Module 7: Variation, Function and Disease

Aim

Learn how to explore variation and the relationship between genotype and phenotype / disease using the following tools and databases:

- The Ensembl Variation Effect Predictor (VEP)
- PolyPhen-2
- OMIM
- GEO Profiles and Gene Expression Atlas
- COSMIC
- DECIPHER
- Ontologies

Often the most valuable information to know about a variant is the effect the observed alleles have on genes, transcripts and proteins. This information can be very helpful to prioritise any variants for further investigation. To determine this effect, several tools are available. One should keep in mind though that all these tools do is make predictions and consequently findings should always be confirmed by experiments.

The Ensembl Variant Effect Predictor (VEP)

The Ensembl Variant Effect Predictor (VEP) determines the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. It also calculates SIFT and PolyPhen scores for changes to protein sequence.

Worked example 1: The Ensembl Variant Effect Predictor (VEP)

In this worked example we will study four newly found variants in human:

Deletion of an A at position 128328461 on chromosome 9
Substitution C>A at position 128322349 on chromosome 9
Substitution C>G at position 128323079 on chromosome 9
Substitution G>A at position 128322917 on chromosome 9

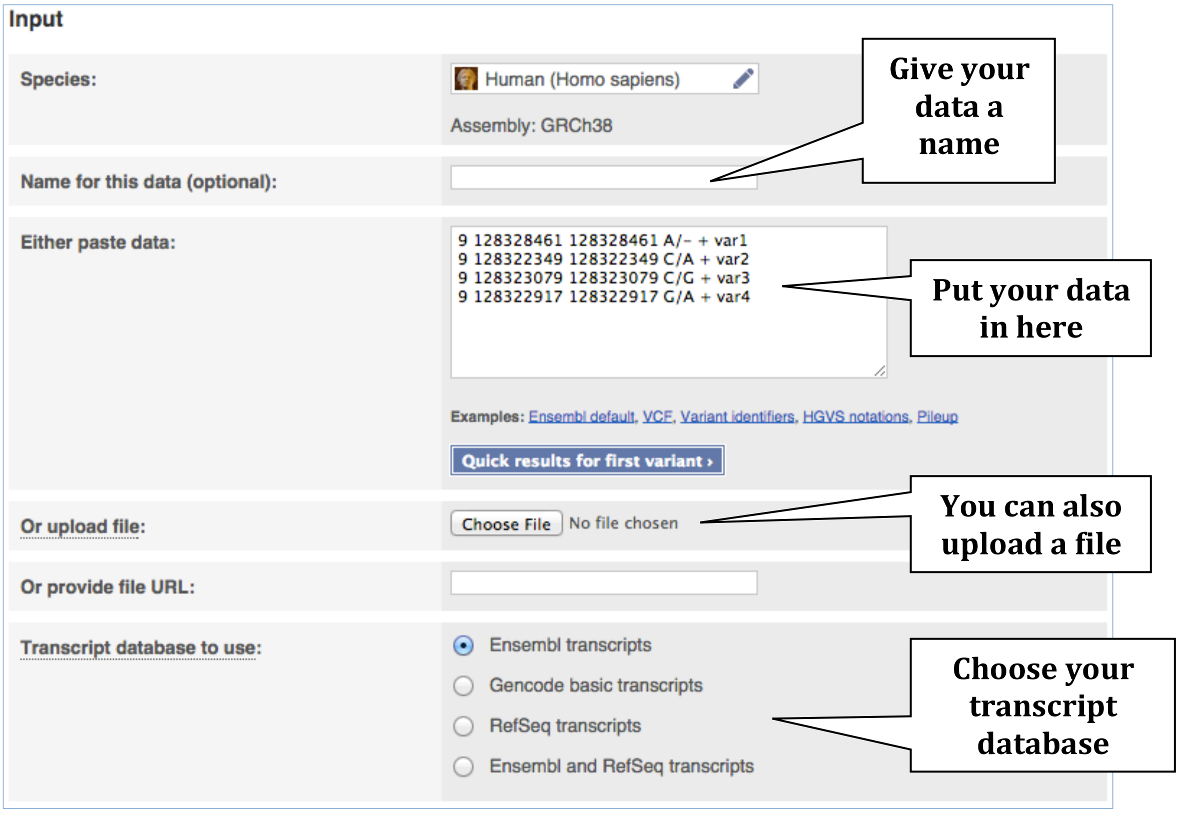
We will use the **Ensembl VEP** to answer the following questions:

- Have my variants already been annotated in Ensembl?

- What genes are affected by my variants?
- Do my variants result in protein changes?
- Do any of my variants affect gene regulation?

Go to the Ensembl Variant Effect Predictor page (http://www.ensembl.org/info/docs/tools/vep/index.html).

This page contains information about the VEP, including links to download the script version of the tool. Click on “Launch Ve!P” to open the input form.



Input

Species:	Human (Homo sapiens) Assembly: GRCh38
Name for this data (optional):	
Either paste data:	9 128328461 128328461 A/- + var1 9 128322349 128322349 C/A + var2 9 128323079 128323079 C/G + var3 9 128322917 128322917 G/A + var4 Examples: Ensembl default, VCF, Variant identifiers, HGVS notations, Pileup Quick results for first variant ›
Or upload file:	Choose File No file chosen
Or provide file URL:	
Transcript database to use:	● Ensembl transcripts ○ Gencode basic transcripts ○ RefSeq transcripts ○ Ensembl and RefSeq transcripts

Give your data a name

Put your data in here

You can also upload a file

Choose your transcript database

The data should be inputted in the following format:

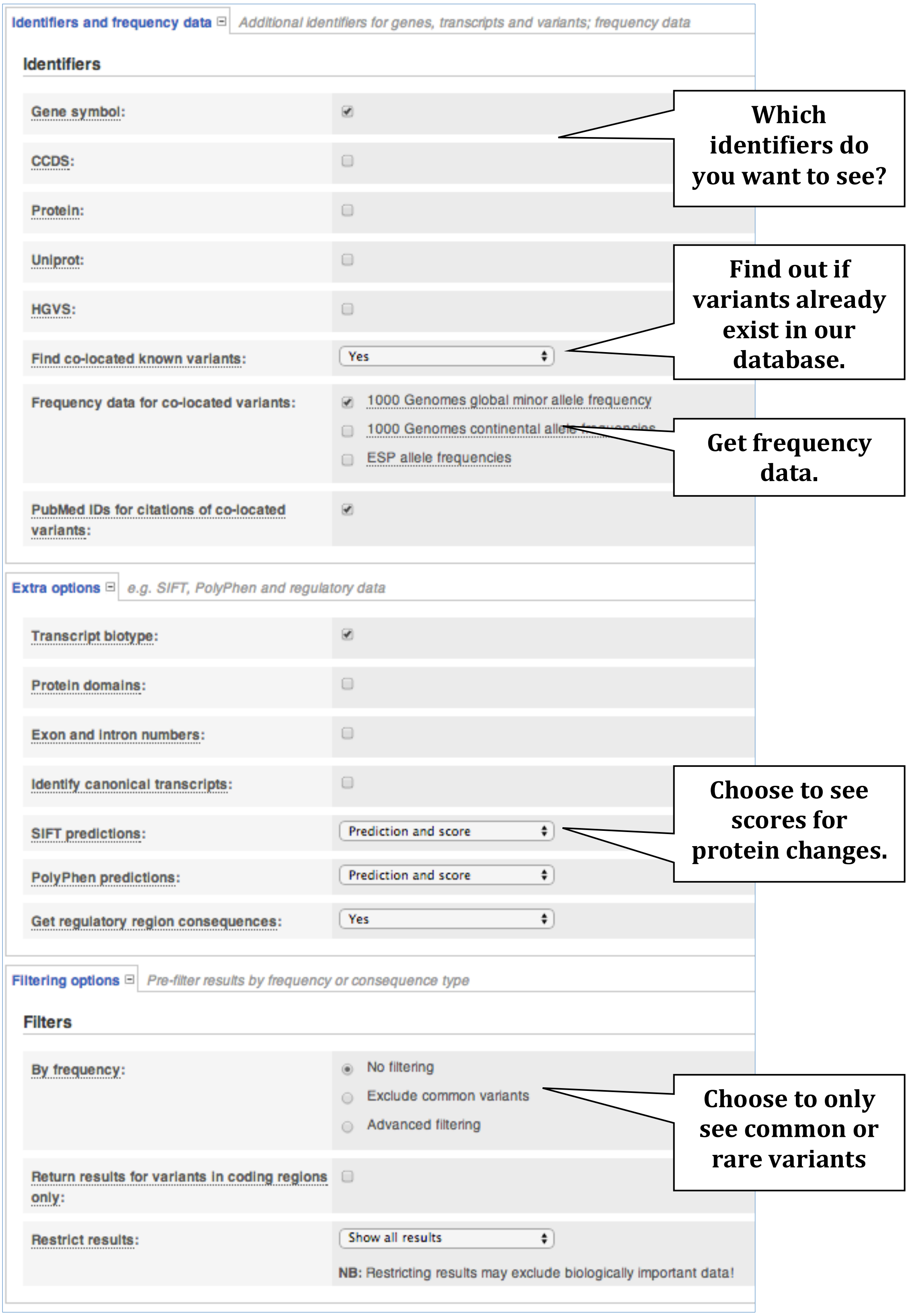
Chromosome Start End Alleles (reference/mutation) Strand Name

Replace the example data in the “Paste data” box with:

```
9 128328461 128328461 A/- + var1
9 128322349 128322349 C/A + var2
9 128323079 128323079 C/G + var3
9 128322917 128322917 G/A + var4
```

The VEP will automatically detect that the data is in “Ensembl default format”.

There are further options that you can choose for your output. These are categorised as “Identifiers and frequency data”, “Filtering options” and “Extra options”. Let’s open all the menus and take a look.



Identifiers and frequency data ⊟ *Additional identifiers for genes, transcripts and variants; frequency data*

Identifiers

Gene symbol: ☑

CCDS: □

Protein: □

Uniprot: □

HGVS: □

Find co-located known variants: Yes

Frequency data for co-located variants:
☑ 1000 Genomes global minor allele frequency
□ 1000 Genomes continental allele frequencies
□ ESP allele frequencies

PubMed IDs for citations of co-located variants: ☑

Extra options ⊟ *e.g. SIFT, PolyPhen and regulatory data*

Transcript biotype: ☑

Protein domains: □

Exon and intron numbers: □

Identify canonical transcripts: □

SIFT predictions: Prediction and score

PolyPhen predictions: Prediction and score

Get regulatory region consequences: Yes

Filtering options ⊟ *Pre-filter results by frequency or consequence type*

Filters

By frequency:
● No filtering
○ Exclude common variants
○ Advanced filtering

Return results for variants in coding regions only: □

Restrict results: Show all results

NB: Restricting results may exclude biologically important data!

Which identifiers do you want to see?

Find out if variants already exist in our database.

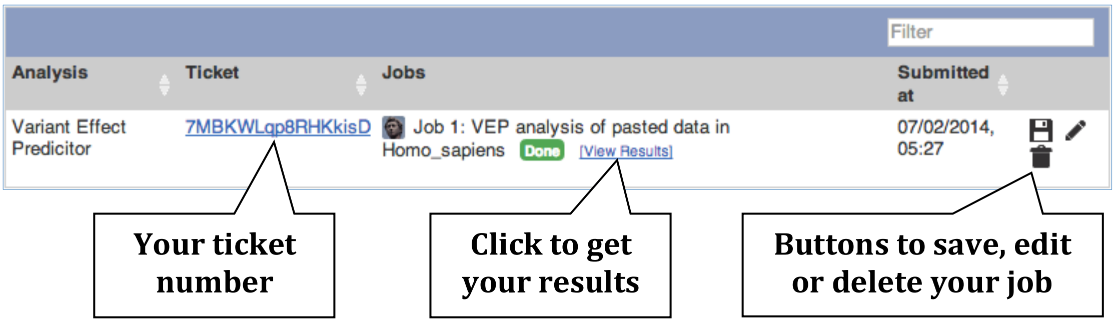
Get frequency data.

Choose to see scores for protein changes.

Choose to only see common or rare variants

Hovering with your mouse over an option will show a pop-up with an explanation of that option.

When you've selected everything you need, scroll right to the bottom and click [Run].



Filter

Analysis	Ticket	Jobs	Submitted at	
Variant Effect Predicitor	7MBKWLqp8RHKkisD	Job 1: VEP analysis of pasted data in Homo_sapiens Done [View Results]	07/02/2014, 05:27	

Your ticket number

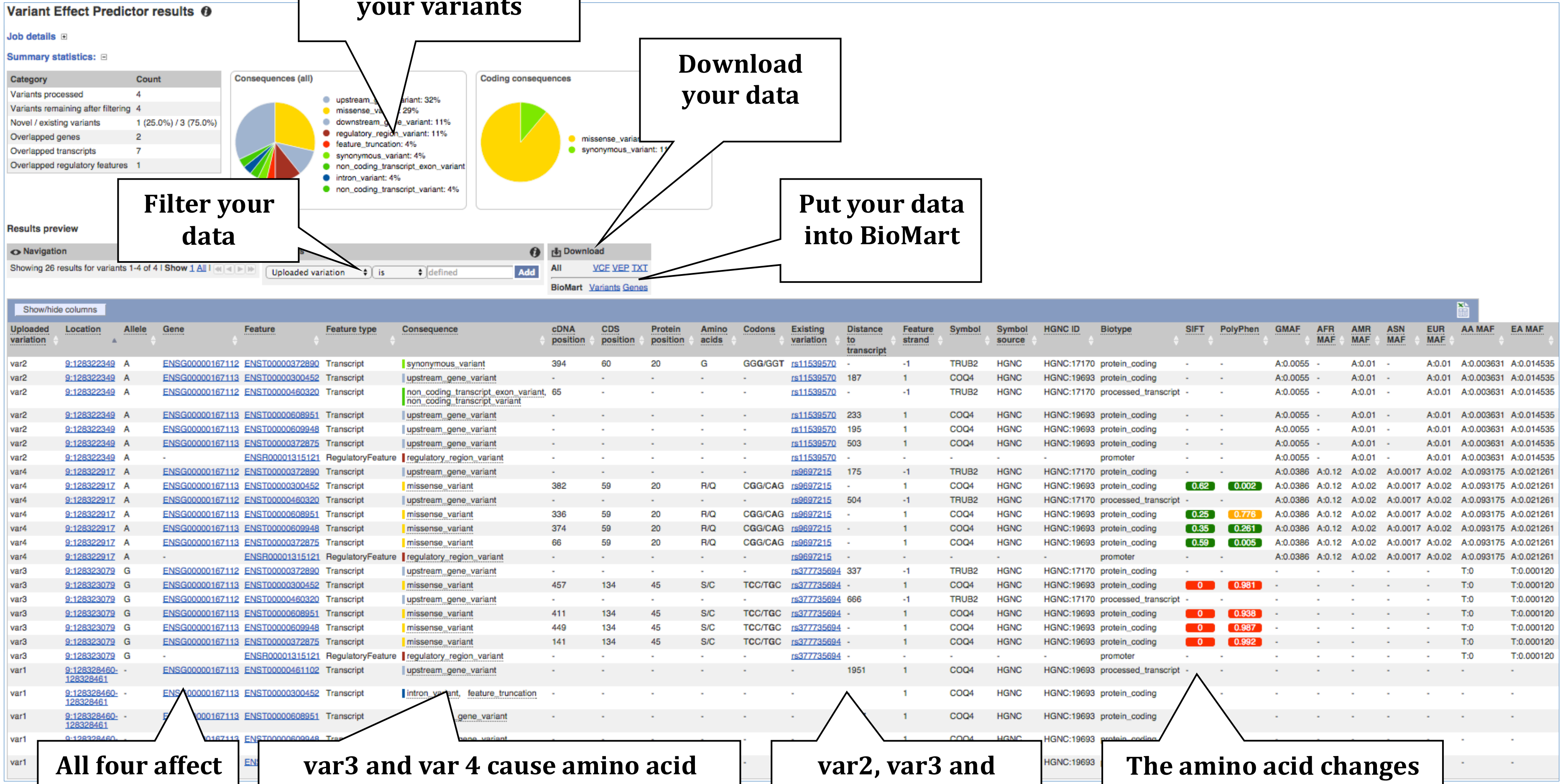
Click to get your results

Buttons to save, edit or delete your job

The display will show you the status of your job. It will say "Queued", then automatically switch to "Done" when the job is done, you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

Click "[View results]" once your job is done.

In your results you will see a graphical summary of your data, as well as a table of your results. (Note that some empty columns in the results table have been hidden in the following screenshot to save space.)



Summary of the consequences of all your variants

Download your data

Filter your data

Put your data into BioMart

Variant Effect Predictor results

Job details

Summary statistics:

Category	Count
Variants processed	4
Variants remaining after filtering	4
Novel / existing variants	1 (25.0%) / 3 (75.0%)
Overlapped genes	2
Overlapped transcripts	7
Overlapped regulatory features	1

Consequences (all)

- upstream_g ariant: 32%
- missense_v 29%
- downstream_g e_variant: 11%
- regulatory_region_variant: 11%
- feature_truncation: 4%
- synonymous_variant: 4%
- non_coding_transcript_exon_variant
- intron_variant: 4%
- non_coding_transcript_variant: 4%

Coding consequences

- missense_varia
- synonymous_variant: 1

Results preview

Navigation

Showing 26 results for variants 1-4 of 4 | Show 1 All |

Uploaded variation | is | defined | Add

Download

All VCF VEP TXT

BioMart Variants Genes

Show/hide columns

Uploaded variation	Location	Allele	Gene	Feature	Feature type	Consequence	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variation	Distance to transcript	Feature strand	Symbol	Symbol source	HGNC ID	Biotype	SIFT	PolyPhen	GMAF	AFR MAF	AMR MAF	ASN MAF	EUR MAF	AA MAF	EA MAF
var2	9:128322349	A	ENSG00000167112	ENST00000372890	Transcript	synonymous_variant	394	60	20	G	GGG/GGT	rs11539570	-	-1	TRUB2	HGNC	HGNC:17170	protein_coding	-	-	A:0.0055	-	A:0.01	-	A:0.01	A:0.003631	A:0.014535
var2	9:128322349	A	ENSG00000167113	ENST00000300452	Transcript	upstream_gene_variant	-	-	-	-	-	rs11539570	187	1	COQ4	HGNC	HGNC:19693	protein_coding	-	-	A:0.0055	-	A:0.01	-	A:0.01	A:0.003631	A:0.014535
var2	9:128322349	A	ENSG00000167112	ENST00000460320	Transcript	non_coding_transcript_exon_variant, non_coding_transcript_variant	65	-	-	-	-	rs11539570	-	-1	TRUB2	HGNC	HGNC:17170	processed_transcript	-	-	A:0.0055	-	A:0.01	-	A:0.01	A:0.003631	A:0.014535
var2	9:128322349	A	ENSG00000167113	ENST00000608951	Transcript	upstream_gene_variant	-	-	-	-	-	rs11539570	233	1	COQ4	HGNC	HGNC:19693	protein_coding	-	-	A:0.0055	-	A:0.01	-	A:0.01	A:0.003631	A:0.014535
var2	9:128322349	A	ENSG00000167113	ENST00000609948	Transcript	upstream_gene_variant	-	-	-	-	-	rs11539570	195	1	COQ4	HGNC	HGNC:19693	protein_coding	-	-	A:0.0055	-	A:0.01	-	A:0.01	A:0.003631	A:0.014535
var2	9:128322349	A	ENSG00000167113	ENST00000372875	Transcript	upstream_gene_variant	-	-	-	-	-	rs11539570	503	1	COQ4	HGNC	HGNC:19693	protein_coding	-	-	A:0.0055	-	A:0.01	-	A:0.01	A:0.003631	A:0.014535
var2	9:128322349	A	-	ENSR00001315121	RegulatoryFeature	regulatory_region_variant	-	-	-	-	-	rs11539570	-	-	-	-	-	promoter	-	-	A:0.0055	-	A:0.01	-	A:0.01	A:0.003631	A:0.014535
var4	9:128322917	A	ENSG00000167112	ENST00000372890	Transcript	upstream_gene_variant	-	-	-	-	-	rs9697215	175	-1	TRUB2	HGNC	HGNC:17170	protein_coding	-	-	A:0.0386	A:0.12	A:0.02	A:0.0017	A:0.02	A:0.093175	A:0.021261
var4	9:128322917	A	ENSG00000167113	ENST00000300452	Transcript	missense_variant	382	59	20	R/Q	CGG/CAG	rs9697215	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0.62	0.002	A:0.0386	A:0.12	A:0.02	A:0.0017	A:0.02	A:0.093175	A:0.021261
var4	9:128322917	A	ENSG00000167112	ENST00000460320	Transcript	upstream_gene_variant	-	-	-	-	-	rs9697215	504	-1	TRUB2	HGNC	HGNC:17170	processed_transcript	-	-	A:0.0386	A:0.12	A:0.02	A:0.0017	A:0.02	A:0.093175	A:0.021261
var4	9:128322917	A	ENSG00000167113	ENST00000608951	Transcript	missense_variant	336	59	20	R/Q	CGG/CAG	rs9697215	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0.25	0.776	A:0.0386	A:0.12	A:0.02	A:0.0017	A:0.02	A:0.093175	A:0.021261
var4	9:128322917	A	ENSG00000167113	ENST00000609948	Transcript	missense_variant	374	59	20	R/Q	CGG/CAG	rs9697215	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0.35	0.261	A:0.0386	A:0.12	A:0.02	A:0.0017	A:0.02	A:0.093175	A:0.021261
var4	9:128322917	A	ENSG00000167113	ENST00000372875	Transcript	missense_variant	66	59	20	R/Q	CGG/CAG	rs9697215	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0.59	0.005	A:0.0386	A:0.12	A:0.02	A:0.0017	A:0.02	A:0.093175	A:0.021261
var4	9:128322917	A	-	ENSR00001315121	RegulatoryFeature	regulatory_region_variant	-	-	-	-	-	rs9697215	-	-	-	-	-	promoter	-	-	A:0.0386	A:0.12	A:0.02	A:0.0017	A:0.02	A:0.093175	A:0.021261
var3	9:128323079	G	ENSG00000167112	ENST00000372890	Transcript	upstream_gene_variant	-	-	-	-	-	rs377735694	337	-1	TRUB2	HGNC	HGNC:17170	protein_coding	-	-	-	-	-	-	-	T:0	T:0.000120
var3	9:128323079	G	ENSG00000167113	ENST00000300452	Transcript	missense_variant	457	134	45	S/C	TCC/TGC	rs377735694	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0	0.981	-	-	-	-	-	T:0	T:0.000120
var3	9:128323079	G	ENSG00000167112	ENST00000460320	Transcript	upstream_gene_variant	-	-	-	-	-	rs377735694	666	-1	TRUB2	HGNC	HGNC:17170	processed_transcript	-	-	-	-	-	-	-	T:0	T:0.000120
var3	9:128323079	G	ENSG00000167113	ENST00000608951	Transcript	missense_variant	411	134	45	S/C	TCC/TGC	rs377735694	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0	0.938	-	-	-	-	-	T:0	T:0.000120
var3	9:128323079	G	ENSG00000167113	ENST00000609948	Transcript	missense_variant	449	134	45	S/C	TCC/TGC	rs377735694	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0	0.987	-	-	-	-	-	T:0	T:0.000120
var3	9:128323079	G	ENSG00000167113	ENST00000372875	Transcript	missense_variant	141	134	45	S/C	TCC/TGC	rs377735694	-	1	COQ4	HGNC	HGNC:19693	protein_coding	0	0.992	-	-	-	-	-	T:0	T:0.000120
var3	9:128323079	G	-	ENSR00001315121	RegulatoryFeature	regulatory_region_variant	-	-	-	-	-	rs377735694	-	-	-	-	-	promoter	-	-	-	-	-	-	-	T:0	T:0.000120
var1	9:128328460-128328461	-	ENSG00000167113	ENST00000461102	Transcript	upstream_gene_variant	-	-	-	-	-	-	1951	1	COQ4	HGNC	HGNC:19693	processed_transcript	-	-	-	-	-	-	-	-	-
var1	9:128328460-128328461	-	ENS 00000167113	ENST00000300452	Transcript	intron_v ant, feature_truncation	-	-	-	-	-	-	[illegible]	1	COQ4	HGNC	HGNC:19693	protein_coding	[illegible]	-	-	-	-	-	-	-	-
var1	9:128328460-128328461	-	E 000167113	ENST00000608951	Transcript	gene_variant	-	-	-	-	-	-	[illegible]	1	COQ4	HGNC	HGNC:19693	protein_coding	[illegible]	-	-	-	-	-	-	-	-
var1	9:128328460-	-	0167113	ENST00000609948	Tra	gene_variant	-	-	-	-	-	-	[illegible]	1	COQ4	HGNC	HGNC:19693	protein_coding	[illegible]	-	-	-	-	-	-	-	-
var1	[illegible]	[illegible]	EN	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-	-	[illegible]	[illegible]	[illegible]	[illegible]	HGNC:19693	[illegible]	[illegible]	-	-	-	-	-	-	-	-

All four affect multiple transcripts of two genes.

var3 and var 4 cause amino acid sequence changes. var2 has a synonymous change in sequence, and var1 falls intronic/ downstream. var2, var3 and var4 all fall within a promoter.

var2, var3 and var4 already exist in our database: click to find out more about them.

The amino acid changes caused by var3 damage the protein product, whereas those caused by var4 do not.

Note that the UCSC Genome Browser has a similar tool, named the Variation Annotation Integrator (http://genome.ucsc.edu/cgi-bin/hgVai), that offers largely the same functionality as the Ensembl Variant Effect Predictor.

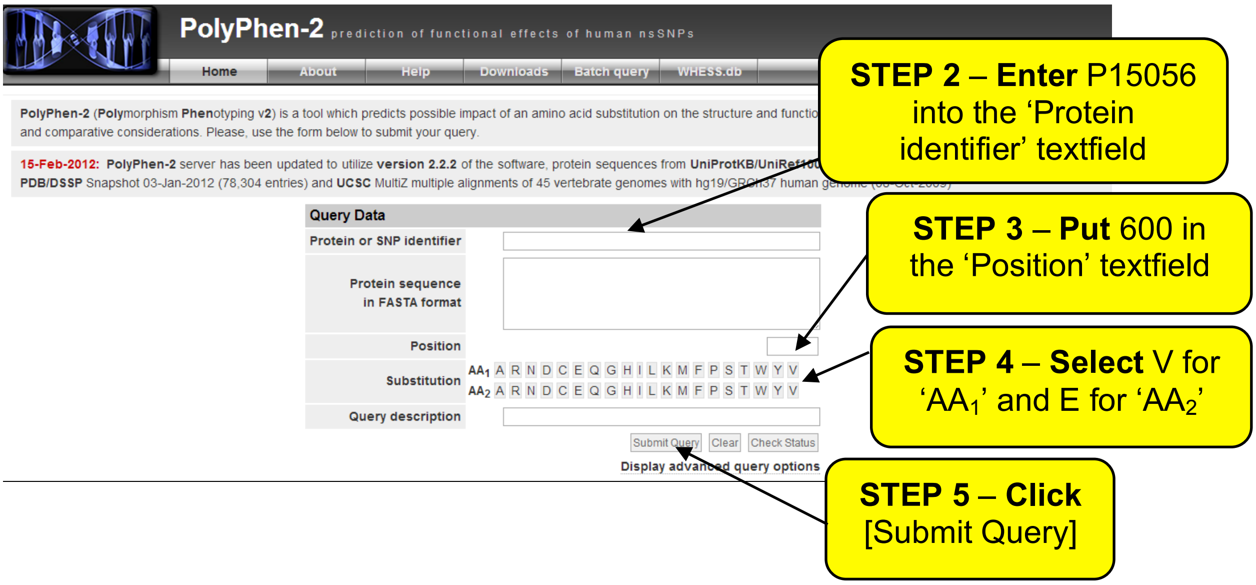
PolyPhen-2

PolyPhen-2 (Polymorphism Phenotyping v2) is a tool that predicts the possible impact of an amino acid substitution on the structure and function of a human protein using physical and comparative considerations.

Worked example 2: PolyPhen-2

In this worked example we will have a look at a variant of the human *BRAF* (B-Raf proto-oncogene, serine/threonine kinase) gene. The most common variant in the BRAF protein is a V to E change at amino acid position 600. Let's assume we did not know the consequence of this variant and use PolyPhen-2 to see if it is deleterious. The UniProt identifier for the BRAF protein is P15056 (http://www.uniprot.org/uniprot/P15056).

STEP 1 – Go to the PolyPhen homepage: http://genetics.bwh.harvard.edu/pph2



PolyPhen-2 (**Poly**morphism **Phen**otyping **v2**) is a tool which predicts possible impact of an amino acid substitution on the structure and functio
and comparative considerations. Please, use the form below to submit your query.

15-Feb-2012: PolyPhen-2 server has been updated to utilize **version 2.2.2** of the software, protein sequences from **UniProtKB/UniRef100**
PDB/DSSP Snapshot 03-Jan-2012 (78,304 entries) and **UCSC** MultiZ multiple alignments of 45 vertebrate genomes with hg19/GRCh37 human genome (05-Oct-2009)

Query Data

Protein or SNP identifier

Protein sequence in FASTA format

Position

Substitution AA_1 A R N D C E Q G H I L K M F P S T W Y V
AA_2 A R N D C E Q G H I L K M F P S T W Y V

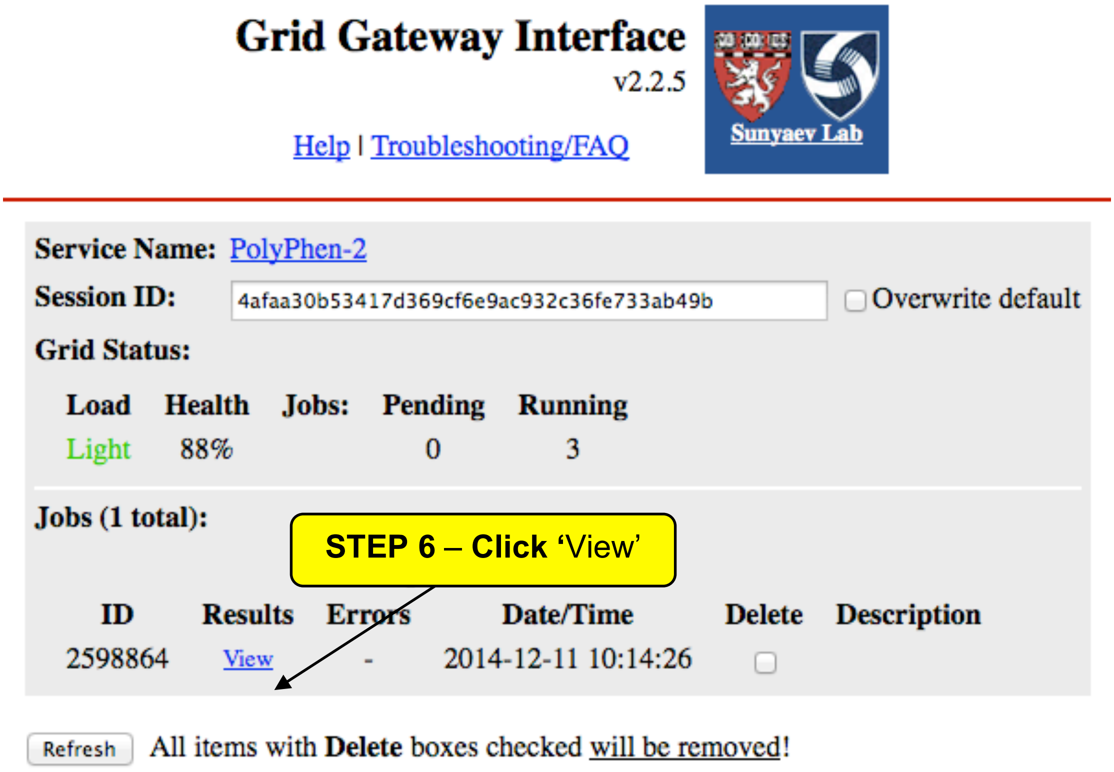
Query description

STEP 2 – Enter P15056 into the 'Protein identifier' textfield

STEP 3 – Put 600 in the 'Position' textfield

STEP 4 – Select V for 'AA_1' and E for 'AA_2'

STEP 5 – Click [Submit Query]



Grid Gateway Interface
v2.2.5

Help | Troubleshooting/FAQ

Service Name: PolyPhen-2

Session ID: 4afaa30b53417d369cf6e9ac932c36fe733ab49b □ Overwrite default

Grid Status:

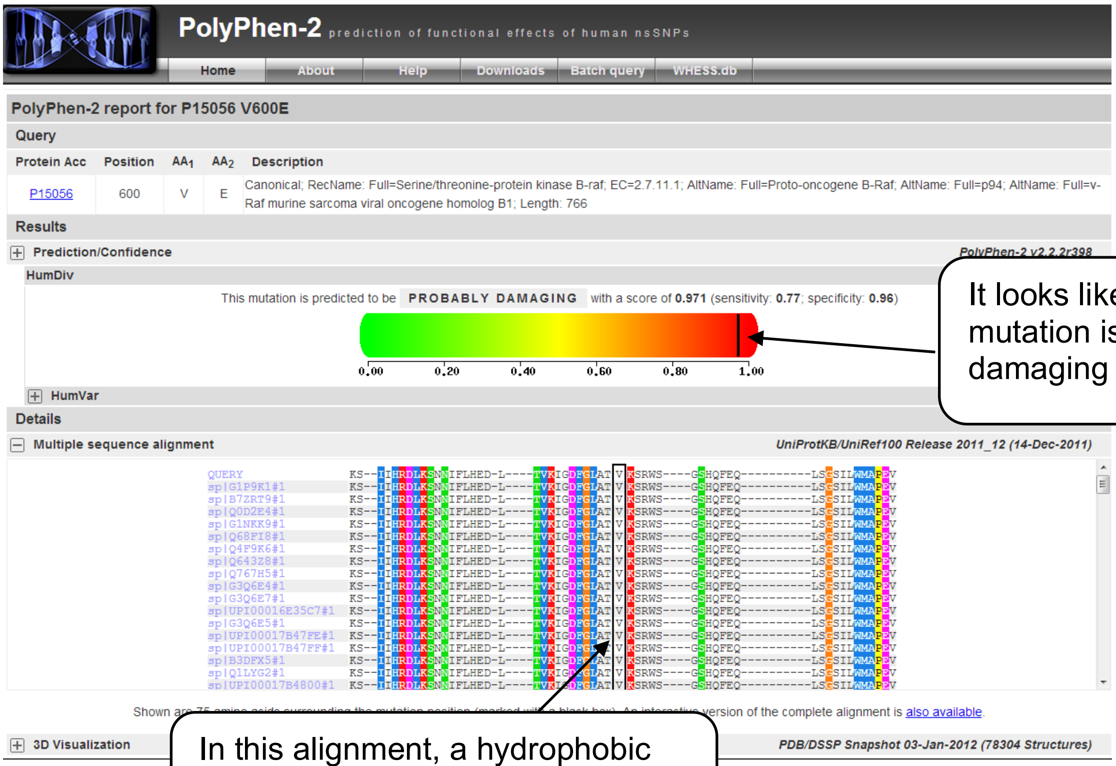
Load	Health	Jobs:	Pending	Running
Light	88%		0	3

Jobs (1 total):

STEP 6 – Click 'View'

ID	Results	Errors	Date/Time	Delete	Description
2598864	View	-	2014-12-11 10:14:26	□	

Refresh All items with **Delete** boxes checked will be removed!



PolyPhen-2 report for P15056 V600E

Query

Protein Acc	Position	AA_1	AA_2	Description
P15056	600	V	E	Canonical; RecName: Full=Serine/threonine-protein kinase B-raf; EC=2.7.11.1; AltName: Full=Proto-oncogene B-Raf; AltName: Full=p94; AltName: Full=v-Raf murine sarcoma viral oncogene homolog B1; Length: 766

Results

Prediction/Confidence — PolyPhen-2 v2.2.2r398

HumDiv

This mutation is predicted to be **PROBABLY DAMAGING** with a score of **0.971** (sensitivity: **0.77**; specificity: **0.96**)

HumVar

It looks like this mutation is probably damaging

Details

Multiple sequence alignment — UniProtKB/UniRef100 Release 2011_12 (14-Dec-2011)

```
QUERY                 KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|G1P9K1#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|B7ZRT9#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|Q0D2E4#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|G1NKK9#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|Q68FI8#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|Q4F9K6#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|Q643Z8#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|Q767H5#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|G3Q6E4#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|G3Q6E7#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|UPI00016E35C7#1    KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|G3Q6E5#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|UPI00017B47FE#1    KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|UPI00017B47FF#1    KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|B3DFX5#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|Q1LYG2#1           KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
sp|UPI00017B4800#1    KS--IIHRDLKSNNIFLHED-L----TVKIGDFGLATVKSRWS----GSHQFEQ---------LSGSILWMAPEV
```

Shown are 75 amino acids surrounding the mutation position (marked with a black box). An interactive version of the complete alignment is also available.

3D Visualization — PDB/DSSP Snapshot 03-Jan-2012 (78304 Structures)

In this alignment, a hydrophobic residue (A, I, L, F, V, P, G) is highly favoured at position 600.

A similar tool to PolyPhen is SIFT (Sorting Intolerant From Tolerant; http://sift.jcvi.org). SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST. SIFT can be applied to naturally occurring nonsynonymous polymorphisms or laboratory-induced missense mutations.

In the following section we will have a look at several resources that focus on the relationship between genotype and phenotype / disease.

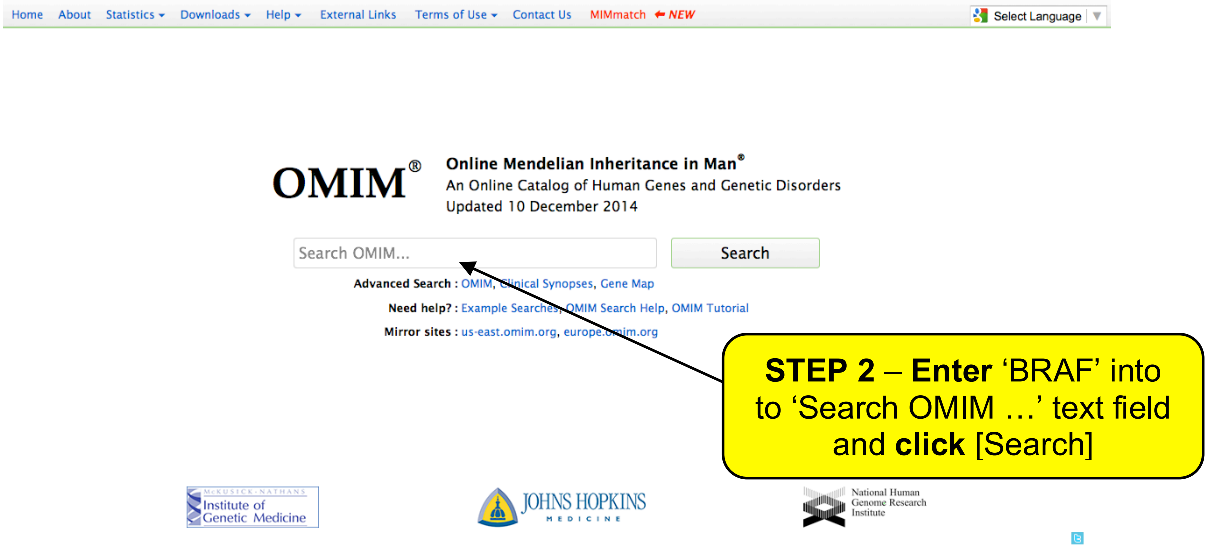
OMIM

The OMIM (Online Mendelian Inheritance in Man; http://omim.org) database is a catalogue of human genes and genetic disorders. OMIM focuses on the relationship between phenotype and genotype.

Worked example 3: OMIM

In this worked example we will have a look what information OMIM contains about the human *BRAF* gene.

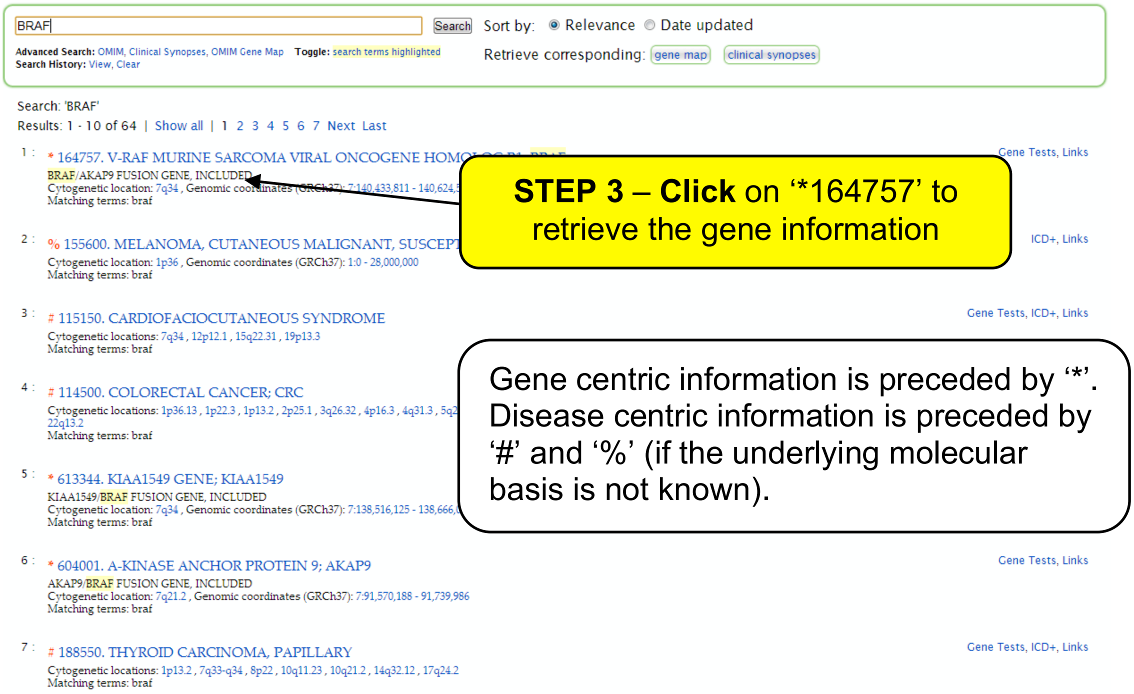
STEP 1 – Go to the OMIM homepage: http://omim.org



Online Mendelian Inheritance in Man®
An Online Catalog of Human Genes and Genetic Disorders
Updated 10 December 2014

Advanced Search : OMIM, Clinical Synopses, Gene Map
Need help? : Example Searches, OMIM Search Help, OMIM Tutorial
Mirror sites : us-east.omim.org, europe.omim.org

STEP 2 – Enter 'BRAF' into to 'Search OMIM …' text field and **click** [Search]



BRAF
Search
Sort by: ● Relevance ○ Date updated
Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map
Toggle: search terms highlighted
Search History: View, Clear
Retrieve corresponding: gene map clinical synopses

Search: 'BRAF'

Results: 1 - 10 of 64 | Show all | 1 2 3 4 5 6 7 Next Last

1 : * 164757. V-RAF MURINE SARCOMA VIRAL ONCOGENE HOMO
BRAF/AKAP9 FUSION GENE, INCLUDED
Cytogenetic location: 7q34 , Genomic coordinates (GRCh37): 7:140,433,811 - 140,624,5
Matching terms: braf
Gene Tests, Links

2 : % 155600. MELANOMA, CUTANEOUS MALIGNANT, SUSCEPT
Cytogenetic location: 1p36 , Genomic coordinates (GRCh37): 1:0 - 28,000,000
Matching terms: braf
ICD+, Links

3 : # 115150. CARDIOFACIOCUTANEOUS SYNDROME
Cytogenetic locations: 7q34 , 12p12.1 , 15q22.31 , 19p13.3
Matching terms: braf
Gene Tests, ICD+, Links

4 : # 114500. COLORECTAL CANCER; CRC
Cytogenetic locations: 1p36.13 , 1p22.3 , 1p13.2 , 2p25.1 , 3q26.32 , 4p16.3 , 4q31.3 , 5q2
22q13.2
Matching terms: braf

5 : * 613344. KIAA1549 GENE; KIAA1549
KIAA1549/BRAF FUSION GENE, INCLUDED
Cytogenetic location: 7q34 , Genomic coordinates (GRCh37): 7:138,516,125 - 138,666,0
Matching terms: braf

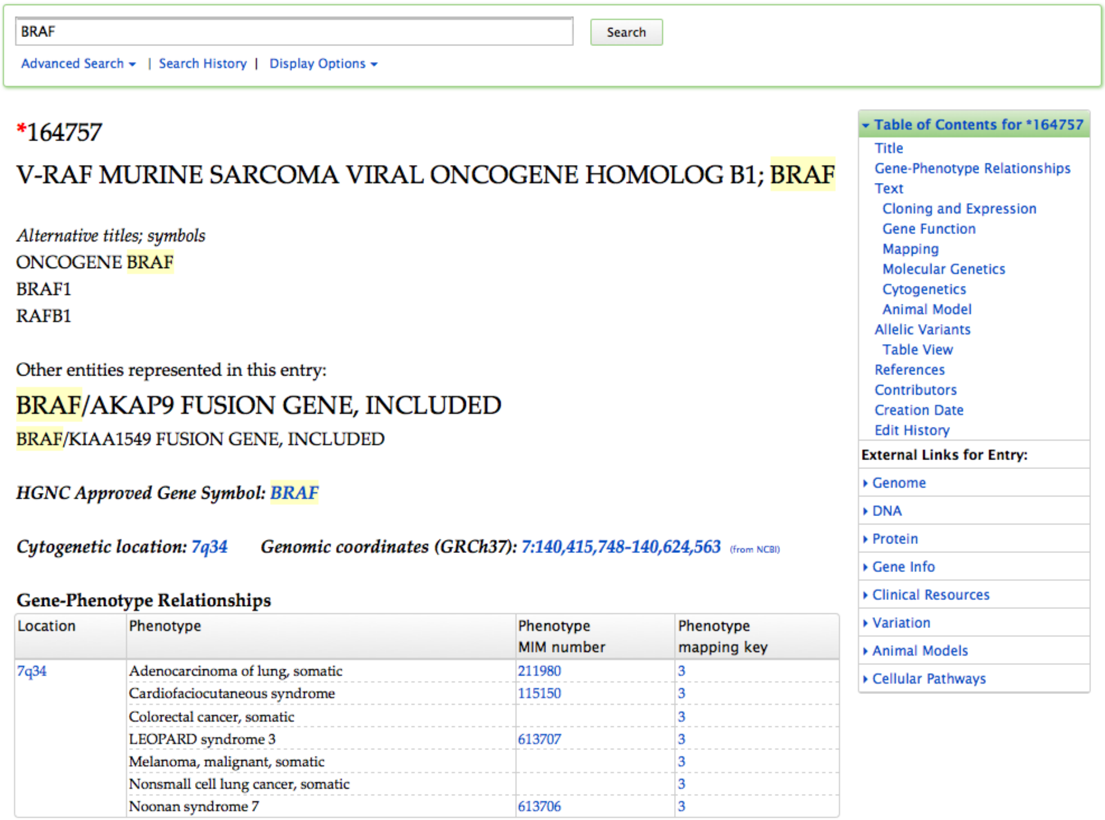
6 : * 604001. A-KINASE ANCHOR PROTEIN 9; AKAP9
AKAP9/BRAF FUSION GENE, INCLUDED
Cytogenetic location: 7q21.2 , Genomic coordinates (GRCh37): 7:91,570,188 - 91,739,986
Matching terms: braf
Gene Tests, Links

7 : # 188550. THYROID CARCINOMA, PAPILLARY
Cytogenetic locations: 1p13.2 , 7q33-q34 , 8p22 , 10q11.23 , 10q21.2 , 14q32.12 , 17q24.2
Matching terms: braf
Gene Tests, ICD+, Links

STEP 3 – Click on '*164757' to retrieve the gene information

Gene centric information is preceded by '*'. Disease centric information is preceded by '#' and '%' (if the underlying molecular basis is not known).

The returned entry is a detailed description of the gene, its location and genetic defects that result in disease. Each piece of clinical and genetic data is cited providing an excellent platform for understanding the function of the gene in a disease setting.



*164757

V-RAF MURINE SARCOMA VIRAL ONCOGENE HOMOLOG B1; BRAF

Alternative titles; symbols

ONCOGENE BRAF
BRAF1
RAFB1

Other entities represented in this entry:

BRAF/AKAP9 FUSION GENE, INCLUDED
BRAF/KIAA1549 FUSION GENE, INCLUDED

HGNC Approved Gene Symbol:* *BRAF

Cytogenetic location:* *7q34 ***Genomic coordinates (GRCh37):* *7:140,415,748-140,624,563*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Phenotype mapping key
7q34	Adenocarcinoma of lung, somatic	211980	3
	Cardiofaciocutaneous syndrome	115150	3
	Colorectal cancer, somatic		3
	LEOPARD syndrome 3	613707	3
	Melanoma, malignant, somatic		3
	Nonsmall cell lung cancer, somatic		3
	Noonan syndrome 7	613706	3

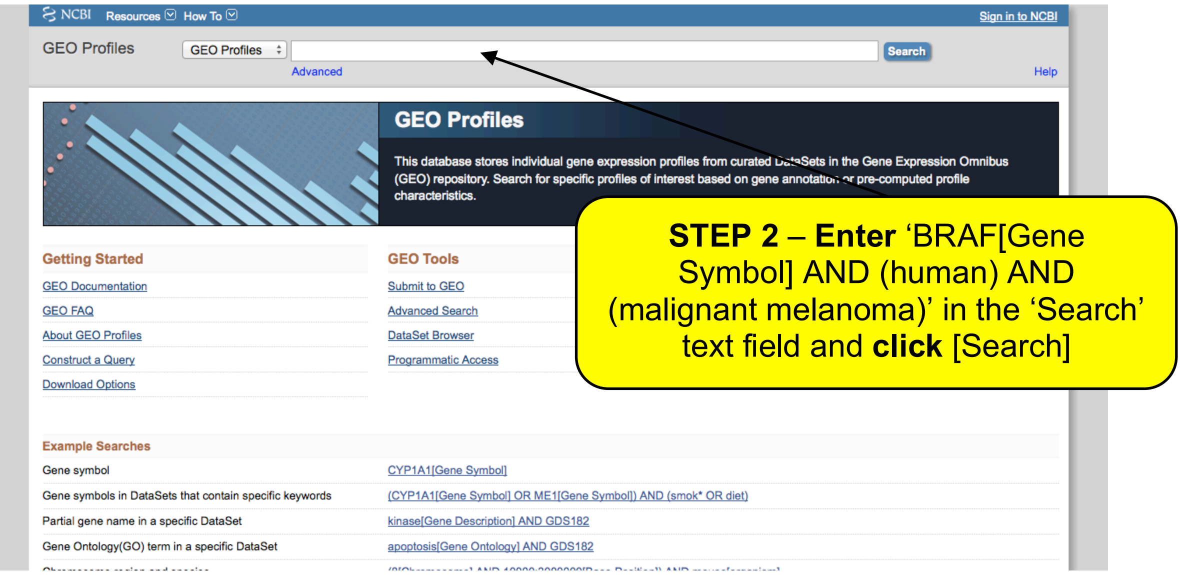
GEO Profiles

The GEO Profiles database (http://www.ncbi.nlm.nih.gov/geoprofiles) stores individual gene expression profiles from curated DataSets in the Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/).

Worked example 4: GEO Profiles

In this worked example we will look whether there are any expression profiles for the human *BRAF* gene for malignant melanoma.

STEP 1 – Go to the GEO Profiles homepage: http://www.ncbi.nlm.nih.gov/geoprofiles



This database stores individual gene expression profiles from curated DataSets in the Gene Expression Omnibus (GEO) repository. Search for specific profiles of interest based on gene annotation or pre-computed profile characteristics.

Getting Started

- GEO Documentation
- GEO FAQ
- About GEO Profiles
- Construct a Query
- Download Options

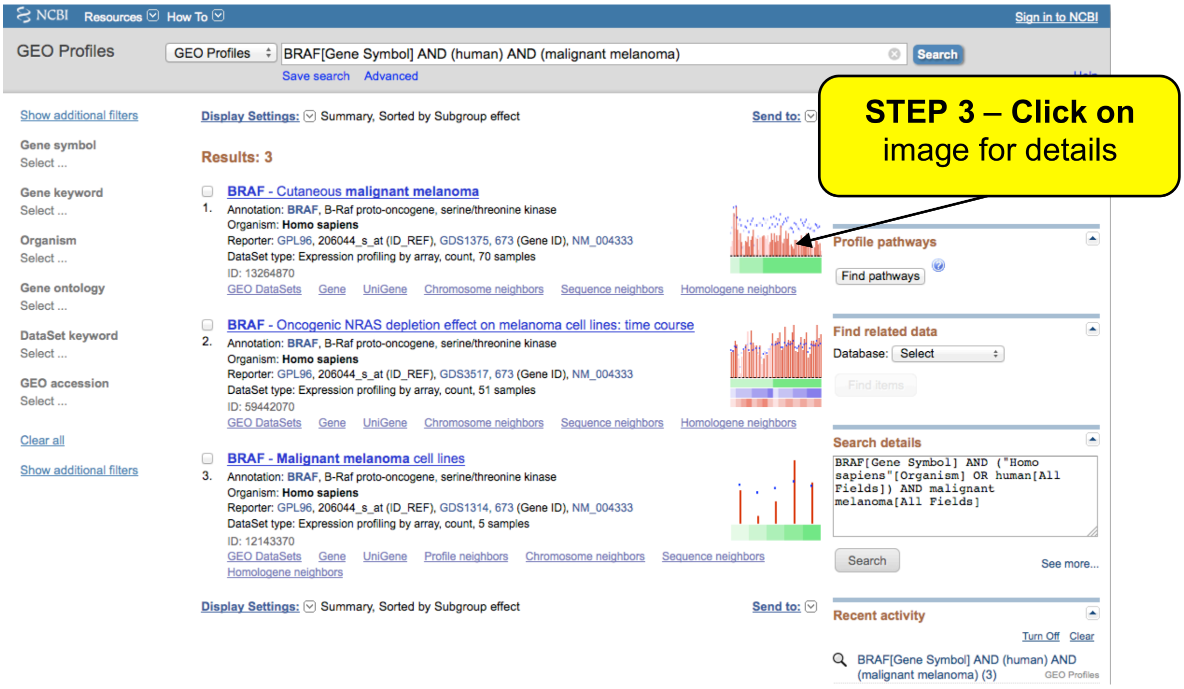
GEO Tools

- Submit to GEO
- Advanced Search
- DataSet Browser
- Programmatic Access

Example Searches

Gene symbol	CYP1A1[Gene Symbol]
Gene symbols in DataSets that contain specific keywords	(CYP1A1[Gene Symbol] OR ME1[Gene Symbol]) AND (smok* OR diet)
Partial gene name in a specific DataSet	kinase[Gene Description] AND GDS182
Gene Ontology(GO) term in a specific DataSet	apoptosis[Gene Ontology] AND GDS182

STEP 2 – Enter ‘BRAF[Gene Symbol] AND (human) AND (malignant melanoma)’ in the ‘Search’ text field and **click** [Search]



BRAF[Gene Symbol] AND (human) AND (malignant melanoma)

Display Settings: Summary, Sorted by Subgroup effect Send to:

Results: 3

□ 1. BRAF - Cutaneous malignant melanoma
Annotation: BRAF, B-Raf proto-oncogene, serine/threonine kinase
Organism: Homo sapiens
Reporter: GPL96, 206044_s_at (ID_REF), GDS1375, 673 (Gene ID), NM_004333
DataSet type: Expression profiling by array, count, 70 samples
ID: 13264870
GEO DataSets Gene UniGene Chromosome neighbors Sequence neighbors Homologene neighbors

□ 2. BRAF - Oncogenic NRAS depletion effect on melanoma cell lines: time course
Annotation: BRAF, B-Raf proto-oncogene, serine/threonine kinase
Organism: Homo sapiens
Reporter: GPL96, 206044_s_at (ID_REF), GDS3517, 673 (Gene ID), NM_004333
DataSet type: Expression profiling by array, count, 51 samples
ID: 59442070
GEO DataSets Gene UniGene Chromosome neighbors Sequence neighbors Homologene neighbors

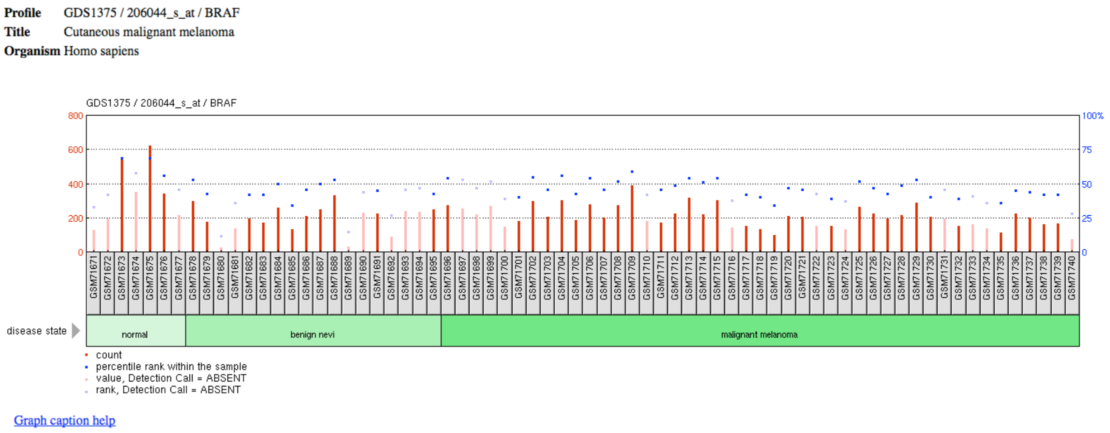
□ 3. BRAF - Malignant melanoma cell lines
Annotation: BRAF, B-Raf proto-oncogene, serine/threonine kinase
Organism: Homo sapiens
Reporter: GPL96, 206044_s_at (ID_REF), GDS1314, 673 (Gene ID), NM_004333
DataSet type: Expression profiling by array, count, 5 samples
ID: 12143370
GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Sequence neighbors Homologene neighbors

Display Settings: Summary, Sorted by Subgroup effect Send to:

```
BRAF[Gene Symbol] AND ("Homo
sapiens"[Organism] OR human[All
Fields]) AND malignant
melanoma[All Fields]
```

BRAF[Gene Symbol] AND (human) AND (malignant melanoma) (3) GEO Profiles

STEP 3 – Click on image for details



Profile GDS1375 / 206044_s_at / BRAF
Title Cutaneous malignant melanoma
Organism Homo sapiens

Graph caption help

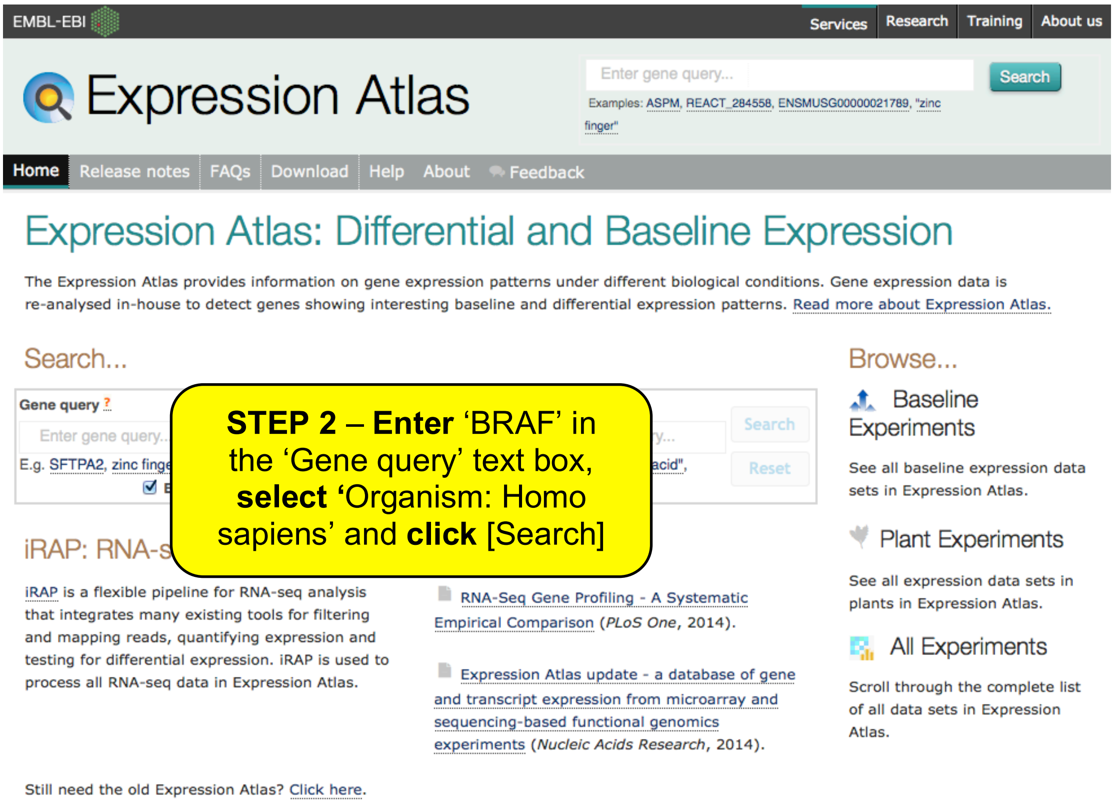
Expression Atlas

Expression Atlas (http://www.ebi.ac.uk/gxa/home) provides information on gene expression patterns under different biological conditions. It includes both microarray and RNA-seq data. The data is re-analysed to detect interesting expression patterns under the conditions of the original experiment. There are two components to the Expression Atlas, i.e. the Baseline Atlas and the Differential Atlas. The Baseline Atlas displays information about which gene products are present (and at what abundance) in "normal" conditions (e.g. tissue, cell type). The Differential Atlas allows users to identify genes that are up- or down-regulated in a wide variety of different experimental conditions.

Worked example 5: Expression Atlas

In this worked example we will have a look what information Expression Atlas contains for the human *BRAF* gene.

STEP 1 – Go to the Expression Atlas homepage: http://www.ebi.ac.uk/gxa/home



Examples: ASPM, REACT_284558, ENSMUSG00000021789, "zinc finger"

Expression Atlas: Differential and Baseline Expression

The Expression Atlas provides information on gene expression patterns under different biological conditions. Gene expression data is re-analysed in-house to detect genes showing interesting baseline and differential expression patterns. Read more about Expression Atlas.

E.g. SFTPA2, zinc finge...

STEP 2 – Enter 'BRAF' in the 'Gene query' text box, **select** 'Organism: Homo sapiens' and **click** [Search]

iRAP: RNA-s...

iRAP is a flexible pipeline for RNA-seq analysis that integrates many existing tools for filtering and mapping reads, quantifying expression and testing for differential expression. iRAP is used to process all RNA-seq data in Expression Atlas.

RNA-Seq Gene Profiling - A Systematic Empirical Comparison (*PLoS One*, 2014).

Expression Atlas update - a database of gene and transcript expression from microarray and sequencing-based functional genomics experiments (*Nucleic Acids Research*, 2014).

Baseline Experiments

See all baseline expression data sets in Expression Atlas.

Plant Experiments

See all expression data sets in plants in Expression Atlas.

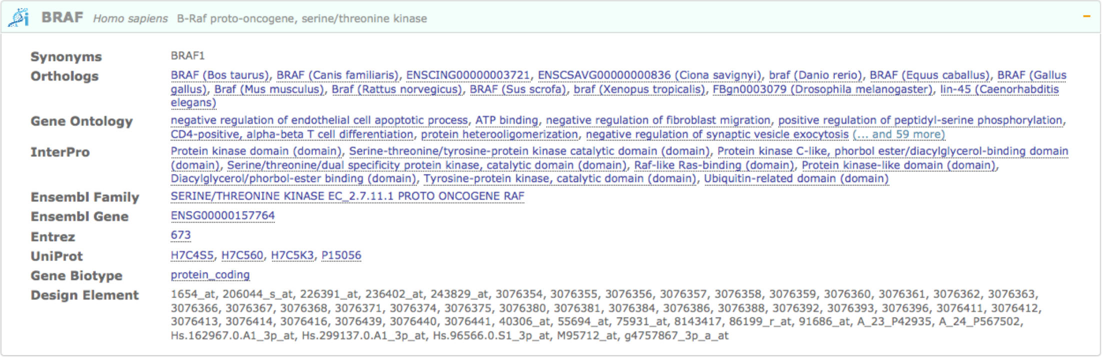
All Experiments

Scroll through the complete list of all data sets in Expression Atlas.

Still need the old Expression Atlas? Click here.

The resulting page consists of three parts.

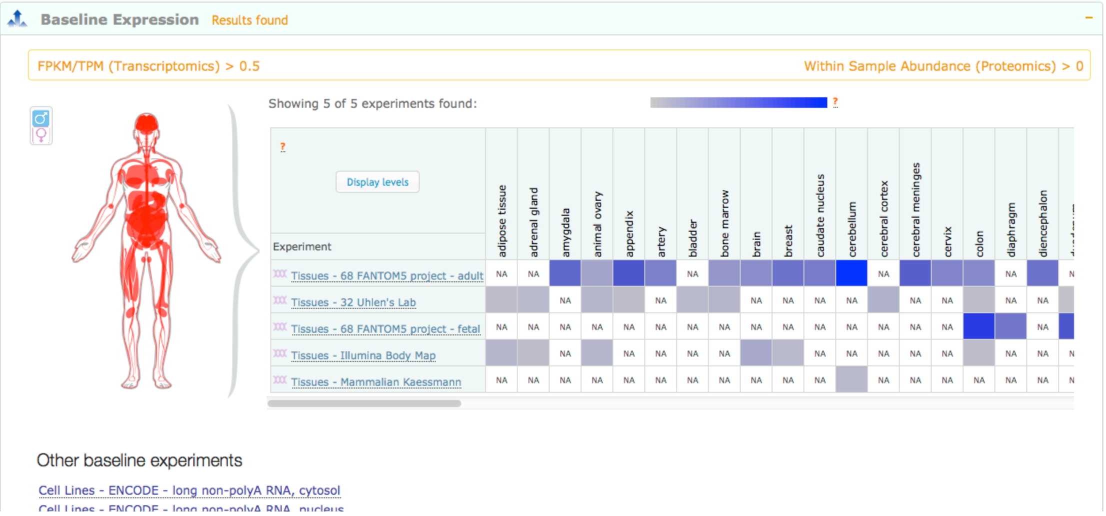
The first part contains general information about the *BRAF* gene:



BRAF *Homo sapiens* B-Raf proto-oncogene, serine/threonine kinase

Synonyms	BRAF1
Orthologs	BRAF (Bos taurus), BRAF (Canis familiaris), ENSCING00000003721, ENSCSAVG00000000836 (Ciona savignyi), braf (Danio rerio), BRAF (Equus caballus), BRAF (Gallus gallus), Braf (Mus musculus), Braf (Rattus norvegicus), BRAF (Sus scrofa), braf (Xenopus tropicalis), FBgn0003079 (Drosophila melanogaster), lin-45 (Caenorhabditis elegans)
Gene Ontology	negative regulation of endothelial cell apoptotic process, ATP binding, negative regulation of fibroblast migration, positive regulation of peptidyl-serine phosphorylation, CD4-positive, alpha-beta T cell differentiation, protein heterooligomerization, negative regulation of synaptic vesicle exocytosis (... and 59 more)
InterPro	Protein kinase domain (domain), Serine-threonine/tyrosine-protein kinase catalytic domain (domain), Protein kinase C-like, phorbol ester/diacylglycerol-binding domain (domain), Serine/threonine/dual specificity protein kinase, catalytic domain (domain), Raf-like Ras-binding (domain), Protein kinase-like domain (domain), Diacylglycerol/phorbol-ester binding (domain), Tyrosine-protein kinase, catalytic domain (domain), Ubiquitin-related domain (domain)
Ensembl Family	SERINE/THREONINE KINASE EC_2.7.11.1 PROTO ONCOGENE RAF
Ensembl Gene	ENSG00000157764
Entrez	673
UniProt	H7C4S5, H7C560, H7C5K3, P15056
Gene Biotype	protein_coding
Design Element	1654_at, 206044_s_at, 226391_at, 236402_at, 243829_at, 3076354, 3076355, 3076356, 3076357, 3076358, 3076359, 3076360, 3076361, 3076362, 3076363, 3076366, 3076367, 3076368, 3076371, 3076374, 3076375, 3076380, 3076381, 3076384, 3076386, 3076388, 3076392, 3076393, 3076396, 3076411, 3076412, 3076413, 3076414, 3076416, 3076439, 3076440, 3076441, 40306_at, 55694_at, 75931_at, 8143417, 86199_r_at, 91686_at, A_23_P42935, A_24_P567502, Hs.162967.0.A1_3p_at, Hs.299137.0.A1_3p_at, Hs.96566.0.S1_3p_at, M95712_at, g4757867_3p_a_at

The second part is the Baseline Expression:



Baseline Expression Results found

FPKM/TPM (Transcriptomics) > 0.5 Within Sample Abundance (Proteomics) > 0

Showing 5 of 5 experiments found:

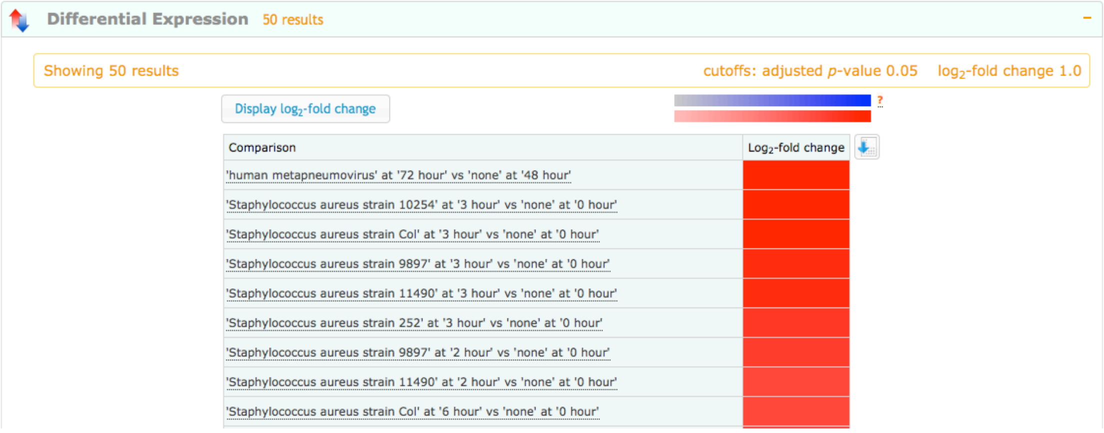
Experiment	adipose tissue	adrenal gland	amygdala	animal ovary	appendix	artery	bladder	bone marrow	brain	breast	caudate nucleus	cerebellum	cerebral cortex	cerebral meninges	cervix	colon	diaphragm	diencephalon
Tissues - 68 FANTOM5 project - adult	NA	NA	■	■	■	■	NA	■	■	■	■	■	NA	■	■	■	NA	■
Tissues - 32 Uhlen's Lab	■	■	NA	■	■	NA	■	■	NA	NA	NA	NA	■	NA	NA	■	NA	NA
Tissues - 68 FANTOM5 project - fetal	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	■	■	NA
Tissues - Illumina Body Map	■	■	NA	■	NA	NA	NA	NA	■	■	NA	NA	NA	NA	NA	■	NA	NA
Tissues - Mammalian Kaessmann	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	■	NA	NA	NA	NA	NA	NA

Other baseline experiments

Cell Lines - ENCODE - long non-polyA RNA, cytosol

Cell Lines - ENCODE - long non-polyA RNA, nucleus

And the third part is the Differential Expression:



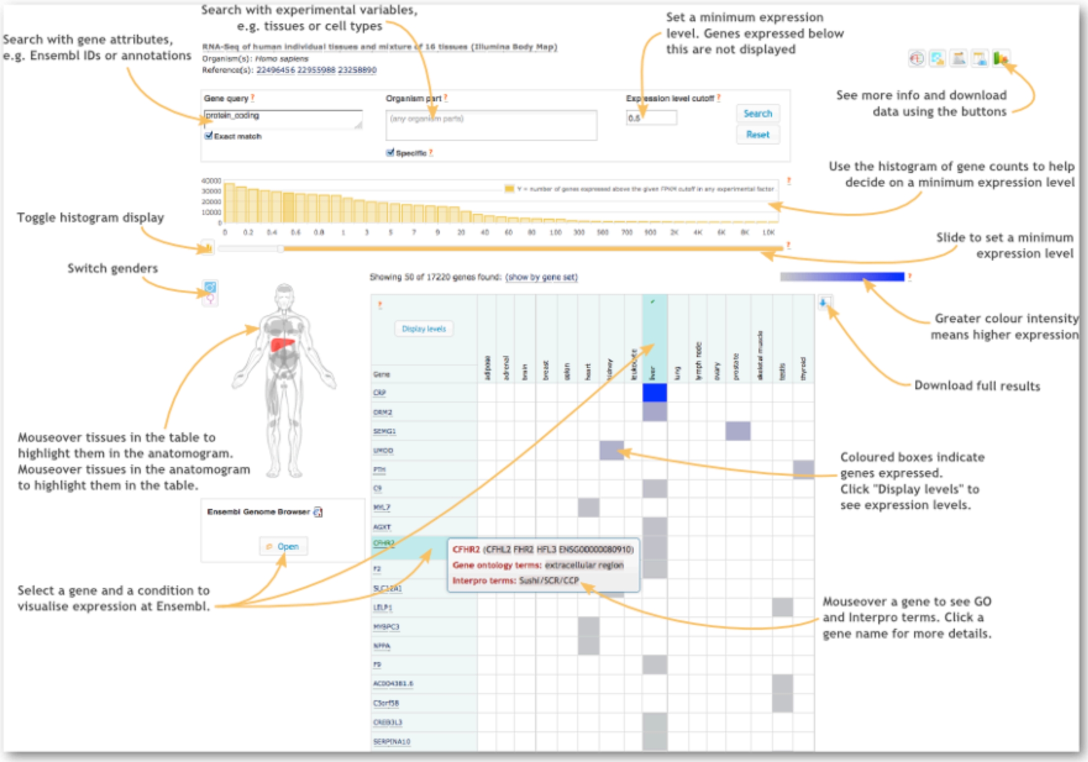
Differential Expression 50 results

Showing 50 results cutoffs: adjusted *p*-value 0.05 $\log_2$-fold change 1.0

Comparison	Log_2-fold change
'human metapneumovirus' at '72 hour' vs 'none' at '48 hour'	■
'Staphylococcus aureus strain 10254' at '3 hour' vs 'none' at '0 hour'	■
'Staphylococcus aureus strain Col' at '3 hour' vs 'none' at '0 hour'	■
'Staphylococcus aureus strain 9897' at '3 hour' vs 'none' at '0 hour'	■
'Staphylococcus aureus strain 11490' at '3 hour' vs 'none' at '0 hour'	■
'Staphylococcus aureus strain 252' at '3 hour' vs 'none' at '0 hour'	■
'Staphylococcus aureus strain 9897' at '2 hour' vs 'none' at '0 hour'	■
'Staphylococcus aureus strain 11490' at '2 hour' vs 'none' at '0 hour'	■
'Staphylococcus aureus strain Col' at '6 hour' vs 'none' at '0 hour'	■

Detailed information about how to work with Expression Atlas, can be found in the Help section:

Baseline Atlas at–a–glance



Search with gene attributes, e.g. Ensembl IDs or annotations

Search with experimental variables, e.g. tissues or cell types

Set a minimum expression level. Genes expressed below this are not displayed

See more info and download data using the buttons

Use the histogram of gene counts to help decide on a minimum expression level

Toggle histogram display

Slide to set a minimum expression level

Switch genders

Greater colour intensity means higher expression

Download full results

Mouseover tissues in the table to highlight them in the anatomogram. Mouseover tissues in the anatomogram to highlight them in the table.

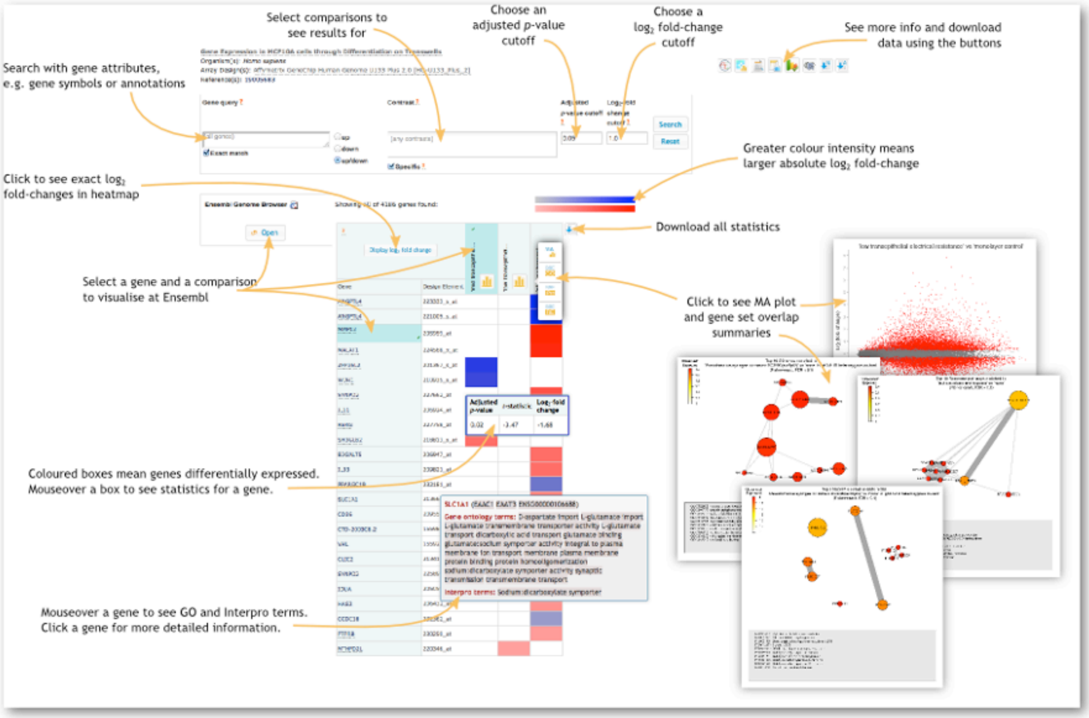
Coloured boxes indicate genes expressed. Click "Display levels" to see expression levels.

Select a gene and a condition to visualise expression at Ensembl.

Mouseover a gene to see GO and Interpro terms. Click a gene name for more details.

RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map)
Organism(s): Homo sapiens
Reference(s): 22496456 22955988 23258890
Gene query
protein_coding
Exact match
Organism part
(any organism parts)
Specific
Expression level cutoff
0.5
Search
Reset
Showing 50 of 17220 genes found: (show by gene set)
Display levels
Gene
adipose
adrenal
brain
breast
colon
heart
kidney
leukocyte
liver
lung
lymph node
ovary
prostate
skeletal muscle
testis
thyroid
CRP
ORM2
SERPINA10
Ensembl Genome Browser
Open
CFHR2 (CFHL2 FHR2 HFL3 ENSG00000080910)
Gene ontology terms: extracellular region
Interpro terms: Sushi/SCR/CCP

Differential Atlas at–a–glance



Select comparisons to see results for

Choose an adjusted p-value cutoff

Choose a $\log_2$ fold-change cutoff

See more info and download data using the buttons

Search with gene attributes, e.g. gene symbols or annotations

Click to see exact $\log_2$ fold-changes in heatmap

Greater colour intensity means larger absolute $\log_2$ fold-change

Download all statistics

Select a gene and a comparison to visualise at Ensembl

Click to see MA plot and gene set overlap summaries

Coloured boxes mean genes differentially expressed. Mouseover a box to see statistics for a gene.

Mouseover a gene to see GO and Interpro terms. Click a gene for more detailed information.

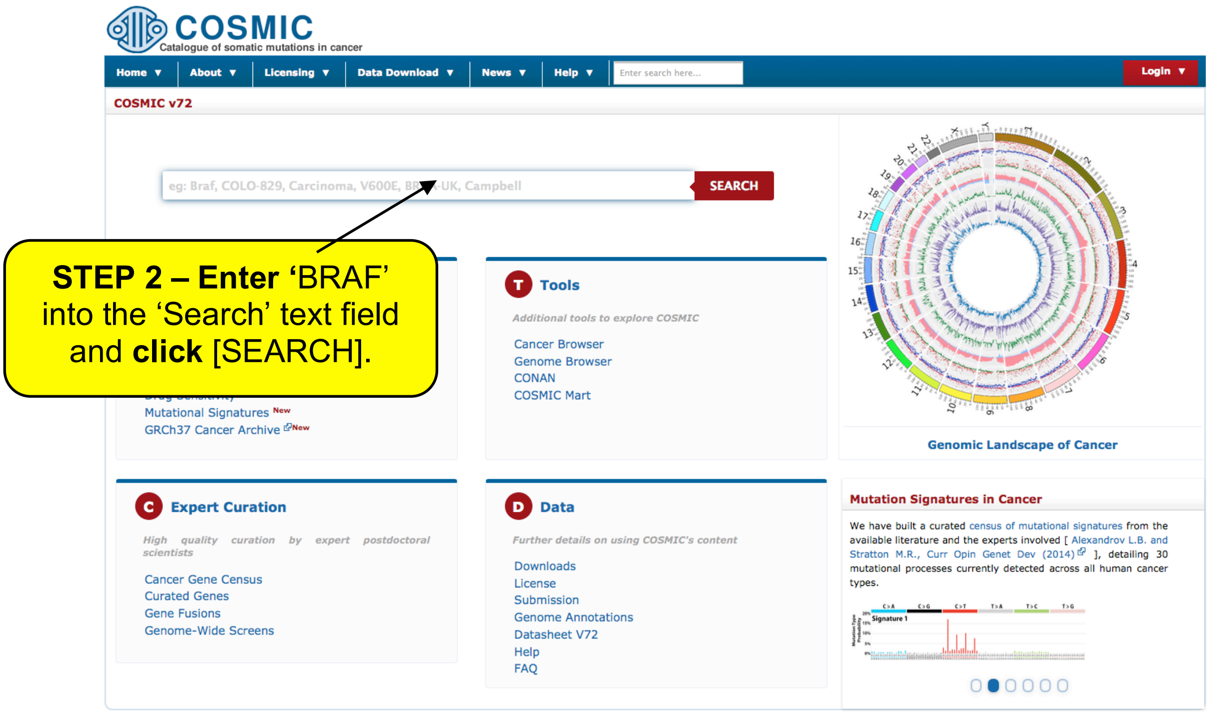
COSMIC

Although OMIM is very detailed, it is not comprehensive. COSMIC, the Catalogue of somatic mutations in cancer (http://cancer.sanger.ac.uk/cosmic/), is a specialist resource that aims to have a comprehensive list of genes and their mutations that are involved in cancer. COSMIC curates data from papers in the scientific literature and large scale experimental screens from the Cancer Genome Project (https://www.sanger.ac.uk/research/projects/cancergenome/) at the Sanger Institute. There are several ways to search COSMIC, in the following worked example the most common search interface will be illustrated.

Worked example 6: COSMIC

In this worked example we will use COSMIC to list all mutations found in the human *BRAF* gene.

STEP 1 – Go to the COSMIC homepage: http://cancer.sanger.ac.uk/cosmic


COSMIC
Catalogue of somatic mutations in cancer
Home ▼
About ▼
Licensing ▼
Data Download ▼
News ▼
Help ▼
Enter search here...
Login ▼
COSMIC v72
eg: Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell
SEARCH
Genomic Landscape of Cancer
C>A
C>G
C>T
T>A
T>C
T>G
Signature 1
Mutation Type Probability
20%
15%
10%
5%
0%

STEP 2 – Enter 'BRAF' into the 'Search' text field and **click** [SEARCH].

Mutational Signatures New
GRCh37 Cancer Archive New

Tools

Additional tools to explore COSMIC

Cancer Browser
Genome Browser
CONAN
COSMIC Mart

Expert Curation

High quality curation by expert postdoctoral scientists

Cancer Gene Census
Curated Genes
Gene Fusions
Genome-Wide Screens

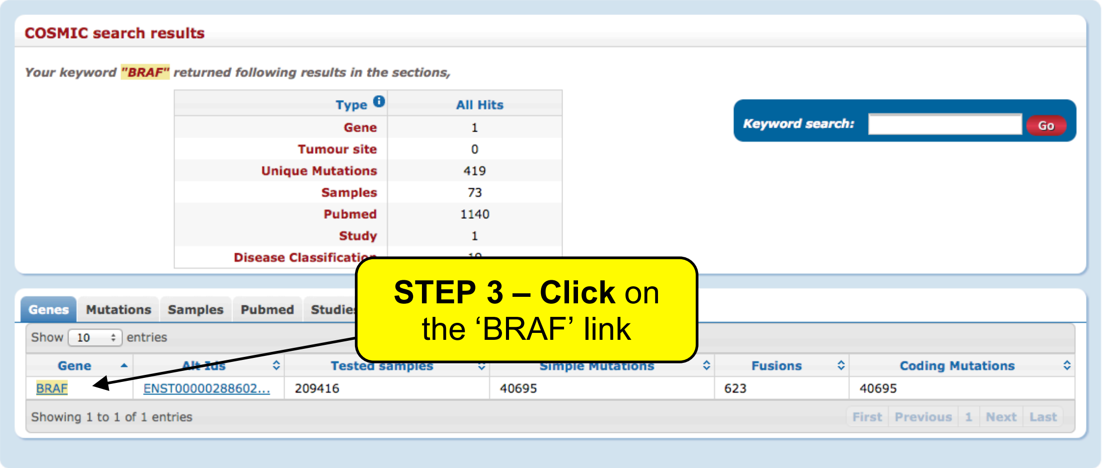
Data

Further details on using COSMIC's content

Downloads
License
Submission
Genome Annotations
Datasheet V72
Help
FAQ

Mutation Signatures in Cancer

We have built a curated census of mutational signatures from the available literature and the experts involved [Alexandrov L.B. and Stratton M.R., Curr Opin Genet Dev (2014)], detailing 30 mutational processes currently detected across all human cancer types.



COSMIC search results

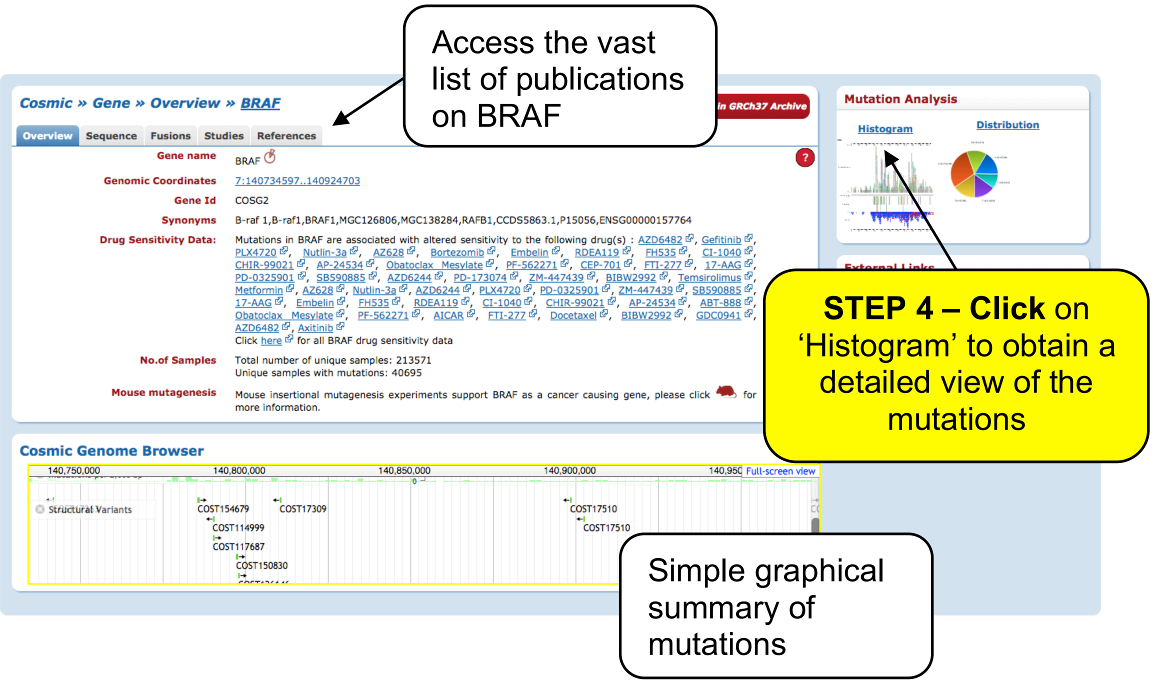
Your keyword "BRAF" returned following results in the sections,

Type	All Hits
Gene	1
Tumour site	0
Unique Mutations	419
Samples	73
Pubmed	1140
Study	1
Disease Classification	10

Gene	Alt Ids	Tested samples	Simple Mutations	Fusions	Coding Mutations
BRAF	ENST00000288602...	209416	40695	623	40695

Showing 1 to 1 of 1 entries

STEP 3 – Click on the 'BRAF' link



Access the vast list of publications on BRAF

Cosmic » Gene » Overview » BRAF

Gene name BRAF

Genomic Coordinates 7:140734597..140924703

Gene Id COSG2

Synonyms B-raf 1,B-raf1,BRAF1,MGC126806,MGC138284,RAFB1,CCDS5863.1,P15056,ENSG00000157764

Drug Sensitivity Data: Mutations in BRAF are associated with altered sensitivity to the following drug(s) : AZD6482, Gefitinib, PLX4720, Nutlin-3a, AZ628, Bortezomib, Embelin, RDEA119, FH535, CI-1040, CHIR-99021, AP-24534, Obatoclax Mesylate, PF-562271, CEP-701, FTI-277, 17-AAG, PD-0325901, SB590885, AZD6244, PD-173074, ZM-447439, BIBW2992, Temsirolimus, Metformin, AZ628, Nutlin-3a, AZD6244, PLX4720, PD-0325901, ZM-447439, SB590885, 17-AAG, Embelin, FH535, RDEA119, CI-1040, CHIR-99021, AP-24534, ABT-888, Obatoclax Mesylate, PF-562271, AICAR, FTI-277, Docetaxel, BIBW2992, GDC0941, AZD6482, Axitinib
Click here for all BRAF drug sensitivity data

No.of Samples Total number of unique samples: 213571
Unique samples with mutations: 40695

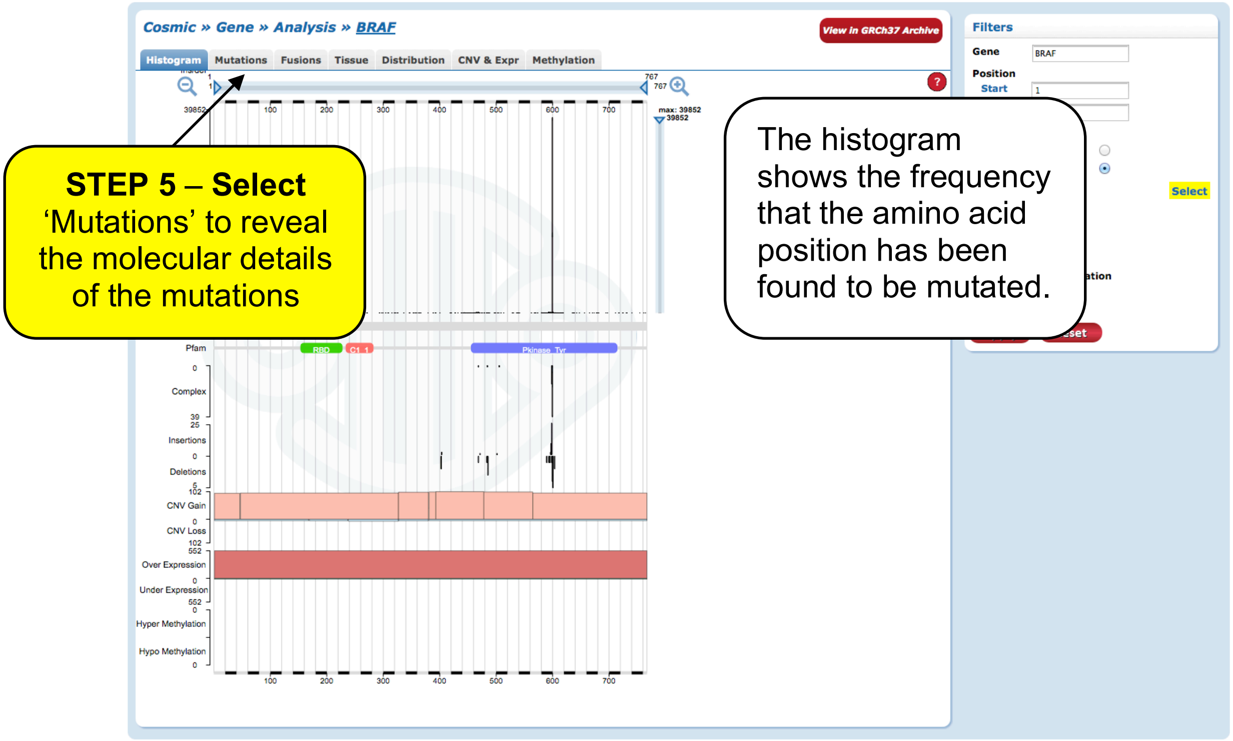
Mouse mutagenesis Mouse insertional mutagenesis experiments support BRAF as a cancer causing gene, please click for more information.

Mutation Analysis

Cosmic Genome Browser

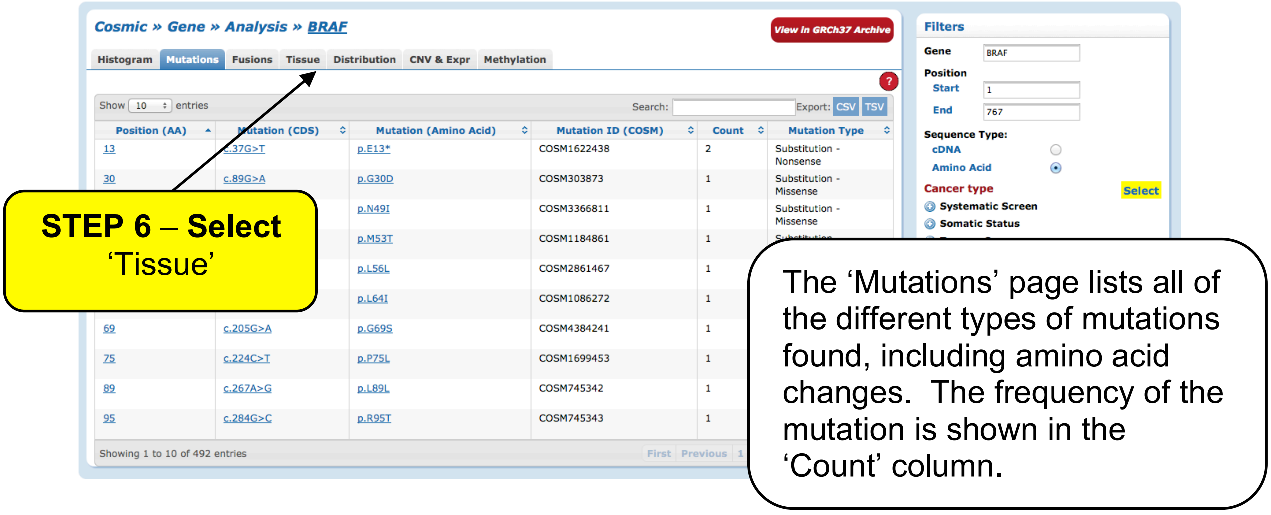
STEP 4 – Click on 'Histogram' to obtain a detailed view of the mutations

Simple graphical summary of mutations



STEP 5 – Select ‘Mutations’ to reveal the molecular details of the mutations

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



Filters

Gene BRAF

Position

Start 1

End 767

Sequence Type:

cDNA ○

Amino Acid ●

Cancer type Select

Systematic Screen

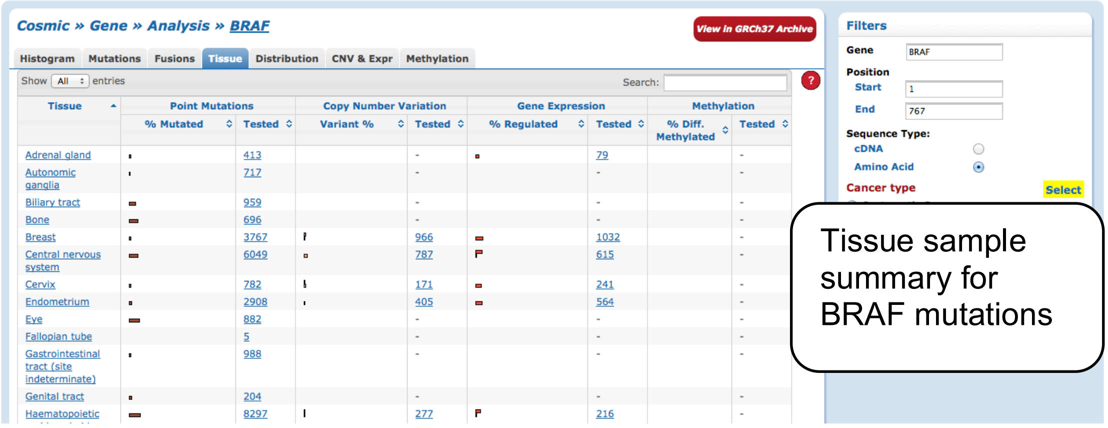
Somatic Status

Position (AA)	Mutation (CDS)	Mutation (Amino Acid)	Mutation ID (COSM)	Count	Mutation Type
13	c.37G>T	p.E13*	COSM1622438	2	Substitution - Nonsense
30	c.89G>A	p.G30D	COSM303873	1	Substitution - Missense
		p.N49I	COSM3366811	1	Substitution - Missense
		p.M53T	COSM1184861	1	Substitution
		p.L56L	COSM2861467	1	
		p.L64I	COSM1086272	1	
69	c.205G>A	p.G69S	COSM4384241	1	
75	c.224C>T	p.P75L	COSM1699453	1	
89	c.267A>G	p.L89L	COSM745342	1	
95	c.284G>C	p.R95T	COSM745343	1	

Showing 1 to 10 of 492 entries

STEP 6 – Select ‘Tissue’

The ‘Mutations’ page lists all of the different types of mutations found, including amino acid changes. The frequency of the mutation is shown in the ‘Count’ column.



Filters

Gene BRAF

Position

Start 1

End 767

Sequence Type:

cDNA ○

Amino Acid ●

Cancer type Select

Tissue	Point Mutations		Copy Number Variation		Gene Expression		Methylation	
	% Mutated	Tested	Variant %	Tested	% Regulated	Tested	% Diff. Methylated	Tested
Adrenal gland	▬	413		-	▬	79		-
Autonomic ganglia	▬	717		-		-		-
Biliary tract	▬	959		-		-		-
Bone	▬	696		-		-		-
Breast	▬	3767	▬	966	▬	1032		-
Central nervous system	▬	6049	▬	787	▬	615		-
Cervix	▬	782	▬	171	▬	241		-
Endometrium	▬	2908	▬	405	▬	564		-
Eye	▬	882		-		-		-
Fallopian tube		5		-		-		-
Gastrointestinal tract (site indeterminate)	▬	988		-		-		-
Genital tract	▬	204		-		-		-
Haematopoietic	▬	8297	▬	277	▬	216		-

Tissue sample summary for BRAF mutations

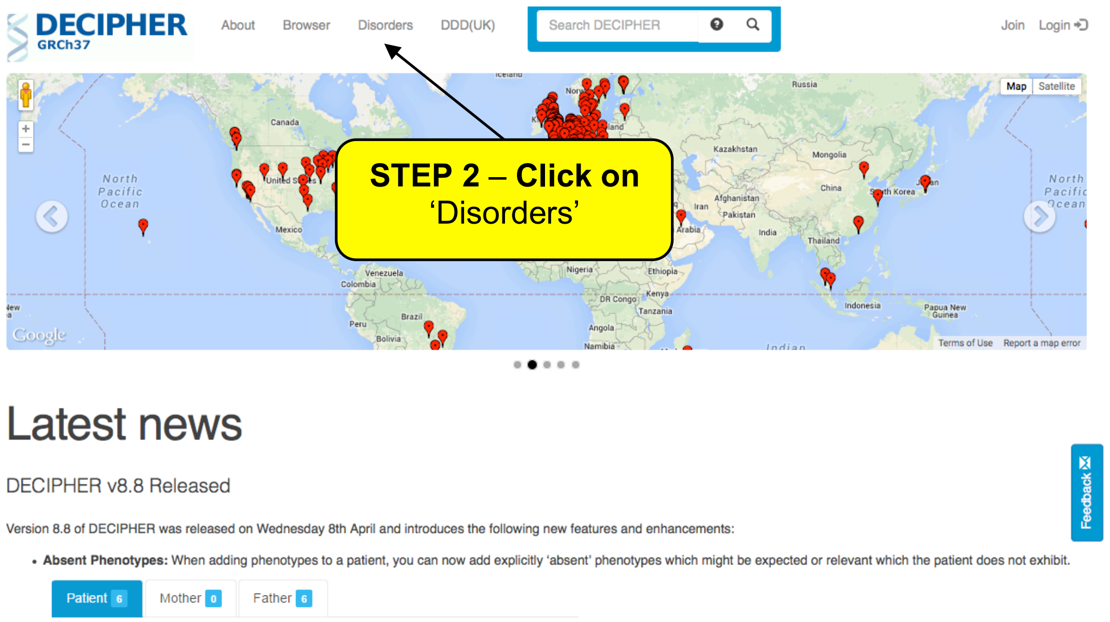
DECIPHER

DECIPHER (DatabasE of Genomic variants and Phenotype in Humans Using Ensembl Resources; https://decipher.sanger.ac.uk/) is a web-based resource and database of genomic variation data from analysis of patient DNA. It documents submicroscopic chromosome abnormalities (microdeletions and duplications) and pathogenic sequence variants (single nucleotide variants - SNVs, Insertions, Deletions, InDels), from over 25,000 patients and maps them to the human genome. In addition it catalogues the clinical characteristics from each patient and maintains a database of microdeletion/duplication syndromes, together with links to relevant scientific reports and support groups.

Worked example 7: DECIPHER

In this worked example, we will use DECIPHER to investigate Williams-Beuren Syndrome (WBS), a rare neurodevelopmental disorder (http://en.wikipedia.org/wiki/Williams_syndrome).

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



STEP 2 – Click on 'Disorders'

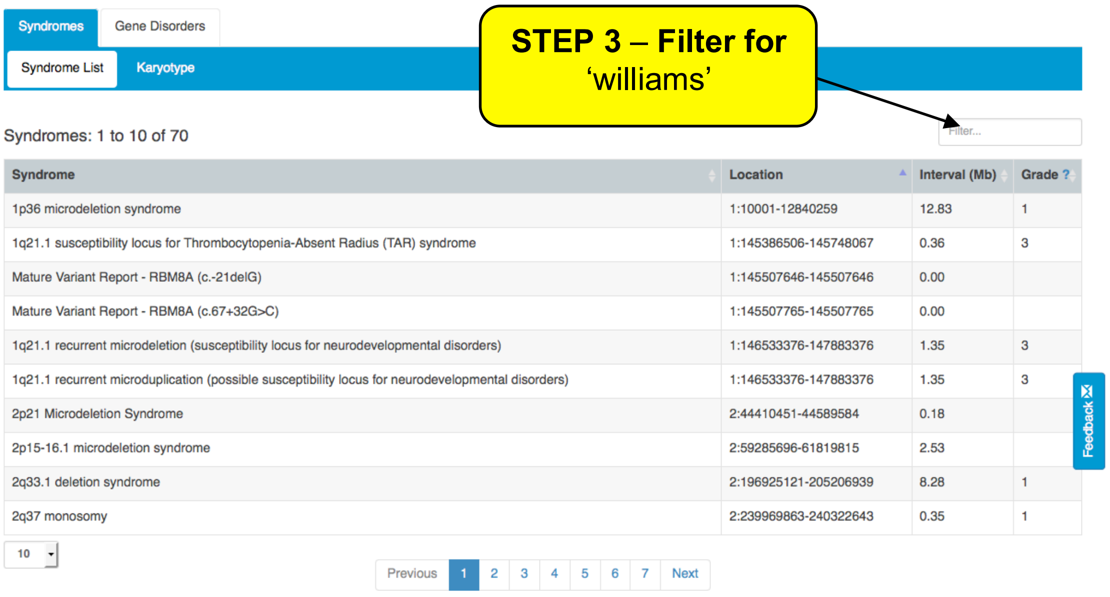
Latest news

DECIPHER v8.8 Released

Version 8.8 of DECIPHER was released on Wednesday 8th April and introduces the following new features and enhancements:

- **Absent Phenotypes:** When adding phenotypes to a patient, you can now add explicitly 'absent' phenotypes which might be expected or relevant which the patient does not exhibit.

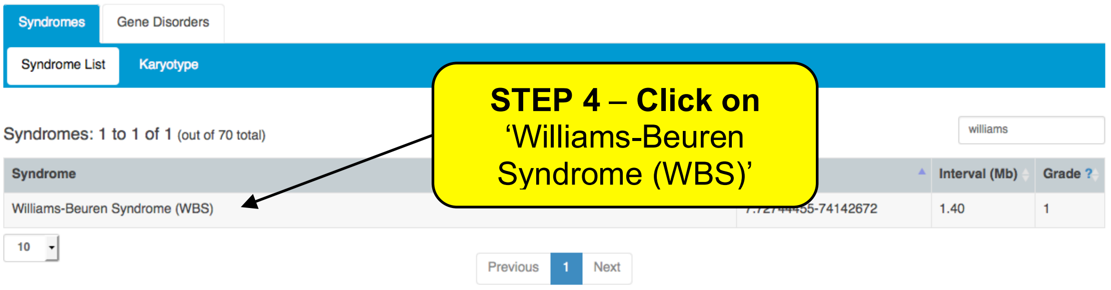
Patient 6 Mother 0 Father 6



STEP 3 – Filter for ‘williams’

Syndromes: 1 to 10 of 70

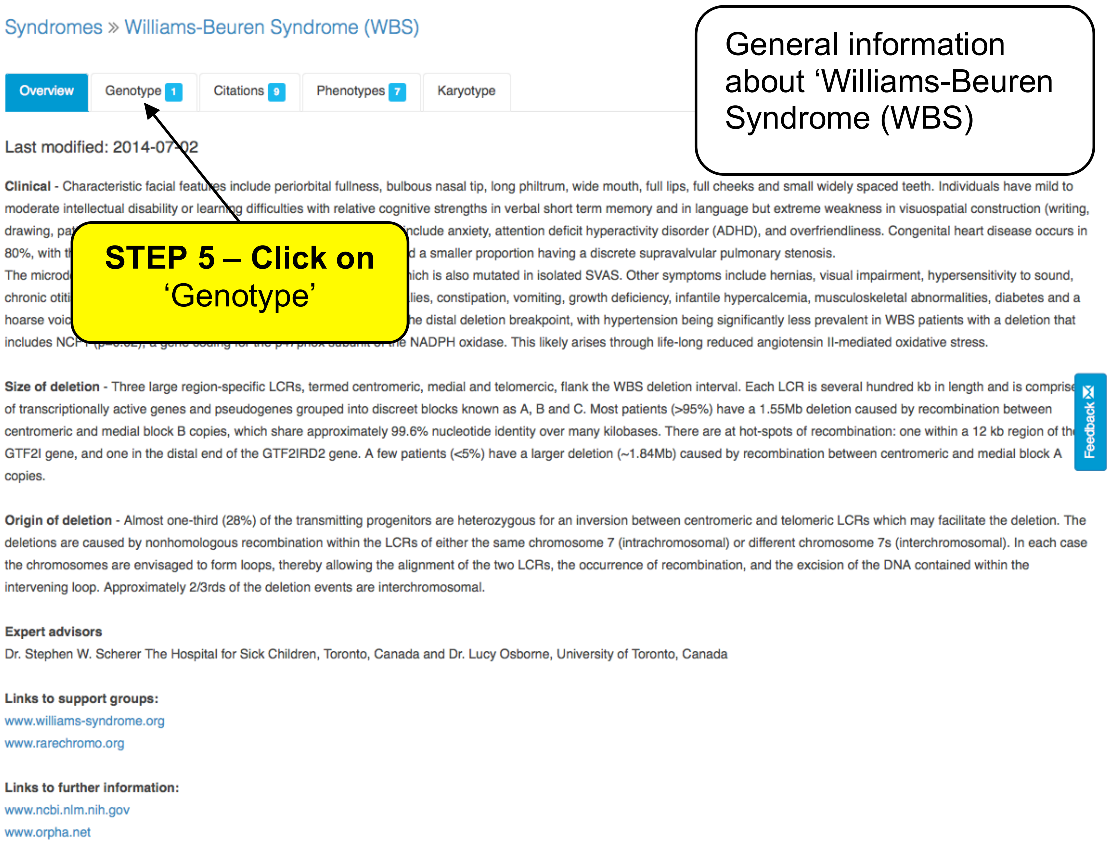
Syndrome	Location	Interval (Mb)	Grade ?
1p36 microdeletion syndrome	1:10001-12840259	12.83	1
1q21.1 susceptibility locus for Thrombocytopenia-Absent Radius (TAR) syndrome	1:145386506-145748067	0.36	3
Mature Variant Report - RBM8A (c.-21delG)	1:145507646-145507646	0.00	
Mature Variant Report - RBM8A (c.67+32G>C)	1:145507765-145507765	0.00	
1q21.1 recurrent microdeletion (susceptibility locus for neurodevelopmental disorders)	1:146533376-147883376	1.35	3
1q21.1 recurrent microduplication (possible susceptibility locus for neurodevelopmental disorders)	1:146533376-147883376	1.35	3
2p21 Microdeletion Syndrome	2:44410451-44589584	0.18	
2p15-16.1 microdeletion syndrome	2:59285696-61819815	2.53	
2q33.1 deletion syndrome	2:196925121-205206939	8.28	1
2q37 monosomy	2:239969863-240322643	0.35	1



STEP 4 – Click on ‘Williams-Beuren Syndrome (WBS)’

Syndromes: 1 to 1 of 1 (out of 70 total)

Syndrome		Interval (Mb)	Grade ?
Williams-Beuren Syndrome (WBS)	7:72744455-74142672	1.40	1



General information about ‘Williams-Beuren Syndrome (WBS)

STEP 5 – Click on ‘Genotype’

Last modified: 2014-07-02

Clinical - Characteristic facial features include periorbital fullness, bulbous nasal tip, long philtrum, wide mouth, full lips, full cheeks and small widely spaced teeth. Individuals have mild to moderate intellectual disability or learning difficulties with relative cognitive strengths in verbal short term memory and in language but extreme weakness in visuospatial construction (writing, drawing, pa include anxiety, attention deficit hyperactivity disorder (ADHD), and overfriendliness. Congenital heart disease occurs in 80%, with t d a smaller proportion having a discrete supravalvular pulmonary stenosis.
The microd hich is also mutated in isolated SVAS. Other symptoms include hernias, visual impairment, hypersensitivity to sound, chronic otiti lies, constipation, vomiting, growth deficiency, infantile hypercalcemia, musculoskeletal abnormalities, diabetes and a hoarse voic he distal deletion breakpoint, with hypertension being significantly less prevalent in WBS patients with a deletion that includes NC e NADPH oxidase. This likely arises through life-long reduced angiotensin II-mediated oxidative stress.

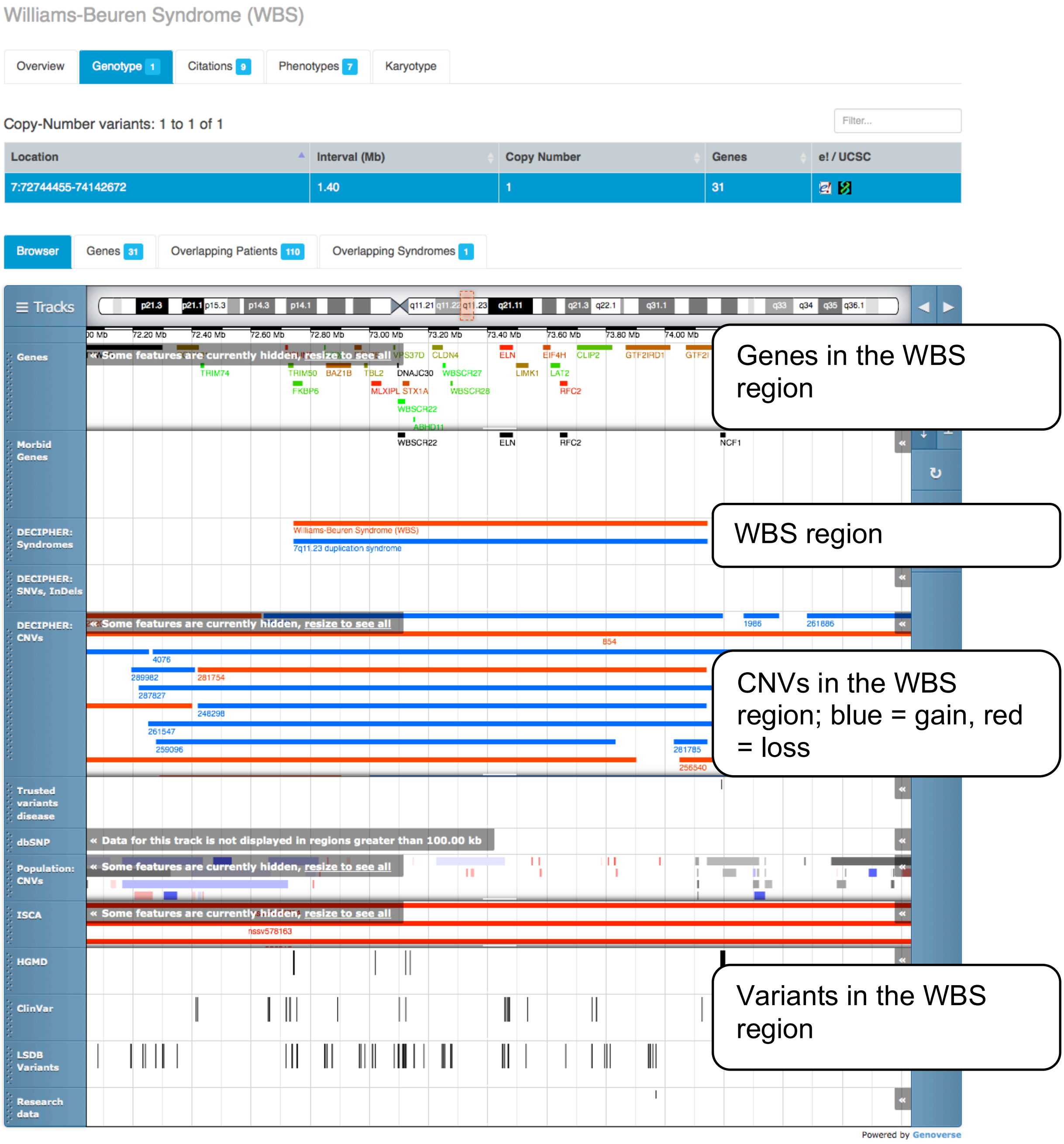
Size of deletion - Three large region-specific LCRs, termed centromeric, medial and telomercic, flank the WBS deletion interval. Each LCR is several hundred kb in length and is comprise of transcriptionally active genes and pseudogenes grouped into discreet blocks known as A, B and C. Most patients (>95%) 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 th 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 support groups:
www.williams-syndrome.org
www.rarechromo.org

Links to further information:
www.ncbi.nlm.nih.gov
www.orpha.net



Williams-Beuren Syndrome (WBS)

Copy-Number variants: 1 to 1 of 1

Location	Interval (Mb)	Copy Number	Genes	e! / UCSC
7:72744455-74142672	1.40	1	31	

Genes in the WBS region

WBS region

CNVs in the WBS region; blue = gain, red = loss

Variants in the WBS region

Biological ontologies

There is no universal standard terminology in biology and related domains, and term usages may be specific to a species, research area or even a particular research group. Different people may use different terms when referring to the same thing and/or may use the same term for different things. This makes communication and sharing of data more difficult. To try to make everyone using the same term when talking about the same thing, biological ontologies have been developed. Two widely-used biological ontologies are the Gene Ontology (GO) and the Sequence Ontology (SO).

Gene Ontology (GO)

The Gene Ontology (GO; http://geneontology.org/) consists of three hierarchically structured, controlled vocabularies that describe gene products in terms of their associated biological processes ("What does a gene product do?"), cellular components ("Where does a gene product do what it does?") and molecular functions ("How does a gene product do what it does?") in a species-independent manner.

Sequence Ontology (SO)

The Sequence Ontology (SO; http://www.sequenceontology.org/) is a set of terms and relationships used to describe the features and attributes of biological sequence.

Exercises

Ensembl Variant Effect Predictor

Resequencing of the genomic region of the human *CFTR* (cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) gene (ENSG00000001626) has revealed the following variants (alleles defined in the forward strand):

- Substitution G/A at position 7: 117,530,985
- Substitution T/C at position 7: 117,531,038
- Substitution T/C at position 7: 117,531,068

Use the Ensembl Variant Effect Predictor to answer the following questions:

(a) What genes are affected by these variants?
(b) Do any of the variants result in protein changes?
(c) If so, are these protein changes predicted to be deleterious or damaging?
(d) Have the variants already been annotated in Ensembl?

OMIM, GEO Profiles and Expression Atlas

Prostate cancer antigen 3 (*PCA3*, also referred to as *DD3*) is a gene that expresses a non-coding RNA. *PCA3* is only expressed in human prostate tissue, and the gene is highly overexpressed in prostate cancer. Because of its restricted expression profile, the *PCA3* RNA is useful as a tumor marker.

Explore what information is available about *PCA3* in OMIM, GEO Profiles and Expression Atlas. Can you find any evidence that this gene is indeed only expressed in prostate and that it is highly overexpressed in prostate cancer?

Exercises answers

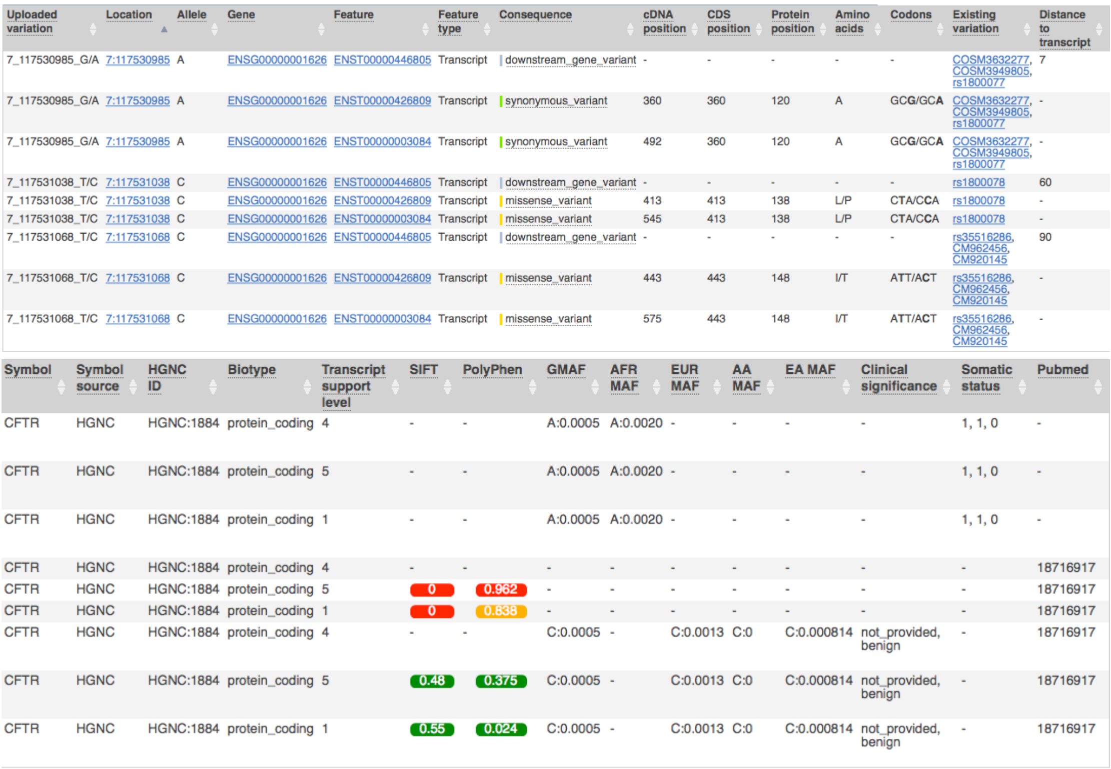
Ensembl Variant Effect Predictor

(a) The only gene affected by these variants is *CFTR*.

(b) The variant at position 117531038 results in a L/P change at position 138 in two of the CFTR proteins. The variant at position 117531068 results in an I/T change at position 148 in two of the CFTR proteins. The variant at position 117530985 doesn't result in a protein change.

(c) The protein change at position 138 is predicted to be deleterious / damaging. The protein change at position 148 is predicted to be tolerated / benign.

(d) All three variations have been already described and are known as rs1800077, rs1800078 and rs35516286 in dbSNP.

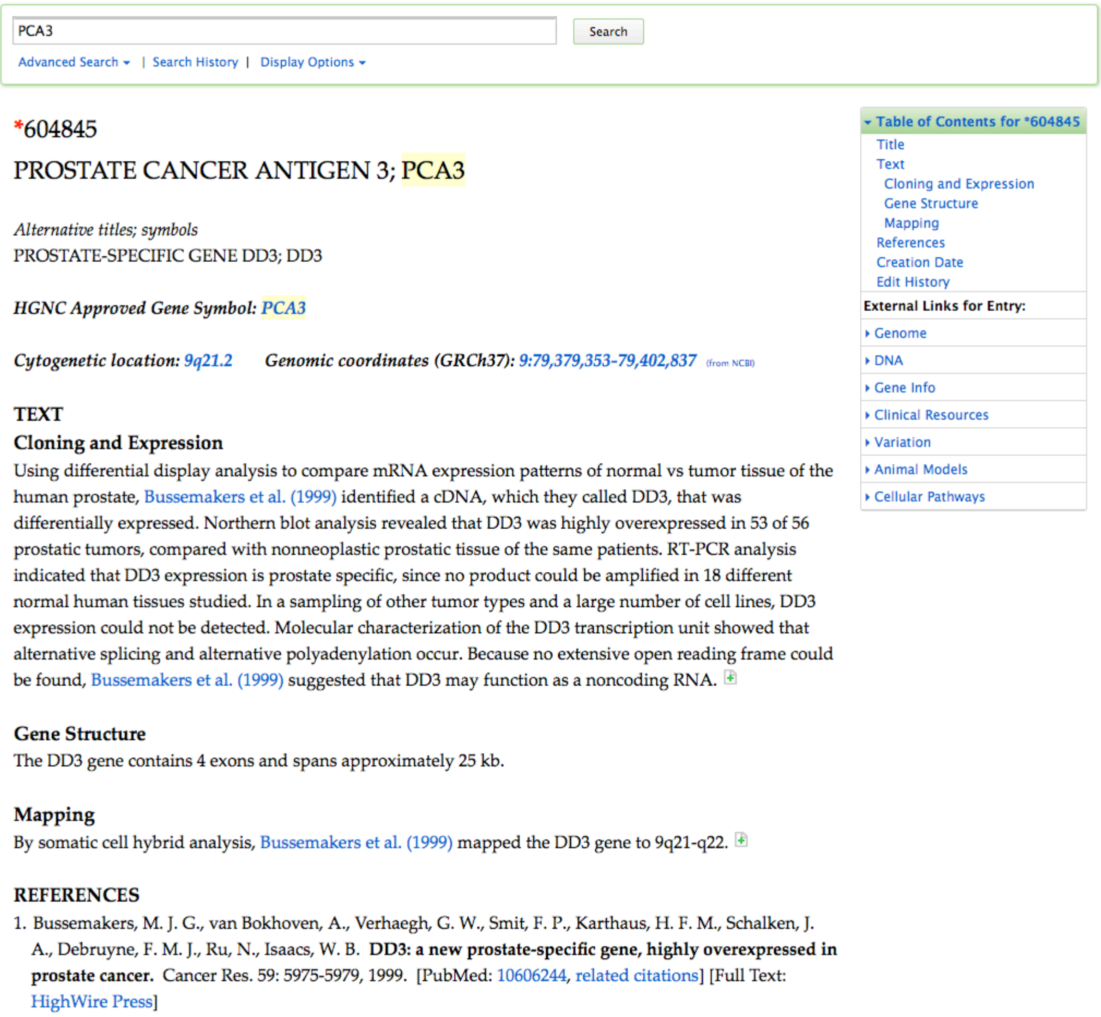


Uploaded variation	Location	Allele	Gene	Feature	Feature type	Consequence	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variation	Distance to transcript
7_117530985_G/A	7:117530985	A	ENSG00000001626	ENST00000446805	Transcript	downstream_gene_variant	-	-	-	-	-	COSM3632277, COSM3949805, rs1800077	7
7_117530985_G/A	7:117530985	A	ENSG00000001626	ENST00000426809	Transcript	synonymous_variant	360	360	120	A	GC**G**/GC**A**	COSM3632277, COSM3949805, rs1800077	-
7_117530985_G/A	7:117530985	A	ENSG00000001626	ENST00000003084	Transcript	synonymous_variant	492	360	120	A	GC**G**/GC**A**	COSM3632277, COSM3949805, rs1800077	-
7_117531038_T/C	7:117531038	C	ENSG00000001626	ENST00000446805	Transcript	downstream_gene_variant	-	-	-	-	-	rs1800078	60
7_117531038_T/C	7:117531038	C	ENSG00000001626	ENST00000426809	Transcript	missense_variant	413	413	138	L/P	C**T**A/C**C**A	rs1800078	-
7_117531038_T/C	7:117531038	C	ENSG00000001626	ENST00000003084	Transcript	missense_variant	545	413	138	L/P	C**T**A/C**C**A	rs1800078	-
7_117531068_T/C	7:117531068	C	ENSG00000001626	ENST00000446805	Transcript	downstream_gene_variant	-	-	-	-	-	rs35516286, CM962456, CM920145	90
7_117531068_T/C	7:117531068	C	ENSG00000001626	ENST00000426809	Transcript	missense_variant	443	443	148	I/T	A**T**T/A**C**T	rs35516286, CM962456, CM920145	-
7_117531068_T/C	7:117531068	C	ENSG00000001626	ENST00000003084	Transcript	missense_variant	575	443	148	I/T	A**T**T/A**C**T	rs35516286, CM962456, CM920145	-

Symbol	Symbol source	HGNC ID	Biotype	Transcript support level	SIFT	PolyPhen	GMAF	AFR MAF	EUR MAF	AA MAF	EA MAF	Clinical significance	Somatic status	Pubmed
CFTR	HGNC	HGNC:1884	protein_coding	4	-	-	A:0.0005	A:0.0020	-	-	-	-	1, 1, 0	-
CFTR	HGNC	HGNC:1884	protein_coding	5	-	-	A:0.0005	A:0.0020	-	-	-	-	1, 1, 0	-
CFTR	HGNC	HGNC:1884	protein_coding	1	-	-	A:0.0005	A:0.0020	-	-	-	-	1, 1, 0	-
CFTR	HGNC	HGNC:1884	protein_coding	4	-	-	-	-	-	-	-	-	-	18716917
CFTR	HGNC	HGNC:1884	protein_coding	5	0	0.962	-	-	-	-	-	-	-	18716917
CFTR	HGNC	HGNC:1884	protein_coding	1	0	0.838	-	-	-	-	-	-	-	18716917
CFTR	HGNC	HGNC:1884	protein_coding	4	-	-	C:0.0005	-	C:0.0013	C:0	C:0.000814	not_provided, benign	-	18716917
CFTR	HGNC	HGNC:1884	protein_coding	5	0.48	0.375	C:0.0005	-	C:0.0013	C:0	C:0.000814	not_provided, benign	-	18716917
CFTR	HGNC	HGNC:1884	protein_coding	1	0.55	0.024	C:0.0005	-	C:0.0013	C:0	C:0.000814	not_provided, benign	-	18716917

OMIM, GEO Profiles and Expression Atlas

OMIM contains basic information about the *PCA3* gene (official and alternative gene symbols, cytogenetic and genomic location) as well as some information from a paper that has shown overexpression of this gene in prostate tumors and has suggested that the gene codes for a noncoding RNA.



*604845

PROSTATE CANCER ANTIGEN 3; PCA3

Alternative titles; symbols
PROSTATE-SPECIFIC GENE DD3; DD3

HGNC Approved Gene Symbol: PCA3

Cytogenetic location: 9q21.2 ***Genomic coordinates (GRCh37):*** 9:79,379,353-79,402,837 (from NCBI)

TEXT

Cloning and Expression

Using differential display analysis to compare mRNA expression patterns of normal vs tumor tissue of the human prostate, Bussemakers et al. (1999) identified a cDNA, which they called DD3, that was differentially expressed. Northern blot analysis revealed that DD3 was highly overexpressed in 53 of 56 prostatic tumors, compared with nonneoplastic prostatic tissue of the same patients. RT-PCR analysis indicated that DD3 expression is prostate specific, since no product could be amplified in 18 different normal human tissues studied. In a sampling of other tumor types and a large number of cell lines, DD3 expression could not be detected. Molecular characterization of the DD3 transcription unit showed that alternative splicing and alternative polyadenylation occur. Because no extensive open reading frame could be found, Bussemakers et al. (1999) suggested that DD3 may function as a noncoding RNA.

Gene Structure

The DD3 gene contains 4 exons and spans approximately 25 kb.

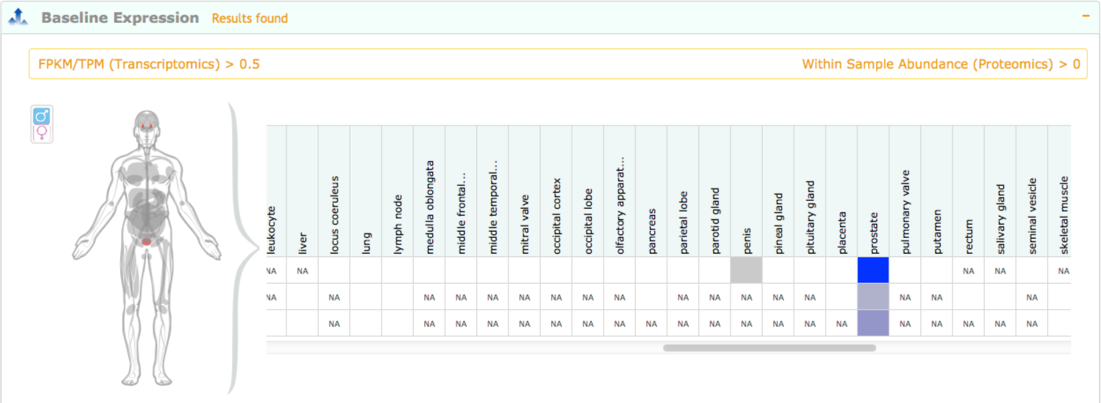
Mapping

By somatic cell hybrid analysis, Bussemakers et al. (1999) mapped the DD3 gene to 9q21-q22.

REFERENCES

1. Bussemakers, M. J. G., van Bokhoven, A., Verhaegh, G. W., Smit, F. P., Karthaus, H. F. M., Schalken, J. A., Debruyne, F. M. J., Ru, N., Isaacs, W. B. **DD3: a new prostate-specific gene, highly overexpressed in prostate cancer.** Cancer Res. 59: 5975-5979, 1999. [PubMed: 10606244, related citations] [Full Text: HighWire Press]

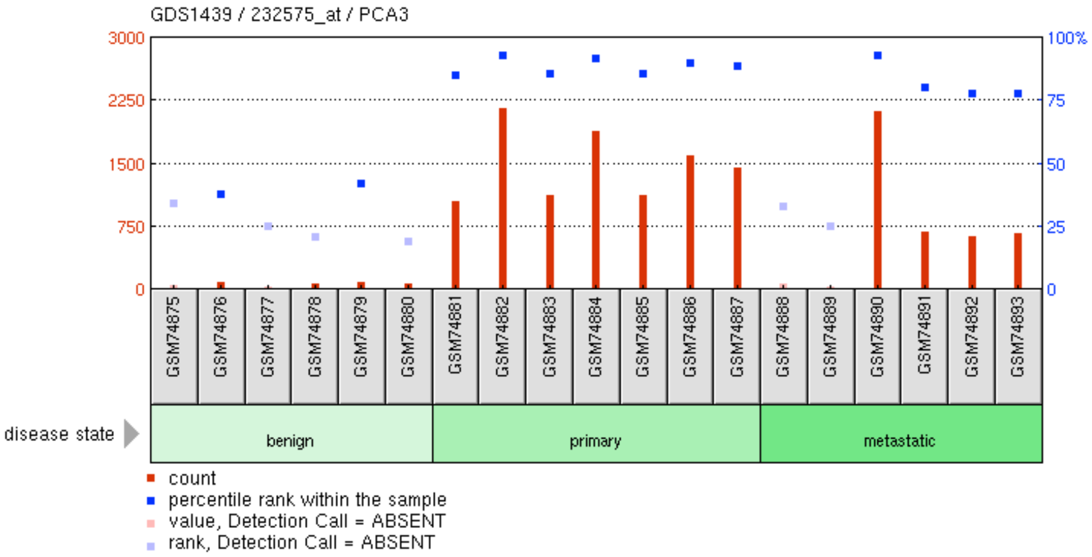
According to the Baseline Expression in Expression Atlas, the *PCA3* gene is indeed only expressed in the prostate:


Baseline Expression
Results found
FPKM/TPM (Transcriptomics) > 0.5
Within Sample Abundance (Proteomics) > 0

leukocyte	liver	locus coeruleus	lung	lymph node	medulla oblongata	middle frontal…	middle temporal…	mitral valve	occipital cortex	occipital lobe	olfactory apparat…	pancreas	parietal lobe	parotid gland	penis	pineal gland	pituitary gland	placenta	prostate	pulmonary valve	putamen	rectum	salivary gland	seminal vesicle	skeletal muscle
NA	NA														▨				■			NA	NA		NA
NA		NA			NA	NA	NA	NA	NA	NA	NA		NA	NA	NA	NA	NA		▨	NA	NA			NA	
		NA			NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	▨	NA	NA	NA	NA	NA	

GEO Profiles shows a number expression profiles from which it is clear that the *PCA3* gene is highly expressed in prostate cancer. For example:

Profile	GDS1439 / 232575_at / PCA3
Title	Prostate cancer progression
Organism	Homo sapiens


GDS1439 / 232575_at / PCA3
GSM74875
GSM74876
GSM74877
GSM74878
GSM74879
GSM74880
GSM74881
GSM74882
GSM74883
GSM74884
GSM74885
GSM74886
GSM74887
GSM74888
GSM74889
GSM74890
GSM74891
GSM74892
GSM74893
disease state
benign
primary
metastatic
count
percentile rank within the sample
value, Detection Call = ABSENT
rank, Detection Call = ABSENT