Genotype-Phenotype Association

CMSC702 Spring 2014

Hector Corrada Bravo
University of Maryland

GWAS

- Genome-wide association studies
- Scans for SNPs (or other structural variants)
- that show *association* with some phenotype
  - categorical phenotypes: e.g., *age-related macular degeneration*
  - continuous phenotypes (QTL): blood pressure
- Commonly: 10^3 samples, 10^6 SNPs

logistic regression

Binary outcome, disease/no disease

Predictors (genotypes)

- Estimate [illegible]o

$$\theta(x) = Pr\{y = 1|x\}$$

$$f(x) = \log \frac{\theta(x)}{1 - \theta(x)}$$

- *f* is linear

logistic regression

$$f(x) = \log \frac{\theta(x)}{1 - \theta(x)} = \beta_0 + \beta_1 x$$

Encoding genotype data

- We usually think of major/minor alleles, where minor allele occurs at a less frequency in the population (e.g., 5%)
- haplotype:
  minor allele: AA, Aa -> x=0; aa -> x=1
  major allele: AA,Aa -> x=1; aa -> x=0
  both:AA->x1=1,x2=1;Aa->x1=1,x2=0,etc...

- genotype (dosage):
  AA -> x=0; Aa -> x=1; aa-> x=2

Interpretation

Odds of outcome for, e.g, genotype AA

$$\frac{P(Y=1|X=0)}{P(Y=0|X=0)} = e^{\beta_0}$$

Odds of outcome for, e.g, genotype Aa

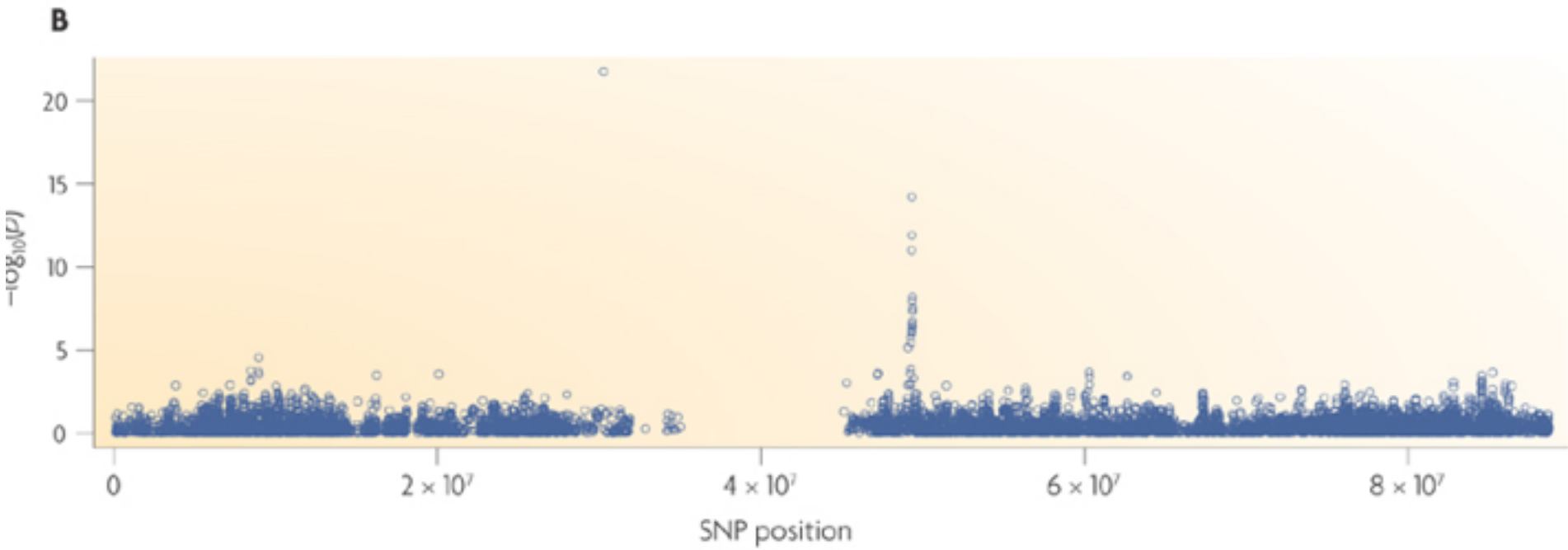
$$\frac{P(Y=1|X=1)}{P(Y=0|X=1)} = e^{\beta_0+\beta_1}$$

Odds-ratio

$$\frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)} = e^{\beta_1}$$

GWAS

Discovering association: how unexpected is this odds ratio?



gwas

Genome.gov | A Catalog of Published Genome-Wide Association Studies

http://www.genome.gov/gwastudies/

nhgri gwas catalog

Bonjour Read Later Strawberry C...acker Crust NextBus:113ToCampus Watch Live ...SPN360.com Hector's Page Academic360... Education DBLP Bibliography MathWorld Developer Doc

Detecting gen... | http://www.n... | Odds ratio – ... | Machine learn... | General Linea... | RAPID detecti... | Article series :... | Genome-wide... | Genome.gov |...

genome.gov
National Human Genome Research Institute
National Institutes of Health

SEARCH

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | For You

Home › About › Organizational Structure › About the Office of the Director › Office of Population Genomics › **A Catalog of Published Genome-Wide Association Studies**

Share Print

Office of Population Genomics

Overview | **A Catalog of Genome-Wide Association Studies** | Research Programs | Publications | Meetings & Workshops | Notices & Funding Opportunities | Staff Biographies | Contact

A Catalog of Published Genome-Wide Association Studies

Potential etiologic and functional implications of genome-wide association loci for human diseases and traits
Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

Keywords: what's this?

- GWAS
- GWAS Catalog
- Office of Population Genomics
- Teri Manolio
- RSS Feeds

Go to the Catalog

- Expensive and pervasive...

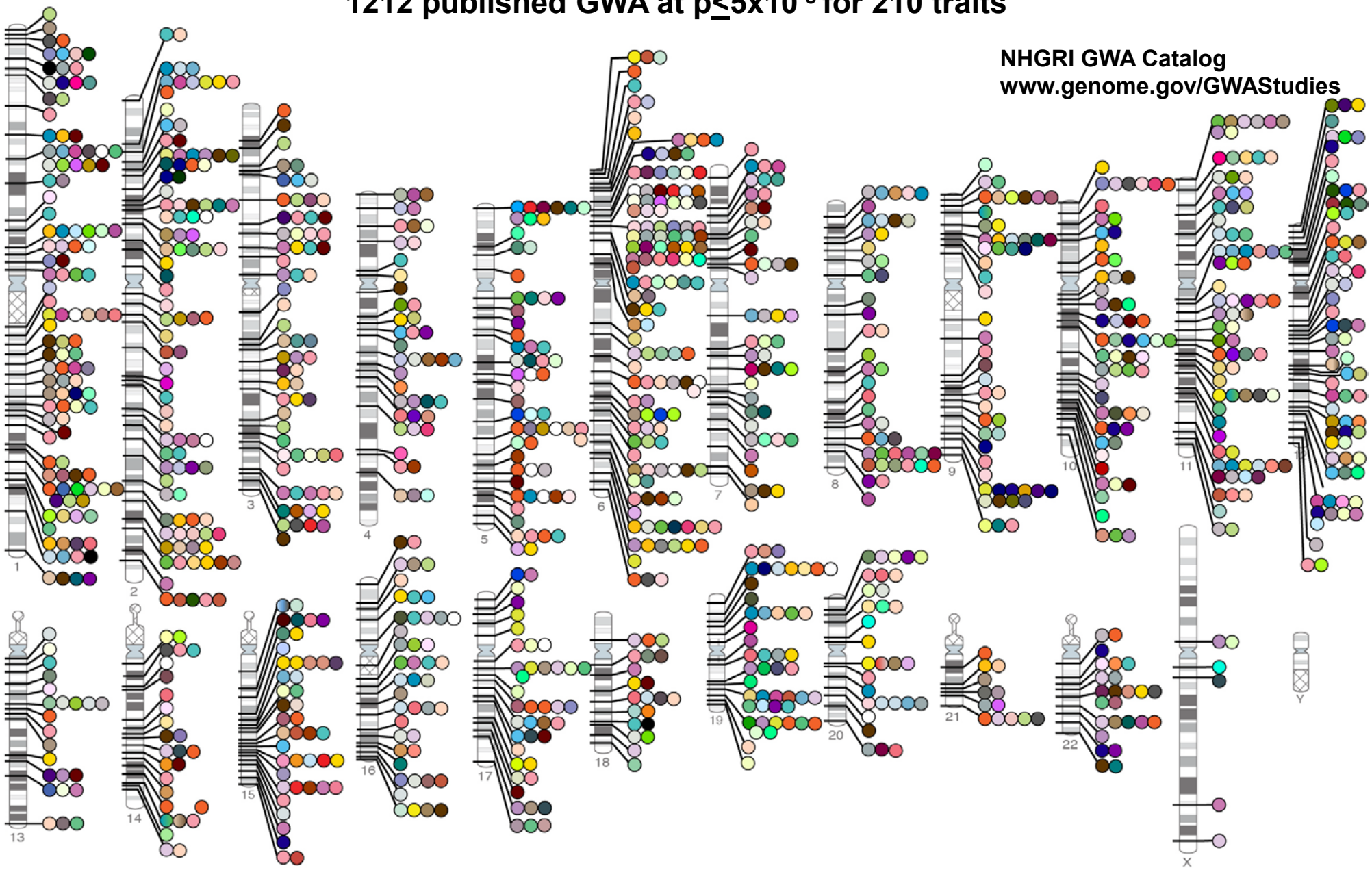
Published Genome-Wide Associations
Credit: Darryl Leja and Teri Manolio

The genome-wide association study (GWAS) publications listed here include only those attempting to assay at least 100,000 single nucleotide polymorphisms (SNPs) in the initial stage. Publications are organized from most to least recent date of publication, indexing from online publication if available. Studies focusing only on candidate genes are excluded from this catalog. Studies are identified through weekly PubMed literature searches, daily NIH-distributed compilations of news and media reports, and occasional comparisons with an existing database of GWAS literature (HuGE Navigator).

SNP-trait associations listed here are limited to those with p-values < 1.0 x 10-5 (see full methods for additional details). Multipliers of powers of 10 in p-values are rounded to the nearest single digit; odds ratios and allele frequencies are rounded to two decimals. Standard errors are converted to 95 percent confidence intervals where applicable. Allele frequencies, p-values, and odds ratios derived from the largest sample size, typically a combined analysis (initial plus replication studies), are recorded below if reported; otherwise statistics from the initial study sample are recorded. For quantitative traits, information on % variance explained, SD increment, or unit difference is reported where available. Odds ratios < 1 in the original paper are converted to OR > 1 for the alternate allele. Where results from multiple genetic models are available, we prioritized effect sizes (OR's or beta-coefficients) as follows: 1) genotypic model, per-allele estimate; 2) genotypic model, heterozygote estimate, 3) allelic model, allelic estimate.

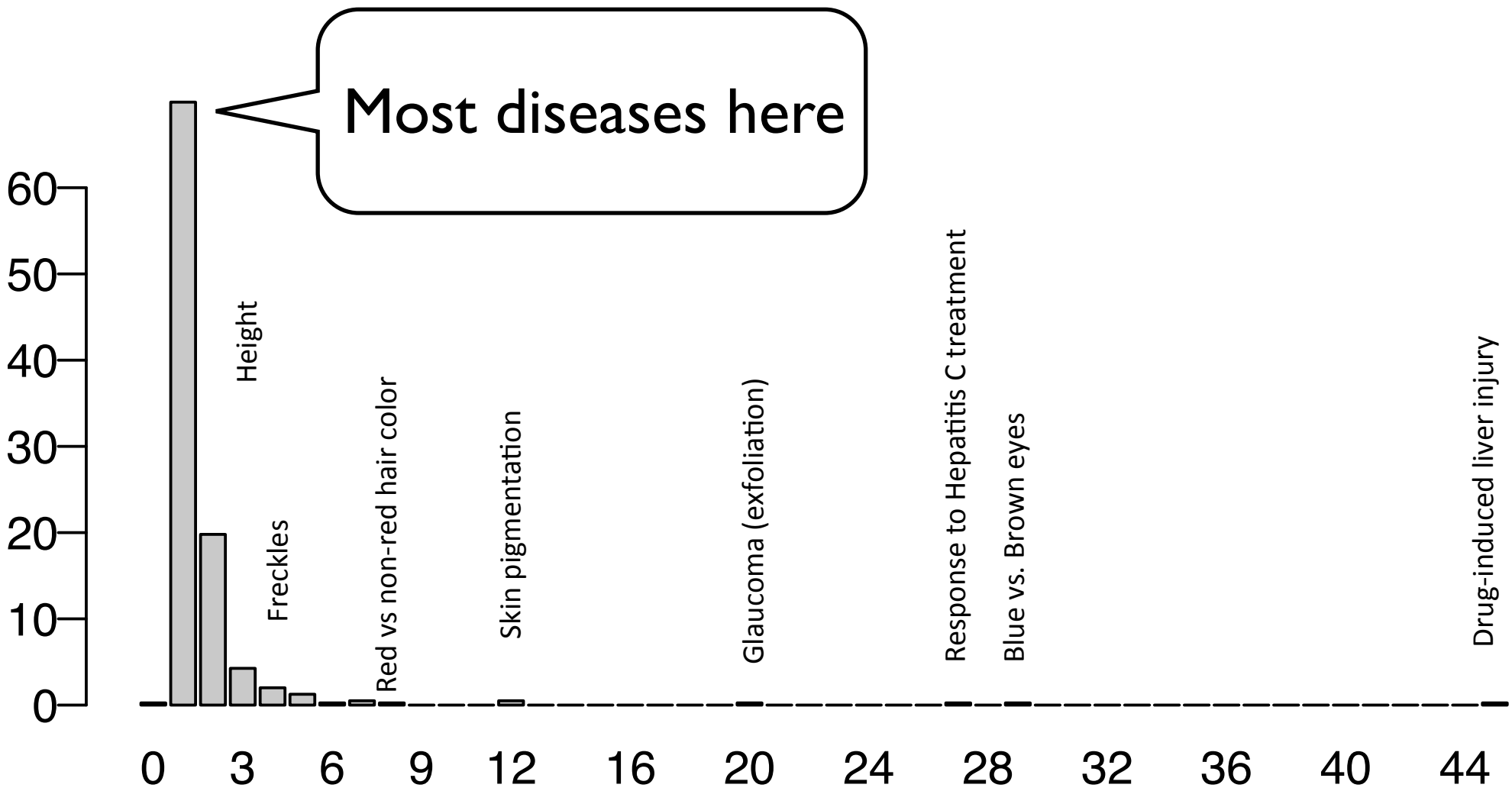
Published Genome-Wide Associations through 12/2010, 1212 published GWA at $p \leq 5 \times 10^{-8}$ for 210 traits

NHGRI GWA Catalog
www.genome.gov/GWAStudies




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

GWAS odds ratios



GWAS

- Testing for *marginal effects* is limited
  - Epistasis, interactions
- Environment/risk factors, unaccounted dependencies
- Not all SNPs are created equal (annotation)

Epistasis

- Testing *marginal effects* is limited
- We want to test interactions (epistasis)
- Modeling is straightforward:
  - add non-linear interaction terms to logistic regression model
- Computationally, it's a problem
  - we started with 10^6 SNPs....

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 22 2010, pages 2856–2862
doi:10.1093/bioinformatics/btq529

Genetics and population analysis

Advance Access publication September 24, 2010

RAPID detection of gene–gene interactions in genome-wide association studies

Dumitru Brinza[1], Matthew Schultz[2], Glenn Tesler[3] and Vineet Bafna[4,*]

[1]Life Technologies, Foster City, CA, [2]Graduate Bioinformatics Program, [3]Department of Mathematics and [4]Department of Computer Science and Engineering, Institute for Genomic Medicine, University of California, San Diego, CA, USA

Associate Editor: Jeffrey Barrett

A filtering approach:

- Discover possible interactions quickly
- Test good candidates completely

Sparsity again

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{j<k} \beta_{jk} x_j x_k$$

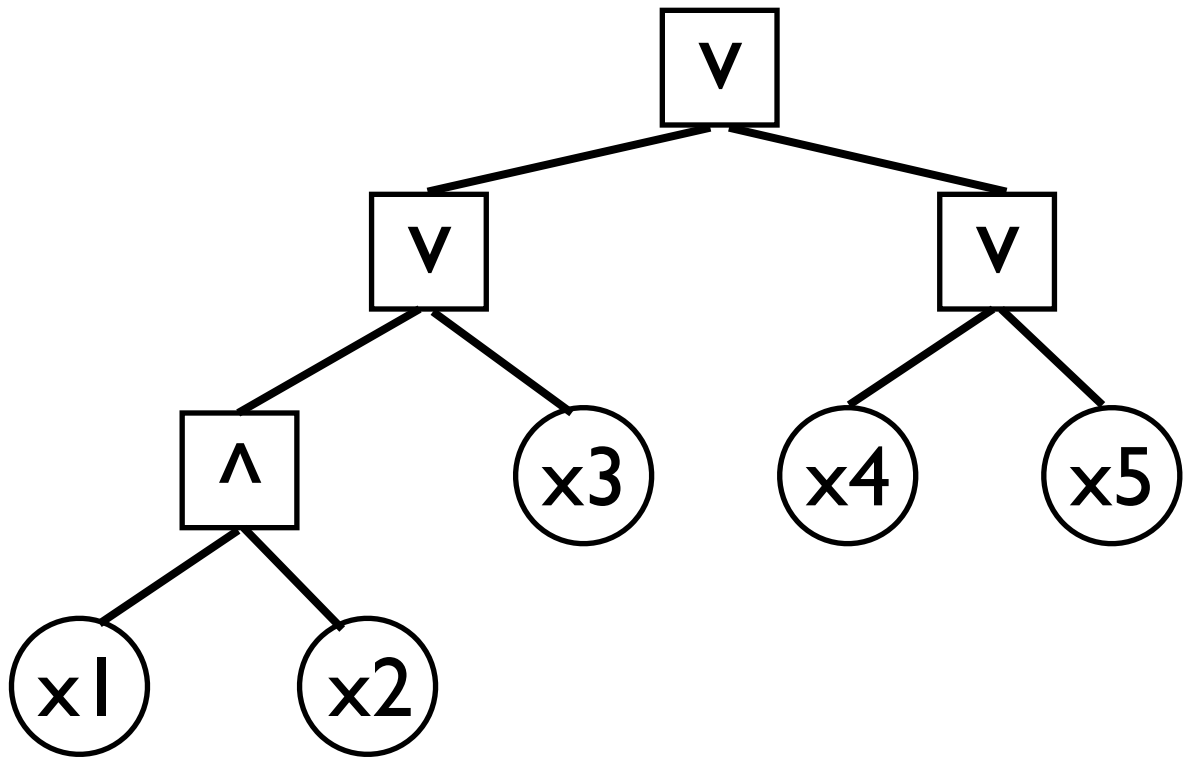
Estimate using penalized likelihood method
using *L1* penalty for sparsity

[Shi, et al., *Statistics and its Interface (SII)* 2008]

Has a *soft rule* interpretation: *soft* disjunction of conjunctions

Logic regression

$$f(x) = \beta_0 + \sum_{l=1}^{t} \beta_l L_l(x)$$



[Ruczinski, et al., *Journal of Computational and Graphical Statistics*, 2003]

RAPID

- If two SNPs (x and y) associate with disease (d) then at least one of the following must hold:
  1. x associates with d
  2. y associates with d
  3. x associates with y in cases
  4. x associates with y in controls
- RAPID finds SNPs where 3 holds

RAPID

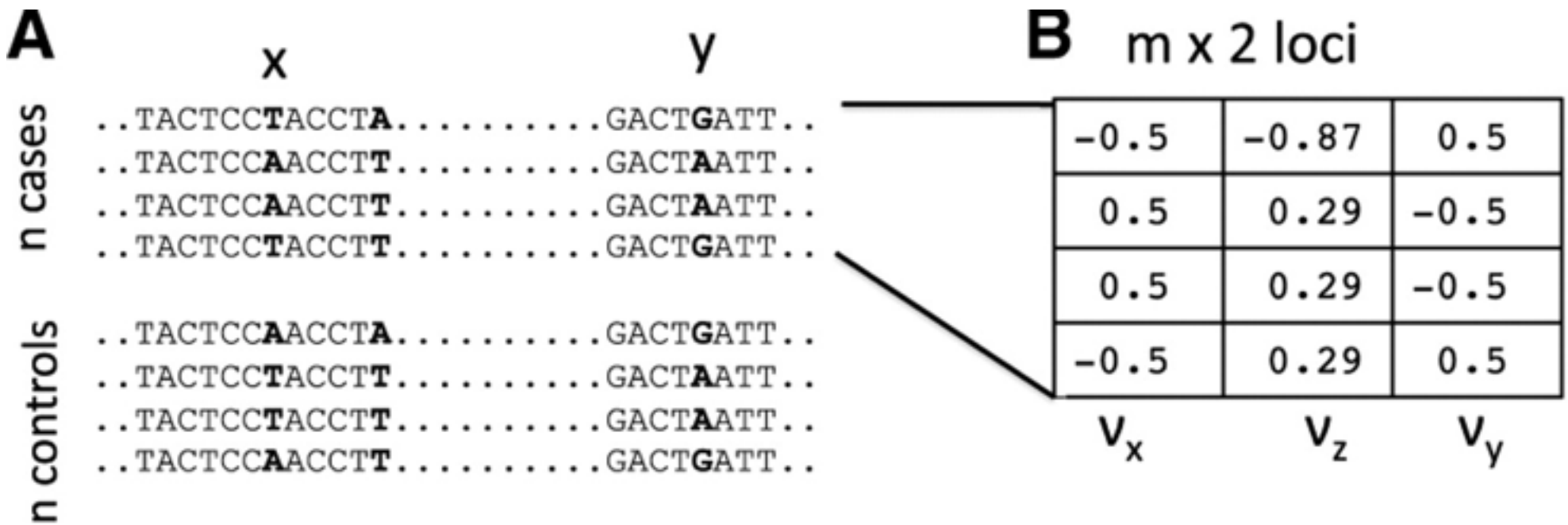
- Look at cases only, and define vector for each SNP as:

$$v_x(a) = \frac{a - P_x}{\sqrt{n}\sqrt{P_x(1 - P_x)}}$$

0, 1

Proportion of 1s

RAPID


A
x
y
n cases
..TACTCCTACCTA.....GACTGATT..
..TACTCCAACCTT.....GACTAATT..
..TACTCCAACCTT.....GACTAATT..
..TACTCCTACCTT.....GACTGATT..
n controls
..TACTCCAACCTA.....GACTGATT..
..TACTCCTACCTT.....GACTAATT..
..TACTCCTACCTT.....GACTAATT..
..TACTCCAACCTT.....GACTGATT..
B
m x 2 loci
-0.5 -0.87 0.5
0.5 0.29 -0.5
0.5 0.29 -0.5
-0.5 0.29 0.5
v_x v_z v_y

RAPID

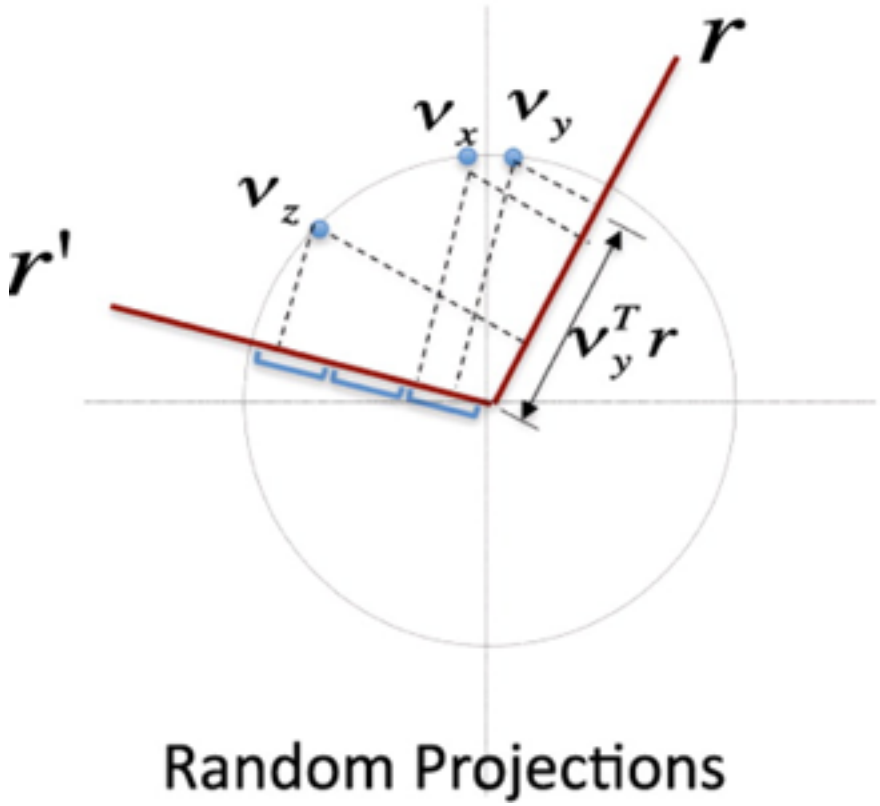
$$dist(v_x, v_y) = \sqrt{2 - 2\sqrt{\chi^2_{x,y}/n}}$$

Association between x and y

Statistical association is now a geometric problem

RAPID

- Use random projections to find possible interacting pairs

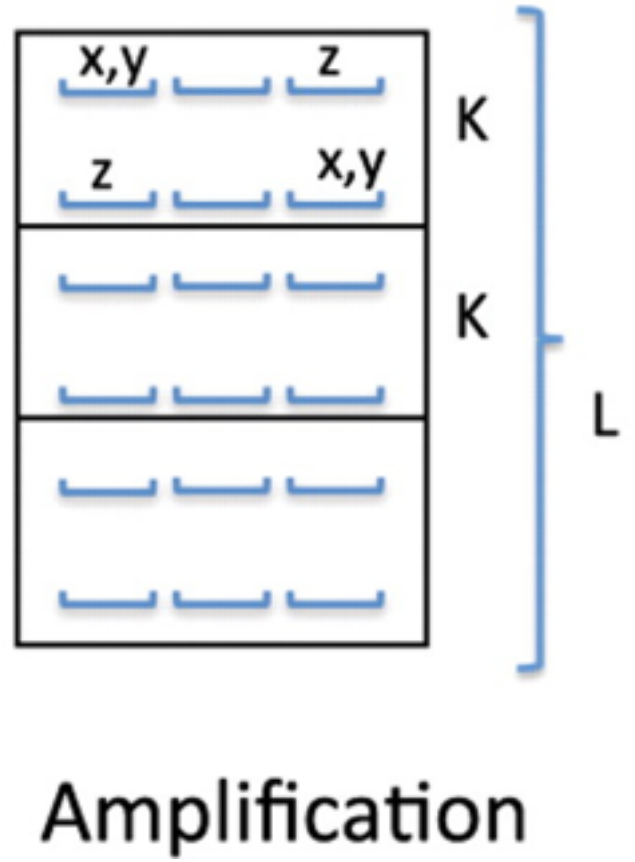


Random Projections

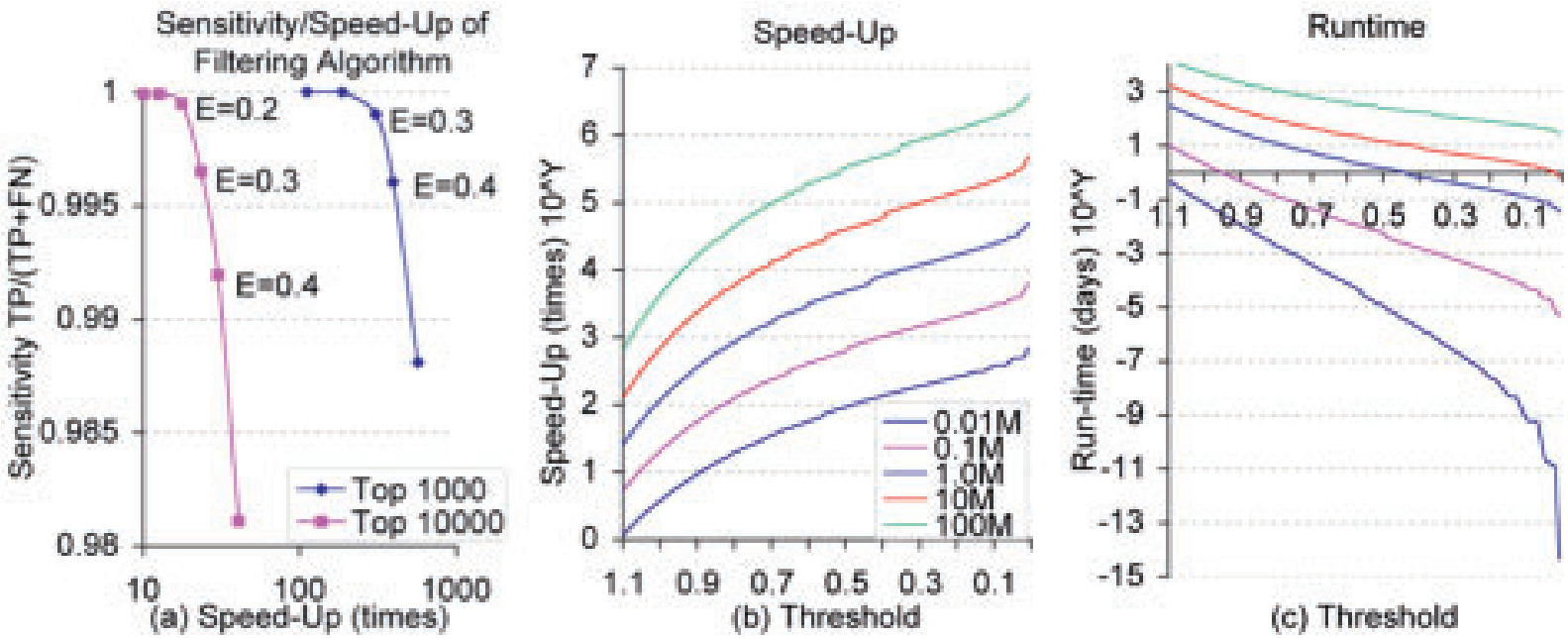
$$\text{Hash}(x, r, B) = \left\lfloor \frac{|v_x \cdot r|}{B} \right\rfloor$$

RAPID

- Do this repeatedly, to avoid false positives



Performance



Interactions/Epistasis

- A MAJOR problem
  - Inherently computational and statistical
  - We are nowhere close
  - We will be inundated with data (sequencing)