

# *Blossoc*

---

---

## Fast association mapping by incompatibilities

*Implementation: [www.daimi.au.dk/~mailund/Blossoc](http://www.daimi.au.dk/~mailund/Blossoc)*

---

*Thomas Mailund  
Søren Besenbacher  
Mikkel H. Schierup*

# Setup: case/control sequences

Sequences of nucleotides  
at known polymorphic sites

Cases (affected)



|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| A | C | A | G | T | C | A |
| T | G | A | G | C | C | A |
| A | G | G | G | C | C | A |
| A | C | A | G | T | C | A |
| T | C | A | G | T | C | A |
| T | C | A | T | T | A | A |

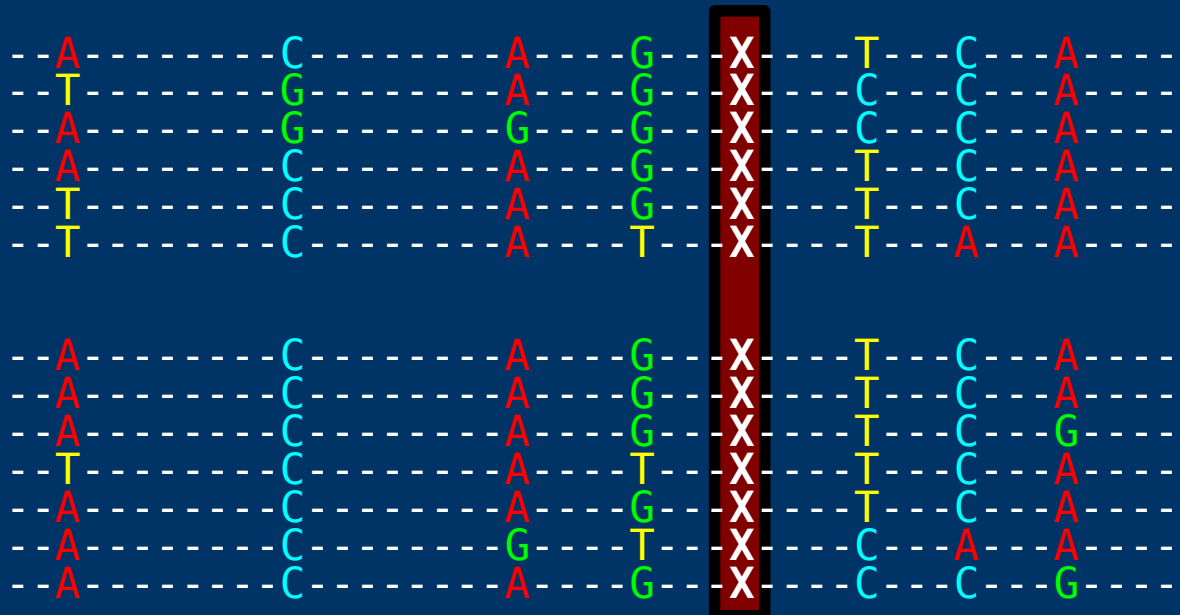
Controls (unaffected)



|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| A | C | A | G | T | C | A |
| A | C | A | G | T | C | A |
| A | C | A | G | T | C | G |
| T | C | A | T | T | C | A |
| A | C | A | G | T | C | A |
| A | C | G | T | C | A | A |
| A | C | A | G | C | C | G |

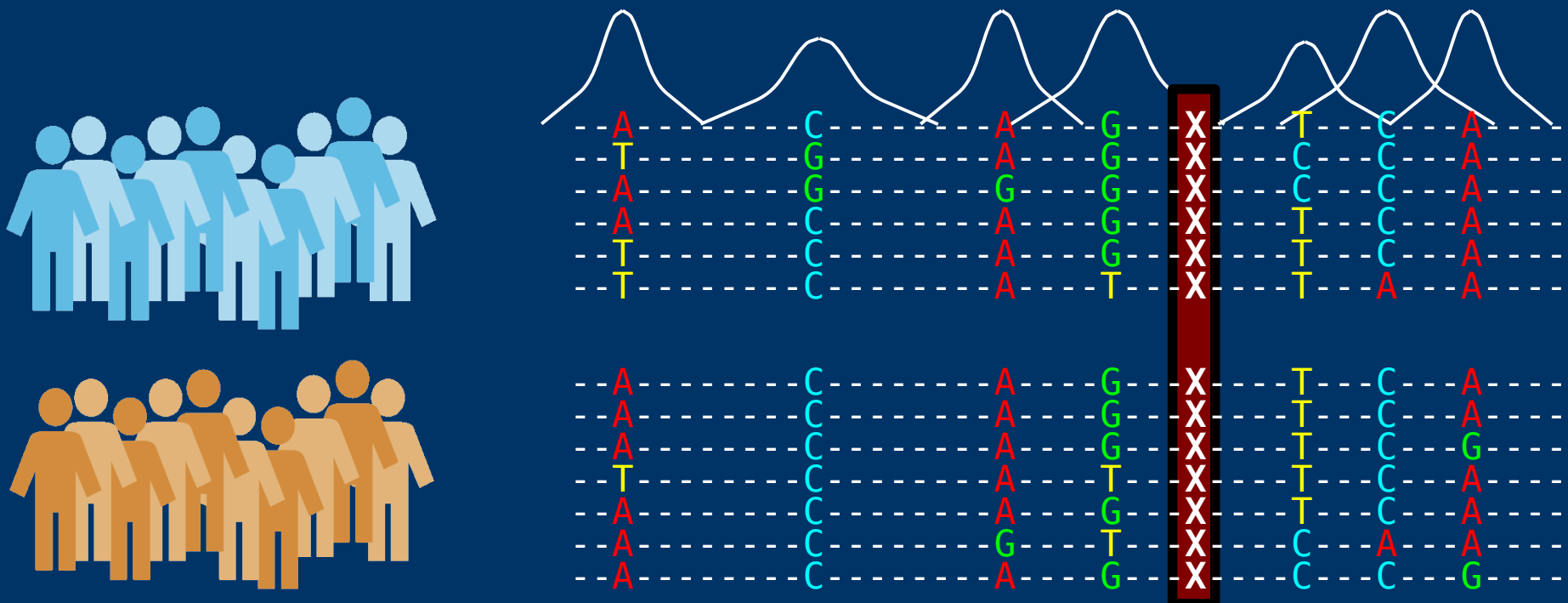
# Unknown causal variation

- Disease site unlikely to be among our markers
  - Might be an unknown polymorphic site
  - Just not part of the chosen markers



# Indirect signal for causal locus

- The markers are *not independent*
  - Knowing one marker is partial knowledge of others
  - This non-independence decreases with distance

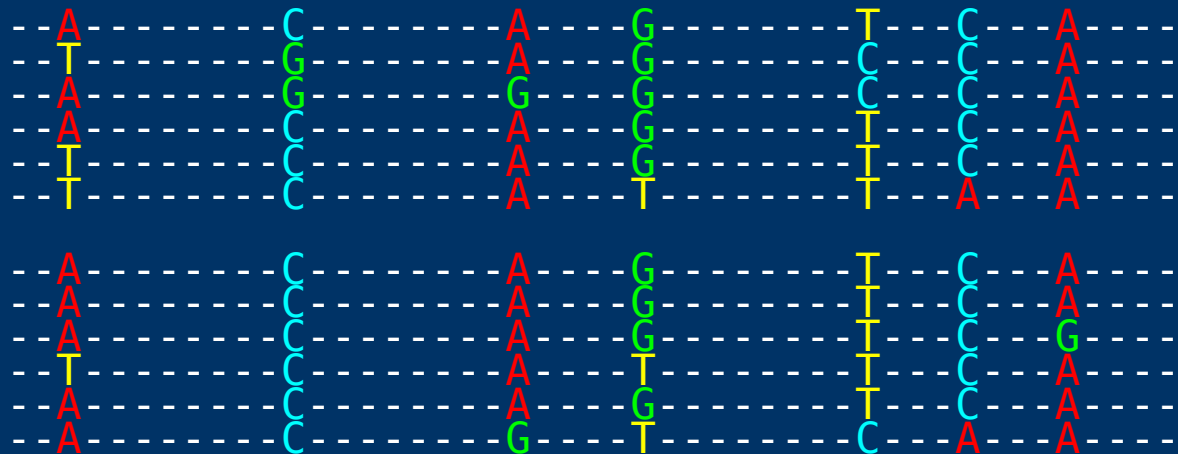




# Haplotype Pattern Mining (HPM)

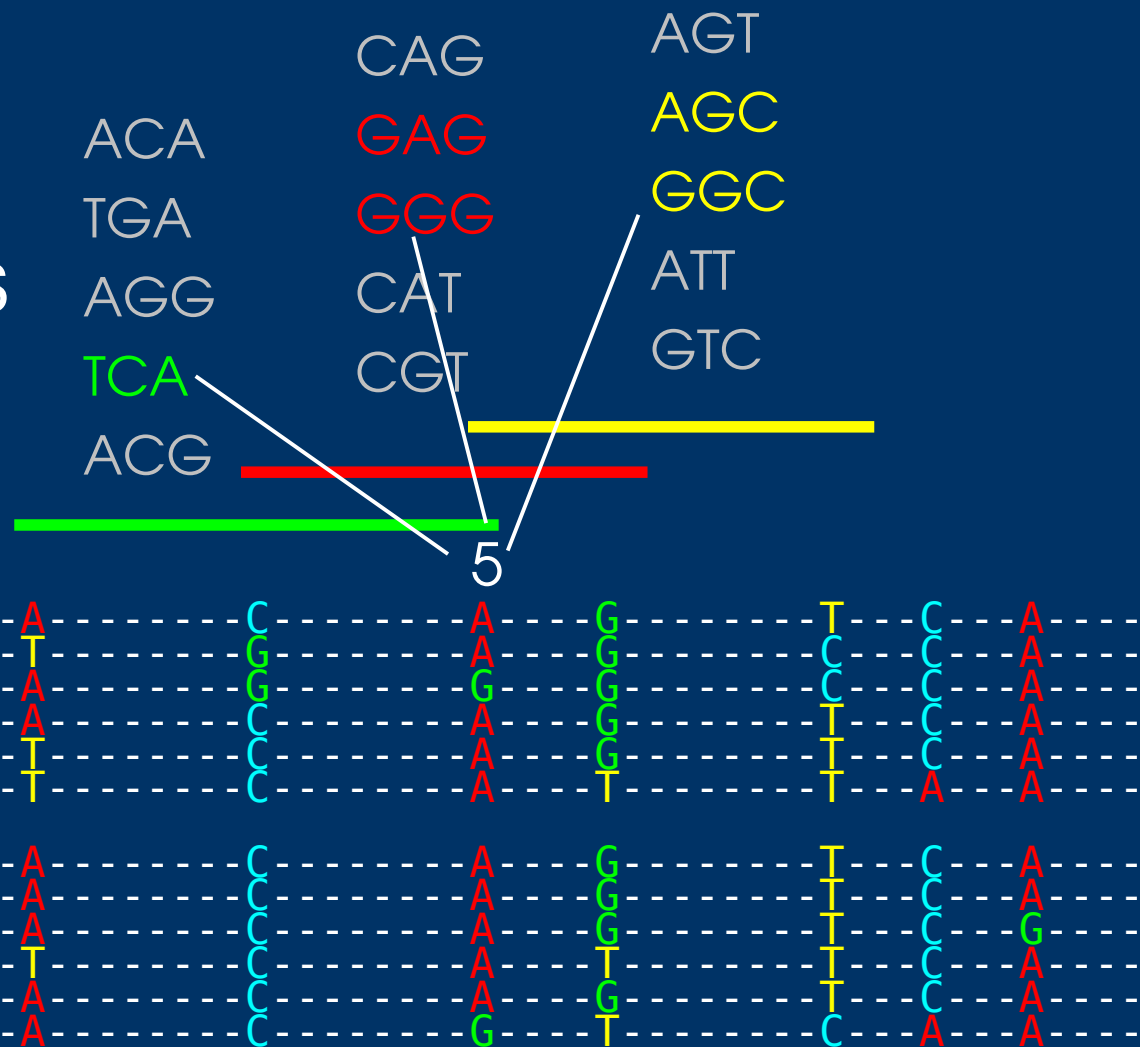
Test all patterns:

|     |     |     |
|-----|-----|-----|
| ACA | CAG | AGT |
| TGA | GAG | AGC |
| AGG | GGG | GGC |
| TCA | CAT | ATT |
| ACG | CGT | GTC |



# Haplotype Pattern Mining (HPM)

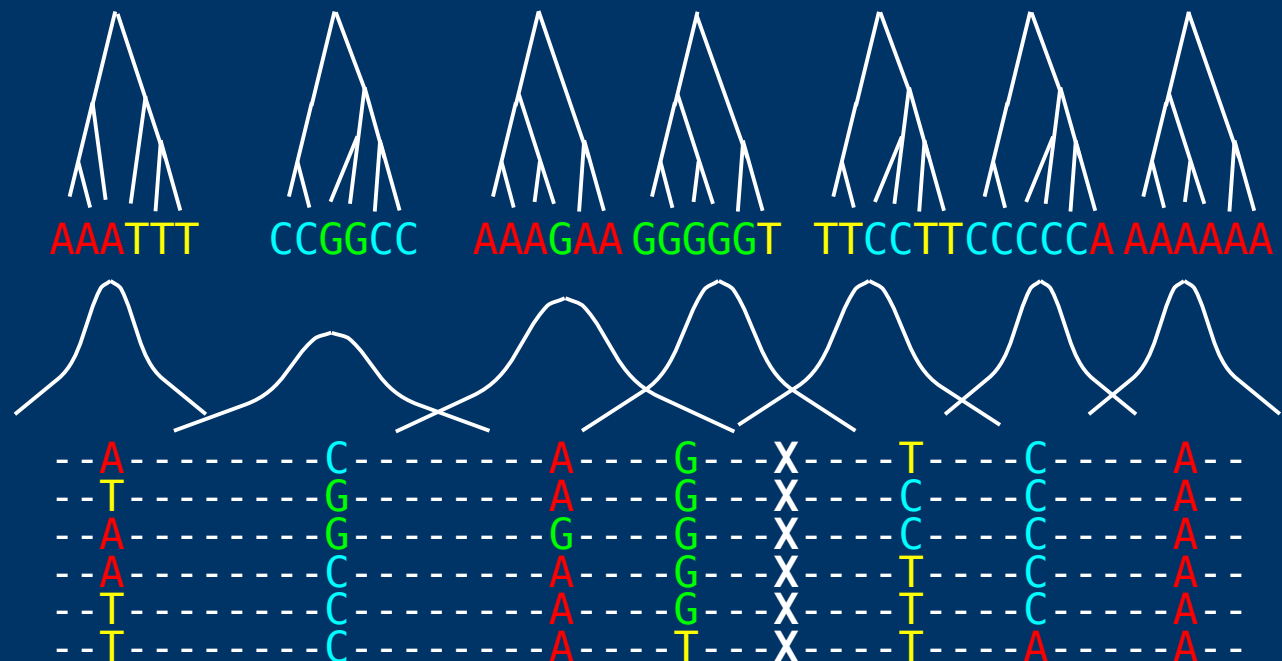
Score each locus  
by the number of  
significant patterns  
overlapping



# Using the (local) genealogy of the locus

- **Local** genealogies
  - Each site a different genealogy
  - Nearby genealogies only slightly different

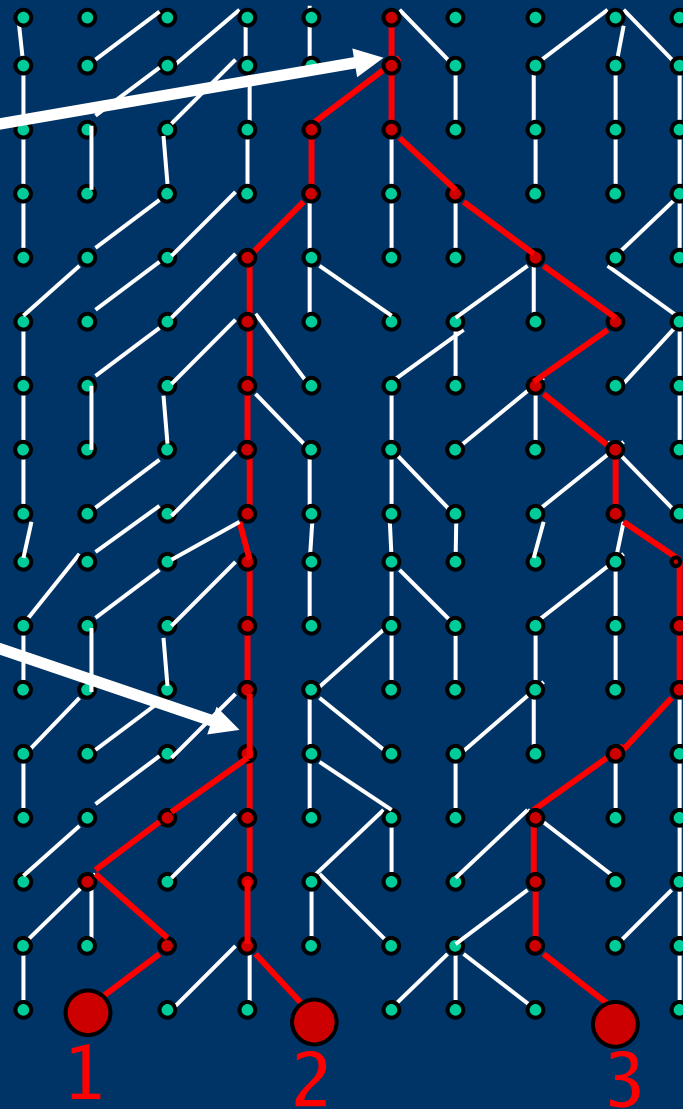
A nearby tree  
an imperfect  
local tree



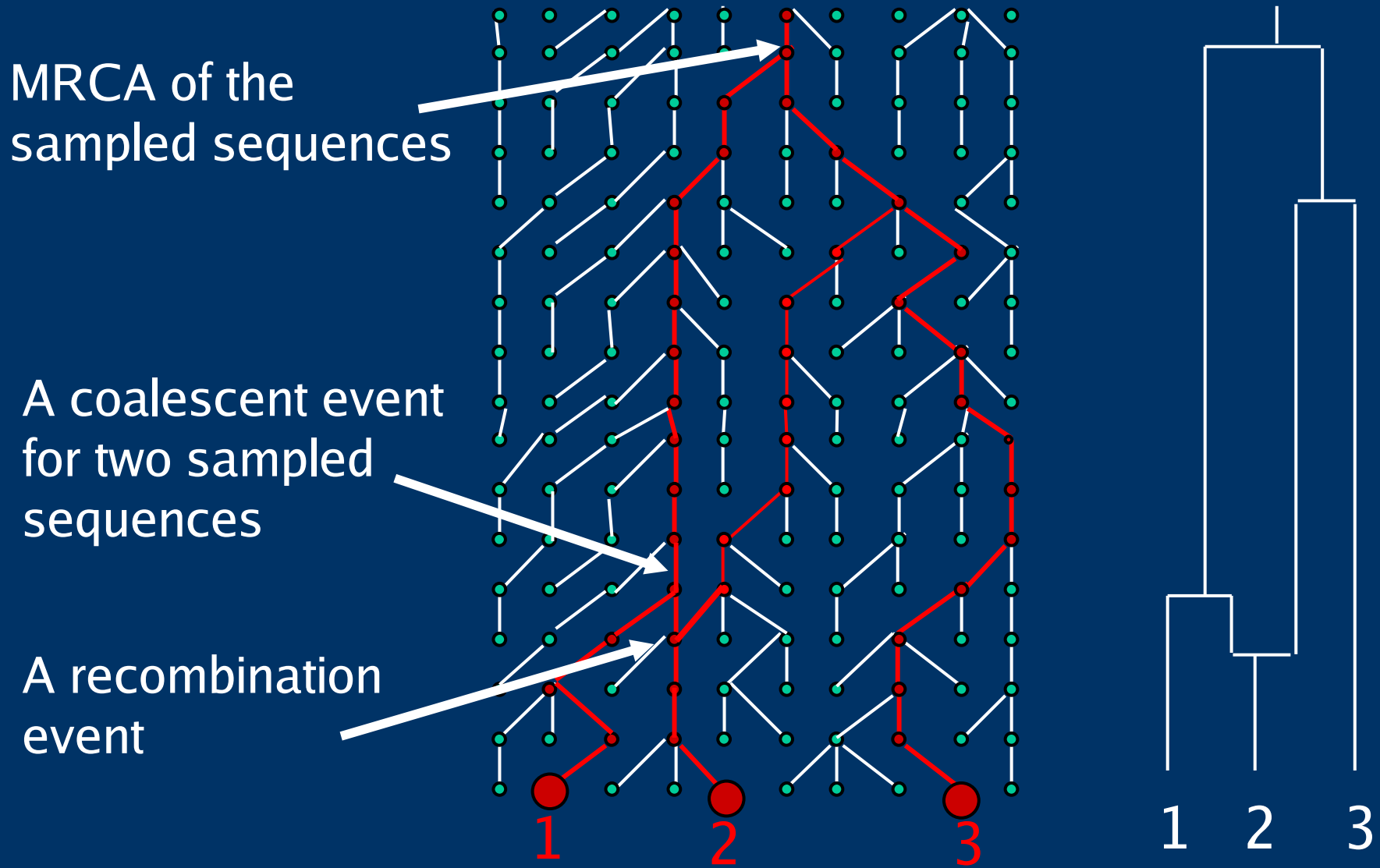
# Detour: Genealogies...

MRCA of the  
sampled sequences

A coalescent event  
for two sampled  
sequences

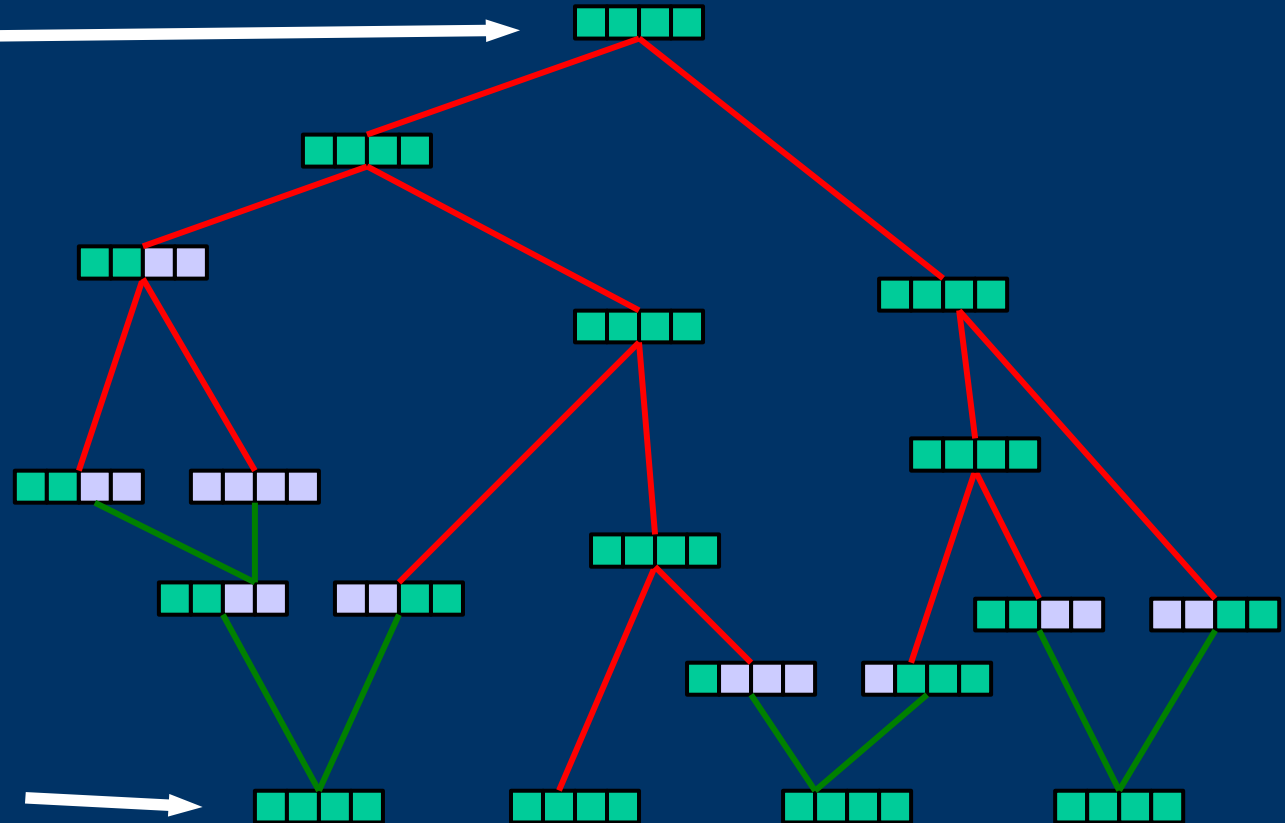


# Detour: Genealogies...



# Ancestral Recombination Graph

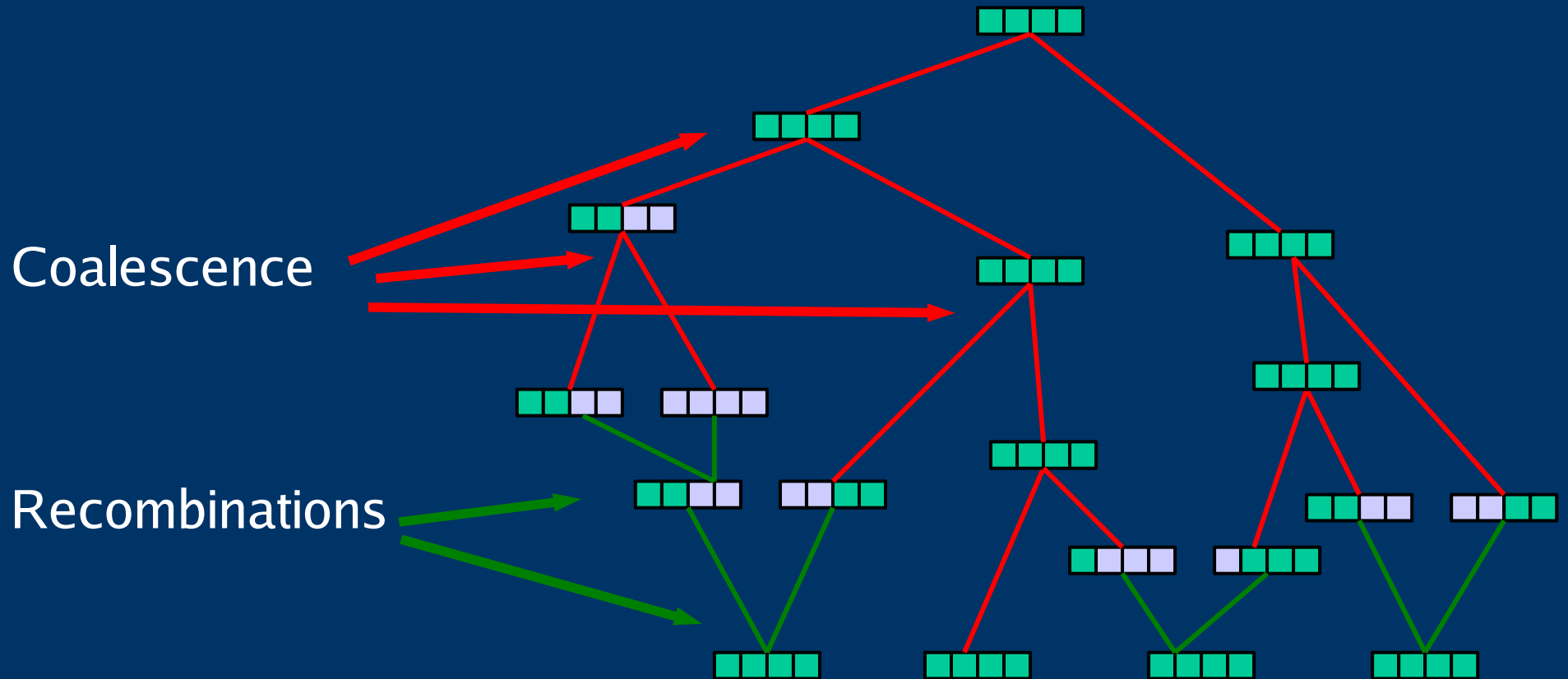
MRCA



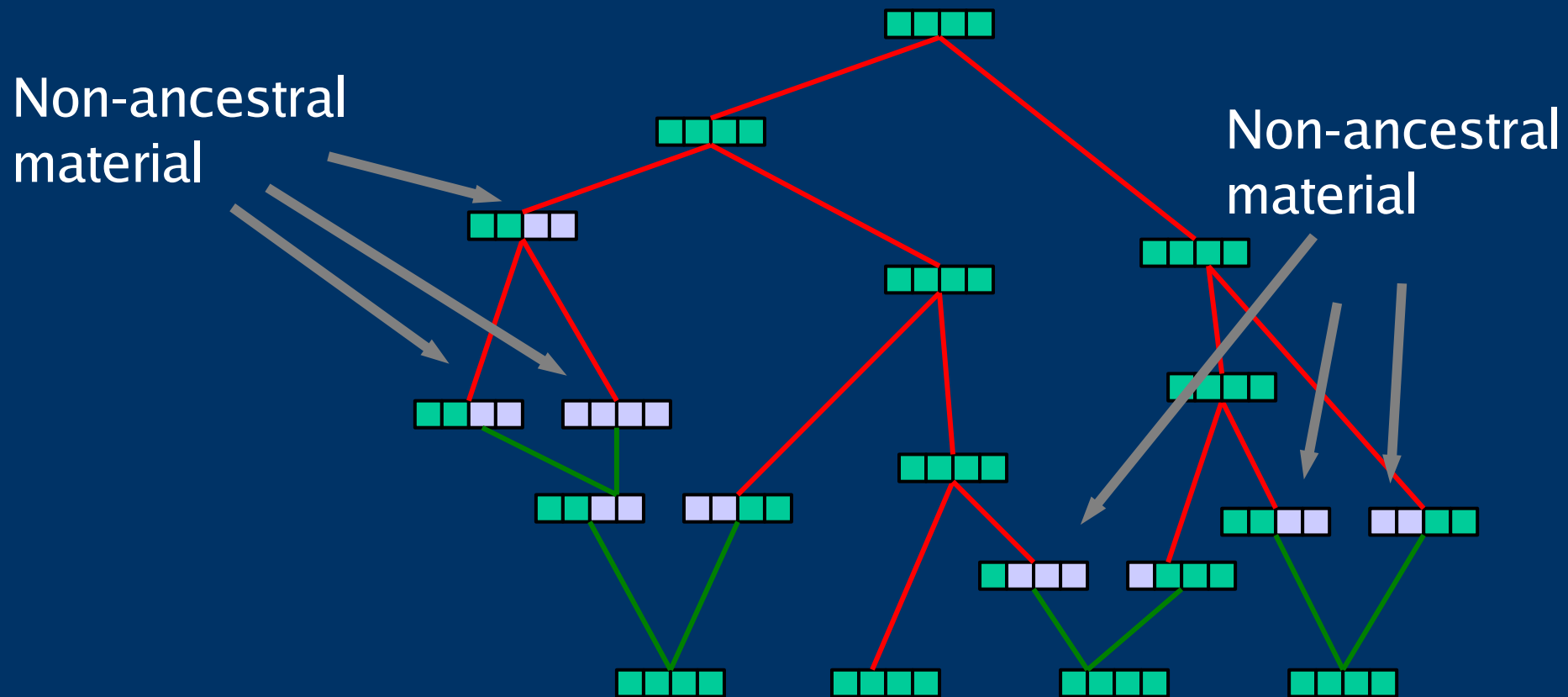
Sampled sequences



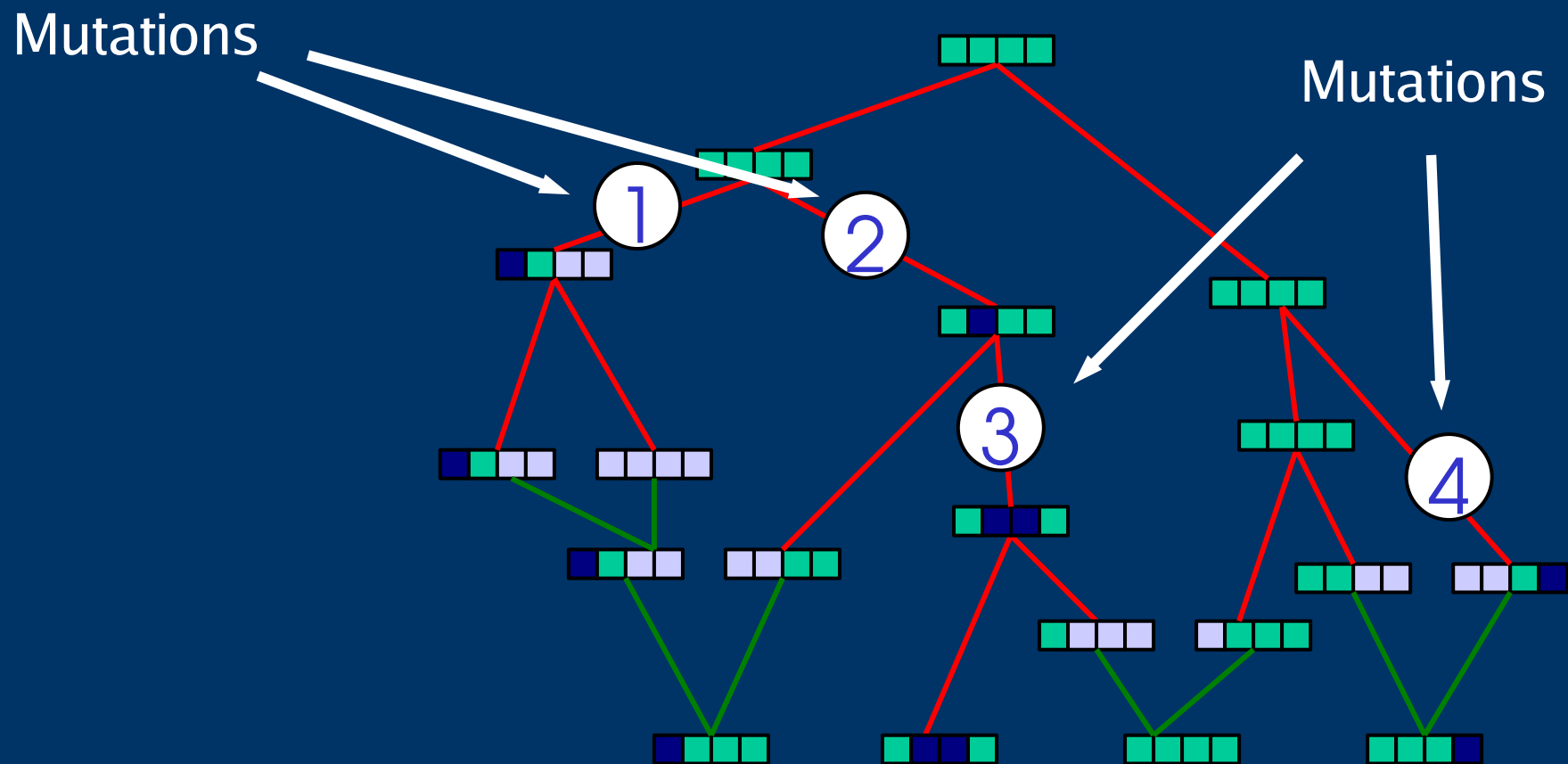
# Ancestral Recombination Graph



# Ancestral Recombination Graph

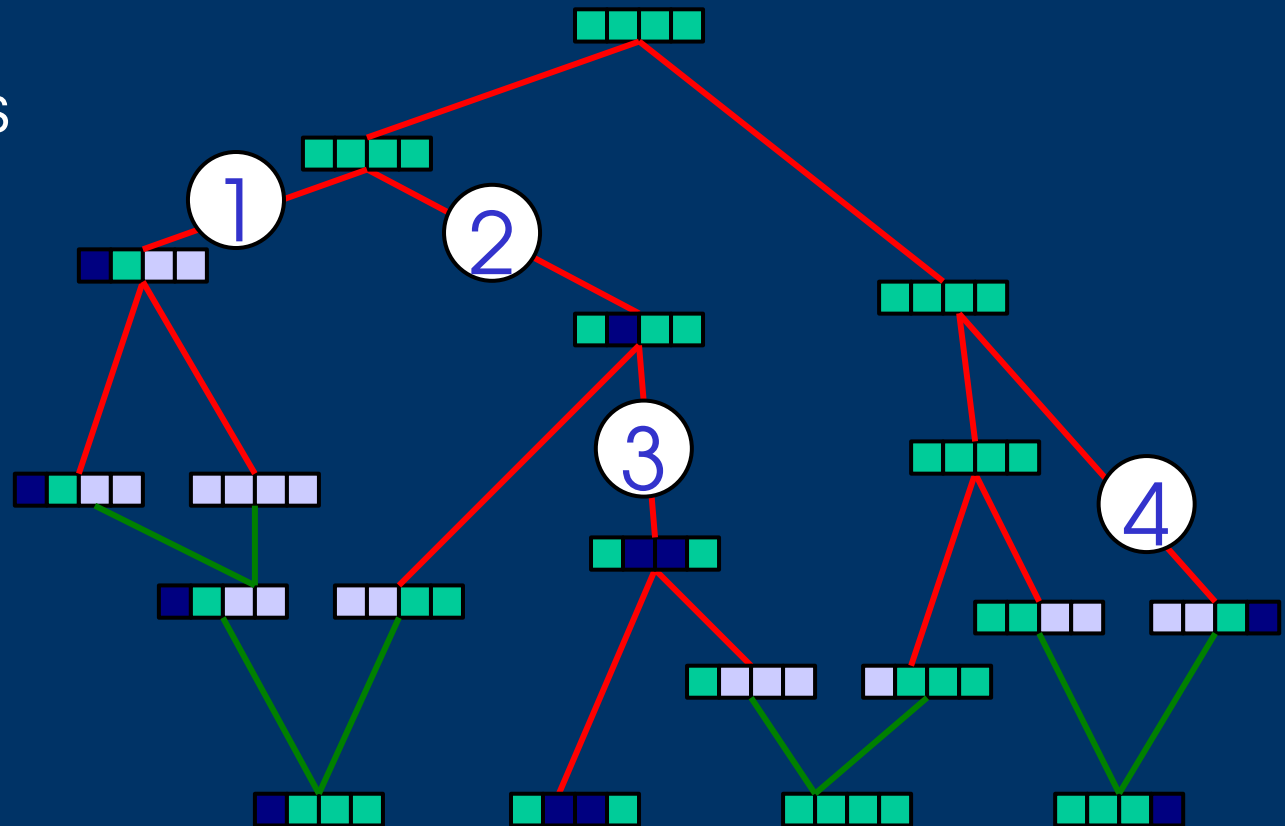


# Ancestral Recombination Graph



# Ancestral Recombination Graph

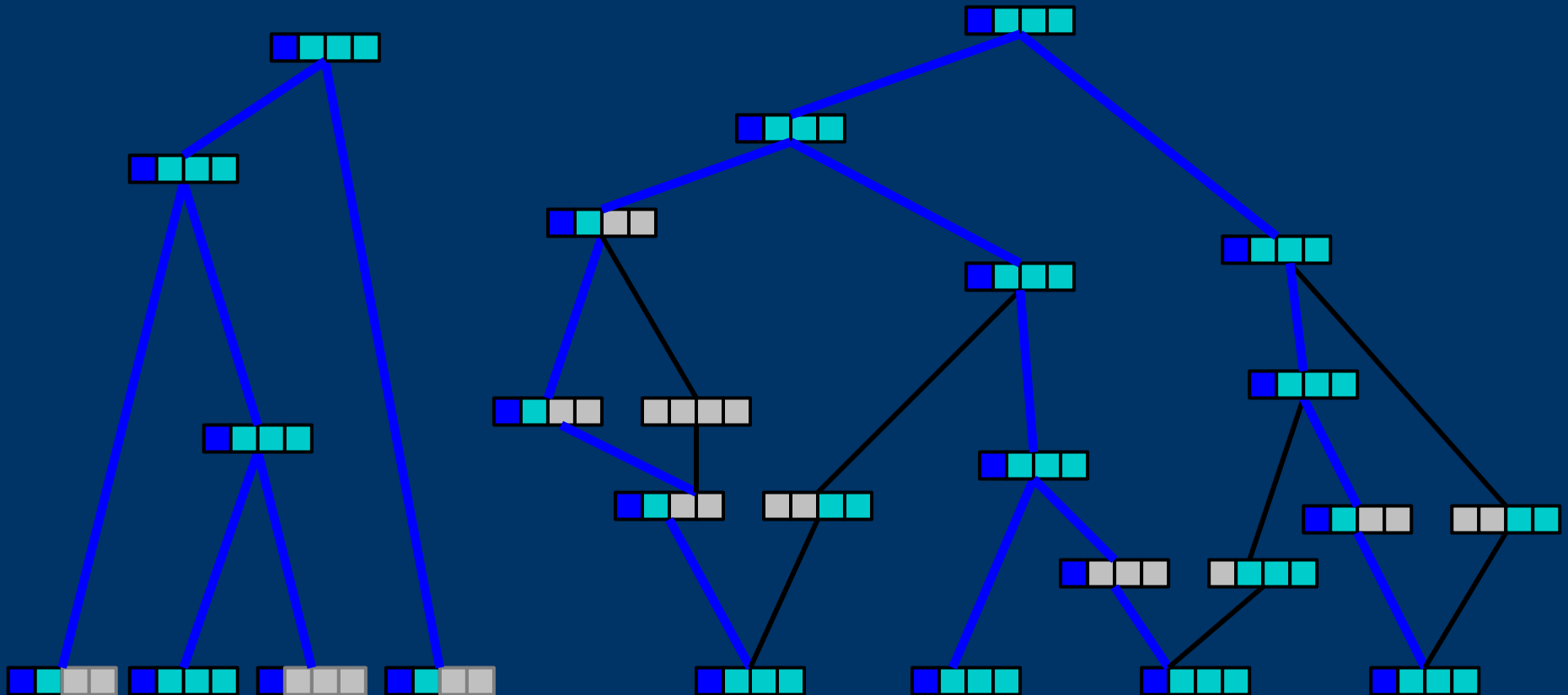
The ARG is a complete genealogy for the sampled sequences



# *“Local” genealogies*

---

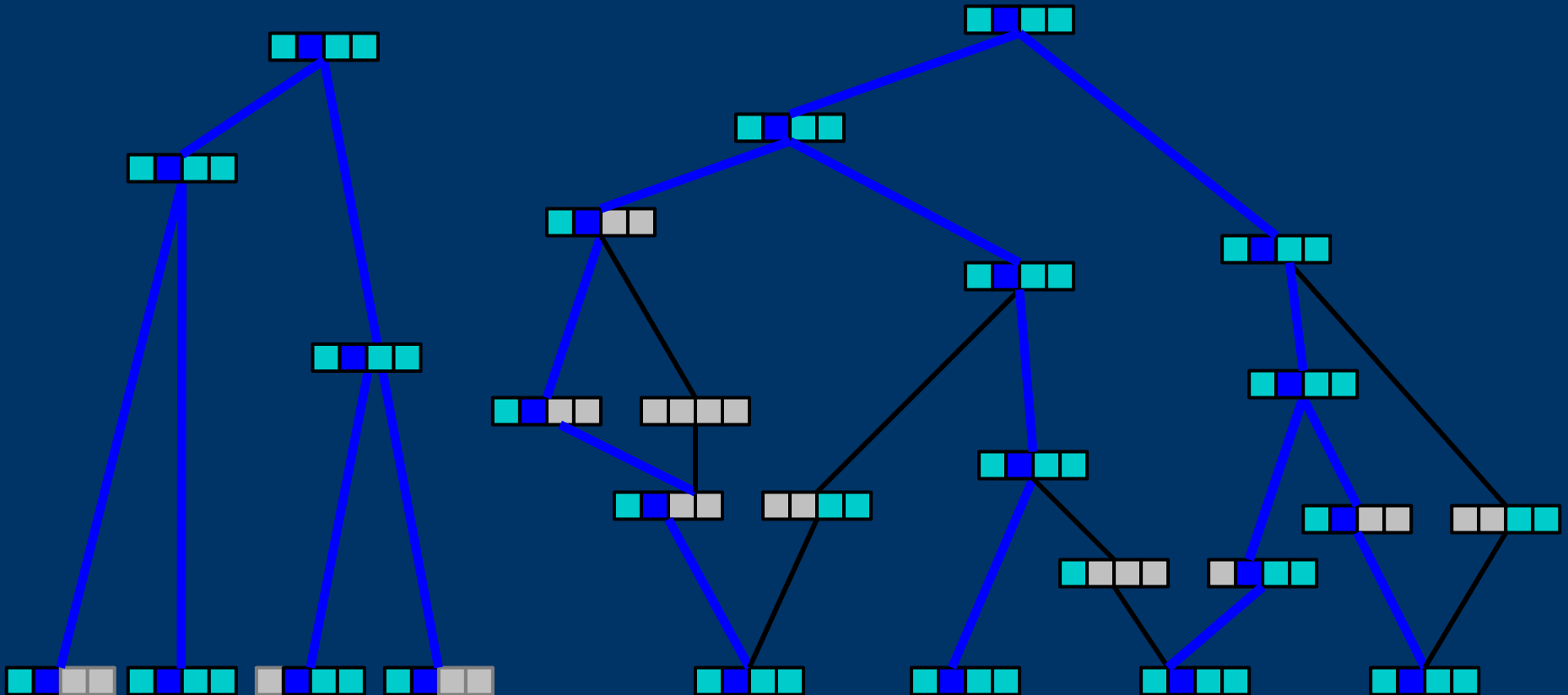
For each “point” on the chromosome, the ARG determines a (local) tree:



# “Local” genealogies

---

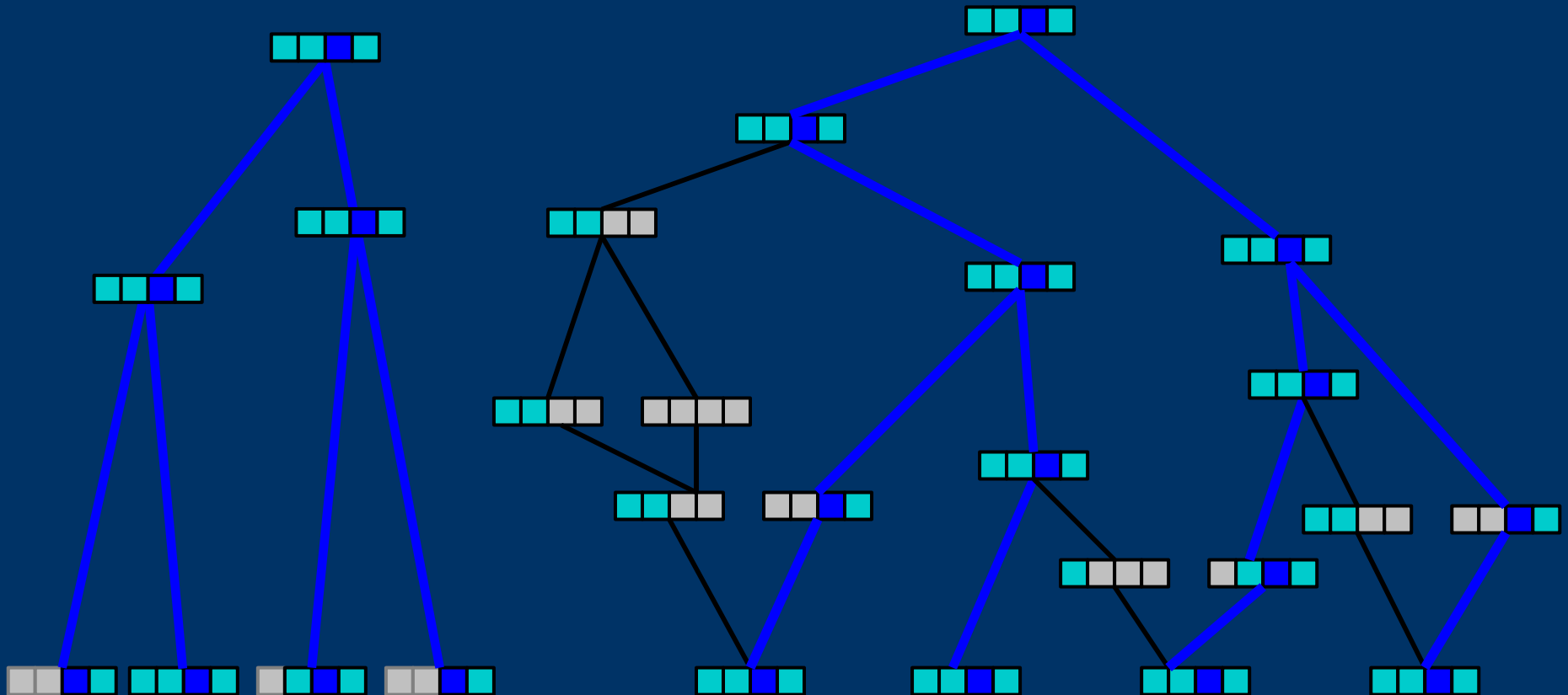
For each “point” on the chromosome, the ARG determines a (local) tree:



# *“Local” genealogies*

---

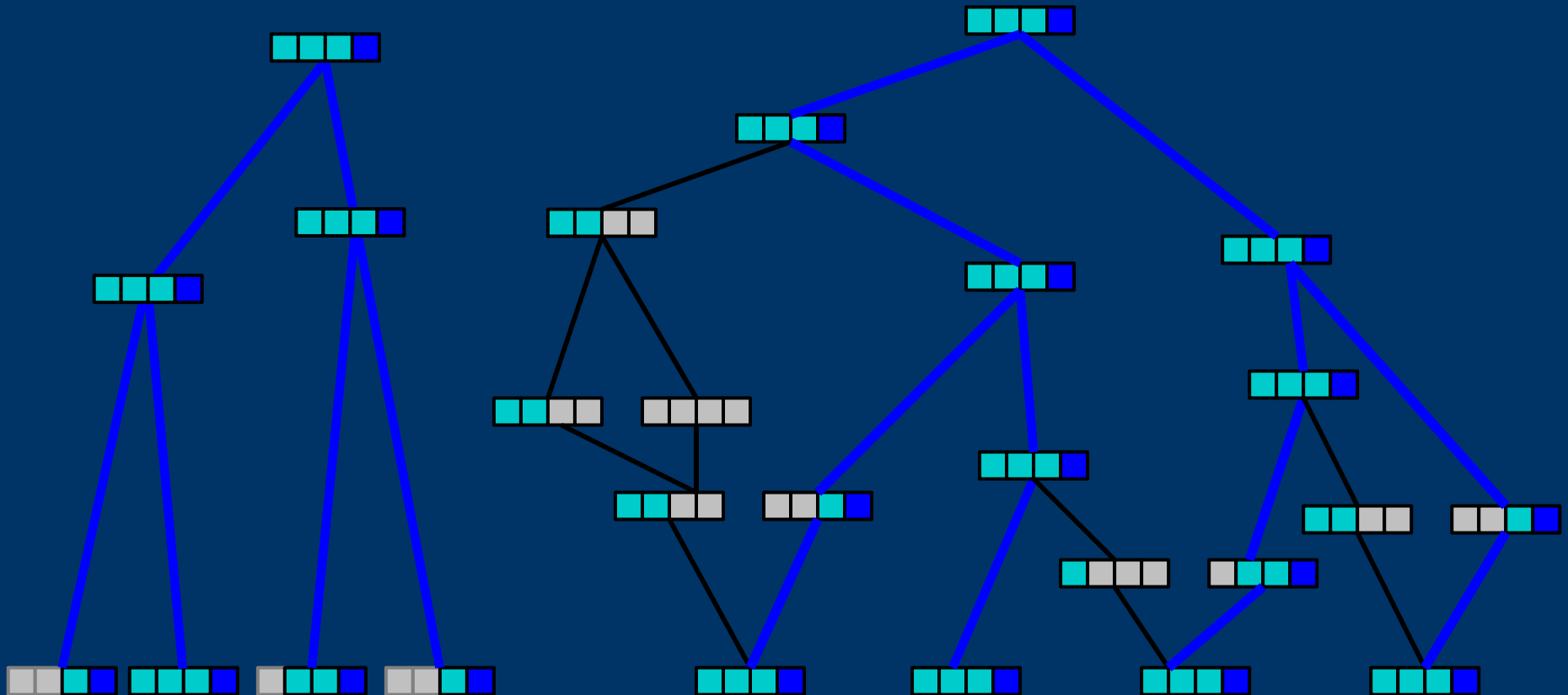
For each “point” on the chromosome, the ARG determines a (local) tree:



# “Local” genealogies

---

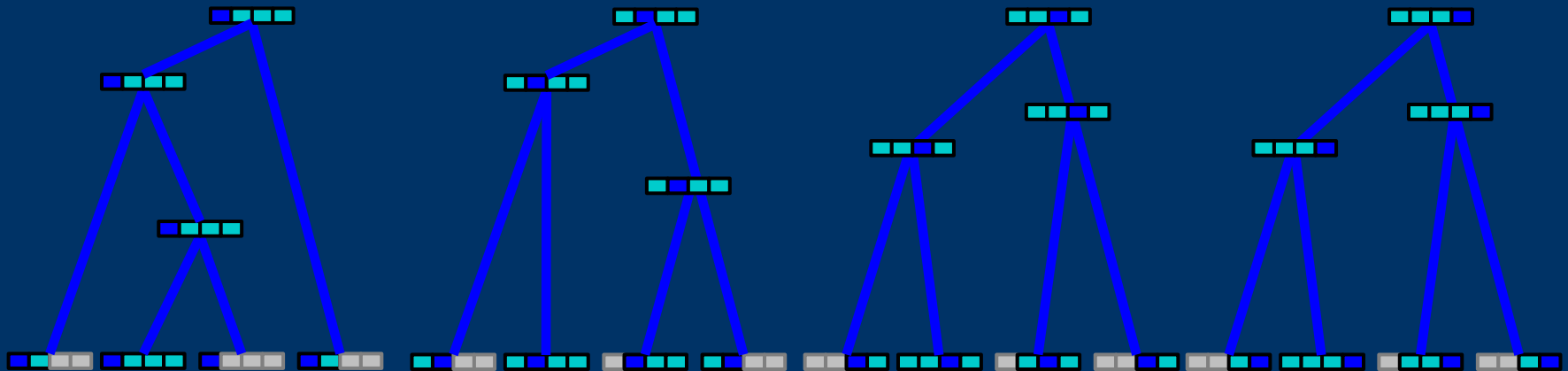
For each “point” on the chromosome, the ARG determines a (local) tree:



# *“Local” genealogies*

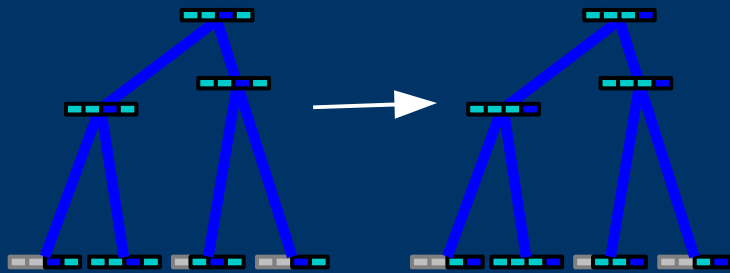
---

---

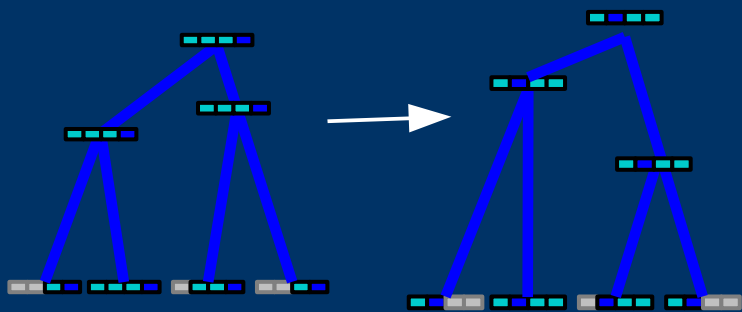


# “Local” genealogies

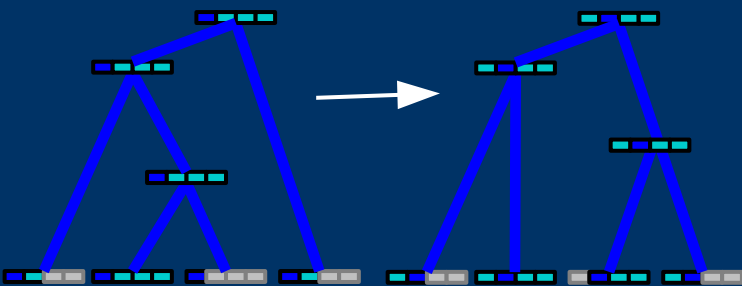
Type 1: No change



Type 2: Change in branch lengths



Type 3: Change in topology



**Table 5.1** Probability of different types of events<sup>a</sup>

| $n$      | $P_n^1$ | $P_n^2$ | $P_n^3$ |
|----------|---------|---------|---------|
| 2        | 0.333   | 0.667   | 0.000   |
| 3        | 0.297   | 0.703   | 0.000   |
| 4        | 0.272   | 0.655   | 0.073   |
| 5        | 0.256   | 0.616   | 0.134   |
| 6        | 0.243   | 0.574   | 0.183   |
| 10       | 0.212   | 0.488   | 0.300   |
| 15       | 0.191   | 0.435   | 0.374   |
| 500      | 0.098   | 0.232   | 0.670   |
| $\infty$ | 0       | 0       | 1       |

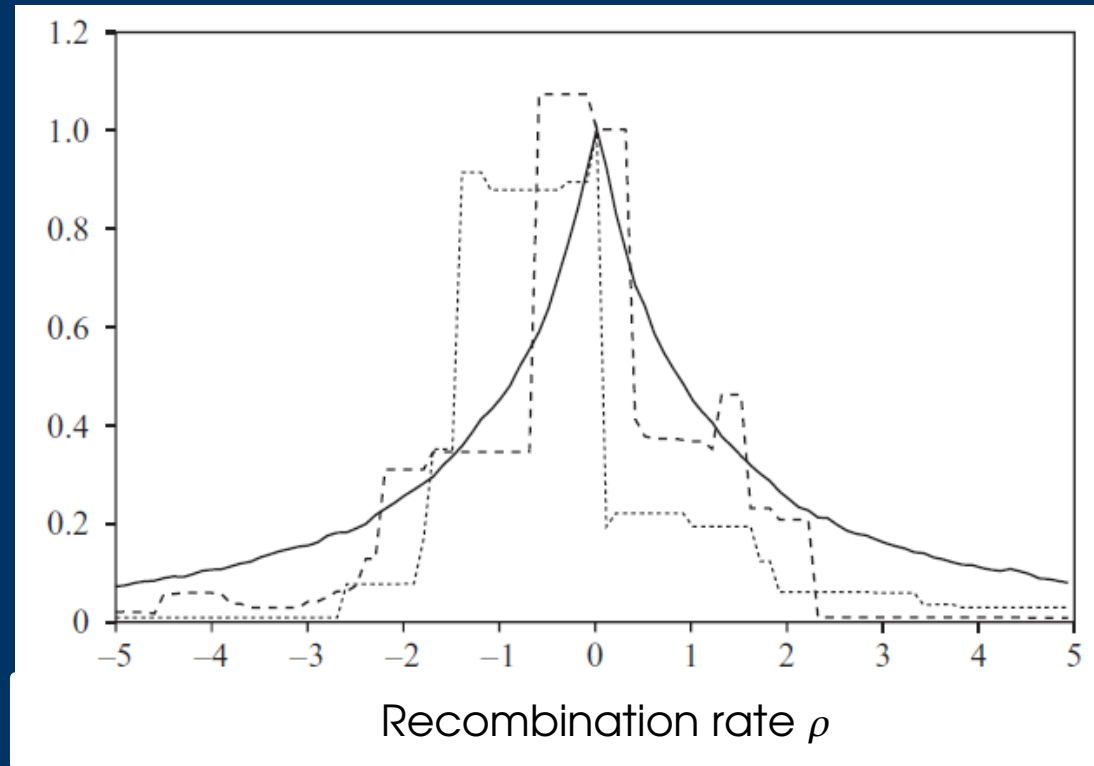
<sup>a</sup>  $P_n^i$  is the probability of a type  $i$  event in a sample of size  $n$ .

# “Local” genealogies

Tree measure:

$$M_{AB} = (\sum_{i,j \{i=j\}} bl(i)bl(j)) / tbl(A)tbl(B)$$

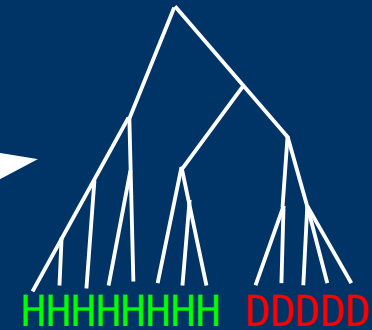
$$S_{AB} = M_{AB} / M_{AA}$$



From Hein *et al.* 2005

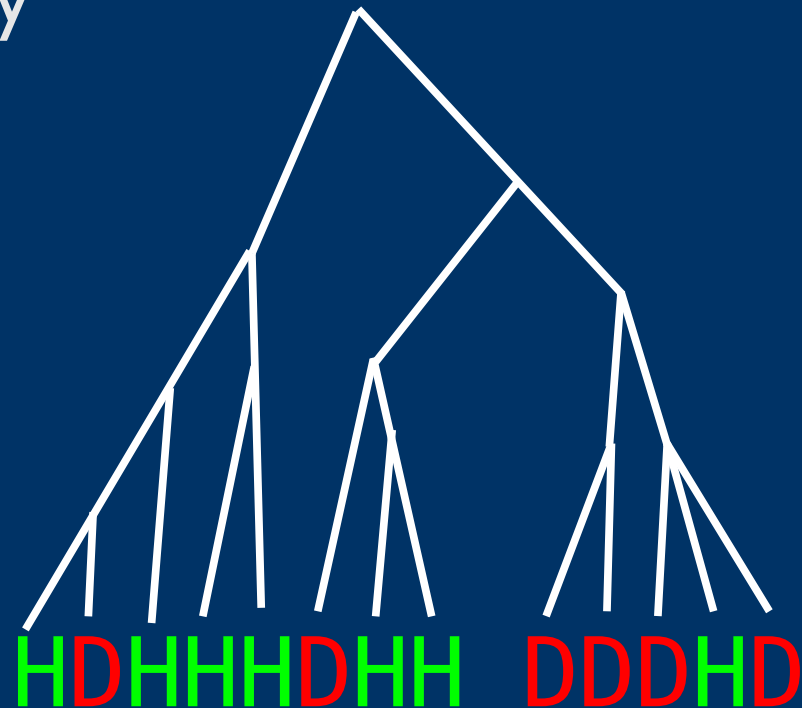
# Using the (local) genealogy of the locus

- Tree at disease site:
  - “Perfect” setup
  - Incomplete penetrance
  - Other disease causes



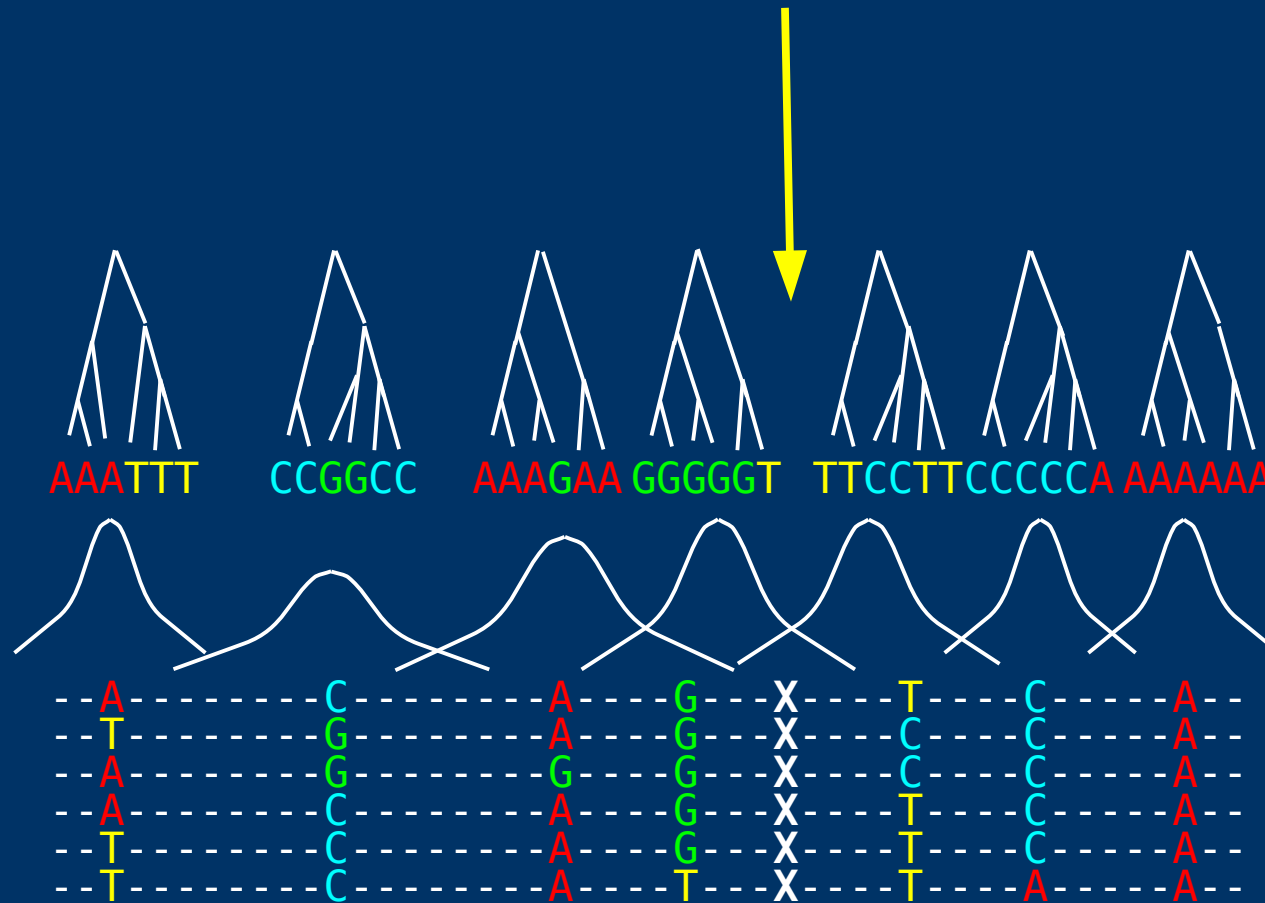
# Using the (local) genealogy of the locus

- At the disease site:
  - A significant clustering of diseased/healthy



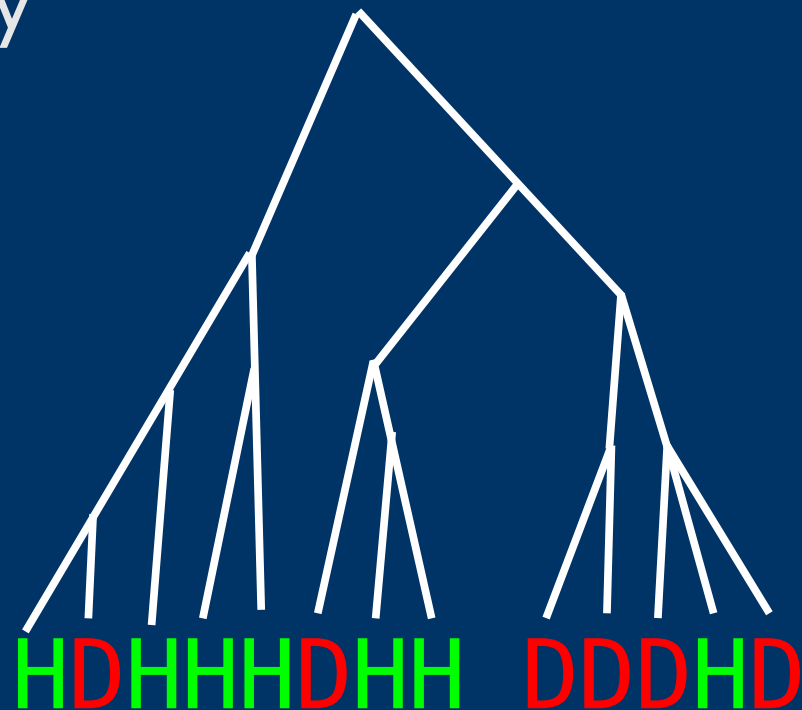
# Using the (local) genealogy of the locus

Tree at disease site resembles neighbours



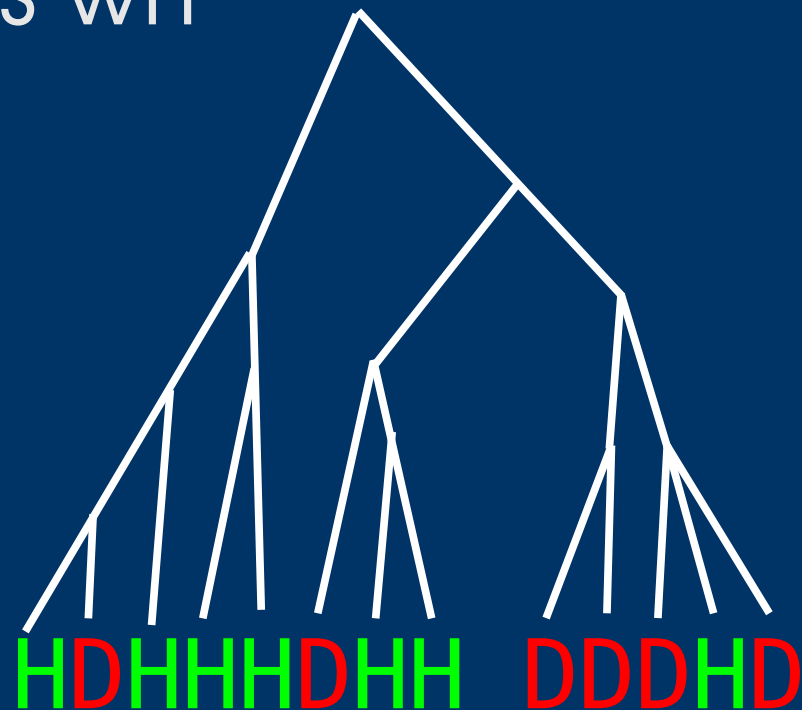
# Using the (local) genealogy of the locus

- **Near the disease site:**
  - A significant clustering of diseased/healthy



# Using the (local) genealogy of the locus

- Approach:
  - Infer trees over regions
  - Score the regions wrt their clustering



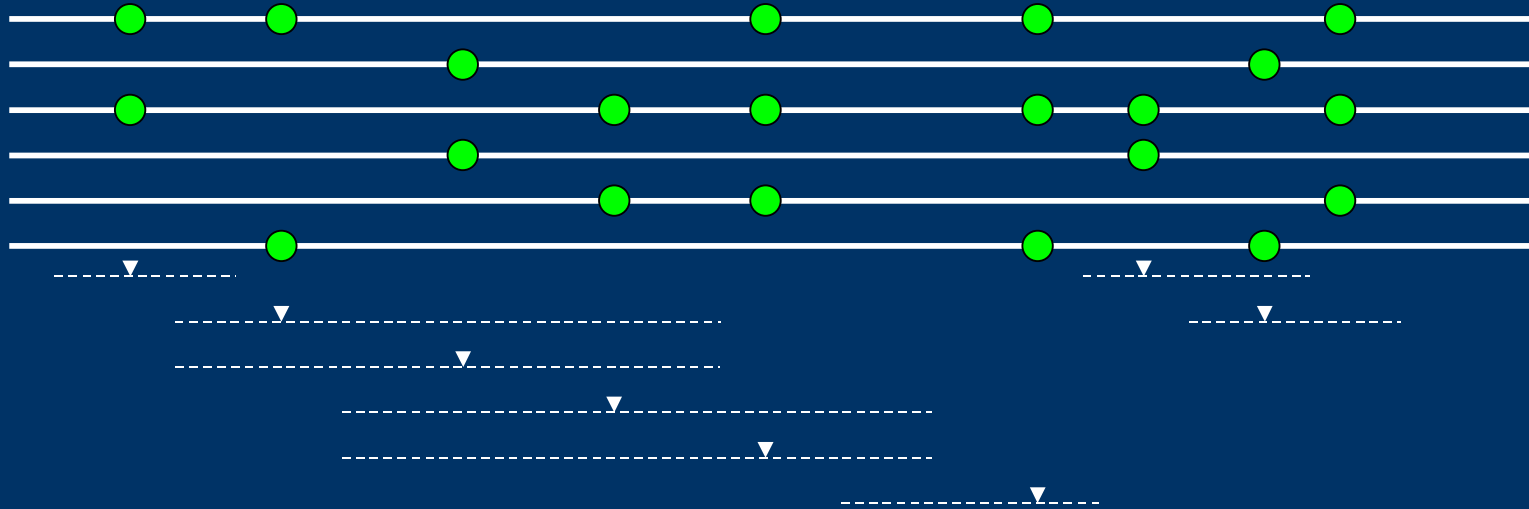
# *BLOck aSSOCiation (BLOSSOC)*

---

- In the *infinite sites* model:
  - Each mutation occurs only once
  - Each mutation splits the sample in two
  - A consistent tree can efficiently be inferred for a region without recombinations

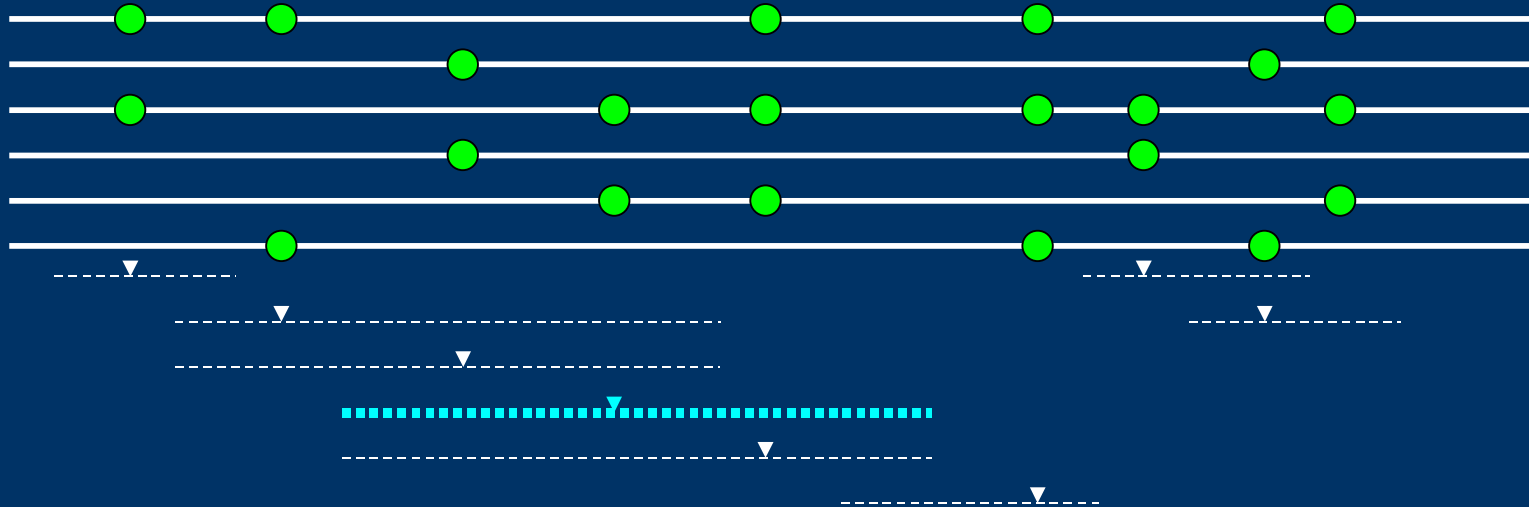
# *BLOck aSSOCiation (BLOSSOC)*

Use the four-gamete test to find regions that can be explained by a tree



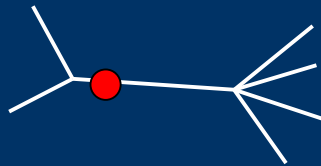
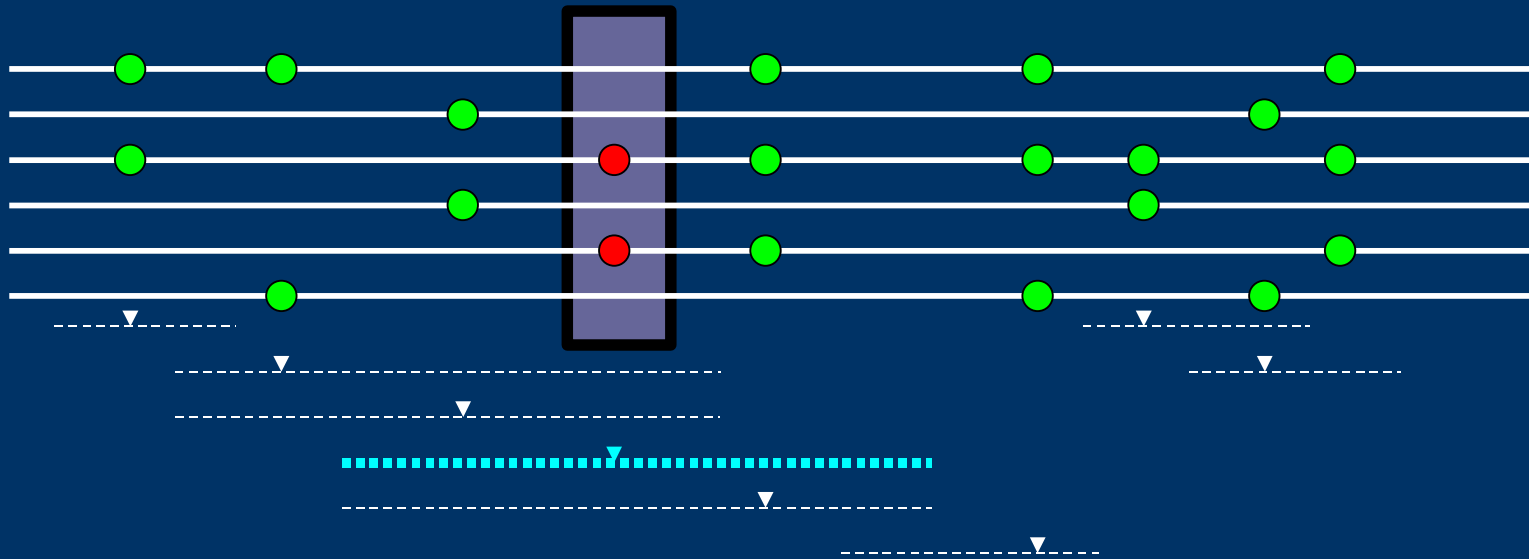
# BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



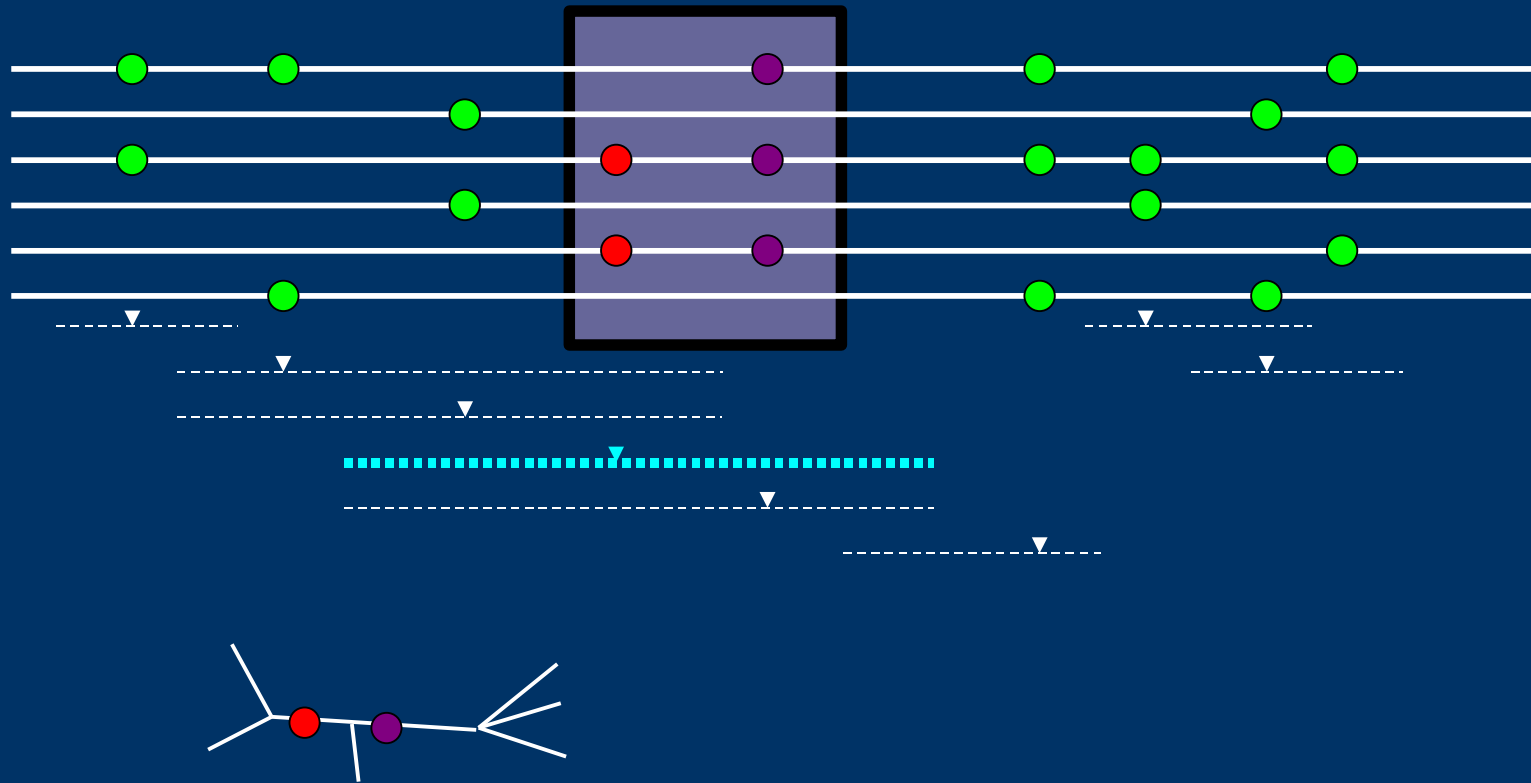
# BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



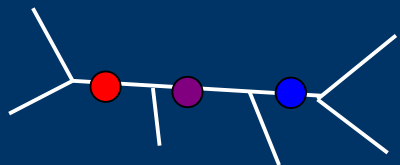
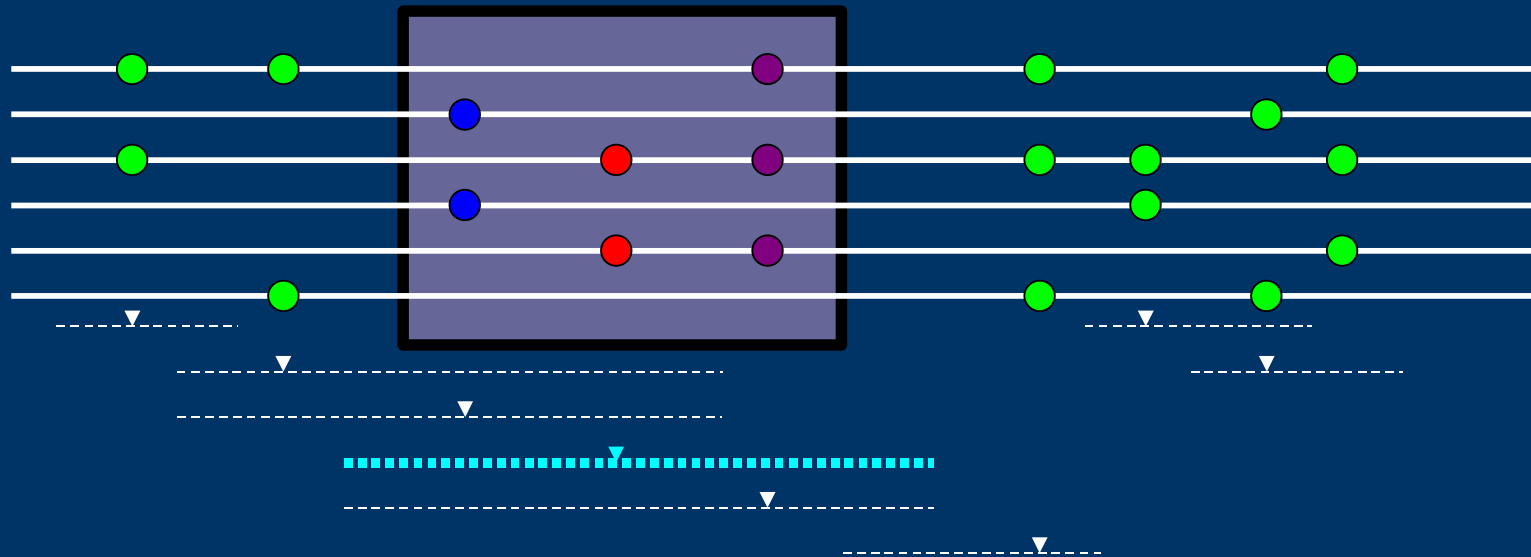
# BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



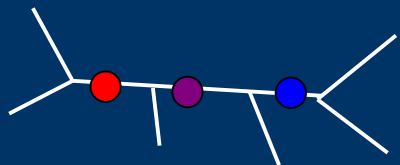
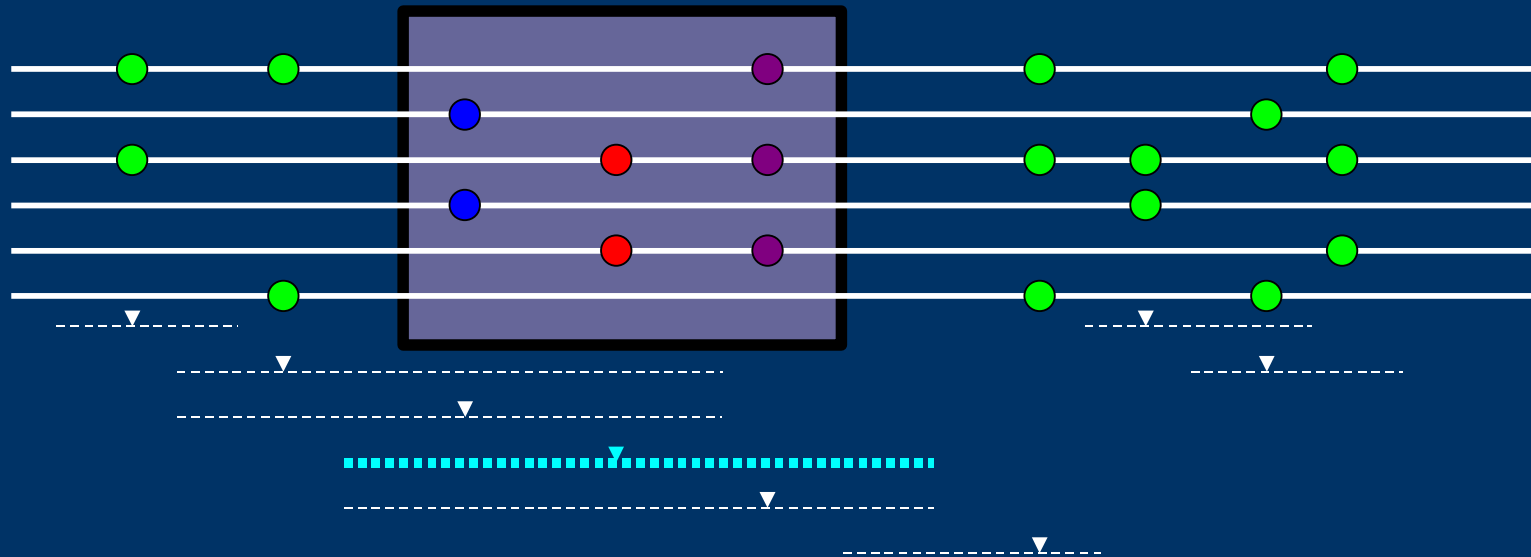
# BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



# BLOck aSSOCiation (BLOSSOC)

Score the tree, and assign the score to the region



# *BLOck aSSOCiation (BLOSSOC)*

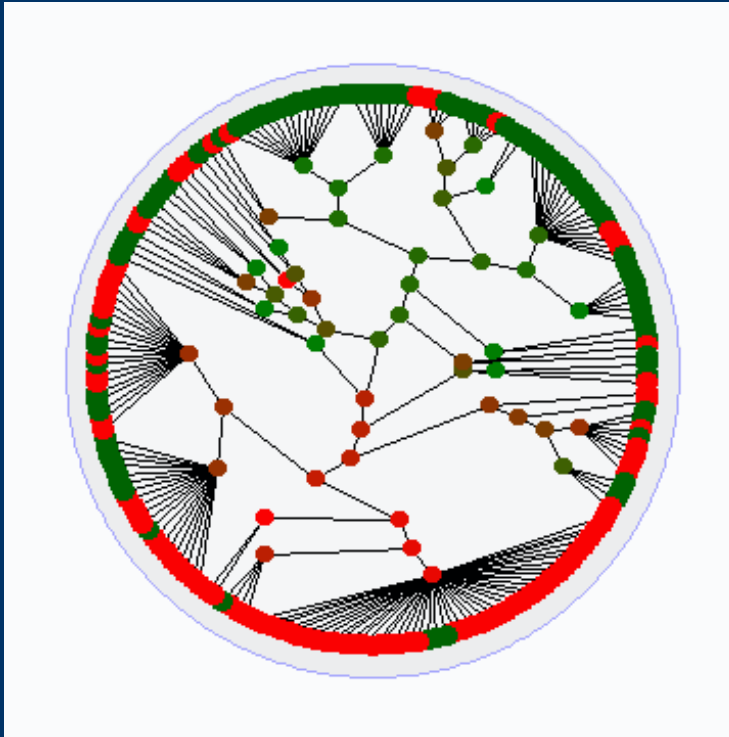
---

- Dealing with many incompatibilities:
  - Repeatedly do this for sub-samples
    - Score regions with each sample's trees
    - Score each locus with the average tree score
  - Insist on minimal number of markers
    - Skip incompatibilities

# Scoring trees...

---

---



Red=cases

Green=controls

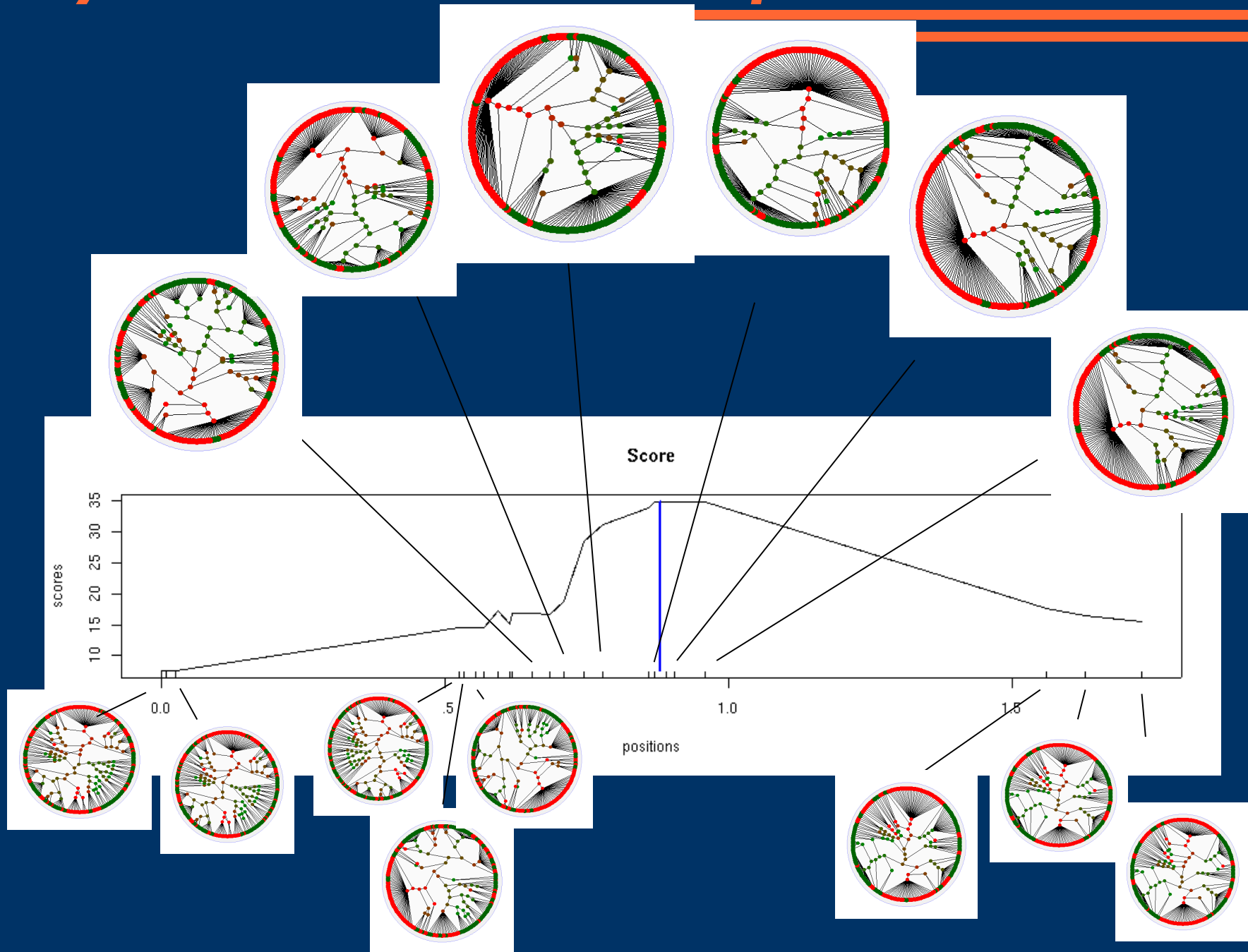
Are the case chromosomes significantly overrepresented in some clusters?

# Scoring trees...

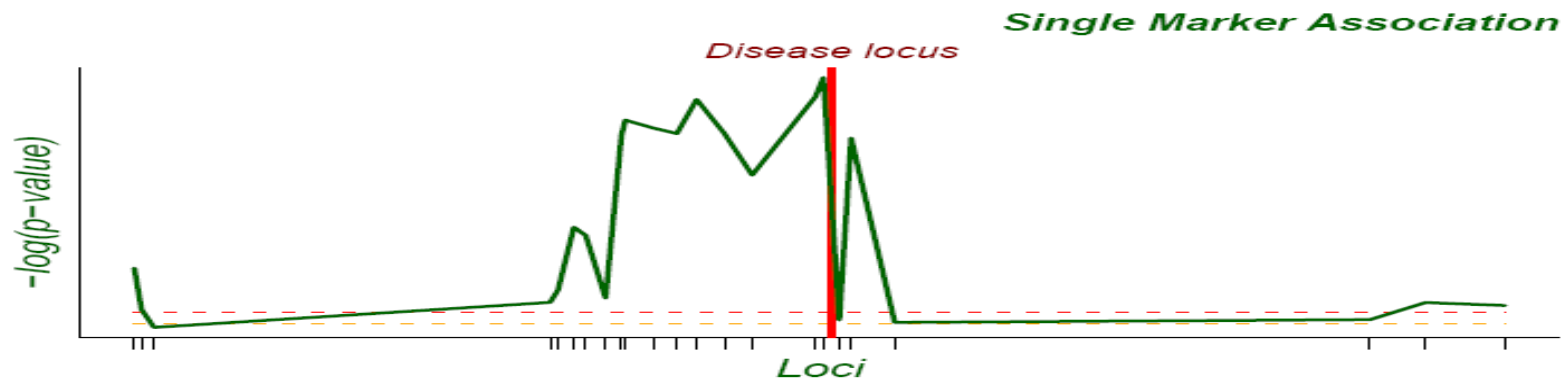
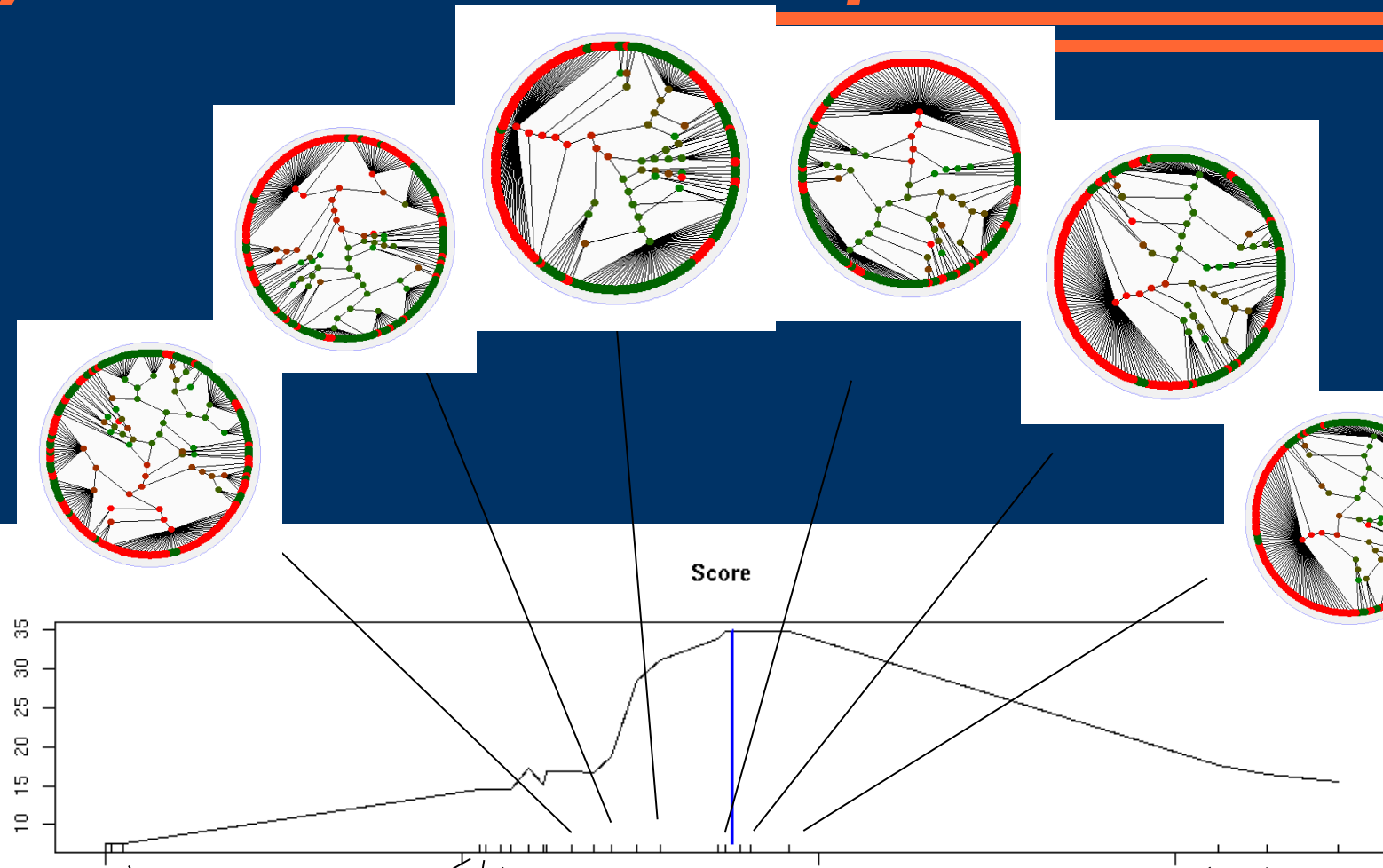
---

- AIC, BIC, HQC:
  - Score =  $-2\ln\text{Pr}(\text{data} \mid \text{model}) - K D(M)$
  - $\text{Pr}(\text{data} \mid \text{model})$  = “product of prob of a leaf, conditional on its cluster”
  - $K$  = number of clusters
  - $M$  = number of samples
  - $D(M)$  = “penalty”
    - $D(M) = 2$  (AIC),  $D(M) = \ln(M)$  (BIC),  $D(M) = \ln(\ln(M))$  (Hanna & Quinn Criteria)

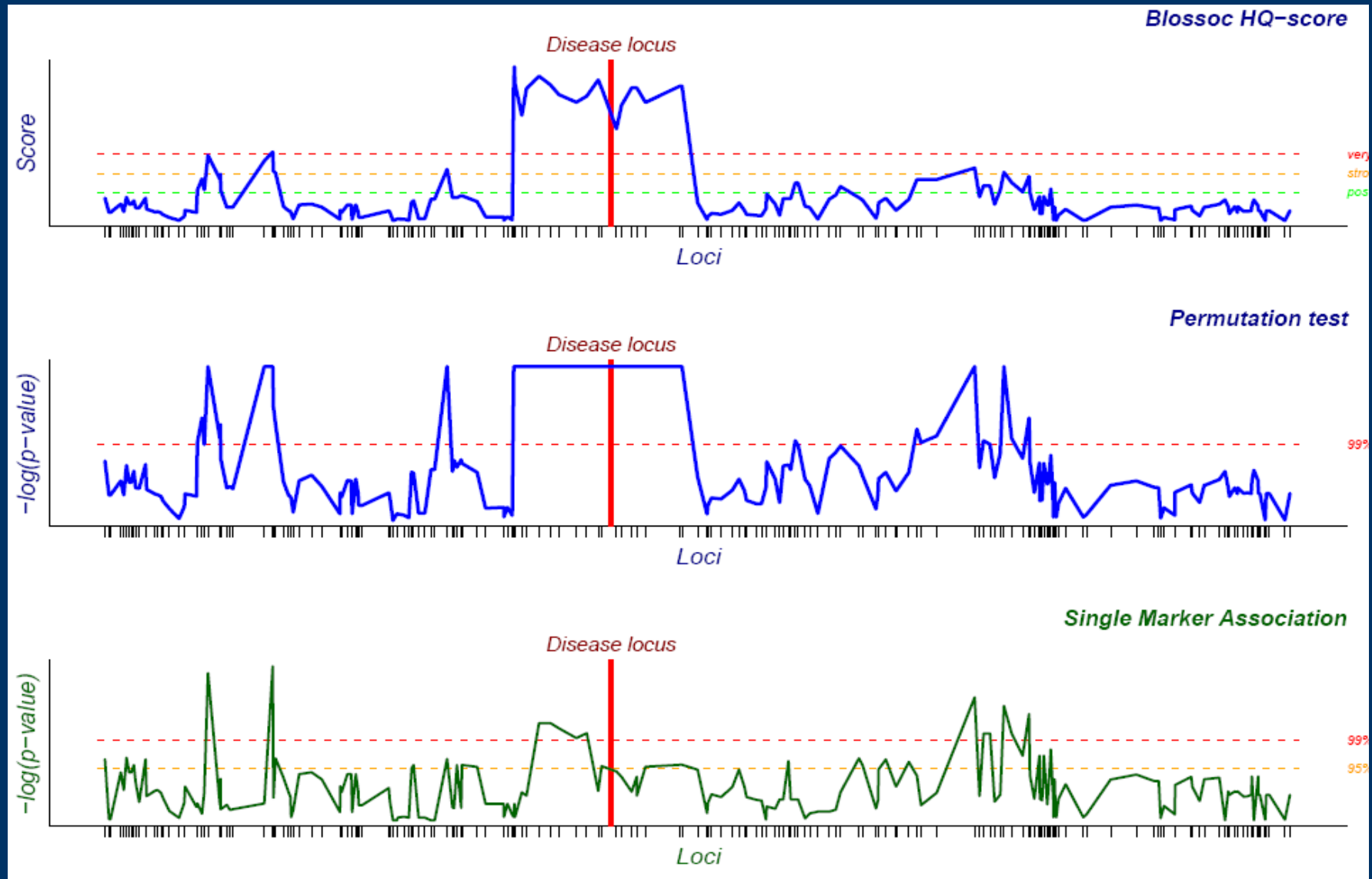
# Cystic Fibrosis example



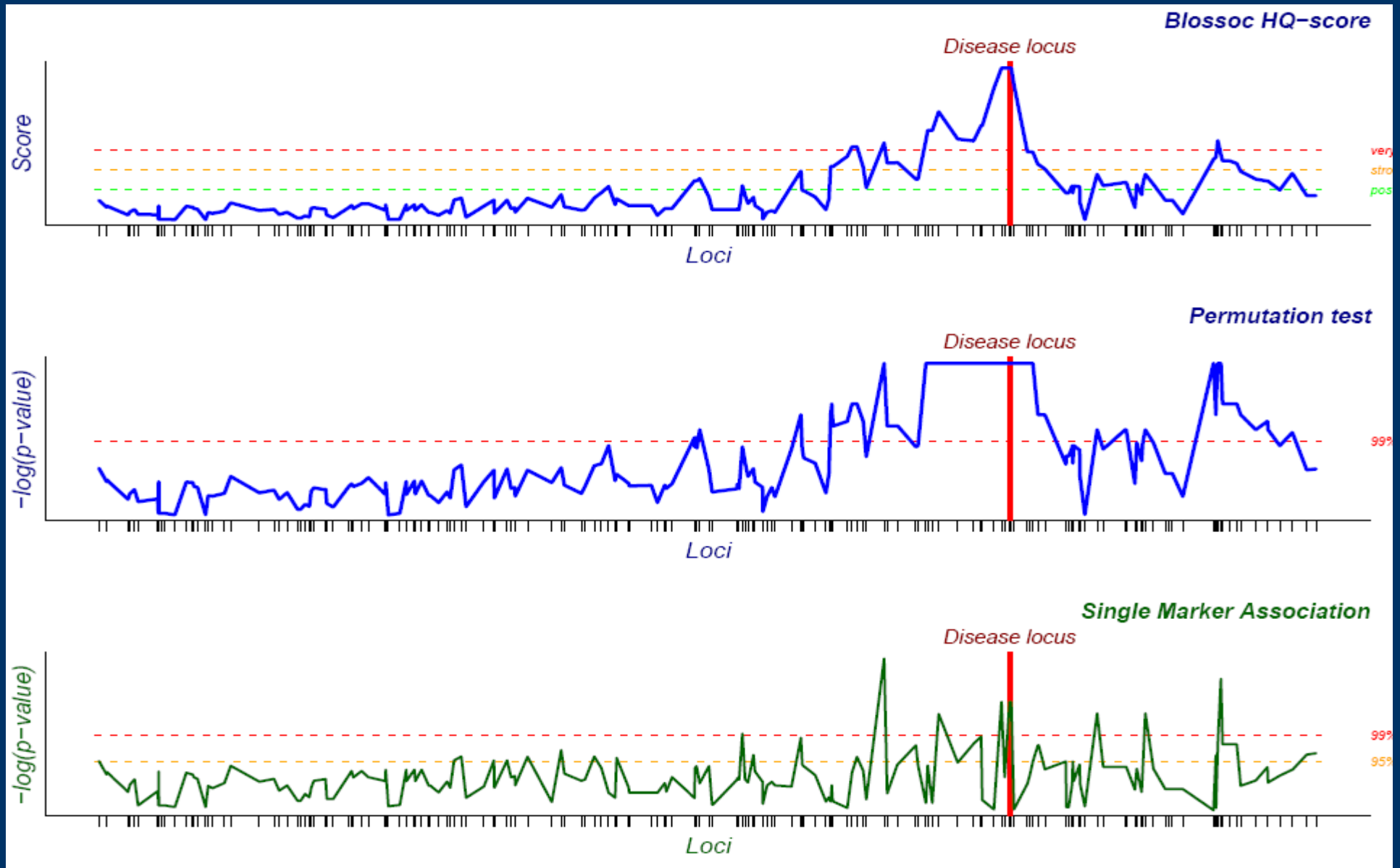
# Cystic Fibrosis example



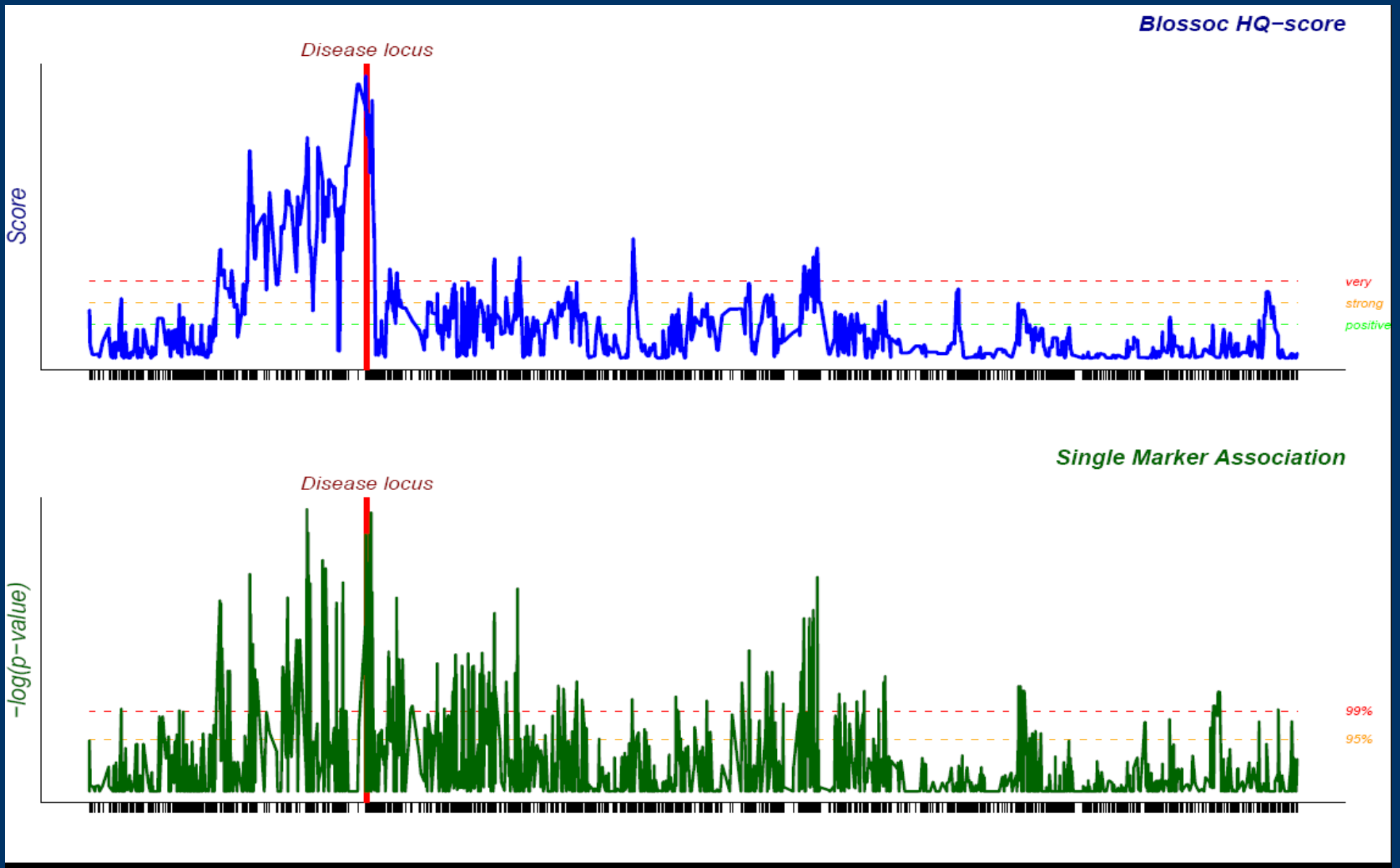
# Simulated Example (CoaSim)



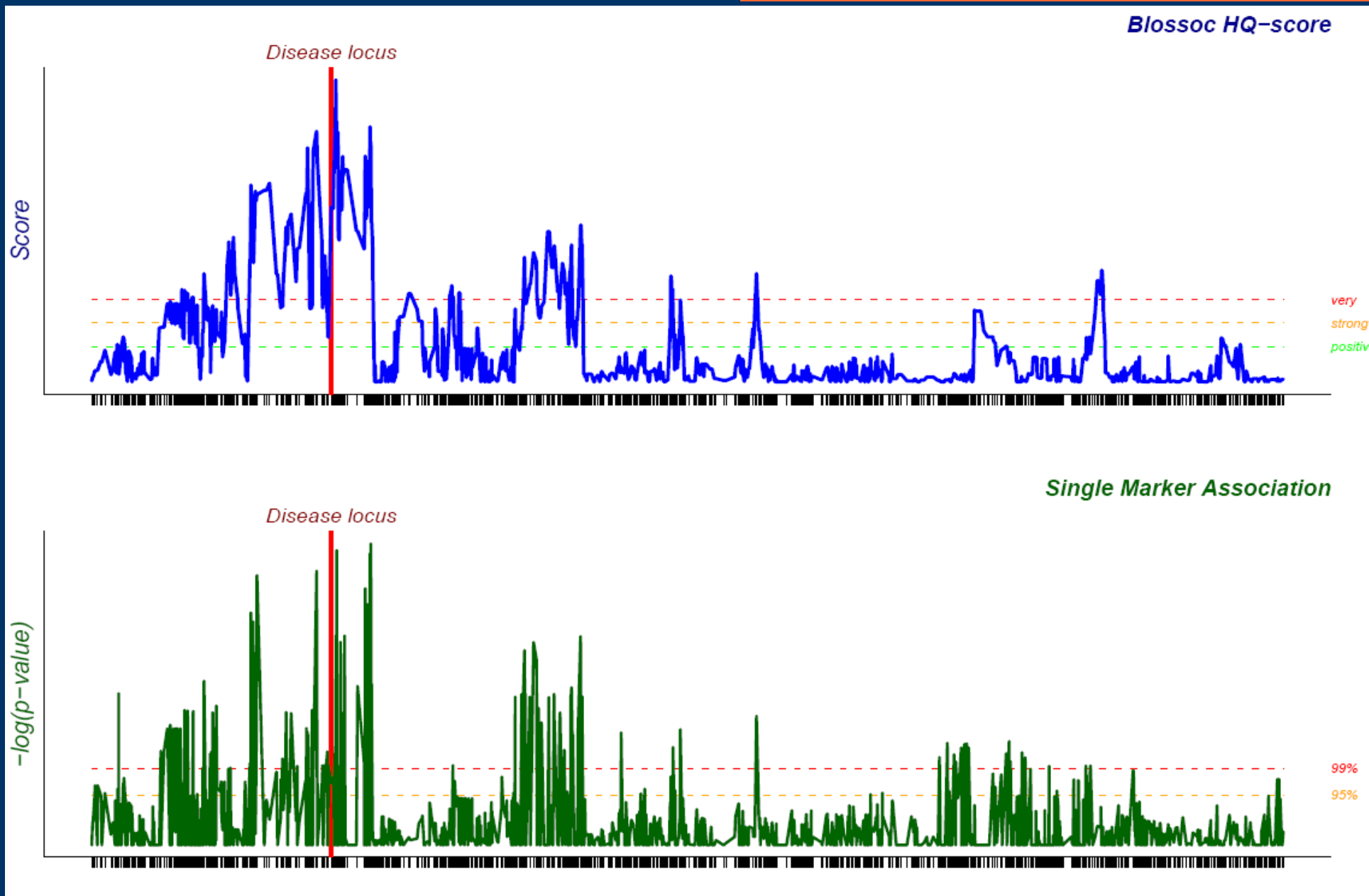
# Simulated Example (CoaSim)



# Augmented HapMap data

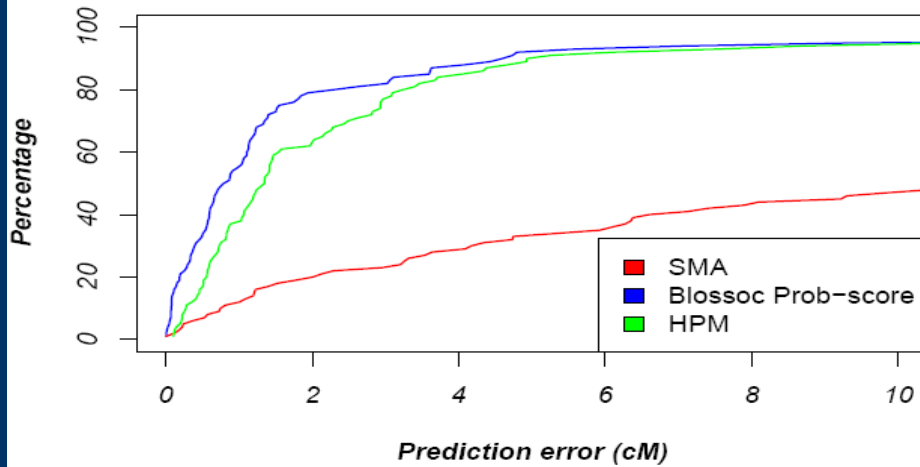


# Augmented HapMap data

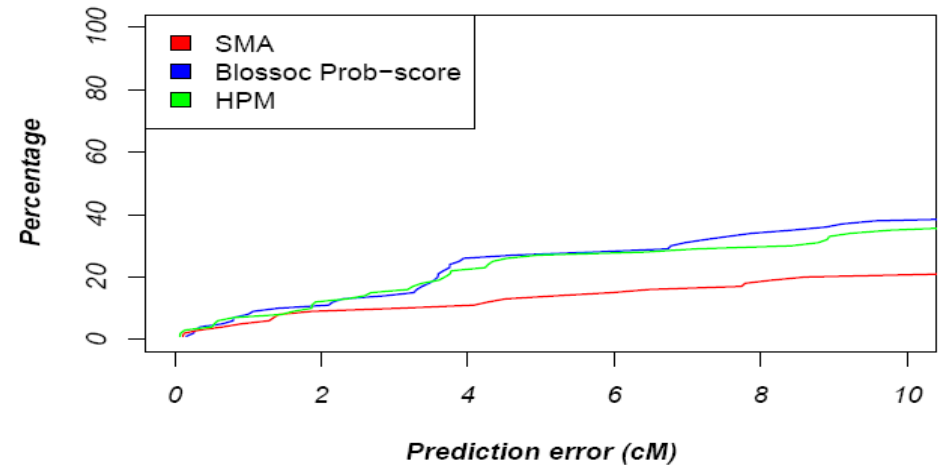


# Comparison with HPM

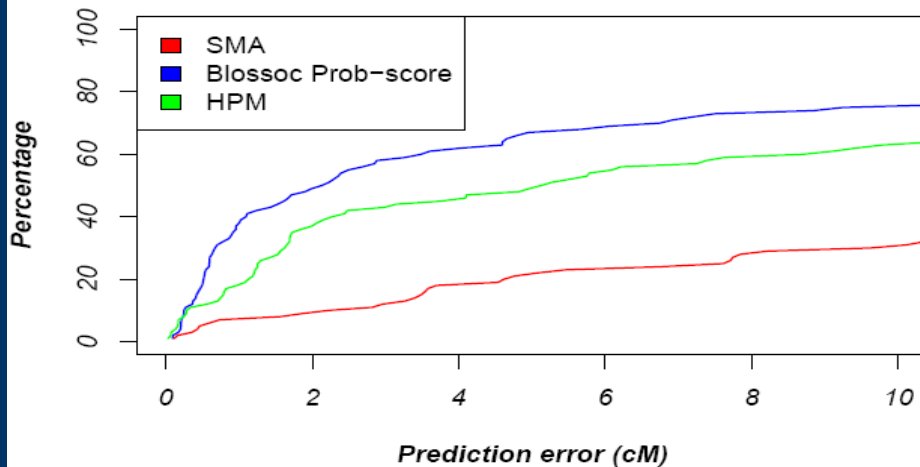
A=10



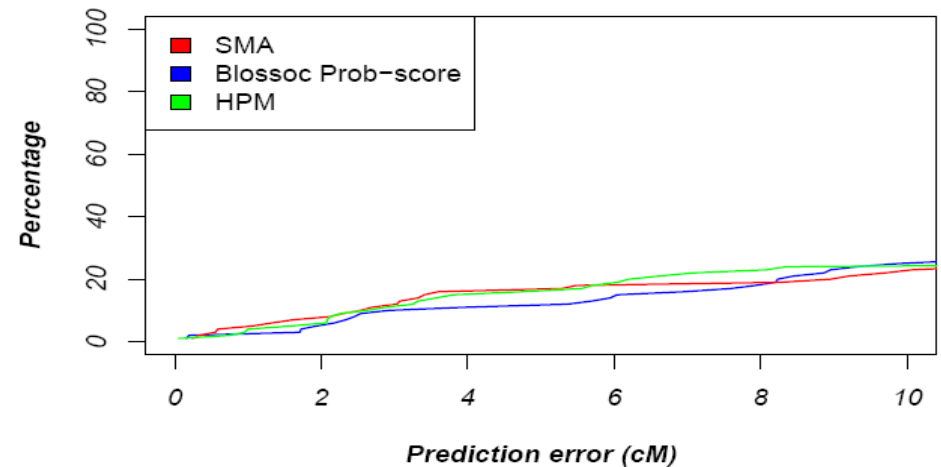
A=5



A=7.5



A=2.5



# Comparison with HapMiner and SMA

200 markers,  $\rho=40$ , 500 cases / 500 controls

MAF=0.1,  $P(A)=18-22\%$ , 1000 replicates

$P(\text{case} | AA)=15\%$

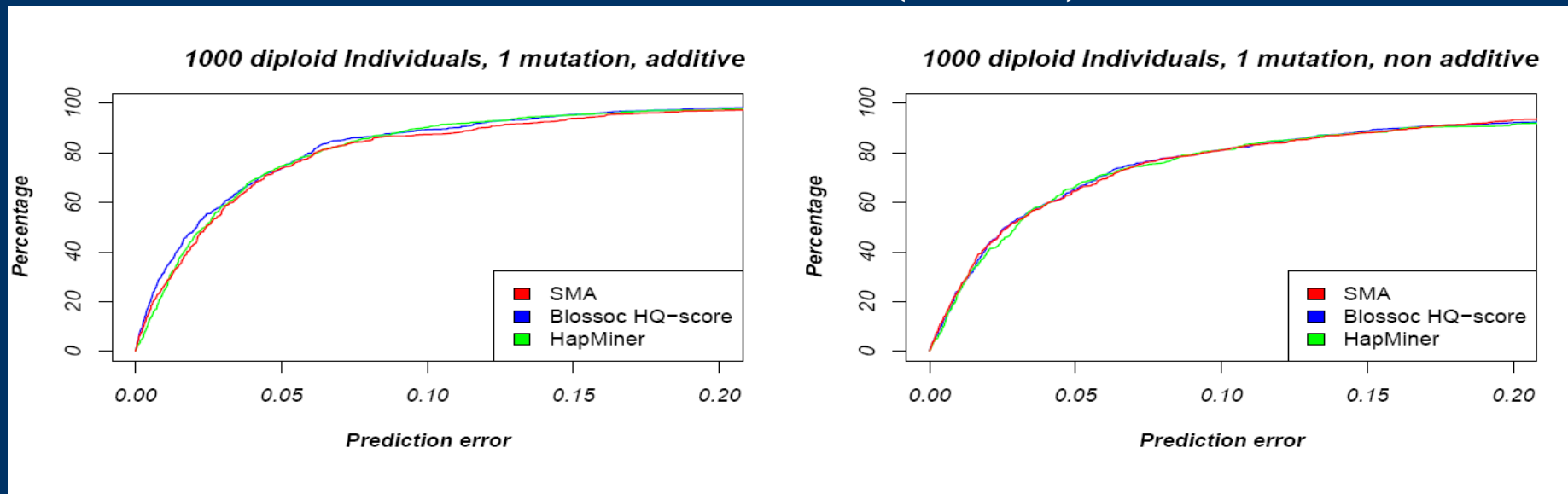
$P(\text{case} | Aa)=10\%$

$P(\text{case} | aa)=5\%$

$P(\text{case} | AA)=20\%$

$P(\text{case} | Aa)=8\%$

$P(\text{case} | aa)=5\%$



BLOSSOC = SMA = HapMiner

# Comparison with HapMiner and SMA

$P(\text{case} \mid A_x A_x) = 15\%$

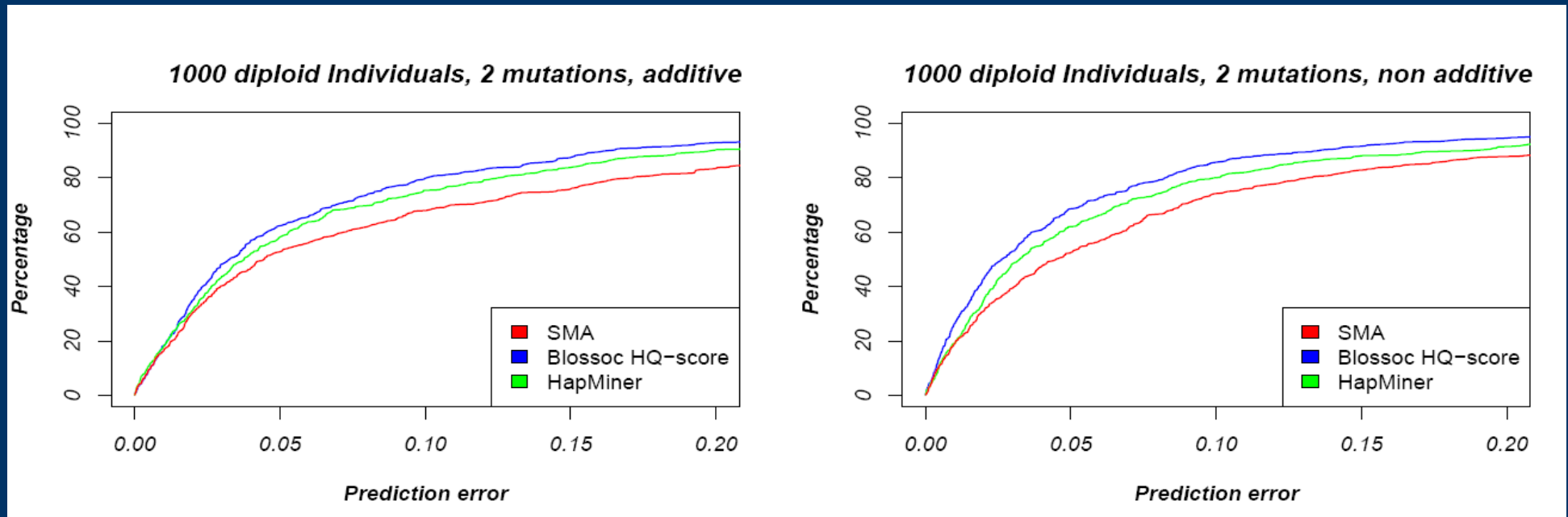
$P(\text{case} \mid A_x a) = 10\%$

$P(\text{case} \mid aa) = 5\%$

$P(\text{case} \mid A_x A_x) = 20\%$

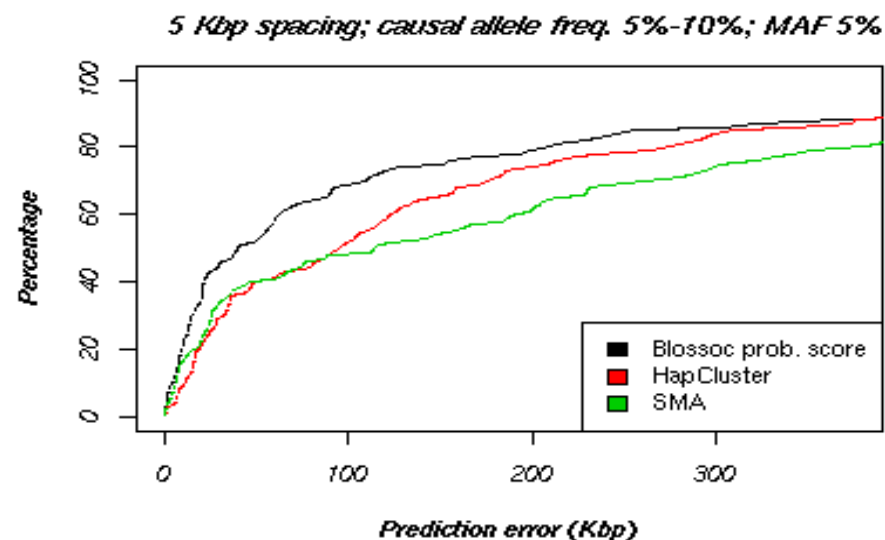
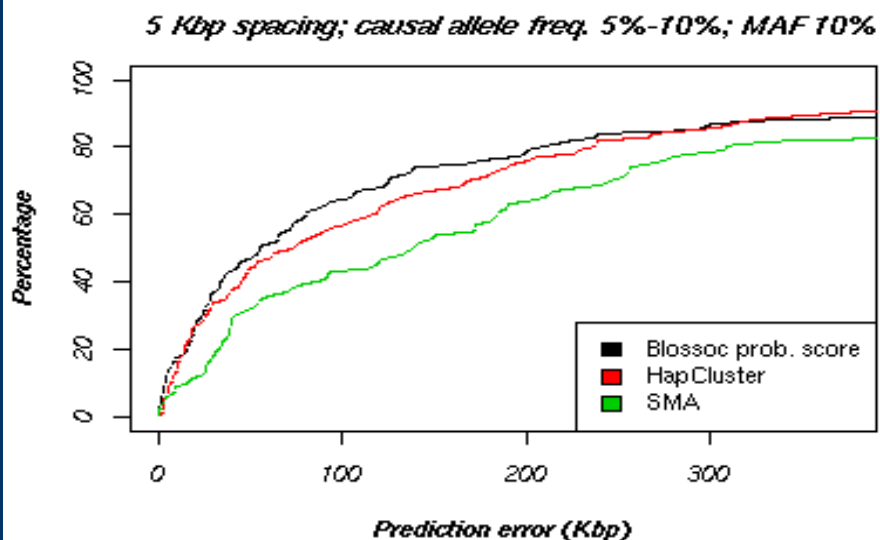
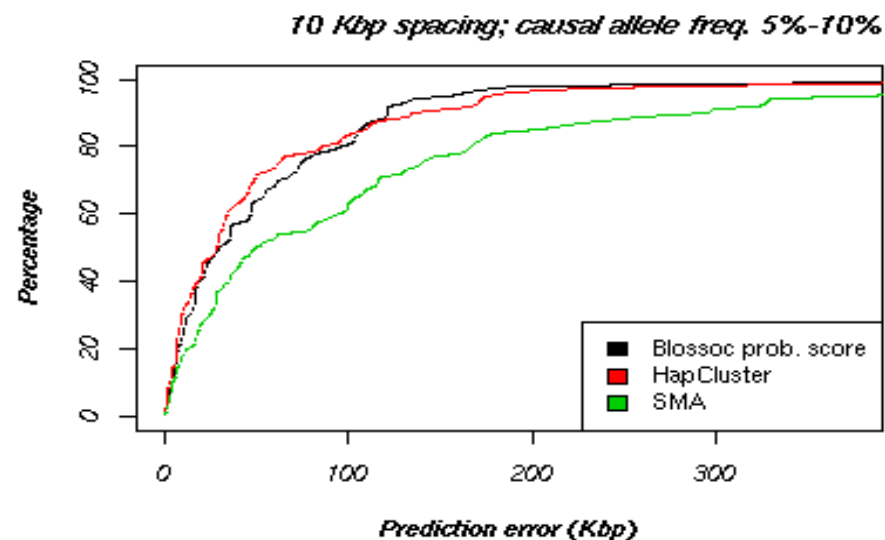
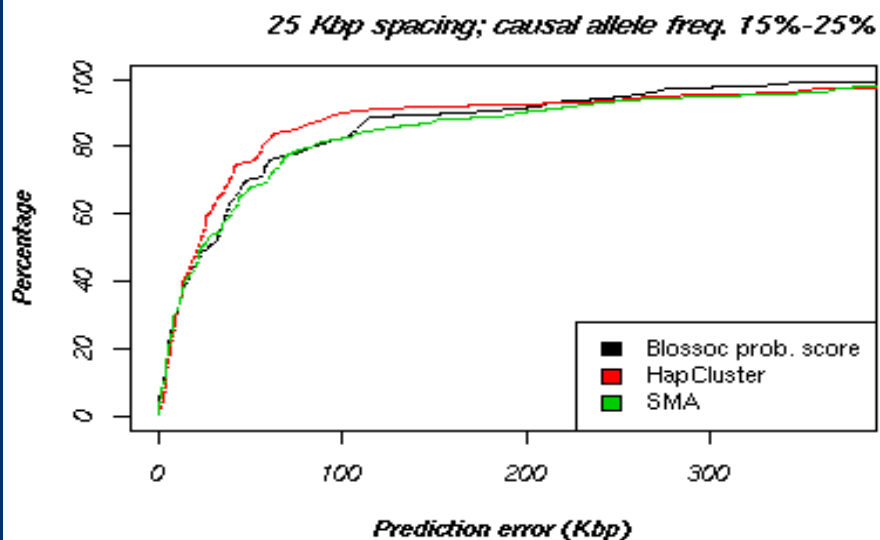
$P(\text{case} \mid A_x a) = 8\%$

$P(\text{case} \mid aa) = 5\%$

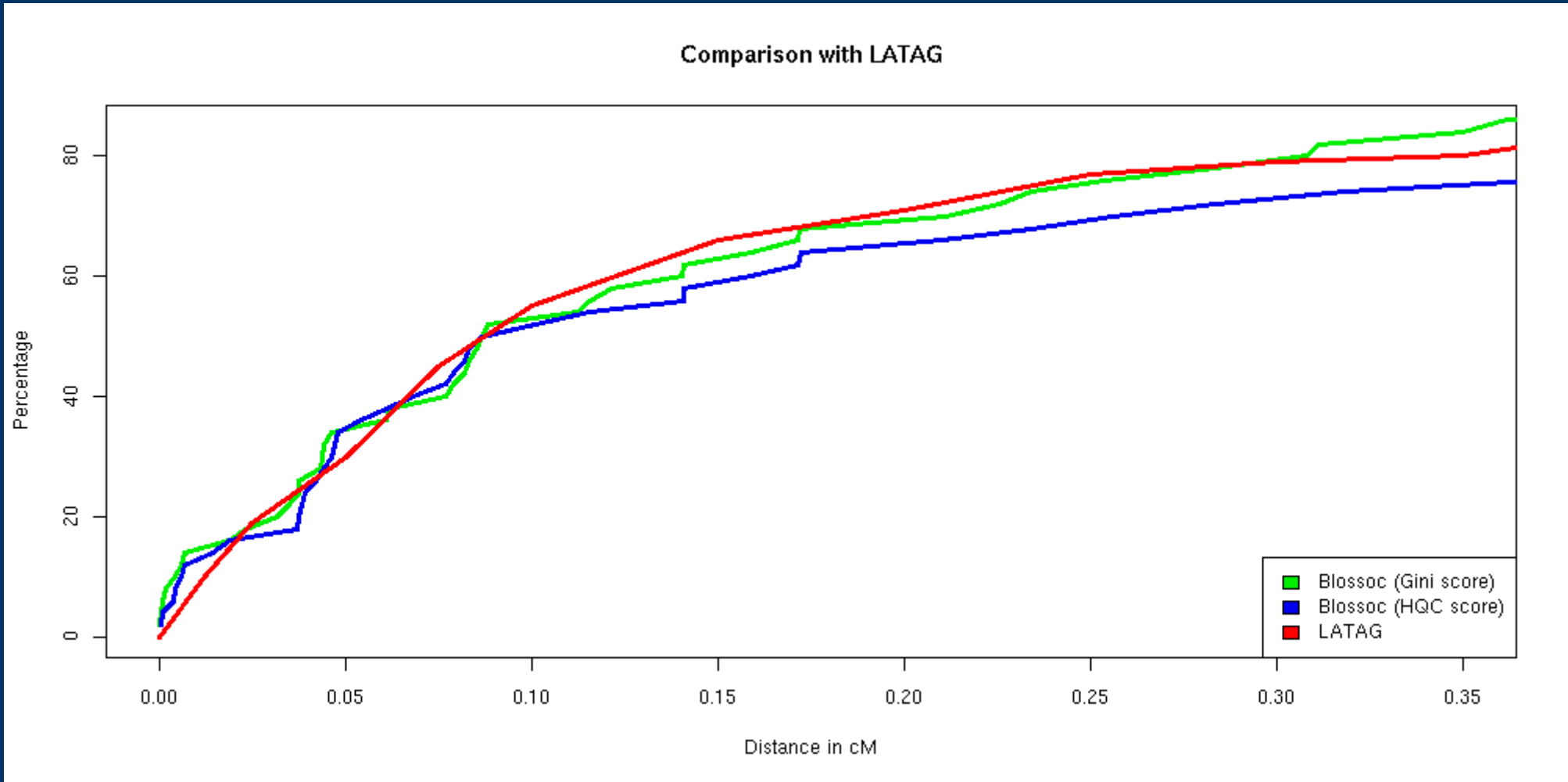


BLOSSOC > HapMiner > SMA

# Comparison with HapCluster



# Comparison with LATAG



# Speed

---

---

3 million SNP markers in 500 cases and 500 controls, i.e. 2000 haplotypes

Estimated running times on 3GHz pentium:

BLOSSOC: 36 hours

HapMiner: >300 days