Population and landscape genomics workshop

Centre for Biodiversity Analysis

ANU

25th to the 28th of March 2014

Program

Tuesday 25th March

8.30am–5.30pm

From raw NGS data to SNPs - *with Matteo Fumagalli*

Location: ANU Commons, Lena Karmel Lodge

Wednesday 26th March

8.30am–5.30pm

Population structure and inference of demography - *with Anders Goncalves da Silva*

Location: ANU Commons, Lena Karmel Lodge

Thursday 27th March

8.30am–5.30pm

Selection scans and admixture analysis - *with Rose Andrew and Justin Borevitz*

Location: University House, ANU

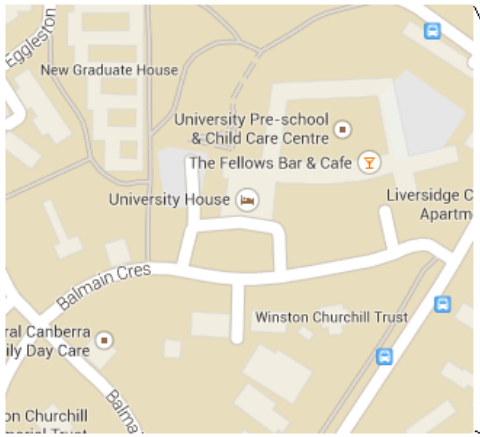
Friday 28th March

8.30am–5.30pm

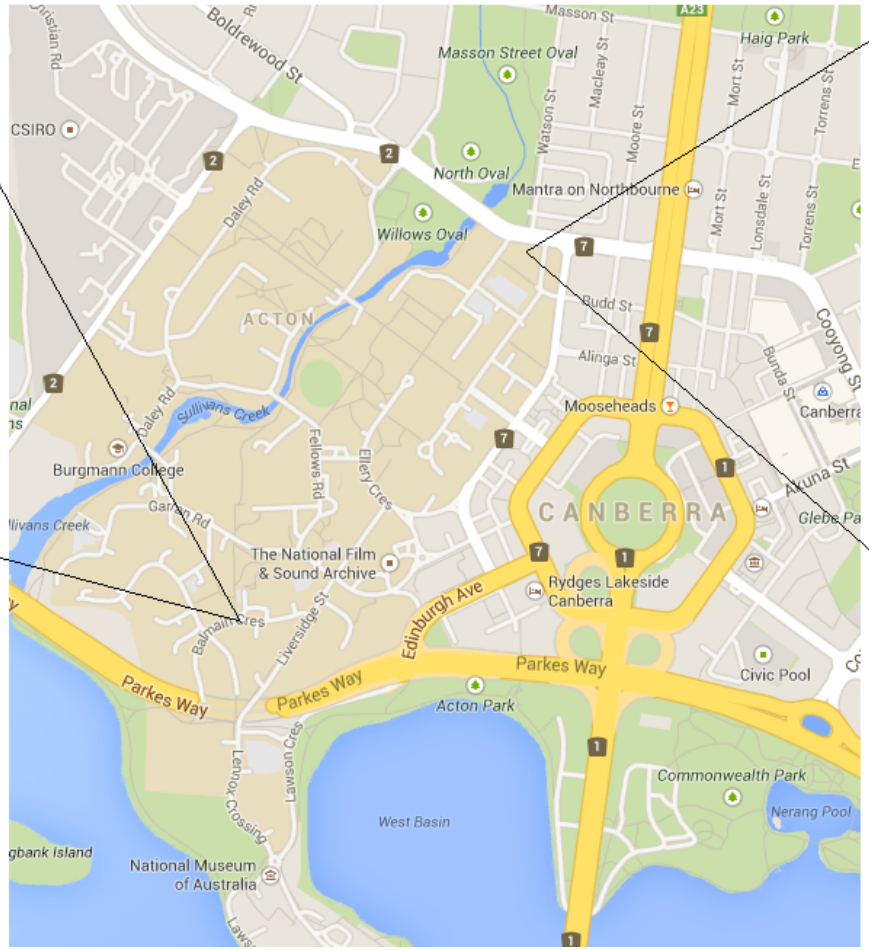
Landscape Genomics - *with Kevin Leempoel*

Location: ANU Commons, Lena Karmel Lodge

Map



University House, ANU



ANU Commons, Lena Karmel Lodge

Day 1: Tuesday 25th March

From raw NGS data to SNPs

Matteo Fumagalli

Day 2: Wednesday 26th March

Population structure and inference of demography

Anders Goncalves da Silva

Morning session

• Introduction to population genetic structure

- *Goal: get everyone on the same page by talking about what is population genetics, defining a locus, a gene, an allele, a genotype, evolutionary forces, and how these are connected to demography, environment, geography.*

• Establish motivation to estimate population genetic structure

- *Goal: stimulate thinking on good reasons for trying to estimate population genetic structure, and when large datasets might be useful in this exercise.*

• A look at inbreeding

- *Goal: generate an understanding of how population genetic structure arises.*

Mid-morning break

• How do we estimate population genetic structure?

- *Goal: instil an understanding of how population structure is estimated. Give a brief overview of the various estimators. Develop a gut feeling for what differences in allele frequencies mean, and the importance of many loci.*

- *Clarify distinction between classical F-statistics-based methods, Bayesian clustering approaches, and PCA-based clustering approaches.*

• What are the pitfalls and bias in estimating population genetic structure

- *Goal: use R to estimate population genetic structure, and demonstrate the balance between individuals, loci, and populations and how that affects Fstatistics.*

• Using STRUCTURE to infer population structure

- *Goal: how to setup an input file, choose a prior, check for convergence, and infer number of units.*

Lunch break

Afternoon session

• When *F-statistics* are just not enough

- *Goal: outline some of the shortcomings of F-statistics, and motive the use of the coalescent*

• Motivate the use of statistical modelling

- *Goal: introduce the use of statistical modelling to infer parameters. Use R to write a small MCMC, and estimate the parameters of a model.*

• What is the coalescent?

- *Goal*: *create an understanding of the use of genealogies within a statistical modelling framework to simultaneously estimate population size, demographic history, migration, natural selection, and population structure.*

• Write your own coalescent simulator, and estimate population size from the data

- *Goal: use R to write a simple coalescent simulator, and explore the consequences of more loci vs more genes in improving estimates of a parameter.*

Afternoon break

• Overview of LAMARC and MIGRATE-N

- *Goal: understand how a genealogy sampler works*

• Use MIGRATE-N to estimate migration and effective population size

- *Goal: prepare an input and parameter file to run in Migrate-N, get it to run, then examine some output files to understand the results*

Day 3: Thursday 27th March

Selection scans and admixture analysis

Justin Borevitz and Rose Andrew

Theory

Session 1
Fst outliers and admixture analysis (Rose)

Session 2
Pairwise haplotype sharing and composite likelihood ratio tests (Justin)

Practical

Session 1
Fst scans BayeScan analysis and hierarchical AMOVA

Session 2
Genomic cline analysis

Session 3
PHS and CLR analysis

Day 4: Friday 28th March

Practical work in landscape genomics

Leempoel Kevin & Joost Stéphane

Theoretical session

1. Introduction
    a. Brief history of Landscape genetics (LG) and its major publications in the last 10 years.
    b. Major differences between adaptive LG and population genetics approaches in the detection of neutral and adaptive variation
    c. Consequences of next generation sequencing datasets on the methods used to detect natural selection

2. Methods
    Our course will be focused on correlative approaches. This part will thus show what kind of genetic and environmental data we need and which are the relevant statistic approaches.

    a. Sources of environmental data
        i. Existing sources
            1. Climatic variables (CRU, WorldClim, interpolated regional datasets)
            2. Satellite imagery, ground cover and soil maps
            3. DEM derived variables (Existing, LIDAR and stereophotogrammetry)
        ii. Fieldwork sources
            1. Loggers (temperature, humidity, soil moisture, solar radiation)
            2. Indirect ecological indicators

    b. Genetic data
        i. Recoding/Filtering alleles and genotypes
        ii. Data transformation
        iii. Note on linkage disequilibrium

    c. Statistical approaches
        i. Computing associations between genetic data and environmental variables
            1. Correlative approaches (GLM) using SAMβada
            2. A note on statistical power (false positives)

ii. Spatial statistics
    1. Concerns regarding spatial autocorrelation. Comparison of results between methods and analysis of spatial dependency between results.
    2. LISA
    3. GWR
    4. LocalDiff
iii. Inclusion of population structure
    1. Admixture
    2. Correlative approaches including population structure (SGLMM, LFMM, Bayenv)
    3. Bivariate models in SAMβada using population structure and environmental variables.

3. Example of studies
    a. Goats in Morocco (NextGen project)
    b. *Arabis Alpina* in the Alps (Intrabiodiv project)
    c. *Biscutella Laevigata* at Les Rochers-de-Naye (CH) (local scale)
    d. Loblolly pine in the US (Eckert et al. 2010)

4. Conclusion and perspectives
    a. Concluding remark on common findings and differences between datasets
    b. Challenges in landscape genomics in the coming years

Practical session

1. Acquiring environmental information at sample's locations
    a. Choice of a coordinate system
    b. Retrieving environmental information from major internet sources
    c. Computing Digital Elevation Model environmental variables
    d. Extracting environmental variables values at sampling locations
    e. Exporting data
    f. PCA to eliminate multi collinearity

2. Identifying loci under selection with SAMβada
    a. Transforming data from PLINK or others to SAMβada and LFMM
    b. Parameters of SAMβada
    c. Console commands

3. Computing membership coefficients and evaluate structure influence using multivariate models.

    a. Admixture
    b. LocalDiff
    c. Multivariate models in SAMβada

4. Identifying loci under selection with approaches including directly population structure
    a. LFMM

5. Comparative analysis of detected loci

6. Spatial structure analysis of detected loci
    a. Indicator of spatial autocorrelation using Univariate LISA (Local Indicator of Spatial Association)
    b. Bivariate LISA of the most relevant associations

7. Genomic position of detected loci and potential gene function
    a. Comparison with results from (Eckert et al. 2010)

8. Producing results maps in a GIS
    a. Opening data in Quantum GIS (shape, raster, text delimited)
    b. Creation of a map (layers, legend, …, output format)

Notes

Notes