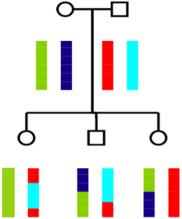
Inference of pairwise relatedness

Ida Moltke, Naples, April 2018



Outline

1. Introduction
   - Goal and motivation
   - Definition of key concepts
   - Overview of current methods for inference

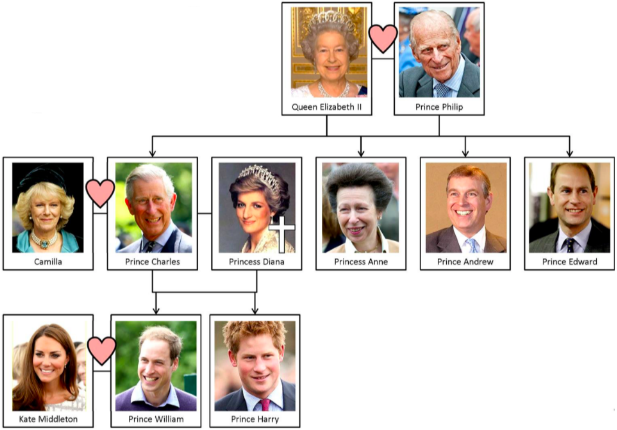
2. A simple maximum likelihood solution for called genotypes
   - Overview and intuition
   - Model
   - ML inference

3. Problem and more advanced methods
   - Problems with standard methods
   - NGSrelate (for NGS data)
   - RelateAdmix (for admixed individuals)

4. Exercise

Goal

- We want to infer how two individuals are related based on DNA
- For example:



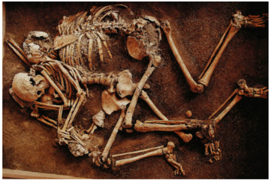
Motivation

- Many methods assume individuals are unrelated
- Violations can lead to wrong conclusions!

Motivation

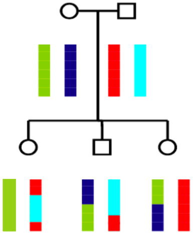
- Many methods assume individuals are unrelated
- Violations can lead to wrong conclusions!
- Can be of interest in itself, e.g.:
    - forensics (paternity testing, DNA evidence)

    - can help reveal past or current cultural/social practices



Key concept: Identity by Descent (IBD)

- One definition: DNA sequence **identity due to recent common ancestry**

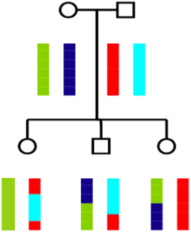


The individuals are IBD in regions where they have the same color

- Note with this definition:
    - Individuals can be identical in a locus without being IBD

Key concept: Identity by Descent (IBD)

- One definition: DNA sequence **identity due to recent common ancestry**

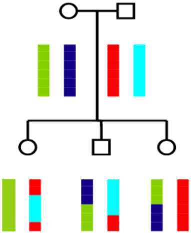


The individuals are IBD in regions where they have the same color

- Note with this definition:
    - Individuals can be identical in a locus without being IBD
    - Hence IBD sharing cannot be observed, **so needs to be inferred**

Key concept: Identity by Descent (IBD)

- One definition: DNA sequence **identity due to recent common ancestry**

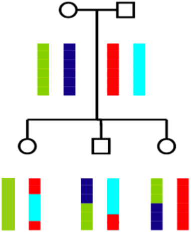


The individuals are IBD in regions where they have the same color

- Note with this definition:
  - Individuals can be identical in a locus without being IBD
  - Hence IBD sharing cannot be observed, **so needs to be inferred**
  - Any non-inbred pair will in a locus **share 0, 1 or 2 alleles IBD**

Key concept: Identity by Descent (IBD)

- One definition: DNA sequence **identity due to recent common ancestry**



The individuals are IBD in regions where they have the same color

- Note with this definition:
  - Individuals can be identical in a locus without being IBD
  - Hence IBD sharing cannot be observed, **so needs to be inferred**
  - Any non-inbred pair will in a locus **share 0, 1 or 2 alleles IBD**
  - And **the closer related, the more IBD sharing**

IBD and relatedness coefficients (R)

- $\mathbf{R} = (\mathbf{k_0}, \mathbf{k_1}, \mathbf{k_2})$: fractions of genome with 0, 1 and 2 alleles IBD

IBD and relatedness coefficients (R)

- $\mathbf{R} = (\mathbf{k_0}, \mathbf{k_1}, \mathbf{k_2})$: fractions of genome with 0, 1 and 2 alleles IBD
- We expect different relationships to have different values of R, e.g.:

Relationship	k_0	k_1	k_2
Monozogotic twins	0	0	1
Parent-offspring	0	1	0
Siblings	0.25	0.5	0.25
Halfsiblings/Uncle-nephew/grandparent-child	0.5	0.5	0
First cousins	0.75	0.25	0
Second cousins	0.9375	0.0625	0
Unrelated	1	0	0

- Hence we can (often) use R to infer how two individuals are related

IBD and relatedness coefficients (R)

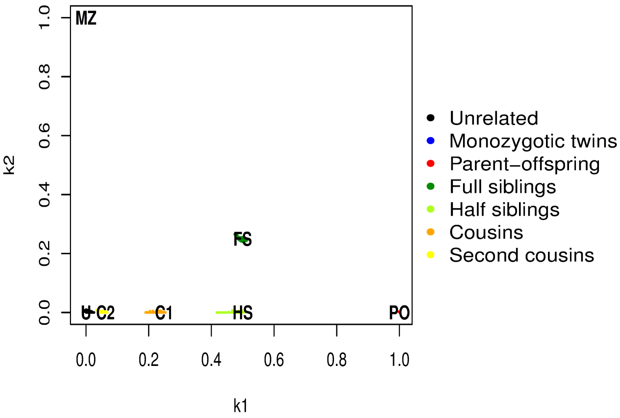
- $\mathbf{R} = (\mathbf{k_0}, \mathbf{k_1}, \mathbf{k_2})$: fractions of genome with 0, 1 and 2 alleles IBD
- We expect different relationships to have different values of R, e.g.:

Relationship	k_0	k_1	k_2
Monozogotic twins	0	0	1
Parent-offspring	0	1	0
Siblings	0.25	0.5	0.25
Halfsiblings/Uncle-nephew/grandparent-child	0.5	0.5	0
First cousins	0.75	0.25	0
Second cousins	0.9375	0.0625	0
Unrelated	1	0	0

- Hence we can (often) use R to infer how two individuals are related
- There are other measures like kinship
- But they are all functions of R so we will focus on R

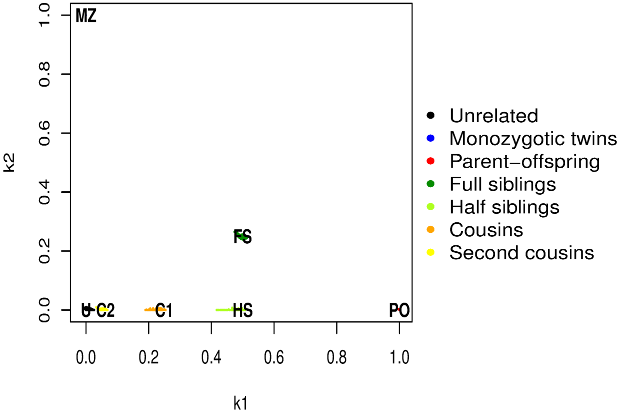
How?

- This can e.g. be done by plotting k_1 against k_2:



How?

- This can e.g. be done by plotting k_1 against k_2:



- However, we need to estimate R (because it can't be observed)

How do we infer R?

There are a number of estimators:

- several method of moments estimators (e.g. PLINK)
- some maximum likelihood (ML) estimators (e.g. Thompson, 1975)
- most well known program is probably PLINK

Importantly, almost all of them are based on several assumptions, namely:

- that the individuals are from a **diploid species**
- that the individuals are **not inbred**
- that we have **called genotypes** for the two individuals in L loci
- that the individuals are from the **same homogenous population**
- that the population **allele frequencies are known**
(or that you have provide enough samples to estimate them)

Outline

1. Introduction
   - Goal and motivation
   - Definition of key concepts
   - Overview of current methods for inference

2. A simple maximum likelihood solution for called genotypes
   - Overview and intuition
   - Model
   - ML inference

3. Problem and more advanced methods
   - Problems with standard methods
   - NGSrelate (for NGS data)
   - RelateAdmix (for admixed individuals)

4. Exercise

A simple ML approach (overview)

1. Define a model that allows us to calculate $P(data|R)$ (likelihood)
2. Estimate R by finding the R value that maximises $P(data|R)$ (MLE)

Intuition behind (what information is used)

Although we cannot observe IBD, only genotypes, we note that

- **If alleles are not identical they cannot be IBD**
- If alleles are identical it could be due to either IBD or chance
- If an allele is frequent, identity will occur often simply by chance
- If an allele is rare, identity will only occur rarely by chance
- **So the rarer the allele, the more will identity makes us believe the individuals are IBD**

Model

- We want a model that allows us to calculate P(data|R) (data=genotypes)
- Notation: for a pair of non-inbred individuals genotyped in L loci we let
  - G_j be their genotypes in locus j, e.g. $G_j = (AA, aa)$ (observed)
  - Z_j indicate how many alleles they share IBD in locus j (unobserved)

Model

- We want a model that allows us to calculate P(data|R) (data=genotypes)
- Notation: for a pair of non-inbred individuals genotyped in L loci we let
  - G_j be their genotypes in locus j, e.g. $G_j = (AA, aa)$ (observed)
  - Z_j indicate how many alleles they share IBD in locus j (unobserved)

- For a single locus, j, we can write:

$$\begin{aligned} P(G_j|R) = & P(Z_j = 0|R)P(G_j|Z_j = 0) + \\ & P(Z_j = 1|R)P(G_j|Z_j = 1) + \\ & P(Z_j = 2|R)P(G_j|Z_j = 2) \quad = \textstyle\sum_{z=0}^{2} P(Z_j = z|R)P(G_j|Z_j = z) \end{aligned}$$

Model

- We want a model that allows us to calculate P(data|R) (data=genotypes)
- Notation: for a pair of non-inbred individuals genotyped in L loci we let
  - G_j be their genotypes in locus j, e.g. $G_j = (AA, aa)$ (observed)
  - Z_j indicate how many alleles they share IBD in locus j (unobserved)

- For a single locus, j, we can write:

$$\begin{aligned} P(G_j|R) = {} & P(Z_j = 0|R)P(G_j|Z_j = 0) + \\ & P(Z_j = 1|R)P(G_j|Z_j = 1) + \\ & P(Z_j = 2|R)P(G_j|Z_j = 2) \quad = \textstyle\sum_{z=0}^{2} P(Z_j = z|R)P(G_j|Z_j = z) \end{aligned}$$

- Note that $P(Z_j = z|R)$ is simply k_z for all $z \in \{0, 1, 2\}$, so we get

$$P(G_j|R) = \textstyle\sum_{z=0}^{2} k_z P(G_j|Z_j = z)$$

Model $(P(G_j|Z_j = z))$

- Assuming HWE we can derive:

G_j	Z_j=0	Z_j=1	Z_j=2	
AA,AA	f_A^4	f_A^3	f_A^2	$\forall A$
AA,aa	$2f_A^2 f_a^2$	0	0	$A \neq a$
AA,Aa	$4f_A^3 f_a$	$2f_A^2 f_a$	0	$A \neq a$
Aa,Aa	$2f_A^2 f_a^2$	$f_A^2 f_a + f_A f_a^2$	$2f_A f_a$	$A \neq a$

Model ($P(G_j|Z_j = z)$)

- Assuming HWE we can derive:

G_j	Z_j=0	Z_j=1	Z_j=2	
AA,AA	f_A^4	f_A^3	f_A^2	$\forall A$
AA,aa	$2f_A^2 f_a^2$	0	0	$A \neq a$
AA,Aa	$4f_A^3 f_a$	$2f_A^2 f_a$	0	$A \neq a$
Aa,Aa	$2f_A^2 f_a^2$	$f_A^2 f_a + f_A f_a^2$	$2f_A f_a$	$A \neq a$

- E.g.
  1. $P(AA, AA|Z_j = 0) = P(AA)P(AA|AA, Z_j = 0) = P(AA)P(AA) = f_A^2 \times f_A^2$

Model ($P(G_j|Z_j = z)$)

- Assuming HWE we can derive:

G_j	Z_j=0	Z_j=1	Z_j=2	
AA,AA	f_A^4	f_A^3	f_A^2	$\forall A$
AA,aa	$2f_A^2 f_a^2$	0	0	$A \neq a$
AA,Aa	$4f_A^3 f_a$	$2f_A^2 f_a$	0	$A \neq a$
Aa,Aa	$2f_A^2 f_a^2$	$f_A^2 f_a + f_A f_a^2$	$2f_A f_a$	$A \neq a$

- E.g.
  1. $P(AA, AA|Z_j = 0) = P(AA)P(AA|AA, Z_j = 0) = P(AA)P(AA) = f_A^2 \times f_A^2$
  2. $P(AA, AA|Z_j = 2) = P(AA)P(AA|AA, Z_j = 2) = P(AA) \times 1 \quad = f_A^2 \times 1$

Model ($P(G_j|Z_j = z)$)

- Assuming HWE we can derive:

G_j	Z_j=0	Z_j=1	Z_j=2	
AA,AA	f_A^4	f_A^3	f_A^2	$\forall A$
AA,aa	$2f_A^2 f_a^2$	0	0	$A \neq a$
AA,Aa	$4f_A^3 f_a$	$2f_A^2 f_a$	0	$A \neq a$
Aa,Aa	$2f_A^2 f_a^2$	$f_A^2 f_a + f_A f_a^2$	$2f_A f_a$	$A \neq a$

- E.g.
  1. $P(AA, AA|Z_j = 0) = P(AA)P(AA|AA, Z_j = 0) = P(AA)P(AA) = f_A^2 \times f_A^2$
  2. $P(AA, AA|Z_j = 2) = P(AA)P(AA|AA, Z_j = 2) = P(AA) \times 1 \quad = f_A^2 \times 1$
  3. $P(AA, Aa\ |Z_j = 2) = P(AA)P(Aa|AA, Z_j = 2) = P(AA) \times 0 \quad = 0$

Model ($P(G_j|Z_j = z)$)

- Assuming HWE we can derive:

G_j	Z_j=0	Z_j=1	Z_j=2	
AA,AA	f_A^4	f_A^3	f_A^2	$\forall A$
AA,aa	$2f_A^2 f_a^2$	0	0	$A \neq a$
AA,Aa	$4f_A^3 f_a$	$2f_A^2 f_a$	0	$A \neq a$
Aa,Aa	$2f_A^2 f_a^2$	$f_A^2 f_a + f_A f_a^2$	$2f_A f_a$	$A \neq a$

- E.g.
  1. $P(AA, AA|Z_j = 0) = P(AA)P(AA|AA, Z_j = 0) = P(AA)P(AA) = f_A^2 \times f_A^2$
  2. $P(AA, AA|Z_j = 2) = P(AA)P(AA|AA, Z_j = 2) = P(AA) \times 1 \quad = f_A^2 \times 1$
  3. $P(AA, Aa\ |Z_j = 2) = P(AA)P(Aa|AA, Z_j = 2) = P(AA) \times 0 \quad = 0$

- Captures connection between IBD, genotypes and allele frequencies:
  - 3) captures that alleles have to be identical to be IBD
  - 1) and 2) captures that the rarer an allele is, the more will observing identity make us believe the individuals are IBD

ML inference based on the model

- **Assuming loci are independent** we get the full likelihood:

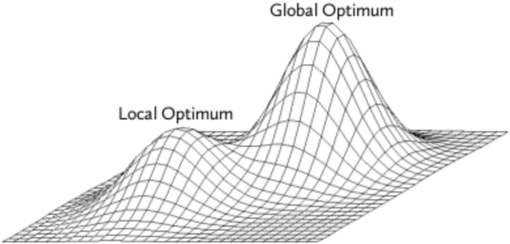
$$P(G_1, G_2, ..., G_M|R) = \prod_{j=1}^{L} P(G_j|R)$$

ML inference based on the model

- **Assuming loci are independent** we get the full likelihood:

  $P(G_1, G_2, ..., G_M|R) = \prod_{j=1}^{L} P(G_j|R)$

- This function is optimized for R and we get MLE of R
- Most often done using an EM algorithm
- NB does not always give you the MLE!



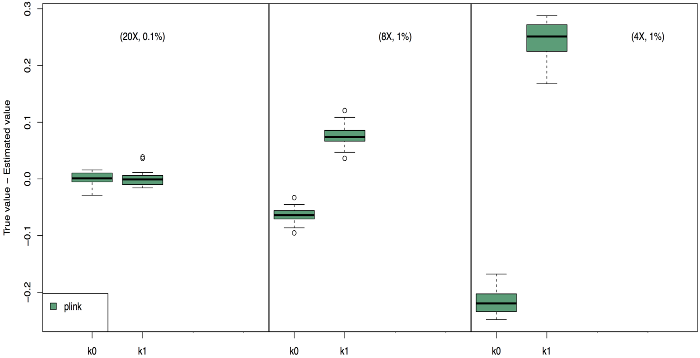
Outline

1. Introduction
   - Goal and motivation
   - Definition of key concepts
   - Overview of current methods for inference
2. A simple maximum likelihood solution for called genotypes
   - Overview and intuition
   - Model
   - ML inference
3. Problem and more advanced methods
   - Problems with standard methods
   - NGSrelate (for NGS data)
   - RelateAdmix (for admixed individuals)
4. Exercise

Problems

Most current methods

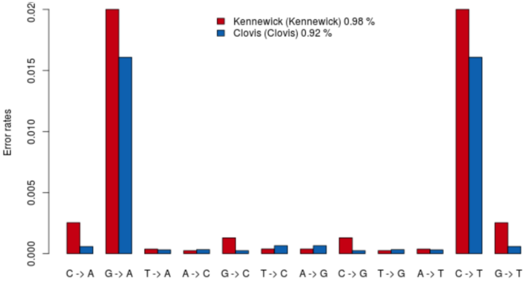
- work on called genotypes and **perform poorly on low depth NGS data**
- E.g. PLINK on simulated data from 20 1st cousins (k_0 is overestimated):



Problems (cont.)

For ancient DNA there is a range of additional issues:

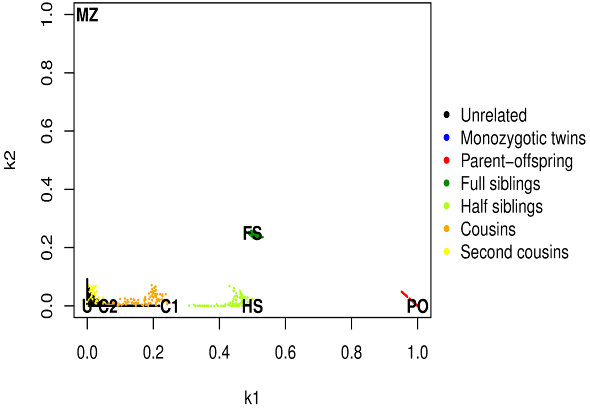
- very low coverage
- high error rates (especially transitions):



- not enough samples available for proper allele frequency estimation

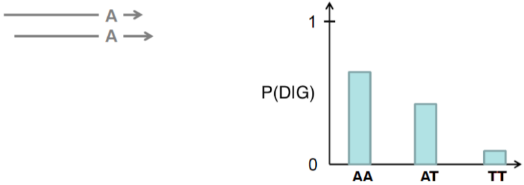
Problems (cont.)

Also most current methods have problems with admixed individuals. E.g. PLINK:



Handling NGS data (NGSrelate)

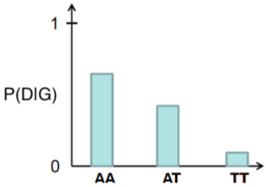
- A solution is to use genotype likelihoods (GLs) instead of called genotypes



- Quantifies the uncertainty in what the true genotype is

Handling NGS data (NGSrelate)

- A solution is to use genotype likelihoods (GLs) instead of called genotypes

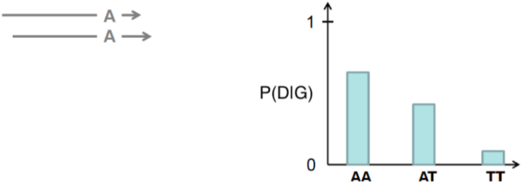


- Quantifies the uncertainty in what the true genotype is
- We can write a new likelihood for a locus j which takes this uncertainty into account using these:

$$P(D_j|R) = \sum_{G_j \in \{\text{all possible genotype pairs}\}} P(D_{j,i1}|G_{j,i1}) P(D_{j,i2}|G_{j,i2}) P(G_j|R)$$

Handling NGS data (NGSrelate)

- A solution is to use genotype likelihoods (GLs) instead of called genotypes



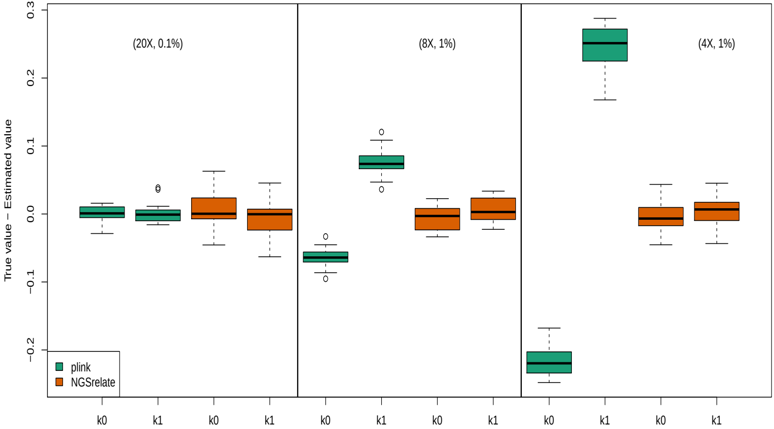
- Quantifies the uncertainty in what the true genotype is
- We can write a new likelihood for a locus j which takes this uncertainty into account using these:

$$P(D_j|R) = \sum_{G_j \in \{\text{all possible genotype pairs}\}} P(D_{j,i1}|G_{j,i1})P(D_{j,i2}|G_{j,i2})P(G_j|R)$$

- NB when genotypes are known the GLs are 0 for all but the true genotype, so the likelihood becomes the same as before!

Handling NGS data (NGSrelate)

- NGSrelate vs. PLINK

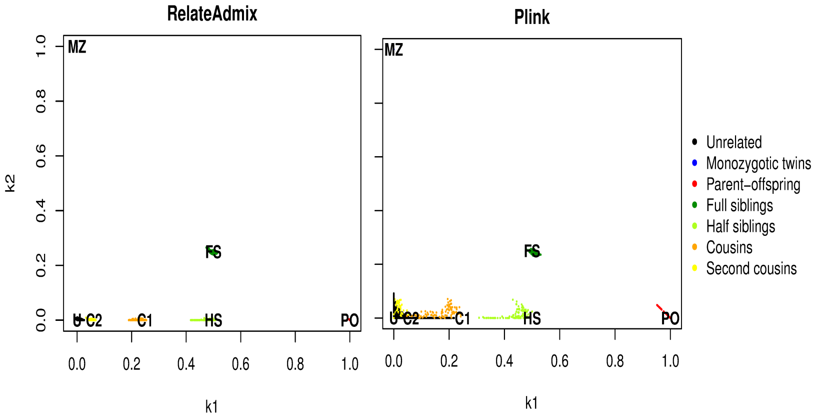


Methods for ancient DNA

- In some cases methods like NGSrelate can be used
  (if frequency info is available and one removes transitions)
- In other cases you need special methods
  (e.g. due to lack of allele frequencies)
- Still highly active research field, recently 3 new methods came out
  (Theunert et al. Genetics, Kuhn et al. BioRXiv, Waples et al. BioRXiv)

Handling admixture

- the likelihood can also be adjusted to take admixture into account
- this is done in the software relateAdmix
- comparison of relateAdmix and PLINK on admixed individuals:



Outline

1. Introduction
   - Goal and motivation
   - Definition of key concepts
   - Overview of current methods for inference
2. A simple maximum likelihood solution for called genotypes
   - Overview and intuition
   - Model
   - ML inference
3. Problem and more advanced methods
   - Problems with standard methods
   - NGSrelate (for NGS data)
   - RelateAdmix (for admixed individuals)
4. Exercise

Exercise with PLINK and NGSrelate

Go to http://popgen.dk/ida/Naples2018/web/ and solve exercises 1 & 2 (run the exercises on the server logged in with ssh -Y)