Relatedness inference (Exercise 1)

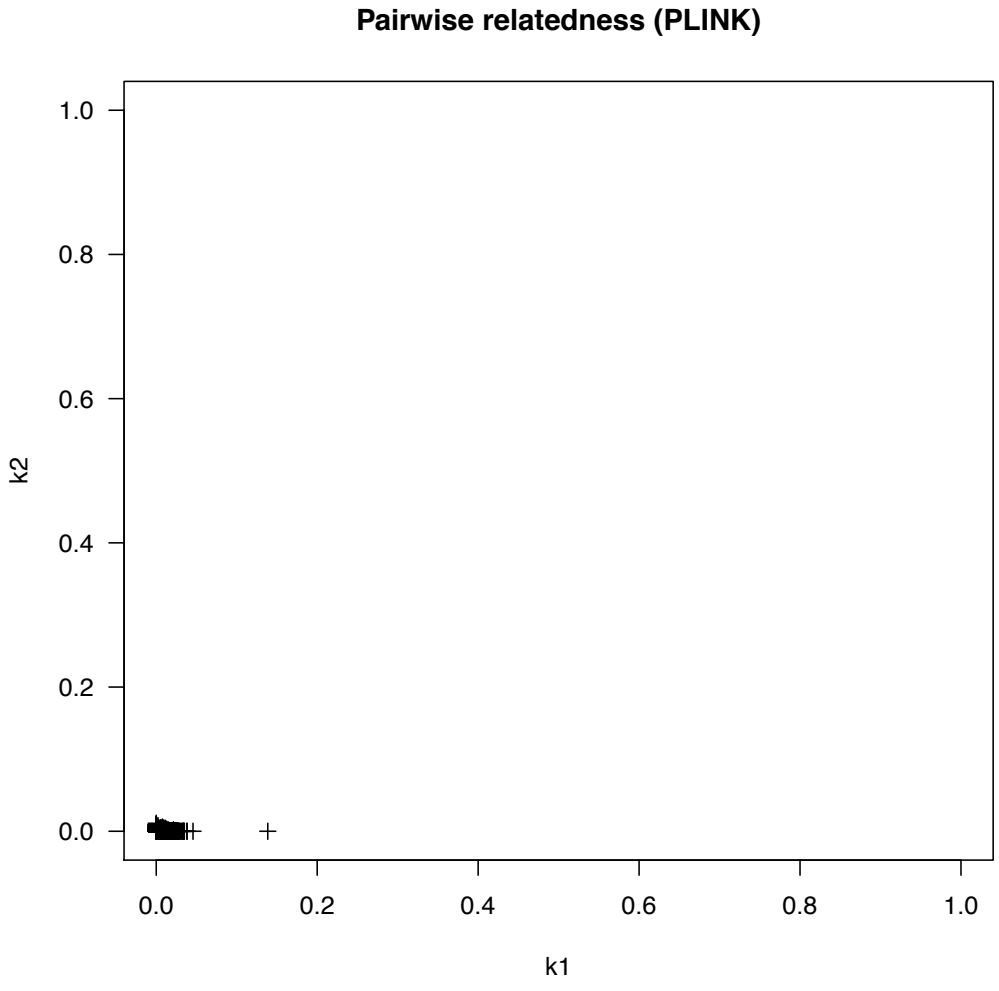
Overview of first dataset

First dataset consists of genotype data in plink format for

- 52 individuals
- 109983 loci

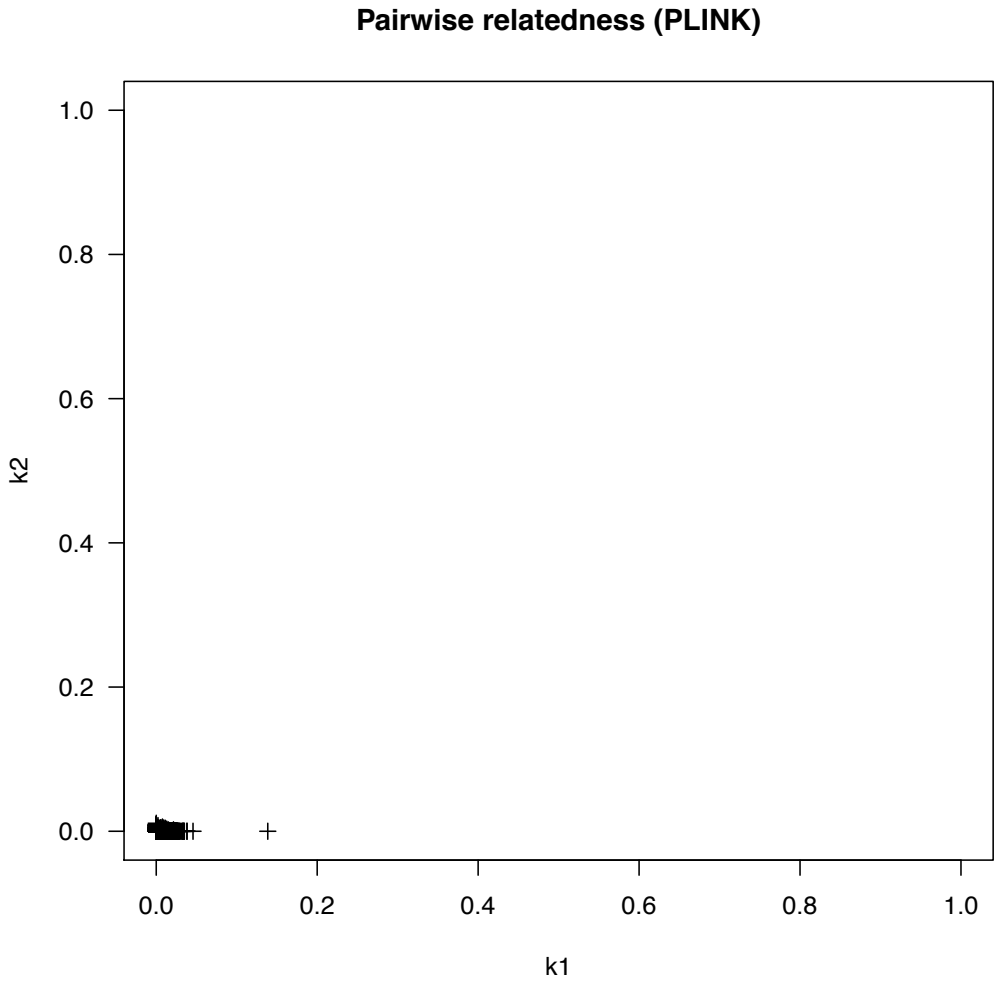
The genotypes called from a simulated 20x NGS dataset.

Applying PLINK to called genotypes (20x)



- What does this suggest (are the individuals related)?

Applying PLINK to called genotypes (20x)



- What does this suggest (are the individuals related)?
- That all pairs are unrelated/very distantly related except for one.

Applying PLINK to called genotypes (20x)

- Which pair is that? And what are the exact estimates for that pair?

Applying PLINK to called genotypes (20x)

- Which pair is that? And what are the exact estimates for that pair?

```
> plinkout20x[1:5,c("IID1","IID2","Z0","Z1","Z2")]
  IID1 IID2     Z0     Z1     Z2
1 ind0 ind1 0.8614 0.1386 0.0000
2 ind0 ind2 0.9947 0.0029 0.0023
3 ind0 ind3 1.0000 0.0000 0.0000
4 ind0 ind4 1.0000 0.0000 0.0000
5 ind0 ind5 1.0000 0.0000 0.0000
```

Applying PLINK to called genotypes (20x)

- Which pair is that? And what are the exact estimates for that pair?

```
> plinkout20x[1:5,c("IID1","IID2","Z0","Z1","Z2")]
  IID1 IID2     Z0     Z1     Z2
1 ind0 ind1 0.8614 0.1386 0.0000
2 ind0 ind2 0.9947 0.0029 0.0023
3 ind0 ind3 1.0000 0.0000 0.0000
4 ind0 ind4 1.0000 0.0000 0.0000
5 ind0 ind5 1.0000 0.0000 0.0000
```

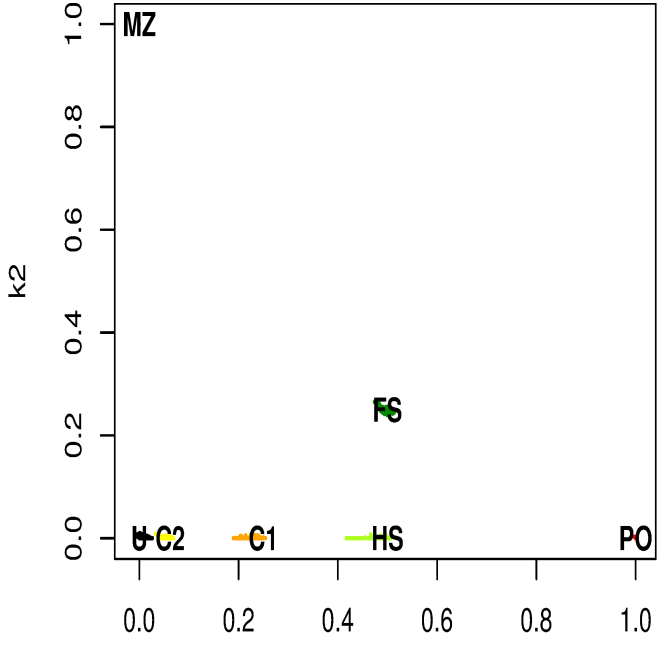
- Suggests that the first pair (ind0 and ind1) is related, but how?

Applying PLINK to called genotypes (20x)

- Which pair is that? And what are the exact estimates for that pair?

```
> plinkout20x[1:5,c("IID1","IID2","Z0","Z1","Z2")]
  IID1 IID2     Z0     Z1     Z2
1 ind0 ind1 0.8614 0.1386 0.0000
2 ind0 ind2 0.9947 0.0029 0.0023
3 ind0 ind3 1.0000 0.0000 0.0000
4 ind0 ind4 1.0000 0.0000 0.0000
5 ind0 ind5 1.0000 0.0000 0.0000
```

- Suggests that the first pair (ind0 and ind1) is related, but how?

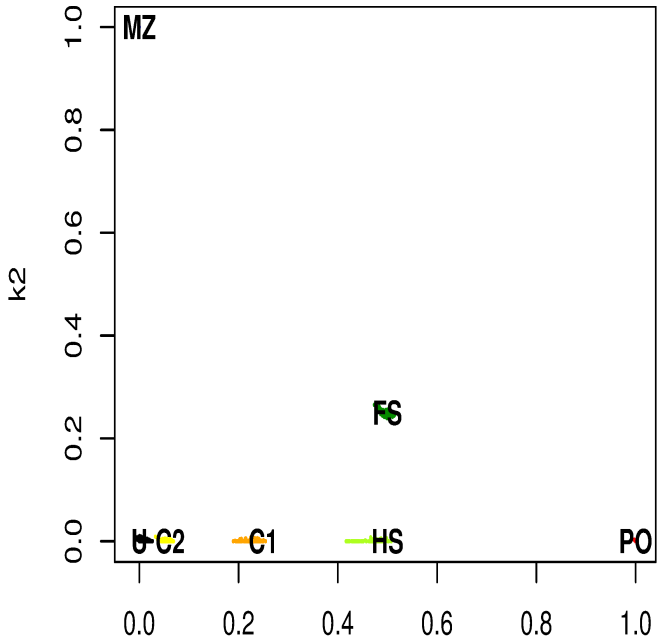


Applying PLINK to called genotypes (20x)

- Which pair is that? And what are the exact estimates for that pair?

```
> plinkout20x[1:5,c("IID1","IID2","Z0","Z1","Z2")]
  IID1 IID2     Z0     Z1     Z2
1 ind0 ind1 0.8614 0.1386 0.0000
2 ind0 ind2 0.9947 0.0029 0.0023
3 ind0 ind3 1.0000 0.0000 0.0000
4 ind0 ind4 1.0000 0.0000 0.0000
5 ind0 ind5 1.0000 0.0000 0.0000
```

- Suggests that the first pair (ind0 and ind1) is related, but how?



- So maybe first cousins (which has expected R of (0.75,0.25,0))

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?
- Fairly well!

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?
- Fairly well!
- Why do you think the true R is not exactly equal to the expected?

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?
- Fairly well!
- Why do you think the true R is not exactly equal to the expected?
- Because the true R will vary due to randomness in recombination

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?
- Fairly well!
- Why do you think the true R is not exactly equal to the expected?
- Because the true R will vary due to randomness in recombination
- Where on a k1 vs k2 plot would you expect identical twins to be? And would you expect the true values for these would vary between different identical twins (why/why not)?

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?
- Fairly well!
- Why do you think the true R is not exactly equal to the expected?
- Because the true R will vary due to randomness in recombination
- Where on a k1 vs k2 plot would you expect identical twins to be? And would you expect the true values for these would vary between different identical twins (why/why not)?
- In the upper left corner and no variation is expected

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?
- Fairly well!
- Why do you think the true R is not exactly equal to the expected?
- Because the true R will vary due to randomness in recombination
- Where on a k1 vs k2 plot would you expect identical twins to be? And would you expect the true values for these would vary between different identical twins (why/why not)?
- In the upper left corner and no variation is expected
- Where on the plot would you expect duplicate samples to be?

Applying PLINK to called genotypes (20x)

- How do the estimates compare to the true R (0.873, 0.127, 0)?
- Fairly well!
- Why do you think the true R is not exactly equal to the expected?
- Because the true R will vary due to randomness in recombination
- Where on a k1 vs k2 plot would you expect identical twins to be? And would you expect the true values for these would vary between different identical twins (why/why not)?
- In the upper left corner and no variation is expected
- Where on the plot would you expect duplicate samples to be?
- Same as monozygotic twins (useful for QC of datasets)

Relatedness inference (Exercise 2)

Overview of second dataset

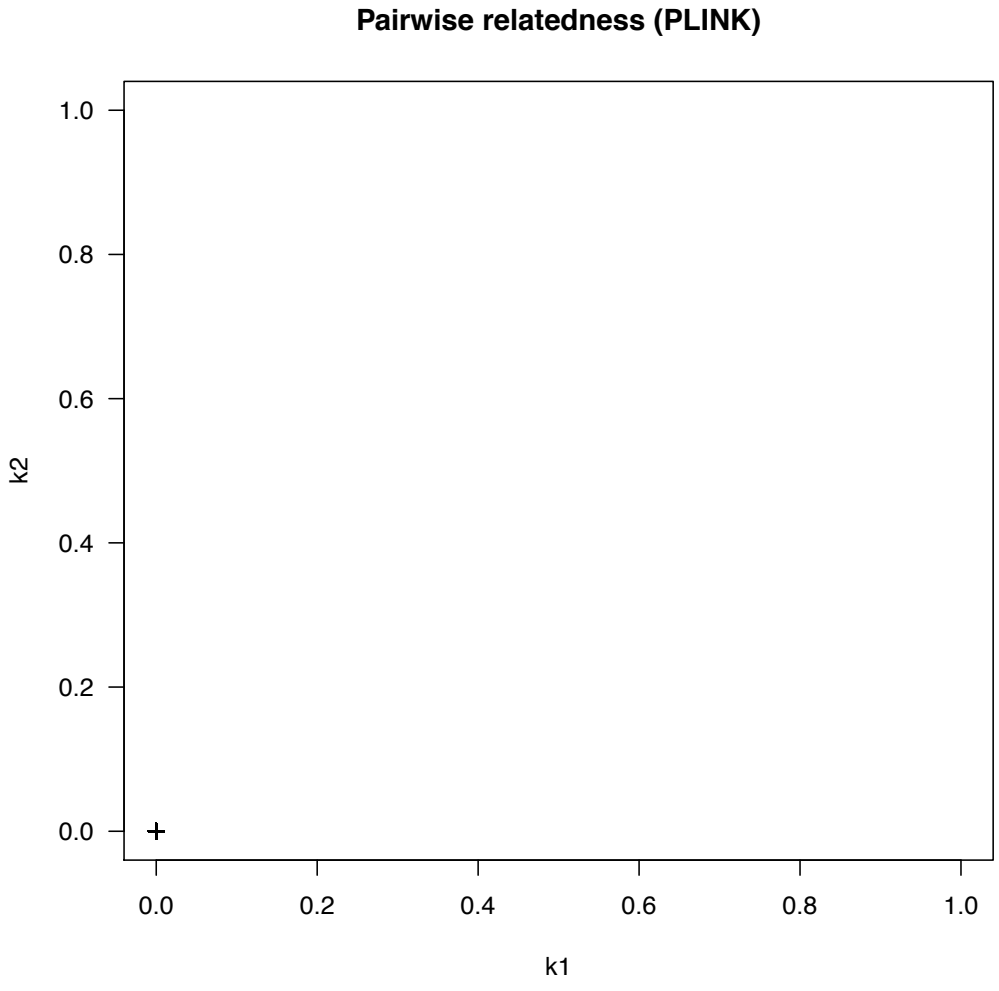
Second dataset consists of genotype data in PLINK format for

- 52 individuals
- 70651 loci

The genotypes were called from a simulated **4x NGS data**.
So fairly **low depth** data.

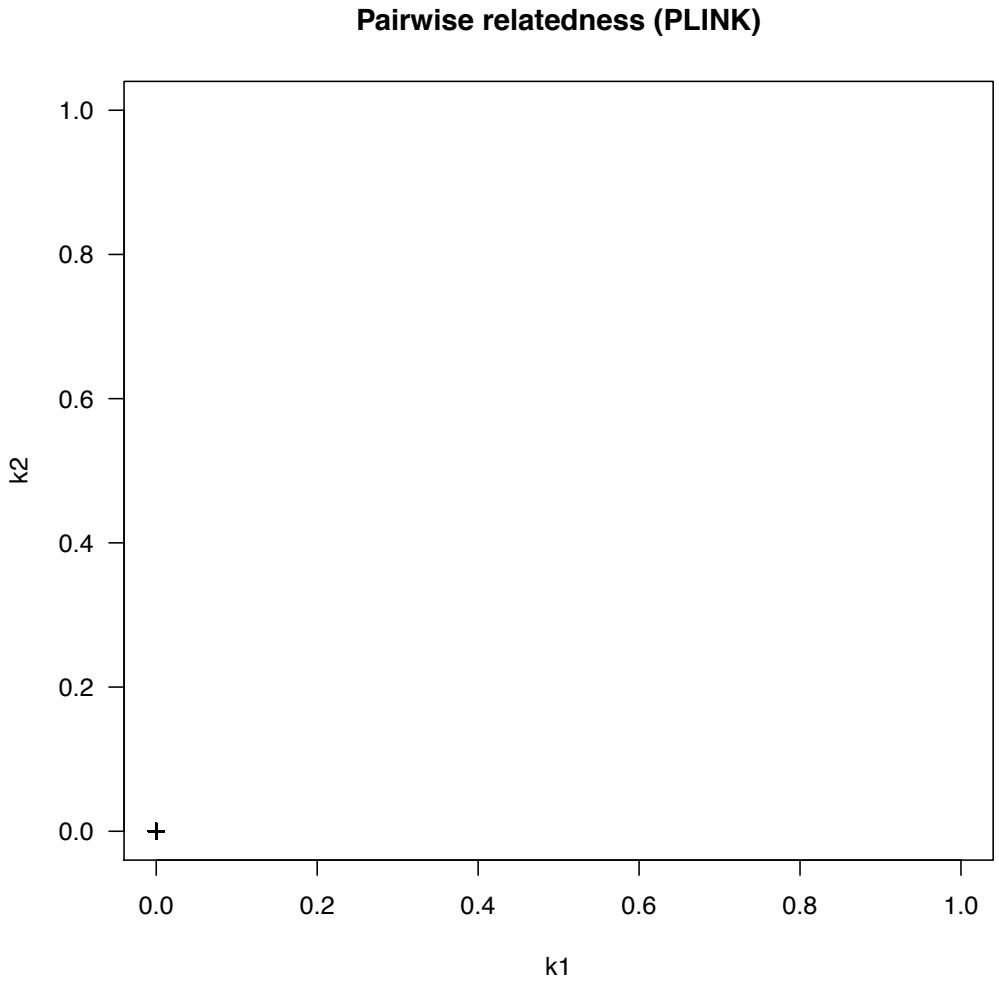
From the same simulated data **we also have genotype likelihoods (GLs)** and allele frequencies in the format needed for NGSrelate.

Applying PLINK to the called genotypes (4x)



- What does this suggest?

Applying PLINK to the called genotypes (4x)



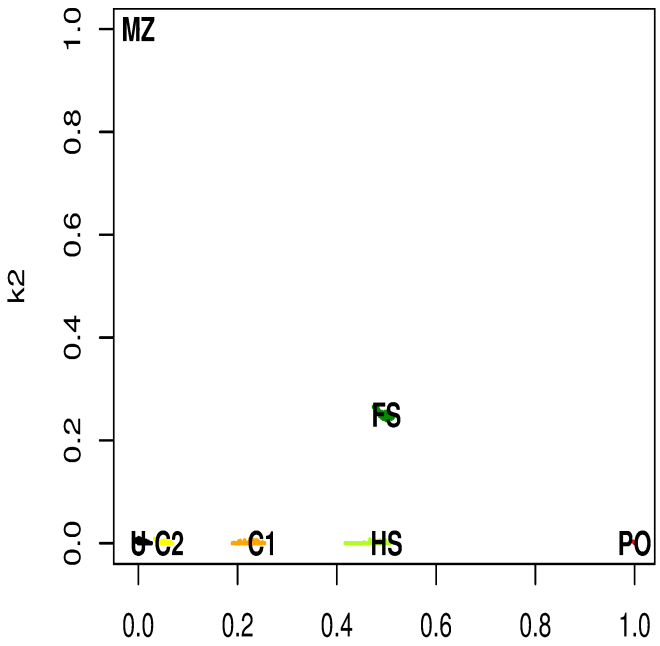
- What does this suggest?
- Suggests all pairs are unrelated

Appplying NGSrelate to GLs (4x)

- Estimated $R = (k_0, k_1, k_2)$ for **the first pair** (0 and 1):

a	b	nSites	k0	k1	k2	loglh	nIter	coverage
0	1	66205	0.854844	0.145154	0.000002	-116979.230583	2456	0.937071

- What does that suggest about this pair?

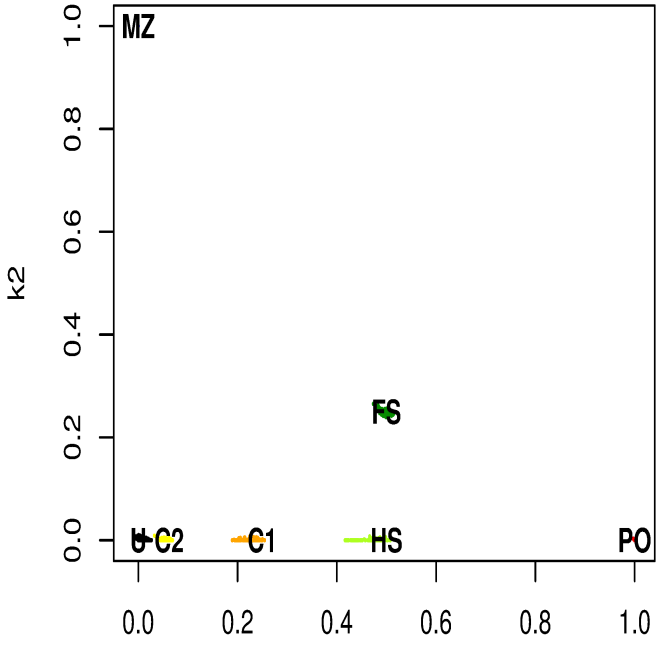


Appplying NGSrelate to GLs (4x)

- Estimated $R = (k_0, k_1, k_2)$ for **the first pair** (0 and 1):

```
a b  nSites k0       k1       k2       loglh          nIter  coverage
0 1  66205 0.854844 0.145154 0.000002 -116979.230583 2456   0.937071
```

- What does that suggest about this pair?



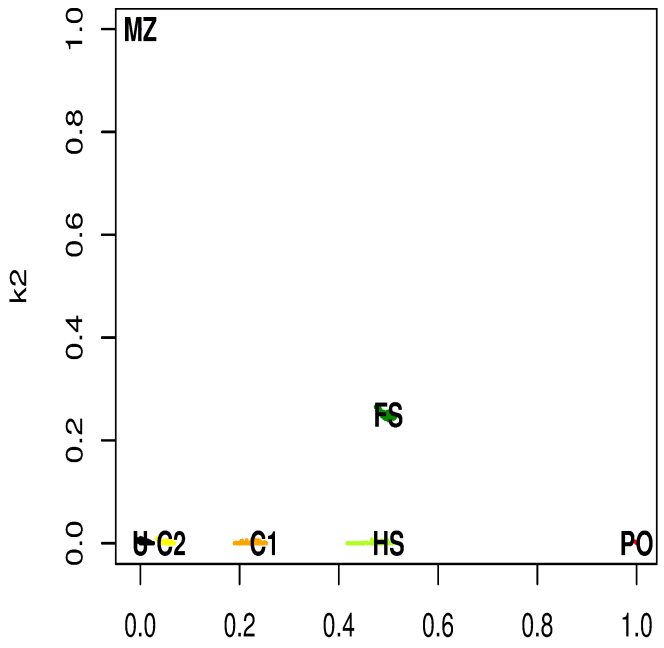
- Suggests that this pair is related (probably first cousins)

Applying NGSrelate to GLs (4x)

- Estimated $R = (k_0, k_1, k_2)$ for **the second pair** (0 and 2):

a	b	nSites	k0	k1	k2	loglh	nIter	coverage
0	2	66194	1.000000	0.000000	0.000000	-117096.981729	-1	0.936915

- What does that suggest about this pair?

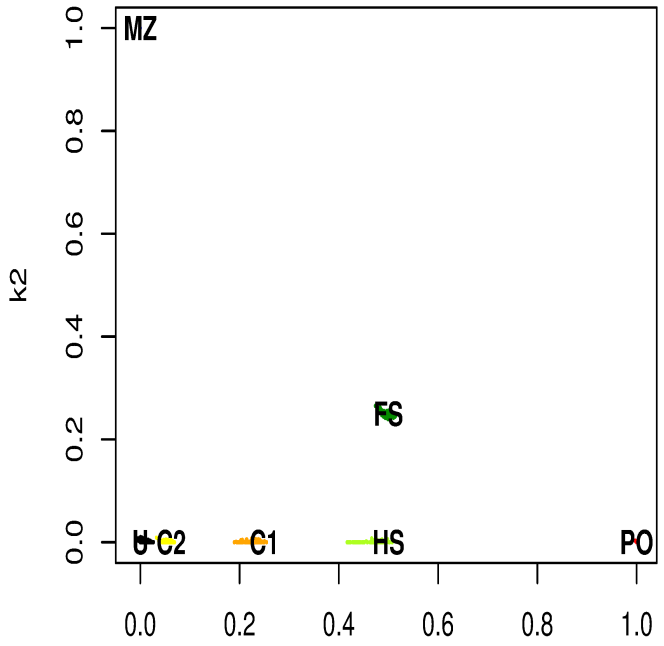


Applying NGSrelate to GLs (4x)

- Estimated $R = (k_0, k_1, k_2)$ for **the second pair** (0 and 2):

```
a b  nSites k0       k1       k2       loglh          nIter  coverage
0 2  66194 1.000000 0.000000 0.000000 -117096.981729    -1  0.936915
```

- What does that suggest about this pair?



- Suggests that this pair is unrelated

Comparing results from PLINK and NGSrelate

- The truth for these two pairs are $R = (0.832, 0.168, 0)$ and $R = (1, 0, 0)$, respectively. What does this suggest about PLINK and NGSrelate?

Comparing results from PLINK and NGSrelate

- The truth for these two pairs are $R = (0.832, 0.168, 0)$ and $R = (1, 0, 0)$, respectively. What does this suggest about PLINK and NGSrelate?
- For low depth data NGSrelate seems to give markedly better results at least for related pairs

Running NGSrelate properly (check for convergence)

- NGSrelate relies on numerical optimisation and is thus not guaranteed to always provide maximum likelihood estimates (like ADMIXTURE and other ML based programs)
- Does it look like the estimates for individuals 0 an 1 you got previously are maximum likelihood estimates?

Running NGSrelate properly (check for convergence)

- NGSrelate relies on numerical optimisation and is thus not guaranteed to always provide maximum likelihood estimates (like ADMIXTURE and other ML based programs)
- Does it look like the estimates for individuals 0 an 1 you got previously are maximum likelihood estimates?
- I ran it ten times with different seeds and got :

```
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ ./ngsRelate -f simdata_4x.freq -g simdata_4x.glf -n 52 -a 0 -b 1 -i 5000 -r 2
        -> Frequency file: 'simdata_4x.freq' contain 70651 number of sites
Pair    k0      k1      k2      loglh   nIter   coverage
(0,1):66205     0.854844        0.145154        0.000002        -116979.230583  3570    0.937071
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ ./ngsRelate -f simdata_4x.freq -g simdata_4x.glf -n 52 -a 0 -b 1 -i 5000 -r 3
        -> Frequency file: 'simdata_4x.freq' contain 70651 number of sites
Pair    k0      k1      k2      loglh   nIter   coverage
(0,1):66205     0.854844        0.145154        0.000002        -116979.230583  2951    0.937071
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ ./ngsRelate -f simdata_4x.freq -g simdata_4x.glf -n 52 -a 0 -b 1 -i 5000 -r 4
        -> Frequency file: 'simdata_4x.freq' contain 70651 number of sites
Pair    k0      k1      k2      loglh   nIter   coverage
(0,1):66205     0.854844        0.145154        0.000002        -116979.230584  3113    0.937071
```

Running NGSrelate properly (check for convergence)

- NGSrelate relies on numerical optimisation and is thus not guaranteed to always provide maximum likelihood estimates (like ADMIXTURE and other ML based programs)
- Does it look like the estimates for individuals 0 an 1 you got previously are maximum likelihood estimates?
- I ran it ten times with different seeds and got :

```
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ ./ngsRelate -f simdata_4x.freq -g simdata_4x.glf -n 52 -a 0 -b 1 -i 5000 -r 2
        -> Frequency file: 'simdata_4x.freq' contain 70651 number of sites
Pair    k0      k1      k2      loglh   nIter   coverage
(0,1):66205     0.854844        0.145154        0.000002        -116979.230583  3570    0.937071
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ ./ngsRelate -f simdata_4x.freq -g simdata_4x.glf -n 52 -a 0 -b 1 -i 5000 -r 3
        -> Frequency file: 'simdata_4x.freq' contain 70651 number of sites
Pair    k0      k1      k2      loglh   nIter   coverage
(0,1):66205     0.854844        0.145154        0.000002        -116979.230583  2951    0.937071
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ ./ngsRelate -f simdata_4x.freq -g simdata_4x.glf -n 52 -a 0 -b 1 -i 5000 -r 4
        -> Frequency file: 'simdata_4x.freq' contain 70651 number of sites
Pair    k0      k1      k2      loglh   nIter   coverage
(0,1):66205     0.854844        0.145154        0.000002        -116979.230584  3113    0.937071
```

- All likelihoods and estimates are very similar, so yes

Taking a closer look at the input format for NGSrelate

The top lines of the frequency file (simdata_4x.freq):

```
0.67633
0.15674
0.796849
0.557128
0.263847
0.900578
0.537322
0.292218
0.83204
```

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP      a1 a2 1                  2                  3
1_30387   0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501   0  1 0.66                0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP      a1 a2 1                  2                  3
1_30387   0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501   0  1 0.66               0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?
- Genotype a1a2 has highest GL and a1=0=A and a2=1=C, so AC

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP      a1 a2 1                  2                  3
1_30387   0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501   0  1 0.66                0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?
- Genotype a1a2 has highest GL and a1=0=A and a2=1=C, so AC
- Does that fit with the called genotype used by PLINK?

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP      a1 a2 1                  2                  3
1_30387   0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501   0  1 0.66               0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?
- Genotype a1a2 has highest GL and a1=0=A and a2=1=C, so AC
- Does that fit with the called genotype used by PLINK?

```
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ head -n 2 simdata_4x.tped | cut -f1-4
1 dummy1_1 0 1  C A     C C     C C
1 dummy1_2 0 2  A A     A A     A C
```

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP      a1 a2 1                  2                  3
1_30387   0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501   0  1 0.66               0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?
- Genotype a1a2 has highest GL and a1=0=A and a2=1=C, so AC
- Does that fit with the called genotype used by PLINK?

```
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ head -n 2 simdata_4x.tped | cut -f1-4
1 dummy1_1 0 1  C A     C C     C C
1 dummy1_2 0 2  A A     A A     A C
```

- So yes

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP      a1 a2 1                  2                  3
1_30387   0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501   0  1 0.66               0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?
- Genotype a1a2 has highest GL and a1=0=A and a2=1=C, so AC
- Does that fit with the called genotype used by PLINK?

```
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ head -n 2 simdata_4x.tped | cut -f1-4
1 dummy1_1 0 1  C A     C C     C C
1 dummy1_2 0 2  A A     A A     A C
```

- So yes
- And would you trust that called genotype (based on the GLs)?

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP     a1 a2 1                   2                   3
1_30387  0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501  0  1 0.66                0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?
- Genotype a1a2 has highest GL and a1=0=A and a2=1=C, so AC
- Does that fit with the called genotype used by PLINK?

```
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ head -n 2 simdata_4x.tped | cut -f1-4
1 dummy1_1 0 1  C A     C C     C C
1 dummy1_2 0 2  A A     A A     A C
```

- So yes
- And would you trust that called genotype (based on the GLs)?
- Yes, to a large degree because the likelihood for AC is fairly high

Taking a closer look at the input format for NGSrelate

- The first lines and columns of the (non-binary version of the) glf file:

```
SNP      a1 a2 1                  2                  3
1_30387   0  1 0.0726538176426983 0.926612305411416 0.000733876945885843
1_64501   0  1 0.66               0.333333333333333 0.00666666666666667
```

- What is the most likely genotype for the 1st SNP of 1st individual?
- Genotype a1a2 has highest GL and a1=0=A and a2=1=C, so AC
- Does that fit with the called genotype used by PLINK?

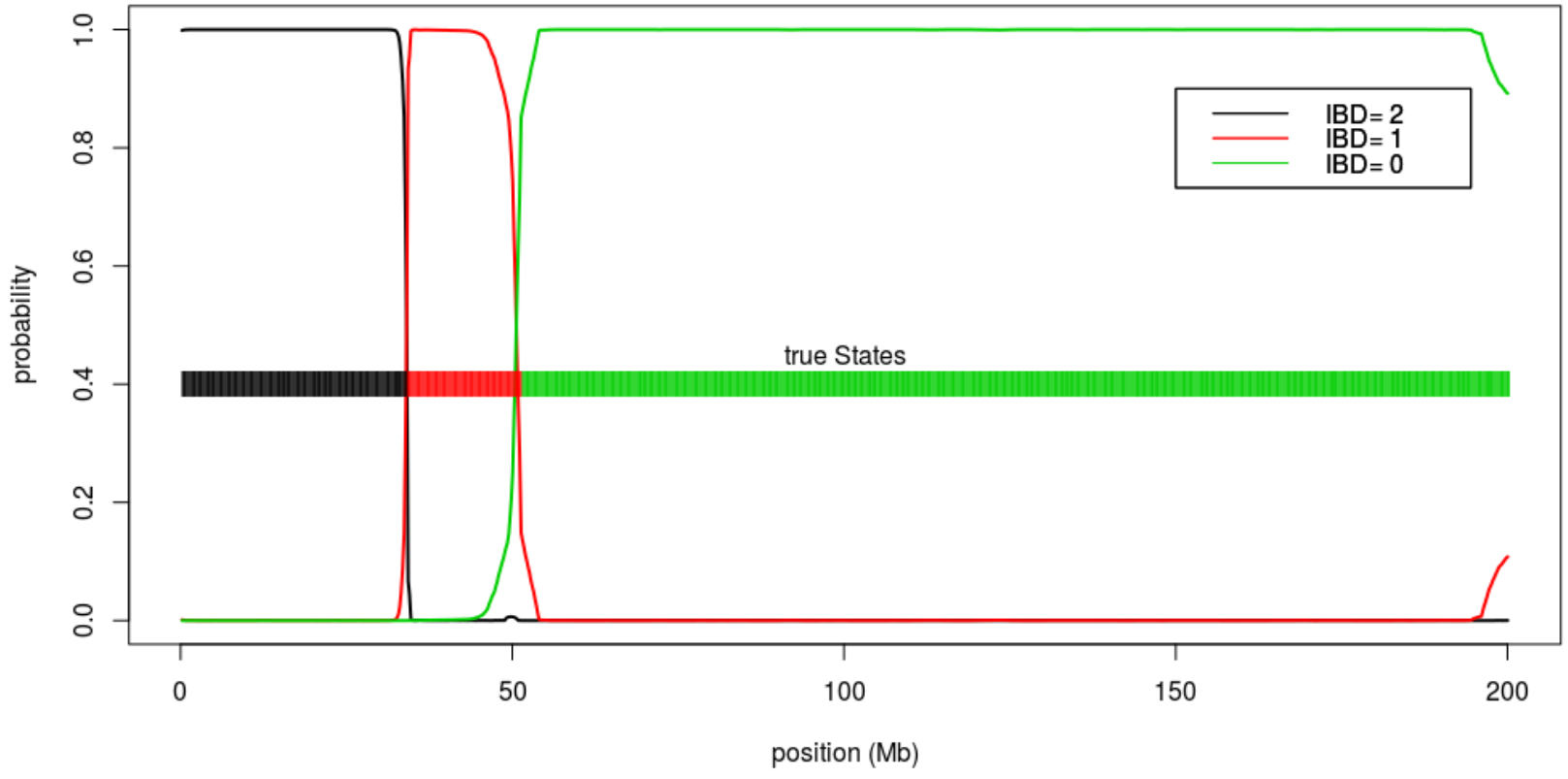
```
ida@pontus:~/Teaching/EMBO2017/RelatednessExerciseV3$ head -n 2 simdata_4x.tped | cut -f1-4
1 dummy1_1 0 1  C A     C C     C C
1 dummy1_2 0 2  A A     A A     A C
```

- So yes
- And would you trust that called genotype (based on the GLs)?
- Yes, to a large degree because the likelihood for AC is fairly high
- Same questions for second SNP

IBD tract inference
(Exercise 3)

Try out Relate in R

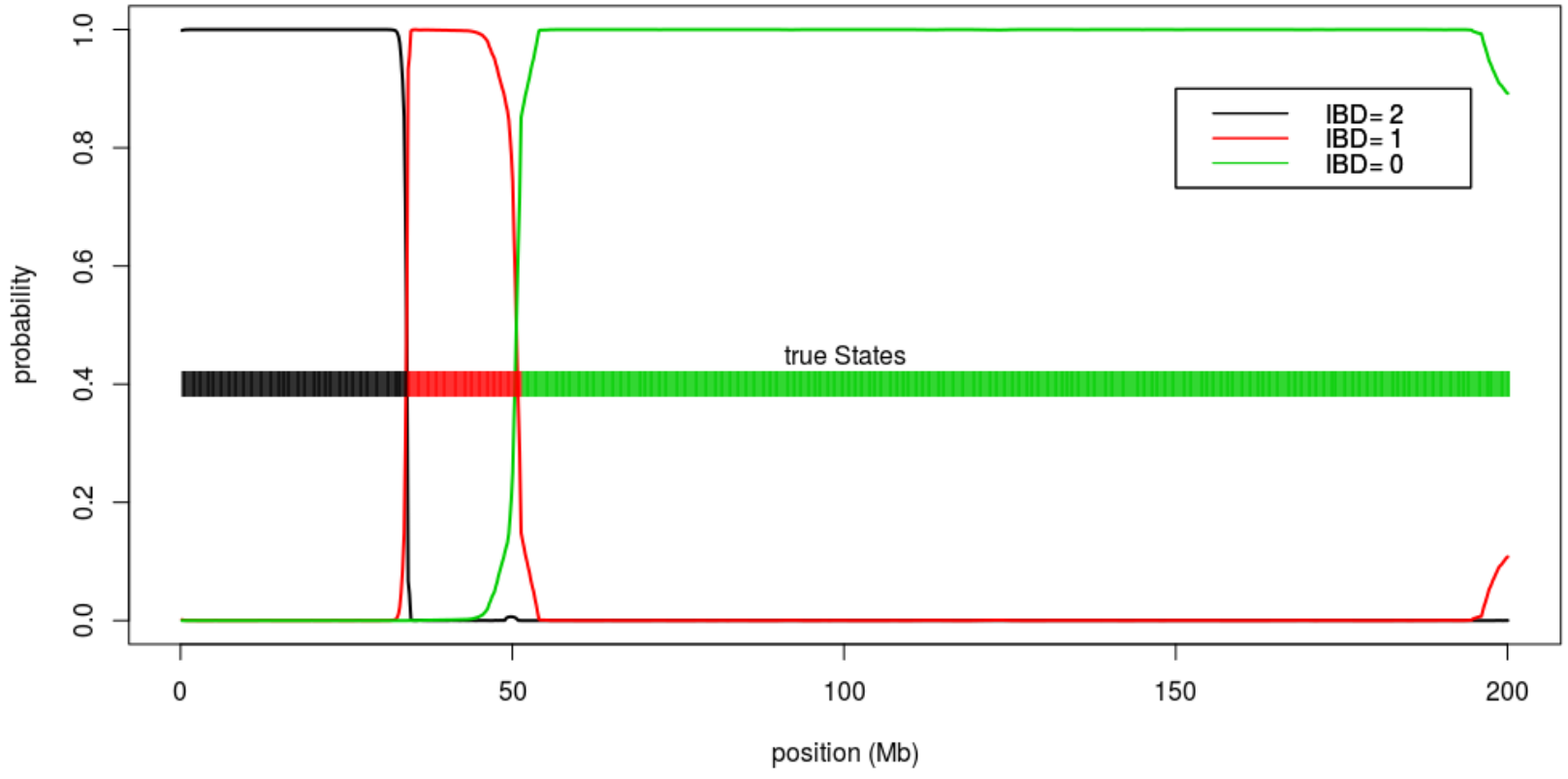
- The example produced the following plot



- Which region do the individuals truly share 2 alleles IBD (IBD=2)?

Try out Relate in R

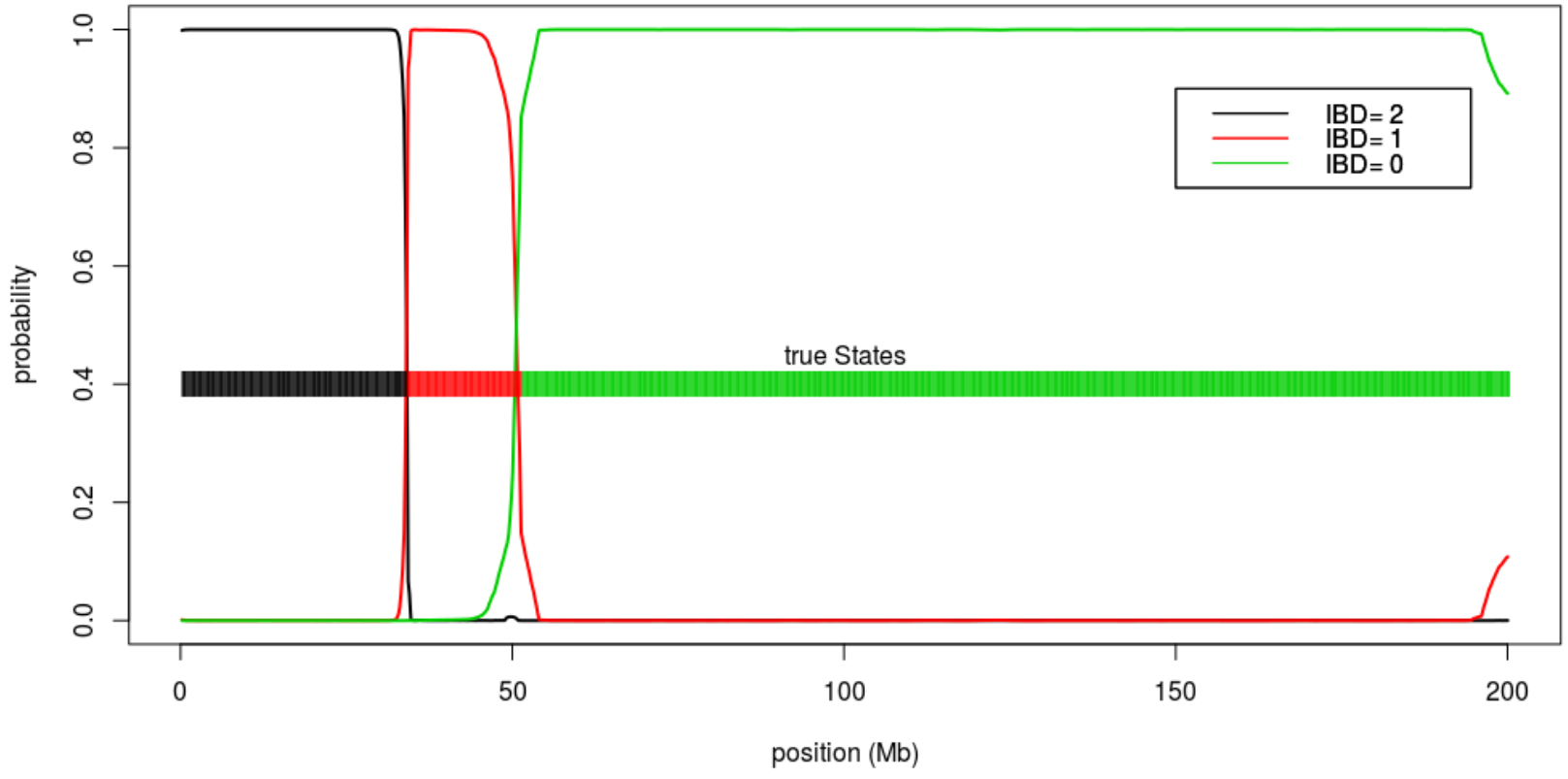
- The example produced the following plot



- Which region do the individuals truly share 2 alleles IBD (IBD=2)?

Try out Relate in R

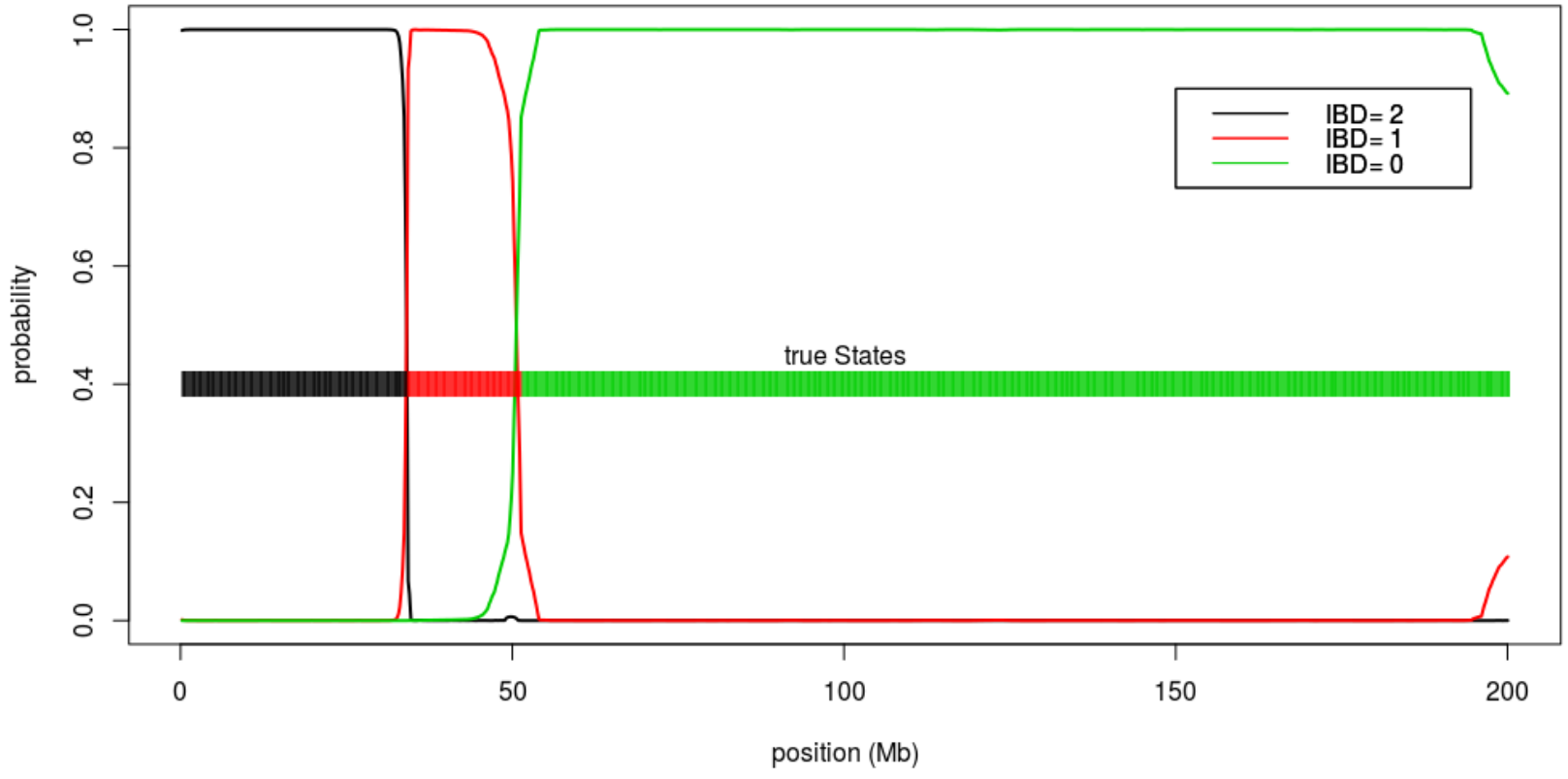
- The example produced the following plot



- Which region do the individuals truly share 2 alleles IBD (IBD=2)?
- Which region does Relate estimate >0.95 probability of IBD=2?

Try out Relate in R

- The example produced the following plot

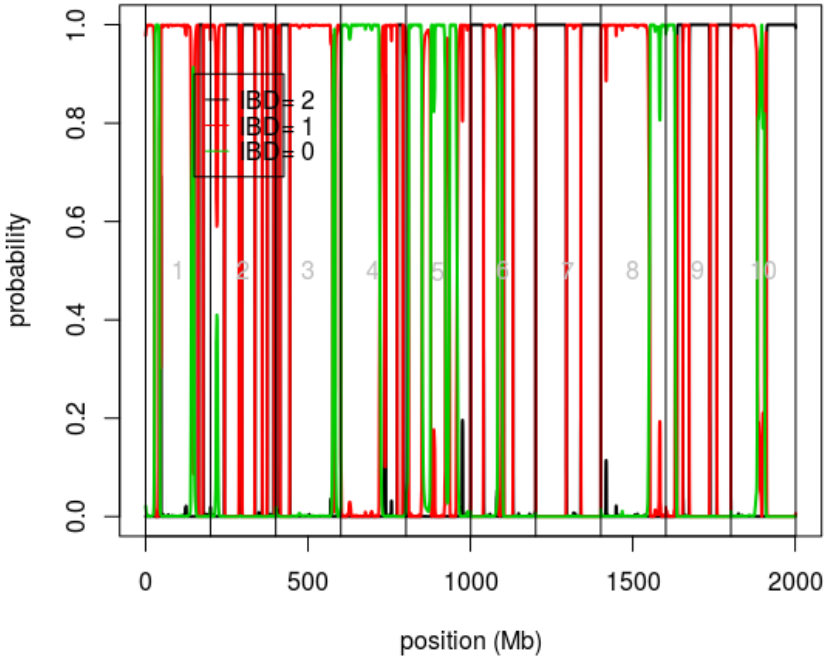


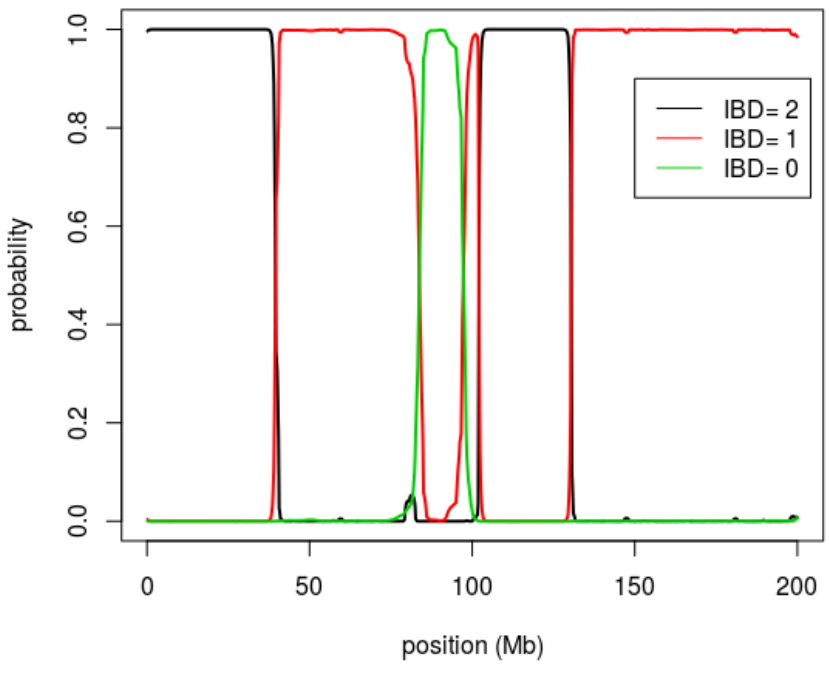
- Which region do the individuals truly share 2 alleles IBD (IBD=2)?
- Which region does Relate estimate >0.95 probability of IBD=2?
- How well does Relate infer the true IBD2 region?

Try on simulated data from siblings

- Running Relate on simulated sibling data we got:

```
k0,k1,k2 = 0.21, 0.51, 0.28
  k.like = 14189.74
     k.r = 0.327
       a = 0.0502
  u.like = 17445.5
 po.like = 16348
```

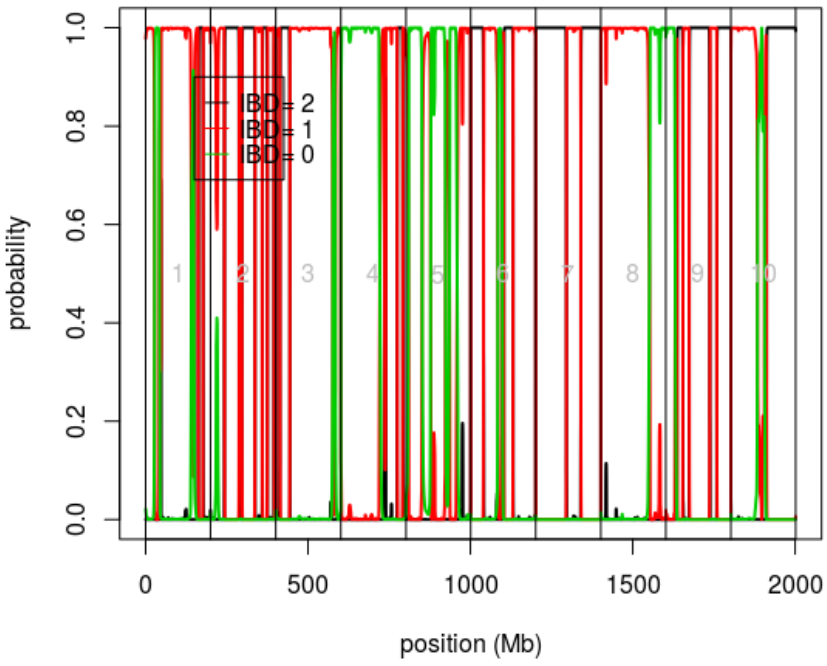


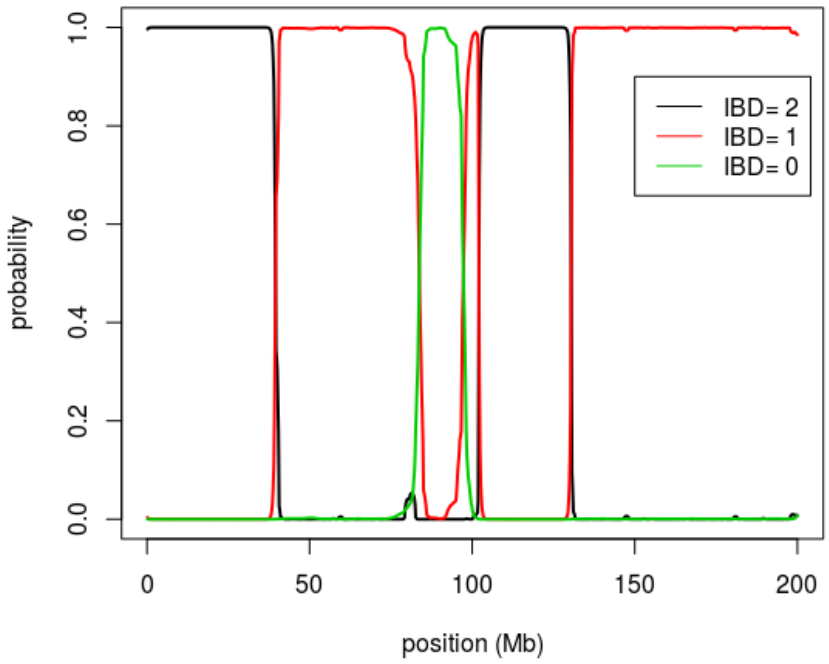


Try on simulated data from siblings

- Running Relate on simulated sibling data we got:

```
k0,k1,k2 = 0.21, 0.51, 0.28
  k.like = 14189.74
     k.r = 0.327
       a = 0.0502
  u.like = 17445.5
 po.like = 16348
```

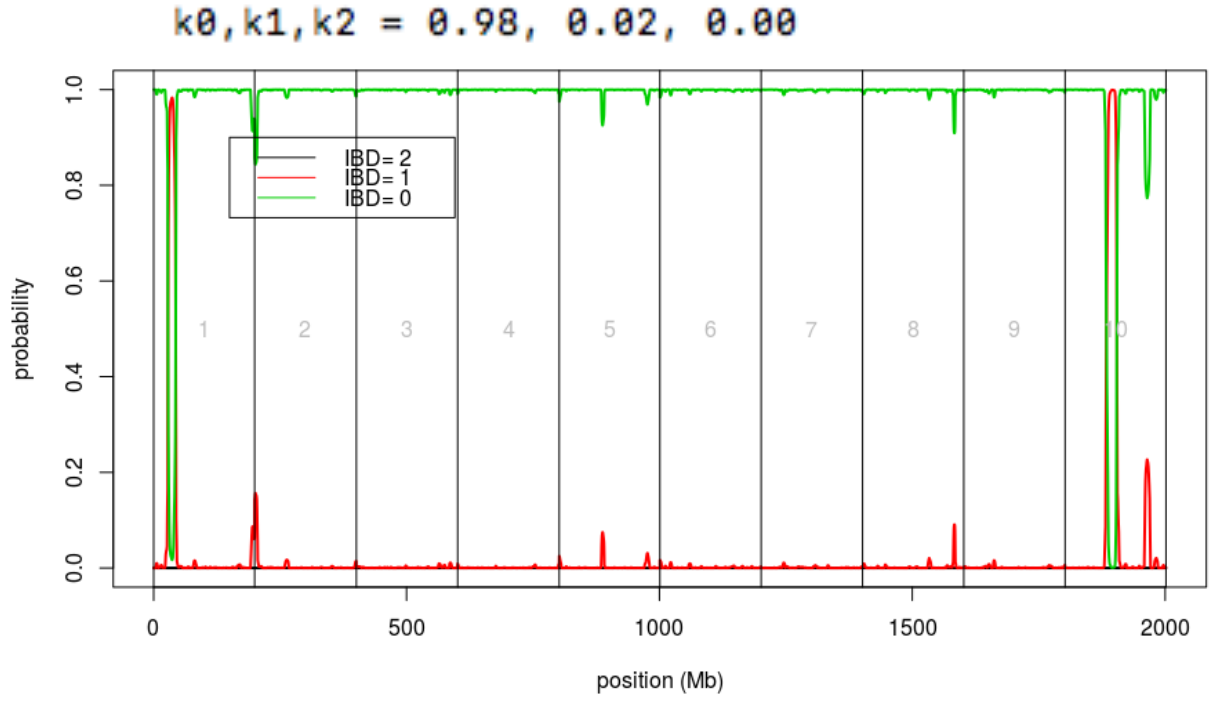




- Numerous long IBD tracts!

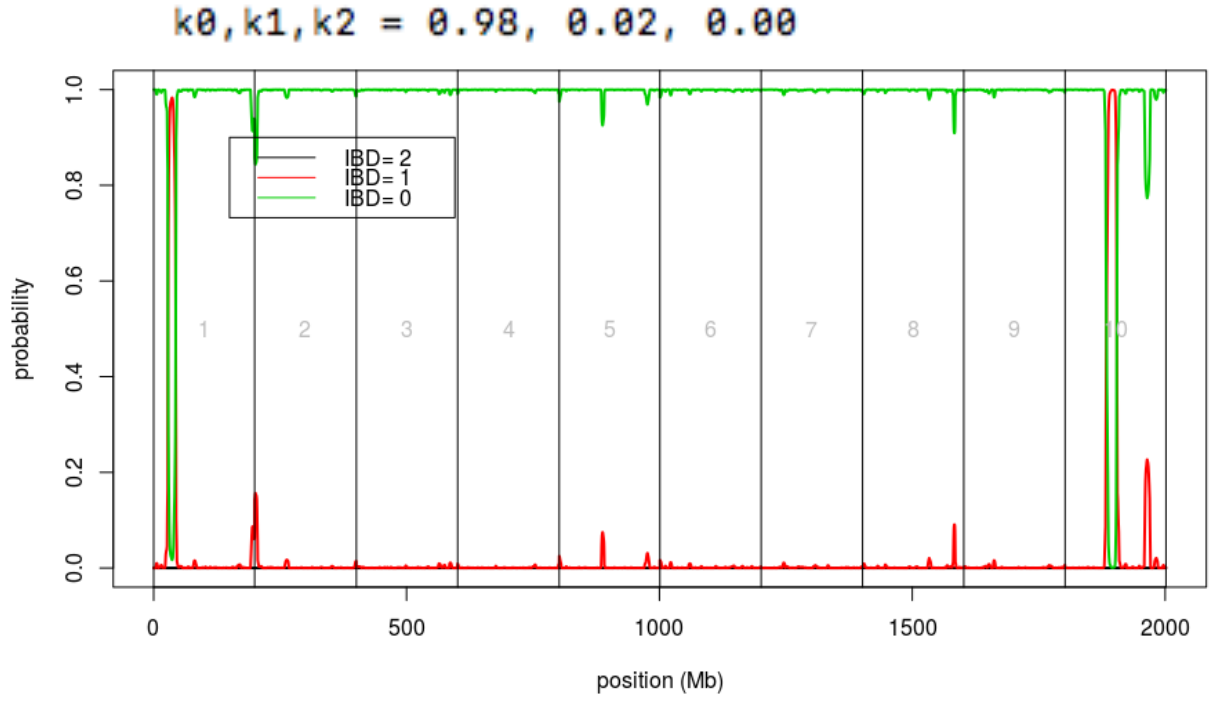
Try on simulated data from distantly related individuals

- Running Relate on the simulated data from distantly related we got:



Try on simulated data from distantly related individuals

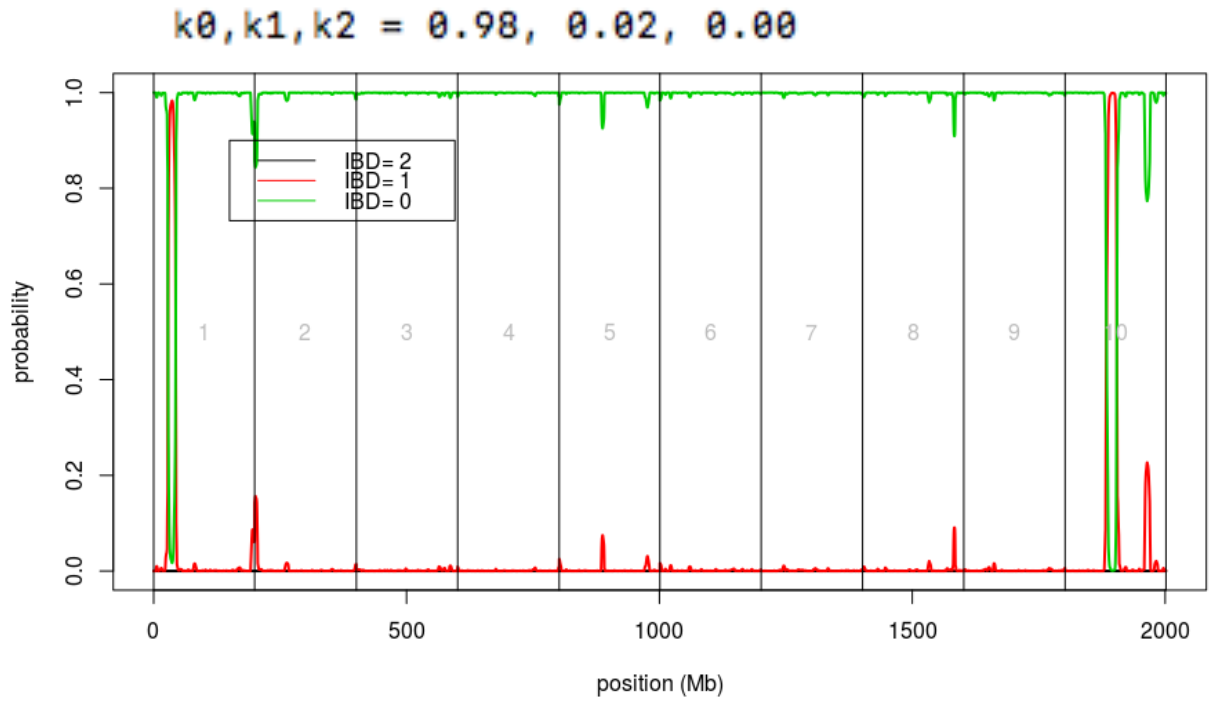
- Running Relate on the simulated data from distantly related we got:



- These individuals seem to share two regions IBD. Imagine they both have a certain disease. What assumptions do we have to make to conclude that the disease causing locus is in one of these regions?

Try on simulated data from distantly related individuals

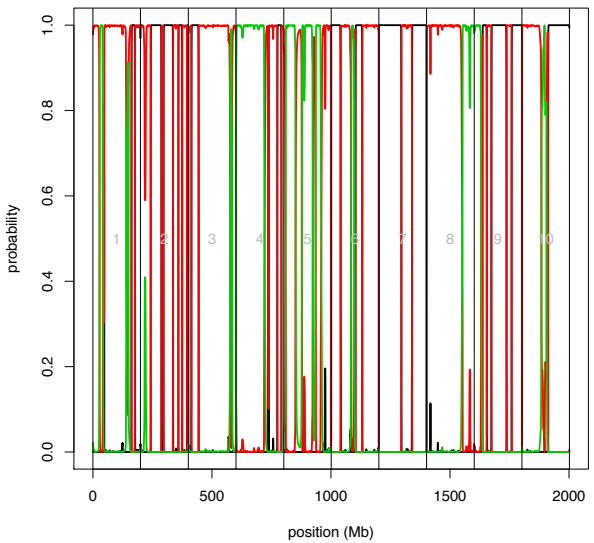
- Running Relate on the simulated data from distantly related we got:

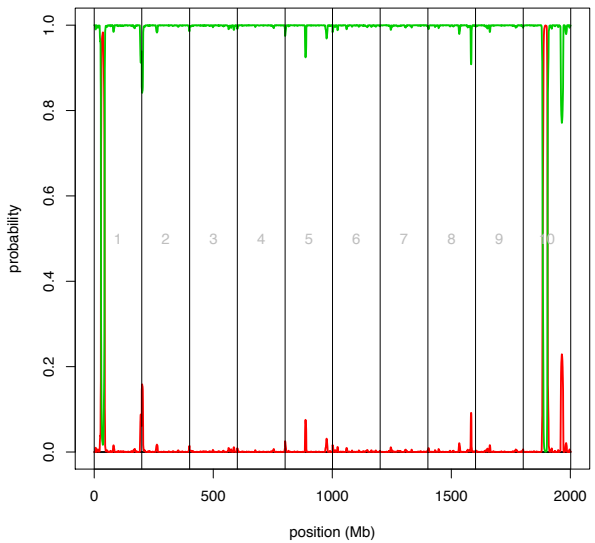


- These individuals seem to share two regions IBD. Imagine they both have a certain disease. What assumptions do we have to make to conclude that the disease causing locus is in one of these regions?
- That the disease is not recessive, that it is caused by the same mutation and that the IBD inference results are correct.

Idea behind the strategy for disease mapping

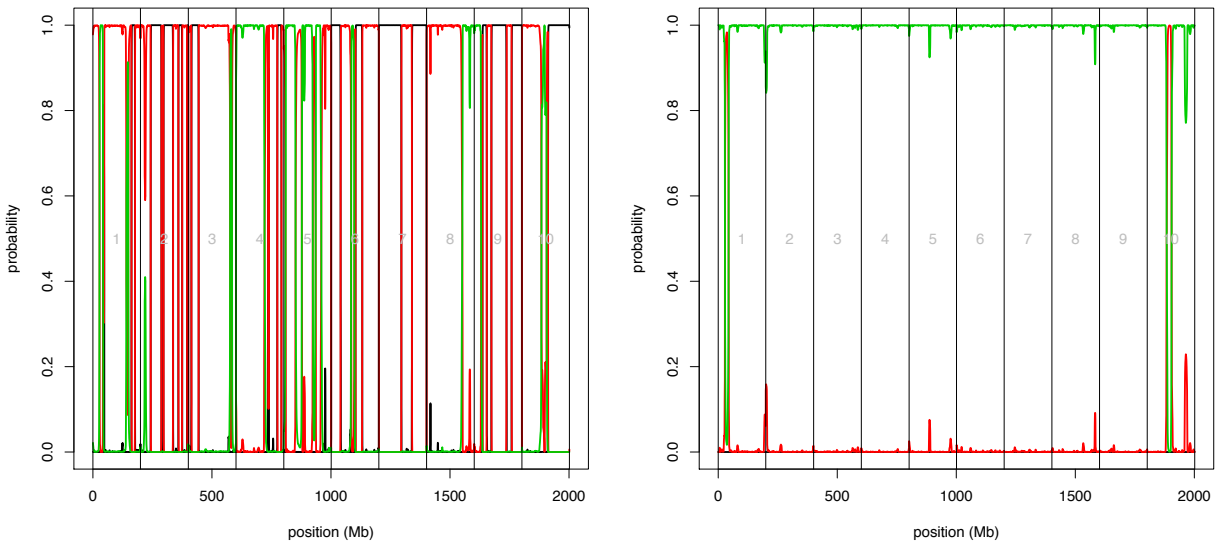
Based on the results you have seen so far: why do you think the idea in relatedness mapping is to use seemingly unrelated/distantly related individuals?





Idea behind the strategy for disease mapping

Based on the results you have seen so far: why do you think the idea in relatedness mapping is to use seemingly unrelated/distantly related individuals?



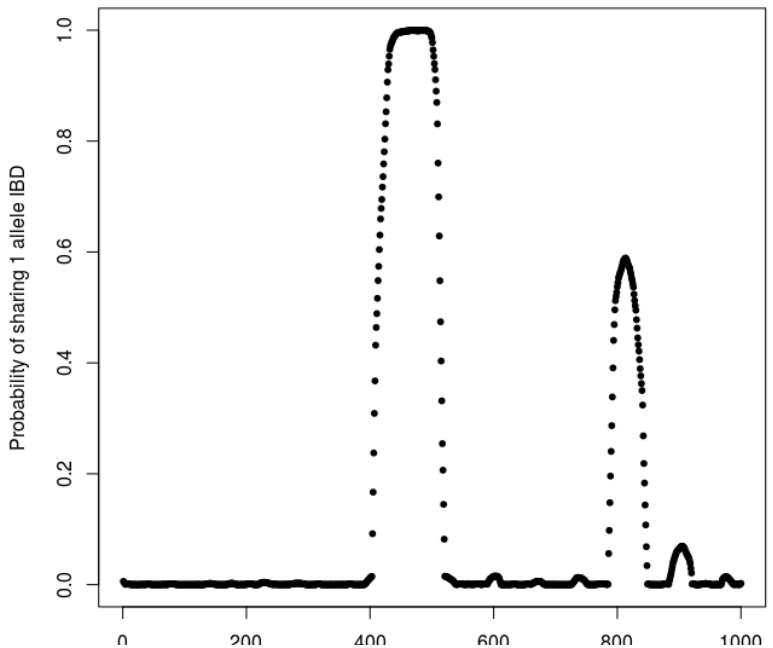
It makes mapping much more accurate and powerful, because distantly related individuals are expected to share much fewer and rather short regions IBD, hence this strategy leads to fewer and shorter candidate regions.

Relatedness mapping
(Exercise 4)

Relatedness mapping

We performed mapping using 10 cases and 10 controls in a few steps:

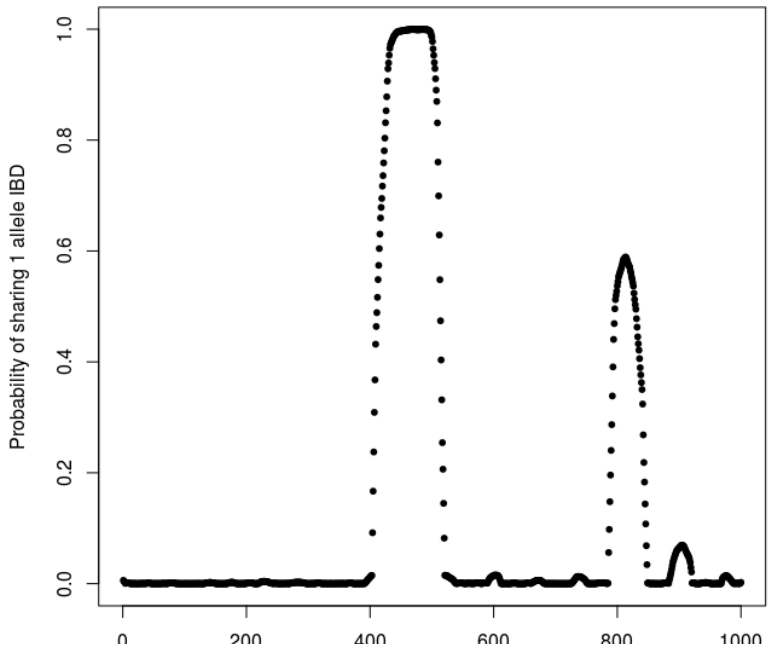
- First we used Relate to estimate the probability of IBD=1 along the genome for all pairs. E.g. for first pair (both cases):



Relatedness mapping

We performed mapping using 10 cases and 10 controls in a few steps:

- First we used Relate to estimate the probability of IBD=1 along the genome for all pairs. E.g. for first pair (both cases):

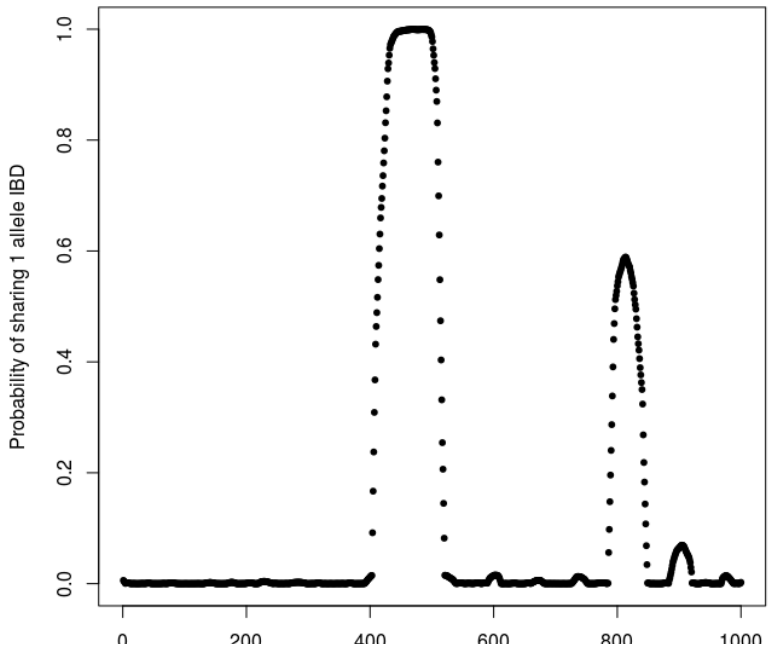


- They seem to share 1-2 regions IBD. What does this suggest about potential disease loci (given that they are both cases)?

Relatedness mapping

We performed mapping using 10 cases and 10 controls in a few steps:

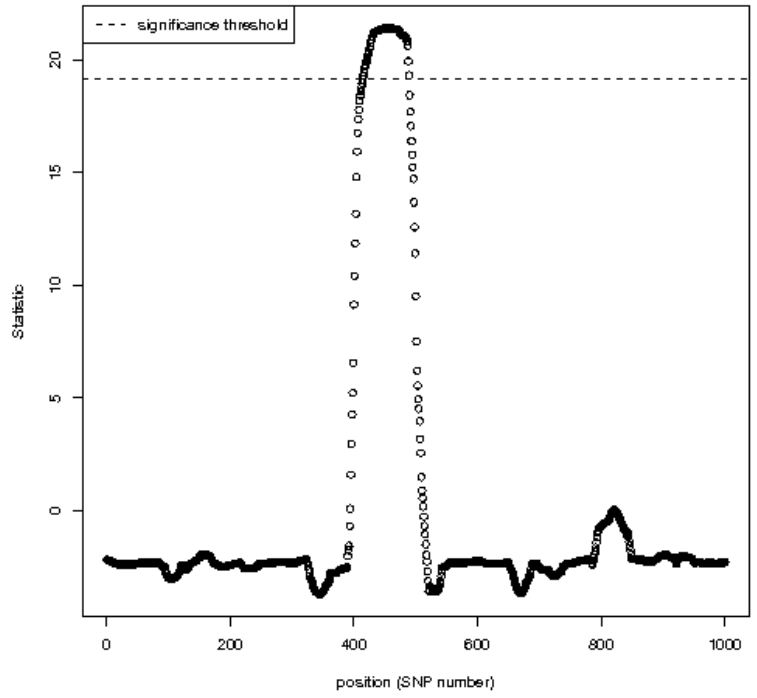
- First we used Relate to estimate the probability of IBD=1 along the genome for all pairs. E.g. for first pair (both cases):



- They seem to share 1-2 regions IBD. What does this suggest about potential disease loci (given that they are both cases)?
- Those two regions are candidate regions

Relatedness mapping

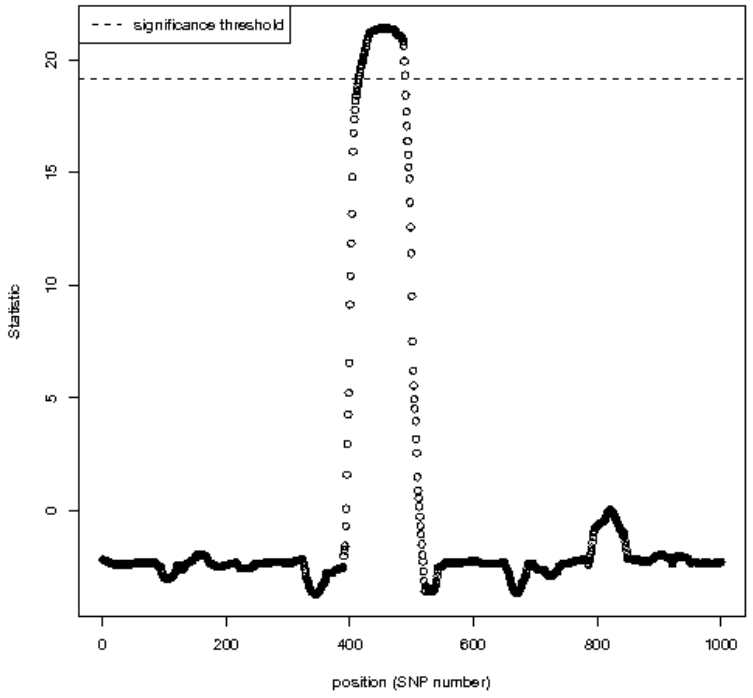
- Next we used Relate to test for a significant difference in IBD sharing among cases versus controls:



Locate the region with the causative SNP disease region

Relatedness mapping

- Next we used Relate to test for a significant difference in IBD sharing among cases versus controls:

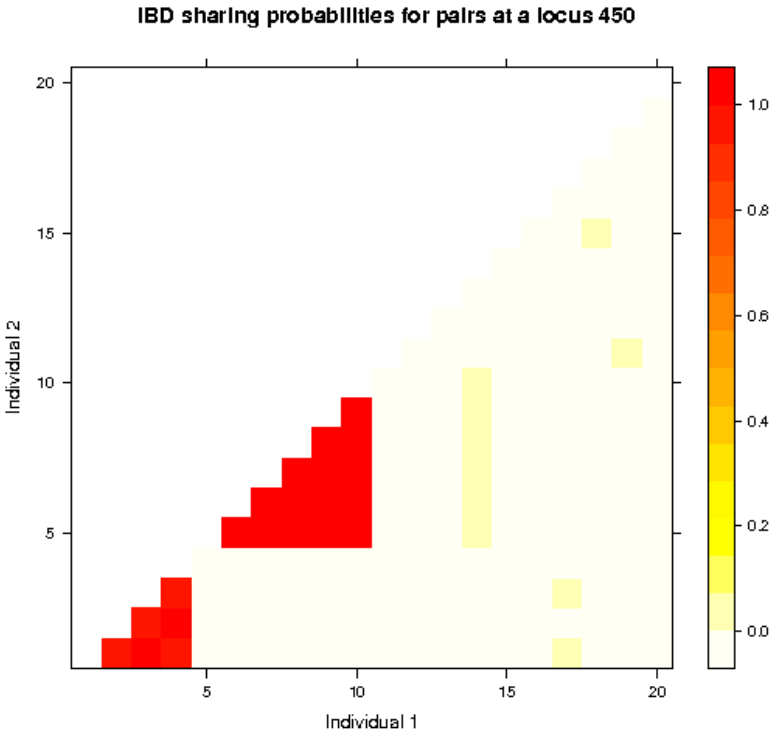


Locate the region with the causative SNP disease region

- Region between SNP 400 and SNP 500 (roughly)

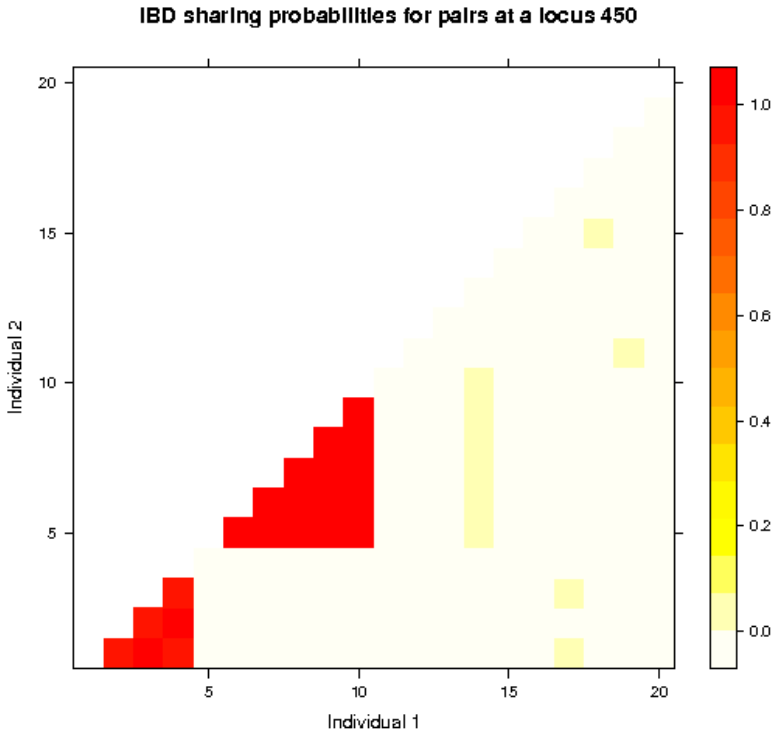
Relatedness mapping - exploring the candidate region

- Which of the individuals are IBD/related in this region?



Relatedness mapping - exploring the candidate region

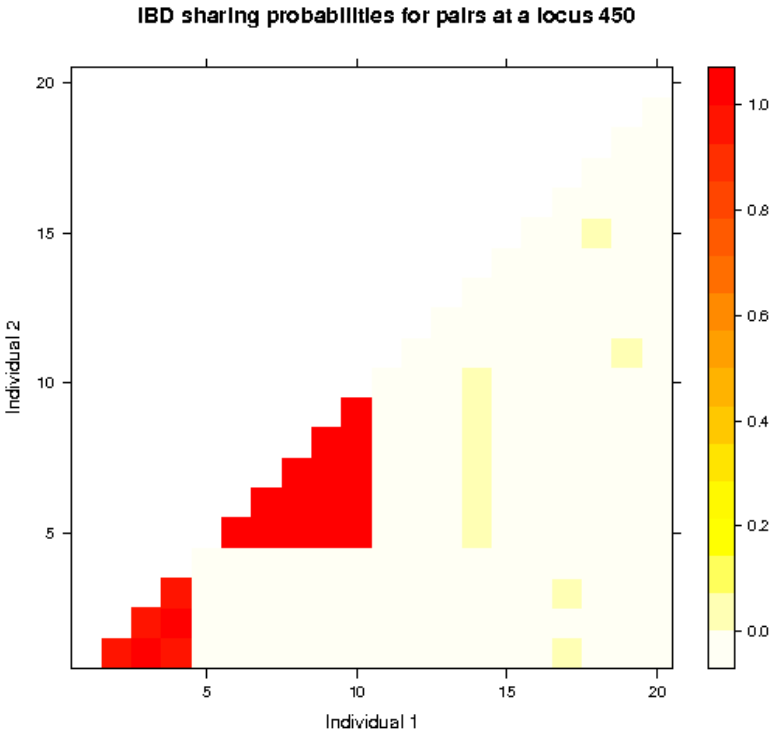
- Which of the individuals are IBD/related in this region?



- Individuals 1,2,3,4 are and individuals 5,6,7,8,9,10 are

Relatedness mapping - exploring the candidate region

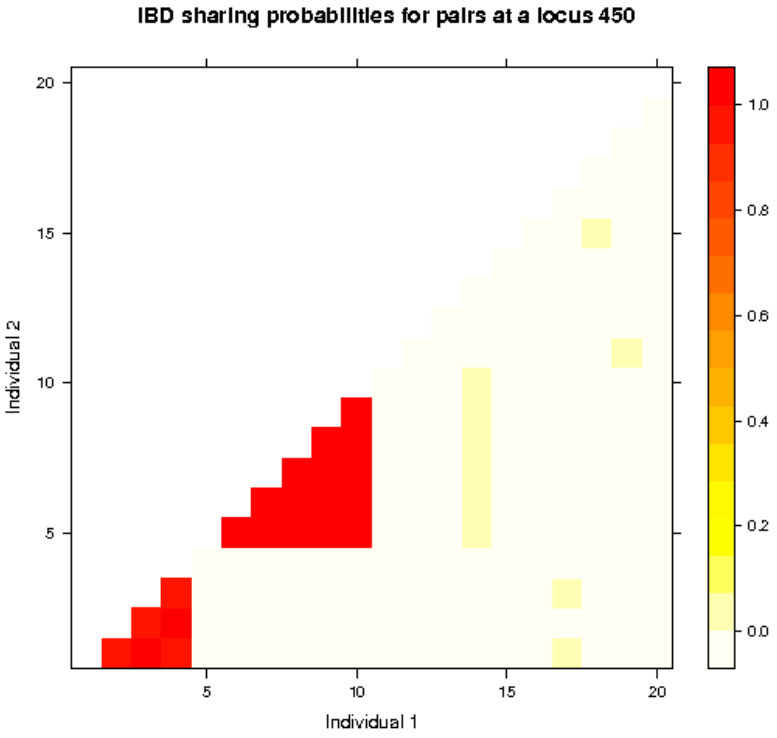
- Which of the individuals are IBD/related in this region?



- Individuals 1,2,3,4 are and individuals 5,6,7,8,9,10 are
- How many disease causing mutations are there?

Relatedness mapping - exploring the candidate region

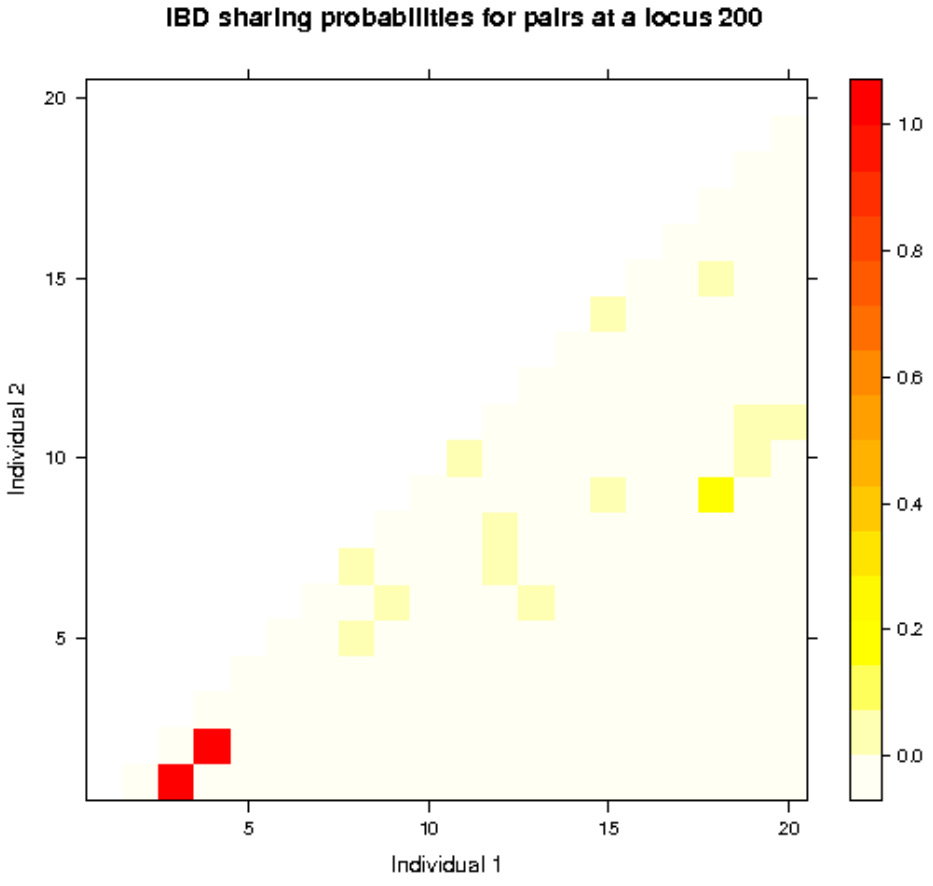
- Which of the individuals are IBD/related in this region?



- Individuals 1,2,3,4 are and individuals 5,6,7,8,9,10 are
- How many disease causing mutations are there?
- Likely 2 (could be 1 though)

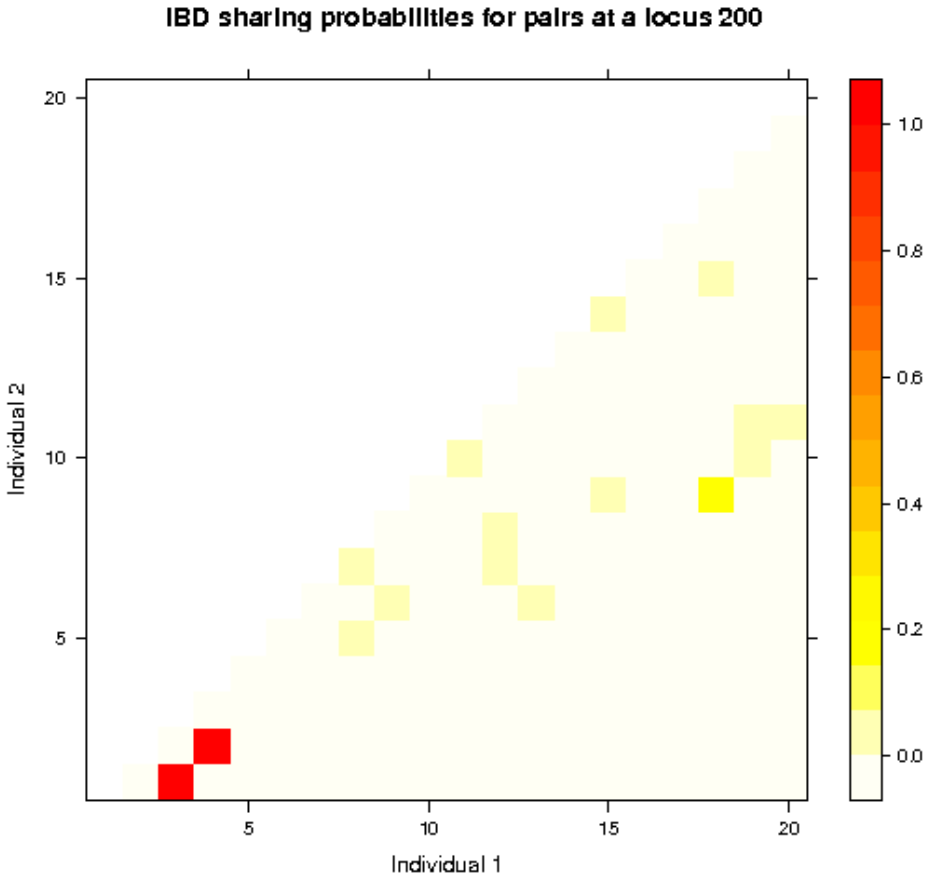
Relatedness mapping - exploring the candidate region

- How are the individuals IBD/related elsewhere?



Relatedness mapping - exploring the candidate region

- How are the individuals IBD/related elsewhere?



- Only a few pairs are IBD, e.g 6 and 9