Gene expression and protein pathways

Joel S Bader
CuraGen
jsbader@curagen.com
IPAM 2000 Nov 8

CuraGen Background

- Product pipeline
  - Therapeutic proteins
  - Therapeutic antibodies
  - Drug targets
- Technology
  - High-throughput biology labs
  - Bioinformatics/information-intensive

Outline

- Clustering gene expression data
  - Interior-node test
  - Statistical significance
  - Power
- Mapping biological pathways
  - Metabolic pathways
  - mRNA coregulation
  - Protein-protein interactions
  - Overlaying
- Genetic studies
  - Disease risk
  - SNPs and association

Clustering

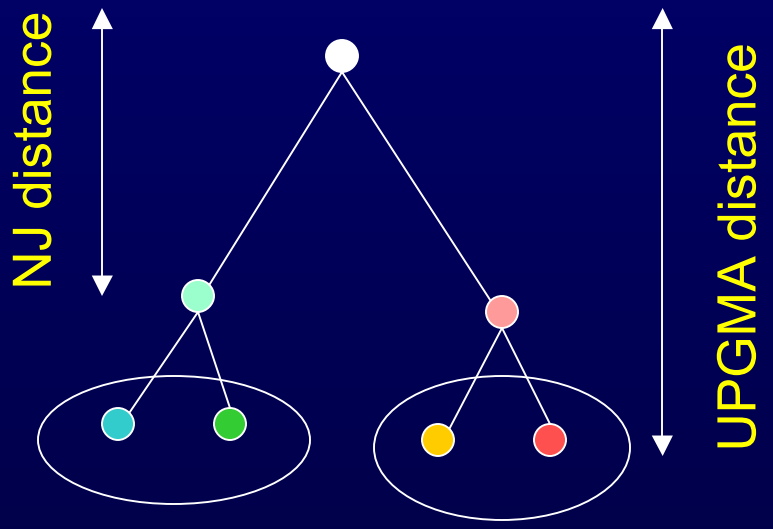
- Standard method (now) for analyzing gene expression data
- Unsupervised algorithms
  - Run to completion
  - Clustering eventually driven by noise, not biology
- Supervised algorithms
  - Inconsistent, irreproducible
  - Not amenable to high-throughput
- Goal: automated, unsupervised, with meaningful p-value for clusters produced
- Collaboration with Rebecca W Doerge and Brian Munneke, Dept of Statistics, Purdue

Hierarchical, distance-based algorithm

- Initialize: each gene in a single cluster
- Repeat
  - Join clusters with shortest distance
  - Re-calculate effective distances
- Until 1 cluster remains

Neighbor-joining: distance is corrected to be distance between ancestors
Studier & Keppler, Mol Biol Evol 5: 729 (1988)

Unweighted pair group method arithmetic mean: distance is mean of all pair-wise distances

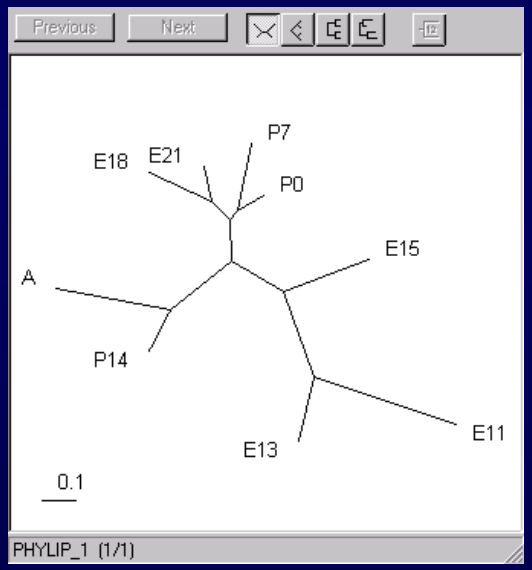


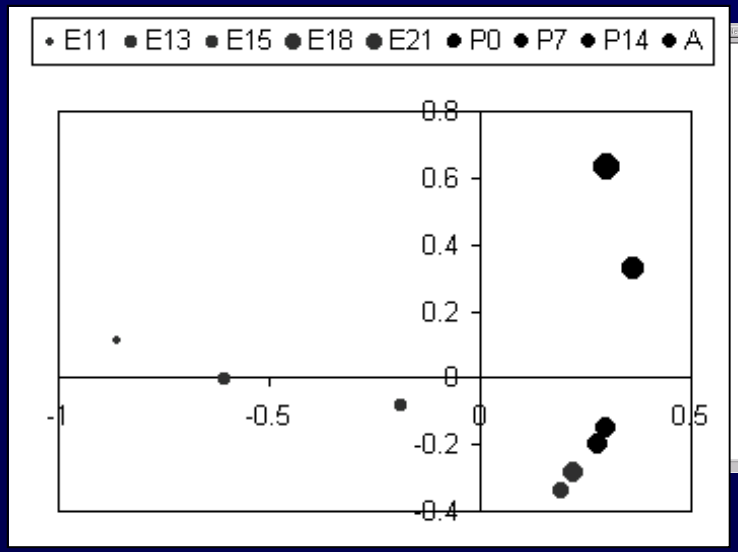
(Typical) results

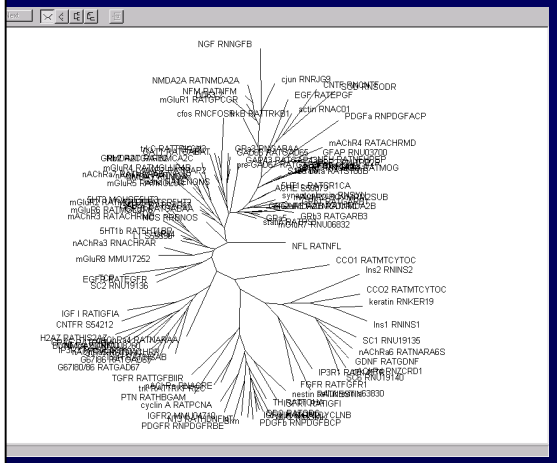
Data taken from X. Wen, …, R Somogyi, PNAS 95: 334 (1998)
Large-scale temporal gene expression mapping of central nervous system development

9 time points, embryonic to adult

112 genes



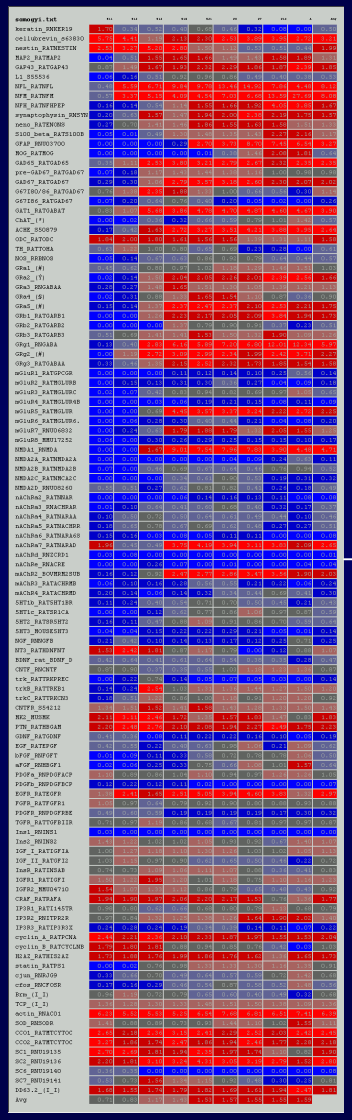


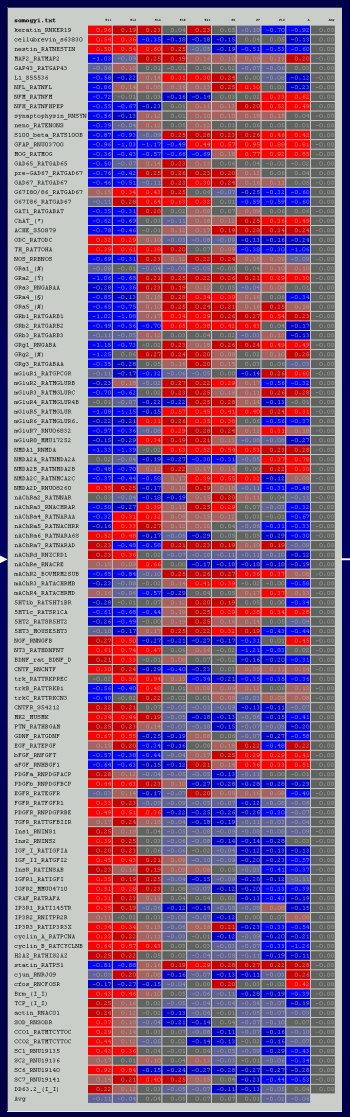


Clustering

Multidimensional scaling, principal component/factor analysis

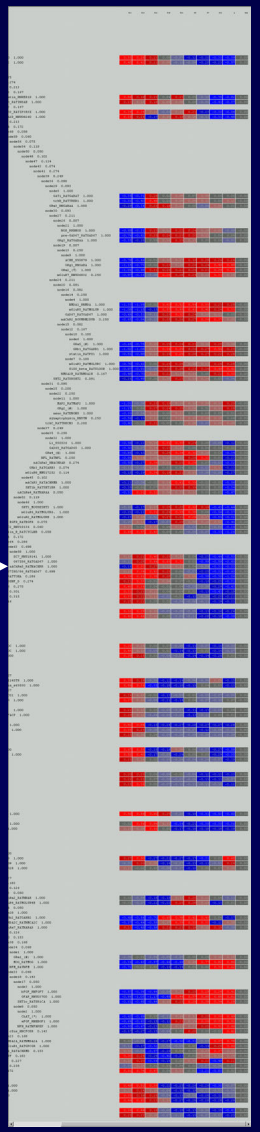
Raw data and clusters





- Set baseline
- Normalize columns (time-points)
- Log-transform
- Subtract row averages (genes)

- Neighbor-joining using correlation distance

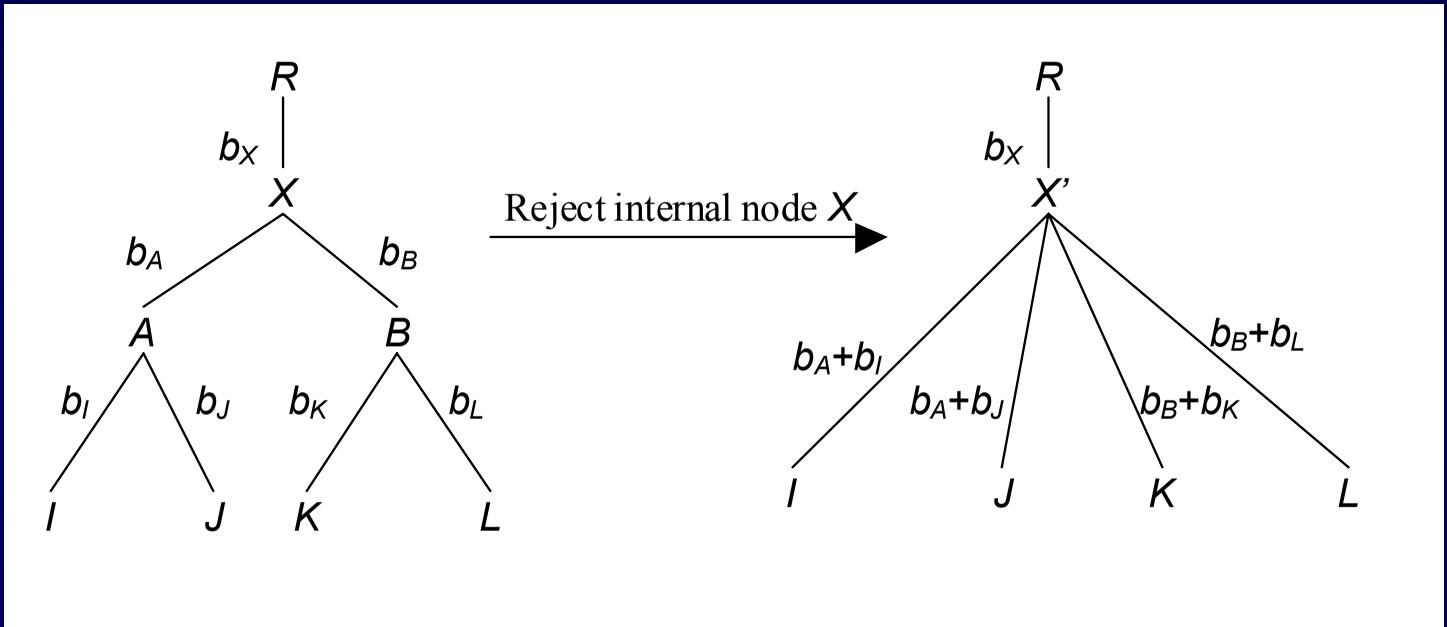


Significance tests

- Interior-branch test
  - Parametric based on branch length and variance estimator
  - Branch length error is not normally distributed
  - Small sample size
  - Global test, draws information from entire tree, doesn't test a node directly
  - Not consistent with neighbor-joining algorithm
- Bootstrap test
  - Resample data matrix by choosing columns at random with replacement
  - Requires underlying data
  - Computationally intensive
  - Also not consistent with neighbor-joining algorithm

Our approach: interior-node test using neighbor-joining statistic and permutation/randomization

Alternative and null hypotheses



Equivalent to interior-branch test for unrooted tree *IJKL*

Test statistic: total length of tree under node *X*

Algorithm

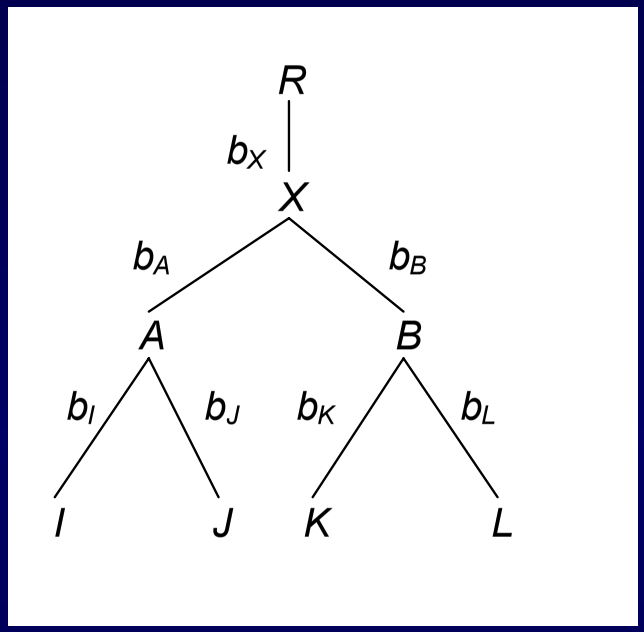
Replace nodes *A* and *B* with star-trees of terminal taxa
Total length = star-tree of *A* + star-tree of *B* + branch length *AB*

$L_A = (N_A - 1)^{-1} T_{AA}$

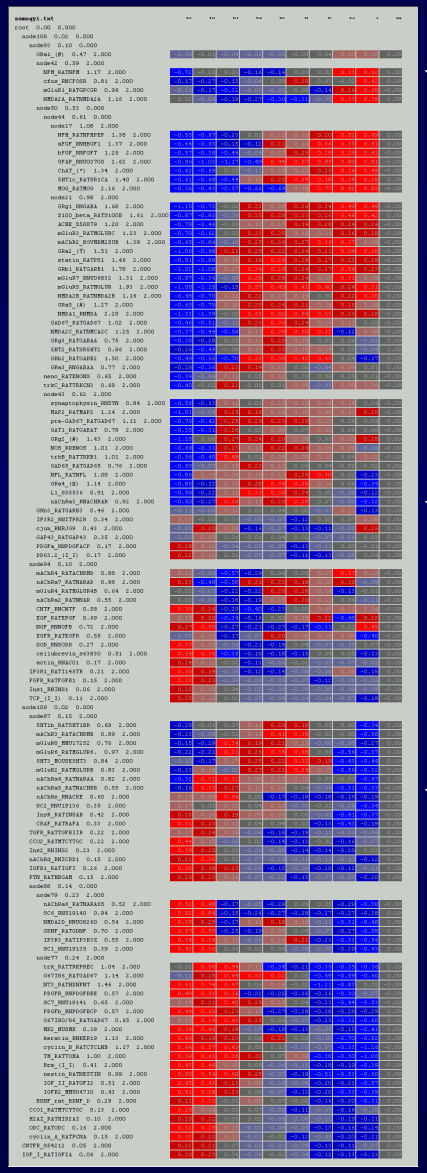
$L_B = (N_A - 1)^{-1} T_{BB}$

$L_{AB} = (N_A N_B)^{-1} T_{AB} - N_A^{-1} L_A - N_B^{-1} L_B$

Calculate p-value from > 1000 re-assignments of terminal taxa with fixed N_A, N_B



Results, p-value = 0.001



Too liberal

Too conservative

Too conservative

Power estimate

Model: Pairwise distance between terminal taxa *m* and *n* is

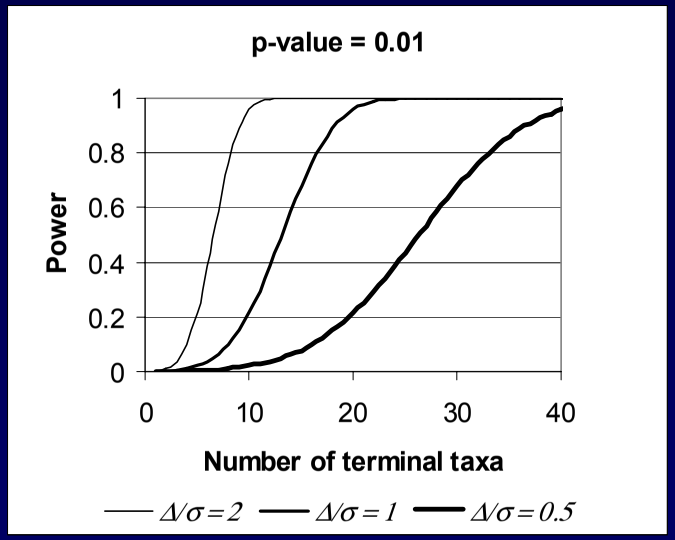
$d_{mn} = d_0 + I_{mn}\Delta d + \varepsilon_{mn}$

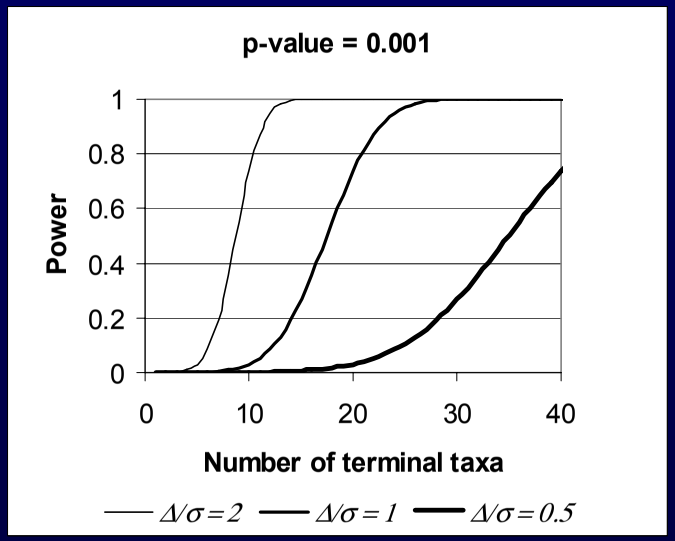
I = 0 or 1 if taxa are in same/different cluster
H = fraction of pairs across clusters
$\mathrm{Var}(\varepsilon) = \sigma^2$

Expectations:

$E(L_{Star}) = Nd_0/2 + NH\Delta d/2$
$E(L_{Alt}) = Nd_0/2 + \Delta d$
$\mathrm{Var}(L_{Alt}) = [\,1 - (N-2)/2N_AN_B\,]\sigma^2 \approx \sigma^2$

$\text{Power} = \Phi[(NH\Delta d/2\sigma) - 2^{-1/2}z_\alpha]$



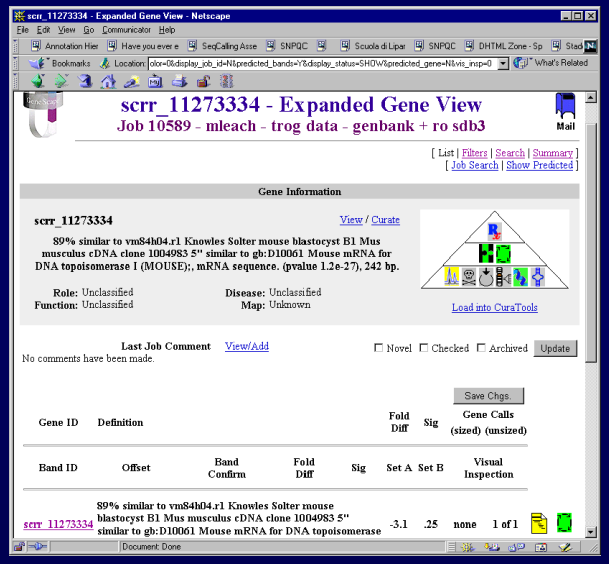


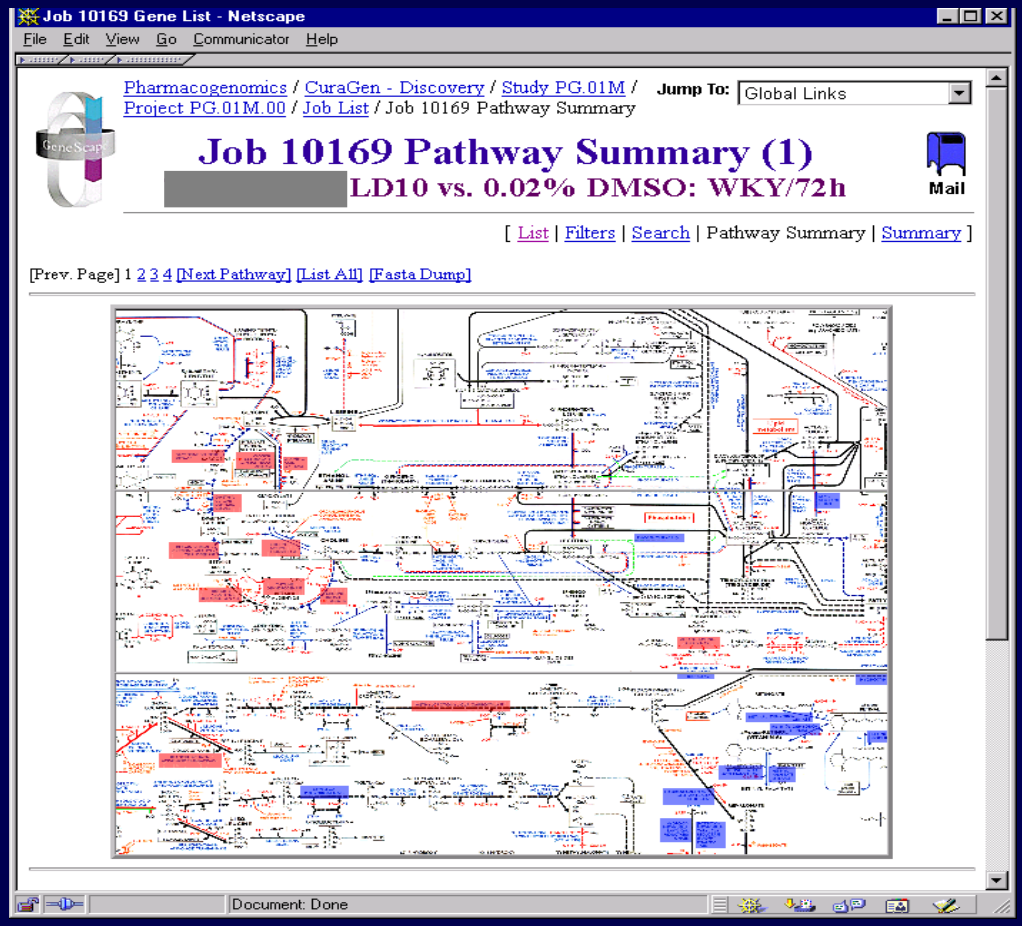
Conclusion

- Interior-node test based on resampling
  - P-value 0.01 to 0.001 gives good results
  - For just-resolved clusters (distance between ~ standard deviation within), need 8-10 terminal taxa per cluster for significance
- Other work
  - Different clustering algorithms, better power
  - 2D extensions
  - PCA, regression, prediction
- Applications
  - Disease-related pathways, target identification/validation
  - Pharmacogenomics: predictive toxicity, efficacy markers (immediately commercializable)
  - Exploration of coregulation pathways

Mapping biological pathways

- Metabolic pathways
- mRNA coregulation





Red: Up Blue: Down

Mapping biological pathways

- Metabolic pathways
- mRNA coregulation
- Protein-protein interactions
  - Yeast two-hybrid system
  - Genome-scale survey of yeast

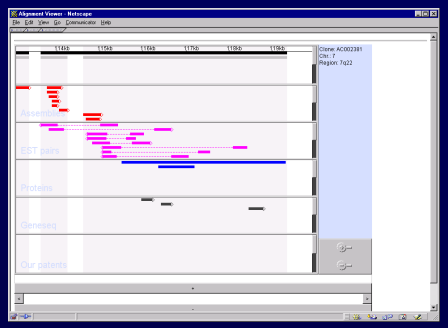
- Overview of PathCalling
- Comparison of pathways from protein-protein interactions and mRNA correlation
- PathCalling bioinformatics: Jim Knight, CuraGen www.curagen.com

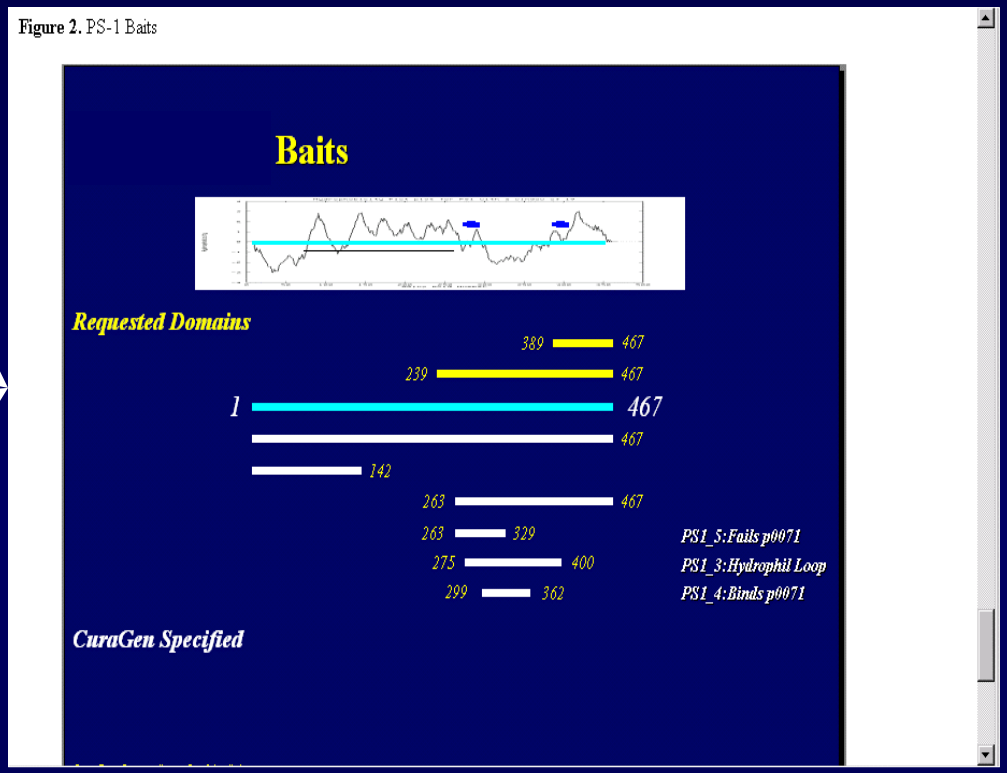
Nature 403: 623 (2000)
Collaboration with Stan Fields

PathCalling process

Candidate genes

Whole-genome ORFS





Bait design

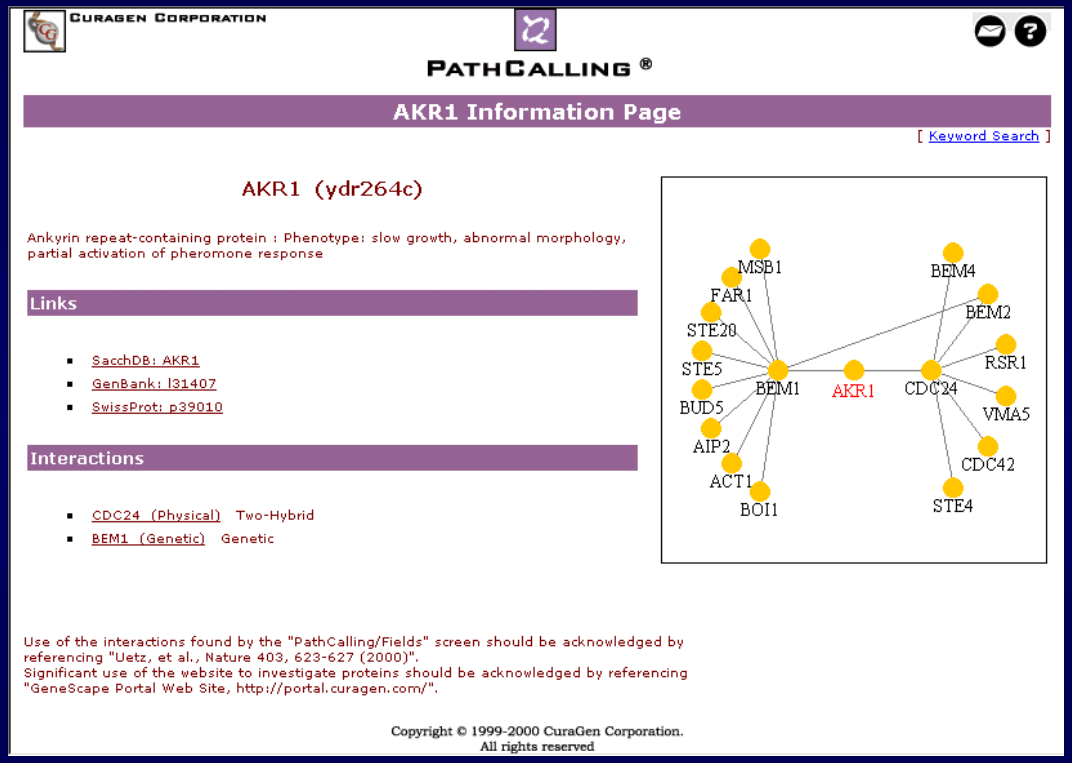
Whole-genome library

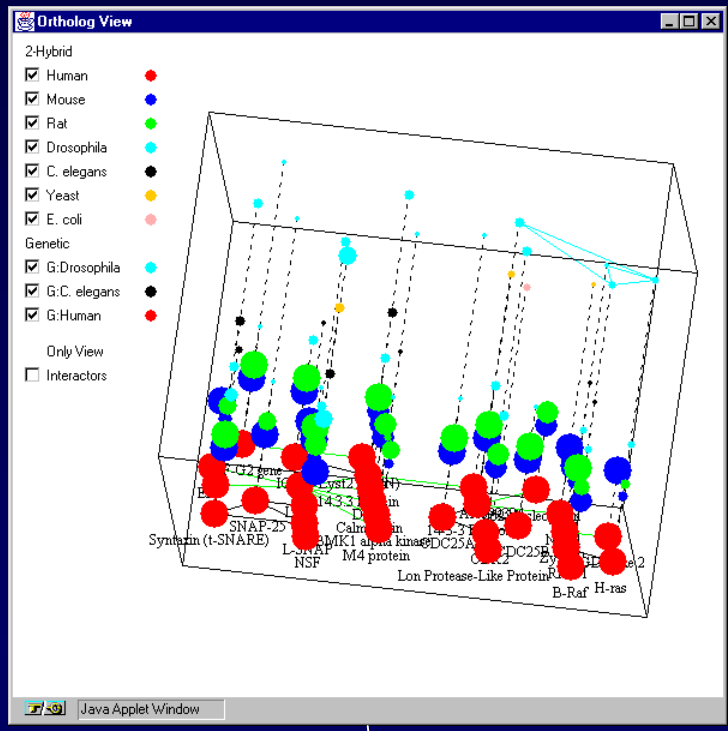
Matings

Interactions (pairwise links)

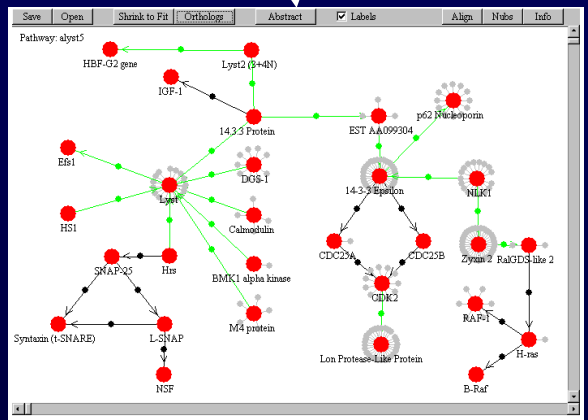
Results

Pairwise interactions
built into pathways





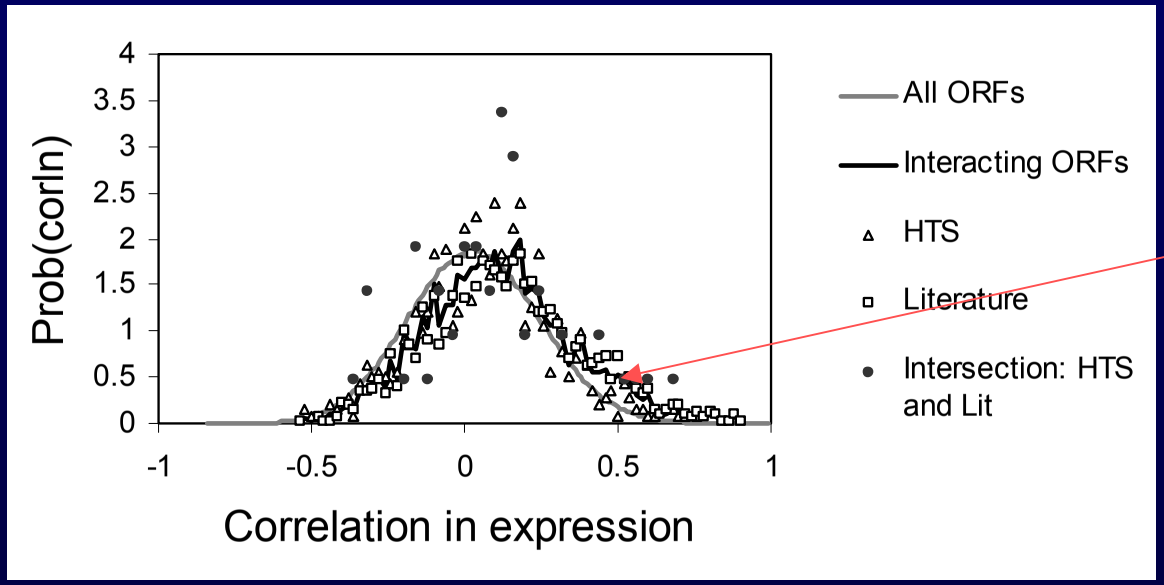
Infer human pathways by homology



Nov 8, 2000

Comparing Interaction and Expression

- Use correlated expression to infer a protein-protein link
- What is the overlap between expression links and interaction links?
  - Yeast expression data from Pat Brown group
  - Yeast interaction data from CuraGen/Fields



Interacting proteins (black line) are slightly more likely to have positive correlation than random ORFs (grey line)

Difference is small

CuraGen

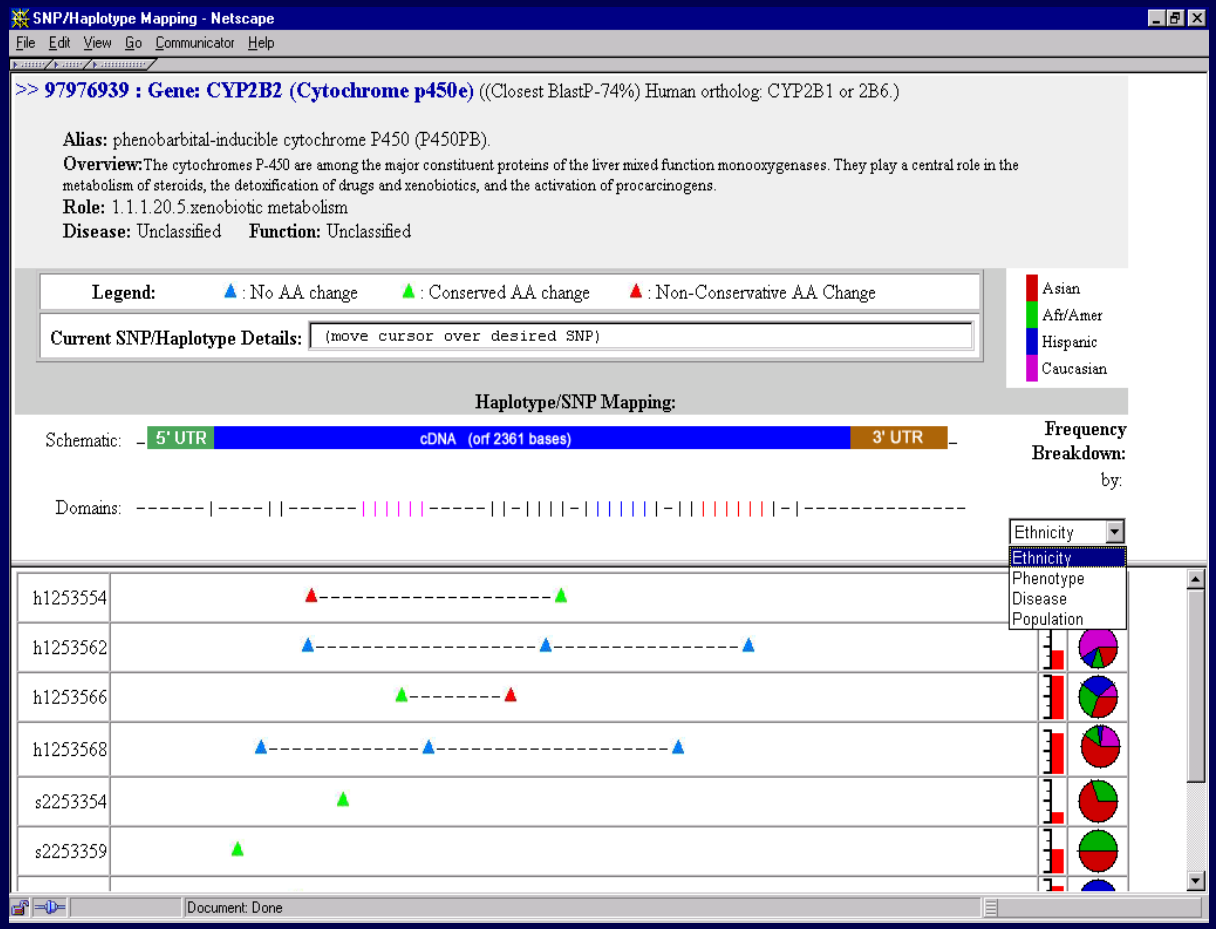
In progress

- Combined visualization of expression/interaction data
- Drosophila whole-genome interaction scan (with Rubin group)

SNP-based association studies

- Cross-validation
  - Candidate genes from expression
  - Independent statistical/biological validation
  - QTLs from genetics
- Genetic determinants of complex disease
  - Risk factors, low penetrance, no clear Mendelian inheritance
  - Traditional linkage analysis has low power
  - Association: direct effects of causative mutations
- Requirements for association tests
  - Causative/dense marker set (SNPs)
  - Large population (1000s to 10,000s)
  - Cheap genotyping

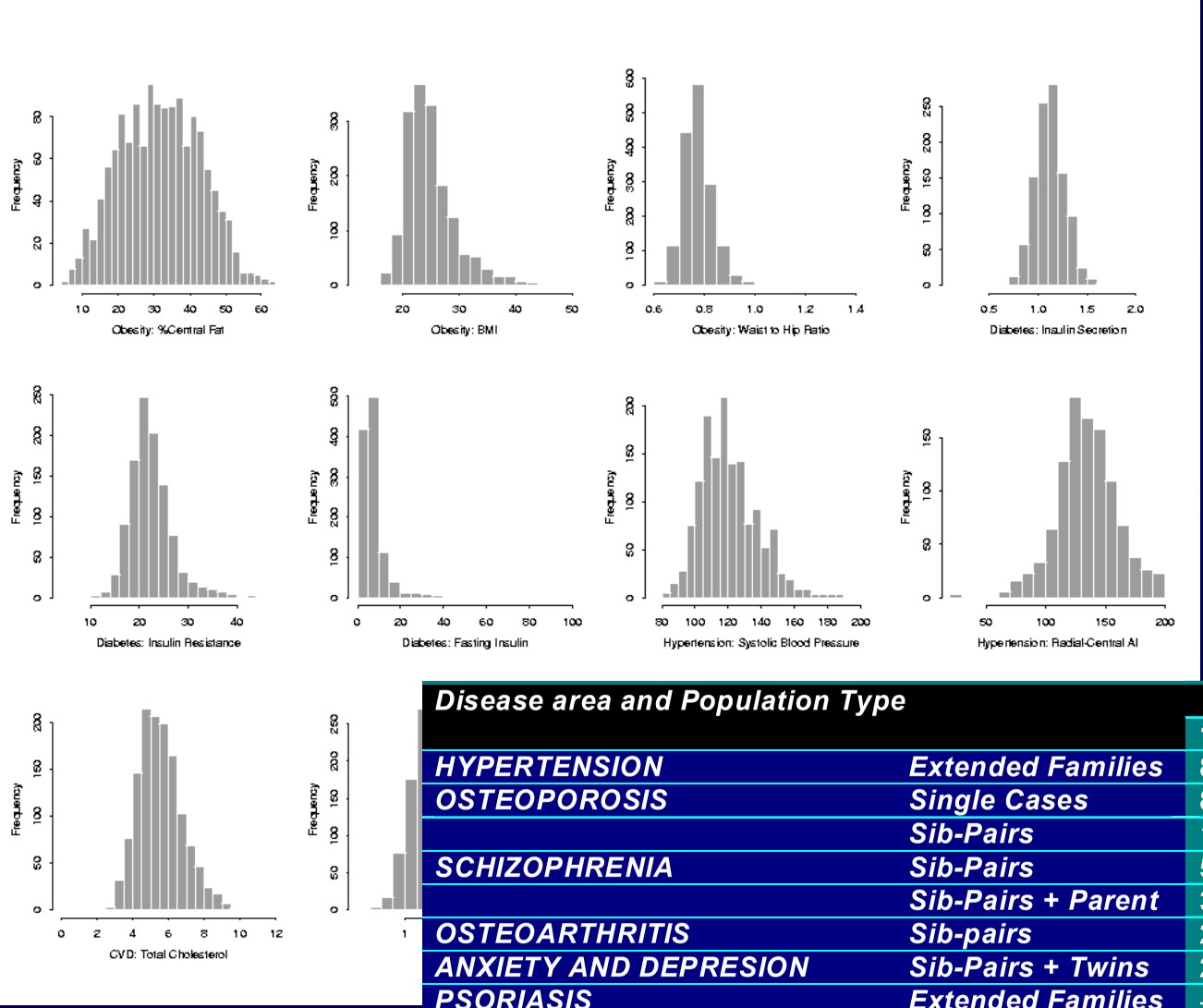
Causative SNPs



Candidates:

- Non-cons aa change
- Drug target
- Disease pathway
- Drug metabolism

Large populations



GEMINI GENOMICS

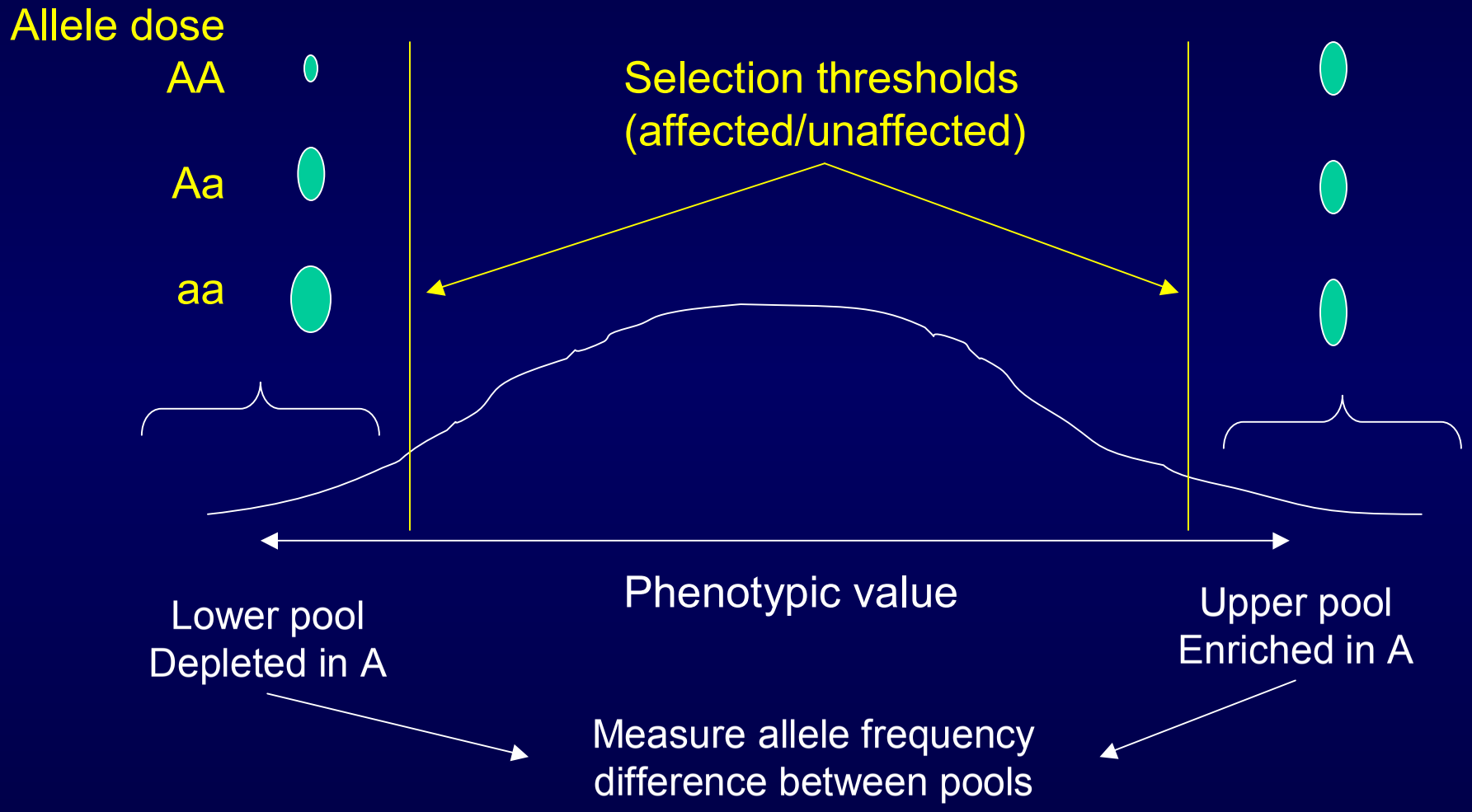
Cost:

2000 genotypes
x 1000 candidates
2 M genotypes

target value ~ $1M

Disease area and Population Type		Year				
		1999	2000	2001	2002	2003
HYPERTENSION	Extended Families	800	1000	1000	1000	1000
OSTEOPOROSIS	Single Cases	800	800	800	800	800
	Sib-Pairs	1200	1500	1500	1500	1500
SCHIZOPHRENIA	Sib-Pairs	500	800	1000	1000	1000
	Sib-Pairs + Parent	300	500	1000	1000	1000
OSTEOARTHRITIS	Sib-pairs	200	200	200	200	200
ANXIETY AND DEPRESION	Sib-Pairs + Twins	2600	2600	2600	2600	2600
PSORIASIS	Extended Families	-	600	900	900	900
TOTAL ACROSS ALL DISEASE-SPECIFIC STUDIES		4500	8000	9900	9900	9900

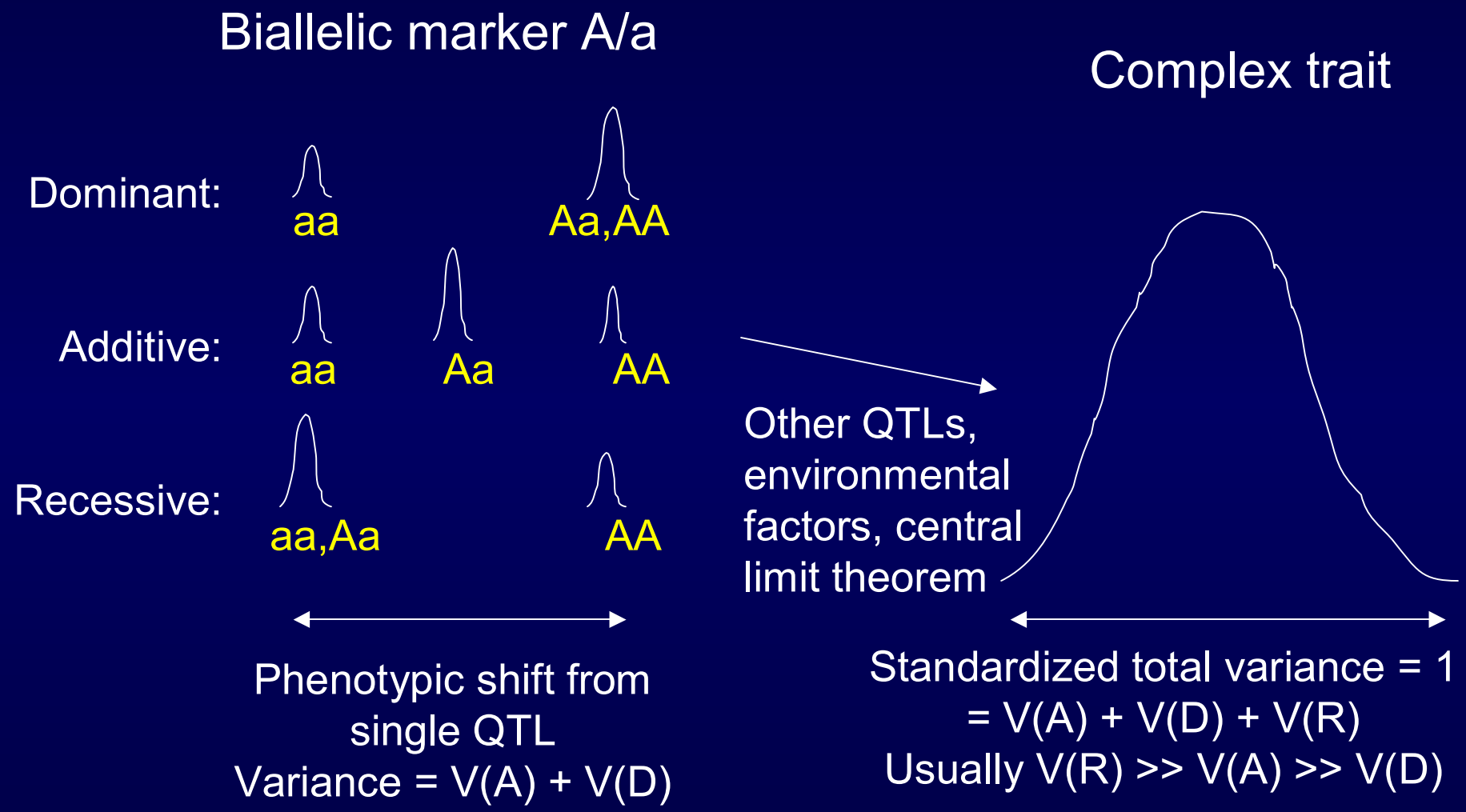
Cheap genotyping: pooling



Choosing a threshold

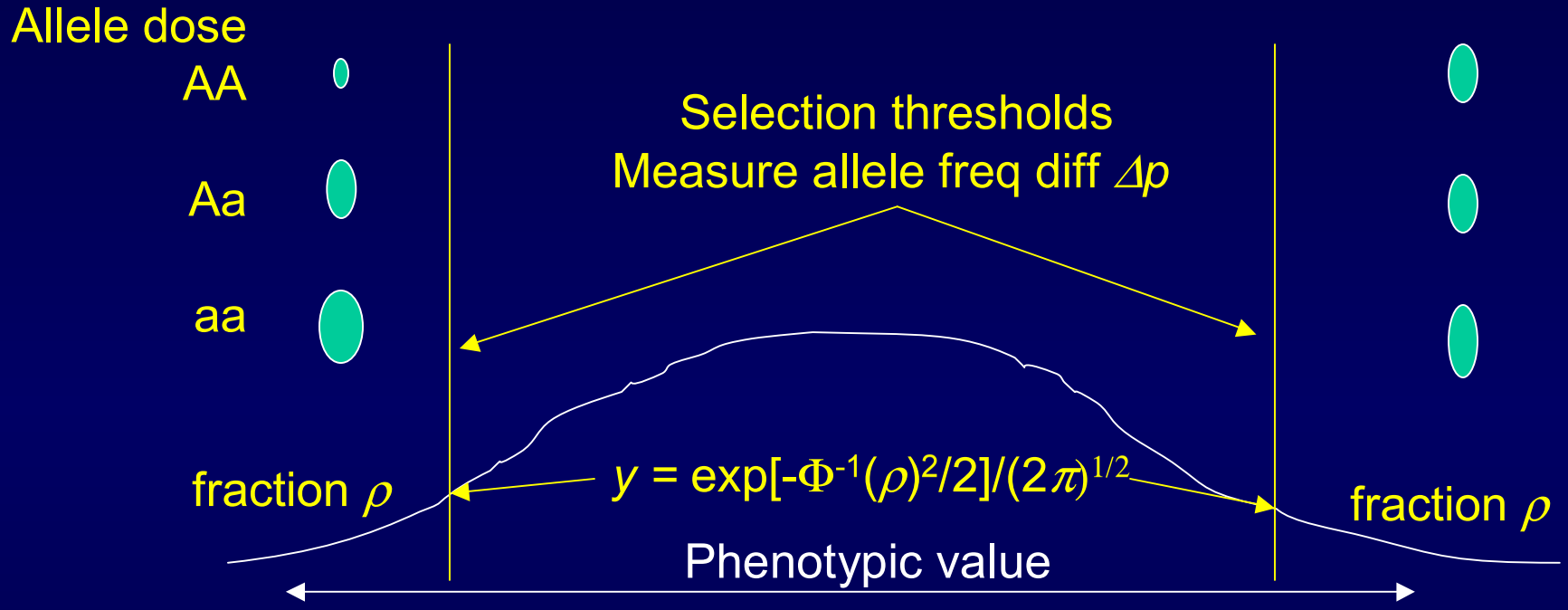
- Optimization: transform a quantitative phenotype into a qualitative phenotype
- How does the optimal threshold depend on
  - Desired false-positive rate
  - Population size
  - Allele frequency
  - Inheritance mode (dominant, additive, recessive)
- What happens when all you have is a qualitative, disease/normal phenotype?
- Collaboration with Aruna Bansal and Pak Sham, Gemini Genomics

Variance components model



Standard model for QTL mapping, Falconer & McKay

Analytic results



$$E(\Delta p) = 2^{1/2} y \sigma_0 \sigma_A / \rho \sigma_R$$

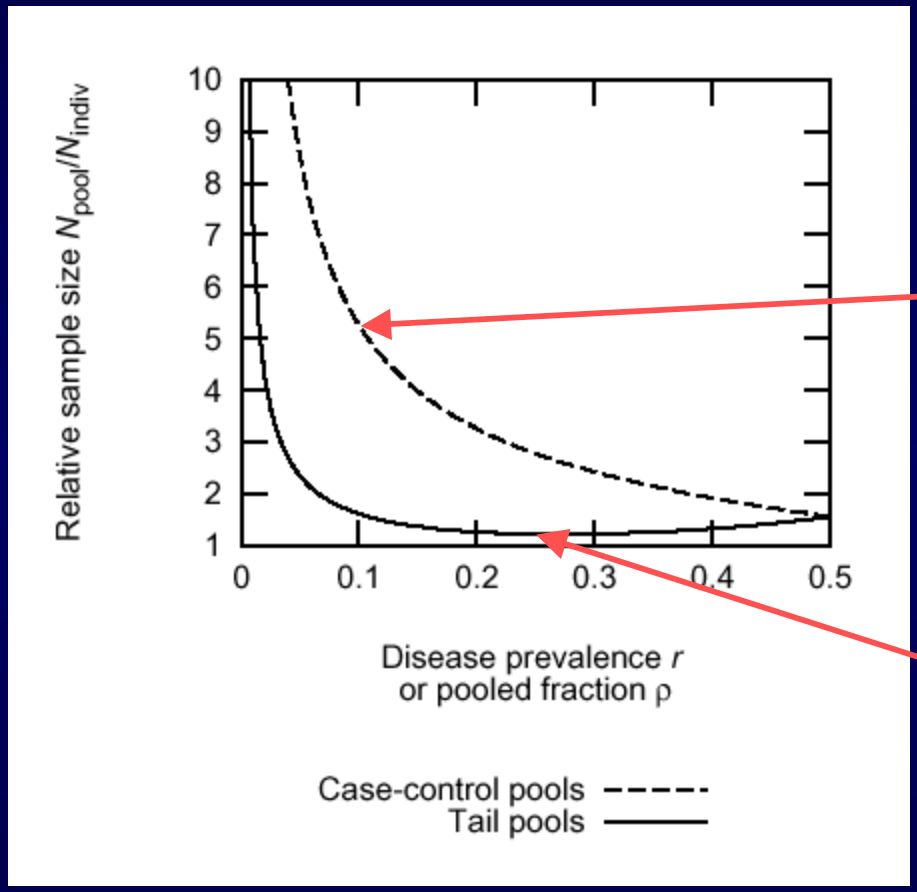
$$\mathrm{Var}(\Delta p) = \sigma_0^2 / \rho N$$

$$N = [z_\alpha - z_{1-\beta}]^2 [V_R / V_A] \cdot \rho / 2y_\rho^2$$

Geometric factor,
1 for indiv genotyping
1.24 for pooled DNA

Relative efficiency

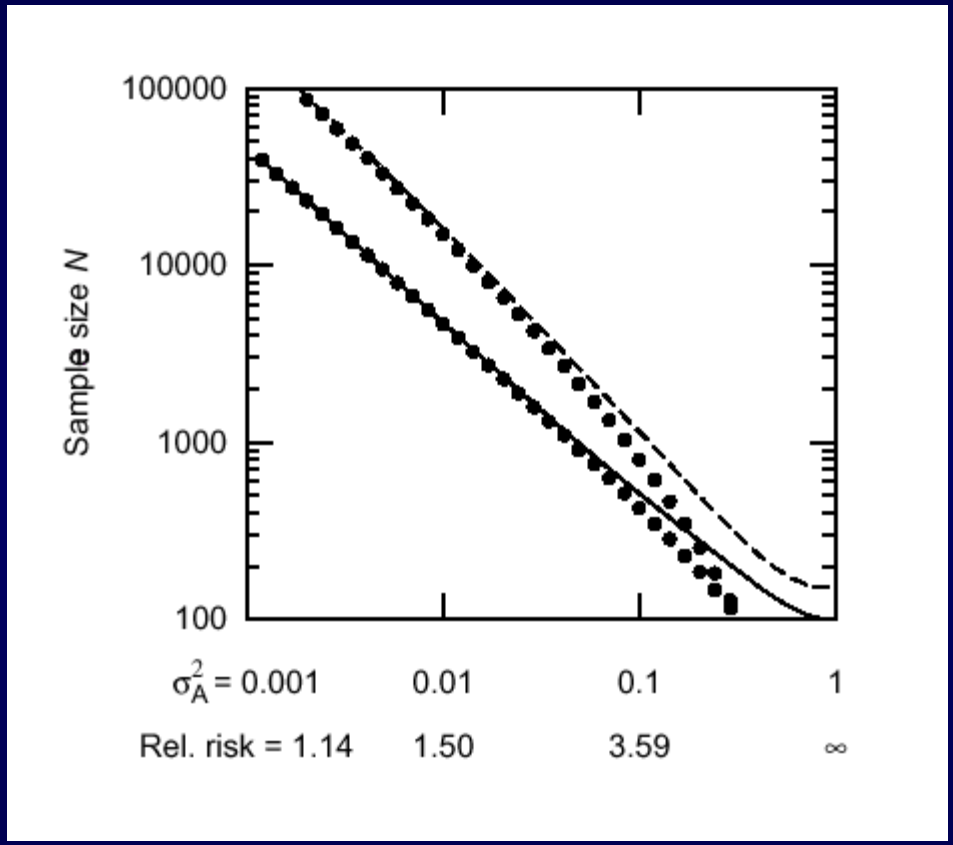
Population required relative to individual genotyping for same type I and type II error rate



Case-control pooling for typical 10% disease incidence is 5X less efficient

Best efficiency at 27%
Only 25% less efficient than individual genotyping

Population size



Genome scan:
Type I error 5×10^{-8}
Type II error 0.2

Lines: exact calculation
Dots: analytic formulas

V(A) ~ 1 / # genes involved

Additive variance is equivalent to a multiplicative haplotype-relative-risk model. Relative risk is calculated at an allele freq of 10%
Monogenic trait: V(A) ~ 0.1 or more

Applications

- Disease-risk markers
  - 11 major areas
  - 100s of phenotypes
  - 1000s of sib-pairs for initial screen
  - Independent disease-specific follow-up populations
- Adverse drug effect markers
- Efficacy markers, personalized medicine

Summary

- Using genomics to improve drug discovery and development
- Exploratory analysis of gene expression
  - Significance thresholds for clustering
  - Identification of disease/drug-response pathways
  - Expression-based markers for drug toxicity, efficacy (pharmacogenomics)
- Protein pathways
  - High-throughput PathCalling Y2H system
  - Overlaying with expression, metabolic pathways
- Genetic variation
  - Large-scale association studies: SNPs, pooled DNA
  - New targets
  - Disease-risk, drug-response markers (pharmacogenetics)

Acknowledgements

- Gene expression
  - CuraGen's GeneCalling bioinformatics group: Darius Dziuda, Shu-Xia Li, Ying Li, Yi Liu, John Tobias, Yi Zhao
  - Prof Rebecca W Doerge and Brian Munneke, Purdue
- Protein pathways
  - Jim Knight (CuraGen)
- Large-scale association studies
  - Pak Sham (Univ of London) and Aruna Bansal (Gemini Genomics)
- CuraGen's genomics facility
- We're hiring

jsbader@curagen.com (203)974-6236