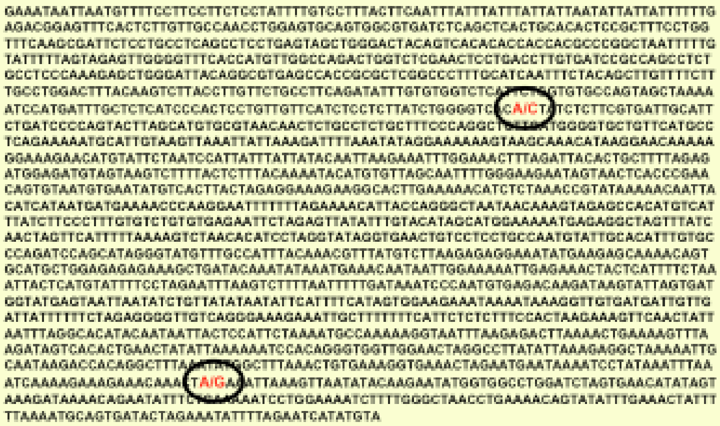
Bayesian Inference of Interactions and Associations

Jun Liu
Department of Statistics
Harvard University

http://www.fas.harvard.edu/~junliu

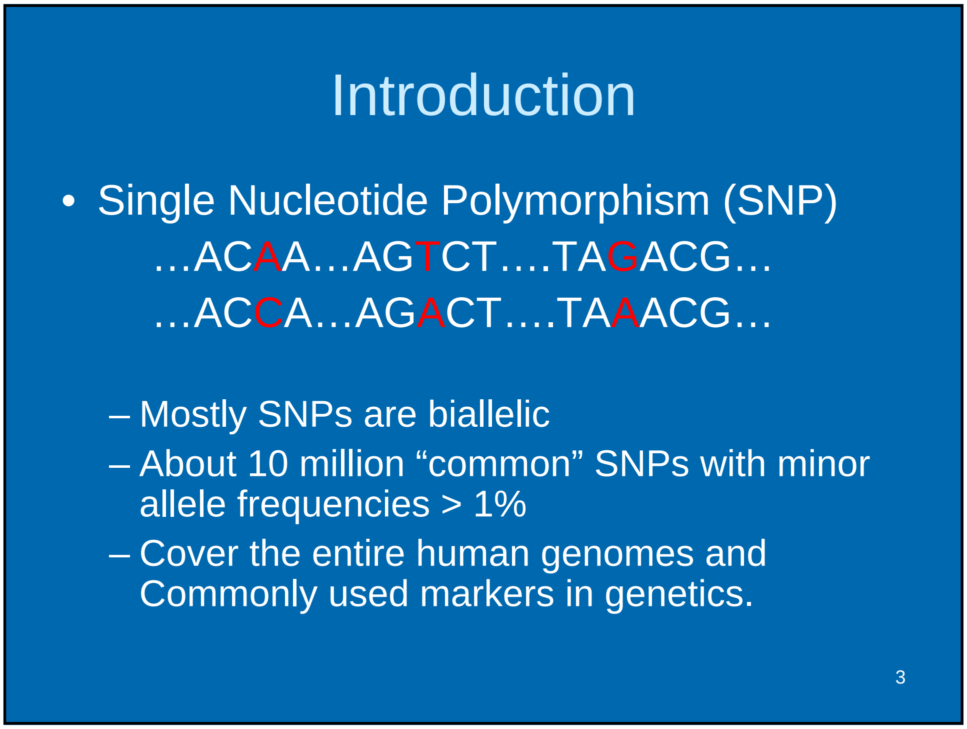
Based on collaborations with Yu Zhang, Jing Zhang, Yuan Yuan, Ke Deng, Zhi Geng

A segment of Chromosome 7 of two random individuals compared



GAAATAATTAATGTTTTCCTTCCTTCTCCTATTTTGTCCTTTACTTCAATTTATTTATTTATTATTAATATTATTATTTTTTG
AGACGGAGTTTCACTCTTGTTGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCACACTCCGCTTTCCTGG
TTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGACTACAGTCACACACCACCACGCCCGGCTAATTTTTG
TATTTTTAGTAGAGTTGGGGTTTCACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCAGCCTCT
GCCTCCCAAAGAGCTGGGATTACAGGCGTGAGCCACCGCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTTTCTT
TGCCTGGACTTTACAAGTCTTACCTTGTTCTGCCTTCAGATATTTGTGTGGTCTCA[illegible]GTGTGCCAGTAGCTAAAA
ATCCATGATTTGCTCTCATCCCACTCCTGTTGTTCATCTCCTCTTATCTGGGGTC[illegible]CA/CT[illegible]TCTCTTCGTGATTGCATT
CTGATCCCCAGTACTTAGCATGTGCGTAACAACTCTGCCTCTGCTTTCCCAGGCT[illegible]TGGGGTGCTGTTCATGCC
TCAGAAAAATGCATTGTAAGTTAAATTATTAAAGATTTTAAATATAGGAAAAAAGTAAGCAAACATAAGGAACAAAAA
GGAAAGAACATGTATTCTAATCCATTATTTATTATACAATTAAGAAATTTGGAAACTTTAGATTACACTGCTTTTAGAG
ATGGAGATGTAGTAAGTCTTTTACTCTTTACAAAATACATGTGTTAGCAATTTTGGGAAGAATAGTAACTCACCCGAA
CAGTGTAATGTGAATATGTCACTTACTAGAGGAAAGAAGGCACTTGAAAAACATCTCTAAACCGTATAAAAACAATTA
CATCATAATGATGAAAACCCAAGGAATTTTTTTAGAAAACATTACCAGGGCTAATAACAAAGTAGAGCCACATGTCAT
TTATCTTCCCTTTGTGTCTGTGTGAGAATTCTAGAGTTATATTTGTACATAGCATGGAAAAATGAGAGGCTAGTTTATC
AACTAGTTCATTTTTAAAAGTCTAACACATCCTAGGTATAGGTGAACTGTCCTCCTGCCAATGTATTGCACATTTGTGC
CCAGATCCAGCATAGGGTATGTTTGCCATTTACAAACGTTTATGTCTTAAGAGAGGAAATATGAAGAGCAAAACAGT
GCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTGGAAAAATTGAGAAACTACTCATTTTCTAA
ATTACTCATGTATTTTCCTAGAATTTAAGTCTTTTAATTTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGAT
GGTATGAGTAATTAATATCTGTTATATAATATTCATTTTCATAGTGGAAGAAATAAAATAAAGGTTGTGATGATTGTTG
ATTATTTTTTCTAGAGGGGTTGTCAGGGAAAGAAATTGCTTTTTTTCATTCTCTCTTTCCACTAAGAAAGTTCAACTATT
AATTTAGGCACATACAATAATTACTCCATTCTAAAATGCCAAAAAGGTAATTTAAGAGACTTAAAACTGAAAAGTTTA
AGATAGTCACACTGAACTATATTAAAAAATCCACAGGGTGGTTGGAACTAGGCCTTATATTAAAGAGGCTAAAAATTG
CAATAAGACCACAGGCTTTAA[illegible]GCTTTAAACTGTGAAAGGTGAAACTAGAATGAATAAAATCCTATAAATTTAA
ATCAAAAGAAAGAAACAAA[illegible]TA/GA[illegible]ATTAAAGTTAATATACAAGAATATGGTGGCCTGGATCTAGTGAACATATAGT
AAAGATAAAACAGAATATTTC[illegible]AATCCTGGAAAATCTTTTGGGCTAACCTGAAAACAGTATATTTGAAACTATTT
TTAAAATGCAGTGATACTAGAAATATTTTAGAATCATATGTA

2200 base pairs



Introduction

- Single Nucleotide Polymorphism (SNP)

  …ACAA…AGTCT….TAGACG…

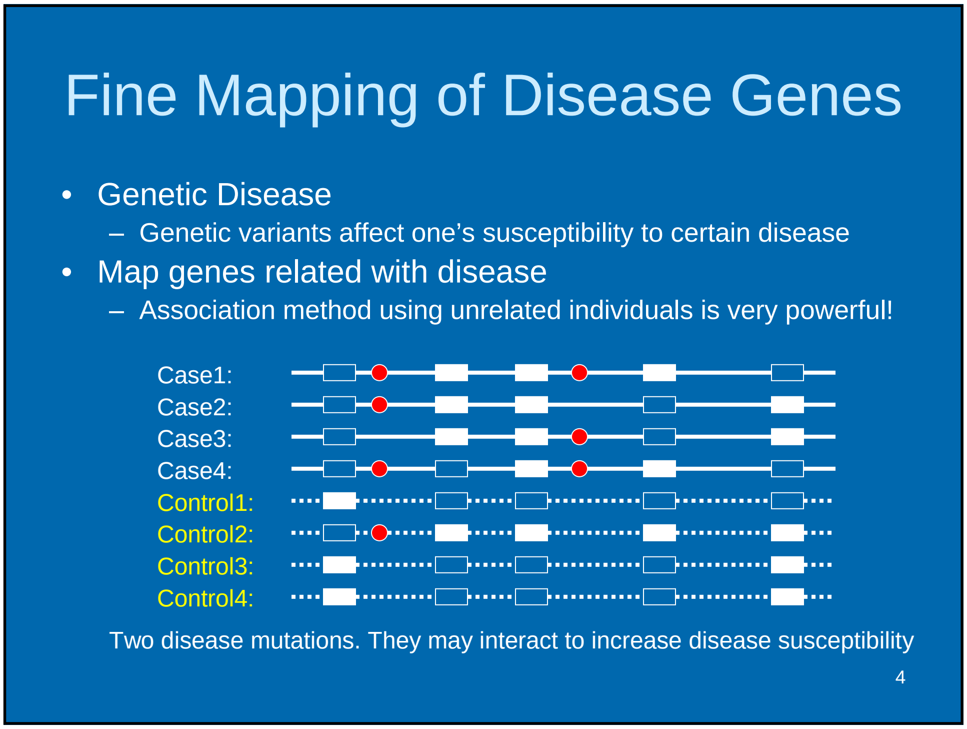
  …ACCA…AGACT….TAAACG…

  – Mostly SNPs are biallelic

  – About 10 million “common” SNPs with minor allele frequencies > 1%

  – Cover the entire human genomes and Commonly used markers in genetics.

3

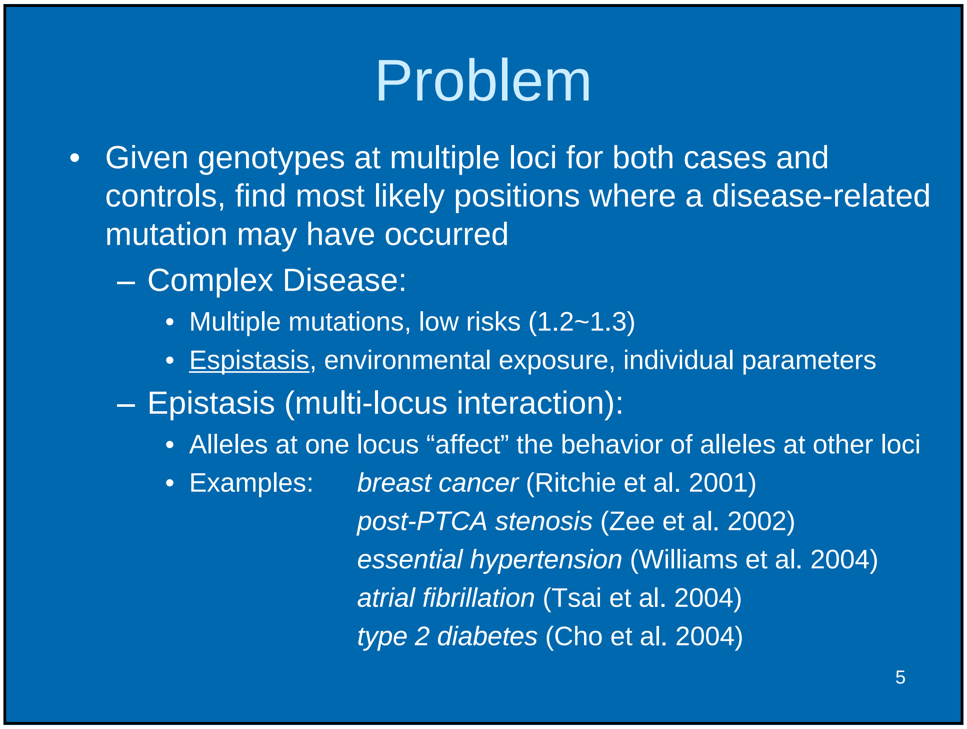


Fine Mapping of Disease Genes

- Genetic Disease
  – Genetic variants affect one’s susceptibility to certain disease
- Map genes related with disease
  – Association method using unrelated individuals is very powerful!

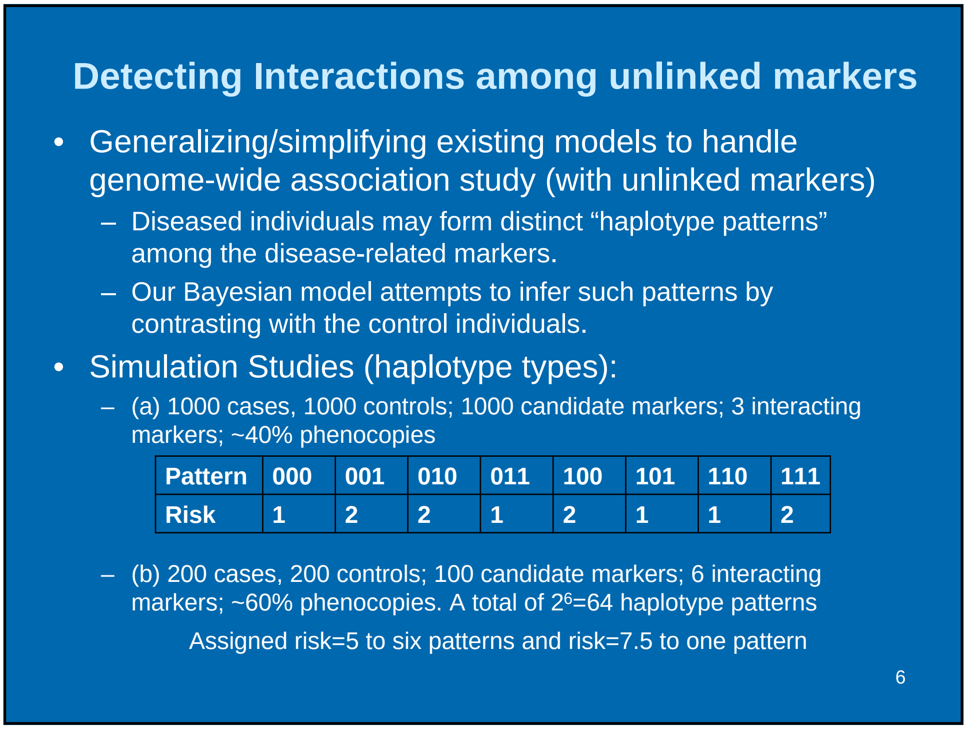
Two disease mutations. They may interact to increase disease susceptibility

4



Problem

- Given genotypes at multiple loci for both cases and controls, find most likely positions where a disease-related mutation may have occurred
  - Complex Disease:
    - Multiple mutations, low risks (1.2~1.3)
    - Espistasis, environmental exposure, individual parameters
  - Epistasis (multi-locus interaction):
    - Alleles at one locus “affect” the behavior of alleles at other loci
    - Examples: *breast cancer* (Ritchie et al. 2001)
      post-PTCA stenosis (Zee et al. 2002)
      essential hypertension (Williams et al. 2004)
      atrial fibrillation (Tsai et al. 2004)
      type 2 diabetes (Cho et al. 2004)



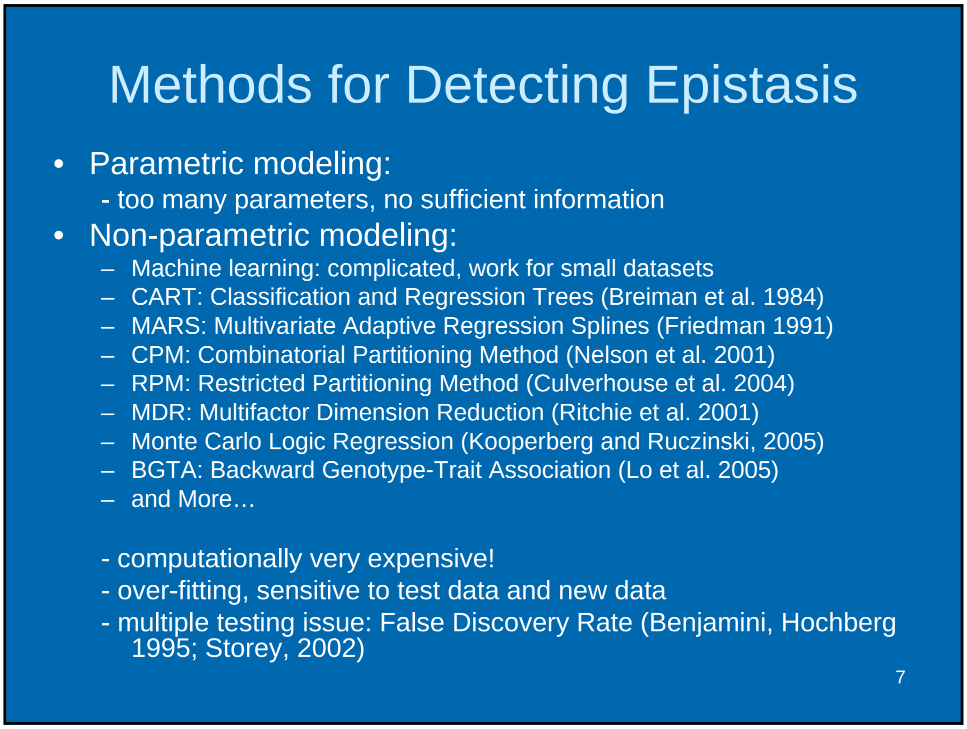
Detecting Interactions among unlinked markers

- Generalizing/simplifying existing models to handle genome-wide association study (with unlinked markers)
  - Diseased individuals may form distinct “haplotype patterns” among the disease-related markers.
  - Our Bayesian model attempts to infer such patterns by contrasting with the control individuals.
- Simulation Studies (haplotype types):
  - (a) 1000 cases, 1000 controls; 1000 candidate markers; 3 interacting markers; ~40% phenocopies

Pattern	**000**	**001**	**010**	**011**	**100**	**101**	**110**	**111**
Risk	**1**	**2**	**2**	**1**	**2**	**1**	**1**	**2**

  - (b) 200 cases, 200 controls; 100 candidate markers; 6 interacting markers; ~60% phenocopies. A total of 2^6=64 haplotype patterns

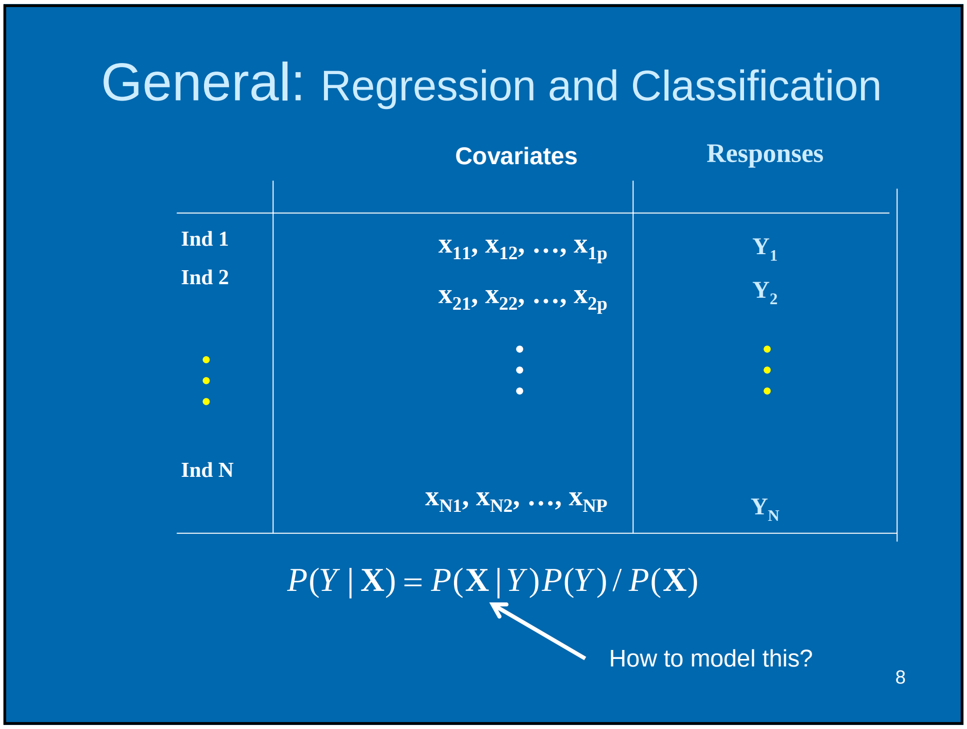
    Assigned risk=5 to six patterns and risk=7.5 to one pattern



Methods for Detecting Epistasis

- Parametric modeling:
  - too many parameters, no sufficient information
- Non-parametric modeling:
  - Machine learning: complicated, work for small datasets
  - CART: Classification and Regression Trees (Breiman et al. 1984)
  - MARS: Multivariate Adaptive Regression Splines (Friedman 1991)
  - CPM: Combinatorial Partitioning Method (Nelson et al. 2001)
  - RPM: Restricted Partitioning Method (Culverhouse et al. 2004)
  - MDR: Multifactor Dimension Reduction (Ritchie et al. 2001)
  - Monte Carlo Logic Regression (Kooperberg and Ruczinski, 2005)
  - BGTA: Backward Genotype-Trait Association (Lo et al. 2005)
  - and More…

- computationally very expensive!
- over-fitting, sensitive to test data and new data
- multiple testing issue: False Discovery Rate (Benjamini, Hochberg 1995; Storey, 2002)

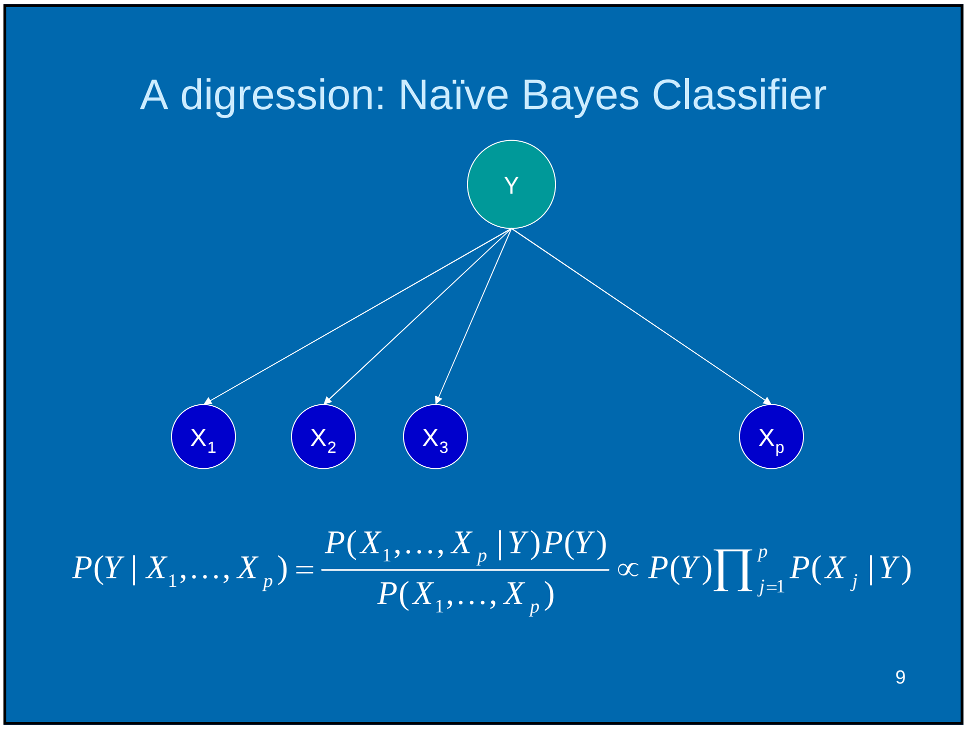


General: Regression and Classification

	Covariates	Responses
Ind 1	$x_{11}, x_{12}, \ldots, x_{1p}$	Y_1
Ind 2	$x_{21}, x_{22}, \ldots, x_{2p}$	Y_2
⋮	⋮	⋮
Ind N	$x_{N1}, x_{N2}, \ldots, x_{NP}$	Y_N

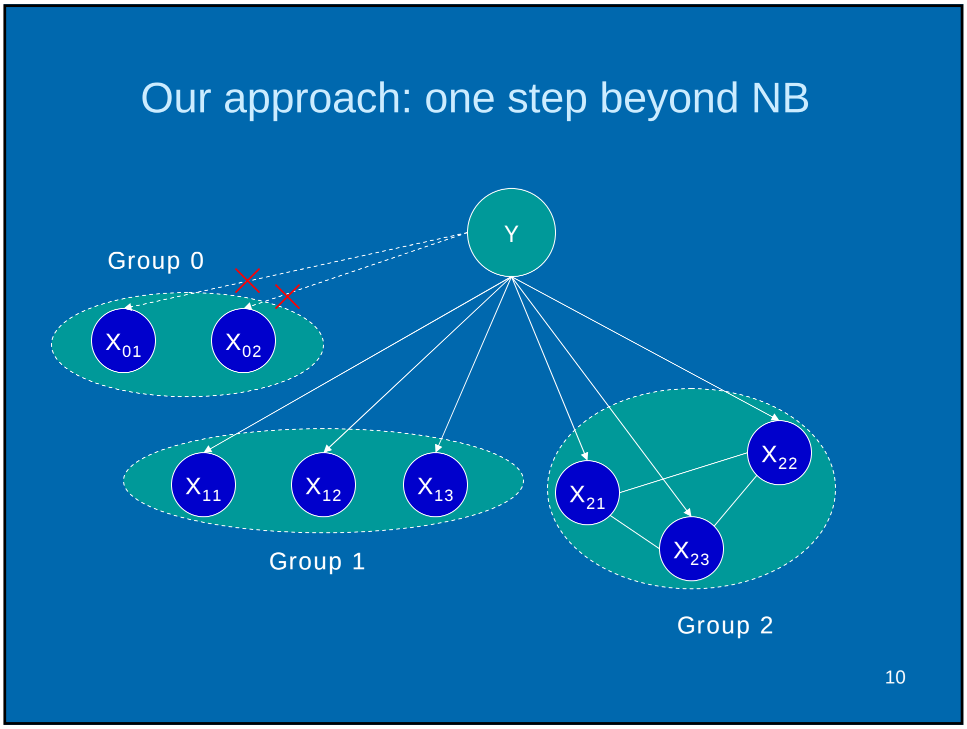
$$P(Y \mid \mathbf{X}) = P(\mathbf{X} \mid Y)P(Y)/P(\mathbf{X})$$

How to model this?



A digression: Naïve Bayes Classifier

$$P(Y\mid X_1,\ldots,X_p)=\frac{P(X_1,\ldots,X_p\mid Y)P(Y)}{P(X_1,\ldots,X_p)}\propto P(Y)\prod\nolimits_{j=1}^{p}P(X_j\mid Y)$$



Our approach: one step beyond NB

Our Approach: Beyond NB

- **Partition *L* markers into three groups**
  - Group 1: l_1 markers have marginal effects only

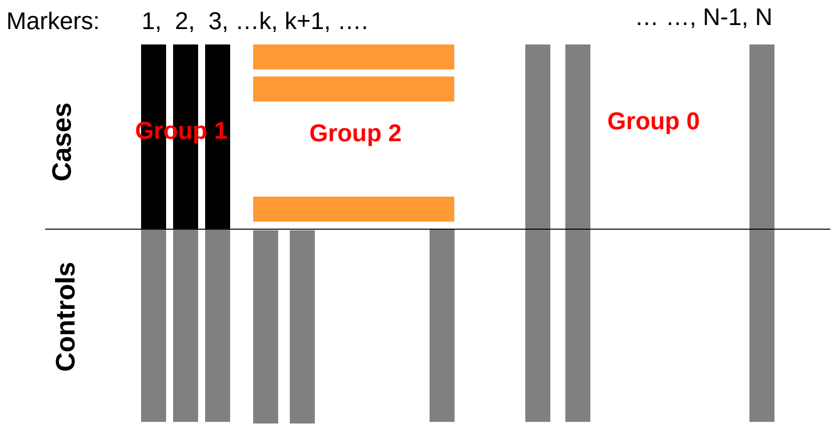
    G_1^d
    - Genotype frequencies are different between cases and controls
  - Group 2: l_2 markers have epistasis effect

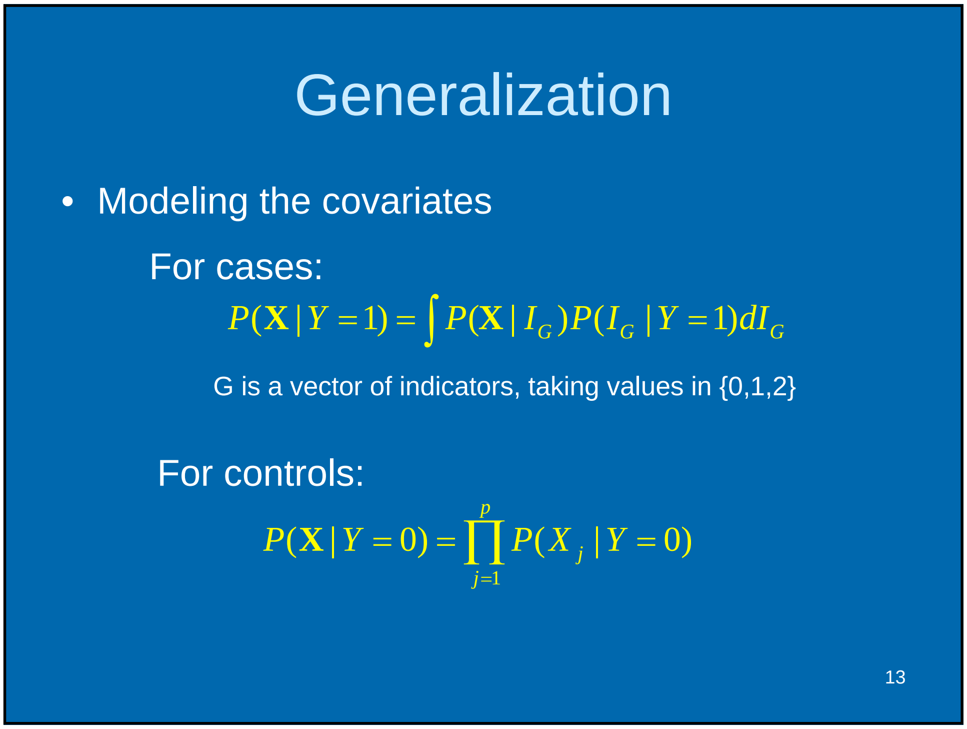
    G_2^d
    - Genotypes of markers are correlated, consider a vector of genotypes with unknown frequencies $\{\rho\}_{1\ldots3^{l_2}}$
    - Different from multiplication of single marker frequencies

      $\rho = f_1 \times \cdots \times f_{l_2}$
  - Group 0: L- l_1 - l_2 markers have no association

    G_0^d
    - Genotype frequencies are the same between cases and controls
    - 1st-order Markov chain to account for Linkage Disequilibrium

Another graphical Illustration





Generalization

- Modeling the covariates

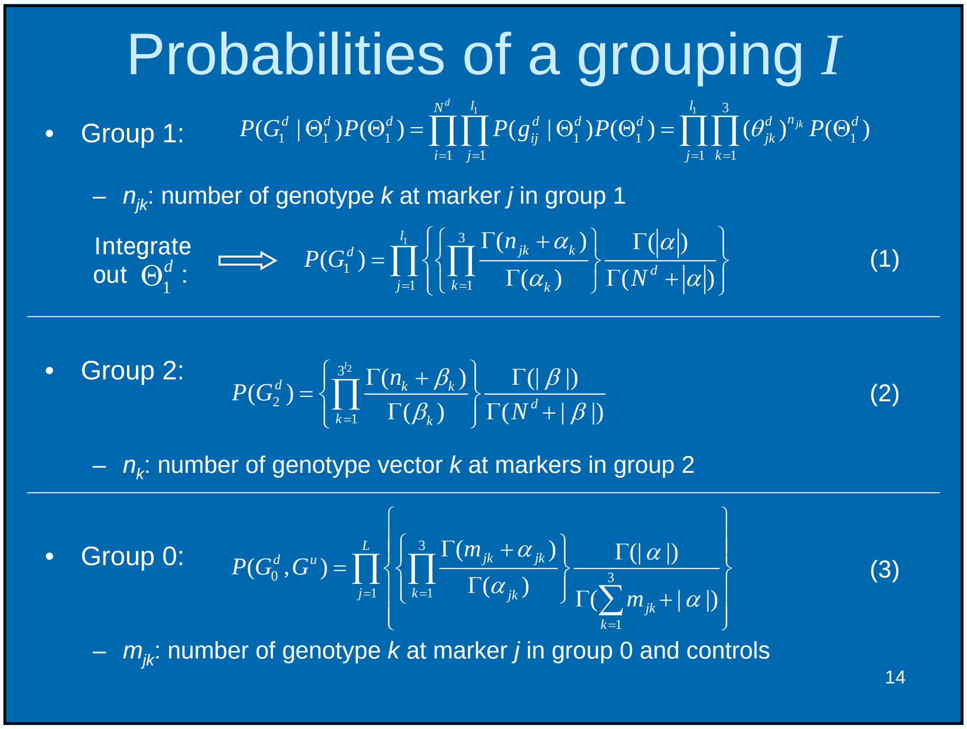
For cases:

$$P(\mathbf{X} \mid Y=1) = \int P(\mathbf{X} \mid I_G) P(I_G \mid Y=1) dI_G$$

G is a vector of indicators, taking values in {0,1,2}

For controls:

$$P(\mathbf{X} \mid Y=0) = \prod_{j=1}^{p} P(X_j \mid Y=0)$$



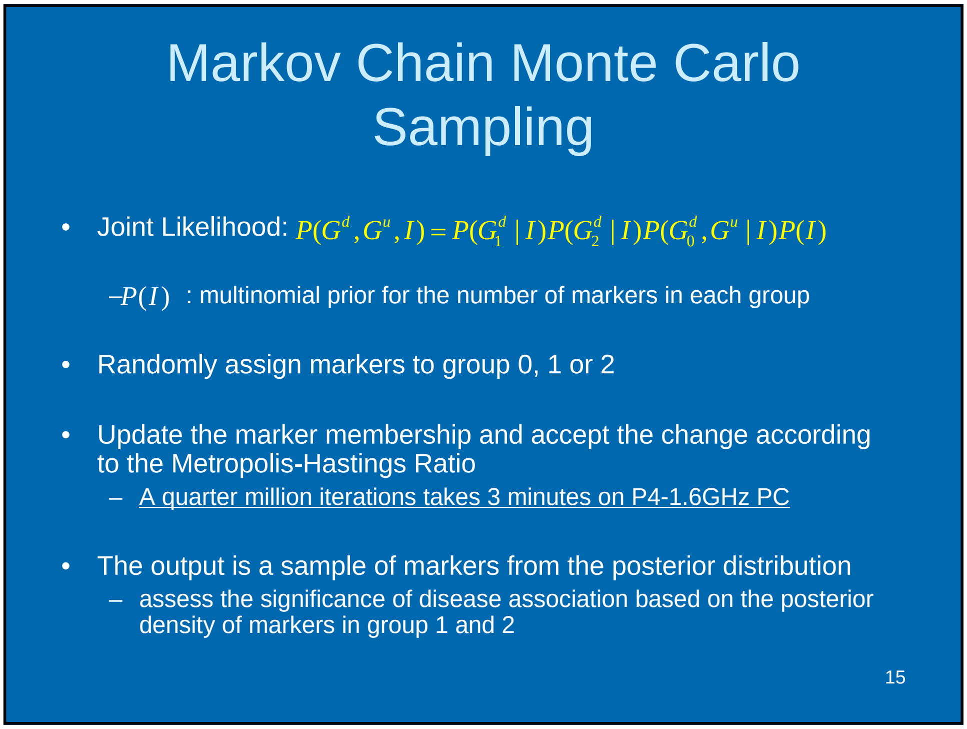
Probabilities of a grouping I

- Group 1: $$P(G_1^d \mid \Theta_1^d) P(\Theta_1^d) = \prod_{i=1}^{N^d} \prod_{j=1}^{l_1} P(g_{ij}^d \mid \Theta_1^d) P(\Theta_1^d) = \prod_{j=1}^{l_1} \prod_{k=1}^{3} (\theta_{jk}^d)^{n_{jk}} P(\Theta_1^d)$$
  - n_{jk}: number of genotype *k* at marker *j* in group 1

Integrate out Θ_1^d : $\Longrightarrow$ $$P(G_1^d) = \prod_{j=1}^{l_1} \left\{ \left\{ \prod_{k=1}^{3} \frac{\Gamma(n_{jk} + \alpha_k)}{\Gamma(\alpha_k)} \right\} \frac{\Gamma(|\alpha|)}{\Gamma(N^d + |\alpha|)} \right\} \qquad (1)$$

- Group 2: $$P(G_2^d) = \left\{ \prod_{k=1}^{3^{l_2}} \frac{\Gamma(n_k + \beta_k)}{\Gamma(\beta_k)} \right\} \frac{\Gamma(|\beta|)}{\Gamma(N^d + |\beta|)} \qquad (2)$$
  - n_k: number of genotype vector *k* at markers in group 2

- Group 0: $$P(G_0^d, G^u) = \prod_{j=1}^{L} \left\{ \left\{ \prod_{k=1}^{3} \frac{\Gamma(m_{jk} + \alpha_{jk})}{\Gamma(\alpha_{jk})} \right\} \frac{\Gamma(|\alpha|)}{\Gamma(\sum_{k=1}^{3} m_{jk} + |\alpha|)} \right\} \qquad (3)$$
  - m_{jk}: number of genotype *k* at marker *j* in group 0 and controls



Markov Chain Monte Carlo Sampling

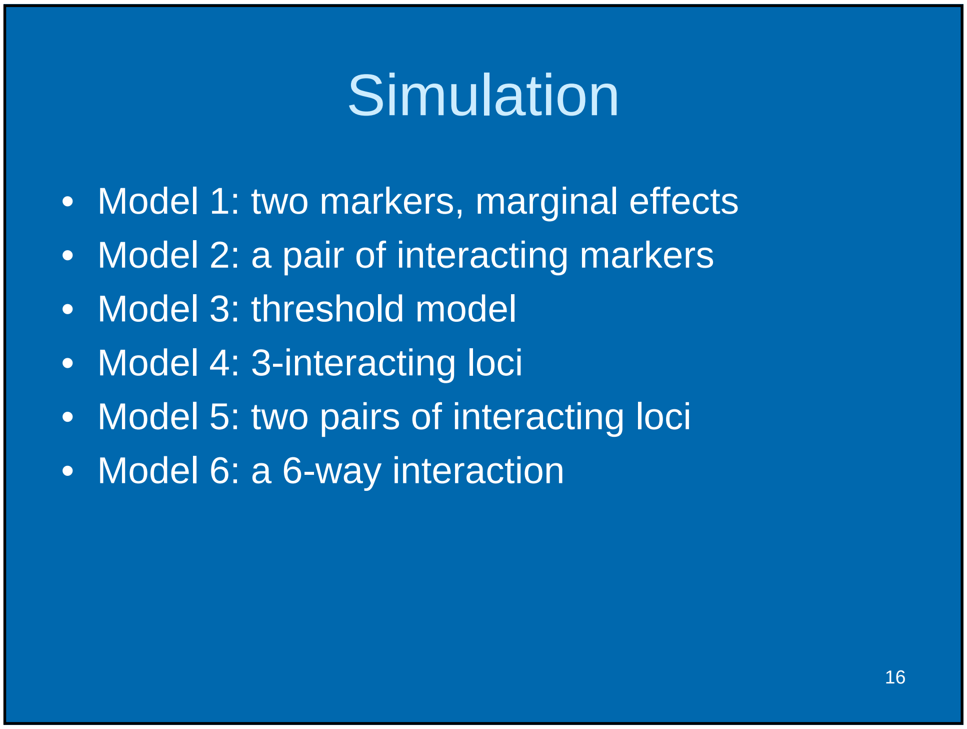
- Joint Likelihood: $P(G^d, G^u, I) = P(G_1^d \mid I)P(G_2^d \mid I)P(G_0^d, G^u \mid I)P(I)$

  – $P(I)$: multinomial prior for the number of markers in each group

- Randomly assign markers to group 0, 1 or 2

- Update the marker membership and accept the change according to the Metropolis-Hastings Ratio
  - A quarter million iterations takes 3 minutes on P4-1.6GHz PC

- The output is a sample of markers from the posterior distribution
  - assess the significance of disease association based on the posterior density of markers in group 1 and 2



Simulation

- Model 1: two markers, marginal effects
- Model 2: a pair of interacting markers
- Model 3: threshold model
- Model 4: 3-interacting loci
- Model 5: two pairs of interacting loci
- Model 6: a 6-way interaction

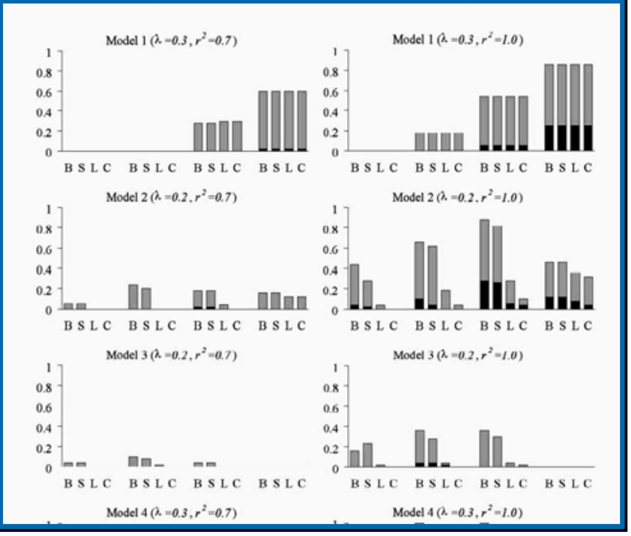
Results

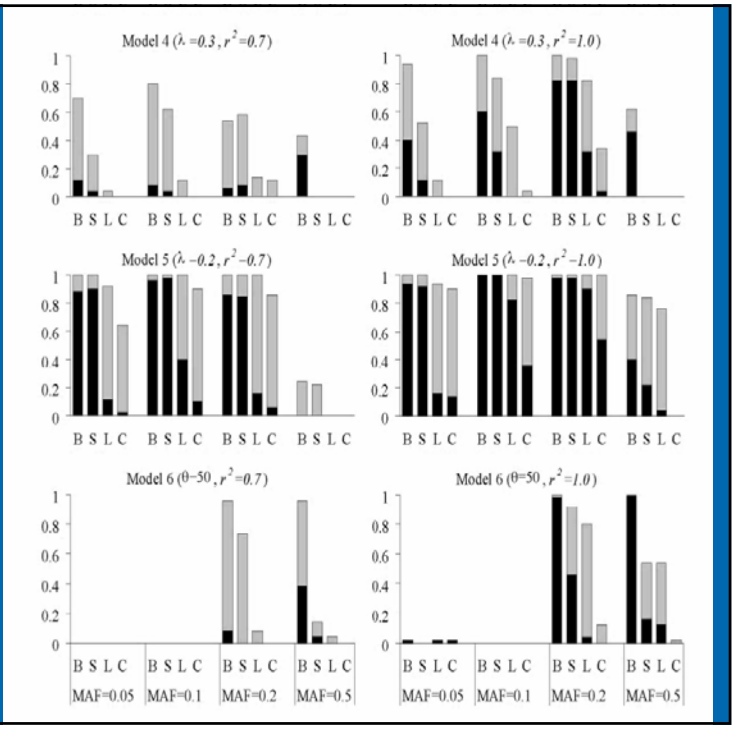
- Compare to and step-wise logistic regression and Chi-square

  B: The full Bayesian model; S: Step-wise B-stat

  L: Step-wise logistic C: Chi-square test

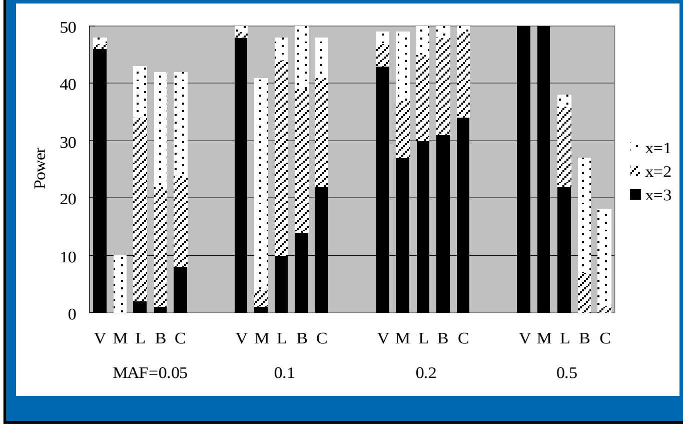
- 1,000 markers in N cases and N controls, where N = 1,000 (black bar) or 2,000 (grey bar)
- Power is averaged over 50 tests
- Type I error rate is at 0.1 with multiple correction





Compare to Other Methods

- MDR (Ritchie et al. 2001)
- Logic Regression (Kooperberg and Ruczinski 2005)
- BGTA (Zheng and Lo 2006)
- Chi-square (a single-marker approach)



40 markers in
400 cases and
400 controls

3-way interaction

Impact of MAF discrepancy

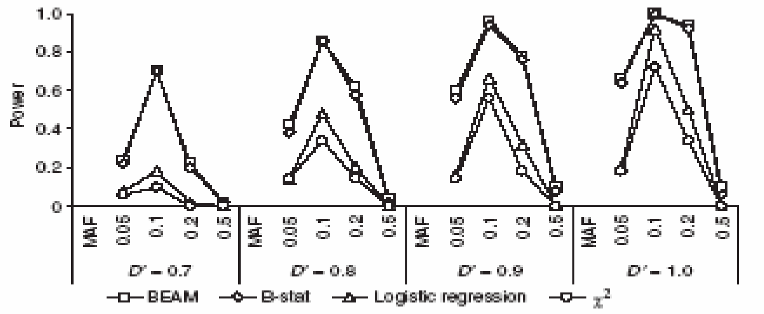
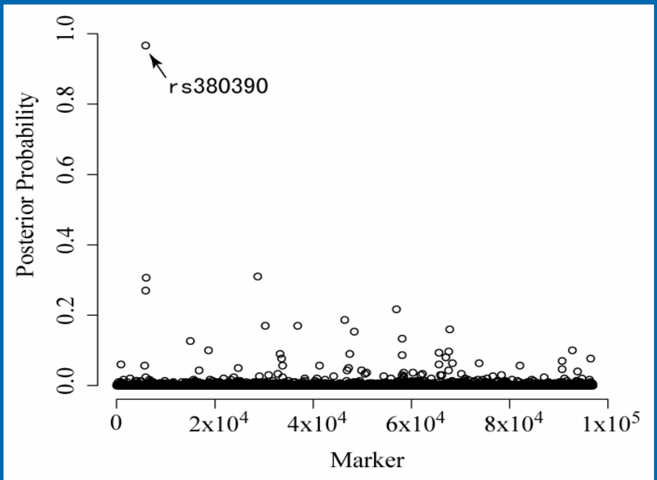


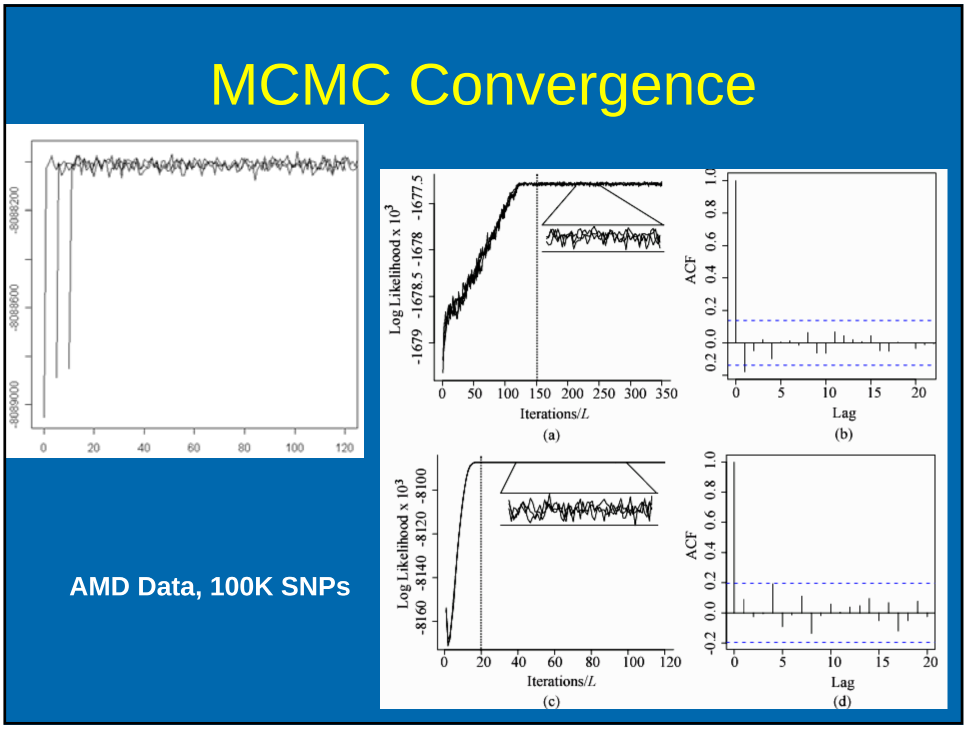
Figure 2 Impact of MAF discrepancy and LD on the powers of BEAM (B), the stepwise B-stat (S), the stepwise logistic regression (L) and the 2-d.f. χ^2 test (C). The comparison is based on model 2, where the allele frequencies of the second disease locus are unmatched by that of the associated marker. The marginal effect size per disease locus is 0.5. Under each setting, the power is calculated from 50 data sets containing 1,000 markers genotyped from 1,000 cases and 1,000 controls. The power is the proportion of 50 data sets in which all associated markers are identified at a significance threshold of 0.1 after Bonferroni correction.

Analyzing the whole-genome AMD data

- From J. Hoh's group (Klein et al. 2005)
- 116,204 SNP markers typed for 96 cases and 50 controls
- After filtering, 96,932 SNPs left for analysis
- We found the two markers reported (marginally significant), but no interactions
- We did further simulations using this data set

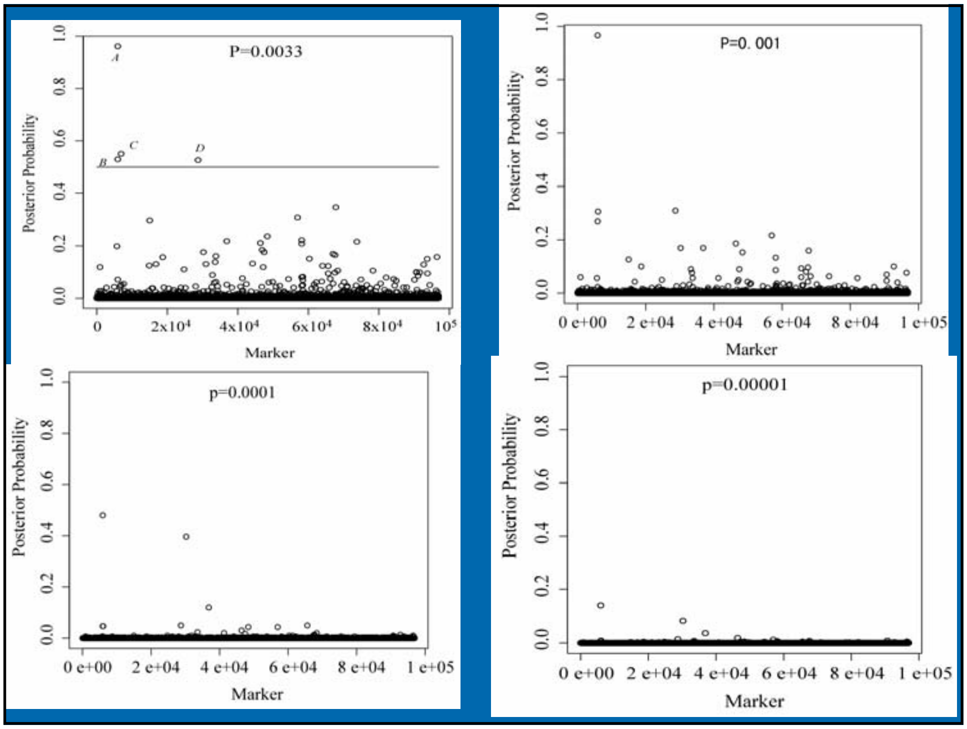
Posterior probabilities prior=10^{-3}



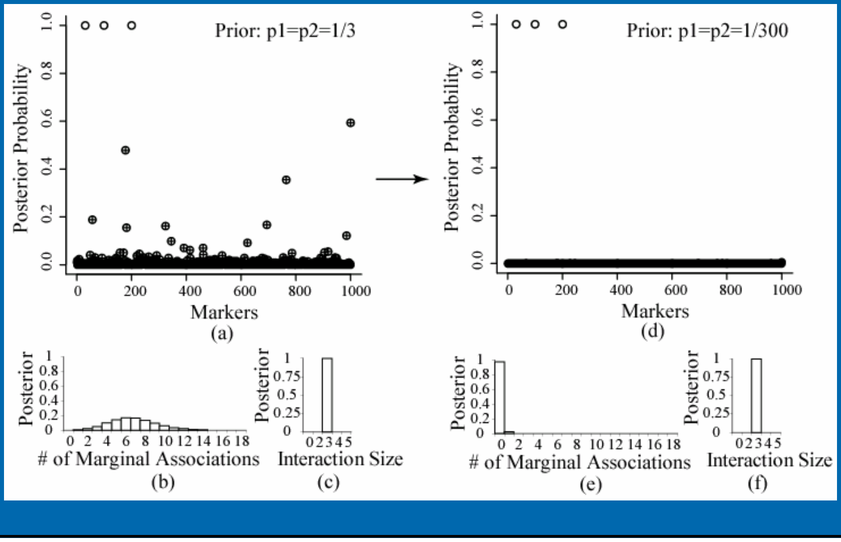


MCMC Convergence

AMD Data, 100K SNPs

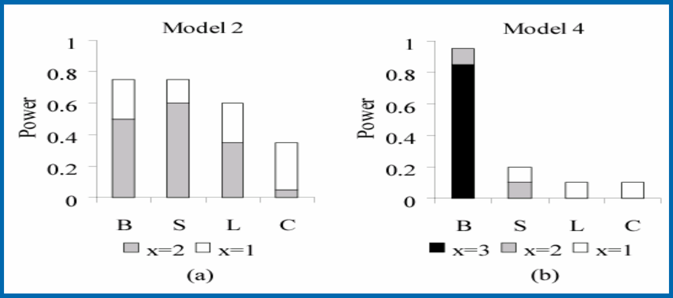


Posterior for a simulated case



Simulating AMD-like data

- 500 cases and 500 controls, 100K SNPs
- With genotype frequencies and LD structures similar to the AMD data
- Insert interactions based on Models 2 & 4
- Both BEAM and Logistic regression runs about 5 hrs

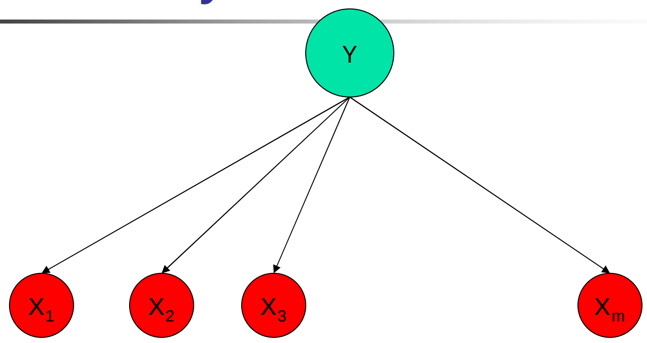


Publication: Zhang and Liu (2007), *Nature Genetics*, **39**(9), 1167-73.

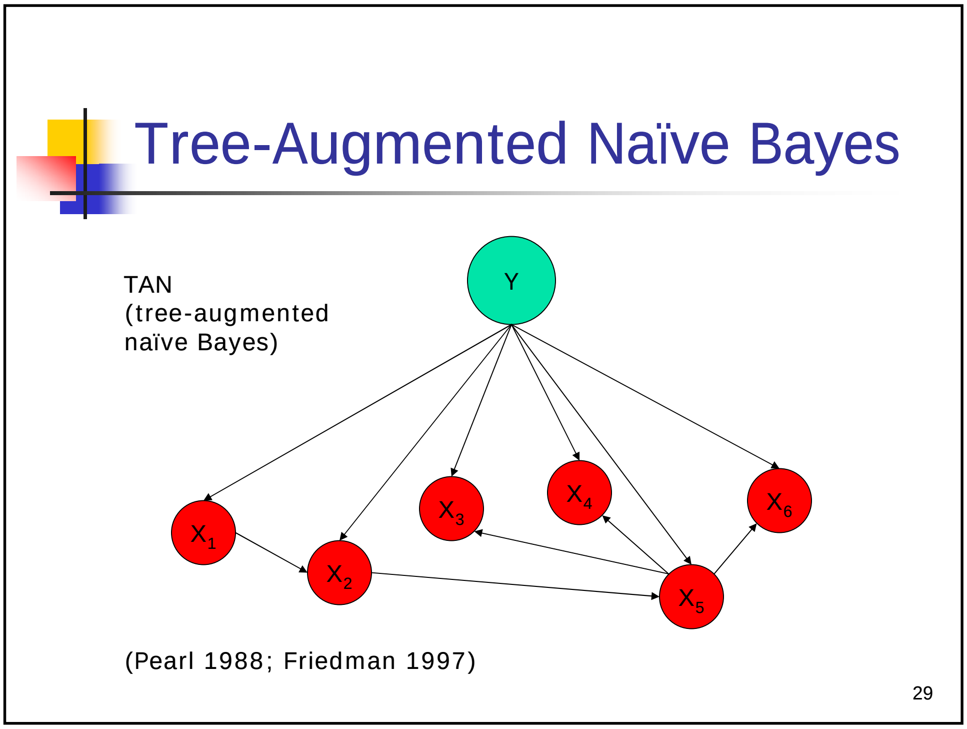
Augmented Naïve Bayes Classifier

- Basic setting for the classification problem:
  - Y: class label (1,2,...,K)
  - X: covariates (1,2,...,m)
    - Discrete valued
- Difficulties
  - Large number of covariates
    - Redundancy and colinearity would affect most classifiers
    - Variable selection is necessary
  - Different classes have different associated covariates.
    - Methods that select one group of variables for all classes would work poorly

Naïve Bayes model



$$P(Y|X_1,\cdots,X_m) = \frac{P(Y)\prod_{j=1}^{m} P(X_j|Y)}{P(X_1,\cdots,X_m)}.$$

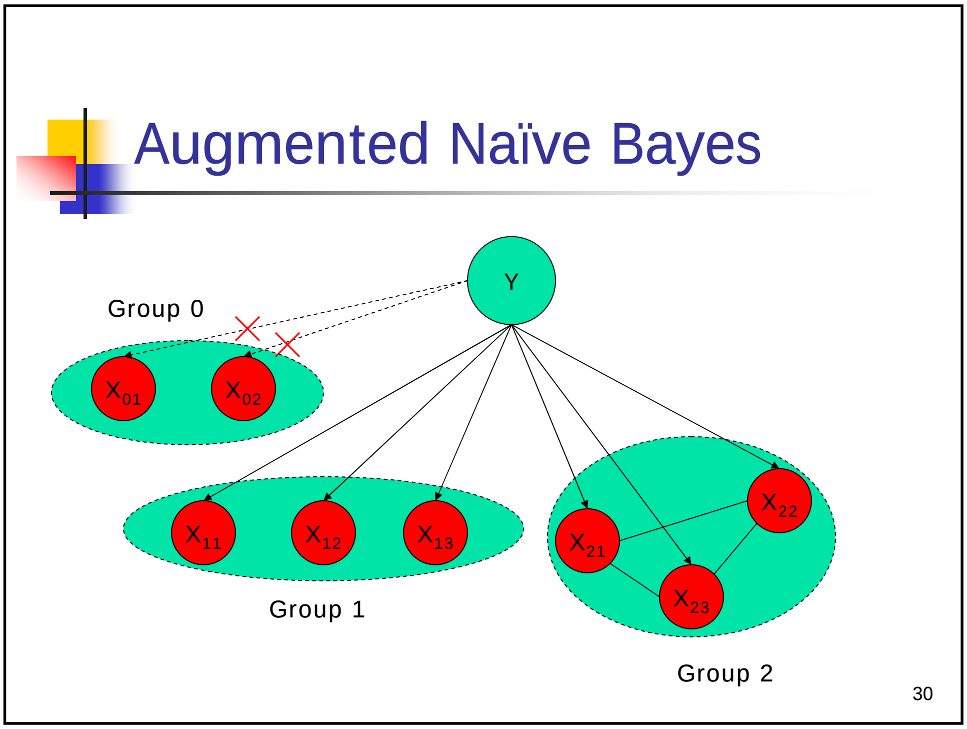


Tree-Augmented Naïve Bayes

TAN
(tree-augmented
naïve Bayes)

(Pearl 1988; Friedman 1997)

29



Augmented Naïve Bayes

30

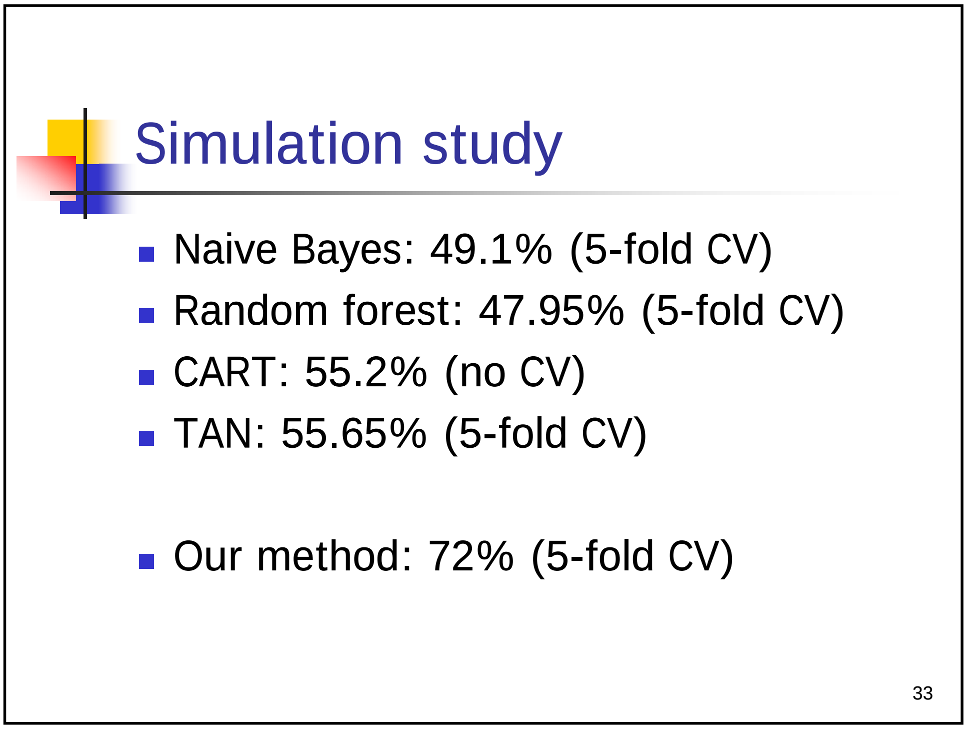
Classification

- Different classes are indicated by I’s
- Sample I from posterior distribution
- Calculate posterior probability ratio for each class k and put sample into the class with the highest ratio.

$$\frac{P(Y_{test}=k|\mathbf{X}_{test},\mathbf{X},\mathbf{y})}{P(Y_{test}\neq k|\mathbf{X}_{test},\mathbf{X},\mathbf{y})}=\frac{P(Y_{test}=k|\mathbf{X},\mathbf{y})}{P(Y_{test}\neq k|\mathbf{X},\mathbf{y})}\frac{P(\mathbf{X}_{test}|Y_{test}=k,\mathbf{X},\mathbf{y})}{P(\mathbf{X}_{test}|Y_{test}\neq k,\mathbf{X},\mathbf{y})}.$$

Simulation study

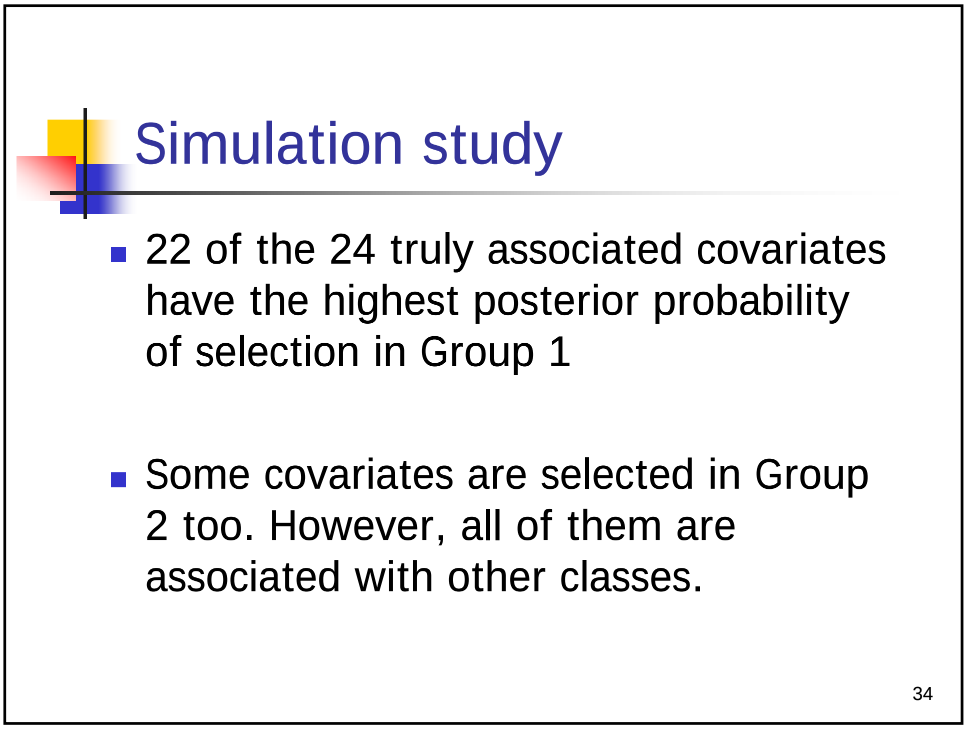
- 5 classes, p=(0.1, 0.1, 0.3, 0.2, 0.3)
- N=2000 samples
- m=200 covariates
- Each class is associated with 5 covariates (1 overlap, total 24)
  - Multinomial distributed with parameters randomly sampled from a Dirichlet distribution



Simulation study

- Naive Bayes: 49.1% (5-fold CV)
- Random forest: 47.95% (5-fold CV)
- CART: 55.2% (no CV)
- TAN: 55.65% (5-fold CV)

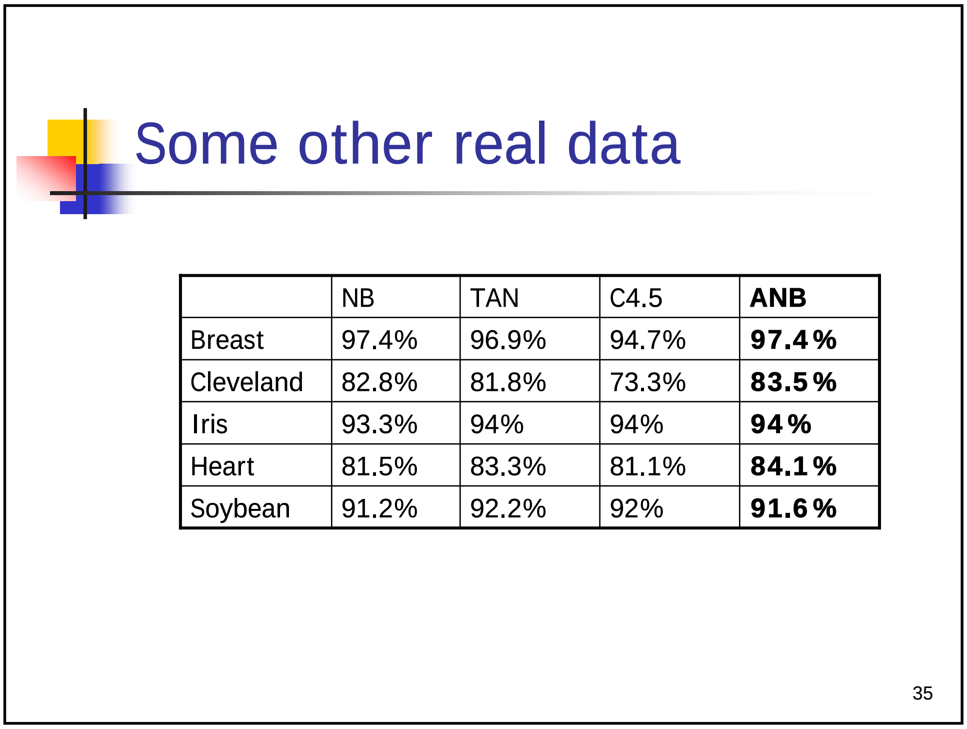
- Our method: 72% (5-fold CV)



Simulation study

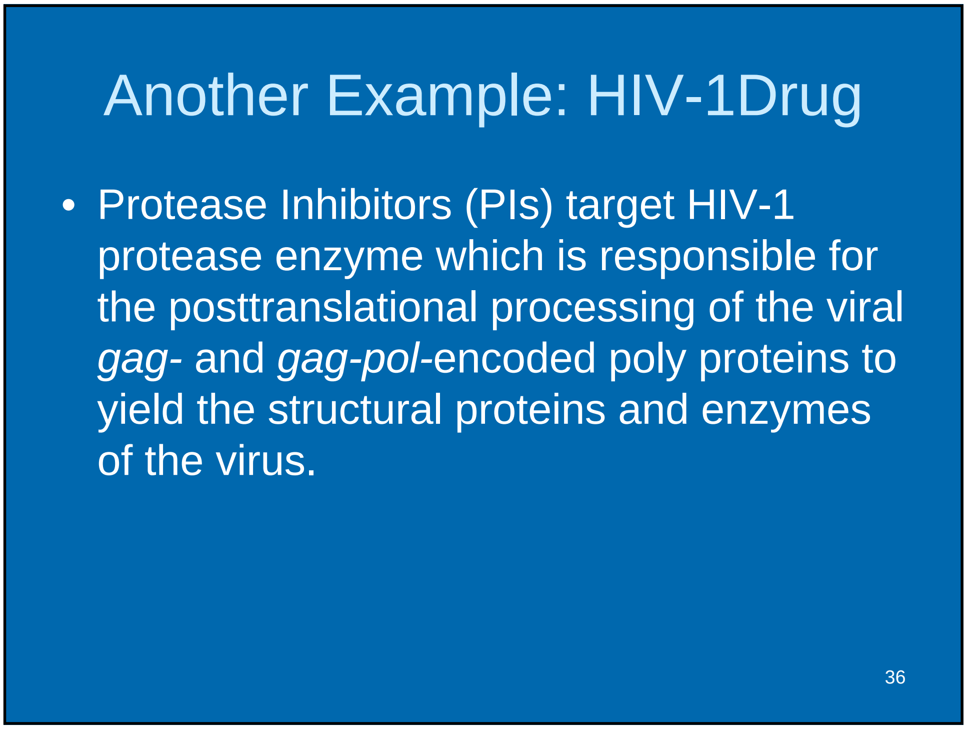
- 22 of the 24 truly associated covariates have the highest posterior probability of selection in Group 1

- Some covariates are selected in Group 2 too. However, all of them are associated with other classes.



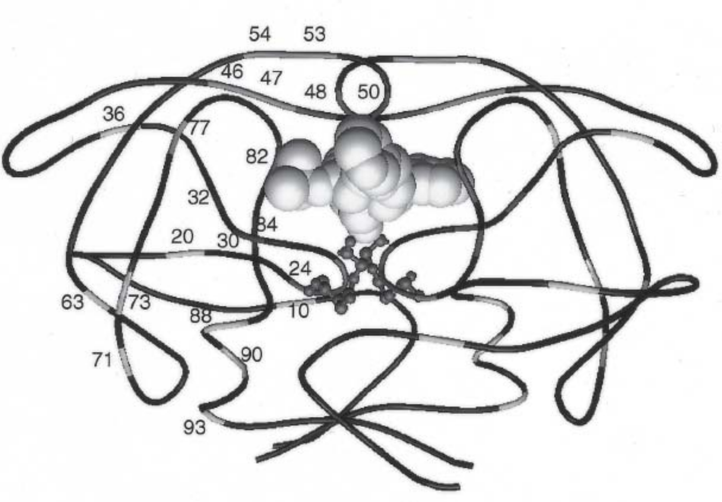
Some other real data

	NB	TAN	C4.5	**ANB**
Breast	97.4%	96.9%	94.7%	**97.4%**
Cleveland	82.8%	81.8%	73.3%	**83.5%**
Iris	93.3%	94%	94%	**94%**
Heart	81.5%	83.3%	81.1%	**84.1%**
Soybean	91.2%	92.2%	92%	**91.6%**



Another Example: HIV-1Drug

- Protease Inhibitors (PIs) target HIV-1 protease enzyme which is responsible for the posttranslational processing of the viral *gag-* and *gag-pol*-encoded poly proteins to yield the structural proteins and enzymes of the virus.



Structural model of HIV-1 protease homodimer labeled with protease inhibitor resistance mutations.

How to detect drug resistance Mutations

- Protease sequences from treated patients (949 cases)

  VVTIRIGGQLKEALLDTGAD
  IVTIRIGGQLKEALLDTGAD
  RVTIRIGGQLREALLDTGAD

- Sequences from untreated patients (4146 controls)

  LVTIRIGGQLREALLDTGAD
  IVTIRIGGQLKEALLDTGAD
  LVTIRIGGQLKEALLDTGAD

Which ones contributes to drug resistance?

Drug resistance mutations

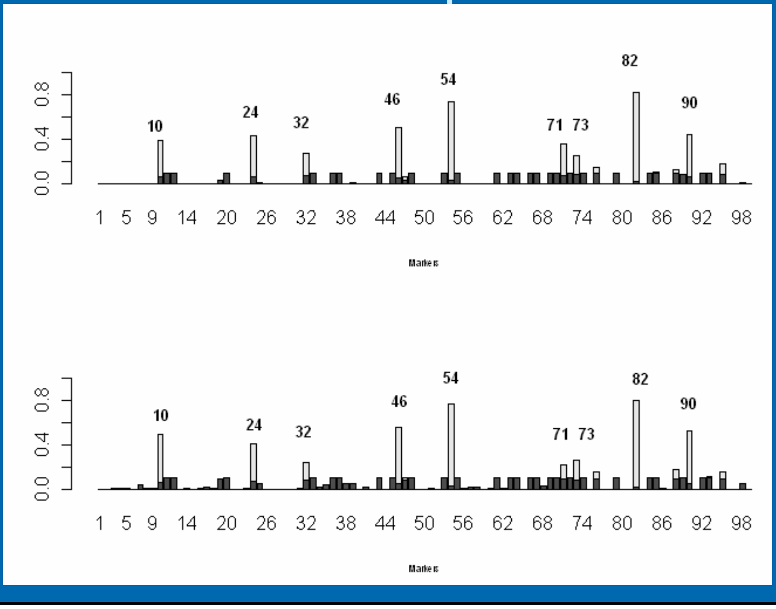
- The IAS-USA Drug Resistance Mutations list in HIV-1 updated in Fall 2006
- For IDV, mutations on the list are

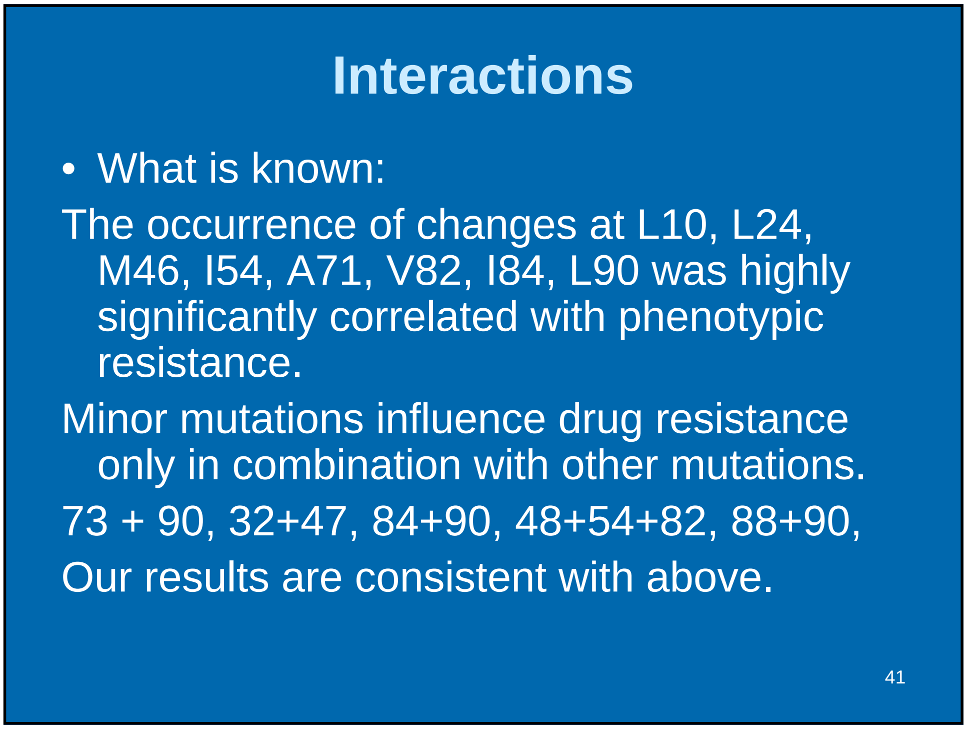
10, 20, 24, 32, 36, 46, 54, 71, 73, 77, 82, 84, 90

- The ones we detect

10, 24, 32, 46, 54, 71, 73, 82, 90

Posterior plots





Interactions

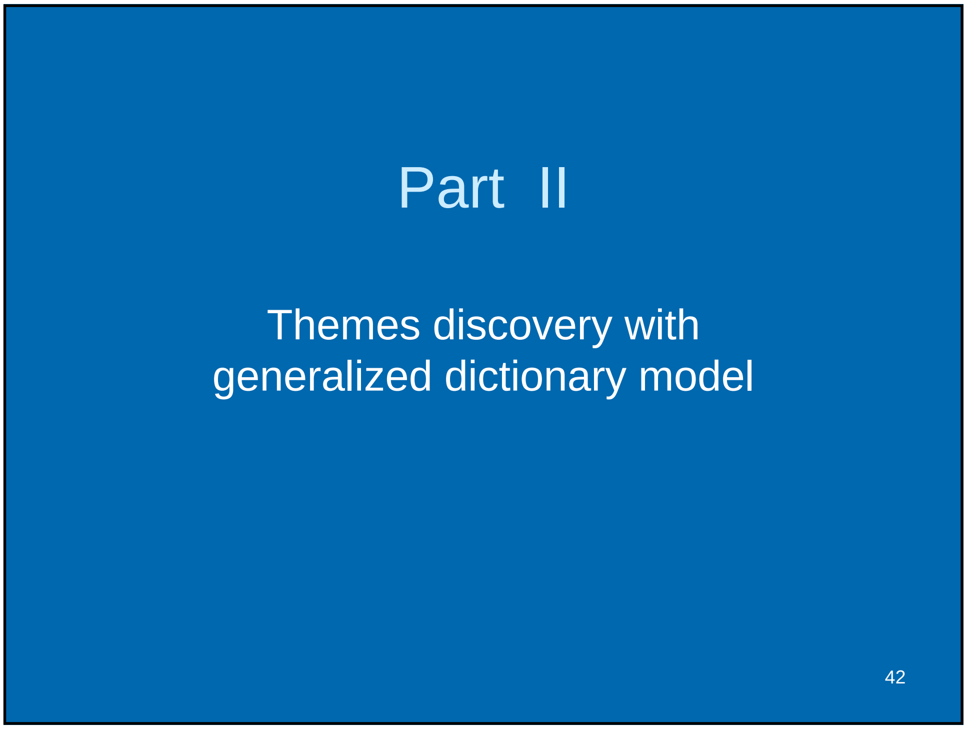
- What is known:

The occurrence of changes at L10, L24, M46, I54, A71, V82, I84, L90 was highly significantly correlated with phenotypic resistance.

Minor mutations influence drug resistance only in combination with other mutations.

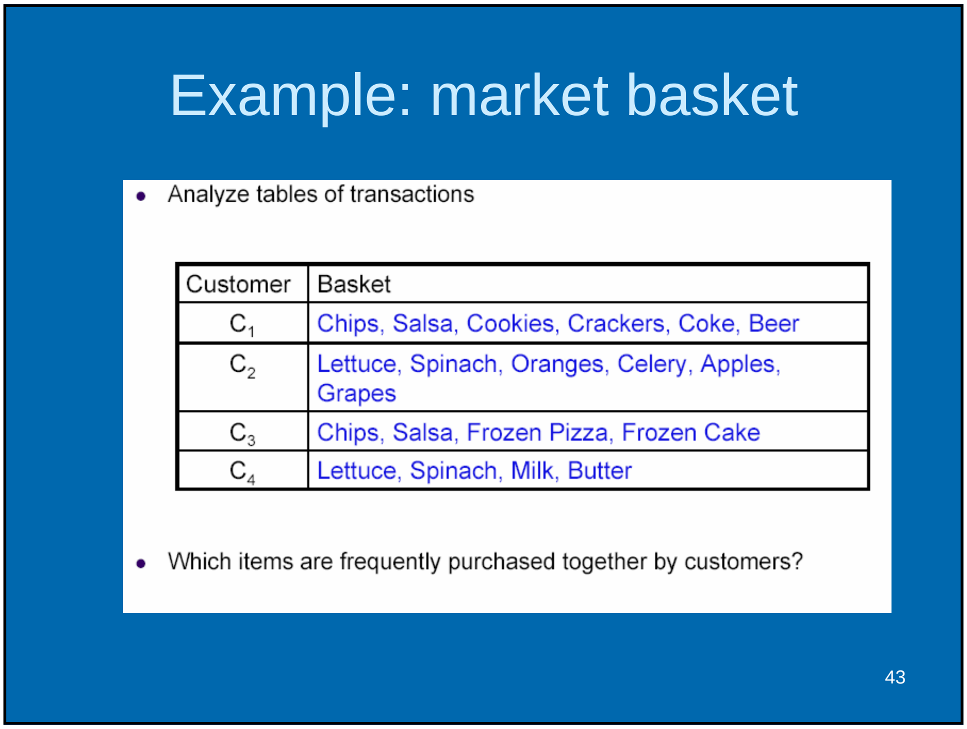
73 + 90, 32+47, 84+90, 48+54+82, 88+90,

Our results are consistent with above.



Part II

Themes discovery with
generalized dictionary model

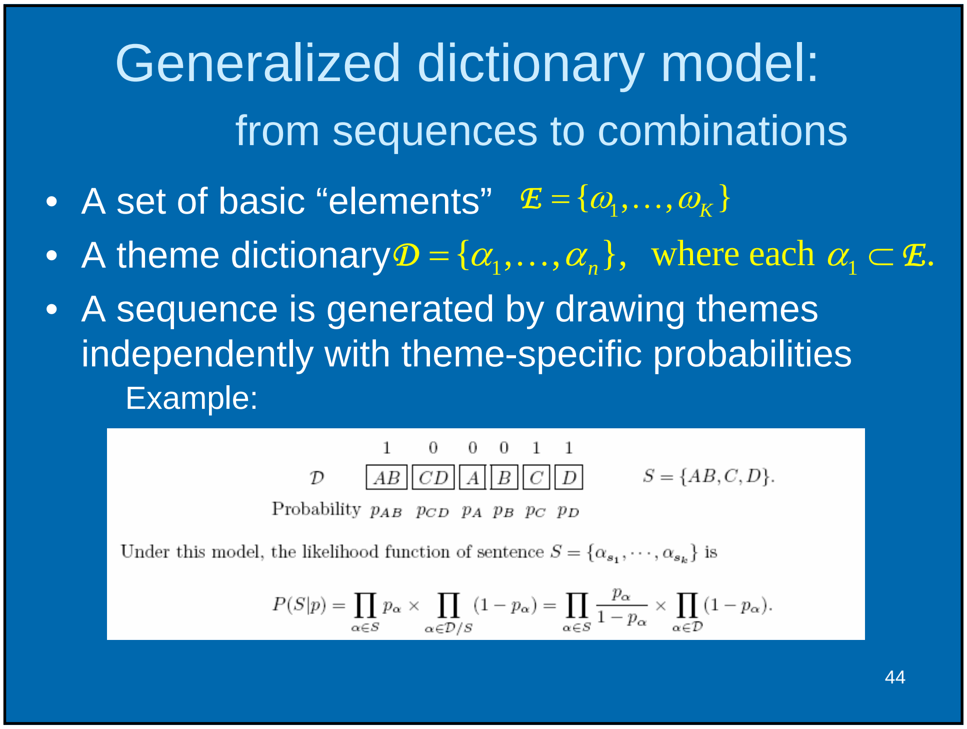


Example: market basket

- Analyze tables of transactions

Customer	Basket
C_1	Chips, Salsa, Cookies, Crackers, Coke, Beer
C_2	Lettuce, Spinach, Oranges, Celery, Apples, Grapes
C_3	Chips, Salsa, Frozen Pizza, Frozen Cake
C_4	Lettuce, Spinach, Milk, Butter

- Which items are frequently purchased together by customers?



Generalized dictionary model:

from sequences to combinations

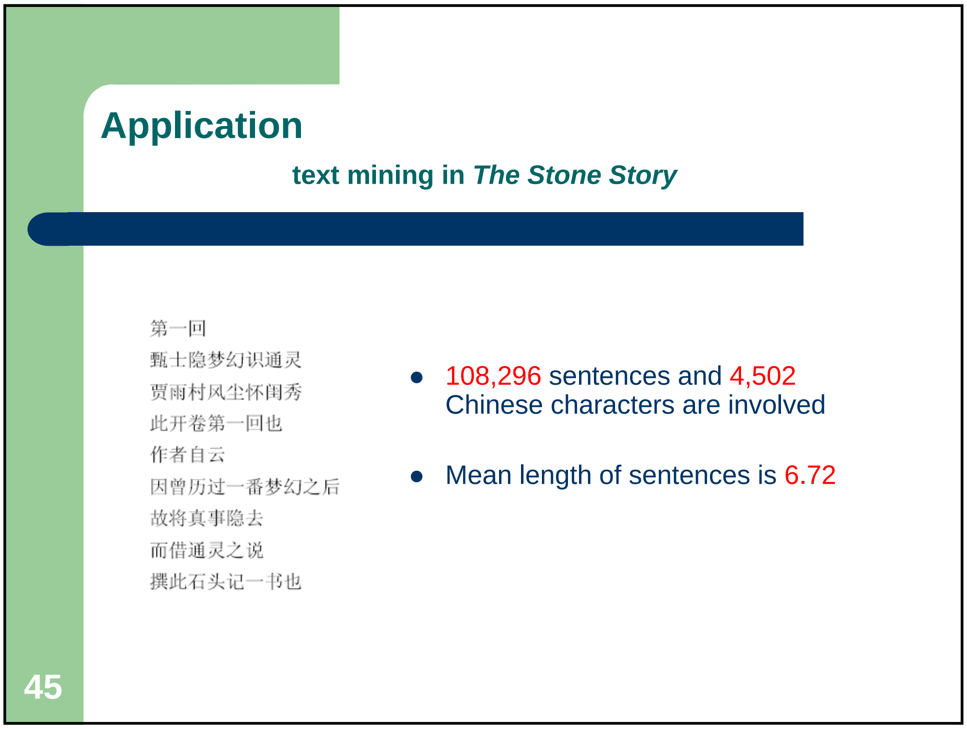
- A set of basic “elements” $\mathcal{E}=\{\omega_1,\dots,\omega_K\}$
- A theme dictionary $\mathcal{D}=\{\alpha_1,\dots,\alpha_n\}$, where each $\alpha_1 \subset \mathcal{E}$.
- A sequence is generated by drawing themes independently with theme-specific probabilities

Example:

$$\begin{array}{lcccccc} & 1 & 0 & 0 & 0 & 1 & 1 \\ \mathcal{D} & \boxed{AB} & \boxed{CD} & \boxed{A} & \boxed{B} & \boxed{C} & \boxed{D} \\ \text{Probability} & p_{AB} & p_{CD} & p_A & p_B & p_C & p_D \end{array} \qquad S=\{AB,C,D\}.$$

Under this model, the likelihood function of sentence $S=\{\alpha_{s_1},\cdots,\alpha_{s_k}\}$ is

$$P(S|p)=\prod_{\alpha\in S} p_\alpha \times \prod_{\alpha\in\mathcal{D}/S}(1-p_\alpha)=\prod_{\alpha\in S}\frac{p_\alpha}{1-p_\alpha}\times\prod_{\alpha\in\mathcal{D}}(1-p_\alpha).$$



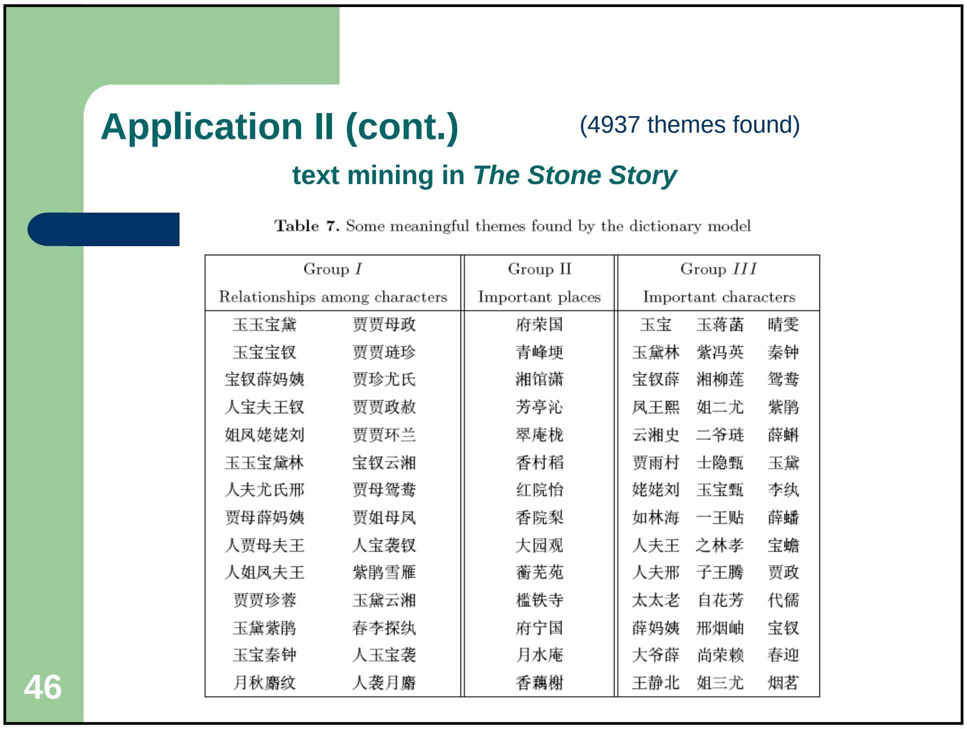
Application

text mining in *The Stone Story*

第一回
甄士隐梦幻识通灵
贾雨村风尘怀闺秀
此开卷第一回也
作者自云
因曾历过一番梦幻之后
故将真事隐去
而借通灵之说
撰此石头记一书也

- 108,296 sentences and 4,502 Chinese characters are involved
- Mean length of sentences is 6.72

45



Application II (cont.)

(4937 themes found)

text mining in *The Stone Story*

Table 7. Some meaningful themes found by the dictionary model

Group *I*		Group II	Group *III*		
Relationships among characters		Important places	Important characters		
玉玉宝黛	贾贾母政	府荣国	玉宝	玉蒋菡	晴雯
玉宝宝钗	贾贾琏珍	青峰埂	玉黛林	紫冯英	秦钟
宝钗薛妈姨	贾珍尤氏	湘馆潇	宝钗薛	湘柳莲	鸳鸯
人宝夫王钗	贾贾政赦	芳亭沁	凤王熙	姐二尤	紫鹃
姐凤姥姥刘	贾贾环兰	翠庵栊	云湘史	二爷琏	薛蝌
玉玉宝黛林	宝钗云湘	香村稻	贾雨村	士隐甄	玉黛
人夫尤氏邢	贾母鸳鸯	红院怡	姥姥刘	玉宝甄	李纨
贾母薛妈姨	贾姐母凤	香院梨	如林海	一王贴	薛蟠
人贾母夫王	人宝袭钗	大园观	人夫王	之林孝	宝蟾
人姐凤夫王	紫鹃雪雁	蘅芜苑	人夫邢	子王腾	贾政
贾贾珍蓉	玉黛云湘	槛铁寺	太太老	自花芳	代儒
玉黛紫鹃	春李探纨	府宁国	薛妈姨	邢烟岫	宝钗
玉宝秦钟	人玉宝袭	月水庵	大爷薛	尚荣赖	春迎
月秋麝纹	人袭月麝	香藕榭	王静北	姐三尤	烟茗

46

Acknowledgement

- Yu Zhang (Penn State U, Statistics Dept)
- Wei Zhang (Harvard Statistics)
- Jing (Maria) Zhang (Harvard Statistics)
- Yuan Yuan
- Ke Deng
- Zhi Geng
- Josephine Hoh