MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2013) **13**, 154–157 doi: 10.1111/1755-0998.12019

Capwire: a R package for estimating population census size from non-invasive genetic sampling

MATTHEW W. PENNELL,*† CARISA R. STANSBURY,‡ LISETTE P. WAITS*‡ and CRAIG R. MILLER*†§

*Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, 441B Life Science South, Moscow, ID 83844, USA, †Department of Biological Sciences, University of Idaho, 252 Life Sciences South, Moscow, ID 83844, USA, ‡Department of Fish and Wildlife Sciences, University of Idaho, 975 West 6th Street, Moscow, ID 83844, USA, §Department of Mathematics, University of Idaho, 300 Brink Hall, Moscow, ID 83844, USA

Abstract

Non-invasive genetic sampling is an increasingly popular approach for investigating the demographics of natural populations. This has also become a useful tool for managers and conservation biologists, especially for those species for which traditional mark–recapture studies are not practical. However, the consequence of collecting DNA indirectly is that an individual may be sampled multiple times per sampling session. This requires alternative statistical approaches to those used in traditional mark–recapture studies. Here we present the R package `capwire`, an implementation of the population size estimators of Miller *et al.* (*Molecular Ecology* 2005; 14: 1991), which were designed to deal specifically with this type of sampling. The aim of this project is to enable users across platforms to easily manipulate their data and interact with existing R packages. We have also provided functions to simulate data under a variety of scenarios to allow for rigorous testing of the robustness of the method and to facilitate further development of this approach.

Keywords: mark–recapture, non-invasive genetic sampling, population size, wildlife management

Received 5 July 2012; revision received 16 August 2012; accepted 20 August 2012

Introduction

Population size is a key demographic parameter and has important implications for the ecology, evolution, management and conservation of a species. Making effective decisions regarding the management of species, such as establishing protected areas or setting quotas for harvesting, requires reliable estimates of population size, and much research effort has been aimed at the development of robust statistical approaches to accomplish this (Otis *et al.* 1978). However, many of these methods (e.g. Program MARK: White & Burnham 1999) rely on the use of multi-session mark–recapture studies. Such an approach has many merits but is often not feasible for some natural populations, such as those of wide-ranging, elusive or rare species (Waits & Paetkau 2005). Furthermore, the species for which mark–recapture approaches are pragmatically difficult to apply are often the species that are of the highest concern for managers and conservation biologists. Technological advances have enabled the use of non-invasive genetic sampling as an alternative to capturing individual animals. With non-invasive genetic sampling, individuals can be identified by DNA obtained from scat, hair or other sources (e.g. feathers; for reviews see Waits & Paetkau 2005; Beja-Pereira *et al.* 2009).

Collecting DNA indirectly often results in an individual being detected multiple times per session requiring the use of alternative statistical approaches for estimating population size. In traditional mark–recapture studies, generally each individual can only be captured once per sampling session and individual capture histories are created across multi-session sampling (Otis *et al.* 1978). However, non-invasive genetic sampling is often conducted in a single session with individuals potentially being detected multiple times (Puechmaille & Petit 2007; Stenglein *et al.* 2010; Mowry *et al.* 2011).

Miller *et al.* (2005) developed an approach (which they called 'capwire', for '*Cap*ture *Wi*th *Re*placement') to

Correspondence: Craig R. Miller, Fax: 208-885-5843;
E-mail: crmiller@uidaho.edu

accommodate non-invasive genetic data that may be sampled in a single session or in multiple sessions (in which case all samples are pooled together). This method has been used to estimate population sizes in a variety of species (Arrendal *et al.* 2007; Puechmaille & Petit 2007; Hájková *et al.* 2009). We briefly review these methods here (for further details see Miller *et al.* 2005) and introduce a new R package `capwire` which implements these methods with additional functions.

The Miller *et al.* (2005) estimators were originally implemented in a stand-alone program written in Visual Basic. This had several limitations. First, it was only available for users working on a PC platform, making it inaccessible and inconvenient for many workers. Second, it was inflexible making it extremely difficult to modify or expand the software. Third, it was impossible for researchers to perform simulation studies to evaluate the robustness of the estimators. The R package we present in this paper is designed to overcome these issues and we hope that the increased accessibility and interoperability of our implementation will facilitate the further development of methods for estimating demographic parameters from non-invasive genetic sampling.

Materials and methods

The capwire estimators (Miller *et al.* 2005) assume that if a population can be modeled as an urn, then the sampling of an individual's DNA can be equated to drawing a ball at random from the urn. The sampled individual is then replaced and another draw is made. If all individuals are assumed to have an equal probability of being sampled (i.e. captured) on each draw, the likelihood $\mathcal{L}$ of obtaining the vector of capture counts $\vec{c} = (c_1, c_2, \ldots, c_T)$ for the T individuals sampled given a total poplulation size N is:

$$\mathcal{L}(N) = \left(\frac{N!}{T!(N-T)!}\right)\left(\frac{S!}{c_1!c_2!\ldots c_T!}\right)\prod_{i=1}^{T}(1/N)^{c_i} \quad \text{(eqn 1)}$$

where S is the total number of samples collected ($S = \sum_{i=1}^{T} c_i$). Miller *et al.* (2005) refer to this as the Equal Capture Model (ECM). However, heterogeneity in capture probabilities is common in non-invasive genetic data sets and this may be related to sex, age or social status or disparities between home range and the distribution of sampling (Piggott & Taylor 2003; Ebert *et al.* 2010). The assumption of equal-capture probabilities can be relaxed by allowing individuals to come from different rate classes (Miller *et al.* 2005). In the Two-Innate Rates Model (TIRM), the population is assumed to contain a mixture of two classes of individuals: those that are easy to capture (denoted N_A by Miller *et al.* 2005) and those that are difficult to capture (denoted N_B). The method does not rely on the researcher understanding the biological reasons for the heterogeneity in capture probabilities but only assumes that it exists. We can compare the fit of this model to the data to that of the ECM (the null model) with a likelihood-ratio test.

Here we present a R package that implements the methods of Miller *et al.* (2005). Using our package, researchers can fit the models to the data to obtain the maximum likelihood estimate (MLE) of the population size, perform a likelihood ratio test (LRT) to select between the ECM and TIRM models, perform a parametric bootstrap to estimate confidence intervals for the MLE and simulate data under these and other models. We have also implemented the 'partitioning' method PART (Stansbury 2012) to handle data sets containing more singletons and frequently captured individuals than the capwire models predict.

The functions for estimating population size in `capwire` require a data set containing the individual identifications, and the number of times each individual was captured. Our package can accommodate any single session sampling data, or multi-session data that has been collapsed into a single session. We caution researchers that the approaches implemented in `capwire` make several assumptions regarding population structure, sampling effort and the distributions of the count data. These assumptions and the robustness of this method to violations are discussed at length by Miller *et al.* (2005) and we again refer readers to this publication for details. The reliability of the estimates obtained from `capwire` depend on whether these assumptions can be justified for a particular data set.

An example

The following is a brief example demonstrating the utility and functionality of `capwire`. For the purposes of this example, we use some data collected from a population of grey wolves (*Canis lupis*) located in western Idaho, USA (Stansbury 2012). Individuals were identified using 8–9 microsatellite loci.

The data are input as a two-column `data.frame`. The first column corresponds to the capture classes (i.e. individuals in class i were captured i times) and the second column corresponds to the number of individuals in each class. Real data sets should be formulated in the same way although column headings do not need to be specified (see Supporting information for details).

```
> wolf.data
```

	capture.class	No.Ind
1	1	17
2	2	1
3	3	1
4	4	4
5	5	6
6	6	1

To fit the ECM to the data and obtain the MLE for the population size, we use the function `fitEcm`:

```
> res.ecm <- fitEcm(data=wolf.data, max.pop=200)
>res.ecm$ml.pop.size
[1] 33
```

The maximum population size is specified to be much larger than we presume the population to be but should take on a reasonable value; this needs to be specified to generate an upper bound for the purposes of optimization. If the data are uninformative (e.g. contains only singletons), there is no finite MLE for population size (Miller *et al.* 2005); in this case, the population estimators in `capwire` return the `max.pop` as the MLE (this is only likely to be the an issue in some simulated data sets). We can then find the MLE for population size under the TIRM.

```
>res.tirm <- fitTirm(data=wolf.data, max.pop=200)
>res.tirm$ml.pop.size
[1] 52
```

The fit of the two models can be compared using a LRT with the function `lrtCapwire`. A *P*-value can be calculated by using a parametric bootstrap approach to estimate the distribution of the LRT for data simulated under the less-parameterized model (ECM)

```
> test.stat <- lrtCapwire(ecm=res.ecm, tirm=res.tirm, bootstraps=100)
>test.stat
$LR
[1] 26.06849
$P.value
[1] 0
```

Here $P \simeq 0$ and we can reject the null hypothesis that the all individuals are equally likely to have been captured. We therefore will use the TIRM for the rest of the analyses. To obtain the 95% confidence intervals for the population size estimate, we perform a parametric bootstrap.

```
>  conf.int <- bootstrapCapwire(x=res.tirm, bootstraps=1000, CI=c(0.025, 0.975))
>conf.int
$ml.pop.size
[1] 52
$conf.int
2.5% 97.5%
41 74
```

In addition to the population size estimators, we have implemented a number of functions to simulate data under a variety of conditions. These will allow researchers to rigorously evaluate the robustness of the models implemented in `capwire`.

Discussion

Non-invasive genetic sampling is becoming an increasingly popular strategy for the study of population genetics and demography of natural populations (DeYoung & Honeycutt 2005; Marucco *et al.* 2011). It is also being utilized as a tool for making management decisions (Waits & Paetkau 2005; Beja-Pereira *et al.* 2009). By implementing the approach of Miller *et al.* (2005) in the R programming environment (R Development Core Team 2012), we aim to make it accessible to users across platforms and allow for easy data manipulation and graphical display of the data. Second, we encourage future developments in population estimation using non-invasive genetic data which incorporate more complex sampling schemes and hope that the flexibility of this package will facilitate this. Third, the capwire methods make several simplifying assumptions and simulation studies investigating the robustness of these assumptions under a variety of more realistic scenarios are needed to further understand the limitations and applicability of these approaches. Last, `capwire` could be extended in multiple ways, such as making the model spatially explicit by incorporating the geospatial data associated with each sample. This could potentially be carried out by linking up `capwire` with R packages developed for the analysis of spatial data (e.g. `spatstat`: Baddeley & Turner 2005; `secr`: Efford 2012) and by developing more complex sampling models. As the cost of obtaining genetic data continues to decrease and improvements are made in analytic techniques (Macbeth *et al.* 2011), non-invasive genetic sampling can be more easily adopted by conservation biologists and wildlife managers. The continued development of statistical methods designed for this type of data is an important area of research.

Notes on implementation and documentation

`capwire` is available on the CRAN repository http://cran.r-project.org/web/packages/capwire/. A tutorial demonstrating the usage of `capwire` is included as Supporting information to this paper and is also available on

the personal website of M.W.P. http://mwpennell.wordpress.com/software/.

Acknowledgements

We thank Paul Joyce for providing insight into these methods. We also thank Frederic Austerlitz and three anonymous reviewers for helpful comments on this manuscript and Andrea Taylor for permission to include her data set in the R package. M.W.P. thanks Luke Harmon for encouragement and support. M.W.P. was funded by NSF grant DEB-0919499. C.R.S. was funded by the Nez Perce Tribe, the Idaho Department of Fish and Game and the University of Idaho Student Grant Program.

References

Arrendal J, Vilá C, Björklund M (2007) Reliability of noninvasive genetic census of otters compared to field censuses. *Conservation Genetics*, **8**, 1097–1107.

Baddeley A, Turner R (2005) spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, **12**, 1–42.

Beja-Pereira A, Oliveira R, Alves PC, Schwartz MK, Luikart G (2009) Advancing ecological understandings through technological transformations in noninvasive genetics. *Molecular Ecology Resources*, **9**, 1279–1301.

DeYoung RW, Honeycutt RL (2005) The molecular toolbox: genetic techniques in wildlife ecology and management. *The Journal of Wildlife Management*, **69**, 1362–1384.

Ebert C, Knauer F, Ilse S, Hohmann U (2010) Individual heterogeneity as a pitfall in population estimates based on non-invasive genetic sampling: a review and recommendations. *Wildlife Biology*, **16**, 225–240.

Efford MG (2012) secr: spatially explicit capture–recapture models. R package version 2.3.2.

Hájková P, Zemanová B, Roche K, Hájek B (2009) An evaluation of field and noninvasive genetic methods for estimating Eurasian otter population size. *Conservation Genetics*, **10**, 1667–1681.

Macbeth GM, Broderick D, Ovenden JR, Buckworth RC (2011) Likelihood-based genetic mark–recapture estimates when genotype samples are incomplete and contain typing errors. *Theoretical Population Biology* **80**, 185–196.

Marucco F, Boitani L, Pletscher D, Schwartz M (2011) Bridging the gaps between non-invasive genetic sampling and population parameter estimation. *European Journal of Wildlife Research*, **57**, 1–13.

Miller CR, Joyce P, Waits LP (2005) A new method for estimating the size of small populations from genetic mark-recapture data. *Molecular Ecology*, **14**, 1991–2005.

Mowry RA, Gompper ME, Beringer J, Eggert LS (2011) River otter population size estimation using noninvasive latrine surveys. *The Journal of Wildlife Management*, **75**, 1625–1636.

Otis DL, Burnham KP, White GC, Anderson DR (1978) Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, **62**, 3–135.

Piggott MP, Taylor AC (2003) Remote collection of animal DNA and its applications in conservation management and understanding the population biology of rare and cryptic species. *Wildlife Research*, **30**, 1–13.

Puechmaille SJ, Petit EJ (2007) Empirical evaluation of non-invasive capture–mark–recapture estimation of population size based on a single sampling session. *Journal of Applied Ecology*, **44**, 843–852.

R Development Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Stansbury CR (2012) *Monitoring Gray Wolves (Canis lupus) Using Noninvasive Genetic Sampling at Rendezvous Sites*. Master's thesis, University of Idaho, Moscow, Idaho.

Stenglein JL, Waits LP, Ausband DE, Zager P, Mack CM (2010) Efficient noninvasive genetic sampling for monitoring reintroduced wolves. *The Journal of Wildlife Management*, **74**, 1050–1058.

Waits LP, Paetkau D (2005) Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. *The Journal of Wildlife Management*, **69**, 1419–1433.

White GC, Burnham KP (1999) Program MARK: survival estimation from populations of marked animals. *Bird Study*, **46**, S120–S139.

M.W.P. wrote the majority of the code for the R package with coding contributions from C.R.M. C.R.M. and M.W.P. developed the PART algorithm. M.W.P., C.R.M., C.R.S., and L.W.P. wrote the paper. The online supplement was written by M.W.P. and C.R.M. C.R.S. and L.W.P. motivated the work by providing data where capwire assumptions were clearly violated.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. capwire tutorial.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.