

# USDA - NSA Workshop

---

## 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](mailto: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 MoPad Etherpad interface. The title bar says "Etherpad is free software" and "MoPad". The main area is a "Public Pad" titled "USDA NRSP-8 / NSA Genomics Resource Coordination Workshop". The content includes:

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

On the right side, there is a sidebar with a user profile for "Steven F" and a link to "Invite other users". At the bottom right, it says "March 25, 2015".

# Today

---

<http://goo.gl/xqf8Og>

Tutorials  
Lightning Talks  
Discussion

Etherpad is free software **MoPad**

Public Pad Pad Options Import/Export

**B I U S** 100%

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 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

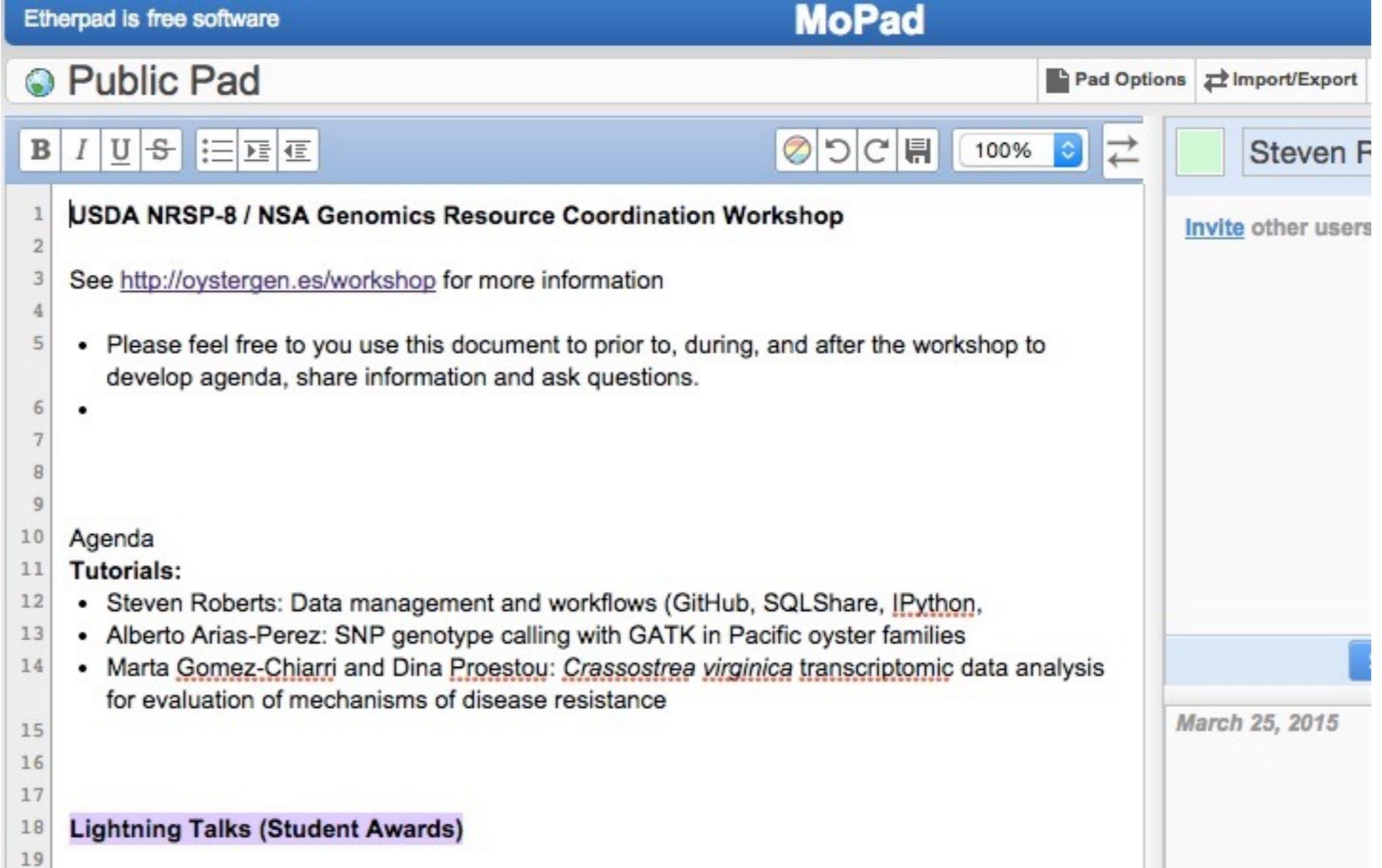
18 **Lightning Talks (Student Awards)**

19

Steven F

Invite other users

March 25, 2015



Today

---

<http://goo.gl/xqf8Og>

*Crassostrea virginica* sequencing

Tutorials

Breeding program

Lightning Talks

Communication Platform

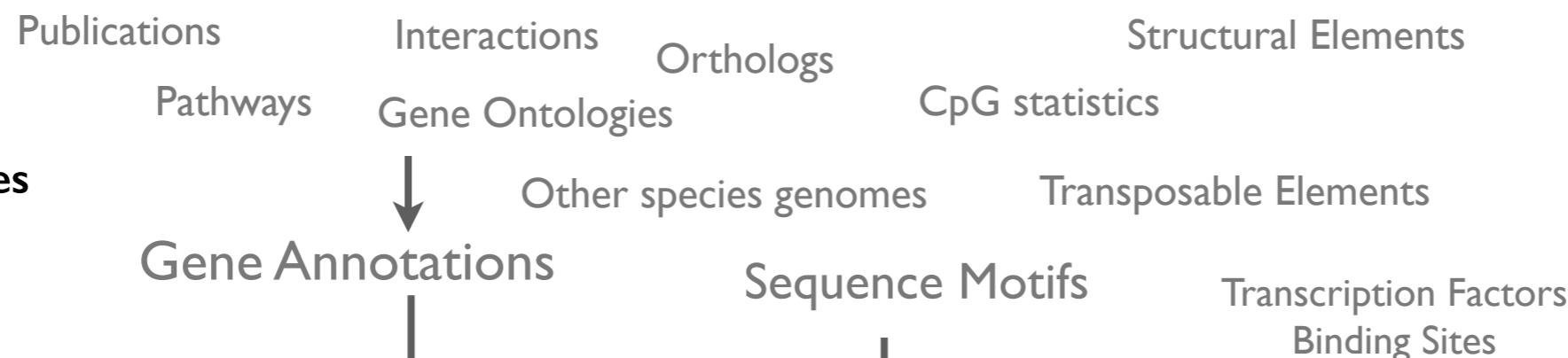
**Discussion**

# Our dataflows & workflow

---

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

static

**Data Tables****Genome***transcripts*

dynamic

**Primary Data Table Groupings**

- Size
- Growth
- Location
- Environment
- Stage
- Treatment
- Tissue
- Trait
- Strain

**Genomic Data Types**

RNA-Sequencing	Single Nucleotide Polymorphisms	DNA Methylation
Expressed Sequence Tags	Amplified Fragment Length Polymorphisms	Histone Modification
Expression Microarrays	Simple Sequence Repeats	miRNA Expression

Biology

Environment

Molecular

## Data Analysis

eScience

iPlant Galaxy

Notebooks

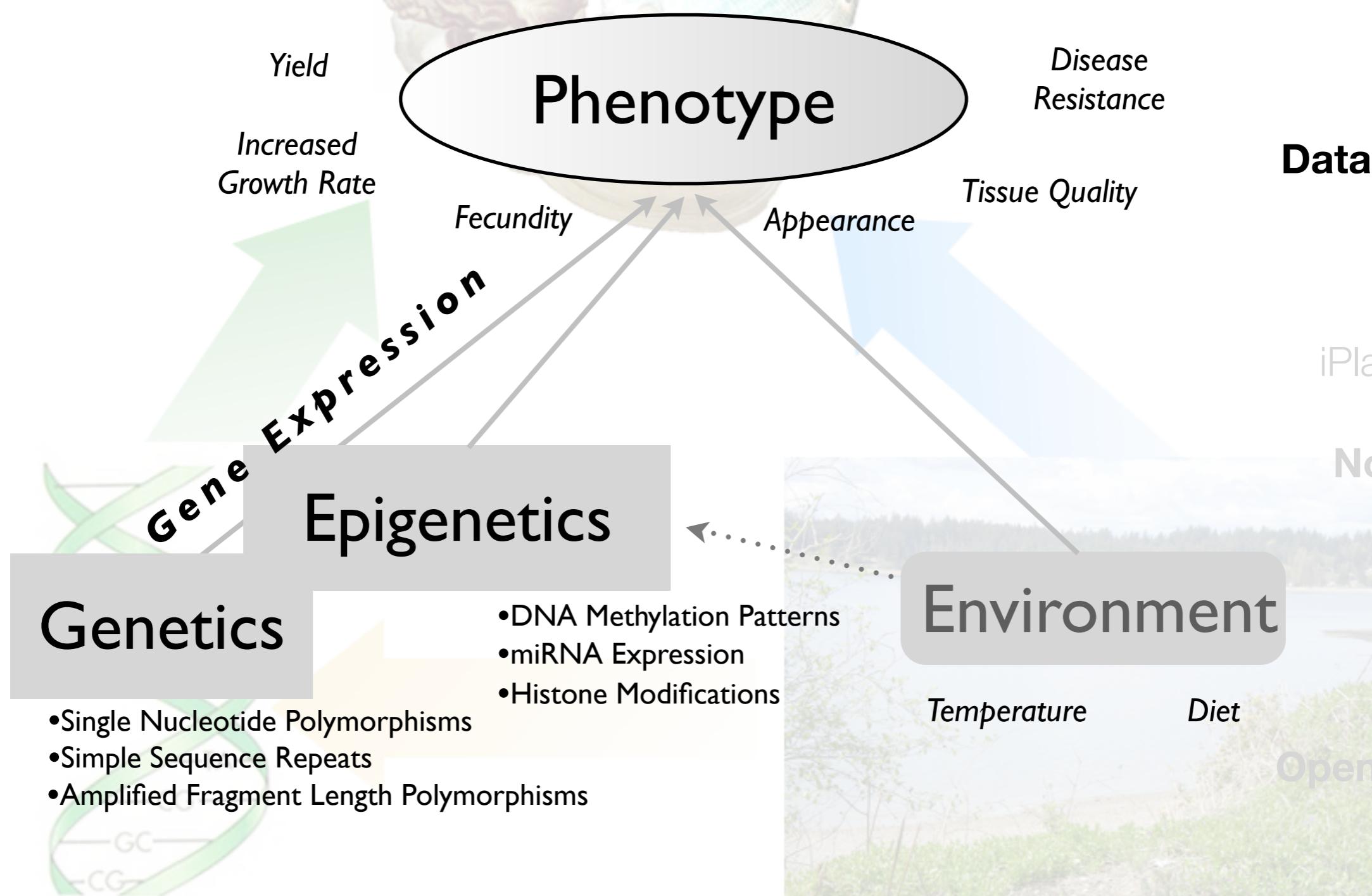
Rationale

Platforms

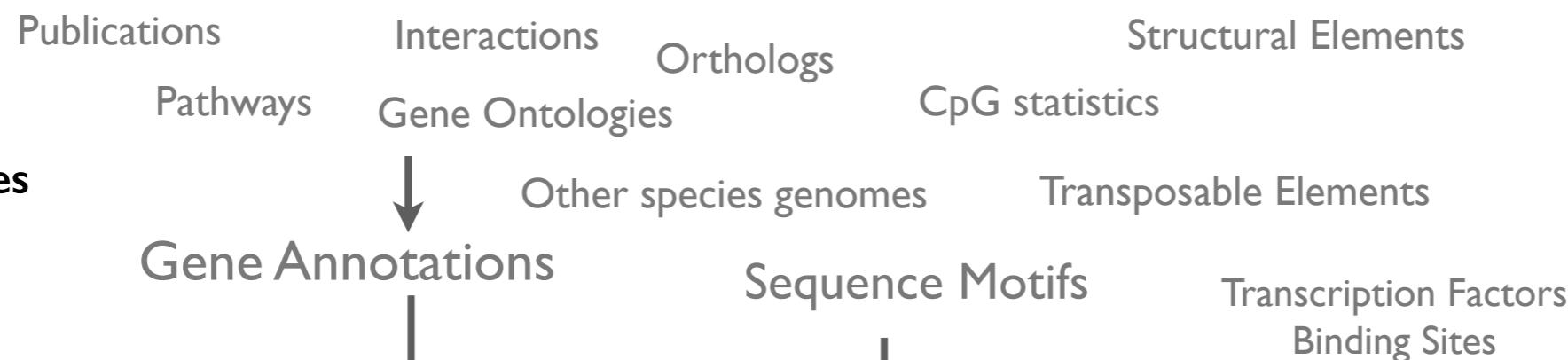
Open Science

Data

everything else...



static

**Data Tables****Genome***transcripts*

dynamic

**Primary Data Table Groupings**

- Size
- Growth
- Location
- Environment
- Stage
- Treatment
- Tissue
- Trait
- Strain

**Genomic Data Types**

RNA-Sequencing	Single Nucleotide Polymorphisms	DNA Methylation
Expressed Sequence Tags	Amplified Fragment Length Polymorphisms	Histone Modification
Expression Microarrays	Simple Sequence Repeats	miRNA Expression



## 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](#).

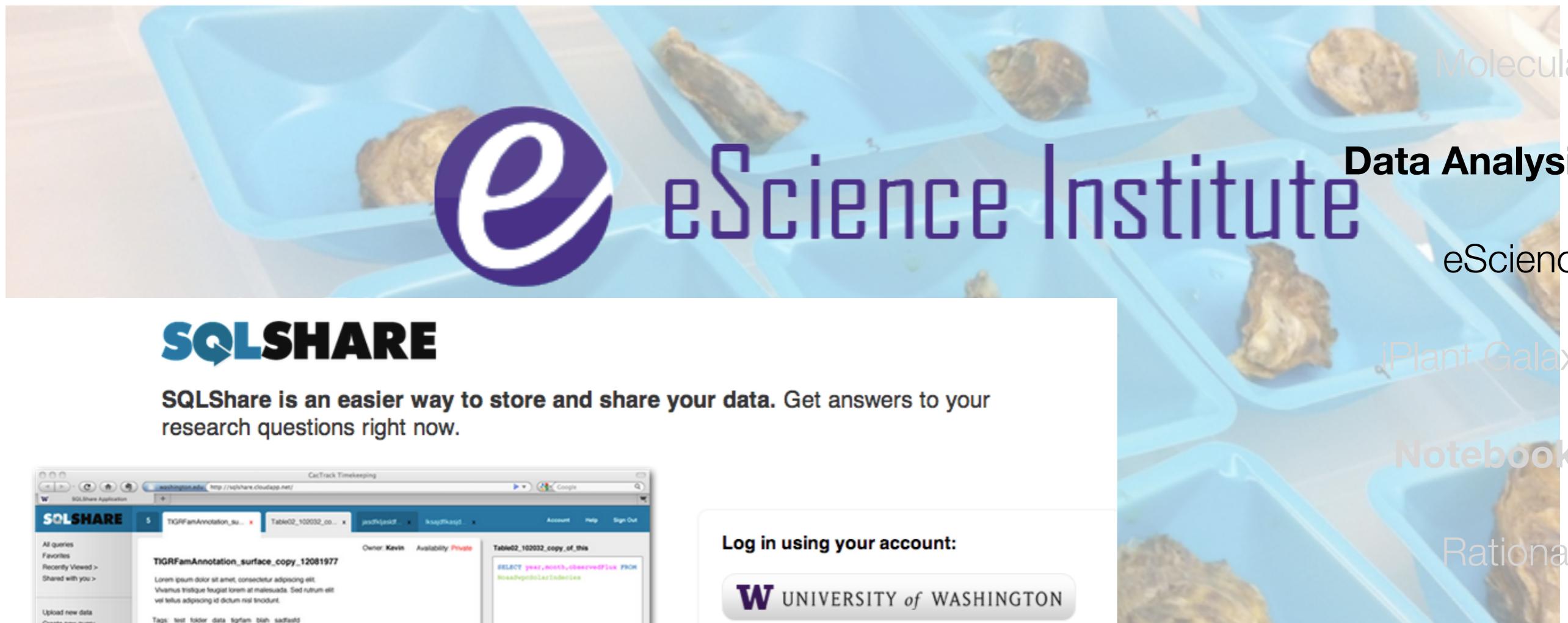
DATA

### qDOD online

qDOD_Cgigas_gene_fasta	sequence fasta file. Exon only.	<a href="http://goo.gl/oqCxl">http://goo.gl/oqCxl</a>	<a href="https://sqlshare.escapevelocity.net/datasets/qdod/qDOD_Cgigas_gene_fasta">https://sqlshare.escapevelocity.net/datasets/qdod/qDOD_Cgigas_gene_fasta</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.escapevelocity.net/datasets/qdod/qDOD_Zhang_Gil_gene_RNA-seq">https://sqlshare.escapevelocity.net/datasets/qdod/qDOD_Zhang_Gil_gene_RNA-seq</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.escapevelocity.net/datasets/qdod/qDOD_Zhang_Mgo_gene_RNA-seq">https://sqlshare.escapevelocity.net/datasets/qdod/qDOD_Zhang_Mgo_gene_RNA-seq</a>

everything else...

One can view (and contribute) using the  
ides you with a screenshot of the data in  
en the table in a new webpage.



## 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 web application. On the left, a sidebar lists 'All queries', 'Favorites', 'Recently Viewed >', 'Shared with you >', 'Upload new data', 'Create new query', and 'Tags' (with entries 'test', 'database', 'blah', 'safefield', 'folder', 'ligfam'). The main area displays two tabs: 'TIGRFamAnnotation\_surface\_copy\_12081977' and 'Table02\_100032\_copy\_of\_this'. The first tab contains a snippet of SQL: 'SELECT year,month,observedFlux FROM RoadvpCpolarIndices'. The second tab also contains a snippet of SQL: 'SELECT year,month,observedFlux FROM RoadvpCpolarIndices'. Below the tabs are two tables:

class_a	class_b	accession	name	class_a	class_b	accession	name
Amino acid biosynthesis	Aromatic amino acid family	TIGR00033	aroC	3R00033	aroC	TIGR00033	aroC
Amino acid biosynthesis	Aromatic amino acid	TIGR00034	aroGH	3R00034	aroGH	TIGR00034	aroGH

At the bottom are buttons for 'DOWNLOAD', 'VISUALIZE', 'COPY', 'QUERY TABLE', and 'DELETE THIS DATASET'.

### Log in using your account:

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.eso.org>  
<https://sqlshare.eso.org>  
<https://sqlshare.eso.org>  
<https://sqlshare.eso.org>  
<https://sqlshare.eso.org>

everything else...

Your Datasets		
		Filter dataset by keyword: <input type="text" value="qdod"/>
All datasets	Name	Sharing / Owner
Shared datasets	qdod_proteome_blast_mouse	 sr320@washington.edu
Recent activity... <span style="color: green;">18</span>	qDOD_v9_gene GFF format file of oyster genes ~28k gene	 sr320@washington.edu
Recently viewed »	_qdod_goslim_graphtest	 sr320@washington.edu
Upload dataset	SNP_RNAseqLibrary_SB_BiGill SNP table from RNA-seq library - SB gill tissue pool (BiGill complement) qdod2	 sr320@washington.edu
New query	BiGill_meth_Zhang_exp Gene-centric data including length, CG, percent methylation (gill) and tissue specific RPKM data from Zhang et al a qdod2	 sr320@washington.edu
YOUR TOP VIEWED	qDOD_Cgigas ... 18	 sr320@washington.edu
	BiGo_Larvae_j... 16	 sr320@washington.edu
	TJGR_CCD_d... 11	 sr320@washington.edu
	BiGill_RNAseq... 10	 sr320@washington.edu
	BiGo_lar_T3D5 10	 sr320@washington.edu
POPULAR TAGS	file0	 sr320@washington.edu
	BiGill meth with SP	 sr320@washington.edu
	SPID and GO Numbers Swiss-Prot IDs and corresponding GO numbers qdod	 sr320@washington.edu
	Cgigas_larvae_RNAseq_OsHV_GO	 sr320@washington.edu
	qDOD_Cgigas_GO_GOsli..._DISTINCT	 sr320@washington.edu
	Cgigas Larvae RNA-Seq OsHV UR10 RNA-seq data with descriptions of larvae exposed to OsHV. (>/= 10 UniqueReads) oyster	 sr320@washington.edu
	Cgigas Larvae RNA-Seq OsHV RNA-seq data with descriptions of larvae exposed to OsHV oshv	 sr320@washington.edu
	Zhang_Mgo_gene_RNA-seq_IGV sperm	 sr320@washington.edu
	Zhang_Gil_gene_RNA-seq_IGV IGV format rna-seq	 sr320@washington.edu
	BiGill_methratio_Gene_Genomic_GFF GFF formatted file indicated DNA methylation on oyster genes qdod	 sr320@washington.edu
	TJGR_GeneBased_CDS_GFF GFF format file with exons indicated for genes in oyster genome qdod	 sr320@washington.edu
	BiGill_Gene_Methratio_VD	 sr320@washington.edu
	oyster_v9_mRNA_GFF GFF (gene) from Zhang et al. Column9 modified for Joining qdod	 sr320@washington.edu
	Cgigas_gene_length CDS only	 sr320@washington.edu

Your datasets	
All datasets	<a href="#">View</a>
Shared datasets	<a href="#">View</a>
Recent activity... <span style="color: green;">18</span>	<a href="#">View</a>
Recently viewed »	<a href="#">View</a>
Upload dataset	<a href="#">Upload</a>
New query	<a href="#">New query</a>

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
seaflow	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	Modified
qdod_proteome_blast_mouse	<a href="#">sr320@washington.edu</a>	Jan 2
qDOD_v9_gene GFF format file of oyster genes ~28k gene	<a href="#">sr320@washington.edu</a>	Nov 1
_qdod_goslim_graphtest		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, percent qdod2		Oct 2
qDOD_Cgigas_gene_fasta Tabular format of Cgigas gene sequence qdod2		Oct 2
qDOD_Cgigas_Gene_Descriptions (Swiss-prot) Description and evaluated blast		Oct 2
file0	<a href="#">sr320@washington.edu</a>	Aug 1
BiGill meth with SP	<a href="#">sr320@washington.edu</a>	Aug 1
SPID and GO Numbers Swiss-Prot IDs and corresponding GO numbers qdod	<a href="#">sr320@washington.edu</a>	Aug 1
Cgigas_larvae_RNAseq_OsHV_GO	<a href="#">sr320@washington.edu</a>	Jul 2
qDOD_Cgigas_GO_GOsliM_DISTINCT	<a href="#">sr320@washington.edu</a>	Jul 2
Cgigas Larvae RNA-Seq OsHV UR10 RNA-seq data with descriptions of larvae exposed to OsHV. (>/= 10 UniqueReads) oyster	<a href="#">sr320@washington.edu</a>	Jul 1
Cgigas Larvae RNA-Seq OsHV RNA-seq data with descriptions of larvae exposed to OsHV oshv	<a href="#">sr320@washington.edu</a>	Jul 1
Zhang_Mgo_gene_RNA-seq_IGV sperm	<a href="#">sr320@washington.edu</a>	Jun 2
Zhang_Gil_gene_RNA-seq_IGV IGV format rna-seq	<a href="#">sr320@washington.edu</a>	Jun 2
BiGill_methratio_Gene_Genomic_GFF GFF formatted file indicated DNA methylation on oyster genes qdod	<a href="#">sr320@washington.edu</a>	May
TJGR_GeneBased_CDS_GFF GFF format file with exons indicated for genes in oyster genome qdod	<a href="#">sr320@washington.edu</a>	May
BiGill_Gene_Methratio_VD	<a href="#">sr320@washington.edu</a>	May
oyster_v9_mRNA_GFF GFF (gene) from Zhang et al. Column9 modified for Joining qdod	<a href="#">sr320@washington.edu</a>	May
Cgigas_gene_length CDS only	<a href="#">sr320@washington.edu</a>	May

# Use Cases

- Joining on Annotations
- File Conversion
- Querying Gene Tables

Your datasets
All datasets
Shared datasets
Recent activity... <span style="float: right;">18</span>
Recently viewed »
Upload dataset
New query

## 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 gene	sr320@washington.edu	Nov 1
_qdod_goslim_graphtest	sr320@washington.edu	Oct 2
SNP_RNAseqLibary_SB_BiGill SNP table from RNA-seq library - SB gill tissue pool / BiGill complement	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 al	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 oyster	sr320@washington.edu	Oct 2
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evaluates associations with Cgigas (~28k genes) Derived using Dataset: Genomic data from the Pacific oyster	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.

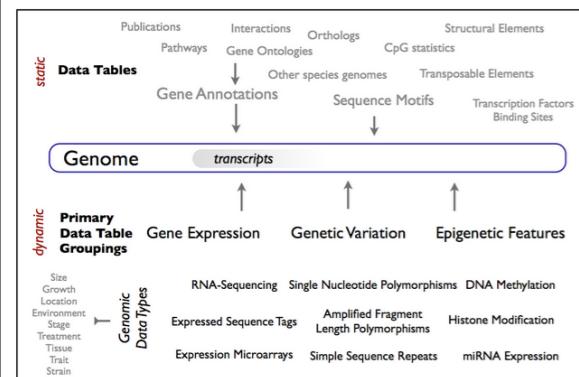
SR discovered that for some proteins, a peptide was sequenced multiple times and so had multiple expression values. From the unique protein associations file in SQLshare, I summed the expression values for all identical peptides.

```
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
```

# 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](#)



Biology

Environment

Molecular

# Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

# Open Science

Data

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.

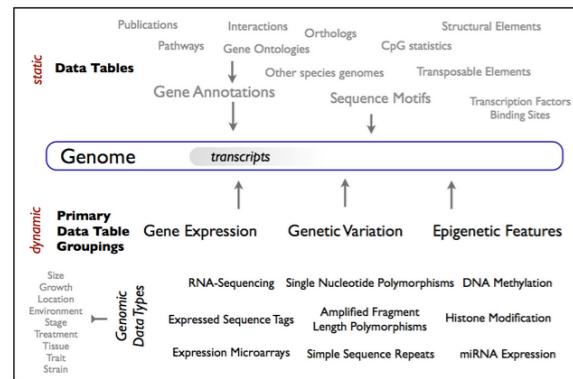
Last edited by sr320, 9 days ago

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>
ETStagseq	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](#)
- [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

Plant Galaxy

Notebooks

Rationale

Platforms

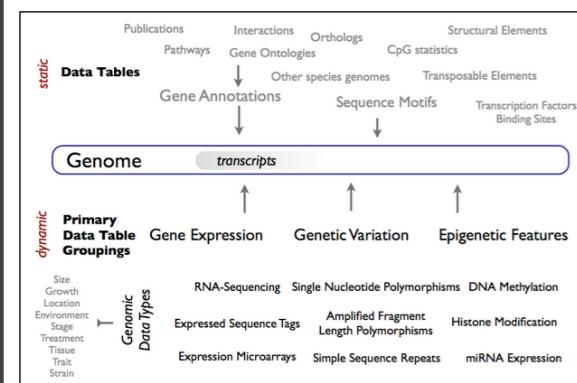
Open Science

Data

## 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</a> ; <a href="#">qual</a>
DH3	SOLiD	RNA	gill	25 x 1	<a href="#">csfasta</a> ; <a href="#">qual</a>

### Data Snapshots

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

### Universal

Dataset	Screenshot	more
<a href="#">UniprotProtNamesReviewed_yes20130610</a>		etc
<a href="#">SPID and GO Numbers</a>		etc

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

Plant Galaxy

ebooks

ationale

platforms

science

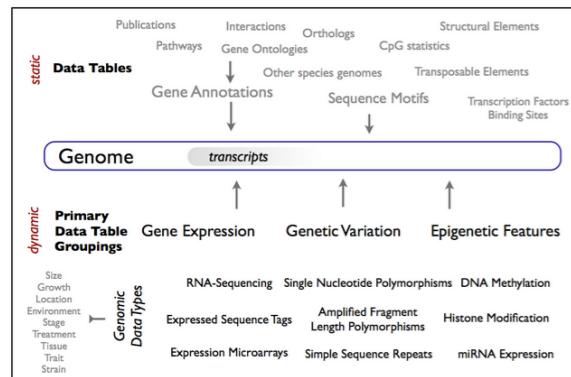
Data

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

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

## 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>

### Data Snapshots

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

### Universal

The screenshot shows the 'Dataset' section for the 'Universal' dataset. It includes a 'Screenshot' preview of a table titled 'UniprotProtNamesReviewed\_yes20130610'. The table has columns: SPID, Status, ProteinName, GenName, Organism, and Length. Below the table is a 'DATASET PREVIEW' section showing rows 1-100 of 28027 columns 4 of 4. The preview table has columns: ID, UniqueReads, TotalReads, and RPKM. The preview table shows data for genes like LDESA, LDHA, etc., with various lengths and expression levels.

### 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 [sr320@washington.edu].[table_solid0078_20091105_RobertsLab_CE_F3 trim]

```

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

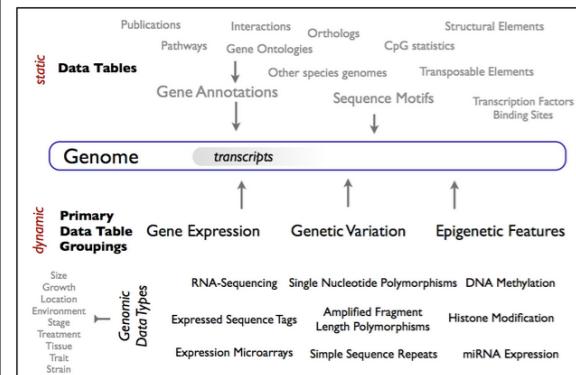
ID	UniqueReads	TotalReads	RPKM
CGI_10000001	0	10	5.23
CGI_10000002	5	5	2.756
CGI_10000003	0	0	0
CGI_10000004	0	0	0
CGI_10000005	0	0	0
CGI_10000006	0	0	0



## 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



eScience Institute Biology Environment Molecular

# Workshop

## Data Snapshots

[Edit Page](#) [Page History](#) [Clone URL](#)

### Universal

The screenshot shows a table titled "Dataset" with a single row labeled "UniprotProtNamesReviewed\_yes20130610". The table includes columns for SPID, Status, ProteinName, GenName, Organism, and Length. The data shows several entries for proteins like LDESA Chor44, LDHA, etc., with their respective lengths and organism information.

## Data Analysis

eScience

Plant Galaxy

ebooks

ationale

platforms

science

Data

g else...

## I: 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

5\_RobertsLab\_GE\_F3 trimmed RNA-Seq.txt Viewable by everyone

Annotation 9 transcriptome

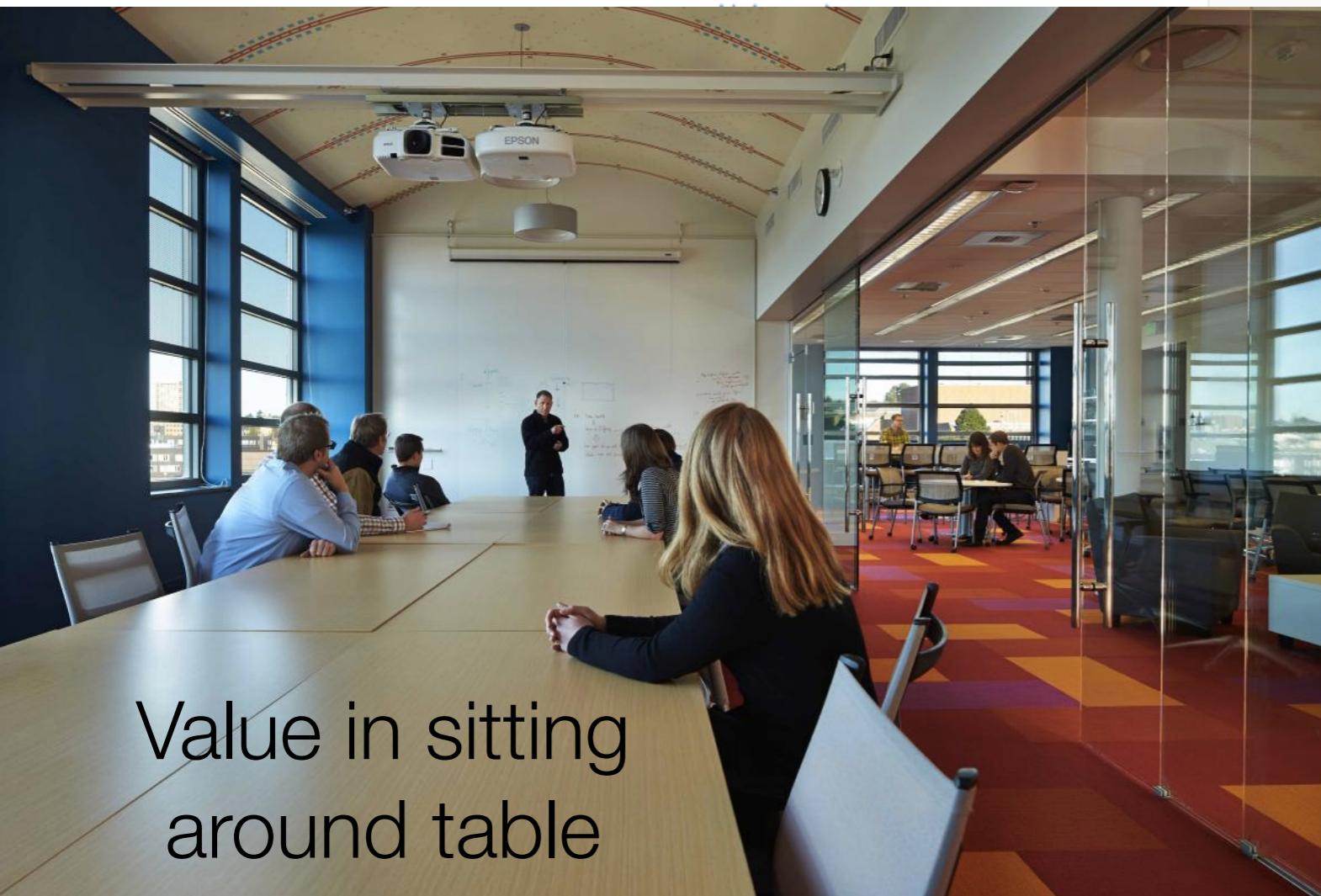
```
as ID,
pads"] as UniqueReads,
pads"] as TotalReads,
4
```

shington.edu.[table\_solid0078\_20091105\_RobertsLab\_CE\_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	



Value in sitting  
around table



iPlant Collaborative™ *Empowering A New Plant Biology*

About Contact Us Login or Register Feedback



## 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...



Biology

Environment

Molecular

Data Analysis

## The iPlant Collaborative

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



Discovery Environment



Data



Apps



Analyses

# Data Analysis

**Discovery Environment**

**Data**

Upload New Refresh Download Edit Share Trash

**Navigation**

- sr320
  - labshare
  - qdod
  - BiGo\_larvae
  - Cgigas\_v9
    - TopHat2-PE\_BiGill\_Bi...
    - Cuffdiff2\_analysis1-2
    - Cuffmerge2\_analysis
    - Cufflinks2\_analysis1
    - TopHat2-SE\_analysis
    - Uncompress\_files\_wi...
    - logs
  - bamout
  - bisque\_data
  - iplant\_ws
  - coge\_data
  - OlyO\_PacBio
  - cacheServiceTempDir
  - OlympiaOyster
  - fish546
  - analyses
  - AltSplice
  - OlyO\_SNPhunt
  - Black\_Abalone
  - ce
- Community Data
- Shared With Me
- Trash

**Details**

Select a file or folder to view its details

**Apps**

Create Use App Request Tool Edit Submit Search Apps

**Categories**

- Workspace (8)
  - Apps under development (6)
  - Favorite Apps (2)
  - My public apps (0)
- Public Apps (434)
  - Archive (27)
  - BEDTools 2.15.0 (22)
  - Beta (56)
  - BLAST (2)
  - Data Sources (2)
  - Experimental (28)
    - Functional Analysis (7)
    - Functional RNA Prediction (5)
  - General Utilities (73)
    - Compress and Decompress
    - Conversion (1)
    - EMBOSS (50)
    - Text and Tabular Data (16)
    - List Analysis (1)
    - Metagenomics (6)
    - microRNA Analysis (3)
    - Molecular Structure (1)
    - Motif Identification (2)
  - NGS (111)
    - Aligners (16)
    - Assemblers (17)
    - Assembly Annotation (9)
    - Bisulfite (3)
    - ChIPseq (4)
    - QC and Processing (21)
    - SAMTools (6)
    - Transcriptome Profiling (11)
      - Misc RNASeq Tools (2)
      - Tuxedo RNA-Seq 1 (4)
      - Tuxedo RNA-Seq 2 (5)
    - Utilities (16)
    - Variant Identification (8)

**Analyses**

Delete Refresh Cancel View Parameters Relaunch

Name	App	Start Date	End Date	Status
genomeCoverageBed_Bi...	genomeCov...	2014 Feb 15 05:11...	2014 Feb 15 05:11...	Completed
genomeCoverageBed_Bi...	genomeCov...	2014 Feb 15 04:14...	2014 Feb 15 04:18...	Completed
RNAseq2bedgraph_anal...	RNAseq2be...	2014 Feb 15 04:07...	2014 Feb 15 09:55...	Completed
RNAseq2bedgraph SE_BB3	RNAseq2be...	2014 Feb 12 14:49...	2014 Feb 12 14:49...	Failed
RNAseq2bedgraph_BiGo...	RNAseq2be...	2014 Feb 12 14:36...	2014 Feb 12 20:00...	Completed
genomeCoverageBed_a...	genomeCov...	2014 Feb 12 11:22...	2014 Feb 12 11:28...	Completed
bamToBed_analysis1	bamToBed	2014 Feb 12 07:50...	2014 Feb 12 07:54...	Completed
VCF to GFF3_analysis1	VCF to GFF3	2014 Feb 11 06:58...	2014 Feb 11 06:59...	Completed
Find SNPs - mpileup_analysis1	Find SNPs - mpileup	2014 Feb 11 04:13...	2014 Feb 11 06:37...	Completed
TopHat2-PE_analysis1	TopHat2-PE	2014 Feb 10 17:23...	2014 Feb 10 21:38...	Completed

# Data Analysis

**Discovery Environment**

**Data**

- [Upload](#) [New](#) [Refresh](#) [Download](#) [Edit](#) [Share](#) [Trash](#)

**Navigation**

- sr320
  - labshare
  - qdod
  - BiGo\_larvae
  - Cgigas\_v9
    - TopHat2-PE\_BiGill\_Bi
    - Cuffdiff2\_analysis1-2
    - Cuffmerge2\_analysis
    - Cufflinks2\_analysis1-
    - TopHat2-SE\_analysis
      - bam
      - logs
    - U1\_out
    - T2\_out
    - T1\_out
    - U2\_out
    - Uncompress\_files\_wi
    - logs
  - bamout
  - bisque\_data
  - iplant\_ws
  - coge\_data
  - OlyO\_PacBio

**Apps**

- [Create](#) [Use App](#) [Request Tool](#) [Edit](#) [Submit](#) [Search Apps](#)

**Categories**

- Workspace (8)
  - Apps under development (6)
  - Favorite Apps (2)
  - My public apps (0)
  - Public Apps (434)
  - Archive (27)

**RNAseq2bedgraph**

Analysis Name: RNAseq2bedgraph\_analysis1

TopHat2-PE for workflows - README

\* TopHat2-PE for workflows - Input data

TopHat2-PE for workflows - Reference Genome (Mandatory)

Select a reference genome from the list or select your own reference genome file. Note one of these two options MUST be selected.

Select a reference genome from the list:

Choose item from list.

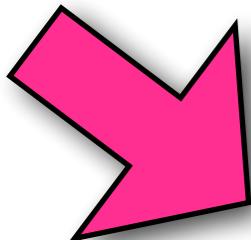
If your species is not in the pull-down menu, try 'Community Data' -> iplant\_training->reference\_genomes. It contains a larger collection. You may also provide your own reference genome in FASTA format.

Provide a reference genome file in FASTA format:

Select a file  [Browse](#)

TopHat2-PE for workflows - Reference Annotations

\* TopHat2-PE for workflows - Analysis Options



**SQLSHARE**

**Galaxy**

 **iPlant  
Collaborative™**

**hyak**



Biology

Environment

Molecular

**Data Analysis**

eScience

iPlant Galaxy

**Notebooks**

Rationale

Platforms

**Open Science**

Data

everything else...

Biology

Environment

Molecular

## Data Analysis

eScience

iPlant Galaxy

## Notebooks

Rationale

Platforms

## Open Science

Data

everything else...



SQLSHARE

Galaxy

iPlant  
Collaborative™

IGV  
Trinity

fastqc

bash (lots)

python APIs

excel

DAVID

Revigo



BLAST

Tophat

DESeq

R

BSMAP

perl scripts

python modules

STACKS

Bedtools

# Jupyter Notebooks (IPython)

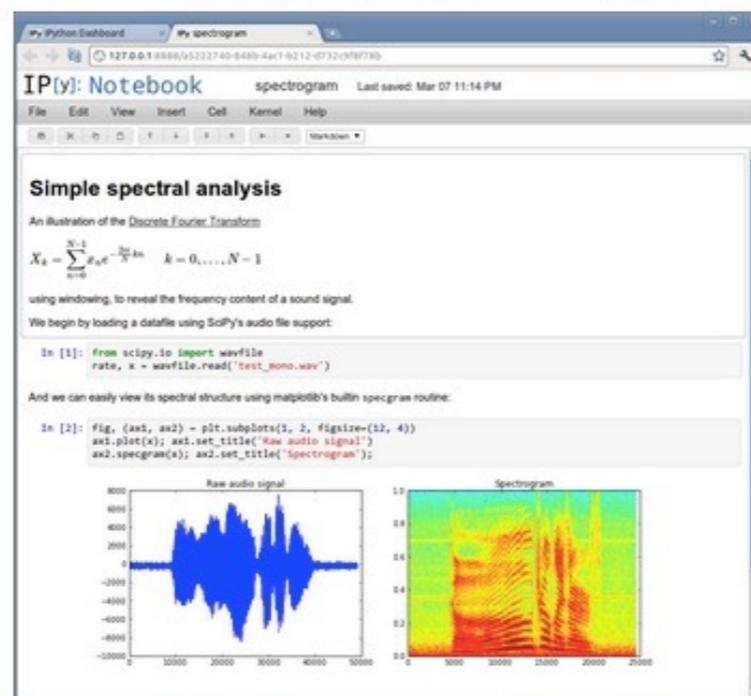
---

**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
ATTACAATACGAAGTAAAATAGATAACGTAAAAATAATCTTGGTGCTGGATGATCGATC
AAGTTCACCAATATTTATTGTAAAAAAATCATTCTAACAGCATGAAATCGTGTACAATG
TATAAACAAAGCAAATATATAACACTAAAGCAAGAGGGCGTAAGTGGGGGGGTGGGTGAGA
GTAAAAAAATTCAAACATGTCAAATACCCGGCGTTAGCCTTAAAGCACCATGGACTTCT
CCCTTCAGTAAAGCTTAAAGCTTAAAGCTTAAAGCTTAAAGCTTAAAGCTTAAAGCTTAAAGCT
```

File Edit View Insert Cell Kernel Help



Run

Run and Select Below

Run and Insert Below

Run All

Run All Above

Run All Below

Cell Toolbar: None

## Fasta2Slim

This IPython notebook is personal preferences. As

- IPython
- NCBI Blast
- SQLShare Python CLI

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

Cell Type

Current Output

All Output

### Instructions for use.

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

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
ATTACAATACGAAGTAAAATAGATAACGTAAAAATAATCTTGGTGCTGGATGATCGATC
AAGTTCACCAATATTTATTGTAAAAAATCATTCTAACAGCATGAAATCGTGTACAATG
TATAAACAAAGCAAATATATAACACTAAAGCAAGAGGGCGTAAGTGGGGGGTGGGTGAGA
GTAAAAAAATTCAAACATGTCAAATACCCCGGCCTAGCCTTAAAAGCACCATGGACTTCT
GCCTTCAATAAGCATAAAATTAAAACACCTAATACACAATGAATATACAGATAAAACAGA
TTTATGAATAGTTGGTGTACATCTTTACAGCCATAAGCCTTCATTGCTTCCAAACG
TATAAAATCTGACTTGGAACAAATATACAGCCATGAGATATGACACAGCGAGCACTACAAT
ATATATTATCTTGTACTATACAGCCTGTACAAGAAAATTCTGGAATTGTCTTCACAAGA
GACAGAAAAATAGTTGCAATGTGAATGCTAGTCTACTATTGATCACAATTGGATAGAAA
```

```
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}].[_blast_sprot]md left join [samwhite@washington.edu].[UniprotProtNamesReviewed_yes20130610]sp on md.Column3=sp.SPID left join [sr320@washington.edu].[SPID and GO Numbers]go on md.Column3=go.SPID left join [sr320@washington.edu].[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

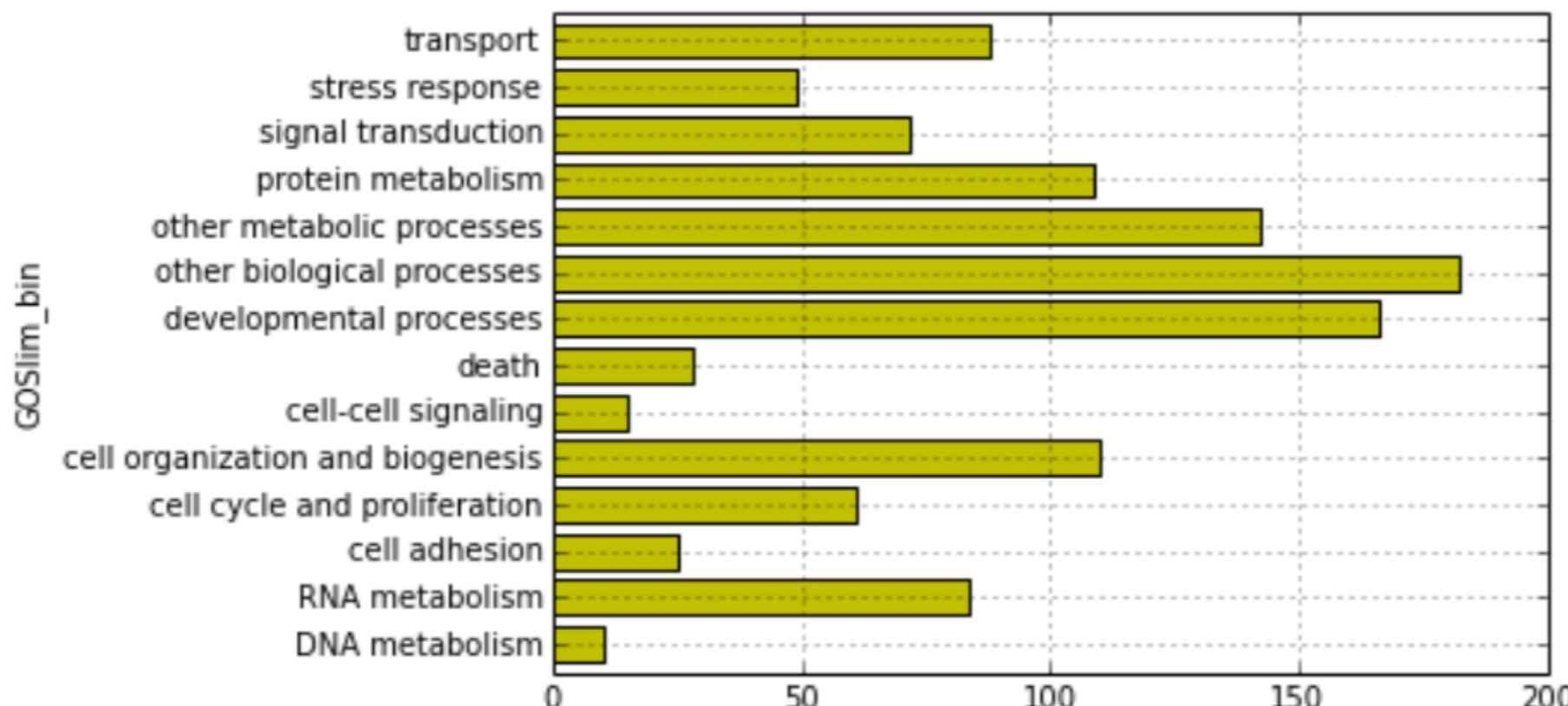
```
In [11]: cd {wd}
```

```
/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>