

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

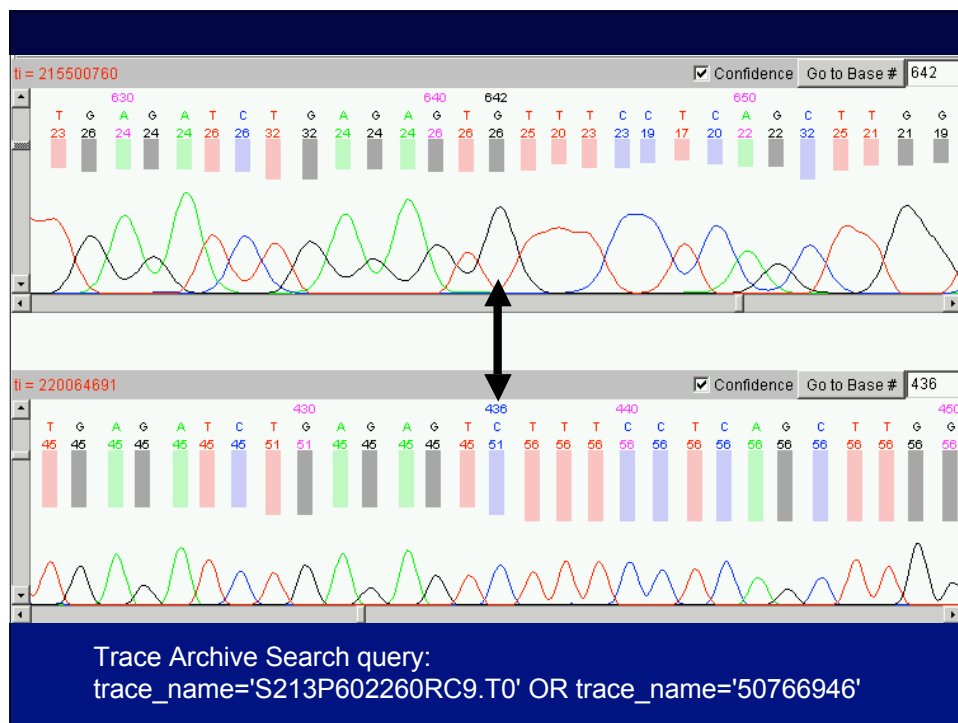
- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project

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.

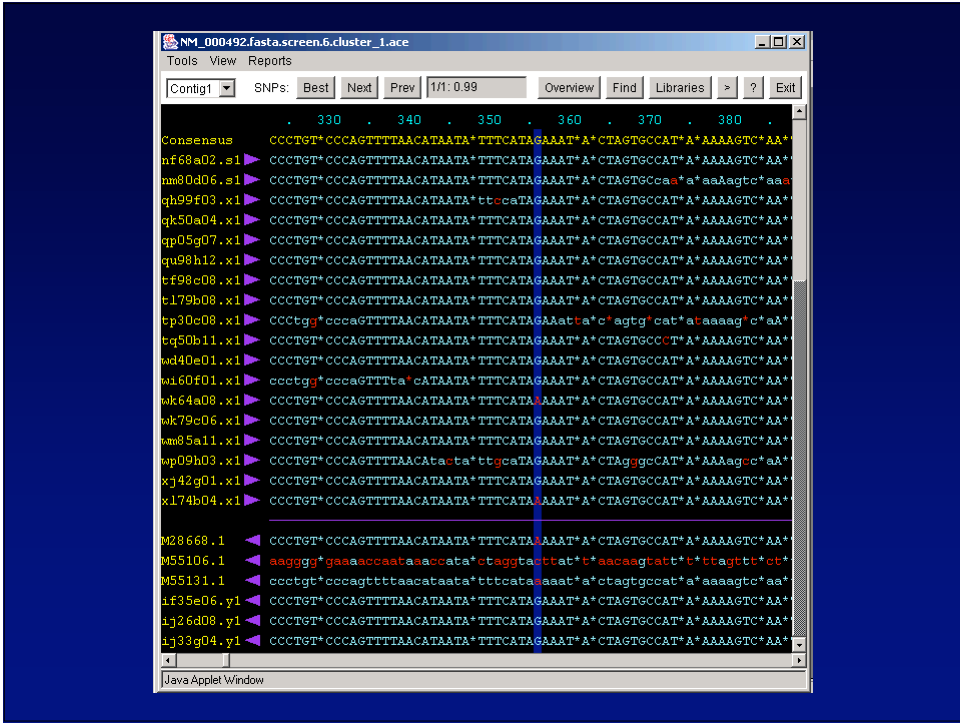


Mining SNPs from sequence

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

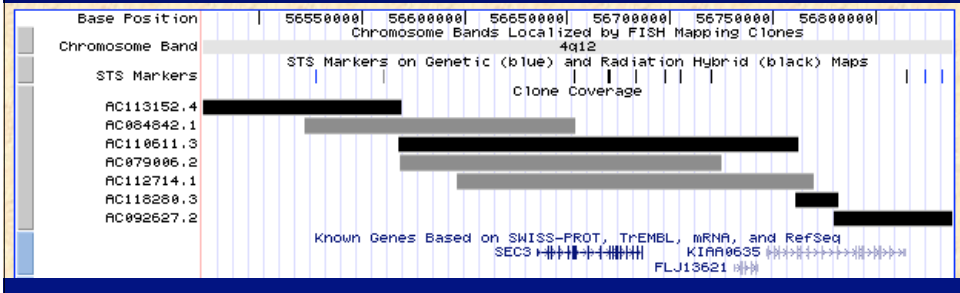
Expressed Sequence Tag Mining

- 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).



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.

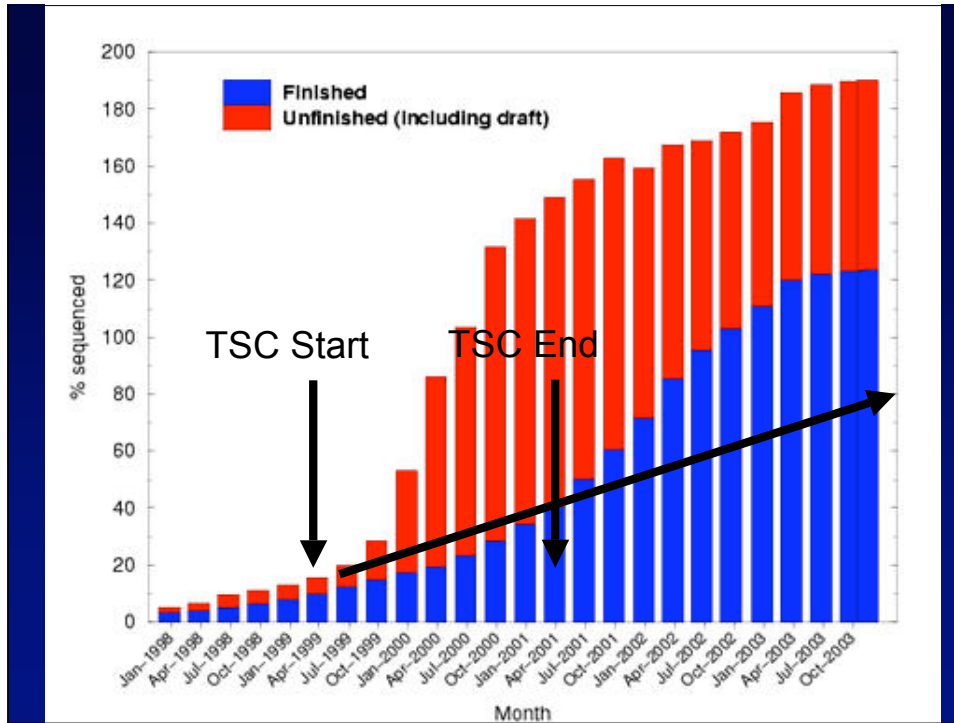


Clone Overlap

- 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.



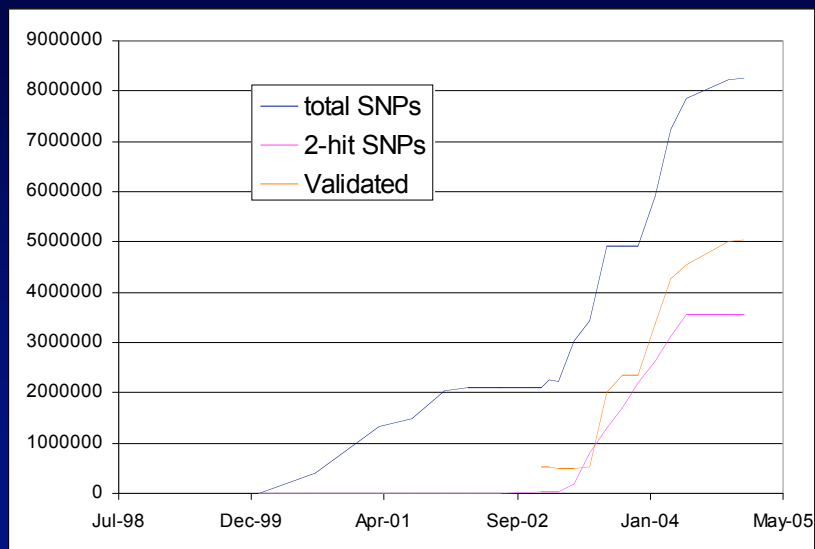
The SNP Consortium

- 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.

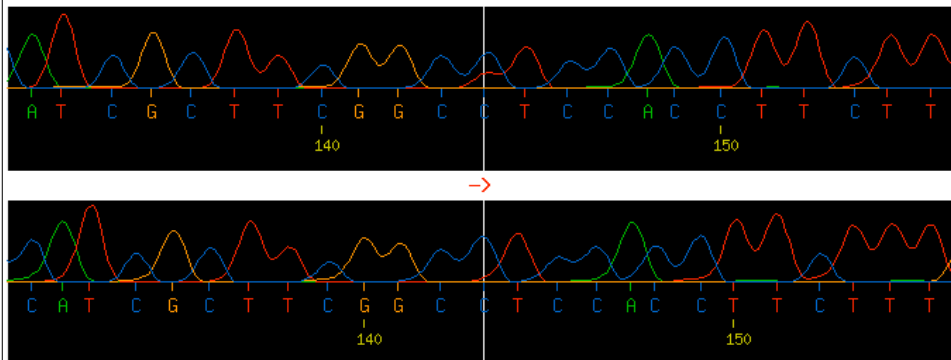
SNPs in dbSNP



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 individual's two haploids, so the special case of heterozygous base positions must be properly processed.

IMS-JST096911



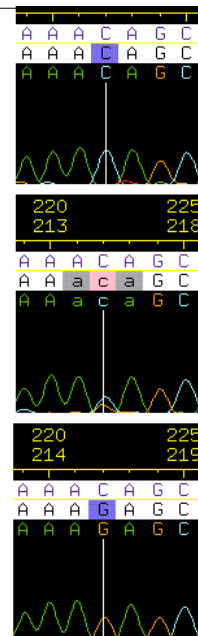
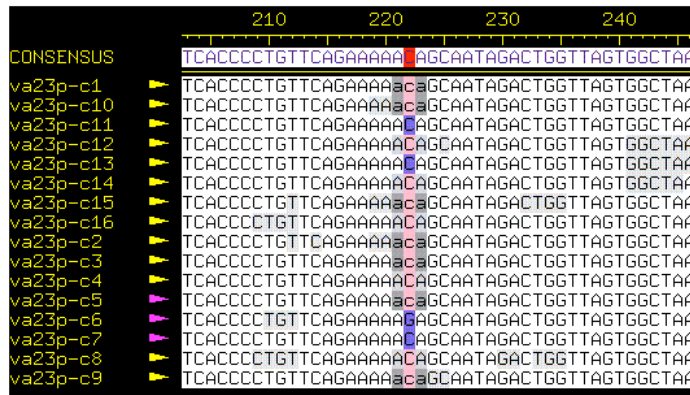
<http://snp.ims.u-tokyo.ac.jp/>

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.

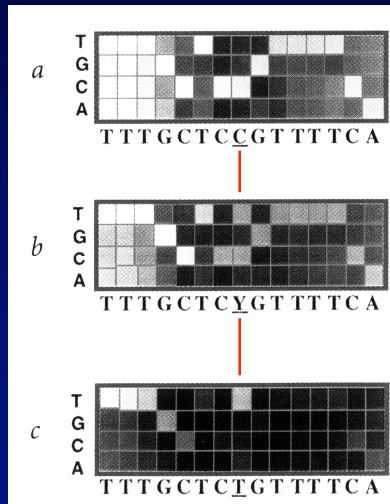
SNP detection by PolyPhred. View of a Consed window with a tag (red=highest ranking SNP tag) marking the consensus position of the SNP in the traces and genotype tags marking each of the samples below (purple=homozygote, pink=heterozygote). On the right trace windows for alternate homozygotes (C/C (top)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



PolyPhred example from their web site.

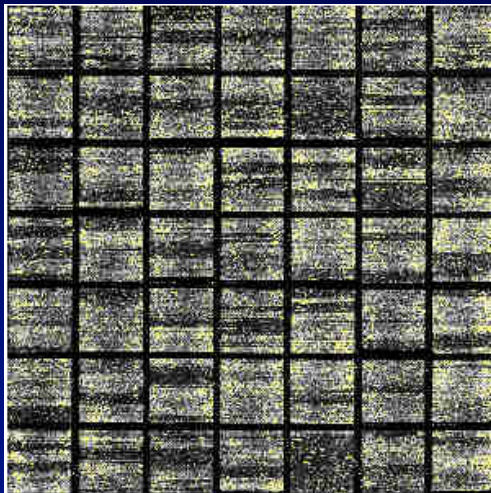
Sequencing Chips

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



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

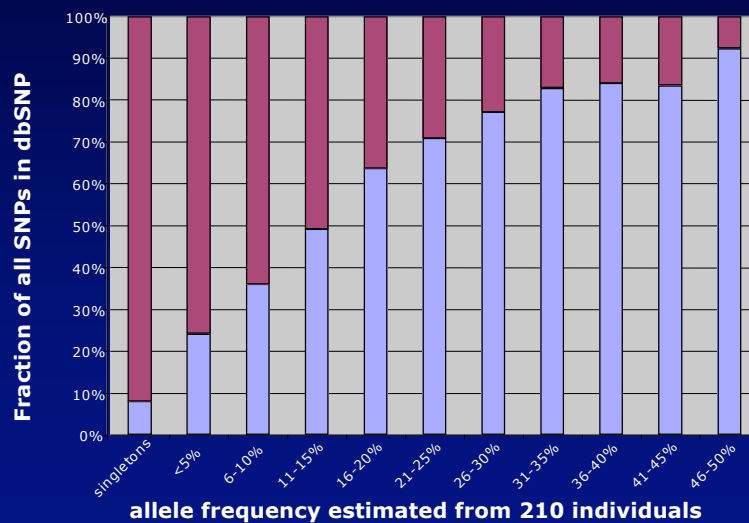
Distribution properties

- 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	Allele
Organism: human (<i>Homo sapiens</i>)	Variation Class: SNP: single nucleotide polymorphism
Molecule Type: Genomic	Alleles: C/G
Created in build: 86	Ancestral Allele: G
Last updated in build: 123	

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		rewT	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss2423851	HGBASE SNP000010888		rewT	C/G	accatgaggtgcatatctatgaaaa	agcggtgccaa
ss2733260	TSC-CSHL TSC0848041		fwd B	C/G	ctcgtgcaccttggtccatbbggcaccgct	tbbcatagat
ss4391917	LEE 151903		rewT	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss4407741	LEE 151902		rewT	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss5815409	8C_JCMINT_007933.10_24217856		rewT	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss14546249	WUGSC_SSAHASNP chr7.NT_007933.13_24217938		rewT	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss16262424	CGAP-GAJ 1525080		rewT	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaa
ss23476794	PERLEGEN afd0546573		rewT	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 AGTGTGACAG CCCTCCATGC TATCagcaaa catgctggag ggcaaaagcca
agaggcagaa aagatggggt cttggctcatg tggagctgct ggatcaagcc tctcctgaag
ccctcaaccc tgtgagtttt tggtaacatg agccaacaca gtccccttaa aattgaagcc
agtttgaatc cgggtttcAC GGTGAGTGGG CAGATGCTCC ACAATCAGTG GCCATGCCCT
GCCTTGCAAC ACCCCCCAA CCCACCCT CTTTCAGGA CGGTGGTCCC AGCCACCCTG
ACATACCTGT CACCTGCCCC TTGTCTCCT TGAGCTCGTG CACCTTGGTC CATTTGGCAC
CGCT
S
TTTTATAGA TATGCACCTC ATGGTTGTTG GGGCAGATGG CAATCTCTGA AGGGGAGATG
GAGGGAGAIT GAGGGGCCCT CTCCATGACT GCCCTCTGCC AGGACACACT ACACAGTGCA
CCTAGGCAAC AACACCTCAC CTTTCATGAC TCAGTCTCTC CTCTTCTGCC TTGAGGGGG
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 [\[all\]](#) [\[cSNP\]](#) [\[has frequency\]](#) [\[double hit\]](#) [\[haplotype tagged\]](#) variations associated with this gene.

Gene Model (contig mRNA transcript) [NT_007933->NM_005720](#); [\[Sequence Viewer\]](#)



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_TTAG	HSC_TTAG
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

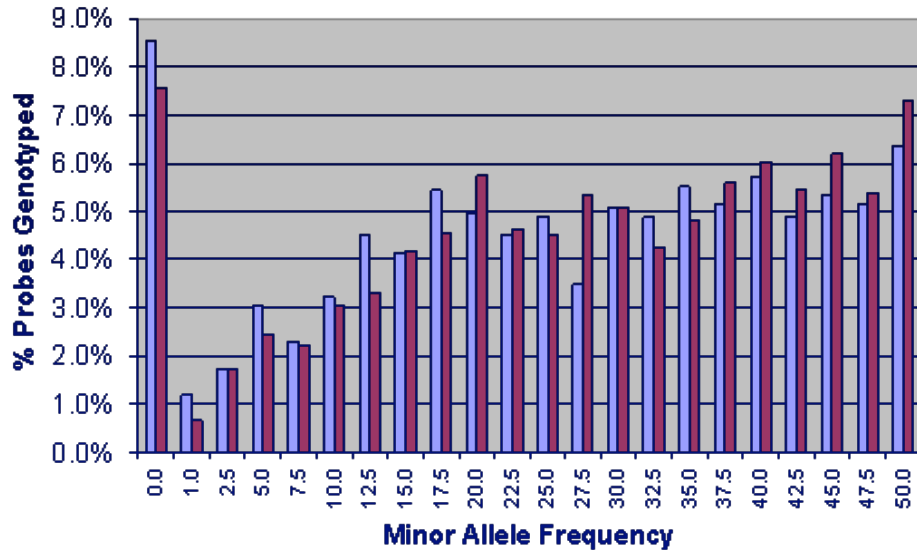
Validation Summary:

Validation status:   	<i>DoubleHit found by:</i> 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
	validated by HapMap project

Double hit SNP minor allele frequency characteristics



Credit: Dr. Paul Hardenbol, Parallele Bioscience

Genotype Detail

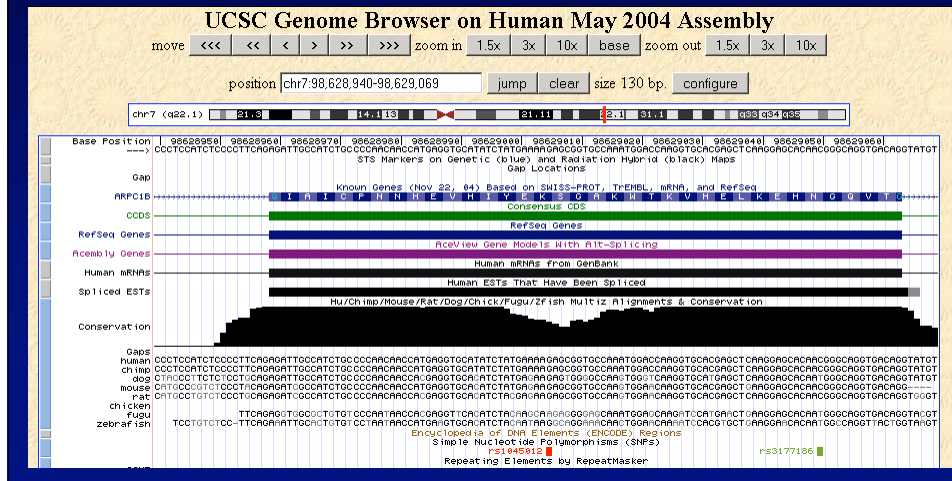
SNP Detail ▲						
rs1045012 ▲						
Assembly	Chromosome	Start	Gene	SNP Type	Orientation	Genotype Freq
35:reference	7	98629005	10095	2	rev	C/G 0.107
35:HSC_TCAG	7	97974083	10095	2	rev	C/C 0.893
35:Celera	7	94177385	10095	2	rev	
ss14546249 Submitter's Id chr7.NT_007933.13_24217938 Orientation to rs rev						
Handle-Population Id		2n	Allele Freq	Genotype Freq	Hardy-Weinberg	
CSHL-HAPMAP-HapMap-CEU		120	C 0.042	C/G 0.083	Chi Square 0.112	
			G 0.958	G/G 0.917		
ss23476794 Submitter's Id af30546573 Orientation to rs rev						
Handle-Population Id		2n	Allele Freq	Genotype Freq	Hardy-Weinberg	
PERLEGEN-AFD EUR PANEL		48	C 0.042	C/G 0.083	Chi Square 0.045	
			G 0.958	G/G 0.917		
PERLEGEN-AFD AFR PANEL		46	C 0.13	C/G 0.261	Chi Square 0.518	
			G 0.87	G/G 0.739		
PERLEGEN-AFD CHN PANEL		48	C 0.021	C/G 0.042	Chi Square 0.011	
			G 0.979	G/G 0.958		

Viewing SNPs in Browsers

NCBI

Ensembl

UCSC



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

The screenshot shows the NCBI Entrez SNP database search results for the query 'clca1'. The search returned 308 items. The results are displayed in a table with columns for SNP ID, organism, and sequence. The first three results are:

SNP ID	Organism	Sequence
rs3820042	Homo sapiens	AACACCCAACTCAGCTGCTCTCTGT[C/G]TCCTCTTTAGGATATGTGGCAACAT
rs3765994	Homo sapiens	ATATTTTCATTGGAGATGGAGAAAAG[A/G]TNANGAAATTGAGATATAGTGAANT
rs3765989	Homo sapiens	TAGACACCATATATTGCCTTGGCAG[A/T]AAGGGTGATTAGTAGTATTTCTTTC

The URL for the search results is <http://www.ncbi.nlm.nih.gov/SNP/index.html>.


Graphic Summary :

-  MapView Mapped to chromosome shown with map weight 1 (single green bar), linkout to MapViewer
-  MapView Mapped to chromosome shown with map weight greater than 1 (two or more green bar)
-  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
-  Validated
-  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:

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	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

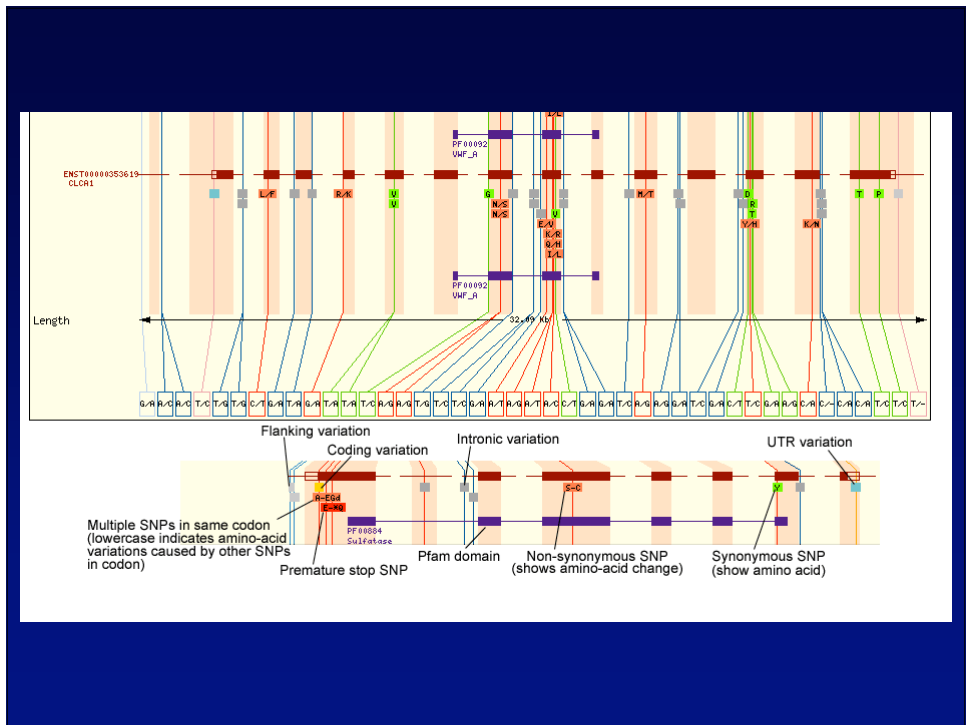
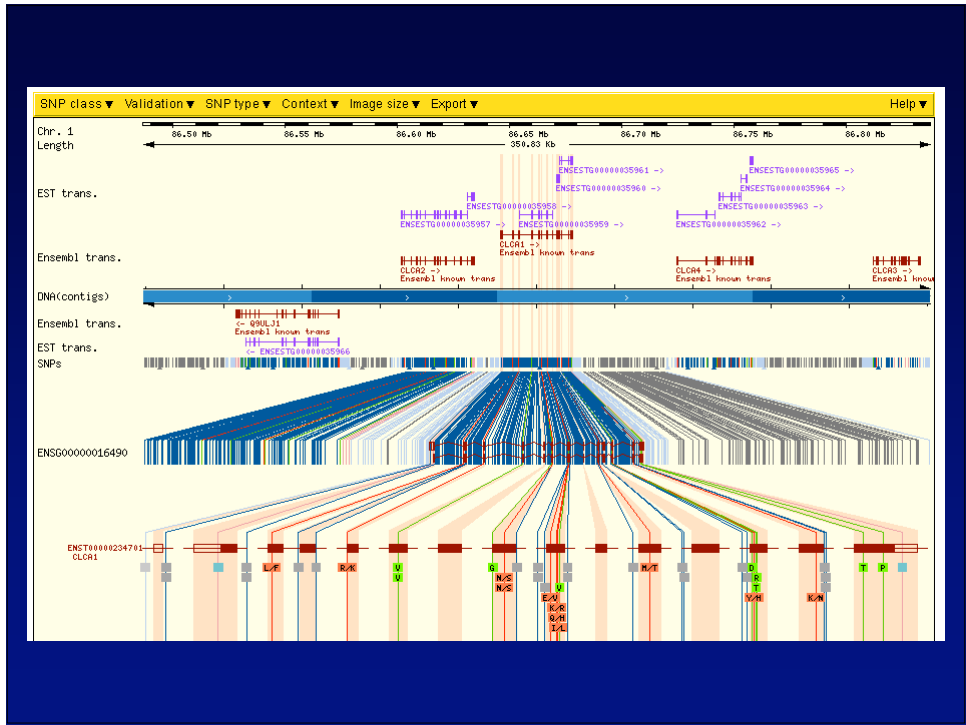
gene model	Contig	mrna	protein	mrna orientation	transcript	snp count
(contig mRNA transcript):	NT_032977	NM_001285	NP_001276	forward	plus strand	18, coding

Contig position	dbSNP rs# cluster id	Heterozygosity	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 [M]	3	215
		0.486			contig reference	A	Val [M]	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 (HGUG ID) (to view all Ensembl genes linked to the name click here) Member of Human CCDS set
Ensembl Gene ID	ENS0000016490
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: nplp1de (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	<p>1: CLCA1 (ENS00000234701) (Transcript information) (Exon information) (Protein information)</p> <p>2: CLCA1 (ENS00000353619) (Transcript information) (Exon information) (Protein information)</p>

http://www.ensembl.org/Homo_sapiens



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	-	-
rs2753396	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)
rs4547852	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	Allele
Organism: human (<i>Homo sapiens</i>)	Variation Class: SNP: single nucleotide polymorphism
Molecule Type: cDNA	Alleles: A/T
Created in build: 86	Ancestral Allele: Not available
Last updated in build: 108	

SNP Details are categorized in the following sections:

[Submission](#) [Fasta](#) [Resource](#) [GeneView](#) [Map](#) [Variation](#) [Validation](#)

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 **rs1142185**

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	ttagggaacaattatccaactgatggatctg	aattgtgctgctgacggatggggaagacaa
ss4435881	LEE 1404930		fwd/B	A/T	tagggaacgaattatccaactgatggatctg	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 TTATCCAAC TATGGATCTG
T
AATTGTGCTG CTGACGGATG GGGGAAGACAA CACTATAAGT GGGTGCTTTA
```

[Home](#) | [Ensembl](#) | [TextSearch](#) | [BlastSearch](#) | [MartSearch](#) | [Download](#)

[new](#) | **START** | [FILTER](#) | [OUTPUT](#) | [export](#)

[refresh](#) | [Online Help](#) | [Help Desk](#)

START [new](#) [next](#)

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

Select the **dataset** for this query
 Focus:

Species:

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

Summary
[start](#) Not yet initialised
[filter](#) Not yet initialised
[output](#) Not yet initialised

FILTER [back](#) [next](#)

Further refine your search or click 'next':

REGION:

Limit to (uncheck for entire genome):
 Chromosome name:

From

To

Limit to ENCODE region
 Type:
 Region:

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

GENERAL SNP FILTERS:

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

SNPs with TSC IDs Only Excluded

SNPs that have been validated Only Excluded

With allele frequency data from population:

Maximum freq of the minor allele:

Minimum freq of the minor allele:

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

Only Excluded

Features
SNPs
Sequences

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

Transcript Structure
 3.75 Kb

Transcript Neighbourhood
 38.97 Mb 38.98 Mb 38.99 Mb 39.00 Mb
 23.75 Kb

Exons - alternating text colour **Codons - alternating background colour**

Synonymous SNP
 Y R Y

Non-synonymous SNP
 CTTACCGTGCCTGGTGGGTTGACTTCA
 -L-T-A-A-A-G-E-V-D-F-

Other variation in coding sequence **Translation**
 Affected residue (Mouse over shows alternative codons)*

Ambiguity code
 Y R

Other variation in UTR **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])
```

Search URL

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. aa[title].
- [Boolean operators](#) AND, OR, NOT must be in upper case.

Function class: clear		Has genotype: clear	
<input type="checkbox"/> coding nonsynonymous	<input type="checkbox"/> reference	<input type="checkbox"/> exception	<input type="checkbox"/> intron
<input type="checkbox"/> coding synonymous	<input type="checkbox"/> locus region	<input type="checkbox"/> mrna utr	<input type="checkbox"/> splice site
<input type="checkbox"/> nucleotide	Heterozygosity(%): clear		Success rate(%): clear
<input type="checkbox"/> omim	<input type="checkbox"/> 0-10	<input type="checkbox"/> 40-50	<input type="checkbox"/> 80-85
<input type="checkbox"/> protein	<input type="checkbox"/> 10-20		<input type="checkbox"/> 85-90
<input type="checkbox"/> structure	<input type="checkbox"/> 20-30		<input type="checkbox"/> 90-95
<input type="checkbox"/> pubmed	<input type="checkbox"/> 30-40		<input type="checkbox"/> 95+
	Het Range from <input type="text"/> to <input type="text"/>		Success Range from <input type="text"/> to <input type="text"/>
SNP class: clear			
<input type="checkbox"/> het	variation has unknown sequence composition, but is observed to be heterozygous		
<input type="checkbox"/> in del	insertion deletion polymorphism, deletions represented by '-' in allele string		
<input type="checkbox"/> microsatt	microsatellite / simple sequence repeat		
<input type="checkbox"/> mixed			
<input type="checkbox"/> mnp	multiple nucleotide polymorphism (all alleles same length where length>1)		
<input type="checkbox"/> named	allele sequences defined by name tag instead of raw sequence, e.g. (Ah)/-		
<input type="checkbox"/> no variation	submission reports invariant region in surveyed sequence		
<input type="checkbox"/> snp	true single nucleotide polymorphism		

Overview of Topics

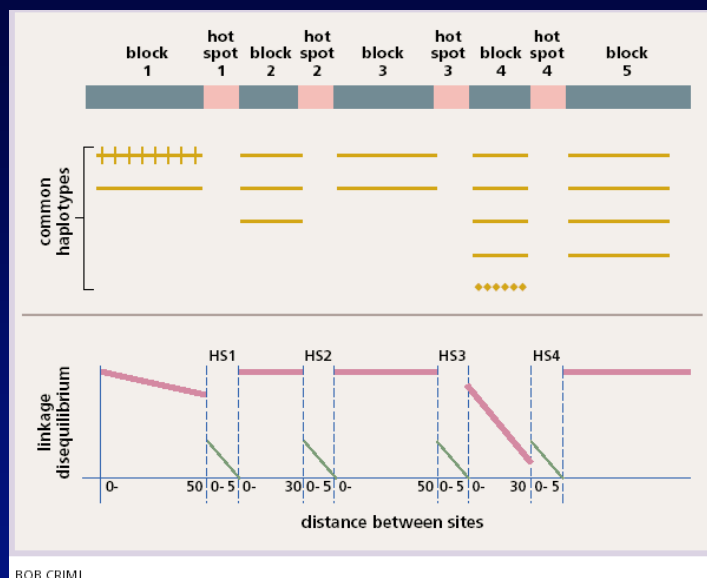
- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project

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.



BOB CRIM

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	0.352	CAACTCAT	.217	$0.352 \times 0.5^7 = 0.00275$
140	0.5	TGGTCTGC	.365	$0.648 \times 0.5^7 = 0.00534$
141	0.499	TGGTCCGC	.127	$0.648 \times 0.5^7 = 0.00534$
142	0.5	TAACTCAT	.266	$0.648 \times 0.5^7 = 0.00534$

0.975

**International
HapMap
Project**



www.hapmap.org



International HapMap Project

International HapMap Project

Home | About the Project | Data

中文 | [English](#) | Français | 日本語 | Yoruba

About the HapMap

- [What is the HapMap?](#)
- [Origins of Haplotypes](#)
- [Health Benefits](#)
- [Populations Sampled](#)
- [Ethical Issues](#)
- [Consent Forms](#)
- [Data Release Policy](#)
- [Guidelines For Data Use](#)

Project Information

- [About the Project](#)
- [Project Data](#)
- [HapMap Mailing List](#)
- [HapMap Project Participants](#)
- [HapMap Mirror Site in Japan](#)

Useful Links

- [HapMap Project Press Release](#)
- [NHGRI HapMap Page](#)

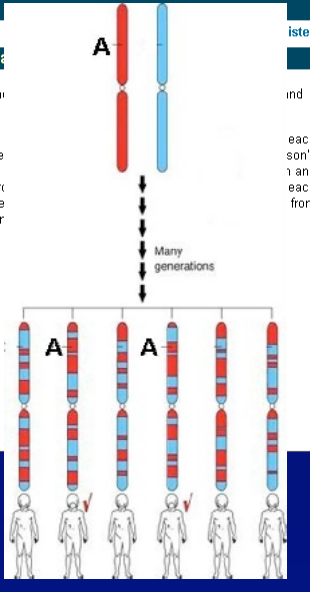
The Origins of Haplotypes

The haplotypes in the human genome have been produced by the recombination of chromosomes over 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 other from the mother. But chromosomes do not pass from each generation to the next unchanged. As chromosomes are being formed, the chromosome pairs undergo a process called recombination. The two members of a chromosome pair come together and exchange pieces. The result is a hybrid chromosome that contains segments from both parents.

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	↓	↓	↓
Chromosome 1	A A C A C G C C A	T T C G G G G T C	A G T C G A C C G
Chromosome 2	A A C A C G C C A	T T C G A G G T C	A G T C A A C C G
Chromosome 3	A A C A T G C C A	T T C G G G G T C	A G T C A A C C G
Chromosome 4	A A C A C G C C A	T T C G G G G T C	A G T C G A C C G

b Haplotypes

Haplotype 1	C	T	C	A	A	A	G	T	A	C	G	G	T	C	A	G	G	C	A
Haplotype 2	T	T	G	A	T	T	G	C	G	C	A	C	A	G	T	A	A	T	A
Haplotype 3	C	C	C	G	A	T	C	T	G	T	G	A	T	A	C	T	G	G	T
Haplotype 4	T	C	G	A	T	T	C	C	G	G	G	T	T	C	A	G	A	C	A

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

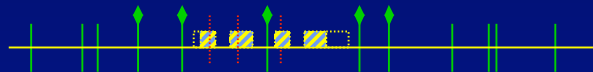
- Fall 2004 – Phase I map of 600,000 SNPs in European samples
- Early 2005 – Phase I map in Asian and African samples
- Spring/summer 2005 – Perlegen will contribute another 3-4M SNPs to the map
- Fall 2005 – Final HapMap, including gap filling
- “HapTag” SNPs will get better with each release, but anticipate 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

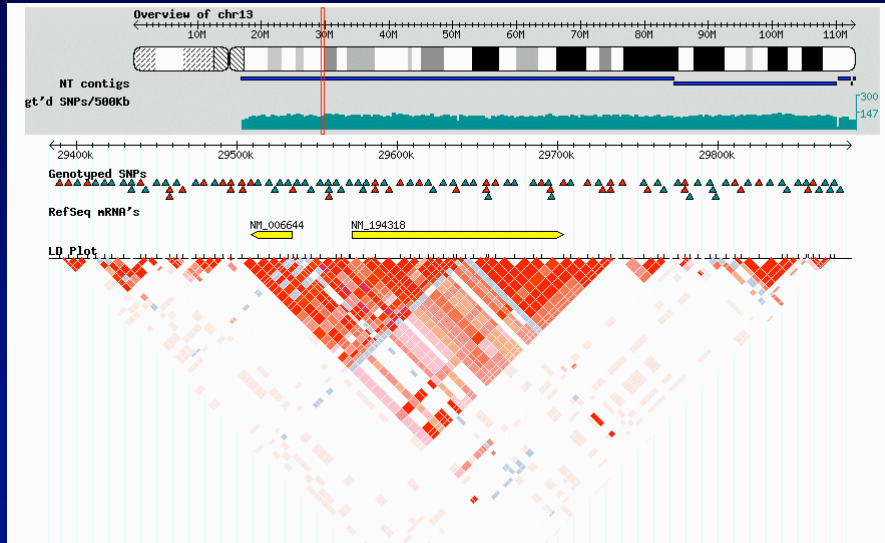
1,000 'haplotype tag' SNPs

Direct analysis: 10,000,000 genotypes

Pooled DNA analysis: 20,000 genotypes

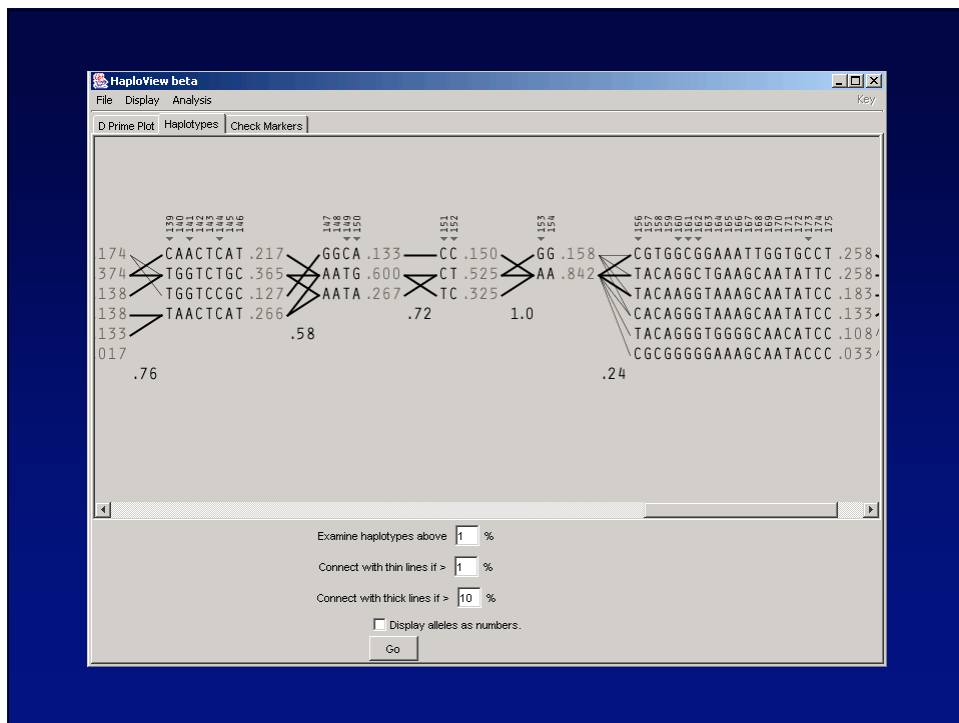
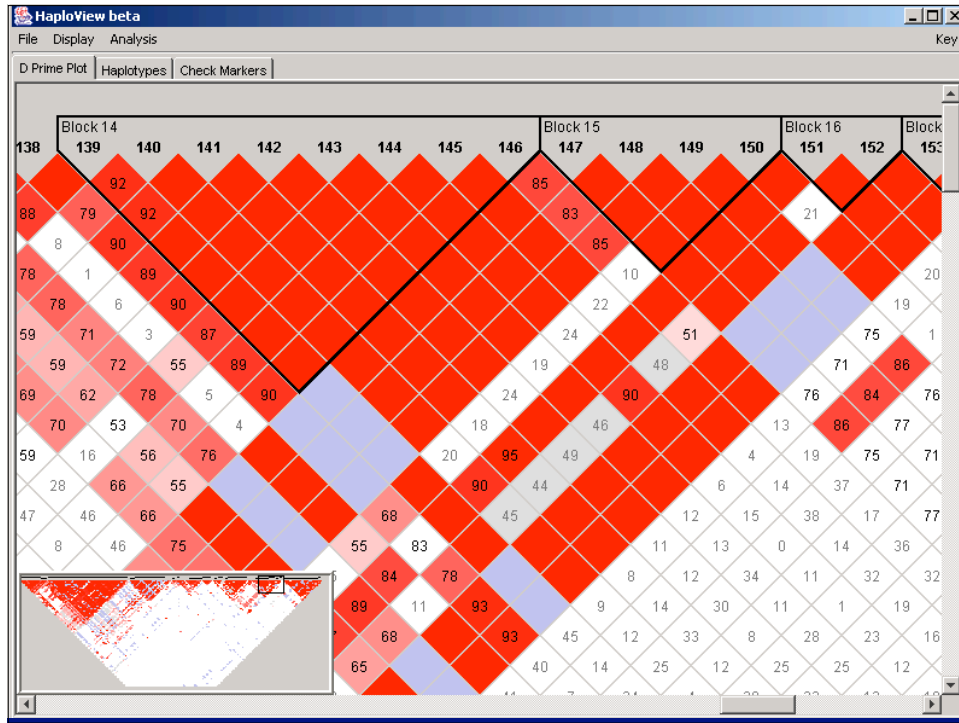
Selected SNPs: 2,000 genotypes

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.



Perlegen
Biosciences:

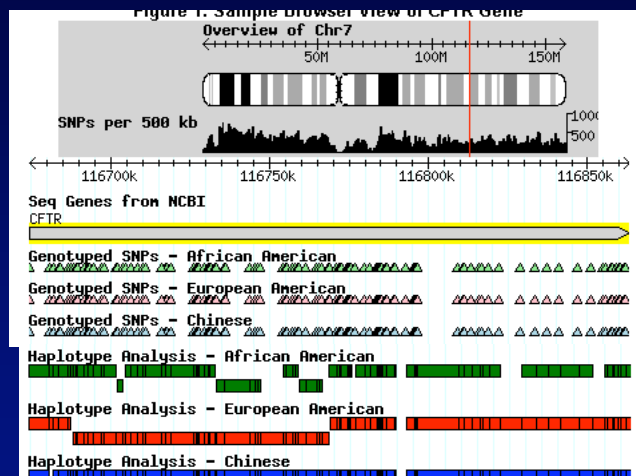
Whole-Genome
Patterns of
Common DNA
Variation in Three
Human Populations

Hinds, et al.

February 14th, 2005



Perlegen's genome browser



<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