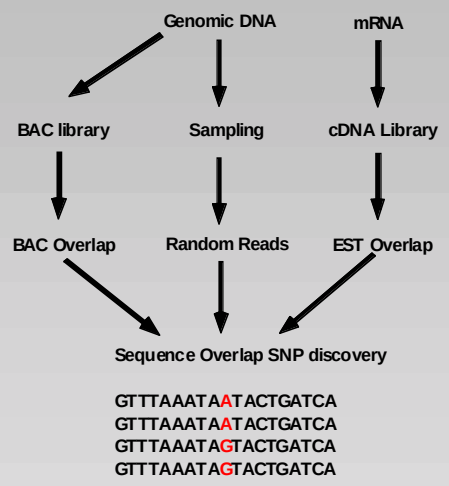
Single Nucleotide Polymorphisms (SNPs)

- How do we identify SNPs?
- How many are there?
- What is their relationship to each other?
- How do we use them?

Single Nucleotide Polymorphisms (SNPs)

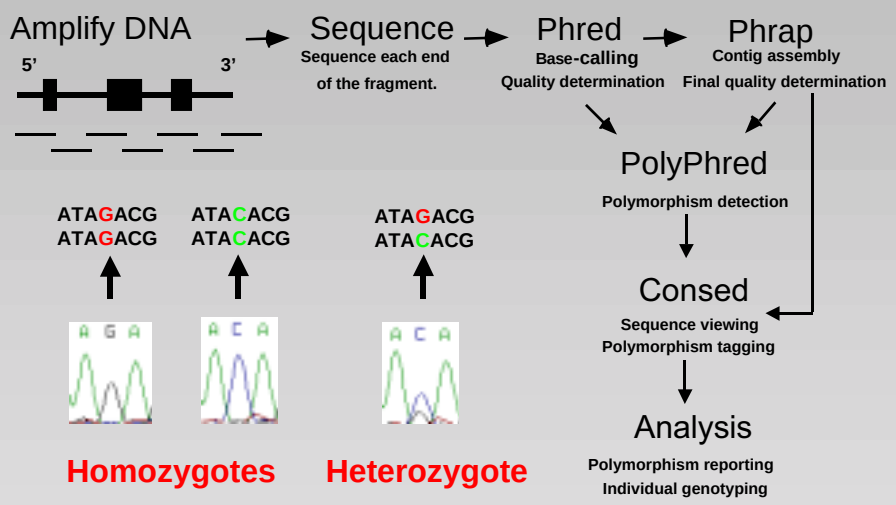
- Single base pair changes or deletions in DNA
- Diallelic
- Approximately 2,000,000-3,000,000 Single Nucleotide Polymorphisms (SNPs) between any two chromosomes (1 every 1,000 bp)
- Can be assessed efficiently with high-throughput genotyping methods

SNP Discovery Approaches

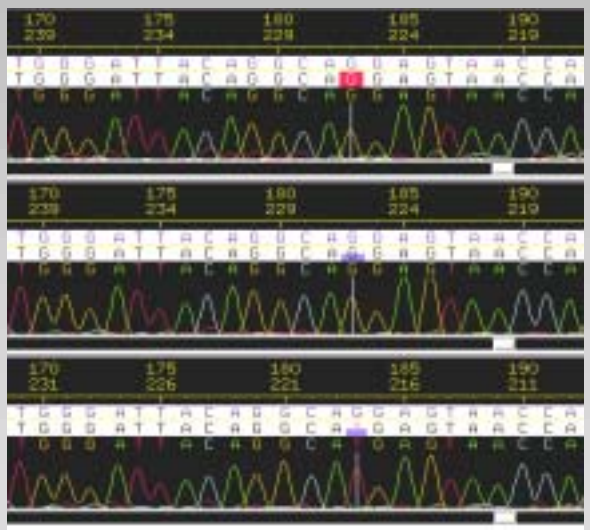


~2.7 Million Mapped SNPs in dbSNP: http://www.ncbi.nlm.gov/SNP/

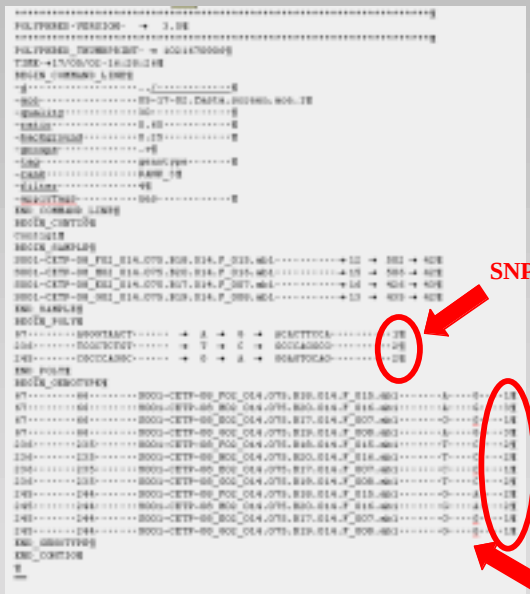
Targeted SNP Discovery



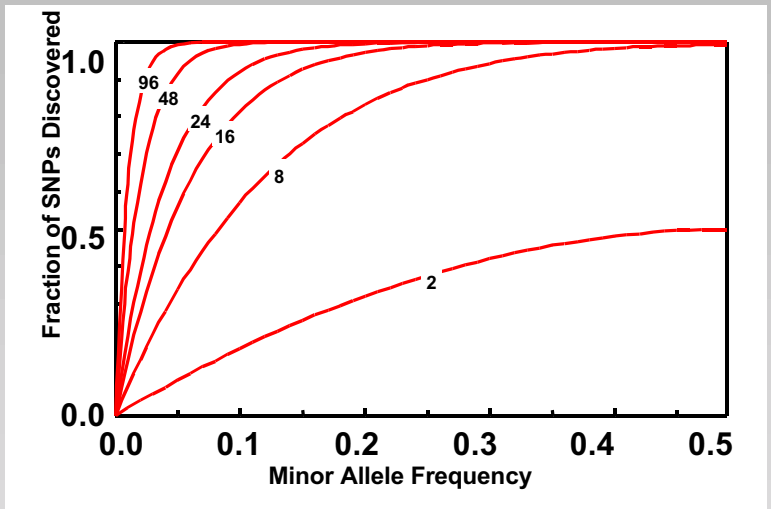
SNP Discovery Visualization (Consed)



SNP Discovery (Polyphred)



SNP Discovery Probability



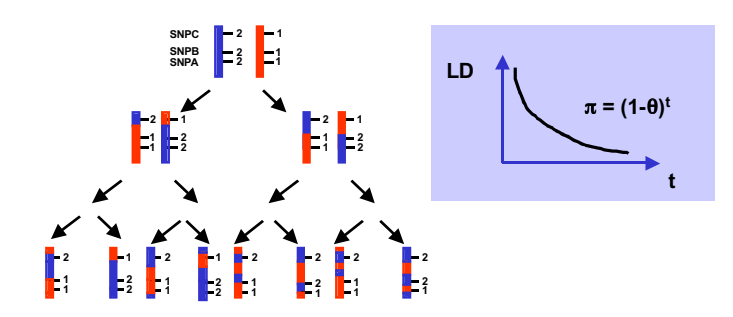
M. Eberle and L. Kruglyak, *Genet Epidemiol* 19 (Suppl 1):S29-S35 2000

SNP Representation in dbSNP

minimal allele frequency	expected SNPs (millions)	expected SNP frequency (bp)	expected % in database
1%	11.0	290	11-12
5%	7.1	450	15-17
10%	5.3	600	18-20
20%	3.3	960	21-25
30%	2.0	1570	23-27
40%	0.97	3280	24-28

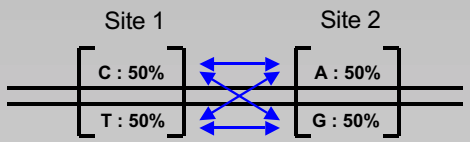
L. Kruglyak and D. Nickerson, *Nat Genet* 27:234-236 2001

Linkage Disequilibrium



- Measure of independence of SNPs from each other
- SNPs that are in complete linkage disequilibrium (LD) predict each other
- SNPs in equilibrium are completely independent from each other

LD and Haplotypes



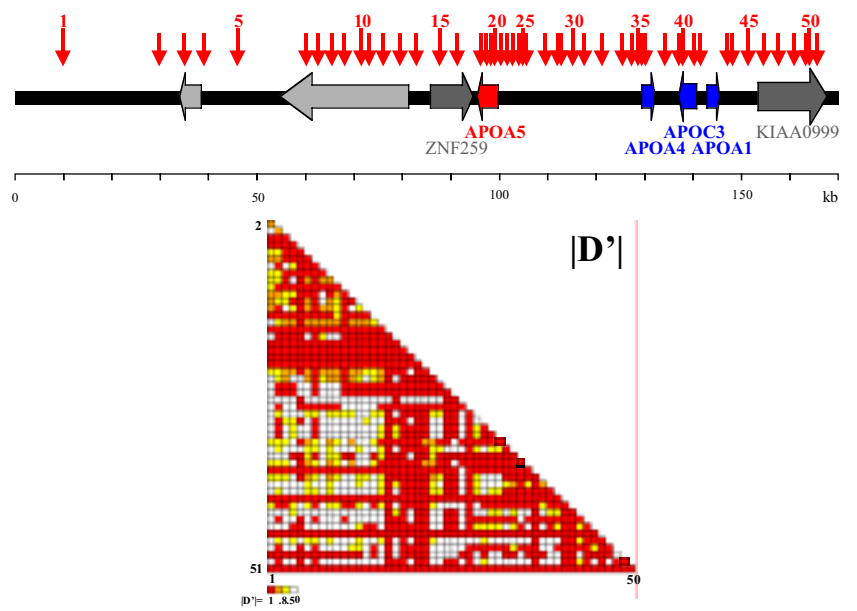
Parental Haplotypes

	Site 1	Site 2
Maternal	**C**	**A**
Paternal	**T**	**G**

Possible Site 1	Possible Site 2	Expected	Observed
C	A	0.5 X 0.5 = 0.25	0.50
C	G	0.5 X 0.5 = 0.25	0
T	A	0.5 X 0.5 = 0.25	0
T	G	0.5 X 0.5 = 0.25	0.50

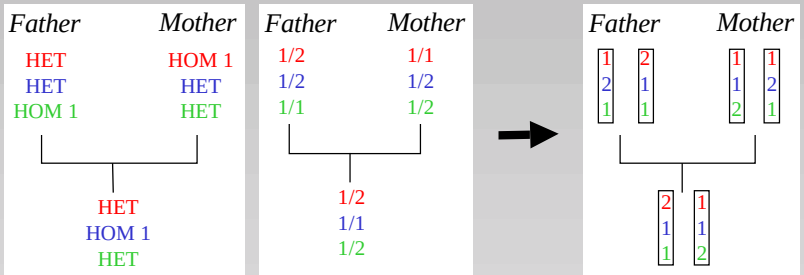
Sites in LD

Apolipoprotein Gene Cluster



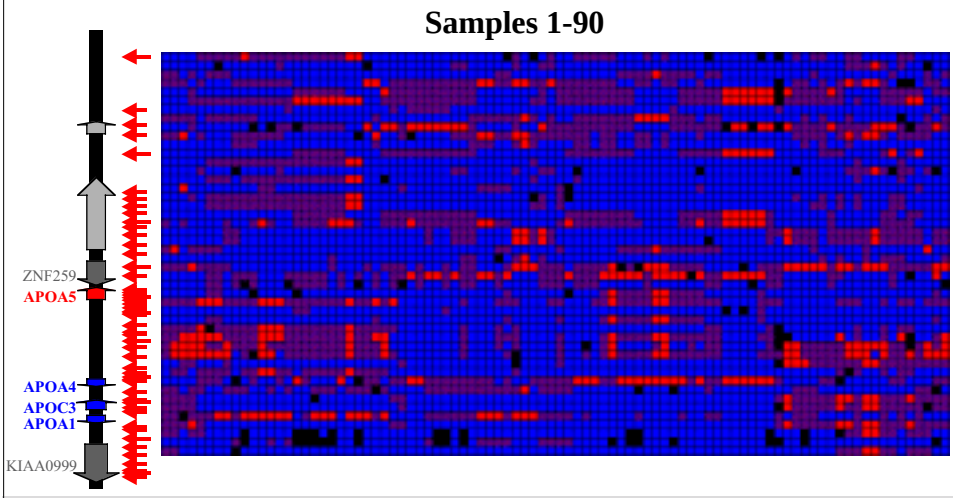
Haplotype Determination

- **If family information is available (2-3 generations), haplotypes can be determined directly:**

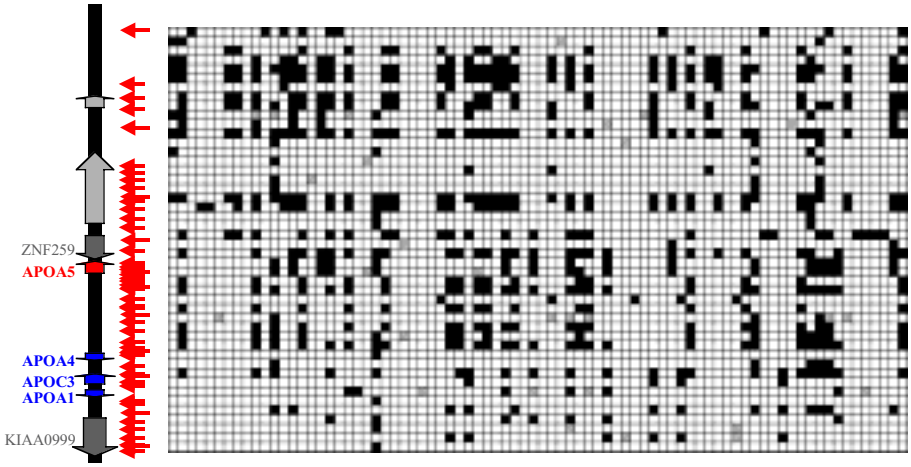


- **If samples are independent, mathematical models can predict the frequency of haplotypes (Expectation Maximization Algorithm)**

Apo Gene Cluster Genotypes



Apo Gene Cluster Haplotypes



Why Do We Need SNPs?

The Next Challenge

Understanding the link between

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

ATTCGCATGGACC

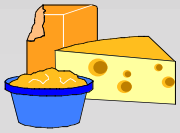
C

A

C G

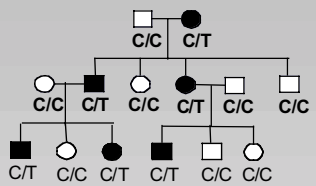
G A

Environment



Approaches for Genetic Analysis

Family Studies



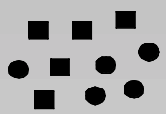
Single Gene

Mendelian Inheritance

Rare (High Penetrance)

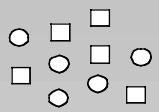
~300 Short Tandem Repeat Markers

Population Studies



40% T, 60% C

Cases



15% T, 85% C

Controls

Polygenic (Genes X Environment)

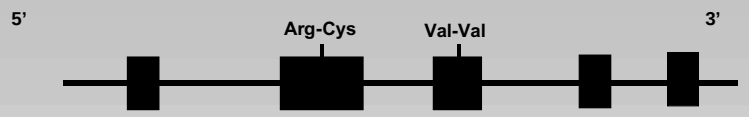
Complex Inheritance (Disease susceptibility)

Common

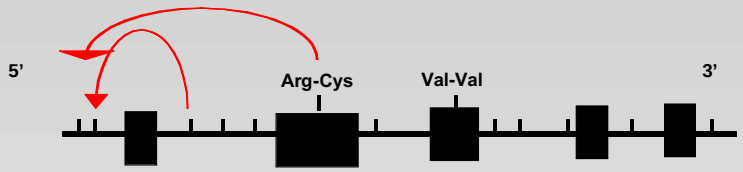
~30,000 to 500,000 or More Polymorphic Markers

SNP Association Studies

Direct: Catalog and test all functional SNPs

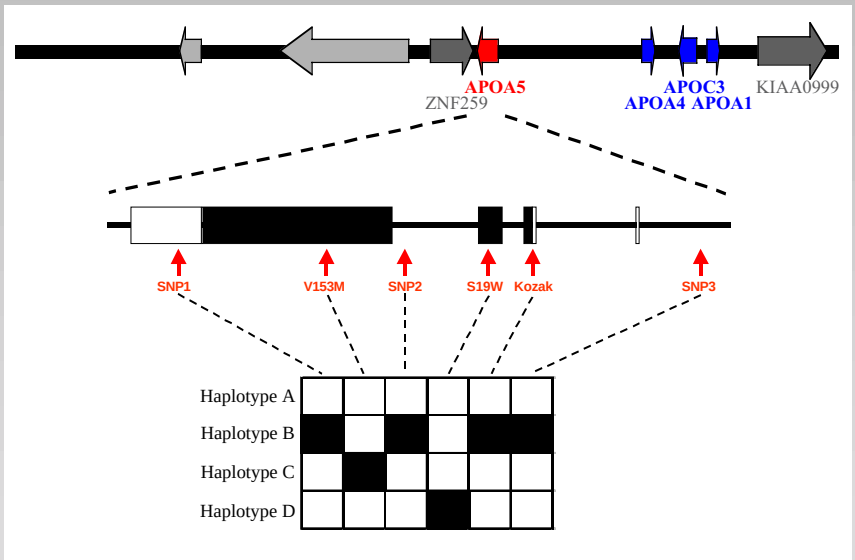


Indirect: Use dense map of SNPs and test for linkage disequilibrium (site association)



Collins, Guyer, Chakravarti *Science* 278:1580-81, 1997

ApoAV Region



Summary

- SNPs are abundant markers for genetic studies (>5,000,000 common SNPs)
- Analysis of linkage disequilibrium allows selection of informative subset of SNPs for genomic regions of interest
- Haplotypes across regions of LD can be used for disease association studies
- Genome-wide association studies will require identification of informative subsets of SNPs and cost-efficient genotyping methods

Tools and Links

- **phredPhrap, consed:** http://depts.washington.edu/ventures/uwtech/license/express/ppccombo.htm
- **Polyphred:** http://droog.mbt.washington.edu/PolyPhred.html
- **GENEHUNTER:** http://www-genome.wi.mit.edu/ftp/distribution/software/genehunter/
- **Arlequin (EM Algorithm):** http://lgb.unige.ch/arlequin/