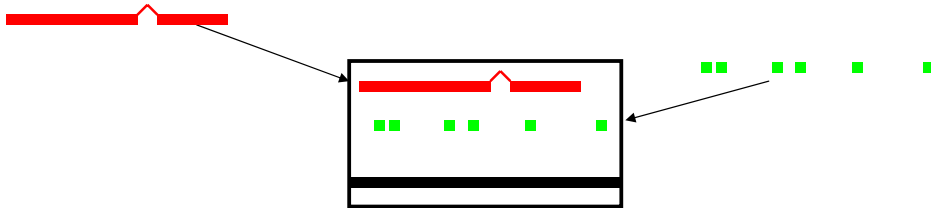


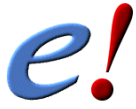
The Distributed Annotation System (DAS) Bringing Data Together



*Dr. Giulietta Spudich
Biosapiens 2007*

Overview

- What is DAS?
- How can I see external sources in Ensembl?
- How can I display my own data on Ensembl?



Genome annotation

Wikipedia:

Genome annotation is the process of attaching biological information to sequences. It consists of two main steps:

1. identifying elements on the genome, a process called Gene Finding, and
2. attaching biological information to these elements.



Genome annotation

Structural genome annotations: Usually positional, i.e. they have a start and stop coordinate and strand position on a reference sequence (*examples: gene, exon, regulatory sequence*).

Functional genome annotations: Usually non-positional, i.e. they are associated with a gene or other entity which in turn has a position on a reference sequence (*examples: biological or biochemical function*).

Protein annotations: Positional or non-positional (*examples: protein domain, biological function*).



Attach data from various sources...

with DAS
Distributed Annotation System

Developed in 1999/2000 by Lincoln Stein (CSHL) and collaborators
Biodas.org (<http://www.biodas.org>):

It allows a single machine (*client*) to gather up genome annotation information from multiple distant databases (*servers*), collate the information, and display it to the user in a single view (*such as ContigView, or BioSapiens DAS Portal*).



BioSapiens DAS Portal



A European Virtual Institute for Genome Annotation

Coordinated by  Funded by 

Home	DAS Portal	Partners	Training	Work Packages	Meetings	News	Restricted Area	Contact
------	------------	----------	----------	---------------	----------	------	-----------------	---------



DAS Resources

- » [BioSapiensDIR](#)
- » [Protein Annotation Ontology](#)

DAS Clients

- » [SPICE Home](#)
- » [Dasty](#)
- » [Ensembl Home](#)

BioSapiens DAS Portal - Alpha Version

This is the BioSapiens DAS portal, the central access point to the annotations provided by the 24 partners in the BioSapiens project, by means of their own DAS servers. The portal allows visitors to view protein annotations, genomic annotations, and structural annotations.

For protein annotation, UniProt accession numbers are currently used. For human genomic annotations, the NCBI35 assembly is used.

For help and information on the portal click [here](#)

Input Accession:


Type

[Advanced DAS Settings](#)

Search Uniprot

Enter search term:

NB. This is a simple alltext search against all of Uniprot. For more complex searching, consider using the [SRS search engine](#) or the [Uniprot site](#).

 The BioSapiens project is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health," contract number LSHG-CT-2003-503265.



Genome annotation

Genome annotations, especially structural annotations, are often stored in **GFF** (General Feature Format) files.

GFF is a simple tab-delimited data format.

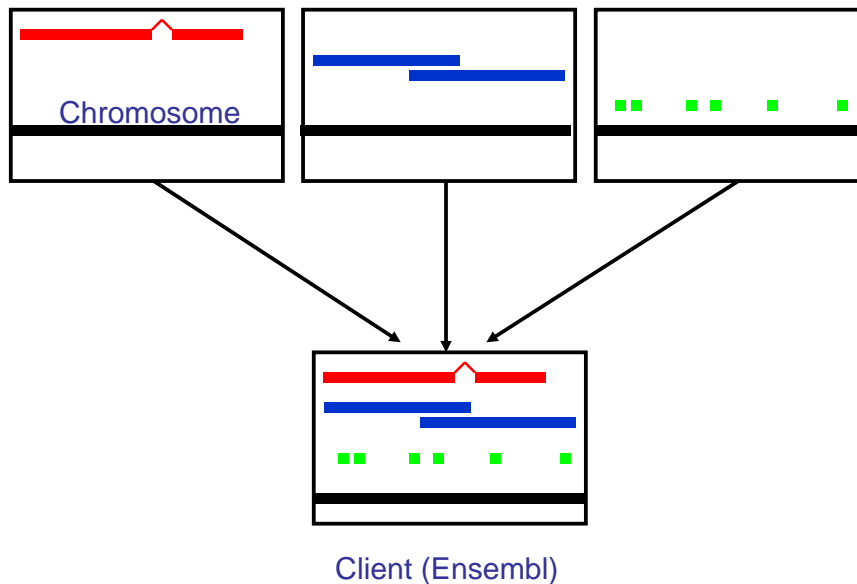
Name	Source	Feature	Start	End	Score	Strand	Frame
SEQ1	EMBL	atg	103	105	.	+	0
SEQ1	EMBL	exon	103	172	.	+	0
SEQ1	EMBL	splice5	172	173	.	+	.
SEQ1	netgene	splice5	172	173	0.94	+	.
SEQ1	genie	sp5-20	163	182	2.3	+	.
SEQ1	genie	sp5-10	168	177	2.1	+	.
SEQ2	grail	ATG	17	19	2.1	-	0

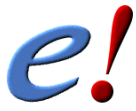
See also: <http://www.sanger.ac.uk/Software/formats/GFF/>



DAS Concept

Server 1 (Genes) Server 2 (DNase I sites) Server 3 (Variations)





DAS Advantages

- Data may be shared more easily.
- The amount of data which needs to be stored locally is decreased.
- Responsibility for updating and maintaining the data is left with the original data provider.
- Conflicting annotations are permitted, encouraging information dissension and dialogue.
- Annotation can be stored in a variety of formats.

8 of 27



DAS Server

Three main categories of DAS data sources:

- **GenomeDAS:** Attach features to chromosomes, super-contigs, contigs or scaffolds.
- **GeneDAS** Attach annotations to gene IDs. (Is not 'locked to' a position on the assembly).
- **ProteinDAS:** Annotating positional features on proteins.

9 of 27



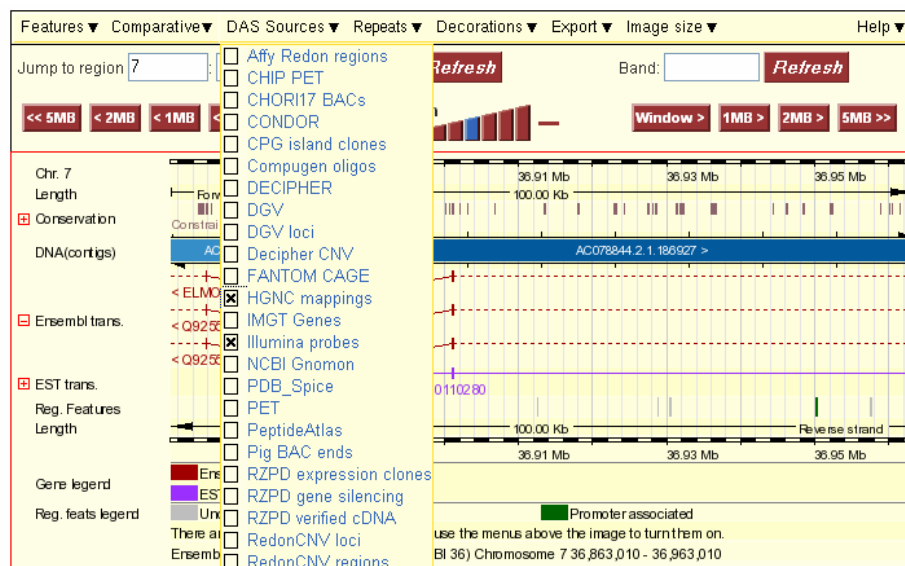
DAS in Ensembl

The following views in Ensembl allow for adding of external data:

- **ContigView:** Displays the genomic sequence in varying degrees of resolution, with things like transcripts, markers, SNPs etc. as features.
- **GeneView:** Displays information about one particular gene.
- **ProtView:** Displays information about one particular peptide.



Pre-configured DAS Sources ContigView



The screenshot displays the Ensembl ContigView interface. At the top, there are navigation tabs: Features, Comparative, DAS Sources, Repeats, Decorations, Export, Image size, and Help. The 'DAS Sources' tab is active, showing a list of pre-configured data sources with checkboxes. The 'Jump to region' field is set to '7'. Below the list, there are navigation buttons for window size: << 5MB, < 2MB, < 1MB, and >> 5MB. The main view shows a genomic track for Chromosome 7, with a 100.00 Kb scale. The track includes various features such as transcripts, conservation, and regulatory elements. A specific region is highlighted with a blue bar, labeled 'AC078844.2.1.186927 >'. The interface also includes a 'Refresh' button and a 'Band' input field.

Category	Feature	Checked	
DAS Sources	Affy Redon regions	<input type="checkbox"/>	
	CHIP PET	<input type="checkbox"/>	
	CHORI17 BACs	<input type="checkbox"/>	
	CONDOR	<input type="checkbox"/>	
	CPG island clones	<input type="checkbox"/>	
	Compugen oligos	<input type="checkbox"/>	
	DECIPHER	<input type="checkbox"/>	
	DGV	<input type="checkbox"/>	
	DGV loci	<input type="checkbox"/>	
	Decipher CNV	<input type="checkbox"/>	
Conservation	FANTOM CAGE	<input type="checkbox"/>	
	ELMO	<input type="checkbox"/>	
	HGNC mappings	<input checked="" type="checkbox"/>	
	IMGT Genes	<input type="checkbox"/>	
	ILLUMINA	<input checked="" type="checkbox"/>	
	NCBI Gnomon	<input type="checkbox"/>	
	PDB_Spice	<input type="checkbox"/>	
	PET	<input type="checkbox"/>	
	PeptideAtlas	<input type="checkbox"/>	
	Pig BAC ends	<input type="checkbox"/>	
Gene legend	RZPD expression clones	<input type="checkbox"/>	
	RZPD gene silencing	<input type="checkbox"/>	
	RZPD verified cDNA	<input type="checkbox"/>	
	RedonCNV loci	<input type="checkbox"/>	
	RedonCNV regions	<input type="checkbox"/>	
	Reg. feats legend	Enr	<input type="checkbox"/>
		ES	<input type="checkbox"/>
		Unr	<input type="checkbox"/>
		There are	<input type="checkbox"/>
		Ensembl	<input type="checkbox"/>
Promoter associated			



ContigView

Features ▾ Comparative ▾ DAS Sources ▾ Repeats ▾ Decorations ▾ Export ▾ Image size ▾ Help ▾

Jump to region 7 : 36863010 - 36963010 Refresh Band: Refresh

<< 5MB < 2MB < 1MB < Window + Zoom Window > 1MB > 2MB > 5MB >>

Chr. 7
Length
Conservation
DNA(contigs)
Ensembl trans.
EST trans.
Reg. Features
HGNC mappings
Illumina probes
Length
Gene legend
Reg. feats legend

There are currently 124 tracks switched off, use the menus above the image to turn them on.
Ensembl Homo sapiens version 46.36h (NCBI 36) Chromosome 7 36,863,010 - 36,963,010



DAS in GeneView

Gene DAS Report

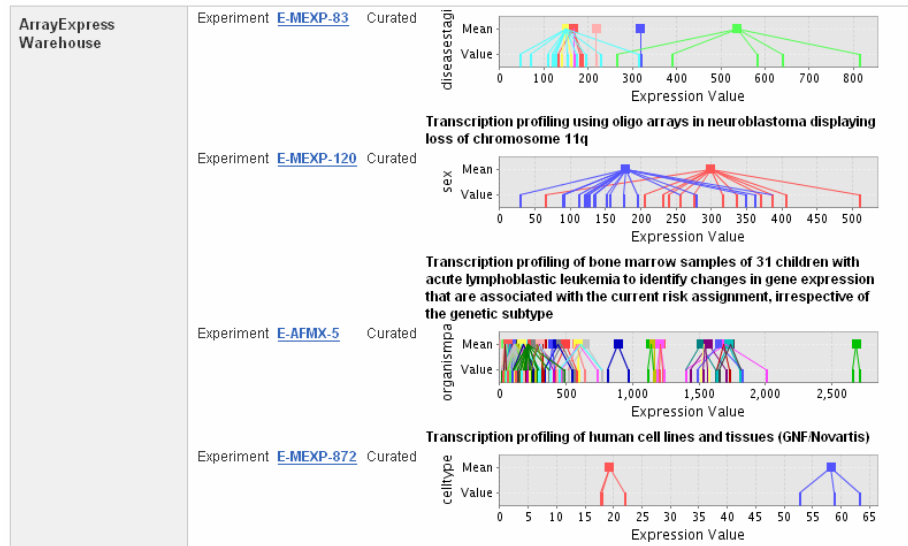
DAS Sources

- [AltSplice](#) (Alternative splice database)
- [AltTrans](#) (Alternative Transcript Diversity Database)
- [ArrayExpress Warehouse](#) (Gene Expression Profiles Database)
- [GAD](#) (Genetic Association Database)
- [HGNC](#) (HUGO Gene Nomenclature Committee)
- [HUGO_text](#) (PubMed text-mining via HGNC symbol)
- [OMA](#) (Orthology Prediction from ETH Zurich)
- [Pancreatic Expression](#) (Pancreatic Expression Database)
- Phenotypes (Associated directly or via orthologues or protein families)
- [Protonet](#) (Global classification of proteins into hierarchical clusters)
- [RZPD verif. cDNA](#) (RZPD sequence verified non-redundant cDNA clone sets)
- [RZPD esiRNA](#) (RZPD gene silencing (RNAi) resources)
- [RZPD Prot Exp](#) (RZPD clones ready for protein expression)
- [Reactome](#) (Knowledgebase of biological processes)
- [UniProt](#) (Protein knowledgebase)
- [aewTest](#) (Test AEW DAS Source)

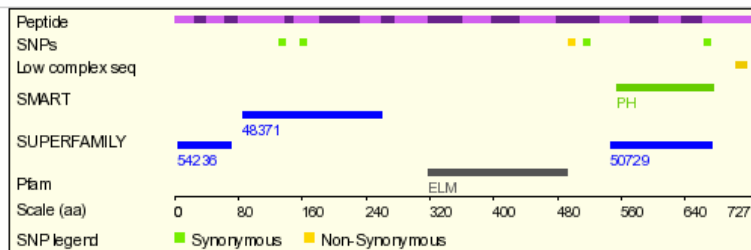
Update
Manage Sources



GeneView



DAS in ProtView

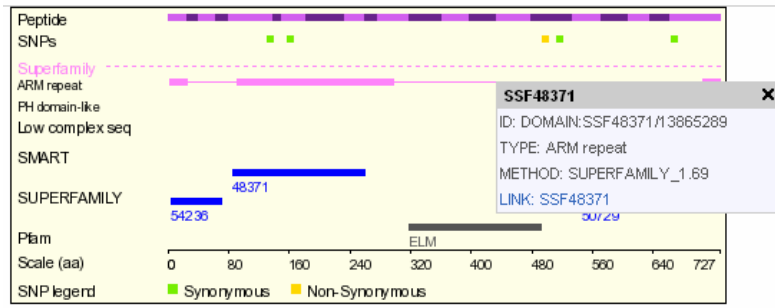


Protein DAS Report

- DAS Sources
- [ArrayExpress Warehouse](#) (Gene Expression Profiles Database)
 - [CBS_func](#) (CBS Protein function and structure predictions)
 - [CBS_ptm](#) (CBS Post-translational modification site predictions)
 - [CBS_sort](#) (CBS Protein sorting predictions)
 - [HGNC](#) (HUGO Gene Nomenclature Committee)
 - [PDB_Splice](#) (Mapping of ENSP protein sequences to protein structures. Based on Compara, MSD.)
 - [Protonet](#) (Global classification of proteins into hierarchical clusters)
 - [Superfamily](#) (structural domain assignments to protein sequences)
 - [UniProt](#) (Protein knowledgebase)
- [Update](#)
- [Manage Sources](#)



ProtView



DAS Sources

- [ArrayExpress Warehouse](#) (Gene Expression Profiles Database)
- [CBS_func](#) (CBS Protein function and structure predictions)
- [CBS_ptm](#) (CBS Post-translational modification site predictions)
- [CBS_sort](#) (CBS Protein sorting predictions)
- [HGNC](#) (HUGO Gene Nomenclature Committee)
- [PDB_Spice](#) (Mapping of ENSP protein sequences to protein structures. Based on Compara, MSD)
- [Protonet](#) (Global classification of proteins into hierarchical clusters)
- [Superfamily](#) (structural domain assignments to protein sequences)
- [UniProt](#) (Protein knowledgebase)

Update

Manage Sources



Where are the sources?

- Click 'Manage sources' for a list of URLs...

DAS sources

Level of annotation...

Name	DAS Server	Data Source	Coordinate System
ArrayExpress	http://www.ebi.ac.uk/microarray-as/aew/das	aew	Ensembl Gene ID
CBS_func	http://genome.cbs.dtu.dk:9000/das	cbs_func	UniProt/Swiss-Prot Acc
CBS_ptm	http://genome.cbs.dtu.dk:9000/das	cbs_ptm	UniProt/Swiss-Prot Acc
CBS_sort	http://genome.cbs.dtu.dk:9000/das	cbs_sort	UniProt/Swiss-Prot Acc
HGNC	http://onyx.gene.ucl.ac.uk:9000/das	HGNC	Entrez Gene ID
PDB_Spice	http://das.sanger.ac.uk/das	ensppdbmapping	Ensembl Peptide ID
Protonet	http://www.protonet.cs.huji.ac.il/das	protonet	UniProt/Swiss-Prot Acc
Superfam	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/cgi-bin/das	hs	Ensembl Peptide ID
UniProt	http://www.ebi.ac.uk/das-sv/uniProt/das	aristotle	UniProt/Swiss-Prot Acc+UniProt/TrEMBL



Display your own data

- By **uploading data** into an internal DAS source within Ensembl
 - For small datasets
- By **attaching new external DAS sources**
 - For larger datasets
 - Allows viewing of the data in other DAS clients
 - Allows usage of the data by other applications
- By **URL-based upload**
 - For small datasets
 - Only ContigView/CytoView



18 of 27



Display Your Own Data

The **'Manage sources'** options in ContigView, GeneView and ProtView lead you to DasConView:

The diagram illustrates the workflow from Ensembl's 'Manage sources' options to the DasConView interface. On the left, two screenshots show the 'Manage sources' menu in a browser-based genome viewer (top) and a desktop application (bottom). Arrows point from these menus to the 'Manage Sources' section of the 'e! Ensembl Human DasconView' interface. This interface includes options to 'Add Data Source' and 'Upload your data', and a table of existing DAS sources.

DAS sources	
Name	DAS
das_COMPUGEN_36	http
das_CPG_36	http
das_Decipher	http
das_GNOMON_36	http
das_HVER_36	http

19 of 27



URL-based Upload

Based on the custom annotation track system of the UCSC browser.

Allowed formats: GFF, GTF, BED, PSL (see also <http://genome.ucsc.edu/goldenPath/help/customTrack.html>).

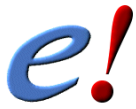
Display data by entering the URL of the data file in the 'DAS Sources' menu (ContigView)



URL-based Upload

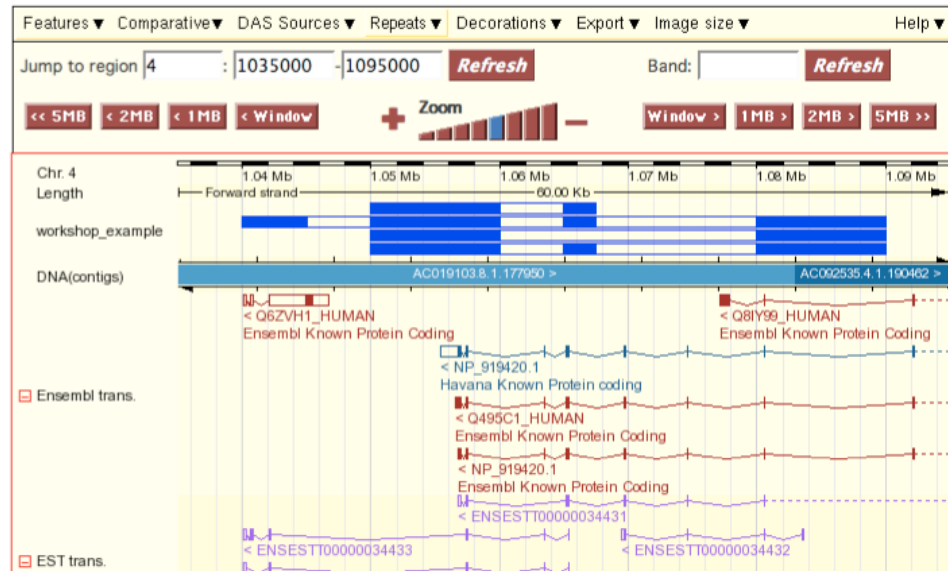
```
browser position chr4:1035000-1095000
track name=workshop_example description="Workshop example" color=blue
url=http://www.ebi.ac.uk/~bert/workshop_example.txt
4     exons  exon_1  1040000 1045000 .      +      .      transcript_1
4     exons  exon_2  1050000 1060000 .      +      .      transcript_1
4     exons  exon_3  1065000 1067500 .      +      .      transcript_1
4     exons  exon_4  1080000 1090000 .      +      .      transcript_1
4     exons  exon_2  1050000 1060000 .      +      .      transcript_2
4     exons  exon_3  1065000 1067500 .      +      .      transcript_2
4     exons  exon_4  1080000 1090000 .      +      .      transcript_2
4     exons  exon_2  1050000 1060000 .      +      .      transcript_3
4     exons  exon_3  1065000 1067500 .      +      .      transcript_3
4     exons  exon_2  1050000 1060000 .      +      .      transcript_4
4     exons  exon_4  1080000 1090000 .      +      .      transcript_4
```

<sequence><source><feature><start><end><score><strand><frame><group>



Example: URL-based Upload

http://www.ebi.ac.uk/~bert/workshop_example.txt



DAS Registry

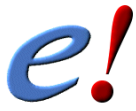
Is a directory of available DAS servers.

Performs regular “health checks” and may contact maintainers of a sources that are down for more than a certain time.

Provides up-time statistics for individual sources.

Standardizes coordinate systems used by the servers.

<http://www.dasregistry.org/>



Ensembl as a DAS source

<http://www.ensembl.org/das/sources>

Genes, karyotypes, other annotation
across species

The reference is the assembly.

24 of 27



DAS - Servers

Dazzle, a server written by Thomas Down (Sanger Institute) in Java, running on Tomcat or Resin.

<http://www.derkholm.net/thomas/dazzle>

ProServer, a light-weight server written by Roger Pettett (Sanger Institute) in Perl.

<http://www.sanger.ac.uk/proserver/>

LDAS, the original server implementation by Lincoln Stein (CSHL), written in Perl.

<http://ww.biodas.org/servers/>

Diplo, a light-weight and modular server written by Andreas Kähäri (EBI) in Perl.

<http://www.ebi.ac.uk/~ak/diplo/>

25 of 27



DAS

For more information:

http://www.ensembl.org/info/data/external_data/das/index.html

For questions and problems:

helpdesk@ensembl.org

das@ebi.ac.uk

26 of 27



Q&A

QUESTIONS

ANSWERS

27 of 27