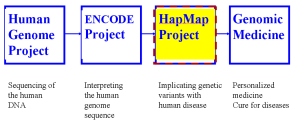
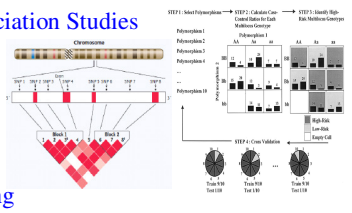
Computational Methods in Genomics
PART THREE

Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

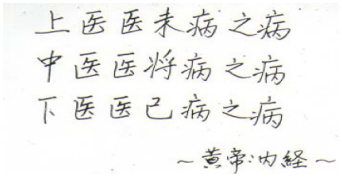
Outline

- HapMap Project
- Human Genetic Variation
  - SNP - Insertion
  - CNV - Deletion
- Genome-Wide Association Studies
- MDR
- Random Forest
- Filtering Algorithms
- Wrapper Algorithms
- Genetic Programming





The Superior Doctor



Superior doctors prevent the disease
Mediocre doctors treat the disease before evident
Inferior doctors treat the full blown disease

-Huang Dee: Nai - Ching
(2600 B.C. 1st Chinese Medical Text)

Preventive Medicine

- Prevent disease from occurring
- Identify the cause of the disease
- Treat the cause of the disease rather than the symptoms
- Genomics identifies the cause of disease
- "All medicine may become pediatrics" Paul Wise
- Effects of environment, accidents, aging, penetrance …
- Health care costs can be greatly reduced if
  - invests in preventive medicine
  - one targets the cause of disease rather than symptoms

Pathway to Genomic Medicine

Human Genome Project	ENCODE Project	HapMap Project	Genomic Medicine
Sequencing of the human DNA	Interpreting the human genome sequence	Implicating genetic variants with human disease	Personalized medicine Cure for diseases

The Reference Human Sequence

- The **reference sequence** for the human genome should not be viewed as just one long string of static characters.
- Instead, it is riddled with variable sites all along the sequence.
- Given that the number of people exceeds the number of bases in the genome, we can imagine that every base in the genome has had its chance to be different. [Baxevanis & Ouellette, 2005]

Genomic Variations

- Collection of genomic variations makes any person a unique human being. It contributes to that person's:
  - Potential to learn
  - Predisposition to disease
  - Predisposition to drug addiction
  - Response to pharmaceutical interventions
- There are variations within, as well as, between populations.
- The variation between individual genomes has sparked a biotech boom in the area of SNP discovery.

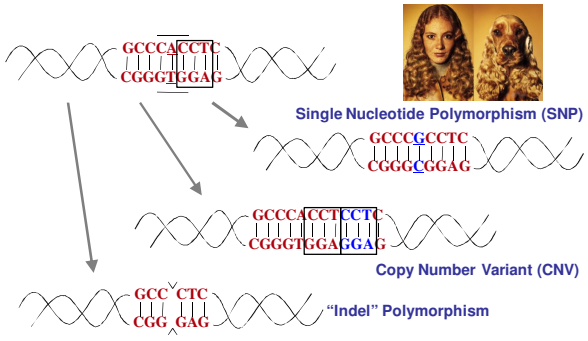
Variation in Human Genome

- How much variation is there in the human genome?
  - The biomedical field is interested in disease-causing variations.
  - What is often considered as a "simple" disease has complex genomic underpinnings.
- How are **genomic variations** used to determine the causes of complex phenotypes?
- How do **genomic variations** influence effective medical interventions?

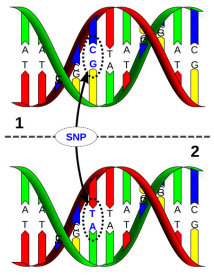
Human Genetic Variation

- Copy Number Variation (CNV)
  - A polymorphism in which the number of repeats of a DNA sequence at a location varies from person to person
- Single Nucleotide Polymorphism (SNP)
  - Major differences between human beings
- Other structural variations
  - Includes deletions, insertions, duplications, inversions, and translocations

Types of Genomic Variations

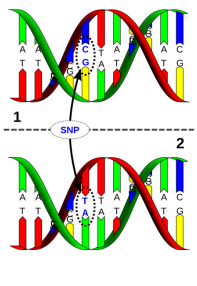


SNPs and Human Variations (I)



- A SNP is a single base-pair mutation that occurs at a specific site in the DNA sequence.
- SNPs are responsible for over 80% of the variation between two individuals
  - ideal for the task of hunting for correlations between genotype phenotype.

SNPs and Human Variations (II)



To classify a variation as a SNP it should occur in at least 1% of the population.

Single Nucleotide Polymorphism

Single Nucleotide Polymorphisms are single bases at a particular locus that are different in different individuals.

```
GCATGCATGCATGCAT
||||||||||||||||
CGTACGTACGTACGTA

GCATGCAaGCATGCAT
||||||||||||||||
CGTACGTtCGTACGTA
```

90% of all human chromosomes have the following sequence at a particular location (i.e., unique locus)

But 10% of all alleles have a slightly different sequence at that particular location (i.e., unique locus)

What is a Polymorphism?

- A **polymorphism** is a difference in DNA sequence among individuals.
- **Genetic variations** occurring in more than 1% of a population would be considered useful polymorphisms for genetic analysis.
- **SNP**: position in a genome at which two or more different bases occur in the population, each with a frequency greater than 1%.

Applications of SNPs (I)

SNPs are useful for several types of research

1) SNPs and the study of **Evolution**
   - **Example**: Different combinations of SNPs of the taste receptor gene: *Tas2R*.

2) SNPs and **Fingerprinting**
   - **Example**: Criminals and Parental Verification.

Applications of SNPs (II)

3) SNPs in **Biomedical Research**

Example: Manufacturing genotype-specific medication

Most genes contain at least one **SNP**, some of which might have functional consequences.

SNPs could be used to determine which combination of coding alleles is associated with a particular disease.

Phenylthiocarbamide (PTC)

Table 4.4 Global PTC taster SNP frequencies.
PAV is the only taster allele. Sample size for each population appears in parentheses.

SNP Combinations	European (200)	West Asian (22)	East Asian (54)	African (24)	SW Native American (18)
AVI	0.47	0.67	0.31	0.25	—
AAV	0.03	—	—	0.04	—
AAI	—	—	—	0.17	—
PVI	—	—	—	0.04	—
PAV	0.49	0.33	0.69	0.50	1.00

To some individuals the chemical compound **phenylthiocarbamide** (**PTC**) has an intensely bitter taste, while to others it is tasteless. It depends on the SNPs that are present in the receptor gene *Tas2R*.

SNPs and Evolution

- **SNP**s can be used in the study of **evolution**.
- Scientists tested 6 nonhumans primates and found that they were all tasters, in other words, they had the PAV form of *Tas2R*.
- Consequently, humans acquired (evolved) the other SNPs: AVI, AAV, AAI and PVI, after the split from our nearest relative, the chimpanzee.

Do SNPs Produce Common Phenotypes?

- Are there point mutations that lead to diseases?
- Yes.
  - **Example**: Sickle Cell Anemia.
- Four more cases:
  - Skin pigmentation
  - Malaria resistance
  - Mitochondrial SNPs
  - Incorrect mRNA splicing

Case 1: Skin Pigmentation

- We do not understand skin coloration.
- Are there 50 or 500 genes involved in skin pigmentation? We do not know.
- Melanin is a polymer of two oxidized derivatives of tyrosine:
  - Pheomelanin which appears in red-yellow
  - Eumelanin which is less soluble and appears in black-brown.
- Mc1R is a gene involved in skin coloration.

Variable Selective Pressures at Mc1R

- In "Evidence for Variable Selective Pressures at Mc1R" by R. Harding et al.
  - "It is widely assumed that genes that influence variation in skin and hair pigmentation are under selection. To date, the melanocortin 1 receptor (Mc1R) is the only gene identified that explains substantial phenotypic variance in human pigmentation."

Eumelanin and Pheomelanin

- The allele for red hair and the allele for blond hair are both found only in Europeans.
- Europeans have more alleles for the Mc1R gene than Africans.
- Africans have only synonymous alleles of Mc1R that all code for eumelanin, a pigment that produces dark skin and hair.
- Eurasians have many alleles for pheomelanin, a red-gold pigment that produces light skin and hair colors.

Africans and Pheomelanin

- Africans lack alleles for pheomelanin because light skin and hair are disadvantageous in Africa.
  - An African who may have acquired them would have been less likely to survive and leave progeny.
- There is a surprising correlation between red hair and resistance to the anesthetic midazolam
  - The clinical investigators did not discern the reason behind this drug resistance.

Case 2: Malaria Resistance

- A SNP in the promoter of the nitric oxide synthase (Nos2) gene may help fight malaria.
- In East African children, a mutation of T→C in the promoter of Nos2 gave more Nitric Acid in the blood
  - their chances of developing fatal malaria were reduced by about 80%.
- Drugs: Can we regulate levels of Nitric Acid through medication?

Case 3: Mitochondrial SNPs

- **Mitochondria** produce most of our cell's ATP.
- Each mitochondrial gene requires the proper function of 22 tRNA genes and 2 rRNA genes that are also encoded in the mitochondrial genome.
- More than 50 different disease-causing mitochondrial SNPs have been identified.
  - This number will probably increase as we become more proficient at detecting SNPs.

Case 4: Incorrect mRNA Splicing

- Research over the past few years has revealed that exons not only specify amino acids, they also contain within their sequences cues necessary for intron removal.
- Chief among these are **exonic splicing enhancer** (**ESE**) motifs--short sequences of about three to eight nucleotides that sit near the ends of the exons and define the exon for the cellular splicing machinery.

"Price of Silent Mutations", Scientific American, June 2009

Exonic Splicing Enhancer Motifs

- The need for **exonic splicing enhancer motifs** can in fact explain a preference for certain nucleotides in human genes.
- Although the codons GGA and GGG, which encode glycine, can both occur in splicing enhancers, GGA acts as a more potent enhancer, leading to more efficient splicing. GGA is also correspondingly more common close to the ends of exons.

"Price of Silent Mutations", Scientific American, June 2009

Exonic Splicing Enhancers & Silencers

- Splicing of RNA to produce a mature mRNA involves the 5' and 3' ends of each exon, but internal sequences are required as well.
- Although the consensus sequences are uncertain, **exonic splicing enhancers** (**ESE**s) and **exonic splicing silencers** (**ESS**s) are located within exons and are distinct from the terminal splicing junctions.

ESE and ESS in BRCA1

- Krainer found examples of ESE and ESS mutations in BRCA1, which probably explains why some women with silent mutations develop breast and ovarian cancer.
- This illustrates that even silent SNPs can have a profound influence on phenotypes, including polygenic traits such as cancer.

SNPs that are Revealed too Late

- We have just studied four cases of SNPs that lead to traits and diseases:

  Skin pigmentation — Malaria resistance
  Mitochondrial SNPs — Incorrect mRNA splicing
- Unfortunately, some SNPs do not reveal themselves until it is too late:
  - Fava bean SNP
    - What is food to some people may be fierce poison to others
  - Variations in medication responsiveness

A SNP in Fava Beans might Kill

- Some people experience a lysis of their red blood cells from the consumption of fava beans.
  - Around 10% of the population cannot produce glucose-6-phosphate dehydrogenase (**G6PD**).
- **G6PD** is a metabolic enzyme found in the cytoplasm of every cell.
- **G6PD** produces nicotinamide adenine dinucleotide phosphate (NADPH) which helps regenerate the enzymes used to neutralize the cellular toxin hydrogen peroxide.

G6PD Deficiency

- **G6PD** deficiency is the most common human enzyme deficiency
  - An estimated 400 million people worldwide are affected by this enzymopathy.
- One benefit of having **G6PD** deficiency is that it confers a resistance to malaria.
- **G6PD** deficiency is also sometimes referred to as favism since some **G6PD** deficient individuals are also allergic to fava beans.

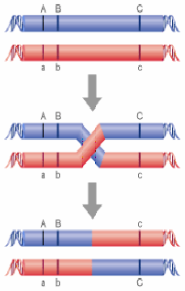
rialto.com/g6pd

G6PD and SNPs

- SNP 376A → G produces G6PD with normal activity.
  - It is found in 20% of African males.
- SNP 202G → A reduces G6PD activity.
  - The reduction in G6PD activity is about 10% in 20% of African males.
- SNP 563C → T produces an enzyme with nearly undetectable activity.
  - It is found in 20% of the alleles of Caucasian males living around the Mediterranean Sea.
  - It is known as the "Mediterranean G6PD".

Linkage Disequilibrium

- **Linkage** refers to how close 2 loci are to each other on a chromosome. If they are near each other, we say the 2 loci are linked.
- **Linkage disequilibrium** describes alleles rather than loci. If 2 alleles (or SNPs) tend to be inherited together more often than would be predicted, we say the SNPs are in **linkage disequilibrium**. In other words, they are inherited together more often than other possible SNP combinations.



genome.wellcome.ac.uk

Genetic Mapping

- Genetic mapping is the localization of genes underlying phenotypes on the basis of correlation with DNA variation, without the need for prior hypotheses about biological function.

Genetic Mapping in Human Disease, Altshuler et al., 2008

Genetic Association in Populations

- A possible path forward emerged from population genetics and genomics.
- Instead of mapping disease genes by tracing transmission in families, one might localize them through association studies—that is, comparisons of frequencies of genetic variants among affected and unaffected individuals.

Genetic Mapping in Human Disease, Altshuler et al., 2008

SNPs and Biomedical Research (I)

Two Populations

Single-locus disease

Frequency P1	Frequency P2	Genomic DNA	SNP Haplotype	Phenotype
81%	49%	SNP1 — allele *A* — SNP2 SNP1 — allele *A* — SNP2	1 – 2 1 – 2	*wt*
18%	42%	SNP1 — allele *A* — SNP2 SNP1' — allele *a* — SNP2'	1 – 2 1' – 2'	*wt*
1%	9%	SNP1' — allele *a* — SNP2' SNP1' — allele *a* — SNP2'	1' – 2' 1' – 2'	disease

SNPs and Biomedical Research (II)

Two Populations

Single-locus disease

Frequency P1	Frequency P2	Genomic DNA	SNP Haplotype	Phenotype
81%	49%	SNP1 — allele *A* — SNP2 SNP1 — allele *A* — SNP2	1 – 2 1 – 2	*wt*
18%	42%	SNP1 — allele *A* — SNP2 SNP1' — allele *a* — SNP2'	1 – 2 1' – 2'	*wt*
1%	9%	SNP1' — allele *a* — SNP2' SNP1' — allele *a* — SNP2'	1' – 2' 1' – 2'	disease

Hypothetical example in which 2 SNPs and one gene are associated with a monogenic disease.

SNPs and Biomedical Research (III)

Frequency P1	Frequency P2	Genomic DNA	SNP Haplotype	Phenotype
81%	49%	SNP1 — allele *A* — SNP2 SNP1 — allele *A* — SNP2	1 – 2 1 – 2	*wt*
18%	42%	SNP1 — allele *A* — SNP2 SNP1' — allele *a* — SNP2'	1 – 2 1' – 2'	*wt*
1%	9%	SNP1' — allele *a* — SNP2' SNP1' — allele *a* — SNP2'	1' – 2' 1' – 2'	disease

The allele and its flanking SNPs define one locus. The two SNPs: 1' and 2', and the recessive allele *a* are in **linkage disequilibrium**.

HapMap Project

- Systematic effort to try to catalogue the common variants that exist across human populations.
- Goal: Implication (Correlation) of genetic variants (SNPs and haplotypes) with human diseases.

International Haplotype Map Project (I)

- The goal of the **International Haplotype Map Project** is to develop a haplotype map of the human genome.
- The "HapMap" describes common patterns of human DNA sequence variation, and is a key source for researchers to find genes affecting health, disease, and responses to drugs, and environmental factors.

[Baxevanis & Ouellette, 2005]

International Haplotype Map Project (II)

The **International Haplotype Map Project** is in the process of refining the ever-increasing number of polymorphisms in the human genome to a more manageable set that still captures the underlying variation information, allowing the design of more cost effective association studies.

[Baxevanis & Ouellette, 2005]

SNP Frequencies and LD Patterns

- The International HapMap Project was launched in 2002, with the goal of characterizing SNP frequencies and local LD patterns across the human genome in 270 samples from Europe, Asia, and West Africa.
- The project genotyped about 1 million SNPs by 2005 and more than 3 million by 2007.

Genetic Mapping in Human Disease, Altshuler et al., 2008

Correlation of Common SNPs

- Sequence data collected by the project confirmed that the vast majority of common SNPs are strongly correlated to one or more nearby proxies: 500,000 SNPs provide excellent power to test over 90% of common SNP variation in out-of-Africa populations, with roughly twice that number required in African populations

International HapMap Project

Home | About the Project | Data | Publications

中文 | English | Français | 日本語 | Yoruba

Participating Groups

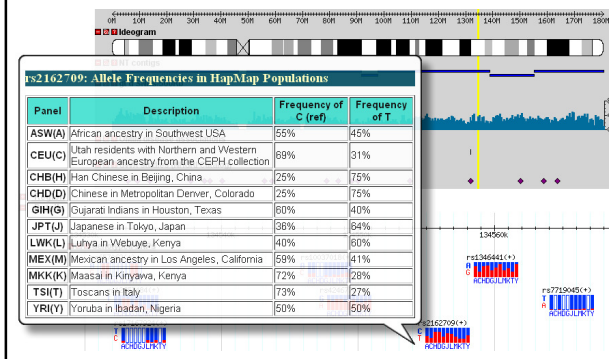
Baylor College of Medicine (USA)
Beijing Genomics Institute (China)
Beijing Normal University (China)
Broad Institute of Harvard and MIT (USA)
Center for Statistical Genetics, University of Michigan (USA)
Chinese National Human Genome Center at Beijing (China)
Chinese National Human Genome Center at Shanghai (China)
Cold Spring Harbor Laboratory (USA)
Eubios Ethics Institute (Japan)
Health Sciences University of Hokkaido (Japan)
Hong Kong University of Science and Technology (China)
Howard University (USA)
Illumina (USA)
Johns Hopkins School of Medicine (USA)
McGill University & Génome Québec Innovation Centre (Canada)
ParAllele BioScience (USA)
Perlegen Science (USA)
RIKEN (Japan)
The Chinese University of Hong Kong (China)
The University of Hong Kong (China)
University of California, San Francisco (USA)
University of Ibadan (Nigeria)
University of Oxford (UK)
University of Oxford / Wellcome Trust Centre for Human Genetics (UK)
University of Tokyo (Japan)
University of Utah (USA)
Washington University, St. Louis (USA)
Wellcome Trust Sanger Institute (UK)

To produce a genome-wide map of common variation

Genotype 6 Million SNPs in Four populations in Two Phases:
- CEPH (CEU) (Europe - n = 90, trios)
- Yoruban (YRI) (Africa - n = 90, trios)
- Japanese (JPT) (Asian - n = 45)
- Chinese (HCB) (Asian - n =45)

Nature 437: 1299-320, 2005

www.hapmap.org



Panel	Description	Frequency of C (ref)	Frequency of T
ASW(A)	African ancestry in Southwest USA	55%	45%
CEU(C)	Utah residents with Northern and Western European ancestry from the CEPH collection	69%	31%
CHB(H)	Han Chinese in Beijing, China	25%	75%
CHD(D)	Chinese in Metropolitan Denver, Colorado	25%	75%
GIH(G)	Gujarati Indians in Houston, Texas	60%	40%
JPT(J)	Japanese in Tokyo, Japan	36%	64%
LWK(L)	Luhya in Webuye, Kenya	40%	60%
MEX(M)	Mexican ancestry in Los Angeles, California	59%	41%
MKK(K)	Maasai in Kinyawa, Kenya	72%	28%
TSI(T)	Toscans in Italy	73%	27%
YRI(Y)	Yoruba in Ibadan, Nigeria	50%	50%

SNP: rs2162709 on chromosome 5

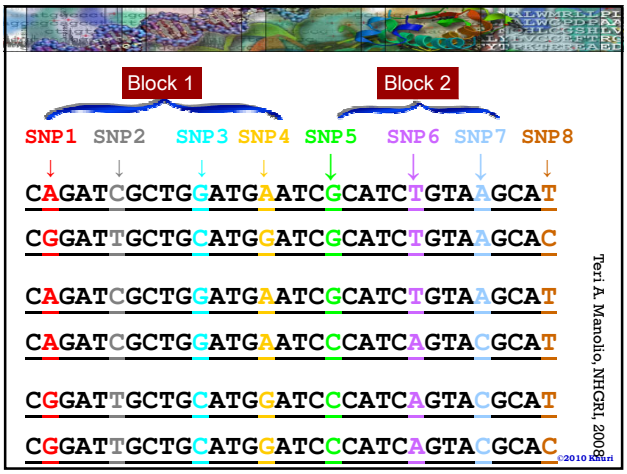
Population Diversity

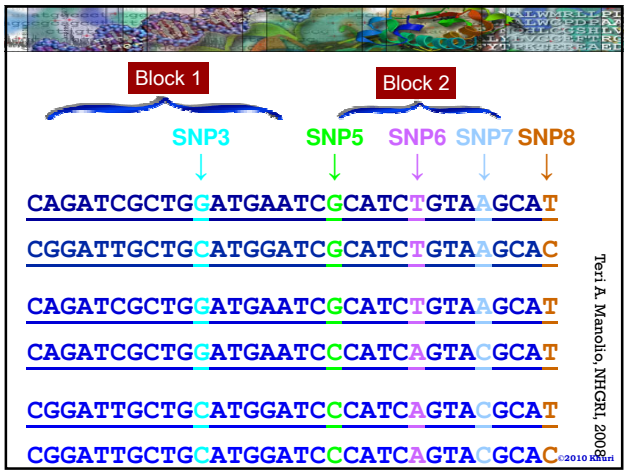
		Sample Ascertainment			Genotype Detail NEW				Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	C/C	C/T	T/T	HWP	C	T
ss23357318	AFD_EUR_PANEL	European	38	IG	0.368	0.526	0.105	0.584	0.632	0.368
	AFD_AFR_PANEL	African American	40	IG	0.200	0.600	0.200	0.371	0.500	0.500
	AFD_CHN_PANEL	Asian	36	IG	0.111	0.278	0.611	0.273	0.250	0.750
ss5212170	HapMap-CEU	European	114	IG	0.333	0.614	0.053	0.020	0.640	0.360
	HapMap-HCB	Asian	90	IG	0.044	0.422	0.533	0.479	0.256	0.744
	HapMap-JPT	Asian	88	IG	0.159	0.500	0.341	1.000	0.409	0.591
	HapMap-YRI	Sub-Saharan African	118	IG	0.220	0.508	0.271	1.000	0.475	0.525

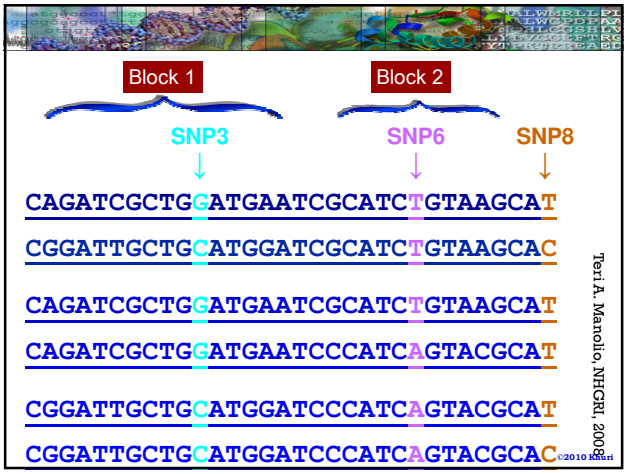
Summary	Average Het.+/- std err:	Individual Count	Founders Count	Individual Overlap	Genotype Conflict
	0.497+/-0.040	1263	1052	15	0

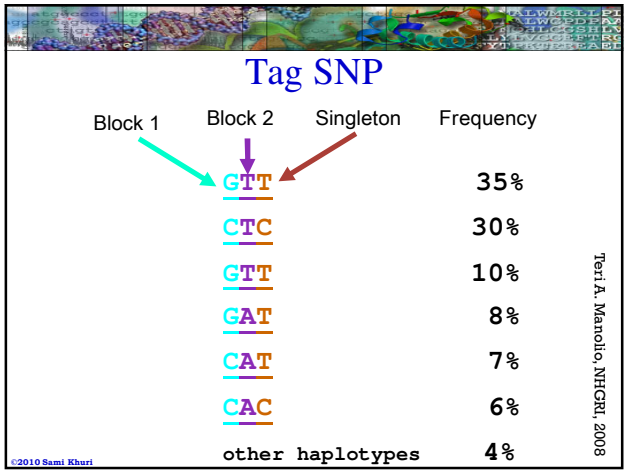
International Haplotype Map Project

- Goal of International Haplotype Map Project
  - Develop a haplotype map of the human genome.
- The "HapMap" describes common patterns of human DNA sequence variation
  - key source for researchers to use to find genes affecting health, disease, and responses to drugs, and environmental factors.
- Haplotypes are groups of SNPs transmitted in "blocks".
- These blocks can be characterized by a subset of their SNPs (tags).









Multi-locus SNP Profiles

- There will be a few hundred to a few thousand SNPs linked to medically important alleles in the next ~10 years.
- Haplotypes will reduce the number that need to be screened (one SNP gives information about a group of linked genes).
- Some genes will turn out to be involved in many important pathways.

©2010 Sami Khuri

GWAS: The New Wave

©2010 Sami Khuri

Hokusai: *The Great Wave*

Genome-Wide Association Study

- Method for interrogating all 10 million variable points across human genome.
- Variation is inherited in groups, or blocks, so not all 10 million points have to be tested.
- NIH is interested in advancing **genome-wide association studies** (GWAS) to identify common genetic factors that influence health and disease.

Teri A. Manolio, NHGRI, 2008

GWAS at NIH (I)

- NIH is interested in advancing **genome-wide association studies** (GWAS) to identify common genetic factors that influence health and disease.
- A **genome-wide association study** is defined as any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition.

GWAS at NIH (II)

- Whole genome information, when combined with clinical and other phenotype data, offers the potential for increased understanding of basic biological processes affecting human health, improvement in the prediction of disease and patient care, and ultimately the realization of the promise of **personalized medicine**.

GWAS at NIH (III)

- Rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying **genetic variants** that contribute to **health** and **disease**.

Testing 10 Million SNPs?

- Would Genome-Wide Association Studies require directly testing each of the nearly 10 million common variants for association to disease?
  - In other words, if only 5% of variants were tested, would 95% of associations be missed?
- Or could a subset serve as reliable proxies for their neighbors?

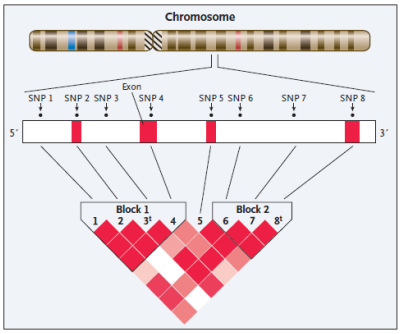
Genetic Mapping in Human Disease, Altshuler et al., 2008

Low Recombination Rates

- Each disease-causing mutation arises on a particular copy of the human genome and bears a specific set of common alleles in cis at nearby loci, termed a haplotype.
- Because the recombination rate is low (about 1 crossover per 100 megabases (Mb) per generation), disease alleles in the population typically show association with nearby marker alleles for many generations, a phenomenon termed linkage disequilibrium (LD)

Genetic Mapping in Human Disease, Altshuler et al., 2008

Mapping Relationships Among SNPs



Christensen and Murray, 2007

Distances Among East Coast Cities

	Boston	Provi-dence	New York	Phila-delphia	Balti-more
Providence	59				
New York	210	152			
Philadelphia	320	237	86		
Baltimore	430	325	173	87	
Washington	450	358	206	120	34

Color Coded Distances

	Boston	Provi-dence	New York	Phila-delphia	Balti-more
Providence	59				
New York	210	152			
Philadelphia	320	237	86		
Baltimore	430	325	173	87	
Washington	450	358	206	120	34

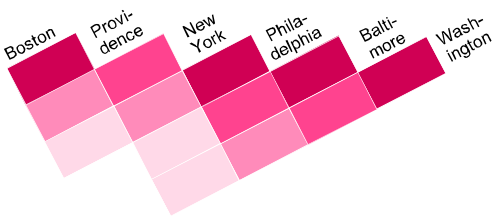
< 100 | 101-200 | 201-300 | 301-400 | > 400

Color Coded Distances (II)

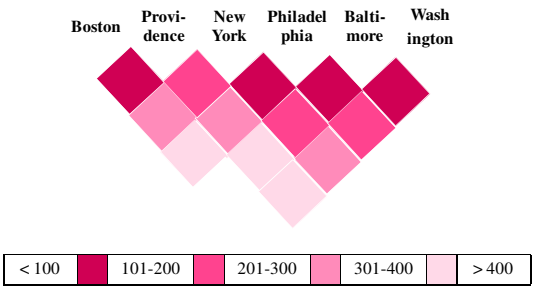
	Boston	Provi-dence	New York	Phila-delphia	Balti-more
Providence	59				
New York	210	152			
Philadelphia	320	237	86		
Baltimore	430	325	173	87	
Washington	450	358	206	120	34

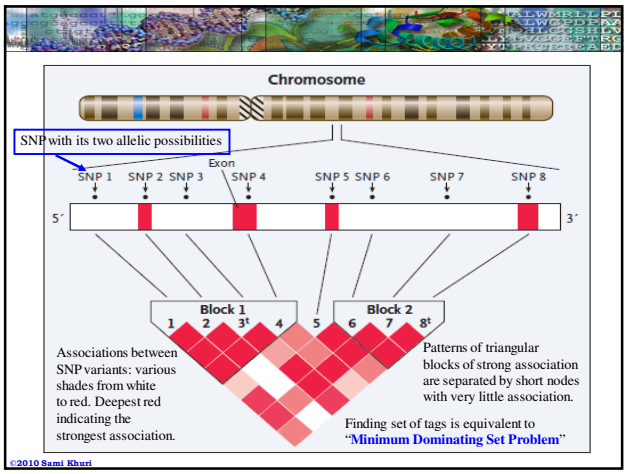
< 100 | 101-200 | 201-300 | 301-400 | > 400

Color Coded Distances (III)

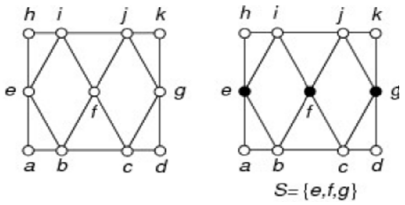


Color Coded Distances (IV)





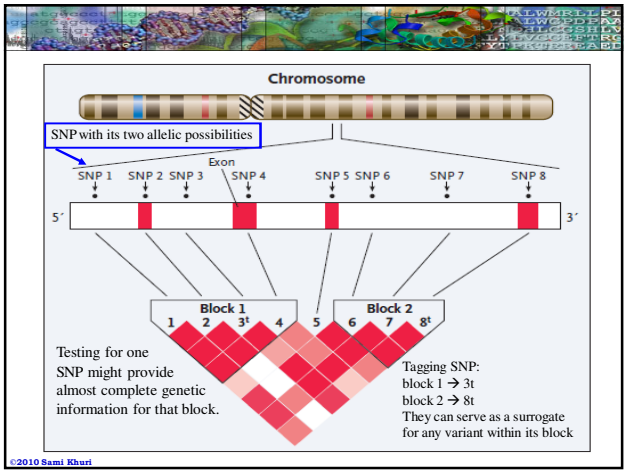
Minimum Dominating Set Problem



A **dominating set** is a set of nodes S such that every node in the network graph G is a neighbor of at least one element of S. The **Minimum Dominating Set** (MDS) problem is to find a minimum such S for a given network graph.

http://www.cs.iastate.edu/~chaudhur/cs611/Sp07/notes/lec22.pdf

©2010 Sami Khuri



Myocardial Infarction and rs1333049

Association of **alleles** of rs1333049 with Myocardial Infarction

	C N (%)	G N (%)	χ^2 (1df)	P-value
Cases	2,132 (55.4)	1,716 (44.6)	55.1	1.2×10^{-13}
Controls	2,783 (47.4)	3,089 (52.6)		

Allelic Odds Ratio = 1.38

Association of **genotypes** of rs1333049 with Myocardial Infarction

	CC N (%)	CG N (%)	GG N (%)	χ^2 (2df)	P-value
Cases	586 (30.5)	960 (49.9)	378 (19.6)	59.7	1.1×10^{-14}
Controls	676 (23.0)	1,431 (48.7)	829 (28.2)		

Heterozygote Odds Ratio = 1.47
Homozygote Odds Ratio = 1.90

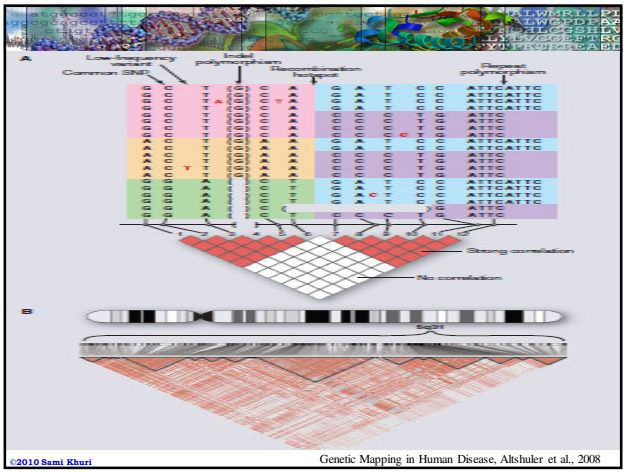
Genome-wide Association Analysis of Coronary Artery Disease, by Samani et al, *NEJM* 2007

©2010 Sami Khuri

Common Disease Common Variant Hypothesis

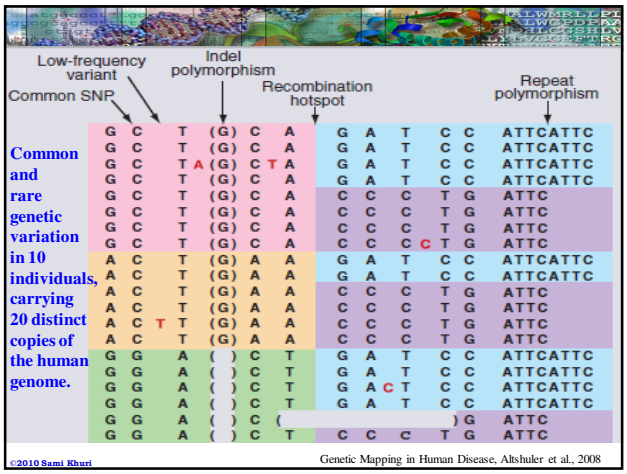
- It is believed that genetic variations with alleles that are common in the population will explain much of the heritability of common diseases.
- These studies were made possible by
  - the sequencing of the human genome (International Human Genome Sequencing Consortium, 2004) and
  - the completion of the subsequent human haplotype mapping (HapMap) project.

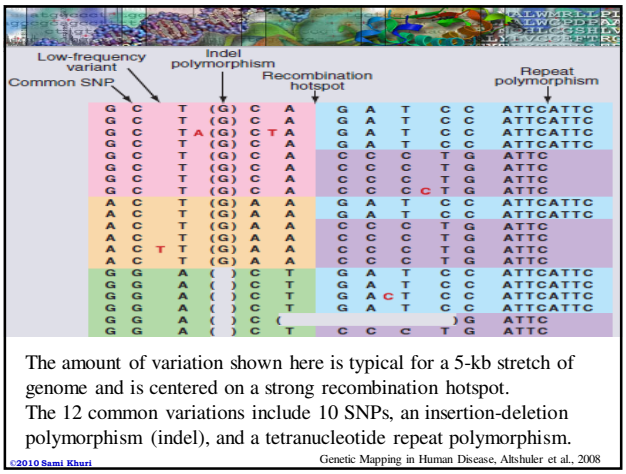
©2010 Sami Khuri

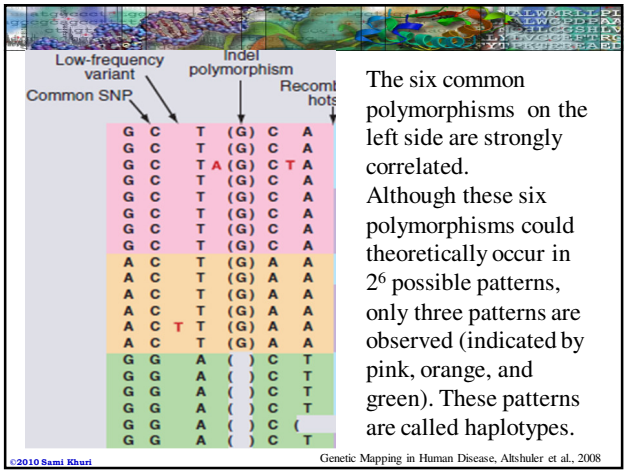


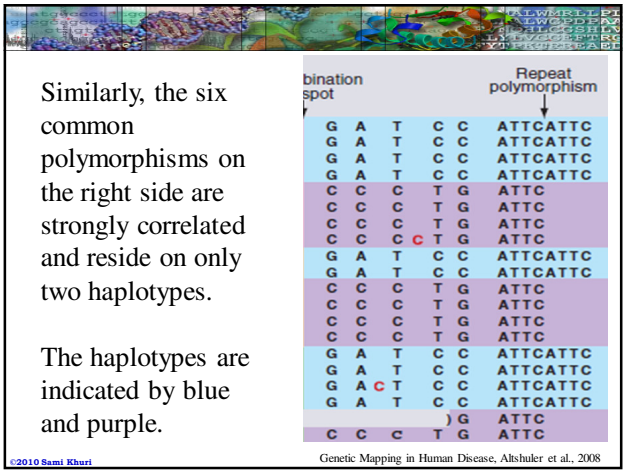
Genetic Mapping in Human Disease, Altshuler et al., 2008

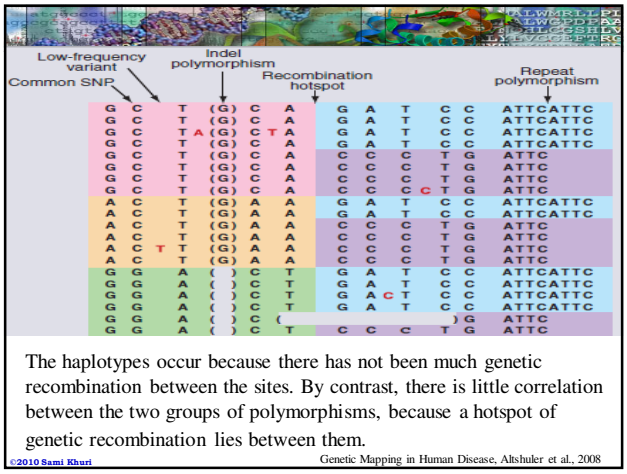
©2010 Sami Khuri

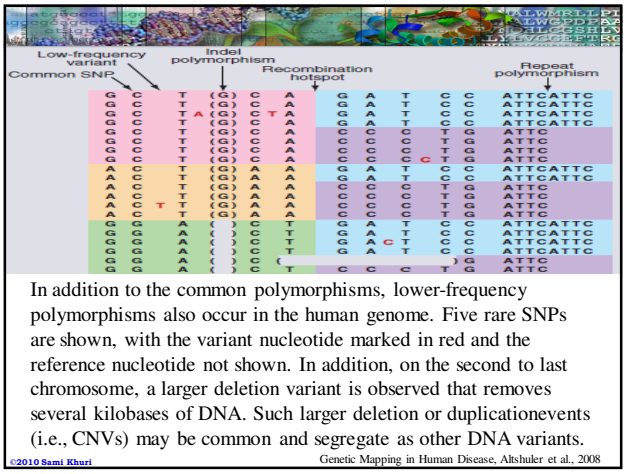

Low-frequency variant
Common SNP
Indel polymorphism
Recombination hotspot
Repeat polymorphism
Common and rare genetic variation in 10 individuals, carrying 20 distinct copies of the human genome.
©2010 Sami Khuri
Genetic Mapping in Human Disease, Altshuler et al., 2008

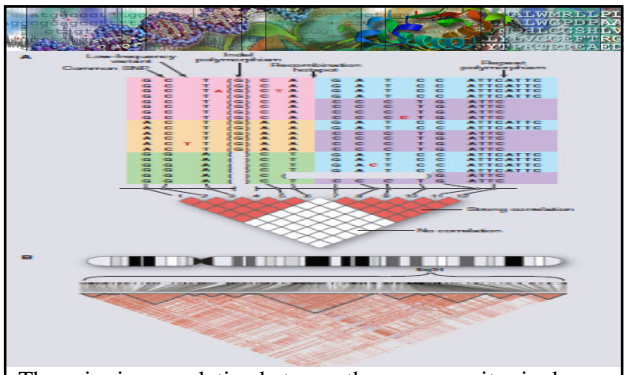

Low-frequency variant
Common SNP
Indel polymorphism
Recombination hotspot
Repeat polymorphism
The amount of variation shown here is typical for a 5-kb stretch of genome and is centered on a strong recombination hotspot.
The 12 common variations include 10 SNPs, an insertion-deletion polymorphism (indel), and a tetranucleotide repeat polymorphism.
©2010 Sami Khuri
Genetic Mapping in Human Disease, Altshuler et al., 2008


Low-frequency variant
Common SNP
Indel polymorphism
The six common polymorphisms on the left side are strongly correlated.
Although these six polymorphisms could theoretically occur in 2^6 possible patterns, only three patterns are observed (indicated by pink, orange, and green). These patterns are called haplotypes.
©2010 Sami Khuri
Genetic Mapping in Human Disease, Altshuler et al., 2008


Repeat polymorphism
Similarly, the six common polymorphisms on the right side are strongly correlated and reside on only two haplotypes.
The haplotypes are indicated by blue and purple.
©2010 Sami Khuri
Genetic Mapping in Human Disease, Altshuler et al., 2008


Low-frequency variant
Common SNP
Indel polymorphism
Recombination hotspot
Repeat polymorphism
The haplotypes occur because there has not been much genetic recombination between the sites. By contrast, there is little correlation between the two groups of polymorphisms, because a hotspot of genetic recombination lies between them.
©2010 Sami Khuri
Genetic Mapping in Human Disease, Altshuler et al., 2008


Low-frequency variant
Common SNP
Indel polymorphism
Recombination hotspot
Repeat polymorphism
In addition to the common polymorphisms, lower-frequency polymorphisms also occur in the human genome. Five rare SNPs are shown, with the variant nucleotide marked in red and the reference nucleotide not shown. In addition, on the second to last chromosome, a larger deletion variant is observed that removes several kilobases of DNA. Such larger deletion or duplicationevents (i.e., CNVs) may be common and segregate as other DNA variants.
©2010 Sami Khuri
Genetic Mapping in Human Disease, Altshuler et al., 2008



The pairwise correlation between the common sites is shown by the red and white boxes below, with red indicating strong correlation and white indicating weak correlation.

Genetic Mapping in Human Disease, Altshuler et al., 2008

MDECODE

- **M**olecular **D**iversity and **E**pidemiology of **Co**mmon **D**isease (**MDECODE**) is a multidisciplinary and multinational project created to gain a greater understanding of the type and amount of human DNA sequence variation, its history, and the relationship of its contemporary organization to the continuous distribution of measures of human health among individuals in the population at large (such as blood pressure or plasma cholesterol levels).

http://droog.mbt.washington.edu/mdecode

Understanding DNA Variations (I)

- An important goal of human genetics and genetic epidemiology is to understand the mapping relationship between interindividual variation in DNA sequences, variation in environmental exposure and variation in disease susceptibility.

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

Understanding DNA Variations (II)

- Stated another way, how do one or more changes in an individual's DNA sequence increase or decrease their risk of developing disease through complex networks of biomolecules that are hierarchically organized, highly interactive and dependent on environmental exposures?

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

Understanding DNA Variations (III)

- Understanding the role of genomic variation and environmental context in disease susceptibility is likely to improve diagnosis, prevention and treatment.

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

The Importance of Non-Linearity (I)

- Success in this important public health endeavor will depend critically on the amount of **non-linearity** in the mapping of genotype to phenotype and our ability to address it.
- An outcome is **non-linear** if it cannot be easily predicted by the sum of the individual genetic markers.

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

The Importance of Non-Linearity (II)

- **Non-linearity** can arise from phenomena such as:
  - **locus heterogeneity** (i.e. different DNA sequence variations leading to the same phenotype),
  - **phenocopy** (i.e. environmentally determined phenotypes that do not have a genetic basis)

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

The Importance of Non-Linearity (III)

- the dependence of genotypic effects on **environmental exposure** (i.e. gene–environment interactions or plastic reaction norms), and
- genotypes at other loci (i.e. **gene–gene interactions** or epistasis).

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

Overcoming Three Challenges (I)

- Three significant challenges that must be overcome if we are to successfully identify those genetic variations that are associated with health and disease using a genome-wide approach.

  1) Powerful **data mining** and **machine learning** methods will need to be developed to computationally model:
  - the relationship between combinations of SNPs,
  - other genetic variations, and
  - environmental exposure with disease susceptibility.

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

Overcoming Three Challenges (II)

2) Accurate and powerful **selection methods** will have to be developed to determine which **subset of SNPs** should be included in the analysis.

- If non-linear interactions between genes explain a significant proportion of the heritability of common diseases, then combinations of SNPs will need to be evaluated from a list of thousands or millions of candidates.
- **Filtering algorithms** and/or **stochastic search** or **wrapper algorithms** will play an important role in GWAS because there are more combinations of SNPs to examine than can be exhaustively evaluated using modern computational horsepower.

"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

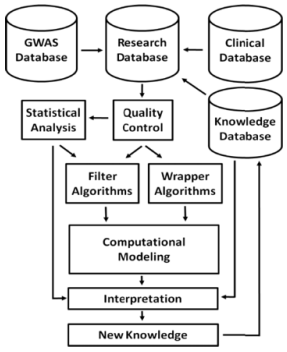
Overcoming Three Challenges (III)

3) Correct **biological interpretation** of non-linear genetic models has to be achieved.

- Even when a computational model can be used to identify SNPs with genotypes that increase susceptibility to disease, the specifics of the mathematical relationships cannot be translated into prevention and treatment strategies without **interpreting the results** in the context of human biology.
- Making **etiological inferences** from computational models may be the most important and the most difficult challenge of all.

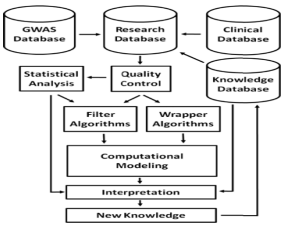
"Bioinformatics challenges for genome-wide association studies" by Moore et al., 2010

Bioinformatics GWAS Pipeline (I)



[MOO10]

Bioinformatics GWAS Pipeline (II)



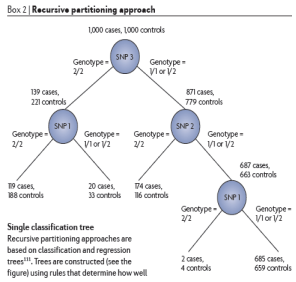
Flowchart for bioinformatics analyses of GWAS data. The use of filter and wrapper algorithms along with computational modeling approaches is recommended in addition to parametric statistical methods. Biological knowledge in public databases has a very important role to play at all levels of the analysis and interpretation. [MOO10]

Data Mining and Machine Learning

- **Data mining** and **machine learning** methods:
  - Will reveal numerous significant interactions and other complex **genotype–phenotype relationships** when they are widely applied to GWAS data
  - Are much more consistent with the idea of letting the data tell us what the model is rather than forcing the data to fit a preconceived notion of what a good model is.

[MOO10]

Methods Used for SNP Analysis



- Regression Models
- Exhaustive Searches
  - Two-Locus Interaction
  - Higher-Order Interaction
- Recursive Partitioning Approach
- Random Forest Approach
- Multifactor Dimensionality Reduction
- Bayesian model selection

Detecting gene-gene interactions that underlie human diseases" by Cordell, Nature, June 2009

Computational Modeling using Decision Trees and Random Forests

- What are Decision Trees?
- An example of a Decision Tree
- What is a Random Forest (RF)?
- How are individual Decision Trees in RF constructed?
- Advantages of Random Forests
- Limitations of Random Forests

Decision Trees

- A **Decision Tree** classifies subjects as case or control by sorting them through a tree from node to node where each node is an **attribute** (example: SNP) with a decision rule that guides that subject through different branches of the tree to a leaf that provides its **classification** (case or control).
- **Decision Trees** are widely used for **modeling** the relationship between one or more attributes and a discrete end point, such as the case-control status.

Advantages of Decision Trees

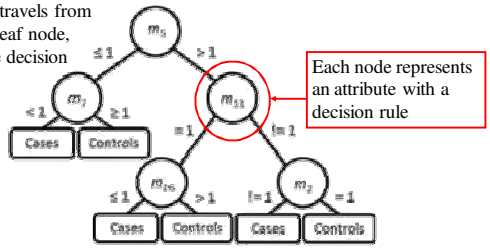
- Advantages of Decision Tree:

  The tree is simple to **visualize** and can be interpreted as a series of IF-Then rules.
- Additional nodes or attributes below the root node allows **hierarchical dependencies** to be modeled.

Decision Tree Classifiers

Decision tree classifies subjects as case or control.

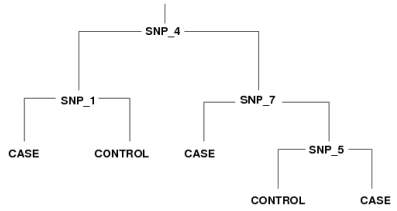
Each subject travels from root node to leaf node, guided by the decision rules.



Leaves of tree represent case or control that provide classification for subject.

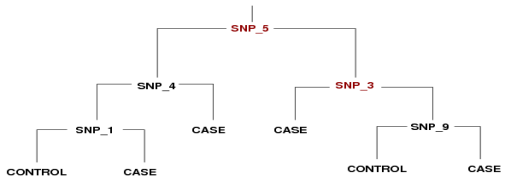
Decision Trees Revisited

- Decision trees offer a series of rules by which samples may be **classified** by SNPs and other variables (attributes).



Random Forests to the Rescue

- Building a **decision tree** is a greedy process by which the best attribute at dividing the classes is chosen at each step.
- However, this may not find the best solution -- the best decision may be based on a combination of attributes.



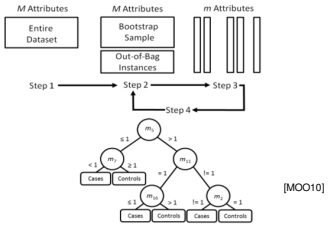
- Solution: Random Forests...

Random Forests

- **Random Forests** extend decision trees for the analysis of more complex data.
- A **Random Forest** is a collection of individual decision tree classifiers, where each tree in the forest has been trained using bootstrap sample of instances from the data, and each attribute in the tree is chosen from among a random subset of attributes (Breiman, 2001).

Random Forests

- **Random Forests** consist of many (100s ~ 1000s) decision trees that are built using subsets of the attributes and data.
- This gives different **combinations of attributes** a chance.



[MOO10]

- The trees in the forest then vote on the best classification of a sample.

Steps for Building Random Forests

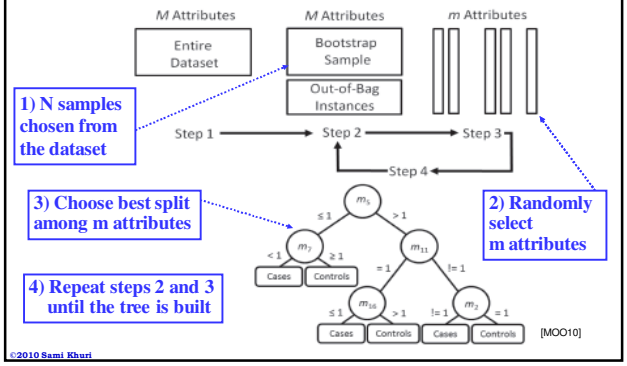
- Steps to construct individual trees from data having N samples and M attributes :

1. Choose a training set by selecting N samples, with replacement, from the data.
2. At each node in the tree, randomly select m attributes from the entire set of M attributes in the data.
3. Choose the best split at that node from among the m attributes.
4. Iterate the second and third steps until the tree is fully grown.

- Repetitions of the algorithm yields a forest of decision trees.

[MOO10]

Random Forest Algorithm



[MOO10]

Advantages of Random Forests (I)

- Advantages of the **Random Forest** approach is that the final decision tree models may uncover interactions among genes and/or environmental factors that do not exhibit strong marginal effects.
- **Random Forests** capitalize on the benefits of decision trees and have been shown excellent predictive performance when the forest is diverse.
- It has also been shown that **Random Forests** are robust in the presence of noisy or potential false positive SNPs.

[MOO10]

Advantages of Random Forests (II)

- **Random Forest** are often used initially for selecting the subset of attributes.
- It has been shown that **Random Forests** have outperformed traditional methods, such as the Fisher's exact test when the 'risk' SNPs interact. Lunetta *et al.* (2004)
  - This study revealed that the relative superiority of the **Random Forest** method increases as more **interacting SNPs** are added to the model.

[MOO10]

Applications of Random Forests

- **Random Forest** have been applied to genetic data in studies of:
  - Asthma (Bureau et al., 2005)
  - Rheumatoid arthritis (Sun et al., 2007)
  - Glioblastoma (Chang et al, 2008)
  - Age-related macular degeneration (Jiang et al., 2009)
  - Vaccination response (McKinney et al., 2009)
- It is expected that random forests will prove to be a useful tool for **detecting gene-gene interaction**.

[MOO10]

Computational Modeling using Multifactor Dimensionality Reduction

- **Multifactor Dimensionality Reduction** (**MDR**) was developed as a **non-parametric** (i.e. no parameters are estimated) and **genetic model-free** (i.e. no genetic model is assumed) **data mining** and **machine learning** strategy for **identifying combinations of discrete genetic and environmental factors** that are predictive of a discrete clinical end point. (Hahn *et al., 2003*)

[MOO10]

Multifactor Dimensionality Reduction

- **MDR** was designed to detect interactions in the absence of detectable marginal effects and thus complements statistical approaches such as logistic regression and machine learning methods such as random forests and neural networks.
- At the heart of the **MDR** approach is a feature or attribute construction algorithm that creates a new variable or attribute by pooling genotypes from multiple SNPs (Moore and White, 2006).

[MOO10]

Constructive Induction

- **MDR** uses **Constructive Induction** (aka Attribute Construction) were a new attribute is defined as a function of two or more other attributes.
- The **MDR** method is based on the idea that **changing** the **representation space** of the data will make it easier for methods such as logistic regression, classification trees or a naive Bayes classifier to **detect attribute dependencies**. [MOO10]

Modifications of MDR (I)

- Many modifications and extensions to **MDR** have been proposed. These include
  - Entropy-based interpretation methods (Moore and White, 2006),
  - The use of odds ratios (Chung *et al., 2007),*
  - Log-linear methods (Lee *et al., 2007),*
  - Generalized linear models *(Lou et al., 2007),*
  - Methods for imbalanced data (Velez *et al., 2007).*

[MOO10]

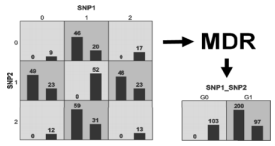
Modifications of MDR (II)

- Permutation testing methods (Greene *et al., 2010a; Pattin et al., 2009),*
- Methods for dealing with missing data (Namkung *et al., 2009a),*
- Model-based methods (Calle *et al., 2008),*
- Parallel implementations *(Bush et al., 2006; Sinnott Arnstrong et al., 2009),* and
- Different evaluation metrics (Bush *et al., 2008; Mei et al., 2007; Namkung et al.,* 2009b).

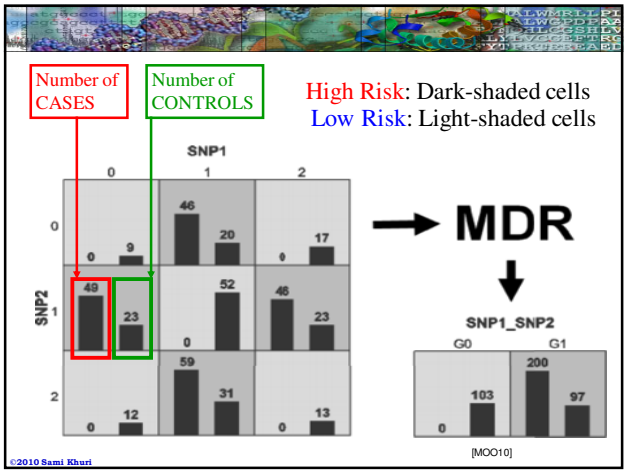
[MOO10]

Multifactor Dimensionality Reduction

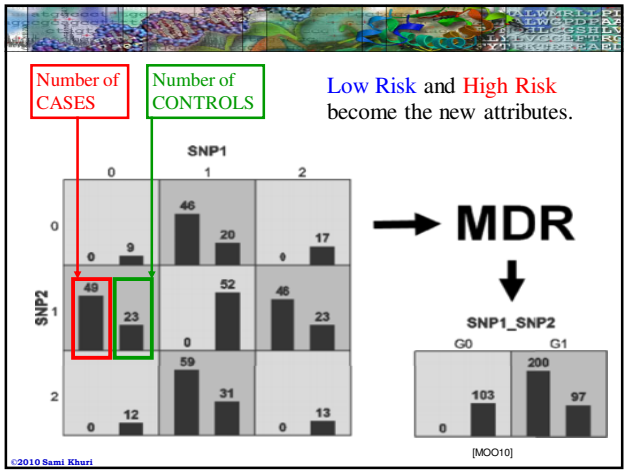
- Creates combinations of attributes to decrease their overall number.
- Attributes are grouped into low risk or high risk based on the ratio of their occurrences in disease cases to control cases.
- Low Risk and High Risk become the new attributes. Statistical analysis are performed on those new attributes.



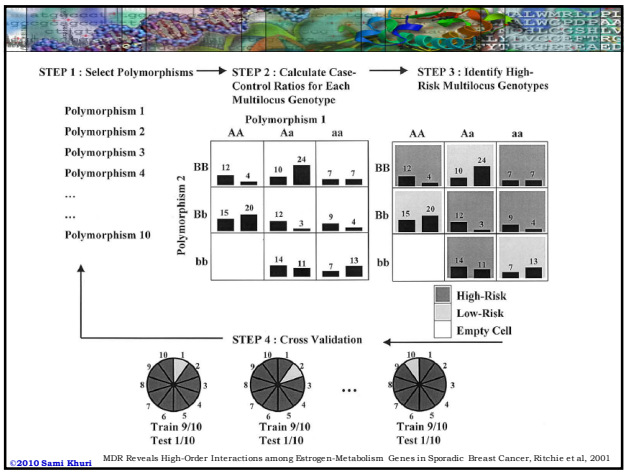
[MOO10]

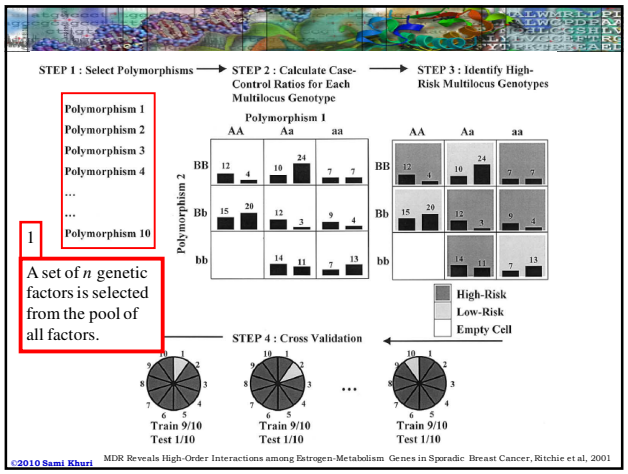


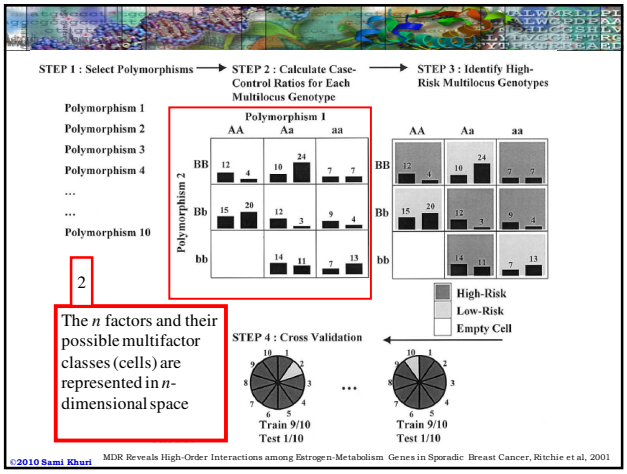
[MOO10]

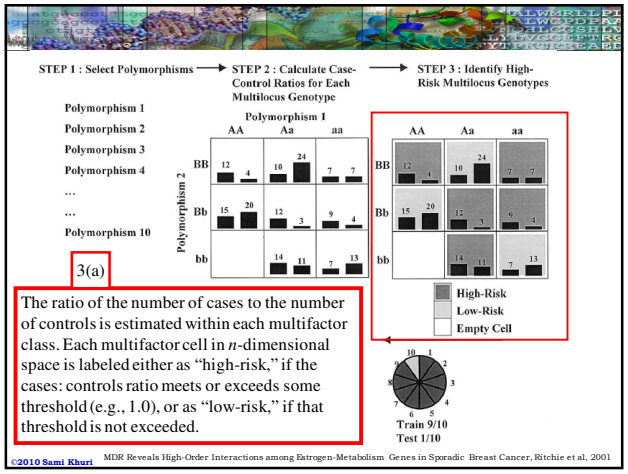


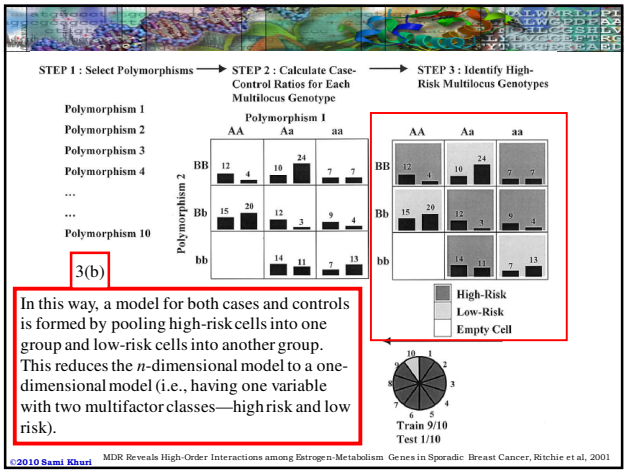
[MOO10]

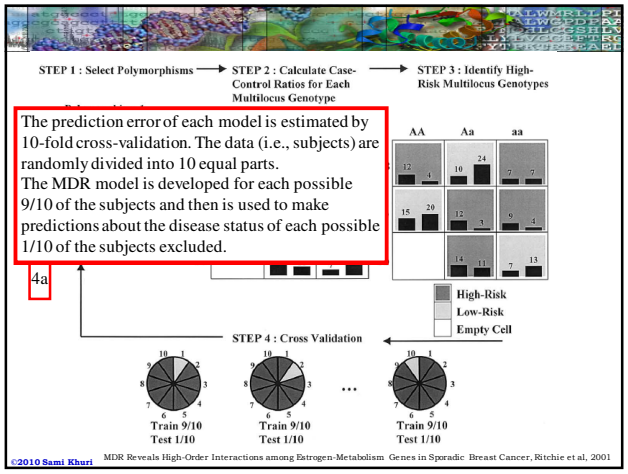

STEP 1 : Select Polymorphisms
STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype
STEP 3 : Identify High-Risk Multilocus Genotypes
Polymorphism 1
Polymorphism 2
Polymorphism 3
Polymorphism 4
...
...
Polymorphism 10
AA Aa aa
BB Bb bb
High-Risk
Low-Risk
Empty Cell
STEP 4 : Cross Validation
Train 9/10 Test 1/10
©2010 Sami Khuri
MDR Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, Ritchie et al, 2001

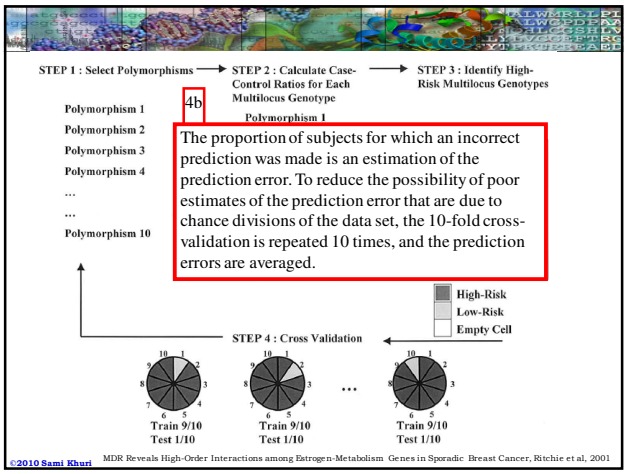

STEP 1 : Select Polymorphisms
STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype
STEP 3 : Identify High-Risk Multilocus Genotypes
1
A set of n genetic factors is selected from the pool of all factors.
STEP 4 : Cross Validation
Train 9/10 Test 1/10
©2010 Sami Khuri
MDR Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, Ritchie et al, 2001


STEP 1 : Select Polymorphisms
STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype
STEP 3 : Identify High-Risk Multilocus Genotypes
2
The n factors and their possible multifactor classes (cells) are represented in n-dimensional space
STEP 4 : Cross Validation
Train 9/10 Test 1/10
©2010 Sami Khuri
MDR Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, Ritchie et al, 2001


STEP 1 : Select Polymorphisms
STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype
STEP 3 : Identify High-Risk Multilocus Genotypes
3(a)
The ratio of the number of cases to the number of controls is estimated within each multifactor class. Each multifactor cell in n-dimensional space is labeled either as "high-risk," if the cases: controls ratio meets or exceeds some threshold (e.g., 1.0), or as "low-risk," if that threshold is not exceeded.
Train 9/10 Test 1/10
©2010 Sami Khuri
MDR Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, Ritchie et al, 2001


STEP 1 : Select Polymorphisms
STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype
STEP 3 : Identify High-Risk Multilocus Genotypes
3(b)
In this way, a model for both cases and controls is formed by pooling high-risk cells into one group and low-risk cells into another group. This reduces the n-dimensional model to a one-dimensional model (i.e., having one variable with two multifactor classes—high risk and low risk).
Train 9/10 Test 1/10
©2010 Sami Khuri
MDR Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, Ritchie et al, 2001


STEP 1 : Select Polymorphisms
STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype
STEP 3 : Identify High-Risk Multilocus Genotypes
The prediction error of each model is estimated by 10-fold cross-validation. The data (i.e., subjects) are randomly divided into 10 equal parts.
The MDR model is developed for each possible 9/10 of the subjects and then is used to make predictions about the disease status of each possible 1/10 of the subjects excluded.
4a
STEP 4 : Cross Validation
Train 9/10 Test 1/10
©2010 Sami Khuri
MDR Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer, Ritchie et al, 2001



The Attribute Selection Challenge

- It is now commonly assumed that at least **one million carefully selected SNPs** are necessary to capture much of the relevant variation across the human genome.
- With this many attributes, the number of higher order combinations is astronomical.
- What is the optimal computational approach to this problem?

[MOO10]

Selecting Attributes for Predictive Models

- There are two general approaches to selecting attributes for predictive models.
- The **filter** approach preprocesses the data by algorithmically assessing the quality or relevance of each variable and then using that information to **select a subset** for analysis.
- The **wrapper** approach **iteratively selects subsets** of attributes for classification using either a deterministic or stochastic algorithm.

[MOO10]

Selecting Attributes: Filtering Algorithms (I)

- It is computationally infeasible to combinatorially explore all high-order interactions among the SNPs in a genome-wide association study.
- A standard statistical strategy in human genetics is to assess the quality of each SNP using a chi-square test of independence followed by a correction of the significance level that takes into account an increased false positive rate due to multiple tests.

[MOO10]

Selecting Attributes: Filtering Algorithms (II)

- This standard statistical strategy is a very efficient filtering method for assessing the independent effects of SNPs on disease susceptibility but it ignores the dependencies or interactions between genes.
- Several filtering algorithms have been devised to solve the gene-gene interaction issue.

[MOO10]

Filtering Algorithms: Relief Family of Algorithms

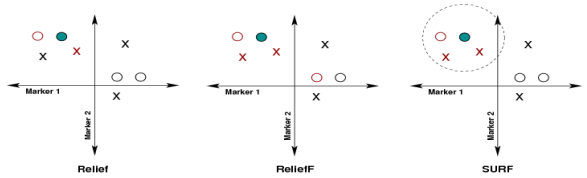
1) Randomly choose a **test sample** from the data set.
2) Find the best matching case and control for that test sample. These are the *nearest neighbors*.
3) Compare the attributes from the test sample to the nearest neighbors to determine quality estimates of those attributes.

– Repeat steps 1-3 for another test sample.

[MOO10]

Relief Variants

- ReliefF: Instead of choosing just one nearest case and control neighbor, use the *n* nearest neighbors.
- SURF (Spatially Uniform ReliefF): Put an upper bound on the distance between the test sample as a nearest neighbor.
- TuRF (Tuned ReliefF): After each round, remove low scoring attributes, so future neighbor-matching will not rely on them.



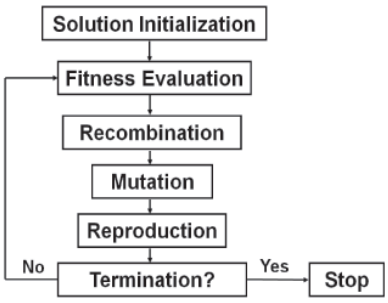
[MOO10]

Wrapping: Genetic Programming

– A collection of programs consisting of lists of SNPs, other attributes, and mathematical functions describing them are randomly generated.
– The programs are evaluated automatically. High scoring programs are recombined, crossbred and mutated.
– Step 2 repeats until some threshold is reached. The result is the best found set of attributes and relations between them.
Results so far have been mixed. Finding good methods to evaluate the programs and intelligent ways of recombining programs is necessary.

[MOO10]

The Genetic Programming



[MOO10]

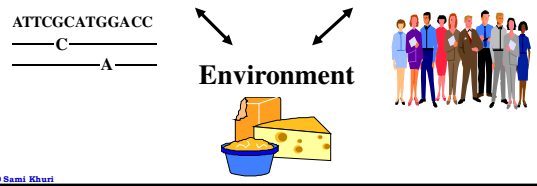
Bioinformatics Challenges for GWAS

- In the past GWAS looked for one to one association between SNP's and disease risk.
- Now need to start looking at gene-gene and gene-environment interaction when conducting GWAS.
- If we know the pathway interactions of a disease. Only look at SNPs in those pathways
- Get information on pathways from biological databases
- Quality of results is dependent on the quality of information in the database.

The Next Challenge

Understanding the link between -

DNA sequence (Genotype) ⟷ **Biology/Disease (Phenotype)**



The Next Wave of GWAS

- To date GWAS have identified a fraction of the genetic relative risk
  – Mostly focused on 'common disease, common variant' hypothesis
- **1000 Genomes Project** is a large sequencing project whose goal is to comprehensively catalogue rare variants
- **Copy number variants** are currently under-represented on products used in GWAS
- **Gene-Gene interactions**

Source: Keith W. Jones, Affymetrix