2013

Microsatellite Data Analysis for Population Genetics

Kyung Seok Kim
Seoul National University

Thomas W. Sappington
U.S. Department of Agriculture, tsapping@iastate.edu

Microsatellite Data Analysis for Population Genetics

Abstract
Theories and analytical tools of population genetics have been widely applied for addressing various questions in the fields of ecological genetics, conservation biology, and any context where the role of dispersal or gene flow is important. Underlying much of population genetics is the analysis of variation at selectively neutral marker loci, and microsatellites continue to be a popular choice of marker. In recent decades, software programs to estimate population genetics parameters have been developed at an increasing pace as computational science and theoretical knowledge advance. Numerous population genetics software programs are presently available to analyze microsatellite genotype data, but only a handful are commonly employed for calculating parameters such as genetic variation, genetic structure, patterns of spatial and temporal gene flow, population demography, individual population assignment, and genetic relationships within and between populations. In this chapter, we introduce statistical analyses and relevant population genetic software programs that are commonly employed in the field of population genetics and molecular ecology.

Keywords
Population genetics, Genetics software, Genetic variation, Genetic structure, Gene flow, Microsatellites

Comments
This is a chapter from Kim, Kyung Seok, and Thomas W. Sappington. "Microsatellite data analysis for population genetics." In Microsatellites, pp. 271-295. Humana Press, Totowa, NJ, 2013. doi: 10.1007/978-1-62703-389-3_19. Posted with permission.

Chapter 19

Microsatellite Data Analysis for Population Genetics

Kyung Seok Kim and Thomas W. Sappington

Abstract

Theories and analytical tools of population genetics have been widely applied for addressing various questions in the fields of ecological genetics, conservation biology, and any context where the role of dispersal or gene flow is important. Underlying much of population genetics is the analysis of variation at selectively neutral marker loci, and microsatellites continue to be a popular choice of marker. In recent decades, software programs to estimate population genetics parameters have been developed at an increasing pace as computational science and theoretical knowledge advance. Numerous population genetics software programs are presently available to analyze microsatellite genotype data, but only a handful are commonly employed for calculating parameters such as genetic variation, genetic structure, patterns of spatial and temporal gene flow, population demography, individual population assignment, and genetic relationships within and between populations. In this chapter, we introduce statistical analyses and relevant population genetic software programs that are commonly employed in the field of population genetics and molecular ecology.

Key words Population genetics, Genetics software, Genetic variation, Genetic structure, Gene flow, Microsatellites

1 Introduction

Population genetics is the study of the frequency and interaction of alleles and genes in populations. It has revolutionized many fields of evolutionary biology over the last 30 years and represents the essence of the modern evolutionary synthesis. Allele frequency in populations can change spatially and temporally under the influence of various evolutionary processes, particularly natural selection, genetic drift, mutation, gene flow, and mating system. Comparative analyses of spatial and temporal patterns in allele frequency provide an important entry point to identify the evolutionary forces that gave rise to them. In large part, however, the fundamental power and premise behind population genetics is that one can compare allele frequencies at selectively neutral marker loci to estimate gene flow, under reasonable assumptions about the rate of drift, mutation, and mating system. Gene flow is a parameter of critical importance in studies related to wildlife conservation, fisheries management,

Stella K. Kantartzi (ed.), *Microsatellites: Methods and Protocols*, Methods in Molecular Biology, vol. 1006,
DOI 10.1007/978-1-62703-389-3_19,

invasion biology and routes of invasion, insect resistance management, pest management, pest eradication programs, population and metapopulation dynamics, phylogeography, biosystematics, and many others. Gene flow is tied in a fundamental way to effective dispersal of the individuals carrying the genes, and information obtained about either for a given species informs understanding of the other. Thus, population genetics analyses have been widely used for examining patterns and magnitude of animal dispersal over both geographic (e.g., 1–5) and temporal dimensions (6).

The ability of population genetics to deliver on its promise of elucidating gene flow has relied on development of suitable molecular markers and population genetics theory for making robust inferences from observed variation in marker loci. Advances in population genetics theory and adoption of new types of molecular markers have been accompanied over the decades by parallel creation and improvement of analytical software programs to effectively calculate population genetics parameters. Although numerous software programs are available, only a handful have been used routinely in studies of natural populations. Microsatellites are a very popular marker for population genetics studies, in part because the abundance of alleles per locus and the ability to distinguish heterozygotes enhance their information content over many other types of markers. This chapter introduces the basic operating procedures of the software programs most commonly used to analyze microsatellite genotype data, including an overview of data formats and parameters for each. We also introduce statistical tests routinely used in the field of population genetics and molecular ecology. Issues commonly encountered and to watch for when conducting genetic analyses based on genotypic data from microsatellite markers are discussed, along with suggestions for troubleshooting.

2 Materials

Some population genetics software programs are designed for comprehensive statistical analyses, but many were specifically produced to calculate particular parameters. The best options for conducting certain types of analyses will often depend on the specific nature of the user's project and the models that the user is assuming. Many programs generate the same or similar population genetics parameters, so the choice of a particular program will depend on personal preference or availability in the user's lab. Nevertheless, basic genetic analyses in empirical population genetics and molecular ecology employ a similar framework across users.

Population genetics analyses based on a microsatellite genotype dataset can be categorized into three sequential phases: (1) Initial data manipulation, including error-checking of the raw genotype dataset and generating correctly formatted input files for other programs; (2) basic genetic analyses for obtaining summary

statistics of common population genetics parameters; and (3) advanced genetic analyses for addressing specific questions or hypothesis testing.

Most of the population genetics software programs in this chapter can be downloaded free of charge from the websites listed in Table 1. Some software programs, e.g., GenAlEx and Microsatellite Toolkit, operate within Microsoft Excel (on Macs and PCs). However, other software such as Arlequin, BOTTLENECK, FSTAT, Genepop, GeneClass, Micro-Checker, and STRUCTURE operate in their own user-friendly platform environments, e.g., Dos and Java. It is important to look over the websites on a regular basis because they are often revised or updated by their curators.

2.1 Formatting and Data Manipulation

Micro-Checker (7) and Microsatellite Toolkit (8) are software programs that can be used for the beginning step of data manipulation. One of the purposes of these software programs is to detect scoring errors and to confirm that the genotype file is correct. Since accurate genotypes are critical for generating reliable results in further statistical analyses, proper and efficient use of software in phase 1 is very important. Microsatellite Toolkit offers additional functions, including the generation of an input file for other programs as an export data option. GeneAlEx (9) also generates input files for many other programs. This capacity to create correctly formatted input files is very useful, because downstream analyses in other programs then become largely a matter of strategy and interpretation of output. That is why much of this chapter, after the first phase of data manipulation is described, is concerned mainly with analytic strategies to obtain desired population genetics output from a microsatellite genotype dataset.

2.2 Basic Population Genetic Analyses

With a correctly formatted genotype file, one can proceed with the basic population genetics analyses of phase 2. Software programs such as Arlequin (10), Cervus (11), FSTAT (12), GeneAlEx (9), Genepop (13), and others provide options for calculating genetic diversity, genetic differentiation, gene flow, partitioning of genetic variation, and so on.

2.3 Advanced Population Genetic Analyses

If the user requires additional population analyses to test specific hypotheses such as population demography, individual/population genetic relationships, isolation by distance, genetic structuring, relatedness, or individual assignment/exclusion, one can use advanced programs such as BOTTLENECK (14), STRUCTURE (15), GeneClass (16), and GeneAlEx (9). Although not every population genetics study requires all such analyses, most studies can benefit from one or more of them. They are useful options to have in one's analytical toolbox, and we present the most commonly used.

Table 1
Characteristics and website information for downloading free software for population genetics studies

Software	Main use	Work environment	Website for download	References
Micro-Checker	Checks for microsatellite null alleles and scoring errors	Unique platform, Windows	http://www.microchecker.hull.ac.uk/	(7)
Microsatellite Toolkit	Error-checking Basic parameters such as diversity measures Input file for other programs	Unique platform in MS Excel, Windows	http://animalgenomics.ucd.ie/sdepark/ms-toolkit/	(8)
GenAlEx	Population genetics software package including import, management, and export of data	Unique platform in MS Excel, Windows	http://www.anu.edu.au/BoZo/GenAlEx/	(9)
Genepop	Population genetics software package, export of data	Unique platform, Windows or Mac	http://genepop.curtin.edu.au/	(13)
FSTAT	Population genetics software package to estimate and test gene diversities and differentiation statistics	Unique platform, Windows	http://www2.unil.ch/popgen/softwares/fstat.htm	(12)
BOTTLENECK	Detects recent effective population size reductions	Unique platform, Windows	http://www.montpellier.inra.fr/URLB/bottleneck/bottleneck.html	(14)
Arlequin	Population genetics software package	Unique platform, Windows or Mac	http://cmpg.unibe.ch/software/arlequin35/	(10)
STRUCTURE	Investigates population structure	Unique platform, Windows	http://pritch.bsd.uchicago.edu/structure.html	(15)
GeneClass	Selects or excludes populations as origins of individuals	Unique platform, Windows	http://www.montpellier.inra.fr/URLB/index.html	(16)

FreeNA	Estimates null allele frequencies Estimates unbiased F_{ST}	Unique platform, Windows	http://www.montpellier.inra.fr/URLB/index.html	(22)
Microsat	Estimates distance measures and diversity indices by population or individual	Unique platform, Windows	http://hpgl.stanford.edu/projects/microsat/	(19)
AGARST	A program for calculating allele frequencies, Gst and Rst from microsatellite data, plus a number of other population genetic estimates	Unique platform, Windows	Requires personal contact with the authors	(17)
DISPAN	Estimates genetic diversity, genetic distance Conducts phylogenetic analysis	Unique platform, Windows	http://www.softsea.com/review/DISPAN-Genetic-Distance-and-Phylogenetic-Analysis.html	(18)
Cervus	Provides statistical method for diversity indices and parentage analysis	Unique platform, Windows	http://www.fieldgenetics.com/pages/home.jsp	(11)
Migrate	Estimates effective population size and past migration rates between populations	Unique platform, Windows or Mac	http://popgen.sc.fsu.edu/Migrate/Migrate-n.html	(20)
RSTCALC	Calculates unbiased estimates of genetic differentiation (Rst, analogous to Fst), with their significance, for microsatellite data	Unique platform, Windows	http://www.biology.ed.ac.uk/archive/software/rst/rst.html	(21)

3 Methods

Manipulation of the genotype dataset and generation of correct input files for analytical software programs are the important initial steps in population genetics analyses. We do not describe detailed population genetics theories and assumptions underlying the specific genetic analyses in the software programs. For this information, it is highly recommended that the user read the information file included in each program's website (Table 1) and the papers it cites.

3.1 Input File and Correct File Extension for Each Program

Figures 1–12 illustrate correctly formatted input files for most of the population genetics software programs described in this chapter. Instructions on formatting are provided on the programs' respective websites (Table 1). Each input file contains the same genotype data for a total of ten individuals from four populations (two individuals for popA, three individuals for popB, two individuals for popC, three individuals for popD) at five microsatellite loci.

- Micro-Checker (7) (Fig. 1): Genepop format with a 3-digit number. Generated from Microsatellite Toolkit.
- Microsatellite Toolkit (8) (Fig. 2): Requires genotype file in Excel.
- GenAlEx (9) (Fig. 3): Requires genotype file in Excel.
- Arlequin (10) (Fig. 4): Requires special format with ".arp" extension. Generated from GenAlEx or Microsatellite Toolkit.
- Genepop (13) (Fig. 1): Requires that files have no extension. Generated from Microsatellite Toolkit or GenAlEx.
- FSTAT (12) (Fig. 5): Requires a ".dat" extension. Generated from Genepop or Microsatellite Toolkit.
- BOTTLENECK (Fig. 1) (14): Requires Genepop or FSTAT format.
- STRUCTURE (15) (Fig. 6): Requires a 3-digit genotype format. Generated from GenAlEx.
- GeneClass (Fig. 1) (16): Requires Genepop or FSTAT format.
- AGARST (17) (Fig. 7): Requires a 3-digit genotype format made manually.
- DISPAN (18) (Fig. 8): Requires special format. Generated from Microsatellite Toolkit.
- Cervus (11) (Fig. 9): Requires special format with ".csv" extension. Generated from GenAlEx.
- Microsat (19) (Fig. 10): Requires special format. Generated from Microsatellite Toolkit.

```
Title line:"3-digit GenePop"
LOC1
LOC2
LOC3
LOC4
LOC5
POP
popA ,  155157  212218  253253  196196  225231
popA ,  155155  212220  253263  178196  231231
POP
popB ,  155155  212212  263263  196196  225231
popB ,  157157  212218  259263  196196  225231
popB ,  155157  220220  253253  178196  225225
POP
popC ,  155157  212220  253259  196196  225225
popC ,  155159  220220  259263  178196  225231
POP
popD ,  157157  212212  245245  196196  225231
popD ,  157157  212220  253259  196196  225225
popD ,  155157  212212  245253  196196  225225
```

Fig. 1 The 3-digit Genepop input file format. FreeNA, BOTTLENECK, GeneClass, and Micro-Checker all use this format, and all look the same

	LOC1		LOC2		LOC3		LOC4		LOC5	
popA1	155	157	212	218	253	253	196	196	225	231
popA2	155	155	212	220	253	263	178	196	231	231
popB1	155	155	212	212	263	263	196	196	225	231
popB2	157	157	212	218	259	263	196	196	225	231
popB3	155	157	220	220	253	253	178	196	225	225
popC1	155	157	212	220	253	259	196	196	225	225
popC2	155	159	220	220	259	263	178	196	225	231
popD1	157	157	212	212	245	245	196	196	225	231
popD2	157	157	212	220	253	259	196	196	225	225
popD3	155	157	212	212	245	253	196	196	225	225

Fig. 2 The Microsatellite Toolkit input file format

5	10	4	2	3	2	3					
			popA	popB	popC	popD					
		LOC1		LOC2		LOC3		LOC4		LOC5	
popA1	popA	155	157	212	218	253	253	196	196	225	231
popA2	popA	155	155	212	220	253	263	178	196	231	231
popB1	popB	155	155	212	212	263	263	196	196	225	231
popB2	popB	157	157	212	218	259	263	196	196	225	231
popB3	popB	155	157	220	220	253	253	178	196	225	225
popC1	popC	155	157	212	220	253	259	196	196	225	225
popC2	popC	155	159	220	220	259	263	178	196	225	231
popD1	popD	157	157	212	212	245	245	196	196	225	231
popD2	popD	157	157	212	220	253	259	196	196	225	225
popD3	popD	155	157	212	212	245	253	196	196	225	225

Fig. 3 The GenAlEx input file format

```
[Profile]
    Title="Arlequin format"
    NbSamples=4
    GenotypicData=1
    GameticPhase=0
    RecessiveData=0
    DataType=STANDARD
    LocusSeparator=WHITESPACE
    MissingData='?'
    CompDistMatrix=1
[Data]
  [[Samples]]   #Data for 5Loci: LOC1 LOC2 LOC3 LOC4 LOC5
    SampleName="popA"
    SampleSize=2
    SampleData=     {
       popA1       1  155 212 253 196 225
                      157 218 253 196 231
       popA2       1  155 212 253 178 231
                      155 220 263 196 231
    }
    SampleName="popB"
    SampleSize=3
    SampleData=     {
       popB1       1  155 212 263 196 225
                      155 212 263 196 231
       popB2       1  157 212 259 196 225
                      157 218 263 196 231
       popB3       1  155 220 253 178 225
                      157 220 253 196 225
    }
    SampleName="popC"
    SampleSize=2
    SampleData=     {
       popC1       1  155 212 253 196 225
                      157 220 259 196 225
       popC2       1  155 220 259 178 225
                      159 220 263 196 231
    }
    SampleName="popD"
    SampleSize=3
     SampleData=     {
       popD1       1  157 212 245 196 225
                      157 212 245 196 231
       popD2       1  157 212 253 196 225
                      157 220 259 196 225
       popD3       1  155 212 245 196 225
                      157 212 253 196 225
     }
[[Structure]]
    StructureName=" Structure"
    NbGroups=1
    IndividualLevel=0
    Group= {

              "popA"
              "popB"
              "popC"
              "popD"
  }
```

Fig. 4 The Arlequin input file format

```
4 5 263 3
LOC1
LOC2
LOC3
LOC4
LOC5
1   155157  212218  253253  196196  225231
1   155155  212220  253263  178196  231231
2   155155  212212  263263  196196  225231
2   157157  212218  259263  196196  225231
2   155157  220220  253253  178196  225225
3   155157  212220  253259  196196  225225
3   155159  220220  259263  178196  225231
4   157157  212212  245245  196196  225231
4   157157  212220  253259  196196  225225
4   155157  212212  245253  196196  225225
```

Fig. 5 The FSTAT input file format

```
LOC1 LOC2 LOC3 LOC4 LOC5
popA1 1 0 155 157 212 218 253 253 196 196 225 231
popA2 1 0 155 155 212 220 253 263 178 196 231 231
popB1 2 0 155 155 212 212 263 263 196 196 225 231
popB2 2 0 157 157 212 218 259 263 196 196 225 231
popB3 2 0 155 157 220 220 253 253 178 196 225 225
popC1 3 0 155 157 212 220 253 259 196 196 225 225
popC2 3 0 155 159 220 220 259 263 178 196 225 231
popD1 4 0 157 157 212 212 245 245 196 196 225 231
popD2 4 0 157 157 212 220 253 259 196 196 225 225
popD3 4 0 155 157 212 212 245 253 196 196 225 225
```

Fig. 6 The STRUCTURE input file format

```
Test data table Five loci, Four pps

population A
a1          155 157 212 218 253 253 196 196 225 231
a2          155 155 212 220 253 263 178 196 231 231
population B
b1          155 155 212 212 263 263 196 196 225 231
b2          157 157 212 218 259 263 196 196 225 231
b3          155 157 220 220 253 253 178 196 225 225
population C
c1          155 157 212 220 253 259 196 196 225 225
c1          155 159 220 220 259 263 178 196 225 231
population D
d1          157 157 212 212 245 245 196 196 225 231
d2          157 157 212 220 253 259 196 196 225 225
d3          155 157 212 212 245 253 196 196 225 225
```

Fig. 7 The AGARST input file format. *Important:* Do not use the words "populations", "population", or "pop" in the title or as part of the population ID

- Migrate (20) (Fig. 11): Requires special format. Generated from AGARST.
- RSTCALC (21) (Fig. 12): Requires special format. Generated from AGARST.
- FreeNA (22) (Fig. 1): Requires Genepop format.

```
#Populations = (popA,popB,popC,popD)
#Monomorphic loci = 0

@Locus 1: LOC1
#Allele = ( 155,     157,     159    )
 0.7500     0.2500     0.0000     4  popA
 0.5000     0.5000     0.0000     6  popB
 0.5000     0.2500     0.2500     4  popC
 0.1667     0.8333     0.0000     6  popD

@Locus 2: LOC2
#Allele = ( 212,     218,     220    )
 0.5000     0.2500     0.2500     4
 0.5000     0.1667     0.3333     6
 0.2500     0.0000     0.7500     4
 0.8333     0.0000     0.1667     6

@Locus 3: LOC3
#Allele = ( 245,     253,     259,     263    )
 0.0000     0.7500     0.0000     0.2500     4
 0.0000     0.3333     0.1667     0.5000     6
 0.0000     0.2500     0.5000     0.2500     4
 0.5000     0.3333     0.1667     0.0000     6

@Locus 4: LOC4
#Allele = ( 178,     196    )
 0.2500     0.7500     4
 0.1667     0.8333     6
 0.2500     0.7500     4
 0.0000     1.0000     6

@Locus 5: LOC5
#Allele = ( 225,     231    )
 0.2500     0.7500     4
 0.6667     0.3333     6
 0.7500     0.2500     4
 0.8333     0.1667     6
```

Fig. 8 The DISPAN input file format

Sample	Sex	LOC1A	LOC1B	LOC2A	LOC2B	LOC3A	LOC3B	LOC4A	LOC4B	LOC5A	LOC5B
popA1	popA	155	157	212	218	253	253	196	196	225	231
popA2	popA	155	155	212	220	253	263	178	196	231	231
popB1	popB	155	155	212	212	263	263	196	196	225	231
popB2	popB	157	157	212	218	259	263	196	196	225	231
popB3	popB	155	157	220	220	253	253	178	196	225	225
popC1	popC	155	157	212	220	253	259	196	196	225	225
popC2	popC	155	159	220	220	259	263	178	196	225	231
popD1	popD	157	157	212	212	245	245	196	196	225	231
popD2	popD	157	157	212	220	253	259	196	196	225	225
popD3	popD	155	157	212	212	245	253	196	196	225	225

Fig. 9 The Cervus input file format

3.2 Converting Genotype File in Excel to Txt File Format

Several programs can generate input files for other programs. Input files generated by MS Excel-based programs must be changed to Txt file format for use in other programs. Genotype data in Excel contains the tab character, which should be eliminated using the following procedures:

```
% individual format

popA1   MS1   155
popA1   MS1   157
popA1   MS2   212
popA1   MS2   218
popA1   MS3   253
popA1   MS3   253
popA1   MS4   196
popA1   MS4   196
popA1   MS5   225
popA1   MS5   231
popA2   MS1   155
popA2   MS1   155
popA2   MS2   212
popA2   MS2   220
popA2   MS3   253
popA2   MS3   263
popA2   MS4   178
popA2   MS4   196
popA2   MS5   231
popA2   MS5   231
popB1   MS1   155
popB1   MS1   155
popB1   MS2   212
popB1   MS2   212
popB1   MS3   263
popB1   MS3   263
popB1   MS4   196
popB1   MS4   196
popB1   MS5   225
popB1   MS5   231
popB2   MS1   157
popB2   MS1   157
popB2   MS2   212
popB2   MS2   218
popB2   MS3   259
popB2   MS3   263
popB2   MS4   196
popB2   MS4   196
popB2   MS5   225
popB2   MS5   231
popB3   MS1   155
popB3   MS1   157
popB3   MS2   220
popB3   MS2   220
popB3   MS3   253
popB3   MS3   253
popB3   MS4   178
popB3   MS4   196
popB3   MS5   225
popB3   MS5   225
popC1   MS1   155
popC1   MS1   157
popC1   MS2   212
popC1   MS2   220
popC1   MS3   253
popC1   MS3   259
popC1   MS4   196
popC1   MS4   196
```

Fig. 10 The Microsat input file format

```
 4  5  . Agarst
 2     population  1
Indiv 1   2.4  2.8  10.10  20.20  2.8
Indiv 2   2.2  2.10  10.20  2.20  8.8
 3     population  2
Indiv 1   2.2  2.2  20.20  20.20  2.8
Indiv 2   4.4  2.8  16.20  20.20  2.8
Indiv 3   2.4  10.10  10.10  2.20  2.2
 2     population  3
Indiv 1   2.4  2.10  10.16  20.20  2.2
Indiv 2   2.6  10.10  16.20  2.20  2.8
 3     population  4
Indiv 1   4.4  2.2  2.2  20.20  2.8
Indiv 2   4.4  2.10  10.16  20.20  2.2
Indiv 3   2.4  2.2  2.10  20.20  2.2
```

Fig. 11 The Migrate input file format

```
Title
  5
  3
  4
  2
  3
  2
  3
Locus1
  2
  154
Locus2
  2
  211
Locus3
  2
  244
Locus4
  18
  177
Locus5
  6
  224
Pop1
  1 2  1 2  2 2  2 2  1 2
  1 1  1 3  2 4  1 2  2 2
Pop2
  1 1  1 1  4 4  2 2  1 2
  2 2  1 2  3 4  2 2  1 2
  1 2  3 3  2 2  1 2  1 1
Pop3
  1 2  1 3  2 3  2 2  1 1
  1 3  3 3  3 4  1 2  1 2
Pop4
  2 2  1 1  1 1  2 2  1 2
  2 2  1 3  2 3  2 2  1 1
  1 2  1 1  1 2  2 2  1 1
```

Fig. 12 The RSTCALC input file format

1. Create input file using Excel-based software, e.g., Microsatellite Toolkit.
2. Copy all of the data on the Excel worksheet and paste into MS word as unformatted text using the “paste special” command.

3. Eliminate all unnecessary keystrokes (all tabs ^t), as some programs are very sensitive in that regard.
4. In general, select "All" to copy and paste the data set into Notepad and save as a *.dat file (or file without extension) to import into the prescribed software program.

3.3 Formatting and Data Manipulation

We list three possible programs for initial data manipulation. These have the advantages of providing options for generating input files for other downstream analysis programs (see Subheading 3.1) and of procedures to check for errors in the dataset. Corrections to data in the genotype file require the process to start from the program selected for error-checking and verification of genotypes. Microsatellite Toolkit and Micro-Checker are commonly used to check for errors in the genotype data. Microsatellite Toolkit and GenAlEx generate input data files for other downstream software programs.

3.3.1 Microsatellite Toolkit

Features:

- Detects invalid alleles, incompletely typed samples (for diploid data), and invalid sample/population names recognized as duplicated or genetically identical samples.
- Calculates allele frequencies per population or locus, heterozygosity, allelic diversity, and individual relationship based on shared allele frequency.
- Creates input files for population genetics analysis programs such as Arlequin, Genepop, FSTAT, DISPAN, and Microsat.

Use:

1. Open **MS_tools.xla** (Excel add-in tools for microsatellite data) and conduct further analysis after selecting "Macro included".
2. Select Diploid one-column format or two-column format in Input data format after selecting "Microsatellite Toolkit" at additional function (two-column format is illustrated in Fig. 2). Toolkit automatically recognizes the number of samples and loci.
3. Click "OK" by selecting "Check data for errors" and by default setting in "Data checking parameters". Then 1 Col data format is created and if click "OK!", "Format options" window will pop up.
4. "Format options" allows data conversion for Arlequin, Genepop, Microsat, FSTAT, and DISPAN and provides summary statistics including allele frequencies and diversity statistics.

3.3.2 Micro-Checker

Features:

- Detects mistyped allele sizes and typographic errors and deviations from a regular repeat motif (suggesting indels or typos).

- Detects evidence of null alleles (one or more alleles fail to amplify during PCR) (see Note 1), stuttering (slight changes occur in the allele sizes during PCR), and large allele dropout (large alleles do not amplify as efficiently as small alleles).

Use:

1. Open **StartMicroChecker.exe** then open the "Data" file. On the lower toolbar, select each locus and identify the repeat motif for each locus. Unless a locus has a size greater than 350 bp, accept the default parameters.
2. For each locus, select "Check" for unusual observations.
3. This will open a display window to the right of the data file window with any unusual observations identified for each locus.
4. Record these unusual genotypes. You will need to return to the "original" worksheet to correct an unusual observation or accept it if you believe it is correct after verification.

3.3.3 GenAlEx

Features:

- The GenAlEx package can be used to generate input files for other useful population genetic software programs including Arlequin, Cervus, GeneClass, Genepop, and STRUCTURE.
- The program also has options to carry out various genetic analyses in MS Excel. Once its own input file is made, the program is a user-friendly package that can perform population genetics analyses including summary statistics such as diversity measures, tests of Hardy–Weinberg equilibrium, as well as advanced statistics such as genetic distance, Analysis of molecular variance (AMOVA), Mantel tests, Principal Coordinates Analysis of multivariate genetic data, estimation of pairwise relatedness among individuals, population assignment, and many more.

Use:

1. Open **GenAlEx6.1.xla** (Excel add-in tools for microsatellite and DNA sequence data) and select "Macro included".
2. Make correct input file. One can start this using the input file for Microsatellite Toolkit because it is regarded as the most basic data format.
3. Make column 2 for population ID using the column inserting option.
4. Make rows for basic information of dataset by selecting "Insert Header Rows" in Parameters.
5. Insert total number of loci by selecting "No. Codominant Loci" in Parameters.
6. Insert information for each population by selecting "Pops from Col2" in Parameters.

7. You can continue to use various options for both basic and advanced population genetic analyses or generate input files for advanced population genetic analyses using the "Export data" option. GenAlEx generates input files in the correct format for other programs.

3.4 Basic Population Genetic Analyses

3.4.1 Indices of Genetic Diversity

Estimation of genetic diversity is an essential component of population genetics analyses of natural organisms. Within-population indices of genetic diversity include the numbers of different alleles per locus, allelic richness, and expected (H_E) and observed (H_O) heterozygosity. The measures of heterozygosity are highly correlated, but expected (H_E) is considered a better estimator of the genetic variability present in a population (23). Since genetic diversity information is the most basic approach in empirical population genetics, numerous software programs are designed to provide such information. Indices of genetic diversity can be calculated using the programs AGARST, Arlequin, Cervus, DISPAN, FSTAT, GenAlEx, Genepop, and Microsatellite Toolkit (see Note 2).

3.4.2 Test of Hardy–Weinberg Equilibrium

A test of Hardy–Weinberg equilibrium (HWE) should be carried out as an initial step of population genetics analyses. Under the Hardy–Weinberg principle, frequencies of alleles remain constant in a population in the absence of selection, mutation, migration, and genetic drift. Thus, tests of Hardy–Weinberg equilibrium interrogate the stability of allele frequencies over time. The Hardy–Weinberg principle concerns the effects of a single generation of random mating where genotype frequencies can be predicted from the allele frequencies. HWE is expected for populations in which mating is random, and such a population should show no significant difference between observed and expected heterozygosity. Excessive deviation from HWE indicates violation of one of the assumptions of population genetics analyses through such processes as nonrandom mating or a lack of selective neutrality. However, significant deviation from HWE can also arise from physical error during genotyping, e.g., null alleles, and data must be interpreted with caution (see Note 1). Tests of HWE and its significance (see Note 3) can be carried out using the programs Arlequin, FSTAT, GenAlEx, and Genepop.

3.4.3 Test for Genotypic Linkage Disequilibrium

Population genetic parameters are calculated from genetic data across multiple loci which are assumed to assort independently of one another during meiosis. If two loci are located too close together on a chromosome, they are considered linked, resulting in genotypic linkage disequilibrium. Tests for genotypic linkage disequilibrium test the null hypothesis that genotypes at one locus are independent from genotypes at the other locus. A test of genotypic linkage disequilibrium and significance should be conducted during the initial step of marker selection or genetic analyses (24).

In the case of significant disequilibrium, the best course of action is to exclude one of the two makers from further population genetic analyses. Tests of genotypic linkage disequilibrium can be carried out using the programs Arlequin, FSTAT, and Genepop.

3.4.4 Measure of Fixation Indices and Genetic Differentiation Between Populations

Genetic differentiation can be measured by difference in frequency distribution of alleles between populations. Information of fixation indices such as *F*-statistics, F_{IS}, F_{IT}, and F_{ST} (25), per locus across all populations, should be investigated at the initial step of population genetics analyses. A significant difference between observed and expected heterozygosity results in a significant F_{IS} value and may indicate the presence of null alleles (see Note 1), the Wahlund effect, or some other anomaly. F_{ST} estimates are potentially in the range 0–1 and are a measure of how genetically different two populations are at selectively neutral loci, with an F_{ST} of 0 indicating that no genetic differentiation has occurred and a value of 1 indicating that the two populations share no genotypes in common. Extent of genetic differentiation between populations using F_{ST} (an estimate of population subdivision under the infinite allele model) and other *F*-statistics (25) per locus across all populations and their respective *p*-values (see Note 3) can be calculated by the programs Arlequin, FSTAT, GenAlEx, Genepop, and many more. AGARST, FSTAT, and RSTCALC can calculate R_{ST} (an estimate of population subdivision for stepwise mutation processes, ref. 26). FSTAT can provide an adjusted *p*-value to derive significance levels for analyses involving multiple comparisons (see Note 4).

3.4.5 Gene Flow Measures

Patterns and extent of gene flow provide important information on dispersal pattern and capacity of the study species. Indirect estimates of gene flow between populations can be measured with different approaches. First, one can calculate population genetic structure-based gene flow according to the relationship $Nem = (1 - F_{ST})/4\ F_{ST}$ (27), where Nem is the effective number of migrants per generation, Ne is the effective population size of each population, and m is the immigration rate. This classical measure of gene flow is based on equilibrium between the forces of immigration and genetic drift under the assumptions of the island model, i.e., that migration occurs among populations of equal size with symmetrical migration rates. Pairwise estimates of genetic differentiation among subpopulations and their significance can be quantified by F_{ST} (25) and R_{ST} (26) using the program FSTAT and RSTCALC, respectively (see Note 5). Second, maximum likelihood estimates of gene flow can be calculated using the coalescent-based Markov Chain Monte Carlo (MCMC) simulation approach, which takes into account the genealogical relationship of the samples and asymmetry in gene flow (20, 28). The necessary migration parameters, such as $4Ne\mu$, where μ is the mutation rate per generation at a locus and $M\ (=m/\mu)$, can be calculated using the program Migrate (20) (see Note 6).

Rate of migration can also be calculated from the frequency of private alleles. A private or rare allele is defined as an allele found in only one subpopulation, but not found in other subpopulations. Estimating gene flow using private alleles was developed by Slatkin (29) based on the following equation: $\ln[p(1)] \approx a \ln(Nm) + b$, where $a = -0.505$ and $b = -2.440$, and $p(1)$ is the frequencies of private alleles. Therefore, the logarithm of expected frequency of a private allele ($p(1)$) is approximately a decreasing linear function of the logarithm of Nm with a slope of −0.505 (Fig. 1 in ref. 29). Simulation showed that this method is relatively insensitive to changes in parameters of the model other than Nm and the number of individuals sampled per population, and the author provided a rough way to correct for differences in sample size (29). Therefore, one can use the value of $p(1)$ to estimate Nm using the program Genepop.

3.4.6 Analysis of Molecular Variance Test

An AMOVA estimates the proportion of genetic diversity within and between populations, or among groups of populations that the user categorizes based on criteria such as region. The AMOVA test is therefore used to evaluate the level of genetic differentiation within and among populations, regions, or other specified hierarchical categories. The partitioning of population genetic variance in such a hierarchical AMOVA can be conducted using the program GenAlEx or Arlequin. The significance of differentiation within and among populations within regions can be determined by permutations of samples, e.g., 1,000 replicates. The AMOVA is calculated based on Euclidean distances between individuals in GenAlEx and the closest model of evolution in Arlequin.

3.5 Advanced Population Genetic Analyses

3.5.1 Bottleneck Tests

Bottleneck tests are commonly used to examine population demography in recent time for evidence of a severe reduction in population size sufficient to leave a genetic signature. Evidence of recent population bottlenecks can be assessed using three different approaches. Three tests, including the Wilcoxon test which produces the most reliable results, are available in the program BOTTLENECK to determine whether deviations of observed heterozygosity (designated H_e in software documentation or H_o in (14)) relative to that expected at drift–mutation equilibrium (designated H_{eq} in software documentation or H_1 in (14)) are significant ($\alpha = 0.05$). Both a strict stepwise mutation model (SMM) (30) and a two-phase model (TPM) (31) with 1,000 iterations can be applied. For the TPM, a generalized stepwise mutation model (GSM), in which a proportion of SMM is set to 0 with a variance in mutation lengths of 0.36 (32), can be applied. Secondly, one can look for a mode-shift in allele frequency distribution from the L-shaped distribution expected under mutation–drift equilibrium, which can be used as a qualitative indicator of population bottlenecks (33). Third, the M value of Garza and Williamson (34) and

its variance across loci are calculated using the program AGARST. *M* is the mean ratio of the number of alleles to the range of allele size. This test is useful for detecting a bottleneck experienced further in the past. After a bottleneck, the *M* statistic will display persistently low values for about 100 generations. When compared to the results of the other two tests, the *M* test can distinguish populations that have been reduced in size recently from those which have been small for a long time (34).

3.5.2 Genetic Relationships Between Samples

Construction of a genetic relationship tree or of a scatter diagram from principal component analysis (PCA) or principal coordinate analysis (PCoA) of a multivariate dataset is performed to visualize pairwise differentiation between individuals or populations. Genetic divergence between populations based on allele frequencies can be calculated as genetic distance (D_A) (35) using the DISPAN computer program. Phylogenetic trees are constructed by neighbor-joining (NJ) clustering (36) or by the unweighted pair group method with the arithmetic mean (UPGMA) (37) using DA distance. Bootstrap resampling ($n = 1{,}000$) is applied to test the robustness of dendrogram topologies. A principal component analysis (PCA) is applied to a covariance matrix of allele frequencies across all variable loci using the program PCAGEN (38). Principal coordinate analysis (PCoA) can be conducted using the program GenAlEx. The geometric relationship among populations is visualized with a scattergram of the factor score data along the two PC axes that account for the most variation. To visualize genetic relationships among individuals, inter-individual genetic distances can be calculated based on the proportion of shared alleles using the Microsat computer program. These distance values are used to construct a UPGMA tree as implemented in the NEIGHBOR module of the PHYLIP software package (39).

3.5.3 Inferring Real-Time Migration Rate

Temporal analyses, the estimation of effective population size (Ne) and the migration rate (m) from samples collected over time, provide a way of measuring real-time migration regardless of population history (40–42). They also provide the most robust estimates possible of effective population size and migration rate (43). Temporal analysis is less sensitive to drift–migration equilibrium than population genetic structure-based gene flow (43), making it useful for estimating gene flow in invasive species or species that have undergone a recent range expansion, where estimates based on spatial data from geographic samples is problematic (see Note 5). The computer program MLNE allows estimation of m and Ne simultaneously using a maximum likelihood strategy (43). This method uses a temporal approach that compares allele frequencies from at least two generations. Simulation studies show that it performs better than other temporal methods (43).

3.5.4 Identification of Migrants in Current Generation

The Monte Carlo simulation approach of Paetkau et al. (42) enables the identification of immigrant individuals in the current generation, allowing an estimate of gene flow among populations at a much narrower time scale. The premise of this approach is based on resampling gametes rather than alleles to preserve linkage disequilibrium in recent immigrants. The analysis can be conducted using the "Detection of first generation migrants" criterion implemented in the program GeneClass, which assigns each potential immigrant to the most likely source population at a specified confidence level (42). First generation (F0) migrants are defined as individuals that traveled from site A to site B in year X (or the current generation) or individuals born in year X to a gravid female that moved from site A to B in year $X-1$ (or the previous generation). Two test statistics (the ratio $L_{\text{home}}/L_{\text{max}}$ and L_{home}) can be used to compute the likelihood of migrant detection (L) (42). In cases where it is unclear whether all potential source populations for immigrants have been sampled, L_{home} is the more appropriate test statistic but has reduced power to identify immigrants (42).

3.5.5 Assignment/ Exclusion Tests

To compute the probability of each individual's belonging to a set of reference (current or potential source) populations, assignment/exclusion tests using the direct and simulation approaches can be conducted using options implemented in the program GeneClass. The direct assignment tests allocate an individual to one of the reference populations without probability computation, thereby simply calculating the proportion of correctly assigned individuals to the most likely population of origin, even if the true population of origin is not among the reference populations. In contrast, the exclusion method uses a simulation approach. This method computes the likelihood of a genotype occurring in the population by simulating multilocus genotypes based on allele frequencies of each reference population and compares the likelihood of the genotype of an individual to the distribution of likelihoods of simulated genotypes for each reference population. If the individual genotype likelihood is below a given threshold (e.g., $\alpha = 0.01$), the population is excluded as a possible origin of the individual (40). Unlike the direct assignment method, the exclusion method does not assume that the true population of origin has been sampled, because each population is treated independently (40). The Bayesian statistical approach of Rannala and Mountain (44), which has proven to be more accurate than frequency and distance based methods (40), is used for both assignment and exclusion tests. Frequency probabilities of multilocus genotypes in each reference population are determined in the exclusion test using Monte Carlo simulations of 10,000 independent individuals for the population (42). The assignment likelihoods of individuals from a geographic population

to putative source populations can be further calculated and averaged using the Bayesian statistical method (44). The statistic $L_{i \text{ to } j}$, the mean individual assignment likelihoods of individuals collected in population i and assigned to population j, provides valuable asymmetric information for the origin of the newly introduced population under the assumptions that it is new in the location and that the putative source population was sampled.

GenAlEx can be used to visualize the relative position of individuals from spatial population data between locations by plotting the log likelihood of an individual's genotype arising in the populations using the frequency-based assignment test (42, 45). If the individual's allele is absent from one of the represented populations, the value can be set to 0.01 and the "leave one out" option (46) is applied for the assignment test.

3.5.6 Inferring the Number of Distinct Genetic Populations

The program STRUCTURE uses a model-based Bayesian clustering method to infer the number of distinct populations (K) from which samples have been drawn and to infer the genetic ancestry of the individuals sampled, based on microsatellite genotypes at multiple loci. This approach provides an independent assessment of these parameters, free of the prior assumption that each sample location constitutes a population. Thus, the results complement those of the genetic tree (Subheading 3.5.2), population structuring (Subheadings 3.4.4 and 3.4.6), and population assignment tests (Subheading 3.5.5) described above. The program is used to estimate $\Pr(X/K)$, the probability of the observed set of genotypes (X), conditional on a given K. The program can be run using different replications for both burn-in and the consequent resampling. An initial burn-in of 100,000 iterations followed by 1,000,000 iterations is common. An admixture model of individual ancestry and correlated allele frequencies among populations are appropriate for most natural populations. Multiple runs are required to test performance for each value of K to verify that estimates of $\Pr(K/X)$ were consistent between runs. The posterior probabilities of K, $\Pr(K/X)$, are calculated according to Pritchard et al. (14). The "real" value of K (number of unique populations represented by the genotypes within the sample) is estimated from the ln $\Pr(K/X)$ values output for each replicate of K using the $m(|L''(K)|)/s[L(K)]$ statistic described by Evanno et al. (47). In brief, the "real" value of K within the dataset is determined as the ln $\Pr(K/X)$ that maximizes the value of $\Delta K = m(|L''(K)|)/s[L(K)]$.

3.5.7 Genetic Isolation by Geographic Distance

A special, but common, problem is to examine gene flow within a continuously distributed population. In such cases, one would expect genetic differentiation between locations within the large continuous population to increase with distance alone. A pattern of

isolation by distance (IBD) can be examined through regression of the genetic distance on geographic distance among locations. Slatkin (48) suggested that a pattern of isolation by distance should be detectable when a population is at or near equilibrium under its current patterns of dispersal. The absence of isolation by distance pattern suggests that the population either is far from equilibrium, and that genetic structuring may reflect a recent range expansion rather than current levels of gene flow, or that the spatial scale sampled was too small relative to normal dispersal distances. IBD (49) is inferred from the relationship between $F_{ST}/(1 - F_{ST})$ (a measure of genetic distance) and the geographic distance between all pairs of sampled locations. It is recommended that untransformed distance (km) be used for a one-dimensional (i.e., linear) sampling scheme and the logarithm of distance be used for two-dimensional sampling schemes in the regression (50). Regression of $F_{ST}/(1 - F_{ST})$ on geographic distance between all pairs of sampling locations and the probability that there is no relationship based on permutations of samples can be calculated using the Matrix Comparison option in Arlequin, FSTAT, Genepop, and GenAlEx (see Note 7).

4 Notes

1. A null allele is caused when nucleotide variation in the flanking region of the microsatellite locus prevents primer binding and PCR amplification, making the locus appear homozygous for the one allele that does amplify (51) or resulting in no amplification at all at the locus if both alleles are null. This functionally recessive behavior leads to a decrease in genotyping accuracy, which in turn can result in a number of artifacts including heterozygote deficiency, inaccurate allele frequency estimates, and inflated F_{IS}, F_{ST}, and genetic distance estimates (22, 51, 52). The extent to which null alleles tend to overestimate the true population differentiation has not been investigated (22) but can lead to overestimates of population differentiation due to effects on subsequent calculations of F_{ST} and genetic distances (53, 54). Therefore, in any population genetics study using microsatellites, the potential for null alleles must be addressed (52, 55).

   Microsatellite loci that deviate significantly from HWE show evidence of null alleles according to the distribution of homozygote-size classes. The program Micro-Checker is used to estimate the frequency of null alleles and other genotyping errors such as stuttering and allele drop out. Null alleles are suspected for a given locus when the Micro-Checker program rejects Hardy–Weinberg equilibrium (HWE) among genotypes and if the excess

homozygote genotypes are evenly distributed among allele size classes. In the case of alleles harboring the potential null alleles, corrected pairwise F_{ST} estimates are calculated for all populations by applying the ENA correction in the FreeNA package.

2. Adjusted allelic diversity to account for variation in sample sizes can be calculated using both bootstrapping and jackknifing techniques implemented in the program AGARST or using allelic richness in the program FSTAT.
3. Determining the significance of differences in multiple comparisons requires a correction to avoid inflated type I error rates. Calculation of population genetics parameters such as genotypic linkage disequilibrium, pairwise F_{ST}, and HWE test often requires multiple tests since multiple populations from different sampling sites are used for calculations in a single table. One of the most popular methods for correcting for such multiple tests is the sequential Bonferroni correction, which provides adjusted p-values to maintain the intended α level of significance (56).
4. The calculations underlying the Bonferroni correction (see Note 3) are appropriate only for multiple independent tests. To account for the presence of multiple dependent tests within pairwise F_{ST} estimates, we suggest a correction to the significance thresholds for the critical value according to the B–Y method of Benjamini and Yekutieli (57).
5. An underlying assumption of spatial pairwise F_{ST} estimates is that the populations are in migration–drift equilibrium. This assumption is most often violated in the case of an invasive species or in a region where a species has undergone a recent range expansion. After a range expansion, F_{ST} values are often low and nonsignificant because of genetic founder effects, even though dispersal and gene flow may be limited. In such cases, estimates of gene flow are best obtained by analyzing temporal genetic data, i.e., data collected over time at the same locations (see Subheading 3.5.3).
6. Both the traditional gene flow measures based on allele frequency distributions and the coalescent-based maximum likelihood estimation of gene flow mainly reflect relatively long-term gene flow and thus may not accurately represent current levels. There are other methods available to determine whether each individual is a resident in the population in which it was sampled or an immigrant, and to estimate the number of immigrant individuals present in the current generation. These are covered in Subheadings 3.5.4 and 3.5.5.
7. Because the pairwise F_{ST} estimates are not independent data, a simple linear regression is not appropriate, and the permutation method is required.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MEST) (No. 2009-0080227). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

References

1. Kim KS, Sappington TW (2006) Molecular genetic variation of boll weevil populations in North America estimated with microsatellites: implications for patterns of dispersal. Genetica 127:143–161
2. Jiang X-F, Luo L-Z, Zhang L (2007) Amplified fragment length polymorphism analysis of *Mythimna separata* (Lepidoptera: Noctuidae) geographic and melanic laboratory populations in China. J Econ Entomol 100: 1525–2532
3. Jiang X-F, Cao W-J, Zhang L, Luo L-Z (2010) Beet webworm (Lepidoptera: Pyralidae) migration in China: evidence from genetic markers. Environ Entomol 39:232–242
4. Nagoshi RN, Fleischer S, Meagher RL (2009) Texas is the overwintering source of fall armyworm in central Pennsylvania: implications for migration into the northeastern United States. Environ Entomol 38:1546–1554
5. Kim KS, Coates BS, Bagley MJ, Hellmich RL, Sappington TW (2011) Genetic structure and gene flow among European corn borer (Lepidoptera: Crambidae) populations from the Great Plains to the Appalachians of North America. Agric For Entomol 13:383–393
6. Kim KS, Bagley MJ, Coates BS, Hellmich RL, Sappington TW (2009) Spatial and temporal genetic analyses show high gene flow among European corn borer (Lepidoptera: Crambidae) populations across the central U.S. Corn Belt. Environ Entomol 38:1312–1323
7. Van Oosterhout C, Hutchinson W, Wills D, Shipley P (2004) Micro-Checker: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Resour 4:535–538
8. Park SDE (2001) Trypanotolerance in West African cattle and the population genetic effects of selection. Ph.D. thesis, University of Dublin
9. Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288–295
10. Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 10:564–567
11. Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program Cervus accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099–1106
12. Goudet J (1995) Fstat version 1.2: a computer program to calculate F statistics (version 2.9.03). J Hered 86:485–486
13. Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Heredity 86:248–249
14. Cornuet J, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144:2001–2014
15. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
16. Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A (2004) GeneClass2: a software for genetic assignment and first-generation migrant detection. Heredity 95:536–539
17. Harley EH (2001) AGARst. A programme for calculating allele frequencies, G_{ST} and R_{ST} from microsatellite data, version 2. University of Cape Town, Cape Town, South Africa

18. Ota T (1993) DISPAN: genetic distance and phylogenetic analysis. Pennsylvania State University, University Park, PA
19. Minch E (1998) MICROSAT version 1.5b. University of Stanford, Stanford, CA
20. Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773
21. Goodman SJ (1997) Rst Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. Mol Ecol 6:881–885
22. Chapuis M-P, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. Mol Biol Evol 24:621–631
23. Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York
24. Kim KS, Stolz U, Miller NJ, Waits ER, Guillemaud T, Sumerford DV, Sappington TW (2008) A core set of microsatellite markers for western corn rootworm (Coleoptera: Chrysomelidae) population genetics studies. Environ Entomol 37:293–300
25. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370
26. Slatkin M (1985) Gene flow in natural populations. Annu Rev Ecol Syst 16:393–430
27. Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159
28. Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. Proc Natl Acad Sci USA 98:4563–4568
29. Slatkin M (1985) Rare alleles as indicators of gene flow. Evolution 39:53–65
30. Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci USA 75:2868–2872
31. Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci USA 91:3166–3170
32. Estoup A, Wilson IJ, Sullivan C, Cornuet JM, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. Genetics 159:1671–1687
33. Luikart G, Allendorf FW, Cornuet JM, Sherwin B (1998) Distortion of allele frequency distributions provides a test for recent population bottlenecks. J Hered 89:238–247
34. Garza JC, Williamson EG (2001) Detection of reduction of population size using data from microsatellite loci. Mol Ecol 10:305–318
35. Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170
36. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425
37. Sneath PHA, Sokal RR (1973) Numerical taxonomy. W.H. Freedman and Co., San Francisco
38. Goudet J (1999) PCAGEN version 1.2. Population genetics laboratory, University of Lausanne, Lausanne, Switzerland
39. Felsenstein J (1993) PHYLIP-phylogenetic inference package, version 3.5c. University of Washington, Seattle, WA
40. Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153:1989–2000
41. Wilson GA, Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163:1177–1191
42. Paetkau D, Slade R, Burdens M, Estoup A (2004) Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation based exploration of accuracy and power. Mol Ecol 13:55–65
43. Wang J, Whitlock MC (2003) Estimating effective population size and migration rates from genetic samples over space and time. Genetics 163:429–446
44. Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci USA 94:9197–9201
45. Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol 4:347–354
46. Efron B (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78:316–331
47. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Mol Ecol 14:2611–2620

48. Slatkin M (1993) Isolation by distance in equilibrium and nonequilibrium populations. Evolution 47:264–279
49. Wright S (1943) Isolation by distance. Genetics 28:114–138
50. Rousset F (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. Genetics 145:1219–1228
51. de Sousa SN, Finkeldey R, Gailing O (2005) Experimental verification of microsatellite null alleles in Norway spruce (*Picea abies* [L.] Karst.): implications for population genetic studies. Plant Mol Biol Rep 23:113–119
52. Girard P, Angers B (2008) Assessment of power and accuracy of methods for detection and frequency-estimation of null alleles. Genetica 134:187–197
53. Slatkin M (1995) Hitchhiking and associative overdominance at a microsatellite locus. Mol Biol Evol 12:473–480
54. Paetkau D, Waits IP, Clarkson PL, Craighead I, Strobeck C (1997) An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics 147:1943–1957
55. Pemberton JM, Slate J, Bancroft DR, Barrett JA (1995) Nonamplifying alleles at microsatellite loci: a caution for parentage and population studies. Mol Ecol 4:249–252
56. Rice WR (1989) Analysing tables of statistical tests. Evolution 43:223–225
57. Benjamini Y, Yekutieli D (2001) The control of false discovery rate under dependency. Ann Stat 29:1165–1188