MicrobeDB Overview

Morgan Langille
morgan.gi.langille@gmail.com

Main Features

- Centralized storage and access to completed archaeal and bacterial genomes
  - Genomes obtained from NCBI RefSeq: http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
  - Genome/Flat files are stored in one central location
    - Including files .gbk, .gff, .fna, .faa, etc.
  - Unpublished genomes can be added as well
  -
- Information at the genome project, chromosome, and gene level are parsed and stored in a MySQL database
- A Perl MicrobeDB API provides non-MySQL interface with the database.

Main MicrobeDB Tables

- **Version**
  - Each download of genomes from NCBI is given a new version numbe
  - Data will not change if you always use the same version number of microbedb
  - Version date can be cited for any method publications
  - A version can be saved by users so not automatically deleted.
- **Genome Project**
  - Contains information about the genome project and the organism that was sequenced
  - Each genome project contains one or more replicons
- **Replicon**
  - Chromosome, plasmids, or contigs
  - Each replicon contains one or more genes
- **Gene**
  - Contains gene annotations and also the DNA and protein sequences (if protein coding gene)

MicrobeDB Annotations

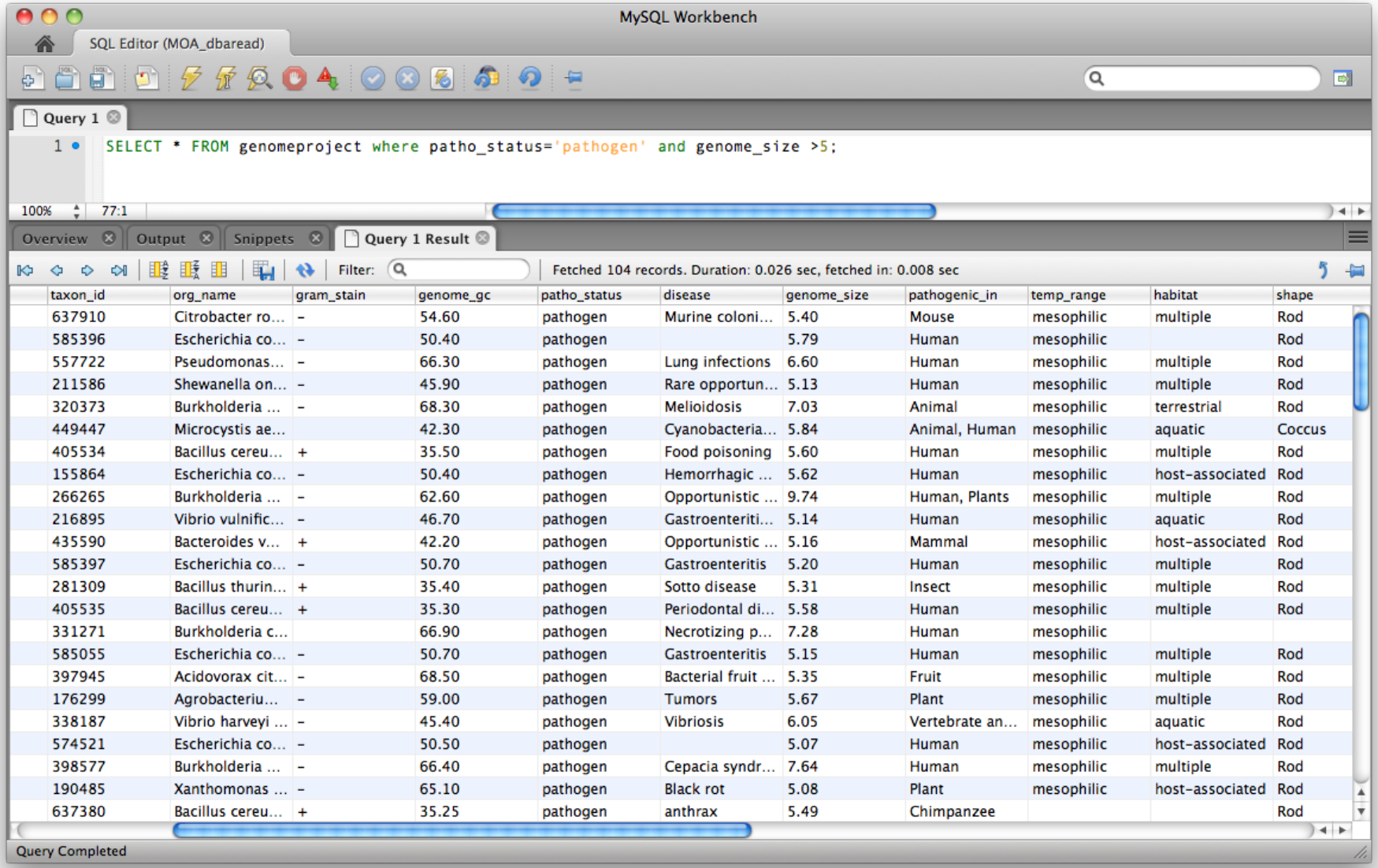
Table/Object*	Field Descriptions*	Example
Genome Project	Organism Name	Pseudomonas aeruginosa LESB58
	NCBI Taxon ID	557722
	Genome Size (Mb)	6.6
	Pathogenic In	Human
	GC %	66.3
	Oxygen Requirements	aerobic
	Sequencing Centre	Wellcome Trust Sanger Institute
Replicon	Replicon Type	Chromosome
	Accession (RefSeq)	NC_011770
	Replicon Size (bp)	6601757
	Number of Genes	6027
	Replicon Sequence	TTTAAAGAG…
Gene	Gene Type	CDS
	Locus ID	PLES_00001
	Start Position	483
	End Position	2027
	Gene Name	dnaA
	Product	chromosomal replication initiation
	DNA Sequence	GTGTCCGT…
	Protein Sequence	MSVELWQQ...
Version	Download Date	2011-12-17
	Flat File Directory	/share/genomes/2011-12-17/
	Used By	Morgan, Matthew

*Not all fields and tables in MicrobeDB are listed.

Accessing MicrobeDB

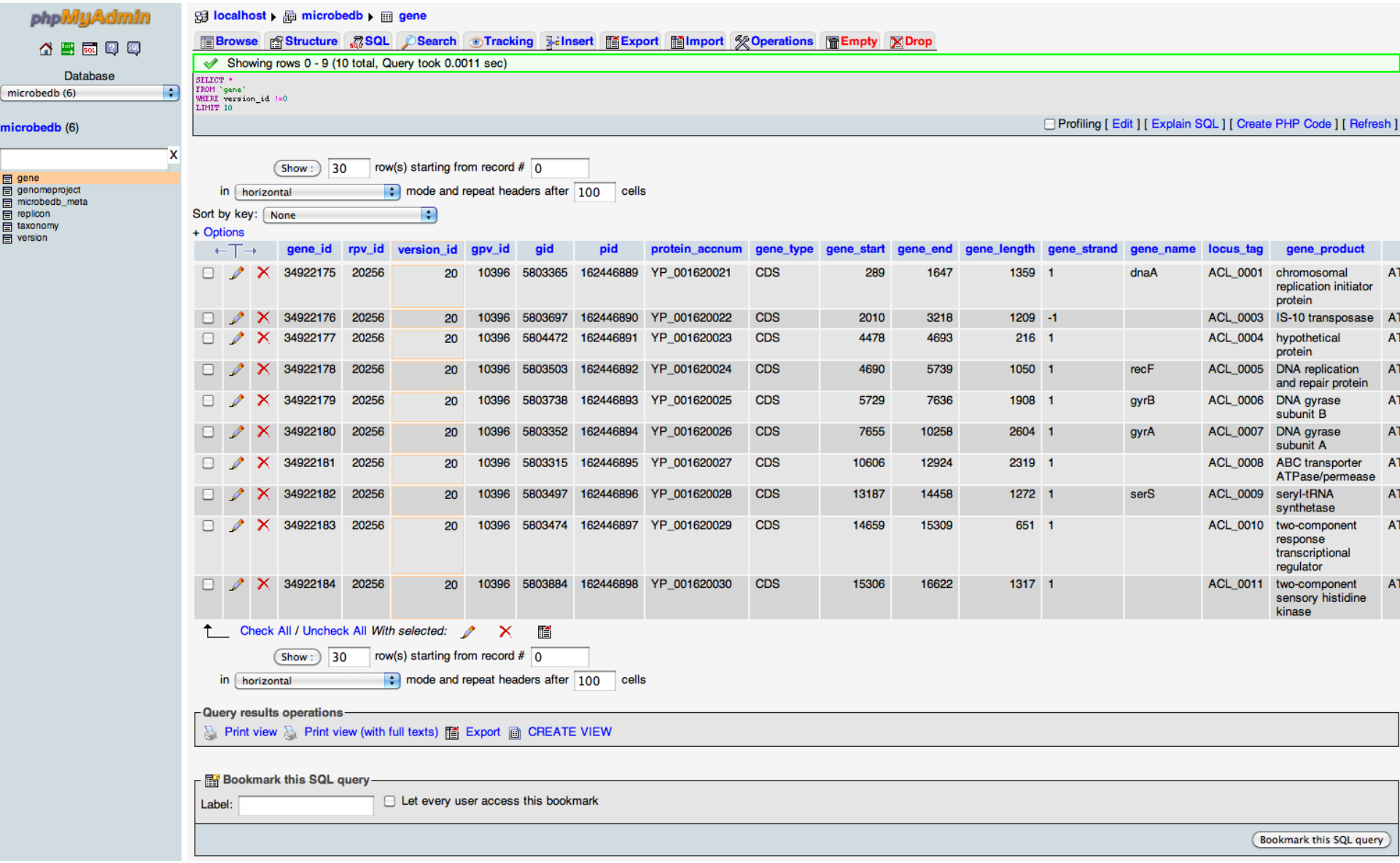
- Any traditional MySQL programs
  - phpMyAdmin:
    - Web-based
    - http://phpmyadmin.net
  - MySQL Workbench
    - Local desktop client
    - http://www.mysql.com/products/workbench/
- MicrobeDB Perl API
    - Allows interaction with database directly from within a Perl script
    - Requires no knowledge of SQL

MySQL Workbench



taxon_id	org_name	gram_stain	genome_gc	patho_status	disease	genome_size	pathogenic_in	temp_range	habitat	shape
637910	Citrobacter ro...	-	54.60	pathogen	Murine coloni...	5.40	Mouse	mesophilic	multiple	Rod
585396	Escherichia co...	-	50.40	pathogen		5.79	Human	mesophilic		Rod
557722	Pseudomonas...	-	66.30	pathogen	Lung infections	6.60	Human	mesophilic	multiple	Rod
211586	Shewanella on...	-	45.90	pathogen	Rare opportun...	5.13	Human	mesophilic	multiple	Rod
320373	Burkholderia ...	-	68.30	pathogen	Melioidosis	7.03	Animal	mesophilic	terrestrial	Rod
449447	Microcystis ae...		42.30	pathogen	Cyanobacteria...	5.84	Animal, Human	mesophilic	aquatic	Coccus
405534	Bacillus cereu...	+	35.50	pathogen	Food poisoning	5.60	Human	mesophilic	multiple	Rod
155864	Escherichia co...	-	50.40	pathogen	Hemorrhagic ...	5.62	Human	mesophilic	host-associated	Rod
266265	Burkholderia ...	-	62.60	pathogen	Opportunistic ...	9.74	Human, Plants	mesophilic	multiple	Rod
216895	Vibrio vulnific...	-	46.70	pathogen	Gastroenteriti...	5.14	Human	mesophilic	aquatic	Rod
435590	Bacteroides v...	+	42.20	pathogen	Opportunistic ...	5.16	Mammal	mesophilic	host-associated	Rod
585397	Escherichia co...	-	50.70	pathogen	Gastroenteritis	5.20	Human	mesophilic	multiple	Rod
281309	Bacillus thurin...	+	35.40	pathogen	Sotto disease	5.31	Insect	mesophilic	multiple	Rod
405535	Bacillus cereu...	+	35.30	pathogen	Periodontal di...	5.58	Human	mesophilic	multiple	Rod
331271	Burkholderia c...		66.90	pathogen	Necrotizing p...	7.28	Human	mesophilic		
585055	Escherichia co...	-	50.70	pathogen	Gastroenteritis	5.15	Human	mesophilic	multiple	Rod
397945	Acidovorax cit...	-	68.50	pathogen	Bacterial fruit ...	5.35	Fruit	mesophilic	multiple	Rod
176299	Agrobacteriu...	-	59.00	pathogen	Tumors	5.67	Plant	mesophilic	multiple	Rod
338187	Vibrio harveyi ...	-	45.40	pathogen	Vibriosis	6.05	Vertebrate an...	mesophilic	aquatic	Rod
574521	Escherichia co...	-	50.50	pathogen		5.07	Human	mesophilic	host-associated	Rod
398577	Burkholderia ...	-	66.40	pathogen	Cepacia syndr...	7.64	Human	mesophilic	multiple	Rod
190485	Xanthomonas ...	-	65.10	pathogen	Black rot	5.08	Plant	mesophilic	host-associated	Rod
637380	Bacillus cereu...	+	35.25	pathogen	anthrax	5.49	Chimpanzee			Rod

phpMyAdmin


phpMyAdmin
Database
microbedb (6)
microbedb (6)
gene
genomeproject
microbedb_meta
replicon
taxonomy
version
localhost ▸ microbedb ▸ gene
Browse Structure SQL Search Tracking Insert Export Import Operations Empty Drop
Showing rows 0 - 9 (10 total, Query took 0.0011 sec)
SELECT *
FROM `gene`
WHERE version_id !=0
LIMIT 10
Profiling [Edit] [Explain SQL] [Create PHP Code] [Refresh]
Show : 30 row(s) starting from record # 0
in horizontal mode and repeat headers after 100 cells
Sort by key: None
+ Options
gene_id rpv_id version_id gpv_id gid pid protein_accnum gene_type gene_start gene_end gene_length gene_strand gene_name locus_tag gene_product
34922175 20256 20 10396 5803365 162446889 YP_001620021 CDS 289 1647 1359 1 dnaA ACL_0001 chromosomal replication initiator protein
34922176 20256 20 10396 5803697 162446890 YP_001620022 CDS 2010 3218 1209 -1 ACL_0003 IS-10 transposase
34922177 20256 20 10396 5804472 162446891 YP_001620023 CDS 4478 4693 216 1 ACL_0004 hypothetical protein
34922178 20256 20 10396 5803503 162446892 YP_001620024 CDS 4690 5739 1050 1 recF ACL_0005 DNA replication and repair protein
34922179 20256 20 10396 5803738 162446893 YP_001620025 CDS 5729 7636 1908 1 gyrB ACL_0006 DNA gyrase subunit B
34922180 20256 20 10396 5803352 162446894 YP_001620026 CDS 7655 10258 2604 1 gyrA ACL_0007 DNA gyrase subunit A
34922181 20256 20 10396 5803315 162446895 YP_001620027 CDS 10606 12924 2319 1 ACL_0008 ABC transporter ATPase/permease
34922182 20256 20 10396 5803497 162446896 YP_001620028 CDS 13187 14458 1272 1 serS ACL_0009 seryl-tRNA synthetase
34922183 20256 20 10396 5803474 162446897 YP_001620029 CDS 14659 15309 651 1 ACL_0010 two-component response transcriptional regulator
34922184 20256 20 10396 5803884 162446898 YP_001620030 CDS 15306 16622 1317 1 ACL_0011 two-component sensory histidine kinase
Check All / Uncheck All With selected:
Show : 30 row(s) starting from record # 0
in horizontal mode and repeat headers after 100 cells
Query results operations
Print view Print view (with full texts) Export CREATE VIEW
Bookmark this SQL query
Label:
Let every user access this bookmark
Bookmark this SQL query

MicrobeDB API Example

```
#Use the MicrobeDB Search library
use MicrobeDB::Search;

#create the search object
my $search_obj= new MicorbeDB::Search();

#Create an object with certain features that we want (i.e. only pathogens)
my $obj = new GenomeProject( version_id => '1', patho_status   => 'pathogen' );

#This does the actual search and returns a list of all genome projects that match search parameters
my @result_objs = $search_obj->object_search($obj);

#Now we can iterate through each genome project
foreach my $gp_obj (@result_objs) {

        #get the name of the genome

        $gp_obj->org_name()

        foreach my $gene_obj ($gp_obj->genes()){

                if($gene_obj->gene_type() eq 'tRNA'){

                        #write the genes in fasta format with gid as the identifier
                        print '>',$gene_obj->gid,"\n",$gene_obj->gene_seq();
                }}}
```