



Stanford - South Africa

*Biomedical Informatics Program*



# Approaches to pharmacogenomics studies

Russ B. Altman, MD, PhD

Professor of Genetics, Bioengineering &  
Medicine

Stanford University



# How can we study PGx?

---

Broadly, there are two approaches:

1. Genotype variation to Phenotype
2. Phenotype variation to Genotype



# Genotype to Phenotype

---

- From the start: Suspicion/knowledge that a gene or gene family is likely to be important for drug response.
- So, look for genetic variation in these genes, and characterize the functional significance
- E.g. Phase I oxoreductase enzymes, Phase II conjugating enzymes, transporters



# Genotype to Phenotype

---

- Screen individuals (by sequencing) from different populations for polymorphisms (SNPs, indels, etc...)
- Polymorphisms with high frequency are then studied phenotypically
  - First with molecular, cellular assays
  - Then, with clinical studies



# Genotype to Phenotype

- Example: new transporter molecule
- Sequence gene in 100 individuals from different ethnic groups
- Find most common variations (coding)
- Put transporter in cell system (e.g. yeast) and measure transport phenotype (e.g. uptake of radioactive small molecule)
- If functional differences, then...
- Study clinically with hypothesis about increased or decreased function in individuals with polymorphism.



# Genotype to Phenotype

---

- Note: common polymorphisms may also be in promoter regions, introns, synonymous coding regions
- Then, studying protein product not directly useful.
- Instead, must study rates of expression, degradation.
- Still can advance to clinical hypotheses, based on accumulated evidence.



# Problems with G to P

---

- How do you choose where to look for variation (exons vs. everything)?
- How do you choose which polymorphisms to followup on functionally?
- How do you know which drugs may be affected by gene and its polymorphisms?
- What if there is no significant variation in the gene? Not much to follow up on...



# Phenotype to Genotype

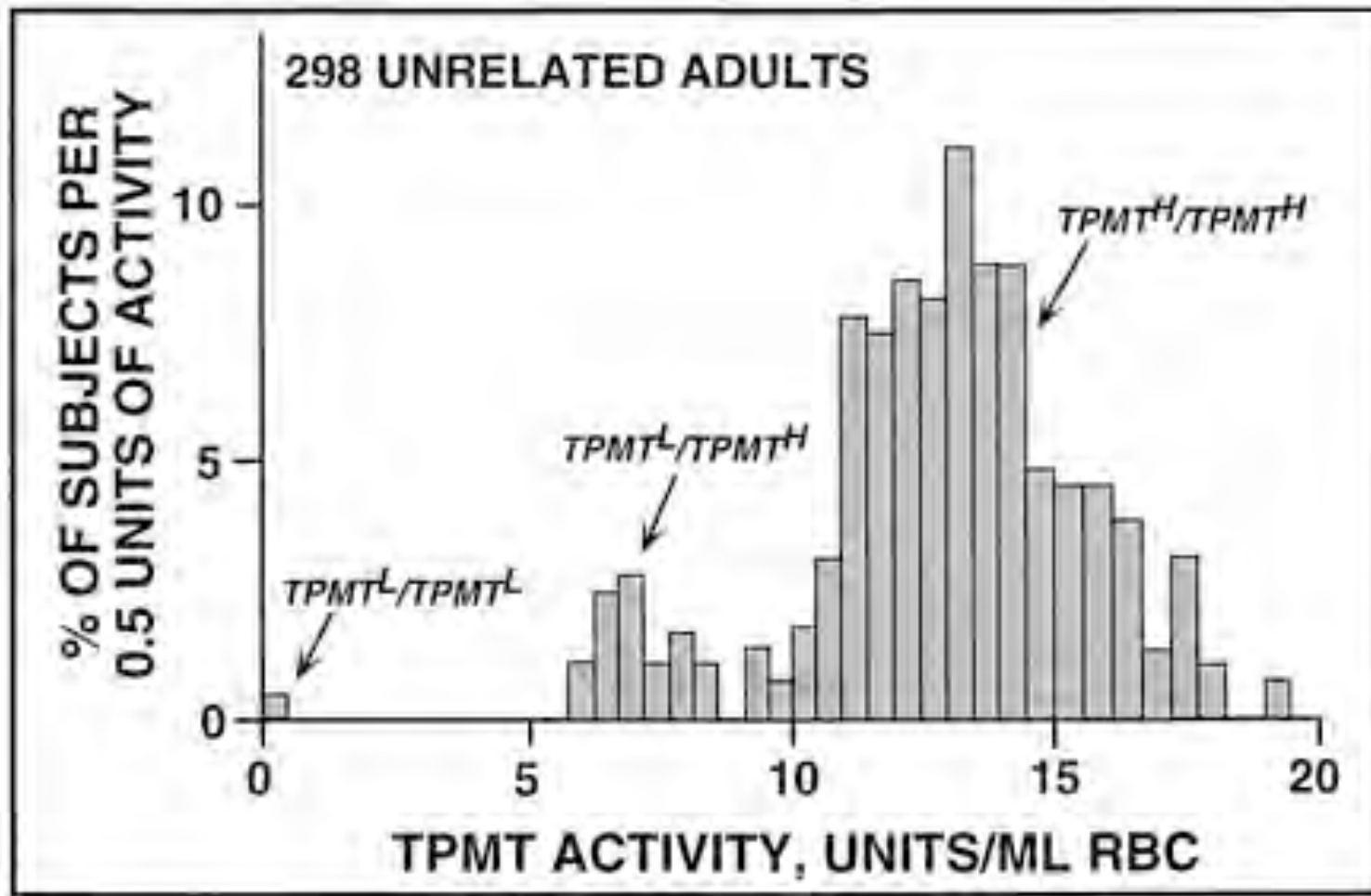
---

- From the start: Suspicion/knowledge that a drug-response phenotype shows marked variation in population. Likely genetic.
- So, find patients with high/low phenotype, and use knowledge of drug pathway to find genotypic variations that explain.

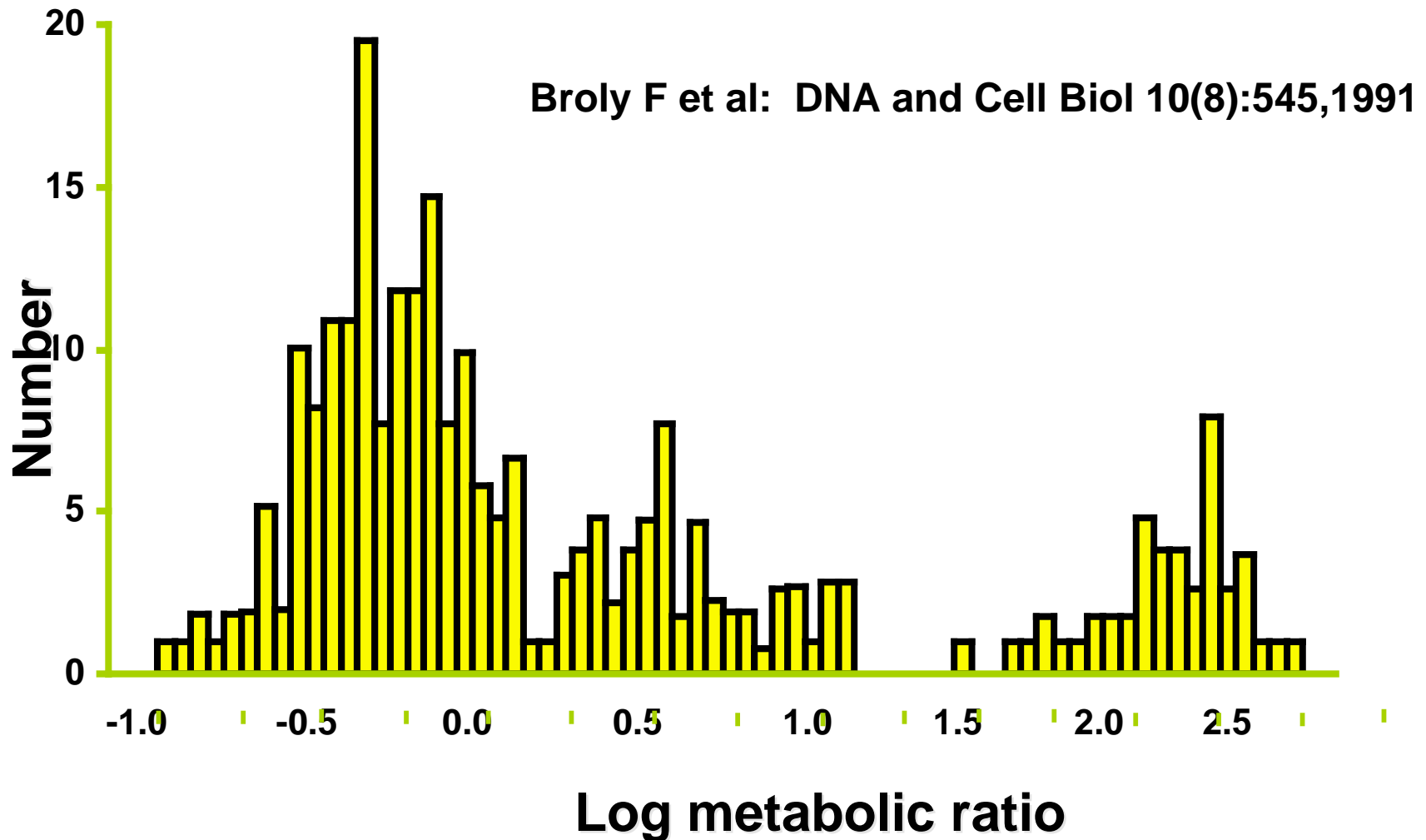


# Variation in TPMT Activity

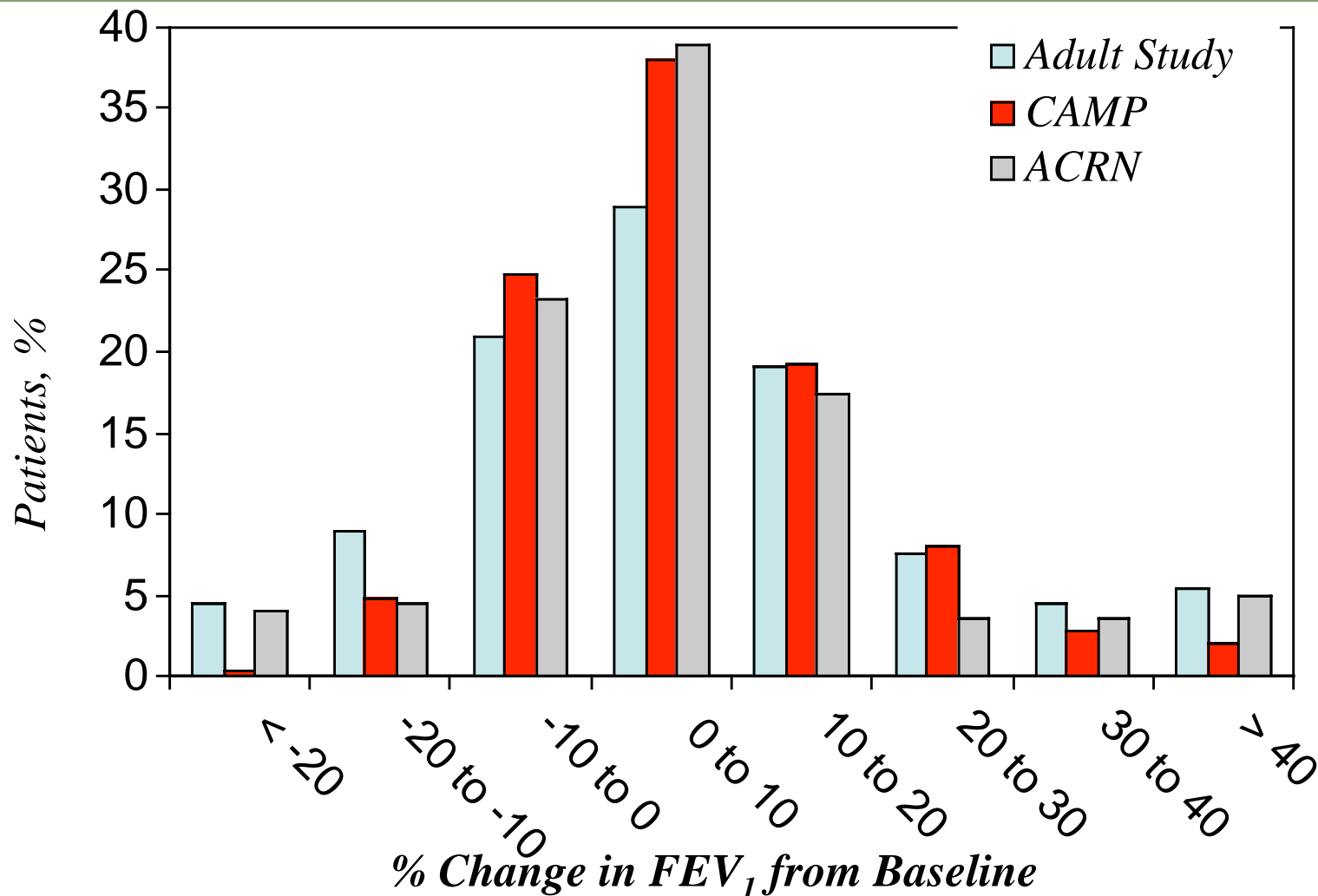
Weinshilboum (Mayo Clinic) 2001



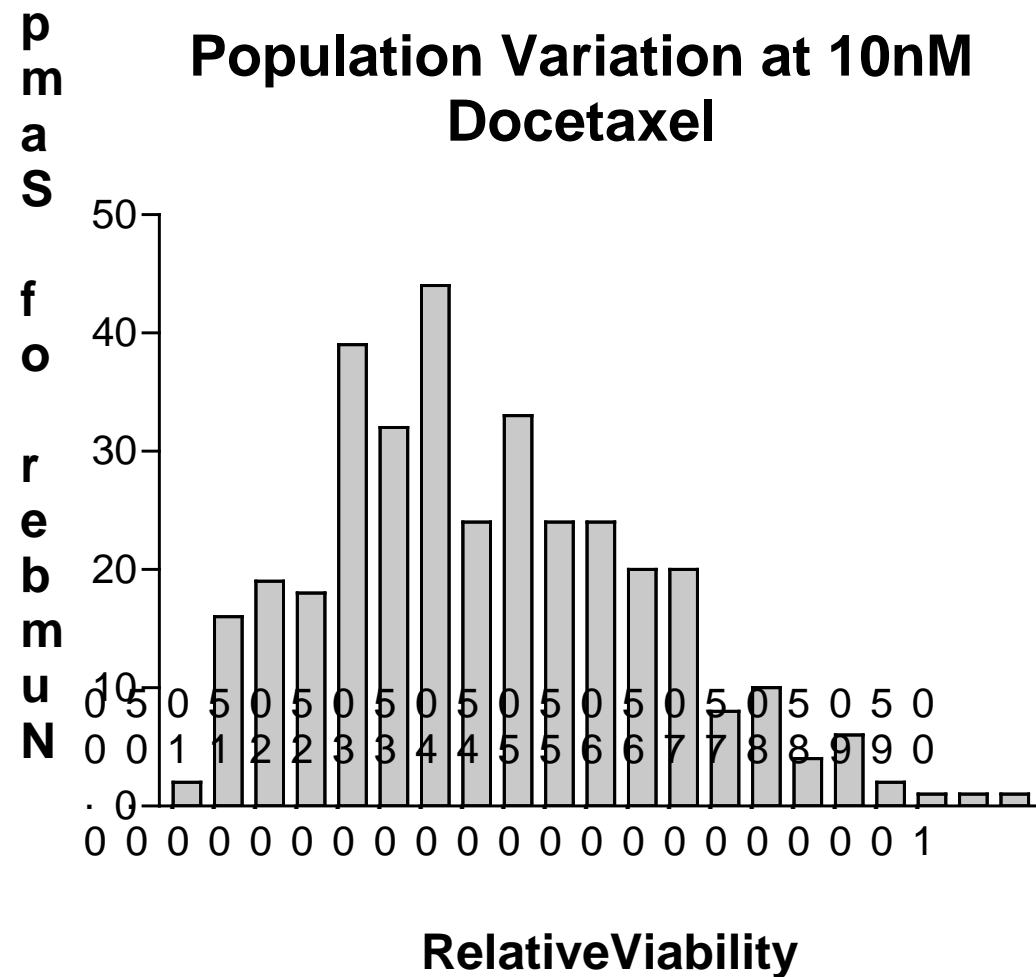
# Distribution of Debrisoquine 4-Hydroxylase Activity



# Distribution of FEV<sub>1</sub> Change in response to inhaled steroid for asthma



# Microarray analysis of extreme samples (McLeod et al)



# Phenotype to Genotype

- Given the phenotype variation, next must find genotypic variation:
  - Candidate genes from known pathways of PK or PD
  - Whole genome assessment of variation (more on this later)
- Often sample the tails of the distribution to find individuals with the most differences



# Problems with P to G

- Need to know something about the genes involved in the drug
- May be multiple genes with small effects
- Need to go back and see how much of the variation is really explained by genetics.  
E.g. Warfarin work focused on CYP2C9 for many years, recently VKORC1 found to explain much more of variation in dosage.



# Future looks good for P to G

---

- Whole genome scans for SNPs allow large populations to be genotyped at MANY positions.
- Then, becomes issue of finding the SNPs that best predict the variability of interest.



# Whole Genome Scan Strategy

---

1. Find population with variability of clinical interest.
2. [Ideally: Establish that variability is heritable]
3. Pick outliers, and perhaps some central individuals.
4. Genotype 100,000 to 300,000 SNPs distributed throughout genome in ~1000 individuals.



# Whole Genome Scan Strategy

5. Compute correlation of individual SNPs with phenotype.
  - Need to choose appropriate measure of correlation
  - Genotype: e.g. AA, AG, GG
  - Phenotype: quantitative [e.g. 0 to 1.0] vs. categorical [e.g. X, Y or Z]
  - Epistasis: interaction of SNPs. May be that need to look at A/a & B/b to see effect (computationally much more complex!)



# Whole Genome Scan Strategy

---

6. Examine regions of genome with most correlated SNPs. May identify numerous regions, if multiple genes are involved.
  - Single gene = strong association (unlikely)
  - Multiple genes = multiple weak associations
7. Use independent sources of data to evaluate the variation genomic regions for supporting evidence.



# Some independent data sources

---

- Expression data: examine response of cells to drug to see what genes are up/down regulated in response to drug.
- Linkage analysis: Look for correlation of phenotype with inherited markers in family studies
- Proteomics: examine proteomic profiles of cells to see if there are phenotypic differences associated with drug



# Whole Genome Scan Strategy

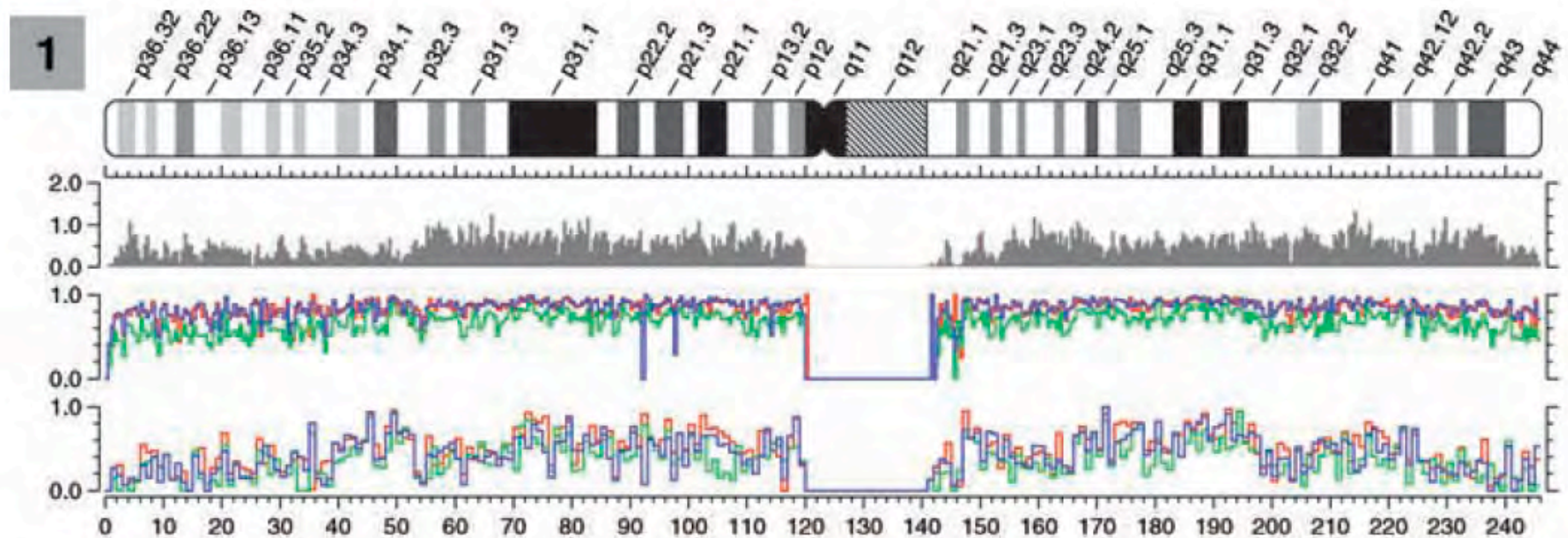
---

8. If able to focus on region that is suggested by independent analyses, then examine genes around the correlated SNP
9. Because of LD, SNP is likely in region, but NOT the functionally important SNP
10. REPLICATION: In a smaller group (~100) of separate cases, do focused sequencing/genotyping at higher density to replicate findings and identify SNPs likely to be functional.



# Will whole genome work?

Perlegen genotypes 1.6 millions SNPs in 71 people.



Hinds et al, Science 307, p 1072

LEGEND  
Top track: genotyped SNPs, per kb  
Middle track: fraction of common SNPs in high LD with another SNP  
Bottom track: fraction of interval covered by LD bins > 50 kb  
Red: European American  
Green: African American  
Blue: Han Chinese

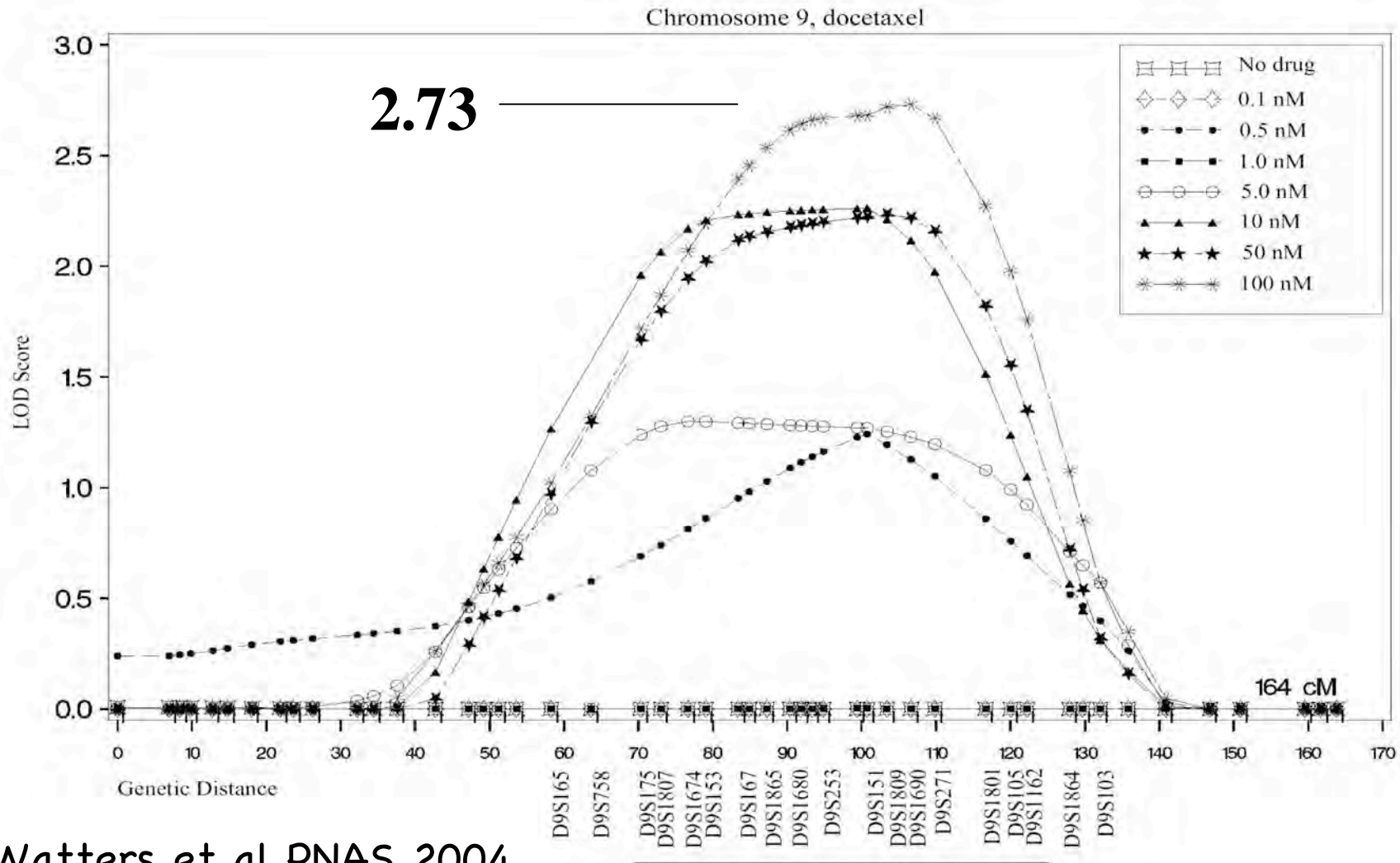


# McLeod et al have used for PGx

1. Use Ceph panel of genotyped individuals (family trios), with cell lines available.
2. Developed a drug assay for sensitivity to drug on cell lines
3. Tested all cell lines for assay
4. Performed linkage analysis to find overall region of phenotype linkage
5. Performed microarray expression to narrow down genes of interest
6. Used SNP data to find correlated SNPs for phenotype



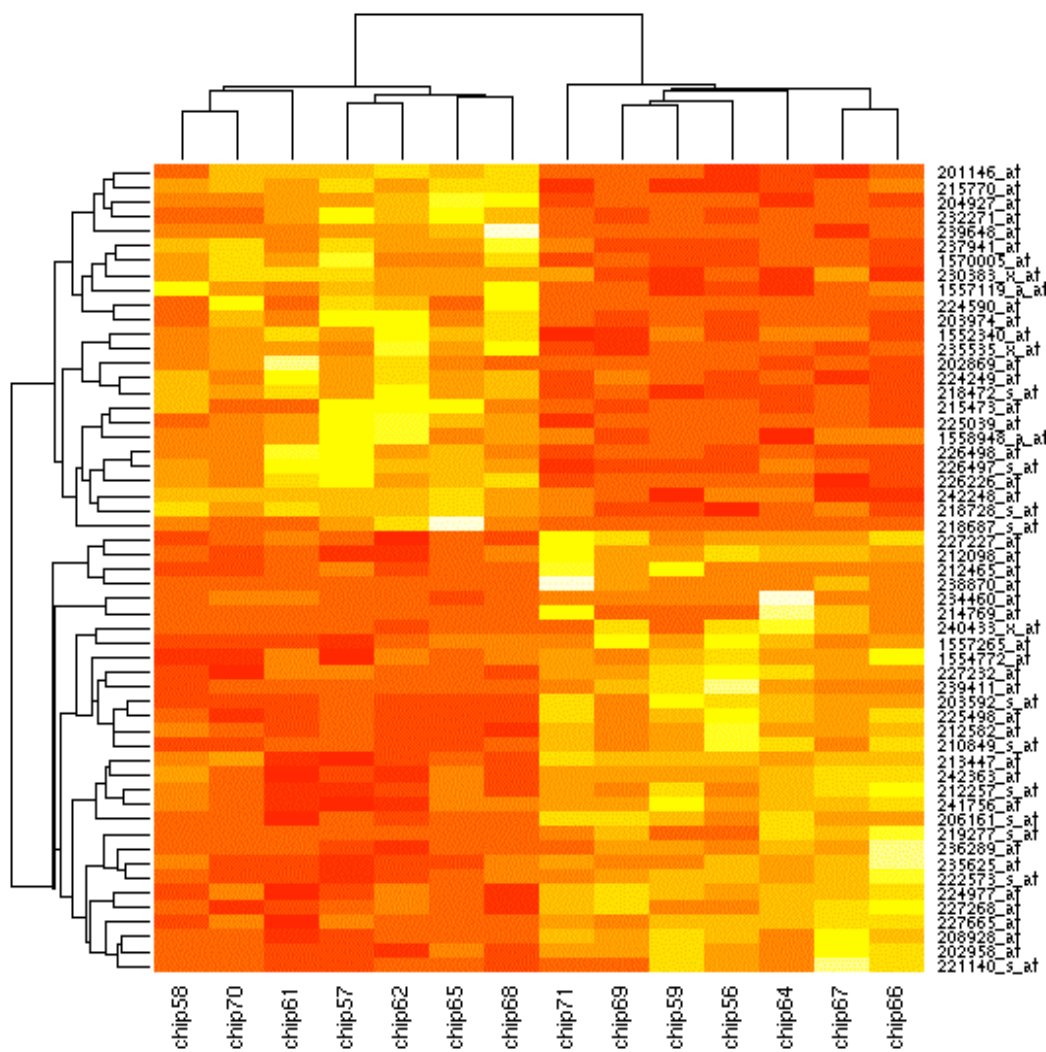
# Linkage analysis



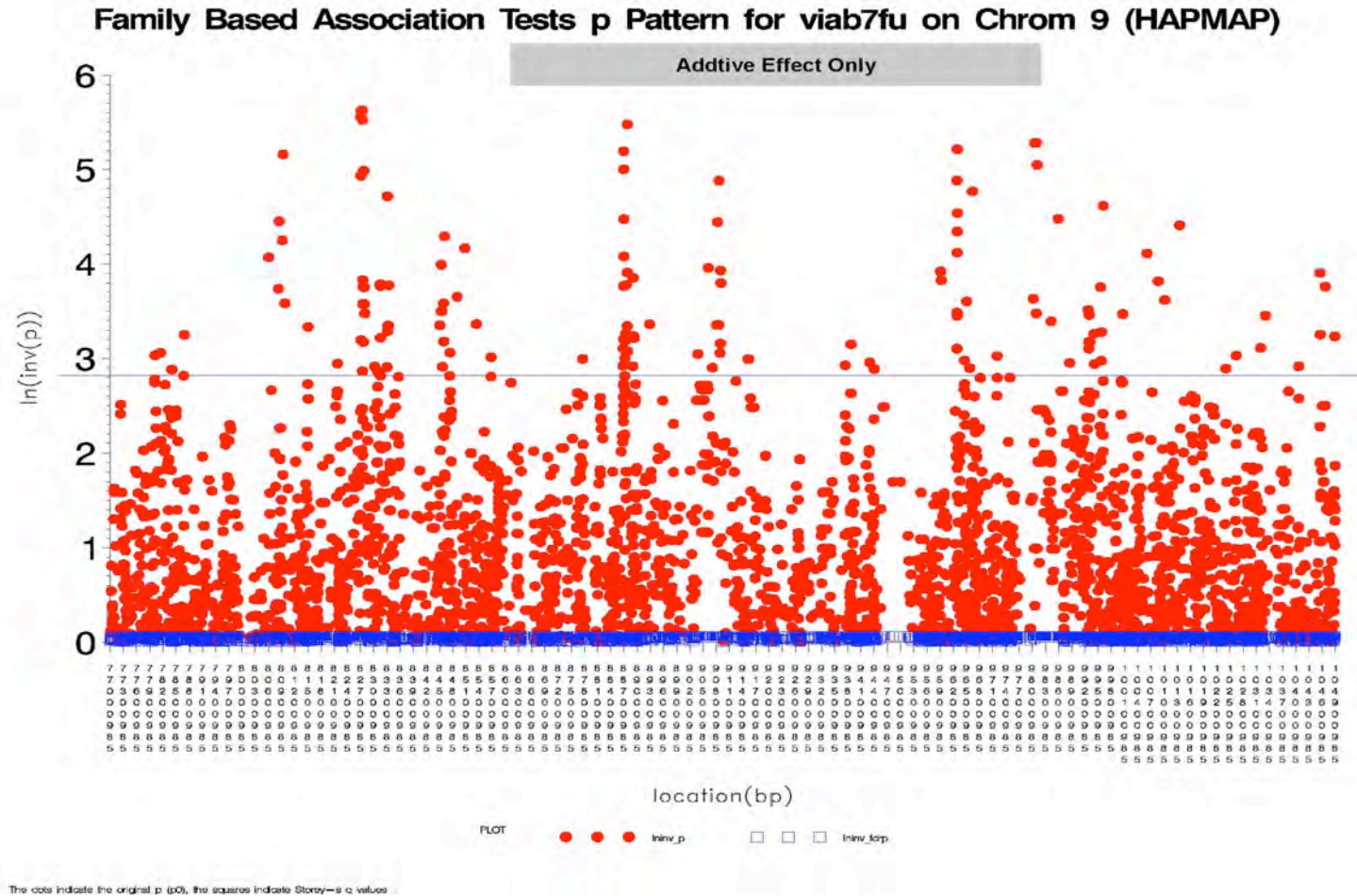
Watters et al PNAS 2004



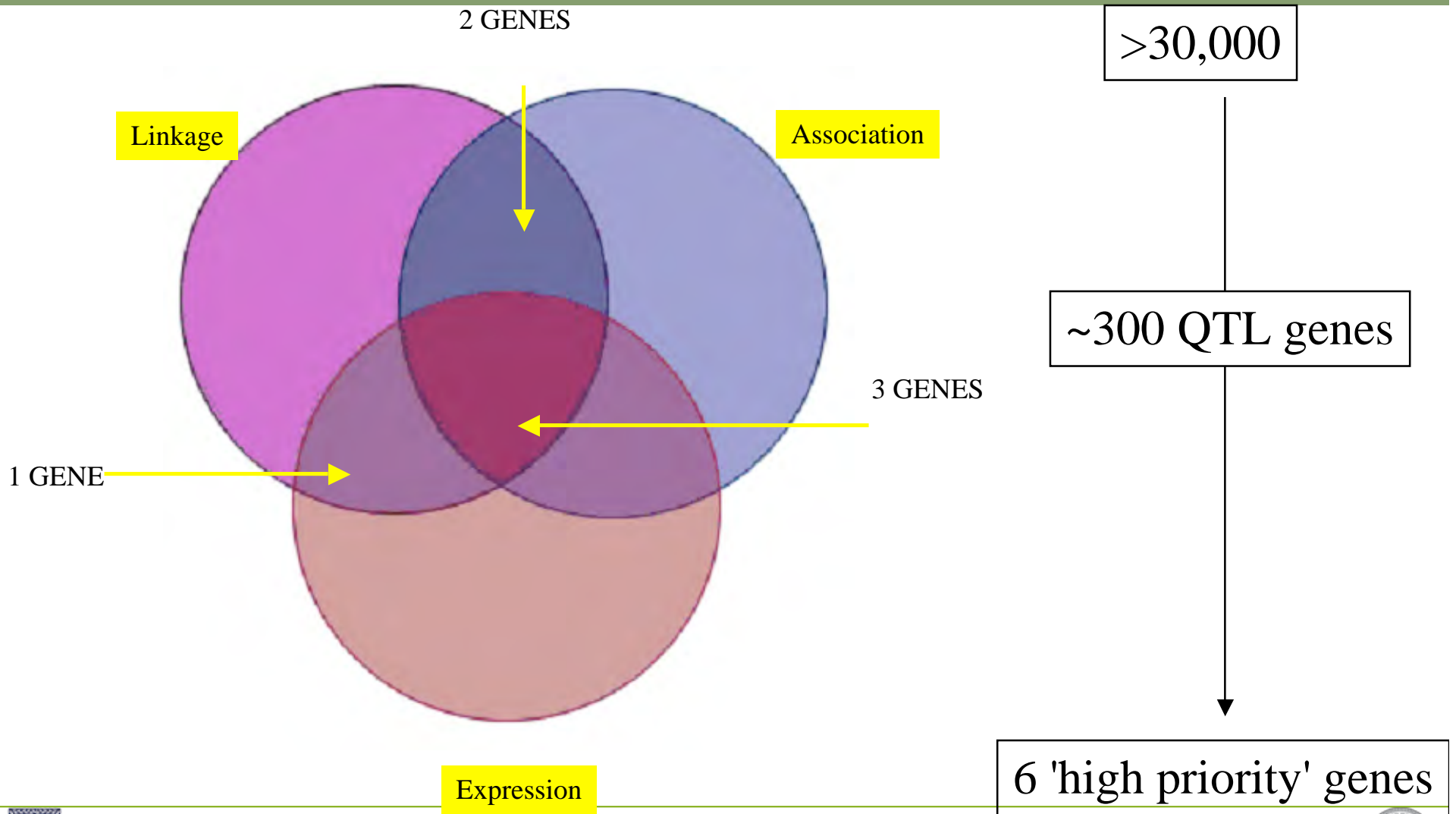
# Genes associated with docetaxel sensitivity



# SNP associations



# Whole Genome Scan Summary



# Analogy to machine learning

---

- Machine learning:
  - Independent features
  - Dependent variables to be predicted
  - Large data sets
    - Web usage
    - Consumer patterns of behavior
    - Large database association mining



# Analogy to machine learning

---

- Genotypes + environmental variables = independent variables
- Phenotypes = dependent variables to be predicted
  
- Phenotype =  $F(\text{Genotype} + \text{Environment})$



# Issues in ML

- Feature selection: which features to include as independent variables?
- Features may be correlated or identical
- Too many features may confuse machine learning algorithm
- Genotype/Phenotype
  - SNPs that are correlated (LD) can be removed



# Issues in ML

- Nature of  $f(\text{genotype} + \text{environment})$ ?
- If  $f$  is linear = weighted combination of genotypes = easier to detect
- If  $f$  is nonlinear = complicated function of genotypes = much harder to detect
- Certain machine learning algorithms better for different situations



# WEKA

---

- Public domain collection of machine learning algorithms
- Provide clustering and classification algorithms
- Relatively easy to use
- Free to download
- Subject of laboratory on Tuesday afternoon.

