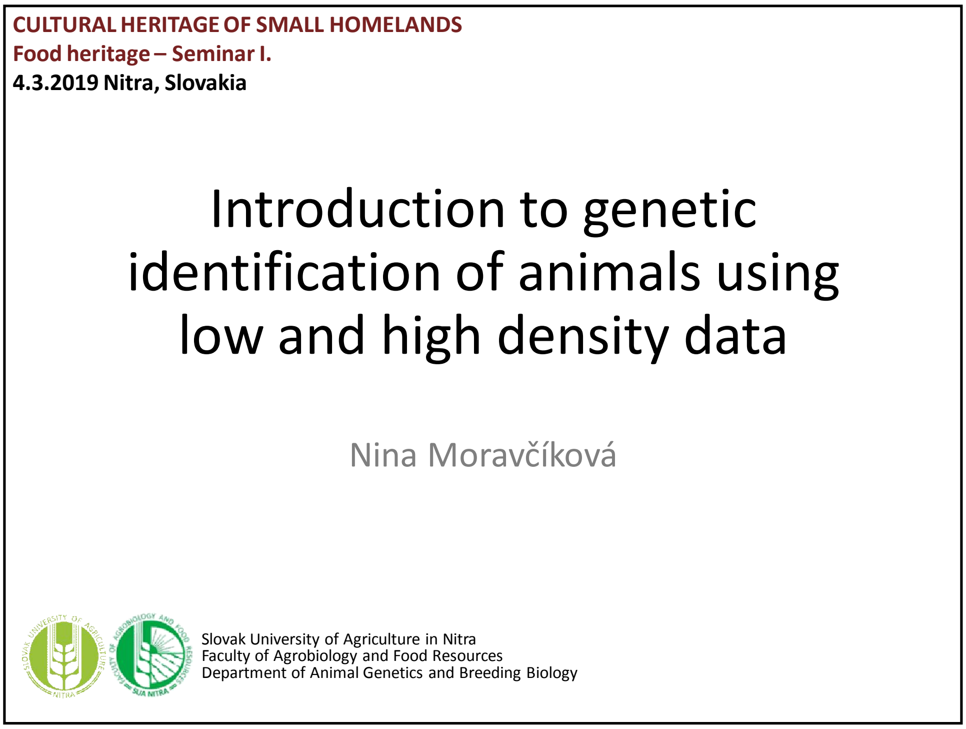
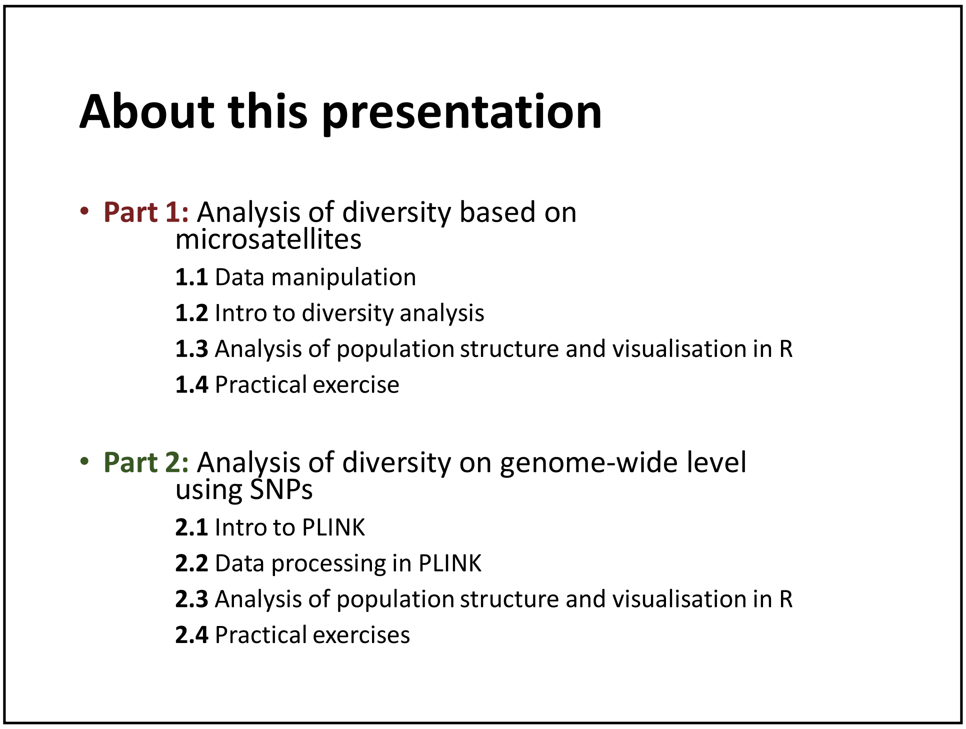


CULTURAL HERITAGE OF SMALL HOMELANDS
Food heritage – Seminar I.
4.3.2019 Nitra, Slovakia

Introduction to genetic identification of animals using low and high density data

Nina Moravčíková

Slovak University of Agriculture in Nitra
Faculty of Agrobiology and Food Resources
Department of Animal Genetics and Breeding Biology



About this presentation

- **Part 1:** Analysis of diversity based on microsatellites
  - **1.1** Data manipulation
  - **1.2** Intro to diversity analysis
  - **1.3** Analysis of population structure and visualisation in R
  - **1.4** Practical exercise
- **Part 2:** Analysis of diversity on genome-wide level using SNPs
  - **2.1** Intro to PLINK
  - **2.2** Data processing in PLINK
  - **2.3** Analysis of population structure and visualisation in R
  - **2.4** Practical exercises

Part 1: Analysis of diversity based on microsatellites

1.1 Data manipulation

From biological samples to genetic analysis

1) isolation of genomic DNA from biological samples (blood, hair roots, semen, ….) in lab
2) lab identification of animals’ genotypes by molecular-genetic methods based on use of genetic markers (microsatellites, SNPs, etc.) showing polymorphism
   - genetic polymorphism – occurrence of two or more genetically determined phenotypes in the same population
3) the quality control of genotyping data and development of input databases depending on the type of data
4) analysis of genetic diversity

1.2 Intro to diversity analysis

- frequency of alleles and genotypes
- Hardy – Weinberg equilibrium in population (χ^2 test, ...)
- heterozygosity and homozygosity
- effective number of alleles
- polymorphic information content
- Wright's F – statistics: F_{IS} – (f, molecular equivalent of pedigree inbreeding), F_{ST} (genetic relationship between pops) a F_{IT} (total inbreeding in metapopulation)
- genetic distance (Nei's D_a, ...)
- analysis of molecular variance (AMOVA)
- analysis of principal components (PCA)

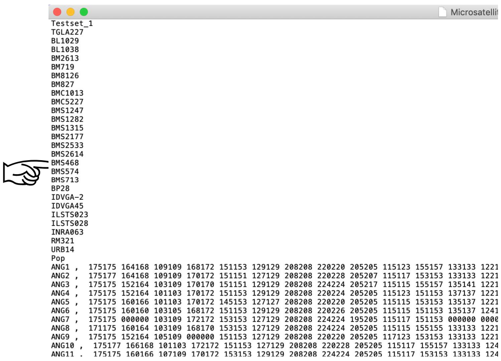
1.2 Intro to diversity analysis

Input data in genepop format

Genepop

- one of the most commonly used format of data
- Genetix, Genepop, Arlequine, various R packages

Microsatellite_testset_1



```
Testset_1
TGLA227
BL1029
BL1038
BM2613
BM719
BM8126
BM827
BMC1013
BMC5227
BMS1247
BMS1282
BMS1315
BMS2177
BMS2533
BMS2614
BMS468
BMS574
BMS713
BP28
IDVGA-2
IDVGA45
ILSTS023
ILSTS028
INRA063
RM321
URB14
Pop
ANG1 ,  175175 164168 109109 168172 151153 129129 208208 220220 205205 115123 155157 133133 1221
ANG2 ,  175177 164168 109109 170172 151151 127129 208208 220228 205207 115117 153153 133133 1221
ANG3 ,  175175 152164 103109 170170 151151 129129 208208 224224 205217 115115 155157 135141 1221
ANG4 ,  175175 152164 101103 170172 151153 129129 208208 220224 205205 115123 151153 137137 1221
ANG5 ,  175175 160166 101103 170172 145153 127127 208208 220220 205205 115115 153153 135137 1221
ANG6 ,  175175 160160 103105 168172 151153 129129 208208 220226 205205 115115 151153 135137 1241
ANG7 ,  175175 000000 103109 172172 153153 127129 208208 224224 195205 115117 151153 000000 0000
ANG8 ,  171175 160164 103109 168170 153153 127129 208208 224224 205205 115115 155155 133133 1221
ANG9 ,  175175 152164 105109 000000 151153 127129 208208 220220 205205 117123 153153 133133 1221
ANG10 ,  175177 166168 101103 172172 151153 127129 208208 220228 205205 115117 155157 133133 124
ANG11 .  175175 160166 107109 170172 153153 129129 208208 224224 205205 115117 153153 133133 124
```

1.2 Intro to diversity analysis

Analysis based on the frequency of alleles

Frequency by Pop: outputs allele frequencies at each locus by population

Frequency by Locus: outputs allele frequencies in each population with loci in columns

1.2 Intro to diversity analysis

Analysis based on the frequency of alleles

Heterozygosity

- Observed
- Expected

Wright's F statistics (F_{IS}, F_{ST} and F_{IT})

1.2 Intro to diversity analysis

Analysis of diversity on intra- and inter-population level

Basic indices:

- **MNA** – mean number of alleles per populations and loci
- **ENA** – effective allele number per populations and loci
- **AR** – total number of alleles per loci within each population
- **Heterozygosity** – average number within and across populations
- **Koeficient of inbreeding (F resp. F_{IS})** – within and across populations
- **F – statistics**
- **AMOVA** (analysis of molecular variance)

1.2 Intro to diversity analysis

Analysis of population genetic structure

Nei's genetic distances

Wright's F_{ST} index

Analysis of principal components (PCA)

- multivariate technique that allows one to find and plot the major patterns within a multivariate dataset e.g. multiple loci and multiple samples

1.3 Analysis of population structure and visualisation in R

- ➢ RStudio

- ➢Free …

  …case sensitive!
  …opposite slashes than Windows defaults

- ➢Package Adegenet

1.4 Practical excercise

- **Your turn!**

- Task description:

  1. How many populations and markers are stored in testset 2?

  2. Compute basic diversity indices for each of analysed population

  3. Make AMOVA analysis

  4. Make PCA analysis.

  5. Make DAPC analysis and visualise the group assignment probability of individuals

  6. Plot the relationships between individuals using NJ unrooted tree

Part 2: Analysis of diversity on genome-wide level using SNPs

2.1 Intro to PLINK

- **PLINK** is a free, open-source whole genome association analysis toolset
- to perform a range of basic, large-scale analyses in a computationally efficient manner

2.1 Intro to PLINK

- Runs in the command line and/or using R
- In Windows, Linux and Mac
- Options preceded by double dash “--”
- All options: http://pngu.mgh.harvard.edu/~purcell/plink/reference.shtml#options

2.2 Data processing in PLINK using R

Data manipulation

- **Task:** Quality control of genotyping data

QC criteria

1. missing genotypes per sample max. 10 %
2. min. SNPs call rate 90 %
3. min. minor allele frequency (MAF) 0.01

2.3 Analysis of population structure and visualisation in R

- **Task:** load data in to R and run analysis

load library for adegenet

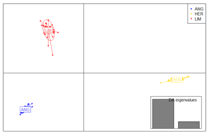
load input data

run DAPC analysis

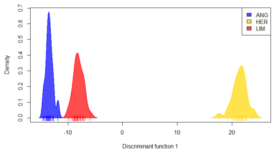
2.3 Analysis of population structure and visualisation in R

- **Task:** load data in to R and run analysis

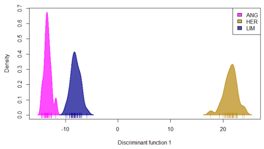
visualise differentiation between pops based on first two DFs

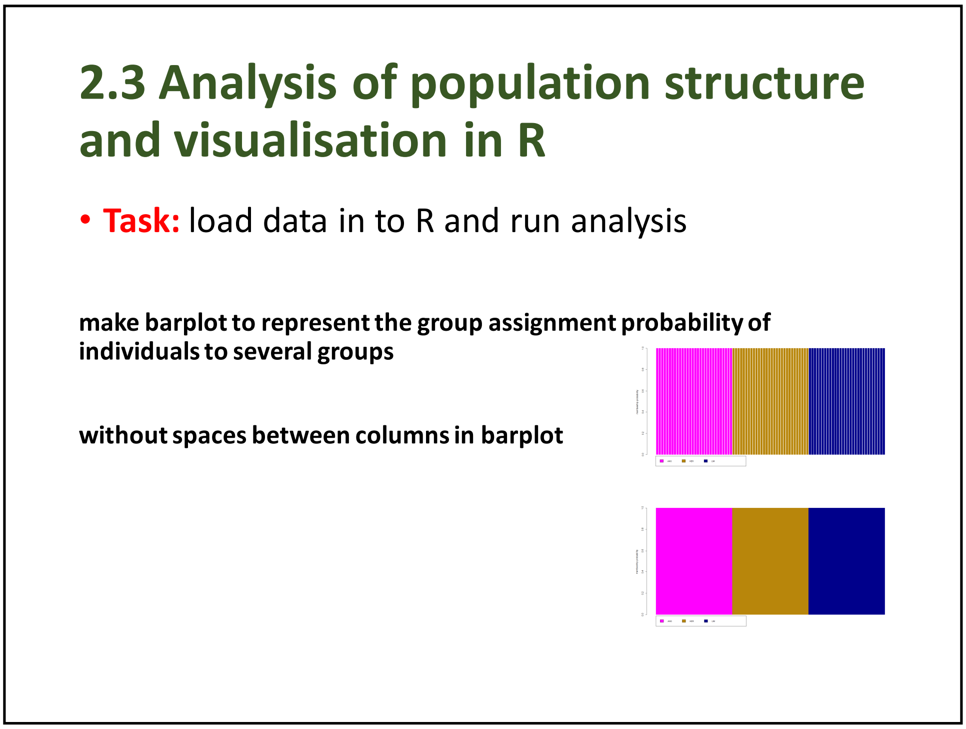


visualise differentiation between pops based on first DF



change colours in picture



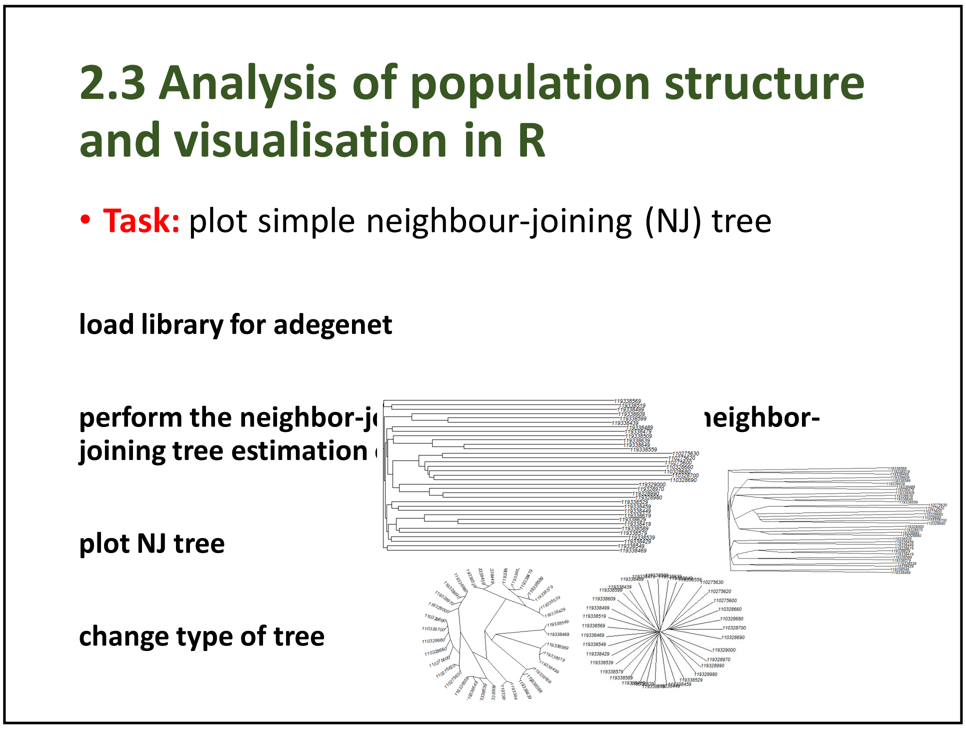


2.3 Analysis of population structure and visualisation in R

- **Task:** load data in to R and run analysis

make barplot to represent the group assignment probability of individuals to several groups

without spaces between columns in barplot



2.3 Analysis of population structure and visualisation in R

- **Task:** plot simple neighbour-joining (NJ) tree

load library for adegenet

perform the neighbor-j ıeighbor- joining tree estimation

plot NJ tree

change type of tree

2.3 Analysis of population structure and visualisation in R

- **Your turn!**

- Task description:

1. Make QC of data and create new dataQC .ped and .map file

- filter out all of animals and SNPs with more than 10% of missing data and MAF<0.01

2. Create input file for adegenet

3. Make DAPC analysis and visualise the group assignment probability of individuals

4. Plot the relationships between individuals using NJ unrooted tree

Thank you for your attention!