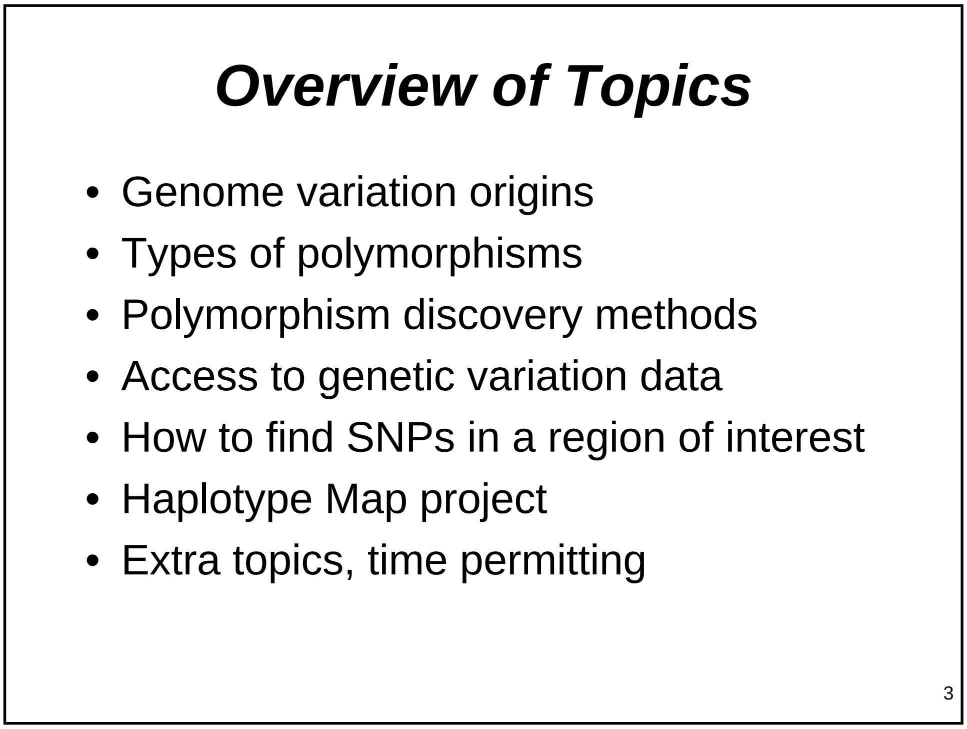
Studying Genetic Variation I: Computational Techniques

Jim Mullikin, PhD
Genome Technology Branch
NHGRI

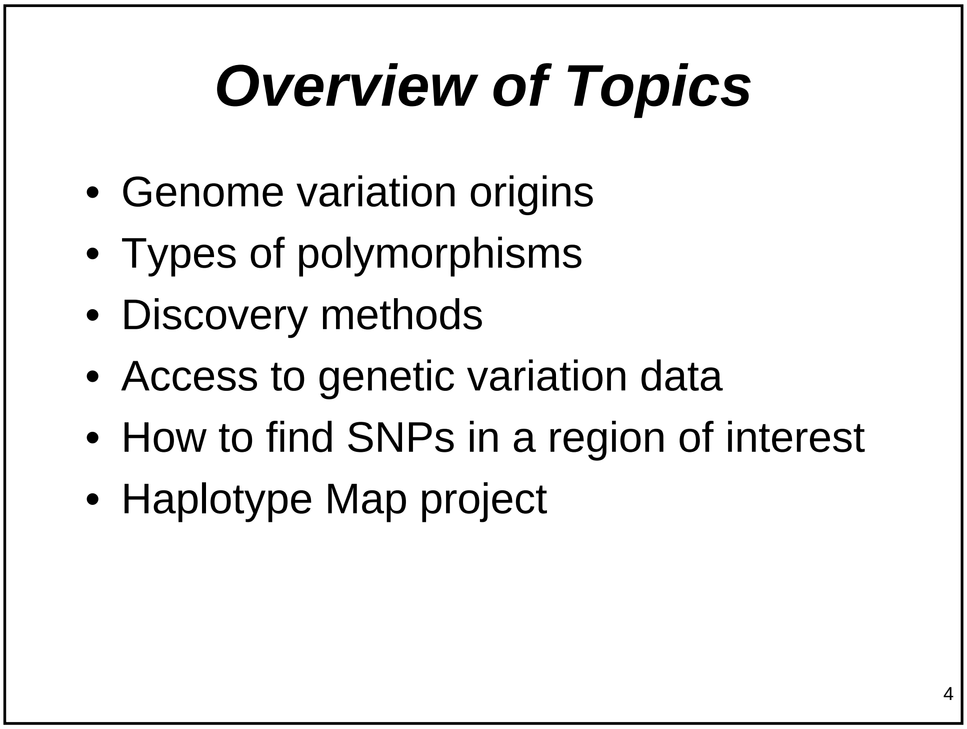
Some points from other lectures

- Population Genetics: Practical Applications by Lynn Jorde
  - Described patterns of human genetic variation among and within populations, linkage disequilibrium and HapMap and how all this relates to the search for complex disease genes.
- Identification of Cancer Susceptibility Genes by Elaine Ostrander
  - Genome wide scans to find cancer susceptibility genes and apply haplotype analyses to identify founder haplotypes.
- Genetic Variation II: Laboratory Techniques by Karen Mohlke
  - Focusing primarily on SNP genotyping methods



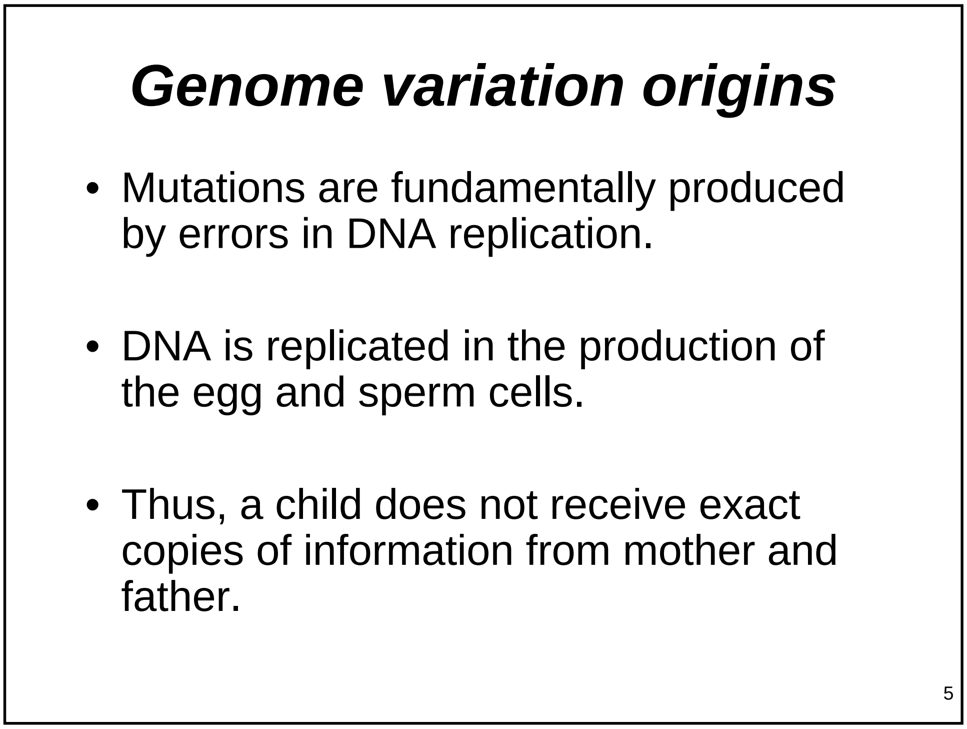
Overview of Topics

- Genome variation origins
- Types of polymorphisms
- Polymorphism discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project
- Extra topics, time permitting



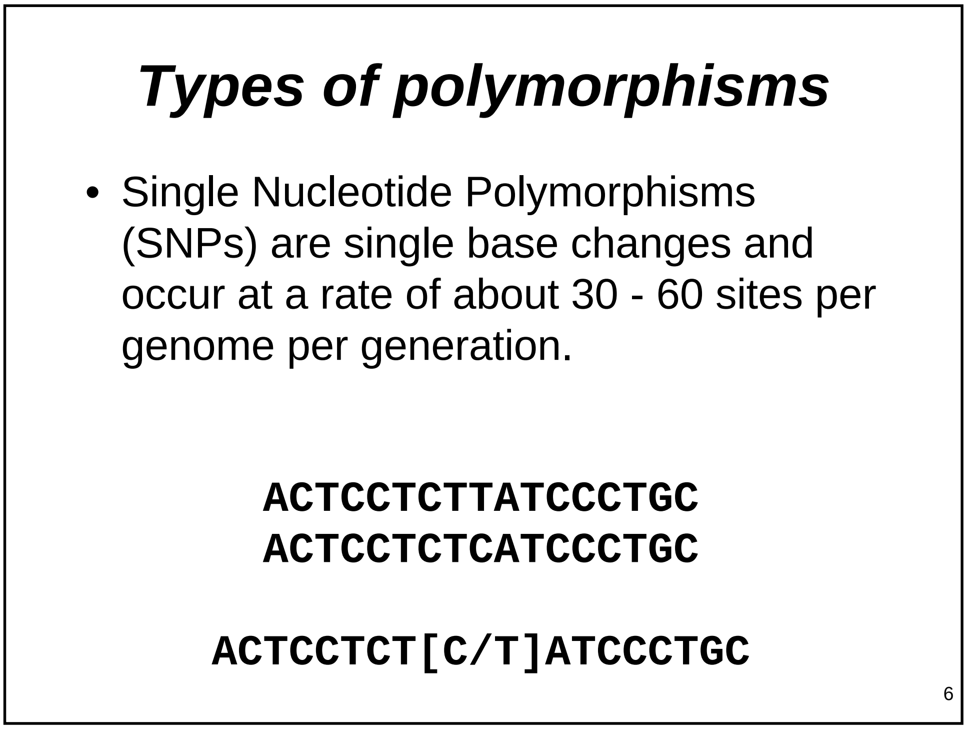
Overview of Topics

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



Genome variation origins

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

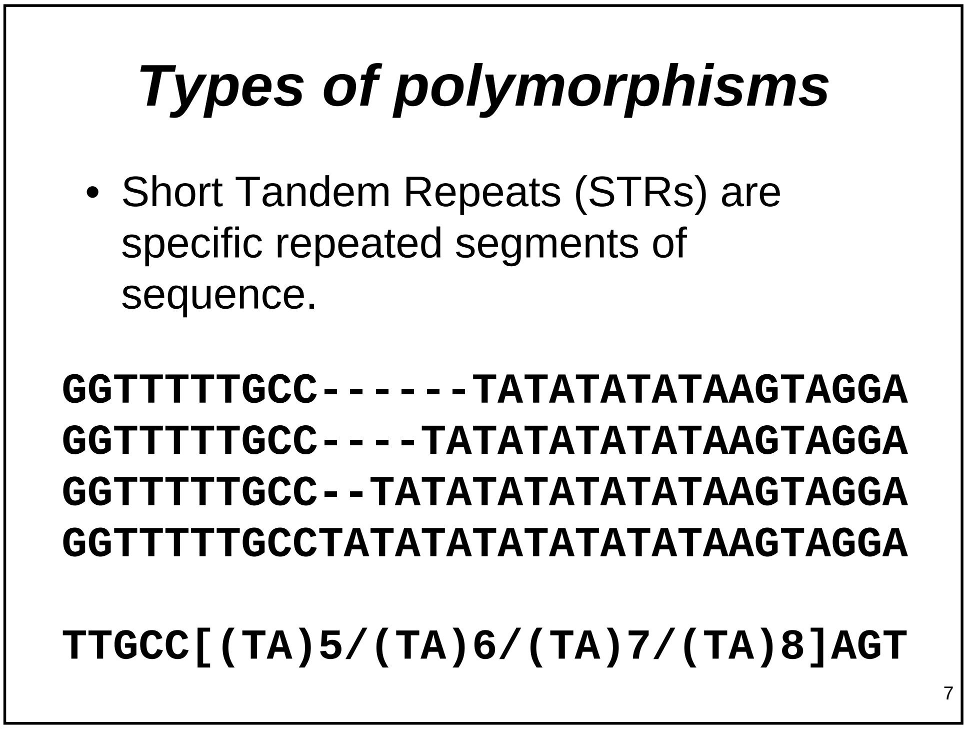


Types of polymorphisms

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

ACTCCTCTTATCCCTGC
ACTCCTCTCATCCCTGC

ACTCCTCT[C/T]ATCCCTGC

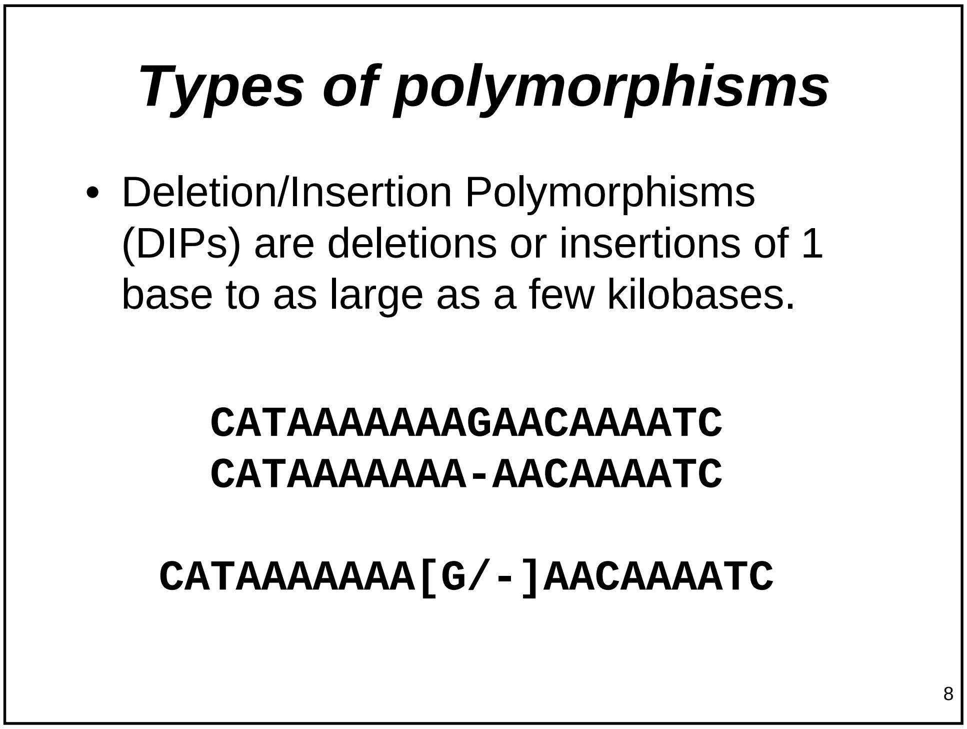


Types of polymorphisms

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

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

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



Types of polymorphisms

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

```
  CATAAAAAAAGAACAAAATC
  CATAAAAAAA-AACAAAATC

CATAAAAAAA[G/-]AACAAAATC
```

Beyond polymorphisms

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

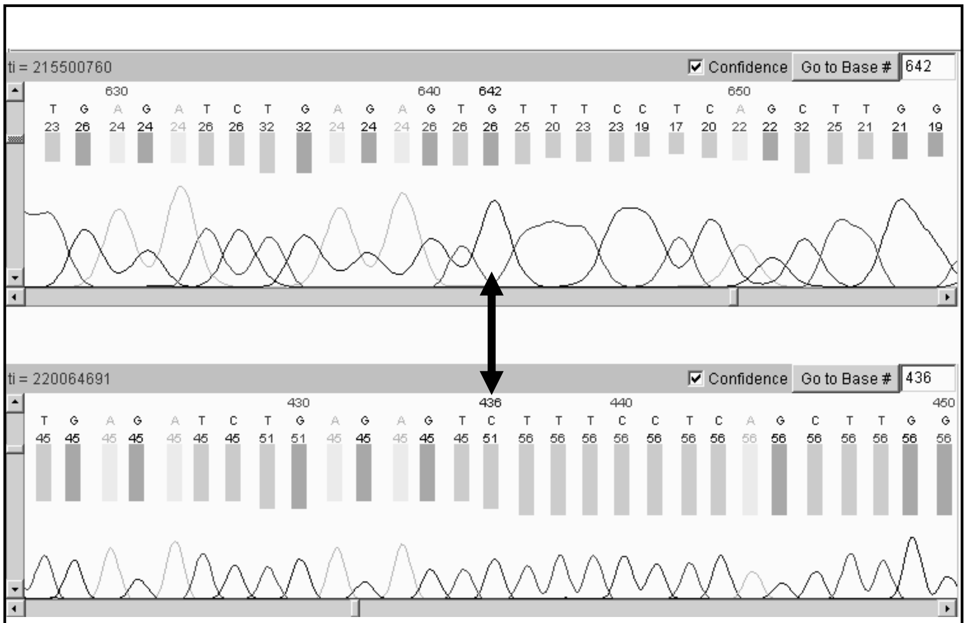
Overview of Topics

- 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.

11



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

12

Mining SNPs from sequence

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

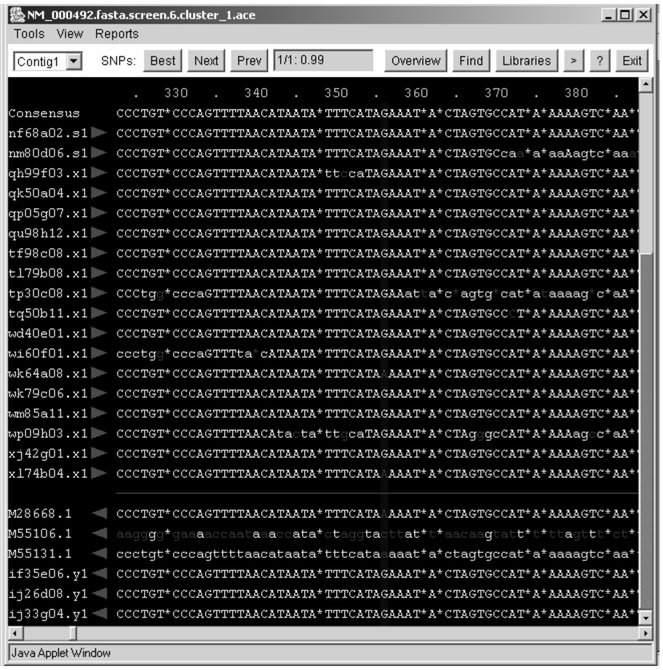
Expressed Sequence Tag Mining

- 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).



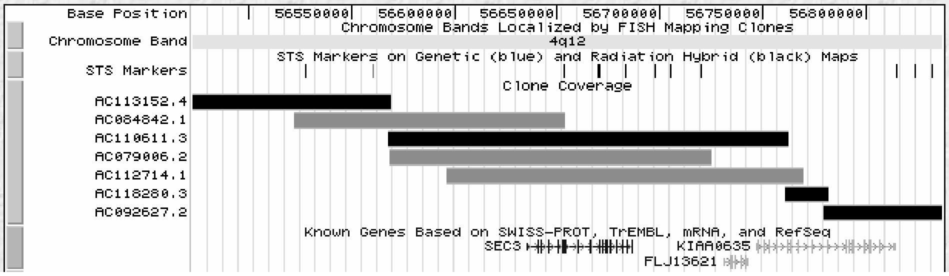
```
                .   330   .   340   .   350   .   360   .   370   .   380   .
Consensus       CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
nf68a02.s1      CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
nm80d06.s1      CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCcaa*a*aaAagtc*aaa
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*aacaagtatt*t*ttagttt*ct*
M55131.1        ccctgt*cccagttttaacataata*tttcataaaaat*a*ctagtgccat*a*aaaagtc*aa*
if35e06.y1      CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
ij26d08.y1      CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
ij33g04.y1      CCCTGT*CCCAGTTTTAACATAATA*TTTCATAGAAAT*A*CTAGTGCCAT*A*AAAAGTC*AA*
```

15

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.

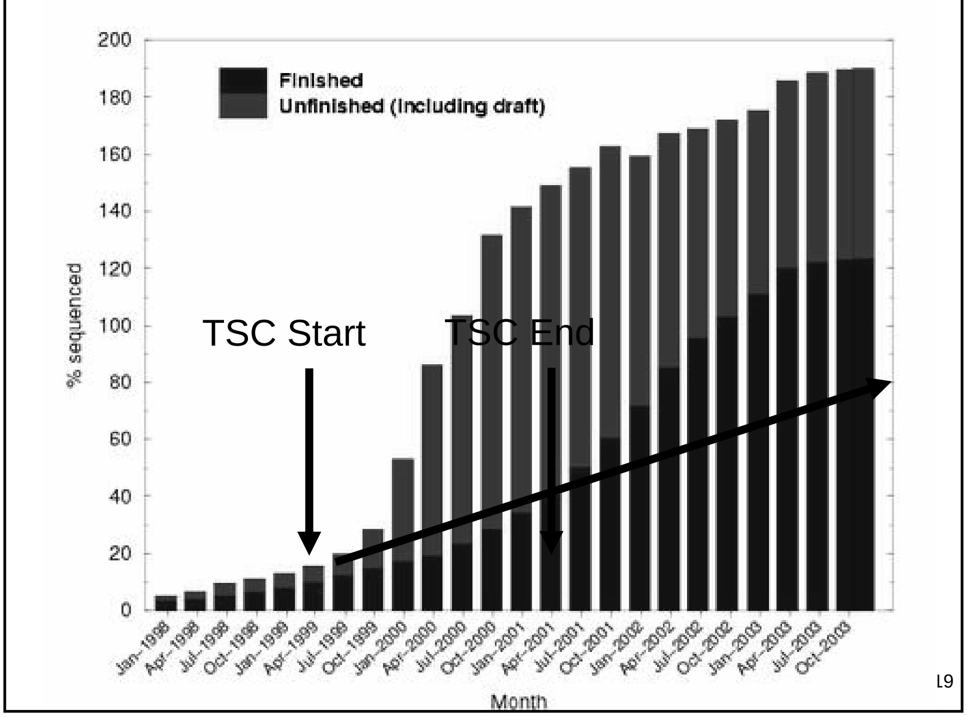


Clone Overlap

- 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.



TSC Start

TSC End

The SNP Consortium

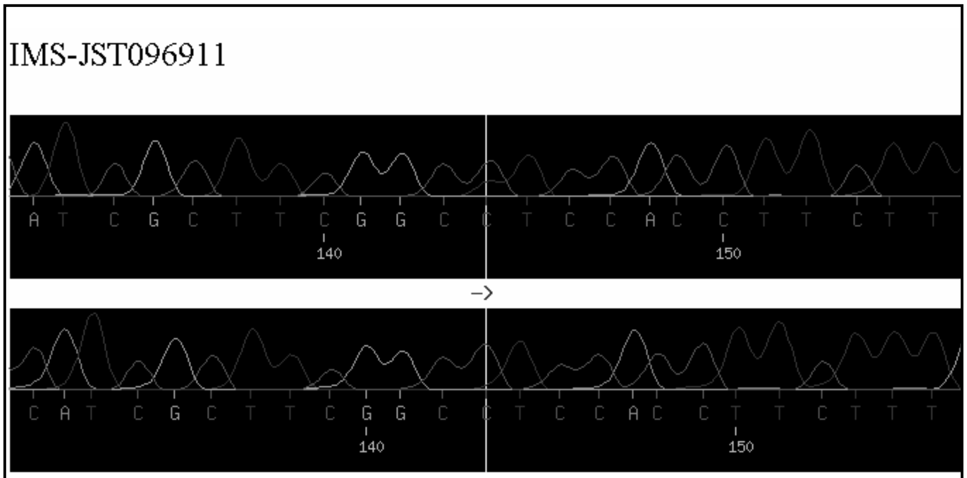
- With the rapid increase in genome coverage from the public Human Genome Project, the strategies changed to take full advantage of the draft and finished sequence.
- The initial target of 300,000 SNP was passed quickly, and now the sequence generated from that project contributes over 1.3M SNPs to the 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 for the HapMap Project.
- It has been estimated that there are perhaps 10M common SNPs (> 5% MAF), so there are many more SNPs yet to discover.

Targeted Resequencing (Medical Sequencing)

- 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.

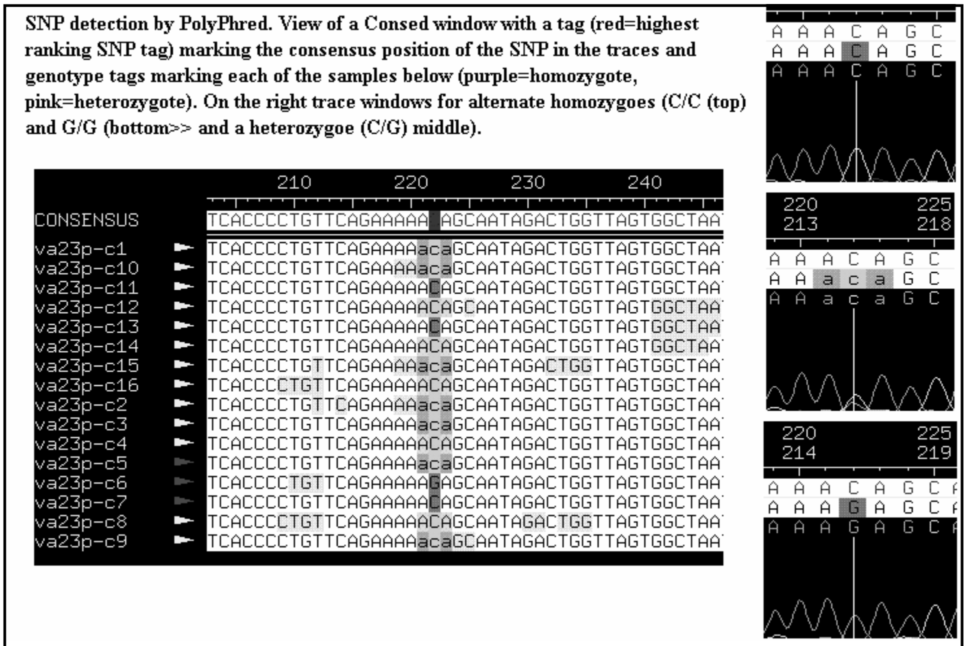

IMS-JST096911
A T C G C T T C G G C C T C C A C C T T C T T
140
150
->
C A T C G C T T C G G C C T C C A C C T T C T T T
140
150

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).

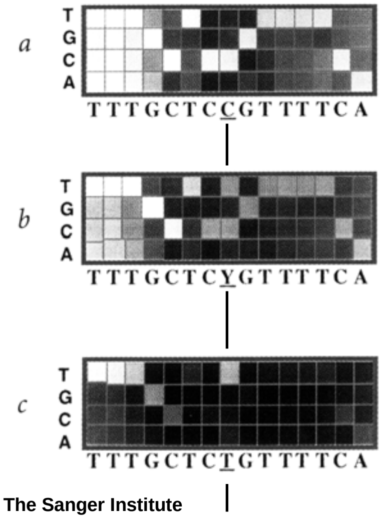
```
                             210       220       230       240
CONSENSUS      TCACCCCTGTTCAGAAAAA AGCAATAGACTGGTTAGTGGCTAA
va23p-c1    ►  TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c10   ►  TCACCCCTGTTCAGAAAAacaGCAATAGACTGGTTAGTGGCTAA
va23p-c11   ►  TCACCCCTGTTCAGAAAAACAGCAATAGACTGGTTAGTGGCTAA
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.

25

Sequencing Chips

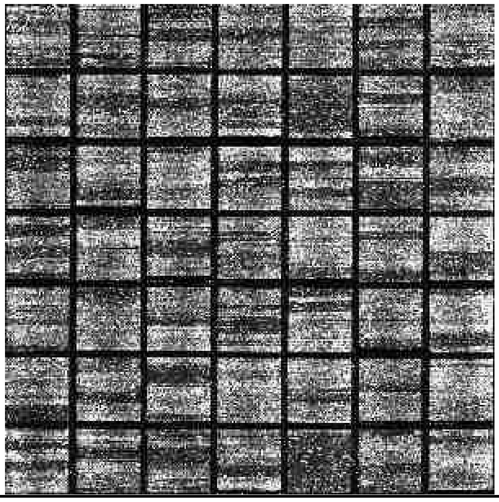
...GCTCCGTTT...
...GCTCTGTTT...


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

26

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.



27

Distribution properties

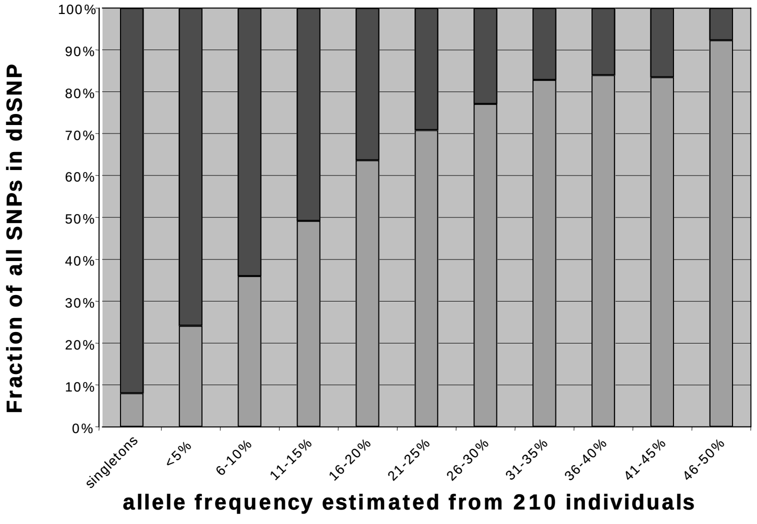
- EST mining
  - Locates SNPs primarily within coding regions.
- Clone overlap
  - High density of SNPs within overlap regions, absent elsewhere.
- The SNP Consortium (TSC)
  - Randomly distributed across the genome, however, total sequence only covers 50% of the genome

28

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

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



Overview of Topics

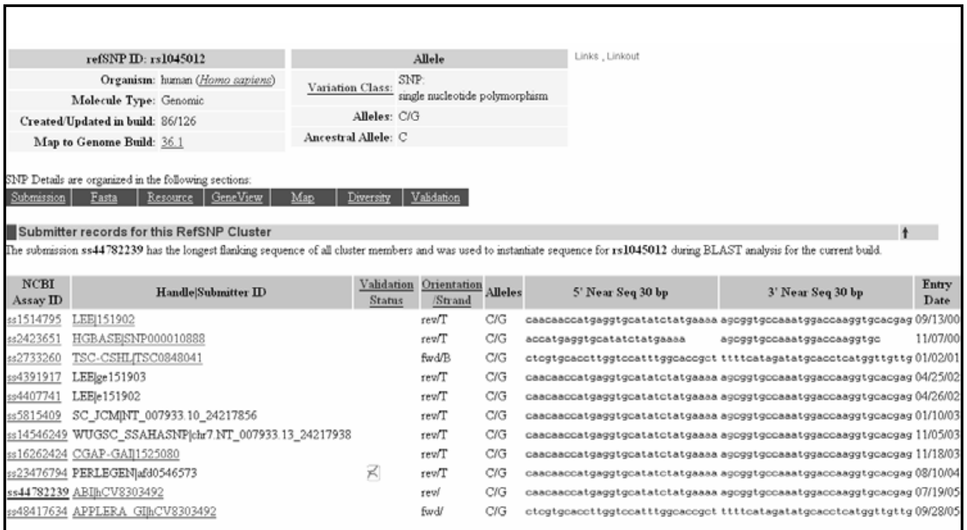
- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project

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.



refSNP ID: rs1045012	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	86/126
Map to Genome Build:	36.1

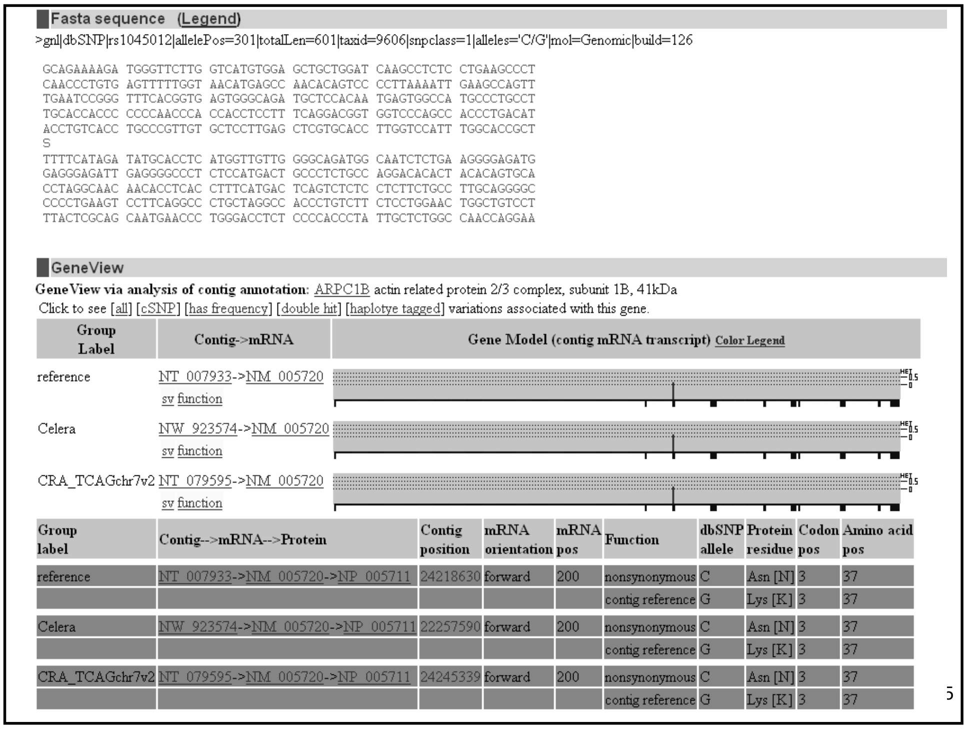
Allele	
Variation Class:	SNP: single nucleotide polymorphism
Alleles:	C/G
Ancestral Allele:	C

SNP Details are organized in the following sections:

Submitter records for this RefSNP Cluster

The submission **ss44782239** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs1045012** during BLAST analysis for the current build.

NCBI Assay ID	Handle|Submitter ID	Validation Status	Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp	Entry Date
ss1514795	LEE|151902		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	09/13/00
ss2423651	HGBASE|SNP000010888		rev/T	C/G	accatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgc	11/07/00
ss2733260	TSC-CSHL|TSC0848041		fwd/B	C/G	ctcgtgcaccttggtccatttggcaccgct	ttttcatagatatgcacctcatggttgttg	01/02/01
ss4391917	LEE|ge151903		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	04/25/02
ss4407741	LEE|e151902		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	04/26/02
ss5815409	SC_JCM|NT_007933.10_24217856		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	01/10/03
ss14546249	WUGSC_SSAHASNP|chr7.NT_007933.13_24217938		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	11/05/03
ss16262424	CGAP-GAI|1525080		rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	11/18/03
ss23476794	PERLEGEN|afd0546573	☒	rev/T	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	08/10/04
ss44782239	ABI|hCV8303492		rev/	C/G	caacaaccatgaggtgcatatctatgaaaa	agcggtgccaaatggaccaaggtgcacgag	07/19/05
ss48417634	APPLERA_GI|hCV8303492		fwd/	C/G	ctcgtgcaccttggtccatttggcaccgct	ttttcatagatatgcacctcatggttgttg	09/28/05



Fasta sequence (Legend)

```
>gnl|dbSNP|rs1045012|allelePos=301|totalLen=601|taxid=9606|snpclass=1|alleles='C/G'|mol=Genomic|build=126

 GCAGAAAAGA TGGGTTCTTG GTCATGTGGA GCTGCTGGAT CAAGCCTCTC CTGAAGCCCT
 CAACCCTGTG AGTTTTTGGT AACATGAGCC AACACAGTCC CCTTAAAATT GAAGCCAGTT
 TGAATCCGGG TTTCACGGTG AGTGGGCAGA TGCTCCACAA TGAGTGGCCA TGCCCTGCCT
 TGCACCACCC CCCCAACCCA CCACCTCCTT TCAGGACGGT GGTCCCAGCC ACCCTGACAT
 ACCTGTCACC TGCCCGTTGT GCTCCTTGAG CTCGTGCACC TTGGTCCATT TGGCACCGCT
 S
 TTTTCATAGA TATGCACCTC ATGGTTGTTG GGGCAGATGG CAATCTCTGA AGGGGAGATG
 GAGGGAGATT GAGGGGCCCT CTCCATGACT GCCCTCTGCC AGGACACACT ACACAGTGCA
 CCTAGGCAAC AACACCTCAC CTTTCATGAC TCAGTCTCTC CTCTTCTGCC TTGCAGGGGC
 CCCCTGAAGT CCTTCAGGCC CTGCTAGGCC ACCCTGTCTT CTCCTGGAAC TGGCTGTCCT
 TTACTCGCAG CAATGAACCC TGGGACCTCT CCCCACCCTA TTGCTCTGGC CAACCAGGAA
```

GeneView

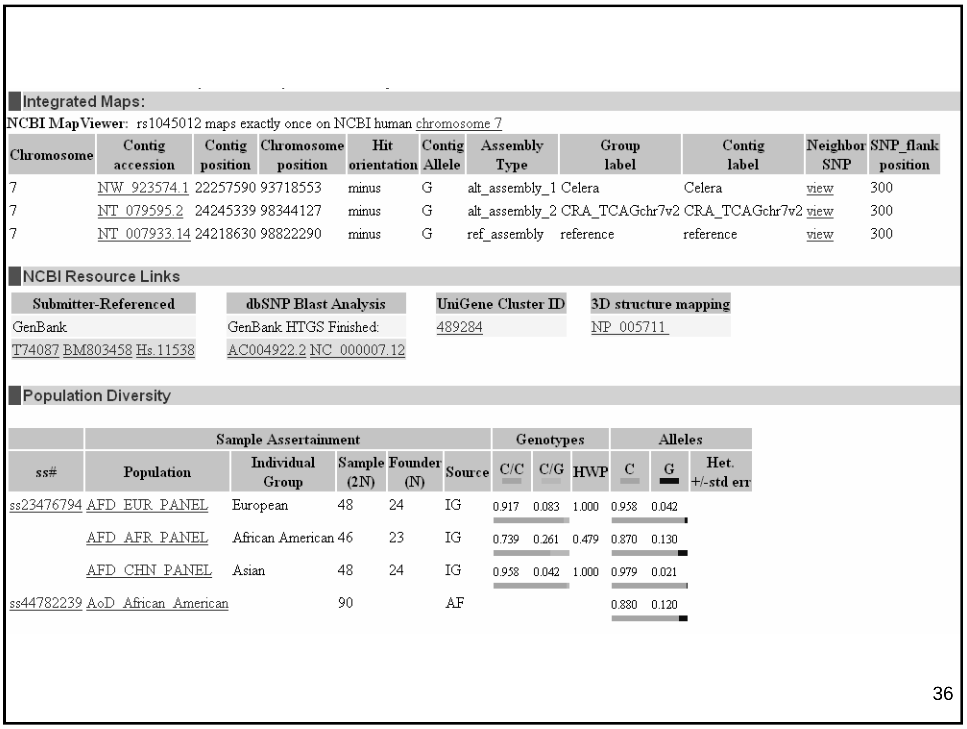
GeneView via analysis of contig annotation: ARPC1B actin related protein 2/3 complex, subunit 1B, 41kDa

Click to see [all] [cSNP] [has frequency] [double hit] [haplotye tagged] variations associated with this gene.

Group Label	Contig->mRNA	Gene Model (contig mRNA transcript) Color Legend
reference	NT_007933->NM_005720 sv function	
Celera	NW_923574->NM_005720 sv function	
CRA_TCAGchr7v2	NT_079595->NM_005720 sv function	

Group label	Contig-->mRNA-->Protein	Contig position	mRNA orientation	mRNA pos	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
reference	NT_007933->NM_005720->NP_005711	24218630	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37
Celera	NW_923574->NM_005720->NP_005711	22257590	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37
CRA_TCAGchr7v2	NT_079595->NM_005720->NP_005711	24245339	forward	200	nonsynonymous	C	Asn [N]	3	37
					contig reference	G	Lys [K]	3	37

5



Integrated Maps:

NCBI MapViewer: rs1045012 maps exactly once on NCBI human chromosome 7

Chromosome	Contig accession	Contig position	Chromosome position	Hit orientation	Contig Allele	Assembly Type	Group label	Contig label	Neighbor SNP	SNP_flank position
7	NW_923574.1	22257590	93718553	minus	G	alt_assembly_1	Celera	Celera	view	300
7	NT_079595.2	24245339	98344127	minus	G	alt_assembly_2	CRA_TCAGchr7v2	CRA_TCAGchr7v2	view	300
7	NT_007933.14	24218630	98822290	minus	G	ref_assembly	reference	reference	view	300

NCBI Resource Links

Submitter-Referenced	dbSNP Blast Analysis	UniGene Cluster ID	3D structure mapping
GenBank	GenBank HTGS Finished:	489284	NP_005711
T74087 BM803458 Hs.11538	AC004922.2 NC_000007.12		

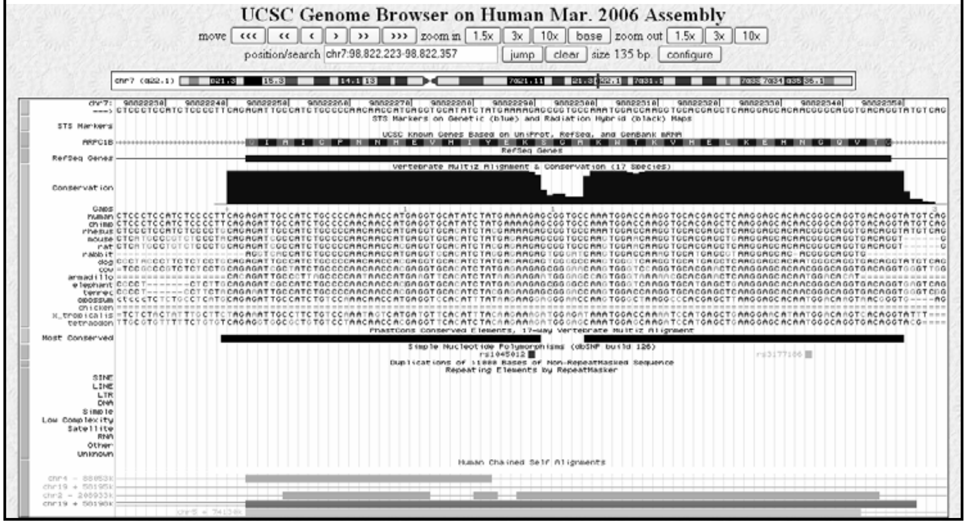
Population Diversity

	Sample Assertainment					Genotypes			Alleles		
ss#	Population	Individual Group	Sample (2N)	Founder (N)	Source	C/C	C/G	HWP	C	G	Het. +/-std err
ss23476794	AFD_EUR_PANEL	European	48	24	IG	0.917	0.083	1.000	0.958	0.042	
	AFD_AFR_PANEL	African American	46	23	IG	0.739	0.261	0.479	0.870	0.130	
	AFD_CHN_PANEL	Asian	48	24	IG	0.958	0.042	1.000	0.979	0.021	
ss44782239	AoD_African_American		90		AF				0.880	0.120	

36

Viewing SNPs in Browsers

NCBI Ensembl UCSC


UCSC Genome Browser on Human Mar. 2006 Assembly
move ‹‹‹ ‹‹ ‹ › ›› ››› zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search chr7:98,822,223-98,822,357 jump clear size 135 bp configure
chr7 (q22.1)
chr7:
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
ARPC1B
RefSeq Genes
RefSeq Genes
Vertebrate Multiz Alignment & Conservation (17 Species)
Conservation
Gaps
human
chimp
rhesus
mouse
rat
rabbit
dog
cow
armadillo
elephant
tenrec
opossum
chicken
X_tropicalis
tetraodon
PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment
Most Conserved
Simple Nucleotide Polymorphisms (dbSNP build 126)
rs1045012
rs3177106
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown
Human Chained Self Alignments

Overview of Topics

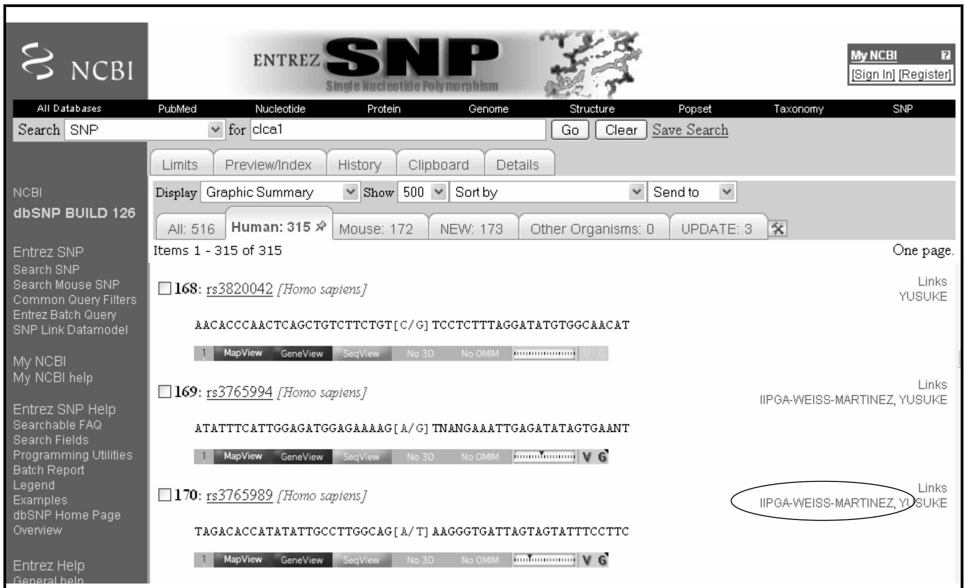
- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project

38

How to find SNPs in a region of interest

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

39



Search SNP for clca1

Display Graphic Summary Show 500 Sort by Send to

Items 1 - 315 of 315

One page.

□ 168: rs3820042 *[Homo sapiens]*

Links
YUSUKE

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

□ 169: rs3765994 *[Homo sapiens]*

Links
IIPGA-WEISS-MARTINEZ, YUSUKE

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

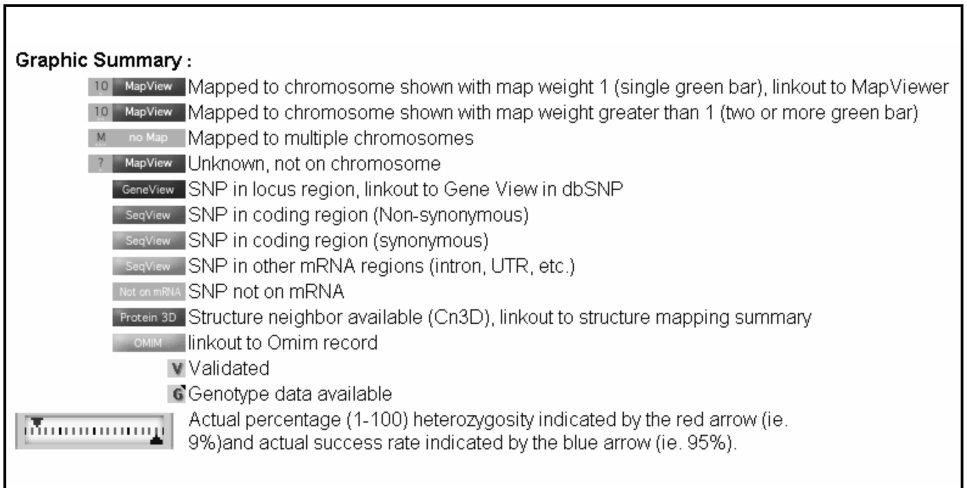
□ 170: rs3765989 *[Homo sapiens]*

Links
IIPGA-WEISS-MARTINEZ, YUSUKE

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

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

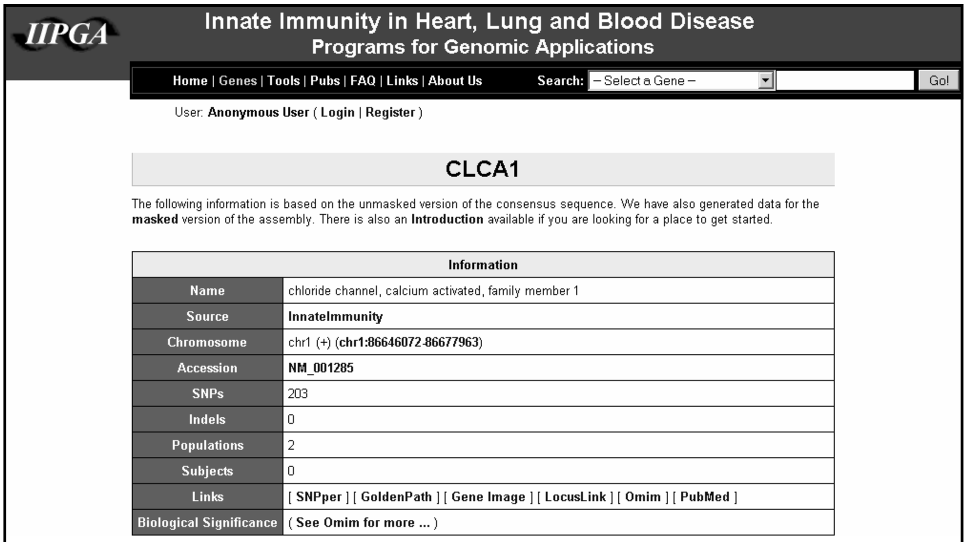
40



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



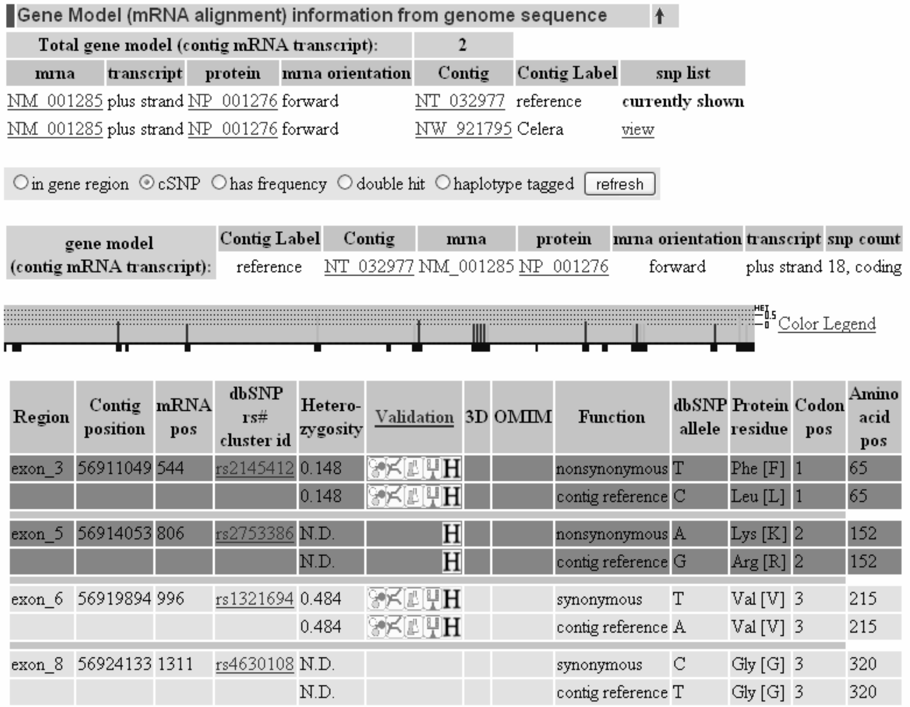
IIPGA
Innate Immunity in Heart, Lung and Blood Disease
Programs for Genomic Applications
Home | Genes | Tools | Pubs | FAQ | Links | About Us
Search: – Select a Gene – Go!
User: Anonymous User (Login | Register)

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

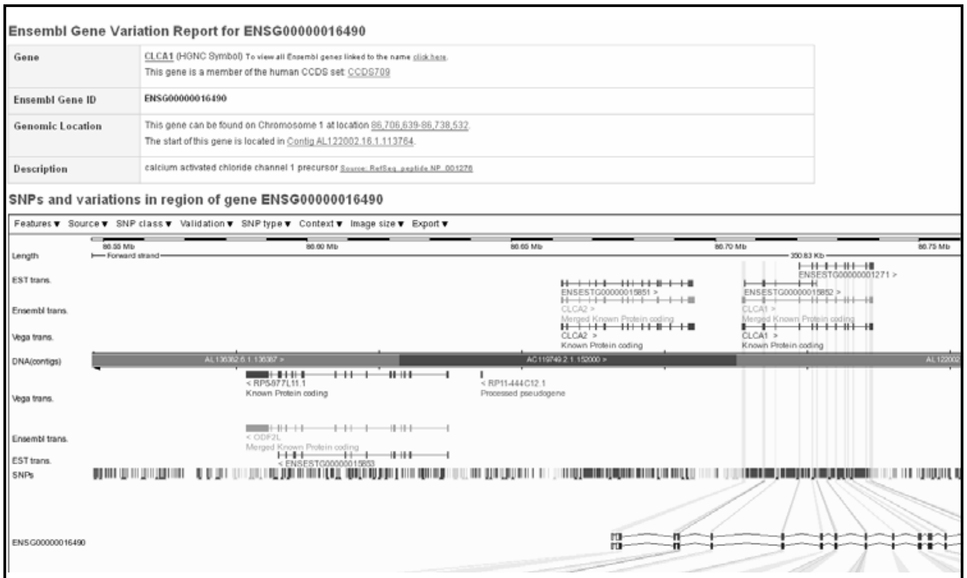
Total gene model (contig mRNA transcript):	2					
mrna	**transcript**	**protein**	**mrna orientation**	**Contig**	**Contig Label**	**snp list**
NM_001285	plus strand	NP_001276	forward	NT_032977	reference	**currently shown**
NM_001285	plus strand	NP_001276	forward	NW_921795	Celera	view

○ in gene region ● cSNP ○ has frequency ○ double hit ○ haplotype tagged [refresh]

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

Region	Contig position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	OMIM	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
exon_3	56911049	544	rs2145412	0.148	H			nonsynonymous	T	Phe [F]	1	65
				0.148	H			contig reference	C	Leu [L]	1	65
exon_5	56914053	806	rs2753386	N.D.	H			nonsynonymous	A	Lys [K]	2	152
				N.D.	H			contig reference	G	Arg [R]	2	152
exon_6	56919894	996	rs1321694	0.484	H			synonymous	T	Val [V]	3	215
				0.484	H			contig reference	A	Val [V]	3	215
exon_8	56924133	1311	rs4630108	N.D.				synonymous	C	Gly [G]	3	320
				N.D.				contig reference	T	Gly [G]	3	320

43



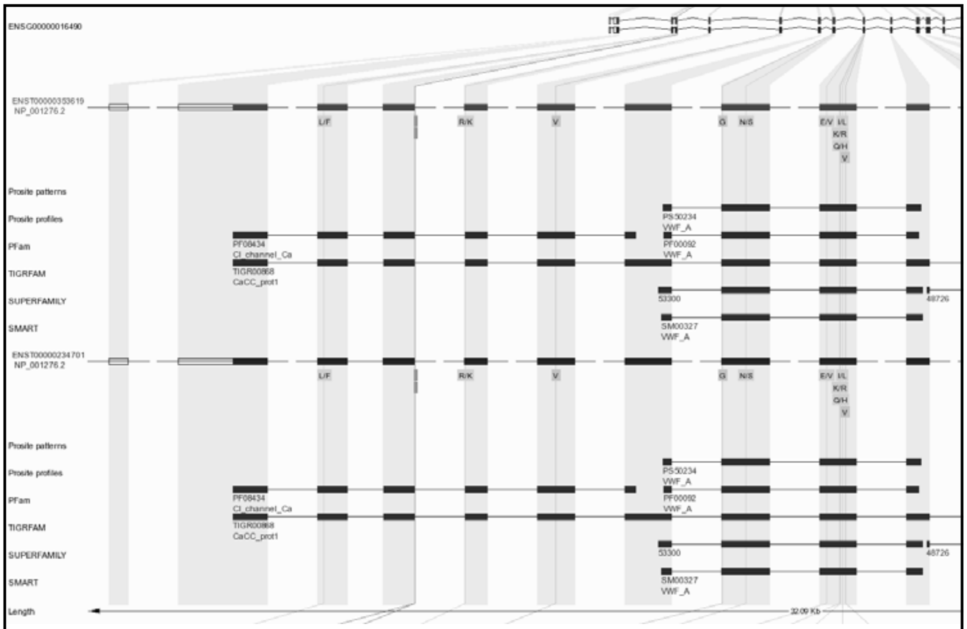
Ensembl Gene Variation Report for ENSG00000016490

Gene	CLCA1 (HGNC Symbol) To view all Ensembl genes linked to the name click here. This gene is a member of the human CCDS set: CCDS709
Ensembl Gene ID	**ENSG00000016490**
Genomic Location	This gene can be found on Chromosome 1 at location 86,706,639-86,738,532. The start of this gene is located in Contig AL122002.16.1.113764.
Description	calcium activated chloride channel 1 precursor Source: RefSeq_peptide NP_001276

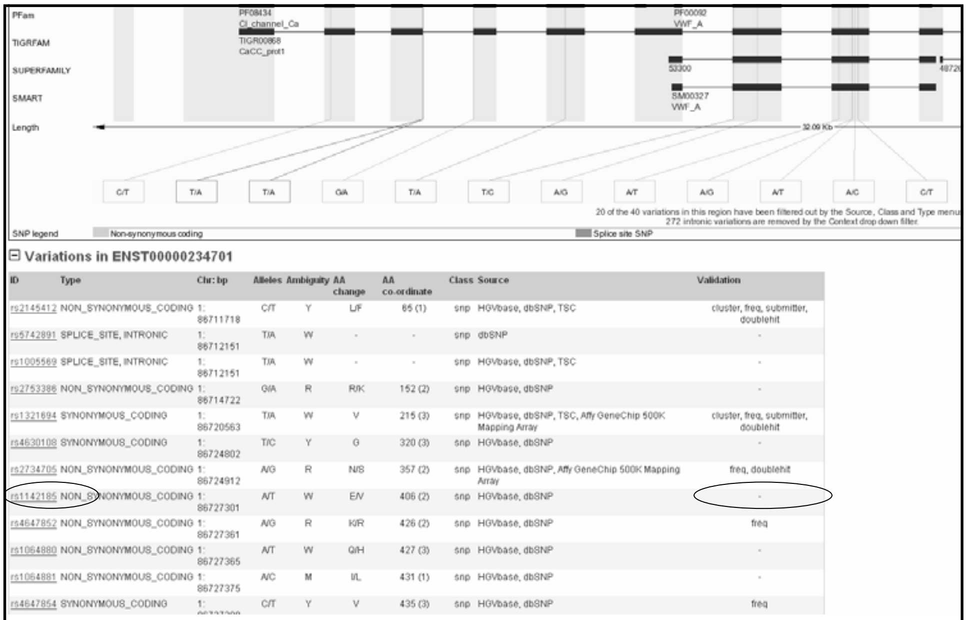
SNPs and variations in region of gene ENSG00000016490

http://www.ensembl.org/Homo_sapiens

44



45

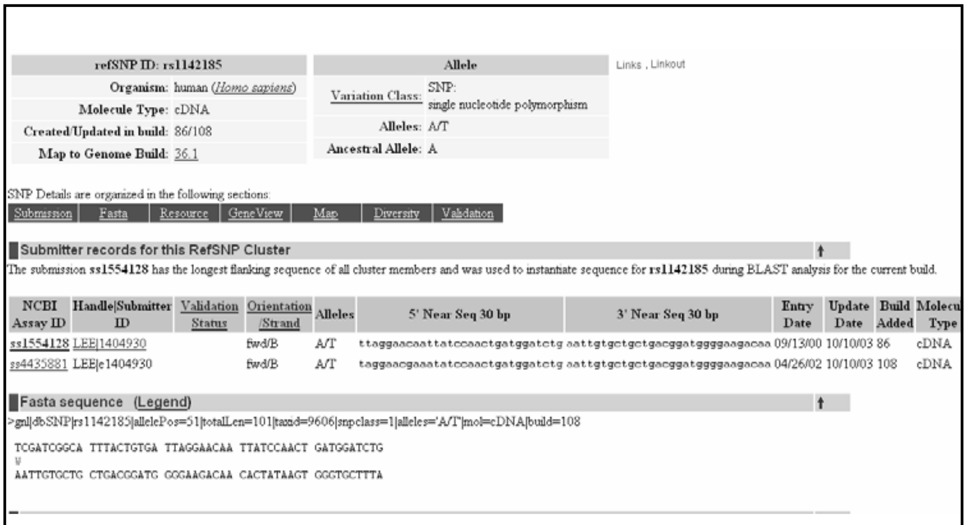


20 of the 40 variations in this region have been filtered out by the Source, Class and Type menu
272 intronic variations are removed by the Context drop down filter.

⊟ **Variations in ENST00000234701**

ID	Type	Chr: bp	Alleles	Ambiguity	AA change	AA co-ordinate	Class	Source	Validation
rs2145412	NON_SYNONYMOUS_CODING	1: 86711718	C/T	Y	L/F	65 (1)	snp	HGVbase, dbSNP, TSC	cluster, freq, submitter, doublehit
rs5742891	SPLICE_SITE, INTRONIC	1: 86712151	T/A	W	-	-	snp	dbSNP	-
rs1005569	SPLICE_SITE, INTRONIC	1: 86712151	T/A	W	-	-	snp	HGVbase, dbSNP, TSC	-
rs2753386	NON_SYNONYMOUS_CODING	1: 86714722	G/A	R	R/K	152 (2)	snp	HGVbase, dbSNP	-
rs1321694	SYNONYMOUS_CODING	1: 86720563	T/A	W	V	215 (3)	snp	HGVbase, dbSNP, TSC, Affy GeneChip 500K Mapping Array	cluster, freq, submitter, doublehit
rs4630108	SYNONYMOUS_CODING	1: 86724802	T/C	Y	G	320 (3)	snp	HGVbase, dbSNP	-
rs2734705	NON_SYNONYMOUS_CODING	1: 86724912	A/G	R	N/S	357 (2)	snp	HGVbase, dbSNP, Affy GeneChip 500K Mapping Array	freq, doublehit
rs1142185	NON_SYNONYMOUS_CODING	1: 86727301	A/T	W	E/V	406 (2)	snp	HGVbase, dbSNP	-
rs4647852	NON_SYNONYMOUS_CODING	1: 86727361	A/G	R	K/R	426 (2)	snp	HGVbase, dbSNP	freq
rs1064880	NON_SYNONYMOUS_CODING	1: 86727365	A/T	W	Q/H	427 (3)	snp	HGVbase, dbSNP	-
rs1064881	NON_SYNONYMOUS_CODING	1: 86727375	A/C	M	I/L	431 (1)	snp	HGVbase, dbSNP	-
rs4647854	SYNONYMOUS_CODING	1:	C/T	Y	V	435 (3)	snp	HGVbase, dbSNP	freq

46



refSNP ID: rs1142185	
Organism:	human (*Homo sapiens*)
Molecule Type:	cDNA
Created/Updated in build:	86/108
Map to Genome Build:	36.1

Allele	
Variation Class:	SNP: single nucleotide polymorphism
Alleles:	A/T
Ancestral Allele:	A

Links , Linkout

SNP Details are organized 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 **rs1142185** during BLAST analysis for the current build.

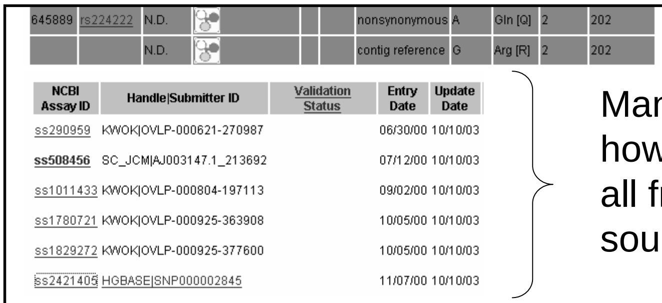
NCBI Assay ID	Handle|Submitter ID	Validation Status	Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp	Entry Date	Update Date	Build Added	Molecu Type
ss1554128	LEE|1404930		fwd/B	A/T	ttaggaacaattatccaactgatggatctg	aattgtgctgctgacggatggggaagacaa	09/13/00	10/10/03	86	cDNA
ss4435881	LEE|e1404930		fwd/B	A/T	taggaacgaaatatccaactgatggatctg	aattgtgctgctgacggatggggaagacaa	04/26/02	10/10/03	108	cDNA

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
```

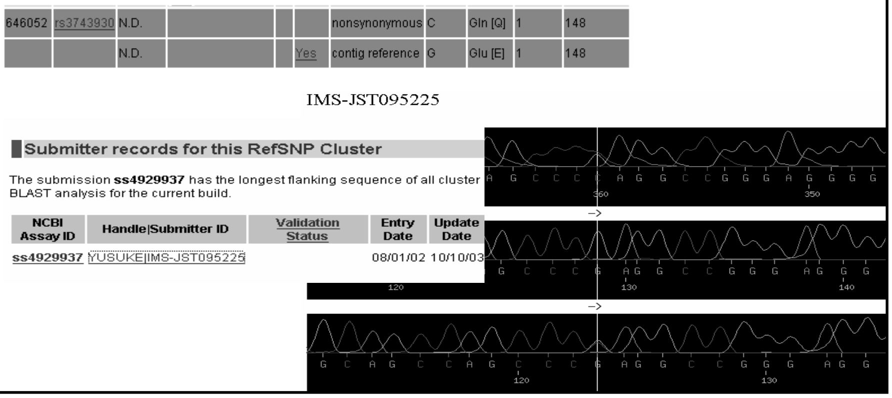
47



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.

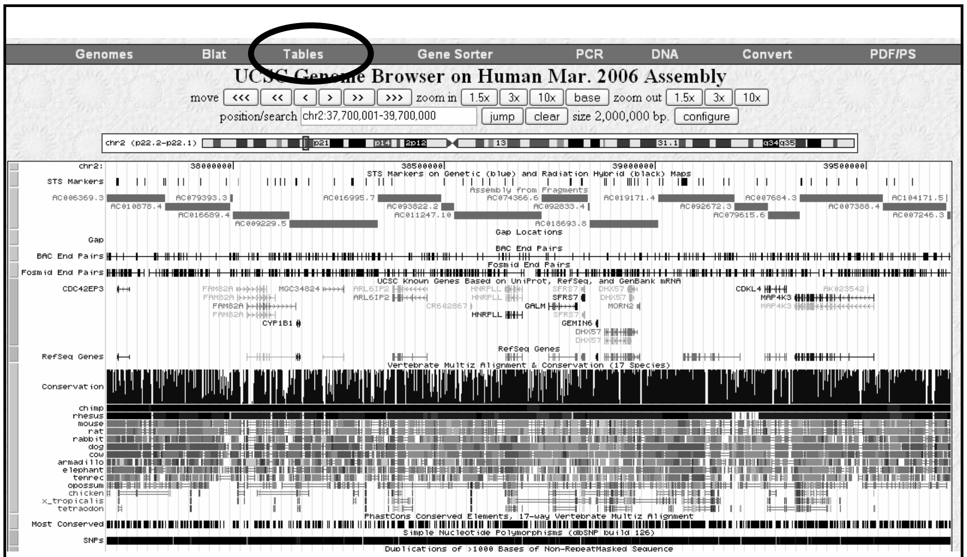
NCBI Assay ID	Handle|Submitter ID	Validation Status	Entry Date	Update Date
ss4929937	YUSUKE|IMS-JST095225		08/01/02	10/10/03

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

49



http://genome.ucsc.edu

50

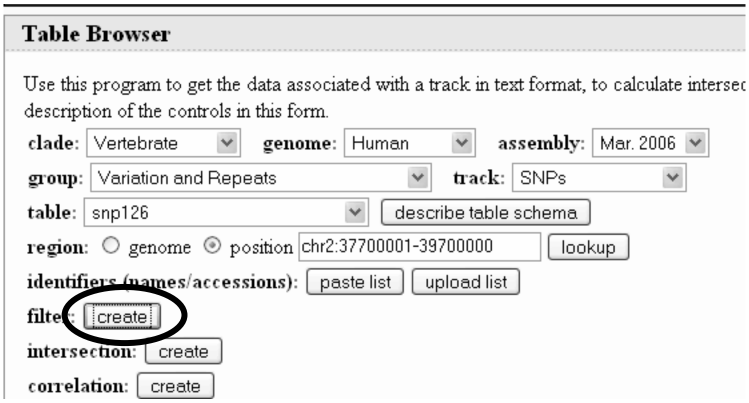


Table Browser

Use this program to get the data associated with a track in text format, to calculate intersec
description of the controls in this form.

clade: Vertebrate genome: Human assembly: Mar. 2006

group: Variation and Repeats track: SNPs

table: snp126 describe table schema

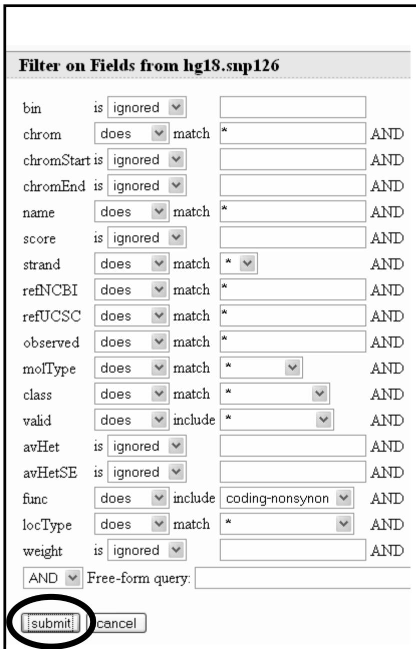
region: ○ genome ● position chr2:37700001-39700000 lookup

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create



Filter on Fields from hg18.snp126

bin	is	ignored		
chrom	does	match	*	AND
chromStart	is	ignored		AND
chromEnd	is	ignored		AND
name	does	match	*	AND
score	is	ignored		AND
strand	does	match	*	AND
refNCBI	does	match	*	AND
refUCSC	does	match	*	AND
observed	does	match	*	AND
molType	does	match	*	AND
class	does	match	*	AND
valid	does	include	*	AND
avHet	is	ignored		AND
avHetSE	is	ignored		AND
func	does	include	coding-nonsynon	AND
locType	does	match	*	AND
weight	is	ignored		AND

AND Free-form query:

submit cancel

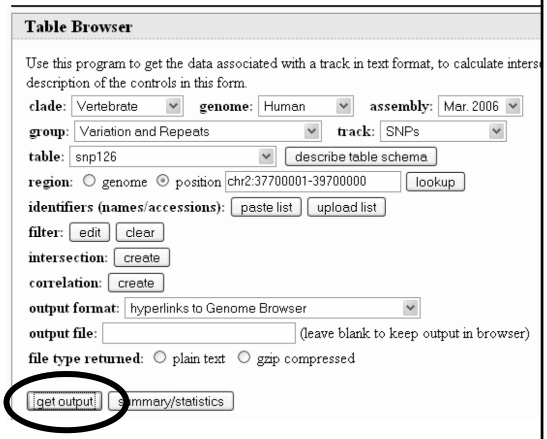


Table Browser

Use this program to get the data associated with a track in text format, to calculate inters
description of the controls in this form.

clade: Vertebrate genome: Human assembly: Mar. 2006

group: Variation and Repeats track: SNPs

table: snp126 describe table schema

region: ○ genome ● position chr2:37700001-39700000 lookup

identifiers (names/accessions): paste list upload list

filter: edit clear

intersection: create

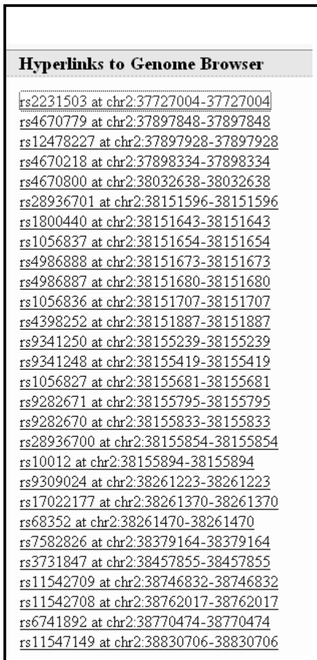
correlation: create

output format: hyperlinks to Genome Browser

output file: (leave blank to keep output in browser)

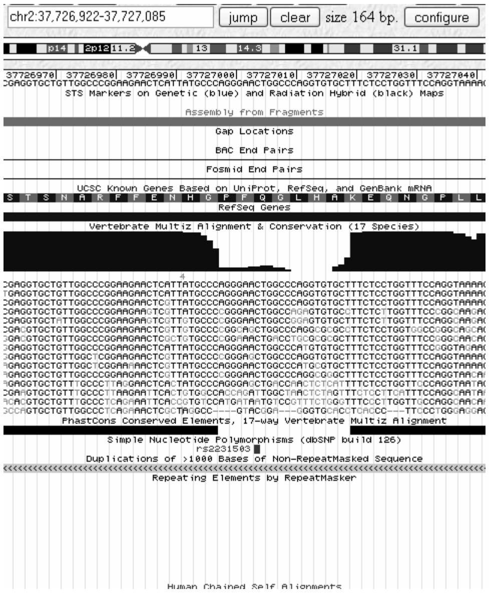
file type returned: ○ plain text ○ gzip compressed

get output summary/statistics



Hyperlinks to Genome Browser

rs2231503 at chr2:37727004-37727004
rs4670779 at chr2:37897848-37897848
rs12478227 at chr2:37897928-37897928
rs4670218 at chr2:37898334-37898334
rs4670800 at chr2:38032638-38032638
rs28936701 at chr2:38151596-38151596
rs1800440 at chr2:38151643-38151643
rs1056837 at chr2:38151654-38151654
rs4986888 at chr2:38151673-38151673
rs4986887 at chr2:38151680-38151680
rs1056836 at chr2:38151707-38151707
rs4398252 at chr2:38151887-38151887
rs9341250 at chr2:38155239-38155239
rs9341248 at chr2:38155419-38155419
rs1056827 at chr2:38155681-38155681
rs9282671 at chr2:38155795-38155795
rs9282670 at chr2:38155833-38155833
rs28936700 at chr2:38155854-38155854
rs10012 at chr2:38155894-38155894
rs9309024 at chr2:38261223-38261223
rs17022177 at chr2:38261370-38261370
rs68352 at chr2:38261470-38261470
rs7582826 at chr2:38379164-38379164
rs3731847 at chr2:38457855-38457855
rs11542709 at chr2:38746832-38746832
rs11542708 at chr2:38762017-38762017
rs6741892 at chr2:38770474-38770474
rs11547149 at chr2:38830706-38830706



chr2:37,726,922-37,727,085

jump
clear
size 164 bp.
configure
p14
2p12
11.2
13
14.3
31.1
37726970
37726980
37726990
37727000
37727010
37727020
37727030
37727040
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
Assembly from Fragments
Gap Locations
BAC End Pairs
Fosmid End Pairs
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
S T S N A R F F E N H G P F Q G L H A K E Q N G P L L
RefSeq Genes
Vertebrate Multiz Alignment & Conservation (17 Species)
PhastCons Conserved Elements, 17-way Vertebrate Multiz Alignment
Simple Nucleotide Polymorphisms (dbSNP build 126)
rs2231503
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Repeating Elements by RepeatMasker
Human Chained Self Alignments

How to find SNPs in a region of interest

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

Selecting SNPs from a list of candidate genes

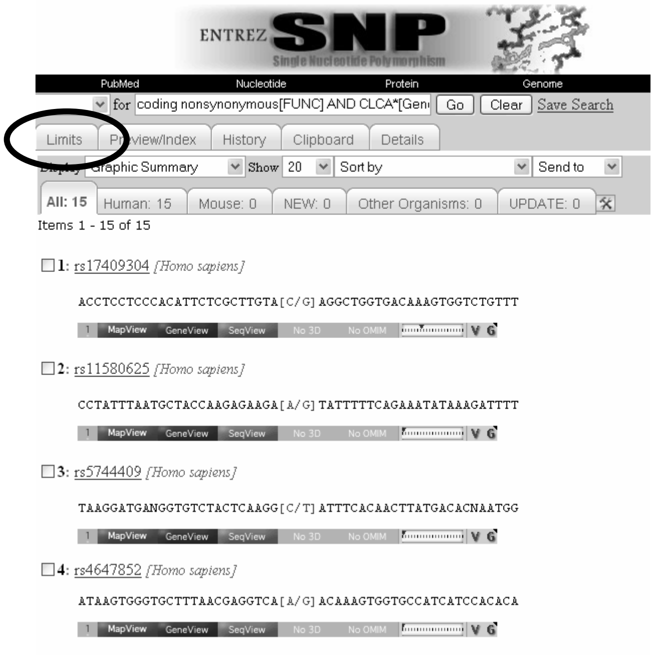
- Use the Entrez SNP query:

  coding nonsynonymous[FUNC] AND CLCA*[Gene name] AND human[orgn]

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

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Snp

55



coding nonsynonymous[FUNC] AND CLCA*[Gen

Display Graphic Summary Show 20

All: 15 Human: 15 Mouse: 0 NEW: 0 Other Organisms: 0 UPDATE: 0

Items 1 - 15 of 15

□ 1: rs17409304 *[Homo sapiens]*

```
ACCTCCTCCCACATTCTCGCTTGTA[C/G]AGGCTGGTGACAAAGTGGTCTGTTT
```

□ 2: rs11580625 *[Homo sapiens]*

```
CCTATTTAATGCTACCAAGAGAAGA[A/G]TATTTTTCAGAAATATAAAGATTTT
```

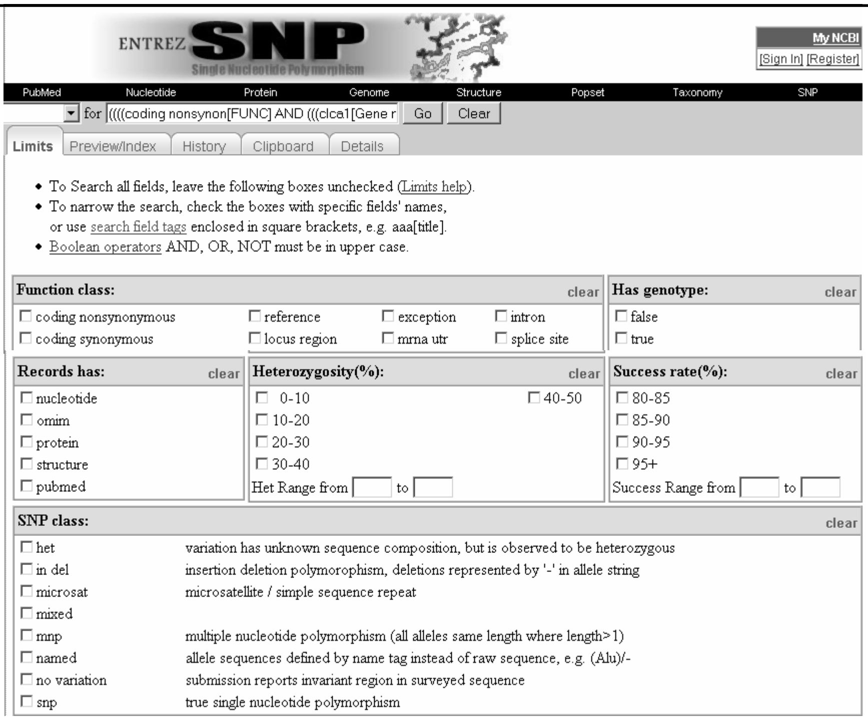
□ 3: rs5744409 *[Homo sapiens]*

```
TAAGGATGANGGTGTCTACTCAAGG[C/T]ATTTCACAACTTATGACACNAATGG
```

□ 4: rs4647852 *[Homo sapiens]*

```
ATAAGTGGGTGCTTTAACGAGGTCA[A/G]ACAAAGTGGTGCCATCATCCACACA
```

56



for ((((coding nonsynon[FUNC] AND (((clca1[Gene r

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

Has genotype: clear

□ false
□ true

Records has: clear

□ nucleotide
□ omim
□ protein
□ structure
□ pubmed

Heterozygosity(%): clear

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

57

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

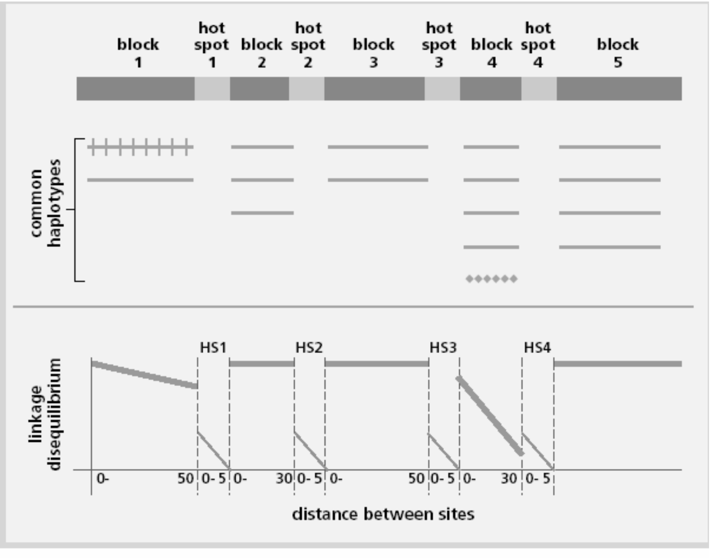
58

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.



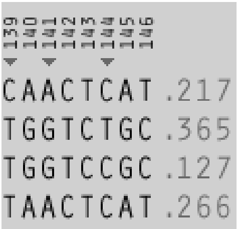
BOB CRIMI

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

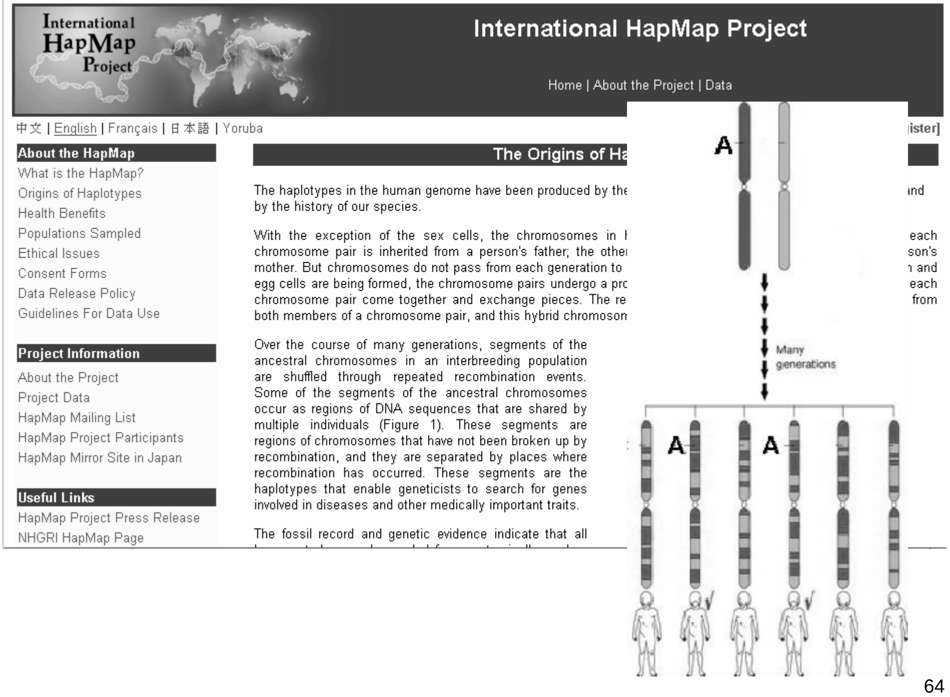
0.648*0.5^7=0.00534

0.648*0.5^7=0.00534

0.648*0.5^7=0.00534

0.975

www.hapmap.org



International HapMap Project

Home | About the Project | Data

中文 | English | Français | 日本語 | Yoruba

ister]

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 Ha

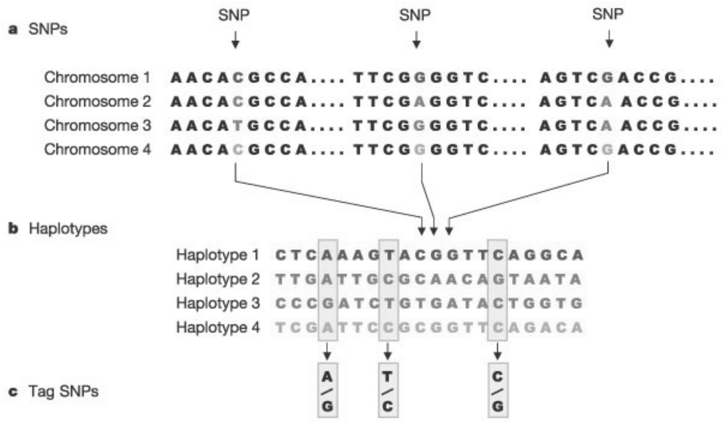
The haplotypes in the human genome have been produced by the and
by the history of our species.

With the exception of the sex cells, the chromosomes in I each
chromosome pair is inherited from a person's father; the other 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.

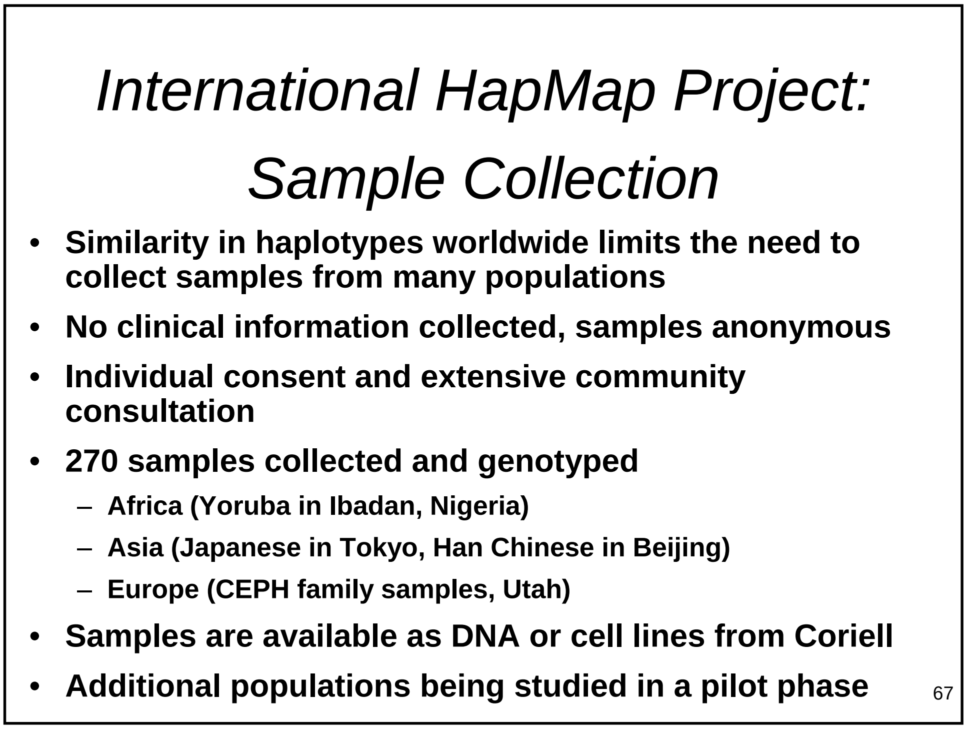
The fossil record and genetic evidence indicate that all

Identification of Haplotypes Through Genotyping


a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
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

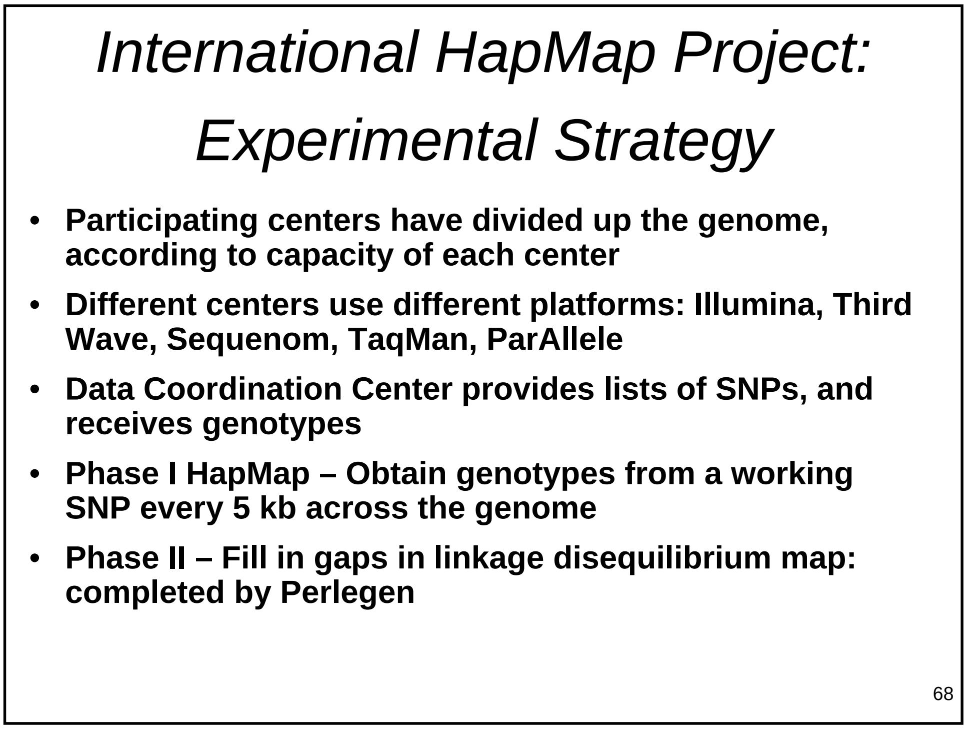
International HapMap Project

- Goal: 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 and completed in October 2005 (Phase I)
- 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



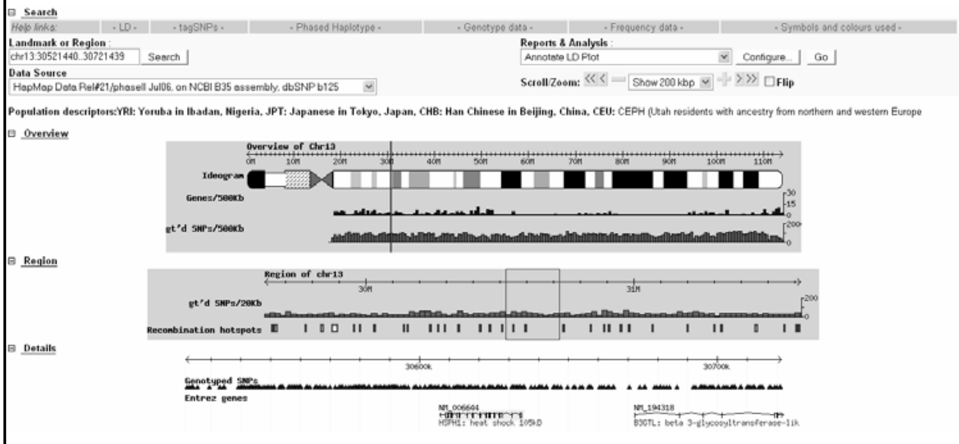
International HapMap Project: Experimental Strategy

- 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: completed by Perlegen

HapMap Milestones

- **Fall 2004 – Phase I map of 600,000 SNPs in European samples**
- **Early 2005 – Phase I map in Asian and African samples**
- **Fall 2005 – Perlegen contributes another 3M SNPs to the map**
- **Fall 2005 – Final HapMap, including gap filling**
- **"HapTag" SNPs able to represent 80-90% of common variation with**
  - **200,000 SNPs for European or Asian samples**
  - **400,000 SNPs for African samples**

HapMap Gbrowse



Landmark or Region : chr13:30521440..30721439

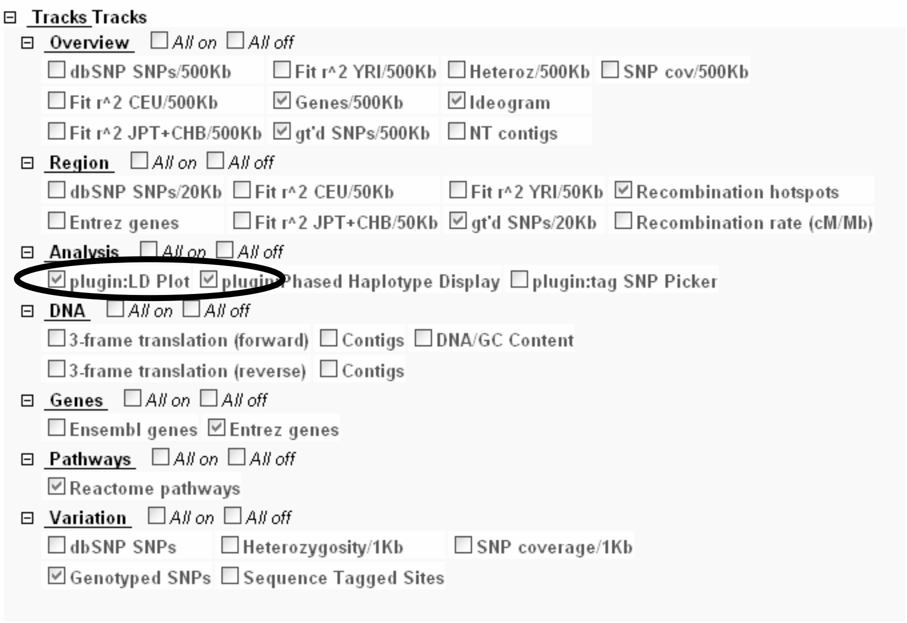
Data Source: HapMap Data Rel#21/phaseII Jul06, on NCBI B35 assembly, dbSNP b125

Reports & Analysis : Annotate LD Plot

Scroll/Zoom: Show 200 kbp □ Flip

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

http://www.hapmap.org/cgi-perl/gbrowse/hapmap_B35/



⊟ Tracks Tracks

⊟ Overview □ *All on* □ *All off*

□ dbSNP SNPs/500Kb □ Fit r^2 YRI/500Kb □ Heteroz/500Kb □ SNP cov/500Kb

□ Fit r^2 CEU/500Kb ☑ Genes/500Kb ☑ Ideogram

□ Fit r^2 JPT+CHB/500Kb ☑ gt'd SNPs/500Kb □ NT contigs

⊟ Region □ *All on* □ *All off*

□ dbSNP SNPs/20Kb □ Fit r^2 CEU/50Kb □ Fit r^2 YRI/50Kb ☑ Recombination hotspots

□ Entrez genes □ Fit r^2 JPT+CHB/50Kb ☑ gt'd SNPs/20Kb □ Recombination rate (cM/Mb)

⊟ Analysis □ *All on* □ *All off*

☑ plugin:LD Plot ☑ plugin:Phased Haplotype Display □ plugin:tag SNP Picker

⊟ DNA □ *All on* □ *All off*

□ 3-frame translation (forward) □ Contigs □ DNA/GC Content

□ 3-frame translation (reverse) □ Contigs

⊟ Genes □ *All on* □ *All off*

□ Ensembl genes ☑ Entrez genes

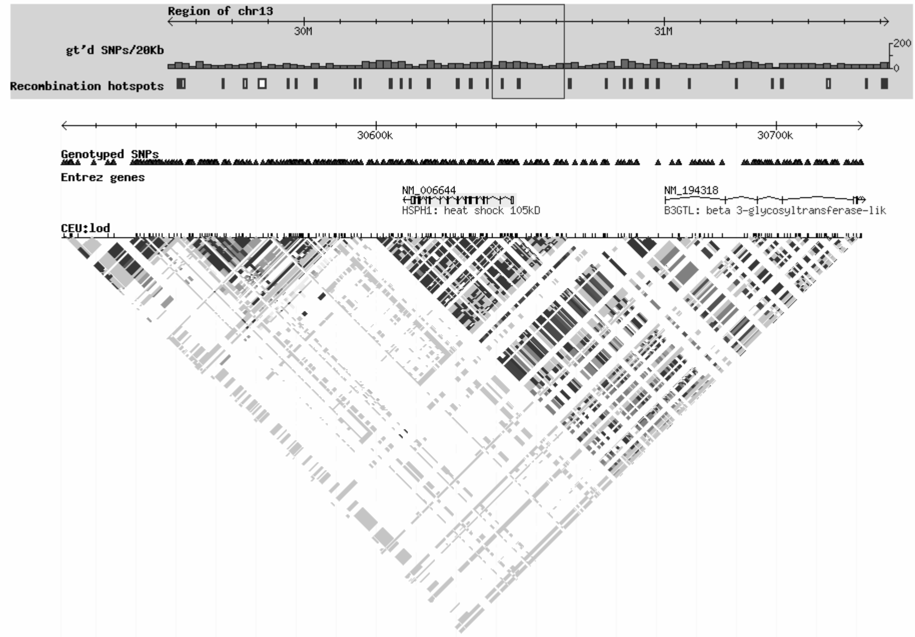
⊟ Pathways □ *All on* □ *All off*

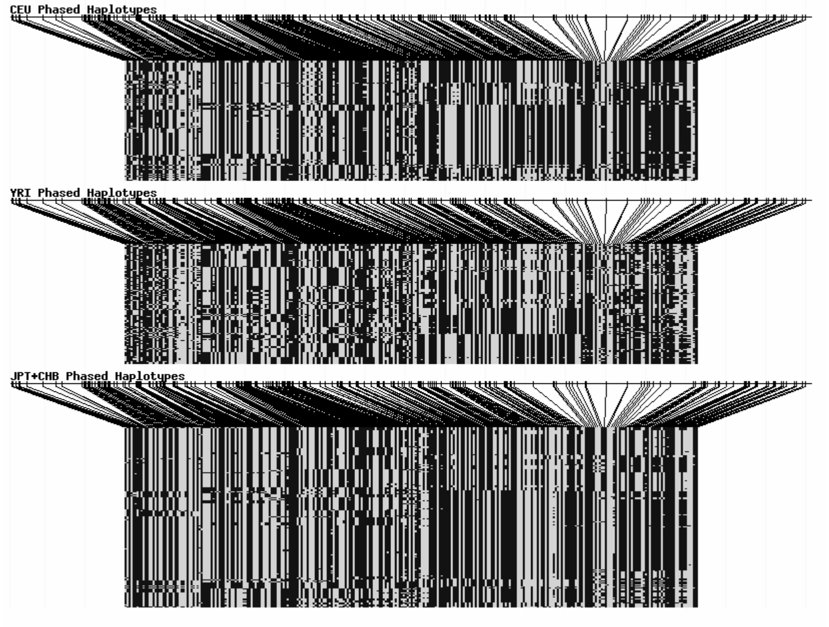
☑ Reactome pathways

⊟ Variation □ *All on* □ *All off*

□ dbSNP SNPs □ Heterozygosity/1Kb □ SNP coverage/1Kb

☑ Genotyped SNPs □ Sequence Tagged Sites





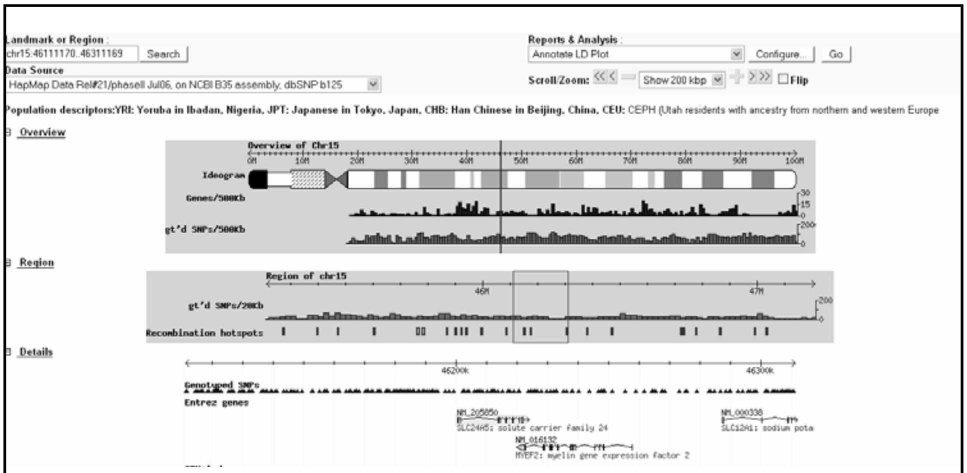
73

SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans

Rebecca L. Lamason,[1*] Manzoor-Ali P.K. Mohideen,[1†]
Jason R. Mest,[1] Andrew C. Wong,[1‡] Heather L. Norton,[6]
Michele C. Aros,[1] Michael J. Jurynec,[8] Xianyun Mao,[6]
Vanessa R. Humphreville,[1§] Jasper E. Humbert,[2,9] Soniya Sinha,[2]
Jessica L. Moore,[1∥] Pudur Jagadeeswaran,[10] Wei Zhao,[3]
Gang Ning,[7] Izabela Makalowska,[7] Paul M. McKeigue,[11]
David O'Donnell,[11] Rick Kittles,[12] Esteban J. Parra,[13]
Nancy J. Mangini,[14] David J. Grunwald,[8] Mark D. Shriver,[6]
Victor A. Canfield,[4] Keith C. Cheng[1,4,5¶]

Science 16 December 2005:
Vol. 310. no. 5755, pp. 1782 - 1786

74



Landmark or Region : chr15:46111170..46311169 Search

Data Source : HapMap Data Rel#21/phaseII Jul06, on NCBI B35 assembly, dbSNP b125

Reports & Analysis : Annotate LD Plot Configure... Go

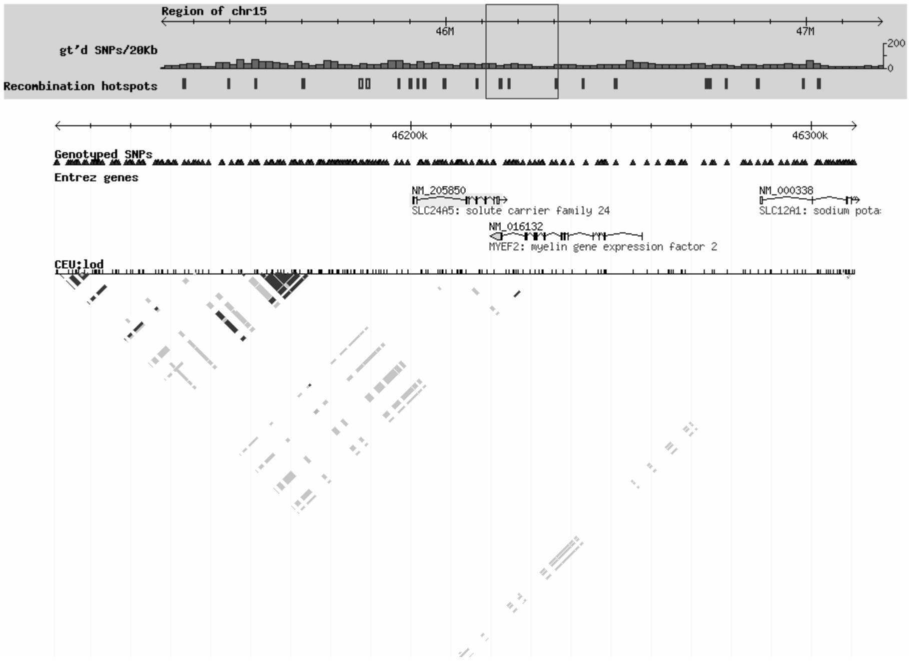
Scroll/Zoom: Show 200 kbp □ Flip

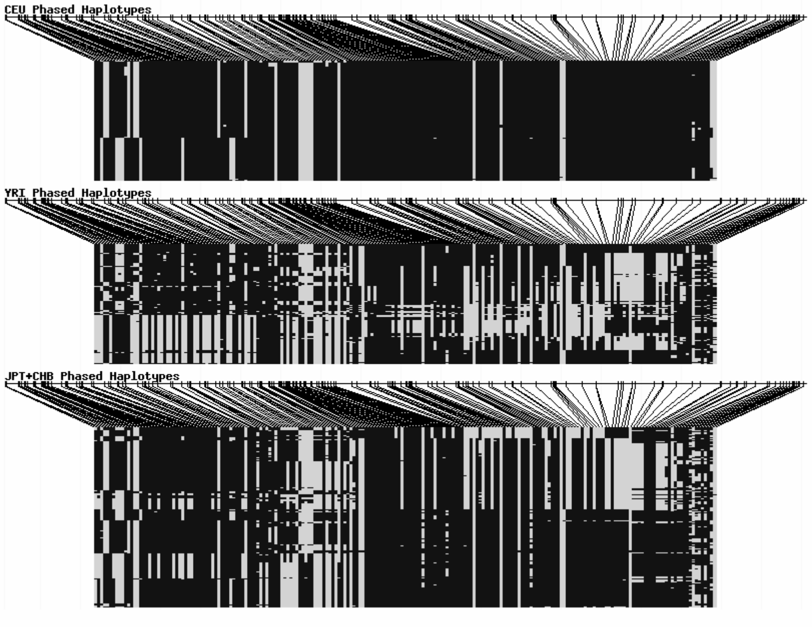
Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

⊟ Overview

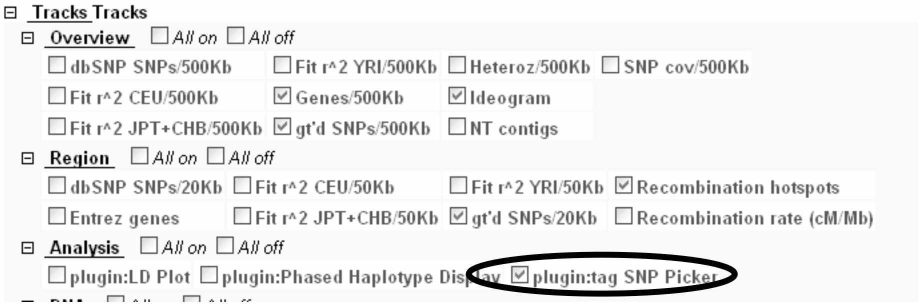
⊟ Region

⊟ Details





77


⊟ Tracks Tracks

⊟ Overview □ All on □ All off

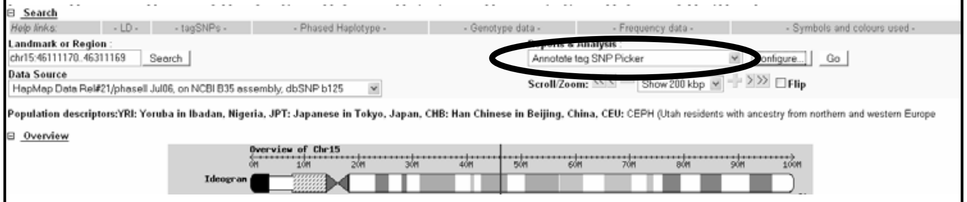
□ dbSNP SNPs/500Kb □ Fit r^2 YRI/500Kb □ Heteroz/500Kb □ SNP cov/500Kb
□ Fit r^2 CEU/500Kb ☑ Genes/500Kb ☑ Ideogram
□ Fit r^2 JPT+CHB/500Kb ☑ gt'd SNPs/500Kb □ NT contigs

⊟ Region □ All on □ All off

□ dbSNP SNPs/20Kb □ Fit r^2 CEU/50Kb □ Fit r^2 YRI/50Kb ☑ Recombination hotspots
□ Entrez genes □ Fit r^2 JPT+CHB/50Kb ☑ gt'd SNPs/20Kb □ Recombination rate (cM/Mb)

⊟ Analysis □ All on □ All off

□ plugin:LD Plot □ plugin:Phased Haplotype Display ☑ plugin:tag SNP Picker


⊟ Search

Help links: - LD - - tagSNPs - - Phased Haplotype - - Genotype data - - Frequency data - - Symbols and colours used -

Landmark or Region : chr15:46111170..46311169 Search

Data Source HapMap Data Rel#21/phaseII Jul06, on NCBI B35 assembly, dbSNP b125

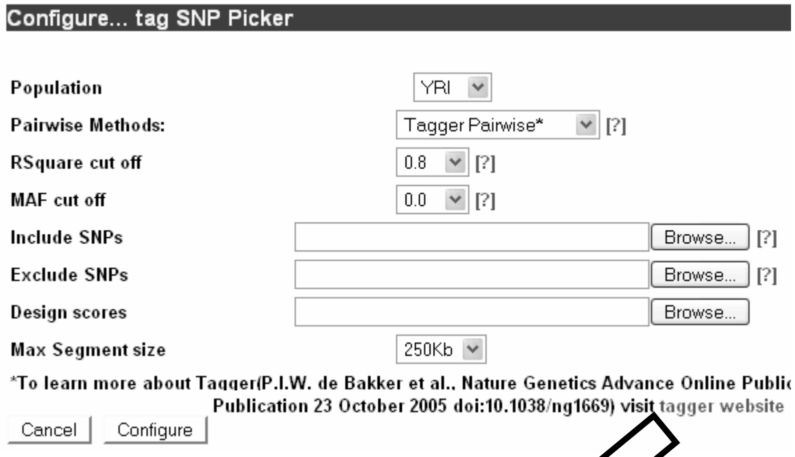
Reports & Analysis : Annotate tag SNP Picker Configure... Go

Scroll/Zoom: Show 200 kbp □ Flip

Population descriptors:YRI: Yoruba in Ibadan, Nigeria, JPT: Japanese in Tokyo, Japan, CHB: Han Chinese in Beijing, China, CEU: CEPH (Utah residents with ancestry from northern and western Europe

⊟ Overview

78

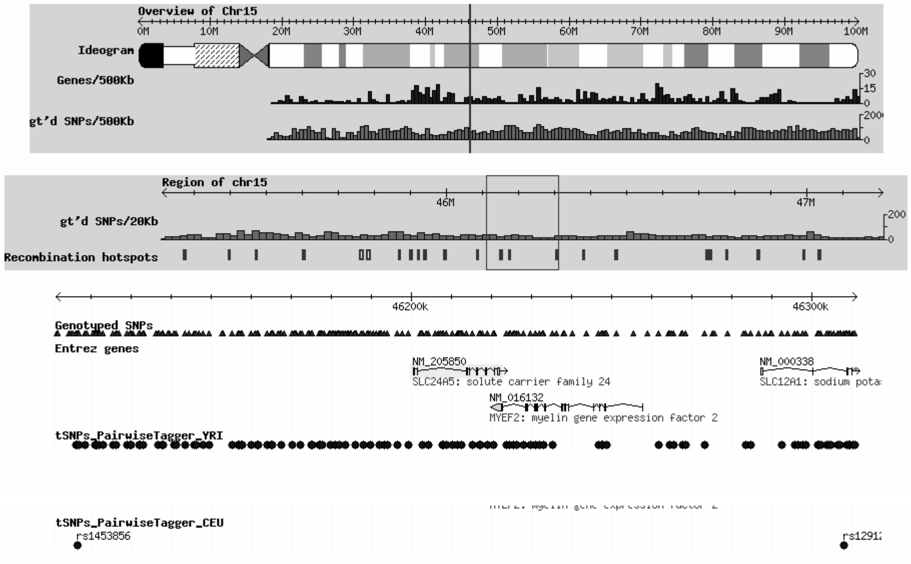


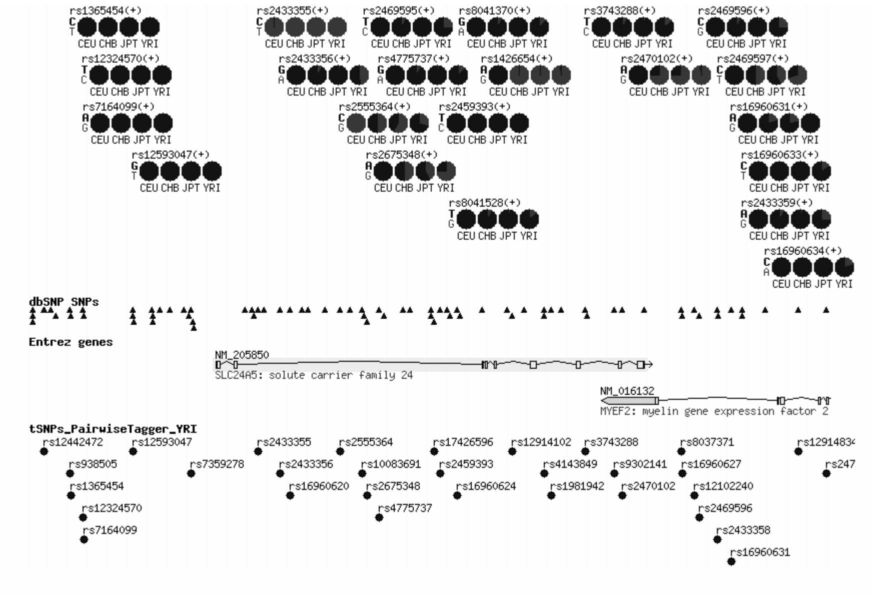
Configure... tag SNP Picker

Population	YRI
Pairwise Methods:	Tagger Pairwise* [?]
RSquare cut off	0.8 [?]
MAF cut off	0.0 [?]
Include SNPs	Browse... [?]
Exclude SNPs	Browse... [?]
Design scores	Browse...
Max Segment size	250Kb

*To learn more about Tagger(P.I.W. de Bakker et al., Nature Genetics Advance Online Public
Publication 23 October 2005 doi:10.1038/ng1669) visit tagger website

http://www.broad.mit.edu/mpg/tagger/





75

31

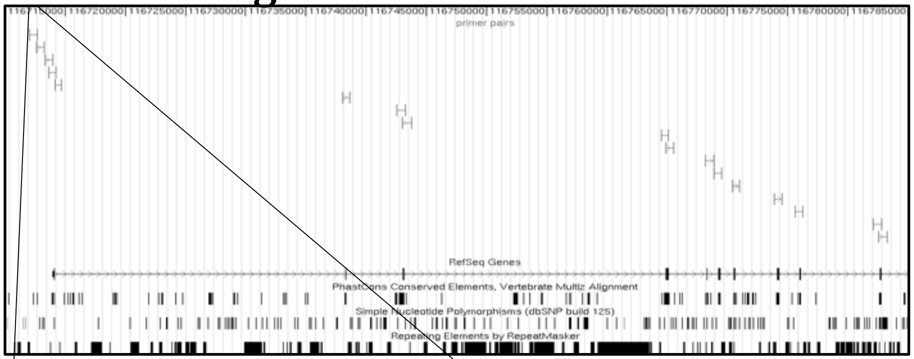
Overview of Topics

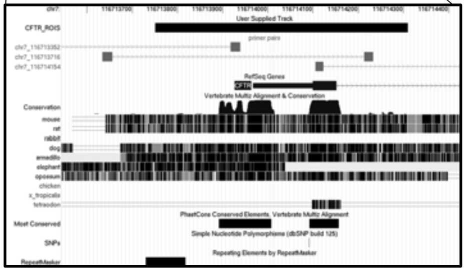
- Genome variation origins
- Types of polymorphisms
- SNP discovery methods
- Access to genetic variation data
- How to find SNPs in a region of interest
- Haplotype Map project
- Medical Sequencing
- SNPs for Other Species
- New Sequencing Technologies

A Brief Tour of a Medical Sequencing Pipeline

83

Primer Design



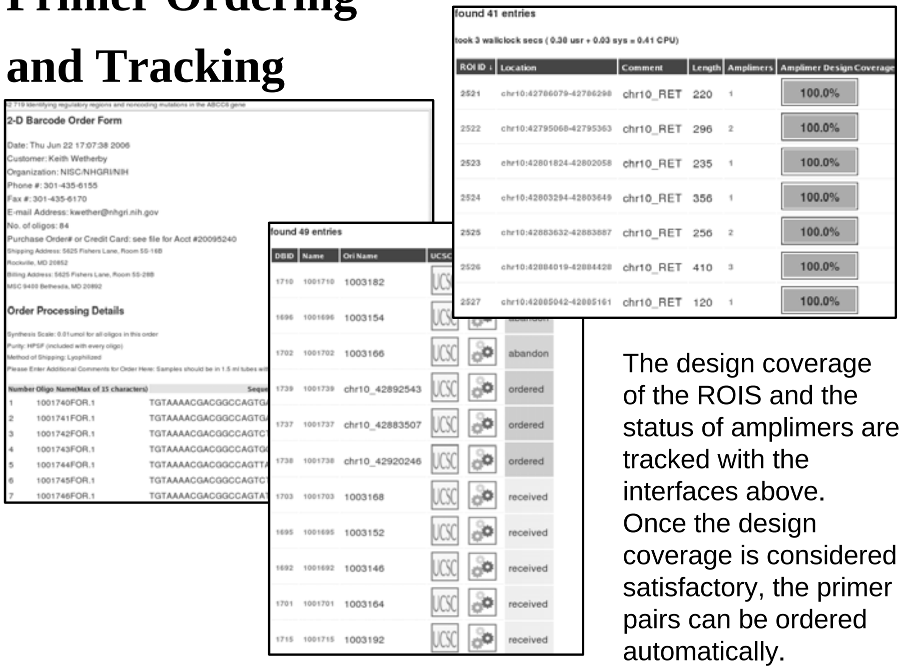


<u>Choice of Genomic Regions</u>

The regions of interest (ROIs) are typically defined by their biological context (coding, conservation, regulatory function, known variation). When features are in close proximity, the number of amplimers is automatically reduced, maintaining optimal coverage.

84

Primer Ordering and Tracking



2 719 Identifying regulatory regions and noncoding mutations in the ABCC6 gene

2-D Barcode Order Form

Date: Thu Jun 22 17:07:38 2006
Customer: Keith Wetherby
Organization: NISC/NHGRI/NIH
Phone #: 301-435-6155
Fax #: 301-435-6170
E-mail Address: kwether@nhgri.nih.gov
No. of oligos: 84
Purchase Order# or Credit Card: see file for Acct #20095240
Shipping Address: 5625 Fishers Lane, Room 5S-16B
Rockville, MD 20852
Billing Address: 5625 Fishers Lane, Room 5S-28B
MSC 9400 Bethesda, MD 20892

Order Processing Details

Synthesis Scale: 0.01umol for all oligos in this order
Purity: HPSF (included with every oligo)
Method of Shipping: Lyophilized
Please Enter Additional Comments for Order Here: Samples should be in 1.5 ml tubes wit

Number	Oligo Name(Max of 15 characters)	Seque
1	1001740FOR.1	TGTAAAACGACGGCCAGTGA
2	1001741FOR.1	TGTAAAACGACGGCCAGTGA
3	1001742FOR.1	TGTAAAACGACGGCCAGTCT
4	1001743FOR.1	TGTAAAACGACGGCCAGTGG
5	1001744FOR.1	TGTAAAACGACGGCCAGTTA
6	1001745FOR.1	TGTAAAACGACGGCCAGTCT
7	1001746FOR.1	TGTAAAACGACGGCCAGTAT

found 49 entries

DBID	Name	Ori Name	UCSC		
1710	1001710	1003182	UCSC		
1696	1001696	1003154	UCSC		abandon
1702	1001702	1003166	UCSC		abandon
1739	1001739	chr10_42892543	UCSC		ordered
1737	1001737	chr10_42883507	UCSC		ordered
1738	1001738	chr10_42920246	UCSC		ordered
1703	1001703	1003168	UCSC		received
1695	1001695	1003152	UCSC		received
1692	1001692	1003146	UCSC		received
1701	1001701	1003164	UCSC		received
1715	1001715	1003192	UCSC		received

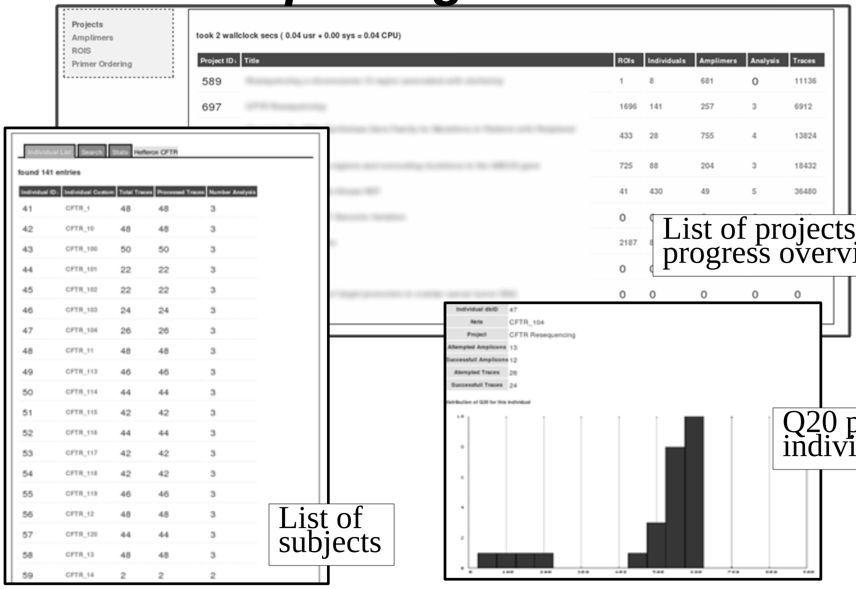
found 41 entries

took 3 wallclock secs (0.38 usr + 0.03 sys = 0.41 CPU)

ROI ID ↓	Location	Comment	Length	Amplimers	Amplimer Design Coverage
2521	chr10:42786079-42786298	chr10_RET	220	1	100.0%
2522	chr10:42795068-42795363	chr10_RET	296	2	100.0%
2523	chr10:42801824-42802058	chr10_RET	235	1	100.0%
2524	chr10:42803294-42803649	chr10_RET	356	1	100.0%
2525	chr10:42883632-42883887	chr10_RET	256	2	100.0%
2526	chr10:42884019-42884428	chr10_RET	410	3	100.0%
2527	chr10:42885042-42885161	chr10_RET	120	1	100.0%

The design coverage of the ROIS and the status of amplimers are tracked with the interfaces above. Once the design coverage is considered satisfactory, the primer pairs can be ordered automatically.

Exploring the data



took 2 wallclock secs (0.04 usr + 0.00 sys = 0.04 CPU)

Project ID↓	Title	ROIs	Individuals	Amplimers	Analysis	Traces
589	[illegible]	1	8	681	0	11136
697	[illegible]	1696	141	257	3	6912
	[illegible]	433	28	755	4	13824
	[illegible]	725	88	204	3	18432
	[illegible]	41	430	49	5	36480
	[illegible]	0	0			
	[illegible]	2187	8			
		0	0			
	[illegible]	0	0	0	0	0

List of projects and progress overview

found 141 entries

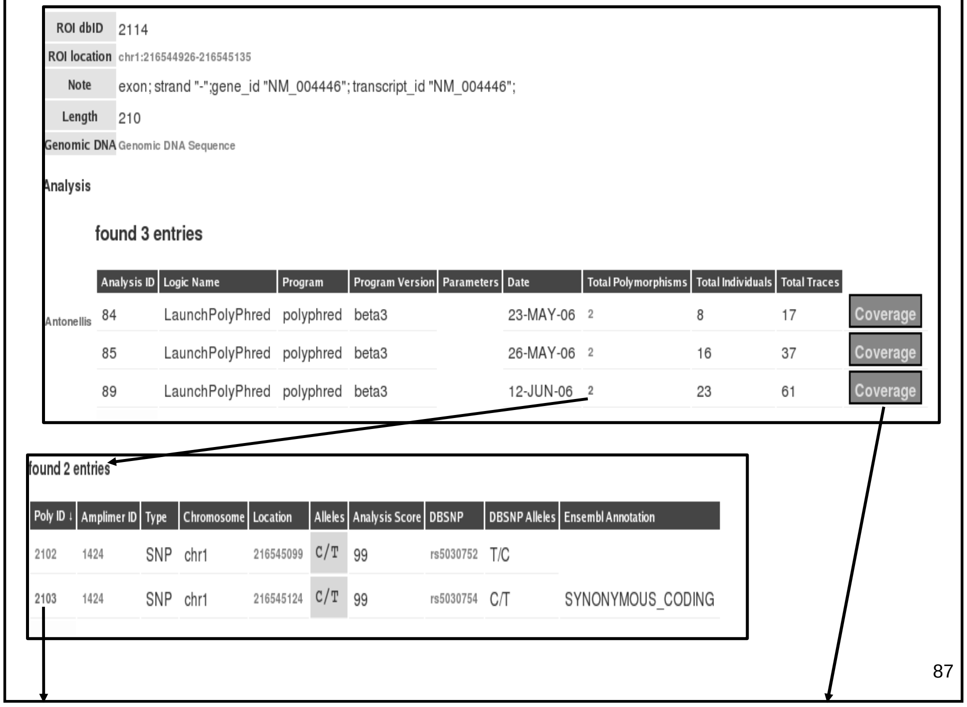
Individual ID↓	Individual Custom	Total Traces	Processed Traces	Number Analysis
41	CFTR_1	48	48	3
42	CFTR_10	48	48	3
43	CFTR_100	50	50	3
44	CFTR_101	22	22	3
45	CFTR_102	22	22	3
46	CFTR_103	24	24	3
47	CFTR_104	26	26	3
48	CFTR_11	48	48	3
49	CFTR_113	46	46	3
50	CFTR_114	44	44	3
51	CFTR_115	42	42	3
52	CFTR_116	44	44	3
53	CFTR_117	42	42	3
54	CFTR_118	42	42	3
55	CFTR_119	46	46	3
56	CFTR_12	48	48	3
57	CFTR_120	44	44	3
58	CFTR_13	48	48	3
59	CFTR_14	2	2	2

List of subjects

Individual dbID	47
Note	CFTR_104
Project	CFTR Resequencing
Attempted Amplicons	13
Successfull Amplicons	12
Atempted Traces	26
Successfull Traces	24

Distribution of Q20 for this individual

Q20 per individual



ROI dbID	2114
ROI location	chr1:216544926-216545135
Note	exon; strand "-";gene_id "NM_004446"; transcript_id "NM_004446";
Length	210
Genomic DNA	Genomic DNA Sequence

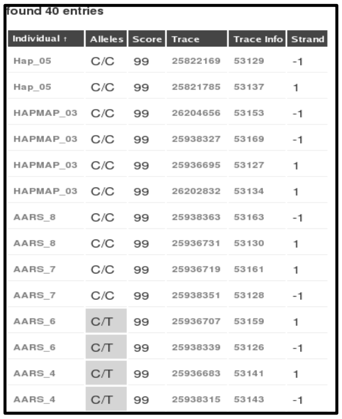
Analysis

found 3 entries

	Analysis ID	Logic Name	Program	Program Version	Parameters	Date	Total Polymorphisms	Total Individuals	Total Traces	
Antonellis	84	LaunchPolyPhred	polyphred	beta3		23-MAY-06	2	8	17	Coverage
	85	LaunchPolyPhred	polyphred	beta3		26-MAY-06	2	16	37	Coverage
	89	LaunchPolyPhred	polyphred	beta3		12-JUN-06	2	23	61	Coverage

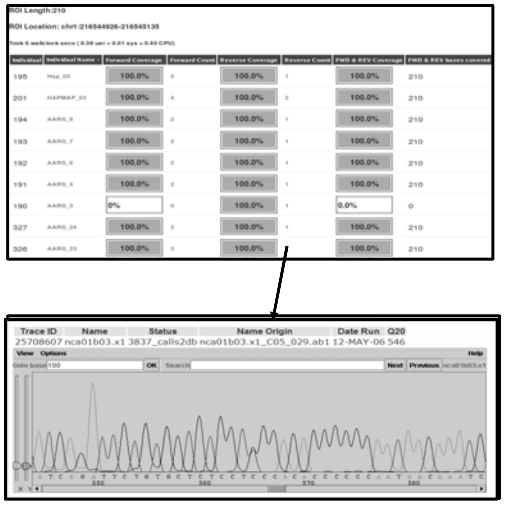
found 2 entries

Poly ID ↓	Amplimer ID	Type	Chromosome	Location	Alleles	Analysis Score	DBSNP	DBSNP Alleles	Ensembl Annotation
2102	1424	SNP	chr1	216545099	C/T	99	rs5030752	T/C	
2103	1424	SNP	chr1	216545124	C/T	99	rs5030754	C/T	SYNONYMOUS_CODING



found 40 entries

Individual ↑	Alleles	Score	Trace	Trace Info	Strand
Hap_05	C/C	99	25822169	53129	-1
Hap_05	C/C	99	25821785	53137	1
HAPMAP_03	C/C	99	26204656	53153	-1
HAPMAP_03	C/C	99	25938327	53169	-1
HAPMAP_03	C/C	99	25936695	53127	1
HAPMAP_03	C/C	99	26202832	53134	1
AARS_8	C/C	99	25938363	53163	-1
AARS_8	C/C	99	25936731	53130	1
AARS_7	C/C	99	25936719	53161	1
AARS_7	C/C	99	25938351	53128	-1
AARS_6	C/T	99	25936707	53159	1
AARS_6	C/T	99	25938339	53126	-1
AARS_4	C/T	99	25936683	53141	1
AARS_4	C/T	99	25938315	53143	-1



ROI Length:210

ROI Location: chr1:216544926-216545135

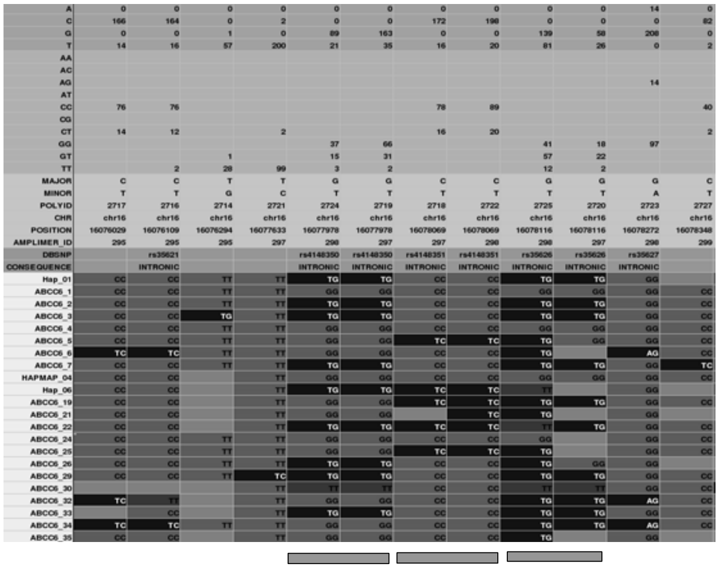
Took 6 wallclock secs (0.39 usr + 0.01 sys = 0.40 CPU)

Individual	Individual Name ↑	Forward Coverage	Forward Count	Reverse Coverage	Reverse Count	FWD & REV Coverage	FWD & REV bases covered
195	Hap_05	100.0%	2	100.0%	1	100.0%	210
201	HAPMAP_03	100.0%	4	100.0%	2	100.0%	210
194	AARS_8	100.0%	2	100.0%	1	100.0%	210
193	AARS_7	100.0%	2	100.0%	1	100.0%	210
192	AARS_6	100.0%	2	100.0%	1	100.0%	210
191	AARS_4	100.0%	2	100.0%	1	100.0%	210
190	AARS_3	0%	0	100.0%	1	0.0%	0
327	AARS_24	100.0%	2	100.0%	1	100.0%	210
326	AARS_23	100.0%	2	100.0%	[illegible]	100.0%	210

Trace ID	Name	Status	Name Origin	Date Run	Q20
25708607	nca01b03.x1	3837_calls2db	nca01b03.x1_C05_029.ab1	12-MAY-06	546

View
Options
Help
Goto base 100
OK
Search
Next
Previous
nca01b03.x1
A T C A G A T T C T G T G C T C T C C T C C C A C A C C C C C C A A T A A C A A A T C
550
560
570
580

The system keeps track of analysis performed on the data and coverage attained for each ROI. It also allows a user to browse the detected genotypes.

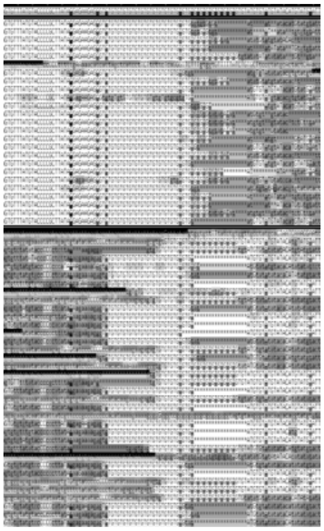


A	0	0	0	0	0	0	0	0	0	0	14	0
C	166	164	0	2	0	0	172	198	0	0	0	82
G	0	0	1	0	89	163	0	0	139	58	208	0
T	14	16	57	200	21	35	16	20	81	26	0	2
AA												
AC												
AG											14	
AT												
CC	76	76					78	89				40
CG												
CT	14	12		2			16	20				2
GG					37	66			41	18	97	
GT			1		15	31			57	22		
TT		2	28	99	3	2			12	2		
MAJOR	C	C	T	T	G	G	C	C	G	G	G	C
MINOR	T	T	G	C	T	T	T	T	T	T	A	T
POLYID	2717	2716	2714	2721	2724	2719	2718	2722	2725	2720	2723	2727
CHR	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16	chr16
POSITION	16076029	16076109	16076294	16077633	16077978	16077978	16078069	16078069	16078116	16078116	16078272	16078348
AMPLIMER_ID	295	295	295	297	298	297	297	298	298	297	298	299
DBSNP		rs35621			rs4148350	rs4148350	rs4148351	rs4148351	rs35626	rs35626	rs35627	
CONSEQUENCE		INTRONIC			INTRONIC	INTRONIC	INTRONIC	INTRONIC	INTRONIC	INTRONIC	INTRONIC	
Hap_01	CC	CC	TT	TT	TG	TG	CC	CC	TG	TG	GG	
ABCC6_1	CC	CC	TT	TT	GG	GG	CC	CC	GG	GG	GG	CC
ABCC6_2	CC	CC	TT	TT	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_3	CC	CC	TG	TT	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_4	CC	CC	TT	TT	GG	GG	CC	CC	GG	GG	GG	CC
ABCC6_5	CC	CC	TT	TT	GG	GG	TC	TC	TG	GG	GG	CC
ABCC6_6	TC	TC	TT	TT	GG	GG	CC	CC	TG		AG	CC
ABCC6_7	CC	CC	TT	TT	TG	TG	CC	CC	TG	TG	GG	TC
HAPMAP_04	CC	CC		TT	GG	GG	CC	CC	GG	GG	GG	CC
Hap_06	CC	CC		TT	TG	TG	TC	TC	TT		GG	
ABCC6_19	CC	CC		TT	GG	GG	TC	TC	TG	TG	GG	CC
ABCC6_21	CC	CC		TT	GG	GG		TC	TG		GG	
ABCC6_22	CC	CC		TT	TG	TG	TC	TC	TT	TG	GG	CC
ABCC6_24	CC	CC	TT	TT	GG	GG	CC	CC	GG		GG	CC
ABCC6_25	CC	CC	TT	TT	GG	GG	TC	TC	TG		GG	CC
ABCC6_26	CC	CC	TT	TT	TG	TG	CC	CC	TG	GG	GG	
ABCC6_29	CC	CC	TT	TC	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_30				TT	TT	TT	CC	CC	TT	TT	GG	CC
ABCC6_32	TC	TT		TT	GG	GG	CC	CC	TG	TG	AG	CC
ABCC6_33		CC		TT	TG	TG	CC	CC	TG	TG	GG	CC
ABCC6_34	TC	TC	TT	TT	GG	GG	CC	CC	TG	TG	AG	CC
ABCC6_35	CC	CC		TT	GG	GG	CC	CC	TG		GG	

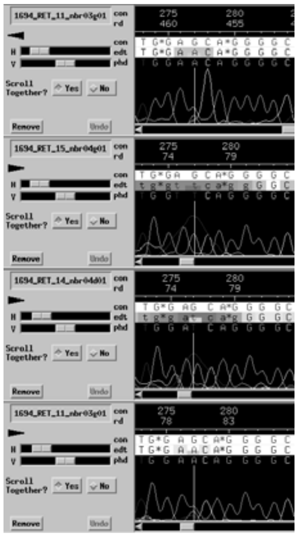
We are developing interfaces that allow exploring the results and identify interesting results as well flag problems.

Three examples of same SNP detected in overlapping amplimers. This information is used to assess accuracy of the detection.

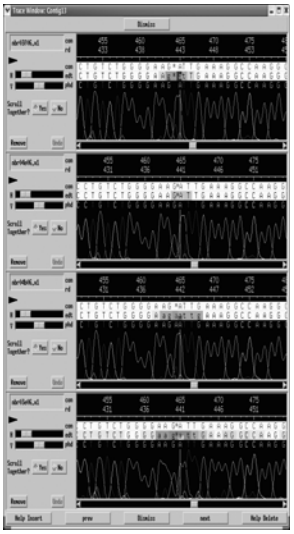
Some of the challenges of variation detection



INDEL



“Dye blob”



Detection saturation

SNPs for Other Species

- Mouse
  - The reference strain sequenced, C57BL/6J, was inbred for sufficient generations to result in a homozygous genome, however, 15 mouse strains have been sequenced and the variations are available from dbSNP (http://www.nih.gov/news/pr/oct2006/niehs-25.htm)
  - This is a great resource for mouse genetics. For example, crossing two different mouse strains where one mouse has given disease causing mutation.
- Dog
  - The reference dog genome sequence comes from a fairly inbred individual (a boxer named Tasha). This individual is 60% homozygous with the heterozygous regions showing 1SNP per 900 bases, giving 770k SNPs.
  - Celera sequenced a poodle, Shadow, and comparing this genome to Tasha's sequence give 1.46M SNPs
  - The public sequencing effort also generated whole genome shotgun sequence from 9 other dogs breeds as well as 4 wolves and a coyote

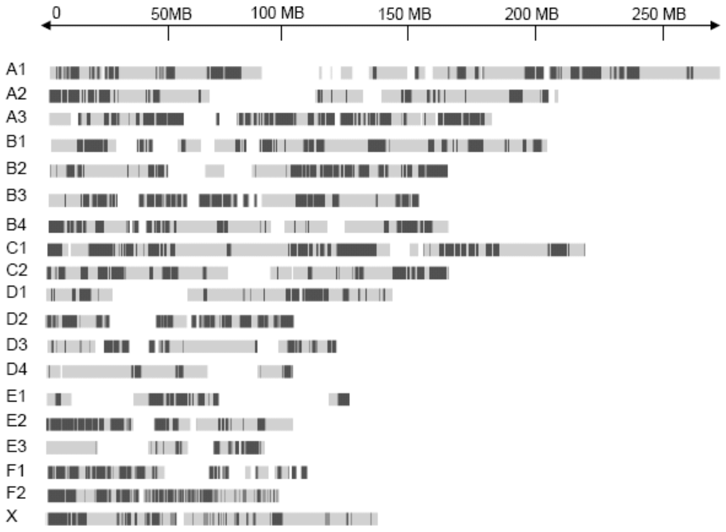
SNPs for Other Species

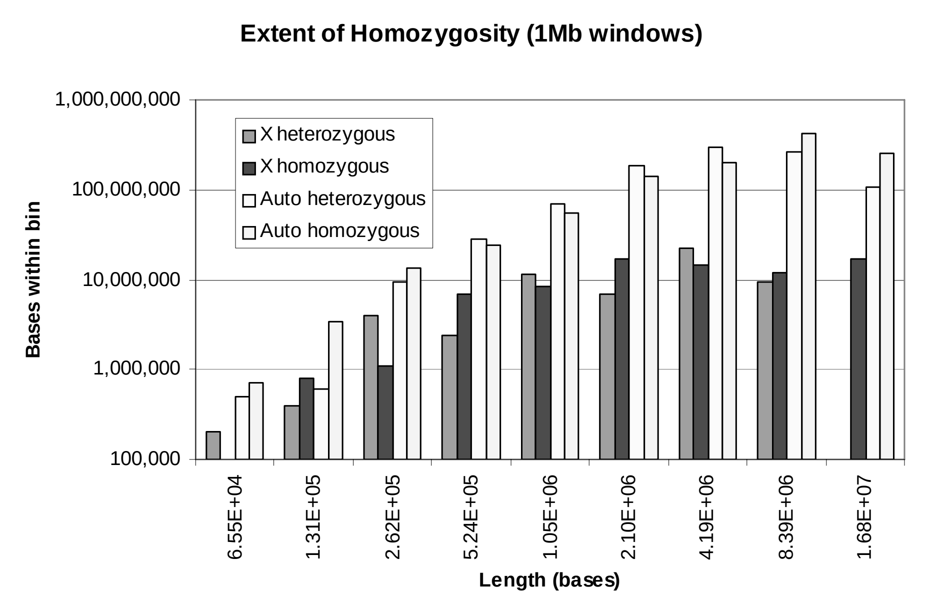
- Chimpanzee
  - The reference sequence in based on Clint along with light WGS of four other West African and three central African chimpanzees giving a total of 1.66M SNPs.
  - Chimpanzee sequence can also be used together with human SNPs to determine the ancestral allele state, as noted in many of the dbSNP records.
- Cat
  - The reference cat sequence, like dog, comes from an inbred individual (an Abyssinian named Cinnamon) which is also about 60% homozygous, with the heterozygous regions showing 1 SNP per 600 bases.

Cat SNP Analysis

- Cinnamon is of the Abyssinian breed, and its genome is diploid
- Thus, when two sequence traces overlap, there is a 50% chance that these two traces came from different chromosomes
- If Cinnamon were an out-bred cat, then traces that arise from different chromosomes should exhibit sequence polymorphisms
- However, due to inbreeding, the locus of these two chromosomes may have been derived from an ancestor’s chromosome only a few generations back, thus exhibiting no polymorphisms

Heterozygosity Profile of Cinnamon



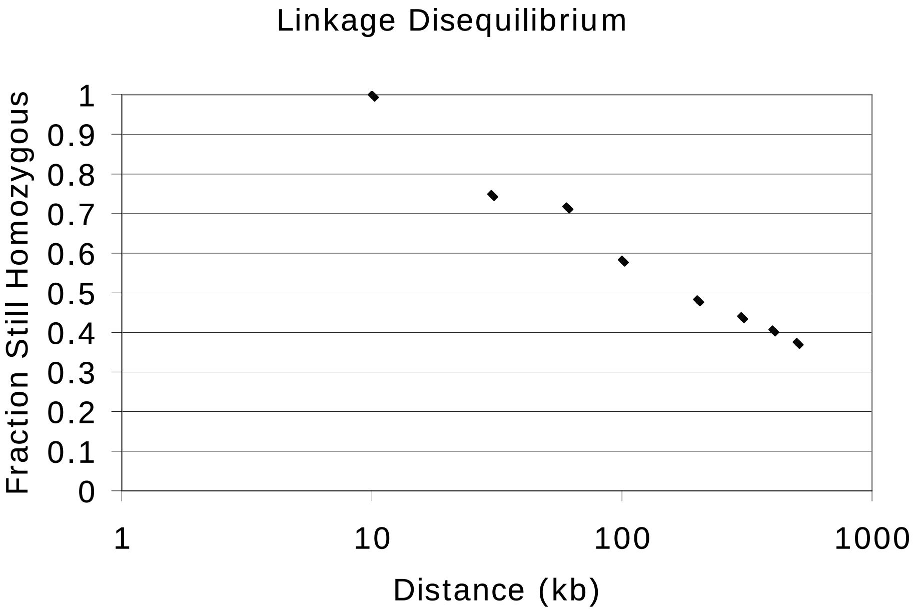


Cinnamon's Polymorphism Statistics

- 57% of Cinnamon's autosomes are homozygous
- Within the heterozygous segments of this individual, we discovered over 325,000 SNPs and over 37,000 deletion/insertion polymorphisms
- The heterozygosity level of heterozygous regions is 0.17%, or about 70% higher than human heterozygosity levels
- Comparing Cinnamon to another cat (Gus), a brown classic tabby (RPCI-86), yields a heterozygosity level of about 0.2%, or about twice the level of humans.

Linkage Disequilibrium Across Cat Breeds

- Selected SNPs detected from Cinnamon’s genome within heterozygous regions on 10 different chromosomes.
- 35 SNPs were selected per chromosome, with the first 8 SNPs within a 15kb window and rest selected every approximately every 15kb away from the previous SNP.
- These SNPs were genotyped across 97 cats from 24 breeds, 7 outbred “alley” cats and 12 wild species.
- Linkage disequilibrium (LD) was calculated for those individuals that were homozygous within the first 15kb window, and the length of LD was derived from the extent of the homozygous interval.



Summary of Cat LD Results

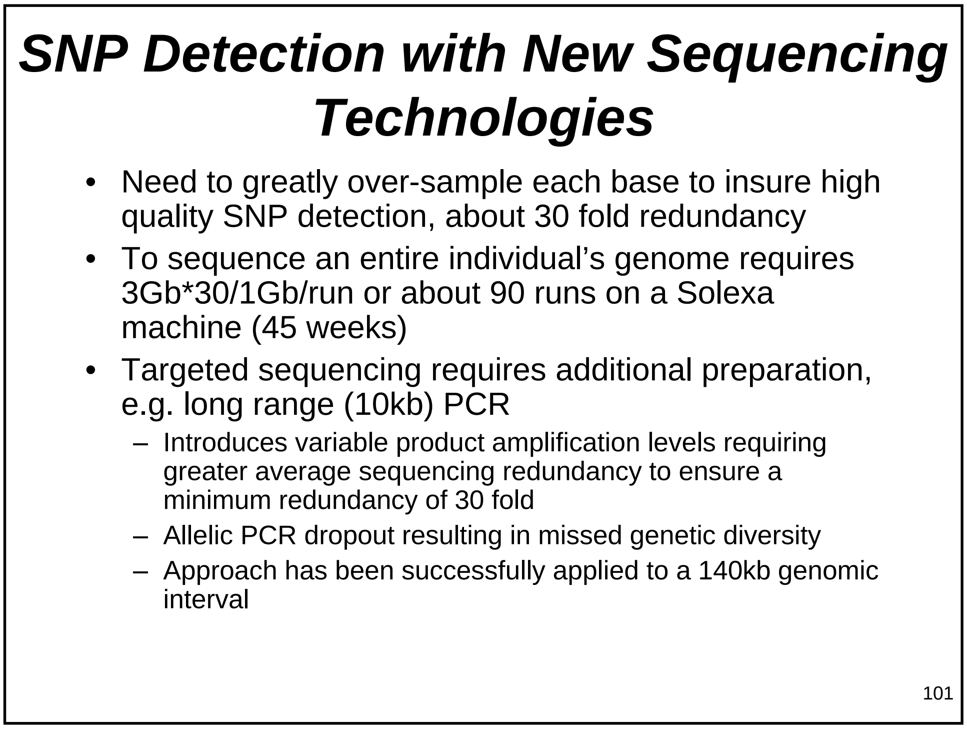
- ~60% of 10 kb regions are homozygous within an individual. This is very similar to dogs.

- Conditional on being homozygous within the 10 kb region, 50% of cases are still homozygous at 150 kb. The extent of linkage disequilibrium is roughly a third that in dogs.

- The number of markers needed for genome-wide association: current estimate about 45k markers.

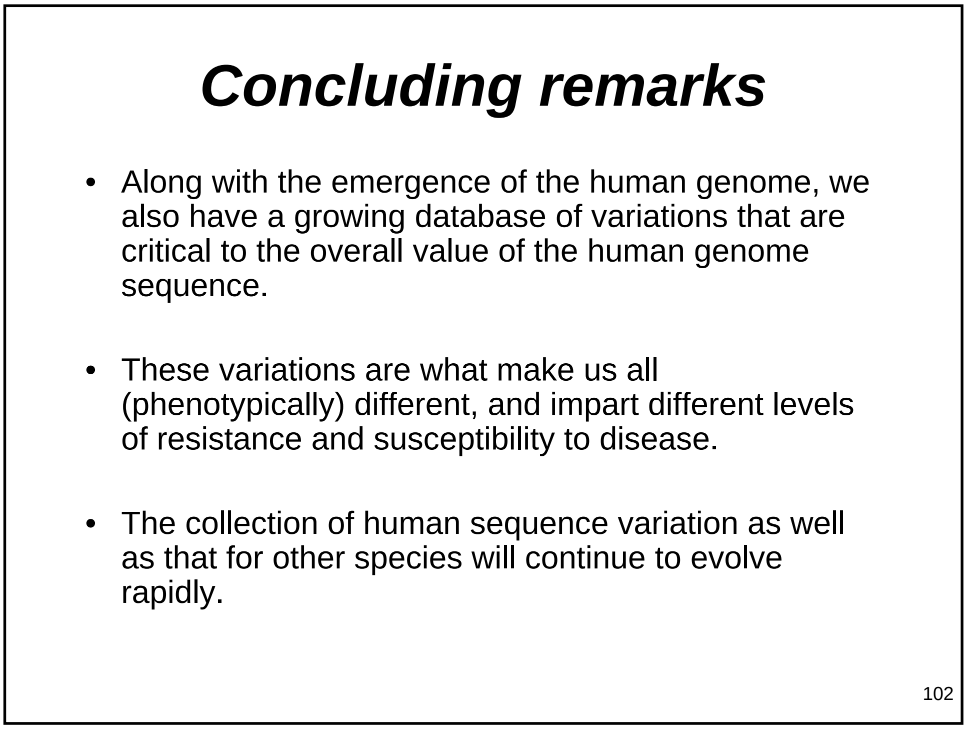
New Sequencing Technologies

- 454 Life Sciences
  - 100-200 base reads
  - 20-40Mb per run
  - 2 runs per day
- Solexa
  - 25-40 base reads
  - 8*125Mb per run
  - 2 runs per week
- ABI SOLiD
  - Similar to Solexa
  - Run performance like Solexa



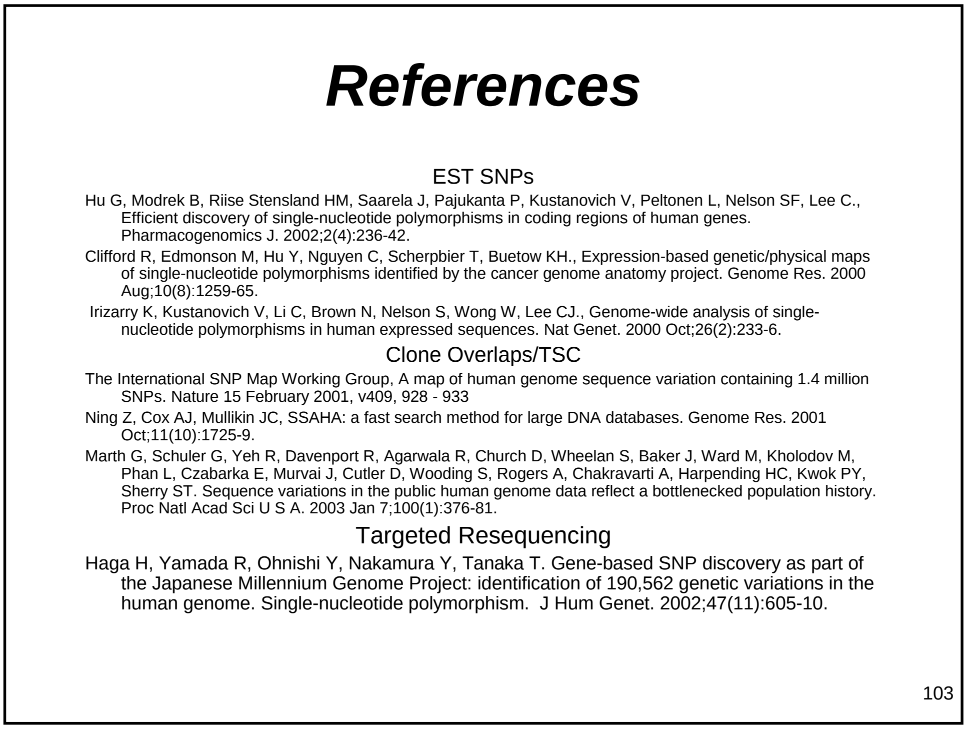
SNP Detection with New Sequencing Technologies

- Need to greatly over-sample each base to insure high quality SNP detection, about 30 fold redundancy
- To sequence an entire individual’s genome requires 3Gb*30/1Gb/run or about 90 runs on a Solexa machine (45 weeks)
- Targeted sequencing requires additional preparation, e.g. long range (10kb) PCR
  - Introduces variable product amplification levels requiring greater average sequencing redundancy to ensure a minimum redundancy of 30 fold
  - Allelic PCR dropout resulting in missed genetic diversity
  - Approach has been successfully applied to a 140kb genomic interval



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 as well as that for other species 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

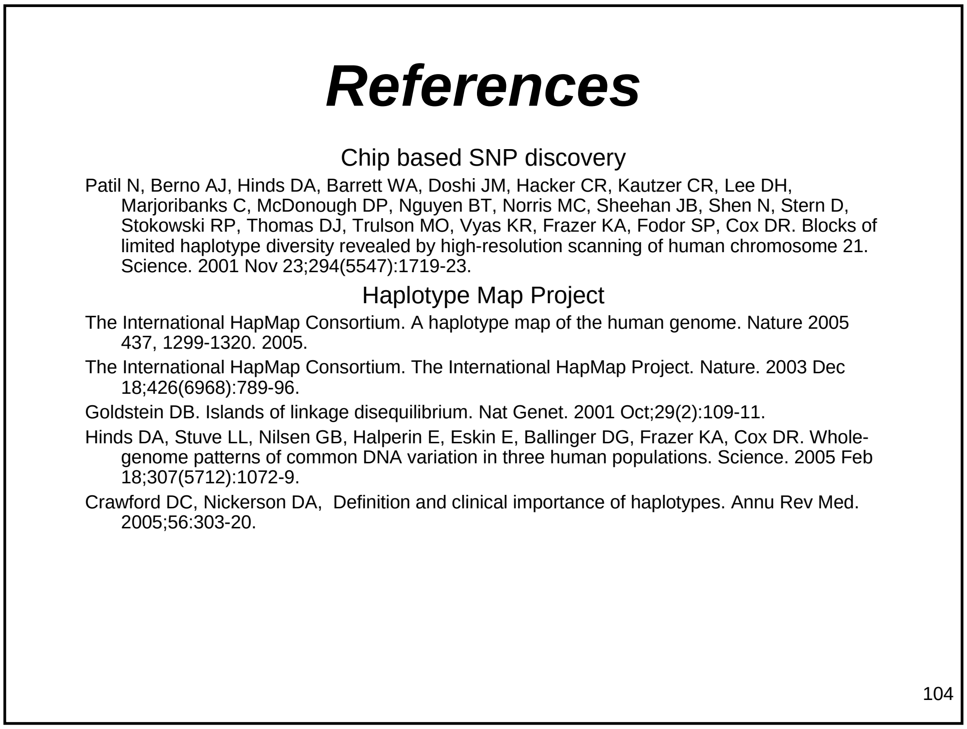
The International SNP Map Working Group, A map of human genome sequence variation containing 1.4 million SNPs. Nature 15 February 2001, v409, 928 - 933

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.

Haplotype Map Project

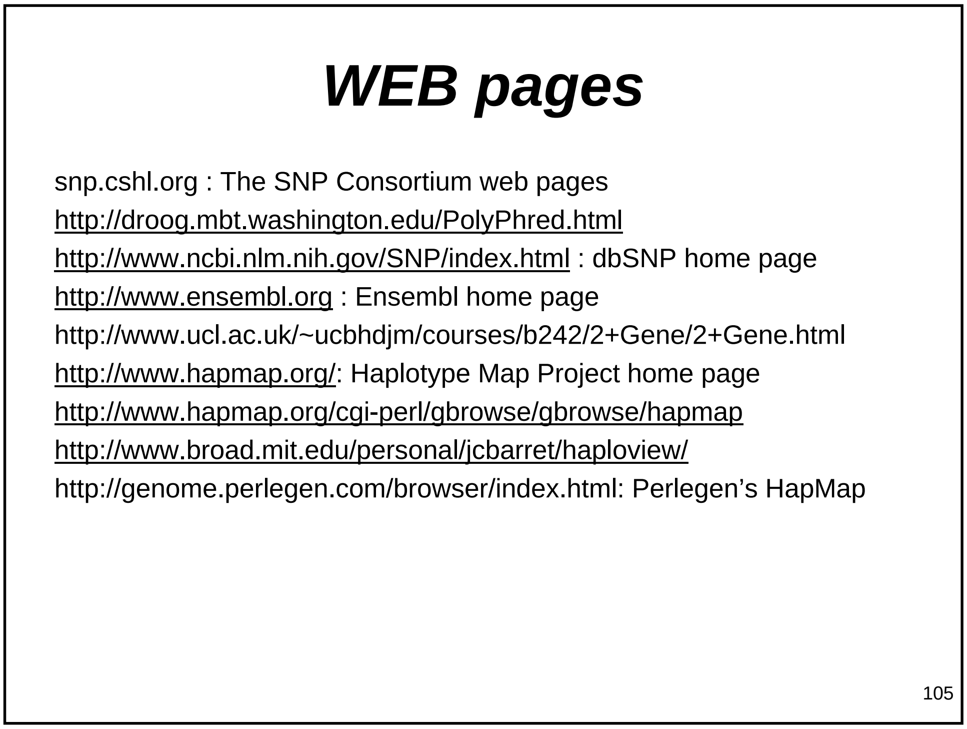
The International HapMap Consortium. A haplotype map of the human genome. Nature 2005 437, 1299-1320. 2005.

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