

Getting starting with a data set

Some basic questions and programs
this week ...

...and the following 2 weeks

ECL 290

Lindsey Clark, Renate Eberl, Andrea Drauch, Mandi
Finger, Rachel Simmons, Joanna Fernandez

Migration?

Population subdivision?

Hybridization?

Inbreeding?

Levels of diversity?

Sources of invasions?

Dispersal?

Mating systems?

How many populations?

Lake Sturgeon microsatellite data

(modified from Andrea's thesis)



Outline

- The data set (preliminary)
 - excel spreadsheet
 - data checking (manual, **Microsatellite toolkit**)
 - different formats (**Microsatellite toolkit, Convert**)
 - characteristics of my data (#alleles, heterozygosity)
 - Which and how many loci to use –**Whichloci**
-

- Linkage Disequilibrium
- Hardy Weinberg
- Population subdivision ...

Next weeks
GenePop, GDA,
Fstat

Different infiles for population genetic software software

GENEPOP

```
begin gdata: [Andreas data]
dimensions npop= 7 nloc= 10;
format missing=? separator=/;
locusAlleleLabels
1 AOX27
2 AFU63
3 AFU68B
4 AFU96
5 AFU160
6 AFU195
7 AFU9
8 AFU204
9 AFU74
10 Sp1120
;

matrix
CANADIAN_LAKE :
CAN_1 130/ 130 139/ 139 182/ 182 266/ 274 134/ 146 161/ 161 268/ 296 141/ 141 226/ 226 254/ 254
CAN_2 130/ 130 133/ 139 166/ 166 274/ 274 134/ 134 161/ 165 288/ 304 141/ 141 226/ 226 283/ 286
CAN_3 130/ 130 135/ 139 166/ 166 266/ 266 134/ 146 165/ 165 292/ 296 141/ 141 218/ 226 254/ 281
CAN_4 130/ 130 139/ 139 182/ 182 266/ 274 134/ 134 161/ 165 296/ 304 141/ 141 218/ 226 254/ 285
CAN_5 130/ 130 133/ 139 166/ 182 262/ 262 134/ 150 161/ 165 288/ 296 141/ 141 226/ 226 254/ 285
CAN_6 130/ 130 139/ 139 182/ 182 266/ 274 134/ 134 161/ 161 296/ 304 141/ 141 226/ 226 254/ 285
CAN_7 130/ 130 139/ 139 170/ 190 274/ 274 134/ 146 161/ 161 296/ 304 141/ 141 226/ 226 254/ 285
CAN_8 130/ 130 139/ 143 182/ 190 266/ 274 134/ 134 161/ 165 288/ 304 141/ 141 226/ 226 254/ 285
CAN_9 130/ 130 139/ 139 190/ 190 266/ 266 134/ 134 161/ 161 296/ 312 141/ 141 226/ 226 254/ 262
CAN_10 130/ 130 135/ 139 139/ 143 166/ 182 262/ 262 134/ 150 161/ 165 288/ 296 141/ 141 226/ 226 254/ 285
CAN_11 130/ 130 139/ 143 170/ 190 262/ 266 134/ 146 165/ 165 296/ 296 141/ 141 222/ 226 262/ 285
CAN_12 130/ 130 139/ 139 166/ 190 274/ 274 134/ 134 161/ 161 298/ 296 141/ 141 218/ 222 254/ 281
CAN_13 130/ 130 139/ 139 182/ 182 266/ 274 134/ 134 161/ 165 288/ 296 141/ 141 226/ 226 254/ 285
CAN_14 130/ 130 139/ 143 166/ 182 266/ 266 134/ 134 161/ 165 288/ 296 141/ 141 226/ 226 285/ 285
CAN_15 130/ 130 139/ 143 170/ 182 266/ 274 134/ 134 161/ 165 288/ 296 141/ 141 226/ 226 254/ 285
CAN_16 130/ 130 139/ 143 166/ 182 266/ 266 134/ 134 161/ 165 288/ 296 141/ 141 226/ 226 285/ 285
CAN_17 130/ 130 139/ 143 166/ 182 266/ 274 134/ 134 161/ 165 288/ 296 141/ 141 226/ 226 281/ 281
CAN_18 130/ 130 139/ 139 190/ 190 274/ 274 134/ 134 161/ 165 268/ 288 141/ 141 222/ 222 281/ 281
CAN_19 130/ 130 139/ 139 190/ 190 266/ 266 134/ 134 161/ 161 296/ 304 141/ 141 222/ 222 254/ 281
CAN_20 130/ 130 139/ 139 166/ 166 266/ 274 134/ 146 161/ 165 296/ 296 141/ 141 222/ 222 281/ 281
CAN_21 130/ 130 139/ 139 166/ 166 266/ 274 134/ 146 161/ 165 296/ 296 141/ 141 222/ 222 281/ 281
CAN_22 130/ 130 139/ 139 166/ 166 266/ 274 134/ 146 161/ 165 296/ 296 141/ 141 222/ 222 281/ 281
CAN_23 130/ 134 139/ 139 166/ 182 266/ 274 134/ 146 161/ 165 296/ 296 141/ 141 226/ 226 281/ 281
CAN_24 130/ 130 139/ 139 166/ 166 266/ 274 134/ 146 161/ 165 296/ 296 141/ 141 222/ 222 281/ 281
CAN_25 130/ 130 139/ 139 166/ 166 266/ 274 134/ 146 161/ 165 296/ 296 141/ 141 222/ 222 281/ 281
CAN_26 130/ 130 135/ 139 190/ 190 266/ 274 134/ 134 165/ 165 288/ 296 141/ 141 226/ 226 254/ 285
CAN_27 130/ 130 135/ 135 190/ 190 266/ 274 134/ 150 161/ 165 296/ 296 141/ 141 226/ 226 254/ 254
CAN_28 130/ 130 139/ 139 170/ 182 266/ 266 134/ 134 161/ 165 272/ 296 141/ 141 222/ 226 281/ 285
CAN_29 130/ 130 139/ 139 178/ 190 266/ 266 134/ 150 161/ 165 296/ 296 141/ 141 226/ 226 285/ 285
CAN_30 130/ 130 139/ 139 182/ 186 266/ 266 134/ 134 161/ 161 296/ 296 141/ 141 226/ 222 254/ 262

WISCONSIN_LAKE :
WILK_1 130/ 130 127/ 139 174/ 178 266/ 266 134/ 146 165/ 165 276/ 284 141/ 145 218/ 218 254/ 254
WILK_2 130/ 138 139/ 139 182/ 190 266/ 266 134/ 146 161/ 165 292/ 292 141/ 145 218/ 226 254/ 277
WILK_3 130/ 139 139/ 139 182/ 190 266/ 266 134/ 146 161/ 161 284/ 292 141/ 145 226/ 226 ?/ ?
WILK_4 130/ 134 127/ 139 178/ 190 266/ 266 134/ 146 161/ 165 284/ 284 141/ 141 218/ 218 254/ 254
WILK_5 130/ 139 139/ 139 182/ 190 266/ 266 134/ 146 161/ 165 292/ 292 141/ 145 218/ 218 254/ 254
WILK_6 130/ 130 139/ 143 178/ 190 262/ 266 134/ 134 161/ 165 288/ 292 141/ 145 218/ 218 254/ 262
WILK_7 130/ 130 127/ 139 178/ 182 266/ 266 134/ 134 161/ 161 284/ 284 141/ 141 218/ 226 254/ 254
WILK_8 130/ 130 139/ 139 166/ 166 266/ 266 134/ 146 161/ 165 284/ 284 141/ 145 218/ 218 254/ 254
WILK_9 130/ 130 139/ 143 174/ 174 266/ 266 134/ 134 161/ 165 284/ 284 145/ 145 218/ 218 277/ 285
WILK_10 130/ 130 139/ 143 178/ 178 266/ 266 134/ 134 161/ 165 284/ 284 141/ 141 218/ 218 254/ 254
WILK_11 130/ 130 139/ 143 178/ 178 266/ 266 134/ 146 161/ 165 284/ 284 141/ 145 218/ 218 254/ 254
WILK_12 130/ 130 135/ 139 178/ 178 266/ 266 134/ 134 165/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_13 130/ 130 139/ 139 166/ 166 266/ 266 134/ 134 161/ 165 284/ 284 141/ 145 218/ 218 254/ 254
WILK_14 130/ 138 139/ 143 178/ 182 262/ 266 134/ 134 165/ 165 284/ 296 141/ 141 218/ 218 254/ 277
WILK_15 130/ 138 139/ 143 178/ 182 262/ 266 134/ 134 165/ 165 284/ 296 141/ 141 218/ 218 254/ 277
WILK_16 130/ 138 139/ 143 178/ 182 262/ 266 134/ 134 165/ 165 284/ 296 141/ 141 218/ 218 254/ 277
WILK_17 130/ 138 139/ 143 178/ 182 262/ 266 134/ 134 165/ 165 284/ 296 141/ 141 218/ 218 254/ 277
WILK_18 130/ 130 174/ 182 266/ 266 134/ 146 161/ 161 276/ 292 141/ 145 218/ 226 254/ 254
WILK_19 130/ 138 139/ 139 182/ 190 266/ 266 134/ 146 161/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_20 130/ 130 135/ 135 190/ 190 266/ 266 134/ 150 161/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_21 130/ 138 139/ 139 182/ 190 266/ 266 134/ 146 161/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_22 130/ 130 135/ 135 190/ 190 266/ 266 134/ 150 161/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_23 130/ 130 139/ 139 166/ 166 266/ 266 134/ 134 161/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_24 130/ 130 139/ 139 166/ 166 266/ 266 134/ 134 161/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_25 130/ 130 139/ 139 166/ 166 266/ 266 134/ 134 161/ 165 284/ 284 141/ 145 218/ 226 254/ 277
WILK_26 130/ 130 172/ 182 266/ 266 134/ 134 161/ 161 284/ 284 141/ 145 218/ 226 277/ 285
WILK_27 130/ 138 139/ 139 182/ 190 266/ 266 134/ 146 161/ 165 284/ 284 141/ 145 218/ 226 285/ 285
WILK_28 130/ 138 127/ 174/ 178 266/ 266 134/ 134 161/ 165 296/ 296 141/ 141 218/ 218 254/ 254
WILK_29 130/ 130 127/ 139 174/ 178 266/ 266 146/ 146 161/ 161 288/ 288 141/ 145 218/ 218 254/ 254
```

GDA

programs for data conversion of microsatellite data

- Microsatellite toolkit
- Convert

	convert	ms toolkit
Arlequin	x	x
Dispan		x
Fstat		x
GDA	x	
Genepop *	x	x
Microsat	x	x
Phylip	x	
PopGene	x	
Structure	x	

Naming and error checking

Population and sample names

(consistent not too long)

Excel spreadsheet

Sample names and populations

PopOne1, popOne2,popFive88

population names should not contain numbers

sample names should not contain unusual characters

Loci

Check for errors in datafile either manually

... or use Microsatellite toolkit

Microsatellite toolkit

<http://animalgenomics.ucd.ie/sdeparc/ms-toolkit/>

An excel add in (free download)

Error checking

- Non-numeric (e.g ?)
- Non-integer
- Negative
- Below above a specified threshold level

Find matching samples (specify min # of nonmatching alleles)

→ redundant worksheet

Data conversion

(Arlequin, GenePop, Microsat, Fstat, Dispan)

creates worksheets in excel datafile, option to save as textfile

Ability to exclude loci or populations

Basic information about diversity in data set

Microsatellite toolkit

Continued

- Allele frequencies and diversity statistics
 - “Alleles by Pop” Allele counts
 - “Allele Fqs” Allele frequencies (%)
 - “Stats” all stats averaged across all loci
 - Unbiased gene diversity = expected heterozygosity (Nei 1987)
 - Observed heterozygosity (Hedrick 1983)
 - Mean number of alleles per locus

Expected Heterozygosity (Nei 1978)

Unbiased gene diversity for each locus is given by:

$$\hat{h} = 2n(1 - \sum \hat{x}_i^2) / (2n - 1)$$

where

$$\hat{x}_i = \hat{X}_{ii} + \sum_{i \neq j} \hat{X}_{ij} / 2$$

where \hat{X}_{ij} is the frequency of genotype $A_i A_j$ in the sample, and n is the number of individuals sampled.

Average gene diversity (across r loci) is given by:

$$\hat{H} = \sum_{j=1}^r \hat{h}_j / r$$

Variance of gene diversity is given by:

$$V(\hat{H}) = V(\hat{h}) / r$$

where

$$V(\hat{h}) = \sum_{j=1}^r (\hat{h}_j - \hat{H})^2 / (r - 1)$$

Observed Heterozygosity (Hedrick 1983)

- # heterozygotes / # individuals typed at a locus

Average observed heterozygosity across r loci is given by:

$$\bar{h}_{obs} = \sum_{j=1}^r h_{obsj} / n_j r$$

The standard deviation of h_{obs} is given by:

$$\sigma(h_{obs}) = \frac{\bar{h}_{obs}(1 - \bar{h}_{obs})}{\bar{n} r}$$

Mean number of alleles per locus

Mean number of alleles per locus is given by:

$$\bar{A} = \sum_{j=1}^r A_j / r$$

where A_j is the number of distinct alleles at locus j and r is the number of loci

Standard deviation of mean number of alleles / locus is calculated as:

$$\sigma(A) = \sqrt{\sum_{j=1}^r (A_j - \bar{A})^2 / (r - 1)}$$

Convert

<http://www.agriculture.purdue.edu/fnr/html/faculty/Rhodes/Students%20and%20Staff/glaubitz/software.htm>

Simple program for data conversion

Input file

```
Data for presentation      *line with descriptive title
npops = 7                  *number of populations
nloci = 10                 *number of loci
      Aox27      Afu63      Afu68b      Afu56      ...
pop = CANADIAN_LAKE
CAN_1      130      130      139      139      182      182      266      274
CAN_2      130      130      135      139      166      166      274      274
CAN_3      130      130      135      139      166      166      266      266
...
```

or GENEPOP format

Data conversion

(GDA, GENEPOP, Arlequin, Structure, Phylip, Microsat, Popgene)

Creates separate output files

Table of allele frequencies, identifies private alleles