The Collaborative Center for Statistics in Science C^2S^2

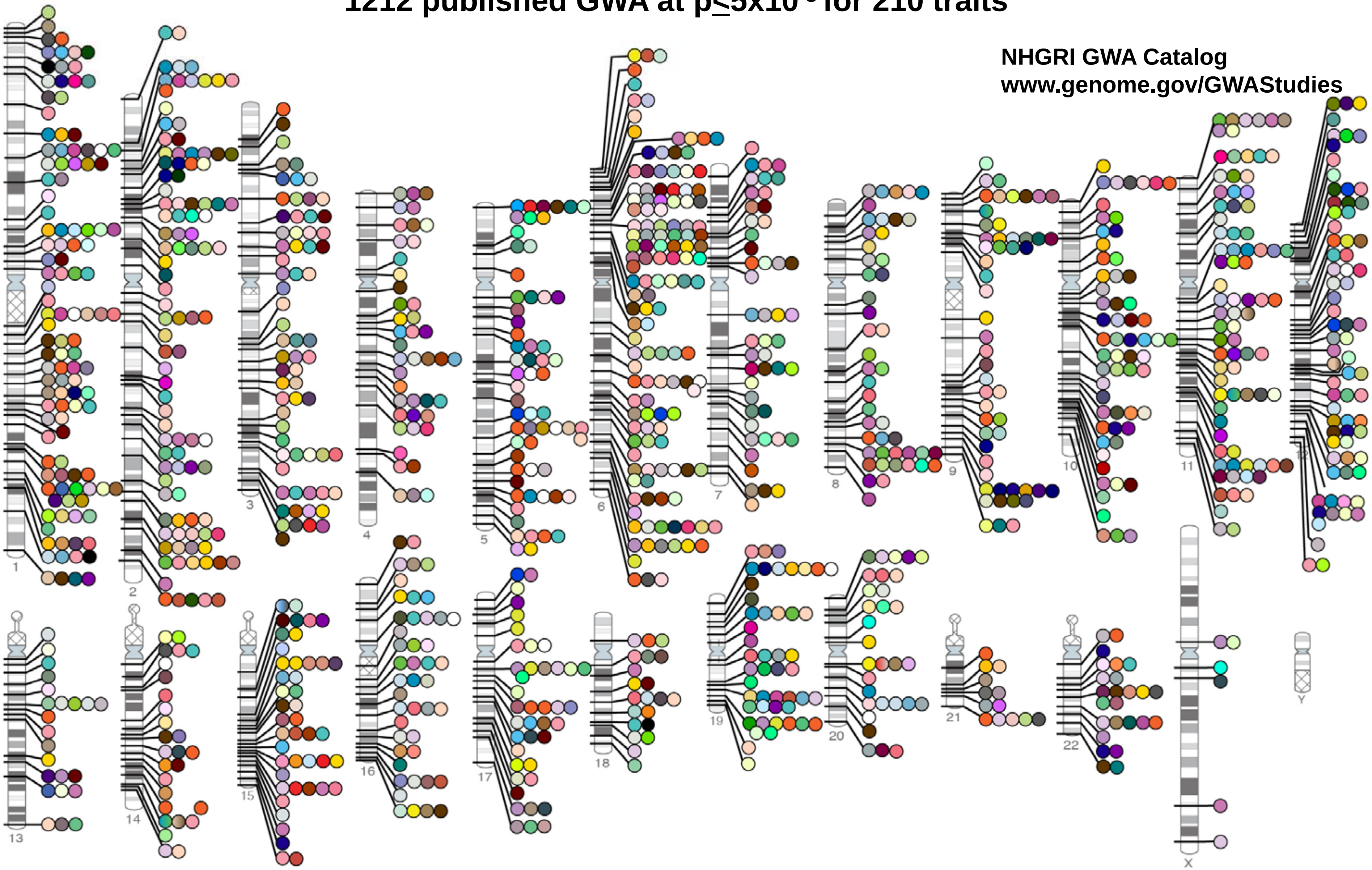
A Forest Approach to Identification of Genes and Gene-Environment Interactions for Complex Diseases

Heping Zhang

School of Public Health
Yale University School of Medicine

Published Genome-Wide Associations through 12/2010, 1212 published GWA at p≤5×10^{-8} for 210 traits

NHGRI GWA Catalog
www.genome.gov/GWAStudies



- Abdominal aortic aneurysm
- Acute lymphoblastic leukemia
- Adhesion molecules
- Adverse response to carbamapezine
- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alcohol dependence
- Alopecia areata
- Alzheimer disease
- Amyloid A levels
- Amyotrophic lateral sclerosis
- Angiotensin-converting enzyme activity
- Ankylosing spondylitis
- Arterial stiffness
- Asparagus anosmia
- Asthma
- Atherosclerosis in HIV
- Atrial fibrillation
- Attention deficit hyperactivity disorder
- Autism
- Basal cell cancer
- Behcet's disease
- Bipolar disorder
- Biliary atresia
- Bilirubin
- Bitter taste response
- Birth weight
- Bladder cancer
- Bleomycin sensitivity
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- C-reactive protein
- Calcium levels
- Cardiac structure/function
- Carnitine levels
- Carotenoid/tocopherol levels
- Celiac disease
- Cerebral atrophy measures
- Chronic lymphocytic leukemia
- Cleft lip/palate
- Cognitive function
- Conduct disorder
- Colorectal cancer
- Corneal thickness
- Coronary disease
- Creutzfeldt-Jakob disease
- Crohn's disease
- Cutaneous nevi
- Dermatitis
- Drug-induced liver injury
- Endometriosis
- Eosinophil count
- Eosinophilic esophagitis
- Erectile dysfunction and prostate cancer treatment
- Erythrocyte parameters
- Esophageal cancer
- Essential tremor
- Exfoliation glaucoma
- Eye color traits
- F cell distribution
- Fibrinogen levels
- Folate pathway vitamins
- Follicular lymphoma
- Fuch's corneal dystrophy
- Freckles and burning
- Gallstones
- Gastric cancer
- Glioma
- Glycemic traits
- Hair color
- Hair morphology
- Handedness in dyslexia
- HDL cholesterol
- Heart failure
- Heart rate
- Height
- Hemostasis parameters
- Hepatic steatosis
- Hepatitis
- Hepatocellular carcinoma
- Hirschsprung's disease
- HIV-1 control
- Hodgkin's lymphoma
- Homocysteine levels
- Hypospadias
- Idiopathic pulmonary fibrosis
- IgA levels
- IgE levels
- Inflammatory bowel disease
- Intracranial aneurysm
- Iris color
- Iron status markers
- Ischemic stroke
- Juvenile idiopathic arthritis
- Keloid
- Kidney stones
- LDL cholesterol
- Leprosy
- Leptin receptor levels
- Liver enzymes
- Longevity
- LP (a) levels
- LpPLA(2) activity and mass
- Lung cancer
- Magnesium levels
- Major mood disorders
- Malaria
- Male pattern baldness
- Matrix metalloproteinase levels
- MCP-1
- Melanoma
- Menarche & menopause
- Meningococcal disease
- Metabolic syndrome
- Migraine
- Moyamoya disease
- Multiple sclerosis
- Myeloproliferative neoplasms
- N-glycan levels
- Narcolepsy
- Nasopharyngeal cancer
- Neuroblastoma
- Nicotine dependence
- Obesity
- Open angle glaucoma
- Open personality
- Optic disc parameters
- Osteoarthritis
- Osteoporosis
- Otosclerosis
- Other metabolic traits
- Ovarian cancer
- Pancreatic cancer
- Pain
- Paget's disease
- Panic disorder
- Parkinson's disease
- Periodontitis
- Peripheral arterial disease
- Phosphatidylcholine levels
- Phosphorus levels
- Photic sneeze
- Phytosterol levels
- Platelet count
- Polycystic ovary syndrome
- Primary biliary cirrhosis
- Primary sclerosing cholangitis
- PR interval
- Progranulin levels
- Prostate cancer
- Protein levels
- PSA levels
- Psoriasis
- Psoriatic arthritis
- Pulmonary funct. COPD
- QRS interval
- QT interval
- Quantitative traits
- Recombination rate
- Red vs.non-red hair
- Refractive error
- Renal cell carcinoma
- Renal function
- Response to antidepressants
- Response to antipsychotic therapy
- Response to hepatitis C treat
- Response to metaformin
- Response to statin therapy
- Restless legs syndrome
- Retinal vascular caliber
- Rheumatoid arthritis
- Ribavirin-induced anemia
- Schizophrenia
- Serum metabolites
- Skin pigmentation
- Smoking behavior
- Speech perception
- Sphingolipid levels
- Statin-induced myopathy
- Stroke
- Systemic lupus erythematosus
- Systemic sclerosis
- T-tau levels
- Tau AB1-42 levels
- Telomere length
- Testicular germ cell tumor
- Thyroid cancer
- Tooth development
- Total cholesterol
- Triglycerides
- Tuberculosis
- Type 1 diabetes
- Type 2 diabetes
- Ulcerative colitis
- Urate
- Venous thromboembolism
- Ventricular conduction
- Vertical cup-disc ratio
- Vitamin B12 levels
- Vitamin D insuffiency
- Vitiligo
- Warfarin dose
- Weight
- White cell count
- YKL-40 levels

Challenges

The identified markers or genes explained a small fraction of the diseases

More markers & GxG?

Environment variables & GxE?

Incorporation of biologic knowledge?

Better characterization and use of traits?

Classic Modeling vs Genomic Association Analysis

In classic statistical modeling, we tend to have an adequate sample size for estimating 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 genomic studies, we have more and more variables (gene based) but the access to the number of study subjects remains the same. One model can no longer provide an adequate summary of the information.

Outline

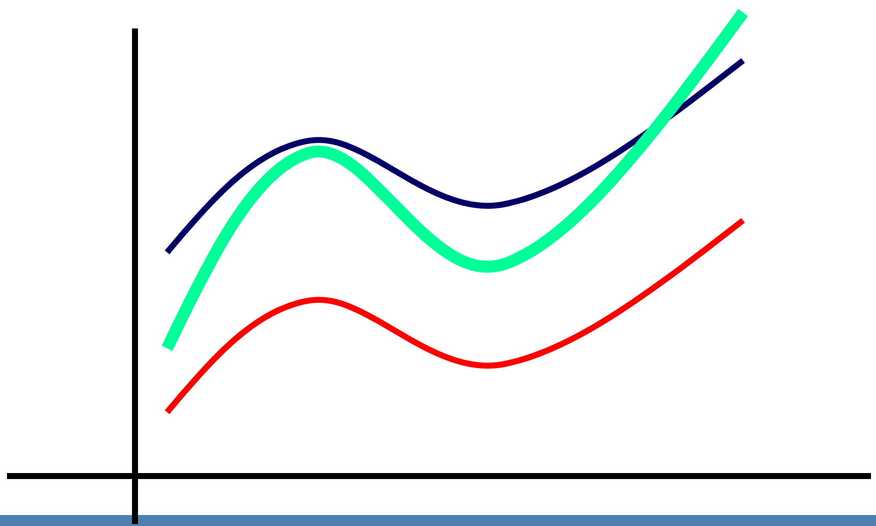
- Background
- Challenges
- **Methods**
  - **Trees and Forests**
  - **Forest Size**
  - **Feature Importance**
  - **Uncertainties in Predictors**
  - **Interactions**
- **Acknowledgement**

Complex Traits

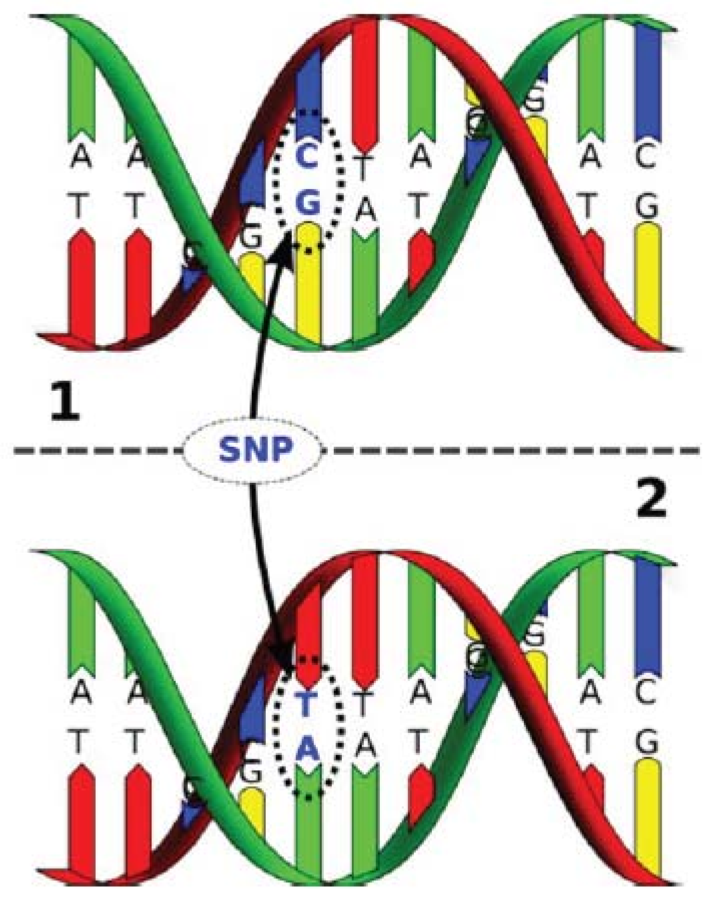
Diseases that do not follow Mendelian Inheritance Pattern

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

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



SNP and Complex Traits



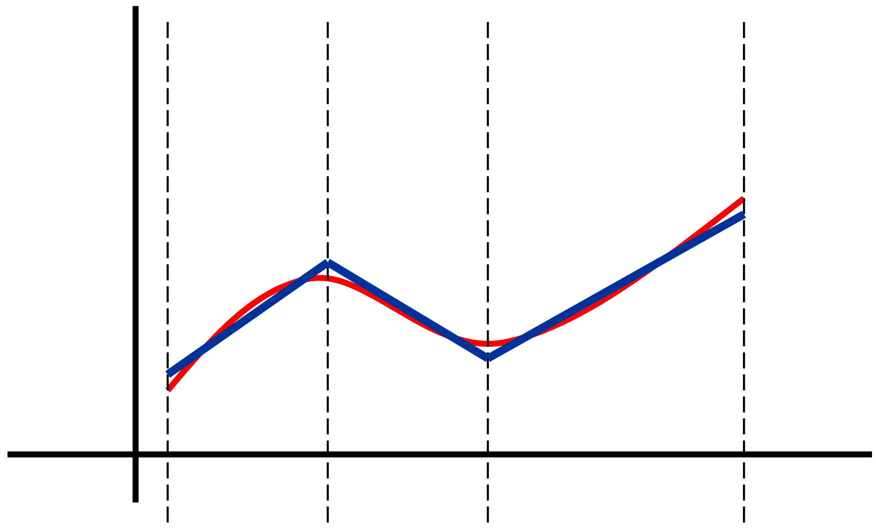
http://en.wikipedia.org/wiki/Single_nucleotide_polymorphism

Regression Approach

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

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.



Leukemia Data

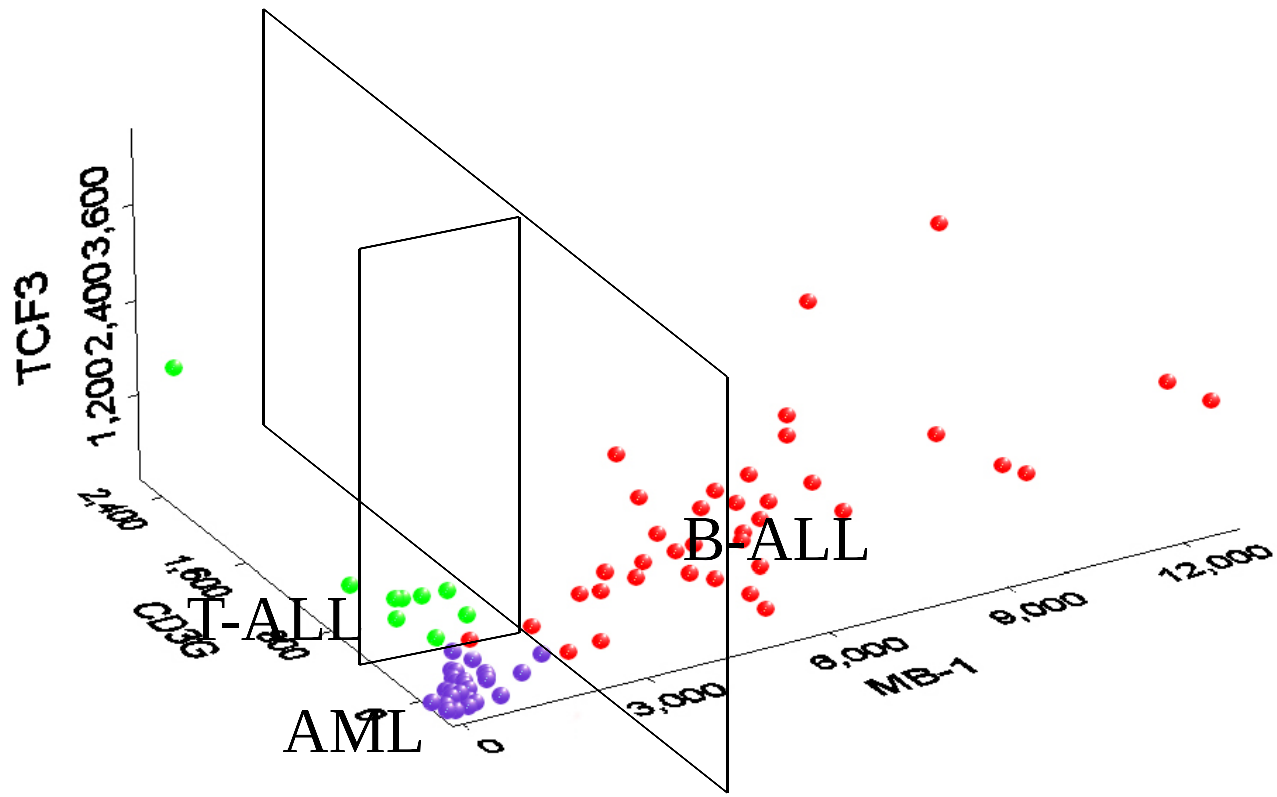
Source: http://www-genome.wi.mit.edu/cancer

Contents:

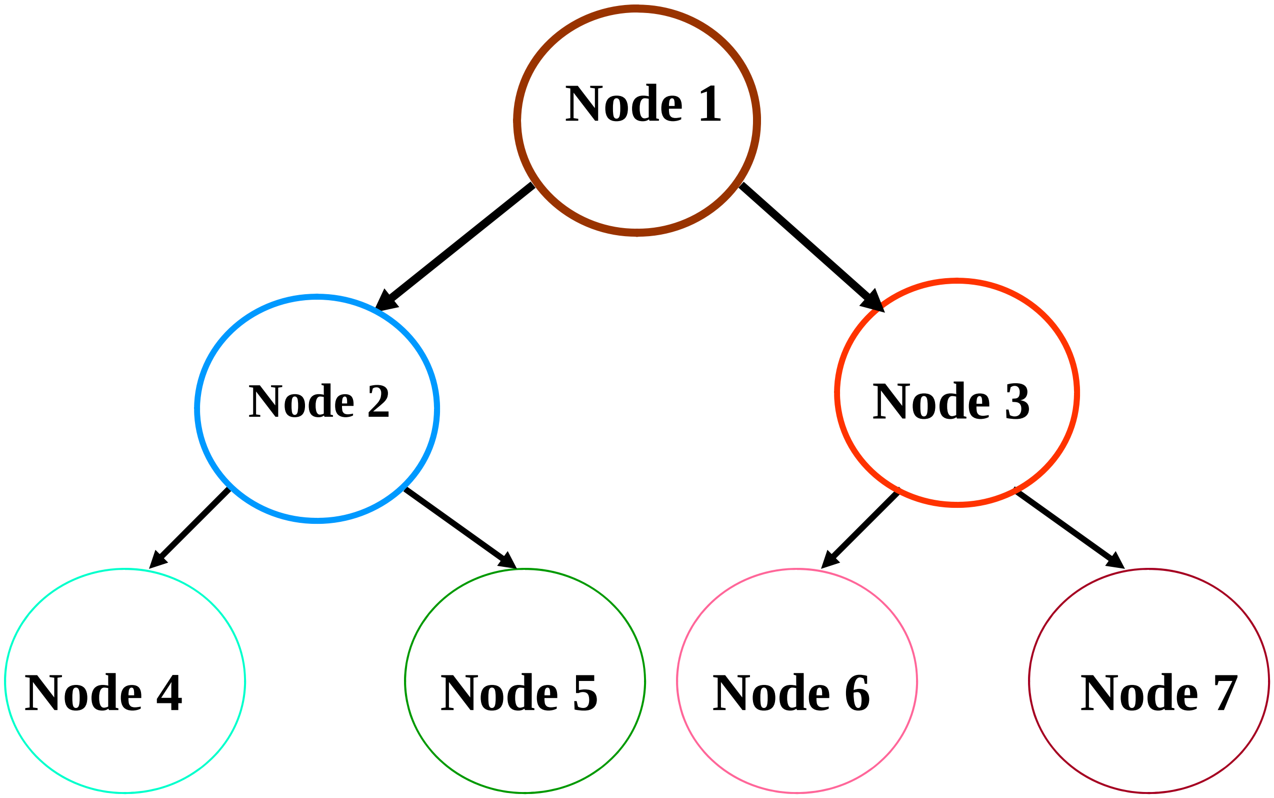
- 25 mRNA - acute myeloid leukemia (AML)
- 38 - B-cell acute lymphoblastic leukemia (B-ALL)
- 9 - T-cell acute lymphoblastic leukemia (T-ALL)
- 7,129 genes

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

3-D View


TCF3
1,200
2,400
3,600
2,400
1,600
800
CD3G
0
3,000
6,000
9,000
12,000
MB-1
T-ALL
B-ALL
AML

Tree Structure

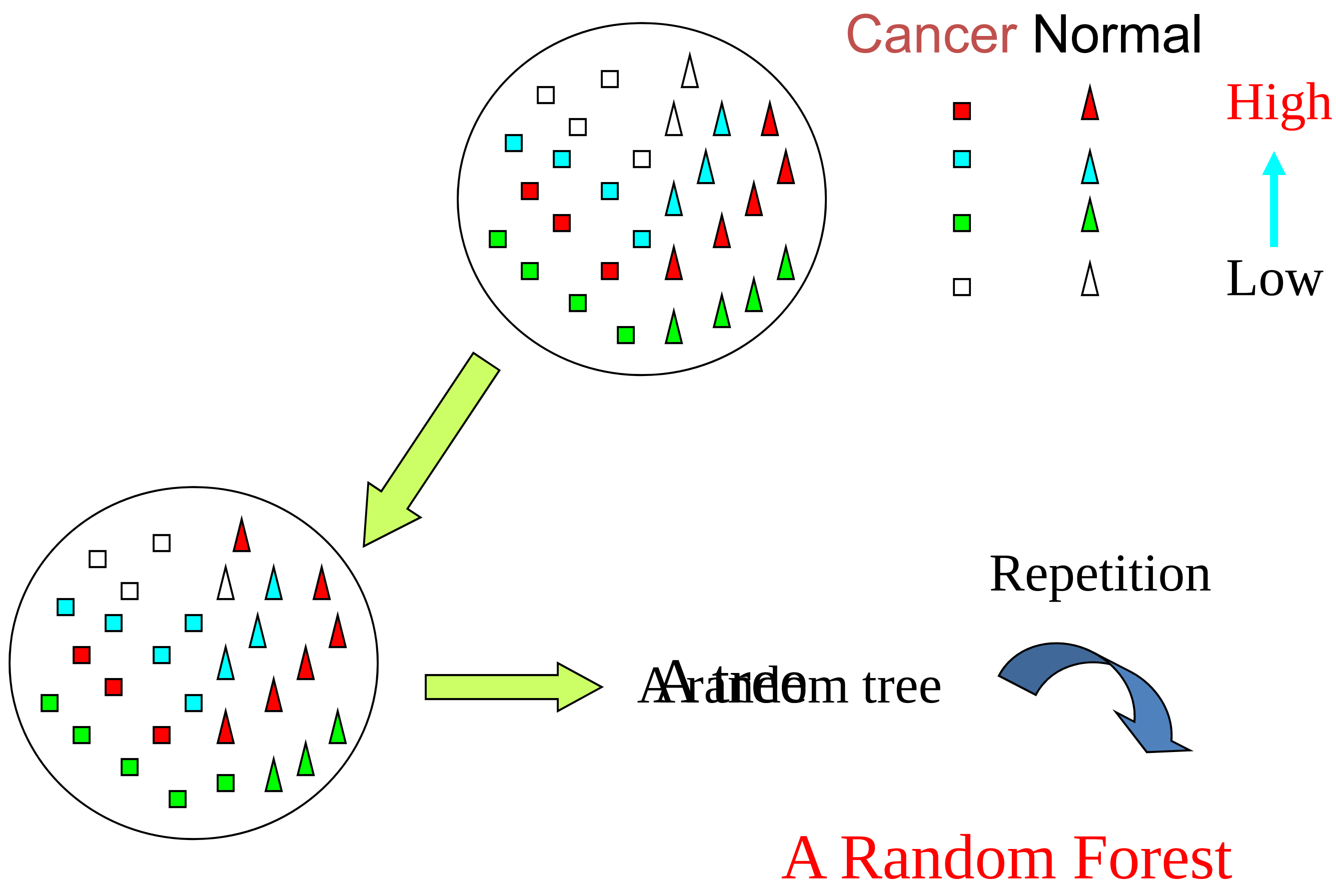


Forests

Random forests have emerged as one of the most commonly used nonparametric statistical methods in many scientific areas, particularly in analysis of high throughput genomic data.

To identify a constellation of models that collectively help us understand the data. For example, in GWAS, we can select and rank the genes that may be highly associated with a trait.

Bagging (Bootstrap Aggregating)



How Big a Forest?

- A general practice in using random forests is to generate a sufficiently large number of trees, although it is subjective as to how large is sufficient.
- Furthermore, random forests are viewed as a "black-box" because of its sheer size.

Forest Size?

- Explore whether it is possible to find a common ground between a forest and a single tree
  - retain the easy interpretability of the tree-based methods
  - avoid the problems that the tree-based methods suffer from.
- Does a forest have to be large, or how small can a forest be?

Shrink a Forest

- Shrink the forest with two objectives
  - maintain a similar (or even better) level of prediction accuracy
  - reduce the number of the trees in the forest to a manageable level

Criteria

- Two measures are considered to determine the importance of a tree in a forest
  - by prediction
  - by similarity

Prediction Based Criterion

- "by prediction" method
  - focuses on the prediction
  - A tree can be removed if its removal from the forest has the minimal impact on the overall prediction accuracy.

Prediction Based Criterion

- "by prediction" method
  - For tree T in forest F, calculate the prediction accuracy of forest $F_{(-T)}$ that excludes T.
  - $\Delta_{(-T)}$ represents the difference in prediction accuracy between F and $F_{(-T)}$.
  - The tree with the smallest $\Delta_{(-T)}$ is the least important one and hence subject to removal.

Similarity Based Criterion

- "by similarity" method
  - is based on the similarity between two trees.
  - A tree can be removed if it is "similar" to other trees in the forest.

Similarity Based Criterion

- "by similarity" method
  - The correlation of the predicted outcomes by two trees gives rise to a similarity between the two trees.
  - For tree *T*, the average of its similarities with all trees in $F_{(-T)}$, denoted by ρ_T, reflects the overall similarity between *T* and $F_{(-T)}$.
  - The tree with the highest ρ_T is the most similar to the trees in $F_{(-T)}$ and hence subject to removal.

Critical Point

- Select the optimal size sub-forest
  - Let *h(i)*, $i=1,\ldots N_f-1$, denote the performance trajectory of a sub-forest of *i* trees
    - N_f is the size of the original random forest.
  - If we have only one realization of *h(i)*, we select the optimal size sub-forest by maximizing *h(i)* over $i=1,\ldots N_f-1$.
  - If we have multiple realizations of *h(i)*, we select the optimal size sub-forest by using the 1-se rule.
- The size of this smallest sub-forest is called the critical point of the performance trajectory.

Simulation Designs

- Simulation Designs
  - For each data set, we generated 500 observations, each of which has one response variable and 30 predictors from Bernoulli distribution with success probability of 0.5.
  - Chose *v* of the 30 variables to determine the response variable.

$$y = \begin{cases} 1, & \text{if } \sum_{i=1}^{v} X_i \;/ v + \sigma > 0.5, \\ 0 & \text{Otherwise.} \end{cases}$$

    - Where σ is a random variable following the normal distribution with mean zero and variance .
  - Considered two choices for *v* (5 and 10) and two choices of σ (0.1 and 0.3).

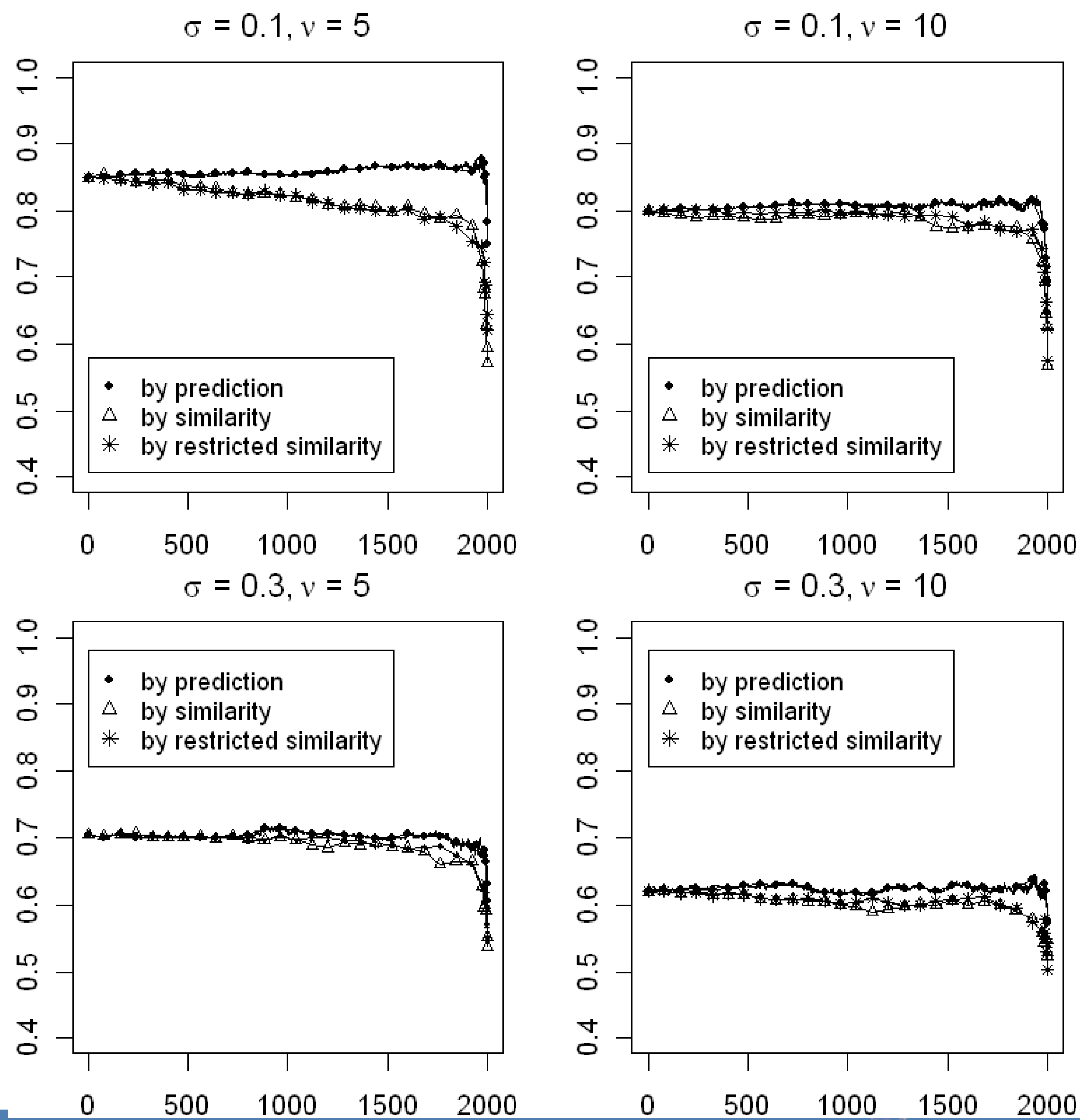
Simulation Designs

- To perform an unbiased comparison of the three tree removal measures, we simulated three independent data sets
  - The training set is used to train the initial random forest
  - The execution set is used to delete trees from the initial forest to produce sub-forests
  - The evaluation set is used to evaluate the prediction performance of the sub-forests
- The generation and use of these three data sets constituted one run of simulation, and we replicated 100 times.

Simulation Results

- Randomly selected one run of simulation and presented the stepwise change in the prediction performance in Figure 1.
- The "by prediction" method is preferable
  - It can identify a critical point during the tree removal process in which the performance of the sub-forest deteriorates very rapidly.
- The performance of the sub-forests may begin to improve before the critical point.

Prediction performance of sub-forests produced from different datasets and methods



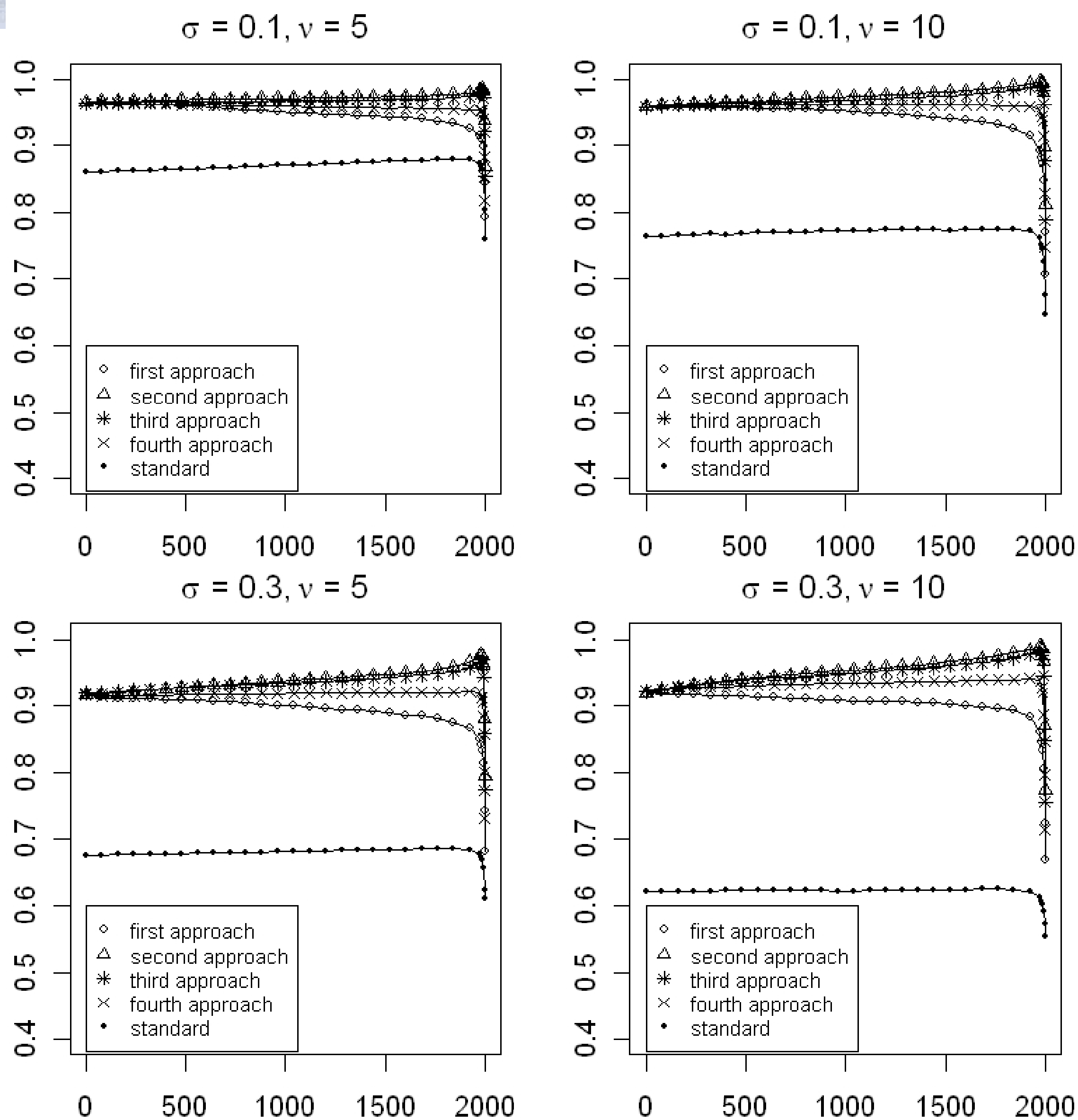
Simulation Designs

- In practice, we generally have one data set only.
- May not have the execution and evaluation data sets as in previous simulation.
- How do we select the optimal sub-forest with only one data set?

Simulation Designs

- After constructing an initial forest using the whole data set as the training data set
  - use one bootstrap data set for execution and the out-of-bag (oob) samples for evaluation.
  - use the oob samples for both execution and evaluation.
  - use the bootstrap samples for both execution and evaluation.
  - re-draw bootstrap samples for execution and re-draw bootstrap samples for evaluation.

A performance summary plot of the “by prediction” method



Simulation Results

- The performance trajectories of the four bootstrap-based approaches may not overlap with the "golden" standard.
- For the selection of the optimal sub-forest, the similarity among the trajectories is most relevant, because it could lead to the same or similar sub-forest.

Simulation Results

- Using the bootstrap samples for execution and the oob samples for evaluation is an effective sample-reuse approach to selecting the optimal sub-forest.

Application

- Dataset
  - the microarray data set of a cohort of 295 young patients with breast cancer, containing expression profiles from 70 previously selected genes.
  - previously studied by van de Vijver *et al*.
- The responses of all patients are defined by whether the patients remained disease-free five years after their initial diagnoses or not.

Application

- Method used
  - The "by prediction" measure
  - The original data set to construct an initial forest
  - A bootstrap data set for execution
  - The oob samples for evaluation.
- The procedure is replicated for a total of 100 times.
  - The oob error rate is used to compare the performance of the initial random forest and the optimal sub-forest.
  - The sizes of the optimal sub-forests fall in a relatively narrow range, of which the 1st quartile, the median, and the 3rd quartile are 13, 26 and 61, respectively.

Application

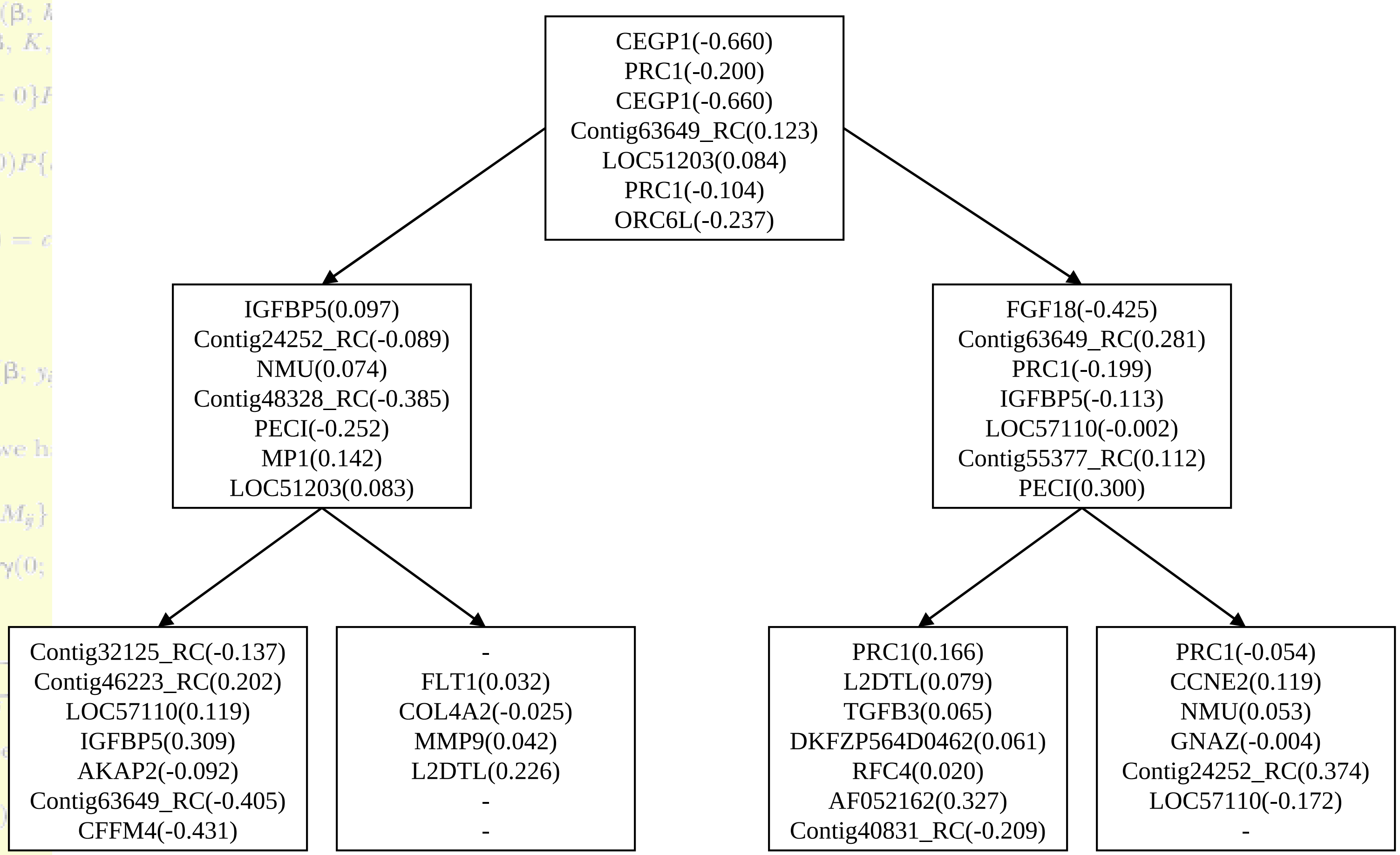
- The smallest optimal sub-forest in the 100 repetitions with the size of 7 is selected.
- As a benchmark, we used the 70-gene classifier proposed by Vijver, *et al*.

Application

- Next table presents the misclassification rates based on the oob samples.
  - The initial forest and the optimal sub-forest achieve almost the same level of performance accuracy.
  - The 70-gene classifier has an out-of-bag error rate which is much higher than those of the forests.

Comparison of prediction performance of the initial random forest, the optimal sub-forest, and a previously established 70-gene classifier

Method	Error rate	Predicted \ True	Good	Poor
Random Forest	26.0%	Good	141	17
		Poor	53	58
Sub-forest	26.0%	Good	146	22
		Poor	48	53
70-gene Classifier	35.3%	Good	103	4
		Poor	91	71



The top three layers of the optimal sub-forest consisting of seven trees

What Did We Learn?

- It is possible to construct a highly accurate random forest consisting of a manageable number of trees.
  - the size of the optimal sub-forest is in the range of tens
  - some sub-forests can even over-perform the original forest in terms of prediction accuracy
- The key advantage
  - the ability to examine and present the forests.
- The limitation
  - future samples and studies are needed to evaluate the performance of the forest-based classifiers.

Interpretation from Forest

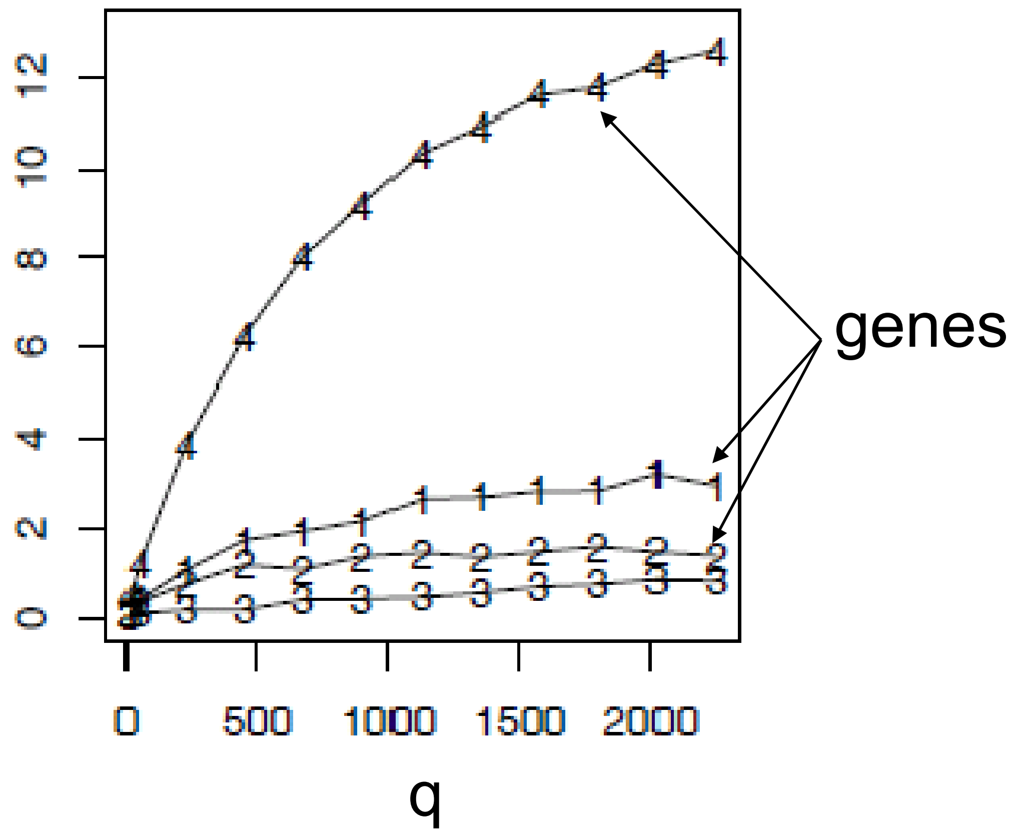
Variable Importance

Permutation importance (Breiman): For each tree in the forest, we count the number of votes cast for the correct class. Then, we randomly permute the values of variable k in the oob cases and recount the number of votes cast for the correct class in the oob cases with the permuted values of variable k. The permutation importance is the average of the differences between the number of votes for the correct class in the variable-k-permuted oob data from the number of votes for the correct class in the original oob data, over all trees in the forest.

Permutation Importance

- Not necessarily positive
- Unbounded
- The magnitudes and relative rankings can be unstable when the number, p, of predictors is large relative to the sample size.
- The magnitudes and relative rankings vary according to the number of trees in the forest and the number, q, of variables that are randomly selected to split a node

Permutation Importance



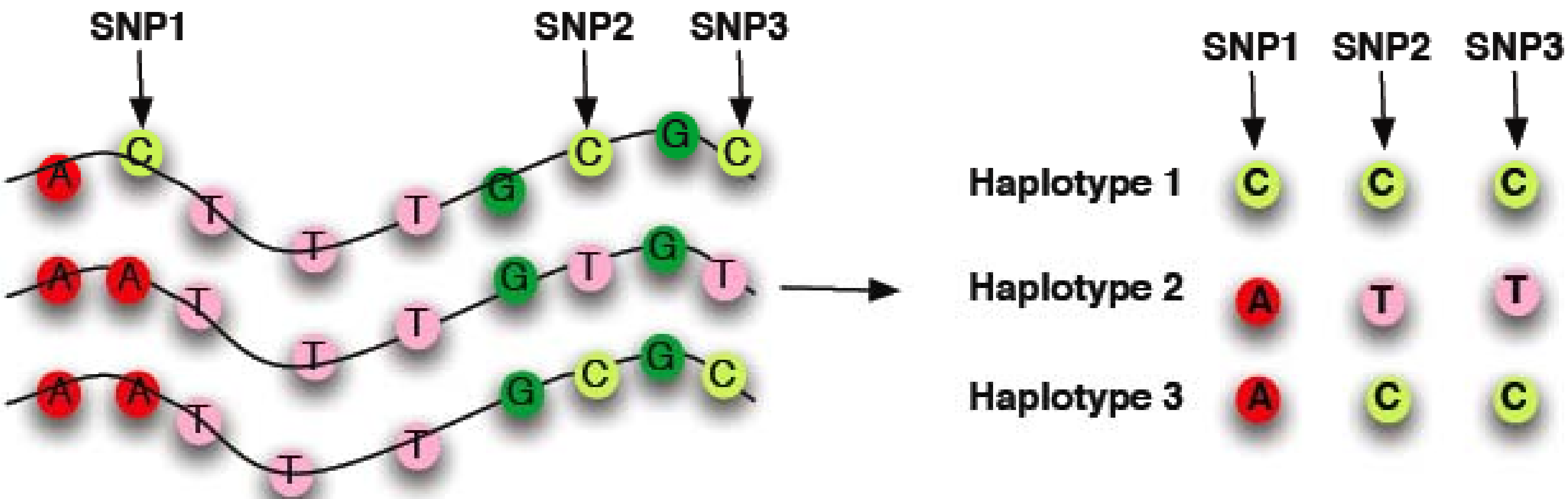
Maximal conditional chi-square (MCC)

Wang et al.(2010) introduced a maximal conditional chi-square (MCC) importance by taking the maximum chi-square statistic resulting from all splits in the forest that use the same predictor

Depth Importance

Chen et al. (2007): Whenever node t is split based on variable k, let L(t) be the depth of the node and S(k,t) be the chi-square test statistic from the variable, then $2^{-L(t)}$ S(k,t) is added up for variable k over all trees in the forest.

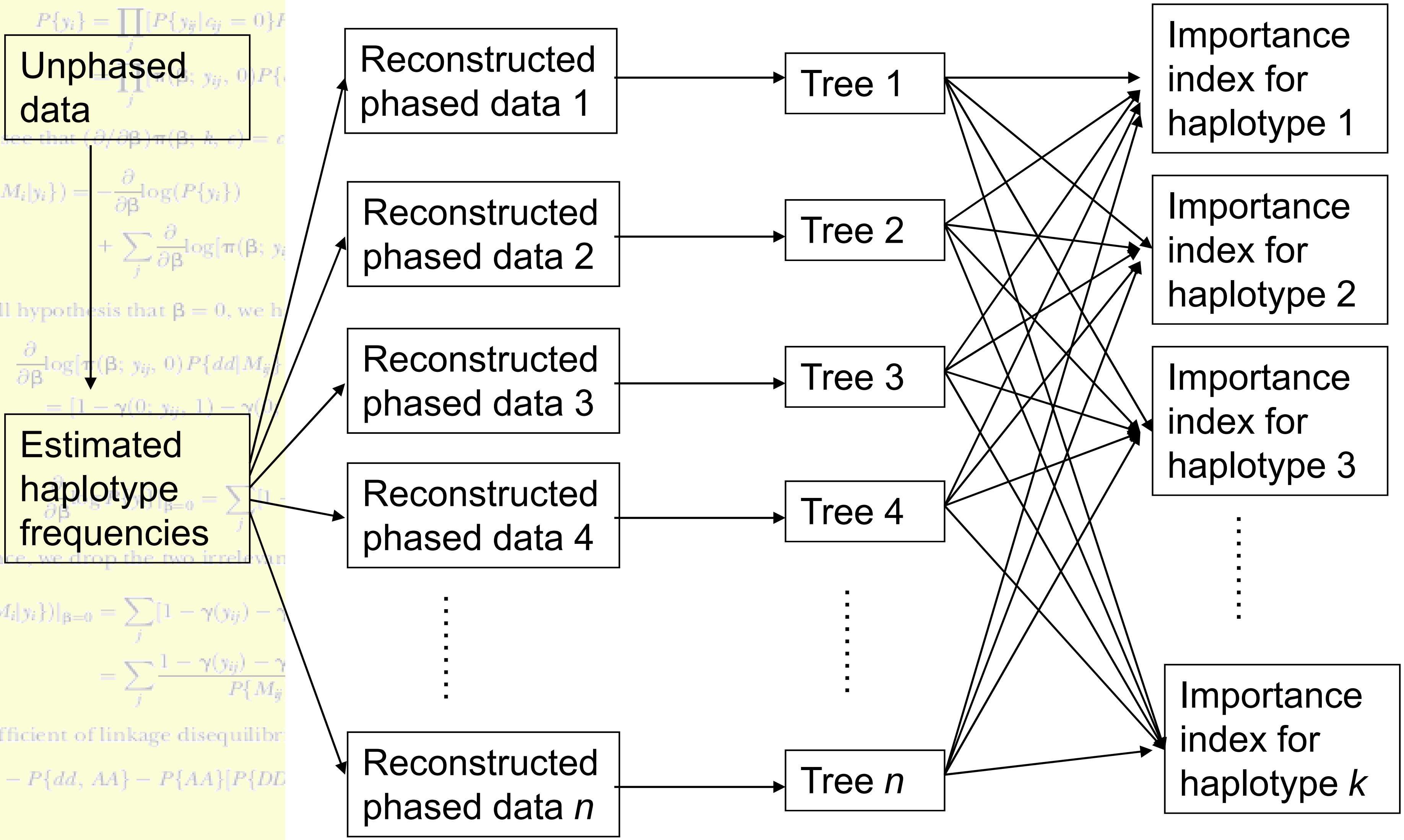
SNPs and Haplotypes



Haplotype Certainty

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

Forest Forming Scheme



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.

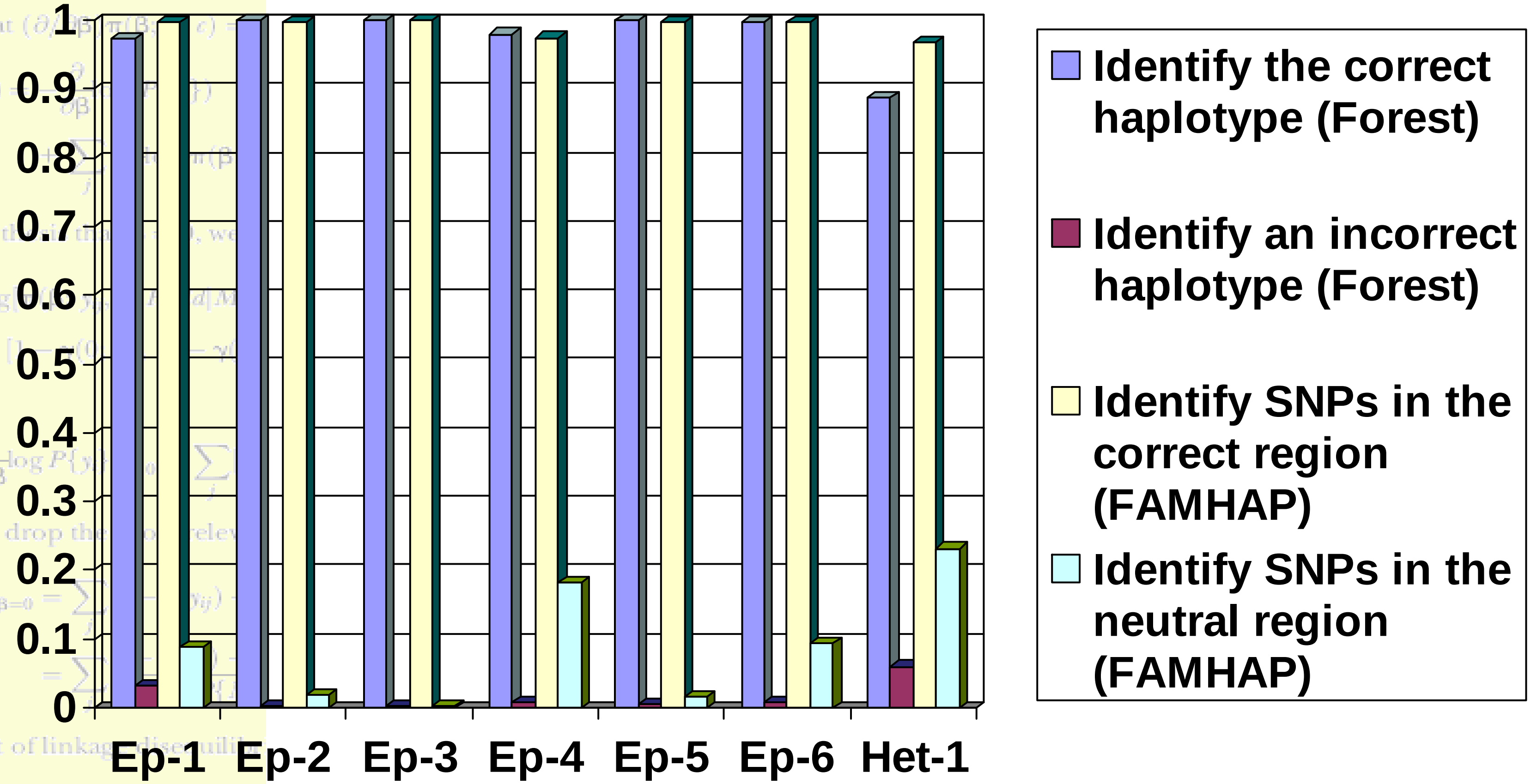
First, disease status is permuted among study subjects while keeping the genome intact for all individuals.

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

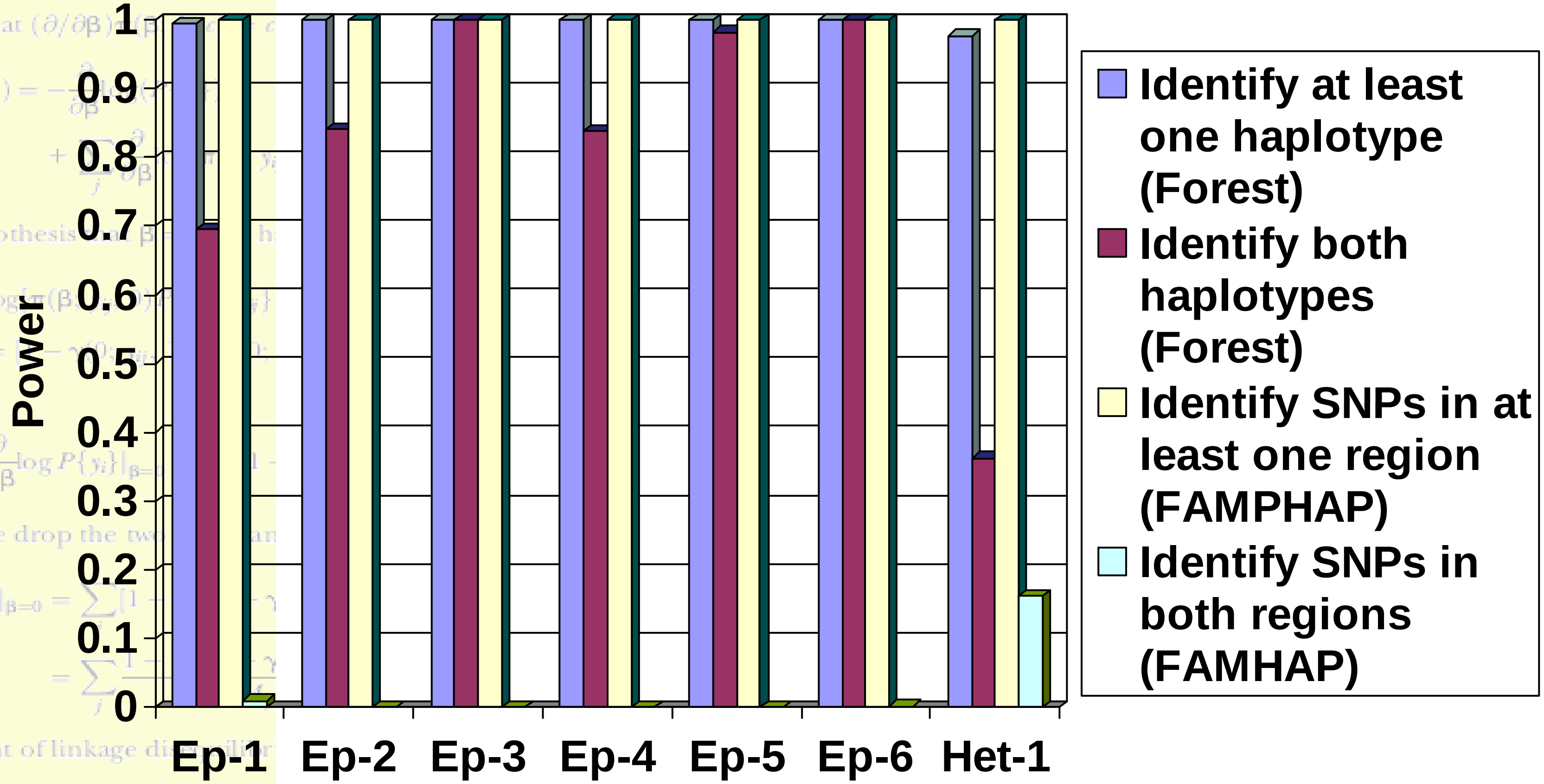
Simulation Studies (2 loci)

- **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

Result for Scenario II



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

Willows 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

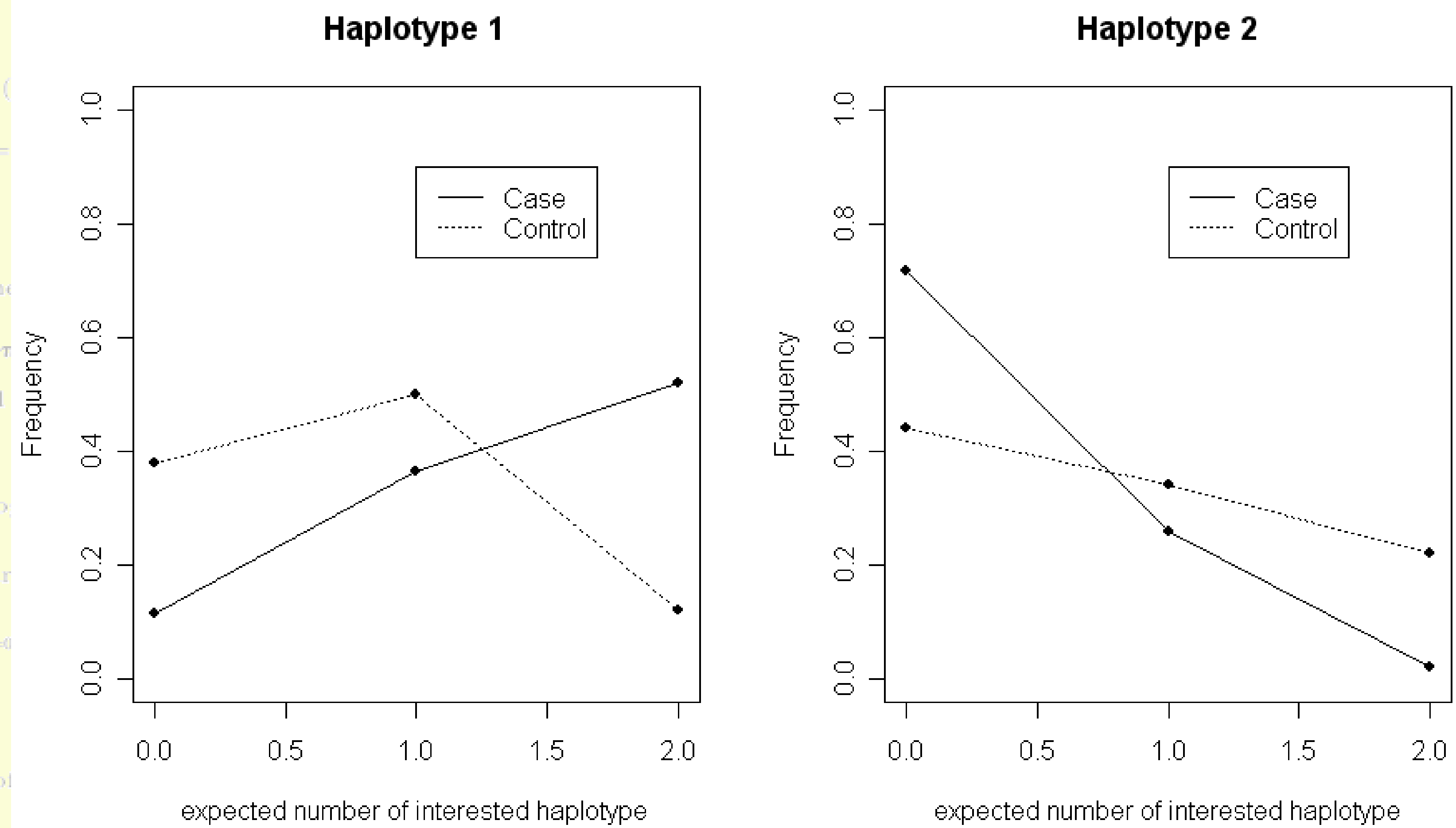
11-SNP block for rs10272438

Result

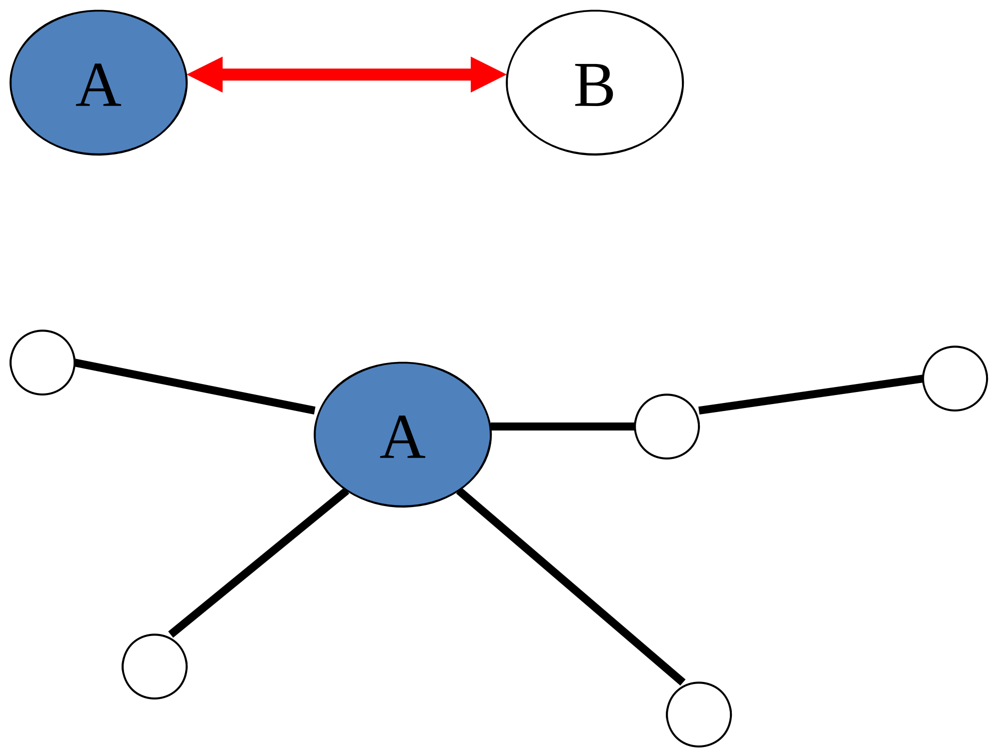
Two haplotypes are identified

- **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

Expected Frequencies

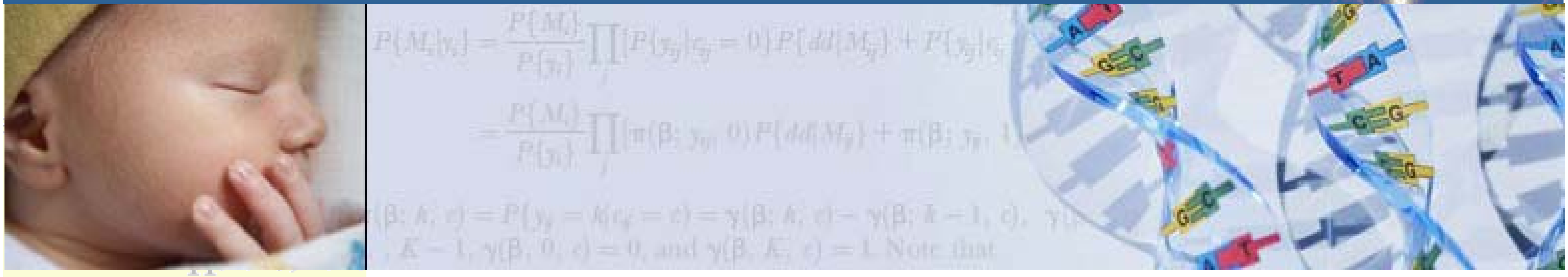


Remarks



Acknowledgement

Minghui Wang, University of Science and Technology in China
Xiang Chen, St. Jude Hospital



Thank You!