

---

## Module 6: Genomic Variation

### Aims

- Introduction to genomic variation
- Introduction to various SNP resources on the web
- Integration of information from various databases to identify SNPs in your favourite gene or chromosomal region
- Choosing SNPs to genotype
- Introduction to genotypes and haplotypes

### Introduction

Genetic variation is at the basis of heritable phenotype. Together with the environment, genetic variation makes each one of us different. A key goal of the human genome project is the compilation of a catalogue of common human sequence polymorphisms. With the exception of identical twins, who have identical genomes, differences between two genomes occur on average between 0.3 and 1 kb, equating to 5 - 10 million differences in a genome of 3.2 billion base pairs. Two types of genetic mutation event give rise to all genetic variants. The simplest type of variant is the substitution of a single nucleotide for another, a so-called single nucleotide polymorphism (SNP). SNPs are the commonest form of variation and when comparing 2 genomes, SNPs with a frequency > 1% typically occur every 1000 bp. Insertions or deletions of a section of DNA, so-called INDELS, account for many other types of variation. Variable number tandem repeats (VNTRs) are the commonest type of INDEL and occur where nucleotide patterns are repeated. The difference in size of VNTRs is used to divide them into minisatellites (10 – 100s bp) and simple tandem repeats (STRs or microsatellites) which are 2 – 6 bp in length. It is these SNPs and INDELS that account for most inherited phenotypes, including disease susceptibility. Analysis of sequence variation provides a powerful tool for understanding susceptibility to disease.

A single nucleotide polymorphism (SNP) is defined as a single base change occurring in a population at a frequency >1%. Single base changes that occur at <1% are often referred to as mutations or rare SNPs. However, there is a

lack of agreement between databases on this terminology and some disease-causing mutations occur with quite a high frequency in some populations. For example, the carrier frequencies of mutations in the CFTR gene that cause cystic fibrosis are around 2% in European populations.

SNPs are highly abundant and are thought to be more stable than STRs due to low mutation rates. Nucleotide diversity is lower in exons and approximately half of the exonic SNPs are non-synonymous. SNPs can act as surrogate markers for an adjacent functional variant or can have direct functional consequences if they occur in coding or regulatory regions. The development of high-throughput genotyping platforms makes SNPs well suited to the identification of factors involved in multi-gene diseases as large sample sizes can be analysed quickly.

It is important to remember that the current SNP maps are not exhaustive and rare nucleotide substitutions, that may be critical for disease, may not be represented in the SNP maps. Re-sequencing of genomic DNA from a large number of individuals can be used to identify sequence variation. One such project is the **1000 Genomes Project**, an international research consortium formed to create the most detailed and medically useful picture to date of human genetic variation. The project involves sequencing the genomes of approximately about 2500 unidentified people from about 25 populations around the world. It's being supported by the Wellcome Trust Sanger Institute in England, the Beijing Genomics Institute Shenzhen in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH) in the US. The project draws on the expertise of multidisciplinary research teams and will develop a new map of the human genome that will provide "a view of biomedically relevant DNA variations at a resolution unmatched by current resources". More information about the 1000 Genome Project can be found at the project website: [www.1000genomes.org](http://www.1000genomes.org)

n	1%	5%	10%	20%	30%	40%
2	.21	.30	.36	.43	.47	.49
3	.32	.46	.55	.65	.71	.74
4	.39	.56	.66	.77	.83	.86
5	.44	.62	.73	.84	.90	.93
6	.48	.68	.78	.89	.94	.96
7	.52	.72	.83	.92	.96	.98
8	.55	.75	.86	.94	.98	.99
9	.57	.78	.88	.96	.98	.99
10	.59	.80	.90	.97	.99	.997
16	.69	.89	.96	.99	.999	>.999
24	.76	.95	.99	.999	>.999	>.999
48	.87	.99	.999	>.999	>.999	>.999
96	.95	.999	>.999	>.999	>.999	>.999
192	.99	>.999	>.999	>.999	>.999	>.999

Kruglyak and Nickerson

Table adapted from Kruglyak and Nickerson, Nature Genetics vol 27, page 234, showing the detection rate for SNPs with a given minimal allele frequency in  $n$  chromosomes.

### The Evolution of SNPs

The appearance of mutations and their evolution to SNPs has been defined in four phases (Miller and Kwok, 2001):

1. Appearance of new variant allele by mutation
2. Survival of allele through early generations against the odds
3. Increase of the allele to a substantial population frequency
4. Fixation of allele in populations

Survival of mutations is limited and a lot are lost in early generations. A heterozygous individual having 2 offspring has a 0.75 probability of passing on the mutation to at least one child. If the mutation is neutral, there is a 94 % probability of loss in 10 generations (approximately 200 years). Deleterious mutations disappear more quickly.

### Ambiguity codes:

<b>M</b> => a/c	<b>V</b> => a/c/g	<b>N</b> => a/c/g/t
<b>H</b> => a/c/t	<b>R</b> => a/g	<b>D</b> => a/g/t
<b>W</b> => a/t	<b>S</b> => c/g	<b>B</b> => c/g/t
<b>Y</b> => c/t	<b>K</b> => g/t	

**Worked Example:**

You have been using Affymetrix GeneChips to identify genes that are differentially regulated in patients and controls. Following analysis, the gene corresponding to the probe set 216025\_x\_at is found to be up-regulated in the patient samples from an Affymetrix gene expression experiment. You need to find out information about this gene, in particular whether there are any polymorphisms in the gene and if these could affect its activity. *To determine a complete SNP map for a gene, information from several databases may need to be combined.*

**Questions:**

1. What is the function of the gene that probe set 216025\_x\_at corresponds to?
2. How many single nucleotide polymorphisms are there in this gene?
3. How many of the SNPs are coding and how many alter the amino acid sequence?

To start, access Ensembl, <http://www.ensembl.org> and select the Human homepage to begin the database search



The screenshot shows the Ensembl website interface. At the top, there is a navigation bar with links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. Below this, the 'Human (GRCh38)' section is visible. A search bar is present with the text '216025\_x\_at' entered. A yellow box with the text 'Enter 216025\_x\_at in search window' and an arrow points to the search bar. Below the search bar, there are several options: 'Genome assembly: GRCh38 (GCA\_000001405.15)', 'More information and statistics', 'Download DNA sequence (FASTA)', 'Convert your data to GRCh38 coordinates', and 'Display your data in Ensembl'. There are also links for 'View karyotype' and 'Example region'.

**e!Ensembl** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh38)

Current selection: **Only searching Human**

Restrict category to: **ProbeFeature 1**

Per page: **10** 25 50 100

Layout: **Standard** Table

Tip: You can choose which results appear near the top of your search by updating your favourite species.

Only searching Human **216025\_x\_at**

1 results match 216025\_x\_at when restricted to **species: Human**

**216025\_x\_at (Human AFFY Probe)**  
**216025\_x\_at**  
 AFFY probeset 216025\_x\_at has probes which hit the genome in 9 locations. They hit transcripts in the following gene: CYP2C9 (ENST00000260662).

<< < 1 >>

Ensembl release 78 - December 2014 © WTSI / EBI  
 Permanent link - View in archive site  
 About Ensembl | Privacy Policy | Disclaimer | Contact Us

**Follow OligoProbe**

Click on the image above to jump to a chromosome, or click and drag to select a region

**Key**

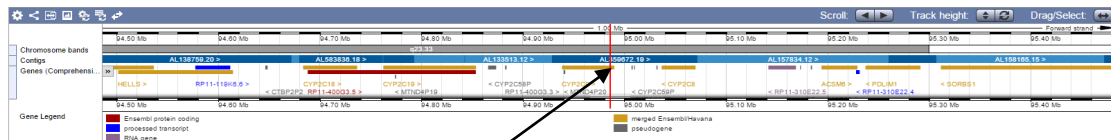
Feature type	Colour
Oligoprobeset	Red
Transcript	Blue

**Oligoprobeset Information**

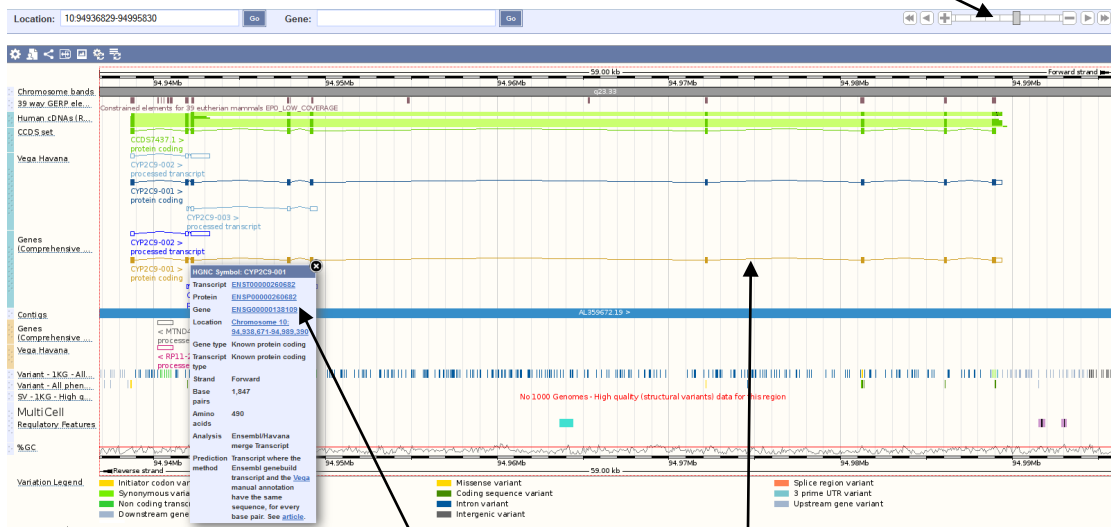
Genomic location (strand)	Length	Ensembl ID
<a href="#">10-94986052-94986076(1)</a>	25	HG-U133A:216025_x_at:323:257 HG-U133A_2:216025_x_at:533:249 HG-U133_Plus_2:216025_x_at:1009:311
<a href="#">10-94986097-94986121(1)</a>	25	HG-U133A:216025_x_at:274:63 HG-U133A_2:216025_x_at:529:61 HG-U133_Plus_2:216025_x_at:1009:311
<a href="#">10-94986148-94986172(1)</a>	25	HG-U133A:216025_x_at:318:455 HG-U133A_2:216025_x_at:317:441 HG-U133_Plus_2:216025_x_at:1009:311
<a href="#">10-94988884-94988908(1)</a>	25	HG-U133A:216025_x_at:310:525 HG-U133A_2:216025_x_at:34:507 HG-U133_Plus_2:216025_x_at:1009:311
<a href="#">10-94988970-94988994(1)</a>	25	HG-U133A:216025_x_at:504:707 HG-U133A_2:216025_x_at:339:685 HG-U133_Plus_2:216025_x_at:1009:311
<a href="#">10-94989078-94989102(1)</a>	25	HG-U133A:216025_x_at:190:607 HG-U133A_2:216025_x_at:686:587 HG-U133_Plus_2:216025_x_at:274:983
<a href="#">10-94989102-94989126(1)</a>	25	HG-U133A:216025_x_at:129:179 HG-U133A_2:216025_x_at:628:173 HG-U133_Plus_2:216025_x_at:1009:311
<a href="#">10-94989137-94989161(1)</a>	25	HG-U133A:216025_x_at:529:103 HG-U133A_2:216025_x_at:303:101 HG-U133_Plus_2:216025_x_at:20:181
<a href="#">10-94989187-94989211(1)</a>	25	HG-U133A:216025_x_at:475:401 HG-U133A_2:216025_x_at:525:389 HG-U133_Plus_2:216025_x_at:732:647

**Check for multiple locations**

**Follow Genomic coordinates**



Drag the Region view, and zoom level in the Detailed view



Follow gene link

Click on gene name to bring up information

Human (GRCh38) Location: 10:94,936,829-94,995,830 Gene: CYP2C9 Transcript: CYP2C9-001

Follow HGNC link

**Gene: CYP2C9** ENSG00000138109

Description: cytochrome P450, family 2, subfamily C, polypeptide 9 [Source:HGNC Symbol;Acc:HGNC:2623]

Synonyms: CYP2C10, P4501IC9

Location: [Chromosome 10: 94,938,658-94,989,390](#) forward strand.

INSDC coordinates: chromosome:GRCh38:CM000672.2:94938658-94989390.1

Transcripts: This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
CYP2C9-001	ENST00000260682	1847	490 aa	Protein coding	CCDS7437	NM_000771 NP_000762	TSL1 GENCODE basic APPRIS PI
CYP2C9-002	ENST00000461906	1431	No protein	Processed transcript	-	-	TSL1
CYP2C9-003	ENST00000473496	841	No protein	Processed transcript	-	-	TSL2

**Summary**

Name: [CYP2C9](#) (HGNC Symbol)

CCDS: This gene is a member of the Human CCDS set: [CCDS7437](#)

UniprotKB: This gene has proteins that correspond to the following Uniprot identifiers: [P11712](#)

RefSeq: Overlapping RefSeq Gene ID [1559](#) matches and has similar biotype of protein\_coding

Ensembl version: ENSG00000138109.9

GRCh37 assembly: This gene maps to [96,698,416-96,749,147](#) in GRCh37 coordinates.

Stable ID ENSG00000138109 not present in GRCh37.

Gene type: Known protein coding

Prediction Method: Annotation for this gene includes both automatic annotation from Ensembl and [Havana](#) manual curation, see [article](#).

Alternative genes: This gene corresponds to the following database identifiers:  
[Havana gene: OTTHUMG0000018805](#)

**HGNC**  
HUGO Gene Nomenclature Committee

Search Genes

Home Search Genes Downloads Gene Families HCOP Useful Links About Contact Us Request Symbol

### Gene Symbol Report

#### CYP2C9

**Approved Symbol** + CYP2C9

**Approved Name** + cytochrome P450, family 2, subfamily C, polypeptide 9

**HGNC ID** + HGNC:2623

**Previous Symbols & Names** + CYP2C10, "cytochrome P450, subfam P4501IC9

**Synonyms** + P4501IC9

**Locus Type** + gene with protein product

**Chromosomal Location** + 10q24.1

**GENE FAMILY** + Cytochrome P450 family C

**HOMOLOGS** + HCOP D TreeFam D

**NUCLEOTIDE SEQUENCES** + GenBank: M61855 EMBL DDBJ C RefSeq: NM\_000771 D CCDS: CCDS7437.1 C Vega: OTTHUMG0000018805 C

**GENE RESOURCES** + Entrez Gene: 1559 C Ensembl: ENSG00000138109 C UCSC: uc001kka.3 D Vega: OTTHUMG0000018805 C

**PROTEIN RESOURCES** + Uniprot: P11712 D Interpro D OMIM D GeneTests D Orphanet D DECIPHER D COSMIC D

**CLINICAL RESOURCES** + LSDB: Human Cytochrome P450 (CYP) PMID: 2009263 PMID: 7841444 CiteXp

**REFERENCES** + GENATLAS D GeneCards D

**GeneCards**  
The Human Gene Compendium Free for academic non-profit institutions. ALL other users need a commercial license from Xenve, Inc.

Home GeneCards Guide Suite Terms and Conditions About Us User Feedback Mirror sites

Set Analyses: [GeneCards](#) [GeneDecks](#) keyword(s) Search Advanced Se

**CYP2C9 Gene**  
protein-coding **GFIS: 63**  
GC10P096688

**cytochrome P450, family 2, subfamily C, polypeptide 9**  
(Previous names: cytochrome P450, subfamily 1IC (mephenytoin 4-hydroxylase), polypeptide 9)  
Symbol approved by the [HUGO Gene Nomenclature Committee \(HGNC\) database](#)  
(Previous symbol: CYP2C10)

**Products**  
SIGMA Pathways Antibodies Peptides / sRNA / esiRNA Small Molecules / sRNA / esiRNA  
M Antibodies / cDNA / RNAi Proteins & Enzymes Assays & Kits / Pathways  
SABiosciences TFBS / miRNA Assays / Genes / sRNA / Primers  
ORIGENE Proteins Antibodies Assays / Genes / sRNA / Primers  
GenScript Proteins Antibody Assays / Cell Lines / Clon

**Aliases & Descriptions for CYP2C9 gene**  
(According to <sup>1</sup>HGNC, <sup>2</sup>Entrez Gene, <sup>3</sup>UniProtKB/Swiss-Prot, <sup>4</sup>UniProtKB/TrEMBL, <sup>5</sup>OMIM, <sup>6</sup>GeneLoc, <sup>7</sup>Ensembl, <sup>8</sup>OMIM, and/or <sup>9</sup>miRBase) [About This Section](#)

**Aliases & Descriptions**  
cytochrome P450, family 2, subfamily C, polypeptide 9<sup>1,2</sup>  
P4501C9<sup>3</sup>  
Cytochrome P-450MP<sup>2,3</sup>  
Cytochrome P450 PB-1<sup>3</sup>  
CYP2C10<sup>3</sup>  
CYP11C9<sup>3</sup>  
cytochrome P450, subfamily 1IC (mephenytoin 4-hydroxylase), polypeptide 9<sup>1</sup>  
CPC9<sup>2</sup>  
CYP2C9  
MGC149605<sup>2</sup>  
MGC88320<sup>2</sup>  
cytochrome P-450 S-mephenytoin 4-hydroxylase<sup>2</sup>  
cytochrome P450 2C9<sup>1</sup>

flavoprotein-linked monooxygenase<sup>2</sup>  
microsomal monooxygenase<sup>2</sup>  
xenobiotic monooxygenase<sup>2</sup>  
(R)-limonene 6-monooxygenase<sup>2</sup>  
(S)-limonene 6-monooxygenase<sup>2</sup>  
(S)-limonene 7-monooxygenase<sup>2</sup>  
EC 1.14.13.48<sup>2</sup>  
EC 1.14.13.49<sup>2</sup>  
EC 1.14.13.80<sup>2</sup>  
Cytochrome P450 MP-4<sup>2</sup>  
Cytochrome P450 MP-9<sup>2</sup>  
S-mephenytoin 4-hydroxylase<sup>2</sup>  
EC 1.14.14.1<sup>1</sup>

**External Ids:** HGNC: 2623<sup>1</sup> Entrez Gene: 1559<sup>2</sup> Ensembl: ENSG00000138109<sup>3</sup> UniProtKB: P11712<sup>4</sup>

[Export aliases for CYP2C9 gene to outside databases](#)

Previous GC identifiers: GC10P095591 GC10P095932 GC10P096829 GC10P096363 GC10P090325

**Summaries for CYP2C9 gene**  
(According to [Entrez Gene](#))

**Entrez Gene summary for CYP2C9:**  
This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. This protein localizes to the endoplasmic reticulum and its expression is induced by rifampin. The enzyme is known to metabolize many xenobiotics, including phenytoin, tolbutamide, ibuprofen and S-warfarin. Studies identifying individuals who are poor metabolizers of phenytoin and tolbutamide suggest that this gene is polymorphic.

Go back to Ensembl GeneView

Human (GRCh38) Location: 10:94,936,829-94,995,830 Gene: CYP2C9 Transcript: CYP2C9-001

**Gene: CYP2C9** ENSG00000138109

Description: cytochrome P450, family 2, subfamily C, polypeptide 9 [Source:HGNC Symbol;Acc:HGNC:2623]

Synonyms: CYP2C10, P450IIC9

Location: [Chromosome 10: 94,938,658-94,989,390](#) forward strand.

INSD coordinates: chromosome:GRCh38:CM000672.2:94938658-94989390:1

Transcripts: This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
CYP2C9-001	ENST00000260682	1847	490 aa	Protein coding	CCDS7437	NM_000771 NP_000762	TSL1 GENCODE basic APPRIS PI
CYP2C9-002	ENST00000461906	1431	No protein	Processed transcript	-	-	TSL1
CYP2C9-003	ENST00000473496	841	No protein	Processed transcript	-	-	TSL2

**Summary**

Name: [CYP2C9](#) (HGNC Symbol)

CCDS: This gene is a member of the Human CCDS set: [CCDS7437](#)

UniProtKB: This gene has proteins that correspond to the following UniProt identifiers: [P11712](#)

RefSeq: Overlapping RefSeq Gene ID [1559](#) matches and has similar biotype of protein\_coding

Ensembl version: ENSG00000138109.3

GRCh37 assembly: This gene maps to [5,698,411-56,749,147](#) in GRCh37 coordinates. Stable ID ENSG00000138109 not present in GRCh37.

Gene type: Known protein coding

Prediction Method: Annotation for this gene includes both automatic annotation from Ensembl and [Havana](#) manual curation, see [article](#).

Alternative genes: This gene corresponds to the following database identifiers:  
[Havana gene: OTHUMG0000018805](#)

Answers 2 & 3 – SNP #'s

COSMIC Variations

Genes: [CYP2C9](#)

Transcript: ENST00000260682 CYP2C9-001

Number of variant consequences: 3102

Number of variant consequences	Type	Description
0	Transcript ablation	A feature ablation whereby the deleted region includes a transcript feature ( <a href="#">GO:0011830</a> )
1	Splice donor variant	A splice variant that changes the 2 base region at the 5' end of an intron ( <a href="#">GO:0011824</a> )
13	Splice acceptor variant	A splice variant that changes the 2 base region at the 3' end of an intron ( <a href="#">GO:0011824</a> )
5	Stop gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript ( <a href="#">GO:0011827</a> )
7	Frameshift variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three ( <a href="#">GO:0011828</a> )
0	Stop lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript ( <a href="#">GO:0011829</a> )
3	Initiator codon variant	A codon variant that changes at least one base of the first codon of a transcript ( <a href="#">GO:0011824</a> )
0	Transcript amplification	A feature amplification of a region containing a transcript ( <a href="#">GO:0011829</a> )
1	Inframe insertion	An inframe non synonymous variant that inserts bases into in the coding sequence ( <a href="#">GO:0011821</a> )
0	Inframe deletion	An inframe non synonymous variant that deletes bases from the coding sequence ( <a href="#">GO:0011822</a> )
183	Missense variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved ( <a href="#">GO:0011823</a> )
18	Splice region variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron ( <a href="#">GO:0011830</a> )
0	Incomplete terminal codon variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed ( <a href="#">GO:0011820</a> )
83	Synonymous variant	A sequence variant where there is no resulting change to the encoded amino acid ( <a href="#">GO:0011819</a> )
1	Stop retained variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains ( <a href="#">GO:0011827</a> )
23	Coding sequence variant	A sequence variant that changes the coding sequence ( <a href="#">GO:0011829</a> )
0	Mature mRNA variant	A transcript variant located with the sequence of the mature mRNA ( <a href="#">GO:0011820</a> )
1	5 prime UTR variant	A UTR variant of the 5' UTR ( <a href="#">GO:0011821</a> )
12	3 prime UTR variant	A UTR variant of the 3' UTR ( <a href="#">GO:0011822</a> )
261	Non coding transcript exon variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript ( <a href="#">GO:0011820</a> )
1631	Intron variant	A transcript variant occurring within an intron ( <a href="#">GO:0011827</a> )
0	NMD transcript variant	A variant in a transcript that is the target of NMD ( <a href="#">GO:0011821</a> )
587	Non coding transcript variant	A transcript variant of a non coding RNA gene ( <a href="#">GO:0011818</a> )
444	Upstream gene variant	A sequence variant located 5' of a gene ( <a href="#">GO:0011821</a> )
437	Downstream gene variant	A sequence variant located 3' of a gene ( <a href="#">GO:0011822</a> )
3102	ALL	All variations



Select "Transcript"

Transcript: CYP2C9-001

Description: cytochrome P450, family 2, subfamily C, polypeptide 9 [Source:HGNC Symbol,Acc:HGNC:2623]  
 CYP2C9, P450C9

Synonyms: CYP2C9

Location: Chromosome 10, 94,338,671-94,359,350 forward strand

Gene: This transcript is a product of gene [ENSG00000138109](#)  
 This gene has 3 transcripts (splice variants) [Hide transcript table](#)

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
CYP2C9-001	ENST00000265882	1847	490 aa	Protein coding	CCDS7437	NM_000771 NP_000762	TSL:1 GENCODE basic APPRIS PI
CYP2C9-002	ENST00000461590	1431	No protein	Processed transcript	-	-	TSL:1
CYP2C9-003	ENST00000473496	841	No protein	Processed transcript	-	-	TSL:2

Follow "Variations"

Variations

Residue	Variation ID	Type	Evidence	Alleles	Ambig. code	Residues	Codons	SIFT	PolyPhen
1	rs114071557	Initiator codon variant		A/G	R	M, V	ATG, GTG	0	0
1	COSM106812	Initiator codon variant		G/A	R	M, I	ATG, ATA	0.02	0
1	rs150891702	Initiator codon variant		G/A	R	M, I	ATG, ATA	0.02	0
2	rs139414138	Missense variant		G/A	R	D, N	GAT, AAT	0.08	0
2	COSM107499	Missense variant		G/A	R	D, N	GAT, AAT	0.08	0
5	rs138957855	Missense variant		T/C	Y	V, A	GTG, GCG	0.1	0
5	rs267502638	Synonymous variant		G/A	R	V	GTG, GTA	-	-
359	rs1057910	Missense variant		A/C	M	I, L	ATT, CTT	0.04	0.45
359	CM960481	Coding sequence variant   Feature elongation		HGMD_MUTATION	-	-	-	-	-
359	rs28371686	Missense variant		C/G	S	D, E	GAC, GAG	0	0.928
360	CM014176	Coding sequence variant   Feature elongation		HGMD_MUTATION	-	-	-	-	-
360	CM090527	Coding sequence variant   Feature elongation		HGMD_MUTATION	-	-	-	-	-
362	rs373048216	Synonymous variant		C/T	Y	L	CTC, CTT	-	-
363	COSM3442014	Missense variant		C/T	Y	P, S	CCC, TCC	0.02	0.998
363	rs11663116	Missense variant		C/T	Y	P, L	CCC, CTC	0	1

Follow rs1057910

rs1057910 SNP

Original source

Alleles

Location

Co-located

Validation status

Clinical significance

Synonyms

HGVs names

Genotyping chips

Variants (including SNPs and indels) imported from dbSNP (release 137) | [View in dbSNP](#)

Reference/Alternative: **A/C** | Ancestral: **A** | Ambiguity code: **M** | MAF: **0.04** (C)

Chromosome **10:96741053** (forward strand) | [View in location tab](#)

with HGMD-PUBLIC [CM960481](#)

This variation is validated by **1000 Genomes**, **HapMap** and also frequency

**pathogenic** (from dbSNP) | [View explanation](#)

This feature has **6** synonyms - click the plus to show

This feature has **3** HGVs names - click the plus to show

This variation has assays on **8** chips - click the plus to show

Follow link to dbSNP

**Build information**

**Alleles**

**Integrated map information**

Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh38	106	10	94981296	NT_030059.14	5328775	Fwd	A	Fwd	view	remap
GRCh37.p13	105	10	96741053	NT_030059.13	47545517	Fwd	A	Fwd	view	blast

Scrolling down the page reveals information about the allele frequency and a link through to genotypes.

**Sequence context of SNP**

NCBI Assay ID	Handle/Submitter ID	Validation Status	SS Assay Orientation (Strand)	Alleles	5' Near Seq 30 bp	3' Near Seq 30 bp	Entry Date	Update Date	Build Added	Molecule Type	Freq Warning	Ancestral Allele	Success Rate
ss1538933	LEE1741019		fwd/TT	A/C	gatgctgtggtgcacaggtccagagatct	tgaccttctccccaccagctgccccatg	09/13/00	10/10/0386		cDNA			unknown
ss2419888	HJ81A3ESHP00000187		fwd/TT	A/C	tgtgtgtgcacaggtccagagatct	tgaccttctccccaccagctgccc	11/07/00	10/10/0389		Genomic			unknown
ss4426472	LEE0741019		fwd/TT	A/C	gatgctgtggtgcacaggtccagagatct	tgaccttctccccaccagctgccccatg	04/26/02	10/10/03106		cDNA			unknown
ss5588419	SNP500CANCERCYP2C9-01		fwd/TT	A/C	gatgctgtggtgcacaggtccagagatct	tgaccttctccccaccagctgccccatg	09/28/02	04/07/04113		Genomic			unknown
ss12588583	EGP_SNPICYP2C9-045324		fwd/TT	A/C	gatgctgtggtgcacaggtccagagatct	tgaccttctccccaccagctgccccatg	08/20/03	04/07/04119		Genomic			unknown
ss28501344	MCJ-GDTIMCJ-CYP2C9_9-AC		fwd/TT	A/C	gatgctgtggtgcacaggtccagagatct	tgaccttctccccaccagctgccccatg	08/20/04	08/20/04126		Genomic			unknown

**Genotype and allele information**

ss#	Population	Group	Sample Cnt.	Source	A/A	A/C	HWP	A	C
<a href="#">ss105107895PA152209538</a>			184	AF				0.918	0.082
<a href="#">ss105108091PA152211301</a>			696	AF				0.953	0.047
<a href="#">ss105109763PA154394460</a>			584	AF				0.938	0.062
<a href="#">ss12588583_PDR90</a>	Global		176	IG	0.920	0.080	0.752	0.960	0.040

Scroll up and go to Gene (ID)

RefSeqGene	Gene (ID)	RefSeqGene Mapping	SNP to RefSeqGene	Position	Allele
<a href="#">NG_008385.1</a>	<a href="#">CYP2C9 (1859)</a>		Fwd	<a href="#">47839</a>	A

**CYP2C9 cytochrome P450, family 2, subfamily C, polypeptide 9 [ *Homo sapiens* (human) ]**

Gene ID: 1559, updated on 7-Dec-2014

**Summary**

**Official Symbol** CYP2C9 provided by HGNC

**Official Full Name** cytochrome P450, family 2, subfamily C, polypeptide 9 provided by HGNC

**Primary source** [HGNC:HGNC:2623](#)

**Locus tag** RP11-208C17.6

**See related** [Ensembl:ENSG00000138103](#); [HPRD:0308](#); [MIM:601130](#); [Ensembl:OTTHUMG00000018805](#)

**Gene type** protein coding

**RefSeq status** REVIEWED

**Organism** [Homo sapiens](#)

**Lineage** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo

**Also known as** CYP2C9, CYP2C, CYP2C10, CYP11C9, P45011C9

**Summary** This gene encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. This protein localizes to the endoplasmic reticulum and its expression is induced by rifampin. The enzyme is known to metabolize many xenobiotics, including phenytoin, tolbutamide, ibuprofen and S-warfarin. Studies identifying individuals who are poor metabolizers of phenytoin and tolbutamide suggest that this gene is polymorphic. The gene is located within a cluster of cytochrome P450 genes on chromosome 10q24. [provided by RefSeq, Jul 2008]

Follow OMIM link

\*601130

CYTOCHROME P450, SUBFAMILY IIC, POLYPEPTIDE 9; CYP2C9

*HGNC Approved Gene Symbol:* CYP2C9

*Cytogenetic location:* 10q23.33 *Genomic coordinates (GRCh37):* 10:96,698,349-96,749,485 View NCBI

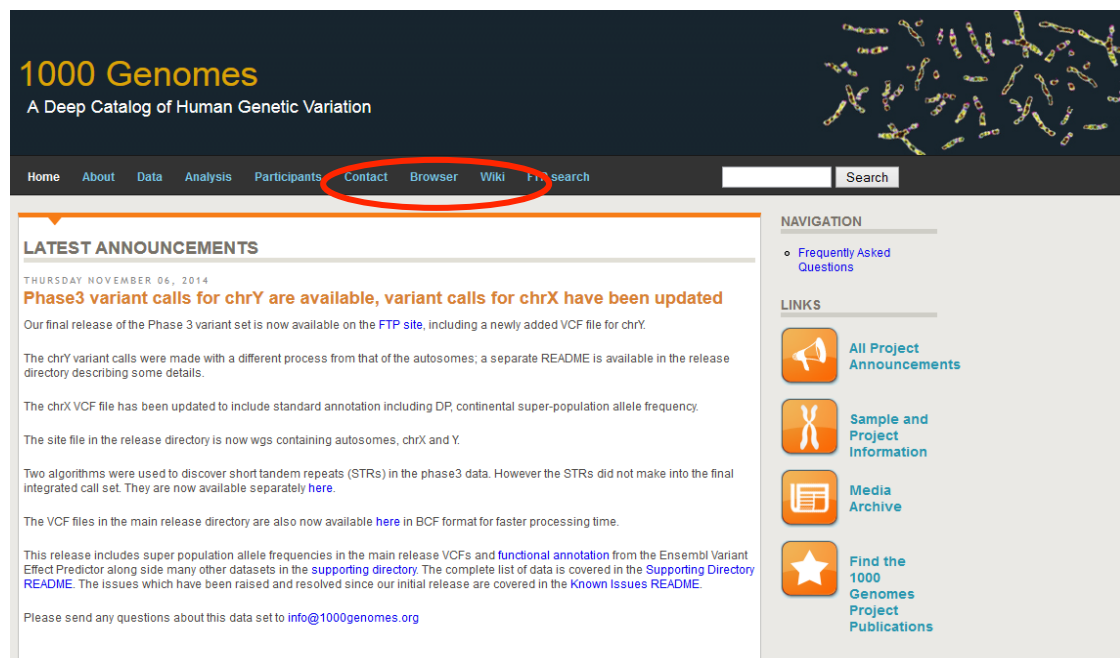
**Gene-Phenotype Relationships**

Location	Phenotype	Phenotype MIM number	Phenotype mapping key
10q23.33	Tolbutamide poor metabolizer Warfarin sensitivity	122700	3 3

## The 1000 Genomes Project

Data from increasing numbers of human genome sequences is currently available through the public website ([www.1000genomes.org](http://www.1000genomes.org)).

These can be searched through the familiar Ensembl browser format for displaying data ([browser.1000genomes.org](http://browser.1000genomes.org))



**1000 Genomes**  
A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants **Contact** Browser Wiki FTP search

**LATEST ANNOUNCEMENTS**

THURSDAY NOVEMBER 06, 2014

**Phase3 variant calls for chrY are available, variant calls for chrX have been updated**

Our final release of the Phase 3 variant set is now available on the [FTP site](#), including a newly added VCF file for chrY.

The chrY variant calls were made with a different process from that of the autosomes; a separate README is available in the release directory describing some details.

The chrX VCF file has been updated to include standard annotation including DP, continental super-population allele frequency.

The site file in the release directory is now wgs containing autosomes, chrX and Y.

Two algorithms were used to discover short tandem repeats (STRs) in the phase3 data. However the STRs did not make into the final integrated call set. They are now available separately [here](#).

The VCF files in the main release directory are also now available [here](#) in BCF format for faster processing time.

This release includes super population allele frequencies in the main release VCFs and [functional annotation](#) from the Ensembl Variant Effect Predictor along side many other datasets in the [supporting directory](#). The complete list of data is covered in the [Supporting Directory README](#). The issues which have been raised and resolved since our initial release are covered in the [Known Issues README](#).

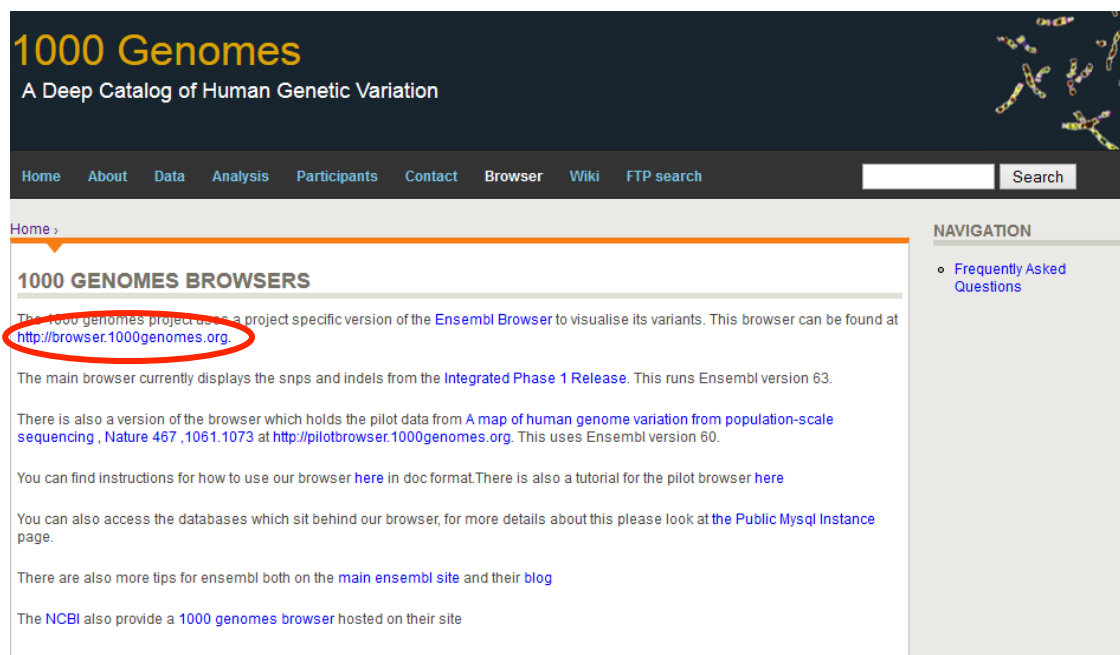
Please send any questions about this data set to [info@1000genomes.org](mailto:info@1000genomes.org)

**NAVIGATION**

- Frequently Asked Questions

**LINKS**

- All Project Announcements
- Sample and Project Information
- Media Archive
- Find the 1000 Genomes Project Publications



**1000 Genomes**  
A Deep Catalog of Human Genetic Variation

Home About Data Analysis Participants Contact **Browser** Wiki FTP search

Home >

**1000 GENOMES BROWSERS**

The 1000 genomes project uses a project specific version of the [Ensembl Browser](#) to visualise its variants. This browser can be found at <http://browser.1000genomes.org>.

The main browser currently displays the snps and indels from the [Integrated Phase 1 Release](#). This runs Ensembl version 63.

There is also a version of the browser which holds the pilot data from [A map of human genome variation from population-scale sequencing](#), *Nature* 467, 1061-1073 at <http://pilotbrowser.1000genomes.org>. This uses Ensembl version 60.

You can find instructions for how to use our browser [here](#) in doc format. There is also a tutorial for the pilot browser [here](#)

You can also access the databases which sit behind our browser, for more details about this please look at the [Public Mysql Instance](#) page.

There are also more tips for ensembl both on the [main ensembl site](#) and their [blog](#)

The NCBI also provide a [1000 genomes browser](#) hosted on their site

**NAVIGATION**

- Frequently Asked Questions

# 1000 Genomes

A Deep Catalog of Human Genetic Variation

**Search 1000 Genomes**

[View details for BRCA2 or Chromosome 6:133098746-133108745](#)

---

**Start Browsing 1000 Genomes data**

[Browse Human](#) →  
GRCh37

[Protein variations](#) →  
View the consequences of sequence variation at the level of each protein in the genome.

[Individual genotypes](#) →  
Show different individual's genotype, for a variant.

## The project uses the familiar Ensembl browser format:

**Search 1000 Genomes**

[New Search](#)

**Results Summary**

[Configure this page](#)

[Add your data](#)

[Export data](#)

[Get VCF data](#)

[Bookmark this page](#)

[Share this page](#)

[View in Ensembl](#)

**Results Summary**

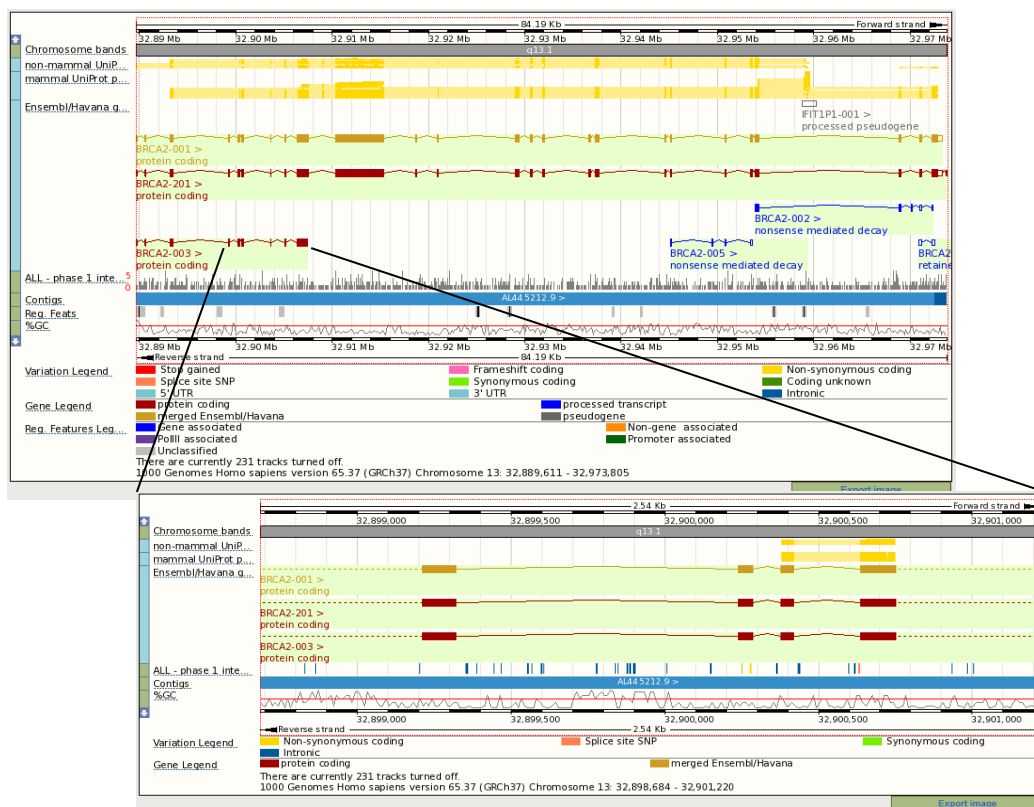
You searched for 'BRCA2'

**Gene or Gene Product**

18 entrie(s) matched your search strings.

1. **Gene:** [ENSG00000170037](#) Region in detail [[17:7835419-7853236](#)] centromerin, centrosomal BRCA2 interacting protein [Source:HGNC Symbol;Acc:29616]
2. **Variations in gene** [ENSG00000170037](#): Variations in gene [[17:7835419-7853236](#)]
3. **Gene:** [ENSG00000107949](#) Region in detail [[10:127512115-127542264](#)] BRCA2 and CDKN1A interacting protein [Source:HGNC Symbol;Acc:978]
4. **Variations in gene** [ENSG00000107949](#): Variations in gene [[10:127512115-127542264](#)]
5. **Gene:** [ENSG00000185515](#) Region in detail [[X:154299695-154351349](#)] BRCA1/BRCA2-containing complex, subunit 3 [Source:HGNC Symbol;Acc:24185]
6. **Variations in gene** [ENSG00000185515](#): Variations in gene [[X:154299695-154351349](#)]
7. **Gene:** [ENSG00000083093](#) Region in detail [[16:23614488-23652631](#)] partner and localizer of BRCA2 [Source:HGNC Symbol;Acc:26144]
8. **Variations in gene** [ENSG00000083093](#): Variations in gene [[16:23614488-23652631](#)]
9. **Gene (from Patch 2013-07-15 14:43:48):** [ENSG00000269884](#) Region in detail [[HG1497\\_PATCH:154239888-154291547](#)] BRCA1/BRCA2-containing complex, subunit 3 [Source:HGNC Symbol;Acc:24185]
10. **Variations in gene** [ENSG00000269884 \(from Patch 2013-07-15 14:43:48\)](#): Variations in gene [[HG1497\\_PATCH:154239888-154291547](#)]
11. **Gene:** [LRG\\_308](#) Region in detail [[LRG\\_308:5001-43196](#)] partner and localizer of BRCA2 [Source:HGNC Symbol;Acc:26144]
12. **Variations in gene** [LRG\\_308](#): Variations in gene [[LRG\\_308:5001-43196](#)]
13. **Gene:** [ENSG00000139618](#) Region in detail [[13:32889611-32973805](#)] BRCA2 - breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]

## Make sure you select *BRCA2 - breast cancer 2, early onset* **ENSG00000139618**



## Choosing SNPs to Genotype

Once you have identified the SNPs associated with your gene(s) of interest it is time to get some data. However, as we have seen there are typically dozens of SNPs in or close to any given gene. So how do you decide which ones to look at? There are no set rules for this, but you should take into consideration:

- 1) Previously published data
- 2) Validation status
- 3) Population frequency
- 4) *In silico* predictions of potential phenotypic effect
- 5) Haplotypes

### Previously published data

The first place to look for SNPs is a literature database such as PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>) using the appropriate search terms (eg 'polymorphism' AND 'CYP17A1'). If somebody has already found a functional SNP, or linked a SNP to a disease then it is an excellent candidate for your study.

The advent of microarrays means that some groups are investigating associations between SNPs and gene expression levels on a large scale, so called expression Quantitative Trait Loci (eQTL). For example the GENE Expression VARIation (GENEVAR) project at the Sanger has looked at the expression of 48,000 genes in all 270 HapMap lymphoblastoid cell lines. A searchable database is under construction to allow you to see what SNPs (if any) are associated with expression of your gene(s) of interest. Data is currently available at: <http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/>

Genome-wide association studies (GWAS) are becoming increasingly common. The US National Human Genome Research Institute maintains a searchable catalogue of GWAS (<http://www.genome.gov/gwastudies/>). If you

know what region or disease you are interested in you can see if any SNPs have already been linked to it:

Division of Genomic Medicine

[Share](#) [Print](#)

A Catalog of Published Genome-Wide Association Studies

[Division Staff](#) | [Funding Opportunities](#) | [Genomic Medicine Activities](#) | [GWAS Catalog](#) | [Meetings & Workshops](#) | [Potential Sample Collections for Sequencing](#) | [Programs](#) | [Publications](#) | [Trans-NIH Sequencing Inventory](#)

Additional information has been added to the HTML catalog columns below. For a description of column headings for the HTML catalog, go to: [Catalog Heading Descriptions](#)

[Potential etiologic and functional implications of genome-wide association loci for human diseases and traits](#)  
Click here to read our recent *Proceedings of the Academy of Sciences (PNAS)* article on catalog methods and analysis.

[View the Full Catalog](#) [Download the Catalog](#) [Search the Catalog](#)

Search By:

Journal:

First Author:

Disease/Trait:

(string search)

or

- Attention deficit hyperactivity disorder symptoms (interaction)
- Autism
- Basal cell carcinoma (cutaneous)
- Behcet's disease
- Beta thalassemia/hemoglobin E disease
- Bilirubin levels
- Biochemical measures
- Biomedical quantitative traits
- Bipolar disorder
- Black vs. blond hair color

(hold Ctrl-key when selecting multiple entries)

Chromosomal Region:

Gene:  (e.g., "LRP5")

SNP:  (e.g., "rs20755555")

OR greater than:

p-Value threshold:  Enter the exponent. For example, enter "5" for  $p < 10^{-5}$







Type in **OXTR**

The search returns details of the study and SNPs identified:

Date Added to Catalog (since 11/25/08)	First Author/Date/Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Mapped Gene (s)	Strongest SNP-Risk Allele	Context	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]	CNV
12/14/11	Edwards AC November 05, 2011 <i>Psychiatr Genet</i> <a href="#">Genome-wide association study of comorbid depressive syndrome and alcohol dependence.</a>	Depression and alcohol dependence	467 European ancestry cases, 407 European ancestry controls	NR	3p25.3	OXTR	OXTR	rs237899-?	intron	NR	$2 \times 10^{-8}$	1.69 [1.36-2.10]	Illumina [876,476]	N

## Validation status

It is important to remember that not all SNPs in dbSNP are useful or even real. Some are sequencing errors and others may be unique to the individual they were found in. Looking at the validation status of a SNP gives an idea of how reliable it is. Unvalidated SNPs should be treated with caution. SNPs with frequency information available are best.

Validation status description	
	Validated by multiple, independent submissions to the refSNP cluster
	Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
	Validated by submitter confirmation
	All alleles have been observed in at least two chromosomes apiece
	Genotyped by HapMap project
	SNP has been sequenced in 1000Genome project.

## Population frequency

The population frequency of a SNP should match the study that you are conducting. How many samples do you have? Do you have the power to detect associations in SNPs with a frequency of 5%, 10% etc? If not then there's little point genotyping SNPs that have such low frequencies.

## *In silico* predictions of potential phenotypic effect

SNPs are more likely to have phenotypic effects if they are:

- 1) Frameshift
- 2) Non-synonymous
- 3) Synonymous (exonic splice enhancer/suppressor)
- 4) Splice site
- 5) Untranslated region (UTR)
- 6) In regulatory regions

Expressed SNPs are easy to identify as shown previously in this module, but determining which of these to prioritize is not always obvious.



## Genotyping data and haplotypes

Consider two SNPs

_____	A	_____	T
_____	C	_____	A

In equilibrium:  
(i.e. no LD)

_____	A	_____	T
_____	A	_____	A
_____	C	_____	T
_____	C	_____	A

In disequilibrium:  
(i.e. LD)

_____	A	_____	T
_____	C	_____	A

## Linkage and Linkage Disequilibrium (LD)

Linkage and linkage-disequilibrium (LD) measure a correlation, co-segregation or association between a genetic marker and disease. They can be distinguished a number of ways:

1. Linkage is focused on a locus whilst LD is focused on an allele
2. Linkage results from recombination events in the last 2-3 generations. LD on the other hand results from much earlier, ancestral recombination events
3. Linkage measures co-segregation in a pedigree. LD measures co-segregation in a population (essentially a very large pedigree)
4. From the dynamical system point of view, Linkage is the "dynamical equation", LD is the "initial condition"
5. In a pedigree likelihood calculation (LOD score), the result tells you whether you have Linkage or not. Conversely, LD is provided by the user prior to performing the calculation.
6. Linkage is usually detected for markers reasonable close to the disease gene (one centiMorgan). LD is detected for markers even closer (0.01-0.02 cM).

Several metrics have been devised to measure linkage disequilibrium (LD). The two most commonly used of these are  $D'$  and  $r^2$ . Both are related to the basic unit of LD,  $D$ .

### **$D$**

$D$  measures the deviation of haplotype frequencies from the equilibrium state. LD occurs when  $D$  is significantly greater than zero. Consider two linked SNPs with alleles ( $A, a$ ) and ( $B, b$ ), resulting in four possible haplotypes:  $AB, Ab, aB$  and  $ab$ .  $D$  can be calculated as in equation 1, where  $f(X)$  represents the frequency of the  $X$  allele or haplotype.

$$D=f(AB)-f(A)f(B) \quad (1)$$

### **$D'$**

$D'$  is the absolute ratio of  $D$  compared with its maximum value,  $D_{\max}$ , when  $D \geq 0$ , or compared with its minimal value,  $D_{\min}$ , when  $D < 0$ .  $D'=1$  denotes complete LD, and historical recombination results in the decay of  $D'$  towards zero.

### **$r^2$**

$r^2$  is the statistical coefficient of determination – a measurement of correlation between a pair of variables (see equation 2).

$$r^2 = \frac{D^2}{f(A) f(a) f(B) f(b)} \quad (2)$$

$r^2$  is of particular importance in genetic mapping as it is inversely related to the required sample size for association mapping, given a fixed genetic effect. For example, if only one pair of SNPs was genotyped and  $r^2$  between the SNPs was found to be 0.5, then to provide the same statistical power for ungenotyped SNP compared with the case where  $r^2=1$ , knowing the genotypes of alleles of one SNP is directly predictive of the genotypes of another SNP. The alternative notation  $R^2$  is used when individual variables are predicted using the multiple regression of a constellation of other variables.

### Relationship between $D'$ and $r^2$

$D'$  and  $r^2$  can be written in terms of each other and allele frequencies. Without losing generality, the four alleles can be chosen such that  $D \geq 0$  and  $f(A) \geq f(B)$ .

So  $D'$  and  $D_{\max}$  have the relations in equations 3 and 4.

$$D' = \frac{D}{D_{\max}} \quad (3)$$

$$D_{\max} = f(a)f(B) \quad (4)$$

and

$$r^2 = (D')^2 \times \frac{f(a)f(B)}{f(A)f(b)} \quad (5)$$

Equation 5 shows the relationship between  $D'$ ,  $r^2$  and allele frequencies. As  $f(A) \geq f(B)$ ,  $r^2$  has the upper bound of  $(D')^2$ , and reaches it only when  $f(A) = f(B)$ . The implication of this is that  $D'$ , a commonly used measure of historical recombination, provides information on the physical extent of useful LD (in terms of association mapping and statistical power) by providing the upper limit of  $r^2$ . Dense LD maps that are based on high frequency SNPs (MAF > 0.1) can reveal regions of historical recombination. Knowing the level of  $D'$  decay in these maps directly provides the maximum potential level of useful LD in association mapping (based on  $r^2$ ) for high-frequency SNPs even if a significant proportion of common SNPs remains undiscovered. For example, if a recombination point resulted in a  $D'$  of 0.7 for SNPs on either side of it, the maximum possible  $r^2$  for these SNPs would be 0.49, and sample sizes would need to be more than doubled to maintain the same statistical power for association mapping. It should be noted that both  $D'$  and  $r^2$  suffer from sampling biases given a small number of individuals and for rare variants. Confidence intervals for  $D'$  have been used by some investigators. LD varies throughout the human genome. Near complete LD has been observed over >800 Kb on chromosome 22, however, these regions of near complete LD are interspersed with regions of little or no LD.

There are a number of web based programs available for determining linkage disequilibrium such as Arlequin (<http://cmpg.unibe.ch/software/arlequin3/>)

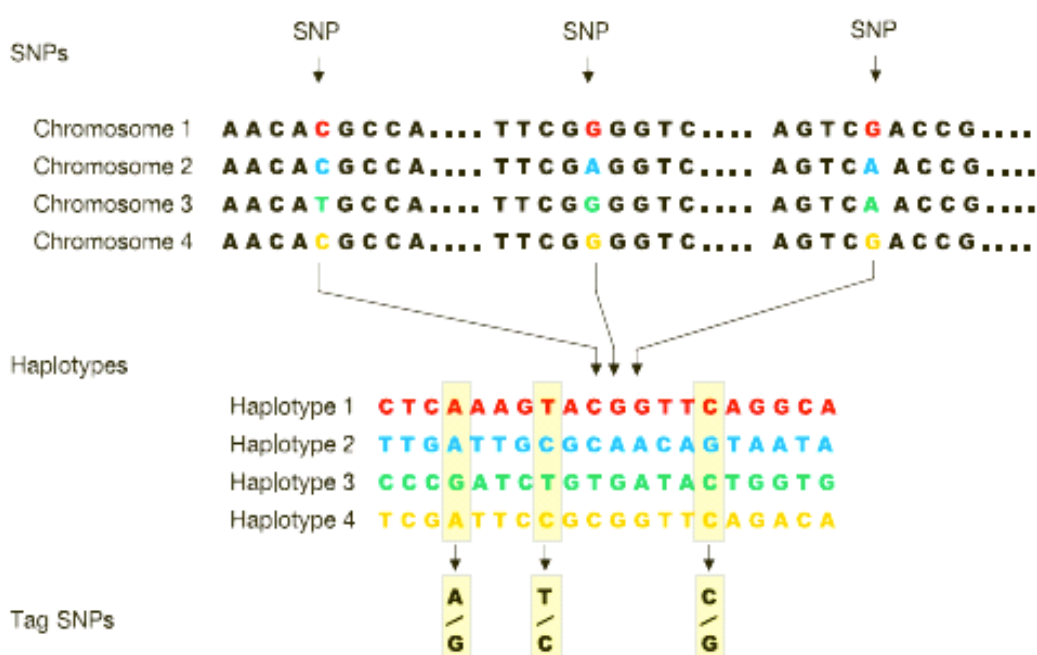
Genepop (<http://genepop.curtin.edu.au/>) and Haploview ([www.broad.mit.edu/personal/jcbarret/haploview/](http://www.broad.mit.edu/personal/jcbarret/haploview/)).

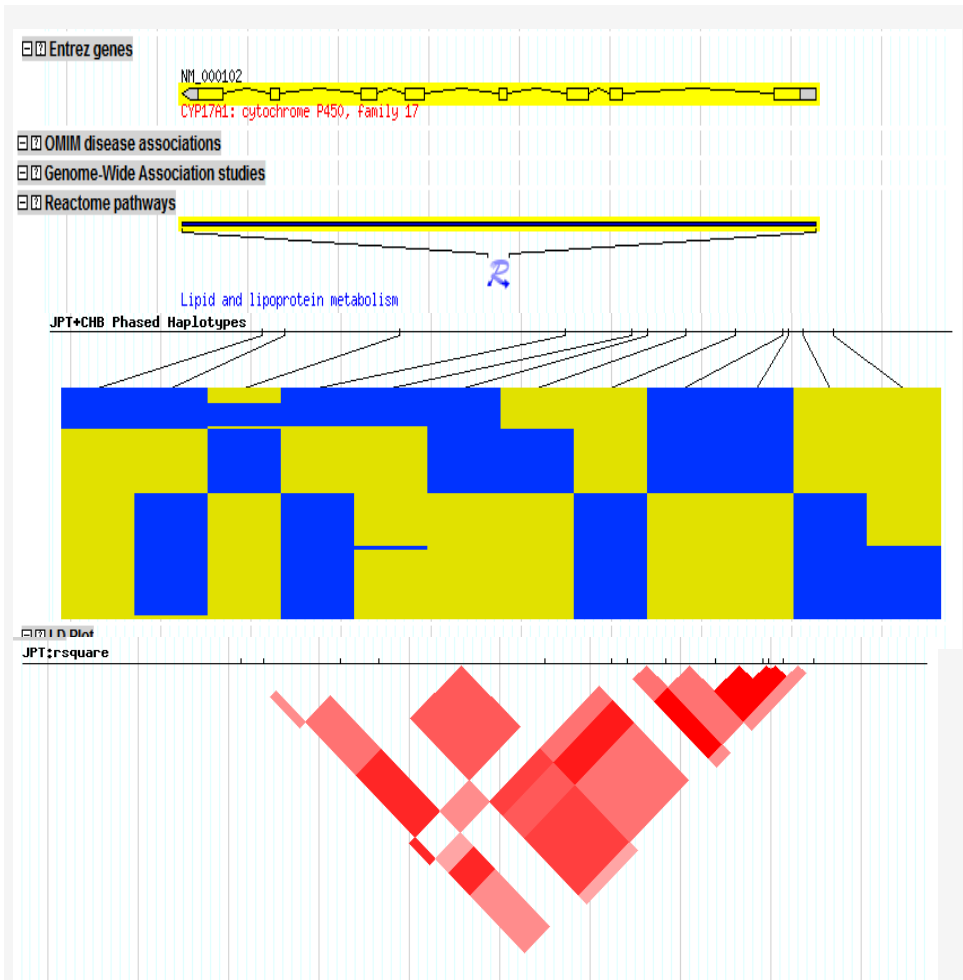
## Haplotypes

While individual SNPs can have an effect on gene expression or protein structure they do not exist in isolation, but as part of haplotypes. Haplotypes are blocks of sequence that derive from the same ancestral chromosome and have not been disrupted by recombination. Haplotypes are defined by groups of closely linked alleles that tend to be inherited together. Thus each SNP investigated is linked to and interacts with a number of other SNPs. As high throughput genotyping platforms become more widely used, attention will inevitably move from individual SNPs to haplotypes.

### Taking Haplotype block structure into account

- Discrete chromosome region of high LD and low haplotype diversity
- All pairs of polymorphisms within a block are in strong LD, whereas other pairs show weaker association
- Blocks hypothesized to be regions of low recombination flanked by recombination hotspots: In other words, SNPs in blocks have had little ancestral recombination happen between them.





	rs10883783	rs17115100	rs4919686	rs3740397	rs1004467	rs755443	rs4919687	rs3781287	rs3824755	rs10786712	rs6163	rs6162	rs762563	rs743572	rs2486758	rs17115149
NA18545_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18608_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18611_c1	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18537_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18526_c1	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18532_c1	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18974_c1	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18953_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18945_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18943_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18635_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18632_c2	A	G	A	G	A	G	A	G	G	T	A	A	G	G	T	T
NA18540_c2	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18555_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18570_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18571_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18572_c2	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18593_c2	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18594_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18603_c2	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18609_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18947_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18960_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18526_c2	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18995_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18994_c1	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18976_c2	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18965_c2	A	G	C	C	A	G	A	G	G	T	A	A	G	G	T	G
NA18577_c1	A	G	A	C	A	A	A	G	G	T	A	A	G	G	T	G
NA18973_c2	A	G	A	C	A	A	A	G	G	T	A	A	G	G	T	G
NA18637_c2	A	G	A	G	G	G	G	G	G	T	A	A	G	G	T	T
NA18940_c1	A	G	A	G	G	G	G	G	G	T	A	A	G	G	T	T
NA18582_c2	T	T	A	C	G	G	G	G	C	T	A	A	C	G	T	G
NA18542_c2	T	T	A	C	G	G	G	G	C	T	A	A	G	G	T	G
NA18545_c1	T	T	A	C	G	G	G	G	C	T	A	A	G	G	T	G

## The International HapMap Project

This is a multi-country project to identify and catalogue genetic similarities and differences in human beings <http://www.hapmap.org>. In phase I & II 270 DNAs haplotyped were from the CEPH, Han Chinese, Japanese and Yoruba populations 6 million SNPs. This means that there will be a common, genotyped SNP every 600bp on average. The Generic Genome Browser enables one to look at the genotyped SNPs associated with a particular region or landmark and provides links to frequency and genotype data.

- Goal: Determine common patterns of DNA sequence variation in human genome in samples from different populations
- Attempts to capture most of the variation due to existing ~10 million SNPs by genotyping 200,000-1,000,000 tag SNPs
- BUT by focusing on common variants, may miss rare, disease-associated variants.

- African ancestry in Southwest USA
  - Utah residents with Northern and Western European ancestry from the CEPH collection
  - Han Chinese in Beijing, China
  - Chinese in Metropolitan Denver, Colorado
  - Gujarati Indians in Houston, Texas
  - Japanese in Tokyo, Japan
  - Luhya in Webuye, Kenya
  - Mexican ancestry in Los Angeles, California
  - Maasai in Kinyawa, Kenya
  - Toscani in Italia
  - Yoruba in Ibadan, Nigeria
- **ASW**
  - **CEU**
  - **CHB**
  - **CHD**
  - **GIH**
  - **JPT**
  - **LWK**
  - **MXL**
  - **MKK**
  - **TSI**
  - **YRI**

A HapMap tutorial section (<http://hapmap.ncbi.nlm.nih.gov/tutorials.html>) includes presentations from the the HapMap tutorials at American Society of Human Genetics Annual Convention on the 27th of October 2005 and a 'Users Guide to the web site'. The Generic Genome Browser enables one to look at the genotyped SNPs associated with a particular region or landmark and provides links to frequency and genotype data.

### Worked Example:

You have identified IL7R as being differentially expressed in patients and controls. You want to see what SNPs have been genotyped for IL7R as part of the HapMap project and to find out their frequencies in the Caucasian population.

### Questions:

1. How many SNPs have been genotyped within 10kb of IL7R?
2. How many of these have minor allele frequencies  $>0.1$  in the Caucasian HapMap population?
3. Are any of the SNPs in linkage disequilibrium?
4. What are the tag SNPs?

To start, access HapMap, <http://hapmap.ncbi.nlm.nih.gov/>

The screenshot shows the International HapMap Project website. A yellow callout box on the right side of the page contains the text: "Step 1: Click on HapMap3 Genome Browser release #27 (Phase 1, 2 & 3 - genotypes, frequencies)". An arrow points from this callout to the corresponding entry in the "Project Data" section of the website. The "Project Data" section lists several releases, including "HapMap3 Genome Browser release #27 (Phase 1, 2 & 3 - merged genotypes & frequencies)".

NB. Since so much has to be precomputed, the visualisation features (especially plots) lag behind the latest data releases. Below we use release 27 containing data from phases 1, 2 & 3, where these tools are available at the time of writing.

**Step 2: Choose 'Annotate LD Plot'. Click Configure**

to change magnification and position.

**Instructions**  
**Searching:** Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed.  
**Navigation:** Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.  
**Examples:** Chr20, Chr9:660,000..760,000, SNP:rs870660, NM\_153254, BRCA2, 5q31, ENM010, gwa\*, PARK3.

[Help] [Reset]  
 Search

Help links:

Landmark or Region: IL7R Search

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

Data Source: HapMap Data Rel 27 PhaseIII, Feb09, on NCBI B36 assembly, dbSNP b126

Population descriptors: ASW: African ancestry in Southwest USA, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing, China, CHD: Chinese in Metropolitan Denver, Colorado, GIH: Gujarati Indians in Houston, Texas, JPT: Japanese in Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MEX: Mexican ancestry in Los Angeles, California, MKK: Maasai in Kinyawa, Kenya, TSI: Tuscans in Italy, YRI: Yoruban in Ibadan, Nigeria.

Make sure 'LD Plot' track is switched on (bottom of page) also 'Phased Haplotype Display' if available for that dataset.

**Variation**  All on  All off

dbSNP SNPs  Genotyped SNPs  Recombination rate (cM/Mb)  Sequence Tagged Sites

**Analysis**  All on  All off

LD Heat Plot  LD Plot

[Configure tracks...] [Update Image]

**Display Settings**

**Step 3: Choose  $r^2$  for LD properties >0.3 and <1.0**

**Configure... LD Plot**

Cancel [Configure]

Max. Segment Size: 250Kb Max. # gt'd SNPs: 200 Box Size: Proportionate

LD Properties: rsquare greater than and less than  
 -1 250

Color: Pairwise plot  
 red

Populations: ASW CEU CHB CHD GIH JPT LWK MEX MKK TSI YRI  
 off  on  off  on  off  on  off  on  off  on  off  on  off  on  off  on  off  on  off  on  off  on

Orientation: invert normal invert normal invert invert invert invert normal invert normal invert

Cancel [Configure]

**Step 4: Turn on the European (CEU) and Japanese (JPT) populations, inverting the latter.**

**Step 5: Click configure**

**Instructions**  
**Searching:** Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed.  
**Navigation:** Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.  
**Examples:** Chr20, Chr9:660,000..760,000, SNP:rs870660, NM\_153254, BRCA2, 5q31, ENM010, gwa\*, PARK3.

[Help] [Reset]  
 Search

Help links:

Landmark or Region: IL7R Search

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

Data Source: HapMap Data Rel 27 PhaseIII, Feb09, on NCBI B36 assembly, dbSNP b126

Population descriptors: ASW: African ancestry in Southwest USA, CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing, China, CHD: Chinese in Metropolitan Denver, Colorado, GIH: Gujarati Indians in Houston, Texas, JPT: Japanese in Tokyo, Japan, LWK: Luhya in Webuye, Kenya, MEX: Mexican ancestry in Los Angeles, California, MKK: Maasai in Kinyawa, Kenya, TSI: Tuscans in Italy, YRI: Yoruban in Ibadan, Nigeria.

**Step 6: Search for IL7R**



**Step 7**  
Show 10kb

**Step 8**  
Click on a SNP for genotype data

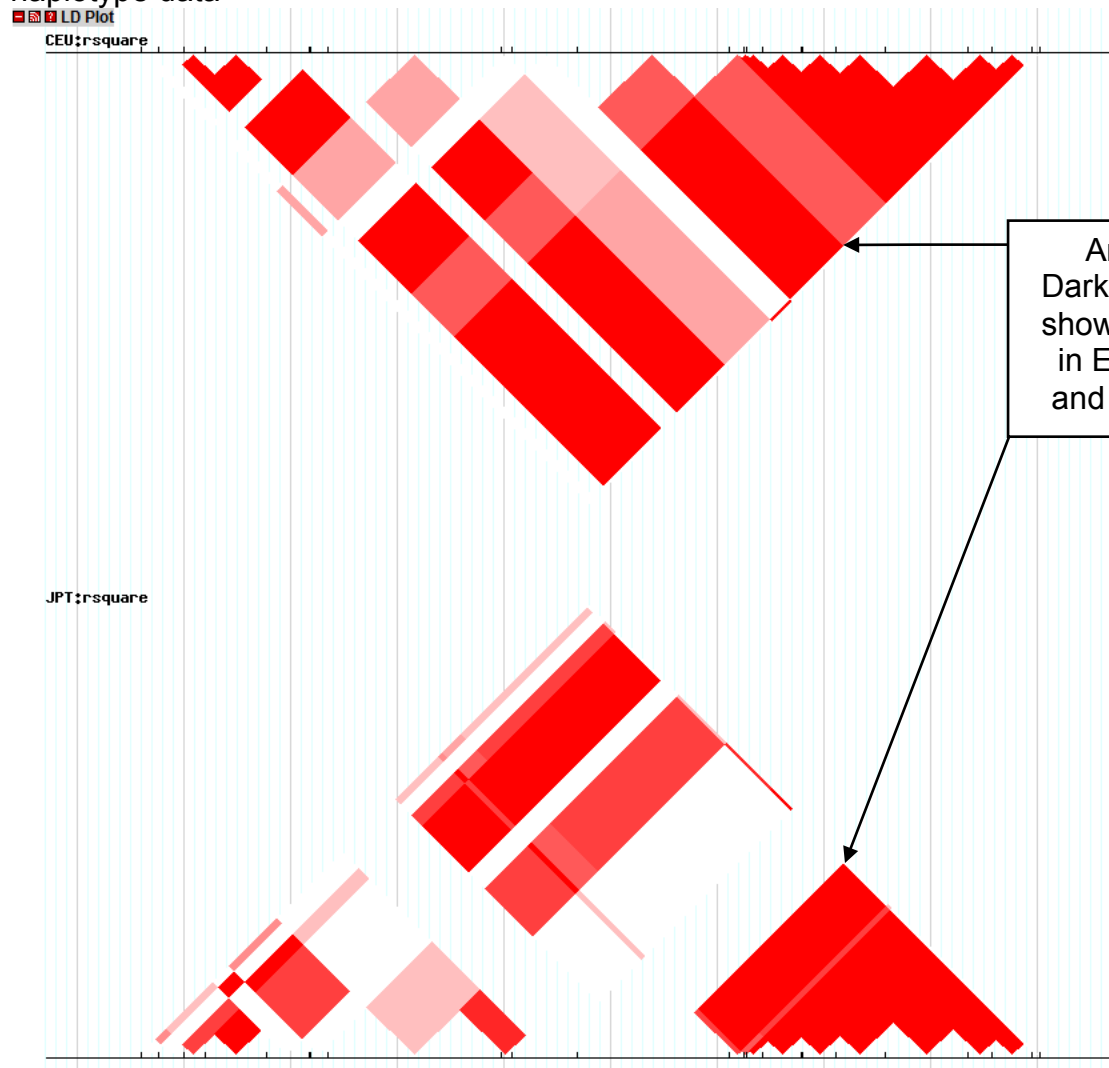
**Answer 1**  
You can select "Download SNP genotype data" to get a text file.

chr5:35910332..35910332, (+) strand relative to the human reference sequence

Population	Genotype frequencies								Allele frequencies				Total	retrieve genotypes				
	genotype	freq	count	genotype	freq	count	genotype	freq	count	Total	Ref-allele	allele			freq	count	Total	
ASW (A)	C/C	0.830	44	C/T	0.132	7	T/T	0.038	2	53	C	0.896	95	T	0.104	11	106	retrieve genotypes
CEU (C)	C/C	0.593	67	C/T	0.327	37	T/T	0.080	9	113	C	0.757	171	T	0.243	55	226	retrieve genotypes
CHB (H)	C/C	0.726	61	C/T	0.202	17	T/T	0.071	6	84	C	0.827	139	T	0.173	29	168	retrieve genotypes
CHD (D)	C/C	0.624	53	C/T	0.341	29	T/T	0.035	3	85	C	0.794	135	T	0.206	35	170	retrieve genotypes
GIH (G)	C/C	0.739	65	C/T	0.261	23	T/T	0	0	88	C	0.869	153	T	0.131	23	176	retrieve genotypes
JPT (J)	C/C	0.663	57	C/T	0.314	27	T/T	0.023	2	86	C	0.820	141	T	0.180	31	172	retrieve genotypes
LWK (L)	C/C	0.899	80	C/T	0.101	9	T/T	0	0	89	C	0.949	169	T	0.051	9	178	retrieve genotypes
MEX (M)	C/C	0.600	30	C/T	0.340	17	T/T	0.060	3	50	C	0.770	77	T	0.230	23	100	retrieve genotypes
MKK (K)	C/C	0.797	114	C/T	0.189	27	T/T	0.014	2	143	C	0.892	255	T	0.108	31	286	retrieve genotypes
TSI (T)	C/C	0.034	3	C/T	0.778	737	T/T	0.188	17	88	C	0.778	737	T	0.222	39	176	retrieve genotypes
YRI (Y)	C/C	0	0	C/T	0	0	T/T	0	0	113	C	0.951	215	T	0.049	11	226	retrieve genotypes

Note: the 're' Allele frequency in pop OR select "Download SNP genotype frequency data"

Step 9 - Go back one page to the chromosome view page and scroll down for haplotype data



Step 10  
Go back to the top of the page and select 'Download tag SNP Data' and click 'configure'

Help links: [Home](#) [About](#) [FAQ](#) [Contact](#)

Landmark or Region: IL7R

Data Source: HapMap Data PhaseII/Rel#2, Feb09, on NCBI B36 assembly, dbSNP b126

Population descriptors: **ASW**: African ancestry in Southwest USA. **CEU**: Utah residents with Han Chinese in Beijing, China. **CHD**: Chinese in Metropolitan Denver, Colorado. **GIH**: Gujarati Webuye, Kenya. **MEX**: Mexican ancestry in Los Angeles, California. **MKK**: Maasai in Kinyawa

Overview

chr5  
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M

Reports & Analysis:

- Annotate LD Plot
- Annotate LD HeatPlot
- Annotate LD Plot
- Annotate Phased Haplotype Display
- Annotate tag SNP Picker
- Download Decorated FASTA File
- Download HapMap GFF File
- Download HapMap LD Data
- Download Impute Data
- Download Phased Haplotype Data
- Download SNP Allele Frequency Data
- Download SNP Genotype Frequency Data
- Download SNP genotype data
- Download tag SNP Data**
- Highlight SNP Properties

Configure... Go

from the CEPH collection. **CH** in Tokyo, Japan, **LWK**: Luhya in Ibadan, Nigeria.

Step 11  
Choose your options  
and click 'Go'

**Configure... tag SNP Data**

Cancel Configure Go

Population: CEU

Pairwise Methods: Tagger Pairwise\* [?]

RSquare cut off: 0.7 [?]

MAF cut off: 0.05 [?]

Include SNPs: [ ] Browse... [?]

Exclude SNPs: [ ] Browse... [?]

Design scores: [ ] Browse... [?]

Output format:  text  Save to Disk

```
#Thu Dec 11 06:23:48 2014: HapMap tag SNPs:7 tag SNPs picked out for population CEU chr5:35897713..35907712 using the algorithm-Tagger-pairwiseTagging
#tag SNPs      Chromosome  Pos      maf
rs1389830      chr5        35899776  0.258
rs7737000      chr5        35907030  0.152
rs7717955      chr5        35898598  0.243
rs10461959     chr5        35904109  0.183
rs6451229     chr5        35901975  0.450
rs11567714     chr5        35898742  0.075
rs11567751     chr5        35907441  0.347
```

```
#captured 21 of 21 alleles at r^2 >= 0.7
#captured 100 percent of alleles with mean r^2 of 1.0
#using 7 tag SNPs in 7 tests.
Allele Best Test      r^2 w/test
rs7717955      rs7717955      1.0
rs11567714     rs11567714     1.0
rs6451226     rs1389830     1.0
rs6893142     rs1389830     1.0
rs1389830     rs1389830     1.0
rs10044838    rs1389830     1.0
rs6451229     rs6451229     1.0
rs6891095     rs7737000     1.0
rs7711202     rs1389830     1.0
rs10461959    rs10461959    1.0
rs11567737    rs1389830     1.0
rs10074127    rs1389830     1.0
rs10074095    rs1389830     1.0
rs1494556     rs1389830     1.0
rs3777090     rs1389830     1.0
rs10941267    rs1389830     1.0
rs9292616     rs1389830     1.0
rs10063445    rs1389830     1.0
rs1494555     rs1389830     1.0
rs7737000     rs7737000     1.0
rs11567751    rs11567751    1.0

Test Alleles Captured
rs1389830      rs10074127,rs3777090,rs11567737,rs1389830,rs10074095,rs6893142,rs6451226,rs7711202,rs10063445,rs10941267,rs1494555,rs1494556,rs10044838,rs9292616
rs7737000      rs6891095,rs7737000
rs7717955      rs7717955
rs10461959    rs10461959
rs6451229     rs6451229
rs11567714    rs11567714
rs11567751    rs11567751
```

Answer 4  
SNPs identified as tag SNPs using  
the above criteria in selected  
genome window.

## Haploview – haplotyping software

Haploview is a user friendly piece of freeware that has been designed to generate haplotypes directly from HapMap or from your own data. It can be downloaded at <http://www.broadinstitute.org/haploview/haploview-downloads> Comprehensive documentation is also available at this web page.

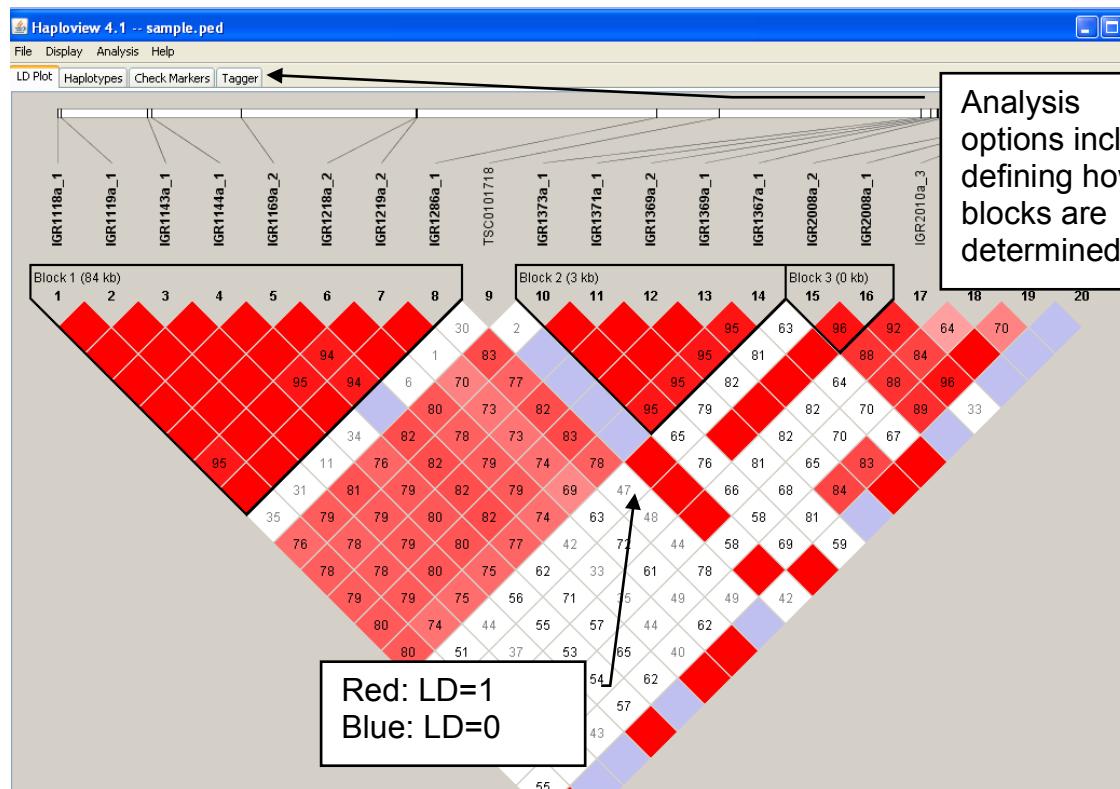
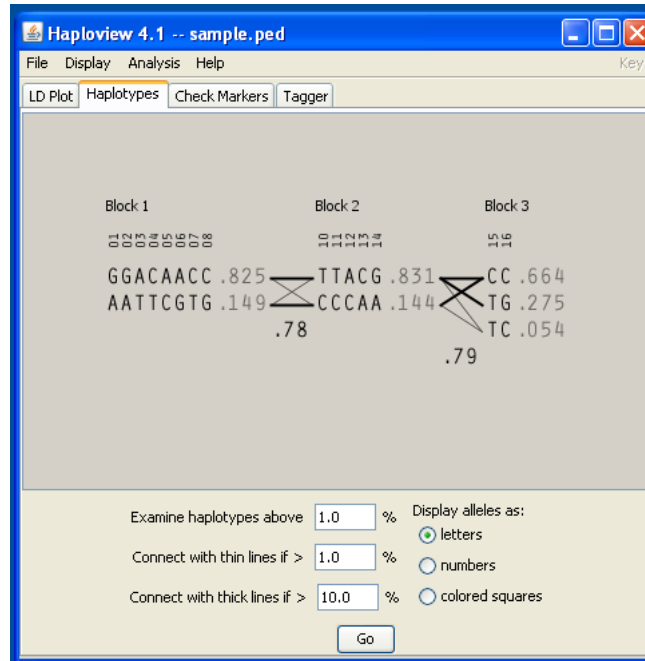
### Worked Example:

Import your data files into Haploview.

The sample files contain data on 40 trios (father, mother and child), so simply click ok.

#	Name	Position	ObsHET	PredH...	HW/pval	%Geno	FamTrio	MendErr	MAF	Rating
1	IGR1118a_1	274044	0.282	0.269	0.762	97.5	39	0	0.16	✓
2	IGR1119a_1	274541	0.267	0.257	0.938	96.7	37	0	0.151	✓
3	IGR1143a_1	286593	0.3	0.289	0.516	100.0	40	0	0.175	✓
4	IGR1144a_1	287261	0.283	0.272	0.696	100.0	40	0	0.162	✓
5	IGR1169a_2	299755	0.268	0.241	0.392	93.3	33	0	0.14	✓
6	IGR1218a_2	324341	0.301	0.284	0.63	94.2	33	0	0.171	✓
7	IGR1219a_2	324379	0.275	0.278	0.711	90.8	31	0	0.167	✓
8	IGR1286a_1	359048	0.263	0.263	1.0	95.0	35	0	0.149	✓
9	TSC0101718	366811	0.132	0.124	1.0	95.0	34	0	0.067	✓
10	IGR1373a_1	395079	0.283	0.272	0.176	100.0	40	0	0.162	✓
11	IGR1371a_1	396363	0.277	0.272	0.215	93.3	33	0	0.162	✓
12	IGR1369a_2	397334	0.311	0.297	0.139	88.3	31	0	0.181	✓
13	IGR1369a_1	397381	0.275	0.264	0.216	100.0	40	0	0.156	✓
14	IGR1367a_1	398352	0.283	0.264	0.216	100.0	40	0	0.156	✓
15	IGR2008a_2	411823	0.393	0.441	0.695	93.3	34	0	0.329	✓
16	IGR2008a_1	411873	0.294	0.403	0.04	85.0	29	0	0.28	✓
17	IGR2010a_3	412456	0.336	0.403	0.143	96.7	38	0	0.279	✓
18	IGR2011b_1	413233	0.489	0.499	0.84	75.0	27	0	0.483	✓
19	IGR2016a_1	415579	0.351	0.422	0.151	95.0	37	0	0.303	✓

There are three haplotype blocks, defined by confidence intervals (Gabriel et al, Science, 2002), which can be genotyped using four tag SNPs.



LD plot shows how the haplotype blocks are composed.

This data set is available online as part of the Haploview tutorial, along with descriptions of file formats and how Haploview analyses data.

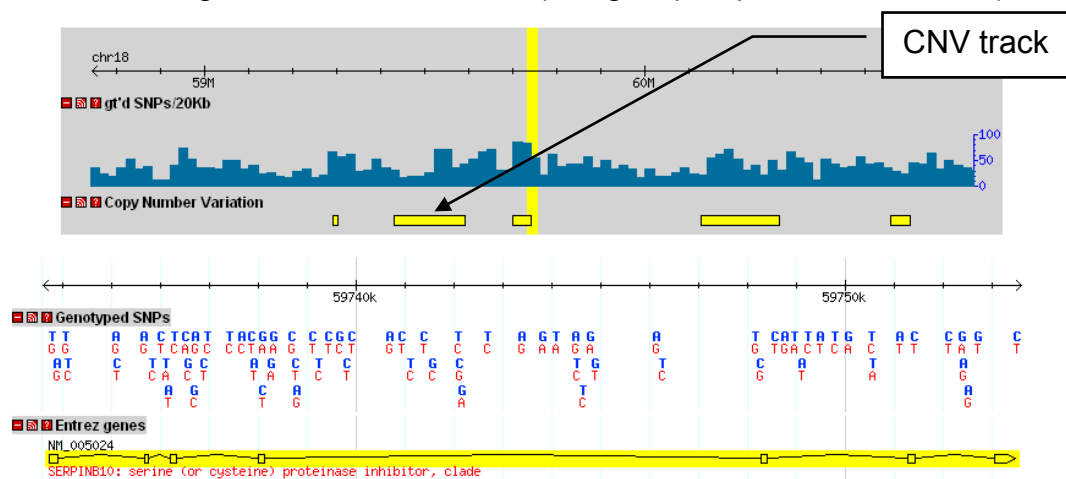
## Copy Number Variation

Another frequent source of polymorphisms is Copy Number Variation (CNVs), which are a type of “structural variation” in the genome. These include anything from small insertions and deletions ( $\geq 1\text{kb}$ ) and duplications to large scale duplications ( $\geq 50\text{kb}$ ). Such CNVs may be on different chromosomes through duplications followed by translocation events, or are segmental duplications arisen through non-allelic homologous recombination. Many of these CNVs have been found through analyses of the HapMap data, and more are being found with the 1000 genome project. For more see Conrad et al. Nature 2009 and the CNV discovery project at WTSI

(<http://www.sanger.ac.uk/humgen/cnv/42mio/> )

While our knowledge of CNVs is far from as complete as with SNPs, they are starting to be implicated in human diseases, often due to their effects via gene dosage. So it's definitely worth checking if any are known in your area of interest.

The HapMap site (<http://hapmap.ncbi.nlm.nih.gov>) we looked at earlier has a track for CNVs e.g. here for SERPINB10 (using HapMap data Release 24).



The Database of Genomic Variance (<http://dgv.tcag.ca/dgv/app/home>) is a more specialised site for finding CNVs.

**Search for SERPINB10**

**Database of Genomic Variants**  
A curated catalogue of human genomic structural variation

About the Project Downloads Links Statistics FAQ  
Genome Browser Query Tool Submissions Contact Us Training Resources

**Keyword, Landmark or Region Search:**  Search NCBI36/hg18

**Examples:** RP11-34P13; CFTR, 7q11.21; chr7:71890181-72690180

**Find DGV Variants**

[by Study](#) [by Sample](#)  
[by Method](#) [by Variant](#)  
[by Platform](#) [by Chromosome](#)

**Summary Statistics**

Stat	Merged-level	Sample-level
CNVs:	109863	2304349
Inversions:	238	3380
<b>Number of Studies:</b>	55	

[News: July 2013 Update and Newsletter has been issued](#)

Opening the results in the genome browser, shows that Database of Genomic Variance has three CNV encompassing at least part of SERPINB10 (Red = gain, Blue = loss). Clicking on the CNV (here the red) gives more data:

Variation: Variation\_67377  
 Landmark: chr18:59,533,336..60,096,344 (Genome Browsers: UCSC , Ensembl)  
 Genomic Position: chr18:59,533,336..60,096,344  
 Variation Type: CopyNumber  
 Cytogenetic Band: 18q21.33-18q22.1  
 Starting position along chromosome (in Mb): 59.5  
 Gap within 100k: No  
 Known Genes: SERPINB2, SERPINB8, C18orf20, SERPINB10, SERPINB7, LOC400654, LOC284294, HMSD, SERPINB11  
 Method: Agilent custom oligo CGH array  
 Individual: NA11995, NA19114, NA19115  
 Comment: CNclass:CNcount-2.442-3.3|CNVRID=CNVR7357\_full  
 Reference: Conrad et al. (2009)  
 Pub Med ID: 19812545  
 Frequency Information:  
 Subject Cohort: Control  
 Sample Size: 450 HapMap Individuals  
 Observed Gain: 3  
 Observed Loss: 0  
 Total Gain/Loss: 3  
 Minor Allele: gain  
 Allele Frequency: 0.013483146  
 Related Locus: chr18:59429883-60096344

**Platform used, and observed changes - here gains.**

It may also be useful to check CNVD <http://202.97.205.78/CNVD/> which text mines CNVs from publications in addition to use large studies.

**Exercises:****1. Analysis of sequence variation at the RUNX1 locus**

Using the Ensembl database determine the number of coding SNPs within the longest transcript of transcription factor RUNX1. How many stop gain, frame shift coding and non-synonymous SNPs are there? What are their ambiguity codes and do they encode amino acid substitutions. How many non-synonymous SNPs have Validation information and which single SNP would you type first?

**2. Sequence variation in SERPINB10**

Microarray data suggests that differential SERPINB10 expression is involved in prostate cancer. You have a medium sized, European, case-control population with 500 cases and 500 controls. You can only afford to genotype a couple of SNPs to see if any are associated with prostate cancer risk in your cohort. Which SNPs do you pick and why?

**3. What are the significant SNPs and MYH11 haplotype that predisposes to disease?**

Your microarray results showed that MYH11 shows differential expression in diseased compared to non-diseased aorta. You decide to genotype 28 SNPs in all of your case and control coronary artery disease samples to identify genetic association of SNPs within the gene. Using Haploview [two files "Genotype\_data.ped" and SNP\_Locations.info] determine how many haplotypes are there? What is the name of the SNP significantly out of HWE? Are any of the SNPs or haplotypes significant or trending in cases than controls (or vice versa)? What are the tag SNPs and how many of them are there?

*Hint: Check "Do association test" and "Case/Control data" from the first window*

*Hint: Find statistical significance under the 'Association' tab, then look at 'Single Marker' and 'Haplotypes' for significant SNPs'*

*Hint: Tag SNP data can be found under the 'Tagger' tab, alter the  $r^2$  to 0.7 then "Run Tagger" at the bottom of the page*

**4. Using dbGaP for identifying heart disease loci.**

You have a heterogeneous Type-1 Diabetes cohort, and you decide to subset your population to help define the multiple genetic components of the disorder.



You segregate a T1D sub-population with myocardial infarction within your cohort. A small, seemingly underpowered screen that you carry out shows some significance on chromosome 16p13.13. Are there any studies within dbGaP that might help you decide if your finding is real or novel?

*Hint: Type 1 Diabetes Genetics Consortium (T1DGC): Genome-Wide Association Study in Type 1 Diabetes, 2008 (pha002862.1)*

*Hint: Try changing the filter of the GWAS to  $<10e-6$  for visualisation*

*Hint: 16p is the petit (shorter) arm.*

## Answers

### Task 1: Variation in the RUNX1 gene

Search Human Ensembl for RUNX1 to determine the gene ID and link to the gene view. You can then link to SNP information in Variation Table for the gene. **Currently**, a total of 153 Stops, 789 Frame shift and 995 non-synonymous (missense). 8 have validation data (multiple observations), possibly rs74315451 because of PolyPhen2 and SIFT predictions, but there are lots of other potentially pathogenic variants!

### Task 2: Sequence variation in SERPINB10

There are no 'right' answers to this question!

Searching PubMed with the terms 'SERPINB10' and 'polymorphism' identifies a paper by Shioji et al (J Hum Genet (2005) 50: 507-515). Two cSNPs, rs8097425 and rs963075, are shown to have significant associations with prostate cancer in a Japanese cohort. These SNPs and three SNPs in SERPINB2 form a haplotype block which can be defined by genotyping just two SNPs (one in SERPINB2 and one in SERPINB10).

### Task 3: Does a MYH11 haplotype predispose to disease?

There are 7 haplotype blocks as first defined in Haploview

The SNP significantly out of HWE is rs7203040 – do you think that this could be important?

The significant SNPs are rs215571 (trending – almost significant) and rs2306860 (p-val 0.02). There are no significant haplotypes, though 1 haplotype in block 5 and block 7 are trending.

#	Name	Assoc Allele	Case, Control Ratios	Chi Square	p value
1	RS1050163	T	0.488, 0.484	0.011	0.916
2	RS1050162	C	0.519, 0.518	0.0020	0.9609
3	RS2075511	G	0.515, 0.500	0.28	0.5965
4	RS11130	A	0.522, 0.500	0.544	0.4605
5	RS12907	A	0.021, 0.020	0.02	0.8881
6	RS16967494	T	0.279, 0.258	0.7	0.4028
7	RS1050113	A	0.337, 0.310	1.005	0.3161
8	RS2272554	G	0.425, 0.393	1.331	0.2486
9	RS4781689	A	0.095, 0.092	0.04	0.8417
10	RS6498574	G	0.368, 0.349	0.474	0.491
11	RS8044595	G	0.334, 0.306	1.097	0.2948
12	RS1050111	A	0.136, 0.115	1.233	0.2669
13	RS7184472	C	0.817, 0.815	0.0050	0.9415
14	RS215590	A	0.416, 0.379	1.646	0.1995
15	RS215581	C	0.786, 0.768	0.529	0.4668
16	RS215579	T	0.700, 0.669	1.335	0.248
17	RS215573	C	0.696, 0.671	0.805	0.3695
18	RS215571	T	0.632, 0.581	3.361	0.0668
19	RS215570	T	0.633, 0.587	2.563	0.1094
20	RS9935015	G	0.795, 0.774	0.822	0.3646
21	RS3851706	G	0.737, 0.701	1.835	0.1755
22	RS2306860	G	0.592, 0.527	5.214	0.0224
23	RS8057023	G	0.943, 0.926	1.546	0.2138
24	RS12446688	G	0.760, 0.753	0.07	0.7912
25	RS3826056	G	0.941, 0.927	1.005	0.3161
27	RS12597051	T	0.944, 0.928	1.368	0.2422
28	RS3213476	A	0.841, 0.821	0.916	0.3384

What happens to the results when you change the LD blocks?

Haplotype	Freq.	Case, Control Ratios	Chi Square	p value
<b>Haplotype Associations</b>				
Block 1				
CC	0.515	0.515, 0.513	0.0070	0.9337
TT	0.485	0.485, 0.487	0.0070	0.9337
Block 2				
GA	0.506	0.513, 0.493	0.471	0.4924
TG	0.476	0.474, 0.480	0.037	0.8479
TA	0.014	0.010, 0.020	2.122	0.1452
Block 3				
CGAG	0.492	0.481, 0.516	1.556	0.2123
TAGG	0.267	0.274, 0.254	0.666	0.4143
CGAA	0.091	0.093, 0.088	0.06	0.7776
CGGC	0.081	0.084, 0.075	0.323	0.5698
CAGG	0.060	0.061, 0.056	0.126	0.7224
Block 4				
AG	0.676	0.666, 0.694	1.097	0.2948
GG	0.195	0.197, 0.192	0.046	0.8306
GA	0.129	0.137, 0.114	1.464	0.2263
Block 5				
ACTTT	0.392	0.408, 0.359	3.067	0.0799
CCTTT	0.204	0.205, 0.203	0.014	0.9049
CTCAC	0.202	0.195, 0.216	0.854	0.3556
CCACC	0.098	0.094, 0.104	0.351	0.5537
CCTCCC	0.060	0.053, 0.072	1.956	0.1619
CTTCT	0.015	0.017, 0.010	1.011	0.3145
Block 6				
GG	0.725	0.737, 0.701	2.017	0.1555
AA	0.212	0.205, 0.225	0.792	0.3735
GA	0.062	0.036, 0.073	1.632	0.2015
Block 7				
GGGTA	0.544	0.560, 0.513	2.716	0.0994
CGACTA	0.218	0.211, 0.233	0.829	0.3627
CGGTG	0.099	0.096, 0.106	0.287	0.5923
CAGCG	0.062	0.057, 0.072	1.274	0.259
CGGTA	0.051	0.045, 0.062	1.756	0.1851
GGAGTA	0.023	0.028, 0.014	2.67	0.1022

#	Name	Position	Design Score	Force Include	Force Exclude	Capture the Allele?
1	RS1050163	15718524	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
2	RS1050162	15718563	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
3	RS2075511	15725442	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
4	RS11130	15725611	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	RS12907	15726154	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
6	RS16967494	15728364	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
7	RS1050113	15746335	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
8	RS2272554	15757705	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
9	RS4781689	15772913	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
10	RS6498574	15797566	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
11	RS8044595	15813631	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
12	RS1050111	15824498	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
13	RS7184472	15826003	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
14	RS215590	15826982	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
15	RS215581	15840675	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
16	RS215579	15843113	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
17	RS215573	15844885	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
18	RS215571	15851834	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
19	RS215570	15852530	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
20	RS9935015	15864402	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
21	RS3851706	15868148	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
22	RS2306860	15868240	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
23	RS8057023	15861611	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
24	RS12446688	15862631	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
25	RS3826056	15867043	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
27	RS12597051	15861106	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
28	RS3213476	15869759	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

There are 15 tag SNPs that capture 27 SNPs typed in the region – notice that the unchecked HWE SNP is not included. Check the “Results” Tab to see the tag SNPs

### Task 4: T1D, MI endophenotype in dbGaP

C16, 11.15Mb, CLEC16A – check OMIM!