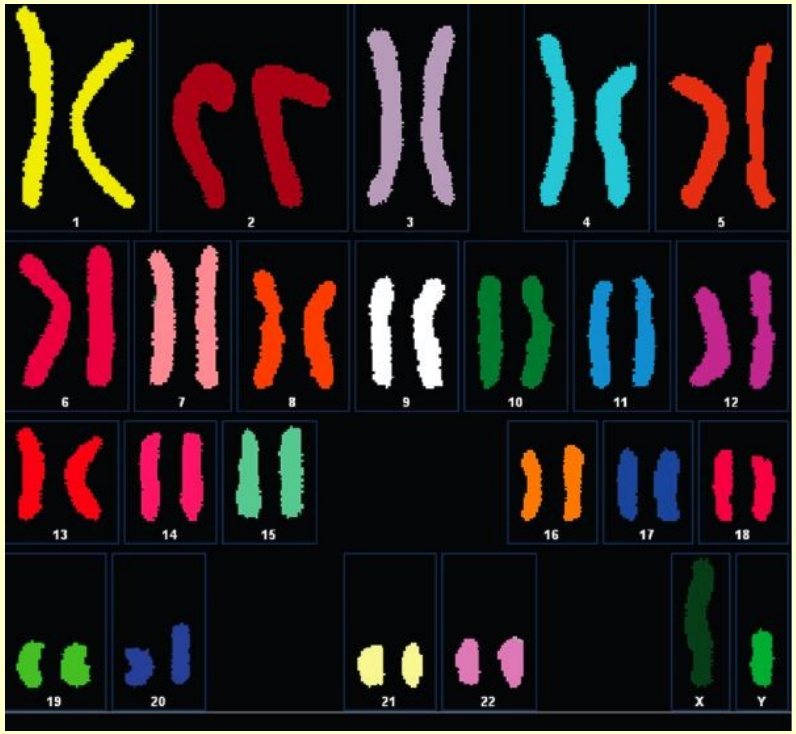
Computational Molecular Biology
Biochem 218 – BioMedical Informatics 231

http://biochem218.stanford.edu/

Simple Nucleotide Polymorphisms (SNPs)



Doug Brutlag
Professor Emeritus
Biochemistry & Medicine (by courtesy)

Simple Nucleotide Polymorphisms (SNPs)

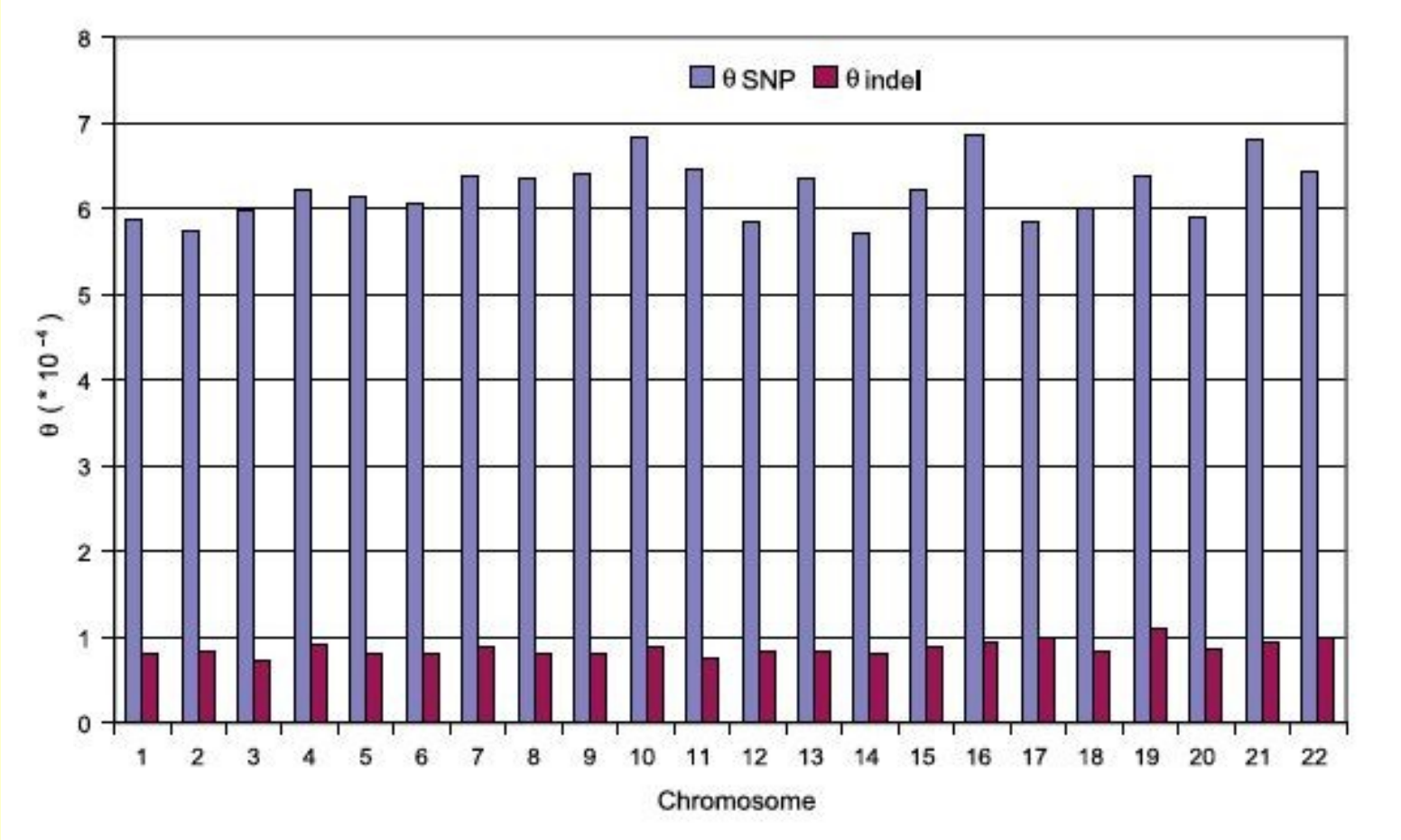
~ 15 million sites in the human genome where SNPs could occur (~ 10 million are common, MAF > 5%)

~ 3.8 million SNPs catalogued in 270 individuals
(~ 1 SNP / 1000 bases)
(International HapMap Consortium, October 2007)

~ Each person has ~ 3-5 million common SNPs (inherited) and ~30 new mutations have arisen
{mutation rate of 10^{-8} / base / generation}

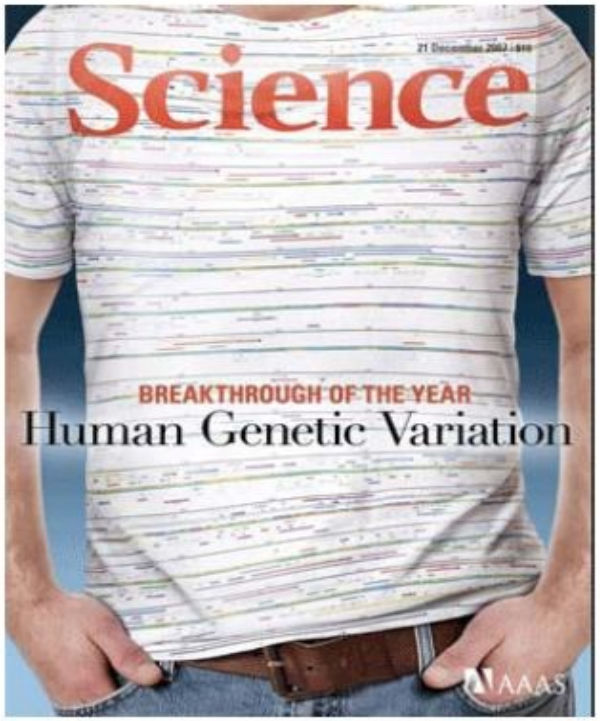
J. Craig Venter INSTITUTE

SNPs & InDels in HuRef Autosomes



2007 SCIENTIFIC BREAKTHROUGH OF THE YEAR

Science Magazine, December 21, 2007



"It's all about me!"

Single Nucleotide Polymorphisms (SNPs)

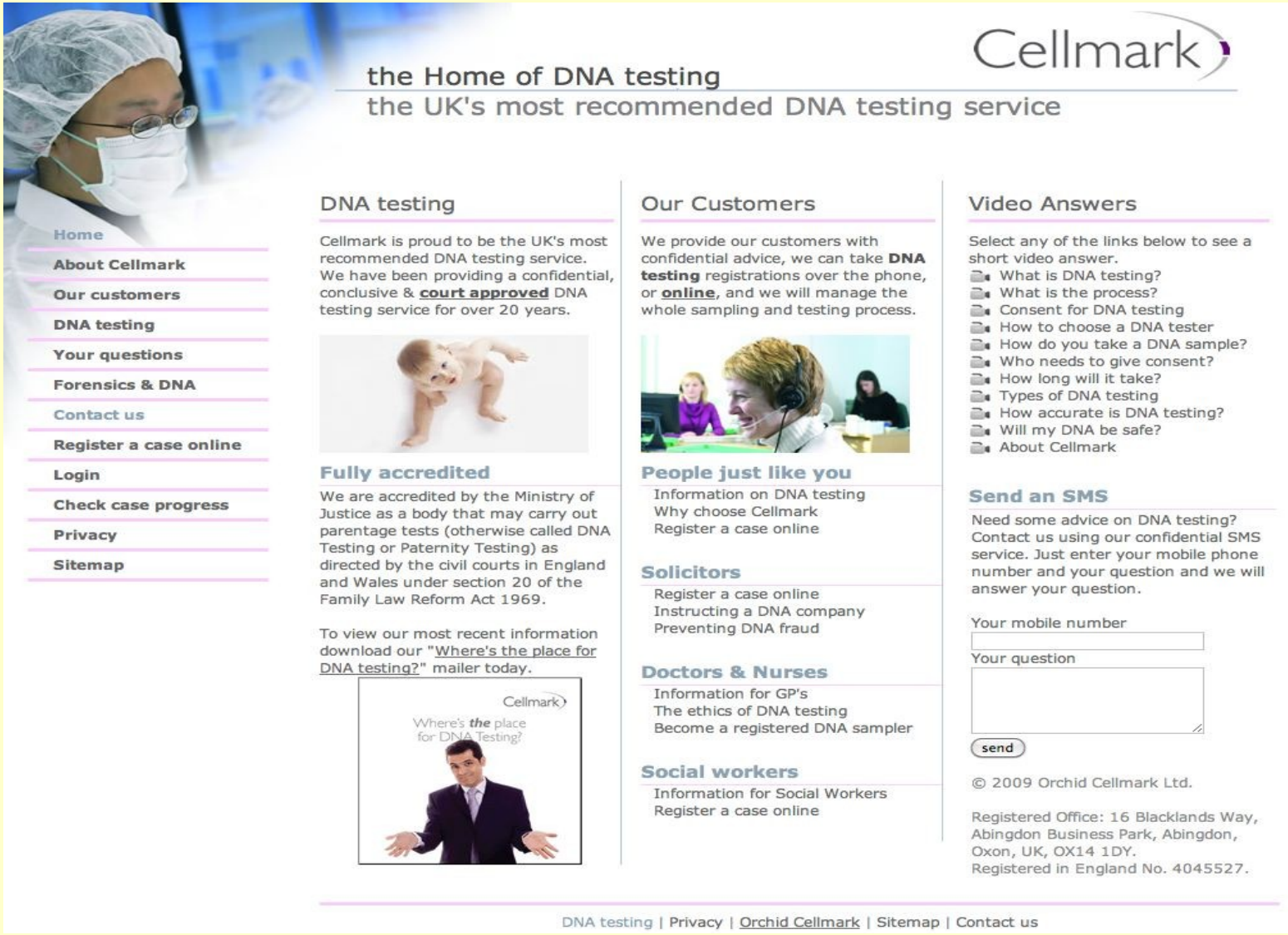
Single (Simple) Nucleotide Polymorphisms (SNPs)

GCTGTATGACTAGAAGATCGAT
GCTGTATGACGAGAAGATCGAT

- SNPs are used for identification and forensics
- SNPs are used for mapping and genome-wide association studies of complex diseases
- SNPs are used for estimating predisposition to disease
- SNPs are used for immigration & citizenship in the UK
- SNPs are used to predict specific genetic traits
- SNPs are used for classifying patients in clinical trials

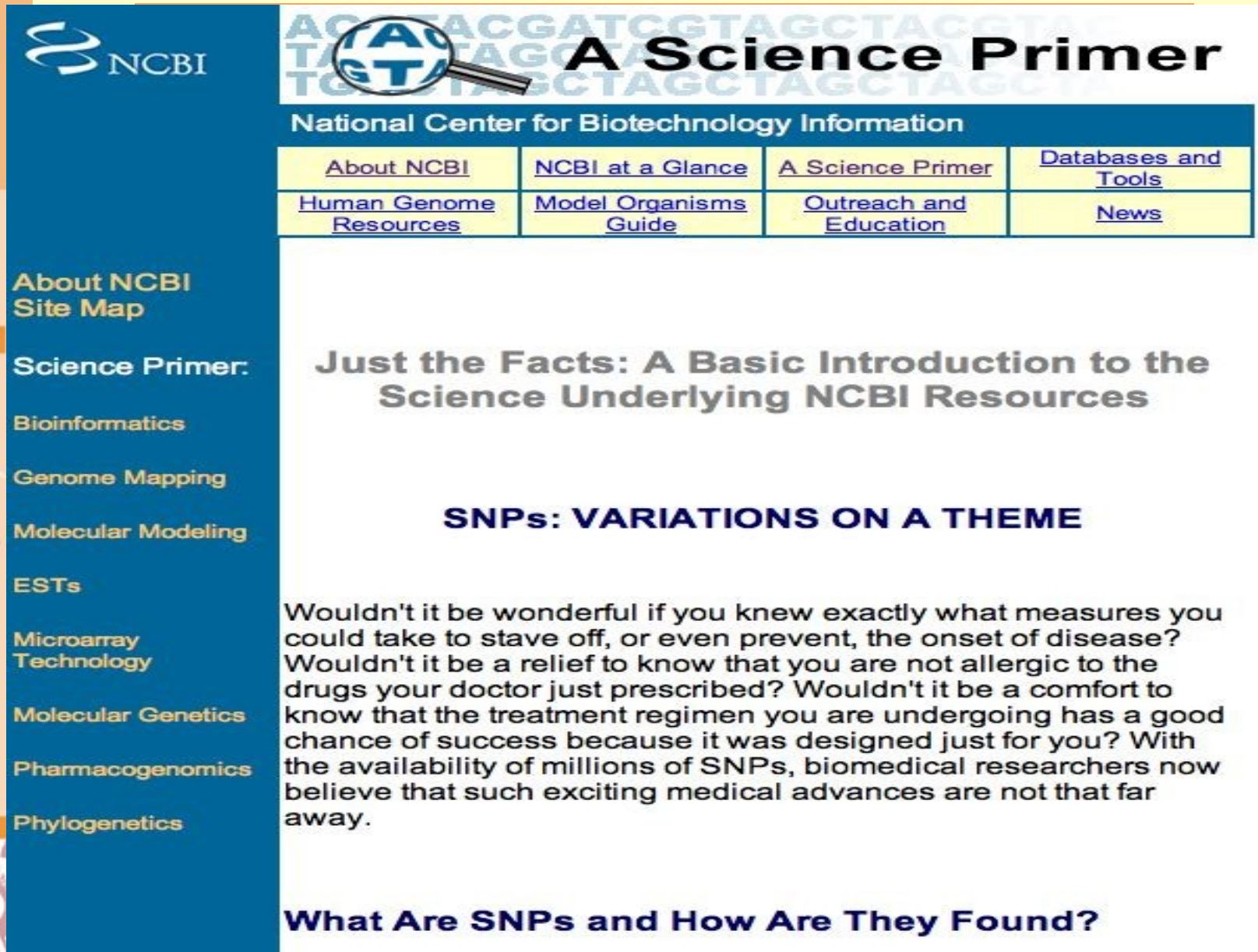
Orchid Cellmark Ltd.

http://www.cellmark.co.uk/



A SNP Primer at NCBI

http://www.ncbi.nlm.nih.gov/About/primer/snps.html



Department of Energy (DOE) SNP Page

http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml

genomics.energy.gov — Human Genome Project Information • Genomics:GTL • Microbial Genome Program • hom

skip navigation

Human Genome Project Information

Home | Site Index | News

About HGP | Research | Education | Ethics | Medicine | Media

SNP Fact Sheet

- Subject Index
- Send the url of this page to a friend

Basic Information
- FAQs
- Glossary
- Acronyms
- Links
- Genetics 101
- Publications
- Meetings Calendar
- Media Guide

About the Project
- What is it?
- Goals
- Progress
- History
- Ethical Issues
- Benefits
- Genetics 101

Medicine & the New Genetics
- Home
- Gene Testing
- Gene Therapy
- Pharmacogenomics
- Disease Information
- Genetic Counseling

Ethical, Legal, Social Issues
- Home
- Privacy Legislation
- Gene Testing
- Patenting
- Forensics
- Genetically Modified Food
- Behavioral Genetics
- Minorities, Race, Genetics
- Genetics in Courtroom

Quick links for this page:

- What are SNPs?
- How can SNPs be used as risk factors in disease development?
- Human Genome Project SNP Mapping Goals
- What is the SNP consortium?
- Who are members of the SNP consortium?
- Why should private companies fund a publicly accessible genome map?
- Whose DNA was analyzed to create the consortium's SNP map?
- Are SNP data available to the public?
- Related Links - SNP basics, articles
- Meeting Proceedings and Reports

What are SNPs?

Single nucleotide polymorphisms or SNPs (pronounced "snips") are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence is altered. For example a SNP might change the DNA sequence AAGGCTAA to ATGGCTAA. For a variation to be considered a SNP, it must occur in at least 1% of the population. SNPs, which make up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement of cytosine (C) with thymine (T). SNPs can occur in both coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function, but scientists believe others could predispose people to disease or influence their response to a drug.

Although more than 99% of human DNA sequences are the same across the population, variations in DNA sequence can have a major impact on how humans respond to disease; environmental insults such as bacteria, viruses, toxins, and chemicals; and drugs and other therapies. This makes SNPs of great value for biomedical research and for developing pharmaceutical products or medical diagnostics. SNPs are also evolutionarily stable --not changing much from generation to generation --making them easier to follow in population studies.

Simple Nucleotide Polymorphisms (SNPs)

- Most SNPs are genetically neutral
  - Forensics
  - Paternity tests
  - Immigration in the United Kingdom
  - Follow ethnic migrations
  - Clinical trials
- However SNPs can reflect distinguishing characteristics
  - Predisposition to disease
  - Insurability and employment => GINA, December 7, 2009
  - Often the basis for discrimination or other stigma
- Some variations cause disease. Unlike SNPs, these variations are usually very rare.
- SNPs can serve as genetic markers for other traits
  - Clinical trials associate SNPs with drug efficacy
  - Clinical trials associate SNPs adverse drug reactions
  - Consumer genomics associate SNPs with many other traits
    - 23andMe, Navigenics, DNADirect

Carrier Frequency of Cystic Fibrosis Alleles

http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=gene&part=cf

Table 5. Carrier Frequency for Mutant *CFTR* Alleles

Population Group	Approximate Carrier Frequency	Reference
Ashkenazi Jewish	1:29	Kerem et al [1997]
North American Caucasian	1:28	Hamosh et al [1998]
African American	1:61	Hamosh et al [1998]

Carrier Frequency of Gaucher's Alleles

http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=gene&part=gaucher

Prevalence

A study from Australia reported a disease frequency of 1:57,000 [Meikle et al 1999]; a similar study from the Netherlands reported 1.16:100,000 [Poorthuis et al 1999].

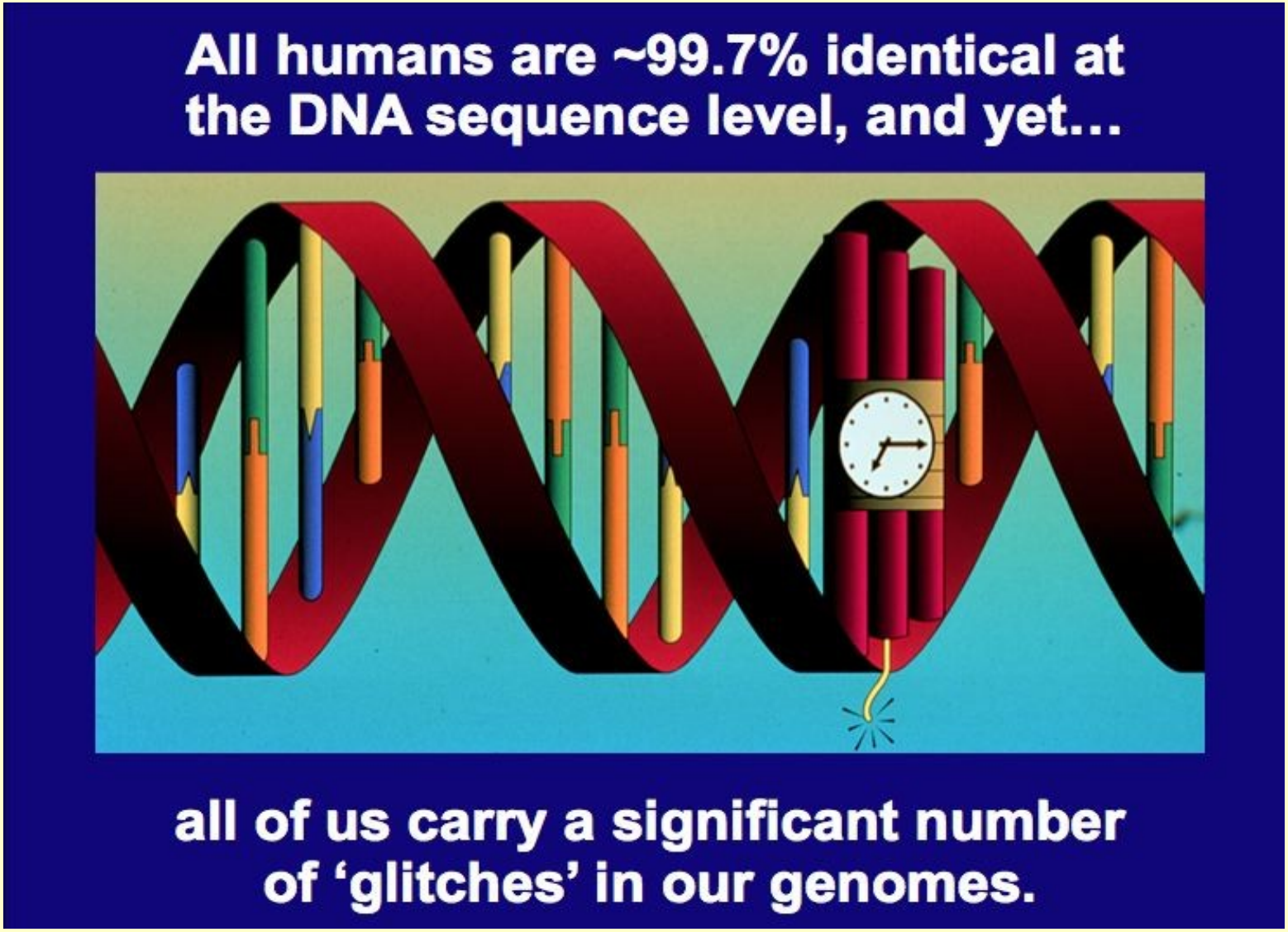
A founder effect for specific alleles underlies the observed occurrence of GD in specific populations:

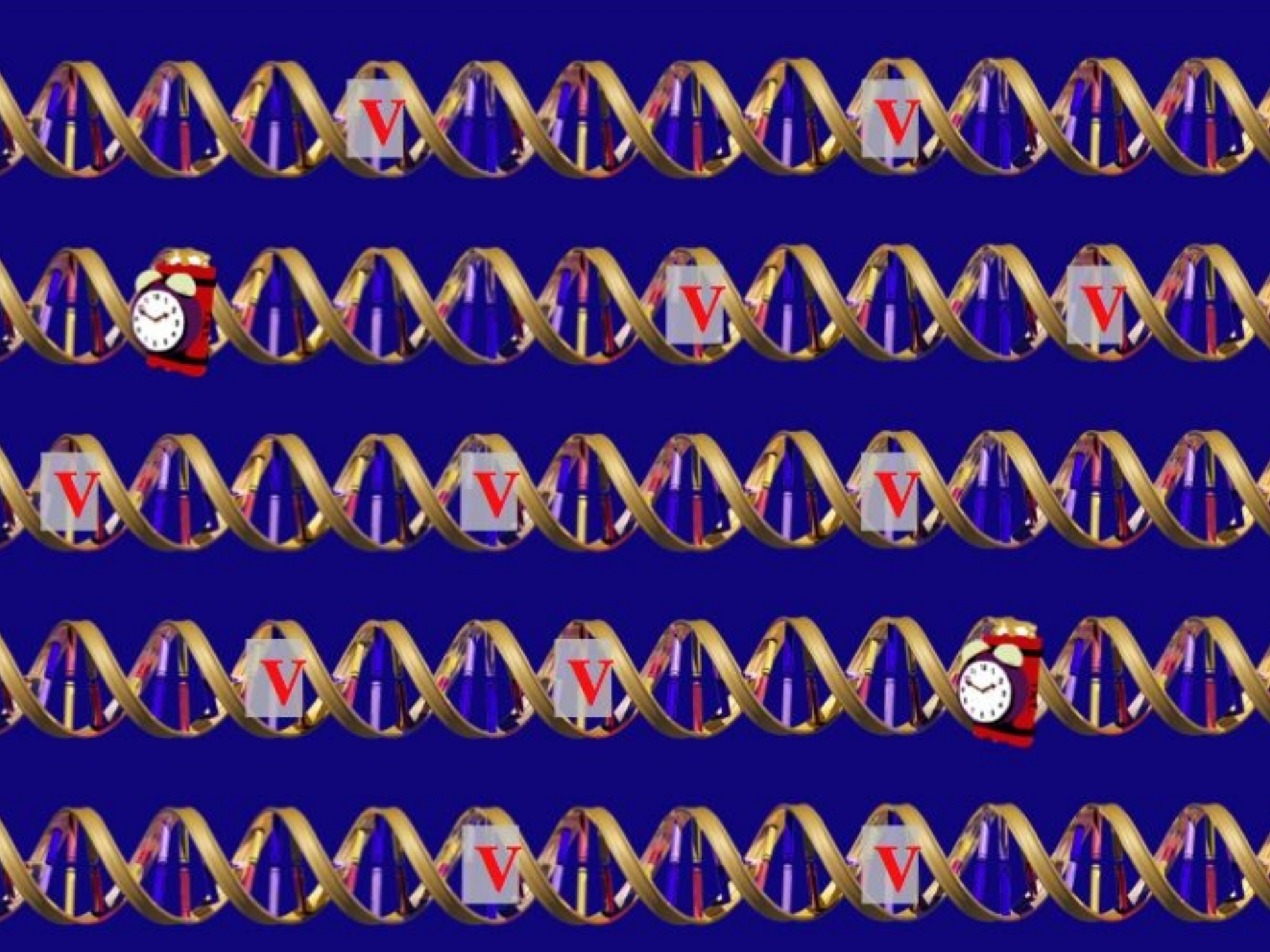
- Ashkenazi Jewish, Spanish, and Portuguese (N370S)
- Swedish (L444P)
- Jenin Arab, Greek, and Albanian (D409H)

Non-neuropathic GD (type 1) is prevalent in the Ashkenazi Jewish population, with a disease prevalence of 1:855 and an estimated carrier frequency of 1:18.

The prevalence of neuropathic GD (types 2 and 3) varies across ethnic groups but appears to be higher among non-Caucasians.

Eric Green's View of SNPs (Director of NHGRI)



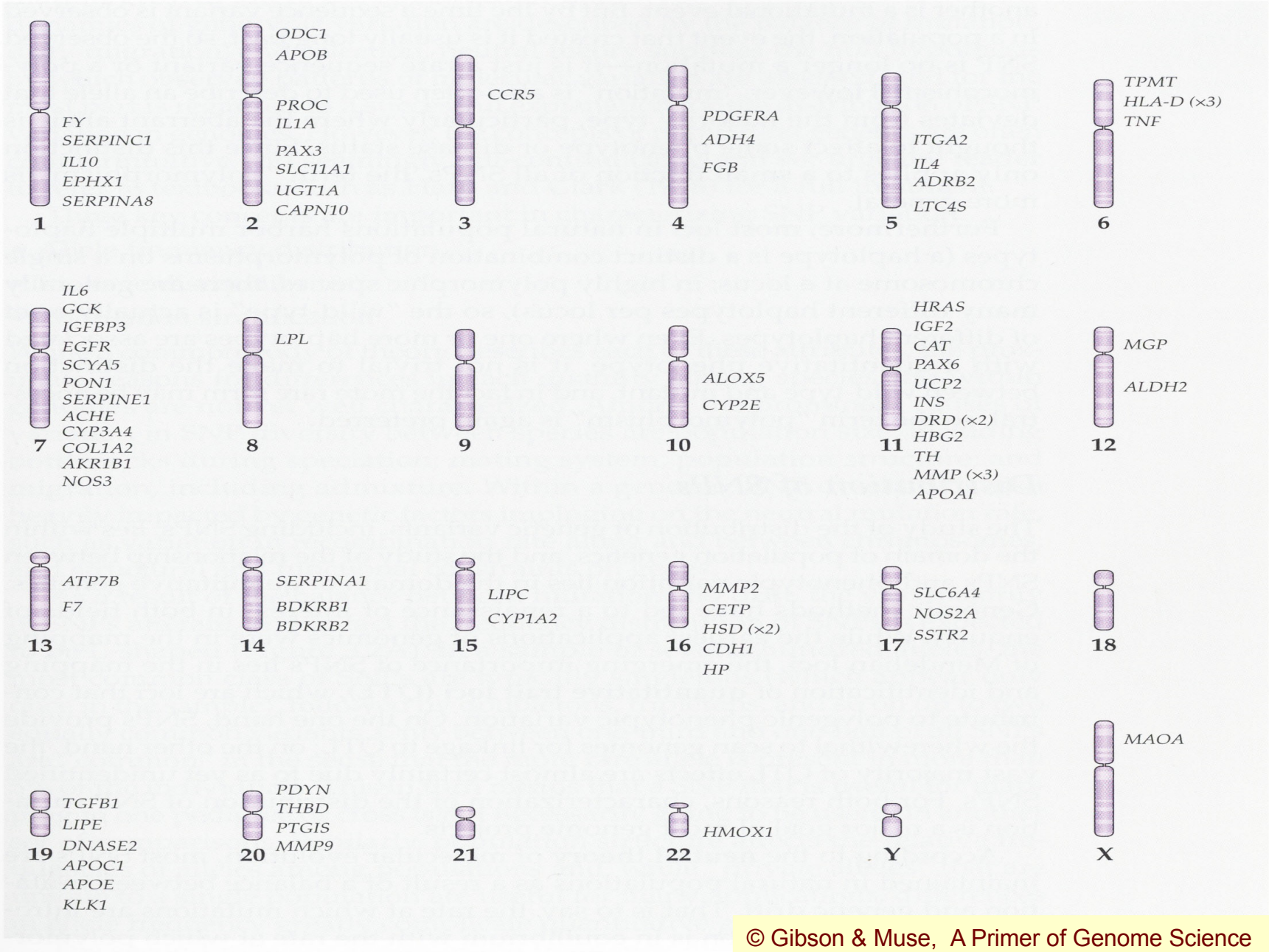

V
V
V
V
V
V
V
V
V
V
V
V
V
V
V

Types of SNPs

http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp

- Noncoding SNPs
  - 5′ UTR
  - 3′ UTR
  - Introns
  - Intergenic Regions
  - Pseudogenes
  - Regulatory
    - Splicing
    - Transcriptional regulation (promoter & TF binding sites)
    - Translational regulation (initiation or termination)
    - Regulatory miRNA target sites
- Coding SNPs
  - Synonymous SNPs (third position variation)
  - Replacement SNPs (change Amino acid)
    - Functional SNPs (acceptable amino acid replacement)
    - Non-functional SNPs (traits & diseases)

Human Promoter SNPs



1: **HBB hemoglobin, beta [*Homo sapiens*]**

GeneID: 3043 updated 17-Feb-2010

Summary

Official Symbol	HBB provided by HGNC
Official Full Name	hemoglobin, beta provided by HGNC
Primary source	HGNC:4827
See related	Ensembl:ENSG00000223609; HPRD:00786; MIM:141900
Gene type	protein coding
RefSeq status	REVIEWED
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	CD113t-C; beta-globin; HBB
Summary	The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains in adult hemoglobin, Hb A. The normal adult hemoglobin tetramer consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell anemia. Absence of beta chain causes beta-zero-thalassemia. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. The order of the genes in the beta-globin cluster is 5'-epsilon -- gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq]

Genomic regions, transcripts, and products

(minus strand) Go to reference sequence details

Try our new Sequence Viewer

Genomic context

chromosome: 11; **Location:** 11p15.5

See HBB in MapViewer

Entrez Gene Hom

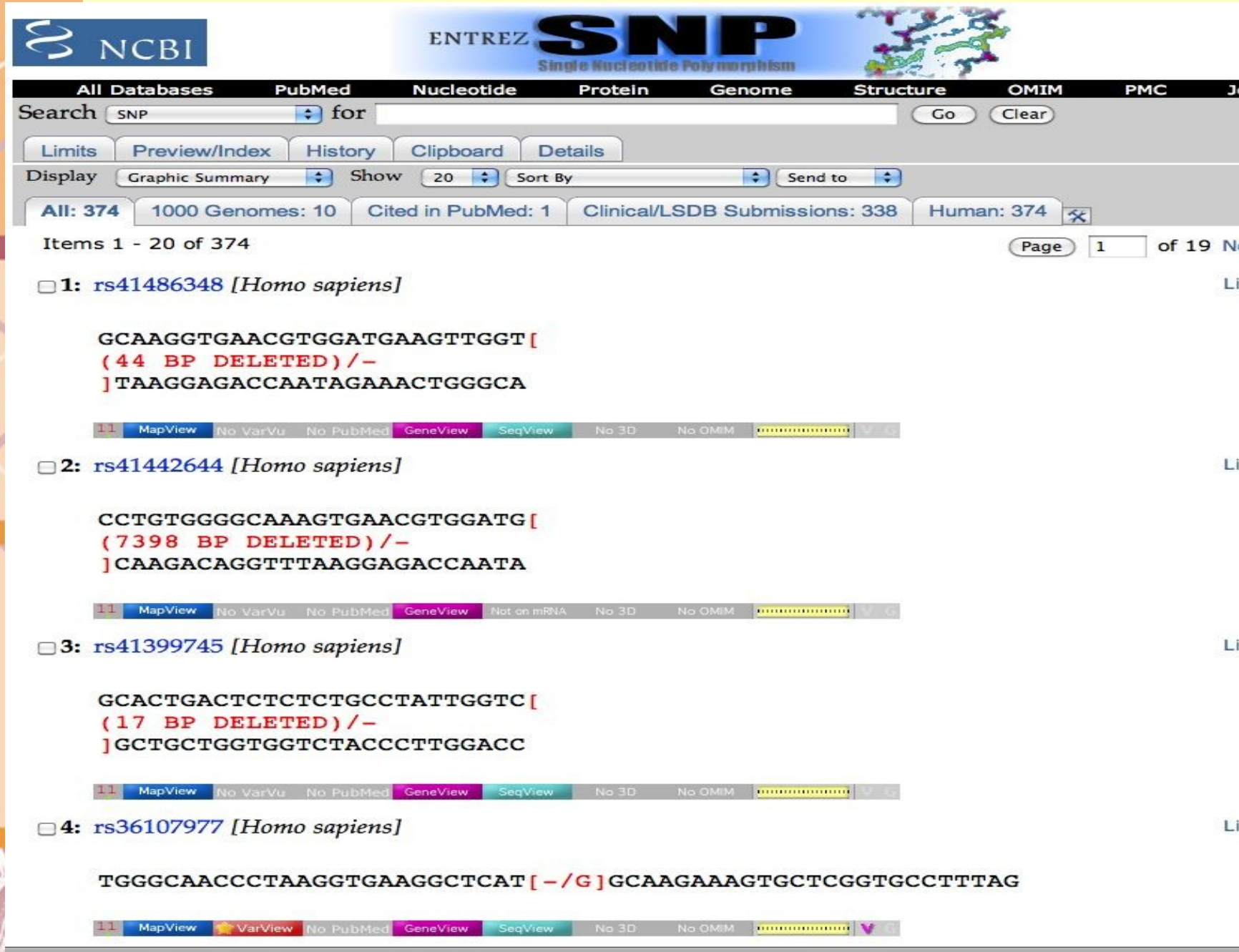
Table Of Conte

Summary
Genomic regio
products
Genomic conte
Bibliography
Interactions
General gene i
General protei
Reference sequ
Related sequer
Additional links

Links

Order cDNA clo
BioAssay, by G
Books
CCDS
Conserved Don
Full text in PMC
GEO Profiles
Genome
HomoloGene
Map Viewer
Nucleotide
OMIM
Probe
Protein
PubChem Comp
PubChem Subs
PubMed
PubMed (Gene
PubMed (OMIM
SNP
SNP: GeneView
SNP: Genotype
SNP: VarView
Taxonomy

Human Beta Hemoglobin Gene SNPs



Gene Model (mRNA alignment) information from genome sequence

Total gene model (contig mRNA transcript):				3		
mrna	**transcript**	**protein**	**mrna orientation**	**Contig**	**Contig Label**	**List SNP**
NM_000518.4	minus strand	NP_000509.1	reverse	NT_009237.17	reference	**<- currently shown**
NM_000518.4	minus strand	NP_000509.1	reverse	NW_925006.1	Celera	View snp on GeneModel
NM_000518.4	minus strand	NP_000509.1	reverse	NW_001838021.1	HuRef	View snp on GeneModel

Include clinically associated | in gene region | cSNP | has frequency | double hit | refresh

gene model (contig mRNA transcript):	Contig Label	Contig	mrna	protein	mrna orientation	transcript	snp count
	reference	NT_009237.17	NM_000518.4	NP_000509.1	reverse	minus strand	12, coding

Color Legend

Region	Contig position	mRNA pos	dbSNP rs# cluster id	Hetero-zygosity	Validation	3D	Clinically Associated	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos
exon_3	4034095	465	rs41405449	N.D.				frame shift	-/AGC		1	139
								frame shift	AGC	Ser [S]	1	139
								contig reference	GCTA	[AN]	1	139
	4034117	430	rs41511744	N.D.				frame shift	(15bp)	[QAAYQ]	2	127
								frame shift	(17 BP DELETED)/-		2	127
								contig reference	(17bp)	[VQAAYQ]	2	127
intron_2	4034199	365	rs71811954	N.D.				frame shift	(14bp)	Thr [T]	3	105
								frame shift	(15bp)		3	105
								frame shift	(LARGEDELETION)/-		3	105
exon_2	4035096	317	rs11549405	N.D.				synonymous	C	Leu [L]	3	89
								contig reference	G	Leu [L]	3	89
	4035119	294	rs11549406	0.005	H			missense	G	Val [V]	1	82
								contig reference	C	Leu [L]	1	82
	4035143	270	rs36209341	N.D.				missense	A		1	74
								missense	T		1	74
								contig reference	G		1	74
	4035210	203	rs17850156	N.D.		Yes		synonymous	C	Thr [T]	3	51
								contig reference	T	Thr [T]	3	51
	4035211	202	rs34676051	N.D.				missense	A	Asn [N]	2	51
								contig reference	C	Thr [T]	2	51

☐ **5:** rs36092904 *[Homo sapiens]* Links

CCTAAGGTGAAGGCTCATGGCAAGAA[A/T]GTGCTCGGTGCCTTTAGTGATGGCC

11 MapView VarView No PubMed GeneView SeqView No 3D No OMIM

HGVS Names: [*NG_000007.3:g.70925A>T*] [*NM_000518.4:c.201A>T*] [*NP_000509.1:p.Lys67Asn*] [*NT_009237.17:g.4035162T>A*]

☐ **6:** rs36081208 *[Homo sapiens]* Links

ACTGAGTGAGCTGCACTGTGACAAGC[C/T]GCACGTGGATCCTGAGAACTTCAGG

11 MapView VarView No PubMed GeneView SeqView No 3D OMIM

HGVS Names: [*NG_000007.3:g.71014T>C*] [*NM_000518.4:c.290T>C*] [*NP_000509.1:p.Leu97Pro*] [*NT_009237.17:g.4035073A>G*]

☐ **7:** rs36038739 *[Homo sapiens]* Links

ACACTGAGTGAGCTGCACTGTGACAA[C/G/T]CTGCACGTGGATCCTGAGAACTTCA

11 MapView VarView No PubMed GeneView SeqView No 3D OMIM

HGVS Names: [*NG_000007.3:g.71012G>C*] [*NG_000007.3:g.71012G>T*] [*NM_000518.4:c.288G>C*] [*NM_000518.4:c.288G>T*] [*NP_000509.1:p.Lys96Asn*] [*NT_009237.17:g.4035075C>A*] [*NT_009237.17:g.4035075C>G*]

GeneView

GeneView via analysis of contig annotation: HBB *hemoglobin, beta*

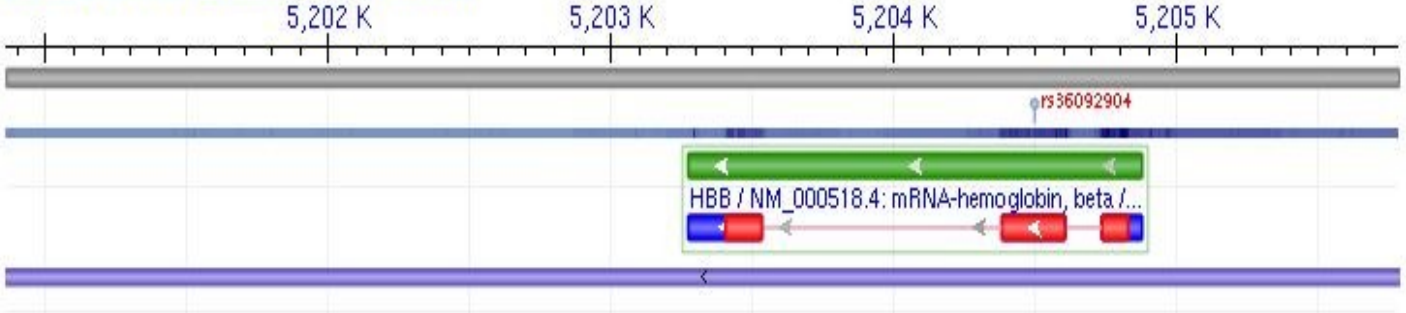
View more variation on this gene (click to hide).

Include clinically associated: in gene region | cSNP | has frequency | double hit | Go

Assembly	SNP to Chr	Chr	Chr position	Contig	Contig position	Allele
reference	-	11	5204497	NT_009237.17	4035162	T

Function	mRNA				Protein		
	mRNA to Chr	Accession	Position	Allele change	Accession	Position	Residue change
missense	-	NM_000518.4	251	AA**A** ⇒ AA**T**	NP_000509.1	67	**K [Lys] ⇒ N [Asn]**

(Open sequence viewer in a new window.)



GeneView via direct blast against RefSeq sequences (used when no gene model is available): N/A

Integrated Maps:

Genome Build	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Group term	Group label	Contig label	Neighbor SNP
36.3	11	5366592	NW_925006.1	866916	-	T	alt_assembly_1	Celera	Celera	view
36.3	11	4907094	NW_001838021.1	876469	-	T	alt_assembly_8	HuRef	HuRef	view
36.3	11	5204497	NT_009237.17	4035162	-	T	ref_assembly	reference	reference	view
Provisional 37.1	11	5247921	NT_009237.18	5187921	-	T	ref_assembly	Primary_Assembly	Primary_Assembly	view

Diversity of One ß-Hemoglobin SNP

Population Diversity

	Sample Ascertainment				Genotype Detail NEW			Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	C/C	C/G	HWP	C	G
ss16249024	HapMap-CEU	European	112	IG	0.982	0.018		0.991	0.009
	HapMap-HCB	Asian	86	IG	1.000			1.000	
	HapMap-JPT	Asian	86	IG	1.000			1.000	
	HapMap-YRI	Sub-Saharan African	114	IG	1.000			1.000	

HBB Variation Viewer

Gene	HBB; hemoglobin, beta
Description	beta globin | beta globin chain | hemoglobin beta chain Also known as: CD113t-C
Species	*Homo sapiens*
Cyto	11p15.5

Gene Reference Sequences	NG_000007.1 genomic NM_000518.4 transcript NP_000509.1 protein *variation locations are ba*
Links	HGMD, Panther, Gene,

Observed Variation | Page 1 of 20 | Displaying results

Var Class	Genomic	Transcript	Protein	Clinical interpretation	...	...	Freq	Pub...	MIM Al Var
DIP		c.-11A>AAAC			2	2			
DIP		c.*108A>AATAA			2	2			141900.0417
SNP	g.70453C>T				2	2			141900.0371
SNP	g.70455C>T				2	2			141900.0425
SNP	g.70457C>A | g....				4	4			141900.0372, ...
SNP	g.70458C>A | g....				6	6			141900.0374
SNP	g.70459C>A | g....				4	4			141900.0375
SNP	g.70513C>A				3	3			141900.0406
SNP	g.70514A>C | g....				4	4			141900.0376
SNP	g.70515T>A | g....				4	4			141900.0377, ...
SNP	g.70516A>G				2	2			141900.0379
SNP	g.70517A>C | g....				4	4			141900.0380, ...
SNP	g.70545A>C	c.-50A>C			2	2			141900.0387
SNP	g.70566G>A	c.-29G>A			2	2			
SNP	g.70577C>G	c.-18C>G			2	2			
SNP	g.70595A>G	c.1A>G	p.M1V		2	2			
SNP	g.70596T>C | g....	c.2T>C | c.2T>G	p.M1T | p.M1R		5	5			141900.0344, ...
SNP	g.70597G>A | g....	c.3G>A | c.3G>...	p.M1I | p.M1I | p...		6	6			141900.043
SNP	g.70598G>A | g....	c.4G>A | c.4G>T	p.V2M | p.V2L		4	4			141900.0266, ...
MIX	g.70599T>A | g....	c.5T>A | c.5T>...	p.V2E | p.V2A | ...		7	7			141900.0069, ...

.0077 HEMOGLOBIN FUKUOKA [HBB, HIS2TYR] dbSNP

See Harano et al. (1990).

.0078 HEMOGLOBIN FUKUYAMA [HBB, HIS77TYR] dbSNP

See Hidaka et al. (1988).

.0079 HEMOGLOBIN G (ACCRA) [HBB, ASP79ASN]

There is no clinical or hematologic abnormality in the homozygote. See Edingto

.0080 HEMOGLOBIN G (COPENHAGEN) [HBB, ASP47ASN] dbSNP

See Sick et al. (1967), Schiliro et al. (1981), and Chen et al. (1985).

.0081 HEMOGLOBIN G (COUSHATTA) [HBB, GLU22ALA] dbSNP

HEMOGLOBIN G (SASKATOON)
HEMOGLOBIN G (HSIN-CHU)
HEMOGLOBIN G (TAEGU)

See Schneider et al. (1964), Bowman et al. (1967), Vella et al. (1967), Blackwe (1989).

.0082 HEMOGLOBIN G (FERRARA) [HBB, ASN57LYS] dbSNP

See Giardina et al. (1978).

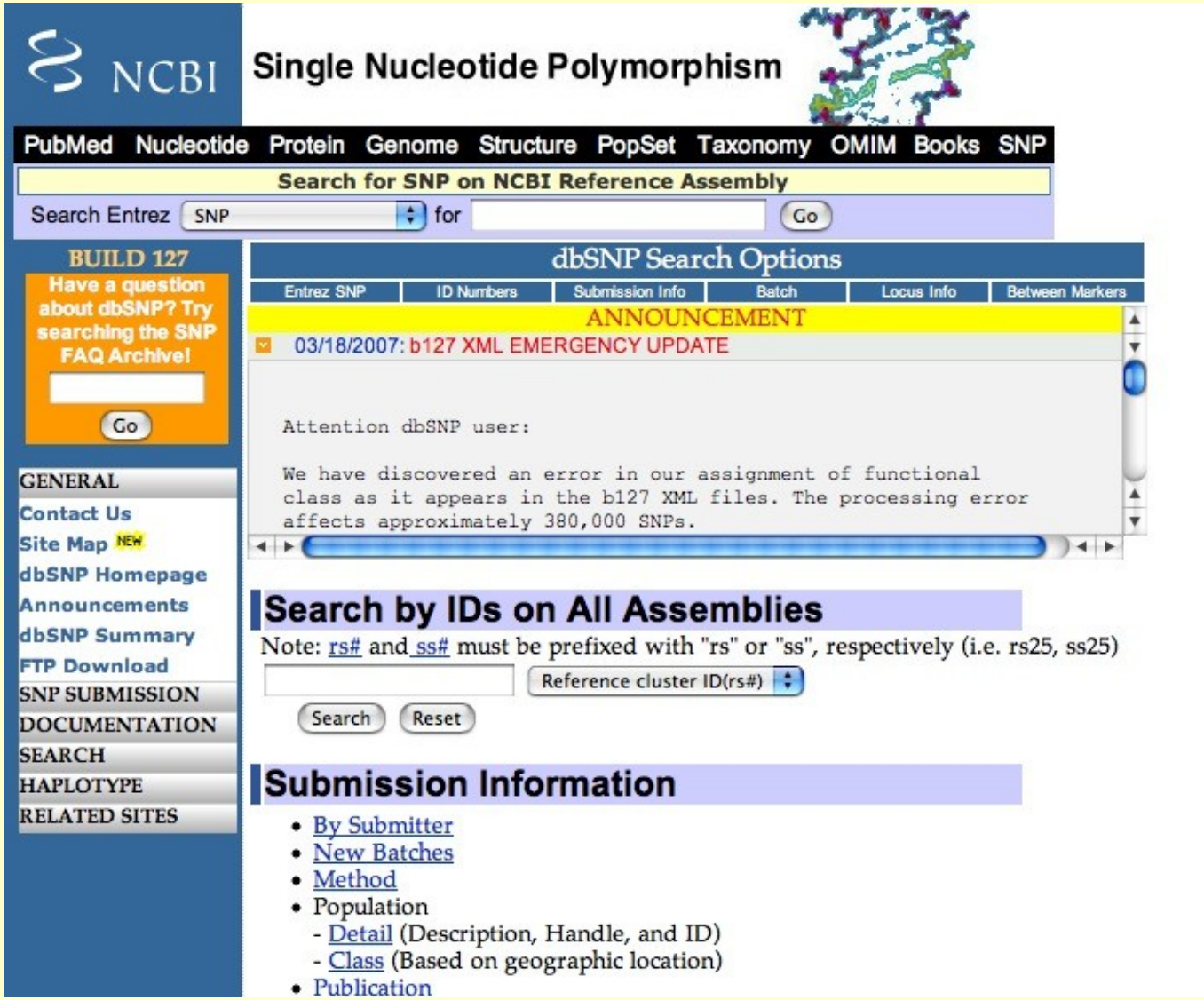
.0083 HEMOGLOBIN G (GALVESTON) [HBB, GLU43ALA] dbSNP

HEMOGLOBIN G (PORT ARTHUR)
HEMOGLOBIN G (TEXAS)

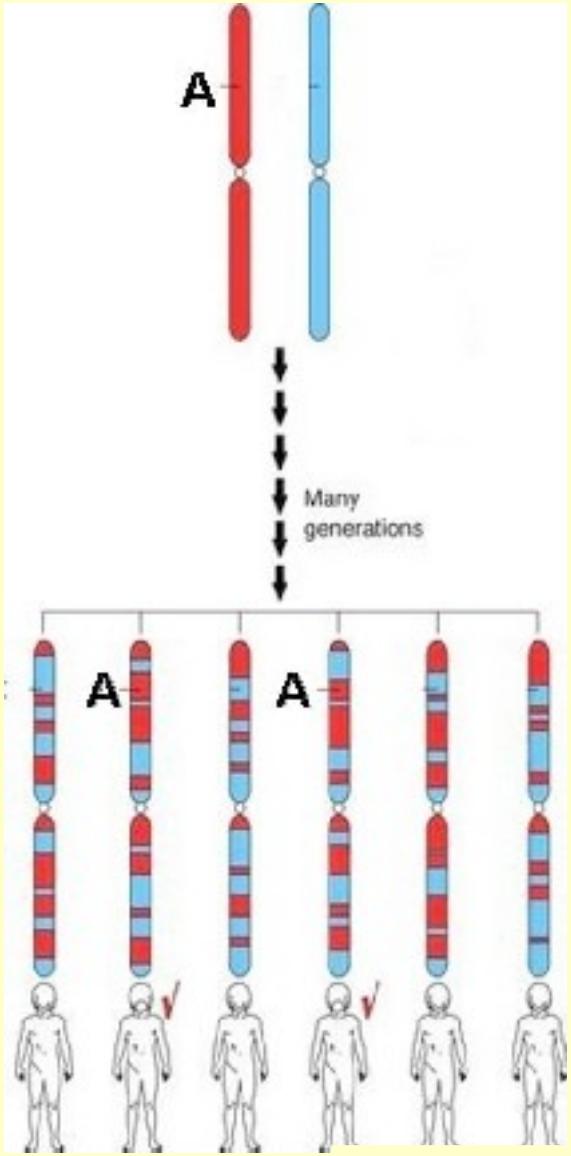
See Bowman et al. (1962, 1964).

dbSNP at NCBI

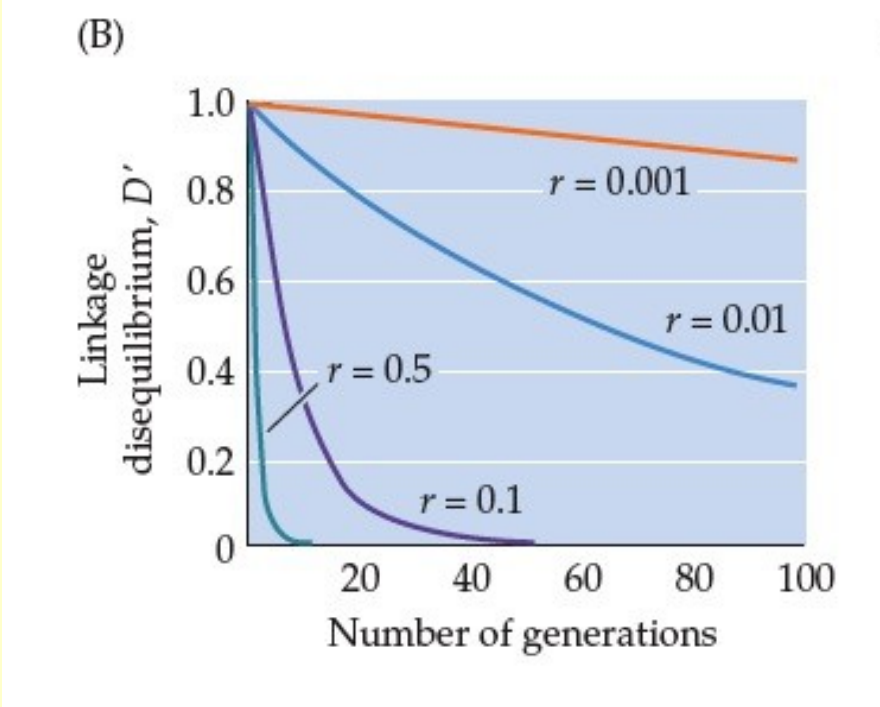
http://www.ncbi.nlm.nih.gov/SNP/



Origin of Haplotypes



Linkage Disequilibrium and Recombination Rate



Linkage Disequilibrium (LD) Across the Human LPL Gene

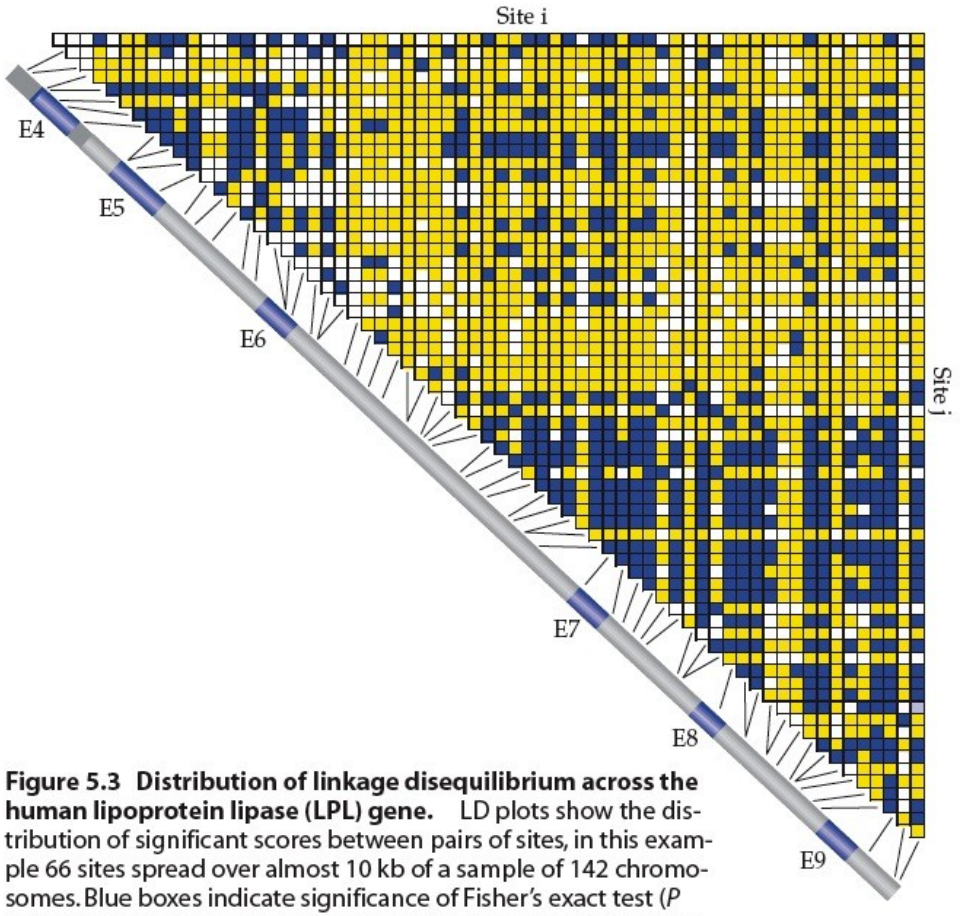
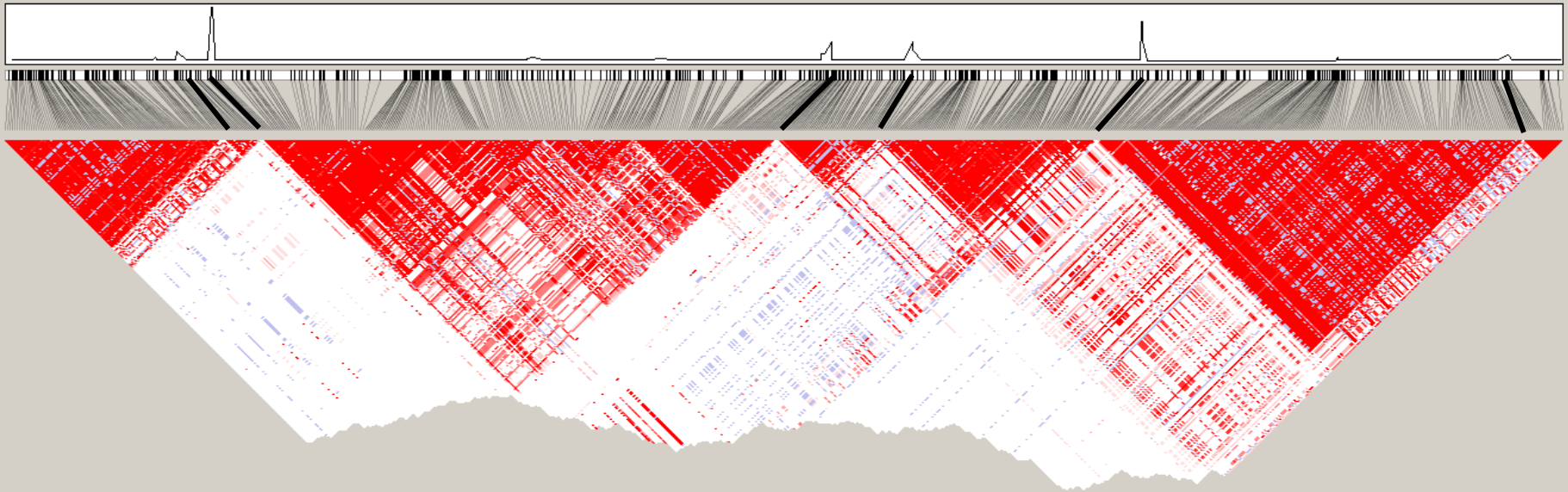


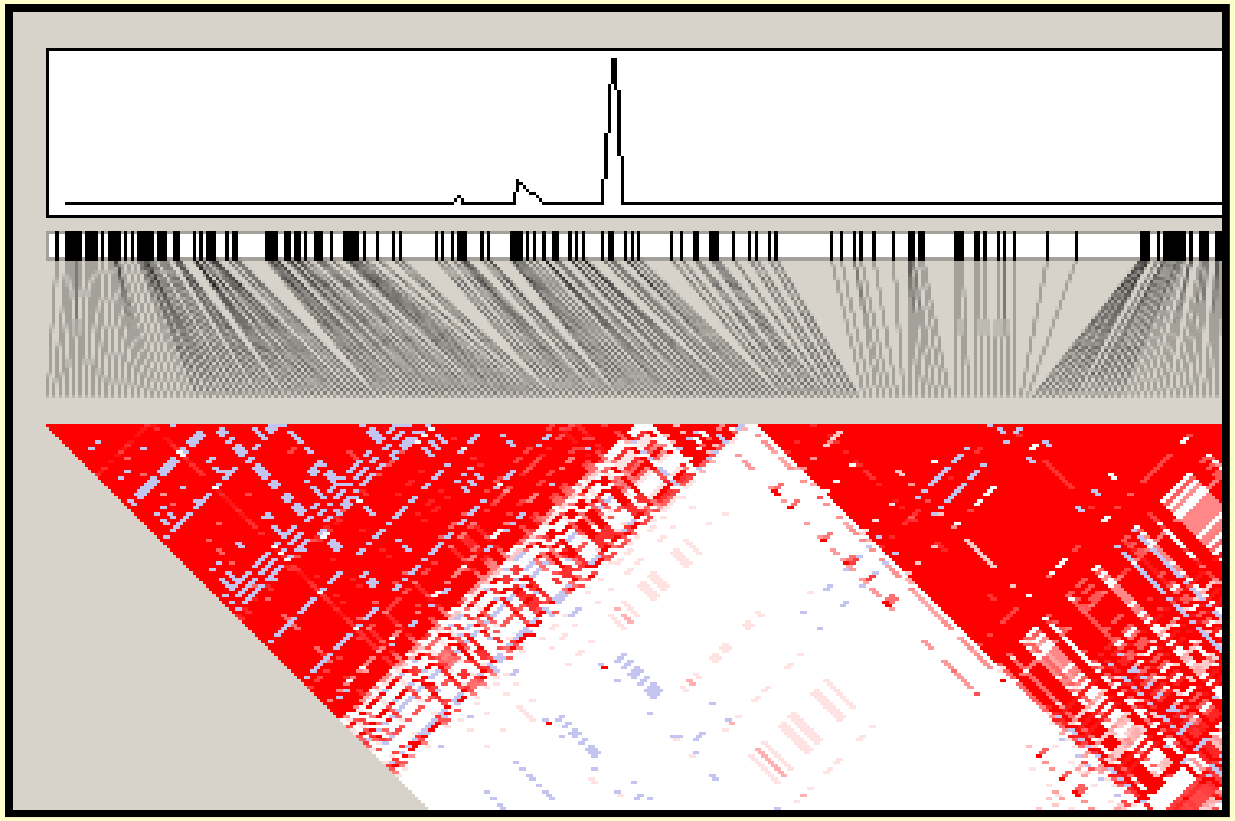
Figure 5.3 Distribution of linkage disequilibrium across the human lipoprotein lipase (LPL) gene. LD plots show the distribution of significant scores between pairs of sites, in this example 66 sites spread over almost 10 kb of a sample of 142 chromosomes. Blue boxes indicate significance of Fisher's exact test ($P < 0.001$), yellow boxes indicate nonsignificance, and white boxes are cases where there was insufficient power to test for LD at this level. Note that the extent of LD varies across the locus, and is not restricted to exon sequences. (Redrawn from Clark et al. 1998.)

Recombination hotspots are widespread and account for LD structure



7q21

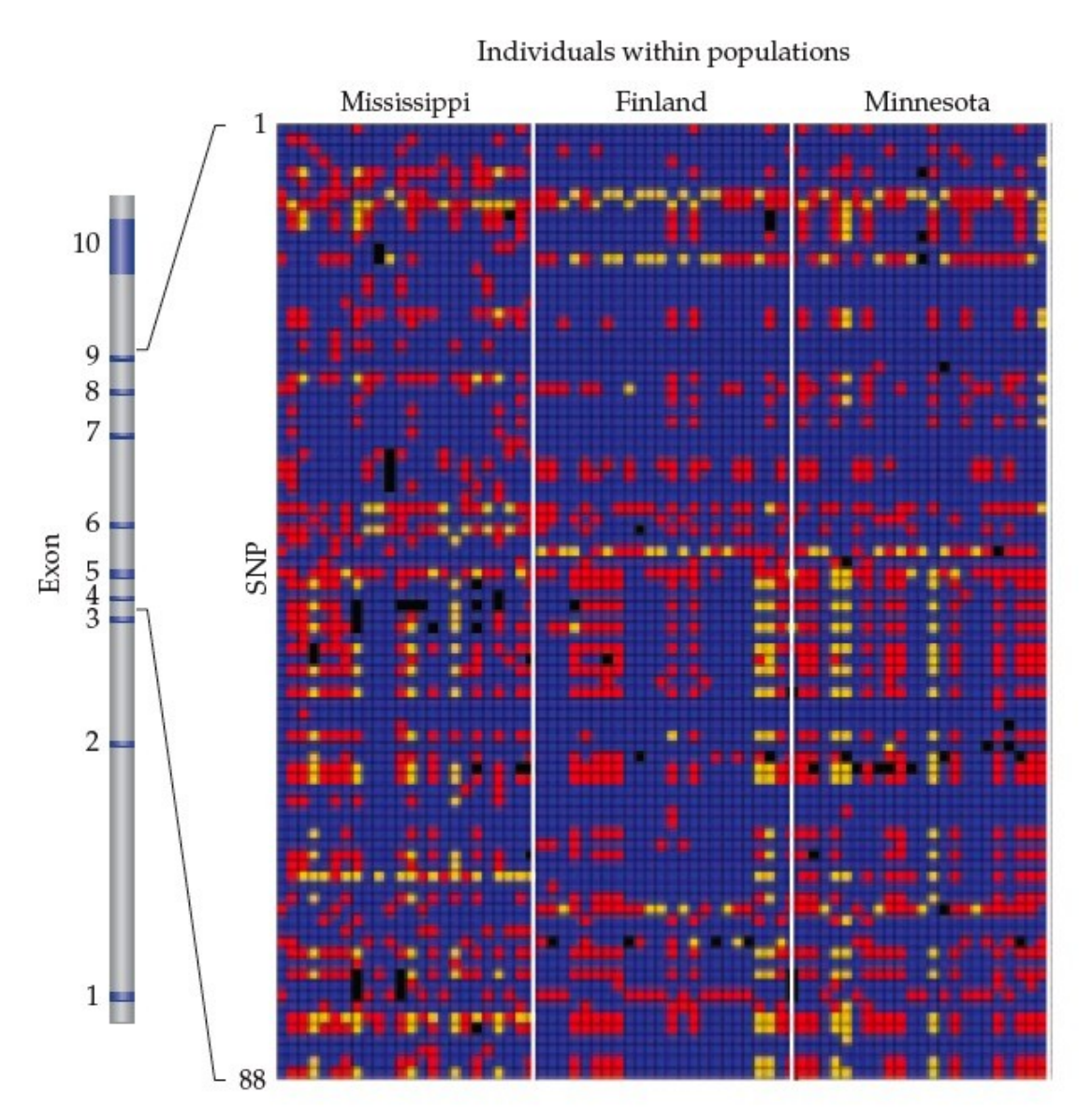
Recombination hotspots are widespread and account for LD structure



7q21

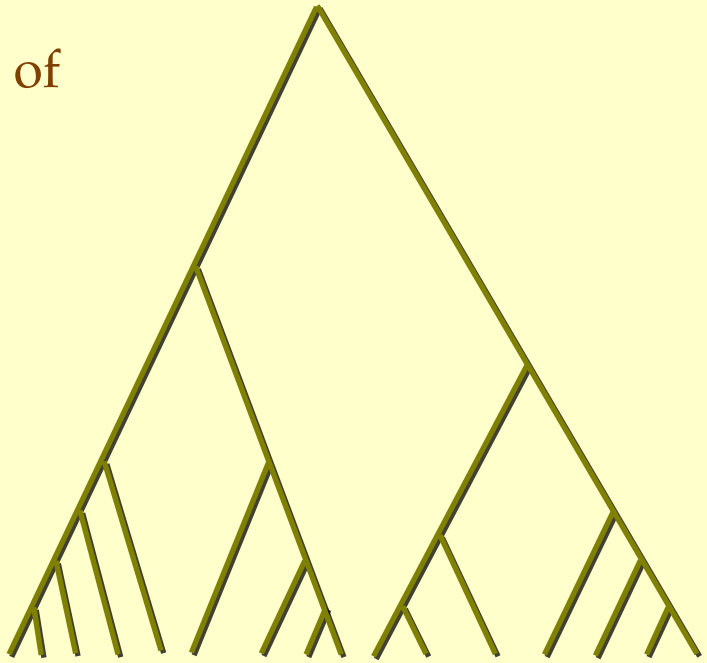
Observation of Haplotypes

SNPs in Populations



Sequence and Distance-Based Phylogenies (evolutionary trees)

- Sequence-Based Methods (Parsimony)
  - Assigns mutations to branches
  - Minimize number of changes
  - Topology maximizes similarity of neighboring leaves
- Distance-based methods
  - Branch lengths = D(i,j)/2 for sequences i, j
  - Distances must be metric
  - Distances can reflect time or number of changes
  - Distances must be relatively constant per unit branch length



is Collins, 2008

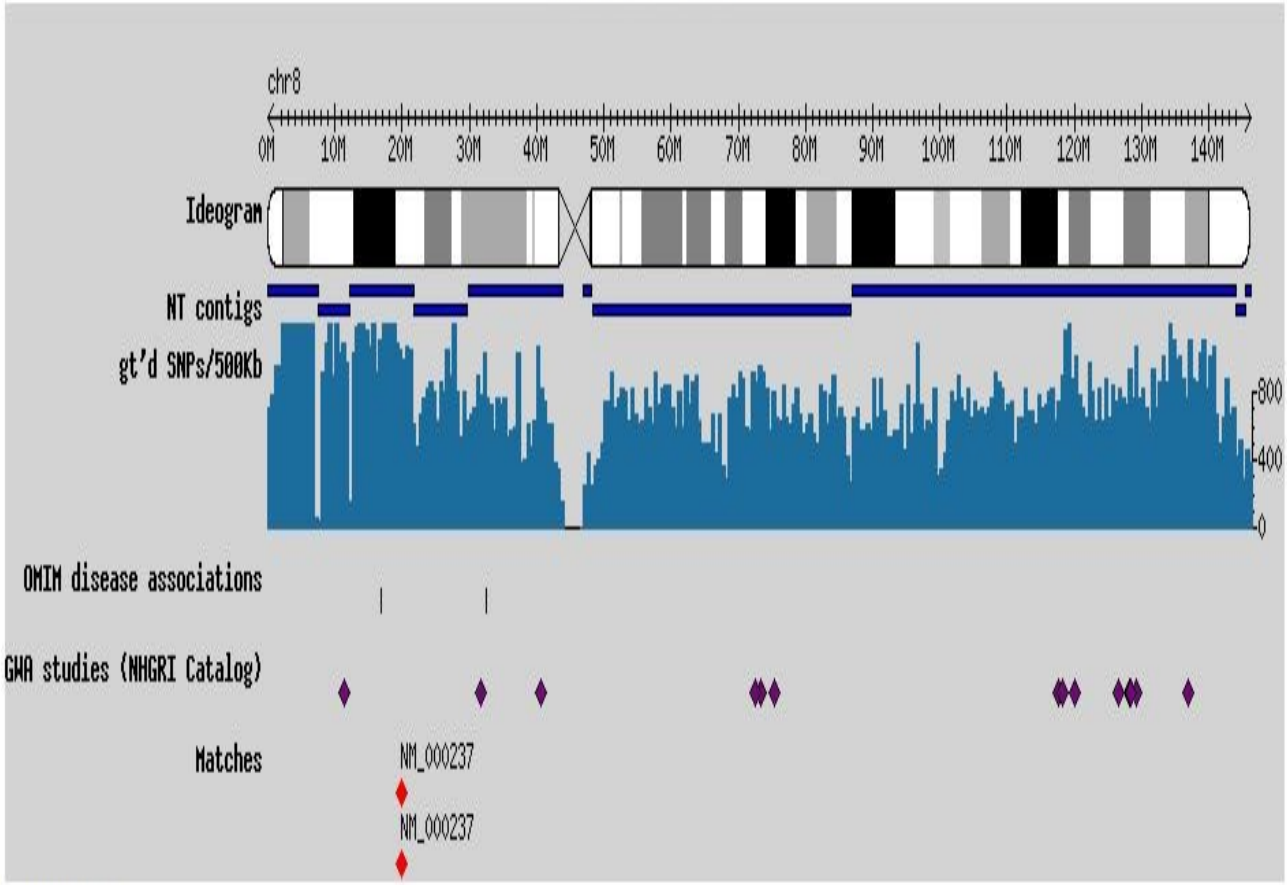
International HapMap Project

http://www.hapmap.org/

International HapMap Project

http://www.hapmap.org/

Matches on chr8



NM_000237 **lipoprotein lipase** precursor

chr8:19.84..19.87 Mbp (27.99 kbp) score=27.49

NM_000237 LPL encodes **lipoprotein lipase**, which is expressed in heart, muscle, and adipose tissue. LPL functions as a homodimer, and has the dual functions of triglyceride hydrolase and ligand/bridging factor for receptor-mediated **lipoprotein** uptake. Severe mutations that cause LPL deficiency result in type I hyper**lipoprotein**emia, while less extreme mutations in LPL are linked to many disorders of **lipoprotein** metabolism. Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Entrez Gene record to access additional publications.

chr8:19.84..19.87 Mbp (27.99 kbp) score=26.49

Overview

Ideogram

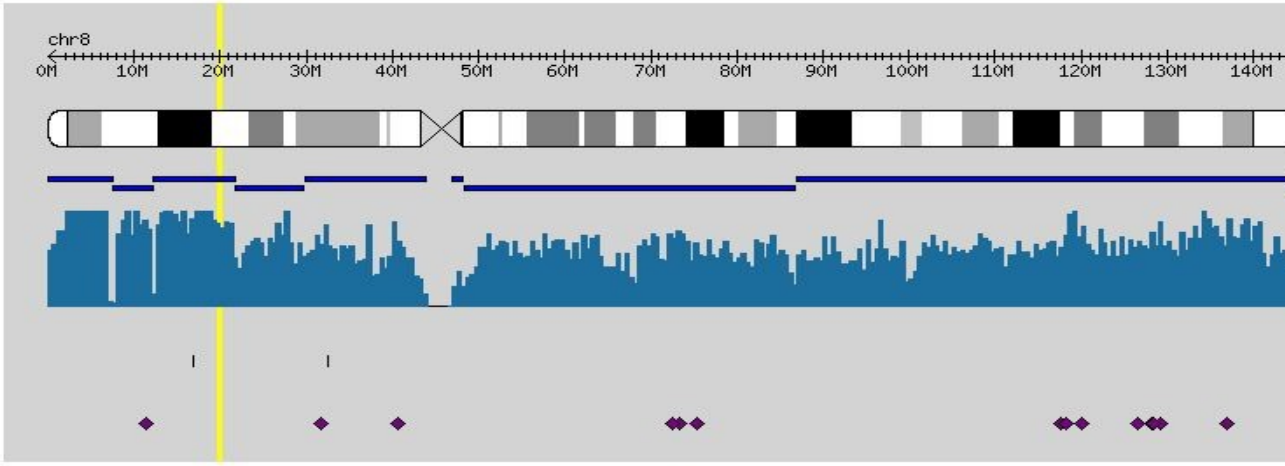
NT contigs

gt'd SNPs/500Kb

OMIM disease associations

GWA studies (NHGRI Catalog)

dbSNP SNPs/500Kb

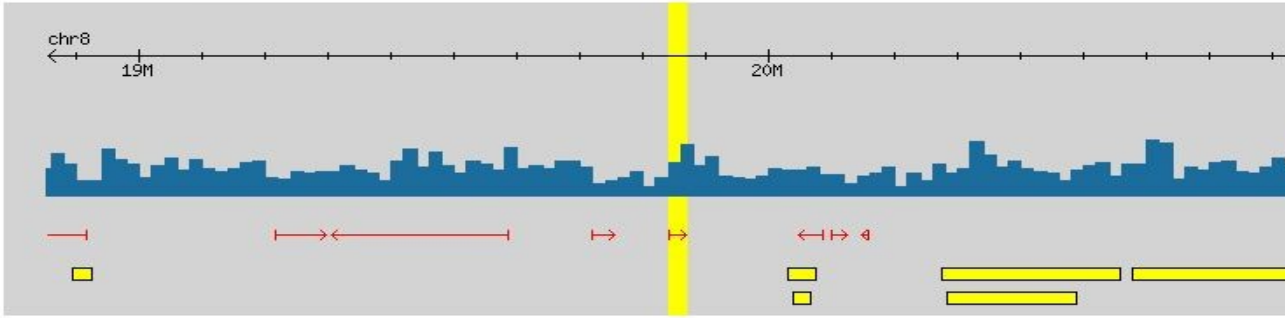


Region

gt'd SNPs/20Kb

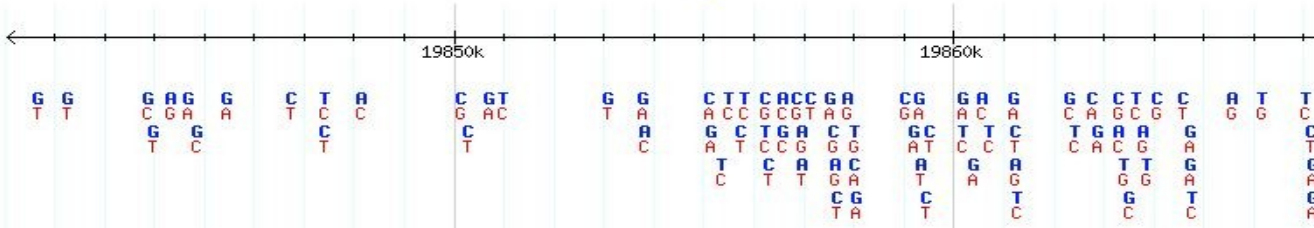
Entrez genes

Copy Number Variation

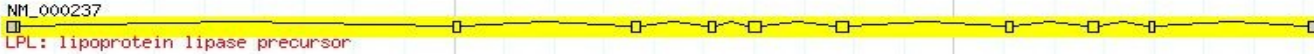


Details

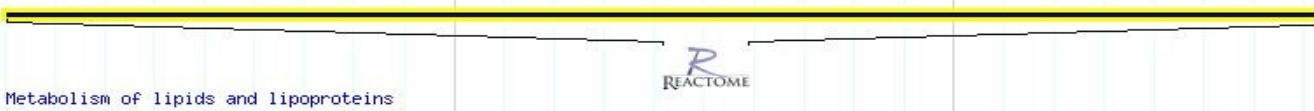
Genotyped SNPs



Entrez genes

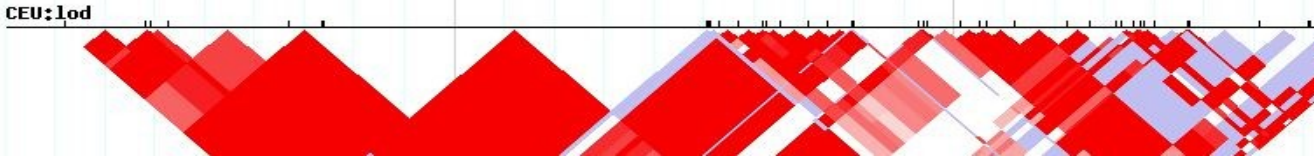


Reactome pathways

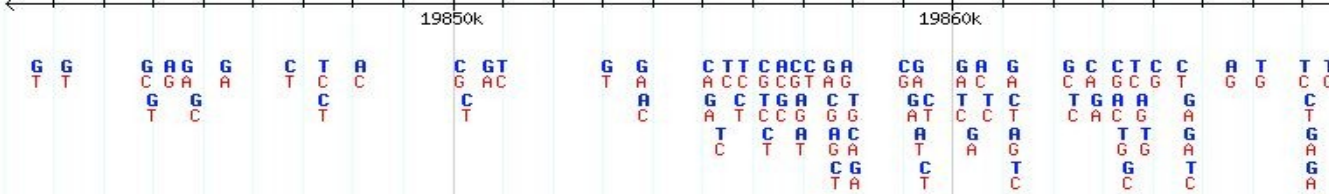


Genome-Wide Association studies (NHGRI Catalog)

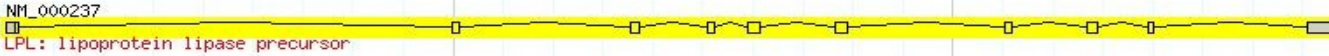
LD Plot



Genotyped SNPs



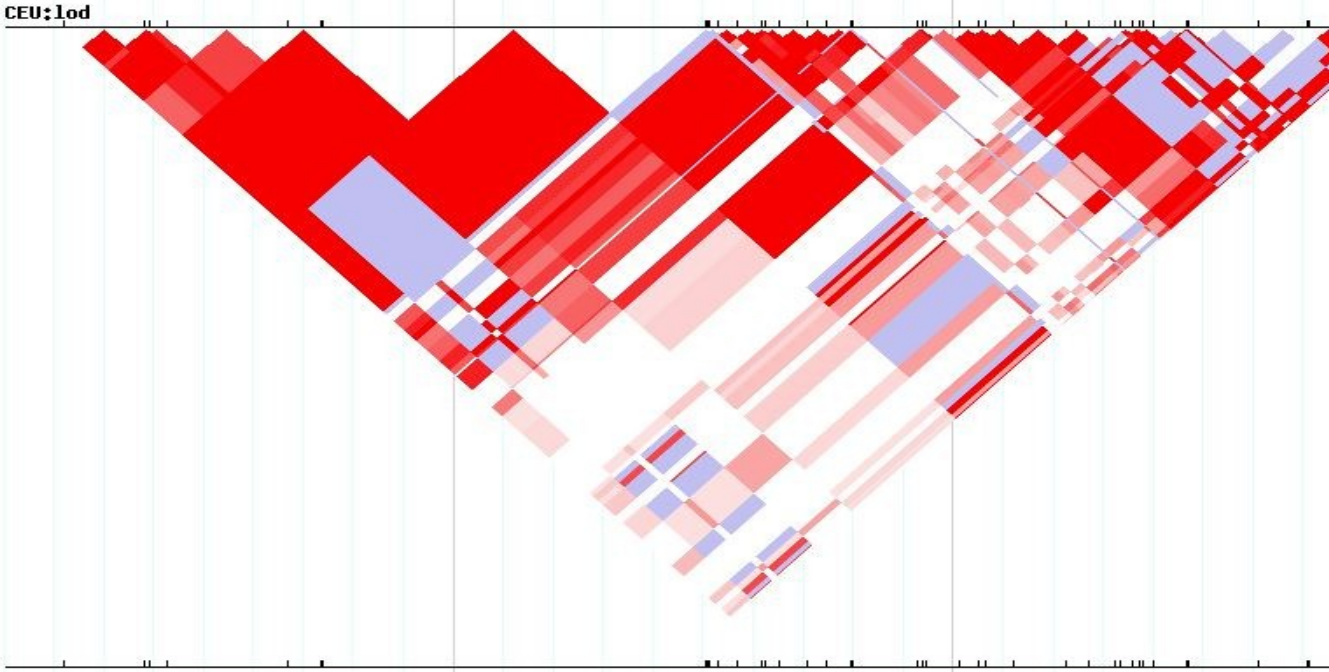
Entrez genes



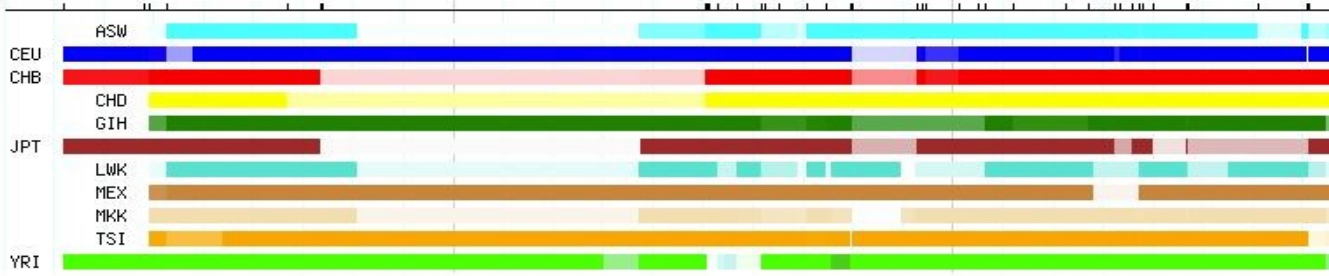
Reactome pathways

Genome-Wide Association studies (NHGRI Catalog)

LD Plot



LD Heat Plot



Clear highlighting

1000 Genomes

A Deep Catalog of Human Genetic Variation

Home About Partners Data Contact Wiki

1000 GENOMES PROJECT DATA RELEASE

SNP data downloads and genome browser representing four high coverage individuals

The first set of SNP calls representing the preliminary analysis of four genome sequences are now available to download through the EBI FTP site and the NCBI FTP site. The README file dealing with the FTP structure will help you find the data you are looking for.

The data can also be viewed directly through the 1000 Genomes browser at http://browser.1000genomes.org. Launch the browser and view a sample region here.

More information about the data release can be found in the data section of this web site.

Download the 1000 Genomes Browser Quick Start Guide

Quick start (pdf)

PRESS RELEASE

WEDNESDAY JUN. 11, 2008
Three Sequencing Companies Join 1000 Genomes Project

TUESDAY JAN. 22, 2008
International Consortium Announces the 1000 Genomes Project

LOG IN

Username:

Password:

Login (I forgot my password)

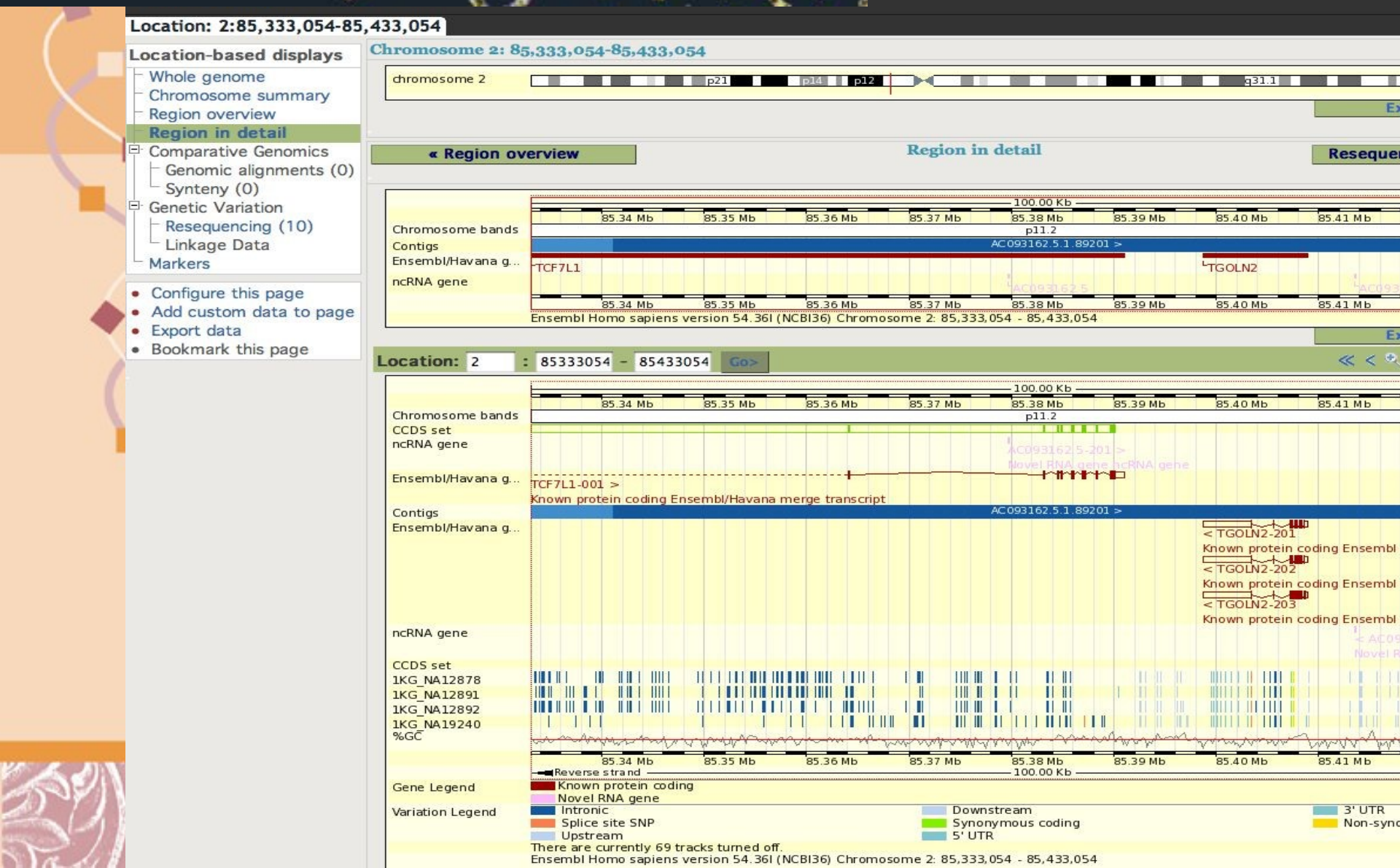
LINKS

Download the meeting report

View the participants

1000 Genomes

A Deep Catalog of Human Genetic Variation



10,000 Genomes Project Evolutionary Biology
http://www.genome.gov/