

# USDA - NSA Workshop

---

<http://oystergen.es/workshop>

<http://goo.gl/xqf8Og>

**OYSTERGEN.ES**

WORKSHOP

PACIFIC

OLYMPIA

---

# **USDA NRSP-8 / NSA Genomics Resource Coordination Workshop**

---

*Wednesday, March 25 - Monterey, CA - Monterey Marriot Hotel - 2pm Los Angeles Room.*

---

## **Draft Agenda**

The Workshop will start with a few "how-to" / tutorial based presentations on how groups are dealing with genomic datasets.

.....

## **Livestream**

We will attempt to broadcast and allow persons to join via Google Hangouts. If you are interested in joining the conversation via Google Hangouts please contact Steven Roberts (sr320@uw.edu). Anyone can watch the livestream

## Rationale

There is a wealth of information regarding **genomic attributes and associated phenotypic traits** that has not been exploited from the enormous datasets being generated from high throughput sequencing. The nature of this data often can answer specific research questions of a project and also contain other valuable information that is not readily of interest in those carrying out the experiments. More importantly, the substantial advance in furthering our **understanding of genomic and phenotypic relationship will only occur once disparate datasets are integrated**. This workshop supported by the USDA-NIFA National Research Support Project 8 will focus on bringing together those in the shellfish community to

- a) discuss challenges and solutions in genomic analyses
- b) improve functional annotation of the genome,
- c) produce a sustaining platform for curation, distribution, and application of genomic datasets.

Today

<http://goo.gl/xqf8Og>

The screenshot shows a web browser window displaying a MoPad Etherpad. The top navigation bar includes the text "Etherpad is free software" and the MoPad logo. Below this, the page title is "Public Pad" with "Pad Options" and "Import/Export" links. A rich text editor toolbar is visible, containing icons for bold, italic, underline, strikethrough, bulleted list, numbered list, and indent. The main editing area contains the following text:

1 **USDA NRSP-8 / NSA Genomics Resource Coordination Workshop**  
2  
3 See <http://oystergen.es/workshop> for more information  
4  
5 • Please feel free to you use this document to prior to, during, and after the workshop to  
6 develop agenda, share information and ask questions.  
7 •  
8  
9  
10 **Agenda**  
11 **Tutorials:**  
12 • Steven Roberts: Data management and workflows (GitHub, SQLShare, IPython,  
13 • Alberto Arias-Perez: SNP genotype calling with GATK in Pacific oyster families  
14 • Marta Gomez-Chiarri and Dina Proestou: *Crassostrea virginica* transcriptomic data analysis  
for evaluation of mechanisms of disease resistance  
15  
16  
17  
18 **Lightning Talks (Student Awards)**  
19

On the right side, a sidebar shows a user profile for "Steven F" with a green status indicator and a link to "Invite other users". At the bottom of the sidebar, the date "March 25, 2015" is displayed.



Today

<http://goo.gl/xqf8Og>

Tutorials

Lightning Talks

Discussion

The screenshot shows the MoPad Etherpad interface. At the top, it says "Etherpad is free software" and "MoPad". Below that, it says "Public Pad" and has buttons for "Pad Options" and "Import/Export". The main editing area contains the following text:

1 **USDA NRSP-8 / NSA Genomics Resource Coordination Workshop**  
2  
3 See <http://oyster.gen.es/workshop> for more information  
4  
5 • Please feel free to you use this document to prior to, during, and after the workshop to  
develop agenda, share information and ask questions.  
6 •  
7  
8  
9  
10 **Agenda**  
11 **Tutorials:**  
12 • Steven Roberts: Data management and workflows (GitHub, SQLShare, IPython,  
13 • Alberto Arias-Perez: SNP genotype calling with GATK in Pacific oyster families  
14 • Marta Gomez-Chiarri and Dina Proestou: *Crassostrea virginica* transcriptomic data analysis  
for evaluation of mechanisms of disease resistance  
15  
16  
17  
18 **Lightning Talks (Student Awards)**  
19

On the right side, there is a user profile for "Steven F" with a green status indicator and a button to "Invite other users". At the bottom right, the date "March 25, 2015" is displayed.

Today

---

<http://goo.gl/xqf8Og>

*Crassostrea virginica* sequencing

Tutorials

Breeding program

Lightning Talks

Communication Platform

**Discussion**

# Our dataflows & workflow

---

- Raw data and management plan
- Analyses
  - **SQLShare [qdod]**
  - **iPlant Collaborative**
  - **IPython / GitHub**
- Publishing
  - Online Notebooks
  - GitHub

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

everything else...

static

Data Tables

Publications

Pathways

Interactions

Gene Ontologies

Orthologs

CpG statistics

Structural Elements

Other species genomes

Transposable Elements

Gene Annotations

Sequence Motifs

Transcription Factors  
Binding Sites

Genome

transcripts

dynamic

Primary  
Data Table  
Groupings

Gene Expression

Genetic Variation

Epigenetic Features

RNA-Sequencing

Single Nucleotide Polymorphisms

DNA Methylation

Expressed Sequence Tags

Amplified Fragment  
Length Polymorphisms

Histone Modification

Expression Microarrays

Simple Sequence Repeats

miRNA Expression

Genomic  
Data Types

Size  
Growth  
Location  
Environment  
Stage  
Treatment  
Tissue  
Trait  
Strain



Biology

Environment

Molecular

**Data Analysis**

eScience

iPlant Galaxy

**Notebooks**

Rationale

Platforms

**Open Science**

Data

everything else...

**Phenotype**

Yield

Disease Resistance

Increased Growth Rate

Tissue Quality

Fecundity

Appearance

**Gene Expression**

**Epigenetics**

**Genetics**

**Environment**

- Single Nucleotide Polymorphisms
- Simple Sequence Repeats
- Amplified Fragment Length Polymorphisms

- DNA Methylation Patterns
- miRNA Expression
- Histone Modifications

Temperature

Diet



Biology

Environment

Molecular

**Data Analysis**

eScience

iPlant Galaxy

**Notebooks**

Rationale

Platforms

**Open Science**

Data

everything else...

*static*

**Data Tables**

Publications

Pathways

Interactions

Gene Ontologies

Orthologs

CpG statistics

Structural Elements

Other species genomes

Transposable Elements

Gene Annotations

Sequence Motifs

Transcription Factors  
Binding Sites

**Genome**

*transcripts*

*dynamic*

**Primary  
Data Table  
Groupings**

Gene Expression

Genetic Variation

Epigenetic Features

RNA-Sequencing

Single Nucleotide Polymorphisms

DNA Methylation

Expressed Sequence Tags

Amplified Fragment  
Length Polymorphisms

Histone Modification

Expression Microarrays

Simple Sequence Repeats

miRNA Expression

**Genomic  
Data Types**

Size  
Growth  
Location  
Environment  
Stage  
Treatment  
Tissue  
Trait  
Strain



## eScience Institute

### Querying Disparate Oyster Datasets | qDOD

The goal of this project is to produce a web-based interface for querying and visualizing *Crassostrea gigas* genomic datasets. This site serves as a portal for documenting our efforts, providing user access, as well as a means to gather feedback.

#### Preliminary Phase: Aggregating Datasets

Using [SQLShare](#) as a platform we have already begun to aggregate and format data. Anyone can view (and contribute) using the tag "qDOD". Below is a table describing some of the relevant datasets. "Snapshot" provides you with a screenshot of the data in SQLShare and "Direct Link" brings you directly to the data in SQLShare. You can also [open the table in a new webpage](#).

qDOD online			
qDOD_Cgigas_gene_fasta	sequence fasta file. Exon only.	<a href="http://goo.gl/ogCxl">http://goo.gl/ogCxl</a>	<a href="https://sqlshare.esc">https://sqlshare.esc</a>
qDOD_Zhang_Gil_gene_RNA-seq	Gill RNA-seq data (gene based)	<a href="http://goo.gl/8oISR">http://goo.gl/8oISR</a>	<a href="https://sqlshare.esc">https://sqlshare.esc</a>
qDOD_Zhang_Mgo_gene_RNA-seq	Male Gonad RNA-seq data (gene based)	<a href="http://goo.gl/6buVz">http://goo.gl/6buVz</a>	<a href="https://sqlshare.esc">https://sqlshare.esc</a>

**DATA**





## SQLSHARE

SQLShare is an easier way to store and share your data. Get answers to your research questions right now.

The screenshot shows the SQLShare interface with a query editor and a results table. The query is: `SELECT year,month,observedFlux FROM BioabvgsolarIndices`. The results table has columns: class\_a, class\_b, accession, name, accession, name, accession, name.

class_a	class_b	accession	name	accession	name	accession	name
Amino acid biosynthesis	Aromatic amino acid family	TIGR00033	arvC	3R00033	arvC	TIGR00033	arvC
Amino acid biosynthesis	Aromatic amino acid	TIGR00034	arvGH	3R00034	arvGH	TIGR00034	arvGH

Log in using your account:

**W** UNIVERSITY of WASHINGTON

Google

Don't have an account?

Create a [Google Account](#) and start using SQLShare quickly.

### Upload

Upload any tabular data and start analyzing instantly. No need to install, configure, or design a database.

### Modify

Exercise the full power of SQL even with zero programming experience: joins, subqueries, set operations.

### Share

Analyze and compare your data collaboratively. Derive new datasets and share them with your colleagues.

<https://sqlshare.esc>  
<https://sqlshare.esc>  
<https://sqlshare.esc>  
<https://sqlshare.esc>

- Your datasets**
- All datasets
  - Shared datasets
  - Recent activity... 18
  - Recently viewed »

Upload dataset  
New query

- YOUR TOP VIEWED**
- qDOD Cgigas ... 18
  - BiGo\_Larvae\_j... 16
  - TJGR\_CCD\_d... 11
  - BiGill\_RNAseq... 10
  - BiGo\_lar\_T3D5 10

- POPULAR TAGS**
- proteomics 318
  - oa 170
  - pnitzsch 139
  - orbitrap 131
  - published 62
  - oyster 51
  - protein 50
  - input 47
  - seaflo 42
  - techtrip 34
  - bioinformatics 26
  - skyline 24
  - oceanography 23
  - ssgcid 18
  - qdod2 18
  - qdod 18
  - swissprot 17
  - sun 16
  - tsg 16

**Your Datasets**

Filter dataset by keyword:

Name	Sharing / Owner	Modified
qdod_proteome_blast_mouse	sr320@washington.edu	Jan 2
qDOD_v9_gene GFF format file of oyster genes ~28k <span style="background-color: #e0f0ff; padding: 2px;">gene</span>	sr320@washington.edu	Nov
_qdod_goslim_graphstest	sr320@washington.edu	Oct 2
SNP_RNAseqLibrary_SB_BiGill SNP table from RNA-seq library - SB gill tissue pool (BiGill complement) <span style="background-color: #e0f0ff; padding: 2px;">qdod2</span>	sr320@washington.edu	Oct 2
BiGill_meth_Zhang_exp Gene-centric data including length, CG, percent methylation (gill) and tissue specific RPKM data from Zhang et a <span style="background-color: #e0f0ff; padding: 2px;">qdod2</span>	sr320@washington.edu	Oct 2
qDOD_Cgigas_gene_fasta Tabular format of Cgigas gene sequence fasta file Derived using Dataset: Genomic data from the Pacific oyste <span style="background-color: #e0f0ff; padding: 2px;">qdod2</span>	sr320@washington.edu	Oct 2
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evalues associated with Cgigas 28k genes Derived using Dataset: Genomik <span style="background-color: #e0f0ff; padding: 2px;">blast</span>	sr320@washington.edu	Oct 2
file0	sr320@washington.edu	Aug 1
BiGill meth with SP	sr320@washington.edu	Aug 1
SPID and GO Numbers Swiss-Prot IDs and corresponding GO numbers <span style="background-color: #e0f0ff; padding: 2px;">qdod</span>	sr320@washington.edu	Aug 1
Cgigas_larvae_RNAseq_OsHV_GO	sr320@washington.edu	Jul 2
qDOD_Cgigas_GO_GOslim_DISTINCT	sr320@washington.edu	Jul 2
Cgigas Larvae RNA-Seq OsHV UR10 RNA-seq data with descriptions of larvae exposed to OsHV. (>= 10 UniqueReads) <span style="background-color: #e0f0ff; padding: 2px;">oyster</span>	sr320@washington.edu	Jul 1
Cgigas Larvae RNA-Seq OsHV RNA-seq data with descriptions of larvae exposed to OsHV <span style="background-color: #e0f0ff; padding: 2px;">oshv</span>	sr320@washington.edu	Jul 1
Zhang_Mgo_gene_RNA-seq_IGV <span style="background-color: #e0f0ff; padding: 2px;">sperm</span>	sr320@washington.edu	Jun 2
Zhang_Gil_gene_RNA-seq_IGV IGV format <span style="background-color: #e0f0ff; padding: 2px;">rna-seq</span>	sr320@washington.edu	Jun 2
BiGill_methratio_Gene_Genomic_GFF GFF formatted file indicated DNA methylation on oyster genes <span style="background-color: #e0f0ff; padding: 2px;">qdod</span>	sr320@washington.edu	May
TJGR_GeneBased_CDS_GFF GFF format file with exons indicated for genes in oyster genome <span style="background-color: #e0f0ff; padding: 2px;">qdod</span>	sr320@washington.edu	May
BiGill_Gene_Methratio_VD	sr320@washington.edu	May
oyster_v9_mRNA GFF GFF (gene) from Zhang et al. Column9 modified for Joining <span style="background-color: #e0f0ff; padding: 2px;">qdod</span>	sr320@washington.edu	May
Cgigas gene length CDS only	sr320@washington.edu	Ma



## Your datasets

All datasets  
Shared datasets  
Recent activity... 18  
Recently viewed »

Upload dataset  
New query

## YOUR TOP VIEWED

qDOD Cgigas ... 18  
BiGo\_Larvae\_j... 16  
TJGR\_CCD\_d... 11  
BiGill\_RNAseq... 10  
BiGo\_lar\_T3D5 10

## POPULAR TAGS

proteomics 318  
oa 170  
pnitzsch 139  
orbitrap 131  
published 62  
oyster 51  
protein 50  
input 47  
seaflo 42  
techtrip 34  
bioinformatics 26  
skyline 24  
oceanography 23  
ssgcid 18  
qdod2 18  
qdod 18  
swissprot 17  
suna 16  
tsg 16

## Your Datasets

Filter dataset by keyword: 

Name	Sharing / Owner	Modi
qdod_proteome_blast_mouse	🔒 sr320@washington.edu	Jan 2
qDOD_v9_gene GFF format file of oyster genes ~28k gene		Nov
_qdod_goslim_graphstest		Oct 2
SNP_RNAseqLibrary_SB_BiGill SNP table from RNA-seq library - SB qdod2		Oct 2
BiGill_meth_Zhang_exp Gene-centric data including length, CG, per qdod2		Oct 2
qDOD_Cgigas_gene_fasta Tabular format of Cgigas gene sequence qdod2		Oct 2
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evalu blast		Oct 2
file0	🔒 sr320@washington.edu	Aug 1
BiGill meth with SP	🔒 sr320@washington.edu	Aug 1
SPID and GO Numbers Swiss-Prot IDs and corresponding GO numbers qdod	🔒 sr320@washington.edu	Aug 1
Cgigas_larvae_RNAseq_OsHV_GO	🔒 sr320@washington.edu	Jul 2
qDOD_Cgigas_GO_GOslim_DISTINCT	🔒 sr320@washington.edu	Jul 2
Cgigas Larvae RNA-Seq OsHV UR10 RNA-seq data with descriptions of larvae exposed to OsHV. (>= 10 UniqueReads) oyster	🔒 sr320@washington.edu	Jul 1
Cgigas Larvae RNA-Seq OsHV RNA-seq data with descriptions of larvae exposed to OsHV	🔒 sr320@washington.edu	Jul 1
Zhang_Mgo_gene_RNA-seq_IGV	🔒 sr320@washington.edu	Jun 2
Zhang_Gil_gene_RNA-seq_IGV IGV format rna-seq	🔒 sr320@washington.edu	Jun 2
BiGill_methratio_Gene_Genomic_GFF GFF formatted file indicated DNA methylation on oyster genes qdod	🔒 sr320@washington.edu	May
TJGR_GeneBased_CDS_GFF GFF format file with exons indicated for genes in oyster genome qdod	🔒 sr320@washington.edu	May
BiGill_Gene_Methratio_VD	🔒 sr320@washington.edu	May
oyster_v9_mRNA GFF GFF (gene) from Zhang et al. Column9 modified for Joining qdod	🔒 sr320@washington.edu	May
Cgigas gene length CDS only	🔒 sr320@washington.edu	Ma

## Use Cases

- Joining on Annotations
- File Conversion
- Querying Gene Tables

Your datasets	
All datasets	
Shared datasets	
Recent activity...	18
Recently viewed »	
Upload dataset	
New query	
YOUR TOP VIEWED	
qDOD Cgigas ...	18
BiGo_Larvae_j...	16
TJGR_CCD_d...	11
BiGill_RNAseq...	10

## Your Datasets

Name	Sharing / Owner	Modi
qdod_proteome_blast_mouse	sr320@washington.edu	Jan 2
qDOD_v9_gene GFF format file of oyster genes ~28k gene	sr320@washington.edu	Nov
_qdod_goslim_graphstest	sr320@washington.edu	Oct 2
SNP_RNAseqLibrary_SB_BiGill SNP table from RNA-seq library - SB gill tissue pool (BiGill complement) qdod2	sr320@washington.edu	Oct 2
BiGill_meth_Zhang_exp Gene-centric data including length, CG, percent methylation (gill) and tissue specific RPKM data from Zhang et a qdod2	sr320@washington.edu	Oct 2
qDOD_Cgigas_gene_fasta Tabular format of Cgigas gene sequence fasta file Derived using Dataset: Genomic data from the Pacific oyst qdod2	sr320@washington.edu	Oct 2
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evaluaes associated with 28k genes. Derived using Dataset: Genomic blast	sr320@washington.edu	Oct 2

## Use Cases

- Joining on Annotations
- File Conversion
- Querying Gene Tables

Secondary stress: proteomics

Original input file had some peptides of charge state >2, so had to redo everything with fixed input file.

SR discovered that for some proteins, a peptide was sequenced multiple times and so had multiple expression values. From the unique protein associations file in SQLshare, I summed the expression values for all identical peptides.

```
SELECT [peptide sequence], SUM([2_01 TotalArea]) AS CG2_01, SUM([2_02 TotalArea]) AS CG2_02, SUM([2_03
TotalArea]) AS CG2_03, SUM([5_01 TotalArea]) AS CG5_01, SUM([5_02 TotalArea]) AS CG5_02, SUM([5_03
TotalArea]) AS CG5_03, SUM([8_01 TotalArea]) AS CG8_01, SUM([8_02 TotalArea]) AS CG8_02, SUM([8_03
TotalArea]) AS CG8_03, SUM([11_01 TotalArea]) AS CG11_01, SUM([11_02 TotalArea]) AS CG11_02, SUM([11_03
TotalArea]) AS CG11_03, SUM([26_01 TotalArea]) AS CG26_01, SUM([26_02 TotalArea]) AS CG26_02, SUM([26_03
TotalArea]) AS CG26_03, SUM([29_01 TotalArea]) AS CG29_01, SUM([29_02 TotalArea]) AS CG29_02, SUM([29_03
TotalArea]) AS CG29_03, SUM([32_01 TotalArea]) AS CG32_01, SUM([32_02 TotalArea]) AS CG32_02, SUM([32_03
TotalArea]) AS CG32_03, SUM([35_01 TotalArea]) AS CG35_01, SUM([35_02 TotalArea]) AS CG35_02, SUM([35_03
TotalArea]) AS CG35_03, SUM([221_01 TotalArea]) AS CG221_01, SUM([221_02 TotalArea]) AS CG221_02,
SUM([221_03 TotalArea]) AS CG221_03, SUM([224_01 TotalArea]) AS CG224_01, SUM([224_02 TotalArea]) AS
CG224_02, SUM([224_03 TotalArea]) AS CG224_03, SUM([227_01 TotalArea]) AS CG227_01, SUM([227_02
TotalArea]) AS CG227_02, SUM([227_03 TotalArea]) AS CG227_03, SUM([230_01 TotalArea]) AS CG230_01,
SUM([230_02 TotalArea]) AS CG230_02, SUM([230_03 TotalArea]) AS CG230_03,
SUM([242_01 TotalArea]) AS CG242_01, SUM([242_02 TotalArea]) AS CG242_02, SUM([242_03 TotalArea]) AS
CG242_03, SUM([245_01 TotalArea]) AS CG245_01, SUM([245_02 TotalArea]) AS CG245_02, SUM([245_03
TotalArea]) AS CG245_03, SUM([248_01 TotalArea]) AS CG248_01, SUM([248_02 TotalArea]) AS CG248_02,
SUM([248_03 TotalArea]) AS CG248_03, SUM([251_01 TotalArea]) AS CG251_01, SUM([251_02 TotalArea]) AS
CG251_02, SUM([251_03 TotalArea]) AS CG251_03
```

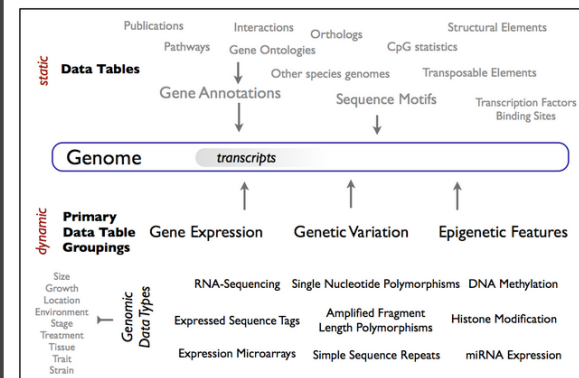
sr320@washington.edu	Aug 1
sr320@washington.edu	Aug 1
sr320@washington.edu	Aug 1
sr320@washington.edu	Jul 2
sr320@washington.edu	Jul 2
sr320@washington.edu	Jul 1
sr320@washington.edu	Jul 1
sr320@washington.edu	Jun 2
sr320@washington.edu	Jun 2
sr320@washington.edu	May
sr320@washington.edu	May
sr320@washington.edu	May
sr320@washington.edu	May
sr320@washington.edu	Ma



# qdod: Querying Disparate Oyster Datasets

This repository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase efficiency of biological discovery. The [wiki](#) associated with this repository will serve as the *primary means for documentation*. Currently the documentation is focused on 1) **describing current datasets** and 2) **providing workflow tutorials**.

In brief, data in the form of delimited text files is aggregated into [SQLShare](#) where they can be easily queried. Below is schematic representation of the different types of datasets.



During the initial phases the focus is on the Pacific oyster and primary data from the [Roberts Lab](#).

## Select IPython Notebooks

- [Static Data Tables - Universal](#)
- [Static Data Tables - Annotations](#)

Currently the documentation is focused on 1) **describing current datasets** and 2) **providing workflow tutorials**.

### A. Raw Data

- [Select NGS Data via Roberts Lab](#)

### B. Datasets in SQLShare

- [Universal](#)
- [Generic Oyster Datasets](#)
- [Tissue Specific Oyster Datasets](#)

### C. Tutorials

- [Simple Gene Search](#)
- [Standard SQLShare Queries](#)
- [Annotating Genes](#)
- [File Format Conversions](#)

### D. Genome Browser Feature Tracks

- [Canonical Tracks](#)
- [Bisulfite sequencing \(gill tissue\)](#)
- [Reference Genome Files](#)

Please use [GitHub's Issue feature](#) to ask question, report problems, or suggest features.

Last edited by sr320, 9 days ago

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

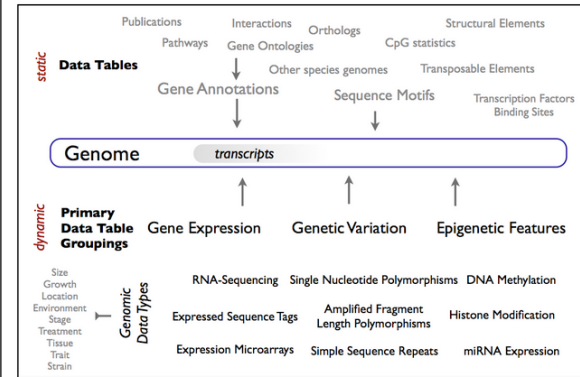
everything else...



# qdod: Querying Disparate Oyster Datasets

This repository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase efficiency of biological discovery. The [wiki](#) associated with this repository will serve as the *primary means for documentation*. Currently the documentation is focused on 1) **describing current datasets** and 2) **providing workflow tutorials**.

In brief, data in the form of delimited text files is aggregated into [SQLShare](#) where they can be easily queried. Below is schematic representation of the different types of datasets.



During the initial phases the focus is on the Pacific oyster and primary data from the [Roberts Lab](#).

## Select IPython Notebooks

- [Static Data Tables - Universal](#)
- [Static Data Tables - Annotations](#)

## Select Genomic Data

ID	Platform	Molecule	Tissue	Length	Files
BB3	SOLiD	RNA	gill	25 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>
DH3	SOLiD	RNA	gill	25 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>
DH2	SOLiD	RNA	gill	25 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>
GE	SOLiD	RNA	larvae	50 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>
GC	SOLiD	RNA	larvae	50 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>
SBunmeth	SOLiD	DNA	gill	25 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>
SBmeth	SOLiD	DNA	gill	25 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>
BSseqGill	Illumina	DNA	gill	36 x 1	<a href="#">fastq</a>
ETStageseq	Illumina	RNA	gill		<a href="#">zip</a>
BSseqSperm	Illumina	DNA	sperm	72 x 2	<a href="#">fastq1</a> ; <a href="#">fastq2</a>
BiGillRNA	Illumina	RNA	gill	50 x 2	<a href="#">fastq1</a> ; <a href="#">fastq2</a>
BiGoRNA	Illumina	RNA	sperm	50 x 2	<a href="#">fastq1</a> ; <a href="#">fastq2</a>

Currently the documentation is focused on

### A. Raw Data

- [Select NGS Data via Roberts Lab](#)

### B. Datasets in SQLShare

- [Universal](#)
- [Generic Oyster Datasets](#)
- [Tissue Specific Oyster Datasets](#)

### C. Tutorials

- [Simple Gene Search](#)
- [Standard SQLShare Queries](#)
- [Annotating Genes](#)
- [File Format Conversions](#)

### D. Genome Browser Feature Tracks

- [Canonical Tracks](#)
- [Bisulfite sequencing \(gill tissue\)](#)
- [Reference Genome Files](#)

Please use [GitHub's Issue feature](#) to ask question, report problems, or suggest features.

Last edited by sr320, 9 days ago

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

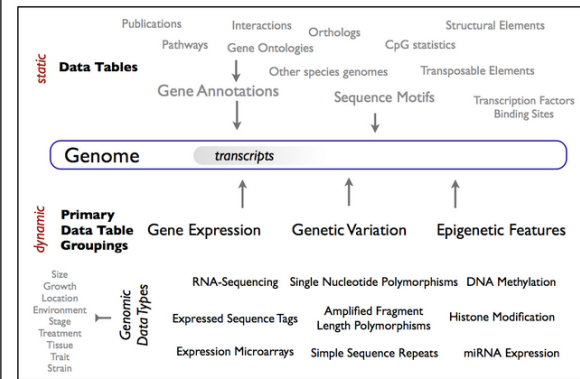
everything else...



# qdod: Querying Disparate Oyster Datasets

This repository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase efficiency of biological discovery. The [wiki](#) associated with this repository will serve as the *primary means for documentation*. Currently the documentation is focused on 1) **describing current datasets** and 2) **providing workflow tutorials**.

In brief, data in the form of delimited text files is aggregated into [SQLShare](#) where they can be easily queried. Below is schematic representation of the different types of datasets.



During the initial phases the focus is on the Pacific oyster and primary data from the [Roberts Lab](#).

## Select IPython Notebooks

- [Static Data Tables - Universal](#)
- [Static Data Tables - Annotations](#)

Currently the documentation is focused on

### A. Raw Data

- [Select NGS Data via Roberts Lab](#)

### B. Datasets in SQLShare

- [Universal](#)
- [Generic Oyster Datasets](#)
- [Tissue Specific Oyster Datasets](#)

### C. Tutorials

- [Simple Gene Search](#)
- [Standard SQLShare Queries](#)
- [Annotating Genes](#)
- [File Format Conversions](#)

### D. Genome Browser Feature Tracks

- [Canonical Tracks](#)
- [Bisulfite sequencing \(gill tissue\)](#)
- [Reference Genome Files](#)

## Select Genomic Data

ID	Platform	Molecule	Tissue	Length	Files
BB3	SOLiD	RNA	gill	25 x 1	<a href="#">csfasta; qual</a>
DH3	SOLiD	RNA	gill	25 x 1	<a href="#">csfasta; qual</a>

## Data Snapshots

Select datasets available from SQLShare. Tag: [qdod2](#)  
<https://sqlshare.escience.washington.edu/sqlshare/#s=tag/qdod2>

## Universal

Biology

Environment

Molecular

Data Analysis

eScience

Plant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

everything else...

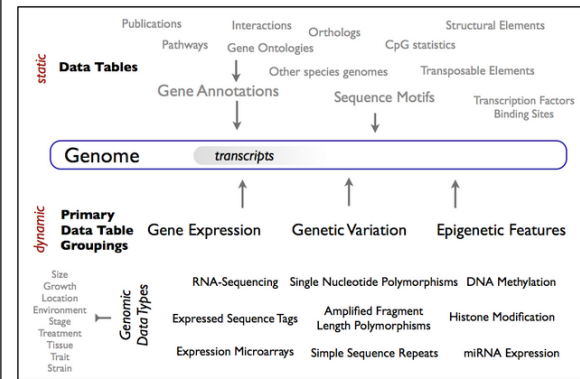




# qdod: Querying Disparate Oyster Datasets

This repository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase efficiency of biological discovery. The [wiki](#) associated with this repository will serve as the *primary means for documentation*. Currently the documentation is focused on 1) **describing current datasets** and 2) **providing workflow tutorials**.

In brief, data in the form of delimited text files is aggregated into [SQLShare](#) where they can be easily queried. Below is schematic representation of the different types of datasets.



During the initial phases the focus is on the Pacific oyster and primary data from the [Roberts Lab](#).

## Select IPython Notebooks

- [Static Data Tables - Universal](#)
- [Static Data Tables - Annotations](#)

## Select Genomic Data

ID	Platform	Molecule	Tissue	Length	Files
BB3	SOLID	RNA	gill	25 x 1	<a href="#">csfasta; qual</a>
DH3	SOLID	RNA	gill	25 x 1	<a href="#">csfasta; qual</a>

## Data Snapshots

Select datasets available from SQLShare. Tag: [qdod2](#)  
<https://sqlshare.escience.washington.edu/sqlshare/#s=tag/qdod2>

## Universal

## Workflow 1: Annotating Oyster Genes

This workflow will take focus on taking a simple SQLShare table that has gene IDs and associated expression data and will take you through the steps of figuring out the name, function, etc of each gene.

### Initial Data Table: Oyster larvae RNA-seq - OsHV exposure

SCREENSHOT

[solid0078\\_20091105\\_RobertsLab\\_GE\\_F3 trimmed RNA-Seq.txt](#) < Viewable by everyone

OsHV RNA-seq on Version 9 transcriptome

[Click here to add a tag](#)

```
SELECT
["Feature ID"] as ID,
["Unique gene reads"] as UniqueReads,
["Total gene reads"] as TotalReads,
["RPKM"] as RPKM
FROM [ar320@washington.edu].[table_solid0078_20091105_RobertsLab_GE_F3 trim
```

DATASET PREVIEW Rows 1 - 100 of 28027 | Columns 4 of 4

ID	UniqueReads	TotalReads	RPKM
CGI_10000001	0	10	5.23
CGI_10000002	5	5	2.756
CGI_10000003	0	0	0
CGI_10000004	0	0	0
CGI_10000005	0	0	0
CGI_10000006	0	0	0

Currently the documentation is focused on

### A. Raw Data

- [Select NGS Data via Roberts Lab](#)

### B. Datasets in SQLShare

- [Universal](#)
- [Generic Oyster Datasets](#)
- [Tissue Specific Oyster Datasets](#)

### C. Tutorials

- [Simple Gene Search](#)
- [Standard SQLShare Queries](#)
- [Annotating Genes](#)
- [File Format Conversions](#)

### D. Genome Browser Feature Tracks

- [Canonical Tracks](#)
- [Bisulfite sequencing \(gill tissue\)](#)
- [Reference Genome Files](#)

Please use [GitHub's Issue feature](#) to ask

Last edited by sr320, 9 days ago



[github.com/sr320/qdod/wiki](https://github.com/sr320/qdod/wiki)

Biology

Environment

Molecular

Data Analysis

eScience

Plant Galaxy

ebooks

ationale

latforms

cience

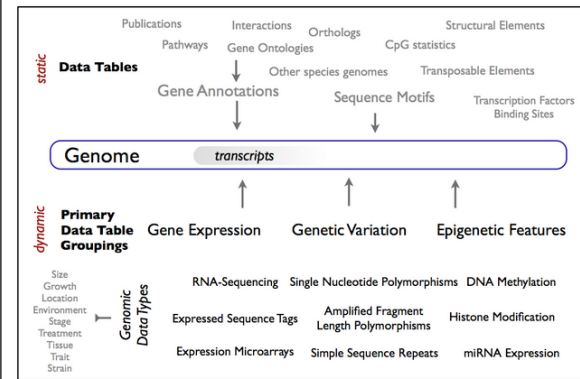
Data

g else...

# qdod: Querying Disparate Oyster Datasets

This repository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase efficiency of biological discovery. The wiki associated with this repository will serve as the primary means for documentation. Currently the documentation is focused on 1) describing current datasets and 2) providing workflow tutorials.

In brief, data in the form of delimited text files is aggregated into SQLShare where they can be easily queried. Below is schematic representation of the different types of datasets.



During the initial phases the focus is on the Pacific oyster and primary data from the Roberts Lab.

## Select IPython Notebooks

- Static Data Tables - Universal
- Static Data Tables - Annotations



# eScience Institute

Biology

Environment

Molecular

## Workshop

### Data Snapshots

Edit Page Page History Clone URL

Select datasets available from SQLShare. Tag: `qdod2`  
<https://sqlshare.escience.washington.edu/sqlshare/#s=tag/qdod2>

### Universal

Dataset	Screenshot	more
UniprotProtNamesReviewed_yes20130610		etc
SPID and GO Number...		

## Data Analysis

eScience

Plant Galaxy

Currently the documentation is focused on

### A. Raw Data

- Select NGS Data via Roberts Lab

### B. Datasets in SQLShare

### Annotating Oyster Genes

Edit Page Page History Clone URL

focus on taking a simple SQLShare table that has gene IDs and associated expression data and will take you through the name, function, etc of each gene.

### Example: Oyster larvae RNA-seq - OsHV exposure

RobertsLab\_GE\_F3 trimmed RNA-Seq.txt

transcriptome

```

select ID,
       reads as UniqueReads,
       reads as TotalReads,
       ...
from [table_solid0078_20091105_RobertsLab_GE_F3 trim

```

1 - 100 of 28027 | Columns 4 of 4

3 4 5 next > last >>

UniqueReads	TotalReads	RPKM
10	5.23	
5	2.756	
0	0	
0	0	
0	0	
0	0	

ebooks

rationale

platforms

science

Data

g else...



Value in sitting around table





## The iPlant Collaborative

The iPlant Collaborative develops cyberinfrastructure and computational tools to solve Grand Challenges in plant science

CHALLENGE

DISCOVER

LEARN

CONNECT

Biology

Environment

Molecular

**Data Analysis**

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

**Open Science**

Data

everything else...

## The iPlant Collaborative

The iPlant Collaborative develops cyberinfrastructure and computational tools to solve Grand Challenges in plant science

Biology

Environment

Molecular

**Data Analysis**

### Discovery Environment



Biology

Environment

Molecular

# Data Analysis

The screenshot displays the Discovery Environment interface, which is used for data analysis. It features a teal header with the 'Discovery Environment' logo and a 'Notifications' badge showing 39 alerts. The interface is divided into three main panels:

- Data Panel:** Shows a file browser with a 'Navigation' pane on the left listing folders like 'labshare', 'qdod', and 'Cgigas\_v9'. The main area shows a table of files with columns for 'Name', 'Last Modified', and 'Details'. Files include 'Cuffdiff2\_analy...', 'Cufflinks2\_anal...', 'Cuffmerge2\_an...', 'logs', 'TopHat2-PE\_Bi...', 'TopHat2-SE\_a...', 'Uncompress\_fil...', 'BiGIIIRNA\_GAC...', 'BiGoRNA\_GTG...', 'Crassostrea\_gi...', 'oyster.v9.glean...', 'T1.fq.gz', 'T2.fq.gz', 'U1.fq.gz', and 'U2.fq.gz'.
- Apps Panel:** Shows a search bar and a list of applications categorized into 'Workspace (8)', 'Public Apps (434)', 'Experimental (28)', 'General Utilities (73)', and 'NGS (111)'. The 'Workspace' section lists apps like 'Uncompress files wi...', 'RNAseq2bedgraph', 'RNAseq2bedgraph SE', 'FastQC 0.10.1 (mul...', 'My phymI', 'BSMAP', 'Workflow test', and 'Copy of TopHat2-P...'. The 'NGS' section includes 'Aligners (16)', 'Assemblers (17)', 'Assembly Annotation (9)', 'Bisulfite (3)', 'ChIPseq (4)', 'QC and Processing (21)', 'SAMTools (6)', 'Transcriptome Profiling (11)', and 'Variant Identification (8)'.
- Analyses Panel:** Shows a table of completed and failed analyses. The table has columns for 'Name', 'App', 'Start Date', 'End Date', and 'Status'. Analyses include 'genomeCoverageBed\_Bi...', 'RNAseq2bedgraph\_anal...', 'RNAseq2bedgraph SE\_BB3', 'RNAseq2bedgraph\_BiGo...', 'genomeCoverageBed\_a...', 'bamToBed\_analysis1', 'VCF to GFF3\_analysis1', 'Find SNPS - mpileup\_analysis1', and 'TopHat2-PE\_analysis1'.



# Data Analysis

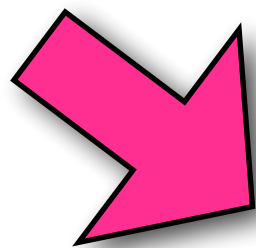


The screenshot displays the Discovery Environment interface, which is used for data analysis. It features a teal header with the 'Discovery Environment' logo on the left. The main workspace is divided into several panels:

- Data Panel:** Located on the left, it includes a 'Navigation' sidebar with a tree view of folders and files. The main area shows a table of files with columns for 'Name', 'Last Modified', and 'Size'. The 'Details' pane on the right of this panel is currently empty, showing the instruction 'Select a file or folder to view its details'.
- Apps Panel:** Located on the right, it has a search bar and a 'Workspace' section. The 'Workspace' section contains a table of installed applications:

Workspace	Name	Integrated by	Rating
Workspace (8)			
Apps under development (6)	Uncompress files wi...	Matthew Vaughn	★★★★★
Favorite Apps (2)	RNAseq2bedgraph	Steven Roberts	★★★★★
My public apps (0)			
Public Apps (434)	RNAseq2bedgraph SE	Steven Roberts	★★★★★
Archive (27)	FastQC 0.10.1 (mul...	Matthew Vaughn	★★★★★

- RNAseq2bedgraph Panel:** A modal window is open for the 'RNAseq2bedgraph' application. It shows the 'Analysis Name' as 'RNAseq2bedgraph\_analysis1' and includes a 'README' section. The 'Input data' section is expanded, showing 'TopHat2-PE for workflows - Reference Genome (Mandatory)'. Below this, there is a dropdown menu for selecting a reference genome and a 'Browse' button for providing a reference genome file in FASTA format.



**SQLSHARE**

**Galaxy**

iPlant Collaborative™

**Hyak**



Biology

Environment

Molecular

**Data Analysis**

eScience

iPlant Galaxy

**Notebooks**

Rationale

Platforms

**Open Science**

Data

everything else...



SQLSHARE

Galaxy

iPlant Collaborative™



BLAST

Tophat

DESeq

R

BSMAP

perl scripts

python modules

STACKS

Bedtools

IGV

Trinity

fastqc

bash (lots)

python APIs

excel

DAVID

Revigo

Biology

Environment

Molecular

**Data Analysis**

eScience

iPlant Galaxy

**Notebooks**

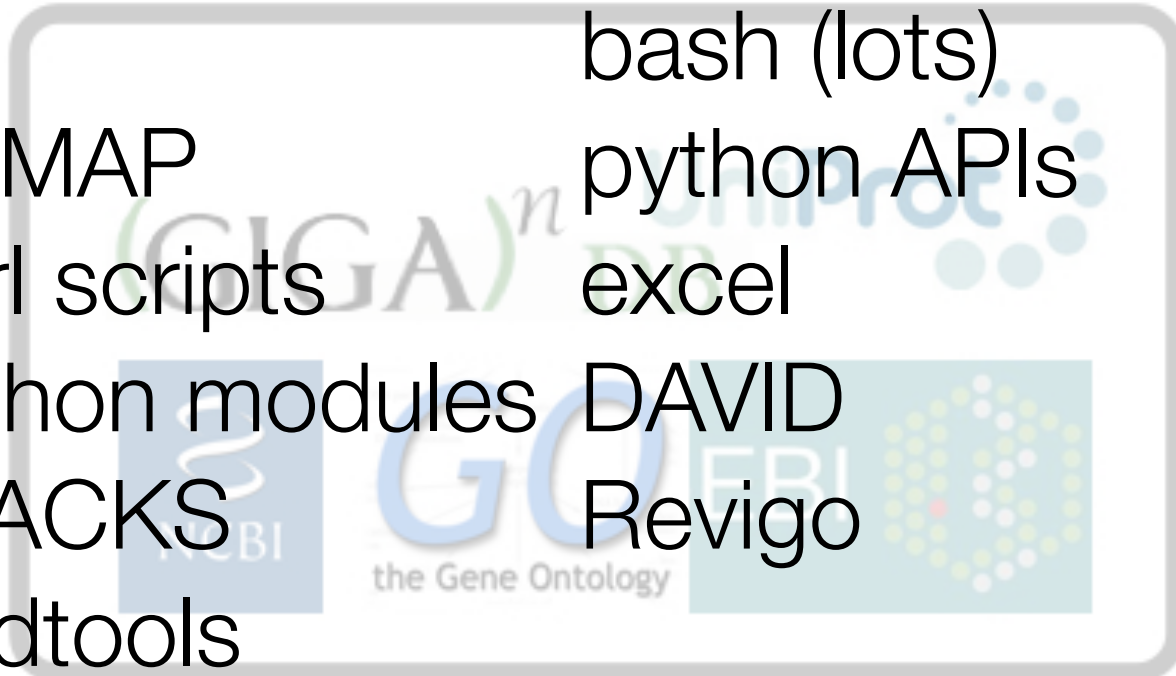
Rationale

Platforms

**Open Science**

Data

everything else...



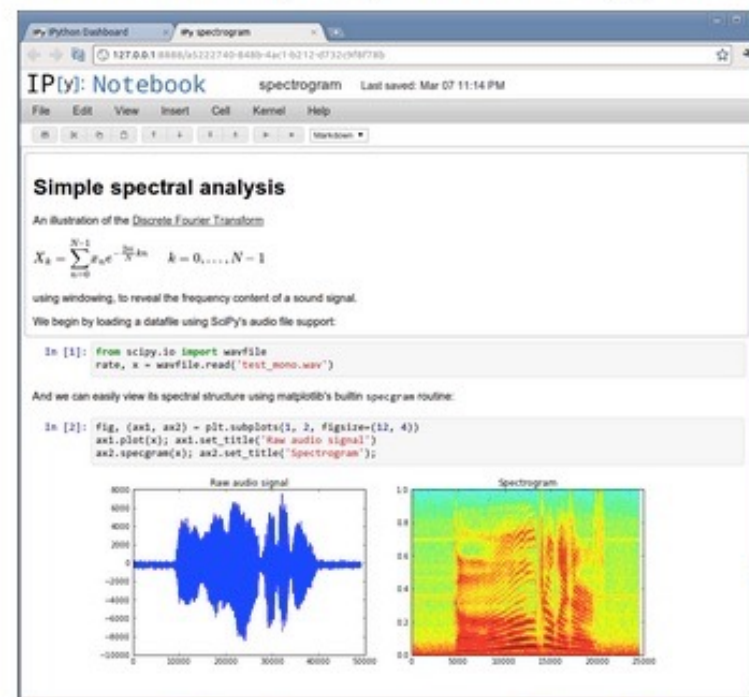
# Jupyter Notebooks (IPython)

**IP[y]:** IPython  
Interactive Computing

[Install](#) · [Docs](#) · [Videos](#) · [News](#) · [Cite](#) · [Sponsors](#) · [Donate](#)

## The IPython Notebook

The IPython Notebook is a web-based interactive computational environment where you can combine code execution, text, mathematics, plots and rich media into a single document:



These notebooks are normal files that can be shared with colleagues, converted to other formats such as HTML or PDF, etc. You can share any publicly available notebook by using the [IPython Notebook Viewer](#) service which will render it as a static web page. This makes it easy to give your colleagues a document





## Fasta2Slim

This IPython notebook is intended to serve as a structured means to annotate sequences using UniProt/SwissProt database. The notebook can be easily modified to personal preferences. As developed, the notebook requires the user has the following software installed ...

- IPython
- NCBI Blast
- SQLShare Python Client

---

### Instructions for use.

In a working directory of your choosing place query fasta file, naming as `query.fa`. Edit the cell below, providing the path to said working directory.

Identify the location of the blast database you would like to use and indicate path in the cell below.

Identify the location of your `sqlshare-pythonclient/tools` and indicate path in the cell below.

Change the input to the `usr` variable to reflect your SQLShare user account.

```
In [2]: #Location Variables
wd="~/Desktop/test/"

db="/Volumes/Bay3/Software/ncbi-blast-2.2.29\+/db/uniprot_sprot_r2013_12"

sqls="~/sqlshare-pythonclient/tools/"

usr="sr320@washington.edu"
```

```
In [254]: !head {wd}query.fa

>PiuraChilensis_v1_contig_1
ATTTACAATACGAAGTAAAATAGATAACGTGAAAATAATCTTGGTGCTGGATGATCGATC
AAGTTCACCAATATTTTATTGTAATAAATCATTCTAAACAGCATGAAATCGTGTACAATG
TATAACAAGCAAATATATAACACTAAAGCAAGAGGGCGTAAGTGGGGGGGTGGGTGAGA
GTAAAAAATTCAAACATGTCAAATACCCCGGCGTTAGCCTTAAAAGCACCATGGACTTCT
CGCTTCATTAAGCAATAAATAAAGAGCTAATAGAGCAATGCAATTAAGCAATAAAGCA
```



File Edit View Insert Cell Kernel Help



Cell Toolbar: None

## Fasta2Slim

This IPython notebook is structured to meet your personal preferences. As

- IPython
- NCBI Blast
- SQLShare Python Client

Run  
Run and Select Below  
Run and Insert Below  
**Run All**  
Run All Above  
Run All Below  
Cell Type ▶  
Current Output ▶  
All Output ▶

structured means to annotate sequences using UniProt/SwissProt. This requires the user has the following software installed ...

### Instructions for use.

In a working directory of your choosing place query fasta file, naming as `query.fa`. Edit the cell below, providing the

Identify the location of the blast database you would like to use and indicate path in the cell below.

Identify the location of your `sqlshare-pythonclient/tools` and indicate path in the cell below.

Change the input to the `usr` variable to reflect your SQLShare user account.

```
In [2]: #Location Variables
wd=~/Desktop/test/

db="/Volumes/Bay3/Software/ncbi-blast-2.2.29\+/db/uniprot_sprot_r2013_12"

sqls=~/sqlshare-pythonclient/tools/

usr="sr320@washington.edu"
```

```
In [254]: !head {wd}query.fa
```

```
In [2]: !head {wd}query.fa
```

```
>PiuraChilensis_v1_contig_1
ATTTACAATACGAAGTAAAATAGATAACGTGAAAATAATCTTGGTGCTGGATGATCGATC
AAGTTCACCAATATTTTATTGTAAAAAATCATTCTAAACAGCATGAAATCGTGTACAATG
TATAACAAGCAAATATATAACACTAAAGCAAGAGGGCGTAAGTGGGGGGGGTGGGTGAGA
GTAAAAAATTCAAACATGTCAAATACCCCGGCGTTAGCCTTAAAAGCACCATGGACTTCT
GCCTTCAATAAGCATAAAATTAAAACACCTAATACACAATGAATATACAGATAAAACAGA
TTTATGAATAGTTGGTGTTACATCTTTTACAGCCATAAGCCTTCATTTTGCTTCCAAACG
TATAAAATCTGACTTGGACAATATACAGCCATGAGATATGACACAGCGAGCACTACAAT
ATATATTTATCTTGTACTATACAGCCTGTACAAGAAAATTCTGGAATTGTCTTCACAAGA
GACAGAAAATAGTTGCAATGTGAATGCTAGTCTACTATTTGATCACAATTGGATAGAAA
```

```
In [3]: #number of sequences
!fgrep -c ">" {wd}query.fa
```

282

## Blast

```
In [4]: !blastx \
-query {wd}query.fa \
-db {db} \
-max_target_seqs 1 \
-max_hsps 1 \
-outfmt 6 \
-evalue 1E-05 \
-num_threads 2 \
-out {wd}blast_sprot.tab
```

## Number of matched sequences:

```
In [5]: !wc -l {wd}blast_sprot.tab
```

```
211 /Users/sr320/Desktop/test/blast_sprot.tab
```

```
In [6]: !tr '|' '\t' <{wd}blast_sprot.tab> {wd}blast_sprot_sql.tab
!head -1 {wd}blast_sprot.tab
!echo SQLShare ready version has Pipes converted to Tabs ....
!head -1 {wd}blast_sprot_sql.tab
```

```
PiuraChilensis_v1_contig_3      sp|Q6P9A1|ZN530_HUMAN  33.33  105  61
3          825      1118      414      516      1e-07  57.4
SQLShare ready version has Pipes converted to Tabs ....
PiuraChilensis_v1_contig_3      sp      Q6P9A1  ZN530_HUMAN      33.33  105
61      3          825      1118      414      516      1e-07  57.4
```

## Joining in SQL Share

```
In [7]: !python {sqls}singleupload.py \
-d _blast_sprot \
{wd}blast_sprot_sql.tab
```

```
processing chunk line 0 to 211 (0.000264167785645 s elapsed)
pushing /Users/sr320/Desktop/test/blast_sprot_sql.tab...
parsing 983DD315...
finished _blast_sprot
```

```
In [8]: !python {sqls}fetchdata.py \
-s "SELECT Column1, term, GOslim_bin, aspect, ProteinName FROM [{usr}].[_b
last_sprot]md left join [samwhite@washington.edu].[UniprotProtNamesReviewe
d_yes20130610]sp on md.Column3=sp.SPID left join [sr320@washington.edu].[S
PID and GO Numbers]go on md.Column3=go.SPID left join [sr320@washington.ed
u].[GO_to_GOslim]slim on go.GOID=slim.GO_id where aspect like 'P'" \
-f tsv \
-o {wd}Godescriptions.txt
```



## Plot GoSlim terms

```
In [10]: pylab inline
```

Populating the interactive namespace from numpy and matplotlib

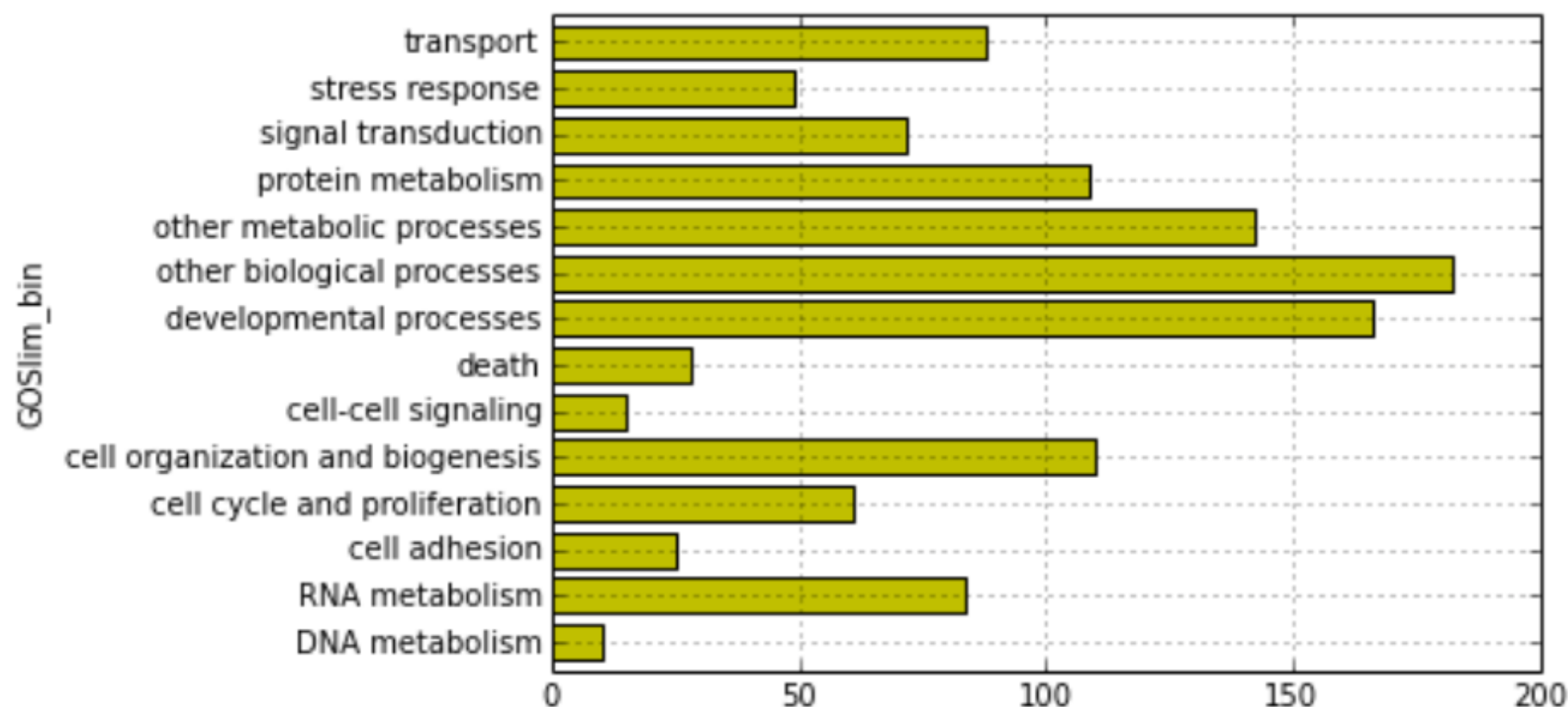
```
In [11]: cd {wd}
```

```
/Users/sr320/Desktop/test
```

```
In [12]: from pandas import *
```

```
gs = read_table('GODESCRIPTIONS.txt')
```

```
In [13]: gs.groupby('GOSlim_bin').Column1.count().plot(kind='barh', color=list('y'))  
savefig('GOSlim.png', bbox_inches='tight')
```



# Value

- Reproducible
- Sharing and Collaborating
- Data Provenance
- Great for teaching / troubleshooting

<https://github.com/sr320>