



Paul Kersey (version1, 29/01/2008)

Integr8 and Genome Reviews

The Integr8 web portal provides easy access to integrated information about deciphered genomes and their corresponding proteomes, incorporating and analysing data from many primary data resources within a single coherent framework. Among the resources presented through Integr8 is Genome Reviews, a database of complete genomes focused on bacteria, archaea, and lower eukaryotes.

You will learn about:

1. What bioinformatics resources are available for complete genome and proteome data, and how Integr8 and Genome Reviews fit into this picture
2. How to query Integr8
3. How to download data from Integr8
4. How to compare complete proteomes using Integr8
5. How to explore relationships between genes using Integr8
6. How to browse complete genomes data in Genome Reviews

Contents:

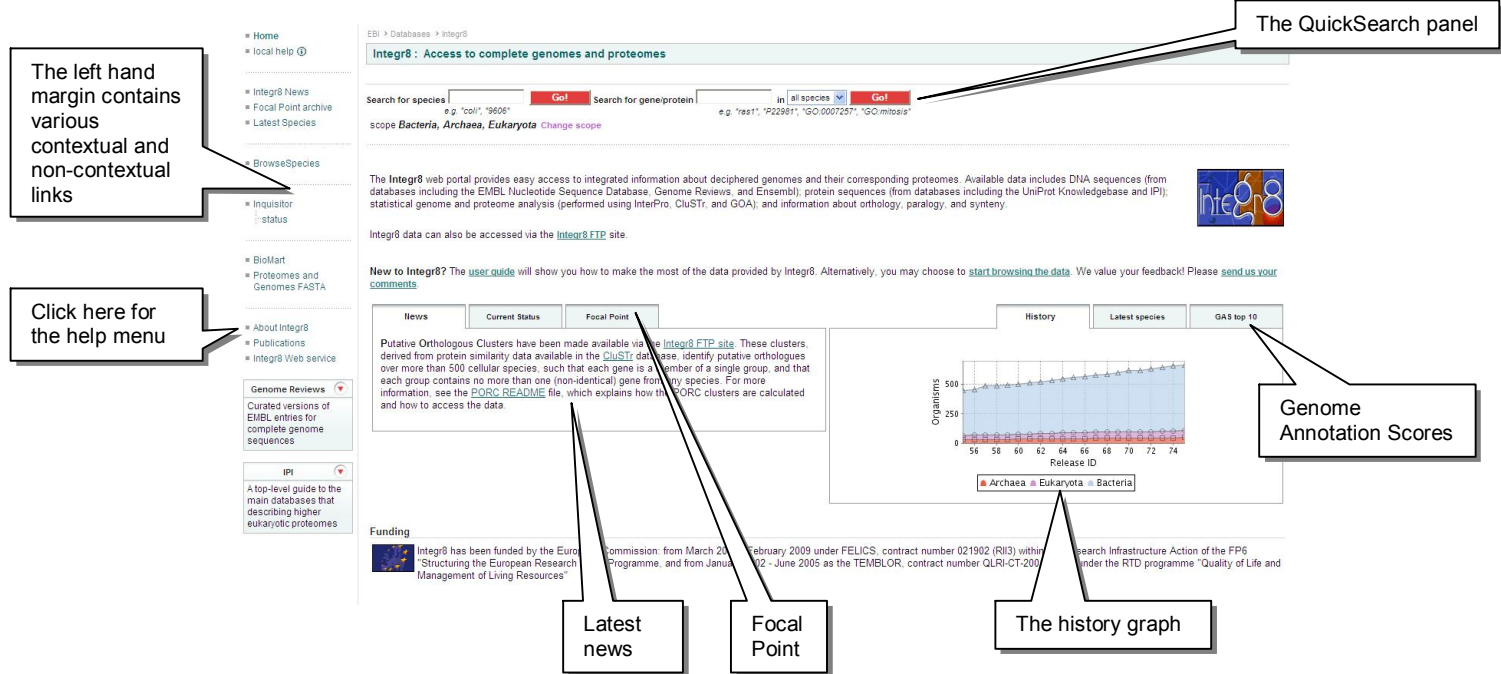
1. The Integr8 home page
2. How to query Integr8
3. Complete proteomes data in Integr8
4. The Integr8 gene view page
5. Browsing complete genomes in Genome Reviews

1 The Integr8 home page

Integr8 (Kersey et al, 2005) offers a single portal for interactive access to data from species with completely deciphered genomes. You can find the Integr8 home page at <http://www.ebi.ac.uk/integr8>.

Figure 1 shows the Integr8 home page and some of the features immediately available. Have a look around and see what some of these features are.





The left hand margin contains various contextual and non-contextual links

The QuickSearch panel

Click here for the help menu

Latest news

Focal Point

The history graph

Genome Annotation Scores

Fig. 1

The Integr8 home page. A number of different features are highlighted. The QuickSearch panel will be considered in more detail in the following section. The history graph displays the number of species in each release of Integr8, broken down by superregnum. The Focal Point is a short article highlighting a feature of the portal, and directing users to more detailed documentation. The Genome Annotation Score is a measure of how complete the annotation of a genome in Integr8 is. Various links are provided in the left hand menu; some of these change, according to context and selection. The “About Integr8” link is always provided, and leads to the Integr8 help menu. Integr8 currently contains data for over 650 cellular organisms (over 600 of which are bacterial or archaeal); and additionally, over 500 bacteriophage.

2 How to query Integr8

Searching in Integr8 revolves around 2 basic concepts; the gene and the species. Often, it makes sense to search for a combination of gene and species. Let’s begin by considering a species-centric search.

1. Open the Integr8 home page, at <http://www.ebi.ac.uk/integr8>, in a Web browser. You’ll see a panel at the top of the page, with two search boxes in it. This is the Integr8 QuickSearch form and it is always accessible when using the Integr8 application. Occasionally, the form collapses, but you can always expand it by clicking on the “show quicksearch” link that appears in it’s place.
2. Try typing “Homo sapiens” (or any other species name) into the species search box on the left hand side of the QuickSearch menu. As you type, an option list may appear, allowing you to select your species of interest without further typing.
3. Click the red “Go!” button to the right of the search box.
4. If you have typed/ selected the complete name of a single species, a page of text describing that species should now appear. If not, a list of potential species matching your query will appear, and you can select the right one by clicking on the species name. For example, try typing “Escherichia”, without using to auto-complete, to see what this page looks like. After you click on the appropriate link, the species home page will appear.

When you select a species, a number of things happen. Firstly, your selected species is now displayed in the QuickSearch panel, beneath the search box you have just used. This will now define the default scope of all subsequent text or sequence searches until you change it.

Secondly, in the left hand margin, a number of new links appear, relating to resources containing data belonging to this species. We'll come back to these links later on.

The screenshot shows the Integr8 QuickSearch interface. At the top, there is a search bar with 'Escherichia' entered and a 'Go!' button. Below the search bar, the selected species is 'E.coli K12' with a 'change scope' link. A list of species names is displayed, including 'Escherichia coli O9:H4 (strain HS)', 'Escherichia coli O139:H28 (strain E24377A / E...', 'Escherichia coli O1:K1 / APEC', 'Escherichia coli (strain K12 / ATCC 27325 / DSM 5911 / ...', 'Escherichia coli O6 (strain UPEC / O6:H1 / ATCC 700928 / CP...', and 'Escherichia coli (strain K42)'. A left-hand menu is visible, showing 'E.coli K12' selected, with sub-options for 'Literature', 'Genome Statistics', 'Proteome Analysis', 'Downloads', and 'Taxonomy'. Three callout boxes provide instructions: one points to the left-hand menu, another points to a species name in the list, and a third points to the search box.

Fig. 2

The results of a textual species query in Integr8. The QuickSearch panel is visible at the top of the page. The currently selected species is given below this, along with the option to change the scope of the search. A list of potential species matching the search term appears in the main window (there are currently 10 different strains of Escherichia coli present in the database. As a species has already been selected in this example, a list of species-specific options is available in the left-hand menu.

By default, species searches in Integr8 are restricted to cellular organisms, but it is also possible to search bacteriophage genomes. To do this, it is necessary to change the scope.

1. Click on the “change scope” link immediately below the species search box.
2. Select bacteriophage.
3. You will now see the Integr8 home page again, but with a difference. Look at the history chart in the lower right hand corner. This now shows the number of bacteriophage present in each release (previously it showed statistics for cellular organisms).
4. The search features available for bacteriophage mirror those available for cellular organisms, but are presented separately owing to the large number of bacteriophage genomes available. For now, let's switch back to cellular organisms (again, use the “change scope” link) and look at other ways one can search.

Of course, sometimes you want to select a species after first seeing what's available. In Integr8, you can do this through the “Browse Species” link in the left hand menu. Click on this now. You can see a list of all species (within the currently defined scope) whose genomes have been sequenced and which are available in Integr8, grouped alphabetically. Click on the name of your species of interest to select it.

You can also browse these species taxonomically. Click on the “taxonomy browser” link at the top of the page. The taxonomy browser shows the distribution of species among the superregna in the form of a pie chart. Click on the chart, or on the adjacent labels, to drill down the tree and see the next layer of the taxonomy.

EMBL-EBI www.ebi.ac.uk

All organisms (665) >> Bacteria (556) >> Firmicutes (128) >> Lactobacillales >>

- Enterococcus (1)
- Lactobacillaceae (13)
- Oenococcus (1)
- Leuconostoc (1)
- Streptococcaceae (29)
- Symbiobacterium (1)

Species name: << 1-20 of 45 >> show all

Enterococcus faecalis (strain V583 / ATCC 700802)	GR
Lactobacillus acidophilus (strain NCFM)	GR
Lactobacillus brevis (strain ATCC 367 / JCM 1170)	GR
Lactobacillus casei (strain ATCC 334)	GR
Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC BAA-365)	GR
Lactobacillus delbrueckii (subsp. bulgaricus, strain ATCC 11842 / DSM 20081)	GR
Lactobacillus gasseri (strain ATCC 33323 / DSM 20243)	GR

Fig. 3

Using the Integr8 taxonomy browser.

Use QuickSearch or the browser to select “*Escherichia coli K12*” as your species of interest. Let’s now look at gene search. Genes can be searched for in the right hand box in the QuickSearch panel. Try typing “ftsZ” in this box.

The first thing you should notice is that the search is set up automatically to occur only in the species you have selected (a menu item to the right of the search box indicates this). Pull down the menu for further options. The “all species” option is instantly available. But it is also possible to restrict your search to any selected portion of the taxonomic tree. Click on “specify”, then on the adjacent “Go!” button, to see what happens.

Search for species Go! Search for gene/protein ftsZ in Go!

Selected species *E.coli K12* Change scope

Escherichia coli (strain K12) - Tax ID: 83333 GAS: ★★★★★ ⓘ

Currently selected species

Searching with this text

Use pull-down menu to change scope of search

Default scope of search is currently selected species

Fig. 4

Selecting the scope of a gene search.

You will now see a familiar looking interface for browsing the taxonomic tree. However, at any node of the taxonomic tree, you can now select all organisms located beneath that node for search, by clicking on the “select all” link. Species from any number of nodes can be selected, or, once selected, individually removed from the search set. When you have chosen whatever species you wish to search in, simply click “continue” at the top of the page. Try performing a search for ftsZ in “gammaproteobacteria” now.

EMBL-EBI www.ebi.ac.uk

show quicksearch

Specify the species within which you wish to search for 'ftsZ' then continue

QuickSearch is collapsed by default on this view; click here to expand

Browse the taxonomic tree to find nodes of interest

Species under the current node can be selected individually or wholesale

Click here once done to perform your search

Selected species can be removed individually or wholesale

All organisms (665) >> Bacteria (556) >> Proteobacteria (296) >> Gammaproteobacteria >>

- Methylococcus (1)
- Aeromonas (2)
- Xanthomonadaceae (8)
- Pasteurellaceae (10)
- Dichelobacter (1)
- Vibrionaceae (8)
- Enterobacteriaceae (45)
- Alteromonadales (21)
- Oceanospirillales (4)
- Chromatiales (3)
- Pseudomonadales (18)
- Candidatus Carsonella (1)
- Legionellales (5)
- Candidatus Baumannia (1)
- Thiotrichales (8)
- sulfur-oxidizing symbionts (2)

<< 1-20 of 138 >>	select all	Selected species: (138)	remove all
Acinetobacter baumannii	selected	Acinetobacter baumannii	remove
Acinetobacter sp.	selected	Acinetobacter sp.	remove
Actinobacillus pleuropneumoniae	selected	Actinobacillus pleuropneumoniae	remove
Actinobacillus succinogenes 130Z	selected	Actinobacillus succinogenes 130Z	remove
Aeromonas hydrophila	selected	Aeromonas hydrophila	remove

Fig. 5

Specifying the taxonomic scope of a query.

Let's now try another search. Supposing we want to find genes linked to mitosis. Potentially, there is a problem in performing a search based on annotation; different genes may be annotated with different terms meaning the same thing, or more/less specific terms. Fortunately, this problem can be solved by searching using the Gene Ontology (GO). GO terms constitute a hierarchical, controlled vocabulary, so by searching for a GO term, it is possible to find all genes annotated with this term, or a term identifying a more specific variant of the search concept. If you don't know the GO term that corresponds to the concept that you want to search for, Integr8 will find it for you. Try typing "GO:" into the gene search box, and then begin typing "mitosis".

Search for gene/protein in

Prefix your search term with "GO:", and an auto complete menu will help you find the GO term you want

Fig. 6

Autocomplete with GO search. If your query is started with the prefix "GO:" an autocomplete menu will offer you a list of GO terms to select for query. Searching with GO is a reliable way to find genes corresponding to your search term at different levels of specificity.

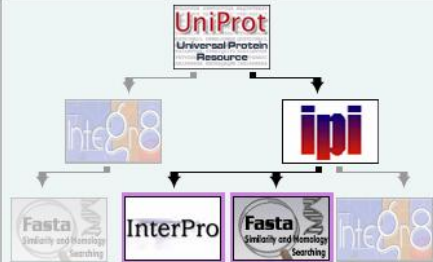
However you perform your search, you will be taken to the gene search results page. We'll come back to this page shortly, but first, let's have a look at sequence search. The Integr8 sequence search tool is called the "Inquisitor": you'll find it in the left-hand margin. Click on this link now. A box should appear for entering your query string. But another box, to the right, indicates the scope of your search. By default, this is remembered from your previous search. If you want to search within a different taxonomic scope, click on the link above the query box, which allows you to specify the taxonomic range. You should find the interface for doing so familiar. Try searching among the *Deinococci*, a group of radiation-sensitive bacteria.

Here's a sequence you can try pasting into the search box.

```
MFPEMELTND AVIKVIGVGG GGGNAVEH MV RERIEGVEFF AVNTDAQALR KTAVGQTIQI
GSGITKGLGA GANPEVGRNA ADEDRDALRA ALEGADMVFI AAGMGGGTGT GAAPVVAEVA
KDLGILTAVV VTKPFNFEGK KRMAFAEQGI TELS KHVDSL ITIPNDKLLK VLGRGISLLD
AFGAANDVLK GAVQGIAELI TRPGLMNVD F ADVRTVMSEM GYAMMGSGVA SGEDRAEEAA
EMAISSPLLE DIDLSGARGV LVNITAGFDL RLDEFETVGN TIRAFASDNA TVVIGTSLDP
DMNDEL RVTV VATGIGMDKR PEITLV TNKQ VQQPVMDRYQ QHGMAPLTQE QKPVAKVVND
NAPQTAKEPD YLDIPAFLRK
```

Then click on submit. A page appears that will show you the current status of your query.

Inquisitor job id: 16 [Return to main status page](#)
Sequence: MFPEMELTND AVIK...



Your protein sequence is not known in the UniProt Knowledgebase or in Integr8.

The Inquisitor is performing a search to identify similar sequences to your query sequence in completed proteomes. The search uses the FASTA algorithm.

Your sequence is also being analysed with InterProScan, to identify protein domains, families and functional sites.

Inquisitor status - Running

Job	Status	Time running (secs)	Time limit (secs)
Proteomes FASTA	Running	20	1200
InterProScan	Running	20	1200

Fig. 7

The Inquisitor job monitoring page. The inquisitor performs a FASTA (global alignment) search against sequences from all selected species. Additionally, it runs InterProScan to analyse the domain composition of your sequence. But to save time, a quick lookup of known results is performed first. This page shows the progress of your running job.

It can take some time for an Inquisitor analysis to complete, but you can continue to browse the Integr8 site, or even submit another Inquisitor query, while your job is running. Once you have submitted a query, an Inquisitor "status" link appears in the left hand menu. Click on this at any time to return to your original search.

When your job is reported as finished, you will be able to access a series of results pages. Each one summarises information about the nearest match to your query sequence in Integr8 (sorted by sequence similarity). Additionally, an InterProScan analysis of your own query sequence is provided.

Match 2 (partial sequence match)

Cell division protein FtsZ in species *Deinococcus geothermalis* has 46.216% identity to the query sequence (over 361 amino acids). View the full [FASTA report](#)

Gene information

Gene [Dgeo_1635](#) from [Deinococcus geothermalis](#).

Full length mRNAs encoding Cell division protein FtsZ have not been sequenced.

Protein information

The protein product of this gene is Cell division protein FtsZ (represented in UniProtKB by Q1IXV4). View [InterProScan](#)

[UniProt](#) [InterPro](#) [Inte8](#)

Protein classification according to InterPro

This sequence belongs to the families:
[Carbohydrate kinase, PfkB](#) [Cell division protein FtsZ](#)
 and contains the following domains:
[Tubulin-FtsZ, C-terminal](#) [Tubulin-FtsZ](#), [GTPase](#)

InterPro record for Q1IXV4

InterPro classification for query sequence

Your query sequence was analysed with InterProScan ([full report](#)).

Protein classification according to InterPro

This sequence belongs to the family:
[Cell division protein FtsZ](#)
 and contains the following domains:
[Tubulin-FtsZ](#), [GTPase](#) [Tubulin-FtsZ, C-terminal](#)

Summary report of the overall match

Summary description of the matched sequence in Integr8

Summary InterPro classification of the matched sequence

Summary InterPro classification of the query sequence

Fig. 8

The Inquisitor results page. A report is shown on the best match to the query sequence within the selected taxonomic scope. (one can also see reports on less good matches). The gene and its product is described; additionally, an InterProScan report is prepared on the query sequence. In the above example, the search sequence and the displayed match show only 46% identity; however, they share 3 of the same protein domains and are classified as members of the same family. However, the matched sequence is additionally a member of a further family, to which the query sequence does not belong.

3 Complete Proteomes Data in Integr8

In this section, we're going to look at complete proteomes in Integr8. Firstly, we're going to see what data is available for download; and then we're going to look at how we can get an overview of complete proteomes within Integr8. Let's see what data we can download for *Deinococcus radiodurans*. Search for this species using the QuickSearch facility, and select it. When you select this strain, the following menu should appear in the left hand margin. The same options also appear at the foot of the species description.

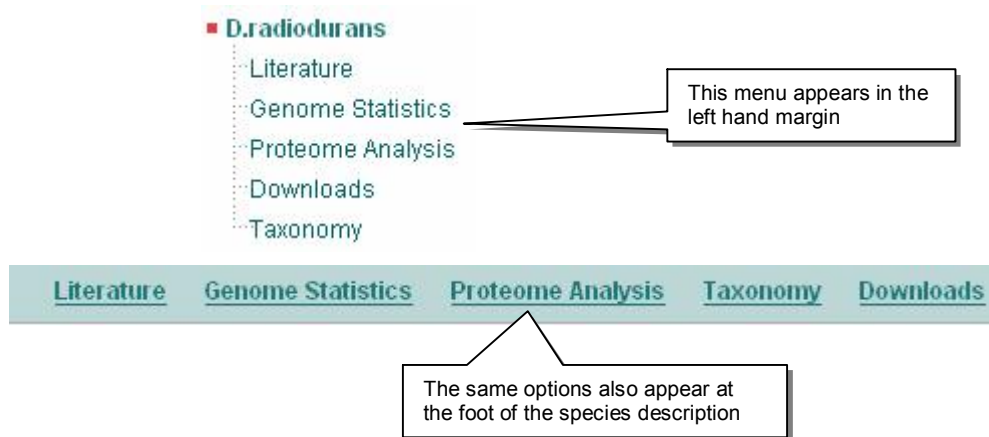


Fig. 9

The species-level menus for the radiation-resistant bacterium Deinococcus radiodurans.

Have a look at what happens when you click on the literature link. Then try clicking on the Genome Statistics link. You should see a page, as follows:

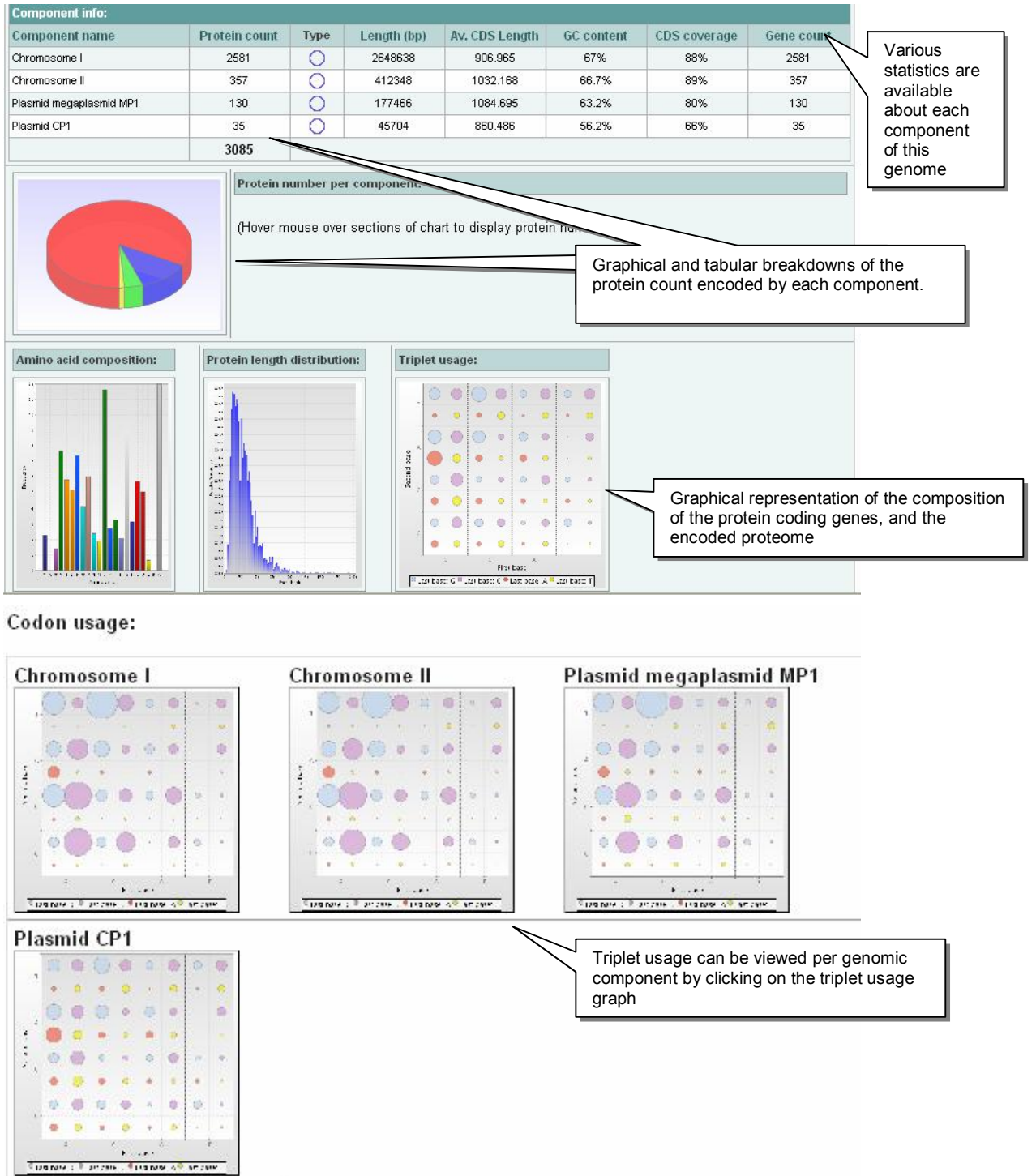


Fig. 10

*The species-level menu for the radiation-resistant bacterium *Deinococcus radiodurans*. The three graphs at the bottom of the page are clickable. Clicking on the triplet usage graph will reveal a display indicating the different pattern of triplet usage on each of the components of this genome; clicking on the other graphs will allow you to see a larger image.*

Now, let's look at the "Download" link, and see what types of data we can download for this bacterium.





Complete proteome - UniProtKB:							UniProt
Proteome sets (Fasta/UniProt/XML format)			Gene sets (Fasta/EMBL format)		InterPro hits	GO annotations	Orthologues
Fasta	UniProt	XML	Fasta	Fasta	Download	Download	Download
Components - UniProtKB:							UniProt
Genome component	EMBL	Genome Reviews.	Chromosome tables	Proteome sets (Fasta/UniProt format)		Gene sets (Fasta/EMBL format)	
Chromosome I	AE000513 	AE000513_GR	Downloads	Fasta	UniProt	Fasta	EMBL
Chromosome II	AE001825 	AE001825_GR	Downloads	Fasta	UniProt	Fasta	EMBL
Plasmid megaplasmid MP1	AE001826 	AE001826_GR	Downloads	Fasta	UniProt	Fasta	EMBL
Plasmid CP1	AE001827 	AE001827_GR	Downloads	Fasta	UniProt	Fasta	EMBL

Fig. 11

The download page for the radiation-resistant bacterium Deinococcus radiodurans.

Here is a summary of the different data types that can be downloaded:

- EMBL archive submissions. These are usually available for each component (i.e. chromosome, plasmid, organelle genome) of the genome, and represent the annotation and sequence originally submitted to the International Nucleotide Sequence Database Consortium (whose members, the EMBL Nucleotide Sequence Database, GenBank, and the DNA Database of Japan, share their data). In some cases, where the sequence in the archival database is known to be out of date, Integr8 is built instead from an alternative source, e.g. a sequence derived from a model organism database. In these cases, a link is provided to the sequence in the source that has been used.
- Genome Reviews entries. Genome Reviews represent re-annotated versions of EMBL submissions. Annotation is imported from the UniProt Knowledgebase, a curated protein sequence database; the representation of certain data types is standardised, and missing annotations (e.g. non-coding RNA genes) are added after the performance of direct sequence analysis. The current scope of Genome Reviews is bacteria, archaea, and bacteriophage; a few lower eukaryotic genomes are also included. Genome Reviews data can also be accessed through an Ensembl-style web interface; Genome Reviews is designed to complement Ensembl's coverage of higher species. Genome Reviews will be explored at greater length later in this tutorial.
- Chromosome tables. Contain information about the protein-coding genes on each component of the genome in a convenient tab-delimited format.
- Gene sets. Gene sets contain the DNA sequence of a (protein-coding) gene, and are available for all species covered by Genome Reviews. Where one gene encodes multiple products, the complete DNA region will be incorporated into a single entity. Gene sets can be downloaded in two formats:
 - EMBL-like flat file format. This contains the same annotation as the source Genome Reviews entry (but recalibrated so that features are expressed in local co-ordinates)
 - FASTA format

Gene sets can be downloaded for a whole genome, or for the subset of genes encoded by a single component of the genome.

- Proteome sets. Non-redundant sets of UniProtKB entries for each complete proteome. In some cases, UniProtKB contains "redundant" sequences representing different submissions for the same protein e.g. variant sequence, fragment sequence, erroneous

sequence etc. The proteome sets are filtered to remove this redundancy. They can be downloaded in 3 formats:

- FASTA format
- UniProtKB flat file format
- UniProtKB XML format

Proteome sets can be downloaded for a whole proteome, or for the subset of proteins encoded by a single component of the genome.

For higher species, IPI sets can also be downloaded – IPI will be discussed later on in this tutorial.

- InterPro hits files. Contain a summary of all InterPro matches for proteins from this proteome.
- GOA files. Contain annotation for all proteins from this proteome using the Gene Ontology as a controlled vocabulary.
- Orthologues files. Contain a list of potential orthologues in all other species in Integr8, identified by protein similarity, for each protein from this proteome.

If you want to access any of this data in bulk, it can be picked up from the Integr8 FTP site at <ftp://ftp.ebi.ac.uk/pub/databases/integr8>. The README file at <ftp://ftp.ebi.ac.uk/pub/databases/integr8/README> contains information on the directory structure, and the file-naming conventions, so that you can write scripts to fetch the files you are interested in.

You can find more information about each complete proteome in Integr8 by clicking on the “Proteome Analysis” link in the left hand menu. Let’s try doing this for the fission yeast, *Schizosaccharomyces pombe*. Select *S.pombe* as your species, then click on the “Proteome Analysis” link. An excerpt from the page you will see is displayed below.

Use tabs to switch between different types of proteome analysis

Use pull down menu to switch between different sub-types of proteome analysis

Click on link to download matching entries from UniProtKB

InterPro ID	Count	Percentage	Sub-type
IPR015943	126	(1.24%)	1 WD40/YVTN repeat-like
IPR011046	126	(1.24%)	2 WD40 repeat-like
IPR001680	126	(1.24%)	2 WD40 repeat
IPR011009	125	(1.23%)	4 Protein kinase-like
IPR000719	111	(1.09%)	5 Protein kinase core
IPR016024	107	(1.05%)	6 Armadillo-type fold
IPR008271	103	(1.01%)	7 Serine/threonine protein kinase, active site
IPR016040	96	(0.94%)	8 NAD(P)-binding
IPR012677	76	(0.75%)	9 Nucleotide-binding, alpha-beta plait
IPR003593	73	(0.72%)	10 AAA+ ATPase, core
IPR002290	73	(0.72%)	10 Serine/threonine protein kinase

Fig. 12


InterPro-based proteome analysis page for the fission yeast Schizosaccharomyces pombe. Proteome analysis is grouped by type and sub-type (controlled by the tabs and the pull-down menu respectively). A statistical breakdown of the proteome is given, and corresponding entries can be downloaded.

The initial page displayed shows the most common InterPro domains in this proteome, and gives you the opportunity to download (from UniProt) the protein entries that contain these domains.

You can see a pull down menu towards the top of the page; using this will give you the opportunity to download other sets of proteins encoded in this proteome defined by various InterPro-based comparisons (for example, the most common families, domains and repeats; also the proteins with the largest number of different InterPro classifications from this proteome)

Above the pull down-menu, there are a number of tabs (note, not all of these tabs are available for every species in Integr8). For example, one of the tab is for a CluSTR-derived analysis. The CluSTR database (<http://www.ebi.ac.uk/clustr>) is built on an all-against-all Smith-Waterman comparison performed between all proteins from UniProtKB and other resources. It has been used to analyse individual proteomes: for example, large clusters of proteins from each species are identified, as are clusters of proteins that are mutually similar but contain no proteins annotated by InterPro (potentially indicating hitherto undescribed families or domains).

Let's have a look at one more tab: the GO tab.



GO Classification for <i>S. pombe</i>		
Term	Proteins	
GO:0003674 molecular_function	3102	62.4%
GO:0003676 nucleic acid binding	647	13.0%
GO:0030528 transcription regulator activity	137	2.7%
GO:0003774 motor activity	23	0.4%
GO:0003824 catalytic activity	1604	32.3%
GO:0030234 enzyme regulator activity	73	1.4%
GO:0005198 structural molecule activity	188	3.7%
GO:0005215 transporter activity	302	6.0%
GO:0005488 binding	1939	39.0%

Fig. 13

GO-based proteome analysis page for the fission yeast Schizosaccharomyces pombe.

The Gene Ontology (GO) defines a controlled, hierarchical vocabulary for the description of gene products. GO is structured as 3 separate vocabularies: one to describe the molecular function of a gene product, one to describe the biological process within which it operates, and one to describe the cellular component within which it is found. Because of its hierarchical nature, it is possible to summarise the contents of a proteome in terms of the proteins it contains at a high level, by combining non-specific annotations with abstractions of specific annotations (the application of which implies the concomitant applicability of all less-specific, ancestral terms). Note that in GO, a term may have 2 parents (e.g. for example, all genes labelled with “JNK cascade” can also be described using the terms “MAPKKK cascade” and “stress-activated protein kinase signalling pathway”; but these two terms are not in an ancestral relationship with respect to each other. Therefore, the proportion of proteins corresponding to a collection of non-overlapping GO terms at a certain depth in the hierarchy may add up to > 100%, because they share common child terms.

The GO vocabulary can be browsed at the EBI using the QuickGO browser (<http://www.ebi.ac.uk/ego>).

4 The Integr8 Gene View page

We're now going to see what information Integr8 displays about a selected gene. Search for 'PAX1' in all species. You should get the following display. Click on the "i8" icon for the human PAX1 gene.

Search for species **Go!** Search for gene/protein in **Go!**
[Change scope](#)

4 genes were found. [Download these UniProtKB entries](#) [UniProt](#) 4 gene(s)

Organism name	Gene	UniProt AC	Protein Description
B.taurus	IGI02664448		
G.gallus	PAX1	P47236	Paired box protein Pax-1
H.sapiens	PAX1	P15863	Paired box protein Pax-1
M.musculus	Pax1	P09084	Paired box protein Pax-1

See this gene in the Integr8or, Integr8's own gene view

See this gene in Ensembl

See the protein encoded by this gene in the UniProt Knowledgebase

Fig. 14

The Integr8 gene search results page. Direct links are provided to a number of resources. Click on a column header to sort by the specified criteria.

Click on the "i8" link to see the human PAX1 gene represented in the Integr8or. You should now see the gene as represented in the following figure.

The screenshot shows the Integr8or gene view for PAX1. At the top, there are tabs for Gene, Results, Context, and History. The main content is divided into several sections:

- Gene Summary:** PAX1, Gene id: IGI01454141, Encoding: Paired box protein Pax-1, Genomic component: Chromosome 20, Homology: Orthologues, Paralogues.
- Gene Ontology:** Function, Biological Process, Cellular component.
- Transcript Overview:** A central graphic showing three transcripts (1, 2, 3) with their respective UniProt and IPI identifiers. Transcript 1 is highlighted.
- Gene, Transcript, Protein:** Three columns of data corresponding to the selected transcript.
- Cross-references:** Three tables showing cross-references for the Gene, Transcript, and Isoform.

Callout boxes provide additional information:

- "Arrows expand or collapse displays" points to the expand/collapse arrows in the Gene Ontology section.
- "Quick link to UniProtKB" points to the UniProtKB logo in the Transcript Overview.
- "Schematic gene.transcript/protein overview. The highlighted boxes indicate the transcript and protein for which cross-references are currently displayed." points to the highlighted transcript and protein in the Transcript Overview.
- "Cross-references specific to the selected transcript." points to the cross-reference table for the selected transcript.
- "Quick link to Ensembl" points to the Ensembl logo in the Gene Summary.
- "For information about orthologues and paralogues, click here" points to the Orthologues and Paralogues links in the Gene Summary.

Fig. 15

The Integr8or gene view. The view is constructed so as to break down cross-references according to the specific entity they refer to. General information about all products of the gene is provided at the top of the page.

Take a time to explore this page. A summary of the gene is given at the top of the page, with annotation provided according to the Gene Ontology. In the middle of the page, the transcripts of the gene (and their products) are displayed. Clicking on a given transcript or protein allows cross-references specific to this product of the gene to be shown in the display at the bottom of the page. Tables can be expanded or contracted by clicking on the arrows.

If we look at the specific data for this gene, we see that there are 3 suggested transcripts. One thing that is visible is that two of these transcripts have names (“isoform1” and “isoform 2” respectively), while the third is identified only by an identifier. This identifier comes from IPI (<http://www.ebi.ac.uk/IPI>), a database that unified information from many alternative sources for certain higher eukaryotic proteomes.

The isoform names are derived from the UniProt Knowledgebase, a highly curated database of protein sequence and function (<http://www.ebi.uniprot.org>). Search for human PAX1 at the UniProt website (you can use the links in the “protein” cross-reference section, in the lower right-hand corner of the page, if you have selected either isoform 1 or isoform 2 to view; there are also quick links to the right of the central display). Scroll down the page until you come to the section entitled “Alternative products”. This explains how the two isoforms differ from each other.

Now, let’s look at the same gene in Ensembl. The Ensembl Genome Annotation system has been used to annotate the human genome (and other metazoan genomes), and this annotation can be visualised in the Ensembl Genome Browser. The quick link from the central panel of the Integr8or takes you to Ensembl “ContigView” page, which shows the gene in its wider context. Click on the graphic of PAX1, and select the “Gene” option from the menu, to get a more detailed view of this gene. You should now be on the “Gene View” page. Compare what Ensembl and UniProt have to say about the gene.

Alternative products

Hide | Top

This entry describes **2** isoforms produced by **alternative splicing**. [Align] [Select]

Isoform 1 (identifier: **P15863-1**)

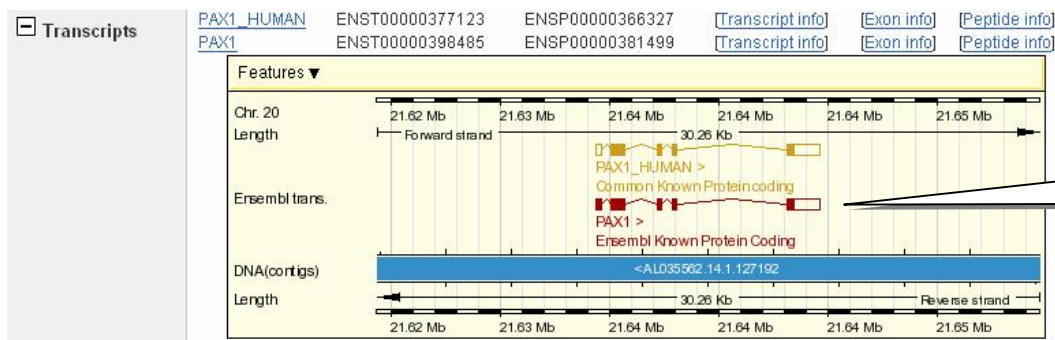
This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Isoform 2 (identifier: **P15863-2**)

The sequence of this isoform differs from the canonical sequence as follows:

330-361 PSREGSLPAPAARPRTPSVAYTDCPSRPRPPR → REGTDRKPPSSGSKAPDALSSLHGLPIPASTS
362-440: Missing.

Alternative product information in the UniProtKB



Transcripts annotated in Ensembl

Fig. 16

Alternative transcript/protein isoform information in the UniProtKB and Ensembl. The UniProtKB describes two isoforms at the protein level. Ensembl displays two transcripts. Note that "isoform 2", as described in the UniProtKB, differs at the C-terminus from the canonical sequence, whereas the two Ensembl transcripts differ at the N-terminus.

UniProt and Ensembl are continuously working to unite their views of metazoan genomes, but there are still differences; the data in the UniProtKB is derived from submission and literature curation, while Ensembl is derived from a computational analysis of the genome. In some cases, such as this gene, perfect agreement does not yet exist. IPI (and Integr8) then offer a combined view of all predicted sequences that have been associated with a gene, although directly redundant representations are still merged (if you look at the cross-references in the Integr8 or display, you can see that isoform 1 is cross-referenced to both UniProtKB and Ensembl; Isoform 2 to UniProtKB only; and the third isoform is derived only from Ensembl).

Let's now have a look at this gene's evolutionary relatives. First of all, let's look at the genes paralogues. Click on the "paralogues" link in the top left hand corner. You should see a page like this, listing PAX family members in human. The Z-score is a measure of the statistical significance of the sequence similarity.

EMBL-EBI  www.ebi.ac.uk

Gene Results Context History

Similar sequences in H.sapiens ⓘ 7 results

Select genes to display Synteny Align

Protein	Chromosome	Organism	Z-Score	Go!
Paired box protein Pax-9	Chromosome 14	H.sapiens	114.0	<input checked="" type="checkbox"/>
Paired box protein Pax-5	Chromosome 9	H.sapiens	51.5	<input checked="" type="checkbox"/>
Paired box protein Pax-2	Chromosome 10	H.sapiens	50.9	<input checked="" type="checkbox"/>
Paired box protein Pax-8	Chromosome 2	H.sapiens	49.0	<input checked="" type="checkbox"/>
Paired box protein Pax-3	Chromosome 2	H.sapiens	48.8	<input checked="" type="checkbox"/>
Paired box protein Pax-7	Chromosome 1	H.sapiens	46.3	<input checked="" type="checkbox"/>
Paired box protein Pax-6	Chromosome 11	H.sapiens	45.5	<input checked="" type="checkbox"/>

2. Chose how you want to compare your sequences

3 Then click "Go!"

1. Click here to select sequences to compare

Fig. 17

Analysis of similar sequences in Integr8. In this example, putative paralogues of the Homo Sapiens PAX1 gene have been retrieved and are being selected for alignment.

Try aligning some of these sequences now. Follow the instructions in the figure to select species for alignment, then click on "Go!".

Let's do one more thing with the gene view. This is most interesting when we have a very large conserved family, so let's search for the *ftsZ* gene in *Escherchia coli K12*. Go to the Integr8or (gene view) page, and click on the "Orthologues" link in the top left hand box. A large number of As you can see, this gene is very widely conserved across the bacterial domain. But how well conserved is the genome around the gene. To investigate this, select a number of potential orthologues from the first page; then a few more from a later page of results. Now, select "synteny", instead of "align", then again click on "Go!". A display like the following page will appear.

Gene	Results	Context	History	Comparative genome view						
	E.coli K12	E.coli O1:H1 / APEC	E.coli RIMD 0509952	E.coli ATCC 700928	B.vietnamiensis	D.vulgaris vulgaris	P.propionicus			
	murD ↑	murD ↑	murD ↑	murD ↑	murD ↑	murD ↑	Ppro_3290			
	ftsW ↑	ftsW ↑	ftsW ↑	ftsW ↑	Bcep1808_0534 ↑	Dvul_0740 ↑	murG			
	murG ↑	murG ↑	murG ↑	murG ↑	murG ↑	murG ↑	murC			
	murC ↑	murC ↑	murC ↑	murC ↑	murC ↑	murC ↑	Ppro_3287 ↓			
	ddlB ↑	ddlB ↑	ddlB ↑	ddlB ↑	Bcep1808_0537 ↑	Dvul_0743 ↑	Ppro_3286 ↓			
	ftsQ ↑	ftsQ ↑	ftsQ ↑	ftsQ ↑	Bcep1808_0538 ↑	Dvul_0744 ↑	Ppro_3285 ↓			
	ftsA ↑	ftsA ↑	ftsA ↑	ftsA ↑	Bcep1808_0539 ↑	Dvul_0745 ↑	Ppro_3284 ↓			
	ftsZ ↑	ftsZ ↑	ftsZ ↑	ftsZ ↑	Bcep1808_0540 ↑	Dvul_0746 ↑	Ppro_3283 ↓			
	lpxC ↑	lpxC ↑	lpxC ↑	lpxC ↑	Bcep1808_0541 ↑	Dvul_0747 ↑	Ppro_3282 ↓			
	secM ↑	secM ↑	secM ↑	secM ↑	lpxC ↑	Dvul_0748 ↓	Ppro_3281 ↓			
	secA ↑	secA ↑	secA ↑	secA ↑	Bcep1808_0543 ↑	Dvul_0749 ↓	Ppro_3280 ↓			
	mutT ↑	mutT ↑	mutT ↑	mutT ↑	Bcep1808_0544 ↑	Dvul_0750 ↓	Ppro_3279 ↓			
	yacG ↓	yacF ↓	ECs0104 ↓	c0118 ↑	Bcep1808_0545 ↑	Dvul_0751 ↓	Ppro_3278 ↑			
	yacF ↓	coaC ↓	yacG ↓	c0119 ↑	Bcep1808_0546 ↑	Dvul_0752 ↓	Ppro_3277 ↑			
	coaE ↓	guaC ↑	yacF ↓	c0120 ↑	Bcep1808_0547 ↑	Dvul_0753 ↓	engB ↓			

The table cells represent genes. Colours represent InterPro domain architectures.

Fig. 18

The Integr8 syntenicity view. A schematic overview of the gene arrangement in a number of genomes is given, centred on putative orthologues. Genomes are sorted according to the degree to which syntenicity is conserved around the selected gene. Colours indicate InterPro domain architecture; thus one can see if syntenicity is conserved, even if gene annotation has not assigned a consistent name between the species (look at *B. vietnamiensis* in the above picture, for example). Clicking on any box takes you to the Integr8 gene view page for that gene. Arrows indicate the relative orientation on the chromosome.

5 Browsing Complete Genomes in Genome Reviews

Genome Reviews (Sterk et al, 2006) is a database of complete genomes, mainly covering bacterial, archaeal and bacteriophage sequences, although some lower eukaryotes are also covered. In Genome Reviews, submissions to the EMBL/Genbank/DDBJ nucleotide sequence repositories are updated with new annotations, either calculated directly from the sequence or imported from actively curated databases. Genome Reviews can be downloaded as a MySQL database. Additionally, individual entries from the Genome Reviews database can be downloaded in the same format as entries from the EMBL Nucleotide Sequence Database. However, the use of the flexible EMBL format is standardised in order to increase the ease of comparison between genomes. We've already seen links for the downloading of individual entries in the course of this tutorial; files for all genomes can also be downloaded directly from the FTP site (ftp://ftp.ebi.ac.uk/pub/databases/genome_reviews). But you can also browse Genome Reviews interactively. We've already looked at the human genome in this tutorial, using the Ensembl browser. Genome Reviews can also be accessed in a similar way.

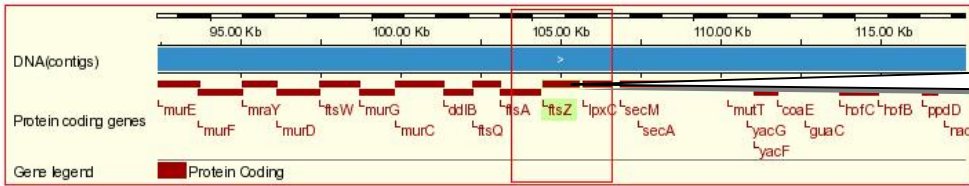
In the left hand margin of Integr8, a menu item appears once you have performed a search labelled "Gene Search Results". Click on this now, and return to the results of your last gene search. You'll notice that, alongside the "i8" link to see a gene in the Integr8, there is also a link labelled "GR". Try clicking on this now: choose the *ftsZ* gene in *E. coli K12* again, so that we can compare the Integr8 and Genome Reviews views. A page should load as follows:

E.coli DSM 5911 Chromosome



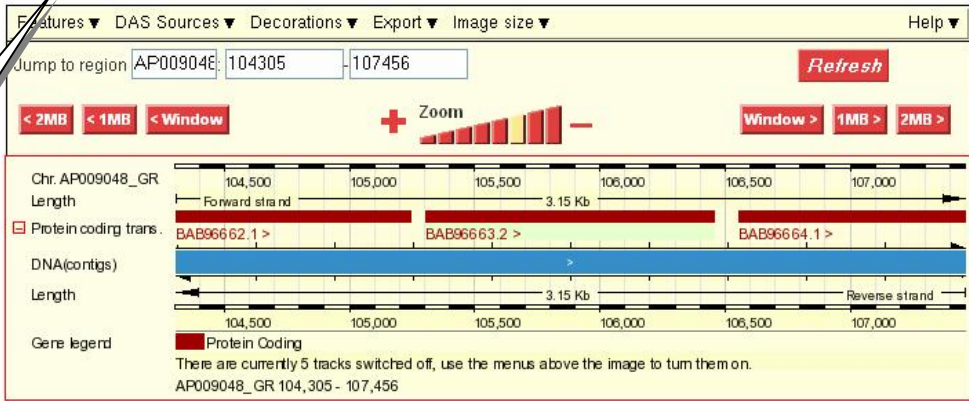
Location of the viewed segment on the chromosome

Overview

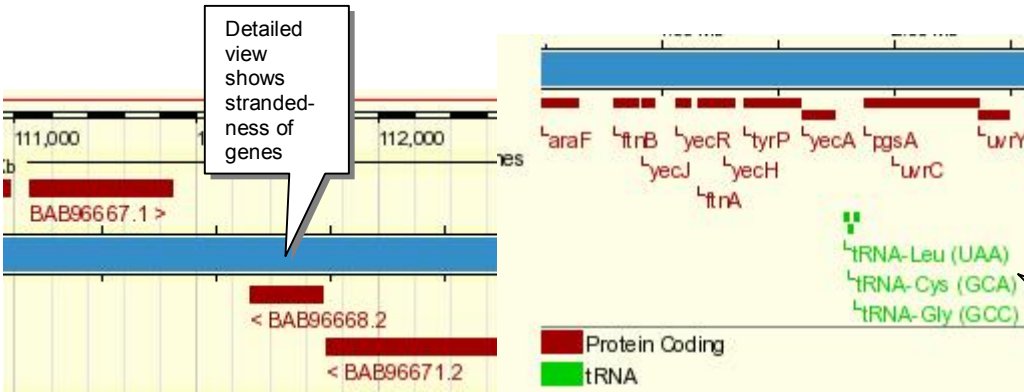


The selected gene

Detailed view



Detailed view with options for zooming and scrolling



Detailed view shows strandedness of genes

Different types of genes appear in different coloured tracks

Fig. 19

The E. coli *ftsZ* gene views in the Genome Reivews graphical browser. Using the same technology as the Ensembl website, the Genome Reviews browser allows you to see a scaled, zoomable representation of each genome. Click on an individual gene, and a menu appears: right click on this menu, and you can access detailed information about each gene. Differnet types of genes (protein-coding, tRNA-encoding, rRNA-encoding) appear in different tracks on the display.

The Genome Reviews website is quite extensive; try exploring to see the different information that is represented here.

Further reading

1. Kersey, PJ et al. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res* 33: D297-D301.
2. Sterk, P. et al. Genome Reviews: Standardizing Content and Representation of Information about Complete Genomes. *OMICS*. 2006 ;10 (2):114-8 16901215