

# Feature Selection in Pharmacogenetics

**Application to Calcium Channel Blockers in  
Hypertension Treatment**

---

IEEE CIS June 2006

Dr. Troy Bremer  
Prediction Sciences



# Pharmacogenetics

## Great potential

- SNPs (Single Nucleotide Polymorphisms)
  - Differences between human genomes underlie differences between individual characteristics



- in susceptibility to or protection from a host of diseases
- enable individualized medicine

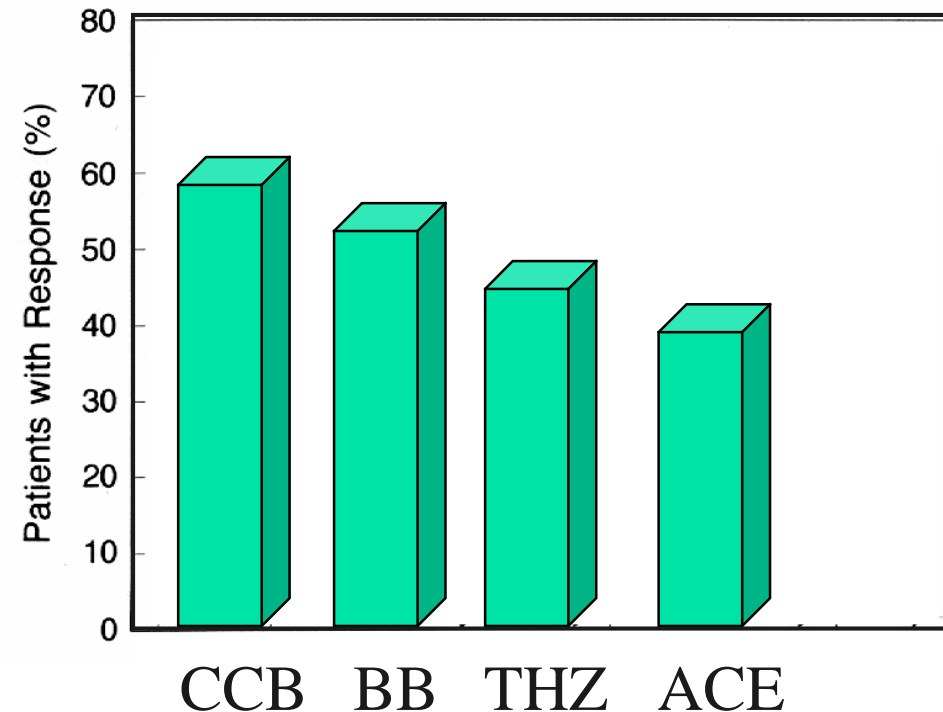
Motivation

Millions of SNPs mapped today, 11 Million estimated to exist

# Treatment Biodiversity

Motivation

- Antihypertensive monotherapy classes
  - Calcium Channel Blocker
  - Beta Blocker
  - Thiazide Diuretic
  - ACE inhibitors
- A given monotherapy is effective in approximately 50% of patients



*Materson NEJM 1993 328:914-921*

# Pharmacogenetics

## Great potential, now which ones?

- SNPs (Single Nucleotide Polymorphisms)
  - Differences between human genomes underlie differences between individual characteristics

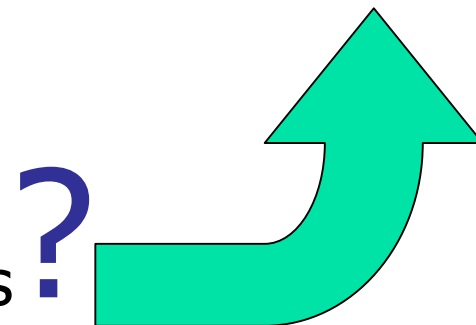


- in susceptibility to or protection from a host of diseases

- Will this drug work for me?

Motivation

Of the Millions identified, which SNPs ?





# Feature Selection in Association Studies

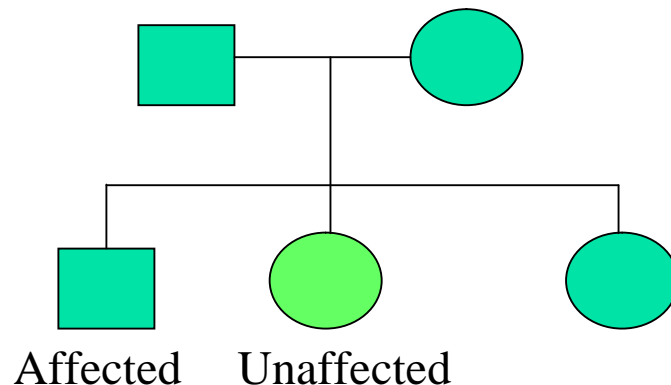
---

- Background
- Definition of Problem
  - Relevance Assessment
  - “Optimal” Subset
- Standard Approaches
  - Filter and Wrapper Based
  - Implications
- Modified Approach
  - Small data sets
- Case Study – Calcium Channel Blocker

# Standard Clinical Approaches

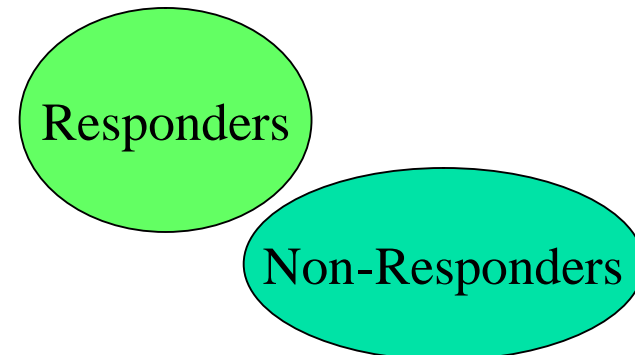
## Linkage Study

- Collect **FAMILIES** with multiple affected responders members
- What regions are consistently transmitted to affected individuals?



## Association Study

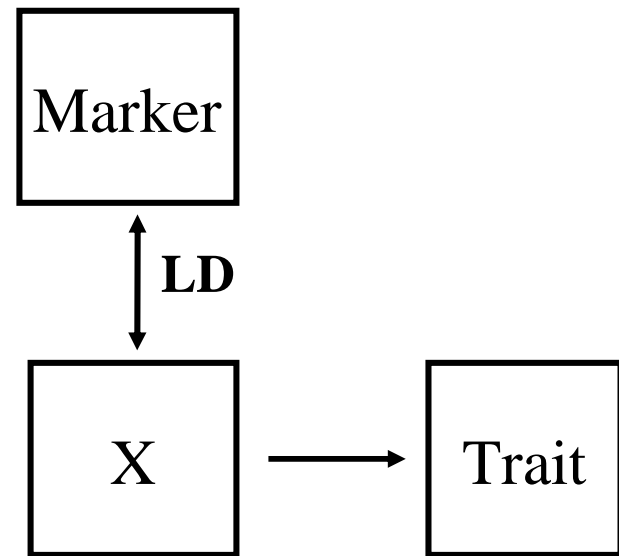
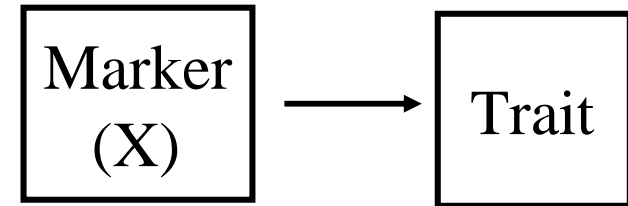
- Collect sample of population (subjects with unknown lineage)
- Is the frequency of a polymorphism associated with response?



# Associations

Background

- **Causative marker locus**
  - Directly influences disease risk
  - Established via functional studies
- **Associated marker locus**
  - Alleles at marker locus only in Linkage Disequilibrium (LD) with alleles at the disease locus (X)
  - Correlate with disease risk
  - Do not directly influence disease risk

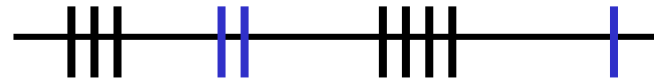


# A "simple" problem

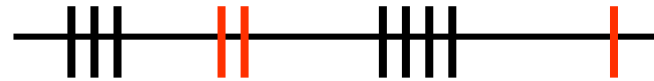
Definition of problem

- Identify a subset of marker(s)
  - that can be used to produce a relevant diagnostic

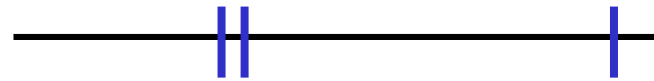
Patients *with* efficacy  
in clinical trials



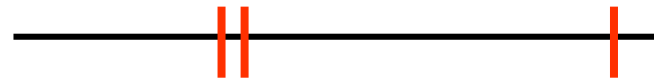
Patients *without* efficacy  
in clinical trials

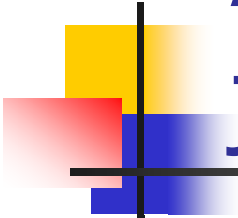


Predictive of *efficacy*



Predictive of *no efficacy*





# A “simple” problem, just a few limitations ...

---

## Definition of problem

- Identify a subset of marker(s)
  - that can be used to produce a relevant diagnostic
- But, ...
  - only a limited number of observations ( $N$ ) are available due to availability
  - choose the subset of markers from many potential markers  $M$ , where usually  $M > N$ , and sometimes  $M \gg N$
- Also, ...
  - markers  $M$  are highly interdependent
  - and they may be both continuous and discrete

# Relevance is ... defined by association

- Is a feature  $x_i$  is relevant?
  - To predict an observed outcome set  $Y$
- One answer
  - A feature  $x_i$  is relevant if it is not statistically independent from  $y$

$$P(y | x_i) \neq P(y)$$

- Chi-squared test

# Relevance is ... defined by model usefulness

- Is a feature  $x_i$  is relevant?
  - To predict an observed outcome set  $Y$
- A simple answer
  - A feature  $x_i$  is relevant if a model  $f(x_i)$  shows “good” performance

$$\min_f (y - f(x_i))^2 \approx 0$$

- But, not likely

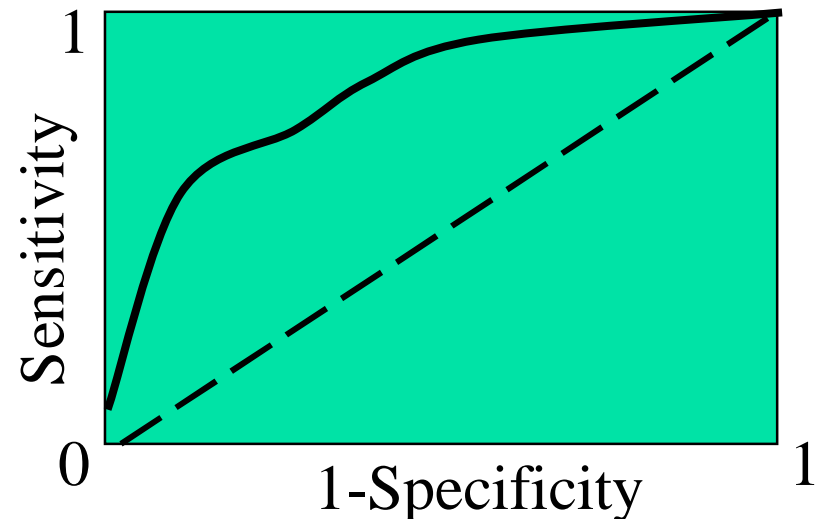
# Relevance is ... diagnostic usefulness

- A model yielding “good” decisions
  - Receiver operator characteristic (ROC) curve statistics
    - Plot of model TPF versus FPF (different thresholds)

	True +	True -
Test +	TP	FP
Test -	FN	TN

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

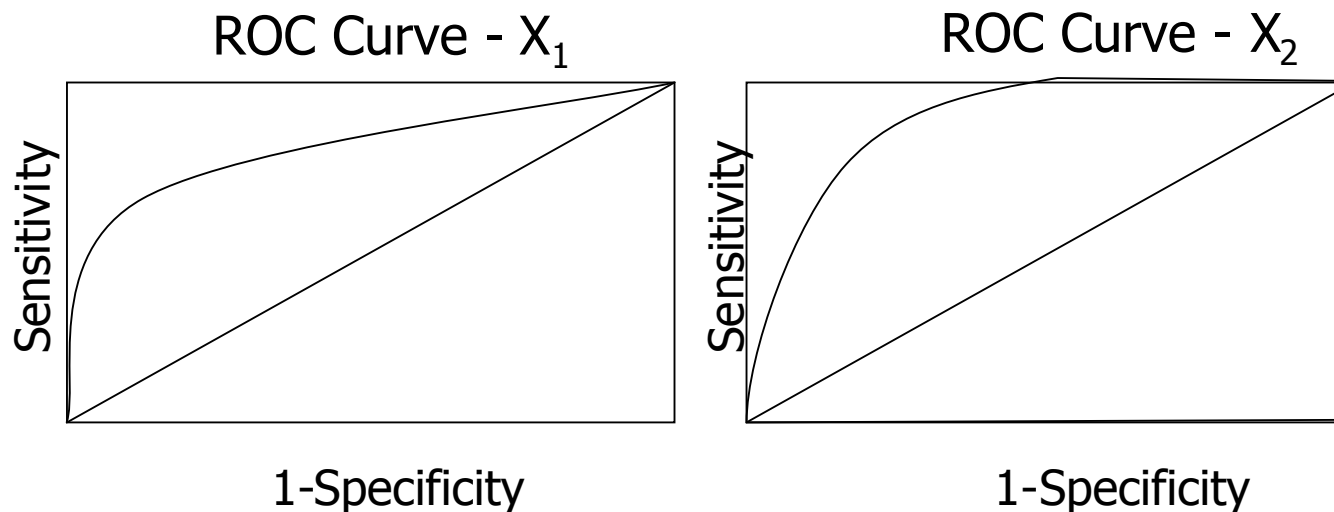
$$\text{Specificity} = \text{TN} / (\text{FP} + \text{TN})$$



# Relevance is ... diagnostic usefulness

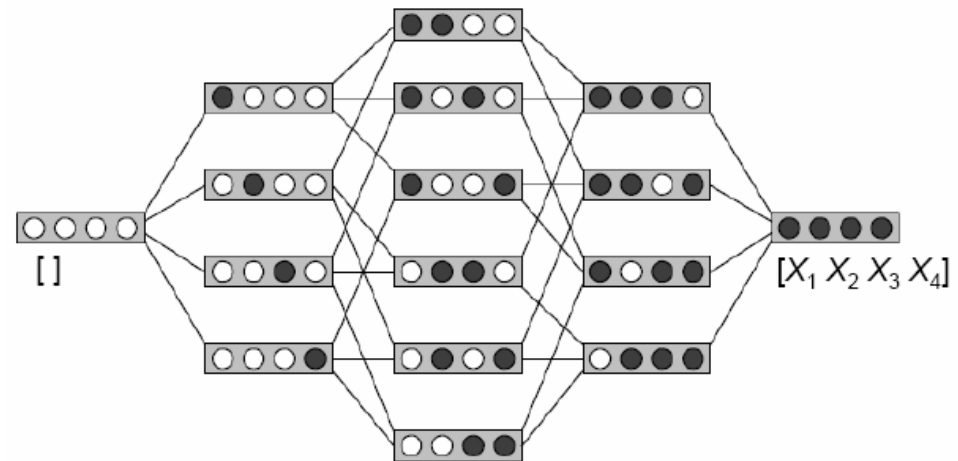
## Relevance Assessment

- Which model is more diagnostically useful?
  - It depends on the problem, but in general
  - Bigger Area Under the Curve (AUC) is better



# Feature Selection

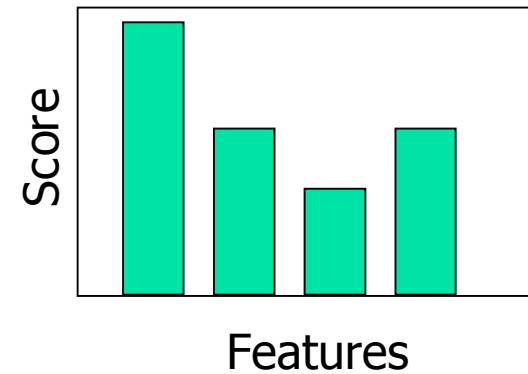
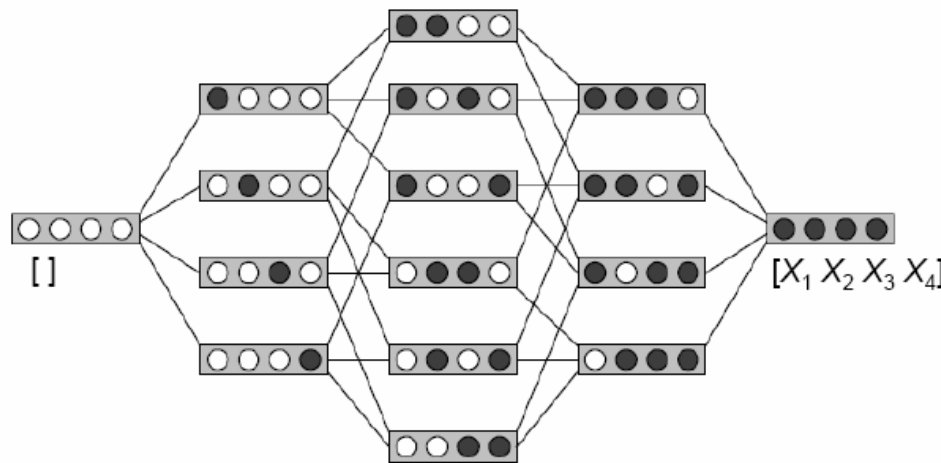
- Select the “best” subset of features  $K$  from set  $M$ 
  - Immense search space --  $M!/K!(M-K)!$ 
    - Even  $M=50, 4 < K < 11$ , Direct Search  $> 1E9$
  - Spurious feature selection
  - Inflated performance estimation
  - Unknown model order



# Feature Selection Ranking

“Optimal” Subset - Filter

- Which subset is most relevant?
  - Hypothesis 1: The set of K most relevant features



- But, the composite of best individual features may not be the best set of features



# Typical Approach Filter Based

---

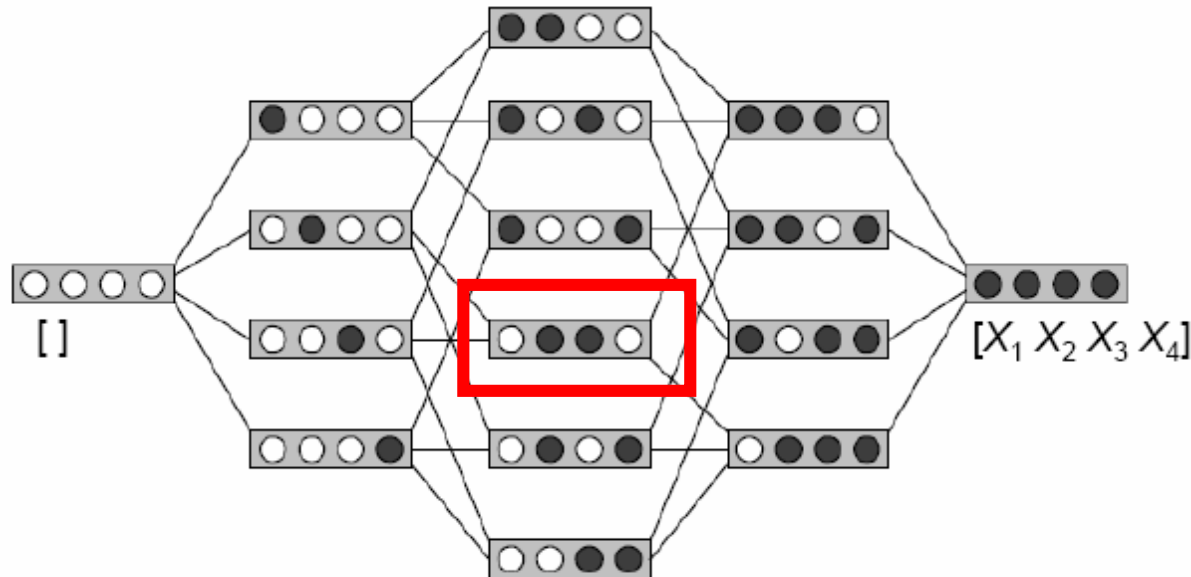
“Optimal” Subset - Filter

- Filter the set of SNPs based on a CHI-squared test and score
- Rank the candidates
  - Rank the scores and select using a cut point
  - Or, generate q-values (FDR) and select using a cut point
- Then ...
  - Build a model from the set and report performance
    - Haplotype
  - Or, run another study with the set to confirm relevance of markers/gene

# Feature Selection Iterative Heuristic

“Optimal” Subset - Wrapper

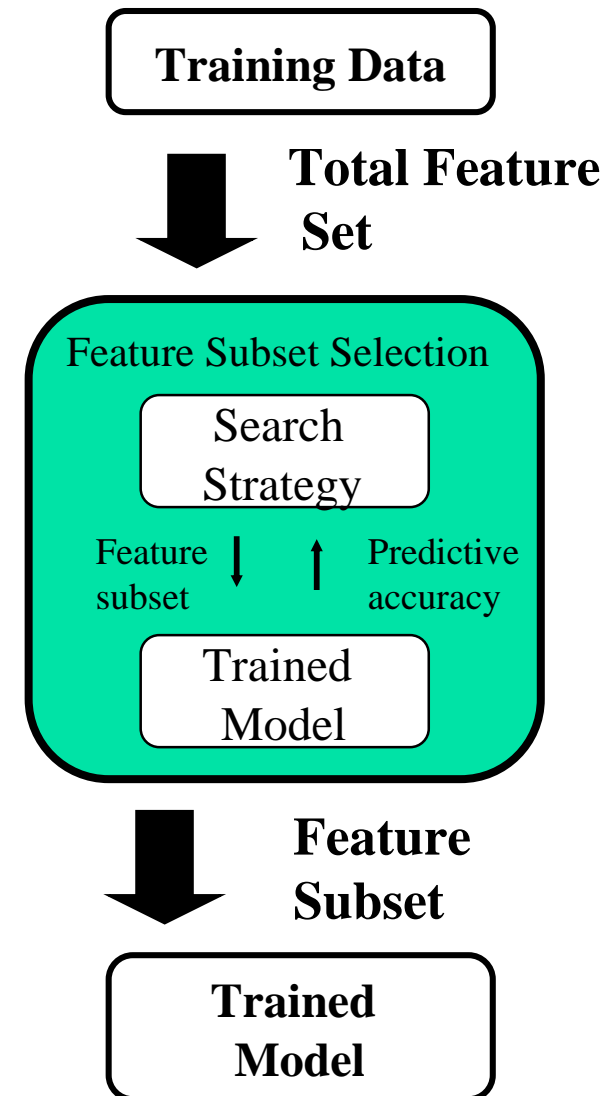
- Which subset is most relevant?
  - Hypothesis 2: Approximate the solution to the N choose K problem



# Methodology – Wrapper based

## “Optimal” Subset - Wrapper

- Advantages
  - Tuned to the interactions between the specific classifier and dataset
    - Better performance
  - Can use cross-validated prediction accuracy measures
    - Avoids over-fitting
- Disadvantage
  - Computationally Intensive
  - The “optimal” feature subset will be specific to the classifier under consideration





# Standard Search Strategy

---

“Optimal” Subset - Wrapper

- Use a greedy search
  - Either Forward or Backward
- Method
  - Start with an initial set
  - Choose variable that maximizes model “generalization performance”
  - Decide when to stop
    - Decreasing “generalization performance”
- Then
  - Report obtained “generalization” performance
  - Run a confirmation study

# Standard Approach Summary

- Filter based Approach
  - Efficient to narrow the field and generate primary features and candidate genes
- However,
  - Identified features may not lead to a useful model
    - Features selected using a univariate approach
    - Feature interdependence
    - Contextual feature identification
    - Not tuned to nuisances of selected model
  - Model performance assessment typically optimistic
    - Feature selection

# Standard Approach Summary

- Wrapper based
  - Contextual features
  - Tuned model performance
- But,
  - Computationally expensive
  - Difficult search problem
    - Backward search may be unfeasible ( $M > N$ )
    - Forward search non-optimality (initialization)
  - Optimistic "Cross Validated performance"
    - Feature Selection

Implications



# Summary of implications

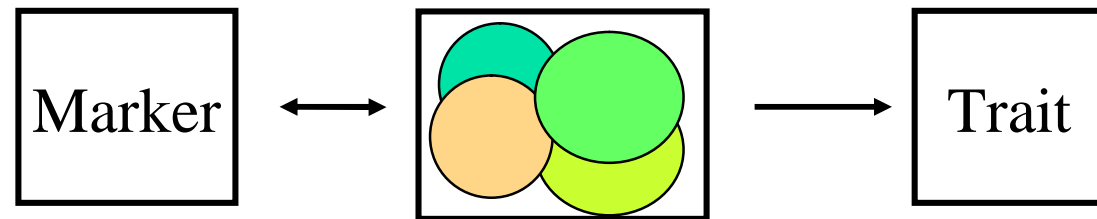
---

- Many studies are not successfully replicated
  - Optimistic Performance Estimate
    - All elements of experimental methodology not accounted for
  - Correlation does not imply causation
    - SNPS

# Issues with Association Studies

- False Positives

- Multiple Test induced Type I errors
- Population strata
  - Different disease and marker genotype frequencies in subpopulations create illusion of marker association
- Robust Validation Required



**Ethnicity**    **Environmental Factors**  
**Geography/Ancestry**

Implications

# Association Study - Methodology

## Modified Approach

- Phase 1 –
  - Subset Approach – Filter based
  - Generate candidate gene list for second study
- Phase 2 –
  - Use candidate gene list
  - Relevance Assessment
    - Naïve Cross-validation
    - Small data set modifications
    - AUC ROC
  - Subset approach - Wrapper based
    - Sequential Floating Forward Search (Pudil)
    - Kernel Based Partial Least Squares (KPLS)

# Naïve Cross-validation

- Nested cross-validation (CV)
  - Outer Loop
    - Test --> Evaluate trained model on fully independent data (feature selection and parameter estimation)
  - Inner Loop
    - Train --> Estimate model parameters given feature set
    - Validate --> Score feature relevance using data independent from the parameter estimation data set





# Small Data Set Modifications

Modified Relevance

- Stratification
  - Make responder rates representative across data sets
  - Based on *a priori* information balance data using key clinical data or biomarkers
- Randomization
  - Enforce a new random seed at each training epoch during feature selection
- Outcome weighting
  - During training and feature selection
  - Re-weight outcome variables to account for unbalanced responder rates

# AUC ROC

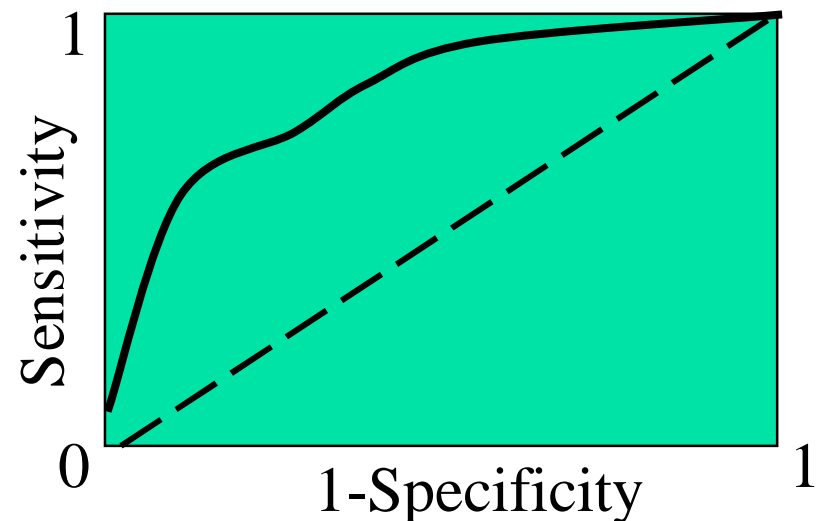
Modified Relevance

- Use AUC ROC for feature selection
  - Maximize General Utility
  - Alternatively add a weighted function of sensitivity at a preset specificity

	True +	True -
Test +	TP	FP
Test -	FN	TN

Sensitivity  $TP/(TP+FN)$

Specificity  $TN/(FP+TN)$





# Search Strategy – SFFS

## Modified Search

- Sequential Floating Forward Search (Pudil)
  - Contains both Forward and Backward Search
  - Generalized (+L) forward (-R) backward Search
- Process
  - 1) Pick feature (x) to maximize  $J(Y,X)$   
Add  $x(k)$  to feature set X
  - 2) Find least useful feature  $x(b)$  in X  
Remove  $x(b)$  if performance  $J(Y, X-x(b))$  is same or better
  - 3) Repeat 2 again if removed feature
  - 4) Repeat 1



# SFFS – Modifications

---

## Modified Search

- Minimize sensitivity to search initialization
  - Force multiple initializations to create separate search sequences
  - Use a higher order search for initialization
  - Follow each search sequence, but eliminate overlapping feature set evaluations



# SFFS - Modifications

---

- Surrogate Data
  - Identify or prevent spurious feature selection
  
- Approach
  - Randomly select a subset of candidate features
  - Randomly re-index this set within each data subset to form a null association set
  - Append null association set to search candidate set
  
- Then
  - Only proceed with feature selection if differential performance is above set % of null distribution
  - OR allow null hypothesis marker to be selected and record for post processing



# Methodology – KPLS

---

- Kernel Partial Least Squares
  - Kernel extension of partial least squares
- Advantages
  - Good for multiple variables with high co linearity (eg height and weight)
  - No nonlinear optimization needed - just linear algebra
- Disadvantage
  - Kernel function must be chosen *a priori*



# Validation

## Primary Feature Validation

- Univariate significance
  - $\chi^2$  tests
    - Either trend (123) or dominant allele response (110,011)
    - Required cell frequencies of 10 or greater
  - False Positive Controls
    - FWER Control
      - Global permutation -
        - For each univariate test re-label outcome (10,000X)
    - FDR Control
      - Q-value analysis in R smoother option

Storey JD and Tibshirani R. (2003) Statistical significance for genome-wide studies. Proceedings of the National Academy of Sciences, 100: 9440-9445.



# Calcium Channel Blocker

---

- Calcium Channel Blocker
  - “Block” the entry of calcium into cardiac and smooth muscle
  - Blood vessels dilate --> reduced blood pressure
- Candidate Gene Approach
  - CACNA1C gene
    - Calcium channel, voltage dependent, L-type alpha-1C subunit
    - 12p13.3
  - 47 SNPs assessed
    - Assays were designed for sixty-two SNPs



# CCB Subjects - Clinical Features

Case Study

	<b>CCB Responders</b>	<b>CCB Non-Responders</b>	<b>p</b>
<b>N</b>	<b>44</b>	<b>76</b>	
<b>Age (years)</b>	<b>58 (11) years</b>	<b>58 (9) years</b>	<b>NS</b>
<b>Smoking</b>	<b>26%</b>	<b>36%</b>	<b>NS</b>
<b>Male</b>	<b>34%</b>	<b>41%</b>	<b>NS</b>
<b>Coffee (cups/day)</b>	<b>1.9 (1.0)</b>	<b>1.7 (1.1)</b>	<b>NS</b>
<b>Non-Essential Hypertension</b>	<b>8%</b>	<b>24%</b>	<b>0.028</b>
<b>Baseline BP (mm Hg)</b>	<b>161(11) / 96 (5)</b>	<b>167 (12) / 100 (6)</b>	<b>NS</b>

# CCB Results

- Identified 7 SNPs (5 Naïve Folds)
- 4 Core Features (gray - yellow)
- 2 Primary Features (gray)

ID rs_gene	chr:bp pos	Role
rs1003306_CACNA1C	12:2624458	intronic
rs1008832_CACNA1C	12:2483782	intronic
rs2238032_CACNA1C	12:2092993	upstream
rs2239050_CACNA1C	12:2317675	intronic
rs2239072_CACNA1C	12:2405419	intronic
rs2283301_CACNA1C	12:2317067	intronic
rs2283287_CACNA1C	12:2168985	intronic

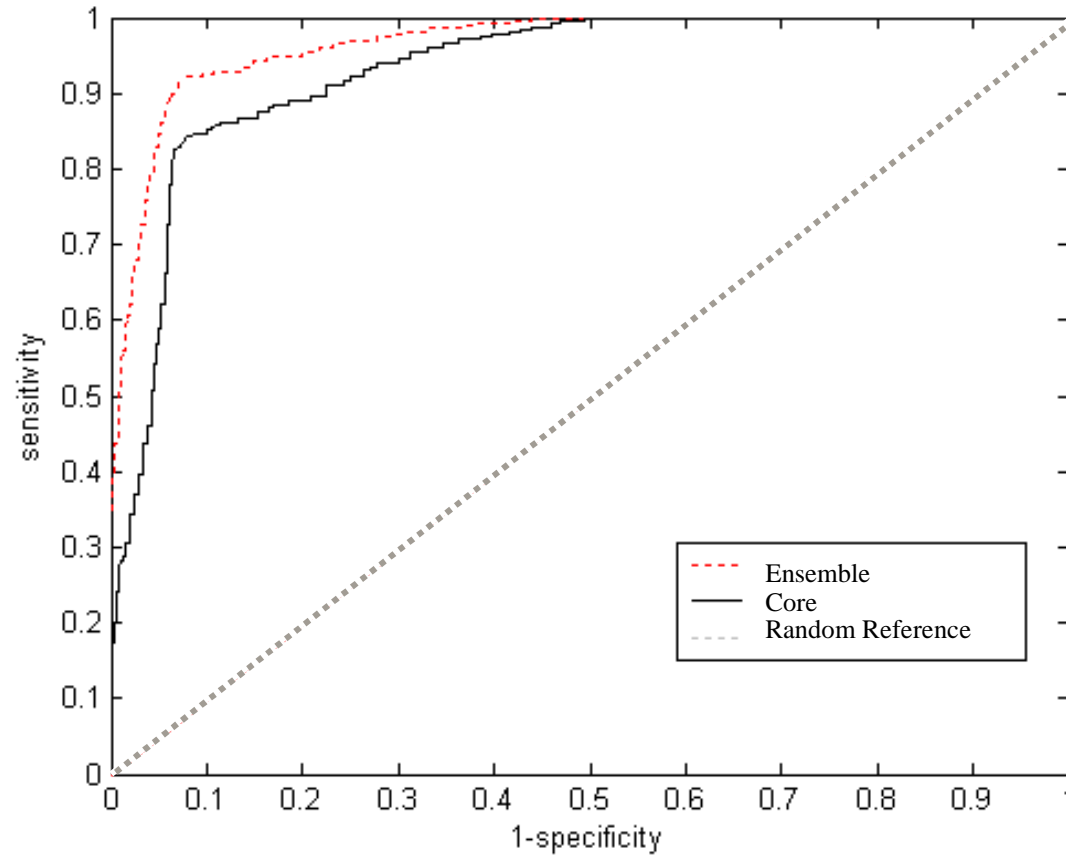
Genotype Coding Scheme	$p^1$	$p^{*2}$	$q^3$	$p_b^{*4}$
Dominant allele	1.2E-13	<1E-4	5.1E-12	5.1E-12
Dominant allele	2.6E-06	<1E-4	5.2E-05	0.00011

- 1 All analyses DF=1 (41 tests)
- 2 Permutation method (10,000X)
- 3 FDR method by Storey
- 4 Bonferroni method

# CCB - Naïve Composite ROC

Case Study

AUC ROC 0.97  
AUC ROC 0.93

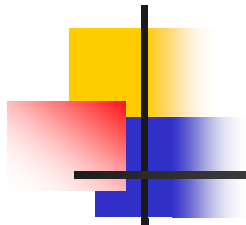




# Summary

---

- Filtering Approach
  - Identified 3 significant SNPs
  - Useful for candidate gene identification
- Wrapper based feature selection and modeling
  - Employed
    - SFFS
    - ROC Relevance Measure
    - Selected modifications for small data sets
  - Yielded high performing model
- Approach appears useful for pharmacogenetics



Thank you