Studying Genetic Variation I: Laboratory Techniques

Karen Mohlke, PhD
Department of Genetics
University of North Carolina

Human Genetic Variation

Variants contribute to rare and common diseases

Variants can be used to trace human origins

Origins of Variation

- **Mutations are produced by errors in DNA replication**
- **Errors in DNA replication during egg or sperm formation lead to new mutations**
- **Average 2.5 x 10^{-8} mutations per nucleotide site or 175 mutations per diploid genome per generation**

Human Genetic Variation

- **What types of variants exist?**
- **How are variants found?**
- **How are variants scored?**
- **How are variants used?**

Human Genetic Variation

- What types of variants exist?
- How are variants found?
- How are variants scored?
- How are variants used?

Human Genetic Variation

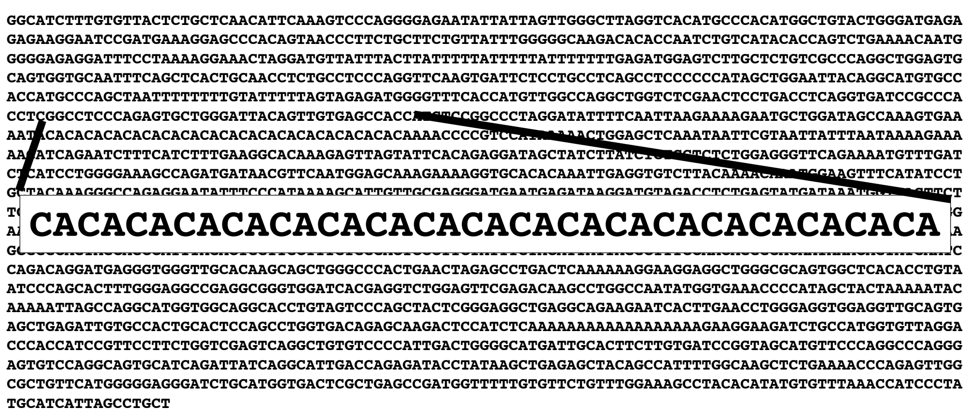
- Sequence repeats
- Single nucleotide polymorphisms
- Insertions and deletions
  - Nucleotides to kilobases
- Rearrangements

A typical sequence from the human genome…

```
GGCATCTTTGTGTTACTCTGCTCAACATTCAAAGTCCCAGGGGAGAATATTATTAGTTGGGCTTAGGTCACATGCCCACATGGCTGTACTGGGATGAGA
GAGAAGGAATCCGATGAAAGGAGCCCACAGTAACCCTTCTGCTTCTGTTATTTGGGGGCAAGACACACCAATCTGTCATACACCAGTCTGAAAACAATG
GGGGAGAGGATTTCCTAAAAGGAAACTAGGATGTTATTTACTTATTTTTATTTTTATTTTTTTGAGATGGAGTCTTGCTCTGTCGCCCAGGCTGGAGTG
CAGTGGTGCAATTTCAGCTCACTGCAACCTCTGCCTCCCAGGTTCAAGTGATTCTCCTGCCTCAGCCTCCCCCCATAGCTGGAATTACAGGCATGTGCC
ACCATGCCCAGCTAATTTTTTTTGTATTTTTAGTAGAGATGGGGTTTCACCATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCCGCCCA
CCTCGGCCTCCCAGAGTGCTGGGATTACAGTTGTGAGCCACCATGTCCGGCCCTAGGATATTTTCAATTAAGAAAAGAATGCTGGATAGCCAAAGTGAA
AATACACACACACACACACACACACACACACACACACACACAAAACCCCGTCCATAAAAACTGGAGCTCAAATAATTCGTAATTATTTAATAAAAGAAA
AACATCAGAATCTTTCATCTTTGAAGGCACAAAGAGTTAGTATTCACAGAGGATAGCTATCTTATCTCTCCTCTCTGGAGGGTTCAGAAAATGTTTGAT
CTCATCCTGGGGAAAGCCAGATGATAACGTTCAATGGAGCAAAGAAAAGGTGCACACAAATTGAGGTGTCTTACAAAACAAATGGAAGTTTCATATCCT
GCTACAAAGGGCCAGAGGAATATTTCCCATAAAAGCATTGTTGCGAGGGATGAATGAGATAAGGATGTAGACCTCTGAGTATGATAAATGGTTAGTTCT
TCCTATTAGTTGTTGTTTCTGATGTAGAAACAGCGTCTTTCTCCCTATATCTGGTCTAAAATCCAACCTGATAGGAGACGTTTTCGTTTGGGATTATGG
AAAGATACAACAGTTCTGGGGGTTGAGTTCAGGGCTAATTTTCTGAAGGATAAGAGAGCAAGCCCCAGCCAAGAGCCAAGAGAAAGCAATGATGAGGAA
GCGGGCAGTAGCAGCCATTTAGACTGGTTGCTTTGTGGGACTCCCTTCTATTTGTACATTATTAGGCTTTCCAACAGGGGACAATAAACAGTATGAATC
CAGACAGGATGAGGGTGGGTTGCACAAGCAGCTGGGCCCACTGAACTAGAGCCTGACTCAAAAAAGGAAGGAGGCTGGGCGCAGTGGCTCACACCTGTA
ATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACGAGGTCTGGAGTTCGAGACAAGCCTGGCCAATATGGTGAAACCCCATAGCTACTAAAAATAC
AAAAATTAGCCAGGCATGGTGGCAGGCACCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGAAGAATCACTTGAACCTGGGAGGTGGAGGTTGCAGTG
AGCTGAGATTGTGCCACTGCACTCCAGCCTGGTGACAGAGCAAGACTCCATCTCAAAAAAAAAAAAAAAAAAAGAAGGAAGATCTGCCATGGTGTTAGGA
CCCACCATCCGTTCCTTCTGGTCGAGTCAGGCTGTGTCCCCATTGACTGGGGCATGATTGCACTTCTTGTGATCCGGTAGCATGTTCCCAGGCCCAGGG
AGTGTCCAGGCAGTGCATCAGATTATCAGGCATTGACCAGAGATACCTATAAGCTGAGAGCTACAGCCATTTTGGCAAGCTCTGAAAACCCAGAGTTGG
CGCTGTTCATGGGGGAGGGATCTGCATGGTGACTCGCTGAGCCGATGGTTTTTGTGTTCTGTTTGGAAAGCCTACACATATGTGTTTAAACCATCCCTA
TGCATCATTAGCCTGCT
```

…from sequence on chromosome 3 stretching from base positions 187543053 to 187545049 of the human genome hg16 (July 2003) assembly.

Microsatellite



A dinucleotide marker named AFM059XA9 and D3S1262 is located at position 187,545,049.

Microsatellites

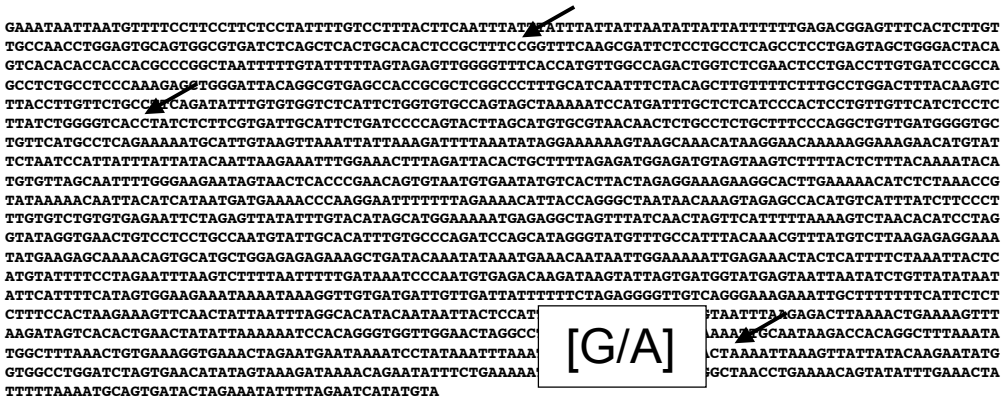
- **Many alleles, highly informative**
- **>50,000 in human genome**
- **Relatively high mutation rate**
- **Used to build first framework map**

More typical sequence …

GAAATAATTAATGTTTTCCTTCCTTCTCCTATTTTGTCCTTTACTTCAATTTATTTATTTATTATTAATATTATTATTTTTTGAGACGGAGTTTCACTCTTGT
TGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCACACTCCGCTTTCTGGTTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGACTACA
GTCACACACCACCACGCCCGGCTAATTTTTGTATTTTTAGTAGAGTTGGGGTTTCACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCA
GCCTCTGCCTCCCAAAGAGCTGGGATTACAGGCGTGAGCCACCGCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTTTCTTTGCCTGGACTTTACAAGTC
TTACCTTGTTCTGCCTTCAGATATTTGTGTGGTCTCATTCTGGTGTGCCAGTAGCTAAAAATCCATGATTTGCTCTCATCCCACTCCTGTTGTTCATCTCCTC
TTATCTGGGGTCACATATCTCTTCGTGATTGCATTCTGATCCCCAGTACTTAGCATGTGCGTAACAACTCTGCCTCTGCTTTCCCAGGCTGTTGATGGGGTGC
TGTTCATGCCTCAGAAAAATGCATTGTAAGTTAAATTATTAAAGATTTTAAATATAGGAAAAAAGTAAGCAAACATAAGGAACAAAAAGGAAAGAACATGTAT
TCTAATCCATTATTTATTATACAATTAAGAAATTTGGAAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAAGTCTTTTACTCTTTACAAAATACA
TGTGTTAGCAATTTTGGGAAGAATAGTAACTCACCCGAACAGTGTAATGTGAATATGTCACTTACTAGAGGAAAGAAGGCACTTGAAAAACATCTCTAAACCG
TATAAAAACAATTACATCATAATGATGAAAACCCAAGGAATTTTTTTAGAAAACATTACCAGGGCTAATAACAAAGTAGAGCCACATGTCATTTATCTTCCCT
TTGTGTCTGTGTGAGAATTCTAGAGTTATATTTGTACATAGCATGGAAAAATGAGAGGCTAGTTTATCAACTAGTTCATTTTTAAAAGTCTAACACATCCTAG
GTATAGGTGAACTGTCCTCCTGCCAATGTATTGCACATTTGTGCCCAGATCCAGCATAGGGTATGTTTGCCATTTACAAACGTTTATGTCTTAAGAGAGGAAA
TATGAAGAGCAAAACAGTGCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTGGAAAAATTGAGAAACTACTCATTTTCTAAATTACTC
ATGTATTTTCCTAGAATTTAAGTCTTTTAATTTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGATGGTATGAGTAATTAATATCTGTTATATAAT
ATTCATTTTCATAGTGGAAGAAATAAAATAAAGGTTGTGATGATTGTTGATTATTTTTTCTAGAGGGGTTGTCAGGGAAAGAAATTGCTTTTTTTCATTCTCT
CTTTCCACTAAGAAAGTTCAACTATTAATTTAGGCACATACAATAATTACTCCATTCTAAAATGCCAAAAAGGTAATTTAAGAGACTTAAAACTGAAAAGTTT
AAGATAGTCACACTGAACTATATTAAAAAATCCACAGGGTGGTTGGAACTAGGCCTTATATTAAAGAGGCTAAAAATTGCAATAAGACCACAGGCTTTAAATA
TGGCTTTAAACTGTGAAAGGTGAAACTAGAATGAATAAAATCCTATAAATTTAAATCAAAAGAAAGAAACAAACTGAAATTAAAGTTATTATACAAGAATATG
GTGGCCTGGATCTAGTGAACATATAGTAAAGATAAAACAGAATATTTCTGAAAAATCCTGGAAAATCTTTTGGGCTAACCTGAAAACAGTATATTTGAAACTA
TTTTTAAAATGCAGTGATACTAGAAATATTTTAGAATCATATGTA

…from sequence on chromosome 7 stretching from base positions 49,719,732 to 49,721,733.

Single nucleotide polymorphisms (SNPs)



Three SNPs are located at positions 49,719,887, 49,720,260 and 49,721,557.

SNPs

- **Less polymorphic/informative**
- **More stable inheritance**
- **~1 SNP / 1,250 nucleotides between any two genomes**
- **Mutation at CpG 10-fold higher rate**
- **2.5 million between two genomes**
- **Exist in coding regions**

DIPs

- **Deletion/insertion polymorphisms**
- **Small or large number of nucleotides**
- **Example chr1:120,506,653-120,506,677**

AGTATCTTCACAGAAATGACCATA

AGTATCTTCACAAGAAATGACCATA

AGTATCTTCACA[-/A]GAAATGACCATA

DIPs

Example: chr7:105,060,001-105,060,023

CAGACTCAATAAGCATGTTTTTA

CAGACTCAATAAGCATGTTTTTTTTTTTTTTTTTTTTTGAGACG
GAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCAGTGGCGCGA
TCTCGGCTCACTGCAAGCTCCGCCTCCCGGGTTCACGCCATT
CTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCTCCC
GCCACCACGCCCGGCTAATTTTTTGTATTTTTAGTAGAGACGG
GGTTAGCATGTTTTT

CAGACTCAATA[LARGEINSERTION/-]AGCATGTTTTT

Human Genetic Variation

- **What types of variants exist?**
- **How are variants found?**
- **How are variants scored?**
- **How are variants used?**

Microsatellite identification

- **Databases/Maps**
  - **deCODE Genetics**
  - **Marshfield Clinic**
  - **Genome DataBase**
  - **Cooperative Human Linkage Center**

Microsatellite identification: deCODE

A high-resolution recombination map of the human genome
Kong et al. (2002) *Nature Genetics* **31**: 241

5,136 microsatellite markers

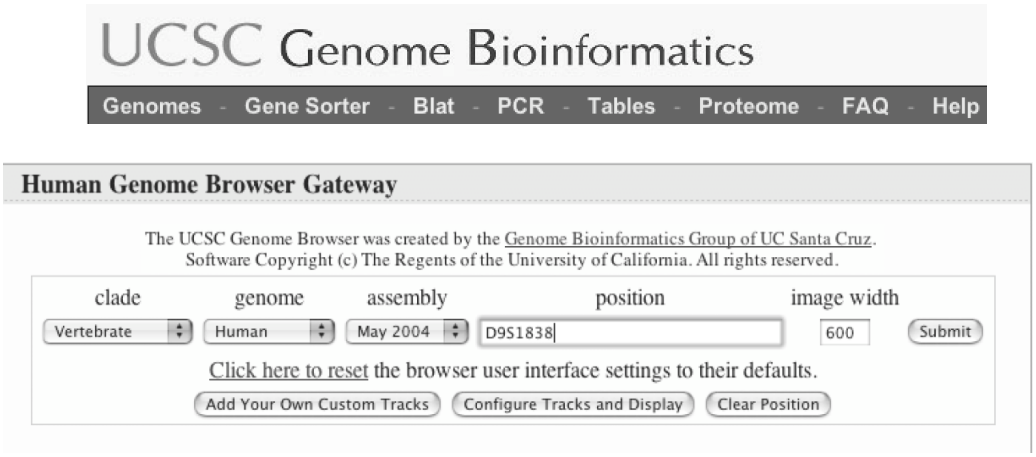
869 individuals from 146 Icelandic families

1,257 meiotic events

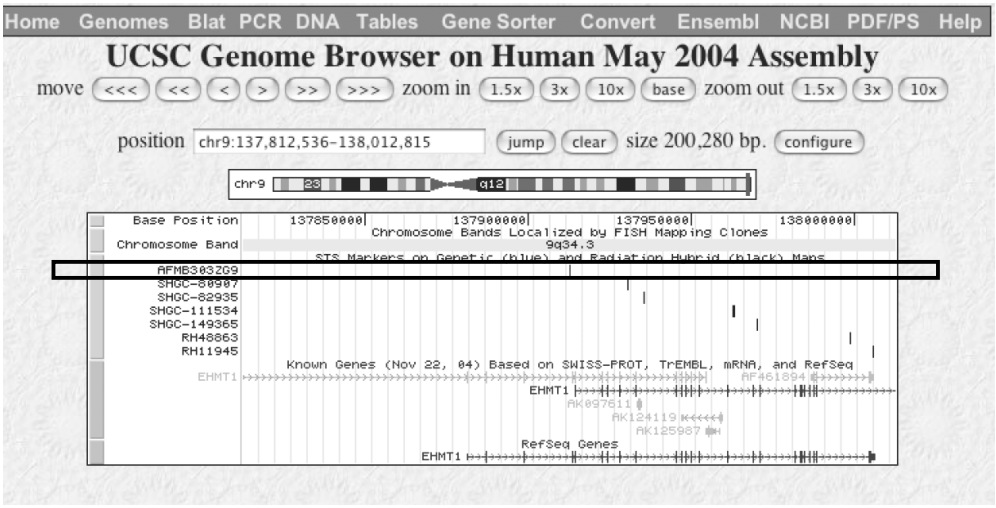
Microsatellite identification: deCODE

Marker retrieval: genome browser

http://genome.ucsc.edu/



Marker retrieval: genome browser



Marker retrieval: genome browser

STS Marker AFMB303ZG9

Chromosome: chr9
Start: 137912536
End: 137912815
Band: 9q34.3

Other names: D9S1838, RH15582, B303ZG9, W3232, RH9769, HSB303ZG9

UCSC STS id: 2879
UniSTS id: 9019
Genbank: Z53450
GDB: GDB:610512 GDB:604229
Organism: Homo sapiens

Left Primer: ACCCAGCTACTGAGGAGGCTT
Right Primer: GCTTCTGCACTTTGTAGAACCAAAT
Distance: 159-175 bps

Genetic Map Positions

	Name	Chromosome	Position
Genethon:	AFMB303ZG9	chr9	166.50
Marshfield:	AFMB303ZG9	chr9	163.84

Marker retrieval: genome browser

Home - Genomes - Genome Browser - Gene Sorter - Blat - PCR - Tables - FAQ - Help

Get DNA in Window

Get DNA for

Position chr9:137,912,342-137,913,009

Note: if you would prefer to get DNA for features of a particular track or table, try the Table Browser using the output format sequence.

Sequence Retrieval Region Options:

Add 0 extra bases upstream (5') and 0 extra dov

Note: if a feature is close to the beginning or end of a chro
they may be truncated in order to avoid extending past the

Sequence Formatting Options:

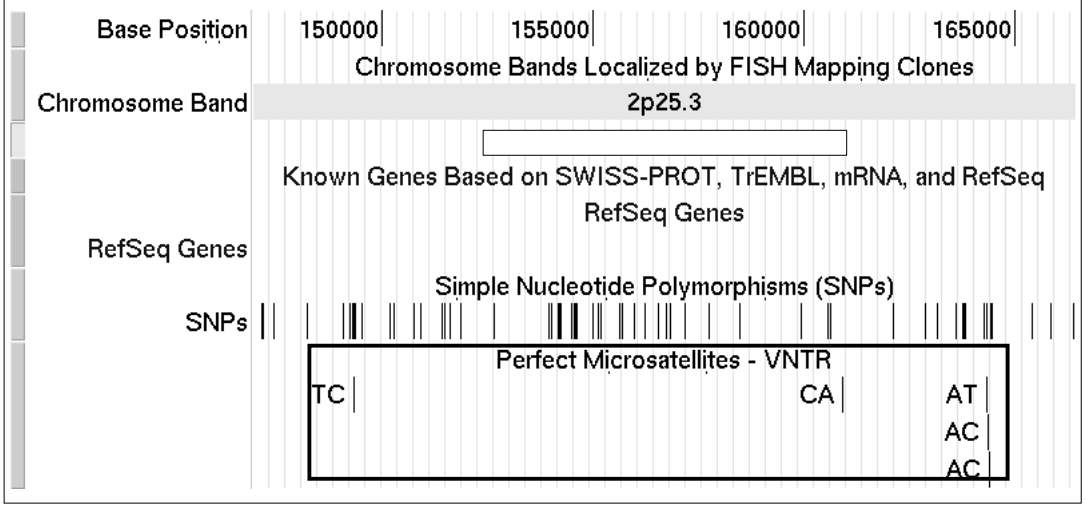
- (•) All upper case.
- ( ) All lower case.
- [ ] Mask repeats: (•) to lower case ( ) to N
- [ ] Reverse complement (get '-' strand sequence)

Get DNA | Extended case/color options

Note: The "Mask repeats" option applies only to "Get DNA", not to "Extended case/color options".

```
>hg17_dna range=chr9:137912342-137913009 5'pad=0 3'
CCAAGTGTCTGCATGTTGGCTGTGTGCTCCGAGCCTGACCCCCATGAACA
TACTGCAGACGCCTGGTGTGATCGTTTCCCAGCGTCCGTGGTCCCAGGCA
CCTCCTTACTCCAGAGCGGATTGCCCAGGCCCCGCGGCGTCTGTGGGTGG
TGCTGTCAAAGGACCTACCCGCTTTGGATGGTTCTCACTCGTTCACGTCC
CTCCAACTGTTGCTTCTGCACTTTGTAGAACCAAATTTGTGTGTGTGTGT
GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTAAGTAAAGGGGGTCTCACTG
TGTTTCCCAGGCTGGGCTTGAACTCCTGGAATGAAGCAAACCCTCCCTCC
CACTCAAGCCTCCTCAGTAGCTGGGTCTTCAGGTGTGAGCTGCTGTGCCC
AGCTTTAAACAGAGTGGATTTTCCCCATCCCTTTAGGAGAGTTTCTTTTA
TGTTAAAGCAGTGGCTTTTAGATCTGTTTTTCTTTAAATCCTGGAACTTA
AAAAAAGTCATGGAGTCTGATTATATAAAACCAGTCGAACCTAGAACTGC
TTTGTTCAGAATGGGGTTGGGAGCCCCCAGGCCCCTTCCTCGCTGCTGCT
CTTGTTTGGAAACCACTGCTCCAGAGCCCCCAGACATTTGCTTCTCCCTC
TCAGCCTCTTATCTTTTA
```

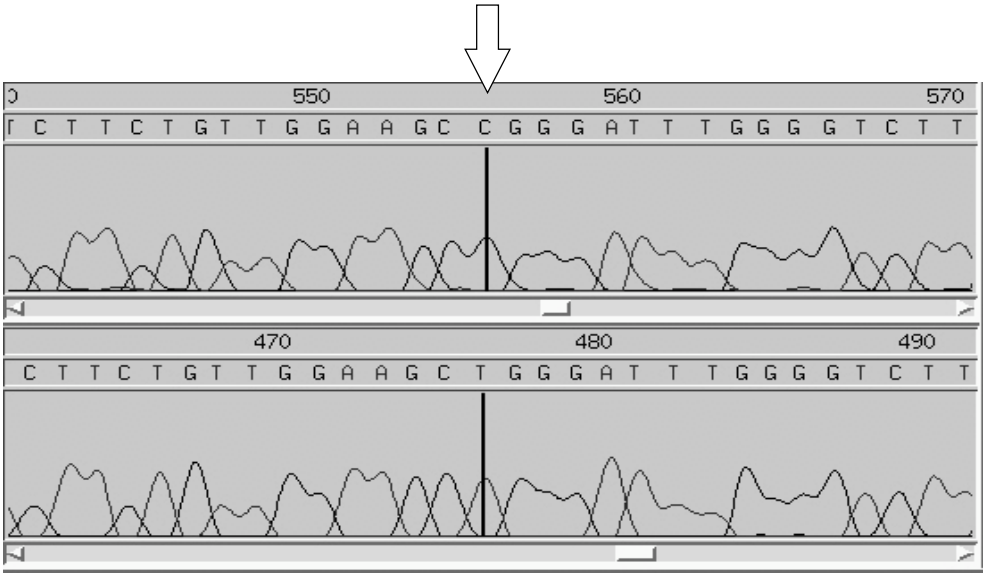
Microsatellite identification from sequence



SNP identification

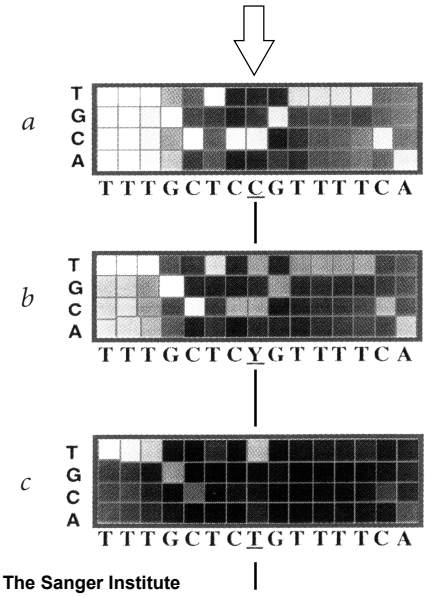
- **Sequencing**
- **Databases**

SNP identification: sequencing



SNP identification: sequencing chips

...GCTCCGTTT...
...GCTCTGTTT...



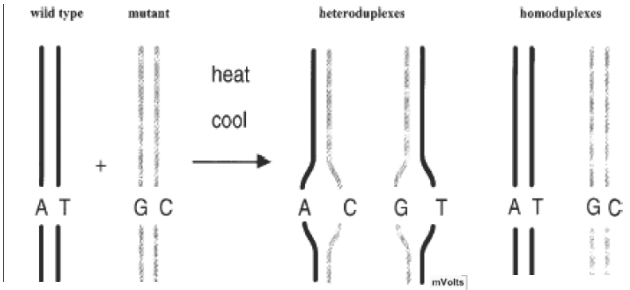
The Sanger Institute

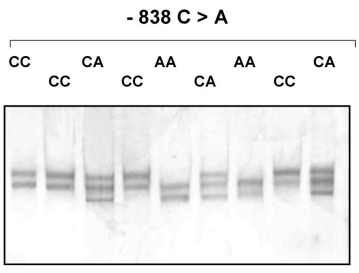
Identification of DNA mismatches

- **Other rapid methods to determine if two sequences differ:**
  - **single strand conformational polymorphism (SSCP)**
  - **denaturing high performance liquid chromatography (dHPLC)**

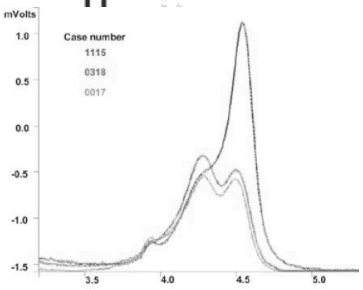
- **These methods do not provide the precise nucleotide change**

Identification of mismatches





González (2004) BMC Biology 2:5

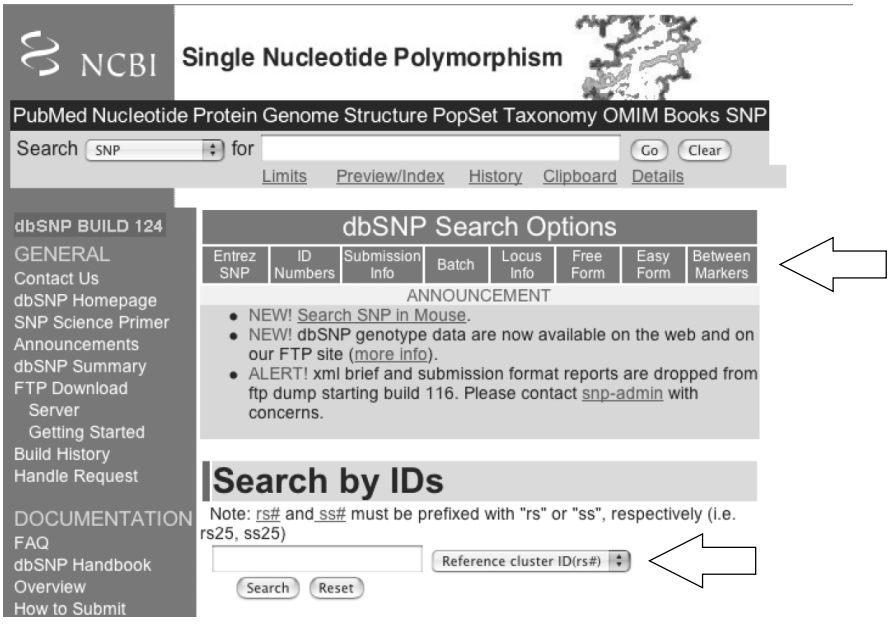


Campbell (2004) Breast Cancer Res 6:R366

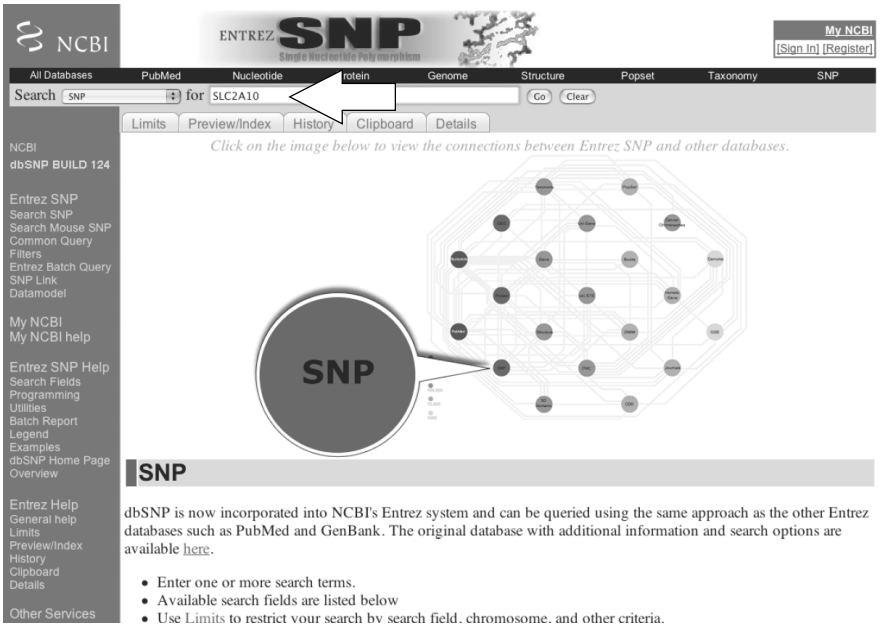
SNP identification: databases

- **dbSNP**
- **The SNP Consortium (TSC)**
- **Human Gene Variation base (HGVbase)**
- **CGAP Genetic Annotation Initiative (CGAP-GAI)**
- **Innate immunity (IIPGA)**
- **Environmental genome project (EGP)**
- **Japanese SNPs (JSNP)**

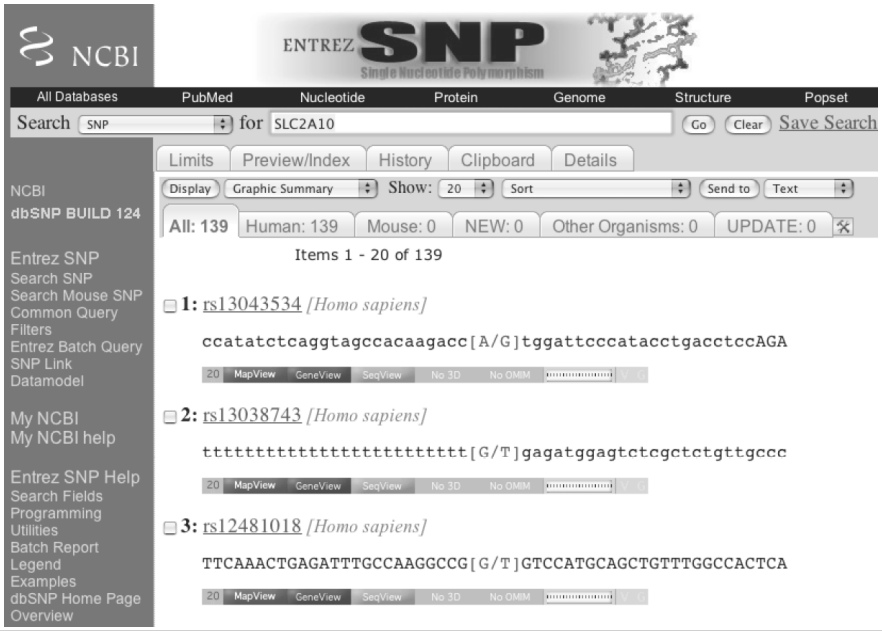
SNP identification: dbSNP



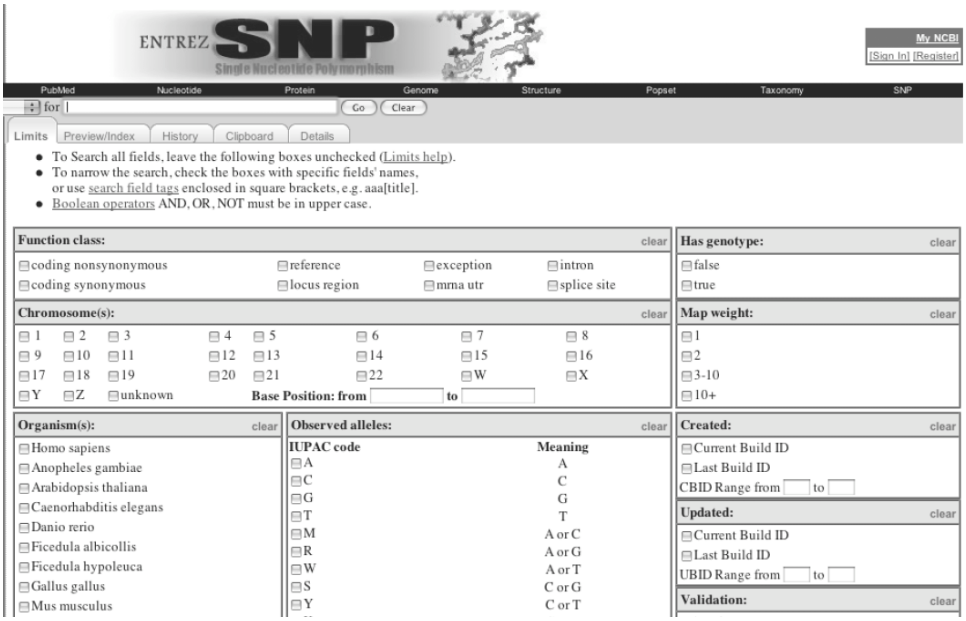
SNP retrieval: Entrez SNP



SNP retrieval: Entrez SNP



Entrez SNP: Limits



Entrez SNP: Limits

SNP class:	
het	variation has unknown sequence composition, but is observed to be heterozygous
in del	insertion deletion polymorophism, deletions represented by '-' in allele string
microsat	microsatellite / simple sequence repeat
mixed	
mnp	multiple nucleotide polymorphism (all alleles same length where length>1)
named	allele sequences defined by name tag instead of raw sequence, e.g. (Alu)/-
no variation	submission reports invariant region in surveyed sequence
snp	true single nucleotide polymorphism

Method class:	
computed	variation was mined from sequence alignment with software
dhplc	Denaturing High Pressure Liquid Chromatography used to detect SNP
hybridize	hybridization method (e.g. chip) was used to assay for variation
other	other method used to detect variation
rflp	variation in enzyme restriction site used to detect variation
sequence	samples were sequenced and resulting alignment used to define variation
sscp	single stranded conformational polymorphism used to detect variation
unknown	

Entrez SNP: GeneView

SNP linked to Gene (geneID:81031)

SNP are linked from gene SLC2A10 via the following methods:

Contig Annotation | GenBank(mrna) Mapping

Send all rs# to Batch Query | Download all rs# to file. | Gene Genotype Report

Gene Model (mRNA alignment) information from genome sequence

Total gene model (contig mRNA transcript):				1	
Contig	mrna	protein	mrna orientation	transcript	snp list
NT_011362	NM_030777	NP_110404	forward	plus strand	currently shown

view rs | in gene region | cSNP | has frequency | double hit | haplotype tagged

gene model (contig mRNA transcript):	Contig	mrna	protein	mrna orientation	transcript	snp count
	NT_011362	NM_030777	NP_110404	forward	plus strand	128, all

Color Legend

Contig position	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
10392176	rs6063016	N.D.				intron				
10392245	rs2425896	N.D.				intron				
10392659	rs6090543	N.D.				intron				
10393800	rs2425897	N.D.				intron				

Entrez SNP: GeneView

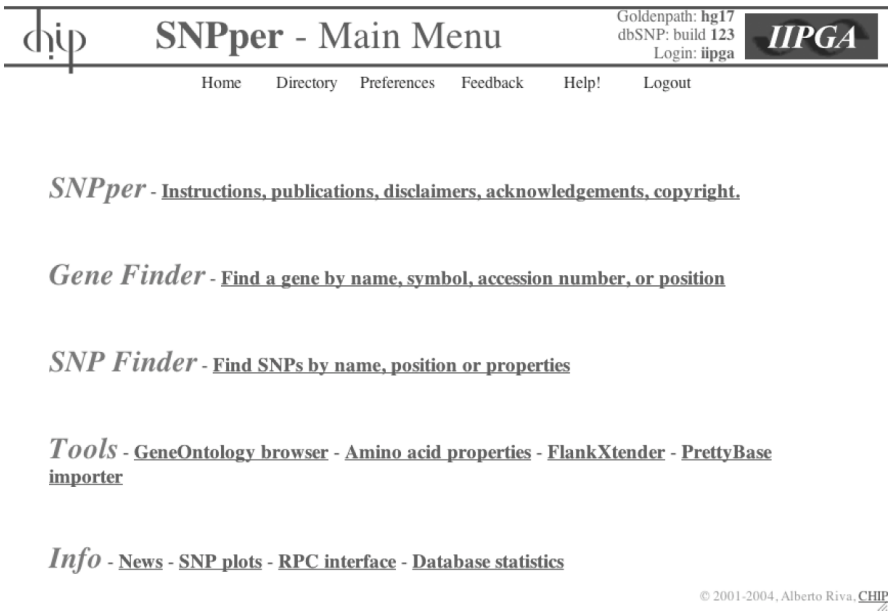
4999137	rs2229683	0.069				untranslated region				
4999698	rs3831326	N.D.				intron				
4999831	rs2236574	0.083				synonymous	T	Ile [I]	3	390
		0.083				contig reference	C	Ile [I]	3	390
5000932	rs2306663	0.091				intron				
5001059	rs2306662	0.094				synonymous	G	Leu [L]	3	355
		0.094				contig reference	A	Leu [L]	3	355
5001814	rs5811	N.D.				nonsynonymous	C	Thr [T]	2	255
		N.D.				contig reference	A	Lys [K]	2	255
5002082	rs4660238	N.D.				synonymous	A	Pro [P]	3	196
		N.D.								
5004532	rs3820546	N.D.								

Fasta sequence (Legend)

>gnl|dbSNP|rs5811|allelePos=61|totalLen=121|taxid=9606|snpclass=1|alleles='A/C'|mol=cDNA|build=52

CTGACGTGAC CCATGACCTG CAGGAGATGA AGCAAGAGAG TCGGCAGATG ATGCGGGAGA
M
GAAGGTCACC ATCCTGGAGC TGTTCCGCTC CCCCGCCTAC CGCCAGCCCA TCCTCATCGC

SNP retrieval: SNPper



Build 124 dbSNP Human Content

21,581,724	SNP submissions (ss#)
10,054,521	RefSNP clusters (rs#)
5,054,675	validated SNPs
2,727,888	SNPs with genotype details

Build 124 dbSNP Human Content

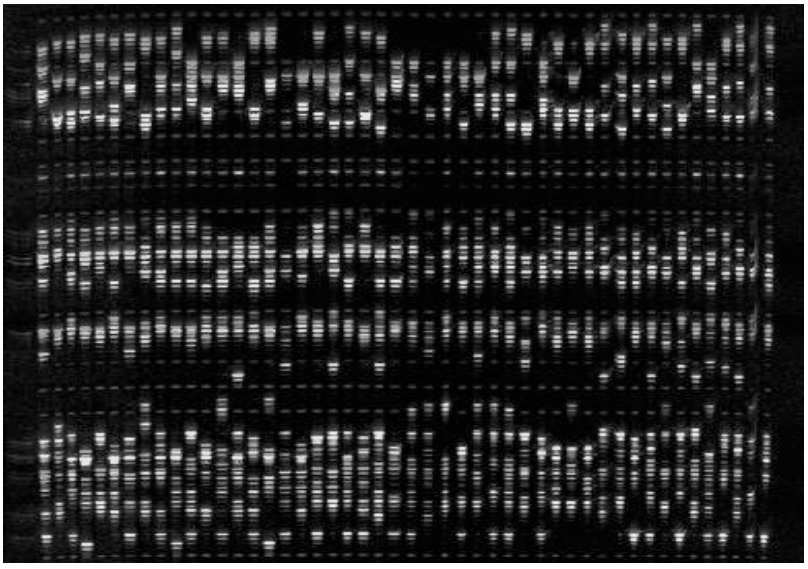
Annotated with human genome build 35.1

FUNCTION	SNPS	GENES
Locus region	338,787	26,210
Synonymous	39,214	14,342
Nonsynonymous	50,772	15,710
Untranslated region	546,961	17,898
Intron	2,932,608	19,448
Splice site	832	769

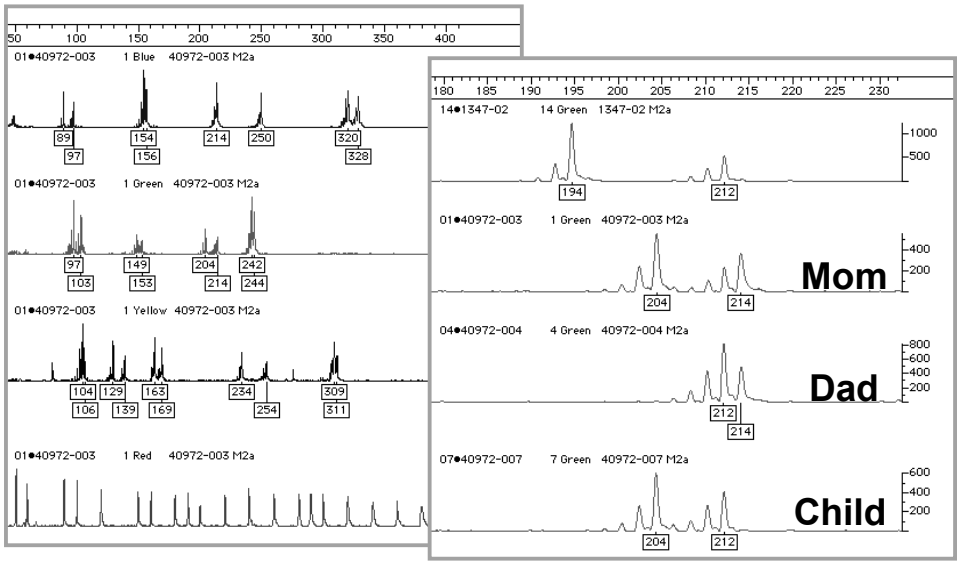
Human Genetic Variation

- **What types of variants exist?**
- **How are variants found?**
- **How are variants scored?**
- **How are variants used?**

Scoring Microsatellites



Scoring Microsatellites



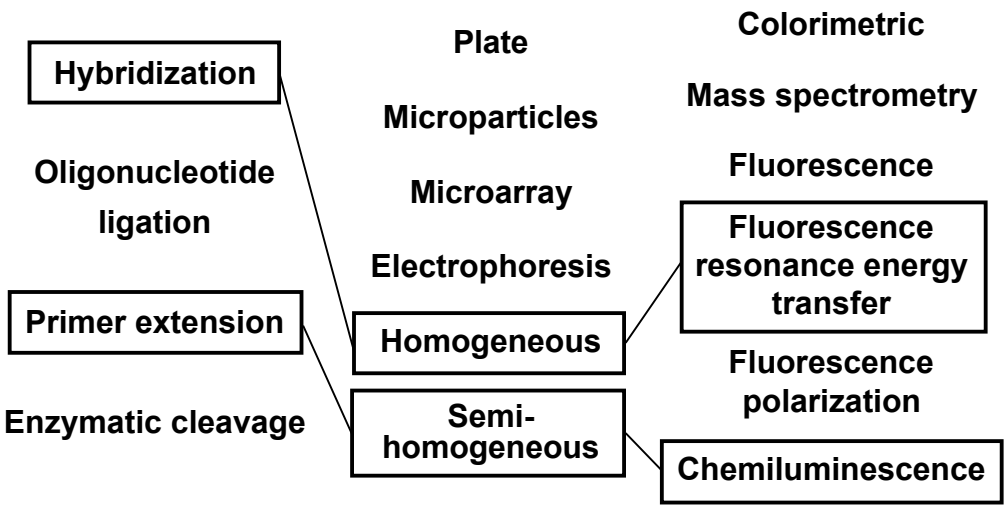
Scoring SNPs

- **Genotype accuracy**
- **Cost of assays and specialized instrument(s)**
- **Assay development time and ease**
- **Ability to automate**

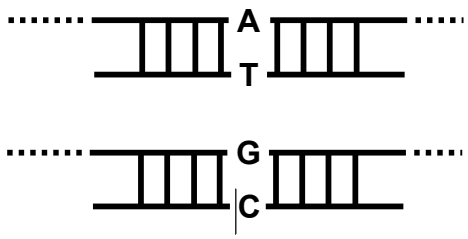
Scoring SNPs (2)

- **Time to perform assays**
- **Ability to multiplex**
- **Data accumulation and analysis**
- **Allele frequency quantification**

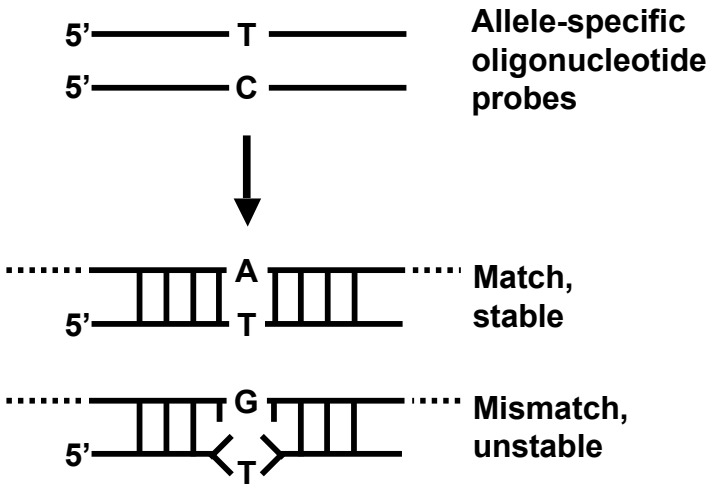
Overview of SNP typing methods



Example SNP



Hybridization



Affymetrix Custom Sequencing Array

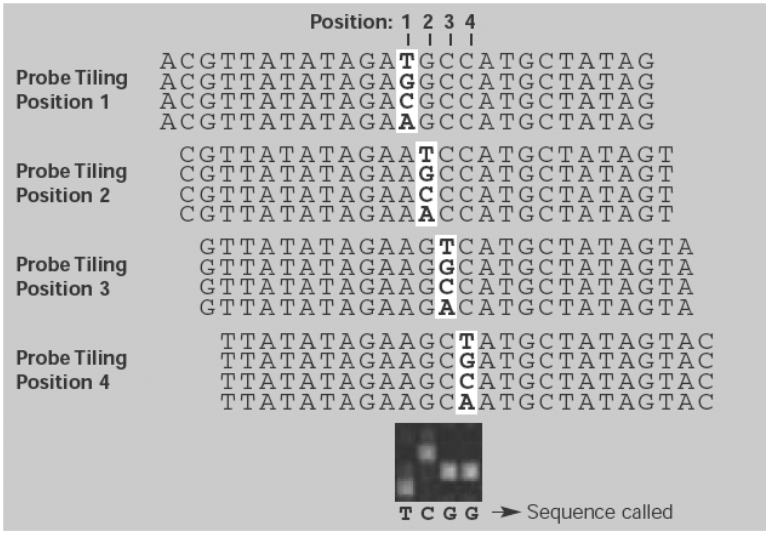


Figure 1: CustomSeq™ arrays tile four probes per strand for each individual base. The central position of each probe varies to incorporate each of the four possible nucleotides—A, C, G, or T.

images from affymetrix.com

Affymetrix GeneChip 10K Array

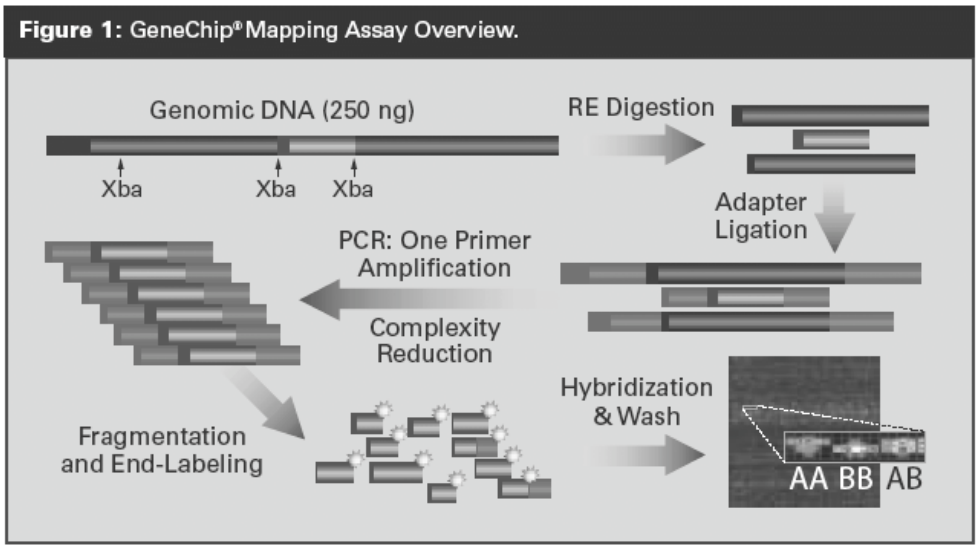
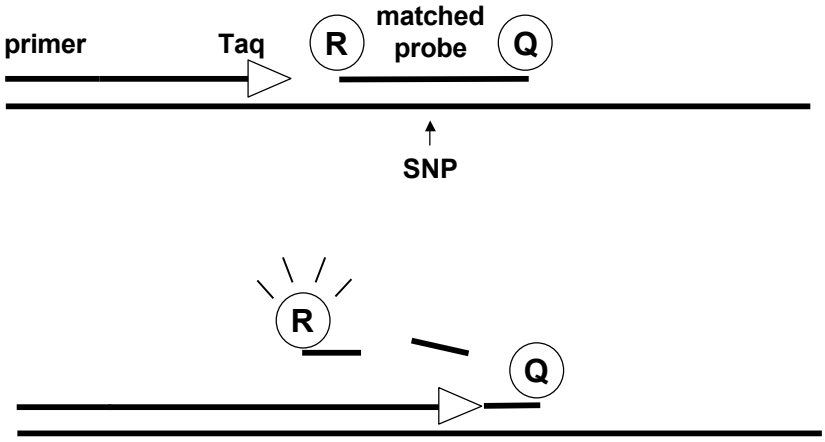


image from affymetrix.com

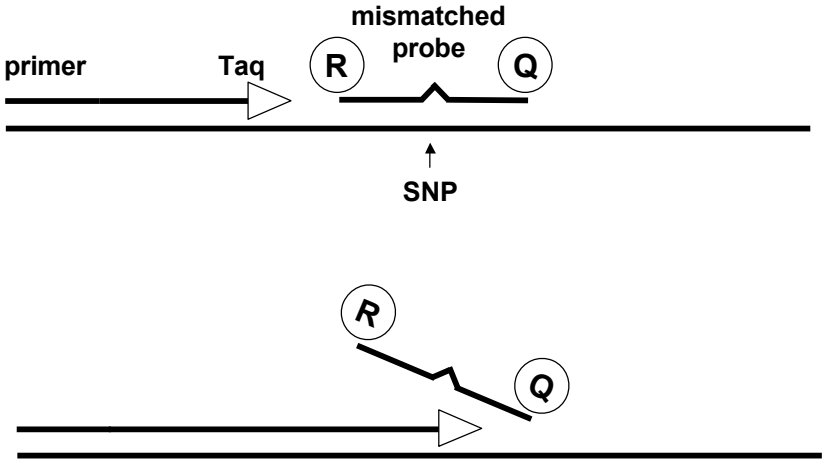
Hybridization to Oligonucleotide Arrays

- **Advantages:**
  - **Simple to perform**
  - **Highly multiplexed**
  - **Automated analysis**
  - **Genome-wide SNPs (mapping chip)**
- **Disadvantages**
  - **Custom chip expensive to design/create**
  - **Mapping chip SNPs pre-selected**
  - **Local sequence affects success**

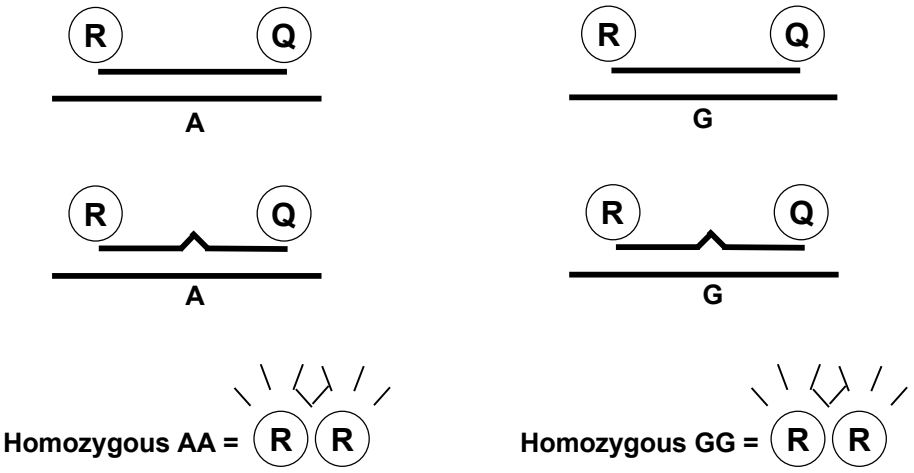
Fluorescence resonance energy transfer (FRET)



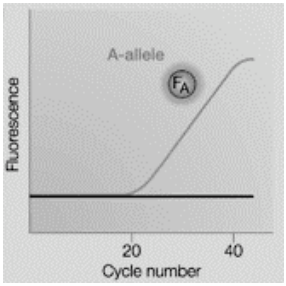
Fluorescence resonance energy transfer (FRET)

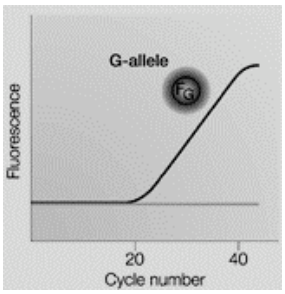


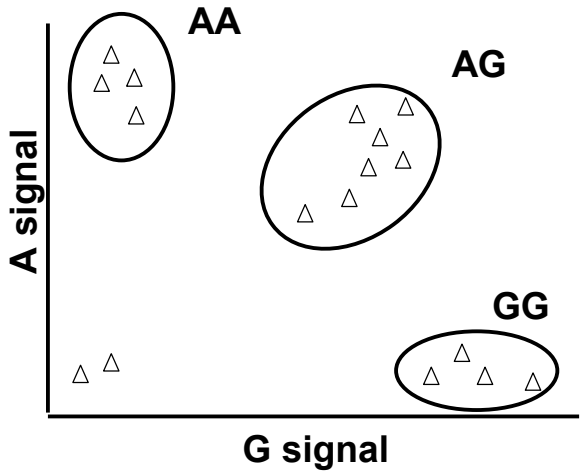
TaqMan competing probes



TaqMan genotype scoring



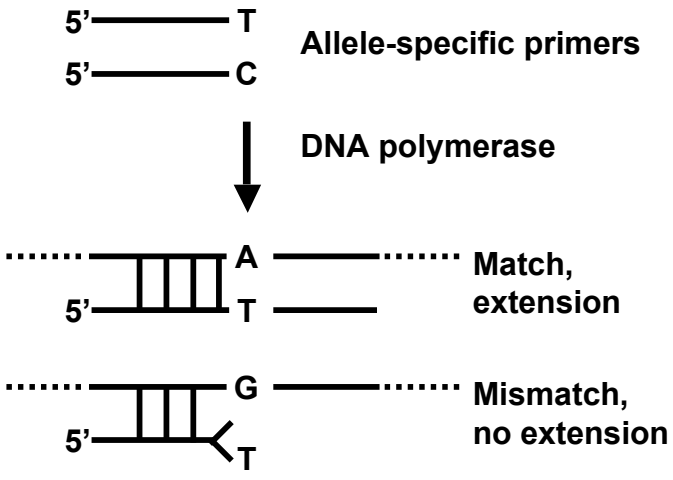




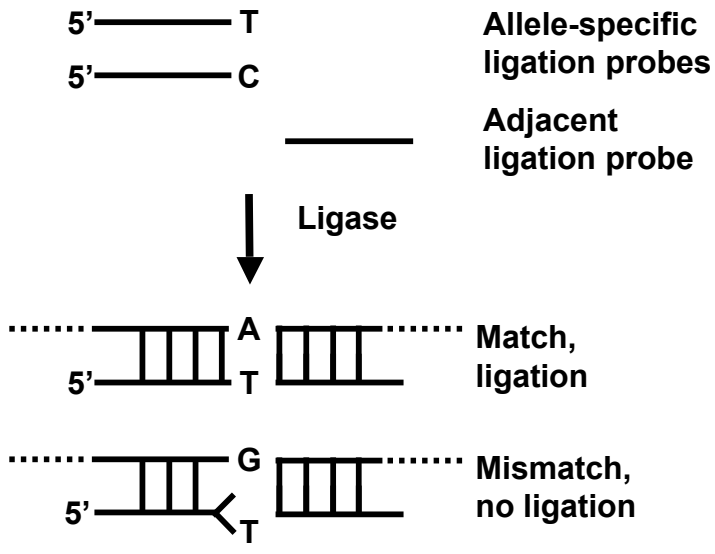
TaqMan

- **Advantages:**
  - **Simple to perform**
  - **Closed-tube system**
  - **Accurate quantification**
- **Disadvantages**
  - **Expensive probes**
  - **Assays require optimization**

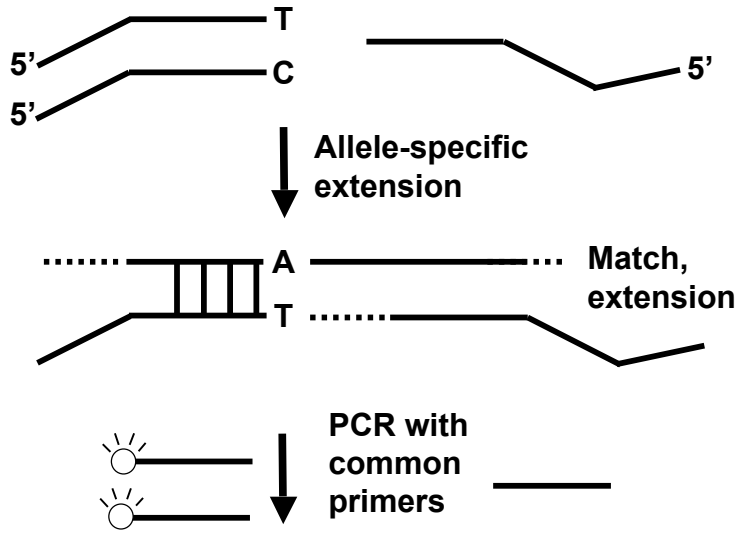
Allele-specific PCR



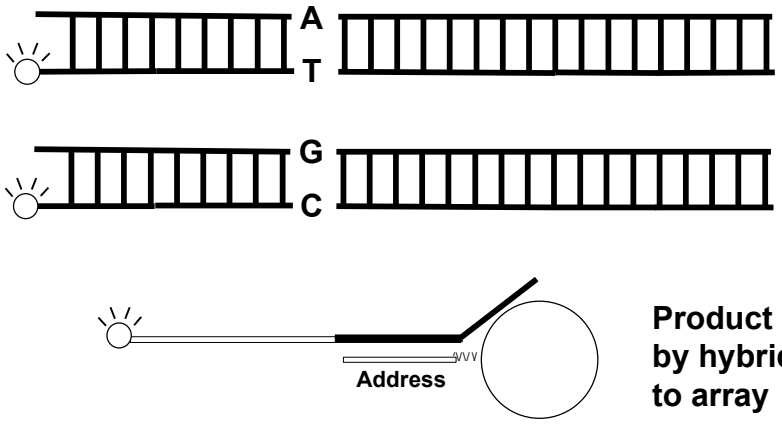
Oligonucleotide Ligation Assay (OLA)



Illumina: Allele-specific extension



Illumina: Allele-specific extension

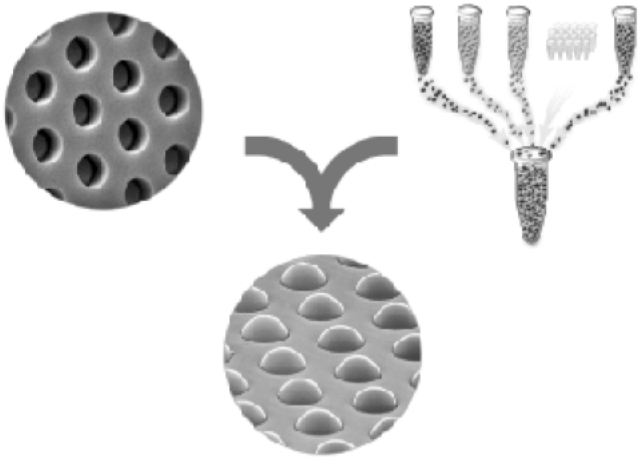


Product capture by hybridization to array

AA = AG = GG =

Readout

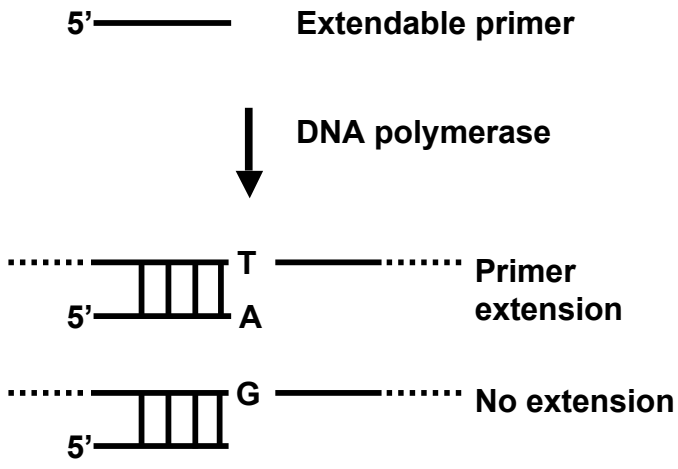
Illumina genotyping technology



Illumina

- **Advantages:**
  - **Very highly multiplexed**
  - **Accurate**
  - **Low cost per genotype**
- **Disadvantages**
  - **Not all SNPs can be designed**
  - **Not flexible**

Primer extension = Minisequencing

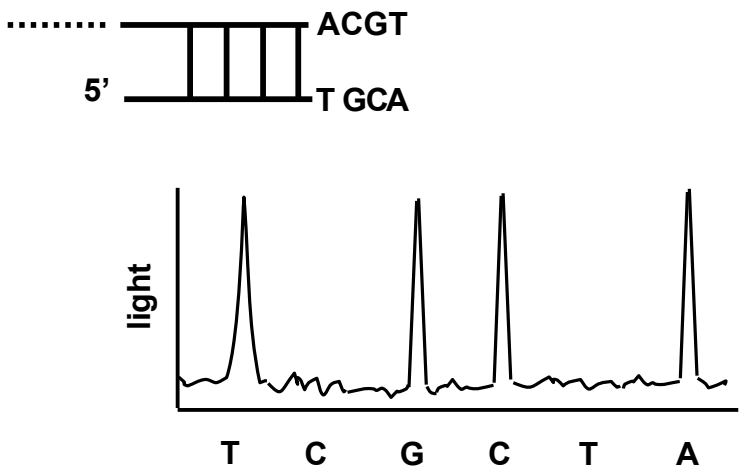


Pyrosequencing

- **Four enzymes**
  - **DNA polymerase**
  - **ATP sulfurylase--converts pyrophosphate to ATP**
  - **Luciferase--converts ATP to light**
  - **Apyrase--degrades excess nucleotides**
- **Nucleotides added sequentially**

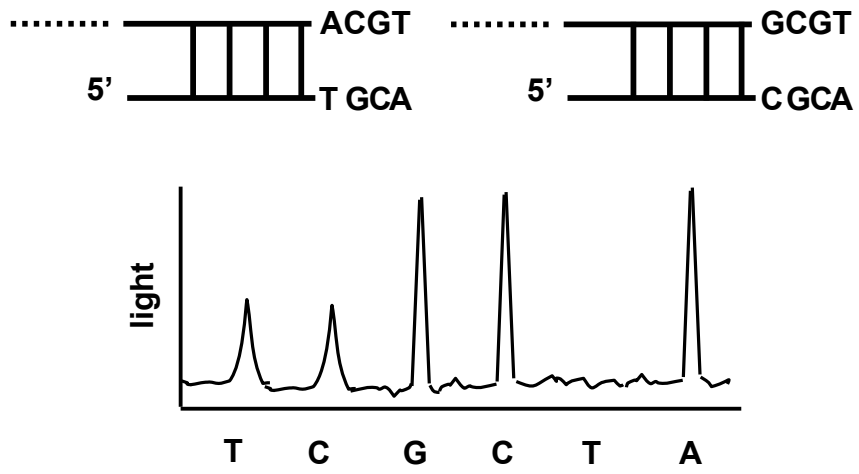
Pyrosequencing

…[A/G]CGT…



Pyrosequencing

…[A/G]CGT…



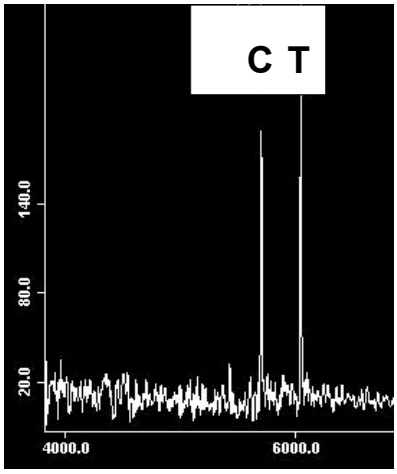
Pyrosequencing

- **Advantages:**
  - **Accurate**
  - **Accurate allele frequency estimation**
  - **Robust for closely spaced SNPs**
- **Disadvantages**
  - **Expensive reagents**
  - **Requires post-PCR processing**

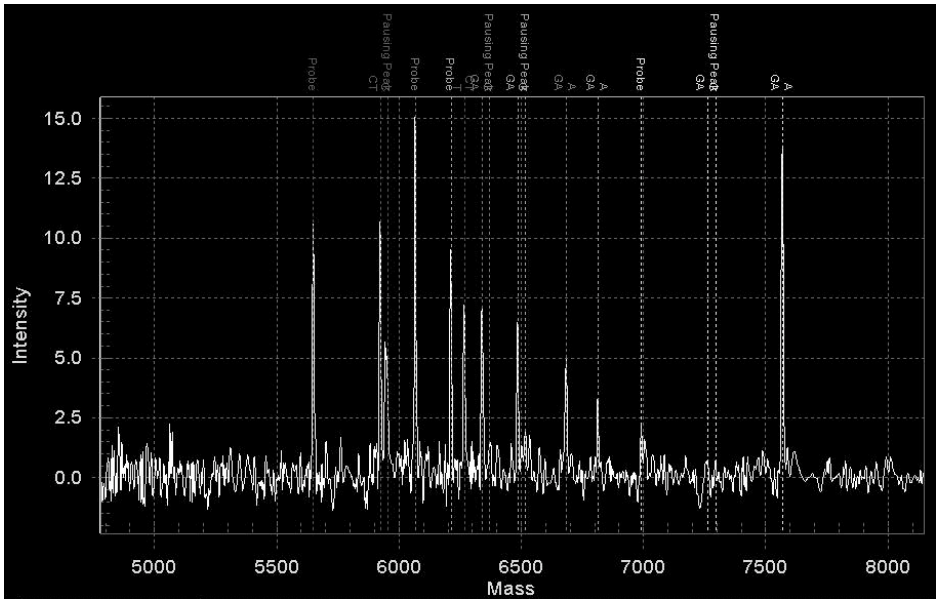
Primer extension mass spectrometry

Primer extension reactions designed to generate different sized products

	Mass in Daltons
GGACCTGGAGCCCCCACC	5430.5
GGACCTGGAGCCCCCACCC	5703.7
GGACCTGGAGCCCCCACCTG	6047.9



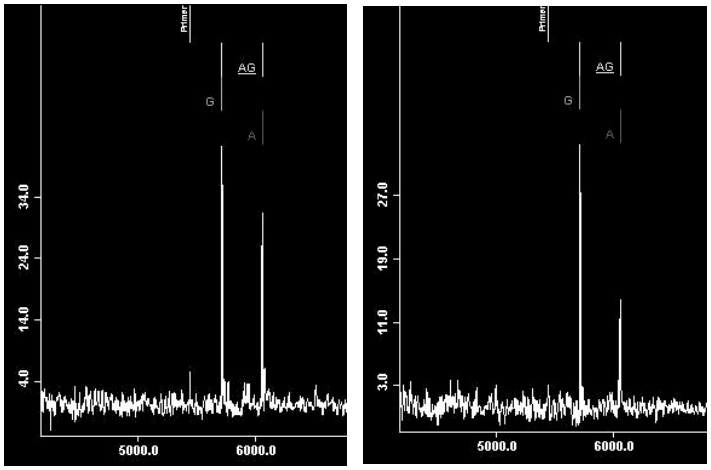
Mass spectrometry multiplexing



Allelic quantification

- **Pools of individual DNAs or tumor sample**
- **Type SNP and determine relative allele frequencies**

Affected cases **Unaffected controls**



Primer extension mass spectrometry

- **Advantages:**
  - **Accurate**
  - **Automated assay design**
  - **Fast automated data collection**
  - **Multiplexing capacity**
- **Disadvantages**
  - **Expensive instruments, consumables**
  - **Extensive post-PCR processing**

Quality control of genotype data

- **High genotype success**
- **Accurate duplicate genotypes**
- **No genotypes in no DNA controls**
- **Allele frequencies similar to databases**
- **Accurate on a second platform**

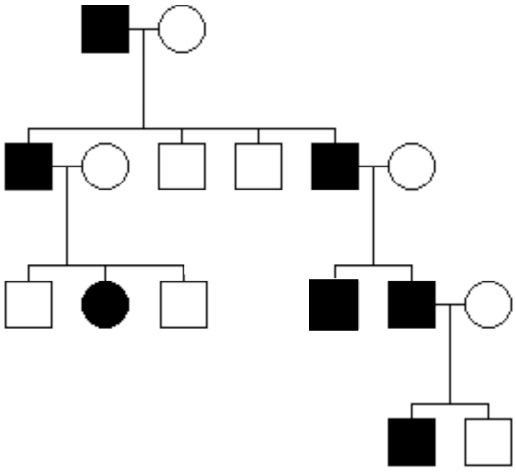
Quality control of genotype data

- **Test whether data are consistent with Hardy-Weinberg Equilibrium (HWE): $p^2 + 2pq + q^2 = 1$**
- **Calculate observed frequencies p and q**
- **Use p and q to calculate expected genotype frequencies**
- **Compare observed and expected genotype frequencies by X^2 test with 1 degree of freedom**

Human Genetic Variation

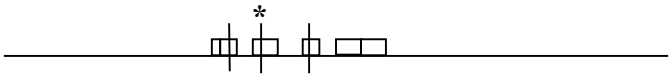
- **What types of variants exist?**
- **How are variants found?**
- **How are variants scored?**
- **How are variants used?**

Linkage analysis

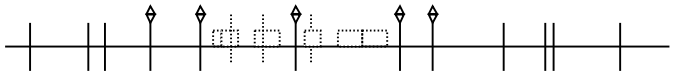


Association Studies

Direct

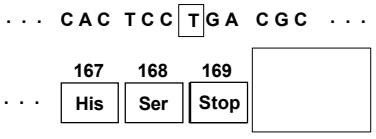


Indirect

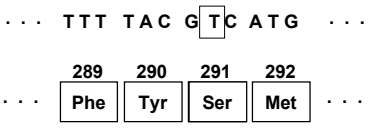


Functional variants

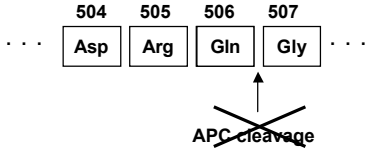
Drug metabolism:
The CYP2D6 gene



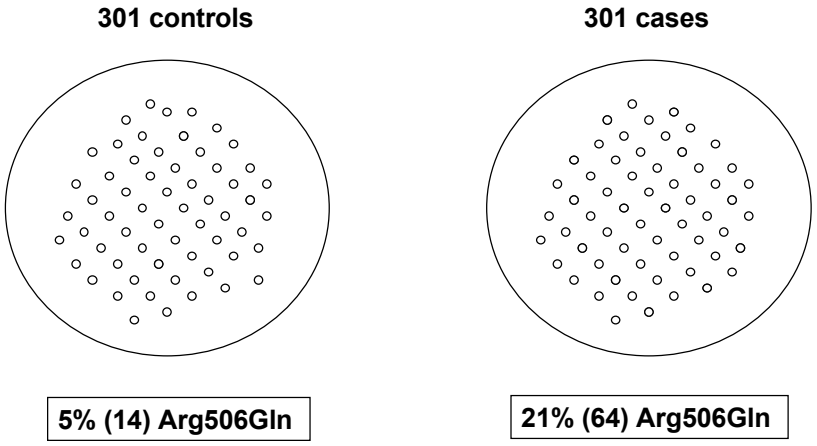
Coronary disease:
LDL receptor gene



Deep-vein thrombosis:
The Factor V gene



Factor V Leiden association study



Case-control association study

	cases	controls
risk allele	a	b
non-risk allele	c	d

$$\text{odds ratio} = \frac{a / c}{b / d} = \frac{ad}{bc}$$

Case-control association study

	cases	controls
risk allele	64	14
non-risk allele	237	287

$$\text{odds ratio} = \frac{64*287}{14*237} = 5.54$$

Disease is 5.54 times as frequent with risk allele

Genome-wide SNP panels

- **10,000 - 500,000 SNPs per experiment**
- **Affymetrix, Illumina, Parallele, Perlegen**
  - **Random SNPs**
  - **Coding SNPs**
  - **Nonsynonymous SNPs**
  - **Selected nonredundant SNPs**

Future

- **Continued identification of SNPs**
- **Faster, cheaper, easier genotyping**
- **More SNP panels for genome-wide association studies**
- **Discovery of new functional variants**

Websites

Marshfield	**research.marshfieldclinic.org/genetics/**
GDB	**www.gdb.org/**
CHLC	**gai.nci.nih.gov/CHLC/**
dbSNP	**www.ncbi.nlm.nih.gov/SNP/**
TSC	**snp.cshl.org/**
HGVbase	**hgvbase.cgb.ki.se**
CGAP	**cgap.nci.nih.gov/**
Innate immunity	**innateimmunity.net/**
EGP	**www.niehs.nih.gov/envgenom/**
JSNP	**snp.ims.u-tokyo.ac.jp/**
Entrez	**www.ncbi.nlm.nih.gov/Entrez**
SNPper	**snpper.chip.org/**

References

Genetic Map

Kong (2002) Nature Genetics 31:241

Mutation Rates

Nachman (2000) Genetics 156: 297

SNP Identification

International SNP mapping group (2001) Nature 409:928

Venter et al. (2001) Science 291:1304

SNP Typing

Syvanen (2001) Nat Review Genet 2:930

Kwok (2001) Ann Rev Genomics Hum Genet 2:235

Gut (2001) Human Mutation 17:475