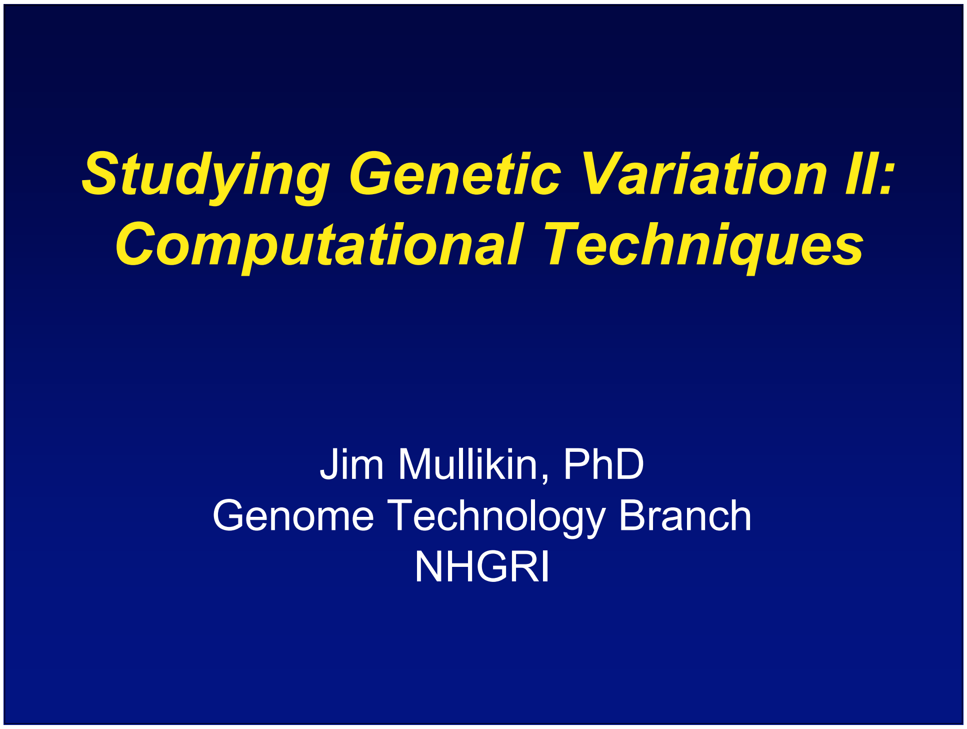
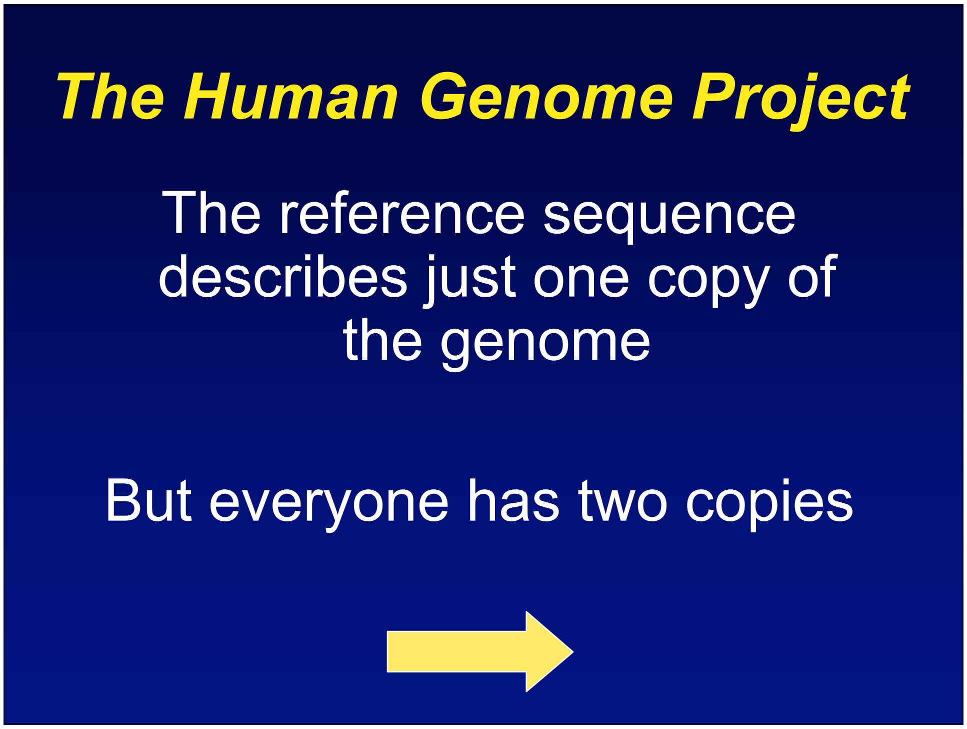


Studying Genetic Variation II: Computational Techniques

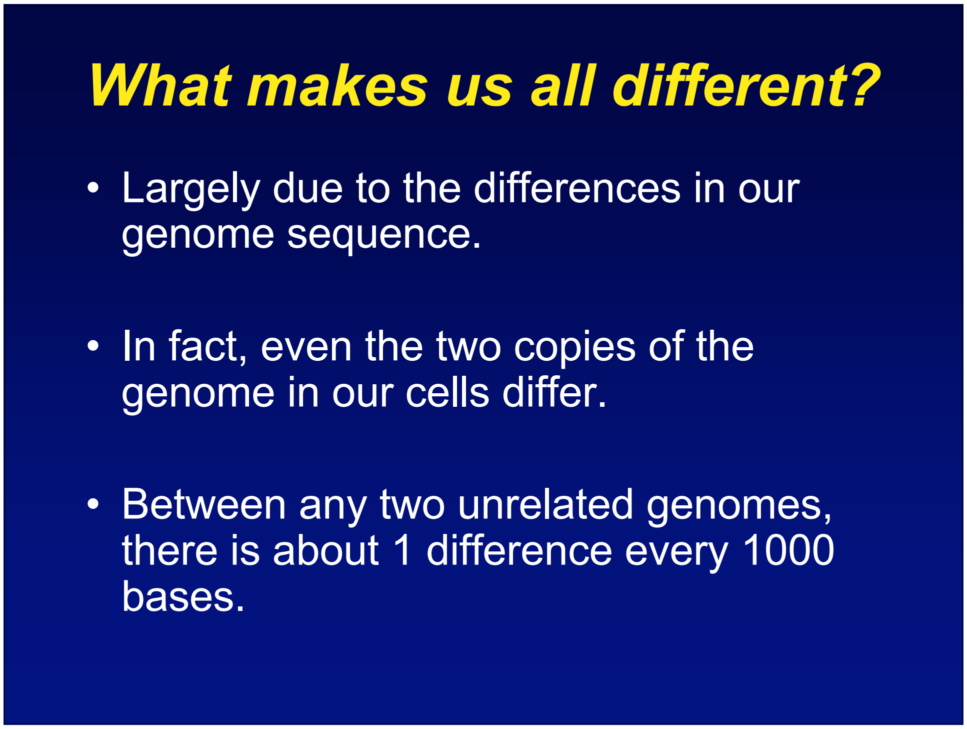
Jim Mullikin, PhD
Genome Technology Branch
NHGRI



The Human Genome Project

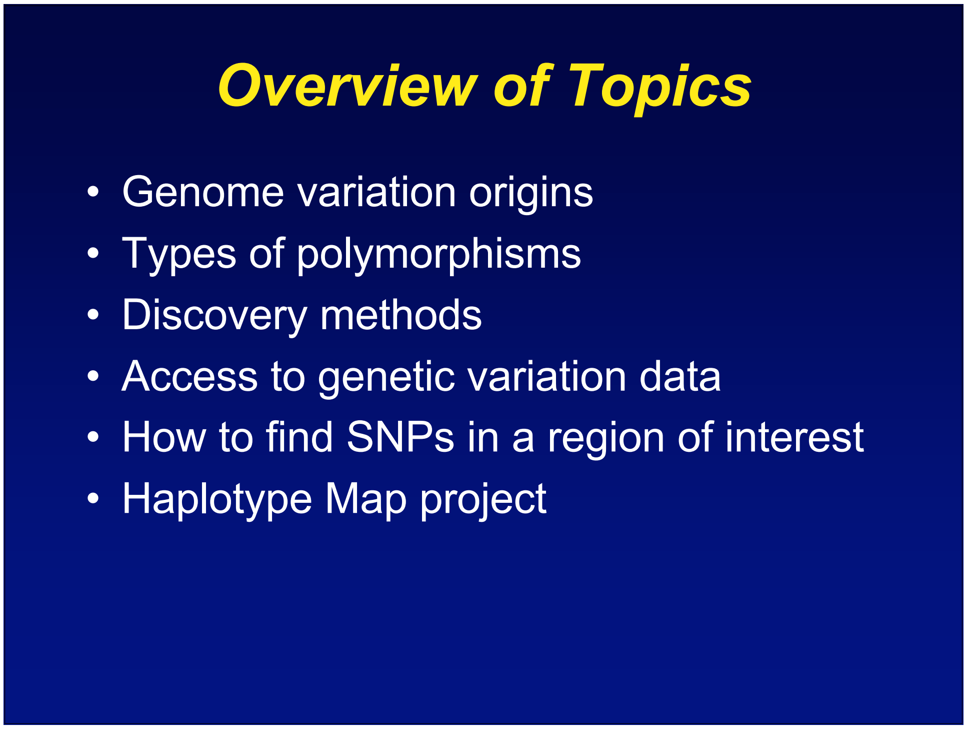
The reference sequence describes just one copy of the genome

But everyone has two copies



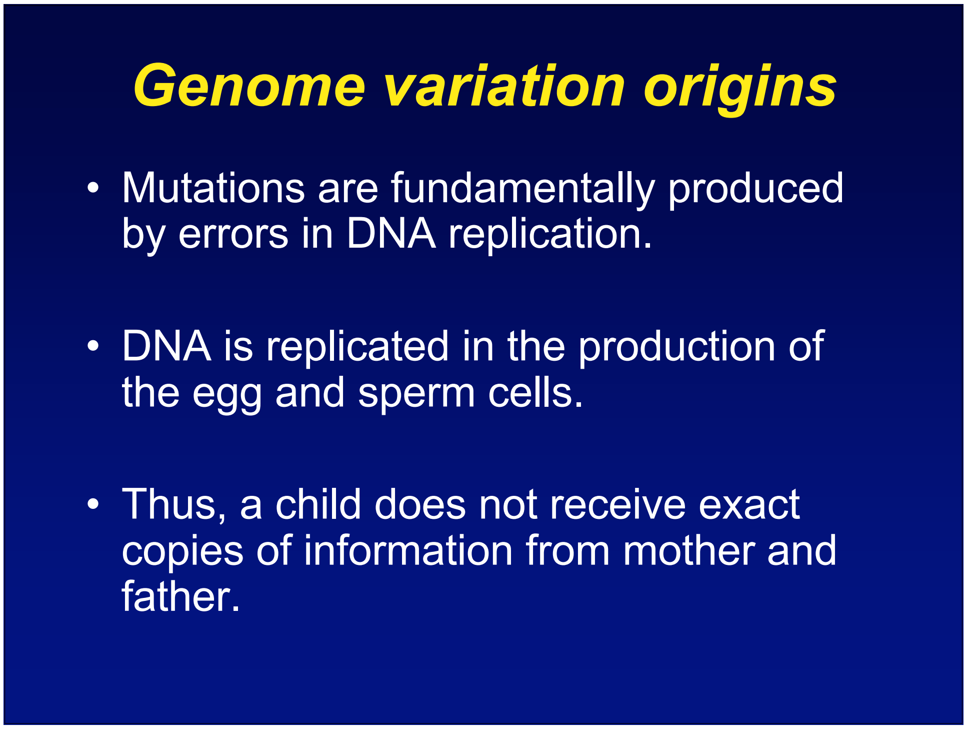
What makes us all different?

- Largely due to the differences in our genome sequence.
- In fact, even the two copies of the genome in our cells differ.
- Between any two unrelated genomes, there is about 1 difference every 1000 bases.



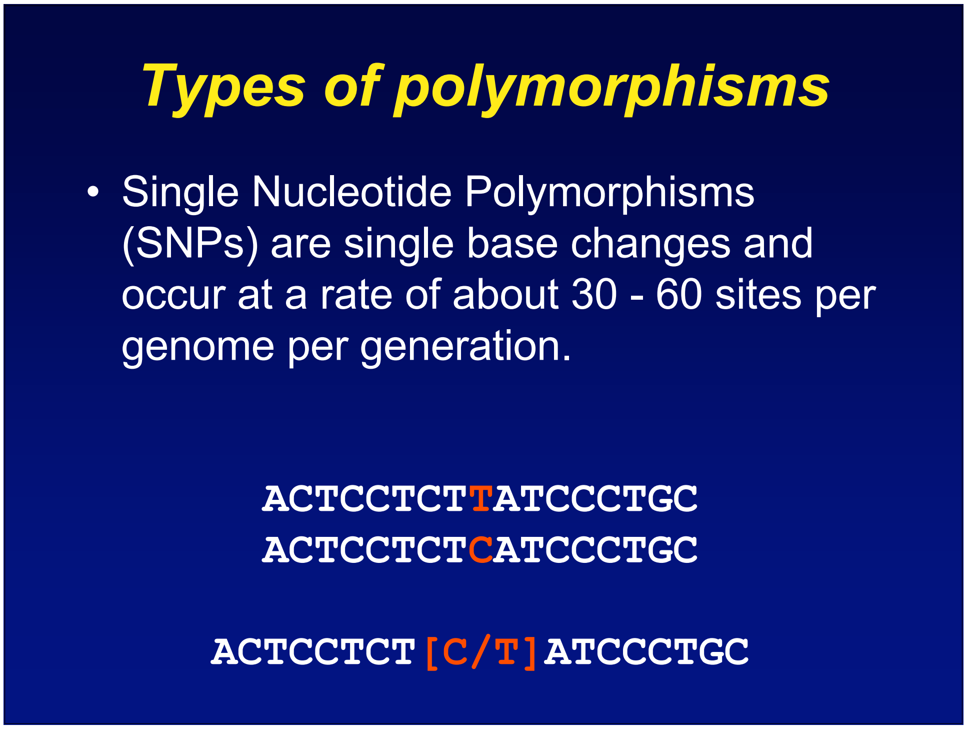
Overview of Topics

- Genome variation origins
- Types of polymorphisms
- Discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project



Genome variation origins

- Mutations are fundamentally produced by errors in DNA replication.
- DNA is replicated in the production of the egg and sperm cells.
- Thus, a child does not receive exact copies of information from mother and father.



Types of polymorphisms

- Single Nucleotide Polymorphisms (SNPs) are single base changes and occur at a rate of about 30 - 60 sites per genome per generation.

```
ACTCCTCTTATCCCTGC
ACTCCTCTCATCCCTGC

ACTCCTCT[C/T]ATCCCTGC
```

Types of polymorphisms

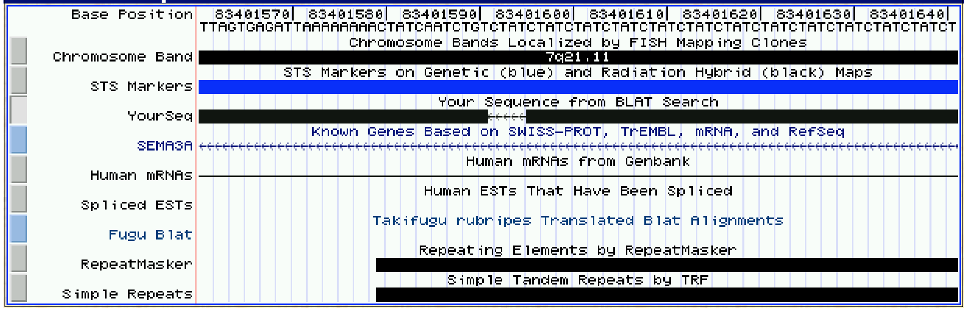
- Short Tandem Repeats (STRs) are specific repeated segments of sequence.

```
GGTTTTTGCC------TATATATATAAGTAGGA
GGTTTTTGCC----TATATATATATAAGTAGGA
GGTTTTTGCC--TATATATATATATAAGTAGGA
GGTTTTTGCCTATATATATATATATAAGTAGGA
```

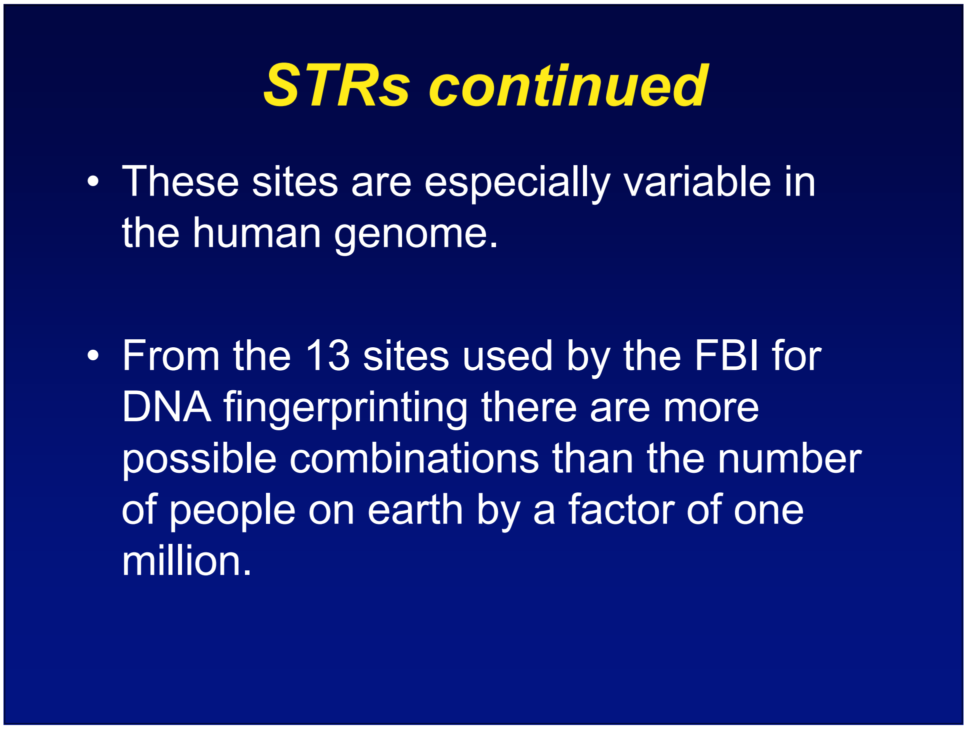
TTGCC[(TA)5/(TA)6/(TA)7/(TA)8]AGT

STRs continued

- These types of polymorphisms are used by the FBI for forensic testing.
- The 13 STRs that are used are all tetrameric, e.g., D7S280 is "gata“ repeated 6-15 times.

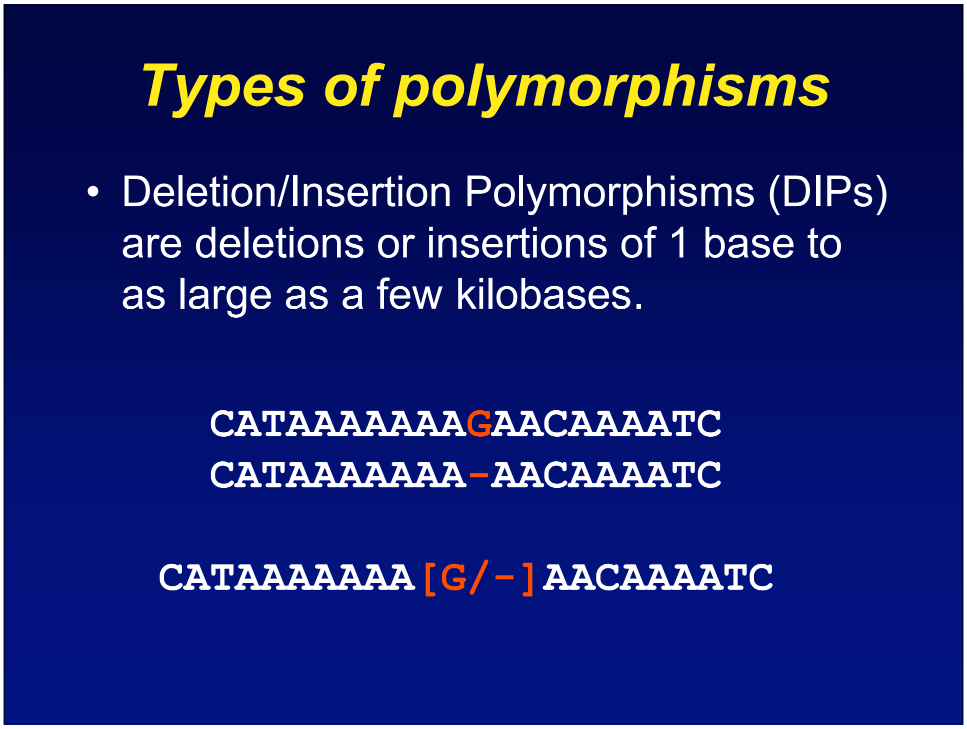


```
Base Position  83401570| 83401580| 83401590| 83401600| 83401610| 83401620| 83401630| 83401640|
TTAGTGAGATTAAAAAAAACTATCAATCTGTCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band  7q21.11
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Your Sequence from BLAT Search
YourSeq
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
SEMA3A
Human mRNAs from Genbank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Takifugu rubripes Translated Blat Alignments
Fugu Blat
Repeating Elements by RepeatMasker
RepeatMasker
Simple Tandem Repeats by TRF
Simple Repeats
```



STRs continued

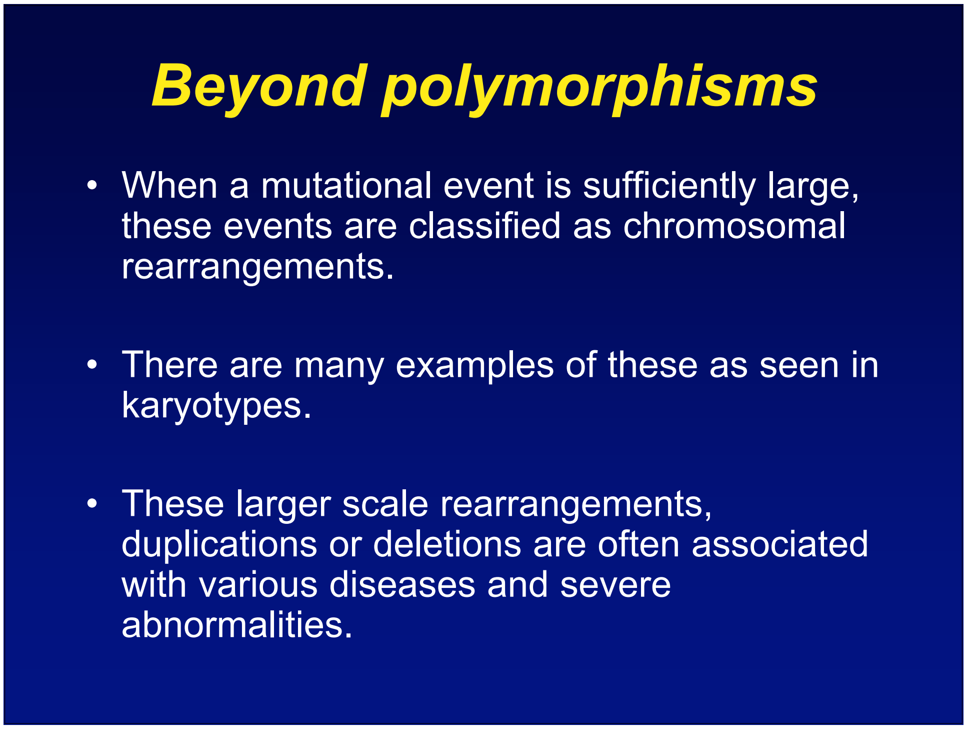
- These sites are especially variable in the human genome.
- From the 13 sites used by the FBI for DNA fingerprinting there are more possible combinations than the number of people on earth by a factor of one million.



Types of polymorphisms

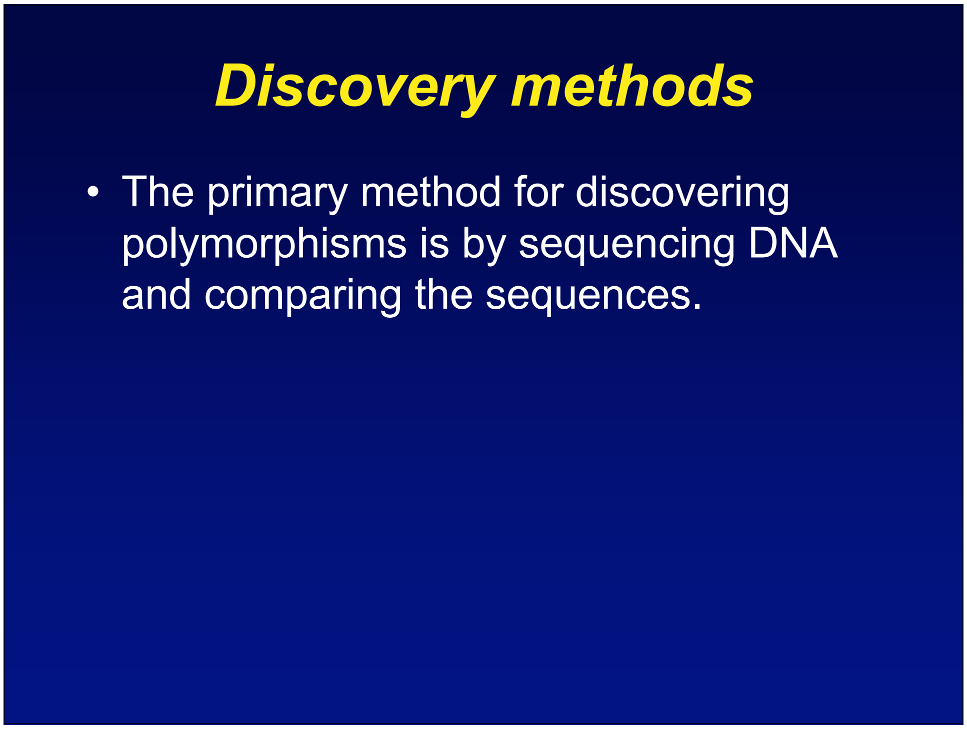
- Deletion/Insertion Polymorphisms (DIPs) are deletions or insertions of 1 base to as large as a few kilobases.

```
CATAAAAAAAGAACAAAATC
CATAAAAAAA-AACAAAATC

CATAAAAAAA[G/-]AACAAAATC
```



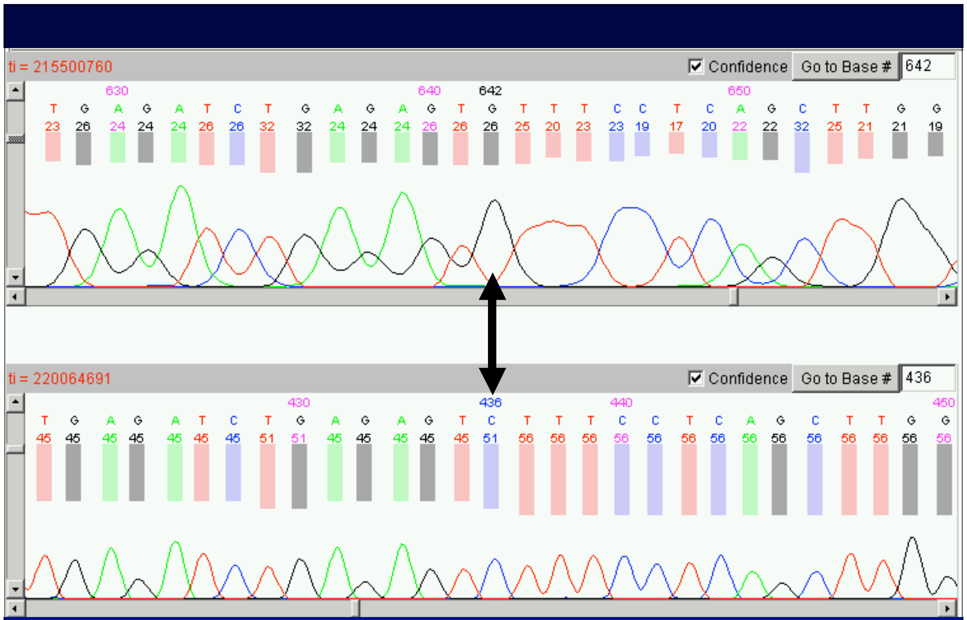
Beyond polymorphisms

- When a mutational event is sufficiently large, these events are classified as chromosomal rearrangements.
- There are many examples of these as seen in karyotypes.
- These larger scale rearrangements, duplications or deletions are often associated with various diseases and severe abnormalities.



Discovery methods

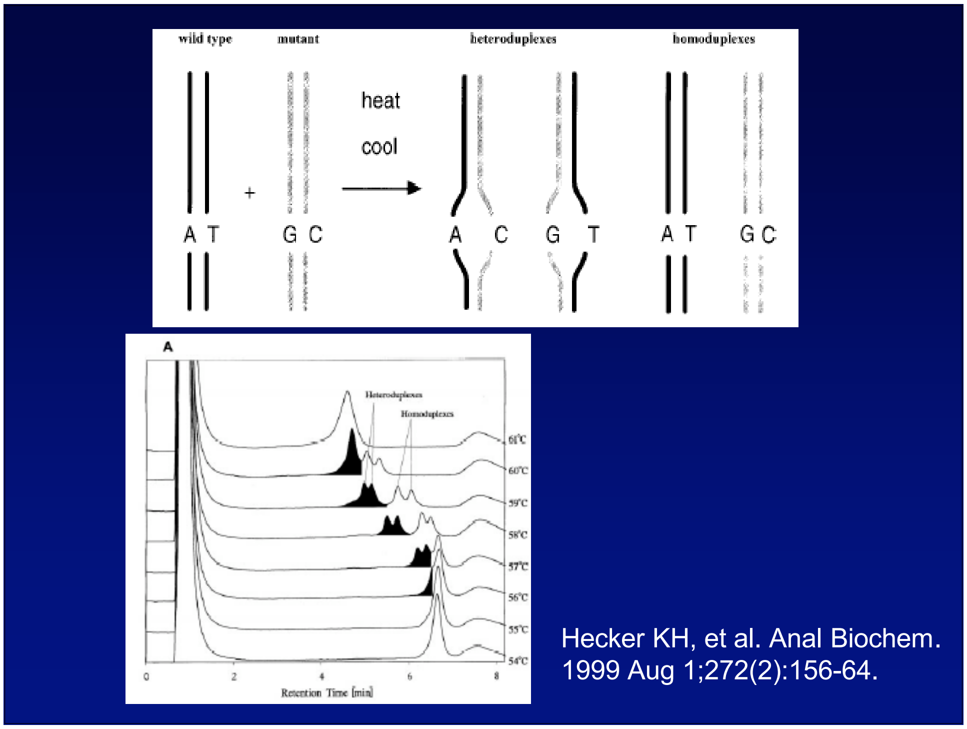
- The primary method for discovering polymorphisms is by sequencing DNA and comparing the sequences.


ti = 215500760
☑ Confidence
Go to Base #
642
630
640
642
650
T G A G A T C T G A G A G T G T T T C C T C A G C T T G G
ti = 220064691
☑ Confidence
Go to Base #
436
430
436
440
450
T G A G A T C T G A G A G T C T T T C C T C A G C T T G G

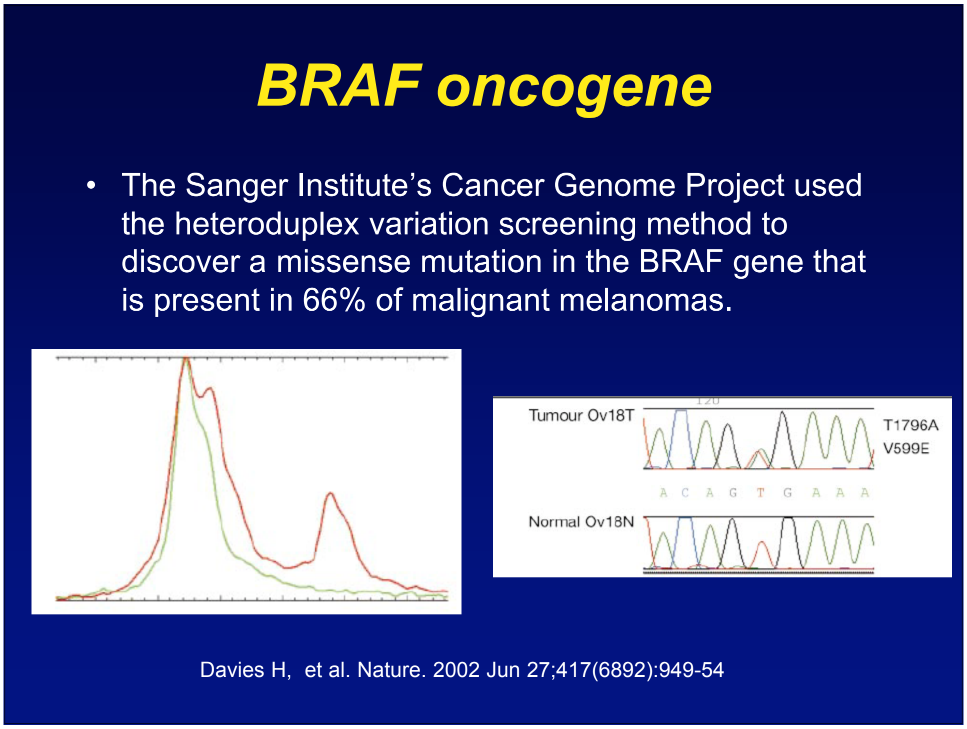
Trace Archive Search query:
trace_name='S213P602260RC9.T0' OR trace_name='50766946'

Discovery methods

- There are other ways to find out if two or more DNAs differ, for example:
  - single strand conformational polymorphism (SSCP)
  - denaturing high performance liquid chromatography (DHPLC)
- These methods do not give specific sequence changes, however, they can be used for rapid mutation screening.



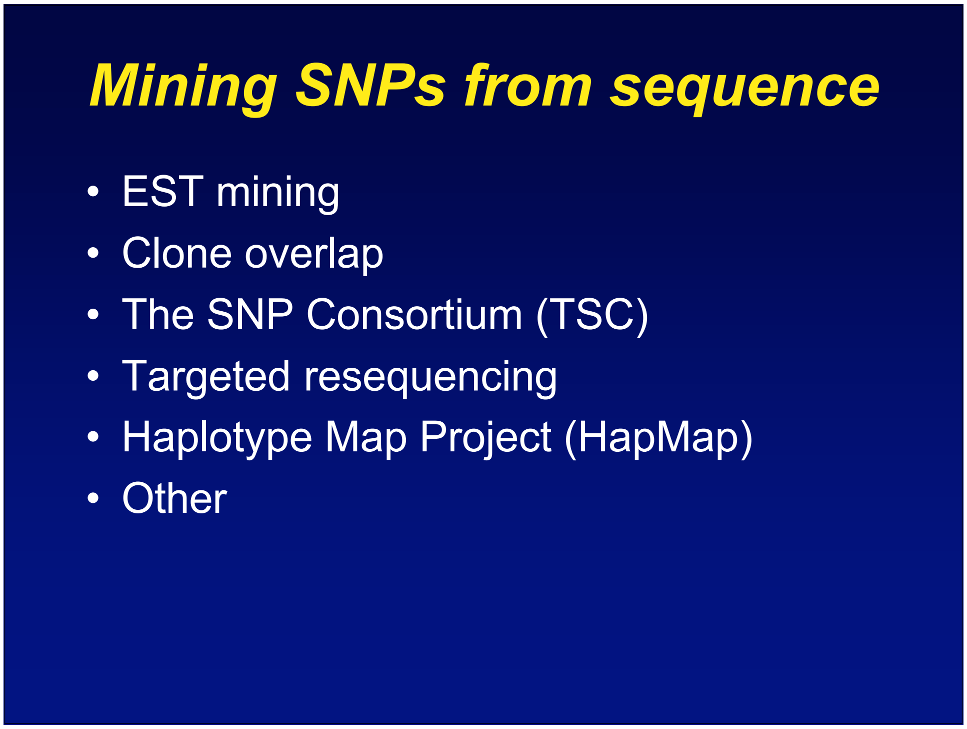
Hecker KH, et al. Anal Biochem. 1999 Aug 1;272(2):156-64.



BRAF oncogene

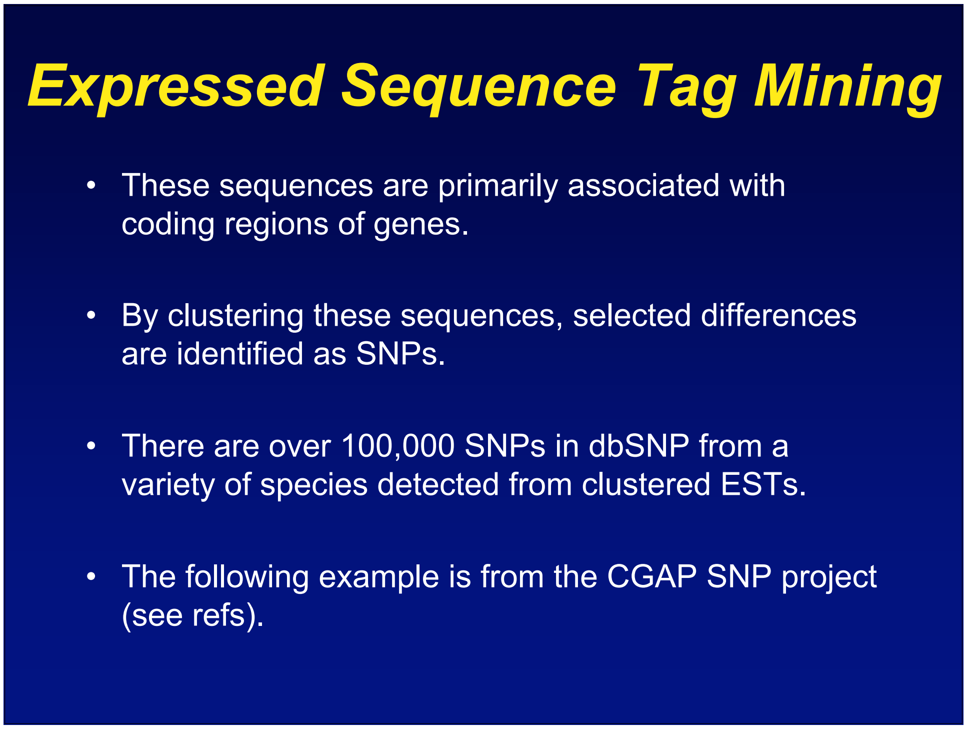
- The Sanger Institute's Cancer Genome Project used the heteroduplex variation screening method to discover a missense mutation in the BRAF gene that is present in 66% of malignant melanomas.

Davies H, et al. Nature. 2002 Jun 27;417(6892):949-54



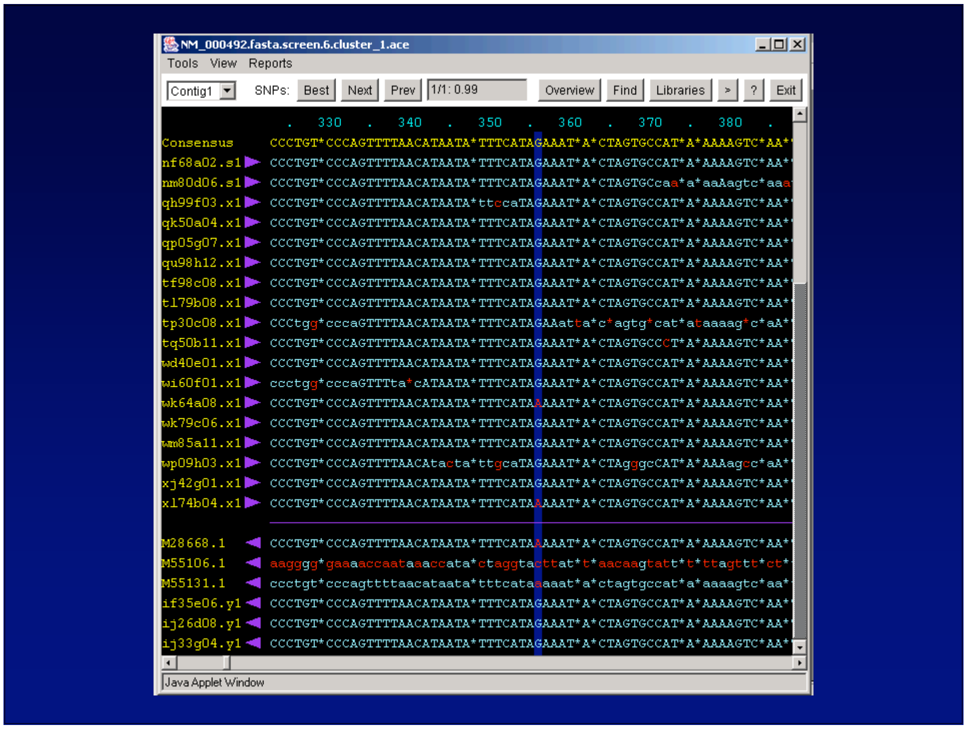
Mining SNPs from sequence

- EST mining
- Clone overlap
- The SNP Consortium (TSC)
- Targeted resequencing
- Haplotype Map Project (HapMap)
- Other



Expressed Sequence Tag Mining

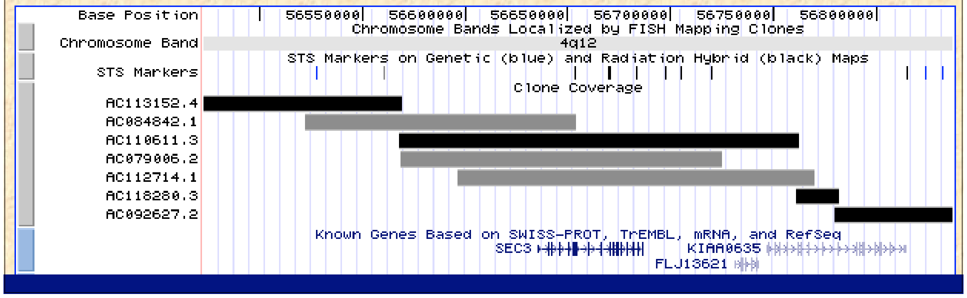
- These sequences are primarily associated with coding regions of genes.
- By clustering these sequences, selected differences are identified as SNPs.
- There are over 100,000 SNPs in dbSNP from a variety of species detected from clustered ESTs.
- The following example is from the CGAP SNP project (see refs).

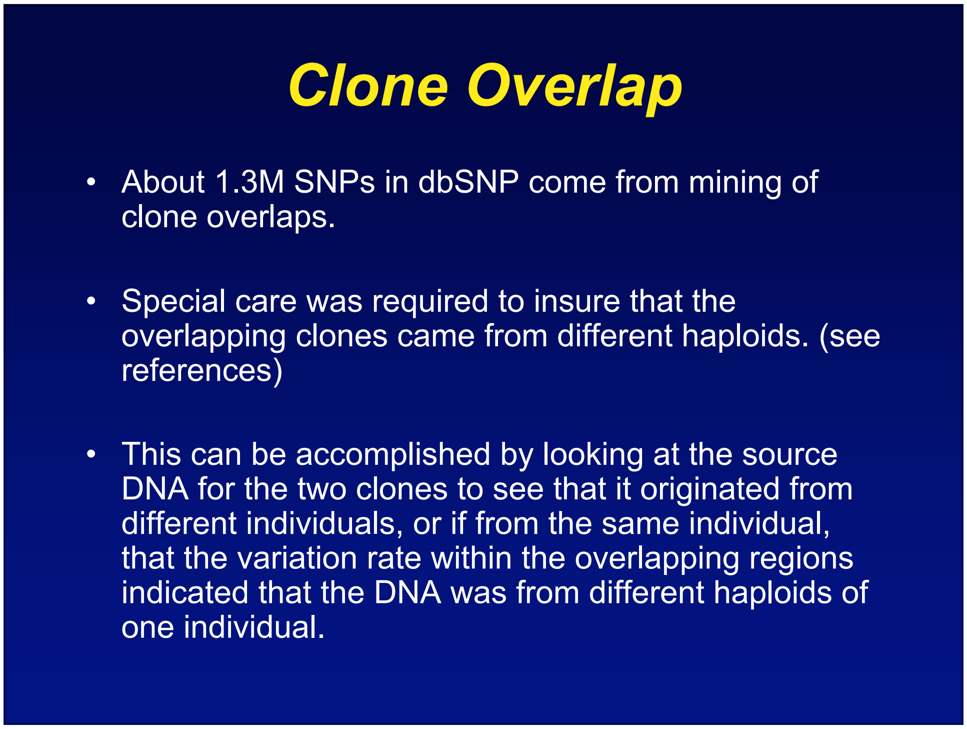


```
               .   330   .   340   .   350   .   360   .   370   .   380   .
Consensus      CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
nf68a02.s1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
nm80d06.s1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCcaa*a*aaAagtc*aaa
qh99f03.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*ttccaTAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
qk50a04.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
qp05g07.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
qu98h12.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
tf98c08.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
tl79b08.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
tp30c08.x1 ▶   CCCtgg*cccaGTTTTAACATAATA*TTTCATAGAAatta*c*agtg*cat*ataaaag*c*aA*
tq50b11.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCCT*A*AAAAGTC*AA*
wd40e01.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
wi60f01.x1 ▶   ccctgg*cccaGTTTta*cATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
wk64a08.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAAAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
wk79c06.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
wm85a11.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
wp09h03.x1 ▶   CCCTGT*CCCAGTTTTAACAtacta*ttgcaTAGAAAT*A*CTAgggcCAT*A*AAAagcc*aA*
xj42g01.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
xl74b04.x1 ▶   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAAAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*

M28668.1   ◀   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAAAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
M55106.1   ◀   aagggg*gaaaaccaataaaccata*ctaggtacttat*t*aacaagtatt*t*ttagttt*ct*
M55131.1   ◀   ccctgt*cccagttttaacataata*tttcataaaaat*a*ctagtgccat*a*aaaagtc*aa*
if35e06.y1 ◀   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
ij26d08.y1 ◀   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
ij33g04.y1 ◀   CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
```

Clone Overlap

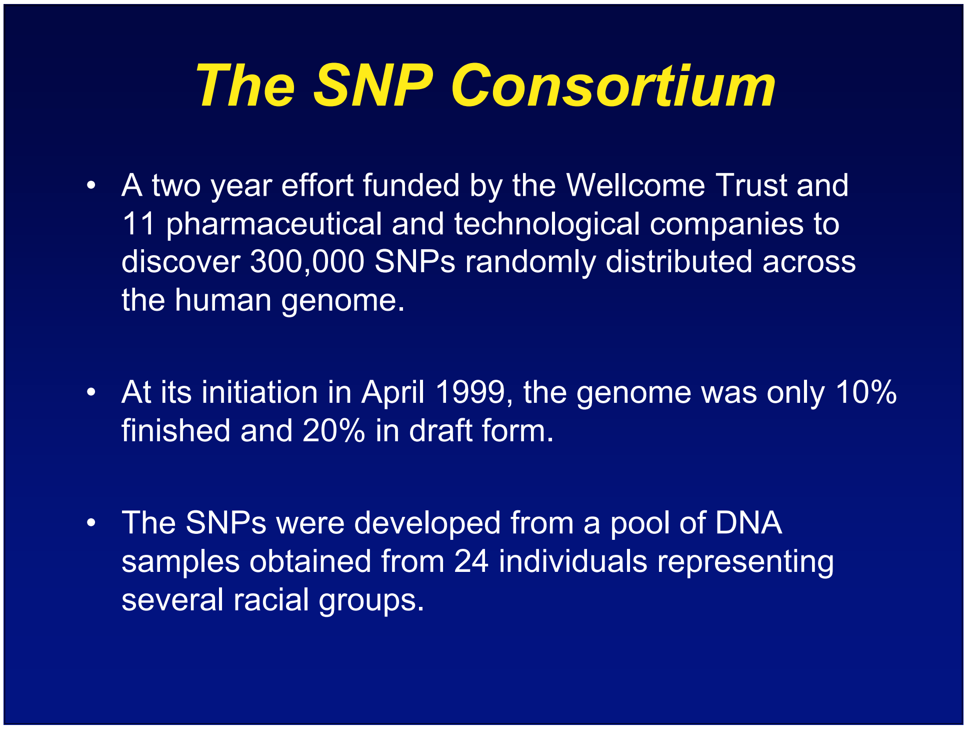
- The human genome was sequenced from BAC clones (containing about 150kb of sequence each).
- These overlapped to various levels, and within the overlap regions, high quality base differences indicated the position and alleles of SNPs.





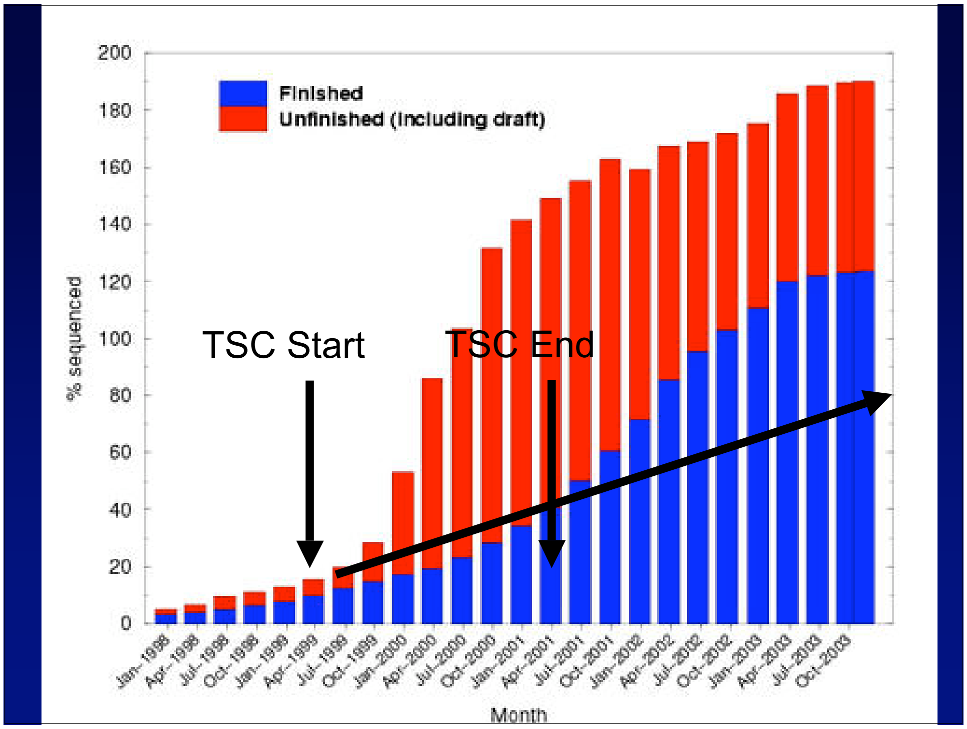
Clone Overlap

- About 1.3M SNPs in dbSNP come from mining of clone overlaps.
- Special care was required to insure that the overlapping clones came from different haploids. (see references)
- This can be accomplished by looking at the source DNA for the two clones to see that it originated from different individuals, or if from the same individual, that the variation rate within the overlapping regions indicated that the DNA was from different haploids of one individual.



The SNP Consortium

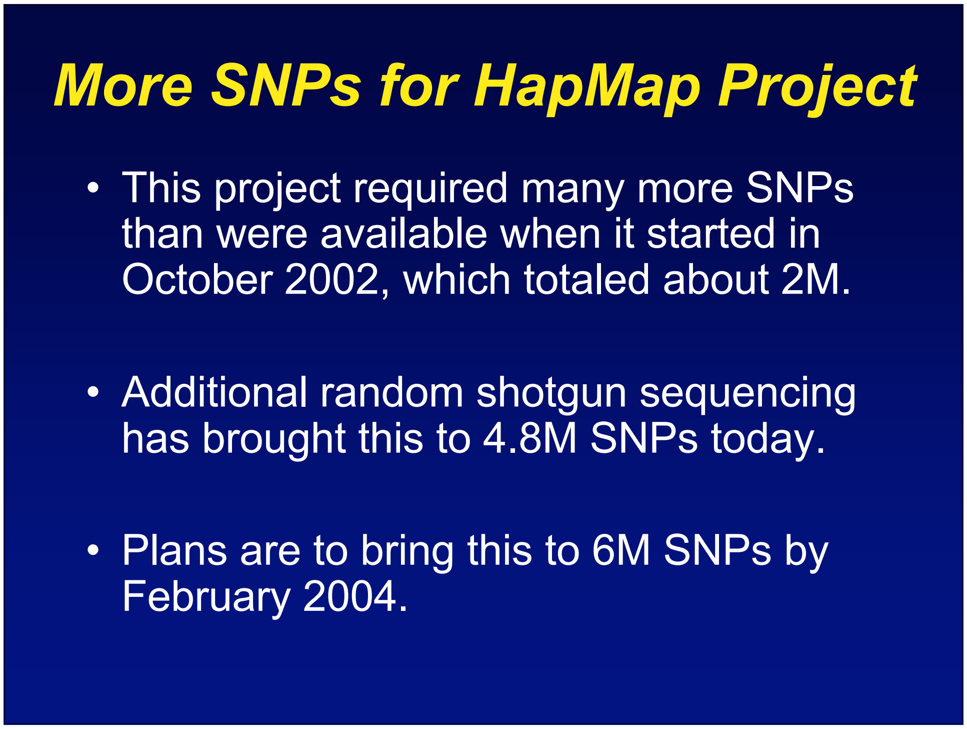
- A two year effort funded by the Wellcome Trust and 11 pharmaceutical and technological companies to discover 300,000 SNPs randomly distributed across the human genome.
- At its initiation in April 1999, the genome was only 10% finished and 20% in draft form.
- The SNPs were developed from a pool of DNA samples obtained from 24 individuals representing several racial groups.



The SNP Consortium

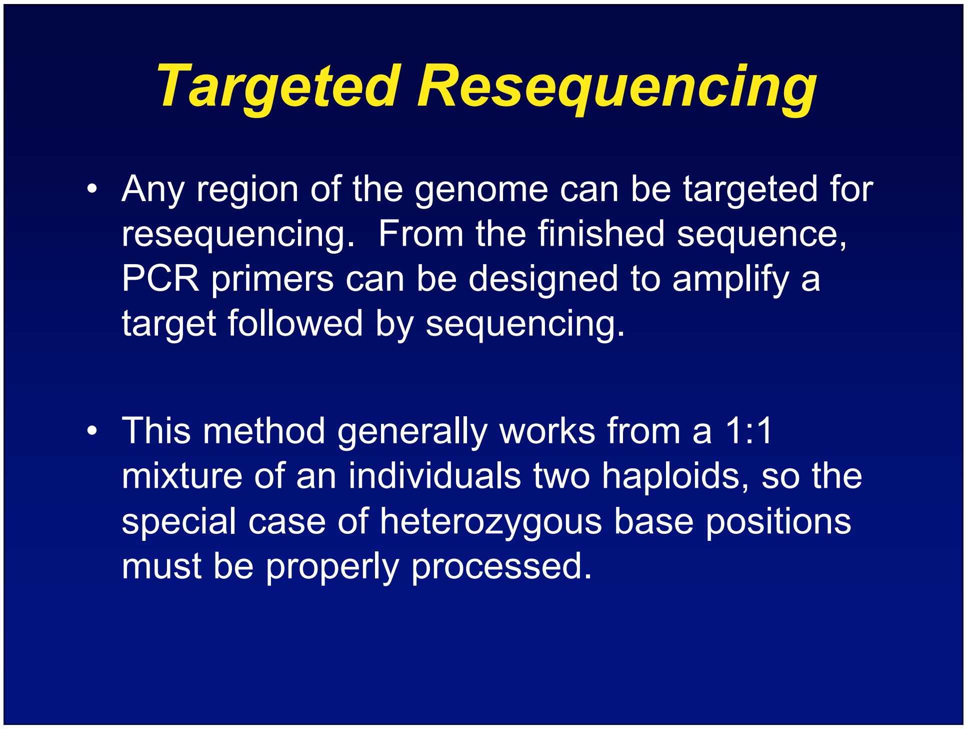
- With the rapid increase in genome coverage from the public Human Genome Project, the strategies changed to take full advantage of the draft and finished sequence.

- The initial target of 300,000 SNP was passed quickly, and now the sequence generated from that project contributes over 1.3M SNPs to the public archives.



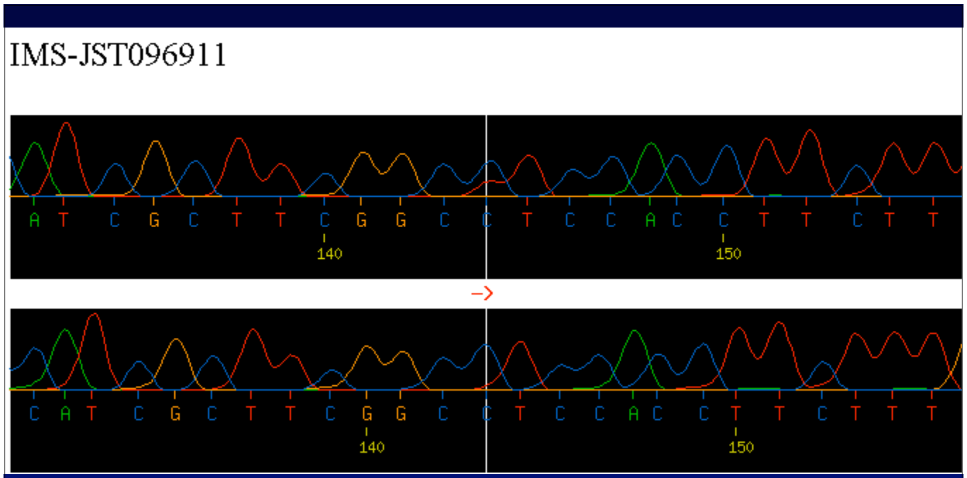
More SNPs for HapMap Project

- This project required many more SNPs than were available when it started in October 2002, which totaled about 2M.
- Additional random shotgun sequencing has brought this to 4.8M SNPs today.
- Plans are to bring this to 6M SNPs by February 2004.



Targeted Resequencing

- Any region of the genome can be targeted for resequencing. From the finished sequence, PCR primers can be designed to amplify a target followed by sequencing.
- This method generally works from a 1:1 mixture of an individuals two haploids, so the special case of heterozygous base positions must be properly processed.



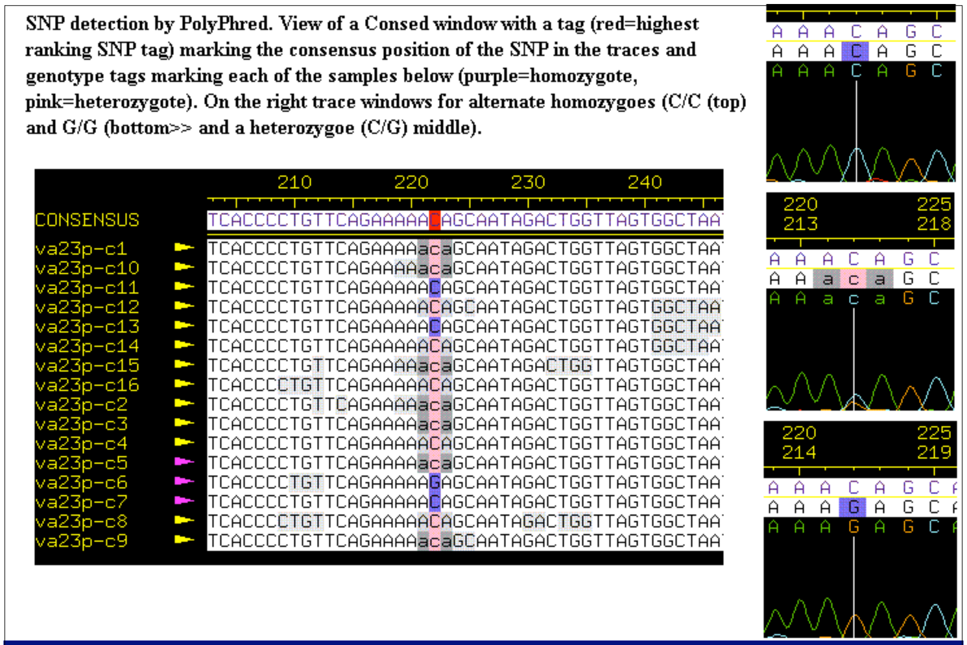
IMS-JST096911

http://snp.ims.u-tokyo.ac.jp/

Chr 19 PTGER1 gcC/gcT A/A

Targeted Resequencing

- JSNP database contains 190,562 SNPs detected from resequencing genomic regions containing genes in DNA from 24 Japanese individuals.
- Many groups use this technique for either SNP discovery in their region of interest, or as a way to validate SNPs.
- PolyPhred (see web links) is commonly used for analyzing resequencing traces.



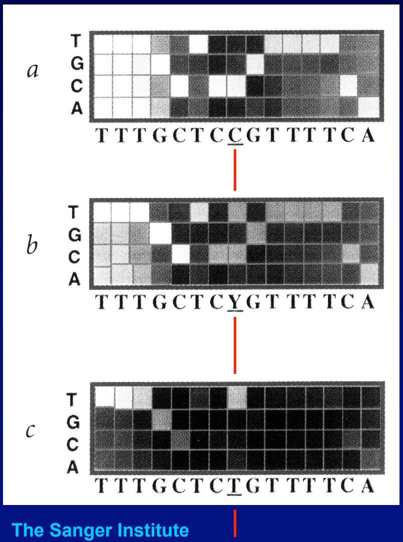
SNP detection by PolyPhred. View of a Consed window with a tag (red=highest ranking SNP tag) marking the consensus position of the SNP in the traces and genotype tags marking each of the samples below (purple=homozygote, pink=heterozygote). On the right trace windows for alternate homozygoes (C/C (top) and G/G (bottom>> and a heterozygoe (C/G) middle).

```
                   210         220         230         240
CONSENSUS      TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c1   ►   TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c10  ►   TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c11  ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c12  ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c13  ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c14  ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c15  ►   TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c16  ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c2   ►   TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c3   ►   TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c4   ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c5   ►   TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c6   ►   TCACCCCTGTTCAGAAAAAGAGCAATAGACTGGTTAGTGGCTAA
va23p-c7   ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c8   ►   TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
va23p-c9   ►   TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
```

PolyPhred example from their web site.

Sequencing Chips

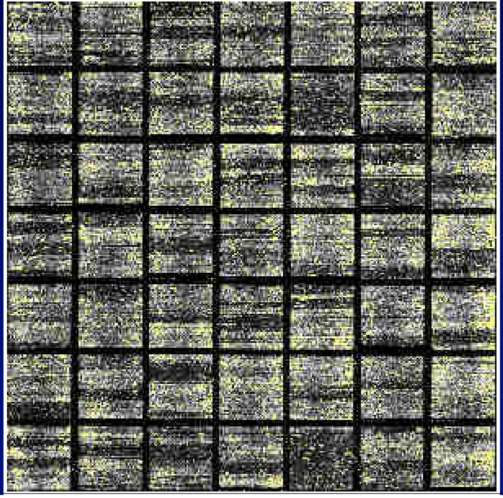
...GCTCCGTTT...
...GCTCTGTTT...



a
T
G
C
A
T T T G C T C C G T T T T C A
b
T
G
C
A
T T T G C T C Y G T T T T C A
c
T
G
C
A
T T T G C T C T G T T T T C A

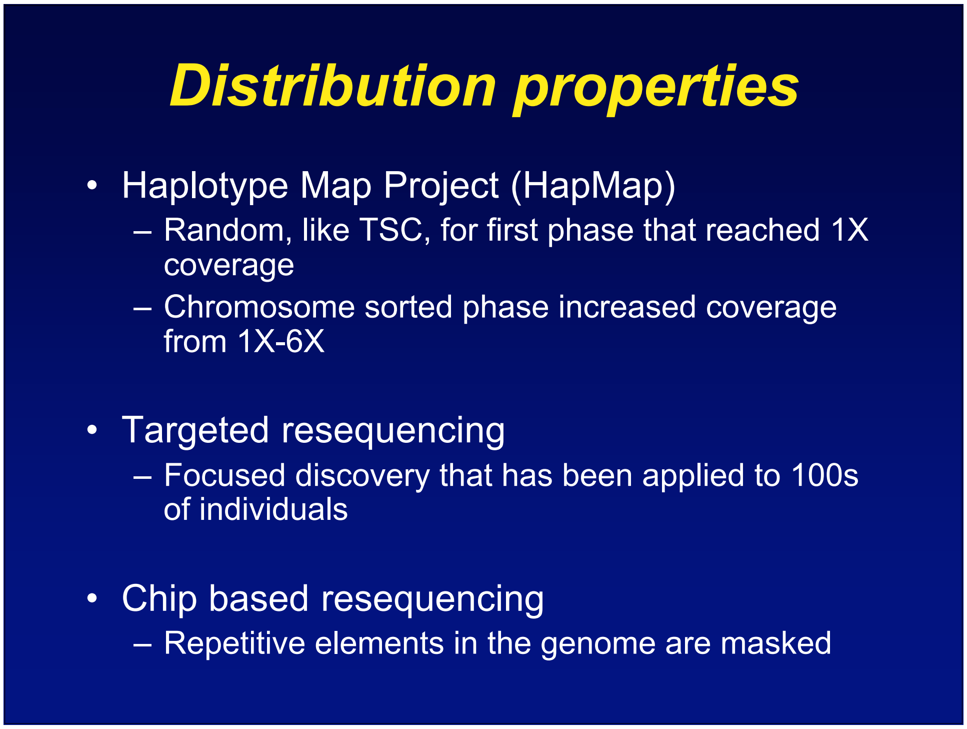
The Sanger Institute

Perlegen used Affymetrix's chip design process to place 60M probes on a 5x5" chip. From 20 single haploid chromosome 21 chromosomes, they discovered 36k SNPs.



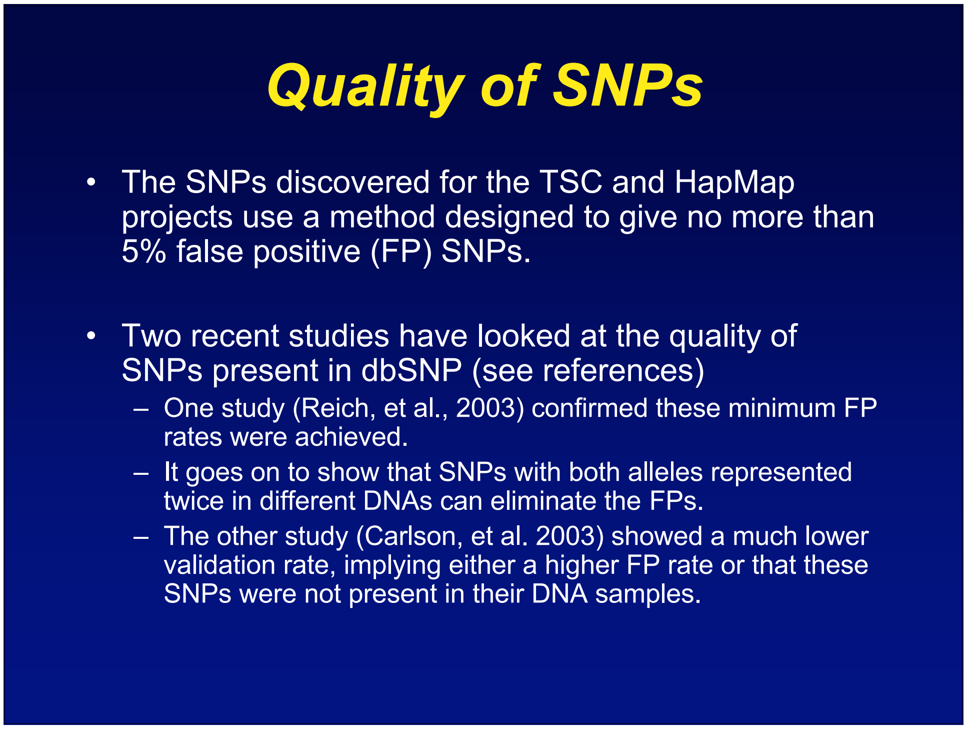
Distribution properties

- EST mining
  - Locates SNPs primarily within coding regions.
- Clone overlap
  - High density of SNPs within overlap regions, absent elsewhere.
- The SNP Consortium (TSC)
  - Randomly distributed across the genome, however, total sequence only covers 50% of the genome



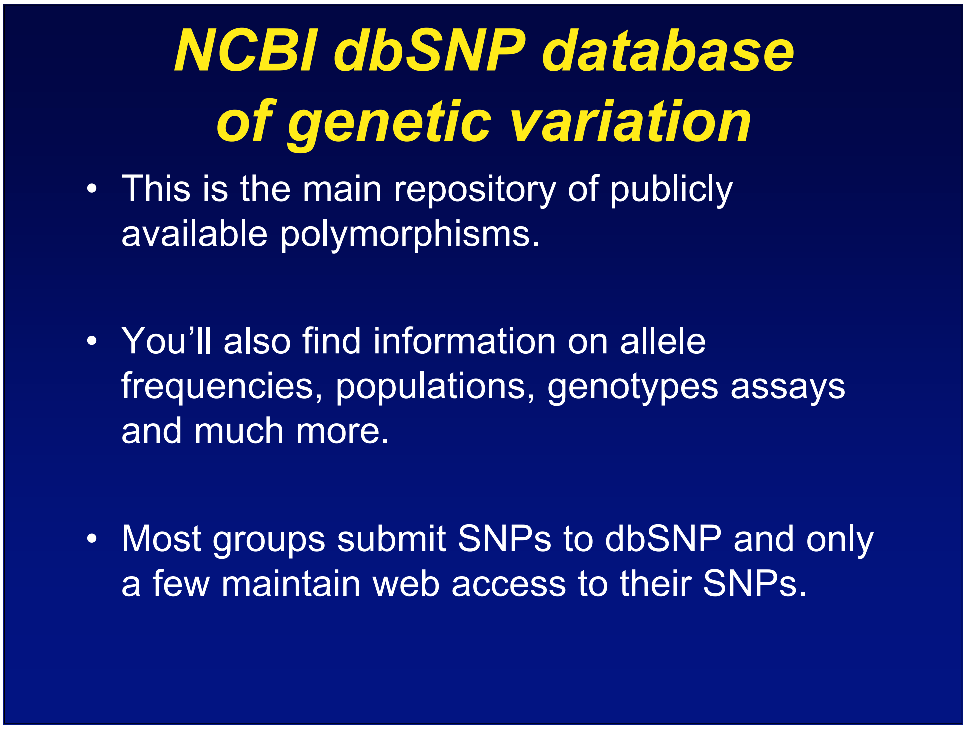
Distribution properties

- Haplotype Map Project (HapMap)
  - Random, like TSC, for first phase that reached 1X coverage
  - Chromosome sorted phase increased coverage from 1X-6X
- Targeted resequencing
  - Focused discovery that has been applied to 100s of individuals
- Chip based resequencing
  - Repetitive elements in the genome are masked



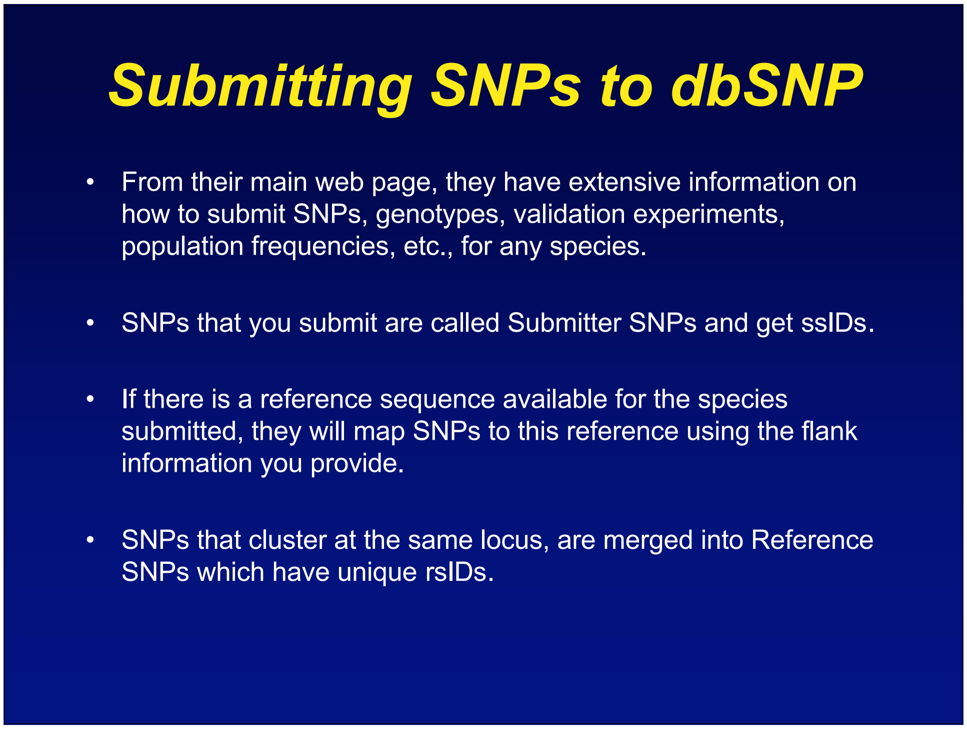
Quality of SNPs

- The SNPs discovered for the TSC and HapMap projects use a method designed to give no more than 5% false positive (FP) SNPs.
- Two recent studies have looked at the quality of SNPs present in dbSNP (see references)
  - One study (Reich, et al., 2003) confirmed these minimum FP rates were achieved.
  - It goes on to show that SNPs with both alleles represented twice in different DNAs can eliminate the FPs.
  - The other study (Carlson, et al. 2003) showed a much lower validation rate, implying either a higher FP rate or that these SNPs were not present in their DNA samples.



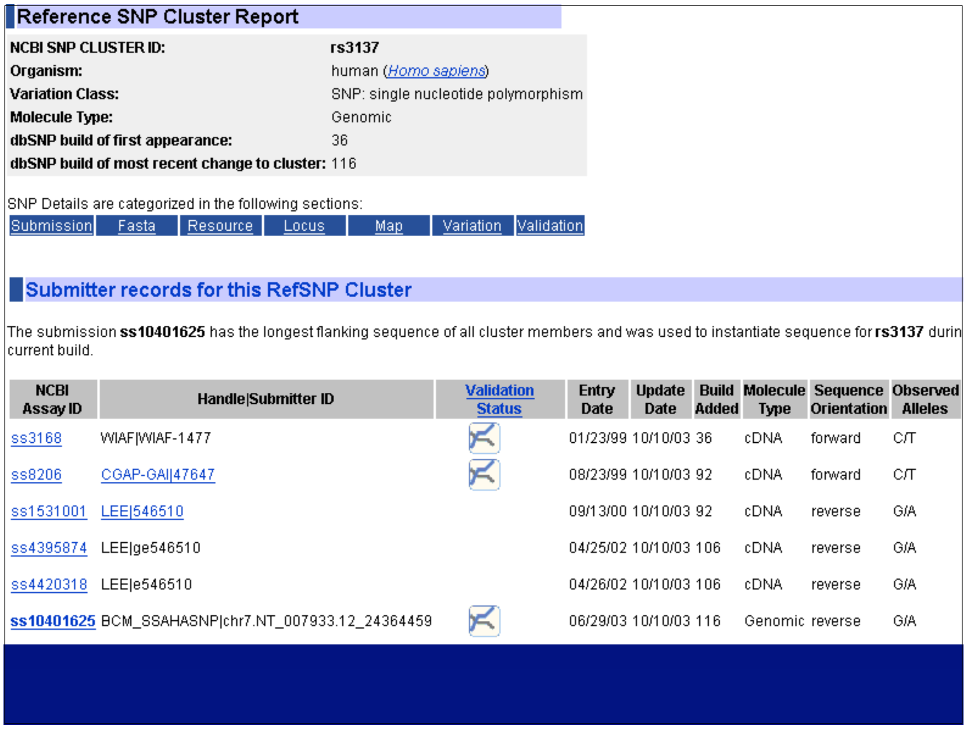
NCBI dbSNP database of genetic variation

- This is the main repository of publicly available polymorphisms.
- You'll also find information on allele frequencies, populations, genotypes assays and much more.
- Most groups submit SNPs to dbSNP and only a few maintain web access to their SNPs.



Submitting SNPs to dbSNP

- From their main web page, they have extensive information on how to submit SNPs, genotypes, validation experiments, population frequencies, etc., for any species.
- SNPs that you submit are called Submitter SNPs and get ssIDs.
- If there is a reference sequence available for the species submitted, they will map SNPs to this reference using the flank information you provide.
- SNPs that cluster at the same locus, are merged into Reference SNPs which have unique rsIDs.



Reference SNP Cluster Report

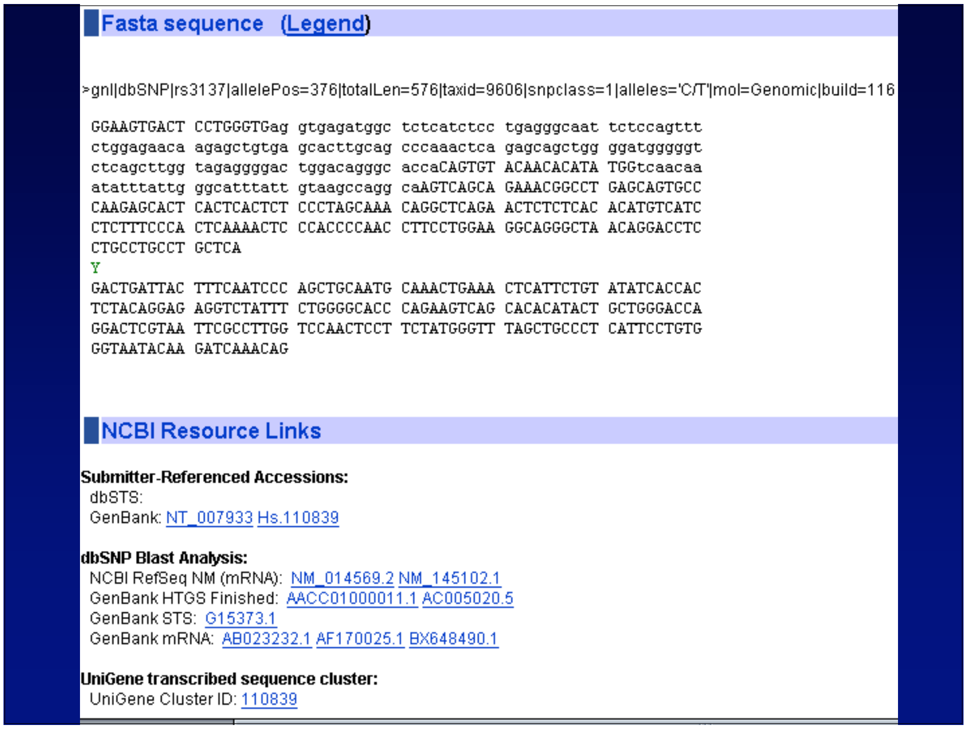
NCBI SNP CLUSTER ID:	**rs3137**
Organism:	human (*Homo sapiens*)
Variation Class:	SNP: single nucleotide polymorphism
Molecule Type:	Genomic
dbSNP build of first appearance:	36
dbSNP build of most recent change to cluster:	116

SNP Details are categorized in the following sections:
Submission | Fasta | Resource | Locus | Map | Variation | Validation

Submitter records for this RefSNP Cluster

The submission **ss10401625** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs3137** durin current build.

NCBI Assay ID	Handle|Submitter ID	Validation Status	Entry Date	Update Date	Build Added	Molecule Type	Sequence Orientation	Observed Alleles
ss3168	WIAF|WIAF-1477	✓	01/23/99	10/10/03	36	cDNA	forward	C/T
ss8206	CGAP-GAI|47647	✓	08/23/99	10/10/03	92	cDNA	forward	C/T
ss1531001	LEE|546510		09/13/00	10/10/03	92	cDNA	reverse	G/A
ss4395874	LEE|ge546510		04/25/02	10/10/03	106	cDNA	reverse	G/A
ss4420318	LEE|e546510		04/26/02	10/10/03	106	cDNA	reverse	G/A
ss10401625	BCM_SSAHASNP|chr7.NT_007933.12_24364459	✓	06/29/03	10/10/03	116	Genomic	reverse	G/A



Fasta sequence (Legend)

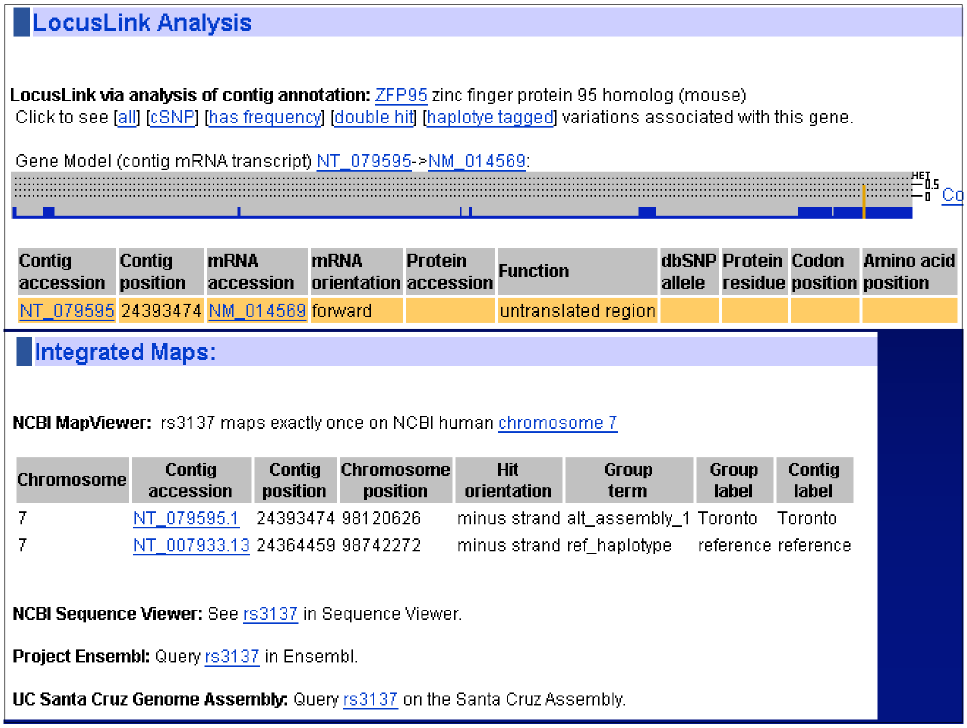
```
>gnl|dbSNP|rs3137|allelePos=376|totalLen=576|taxid=9606|snpclass=1|alleles='C/T'|mol=Genomic|build=116

 GGAAGTGACT CCTGGGTGag gtgagatggc tctcatctcc tgagggcaat tctccagttt
 ctggagaaca agagctgtga gcacttgcag cccaaactca gagcagctgg ggatgggggt
 ctcagcttgg tagaggggac tggacagggc accaCAGTGT ACAACACATA TGGtcaacaa
 atatttattg ggcatttatt gtaagccagg caAGTCAGCA GAAACGGCCT GAGCAGTGCC
 CAAGAGCACT CACTCACTCT CCCTAGCAAA CAGGCTCAGA ACTCTCTCAC ACATGTCATC
 CTCTTTCCCA CTCAAAACTC CCACCCCAAC CTTCCTGGAA GGCAGGGCTA ACAGGACCTC
 CTGCCTGCCT GCTCA
 Y
 GACTGATTAC TTTCAATCCC AGCTGCAATG CAAACTGAAA CTCATTCTGT ATATCACCAC
 TCTACAGGAG AGGTCTATTT CTGGGGCACC CAGAAGTCAG CACACATACT GCTGGGACCA
 GGACTCGTAA TTCGCCTTGG TCCAACTCCT TCTATGGGTT TAGCTGCCCT CATTCCTGTG
 GGTAATACAA GATCAAACAG
```

NCBI Resource Links

Submitter-Referenced Accessions:
dbSTS:
GenBank: NT_007933 Hs.110839

dbSNP Blast Analysis:
NCBI RefSeq NM (mRNA): NM_014569.2 NM_145102.1
GenBank HTGS Finished: AACC01000011.1 AC005020.5
GenBank STS: G15373.1
GenBank mRNA: AB023232.1 AF170025.1 BX648490.1

UniGene transcribed sequence cluster:
UniGene Cluster ID: 110839



LocusLink Analysis

LocusLink via analysis of contig annotation: ZFP95 zinc finger protein 95 homolog (mouse)
Click to see [all] [cSNP] [has frequency] [double hit] [haplotye tagged] variations associated with this gene.

Gene Model (contig mRNA transcript) NT_079595->NM_014569:

Contig accession	Contig position	mRNA accession	mRNA orientation	Protein accession	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
NT_079595	24393474	NM_014569	forward		untranslated region				

Integrated Maps:

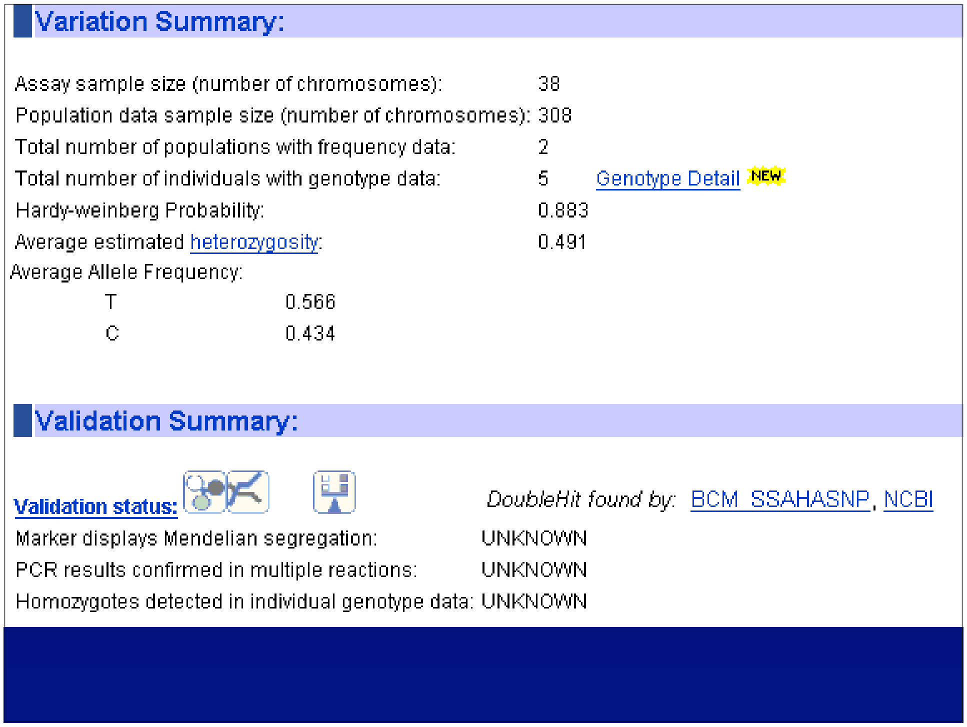
NCBI MapViewer: rs3137 maps exactly once on NCBI human chromosome 7

Chromosome	Contig accession	Contig position	Chromosome position	Hit orientation	Group term	Group label	Contig label
7	NT_079595.1	24393474	98120626	minus strand	alt_assembly_1	Toronto	Toronto
7	NT_007933.13	24364459	98742272	minus strand	ref_haplotype	reference	reference

NCBI Sequence Viewer: See rs3137 in Sequence Viewer.

Project Ensembl: Query rs3137 in Ensembl.

UC Santa Cruz Genome Assembly: Query rs3137 on the Santa Cruz Assembly.



Variation Summary:

Assay sample size (number of chromosomes):	38	
Population data sample size (number of chromosomes):	308	
Total number of populations with frequency data:	2	
Total number of individuals with genotype data:	5	Genotype Detail NEW
Hardy-weinberg Probability:	0.883	
Average estimated heterozygosity:	0.491	

Average Allele Frequency:

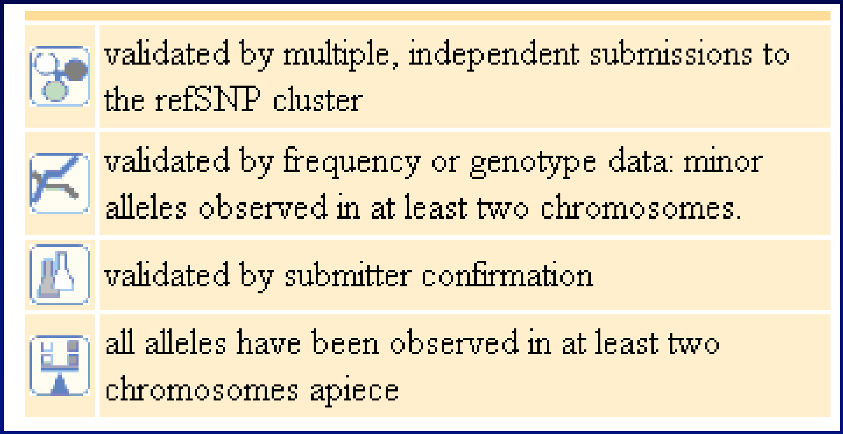
T	0.566
C	0.434

Validation Summary:

Validation status: *DoubleHit found by:* BCM_SSAHASNP, NCBI

Marker displays Mendelian segregation:	UNKNOWN
PCR results confirmed in multiple reactions:	UNKNOWN
Homozygotes detected in individual genotype data:	UNKNOWN

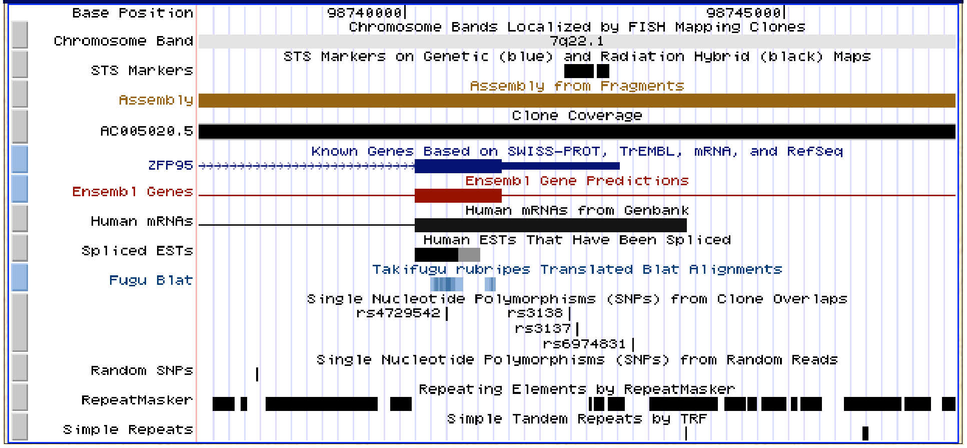
Validation summary



	validated by multiple, independent submissions to the refSNP cluster
	validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
	validated by submitter confirmation
	all alleles have been observed in at least two chromosomes apiece

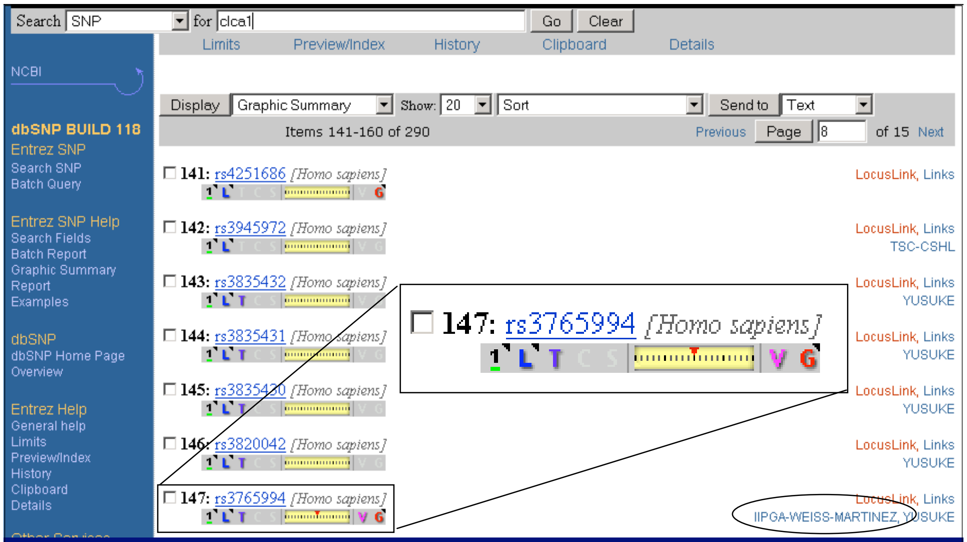
Viewing SNPs in Browsers

NCBI Ensembl UCSC


Base Position
98740000
98745000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
7q22.1
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Assembly from Fragments
Assembly
Clone Coverage
AC005020.5
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
ZFP95
Ensembl Gene Predictions
Ensembl Genes
Human mRNAs from Genbank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Takifugu rubripes Translated Blat Alignments
Fugu Blat
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
rs4729542
rs3138
rs3137
rs6974831
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Repeating Elements by RepeatMasker
RepeatMasker
Simple Tandem Repeats by TRF
Simple Repeats

How to find SNPs in a region of interest

- Gene based example
- A 2 Mbp region



Items 141-160 of 290

□ 141: rs4251686 *[Homo sapiens]* LocusLink, Links

□ 142: rs3945972 *[Homo sapiens]* LocusLink, Links
TSC-CSHL

□ 143: rs3835432 *[Homo sapiens]* LocusLink, Links
YUSUKE

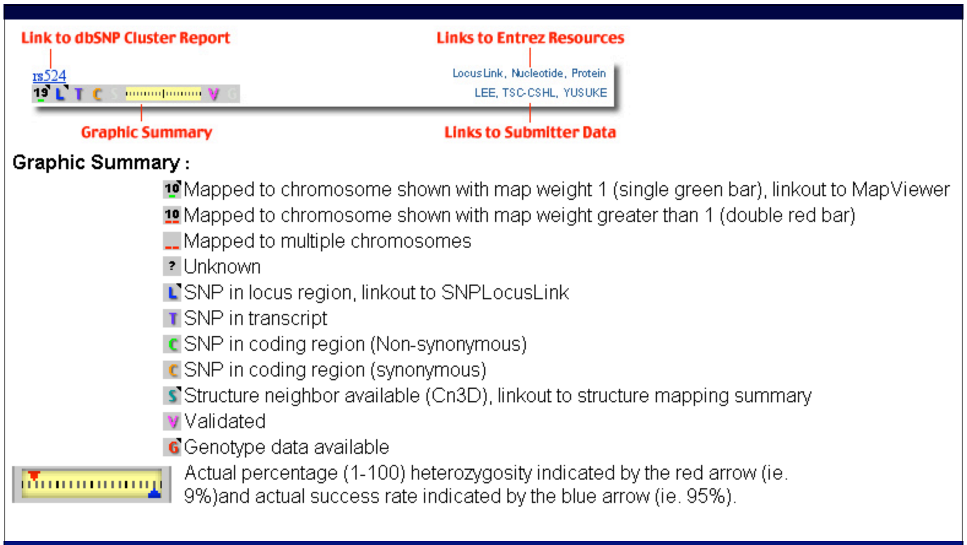
□ 144: rs3835431 *[Homo sapiens]* LocusLink, Links
YUSUKE

□ 145: rs3835430 *[Homo sapiens]* LocusLink, Links
YUSUKE

□ 146: rs3820042 *[Homo sapiens]* LocusLink, Links
YUSUKE

□ 147: rs3765994 *[Homo sapiens]* LocusLink, Links
IIPGA-WEISS-MARTINEZ, YUSUKE

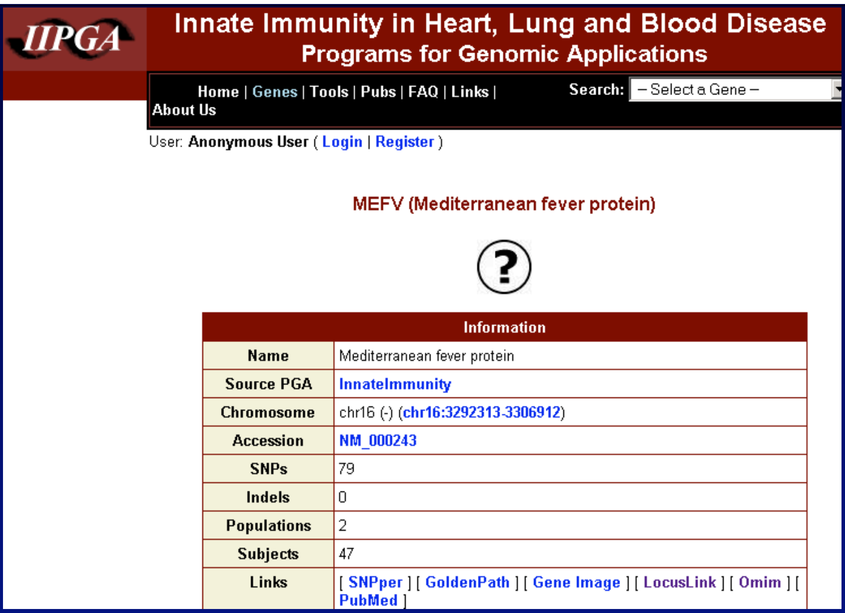
http://www.ncbi.nlm.nih.gov/SNP/index.html


Link to dbSNP Cluster Report
rs524
19 L T C S V G
Links to Entrez Resources
LocusLink, Nucleotide, Protein
LEE, TSC-CSHL, YUSUKE
Graphic Summary
Links to Submitter Data

Graphic Summary :

- 10 Mapped to chromosome shown with map weight 1 (single green bar), linkout to MapViewer
- 10 Mapped to chromosome shown with map weight greater than 1 (double red bar)
- Mapped to multiple chromosomes
- ? Unknown
- L SNP in locus region, linkout to SNPLocusLink
- T SNP in transcript
- C SNP in coding region (Non-synonymous)
- C SNP in coding region (synonymous)
- S Structure neighbor available (Cn3D), linkout to structure mapping summary
- V Validated
- G Genotype data available
- Actual percentage (1-100) heterozygosity indicated by the red arrow (ie. 9%)and actual success rate indicated by the blue arrow (ie. 95%).

http://www.ncbi.nlm.nih.gov/entrez/query/Snp/EntrezSNPlegend.html


IIPGA
Innate Immunity in Heart, Lung and Blood Disease
Programs for Genomic Applications
Home | Genes | Tools | Pubs | FAQ | Links | About Us
Search: -- Select a Gene --

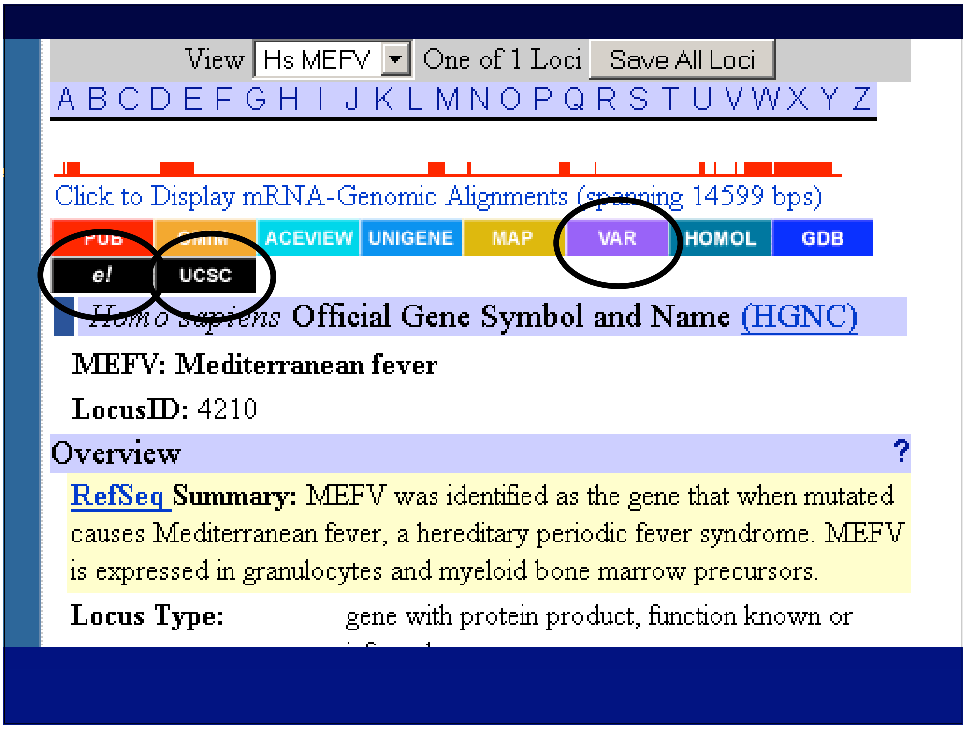
User: **Anonymous User** (**Login** | **Register**)

MEFV (Mediterranean fever protein)

?

Information	
Name	Mediterranean fever protein
Source PGA	**InnateImmunity**
Chromosome	chr16 (-) (**chr16:3292313-3306912**)
Accession	**NM_000243**
SNPs	79
Indels	0
Populations	2
Subjects	47
Links	[**SNPper**] [**GoldenPath**] [**Gene Image**] [**LocusLink**] [**Omim**] [**PubMed**]

http://innateimmunity.net/IIPGA2/PGAs/InnateImmunity/MEFV/



Click to Display mRNA-Genomic Alignments (spanning 14599 bps)

Homo sapiens **Official Gene Symbol and Name** (HGNC)

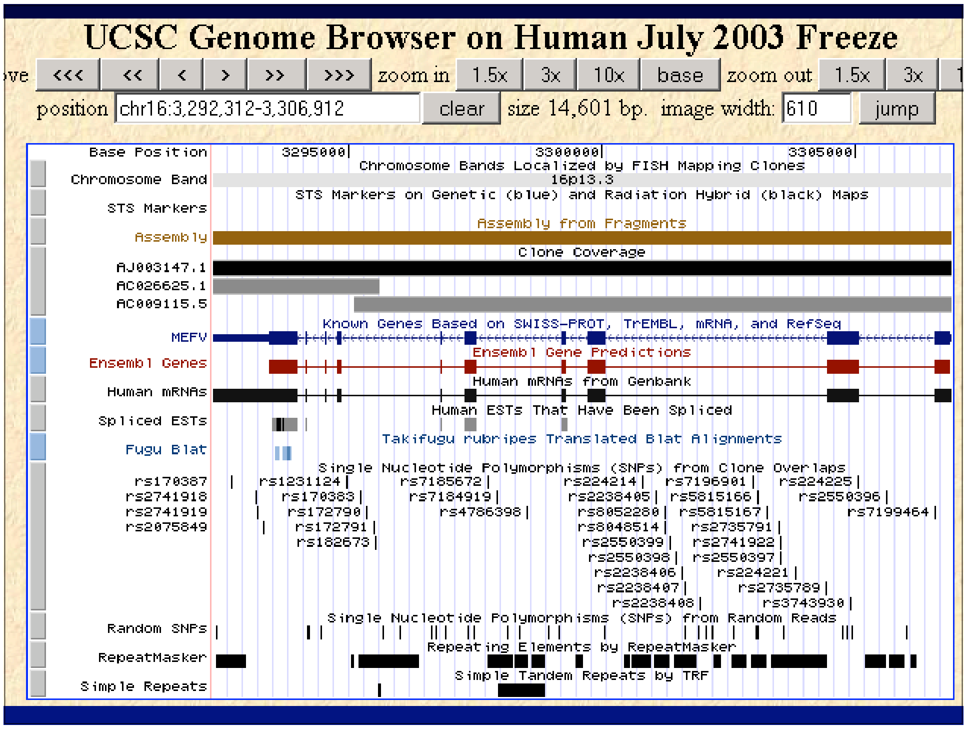
MEFV: Mediterranean fever

LocusID: 4210

Overview ?

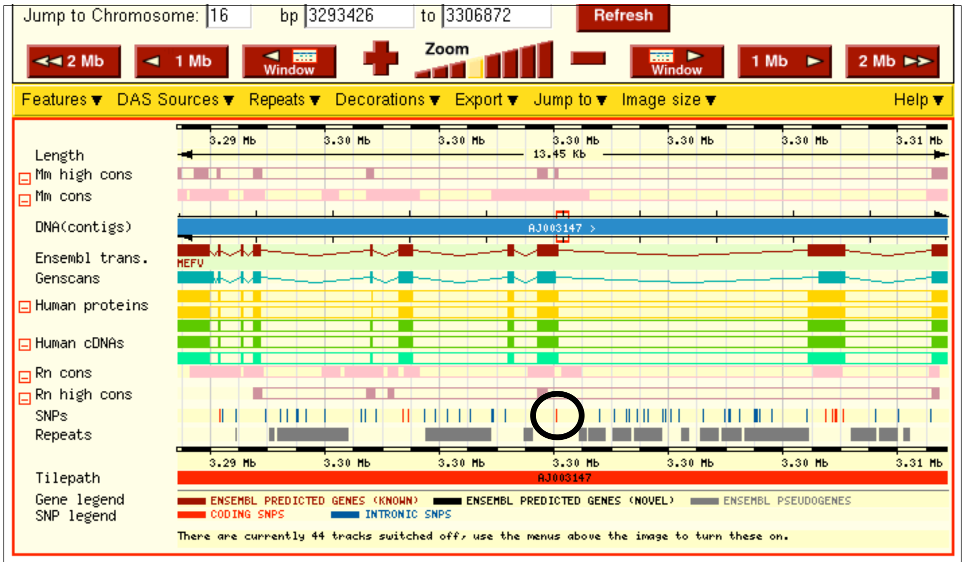
RefSeq Summary: MEFV was identified as the gene that when mutated causes Mediterranean fever, a hereditary periodic fever syndrome. MEFV is expressed in granulocytes and myeloid bone marrow precursors.

Locus Type: gene with protein product, function known or



UCSC Genome Browser on Human July 2003 Freeze

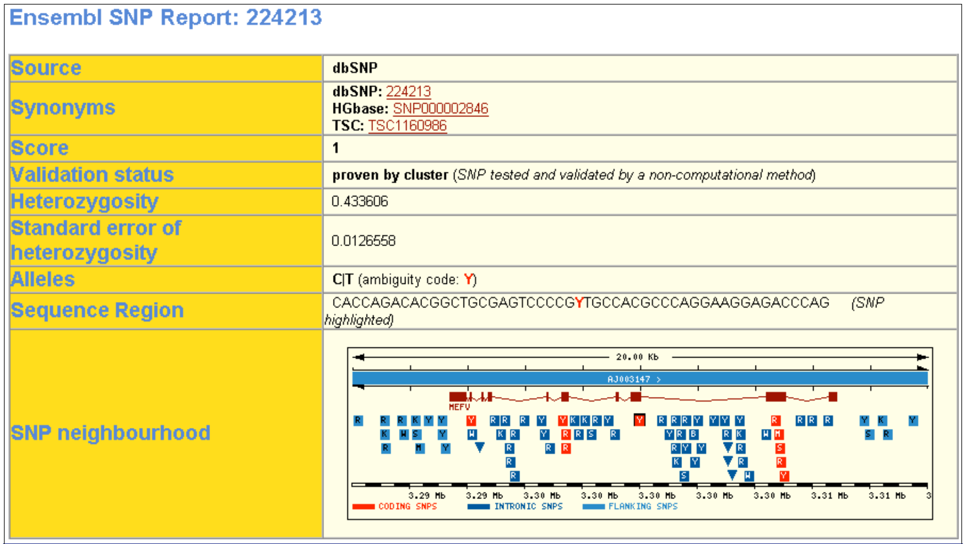
position chr16:3,292,312-3,306,912 size 14,601 bp. image width: 610



Jump to Chromosome: 16 bp 3293426 to 3306872

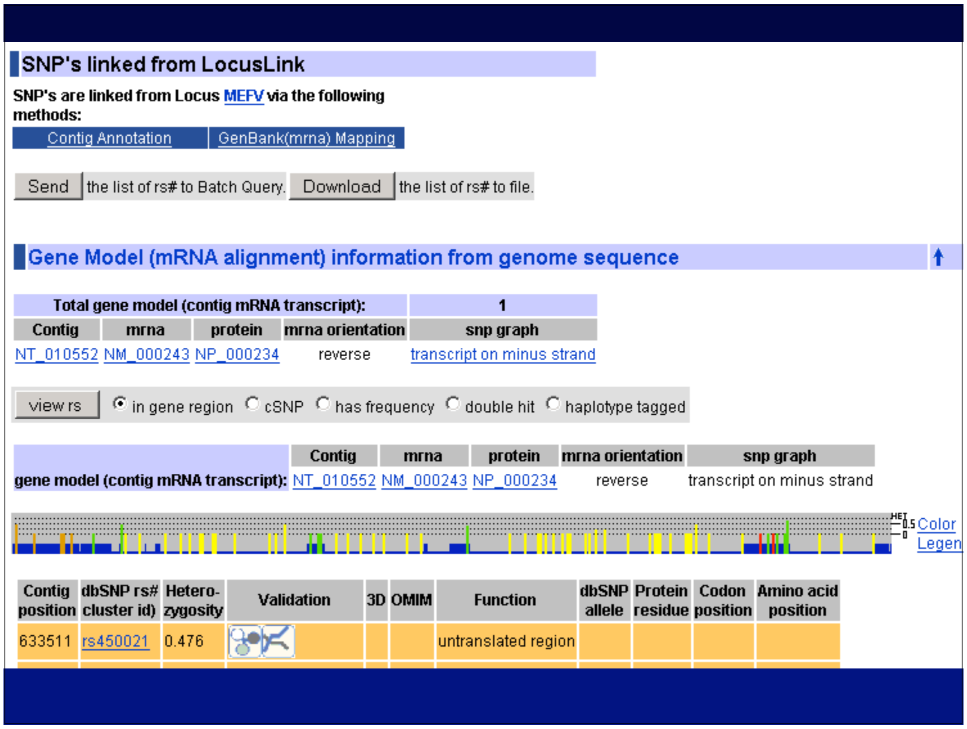
There are currently 44 tracks switched off, use the menus above the image to turn these on.

http://www.ensembl.org/Homo_sapiens



Ensembl SNP Report: 224213

Source	**dbSNP**
Synonyms	**dbSNP:** 224213 **HGbase:** SNP000002846 **TSC:** TSC1160986
Score	**1**
Validation status	**proven by cluster** (*SNP tested and validated by a non-computational method*)
Heterozygosity	0.433606
Standard error of heterozygosity	0.0126558
Alleles	**C|T** (ambiguity code: Y)
Sequence Region	CACCAGACACGGCTGCGAGTCCCCGYTGCCACGCCCAGGAAGGAGACCCAG (*SNP highlighted*)
SNP neighbourhood	



SNP's linked from LocusLink

SNP's are linked from Locus MEFV via the following methods:

Contig Annotation
GenBank(mrna) Mapping
Send
the list of rs# to Batch Query.
Download
the list of rs# to file.

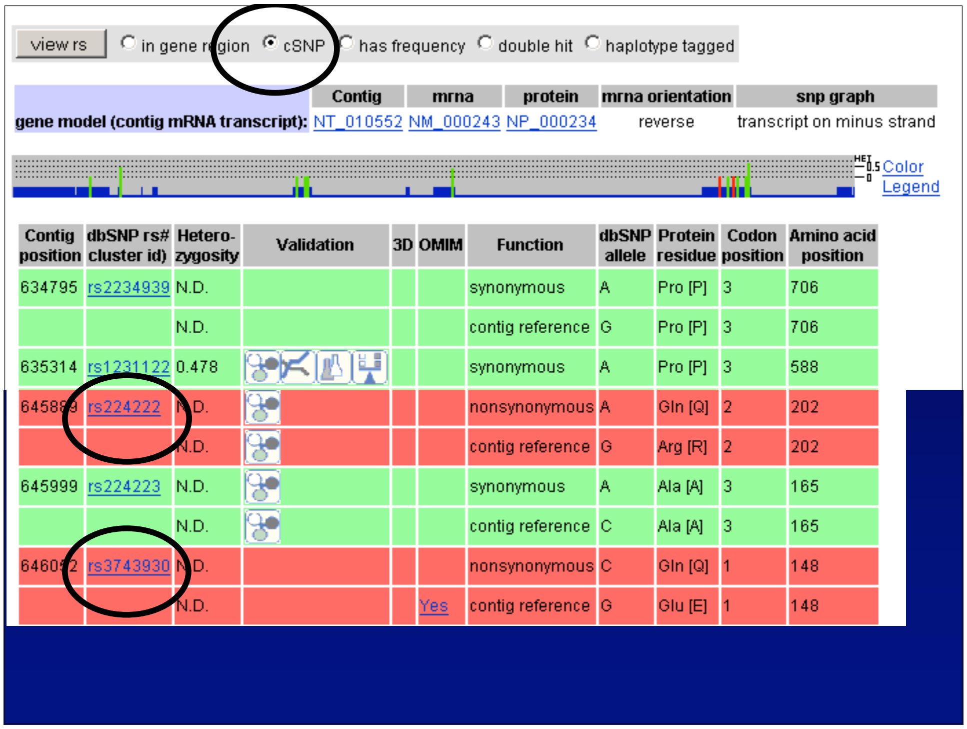
Gene Model (mRNA alignment) information from genome sequence

Total gene model (contig mRNA transcript):				1
Contig	mrna	protein	mrna orientation	snp graph
NT_010552	NM_000243	NP_000234	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_010552	NM_000243	NP_000234	reverse	transcript on minus strand

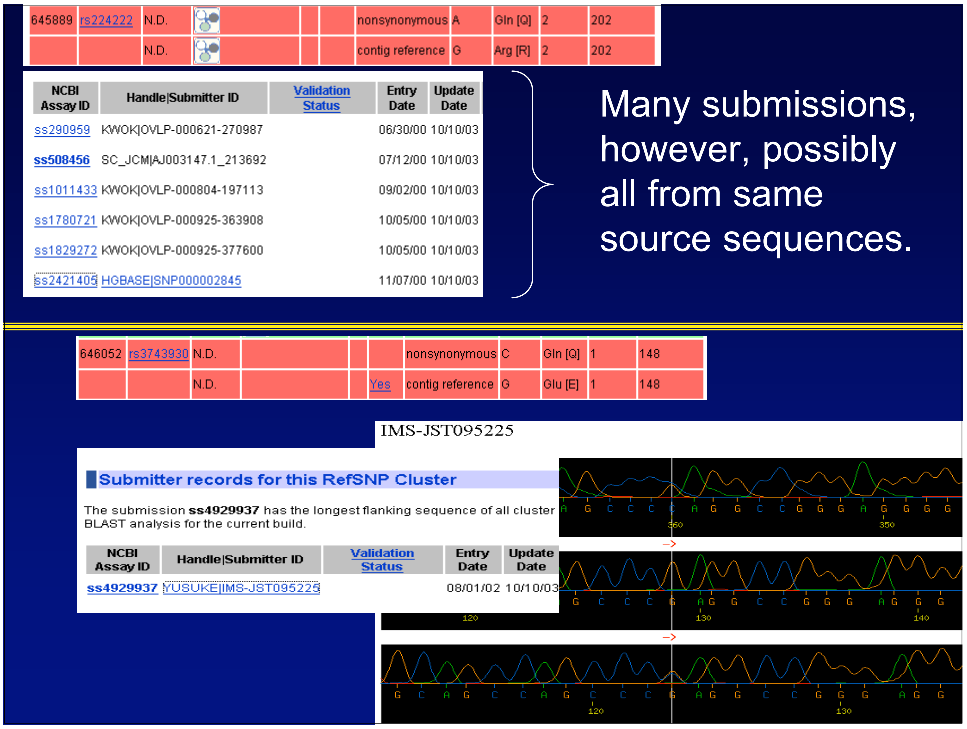
Contig position	dbSNP rs# cluster id)	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
633511	rs450021	0.476	● ●			untranslated region				



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_010552	NM_000243	NP_000234	reverse	transcript on minus strand

Contig position	dbSNP rs# cluster id)	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
634795	rs2234939	N.D.				synonymous	A	Pro [P]	3	706
		N.D.				contig reference	G	Pro [P]	3	706
635314	rs1231122	0.478	● ● ● ●			synonymous	A	Pro [P]	3	588
645889	rs224222	N.D.	●			nonsynonymous	A	Gln [Q]	2	202
		N.D.	●			contig reference	G	Arg [R]	2	202
645999	rs224223	N.D.	●			synonymous	A	Ala [A]	3	165
		N.D.	●			contig reference	C	Ala [A]	3	165
6460[illegible]2	rs3743930	N.D.				nonsynonymous	C	Gln [Q]	1	148
		N.D.			Yes	contig reference	G	Glu [E]	1	148



645889	rs224222	N.D.					nonsynonymous	A	Gln [Q]	2	202
		N.D.					contig reference	G	Arg [R]	2	202

NCBI Assay ID	Handle|Submitter ID	Validation Status	Entry Date	Update Date
ss290959	KWOK|OVLP-000621-270987		06/30/00	10/10/03
ss508456	SC_JCM|AJ003147.1_213692		07/12/00	10/10/03
ss1011433	KWOK|OVLP-000804-197113		09/02/00	10/10/03
ss1780721	KWOK|OVLP-000925-363908		10/05/00	10/10/03
ss1829272	KWOK|OVLP-000925-377600		10/05/00	10/10/03
ss2421405	HGBASE|SNP000002845		11/07/00	10/10/03

Many submissions, however, possibly all from same source sequences.

646052	rs3743930	N.D.				nonsynonymous	C	Gln [Q]	1	148
		N.D.			Yes	contig reference	G	Glu [E]	1	148

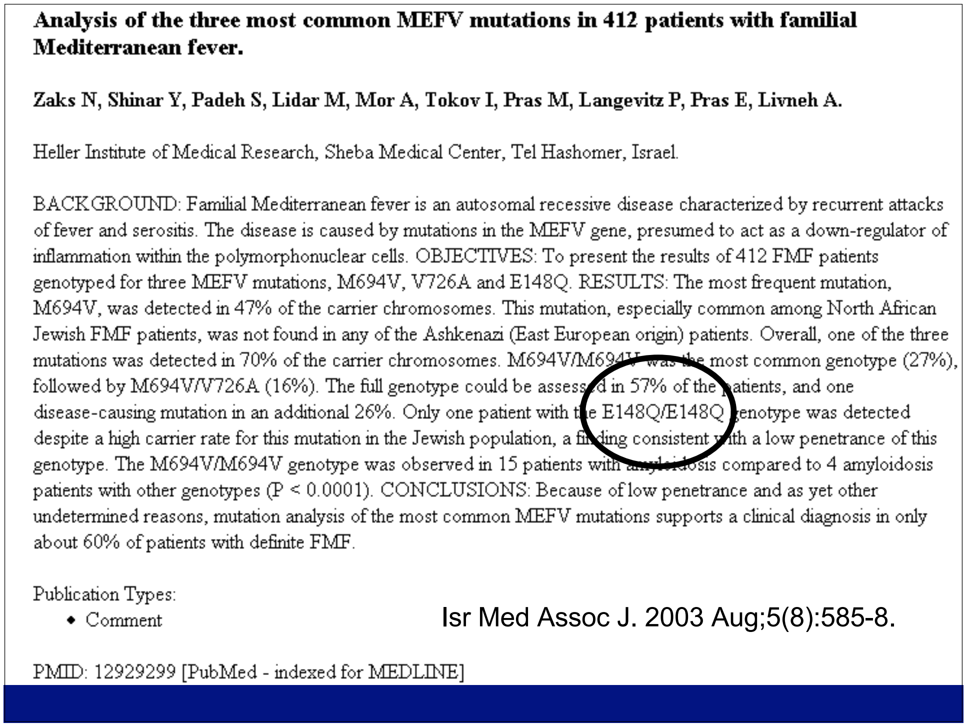
IMS-JST095225

Submitter records for this RefSNP Cluster

The submission **ss4929937** has the longest flanking sequence of all cluster BLAST analysis for the current build.

NCBI Assay ID	Handle|Submitter ID	Validation Status	Entry Date	Update Date
ss4929937	YUSUKE|IMS-JST095225		08/01/02	10/10/03

A G C C C C A G G C C G G G A G G G G
360
350
G C C C G A G G C C G G G A G G G
120
130
140
G C A G C C A G C C C G A G G C C G G G A G G
120
130



Analysis of the three most common MEFV mutations in 412 patients with familial Mediterranean fever.

Zaks N, Shinar Y, Padeh S, Lidar M, Mor A, Tokov I, Pras M, Langevitz P, Pras E, Livneh A.

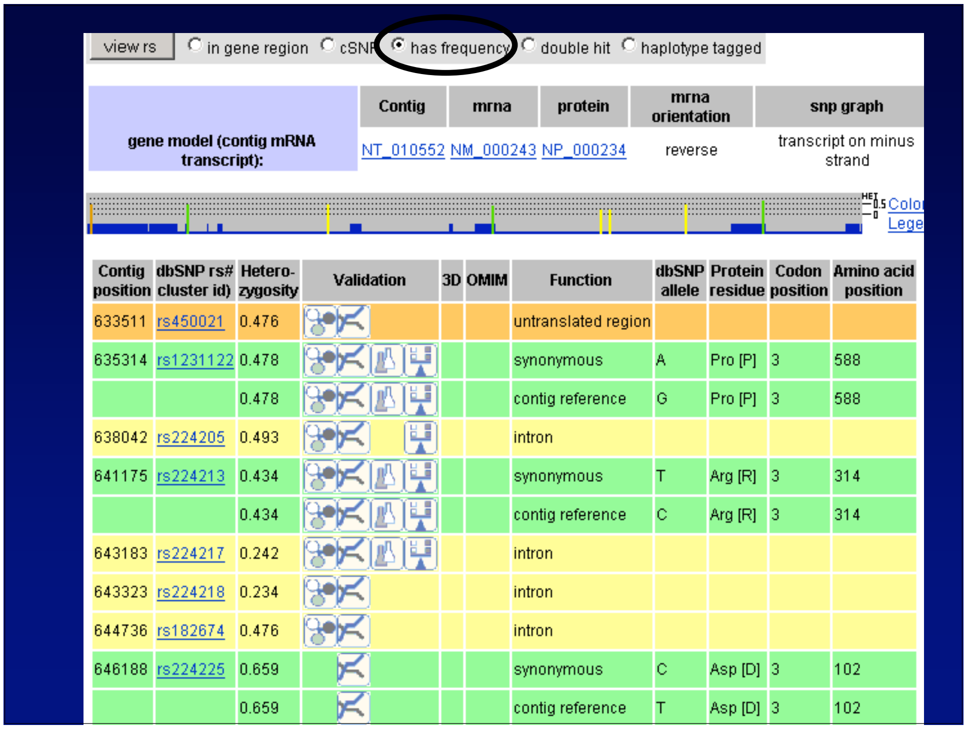
Heller Institute of Medical Research, Sheba Medical Center, Tel Hashomer, Israel.

BACKGROUND: Familial Mediterranean fever is an autosomal recessive disease characterized by recurrent attacks of fever and serositis. The disease is caused by mutations in the MEFV gene, presumed to act as a down-regulator of inflammation within the polymorphonuclear cells. OBJECTIVES: To present the results of 412 FMF patients genotyped for three MEFV mutations, M694V, V726A and E148Q. RESULTS: The most frequent mutation, M694V, was detected in 47% of the carrier chromosomes. This mutation, especially common among North African Jewish FMF patients, was not found in any of the Ashkenazi (East European origin) patients. Overall, one of the three mutations was detected in 70% of the carrier chromosomes. M694V/M694V was the most common genotype (27%), followed by M694V/V726A (16%). The full genotype could be assessed in 57% of the patients, and one disease-causing mutation in an additional 26%. Only one patient with the E148Q/E148Q genotype was detected despite a high carrier rate for this mutation in the Jewish population, a finding consistent with a low penetrance of this genotype. The M694V/M694V genotype was observed in 15 patients with amyloidosis compared to 4 amyloidosis patients with other genotypes (P < 0.0001). CONCLUSIONS: Because of low penetrance and as yet other undetermined reasons, mutation analysis of the most common MEFV mutations supports a clinical diagnosis in only about 60% of patients with definite FMF.

Publication Types:
- Comment

Isr Med Assoc J. 2003 Aug;5(8):585-8.

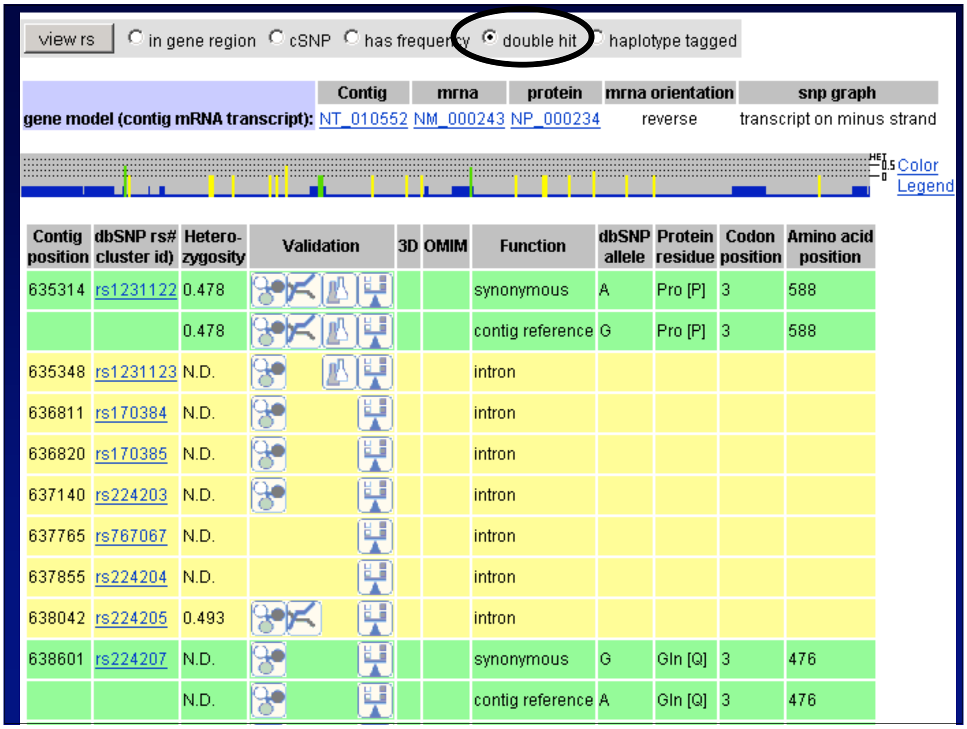
PMID: 12929299 [PubMed - indexed for MEDLINE]



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

gene model (contig mRNA transcript):	Contig	mrna	protein	mrna orientation	snp graph
	NT_010552	NM_000243	NP_000234	reverse	transcript on minus strand

Contig position	dbSNP rs# cluster id)	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
633511	rs450021	0.476				untranslated region				
635314	rs1231122	0.478				synonymous	A	Pro [P]	3	588
		0.478				contig reference	G	Pro [P]	3	588
638042	rs224205	0.493				intron				
641175	rs224213	0.434				synonymous	T	Arg [R]	3	314
		0.434				contig reference	C	Arg [R]	3	314
643183	rs224217	0.242				intron				
643323	rs224218	0.234				intron				
644736	rs182674	0.476				intron				
646188	rs224225	0.659				synonymous	C	Asp [D]	3	102
		0.659				contig reference	T	Asp [D]	3	102

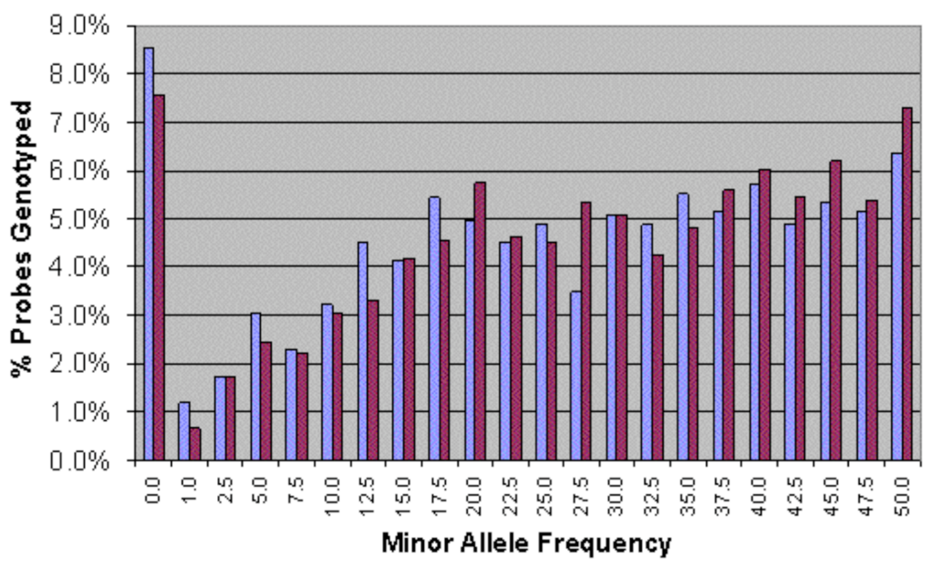


view rs ○ in gene region ○ cSNP ○ has frequency ● double hit ○ haplotype tagged

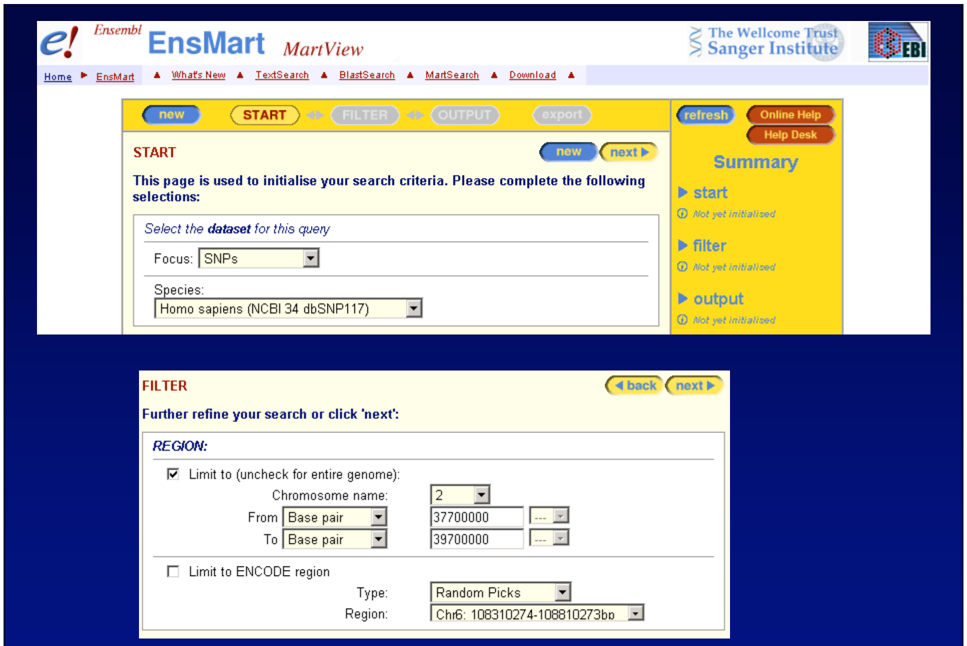
gene model (contig mRNA transcript):	Contig	mrna	protein	mrna orientation	snp graph
	NT_010552	NM_000243	NP_000234	reverse	transcript on minus strand

Contig position	dbSNP rs# cluster id)	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
635314	rs1231122	0.478				synonymous	A	Pro [P]	3	588
		0.478				contig reference	G	Pro [P]	3	588
635348	rs1231123	N.D.				intron				
636811	rs170384	N.D.				intron				
636820	rs170385	N.D.				intron				
637140	rs224203	N.D.				intron				
637765	rs767067	N.D.				intron				
637855	rs224204	N.D.				intron				
638042	rs224205	0.493				intron				
638601	rs224207	N.D.				synonymous	G	Gln [Q]	3	476
		N.D.				contig reference	A	Gln [Q]	3	476

Double hit SNP minor allele frequency characteristics



Credit: Dr. Paul Hardenbol, Parallele Bioscience



START

This page is used to initialise your search criteria. Please complete the following selections:

*Select the **dataset** for this query*

Focus: SNPs

Species: Homo sapiens (NCBI 34 dbSNP117)

Summary

- start
  Not yet initialised
- filter
  Not yet initialised
- output
  Not yet initialised

FILTER

Further refine your search or click 'next':

REGION:

☑ Limit to (uncheck for entire genome):

Chromosome name: 2

From Base pair 37700000 ---

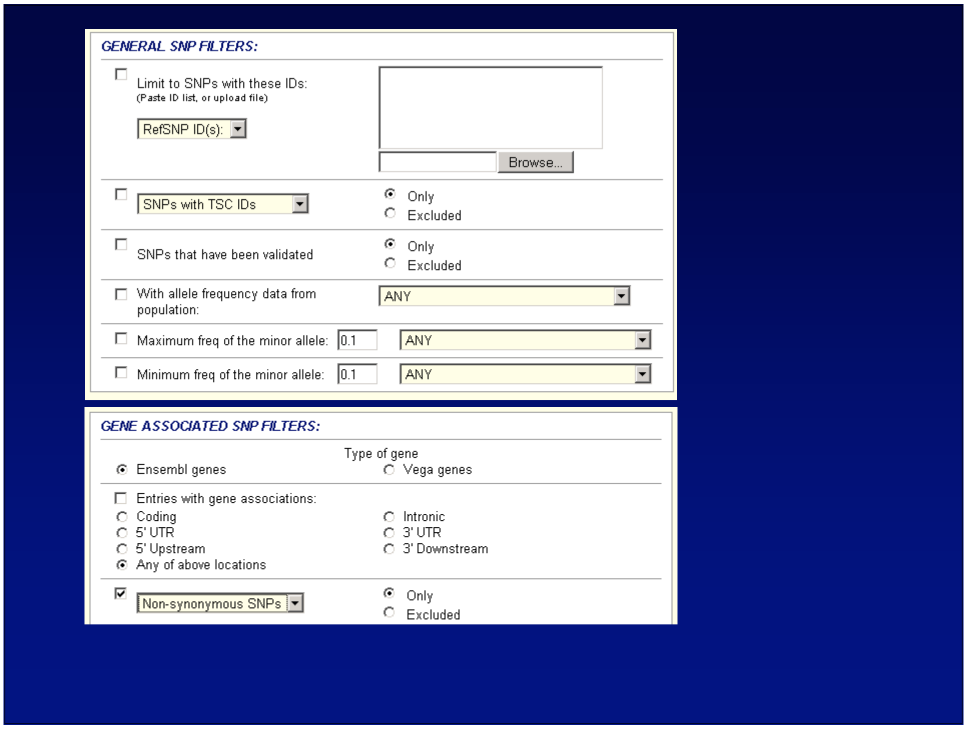
To Base pair 39700000 ---

☐ Limit to ENCODE region

Type: Random Picks

Region: Chr6: 108310274-108810273bp

http://www.ensembl.org/Multi/martview?species=Homo_sapiens



GENERAL SNP FILTERS:

□ Limit to SNPs with these IDs:
(Paste ID list, or upload file)
RefSNP ID(s):

□ SNPs with TSC IDs — ● Only ○ Excluded

□ SNPs that have been validated — ● Only ○ Excluded

□ With allele frequency data from population: ANY

□ Maximum freq of the minor allele: 0.1 ANY

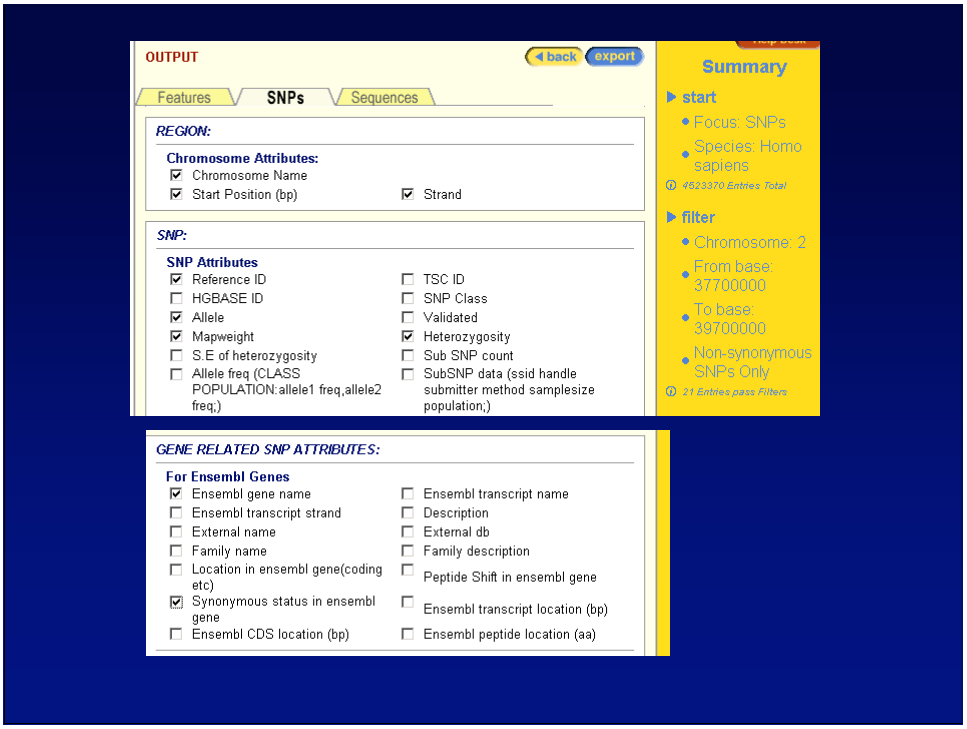
□ Minimum freq of the minor allele: 0.1 ANY

GENE ASSOCIATED SNP FILTERS:

Type of gene
● Ensembl genes ○ Vega genes

□ Entries with gene associations:
○ Coding ○ Intronic
○ 5' UTR ○ 3' UTR
○ 5' Upstream ○ 3' Downstream
● Any of above locations

☑ Non-synonymous SNPs — ● Only ○ Excluded



OUTPUT

REGION:

Chromosome Attributes:
☑ Chromosome Name
☑ Start Position (bp) ☑ Strand

SNP:

SNP Attributes
☑ Reference ID ☐ TSC ID
☐ HGBASE ID ☐ SNP Class
☑ Allele ☐ Validated
☑ Mapweight ☑ Heterozygosity
☐ S.E of heterozygosity ☐ Sub SNP count
☐ Allele freq (CLASS POPULATION:allele1 freq,allele2 freq;) ☐ SubSNP data (ssid handle submitter method samplesize population;)

Summary
▶ start
- Focus: SNPs
- Species: Homo sapiens

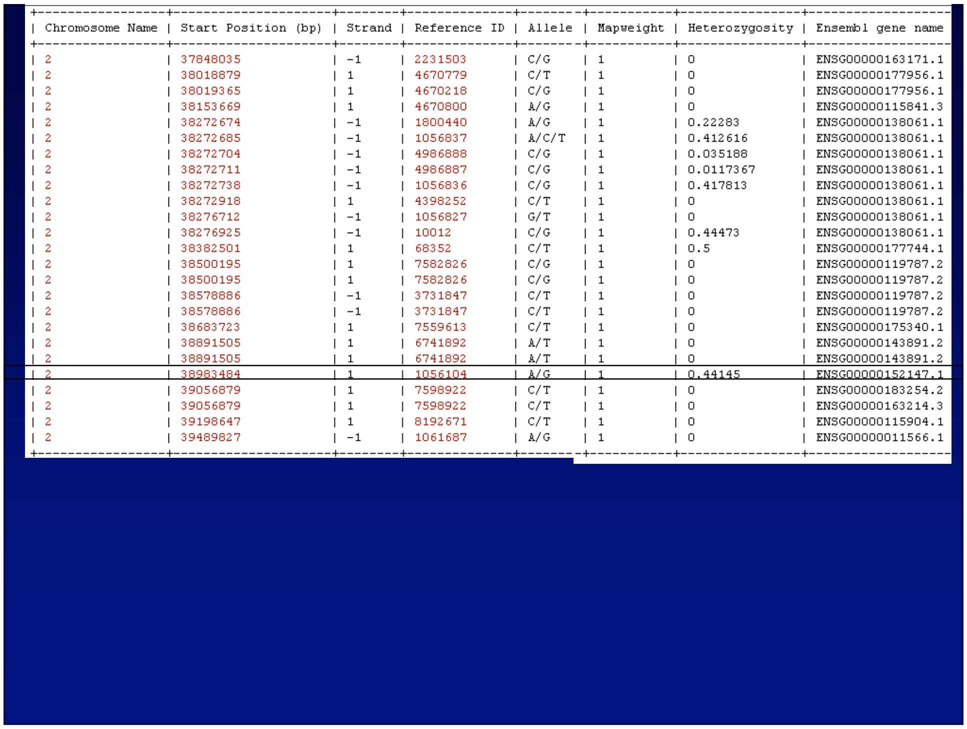
ⓘ 4523370 Entries Total

▶ filter
- Chromosome: 2
- From base: 37700000
- To base: 39700000
- Non-synonymous SNPs Only

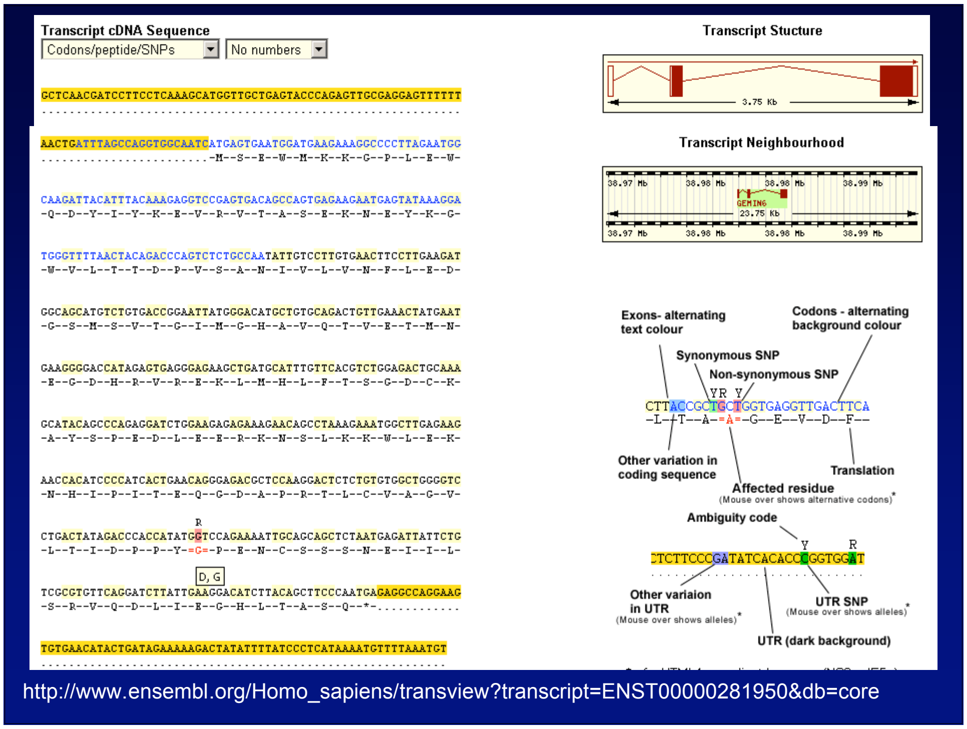
ⓘ 21 Entries pass Filters

GENE RELATED SNP ATTRIBUTES:

For Ensembl Genes
☑ Ensembl gene name ☐ Ensembl transcript name
☐ Ensembl transcript strand ☐ Description
☐ External name ☐ External db
☐ Family name ☐ Family description
☐ Location in ensembl gene(coding etc) ☐ Peptide Shift in ensembl gene
☑ Synonymous status in ensembl gene ☐ Ensembl transcript location (bp)
☐ Ensembl CDS location (bp) ☐ Ensembl peptide location (aa)



```
+-----------------+---------------------+--------+----------------+---------+------------+------------------+-------------------
| Chromosome Name | Start Position (bp) | Strand | Reference ID   | Allele  | Mapweight  | Heterozygosity   | Ensembl gene name
+-----------------+---------------------+--------+----------------+---------+------------+------------------+-------------------
| 2               | 37848035            | -1     | 2231503        | C/G     | 1          | 0                | ENSG00000163171.1
| 2               | 38018879            | 1      | 4670779        | C/T     | 1          | 0                | ENSG00000177956.1
| 2               | 38019365            | 1      | 4670218        | C/G     | 1          | 0                | ENSG00000177956.1
| 2               | 38153669            | 1      | 4670800        | A/G     | 1          | 0                | ENSG00000115841.3
| 2               | 38272674            | -1     | 1800440        | A/G     | 1          | 0.22283          | ENSG00000138061.1
| 2               | 38272685            | -1     | 1056837        | A/C/T   | 1          | 0.412616         | ENSG00000138061.1
| 2               | 38272704            | -1     | 4986888        | C/G     | 1          | 0.035188         | ENSG00000138061.1
| 2               | 38272711            | -1     | 4986887        | C/G     | 1          | 0.0117367        | ENSG00000138061.1
| 2               | 38272738            | -1     | 1056836        | C/G     | 1          | 0.417813         | ENSG00000138061.1
| 2               | 38272918            | 1      | 4398252        | C/T     | 1          | 0                | ENSG00000138061.1
| 2               | 38276712            | -1     | 1056827        | G/T     | 1          | 0                | ENSG00000138061.1
| 2               | 38276925            | -1     | 10012          | C/G     | 1          | 0.44473          | ENSG00000138061.1
| 2               | 38382501            | 1      | 68352          | C/T     | 1          | 0.5              | ENSG00000177744.1
| 2               | 38500195            | 1      | 7582826        | C/G     | 1          | 0                | ENSG00000119787.2
| 2               | 38500195            | 1      | 7582826        | C/G     | 1          | 0                | ENSG00000119787.2
| 2               | 38578886            | -1     | 3731847        | C/T     | 1          | 0                | ENSG00000119787.2
| 2               | 38578886            | -1     | 3731847        | C/T     | 1          | 0                | ENSG00000119787.2
| 2               | 38683723            | 1      | 7559613        | C/T     | 1          | 0                | ENSG00000175340.1
| 2               | 38891505            | 1      | 6741892        | A/T     | 1          | 0                | ENSG00000143891.2
| 2               | 38891505            | 1      | 6741892        | A/T     | 1          | 0                | ENSG00000143891.2
| 2               | 38983484            | 1      | 1056104        | A/G     | 1          | 0.44145          | ENSG00000152147.1
| 2               | 39056879            | 1      | 7598922        | C/T     | 1          | 0                | ENSG00000183254.2
| 2               | 39056879            | 1      | 7598922        | C/T     | 1          | 0                | ENSG00000163214.3
| 2               | 39198647            | 1      | 8192671        | C/T     | 1          | 0                | ENSG00000115904.1
| 2               | 39489827            | -1     | 1061687        | A/G     | 1          | 0                | ENSG00000011566.1
+-----------------+---------------------+--------+----------------+---------+------------+------------------+-------------------
```



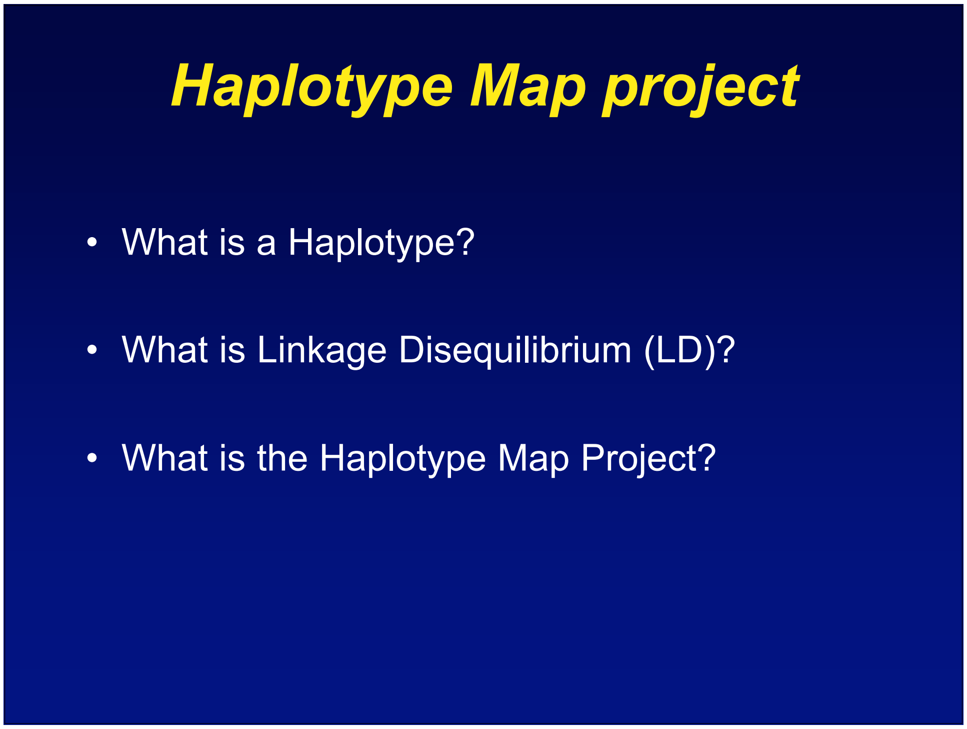
Transcript cDNA Sequence

Codons/peptide/SNPs ▾ No numbers ▾

```
GCTCAACGATCCTTCCTCAAAGCATGGTTGCTGAGTACCCAGAGTTGCGAGGAGTTTTTT
------------------------------------------------------------

AACTGATTTAGCCAGGTGGCAATCATGAGTGAATGGATGAAGAAAGGCCCCTTAGAATGG
........................-M--S--E--W--M--K--K--G--P--L--E--W-

CAAGATTACATTTACAAAGAGGTCCGAGTGACAGCCAGTGAGAAGAATGAGTATAAAGGA
-Q--D--Y--I--Y--K--E--V--R--V--T--A--S--E--K--N--E--Y--K--G-

TGGGTTTTAACTACAGACCCAGTCTCTGCCAATATTGTCCTTGTGAACTTCCTTGAAGAT
-W--V--L--T--T--D--P--V--S--A--N--I--V--L--V--N--F--L--E--D-

GGCAGCATGTCTGTGACCGGAATTATGGGACATGCTGTGCAGACTGTTGAAACTATGAAT
-G--S--M--S--V--T--G--I--M--G--H--A--V--Q--T--V--E--T--M--N-

GAAGGGGACCATAGAGTGAGGGAGAAGCTGATGCATTTGTTCACGTCTGGAGACTGCAAA
-E--G--D--H--R--V--R--E--K--L--M--H--L--F--T--S--G--D--C--K-

GCATACAGCCCAGAGGATCTGGAAGAGAGAAAGAACAGCCTAAAGAAATGGCTTGAGAAG
-A--Y--S--P--E--D--L--E--E--R--K--N--S--L--K--K--W--L--E--K-

AACCACATCCCCATCACTGAACAGGGAGACGCTCCAAGGACTCTCTGTGTGGCTGGGGTC
-N--H--I--P--I--T--E--Q--G--D--A--P--R--T--L--C--V--A--G--V-

                           R
CTGACTATAGACCCACCATATGGTCCAGAAAATTGCAGCAGCTCTAATGAGATTATTCTG
-L--T--I--D--P--P--Y-=G=-P--E--N--C--S--S--S--N--E--I--I--L-

                           D, G
TCGCGTGTTCAGGATCTTATTGAAGGACATCTTACAGCTTCCCAATGAGAGGCCAGGAAG
-S--R--V--Q--D--L--I--E--G--H--L--T--A--S--Q--*-............

TGTGAACATACTGATAGAAAAAGACTATATTTTATCCCTCATAAAATGTTTTAAATGT
----------------------------------------------------------
```

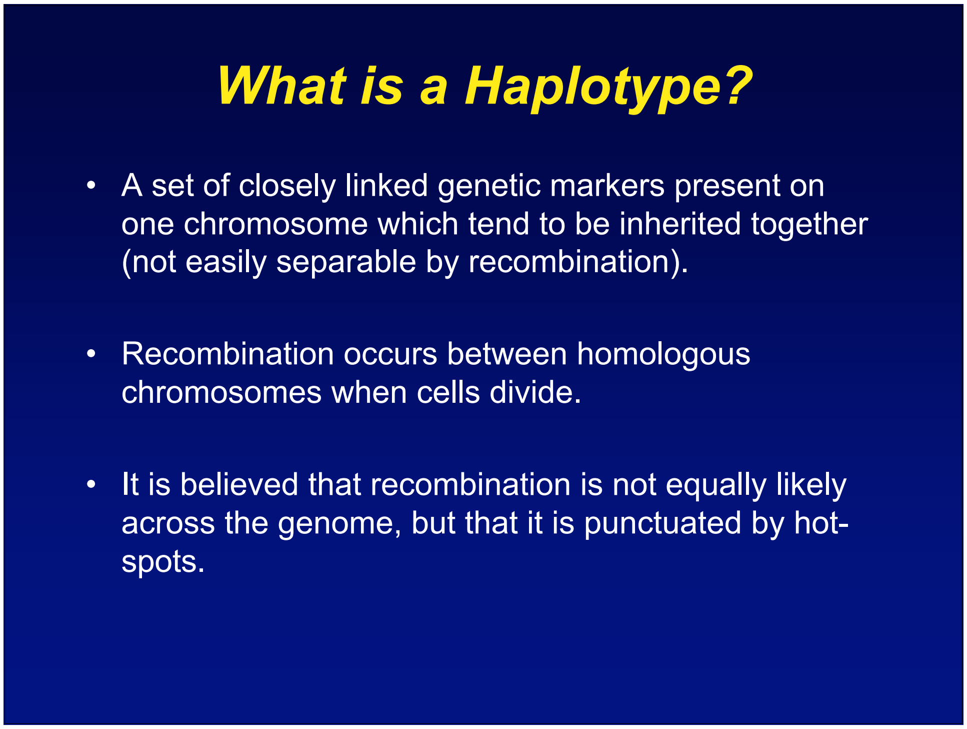
Transcript Stucture
3.75 Kb
Transcript Neighbourhood
38.97 Mb
38.98 Mb
38.98 Mb
38.99 Mb
GEMIN6
23.75 Kb
38.97 Mb
38.98 Mb
38.98 Mb
38.99 Mb
Exons- alternating text colour
Codons - alternating background colour
Synonymous SNP
Non-synonymous SNP
YR Y
CTTACCGCTGCTGGTGAGGTTGACTTCA
-L--T--A-=A=-G--E--V--D--F--
Other variation in coding sequence
Translation
Affected residue
(Mouse over shows alternative codons)*
Ambiguity code
Y R
CTCTTCCCGATATCACACCCGGTGGAT
Other variaion in UTR
(Mouse over shows alleles)*
UTR SNP
(Mouse over shows alleles)*
UTR (dark background)

http://www.ensembl.org/Homo_sapiens/transview?transcript=ENST00000281950&db=core



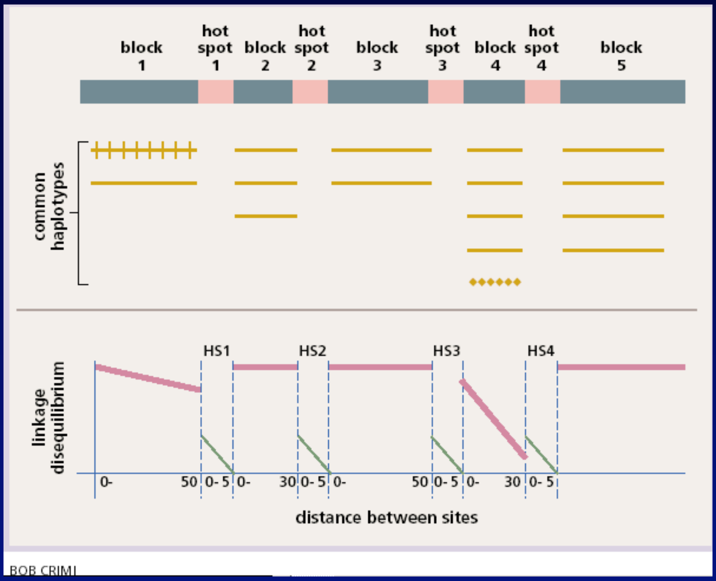
Haplotype Map project

- What is a Haplotype?
- What is Linkage Disequilibrium (LD)?
- What is the Haplotype Map Project?



What is a Haplotype?

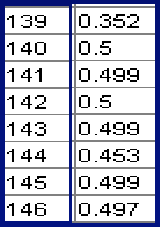
- A set of closely linked genetic markers present on one chromosome which tend to be inherited together (not easily separable by recombination).
- Recombination occurs between homologous chromosomes when cells divide.
- It is believed that recombination is not equally likely across the genome, but that it is punctuated by hot-spots.



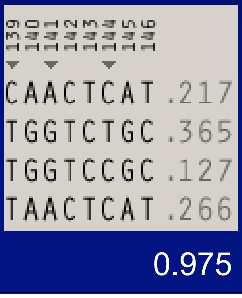
From: Goldstein DB. Islands of linkage disequilibrium. Nat Genet. 2001 Oct;29(2):109-11.

What is Linkage Disequilibrium?

- When the observed frequencies of genetic markers in a population does not agree with haplotype frequencies predicted by multiplying together the frequency of individual genetic markers in each haplotype.



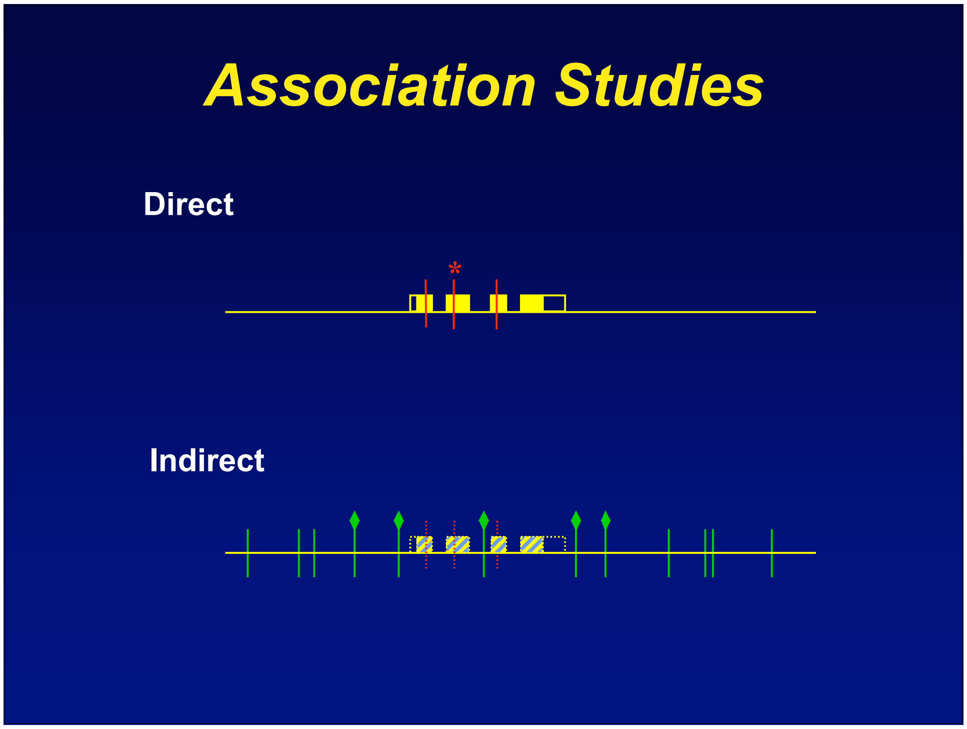
139	0.352
140	0.5
141	0.499
142	0.5
143	0.499
144	0.453
145	0.499
146	0.497



CAACTCAT .217
TGGTCTGC .365
TGGTCCGC .127
TAACTCAT .266

0.352*0.5^7=0.00275
0.648*0.5^7=0.00534
0.648*0.5^7=0.00534
0.648*0.5^7=0.00534

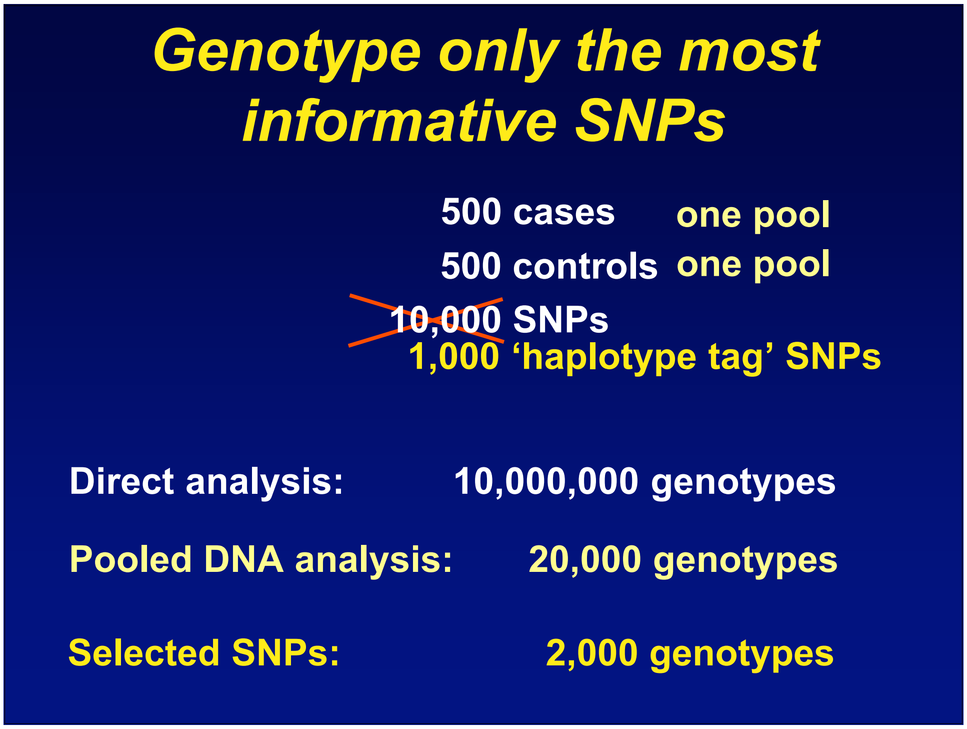
0.975



Association Studies

Direct

Indirect

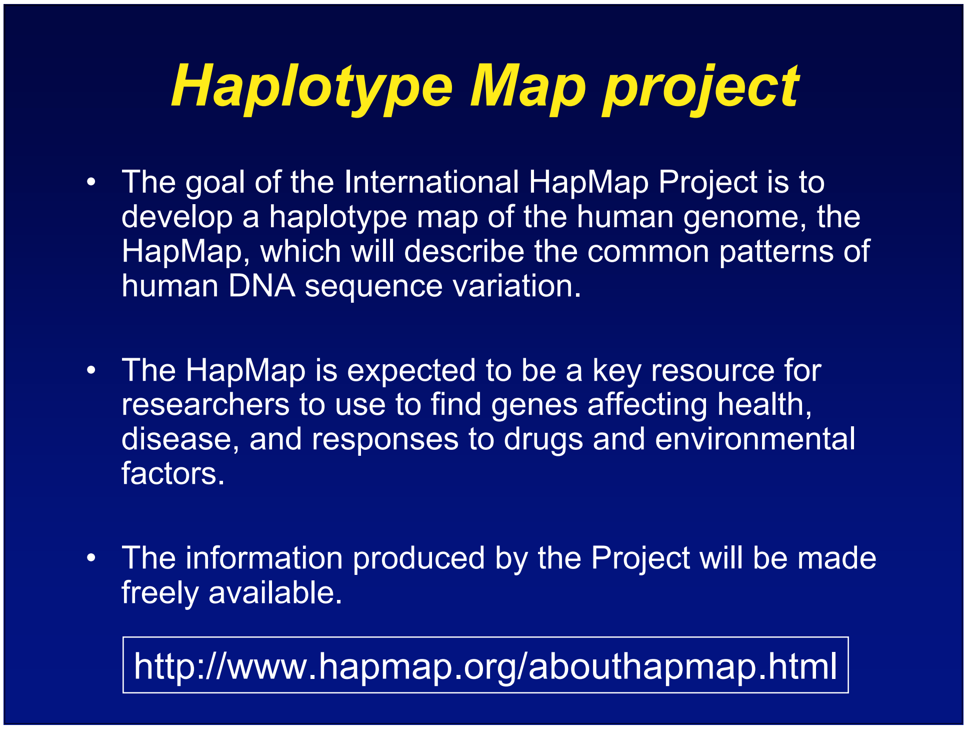


Genotype only the most informative SNPs

500 cases	one pool
500 controls	one pool

~~10,000~~ SNPs
1,000 'haplotype tag' SNPs

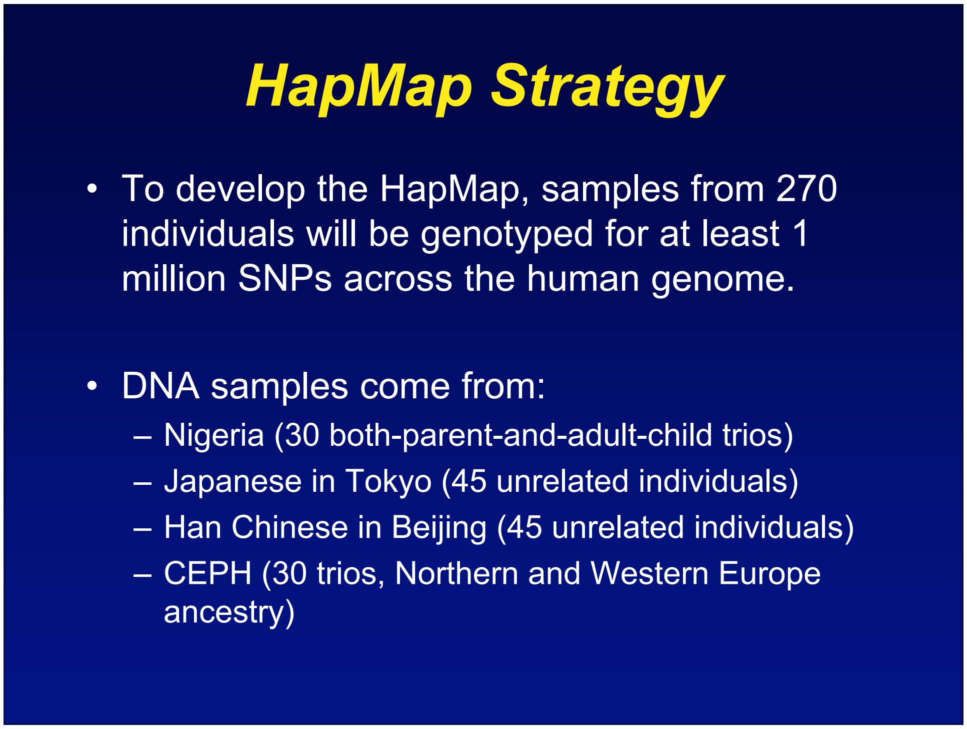
Direct analysis:	10,000,000 genotypes
Pooled DNA analysis:	20,000 genotypes
Selected SNPs:	2,000 genotypes



Haplotype Map project

- The goal of the International HapMap Project is to develop a haplotype map of the human genome, the HapMap, which will describe the common patterns of human DNA sequence variation.
- The HapMap is expected to be a key resource for researchers to use to find genes affecting health, disease, and responses to drugs and environmental factors.
- The information produced by the Project will be made freely available.

http://www.hapmap.org/abouthapmap.html



HapMap Strategy

- To develop the HapMap, samples from 270 individuals will be genotyped for at least 1 million SNPs across the human genome.
- DNA samples come from:
  - Nigeria (30 both-parent-and-adult-child trios)
  - Japanese in Tokyo (45 unrelated individuals)
  - Han Chinese in Beijing (45 unrelated individuals)
  - CEPH (30 trios, Northern and Western Europe ancestry)

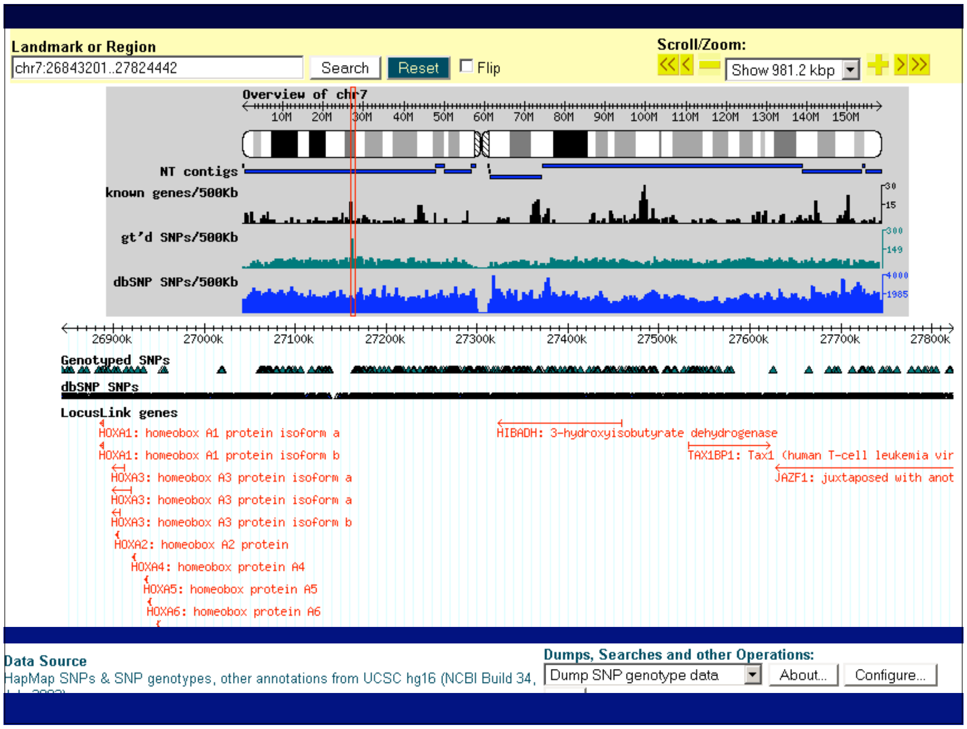
Data Analysis

- Genotyped SNPs are analyzed for association using standard measures, such as *D'* and r^2.
  - Deviation from equilibrium between two markers is denoted by *D*.
  - When normalized it is called *D'* and has a range from -1 to +1.
  - r^2 uses a different normalization method and ranges between 0 and 1.
  - See URL below for a good description of these measures.

http://www.ucl.ac.uk/~ucbhdjm/courses/b242/2+Gene/2+Gene.html

Current status of HapMap

- November 1st, 2003: First major public data release!
  - Over 13 million genotypes from 145,554 SNPs
  - Associated allele frequency and assay data have been released for public download
- Here's an example of generating haplotype information from the current data release…



Landmark or Region: chr7:26843201..27824442

Scroll/Zoom: Show 981.2 kbp

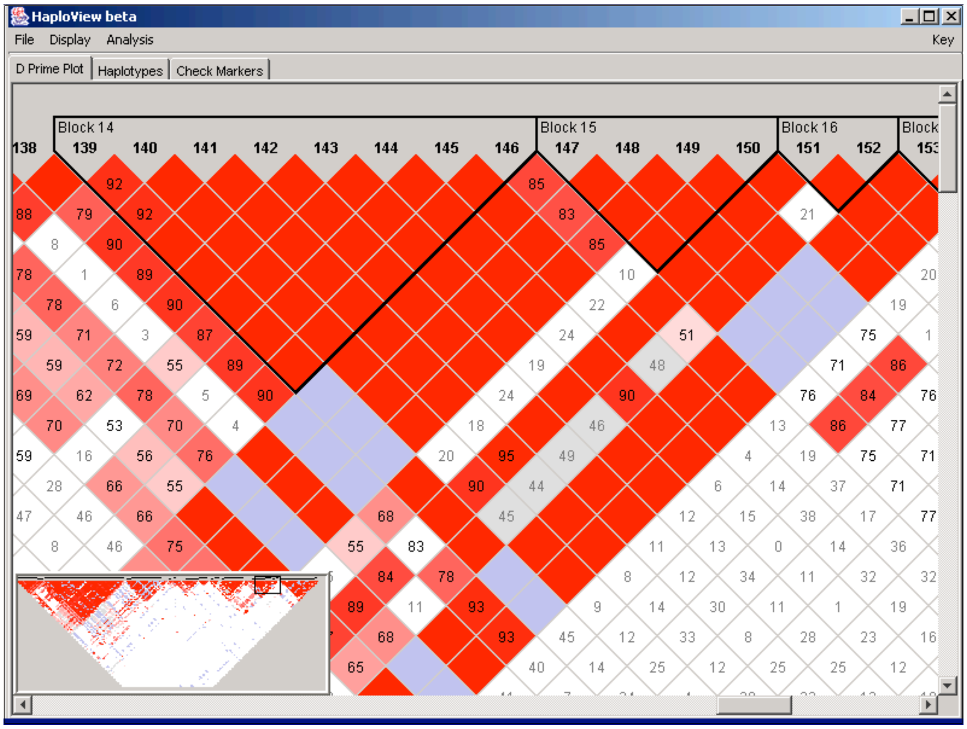
Data Source
HapMap SNPs & SNP genotypes, other annotations from UCSC hg16 (NCBI Build 34,

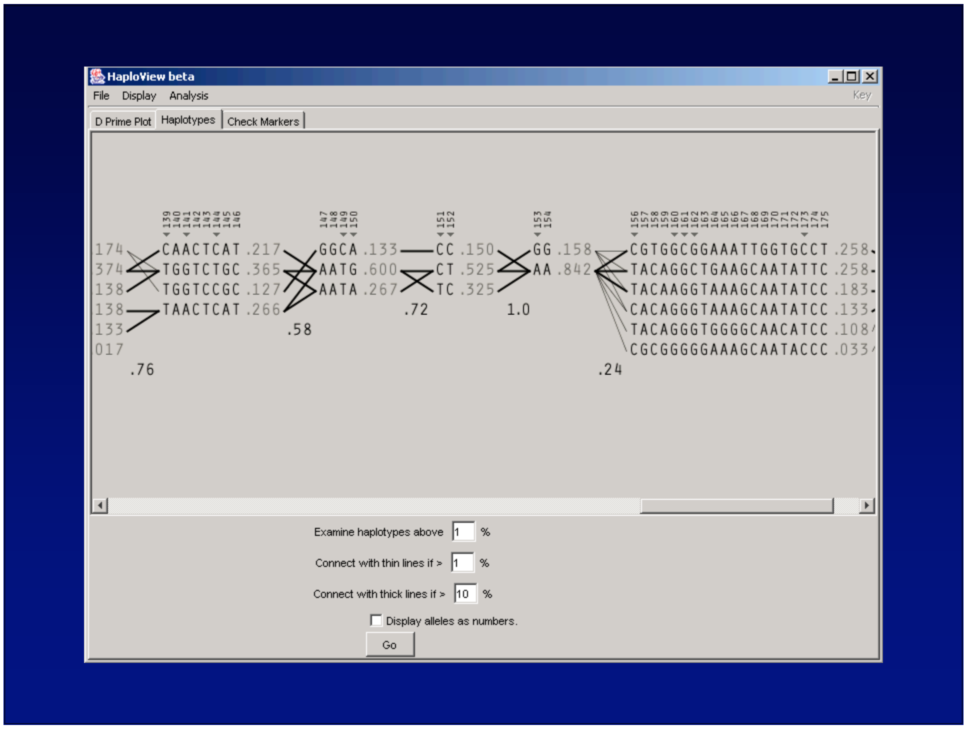
Dumps, Searches and other Operations: Dump SNP genotype data

Search
Reset
□ Flip
About...
Configure...
Overview of chr7
10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M
NT contigs
known genes/500Kb
gt'd SNPs/500Kb
dbSNP SNPs/500Kb
30
15
300
149
4000
1985
26900k 27000k 27100k 27200k 27300k 27400k 27500k 27600k 27700k 27800k
Genotyped SNPs
dbSNP SNPs
LocusLink genes
HOXA1: homeobox A1 protein isoform a
HOXA1: homeobox A1 protein isoform b
HOXA3: homeobox A3 protein isoform a
HOXA3: homeobox A3 protein isoform a
HOXA3: homeobox A3 protein isoform b
HOXA2: homeobox A2 protein
HOXA4: homeobox protein A4
HOXA5: homeobox protein A5
HOXA6: homeobox protein A6
HIBADH: 3-hydroxyisobutyrate dehydrogenase
TAX1BP1: Tax1 (human T-cell leukemia vir
JAZF1: juxtaposed with anot

HaploView

- Developed and maintained by Jeffrey Barrett in Mark Daly's lab at The Broad Institute.
- Haploview currently allows users to:
  - – examine block structures
  - – generate haplotypes in these blocks
  - – run association tests
  - – and save the data in a number of formats.





```
174   CAACTCAT .217   GGCA .133   CC .150   GG .158   CGTGGCGGAAATTGGTGCCT .258
374   TGGTCTGC .365   AATG .600   CT .525   AA .842   TACAGGCTGAAGCAATATTC .258
138   TGGTCCGC .127   AATA .267   TC .325             TACAAGGTAAAGCAATATCC .183
138   TAACTCAT .266                                   CACAGGGTAAAGCAATATCC .133
133                                                   TACAGGGTGGGGCAACATCC .108
017                                                   CGCGGGGGAAAGCAATACCC .033
                  .58          .72        1.0
    .76                                          .24
```

Examine haplotypes above 1 %

Connect with thin lines if > 1 %

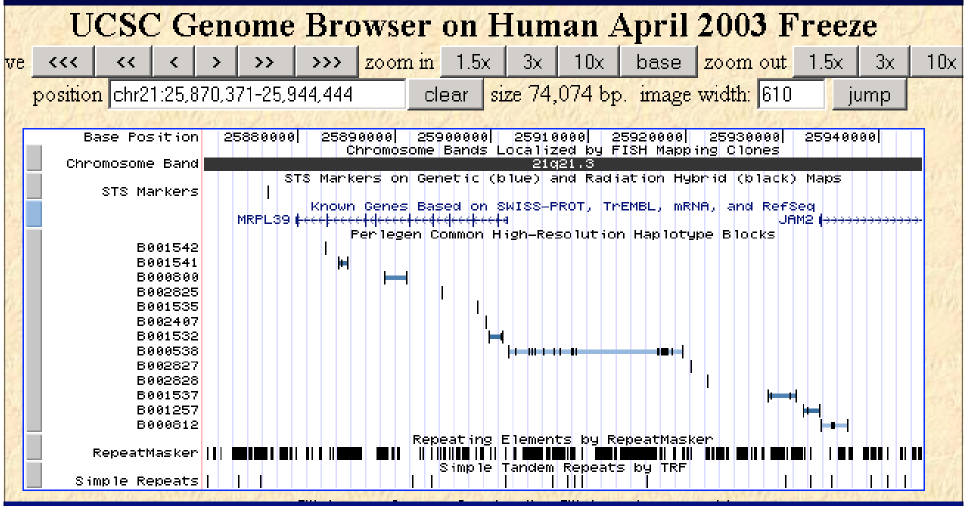
Connect with thick lines if > 10 %

☐ Display alleles as numbers.

Go

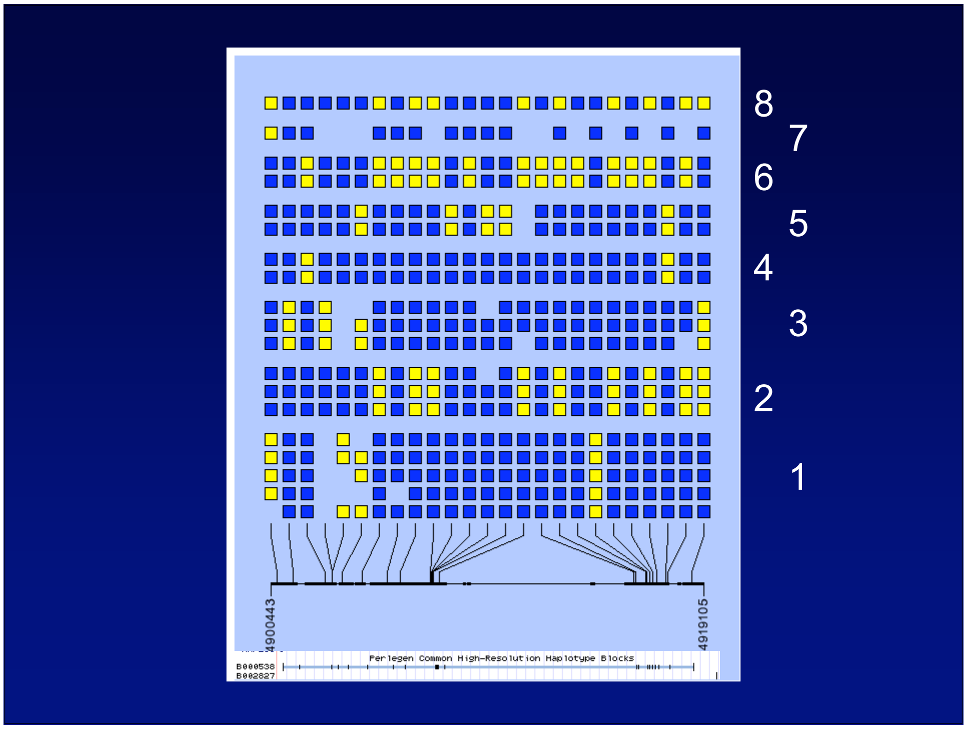
Perlegen's haplotype map

- Used chip based resequencing to discover SNPs across the genome.
- Applied this technology to single haploid copies of each chromosome from a number of different individuals.
- From this information, haplotypes can be deduced directly from the data.
- They have released data from chromosome 21 for public use.



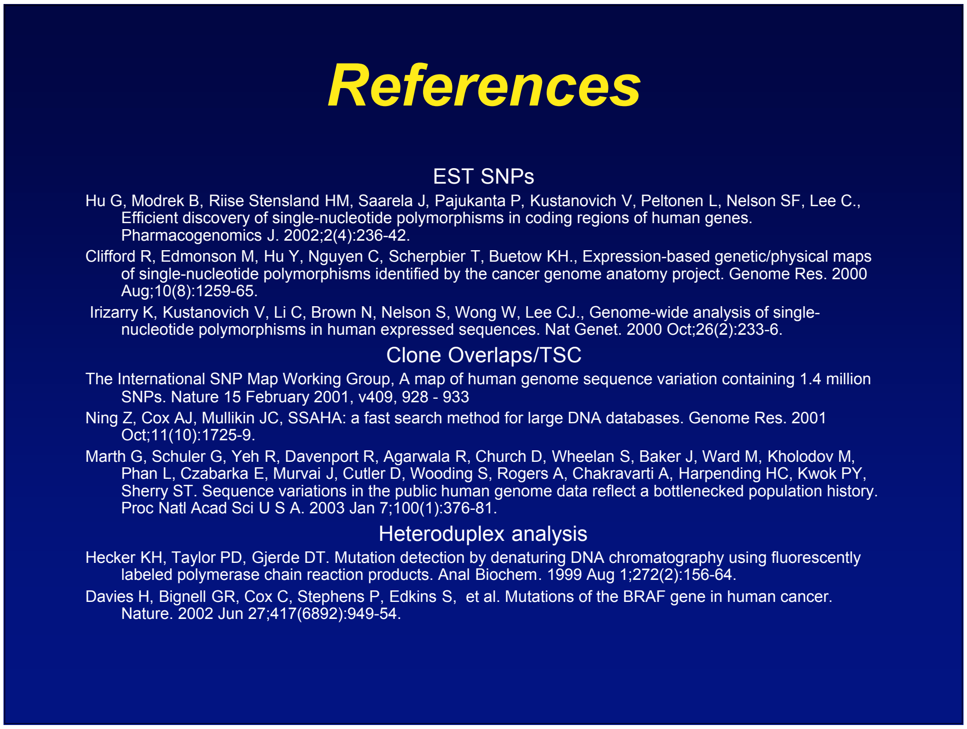
UCSC Genome Browser on Human April 2003 Freeze

position chr21:25,870,371-25,944,444 size 74,074 bp. image width: 610



Concluding remarks

- Along with the emergence of the human genome, we also have a growing database of variations that are critical to the overall value of the human genome sequence.
- These variations are what make us all (phenotypically) different, and impart different levels of resistance and susceptibility to disease.
- The collection of human sequence variation information will continue to evolve rapidly.



References

EST SNPs

Hu G, Modrek B, Riise Stensland HM, Saarela J, Pajukanta P, Kustanovich V, Peltonen L, Nelson SF, Lee C., Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes. Pharmacogenomics J. 2002;2(4):236-42.

Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, Buetow KH., Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. Genome Res. 2000 Aug;10(8):1259-65.

Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, Lee CJ., Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. Nat Genet. 2000 Oct;26(2):233-6.

Clone Overlaps/TSC

The International SNP Map Working Group, A map of human genome sequence variation containing 1.4 million SNPs. Nature 15 February 2001, v409, 928 - 933

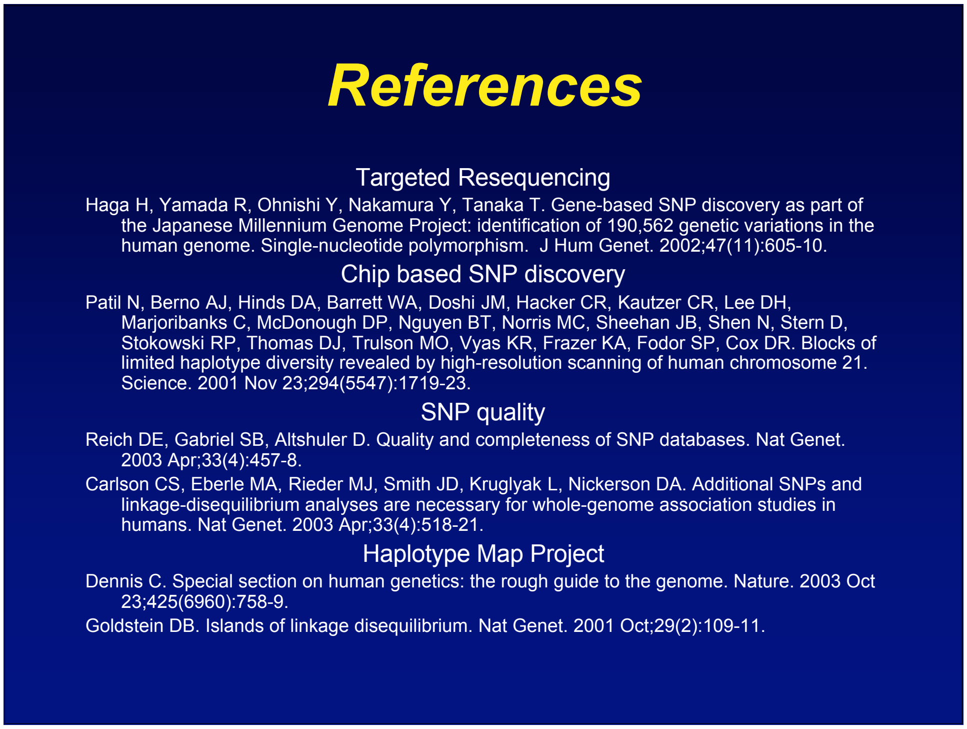
Ning Z, Cox AJ, Mullikin JC, SSAHA: a fast search method for large DNA databases. Genome Res. 2001 Oct;11(10):1725-9.

Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, Church D, Wheelan S, Baker J, Ward M, Kholodov M, Phan L, Czabarka E, Murvai J, Cutler D, Wooding S, Rogers A, Chakravarti A, Harpending HC, Kwok PY, Sherry ST. Sequence variations in the public human genome data reflect a bottlenecked population history. Proc Natl Acad Sci U S A. 2003 Jan 7;100(1):376-81.

Heteroduplex analysis

Hecker KH, Taylor PD, Gjerde DT. Mutation detection by denaturing DNA chromatography using fluorescently labeled polymerase chain reaction products. Anal Biochem. 1999 Aug 1;272(2):156-64.

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, et al. Mutations of the BRAF gene in human cancer. Nature. 2002 Jun 27;417(6892):949-54.



References

Targeted Resequencing

Haga H, Yamada R, Ohnishi Y, Nakamura Y, Tanaka T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. J Hum Genet. 2002;47(11):605-10.

Chip based SNP discovery

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science. 2001 Nov 23;294(5547):1719-23.

SNP quality

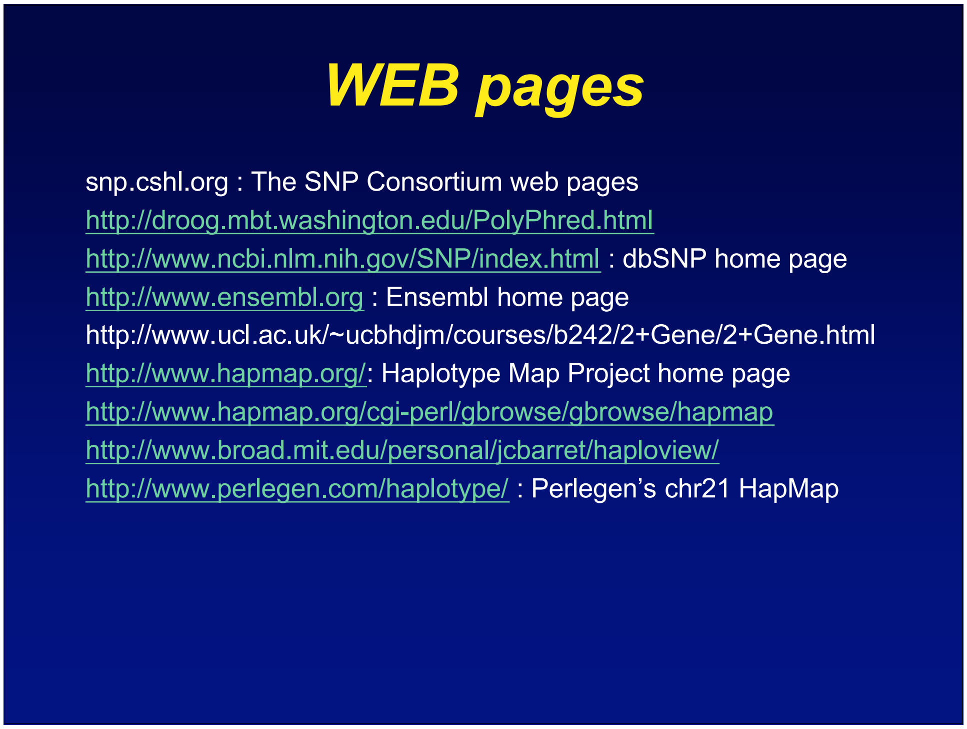
Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. Nat Genet. 2003 Apr;33(4):457-8.

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet. 2003 Apr;33(4):518-21.

Haplotype Map Project

Dennis C. Special section on human genetics: the rough guide to the genome. Nature. 2003 Oct 23;425(6960):758-9.

Goldstein DB. Islands of linkage disequilibrium. Nat Genet. 2001 Oct;29(2):109-11.



WEB pages

snp.cshl.org : The SNP Consortium web pages
http://droog.mbt.washington.edu/PolyPhred.html
http://www.ncbi.nlm.nih.gov/SNP/index.html : dbSNP home page
http://www.ensembl.org : Ensembl home page
http://www.ucl.ac.uk/~ucbhdjm/courses/b242/2+Gene/2+Gene.html
http://www.hapmap.org/: Haplotype Map Project home page
http://www.hapmap.org/cgi-perl/gbrowse/gbrowse/hapmap
http://www.broad.mit.edu/personal/jcbarret/haploview/
http://www.perlegen.com/haplotype/ : Perlegen’s chr21 HapMap