Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 299–303 doi: 10.1111/2041-210x.12025

APPLICATION

Sample Planning Optimization Tool for conservation and population Genetics (SPOTG): a software for choosing the appropriate number of markers and samples

Sean Hoban[1]*, Oscar Gaggiotti[2]†, ConGRESS Consortium‡ and Giorgio Bertorelle[1]

[1]*Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, Via L Borsari 46, Ferrara, 44100, Italy; and*
[2]*Université Josef Fourier, Grenoble, 38041, France*

Summary

1. Genetic data are frequently used to make inferences about evolutionary and ecological processes, but the choice of the number of genetic markers and samples for such studies is usually *ad hoc*. Unfortunately, suboptimal sampling routinely leads to ambiguous results.

2. SPOTG is a user-friendly software for optimizing sampling strategy for five common genetic study topics: hybridization, temporal sampling, bottlenecks, connectivity and assignment. SPOTG facilitates formal evaluation of the expected statistical power of proposed sampling strategies before project implementation, by using stochastic genetic simulations of realistic population scenarios and various sampling schemes.

3. We demonstrate use of the tool with two example species (lynx and bison) in which demographic history differs; the appropriate sampling strategy for detecting a genetic bottleneck differs dramatically between the two cases, with important implications for sample planning.

4. SPOTG has an interactive graphical tool for exploring results, and extensive documentation, tips and tutorials to enable use by conservation managers, ecologists beginning to use genetics and students.

Key-words: data analysis, statistical power, conservation interventions, monitoring, simulation, molecular ecology, management

Introduction

Population and conservation genetic studies routinely utilize genotypic data to detect processes such as hybridization or migration. However, choosing the sampling strategy for these studies (number of individuals and populations, number and type of molecular markers) is often *ad hoc*. Some studies obtain as many markers and samples as possible, after which, "statistical power (or 'resolving power') is expected to be 'high' because of the use of large sample sizes, many loci, or some particular type of genetic marker" (Ryman & Palm 2006). Other studies follow rules-of-thumb, such as >30 individuals for estimating differentiation (Ward & Jasieniuk 2009). General guidelines exist for some study goals. For example, one bottleneck detection test recommends 5–20 polymorphic loci and 20–30 individuals (Luikart & Cornuet 1998), while suggestions for a common assignment test are 10 loci and 30–50 individuals for highly differentiated populations (Manel, Berthier & Luikart 2002). However, these guidelines are based on previous experience on particular species, or simulations that explore few, general cases, and cannot be considered comprehensive guidelines for all possible population systems, marker characteristics, etc. Generally, the probability of detecting a given genetic pattern depends on the number of markers and samples (of which many combinations are possible), and on the strength of the genetic pattern (Ward & Jasieniuk 2009). For example, detection of significant differences in allele frequencies requires larger sampling efforts when divergence is low (Kalinowski 2005), as does assignment of individuals to source populations. Chances of bottleneck detection depend on timing and severity of the bottleneck (Girod *et al.* 2011). Even the simple question, 'Is statistical power increased more by doubling the number of markers or the number of individuals?' has different answers depending on the type of markers, the goal, and the level of polymorphisms (Heled & Drummond 2008; Morin, Martien & Taylor 2009). Thus, the appropriate sampling strategy for a molecular ecology study depends on many factors.

Due to limited resources, scientific investigators (and funding agencies) could use knowledge of the expected statistical power (hereafter, simply 'power') of potential sampling strategies, when planning a study, to ensure that sampling is sufficient (high probability of detecting a genetic pattern of interest, e.g. $P > 0{\cdot}90$) and efficient (Peterman 1990). Avoiding oversampling allows more resources to be allocated to other

*Correspondence author. E-mail: shoban@alumni.nd.edu
†Present address: School of Biology, Scottish Oceans Institute, University of St Andrews, St Andrews, Fife, KY16 8LB, UK
‡Members of ConGRESS are listed in online supplemental materials

projects (and minimizes disturbance of sensitive natural populations). Avoiding undersampling helps prevent negative or ambiguous results (Swatdipong, Primmer & Vasemägi 2010). For example, some shallow bottlenecks may be undetectable under any feasible sampling scheme (Girod *et al.* 2011), in which case a study is likely to be fruitless (see example, below).

The expected relationship between power and sampling strategy for particular studies is quantifiable, but this remains a rare component of molecular ecology studies (Ryman & Palm 2006; Ward & Jasieniuk 2009). Thus, it is not clear how many published studies have optimal sampling, and even less clear how many studies were not published because suboptimal sampling led to inconclusive results. An estimate of power of a given sampling strategy can be made with forward or backward simulation software over a restricted parameter space (Ryman & Palm 2006; Swatdipong, Primmer & Vasemägi 2010) with the following procedure. First, propose a limited number of population/evolutionary histories and sampling strategies (hereafter, a 'scenario'). Then for each scenario: (i) perform a simulation (a simplified representation of the real world) of the populations, (ii) sample individuals and obtain their genotypes, (iii) with genetic data, perform the analytical test(s) that would be performed in the real study, (iv) reach a conclusion based on this test (e.g. a bottleneck did occur, or FST = 0·05), and (v) determine the error from the simulation (is the conclusion correct, how far from the true value is the estimate). This process is performed many times to incorporate genetic, demographic and sampling stochasticity (Marjoram & Tavare 2006; Hoban, Bertorelle & Gaggiotti 2012a; Hoban *et al.* 2012b). Based on many replicates, an estimate of the statistical power of a data set is obtained (an expectation of the relative performance of different strategies to detect particular genetic effects in real population systems).

This is how simulations are used to test new analytical procedures, and how *post hoc* power is estimated to determine the meaning of non-significant results. Many simulation software are available (Hoban, Bertorelle & Gaggiotti 2012a; Hoban *et al.* 2012b), so why is quantification of power not a routine part of planning a genetic study? First, constructing simulations often requires complex input files and numerous parameters. Second, the process requires substantial automation (repeated simulations, analysing simulated data, collating results across replicates), and thus some bioinformatics. A user-friendly software for calculating power would benefit the population genetics community (and conservation managers, ecologists who are beginning to use genetics and students) as it would encourage and facilitate power calculation before performing genetic studies.

We present SPOTG, a software for estimating statistical power for five common genetic study goals: connectivity, bottlenecks, assignment tests, hybridization and temporal sampling/monitoring. This software is user-friendly for entering parameters and interpreting results (Fig. 1), and is flexible and realistic (e.g. arbitrary number of populations, asymmetrical migration

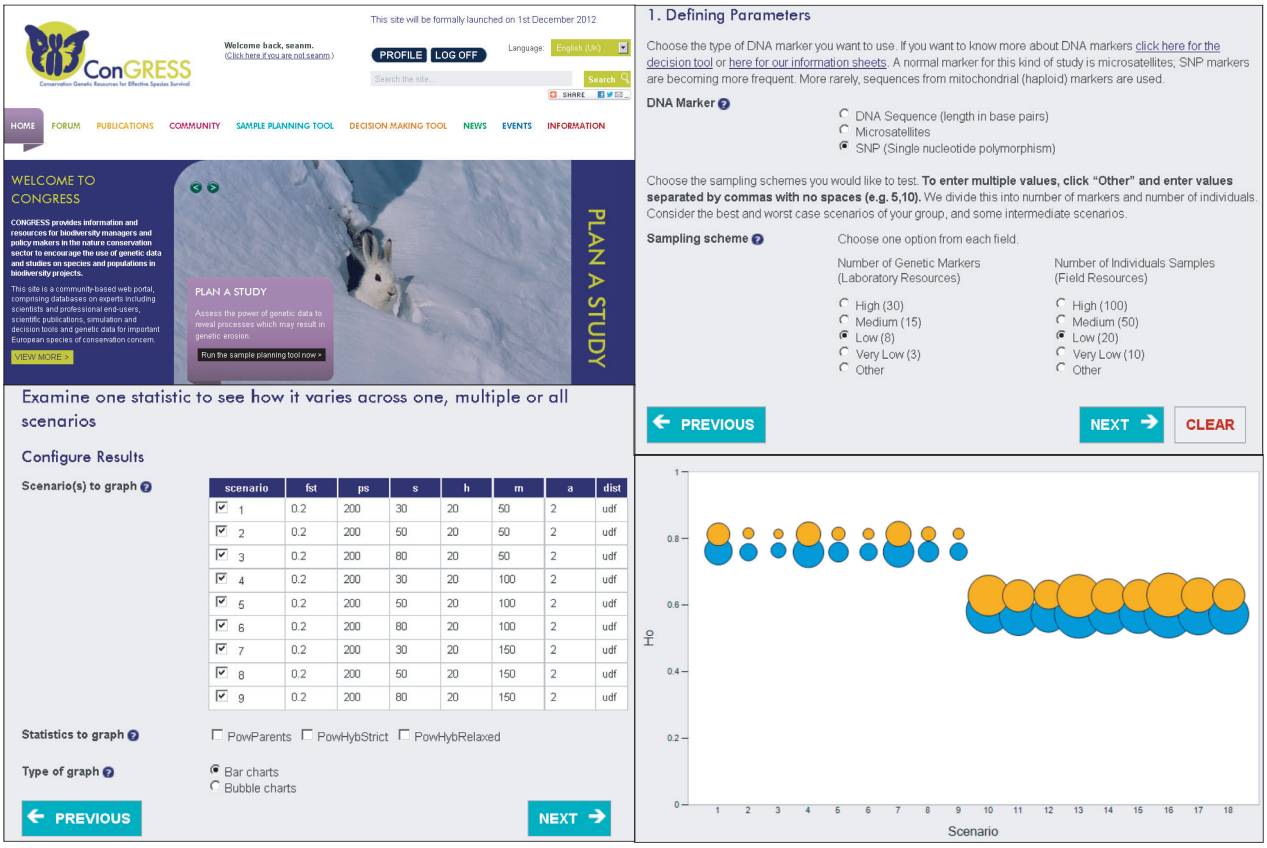

Fig. 1. Screenshots (clockwise from upper left) for homepage, parameter entry for connectivity module, example results, graphing tool.

rates). One software, POWSIM (Ryman & Palm 2006) exists for estimating the power to detect significant genetic differentiation. SPOTG is a major advance because POWSIM is only for connectivity studies, and only simulates two populations of equal size. SPOTG has been developed in the context of the ConGRESS (Conservation Genetic Resources for Effective Species Survival) project, and is connected to an online community of conservation practitioners.

Software details

There are five modules, each aimed at evaluating power for one particular goal or analysis. Below we briefly describe the simulation methods and statistical analyses implemented. More details and examples are provided in the online software documentation. For each module, the output is a measure of the accuracy of an estimate or inference under the sampling schemes and scenarios assumed by the simulations, which represents the expected probability of making a correct conclusion in a real study.

BOTTLENECK

A common use of genetic data is to detect past fluctuations in population size, such as to reconstruct evolutionary history or to determine the impact of human exploitation. This module allows simulation of various bottleneck scenarios to determine the power of a proposed study to detect bottlenecks.

Data are simulated with the coalescent simulator SIMCOAL2 (Laval & Excoffier 2004; with permission) under simple bottleneck scenarios defined by five parameters for evolutionary history (pre-, during-, and post-bottleneck population size, and timing of the bottleneck and recovery) plus two parameters for sampling (number of markers and sampled individuals). Data are also simulated for a population of constant size, with the same sampling. Specifically, the M-ratio (Garza & Williamson 2001), a statistic sensitive to demographic decline, is calculated using arlecore (Excoffier & Lischer 2010), for every replicate of the bottleneck and non-bottlenecked scenarios. Based on many simulations, power equals the proportion of bottleneck replicates that result in M-ratios smaller than the lower 5% of the non-bottlenecked population distribution.

CONNECTIVITY

A common topic in genetics is population divergence, such as caused by low gene flow, small population size, or selection. A reliable estimate of genetic connectivity can reveal the influence of human-made barriers, local environmental conditions and other evolutionary and ecological processes. We allow simulation of a simple or complex population system: an arbitrary number of populations in an island model or a series of up to seven populations, each having different size and specific migration rates to each of the others.

Data are simulated under the population sizes, migration rates and sampling parameters defined by the user. The goal is to determine power and accuracy of a measure of genetic divergence calculated from the sample. Therefore, a scenario is simulated where the entire population is genotyped at 100 markers, providing an FST estimate based on exact allele frequencies (no sampling error), which is the 'truth' against which error is calculated. We use arlecore to calculate FST and to perform an exact test of population differentiation, for sampled and 'truth' situations. Power is the proportion of replicates that correctly identify significant differentiation when it exists. We also report relative standard deviation, across replicates, of FST (reported for each population pair and globally).

ASSIGNMENT

Assignment of an individual to its source (natal) population is used for providing evidence of poaching, detecting fish escaped from hatcheries and quantifying recent migration events. The procedure is to match the genotype of a focal individual to the region or population it came from (assignment), out of a series of possible, sampled populations, or to safely exclude all sampled populations, if it did not come from any (exclusion).

Assignment tests depend on polymorphism of the markers and shape of allele frequency distributions (Manel, Berthier & Luikart 2002), so the user provides global allele frequencies (or chooses from a distribution), and degree of divergence (FST) for possible source populations. *In-silico* population allele frequencies are 'built' using the sampling formula for FST (Balding & Nichols 1997; Gaggiotti *et al*. 2004), which generates frequency distributions for each deme of a subdivided population with chosen genetic differentiation and marker characteristics. Once population allele frequencies are defined, a number of alleles (sample size, times two for diploids) is sampled probabilistically, and genotypes are 'created' by randomly combining alleles. Simple assignment and exclusion tests are performed (Paetkau *et al*. 1995) for every individual. We report the mean and standard deviation (across replicates) of three types of errors (mis-assignment, incorrect inclusion and incorrect exclusion).

HYBRIDIZATION

Hybridization can have short and long-term consequences for a species' genetic makeup, and is a conservation concern between rare and common species, and between native and non-native species (Hoban, Bertorelle & Gaggiotti 2012a; Hoban *et al*. 2012b). On the other hand, intra- and inter- specific hybridization may be used to introduce genes for resistance to introduced pests and diseases. In either case, detecting hybrid individuals is a common task in molecular ecology.

Using the sampling formula for FST (see above), the allele frequencies of two populations/species are created with defined divergence. First-generation hybrid genotypes are created by simulated random mating between the populations/species.

Second-generation hybrids are created by simulated random mating of F1 hybrids. Assignment tests are used to assign individuals to parent, F1, or F2 groups, and several error rates are reported (strict and relaxed hybrid misidentification, and parental misidentification).

TEMPORAL SAMPLING

The use of DNA samples from the past (ancient DNA, museum/herbaria samples) can provide insight into genetic diversity in the distant or recent past. We focus on detecting a significant decline in diversity between two temporally spaced samples.

The user enters parameters similar to the bottleneck module, plus the number and time point of temporal samples. For this module we use fastsimcoal (Excoffier & Foll 2011), which allows the simulation of modern and ancient samples along the same coalescent tree. Similar to the bottleneck module, a population of constant size and a declining population are simulated. For each case, the difference in number of alleles and heterozygosity between the samples at the two time points is calculated. Power is the proportion of differences from the population decline that fall below the lower 5% of the differences from the constant population.

Example use

We present an example use of the software for planning a study to detect genetic bottlenecks in two species for which the bottleneck timing and severity differ: a recent, strong decline with no recovery, and an older, drastic decline with some recovery. We assume some demographic information is available suggesting a possible demographic bottleneck, but its severity and genetic impact are not known. The Iberian lynx has shown continual decline over the last century due to disease in its main prey (rabbits), and habitat loss. In 1960, there were 5000 lynx, and now there are <150. The European bison experienced an extended decline over 500 years due to hunting and habitat loss. A small herd was founded by the last 12 bison around 1920. After captive breeding and reintroduction, there are now >1500 free-living bison.

The first step is to propose parameters for the bottleneck model. This example will propose a strong and a weak bottleneck for each species, to incorporate uncertainty about pre-bottleneck population sizes (corresponding to liberal and conservative estimates of power). For the lynx, plausible values for pre-bottleneck effective population size (Ne) are 6500 and 1500, whereas for the bison, plausible values are 50 000 and 5000. Current population sizes are known. We propose nine combinations of markers and samples for each situation.

From the example results (Fig. 2), we emphasize the following points.

1 Gain of power with sampling effort is often nonlinear. Thus, thresholds can be identified, where increased sampling does not substantially improve power. Thresholds are most clear for the bison, in which sampling > 10 markers and 20 samples is unnecessary even for the conservative scenario.

2 Gain of power with sampling effort is species- and situation-specific

- **a.** In some cases, the genetic effect cannot be detected for any realistic sampling. For the conservative scenario for the lynx, power never exceeds ~0·60, even with 50 markers and 100 individuals.
- **b.** In other cases, small sampling effort can suffice (e.g. 10 individuals and 5 markers, much smaller than is typical for molecular ecology studies).

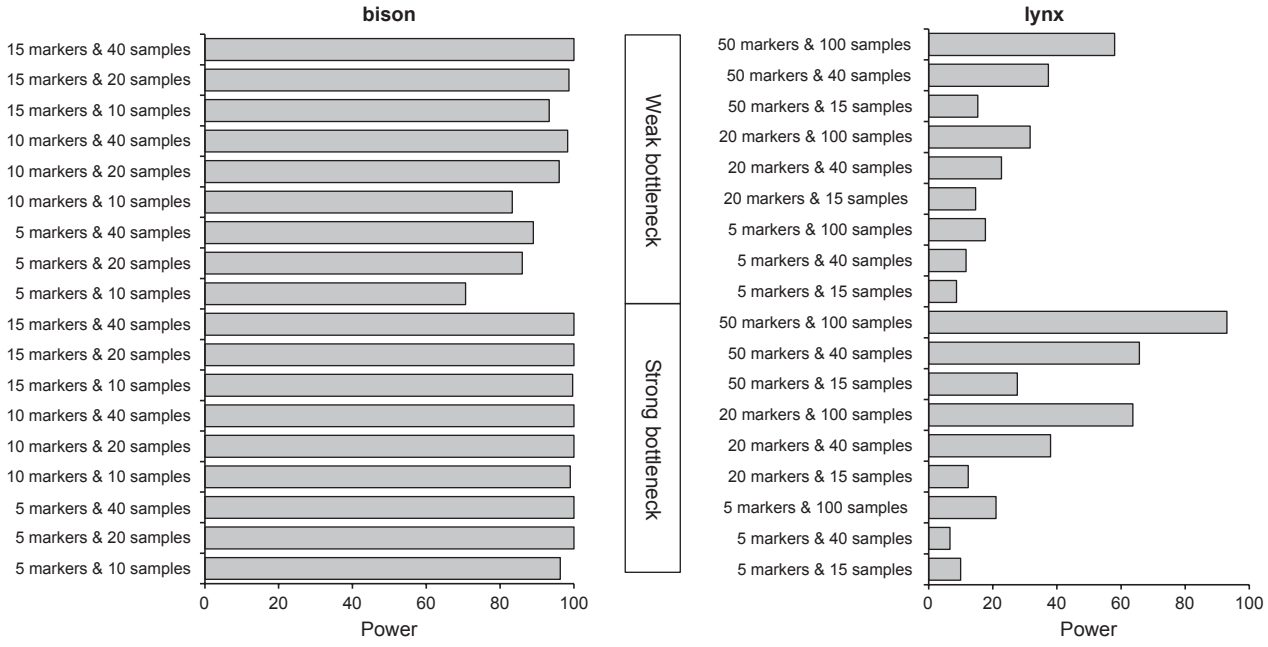

Fig. 2. Power to detect a bottleneck under conservative and liberal scenarios for nine combinations of markers and sampled individuals, for two species.

3 Sometimes more markers are more effective, sometimes more samples.

For practical use of the tool, we suggest working in two steps. First, propose a series of scenarios with several values for each major parameter, to determine whether or not any particular parameter has a strong effect. This is also important when little is known a priori about the species' biology and history. Then, in a second stage, set some parameters at constant values (those with less effect), while varying other parameters. Also, for our example, it would be worthwhile to further simulate intermediate values for the lynx (e.g. 20 samples, 60 samples). For pre-project planning, the focus is to try a few evolutionary histories and many combinations of sampling strategy, whereas for post-project power analysis, one sampling strategy is proposed (that which was used), along with many, varied evolutionary histories.

Conclusion

SPOTG is written in Java, and has a graphical web-based interface (http://www.congressgenetics.eu, command-line version available at https://sites.google.com/site/hoban3/scripts), so it can run on any computer, without downloading/installation (sometimes restricted on shared or agency computers). There are tips, tutorials and guides to interpreting results, so SPOTG is usable by conservation managers and graduate students, and also for teaching purposes. The documentation provides guidance on choosing appropriate parameters such as mutation rate, effective population size and migration rates, but collaboration or supervision with an experienced population geneticist may be useful.

SPOTG should be used as a rough guide for study planning, to be complemented with expert knowledge. Scenarios are simplified from the real world (e.g. cannot specify mating patterns or overlapping generations). Analyses are also basic, but are employed in many population genetics studies (e.g. FST, assignment). More sophisticated analyses can extract more information from the data and, therefore, may perform better under the sampling schemes identified as optimal by SPOTG. A future improvement could be to incorporate costs of developing markers, genotyping samples and field sampling.

The simulation approach to project planning (to simulate a given effect from a biological process, simulate collection of data and analysis to measure that effect with different sampling schemes) can be applied in many studies, even outside genetic studies (Peterman 1990).

Acknowledgements

We thank Laurent Excoffier for permission to incorporate SIMCOAL2, Arlecore and fastsimcoal into SPOTG. We also thank two anonymous reviewers, and many volunteers who tested the software. This work was funded by European Commission grant FP7-ENV-2009-1 244250 (Knowledge Transfer and Uptake of EU Research Results). We declare no conflicts of interest.

References

Balding, D.J. & Nichols, R.A. (1997) Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity*, **78**, 583–589.

Excoffier, L. & Foll, M. (2011) Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, **27**, 1332–1334.

Excoffier, L. & Lischer, H.E.L. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Gaggiotti, O.E., Brooks, S.P., Amos, W. & Harwood, J. (2004) Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Molecular Ecology*, **13**, 811–825.

Garza, J.C. & Williamson, E.G. (2001) Detection of reduction in population size using data from microsatellite loci. *Molecular Ecology*, **10**, 305–318.

Girod, C., Vitalis, R., Leblois, R. & Freville, H. (2011) Inferring population decline and expansion from microsatellite data: a simulation-based evaluation of the Msvar method. *Genetics*, **188**, 165–179.

Heled, J. & Drummond, A.J. (2008) Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, **8**, 289.

Hoban, S., Bertorelle, G. & Gaggiotti, O.E. (2012a) Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, **73**, 2–14.

Hoban, S., McCleary, T., Schlarbaum, S., Anagnostikas, S. & Romero-Severson, J. (2012b) Human impacted landscapes facilitate hybridization between a native and an introduced tree species. *Evolutionary Applications*. **5**, 720–731.

Kalinowski, S.T. (2005) Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity*, **94**, 62–65.

Laval, G. & Excoffier, L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–7.

Luikart, G. & Cornuet, J. (1998) Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology*, **12**, 228–237.

Manel, S., Berthier, P. & Luikart, G. (2002) Detecting wildlife poaching: identifying the origin of multilocus genotypes. *Conservation Biology*, **16**, 650–659.

Marjoram, P. & Tavare, S. (2006) Modern computational approaches for analysing molecular genetic variation data. *Nature Reviews Genetics*, **7**, 759–770.

Morin, P.A., Martien, K.K. & Taylor, B.L. (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Paetkau, D., Calvert, W., Stirling, I. & Strobeck, C. (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Peterman, R.M. (1990) Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 2–15.

Ryman, N. & Palm, S. (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology*, **6**, 600–602.

Swatdipong, A., Primmer, C. & Vasemägi, A. (2010) Historical and recent genetic bottlenecks in European grayling, Thymallus thymallus. *Conservation Genetics*, **11**, 279–292.

Ward, S.M. & Jasieniuk, M. (2009) Review: sampling weedy and invasive plant populations for genetic diversity analysis. *Weed Science*, **57**, 593–602.

Received 22 October 2012; accepted 16 December 2012
Handling Editor: Robert B. O'Hara

Supporting Information

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

Data S1. The ConGRESS Consortium.