

Blossoc

Fast association mapping by incompatibilities

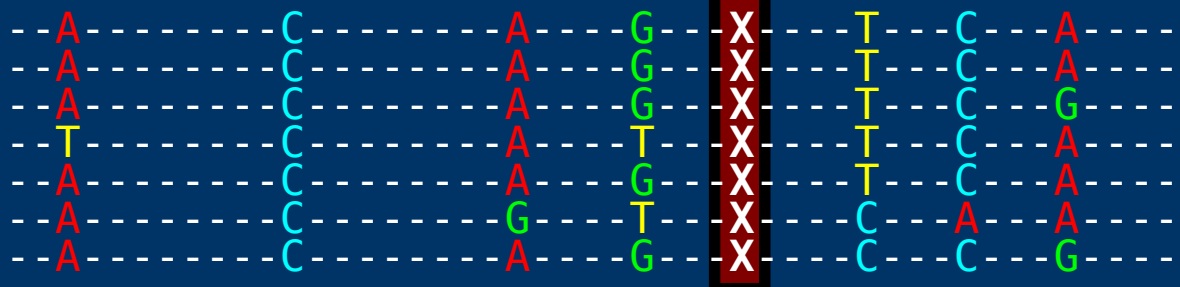
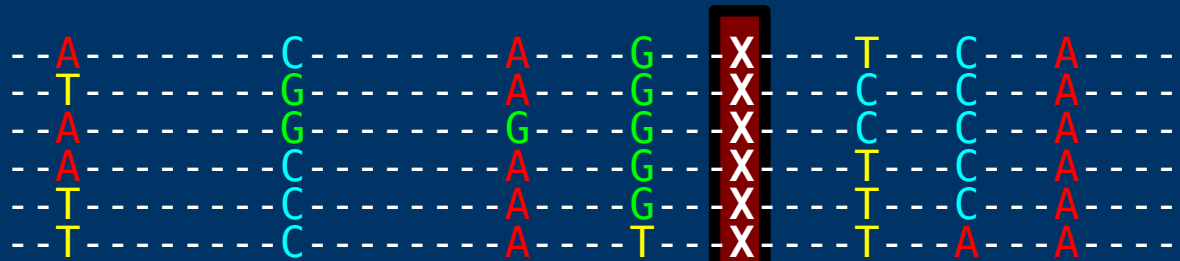
Implementation: www.daimi.au.dk/~mailund/Blossoc

*Thomas Mailund
Søren Besenbacher
Mikkel H. Schierup*

Disease mapping...

- Locate disease locus
 - Unlikely to be among our markers
 - Use information from available markers

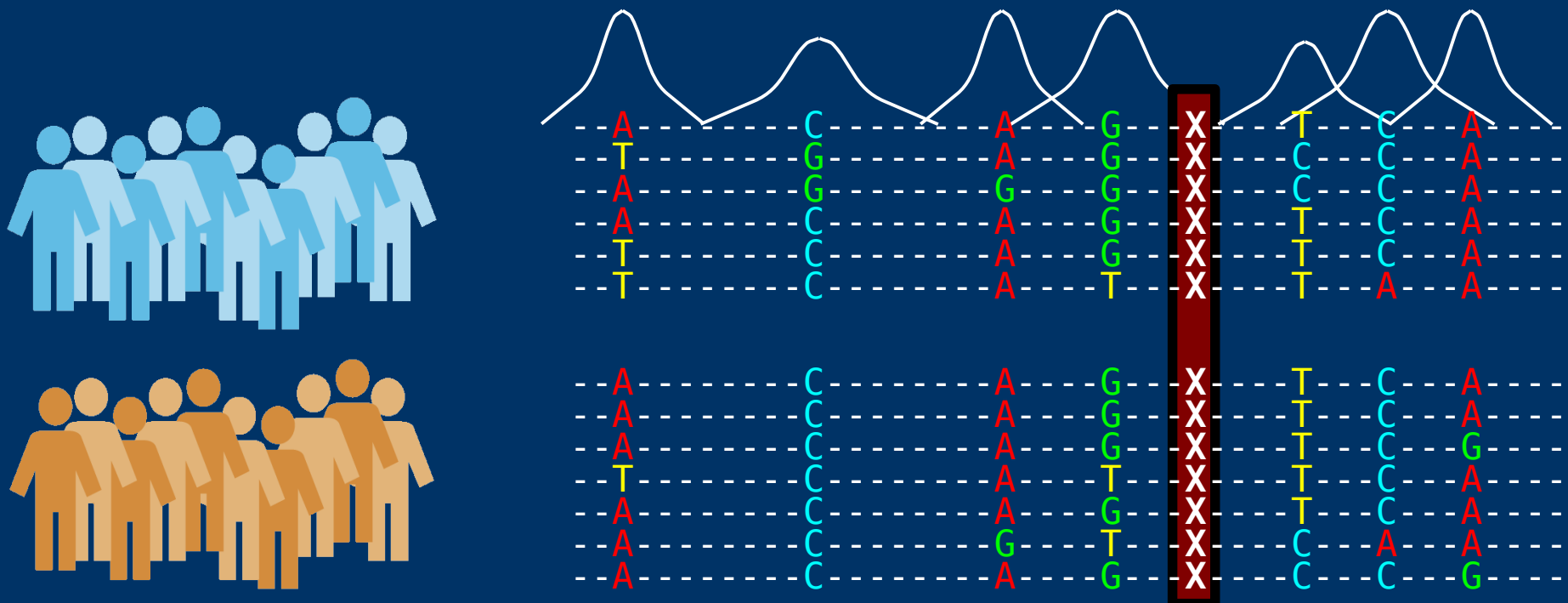
Cases (affected)



Controls (unaffected)

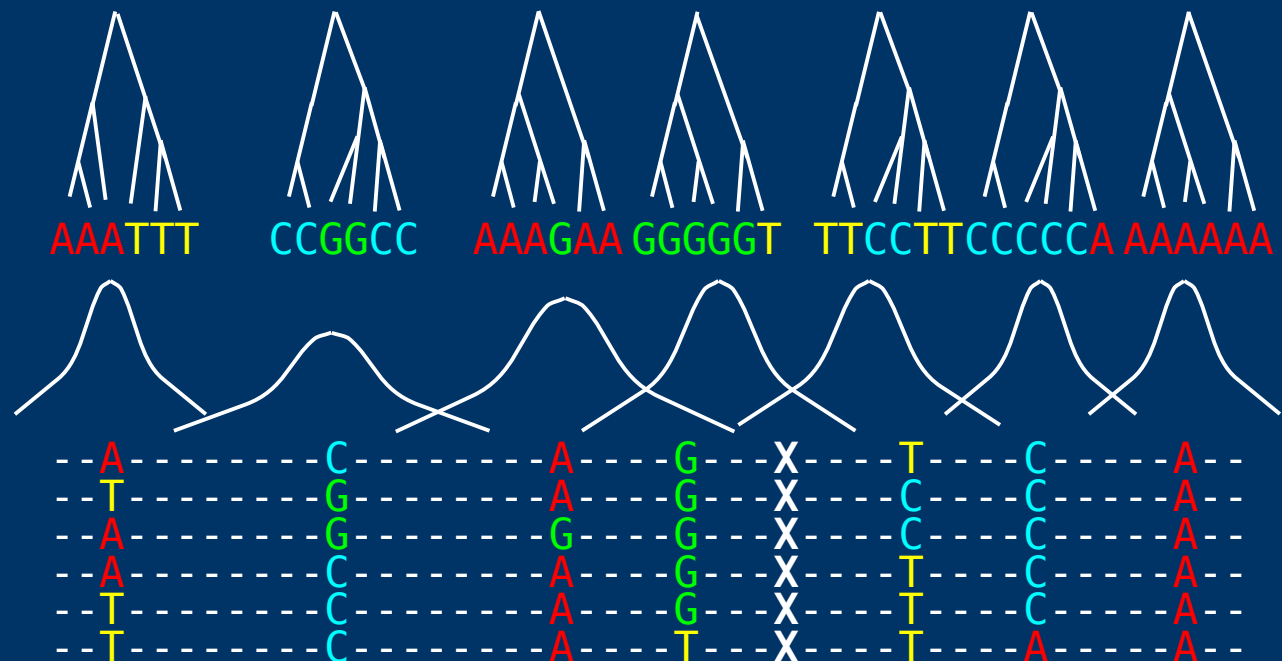
Indirect signal for causal locus

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



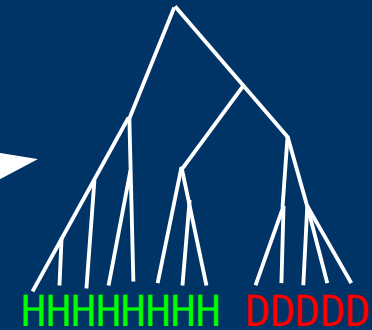
Using the (local) genealogy of the locus

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



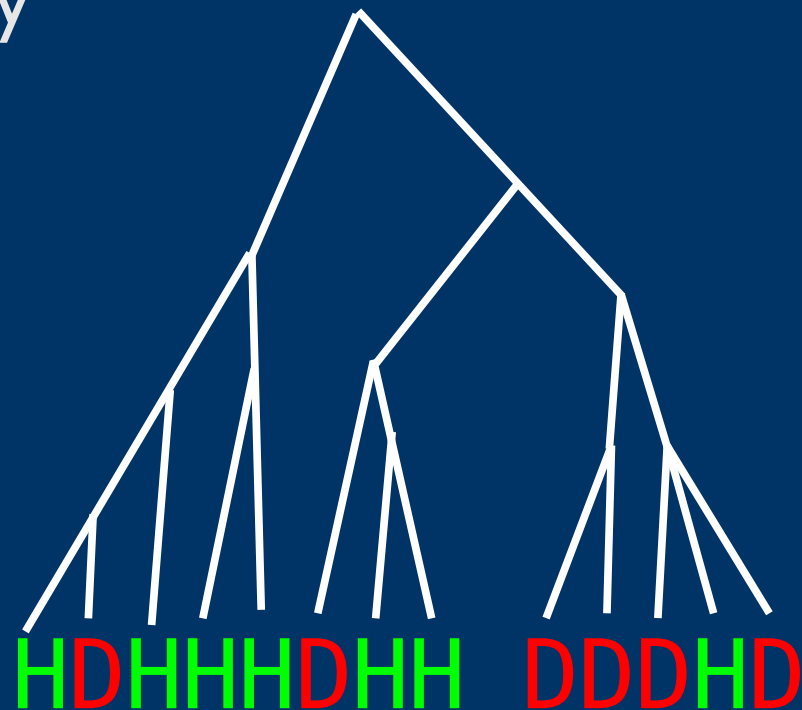
Using the (local) genealogy of the locus

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



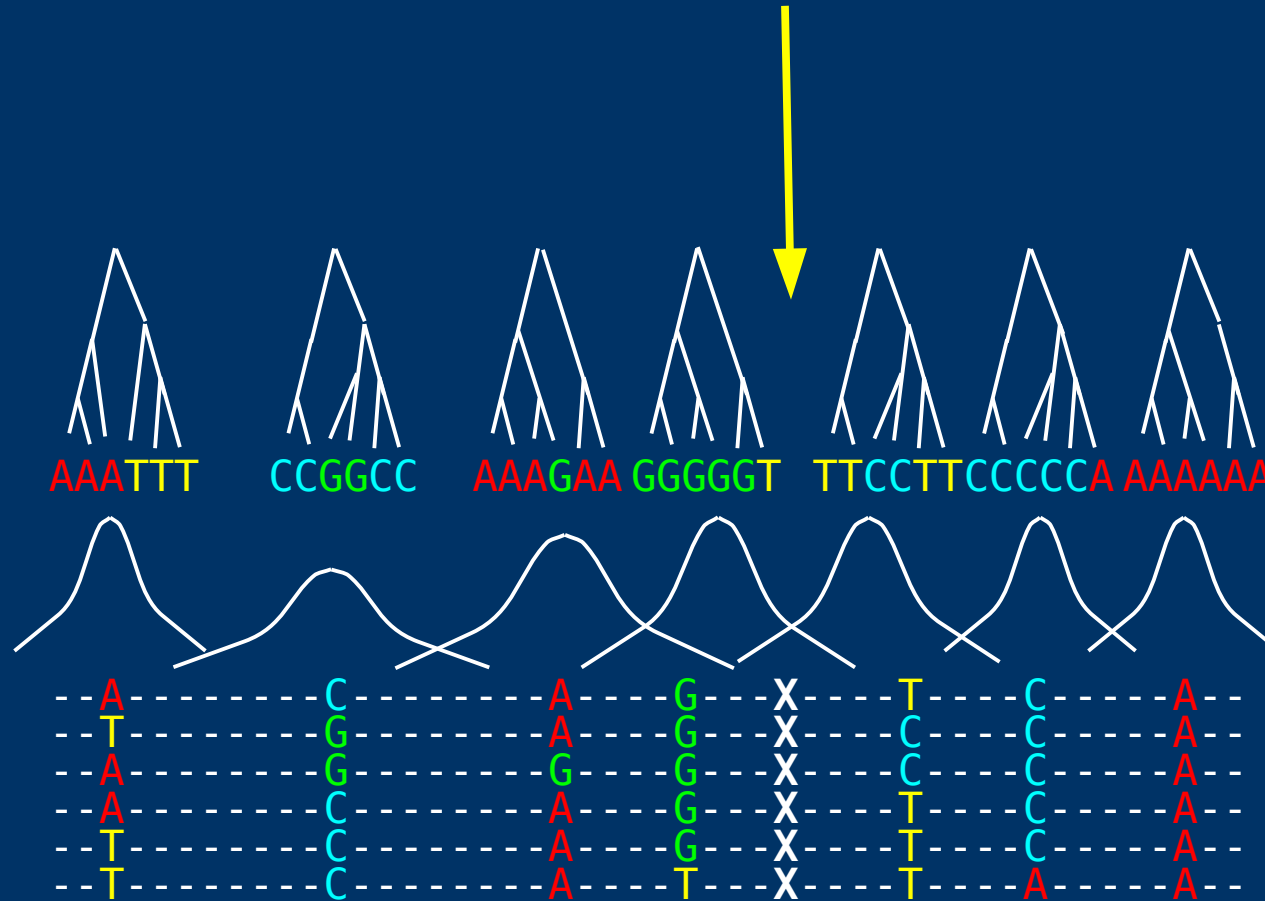
Using the (local) genealogy of the locus

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



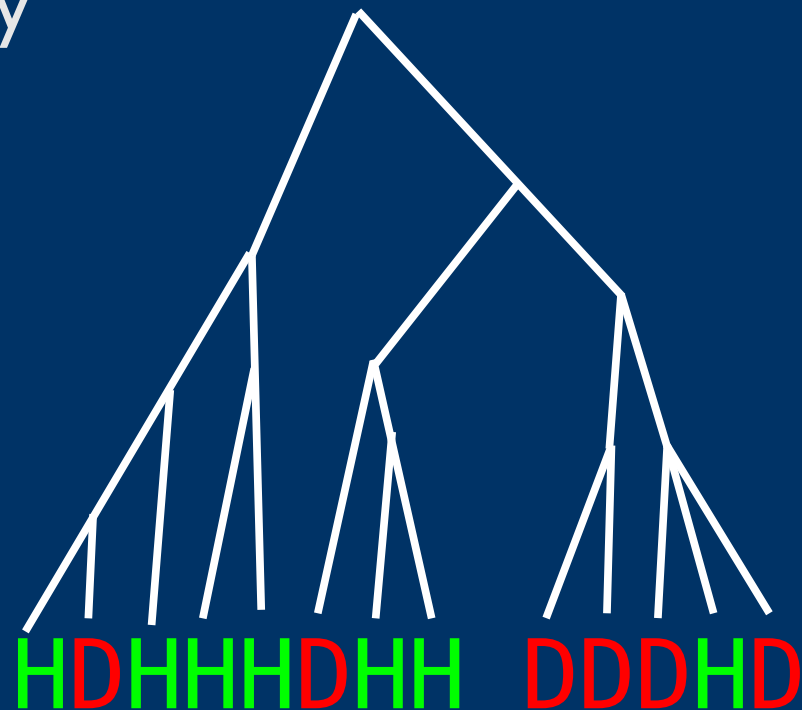
Using the (local) genealogy of the locus

Tree at disease site resembles neighbours



Using the (local) genealogy of the locus

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

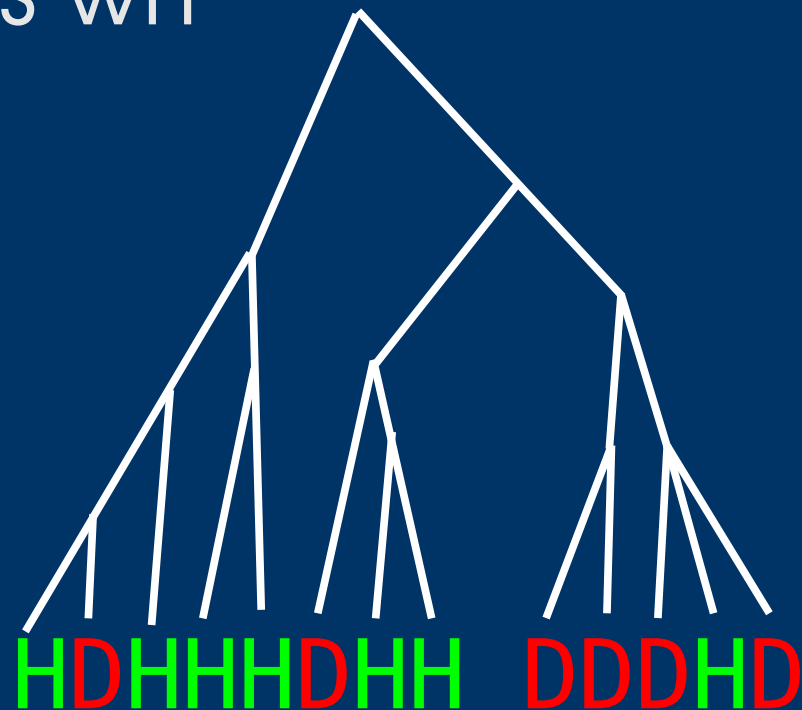


Templeton et al 1987

Zöllner&Pritchard 2005

Using the (local) genealogy of the locus

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

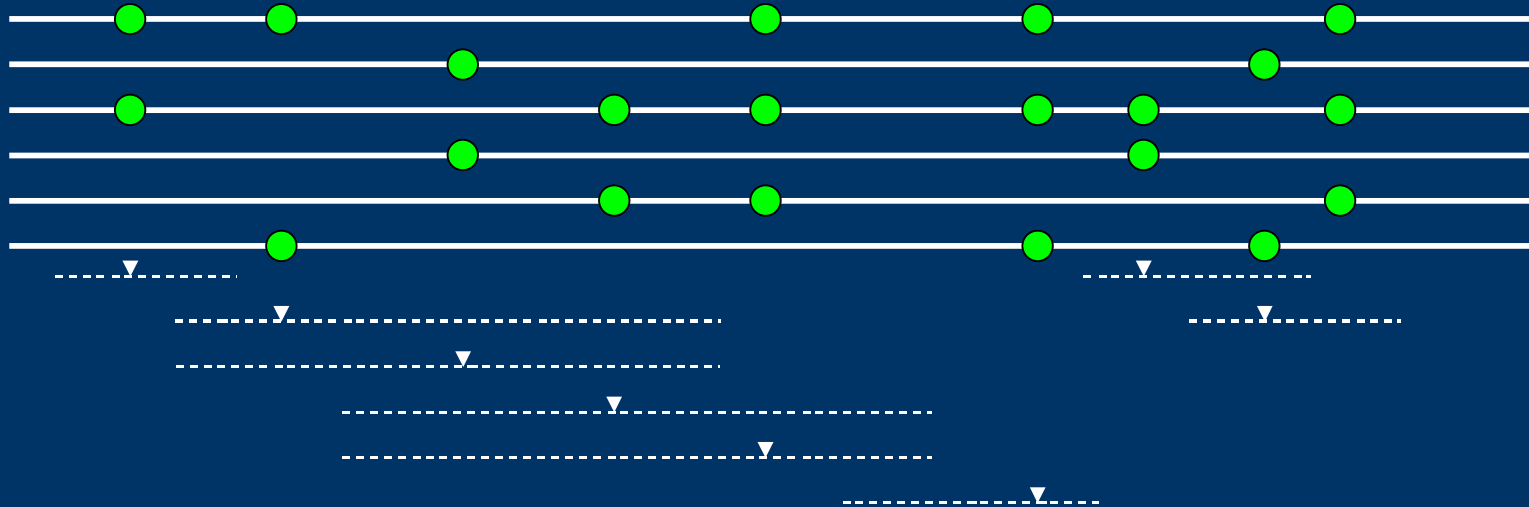


BLOck aSSOCiation (BLOSSOC)

- 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

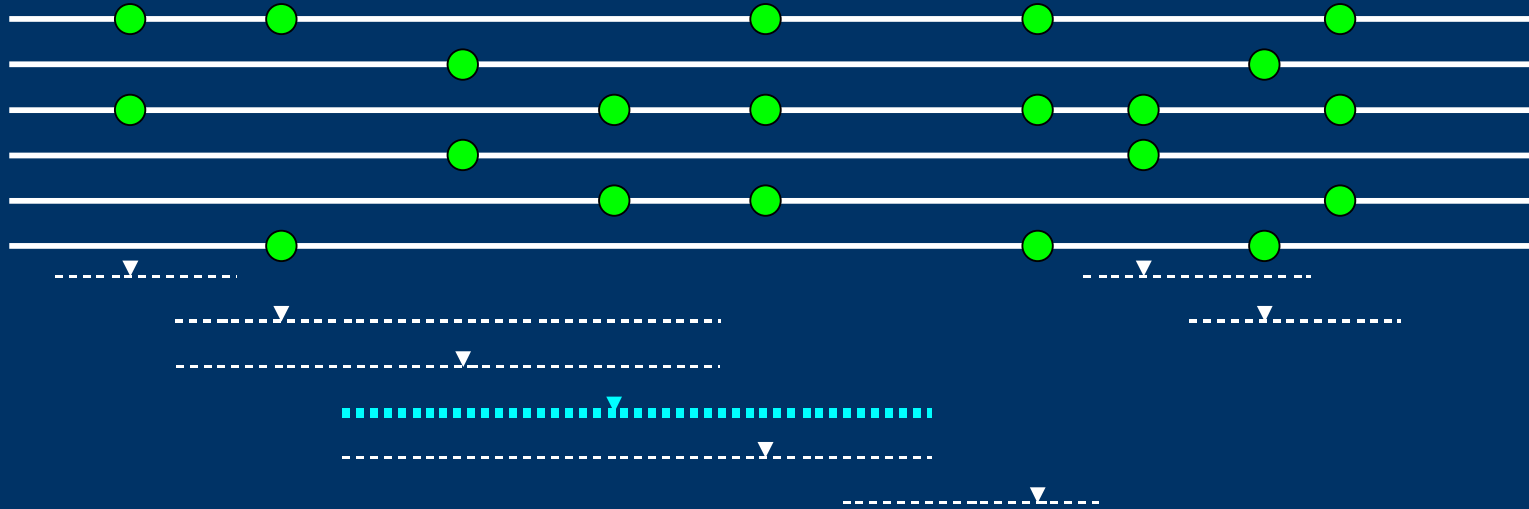
BLOck aSSOCiation (BLOSSOC)

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



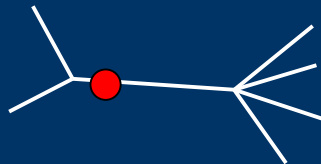
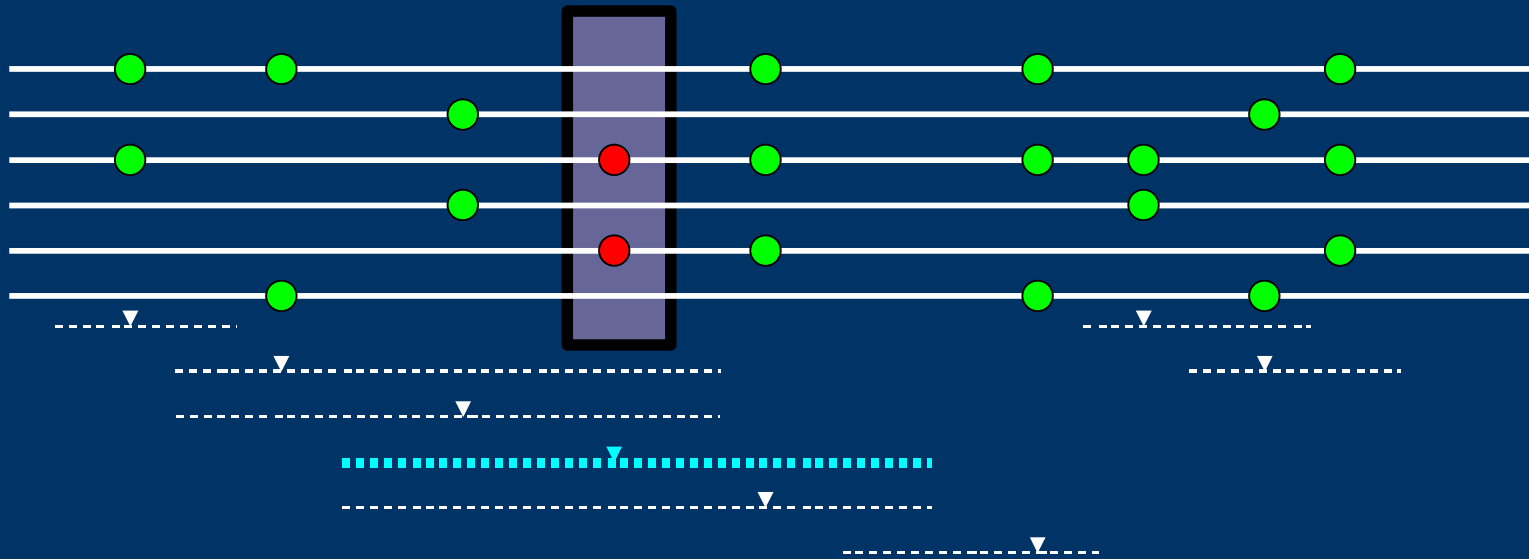
BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



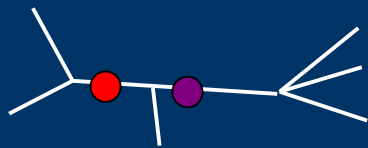
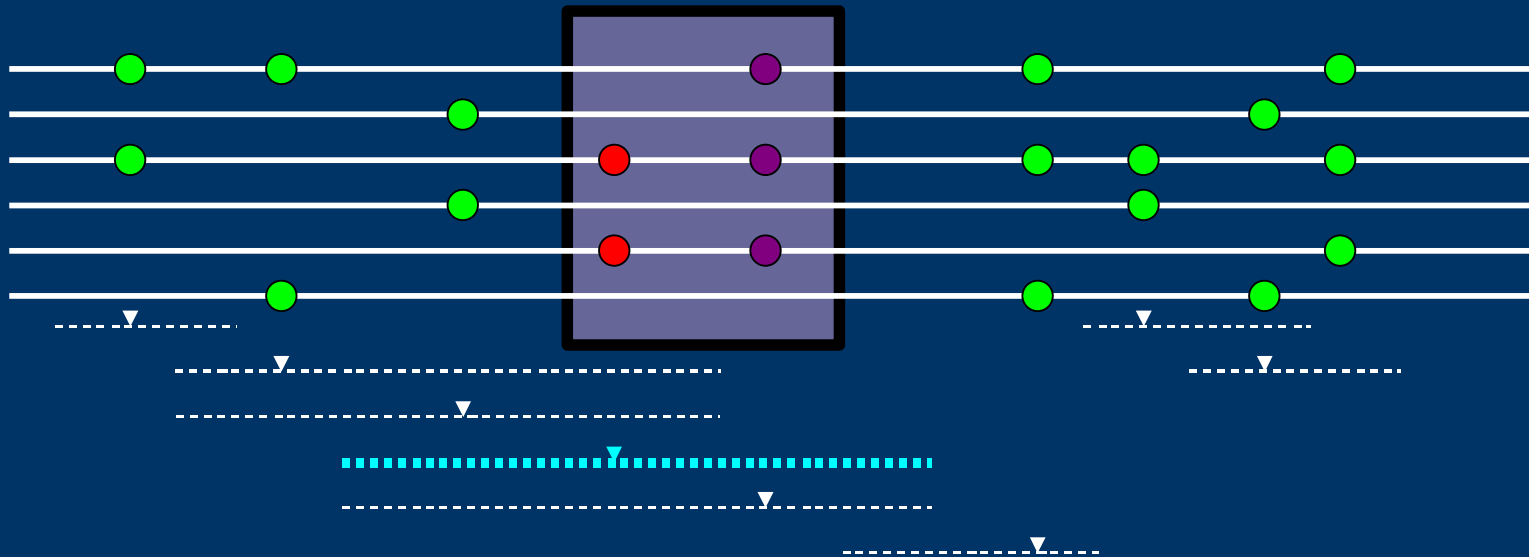
BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



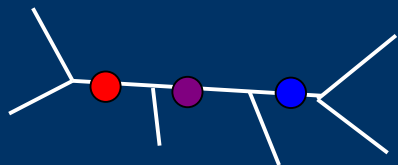
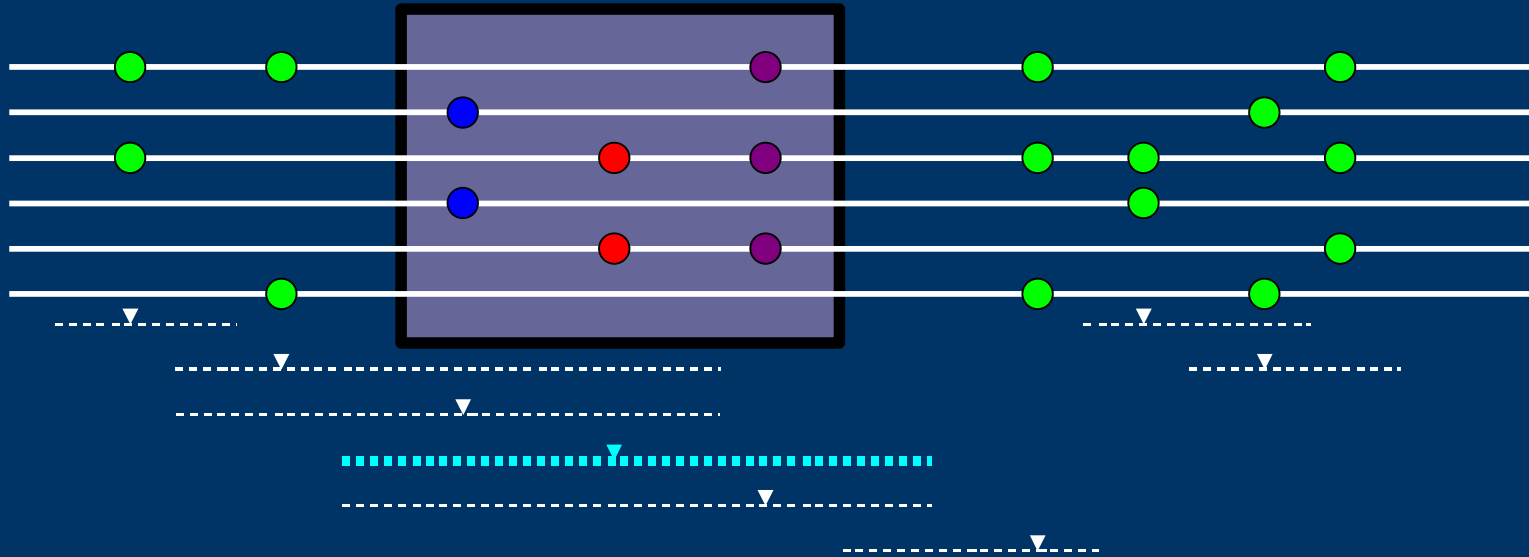
BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



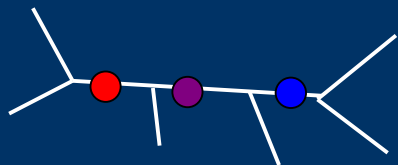
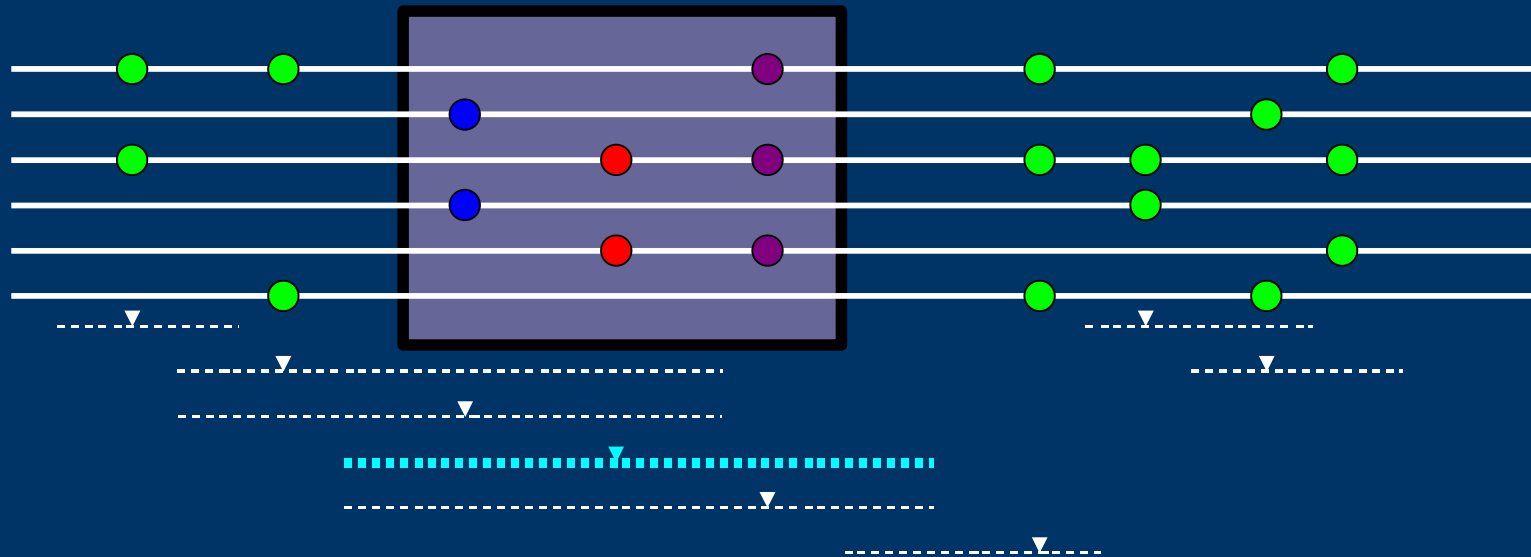
BLOck aSSOCiation (BLOSSOC)

Build a tree for each such region



BLOck aSSOCiation (BLOSSOC)

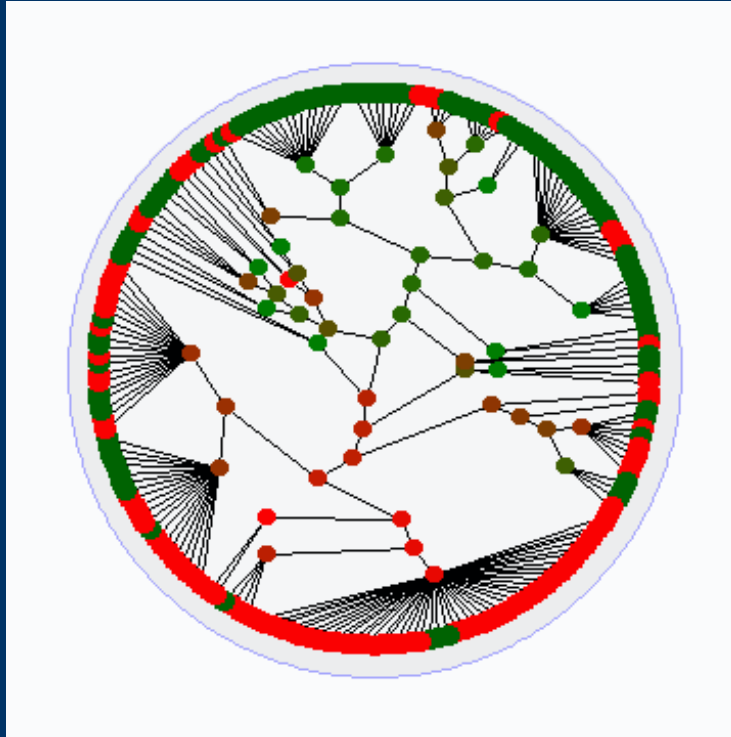
Score the tree, and assign the score to the region



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...

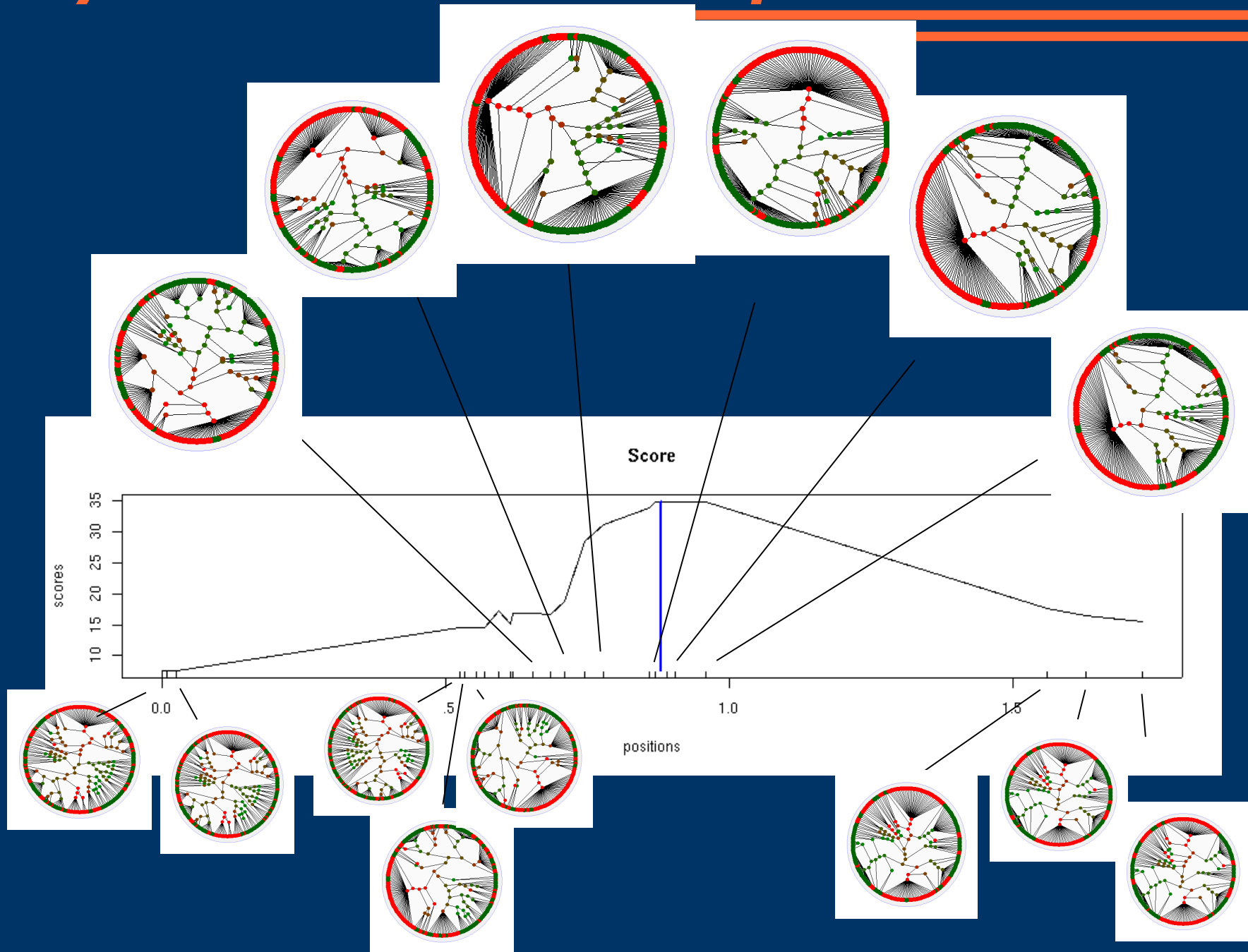
- AIC, BIC, HQC:
 - Score = $-2\ln\text{Pr}(\text{data} \mid \text{model}) - K D(M)$
 - $\text{Pr}(\text{data} \mid \text{model})$ = “product of prob of a leaf, conditional on its cluster”
 - K = number of clusters
 - M = number of samples
 - $D(M)$ = “penalty”
 - $D(M) = 2$ (AIC), $D(M) = \ln(M)$ (BIC),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

Scoring a cluster...

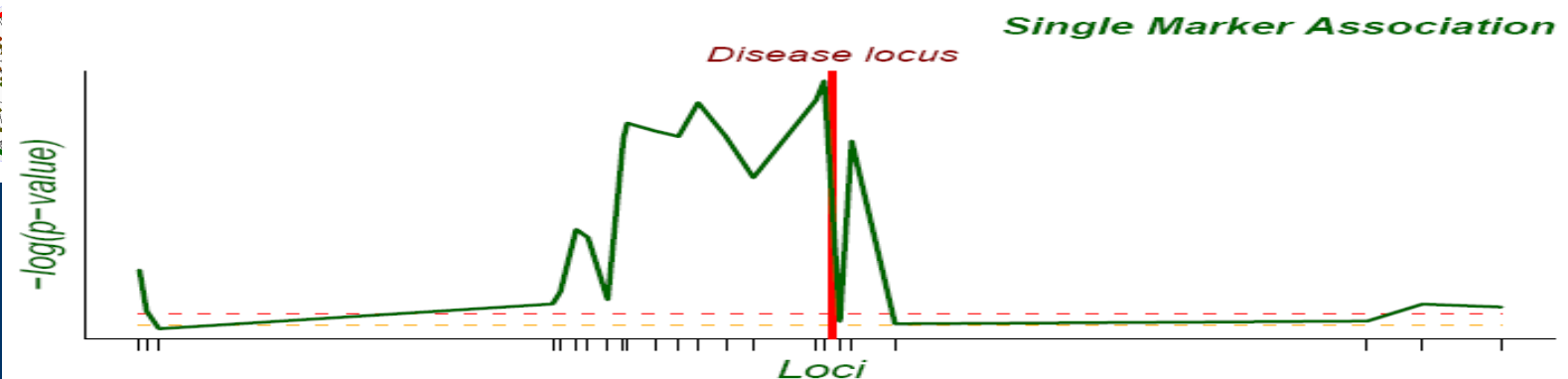
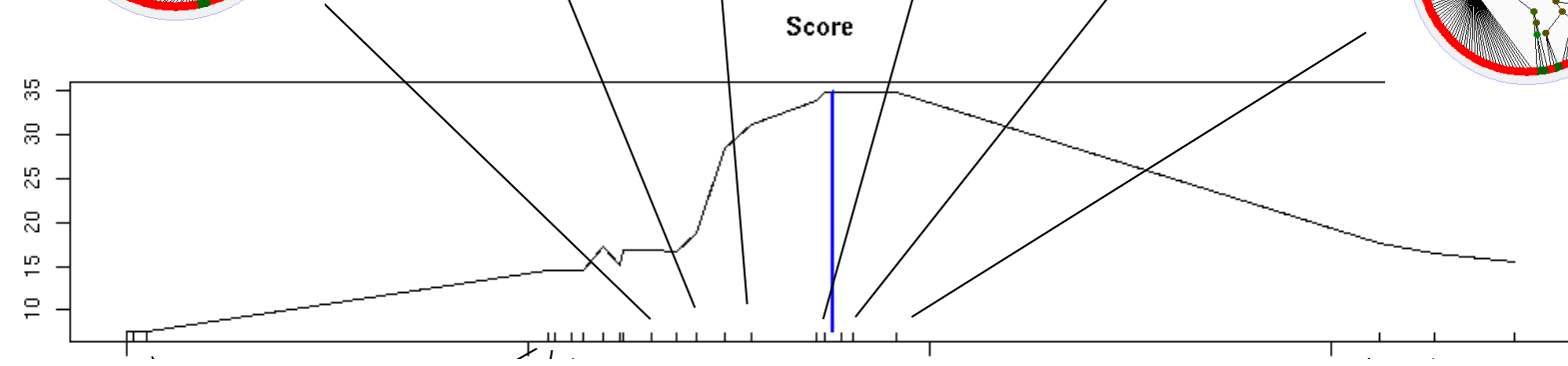
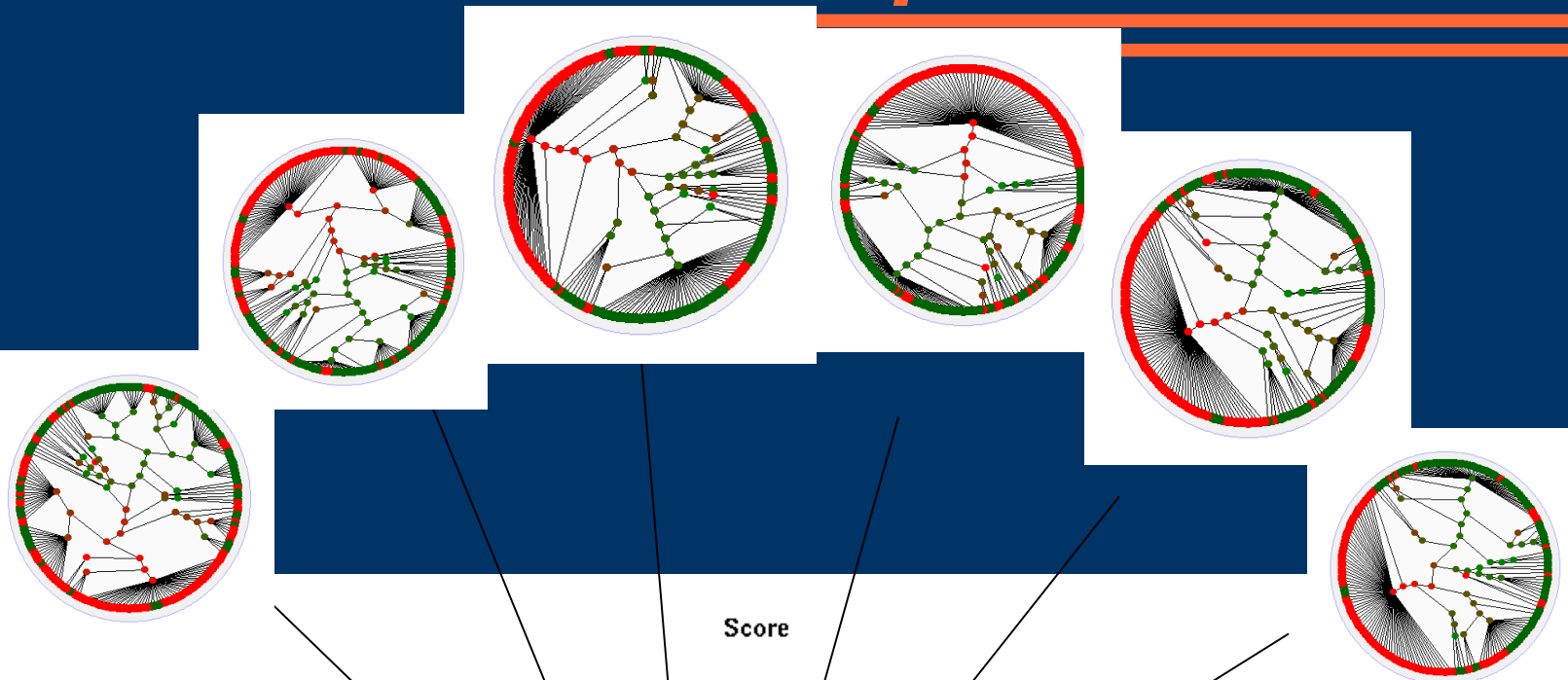
Open Office on DAIMI suck ***ass!!!***

... so this part goes on the
whiteboard...

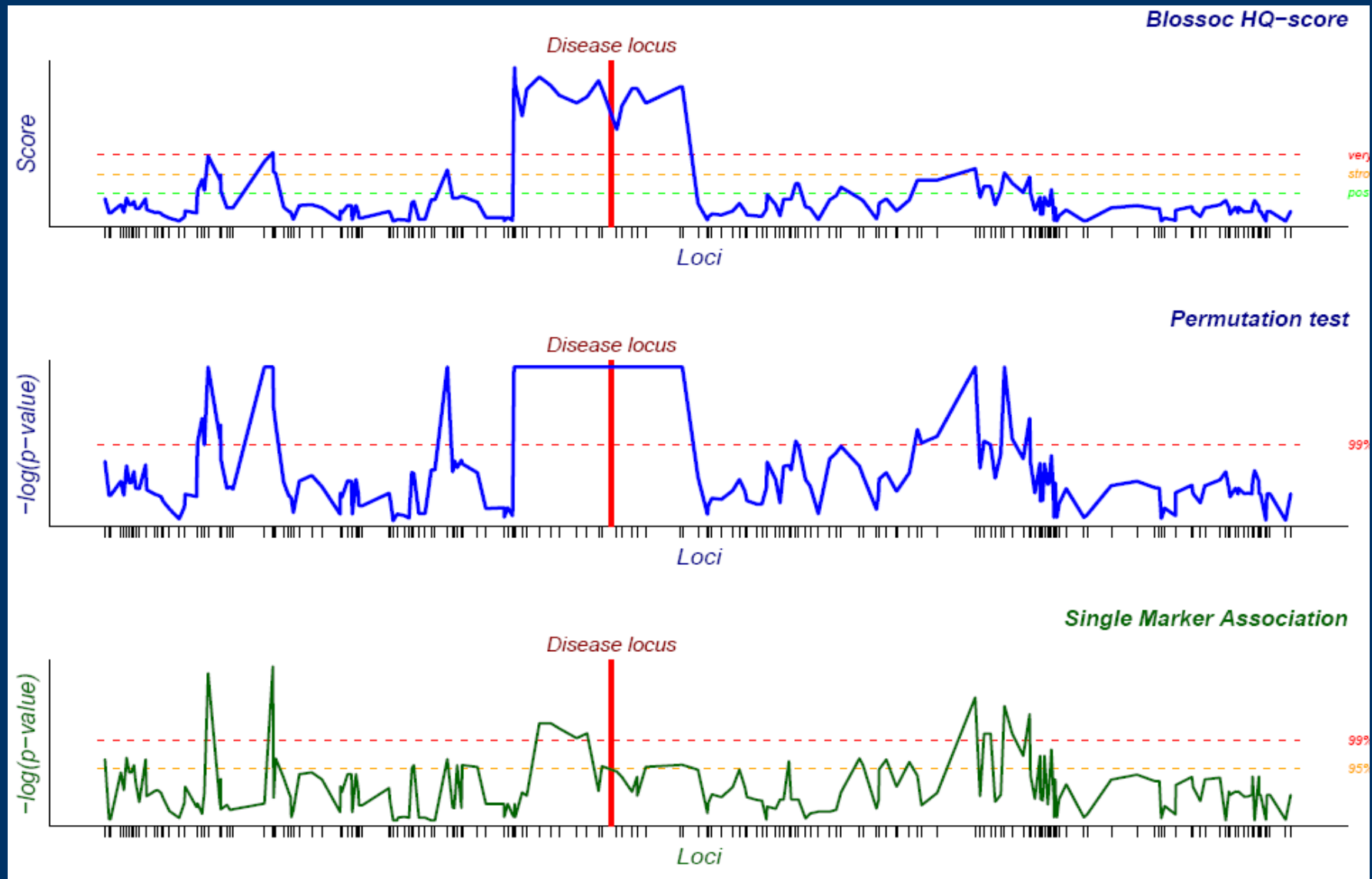
Cystic Fibrosis example



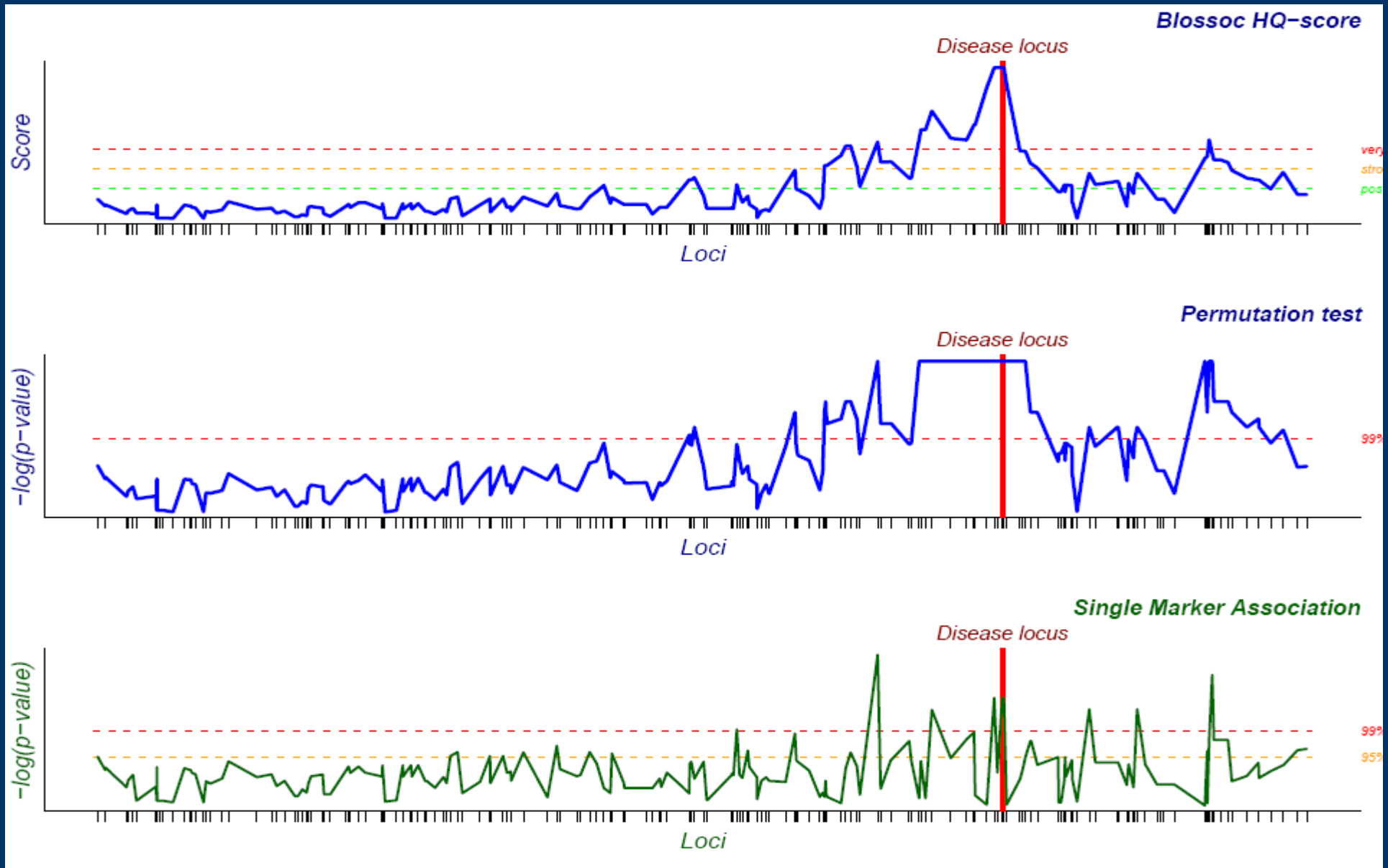
Cystic Fibrosis example



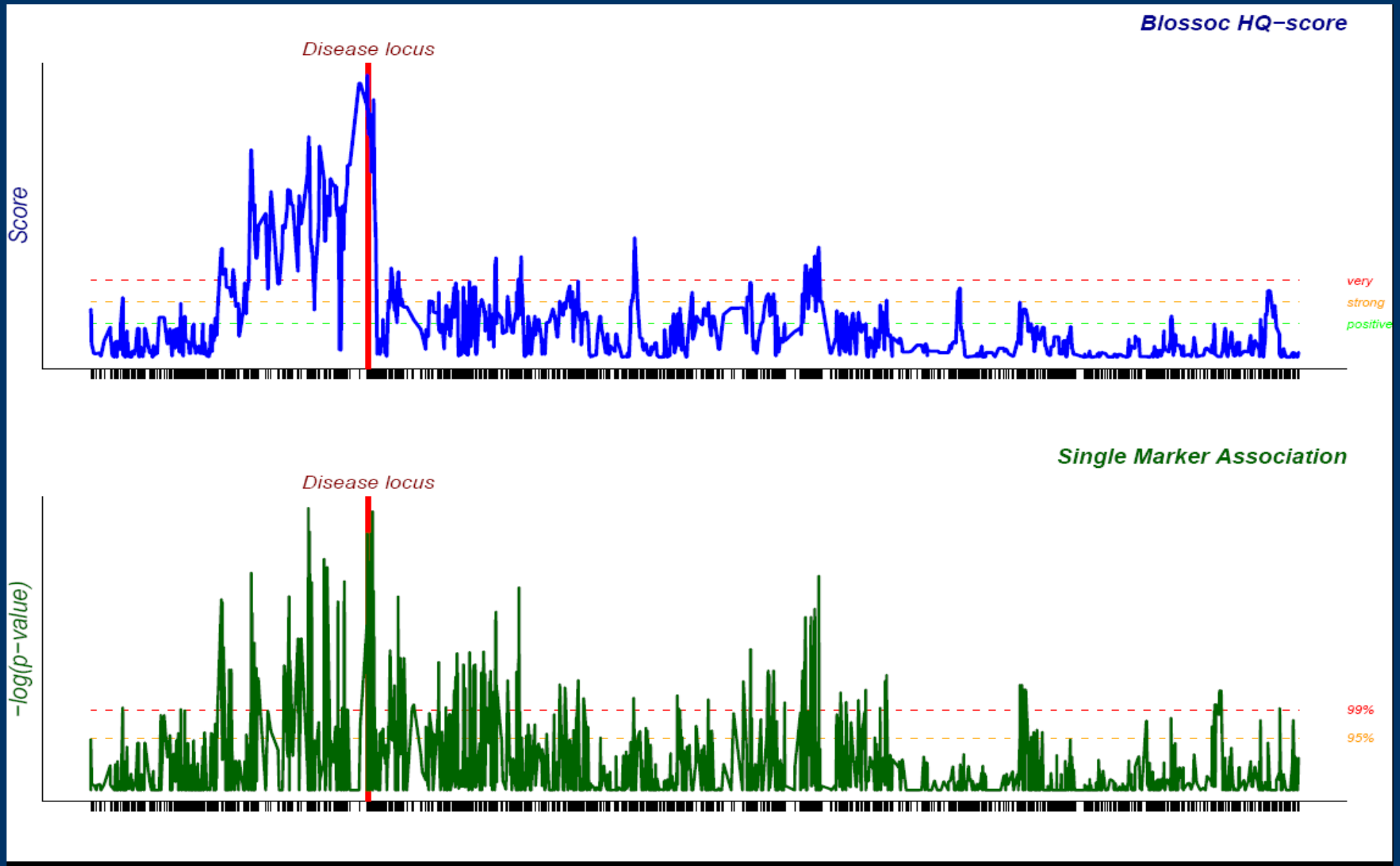
Simulated Example (CoaSim)



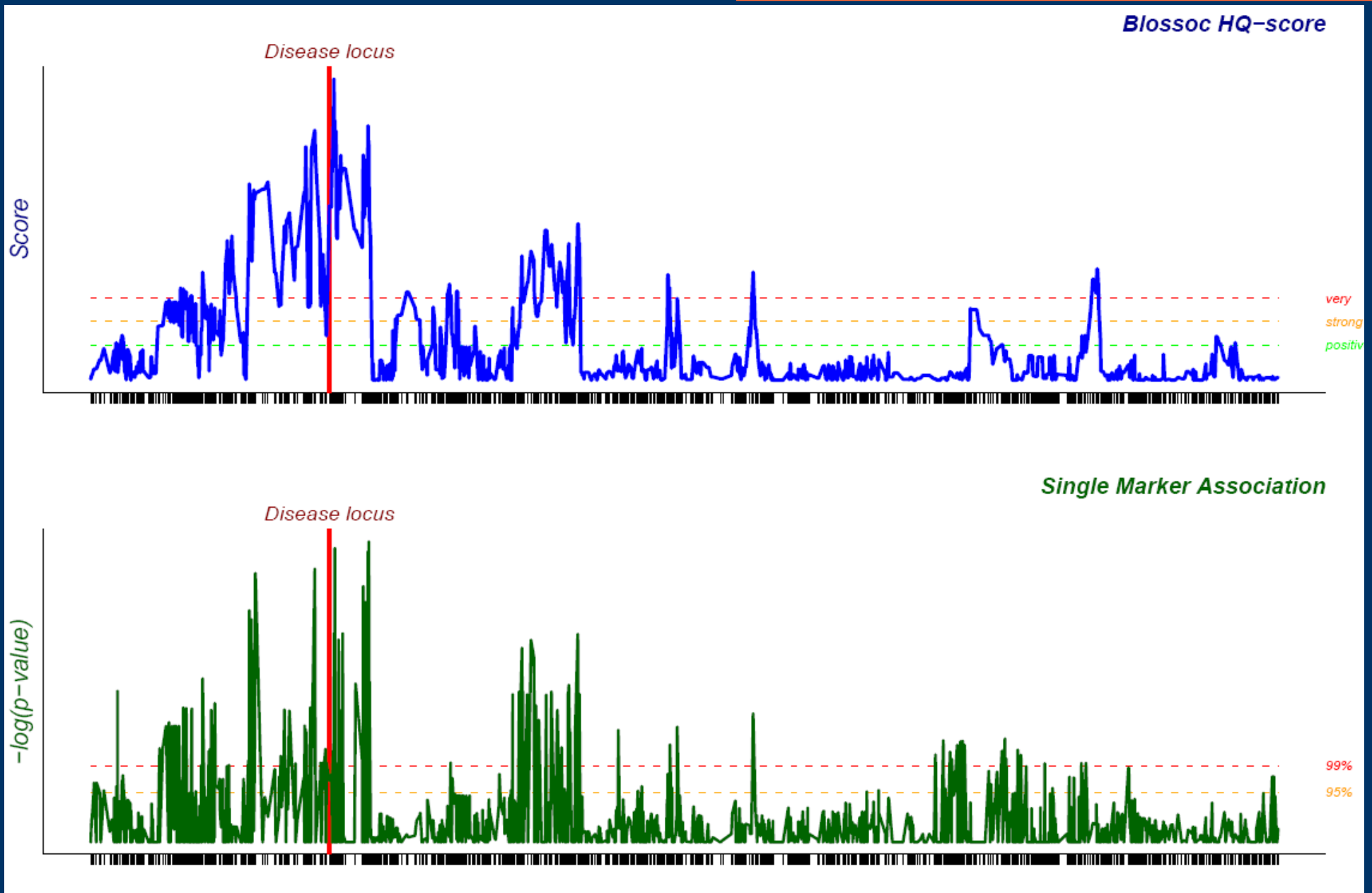
Simulated Example (CoaSim)



Augmented HapMap data



Augmented HapMap data



Implementation...

Homepage: www.daimi.au.dk/~mailund/Blossoc

Command line and graphical user interface...



ABOUT BLOSSOC

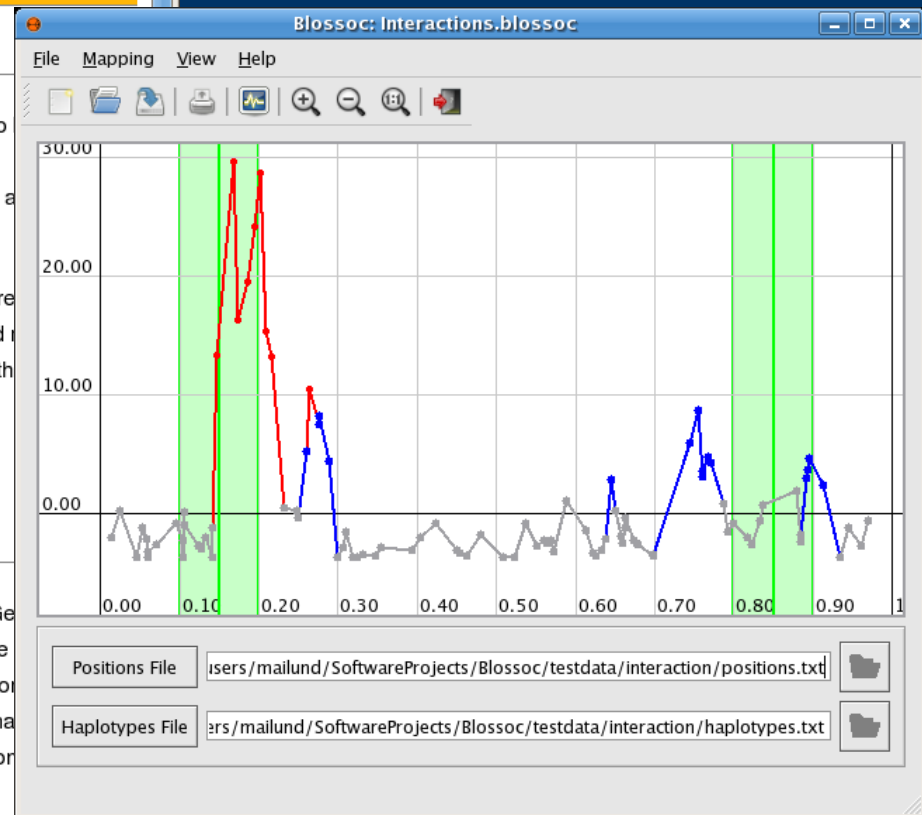
Blossoc is a linkage disequilibrium association mapping tool that attempts to (perfect) genealogies for each site in the input and score these according to non-random clustering of affected individuals, and judge high-scoring areas as candidates for containing disease affecting variation.

Building the local genealogy trees is based on a number of heuristics that are guaranteed to build true trees, but have the advantage of more sophisticated of being extremely fast. Blossoc can therefore handle much larger datasets than sophisticated tools, but at the cost of sacrificing some accuracy.

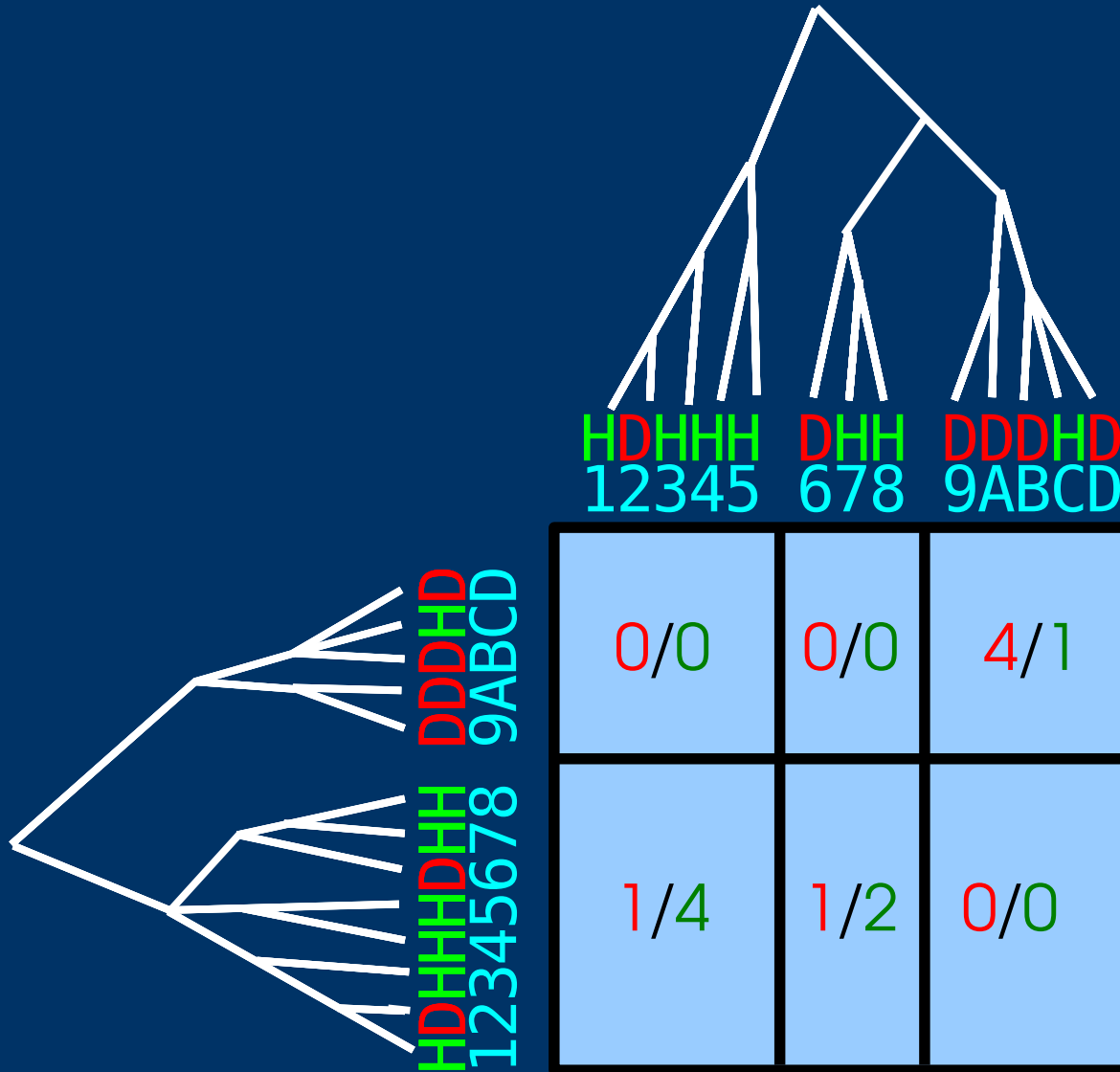
Blossoc is released under the [GNU General Public License](#).

INSTALLATION

Blossoc is written in C++ and is available as source code (under the GNU General Public License, GPL) and as binary versions as Linux RPM files. The source has been successfully compiled on various Linux and UNIX systems. As I have only limited access to architectures other than Linux, it is not possible for me to make distributions for other platforms, but if anyone is willing to build the distribution, I would be more than happy to put them on this site.



Interactions (pairwise)



Simple Idea:

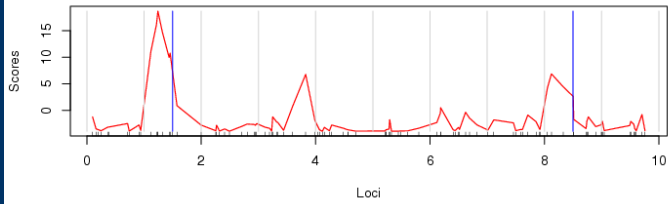
Collect intersections
of clusters...

Scoring a cluster pair...

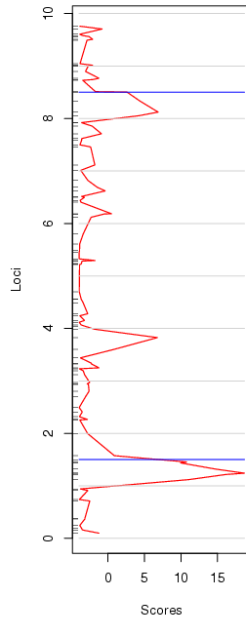
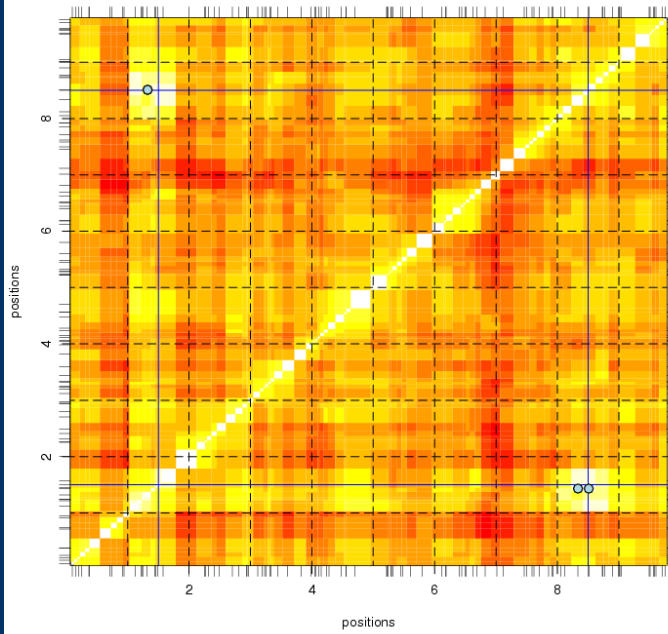
Once again, due to the ultimate suckitude of OO @ DAIMI, we move to the whiteboard...

Pairwise interactions...

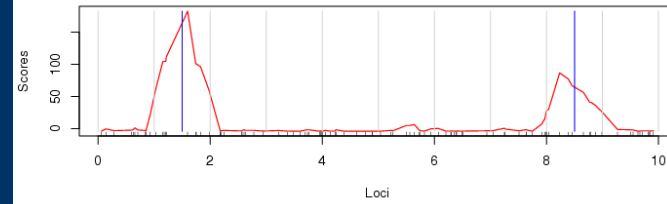
Marginal Blossoc scores



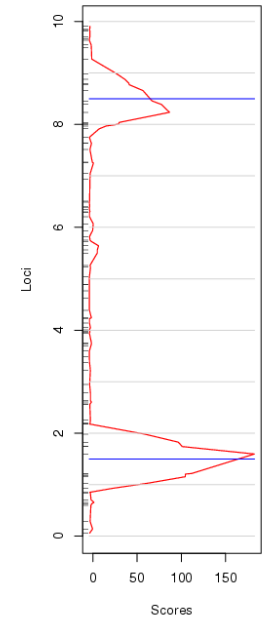
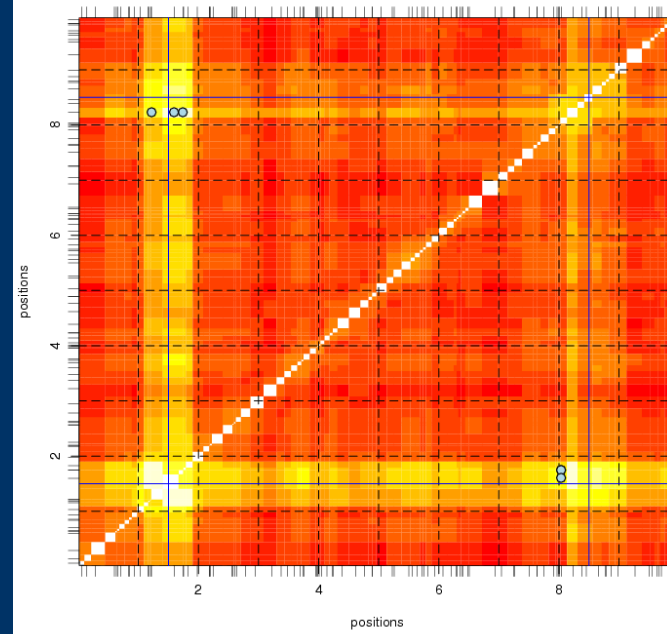
Pairwise scores



Marginal Blossoc scores

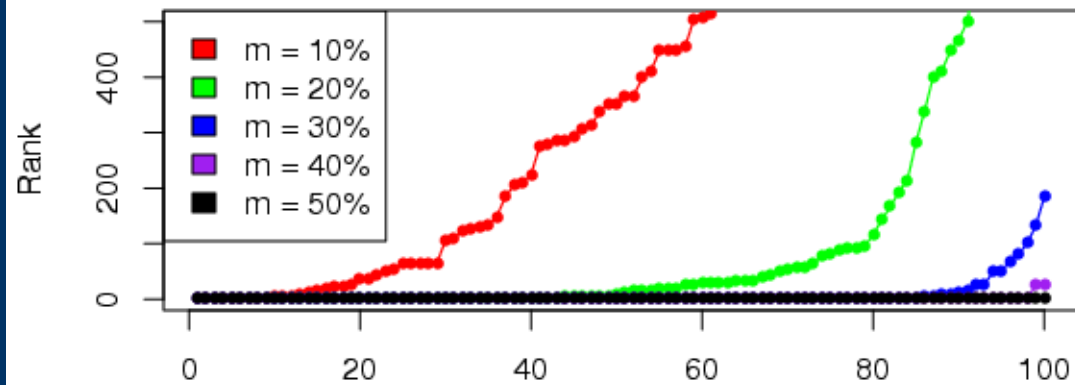


Pairwise scores

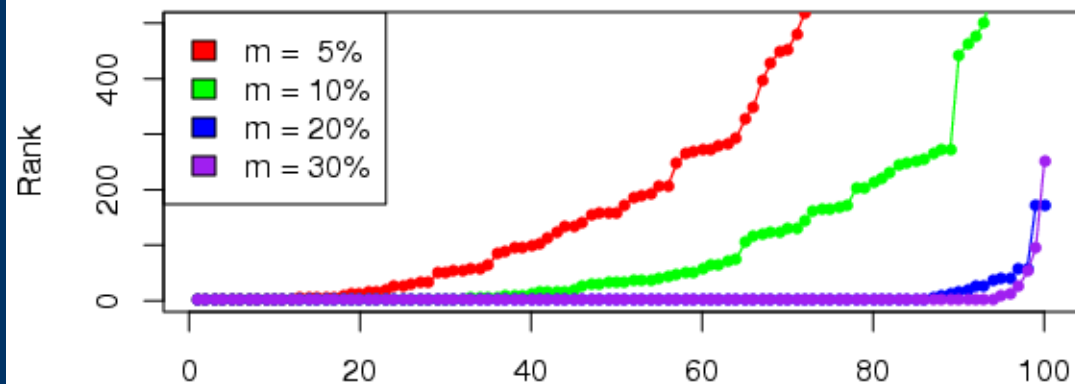


Rank of first pair in the right region...

Rank of first pair in the correct region; wt = 5%



Rank of first pair in the correct region; wt = 2%



Future work...

- Interactions:
 - Improve scoring
 - Heuristic search methods to move beyond pair-wise interactions
- Whole genome scans:
 - Algorithmic speedups
 - Parallelisation