Studying Genetic Variation I: Laboratory Techniques

Karen Mohlke, PhD
Genome Technology Branch
NHGRI

Human Genetic Variation

Variants contribute to rare and common diseases

Variants can be used to trace human origins

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**
- **Insertion/deletion**
  - **Nucleotide(s)**
  - **Alu element**

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
GTCACACACCACCACGCCCGGCTAATTTTTTGTATTTTTAGTAGAGTTGGGGTTTCACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCA
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)

GAAATAATTAATGTTTTCCTTCCTTCTCCTATTTTGTCCTTTACTTCAATTTATTTATTTATTATTAATATTATTATTTTTTGAGACGGAGTTTCACTCTTGT
TGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCACACTCCGCTTTCGGTTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGACTACA
GTCACACACCACCACGCCCGGCTAATTTTTTGTATTTTTAGTAGAGTTGGGGTTTCACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCA
GCCTCTGCCTCCCAAAGAGCTGGGATTACAGGCGTGAGCCACCGCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTTTCTTTGCCTGGACTTTACAAGTC
TTACCTTGTTCTGCCTTCAGATATTTGTGTGGTCTCATTCTGGTGTGCCAGTAGCTAAAAATCCATGATTTGCTCTCATCCCACTCCTGTTGTTCATCTCCTC
TTATCTGGGGTCACTATCTCTTCGTGATTGCATTCTGATCCCCAGTACTTAGCATGTGCGTAACAACTCTGCCTCTGCTTTCCCAGGCTGTTGATGGGGTGC
TGTTCATGCCTCAGAAAAATGCATTGTAAGTTAAATTATTAAAGATTTTAAATATAGGAAAAAAGTAAGCAAACATAAGGAACAAAAAGGAAAGAACATGTAT
TCTAATCCATTATTTATTATACAATTAAGAAATTTGGAAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAAGTCTTTTACTCTTTACAAAATACA
TGTGTTAGCAATTTTGGGAAGAATAGTAACTCACCCGAACAGTGTAATGTGAATATGTCACTTACTAGAGGAAAGAAGGCACTTGAAAAACATCTCTAAACCG
TATAAAAACAATTACATCATAATGATGAAAACCCAAGGAATTTTTTTAGAAAACATTACCAGGGCTAATAACAAAGTAGAGCCACATGTCATTTATCTTCCCT
TTGTGTCTGTGTGAGAATTCTAGAGTTATATTTGTACATAGCATGGAAAAATGAGAGGCTAGTTTATCAACTAGTTCATTTTTAAAAGTCTAACACATCCTAG
GTATAGGTGAACTGTCCTCCTGCCAATGTATTGCACATTTGTGCCCAGATCCAGCATAGGGTATGTTTGCCATTTACAAACGTTTATGTCTTAAGAGAGGAAA
TATGAAGAGCAAAACAGTGCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTGGAAAAATTGAGAAACTACTCATTTTCTAAATTACTC
ATGTATTTTCCTAGAATTTAAGTCTTTTAATTTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGATGGTATGAGTAATTAATATCTGTTATATAAT
ATTCATTTTCATAGTGGAAGAAATAAAATAAAGGTTGTGATGATTGTTGATTATTTTTTCTAGAGGGGTTGTCAGGGAAAGAAATTGCTTTTTTTCATTCTCT
CTTTCCACTAAGAAAGTTCAACTATTAATTTAGGCACATACAATAATTACTCCATTCTAAAATGCCAAAAAGGTAATTTAAGAGACTTAAAACTGAAAAGTTT
AAGATAGTCACACTGAACTATATTAAAAAATCCACAGGGTGGTTGGAACTAGGCCTTATATTAAAGAGGCTAAAAATTGCAATAAGACCACAGGCTTTAAATA
TGGCTTTAAACTGTGAAAGGTGAAACTAGAATGAATAAAATCCTATAAATTTAAATCAAAAGAAAGAAACAAACTAAAATTAAAGTTATTATACAAGAATATG
GTGGCCTGGATCTAGTGAACATATAGTAAAGATAAAACAGAATATTTCTGAAAAATCCTGGAAAATCTTTTGGGCTAACCTGAAAACAGTATATTTGAAACTA
TTTTTAAAATGCAGTGATACTAGAAATATTTTAGAATCATATGTA

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
- 2.5 million between two genomes
- Exist in coding regions

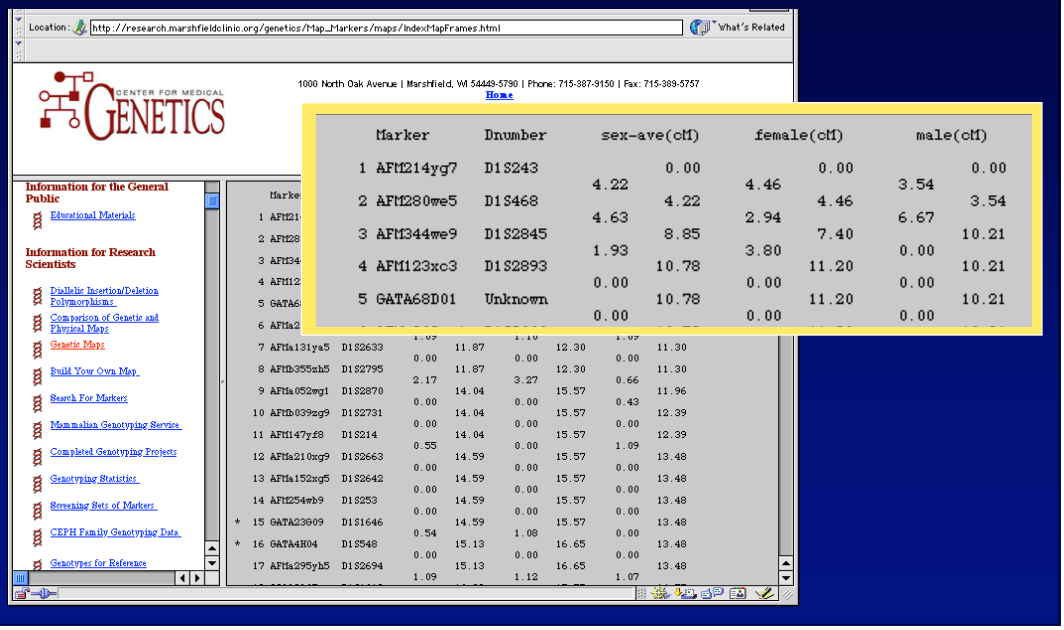
Human Genetic Variation

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

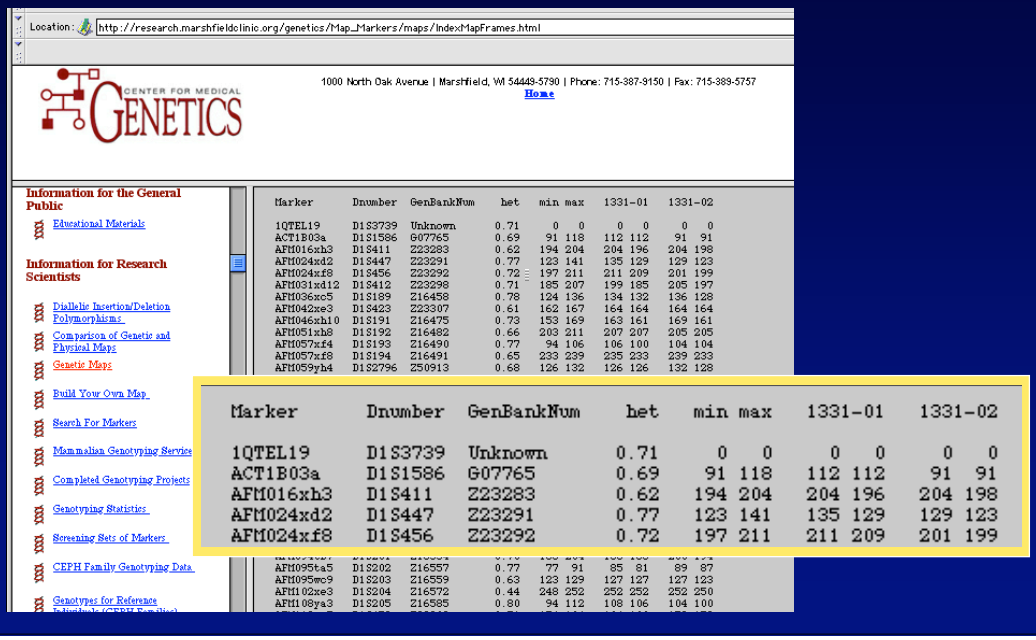
Microsatellite identification

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

Microsatellite identification: databases



Microsatellite identification: databases



Marker	Dnumber	GenBankNum	het	min	max	1331-01	1331-02
1QTEL19	D1S3739	Unknown	0.71	0	0	0	0
ACT1B03a	D1S1586	G07765	0.69	91	118	112 112	91 91
AFM016xh3	D1S411	Z23283	0.62	194	204	204 196	204 198
AFM024xd2	D1S447	Z23291	0.77	123	141	135 129	129 123
AFM024xf8	D1S456	Z23292	0.72	197	211	211 209	201 199

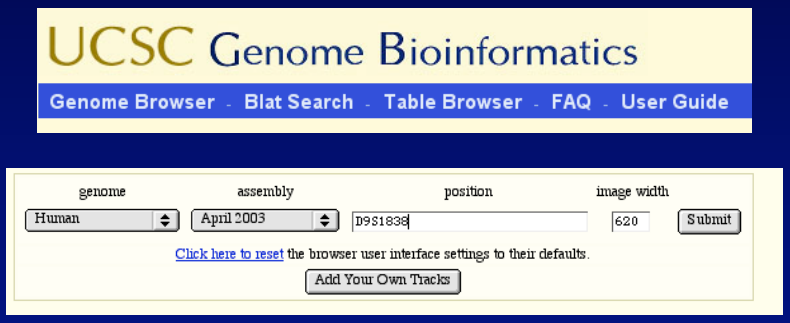
Microsatellite identification from sequence

Sputnik: searches DNA sequence files in Fasta format for microsatellite repeats.

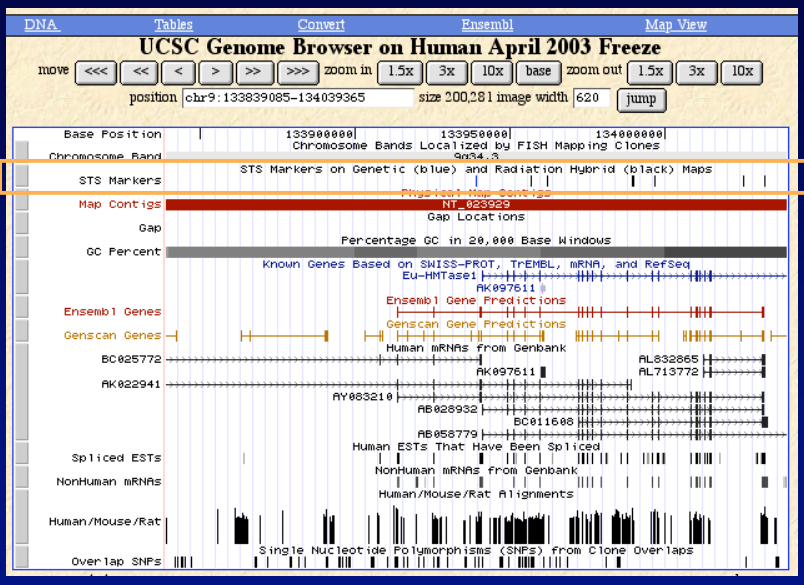
```
>bK2653D5.00294  Unfinished sequence: bK2653D5  Contig_ID: 00294  acc=
Length: 1604 bp dinucleotide from Sputnik: bases 519-1080
tcttaggtagaataagatccagtaagtatagacacttttgcggcatccaaagaattaacc
cttcactcatttactcacctggtaagagatacagggagaaagctgtggagtaactcaggg
agctggagcccataaggcaggaaacccatgcccattcattcaacaaacttgtattgagct
ccttttgatgcatcccccatccactataagcacttggagacccacacagatgtggtttc
tgctcccatagtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt
gtgtgtgtgtgtaggggaatgtaaacaggaaaacagatatgcaaaacaatttcagatcgc
ggtaagtgctaggaacagaatgaaataggataggagtgatggacaggggagacttcaggt
ggagtcatcgggaaaagcctctccataaagtgaccttctgggagaaaaccgaggggtaag
aatctggtcctgcaaagatctgggcaagaaatgtccaggtgtagggaacagcgaggtcaa
agtcaccatcacaaggaaacgc
```

Marker retrieval: genome browser

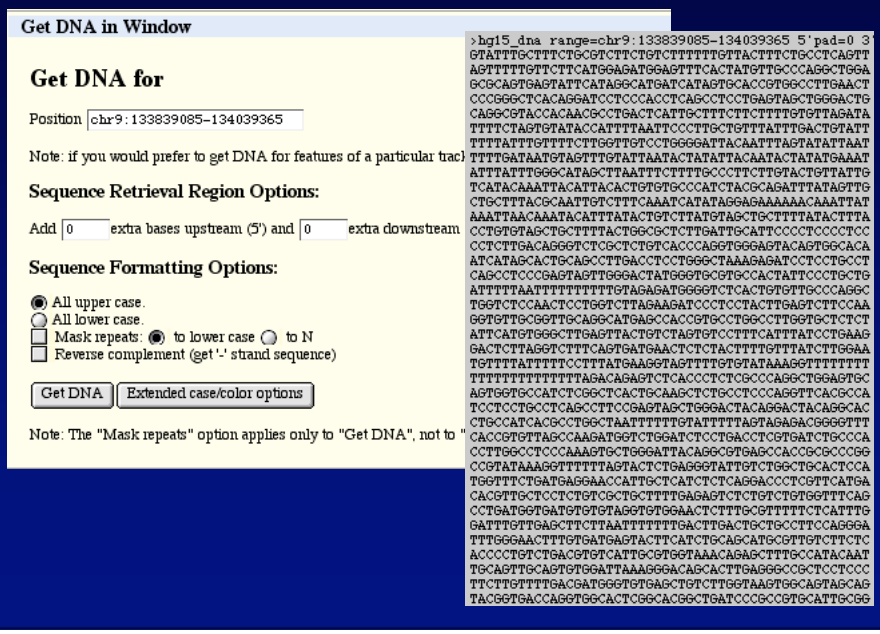
http://genome.ucsc.edu/



Marker retrieval: genome browser



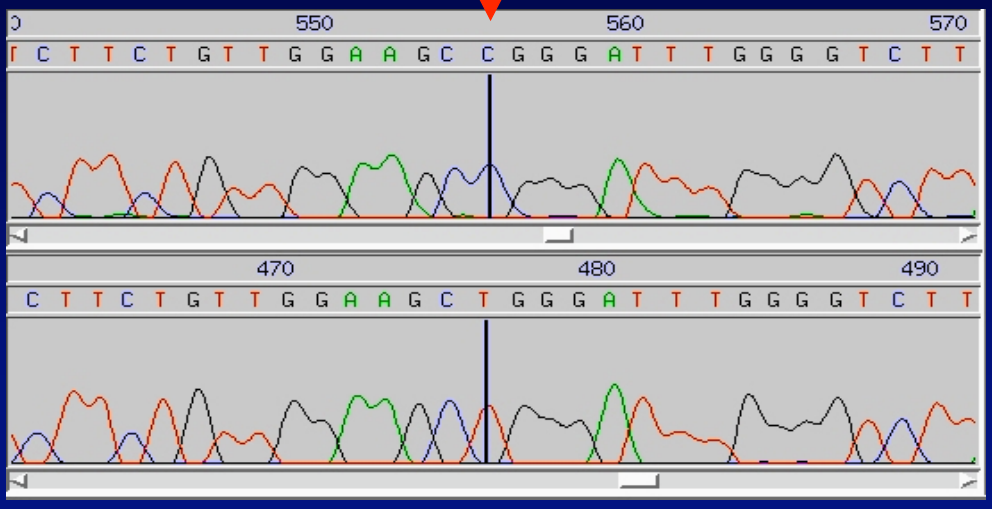
Marker retrieval: genome browser



SNP identification

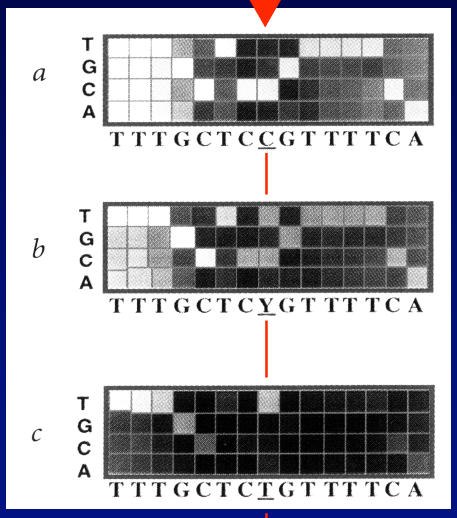
- Sequencing
- Databases

SNP identification: sequencing



SNP identification: sequencing chips

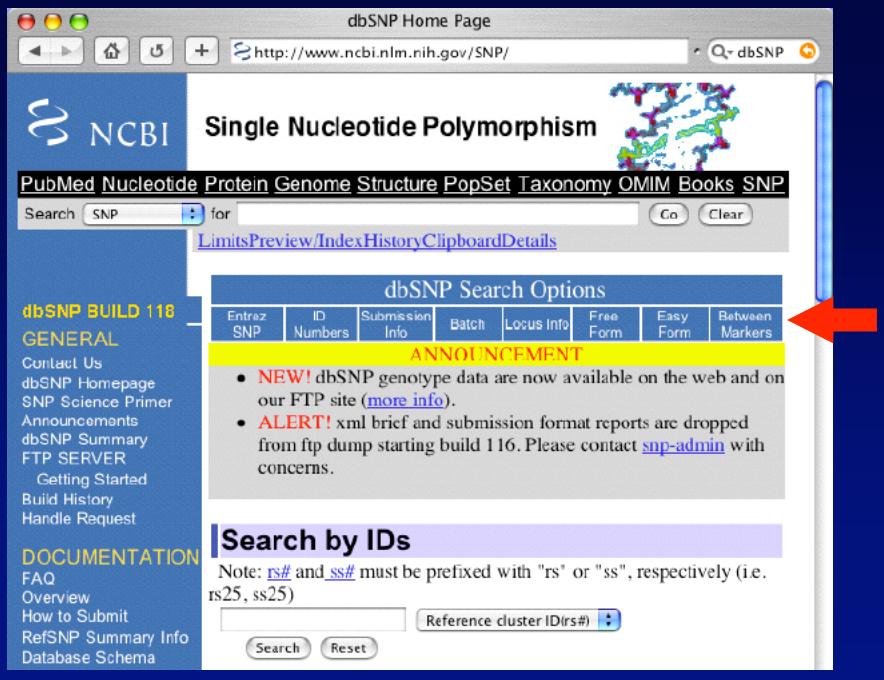
...GCTCCGTTT...
...GCTCTGTTT...



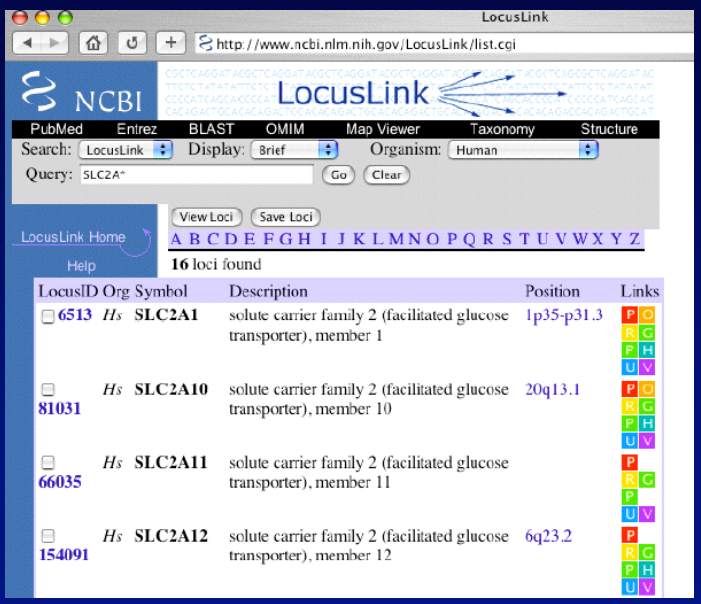
The Sanger Institute

SNP identification: databases

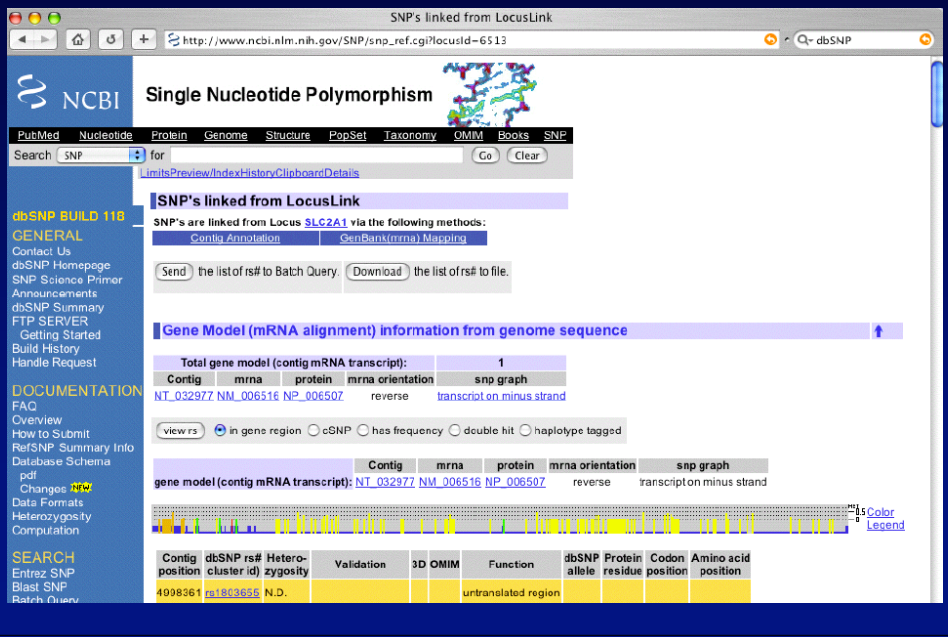
- dbSNP — Nov 2003
  - 10,384,535 submitted; 5,798,183 reference; 288,265 with frequencies
- The SNP Consortium (TSC)
  - ~1.8 million SNPs
- Human Gene Variation base (HGVbase)
  - 2,859,131 entries
- CGAP Genetic Annotation Initiative (CGAP-GAI)
  - 23,039 total (1% confirmed, 27% validated, 72% candidate)
- Japanese SNPs (JSNP)
  - 195,059 total; 84,560 with allele frequencies



SNP retrieval: LocusLink



SNP retrieval: LocusLink



SNP retrieval: LocusLink

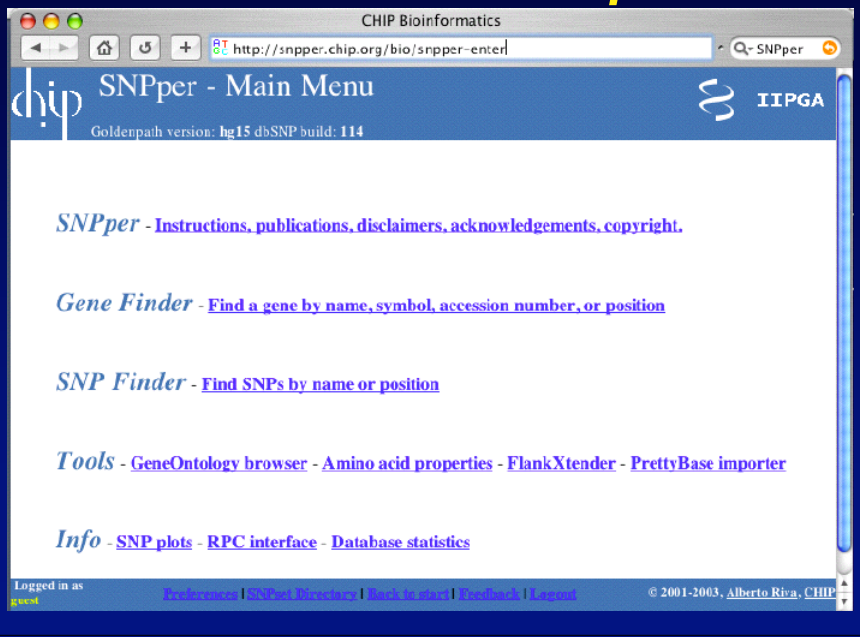
4999137	rs2229683	0.069				untranslated region			
4999698	rs3831326	N.D.				intron			
4999831	rs2236574	0.083				synonymous	T	Ile [I]	3	390
		0.083				contig reference	C	Ile [I]	3	390
5000932	rs2306663	0.091				intron			
5001059	rs2306662	0.094				synonymous	G	Leu [L]	3	355
		0.094				contig reference	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 117 dbSNP content

4,678,153 SNPs
unique in human reference genome build 34

Intergenic:	3,108,657	(66.5%)
Intragenic:	1,569,496	(33.5%)
Exonic	269,357	(5.8%)
Intronic	1,300,139	(27.8%)
Splice A/D	319	

Build 117 dbSNP content

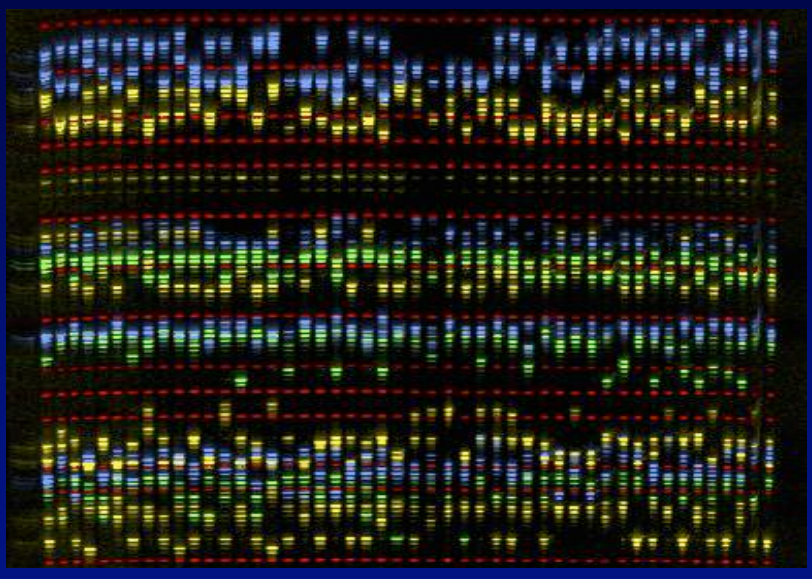
Of the 43,506 coding SNPs:

Synonymous	20,728	48%
Nonsynonymous	22,778	52%

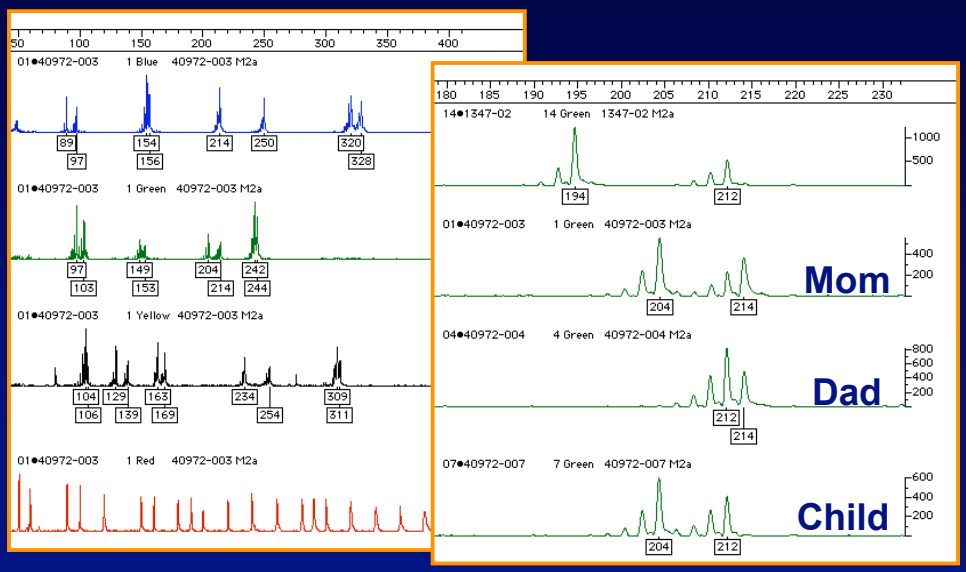
Human Genetic Variation

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

Scoring Microsatellites



Scoring Microsatellites



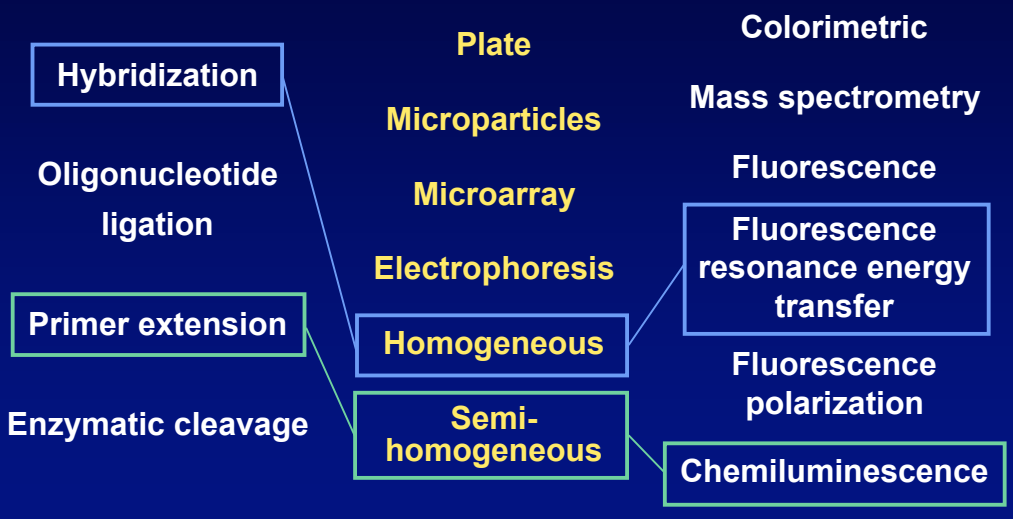
Scoring SNPs

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

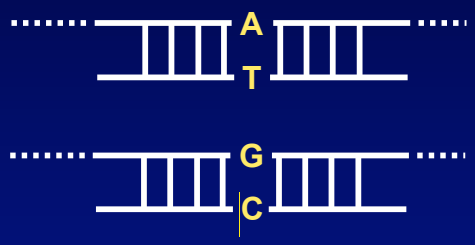
Scoring SNPs (2)

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

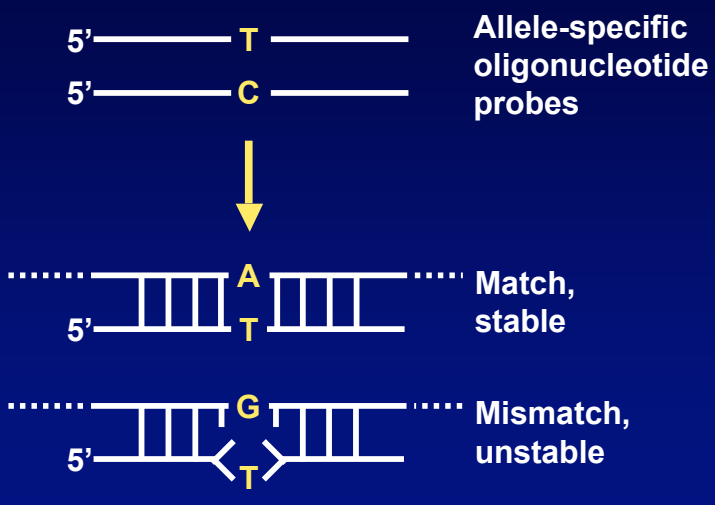
Overview of SNP typing methods



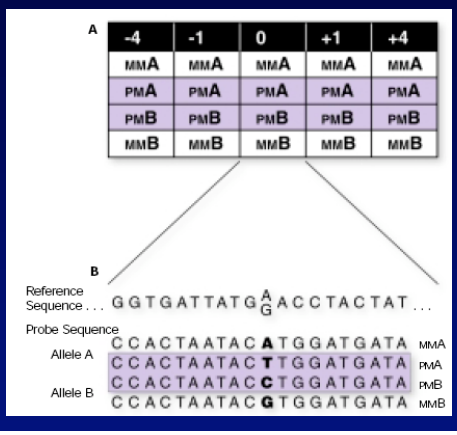
Example SNP

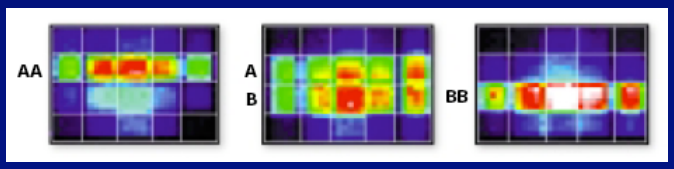


Hybridization



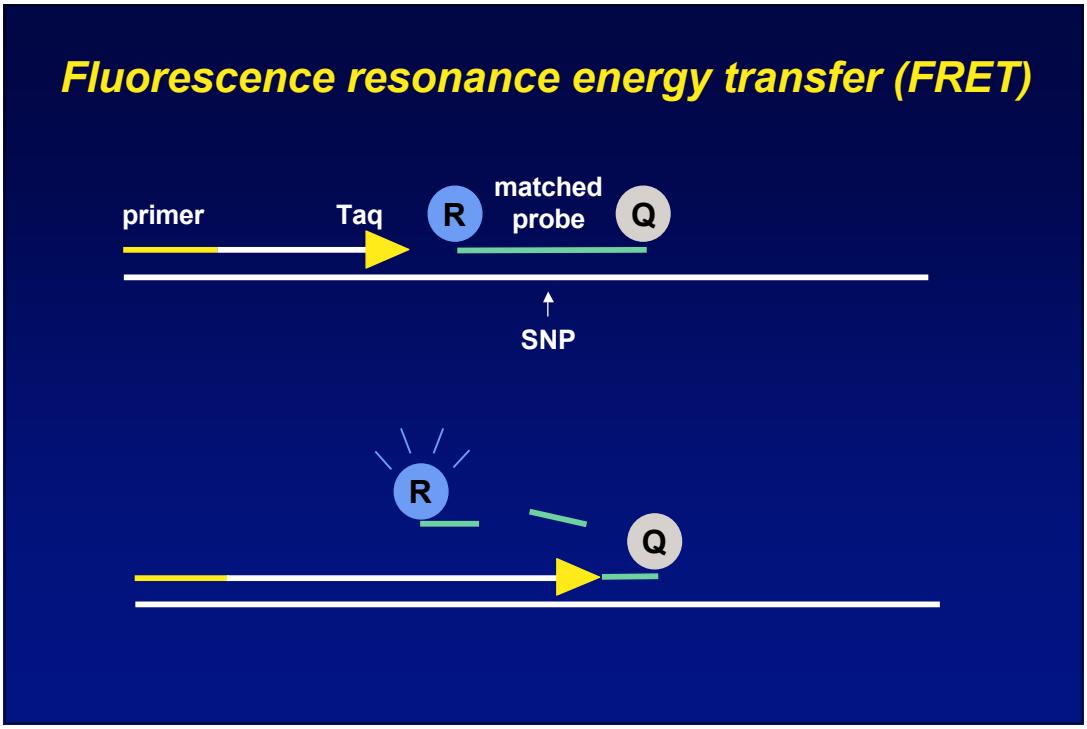
Affymetrix HuSNP Mapping Assay

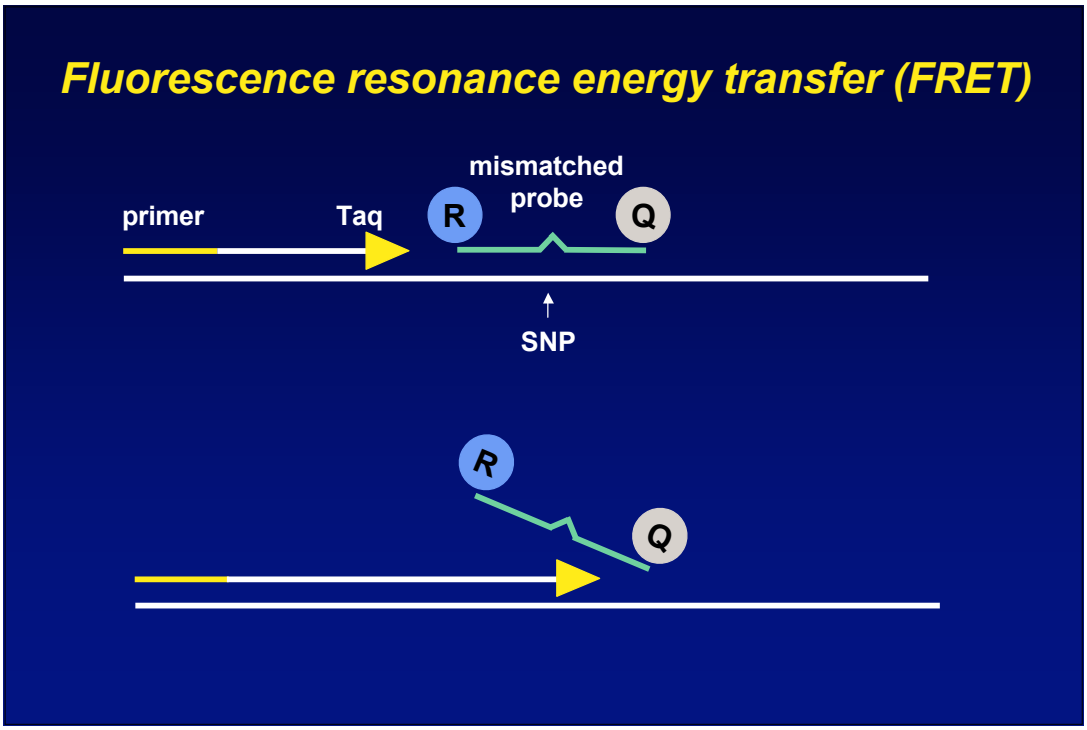


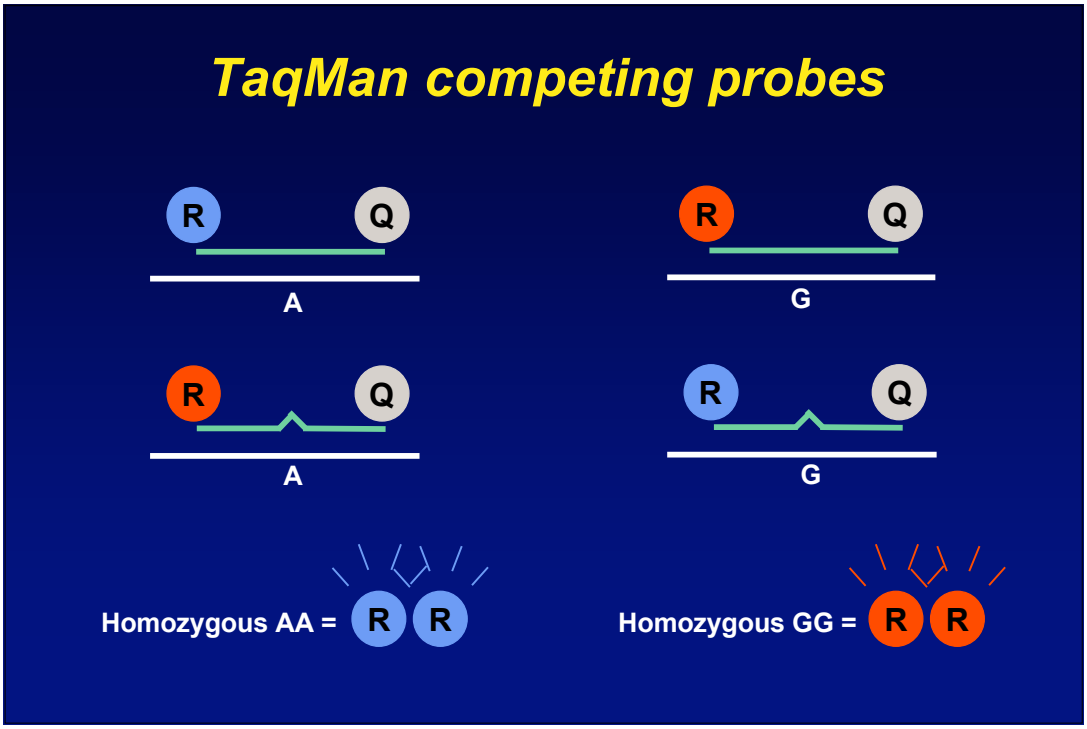


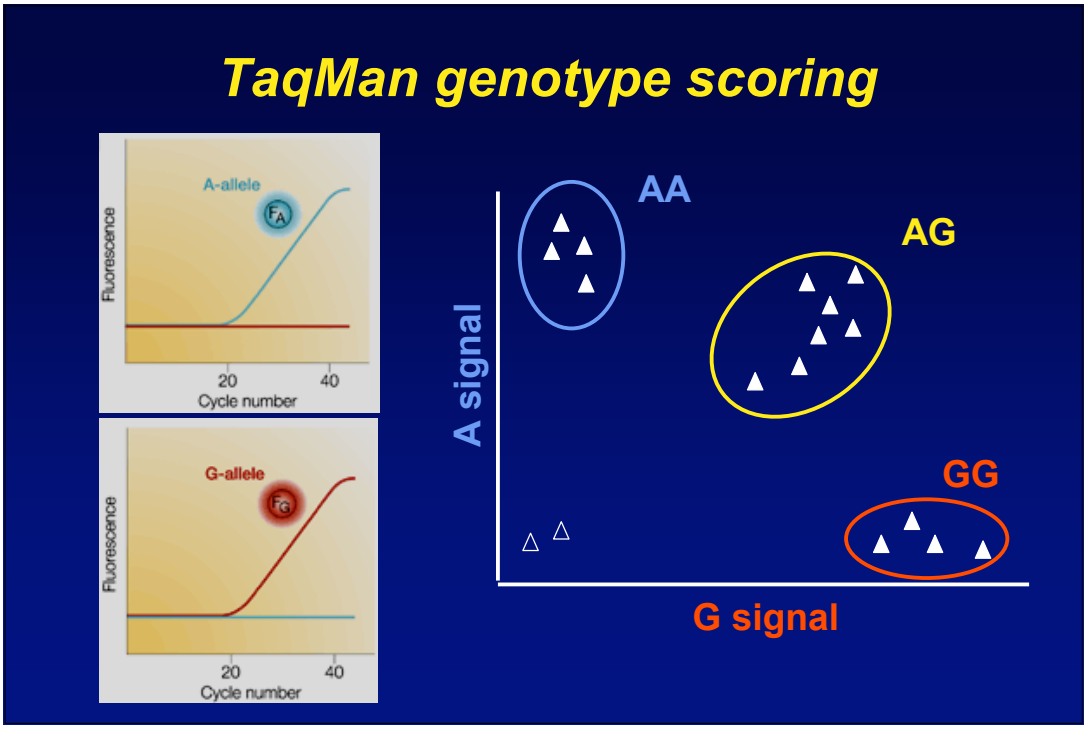
Hybridization to Oligonucleotide Arrays

- **Advantages:**
  - **Simple to perform**
  - **Highly multiplexed**
  - **Automated analysis**
- **Disadvantages**
  - **Expensive to design/create chip**
  - **Local sequence affects success**


Fluorescence resonance energy transfer (FRET)
primer
Taq
R
matched
probe
Q
SNP
R
Q


Fluorescence resonance energy transfer (FRET)
primer
Taq
R
mismatched
probe
Q
SNP
R
Q

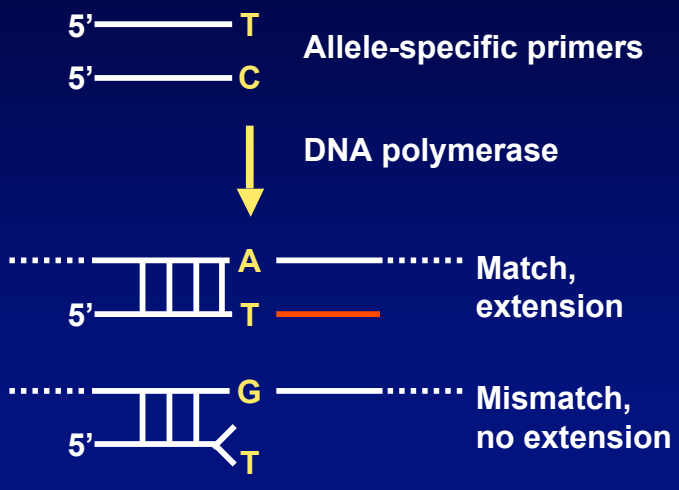

TaqMan competing probes
R
Q
A
R
Q
G
R
Q
A
R
Q
G
Homozygous AA =
R
R
Homozygous GG =
R
R

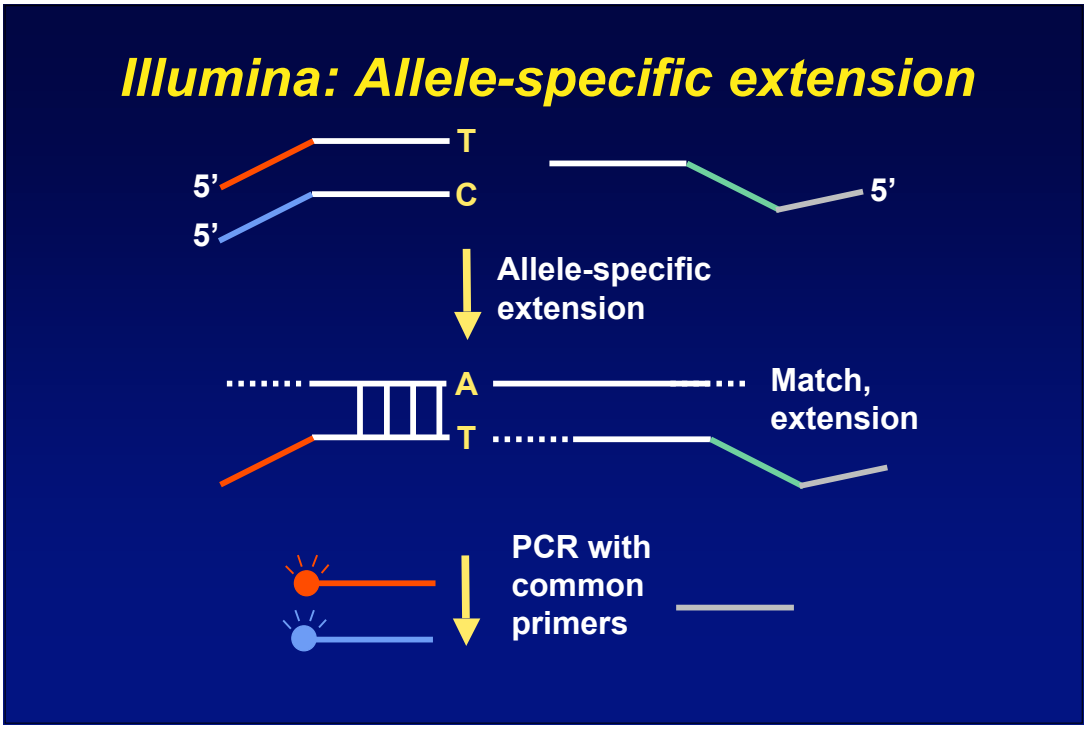

TaqMan genotype scoring
A-allele
Fluorescence
20
40
Cycle number
G-allele
Fluorescence
20
40
Cycle number
AA
AG
GG
A signal
G signal

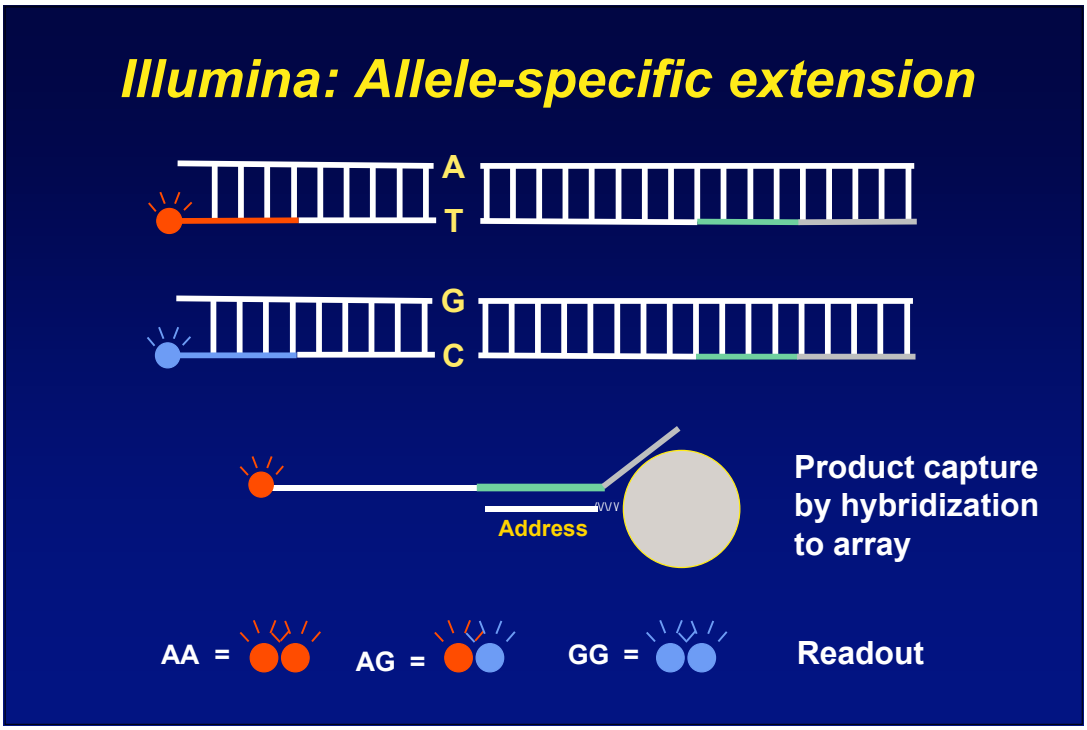
TaqMan

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

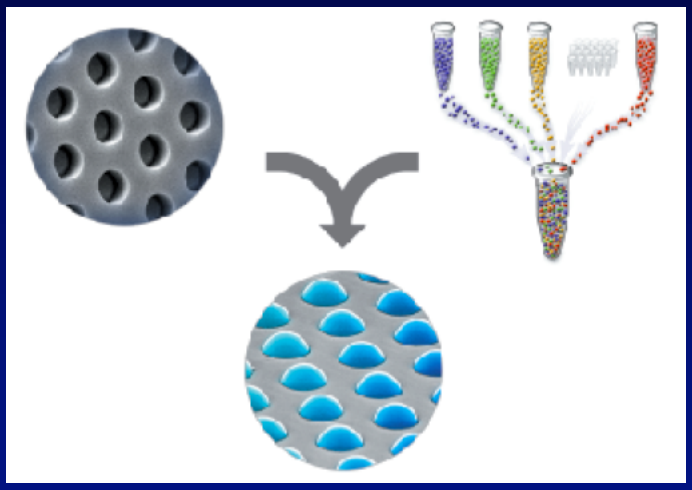
Allele-specific PCR




Illumina: Allele-specific extension
T
C
5'
5'
5'
Allele-specific extension
A
T
Match, extension
PCR with common primers


Illumina: Allele-specific extension
A
T
G
C
Address
Product capture by hybridization to array
AA =
AG =
GG =
Readout

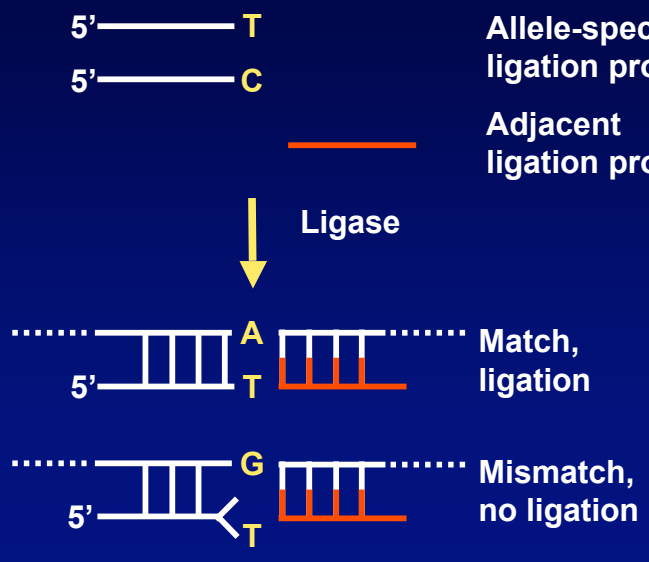
Illumina genotyping technology



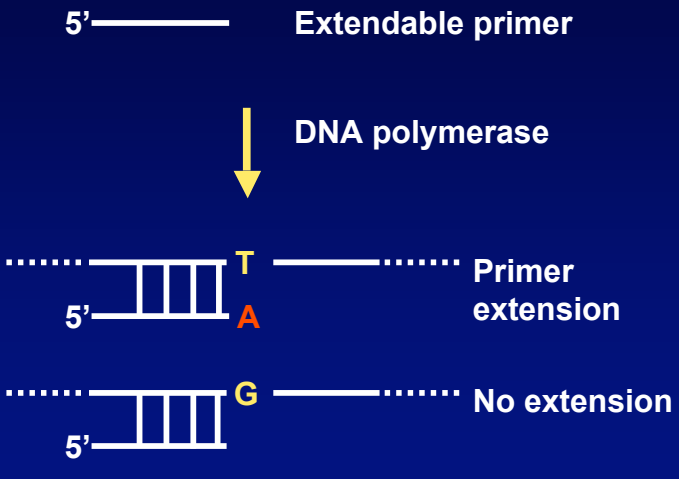
Illumina

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

Oligonucleotide Ligation Assay (OLA)



Primer extension = Minisequencing

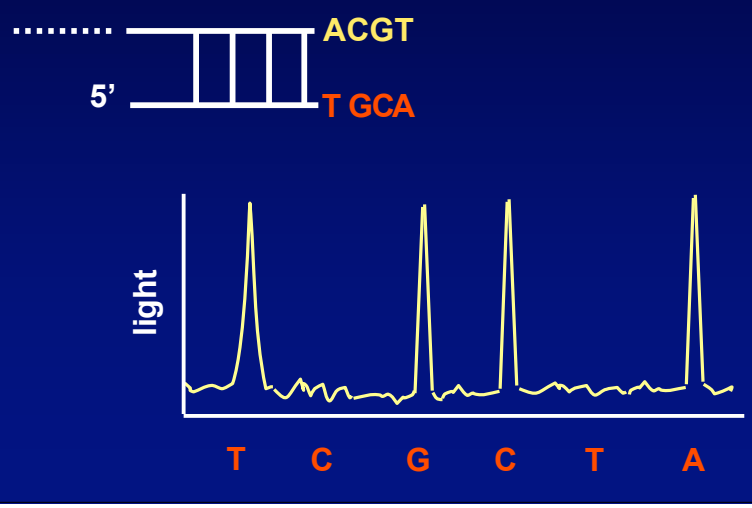


Pyrosequencing

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

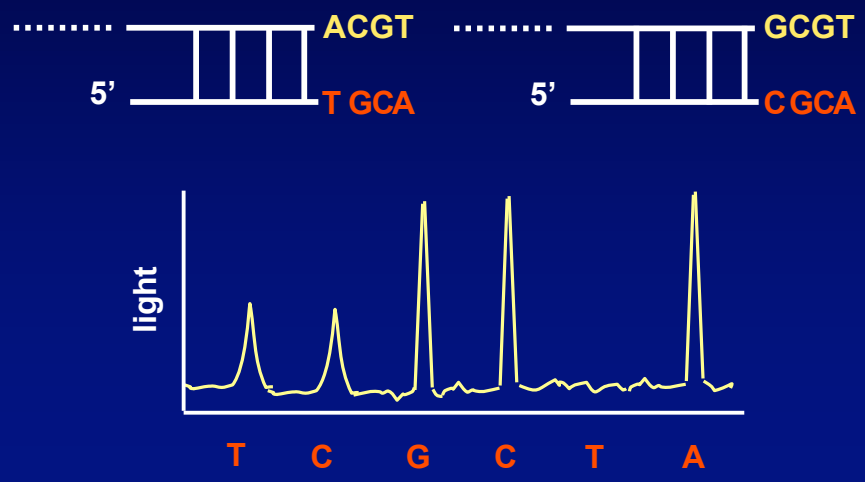
Pyrosequencing

…[A/G]CGT…



Pyrosequencing

…[A/G]CGT…



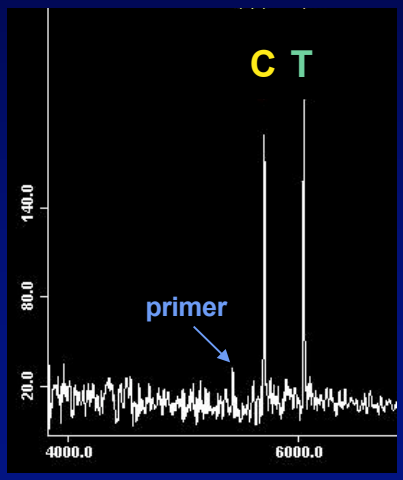
Pyrosequencing

- **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

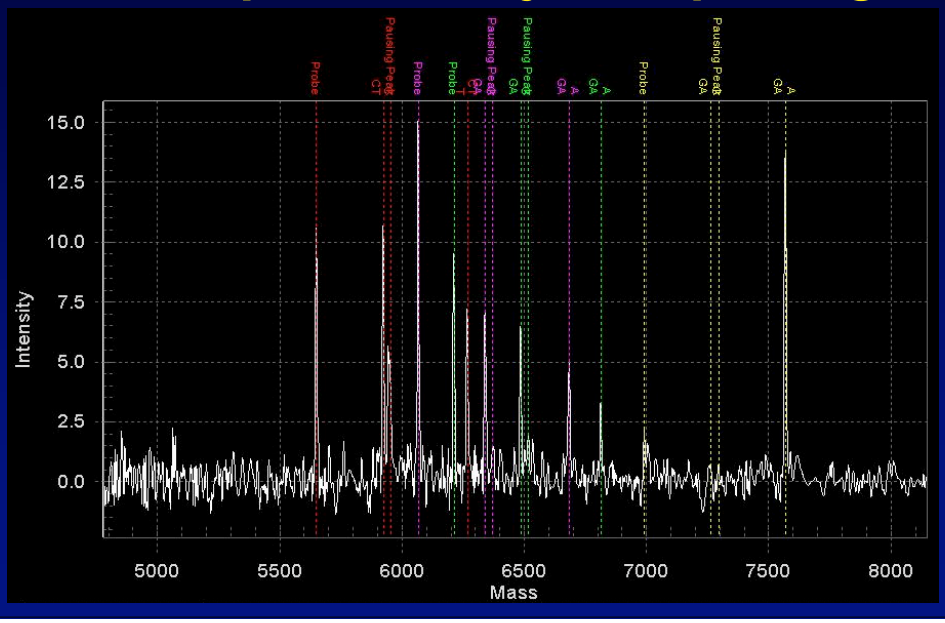
	Mass in Daltons
GGACCTGGAGCCCCCACC	5430.5
GGACCTGGAGCCCCCACCC	5703.7
GGACCTGGAGCCCCCACCTG	6047.9



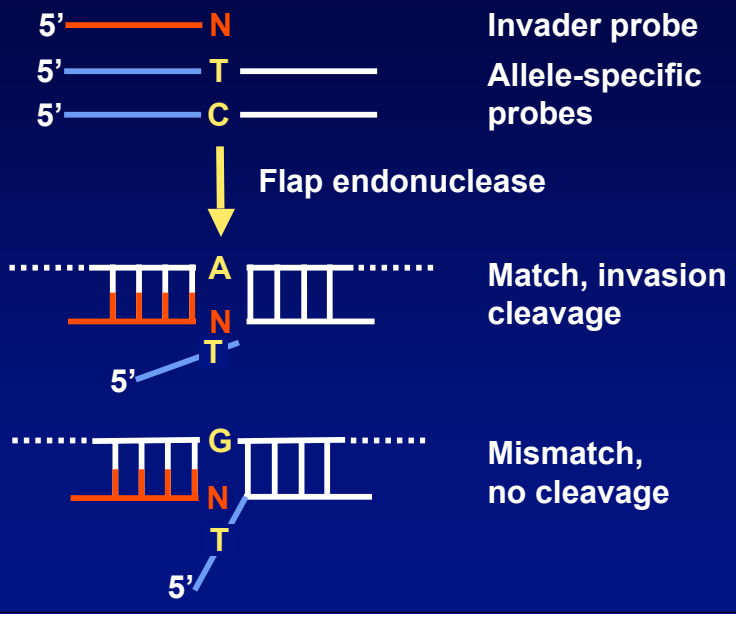
Primer extension mass spectrometry

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

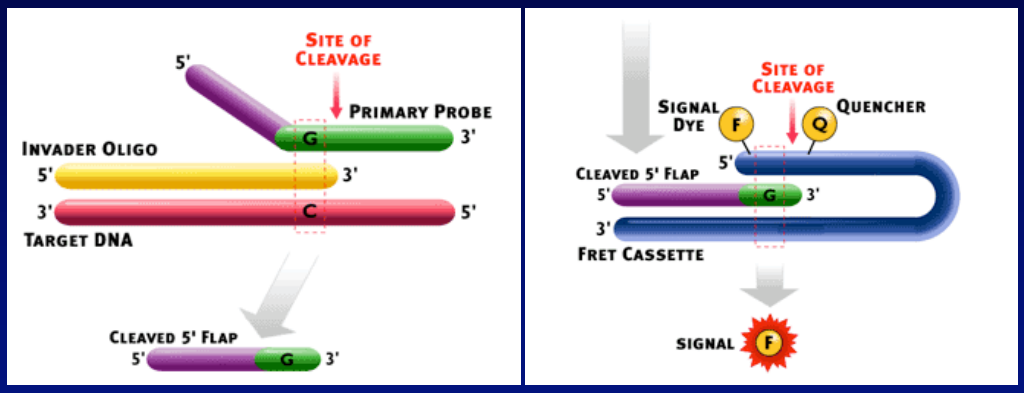
Mass spectrometry multiplexing



Invasive cleavage of oligo probes



Invader ® Technology



Third Wave Technologies, Inc.

Invasive cleavage of oligo probes

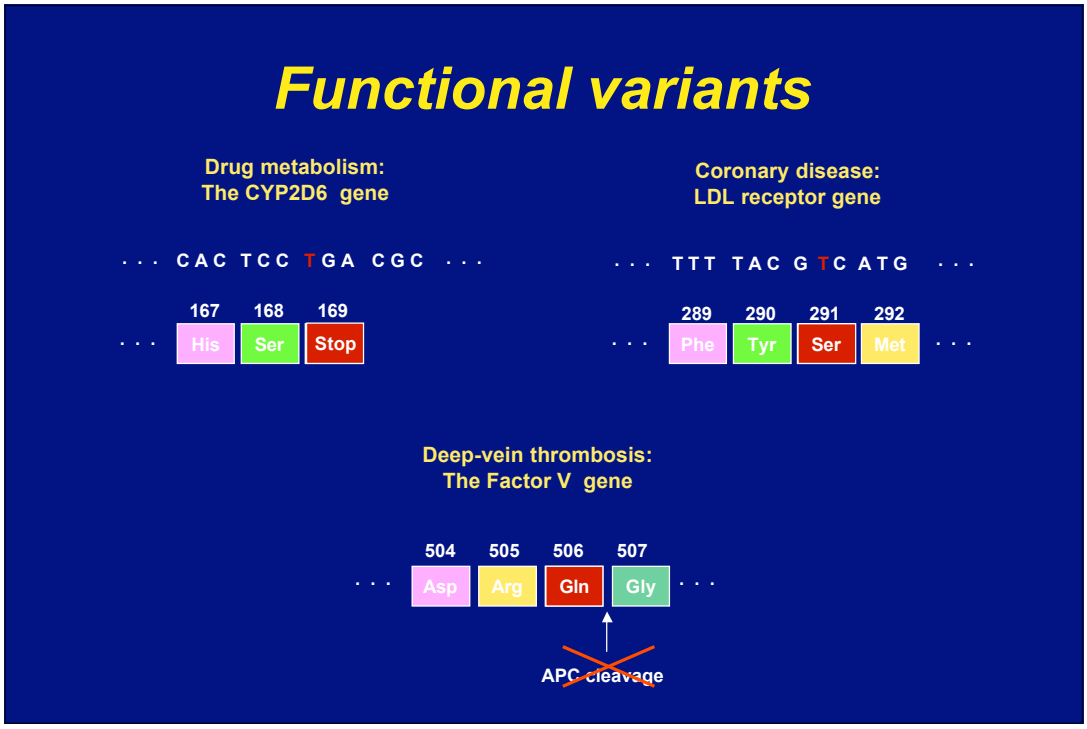
- Advantages
  - Avoids need for PCR
- Disadvantages
  - Still requires larger amount of DNA
  - Tricky probe design

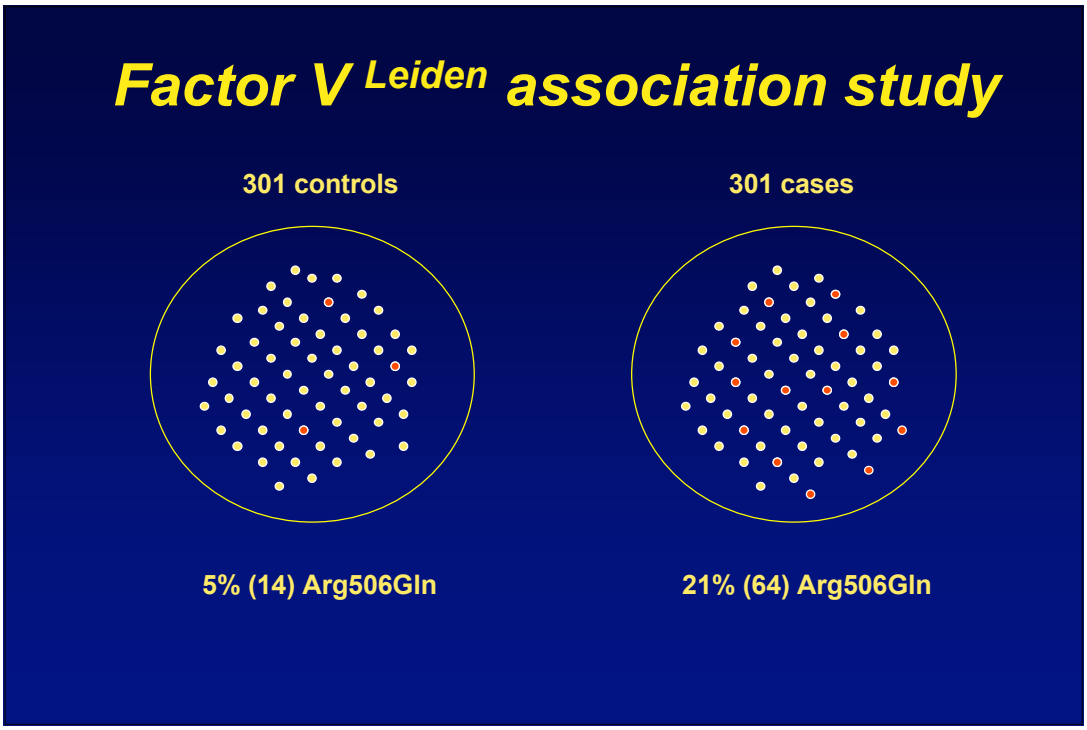
Quality control of genotype data

- High genotype success
- Accurate duplicate genotypes
- Consistent with Hardy-Weinberg Equilibrium: $p^2 + 2pq + q^2 = 1$
- Accurate on a second platform

Human Genetic Variation

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

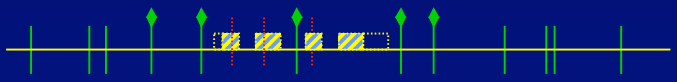

Functional variants
Drug metabolism:
The CYP2D6 gene
. . . CAC TCC TGA CGC . . .
167 168 169
. . . His Ser Stop
Coronary disease:
LDL receptor gene
. . . TTT TAC G TC ATG . . .
289 290 291 292
. . . Phe Tyr Ser Met . . .
Deep-vein thrombosis:
The Factor V gene
504 505 506 507
. . . Asp Arg Gln Gly . . .
APC cleavage


Factor V Leiden association study
301 controls
301 cases
5% (14) Arg506Gln
21% (64) Arg506Gln

Association Studies

Direct

Indirect



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	167	148
non-risk allele	133	152

$$\text{odds ratio} = \frac{167*152}{148*133} = 1.33$$

Disease is 1.33 times as frequent with risk allele

Example case-control association study

500 cases
500 controls
Prior evidence suggests 10 Mb candidate region

In 10Mb, expect ~10,000 SNPs, ~100 genes

Need:
- Efficient way to screen SNPs
- Knowledge of most useful SNPs

Screen SNPs using pooled DNA

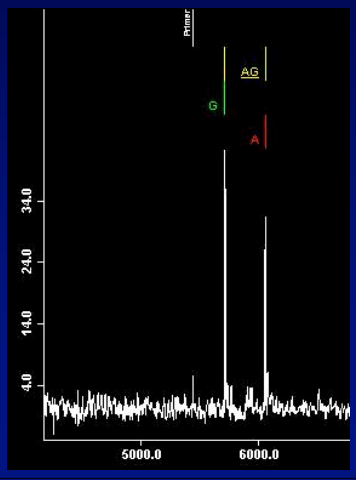
500 cases one pool
500 controls one pool
10,000 SNPs

Direct analysis: 10,000,000 genotypes
Pooled DNA analysis: 20,000 genotypes

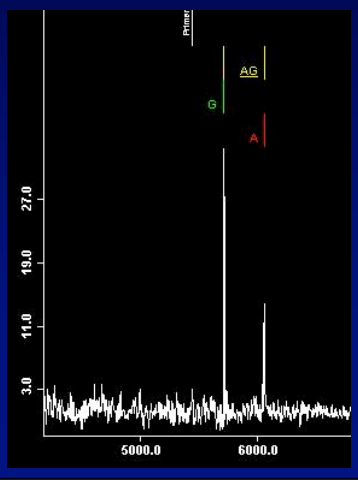
Genotyping of DNA pools

- Create equimolar pools of individual DNAs
- Type SNP and determine relative allele frequencies

Affected cases



Unaffected controls



Example case-control association study

500 cases
500 controls
10 Mb candidate region

In 10Mb, expect ~10,000 SNPs, ~100 genes

Need:
- **Efficient way to screen SNPs**
- **Knowledge of most useful SNPs**

Variation at adjacent sites tends to correlate

GAAATAATTAATGTTTTCCTTCCTTCTCCTATTTTGTCCTTTACTTCAATTTATTTTTTTTATTATTAATATTATTATTTTTTTGAGACGGAGTTTCACTCTTGT
TGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCACACTCCGCTTTC[C/T]GGTTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGAC
TACAGTCACACACCACCACGCCCGGCTAATTTTTGTATTTTTAGTAGAGTTGGGGTTTCACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCC
GCCAGCCTCTGCCTCCCAAAGAGCTGGGATTACAGGCGTGAGCCACCGCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTTTCTTTGCCTGGACTTTACA
AGTCTTACCTTGTTCTGCCTTGGATATTTGTGTGGTCTCATTCTGGTGTGCCAGTAGCTAAAAATCCATGATTTGCTCTCATCCCACTCCTGTTGTTCATCT
CCTCTTATCTGGGGTCAC[A/C]TATCTCTTCGTGATTGCATTCTGATCCCCAGTACTTAGCATGTGCGTAACAACTCTGCCTCTGCTTTCCCAGGCTGTTGA
TGGGGTGCTGTTCATGCCTCAGAAAAATGCATTGTAAGTTAAATTATTAAAGATTTTAAATATAGGAAAAAAGTAAGCAAACATAAGGAACAAAAAGGAAAGA
ACATGTATTCTAATCCATTATTTATTATACAATTAAGAAATTTGGAAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAAGTCTTTTACTCTTTAC
AAAATACATGTGTTAGCAATTTTGGGAAGAATAGTAACTCACCCGAACAGTGTAATGTGAATATGTCACTTACTAGAGGAAAGAAGGCACTTGAAAAACATCT
CTAAACCGTATAAAAACAATTACATCATAATGATGAAAACCCAAGGAATTTTTTTAGAAAACATTACCAGGGCTAATAACAAAGTAGAGCCACATGTCATTTA
TCTTCCCTTTGTGTCTGTGTGAGAATTCTAGAGTTATATTTGTACATAGCATGGAAAAATGAGAGGCTAGTTTATCAACTAGTTCATTTTTAAAAGTCTAACA
CATCCTAGGTATAGGTGAACTGTCCTCCTGCCAATGTATTGCACATTTGTGCCCAGATCCAGCATAGGGTATGTTTGCCATTTACAAACGTTTATGTCTTAAG
AGAGGAAATATGAAGAGCAAAACAGTGCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTGGAAAAATTGAGAAACTACTCATTTTCTA
AATTACTCATGTATTTTCCTAGAATTTAAGTCTTTTAATTTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGATGGTATGAGTAATTAATATCTGT
TATATAATATTCATTTTCATAGTGGAAGAAATAAAATAAAGGTTGTGATGATTGTTGATTATTTTTTCTAGAGGGGTTGTCAGGGAAAGAAATTGCTTTTTTT
CATTCTCTCTTTCCACTAAGAAAGTTCAACTATTAATTTAGGCACATACAATAATTACTCCATTCTAAAATGCCAAAAAGGTAATTTAAGAGACTTAAAACTG
AAAAGTTTAAGATAGTCACACTGAACTATATTAAAAAATCCACAGGGTGGTTGGAACTAGGCCTTATATTAAAGAGGCTAAAAATTCAATAAGACCACAGGC
TTTAAATATGGCTTTAAACTGTGAAAGGTGAAACTAGAATGAATAAAATCCTATAAATTTAAATCAAAAGAAAGAAACAAACT[A/G]AAATTAAAGTTATTA
TACAAGAATATGGTGGCCTGGATCTAGTGAACATATAGTAAAGATAAAACAGAATATTTCTGAAAAATCCTGGAAAATCTTTTGGGCTAACCTGAAAACAGTA
TATTTGAAACTATTTTTAAAATGCAGTGATACTAGAAATATTTTAGAATCATATGTA

[C/T] [A/C] [A/G]

Linkage disequilibrium

[C/T] [A/C] [A/G]

CAA TAA
CAG TAG
CCA TCA
CCG TCG

Linkage disequilibrium

[C/T] [A/C] [A/G]

CAA

TCG

Genotype only the most useful SNPs

500 cases one pool
500 controls one pool
~~10,000~~ SNPs
1,000 'haplotype tag' SNPs

Direct analysis:	10,000,000 genotypes
Pooled DNA analysis:	20,000 genotypes
Selected SNPs:	2,000 genotypes

Future

- Continued identification of SNPs
- Faster, cheaper, easier genotyping
- Genome haplotype map
- SNP panel(s) for association studies
- Discovery of new functional variants

Websites

GDB	http://www.gdb.org/
Marshfield	http://research.marshfieldclinic.org/genetics/
CHLC	http://gai.nci.nih.gov/CHLC/
Genethon	http://www.genethon.fr/php/index_us.php
Sputnik	http://espressosoftware.com/pages/sputnik.jsp
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
TSC	http://snp.cshl.org/
HGVbase	http://hgvbase.cgb.ki.se
CGAP	http://cgap.nci.nih.gov/
JSNP	http://snp.ims.u-tokyo.ac.jp/
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
SNPper	http://snpper.chip.org/

References

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