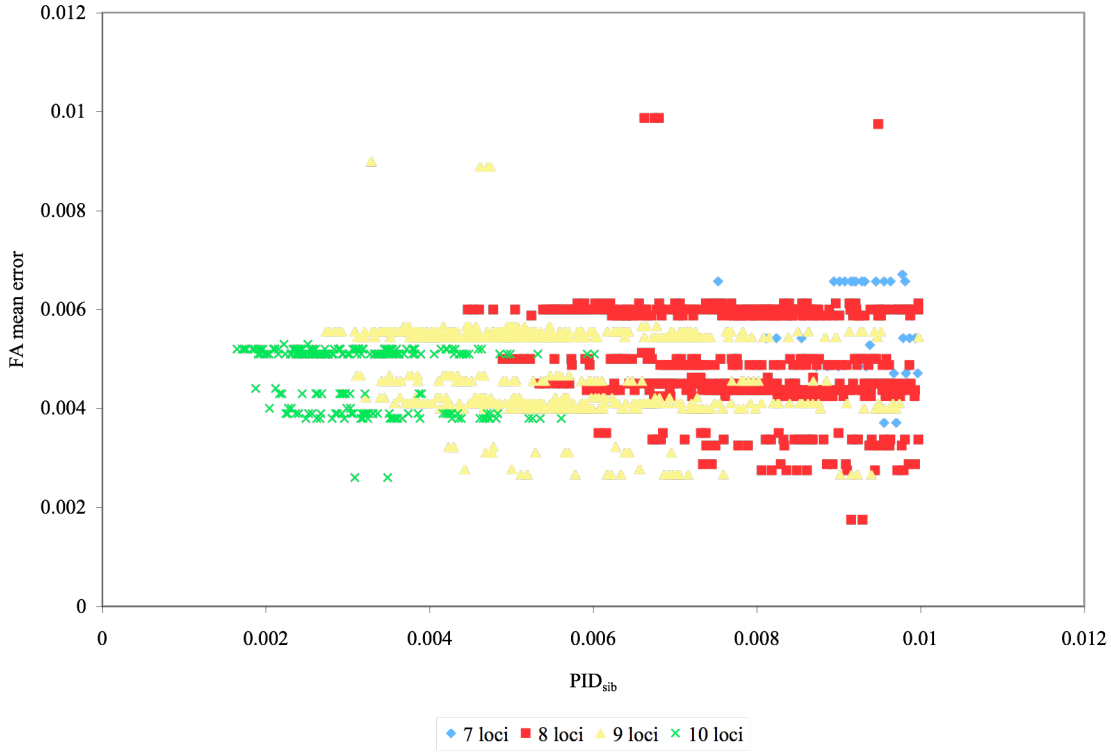


Figure 4.4. Graphical representation of all possible marker panel sets' overall PID_{sib}, mean FA error, and number of loci in each panel including genetic error estimates.

CHAPTER 5

CONCLUSIONS

Advances within the fields of molecular and genetic biology have increased the ability to use genetic analyses in wildlife studies, especially with non-invasive sampling methods. The presence of genetic error, namely allelic dropout and false alleles, decreases accuracy in demographic estimates from non-invasive studies. Genetic errors can occur at various steps within a genetic study, and result from many different sources. Reported error rates vary by allele base pair length, locus, study, and sample type. Although the methods of quantifying genotyping errors vary, studies report higher rates of allelic dropout than false alleles, typically. Several methods of minimizing error exist within categories of laboratory methods, pilot studies and simulations. Often laboratory methods of minimizing error increase the cost of studies substantially. The balance of accepting certain levels of error versus genotyping additional loci has also not been quantified in the literature. Many methods are specific to the study organism and genetic project, and there is minimal generality among studies and laboratories. Statistical methods providing a general framework to assess error would allow comparisons to be made across projects and species.

The main objective of this study was to develop a cost-effective method of estimating misidentification error from non-invasive sampling, which could be incorporated into estimates of abundance for a case study of black bears in the central Georgia population in a hierarchical

Bayesian framework. Calibration samples of known individuals from tissue and hair samples collected from the CGP were used. A calibration sample is more cost-efficient than regenotyping an entire sample set (often >200 samples in non-invasive studies), and still estimates the genetic error in a laboratory with a subset of the entire sample. Tissue samples are higher quality samples and provide a reference point for the hair samples of lower quality.

Model verification was conducted with both simulated data and data from the case study. The Bayesian estimation model described in chapter 2, essentially involves the likelihood of the observed genotypes from the hair samples in the calibration sample given the tissue genotypes of the calibration sample, classified as the true genotypes, and the error parameters. Informative priors, based on several studies that report error rates, were used for the error parameters of allelic dropout and false alleles. Uninformative priors were also explored in the simulation study. The simulation study with the estimation model incorporated 81 different parameter levels involving combinations of population size, calibration sample size, alleles per locus and number of loci. The frequentist properties of Bayesian credible interval percent coverage, BCI length, relative root mean square error, and relative bias were evaluated for the error parameters in the model for all 81 parameter combinations of the simulation study. The sample size of the population/calibration sample influences the frequentist properties slightly. There was a mild sensitivity to prior distributions, and starting values, and should be considered when applying real data to this model.

The final step for model verification included applying the estimation model to the black bear case study. Four subsets of models were used with the real data. Two candidate subsets included all possible loci in the analysis, one including records classified as ‘no data’ and the other without records classified as ‘no data’. Since there is no way of separating out true double

allele dropout events from other laboratory events with this model, including 'no data' records may overestimate the error in the study. Therefore, the true error is likely in between the two models. The last two candidate model sets included 9 loci that were selected for unknown hair samples in the entire study. The most likely model with records classified as 'no data' for both 16 and 9 loci was the model with locus specific rates of allelic dropout and false allele probability. The most likely model without records classified as 'no data' for both 16 and 9 loci was the model with locus specific rates of false allele probability, and an allelic dropout probability that was the same for all loci. The locus specific probabilities indicated that at least half of the loci had very low rates of error (<0.01), with only a few exceptions, and in general, allelic dropout was more common than false allele events; this is consistent with other studies. Based on analyses from chapter 3 and chapter 4 (see below), there is evidence that one of the markers (Bear 27B) should be removed from the marker panel of 9 loci from chapter 3, since both types of error rates were high, and it provided less accuracy per cost (both in money and genetic error cost) than other more suitable markers. There was also evidence of poor samples included in the analysis of chapter 3, since three bears contributed to 50% of the observed errors.

In addition to the estimation model, an algorithm for evaluating the optimal, in terms of cost and accuracy, number of genetic markers necessary for genetic analysis with the CGP, was evaluated. For comparison, an objective function for optimal selection of a marker panel was developed with and without estimates of error probabilities from allelic dropout and false alleles. The objective function included cost with regards to an initial fixed cost for the first locus, plus a second cost for any additional loci. Specific costs were not explicitly stated in the chapter, but the model was based upon actual estimates of genetic laboratory costs and the budget for the CGP case study. Estimates of genetic error were derived from the most likely model of 16 loci

including records classified as ‘no data’ from chapter 3. The probabilities of identity among siblings, or the ability to distinguish the true identities of individuals, were derived from tissue samples from the CGP. The optimal number of loci was the same when error estimates were and were not included. However, the composition of the optimal marker panel differed by two to four loci, depending on how conservative restrictions were set for acceptable rates of allelic dropout. This has implications in optimal marker selection for non-invasive studies, especially when genetic error is a concern.

In conclusion, the main objective of this study of developing a cost-effective method of estimating misidentification error from non-invasive sampling was met by use of a calibration sample of tissue and hair samples of known identities. These estimates could be incorporated into estimates of abundance for a case study of black bears in the CGP. Based on the calibration sample simulation study, the proposed model is valid and has reasonable estimates of genetic error in the case study compared to other methods of error estimation and studies. Finally, an additional concern with non-invasive studies is the reduction in genetic error based on marker panel selection. The objective function and algorithm for selecting an optimal marker panel is discussed. The combination of the calibration sample for genetic error estimation and the use of these estimates in the optimal marker selection process provide an overall cost-efficient method of reducing genetic error in non-invasive studies.

Appendix A. Non-invasive genetic studies that report error rates and/or include rate in estimates.

Year	Paper	Species	Sample type *	Includes error rate in estimates	States error estimate
1996	Taberlet et al.	brown bear	F	0	1
1997	Gagneux et al.	chimpanzees	H	0	1
1997	Paxinos et al.	kit fox, red fox, gray fox, coyote, domestic dog	F	0	0
1997	Taberlet et al.	brown bears	H, F	0	0
1998	Goossens et al.	alphine marmot	H	0	1
1998	Launhardt	Hanuman langurs	F	0	1
1999	Kohn et al.	coyote	F	0	1
1999	Nota & Takenaka	White Leghorn chicken	U	0	1
1999	Woods et al.	black bears, brown bears	H, T	0	0
2000	Bayes et al.	savannah baboon	F	0	1
2000	Bradley et al.	chimpanzee, gorilla	F	0	1
2000	Ernest et al.	mountain lions	F	0	1
2000	Goossens et al.	orang-utans	F	0	1
2000	Mowat & Strobeck	grizzly bears	H	0	0
2000	Sloane et al.	wombats	H	0	1
2000	Valiere and Taberlet	gray wolf	U	0	1
2001	Bradley et al	chimpanzees, gorillas, gibbons	F	0	1
2001	Constable et al.	chimpanzees	H, F	1	1
2001	Garnier et al.	black rhinoceros	F	0	1
2001	Latuhulliere et al.	Barbary macaques	F	0	1
2001	Morin et al.	chimpanzees	F	0	1
2001	Oka and Takenaka	gibbons	H,F	0	0
2001	Parsons	bottlenose dolphins	F	0	1
2002	Lucchini et al.	wolf	F	0	1
2002	Murphy et al.	brown bear	F	0	1
2002	Segelbacher	Capercaillie	Fe	0	1
2003	Creel et al.	wolves	F	1	1
2003	Frantz et al.	Eurasian badger	F	0	1
2004	Banks et al.	northern hairy-nosed wombat	H	0	1
2004	Bonin et al.	brown bear	H	0	1
2004	Lorenzini et al.	brown bear	H	0	1
2004	Piggot et al.	brush-tailed rock-wallaby, spotted-tailed quoll, eastern quoll	F	0	1

* Sample types: H=hair, F=feces, S=sloughed skin, U= urine, T= tissue, Fe= feather, B= blood, C= cloacal swab, Bu= buccal swab, Sp= spraints, Fo= foot mucus

Appendix A. (continued) Non-invasive genetic studies that report error rates and/or include rate in estimates.

Year	Paper	Species	Sample type *	Includes error rate in estimates	States error estimate
2004	Proctor et al.	grizzly bears	T, F	0	0
2004	Schwartz et al.	Canada lynx, bobcats	H, F	0	1
2004	Triant et al.	black bears	H, B, T	0	0
2005	Belant et al.	black bears	H	0	0
2005	Bellemain et al.	brown bears	F, T	0	1
2005	Dobey et al	black bears	H	0	0
2005	Prugh et al.	coyote	F	0	1
2005	Thompson et al.	black bears	H	0	0
2006	Dixon et al.	black bears	H	0	0
2006	Fernando et al.	rhinoceros	F	0	0
2006	Hedmark & Ellegren	wolverine	feces	0	1
2006	Miller	tuatara	C, Bu	0	1
2006	Regnaut et al.	capercaillie	F	0	1
2006	Schwartz et al.	black bears	H	1	0
2006	Smith et al.	kit fox	F	0	1
2007	Adams and Waits	red wolves	F, T	0	1
2007	Arrebdal et al.	Eurasian otter	F	0	1
2007	Ball et al.	woodland caribou, swift fox	F	0	1
2007	Broquet et al.	Alpine newt, green tree frog	Bu	1	1
2007	Dixon et al.	black bears	H, T	0	0
2007	Hedmerk & Ellegren	wolverine	F	0	1
2007	Livia et al.	pine and beech marten	F	0	0
2007	Mukherjee et al.	tigers	F	0	0
2007	Ruell and Crooks	bobcat	H, F	0	1
2007	Valiere et al.	red deer	H, F	0	1
2008	Ferrando et al.	otters	Sp	1	1
2008	He et al.	giant panda	F	0	0
2008	Janecka et al.	snow leopard	F	0	1
2008	Kendall et al.	grizzly bear	H	0	0
2008	Lampa et al.	otter	F (Sp)	0	1
2008	Palmer er al.	mollusc	Fo	0	0
2009	Bowkett et al.	antelope	F	0	0

* Sample types: H=hair, F=feces, S=sloughed skin, U= urine, T= tissue, Fe= feather, B= blood, C= cloacal swab, Bu= buccal swab, Sp= spraints, Fo= foot mucus

Appendix A. (continued) Non-invasive genetic studies that report error rates and/or include rate in estimates.

Year	Paper	Species	Sample type *	Includes error rate in estimates	States error estimate
2009	Hajkova et al.	Eurasian otters	Sp	0	1
2009	Kendall et al.	grizzly bear	H	0	1
2009	Kruckenhauser et al.	brown bears	H, F	0	0
2009	Perez et al.	brown bear	H, F	0	1
2009	Williams et al.	fishers, martens	H	1	0

* Sample types: H=hair, F=feces, S=sloughed skin, U= urine, T= tissue, Fe= feather, B= blood, C= cloacal swab, Bu= buccal swab, Sp= spraints, Fo= foot mucus

Literature Cited

Adams, J. R., and L.P. Waits. 2007. An efficient method for screening faecal DNA genoty and detecting new individuals and hybrids in the red wolf (*Canis rufus*) experimental population area. Conservation Genetics 8: 123-131. pes

Arrendal, J., C. Vila, and M. Bjorklund. 2007. Reliability of noninvasive genetic census of otters compared to field censuses. Conservation Genetics 8: 1097-1107.

Ball, M. C., R. Pither, M. Manseau, J. Clark, S. D. Petersen, S. Kingston, N. Morrill, and P. Wilson. 2007. Conservation Genetics 8: 577-586.

Banks, S.C., S.D. Hoyle, A. Horsup, P. Sunnucks and A.C. Taylor. 2003. Demographic monitoring of an entire species (the northern hairy-nosed wombat, *Lasiorhinus krefftii* by genetic analysis of non-invasively collected material. Animal Conservation 6: 101 107.)

Bayes, M.K., K.L. Smith, S.C. Alberts, J. Altmann, and M.W. Bruford. 2000. Testing the - reliability of microsatellite typing from faecal DNA in the savannah baboon. Conservation Genetics 1: 173-176.

Belant, J., J.F.Van Stappen, D. Paetkau. 2005. American black bear population size and genetic diversity at Apostle Islands National Lakeshore. Ursus 16:85-92.

Bellemain, E., J.E. Swenson, D. Tallmon, S. Brunberg, and P. Taberlet. 2005. Estimating population size of elusive animals with DNA from hunter-collected feces: four methods for brown bears. Conservation Biology 19: 150-161.

Bonin, A., E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. 2004. How to track and assess genotyping errors in population genetics studies. Molecular Ecology 13:3261-3273.

Bowkett, A. E., A. B. Plowman, J. R. Stevens, T. R. B. Davenport, and B. J. van Vuuren. 2009. Genetic testing of dung identification for antelope surveys in the Udzungwa Mountains, Tanzania. Conservation Genetics 10: 251-255.

Bradley, B.J., C. Boesch, and L. Vigilant. 2000. Identification and redesign of human microsatellite markers for genotyping wild chimpanzee (*Pan troglodytes verus*) and gorilla (*Gorilla gorilla gorilla*) DNA from faeces. Conservation Genetics 1: 289-292.

Bradley, B. J., K. E. Chambers, and L. Vigilant. 2001. Accurate DNA-based sex identification of apes using non-invasive samples. Conservation Genetics 2: 179-181.

Broquet, Thomas and Eric Petit. 2004. Quantifying genotyping errors in noninvasive population genetics. Molecular Ecology 13: 3601-3608.

Constable, J. L., M. V. Ashley, J. Goodall, and A. E. Pusey. 2001. Noninvasive paternity assignment in Gombe chimpanzees. Molecular Ecology 10: 1279-1300.

Creel, Scott, G. Spong, J.L. Sands et al. 2003. Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12: 2003-2009.

Dixon, J.D., M.O. Oli, M.C. Wooten, T.H. Eason, J.W. McCown, and D. Paetkau. 2006. Effectiveness of a regional corridor in connecting two Florida black bear populations. Conservation Biology 20: 155-162.

Dobey, S., D.V. Masters, B.K. Scheick, J.D. Clark, M.R. Pelton, and M.E. Sunquist. 2005. Ecology of Florida black bears in the Okefenokee-Osceola Ecosystem. Wildlife Monographs 158: 1-41.

Ernest, H.B., M.C.T. Penedo, B.P. May, M. Syvanen, and W.M. Boyce. 2000. Molecular tracking of mountain lions in the Yosemite Valley region in California: genetic analysis using microsatellites and faecal DNA. Molecular Ecology 9: 433-441.

Fernando, P., G. Polet, N. Foead, L.S. Ng, J. Pastorini, and D. J. Melnick. 2006. Genetic diversity, phylogeny, and conservation of the Javan rhinoceros (*Rhinoceros sondaicus*). Conservation Genetics 7: 439-448.

Ferrando, A., R. Lecis, X. Domingo-Roura, and M. Ponsa. 2008. Genetic diversity and individual identification of reintroduced otters (*Lutra lutra*) in north-eastern Spain by DNA genotyping of spraints. Conservation Genetics 9: 129-139.

Frantz, A.C., L.C. Pope, P.J. Carpenter, T.J. Roper, G.J. Wilson, R.J. Delahay, and T. Burke. 2003. Reliable microsatellite genotyping of the Eurasian badger (Meles meles) using faecal DNA. Molecular Ecology 12: 1649-1661.

Gagneux, P., C. Boesch, and D.S. Woodruff. 1997. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6: 861-868.

Garnier, J. N., M. W. Bruford, and B. Goossens. 2001. Mating system and reproductive skew in the black rhinoceros. Molecular Ecology 10: 2031-2041.

Goossens, B., L.P. Waits, and P. Taberlet. 1998. Plucked hair samples as a source of DNA: reliability of dinuccleotide microsatellite genotyping. Molecular Ecology 7: 1237-1241.

Hajkova, P., B. Zemanova, K. Roche, and B. Hajek. An evaluation of field and noninvasive genetic methods for estimating Eurasian otter population size. Conservation Genetics, published online. DOI 10.1007/s10592-008-9745-4

He, W., L. Lin, F. Shen, W. Zhang, Z. Zhang, E. King, and B. Yue. 2008. Genetic diversities of the giant panda (*Ailuropoda melanoleuca*) in Wanglang and Baoxing Nature Reserves. Conservation Genetics 9: 1541-1546.

Hedmark, E. and H. Ellegren. 2006. A test of the multiplex pre-amplification approach in microsatellite genotyping of wolverine faecal DNA. Conservation Genetics 7: 289-293.

Hedmark, E. and H. Ellegren. 2007. DNA-based monitoring of two newly founded Scandinavian wolverine populations. Conservation Genetics 8: 843-852.

Janecka, J. E., R. Jackson, Z. Yuquang, L. Diqiang, B. Munkhtsog, V. Buckley-Beason, and W. J. Murphy. 2008. Population monitoring of snow leopards using noninvasive collection of scat samples: a pilot study. Animal Conservation 11: 401-411.

Kendall, K. C., J. B. Stetz, D. A. Roon, L. P. Waits, J. B. Boulanger, and D. Paetkau. 2008. Grizzly bear density in Glacier National Park, Montana. Journal of Wildlife Management 72: 1693-1705.

Kendall, K. C., J. B. Stetz, J. Boulanger, A. C. Macleod, D. Paetkau, and G. C. White. 2009. Demography and genetic structure of a recovering grizzly bear population. Journal of Wildlife Management 73: 3-17.

Kohn, M.H., E. York, D.A. Kamradt, G. Haught, R. Sauvajot, and R.K. Wayne. 1999. Estimating population size by genotyping feces. Proceedings of the Royal Society of London, Series B 266: 657-663.

Kruckenhauser, L., G. Rauer, B. Daubi, and E. Haring. Genetic monitoring of a founder population of brown bears (*Ursos arctos*) in central Austria. Conservation Genetics, published online. DOI 10.1007/s10592-008-9654-6

Lampa, S., B. Gruber, K. Henle, and M. Hoehn. 2008. An optimization approach to increase DNA amplification success of otter faeces. Conservation Genetics 9: 201-210.

Lathuilliere, M., N. Menard. A. Gautier-Hion, and B. and Crouau-Roy. 2001. Testing the reliability of noninvasive genetic sampling by comparing analyses of blood and fecal samples in Barbary Macaques (*Macaca sylvanus*). American Journal of Primatology 55: 151-158.

Launhardt, K., C. Epplen, J.T. Epplen, and P. Winkler. 1998. Amplification of microsatellites adapted from human systems in faecal DNA of wild Hanuman langurs (*Presbytis entellus*). Electrophoresis 19: 1356-1361.

Livia, L., V. Francesca, P. Antonella, P. Fausto, and R. Bernardino. 2007. A PCR-RFLP method on faecal samples to distinguish *Martes martes*, *Martes foina*, *Mustela putorius* and *Vulpes vulpes*. Conservation Genetics 8: 7575-759.

Lorenzini, R., M. Posillico, S. Lovari, and A. Petrella. 2004. Non-invasive genotyping of the endangered Apennine brown bear: a case study not to let one's hair down. Animal Conservation 7: 199-209.

Lucchini, V., E. Fabbri, F. Marucco, S. Ricci, L. Boitani, and E. Randi. 2002. Noninvasive molecular tracking of colonizing wolf (Canis lupus) packs in the western Italian Alps. Molecular Ecology 11: 857-868.

Miller, H. C. 2006. Cloacal and buccal swabs are a reliable source of DNA for microsatellite genotyping of reptiles. Conservation Genetics 7: 1001-1003.

Morin. P. A., K. E. Chambers, C. Boesch and L. Vigilant. 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). Molecular Ecology 10: 1835-1844.

Mowat, G. and C. Strobeck. 2000. Estimating population size of grizzly bears using hair capture, DNA profiling, and mark-recapture analysis. Journal of Wildlife Management 64: 183-193.

Mukherjee, N., S. Mondol, A. Andheria, and U. Ramakrishnan. 2007. Rapid multiplex PCR based species identification of wild tigers using non-invasive samples. Conservation Genetics 8: 1465-1470.

Murphy, M. A., L. P. Waits, K. C. Kendall, S. K. Wasser, J. A. Higbee, and R. Bogden. 2002. An evaluation of long-term preservation methods for brown bear (*Ursus arctos*) faecal DNA samples. Conservation Genetics 3 : 435-440.

Oka, T. and O. Takenaka. 2001. Wild Gibbons' parentage tested by non-invasive DNA sampling and PCR-amplified polymorphic microsatellites. Primates 42: 67-73.

Palmer, A. N. S., C. A. Styan, and D. C. A. Shearman. 2008. Foot mucus is a good source for non-destructive genetic sampling in Polyplacophora. Conservation Genetics 9: 229-231.

Parsons, K.M. 2001. Reliable microsatellite genotyping of dolphin DNA from faeces. Molecular Ecology 1: 341-344.

Paxinos, E., C. Mcintosh, K. Ralls, and R. Fleischer. 1997. A noninvasive method for distinguishing among canid species: amplification and enzyme restriction of DNA from dung. Molecular Ecology 6: 483- 486.

Perez, T., F. Vazquez, J. Naves, A. Fernandez, A. Corao, J. Albornoz, and A. Dominguez. 2009.

Non-invasive genetic study of the endangered Cantabrian brown bear (*Ursos arctos*). Conservation Genetics 10: 291-301.

Piggott, M. P., E. Gellemain, P. Taberlet, and A.C. Taylor. 2004. A multiplex pre-amplification method that significantly improves microsatellite amplification and error rates for faecal DNA in limiting conditions. Conservation Genetics 5: 417-420.

Proctor, M.F., B.N. McLellan, C. Strobeck, and R.M.R. Barclay. 2004. Gender-specific dispersal distances of grizzly bears estimated by genetic analysis. Canadian Journal of Zoology 82: 1108-1118.

Prugh, L. R., C. E. Ritland, S. M. Arthur, and C. J. Krebs. 2005. Monitoring coyote popoulation dynamics by genotyping faeces. Molecular Ecology 14: 1585-1596.

Regnaut, S., P. Christe, M. Chapuisat, and L. Fumagalli. 2006. Genotyping faeces reveals facultative kin association on capercaillie's leks. Conservation Genetics 7: 665-674.

Ruell, E. W. and K. R. Crooks. 2007. Evaluation of noninvasive genetic sampling methods for felid and canid populations. Journal of Wildlife Management 71: 1690-1694.

Schwartz, M. K., K. L. Pilgrim, K. S. McKelvey, E. L. Lindquist, J. J. Claar, S. Loch, and L. F. Ruggiero. 2004. Hybridization between Canada lynx and bobcats: genetic results and management implications. Conservation Genetics 5: 349-355.

Schwartz, M.K., S.A. Cushman, K.S. McKelvey, J. Hayden, and C. Engkjer. 2006. Detecting genotyping errors and describing American black bear movement in northern Idaho. Ursus 17:138-148.

Segelbacher, G. 2002. Noninvasive genetic analysis in birds: testing reliability of feather samples. Molecular Ecology Notes 2: 367-369.

Sloane, M.A., P. Sunnucks, D. Alpers, B. Beheregaray, and A.C. Taylor. 2000. Highly reliable

genetic identification of individual northern hairy-nosed wombats from single remotely collected hairs: a feasible censusing method. Molecular Ecology 9: 123-124.

Smith, D. A., K. Ralls, A. Hurt, B. Adams, M. Parker, and J. E. Maldonado. 2006. Assessing reliability of microsatellite genotypes from kit fox faecal samples using genetic and GIS analyses. Molecular Ecology 15: 387-406.

Taberlet, P. ,S. Griffen, B. Goossens, S. Questiau, V. Manceau, N. Escaravage, L. P. Waits and J. Bouvet. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research 24: 3189-3194.

Taberlet, P., J. Camarra, S. Griffin, E. Uhres, O. Hanotte, L.P. Waits, C. Dubois-Paganon, T. Burke, and J. Bouvet. 1997. Noninvasive genetic tracking of the endangered Pyrenean brown bear population. Molecular Ecology 6: 869-876.

Thompson, L.M., F.T. van Manen, and T. L. King. 2005. Geostatistical analysis of allele presence patterns among American black bears in eastern North Carolina. Ursus 16: 59-69.

Triant, D.A., R.M. Pace and M. Stine. 2004. Abundance, genetic diversity and conservation of Louisiana black bears (Ursus americanus luteolus) as detected through noninvasive sampling. Conservation Genetics 5: 647-659.

Valiere, N. and P. Taberlet. 2000. Urine collected in the field as a source of DNA for species and individual identification. Molecular Ecology 9: 2149-2154.

Valiere, N. , C. Bonenfant, C. Toigo, G. Luikart, J-M. Gaillard, and F. Klein. 2007. Importance of a pilot study for non-invasive genetic sampling: genotyping errors and population size estimation in red deer. Conservation Genetics 8: 69-78.

Williams, B. W., D. R. Etter, D. W. Linden, K. F Millenbah, S. R. Winterstein, and K. T.

Scribner. 2009. Noninvasive hair sampling and genetic tagging of co-disctributed fishers and American martens. Journal of Wildlife Management 73: 26-34.

Woods, J.G., D. Paetkau, D. Lewis, B.N. McLellan, M. Proctor, and C. Strobeck. 1999. Genetic tagging of free-ranging black and brown bears. Wildlife Society Bulletin 27: 616-627.

Appendix B. Reported genotyping error rates in the literature used in chapter 2 and chapter 3 (studies with * indicated additional references used in the prior distributions of chapter 3). Sample types include: H (hair), F (faecal), S (sloughed skin), Fe (feather), B (blood), Bu (buccal swab), Sp (Spraints), U (urine), C (cloacal swab), and T (tissue). Error types include: AD (allele dropout), FA (false allele), S (slippage), SC (scoring error), H (human error), and CONT (contamination). See text of chapter 2 for ADO_u, ADO_1, ADO_2, FA_u, FA_1, FA_2. See Creel et al. (2003) for 'allele specific' calculations. See Palsbøll et al. (1997) for 'locus specific' calculation. R_{PCR}= number of genotype errors (FA or ADO) out of total number of repeat PCRs (same DNA extract of sample, different PCR and genotyping). R_{ext}= number of genotype errors (FA or ADO) out of total repeat DNA extractions (different DNA extracts, PCR, and genotyping of sample). R_{gen}= number of genotype errors (FA or ADO) out of total repeat genotyping (same DNA extract and PCR run of sample, different genotyping). R_{score}= number of genotype errors (FA or ADO) out of total repeat scoring session (same DNA extracts, PCR run, and genotyping of sample, different individuals or sessions of scoring genotype).

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
red wolf	Adams and Waits 2007*	105 F, 15 B	F, B	7	0.008 0.259	FA_2 ADO_u	AD	FA
Eurasian otter	Arrendal et al. 2007*	46 20	F T	8	0.077	ADO_u	AD	FA
woodland caribou	Ball et al. 2007*	60	F	6	0	ADO unknown		
Swift fox		50	F	1	0	FA		

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
savannah baboon	Bayes et al. 2000	12	F B	8	0.08 0.01	ADO_2	AD	FA
wombats	Banks et al. 2003	406 H	H	10	0.002	R_{PCR}		FA,AD, CONT
northern hairy-nosed wombat	Banks et al. 2003	44	H	10	0.002	ADO+FA		
brown bears	Bellemain et al. 2005	1904	F	6	0.02	R_{PCR}		AD,FA, CONT
frog	Bonin et al. 2004	192	T	5	0.02	R_{PCR}, R_{score}	SC	H, difference in peak intensities
		34	T	8	0.008	R_{gen}		
brown bear		47	F	6	0.02	R_{ext}	AD, H	FA
		96			0.012	R_{PCR}		
chimpanzees	Bradley et al. 2000	48	F	9				
gorillas		26	F	10	<0.01	ADO or FA		
chiimpanzees	Bradley et al. 2001*	84	F	1	0.02	ADO_u		
					0.012	FA_u	AD	FA
gorillas		74	F	1	0.03	ADO_u		
					0.011	FA_u	AD	FA
Alpine newt	Broquet et al. 2007*	12	Bu	7	0	ADO_u		
					0	FA_u		
green tree frog		12	Bu	6	0.021	ADO_u		
					0	FA_u	AD	

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
elephants	Buchan et al. 2005		F	11	0.19	ADO_u	AD	S, CONT, FA
baboons			F	14	0.21	ADO_u	AD	CONT (human DNA, other), S, FA
chimpanzees	Constable et al. 2001*	14	F	16	0.3	ADO among samples	AD	
wolves	Creel et al. 2003	227	F	13	0.111	allelle specific	AD	
					0.056	allelle specific	FA	
otters	Dallas et al. 2003	65 F, 3 B	F	9	0.021	ADO_2	ADO	FA
					0.012	FA_u		
elephants	Eggert et al. 2003	124 F	F	7	0.0039	cummulative R_{ext}	FA	ADO
mountain lions	Ernest et al. 2000	15	F	12	0.08	ADO_1	AD	
		62	T	12	>0.01	ADO_1		
Asian elephants	Fernando et al. 2003	20	F,B	6	0.019	ADO_1, FA_1	FA	AD, nondetection of alleles

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
					0.158	ADO_u among loci, GIMLET		
otters	Ferrando et al. 2008*	39	Sp	10	0.14	ADO_u among samples, GIMLET		
					<0.02	FA_U across loci and samples	AD	FA
reindeer	Flagstad et al. 1999	10	F	6	0.02	ADO_1	AD	none
domestic sheep		42			0.015			
Eurasian badger	Frantz et al. 2003	33	F	7	0.08	FA_2	AD	FA
					0.27	ADO_u		
chimpanzees	Gagneux et al. 1997*	791 H, 18 indiv.	H	11	0.31	ADO		
					0.056	F (CONT)	AD	FA
black rhinoceros	Garnier et al. 2001*	19	F	10	0.3	ADO		
					0	FA	AD	FA
alpine marmot	Goossens et al. 1998	50 (1 H)	H	1	0.14	ADO_1, FA_1	AD	FA
		50 (3 H)			0.0486	ADO_u, FA_u		
		50 (10 H)			0.0029			
orang-utans	Goossens et al. 2000	16 indiv.	F	1	0.0295	FA_u	AD	FA
					0.042	ADO_u		

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
Seychelles warblers	Hadfield et al. 2006	319	B	14	0.029	Bayesian analysis	AD	SC
					0.01	Bayesian analysis	SC	
Eurasian otter	Hajkova et al. 2009*	262	Sp	10	0.18	$ADO_{u\ (among\ loci)}$, GEMINI		
wolverine	Hedmark & Ellegren 2006*	48	F	18	0.029	FA_u	AD	FA
					0.024	ADO		
					<0.01	FA	AD	FA
wolverine	Hedmark & Ellegen 2007*	249	F	10	0.13	ADO among loci (first marker set)		
					<0.01	FA_u	AD	FA
			F	10	0.14	ADO among loci (second marker set)		
					<0.01	FA_u	AD	FA
Antarctic fur seals	Hoffman and Amos 2005	1763	T	9	0.0013-0.0074	R_{gen}	SC	data input, AD, unkown, SC
European hares	Huber et al. 2003	11	B,F	4	0.138	FA_u	FA	AD
					0.023	ADO_u		
red deer		11	B,F	3	0.164	FA_u	FA	AD
					0.026	ADO_u		

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
		16	F		0.14	$ADO_{u\ (FCA\ primer)}$		
snow leopard	Jenecka et al. 2008*			6	0.043	$ADO_{u\ (PUN\ primer)}$		
		25	F		0.031	FA_u (FCA primer)		
		38			0.008	FA_u (PUN primer)	AD	FA
coyote	Kohn et al. 1999	59 F	F	3	0.05	ADO_1,FA_1	AD,FA	
					0.27	ADO_u (optimized single locus protocol)		
					0.38	ADO_u (single locus protocol)		
otter	Lampa et al. 2008*	unknown	F, Sp	6	0.29	ADO_u (multiplex protocol)		
					0	FA_u (optimized single locus protocol)		
					0.02	FA_u (multiplex protocol)	AD	FA

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
Barbary macaques	Lathuilliere et al. 2001	11	B,F	3	0.153	FA_u	FA	AD
					0.03	ADO_u		
Hanuman langurs	Launhardt et al. 1998*	178	F	5	0.05	ADO_u		
		12	T		0	FA_u	AD	
brown bear	Lorenzini et al. 2004*	44	H	12	0.06	ADO_1	AD	
wolf	Lucchini et al. 2002	40 F	F	6	0.003	FA_1	AD	FA
					0.18	ADO_u		
tuatara	Miller 2006*	6	Bu, C	2 mito., 3 microsat.	0	ADO_u		
chimpanzees	Morin et al. 2001	90 F	F	9	0.24	ADO_u	AD	none
brown bear	Murphy et al. 2002*	50	F	1	0.07	ADO		
					0.18	FA (multiple alleles)		
					0.06	FA	FA	AD
brown bear	Murphy et al. 2007*	53	F	2	0.0928	FA_u		
					0.0515	ADO_u		
					0.2474	CONT	FA	FA, CONT

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
White Leghorn chickens	Nota & Takenaka 1999*	12	U	1	0.083	ADO per sample	AD	
Bear	Paetkau 2003	6638	H	5-7	0.018 0.016	R_{score}	SC AD,FA	AD,FA
bottlenose dolphins	Parsons et al. 2001	12	B,T,F	3	0.0097	ADO_u, FA_u	FA	none
humpback whale	Palsbøll et al. 1997	2368 S	S	6	0.0011	locus specific		
brown bear	Perez et al. 2009*	92	F	18	0.0375 0.0146	ADO FA	AD	FA
		41	H		0.0143 0.0113	ADO FA	AD	FA
brush tailed wallaby		10	F	6	0.92 (0) 0.0372 (0.02)	ADO (after multiplex) FA (after multiplex)	AD	FA
spotted-tailed quoll	Piggot et al. 2004*	10	F	6	0.0277 (0) 0.0759 (0)	ADO (after multiplex) FA (after multiplex)	FA	AD
eastern quoll		10	F	6	0.41 (0.21) 0.0256 (0)	ADO (after multiplex) FA (after multiplex)	AD	FA
coyote	Prugh et al. 2005*	834	F	6	0.045	ADO_2		
		22	T		0.018	FA_u	AD	FA

Appendix B. (continued)

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
capercaillie	Regnaut et al. 2006*	57	F	11	0.21	ADO_U, GIMLET		
					0.03	FA_U, GIMLET	AD	FA
bobcats	Ruell & Crooks 2007*	25	F	5	0.094	ADO_u		
					0.012	FA_u	AD	FA
		31	H	5	0.1	ADO_u		
					0.023	FA_u	AD	FA
wolf	Scandura et al. 2006	141 F, 87 H, 24 B, 17 T	H,F,B, T	10	0.029	ADO_u		
					0.016	FA_u	AD	FA
black bear	Schwartz et al. 2006	245 H	H	9	120 errors	program DROPOUT with DCH test	SC	AD, FA, SC
African Indigo birds	Sefc et al. 2003	128 (1 fe)	Fe	9	0.192	ADO_u		
					0.049	FA_u	AD	FA
		128 (2 fe)			0.121	ADO_u		
					0.01	FA_u		
capercailie	Segelbacher 2002	20	Fe	10	0.011	ADO_l	AD	

Appendix B. (continued).

Species	Reference	Sample Size	Sample type	Number of Loci	Reported error rate	Method(s) of calculating error	Most common error	Other error
wombats	Sloane et al. 2000	284	H	12	0.003	ADO_u, FA_u		
savannah baboon	Smith et al. 2000	87 F	F,B	5	0.48	ADO_2	AD	none
fox	Soulsbury et al. 2007*	30	T	10	0.21-0.57	ADO_u	AD	
red deer	Valiere et al. 2007*	40	T	9	0.2	ADO_u, GIMLET		
		12	F		0.02	FA, GIMLET	AD	FA
wolf	Valiere and Taberlet 2000*	5	U	3	0.0420	ADO_u FA	AD	FA
				2	0.7460	ADO FA	AD	FA

Literature Cited

Adams, J. R., and L.P. Waits. 2007. An efficient method for screening faecal DNA genotypes and detecting new individuals and hybrids in the red wolf (*Canis rufus*) experimental population area. Conservation Genetics 8: 123-131.

Arrendal, J., C. Vila, and M. Bjorklund. 2007. Reliability of noninvasive genetic census of otters compared to field censuses. Conservation Genetics 8: 1097-1107.

Ball, M. C., R. Pither, M. Manseau, J. Clark, S. D. Petersen, S. Kingston, N. Morrill, and P. Wilson. 2007. Conservation Genetics 8: 577-586.

Banks, S.C., S.D. Hoyle, A. Horsup, P. Sunnucks and A.C. Taylor. 2003. Demographic monitoring of an entire species (the northern hairy-nosed wombat, *Lasiorhinus krefftii* by genetic analysis of non-invasively collected material. Animal Conservation 6: 101 107.)

Bayes, M.K., K.L. Smith, S.C. Alberts, J. Altmann, and M.W. Bruford. 2000. Testing the - reliability of microsatellite typing from faecal DNA in the savannah baboon. Conservation Genetics 1: 173-176.

Bellemain, E., J.E. Swenson, D. Tallmon, S. Brunberg, and P. Taberlet. 2005. Estimating population size of elusive animals with DNA from hunter-collected feces: four methods for brown bears. Conservation Biology 19: 150-161.

Bonin, A., E. Bellemain, P. Bronken Eidesen, F. Pompanon, C. Brochmann, and P. Taberlet. 2004. How to track and assess genotyping errors in population genetics studies. Molecular Ecology 13:3261-3273.

Bradley, B.J., C. Boesch, and L. Vigilant. 2000. Identification and redesign of human microsatellite markers for genotyping wild chimpanzee (*Pan troglodytes verus*) and gorilla (*Gorilla gorilla gorilla*) DNA from faeces. Conservation Genetics 1: 289-292.

Bradley, B. J., K. E. Chambers, and L. Vigilant. 2001. Accurate DNA-based sex identification of apes using non-invasive samples. Conservation Genetics 2: 179-181.

Broquet, T., L. B. Braendli, G. Emaresi, and L. Fumagalli. 2007. Buccal swabs allow efficient and reliable microsatellite genotyping in amphibians. Conservation Genetics 8: 509-511.

Buchan, J.C., E. A. Archie, R.C. Van Horn, C.J. Moss, and S.C. Alberts. 2005. Locus effects and sources of error in noninvasive genotyping. Molecular Ecology Notes 5: 680-683.

Constable, J. L., M. V. Ashley, J. Goodall, and A. E. Pusey. 2001. Noninvasive paternity assignment in Gombe chimpanzees. Molecular Ecology 10: 1279-1300.

Creel, Scott, G. Spong, J.L. Sands et al. 2003. Population size estimation in Yellowstone wolves with error-prone noninvasive microsatellite genotypes. Molecular Ecology 12: 2003-2009.

Dallas, J. F., K. E. Coxon, T. Sykes, P.R. F. Chanin, F. Marchall, D. N. Carss, P. J. Bacon, S. B. Piertney, and P. A. Racey. 2003. Similar estimates of population genetic composition and sex ratio derived from carcasses and faeces of Eurasian otter *Lutra lutra*. Molecular Ecology 12: 275-283.

Eggert, L.S. , J.A. Eggert. And D.S. Woodruff. 2003. Estimating population sizes for elusive animals: the forest elephants of Kakum National Park, Ghana. Molecular Ecology 12: 1389-1402.

Ernest, H.B., M.C.T. Penedo, B.P. May, M. Syvanen, and W.M. Boyce. 2000. Molecular tracking of mountain lions in the Yosemite Valley region in California: genetic analysis using microsatellites and faecal DNA. Molecular Ecology 9: 433-441.

Fernando, P., T.N.C. Vidya, C. Rajapakse, A. Dangolla, and D.J. Melnick. 2003. Reliable noninvasive genotyping: fantasy or reality? Journal of Heredity 94: 115-123.

Ferrando, A., R. Lecis, X. Domingo-Roura, and M. Ponsa. 2008. Genetic diversity and individual identification of reintroduced otters (*Lutra lutra*) in north-eastern Spain by DNA genotyping of spraints. Conservation Genetics 9: 129-139.

Flagstad, Ø., K. Røed, J.E. Stacy, and K.S. Jakobsen. 1999. Reliable noninvasive genotyping based on excremental PCR of nuclear DNA purified with a magnetic bear protocol. Molecular Ecology 8:879-883.

Frantz, A.C., L.C. Pope, P.J. Carpenter, T.J. Roper, G.J. Wilson, R.J. Delahay, and T. Burke. 2003. Reliable microsatellite genotyping of the Eurasian badger (Meles meles) using faecal DNA. Molecular Ecology 12: 1649-1661.

Gagneux, P., C. Boesch, and D.S. Woodruff. 1997. Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6: 861-868.

Goossens, B., L.P. Waits, and P. Taberlet. 1998. Plucked hair samples as a source of DNA: reliability of dinuccleotide microsatellite genotyping. Molecular Ecology 7: 1237-1241.

Goosens, B., L. Chikhi, S.S. Utami, J. de Ruiter, and M.W. Bruford. 2000. A multi-samples, multi-extracts approach for microsatellite analysis of faecal samples in an arboreal ape. Conservation Genetics 1: 157-162.

Hadfield, J.D., D.S. Richardson, and T. Burke. 2006. Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. Molecular Ecology 15: 3715-3730.

Hajkova, P., B. Zemanova, K. Roche, and B. Hajek. An evaluation of field and noninvasive genetic methods for estimating Eurasian otter population size. Conservation Genetics, published online. DOI 10.1007/s10592-008-9745-4

Hedmark, E. and H. Ellegren. 2006. A test of the multiplex pre-amplification approach in microsatellite genotyping of wolverine faecal DNA. Conservation Genetics 7: 289-293.

Hedmark, E. and H. Ellegren. 2007. DNA-based monitoring of two newly founded Scandinavian wolverine populations. Conservation Genetics 8: 843-852.

Hoffman, J. I. and W. Amos. 2005. Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology 14:599-612.

Huber, S., U. Bruns, and W. Arnold. 2003. Genotyping herbivore feces facilitating their further analyses. Wildlife Society Bulletin 31: 692-697.

Janecka, J. E., R. Jackson, Z. Yuquang, L. Diqiang, B. Munkhtsog, V. Buckley-Beason, and W. J. Murphy. 2008. Population monitoring of snow leopards using noninvasive collection of scat samples: a pilot study. Animal Conservation 11: 401-411.

Kohn, M.H., E. York, D.A. Kamradt, G. Haught, R. Sauvajot, and R.K. Wayne. 1999. Estimating population size by genotyping feces. Proceedings of the Royal Society of London, Series B 266: 657-663.

Lampa, S., B. Gruber, K. Henle, and M. Hoehn. 2008. An optimization approach to increase DNA amplification success of otter faeces. Conservation Genetics 9: 201-210.

Lathuilliere, M., N. Menard. A. Gautier-Hion, and B. and Crouau-Roy. 2001. Testing the reliability of noninvasive genetic sampling by comparing analyses of blood and fecal samples in Barbary Macaques (*Macaca sylvanus*). American Journal of Primatology 55: 151-158.

Launhardt, K., C. Epplen, J.T. Epplen, and P. Winkler. 1998. Amplification of microsatellites adapted from human systems in faecal DNA of wild Hanuman langurs (*Presbytis entellus*). Electrophoresis 19: 1356-1361.

Lorenzini, R., M. Posillico, S. Lovari, and A. Petrella. 2004. Non-invasive genotyping of the endangered Apennine brown bear: a case study not to let one's hair down. Animal Conservation 7: 199-209.

Lucchini, V., E. Fabbri, F. Marucco, S. Ricci, L. Boitani, and E. Randi. 2002. Noninvasive molecular tracking of colonizing wolf (Canis lupus) packs in the western Italian Alps. Molecular Ecology 11: 857-868.

Miller, H. C. 2006. Cloacal and buccal swabs are a reliable source of DNA for microsatellite genotyping of reptiles. Conservation Genetics 7: 1001-1003.

Morin. P. A., K. E. Chambers, C. Boesch and L. Vigilant. 2001. Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). Molecular Ecology 10: 1835-1844.

Murphy, M. A., L. P. Waits, K. C. Kendall, S. K. Wasser, J. A. Higbee, and R. Bogden. 2002. An evaluation of long-term preservation methods for brown bear (*Ursus arctos*) faecal DNA samples. Conservation Genetics 3 : 435-440.

Murphy, M.A. , K.C. Kendall, A. Robinson, and L. P. Waits. 2007. The impact of time and field conditions on brown bear (*Ursus arctos*) faecal DNA amplification. Conservation Genetics 8: 1219-1224.

Paetkau, D. 2003. An empirical exploration of data quality in DNA-based population inventories. Molecular Ecology 12:1375-1387.

Parsons, K.M. 2001. Reliable microsatellite genotyping of dolphin DNA from faeces. Molecular Ecology 1: 341-344.

Palsbøll, P.J., J. Allen, M. Bérubé, P.J. Clapham, T.P. Feddersen, P.S. Hammond, R.R. Hudson, H. Jørgensen, S. Katona, A.H. Larsen, F. Larsen, J. Lien, D.K. Mattila, H. Sigurjónsson, R. Sears, T. Smith, R. Sponer, P. Stevick, and N. Øien. 1997. Genetic tagging of humpback whales. Nature 388: 767-769.

Perez, T., F. Vazquez, J. Naves, A. Fernandez, A. Corao, J. Albornoz, and A. Dominguez. 2009. Non-invasive genetic study of the endangered Cantabrian brown bear (*Ursos arctos*). Conservation Genetics 10: 291-301.

Piggott, M. P., E. Gellemain, P. Taberlet, and A.C. Taylor. 2004. A multiplex pre-amplification method that significantly improves microsatellite amplification and error rates for faecal DNA in limiting conditions. Conservation Genetics 5: 417-420.

Prugh, L. R., C. E. Ritland, S. M. Arthur, and C. J. Krebs. 2005. Monitoring coyote popoulation dynamics by genotyping faeces. Molecular Ecology 14: 1585-1596.

Regnaut, S., P. Christe, M. Chapuisat, and L. Fumagalli. 2006. Genotyping faeces reveals facultative kin association on capercaillie's leks. Conservation Genetics 7: 665-674.

Ruell, E. W. and K. R. Crooks. 2007. Evaluation of noninvasive genetic sampling methods for felid and canid populations. Journal of Wildlife Management 71: 1690-1694.

Scandura, M., C. Capitani, L. Iacolina, and A. Marco. 2006. An empirical approach for reliable microsatellite genotyping of wolf DNA from multiple noninvasive sources. Conservation Genetics 7: 813-823.

Schwartz, M.K., S.A. Cushman, K.S. McKelvey, J. Hayden, and C. Engkjer. 2006. Detecting genotyping errors and describing American black bear movement in northern Idaho. Ursus 17:138-148.

Sefc, K., R. Payne, and M.D. Sorenson. 2003. Microsatellite amplification from museum feather samples: effects of fragment size and template concentration on genotyping errors. The Auk 120: 982-989.

Segelbacher, G. 2002. Noninvasive genetic analysis in birds: testing reliability of feather samples. Molecular Ecology Notes 2: 367-369.

Sloane, M.A., P. Sunnucks, D. Alpers, B. Beheregaray, and A.C. Taylor. 2000. Highly reliable genetic identification of individual northern hairy-nosed wombats from single remotely collected hairs: a feasible censusing method. Molecular Ecology 9: 123-124.

Smith, K. L., S.C. Alberts, M.K. Bayes, M.W. Bruford, J. Altmann, and C. Ober. 2000. Cross-species amplification, non-invasive genotyping, and non-Mendelian inheritance of human STRPs in Savannah baboons. American Journal of Primatology 51: 219-227.

Soulsbury, C. D. , G. Iossa, K. J. Edwards, P. J. Baker, and S. Harris. 2007. Allelic dropout from a high-quality DNA source. Conservation Genetics 8: 733-738.

Valiere, N. and P. Taberlet. 2000. Urine collected in the field as a source of DNA for species and individual identification. Molecular Ecology 9: 2149-2154.

Valiere, N. , C. Bonenfant, C. Toigo, G. Luikart, J-M. Gaillard, and F. Klein. 2007. Importance of a pilot study for non-invasive genetic sampling: genotyping errors and population size estimation in red deer. Conservation Genetics 8: 69-78.

Appendix C. Genetic error proof.

Part 1. General form_

A general form of the possibilities of observed genotypes, when there are 3 or more possible alleles at a given locus ($n \geq 3$), and associated probabilities are in Figure 1.

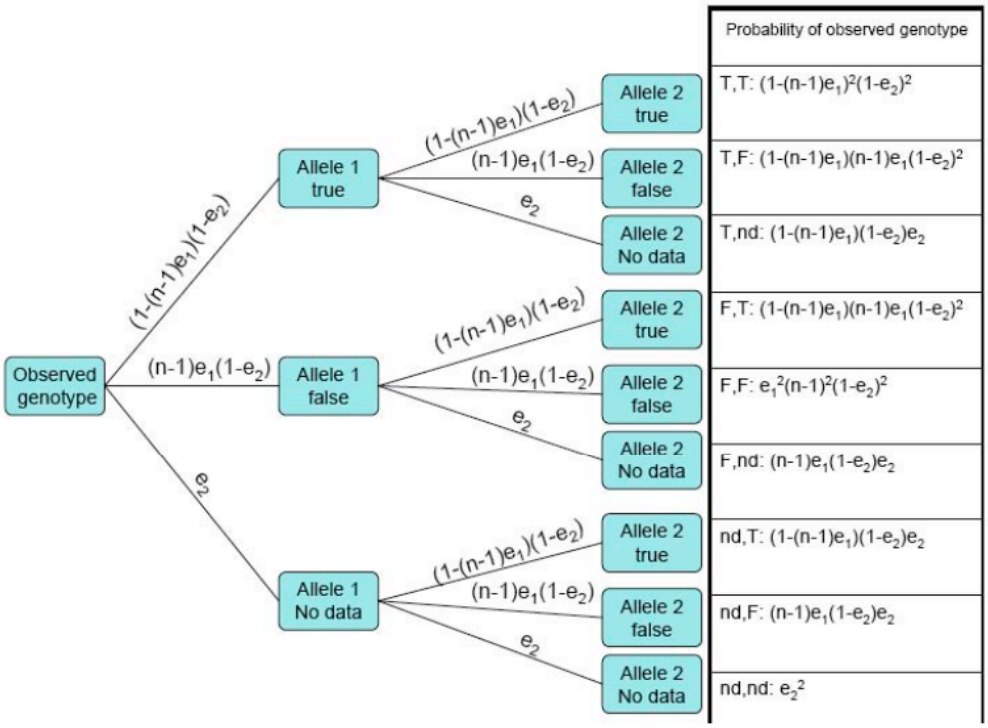


Figure 1. Observed genotype possibilities and respective probabilities when there are 3 or more possible alleles per locus ($n \geq 3$)

For the first allele at a locus, the sum of the probabilities of observing each of the 3 possibilities (true, false, no data) must sum to one.

$(1-(n-1)e_1)(1-e_2)+e_1(n-1)(1-e_2)+e_2$

$=(1-(n-1)\,e_1)-e_2\,(1-(n-1)\,e_1)+e_1\,(n-1)-e_1\,(n-1)\,e_2+e_2$

$=1-(n-1)\,e_1-e_2+(n-1)\,e_1\,e_2+(n-1)\,e_1-(n-1)\,e_1\,e_2+e_2$

$=1$

Since this is a diploid organism there are 2 alleles at a locus. The sum of the probabilities of observing each of the 9 combinations (TT,TF, FT, FF, Tnd, ndT, Fnd, ndF, ndnd) must sum to one. For simplification, this was divided into 3 categories, with the sum of the 3 categories as the last step.

a) $(1-(n-1)e_1)^2(1-e_2)^2+(n-1)\,e_1\,(1-(n-1)\,e_1)(1-e_2)^2+(1-(n-1)\,e_1)(1-e_2)\,e_2$

$=(1-(n-1)\,e_1)\,(1-(n-1)\,e_1)\,(1-e_2)^2+(n-1)\,e_1\,(1-(n-1)\,e_1)(1-e_2)^2+(1-(n-1)\,e_1)(1-e_2)\,e_2$

$=(1-(n-1)\,e_1)\,(1-e_2)^2-(n-1)\,e_1\,(1-(n-1)\,e_1)\,(1-e_2)^2+(n-1)\,e_1\,(1-(n-1)e_1)(1-e_2)^2+(1-($
$e_2)\,e_2$

$=(1-(n-1)\,e_1)\,(1-e_2)\,(1-e_2)+(1-(n-1)\,e_1)(1-e_2)\,e_2=(1-(n-1)$

$e_1)\,(1-e_2)-(1-(n-1)\,e_1)\,(1-e_2)\,e_2+(1-(n-1)\,e_1)(1-e_2)\,e_2$

$=(1-(n-1)\,e_1)\,(1-e2)$

b) $(n-1)\,e_1\,(1-(n-1)\,e_1)(1-e_2)^2+(n-1)^2\,e_1^2(1-e_2)^2+(n-1)\,e_1\,(1-e_2)\,e_2$

$=((n-1)\,e_1-(n-1)^2\,e_1^2))(1-e_2)^2+(n-1)^2\,e_1^2(1-e_2)^2+(n-1)\,e_1\,(1-e_2)\,e_2$

$= (n-1)e_1(1-e_2)^2 -(n-1)^2 e_1^2 (1- e_2)^2 +(n-1)^2 e_1^2 (1- e_2)^2 + (n-1) e_1 (1- e_2) e_2$

$= (n-1) e_1(1- e_2) (1- e_2)+ (n-1) e_1 (1- e_2) e_2$

$= (n-1) e_1 (1- e_2)-(n-1) e_1 e_2 (1- e_2)+ (n-1) e_1 (1- e_2) e_2$

$= (n-1) e_1 (1- e_2)$

c) $e_2(1-(n-1)e_1)(1- e_2)+ (n-1) e_1 e_2 (1- e_2)+ e_2^2$

$= e_2 (1-e_2)-(n-1) e_1 e_2 (1-e_2)+ (n-1) e_1 e_2 (1- e_2)+ e_2^2$

$= e_2-e_2^2 + e_2^2$

$= e_2$

d) parts a+b+c= $(1-(n-1)e_1) (1-e_2)+ (n-1) e_1 (1- e_2)+ e_2$

$=(1- e_2) – (n-1) e_1 (1- e_2) +(n-1) e_1 (1- e_2)+ e_2$

$=1$

The 9 combinations can be combined (based on true homozygous or heterozygous), since an observed genotype may be different than the underlying genotype. When n>2, and the true genotype is heterozygous, there are 6 observed categories (with respective probabilities). The case of 2 false alleles can create additions/subtractions from the categories from what would be expected.

1) observed heterozygote= true heterozygote (this accounts for the one case in which 2 false alleles can lead to an observed true heterozygote)

$$(1-(n-1)e_1)^2(1-e_2)^2+e_1^2(1-e_2)^2$$

2) observed heterozygote does not equal heterozygote with one true allele (this accounts for the 2*(n-2) ways that a F/F could result in this category and the 2(n-2) ways a T/F could result in this category)

$$2(n-2)(1-(n-1)e_1)\, e_1\, (1-e_2)^2+2(n-2)e_1^2(1-e_2)^2$$

3) observed heterozygote does not equal heterozygote with no alleles true (this accounts for F/F where the first allele there are n-2 ways not to be true, and the second allele there are n-3 ways not to be true and still be heterozygous)

$$(n-2)(n-3)\, e_1^2(1-e_2)^2$$

4) observed homozygote with only 1 true allele (this accounts for the 2 ways that T/nd leads to this scenario, the 2 ways that F/nd leads to this scenario, the 2 ways that TF leads to this scenario)

$$2(1-(n-1)\, e_1)\, e_2\, (1-e_2)+2\, e_1\, e_2\, (1-e_2)+2\, e_1\, (1-(n-1)\, e_1)(1-e_2)^2$$

5) observed homozygote with no true alleles (this accounts for n-2 ways that 2 false alleles at locus could be observed homozygous with no true alleles, and the 2(n-2) ways that a F/nd could result in this scenario)

$$2(n-2)e_1e_2(1-e_2)+(n-2)\, e_1^2(1-e_2)^2$$

6) no data

e_2^2

The heterozygous categories must sum to one.

a) category 4 + category 5

$2(1-(n-1)e_1)e_2(1-e_2)+2\,e_1\,e_2\,(1-e_2)+2\,e_1\,(1-(n-1)\,e_1)(1-e_2)^2+2(n-2)\,e_1\,e_2\,(1-e_2)+(n-2)\,e_1^2(1e_2)^2$

$=2(1-(n-1)e_1)e_2(1-e_2)+2(n-1)\,e_1\,e_2\,(1-e_2)+2\,e_1\,(1-(n-1)\,e_1)(1-e_2)^2+(n-2)\,e_1^2(1-e2)^2$

b) a + category 2

$2(1-(n-1)e_1)e_2(1-e_2)+2(n-1)\,e_1\,e_2\,(1-e_2)+2\,e_1\,(1-(n-1)\,e_1)(1-e_2)^2+(n-2)\,e_1^2(1-e_2)^2+2(n-2)(1(n-1)\,e_1)\,e_1\,(1-e_2)^2+2(n-2)\,e_1^2(1-e_2)^2$

$=2(1-(n-1)\,e_1)\,e_2\,(1-e_2)+2(n-1)\,e_1\,e_2\,(1-e_2)+2(n-1)\,e_1\,(1-(n-1)\,e_1)(1-e_2)^2+3(n-2)\,e_1^2(1-e_2)^2$

b) b+ category 1

$=2(1-(n-1)e1)e2(1-e2)+2(n-1)e1e2(1-e2)+2(n-1)e1(1-(n-1)e1)(1-e2)^2+3(n-2)\,e1^2(1\,e2)^2+(1-(n-1)e1)^2(1-e2)^2+e1^2(1-e2)^2$

$=2(1-(n-1)e1)e2(1-e2)+2(n-1)e1e2(1-e2)+2(n-1)e1(1-(n-1)e1)(1-e2)^2+(3n-5)e1^2(1e2)^2+(1-(n-1)e1)^2(1-e2)^2$

d) c+ category 3 $2(1-(n-1)e1)e2(1-e2)+2(n-1)e1e2(1-e2)+2(n-1)e1(1-(n-1)e1)(1-e2)^2+(3n-5)e1^2(1-e2)^2+(1(n-1)e1)^2(1-e2)^2+(n-2)(n-3)e1^2(1-e2)^2$

$=2(1-(n-1)e1)e2(1-e2)+ 2(n-1)e1e2(1-e2)+ 2(n-1)e1(1-(n-1)e1)(1-e2)^2+(n-1)^2e1^2(1-e2)^2+(1-(n-1)e1)^2(1-e2)^2$

$=2e2(1-e2)-\underline{2(n-1)e2e1(1-e2)}+ \underline{2(n-1)e1e2(1-e2)}+ 2(n-1)e1(1-(n-1)e1)(1-e2)^2+(n-1)^2e1^2(1e2)^2+(1-(n-1)e1)^2(1-e2)^2$

$=2e2(1-e2)+ 2(n-1)e1(1-(n-1)e1)(1-e2)^2+(n-1)^2e1^2(1-e2)^2+(1-(n-1)e1)^2(1-e2)^2$

$=2e2(1-e2)+ 2(n-1)e1(1-(n-1)e1)(1-e2)^2+(n-1)^2e1^2(1-e2)^2+ (1-(n-1)e1)(1-(n-1)e1)(1-e2)^2$

$=2e2(1-e2)+ 2(n-1)e1(1-(n-1)e1)(1-e2)^2+(n-1)^2e1^2(1-e2)^2+[1-2(n-1)e1+(n-1)^2e1^2](1-e2)^2$

$=2e2(1-e2)+ 2(n-1)e1(1-(n-1)e1)(1-e2)^2\ \underline{+(n-1)^2e1^2(1-e2)^2}+(1-e2)^2-2(n-1)e1(1-e2)^2\ \underline{+(n-1)^2e1^2(1-e2)^2}$

$=2e2(1-e2)+ 2(n-1)e1(1-(n-1)e1)(1-e2)^2+2(n-1)^2e1^2(1-e2)^2+(1-e2)^2-2(n-1)e1(1-e2)^2$

$=2e2(1-e2)+ \underline{2(n-1)e1(1-e2)^2-2(n-1)^2e1^2(1-e2)^2}+\underline{2(n-1)^2e1^2(1-e2)^2}+(1-e2)^2\ \underline{-2(n-1)e1(1-e2)^2}$

$=2e2(1-e2)+(1-e2)^2$

$=\underline{2e2-2e_2^2}+ 1\ \underline{-2e2}+\underline{e_2^2}$

$=1-e_2^2$

e) d + category 6

$1-e2^2+e2^2$

When n>2, and the true genotype is homozygous, there are 5 observed categories (with respective probabilities):

1) observed heterozygote= true heterozygote

$(1-(n-1)e1)^2(1-e2)^2+2(1-(n-1)e1)(1-e2)\ e2$

2) observed homozygote not equal to the true homozygote (this includes n-1 ways that the F/F results in a homozygote in this category)

$2e1(n-1)(1-e2)e2+(n-1)e1^{2}(1-e2)^{2}$

3) observed heterozygote with one allele equal to the true homozygote

$2(1-(n-1)e1)(n-1)e1(1-e2)^{2}$

4) observed heterozygote with no alleles equal to true homozygote (this takes in account the n-1 ways that the F/F falls into the 2^{nd} category)

$[(n-1)^{2}-(n-1)]e1^{2}(1-e2)^{2}$

5) no data

$e_{2}{}_{2}$

The homozygous categories must sum to one.

a) category 2+4

$2e1(n-1)(1-e2)e2+(n-1)e1^{2}(1-e2)^{2}+[(n-1)^{2}-(n-1)]e1^{2}(1-e2)^{2}$

$=2e1(n-1)(1-e2)e2+(n-1)^{2}e1^{2}(1-e2)^{2}$

b) a + category 3

$2e1(n-1)(1-e2)e2+(n-1)^{2}e1^{2}(1-e2)^{2}+2(1-(n-1)e1)(n-1)e1(1-e2)^{2}$

$=2e1(n-1)(1-e2)e2+(n-1)^{2}e1^{2}(1-e2)^{2}+2[(n-1)e1(1-e2)^{2}-(n-1)^{2}e1^{2}(1-e2)^{2}]$

$=2e1(n-1)(1-e2)e2+2(n-1)e1(1-e2)^{2}-(n-1)^{2}e1^{2}(1-e2)^{2}$

$=2e1(n-1)(1-e2)e2+2(n-1)e1(1-e2)(1-e2)-(n-1)^{2}e1^{2}(1-e2)^{2}$

$=2e1(n-1)(1-e2)e2+2(n-1)e1(1-e2)-2(n-1)e1(1-e2)e2-(n-1)^2 e1^2 (1-e2)^2$

$=2(n-1)e1(1-e2) -(n-1)^2 e1^2 (1-e2)^2$

c) b + category 1

$2(n-1)e\ 1(1-e2) -(n-1)^2 e1^2 (1-e2)^2+(1-(n-1)e1)^2 (1-e2)^2+2(1-(n-1)e1)(1-e2)\ e2$

$=2(n-1)e\ 1(1-e2) -(n-1)^2 e1^2 (1-e2)^2+(1-(n-1)e1)\ (1-(n-1)e1)\ (1-e2)^2+2(1-(n-1)e1)(1-e2)\ e2$

$=2(n-1)e\ 1(1-e2) -(n-1)^2 e1^2 (1-e2)^2+(1-(n-1)e1)\ (1-e2)^2-(n-1)e1(1-(n-1)e1)\ (1e2)^2+2(1-(n-1)e1)(1-e2)\ e2$

$=2(n-1)e1(1-e2)\ \underline{-(n-1)^2 e1^2 (1-e2)^2}+(1-e2)^2-(n-1)e1(1-e2)^2-(n-1)e1(1-e2)^2\ \underline{+(n-1)^2 e1^2 (1-e2)^2}+2(1-(n-1)e1)(1-e2)\ e2$

$=2(n-1)e\ 1(1-e2) +(1-e2)^2-2(n-1)e1(1-e2)^2+2(1-(n-1)e1)(1-e2)\ e2$

$=2(n-1)e\ 1(1-e2) +(1-e2)^2-2(n-1)e1(1-e2)^2+2(1-e2)e2-2(1-e2)e2(n-1)e1$

$=2(n-1)e\ 1(1-e2) +(1-e2)^2-2(n-1)e1(1-e2)(1-e2)+2(1-e2)e2-2(1-e2)e2(n-1)e1$

$=\underline{2(n-1)e1(1-e2)}\ +(1-e2)^2\ \underline{-2(n-1)e1(1-e2)}+\underline{2(n-1)e1(1-e2)e2}+2(1-e2)e2-\underline{2(1-e2)e2(n1)e1}$

$=(1-e\ 2)^2+2(1-e2)e2$

d) c + category 5

$(1-e2)^2+2(1-e2)e2+\ e2^2$

$=(1-e2)\ (1-e2)\ +2\ e2-2e2^2+\ e2^2$

$=1-2e2+\ e2^2+2e2-2e2^2+\ e2^2$

$=1$

Here is an example with 3 alleles in Figure 2 with a true heterozygote and Figure 3 with a true homozygote. By deduction, the above should be true for all n>3 alleles at a given locus.

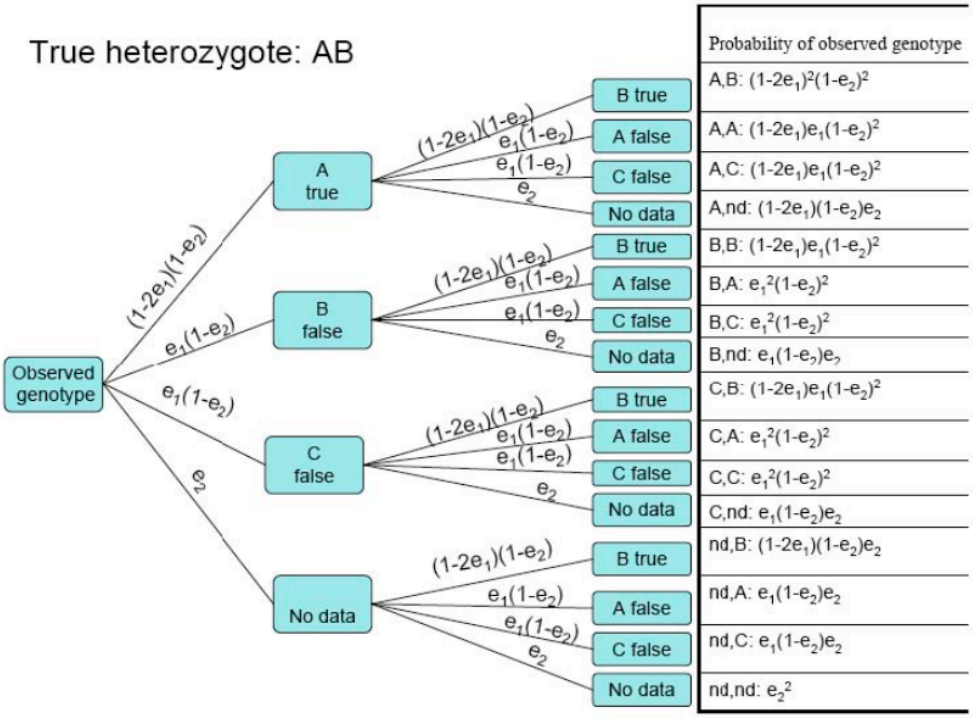


Figure 2. Example with 3 alleles with a true heterozygote.

Under the above scenario, the probability of observing each category is summarized below:

1) observed heterozygote= true heterozygote

$(1-2e_1)^2(1-e_2)^2+e_1^2(1-e_2)^2$

2) observed heterozygote does not equal heterozygote with one true allele

$2(1-2e1)e1(1-e2)^2+2e1^2(1-e2)^2$

3) observed heterozygote does not equal heterozygote with no alleles true (this category would be filled with n>3 alleles)

0

4) observed homozygote with only 1 true allele

$2(1-2e1)e1(1-e2)^2+2(1-2e1)e2(1-e2)+2e1e2(1-e2)$

5) observed homozygote with no true alleles

$2e1e2(1-e2)+e1^2(1-e2)^2$

6) no data

e_2^2

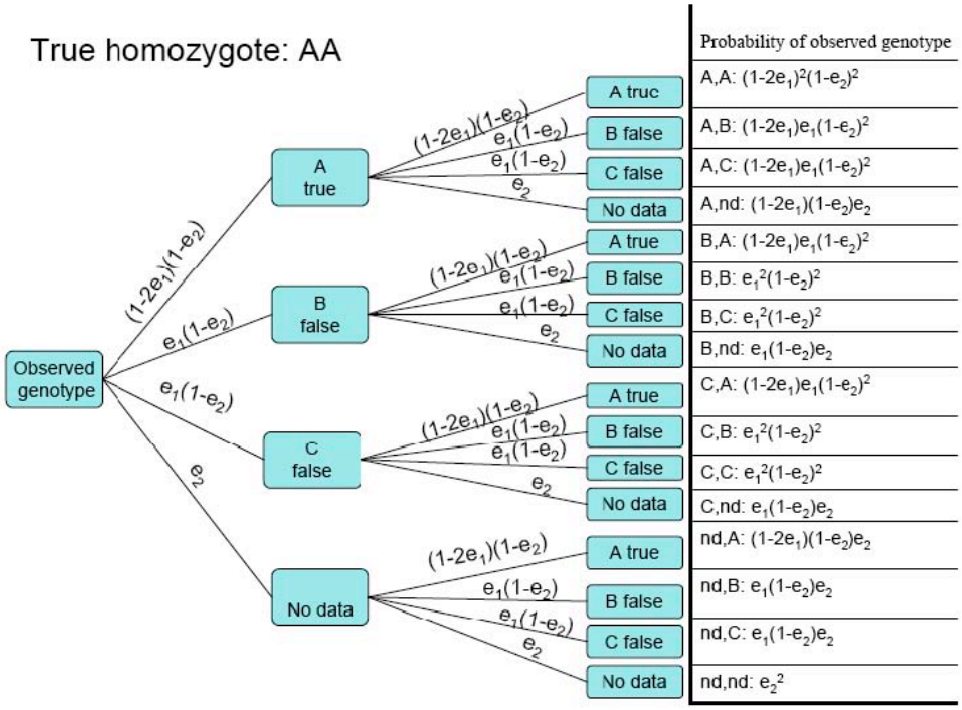


Figure 3. Example with 3 alleles of a true homozygote.

Under the above scenario, the probability of observing each of the 5 categories is summarized below:

1) observed heterozygote= true heterozygote

$(1-2e1)^2(1-e2)^2+2(1-2e1)(1-e2)\,e2$

2) observed homozygote not equal to the true homozygote

$4e1(1-e2)e2+2e1^2(1-e2)^2$

3) observed heterozygote with one allele equal to the true homozygote

$4(1-2e1)e1(1-e2)^2$

4) observed heterozygote with no alleles equal to true homozygote

$2e1^2(1-e2)^2$

5) no data

$e2^2$

Part 2. Special case

A special case occurs when there are only 2 alleles (n=2) possible at one locus. A general form of the possibilities, and associated probabilities are in Figure 4.

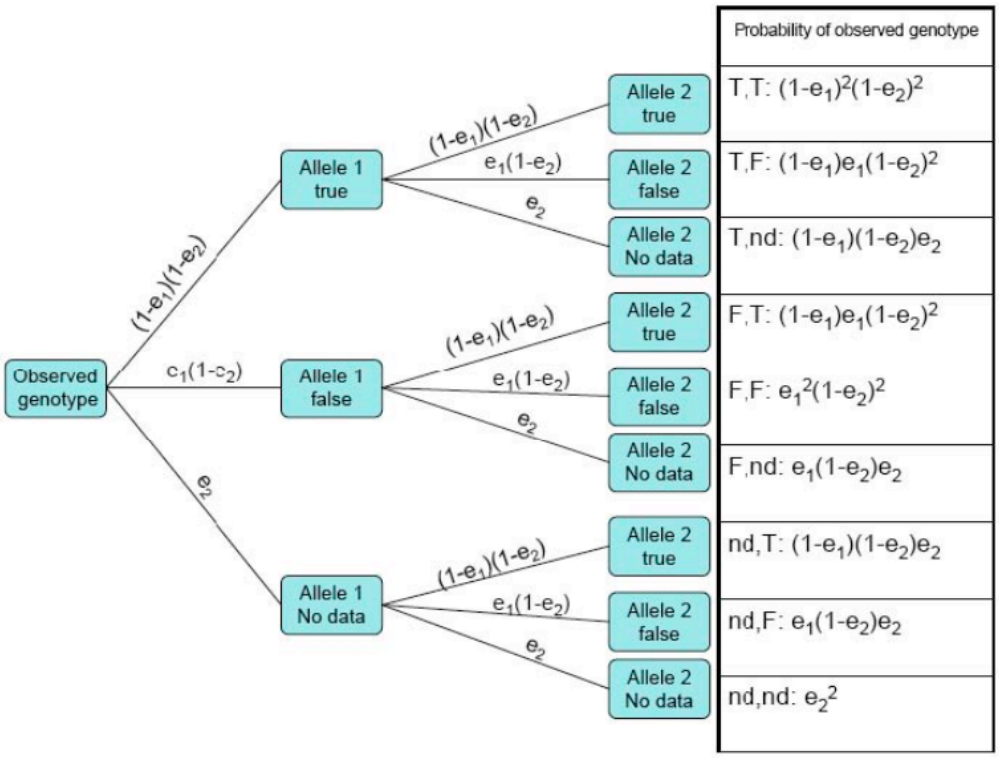


Figure 4. Observed genotype possibilities and respective probabilities when there are 2 possible alleles per locus (n=2)

For the first allele at a locus, the sum of the probabilities (true, false, no data) of observing each of the 3 possibilities must sum to one.

(1-e1)(1-e2)+e1(1-e2)+e2

=(1-e 1)-e2(1-e1)+e1-e1e2+e2

=1-e 1-e2+e1e2+e1-e1e2+e2

=1

For 2 alleles at a locus, the sum of the probabilities of observing each of the 9 combinations (TT,TF,FT,FF,Tnd,ndT,Fnd,ndF,ndnd) must sum to one. For simplification, this was divided into 3 categories, with the sum of the 3 categories as the last step.

a) $(1-e1)^2(1-e2)^2+e1(1-e1)(1-e2)^2+(1-e1)(1-e2)e2$

$=(1-e1)(1-e1)(1-e2)^2+e1(1-e1)(1-e2)^2+(1-e1)(1-e2)e2$

$=(1-e1)(1-e2)^2-e1(1-e1)(1-e2)^2+e1(1-e1)(1-e2)^2+(1-e1)(1-e2)e2$

$=(1-e1)(1-e2)(1-e2)+(1-e1)(1-e2)e2 =(1-e1)(1-e2)-(1-e1)(1-e2)$

$e2+(1-e1)(1-e2)e2$

$=(1-e1)(1-e2)$

b) $e1(1-e1)(1-e2)^2+e1^2(1-e2)^2+e1(1-e2)e2$

$=(e1-e1^2)(1-e2)^2+e1^2(1-e2)^2+e1(1-e2)e2$

$=e1(1-e2)^2-e1^2(1-e2)^2+e1^2(1-e2)^2+e1(1-e2)e2$

$=e1(1-e2)(1-e2)+e1(1-e2)e2$

$=e1(1-e2)-e1e2(1-e2)+e1(1-e2)e2$

$=e1(1-e2)$

c) $e2(1-e1)(1-e2)+e1e2(1-e2)+ e2^2$

$= e2(1-e2)-e1e2(1-e2)+ e1e2(1-e2)+ e2^2$

$= e2-e2^2 + e2^2$

$= e2$

d) parts a+b+c= $(1-e1)(1-e2)+ e1(1-e2)+ e2$

$=(1-e2) – e1(1-e2) + e1(1-e2)+ e2$

$=1$

The 9 combinations can be combined (based on true homozygous or heterozygous), since an observed genotype may be different than the underlying genotype. When n=2, and the true genotype is heterozygous, there are 3 observed categories (with respective probabilities):

6) observed heterozygote= true heterozygote

$(1-e1)^2(1-e2)^2+e1^2(1-e2)^2$

7) observed homozygote with only 1 true allele

$2e1(1-e1)(1-e2)^2+2(1-e1)(1-e2)e2+2e1(1-e2)e2$

8) no data

$e_2{}^2$

When n=2, and the true genotype is homozygous, there are 4 observed categories (with respective probabilities):

9) observed heterozygote= true heterozygote

$(1-e1)^2 (1-e2)^2 + 2(1-e1)(1-e2)\, e2$

10) observed homozygote not equal to the true homozygote

$2e1(1-e2)e2 + e1^2 (1-e2)^2$

11) observed heterozygote with one allele equal to the true homozygote

$2(1-e1)e1(1-e2)^2$

12) no data

e_2^2

Here is an example with 2 alleles in Figure 5 with a true heterozygote and Figure 6 with a true homozygote.

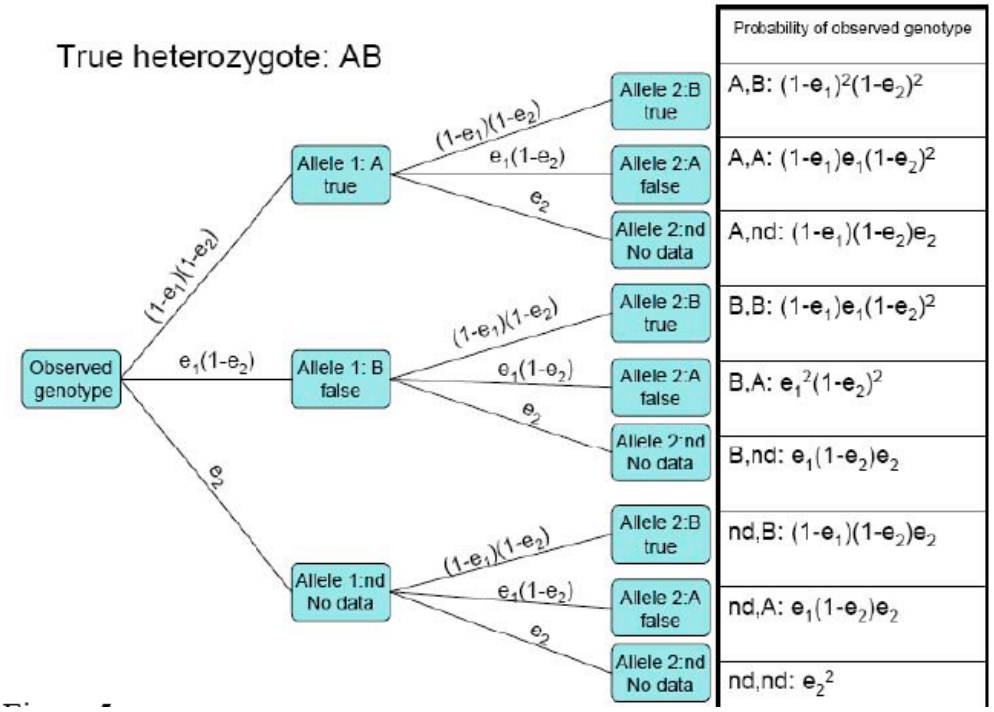


Figure 5. Example of true heterozygote with 2 alleles.

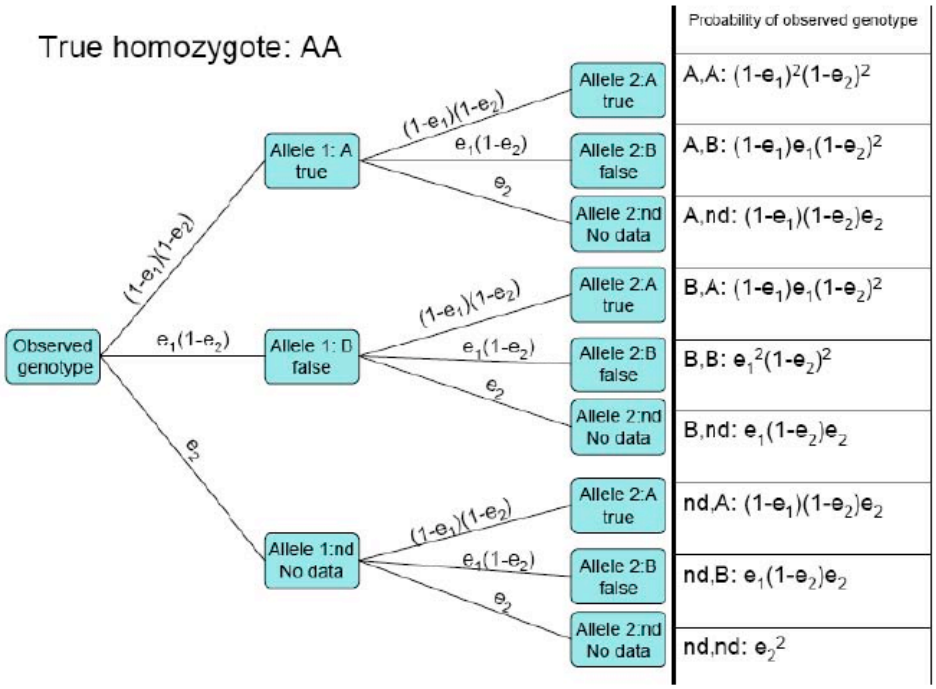


Figure 6. Example of true homozygote with 2 alleles.