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

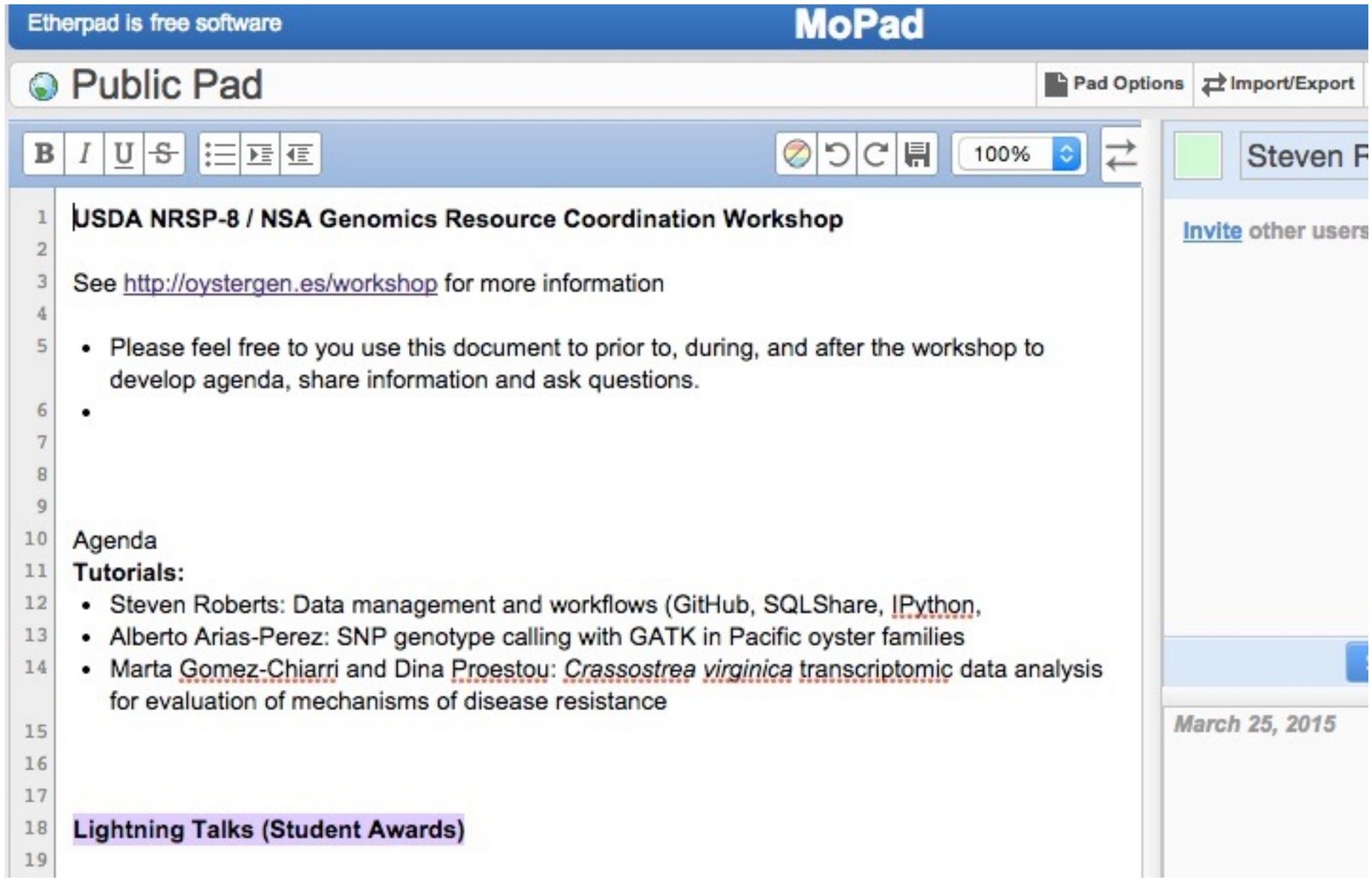
Rationale

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



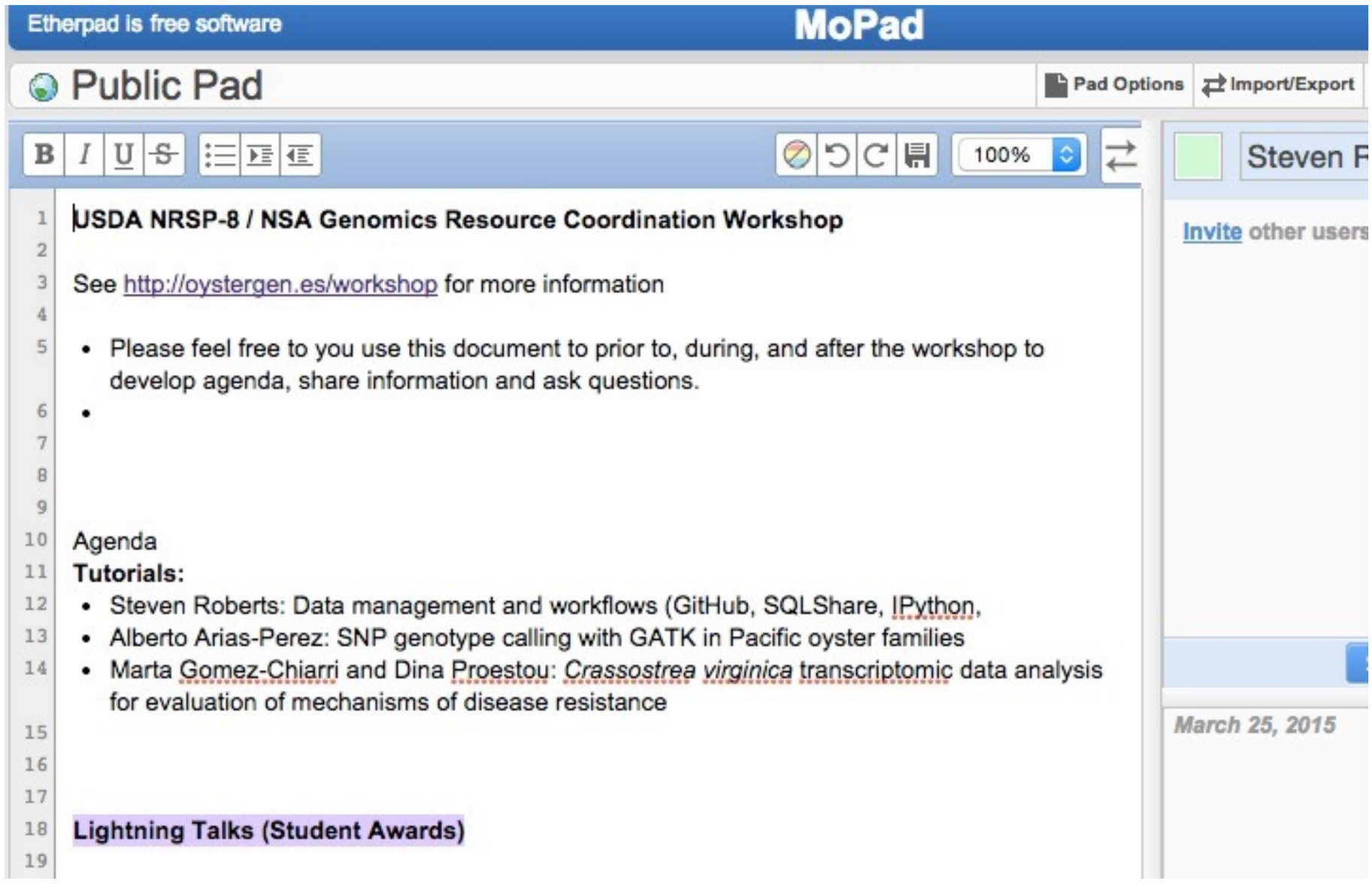
Today

http://goo.gl/xqf8Og

Tutorials

Lightning Talks

Discussion



Today

http://goo.gl/xqf8Og

Crassostrea virginica sequencing

Tutorials

Breeding program

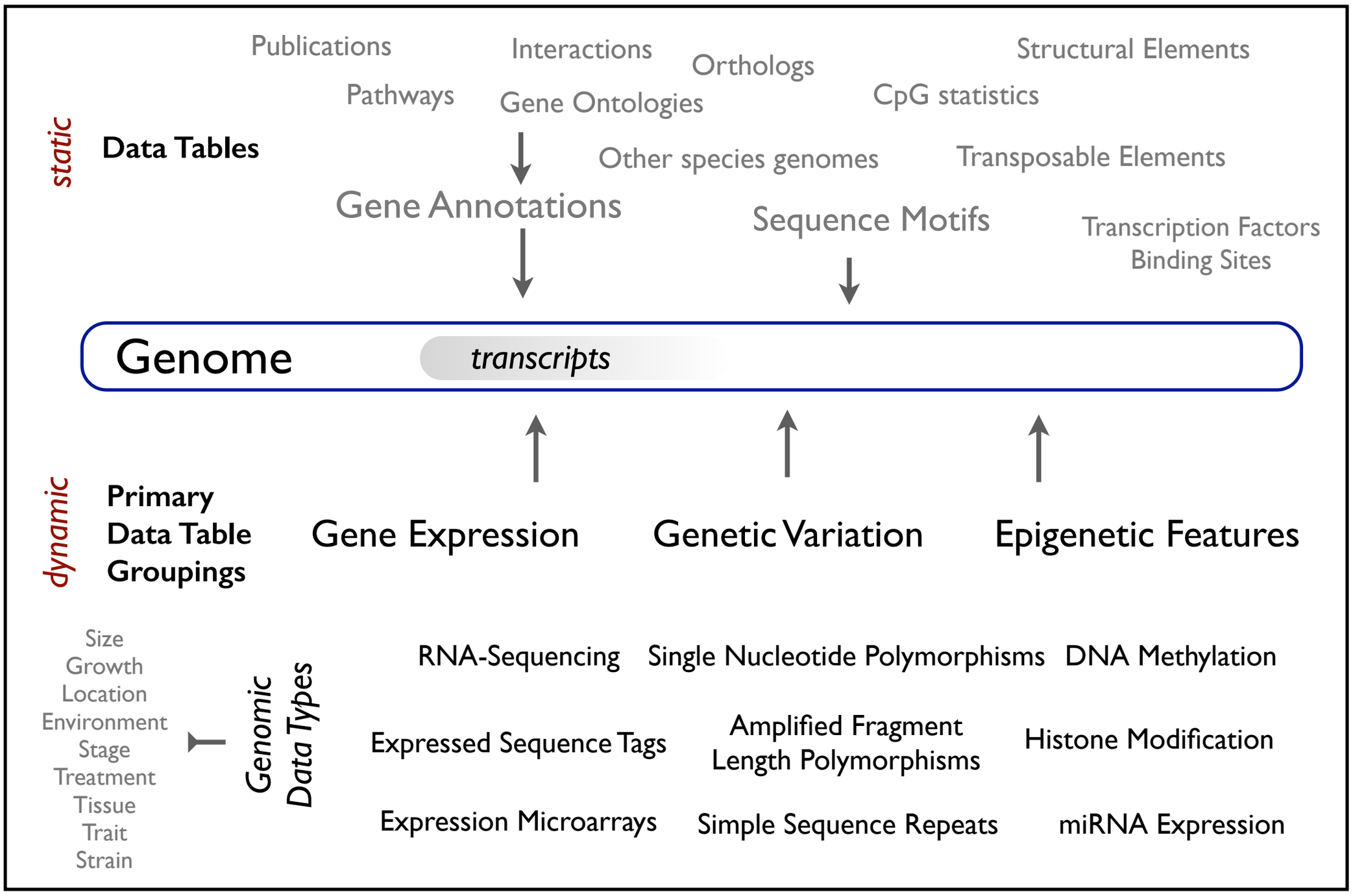
Lightning Talks

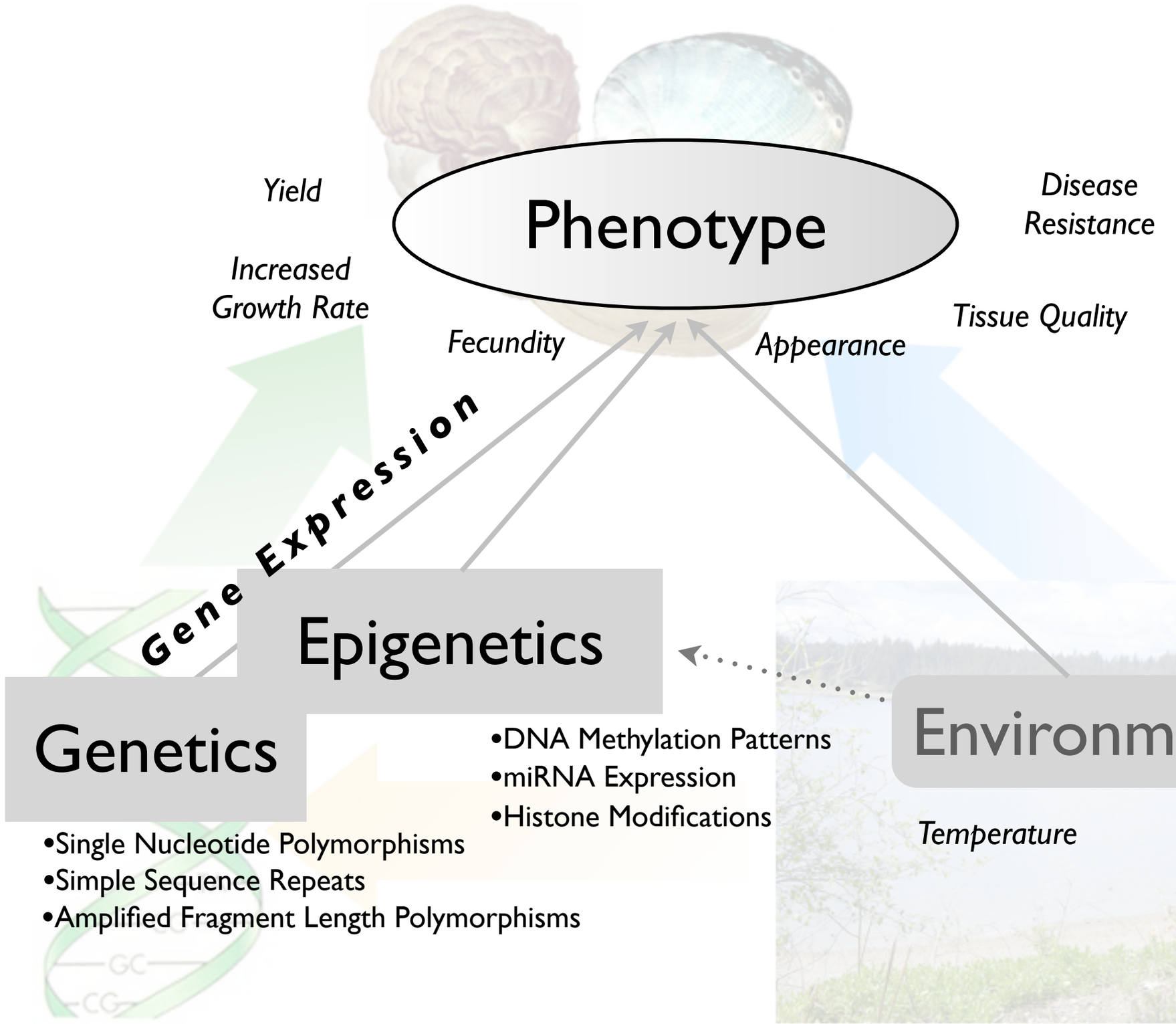
Communication Platform

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

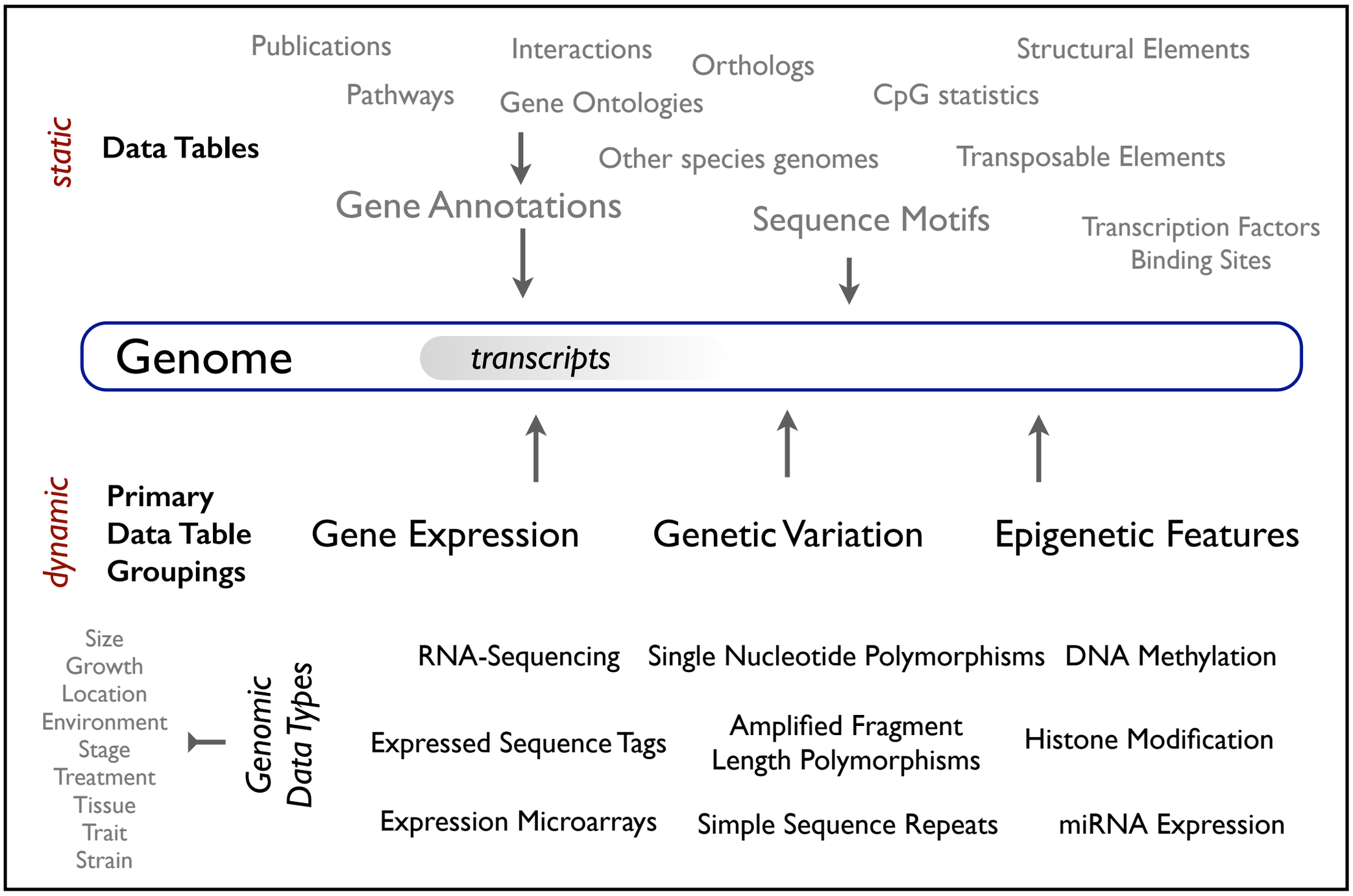
Rationale

Platforms

Open Science

Data

everything else...



OYSTERGEN.ES

PACIFIC OLYMPIA

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.

Biology

Environment

Molecular

Data Analysis

eScience

jPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

everything else...

DATA

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.	http://goo.gl/oqCxl	https://sqlshare.esc
qDOD_Zhang_Gil_gene_RNA-seq	Gill RNA-seq data (gene based)	http://goo.gl/8oISR	https://sqlshare.esc seq
qDOD_Zhang_Mgo_gene_RNA-seq	Male Gonad RNA-seq data (gene based)	http://goo.gl/6buVz	https://sqlshare.esc seq

OYSTERGEN.ES /query

PACIFIC OLYMPIA

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

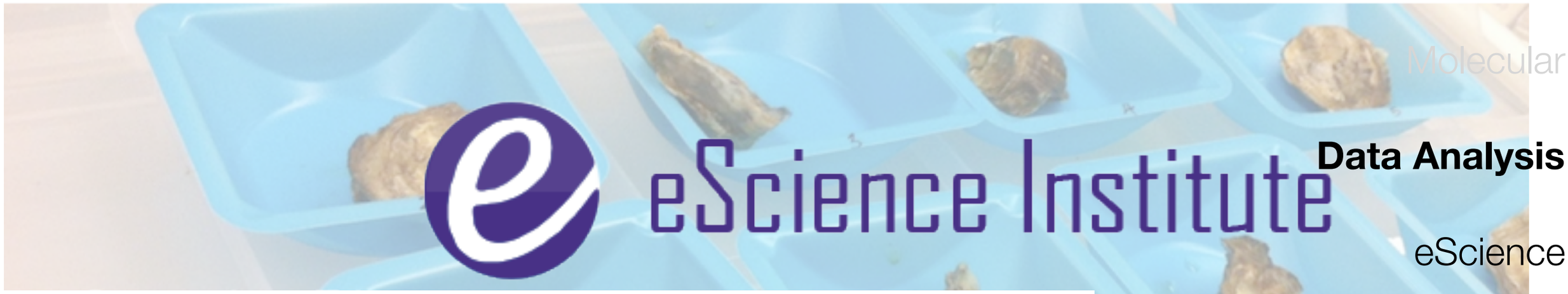
Rationale

Platforms

Open Science

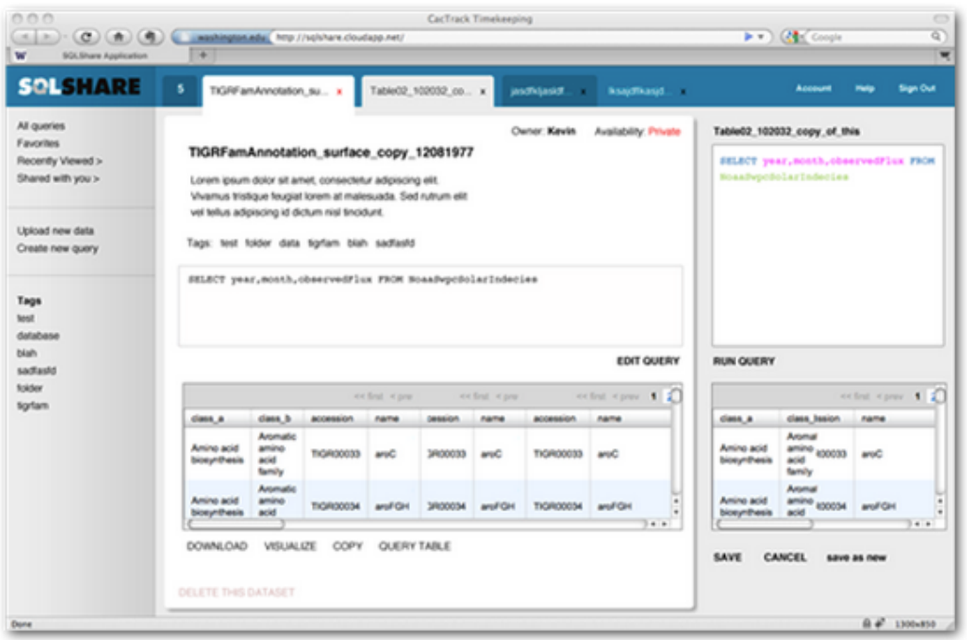
Data

everything else…



SQLSHARE

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



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.

SQLSHARE

Logged in:

- 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
- 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: qdod

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_graphtest	sr320@washington.edu	Oct 2
SNP_RNAseqLibary_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 oyste qdod2	sr320@washington.edu	Oct 2
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evalues associated with Cgigas 28k genes Derived using Dataset: Genomic blast	sr320@washington.edu	Oct 2
file0	sr320@washington.edu	Aug
BiGill meth with SP	sr320@washington.edu	Aug
SPID and GO Numbers Swiss-Prot IDs and corresponding GO numbers qdod	sr320@washington.edu	Aug
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 oshv	sr320@washington.edu	Jul 1
Zhang_Mgo_gene_RNA-seq_IGV sperm	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	May

SQLSHARE

Logged in:

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
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: qdod

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_graphtest		Oct 2
SNP_RNAseqLibary_SB_BiGill SNP table from RNA-seq library - SB qdod2		Oct 2
BiGill_meth_Zhang_exp Gene-centric data including length, CG, perc qdod2		Oct 2
qDOD_Cgigas_gene_fasta Tabular format of Cgigas gene sequence qdod2		Oct 2
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evalue blast		Oct 2
file0	sr320@washington.edu	Aug 3
BiGill meth with SP	sr320@washington.edu	Aug 3
SPID and GO Numbers Swiss-Prot IDs and corresponding GO numbers qdod	sr320@washington.edu	Aug 3
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 oshv	sr320@washington.edu	Jul 1
Zhang_Mgo_gene_RNA-seq_IGV sperm	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	May

Use Cases

- **Joining on Annotations**
- **File Conversion**
- **Querying Gene Tables**

SQLSHARE

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

Filter dataset by keyword: qdod

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_graphtest	sr320@washington.edu	Oct 2
SNP_RNAseqLibary_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 oyste qdod2	sr320@washington.edu	Oct 2
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evalues Associations (28K) 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.

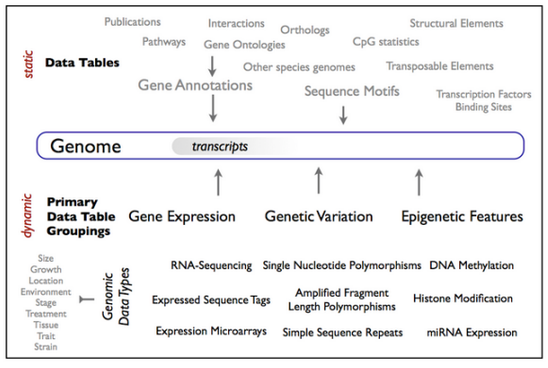
SR discovered that for some proteins, a peptide was sequenced multiple times and so had multiple expression values. From the unique protein associations file in SQLshare, I summed the expression values for all identical peptides.

```
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
```

qdod: Querying Disparate Oyster Datasets

This respository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase effeciency 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

USDA

github.com/sr320/qdod/wiki

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

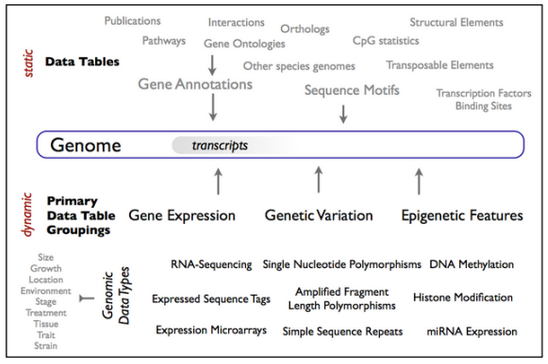
Data

everything else...

qdod: Querying Disparate Oyster Datasets

This respository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase effeciency 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 question, report problems, or suggest features.

Last edited by sr320, 9 days ago

Select Genomic Data

ID	Platform	Molecule	Tissue	Length	Files
BB3	SOLiD	RNA	gill	25 x 1	csfasta; qual
DH3	SOLiD	RNA	gill	25 x 1	csfasta; qual
DH2	SOLiD	RNA	gill	25 x 1	csfasta; qual
GE	SOLiD	RNA	larvae	50 x 1	csfasta; qual
GC	SOLiD	RNA	larvae	50 x 1	csfasta; qual
SBunmeth	SOLiD	DNA	gill	25 x 1	csfasta; qual
SBmeth	SOLiD	DNA	gill	25 x 1	csfasta; qual
BSseqGill	Illumina	DNA	gill	36 x 1	fastq
ETStagseq	Illumina	RNA	gill		zip
BSseqSperm	Illumina	DNA	sperm	72 x 2	fastq1; fastq2
BiGillRNA	Illumina	RNA	gill	50 x 2	fastq1; fastq2
BiGoRNA	Illumina	RNA	sperm	50 x 2	fastq1; fastq2

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

everything else...

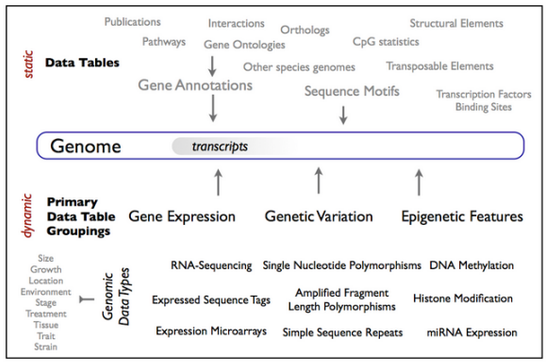
USDA

github.com/sr320/qdod/wiki

qdod: Querying Disparate Oyster Datasets

This respository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase effeciency 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 question, report problems, or suggest features.

Last edited by sr320, 9 days ago

Select Genomic Data

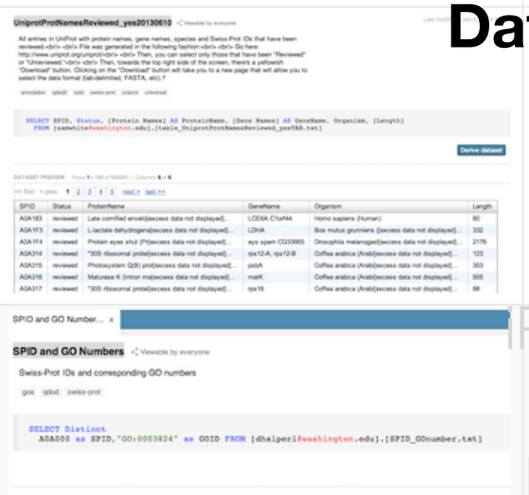
ID	Platform	Molecule	Tissue	Length	Files
BB3	SOLiD	RNA	gill	25 x 1	csfasta; qual
DH3	SOLiD	RNA	gill	25 x 1	csfasta; qual

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 Numbers		etc

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

everything else...

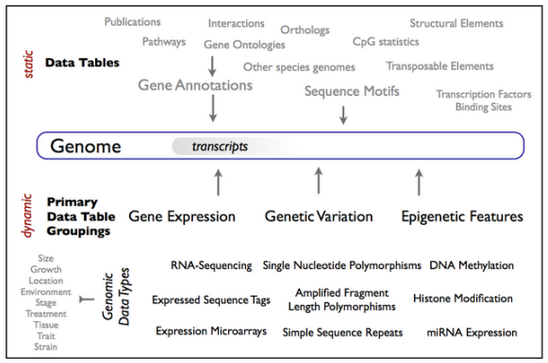
USDA

github.com/sr320/qdod/wiki

qdod: Querying Disparate Oyster Datasets

This respository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase effeciency 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

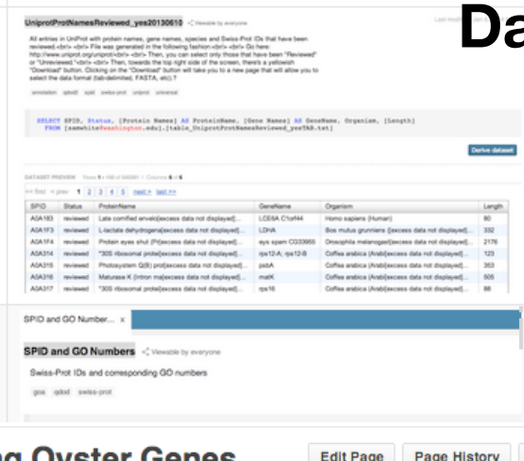
ID	Platform	Molecule	Tissue	Length	Files
BB3	SOLiD	RNA	gill	25 x 1	csfasta; qual
DH3	SOLiD	RNA	gill	25 x 1	csfasta; qual

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

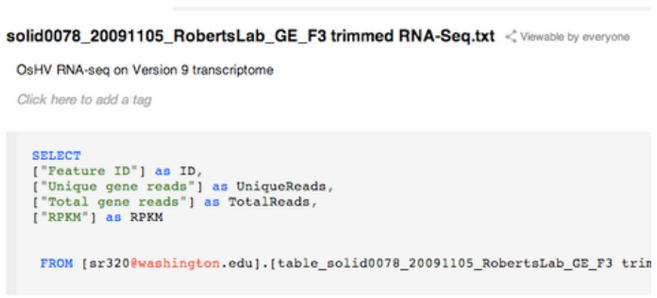
Workflow 1: Annotating Oyster Genes

Edit Page | Page History | Clone URL

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



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

<< first < prev 1 2 3 4 5 next > last >>

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

Biology
Environment
Molecular
Data Analysis
eScience
iPlant Galaxy
ebooks
ationale
latforms
Science
Data
g else...

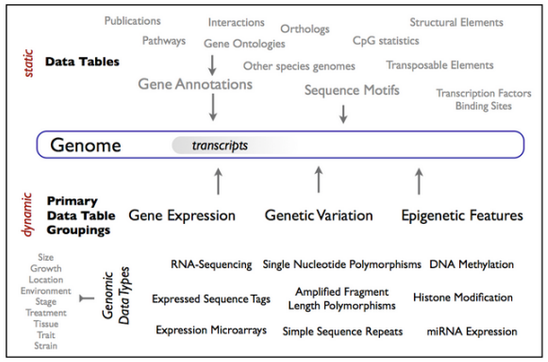
USDA

github.com/sr320/qdod/wiki

qdod: Querying Disparate Oyster Datasets

This respository provides access to genomic data and workflows (IPython notebooks) that are being integrated as part of effort to increase effeciency 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 Workshop

Currently the documentation is focused on

A. Raw Data

- Select NGS Data via Roberts Lab

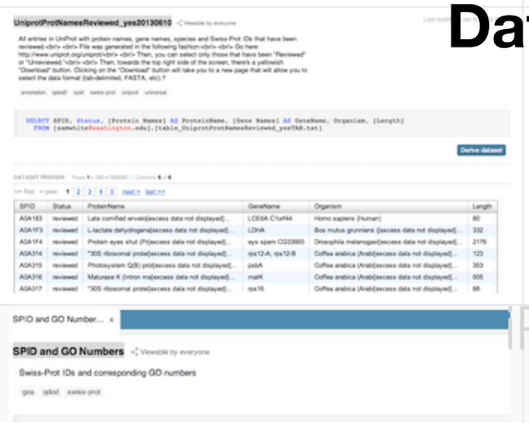
B. Datasets in SQLShare

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

: Annotating Oyster Genes

Edit Page | Page History | Clone URL

ocus 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.

le: Oyster larvae RNA-seq - OsHV exposure

5_RobertsLab_GE_F3 trimmed RNA-Seq.txt Viewable by everyone

ion 9 transcriptome

```
s ID,
ads"] as UniqueReads,
ds"] as TotalReads,
shington.edu].[table_solid0078_20091105_RobertsLab_GE_F3 trim
```

1 - 100 of 28027 | Columns 4 of 4

3 4 5 next > last >>

ueReads	TotalReads	RPKM
	10	5.23
	5	2.756
	0	0
	0	0
	0	0

Value in sitting around table

Biology
Environment
Molecular
Data Analysis
eScience
iPlant Galaxy
ebooks
ationale
latforms
Science
Data
g else...

Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

everything else...

Biology

Environment

Molecular

Data Analysis

Discovery Environment

Data

Apps

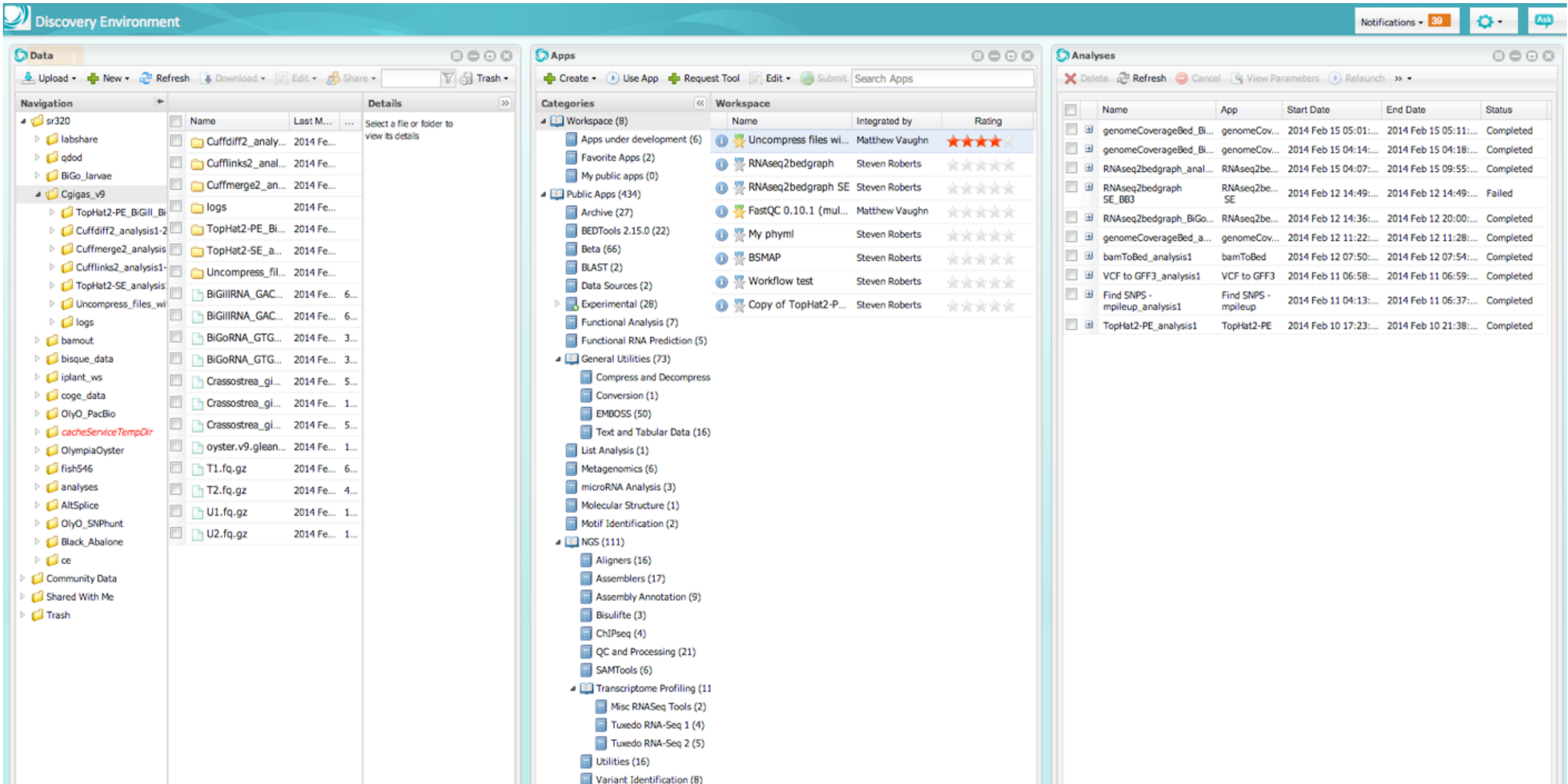
Analyses

Biology

Environment

Molecular

Data Analysis



Biology

Environment

Molecular

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

Name	Last M...	...
condor-0-input...	2014 Fe...	0...
condor-0-input...	2014 Fe...	3...
condor-0-input...	2014 Fe...	0...
condor-0-input...	2014 Fe...	3...
condor-0-input...	2014 Fe...	0...
condor-0-input...	2014 Fe...	3...
condor-0-input...	2014 Fe...	0...
condor-0-input...	2014 Fe...	3...
condor-0-input...	2014 Fe...	0...
condor-0-input...	2014 Fe...	3...
condor-0-input...	2014 Fe...	0...
condor-0-input...	2014 Fe...	3...
condor-stderr-0	2014 Fe...	2...
condor-stdout-0	2014 Fe...	1...
iplant.cmd	2014 Fe...	6...
iplant.sh	2014 Fe...	5...
output-last-stderr	2014 Fe...	0...
output-last-std...	2014 Fe...	6...
script-condor-log	2014 Fe...	3...

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)

Workspace

Name	Integrated by	Rating
Uncompress files wi...	Matthew Vaughn	
RNAseq2bedgraph	Steven Roberts	
RNAseq2bedgraph SE	Steven Roberts	
FastQC 0.10.1 (mul...	Matthew Vaughn	

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:

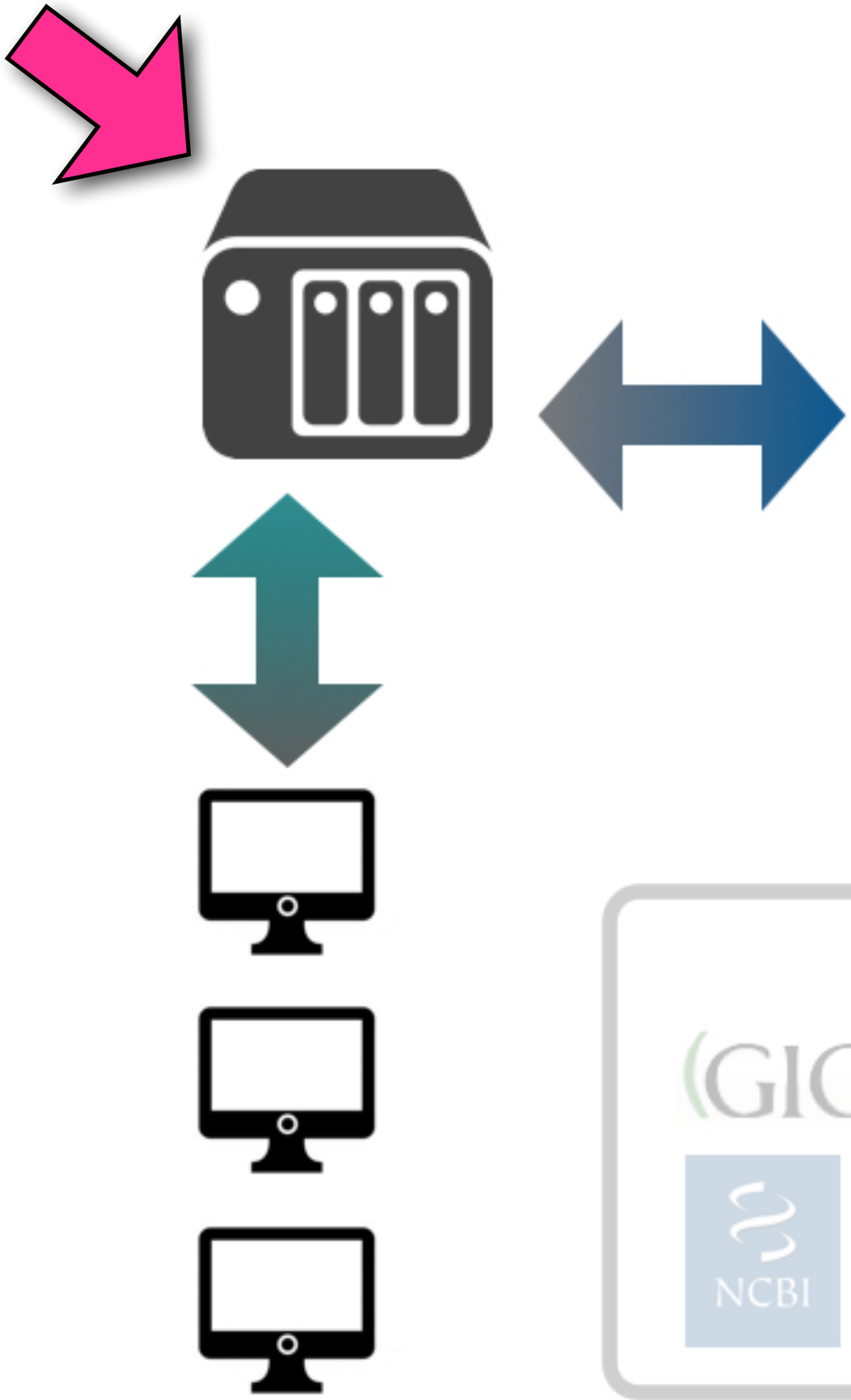
Choose item from list.

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



Biology

Environment

Molecular

Data Analysis

eScience

iPlant Galaxy

Notebooks

Rationale

Platforms

Open Science

Data

everything else…

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...

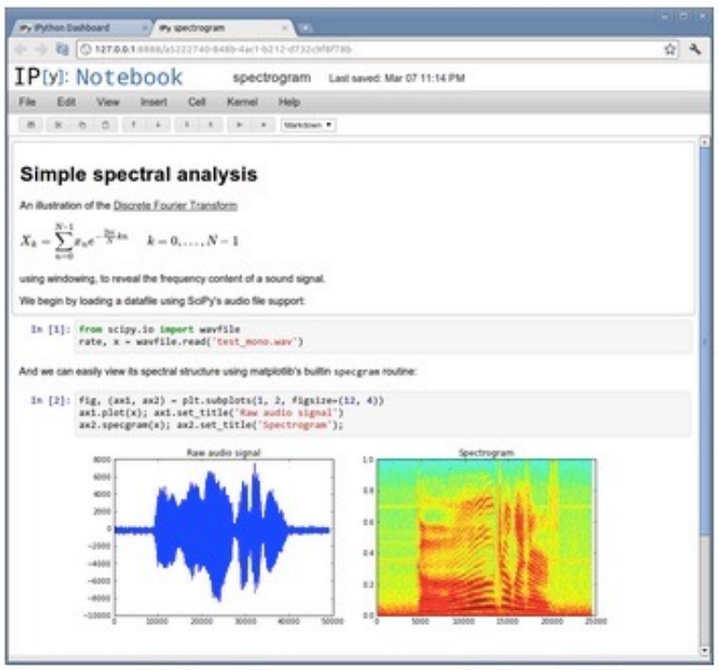
Jupyter Notebooks (IPython)

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
AAGTTCACCAATATTTTATTGTAAAAAATCATTCTAAACAGCATGAAATCGTGTACAATG
TATAAACAAGCAAATATATAACACTAAAGCAAGAGGGCGTAAGTGGGGGGGTGGGTGAGA
GTAAAAAATTCAAACATGTCAAATACCCCGGCGTTAGCCTTAAAAGCACCATGGACTTCT
```

IP[y]: Notebook fasta2slim Last Checkpoint: Jan 05 15:31 (autosaved)

File Edit View Insert Cell Kernel Help

Cell Toolbar: None

- Run
- Run and Select Below
- Run and Insert Below
- Run All
- Run All Above
- Run All Below
- Cell Type
- Current Output
- All Output

Fasta2Slim

This IPython notebook is ructured means to annotate sequences using UniProt/SwissPr
personal preferences. As equires the user has the following software installed ...

- IPython
- NCBI Blast
- SQLShare Python Cl

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
ATTTACAATACGAAGTAAAATAGATAACGTGAAAATAATCTTGGTGCTGGATGATCGATC
AAGTTCACCAATATTTTATTGTAAAAAATCATTCTAAACAGCATGAAATCGTGTACAATG
TATAAACAAGCAAATATATAACACTAAAGCAAGAGGGCGTAAGTGGGGGGGTGGGTGAGA
GTAAAAAATTCAAACATGTCAAATACCCCGGCGTTAGCCTTAAAAGCACCATGGACTTCT
GCCTTCAATAAGCATAAAATTAAAACACCTAATACACAATGAATATACAGATAAAACAGA
TTTATGAATAGTTGGTGTTACATCTTTTACAGCCATAAGCCTTCATTTTGCTTCCAAACG
TATAAAATCTGACTTGGAACAATATACAGCCATGAGATATGACACAGCGAGCACTACAAT
ATATATTTATCTTGTACTATACAGCCTGTACAAGAAAATTCTGGAATTGTCTTCACAAGA
GACAGAAAAATAGTTGCAATGTGAATGCTAGTCTACTATTTGATCACAATTGGATAGAAA
```

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
```

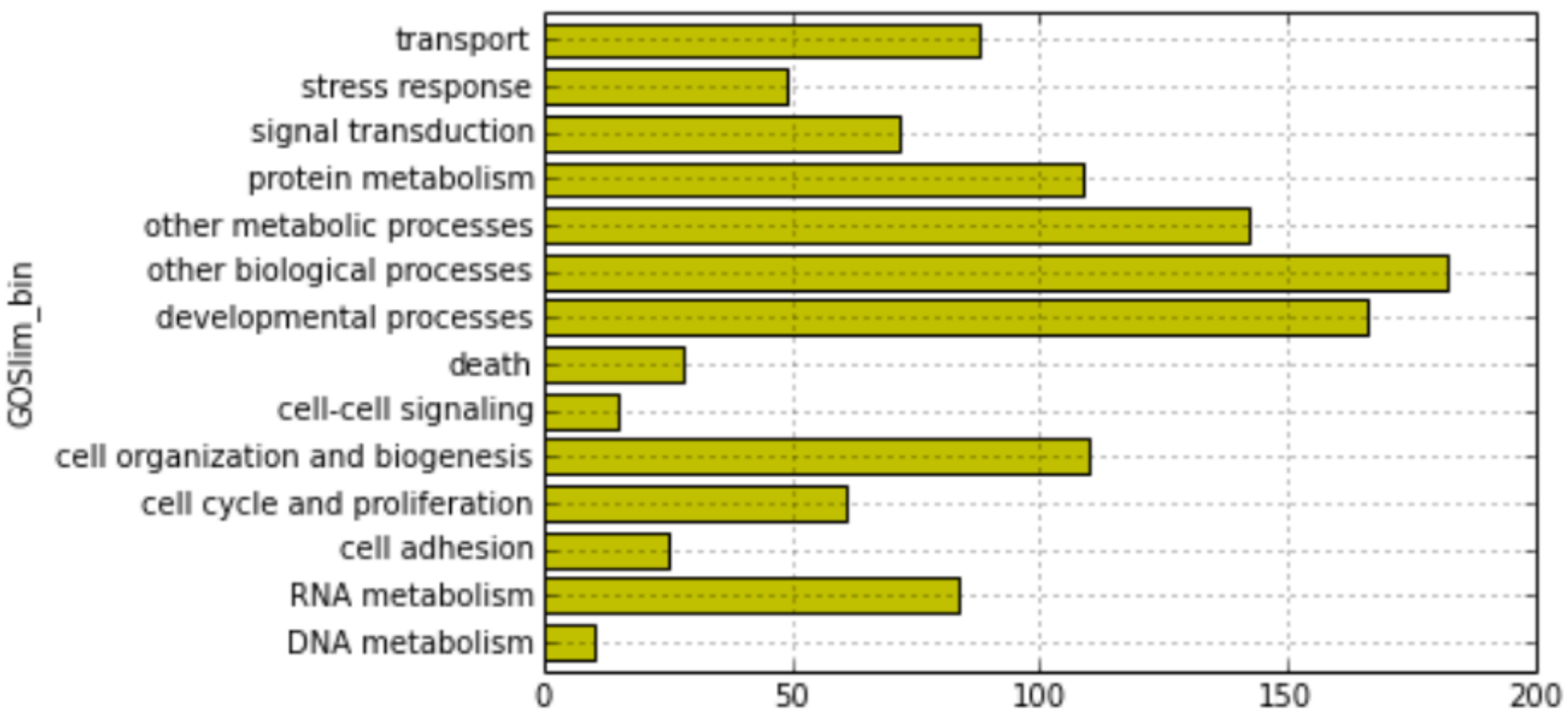
In [11]:
```
cd {wd}
```

```
/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