

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 2.5×10^{-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...

```
GGCATCTTTGTGTTACTCTGCTCAACATTCAAAGTCCCAGGGGAGAATATTTAGTTGGGCTTAGGTCACATGCCACATGGCTGTAAGGGATGAGA
GAGAAGGAATCCGATGAAAGGAGCCACAGTAACCTTCTGCTTCTGTTATTTGGGGCAAGACACACCAATCTGTCTACACACAGCTGAAAAAATG
GGGGAGAGGATTTCTAAAAGGAACTAGGATGTTATTTACTTATTTTTATTTTTATTTTTTTGAGATGGAGTCTTGCTCTGTCGCCAGGCTGGAGTG
CAGTGGTGAATTTCACTCACTGCAACCTCTGCCTCCCAGGTTCAAGTGATTCTCTGCTCAGCCTCCCCCATAGCTGGAATACAGGCATGTGCC
ACCATGCCAGCTAATTTTTTTGATTTTTTAGTAGAGATGGGGTTTCACCATGTTGGCCAGGCTGGTCTCGAACTCTGACCTCAGGTGATCCGCCCA
CTCGGCCCTCCAGAGTGTGGGATTACAGTGTGTAGCCACCATGTCCGGCCCTAGGATATTTCAATTAAGAAAAGAAATGCTGGATAGCCAAAGTGAA
AATACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
AACATCAGAATCTTTCATCTTTGAAGGCACAAAAGTGTAGTATTACACAGAGGATAGCTATCTTATCTCTCCCTCTCGGAGGTTTCAGAAAATGTTTGAT
ATCATCTGGGGAAGCCAGATGATAACGTTCAATGGAGCAAAAGAAAAGGTGCACACAAATGAGGTGCTTACAAAAAATGGAAGTTTCATATCTCT
GCTACAAAGGGCCAGAGGAATATTTCCCAATAAAGCATTGTTGCCAGGGATGAATGAGATAAGGATGAGACCTCTGATGATAAAATGGTTAGTTCT
TCCTATTAGTTGTTGTTCTGATGTAGAAACAGCCTCTTCTCCCTATATCTGCTTAAATCCAACCTGATAGGAGACGTTTTCCGTTGGGATTAGG
AAAGATACAACAGTTCTGGGGTTGAGTTCAGGGCTAATTTTTCTGAAGGATAAGAGAGCAAGCCCAAGCAAGAGCCAAAGAAAGCAATGATGAGGAA
GCGGCGTAGCAGCCATTTAGACTGGTTGCTTTGTGGGACTCCCTCTTATTTGTACATTATTAGGCTTTCCAACAGGGGACAAATAACAGTATGAATC
CAGACAGGATGAGGGTGGGTTGCACAAGCAGCTGGGCCACTGAACCTAGAGCCTGACTCAAAAAGGAAGGAGGCTGGGCGCAGTGGCTCACACCTGTA
ATCCAGCACTTTGGGAGGCCAGGCGGGTGGATCAGAGGCTGGAGTTTCGAGACAAGCCCTGGCCAAATGTTGTAACCCCATAGCTACTAAAAATAC
AAAAATTAGCCAGGCATGGTGGCAGGCACCTGTAGTCCAGCTACTCGGGAGGCTGAGGCAGAAGAATCACTTGAACCTGGGAGGTGGAGTTCCAGTG
AGCTGAGATTGTCCACTGCATCCAGCCTGGTACAGAGCAAGCTCCATCTCAAAAAAAAAAAAAAAAAAAGGAAGATCTGCCATGGTGTAGGA
CCACCATCCGTTCTCTGGTCGAGTCAGGCTGTGCCCATTTGACTGGGGCATGATGCACTTCTTGTGATCCGATGACATGTTCCAGGCCCCAGGG
AGTGTCCAGGCAGTGCATCAGATTATCAGGCATTGACAGAGATACCTATAAGCTGAGAGCTACAGCCATTTTGGCAAGCTCTGAAAACCCAGAGTTGG
CGCTGTTTCATGGGGGAGGATCTGCATGGTACTCGCTGAGCCATGTTTTTTGTGTTCTGTTTGGAAAGCCTACACATATGTGTTTAAACCATCCCTA
TGCATCATTAGCCTGCT
```

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

Microsatellite

```
GGCATCTTTGTGTTACTCTGCTCAACATTCAAAGTCCCAGGGGAGAATATTTAGTTGGGCTTAGGTCACATGCCACATGGCTGTAAGGGATGAGA
GAGAAGGAATCCGATGAAAGGAGCCACAGTAACCTTCTGCTTCTGTTATTTGGGGCAAGACACCAATCTGTCTACACACAGCTGAAAAAATG
GGGGAGAGGATTTCTAAAAGGAACTAGGATGTTATTTACTTATTTTTATTTTTATTTTTTTGAGATGGAGTCTTGCTCTGTCGCCAGGCTGGAGTG
CAGTGGTGAATTTCACTCACTGCAACCTCTGCCTCCCAGGTTCAAGTGATTCTCTGCTCAGCCTCCCCCATAGCTGGAATACAGGCATGTGCC
ACCATGCCAGCTAATTTTTTTGATTTTTTAGTAGAGATGGGGTTTCACCATGTTGGCCAGGCTGGTCTCGAACTCTGACCTCAGGTGATCCGCCCA
CTCGGCCCTCCAGAGTGTGGGATTACAGTGTGTAGCCACCTGCTCGGGCCCTAGGATATTTCAATTAAGAAAAGAAATGCTGGATAGCCAAAGTGAA
AATACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC
AAATCAGAACTTTCATCTTTGAAGGCACAAAAGTGTAGTATTACACAGAGGATAGCTATCTTATCTCTCTCGGAGGTTTCAGAAAATGTTTGAT
CTCATCTGGGGAAGCCAGATGATAACGTTCAATGGAGCAAAAGGAGTGCACACAAATGAGGTGCTTACAAAAAATGGAAGTTTCATATCTCT
GTTACAAAGGGCCAGAGGAATATTTCCCAATAAAGCATTGTTGGCCAGGCTGAAATGAGATAAGGATGATAGCCTCTGATGATAAAATGTTGCT
TCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCTCTGCT
ATCCAGCACTTTGGGAGCCGAGGCGGGTGGATCAGAGGCTGGAGTTCGAGACAAGCCTGGCCAAATGTTGAAACCCCATAGCTACTAAAAATAC
AAAATTAGCCAGGCATGGTGGCAGGCACCTGTAGTCCAGCTACTCGGGAGGCTGAGGCAGAAGAATCACTTGAACCTGGGAGGTGGAGTTGCAGTG
AGCTGAGATTGTGCCACTGCATCCAGCCTGGTGCAGAGCAAGACTCCATCTCAAAAAAAAAAAAAAAAAAAGGAAGATCTGCCATGGTGTAGGA
CCACCATCCGTTCTCTGGTCGAGTCAGGCTGTGCCCATTTGACTGGGGCATGATGCACTTCTTGTGATCCGATGACATGTTCCAGGCCCCAGGG
AGTGTCCAGGCAGTGCATCAGATTATCAGGCATTGACAGAGATACCTATAAGCTGAGAGCTACAGCCATTTTGGCAAGCTCTGAAAACCCAGAGTTGG
CGCTGTTTCATGGGGGAGGATCTGCATGGTACTCGCTGAGCCATGTTTTTTGTGTTCTGTTTGGAAAGCCTACACATATGTGTTTAAACCATCCCTA
TGCATCATTAGCCTGCT
```

CA

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 ...

```
GAAATAATTAAGTTTTCTTCTCTCTCTATTTTGTCTCTTACTTCAATTTATTTATTTATTAATAATTATTTTGTGAGCGGAGTTTCACTCTTGT
TGCCAACTGGAGTGCAGTGGCGTGATCTCAGCTCACTGCACACTCCGCTTCTGGTTTCAAGCGATTCTCTGCCCTCAGCCTCCTGAGTAGTGGGACTACA
GTCACACACCACCACGCCCGGCTAATTTTTGTATTTTTAGTAGAGTTGGGGTTTACCATGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCA
GCCTCTGCCTCCCAAAGAGCTGGGATTACAGCGGTGAGCCACCCGCTCGGCCCTTGCATCAATTTCTACAGCTTGTTTTCTTGGCTGGACTTTACAAGTC
TTACCTTGTCTGCCTCAGATATTTGTGTGTCTCATTTCTGGTGTGCCAGTAGCTAAAAATCCATGATTTGCTCTCATCCCACTCCTGTTGTTCARCTCCTC
TTACTGGGGTCACATATCTTTCGTGATTGCATTCGATCCCCAGTACTTAGCATGTGCGTAAACAACCTGCTCTGCTTTCCCAGGCTGTGATGGGGTGC
TGTTCACTCCCTCAGAAAAATGCATTGTAAGTTAAATTTAAAAGATTTTAAATATAGGAAAAAAGTAAGCAAAACATAAGGAACAAAAAGGAAAGAACATGTAT
TCTAATCCATTTATTTATTAACAATTAAGAAATTTGGAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAAGTCTTTTACTCTTTACAAAATACA
TGTGTAGCAATTTTGGGAAGAATAGTAACCTACCCGAAACAGTGAATGTGAATATGTCACTTACTAGAGGAAAGAAGGCCTTGAAAAACATCTTAAACCG
TATAAAAACAAATTCATCATATAATGATGAAACCCAGGAATTTTTTAGAAAAACATTACCAGGGCTAATAACAAGTAGAGCCACATGTCTATTTCTTCCCT
TGTGTCTGTGAGAAATCTAGAGTTATTTTACATAGCATGGA AAAAATGAGAGGCTAGTTTATCAACTAGTTCATTTTAAAAGTCTAACACATCCTAG
GTATAGTGAAGTCTCTCCTGCCAATGTATTGCACATTTGTGCCAGATCCAGCATAGGGTATGTTTGCCATTACAAAAGTTTATGTCTTAAGAGAGGAAA
TATGAGAGCAAAACAGTGCATGCTGGAGAGAGAAGCTGATACAAATATAAATGAAACAATAATGGAAAAATGAGAACTACTCATTTCATAATTAATCTC
ATGTATTTTCTAGAAATTAAGTCTTTAATTTTTGATAAAATCCCAATGTGAGACAAGATAAGTATTAGTGTGGTATGAGTAATTAATATCTGTATATAAT
ATTCAATTTCTAGTGGAGAAATAAAATAAAGTTGTGATGATTTGTTGATTTATTTTTCTAGAGGGGTTGTGAGGGAAGAAATGTCTTTTTTCTCATCTCT
CTTCCACTAAGAAAGTCACTATTAATTTAGGCACATACAAATTAATCTCCATTTCTAAAATGCCAAAAGGTAATTTAAGAGACTTAAAACGAAAAGTTT
AAGATAGTCACTGAACTATATAAAAAATCCACAGGGTGGTTGGAAC TAGGCCTTATATTAAGAGGCTAAAAATGCAATAAGCCACAGGCTTTAATA
TGGCTTTAACTGTGAAAGGTGAACTAGAAATGAATAAAATCCTATAAAATTTAAATCAAAGAAAGAAACAACTGAAATTAAGTTATTATACAAGAAATAG
GTGGCTGGACTAGTGAACATATAGTAAAGATAAAAACAGAAATTTCTGAAAAATCCTGAAAAATCTTTGGGCTAACCTGAAAAACAGTATATTTGAAACTA
TTTTTAAAATGCACTGATCTAGAAATTTTTAGAATCATATGTA
```

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

Single nucleotide polymorphisms (SNPs)

```

GAAATAATTAATGTTTTCTTCCTTCCTCAATTTTGGTCTTTACTTCAATTTTATTTTATTATTAATATTATTATTTTTGAGACGGGAGTTTCACTCTTGT
TGCCAACCTGGAGTGCAGTGGCGTGAATCTCAGCTCACTGCACACTCCGCTTTCGGTTCAAGCGATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGACTACA
GTCACACACCACCACGCCCCGCTAATTTTTGTATTTTTAGTAGAGTTGGGGTTTACCATTGTTGGCCAGACTGGTCTCGAACTCCTGACCTTGTGATCCGCCA
GCCTCTGCCTCCCAAAGAGTGGGATTACAGCGTGAAGCCACCCGCTCGGCCCTTTGCATCAATTTCTACAGCTTGTTCCTTTGCTGGACTTTTACAAGTC
TTACCTTGTCTCGCCAGATATTTGTGTGGTCTCATTCTGGTGTGCCAGTAGCTAAAAATCCATGATTTGCTCTCATCCCACTCCTGTTGTCATCTCCTC
TTATCTGGGGTCACTATCTCTTCGTGATTGCATTCGTATCCCCAGTACTTAGCATGTGCGTAACTCAACTCTGCCTCTGCTTTCCAGGCTGTTGATGGGGTGC
GTCATCGCCTCAGAAAAATGCATCTAAGTTAAATTTAAGATTTTAAATATAGGAAAAAAGTAAACAAATAGGAACAAAAAGGAAGAACATGTAT
TCTAATCCATTATTTATTATACAATTAAGAAATTTGAAACTTTAGATTACACTGCTTTTAGAGATGGAGATGTAGTAACTTTTACTCTTTTACAAAAATACA
TGTGTTAGCAATTTGGGAAGAATAGTAACTCACCAGAACAGTGAATGTGAATATGTCACCTACTAGAGGAAGAAGGCCTTGA AAAACATCTTAAACCG
TATAAAAACAATTACATCAATGATGAAACCCAGGAATTTTTTAGAAAAATTACCAGGGCTAATAACAAGTAGAGCCACATGTCATTTATCTCCCT
TTGTGCTGTGTGAGAATTCTAGAGTTATTTGTACATAGCATGGA AAAATGAGAGGCTAGTTTATCRACTAGTTCAATTTTAAAAGTCTAACACATCCTAG
GTATAGTGAACCTCTCCTGCCAATGTATTGCACATTTGTGCCAGATCCAGCATAGGGTATGTTTGCATTTTCAAACTTTATGCTTTAAGAGAGGAAA
TATGAGAGCAAAACAGTGCATGCTGGAGAGAGAAAGCTGATACAAATATAAATGAAACAATAATTTGAAAAATTGAGAACTACTCATTTC TAAATTACTC
ATGATTTTCTAGAAATTAAGTCTTTTAAATTTTGATAAATCCCAATGTGAGACAAGATAAGTATTAGTGATGGTATGAGTAAATTAATCTGTTATATAAT
ATTCAATTTCTAGTGGAGAAATAAAATAAAGGTGTGATGATTTGTTGATTTATTTTTCTAGAGGGTTGTCAGGGAAAGAAATGCTTTTTTTCATCTCT
CTTCCACTAAGAAAGTCAACTATTAATTTAGGCACATACAATAATTACTCCATTAATTTAAGACTTAAAACGAAAGTTT
AAGTAGTACACACTGAACATATTA AAAATCCAGGGTGGTGGAACTAGGCCCTAATTTAAGACTTAAAACGAAAGTTT
TGGCTTTAAACTGTGAAAGGTGAACTAGAATGAATAAAATCCTATAAATTTAAATCTAATTAAGACTTAAAACGAAAGTTT
GTGGCTGGATCTAGTGAACATATAGTAAAGATAAAACAGAATATTTCTGAAAAATGCTAATTAAGACTTAAAACGAAAGTTT
TTTTAAATGCACTAGATACTAGAAATTTTTAGAAATCATATGTA
    
```

[G/A]

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

CAGACTCAATAAGCATGTTTTTTTTTTTTTTTTTTTTTTTTTTGAGACG
GAGTCTCGCTCTGTCGCCAGGCTGGAGTGCAAGTGGCGCGA
TCTCGGCTCACTGCAAGCTCCGCCTCCGGGTTACGCCATT
CTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCTCCC
GCCACCACGCCCGGCTAATTTTTGTATTTTAGTAGAGACGG
GGTTAGCATGTTTT

CAGACTCAATA[LARGEINSERTION/-]AGCATGTTTT

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

chr	marker	primer	phys.loc	gen.loc	gen.female	gen.male
1	D1S468	AFM280we5	3766906	4.00	4.46	3.54
1	D1S2845	AFM344we9	4604712	6.43	6.47	6.39
1	D1S2893	AFM123xc3	4737957	6.43	6.47	6.39
1	D1S2660	AFMa203yc1	4950934	7.09	7.78	6.40
1	D1S2132	GATA68D01	NA	8.06	8.66	7.46
1	D1S2633	AFMa131ya5	6063900	10.11	9.72	10.49
1	D1S2870	AFMa052wg1	6276124	11.37	11.41	11.32
1	D1S2731	AFMb039zg9	6687488	12.00	12.35	11.64
1	D1S2642	AFMa152xg5	6876371	12.00	12.35	11.64
1	D1S1646	GATA23G09	7505668	12.14	12.36	11.92
1	D1S2663	AFMa210xg9	7620833	12.84	13.75	11.92
1	D1S2694	AFMa295yh5	7705014	12.90	13.89	11.92
1	D1S548	GATA4H04	7806005	13.08	13.89	12.27

Marker retrieval: genome browser

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

UCSC Genome Bioinformatics

Genomes - Gene Sorter - Blat - PCR - Tables - Proteome - FAQ - Help

Human Genome Browser Gateway

The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#).
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade genome assembly position image width

Vertebrate Human May 2004 D951838 600 Submit

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

[Add Your Own Custom Tracks](#) [Configure Tracks and Display](#) [Clear Position](#)


Marker retrieval: genome browser

Home Genomes Blat PCR DNA Tables Gene Sorter Convert Ensembl NCBI PDF/PS Help

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position chr9:137,812,536-138,012,815 jump clear size 200,280 bp. configure



Base Position	137850000	137900000	137950000	138000000
Chromosome Band		Chromosome Bands Localized by FISH Mapping Clones 9q34.3		
		STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps		
RFMB303209				
SHGC-30997				
SHGC-82935				
SHGC-111534				
SHGC-149365				
RH48863				
RH11945				
		Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq		
	EHMT1	EHMT1	EHMT1	RP451334
		AK097611		
		AK124119		
		AK125987		
		RefSeq Genes		
	EHMT1			

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: GCTTCTGCACCTTGTAGAACCAAT
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

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 extra bases upstream (5') and extra downstream (3')

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

Sequence Formatting Options:

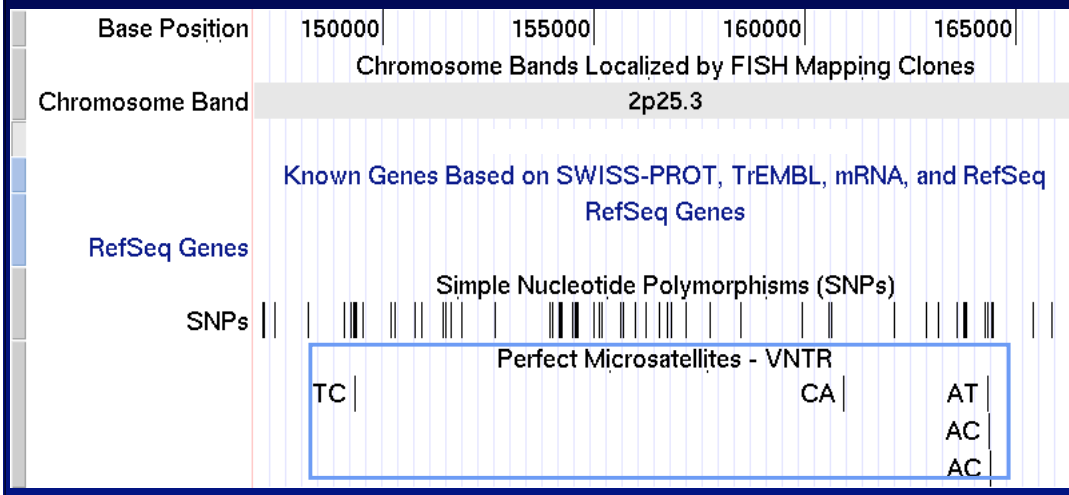
All upper case.
 All lower case.
 Mask repeats: to lower case to N
 Reverse complement (get '-' strand sequence)

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'pad=0
CCAAAGTGTCTGCATGTTGGCTGTGTGCTCCGAGCCTGACCCCCATGAACA
TACTGCAGACGCCTGGTGTGATCGTTTCCAGCGTCCGTGGTCCCAGGCA
CCTCCTTACTCCAGAGCGGATTGCCAGGCCCGCGGCGTCTGTGGGTGG
TGCTGTCAAAGGACCTACCCGCTTTGGATGGTTCTCAGCTGTTACGTTCC
CTCCAAGTGTGCTTCTGCACCTTGTAGAACCAATGTGTGTGTGTGTGTGT
GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT
AAGTAAAGGGGGTCTCACTG
TGTTTCCAGGCTGGGCTTGAACCTCCTGGAATGAAGCAAACCTCCCTCC
CACTCAAGCCTCCTCAGTAGCTGGGTCTTCCAGGTGTGAGCTGCTGTGCC
AGCTTAAACAGAGTGGATTTCCCATCCCTTTAGGAGAGTTTCTTTTA
TGTTAAAGCAGTGGCTTTAGATCTGTTTCTTTAAATCCTGGAACCTTA
AAAAAAGTCATGGAGTCTGATTATATAAAACAGTCGAACCTAGAAGTGC
TTTGTTCAGAATGGGTTGGGAGCCCCAGGCCCTTCTCGCTGCTGCT
CTTGTTTGGAACCACTGTCTCCAGAGCCCCAGACATTTGCTTCTCCCTC
TCAGCCTCTTATCTTTTA
```

Microsatellite identification from sequence

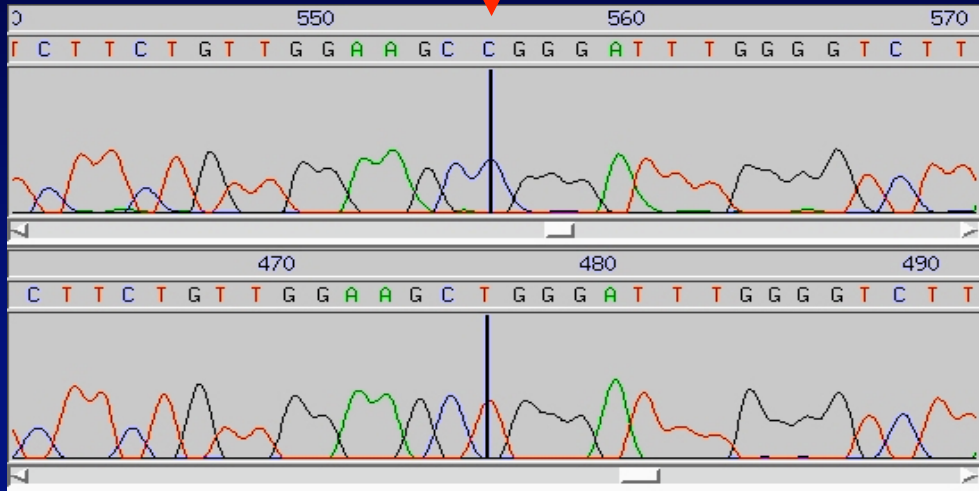
Chr1:146,000 - 166,000



SNP identification

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

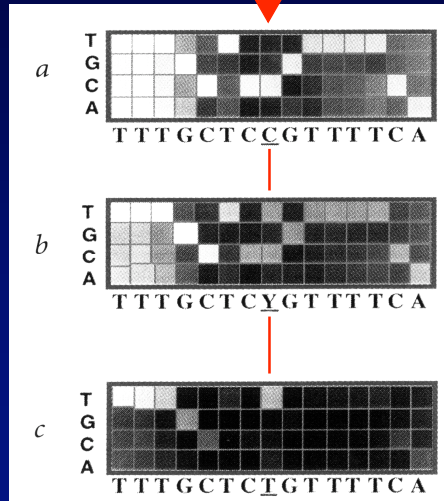
SNP identification: sequencing



SNP identification: sequencing chips



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

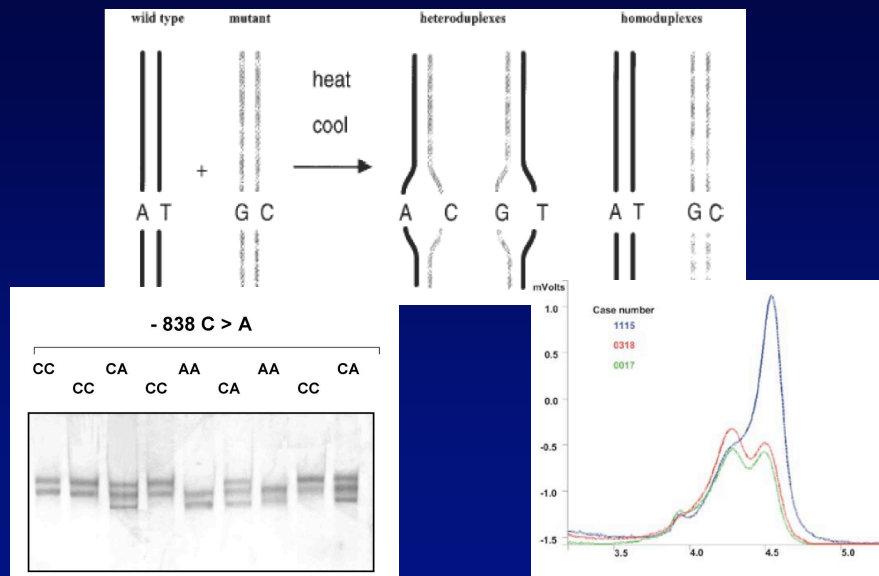


The Sanger Institute

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

NCBI Single Nucleotide Polymorphism

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for

[Limits](#) [Preview/Index](#) [History](#) [Clipboard](#) [Details](#)

dbSNP BUILD 124

GENERAL

- Contact Us
- dbSNP Homepage
- SNP Science Primer
- Announcements
- dbSNP Summary
- FTP Download Server
- Getting Started
- Build History
- Handle Request

DOCUMENTATION

- FAQ
- dbSNP Handbook Overview
- How to Submit

dbSNP Search Options

Entrez SNP	ID Numbers	Submission Info	Batch	Locus Info	Free Form	Easy Form	Between Markers
------------	------------	-----------------	-------	------------	-----------	-----------	-----------------

ANNOUNCEMENT

- **NEW!** [Search SNP in Mouse](#).
- **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.

Search by IDs

Note: [rs#](#) and [ss#](#) must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

Reference cluster ID(rs#)

SNP retrieval: Entrez SNP

ENTREZ **SNP**
Single Nucleotide Polymorphism

All Databases PubMed Nucleotide Protein Genome Structure Popset Taxonomy SNP

Search SNP for SLC2A10 [Go] [Clear]

Limits Preview/Index History Clipboard Details

Click on the image below to view the connections between Entrez SNP and other databases.

SNP

dbSNP is now incorporated into NCBI's Entrez system and can be queried using the same approach as the other Entrez databases such as PubMed and GenBank. The original database with additional information and search options are available [here](#).

- Enter one or more search terms.
- Available search fields are listed below
- Use [Limits](#) to restrict your search by search field, chromosome, and other criteria.

SNP retrieval: Entrez SNP

ENTREZ **SNP**
Single Nucleotide Polymorphism

All Databases PubMed Nucleotide Protein Genome Structure Popset

Search SNP for SLC2A10 [Go] [Clear] [Save Search]

Limits Preview/Index History Clipboard Details

Display Graphic Summary Show: 20 Sort Send to Text

All: 139 Human: 139 Mouse: 0 NEW: 0 Other Organisms: 0 UPDATE: 0

Items 1 - 20 of 139

1: [rs13043534](#) [Homo sapiens]
ccatatctcaggtagccacaagacc [A/G] tggattcccatacctgacctccAGA
[MapView] [GeneView] [SeqView] [No 3D] [No OMM]

2: [rs13038743](#) [Homo sapiens]
tttttttttttttttttttttttttttttt [G/T] gagatggagtctegctctgttcccc
[MapView] [GeneView] [SeqView] [No 3D] [No OMM]

3: [rs12481018](#) [Homo sapiens]
TTCAAAC TGAGATTG CCAAGGCCG [G/T] GTCCATGCAGCTGTTGGCCACTCA
[MapView] [GeneView] [SeqView] [No 3D] [No OMM]

Entrez SNP: Limits

Entrez SNP: Limits

SNP class:	
<input type="checkbox"/> het	variation has unknown sequence composition, but is observed to be heterozygous
<input type="checkbox"/> in del	insertion deletion polymorphism, deletions represented by '-' in allele string
<input type="checkbox"/> microsat	microsatellite / simple sequence repeat
<input type="checkbox"/> mixed	
<input type="checkbox"/> mnp	multiple nucleotide polymorphism (all alleles same length where length>1)
<input type="checkbox"/> named	allele sequences defined by name tag instead of raw sequence, e.g. (Alu)-
<input type="checkbox"/> no variation	submission reports invariant region in surveyed sequence
<input type="checkbox"/> snp	true single nucleotide polymorphism
Method class:	
<input type="checkbox"/> computed	variation was mined from sequence alignment with software
<input type="checkbox"/> dhplc	Denaturing High Pressure Liquid Chromatography used to detect SNP
<input type="checkbox"/> hybridize	hybridization method (e.g. chip) was used to assay for variation
<input type="checkbox"/> other	other method used to detect variation
<input type="checkbox"/> rflp	variation in enzyme restriction site used to detect variation
<input type="checkbox"/> sequence	samples were sequenced and resulting alignment used to define variation
<input type="checkbox"/> sscp	single stranded conformational polymorphism used to detect variation
<input type="checkbox"/> 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	protein	mrna orientation	transcript	snp count
(contig mRNA transcript):	NT_011362	NM_030777	NP_110404	forward	plus strand	128, all

Contig position	dbSNP rs#	Heterozygosity	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	300
		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	195

Fasta sequence (Legend)

```
>gn|dbSNP|rs5811|allelePos=61|totalLen=121|taxid=9606|snpclass=1|alleles='A/C'|mol=cDNA|build=52
CTCACGTGAC CCAATGACCTG CAGGAGATGA AGCAAGAGAG TCGGCAGATG ATGCGGAGA
#
GAAGGTACCC ATCCTGGAGC TGTTCGCTC CCCCCTAC CGCCAGCCCA TCCTCATCGC
```

SNP retrieval: SNPper

chip **SNPper - Main Menu** Goldenpath: hg17
dbSNP: build 123
Login: iipga **IIPGA**

[Home](#) [Directory](#) [Preferences](#) [Feedback](#) [Help!](#) [Logout](#)

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

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

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

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

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

© 2001-2004, Alberto Riva, [CHIP](#)

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

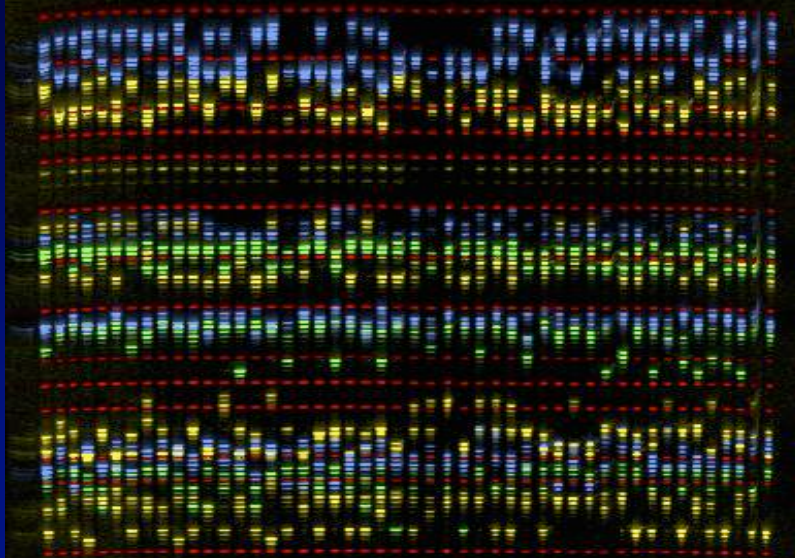
Annotated with human genome build 35.1

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



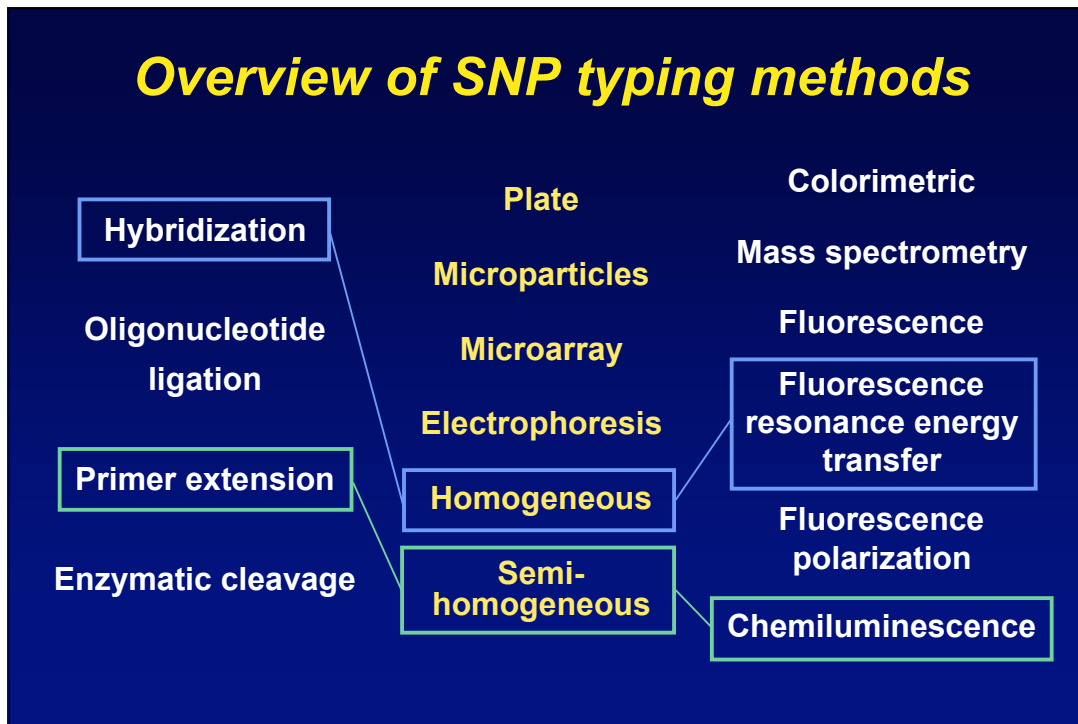
Scoring SNPs

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

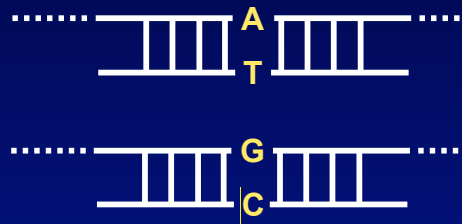
Scoring SNPs (2)

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

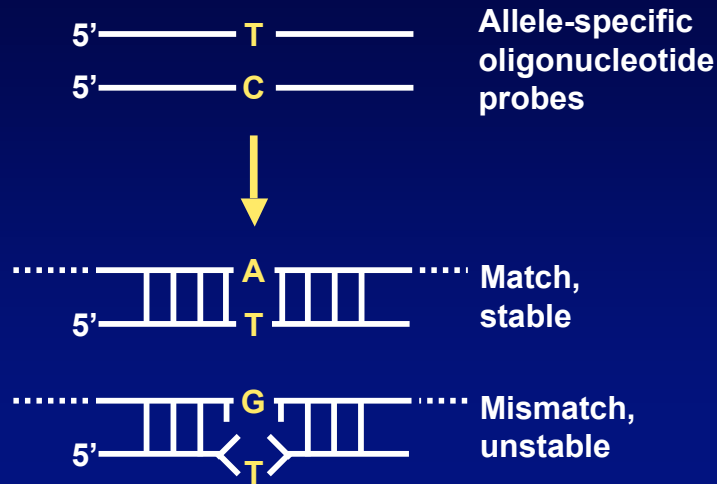
Overview of SNP typing methods



Example SNP



Hybridization



Affymetrix Custom Sequencing Array

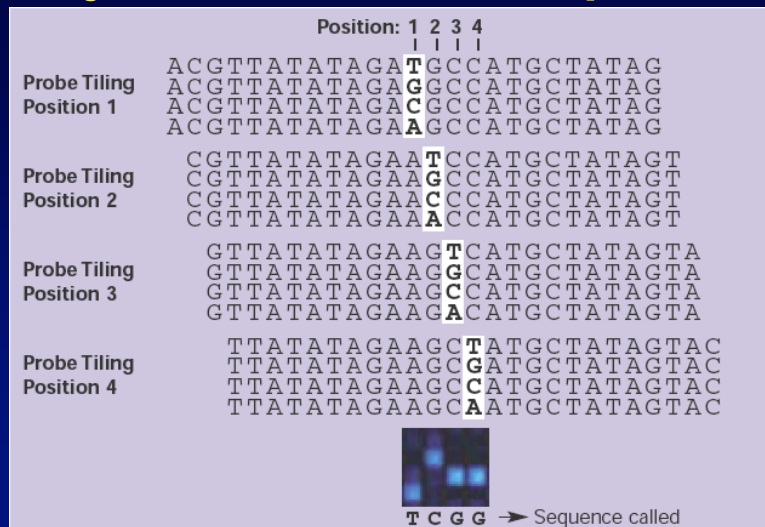


Figure 1: CustomSeq™ arrays tile four probes per strand for each individual base. The central position of each probe varies to incorporate each of the four possible nucleotides—A, C, G, or T.



images from
affymetrix.com

Affymetrix GeneChip 10K Array

Figure 1: GeneChip® Mapping Assay Overview.

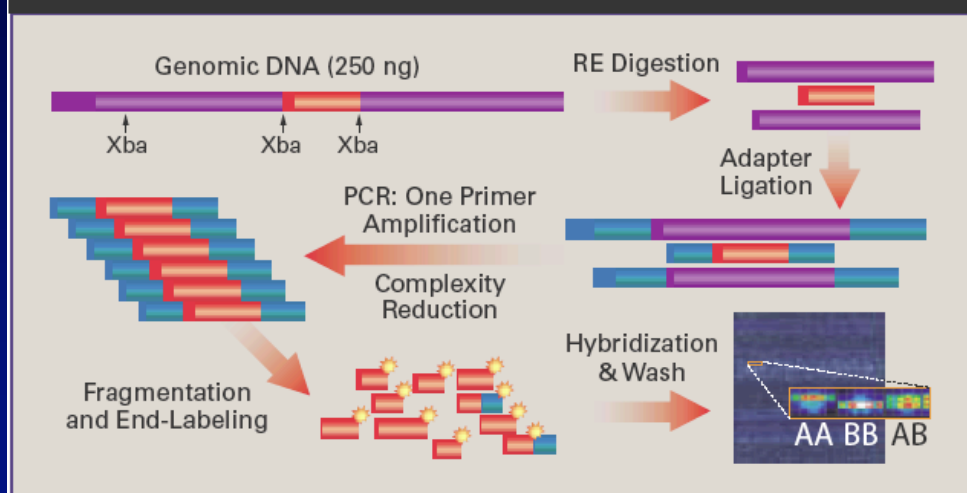
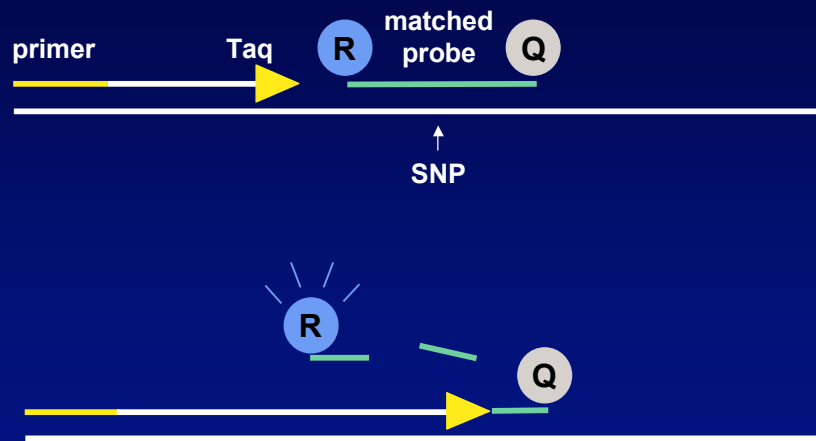


image from affymetrix.com

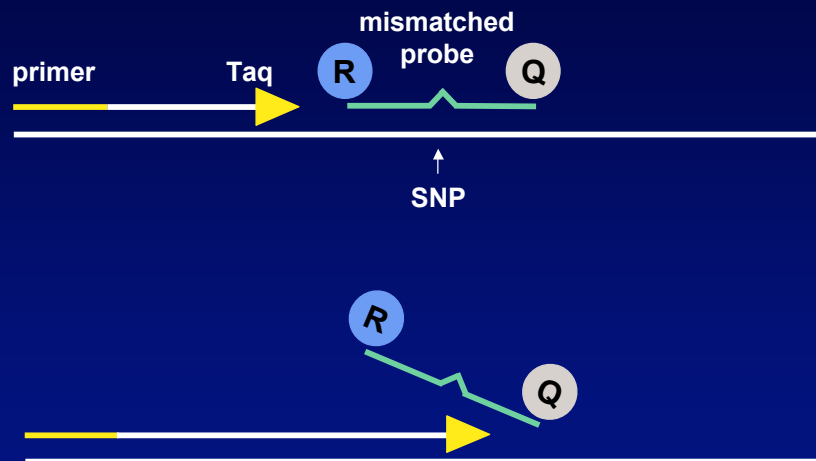
Hybridization to Oligonucleotide Arrays

- **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

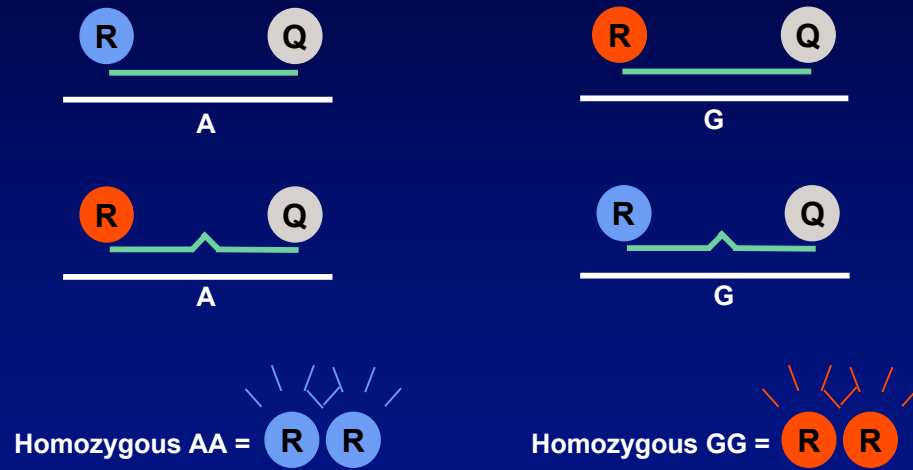
Fluorescence resonance energy transfer (FRET)



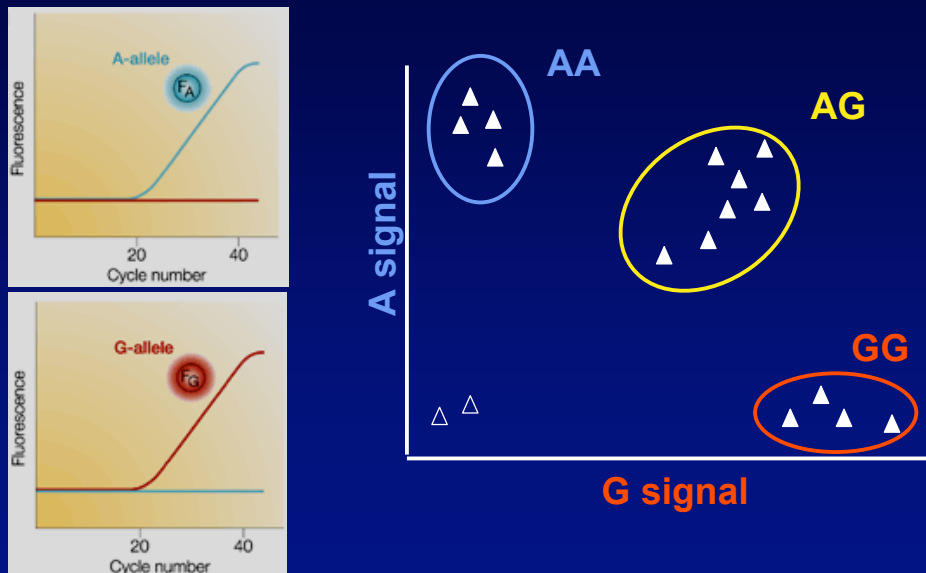
Fluorescence resonance energy transfer (FRET)



TaqMan competing probes



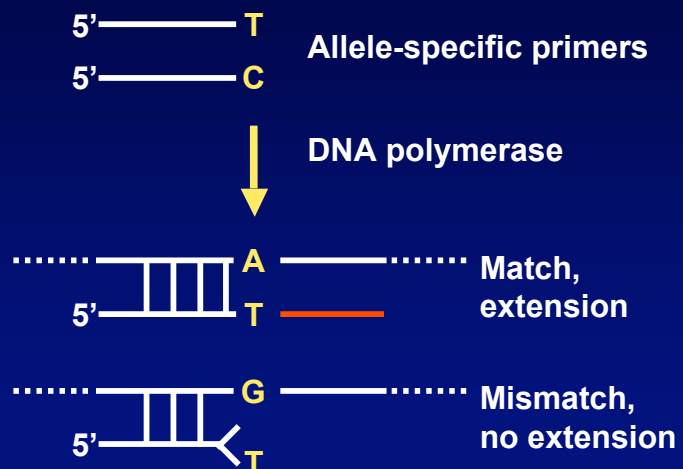
TaqMan genotype scoring



TaqMan

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

Allele-specific PCR



Oligonucleotide Ligation Assay (OLA)

5' ——— T
5' ——— C

Allele-specific
ligation probes

Adjacent
ligation probe

Ligase

..... A
5' ——— T
Match,
ligation

..... G
5' ——— T
Mismatch,
no ligation

Illumina: Allele-specific extension

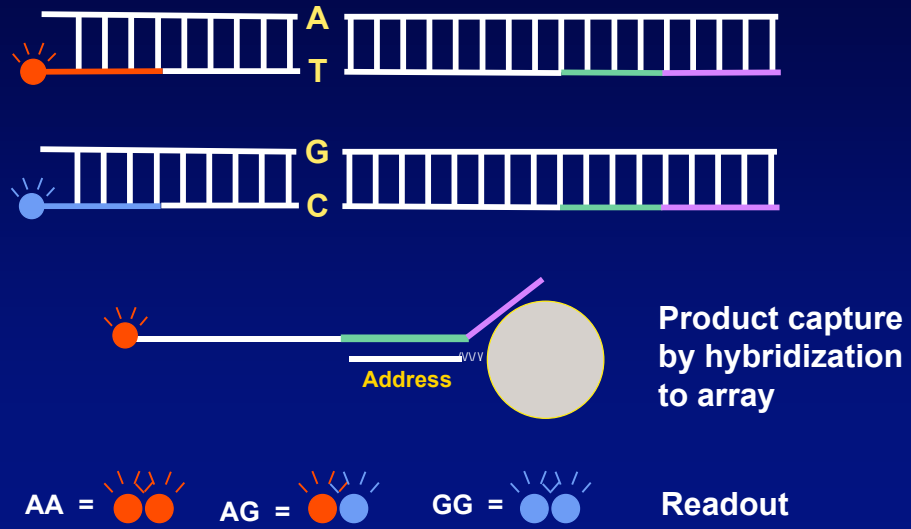
5' ——— T
5' ——— C

Allele-specific
extension

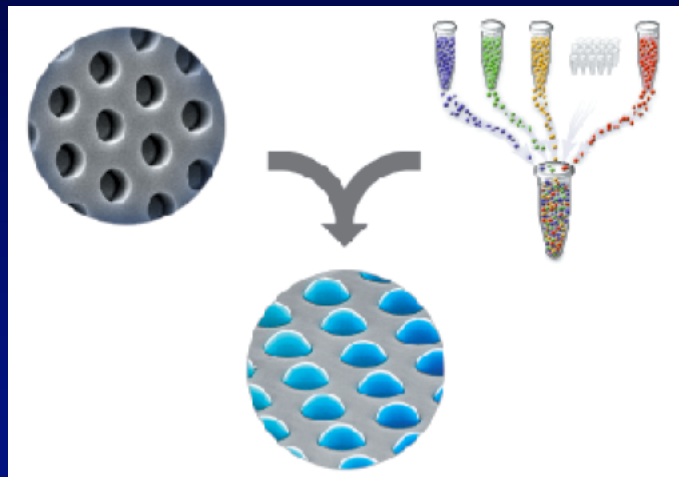
..... A
..... T
Match,
extension

PCR with
common
primers

Illumina: Allele-specific extension



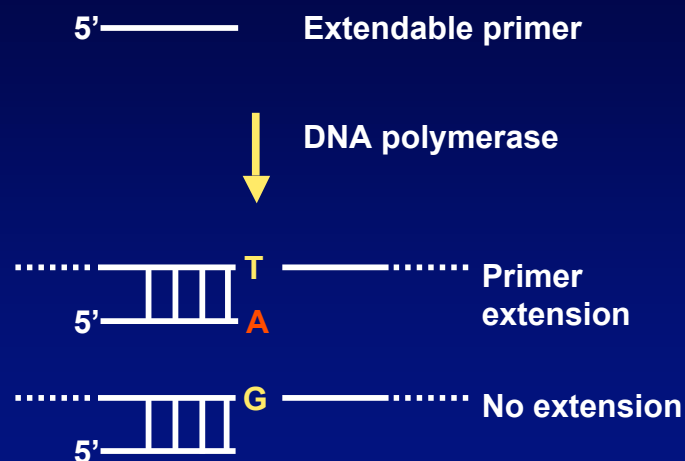
Illumina genotyping technology



Illumina

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

Primer extension = Minisequencing

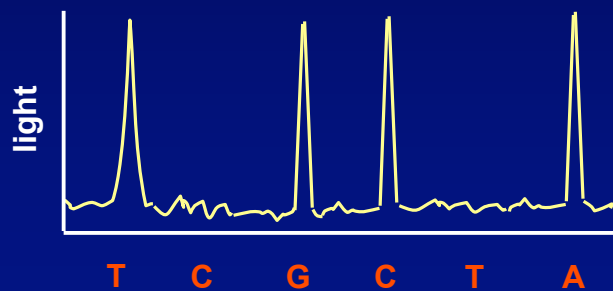


Pyrosequencing

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

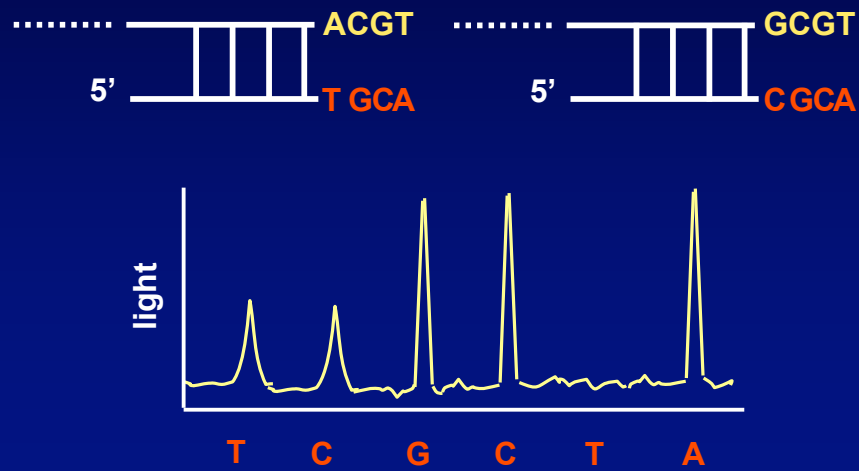
Pyrosequencing

...[A/G]CGT...



Pyrosequencing

...[A/G]CGT...

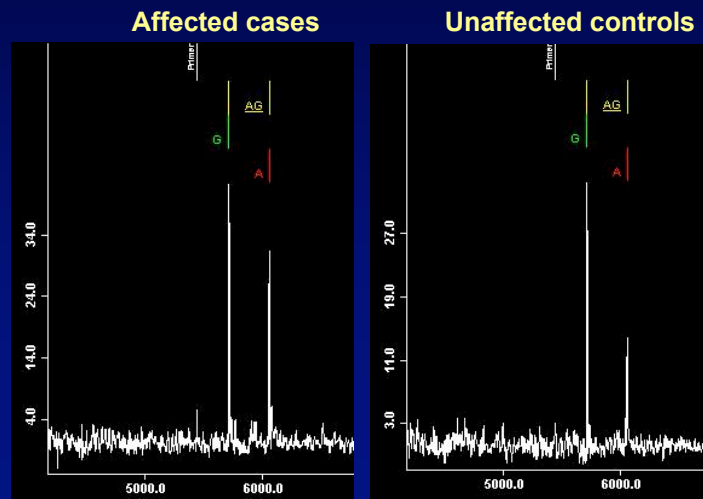


Pyrosequencing

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

Allelic quantification

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



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

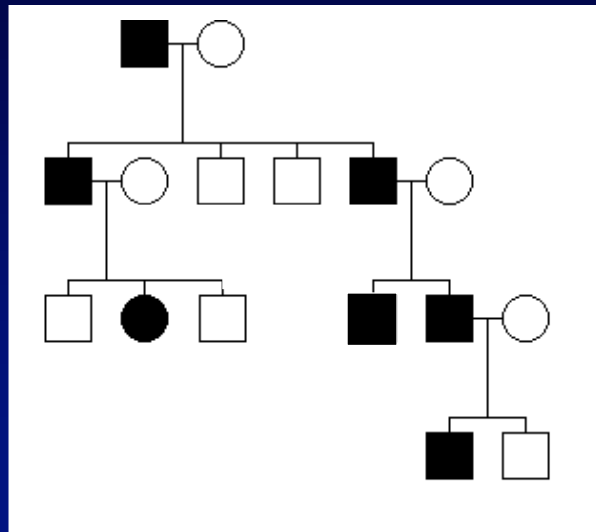
Quality control of genotype data

- 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 χ^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

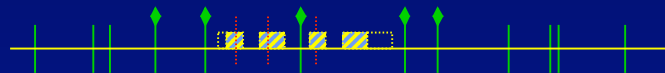


Association Studies

Direct



Indirect



Functional variants

Drug metabolism:
The CYP2D6 gene

... CAC TCC TGA CGC ...

167 168 169
His Ser Stop

Coronary disease:
LDL receptor gene

... TTT TAC GTC ATG ...

289 290 291 292
Phe Tyr Ser Met

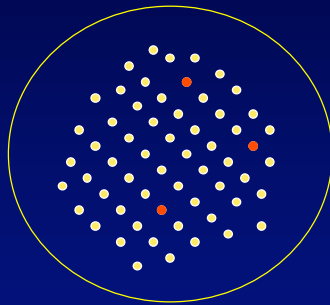
Deep-vein thrombosis:
The Factor V gene

504 505 506 507
Asp Arg Gln Gly

APC cleavage

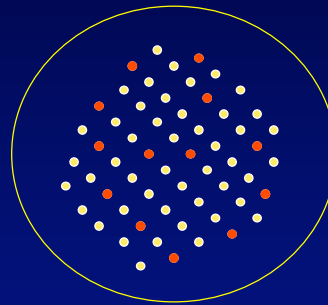
Factor V^{Leiden} association study

301 controls



5% (14) Arg506Gln

301 cases



21% (64) Arg506Gln

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

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

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	hgibase.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