Database Integration

Paul Flicek

Vertebrate Genomics

EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

(Dramatically) Simplified Clinical Workflow

Identify variants

Technically easy and getting easier

Use what we already know to make some sense of them

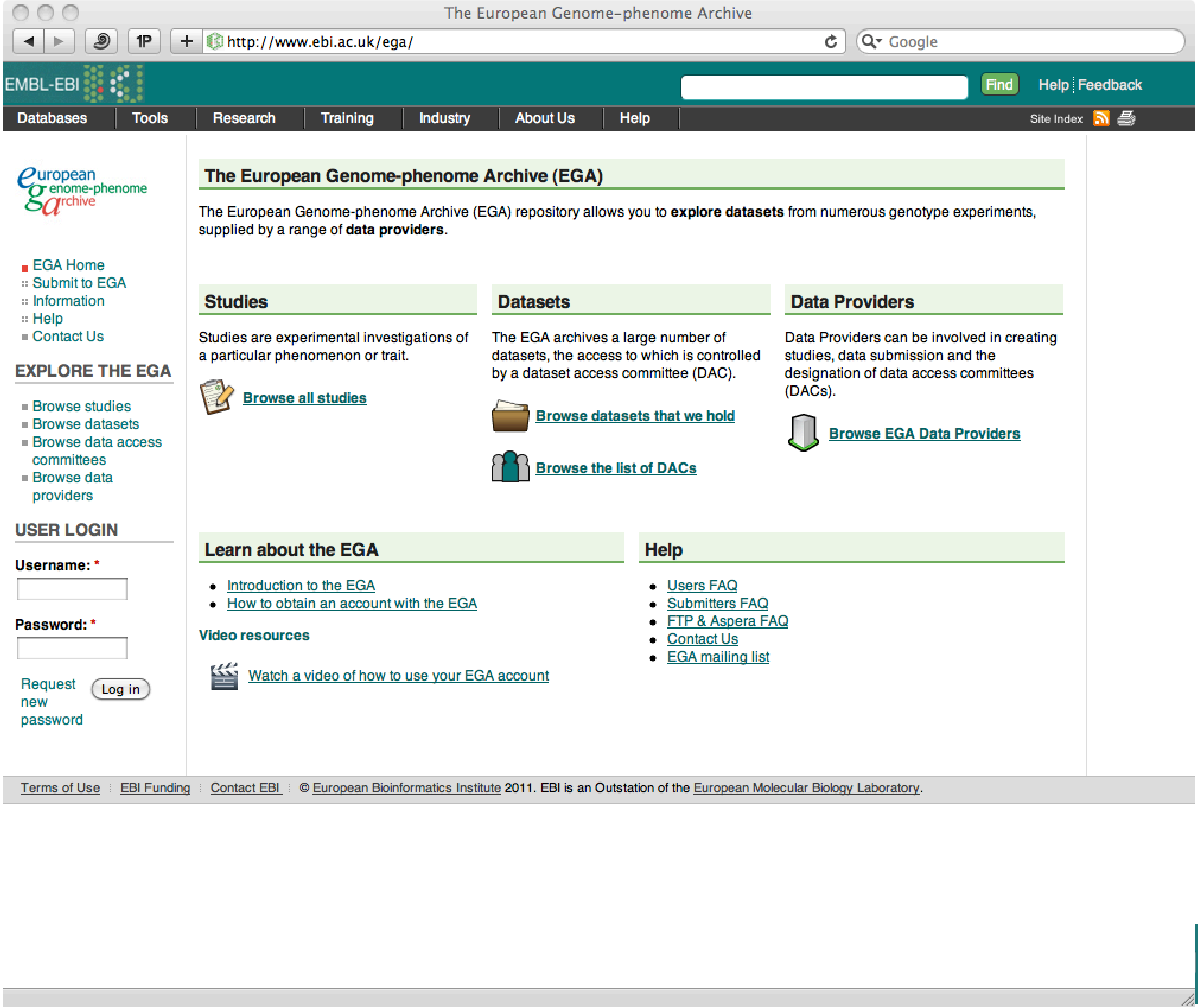
nething
ut it

Data interpretation: beyond research toward medical practice

- Needs:
  - Consistent, traceable data generation and analysis routines
  - Robust annotation based on public information sources such as those at the EBI and NCBI
    - Probably 95% of all information that could be used to understand and interpret human variation is already in the public domain
  - Reporting into medical records

Database integration

- Part 1: Continually update the existing information to ensure it is accurate and comprehensive
- Part 2: Provide some method to search relevant resources using variants and/or whole genomes as input


The European Genome-phenome Archive
http://www.ebi.ac.uk/ega/
Google
EMBL-EBI
Find
Help
Feedback
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
european genome-phenome archive

- EGA Home
- Submit to EGA
- Information
- Help
- Contact Us

EXPLORE THE EGA

- Browse studies
- Browse datasets
- Browse data access committees
- Browse data providers

USER LOGIN

Username: *

Password: *

Request new password

Log in

The European Genome-phenome Archive (EGA)

The European Genome-phenome Archive (EGA) repository allows you to **explore datasets** from numerous genotype experiments, supplied by a range of **data providers**.

Studies

Studies are experimental investigations of a particular phenomenon or trait.

Browse all studies

Datasets

The EGA archives a large number of datasets, the access to which is controlled by a dataset access committee (DAC).

Browse datasets that we hold

Browse the list of DACs

Data Providers

Data Providers can be involved in creating studies, data submission and the designation of data access committees (DACs).

Browse EGA Data Providers

Learn about the EGA

- Introduction to the EGA
- How to obtain an account with the EGA

Video resources

Watch a video of how to use your EGA account

Help

- Users FAQ
- Submitters FAQ
- FTP & Aspera FAQ
- Contact Us
- EGA mailing list

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2011. EBI is an Outstation of the European Molecular Biology Laboratory.

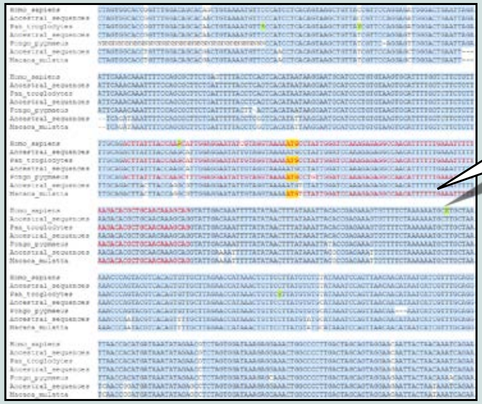
The European Genome-phenome Archive

- Secure storage and authorised access to all types of data sets that might be generated in the context of research into molecular medicine
  - DNA sequence; Array-based genotypes; epigenetic data
  - Transcriptomics; Proteomics
  - Phenotype data

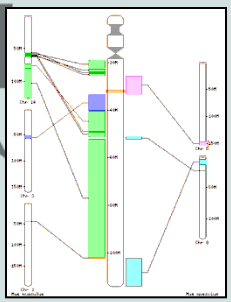
- Used for GWAS, ICGC, IHEC, IHMC, UK10K and data

- EGA supports only data access decisions that are based on original consent
  - Authorized users have personal accounts in our system
  - Access to the data requires account password
  - Data decryption requires a separate key that must be requested and is sent off line

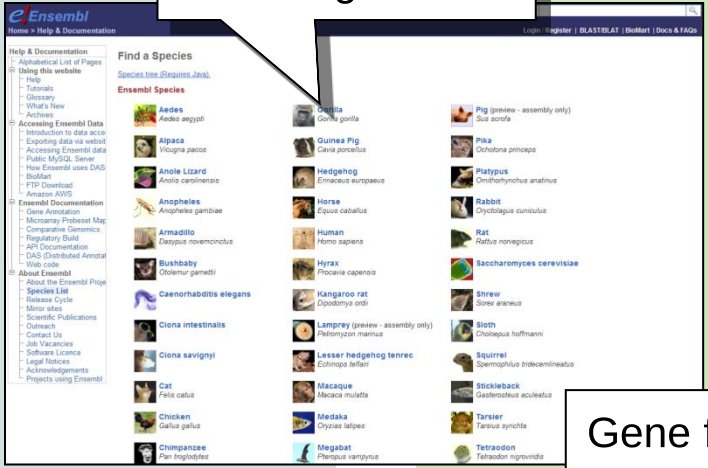
Ensembl genome-wide annotation



Genomic alignments



Synteny



Pick a genome

e! Ensembl

Home > Help & Documentation

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Help & Documentation

- Alphabetical List of Pages
- Using this website
  - Help
  - Tutorials
  - Glossary
  - What's New
  - Archives
- Accessing Ensembl Data
  - Introduction to data acce
  - Exporting data via websit
  - Accessing Ensembl data
  - Public MySQL Server
  - How Ensembl uses DAS
  - BioMart
  - FTP Download
  - Amazon AWS
- Ensembl Documentation
  - Gene Annotation
  - Microarray Probeset Map
  - Comparative Genomics
  - Regulatory Build
  - API Documentation
  - DAS (Distributed Annotat
  - Web code
- About Ensembl
  - About the Ensembl Proje
  - Species List
  - Release Cycle
  - Mirror sites
  - Scientific Publications
  - Outreach
  - Contact Us
  - Job Vacancies
  - Software Licence
  - Legal Notices
  - Acknowledgements
  - Projects using Ensembl

Find a Species

Species tree (Requires Java).

Ensembl Species

Aedes *Aedes aegypti*

Alpaca *Vicugna pacos*

Anole Lizard *Anolis carolinensis*

Anopheles *Anopheles gambiae*

Armadillo *Dasypus novemcinctus*

Bushbaby *Otolemur garnettii*

Caenorhabditis elegans

Ciona intestinalis

Ciona savignyi

Cat *Felis catus*

Chicken *Gallus gallus*

Chimpanzee *Pan troglodytes*

Gorilla *Gorilla gorilla*

Guinea Pig *Cavia porcellus*

Hedgehog *Erinaceus europaeus*

Horse *Equus caballus*

Human *Homo sapiens*

Hyrax *Procavia capensis*

Kangaroo rat *Dipodomys ordii*

Lamprey (preview - assembly only) *Petromyzon marinus*

Lesser hedgehog tenrec *Echinops telfairi*

Macaque *Macaca mulatta*

Medaka *Oryzias latipes*

Megabat *Pteropus vampyrus*

Pig (preview - assembly only) *Sus scrofa*

Pika *Ochotona princeps*

Platypus *Ornithorhynchus anatinus*

Rabbit *Oryctolagus cuniculus*

Rat *Rattus norvegicus*

Saccharomyces cerevisiae

Shrew *Sorex araneus*

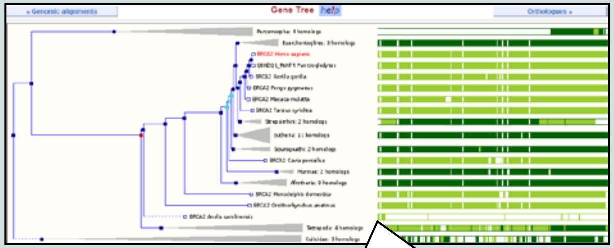
Sloth *Choloepus hoffmanni*

Squirrel *Spermophilus tridecemlineatus*

Stickleback *Gasterosteus aculeatus*

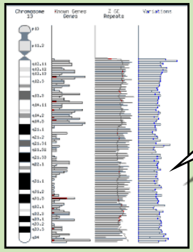
Tarsier *Tarsius syrichta*

Tetraodon *Tetraodon nigroviridis*

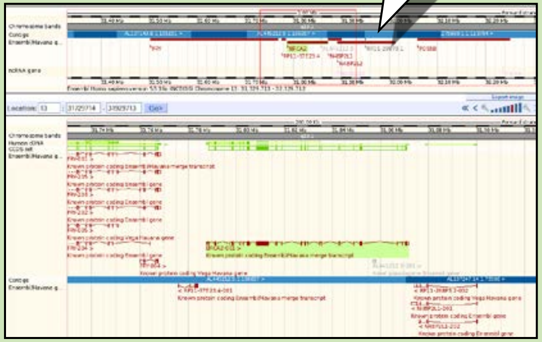


Orthology

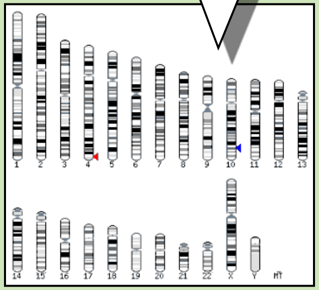
Across species



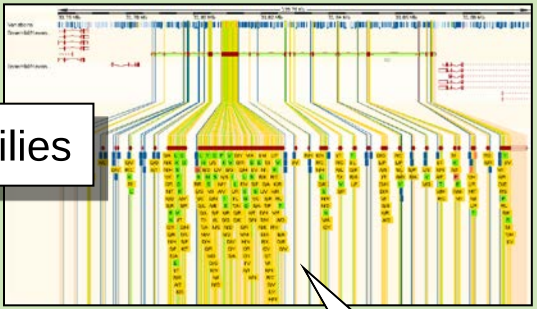
Chromosomes



Genes



Gene families



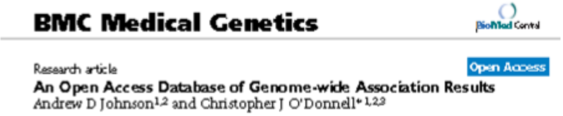
SNPs

Within species

Integrating variation data across the genome

- Polymorphism data (from dbSNP)
  - SNPs and indels for 14 species including 1000 Genomes
  - Allele and genotype frequencies by population
  - Locus-specific data from LRG
- Structural polymorphism data
- Mutation data (human)
  - Somatic mutation data (from COSMIC)
  - Human Gene Mutation Database (HGMD) IDs

- Phenotype associations: OMIM, UniProt, GWAS
- Affymetrix and Illumina chipsets



BMC Medical Genetics

BioMed Central

Research article

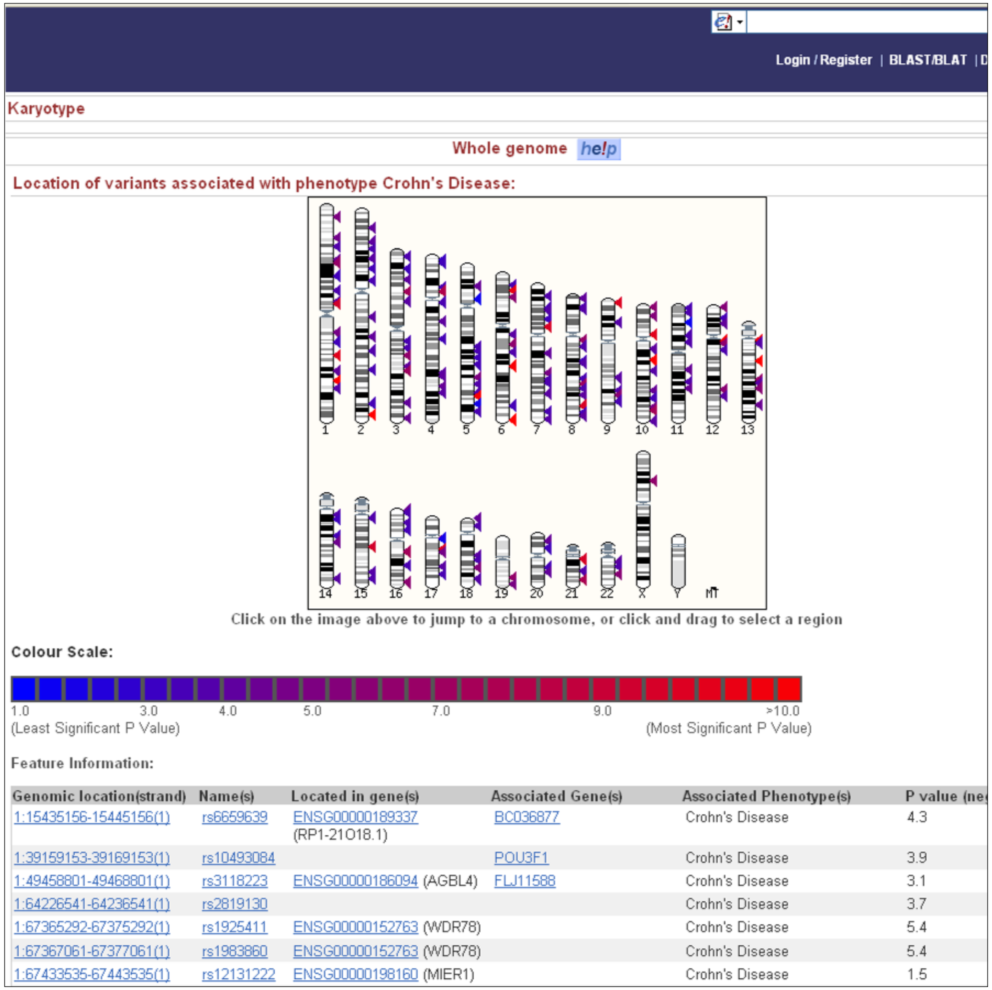
Open Access

An Open Access Database of Genome-wide Association Results

Andrew D Johnson[1,2] and Christopher J O'Donnell*[1,2,3]

Variation annotation – phenotype data

- 37,964 somatic mutations:
  - COSMIC
- 57,930 germline mutations:
  - HGMD
- 56,177 literature curation:
  - OMIM
  - UniProt
- 62,737 GWAS data:
  - NHGRI GWAS catalog
  - Open Access DB
  - EGA
- 22,449 from SNPedia by DAS



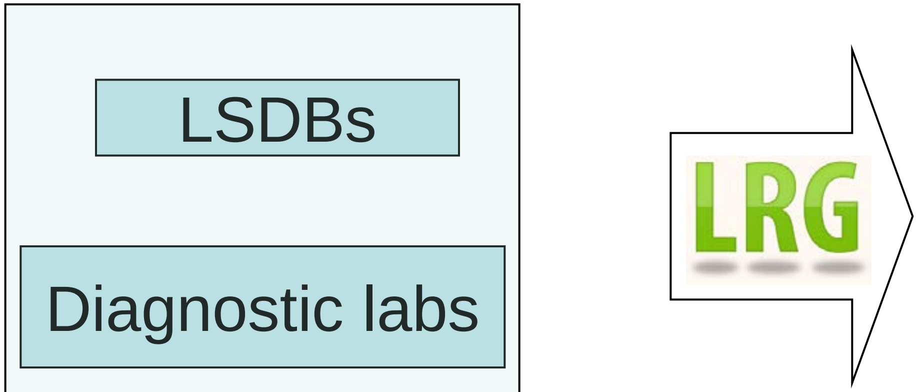
Location of variants associated with phenotype Crohn's Disease:

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

Feature Information:

Genomic location(strand)	Name(s)	Located in gene(s)	Associated Gene(s)	Associated Phenotype(s)	P value (ne
1:15435156-15445156(1)	rs6659639	ENSG00000189337 (RP1-21O18.1)	BC036877	Crohn's Disease	4.3
1:39159153-39169153(1)	rs10493084		POU3F1	Crohn's Disease	3.9
1:49458801-49468801(1)	rs3118223	ENSG00000186094 (AGBL4)	FLJ11588	Crohn's Disease	3.1
1:64226541-64236541(1)	rs2819130			Crohn's Disease	3.7
1:67365292-67375292(1)	rs1925411	ENSG00000152763 (WDR78)		Crohn's Disease	5.4
1:67367061-67377061(1)	rs1983860	ENSG00000152763 (WDR78)		Crohn's Disease	5.4
1:67433535-67443535(1)	rs12131222	ENSG00000198160 (MIER1)		Crohn's Disease	1.5

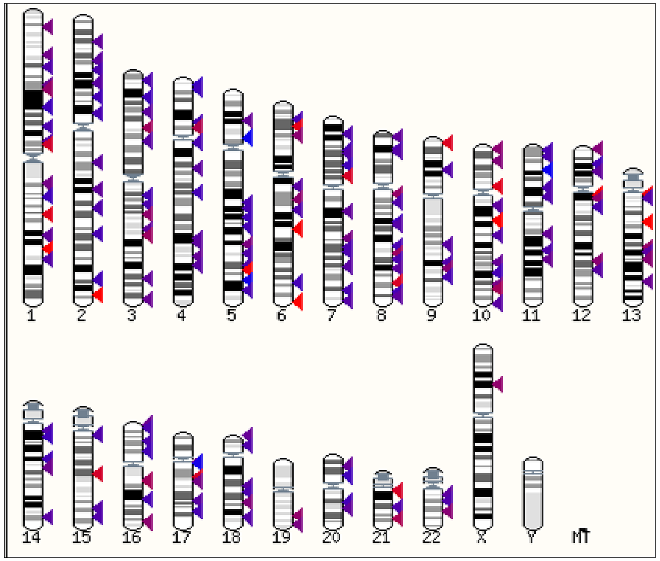
Variation annotation – phenotype data


LSDBs

Diagnostic labs

Locus-specific information

Dalgleish, et al. *Genome Medicine* 2010



Genome-wide information

- LRG project- Locus Reference Genomic
  - Create stable reference sequences (LRGs)
  - Use LRGs for exchange of variation data

Database integration

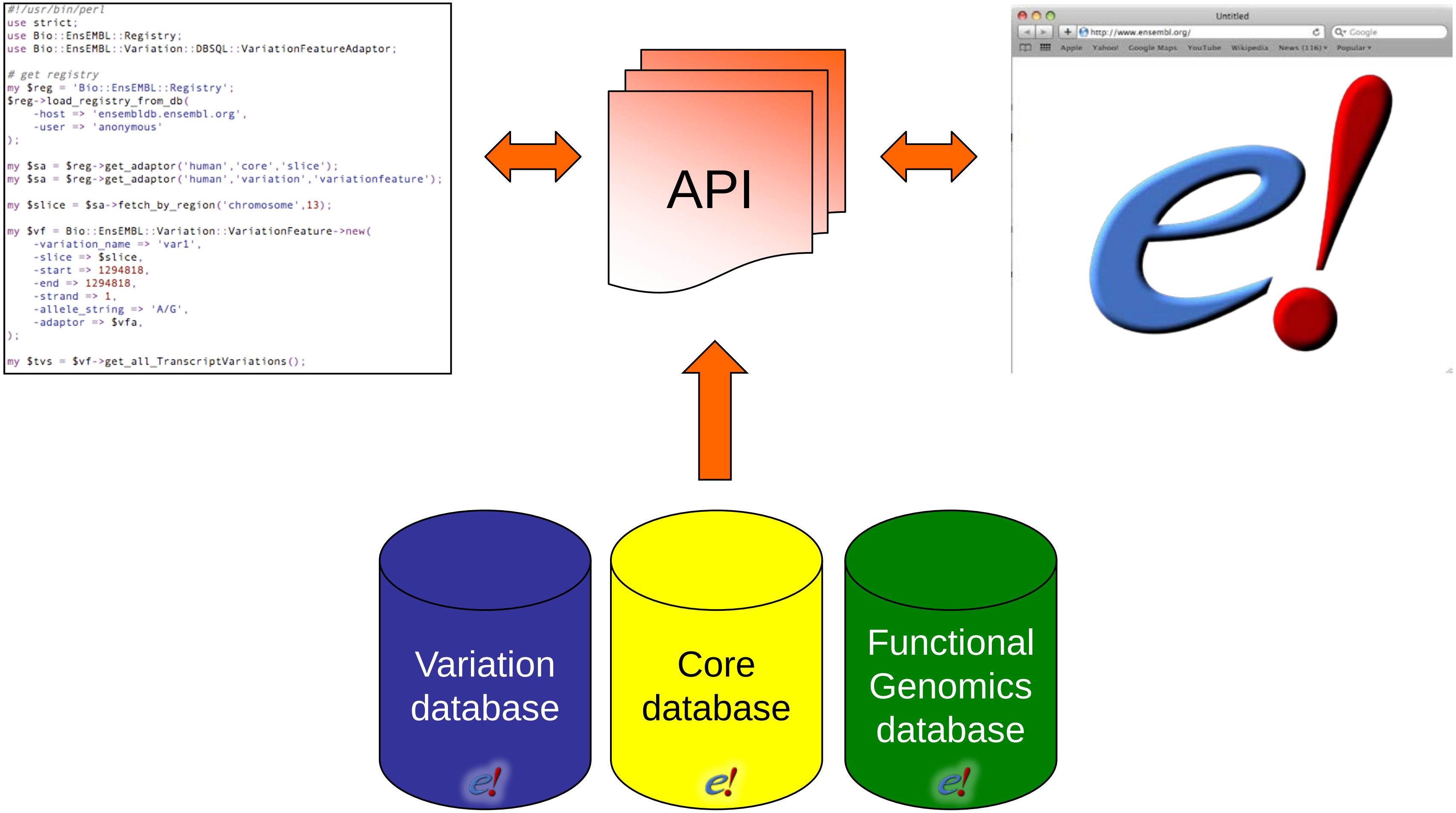
- Part 1: Continually update the existing information to ensure it is accurate and comprehensive
- Part 2: Provide some method to search relevant resources using variants and/or whole genomes as input

Ensembl Variant Effect Prediction (VEP) tool

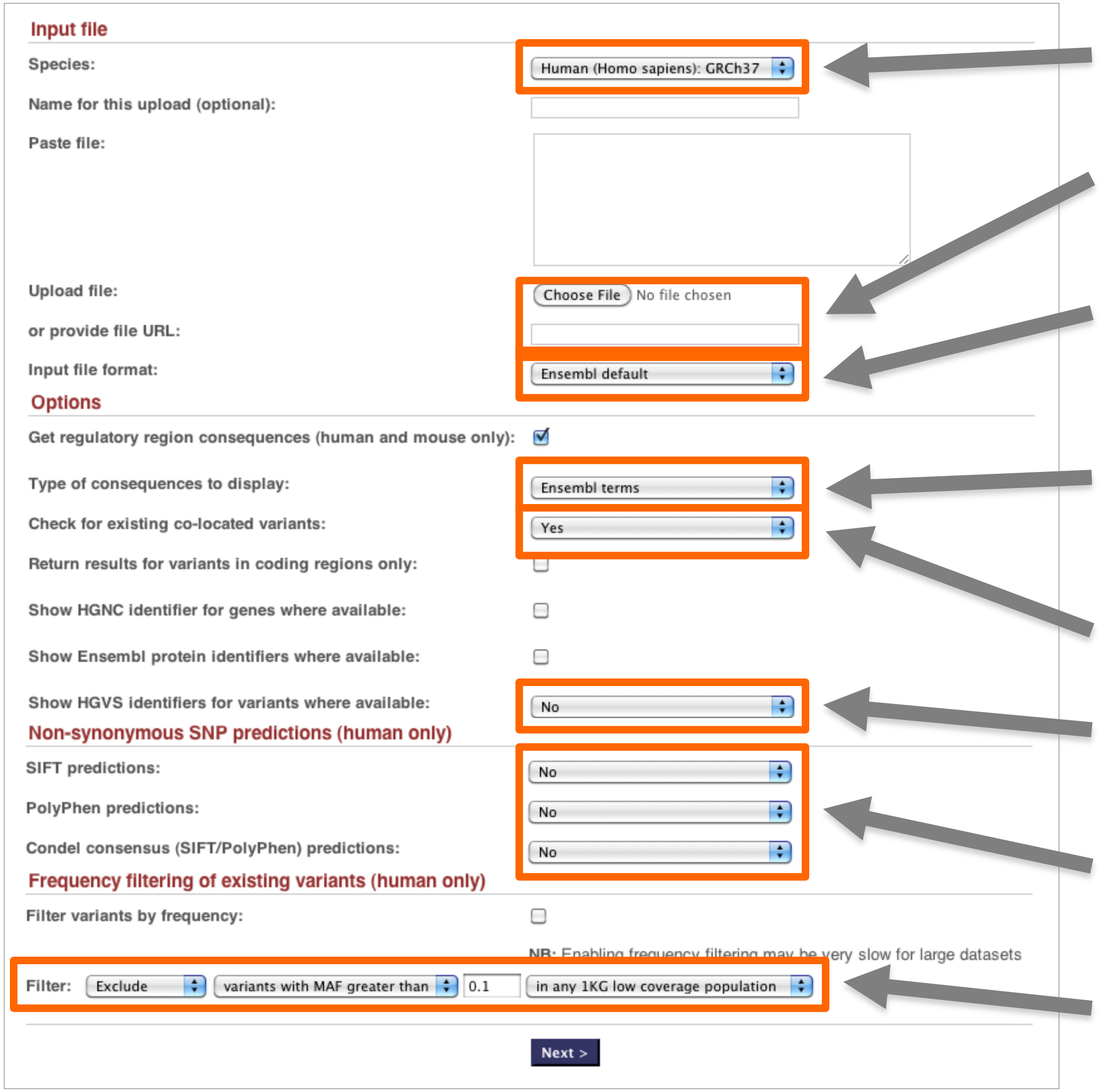
- Calculates the effect of SNPs in the context of Ensembl genes and regulatory features
  - Web and API interface
  - Code back-ported to support NCBI36 assembly
  - Programmatic support for tab-delimited and VCF files
  - Easily integrated into analysis pipelines
- Working within ICGC to capture structural and other genome rearrangements
- Disruption of experimentally observed TF binding sites and conserved regions
- Ability to run without connection to the internet
- Support for user defined analysis plug-ins coming in January 2012
- Will return if variant is present in EGA dataset in 2012
- Effectively a variant based search of EBI’s data resources

McLaren, et al. *Bioinformatics.* 2010

Ensembl VEP Implementation



```perl
#!/usr/bin/perl
use strict;
use Bio::EnsEMBL::Registry;
use Bio::EnsEMBL::Variation::DBSQL::VariationFeatureAdaptor;

# get registry
my $reg = 'Bio::EnsEMBL::Registry';
$reg->load_registry_from_db(
    -host => 'ensembldb.ensembl.org',
    -user => 'anonymous'
);

my $sa = $reg->get_adaptor('human','core','slice');
my $sa = $reg->get_adaptor('human','variation','variationfeature');

my $slice = $sa->fetch_by_region('chromosome',13);

my $vf = Bio::EnsEMBL::Variation::VariationFeature->new(
    -variation_name => 'var1',
    -slice => $slice,
    -start => 1294818,
    -end => 1294818,
    -strand => 1,
    -allele_string => 'A/G',
    -adaptor => $vfa,
);

my $tvs = $vf->get_all_TranscriptVariations();
```



Input file

Species: Human (Homo sapiens): GRCh37

Name for this upload (optional):

Paste file:

Upload file: Choose File No file chosen

or provide file URL:

Input file format: Ensembl default

Options

Get regulatory region consequences (human and mouse only): ☑

Type of consequences to display: Ensembl terms

Check for existing co-located variants: Yes

Return results for variants in coding regions only: □

Show HGNC identifier for genes where available: □

Show Ensembl protein identifiers where available: □

Show HGVS identifiers for variants where available: No

Non-synonymous SNP predictions (human only)

SIFT predictions: No

PolyPhen predictions: No

Condel consensus (SIFT/PolyPhen) predictions: No

Frequency filtering of existing variants (human only)

Filter variants by frequency: □

NB: Enabling frequency filtering may be very slow for large datasets

Filter: Exclude variants with MAF greater than 0.1 in any 1KG low coverage population

50+ species at www.ensembl.org
300+ at www.ensemblgenomes.org

Data input by file upload or external URL

Support for multiple file formats:
VCF, Pileup, HGVS, dbSNP rsID

Output Ensembl, Sequence Ontology (SO) or NCBI consequence terms

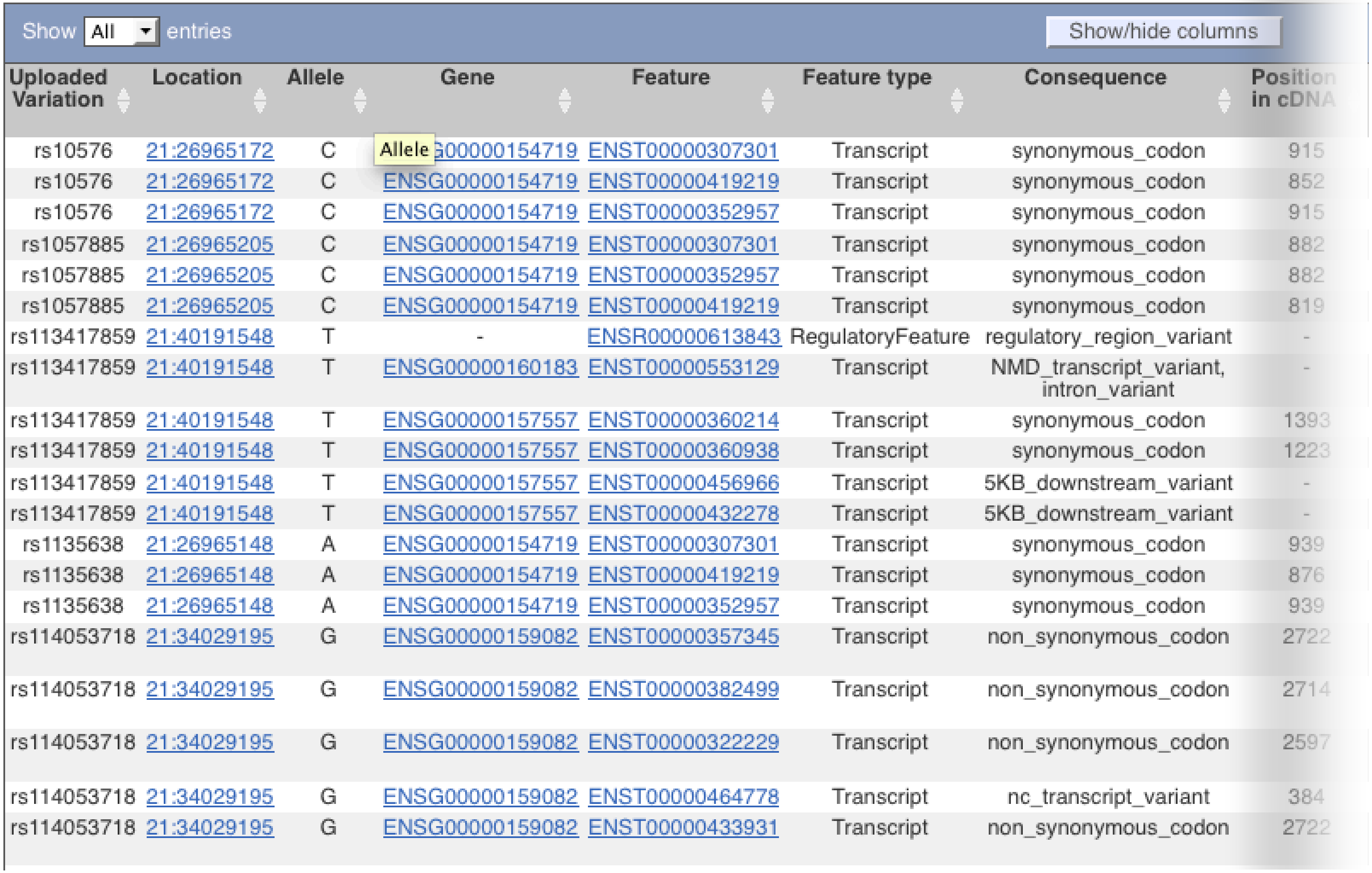
Find existing overlapping variants annotated by Ensembl

Create HGVS notations

Include **SIFT**, **PolyPhen** and **Condel** predictions for non-synonymous changes in human

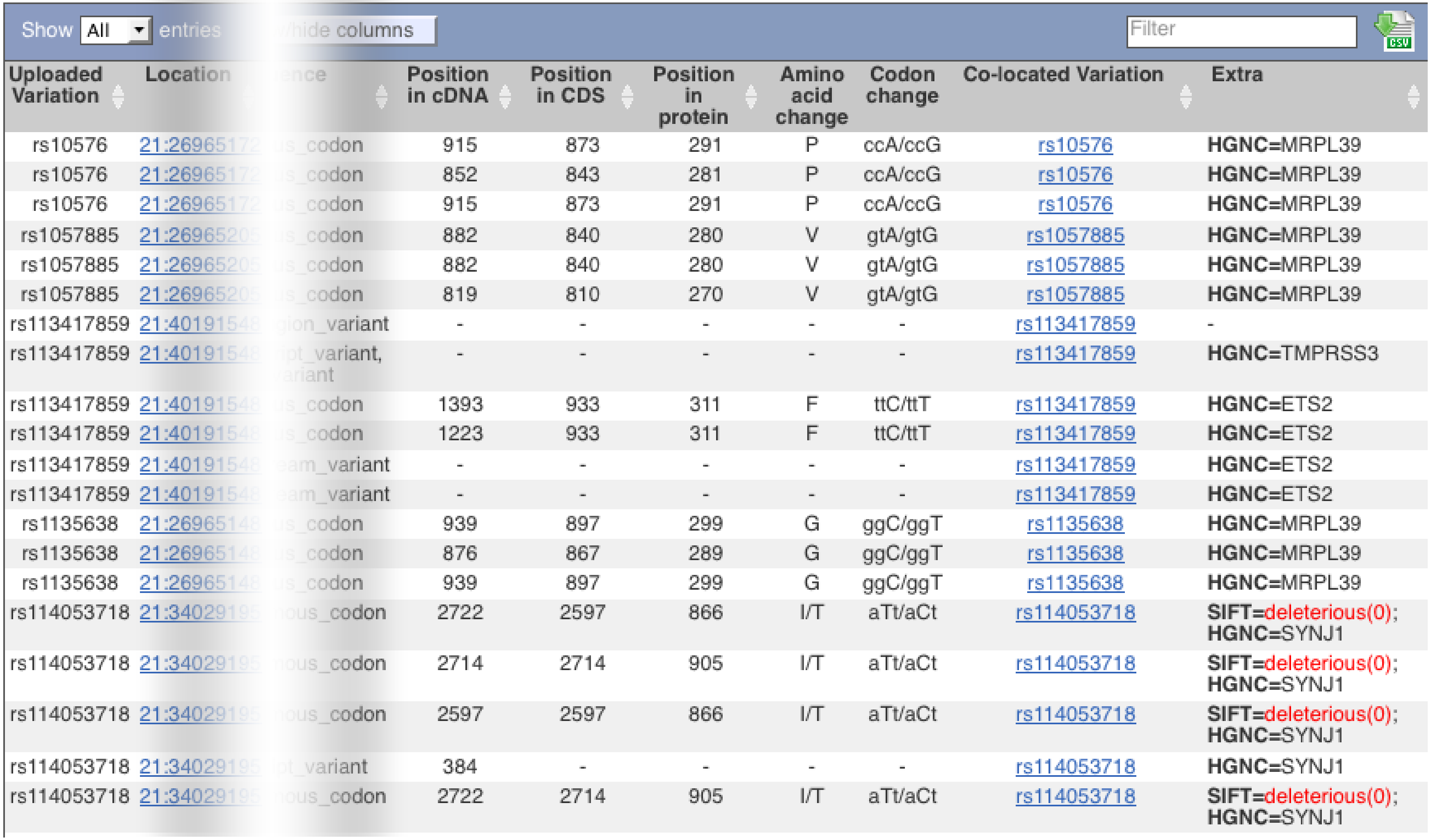
Filter input against **HapMap** or **1000 genomes** frequency data

Output



Uploaded Variation	Location	Allele	Gene	Feature	Feature type	Consequence	Position in cDNA
rs10576	21:26965172	C	G00000154719	ENST00000307301	Transcript	synonymous_codon	915
rs10576	21:26965172	C	ENSG00000154719	ENST00000419219	Transcript	synonymous_codon	852
rs10576	21:26965172	C	ENSG00000154719	ENST00000352957	Transcript	synonymous_codon	915
rs1057885	21:26965205	C	ENSG00000154719	ENST00000307301	Transcript	synonymous_codon	882
rs1057885	21:26965205	C	ENSG00000154719	ENST00000352957	Transcript	synonymous_codon	882
rs1057885	21:26965205	C	ENSG00000154719	ENST00000419219	Transcript	synonymous_codon	819
rs113417859	21:40191548	T	-	ENSR00000613843	RegulatoryFeature	regulatory_region_variant	-
rs113417859	21:40191548	T	ENSG00000160183	ENST00000553129	Transcript	NMD_transcript_variant, intron_variant	-
rs113417859	21:40191548	T	ENSG00000157557	ENST00000360214	Transcript	synonymous_codon	1393
rs113417859	21:40191548	T	ENSG00000157557	ENST00000360938	Transcript	synonymous_codon	1223
rs113417859	21:40191548	T	ENSG00000157557	ENST00000456966	Transcript	5KB_downstream_variant	-
rs113417859	21:40191548	T	ENSG00000157557	ENST00000432278	Transcript	5KB_downstream_variant	-
rs1135638	21:26965148	A	ENSG00000154719	ENST00000307301	Transcript	synonymous_codon	939
rs1135638	21:26965148	A	ENSG00000154719	ENST00000419219	Transcript	synonymous_codon	876
rs1135638	21:26965148	A	ENSG00000154719	ENST00000352957	Transcript	synonymous_codon	939
rs114053718	21:34029195	G	ENSG00000159082	ENST00000357345	Transcript	non_synonymous_codon	2722
rs114053718	21:34029195	G	ENSG00000159082	ENST00000382499	Transcript	non_synonymous_codon	2714
rs114053718	21:34029195	G	ENSG00000159082	ENST00000322229	Transcript	non_synonymous_codon	2597
rs114053718	21:34029195	G	ENSG00000159082	ENST00000464778	Transcript	nc_transcript_variant	384
rs114053718	21:34029195	G	ENSG00000159082	ENST00000433931	Transcript	non_synonymous_codon	2722

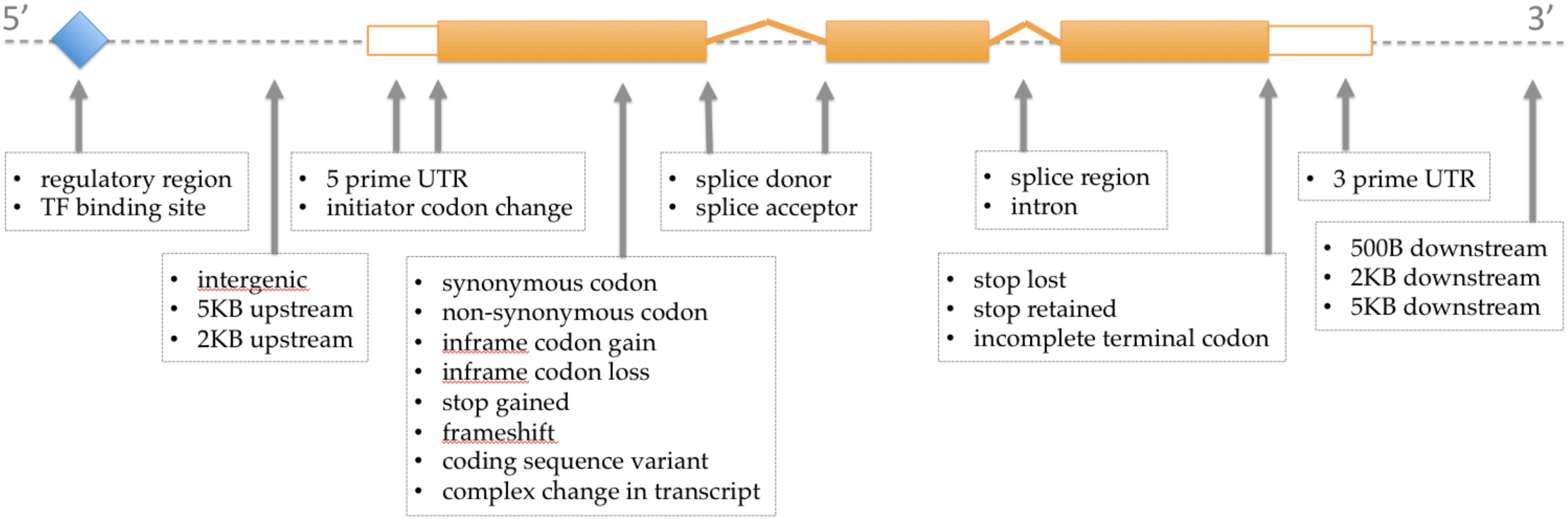
Output



Uploaded Variation	Location	ence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
rs10576	21:26965172	us_codon	915	873	291	P	ccA/ccG	rs10576	HGNC=MRPL39
rs10576	21:26965172	us_codon	852	843	281	P	ccA/ccG	rs10576	HGNC=MRPL39
rs10576	21:26965172	us_codon	915	873	291	P	ccA/ccG	rs10576	HGNC=MRPL39
rs1057885	21:26965205	us_codon	882	840	280	V	gtA/gtG	rs1057885	HGNC=MRPL39
rs1057885	21:26965205	us_codon	882	840	280	V	gtA/gtG	rs1057885	HGNC=MRPL39
rs1057885	21:26965205	us_codon	819	810	270	V	gtA/gtG	rs1057885	HGNC=MRPL39
rs113417859	21:40191548	gion_variant	-	-	-	-	-	rs113417859	-
rs113417859	21:40191548	ipt_variant, variant	-	-	-	-	-	rs113417859	HGNC=TMPRSS3
rs113417859	21:40191548	us_codon	1393	933	311	F	ttC/ttT	rs113417859	HGNC=ETS2
rs113417859	21:40191548	us_codon	1223	933	311	F	ttC/ttT	rs113417859	HGNC=ETS2
rs113417859	21:40191548	eam_variant	-	-	-	-	-	rs113417859	HGNC=ETS2
rs113417859	21:40191548	eam_variant	-	-	-	-	-	rs113417859	HGNC=ETS2
rs1135638	21:26965148	us_codon	939	897	299	G	ggC/ggT	rs1135638	HGNC=MRPL39
rs1135638	21:26965148	us_codon	876	867	289	G	ggC/ggT	rs1135638	HGNC=MRPL39
rs1135638	21:26965148	us_codon	939	897	299	G	ggC/ggT	rs1135638	HGNC=MRPL39
rs114053718	21:34029195	nous_codon	2722	2597	866	I/T	aTt/aCt	rs114053718	SIFT=deleterious(0); HGNC=SYNJ1
rs114053718	21:34029195	nous_codon	2714	2714	905	I/T	aTt/aCt	rs114053718	SIFT=deleterious(0); HGNC=SYNJ1
rs114053718	21:34029195	nous_codon	2597	2597	866	I/T	aTt/aCt	rs114053718	SIFT=deleterious(0); HGNC=SYNJ1
rs114053718	21:34029195	ipt_variant	384	-	-	-	-	rs114053718	HGNC=SYNJ1
rs114053718	21:34029195	nous_codon	2722	2714	905	I/T	aTt/aCt	rs114053718	SIFT=deleterious(0); HGNC=SYNJ1

Sequence Ontology consequences

- Provides a structured controlled vocabulary for the description of mutations at both the sequence and more gross level in the context of genomic databases



- regulatory region
- TF binding site

- intergenic
- 5KB upstream
- 2KB upstream

- 5 prime UTR
- initiator codon change

- synonymous codon
- non-synonymous codon
- inframe codon gain
- inframe codon loss
- stop gained
- frameshift
- coding sequence variant
- complex change in transcript

- splice donor
- splice acceptor

- splice region
- intron

- stop lost
- stop retained
- incomplete terminal codon

- 3 prime UTR

- 500B downstream
- 2KB downstream
- 5KB downstream

SIFT, PolyPhen and Condel in practice

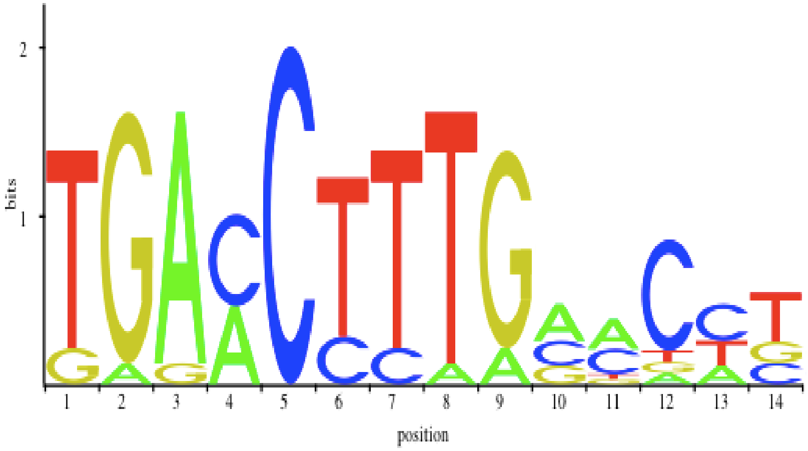
- Store every possible score for every* protein

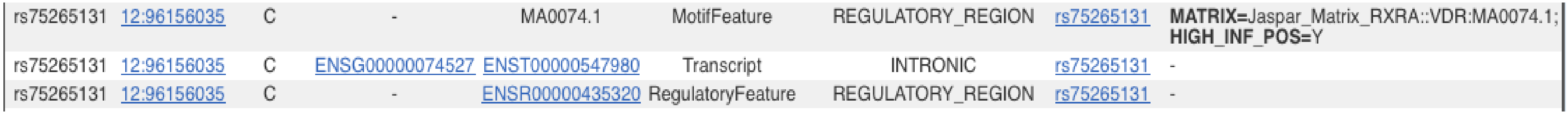
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1	0.001	0.047	0.007	0.007	0.007	0.002	0.047	0.001	0.002	0.001	-	0.007	0.007	0.007	0.007	0.002	0.002	0.001	0.094	0.017
2	0.081	0.547	0.547	0.348	0.201	0.348	0.817	0.081	0.348	-	0.348	0.547	0.547	0.547	0.547	0.201	0.201	0.081	0.817	0.547
3	0.007	0.191	0.007	0.002	0.094	0.017	0.094	0.047	0.002	0.017	0.094	0.017	0.017	-	0.007	0.007	0.017	0.017	0.191	0.047
4	0.017	0.362	0.201	0.106	0.106	0.106	0.362	0.017	0.106	0.017	0.201	0.362	0.201	0.362	0.362	0.106	0.04	-	0.677	0.201
5	0.017	0.362	0.201	0.106	0.106	0.106	0.362	0.017	0.106	0.017	0.201	0.362	0.201	0.362	0.362	0.106	0.04	-	0.677	0.201
6	0.007	0.191	0.007	0.002	0.094	0.017	0.094	0.047	0.002	0.017	0.094	0.017	0.017	-	0.007	0.007	0.017	0.017	0.191	0.047
7	0.081	0.817	0.035	-	0.547	0.081	0.547	0.547	0.081	0.201	0.547	0.201	0.201	0.081	0.201	0.081	0.081	0.201	0.817	0.547
8	0.663	0.99	0.964	0.964	0.964	-	0.99	0.964	0.964	0.964	0.99	0.922	0.964	0.964	0.964	0.848	0.964	0.964	0.99	0.99
9	0.081	0.817	0.081	0.081	0.547	0.081	0.348	0.547	0.081	0.201	0.547	-	0.348	0.201	0.201	0.081	0.081	0.201	0.817	0.547
…																				

- Condel scores are an algorithmic function of SIFT and Polyphen scores

Regulatory region consequences

- Variant within a regulatory feature = RegulatoryFeature
- Variant within a transcription factor binding motif = MotifFeature
- Variant in an “informative position” = HIGH_INF_POS





rs75265131	12:96156035	C	-	MA0074.1	MotifFeature	REGULATORY_REGION	rs75265131	**MATRIX=**Jaspar_Matrix_RXRA::VDR:MA0074.1; **HIGH_INF_POS=**Y
rs75265131	12:96156035	C	ENSG00000074527	ENST00000547980	Transcript	INTRONIC	rs75265131	-
rs75265131	12:96156035	C	-	ENSR00000435320	RegulatoryFeature	REGULATORY_REGION	rs75265131	-

Has this variant ever been seen before?

- Quickly becoming the most common question in human genomics
  - Incredibly hard to answer
- Nature said (in the October 2010 1000 Genomes issue) that about 2700 genomes had been sequenced and estimate 30,000 by the end of 2011
  - Beyond the those currently in the 1000 Genomes project (~2000)relatively few of these genomes are easily accessible
- There are many more exomes
  - Access here can be a problem as well

- Some data is available under controlled access and the fraction of data in this category is expected to increase

Future

- Ensembl is not a clinical decision support tool and only a fraction of the important resources were presented

- It does show the way forward
  - Comprehensive
  - Versioned
  - Standardized
  - Using controlled terminology
  - Regularly updated
  - Evidence based and algorithmic
  - Fully open

- There is uncertainty at every step in the process from the genome reference to the gene set to the interpretation and we have to work in this environment

Acknowledgements

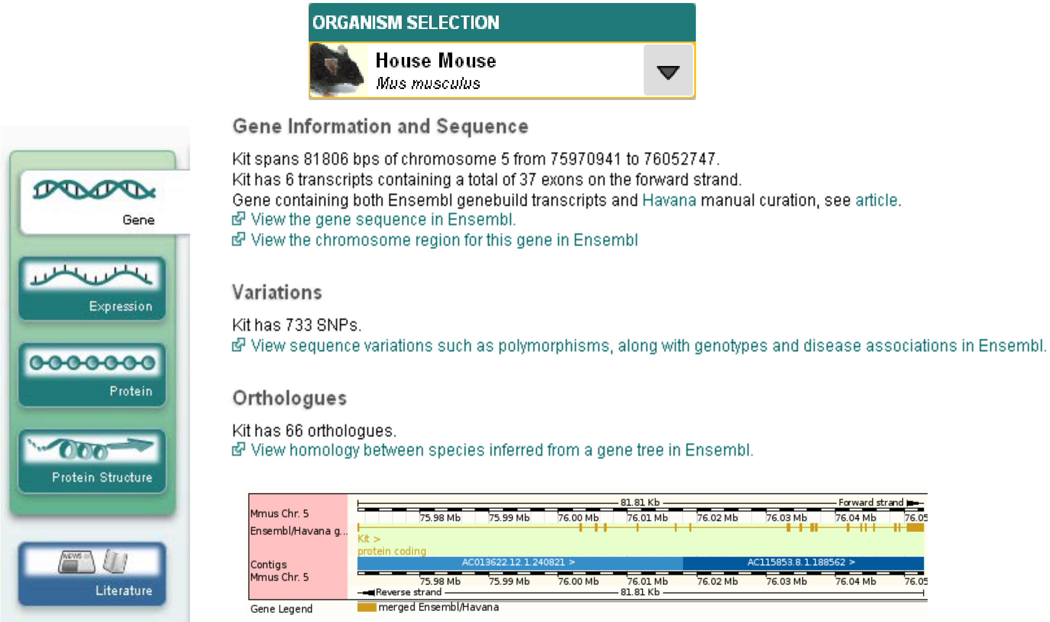
- Ensembl Annotation and VEP: Will McLaren, Graham Ritchie, Pontus Larsson, Daniel Sobral, Bethan Yates, Anne Parker, Jackie MacArthur, Fiona Cunningham
- EBI Variation Archives: Ilkka Lappalainen, Vasudev Kumanduri, Dylan Spalding, Mick Maguire, Lisa Skipper, Jeff Almeida-King

- Funding: Wellcome Trust, European Commission, NHGRI, British Heart Foundation, EMBL

EMBL-EBI

EBI data integration and added value

- EBI search provides integration into EBI existing spines (DAS based)
- Development of new spines diseases, cell type, tissue, tools
- User focussed design with general and specific user groups
- Added value - terminology, literature searching, pathways etc (user defined)
- Reciprocal integration between KOMP2 web portal and EBI resources



ORGANISM SELECTION

House Mouse
Mus musculus

Gene Information and Sequence

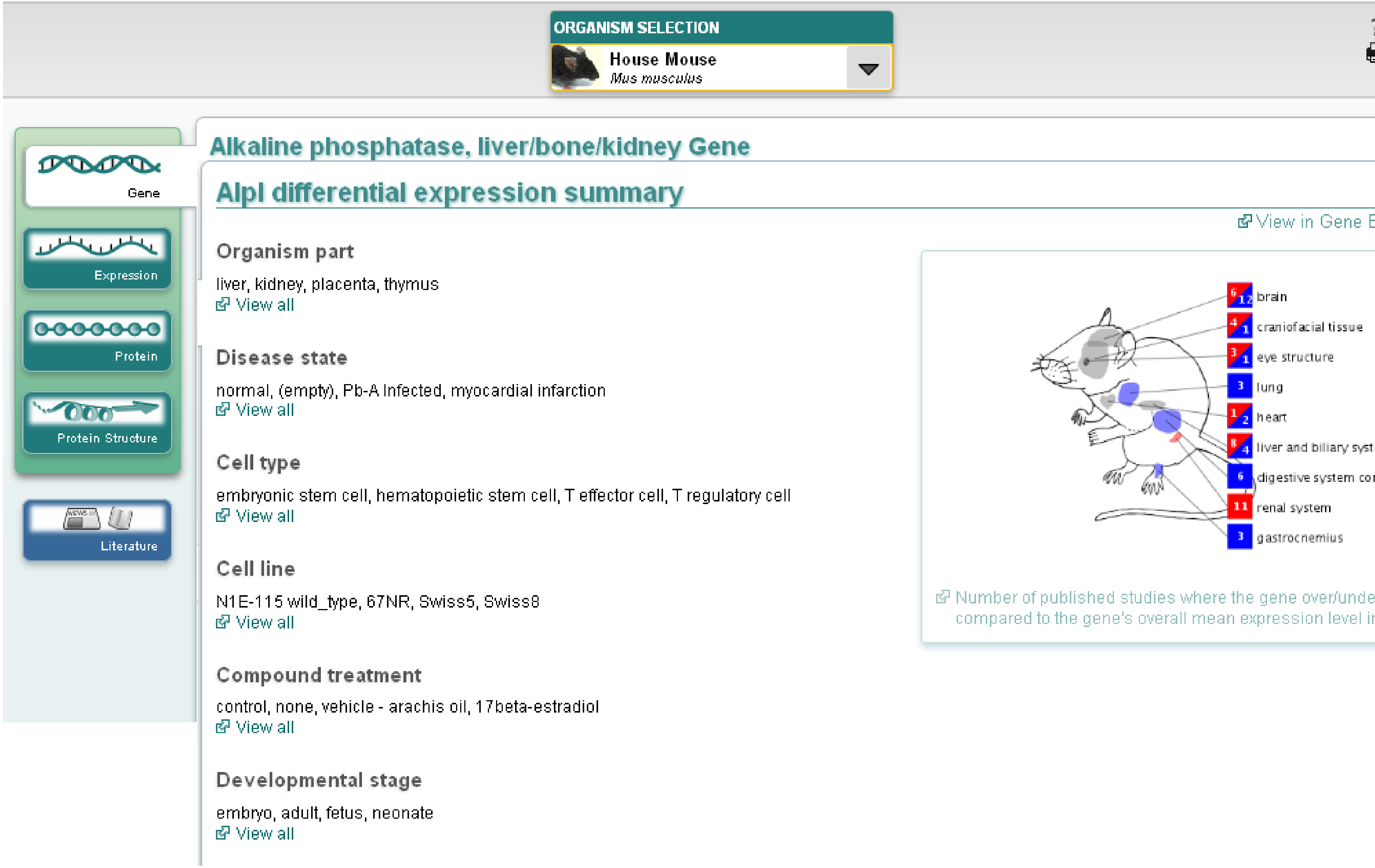
Kit spans 81806 bps of chromosome 5 from 75970941 to 76052747.
Kit has 6 transcripts containing a total of 37 exons on the forward strand.
Gene containing both Ensembl genebuild transcripts and Havana manual curation, see article.
View the gene sequence in Ensembl.
View the chromosome region for this gene in Ensembl

Variations

Kit has 733 SNPs.
View sequence variations such as polymorphisms, along with genotypes and disease associations in Ensembl.

Orthologues

Kit has 66 orthologues.
View homology between species inferred from a gene tree in Ensembl.



Alkaline phosphatase, liver/bone/kidney Gene

Alpl differential expression summary

View in Gene E

Organism part

liver, kidney, placenta, thymus
View all

Disease state

normal, (empty), Pb-A Infected, myocardial infarction
View all

Cell type

embryonic stem cell, hematopoietic stem cell, T effector cell, T regulatory cell
View all

Cell line

N1E-115 wild_type, 67NR, Swiss5, Swiss8
View all

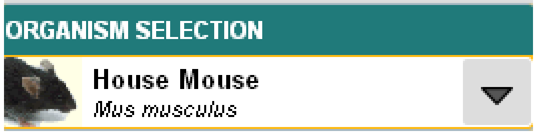
Compound treatment

control, none, vehicle - arachis oil, 17beta-estradiol
View all

Developmental stage

embryo, adult, fetus, neonate
View all

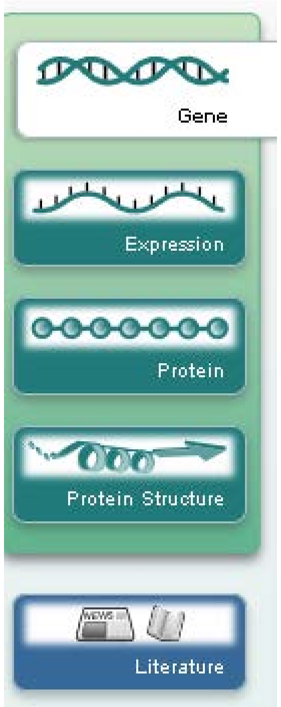
Number of published studies where the gene over/unde
compared to the gene's overall mean expression level i



ORGANISM SELECTION

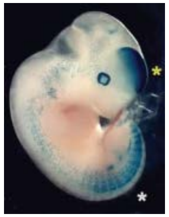
House Mouse

Mus musculus



KOMP2 Ensembl links

LacZ summaries, image links



Disease → Mouse models of disease, phenotype summaries

Pathways

Tissues → Expression summaries, phenotype links

Chemistry

Tools → Mouse knockouts, phenotype summaries, CDA links