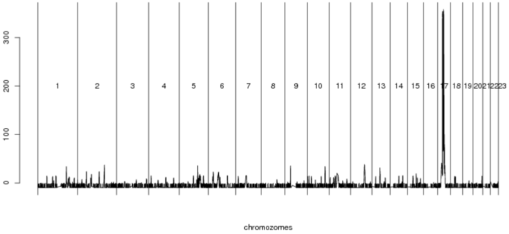
Local IBD inference

Ida Moltke, Naples, May 2017



Outline

1. Introduction
   - Goal
   - Motivation

2. Inferring local IBD sharing between pairs of individuals
   - Current solutions
   - An HMM based solution
   - Example of use: disease mapping

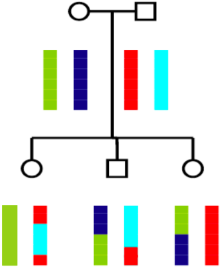
3. Exercises

Outline

1. Introduction
   - Goal
   - Motivation
2. Inferring local IBD sharing between pairs of individuals
   - Current solutions
   - An HMM based solution
   - Example of use: disease mapping
3. Exercises

Goal

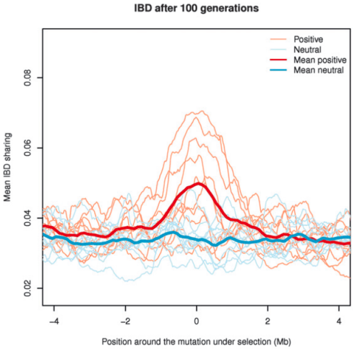
- We want to infer tracts of IBD sharing along the genome
- For example:



- For unphased data: to infer if individuals share 0, 1 or 2 alleles IBD locally

Motivation

- Can be used for population genetic analyses, e.g. detection of selection:



- Can be used for disease mapping (will return to this later if time allows)

Outline

1. Introduction
   - Goal
   - Motivation
2. Inferring local IBD sharing between pairs of individuals
   - Current solutions
   - An HMM based solution
   - Example of use: disease mapping
3. Exercises

Current solutions

There are several methods for doing this:

- Non-probabilistic, including but not limited to:
  - GERMLINE (Gusev et al. 2009)
    Finds stretches of identity above a certain (user specified) length
- Probabilistic, including but not limited to:
  - BEAGLE (Browning et al. 2010-2016)
  - Relate (Albrechtsen et al. 2009)

We will look at a fairly simple probabilistic solution
(a simple version of Relate)

Intuition behind

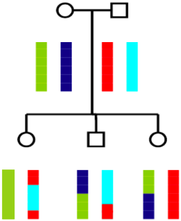
Info source 1: genotypes and allele frequencies (as before!)

- **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 occur rarely by chance
- **So the rarer the allele, the more identity makes us believe the individuals are IBD**

Intuition behind (cont')

Info source 2: The length of the stretches of identity (new!)

- IBD is broken up recombination and thus comes in consecutive tracts, hence so will IBD 0, 1 and 2

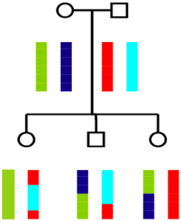


- Identity that occurs by chance does not have this property.
- In fact, often very unlikely to see long tracts of identity by chance

Intuition behind (cont')

Info source 2: The length of the stretches of identity (new!)

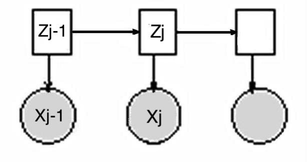
- IBD is broken up recombination and thus comes in consecutive tracts, hence so will IBD 0, 1 and 2



- Identity that occurs by chance does not have this property.
- In fact, often very unlikely to see long tracts of identity by chance
- The more distantly related, the shorter IBD tracts are expected to be

Using an Hidden Markov Model (HMM)

- We use these observations for inference based on a HMM
- Similar to the model we just looked at! E.g. has 2 variables per locus j:
  1. an observed variable, G_j (here genotype data, e.g. (AA,aa)).
  2. a hidden variable, Z_j, that indicates #alleles shared IBD in locus j.
- What is different: we no longer assume the loci are independent
- Instead Z_j depends on Z_{j-1}



- Mathematical formulation for L loci:

$$P(G,Z) = P(Z)P(G|Z) = P(Z_1)\Big(\prod_{j=2}^{L} P(Z_j|Z_{j-1})\Big)\Big(\prod_{j=1}^{L} P(G_j|Z_j)\Big)$$

with $G = (G_1, G_2, ..., G_L)$ and $Z = (Z_1, Z_2, ..., Z_L)$.

Emission probabilities ($P(G_j|Z_j)$)

- $P(G_j|Z_j)$ is exactly the same as before:

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

- Captures connection between IBD, genotypes and allele frequencies:
  - that alleles have to be identical to be IBD
  - that the rarer an allele, the more will observing IBS make us believe the individuals are IBD

Transition probabilities ($P(Z_j|Z_{j-1})$)

The probability of a change in IBD state from locus j-1 to j:

- We let these depend on the distance between markers, d, and R
- E.g. we set

  $$P(Z_j = 0|Z_{j-1} = 1) = (1 - e^{-ad})k_0$$

  where a is a non-negative parameter of the model

Transition probabilities ($P(Z_j|Z_{j-1})$)

The probability of a change in IBD state from locus j-1 to j:

- We let these depend on the distance between markers, d, and R
- E.g. we set
$$P(Z_j = 0|Z_{j-1} = 1) = (1 - e^{-ad})k_0$$
where a is a non-negative parameter of the model
- Dependence on the distance d:
    - The larger d is, the larger is the probability of a change
    - Reflects that the probability of recombination increases with d
    - Captures that IBD occurs in consecutive tracts that are broken up by recombination

Transition probabilities ($P(Z_j|Z_{j-1})$)

The probability of a change in IBD state from locus j-1 to j:

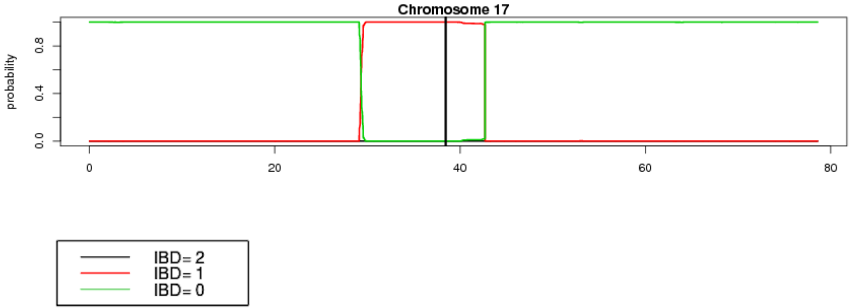
- We let these depend on the distance between markers, d, and R
- E.g. we set

  $$P(Z_j = 0|Z_{j-1} = 1) = (1 - e^{-ad})k_0$$

  where a is a non-negative parameter of the model
- Dependence on the distance d:
  - The larger d is, the larger is the probability of a change
  - Reflects that the probability of recombination increases with d
  - Captures that IBD occurs in consecutive tracts that are broken up by recombination
- Dependence on R (in this case k_0)
  - The higher k_0 is (so the more distantly related), the higher is the probability of a change to no IBD sharing ($Z_j = 0$)
  - Captures that more distantly related tend to have shorter IBD tracts

The inference

- We can now use standard inference methods for HMMs
- We can e.g. (using the forward-backward algorithm) get the posterior probability of the 3 possible IBD states $P(Z_j|G)$:

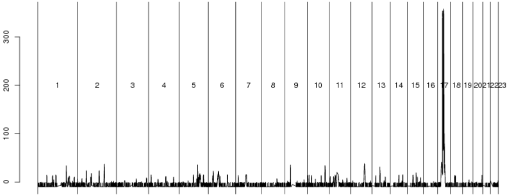


Example of use: disease mapping

- Individuals with a disease caused by the same mutation are IBD in a region harbouring the mutation
- Can identify such mutations by finding regions where cases tend to be IBD

Example of use: disease mapping

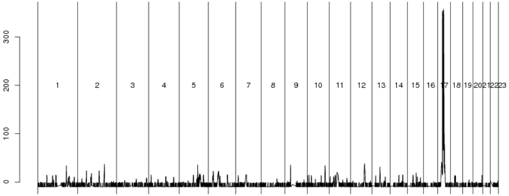
- Individuals with a disease caused by the same mutation are IBD in a region harbouring the mutation
- Can identify such mutations by finding regions where cases tend to be IBD
- E.g. in Albrechtsen et al. 2009:
  - 7 seemingly unrelated Danish individuals with breast cancer
  - All genotyped for 225,000 SNPs across the genome
  - Used an HMM to infer IBD 0,1,2 along the genome
  - Looked for excess IBD sharing in cases (vs. controls)



  - Only 1 region, where all cases carried same cancer causing mutation

Example of use: disease mapping

- Individuals with a disease caused by the same mutation are IBD in a region harbouring the mutation
- Can identify such mutations by finding regions where cases tend to be IBD
- E.g. in Albrechtsen et al. 2009:
    - 7 seemingly unrelated Danish individuals with breast cancer
    - All genotyped for 225,000 SNPs across the genome
    - Used an HMM to infer IBD 0,1,2 along the genome
    - Looked for excess IBD sharing in cases (vs. controls)



    - Only 1 region, where all cases carried same cancer causing mutation
- So mapping can be done w. very few individuals (and SNPs) (<<GWAS)

Outline

1. Introduction
   - Goal
   - Motivation

2. Inferring local IBD sharing between pairs of individuals
   - Current solutions
   - An HMM based solution
   - Example of use: disease mapping

3. Exercises

Exercises

Go to http://popgen.dk/ida/EMBONaples2017/web/ and solve exercise 3 & 4

(run the exercises on the server logged in with ssh -Y)