The Collaborative Center for Statistics in Science





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

ole to see that $(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$

$$og(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}log(P\{y_i\})$$

$$+ \sum_i \frac{\partial}{\partial \beta}log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0}$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

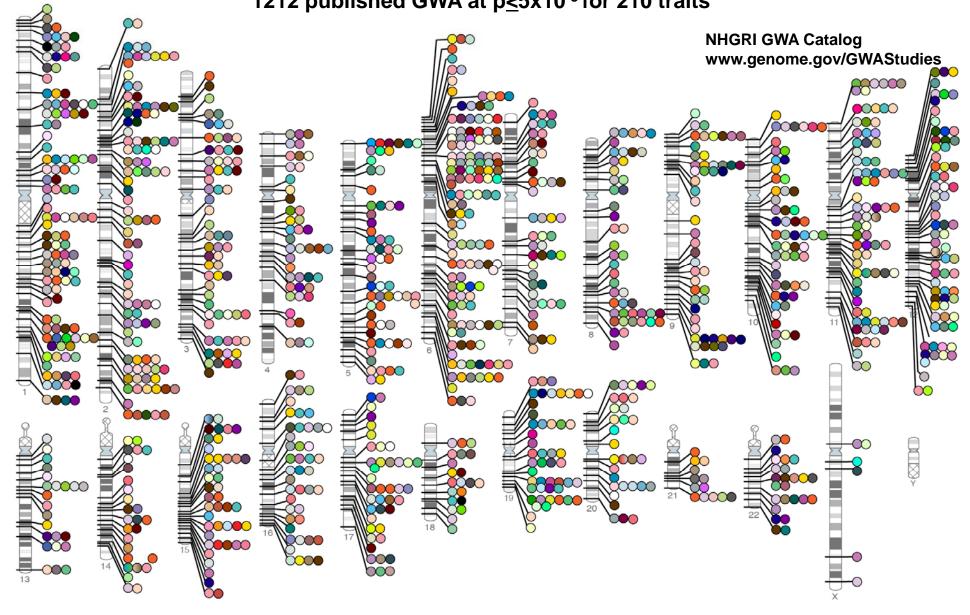
e coefficient of linkage disequilibr

$$, AA \} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Heping Zhang

School of Public Health
Yale University School of Medicine

Published Genome-Wide Associations through 12/2010, 1212 published GWA at p<5x10⁻⁸ for 210 traits



	Abdominal aortic aneurysm	0		0.22	Homocysteine levels	•	Osteoarthritis		
C	[1] - Y [1] S. B.		Cognitive function		Hypospadias	0	Osteoporosis		
	Adhesion molecules	0	Conduct disorder		Idiopathic pulmonary fibrosis		Otosclerosis		
	Adverse response to carbamapezine	_		0	IgA levels	0	Other metabolic traits		
	Adiponectin levels	_	Corneal thickness	_	IgE levels		Ovarian cancer		
			Coronary disease	\bigcirc	Inflammatory bowel disease		Pancreatic cancer		
\circ	AIDS progression	\bigcirc	Creutzfeldt-Jakob disease	\bigcirc	Intracranial aneurysm		Pain		
	Alcohol dependence		Crohn's disease		Iris color		Paget's disease	_	
	Alopecia areata		Cutaneous nevi		Iron status markers		Panic disorder		Ribavirin-induced anemia
	Alzheimer disease		Dermatitis		Ischemic stroke		Parkinson's disease	\circ	Schizophrenia
	Amyloid A levels		Drug-induced liver injury	\bigcirc	Juvenile idiopathic arthritis	\bigcirc	Periodontitis		Serum metabolites
	Amyotrophic lateral sclerosis		Endometriosis		Keloid		Peripheral arterial disease		Skin pigmentation
\subset	Angiotensin-converting enzyme activity		Eosinophil count	\bigcirc	Kidney stones	\bigcirc	Phosphatidylcholine levels		Smoking behavior
	Ankylosing spondylitis		Eosinophilic esophagitis		LDL cholesterol		Phosphorus levels		Speech perception
	Arterial stiffness		Erectile dysfunction and prostate cancer treatment	\bigcirc	Leprosy	\bigcirc	Photic sneeze	\circ	Sphingolipid levels
	Asparagus anosmia		Erythrocyte parameters	\bigcirc	Leptin receptor levels		Phytosterol levels		Statin-induced myopathy
) Asthma		Esophageal cancer		Liver enzymes		Platelet count		Stroke
	Atherosclerosis in HIV	\bigcirc	Essential tremor		Longevity		Polycystic ovary syndrome	\circ	Systemic lupus erythematosus
	Atrial fibrillation		Exfoliation glaucoma	\bigcirc	LP (a) levels		Primary biliary cirrhosis		Systemic sclerosis
	Attention deficit hyperactivity disorder	\bigcirc	Eye color traits	\bigcirc	LpPLA(2) activity and mass		Primary sclerosing cholangitis		T-tau levels
) Autism	\bigcirc	F cell distribution		Lung cancer		PR interval		Tau AB1-42 levels
	Basal cell cancer	\circ	Fibrinogen levels	\bigcirc	Magnesium levels	\bigcirc	Progranulin levels	\circ	Telomere length
	Behcet's disease		Folate pathway vitamins		Major mood disorders		Prostate cancer	\circ	Testicular germ cell tumor
\subset) Bipolar disorder		Follicular lymphoma		Malaria		Protein levels		Thyroid cancer
	Biliary atresia		Fuch's corneal dystrophy	\bigcirc	Male pattern baldness		PSA levels		Tooth development
	Bilirubin	\bigcirc	Freckles and burning		Matrix metalloproteinase levels	0	Psoriasis		Total cholesterol
	Bitter taste response		Gallstones	\bigcirc	MCP-1		Psoriatic arthritis		Triglycerides
\subset	Birth weight		Gastric cancer	\bigcirc	Melanoma		Pulmonary funct. COPD		Tuberculosis
	Bladder cancer		Glioma	\bigcirc	Menarche & menopause		QRS interval	\circ	Type 1 diabetes
	Bleomycin sensitivity		Glycemic traits	\bigcirc	Meningococcal disease		QT interval		Type 2 diabetes
	Blond or brown hair	\circ	Hair color	\bigcirc	Metabolic syndrome		Quantitative traits		Ulcerative colitis
	Blood pressure	\bigcirc	Hair morphology	\bigcirc	Migraine		Recombination rate	\circ	Urate
	Blue or green eyes		Handedness in dyslexia	\bigcirc	Moyamoya disease		Red vs.non-red hair		Venous thromboembolism
	BMI, waist circumference	\circ	HDL cholesterol	\bigcirc	Multiple sclerosis		Refractive error		Ventricular conduction
\subset	Bone density	\circ	Heart failure	\bigcirc	Myeloproliferative neoplasms		Renal cell carcinoma	\bigcirc	Vertical cup-disc ratio
	Breast cancer		Heart rate	\bigcirc	N-glycan levels		Renal function		Vitamin B12 levels
	C-reactive protein		Height	\bigcirc	Narcolepsy		Response to antidepressants		Vitamin D insuffiency
	Calcium levels		Hemostasis parameters	0	Nasopharyngeal cancer		Response to antipsychotic therapy		Vitiligo
	Cardiac structure/function		Hepatic steatosis		Neuroblastoma		Response to hepatitis C treat		Warfarin dose
	Carnitine levels	\circ	Hepatitis		Nicotine dependence		Response to metaformin		Weight
	Carotenoid/tocopherol levels		Hepatocellular carcinoma	\bigcirc	Obesity		Response to statin therapy	\circ	White cell count
		\circ	Hirschsprung's disease		Open angle glaucoma	\bigcirc	Restless legs syndrome		YKL-40 levels
	Cerebral atrophy measures	\circ			Open personality	\bigcirc	Retinal vascular caliber		
	Chronic lymphocytic leukemia	0	Hodgkin's lymphoma	0	Optic disc parameters	0	Rheumatoid arthritis		

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P^2$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

$-\frac{P\{y_i\}}{P\{y_i\}}\prod_{j}|P\{y_j|a_j=0\}P$ - Challenges



3;
$$k$$
, c) = $P\{y_{ij} = k | c_{ij} = c\} = \gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, K, K) \}$

$$P\{y_i\} = \prod_j [P\{y_{ij} | q_j = 0\}]$$

=
$$\prod_i [\pi(\beta; y_{ij}, 0)]P\{$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,c)=$$

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij} = [1 - \gamma(0; y_{ij}, 1) - \gamma(0 + \gamma(0))] + \gamma(0) + \gamma(0)$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1$$

enience, we drop the two irreleva-

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DI]$$

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?

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{P\{y_{i}\}}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}P]^{2} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta;|y_{ij},|0)P\{c_{ij}\}] \end{aligned}$$

Classic Modeling vs Genomic Association Analysis

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\}$ = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, P\{y_i\} = \prod [P\{y_{ij} | e_{ij} = 0\} P\{y_i\}]$

$$= \prod_{j} (\pi(\beta; y_{ij}, 0)) P\{$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; \ y_{ij}, \ \boldsymbol{0}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{0}; \ y_{ij}, \ \boldsymbol{1}) - \gamma(\boldsymbol{0}; \ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irreleval

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

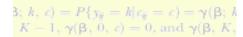
$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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.

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\} P P\{M_{i}|y_{i}\} &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{i}|\beta_{i}, k, c) - P\{y_{i}|\beta_{i}, k, c\}] + P\{y_{i}|\beta_{i}, k, c\} - P\{y$$

Outline



$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] P\{y_i\} = \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{y_i\}$$

ole to see that
$$(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \text{log}[\boldsymbol{\pi}(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ & = [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

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



$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

-P{M} Π(π(β:)η, 0)P{



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | a_{ij} = 0\} P \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P \{ e^{-i\beta_{ij}} \} P \} \end{split}$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) = c$$

$$\begin{aligned}
\log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\
&+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]
\end{aligned}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

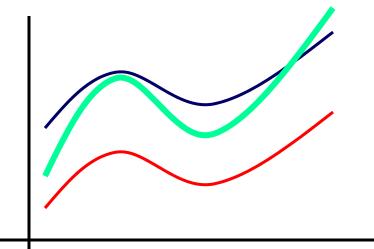
= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

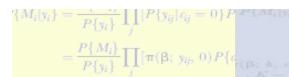
$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

Diseases that do not follow Mendelian Inheritance Pattern

Genetic factors, Environment factors, G-G and G-E $+\sum \frac{\partial}{\partial \beta} \log[\pi(\beta)]$ interactions

^θ_{λβ}log[π(β, y₀, 0)P{dd[M_g]</sub> Interactions: effects that deviate from the additive effects of single effects





SNP and Complex Traits

3; k, e) = $P\{y_{\theta} = k | e_{\theta} = e\} = \gamma(\beta; k K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij}|q_j\}] = \prod_{j} [\pi(\beta; y_i)]$$

ole to see that $(\partial/\partial \beta)\pi(\beta; k)$

$$og(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}log(P\{y_i\})$$

 $+\sum_i \frac{\partial}{\partial \beta}log$

he null hypothesis that $\beta =$

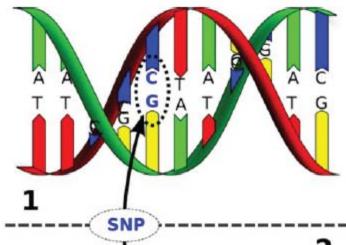
$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0)P\{ -\frac{1}{\beta} \}]$$

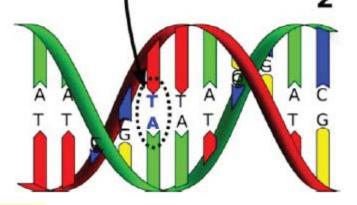
$$= [1 - \gamma(0; y_{ij}, 1)]$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0}$$

enience, we drop the two ir

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma]$$
$$= \sum_j \frac{1 - \gamma}{1 - \gamma}$$







e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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

$$\begin{aligned} \{M_i|y_i\} &= \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij}=0\}P P\{M_i|y_i\}] &= \frac{P\{M_i\}}{P} \\ &= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{i(\beta; k, c)} - P\{M_i\}]\} \\ \end{aligned}$$

Regression Approach



$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{ij}, c_{ij}, c_{ij}\}] = 0$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta;\,h,\,c)=c$$



$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\}] = [1 - \gamma(0; y_{ij}, 1) - 25;$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \frac{26}{10}$$

senience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$
$$= \sum_{j} \frac{1 - \gamma(y_{ij})}{P\{M_{ij}\}} \frac{2\gamma}{M_{ij}}$$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DE]$$













$$\begin{aligned} \{M_i|y_i\} &= \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P \underbrace{P\{M_i\}y_i\}}_{P\{y_i\}} &= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\boldsymbol{\beta}; y_{ij}, 0)P\{c_{ij}, k, c) = P\{y_{ij}\}\} \end{aligned}$$

Recursive Partitioning



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} P\{y_{ij} | c_{ij} = 0\} P\{y_{ij}, 0\} P\{y_{i$$

ole to see that
$$(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{log}[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevan

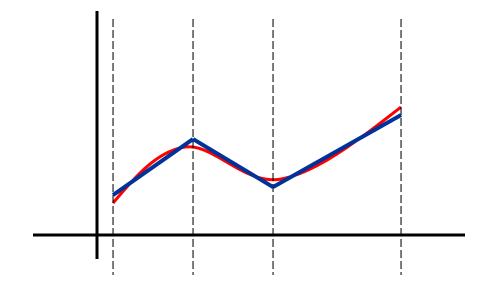
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

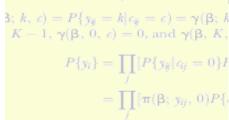
A technique to identify heterogeneity in the data and fit a simple model (such as constant or linear) locally, and this avoids pre-specifying a systematic component.



$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\}P$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}] P\{x_{ij}\}$

$\frac{\{M_i|y_i\} = \frac{1}{P\{y_i\}}\prod_{j}|P\{y_{ij}|c_{ij}=0\}P}{-\frac{P\{M_i\}}{P\{y_i\}}\prod_{j}|\pi(\beta;\;y_{ij},\;0)P\{c_{ij}=0\}P}$ Leukemia Data



ple to see that
$$(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)] - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)] + \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 -$$

enience, we drop the two irrelevar

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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 $-\sum_{p \in \mathcal{P}(M)} -\sum_{p \in \mathcal{P}(M)} -\sum_{$

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\}P \underbrace{P\{M_{i}\}y_{i}\}}_{P\{y_{i}\}} &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} |\pi(\beta; y_{ij}, 0)P\{c_{i}\}_{E=1, P(\beta; \beta)} &= \frac{P\{y_{i}\}}{E=1, P(\beta; \beta)} \end{aligned}$$

3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{ and$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{ c_{ij} = 0 \} F = 0 \} F$$

ole to see that (∂/∂β)π(β; k, ε) = ε

he null hypothesis that $\beta = 0$, we have

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{0}) P\{dd|M_{ij}\} \\ &= [1 - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1})] + \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1})] + \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1})] + \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}; \ \boldsymbol{0}) - \boldsymbol$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0}$$

enience, we drop the two irrelevan

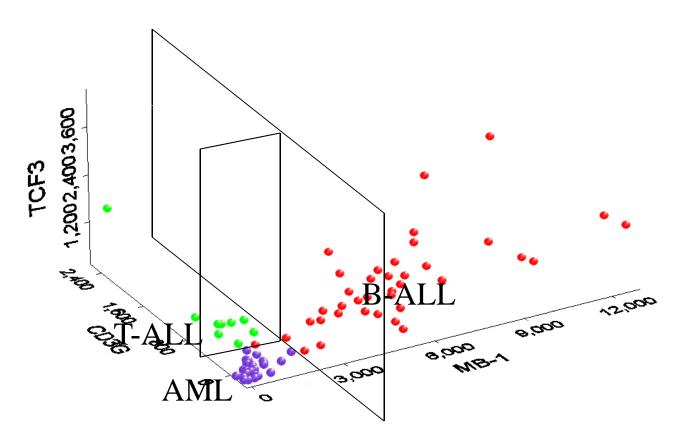
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

3-D View





$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\} P$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0) P\{a_i\}]$

$-\frac{P\{M_i\}}{P\{y_i\}}\prod_{j}[\pi(\beta;y_i,0)P\{c_{(\beta;A_i,0)}-P\{y_i-A_{(\beta,A_i,0)}\}}]Tree Structure$



$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

ole to see that $(\partial/\partial \beta)\pi(\beta; k, c) = c$

$$\begin{aligned}
\log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\
&+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]
\end{aligned}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)] - \gamma(0; y_{ij}, 1) - \gamma(0;$$

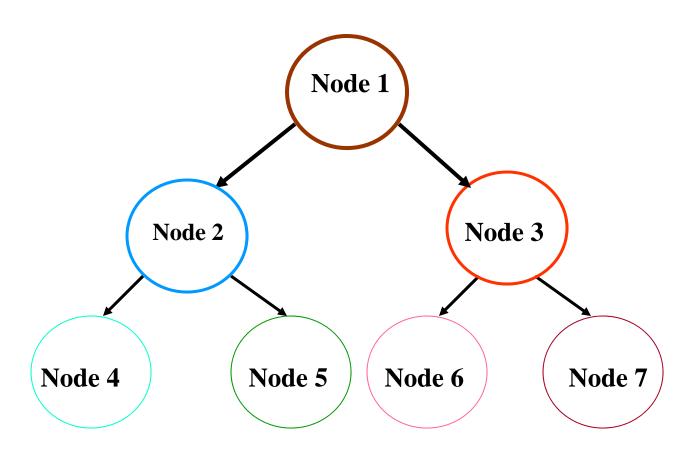
$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{\beta}]$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$





$$\{M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\} P P\{M_{i}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} |\pi(\beta; y_{ij}, 0)P\{c_{i}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} |\pi(\beta; y_{ij}, 0)P\{c_{i}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} |\pi(\beta; y_{ij}, 0)P\{c_{j}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} |\pi(\beta; y_{i})P\{c_{j}\} = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} |\pi(\beta; y_{ij}, 0)P\{c_{j}$$



3;
$$k, c) = P\{y_{ij} = k | c_{ij} = c\} = \gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, -1, \gamma(\beta, 0, c) = 0)\}$$

$$P\{y_i\} = \prod_j [P\{y_{ij}|a_{ij} = 0\}F]$$

= $\prod_j [\pi(\beta; y_{ij}, 0)P\{$

$$g(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta} log(P\{y_i\})$$

$$+ \sum_i \frac{\partial}{\partial \beta} log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0}$$

renience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

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.

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\}P^{PP} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}] \end{aligned}$$

Bagging (Bootstrap Aggregating)

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, i)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

ole to see that $(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \text{log}[\boldsymbol{\pi}(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ & = [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

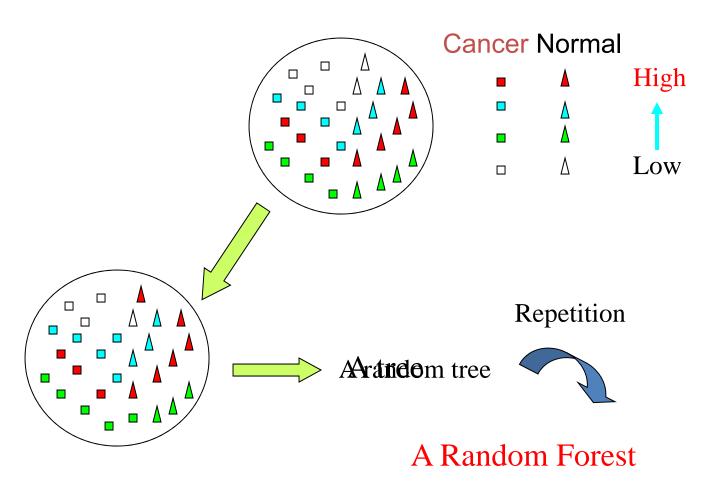
enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$



$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_i\}]$

$-\frac{P\{y_i\}}{P\{y_i\}}\prod_{j=1}^{|P\{y_i|Q_j=0\}P}$ How Big a Forest?

3; k, e) = $P\{y_{\theta} = k | e_{\theta} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K)\}$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | c_{ij} = 0\} P \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P \{ e^{-i\beta_{ij}} \} P \} \end{split}$$

ole to see that $(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$

he null hypothesis that $\beta = 0$, we h

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd] M_{ij}\}$$

$$= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; y_{ij}, 1)]$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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.





$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Forest Size?



3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | q_j = 0\} I]$$

= $\prod_{j} [\pi(\beta; y_{ij}, 0) P\{$

ble to see that $(\partial/\partial\beta)\pi(\beta; k, c) = c$

$$\begin{aligned} \log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\ &+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)] \end{aligned}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$

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?

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_i\}]$

Shrink a Forest



3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E) = 0\}$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

ole to see that $(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$

$$\begin{aligned}
\log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\
&+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]
\end{aligned}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}]$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

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

$$\{M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\} P P\{M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\} P\{M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} |\pi(\beta; y_{ij}, 0) P\{a P\{y_{ij} = \alpha\} = \gamma(\beta; k, c) = \gamma$$

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, \beta, \epsilon) = 0, \text{ and } \gamma(\beta, K, \epsilon)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\}] P\{y_{ij} | a_{ij} = 0\} P\{y$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we have

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0)P\{dd|M_{ij}\}] - \text{by similarity}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA - P\{dd, AA\} - P\{AA\}[P\{DE]$$

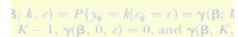
Two measures are considered to determine the importance of a tree in a forest

- by prediction

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Prediction Based Criterion



$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] P\{y_{ij} | c_{ij} = 0\} P\{y$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,\epsilon)=\epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_{i} [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0} = \sum_{i} [1 - \frac{\partial}{$$

renience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DE]$$

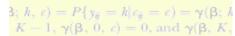
ые то see that (0/0β) π(β: A. ♣— "by prediction" method

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

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0) P\{z_i\}]$

Prediction Based Criterion



$$P\{y_i\} = \prod_j [P\{y_i|c_j = 0\}]$$

$$= \prod_j [\pi(\beta; y_{ij}, 0)P\{$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,\epsilon)=\epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{0}) P\{dd|M_{ij}\} \\ &= [1 - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1})] + \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0}; \ \boldsymbol{y}_{ij}, \ \boldsymbol{1})] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{\beta}]$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

"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.

$$\{M_i|y_i\} = \frac{G}{P\{y_i\}} \prod_j [P\{y_{ij}|z_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{z_i\}]$

Similarity Based Criterion



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{$$

be to see that (θ/θβ)π(β: λ. •) = "by similarity" method

- $og(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}log(P\{y_i\})$ $+\sum_{i}\frac{\partial}{\partial \beta}\log[\pi(\beta;y_{i})]$

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; \, y_{ij}, \, \boldsymbol{\theta}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{\theta}; \, y_{ii}, \, \boldsymbol{\theta}) - \gamma(\boldsymbol{\theta}; \, \boldsymbol{\theta})] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

renience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$\{AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

- is based on the similarity between two trees.
- be null hypothesis that β = 0, we h A tree can be removed if it is "similar" to other trees in the forest.

$$\{M_i|y_i\} = \frac{0}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P\{y_i\} = \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{a\}]$$

Similarity Based Criterion



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | q_j = 0\}] I$$
$$= \prod_{j} [\pi(\beta; y_{ij}, 0)] P\{.$$

by similarity" method

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

 $\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial\beta}\log(P\{y_i\})$
 $+\sum_j \frac{\partial}{\partial\beta}\log[\pi(\beta; y_i)]$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{\beta}] = \sum_i [1 - \frac{1}{\beta}$$

enience, we drop the two irrelevat

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

- 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_{τ} is the most similar to the trees in $F_{(-T)}$ and hence subject to removal.

$$\begin{cases} M_{i}|y_{i}\} = \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\}P \\ = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} (\pi(\beta; y_{ij}, 0)P\{c_{ij}, c_{ij} = 0\}P \end{cases}$$

$$\begin{aligned} & = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} (\pi(\beta; y_{ij}, 0)P\{c_{ij}, c_{ij} = 0\}P \\ & = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} (\pi(\beta; y_{ij}, 0)P\{c_{ij}, c_{ij} = 0\}P \end{aligned}$$

$$\begin{aligned} & = \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} (\pi(\beta; y_{ij}, 0)P\{c_{ij}, c_{ij} = 0\}P \end{aligned}$$



3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$P\{y_i\} = \prod_j [P\{y_{ij} | a_{ij} = 0\}] P$$
$$= \prod_j [\pi(\beta; y_{ij}, 0)] P\{.$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,c)=$$

$$og(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}log(P\{y_i\})$$

$$+ \sum_{i} \frac{\partial}{\partial \beta}log[\pi(\beta; y_i - y_i)]$$

he null hypothesis that
$$\beta = 0$$
, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; \, y_{ij}, \, \boldsymbol{\theta}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{\theta}; \, y_{ij}, \, \boldsymbol{\theta}) - \gamma(\boldsymbol{\theta}; \, \boldsymbol{\theta})] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{j} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_i\}}$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

Select the optimal size sub-forest

- I response Let h(i), $i=1,...N_f$ -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 $i=1,...N_f-1.$
- $-11-\gamma^{(0)}$ 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.

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\} P P\{M_{i}\}\} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{i}, k, c) = K - 1] \end{aligned}$$

Simulation Designs



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0$, and $\gamma(\beta, K, K, e)$

$$P\{y_i\} = \prod_j [P\{y_{ij} | c_{ij} = 0\}]$$

=
$$\prod_j [\pi(\beta; y_{ij}, 0)]P\{y_{ij} = 0\}$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) = c$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

- Ππ(β: χης Θ)P{ 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.
- $+\sum_{i=0}^{n} \frac{\partial}{\partial \beta^{\log(\pi(\beta))}}$ Chose ν 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 ν (5 and 10) and two choices of σ (0.1 and 0.3).

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0) P\{c_i\}]$

Simulation Designs



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_j [P\{y_{ij} | a_{ij} = 0\}]$$
$$= \prod_j [\pi(\beta; y_{ij}, 0)]P\{$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta;\,\hbar,\,c)=$$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{log}[\boldsymbol{\pi}(\boldsymbol{\beta};\; \mathbf{y}_{ij},\; \boldsymbol{0}) P\{dd|M_{ij}\} \\ &= [1 - \boldsymbol{\gamma}(\boldsymbol{0};\; \mathbf{y}_{ij},\; \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{0};\; \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma_j]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma_j}{P\{M_{ij}\}}$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

To perform an unbiased comparison of the three tree removal measures, we simulated 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.

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_i |\pi(\beta; y_{ij}, 0)P\{c_{ij}\}|$

Simulation Results



3;
$$k$$
, e) = $P\{y_{\theta} = k | e_{\theta} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | q_j = 0\}]$$

=
$$\prod_{i} [\pi(\beta; y_{ij}, 0)]P\{$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) =$$

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}\log(P\{y_i\}) \quad \Phi$$

$$+ \sum_i \frac{\partial}{\partial \beta}\log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{log}[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 -$$

enience, we drop the two irreleval

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

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.



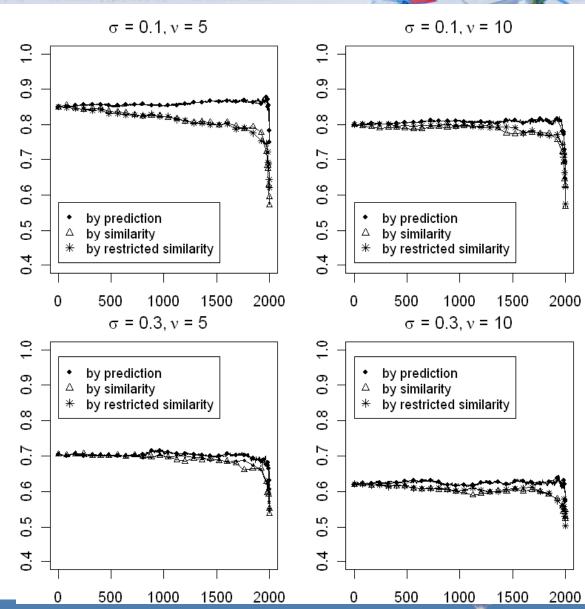
3;
$$k$$
, ϵ) = $P\{y_{ij} = k | \epsilon_{ij} = \epsilon\} = \gamma(\beta; k K - 1, \gamma(\beta, 0, \epsilon) = 0, \text{ and } \gamma(\beta, K, E)\}$

$$P\{y_{i}\} = \prod_{i} [P\{y_{ij} | \epsilon_{ij} = 0\}F]$$

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

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1 - \frac{1}{2}]$$
 senience, we drop the two irrelevants
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

$$= \sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$$
 we coefficient of linkage disequilibrity,
$$AA\} - P\{dd, AA\} - P\{AA\}[P\{DE\}]$$



$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\} P^{3} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{i}\}] \end{aligned}$$

Simulation Designs



3;
$$k, c$$
) = $P\{y_{ij} = k | c_{ij} = c$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, c)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} P\{y_{ij} | c_{ij} = 0\} P\{y_$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,\epsilon)=$$

$$\begin{aligned}
\log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\
&+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]
\end{aligned}$$

he null hypothesis that
$$\beta = 0$$
, we n

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{\beta}]$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilib

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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?

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\} P P \{M_{i}\}\} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{i}, k_{i}, c_{j}\}] \end{aligned}$$

Simulation Designs



3;
$$k$$
, ϵ) = $P\{y_{ij} = k | \epsilon_{ij} = \epsilon\}$ = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, \epsilon) = 0, \text{ and } \gamma(\beta, K, \epsilon)$

$$P\{y_i\} = \prod_j [P\{y_{ij}|\, q_j = 0\} F$$

$$= \prod_{j} [\pi(\beta; y_{ij}, \bullet)]P$$

to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) =$$

$$\begin{aligned} \log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\ &+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)] \end{aligned}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} \text{log}[\boldsymbol{\pi}(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \boldsymbol{\gamma}(0; y_{ij}, 1) - \boldsymbol{\gamma}(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \! \log P\{y_i\}|_{\beta=0} = \sum_j [1$$

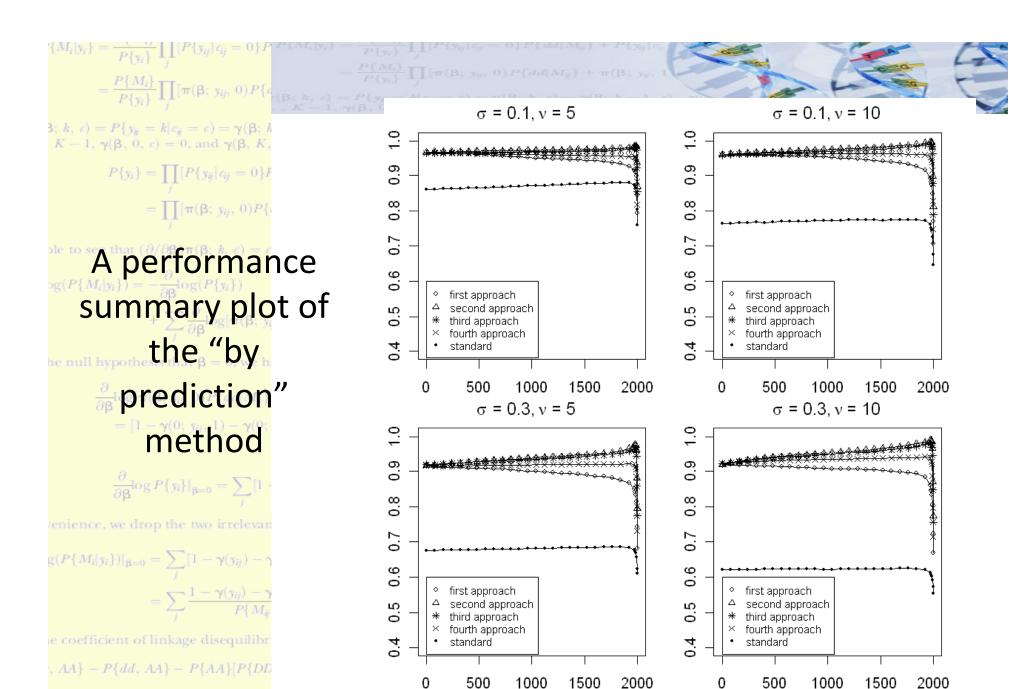
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{j} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

After constructing an initial forest using the whole data set as the training data set

- $-\sum_{l=0}^{n} \frac{\partial l}{\partial \beta^{\log l}} = 0$ 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.
 - $\frac{\partial}{\partial B} \log P(y)|_{B=0} = \sum_{i=1}^{n}$ use the bootstrap samples for both execution and evaluation.
- re-draw bootstrap samples for execution and re-



The Collaborative Center for Statistics in Science



$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}] P\{x_{ij}\}$

Simulation Results



3;
$$k, e$$
) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e)$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | q_{ij} = 0\} P_i] \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P_i] \end{split}$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) =$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; \, y_{ij}, \, \boldsymbol{\theta}) P\{dd| \boldsymbol{\mathcal{M}}_{ij} \\ &= [1 - \gamma(\boldsymbol{\theta}; \, y_{ij}, \, \boldsymbol{\theta}) - \gamma(\boldsymbol{\theta}) \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1$$

$$\frac{|g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} (1 - \gamma(y_{ij}) - \gamma)}{\sum_{j} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}} \text{ forest.}$$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

The performance trajectories of the four bootstrap-based approaches may not overlap with the "golden" standard.

For the selection of the optimal subforest, the similarity among the trajectories is most relevant, because it could lead to the same or similar sub-

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P^{s}$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Simulation Results



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}]P\{y_{ij} = 0\}P\{y_{ij} = 0\}P\{y$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) = 0$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA \} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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

$$\{M_i|y_i\} = \frac{G}{P\{y_i\}} \prod_j |P\{y_{ij}|e_{ij} = 0\}P\}$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{e_i]$

Application



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{ and } \gamma($

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}]$$

$$= \prod_{j} [\pi(\beta; y_{ij}, 0)] P$$

<u>- π-⊪⊶•</u> Dataset

ole to see that
$$(\partial/\partial\beta)\pi(\beta; k, e) = e$$

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial\beta}\log(P\{y_i\})$$

$$+ \sum_j \frac{\partial}{\partial\beta}\log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we h

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\}]$$

- 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.

 $AA - P\{dd, AA\} - P\{AA\}[P\{DD]$

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{\Box}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}P] \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta;|y_{ij},|0)P\{c_{i}\}] \end{aligned}$$

$= \frac{1}{P\{y_i\}} \prod_{j} P\{y_{ij} \mid c_{ij} = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c] = 0\} P$ $= \frac{P\{M_i\}}{P\{y_i\}} \prod_{j} [\pi(\beta; y_{i$



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{$

$$P\{y_i\} = \prod_{i} [P\{y_{ij} | c_{ij} = 0\}] P\{y_{ij} | c_{ij} = 0\} P\{y$$

$$g(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}log(P\{y_i\})$$

$$+ \sum_{j} \frac{\partial}{\partial \beta} log[\pi(\beta)]$$

The "by prediction" measure

- The original data set to construct an initial forest
 - $+\sum_{j}\frac{\partial}{\partial\beta}\log[\pi(\beta;y]]$ A bootstrap data set for execution
 - The oob samples for evaluation.

he null hypothesis that $\beta = 0$, we l

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{\partial}{\partial \beta} \log P\{y_i\}]_{\beta=0}$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

The procedure is replicated for a total of 100 times.

- $\frac{1}{1-\gamma(0)(3)(3)(1)-\gamma(0)}$ 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.

$$\{M_i|y_i\} = \frac{e^{-i\beta}}{P\{y_i\}} \prod_j |P\{y_{ij}|e_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_i [\pi(\beta; y_{ij}, 0)P\{a_i\}]$

Application



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} P\{y_{ij} | c_{ij} = 0\} P\{y_$$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; h, c) =$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd] M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{aligned}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \mathbf{Q}].$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$

The smallest optimal sub-forest in the 100 repetitions with the size of $+\sum_{\beta\beta}^{\beta}\log \pi(\beta)$ 7 is selected.

As a benchmark, we used the 70gene classifier proposed by Vijver, et

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|\epsilon_{ij} = 0\}P^i$$

 $= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{\epsilon_i\}]$

Application



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{ and$

$$P\{y_i\} = \prod_i [P\{y_{ij}|c_{ij} = 0\}P$$

Next table presents the misclassification ole to see that $(\partial/\partial \beta)\pi(\beta; k, c)$: rates based on the oob samples. $\operatorname{og}(P\{M_i|y_i\}) = -\frac{\partial}{\partial \mathbf{R}} \operatorname{log}(P\{y_i\})$

$$+ \sum_{j} \frac{\partial}{\partial \beta} log[\pi(\beta; y_{i}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;)] \end{split}$$

$$\frac{\partial}{\partial\beta}{\rm log}\,P\{y_i\}|_{\beta=0}=\sum_j[1]$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

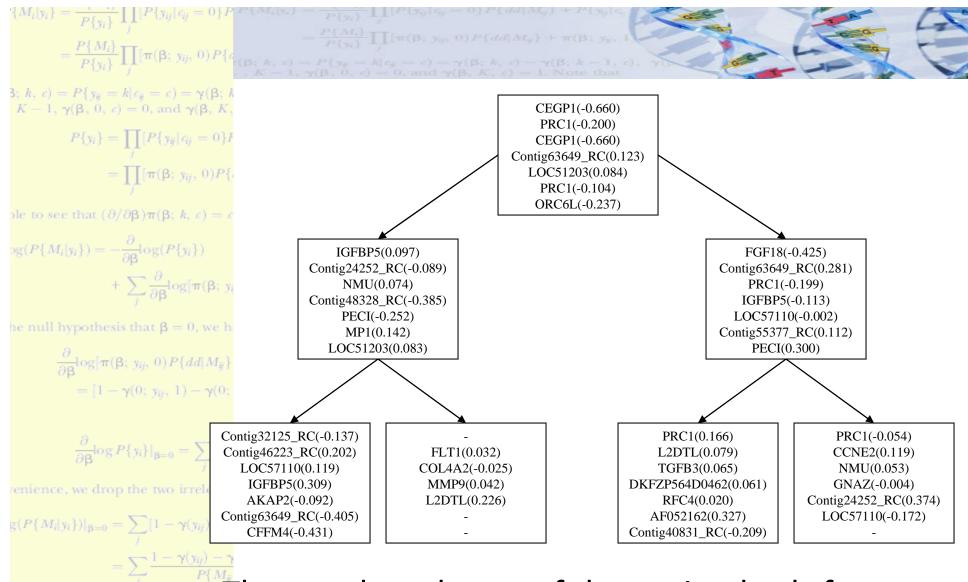
= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

- 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	True	Good	Poor
	at $(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$ $) = -\frac{\partial}{\partial \beta}\log(P\{y_i\})$		Predicted		
	Random Forest	26.0%	Good	141	17
he null hyp $\frac{\partial}{\partial \beta}$	othesis that $\beta = 0$, we h $ \log[\pi(\beta; y_{ij}, 0)P\{dd M_{ij}\} $		Poor	53	58
	Sub-forest	26.0%	Good	146	22
enience, w	$\frac{\gamma}{\beta} \log P\{y_i\} _{\beta=0} = \sum_{j} [1 - \frac{\gamma}{\beta}]$ e drop the two irrelevan		Poor	48	53
$g(P\{M_i y_i\}$	70-gene Classifier	35.3%	Good	103	4
	$= \sum_{j} \frac{P\{M_{ij}\}}{P\{M_{ij}\}}$ it of linkage disequilibries		Poor	91	71



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

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

What Did We Learn?



3;
$$k, e$$
) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, P\{y_i\} = \prod [P\{y_{ij} | e_{ij} = 0\} P\{y_i\}] = 0\}$

$$P\{y_i\} = \prod_{f} [P\{y_{ij} | q_i = 0\}]$$
$$= \prod_{i} [\pi(\beta; y_{ij}, 0) P\{y_{ij} | q_i = 0\}]$$

ole to see that
$$(\partial/\partial \beta)\pi(\beta; k, c) = \log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}\log(P\{y_i\})$$

he null hypothesis that
$$\beta = 0$$
, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} & \log[\pi(\boldsymbol{\beta}; \, y_{ij}, \, \boldsymbol{\theta}) P\{dd|M_{ij}\} \\ &= [1 - \gamma(\boldsymbol{\theta}; \, y_{ij}, \, \boldsymbol{1}) - \gamma(\boldsymbol{\theta}; \, \boldsymbol{\theta})] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 -$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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.

enience, we drop the two irrelevant The limitation

 future samples and studies are needed to evaluate the performance of the forest-based classifiers.

$$\begin{split} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}P(M_{i}|y_{i}) - \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}P\{dd(M_{ij}) + P\{y_{ij}|c_{ij} = 0\}P\{dd(M_{ij}) + P\{y_{ij}|c_{ij} = 0\}P\{dd(M_{ij}) + \pi(\beta; y_{ij}, 0)P\{dd(M_{ij}) + \pi(\beta; y_{ij}, 0)P\{dd(M_{ij}, 0)P\{dd(M_{ij}, 0)P\{dd(M_{ij}, 0)P\{dd(M_{ij}, 0)P\{dd(M_{ij}, 0)P\{dd(M_{i$$

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, i)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} I]$$

= $\prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{ij} | c_{ij} = 0\} I]$

ole to see that (∂/∂β)π(β; k, ε) = ε

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$

Interpretation from Forest



$$\{y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}| c_{ij} = 0\} P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0) P\{c_{ij}\}] P\{c_{ij}\}$

Variable Importance

3; k, e) = $P\{y_{ij} = k | e_{ij} = e\}$ = $\gamma(\beta)$; i K = 1, $\gamma(\beta)$, 0, e) = 0, and $\gamma(\beta)$, K. $P\{y_i\} = \prod [P\{y_{ij}|c_{ij} = 0\}I$

ole to see that
$$(\partial/\partial\beta)\pi(\beta; h, e) = e$$

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial\beta}\log(P\{y_i\})$$

$$+ \sum_j \frac{\partial}{\partial\beta}\log[\pi(\beta; j)]$$

he null hypothesis that
$$\beta = 0$$
, we have
$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd] M_{ij}\}$$

$$= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)]$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1$$

$$\xi(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - 1 - \gamma(y_{ij}) - 1]$$

$$=\sum_{j}\frac{1-\gamma(y_{ij})-\gamma(y_{ij})-\gamma(y_{ij})}{P\{M_{ij}\}}$$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

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 entence, we drop the two irrelevant 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.

$$\{M_i|y_i\} = \frac{e^{-i\beta}}{P\{y_i\}} \prod_j [P\{y_{ij}|e_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{e^{-i\beta}\}]$

Permutation Importance



3;
$$k$$
, c) = $P\{y_{ij} = k | e_{ij} = c\} = \gamma(\beta; k - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, P\{y_i\} = \prod_j [P\{y_{ij} | e_{ij} = 0\} F = \prod_j [\pi(\beta; y_{ij}, 0) P\{x_i\}] = \prod_j [\pi(\beta; y_{ij}, 0) P\{x_i\}]$

- Not necessarily positive
- $\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}\log(P\{y_i\})$ Unbounded
- The magnitudes and relative rankings can be be null hypothesis that β = 0, we under 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 entence, we drop the two trees and the number, q, of variables that are randomly selected to split a node

$$=\sum_{i}\frac{1-\gamma(y_{ij})-\gamma}{P\{M_{ij}}$$

$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\}P\}$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Permutation Importance

3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{ij} | c_{ij} = 0\} F = 0\} F$$

ole to see that $(\partial/\partial \beta)\pi(\beta; k, c) = c$

he null hypothesis that $\beta = 0$, we h

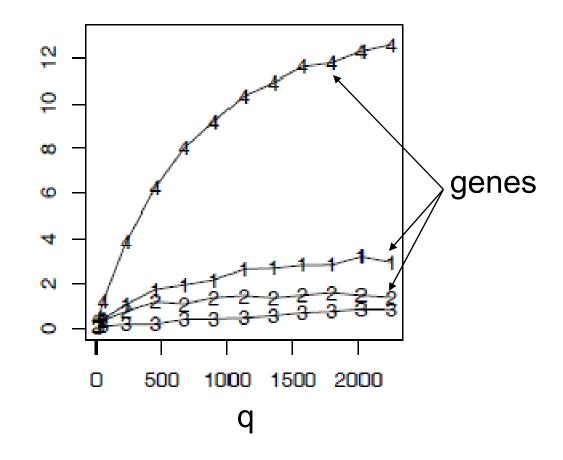
$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$



Maximal conditional chi-square

3;
$$k$$
, e) = $P\{y_{\theta} = k | e_{\theta} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e) = 0, \text{ and$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\}] P\{y_{ij} | a_{ij} = 0\} P\{y$$

ole to see that
$$(\partial/\partial \beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd] M_{ij}\}$$

$$= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)]$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

(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 renience, we drop the two irrelevan predictor

$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j [P\{y_{ij}|c_{ij} = 0\}P]$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{c_{ij}\}]$

Depth Importance

3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$ $P\{y_i\} = \prod [P\{y_{ij}|c_{ij}=0\}F$

ole to see that $(\partial/\partial \beta)\pi(\beta; k, c) = c$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} &\frac{\partial}{\partial \boldsymbol{\beta}} log[\boldsymbol{\pi}(\boldsymbol{\beta};\; \mathbf{y}_{ij},\; \boldsymbol{\theta}) P\{dd|M_{ij}\} \\ &= [1 - \boldsymbol{\gamma}(\boldsymbol{\theta};\; \mathbf{y}_{ij},\; \boldsymbol{1}) - \boldsymbol{\gamma}(\boldsymbol{\theta};\; \boldsymbol{\eta})] \end{split}$$

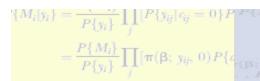
$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

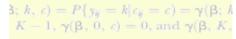
= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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 chisquare 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



$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\}] P\{y_{ij} | a_{ij} = 0\} P\{y$$

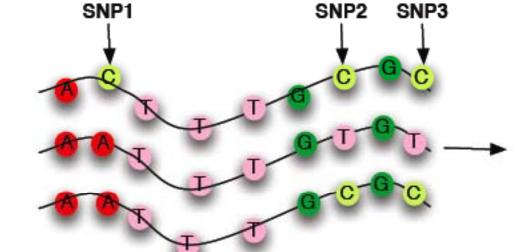
ole to see th

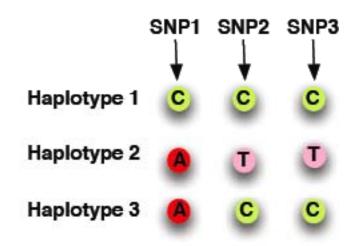
 $og(P\{M_i|y_i\})$

he null hypo

 $\frac{\partial}{\partial \beta}$ log
=

enience, we





$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_j \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA - P\{dd, AA\} - P\{AA\}[P\{DE]$$

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|c_{ij} = 0\}P \underbrace{P\{M_{i}|y_{i}\}}_{P} &= \frac{P\{M_{i}\}}{P} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{a_{i}, k, c) = P\{A_{i}, k, c\}]_{P} \end{aligned}$$

Haplotype Certainty

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$\begin{split} P\{y_i\} &= \prod_{j} [P\{y_{ij} | a_{ij} = 0\} P \\ &= \prod_{j} [\pi(\beta; y_{ij}, 0) P \{ e^{-i\beta_{ij}} \} P \} \end{split}$$

ole to see that $(\partial/\partial\beta)\pi(\beta; h, \epsilon)$ **SNPS**

$$\log(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta}\log(P\{y_i\}) + \sum_i \frac{\partial}{\partial \beta}\log[\pi(\beta; y_i)]$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

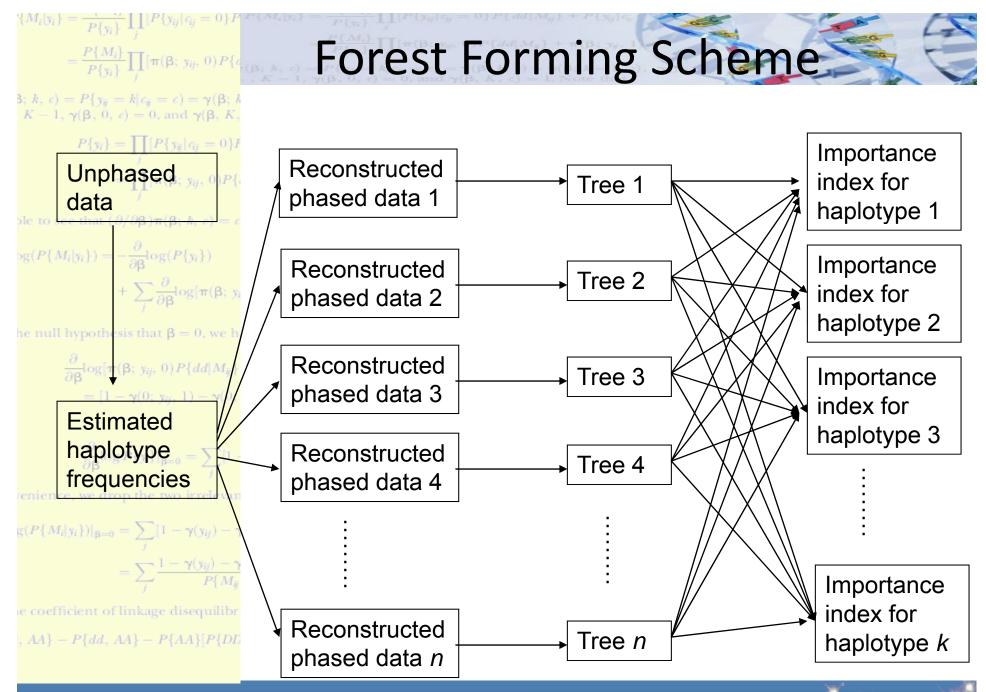
$$AA = P\{dd, AA\} - P\{AA\}[P\{DD]$$

✓ Directly observed

- **✓** No uncertainty
- **X** Less informative
- *****Tree approaches

Haplotypes

- **✗** Inferred from SNPs
- **X** Uncertain
- **✓** More informative
- Forest approaches



$$\{M_i|y_i\} = \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|e_{ij} = 0\}P$$

= $\frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0)P\{e_{ij}\}]$

Inference from the Forest



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\}] P\{y_{ij} | a_{ij} = 0\} P\{y$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, c) = c$$

$$\begin{aligned}
\log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\
&+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)]
\end{aligned}$$

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_j [1 - \frac{1}{\beta}]$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Importance of haplotype h in tree T

be null hypothesis that
$$\beta=0$$
, we have $V_h=\sum_{t\in T,t}2^{-L_t}\chi_t^2$,

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

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} [P\{y_{ij}|c_{ij} = 0\}PP\{M_{i}|y_{i}\} = \frac{P}{I} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{c_{i}, k, c) = P\} \end{aligned}$$

Significance Level

3;
$$k, c) = P\{y_{ij} = k | c_{ij} = c\} = \gamma(\beta; k | K - 1, \gamma(\beta, 0, c) = 0, \text{ and } \gamma(\beta, K, 0) = 0 \}$$

$$P\{y_{i}\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] P\{y_{ij}\} = 0$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta;\,k,\,\epsilon)=\epsilon$$

$$\begin{aligned} \log(P\{M_i|y_i\}) &= -\frac{\partial}{\partial \beta} \log(P\{y_i\}) \\ &+ \sum_j \frac{\partial}{\partial \beta} \log[\pi(\beta; y_i)] \end{aligned}$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; 0)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 -$$

enience, we drop the two irreleval

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

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.

$$\begin{aligned} \{M_i|y_i\} &= \frac{G}{P\{y_i\}} \prod_j |P\{y_{ij}| \, c_{ij} = 0\} P \\ &= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta; y_{ij}, 0) P\{c_{ij}\}] \end{aligned}$$

Simulation Studies (2 loci)

- 3; k, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$
 - $P\{y_i\} = \prod [P\{y_{ij}|c_{ij}=0\}F$

$$g(P\{M_i|y_i\}) = -\frac{\partial}{\partial \beta} log(P\{y_i\}) + \sum_i \frac{\partial}{\partial \beta} log[\pi(\beta)]$$

- - Blog[π(β: y₁₅ 0)P[dd]M_g 3 scenarios $= [1 - \gamma(0; y_B, 1) - \gamma(0;$

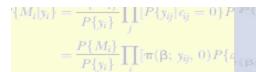
$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_{j} [1 - \gamma(y_{ij}) - \gamma]$$

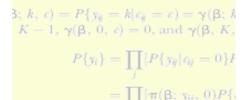
= $\sum_{i} \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

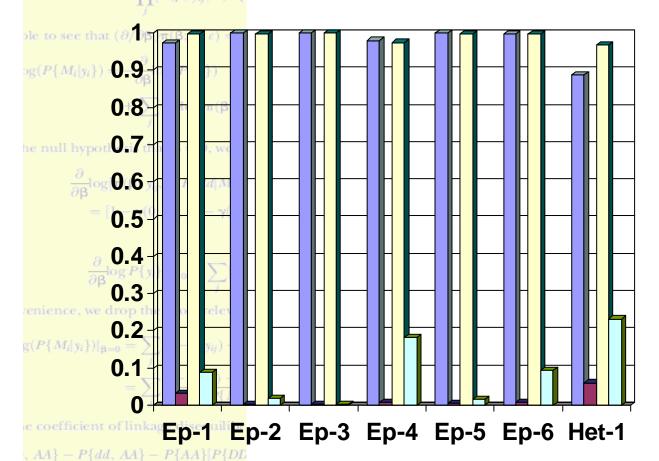
$$AA$$
} - P { dd , AA } - P { AA }[P { DL

- ble to see that (θ/θβ)π(β; λ, ε) = P Each region has 3 SNPs
- 12 interaction models from Knapp et. al. (1994) $+\sum_{0}^{0}$ and Becker et. al. (2005)
- be null hypothesis that β = 0, we let 2 additive models with background penetrance
 - - 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

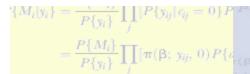


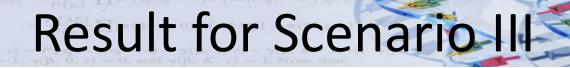
Result for Scenario H

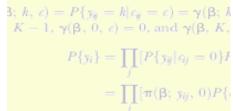


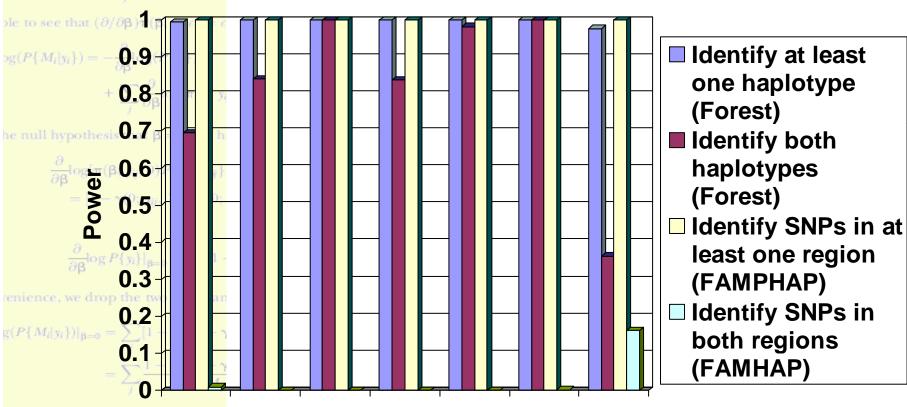


- Identify the correct haplotype (Forest)
- Identify an incorrect haplotype (Forest)
- Identify SNPs in the correct region (FAMHAP)
- Identify SNPs in the neutral region (FAMHAP)





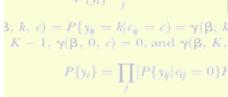




Ep-2 Ep-3 Ep-4 Ep-5 Ep-6 Het-1 $AA = P\{dd, AA\} - P\{AA\}[P\}DE$

$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{e^{-i\beta}}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|e_{ij} = 0\} P^{i} \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{a_{i}\}] \end{aligned}$$

-P{M} Π(π(β. γμ. 0)P{c} Real Case Study



ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$

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



$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}|\epsilon_{ij} = 0\}P \underbrace{P\{M_{i}|y_{i}\}}_{P\{y_{i}\}} &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0)P\{a_{i}, k, e) - P\{y_{i}\}]_{P\{y_{i}\}} \end{aligned}$$

Analysis Procedure

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, e)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | a_{ij} = 0\} P\{y_{ij} = 0\} P\{y_{ij} | a_{ij} = 0\} P\{y_{ij} | a_$$

ole to see that
$$(\partial/\partial \beta)\pi(\beta; k, c) = c$$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 + \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevar

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$|AA| = P\{dd, |AA| = P\{AA\}[P\{DD]$$

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



$$\begin{aligned} \{M_i|y_i\} &= \frac{1}{P\{y_i\}} \prod_j |P\{y_{ij}|c_{ij} = 0\} P P\{M_i|y_i\} = \frac{P\{M_i\}}{P\{y_i\}} \prod_j |P(y_{ij}|c_{ij}) = \frac{P\{M_i\}}{P\{y_i\}} \prod_j |\pi(\mathbf{\beta}; y_{ij}, 0) P\{c_{ij}, k, c\} = P\{y_{ij} = k|c_{ij} = c\} = 0 \end{aligned}$$

Result



3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\}] P\{y_{ij} | c_{ij} = 0\} P\{y$$

ble to see that
$$(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} &\log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevar

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$|AA| = P\{dd, |AA| = P\{AA\}[P\{DL]]$$

Two haplotypes are identified

Most significant: ACTCCG in region 1

$$(p-value = 2e-6)$$

Identical to Klein et. al. (2005)

Located in CFH gene

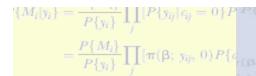
Another significant haplotype:

TCTGGACGACA, in region 2 (p-value = 0.0024)

Not reported before

Protective

Located in BBS9 gene

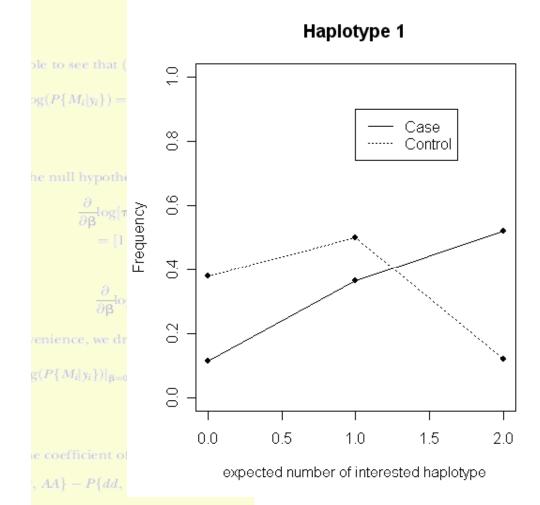


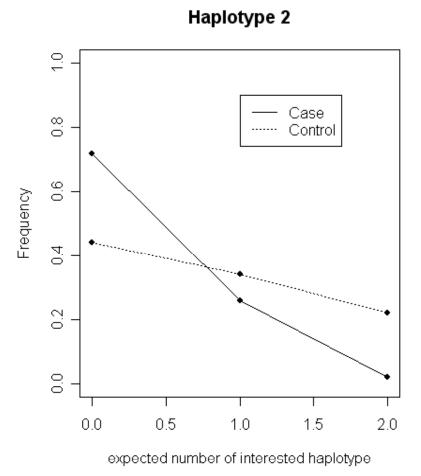
Expected Frequencies



3;
$$k, e$$
) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k | K - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e)$

 $P\{y_i\} = \prod [P\{y_{ij}|c_{ij} = 0\}P$







$$\{M_i|y_i\} = \frac{1}{P\{y_i\}}\prod_j |P\{y_{ij}|c_{ij}=0\}P\{M_i|x_i\} - \frac{P\{M_i\}}{P\{y_i\}}\prod_j |\pi(\beta;y_{ij},0)P\{c_{i\beta;k_i,0}-P\{y_i-k_{i\beta}-0\}-\gamma(\beta;k_i)\} + \frac{P\{M_i\}}{P\{y_i\}}\prod_j |\pi(\beta;y_{i\beta}-0)P\{c_{i\beta;k_i,0}-P\{y_i-k_{i\beta}-0\}-\gamma(\beta;k_i)\} + \frac{P\{M_i\}}{P\{y_i\}}\prod_j |\pi(\beta;y_i-\beta;k_$$

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e$) = $\gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, e)$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} F$$

= $\prod_{j} [\pi(\beta; y_{ij}, 0) P\{$.

ole to see that $(\partial/\partial \beta)\pi(\beta; k, c) = c$

he null hypothesis that $\beta = 0$, we have

$$\begin{split} \frac{\partial}{\partial \boldsymbol{\beta}} \log[\pi(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] \frac{\partial}{\partial \beta} \log P\{y_i\}|_$$

enience, we drop the two irrelevan

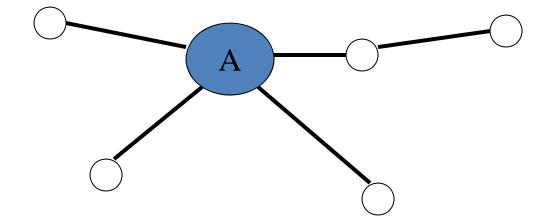
$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DE$$







$$\begin{aligned} \{M_{i}|y_{i}\} &= \frac{1}{P\{y_{i}\}} \prod_{j} |P\{y_{ij}| c_{ij} = 0\} P \\ &= \frac{P\{M_{i}\}}{P\{y_{i}\}} \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{ij}\}] \end{aligned}$$

Acknowledgement

3;
$$k$$
, e) = $P\{y_{ij} = k | e_{ij} = e\} = \gamma(\beta; k - 1, \gamma(\beta, 0, e) = 0, \text{ and } \gamma(\beta, K, K, E)\}$

$$P\{y_i\} = \prod_{j} [P\{y_{ij} | c_{ij} = 0\} I = \prod_{j} [\pi(\beta; y_{ij}, 0) P\{c_{ij} | c_{ij} = 0\} I = 0\} I$$

ble to see that $(\partial/\partial\beta)\pi(\beta; k, c) = c$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} & \frac{\partial}{\partial \boldsymbol{\beta}} \text{log}[\boldsymbol{\pi}(\boldsymbol{\beta}; y_{ij}, 0) P\{dd|M_{ij}\} \\ & = [1 - \gamma(0; y_{ij}, 1) - \gamma(0; \\ \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{1}{2}] = \sum_i [1 - \frac{1}{2}$$

enience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$AA - P\{dd, AA\} - P\{AA\}[P\{DD]$$

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

The Collaborative Center for Statistics in Science



$$P\{M_i|y_i\} = \frac{P\{M_i\}}{P\{y_i\}} \prod_j [P\{y_j|x_j = 0\} P\{dd\{M_j\}\} + P\{y_j|x_j\}]$$

$$= \frac{P\{M_i\}}{P\{y_i\}} \prod_j [\pi(\beta;y_{ij}|0) P\{dd\{M_j\}\} + \pi(\beta;y_{ij},1)]$$

$$\epsilon(\beta;h,v) = P\{y_j = h(x_i = v) = \gamma(\beta;h,v) - \gamma(\beta;k-1,v),$$

to see that $(\partial/\partial\beta)\pi(\beta; k, \epsilon) = \epsilon$

he null hypothesis that $\beta = 0$, we h

$$\begin{split} \frac{\partial}{\partial \beta} \log[\pi(\beta; y_{ij}, 0) P\{dd|M_{ij}\} \\ &= [1 - \gamma(0; y_{ij}, 1) - \gamma(0;)] \end{split}$$

$$\frac{\partial}{\partial \beta} \log P\{y_i\}|_{\beta=0} = \sum_i [1 - \frac{\partial}{\partial \beta}] \log P\{y_i\}|_{\beta=0}$$

renience, we drop the two irrelevan

$$g(P\{M_i|y_i\})|_{\beta=0} = \sum_j [1 - \gamma(y_{ij}) - \gamma]$$

= $\sum_i \frac{1 - \gamma(y_{ij}) - \gamma}{P\{M_{ij}\}}$

e coefficient of linkage disequilibr

$$, AA\} = P\{dd, AA\} = P\{AA\}[P\{DD]$$

Thank You!