Module 3: Exploring Function and Disease

Aims

- To look at the information available to determine the possible function of a gene product
- To highlight various inter-linked information resources that are available for this purpose
- Worked and task examples to help illustrate these resources

By the end of this module you should be able to go from obtaining a gene structure, via various routes, to finding out about:

- published information on the gene
- known mendelian inherited disorder(s) associated with the gene
- summary of predicted function from several linked databases
- domains found within the protein
- other predicted proteins also containing any domains found
- Viewing structural information if available.

This module will concentrate on looking at human data as this is this is what the disease databases are primarily concerned with.

Introduction

Once we have located a gene and obtained its sequence and structure, how can we go about finding out more about the possible function of its protein product(s)?

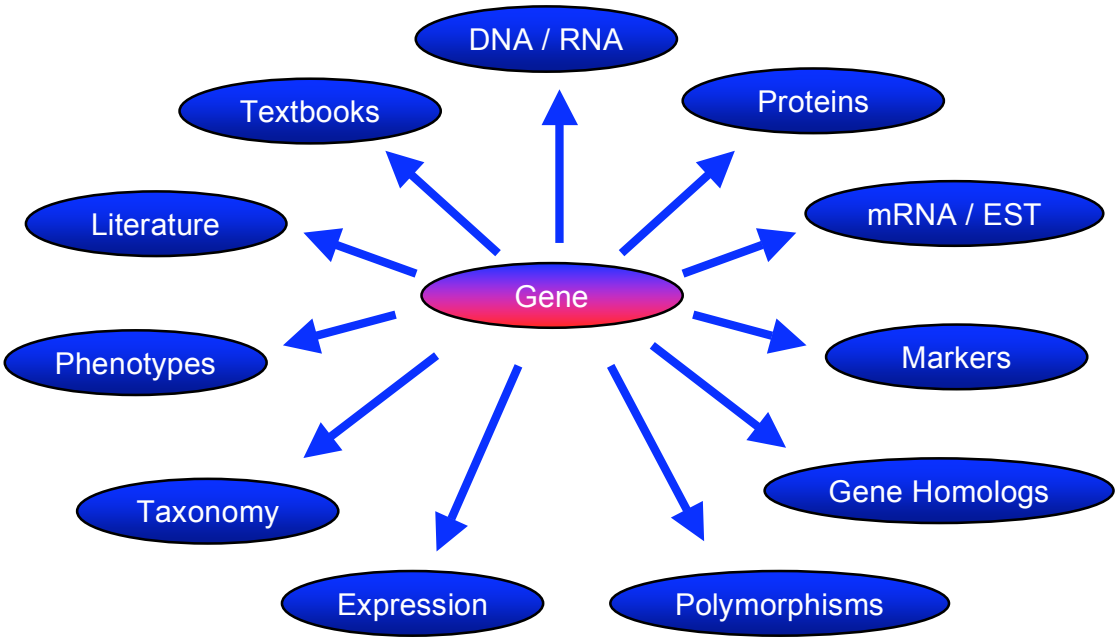
This module will take us through various inter-linked information resources that are now available, enabling the user to find out more about a given gene product. Protein “function” is of course an open-ended issue; there are many different levels, ranging from the biochemical functions such as kinase activity, to physiological function such as a role in an immunological signalling cascade. Information from various sources needs to be collated to piece together a picture of a protein’s potential function.

By its nature, this is largely restricted to information that is already “known”, and is dependent on regular updates of databases. In addition, *ab-initio* analysis of novel sequences can provide clues as to the function of a protein, through homologies to proteins for which some functional information is available and from discovery of conserved domains within the sequence. This type of analysis will increase further in effectiveness as further genome sequencing ties in with mutational studies and protein structure determination. Later modules will involve further investigation of novel genes to find homologues/orthologues.

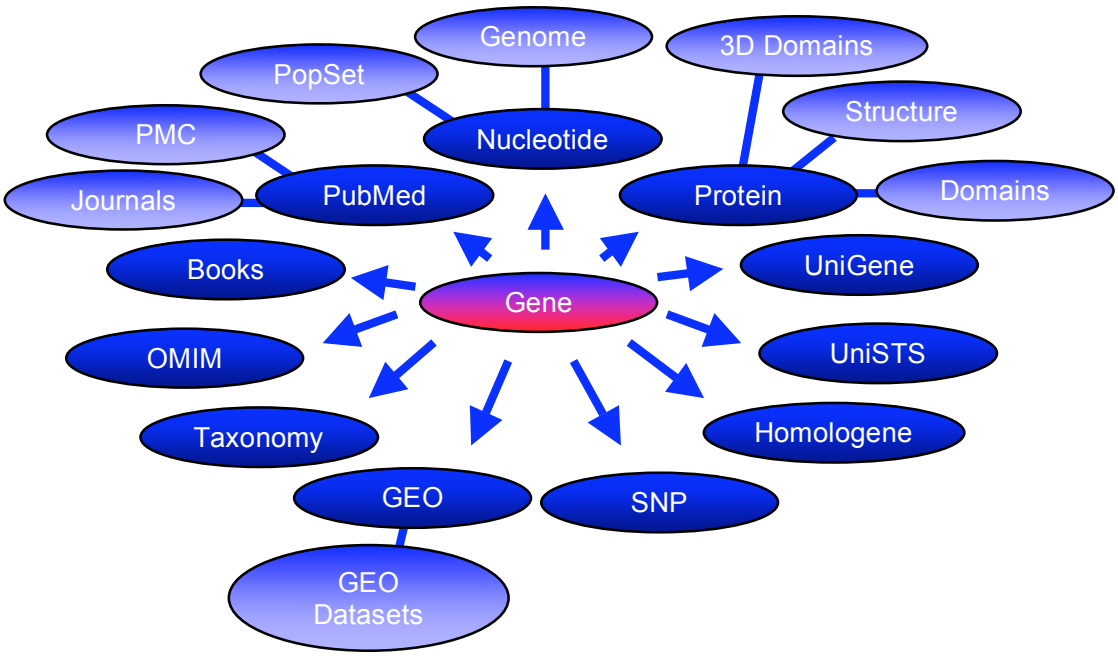
The NCBI Web Server

The National Center for Biotechnology Information (NCBI) is one of the world's premier web sites for biomedical and bioinformatical research. Based within the National Library of Medicine at the National Institutes of Health, USA, the NCBI hosts many databases used by biomedical and research professionals. The services include PubMed, the bibliographic database; GenBank, the nucleotide sequence database; and the BLAST algorithm for sequence comparison, among many others.

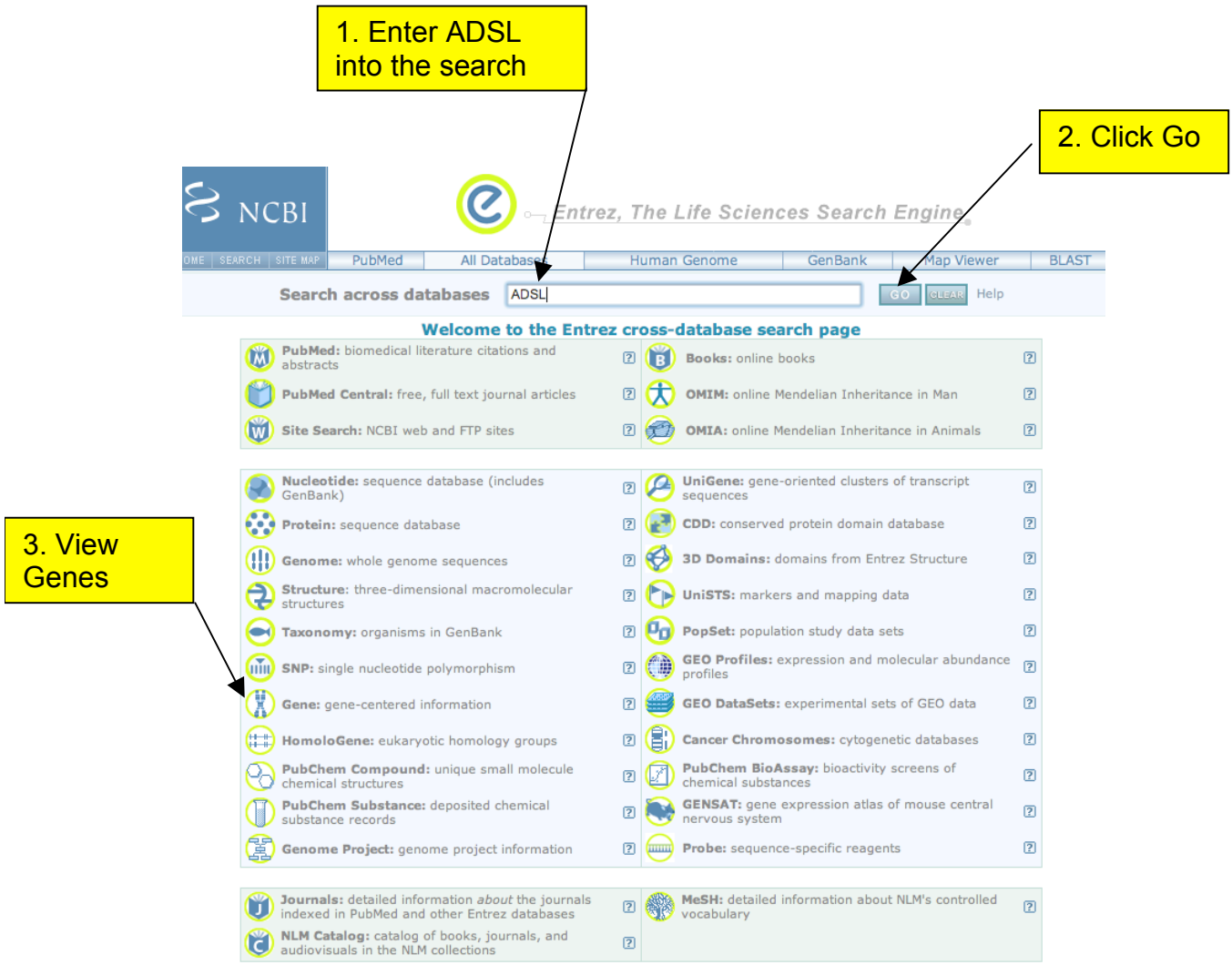
Information Associated With a Gene Locus

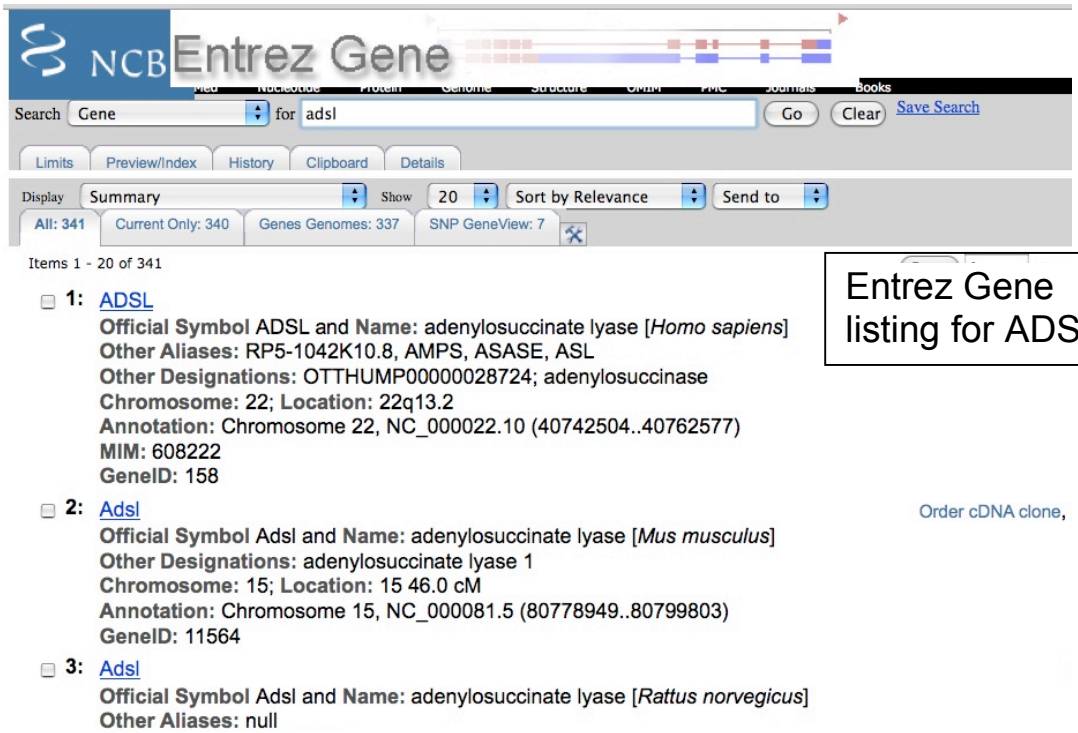


Linking to further information

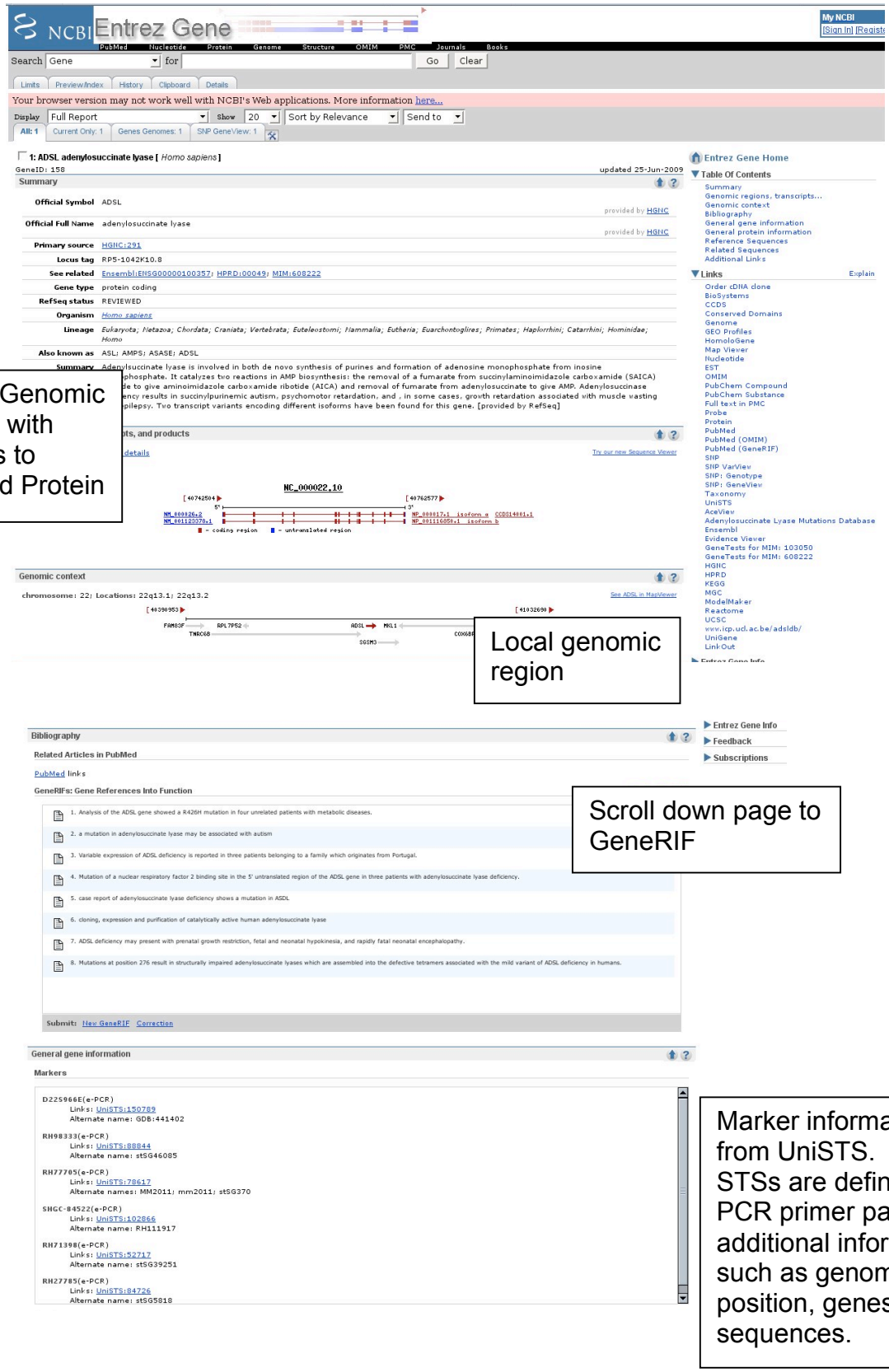


Worked Example: Use the Entrez system to explore function and disease information for human adenylosuccinate lyase gene (ADSL).





Entrez Gene listing for ADSL

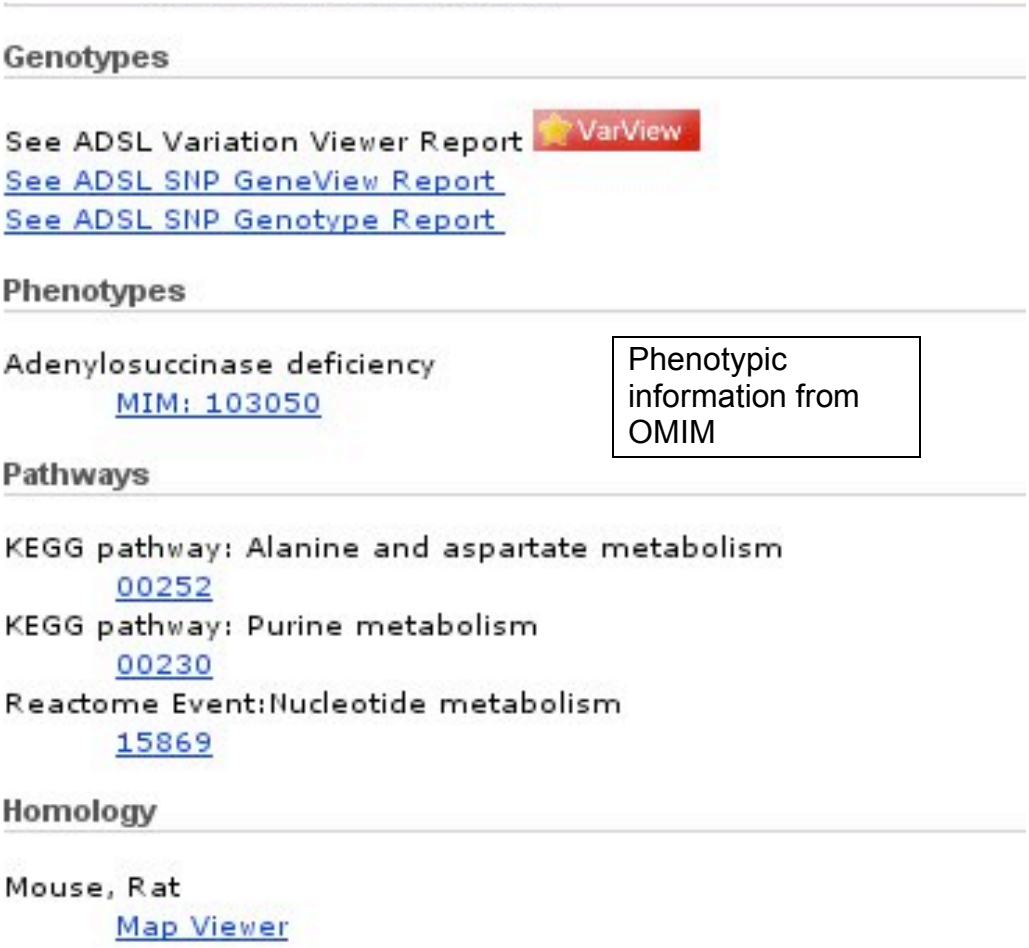


mRNA to Genomic alignment with quick links to mRNA and Protein

Local genomic region

Scroll down page to GeneRIF

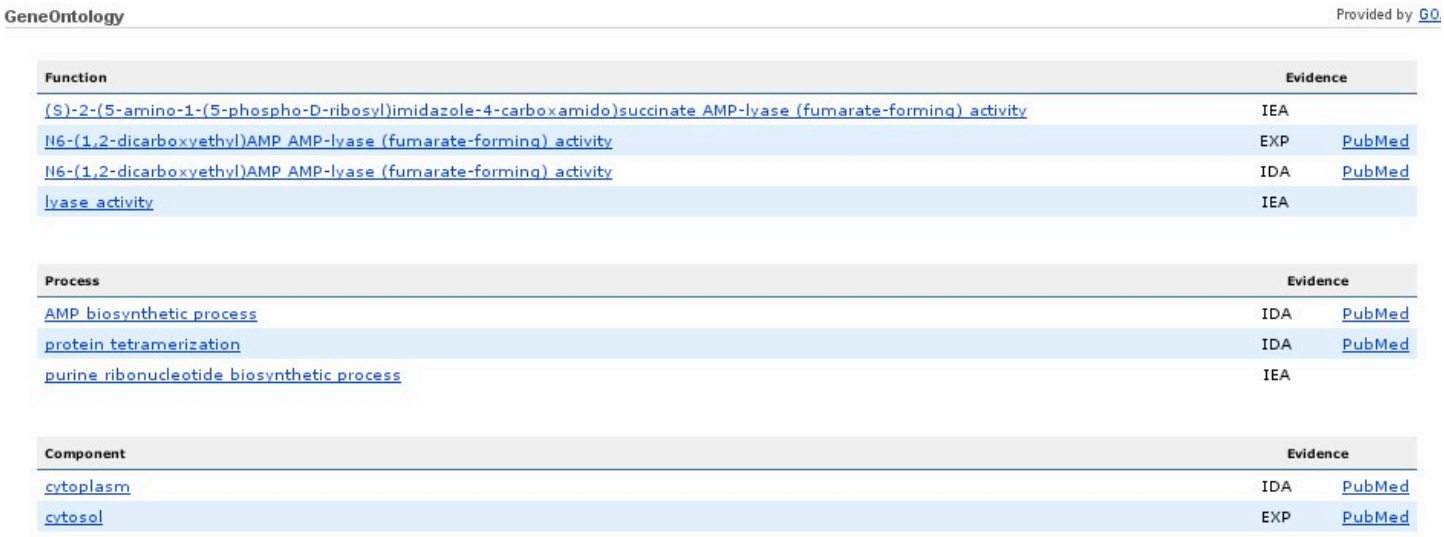
Marker information from UniSTS.
STSs are defined by PCR primer pairs plus additional information, such as genomic position, genes and sequences.



Links to KEGG: aim for complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information.
Links to Reactome: The Reactome project is a collaboration among Cold Spring Harbor Laboratory, The European Bioinformatics Institute, and The Gene Ontology Consortium to develop a curated resource of core pathways and reactions in human biology.

Gene Ontology links:

- A collaborative effort to address the need for consistent descriptions of gene products in different databases.
- Three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner
- Ontologies are 'specifications of a relational vocabulary'
- Terms in a particular vocabulary are restricted to a particular field. GO terms are all biological.



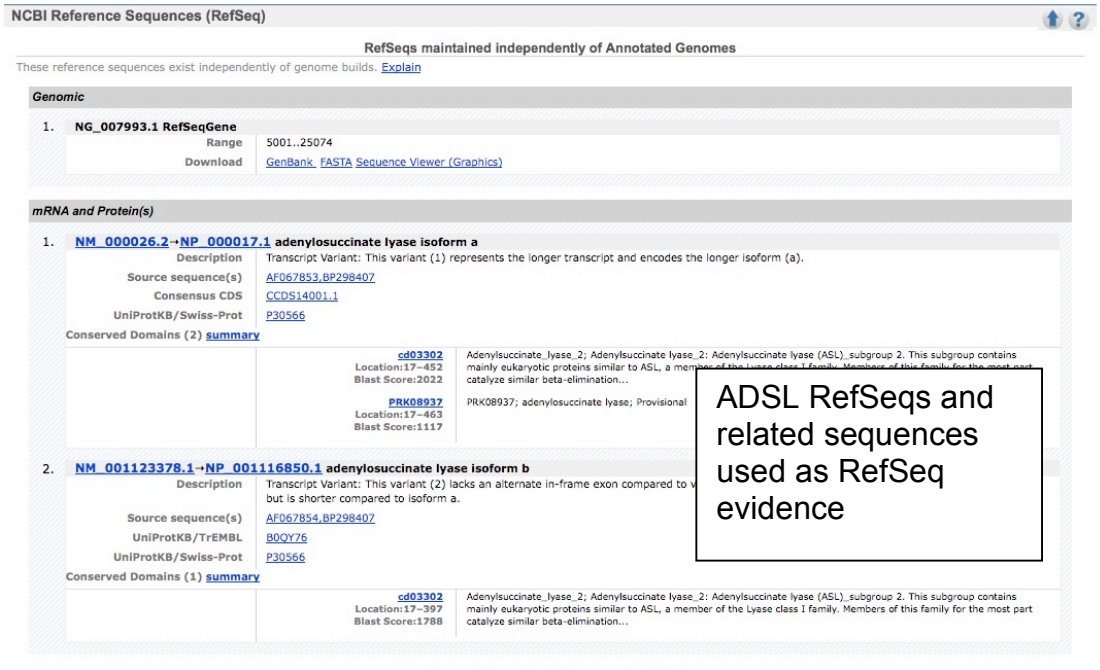
GeneOntology

Provided by GO

Function	Evidence	
(S)-2-(5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamido)succinate AMP-lyase (fumarate-forming) activity	IEA	
N6-(1,2-dicarboxyethyl)AMP AMP-lyase (fumarate-forming) activity	EXP	PubMed
N6-(1,2-dicarboxyethyl)AMP AMP-lyase (fumarate-forming) activity	IDA	PubMed
lyase activity	IEA	

Process	Evidence	
AMP biosynthetic process	IDA	PubMed
protein tetramerization	IDA	PubMed
purine ribonucleotide biosynthetic process	IEA	

Component	Evidence	
cytoplasm	IDA	PubMed
cytosol	EXP	PubMed



NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

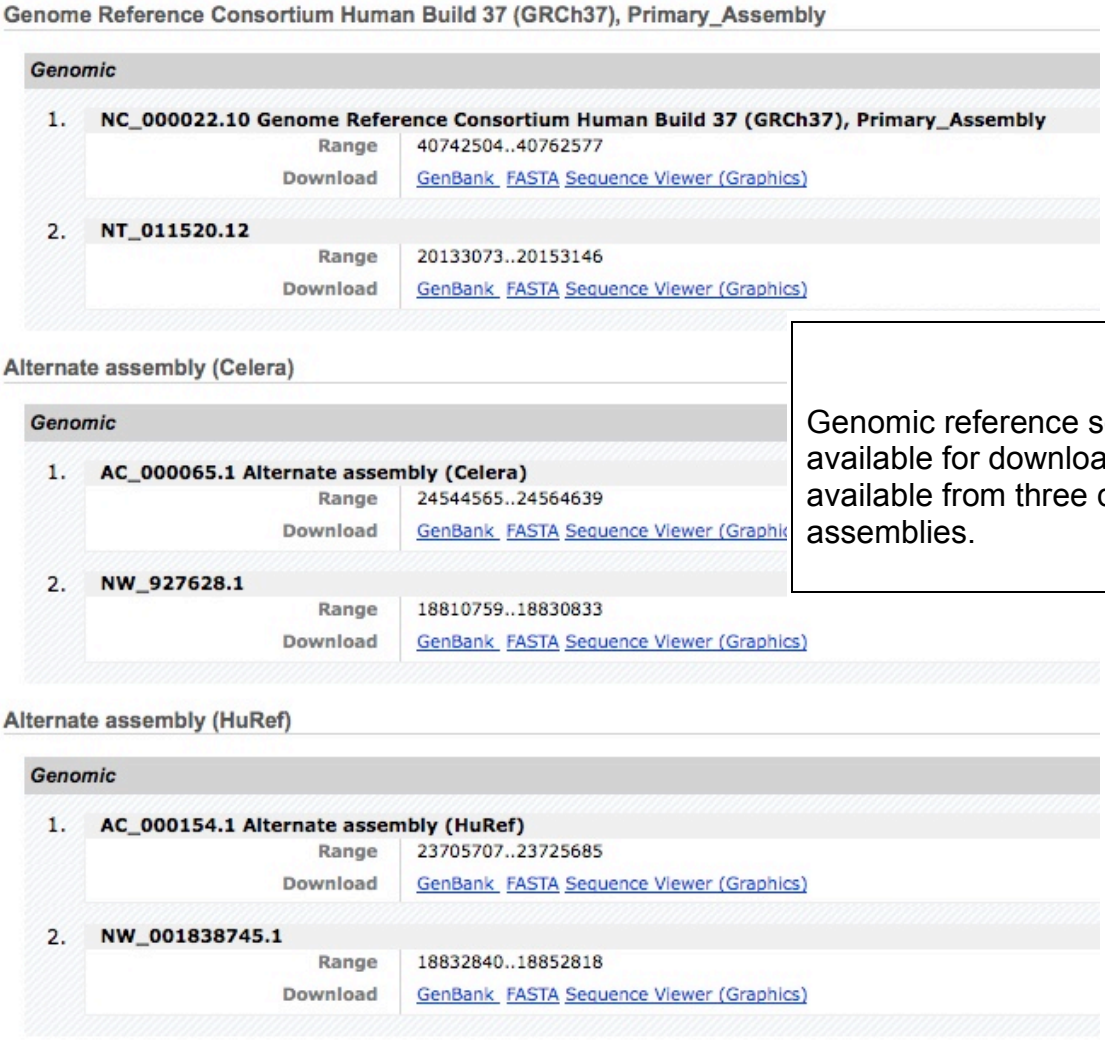
Genomic

1. NG_007993.1 RefSeqGene
   - Range: 5001..25074
   - Download: GenBank, FASTA, Sequence Viewer (Graphics)

mRNA and Protein(s)

1. NM_000026.2→NP_000017.1 adenylosuccinate lyase isoform a
   - Description: Transcript Variant: This variant (1) represents the longer transcript and encodes the longer isoform (a).
   - Source sequence(s): AF067853,BP298407
   - Consensus CDS: CCDS14001.1
   - UniProtKB/Swiss-Prot: P30566
   - Conserved Domains (2) summary
     - cd03302, Location:17–452, Blast Score:2022: Adenylsuccinate_lyase_2; Adenylsuccinate lyase_2: Adenylsuccinate lyase (ASL)_subgroup 2. This subgroup contains mainly eukaryotic proteins similar to ASL, a member of the Lyase class I family. Members of this family for the most part catalyze similar beta-elimination...
     - PRK08937, Location:17–463, Blast Score:1117: PRK08937; adenylosuccinate lyase; Provisional
2. NM_001123378.1→NP_001116850.1 adenylosuccinate lyase isoform b
   - Description: Transcript Variant: This variant (2) lacks an alternate in-frame exon compared to v... but is shorter compared to isoform a.
   - Source sequence(s): AF067854,BP298407
   - UniProtKB/TrEMBL: B0QY76
   - UniProtKB/Swiss-Prot: P30566
   - Conserved Domains (1) summary
     - cd03302, Location:17–397, Blast Score:1788: Adenylsuccinate_lyase_2; Adenylsuccinate lyase_2: Adenylsuccinate lyase (ASL)_subgroup 2. This subgroup contains mainly eukaryotic proteins similar to ASL, a member of the Lyase class I family. Members of this family for the most part catalyze similar beta-elimination...

ADSL RefSeqs and related sequences used as RefSeq evidence



Genome Reference Consortium Human Build 37 (GRCh37), Primary_Assembly

Genomic

1. **NC_000022.10 Genome Reference Consortium Human Build 37 (GRCh37), Primary_Assembly**
   - Range: 40742504..40762577
   - Download: GenBank FASTA Sequence Viewer (Graphics)
2. **NT_011520.12**
   - Range: 20133073..20153146
   - Download: GenBank FASTA Sequence Viewer (Graphics)

Alternate assembly (Celera)

Genomic

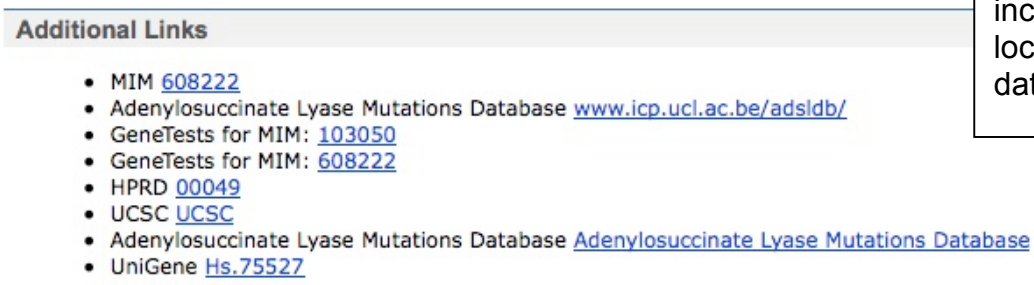
1. **AC_000065.1 Alternate assembly (Celera)**
   - Range: 24544565..24564639
   - Download: GenBank FASTA Sequence Viewer (Graphic
2. **NW_927628.1**
   - Range: 18810759..18830833
   - Download: GenBank FASTA Sequence Viewer (Graphics)

Alternate assembly (HuRef)

Genomic

1. **AC_000154.1 Alternate assembly (HuRef)**
   - Range: 23705707..23725685
   - Download: GenBank FASTA Sequence Viewer (Graphics)
2. **NW_001838745.1**
   - Range: 18832840..18852818
   - Download: GenBank FASTA Sequence Viewer (Graphics)

Genomic reference sequence available for download. Here available from three different assemblies.

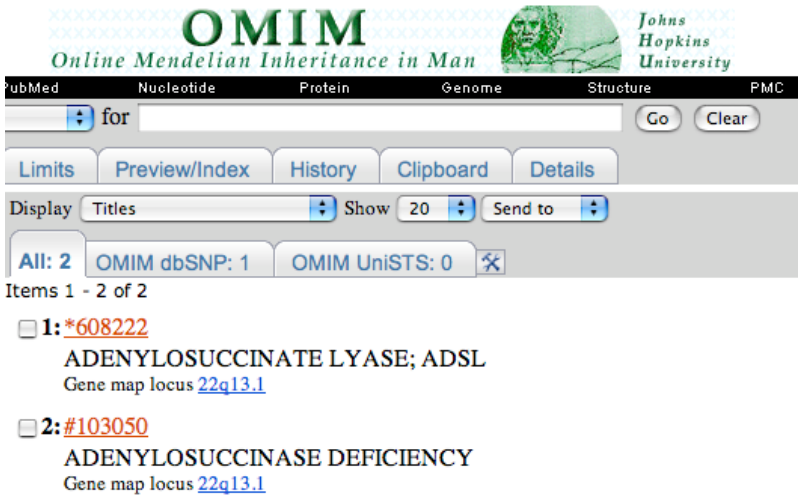


Additional Links

- MIM 608222
- Adenylosuccinate Lyase Mutations Database www.icp.ucl.ac.be/adsldb/
- GeneTests for MIM: 103050
- GeneTests for MIM: 608222
- HPRD 00049
- UCSC UCSC
- Adenylosuccinate Lyase Mutations Database Adenylosuccinate Lyase Mutations Database
- UniGene Hs.75527

Further links, including locus specific databases

4. Look more closely at the disease related information in OMIM.



Clinical literature database curated at Johns Hopkins Univ

Links human genes and genetic disorders to human disease

Lists allelic variants that have clinical consequences

5. Select *608222 for information on ADSL or #103050 for ADENYLOSUCCINASE DEFICIENCY

*608222
ADENYLOSUCCINATE LYASE; ADSL

Alternative titles; symbols

ADENYLOSUCCINASE

Gene map locus 22q13.1

TEXT

DESCRIPTION

Adenylosuccinase (ADSL; EC 4.3.2.2) is an enzyme involved in both the de novo synthesis of purines and the formation of adenosine monophosphate from inosine monophosphate.

CLONING

Using an avian liver ADSL cDNA as a probe to screen a human liver cDNA library, Stone et al. (1992) isolated an ADSL cDNA encoding a 459-amino acid protein with a molecular mass of 52 kD. The enzyme has a homotetrameric structure.

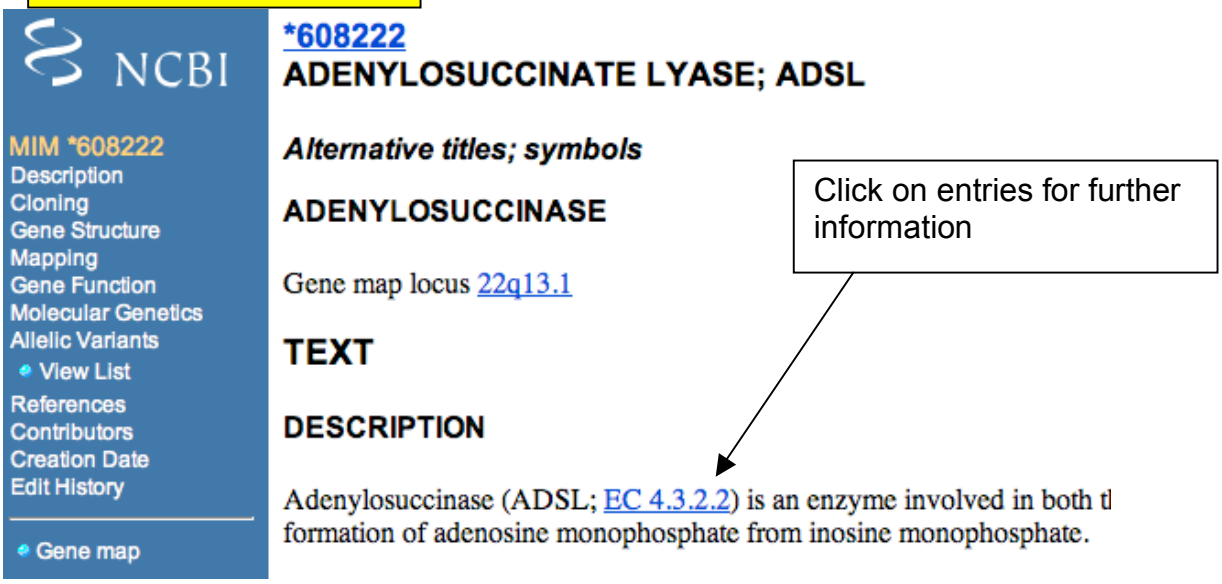
Marie et al. (1999) found that the human ADSL cDNA contains an additional segment at the 5-prime end, encoding a protein of 484 amino acids, rather than 459 as previously reported. Kmoch et al. (2000) reported the complete human ADSL cDNA sequence, which revealed the novel 52-bp sequence at the 5-prime end of the ADSL gene, containing an alternate initiation codon. This longer sequence was termed 'M1,' and the shorter one 'M2.' Expression studies showed that the M1 protein was soluble, active, and stable, in contrast to M2, which was insoluble and inactive. The authors noted that the native human protein is composed of 484 amino acids, the same as murine ADSL. In addition, Kmoch et al. (2000) found 2 ADSL isoforms produced by alternative splicing of exon 12. Both transcripts were expressed in all tissues studied, with the unspliced form being about 10-fold more abundant. The authors hypothesized that the inactive isoform may be able to form tetramers with the active isoform, forming an array of enzymes with different activities depending on the composition of the tetramer.

Wong and O'Brien (1995) cloned the mouse ADSL gene, and found that the human and mouse ADSL proteins share 94% identity.

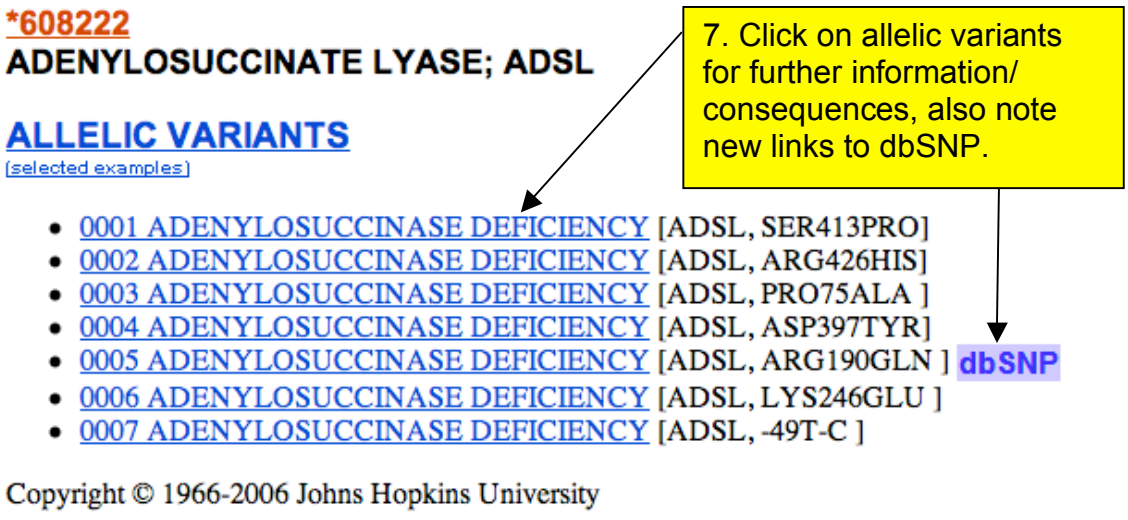
GENE STRUCTURE

Kmoch et al. (2000) determined that the human ADSL gene contains 13 exons.

6. View list of Allelic Variants from left menu

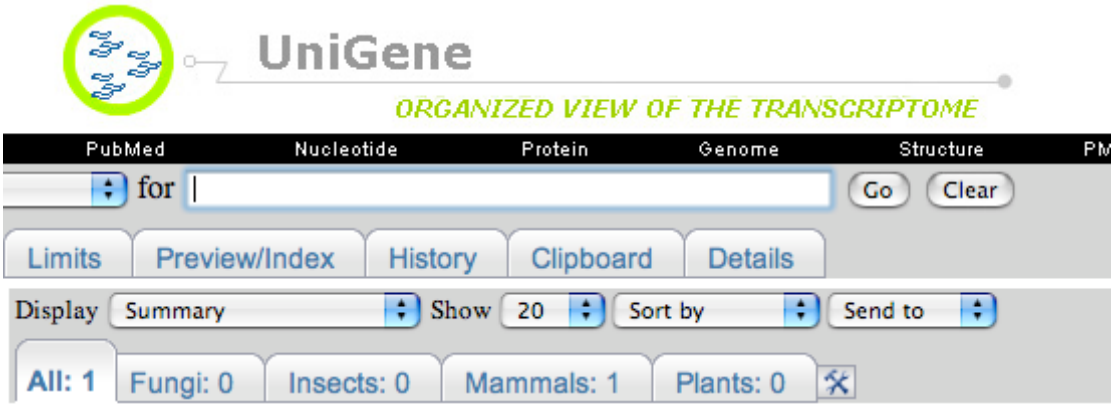


Click on entries for further information



7. Click on allelic variants for further information/ consequences, also note new links to dbSNP.

8. Click back into EntrezGene and explore Unigene from the links menu



1: Hs.75527

ADSL: Adenylosuccinate lyase

Homo sapiens, 592 sequence(s)

Adenylosuccinate lyase (ADSL)

SELECTED PROTEIN SIMILARITIES

Comparison of sequences in UniGene with proteins supported by a complete genome. The alignments can suggest function of a gene.

C. elegans	ref:NP_492049.1 - adenylosuccinate lyase [Caenorhabditis elegans]	46.68 % / 464 aa (see ProtEST)
H. sapiens	ref:NP_000017.1 - adenylosuccinate lyase; adenylosuccinase [Homo sapiens]	100 % / 484 aa (see ProtEST)
M. musculus	sp:P54822 - PUR8_MOUSE Adenylosuccinate lyase	93.6 % / 484 aa (see ProtEST)
S. cerevisiae	pir:S51377 - S51377 probable membrane protein YLR359w - yeast	64.39 % / 466 aa (see ProtEST)

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

Expression Profile:	View expression levels using UniGene's EST ProfileViewer [Show more entries with profiles like this]
GEO profiles:	Gene expression profiles in the NCBI Gene Expression Omnibus database
cDNA Sources:	uncharacterized tissue; brain; placenta; lung; mixed; embryonic tissue; testis; kidney; colon; liver; uterus; lymph node; eye; muscle; skin; whole brain; blood; prostate; mammary gland; bone; ovary; whole body; pancreas; connective tissue; heart; stomach; cervix; adipose tissue; small intestine; thyroid; esophagus; bladder; pituitary gland; tonsil; salivary gland; thymus; trachea; cochlea; ascites; pharynx; lymph; bone marrow; vascular; mouth; spleen; rectum; dorsal root ganglion; parathyroid

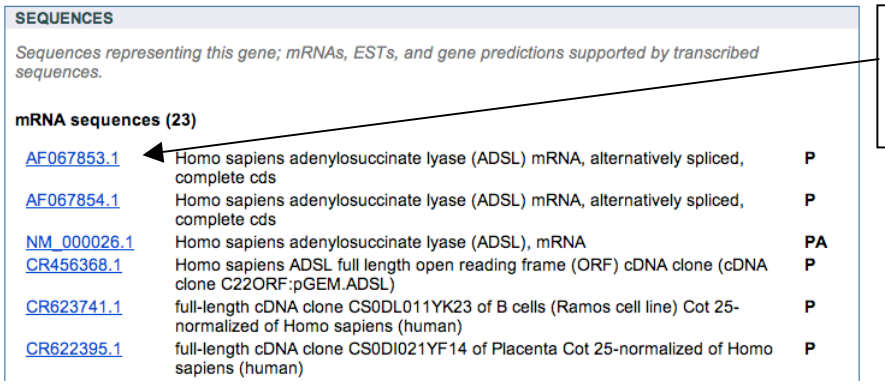
MAPPING POSITION

Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.

Chromosome:	22		
Map position:	22q13.1|22q13.2		
UniSTS entry:	Chr 22	RH71398	[Map Viewer]
UniSTS entry:		D22S966E	
UniSTS entry:		RH27785	
UniSTS entry:	Chr 22	RH77705	[Map Viewer]
UniSTS entry:	Chr 22	RH98333	

9. Scrolling through the UniGene page for ADSL you can see the mRNAs and ESTs clustering to the ADSL locus. Click on any one of these to see sequence information and links to the Trace archive.

10. Also from this page is a link to UniGene's EST profile viewer and a link to GEO profiles which can also be accessed from the links menu.



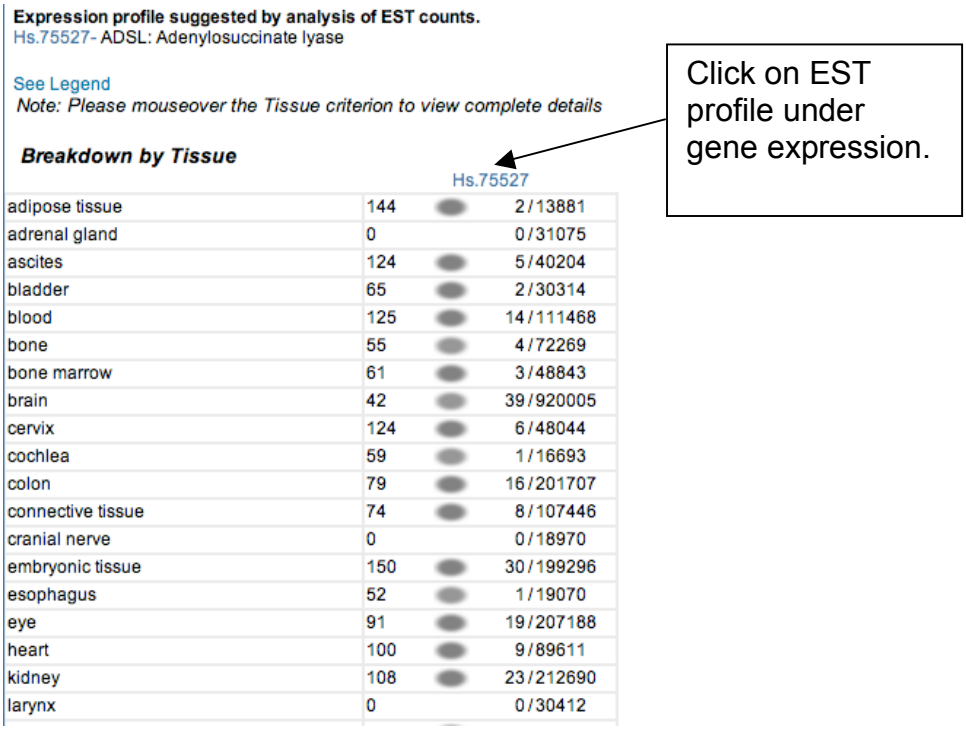
SEQUENCES

Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.

mRNA sequences (23)

AF067853.1	Homo sapiens adenylosuccinate lyase (ADSL) mRNA, alternatively spliced, complete cds	P
AF067854.1	Homo sapiens adenylosuccinate lyase (ADSL) mRNA, alternatively spliced, complete cds	P
NM_000026.1	Homo sapiens adenylosuccinate lyase (ADSL), mRNA	PA
CR456368.1	Homo sapiens ADSL full length open reading frame (ORF) cDNA clone (cDNA clone C22ORF:pGEM.ADSL)	P
CR623741.1	full-length cDNA clone CS0DL011YK23 of B cells (Ramos cell line) Cot 25-normalized of Homo sapiens (human)	P
CR622395.1	full-length cDNA clone CS0DI021YF14 of Placenta Cot 25-normalized of Homo sapiens (human)	P

Click on entries for further information plus links to the trace archive.



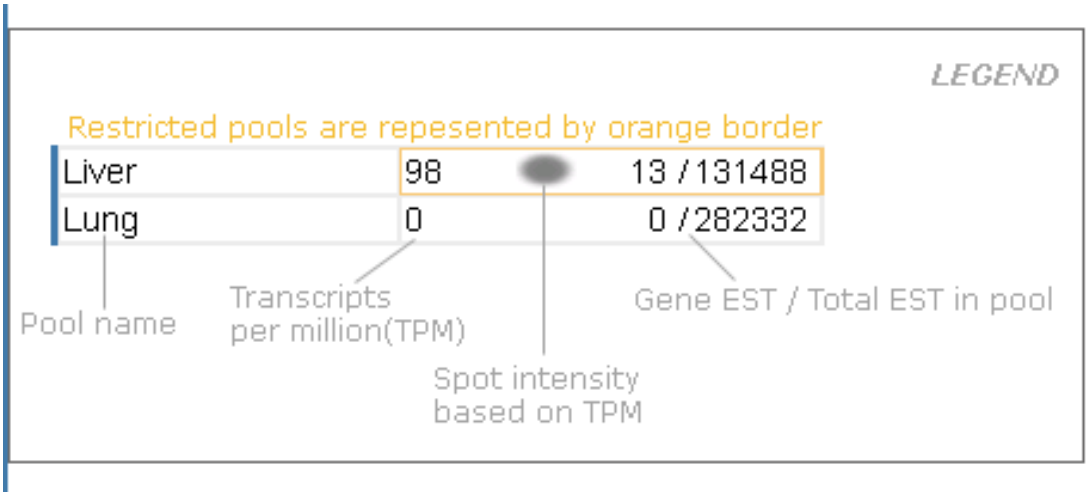
Expression profile suggested by analysis of EST counts.
Hs.75527- ADSL: Adenylosuccinate lyase

See Legend
Note: Please mouseover the Tissue criterion to view complete details

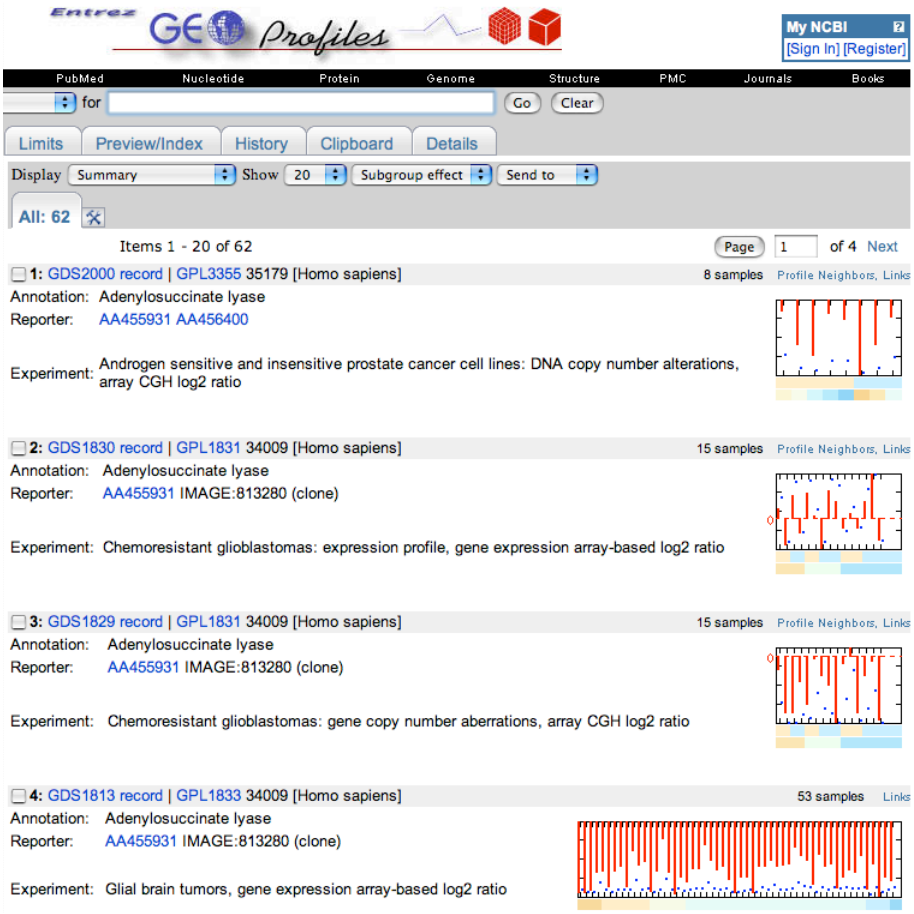
Breakdown by Tissue

		Hs.75527
adipose tissue	144	2/13881
adrenal gland	0	0/31075
ascites	124	5/40204
bladder	65	2/30314
blood	125	14/111468
bone	55	4/72269
bone marrow	61	3/48843
brain	42	39/920005
cervix	124	6/48044
cochlea	59	1/16693
colon	79	16/201707
connective tissue	74	8/107446
cranial nerve	0	0/18970
embryonic tissue	150	30/199296
esophagus	52	1/19070
eye	91	19/207188
heart	100	9/89611
kidney	108	23/212690
larynx	0	0/30412

Click on EST profile under gene expression.



11. For more expression profiles scroll back to the link for GEO profiles. (Which can also be accessed from the Entrez Gene page)



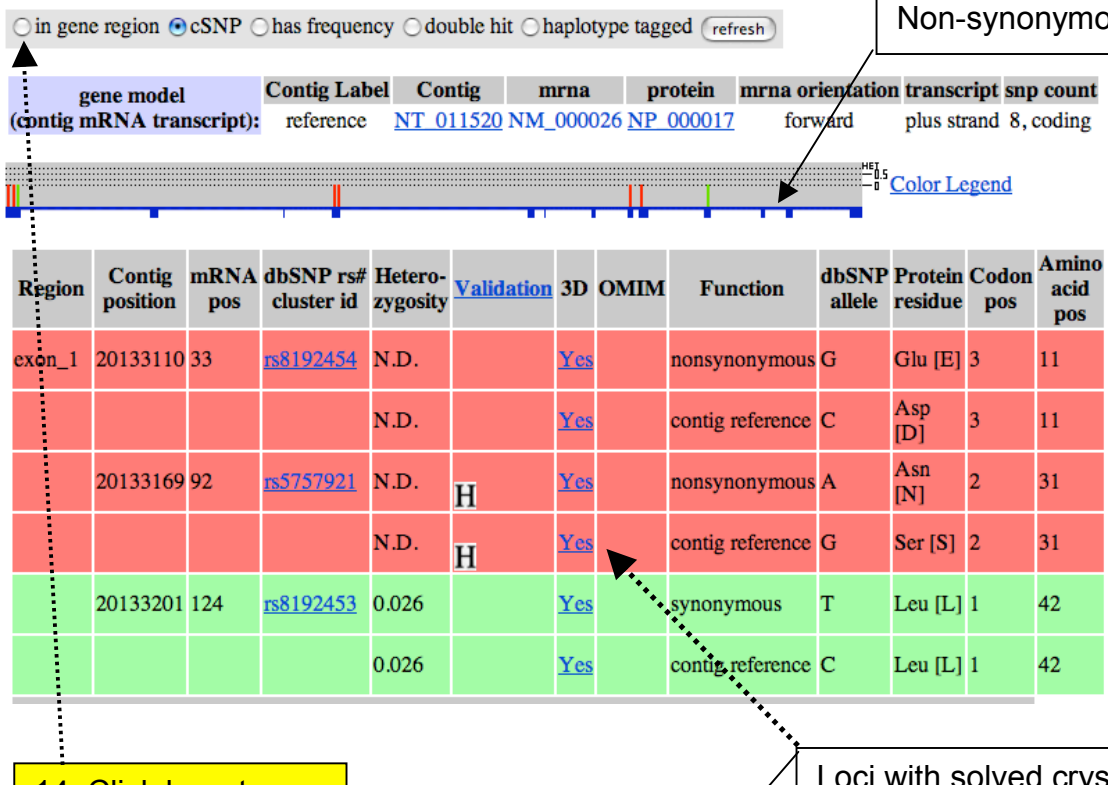
The Geo profiles database stores individual gene expression and molecular abundance profiles assembled from the Gene Expression Omnibus (GEO) repository. Search for specific profiles of interest based on gene annotation or pre-computed profile characteristics. GEO Profiles facilitates powerful searching and linking to additional information sources.

12. From the ADSL links in EntrezGene click on SNP, to get a list of the SNP associated with ADSL

Quick links to NCBI datasets

13. Display SNPs in a gene centric view by clicking on SNP:Geneview

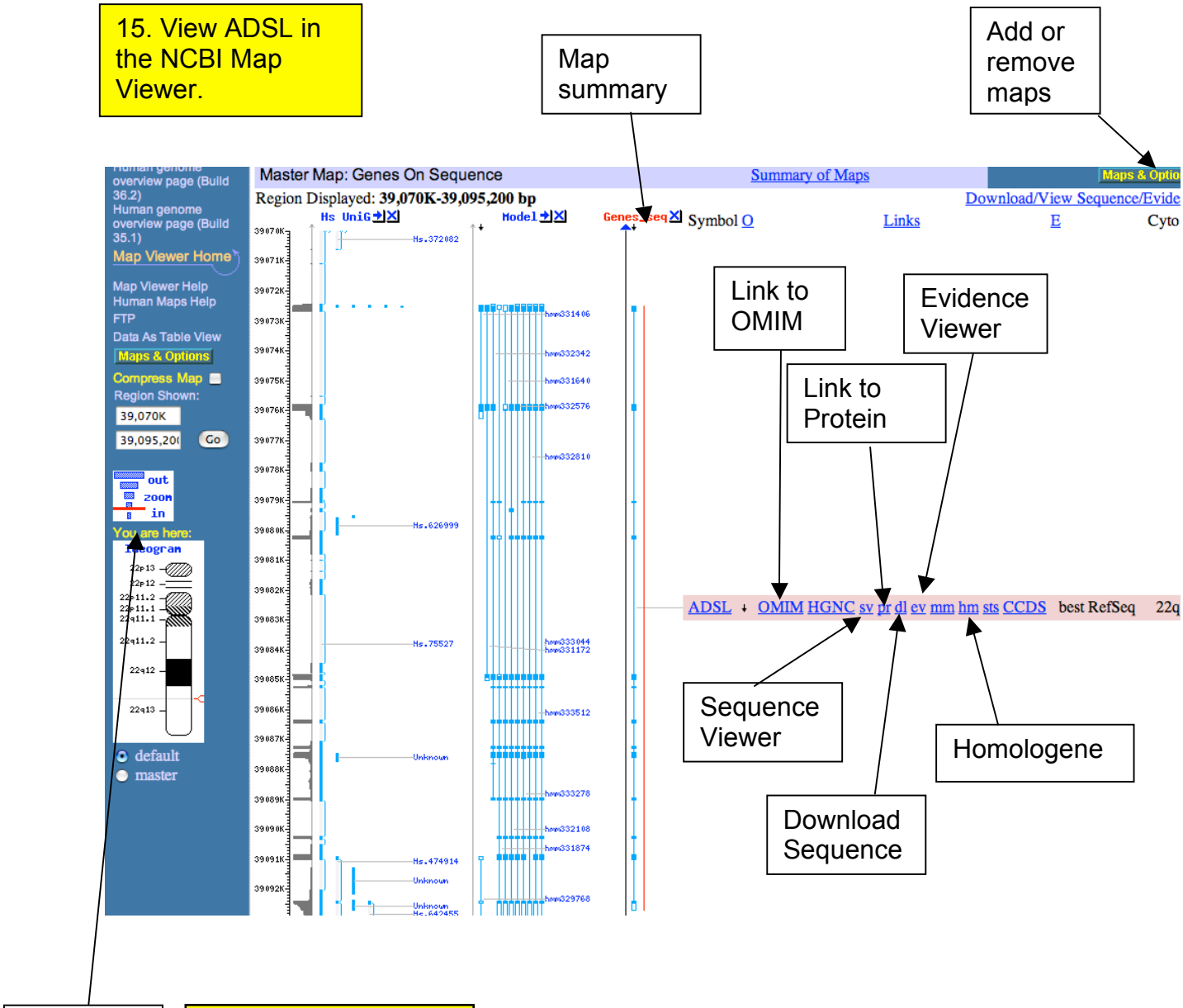
mRNA/ Genomic alignment of ADSL also showing location of Non-synonymous SNP



14. Click here to view all SNPs in gene region

Loci with solved crystal structures can be viewed with coding SNPs highlighted. See Cn3D at the NCBI for more details.

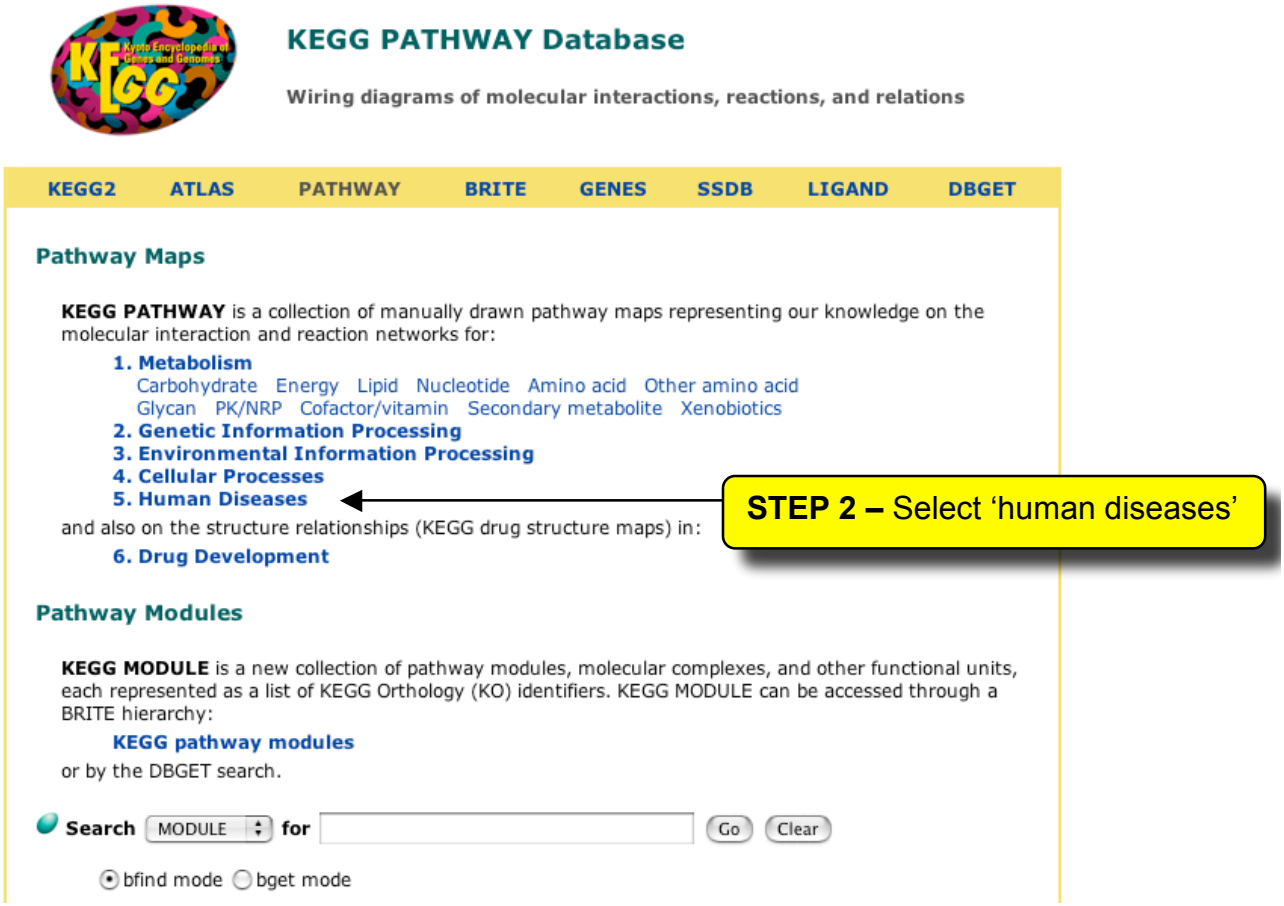
15. View ADSL in the NCBI Map Viewer.



16. Go to maps and options and show the following:
Human Uni Gene
Ab initio models
GenBank DNA
Human ESTs

KEGG – The KEGG database contains a description of cellular pathways. It is more commonly used to analyse metabolic pathways, but it also contains disease related pathways. In the following **worked example** you will be shown how to find information on disease related pathways.

STEP 1 – Go to the KEGG homepage at:
http://www.genome.ad.jp/kegg/pathway.html



5. Human Diseases

5.1 Cancers
- Colorectal cancer
- Pancreatic cancer
- Glioma
- Thyroid cancer
- Acute myeloid leukemia
- Chronic myeloid leukemia
- Basal cell carcinoma
- Melanoma
- Renal cell carcinoma
- Bladder cancer
- Prostate cancer
- Endometrial cancer
- Small cell lung cancer
- Non-small cell lung cancer

5.2 Immune Disorders
- Asthma *New!*

5.3 Neurodegenerative Diseases
- Alzheimer's disease
- Parkinson's disease
- Amyotrophic lateral sclerosis (ALS)
- Huntington's disease
- Dentatorubropallidoluysian atrophy (DRPLA)
- Prion diseases

5.4 Metabolic Disorders
- Type I diabetes mellitus
- Type II diabetes mellitus
- Maturity onset diabetes of the young

5.5 Infectious Diseases
- Cholera
- Epithelial cell signaling in Helicobacter pylori infection
- Pathogenic Escherichia coli infection

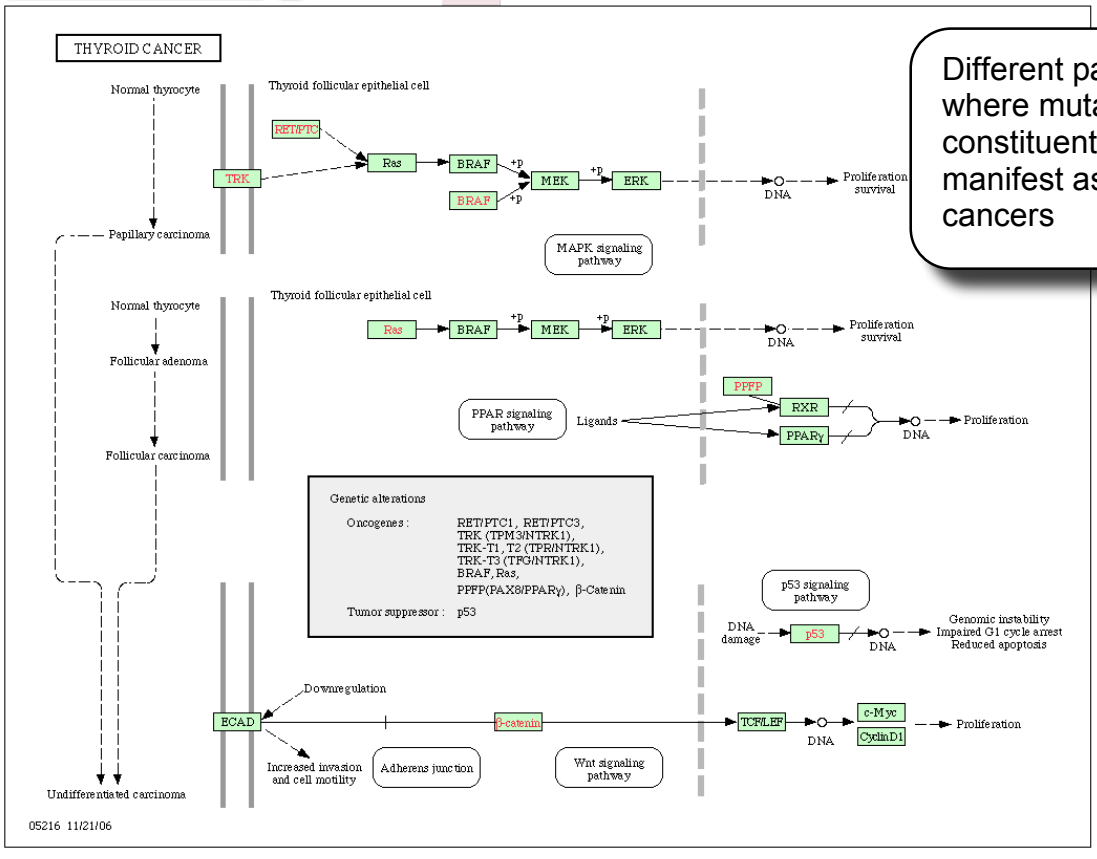
Infectious diseases

STEP 3 – Select 'thyroid cancer'

KEGG **Thyroid cancer - Homo sapiens (human)**

[Pathway menu | Reference list]

Homo sapiens (human) Go Current selection Select



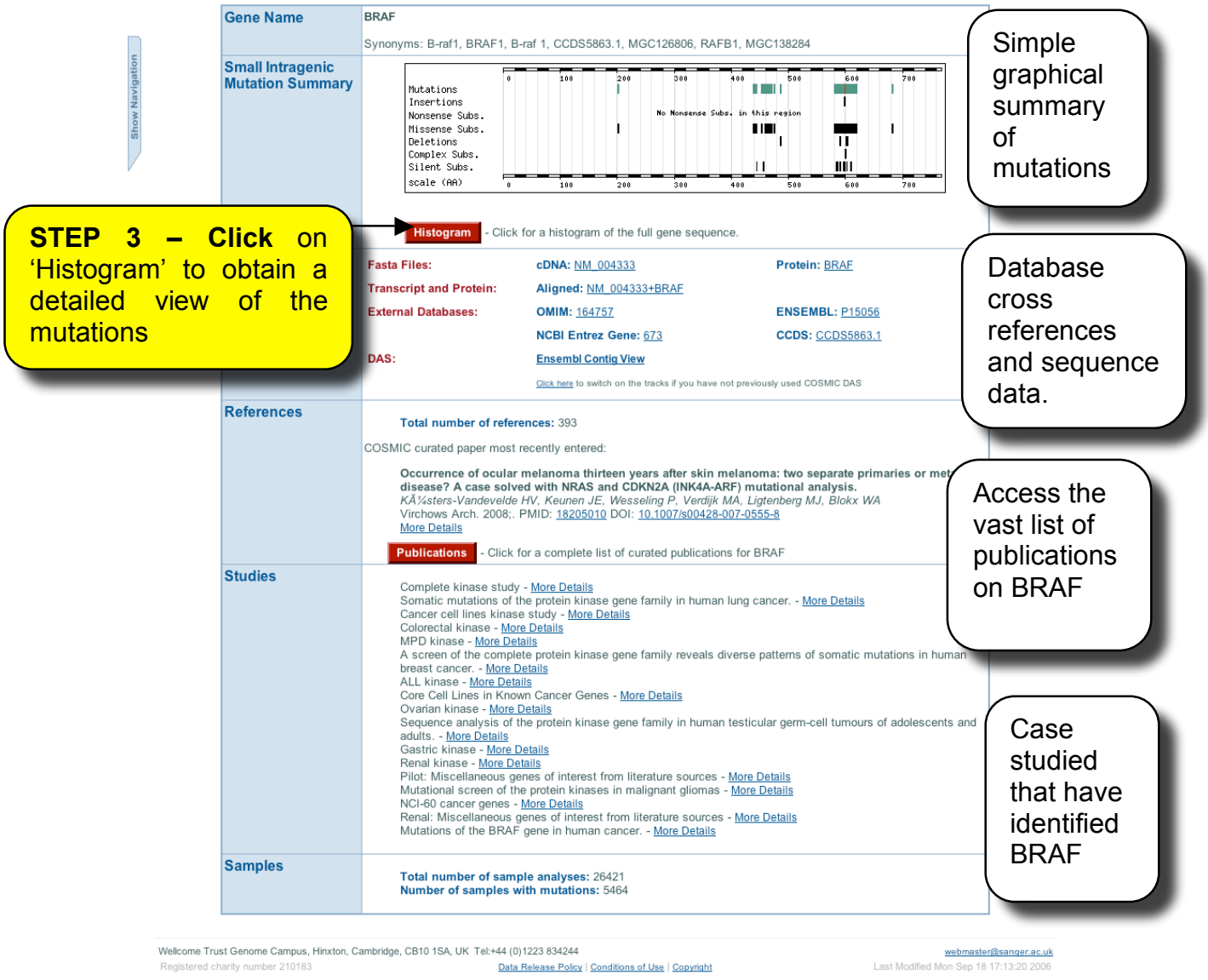
Different pathways where mutations in constituent proteins manifest as thyroid cancers

COSMIC – Although OMIM is very detailed, it is not comprehensive. COSMIC, the catalogue of somatic mutations in cancer, is a specialist resource that aims to have a comprehensive list of genes and their mutations that are involved in cancer. There are several different ways to search COSMIC, in the following worked example, the most common search interface will be illustrated.

Worked Example - List all mutations found in the BRAF gene.

STEP 1 – Go to the COMSIC homepage: http://www.sanger.ac.uk/genetics/CGP/cosmic/

STEP 2 – Enter BRAF into the textfield and **click** search.



STEP 3 – Click on 'Histogram' to obtain a detailed view of the mutations

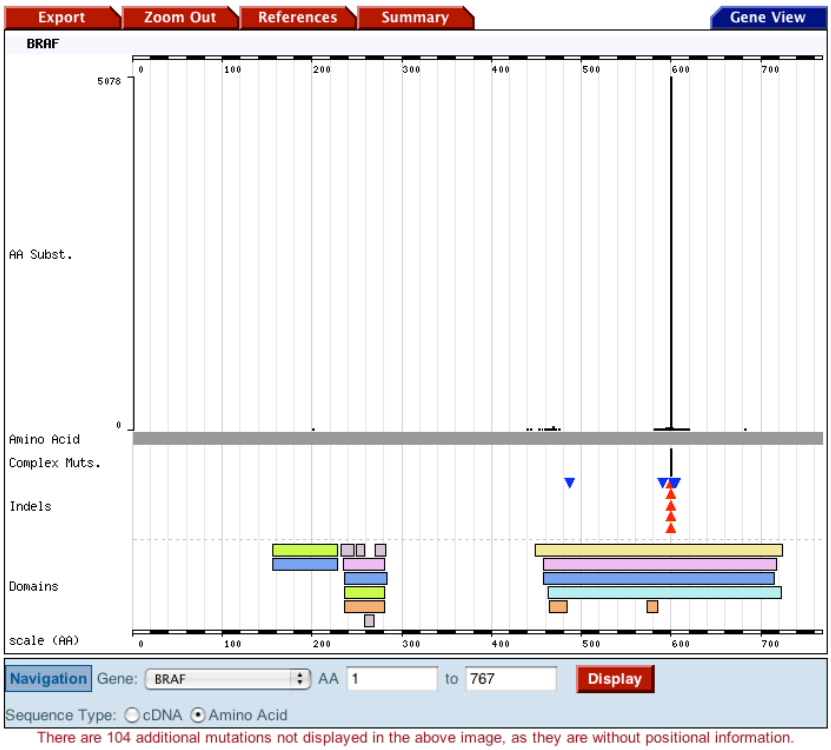
Simple graphical summary of mutations

Database cross references and sequence data.

Access the vast list of publications on BRAF

Case studied that have identified BRAF

Distribution of somatic mutations in BRAF



There are 104 additional mutations not displayed in the above image, as they are without positional information.

Details | Mutations

The histogram shows the frequency that amino acid position has been found to be mutated.

Blue triangles denote deletions, red insertions

STEP 4 – select 'Mutations' to reveal the molecular **details** of the mutations

Tissue sample summary for BRAF mutations

Details for BRAF

Primary Tissue	Mutated Samples	% Mutated	All Samples	Mutation Data
NS	199	60%	332	More Details
adrenal gland	0	0%	4	More Details
autonomic ganglia	1	1%	129	More Details
biliary tract	23	14%	159	More Details
bone	0	0%	61	More Details
breast	4	2%	171	More Details
central nervous system	16	3%	488	More Details
cervix	5	1%	368	More Details
endometrium	8	1%	561	More Details
eye	28	7%	396	More Details
gastrointestinal tract (site indeterminate)	0	0%	1	More Details
genital tract	0	0%	22	More Details
haematopoietic and lymphoid tissue	13	2%	788	More Details
kidney	1	0%	236	More Details
large intestine	1009	14%	7424	More Details
liver	2	3%	61	More Details
lung	26	2%	1310	More Details
oesophagus	3	2%	138	More Details
ovary	125	13%	982	More Details
pancreas	5	2%	227	More Details
pituitary	1	2%	50	More Details
placenta	0	0%	3	More Details
pleura	0	0%	79	More Details
prostate	21	6%	343	More Details
salivary gland	0	0%	3	More Details
skin	1919	41%	4631	More Details
small intestine	2	5%	42	More Details
soft tissue	10	3%	289	More Details
stomach	10	1%	816	More Details
testis	0	0%	24	More Details
thyroid	2024	36%	5628	More Details
upper aerodigestive tract	9	2%	444	More Details
urinary tract	0	0%	208	More Details
vulva	0	0%	3	More Details
Totals	5464	21%	26421	More Details

Substitutions

Position	Mutation(n)
201	p.Q201H(1)
439	p.K439Q(1) p.K439T(1)
440	p.T440P(1)
443	p.R443T(1)
444	p.R444Q(1) p.R444R(1) p.R444W(2)
453	p.P453T(2)
456	p.Q456Q(1)
459	p.V459L(1)
462	p.R462I(2)
463	p.I463S(1)
464	p.G464E(4) p.G464R(1) p.G464V(3)
466	p.G466A(1) p.G466E(4) p.G466R(2) p.G466V(5)
468	p.F468C(1)
469	p.G469A(13) p.G469E(5) p.G469R(1) p.G469R(4) p.G469S(3) p.G469S(1) p.G469S(1) p.G469V(7)
471	p.V471F(3)
475	p.K475M(1)
581	p.N581I(1) p.N581S(2)

⋮ ⋮

Insertions

Position	Mutation(n)
598	p.A598_T599insV(1)
599	p.T599_V600insDFGLAT(1) p.T599_V600insTT(2) p.T599_V600insTT(1) p.T599_V600insV(1)

Deletions

Position	Mutation(n)
486	p.N486_P490del(1)
590	p.V590fs*3(1)
601	p.K601del(3)
604	p.W604del(1)

Complex

Position	Mutation(n)
600	p.V600_K601>E(3) p.V600_S605>D(1)

Fusion Mutations

Mutation(n)
No Fusion Mutations in Current Selection

Other Mutations

Position	Mutation(n)
	p.?(1) p.?(1) p.?(102)

The ‘Mutations’ page lists all of the different types of mutations found, including amino acid transitions and frequencies in brackets().

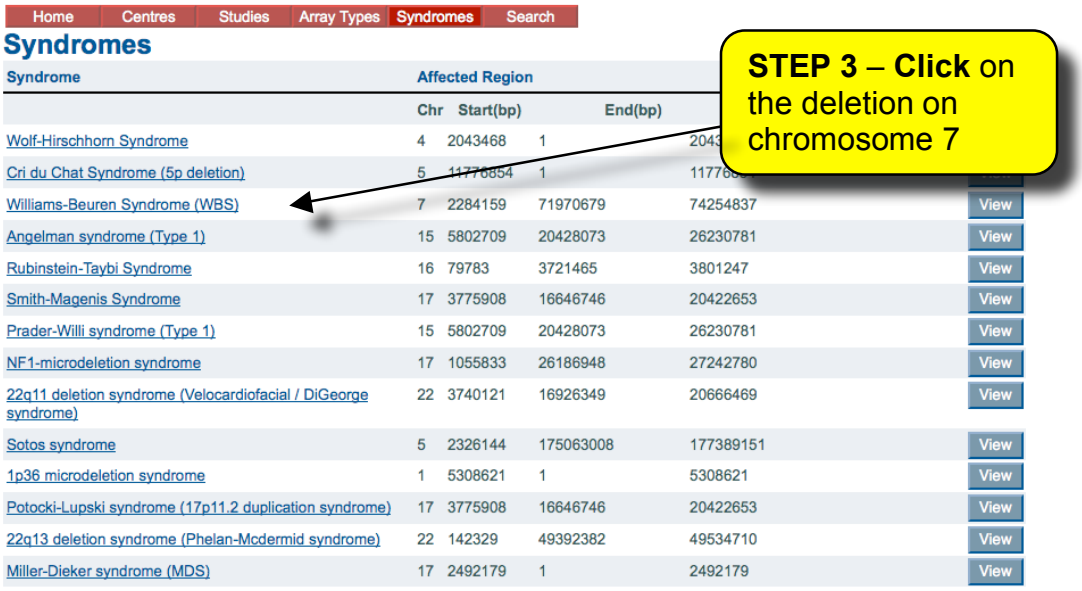
Disease Phenotype Resources

DECIPHER is a database of microscopic chromosomal imbalances and phenotypes that integrates into Ensembl. This database is a departure from traditional bioinformatics resources, where the focus is primarily on the description of the phenotype caused by a genetic defect by clinicians. As there is patient data within the system, there are different levels of access. We will be using the *Guest Access*.

Worked example: In the following example, you will use DECIPHER to investigate Williams-Beuren Syndrome.

STEP 1 – Go to the DECIPHER homepage
https://decipher.sanger.ac.uk/

The following page lists all of the syndromes represented by the karyotype view. Note, red indicates deletions and green inserts.



Syndrome	Affected Region				
	Chr	Start(bp)		End(bp)	
Wolf-Hirschhorn Syndrome	4	2043468	1	2043	
Cri du Chat Syndrome (5p deletion)	5	11776854	1	1177685	
Williams-Beuren Syndrome (WBS)	7	2284159	71970679	74254837	View
Angelman syndrome (Type 1)	15	5802709	20428073	26230781	View
Rubinstein-Taybi Syndrome	16	79783	3721465	3801247	View
Smith-Magenis Syndrome	17	3775908	16646746	20422653	View
Prader-Willi syndrome (Type 1)	15	5802709	20428073	26230781	View
NF1-microdeletion syndrome	17	1055833	26186948	27242780	View
22q11 deletion syndrome (Velocardiofacial / DiGeorge syndrome)	22	3740121	16926349	20666469	View
Sotos syndrome	5	2326144	175063008	177389151	View
1p36 microdeletion syndrome	1	5308621	1	5308621	View
Potocki-Lupski syndrome (17p11.2 duplication syndrome)	17	3775908	16646746	20422653	View
22q13 deletion syndrome (Phelan-Mcdermid syndrome)	22	142329	49392382	49534710	View
Miller-Dieker syndrome (MDS)	17	2492179	1	2492179	View

From this link, there is a vast amount of detail about the mutation, how is was identified, literature references, phenotypic and genetic information.

Clinical - Characteristic facial features include periorbital fullness, bulbous nasal tip, long philtrum, wide mouth, full lips, full cheeks a
spaced teeth. Individuals have mild to moderate intellectual disability or learning difficulties with relative cognitive strengths in verbal
and in language but extreme weakness in visuospatial construction (writing, drawing, pattern construction). Distinctive behavioural ch
include anxiety, attention deficit hyperactivity disorder (ADHD), and overfriendliness. Congenital heart disease occurs in 80%, with th
supravalvular aortic stenosis (SVAS), and a smaller proportion having a discrete supravalvular pulmonary stenosis.
The microdeletion on 7q11.23 encompasses the elastin gene (ELN) which is also mutated in isolated SVAS. Other symptoms include
impairment, hypersensitivity to sound, chronic otitis media, malocclusion, small or missing teeth, renal anomalies, constipation, vomi
deficiency, infantile hypercalcemia, musculoskeletal abnormalities, diabetes and a hoarse voice. Risk for hypertension has been linke
the distal deletion breakpoint, with hypertension being significantly less prevalent in WBS patients with a deletion that includes NCF1
coding for the p47phox subunit of the NADPH oxidase. This likely arises through life-long reduced angiotensin II-mediated oxidative

Size of deletion - Three large region-specific LCRs, termed centromeric, medial and telomercic, flank the WBS deletion interval. Eac
hundred kb in length and is comprised of transcriptionally active genes and pseudogenes grouped into discreet blocks known as A, B an
patients (>95%) have a 1.55Mb deletion caused by recombination between centromeric and medial block B copies, which share approximately 99.6%
nucleotide identity over many kilobases. There are at hot-spots of recombination: one within a 12 kb region of the GTF2I gene, and one in the distal end
of the GTF2IRD2 gene. A few patients (<5%) have a larger deletion (~1.84Mb) caused by recombination between centromeric and medial block A
copies.

Origin of deletion - Almost one-third (28%) of the transmitting progenitors are heterozygous for an inversion between centromeric and telomeric LCRs which may facilitate the deletion. The deletions are caused by nonhomologous recombination within the LCRs of either the same chromosome 7 (intrachromosomal) or different chromosome 7s (interchromosomal). In each case the chromosomes are envisaged to form loops, thereby allowing the alignment of the two LCRs, the occurrence of recombination, and the excision of the DNA contained within the intervening loop. Approximately 2/3rds of the deletion events are interchromosomal.

Expert advisors
Dr. Stephen W. Scherer The Hospital for Sick Children, Toronto, Canada and Dr. Lucy Osborne, University of Toronto, Canada
Links to further information and support groups:
http://williams-syndrome.org/

http://www.williams-syndrome.org/fordoctors/growthcharts.html

http://www.geneclinics.org/servlet/access?db=geneclinics&site=gt&id=8888892&key=-OsGtBoTltKT2&gry=&fcn=y&fw=aqRv&filename=/profiles/williams/index.html

Citations (9)

Literature references

American Academy of Pediatrics: Health care supervision for children with Williams syndrome.
Committee on Genetics
Pediatrics. 2001;107;1192-204. PMID: 11331709
A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome.
Osborne LR, Li M, Pober B, Chitayat D, Bodurtha J, Mandel A, Costa T, Grebe T, Cox S, Tsui LC, Scherer SW
Nat Genet. 2001;29;321-5. PMID: 11685205
Cardiovascular manifestations in 75 patients with Williams syndrome.
Eronen M, Peippo M, Hiippala A, Raatikka M, Arvio M, Johansson R, Kähkönen M
J Med Genet. 2002;39;554-8. PMID: 12161592
Mutational mechanisms of Williams-Beuren syndrome deletions.
Bayés M, Magano LF, Rivera N, Flores R, Pérez Jurado LA
Am J Hum Genet. 2003;73;131-51. PMID: 12796854
Williams-Beuren syndrome: a challenge for genotype-phenotype correlations.
Tassabehji M
Hum Mol Genet. 2003;12 Spec No 2;R229-37. PMID: 12952863
GTF2I hemizygosity implicated in mental retardation in Williams syndrome: genotype-phenotype analysis of five families with deletions in the Williams syndrome region.
Morris CA, Mervis CB, Hobart HH, Gregg RG, Bertrand J, Ensing GJ, Sommer A, Moore CA, Hopkin RJ, Spallone PA, Keating MT, Osborne L, Kimberley KW, Stock AD
Am J Med Genet A. 2003;123;45-59. PMID: 14556246
GTF2IRD1 in craniofacial development of humans and mice.
Tassabehji M, Hammond P, Karmiloff-Smith A, Thompson P, Thorgeirsson SS, Durkin ME, Popescu NC, Hutton T, Metcalfe K, Rucka A, Stewart H, Read AP, Maconochie M, Donnai D
Science. 2005;310;1184-7. PMID: 16293761 DOI: 10.1126/science.1116142
Hemizygosity at the NCF1 gene in patients with Williams-Beuren syndrome decreases their risk of hypertension.
Del Campo M, Antonell A, Magano LF, Muñoz FJ, Flores R, Bayés M, Pérez Jurado LA
Am J Hum Genet. 2006;78;533-42. PMID: 16532385 DOI: 10.1086/501073
Neural mechanisms in Williams syndrome: a unique window to genetic influences on cognition and behaviour.
Meyer-Lindenberg A, Mervis CB, Berman KF
Nat Rev Neurosci. 2006;7;380-93. PMID: 16760918 DOI: 10.1038/nrn1906

Phenotypes (7)

Primary	Secondary	Tertiary
STATURE	**Short stature, general abnormalities**	
FACE	Malar region, general abnormalities	**Flat malar region**
MOUTH	Lower lip, general abnormalities	**Prominent/everted lower lip**
TEETH	Teeth, general abnormalities	**Small teeth**
THORAX	Heart, general abnormalities	**Aortic stenosis**
NEUROLOGY	MENTAL,COGNITIVE FUNCTION, general abnormalities	**Mental retardation/developmental delay**
NEUROLOGY	BEHAVIOURAL PROBLEMS, general abnormalities	**Short attention span**

Features

Chromosome
7
Start Position(bp)
71970679
End Position (bp)
74254837
Copy Number
1

e! cytoview

View deletion in detail in EnsEMBL

Graph HGNC OMIM (HGNC) Imprinted (HGNC) Ensembl Novel
Prioritise All Phenotypes Prioritise Individual Phenotypes Overlapping Syndromes
Overlapping Patients

4 listed

List of effected genes

ELN Chr:7Start:73080367End:73122173

elastin (supravalvular aortic stenosis, Williams-Beuren syndrome). Aliases: WBS, WS, SVAS
Ensembl:ELN Ensembl:ENSG00000049540 OMIMMorbid:130160

GTF2IRD1 Chr:7Start:73506056End:73654846

GTF2I repeat domain containing 1. Aliases: MusTRD1, RBAP2, GTF3, WBSCR12, BEN, Cream1
Ensembl:ENSG00000006704 Ensembl:GTF2IRD1 OMIMMorbid:604318

GTF2I Chr:7Start:73709966End:73812956

general transcription factor II, i. Aliases: TFII-I, BAP-135, SPIN, BTKAP1, DIWS, IB291
Ensembl:ENSG00000077809 Ensembl:GTF2I OMIMMorbid:601679

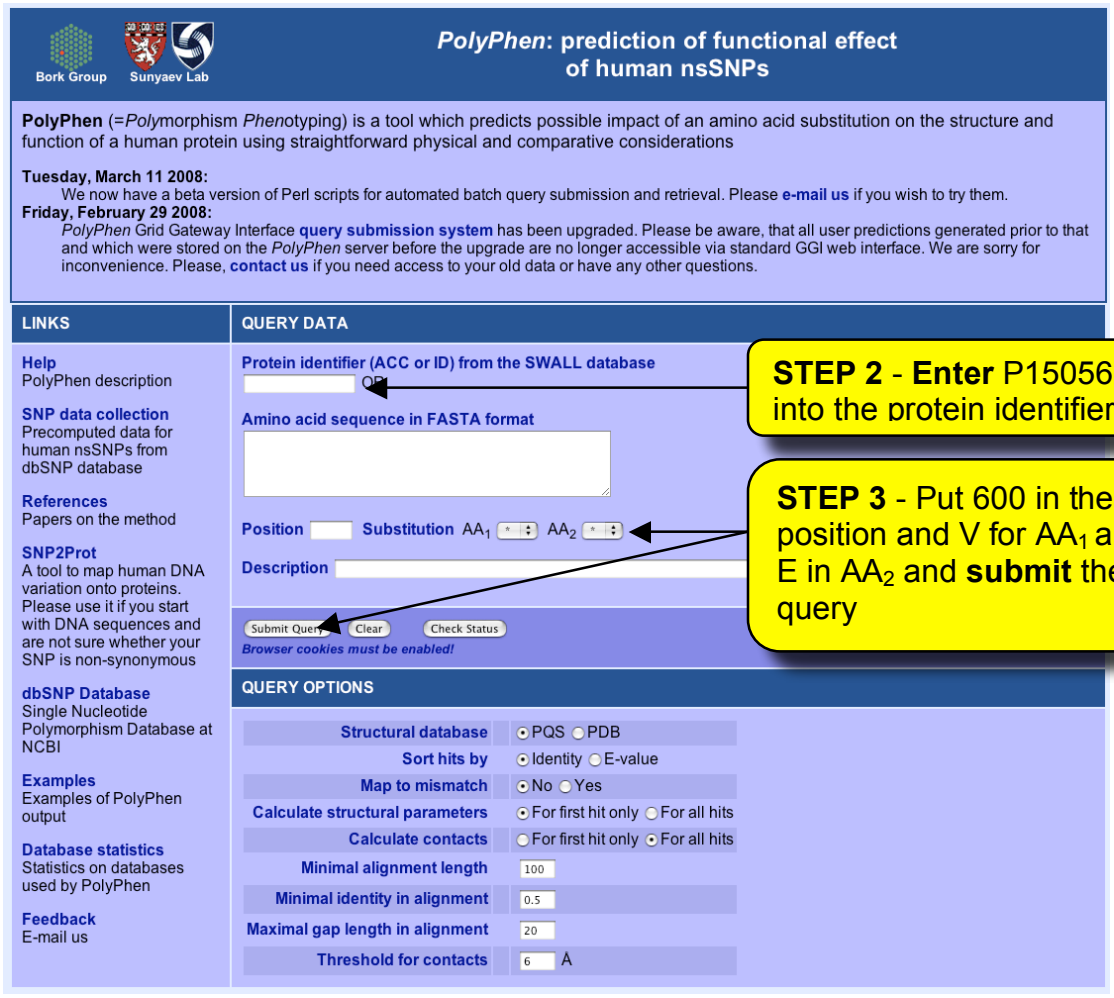
NCF1 Chr:7Start:73826245End:73841594

neutrophil cytosolic factor 1, (chronic granulomatous disease, autosomal 1). Aliases: p47phox, NOXO2, NCF1A, SH3PXD1A
Ensembl:ENSG00000158517 Ensembl:NCF1 OMIMMorbid:608512

SNPs Analysis – Having found a non-synonymous single nucleotide polymorphism (nsSNP), researches often want to know whether this is natural, tolerable variation, or whether that the SNP is potentially deleterious. Unfortunately, there is no simple answer to this question. Using methods such as database searching, homology modelling and literature searching should be used. However, the PolyPhen server does provide a tool for trying to establish the nature of a nsSNP. PolyPhen uses a variety of rules to assess the nsSNP, including sequence conservation, amino acid properties and structural context.

Worked Example – In the last section, we COSMIC was used to investigate the mutations in BRAF. The most common point mutation was a V to E transition at amino acid position 600. Lets assume we did not know the consequence of this mutation and use PolyPhen to see if it is deleterious.

STEP 1 – Go to the PolyPhen homepage:
http://genetics.bwh.harvard.edu/pph/



STEP 2 - **Enter** P15056 into the protein identifier

STEP 3 - Put 600 in the position and V for AA_1 and E in AA_2 and **submit** the query

Query

Acc number	Position	AA_1	AA_2	Description
P15056	600	V	E	B-Raf proto-oncogene serine/threonine-protein kinase (EC 2.7.11.1) LENGTH: 766 AA

Prediction

This variant is predicted to be probably damaging

Prediction	Available data	Prediction basis	Substitution effect	Prediction data
probably damaging	FT alignment structure	alignment	N/A	PSIC score difference: 2.120

Remarks

Charge change at exposed site: substitution V -> E, normed accessibility: 0.93

Details

SEQUENCE FEATURES OF THE SUBSTITUTION SITE

Region	Site	Feature table	Critical sites
N/A	N/A	show FT fields for P15056	235, 248, 251, 261, 264, 269, 272, 280, 483, 576

PSIC PROFILE SCORES FOR TWO AMINO ACID VARIANTS

Score1	Score2	|Score1-Score2|	Observations	Diagnostics	Multiple alignment around substitution position
+1.418	-0.702	**2.120**	50	precomputed	Sequences: all Flanks: 25 Show alignment

MAPPING OF THE SUBSTITUTION SITE TO KNOWN PROTEIN 3D STRUCTURES

Database	Initial number of structures	Number of structures
PQS	500	4

Num	ID		Res	AA	E-value	Len	Ide	Gaps	Params	Cont	PDB TITLE
1	1uwh	B	599	V	2.0e-156	276	1.00	12	Params		THE COMPLEX OF WILD TYPE B-RAF AND BAY439006
2	1uwh	A	599	V	2.0e-156	276	1.00	12			THE COMPLEX OF WILD TYPE B-RAF AND BAY439006
3	2fb8_2	B	600	V	5.1e-154	272	1.00	13			STRUCTURE OF THE B-RAF KINASE DOMAIN BOUND TO SB-590885
4	2fb8_1	A	600	V	5.1e-154	272	1.00	13			STRUCTURE OF THE B-RAF KINASE DOMAIN BOUND TO SB-590885

STRUCTURAL PARAMETERS

Num	ID		Res	SecStr	Acc	Acc Normed	dPropens	(Phi, Psi)	Map Region	dVol	Normed B-factor
1	**1uwh**	B	599	.	112	0.93	1.48	(-113.2, 360.0)	?	-2	2.12

STEP 4 - **Select** 10 sequences and view the alignment. **Flanks** controls how many amino acids either side of the mutated amino acid are displayed.

Fragment of multiple alignment around position 600:

0		QUERY:	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
1	sp|P34908|BRAF1_COTJA	B-Raf proto-oncogene serine/threonine-prot...	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
2	ref|XP_001070228.1|	PREDICTED: similar to v-raf murine sarcoma v...	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
3	ref|XP_001375430.1|	PREDICTED: hypothetical protein [Monodelphis...	...TNIKCRNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
4	dbj|BAD16727.1|	serine/threonine protein kinase BRAF [Danio rerio]	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
5	gb|AAI21877.1|	LOC779570 protein [Xenopus tropicalis]	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
6	gb|AAZ06667.1|	B-Raf [Xenopus laevis]	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
7	ref|NP_001032957.1|	serine/threonine protein kinase BRAF [Takifu...	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
8	dbj|BAD16728.1|	serine/threonine protein kinase BRAF	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
9	emb|CAF96750.1|	unnamed protein product [Tetraodon	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...
10	gb|AAD43193.1|AC006344_3	serine/threonine protein kinase; simi	...RDLKSNNIFLHEDLTVKIGDFGLAT V KSRWSGSHQFEQLSGSILWMAPEVI...

In this alignment, the valine is invariant.

TASKS

A series of related individuals exhibit a variety of clinical phenotypes, including early onset of colon cancer and mental retardation. Analysis of their DNA reveals a homozygous deletion of a 1 Mb region on human chromosome 11, between D11S4379 and D11S1091. The aim of this exercise is to understand the region and the genes contained within. By the end, you should have obtained enough information to be able to carry out experimental analysis of the genes and their protein products, and to further understand how the deleted region of DNA may contribute to the disease.

1) Content of the region:
    a. Search with both markers in Ensembl
    b. How many genes are there in this region.
    c. Do any genes show evidence of alternative splicing?
    d. How does the region compare in the UCSC database?

2) EntrezGene
    a. Are the genes catalogued in EntrezGene?
    b. What are their preferred symbols and full names?
    c. What are the REFSEQ entries?
    d. Is there any gene ontology information listed

3) Function of genes:
    a. Do any of the genes have an experimentally determined function?
    b. Are any genes listed in OMIM, if so what is the information
    c. For the unknown genes, are any protein domains predicted?
    d. Can you conclude what the likely function is for the genes in the region?

Polyphen exercise:
OMIM suggests that SH2D1A interacts via its SH2 domain with a motif (TIYXXV) present in the cytoplasmic tail of the cell-surface receptors CD150 (SLAM), Sayos et al. (1998) showed that SAP cDNAs isolated from the blood cells of patients with X-linked lymphoproliferative syndrome did not bind SLAM. OMIM lists 11 allelic variants of SH2D1A which result in the inability of SH2D1A to bind CD150 (SLAM)

We now want to find out whether or not any of the substitutions have an effect on the protein structure/function.
Characterise the following allelic variants:

- In a male with XLP (308240), Coffey et al. (1998) identified a 394G-C transversion in the SH2D1A gene, resulting in an arg32-to-thr (R32T) substitution.

- In a patient with XLP (308240), Coffey et al. (1998) identified a 502C-T transition in the SH2D1A gene, resulting in a thr68-to-ile (T68I) amino acid substitution.

Use PolyPhen to predict the effect of the substitutions on protein structure and function.