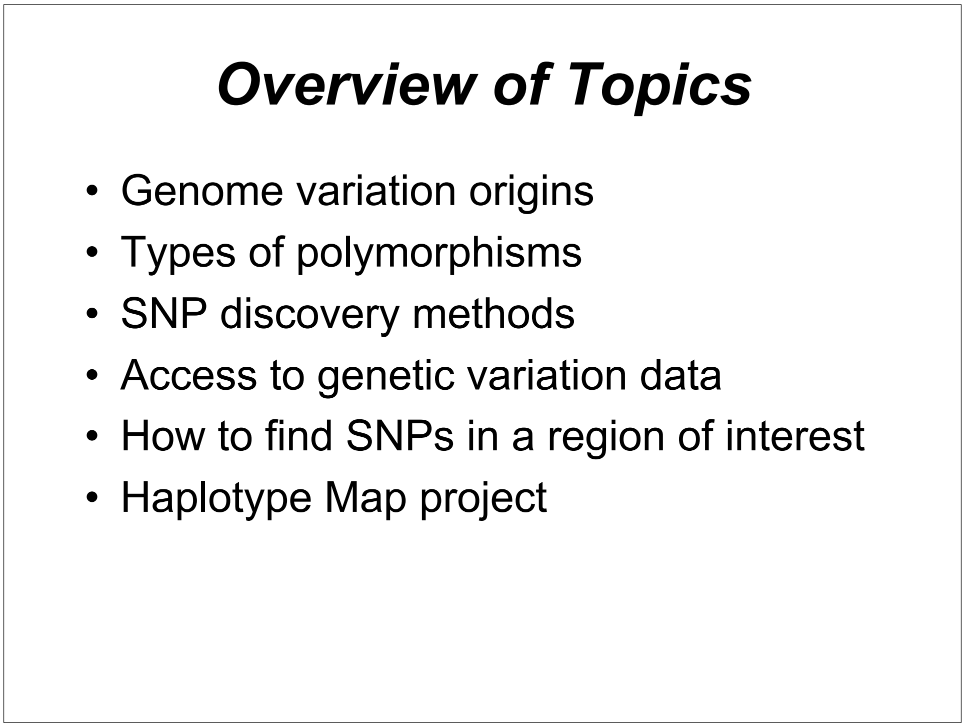
Studying Genetic Variation II: Computational Techniques

Jim Mullikin, PhD
Genome Technology Branch
NHGRI

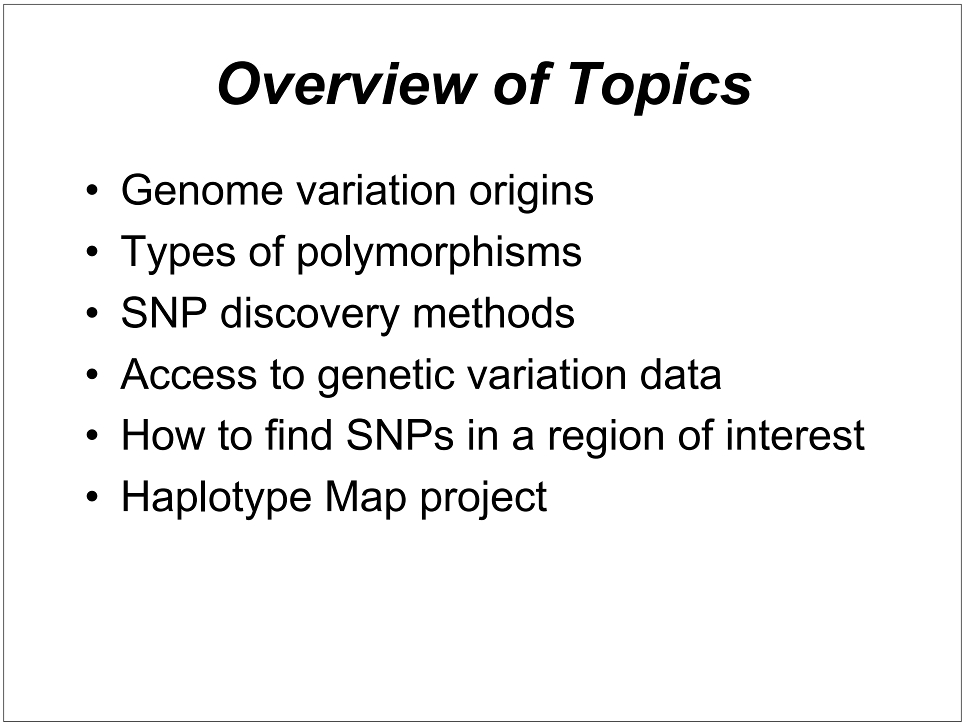
Some points from the previous two lectures

- Genetic maps, markers and linkage analysis by Elaine Ostrander
  - Genome wide scans for Mendelian inherited disease, microsatellites are still an effective marker to use
- Genetic Variation I: Laboratory Techniques by Karen Mohlke
  - Types of polymorphisms and genotyping methods, focusing primarily on SNP genotyping



Overview of Topics

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

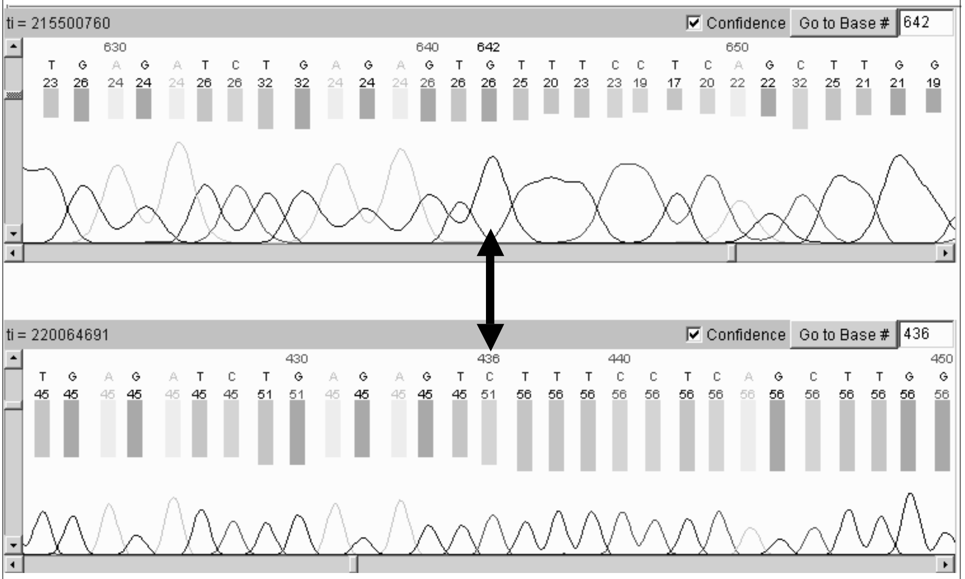


Overview of Topics

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

Discovery methods

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



ti = 215500760 ☑ Confidence Go to Base # 642

ti = 220064691 ☑ Confidence Go to Base # 436

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

Mining SNPs from sequence

- EST mining
- Clone overlap
- The SNP Consortium (TSC)
- Targeted resequencing
- Haplotype Map Project (HapMap)
- Chip based sequencing arrays

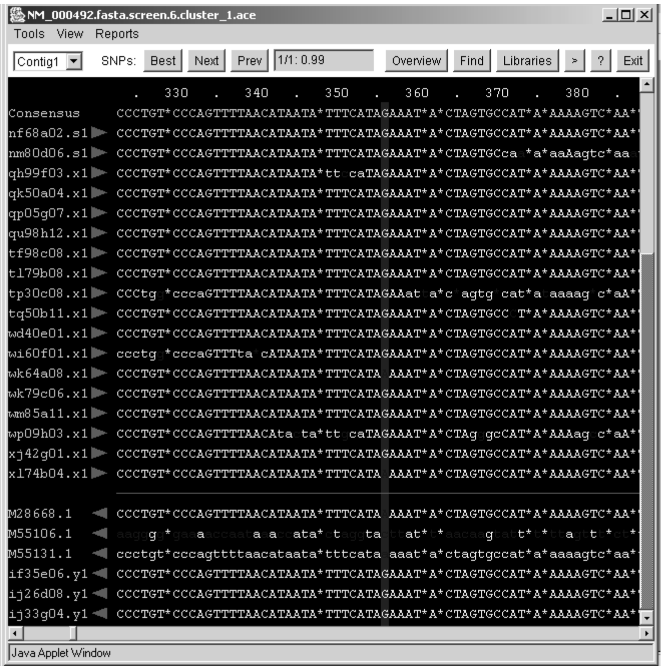
Expressed Sequence Tag Mining

- 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.

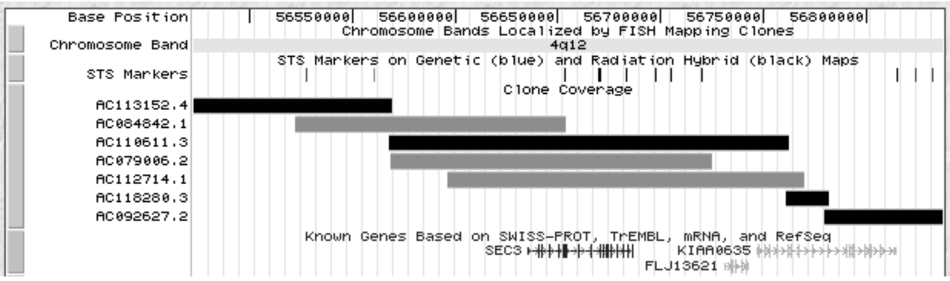
- The following example is from the CGAP SNP project (see refs).



```
                  .  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*aacaaggtatt*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.

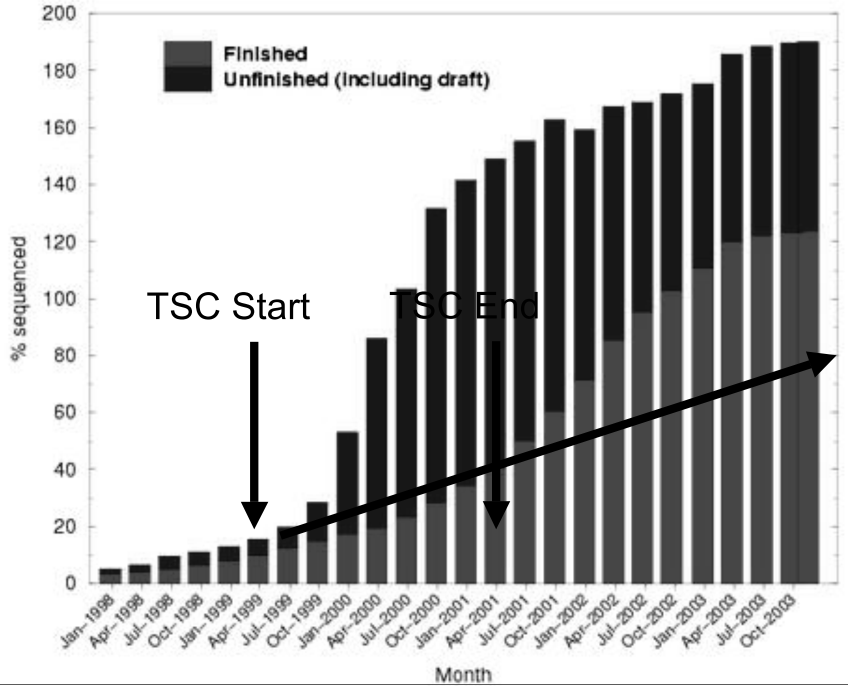


Clone Overlap

- 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 ethnic groups.



The SNP Consortium

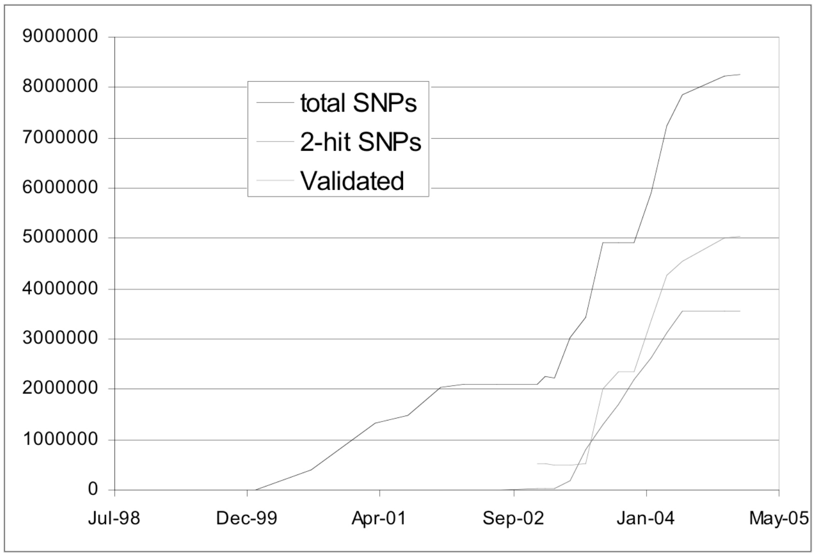
- With the rapid increase in genome coverage from the public Human Genome Project, the strategies changed to take full advantage of the draft and finished sequence.

- 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 8.2M SNPs today.
- It has been estimated that there are perhaps 10M common SNPs (> 5% MAF), so there are many more SNPs yet to discover.

SNPs in dbSNP

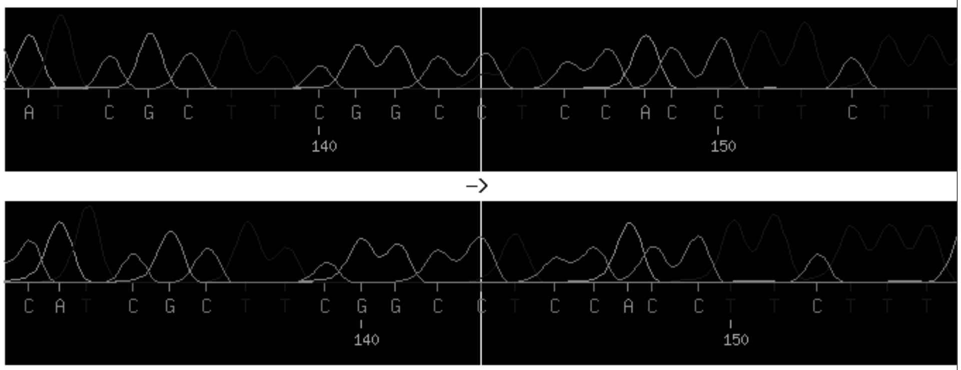


Targeted Resequencing

- 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

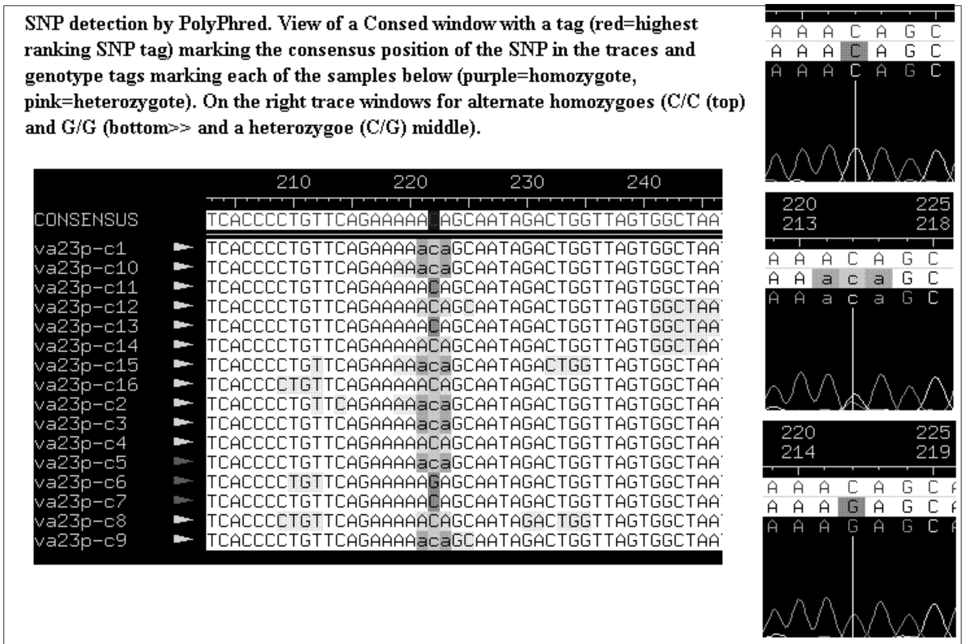


http://snp.ims.u-tokyo.ac.jp/

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.



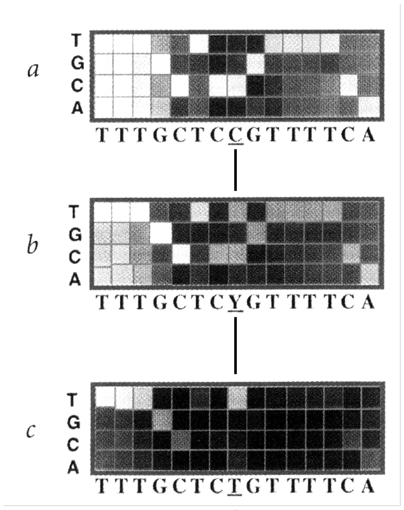
SNP detection by PolyPhred. View of a Consed window with a tag (red=highest ranking SNP tag) marking the consensus position of the SNP in the traces and genotype tags marking each of the samples below (purple=homozygote, pink=heterozygote). On the right trace windows for alternate homozygoes (C/C (top)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

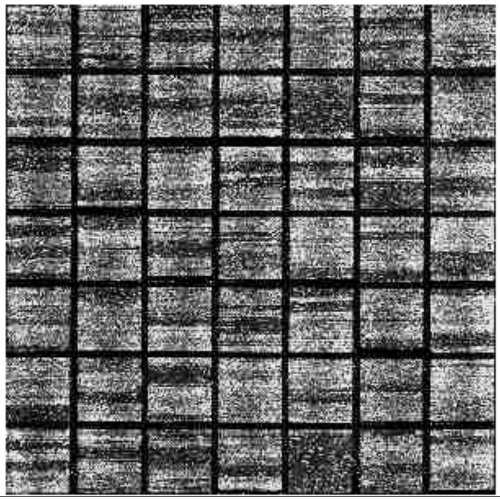
Sequencing Chips

...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

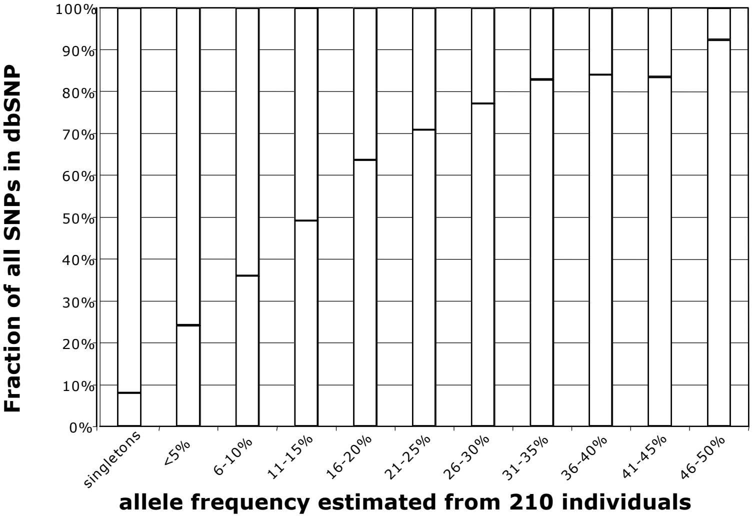
Distribution properties

- Haplotype Map Project (HapMap)
  - Random, like TSC, for first phase that reached 2X 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 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.

SNPs detected from 48 HapMap individuals gives an estimate dbSNP build 121 completeness



Overview of Topics

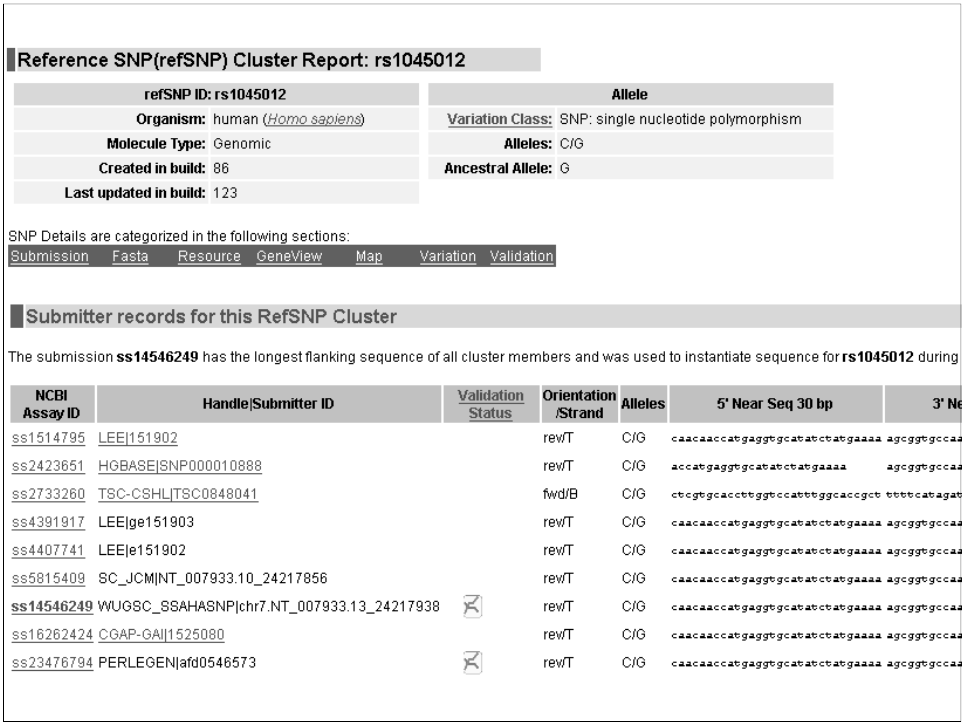
- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project

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(refSNP) Cluster Report: rs1045012

refSNP ID:	rs1045012
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created in build:	86
Last updated in build:	123

Allele	
Variation Class:	SNP: single nucleotide polymorphism
Alleles:	C/G
Ancestral Allele:	G

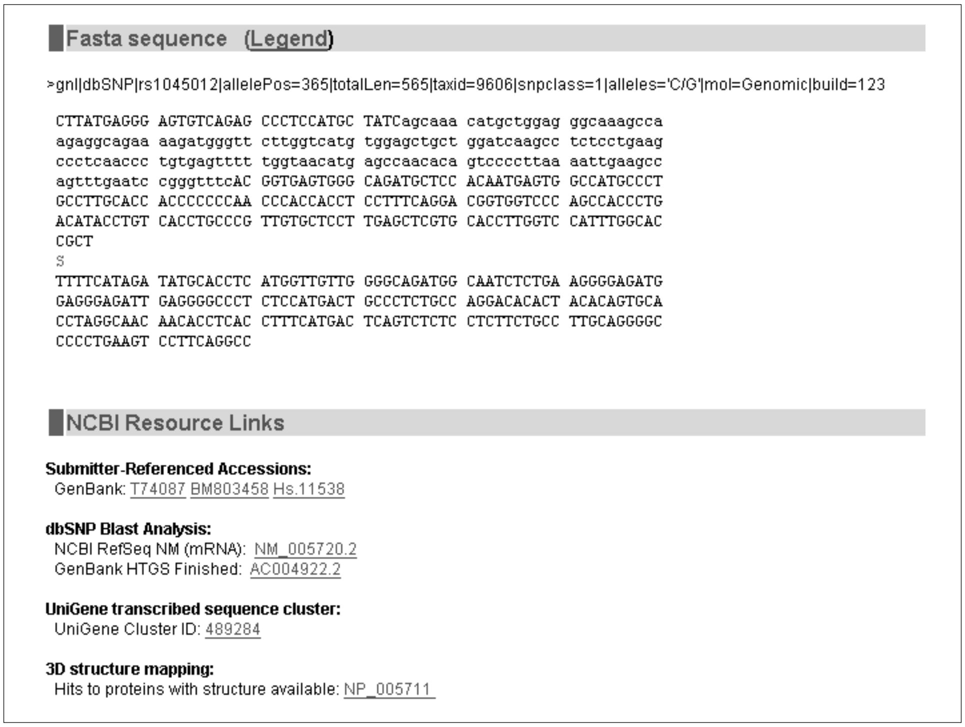
SNP Details are categorized in the following sections:

Submission Fasta Resource GeneView Map Variation Validation

Submitter records for this RefSNP Cluster

The submission **ss14546249** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs1045012** during

NCBI Assay ID	Handle|Submitter ID	Validation Status	Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Ne
ss1514795	LEE|151902		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss2423651	HGBASE|SNP000010888		rev/T	C/G	accatgaggtgcatatctatgaaaa	agcggtgccaa
ss2733260	TSC-CSHL|TSC0848041		fwd/B	C/G	ctcgtgcaccttggtccatttggcaccgct	ttttcatagat
ss4391917	LEE|ge151903		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss4407741	LEE|e151902		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss5815409	SC_JCM|NT_007933.10_24217856		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss14546249	WUGSC_SSAHASNP|chr7.NT_007933.13_24217938	⧖	rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss16262424	CGAP-GAI|1525080		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss23476794	PERLEGEN|afd0546573	⧖	rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa



Fasta sequence (Legend)

```
>gnl|dbSNP|rs1045012|allelePos=365|totalLen=565|taxid=9606|snpclass=1|alleles='C/G'|mol=Genomic|build=123

 CTTATGAGGG AGTGTCAGAG CCCTCCATGC TATCagcaaa catgctggag ggcaaagcca
 agaggcagaa aagatgggtt cttggtcatg tggagctgct ggatcaagcc tctcctgaag
 ccctcaaccc tgtgagtttt tggtaacatg agccaacaca gtccccttaa aattgaagcc
 agtttgaatc cgggtttcAC GGTGAGTGGG CAGATGCTCC ACAATGAGTG GCCATGCCCT
 GCCTTGCACC ACCCCCCCAA CCCACCACCT CCTTTCAGGA CGGTGGTCCC AGCCACCCTG
 ACATACCTGT CACCTGCCCG TTGTGCTCCT TGAGCTCGTG CACCTTGGTC CATTTGGCAC
 CGCT
 S
 TTTTCATAGA TATGCACCTC ATGGTTGTTG GGGCAGATGG CAATCTCTGA AGGGGAGATG
 GAGGGAGATT GAGGGGCCCT CTCCATGACT GCCCTCTGCC AGGACACACT ACACAGTGCA
 CCTAGGCAAC AACACCTCAC CTTTCATGAC TCAGTCTCTC CTCTTCTGCC TTGCAGGGGC
 CCCCTGAAGT CCTTCAGGCC
```

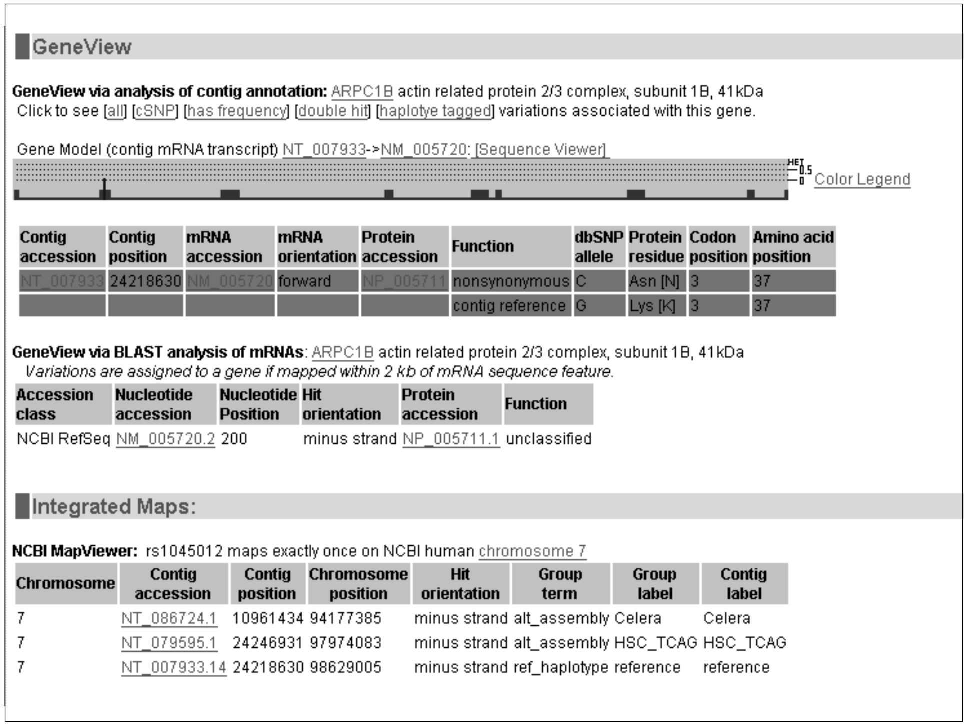
NCBI Resource Links

Submitter-Referenced Accessions:
GenBank: T74087 BM803458 Hs.11538

dbSNP Blast Analysis:
NCBI RefSeq NM (mRNA): NM_005720.2
GenBank HTGS Finished: AC004922.2

UniGene transcribed sequence cluster:
UniGene Cluster ID: 489284

3D structure mapping:
Hits to proteins with structure available: NP_005711



GeneView

GeneView via analysis of contig annotation: ARPC1B actin related protein 2/3 complex, subunit 1B, 41kDa
Click to see [all] [cSNP] [has frequency] [double hit] [haplotye tagged] variations associated with this gene.

Gene Model (contig mRNA transcript) NT_007933->NM_005720: [Sequence Viewer]

Color Legend

Contig accession	Contig position	mRNA accession	mRNA orientation	Protein accession	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
NT_007933	24218630	NM_005720	forward	NP_005711	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37

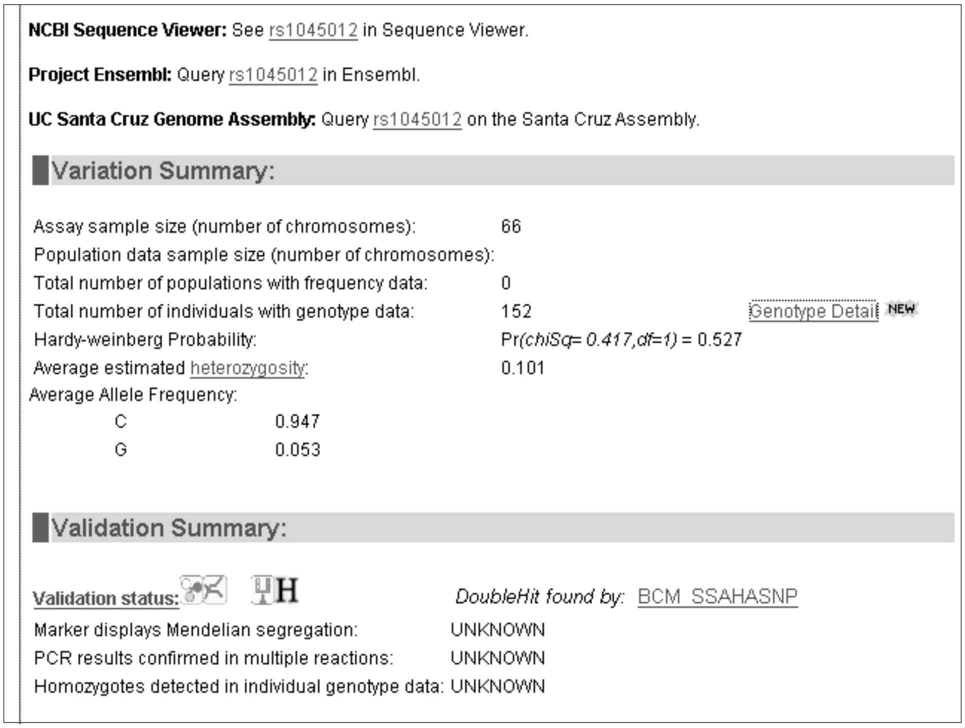
GeneView via BLAST analysis of mRNAs: ARPC1B actin related protein 2/3 complex, subunit 1B, 41kDa
Variations are assigned to a gene if mapped within 2 kb of mRNA sequence feature.

Accession class	Nucleotide accession	Nucleotide Position	Hit orientation	Protein accession	Function
NCBI RefSeq	NM_005720.2	200	minus strand	NP_005711.1	unclassified

Integrated Maps:

NCBI MapViewer: rs1045012 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_086724.1	10961434	94177385	minus strand	alt_assembly	Celera	Celera
7	NT_079595.1	24246931	97974083	minus strand	alt_assembly	HSC_TCAG	HSC_TCAG
7	NT_007933.14	24218630	98629005	minus strand	ref_haplotype	reference	reference



NCBI Sequence Viewer: See rs1045012 in Sequence Viewer.

Project Ensembl: Query rs1045012 in Ensembl.

UC Santa Cruz Genome Assembly: Query rs1045012 on the Santa Cruz Assembly.

Variation Summary:

Assay sample size (number of chromosomes):	66	
Population data sample size (number of chromosomes):		
Total number of populations with frequency data:	0	
Total number of individuals with genotype data:	152	Genotype Detail NEW
Hardy-weinberg Probability:	Pr(*chiSq= 0.417,df=1*) = 0.527	
Average estimated heterozygosity:	0.101	

Average Allele Frequency:

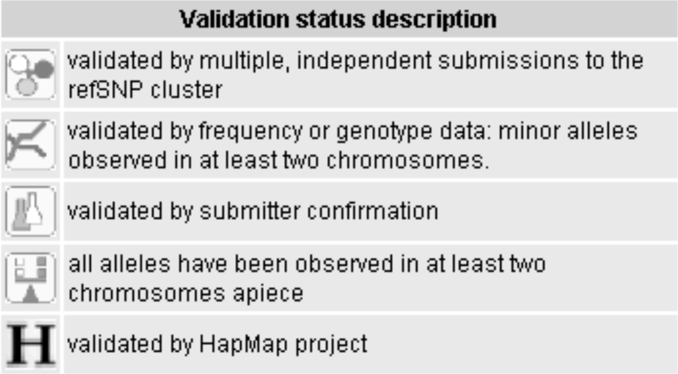
C	0.947
G	0.053

Validation Summary:

Validation status: H *DoubleHit found by:* BCM_SSAHASNP

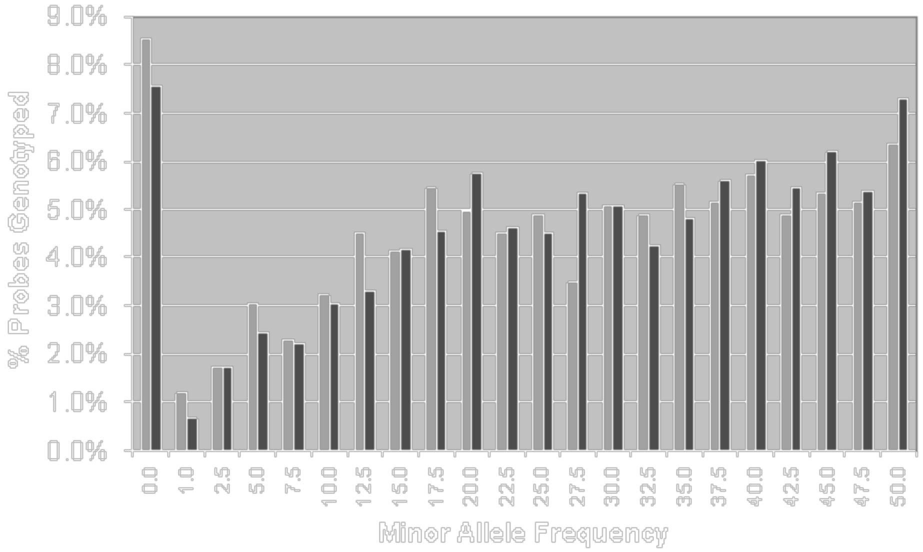
Marker displays Mendelian segregation:	UNKNOWN
PCR results confirmed in multiple reactions:	UNKNOWN
Homozygotes detected in individual genotype data:	UNKNOWN

Validation summary



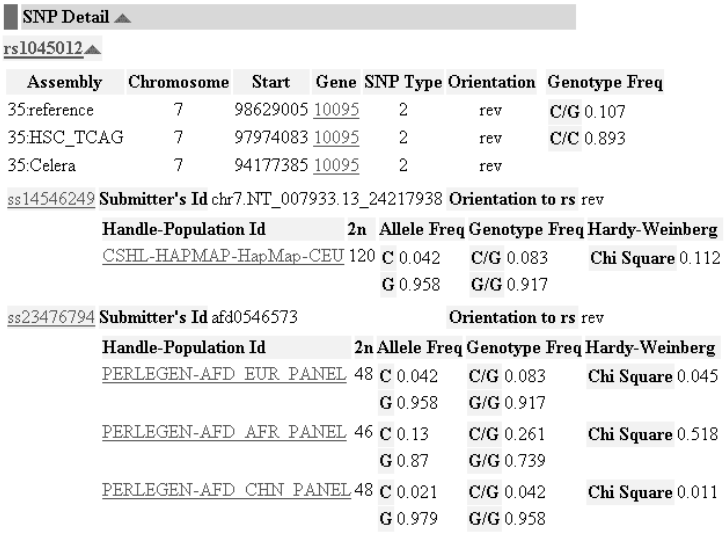
Validation status description	
	validated by multiple, independent submissions to the refSNP cluster
	validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
	validated by submitter confirmation
	all alleles have been observed in at least two chromosomes apiece
H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

Double hit SNP minor allele frequency characteristics



Credit: Dr. Paul Hardenbol, Parallele Bioscience

Genotype Detail



SNP Detail

rs1045012

Assembly	Chromosome	Start	Gene	SNP Type	Orientation	Genotype Freq
35:reference	7	98629005	10095	2	rev	C/G 0.107
35:HSC_TCAG	7	97974083	10095	2	rev	C/C 0.893
35:Celera	7	94177385	10095	2	rev	

ss14546249 **Submitter's Id** chr7.NT_007933.13_24217938 **Orientation to rs** rev

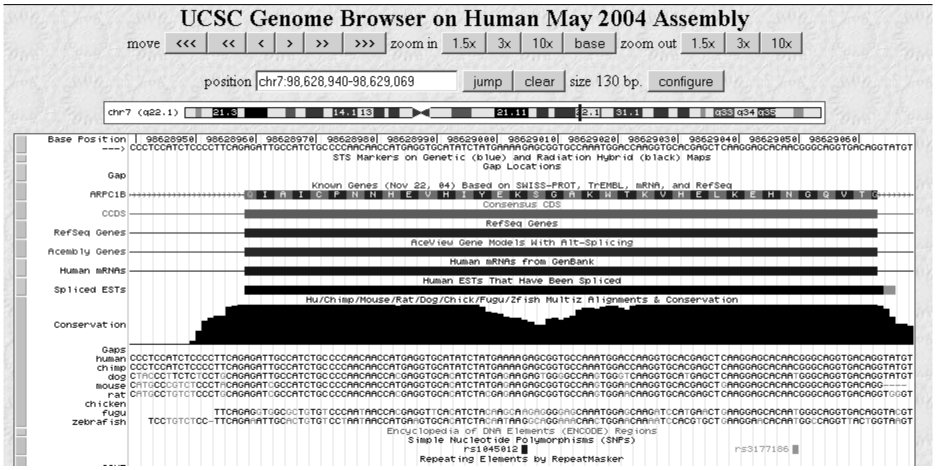
Handle-Population Id	2n	Allele Freq	Genotype Freq	Hardy-Weinberg
CSHL-HAPMAP-HapMap-CEU	120	C 0.042 G 0.958	C/G 0.083 G/G 0.917	Chi Square 0.112

ss23476794 **Submitter's Id** afd0546573 **Orientation to rs** rev

Handle-Population Id	2n	Allele Freq	Genotype Freq	Hardy-Weinberg
PERLEGEN-AFD_EUR_PANEL	48	C 0.042 G 0.958	C/G 0.083 G/G 0.917	Chi Square 0.045
PERLEGEN-AFD_AFR_PANEL	46	C 0.13 G 0.87	C/G 0.261 G/G 0.739	Chi Square 0.518
PERLEGEN-AFD_CHN_PANEL	48	C 0.021 G 0.979	C/G 0.042 G/G 0.958	Chi Square 0.011

Viewing SNPs in Browsers

NCBI Ensembl UCSC



UCSC Genome Browser on Human May 2004 Assembly

position chr7:98,628,940-98,629,069 size 130 bp.

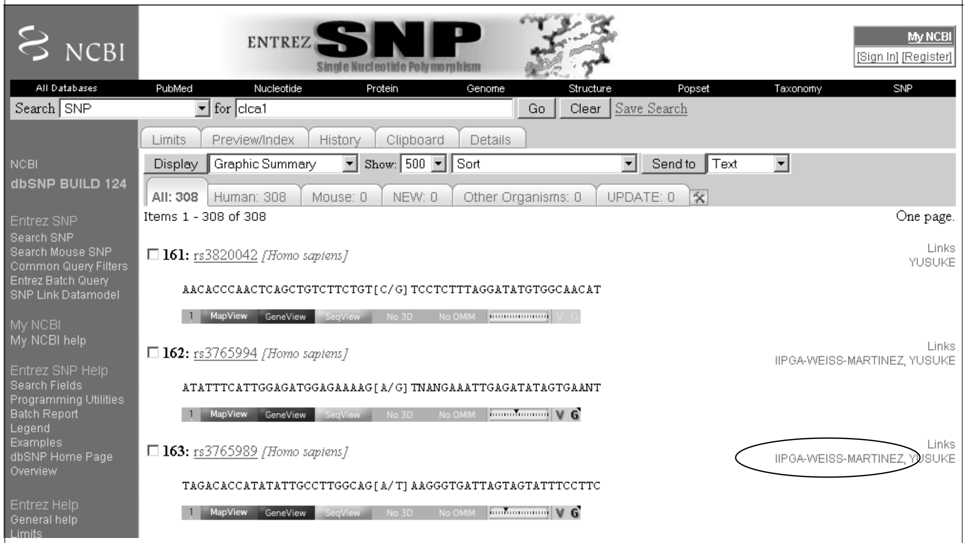
```
Base Position | 98628950| 98628960| 98628970| 98628980| 98628990| 98629000| 98629010| 98629020| 98629030| 98629040| 98629050| 98629060|
         ---> CCCTCCATCTCCCCTTCAGAGATTGCCATCTGCCCCAACAACCATGAGGTGCATATCTATGAAAAGAGCGGTGCCAAATGGACCAAGGTGCACGAGCTCAAGGAGCACAACGGGCAGGTGACAGGTATGT
                        STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
                                                  Gap Locations
          Gap
                         Known Genes (Nov 22, 04) Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
       ARPC1B                      Q I A I C P N N H E V H I Y E K S G A K W T K V H E L K E H N G Q V T G
                                                  Consensus CDS
         CCDS
                                                  RefSeq Genes
 RefSeq Genes
                                        AceView Gene Models With Alt-Splicing
Acembly Genes
                                             Human mRNAs from GenBank
  Human mRNAs
                                          Human ESTs That Have Been Spliced
 Spliced ESTs
                          Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
 Conservation
         Gaps
        human CCCTCCATCTCCCCTTCAGAGATTGCCATCTGCCCCAACAACCATGAGGTGCATATCTATGAAAAGAGCGGTGCCAAATGGACCAAGGTGCACGAGCTCAAGGAGCACAACGGGCAGGTGACAGGTATGT
        chimp CCCTCCATCTCCCCTTCAGAGATTGCCATCTGCCCCAACAACCATGAGGTGCATATCTATGAAAAGAGCGGTGCCAAATGGACCAAGGTGCACGAGCTCAAGGAGCACAACGGGCAGGTGACAGGTATGT
          dog CTACCCTTCTCTCCTGCAGAGATTGCCATCTGCCCCAACAACCACGAGGTGCACATCTATGAGAAGAGTGGGGCCAAGTGGGTCAAGGTGCATGAGCTCAAGGAGCACAATGGGCAGGTGACAGGTATGT
        mouse CATGCCCGTCTCCCTACAGAGATCGCCATCTGCCCCAACAACCATGAGGTGCACATCTATGAGAAGAGCGGTGCCAAGTGGAACAAGGTGCACGAGCTGAAGGAGCACAACGGGCAGGTGACAGG----
          rat CATGCCTGTCTCCCTGCAGAGATCGCCATCTGCCCCAACAACCACGAGGTGCACATCTACGAGAAGAGCGGTGCCAAGTGGAACAAGGTGCACGAGCTCAAGGAGCACAACGGGCAGGTGACAGGTGGGT
      chicken
         fugu                   TTCAGAGGTGGCGCTGTGTCCCAATAACCACGAGGTTCACATCTACAAGCAAGAGGGGAGCAAATGGAGCAAGATCCATGAACTGAAGGAGCACAATGGGCAGGTGACAGGTACGT
    zebrafish     TCCTGTCTCC-TTCAGAAATTGCACTGTGTCCTAATAACCATGAAGTGCACATCTACAATAAGGCAGGAAACAACTGGAACAAAATCCACGTGCTGAAGGAACACAATGGCCAGGTTACTGGTAAGT
                                    Encyclopedia of DNA Elements (ENCODE) Regions
                                         Simple Nucleotide Polymorphisms (SNPs)
                                          rs1045012                          rs3177186
                                           Repeating Elements by RepeatMasker
```

Overview of Topics

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

How to find SNPs in a region of interest

- Gene based example
- A 2 Mbp region
- From a list of candidate genes



Search SNP for clca1

Display Graphic Summary Show: 500 Sort Send to Text

Items 1 - 308 of 308 One page.

□ **161:** rs3820042 *[Homo sapiens]* Links YUSUKE

```
AACACCCAACTCAGCTGTCTTCTGT[C/G] TCCTCTTTAGGATATGTGGCAACAT
```

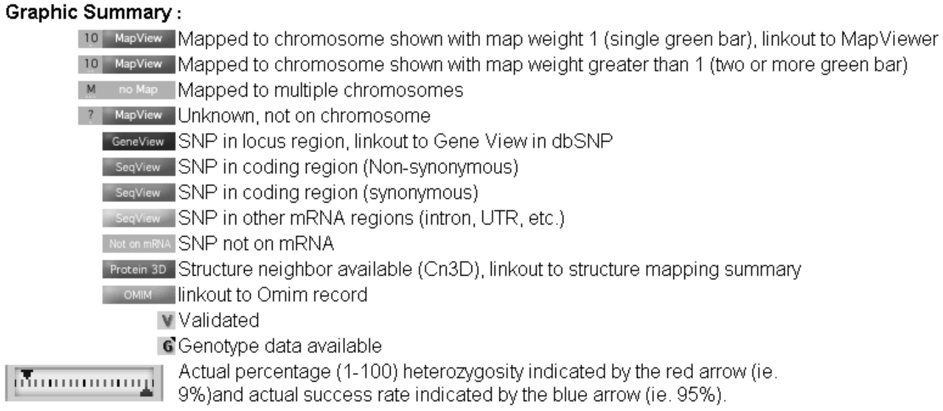
□ **162:** rs3765994 *[Homo sapiens]* Links IIPGA-WEISS-MARTINEZ, YUSUKE

```
ATATTTCATTGGAGATGGAGAAAAG[A/G] TNANGAAATTGAGATATAGTGAANT
```

□ **163:** rs3765989 *[Homo sapiens]* Links IIPGA-WEISS-MARTINEZ, YUSUKE

```
TAGACACCATATATTGCCTTGGCAG[A/T] AAGGGTGATTAGTAGTATTTCCTTC
```

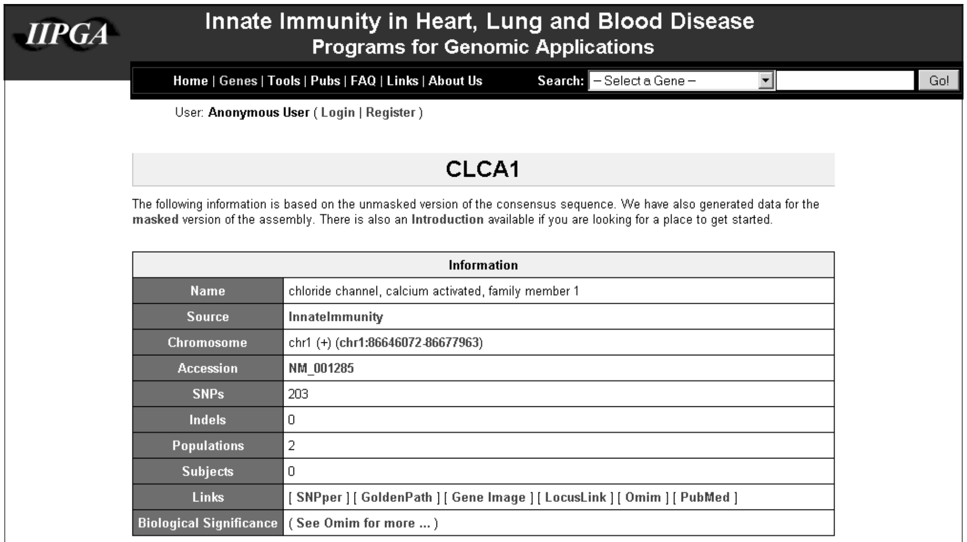
http://www.ncbi.nlm.nih.gov/SNP/index.html



Graphic Summary :

10 MapView	Mapped to chromosome shown with map weight 1 (single green bar), linkout to MapViewer
10 MapView	Mapped to chromosome shown with map weight greater than 1 (two or more green bar)
M no Map	Mapped to multiple chromosomes
? MapView	Unknown, not on chromosome
GeneView	SNP in locus region, linkout to Gene View in dbSNP
SeqView	SNP in coding region (Non-synonymous)
SeqView	SNP in coding region (synonymous)
SeqView	SNP in other mRNA regions (intron, UTR, etc.)
Not on mRNA	SNP not on mRNA
Protein 3D	Structure neighbor available (Cn3D), linkout to structure mapping summary
OMIM	linkout to Omim record
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



Innate Immunity in Heart, Lung and Blood Disease
Programs for Genomic Applications

Home | Genes | Tools | Pubs | FAQ | Links | About Us
Search: -- Select a Gene --
Go!

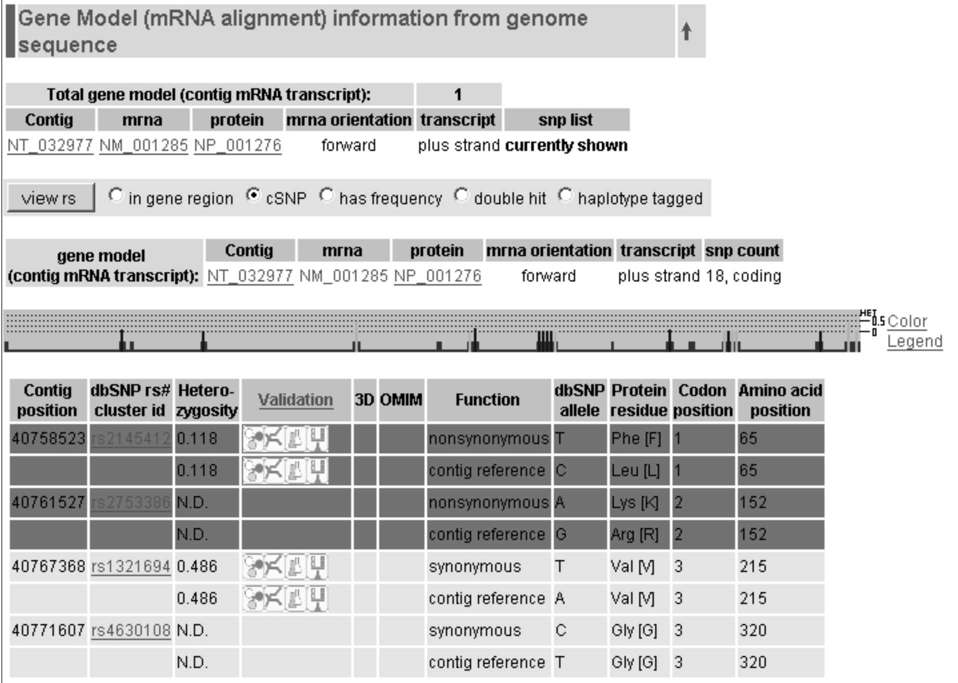
User: **Anonymous User** (Login | Register)

CLCA1

The following information is based on the unmasked version of the consensus sequence. We have also generated data for the **masked** version of the assembly. There is also an **Introduction** available if you are looking for a place to get started.

Information	
Name	chloride channel, calcium activated, family member 1
Source	**InnateImmunity**
Chromosome	chr1 (+) (**chr1:86646072-86677963**)
Accession	**NM_001285**
SNPs	203
Indels	0
Populations	2
Subjects	0
Links	[SNPper] [GoldenPath] [Gene Image] [LocusLink] [Omim] [PubMed]
Biological Significance	(See Omim for more ...)

http://innateimmunity.net/IIPGA/PGAs/InnateImmunity/CLCA1



Gene Model (mRNA alignment) information from genome sequence

Total gene model (contig mRNA transcript):			1	
Contig	**mrna**	**protein**	**mrna orientation**	**transcript**

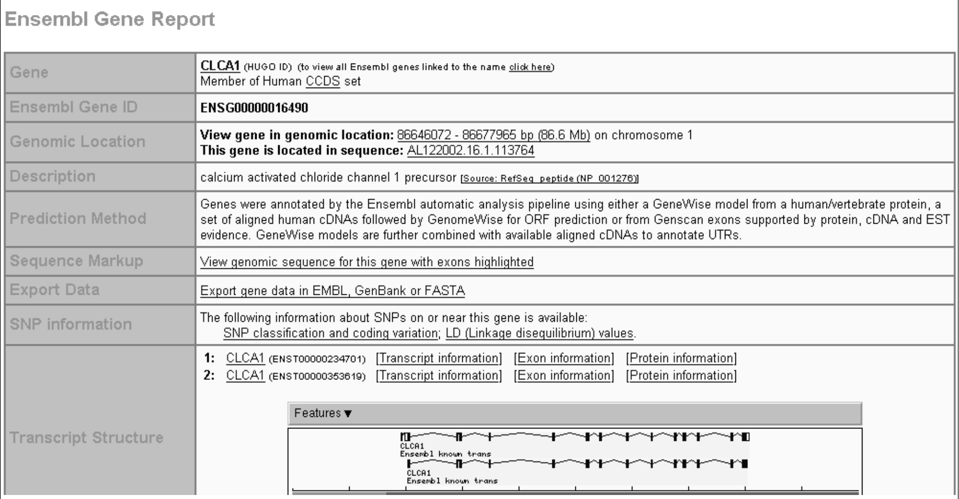
Contig	mrna	protein	mrna orientation	transcript	snp list
NT_032977	NM_001285	NP_001276	forward	plus strand	**currently shown**

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

gene model (contig mRNA transcript):	Contig	mrna	protein	mrna orientation	transcript	snp count
	NT_032977	NM_001285	NP_001276	forward	plus strand	18, coding

Color Legend

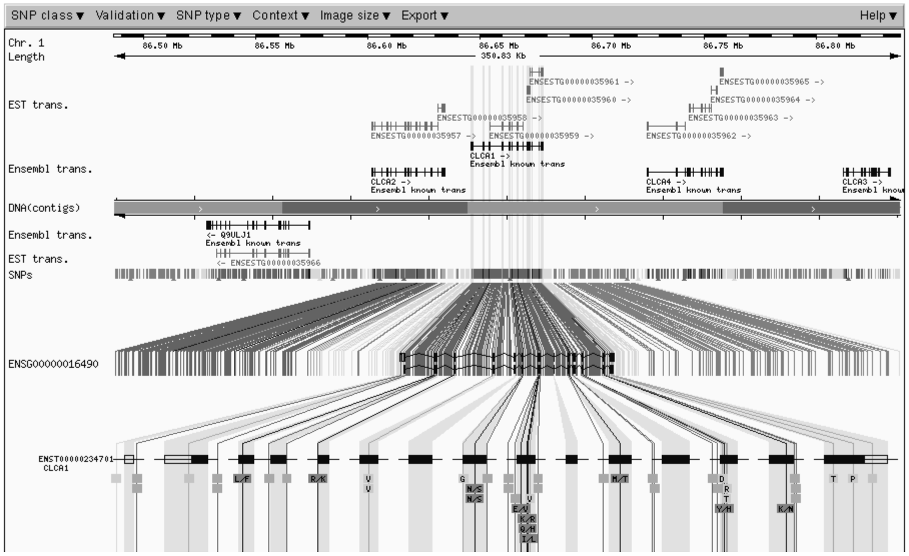
Contig position	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon position	Amino acid position
40758523	rs2145412	0.118				nonsynonymous	T	Phe [F]	1	65
		0.118				contig reference	C	Leu [L]	1	65
40761527	rs2753386	N.D.				nonsynonymous	A	Lys [K]	2	152
		N.D.				contig reference	G	Arg [R]	2	152
40767368	rs1321694	0.486				synonymous	T	Val [V]	3	215
		0.486				contig reference	A	Val [V]	3	215
40771607	rs4630108	N.D.				synonymous	C	Gly [G]	3	320
		N.D.				contig reference	T	Gly [G]	3	320

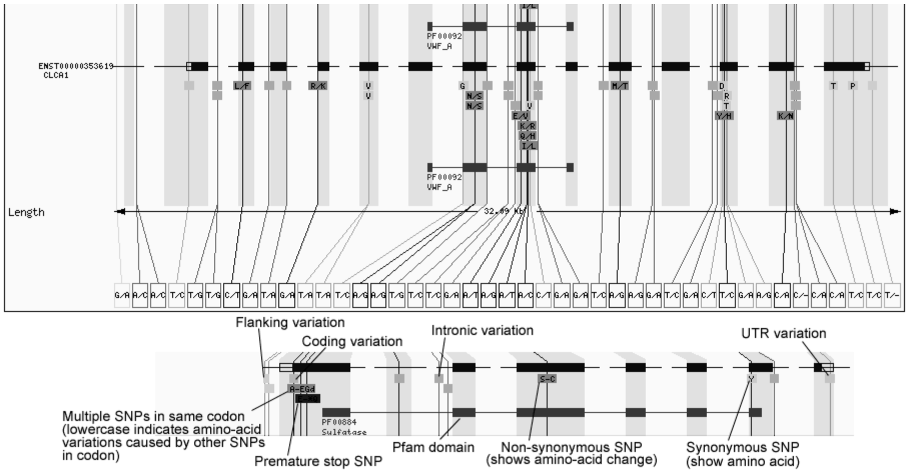


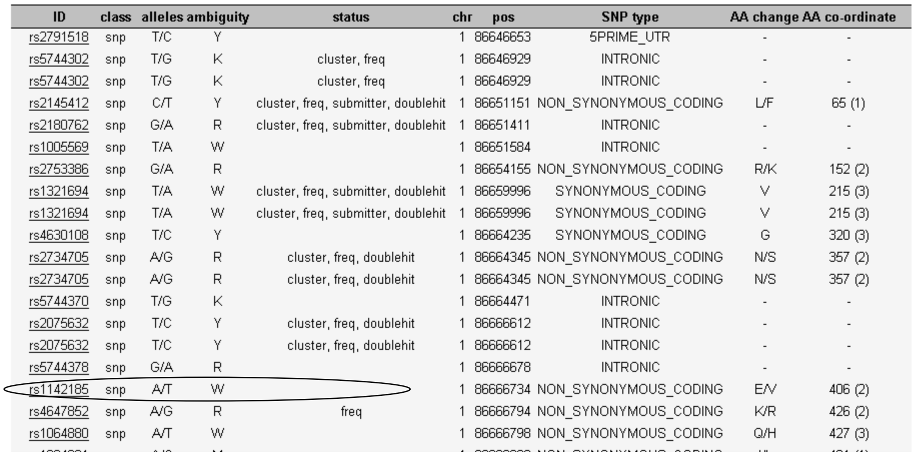
Ensembl Gene Report

Gene	CLCA1 (HUGO ID) (to view all Ensembl genes linked to the name click here) Member of Human CCDS set
Ensembl Gene ID	**ENSG00000016490**
Genomic Location	**View gene in genomic location:** 86646072 - 86677965 bp (86.6 Mb) on chromosome 1 **This gene is located in sequence:** AL122002.16.1.113764
Description	calcium activated chloride channel 1 precursor [Source: RefSeq_peptide (NP_001276)]
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise model from a human/vertebrate protein, a set of aligned human cDNAs followed by GenomeWise for ORF prediction or from Genscan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.
Sequence Markup	View genomic sequence for this gene with exons highlighted
Export Data	Export gene data in EMBL, GenBank or FASTA
SNP information	The following information about SNPs on or near this gene is available: SNP classification and coding variation; LD (Linkage disequilibrium) values.
Transcript Structure	**1:** CLCA1 (ENST00000234701) [Transcript information] [Exon information] [Protein information] **2:** CLCA1 (ENST00000353619) [Transcript information] [Exon information] [Protein information]

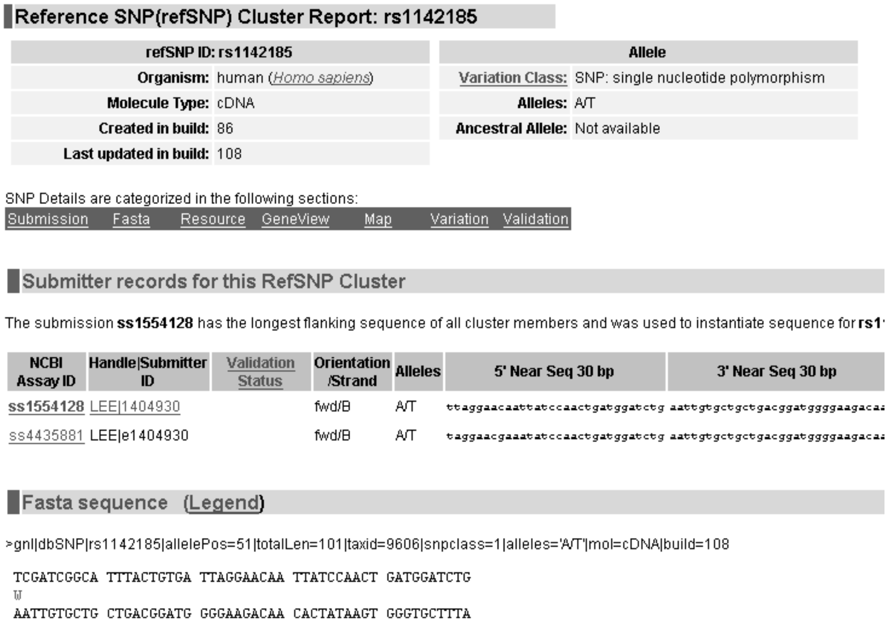
http://www.ensembl.org/Homo_sapiens


SNP class ▾ Validation ▾ SNP type ▾ Context ▾ Image size ▾ Export ▾ Help ▾
Chr. 1
Length
86.50 Mb
86.55 Mb
86.60 Mb
86.65 Mb
86.70 Mb
86.75 Mb
86.80 Mb
350.83 Kb
EST trans.
Ensembl trans.
DNA(contigs)
Ensembl trans.
EST trans.
SNPs
ENSG00000016490
ENST00000234701
CLCA1
CLCA1 ->
Ensembl known trans
CLCA2 ->
CLCA4 ->
CLCA3 ->
<- Q9ULJ1
ENSESTG00000035957 ->
ENSESTG00000035958 ->
ENSESTG00000035959 ->
ENSESTG00000035960 ->
ENSESTG00000035961 ->
ENSESTG00000035962 ->
ENSESTG00000035963 ->
ENSESTG00000035964 ->
ENSESTG00000035965 ->
<- ENSESTG00000035966


PF00092
VWF_A
ENST00000353619
CLCA1
Length
32.09 Kb
Flanking variation
Coding variation
Intronic variation
UTR variation
Multiple SNPs in same codon (lowercase indicates amino-acid variations caused by other SNPs in codon)
Premature stop SNP
Pfam domain
Non-synonymous SNP (shows amino-acid change)
Synonymous SNP (show amino acid)
PF00884
Sulfatase



ID	class	alleles	ambiguity	status	chr	pos	SNP type	AA change	AA co-ordinate
rs2791518	snp	T/C	Y		1	86646653	5PRIME_UTR	-	-
rs5744302	snp	T/G	K	cluster, freq	1	86646929	INTRONIC	-	-
rs5744302	snp	T/G	K	cluster, freq	1	86646929	INTRONIC	-	-
rs2145412	snp	C/T	Y	cluster, freq, submitter, doublehit	1	86651151	NON_SYNONYMOUS_CODING	L/F	65 (1)
rs2180762	snp	G/A	R	cluster, freq, submitter, doublehit	1	86651411	INTRONIC	-	-
rs1005569	snp	T/A	W		1	86651584	INTRONIC	-	-
rs2753386	snp	G/A	R		1	86654155	NON_SYNONYMOUS_CODING	R/K	152 (2)
rs1321694	snp	T/A	W	cluster, freq, submitter, doublehit	1	86659996	SYNONYMOUS_CODING	V	215 (3)
rs1321694	snp	T/A	W	cluster, freq, submitter, doublehit	1	86659996	SYNONYMOUS_CODING	V	215 (3)
rs4630108	snp	T/C	Y		1	86664235	SYNONYMOUS_CODING	G	320 (3)
rs2734705	snp	A/G	R	cluster, freq, doublehit	1	86664345	NON_SYNONYMOUS_CODING	N/S	357 (2)
rs2734705	snp	A/G	R	cluster, freq, doublehit	1	86664345	NON_SYNONYMOUS_CODING	N/S	357 (2)
rs5744370	snp	T/G	K		1	86664471	INTRONIC	-	-
rs2075632	snp	T/C	Y	cluster, freq, doublehit	1	86666612	INTRONIC	-	-
rs2075632	snp	T/C	Y	cluster, freq, doublehit	1	86666612	INTRONIC	-	-
rs5744378	snp	G/A	R		1	86666678	INTRONIC	-	-
rs1142185	snp	A/T	W		1	86666734	NON_SYNONYMOUS_CODING	E/V	406 (2)
rs4647852	snp	A/G	R	freq	1	86666794	NON_SYNONYMOUS_CODING	K/R	426 (2)
rs1064880	snp	A/T	W		1	86666798	NON_SYNONYMOUS_CODING	Q/H	427 (3)



Reference SNP(refSNP) Cluster Report: rs1142185

refSNP ID:	rs1142185
Organism:	human (*Homo sapiens*)
Molecule Type:	cDNA
Created in build:	86
Last updated in build:	108

Allele	
Variation Class:	SNP: single nucleotide polymorphism
Alleles:	A/T
Ancestral Allele:	Not available

SNP Details are categorized in the following sections:

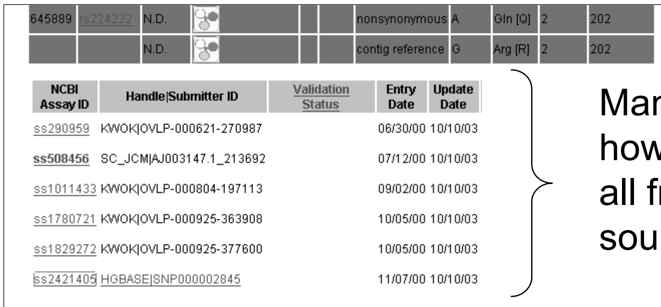
Submitter records for this RefSNP Cluster

The submission **ss1554128** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs1**

NCBI Assay ID	Handle|Submitter ID	Validation Status	Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp
ss1554128	LEE|1404930		fwd/B	A/T	ttaggaacaattatccaactgatggatctg	aattgtgctgctgacggatggggaagacaa
ss4435881	LEE|e1404930		fwd/B	A/T	taggaacgaaatatccaactgatggatctg	aattgtgctgctgacggatggggaagacaa

Fasta sequence (Legend)

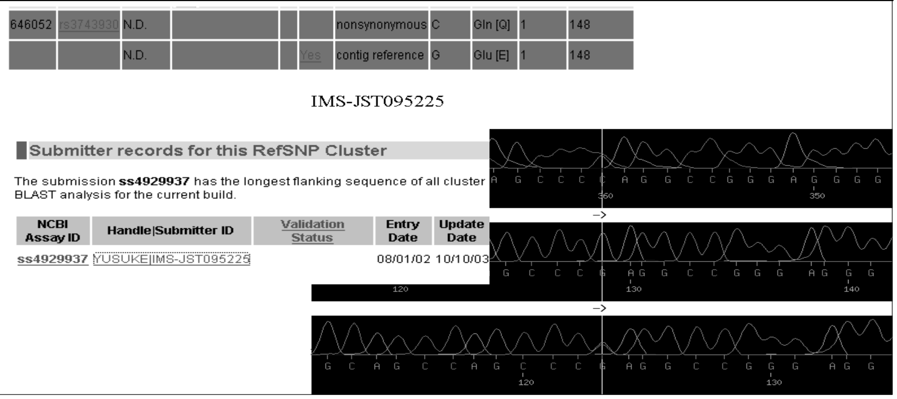
```
>gnl|dbSNP|rs1142185|allelePos=51|totalLen=101|taxid=9606|snpclass=1|alleles='A/T'|mol=cDNA|build=108

 TCGATCGGCA TTTACTGTGA TTAGGAACAA TTATCCAACT GATGGATCTG
 W
 AATTGTGCTG CTGACGGATG GGGAAGACAA CACTATAAGT GGGTGCTTTA
```



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

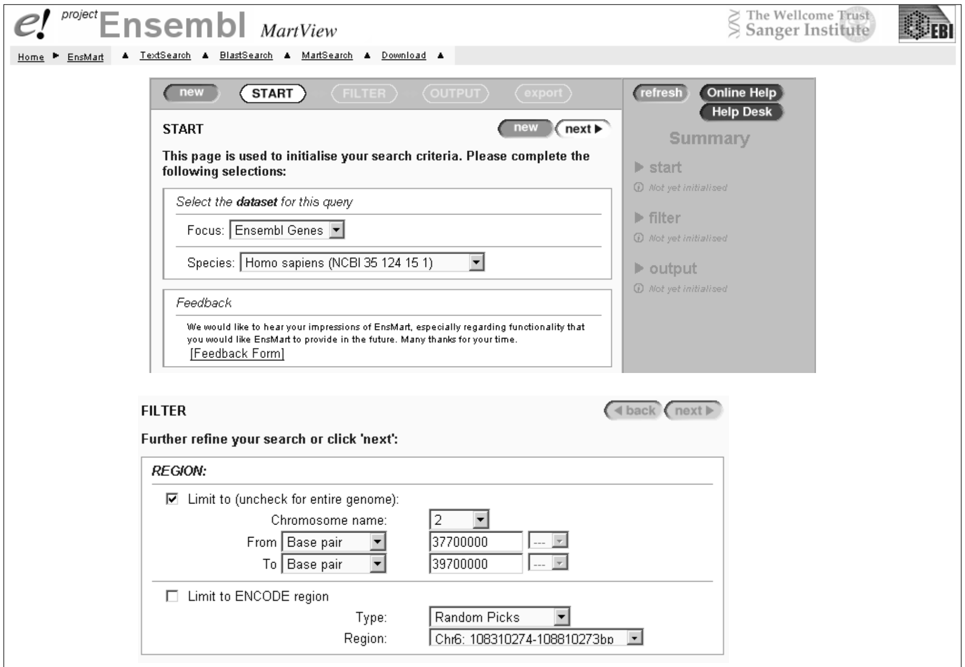
The submission **ss4929937** has the longest flanking sequence of all cluster BLAST analysis for the current build.

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

How to find SNPs in a region of interest

- Gene based example
- A 2 Mbp region
- From a list of candidate genes



START

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

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

Focus: Ensembl Genes

Species: Homo sapiens (NCBI 35 124 15 1)

Feedback

We would like to hear your impressions of EnsMart, especially regarding functionality that you would like EnsMart to provide in the future. Many thanks for your time.
[Feedback Form]

FILTER

Further refine your search or click 'next':

REGION:

☑ Limit to (uncheck for entire genome):

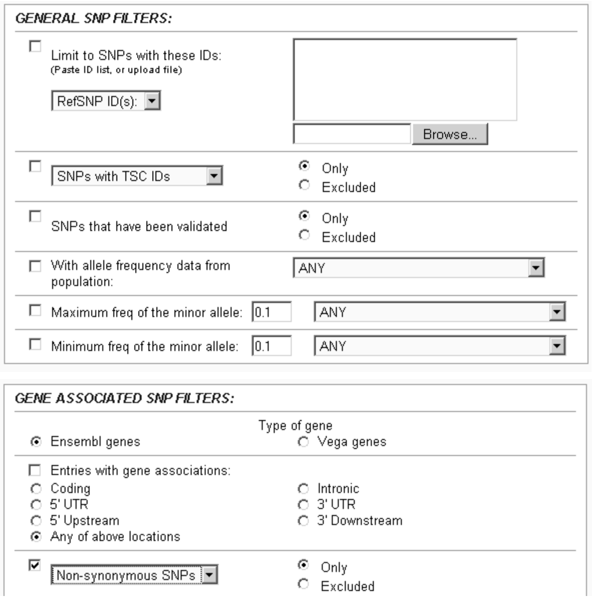
Chromosome name:	2	
From Base pair	37700000	---
To Base pair	39700000	---

☐ 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

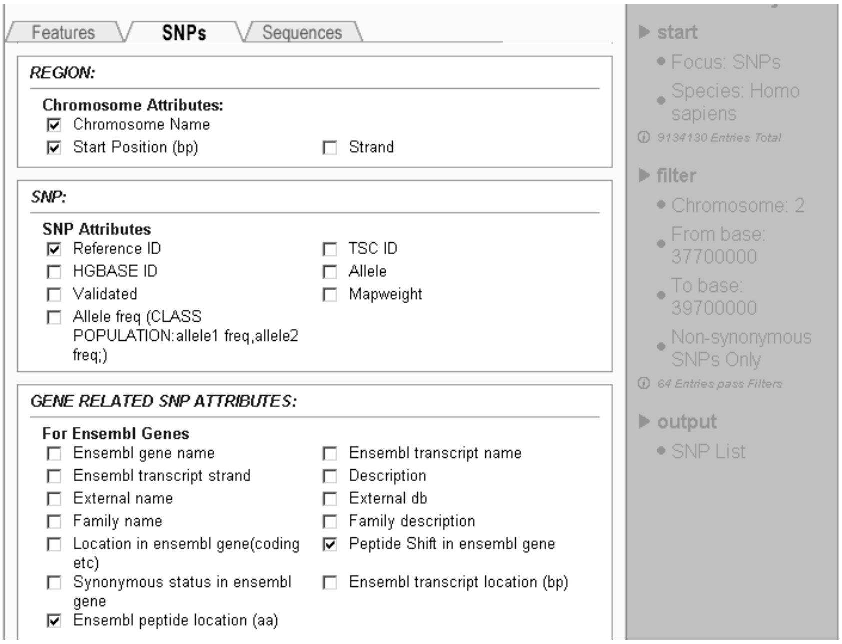
GENE ASSOCIATED SNP FILTERS:

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



REGION:

Chromosome Attributes:

☑ Chromosome Name

☑ Start Position (bp) ☐ Strand

SNP:

SNP Attributes

☑ Reference ID ☐ TSC ID

☐ HGBASE ID ☐ Allele

☐ Validated ☐ Mapweight

☐ Allele freq (CLASS POPULATION:allele1 freq,allele2 freq;)

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 peptide location (aa)

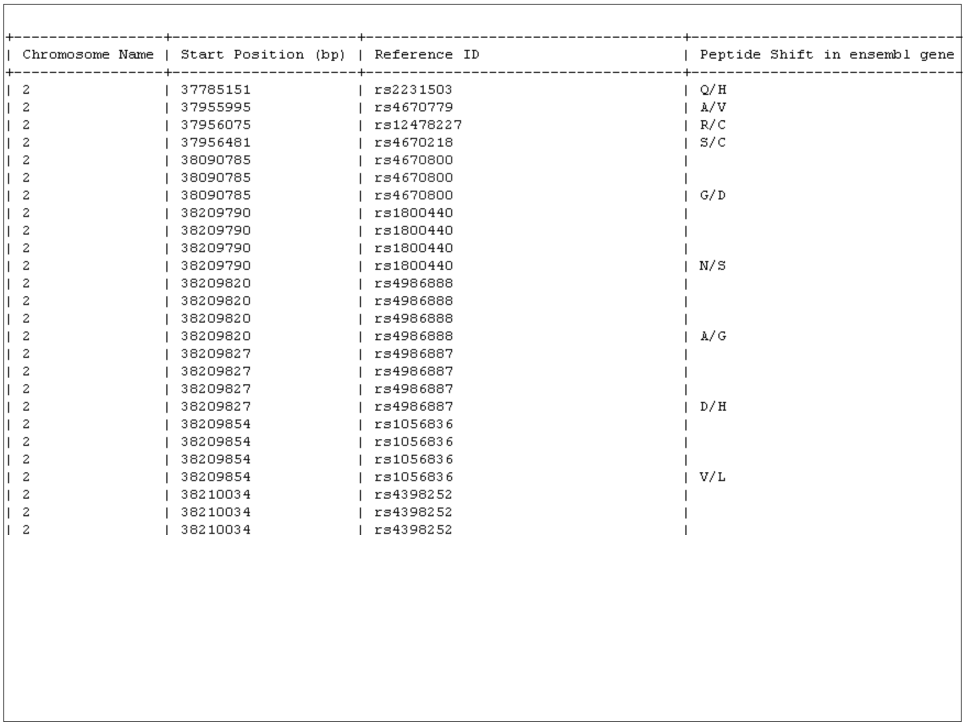
▶ start
- Focus: SNPs
- Species: Homo sapiens

9134130 Entries Total

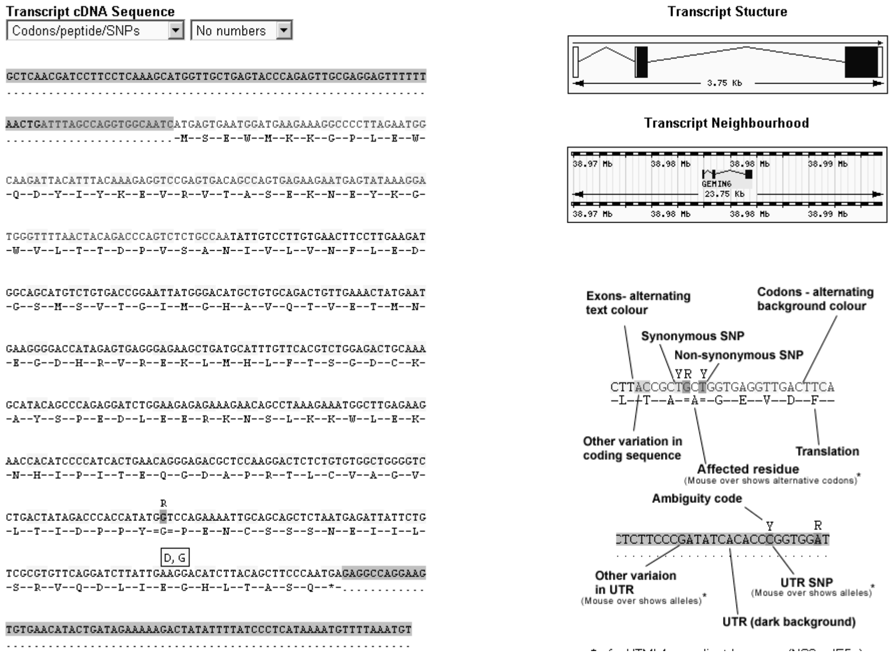
▶ filter
- Chromosome: 2
- From base: 37700000
- To base: 39700000
- Non-synonymous SNPs Only

64 Entries pass Filters

▶ output
- SNP List



```
+-----------------+---------------------+-----------------------------------+------------------------------
| Chromosome Name | Start Position (bp) | Reference ID                      | Peptide Shift in ensembl gene
+-----------------+---------------------+-----------------------------------+------------------------------
| 2               | 37785151            | rs2231503                         | Q/H
| 2               | 37955995            | rs4670779                         | A/V
| 2               | 37956075            | rs12478227                        | R/C
| 2               | 37956481            | rs4670218                         | S/C
| 2               | 38090785            | rs4670800                         |
| 2               | 38090785            | rs4670800                         |
| 2               | 38090785            | rs4670800                         | G/D
| 2               | 38209790            | rs1800440                         |
| 2               | 38209790            | rs1800440                         |
| 2               | 38209790            | rs1800440                         |
| 2               | 38209790            | rs1800440                         | N/S
| 2               | 38209820            | rs4986888                         |
| 2               | 38209820            | rs4986888                         |
| 2               | 38209820            | rs4986888                         |
| 2               | 38209820            | rs4986888                         | A/G
| 2               | 38209827            | rs4986887                         |
| 2               | 38209827            | rs4986887                         |
| 2               | 38209827            | rs4986887                         |
| 2               | 38209827            | rs4986887                         | D/H
| 2               | 38209854            | rs1056836                         |
| 2               | 38209854            | rs1056836                         |
| 2               | 38209854            | rs1056836                         |
| 2               | 38209854            | rs1056836                         | V/L
| 2               | 38210034            | rs4398252                         |
| 2               | 38210034            | rs4398252                         |
| 2               | 38210034            | rs4398252                         |
```



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

Transcript Neighbourhood

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

Selecting SNPs from a list of candidate genes

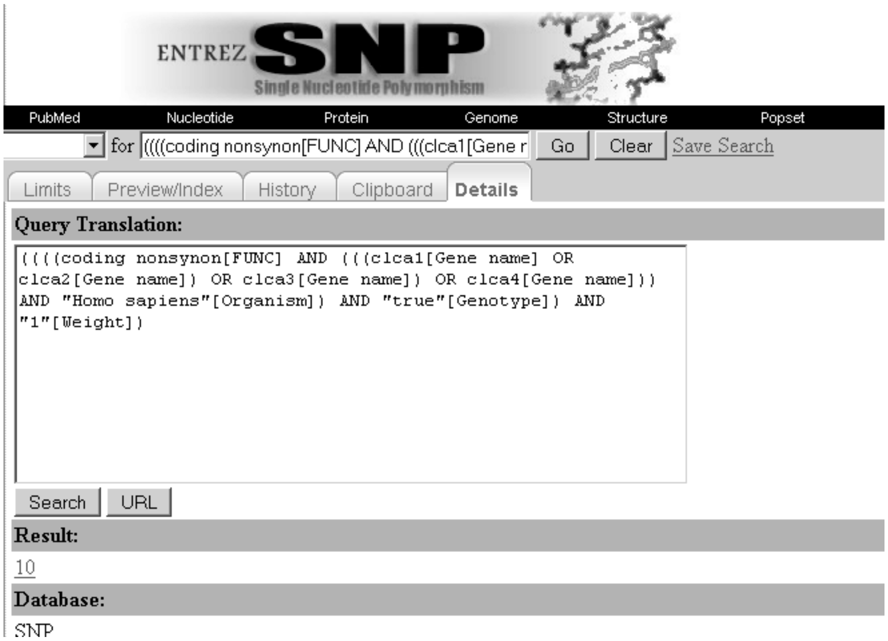
- Use the Entrez SNP query:

  coding nonsynon[FUNC] AND

  CLCA*[Gene name] AND

  human[orgn]

- Download dbSNP database and cross reference with candidate gene list coordinates


ENTREZ SNP
Single Nucleotide Polymorphism
PubMed Nucleotide Protein Genome Structure Popset
for ((((coding nonsynon[FUNC] AND (((clca1[Gene r
Go
Clear
Save Search
Limits Preview/Index History Clipboard Details

Query Translation:

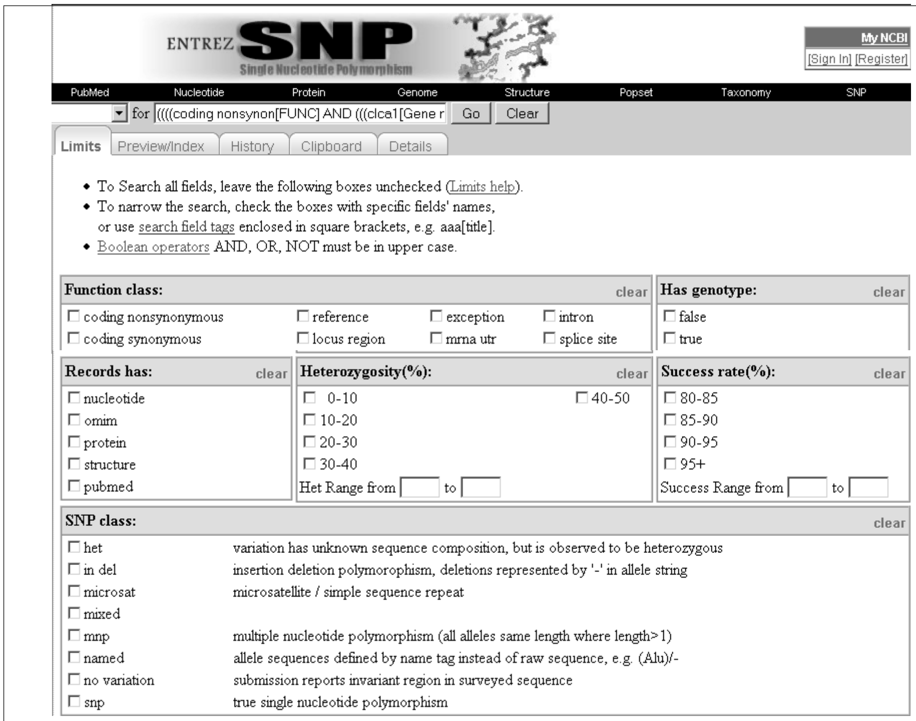
```
((((coding nonsynon[FUNC] AND (((clca1[Gene name] OR
clca2[Gene name]) OR clca3[Gene name]) OR clca4[Gene name]))
AND "Homo sapiens"[Organism]) AND "true"[Genotype]) AND
"1"[Weight])
```

Result:

10

Database:

SNP


ENTREZ SNP
Single Nucleotide Polymorphism
My NCBI
[Sign In] [Register]
PubMed Nucleotide Protein Genome Structure Popset Taxonomy SNP
for ((((coding nonsynon[FUNC] AND (((clca1[Gene r
Go
Clear
Limits Preview/Index History Clipboard Details

- To Search all fields, leave the following boxes unchecked (Limits help).
- To narrow the search, check the boxes with specific fields' names, or use search field tags enclosed in square brackets, e.g. aaa[title].
- Boolean operators AND, OR, NOT must be in upper case.

Function class: clear

□ coding nonsynonymous
□ coding synonymous
□ reference
□ locus region
□ exception
□ mrna utr
□ intron
□ splice site

Has genotype: clear

□ false
□ true

Records has: clear

□ nucleotide
□ omim
□ protein
□ structure
□ pubmed

Heterozygosity(%): clear

□ 0-10
□ 10-20
□ 20-30
□ 30-40
□ 40-50

Het Range from ____ to ____

Success rate(%): clear

□ 80-85
□ 85-90
□ 90-95
□ 95+

Success Range from ____ to ____

SNP class: clear

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

Overview of Topics

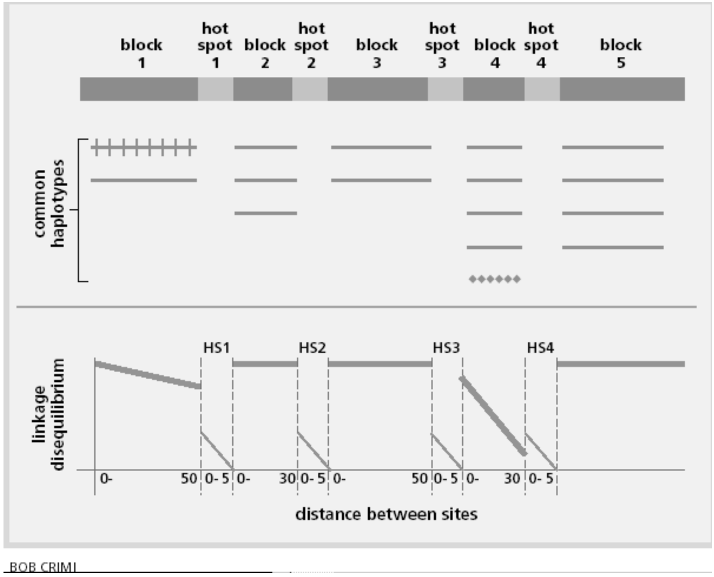
- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project

Haplotype Map project

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

What is a Haplotype?

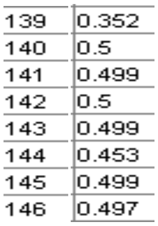
- A set of closely linked genetic markers present on one chromosome which tend to be inherited together (not easily separable by recombination).
- Recombination occurs between homologous chromosomes when cells divide.
- It is believed that recombination is not equally likely across the genome, but that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



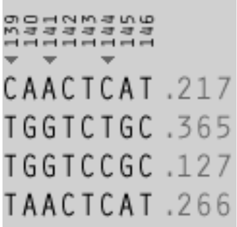
From: Goldstein DB. Islands of linkage disequilibrium. Nat Genet. 2001 Oct;29(2):109-11.

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

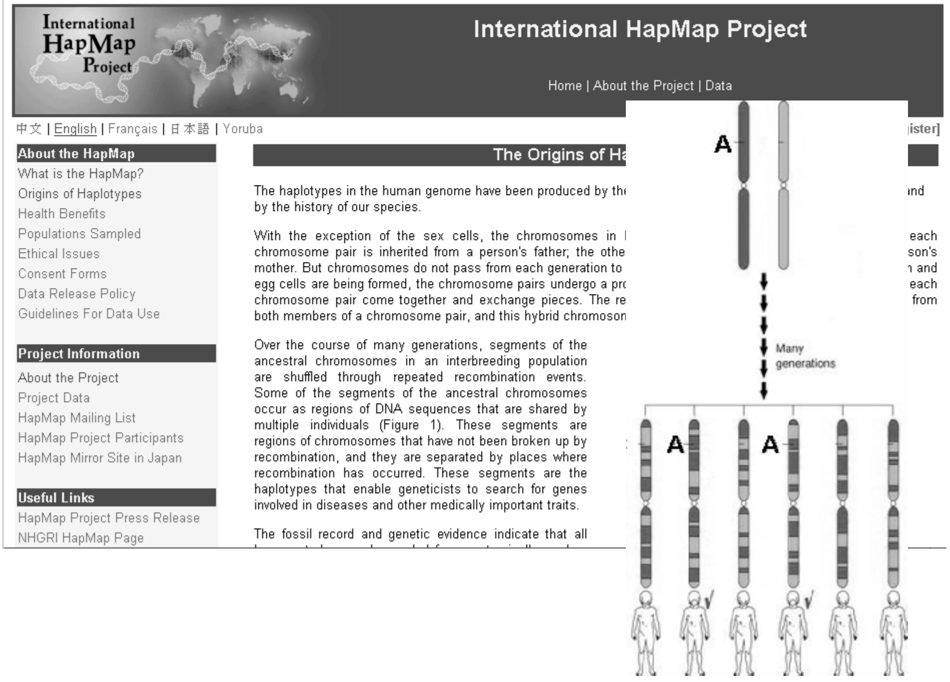


```
139 140 141 142 143 144 145 146
CAACTCAT .217
TGGTCTGC .365
TGGTCCGC .127
TAACTCAT .266
```

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

www.hapmap.org



International HapMap Project

The Origins of Ha

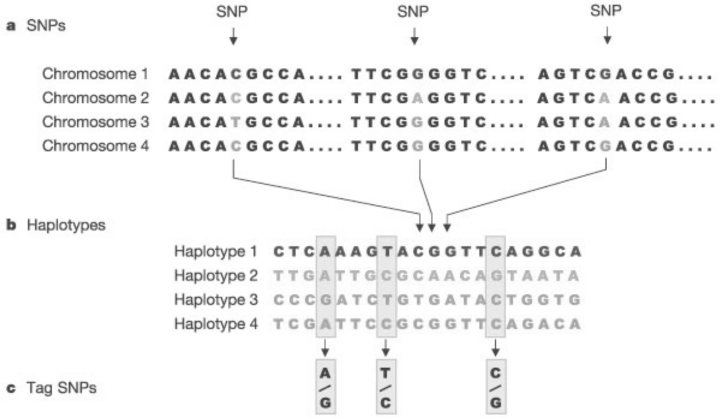
The haplotypes in the human genome have been produced by the and by the history of our species.

With the exception of the sex cells, the chromosomes in each chromosome pair is inherited from a person's father; the othe son's mother. But chromosomes do not pass from each generation to n and egg cells are being formed, the chromosome pairs undergo a pro each chromosome pair come together and exchange pieces. The re from both members of a chromosome pair, and this hybrid chromoson

Over the course of many generations, segments of the ancestral chromosomes in an interbreeding population are shuffled through repeated recombination events. Some of the segments of the ancestral chromosomes occur as regions of DNA sequences that are shared by multiple individuals (Figure 1). These segments are regions of chromosomes that have not been broken up by recombination, and they are separated by places where recombination has occurred. These segments are the haplotypes that enable geneticists to search for genes involved in diseases and other medically important traits.

The fossil record and genetic evidence indicate that all

Identification of Haplotypes Through Genotyping



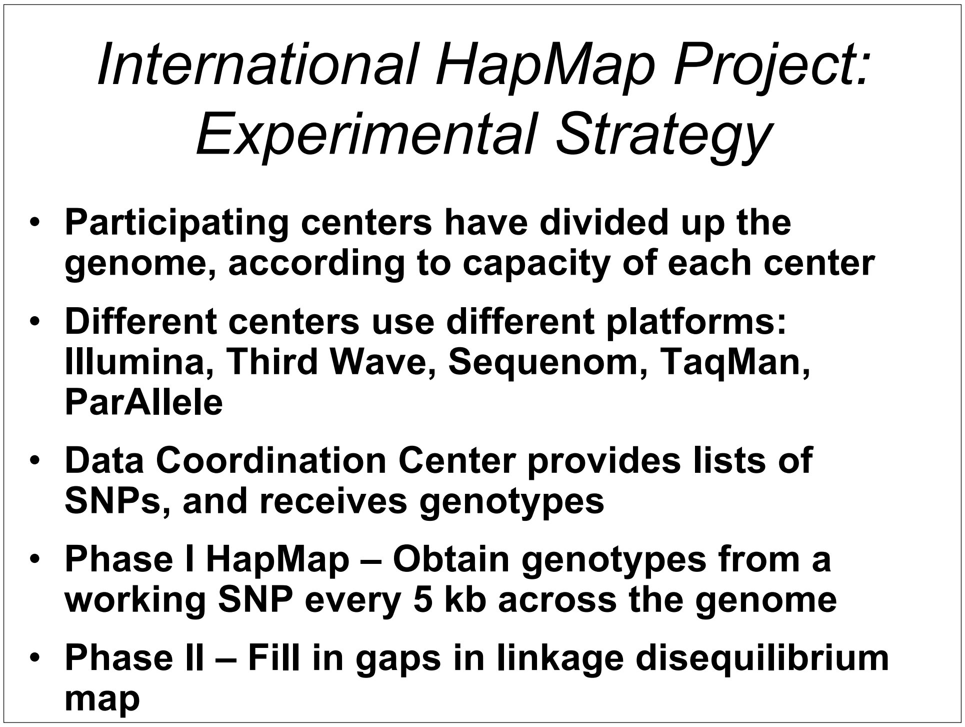
a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCAACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCAACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

International HapMap Project

- **Goal is to develop a haplotype map covering 80 - 90% of the genome**
- **The map should be usable in all populations**
- **Three year project started October 2002**
- **International collaboration, involving Canada, China, Nigeria, Japan, the United Kingdom, and the United States**
- **All data publicly accessible at www.hapmap.org**

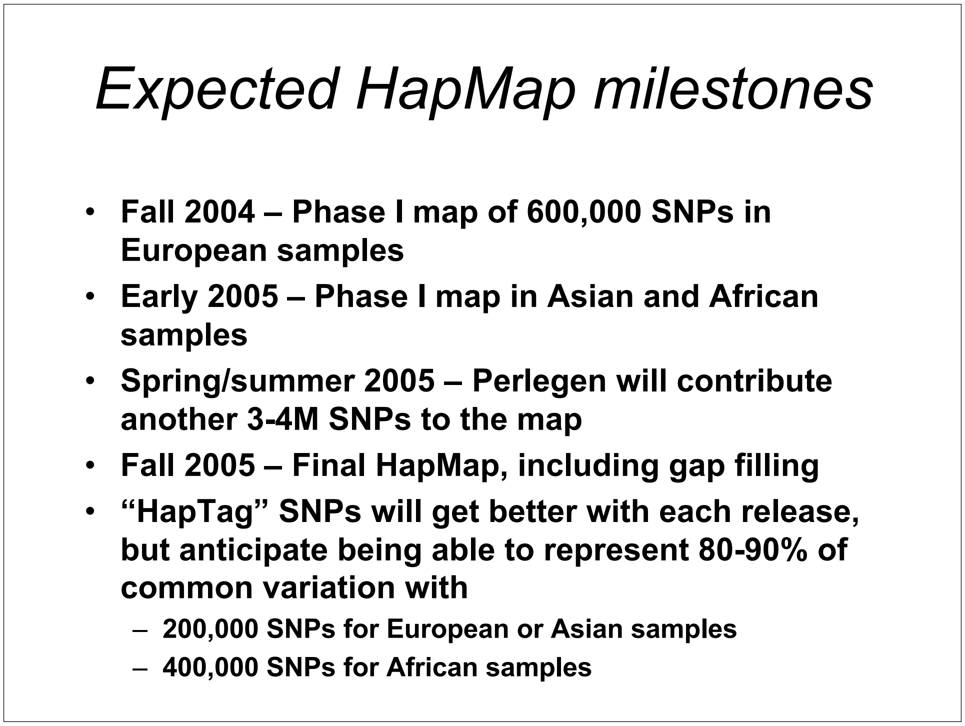
International HapMap Project: Sample Collection

- **Similarity in haplotypes worldwide limits the need to collect samples from many populations**
- **No clinical information collected, samples anonymous**
- **Individual consent and extensive community consultation**
- **270 samples collected and genotyped**
  - **Africa (Yoruba in Ibadan, Nigeria)**
  - **Asia (Japanese in Tokyo, Han Chinese in Beijing)**
  - **Europe (CEPH family samples, Utah)**
- **Samples are available as DNA or cell lines from Coriell**
- **Additional populations being studied in a pilot phase**



International HapMap Project: Experimental Strategy

- **Participating centers have divided up the genome, according to capacity of each center**
- **Different centers use different platforms: Illumina, Third Wave, Sequenom, TaqMan, ParAllele**
- **Data Coordination Center provides lists of SNPs, and receives genotypes**
- **Phase I HapMap – Obtain genotypes from a working SNP every 5 kb across the genome**
- **Phase II – Fill in gaps in linkage disequilibrium map**

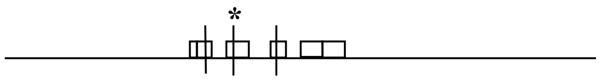


Expected HapMap milestones

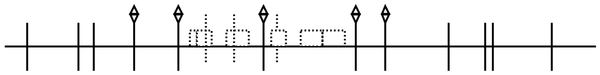
- **Fall 2004 – Phase I map of 600,000 SNPs in European samples**
- **Early 2005 – Phase I map in Asian and African samples**
- **Spring/summer 2005 – Perlegen will contribute another 3-4M SNPs to the map**
- **Fall 2005 – Final HapMap, including gap filling**
- **“HapTag” SNPs will get better with each release, but anticipate being able to represent 80-90% of common variation with**
  - **200,000 SNPs for European or Asian samples**
  - **400,000 SNPs for African samples**

Association Studies

Direct



Indirect



Genotype only the most informative SNPs

500 cases one pool

500 controls one pool

~~10,000~~ SNPs

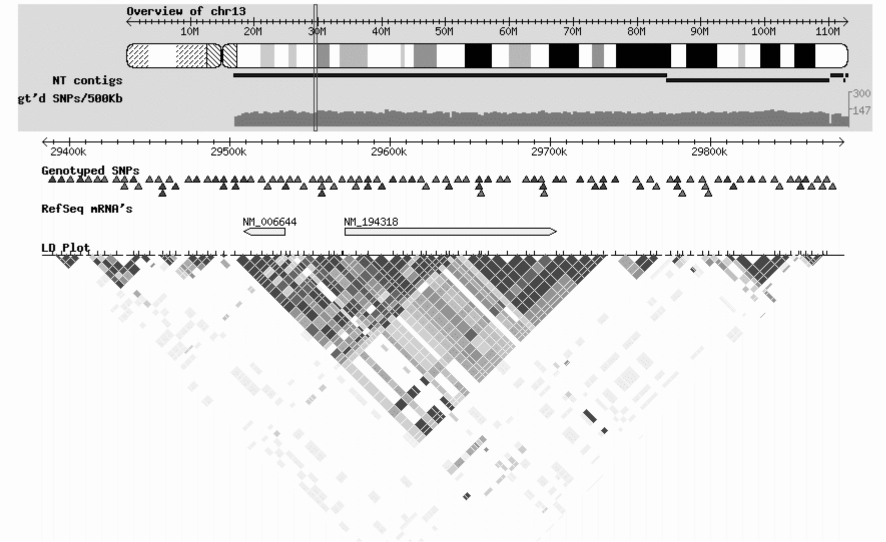
1,000 ‘haplotype tag’ SNPs

Direct analysis: 10,000,000 genotypes

Pooled DNA analysis: 20,000 genotypes

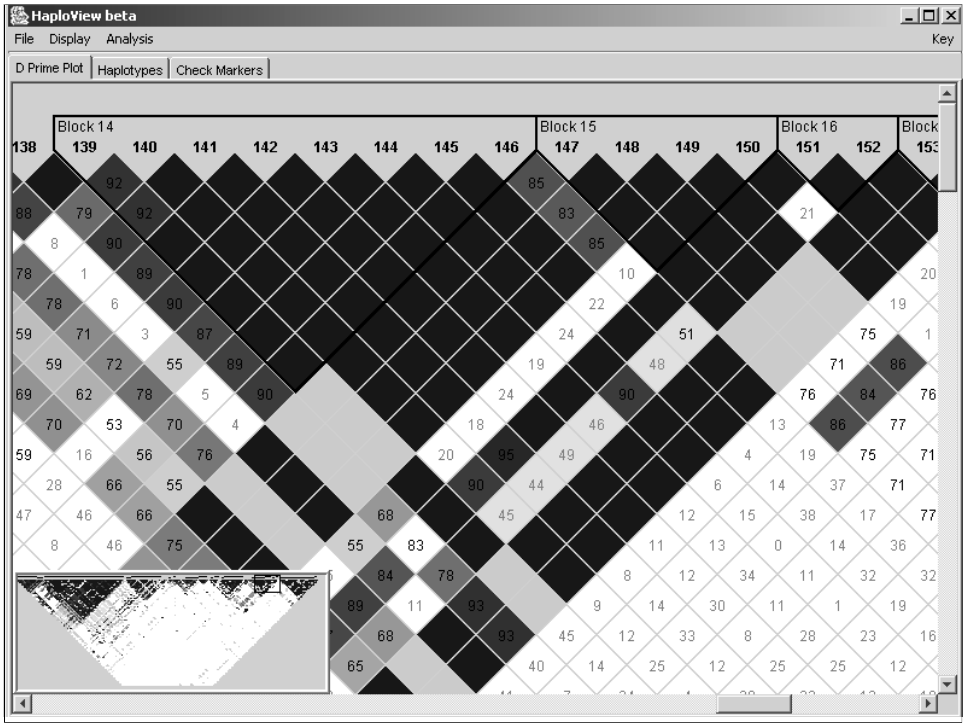
Selected SNPs: 2,000 genotypes

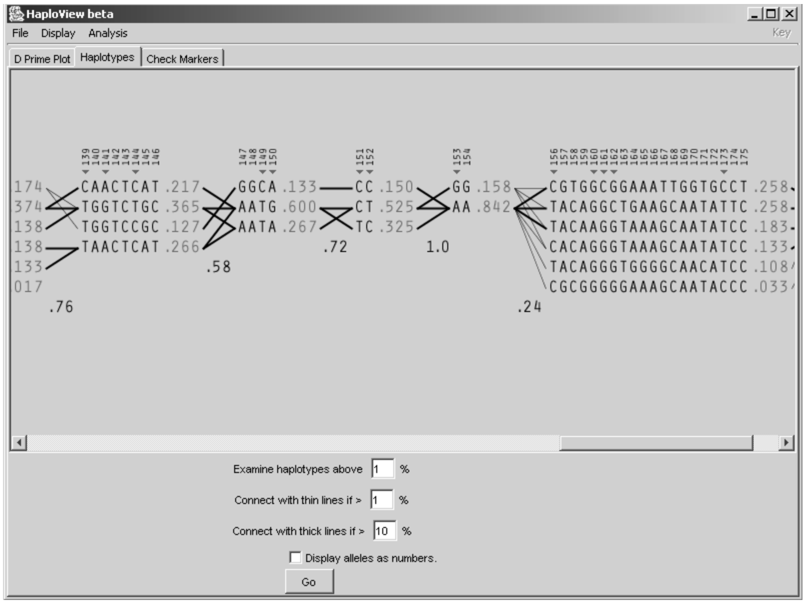
LD Display



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.





```
       139-146             147-150          151-152       153-154        156-175
.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                        .72           1.0            CACAGGGTAAAGCAATATCC .133
.133                       .58                                           TACAGGGTGGGGCAACATCC .108
.017                                                                     CGCGGGGGAAAGCAATACCC .033
  .76                                                                .24
```

Examine haplotypes above 1 %

Connect with thin lines if > 1 %

Connect with thick lines if > 10 %

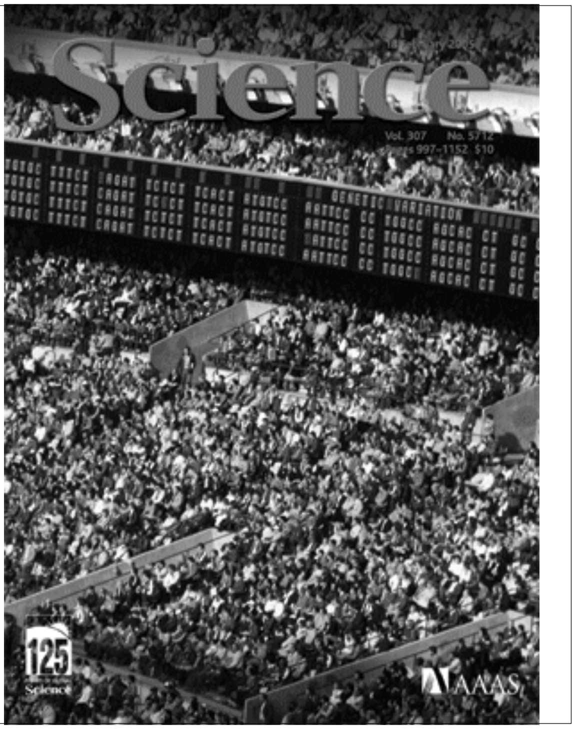
☐ Display alleles as numbers.

Perlegen Biosciences:

Whole-Genome Patterns of Common DNA Variation in Three Human Populations

Hinds, et al.

February 14th, 2005



Perlegen's genome browser

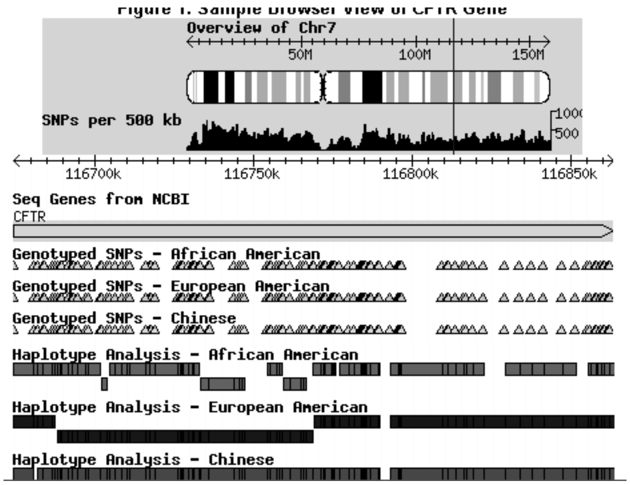


Figure 1. Sample Browser View of CFTR Gene
Overview of Chr7
50M
100M
150M
SNPs per 500 kb
1000
500
116700k
116750k
116800k
116850k
Seq Genes from NCBI
CFTR
Genotyped SNPs - African American
Genotyped SNPs - European American
Genotyped SNPs - Chinese
Haplotype Analysis - African American
Haplotype Analysis - European American
Haplotype Analysis - Chinese

http://genome.perlegen.com/browser/index.html

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.

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.

References

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

The International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789-96.

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

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. Whole-genome patterns of common DNA variation in three human populations. Science. 2005 Feb 18;307(5712):1072-9.

Crawford DC, Nickerson DA, Definition and clinical importance of haplotypes. Annu Rev Med. 2005;56:303-20.

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://genome.perlegen.com/browser/index.html: Perlegen’s HapMap