

Association mapping

Searching for Disease Causing Genes

Thomas Mailund
Bioinformatics ApS

Defining the problem

Looking for a cure for cancer?

Defining the problem

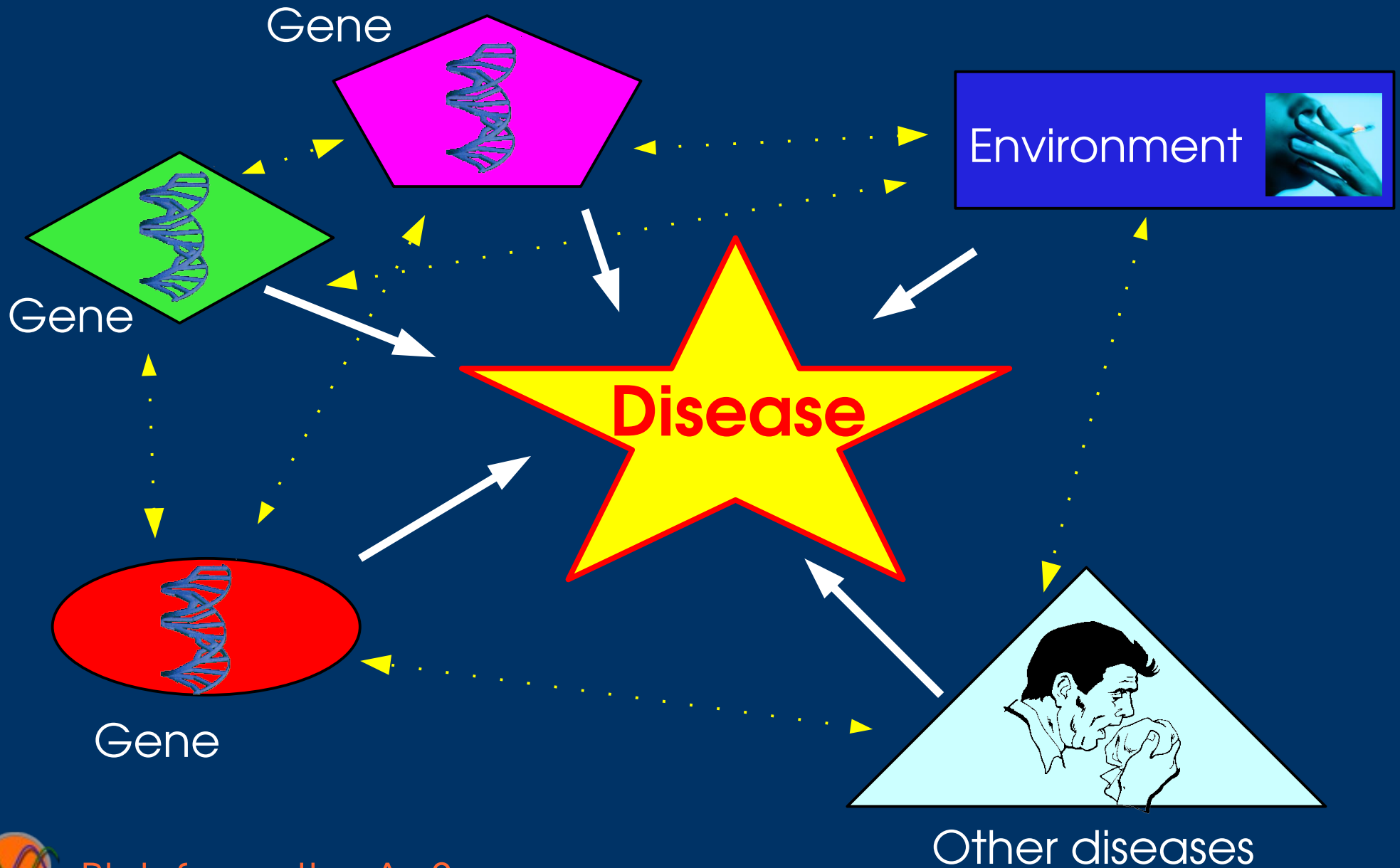
~~Looking for a cure for cancer?~~

Looking for a *cause* for cancer!

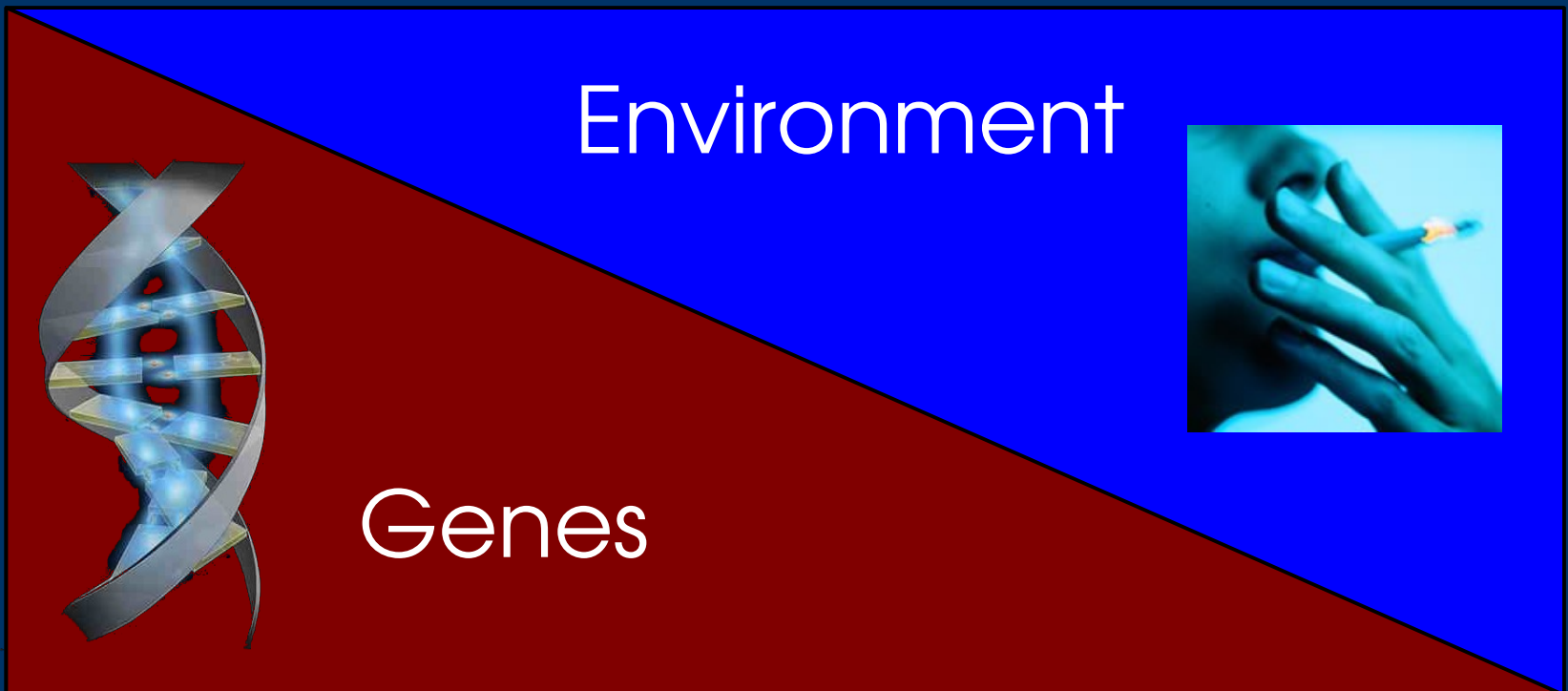
Defining the problem

- Know the causes of a disease to:
 - Understand the disease
 - Attack the causes, not the symptoms
 - Identify risks or people at risk (early warning system)

What are the causes?



Genes vs environment



Gunshot wounds

Car accidents

Smoking induced

lung cancer

Cardiovascular

disease

Obesity

Diabetes 2

Alzheimer

Schizophrenia

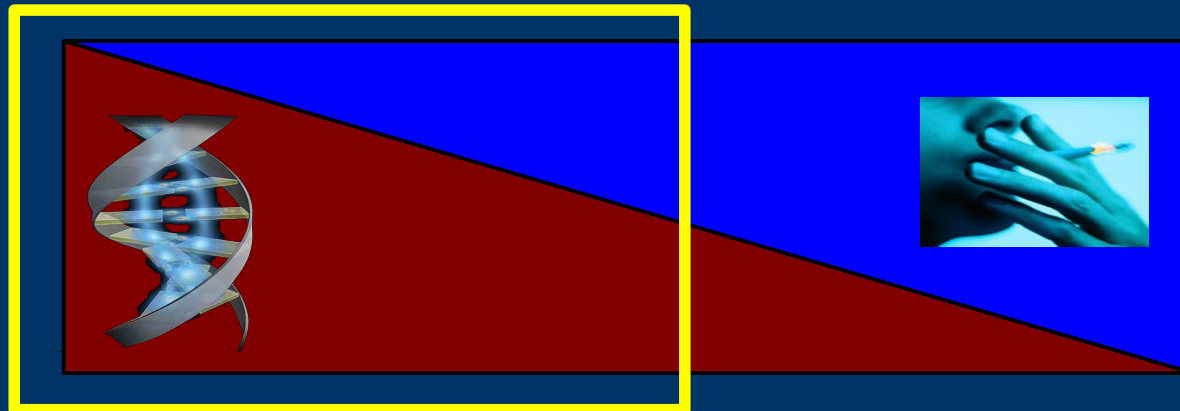
BRCA1 breast
cancer

Cystic fibrosis
Hemophilia



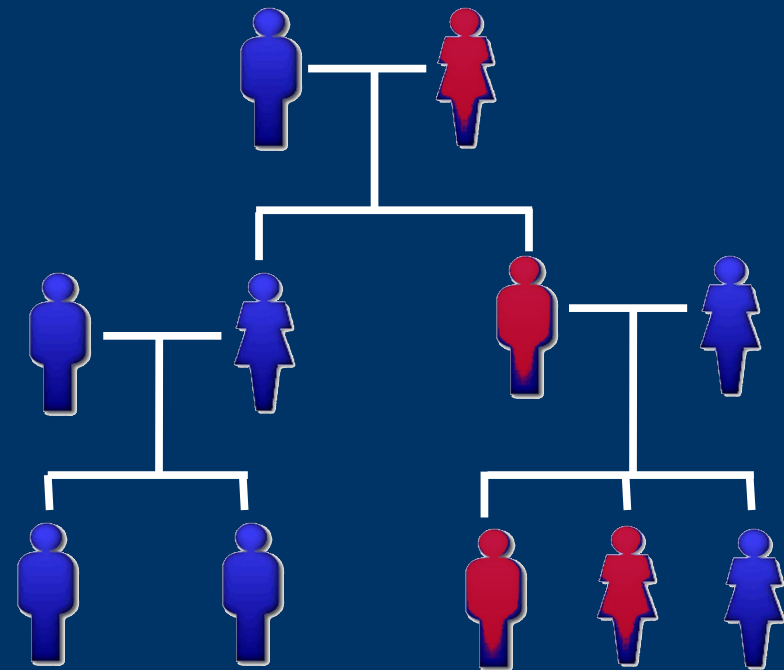
Bioinformatics ApS

We are in the business of
locating disease affecting genes
and the
disease affecting variations



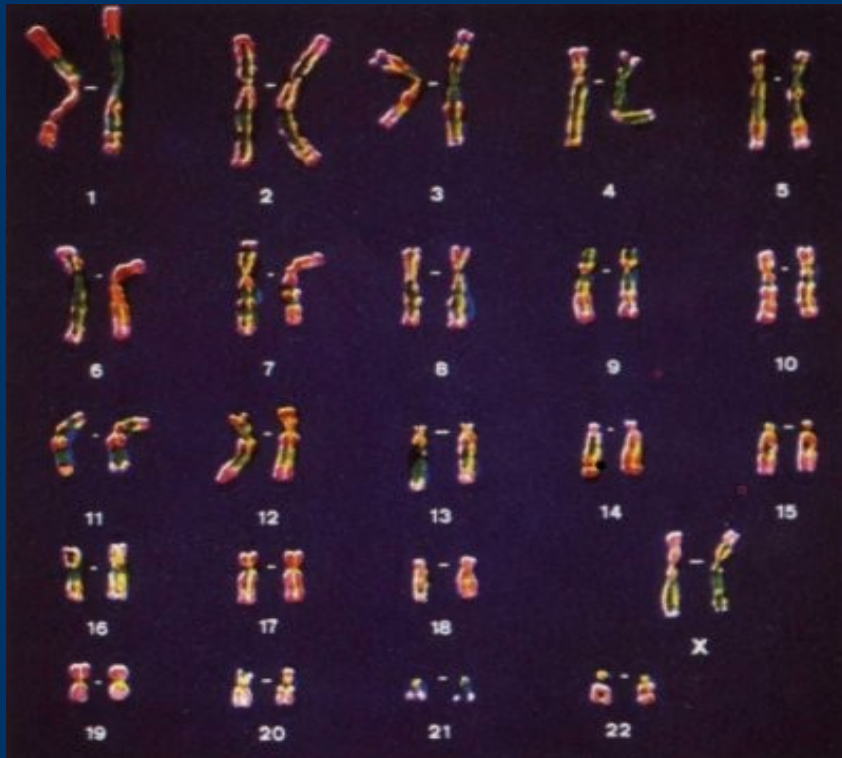
In the genes?

- How do we know there is a genetic component?
 - Clustering in families
 - Twin/adoption studies



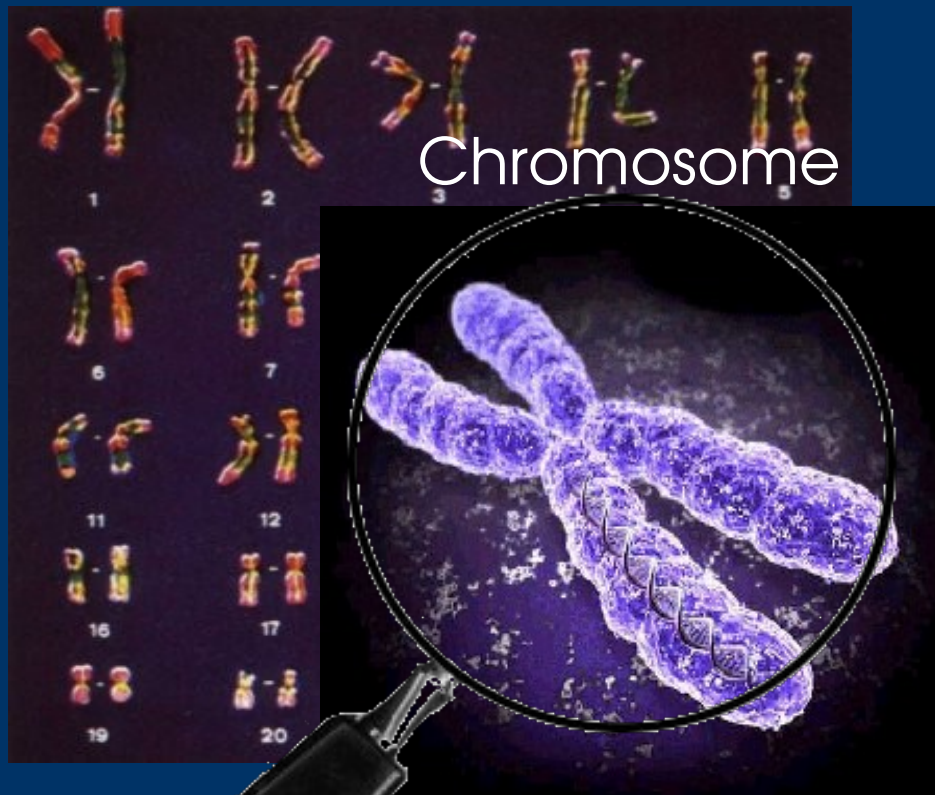
Searching for the disease genes...

Genome



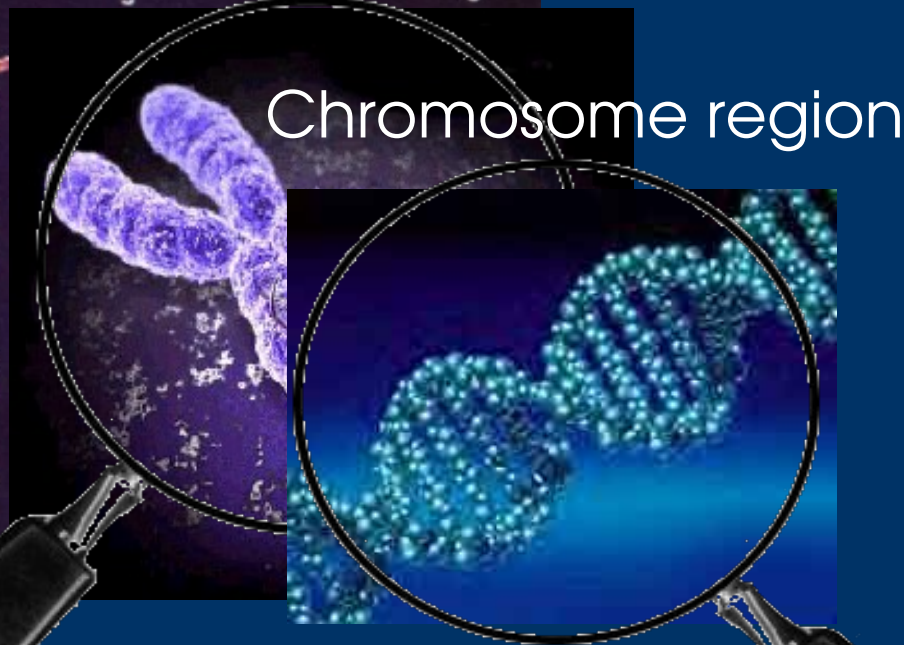
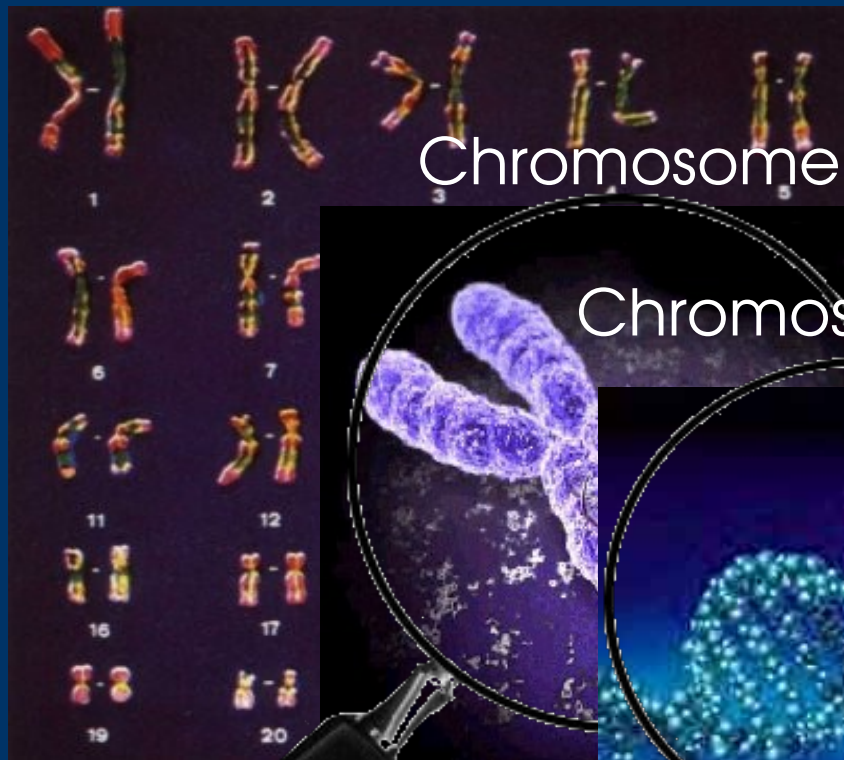
Searching for the disease genes...

Genome



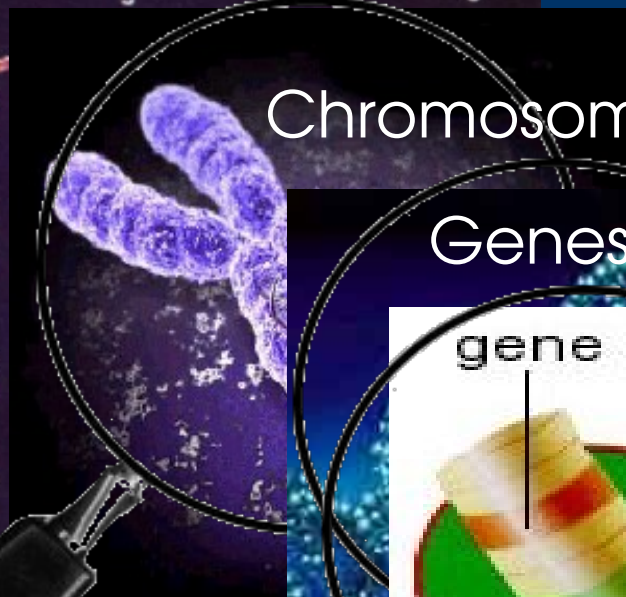
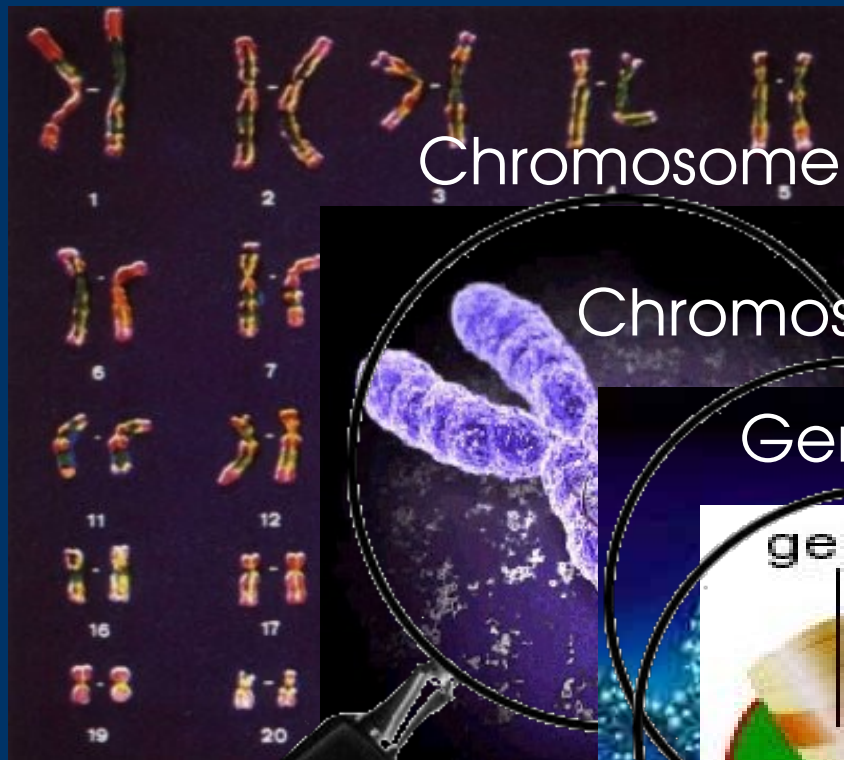
Searching for the disease genes...

Genome



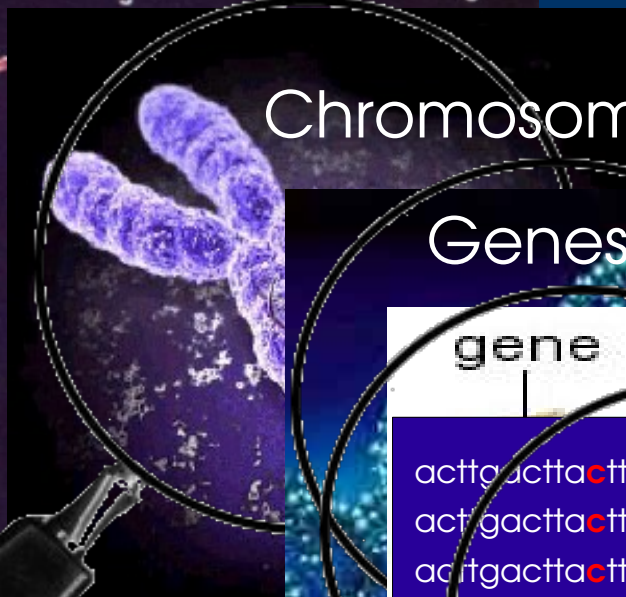
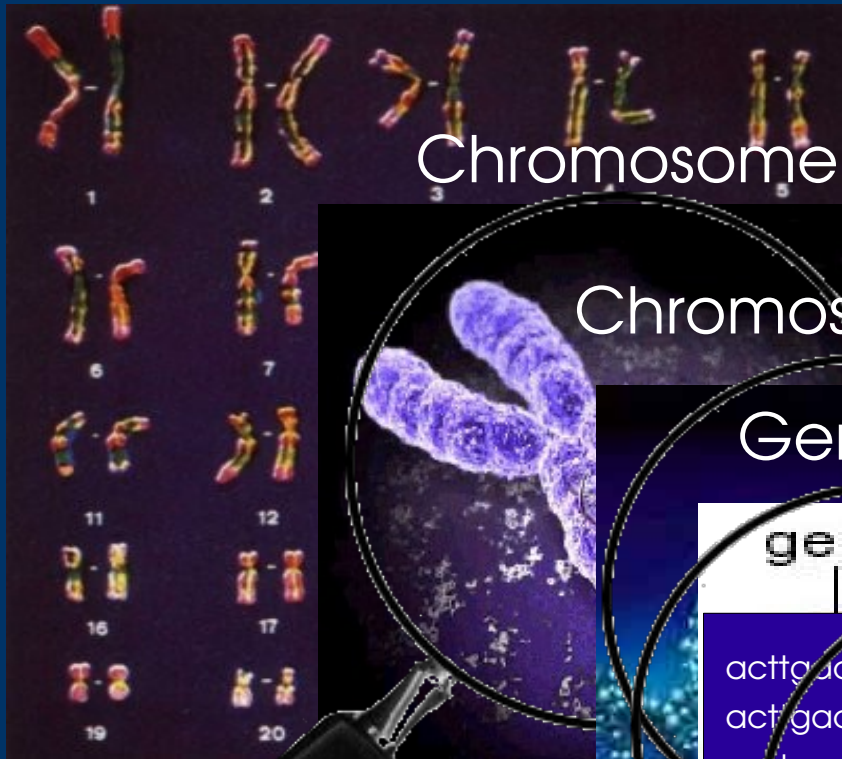
Searching for the disease genes...

Genome



Searching for the disease genes...

Genome



Genes

gene

```
actgacttacttaccggttgacttactaccagac  
actgacttacttaccggttgacttactaccagac  
actgacttacttaccggttgacttgctaccagac  
actgacttagttaccggttgacttactaccagac  
actgacttacttaccggttgacttactaccagac  
actgacttagttaccggttgacttactaccagac  
actgacttagttaccggttgacttactaccagac
```

Polymorphisms



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

Enabling technology

- Sequencing chips:



Affymetrix: 100K, next version 500K



Illumina: 100K, next version 250K

- Price:

- ~ \$1000 per chip

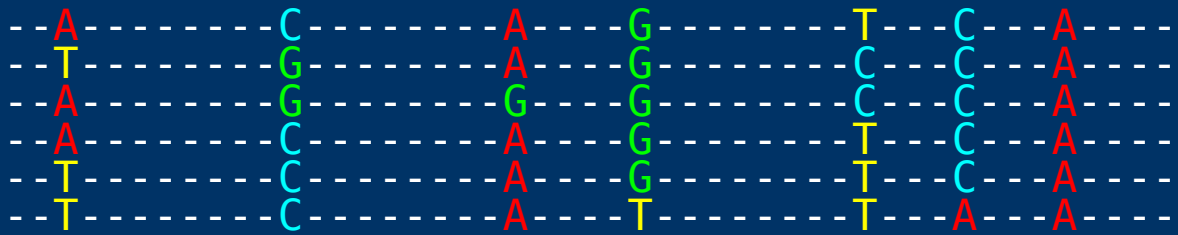
- ~ ¢0.01-0.005 per marker

- ~ \$1,000,000 for 1000 individuals

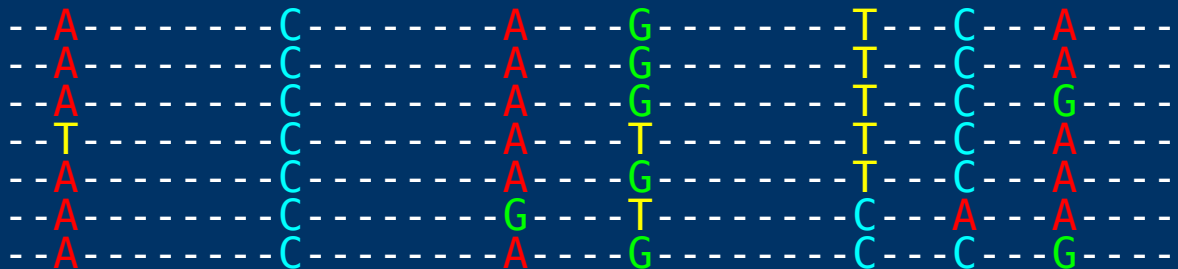
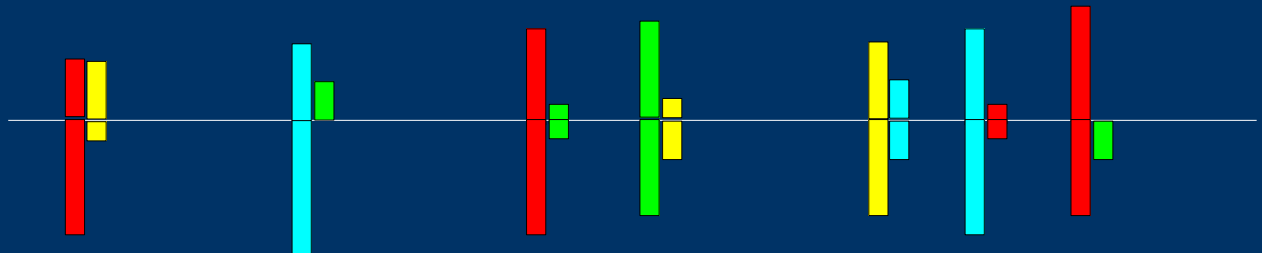


Association mapping

We are searching for an association between variant and disease status

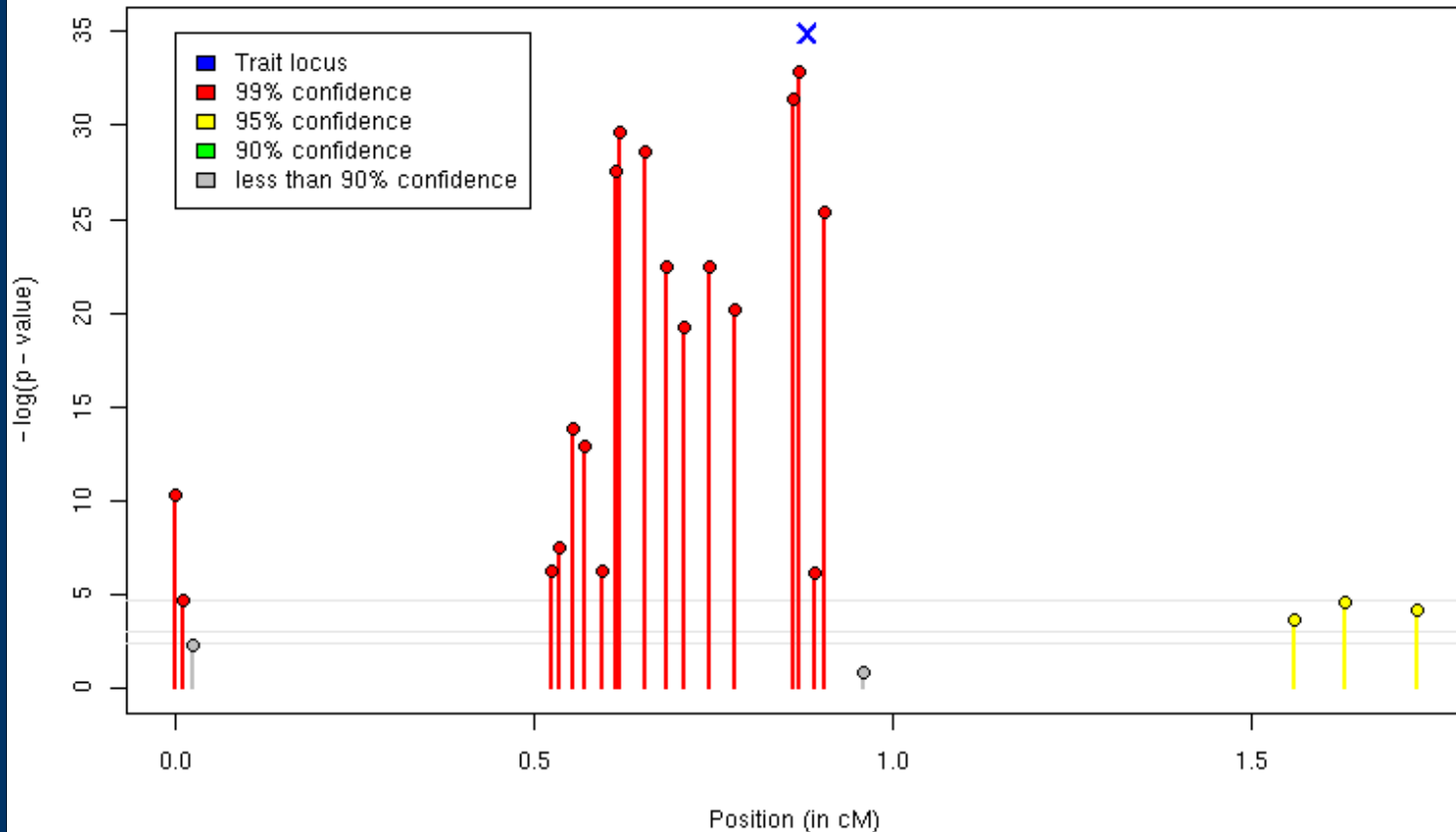


Significant difference in distributions?



Example: Cystic fibrosis

χ^2 -test for different distributions



Kerem et al. (1989)

Control group: 92 SNP Haplotypes

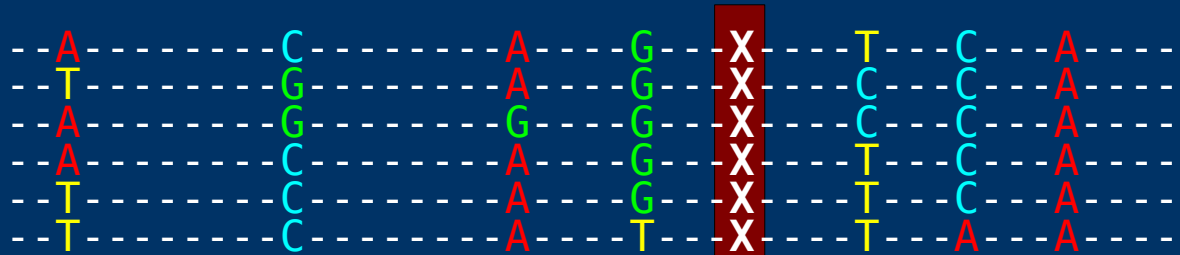
Case group: 94 SNP Haplotypes

23 SNP Markers



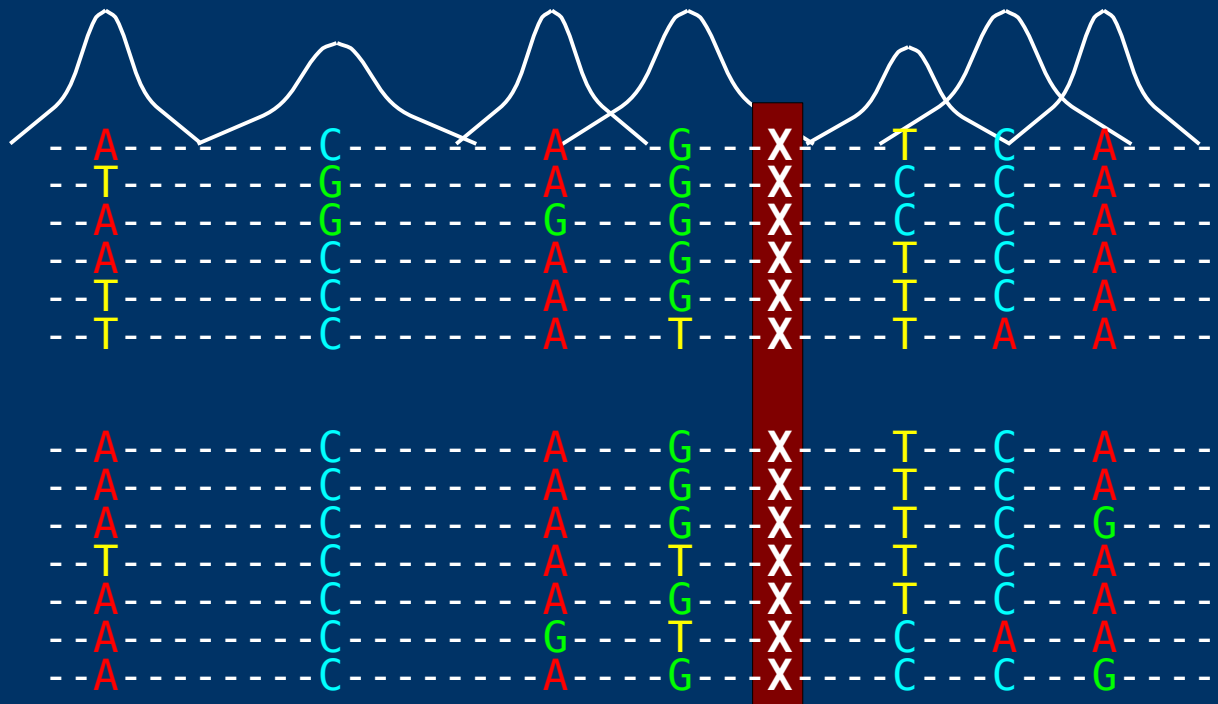
An indirect approach

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



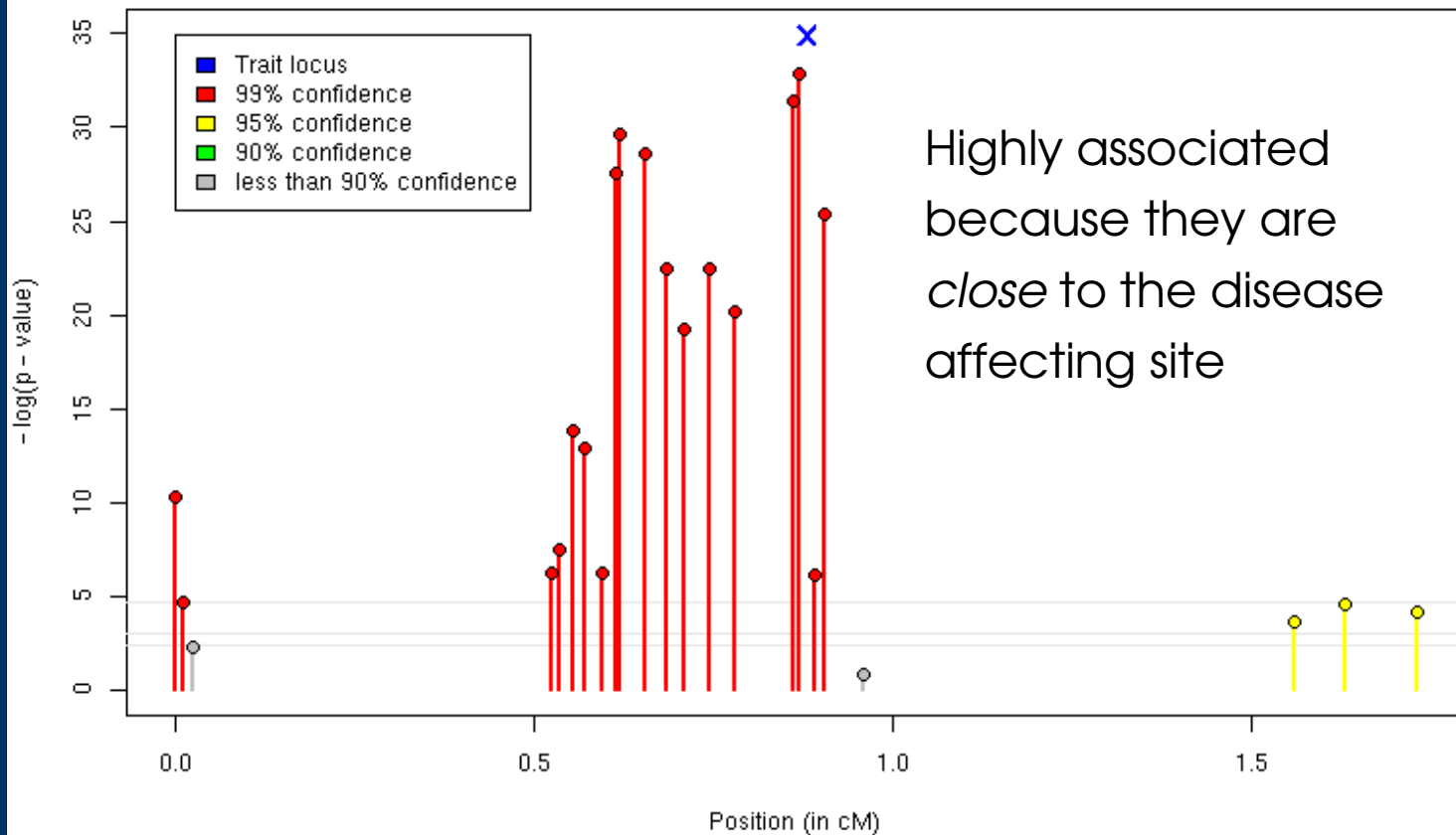
An indirect approach

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



An indirect approach

χ^2 -test for different distributions



Extend of relatedness

- “Nearby” is $\sim 0.1\text{--}0.01$ cM
 - $\sim 100\text{--}10$ Kbp
 - $\sim 1/30,000 - 1/300,000$ of the genome
- Closer spacing needed for accuracy
 - $\sim 500,000\text{--}1,000,000$ for whole genome
 - $\sim 10\text{--}100$ for typical gene



Extend of relatedness

- Population dependent
 - Founding age
 - Isolation (inbreeding)

Isolated recently founded

Quebec, Cajun Acadiana
Utah
Amish
Iceland
Faroese Islands

Isolated relatively old

Kainuu (Finland)
North Karelia (Finland)
Sardinien
Ashkenazi Jews

non-isolated relatively old (bottlenecks)

European, Asian

Africa



extends over longer distances

“low” density marker map

decreased complexity

low resolution

extends over shorter distances

“high” density marker map

increased complexity

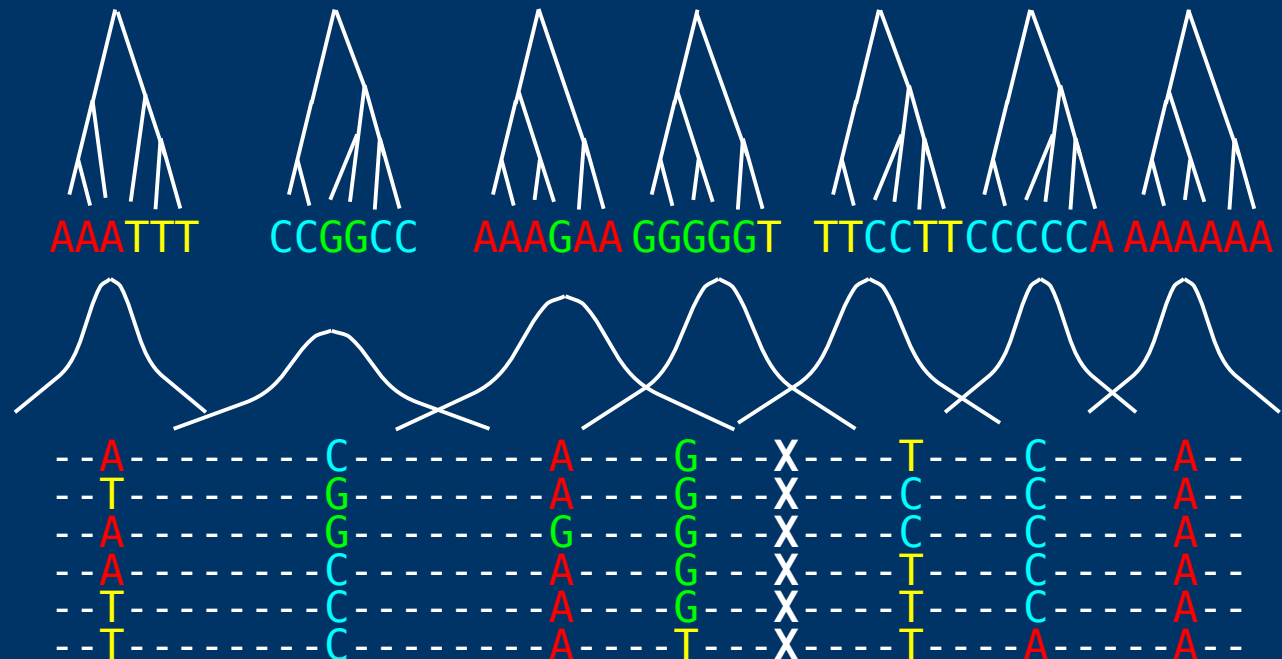
high resolution



Local genealogies

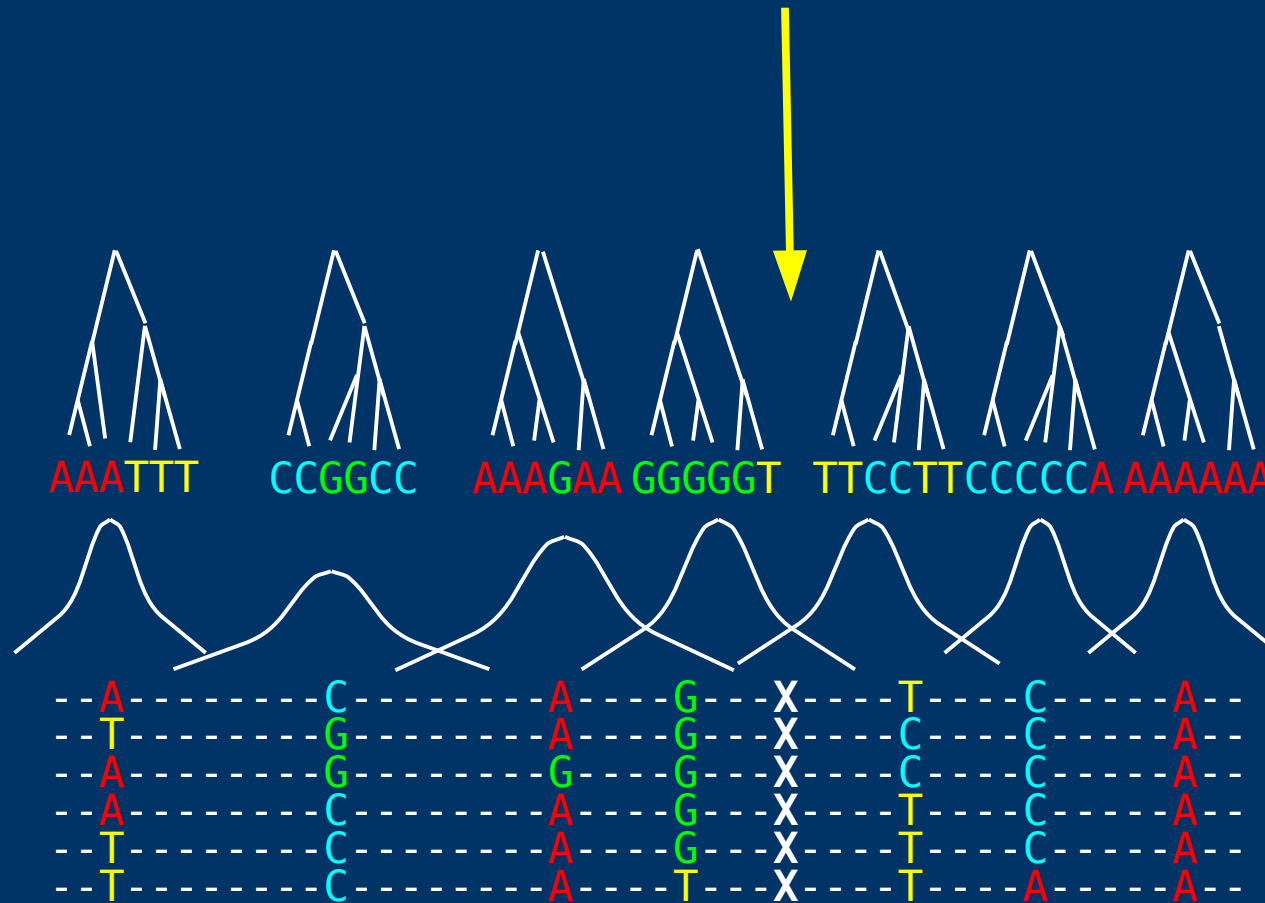
- Recombinations
 - Each site a different genealogy
 - Nearby genealogies only slightly different

A nearby tree
an imperfect
local tree



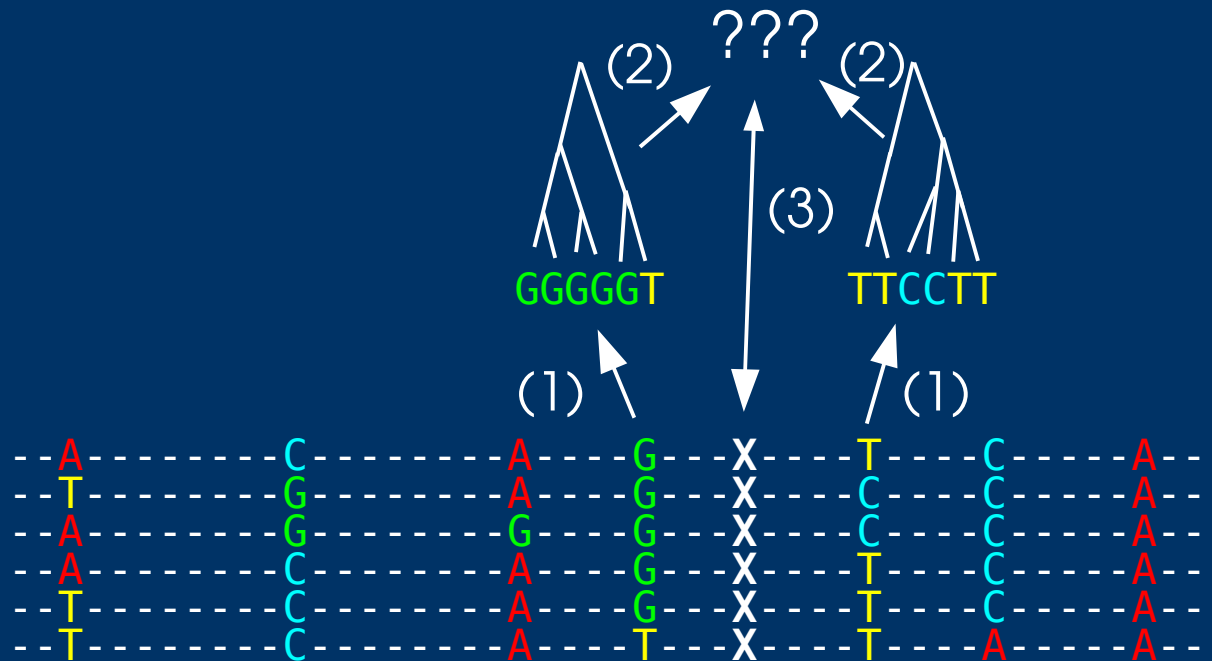
Local genealogies

Tree at disease site resembles neighbours



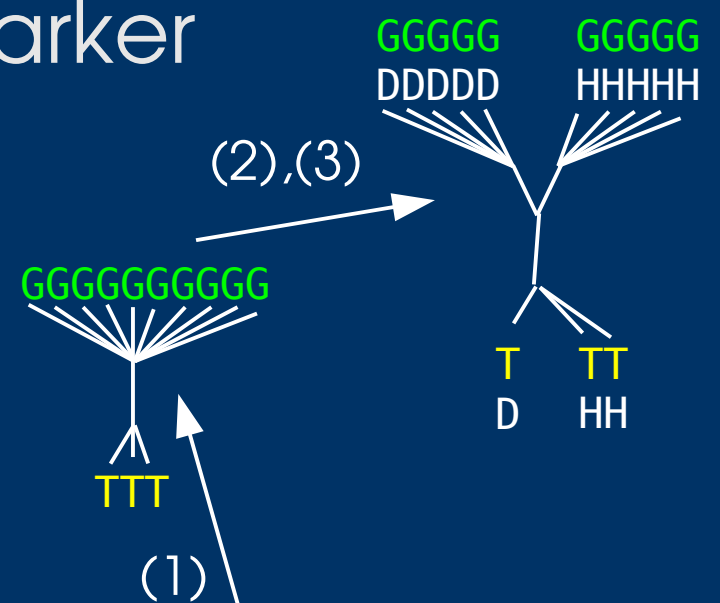
Local genealogies

- Some information loss due to partial knowledge
 - (1) Not *knowing* the trees
 - (2) Using *neighbouring* trees
 - (3) Disease locus and status not identical

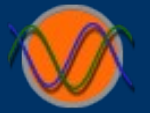


Single marker association

- Limited power in a single marker
 - Simple tree (single split)
 - Assumed status=variation



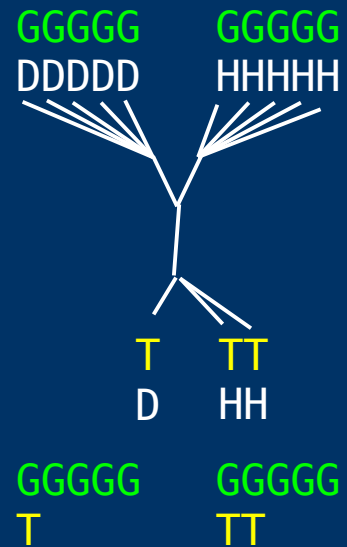
A	C	A	G	D	T	C	A
T	G	A	G	D	C	C	A
A	G	G	G	D	C	C	A
A	C	A	G	D	T	C	A
T	C	A	G	D	T	C	A
T	C	A	T	D	T	A	A
A	C	A	G	H	T	C	A
A	C	A	G	H	T	C	A
A	C	A	T	H	T	C	G
A	C	A	G	H	T	C	A
A	C	G	T	H	C	A	A
A	C	A	G	H	C	A	G



Single marker association

- Limited power in a single marker

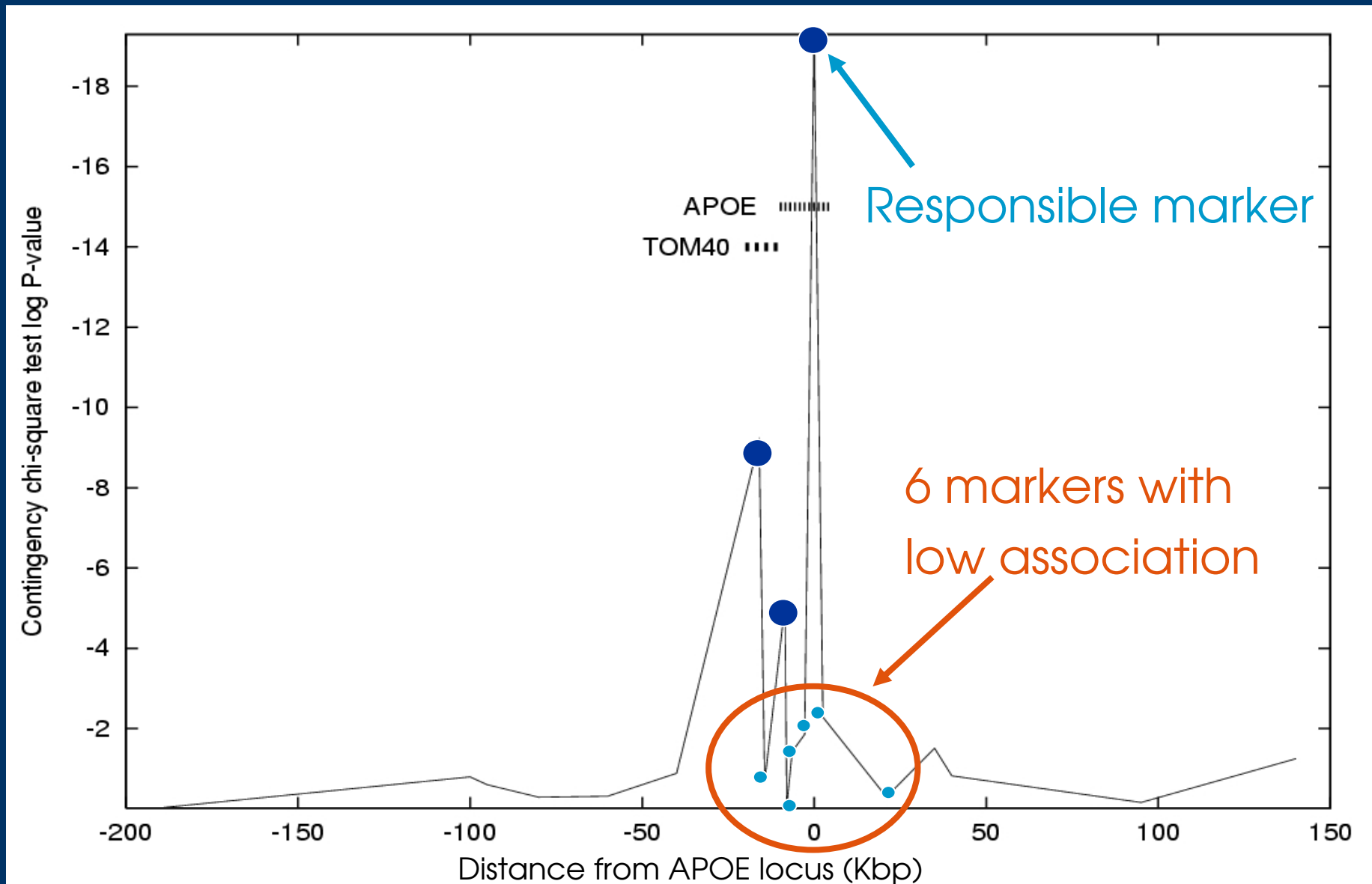
- Simple tree (single split)
- Assumed status=variation
- Genealogy → frequency
- Marker/disease distance not modeled



A	C	A	G	D	T	C	A
T	G	A	G	D	C	C	A
A	G	A	G	D	C	C	A
A	C	A	G	D	T	C	A
T	C	A	T	D	T	C	A
T	C	A	T	D	T	C	A
A	C	A	G	H	T	C	A
A	C	A	G	H	T	C	G
A	C	A	T	H	T	C	A
A	C	A	G	H	T	C	A
A	C	A	T	H	C	A	A
A	C	A	G	H	C	A	G



Alzheimer and ApoE



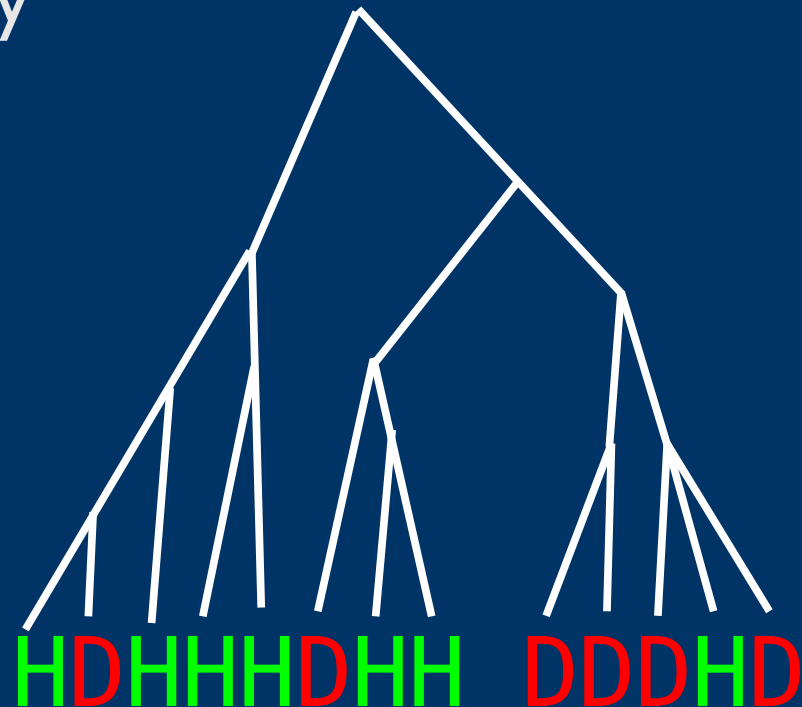
Local genealogies revisited

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



Local genealogies revisited

- Closeness to disease site:
 - A significant clustering of diseased/healthy



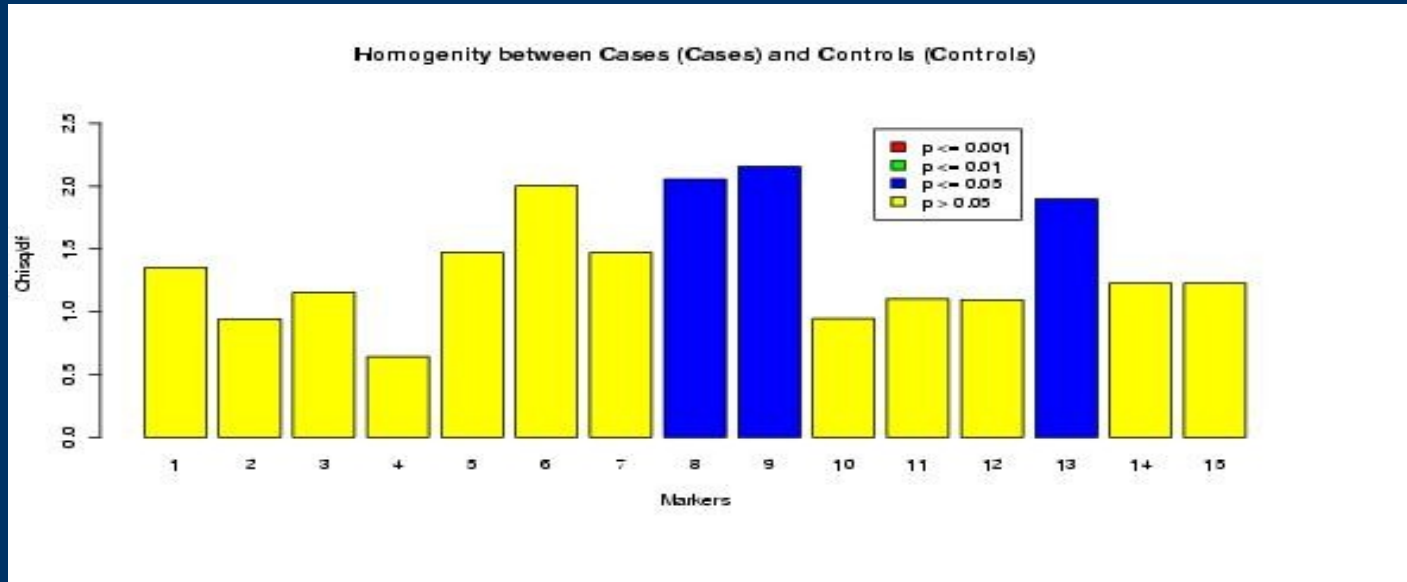
Our approach...

- Multi-marker association
 - Windows
 - Distance-weighting
- Explicit modeling of genealogies
 - Full sample or sub-samples
 - All or just cases
- Search through space of positions and genealogies
 - Statistical methods, e.g. MCMC
 - Heuristics based on recombination evidence

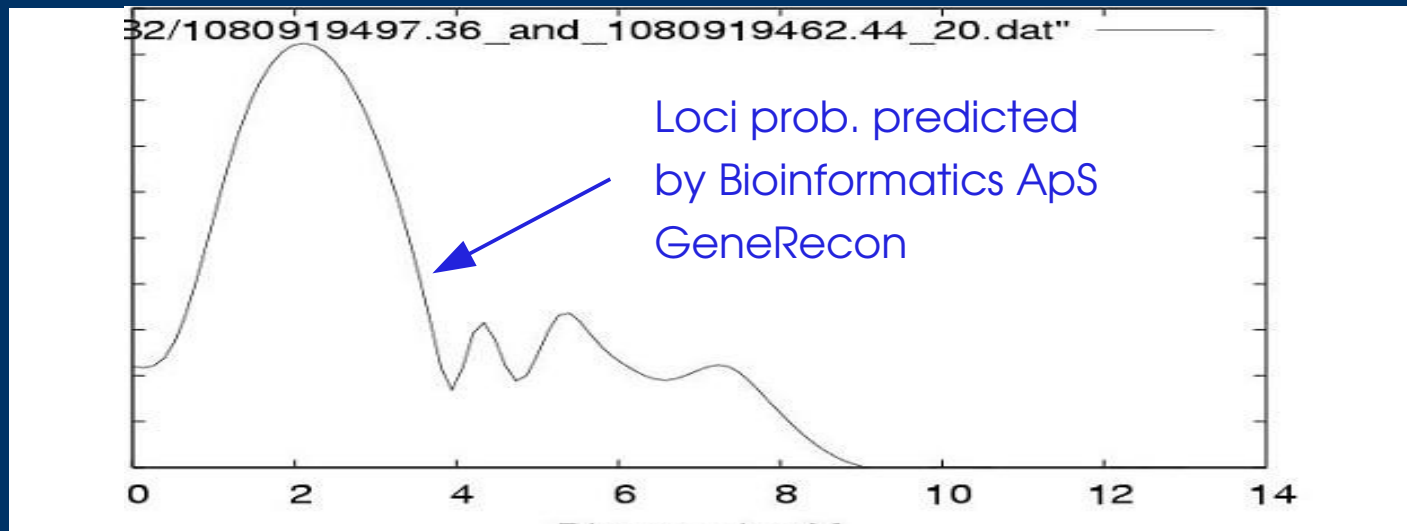


Example: BRCA 2

Single marker association



GeneRecon



Steinunn et al, pers comm

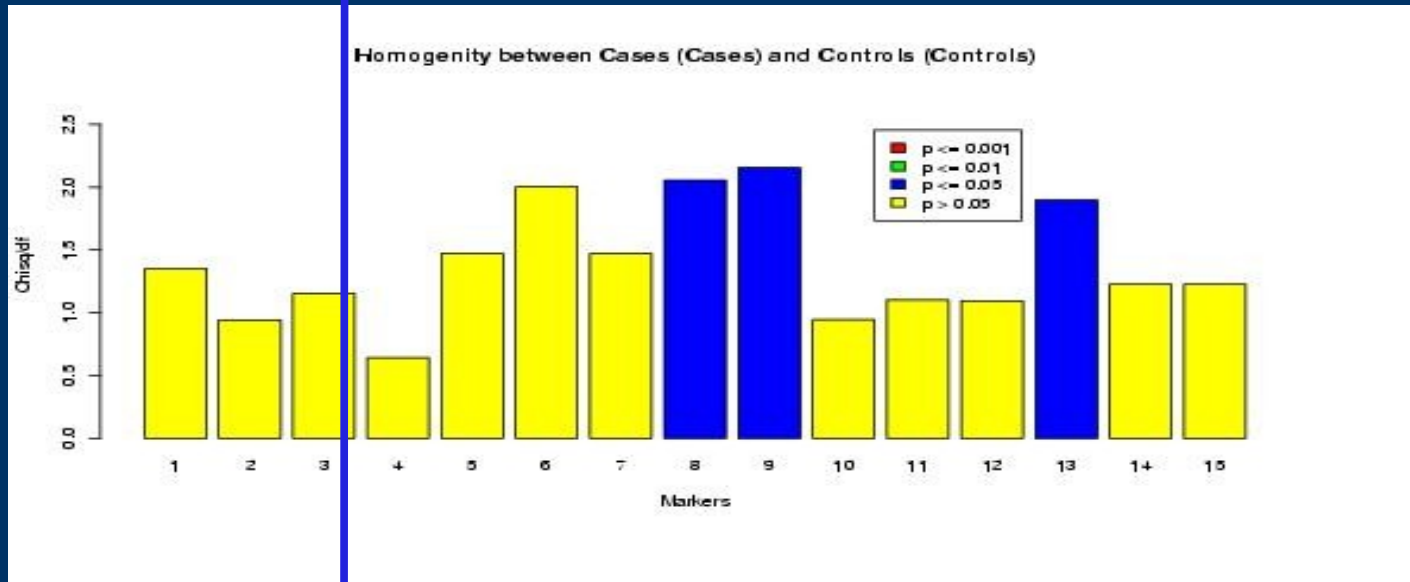
1000 cases, 1000 controls

15 microsatellite markers

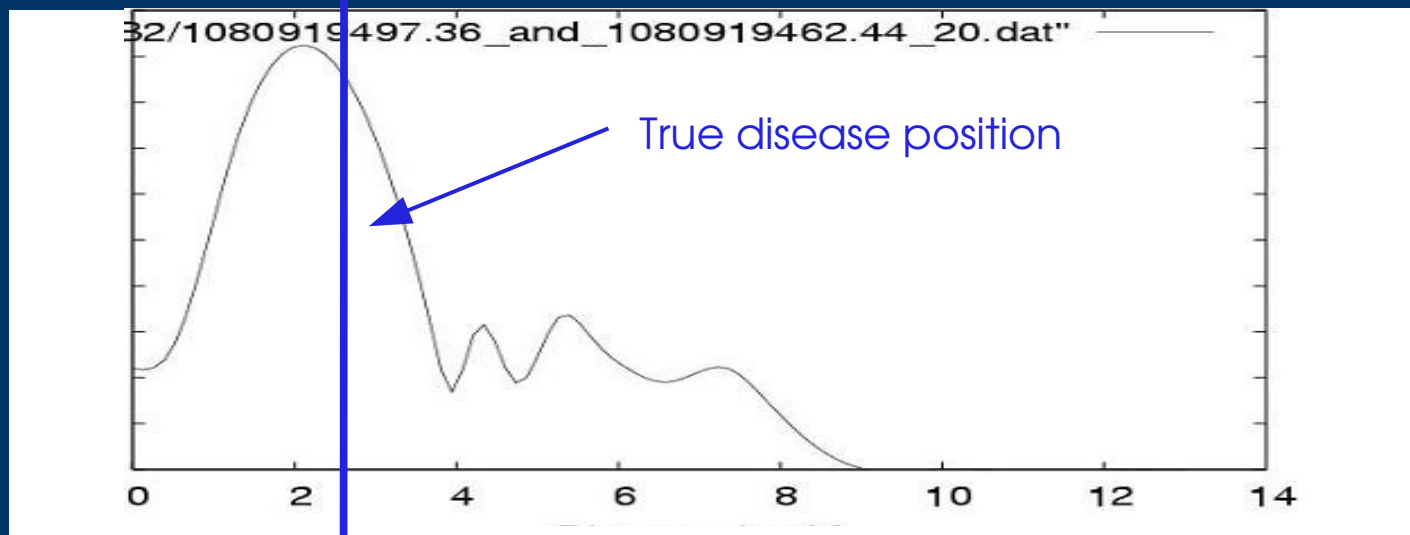


Example: BRCA 2

Single marker association



GeneRecon



Steinunn et al, pers comm

1000 cases, 1000 controls

15 microsatellite markers



Analysis time...

- Very CPU intensive!
 - 1000 cases/controls, 100 (SNP) markers: ~1 week
 - Several runs usually needed
- Collaboration with *Minimal Intrusion Grid* (MiG)
 - Parallelization on computer clusters
 - “Screen saver science”
- “Quick and Dirty” heuristics for initial mapping
 - Locate candidate regions



Bioinformatics ApS software



CoaSim

→ Simulator for generating realistic data



GeneRecon

→ MCMC based statistical search

→ Fine-scale mapping



Blossoc

→ Heuristic search

→ Medium-scale mapping

