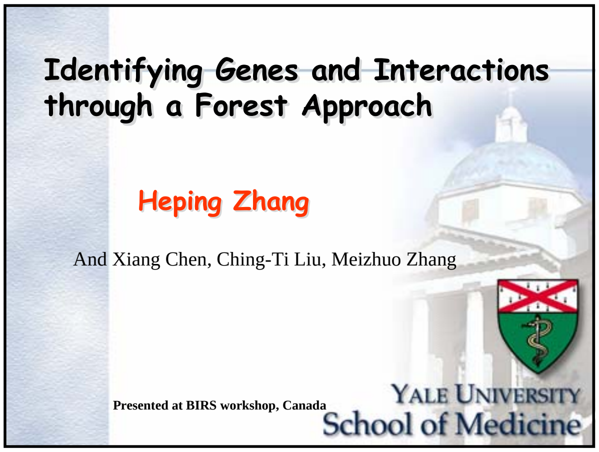


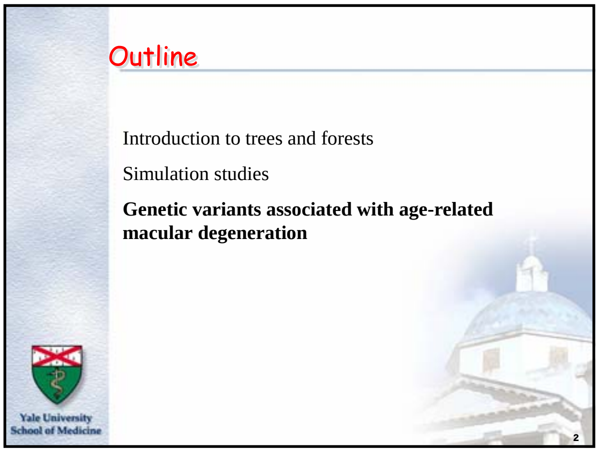
Identifying Genes and Interactions through a Forest Approach

Heping Zhang

And Xiang Chen, Ching-Ti Liu, Meizhuo Zhang

Presented at BIRS workshop, Canada

YALE UNIVERSITY
School of Medicine



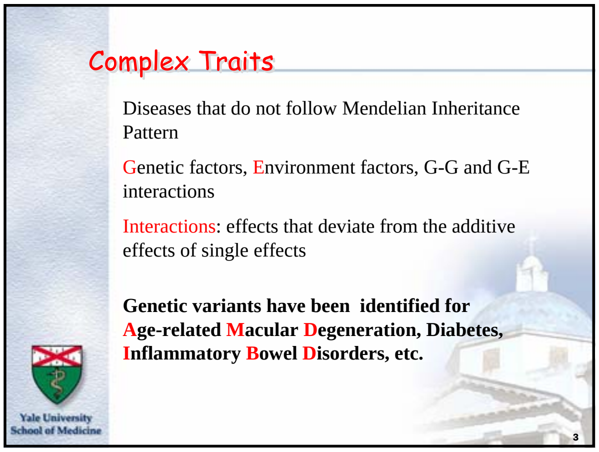
Outline

Introduction to trees and forests

Simulation studies

Genetic variants associated with age-related macular degeneration

2



Complex Traits

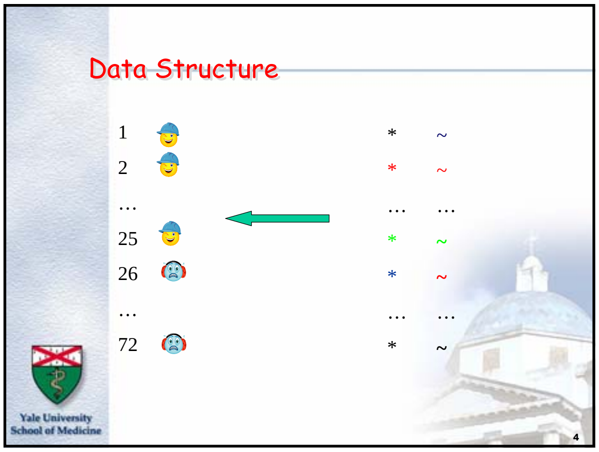
Diseases that do not follow Mendelian Inheritance Pattern

Genetic factors, Environment factors, G-G and G-E interactions

Interactions: effects that deviate from the additive effects of single effects

Genetic variants have been identified for Age-related Macular Degeneration, Diabetes, Inflammatory Bowel Disorders, etc.

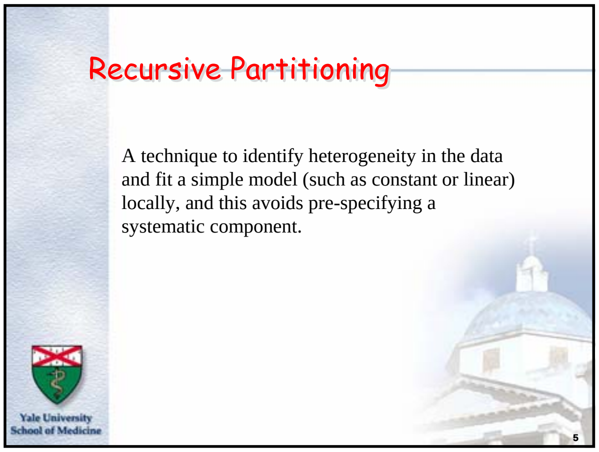
3



Data Structure

1	🙂	*	~
2	🙂	*	~
…		…	…
25	🙂	*	~
26	🙁	*	~
…		…	…
72	🙁	*	~

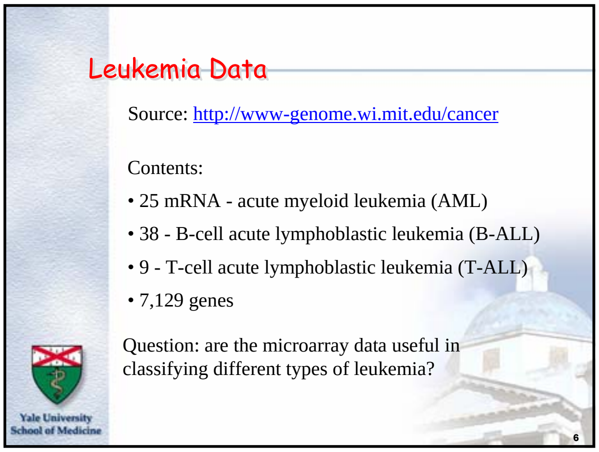
4



Recursive Partitioning

A technique to identify heterogeneity in the data and fit a simple model (such as constant or linear) locally, and this avoids pre-specifying a systematic component.

5



Leukemia Data

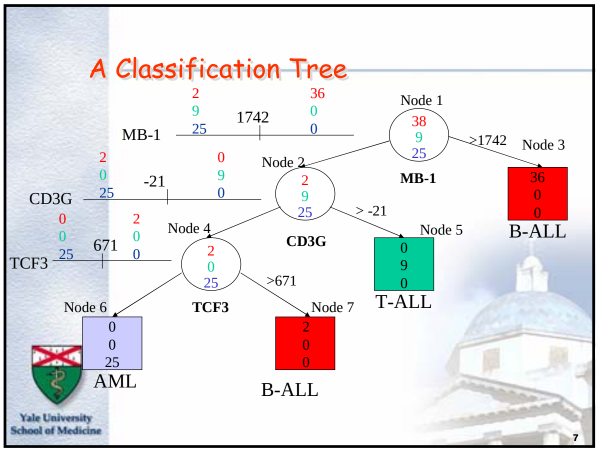
Source: http://www-genome.wi.mit.edu/cancer

Contents:

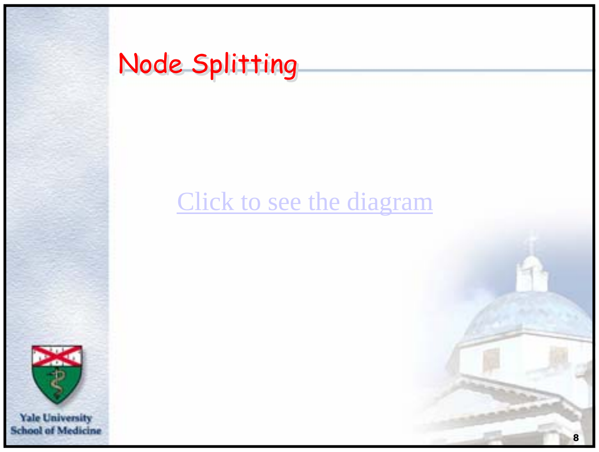
- 25 mRNA - acute myeloid leukemia (AML)
- 38 - B-cell acute lymphoblastic leukemia (B-ALL)
- 9 - T-cell acute lymphoblastic leukemia (T-ALL)
- 7,129 genes

Question: are the microarray data useful in classifying different types of leukemia?

6



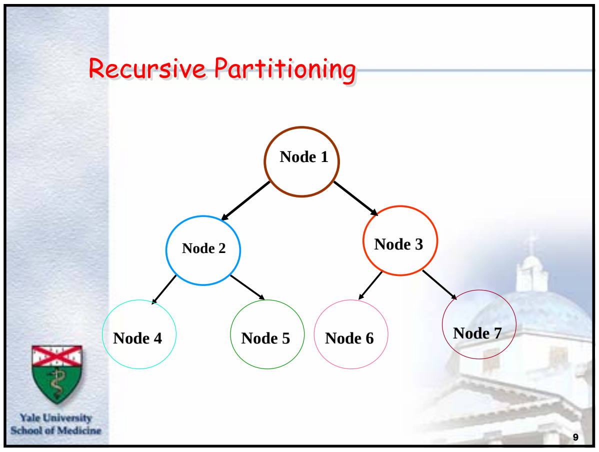
A Classification Tree



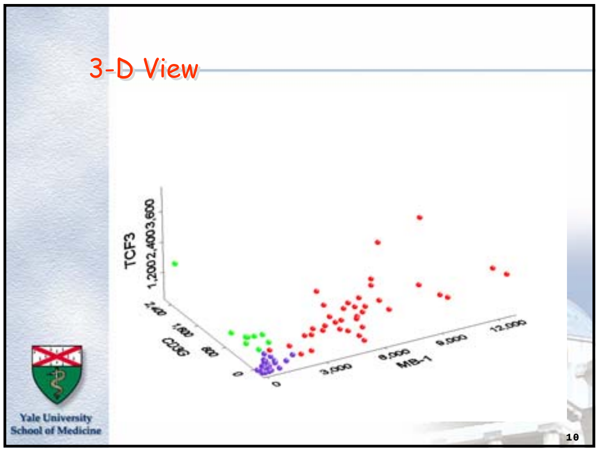
Node Splitting

Click to see the diagram

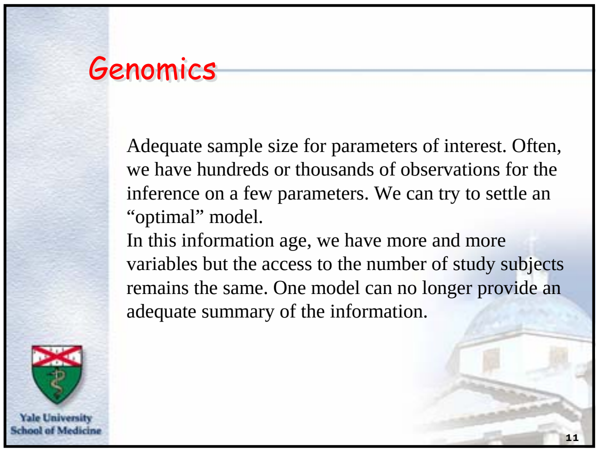
Yale University School of Medicine
8



Recursive Partitioning



3-D View

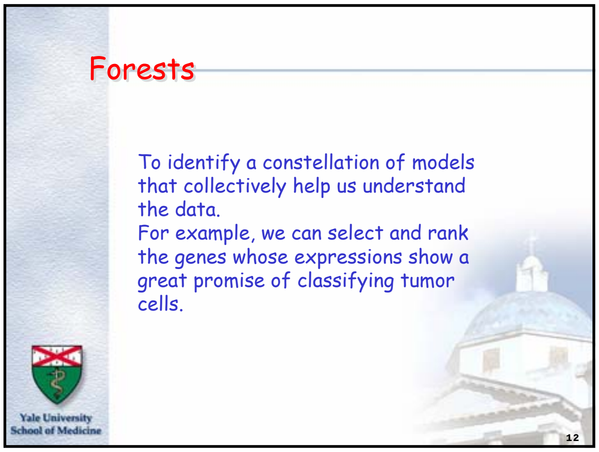


Genomics

Adequate sample size for parameters of interest. Often, we have hundreds or thousands of observations for the inference on a few parameters. We can try to settle an “optimal” model.

In this information age, we have more and more variables but the access to the number of study subjects remains the same. One model can no longer provide an adequate summary of the information.

Yale University School of Medicine
11

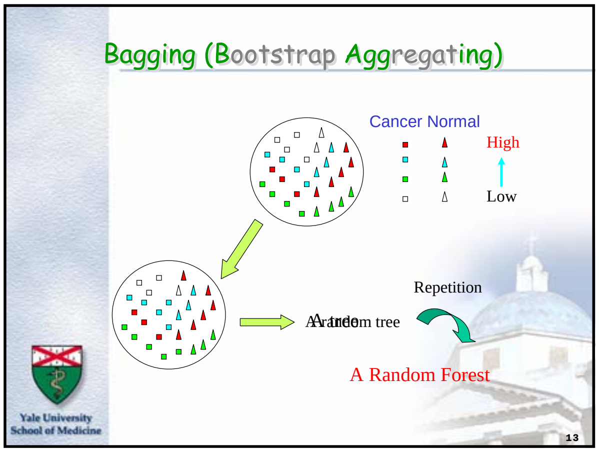


Forests

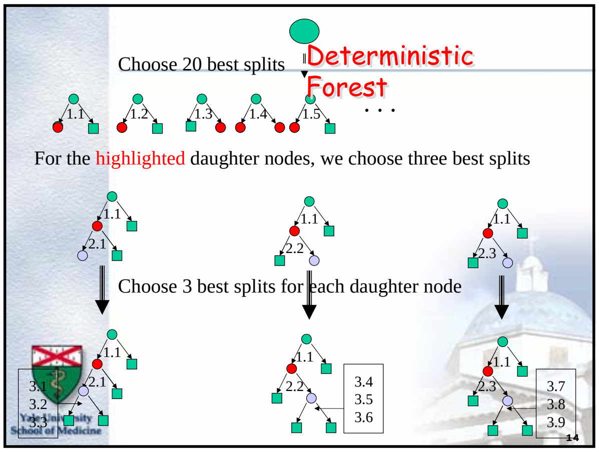
To identify a constellation of models that collectively help us understand the data.

For example, we can select and rank the genes whose expressions show a great promise of classifying tumor cells.

Yale University School of Medicine
12



Bagging (Bootstrap Aggregating)

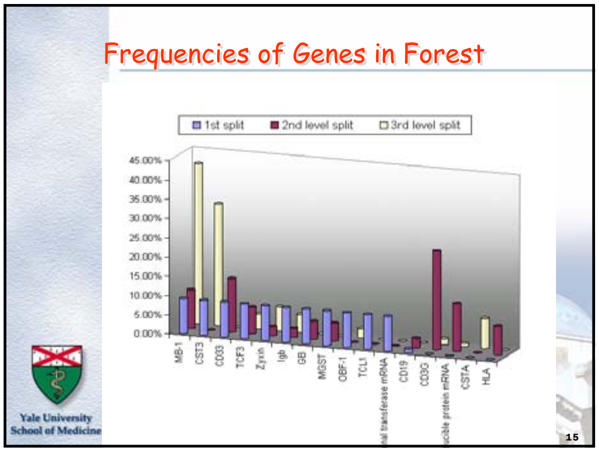


Deterministic Forest

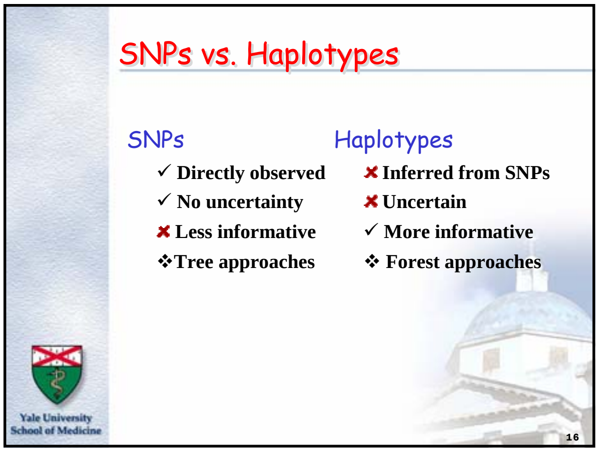
Choose 20 best splits

For the highlighted daughter nodes, we choose three best splits

Choose 3 best splits for each daughter node



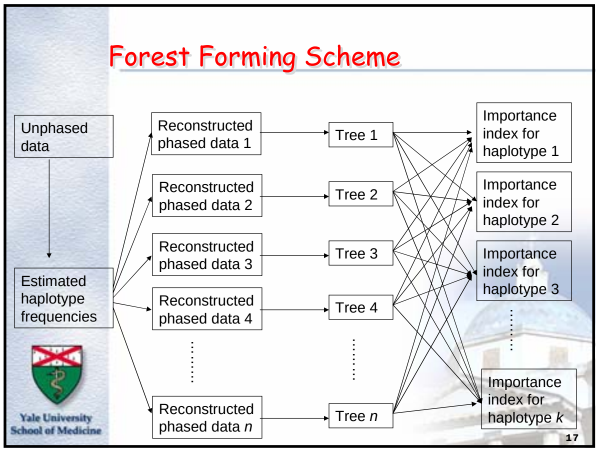
Frequencies of Genes in Forest



SNPs vs. Haplotypes

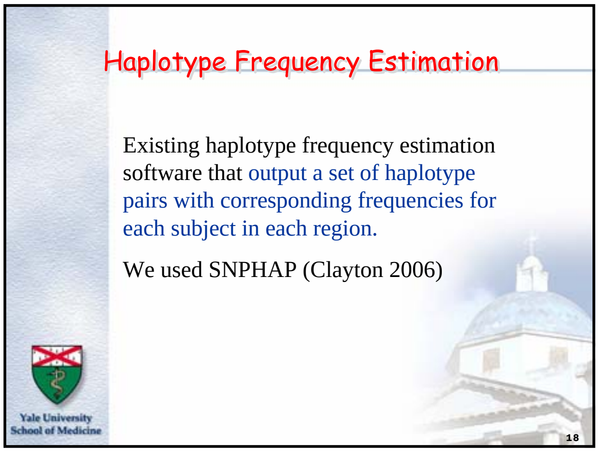
SNPs	Haplotypes
✓ Directly observed	✗ Inferred from SNPs
✓ No uncertainty	✗ Uncertain
✗ Less informative	✓ More informative
❖ Tree approaches	❖ Forest approaches

Yale University School of Medicine
16



Forest Forming Scheme

Unphased data
Estimated haplotype frequencies
Reconstructed phased data 1
Reconstructed phased data 2
Reconstructed phased data 3
Reconstructed phased data 4
Reconstructed phased data *n*
Tree 1
Tree 2
Tree 3
Tree 4
Tree *n*
Importance index for haplotype 1
Importance index for haplotype 2
Importance index for haplotype 3
Importance index for haplotype *k*
Yale University School of Medicine
17

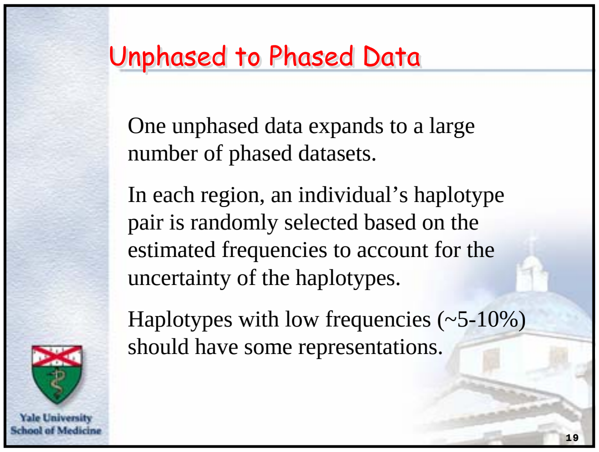


Haplotype Frequency Estimation

Existing haplotype frequency estimation software that output a set of haplotype pairs with corresponding frequencies for each subject in each region.

We used SNPHAP (Clayton 2006)

Yale University School of Medicine
18

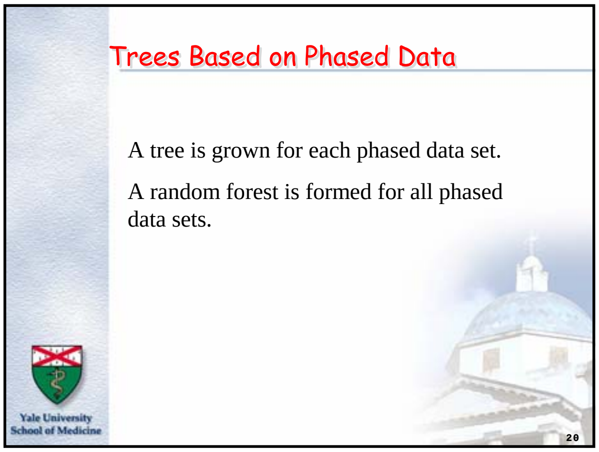


Unphased to Phased Data

One unphased data expands to a large number of phased datasets.

In each region, an individual's haplotype pair is randomly selected based on the estimated frequencies to account for the uncertainty of the haplotypes.

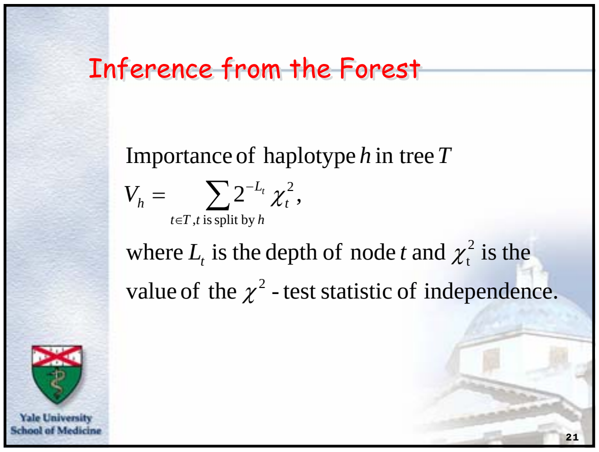
Haplotypes with low frequencies (~5-10%) should have some representations.



Trees Based on Phased Data

A tree is grown for each phased data set.

A random forest is formed for all phased data sets.

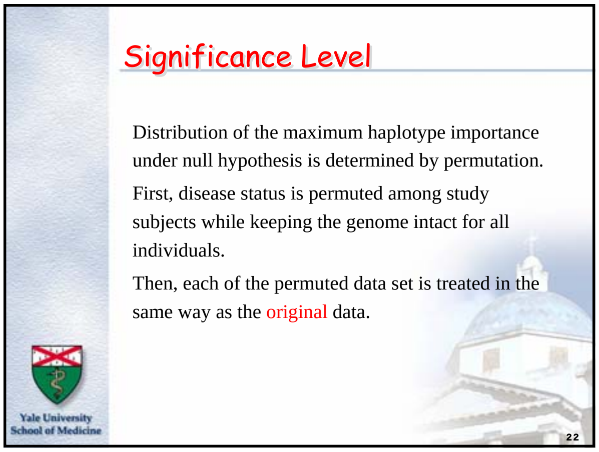


Inference from the Forest

Importance of haplotype h in tree T

$$V_h = \sum_{t\in T, t \text{ is split by } h} 2^{-L_t}\chi_t^2,$$

where L_t is the depth of node t and χ_t^2 is the value of the χ^2 - test statistic of independence.

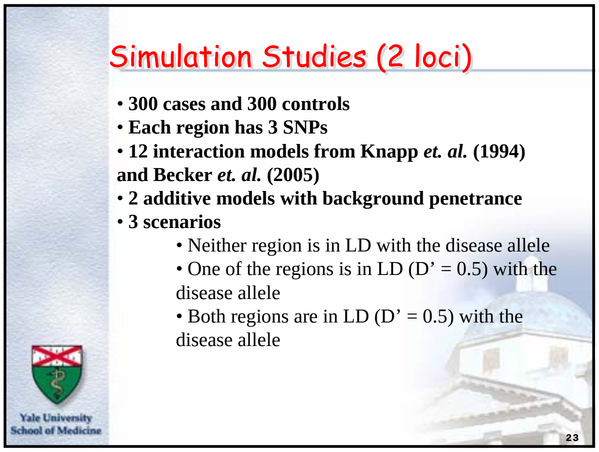


Significance Level

Distribution of the maximum haplotype importance under null hypothesis is determined by permutation.

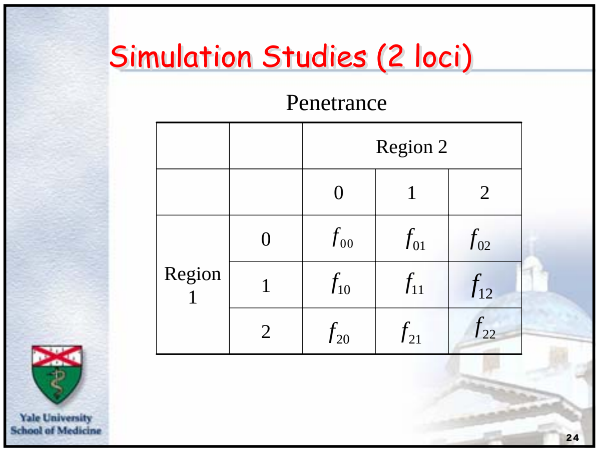
First, disease status is permuted among study subjects while keeping the genome intact for all individuals.

Then, each of the permuted data set is treated in the same way as the original data.



Simulation Studies (2 loci)

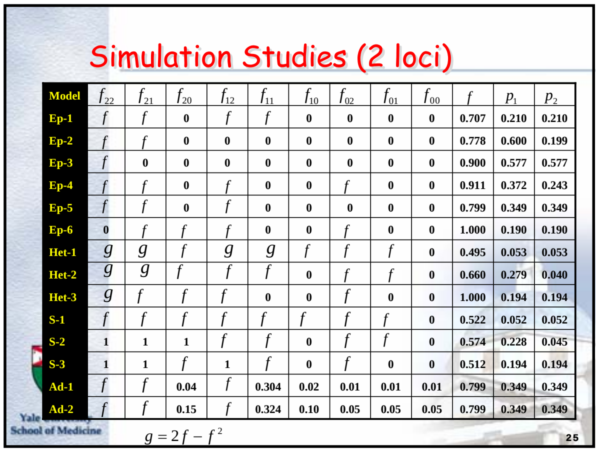
- **300 cases and 300 controls**
- **Each region has 3 SNPs**
- **12 interaction models from Knapp *et. al.* (1994) and Becker *et. al.* (2005)**
- **2 additive models with background penetrance**
- **3 scenarios**
  - Neither region is in LD with the disease allele
  - One of the regions is in LD (D' = 0.5) with the disease allele
  - Both regions are in LD (D' = 0.5) with the disease allele



Simulation Studies (2 loci)

Penetrance

		Region 2		
		0	1	2
Region 1	0	f_{00}	f_{01}	f_{02}
	1	f_{10}	f_{11}	f_{12}
	2	f_{20}	f_{21}	f_{22}

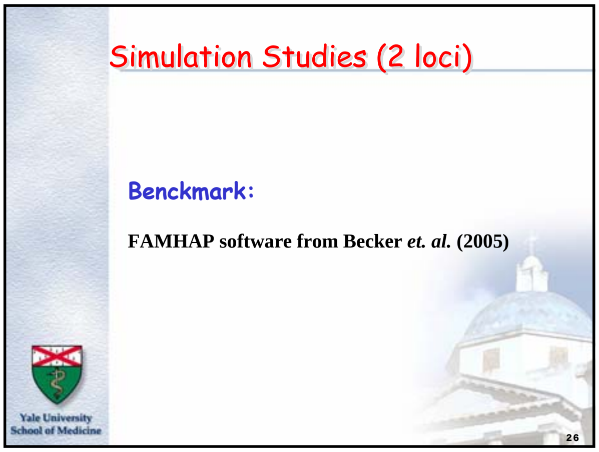


Simulation Studies (2 loci)

Model	f_{22}	f_{21}	f_{20}	f_{12}	f_{11}	f_{10}	f_{02}	f_{01}	f_{00}	f	p_1	p_2
Ep-1	f	f	0	f	f	0	0	0	0	0.707	0.210	0.210
Ep-2	f	f	0	0	0	0	0	0	0	0.778	0.600	0.199
Ep-3	f	0	0	0	0	0	0	0	0	0.900	0.577	0.577
Ep-4	f	f	0	f	0	0	f	0	0	0.911	0.372	0.243
Ep-5	f	f	0	f	0	0	0	0	0	0.799	0.349	0.349
Ep-6	0	f	f	f	0	0	f	0	0	1.000	0.190	0.190
Het-1	g	g	f	g	g	f	f	f	0	0.495	0.053	0.053
Het-2	g	g	f	f	f	0	f	f	0	0.660	0.279	0.040
Het-3	g	f	f	f	0	0	f	0	0	1.000	0.194	0.194
S-1	f	f	f	f	f	f	f	f	0	0.522	0.052	0.052
S-2	1	1	1	f	f	0	f	f	0	0.574	0.228	0.045
S-3	1	1	f	1	f	0	f	0	0	0.512	0.194	0.194
Ad-1	f	f	0.04	f	0.304	0.02	0.01	0.01	0.01	0.799	0.349	0.349
Ad-2	f	f	0.15	f	0.324	0.10	0.05	0.05	0.05	0.799	0.349	0.349

$g = 2f - f^2$

25

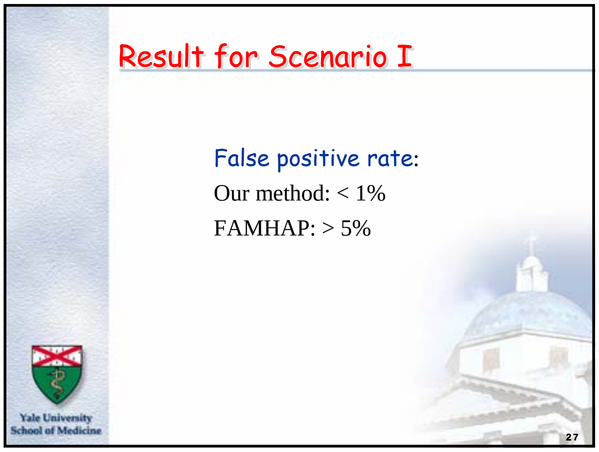


Simulation Studies (2 loci)

Benckmark:

FAMHAP software from Becker *et. al.* (2005)

26



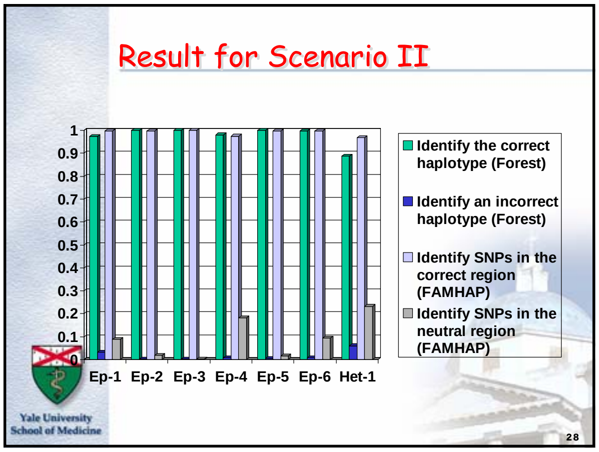
Result for Scenario I

False positive rate:

Our method: < 1%

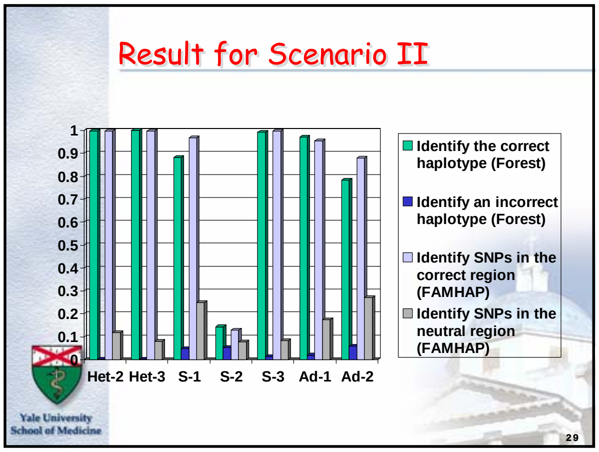
FAMHAP: > 5%

27



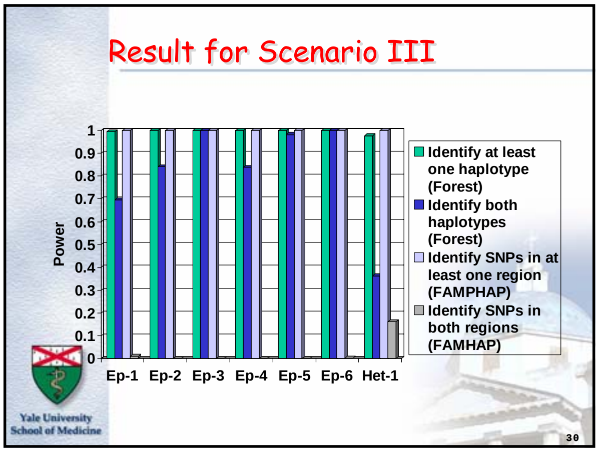
Result for Scenario II

28



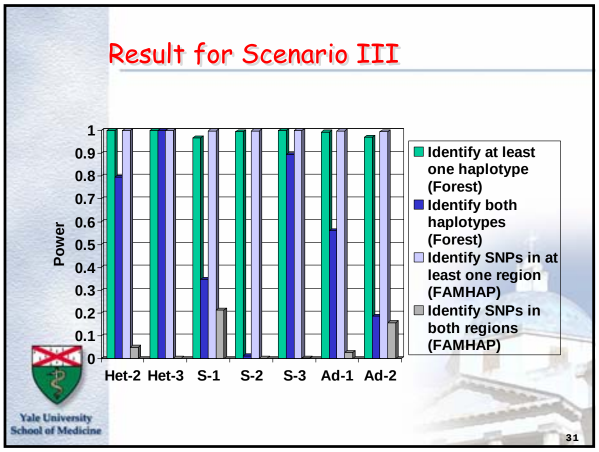
Result for Scenario II

29

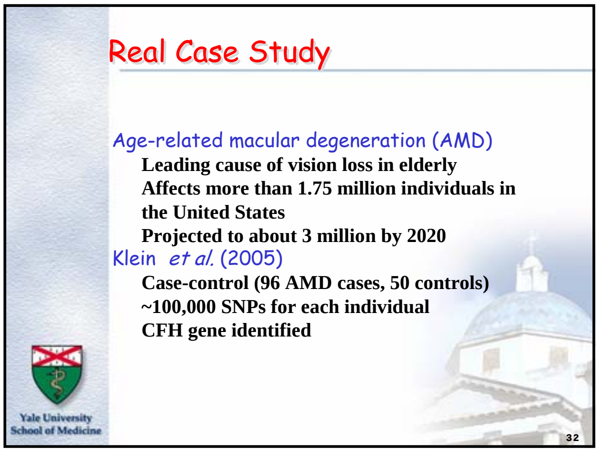


Result for Scenario III

30



Result for Scenario III



Real Case Study

Age-related macular degeneration (AMD)

Leading cause of vision loss in elderly

Affects more than 1.75 million individuals in the United States

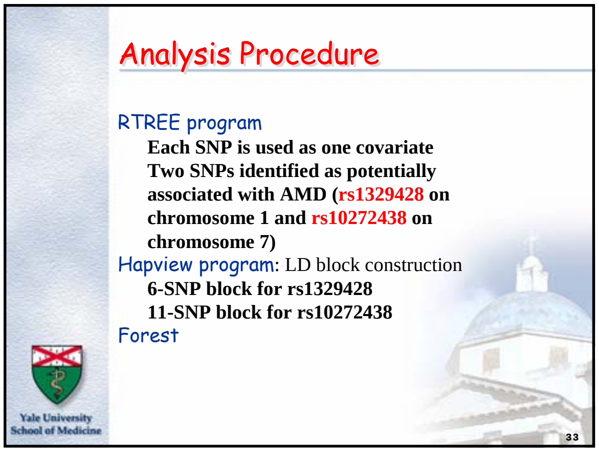
Projected to about 3 million by 2020

Klein *et al.* (2005)

Case-control (96 AMD cases, 50 controls)

~100,000 SNPs for each individual

CFH gene identified



Analysis Procedure

RTREE program

Each SNP is used as one covariate

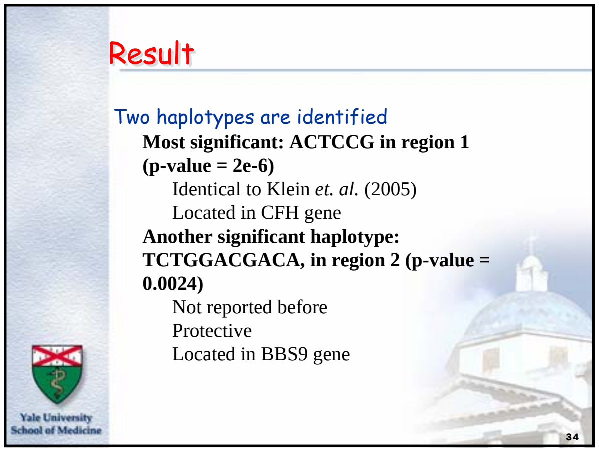
Two SNPs identified as potentially associated with AMD (rs1329428 on chromosome 1 and rs10272438 on chromosome 7)

Hapview program: LD block construction

6-SNP block for rs1329428

11-SNP block for rs10272438

Forest



Result

Two haplotypes are identified

Most significant: ACTCCG in region 1 (p-value = 2e-6)

Identical to Klein *et. al.* (2005)

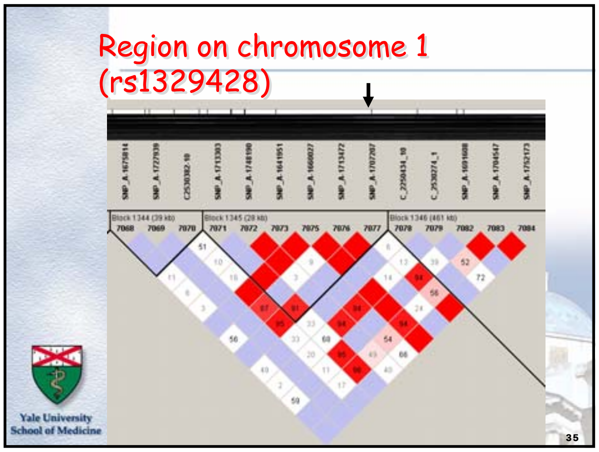
Located in CFH gene

Another significant haplotype: TCTGGACGACA, in region 2 (p-value = 0.0024)

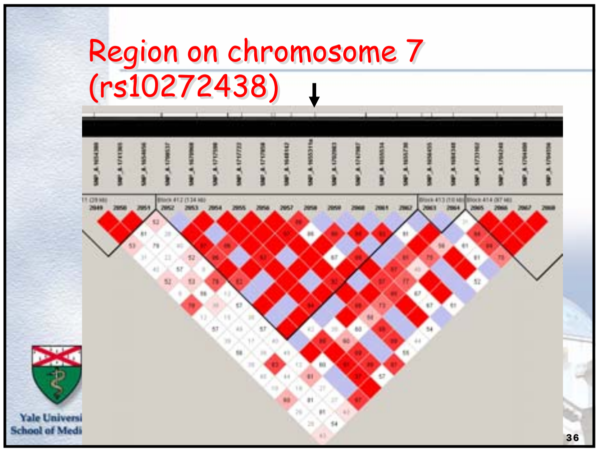
Not reported before

Protective

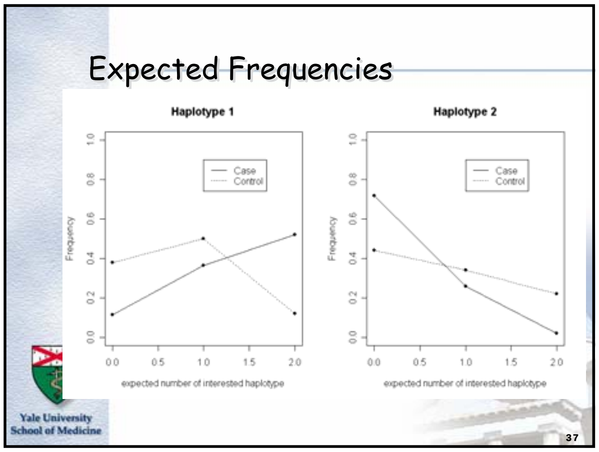
Located in BBS9 gene



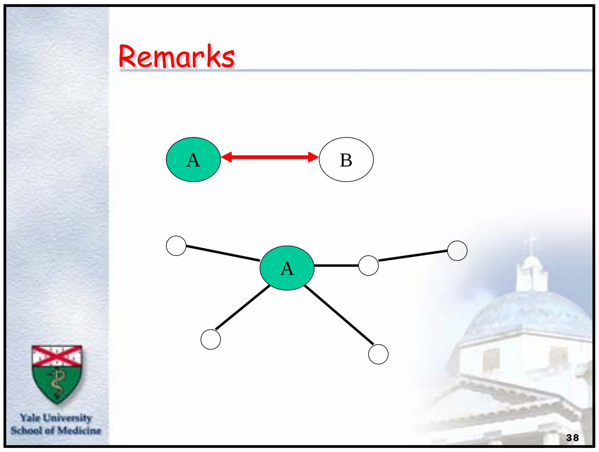
Region on chromosome 1 (rs1329428)



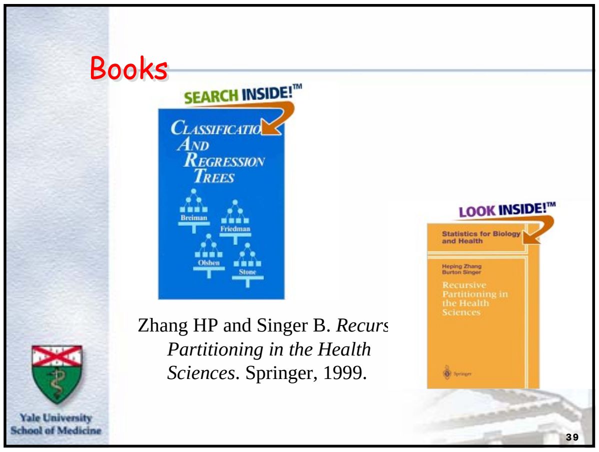
Region on chromosome 7 (rs10272438)



Expected Frequencies



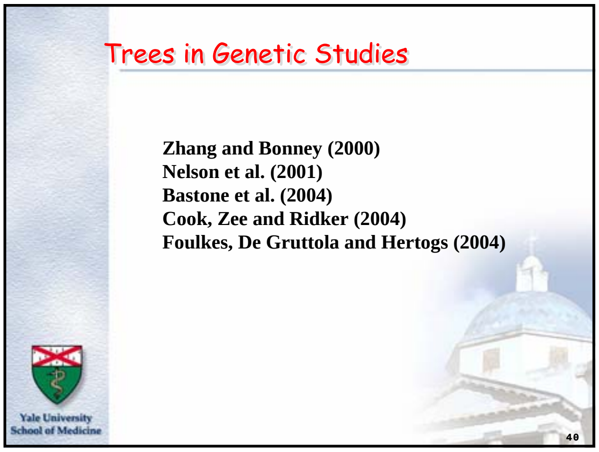
Remarks



Books

Zhang HP and Singer B. *Recurs Partitioning in the Health Sciences*. Springer, 1999.

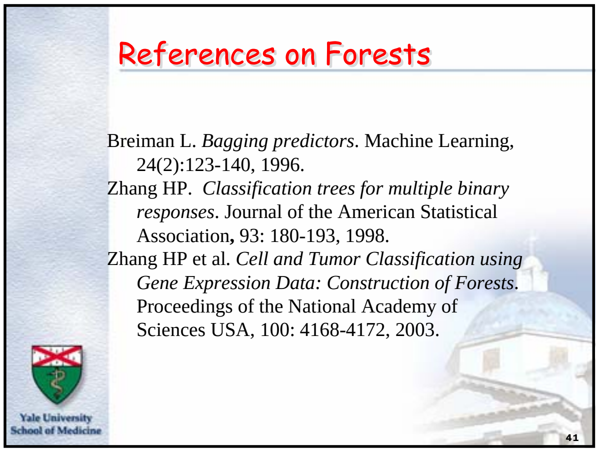
SEARCH INSIDE!™
Classification And Regression Trees
Breiman
Friedman
Olshen
Stone
LOOK INSIDE!™
Statistics for Biology and Health
Heping Zhang Burton Singer
Recursive Partitioning in the Health Sciences
Springer
Yale University School of Medicine
39



Trees in Genetic Studies

Zhang and Bonney (2000)
Nelson et al. (2001)
Bastone et al. (2004)
Cook, Zee and Ridker (2004)
Foulkes, De Gruttola and Hertogs (2004)

Yale University School of Medicine
40



References on Forests

Breiman L. *Bagging predictors*. Machine Learning, 24(2):123-140, 1996.

Zhang HP. *Classification trees for multiple binary responses*. Journal of the American Statistical Association**,** 93: 180-193, 1998.

Zhang HP et al. *Cell and Tumor Classification using Gene Expression Data: Construction of Forests*. Proceedings of the National Academy of Sciences USA, 100: 4168-4172, 2003.