A bit about me... PhD
A bit about me…PhD
A bit about me...Postdoc 1
A bit about me… Postdoc 1
A bit about me…Postdoc 2
A bit about me…Selling Out

<table>
<thead>
<tr>
<th>SNP:</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ind 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ind 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Ind 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Ind 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
A bit about me… Selling Out

<table>
<thead>
<tr>
<th>Species:</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Site 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Site 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Site 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Population Genetics – What the?!

How is genetic variation distributed across a species’ range?

How is genetic variation distributed across the genome?

Where does genetic variation come from? Why does it persist?
What do we mean by a ‘population’

Panmixia

\[ M = M \]
What do we mean by a ‘population’

Continuous spatial populations

M < M
What do we mean by a ‘population’

Continuous spatial populations
What do we mean by a ‘population’

Continuous spatial populations
What do we mean by a ‘population’

Discrete spatial populations
What do we mean by a ‘population’

Discrete spatial populations
Population Genetics – What the?!

- Recombination
- Selection
- Mutation
- Gene Flow
- Drift
Mutation

Mutation is **RANDOM** (with respect to fitness)

Mutation is **NON-RANDOM** (with respect to location)

The ultimate source of novel variation

Often we are interested in:

- \( \text{Pop-scaled mutation rate} = 4N\mu \)
- \( \text{Distribution of fitness effects} = DFE \)
DNA fragility in the parallel evolution of pelvic reduction in stickleback fish

Kathleen T. Xie¹,²,³, Guliang Wang⁴, Abbey C. Thompson¹,⁵, Julia I. Wucherpfennig¹, Thomas E. Reimchen⁶, Andrew D. C. MacColl⁷, Dolph Schluter⁸, Michael A. Bell⁹, Karen M. Vasquez⁴, David M. Kingsley¹,¹⁺
Selection - Positive

Reduces variation within populations (linked selection)

Increases variation among populations (unless ‘global’)
Purifying selection removes deleterious variants

Deleterious mutations take linked neutral variation with them

Gene density increases deleterious likelihood

Low recombination/tight linkage removes more neutral variation
Hotspots and maintenance of genetic variation in the genome can be driven by different forms of selection, e.g. Negative Frequency Dependent Selection or Heterozygote Advantage. Commonly associated with immune genes, e.g. MHC in response to pathogens.
A large and diverse autosomal haplotype is associated with sex-linked colour polymorphism in the guppy

Gene Flow

Transfer of alleles among populations.

Can be asymmetric, unidirectional and episodic.

Works against population sub-division.

Inflates diversity by increasing effective population size (novel diversity)
Genetic Drift

Stochastic fluctuations in allele frequency

Probabilistic, determined by population size and selection

Drives population subdivision and neutral divergence

Neutral fixation = $1/2N$
Recombination

~1 CO per chromosome = more recombination on smaller chromosomes

Moderates linkage across the genome

Low recombination resists introgression

Recombination generates novel haplotypes

Conserved landscapes, but variation as well
What determines genetic diversity?

Figure 2 | Overview of determinants of genetic diversity. Effective population size, mutation rate and linked selection are the main factors affecting diversity. These factors are in turn governed by several other parameters. The direction of correlation is indicated by the + and – symbols. Selfing, self-fertilization.

Ellegren and Galtier 2017
Applications for population genomics

Conservation genomics
• Understanding connectivity and structure among populations
• Quantifying temporal changes in population diversity and extinction risk

Agricultural genomics
• Identifying novel variation within wild crop relatives
• Genomic selection and guided breeding

Medical genomics
• Tracking pathogen evolution (e.g. Covid!)
• Understanding disease risk of individuals and populations

Evolutionary biology
• Identifying candidate loci under selection and adaptive genes
• Identifying barriers to gene flow and understanding speciation
• Understanding the evolutionary history of species
Pi (nucleotide diversity)

(a) Pairwise differences between genotypes

Total differences

No. comparisons (n choose 2)

SEQUENCE LENGTH = 12

CASE 1: MISSING DATA ASSUMED IN Variant

STRUCTION

Mean numerator 4/12 = 0.333
Mean denominator 72/12 = 6.000

πraw 0.333/6.000 = 0.055

πper site 0.055/12 = 0.004

SEQUENCE LENGTH = 9

CASE 2: CALCULATION ADJUSTED FOR MISSING DATA (pixy)

Mean numerator 4/9 = 0.444
Mean denominator 40/9 = 4.444

πraw 0.444/4.444 = 0.099

πper site 0.099/9 = 0.011

PIXY: Korunes and Samuk 2021
Pi (nucleotide diversity)

Tajima’s D

\[ D = \frac{\pi - S/a_1}{\sqrt{V}} \]

van der Zee et al. 2022
PopGen Toolkit – Summary Statistics

Pi (nucleotide diversity)

Tajima’s D

\( F_{ST} \)

\[ \pi_{\text{Between}} - \pi_{\text{Within}} \]

\( \pi_{\text{Between}} \)

Hohenlohe et al. 2010
PopGen Toolkit – Summary Statistics

Pi (nucleotide diversity)

Tajima’s D

$F_{ST}$

ROH (Runs of Homozygosity)

Ceballos et al. 2018
PopGen Toolkit – Site Frequency Spectrum

A summary of the frequency of allele counts within a population

Folded (based on maf)

Unfolded (based on derived)

Do we know ancestral/derived allele?
PopGen Toolkit – Site Frequency Spectrum

Folded - Neutral

Folded – Selection

Folded – Balancing
PopGen Toolkit – Site Frequency Spectrum

2D/Joint
Ancestral recombination graphs

- ARGweaver
- RELATE
- tsinfer

Hejase et al. 2020
Ancestral recombination graphs

- ARGweaver
- RELATE
- tsinfer

Machine-learning

- Classification algorithms (e.g. diploS/HIC)
- Neural networks (e.g. RELERNN)
- Image processing (e.g. IntroUNET)

Adrion et al 2020

Ray et al 2023
PopGen Toolkit – The future... (is now)

Ancestral recombination graphs
- ARGweaver
- RELATE
- tsinfer

Machine-learning
- Classification algorithms
- Neural networks (e.g. RELERNN)
- Image processing (e.g. IntroUNET)

Pangenomics/SV (beyond the SNP...)
- How applicable are our models?
- What is a population/species in a pangenome sense?
Using a simulated VCF, we’ll work through a simple popgen workflow covering:

• Describing population structure
  • PCA
  • DAPC

• Quantifying population diversity
  • Nucleotide diversity

• Identifying selective sweeps
  • SweepFinder2

• Testing adaptive introgression
  • D-Suite (ABBA-BABA / D-statistics)