

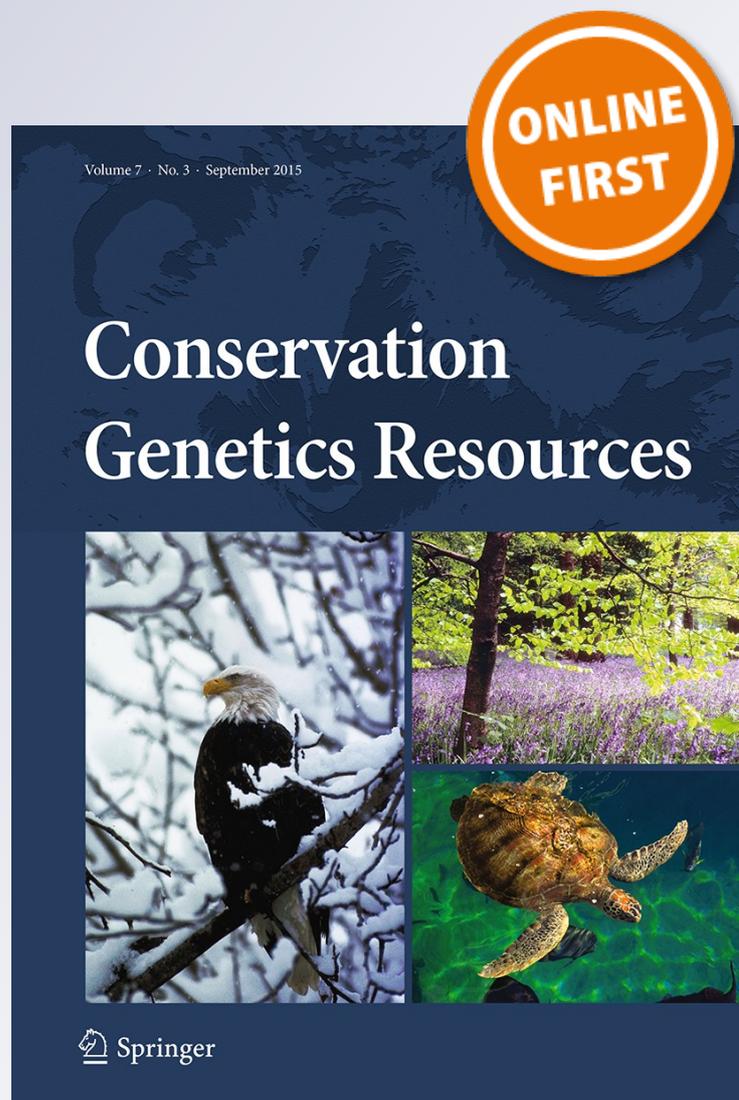
*ConGenR: rapid determination of
consensus genotypes and estimates of
genotyping errors from replicated genetic
samples*

Robert C. Lonsinger & Lisette P. Waits

Conservation Genetics Resources

ISSN 1877-7252

Conservation Genet Resour
DOI 10.1007/s12686-015-0506-7



Your article is protected by copyright and all rights are held exclusively by Springer Science +Business Media Dordrecht. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your article, please use the accepted manuscript version for posting on your own website. You may further deposit the accepted manuscript version in any repository, provided it is only made publicly available 12 months after official publication or later and provided acknowledgement is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be accompanied by the following text: "The final publication is available at link.springer.com".

ConGenR: rapid determination of consensus genotypes and estimates of genotyping errors from replicated genetic samples

Robert C. Lonsinger¹ · Lisette P. Waits¹

Received: 31 May 2015 / Accepted: 4 November 2015
© Springer Science+Business Media Dordrecht 2015

Abstract ConGenR (available at <http://www.uidaho.edu/cnr/research-outreach/facilities/leecg/publications-and-software>) is an R based conservation genetics script that facilitates rapid determination of consensus genotypes from replicated samples, determines overall (successful amplifications/amplification attempted) and individual sample level (proportion of samples with successful amplifications at n loci) amplification success rates, and quantifies genotyping error rates. ConGenR is intended for use with codominant, multilocus microsatellite data generated primarily through noninvasive genetic sampling and processed with a multi-tubes approach. ConGenR handles input that can be easily exported from GENEMAPPER, a program commonly used to score allele sizes. Amplification success and genotyping error rates can be evaluated by sample class (i.e., any identifiable and meaningful subdivision of samples; e.g., sex, season, region, or sample condition), offering insights into processes driving amplification success and genotyping error rates. Additionally, amplification success and genotyping error rates are calculated by locus, expediting the identification of problematic loci during pilot studies.

Keywords Allelic dropout · Consensus genotypes · False alleles · Genotyping errors · Noninvasive genetic sampling · PCR success

Noninvasive genetic sampling is an appealing monitoring strategy when working with species that are difficult to

observe or capture and provides the opportunity to identify individuals (Waits et al. 2001), estimate population demographic parameters (Marucco et al. 2011), and evaluate genetic health without observing or handling individuals (Waits and Paetkau 2005; Beja-Pereira et al. 2009). Noninvasive genetic samples are typically characterized by low quantity and quality DNA, leading to low polymerase chain reaction (PCR) success and the presence of genotyping errors, making it challenging to obtain reliable genotypes (Pompanon et al. 2005; Waits and Paetkau 2005; Broquet et al. 2006). A multi-tubes approach is frequently used to establish reliable consensus genotypes and minimize the influence of genotyping errors (Taberlet et al. 1996). Genotyping errors are typically classified as a false allele (FA), where an allele is observed within a replicate that is not present in the consensus or reference genotype, or allelic dropout (ADO), where an allele present in the consensus or reference genotype fails to amplify in a successful PCR replicate (Broquet and Petit 2004). Prior to initiating noninvasive genetic monitoring, pilot studies are recommended to quantify PCR and genotyping error rates (Bonin et al. 2004; Valière et al. 2006), determine sufficient sampling designs under observed rates (Rodgers and Janečka 2012; Lonsinger et al. 2015), and evaluate if errors are sufficiently low to avoid substantial biases (Waits and Leberg 2000; Luikart et al. 2010). Still, genotyping error is often neglected or not reported (Pompanon et al. 2005).

The program GENEMAPPER (Applied Biosystems, Foster City, CA, USA) is widely used for scoring alleles. Efficiently handling results files generated by GENEMAPPER can be cumbersome; particularly when working with noninvasive samples analyzed using a multi-tubes approach to achieve reliable consensus genotypes (Taberlet et al. 1996) and analyzed across a sufficient number of loci for individual identification (Waits et al. 2001).

✉ Robert C. Lonsinger
Lons1663@vandals.uidaho.edu

¹ Department of Fish and Wildlife Sciences, University of Idaho, Moscow, ID 83844-1136, USA

GENEMAPPER results files include a row for each PCR reaction-locus combination, leading to large files (e.g., 1000 samples each analyzed in four replicates at 10 microsatellite loci would yield 40,000 rows of data; in practice, >4 replicates are usually required for reliable consensus genotypes and >10 loci may be necessary for individual identification). Decreasing costs are making noninvasive genetic sampling applicable to large scale surveys (Beja-Pereira et al. 2009) and consequently, non-invasive genetic monitoring projects collecting thousands of samples have become common (e.g., Kendall et al. 2008; Brinkman et al. 2010). Thus, GENEMAPPER results files can contain hundreds of thousands of lines, making data handling, comparing replicates and determining consensus genotypes, and calculating genotyping error rates time-consuming and error-prone if completed manually.

Our goal was to develop a method to quickly handle result files from GENEMAPPER and evaluate replicated genotype data. To this end, ConGenR facilitates the rapid determination of consensus genotypes from replicated samples, determines overall and individual sample level PCR success, and calculates observed genotyping error rates (i.e., ADO and FA rates). Additionally, ConGenR can be used to compare multilocus consensus genotypes across samples, identify samples that match at all or a specified number of loci, and evaluate the spatial relationship between matches and near matches. ConGenR is intended for use with codominant, multilocus microsatellite data generated primarily through noninvasive genetic sampling and processed with a multi-tubes approach (Taberlet et al. 1996). ConGenR is written in the R programming language (R Core Team 2015) and is designed to handle input in a format that can be easily exported from GENEMAPPER or alternative databases such as Microsoft Access (Microsoft, Redmond, WA, USA). ConGenR allows users to efficiently compare overall and individual sample level PCR success rates, as well as genotyping error rates, by sample class (i.e., any identifiable and meaningful subdivision of samples such as sex, season, region, or sample condition). ConGenR also estimates PCR success and genotyping error rates by locus, expediting the identification of problematic loci during pilot studies. Researchers interested in calculating genotyping error rates by comparing low quality samples (e.g., noninvasive samples) to high quality reference samples (e.g., blood or tissue samples) can do so by directly calling the genotyping error function; this is particularly useful when conducting pilot studies to evaluate genotyping error rates using noninvasive samples collected from known individuals from which high quality samples have been obtained.

To determine consensus genotypes, ConGenR employs common protocols for replicated DNA samples (e.g., Frantz et al. 2003; Flagstad et al. 2004). Specifically, ConGenR requires that each allele of heterozygous

genotypes be observed ≥ 2 times, while single alleles must be observed ≥ 3 times to confirm a homozygous genotype. ConGenR calculates an overall assessment of PCR success (the number of successful amplifications/the total number of amplifications attempted) and an individual sample level PCR success rate (the proportion of samples with successful amplifications at n loci; n will most appropriately be set to the number of loci required for individual identification [Waits et al. 2001]). ConGenR quantifies genotyping error rates by comparing each replicated genotype to the consensus or reference genotype (Lampa et al. 2013) and generally follows procedures detailed by Broquet and Petit (2004), except ConGenR allows ADOs to be scored for genotypes confirmed as homozygous, when the presence of the FA indicates a successful PCR amplification but the confirmed allele fails to amplify. The ConGenR script, including source code, a supporting user manual, and example input files are available at <http://www.uidaho.edu/cnr/research-outreach/facilities/leecg/publications-and-software>.

Acknowledgments This project was funded by the U.S. Department of Defense's Environmental Security Technology Certification (12 EB-RC5-006) and Legacy Resource Management (W9132T-12-2-0050) programs. We thank K Cleary and S Woodruff for providing data and opportunities to test and improve ConGenR.

References

- Beja-Pereira A, Oliveira R, Alves PC, Schwartz MK, Luikart G (2009) Advancing ecological understandings through technological transformations in noninvasive genetics. *Mol Ecol Resour* 9:1279–1301
- Bonin A, Bellemain E, Bronken Eidesen P, Pompanon F, Brochmann C, Taberlet P (2004) How to track and assess genotyping errors in population genetics studies. *Mol Ecol* 13:3261–3273
- Brinkman TJ, Person DK, Schwartz MK, Pilgrim KL, Colson KE, Hundertmark KJ (2010) Individual identification of sitka black-tailed deer (*Odocoileus hemionus sitkensis*) using DNA from fecal pellets. *Conserv Genet Resour* 2:115–118
- Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. *Mol Ecol* 13:3601–3608
- Broquet T, Ménard N, Petit E (2006) Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates. *Conserv Genet* 8:249–260
- Flagstad Ø, Hedmark E, Landa A, Brøseth H, Persson J, Andersen R, Segerström P, Ellegren H (2004) Colonization history and noninvasive monitoring of a reestablished wolverine population. *Conserv Biol* 18:676–688
- Frantz AC, Pope LC, Carpenter PJ, Roper TJ, Wilson GJ, Delahay RJ, Burke T (2003) Reliable microsatellite genotyping of the Eurasian badger (*Meles meles*) using faecal DNA. *Mol Ecol* 12:1649–1661
- Kendall KC, Stetz JB, Roon DA, Waits LP, Boulanger JB, Paetkau D (2008) Grizzly bear density in Glacier National Park, Montana. *J Wildl Manage* 72:1693–1705
- Lampa S, Henle K, Klenke R, Hoehn M, Gruber B (2013) How to overcome genotyping errors in non-invasive genetic mark-

- recapture population size estimation—a review of available methods illustrated by a case study. *J Wildl Manage* 77:1490–1511
- Lonsinger RC, Gese EM, Dempsey SJ, Kluever BM, Johnson TR, Waits LP (2015) Balancing sample accumulation and DNA degradation rates to optimize noninvasive genetic sampling of sympatric carnivores. *Mol Ecol Resour* 15:831–842
- Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW (2010) Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conserv Genet* 11:355–373
- Marucco F, Boitani L, Pletscher DH, Schwartz MK (2011) Bridging the gaps between non-invasive genetic sampling and population parameter estimation. *Eur J Wildl Res* 57:1–13
- Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* 6:847–859
- R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Rodgers TW, Janečka JE (2012) Applications and techniques for non-invasive faecal genetics research in felid conservation. *Eur J Wildl Res* 59:1–16
- Taberlet P, Griffin S, Goossens B, Questiau S, Manceau V, Escaravage N, Waits LP, Bouvet J (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res* 24:3189–3194
- Valière N, Bonenfant C, Toïgo C, Luikart G, Gaillard J, Klein F (2006) Importance of a pilot study for non-invasive genetic sampling: genotyping errors and population size estimation in red deer. *Conserv Genet* 8:69–78
- Waits JL, Leberg PL (2000) Biases associated with population estimation using molecular tagging. *Anim Conserv* 3:191–199
- Waits LP, Paetkau D (2005) Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. *J Wildl Manage* 69:1419–1433
- Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Mol Ecol* 10:249–256