Inbreeding and runs of homozygosity
Definitions

• Inbreeding, the mating between two related individuals
  – More related than the average relatedness in the population

• Inbreeding coefficient: an individual measure
  – See later

• Inbreeding depression
  – Reduced fitness because of lower survival, mating, and/or reproduction in the progeny of related individuals compared to that of unrelated individuals
Inbreeding coefficient

• The correlation between parental gametes that unite to form an individual (Wright, 1922)
  – relative to the total array of such gametes in a random sample from the reference population
Coefficient of Coancestry

- If two alleles are randomly drawn in two individuals $x$ and $y$ (one per individual), the probability $\Theta_{xy}$ that these alleles are IBD is called the coefficient of coancestry.
- Sometimes called coefficient of kinship or coefficient de parente.
Inbreeding coefficient

• This coefficient is also the probability that the two alleles of a gene (one locus) inherited by one individual $z$, offspring of parents $x$ and $y$, are IBD

• This quantity is Wright’s (1922) inbreeding coefficient, $f_z$

• The parent’s coefficient of coancestry is equal to the inbreeding coefficient

• The probability that an individual carries two IBD copies at a given neutral locus (Malécot, 1948)
Estimation of genetic relatedness $\Theta_{xy}$

- Using path analysis:
  - Identify all lines (paths) connecting $x$ and $y$ through a common ancestor $a$ within the genealogy
  - $n$ is the number of individuals in the path (including $x$ and $y$)
  - $F_a$ is the inbreeding coefficient of the common ancestor $a$

$$\theta_{xy} = \left(\frac{1}{2}\right)^n (1 + F_a)$$
Path analysis

• Example: parent – child

\[ \theta_{xy} = \left( \frac{1}{2} \right)^n (1 + F_a) = 0.25 \]
Path analysis

- Example: half-sibs

\[ \theta_{xy} = \left( \frac{1}{2} \right)^n (1 + F_a) = 0.125 \]

- \( n = 3 \)
Extension

• Extension to the entire pedigree:
  – Identify all common ancestors and their inbreeding coefficient $F$
  – Identify all path between $x$ and $y$ going through $a$
  – Determine $n$ for each path
  – The coefficient of coancestry is obtained by summing all probabilities:

$$\theta_{xy} = \sum \left( \frac{1}{2} \right)^n (1 + F_a)$$
Example

\[ \theta_{15} = 0.125 \quad \text{Grand-parent} \]
\[ \theta_{45} = 0.0625 \]
\[ \theta_{56} = 0.03125 \]
Example

\[ \theta_{15} = 0.125 \quad \text{Grand-parent} \]
\[ \theta_{45} = 0.0625 + 0.0625 = 0.125 \quad \text{Uncle} \]
\[ \theta_{56} = 0.03125 + 0.03125 = 0.0625 \quad \text{Cousin} \]
Estimating the inbreeding coefficient

- Using path analysis:
  - The inbreeding coefficient of \( I \) is the coefficient of coancestry between its parents (\( x \) et \( y \))
  - Identify all lines (paths) connecting \( x \) and \( y \) through a common ancestor \( a \) within the genealogy
  - \( n \) is the number of individuals in the path (including \( x \) and \( y \))
  - \( F_a \) is the inbreeding coefficient of the common ancestor \( a \)

\[
F_I = \left( \frac{1}{2} \right)^n \left( 1 + F_a \right)
\]
Example

\[ \theta_{56} = F_7 = 0.03125 + 0.03125 = 0.0625 \]
Causes of inbreeding

• Genetic drift
  – Related to effective population size
  – Demographic events: bottleneck, founder effects, population decline
  – Breed creation, selection

• Non-random mating
  – Selfing (reproduction system)
  – Experimental lines
  – Selection
  – Cultural aspect (family, communities)
  – Sub-populations (no mixing)
Consequences of inbreeding

- Increased homozygosity
  - More chances for homozygosity at rare deleterious alleles
  - Genetic defects (numerous examples) – most striking effects
  - Many examples in livestock and dogs but also human

- Decreased heterozygosity
  - Reduced heterozygosity for alleles presenting overdominance / heterozygous advantage

- Inbreeding depression
  - More general consequence
Inbreeding depression

• Definition
  – Reduced fitness because of lower survival, mating, and/or reproduction in the progeny of related individuals compared to that of unrelated individuals
  – Reduction of the mean phenotypic value for fitness trait and characters connected with reproductive capacity or survival

• Inbreeding depression may impact any trait
Inbreeding depression

• Genetic basis (Charlesworth & Willis, 2009)
  – Partial dominance: increased expression of recessive deleterious alleles
  – Overdominance: superiority of heterozygotes over two kinds of homozygotes (balancing selection)
  – (Epistasis)
Inbreeding depression

<table>
<thead>
<tr>
<th>Model</th>
<th>Parent genotypes</th>
<th>F₁ hybrid genotypes and their fitness (or quality) relative to the parent genotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recessive deleterious mutations</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dominance hypothesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single locus</td>
<td>(A/A) x (a/a)</td>
<td>(A/a)</td>
</tr>
<tr>
<td>Intermediate fitness but above the parental average (homozygote shows inbreeding depression)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple loci (effects of different mutant alleles marked in hybrids)</td>
<td>(A/A) x (b/b)</td>
<td>(A/a) (B/b)</td>
</tr>
<tr>
<td>High fitness (heterosis and inbreeding depression)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recessive deleterious mutations</strong> at closely linked loci</td>
<td>(A/b) x (a/b)</td>
<td>(A/b) (a/b)</td>
</tr>
<tr>
<td>Pseudo-overdominance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Higher fitness than the parent genotypes</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Single loci with heterozygous advantage</strong></td>
<td>(A₂/A₁) x (A₂/A₂)</td>
<td>(A₂/A₂)</td>
</tr>
<tr>
<td>True overdominance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High fitness (heterosis and inbreeding depression in homozygote)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Estimating the inbreeding coefficient

• Pedigree-based methods
  – Tabular method, path analysis
  – Pedigree required (errors)
  – Expected inbreeding

• Marker-based approaches
  – Realized inbreeding
  – Global and locus-specific estimators
  – Numerous approaches
Marker-based approaches

• Similar to marker-based genetic relatedness measures
• Maximum-likelihood estimators
  – Coancestry or related R-package (see also exercises)
  – Computationally intensive (slow)
• Method-of-moment estimators
  – Regression
  – Correlations (Ritland, 1996) / Uniting gametes (GCTA, plink)
  – Expected / observed homozygosity
• Diagonal elements of the genetic relationship matrix
  – Similar to pedigree-based
General principle

• In absence of inbreeding, genotypes are assumed under HWE
• The inbreeding coefficient $F$, also represents the proportion of the genome IBD
  – For IBD loci, heterozygous are not observed and homozygous according to the respective genotype frequencies
• Find $F$ that best fit the data (ML, MSE, ...)
• For instance:

$$P(AA|F) = (1 - F)f_A^2 + Ff_A$$
Maximum-likelihood

- Requires the nine condensed identity states
- Only states 1-6 include inbreeding
Expected / observed heterozygosity

- Count number of observed homozygous genotypes $H_O$
- Estimate expected number of homozygous genotypes $H_E$
  - Using allele frequencies
  - Using HWE
    \[
    H_E = \sum_i^N (1 - 2f_i (1 - f_i))
    \]
- Method-of-moments estimator:
  \[
  H_O = NF + (1 - F)H_E
  \]
  \[
  F = \frac{H_O - H_E}{N - H_E}
  \]
Softwares

• R package related (several estimators)

• PLINK and GCTA
  – Three shared estimators: GRM-based, excess homozygosity, uniting gametes
  – Option in PLINK --ibc

• PLINK:
  – Expected / observed homozygosity (option --het)
  – ROH (see later)
$F$ is associated with segments

- “Probability to sample two IBD alleles” (Malécot, 1948)
Inbreeding creates ROH

- Runs of homozygosity (ROH)
  - Autozygous or homozygous-by-descent (HBD) segments
Introduction

• Runs of homozygosity (ROH)
  – Autozygous or homozygous-by-descent (HBD) segments

![Diagram showing the relationship between ancestral chromosome and HBD segments, ROH, and N1 and N2 ancestors.]

- Ancestral chromosome
- HBD segments, ROH
- Path 1: $N_1$ ancestors
- Path 2: $N_2$ ancestors

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB

AABABBABABABB
Example of ROH

- Runs of homozygosity (ROH)
  - ‘Long’ series of successive homozygous genotypes in one individual

```
200210012001220010021100001200102201021022200021202200010221100220021210002
10012000220020022202000200002202202200020202200020220000202200022002021002
2001200001022110022002121000220220220022010210222000212010020220220002
0222000220220002220000220020210001210210012000220020022022000220002201010201202
220002201021022200021201002022022000100211000012001022010210222000212022121
```
Genomics advances the study of inbreeding depression in the wild
Genomics advances the study of inbreeding depression in the wild
Some properties

- Distribution of ROH / HBD segments:
  - Can be used to estimate $F$
  - Proportions of the genome in ROH
  - The sequence of observations (link) is used whereas other methods consider marker independent
  - Provides also local estimates: for a specific locus (useful for autozygosity mapping)
Applications using HBD segments

- Estimate inbreeding coefficient
- Study inbreeding depression
- Homozygosity mapping (recessive effects)
- Measure genetic diversity
- Reveal population demographic history
- Identify signatures of selection
Figure 1. Global distribution of long runs of homozygosity.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013996
Identification of ROH

• Rule-based methods
  – Arbitrary definitions: min window length, win number of SNPs, max heterozygotes, max spacing, min density, etc.
  – Ideally, redefine for each data set (population, marker map, marker informativity)
  – Allele frequencies and genetic map not used
  – Inappropriate for low-fold sequencing
  – Very popular (implemented in PLINK)
Identification of ROH

- Likelihood-based approaches
  - Computes the probability to observe the genotypes in HBD versus non-HBD segments
  - Performs a LOD score using allele frequencies and genotyping error rates ($\log_{10}(H1/H0)$)
  - Requires the definition of an optimal window size (bimodal distribution)
  - Genetic map not used
  - Inappropriate for low-fold sequencing
Conditional probabilities

- Probability of genotype given HBD status (emission prob.):

<table>
<thead>
<tr>
<th></th>
<th>HBD</th>
<th>Non-HBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_iA_i$</td>
<td>$(1-\varepsilon)p_i$</td>
<td>$p_i^2$</td>
</tr>
<tr>
<td>$A_iA_j$</td>
<td>$\varepsilon$</td>
<td>$2p_ip_j$</td>
</tr>
</tbody>
</table>
Identification of ROH

• Hidden Markov model
  – Uses marker allele frequencies, genotyping error rates, genetic map and expected length of HBD segments
  – Provides a probability for HBD
  – Test automatically for all windows
  – Genetic map not used
  – Can handle low-fold sequencing and heterogeneous data
Hidden Markov models

- Models the genome as a mosaic of HBD (inbred) and non-HBD segments (e.g., Leutenegger, 2003 - AJHG)

10020110102111100200202021211012110210110120101210011
Hidden Markov models

• Models the genome as a mosaic of HBD (inbred) and non-HBD segments (e.g., Leutenegger, 2003 - AJHG)
Emission probabilities

- Probability of genotype given HBD status (emission prob.):

<table>
<thead>
<tr>
<th></th>
<th>HBD</th>
<th>Non-HBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_iA_i$</td>
<td>$(1-\varepsilon)p_i$</td>
<td>$p_i^2$</td>
</tr>
<tr>
<td>$A_iA_j$</td>
<td>$\varepsilon$</td>
<td>$2p_i p_j$</td>
</tr>
</tbody>
</table>
Transition probabilities

• Absence of coancestry change is $e^{-\alpha}$ ($\alpha$ is the transition rate: recombination rate & time to common ancestor)
• Prob. new coancestry is HBD is $F$
• Prob. New coancestry is non-HBD equals $(1-F)$
Transition probabilities

- Transition matrix:

<table>
<thead>
<tr>
<th></th>
<th>HBD</th>
<th>Non-HBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBD</td>
<td>$1 - e^{-\alpha}$ $(1 - F)$</td>
<td>$(1 - e^{-\alpha})(1 - F)$</td>
</tr>
<tr>
<td>Non-HBD</td>
<td>$(1 - e^{-\alpha})F$</td>
<td></td>
</tr>
</tbody>
</table>
## Transition probabilities

- Transition matrix:

<table>
<thead>
<tr>
<th></th>
<th>HBD</th>
<th>Non-HBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBD</td>
<td>$e^{-\alpha}$</td>
<td>$(1-e^{-\alpha})(1-F)$</td>
</tr>
<tr>
<td>Non-HBD</td>
<td>$(1-e^{-\alpha})F$</td>
<td>$e^{-\alpha}$</td>
</tr>
</tbody>
</table>
Transition probabilities

- Transition matrix:

<table>
<thead>
<tr>
<th></th>
<th>HBD</th>
<th>Non-HBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBD</td>
<td>(e^{-\alpha} + (1-e^{-\alpha})F)</td>
<td>((1-e^{-\alpha})(1-F))</td>
</tr>
<tr>
<td>Non-HBD</td>
<td>((1-e^{-\alpha})F)</td>
<td>(e^{-\alpha} +(1-e^{-\alpha})(1-F))</td>
</tr>
</tbody>
</table>
Transition probabilities

• Absence of coancestry change is $e^{-\alpha}$ ($\alpha$ is the transition rate: recombination rate & time to common ancestor)
• Prob. new coancestry is HBD is $F$
• Prob. New coancestry is non-HBD equals $(1-F)$

$F$ is the inbreeding coefficient: the equilibrium HBD probability (proportion HBD segments)
Transition probabilities

• Absence of coancestry change is $e^{-\alpha}$ ($\alpha$ is the transition rate: recombination rate & time to common ancestor)

• Prob. new coancestry is HBD is $F$

• Prob. New coancestry is non-HBD equals $(1-F)$

• $F$ is the inbreeding coefficient: the equilibrium HBD probability (proportion HBD segments)

• $\alpha$ is equal to $Rr_{kl}$ ($r_{kl}$ is the recombination rate in Morgans, and $R$ is the rate of ‘co-ancestry’ change per Morgans)
Length of HBD segment

\[ G = N1 + N2 \] (size of the inbreeding loop in generations / \( N_i \) includes MRCA)

Expected \( L = 1/G \) (Morgans) – exponential distribution with rate \( G \)
Extension to WGS data

• Replace genotypes in emission probabilities:
  – Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF):

\[
P(\text{Data} \mid \text{HBD}) = p_i P(A_iA_i \mid \text{Data}) + p_j P(A_jA_j \mid \text{Data}) + \epsilon P(A_iA_j \mid \text{Data})
\]
Extension to WGS data

• Replace genotypes in emission probabilities:
  – Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF)
  – Use allele counts (allele depth – AD)

\[ P(AD \mid \text{HBD}) = p_i P(AD \mid A_i A_i) + p_j P(AD \mid A_j A_j) \]

\( \epsilon \) included
Extension to WGS data

• Replace genotypes in emission probabilities :
  – Use genotype likelihoods or phred scores incorporating uncertainty on genotype calls (from VCF)
  – Use allele counts (allele depth – AD)

• Recent implementations:
  – BCFtools / RoH (Narasimhan et al. – Bionformatics, 2016)
  – ngsF-HMM (Viera et al. – Bionformatics, 2016)
Hidden Markov models

- Uses marker allele frequencies, genotyping error rates, genetic map and expected length of HBD segments
- Provides a probability for HBD
- Test automatically for all windows
- Genetic map not used
- Can handle low-fold sequencing and heterogeneous data
Conclusions

• Management of inbreeding
  – Livestock species
  – Conservation program

• Studying inbreeding depression

• HBD segments or ROH
  – Information on demographic history and trait architecture