

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...

```
GGCATCTTTGTTACTCTGCTCAACATTCAAAGTCCAGGGGAGAATATTATTAGTTGGGCTTAGGTACATGCCACATGGCTGTACTGGGATGAGA
GAGAAGGAATCCGATGAAAGGAGCCACAGTAACCCCTTCGCTTCTGTTATTTGGGGCAAGACACACCAATCTGCATACACAGCTCTGAAAACAATG
GGGGAGAGATTTCCTAAAAGGAACTAGGATGTTATTACTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTTATTTT
CAGTGGTGCATTTTCAGTCTACTGCAACCTCTGCCTCCAGGTTCAAGTGATTCTCCTGCCTCAGCCTCCCCCATAGCTGGAATTCAGGCATGTGCC
ACCATGCCAGCTAATTTTTTTGATTTTTAGTAGAGATGGGTTCCACCATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCAGTGATCCGCCCA
CCTCGGCCCTCCAGAGTGTGGGATTACAGTTGTGAGCCACCATGTCCGGCCCTAGGATATTTTCAATTAAGAAAAGAATGCTGGATAGCCAAAGTGAA
AATACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
AACATCAGAACTTTTCATCTTTGAAGGCACAAAGAGTTAGTATTCACAGAGGATAGCTAATCTTATCTCTCCTCTCGGAGGTTTCAGAAAATGTTTAT
CTCATCCTGGGGAAAGCCAGATGATAACGTTCAATGGAGCAAGAAAAGGTCACACAAAATGAGGTGTCTTACAAAACAAATGGAAATTTTCATATCCT
GCTCAAAGGGCCAGAGGATATTTCCCAATAAAGCATTTGTCGAGGGATGAATGAGATAGGATCTAGACCTCTGAGTATGATAAATGGTTAGTTCT
TCCTATTAGTTGTTGTTCTGATGTAGAAAACAGGCTTTTCTCCCTATATCTGGTCTAAAAATCCAACTGATAGGAGACGTTTTTCGTTTGGGATTATGG
AAAGATACACAGTTCTGGGGTTGAGTTCAGGGCTAATTTTCTGAAGGATAAGAGAGCAAGCCCAAGCCAAAGAGCCAAGAGAAAAGCAATGATGAGGAA
CGGGCAGTAGCAGCCATTAGACTGGTGTCTTTGTTGGACTCCCTTCTATTGTCATATTATAGGCTTTCCAAACAGGGGACAAATAACAGTATGAATC
CAGACAGGATGAGGTGGTTGCACAAGCAGCTGGGCCACTGAACTAGAGCCTGACTCAAAAAGGAAGGAGGCTGGGGCAGTGGCTCACACCTGTA
ATCCACAGCTTTGGGAGCCGAGGGGTGGATCAGAGGCTGGAGTTCGAGACRAGCCTGGCCRAATATGGTGAACCCCATAGCTACTAAAATAC
AAAAATTAGCCAGGATGGTGGCAGGCACCTGTAGTCCAGCTACTCGGAGGCTGAGGCAGAAGAACTACTGAACTGGGAGGTGGAGGTTGCAGTG
AGCTGAGATTGTCCTGCACTCCAGCCTGGTGACAGGCAAGACTCCATCTCAAAAAAAAAAAAAAAAAAAGGAAGATCTGCCATGGTGTAGGA
CCCCCATCCGTTCTCTGGTCCAGTCAAGCTGTGCCCATTTGACTGGGGCATGATGCACTTCTTGTATCCGGTAGCATGTTCCAGGCCAGGG
AGTGTCCAGGCAGTGCATCAGATTATCAGGCATTGACCAGAGATACCTATAAGCTGAGAGCTACAGCCATTTTGGCAAGCTCTGAAAACCCAGAGTTGG
CGCTGTTCAATGGGGAGGATCTGCATGGTACTCGCTGAGCCGATGTTTTTGTGTTCTGTTTGGAAAGCCTACACATATGTGTTAAACCATCCCTA
TGCATCATTAGCCTGCT
```

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

More typical sequence ...

```
GAAAAAATAATTAAGTTTTCCCTTCCTCCTCAATTTTGGCTTACTTCAATTTATTTATTTATTTATTAATATATTTATTTTTTTGAGACGGAGTTTCACTCTTGT  
TGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCAGTGCACACACCCGCTTTCGGTTTTCAAGCGATTCTCCTGCCTCAGCTCCTGAGTAGTGGGACTACA  
GTCACACACCACCACGCCCCGGCTAATTTTTGTATTTTTAGTAGAGTTGGGGTTTCCACATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCA  
GCCCTGCCTCCCAAAGAGCTGGGATTACAGGCGTGAGCCACCGCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTTTTCTTTGGCTGGACTTTACAAGTC  
TTACCTTGTCTGCCTTCAGATATTTGTGTGCTCATTCTGGTGTGCCAGTAGTAAAAATCCATGATTTGCTCTCATCCACCTCCTGTTGTTTCATCTCCTC  
TTATCTGGGTACATATCTCTTGGTATTGCAATCTGATCCCACTACTTAGCATGTGCGTAACTCTGCCTCTGCTTTCCAGGCTGTTGATGGGGTGC  
TGTTCATGCCCTCAGAAAAATGCAATTTAAGTAAATTTAAAGATTTAAATATAGGAAAAAAGTAAAGCAACATAAGGAAACAAAAAGGAAAGACATGTAT  
TCTAATCCATTTATTTATACAATTAAGAAATTTGGAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAAGTCTTTTACTCTTTACAAAATACA  
TGTGTTAGCAATTTGGGAAGAAATAGTAACACCCGAACTGTAATGTGAATATGTCACCTTACTAGAGGAAAGAGGCACTTGA AAAACATCTCTAAACCG  
TATAAAAACAATTCATCATAATGATGAAACCCCAAGGAATTTTTTAGAAAACATACCAGGGCTAATAACAAAGTAGAGCCACATGCTATTTACTCTCCCT  
TTGTCTGTGTGAGAAATCTAGACTTATTTGTACATAGCATGGA AAAATGAGAGGCTAGTTTATCAACTAGTTCAATTTTAAAAGTCTAACACATCCTAG  
GTATAGTGAACCTGCTCCTGCCAATGTATTGCACATTTGTGCCAGATCCAGCATAGGGTATGTTGGCATTTCACAAACGTTTATGTCTTAAGAGAGGAAA  
TATGAGAGCAAAAACAGTGCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTTGAAAAATTTGAGAACTACTCATTCTTAAATTACTC  
ATGATTTTTCCTAGAATTTAAGTCTTTAATTTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGATGGTATGAGTAATTAATATCTGTTATATAAT  
ATTCATTTTCATAGTGGAAAGAAATAAAATAAGGTTGTGATGATGTTGATTTATTTTTCTAGAGGGTGTGTCAGGAAAGAAATGCTTTTTTTCATCTCTC  
CTTTCCACTAAGAAAGTTCACTATTAATTTAGGCACATACATAAATTTACTCATTCTAAAATGCCAAAAAGGTAATTTAAGAGACTTAAAACGAAAAATTT  
AAGATAGTCACTGAACTATTTAAAATCCACAGGGTGTGGAAC TAGGCCTTATATTAAGAGGCTAAAATTTGCAATAAGCCACAGGCTTTAAATA  
TGGCTTTAACTGTGAAAGGTGAACTAGAATGAATAAATCCTATAAATTTAAATCAAAAAGAAACAACTGAAATTTAAAGTTATTTATACAAGAAATAG  
GTGGCTGGATCTAGTGAAACATATAGTAAAGATAAAACAGAATATTTCTGAAAAATCCTGAAAAATCTTTTGGGCTAACCTGAAAAACAGTATATTTGAAACTA  
TTTTTAAAATGCACTGATCTAGAAATATTTAGAACTATATGTA
```

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

Single nucleotide polymorphisms (SNPs)

```
GAAAAAATAATTAAGTTTTCCCTTCCTCCTCAATTTTGGCTTACTTCAATTTATTTATTTATTTATTAATATATTTATTTTTTTGAGACGGAGTTTCACTCTTGT  
TGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCAGTGCACACACCCGCTTTCGGTTTTCAAGCGATTCTCCTGCCTCAGCTCCTGAGTAGTGGGACTACA  
GTCACACACCACCACGCCCCGGCTAATTTTTGTATTTTTAGTAGAGTTGGGGTTTCCACATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCA  
GCCCTGCCTCCCAAAGAGCTGGGATTACAGGCGTGAGCCACCGCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTTTTCTTTGGCTGGACTTTACAAGTC  
TTACCTTGTCTGCCTTCAGATATTTGTGTGCTCATTCTGGTGTGCCAGTAGTAAAAATCCATGATTTGCTCTCATCCACCTCCTGTTGTTTCATCTCCTC  
TTATCTGGGTACATCTCTTGGTATTGCAATCTGATCCCACTACTTAGCATGTGCGTAACTCTGCCTCTGCTTTCCAGGCTGTTGATGGGGTGC  
TGTTCATGCCCTCAGAAAAATGCAATTTAAGTAAATTTAAAGATTTAAATATAGGAAAAAAGTAAAGCAACATAAGGAAACAAAAAGGAAAGACATGTAT  
TCTAATCCATTTATTTATACAATTAAGAAATTTGGAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAAGTCTTTTACTCTTTACAAAATACA  
TGTGTTAGCAATTTGGGAAGAAATAGTAACACCCGAACTGTAATGTGAATATGTCACCTTACTAGAGGAAAGAGGCACTTGA AAAACATCTCTAAACCG  
TATAAAAACAATTCATCATAATGATGAAACCCCAAGGAATTTTTTAGAAAACATACCAGGGCTAATAACAAAGTAGAGCCACATGCTATTTACTCTCCCT  
TTGTCTGTGTGAGAAATCTAGACTTATTTGTACATAGCATGGA AAAATGAGAGGCTAGTTTATCAACTAGTTCAATTTTAAAAGTCTAACACATCCTAG  
GTATAGTGAACCTGCTCCTGCCAATGTATTGCACATTTGTGCCAGATCCAGCATAGGGTATGTTGGCATTTCACAAACGTTTATGTCTTAAGAGAGGAAA  
TATGAGAGCAAAAACAGTGCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTTGAAAAATTTGAGAACTACTCATTCTTAAATTACTC  
ATGATTTTTCCTAGAATTTAAGTCTTTAATTTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGATGGTATGAGTAATTAATATCTGTTATATAAT  
ATTCATTTTCATAGTGGAAAGAAATAAAATAAGGTTGTGATGATGTTGATTTATTTTTCTAGAGGGTGTGTCAGGAAAGAAATGCTTTTTTTCATCTCTC  
CTTTCCACTAAGAAAGTTCACTATTAATTTAGGCACATACATAAATTTACTCATTCTAAAATGCCAAAAAGGTAATTTAAGAGACTTAAAACGAAAAATTT  
AAGATAGTCACTGAACTATTTAAAATCCACAGGGTGTGGAAC TAGGCCTTATATTAAGAGGCTAAAATTTGCAATAAGCCACAGGCTTTAAATA  
TGGCTTTAACTGTGAAAGGTGAACTAGAATGAATAAATCCTATAAATTTAAATCAAAAAGAAACAACTGAAATTTAAAGTTATTTATACAAGAAATAG  
GTGGCTGGATCTAGTGAAACATATAGTAAAGATAAAACAGAATATTTCTGAAAAATCCTGAAAAATCTTTTGGGCTAACCTGAAAAACAGTATATTTGAAACTA  
TTTTTAAAATGCACTGATCTAGAAATATTTAGAACTATATGTA
```

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

Location: http://research.marshfieldclinic.org/genetics/Map_Markers/maps/indexMapFrames.html

1000 North Oak Avenue | Marshfield, WI 54449-5790 | Phone: 715-387-9150 | Fax: 715-389-5757

Marker **Dnumber** **sex-ave(ctf)** **female(ctf)** **male(ctf)**

1 AFM214yg7	D1S243	4.22	0.00	4.46	0.00	3.54	0.00
2 AFM280we5	D1S468	4.63	4.22	2.94	4.46	6.67	3.54
3 AFM344we9	D1S2845	1.93	8.85	3.80	7.40	0.00	10.21
4 AFM123xc3	D1S2893	0.00	10.78	0.00	11.20	0.00	10.21
5 GATA6	Unknown	0.00	10.78	0.00	11.20	0.00	10.21

Microsatellite identification: databases

Location: http://research.marshfieldclinic.org/genetics/Map_Markers/maps/IndexMapFrames.html

1000 North Oak Avenue | Marshfield, WI 54448-5780 | Phone: 715-387-8150 | Fax: 715-388-5757

Information for the General Public

- [Educational Materials](#)

Information for Research Scientists

- [Diallelic Insertion/Deletion Polymorphisms](#)
- [Comparison of Genetic and Physical Maps](#)
- [Genetic Maps](#)
- [Build Your Own Map](#)
- [Search For Markers](#)
- [Mammalian Genotyping Services](#)
- [Complated Genotyping Projects](#)
- [Genotyping Statistics](#)
- [Screening Sets of Markers](#)
- [CEPH Family Genotyping Data](#)
- [Genotypes for Reference](#)

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
AFM016xb3	D1S411	Z23283	0.62	194	204	204	196
AFM044xd2	D1S447	Z23291	0.77	123	141	135	129
AFM024xf8	D1S456	Z23292	0.72	197	211	211	209
AFM031xd12	D1S412	Z23298	0.71	185	207	199	185
AFM036xc5	D1S189	Z16438	0.78	124	136	134	132
AFM042xe3	D1S423	Z23307	0.61	152	167	164	164
AFM046xb10	D1S191	Z16475	0.73	153	169	163	161
AFM051xb8	D1S192	Z16482	0.66	203	211	207	205
AFM057xf4	D1S193	Z16490	0.77	94	106	106	104
AFM057xf8	D1S194	Z16491	0.65	233	239	235	233
AFM059yb4	D1S2796	Z50913	0.68	126	132	126	126

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
AFM016xb3	D1S411	Z23283	0.62	194	204	204	196
AFM044xd2	D1S447	Z23291	0.77	123	141	135	129
AFM024xf8	D1S456	Z23292	0.72	197	211	211	209

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
tcttagtagaataagatccagtagtatagacacttttgcggcatccaaagaattaacc
cttcactcatttactcacctggtaagagatacaggggaaaagctgtggagtaactcagg
agctggagccataaggcaggaaaccatgccattcattcaaaaacttgattgagct
cctttttagtgcacccccatccactataagcacttggagaccacacagatgtggttcc
tgcctccatcgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt
gtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt
aggggaatgtaaacaggaaaacagatatgcaaaaacaatttcagatcgc
ggtaagtgtctaggaacagaatgaaataggataggagtgatggacaggggagacttcagg
ggagtcacatcgggaaaagcctccataaaagtaccttctgggagaaaaccgagggtaag
aatctggtcctgcaaagatctgggcaagaaatgtccagggttagggaacagcgaggtcaa
agtcaccatcacaaaggaaaccg
```


Marker retrieval: genome browser

<http://genome.ucsc.edu/>

UCSC Genome Bioinformatics

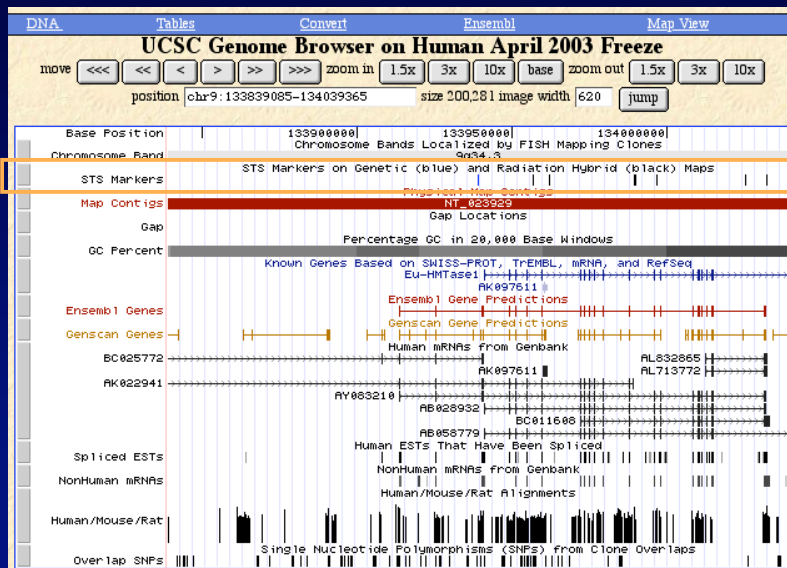
Genome Browser - Blat Search - Table Browser - FAQ - User Guide

genome assembly position image width
Human April 2003 0951838 620 Submit

[Click here to reset](#) the browser user interface settings to their defaults.

Add Your Own Tracks

Marker retrieval: genome browser



Marker retrieval: genome browser

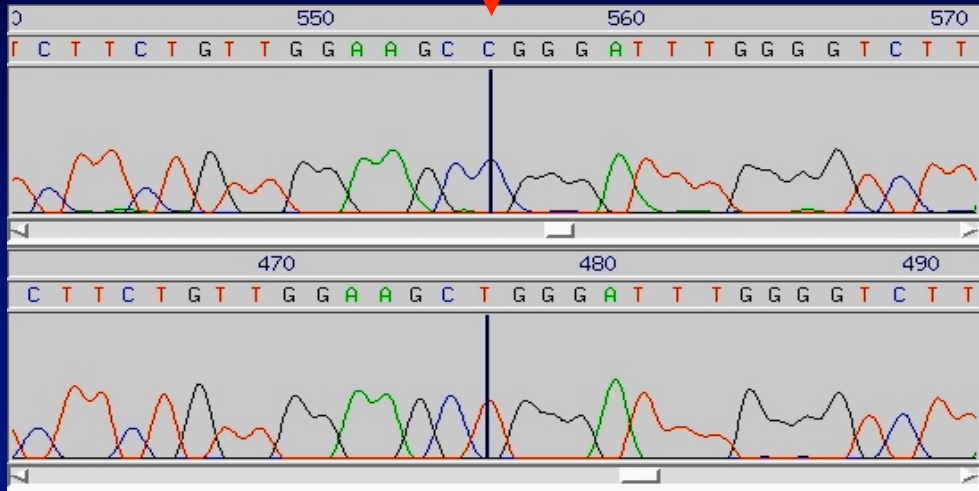
The screenshot shows a web interface titled "Get DNA in Window". It includes a "Get DNA for" section with a "Position" field containing "chr9:133839085-134039365". Below this are "Sequence Retrieval Region Options" with fields for "Add" extra bases upstream and downstream, both set to 0. The "Sequence Formatting Options" section has radio buttons for "All upper case" (selected), "All lower case", and "Mask repeats" (with sub-options for "to lower case" and "to N"), and a checkbox for "Reverse complement (@ '-' strand sequence)". There are two buttons: "Get DNA" and "Extended case/color options". A note at the bottom states: "Note: The 'Mask repeats' option applies only to 'Get DNA', not to".

```
>hg15_dna range=chr9:133839085-134039365 5'pad=0 3
GTATTGCTTCTGCTCTCTGCTCTTTTGTACTTCTGCTCAGTT
AGTTTTTGTCTTCATGGAGATGGAGTTCACTATGTTGCCAGGCTGG
GCGCAGTGAATTCATAGGATGATCATAAGTGCACCGTGGCTTGAAC
CCGGGCTCACAGGATCCTCCACCTCAGGCTCCTGAGTAGCTGGGACT
CAGGCTACCAACAGGCTGACTGATTGCTTTCTTTTGTGTAGATA
TTTTCTAGTATACCAATTTAAATCCCTGCTGTTTATGACTATAT
TTTTTTTTTTCTTGGTGTCTGGGATTAACAATTAATTAATTAAT
TTTTGATAATGTAGTTGATTAATACTATATTACAACTATATGAAAT
ATTTATTTGGGCATAGCTTAATTTCTTTGGCCCTTGTACTGTATTG
TCATACAAATACATTACACTGTGTGCCCATACAGCAGATTTATAGTTG
CTGCTTACGCAATTGTCTTCAAATCATATAGGAGAAAAACAATTA
AAATTTACAAATACATTTATACTGTCTATGTAGCTGCTTTTATACTTA
CCTGTGACTGCTTTTACTGGGCTCTGATTGCTTCCCTCCCTCC
CCTCTGACAGGCTCTGCTCTGTACCCAGGTTGGGACTACAGTGGCACA
ATCATASCACTGCAGGCTTGAACCTCCTGGGCTAAAGAGATCTCCTGCT
CAGCCTCCGAGTAGTTGGGACTATGGGTGCTGCCACTATCCCTGCTG
ATTTTTAAATTTTTTTTTGTAGAGATGGGCTCTACTGTGTGGCCAGGC
TGGTCCCACTCCTGGCTTGAAGATCCCTCCTACTGAGTCTTCCAA
GGTGTGCGTTGACGGCATTGAGCCACCGTGGCTGGCTGTCTCT
ATTCAATGGGCTTGAATTAAGTGTCTAGTGTCTTCAATTTACTGAA
GACTCTAGGCTCTTCAATGATGACTCTCTACTTTTGTATCTGGAA
TGTTTTTTTTTTCTTATGAAGTAGTTTTGTGTATAAAGGTTTTTTTT
TTTTTTTTTTTTTGAAGAGATCTCAACCTCCTGCGAGGCTGGAGTGC
AGTGGTGCATCTGGCTCACGCAAGCTCTGCTCCAGGTTACAGCCA
TCTCTGCTGCTCAGGCTCCGAGTAGCTGGGACTACAGGACTACAGGAC
CTGCCATCAGGCTGGCTAATTTTTGTATTTTTAGTAGAGACGGGTTT
CACCGTGTAGCCAGATGCTGGATCTCCTGACTCTGATCTGCCCCA
CCTTGGCTCCCAAGTGTCTGGGATTACAGGCTGAGCCACCGGCGCCGG
CCGTATAAAGGTTTTTAACTACTGAGGATTTGTCTGGCTGACTCCA
TGGTTCTGATGAGGACCAATGCTCATCTCAGGACCTCGTTGATGA
CAGGTTGCTCCTGTTGCTGTTTTGAGATCTCTGCTGTGTTTCA
CCTGATGATGATGTTGAGGTTGGGACTTTTGGCTTTTCTGATTTG
GATTTGTGASCTCTTAATTTTTGACTGACTGCTGCTTCCAGGGA
TTGGGAACTTTGTAGTACTTCACTGAGCATGCTGTCTCTC
ACCCCTGTGAGCTGCTATTGCTGTTAAACAGAGCTTTGCCATACAA
TGAGTTGAGTGTGGATTAAGGGACAGCACTTGAAGGCGCTCTCC
TTCTTGTGAGGATGGGTTGAGCTGTCTTGGTAAAGTGGCAGTACAG
TACGTTGACAGGTTGACTCGGACCGCTGATCCCGCTGCTTGG
```

SNP identification

- Sequencing
- Databases

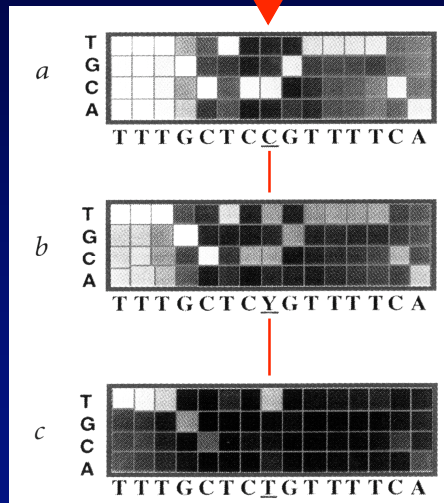
SNP identification: sequencing



SNP identification: sequencing chips



...GCTC**C**GTTT...
...GCTC**T**GTTT...



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

The screenshot shows the dbSNP Home Page in a web browser. The browser address bar displays "http://www.ncbi.nlm.nih.gov/SNP/". The page features the NCBI logo and the title "Single Nucleotide Polymorphism". A navigation menu includes "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "PopSet", "Taxonomy", "OMIM", and "Books", with "SNP" selected. A search bar contains "SNP" and has "Go" and "Clear" buttons. Below the search bar are links for "Limits", "Preview", "Index", "History", "Clipboard", and "Details".

The "dbSNP Search Options" section contains a table with the following columns: "Entrez SNP", "ID Numbers", "Submission Info", "Batch", "Locus Info", "Free Form", "Easy Form", and "Between Markers". A red arrow points to the "Between Markers" column.

An "ANNOUNCEMENT" section contains the following text:

- **NEW!** dbSNP genotype data are now available on the web and on our FTP site ([more info](#)).
- **ALERT!** xml brief and submission format reports are dropped from ftp dump starting build 116. Please contact [snp-admin](#) with concerns.

The "Search by IDs" section includes a note: "Note: *rs#* and *ss#* must be prefixed with 'rs' or 'ss', respectively (i.e. rs25, ss25)". Below the note is a search input field, a "Reference cluster ID(rs#)" dropdown menu, and "Search" and "Reset" buttons.

SNP retrieval: LocusLink

The screenshot shows the LocusLink web interface. At the top, there is a search bar with the query "SLC2A*" and buttons for "Go" and "Clear". Below the search bar, there are navigation tabs for "PubMed", "Entrez", "BLAST", "OMIM", "Map Viewer", "Taxonomy", and "Structure". The main content area displays "16 loci found" and a table of results:

LocusID	Org	Symbol	Description	Position	Links
6513	Hs	SLC2A1	solute carrier family 2 (facilitated glucose transporter), member 1	1p35-p31.3	P, O, R, C, P, H, U, V
81031	Hs	SLC2A10	solute carrier family 2 (facilitated glucose transporter), member 10	20q13.1	P, O, R, C, P, H, U, V
66035	Hs	SLC2A11	solute carrier family 2 (facilitated glucose transporter), member 11		P, O, R, C, P, U, V
154091	Hs	SLC2A12	solute carrier family 2 (facilitated glucose transporter), member 12	6q23.2	P, O, R, C, P, U, V

SNP retrieval: LocusLink

The screenshot shows the LocusLink SNP retrieval interface. The search bar contains "SNP" and the URL is "http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusid=6513". The main content area displays "SNP's linked from LocusLink" and provides detailed information for the gene SLC2A1:

SNP's are linked from Locus **SLC2A1** via the following methods:
[Contig Annotation](#) [GenBank\(mrna\) Mapping](#)

Send the list of rs# to Batch Query. Download the list of rs# to file.

Gene Model (mRNA alignment) information from genome sequence

Total gene model (contig mRNA transcript): 1

Contig	mrna	protein	mrna orientation	snp graph
NT_032977	NM_006516	NP_006507	reverse	transcript on minus strand

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

Contig	mrna	protein	mrna orientation	snp graph	
gene model (contig mRNA transcript):	NT_032977	NM_006516	NP_006507	reverse	transcript on minus strand

Color Legend

Contig position	dbSNP cluster id	Heterozygosity	Validation	3D OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
4098361	rs1803658	N.D.			untranslated region				

SNP retrieval: LocusLink

4999137	rs2229683	0.069			untranslated region				
4999698	rs3831326	N.D.			intron				
4999831	rs2238574	0.083			synonymous	T	Ile [I]	3	300
		0.083			contig reference	C	Ile [I]	3	390
5000932	rs2305663	0.091			intron				
5001059	rs2305662	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
CTCACGTGAC CCAAGACCTG CAGGAGATGA AGGAAGAGAG TCGGCAGATG ATGCGGGAGA
M
GAAGGTACCC ATCCTGGAGC TGTTCGCTC CCCCCTAC CCGCAGCCCA TCCTCATCGC
```

SNP retrieval: SNPper

CHIP Bioinformatics

http://snpper.chip.org/bio/snpper-enter/

SNPper - Main Menu

Goldenpath version: hg15 dbSNP build: 114

[SNPper - Instructions, publications, disclaimers, acknowledgements, copyright.](#)

[Gene Finder - Find a gene by name, symbol, accession number, or position](#)

[SNP Finder - Find SNPs by name or position](#)

[Tools - GeneOntology browser - Amino acid properties - FlankXtender - PrettyBase importer](#)

[Info - SNP plots - RPC interface - Database statistics](#)

Logged in as [guest](#) | [Products](#) | [Help](#) | [Feedback](#) | [Logout](#)

© 2001-2003, Alberto Riva, CHIP

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

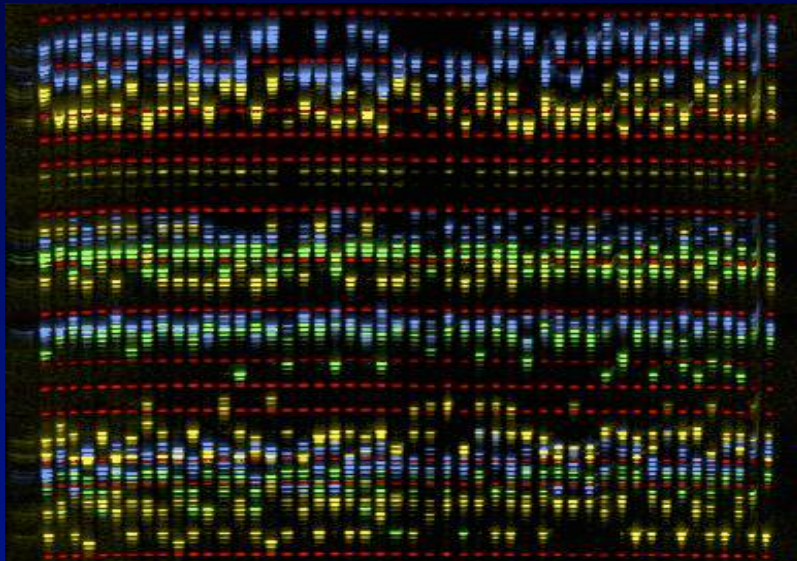
Of the 43,506 coding SNPs:

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

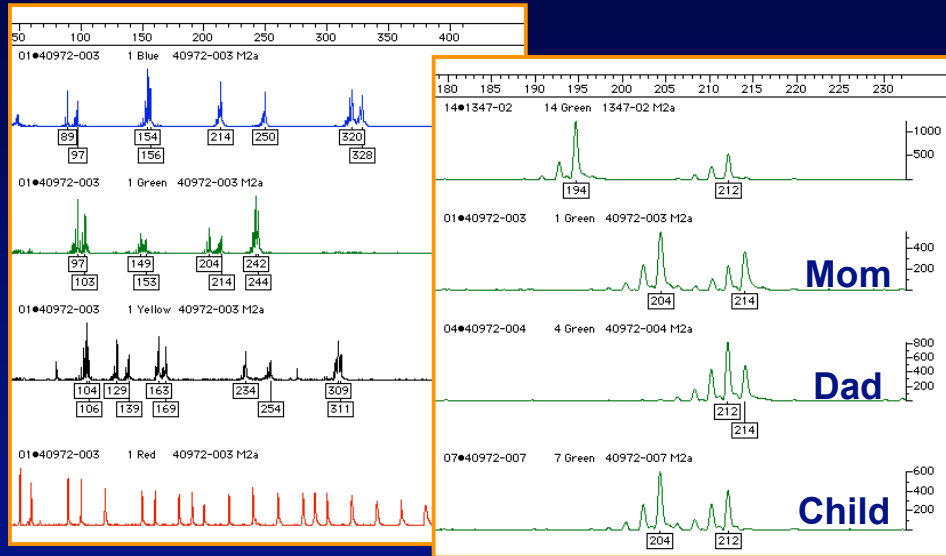
Human Genetic Variation

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

Scoring Microsatellites



Scoring Microsatellites



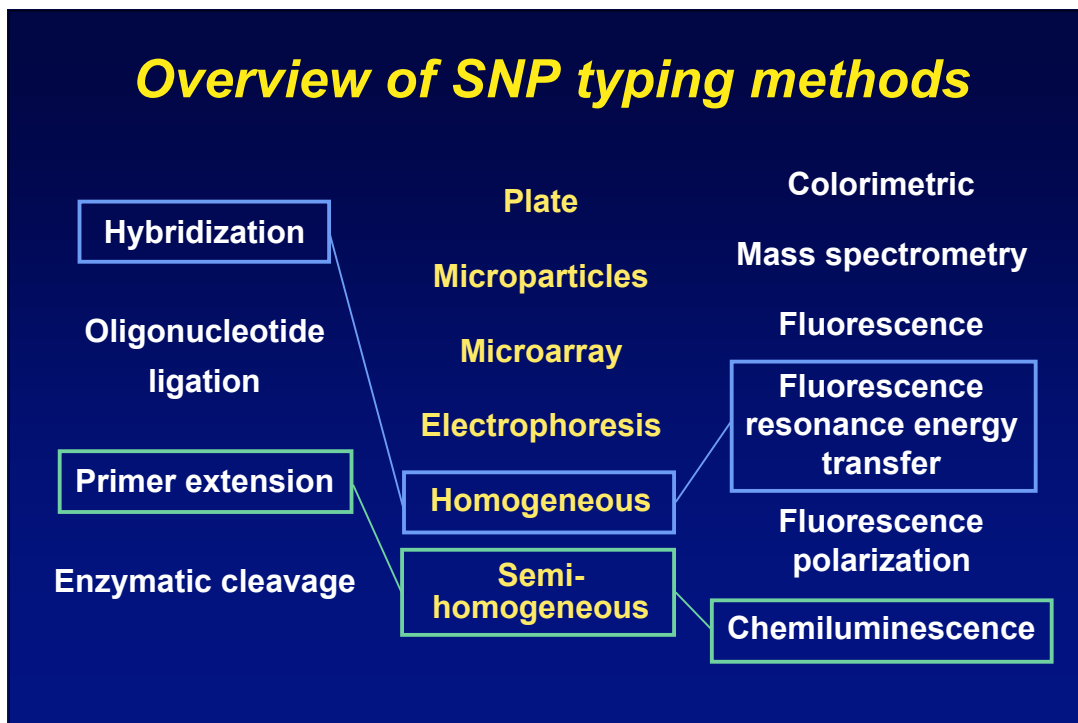
Scoring SNPs

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

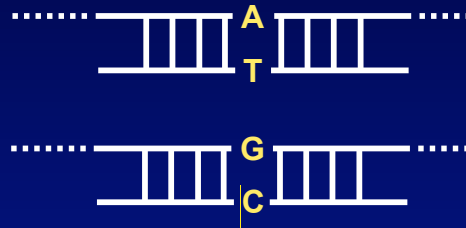
Scoring SNPs (2)

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

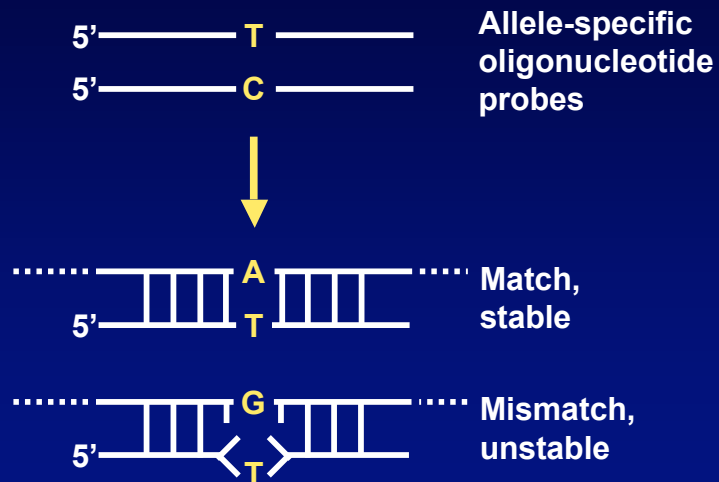
Overview of SNP typing methods



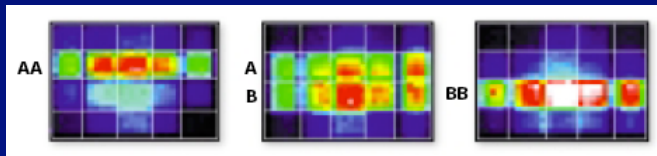
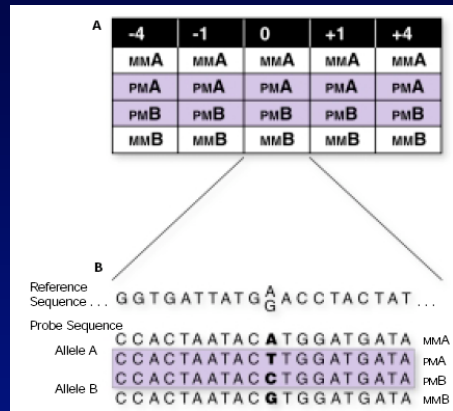
Example SNP



Hybridization



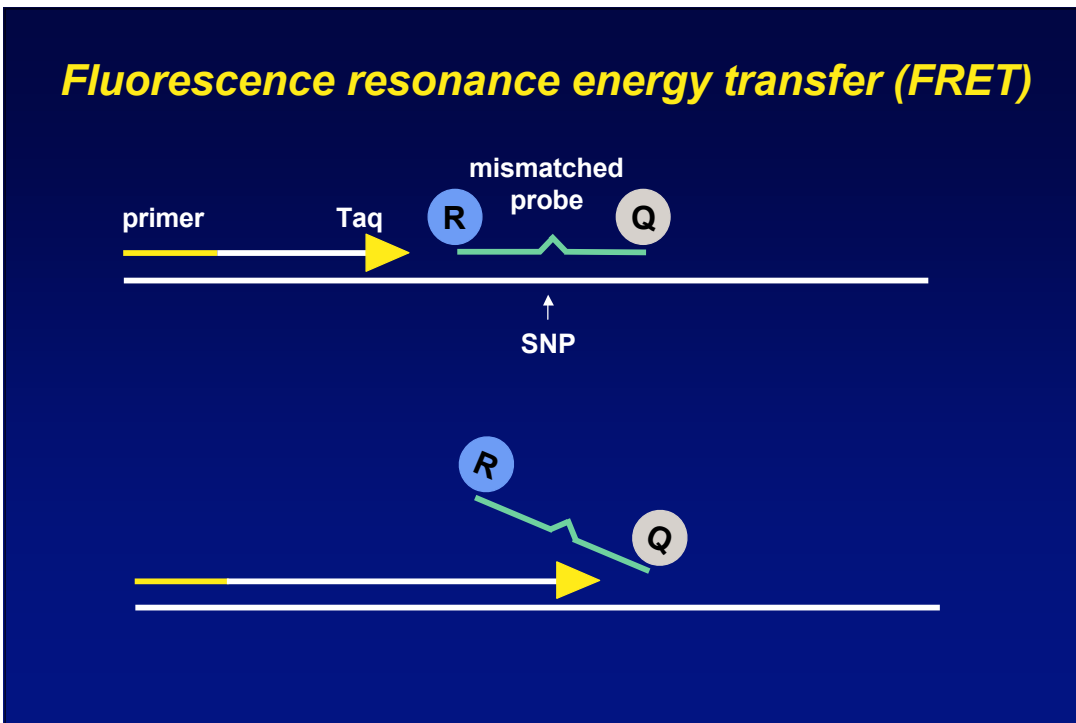
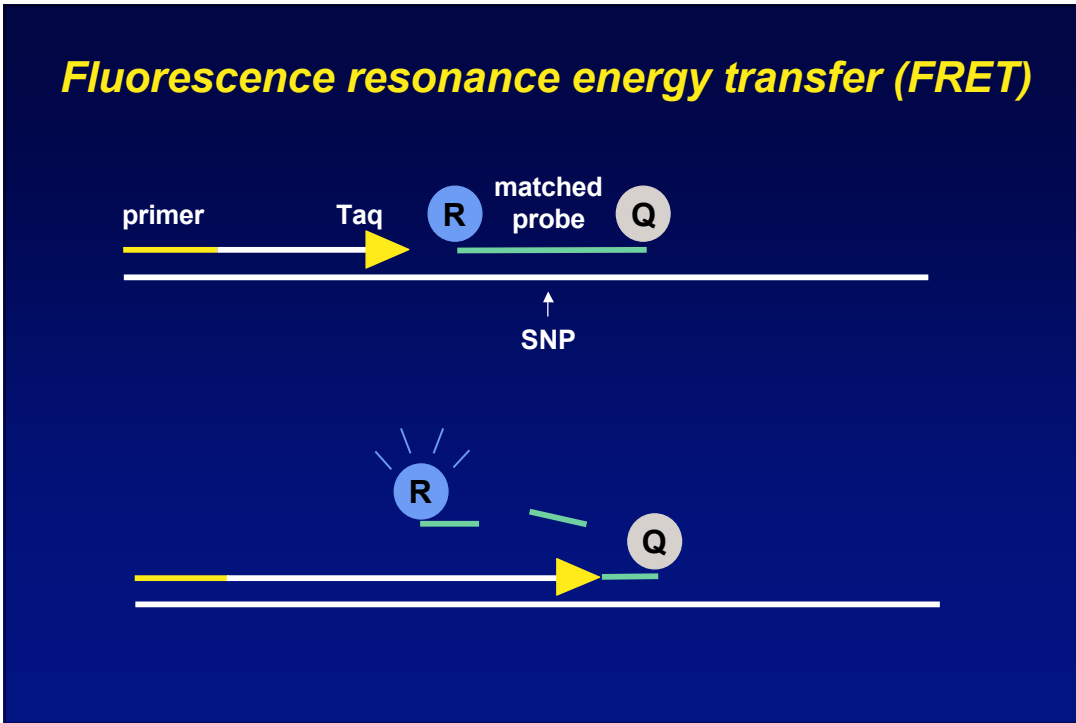
Affymetrix HuSNP Mapping Assay



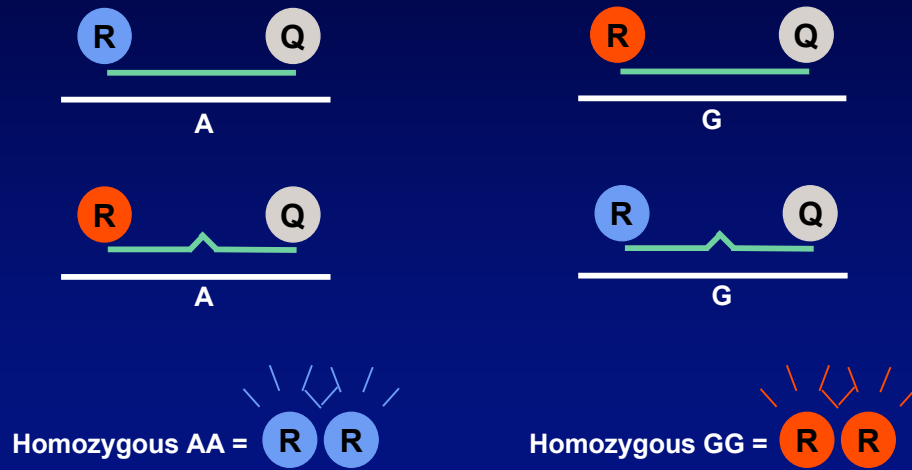
Hybridization to Oligonucleotide Arrays

- **Advantages:**
 - Simple to perform
 - Highly multiplexed
 - Automated analysis

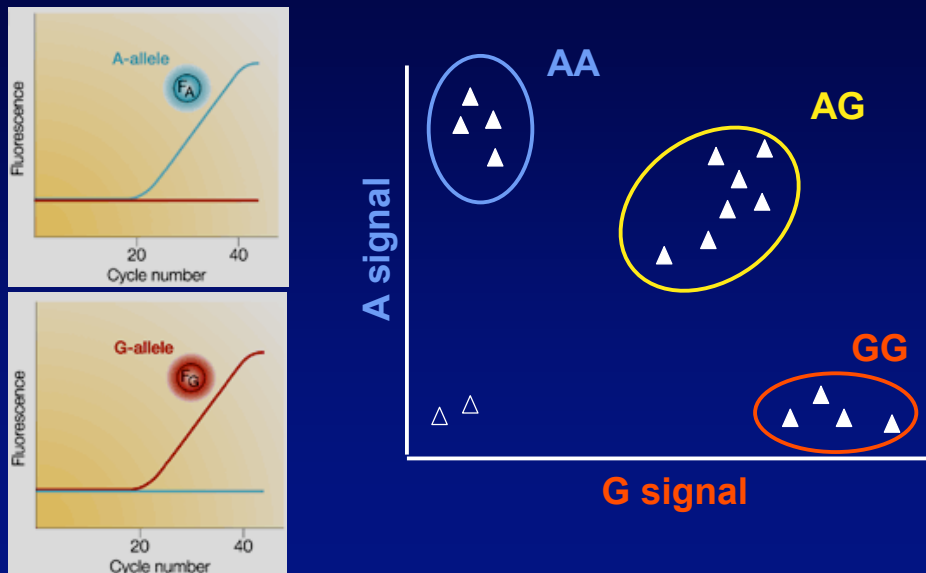
- **Disadvantages**
 - Expensive to design/create chip
 - Local sequence affects success



TaqMan competing probes



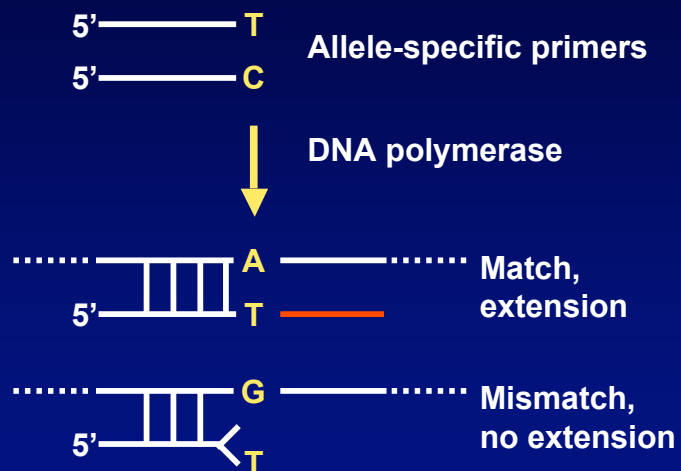
TaqMan genotype scoring



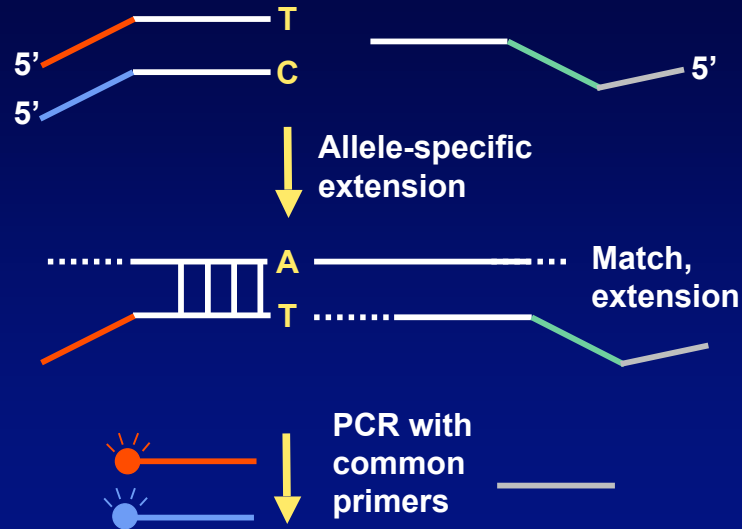
TaqMan

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

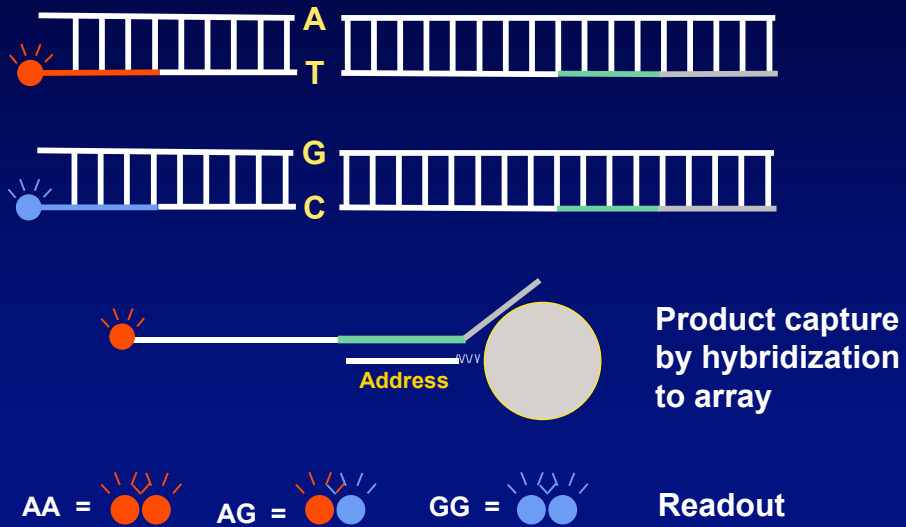
Allele-specific PCR



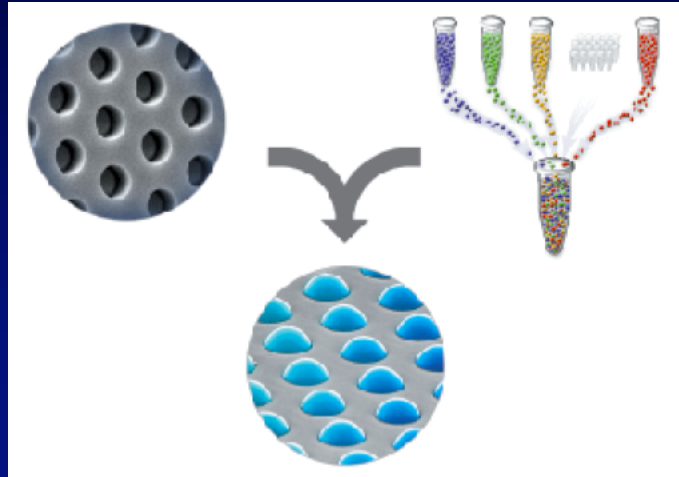
ILLUMINA: Allele-specific extension



ILLUMINA: Allele-specific extension



Illumina genotyping technology



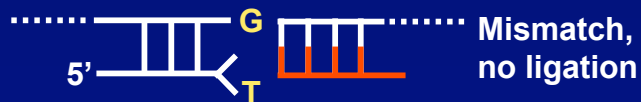
Illumina

- **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)



Ligase



Primer extension = Minisequencing



DNA polymerase

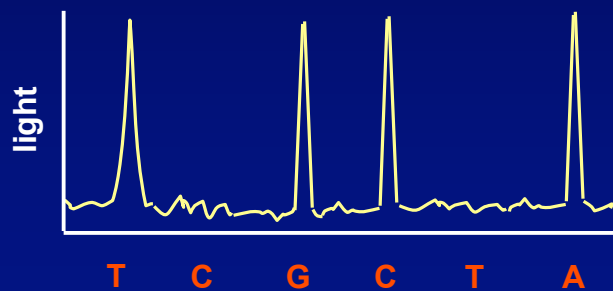


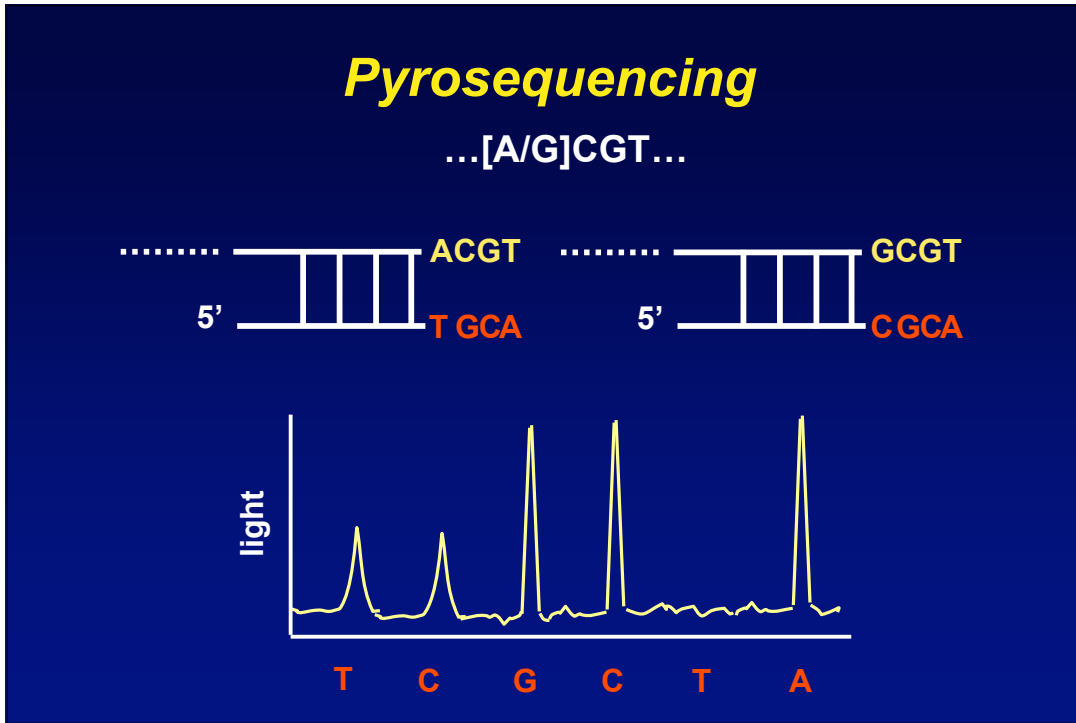
Pyrosequencing

- 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...





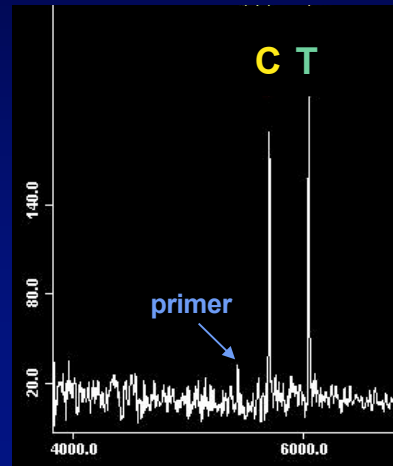
- ### *Pyrosequencing*
- **Advantages:**
 - Accurate
 - Accurate allele frequency estimation
 - Robust for closely spaced SNPs

 - **Disadvantages**
 - Expensive reagents
 - Requires post-PCR processing

Primer extension mass spectrometry

Primer extension reactions designed to generate different sized products

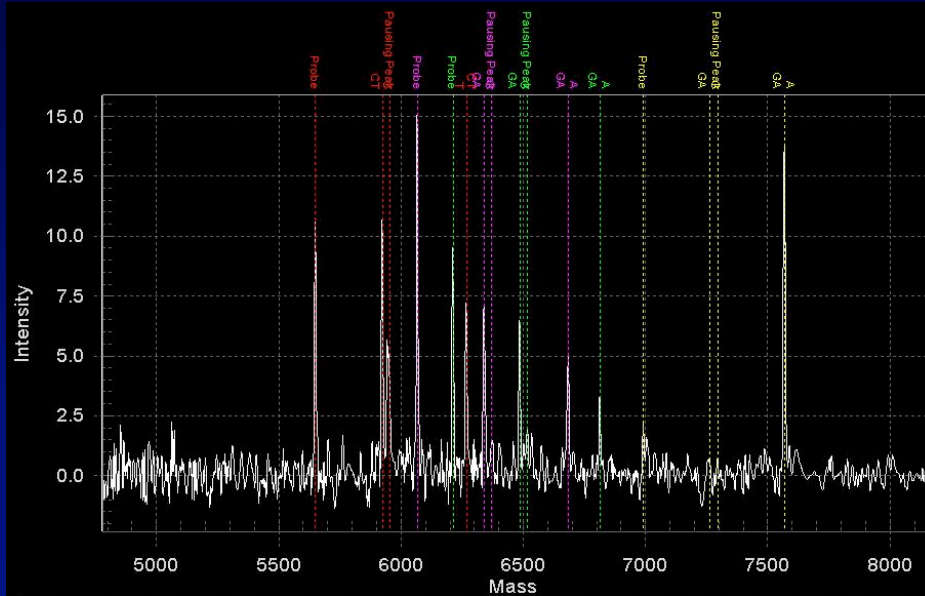
	Mass in Daltons
GGACCTGGAGCCCCCACC	5430.5
GGACCTGGAGCCCCCACC C	5703.7
GGACCTGGAGCCCCCACC TG	6047.9



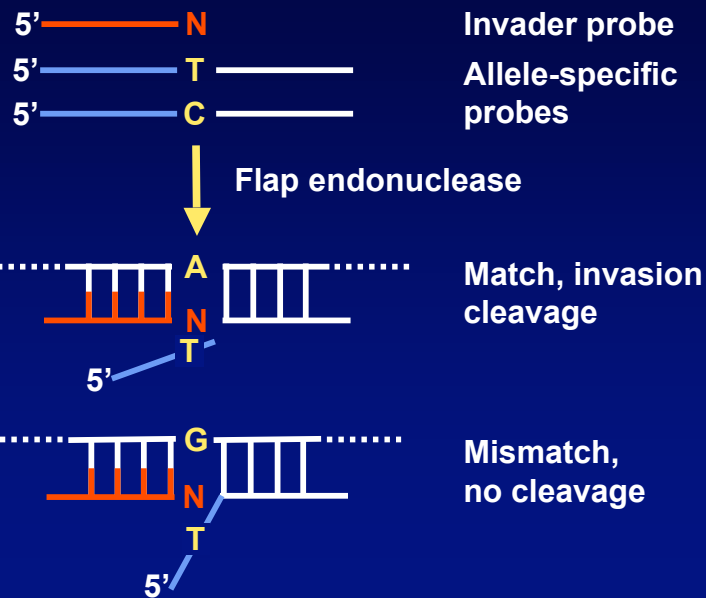
Primer extension mass spectrometry

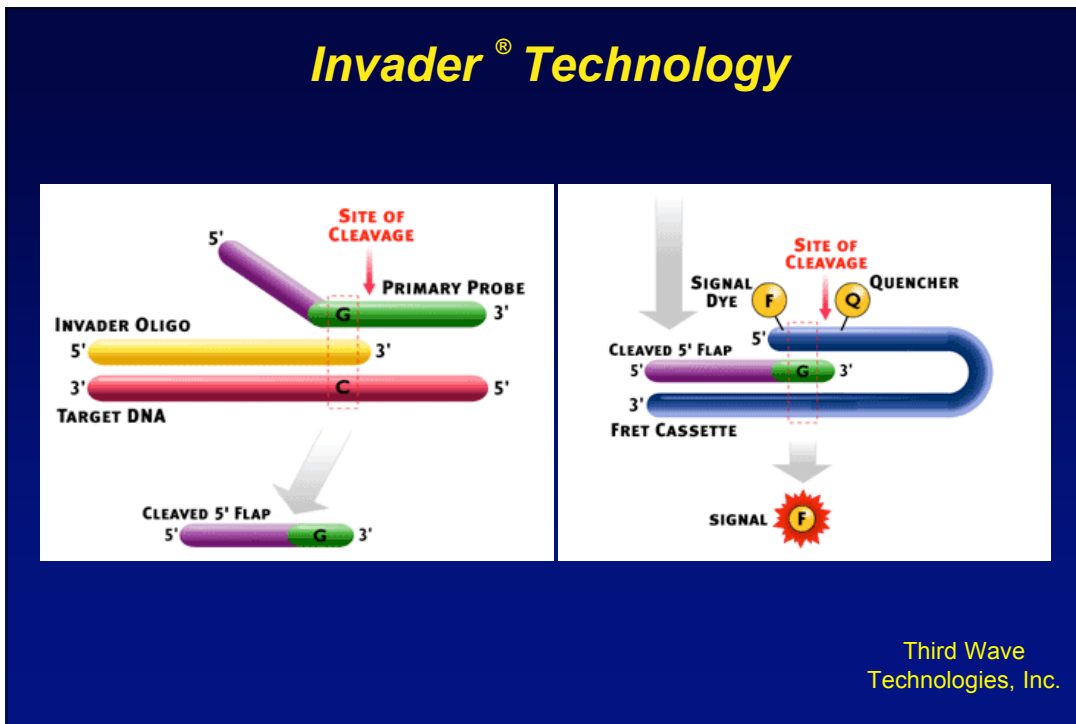
- **Advantages:**
 - Accurate
 - Automated assay design
 - Fast automated data collection
 - Multiplexing capacity
- **Disadvantages**
 - Expensive instruments, consumables
 - Extensive post-PCR processing

Mass spectrometry multiplexing



Invasive cleavage of oligo probes





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

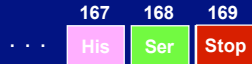
Human Genetic Variation

- 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 ...



Coronary disease:
LDL receptor gene

... TTT TAC GTC ATG ...



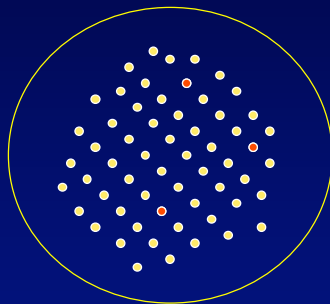
Deep-vein thrombosis:
The Factor V gene



~~APC cleavage~~

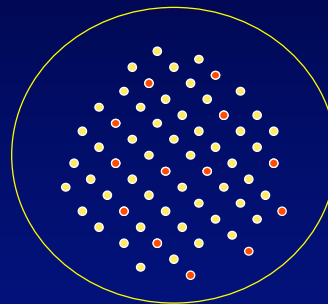
Factor V^{Leiden} association study

301 controls



5% (14) Arg506Gln

301 cases



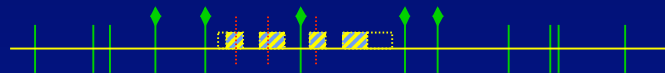
21% (64) Arg506Gln

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

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

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

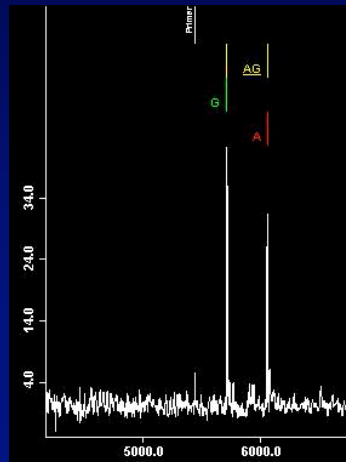
500 cases one pool
500 controls one pool
10,000 SNPs

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

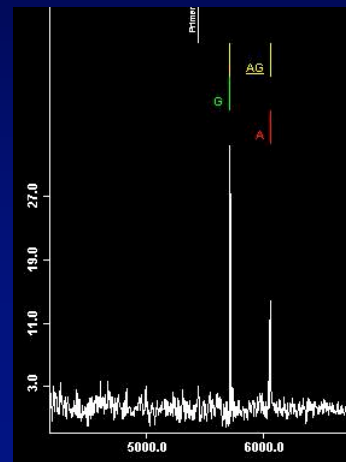
Genotyping of DNA pools

- 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

```
GAAAAAATAAAGTTTTCCCTTCCTCCTATTTTGTCCCTTACTTCAATTTATTTTATTATTAATATATATATTTTTTGAGACGGAGTTTCACTCTGTG  
TGCCAACCTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCACACTCCGCTTTC [C/T] GGTTC AAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGAC  
TACAGTCACACACCACCCGCCCCGCTAATTTTTGTATTTTTAGTAGAGTTGGGGTTTCCACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCC  
GCCAGCCTCTGCCCTCCCAAAGAGCTGGGATACAGGGCGTGAGCCACCGCGCTCGGCCCTTGCATCAATTTCTACAGCTTGTTCCTTGCCTGGACTTTACA  
AGTCTTACCTTGTCTGCCCTTGCATATTTGTGTGGTCTCATTCTGGTGTGCCAGTAGCTAAAAATCCATGATTTGCTCTCATCCCACTCCTGTGTTCATCT  
CCTCTTATCTGGGGTCAC [A/C] TATCTCTTCGTTGATTGCATTCRGATCCCACTACTTAGCATGTGGTAAACAACCTGCCTCTGCTTCCGAGGCTGTTGA  
TGGGGTGTCTTTCATGCCTCAGAAAAATGCATTGTAAGTTAAATTTAAAGATTTTAAATATAGGAAAAAGTAAGCAACATAAGGAACAAAAAGGAAAGA  
ACATGTATCTAATCCATTATTTATATACAATTAAGAAATTTGAAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAGTCTTTTACTCTTTAC  
AAAAACATGTGTAGCAATTTGGGAAGAATAGTAACTCACCAGAACAGTGAATGTGAATATGTCACTTACTAGAGGAAAGAACGCACTTGAAAAACATCT  
CTAAACCGTATAAAAACAATTACATCAATATGATGAAACCCAGGAATTTTTTAGAAAAATACAGGGCTAATAACAAGTAGAGCCACATGTCAATTA  
TCTCCCTTGTGCTGTGTGAGAAATCTAGAGTTATATTTGTACATAGCATGGAAAAATGAGAGGCTAGTTTATCAACTAGTTCATTTTTAAAGTCTAACA  
CATCCTAGGTATAGGTGAACCTGCTCCTGCCAATGTATTGCACATTTGTGCCAGATCCAGCATAGGGTATGTTTGCCATTTACAAAAGTCTTATGCTTAAAG  
AGAGGAAATATGAAGAGCAAAACAGTGCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATGGAAAAATGAGAACTACTCATTTCTA  
AATTACTCATGTATTTTCCAGAAATTAAGTCTTTTAAATTTTGTATAAAATCCCAATGTGAGACAAGATAAGTATTAGTATGGTATGAGTAATTAATCTGT  
TATAAATATTCATTTTCTAGTGAAGAAATAAATAAAGTTGTGATGATTTGTTGATTTTTTTCTAGAGGGGTGTGAGGAAAGAAATGCTTTTTTT  
CATCTCTCTTTCCACTAAGAAAGTTCARCTATTAATTTAGGCACATACAAATAATTTCTCCATTTCAAATGCCCCAAAAGGTAATTTAAGAGACTTAAACTG  
AAAAGTTTAAAGATAGTCACACTGAATATATTAATAAATCCACAGGGTGTGTTGAACTAGGCCCTTATATTAAGAGGCTAAAAATTAATAAGACCCACAGGC  
TTTAAATATGGCTTTAACTGTGAAAGGTGAACTAGAAATGAATAAATCCATAAATTTAAATCAAAGAAAGAAACAACT [A/G] AAATTAAGTTATTA  
TACAAGAAATATGGTGGCCGGATCTAGTGAACATATAGTAAAGATAAAACAGAAATTTCTGAAAAATCCTGAAAAATCTTTGGGCTAACCTGAAAAACGTA  
TATTTGAACTATTTTTAAATGCAGTGATCTAGAAATATTTAGAAATCATATGTA
```

[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://espressoftware.com/pages/sputnik.jsp
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
TSC	http://snp.cshl.org/
HGVbase	http://hgibase.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