End User Documentation

Metagenomic Clone Bank:

A storage and analysis tool for metagenomic clones

Katelyn Davis

Kathy Lam

Nina Nissan

Rene Quevedo

Philip Rees

Table of Contents

1 Website motivation and overview

1.1 Purpose and benefits

1.2 Description of research activities

1.2.1 Function-based screening of metagenomic libraries

1.2.2 Biochemical assays and targeted sub-cloning of Cosmid clone genes

1.2.3 Pooled Illumina sequencing of Cosmid clones

1.3 Motivation for web-accessible database and custom analysis tools

1.3.1 Data storage needs

1.3.2 Data analysis needs

1.4 Summary of data relationships

1.5 Scope and limitations

2 Website administration

2.1 User accounts

2.2 Web browser recommendations

2.3 Login and user settings

2.4 User permissions

3 Website content and navigation

3.1 Navigation summary

3.2 Homepage

3.3 View All pages

3.3.2 Static Data

3.3.1 Main data

3.3.3 Sorting table data

3.3.4 Exporting table data

3.4 Detail pages

3.4.1 Cosmid and Contig detail pages

3.4.2 Other detail pages

3.5 Search pages

3.5.1 Basic search

3.5.2 Advanced search

3.5.3 Local BLAST search

3.6 Add pages

3.6.1 Add a Cosmid

3.6.2 Add a Subclone, Subclone Assay or Cosmid Assay

3.6.3 Add an ORF

3.6.4 Add a FASTA file of Contigs

3.6.5 Static data

3.7 Edit pages

4 Website custom tools

4.1 Contig retrieval tool

4.2 ORF finding and annotation tool

4.2.1 Usage of the tool

4.2.2 Viewing the Data

4.2.3 Re-generation of image upon adding, editing, or removing ORFs from a Contig

5 Maintenance and troubleshooting

5.1 Server specifications and details

5.2 Common errors and troubleshooting

5.2.1 Contig retrieval tool

5.2.2 ORF prediction and annotation tool

5.2.3 Errors resulting from adding information to the database

5.2.4 404 error

5.2.5 403 error

5.2.6 500 error

6 Technical details

6.1 Database

6.1.1 Database entity-relationship diagram

6.1.2 Website—database communication through Django

6.2 Communication between the Metagenomics and SHARCNET servers

6.2.1 Setting up a connection

6.3 Contig retrieval tool

6.3.1 Generation of user-specified database information

6.3.2 Execution of the Contig retrieval tool

6.3.3 Processing the returned data

6.4 ORF finding and annotation tool

6.4.1 ORF finding and annotation tool

6.4.2 Complete annotations from scratch

6.4.2.1 Generation of user-specified request

6.4.2.2 Complete annotations from scratch

6.4.3 Updated user annotations

6.4.3.1 Generation of user-specified request

6.4.3.2 Updated user annotations

6.4.4 Processing the returned data

7 References

List of Appendices

Appendix 1: Definition of acronyms

Appendix 2: Field requirements for data entry

1 Website motivation and overview

1.1 Purpose and benefits

This website is specifically designed for use by members of the laboratory group of **Trevor C. Charles** at the **University of Waterloo**. Its interface can be easily understood by those with intimate knowledge of data obtained from the Charles Laboratory research activities. Briefly, the website provides researchers with two important functions:

- a web interface to a database so that members can easily store and modify their data
- a suite of custom tools to allow members to analyze their data, particularly DNA sequence data

The following **Section 1.2** briefly describes the nature of research conducted in the Charles laboratory. It provides the minimum amount of background necessary to understand the functions available on the website. However, if you are already familiar the research activities of the laboratory group, please skip the following section.

1.2 Description of research activities

The Charles laboratory is a bacterial genetics research group. One of their research goals is to examine bacterial communities at the DNA level in order to discover novel genes. These discoveries are interesting in their own right, from an academic point of view, but can also be useful for industrial applications.

They are particularly interested in soil communities because soil is the most diverse environment on Earth in terms of bacterial species. The lab, working with other groups, has started an endeavour to sample soils from diverse environments across Canada (http://www.cm2bl.org).

1.2.1 Function-based screening of metagenomic libraries

Using the collects oil samples, the lab extracts the environmental DNA and clones it into cosmid vectors to make DNA libraries. **Figure 1-1** depicts the construction of a metagenomic library; the library is described as "metagenomic" because the genetic material is derived from more than one organism. The libraries are housed in *E. coli* cells, which can be frozen and stored indefinitely.

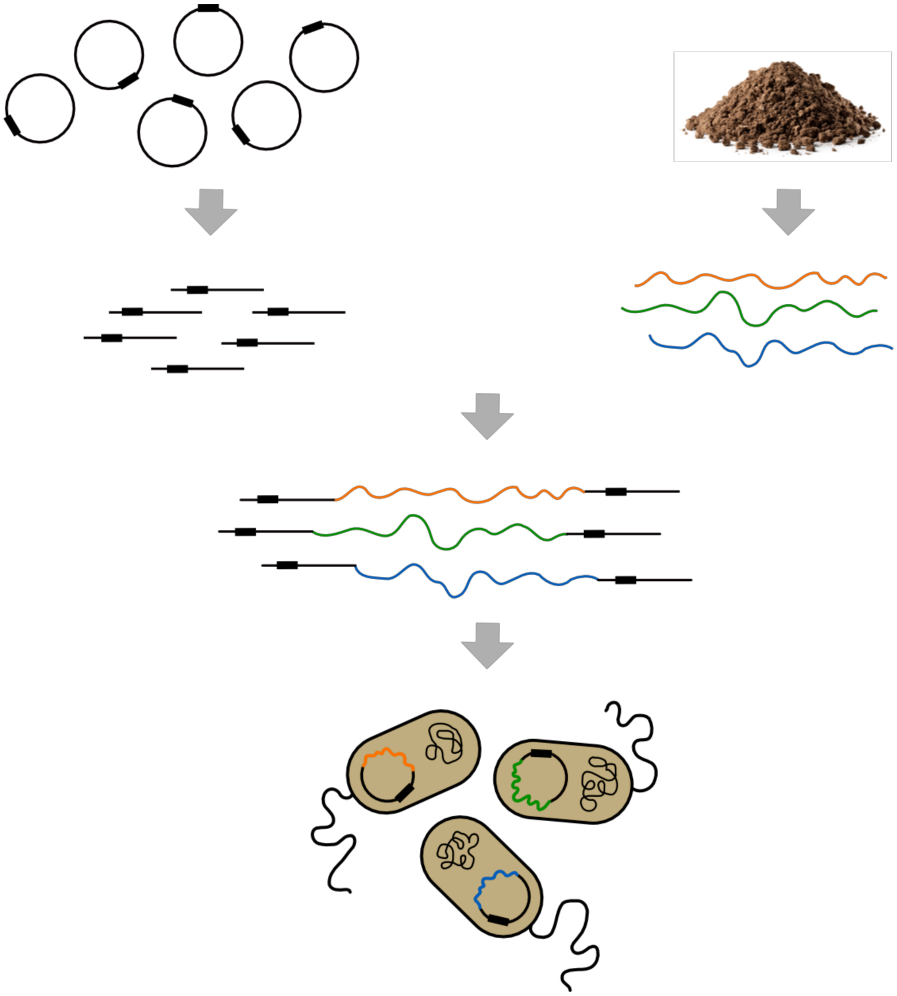


Figure 1-1: Construction of a metagenomic library from an environmental soil sample

Once a metagenomic library is made, the lab will begin screening it for interesting functions. Specifically, the genes carried on the DNA can be expressed in surrogate bacterial hosts, such as *E. coli* ; the surrogate host expresses the foreign genes, and any gene functions are conferred onto the host. For example, it is possible to look for genes that confer antibiotic resistance (**Figure 1-2**): the library-containing *E. coli* cells can simply be grown on media containing the antibiotic of interest; only those cells that carry a resistance gene will be able to survive and form a colony on the media. The colony can be isolated using traditional microbiological approaches, and the cosmid DNA can be purified from the *E. coli* to characterize further.

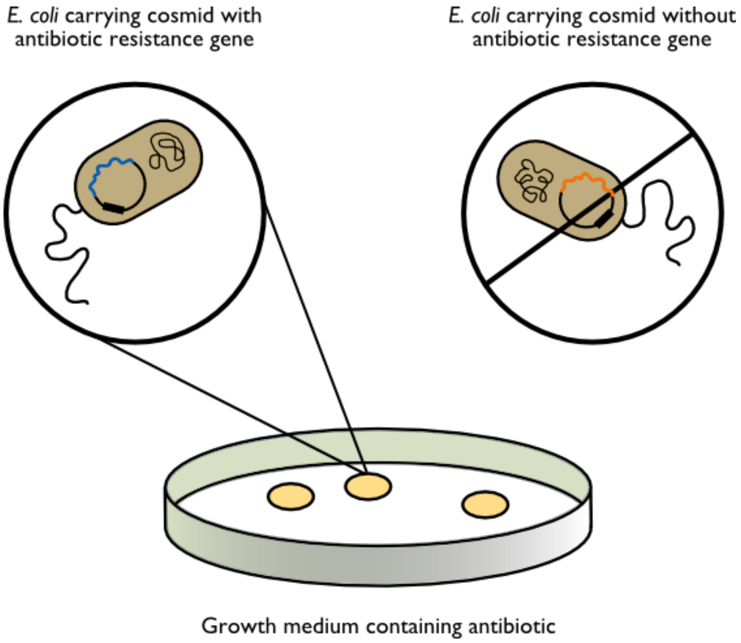

E. coli carrying cosmid with antibiotic resistance gene
E. coli carrying cosmid without antibiotic resistance gene
Growth medium containing antibiotic

Figure 1-2: Example of a function-based screen of a metagenomic library to find antibiotic resistance genes

By using a purely function-based approach, novel genes can be discovered that have low sequence similarity to already known genes, and which may not have been found using only a sequence-based homology approach. After identifying clones from a screen, the lab tries to identify which gene carried on the clone is responsible for the function of interest. However, the cloned DNA is very large (on average, over 30,000 base pairs) and can therefore carry many genes. To find the gene(s) of interest, it is necessary to: (1) perform subcloning and biochemical assays (**Section 1.2.2**), and, most importantly, (2) to analyze the cloned DNA at the sequence level (**Section 1.2.3**).

1.2.2 Biochemical assays and targeted sub-cloning of Cosmid clone genes

Once a cosmid is isolated from *E. coli*, it can be studied further. Specific genes can be extracted from the cosmid and cloned into new vectors to generate subclones. Both cosmids and subclones can be subjected to function-based biochemical assays to confirm the gene functions of interest, and to get various measures on the proteins expressed from the genes.

1.2.3 Pooled Illumina sequencing of Cosmid clones

The Charles lab uses an economical strategy for sequencing their cosmid clones: a pooled sequencing strategy (**Figure 1-3**). This strategy is a hybrid Sanger/Illumina sequencing approach. As part of the sequencing strategy, end sequences for every clone are generated by Sanger-sequencing; these sequences are called “end-tags” to describe their role in the downstream sequence retrieval process in which clones are matched to contigs. In the pooled sequencing method, clones are sequenced together and the lab relies on the post-sequencing assembly process to generate contigs that represent individual clones. After assembly, contigs exist in a pool; to retrieve a specific clone’s contig, they use the clone’s end-tags to query the pool.

Currently, the end-tag-based contig retrieval is done manually using BLAST. It is a slow and tedious procedure to perform for the large number of clones that is sequenced in one pool.

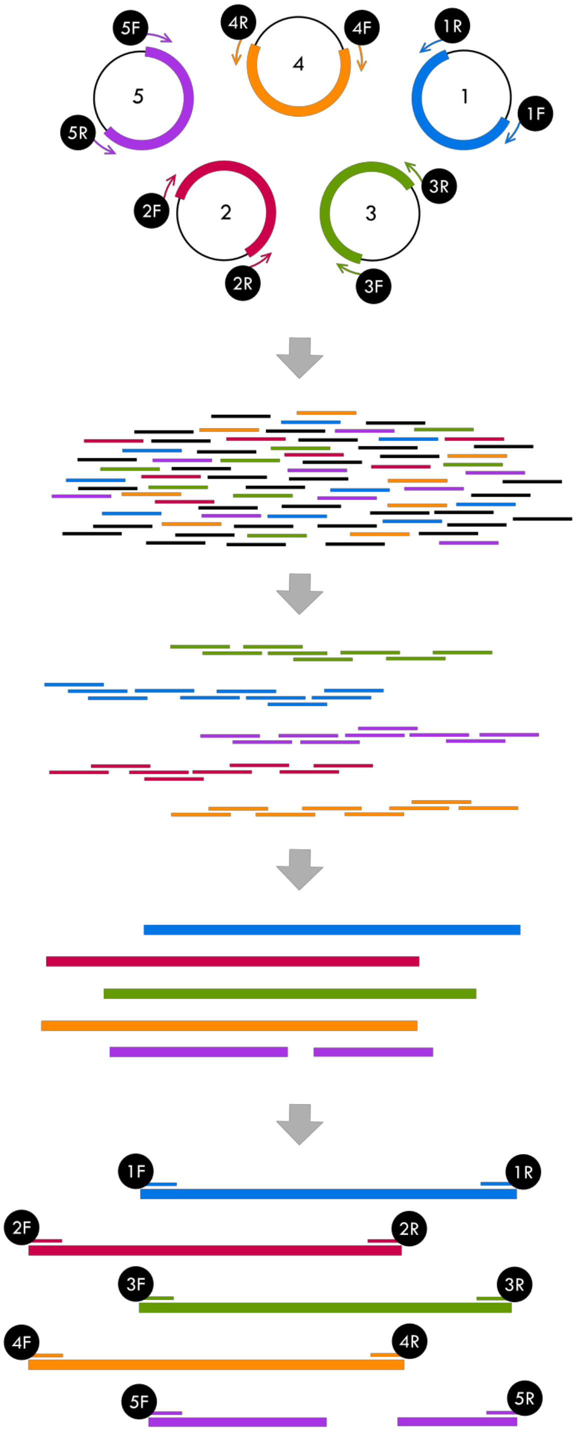


Figure 1-3: Pooled Illumina sequencing of Cosmids with subsequent Contig retrieval using Cosmid end-tags

1.3 Motivation for web-accessible database and custom analysis tools

The Charles laboratory is a relatively small bacterial genetics lab, lacking expertise in bioinformatics. However, they are quickly accumulating genera experimental data and sequence data more quickly than they are able to deal with. At this moment, they are in particular need of an effective way to: (1) store and access their data, and (2) analyze their data.

1.3.1 Data storage needs

The Charles lab requires an organized data management system. Currently, metadata and experimental data are scattered in personal files and lab books; furthermore, sequence data are frequently stored in Microsoft Word Excel files. Unfortunately, both of these storage methods are disorganized, unsustainable and may lead to data loss. This is an undesirable outcome for the lab.

With a web-accessible content management system, the lab would be in a position to better handle their meta- and sequence data. Data would be stored, organized, sortable, queryable, and easily understood by current and new lab members. This would provide a huge advantage to the lab, and aid in lab research productivity.

1.3.2 Data analysis needs

In addition to a data management system, the lab is in particular need of a data analysis pipeline. This pipeline is necessary to be able to handle high-throughput Illumina sequencing data from sequencing Cosmids (see **Section 1.2.3**). Given the lab's goal of identifying novel genes from Cosmids, an automated method is required to identify possible gene functions in sequence data.

Custom tools that assist in: (1) handling Illumina sequencing data, and (2) finding potential genes and gene functions, would be of great benefit to Charles lab members. It would help them move their research forward more quickly, not to mention enabling them to handle and analyze their sequence data independently without external assistance.

1.4 Summary of data relationships

The following statements quickly summarize the note-worthy relationships between objects commonly referenced in the website content and in this user manual. In particular, the user should be familiar with these relationships in order to understand how the custom tools operate.

- Cosmids have End-Tag sequences
- Cosmids are assigned to Sequencing Pools for pooled sequencing
- Contigs are a result of pooled sequencing and are always associated with a Sequencing Pool
- Cosmids can be associated with specific Contigs, through their End-Tags
- The Contig sequence represents the cloned DNA carried on the Cosmid
- Contigs can be associated with more than one Cosmid
- ORFs are found on (i.e. are subsequences) of Contigs
- Specific ORFs are cloned from Cosmids to generate Subclones
- Subclones can be subjected to a Subclone Assay
- Cosmids can be subjected to a Cosmid Assay

Information regarding the entity-relationship diagram (**Figure 6-1**) is provided in **Section 6.1.1**.

1.5 Scope and limitations

A few limitations currently exist, due to the time limitations of this project.

- All data can only be added as individual records, except for Contigs
- The Local BLAST search only has most common parameters available, not all parameters
- There is automated analysis of Contig DNA sequence data, but not of end-tag sequence data
- There is currently no way to edit sets of a records; for example, it is not possible to associate multiple Cosmids with a sequencing pool without editing each Cosmid individually

2 Website administration

2.1 User accounts

Users must have an account to access the website. In order to obtain an account, a request must be made to the Principal Investigator, **Trevor Charles**.

2.2 Web browser recommendations

The website is best viewed in Google Chrome or Mozilla Firefox web browsers; however, it should function in all common web browsers as of April 2014. The website navigation and content were designed to be viewed at a resolution of 1024Loading...768 or higher; at lower resolutions, navigation controls and content may not display appropriately.

2.3 Login and user settings

To access the website, open a web-browser and go to the URL www.metagenomics.uwaterloo.ca. You will be immediately prompted for your account information (**Figure 2-1**). Enter your username and password to access the website.

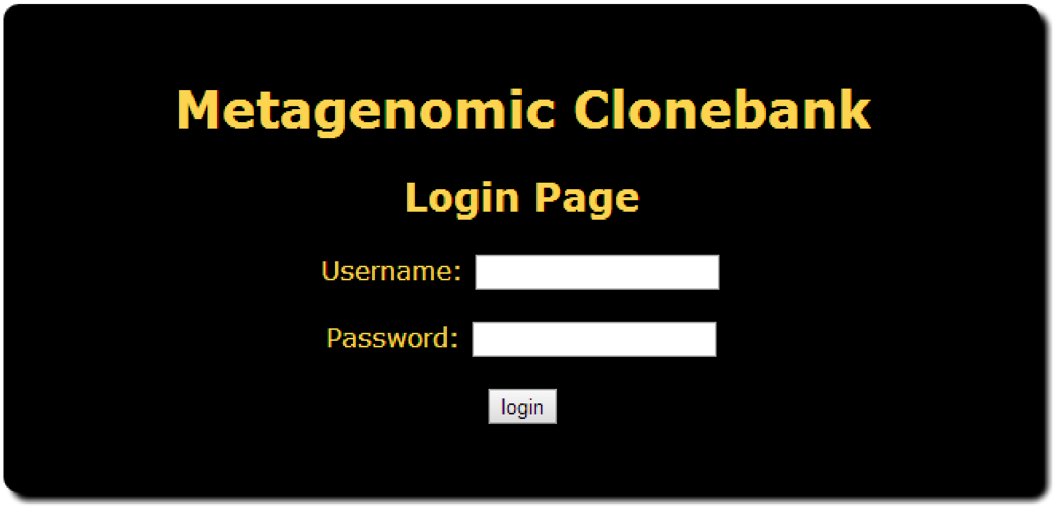


Figure 2-1: Logging in to the website at www.metagenomics.uwaterloo.ca

Note: The username is not case-sensitive but the password is case-sensitive.

Once you have logged in, any of your user settings can be edited by clicking on the gear in the upper right corner of any page (**Figure 2-2**). You will be redirected to a page where you are able to edit: your first name, last name, username and e-mail address. To change your password, follow the **Change Password** link.

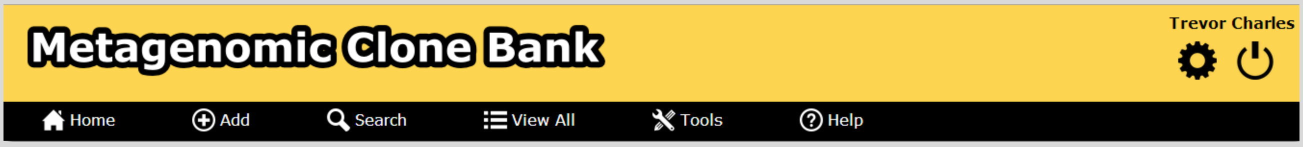


Figure 2-2: Page header with user settings and logout at the top right-hand corner.

To logout from the website, click the logout icon in the top right-hand corner of any page (**Figure 2-2**). For security reasons, we recommend that users log out after every session.

Relevant icons:

⏻ Log in and log out

⚙ Change user settings

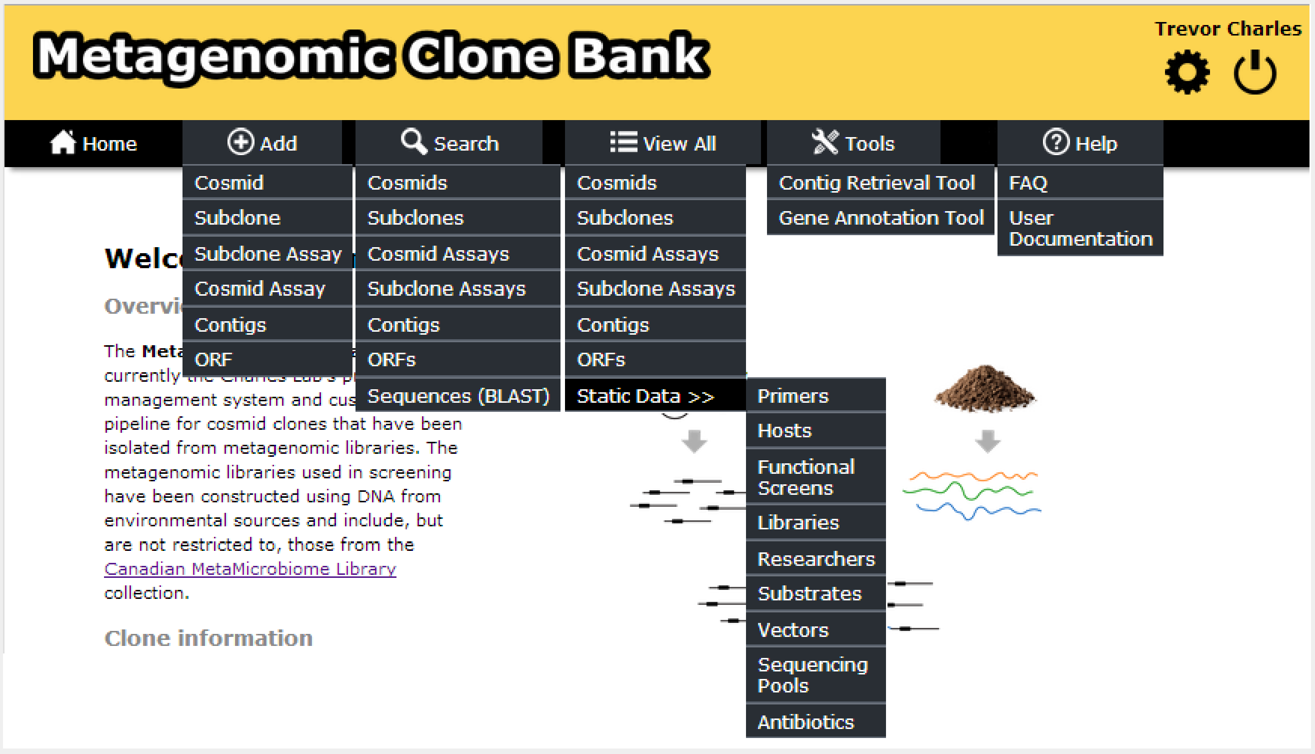
2.4 User permissions

Although all users may view data, not all users may be able to edit database content or execute tools that affect or operate on database content. Depending on your account permissions, you will have different privileges within the website. These permissions are controlled by website administrator. Please refer to **Section 3.7** for more details.

3 Website content and navigation

3.1 Navigation summary

The navigation menu is available on all pages on the website. It provides the following menu options: **Home**, **Add**, **Search**, **View**, **Tools**, and **Help**. Each menu option has further sub-options which become visible when the option is hovered over; **Figure 3-1** shows all available options from the navigation bar.



The Meta
currently
management system and cus
pipeline for cosmid clones that have been isolated from metagenomic libraries. The metagenomic libraries used in screening have been constructed using DNA from environmental sources and include, but are not restricted to, those from the Canadian MetaMicrobiome Library collection.

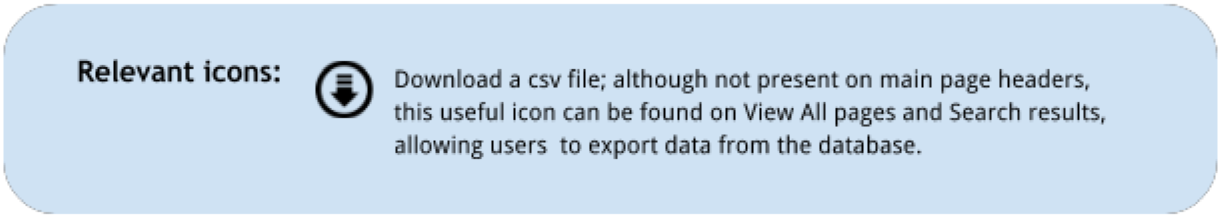
Figure 3-1: Full view of navigation bar menu options: Add, Search, View All, Tools, and Help.

The menu options are summarized in this **Section 3.1**, and are discussed in detail in the remaining **Sections 3.2** to **3.7**. Each detailed section provides instructions on the use of sub-options.

The menu options in the main navigation bar are accompanied by navigation icons, most of which are described in more detail in **Table 3-1**, along with other relevant icons found on the website. All icons listed can be found on select page headings on the website to remind users about the nature of the action that they are performing.

Table 3-1: Description of navigation icons

Icon	Name	Description
🏠	Home icon	Return to homepage
⊕	Add icon	Add new data to database; includes both manual data entry and uploading of sequence data files
🔍	Search icon	Search all data in the database; includes general key-word search, advanced field-based searches, and sequence-based BLAST search
☰	View All icon	View table-based summaries of data records in the database; sorted by categories
📄	Detail icon	View a specific record in the database; linked to from the View All pages
✎	Edit icon	Edit a specific record in the database; linked to from the Detail pages
🛠	Tool icon	Use custom tools to analyze Cosmid-related sequence data; includes retrieving sequence data from Illumina sequencing, as well as executing gene prediction and annotation on sequence data



Relevant icons: ⬇ Download a csv file; although not present on main page headers, this useful icon can be found on View All pages and Search results, allowing users to export data from the database.

3.2 Homepage

The **Homepage** is where users will be redirected upon logging in to the site. It displays a brief overview of the function of Metagenomic Clone Bank. From here, you can navigate the site and access all features that your authentication level allows you. Different tasks are made available depending on the amount of permissions granted for your account.

3.3 View All pages

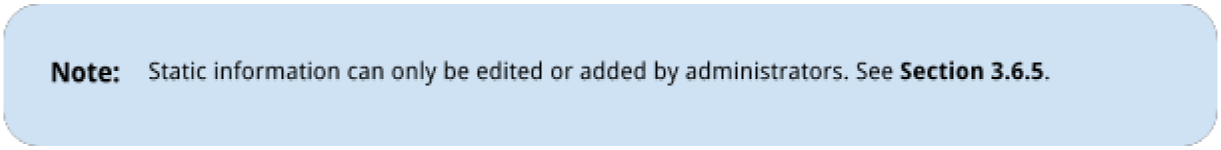
The **View All** option under the main navigation menu allows you to view all of the entries for a specific category of data. The following categories of data are available:

- Cosmids
- Subclones
- Cosmid Assays
- Subclone Assays
- Contigs
- ORFs
- Static Data (contains further options)

The first 6 categories constitute the main data while the last category contains static data. The difference between these are discussed below.

3.3.2 Static Data

Static Data is a field that is unique to the **View All** menu option. These are data that are used as reference data throughout the website. They are added by a website administrator, and are carefully maintained and curated.

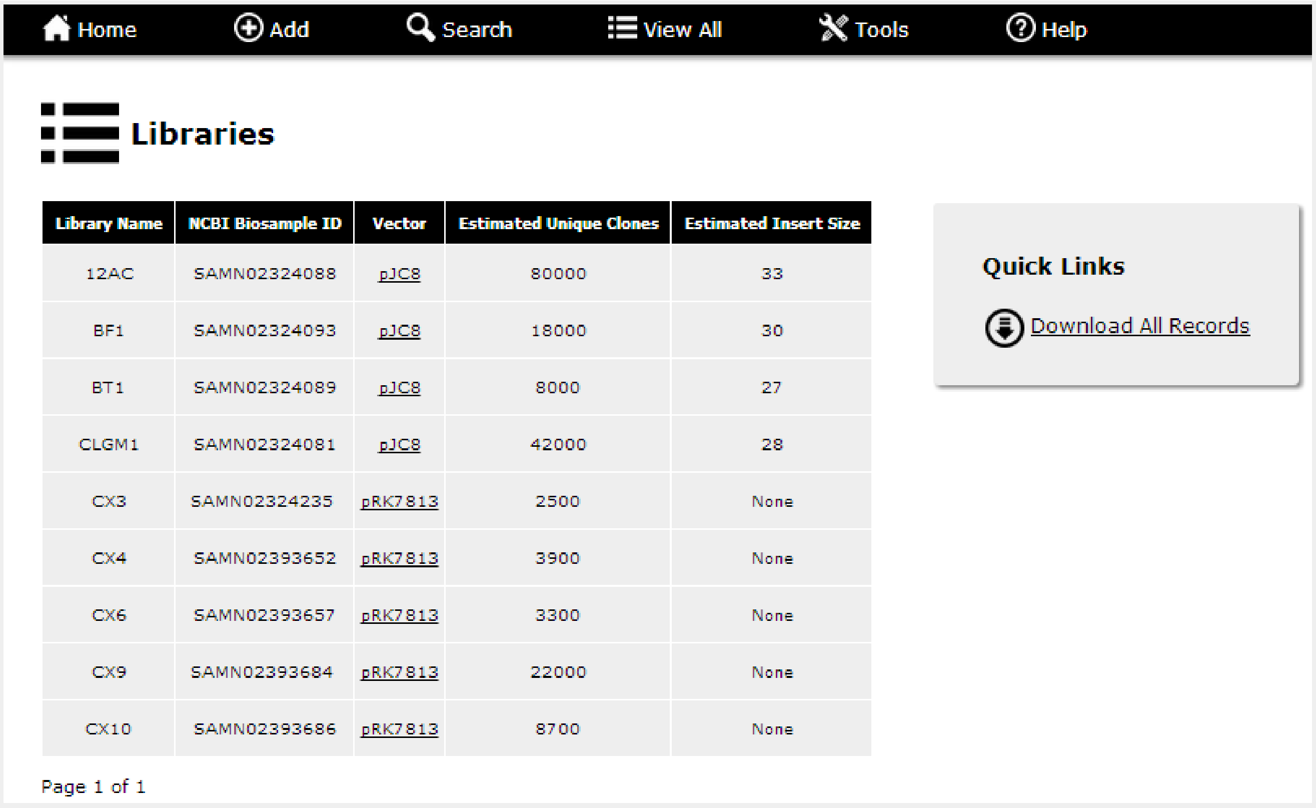


Note: Static information can only be edited or added by administrators. See **Section 3.6.5**.

The following are categories of static data:

- Primers
- Hosts
- Functional Screens
- Libraries
- Researchers
- Substrates
- Vectors
- Sequencing Pools
- Antibiotics

Figure 3-2 shows a View All page for one of the static data categories.



Libraries

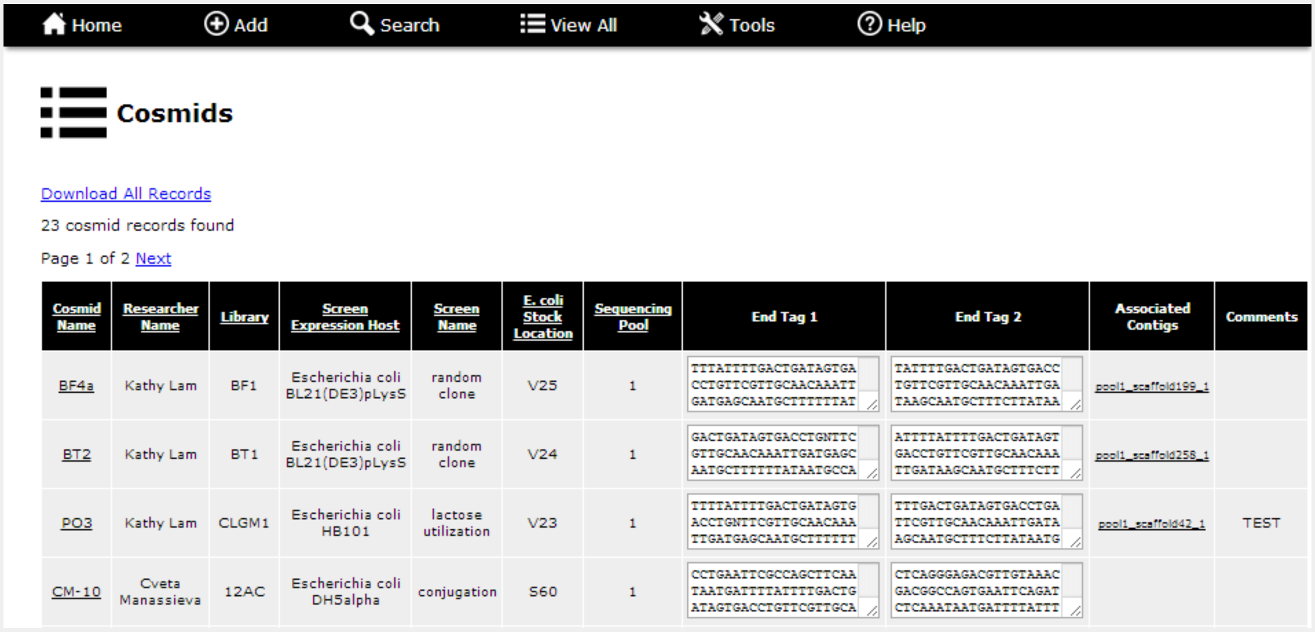
Library Name	NCBI Biosample ID	Vector	Estimated Unique Clones	Estimated Insert Size
12AC	SAMN02324088	pJC8	80000	33
BF1	SAMN02324093	pJC8	18000	30
BT1	SAMN02324089	pJC8	8000	27
CLGM1	SAMN02324081	pJC8	42000	28
CX3	SAMN02324235	pRK7813	2500	None
CX4	SAMN02393652	pRK7813	3900	None
CX6	SAMN02393657	pRK7813	3300	None
CX9	SAMN02393684	pRK7813	22000	None
CX10	SAMN02393686	pRK7813	8700	None

Page 1 of 1

Figure 3-2: An example of a static data page

3.3.1 Main data

Main data are records that are added to the database by active users. These constitute the bulk of the data of the Charles Lab, and the **View All** pages will likely be constantly changing as a result. To combat the large number of records, all View All pages are paginated for ease of viewing.



Cosmids

Download All Records

23 cosmid records found

Page 1 of 2 Next

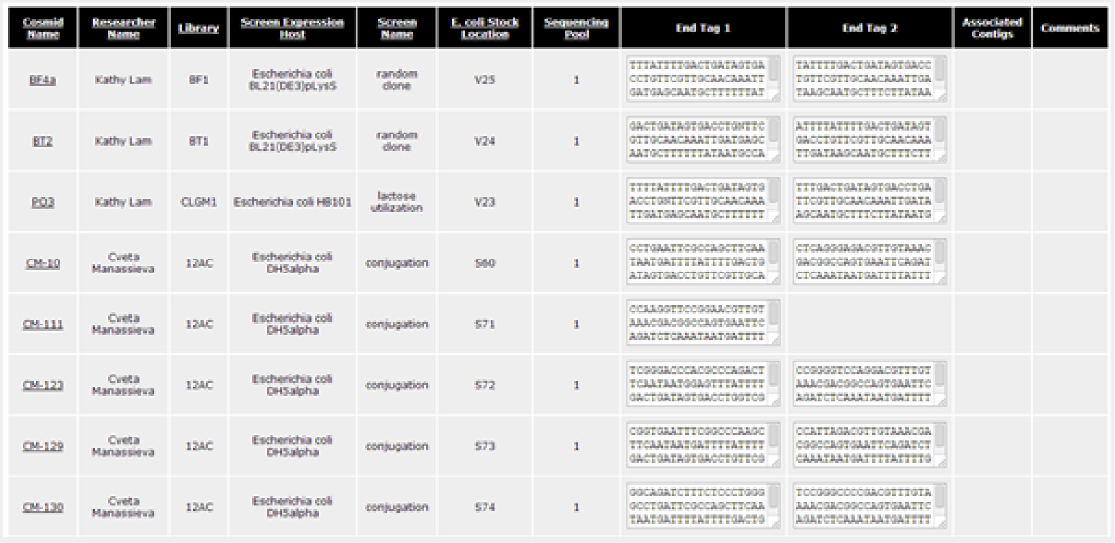
Cosmid Name	Researcher Name	Library	Screen Expression Host	Screen Name	E. coli Stock Location	Sequencing Pool	End Tag 1	End Tag 2	Associated Contigs	Comments
BF4a	Kathy Lam	BF1	Escherichia coli BL21(DE3)pLysS	random clone	V25	1	TTTATTTTGACTGATAGTGACCTGTTCGTTGCAACAAATTGATGAGCAATGCTTTTTTAT	TATTTTGACTGATAGTGACCTGTTCGTTGCAACAAATTGATAAGCAATGCTTTCTTATAA	pool1_scaffold199_1	
BT2	Kathy Lam	BT1	Escherichia coli BL21(DE3)pLysS	random clone	V24	1	GACTGATAGTGACCTGNTTCGTTGCAACAAATTGATGAGCAATGCTTTTTTATAATGCCA	ATTTTATTTTGACTGATAGTGACCTGTTCGTTGCAACAAATTGATAAGCAATGCTTTCTT	pool1_scaffold258_1	
PO3	Kathy Lam	CLGM1	Escherichia coli HB101	lactose utilization	V23	1	TTTTATTTTGACTGATAGTGACCTGNTTCGTTGCAACAAATTGATGAGCAATGCTTTTTT	TTTGACTGATAGTGACCTGATTCGTTGCAACAAATTGATAAGCAATGCTTTCTTATAATG	pool1_scaffold42_1	TEST
CM-10	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S60	1	CCTGAATTCGCCAGCTTCAATAATGATTTTATTTTGACTGATAGTGACCTGTTCGTTGCA	CTCAGGGAGACGTTGTAAACGACGGCCAGTGAATTCAGATCTCAAATAATGATTTTATTT		

Figure 3-3: View All page for Cosmid records

Note: A particular useful function of View All pages is the ability to export all records of one specific type of data. See **Section 3.3.4** for more details on exporting data.

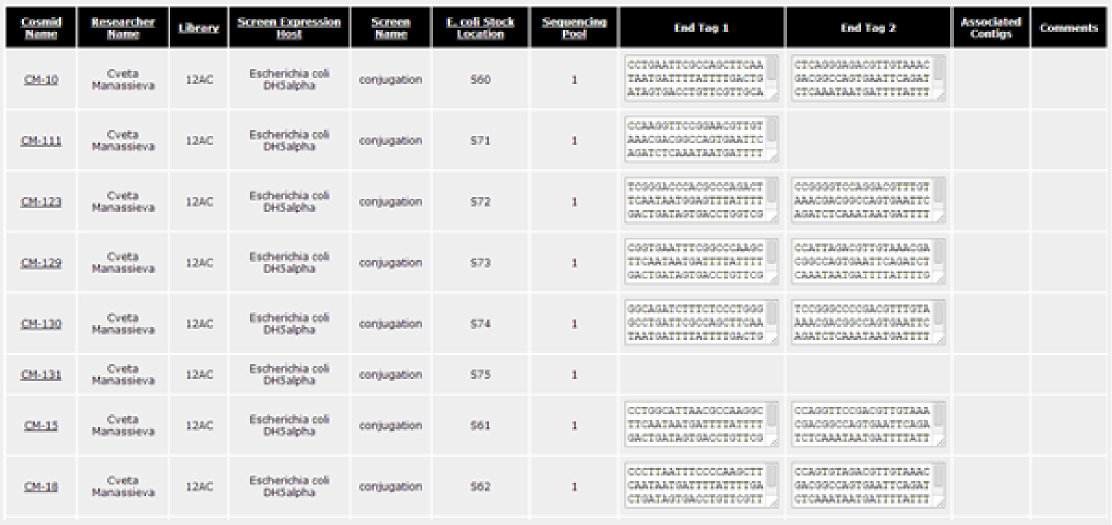
3.3.3 Sorting table data

The information displayed in the table is not sorted by default (**Figure 3-4**). Non-static tables can be sorted by any column where the heading is underlined in white. To sort by a given column simply click on the column heading and you will be directed to the first page of the now sorted results (**Figure 3-5**). Columns that contain solely numerical data will be sorted numerically, all other columns will be sorted alphanumerically. An example of sorting can be seen below.



Cosmid Name	Researcher Name	Library	Screen Expression Host	Screen Name	E. coli Stock Location	Sequencing Pool	End Tag 1	End Tag 2	Associated Contigs	Comments
BF4a	Kathy Lam	BF1	Escherichia coli BL21(DE3)pLysS	random clone	V25	1	TTTATTTTGACTGATAGTGA CCTGTTCGTTGCAACAAATT GATGAGCAATGCTTTTTTAT	TATTTTGACTGATAGTGACC TGTTCGTTGCAACAAATTGA TAAGCAATGCTTTCTTATAA		
BT2	Kathy Lam	BT1	Escherichia coli BL21(DE3)pLysS	random clone	V24	1	GACTGATAGTGACCTGTTTC GTTGCAACAAATTGATGAGC AATGCTTTTTTATAATGCCA	ATTTTATTTTGACTGATAGT GACCTGTTCGTTGCAACAAA TTGATAAGCAATGCTTTCTT		
PQ3	Kathy Lam	CLGM1	Escherichia coli HB101	lactose utilization	V23	1	TTTTATTTTGACTGATAGTG ACCTGTTTCGTTGCAACAAA TTGATGAGCAATGCTTTTTT	TTTGACTGATAGTGACCTGA TTCGTTGCAACAAATTGATA AGCAATGCTTTCTTATAATG		
CM-10	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S60	1	CCTGAATTCGCCAGCTTCAA TAATGATTTTATTTTGACTG ATAGTGACCTGTTCGTTGCA	CTCAGGGAGACGTTGTAAAC GACGGCCAGTGAATTCAGAT CTCAAATAATGATTTTATTT		
CM-111	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S71	1	CCAAGGTTCCGGAACGTTGT AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT			
CM-123	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S72	1	TCGGGACCCACGCCCAGACT TCAATAATGGAGTTTATTTT GACTGATAGTGACCTGGTCG	CCGGGGTCCAGGACGTTTGT AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT		
CM-129	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S73	1	CGGTGAATTTCGGCCCAAGC TTCAATAATGATTTTATTTT GACTGATAGTGACCTGTTCG	CCATTAGACGTTGTAAACGA CGGCCAGTGAATTCAGATCT CAAATAATGATTTTATTTTG		
CM-130	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S74	1	GGCAGATCTTTCTCCCTGGG GCCTGATTCGCCAGCTTCAA TAATGATTTTATTTTGACTG	TCCGGGCCCCGACGTTTGTA AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT		

Figure 3-4 : Default unsorted Cosmid table

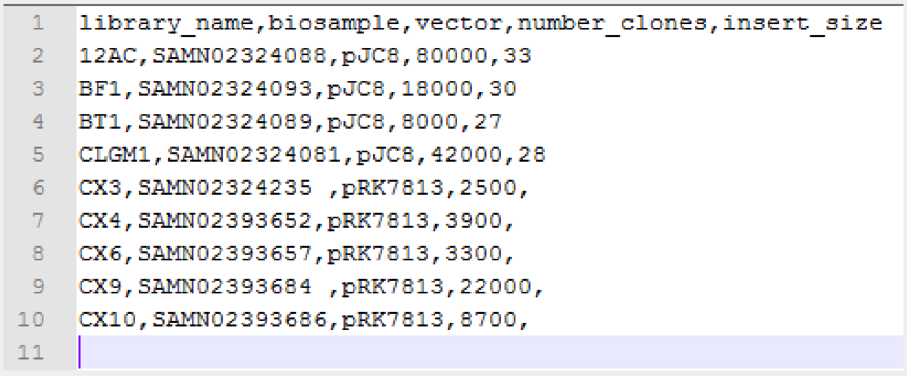


Cosmid Name	Researcher Name	Library	Screen Expression Host	Screen Name	E. coli Stock Location	Sequencing Pool	End Tag 1	End Tag 2	Associated Contigs	Comments
CM-10	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S60	1	CCTGAATTCGCCAGCTTCAA TAATGATTTTATTTTGACTG ATAGTGACCTGTTCGTTGCA	CTCAGGGAGACGTTGTAAAC GACGGCCAGTGAATTCAGAT CTCAAATAATGATTTTATTT		
CM-111	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S71	1	CCAAGGTTCCGGAACGTTGT AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT			
CM-123	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S72	1	TCGGGACCCACGCCCAGACT TCAATAATGGAGTTTATTTT GACTGATAGTGACCTGGTCG	CCGGGGTCCAGGACGTTTGT AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT		
CM-129	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S73	1	CGGTGAATTTCGGCCCAAGC TTCAATAATGATTTTATTTT GACTGATAGTGACCTGTTCG	CCATTAGACGTTGTAAACGA CGGCCAGTGAATTCAGATCT CAAATAATGATTTTATTTTG		
CM-130	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S74	1	GGCAGATCTTTCTCCCTGGG GCCTGATTCGCCAGCTTCAA TAATGATTTTATTTTGACTG	TCCGGGCCCCGACGTTTGTA AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT		
CM-131	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S75	1				
CM-15	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S61	1	CCTGGCATTAACGCCAAGGC TTCAATAATGATTTTATTTT GACTGATAGTGACCTGTTCG	CCAGGTTCCGACGTTGTAAA CGACGGCCAGTGAATTCAGA TCTCAAATAATGATTTTATT		
CM-18	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S62	1	CCCTTAATTTCCCCAAGCTT CAATAATGATTTTATTTTGA CTGATAGTGACCTGTTCGTT	CCAGTGTAGACGTTGTAAAC GACGGCCAGTGAATTCAGAT CTCAAATAATGATTTTATTT		

Figure 3-5: Cosmid table after having been sorted by library.

3.3.4 Exporting table data

From any table-based page, the table of information can be exported. Simply click on the download icon (recall the export icon from icons discussed in **Section 3.1)**. This will download a file to your computer in the standard comma separated value (CSV) format:



```
library_name,biosample,vector,number_clones,insert_size
12AC,SAMN02324088,pJC8,80000,33
BF1,SAMN02324093,pJC8,18000,30
BT1,SAMN02324089,pJC8,8000,27
CLGM1,SAMN02324081,pJC8,42000,28
CX3,SAMN02324235 ,pRK7813,2500,
CX4,SAMN02393652,pRK7813,3900,
CX6,SAMN02393657,pRK7813,3300,
CX9,SAMN02393684 ,pRK7813,22000,
CX10,SAMN02393686,pRK7813,8700,

```

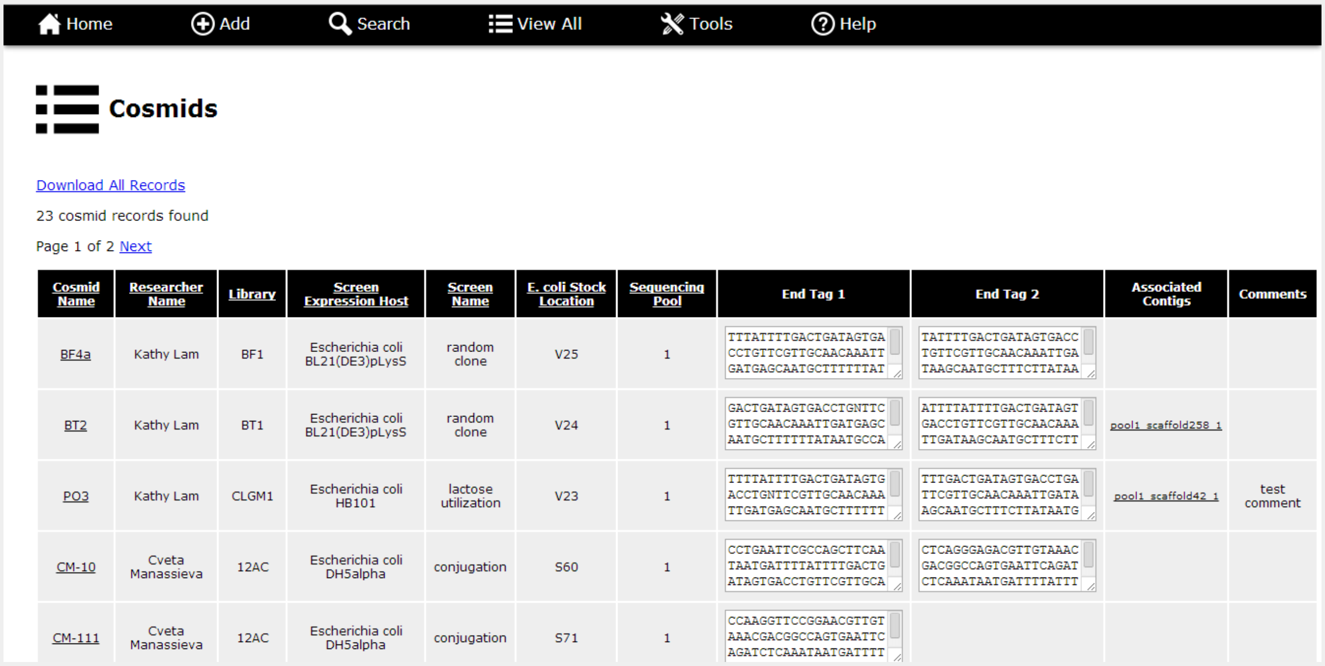
Figure 3-6: Example of an exported CSV file, using data from Figure 3-5.

Caution: It is highly recommended that you do not use Microsoft Excel to view exported sequence data as MS Excel data cells have a maximum capacity of 32 767 characters per cell. Frequently, Contig sequences are exceed this limit, which causes truncation or unexpected formatting during viewing. As an alternative, we recommend using Notepad++ to view files containing long string values.

3.4 Detail pages

From any table-based page, including both **View All** pages or **Search** results pages, it is possible to click on a specific record to navigate to a page with more detailed information. **Figure 3-7** shows an example using Cosmids, but this navigation is set up for all of the table-based pages:

- ○ Cosmid
- ○ Subclone
- ○ Cosmid Assay
- ○ Subclone Assay
- ○ Open Reading Frame (ORF)
- ○ Contigs



Cosmids

Download All Records

23 cosmid records found

Page 1 of 2 Next

Cosmid Name	Researcher Name	Library	Screen Expression Host	Screen Name	E. coli Stock Location	Sequencing Pool	End Tag 1	End Tag 2	Associated Contigs	Comments
BF4a	Kathy Lam	BF1	Escherichia coli BL21(DE3)pLysS	random clone	V25	1	TTTATTTTGACTGATAGTGA CCTGTTCGTTGCAACAAATT GATGAGCAATGCTTTTTTAT	TATTTTGACTGATAGTGACC TGTTCGTTGCAACAAATTGA TAAGCAATGCTTTCTTATAA		
BT2	Kathy Lam	BT1	Escherichia coli BL21(DE3)pLysS	random clone	V24	1	GACTGATAGTGACCTGNTTC GTTGCAACAAATTGATGAGC AATGCTTTTTTATAATGCCA	ATTTTATTTTGACTGATAGT GACCTGTTCGTTGCAACAAA TTGATAAGCAATGCTTTCTT	pool1_scaffold258_1	
PO3	Kathy Lam	CLGM1	Escherichia coli HB101	lactose utilization	V23	1	TTTTATTTTGACTGATAGTG ACCTGNTTCGTTGCAACAAA TTGATGAGCAATGCTTTTTT	TTTGACTGATAGTGACCTGA TTCGTTGCAACAAATTGATA AGCAATGCTTTCTTATAATG	pool1_scaffold42_1	test comment
CM-10	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S60	1	CCTGAATTCGCCAGCTTCAA TAATGATTTTATTTTGACTG ATAGTGACCTGTTCGTTGCA	CTCAGGGAGACGTTGTAAAC GACGGCCAGTGAATTCAGAT CTCAAATAATGATTTTATTT		
CM-111	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S71	1	CCAAGGTTCCGGAACGTTGT AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT			

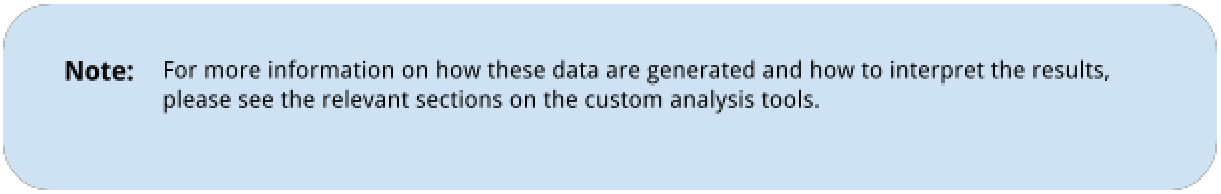
Figure 3-7: Click on the Cosmid name to navigate to a Cosmid Detail page from the View All Cosmids page

3.4.1 Cosmid and Contig detail pages

The most important Detail pages are the Cosmid Detail page (**Figure 3-8)** and the Contig Detail page. The latter is very similar to the former, except that the same Contig can be associated with different Cosmids (as explained in **Section 1.2.3** on pooled sequencing, as well as in **Section 4.1** on the Contig Retrieval Tool).

The Cosmid Detail page is likely the main way that users will be viewing their data. It contains the following:

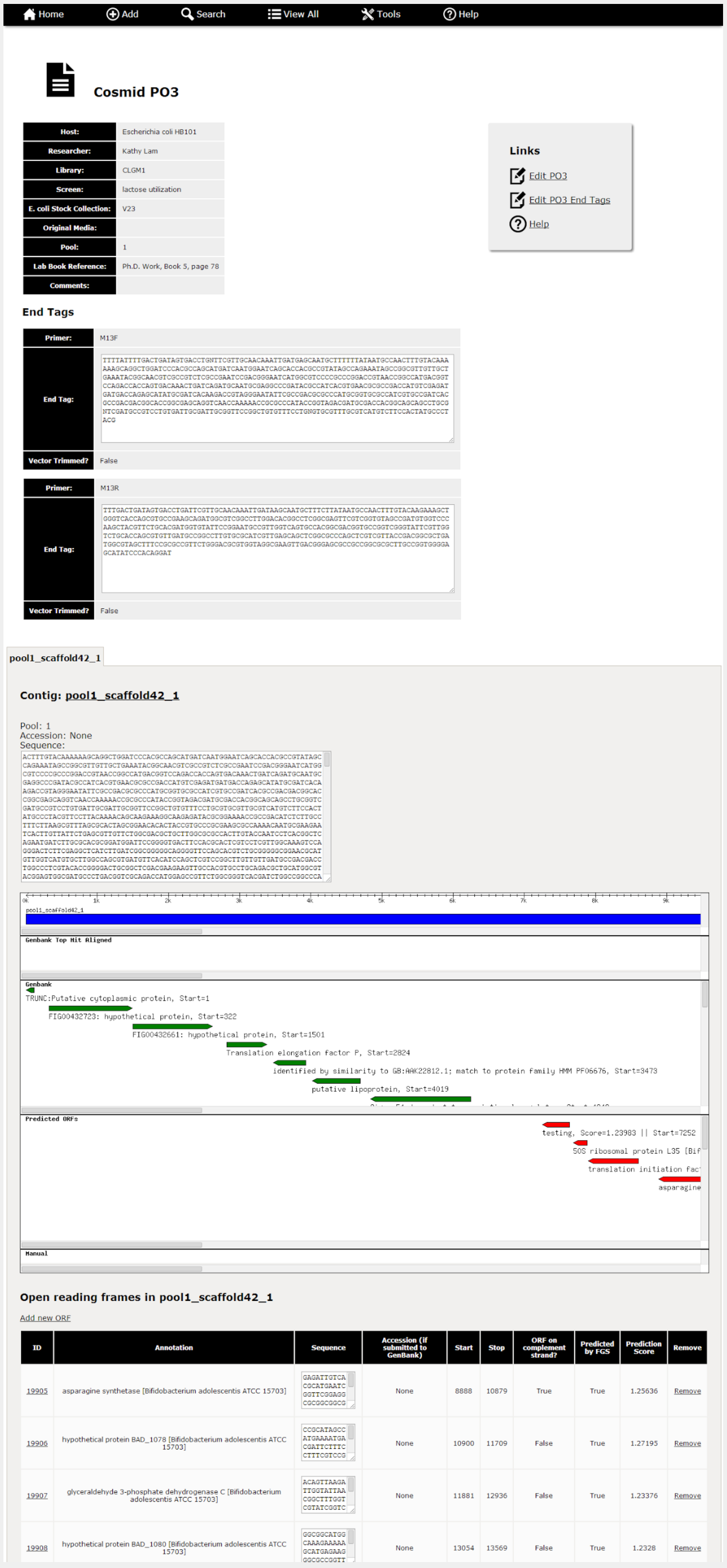
- a summary of Cosmid information, including end-tag sequences
- associated Contigs and their sequences -- up to two Contigs from execution of the Contig Retrieval tool (**Section 4.1**)
- ORF information in the form of images and records -- from execution of the ORF Finding and Annotation tool (**Section 4.2**)



Note: For more information on how these data are generated and how to interpret the results, please see the relevant sections on the custom analysis tools.

3.4.2 Other detail pages

The remaining Detail pages are much more straight-forward than either Cosmid or Contig, and do not require much in the way of explanation. They simple list various features of each record.



Cosmid PO3

Host:	Escherichia coli HB101
Researcher:	Kathy Lam
Library:	CLGM1
Screen:	lactose utilization
E. coli Stock Collection:	V23
Original Media:	
Pool:	1
Lab Book Reference:	Ph.D. Work, Book 5, page 78
Comments:	

Links

Edit PO3

Edit PO3 End Tags

Help

End Tags

Primer: M13F

End Tag:

```
TTTTATTTTGACTGATAGTGACCTGNTTCGTTGCAACAAATTGATGAGCAATGCTTTTTTATAATGCCAACTTTGTACAAA
AAAGCAGGCTGGATCCCACGCCAGCATGATCAATGGAATCAGCACCACGCCGTATAGCCAGAAATAGCCGGCGTTGTTGCT
GAAATACGGCAACGTCGCCGTCTCGCCGAATCCGACGGGAATCATGGCGTCCCCGCCCGGACCGTAACCGGCCATGACGGT
CCAGACCACCAGTGACAAACTGATCAGATGCAATGCGAGGCCCGATACGCCATCACGTGAACGCGCCGACCATGTCGAGAT
GATGACCAGAGCATATGCGATCACAAGACCGTAGGGAATATTCGCCGACGCGCCCATGCGGTGCGCCATCGTGCCGATCAC
GCCGACGACGGCACCGGCGAGCAGGTCAACCAAAAACCGCGCCCATACCGGTAGACGATGCGACCACGGCAGCAGCCTGCG
NTCGATGCCGTCCTGTGATTGCGATTGCGGTTCCGGCTGTGTTTCCTGNGTGCGTTTGCGTCATGTCTTCCACTATGCCCT
ACG
```

Vector Trimmed? False

Primer: M13R

End Tag:

```
TTTGACTGATAGTGACCTGATTCGTTGCAACAAATTGATAAGCAATGCTTTCTTATAATGCCAACTTTGTACAAGAAAGCT
GGGTCACCAGCGTGCCGAAGCAGATGGCGTCGGCCTTGGACACGGCCTCGGCGAGTTCGTCGGTGTAGCCGATGTGGTCCC
AAGCTACGTTCTGCACGATGGTGTATTCCGGAATGCCGTTGGTCAGTGCCACGGCGACGGTGCCGGTCGGGTATTCGTTGG
TCTGCACCAGCGTGTTGATGCCGGCCTTGTGCGCATCGTTGAGCAGCTCGGCGCCCAGCTCGTCGTTACCGACGGCGCTGA
TGGCGTAGCTTTCCGCGCCGTTCTGGGACGCGTGGTAGGCGAAGTTGACGGGAGCGCCGCCGGCGCGCTTGCCGGTGGGGA
GCATATCCCACAGGAT
```

Vector Trimmed? False

pool1_scaffold42_1

Contig: pool1_scaffold42_1

Pool: 1
Accession: None
Sequence:

```
ACTTTGTACAAAAAAGCAGGCTGGATCCCACGCCAGCATGATCAATGGAATCAGCACCACGCCGTATAGC
CAGAAATAGCCGGCGTTGTTGCTGAAATACGGCAACGTCGCCGTCTCGCCGAATCCGACGGGAATCATGG
CGTCCCCGCCCGGACCGTAACCGGCCATGACGGTCCAGACCACCAGTGACAAACTGATCAGATGCAATGC
GAGGCCCGATACGCCATCACGTGAACGCGCCGACCATGTCGAGATGATGACCAGAGCATATGCGATCACA
AGACCGTAGGGAATATTCGCCGACGCGCCCATGCGGTGCGCCATCGTGCCGATCACGCCGACGACGGCAC
CGGCGAGCAGGTCAACCAAAAACCGCGCCCATACCGGTAGACGATGCGACCACGGCAGCAGCCTGCGGTC
GATGCCGTCCTGTGATTGCGATTGCGGTTCCGGCTGTGTTTCCTGCGTGCGTTGCGTCATGTCTTCCACT
ATGCCCTACGTTCCTTACAAAACAGCAAGAAAGGCAAGAGATACGCGGAAAACCGCCGACATCTCTTGCC
TTTCTTAAGCGTTTAGCGCACTAGCGGAACACACTACCGTGCCCGCGAAGCGCCAAAACAATGCGAAGAA
TCACTTGTTATTCTGAGCGTTGTTCTGGCGACGCTGCTTGGCGCGCCACTTGTACCAATCCTCACGGCTC
AGAATGATCTTGCGCACGCGGATGGATTCCGGGGTGACTTCCACGCACTCGTCCTCGTTGGCAAAGTCCA
GGGACTCTTCGAGGCTCATCTTGATCGGCGGGGGCAGGGGTTCCAGCACGTCTGCGGGGGCGGAACGCAT
GTTGGTCATGTGCTTGGCCAGCGTGATGTTCACATCCAGCTCGTCCGGCTTGTTGTTGATGCCGACGACC
TGGCCCTCGTACACCGGGGACTGCGGCTCGACGAAGAAGTTGCCACGTGCCTGCAGACGCTGCATGGCGT
ACGGAGTGGCGATGCCCTGACGGTCGCAGACCATGGAGCCGTTCTGGCGGGTCACGATCTGGCCGGCCCA
```

0k
1k
2k
3k
4k
5k
6k
7k
8k
9k
pool1_scaffold42_1
Genbank Top Hit Aligned
Genbank
TRUNC:Putative cytoplasmic protein, Start=1
FIG00432723: hypothetical protein, Start=322
FIG00432661: hypothetical protein, Start=1501
Translation elongation factor P, Start=2824
identified by similarity to GB:AAK22812.1; match to protein family HMM PF06676, Start=3473
putative lipoprotein, Start=4019
Predicted ORFs
testing, Score=1.23983 || Start=7252
50S ribosomal protein L35 [Bif
translation initiation Fac
asparagine
Manual

Open reading frames in pool1_scaffold42_1

Add new ORF

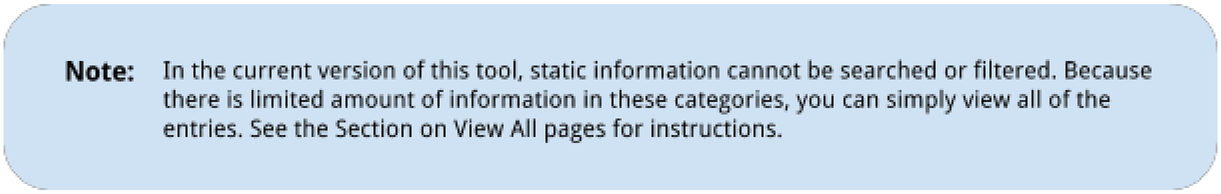
ID	Annotation	Sequence	Accession (if submitted to GenBank)	Start	Stop	ORF on complement strand?	Predicted by FGS	Prediction Score	Remove
19905	asparagine synthetase [Bifidobacterium adolescentis ATCC 15703]	GAGATTGTCA CGCATGAATC GGTTCGGAGG CGCGGCGGCG	None	8888	10879	True	True	1.25636	Remove
19906	hypothetical protein BAD_1078 [Bifidobacterium adolescentis ATCC 15703]	CCGCATAGCC ATGAAAATGA CGATTCTTTC CTTTCGTCCG	None	10900	11709	False	True	1.27195	Remove
19907	glyceraldehyde 3-phosphate dehydrogenase C [Bifidobacterium adolescentis ATCC 15703]	ACAGTTAAGA TTGGTATTAA CGGCTTTGGT CGTATCGGTC	None	11881	12936	False	True	1.23376	Remove
19908	hypothetical protein BAD_1080 [Bifidobacterium adolescentis ATCC 15703]	GGCGGCATGG CAAAGAAAAA GCATGAGAAG GGCGCCGGTT	None	13054	13569	False	True	1.2328	Remove

Figure 3-8: Cosmid Detail page

3.5 Search pages

All informations on the database can be accessed via the **Search** menu item on the navigation bar. The search functionality is divided based on the type of information that you are interested in viewing. The following is a list of all searchable categories:

- Cosmid
- Subclone
- Cosmid Assay
- Subclone Assay
- Open Reading Frame (ORF)
- Contigs



Note: In the current version of this tool, static information cannot be searched or filtered. Because there is limited amount of information in these categories, you can simply view all of the entries. See the Section on View All pages for instructions.

For each category, there are several types of searches available: **Basic Search, Advanced Search** and **Local BLAST**. Several options of the Basic Search and the Advanced Search are identical to the functionality of the View All (**Section 3.3**) pages, specifically sorting (**Section 3.3.3**) and exporting (**Section 3.3.4**).

3.5.1 Basic search

This will perform a simple keyword search of all fields associated with the category. For example, you can type in a Cosmid name, host name and a researcher's name and search. This will search all possible fields for all keywords that you enter. In general, all values that are associated with each record are searched. However, to see a list of all searchable fields, refer to the Advanced Search tab (**Section 3.5.2**).

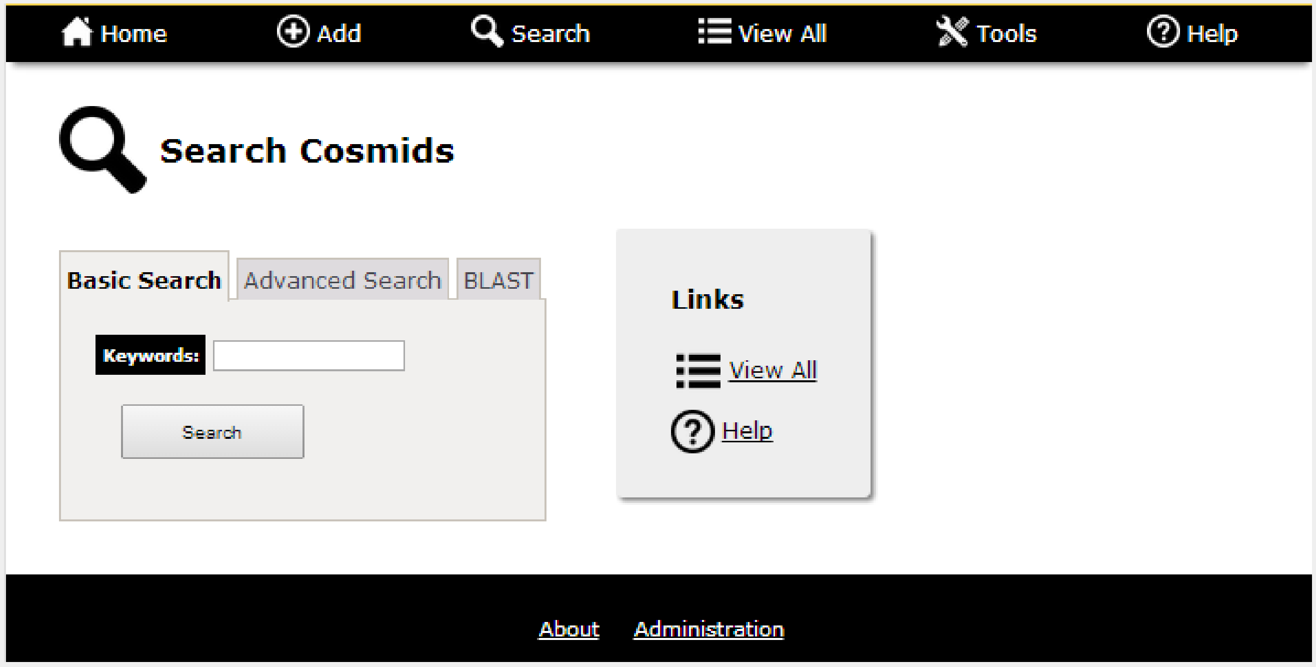
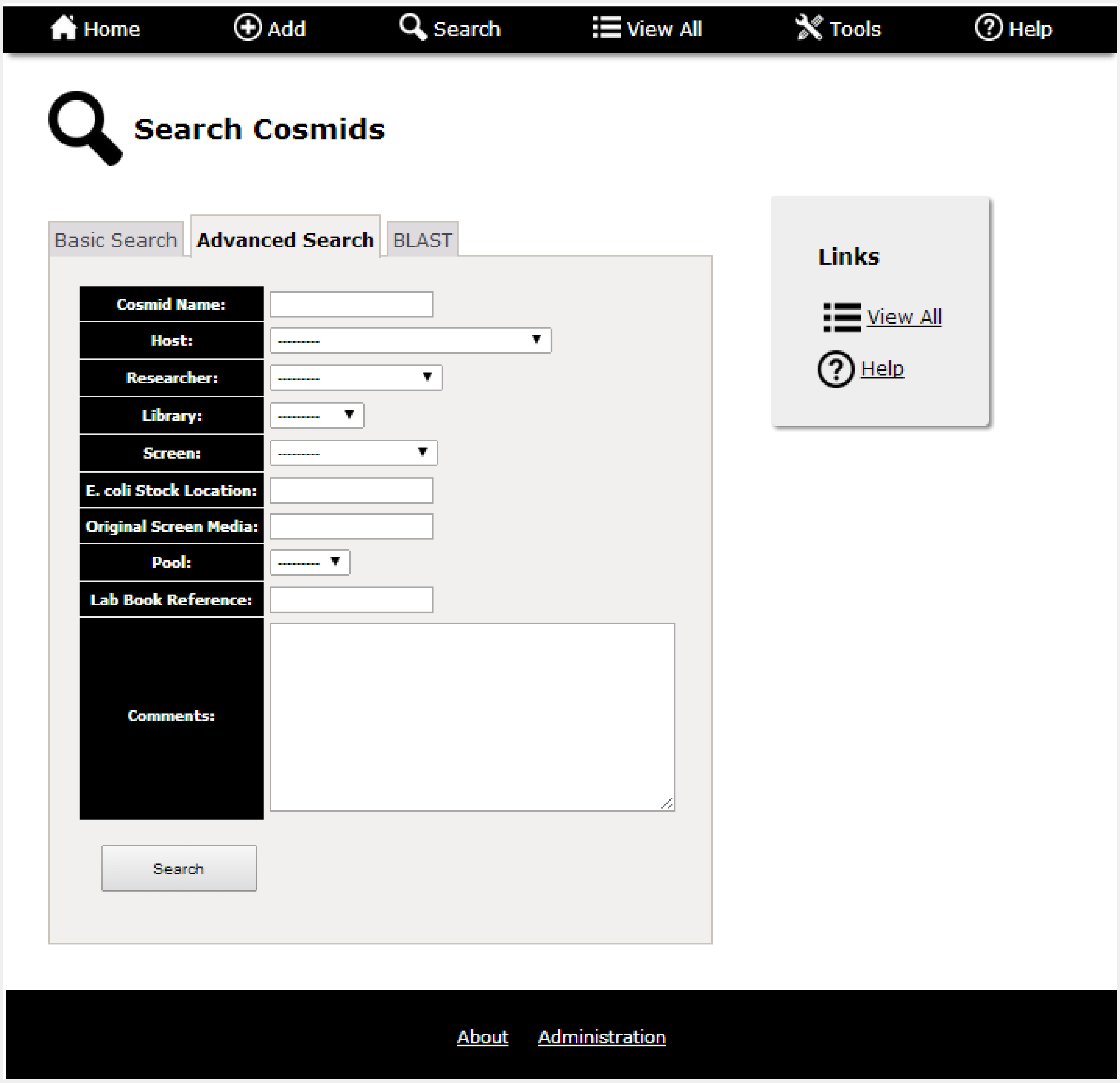


Figure 3-9: Basic search of Cosmid records

3.5.2 Advanced search

Advanced search allows you to search by specific values for each given field (**Figure 3-10**). For example, you can search for Cosmids by typing in a Cosmid name, then selecting a host and a researcher. Each submitted value will only be searched for in its designated field.

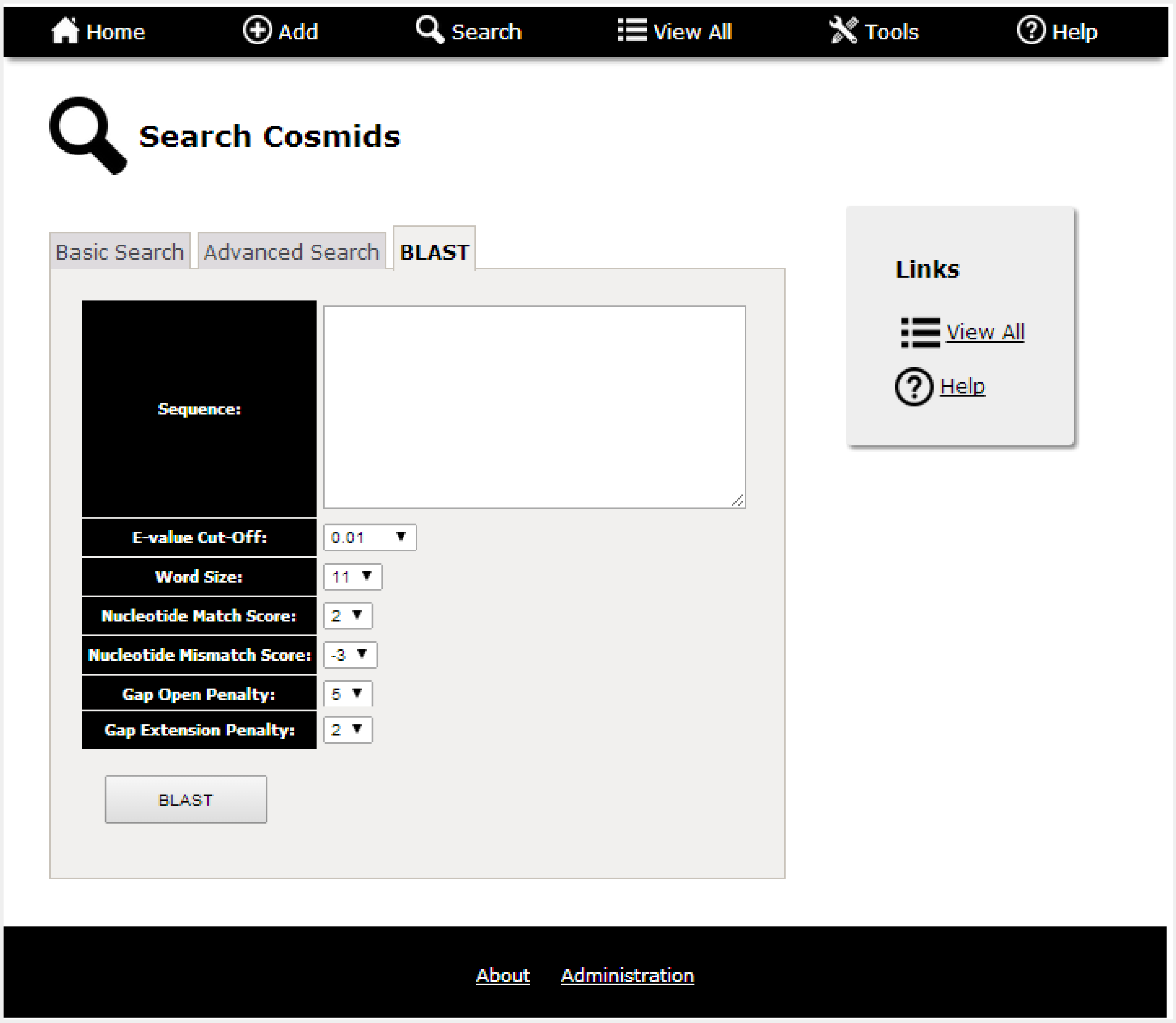


Cosmid Name:	
Host:	---------
Researcher:	---------
Library:	---------
Screen:	---------
E. coli Stock Location:	
Original Screen Media:	
Pool:	---------
Lab Book Reference:	
Comments:	

Figure 3-10: Advanced search of Cosmid records

3.5.3 Local BLAST search

The local BLAST search is only available for entries that are associated with sequences -- that is, Cosmids, Subclones, Contigs and ORFs. This search utilizes a locally installed version of BLAST. If you know exactly which type of record you are looking for (e.g., Cosmid, Contig, etc.), you can click on the appropriate field under the Search menu option and run a specific BLAST search. An example for Cosmid is shown in **Figure 3-11**.



Sequence:

E-value Cut-Off: 0.01

Word Size: 11

Nucleotide Match Score: 2

Nucleotide Mismatch Score: -3

Gap Open Penalty: 5

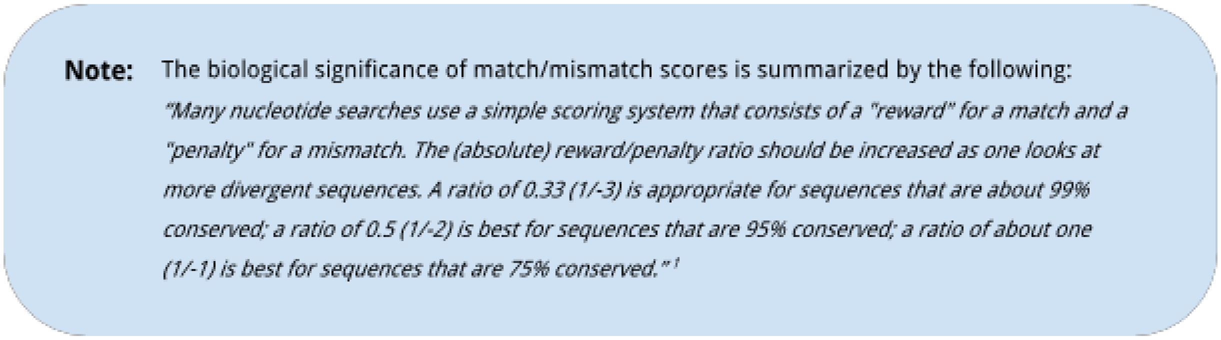
Gap Extension Penalty: 2

Figure 3-11: BLAST search of Cosmid sequences, which includes End-Tag, Contig, and ORF sequences.

If you would like to search all sequence-based fields (that is, all of Cosmids, Contigs, Subclones, and ORFs), you can use a general BLAST search using the **Sequences (BLAST)** page, again under the main Search menu option.

For all local BLAST searches the following parameters are available:

- ***E-value cut-off***: The statistical significance threshold for reporting matches against the sequences in the database.
- ***Word size***: BLAST uses this value as the minimum length of bases to align for the query and subject sequence on its "first pass". If there are any aligned areas, further BLAST algorithms are used to extend this initial alignment to obtain a more complete match. Increasing this value will result in more sensitive queries returned at the cost of processing time. Conversely, decreasing this value with result in a less sensitive query but will return quicker results.
- ***Nucleotide match score***: The value that the BLAST algorithm assigns to matching bases during sequence alignment.
- ***Nucleotide mismatch score***: The value that the BLAST algorithm assigns to bases that do not match (mismatch) during sequence alignment.
- ***Gap open penalty***: This is the value BLAST uses to subtract from a score if a gap is introduced to the sequence alignment.
- ***Gap extension penalty***: This is similar to 'Gap Open Penalty' except that this value is used to subtract from a score when a gap is extended.

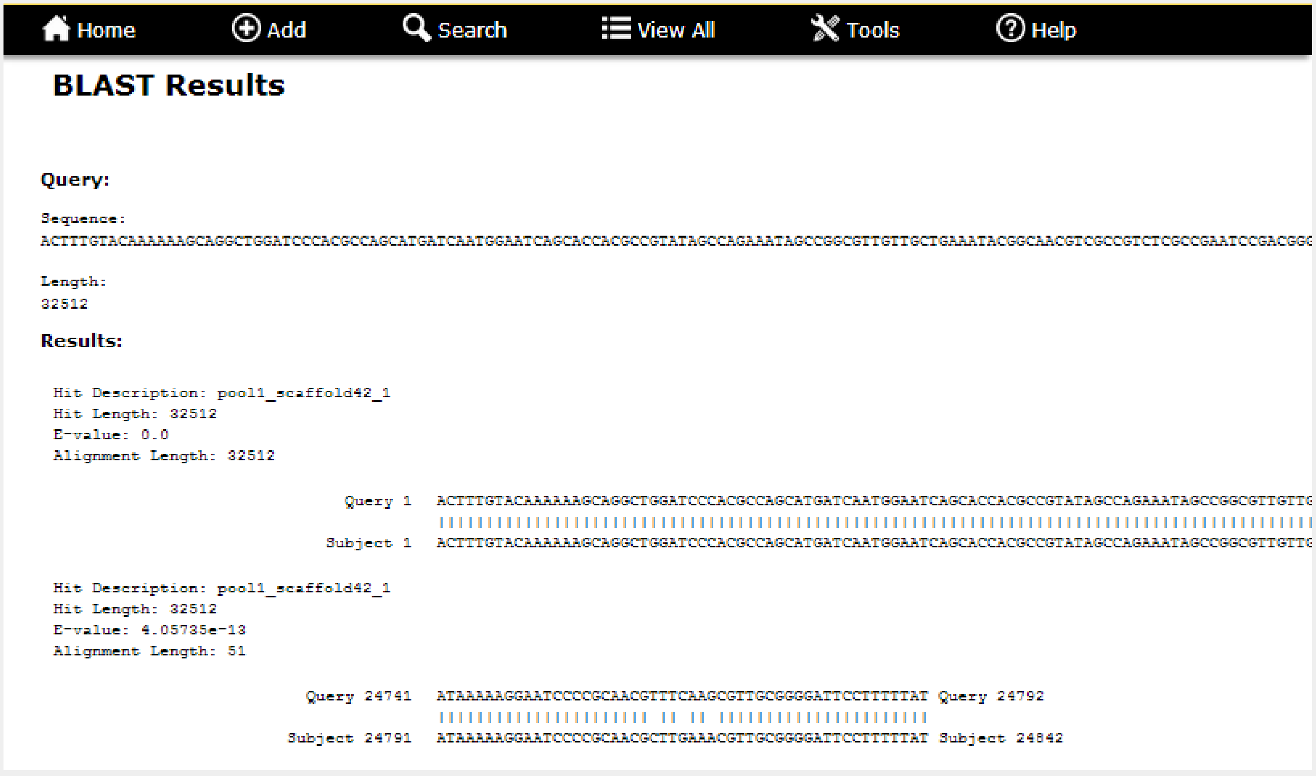


Note: The biological significance of match/mismatch scores is summarized by the following:
"Many nucleotide searches use a simple scoring system that consists of a "reward" for a match and a "penalty" for a mismatch. The (absolute) reward/penalty ratio should be increased as one looks at more divergent sequences. A ratio of 0.33 (1/-3) is appropriate for sequences that are about 99% conserved; a ratio of 0.5 (1/-2) is best for sequences that are 95% conserved; a ratio of about one (1/-1) is best for sequences that are 75% conserved." [1]

In a biological context, it is more likely to have several large gaps that result from extension of one gap opening, rather than many smaller dispersed gaps. In other words, the 'cost' of extending a gap is generally considered to be less than it is to initially introduce a new gap. More information regarding BLAST can be found at the NCBI website about BLAST parameters and the BLAST algorithm.

Note: All other parameters not specified here are set to their default values for BLAST. Additionally, the values of all parameters when a BLAST page is initially loaded are set to the default values used by NCBI blastn.

After you perform a BLAST search, you will be redirected to a results page (**Figure 3-12**). You will be able to see a list of sequences that have sequence similarity to your query sequence. The results include: the hit name, hit length, e-value of the HSP, and the alignment length.



BLAST Results

Query:

```
Sequence:
ACTTTGTACAAAAAAGCAGGCTGGATCCCACGCCAGCATGATCAATGGAATCAGCACCACGCCGTATAGCCAGAAATAGCCGGCGTTGTTGCTGAAATACGGCAACGTCGCCGTCTCGCCGAATCCGACGGG

Length:
32512
```

Results:

```
 Hit Description: pool1_scaffold42_1
 Hit Length: 32512
 E-value: 0.0
 Alignment Length: 32512

                              Query 1   ACTTTGTACAAAAAAGCAGGCTGGATCCCACGCCAGCATGATCAATGGAATCAGCACCACGCCGTATAGCCAGAAATAGCCGGCGTTGTTG
                                        |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
                            Subject 1   ACTTTGTACAAAAAAGCAGGCTGGATCCCACGCCAGCATGATCAATGGAATCAGCACCACGCCGTATAGCCAGAAATAGCCGGCGTTGTTG

 Hit Description: pool1_scaffold42_1
 Hit Length: 32512
 E-value: 4.05735e-13
 Alignment Length: 51

                          Query 24741   ATAAAAAGGAATCCCCGCAACGTTTCAAGCGTTGCGGGGATTCCTTTTTAT Query 24792
                                        |||||||||||||||||||||| || || ||||||||||||||||||||||
                        Subject 24791   ATAAAAAGGAATCCCCGCAACGCTTGAAACGTTGCGGGGATTCCTTTTTAT Subject 24842
```

Figure 3-12: Example of BLAST results after querying with a Contig sequence.

3.6 Add pages

Through the main navigation menu, any type of information can be added to the database by accessing the appropriate page through the **Add** menu option. The data that can be added are discussed in detail in the following subsections, and are summarized in the list below. For specific information regarding the requirements of each field see **Appendix 2**.

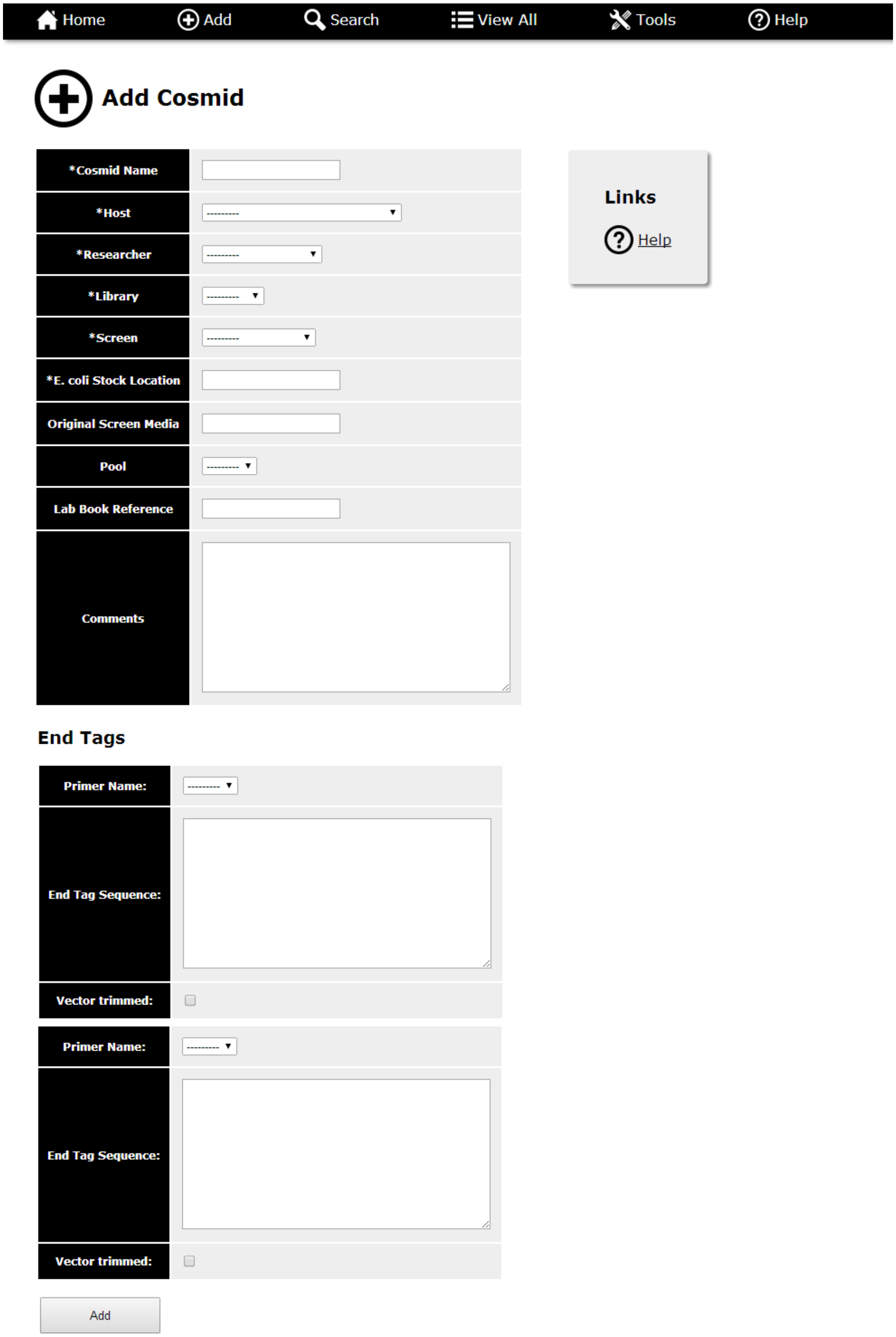
- Cosmid
- Subclone
- Cosmid Assay
- Subclone Assay
- ORF
- Contigs

Caution: Although data can be easily added and edited, they are not as easily removed. Deletion of data can only be done by a system administrator. Please be certain about your data before adding it to the website.

3.6.1 Add a Cosmid

The Add Cosmid page is likely to be the most heavily used Add page on this website (Figure 3-13). Though only several fields are required for addition of a Cosmid to the website, it is recommended that you add as much information as possible. This includes end-tag sequence data and sequencing pool information, if known. Upon clicking ‘Add’, you will be redirected to the Detail View for the Cosmid you have just added (recall Section 3.4).

Note: There is currently no way to associate multiple cosmids with a sequencing pool. It is recommended you add the pool information when you add the cosmids.



Add Cosmid

*Cosmid Name	
*Host	---------
*Researcher	---------
*Library	---------
*Screen	---------
*E. coli Stock Location	
Original Screen Media	
Pool	---------
Lab Book Reference	
Comments	

End Tags

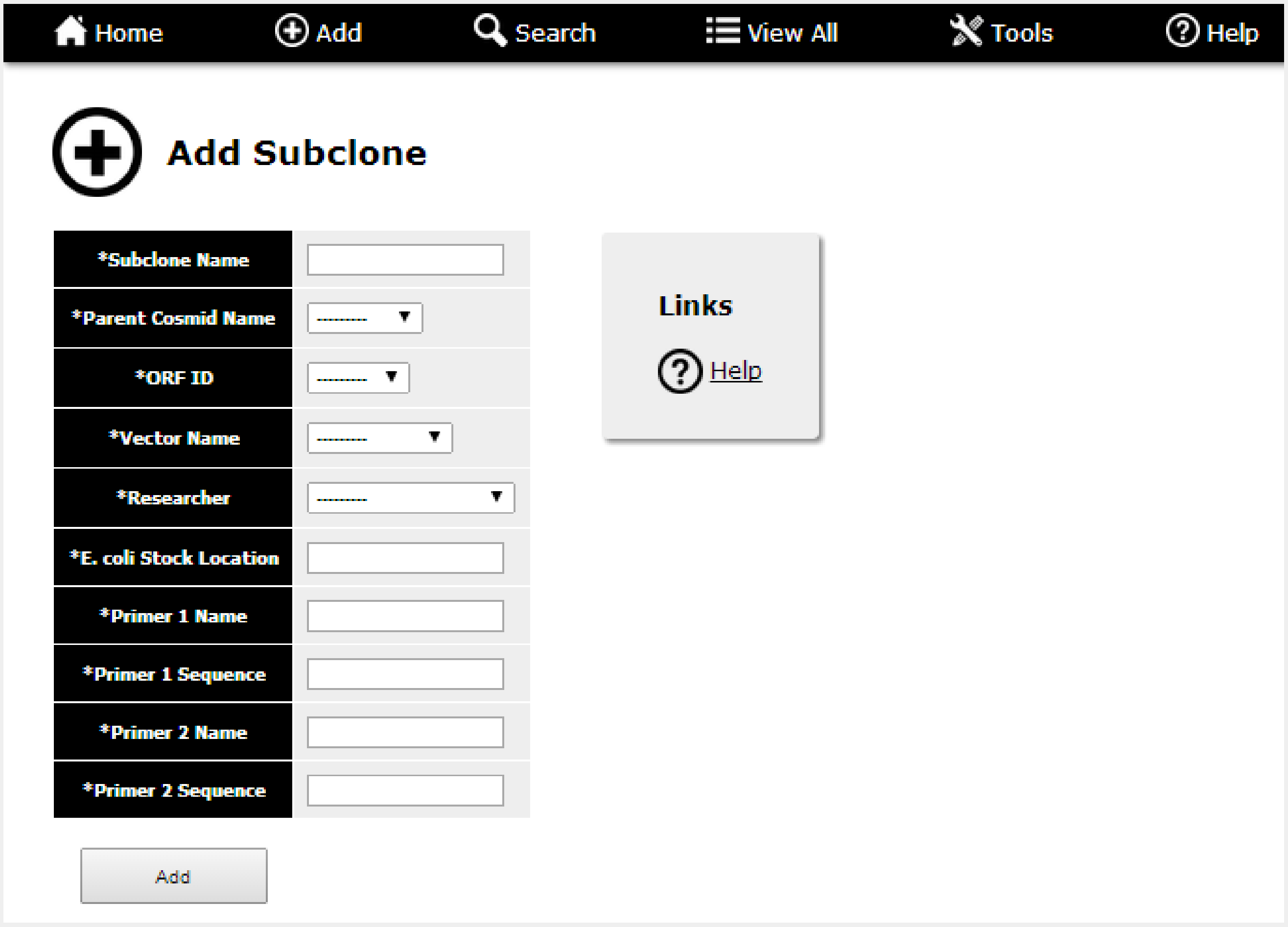
Primer Name:	---------
End Tag Sequence:	
Vector trimmed:	□
Primer Name:	---------
End Tag Sequence:	
Vector trimmed:	□

Figure 3-13: Add Cosmid page.

Note: Required field on Add pages are always indicated by an asterisk (*).

3.6.2 Add a Subclone, Subclone Assay or Cosmid Assay

The three Add pages for Subclone, Subclone Assay and Cosmid Assay have very similar forms and are all straight-forward as they contain fewer fields (unlike the Add Cosmid form). Upon clicking “Add”, you will be redirected to the View All page for the respective category type (recall **Section 3.3**).



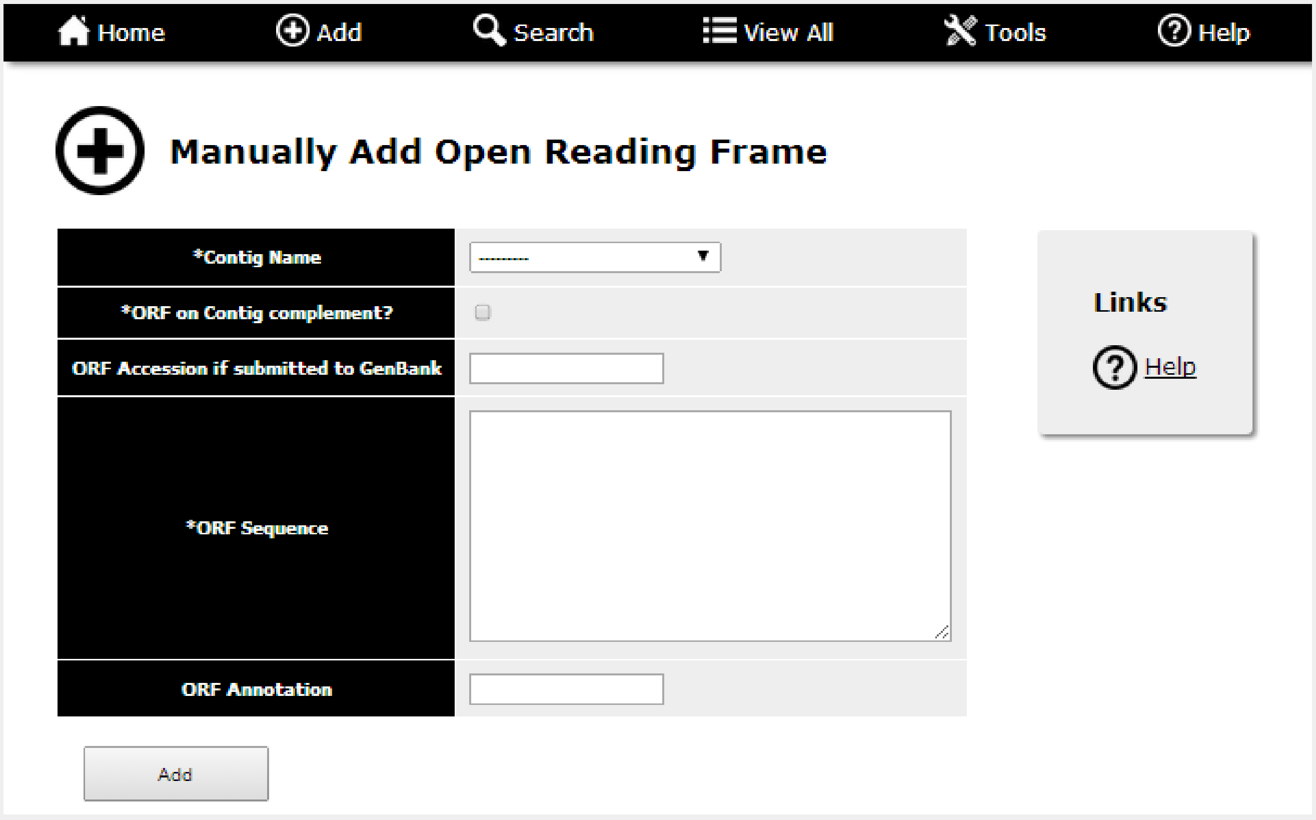
Add Subclone

*Subclone Name	
*Parent Cosmid Name	---------
*ORF ID	---------
*Vector Name	---------
*Researcher	---------
*E. coli Stock Location	
*Primer 1 Name	
*Primer 1 Sequence	
*Primer 2 Name	
*Primer 2 Sequence	

Figure 3-14: Add Subclone page.

3.6.3 Add an ORF

Although Contigs can be submitted for automatic ORF finding and annotation (**Section 4.2**), there may be instances when you wish to add an ORF manually. To do so, use the Add ORF form (**Figure 3-15**). After selecting a Contig and the correct a strand, inputting an ORF sequence, and clicking “Add”, the website will validate that the ORF sequence you specified is found on the Contig. It will also automatically save the start and stop positions t be used for Contig image generation (see **Section 4.2** for more details on image generation).



Manually Add Open Reading Frame

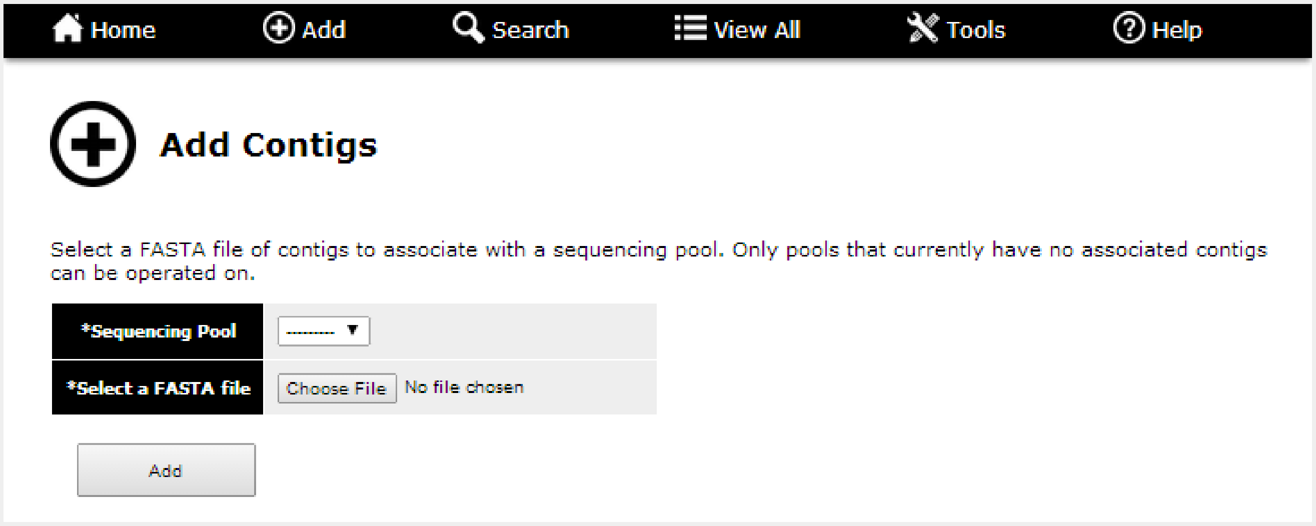
*Contig Name	--------- ▼
*ORF on Contig complement?	□
ORF Accession if submitted to GenBank	
*ORF Sequence	
ORF Annotation	

Figure 3-15: Add Open Reading Frame page.

Note: This Add page is for manually adding ORFs to specific Contigs. This is to be used when the automatic ORF Finding and Annotation tool does not locate an ORF that you have evidence to believe is present. Please see **Section 4.2** for more information on the ORF Finding tool.

3.6.4 Add a FASTA file of Contigs

New sequence information obtained from an Illumina pooled sequencing run can be added through the Add Contigs form. In order to add a set of Contigs, the user must first select the appropriate sequencing pool that the Contigs are to be associated with. This sequencing pool must be added prior to the Contigs, and must be done by an administrator. Once the user has selected a pool and their desired FASTA file containing all of the Contig sequences, clicking “Add” will ensure the file is in proper FASTA format and, if so, will store all information in the database. If the uploaded FASTA file is not in the proper FASTA format, an error will be returned.



Add Contigs

Select a FASTA file of contigs to associate with a sequencing pool. Only pools that currently have no associated contigs can be operated on.

*Sequencing Pool	--------- ▼
*Select a FASTA file	Choose File No file chosen

Figure 3-16: Add Contigs page.

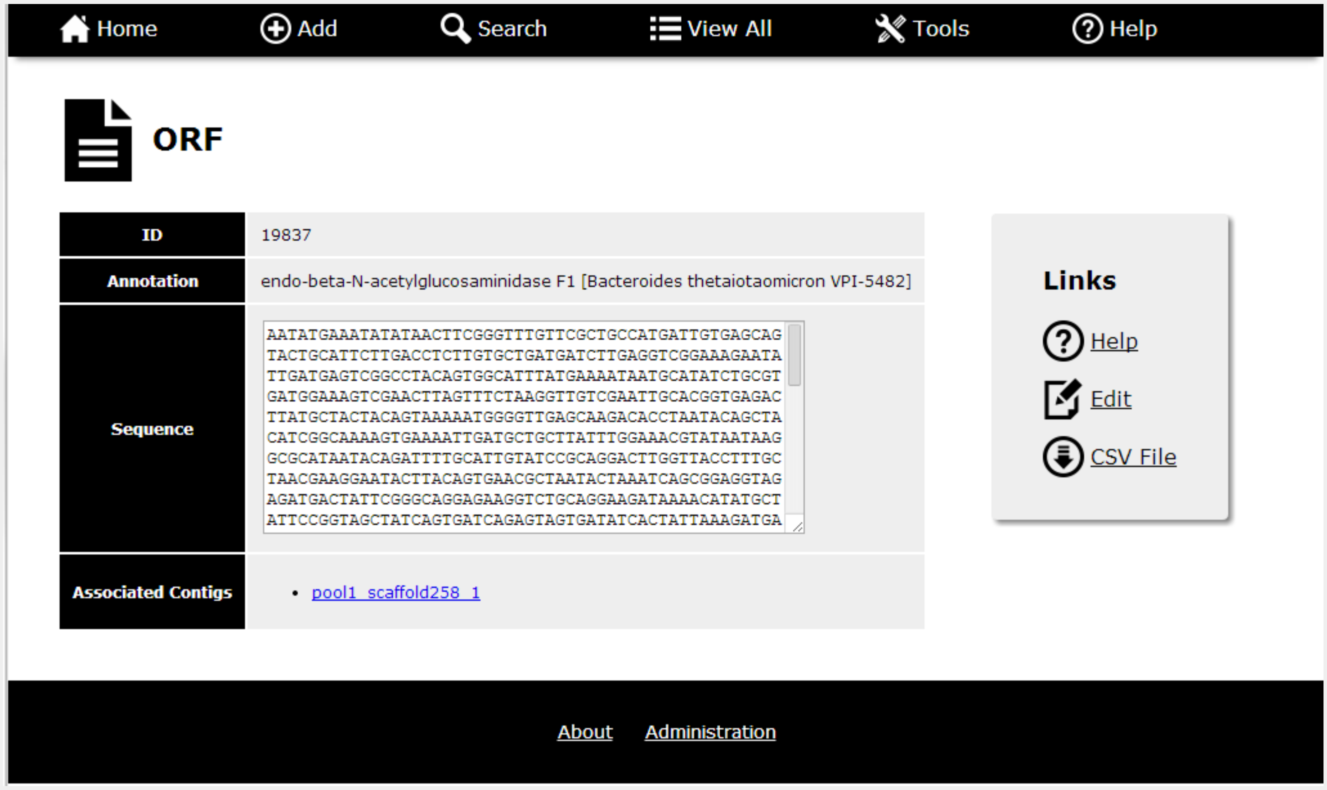
3.6.5 Static data

As mentioned in **Section 3.3.2**, reference data cannot be added by the typical user because these data require careful curation and validation.

Note: This information can only be edited or added by administrators. If you find that one these entries is incorrect or if you need to use one of these entries and it does not exist, please contact a site administrator.

3.7 Edit pages

From any Detail page, you can navigate to the respective Edit page. For example, to edit a particular ORF, locate the ORF record of interest (e.g., **Figure 3-17**) and look for the edit icon under the Links section of the page. You will be directed to a page where you can make changes (**Figure 3-18**).

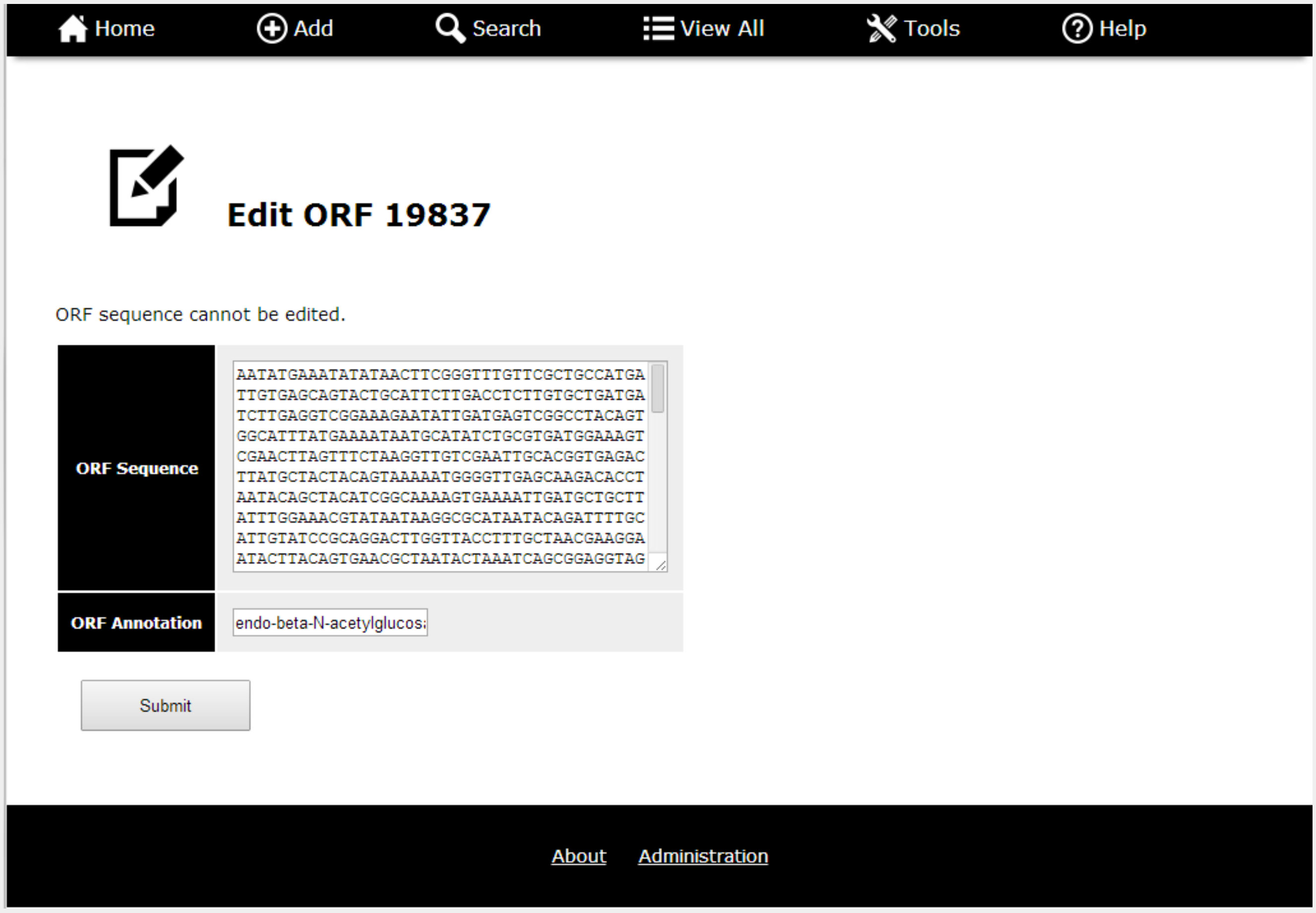


ID	19837
Annotation	endo-beta-N-acetylglucosaminidase F1 [Bacteroides thetaiotaomicron VPI-5482]
Sequence	AATATGAAATATATAACTTCGGGTTTGTTCGCTGCCATGATTGTGAGCAG TACTGCATTCTTGACCTCTTGTGCTGATGATCTTGAGGTCGGAAAGAATA TTGATGAGTCGGCCTACAGTGGCATTTATGAAAATAATGCATATCTGCGT GATGGAAAGTCGAACTTAGTTTCTAAGGTTGTCGAATTGCACGGTGAGAC TTATGCTACTACAGTAAAAATGGGGTTGAGCAAGACACCTAATACAGCTA CATCGGCAAAAGTGAAAATTGATGCTGCTTATTTGGAAACGTATAATAAG GCGCATAATACAGATTTTGCATTGTATCCGCAGGACTTGGTTACCTTTGC TAACGAAGGAATACTTACAGTGAACGCTAATACTAAATCAGCGGAGGTAG AGATGACTATTCGGGCAGGAGAAGGTCTGCAGGAAGATAAAACATATGCT ATTCCGGTAGCTATCAGTGATCAGAGTAGTGATATCACTATTAAAGATGA
Associated Contigs	- pool1_scaffold258_1

Figure 3-17: Detail page for ORF record.

Note: If you do not see an Edit icon, it is likely that your account does not have permission to edit this information. Contact an administrator to request the appropriate permissions to edit database information.

Much of the information found on the website is editable. However, in some special cases, information can not be edited. For example, the ORF sequence cannot be edited by any users (**Figure 3-18**); this is because it has been identified by software from the ORF Finding and Annotation tool.



Edit ORF 19837

ORF sequence cannot be edited.

ORF Sequence	AATATGAAATATATAACTTCGGGTTTGTTCGCTGCCATGA TTGTGAGCAGTACTGCATTCTTGACCTCTTGTGCTGATGA TCTTGAGGTCGGAAAGAATATTGATGAGTCGGCCTACAGT GGCATTTATGAAAATAATGCATATCTGCGTGATGGAAAGT CGAACTTAGTTTCTAAGGTTGTCGAATTGCACGGTGAGAC TTATGCTACTACAGTAAAAATGGGGTTGAGCAAGACACCT AATACAGCTACATCGGCAAAAGTGAAAATTGATGCTGCTT ATTTGGAAACGTATAATAAGGCGCATAATACAGATTTTGC ATTGTATCCGCAGGACTTGGTTACCTTTGCTAACGAAGGA ATACTTACAGTGAACGCTAATACTAAATCAGCGGAGGTAG
ORF Annotation	endo-beta-N-acetylglucos:

Figure 3-18: Edit page for ORF record.

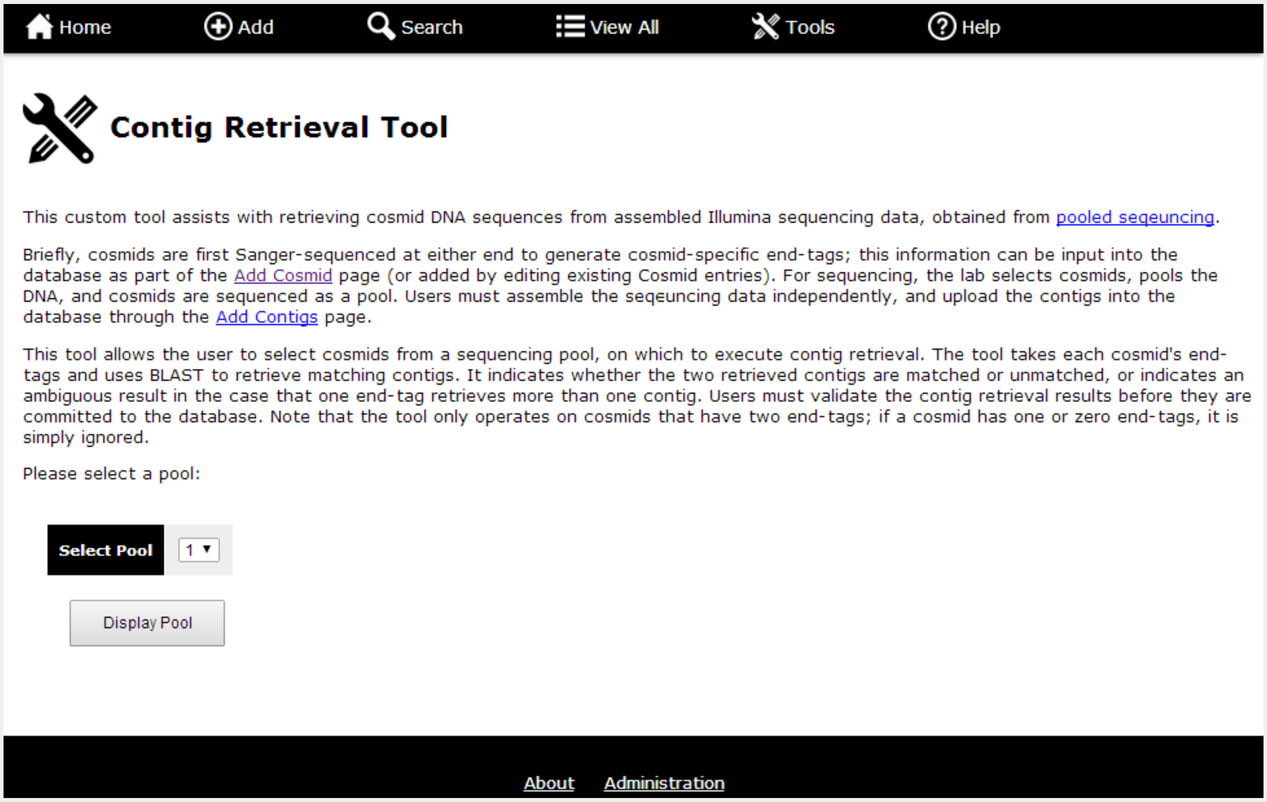
4 Website custom tools

4.1 Contig retrieval tool

This tool matches Contigs to Cosmids within a particular pool by aligning the Cosmid end tag's to Contig sequences and determining the best matches between them. A BLAST search is performed with each Cosmid's end-tags to retrieve the Contigs that match each end-tag. The result is a list of the top matches for each end-tag, of each Cosmid.

These matches are categorized into three types: a match when both end-tags retrieve the same Contig, a mismatch when both end-tags retrieve different Contigs and ambiguous when one end tag retrieves more than one Contig. Users must validate the Contig retrieval results before they are committed to the database.

Select a sequencing pool from the dropdown menu on the Contig Retrieval Tool page (**Figure 4-1**).



Contig Retrieval Tool

This custom tool assists with retrieving cosmid DNA sequences from assembled Illumina sequencing data, obtained from pooled seqeuncing.

Briefly, cosmids are first Sanger-sequenced at either end to generate cosmid-specific end-tags; this information can be input into the database as part of the Add Cosmid page (or added by editing existing Cosmid entries). For sequencing, the lab selects cosmids, pools the DNA, and cosmids are sequenced as a pool. Users must assemble the seqeuncing data independently, and upload the contigs into the database through the Add Contigs page.

This tool allows the user to select cosmids from a sequencing pool, on which to execute contig retrieval. The tool takes each cosmid's end-tags and uses BLAST to retrieve matching contigs. It indicates whether the two retrieved contigs are matched or unmatched, or indicates an ambiguous result in the case that one end-tag retrieves more than one contig. Users must validate the contig retrieval results before they are committed to the database. Note that the tool only operates on cosmids that have two end-tags; if a cosmid has one or zero end-tags, it is simply ignored.

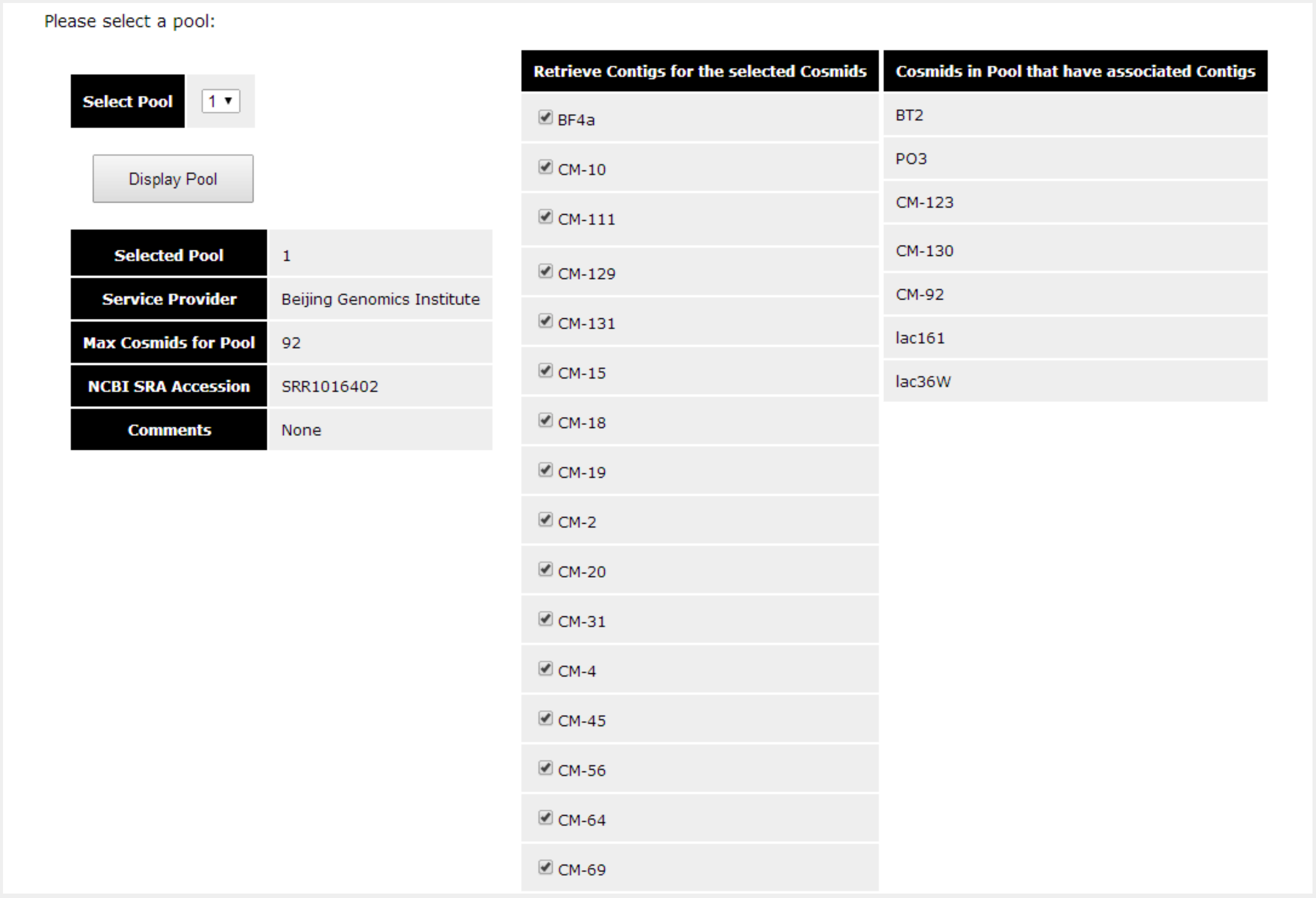
Please select a pool:

Select Pool 1 ▼

Figure 4-1: Selecting a specific sequencing pool to use for Contig tool retrieval tool.

To show the details for the selected pool, click the **Display Pool** button (**Figure 4-1**). A list of Cosmids associated with the pool will be displayed which the user can select from (**Figure 4-2**). If no Cosmids are displayed or if a particular Cosmid is not present, you must either add a new Cosmid or edit an existing one in the selected pool. All Cosmids that are already associated with Contigs within the selected pool will be displayed in a table to the right of the pool information.

Caution: This tool can only operate on cosmids that have two end-tags; if a cosmid has zero or one end-tag, it will be ignored. Refer to the cosmid page to ensure there is a forward and reverse end-tag associated to your cosmid.



Please select a pool:

Select Pool	1 ▾

Selected Pool	1
Service Provider	Beijing Genomics Institute
Max Cosmids for Pool	92
NCBI SRA Accession	SRR1016402
Comments	None

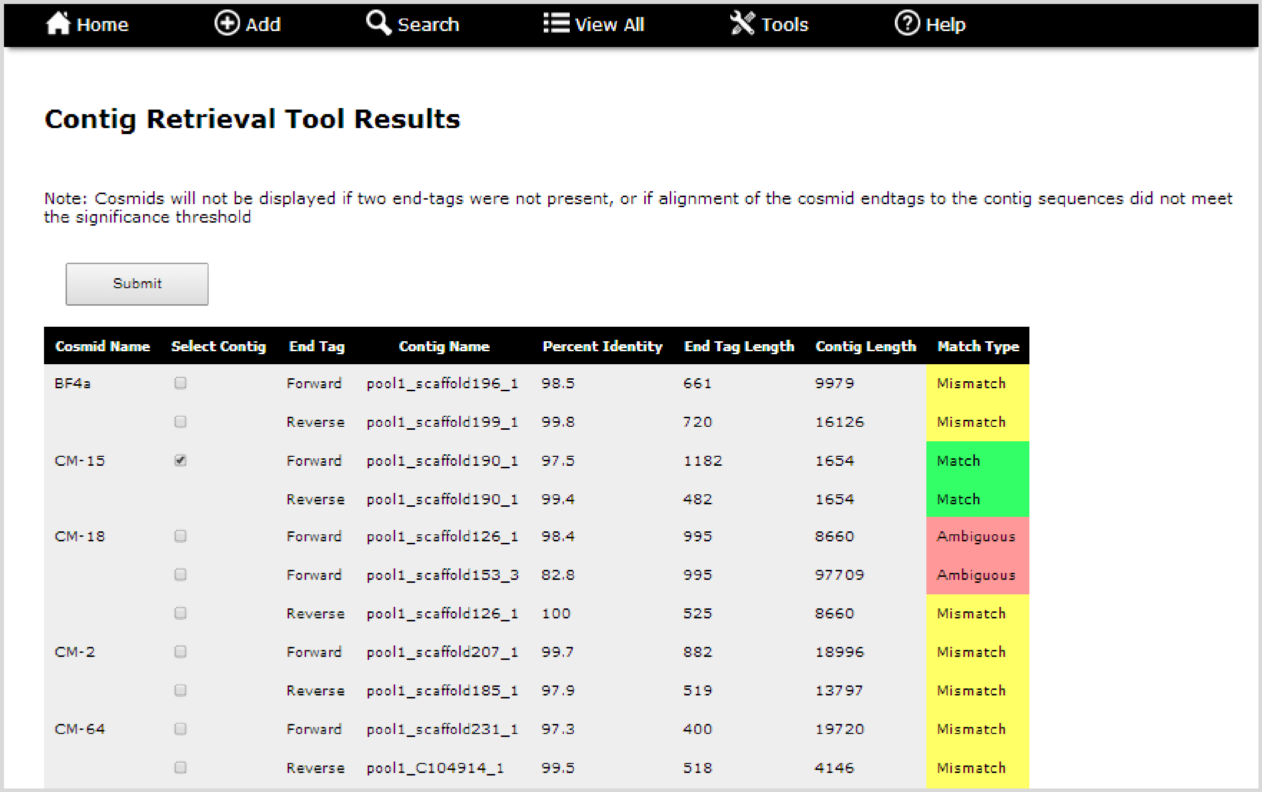
Retrieve Contigs for the selected Cosmids
☑ BF4a
☑ CM-10
☑ CM-111
☑ CM-129
☑ CM-131
☑ CM-15
☑ CM-18
☑ CM-19
☑ CM-2
☑ CM-20
☑ CM-31
☑ CM-4
☑ CM-45
☑ CM-56
☑ CM-64
☑ CM-69

Cosmids in Pool that have associated Contigs
BT2
PO3
CM-123
CM-130
CM-92
lac161
lac36W

Figure 4-2: Selecting Cosmids to retrieve their associated Contigs for Pool 1 in the middle table. The table to the right shows Cosmids that already have Contigs associated with them.

After clicking on **Submit** (**Figure 4-2**), both end-tags for the selected Cosmid(s) will run through the Contig-Retrieval step which uses BLAST to align them to Contigs in the database. The user will be redirected to a page where the results will be displayed in a table-format; this results require user validation before they are inputted into the database (**Figure 4-3**). A specific "Match Type" tag will be displayed next to Contig name indicating whether the two retrieved Contigs match, mismatch, or if the results are ambiguous.

Caution: Ambiguous matches will be coloured in red and will require user intervention. This indicates that one endtag significantly aligned to two or more contigs. The percent identity and sequence length is provided to aid with this decision.



Contig Retrieval Tool Results

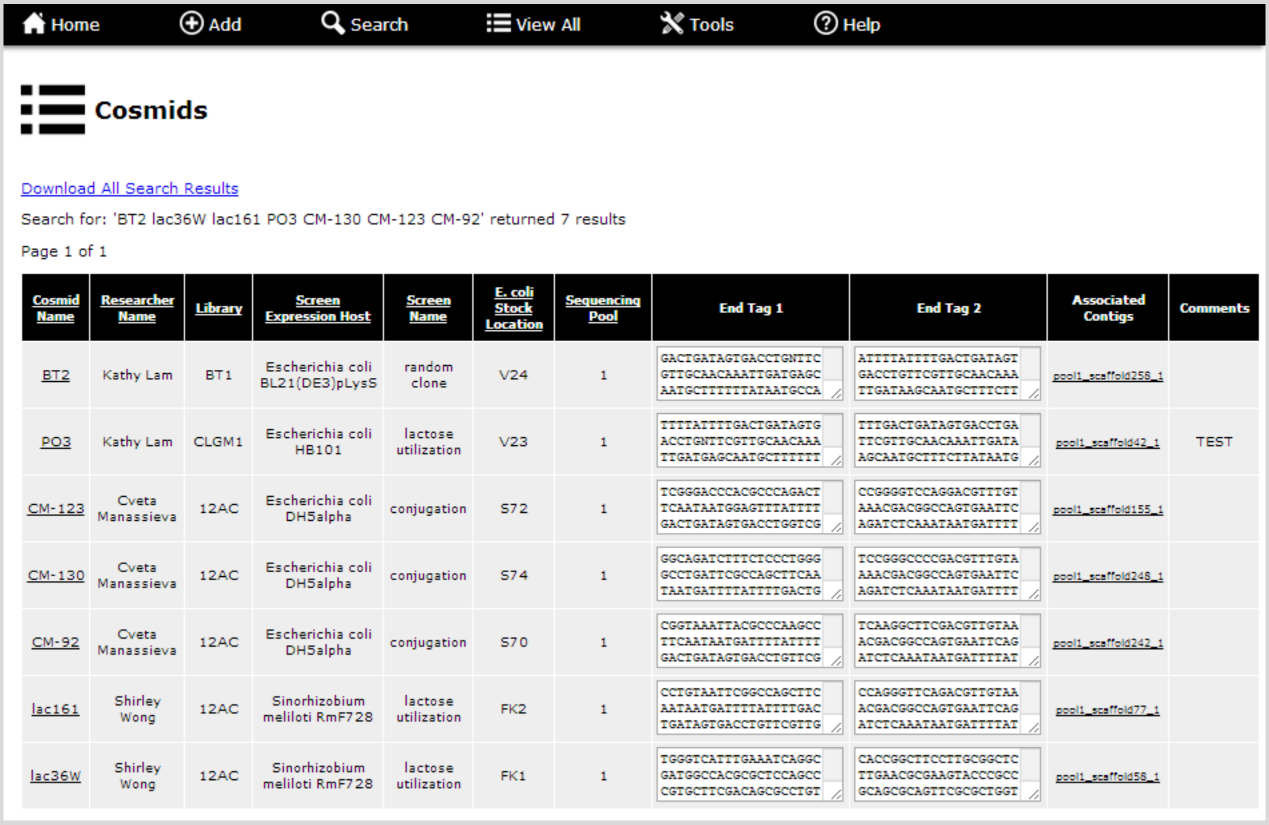
Note: Cosmids will not be displayed if two end-tags were not present, or if alignment of the cosmid endtags to the contig sequences did not meet the significance threshold

Cosmid Name	Select Contig	End Tag	Contig Name	Percent Identity	End Tag Length	Contig Length	Match Type
BF4a	□	Forward	pool1_scaffold196_1	98.5	661	9979	Mismatch
	□	Reverse	pool1_scaffold199_1	99.8	720	16126	Mismatch
CM-15	☑	Forward	pool1_scaffold190_1	97.5	1182	1654	Match
		Reverse	pool1_scaffold190_1	99.4	482	1654	Match
CM-18	□	Forward	pool1_scaffold126_1	98.4	995	8660	Ambiguous
	□	Forward	pool1_scaffold153_3	82.8	995	97709	Ambiguous
	□	Reverse	pool1_scaffold126_1	100	525	8660	Mismatch
CM-2	□	Forward	pool1_scaffold207_1	99.7	882	18996	Mismatch
	□	Reverse	pool1_scaffold185_1	97.9	519	13797	Mismatch
CM-64	□	Forward	pool1_scaffold231_1	97.3	400	19720	Mismatch
	□	Reverse	pool1_C104914_1	99.5	518	4146	Mismatch

Figure 4-3: Contig retrieval results requiring user validation. Match type is indicated in green, yellow, and red.

A “match” scenario is one where the Contigs retrieved for both end-tags are the same, e.g. BF4 endtag 1 and endtag 2 both retrieve scaffold15_1. A “mismatch” scenario is one where the Contigs retrieved for both end-tags are different from each other, e.g. BT2 endtag 1 retrieves scaffold77_1 while BT2 endtag 2 retrieves scaffold258_1. An “ambiguous” scenario is where one or both end-tags retrieves two or more Contigs, thus requiring the end-user to analyze the retrieval, e.g. CM10 endtag 1 retrieves scaffold42_1, while CM10 endtag 2 retrieves scaffold248_1 and scaffold48_1. Only Cosmids with the match scenario will be checked by default, however any Contigs can be added or removed from any matched Cosmids. To aid the user in selecting significant retrievals, the percent identity and lengths for each sequence are also displayed.

Upon clicking **Submit** (**Figure 4-3**), the now user-authenticated Cosmid-Contig pairs will be stored in the database. Once the Contigs have been validated and submitted, the user will be redirected to the Cosmid page to view the finalized Cosmid-Contig pairs (**Figure 4-4**).



Cosmids

Download All Search Results

Search for: 'BT2 lac36W lac161 PO3 CM-130 CM-123 CM-92' returned 7 results

Page 1 of 1

Cosmid Name	Researcher Name	Library	Screen Expression Host	Screen Name	E. coli Stock Location	Sequencing Pool	End Tag 1	End Tag 2	Associated Contigs	Comments
BT2	Kathy Lam	BT1	Escherichia coli BL21(DE3)pLysS	random clone	V24	1	GACTGATAGTGACCTGNTTC GTTGCAACAAATTGATGAGC AATGCTTTTTTATAATGCCA	ATTTTATTTTGACTGATAGT GACCTGTTCGTTGCAACAAA TTGATAAGCAATGCTTTCTT	pool1_scaffold258_1	
PO3	Kathy Lam	CLGM1	Escherichia coli HB101	lactose utilization	V23	1	TTTTATTTTGACTGATAGTG ACCTGNTTCGTTGCAACAAA TTGATGAGCAATGCTTTTTT	TTTGACTGATAGTGACCTGA TTCGTTGCAACAAATTGATA AGCAATGCTTTCTTATAATG	pool1_scaffold42_1	TEST
CM-123	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S72	1	TCGGGACCCACGCCCAGACT TCAATAATGGAGTTTATTTT GACTGATAGTGACCTGGTCG	CCGGGGTCCAGGACGTTTGT AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT	pool1_scaffold155_1	
CM-130	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S74	1	GGCAGATCTTTCTCCCTGGG GCCTGATTCGCCAGCTTCAA TAATGATTTTATTTTGACTG	TCCGGGCCCCGACGTTTGTA AAACGACGGCCAGTGAATTC AGATCTCAAATAATGATTTT	pool1_scaffold248_1	
CM-92	Cveta Manassieva	12AC	Escherichia coli DH5alpha	conjugation	S70	1	CGGTAAATTACGCCCAAGCC TTCAATAATGATTTTATTTT GACTGATAGTGACCTGTTCG	TCAAGGCTTCGACGTTGTAA ACGACGGCCAGTGAATTCAG ATCTCAAATAATGATTTTAT	pool1_scaffold242_1	
lac161	Shirley Wong	12AC	Sinorhizobium meliloti RmF728	lactose utilization	FK2	1	CCTGTAATTCGGCCAGCTTC AATAATGATTTTATTTTGAC TGATAGTGACCTGTTCGTTG	CCAGGGTTCAGACGTTGTAA ACGACGGCCAGTGAATTCAG ATCTCAAATAATGATTTTAT	pool1_scaffold77_1	
lac36W	Shirley Wong	12AC	Sinorhizobium meliloti RmF728	lactose utilization	FK1	1	TGGGTCATTTGAAATCAGGC GATGGCCACGCGCTCCAGCC CGTGCTTCGACAGCGCCTGT	CACCGGCTTCCTTGCGGCTC TTGAACGCGAAGTACCCGCC GCAGCGCAGTTCGCGCTGGT	pool1_scaffold58_1	

Figure 4-4: After confirming Contig retrieval results the user will be redirected to a list of all Cosmids that now have associated Contigs. Notice on the right side of the table it will show the newly associated Contigs.

4.2 ORF finding and annotation tool

This tool assists with annotating Contig sequences to identify genes responsible for the function of interest. The tool can only operate on Contigs that have been associated to a Cosmid clone through the use of the Contig Retrieval tool. This tool can only run on Contigs that have not yet been annotated, therefore, it can only run once per Contig.

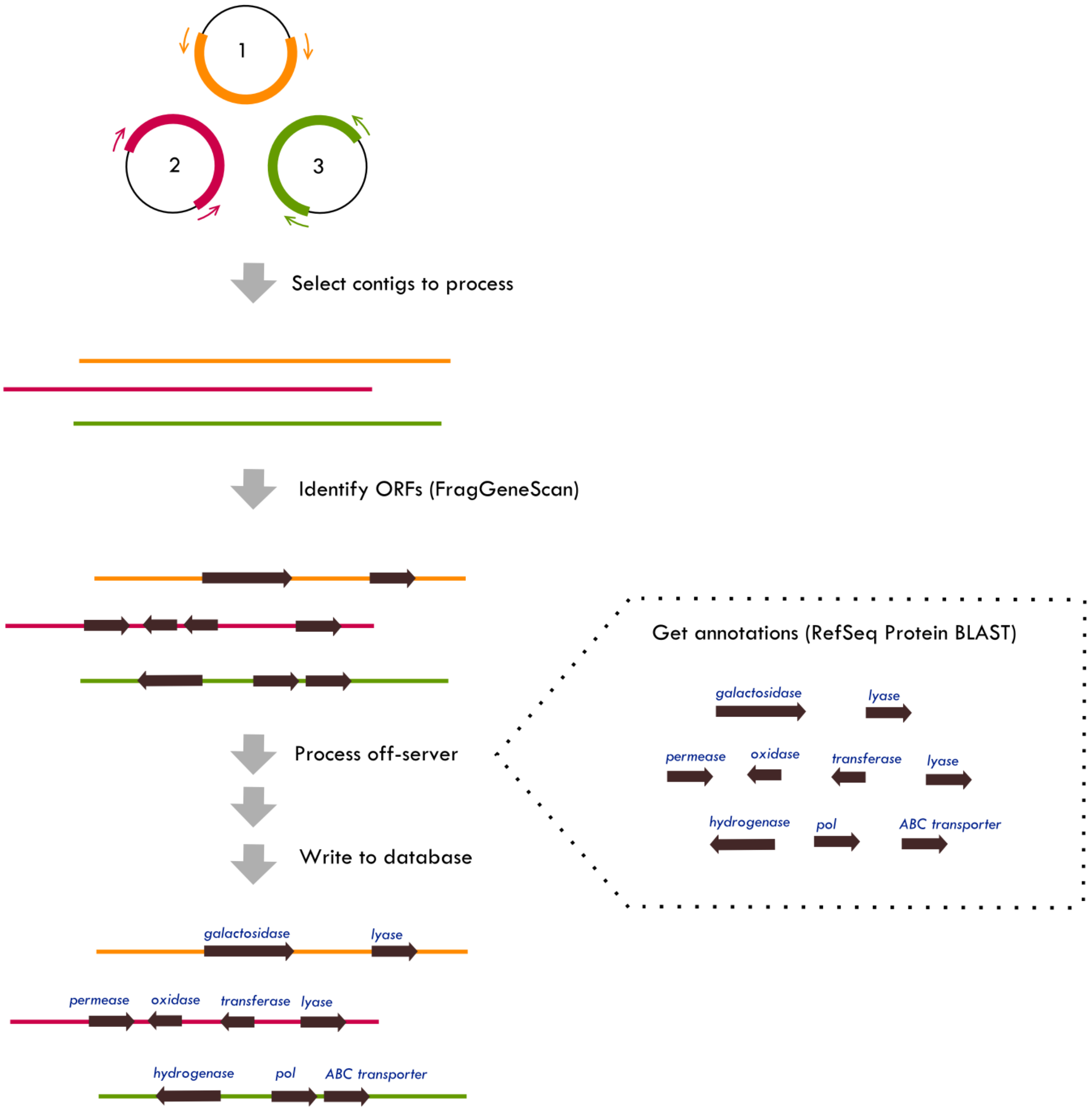
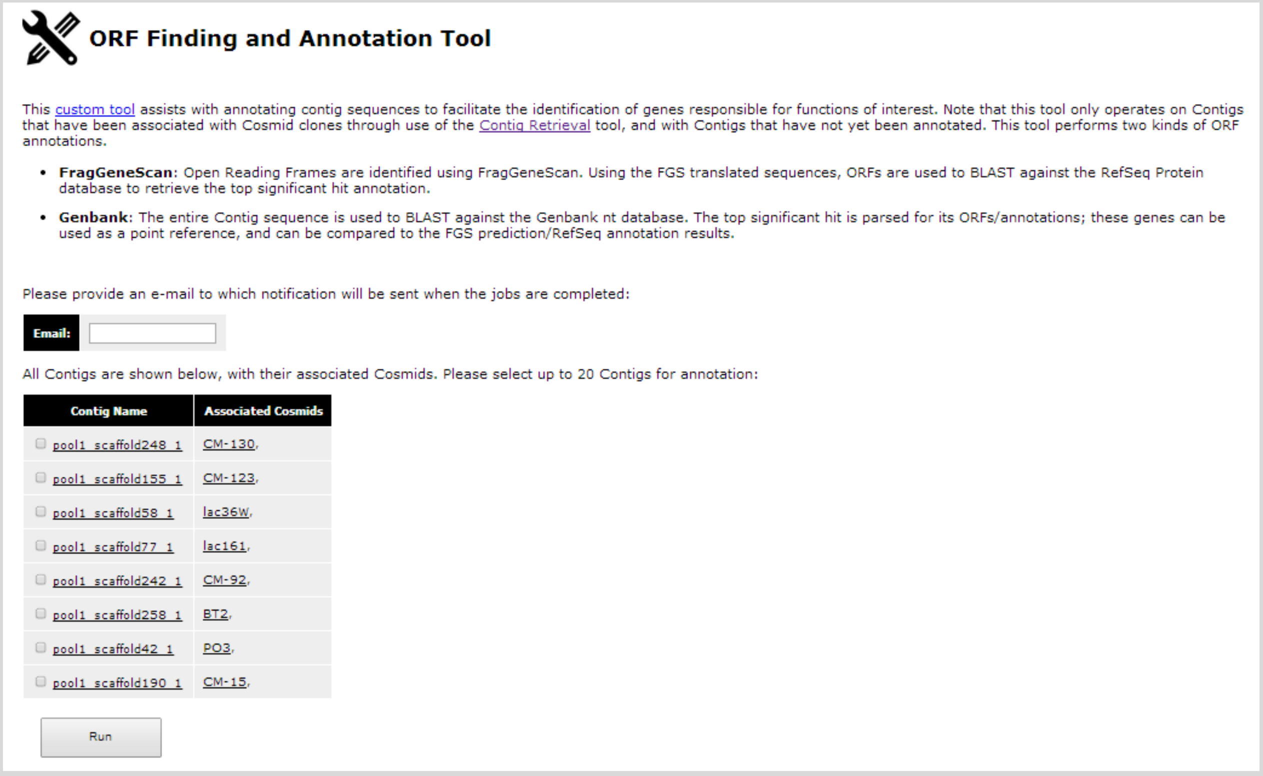


Figure 4-5: Overview of ORF finding and gene annotation tool, using two distinct computing servers

4.2.1 Usage of the tool

Upon loading the ORF Finding and Annotation Tool page a list of potential Contigs to be annotated are listed for selection (**Figure 4-6**). Additionally, an e-mail field is provided for the purpose of automatic notifications to be sent out updating the user regarding which Contigs they queued and when the job is complete.



ORF Finding and Annotation Tool

This custom tool assists with annotating contig sequences to facilitate the identification of genes responsible for functions of interest. Note that this tool only operates on Contigs that have been associated with Cosmid clones through use of the Contig Retrieval tool, and with Contigs that have not yet been annotated. This tool performs two kinds of ORF annotations.

- **FragGeneScan**: Open Reading Frames are identified using FragGeneScan. Using the FGS translated sequences, ORFs are used to BLAST against the RefSeq Protein database to retrieve the top significant hit annotation.
- **Genbank**: The entire Contig sequence is used to BLAST against the Genbank nt database. The top significant hit is parsed for its ORFs/annotations; these genes can be used as a point reference, and can be compared to the FGS prediction/RefSeq annotation results.

Please provide an e-mail to which notification will be sent when the jobs are completed:

Email:

All Contigs are shown below, with their associated Cosmids. Please select up to 20 Contigs for annotation:

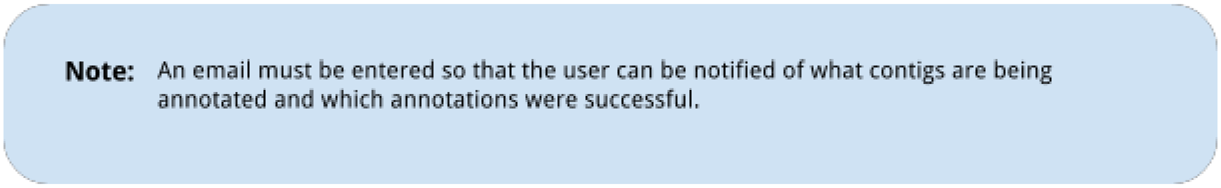
Contig Name	Associated Cosmids
□ pool1_scaffold248_1	CM-130,
□ pool1_scaffold155_1	CM-123,
□ pool1_scaffold58_1	lac36W,
□ pool1_scaffold77_1	lac161,
□ pool1_scaffold242_1	CM-92,
□ pool1_scaffold258_1	BT2,
□ pool1_scaffold42_1	PO3,
□ pool1_scaffold190_1	CM-15,

Figure 4-6: Main page of the ORF Finding and Annotation Tool.

Note: Only a fraction of all Contigs in the database are listed in the Gene Annotation Tool. This is because the tool can only be executed on Contigs that meet the following two criteria:

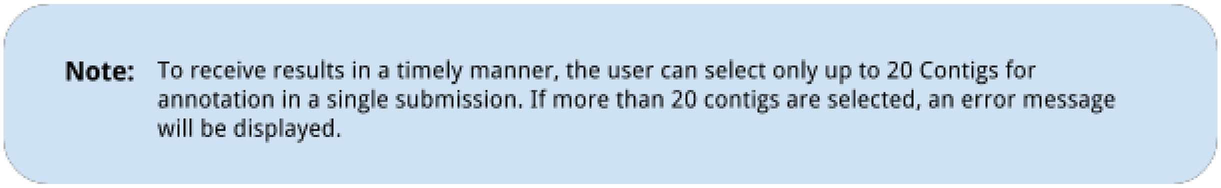
1. The Contig must be associated with at least one Cosmid clone through use of the Contig Retrieval Tool.
2. The Contig must not have been previously annotated. That is, the tool can only be executed once per Contig.

Once the user clicks **Run**, they will be redirected to an authentication message stating which Contigs are queued for processing and the e-mail address that the notifications will be sent to.



Note: An email must be entered so that the user can be notified of what contigs are being annotated and which annotations were successful.

Shortly after running the tool, the user will receive an email indicating the selected Contigs that are being processed. The user will receive another e-mail within 12 hours indicating the Contigs that have successfully been annotated or will display a message stating the job was unsuccessful. As a point of reference, every 5 Contigs will take approximately 2 hours to be completed. Upon successful completion of the annotation process, all ORF information for the Contig will be uploaded to the database, and images displaying detailed Contig information will be generated.



Note: To receive results in a timely manner, the user can select only up to 20 Contigs for annotation in a single submission. If more than 20 contigs are selected, an error message will be displayed.

4.2.2 Viewing the Data

All predicted ORF information and data retrieved about the Contig can be viewed under the Cosmid or Contig page for the Contig of interest; Five images will be available to the user in order to view a detailed display of all information processed in the ORF finding and annotation tool (**Figure 4-7**).

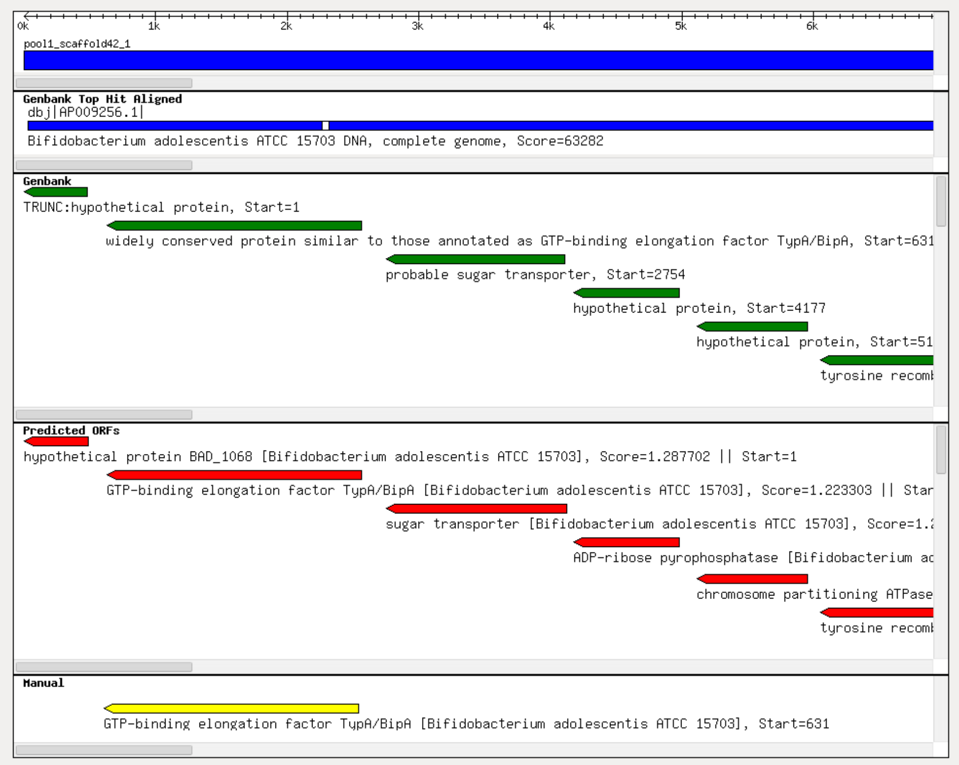
1) ***Contig*** - Displays a single track containing the Contig sequence and a scale. The Contig name is displayed above the bar.

2) ***Genbank Top Hit Aligned*** **-** Based on the results of running a blast search of the Contig on the nt/nr database, an image is generated displaying all the HSPs for that hit. The accession ID for the top hit is displayed above the bar, while the organism is displayed below the bar. HSPs with a higher alignment score are shaded in a darker blue than HSPs that have a lower alignment score. Additionally, the displayed hit is aligned to the Contig sequence displayed in the first image.

3) ***Genbank*** **-** The genbank image displays all CDS that are found within the top-hit sequence that has been aligned to the Contig sequence. For instance, given the alignment of the top HSP for the Contig to the top hit on the genome retrieved from the nt database, the range of the predicted-genbank CDS is extrapolated to predict the start and stop site of the Contig sequence. The annotation for the CDS, as well as the start position is displayed below the bar with all CDSs being aligned to the Contig sequence.

4) ***Predicted ORFs*** **-** Displays all annotated-ORFs that were predicted using FragGeneScan. Below the bar, the annotation is displayed, followed by the FragGeneScan prediction score, and the start of the ORF. All the ORFs are aligned to the Contig sequence. The ORF descriptions are retrieved from BLAST against the RefSeq Protein database.

5) ***Manual*** **-** Displays all annotated-ORFs that were manually inputted by the user. The annotation for the ORF and the start site is displayed below the bar. All ORFs are aligned to the Contig sequence.



Genbank Top Hit Aligned

dbj|AP009256.1|

Bifidobacterium adolescentis ATCC 15703 DNA, complete genome, Score=63282

Genbank

TRUNC:hypothetical protein, Start=1

widely conserved protein similar to those annotated as GTP-binding elongation factor TypA/BipA, Start=631

probable sugar transporter, Start=2754

hypothetical protein, Start=4177

hypothetical protein, Start=51

tyrosine recomb

Predicted ORFs

hypothetical protein BAD_1068 [Bifidobacterium adolescentis ATCC 15703], Score=1.287702 || Start=1

GTP-binding elongation factor TypA/BipA [Bifidobacterium adolescentis ATCC 15703], Score=1.223303 || Star

sugar transporter [Bifidobacterium adolescentis ATCC 15703], Score=1.2

ADP-ribose pyrophosphatase [Bifidobacterium ad

chromosome partitioning ATPase

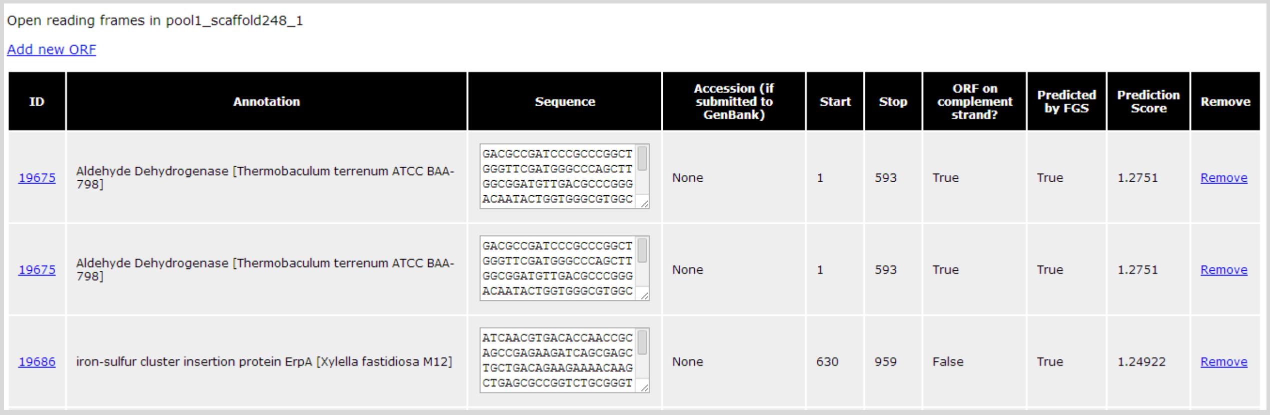
tyrosine recomb

Manual

GTP-binding elongation factor TypA/BipA [Bifidobacterium adolescentis ATCC 15703], Start=631

Figure 4-7: View of the Contig browser which displays: Contig, genbank top-hit alignments, Genbank-predicted CDSs, FragGeneScan annotated predicted-ORFs, user manually inputted annotated ORFs.

A detailed table located below the images is provided to allow for viewing detailed information regarding the FragGeneScan predicted ORFs and user-added ORFs (**Figure 4-8**).



Open reading frames in pool1_scaffold248_1

Add new ORF

ID	Annotation	Sequence	Accession (if submitted to GenBank)	Start	Stop	ORF on complement strand?	Predicted by FGS	Prediction Score	Remove
19675	Aldehyde Dehydrogenase [Thermobaculum terrenum ATCC BAA-798]	GACGCCGATCCCGCCCGGCT GGGTTCGATGGGCCCAGCTT GGCGGATGTTGACGCCCGGG ACAATACTGGTGGGCGTGGC	None	1	593	True	True	1.2751	Remove
19675	Aldehyde Dehydrogenase [Thermobaculum terrenum ATCC BAA-798]	GACGCCGATCCCGCCCGGCT GGGTTCGATGGGCCCAGCTT GGCGGATGTTGACGCCCGGG ACAATACTGGTGGGCGTGGC	None	1	593	True	True	1.2751	Remove
19686	iron-sulfur cluster insertion protein ErpA [Xylella fastidiosa M12]	ATCAACGTGACACCAACCGC AGCCGAGAAGATCAGCGAGC TGCTGACAGAAGAAAACAAG CTGAGCGCCGGTCTGCGGGT	None	630	959	False	True	1.24922	Remove

Figure 4-8: Detailed view of FragGeneScan predicted and manually inputted ORFs.

4.2.3 Re-generation of image upon adding, editing, or removing ORFs from a Contig

Editing the ORFs on a specific Contig will initiate auto-updating of the images generated by the tool. Users can do this on either the Contig or Cosmid detail page that has an ORF table (**Figure 4-8**).

Clicking **Add new ORF** will redirect the user to the Manually Add Open Reading Frame page, while clicking **Remove** will deleted the ORF from the database. Additionally, to edit any existing ORF, the user must click on the unique ORF ID number which will redirect the user to the unique ORF page. By clicking **Edit**, the ORF annotation can be manually changed and the updated version can be submitted to the database. By instantiating either of the three processes, the images will be regenerated with the appropriate data and the table will be updated.

5 Maintenance and troubleshooting

5.1 Server specifications and details

This website, database and all related scripts are hosted by the Science Computing group at the University of Waterloo on a Debian (v. 7 'wheezy') Linux server. All of the scientific data and information is stored in a MySQL database. The website itself uses the Django framework (v. 1.6) with Python (v 2.7). Additional Django packages used were the django-reversion and django-extensions. The website pages are written to meet both HTML5 and CSS3 standards. The webserver itself is Apache (v 2.2.22). The Contig retrieval tool and gene annotation tool are written in Perl (5.14.2) and use the BioPerl API for several functions. Specifically the BioGraphics (v. 1.6.923) module is used to generate the annotated images of the Contigs. The command line version of BLAST (2.2.29+) is used in both the annotation tool and the Contig retrieval tool, as well as on the website interface for aligning queries to sequences locally stored within the database. The software used for ORF prediction is FragGeneScan (v. 1.16). The computing resources needed for the tools are extremely intensive and in order to minimize the amount of time these take the SHARCNET cluster of computers is utilized for these tasks.

5.2 Common errors and troubleshooting

5.2.1 Contig retrieval tool

If no Cosmids are selected and the **Submit** button is pressed, an error message will display. If a particular Cosmid is not present within the desired pool, the user must manually add the Cosmid to the pool through the Add Cosmid page. Selected Cosmids may be missing from the validation page for one of two reasons: either the Cosmid is missing the forward or reverse end-tag, or the ratio of the HSP length to the end-tag sequence length is below the significance threshold of 65%. Refer to **Section 6.3** for a detailed overview of the Contig-Retrieval process.

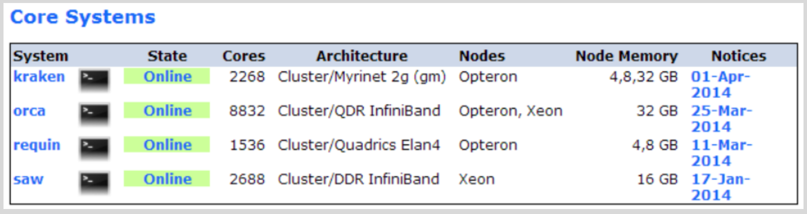
5.2.2 ORF prediction and annotation tool

This tool will only annotate 20 Contigs per run, if more are selected an error message will display. Only Contigs that have been associated with a Cosmid through the Contig retrieval tool are available for annotating. If a Contig has already been annotated with this tool it cannot be run through again until all ORFs are removed from the associated Cosmid and Contig.

It is possible to receive an email saying that not all of the Contigs you submitted were annotated. There are two reasons this could happen:

- The SHARCNET server is down or the status is “conditional”
- You discovered a new bug in the software that does the analysis

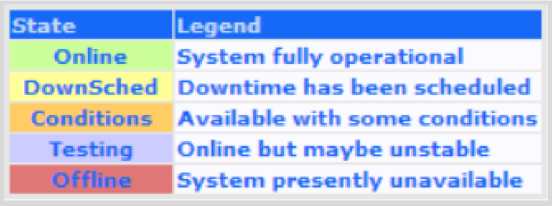
You can check the status of the SHARCNET server at https://www.sharcnet.ca/my/systems. This tool currently uses the ‘saw’ server. At the top of the page under ‘Core Systems’, you can view the status of the server (**Figure 5-1)**. The legend for this status is at the very bottom of the page. If the status is anything other than ‘Online’, it is likely this is the source of the problem. In this case it is best to check back to this website for updates as to when it will be back online.



Core Systems

System		State	Cores	Architecture	Nodes	Node Memory	Notices
kraken	>_	Online	2268	Cluster/Myrinet 2g (gm)	Opteron	4,8,32 GB	01-Apr-2014
orca	>_	Online	8832	Cluster/QDR InfiniBand	Opteron, Xeon	32 GB	25-Mar-2014
requin	>_	Online	1536	Cluster/Quadrics Elan4	Opteron	4,8 GB	11-Mar-2014
saw	>_	Online	2688	Cluster/DDR InfiniBand	Xeon	16 GB	17-Jan-2014

Figure 5-1: The core systems on SHARCNET and their statuses. The only cluster used by this website is saw. This can be found at https://www.sharcnet.ca/my/systems



State	Legend
Online	System fully operational
DownSched	Downtime has been scheduled
Conditions	Available with some conditions
Testing	Online but maybe unstable
Offline	System presently unavailable

Figure: The legend used to indicate the status of the SHARCNET servers.

If the saw server is shown to be online then the source of the issue is likely within the actual website or tool itself. It is best to simply rerun the tool on the Contigs that were not properly annotated. If the issue persists with a particular set of data, it is possible there is something specific regarding your data that has not been accounted for in the tool. At this point it is best to record all the steps you took and what information you are attempting to use and contact a website administrator.

5.2.3 Errors resulting from adding information to the database

Data will not be added unless it matches the specified criteria required. For example, a Cosmid name has a maximum length of 50 characters. For a full list of each specified requirements for all values, see **Appendix 1**.

5.2.4 404 error

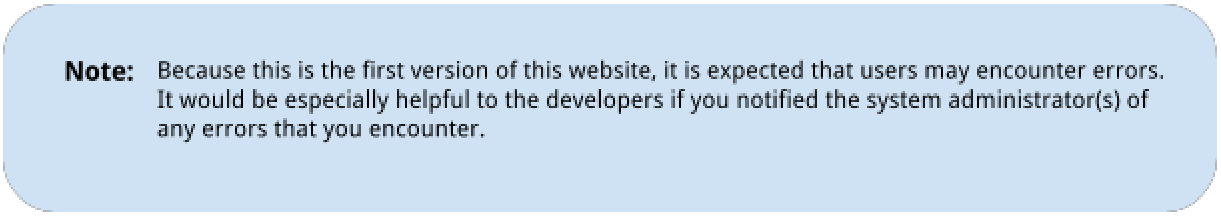
You went to a page that doesn't exist. If you think this page should actually exist contact a website administrator and make sure to record the exact URL that returns the error.

5.2.5 403 error

You attempted to access a page that you do not have permission to access. If you think you should be able to access this page then contact a website administrator. If not, well then, just carry on.

5.2.6 500 error

Something internal within the website has gone wrong. It is best to record the steps you took with as much detail as possible and report this to a website administrator.



Note: Because this is the first version of this website, it is expected that users may encounter errors. It would be especially helpful to the developers if you notified the system administrator(s) of any errors that you encounter.

6 Technical details

This section provides in-depth technical information for the more advanced user and possibly for future developers of this website. Understanding of this section is not required for interaction with data on the website or operation of the custom tools.

6.1 Database

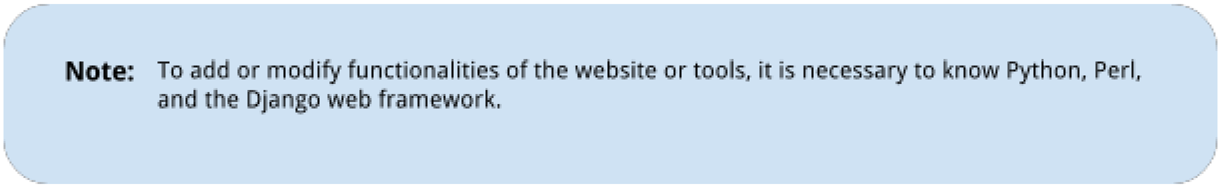
6.1.1 Database entity-relationship diagram

All data accessed through the web interface is stored in a MySQL database. Relationships between tables in the database are complex; the entity relationship diagram (ERD) is provided **(Figure 6-1**).

6.1.2 Website—database communication through Django

This website is built on a Django web framework, which conveniently using object-oriented Python to manipulate records in the database. In the framework, each MySQL table is a Python class, and each record in the MySQL table is an object of the respective class.

Object-oriented control of the database is extremely powerful, and many of the complex functions of the website make use of this power. To understand data manipulation on the server-side or to modify functionality of the website, it is necessary to know Python and understand the DJango framework. It is not necessary to know SQL, although knowledge of SQL may be helpful.



Note: To add or modify functionalities of the website or tools, it is necessary to know Python, Perl, and the Django web framework.

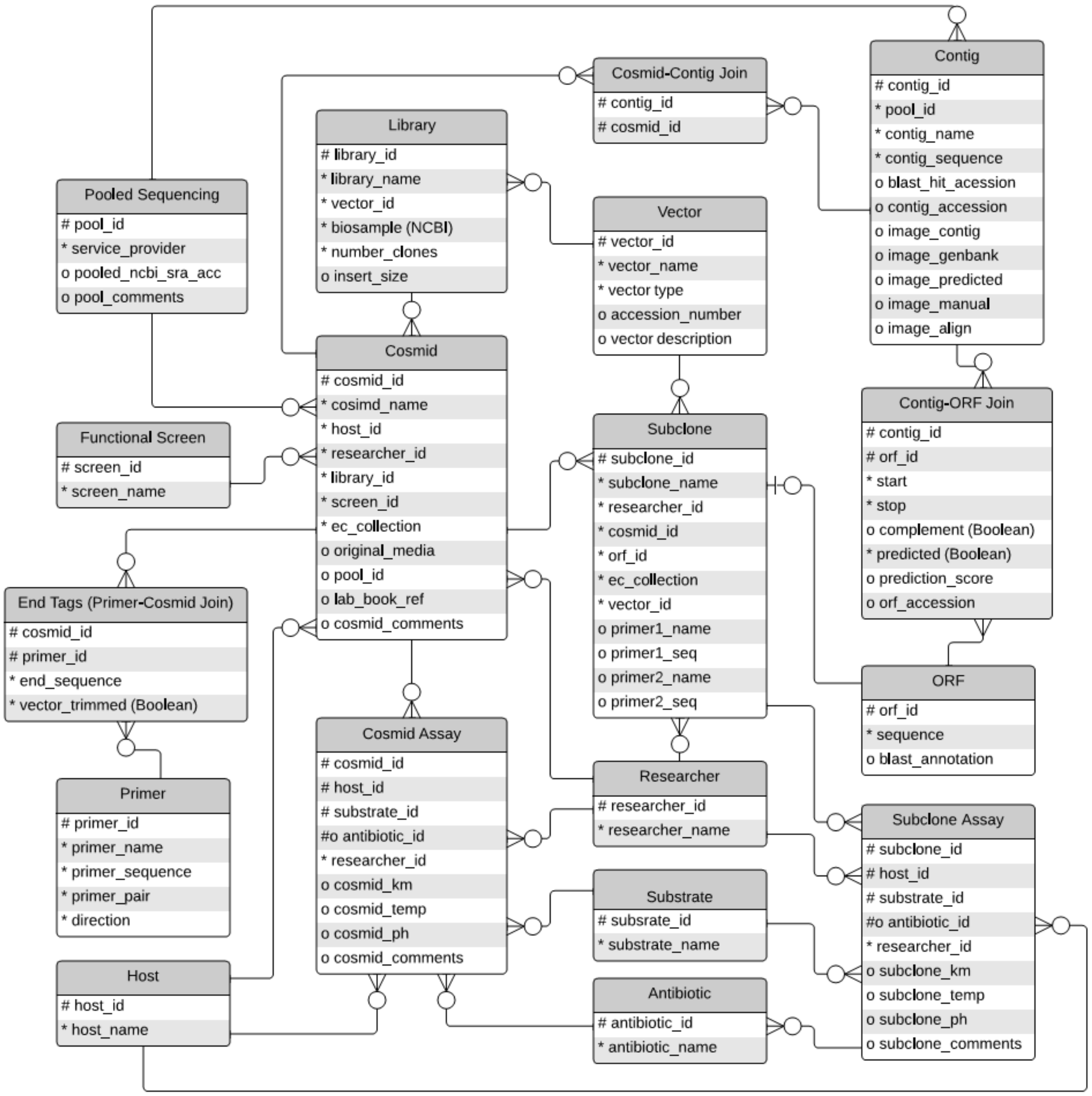


Figure 6-1: Entity relationship diagram for MySQL database on Metagenomics server

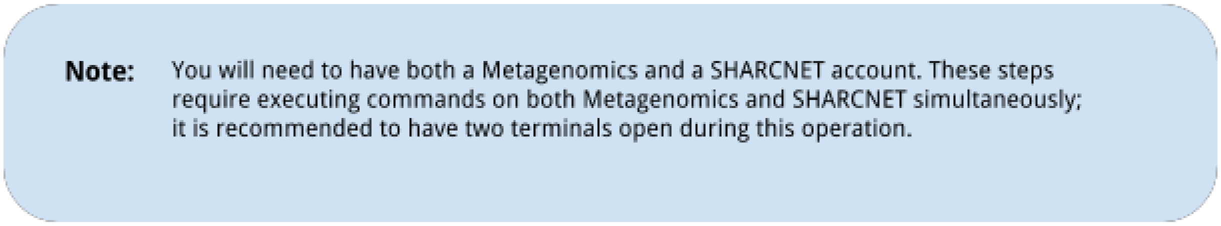
6.2 Communication between the Metagenomics and SHARCNET servers

The website and set of tools described in this document are actually stored and executed on two physically separate servers. The main server is Metagenomics, which contains the database and the website. The other server is SHARCNET which is actually a cluster of computers shared by many institutions; this server is used for the more computationally intensive processing of the data that is done in the ORF prediction and annotation tool.

In order for the tool to execute seamlessly for a user on the website, a secure connection must be set between the two servers to allow for data to be transferred. Once established, the connection settings will not need to be re-established unless there are changes made to the files or directories that are set up during this process (e.g., if the Metagenomics server is reconfigured).

6.2.1 Setting up a connection

To establish a working pipeline between the Metagenomics server and SHARCNET, the user will first require an account on Metagenomics, as well as a Level-1 certified account on SHARCNET; the former can be requested from the system administrator at Waterloo Science Computing and the latter is obtained by application to SHARCNET. The following pages outline steps that are required to set up a connection:



Note: You will need to have both a Metagenomics and a SHARCNET account. These steps require executing commands on both Metagenomics and SHARCNET simultaneously; it is recommended to have two terminals open during this operation.

1) **Install a bilateral Public Key Authentication between Metagenomics and SHARCNET**.

- ○ Log in to the Metagenomics server.
- ○ Create a .ssh folder in your home directory:

```
$ cd ~
$ mkdir .ssh
$ chmod 700 .ssh
```

- ○ Generate your own personal set of keys on Metagenomics:

```
$ cd .ssh
$ ssh-keygen -b 1024 -t dsa
```

- ○ Log in to the SHARCNET server
- ○ Create a .ssh folder in your home directory:

```
$ cd ~
$ mkdir .ssh
$ chmod 700 .ssh
```

- ○ On the Metagenomics server, secure copy over the “id_dsa.pub” key file to Sharcnet:

```
$ scp id_dsa.pub username@saw.sharcnet.ca:/path/to/home/.ssh/
```

- ○ On the SHARCNET server, add the contents of “id_dsa.pub” to a file named “authorized_keys”:

```
$ cd /home/username/.ssh
$ cat id_dsa.pub >> authorized_keys
$ rm id_dsa.pub
```

- Generate your personal set of keys on SHARCNET, and secure copy the key to Metagenomics:

```
$ ssh-keygen -b 1024 -t rsa
$ scp -P 8022 id_rsa.pub
  username@metagenomics.uwaterloo.ca:/home/.ssh
```

- On the Metagenomics server, add the contents of “id_dsa.pub” to a file name “authorized_keys”:

```
$ cd /home/username/.ssh
$ cat id_rsa.pub >> authorized_keys
$ rm id_rsa.pub
```

2) **Transfer the SHARCNET pipeline file to your SHARCNET scratch directory**.

- On the Metagenomics server, locate the “sharcnet_pipe.tar” and secure copy it to SHARCNET:

```
        $ cd /home/trevor/sharcnet_files
        $ scp sharcnet_pipe.tar
username@saw.sharcnet.ca:/scratch/username/
```

- On the SHARCNET server, untar the transferred “sharcnet_pipe.tar” file:

```
$ cd /scratch/username
$ tar -xvf sharcnet_pipe.tar
$ rm sharcnet_pipe.tar
```

- Check that the pipeline was extracted:

```
$ ls metagenomics/sharc_mg_pipe.pl
```

3) Set up the configuration files for SHARCNET

- On the Metagenomics server, locate the “sharcnet_config.tar” and secure copy it to SHARCNET:

```
$ cd /home/trevor/sharcnet_files
$ scp sharcnet_config.tar
  username@saw.sharcnet.ca:/home/username/
```

- On the SHARCNET server, untar the transferred “sharcnet_config.tar” file:

```
$ cd ~
$ tar -xvf sharcnet_config.tar
```

- Check that the “.bashrc” file was created:

```
$ ls -a .bashrc
```

6.3 Contig retrieval tool

6.3.1 Generation of user-specified database information

The pool and the Cosmids selected on the tool page are sent to the Contig_pipeline function. All Contigs associated with the selected pool are taken from the database and written to a fasta file under 'Contig_retrieval_tool/Contigs.fa'. All end-tags associated with the selected Cosmids are taken from the database and written to two separate CSV files: one for the forward end-tags under ‘Contig_retrieval_tool/primers_1.CSV’, and the other for reverse end-tags under ‘Contig_retrieval_tool/primers_2.CSV’. These files are used to execute the Contig retrieval pipeline.

6.3.2 Execution of the Contig retrieval tool

The Contig retrieval tool is implemented in perl and runs on a linux-based environment. The command, *perl retrieval_pipeline.pl [end_tags_1.CSV] [end_tags_2.CSV] [database_file.fa]*, is used to initiate the pipeline. Previous work in the working directory will be detected and removed, and the necessary work environment will be recreated. Once the environment is set up, both end-tag files are parsed with the relevant matching-pair information being stored in an internal data structure. The local database is formatted from the specified *[database_file.fa]* using the NCBI makeblastdb tool. Each end-tag sequence which has a matching pair is aligned to sequences from the locally created database using blastn. The following information is retrieved and stored to the data structure:

- Cosmid name
- Which end-tag is being used
- Contig hit name
- Algorithm
- E-Value
- Percent identity
- Number of identical residues
- Query length
- Top HSP length
- Range of top HSP alignment on the query sequence
- Range of top HSP alignment on the hit sequence
- Contig sequence

Using the previously retrieved information, the tool will then determine whether the Contig retrieved from the local database BLAST is of any significance. It will filter out spurious elements that aligned and will report whether both end-tags for a given Cosmid retrieve the same Contig, whether they retrieve different Contigs, or whether there was a significant ambiguity that warrants further investigation. The following conditions are used to match the end-tag sequences for any given Cosmid to the Contigs:

- If both end-tags align to the same Contig sequence.
- Determines if the ratio of the HSP length to the query sequence length is above a preset significant level (Default = 65%).
- Identifies ambiguities if one query sequence aligns significantly to two or more Contigs.

6.3.3 Processing the returned data

Once the pipeline has finished executing, a CSV file will output to ‘Contig_retrieval_tool/tmp/out/retrieval.CSV’. This file contains the results of the Contig retrieval tool, with each row in the file representing a new Cosmid-Contig association. The CSV content will be displayed for user validation and the selected Contigs are then stored in the database. The user is then redirected to a Cosmid table where the new Contigs are displayed.

6.4 ORF finding and annotation tool

6.4.1 ORF finding and annotation tool

The ORF finding and annotation tool is implemented in Perl and runs on a Linux environment, specifically the SHARCNET server.

6.4.2 Complete annotations from scratch

6.4.2.1 Generation of user-specified request

All Contigs that have been associated with a Cosmid through the Contig Retrieval Tool and Contigs that have not yet been annotated will be displayed. Once the Contigs have been selected, an email address is entered and the **Run** button is pressed, a “data.lib” file is created for the execution of the perl script in the annotation pipeline. The “data.lib” file takes all Contig and ORF information from the Contigs that were selected and stores it in a perl data structure.

The user is then emailed through the Linux mail command to notify them of their selection. The annotation pipeline is executed in the background through the command: “tsp perl annotation_tool/annotation_pipeline.pl -annotate &”. As well, a daemon process is run in the background with the command: “python manage.py runscript annotation_processor %s & %email”. The daemon process checks every five minutes for the annotations output, which can be found under 'annotation_tool/tool/out/annotations.CSV'. This file contains a new annotation in each row, these rows are read and checked against the database to confirm they are novel annotations for the Contig and then stored into the database’s Contig_ORF_join table. The annotation tool also outputs images of its results under "annotation_tool/tool/img/", these images are stored in the database’s Contig table. Each successful Contig will generate five images to display all annotation results.

Once everything has been stored into the database, the user receives another email stating which Contigs were successful or that the process was unsuccessful.

6.4.2.2 Complete annotations from scratch

A complete annotation pipeline is initialized using the *-annotate* argument. The *data.lib* file is analyzed to retrieve the Contigs and sequences that require annotation. The Contig sequences are passed into the ORF predictor tool, FragGeneScan, where three files are generated per Contig:

- an *.faa* file containing the amino acid ORF sequences
- an *.fna* file containing the DNA ORF sequences
- an *.out* file containing a summary of each ORF

Files are copied to the 'saw' cluster of SHARCNET, an academic consortium that shares a network of high performance computers, where all complete Contig sequences and ORF sequences are aligned to either the NCBI nucleotide (nr/nt) database, or the Refseq protein (refseq_protein) database respectively. Jobs on SHARCNET are run in groups of 5, with each BLAST threaded across 4 cores. The estimated amount of time to complete the Contig sequence alignment is approximately 25 minutes, while the estimated amount of time to complete alignments of all predicted ORFs to their respective Contigs is approximately 1.5 hours.

Once all the jobs have been successfully executed, the newly created .blast files are packaged and transferred back to the metagenomics server. Genbank and ORF prediction annotations are handled separately:

> **Genbank Annotations**: The blast files for the Contig sequence queries are parsed for the accession number and the range of the HSP alignment(s) for the top hits. The genbank files associated with each of the top hits for the Contig sequence alignments are retrieved and further parsed for all CDS within the range of the alignment.
>
> **ORF Annotations**: The blast files for each of the Contigs contain all the alignment information for every ORF associated with that Contig. Each .blast file is parsed for the top hit for every result. The annotations associated with that top hit is then associated with its respective ORF.

Once all annotations have been acquired and associated with either the Genbank CDS or the predicted ORFs, images are generated using the BioGraphics module of Perl.

Five images of .png format are created for every Contig:

- ○ Contig sequence including a scale to identify base pair location.
- ○ Alignment of the HSP(s) for the top hit using the entire Contig sequence as a query.
- ○ All Genbank CDSs and annotations that align themselves between the range of the Contig sequence.
- ○ All predicted ORFs and annotations associated with the Contig.
- ○ All manually inserted ORFs and annotations.

6.4.3 Updated user annotations

6.4.3.1 Generation of user-specified request

If an ORF is added, edited or deleted the images will be updated. Upon editing, the “data.lib” file is created to represent the new Contig-ORF information within a perl data structure, it will run the ORF_data_update function. This file is used in the annotation pipeline through the command: "perl annotation_tool/annotation_pipeline.pl -update". The script will output images of the updated annotations and store these images into the database.

6.4.3.2 Updated user annotations

An update annotation pipeline is initialized using the *-annotate* argument. The *data.lib* file is analyzed to retrieve the Contigs and their associated ORF with annotations, including manually inserted ORFs with annotations from the end-user. The accession number for each Contig is retrieved and used to retrieve the newest version of the Contigs associated Genbank file. The Contig is then aligned against the sequence retrieved from the Genbank file. All CDS that are found with the top alignment are retrieved and stored for image generation.

Once all updated CDSs have been acquired, images are generated using the BioGraphics module of Perl. Five images of .png format are created for every Contig:

- ○ Contig sequence including a scale to identify base pair location.
- ○ Alignment of the HSP(s) for the top hit using the entire Contig sequence as a query.
- ○ All Genbank CDSs and annotations that align themselves between the range of the Contig sequence.
- ○ All predicted ORFs and annotations associated with the Contig.
- ○ All manually inserted ORFs and annotations.

6.4.4 Processing the returned data

If the edit has been submitted the user will be redirected to a page stating which Contig has been successfully edited. If a new ORF is added to a Contig, the user will be redirected to the Contig detail page to view the new ORF under the manual image. After pressing the **Submit** button the Contig or ORF table will be updated and the annotation update pipeline will run. The pipeline will return updated images to store in the database for the edited Contig.

7 References

1. States DJ, Gish W, and Altschul SF (1991) METHODS: A companion to Methods in Enzymology 3:66-70.

Appendix 1: Terminology and definitions

HSP	high-scoring pair
SHARCNET	shared hierarchical academic research computing network
BLAST	basic local alignment search tool
ORF	open reading frame
CDS	coding DNA sequence
CSV	comma separated values
ERD	entity relationship diagram
PI	principal investigator
HTML	hypertext markup language
CSS	cascading style sheets
SQL	structured query language

Appendix 2: Requirements for data entry

Table 1. All of the requirements for any data entry into the database.

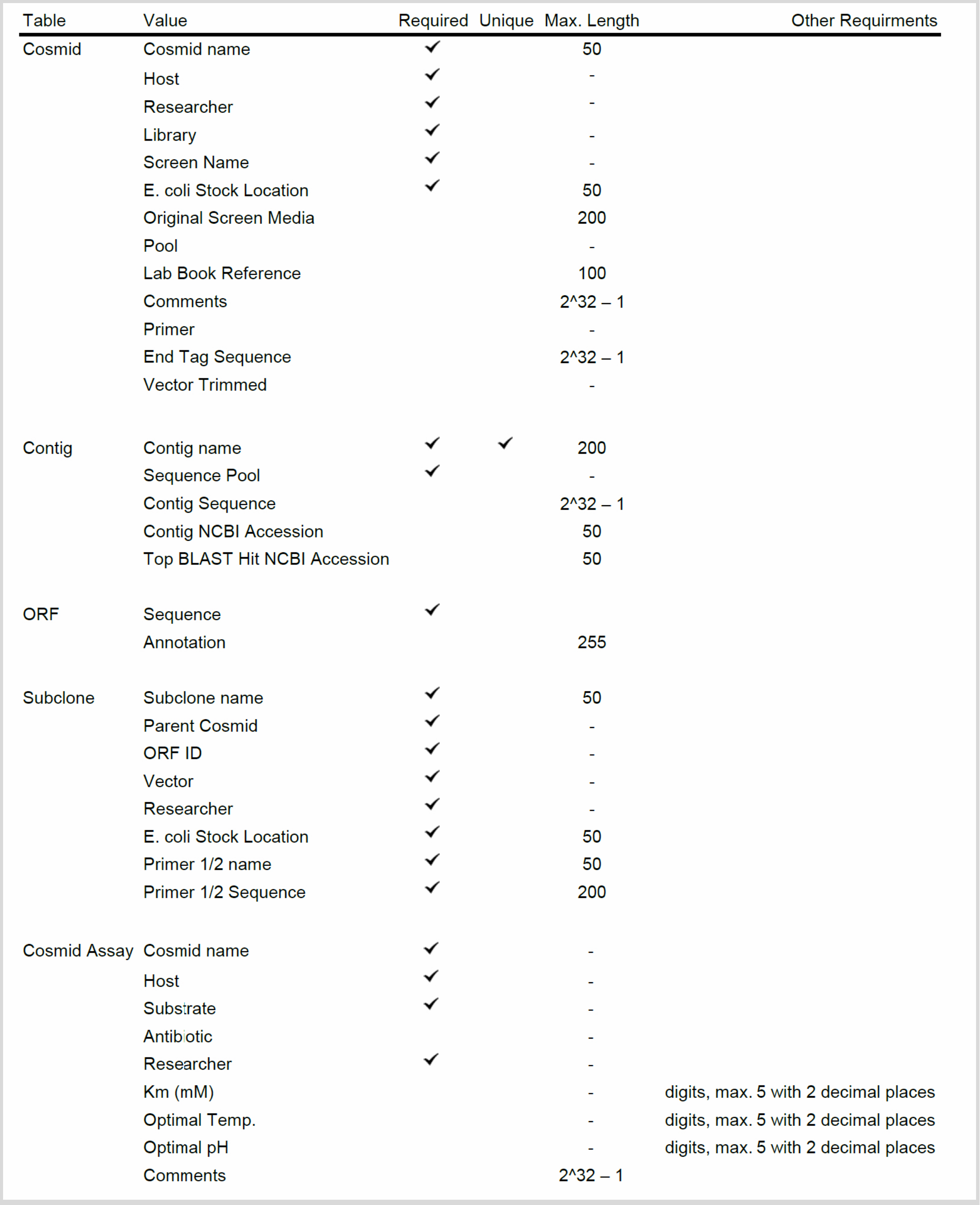


Table	Value	Required	Unique	Max. Length	Other Requirments
Cosmid	Cosmid name	✓		50	
	Host	✓		-	
	Researcher	✓		-	
	Library	✓		-	
	Screen Name	✓		-	
	E. coli Stock Location	✓		50	
	Original Screen Media			200	
	Pool			-	
	Lab Book Reference			100	
	Comments			2^32 – 1	
	Primer			-	
	End Tag Sequence			2^32 – 1	
	Vector Trimmed			-	
Contig	Contig name	✓	✓	200	
	Sequence Pool	✓		-	
	Contig Sequence			2^32 – 1	
	Contig NCBI Accession			50	
	Top BLAST Hit NCBI Accession			50	
ORF	Sequence	✓			
	Annotation			255	
Subclone	Subclone name	✓		50	
	Parent Cosmid	✓		-	
	ORF ID	✓		-	
	Vector	✓		-	
	Researcher	✓		-	
	E. coli Stock Location	✓		50	
	Primer 1/2 name	✓		50	
	Primer 1/2 Sequence	✓		200	
Cosmid Assay	Cosmid name	✓		-	
	Host	✓		-	
	Substrate	✓		-	
	Antibiotic			-	
	Researcher	✓		-	
	Km (mM)			-	digits, max. 5 with 2 decimal places
	Optimal Temp.			-	digits, max. 5 with 2 decimal places
	Optimal pH			-	digits, max. 5 with 2 decimal places
	Comments			2^32 – 1	

Subclone Assay	Subclone name	✓		-	
	Host	✓		-	
	Substrate	✓		-	
	Antibiotic			-	
	Researcher	✓		-	
	Km (mM)			-	digits, max. 5 with 2 decimal places
	Optimal Temp.			-	digits, max. 5 with 2 decimal places
	Optimal pH			-	digits, max. 5 with 2 decimal places
	Comments			2^32 – 1	
Primer	Primer name	✓	✓	50	
	Primer Pair ID	✓		-	integer -2147483648 to 2147483647
	Direction	✓		1	F' or 'R'
	Sequence	✓		200	
Vector	Name	✓	✓	50	
	Type	✓		50	
	NCBI Accession			50	
	Description			255	
Library	Name	✓	✓	100	
	NCBI BioSample ID	✓	✓	50	
	Vector	✓		-	
	Est. Number of Unique Clones	✓		-	integer -2147483648 to 2147483647
	Est Insert Size			-	integer -2147483648 to 2147483647
Sequencing Pool	Service Provider Name	✓		200	
	NCBI SRA Accession			100	
	Maximum number of clones	✓		-	integer -2147483648 to 2147483647
	comments			2^32 – 1	
Substrate	Substrate name	✓	✓	100	
Antibiotic	Antibiotic name	✓	✓	100	
Researcher	Name	✓	✓	100	
Host	Host Name	✓	✓	150	
Screen	Screen Name	✓	✓	100	