Collaborative Genomic Data Analyses in the Cloud

Steven B. Roberts
Associate Professor
School of Aquatic and Fishery Sciences
University of Washington

robertslab.info

Open Science

- You are free to Share!

cc BY SA

- Our lab practices open notebook science

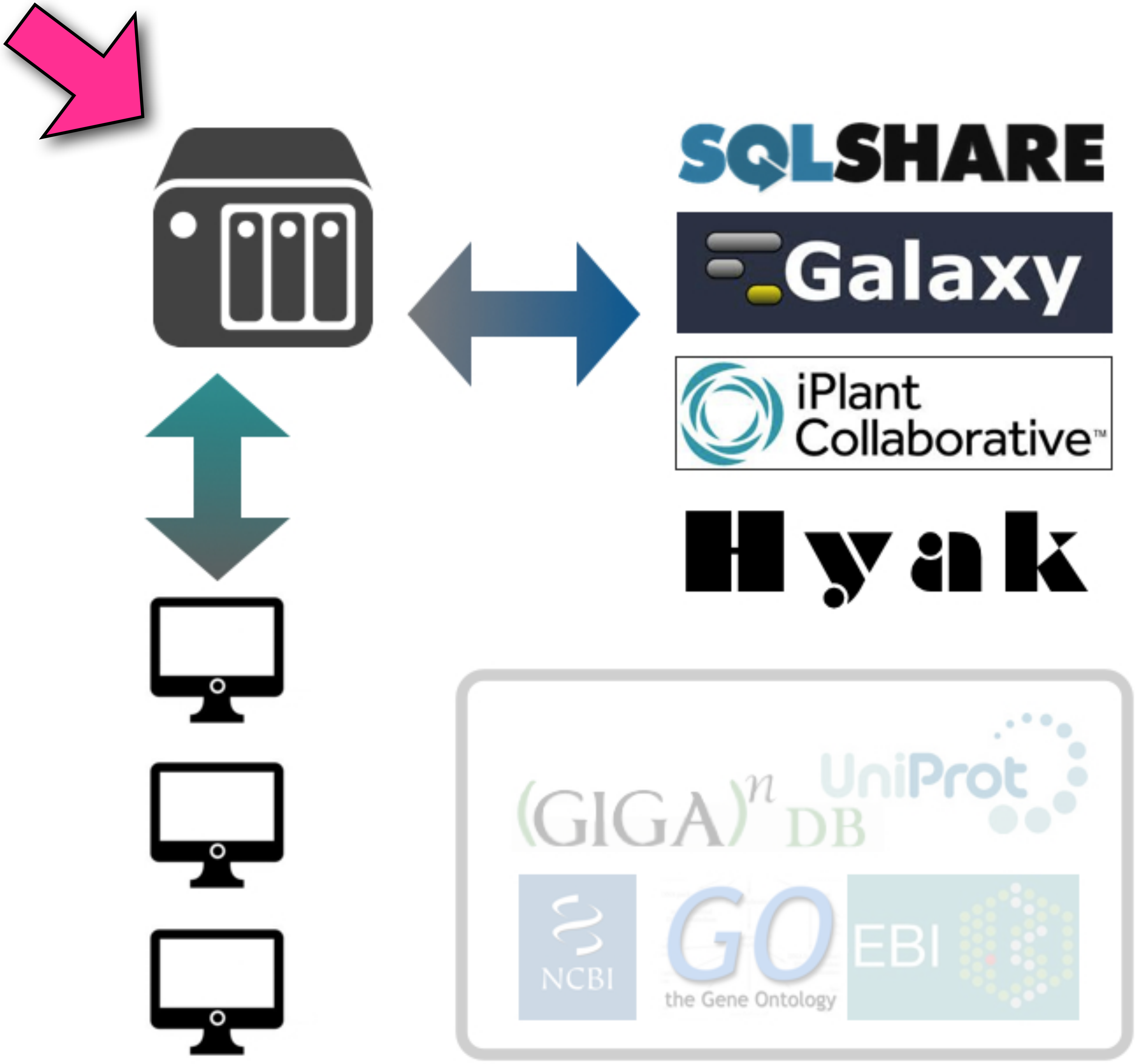
wikispaces
wikis for everyone

IP[y]: IPython
Interactive Computing

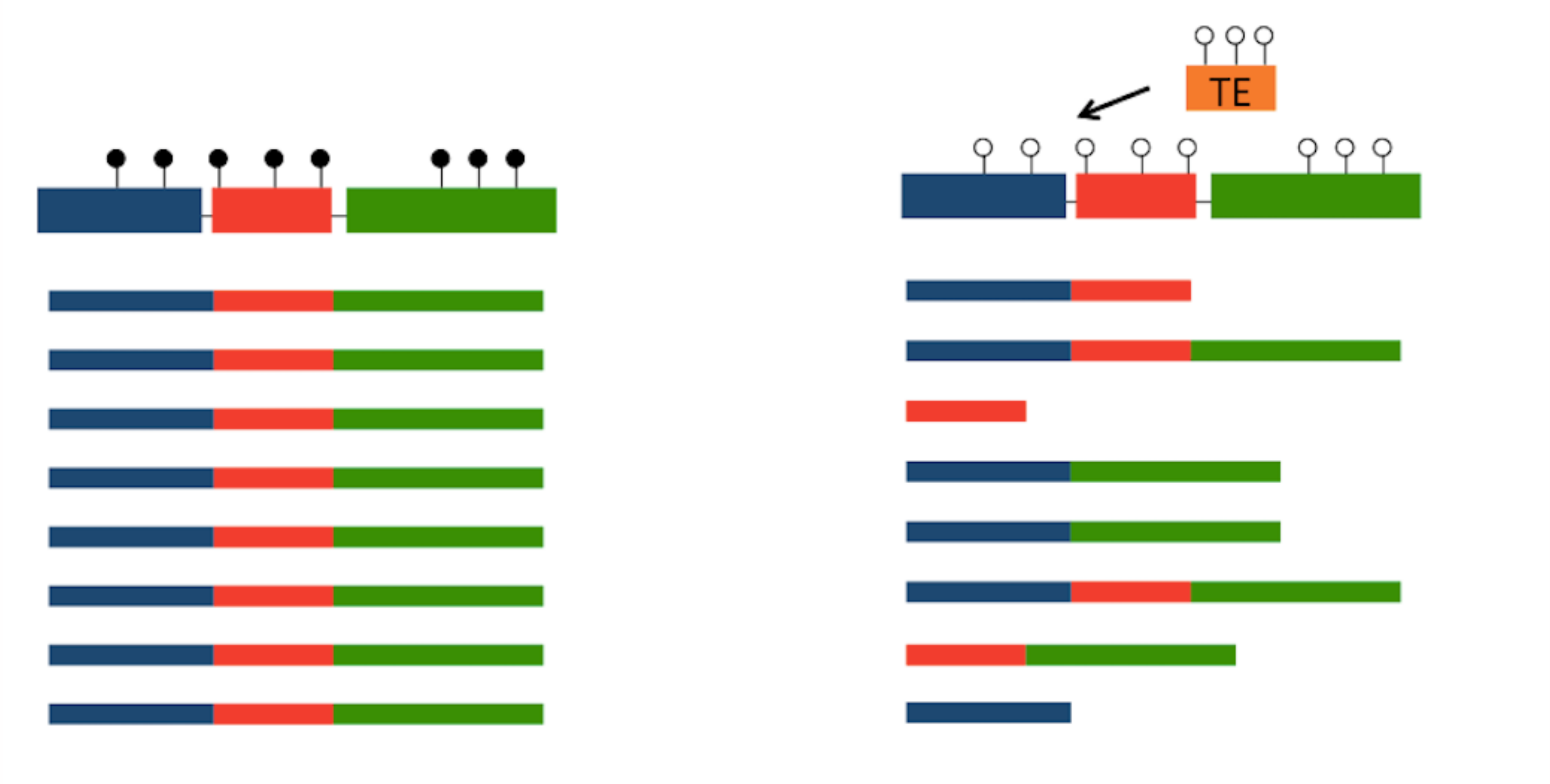
EVERNOTE

robertslab.info

sr320@uw.edu


SQLSHARE
Galaxy
iPlant Collaborative™
Hyak
(GIGA)^n DB
UniProt
NCBI
GO
the Gene Ontology
EBI

Stochastic Variation

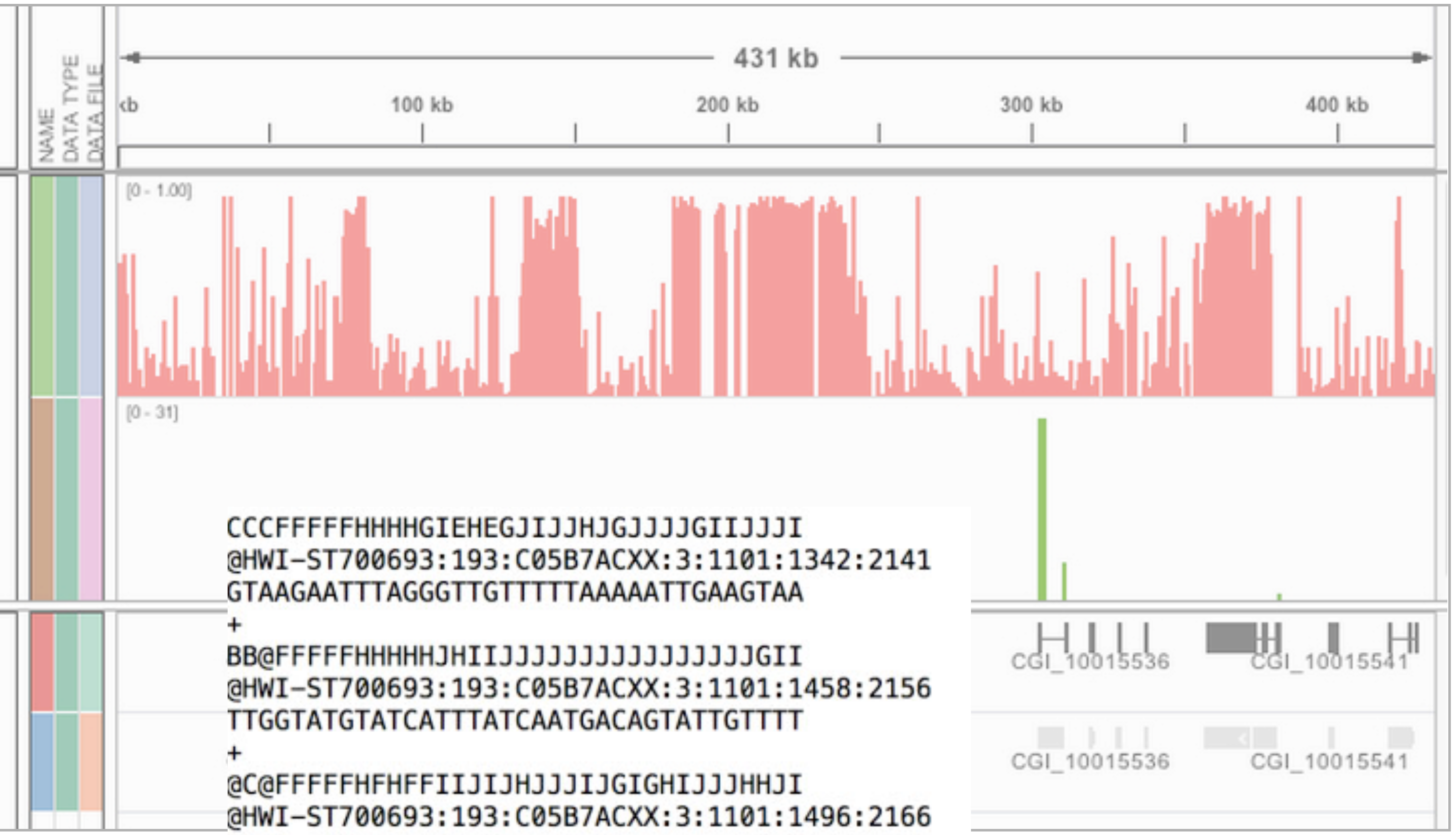


A context dependent role for DNA methylation in bivalves

Mackenzie R. Gavery and Steven B. Roberts

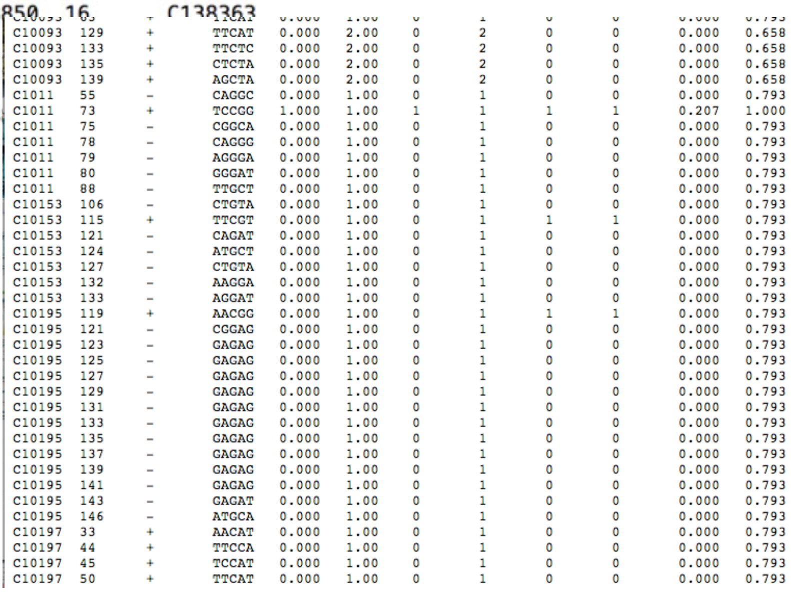
10.1093/bfgp/elt054

10.6084/m9.figshare.880763



```
CCCFFFFFHHHHHGIEHEGJIJJHJGJJJJGIIJJJI
@HWI-ST700693:193:C05B7ACXX:3:1101:1342:2141
GTAAGAATTTAGGGTTGTTTTTAAAAATTGAAGTAA
+
BB@FFFFFHHHHHJHIIJJJJJJJJJJJJJJJJGII
@HWI-ST700693:193:C05B7ACXX:3:1101:1458:2156
TTGGTATGTATCATTTATCAATGACAGTATTGTTTT
+
@C@FFFFFHFHFFIIJIJHJJJIJGIGHIJJJHHJI
@HWI-ST700693:193:C05B7ACXX:3:1101:1496:2166
TATATTTAATATTTTGTGTTATTTAGTTTTGTAAAG
+
CCCFFFFFGHGHHJJJGHIJJJJJGIJJJJJIJJJJ
@HWI-ST700693:193:C05B7ACXX:3:1
TTTAAAAGAATGTTTTTTTTTATAAATAAA
+
```

```
HWI-ST700693:193:C05B7ACXX:3:2308:17487:198950  272     C135321
011     255     36M     *       0       0       TAAAAAAAAATTGAA
AAAATACTACTTAAAATATAA   JJJJJJJJJJJJJJJJJJIHHHIFHHHHFFFFFCCC  N
M:i:1   ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:17728:198909  272     scaffol
d124    1441    255     36M     *       0       0       CAAATTT
TAACGAATTTTCATTAAATATATACCAAA   JJIJJJIJJJJJJJJJIJJJIJJHHHHGFFF
FFCCC   NM:i:0  ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:17749:198922  272     C149943
572     255     36M     *       0       0       TACTATATACTCATT
ACGTCATTAATTATCAAAAAC   GEJJJJIHFJIJJJIHIIIJJHGHHAHGFFFFFBBB  N
M:i:0   ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:17857:198878  256     C149349
54      255     36M     *       0       0       TTAAGTTTGGTTGAA
ATTGGTTTAGTGATTTTGGAG   @@@DDDDFHHDHH@FGHIHHEFHHIHHIIEIGHIDG  N
M:i:0   ZS:Z:++
HWI-ST700693:193:C05B7ACXX:3:2308:18471:198850  16      C138363
705     255     36M     *       0       0
ACGCATATTATTTAATCTCCA   HGGGEHHJHHGIHFJIGII
M:i:0   ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:18302:198
```



```
C10093  129     +       TTCAT   0.000   2.00    0       2       0       0       0.000   0.658
C10093  133     +       TTCTC   0.000   2.00    0       2       0       0       0.000   0.658
C10093  135     +       CTCTA   0.000   2.00    0       2       0       0       0.000   0.658
C10093  139     +       AGCTA   0.000   2.00    0       2       0       0       0.000   0.658
C1011   55      -       CAGGC   0.000   1.00    0       1       0       0       0.000   0.793
C1011   73      +       TCCGG   1.000   1.00    1       1       1       1       0.207   1.000
C1011   75      -       CGGCA   0.000   1.00    0       1       0       0       0.000   0.793
C1011   78      -       CAGGG   0.000   1.00    0       1       0       0       0.000   0.793
C1011   79      -       AGGGA   0.000   1.00    0       1       0       0       0.000   0.793
C1011   80      -       GGGAT   0.000   1.00    0       1       0       0       0.000   0.793
C1011   88      -       TTGCT   0.000   1.00    0       1       0       0       0.000   0.793
C10153  106     -       CTGTA   0.000   1.00    0       1       0       0       0.000   0.793
C10153  115     +       TTCGT   0.000   1.00    0       1       1       1       0.000   0.793
C10153  121     -       CAGAT   0.000   1.00    0       1       0       0       0.000   0.793
C10153  124     -       ATGCT   0.000   1.00    0       1       0       0       0.000   0.793
C10153  127     -       CTGTA   0.000   1.00    0       1       0       0       0.000   0.793
C10153  132     -       AAGGA   0.000   1.00    0       1       0       0       0.000   0.793
C10153  133     -       AGGAT   0.000   1.00    0       1       0       0       0.000   0.793
C10195  119     +       AACGG   0.000   1.00    0       1       1       1       0.000   0.793
C10195  121     -       CGGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  123     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  125     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  127     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  129     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  131     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  133     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  135     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  137     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  139     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  141     -       GAGAG   0.000   1.00    0       1       0       0       0.000   0.793
C10195  143     -       GAGAT   0.000   1.00    0       1       0       0       0.000   0.793
C10195  146     -       ATGCA   0.000   1.00    0       1       0       0       0.000   0.793
C10197  33      +       AACAT   0.000   1.00    0       1       0       0       0.000   0.793
C10197  44      +       TTCCA   0.000   1.00    0       1       0       0       0.000   0.793
C10197  45      +       TCCAT   0.000   1.00    0       1       0       0       0.000   0.793
C10197  50      +       TTCAT   0.000   1.00    0       1       0       0       0.000   0.793
```

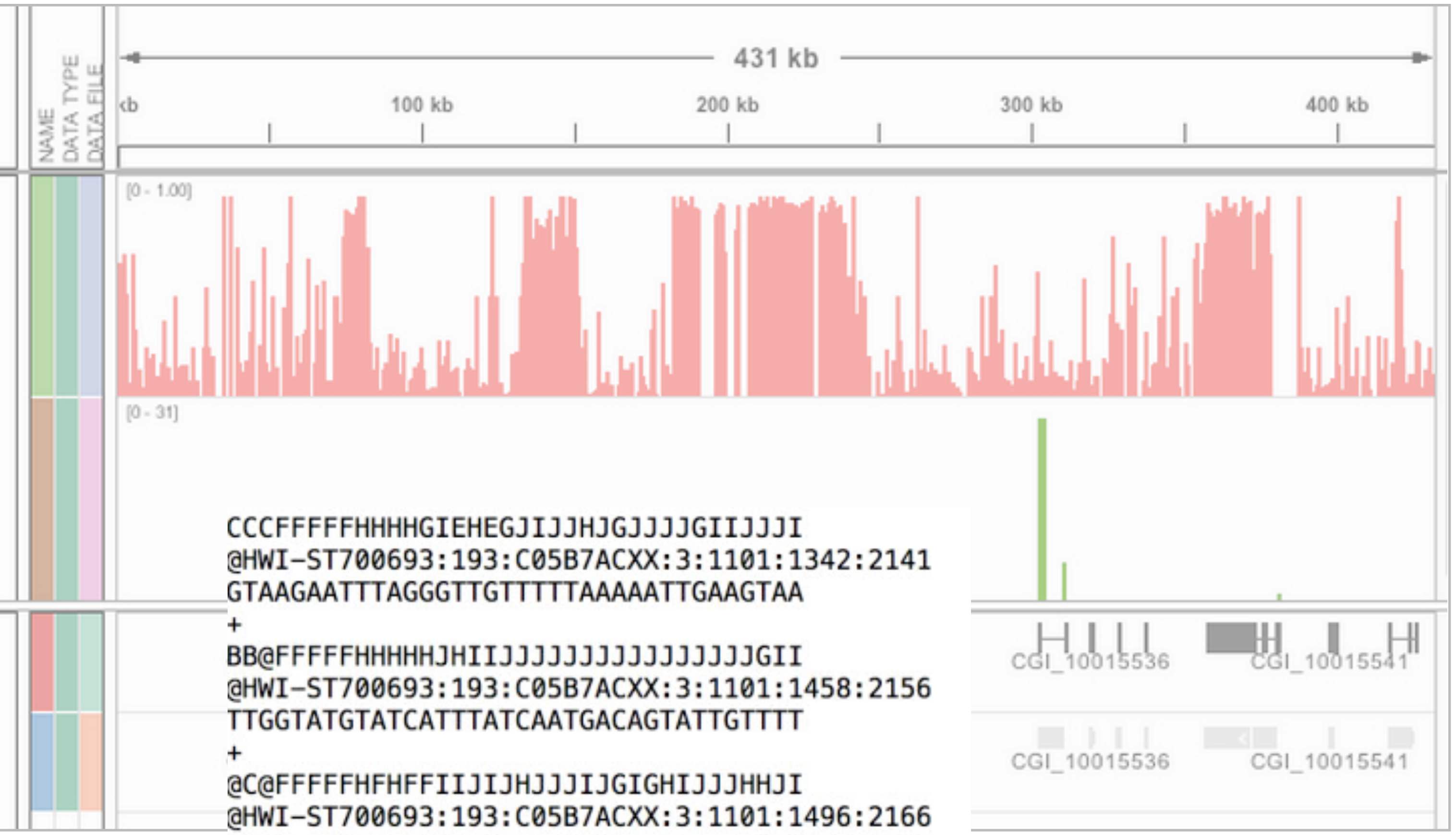
Why cloud?

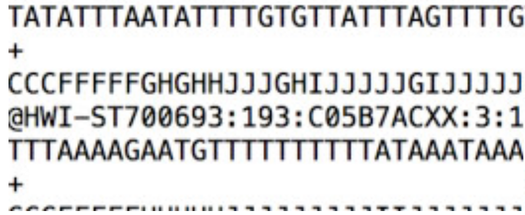
big,

big,

compute intensive,

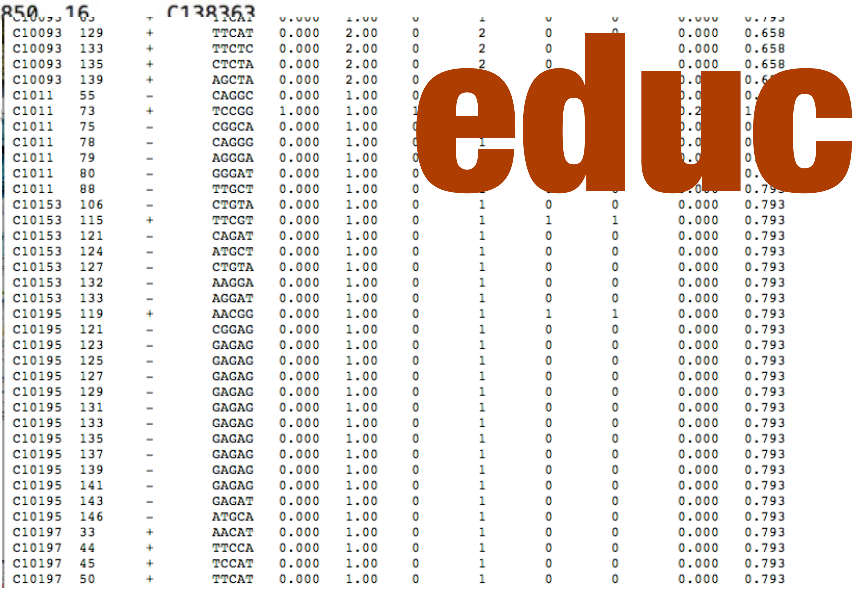
education





```
CCCFFFFFHHHHGIEHEGJIJJHJGJJJJGIIJJJI
@HWI-ST700693:193:C05B7ACXX:3:1101:1342:2141
GTAAGAATTTAGGGTTGTTTTTAAAAATTGAAGTAA
+
BB@FFFFFHHHHHJHIIJJJJJJJJJJJJJJJJJGII
@HWI-ST700693:193:C05B7ACXX:3:1101:1458:2156
TTGGTATGTATCATTTATCAATGACAGTATTGTTTT
+
@C@FFFFFHFHFFIIJIJHJJJIJGIGHIJJJHHJI
@HWI-ST700693:193:C05B7ACXX:3:1101:1496:2166
TATATTTAATATTTTGTGTTATTTAGTTTTGTAAAG
+
CCCFFFFFGHGHHJJJGHIJJJJJGIJJJJJIJJJJ
@HWI-ST700693:193:C05B7ACXX:3:1
TTTAAAAGAATGTTTTTTTTTTATAAATAAA
+
```

```
HWI-ST700693:193:C05B7ACXX:3:2308:17487:198950  272     C135321
011     255     36M     *       0       0       TAAAAAAAAATTGAA
AAAATACTACTTAAAATATAA   JJJJJJJJJJJJJJJJJJJIHHHIFHHHHFFFFFCCC  N
M:i:1   ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:17728:198909  272     scaffol
d124    1441    255     36M     *       0       0       CAAATTT
TAACGAATTTTCATTAAATATATACCAAA   JJIJJJIJJJJJJJJJIJJJIJJHHHHGFFF
FFCCC   NM:i:0  ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:17749:198922  272     C149943
572     255     36M     *       0       0       TACTATATACTCATT
ACGTCATTAATTATCAAAAAC   GEJJJJIHFJIJJJIHIIIJJHGHHAHGFFFFFBBB  N
M:i:0   ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:17857:198878  256     C149349
54      255     36M     *       0       0       TTAAGTTTGGTTGAA
ATTGGTTTAGTGATTTTGGAG   @@@DDDDFHHDHH@FGHIHHEFHHIHHIIEIGHIDG  N
M:i:0   ZS:Z:++
HWI-ST700693:193:C05B7ACXX:3:2308:18471:198850  16      C138363
705     255     36M     *       0       0
ACGCATATTATTTAATCTCCA   HGGGEHHJHHGIHFJIGII
M:i:0   ZS:Z:-+
HWI-ST700693:193:C05B7ACXX:3:2308:18302:198
```

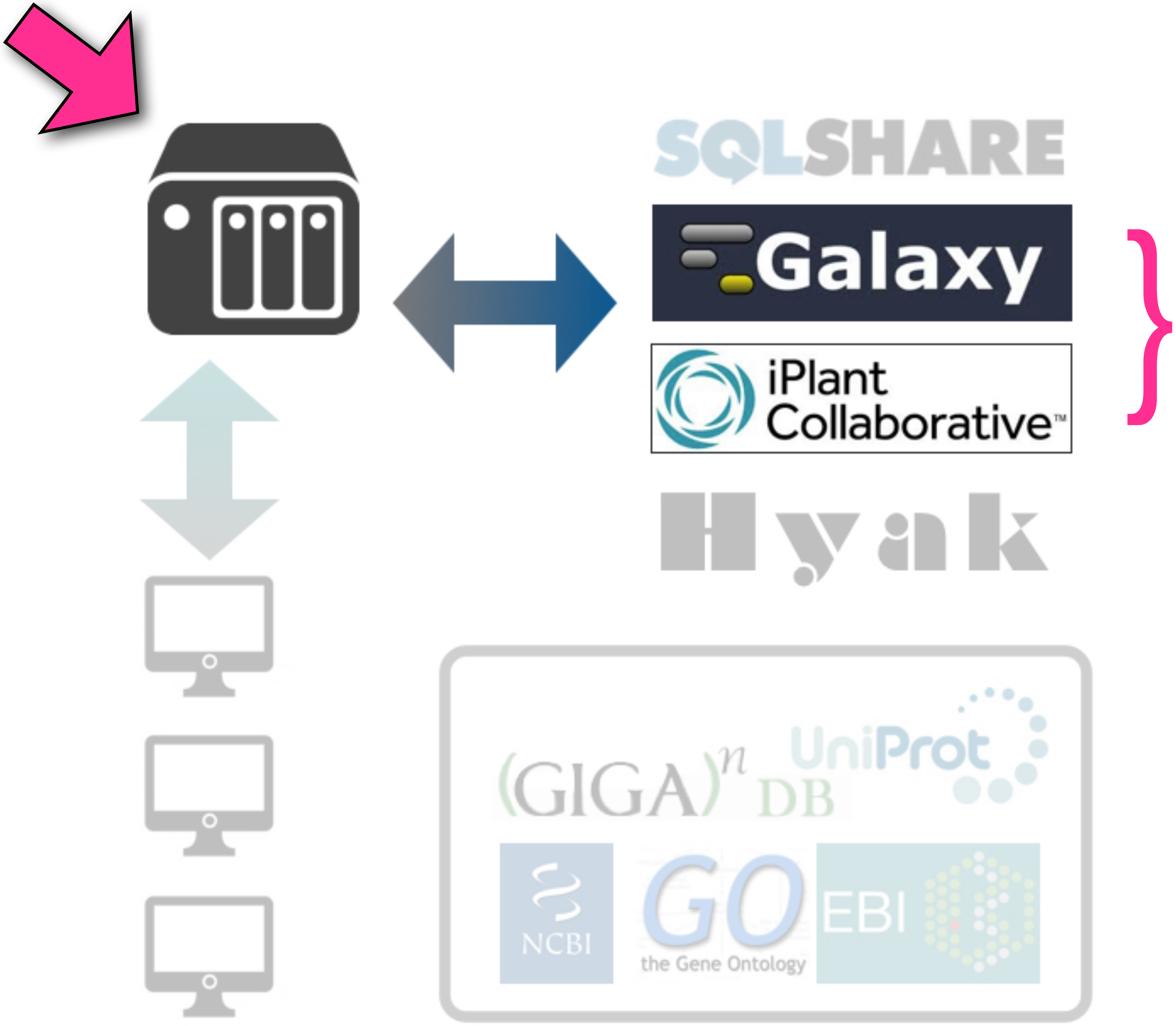


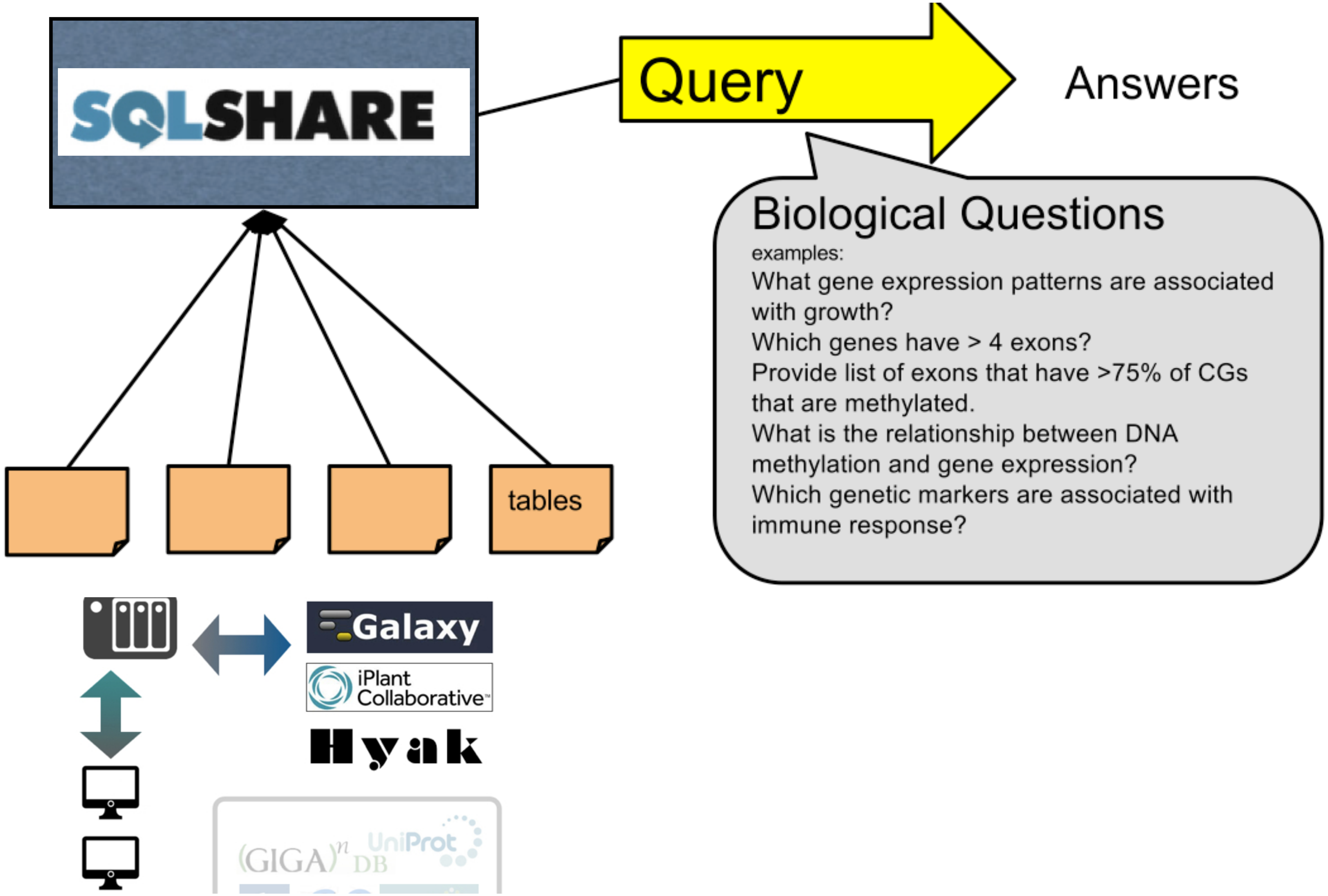
```
C10093  [illegible]  +  TTCAT  0.000  1.00  0  1  0  0  0.000  0.793
C10093  129  +  TTCAT  0.000  2.00  0  2  0  0  0.000  0.658
C10093  133  +  TTCTC  0.000  2.00  0  2  0  [illegible]  0.000  0.658
C10093  135  +  CTCTA  0.000  2.00  0  2  0  [illegible]  0.000  0.658
C10093  139  +  AGCTA  0.000  2.00  0  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]
C1011   55   -  CAGGC  0.000  1.00  0  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]
C1011   73   +  TCCGG  1.000  1.00  1  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]
C1011   75   -  CGGCA  0.000  1.00  0  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]
C1011   78   -  CAGGG  0.000  1.00  0  1  [illegible]  [illegible]  [illegible]  [illegible]
C1011   79   -  AGGGA  0.000  1.00  0  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]
C1011   80   -  GGGAT  0.000  1.00  0  [illegible]  [illegible]  [illegible]  [illegible]  [illegible]
C1011   88   -  TTGCT  0.000  1.00  0  [illegible]  [illegible]  [illegible]  [illegible]  0.793
C10153  106  -  CTGTA  0.000  1.00  0  1  0  0  0.000  0.793
C10153  115  +  TTCGT  0.000  1.00  0  1  1  1  0.000  0.793
C10153  121  -  CAGAT  0.000  1.00  0  1  0  0  0.000  0.793
C10153  124  -  ATGCT  0.000  1.00  0  1  0  0  0.000  0.793
C10153  127  -  CTGTA  0.000  1.00  0  1  0  0  0.000  0.793
C10153  132  -  AAGGA  0.000  1.00  0  1  0  0  0.000  0.793
C10153  133  -  AGGAT  0.000  1.00  0  1  0  0  0.000  0.793
C10195  119  +  AACGG  0.000  1.00  0  1  1  1  0.000  0.793
C10195  121  -  CGGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  123  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  125  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  127  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  129  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  131  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  133  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  135  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  137  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  139  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  141  -  GAGAG  0.000  1.00  0  1  0  0  0.000  0.793
C10195  143  -  GAGAT  0.000  1.00  0  1  0  0  0.000  0.793
C10195  146  -  ATGCA  0.000  1.00  0  1  0  0  0.000  0.793
C10197  33   +  AACAT  0.000  1.00  0  1  0  0  0.000  0.793
C10197  44   +  TTCCA  0.000  1.00  0  1  0  0  0.000  0.793
C10197  45   +  TCCAT  0.000  1.00  0  1  0  0  0.000  0.793
C10197  50   +  TTCAT  0.000  1.00  0  1  0  0  0.000  0.793
```

raw - 70G

mapping - 60G

tables - 40G ....


SQLSHARE
Galaxy
iPlant Collaborative™
Hyak
(GIGA)ⁿ DB
UniProt
NCBI
GO the Gene Ontology
EBI



Biological Questions

examples:

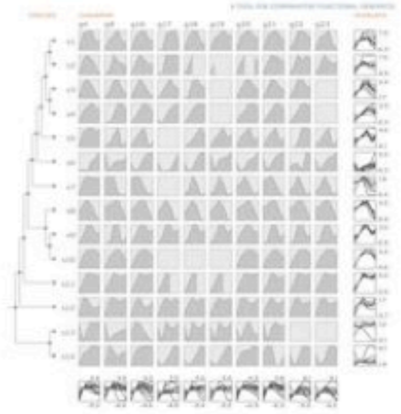
What gene expression patterns are associated with growth?

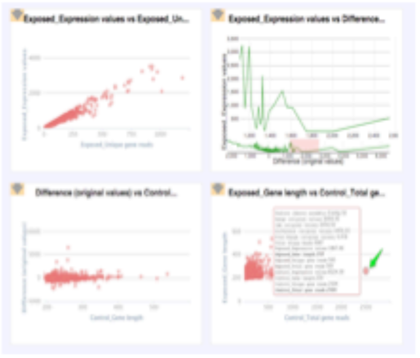
Which genes have > 4 exons?

Provide list of exons that have >75% of CGs that are methylated.

What is the relationship between DNA methylation and gene expression?

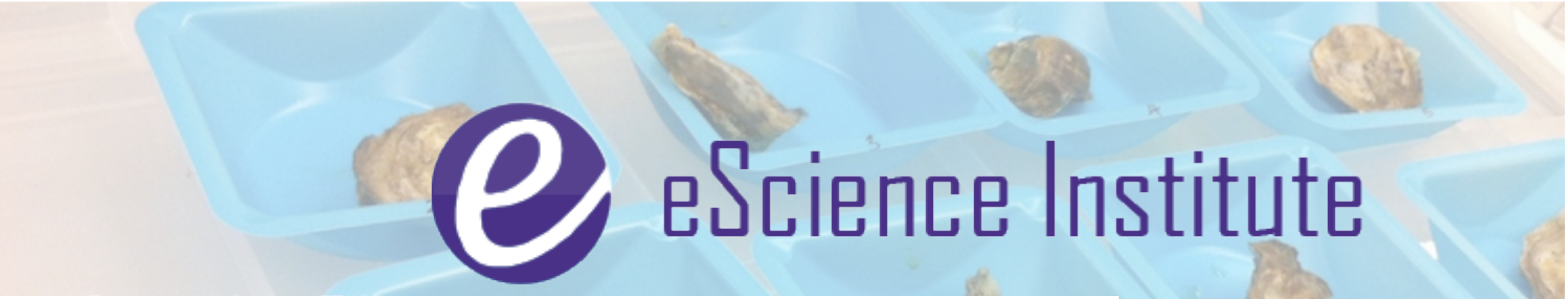
Which genetic markers are associated with immune response?





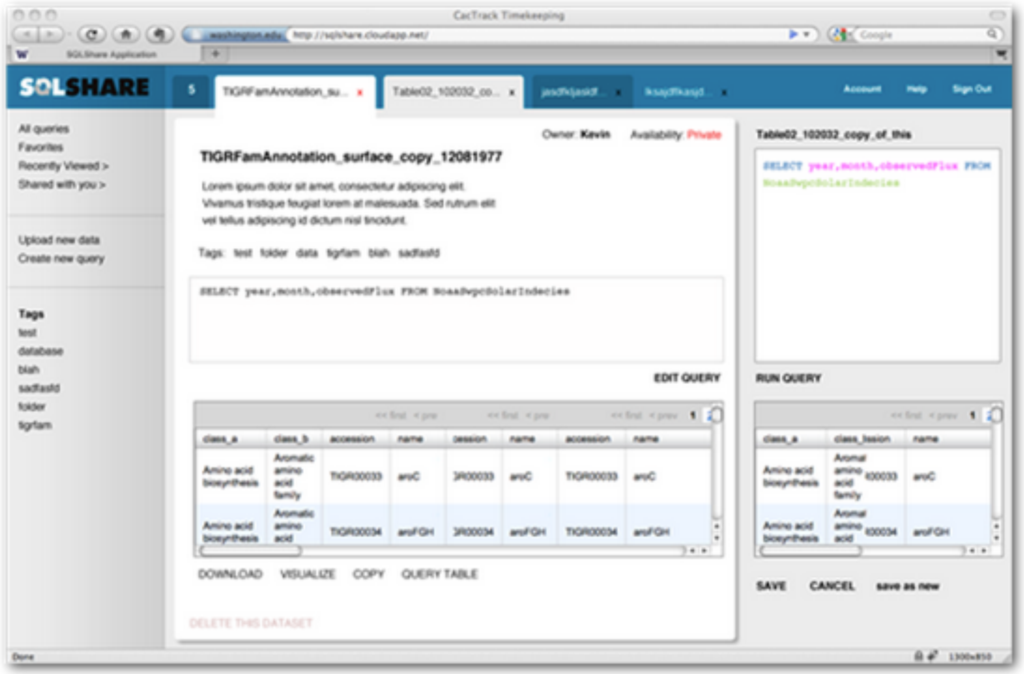
OYSTERGEN.ES

PACIFIC OLYMPIA


eScience Institute

SQLSHARE

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



Owner: Kevin Availability: Private

TIGRFamAnnotation_surface_copy_12081977

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus tristique feugiat lorem at malesuada. Sed rutrum elit vel tellus adipiscing id dictum nisl tincidunt.

Tags: test folder data tigrfam blah sadfasfd

```
SELECT year,month,observedFlux FROM NoaaWgcSolarIndecies
```

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

Table02_102032_copy_of_this

```
SELECT year,month,observedFlux FROM
NoaaWgcSolarIndecies
```

class_a	class_bssion	name
Amino acid biosynthesis	Aromat amino acid family 00033	aroC
Amino acid biosynthesis	Aromat amino acid 00034	aroFGH

Log in using your account:

UNIVERSITY *of* WASHINGTON

Google

Don't have an account?

Create a Google Account and start using SQLShare quickly.

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

ttps://sqlshare.esc
ttps://sqlshare.esc
eq
ttps://sqlshare.esc
eq

Upload

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

Modify

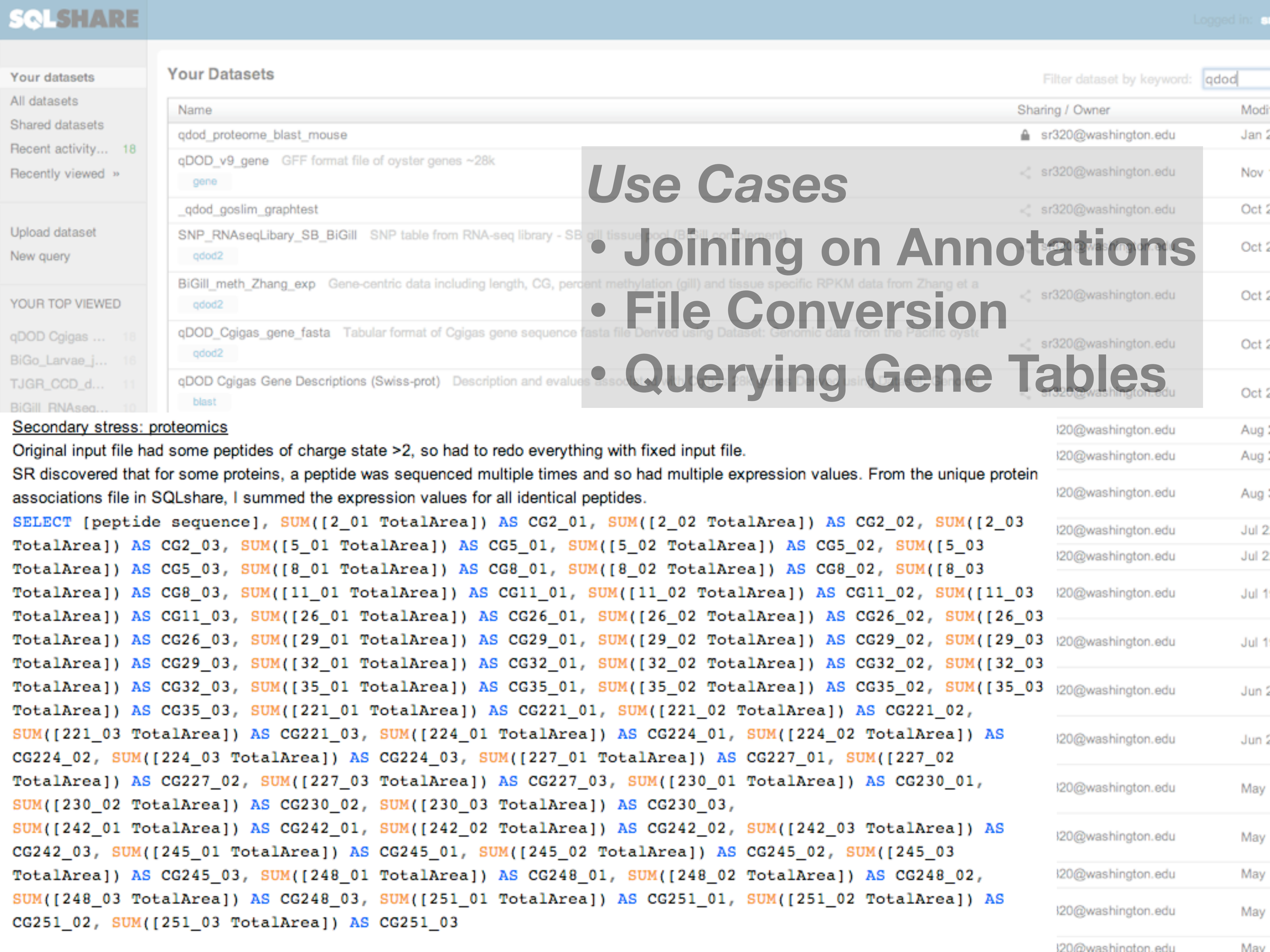
Exercise the full power of SQL even with zero programming experience: joins, subqueries, set operations.

Share

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

Use Cases

- Joining on Annotations
- File Conversion
- Querying Gene Tables



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

Your Datasets

Filter dataset by keyword: qdod

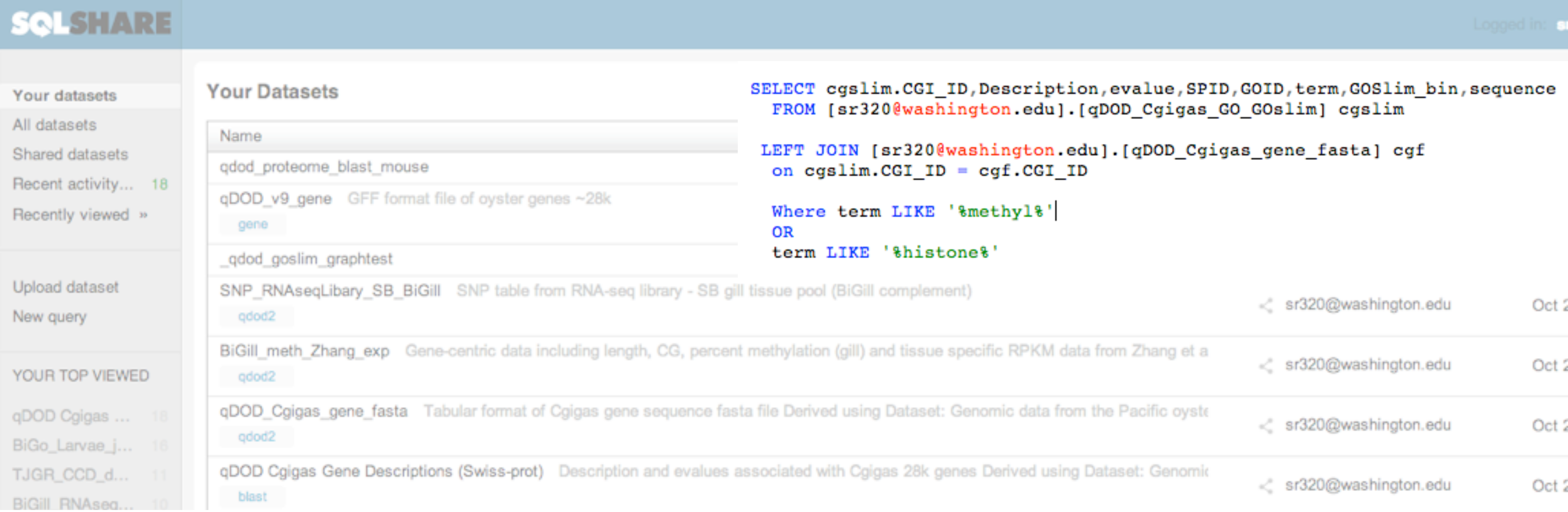
Name	Sharing / Owner	Modi
qdod_proteome_blast_mouse	🔒 sr320@washington.edu	Jan
qDOD_v9_gene GFF format file of oyster genes ~28k gene	sr320@washington.edu	Nov
_qdod_goslim_graphtest	sr320@washington.edu	Oct
SNP_RNAseqLibary_SB_BiGill SNP table from RNA-seq library - SB gill tissue pool (BiGill complement) qdod2	sr320@washington.edu	Oct
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
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
qDOD Cgigas Gene Descriptions (Swiss-prot) Description and evalues associated with Cgigas 28k genes Derived using Dataset: Genomi blast	sr320@washington.edu	Oct
	320@washington.edu	Aug
	320@washington.edu	Aug
	320@washington.edu	Aug
	320@washington.edu	Jul
	320@washington.edu	Jul
	320@washington.edu	Jul
	320@washington.edu	Jul
	320@washington.edu	Jun
	320@washington.edu	Jun
	320@washington.edu	May
	320@washington.edu	May
	320@washington.edu	May
	320@washington.edu	May
	320@washington.edu	May

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
```



SQLSHARE

Your Datasets

Name

qdod_proteome_blast_mouse

qDOD_v9_gene GFF format file of oyster genes ~28k
gene

_qdod_goslim_graphtest

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: Genomi
blast sr320@washington.edu Oct 2

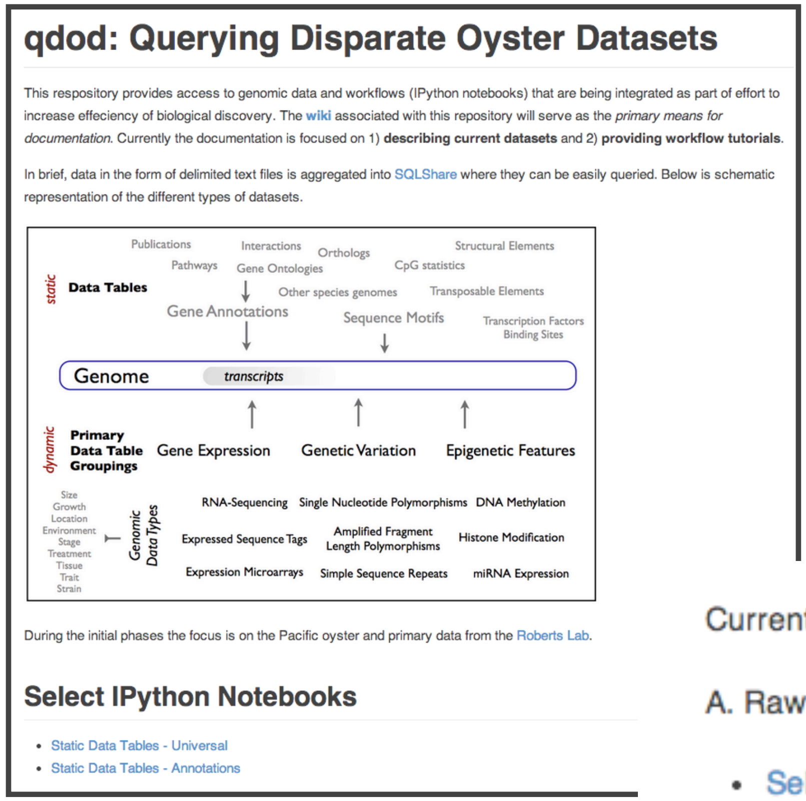
```
SELECT cgslim.CGI_ID,Description,evalue,SPID,GOID,term,GOSlim_bin,sequence
  FROM [sr320@washington.edu].[qDOD_Cgigas_GO_GOslim] cgslim

 LEFT JOIN [sr320@washington.edu].[qDOD_Cgigas_gene_fasta] cgf
  on cgslim.CGI_ID = cgf.CGI_ID

  Where term LIKE '%methyl%'
  OR
  term LIKE '%histone%'
```

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**.

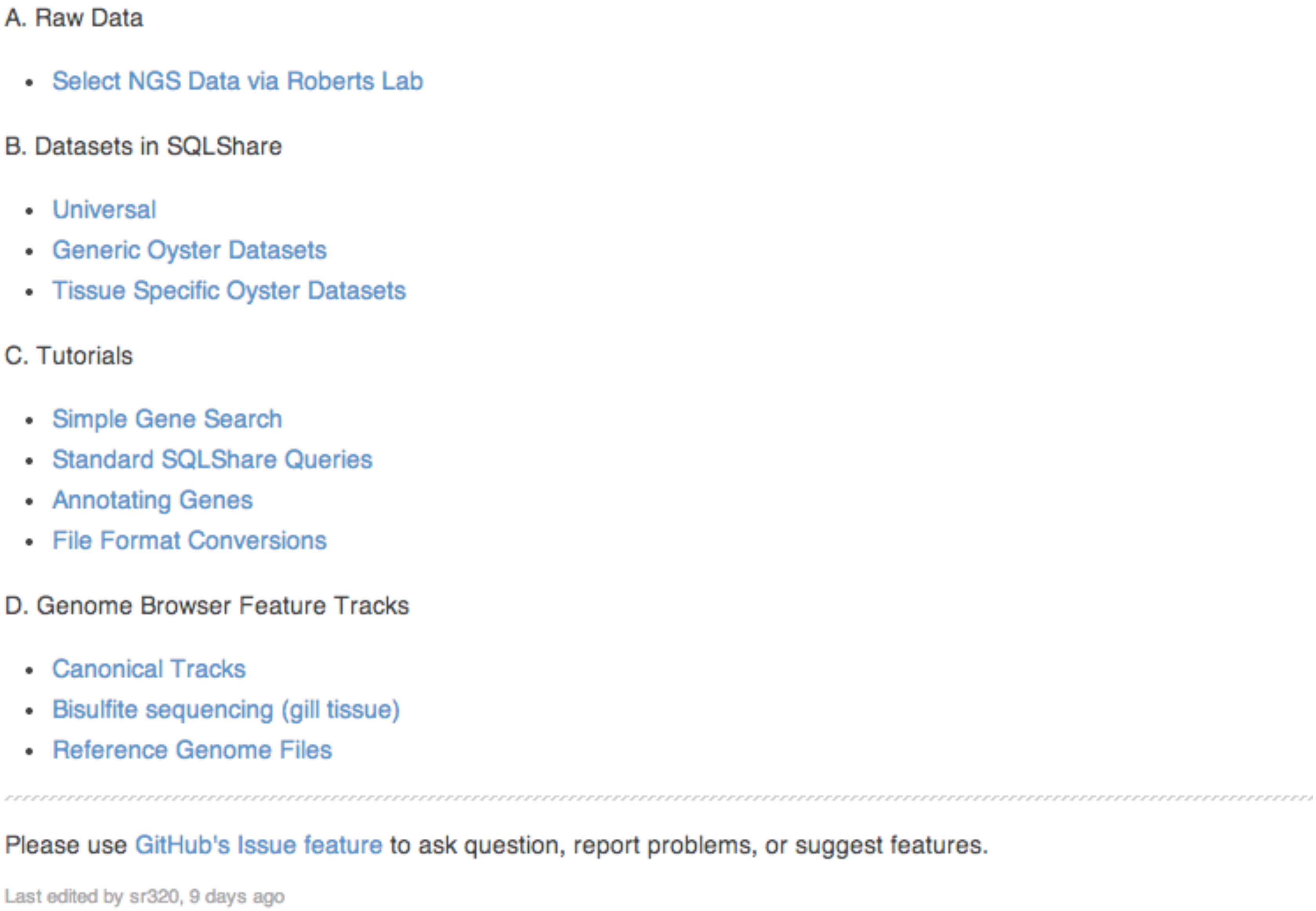
In brief, data in the form of delimited text files is aggregated into SQLShare where they can be easily queried. Below is schematic representation of the different types of datasets.

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

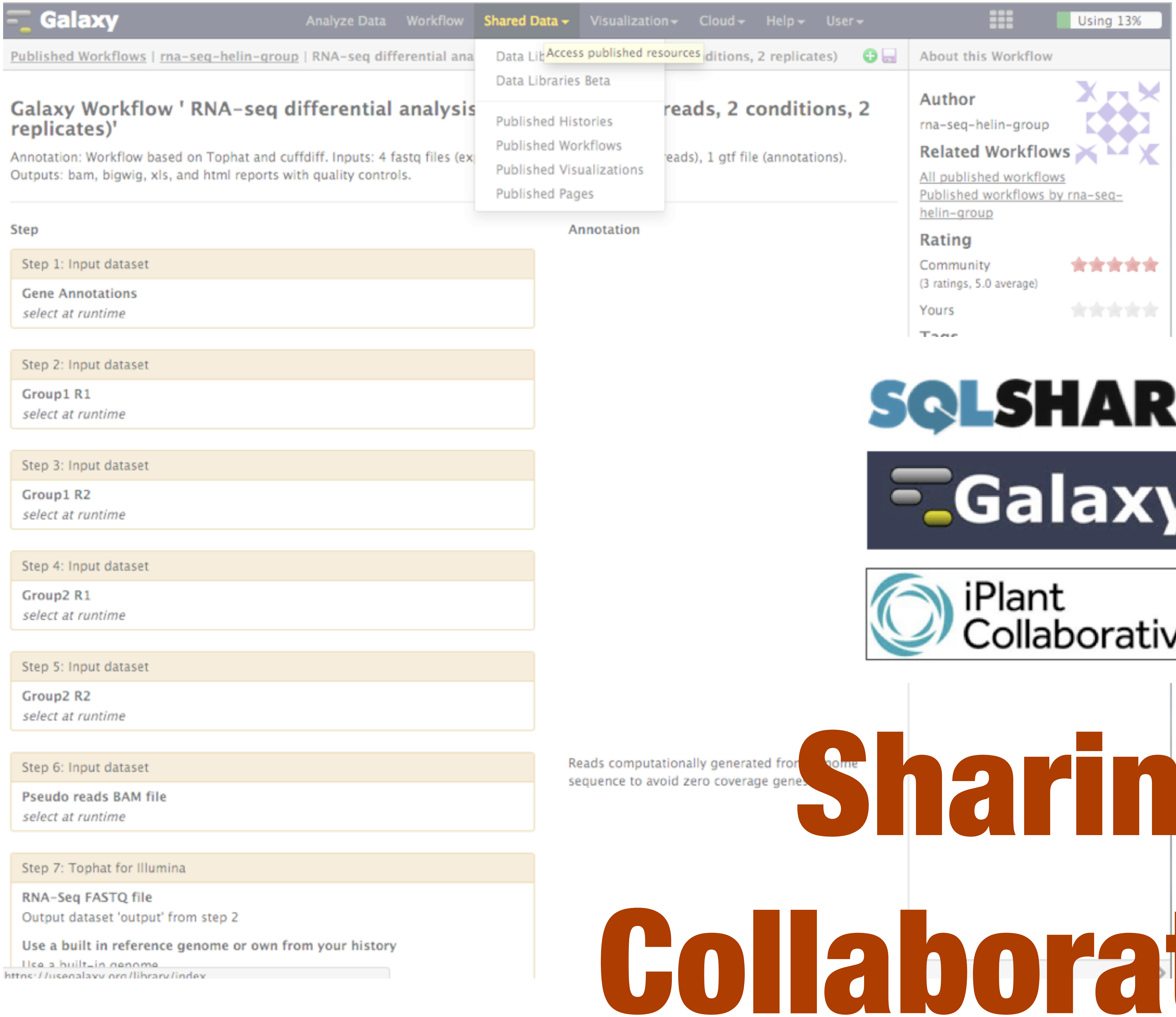
D. Genome Browser Feature Tracks

- 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

github.com/sr320/qdod/wiki


Galaxy

Analyze Data | Workflow | Shared Data ▾ | Visualization ▾ | Cloud ▾ | Help ▾ | User ▾

Using 13%

Published Workflows | rna-seq-helin-group | RNA-seq differential ana…ditions, 2 replicates)

Data Lib… Access published resources

Data Libraries Beta

Published Histories

Published Workflows

Published Visualizations

Published Pages

Galaxy Workflow ' RNA-seq differential analysis … reads, 2 conditions, 2 replicates)'

Annotation: Workflow based on Tophat and cuffdiff. Inputs: 4 fastq files (ex… reads), 1 gtf file (annotations). Outputs: bam, bigwig, xls, and html reports with quality controls.

Step	Annotation
Step 1: Input dataset **Gene Annotations** *select at runtime*	
Step 2: Input dataset **Group1 R1** *select at runtime*	
Step 3: Input dataset **Group1 R2** *select at runtime*	
Step 4: Input dataset **Group2 R1** *select at runtime*	
Step 5: Input dataset **Group2 R2** *select at runtime*	
Step 6: Input dataset **Pseudo reads BAM file** *select at runtime*	Reads computationally generated fro… …ome sequence to avoid zero coverage gene…
Step 7: Tophat for Illumina **RNA-Seq FASTQ file** Output dataset 'output' from step 2 **Use a built in reference genome or own from your history** Use a built-in genome	

https://usegalaxy.org/library/index

About this Workflow

Author

rna-seq-helin-group

Related Workflows

All published workflows

Published workflows by rna-seq-helin-group

Rating

Community ★★★★★

(3 ratings, 5.0 average)

Yours ☆☆☆☆☆

Tags

SQLSHARE

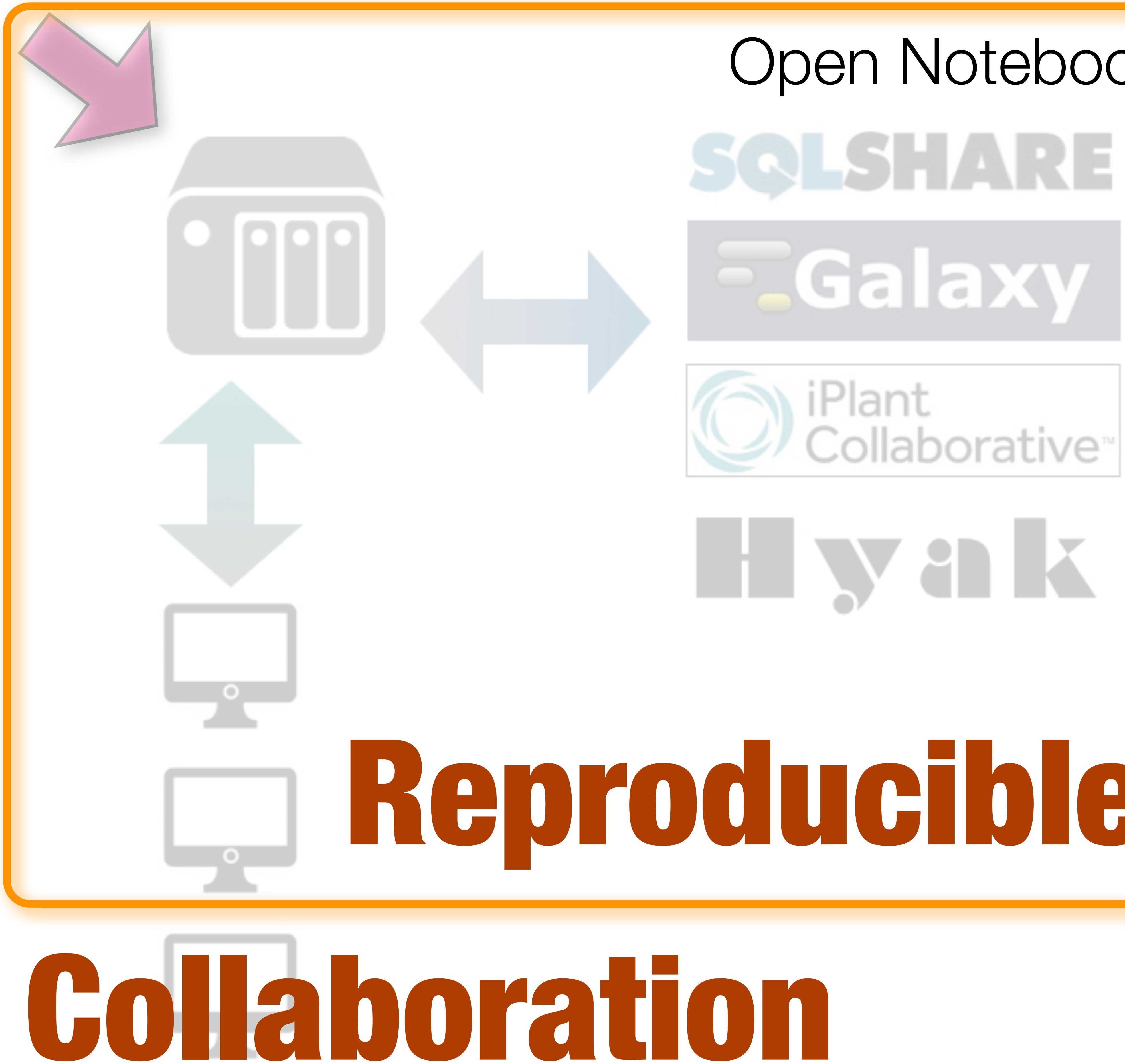
Galaxy

iPlant Collaborative™

Sharing

Collaboration*

Open Notebook Science



Reproducible

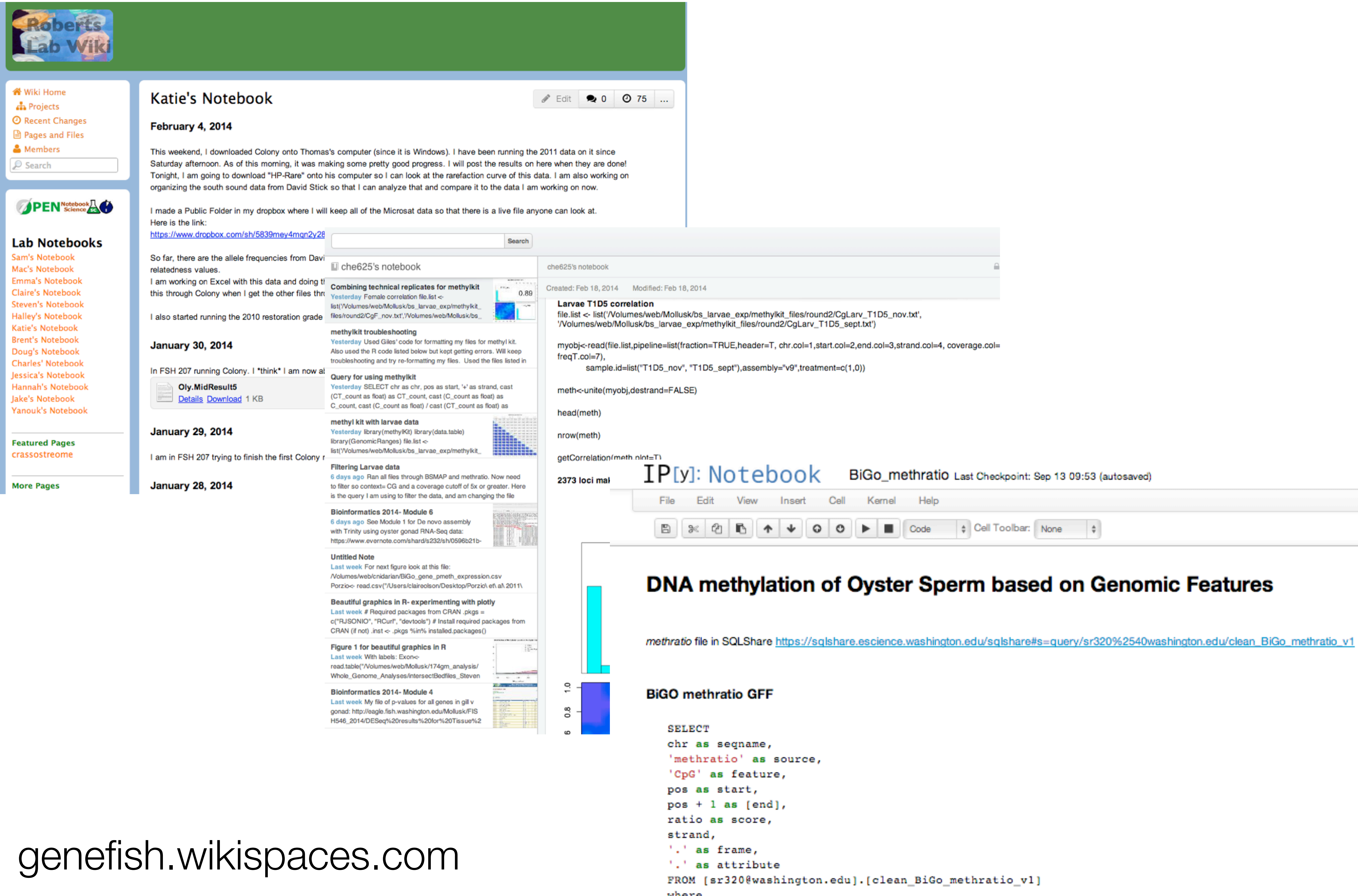
Collaboration

Open

Open Notebook Science

> ... there is a URL to a laboratory notebook that is freely available and indexed on common search engines. It does not necessarily have to look like a paper notebook but it is essential that all of the information available to the researchers to make their conclusions is equally available to the rest of the world.
>
> —Jean-Claude Bradley

Open Notebook Science



Lab Notebooks

Sam's Notebook
Mac's Notebook
Emma's Notebook
Claire's Notebook
Steven's Notebook
Halley's Notebook
Katie's Notebook
Brent's Notebook
Doug's Notebook
Charles' Notebook
Jessica's Notebook
Hannah's Notebook
Jake's Notebook
Yanouk's Notebook

Featured Pages
crassostreome

More Pages

Katie's Notebook

February 4, 2014

This weekend, I downloaded Colony onto Thomas's computer (since it is Windows). I have been running the 2011 data on it since Saturday afternoon. As of this morning, it was making some pretty good progress. I will post the results on here when they are done! Tonight, I am going to download "HP-Rare" onto his computer so I can look at the rarefaction curve of this data. I am also working on organizing the south sound data from David Stick so that I can analyze that and compare it to the data I am working on now.

I made a Public Folder in my dropbox where I will keep all of the Microsat data so that there is a live file anyone can look at.
Here is the link:
https://www.dropbox.com/sh/5839mey4mqn2y28

So far, there are the allele frequencies from Davi
relatedness values.
I am working on Excel with this data and doing th
this through Colony when I get the other files thr

I also started running the 2010 restoration grade

January 30, 2014

In FSH 207 running Colony. I *think* I am now ab

Oly.MidResult5
Details Download 1 KB

January 29, 2014

I am in FSH 207 trying to finish the first Colony r

January 28, 2014

Combining technical replicates for methylkit
Yesterday Female correlation file.list <- list('/Volumes/web/Mollusk/bs_larvae_exp/methylkit_files/round2/CgF_nov.txt','/Volumes/web/Mollusk/bs_

methylkit troubleshooting
Yesterday Used Giles' code for formatting my files for methyl kit. Also used the R code listed below but kept getting errors. Will keep troubleshooting and try re-formatting my files. Used the files listed in

Query for using methylkit
Yesterday SELECT chr as chr, pos as start, '+' as strand, cast (CT_count as float) as CT_count, cast (C_count as float) as C_count, cast (C_count as float) / cast (CT_count as float) as

methyl kit with larvae data
Yesterday library(methylKit) library(data.table) library(GenomicRanges) file.list <- list('/Volumes/web/Mollusk/bs_larvae_exp/methylkit_

Filtering Larvae data
6 days ago Ran all files through BSMAP and methratio. Now need to filter so context= CG and a coverage cutoff of 5x or greater. Here is the query I am using to filter the data, and am changing the file

Bioinformatics 2014- Module 6
6 days ago See Module 1 for De novo assembly with Trinity using oyster gonad RNA-Seq data: https://www.evernote.com/shard/s232/sh/0596b21b-

Untitled Note
Last week For next figure look at this file: /Volumes/web/cnidarian/BiGo_gene_pmeth_expression.csv Porzio<- read.csv("/Users/claireolson/Desktop/Porzio\ et\ al\ 2011\

Beautiful graphics in R- experimenting with plotly
Last week # Required packages from CRAN .pkgs = c("RJSONIO", "RCurl", "devtools") # Install required packages from CRAN (if not) .inst <- .pkgs %in% installed.packages()

Figure 1 for beautiful graphics in R
Last week With labels: Exon<- read.table("/Volumes/web/Mollusk/174gm_analysis/Whole_Genome_Analyses/intersectBedfiles_Steven

Bioinformatics 2014- Module 4
Last week My file of p-values for all genes in gill v gonad: http://eagle.fish.washington.edu/Mollusk/FISH546_2014/DESeq%20results%20for%20Tissue%2

che625's notebook

Created: Feb 18, 2014 Modified: Feb 18, 2014

Larvae T1D5 correlation

```
file.list <- list('/Volumes/web/Mollusk/bs_larvae_exp/methylkit_files/round2/CgLarv_T1D5_nov.txt',
'/Volumes/web/Mollusk/bs_larvae_exp/methylkit_files/round2/CgLarv_T1D5_sept.txt')

myobj<-read(file.list,pipeline=list(fraction=TRUE,header=T, chr.col=1,start.col=2,end.col=3,strand.col=4, coverage.col=
freqT.col=7),
        sample.id=list("T1D5_nov", "T1D5_sept"),assembly="v9",treatment=c(1,0))

meth<-unite(myobj,destrand=FALSE)

head(meth)

nrow(meth)

getCorrelation(meth,plot=T)
```

2373 loci mal

BiGo_methratio Last Checkpoint: Sep 13 09:53 (autosaved)

DNA methylation of Oyster Sperm based on Genomic Features

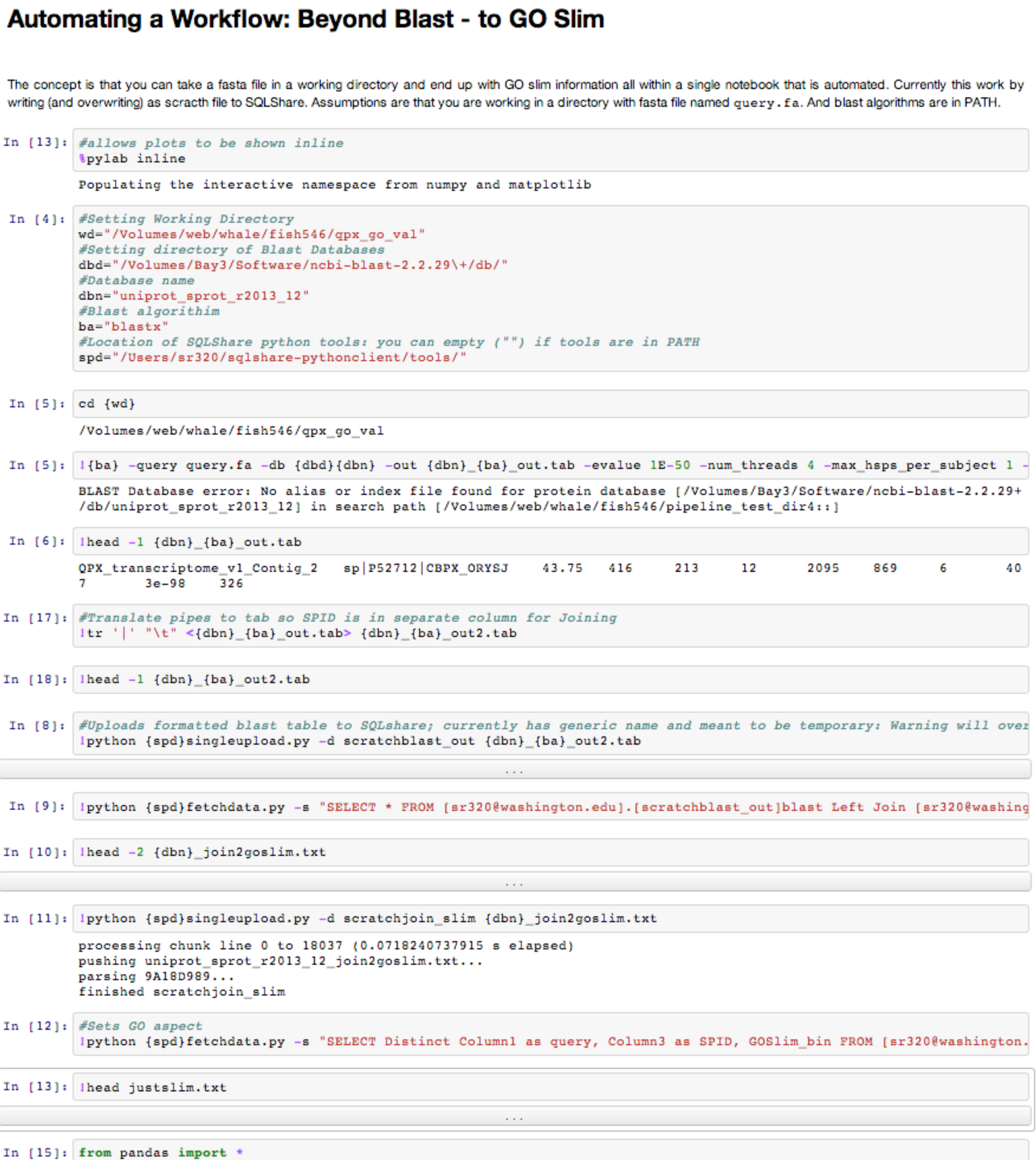
methratio file in SQLShare https://sqlshare.escience.washington.edu/sqlshare#s=query/sr320%2540washington.edu/clean_BiGo_methratio_v1

BiGO methratio GFF

```
SELECT
chr as seqname,
'methratio' as source,
'CpG' as feature,
pos as start,
pos + 1 as [end],
ratio as score,
strand,
'.' as frame,
'.' as attribute
FROM [sr320@washington.edu].[clean_BiGo_methratio_v1]
where
```

genefish.wikispaces.com

Open Notebook Science



Automating a Workflow: Beyond Blast - to GO Slim

The concept is that you can take a fasta file in a working directory and end up with GO slim information all within a single notebook that is automated. Currently this work by writing (and overwriting) as scracth file to SQLShare. Assumptions are that you are working in a directory with fasta file named `query.fa`. And blast algorithms are in PATH.

```
In [13]: #allows plots to be shown inline
         %pylab inline

         Populating the interactive namespace from numpy and matplotlib

In [4]:  #Setting Working Directory
         wd="/Volumes/web/whale/fish546/qpx_go_val"
         #Setting directory of Blast Databases
         dbd="/Volumes/Bay3/Software/ncbi-blast-2.2.29\+/db/"
         #Database name
         dbn="uniprot_sprot_r2013_12"
         #Blast algorithim
         ba="blastx"
         #Location of SQLShare python tools: you can empty ("") if tools are in PATH
         spd="/Users/sr320/sqlshare-pythonclient/tools/"

In [5]:  cd {wd}

         /Volumes/web/whale/fish546/qpx_go_val

In [5]:  !{ba} -query query.fa -db {dbd}{dbn} -out {dbn}_{ba}_out.tab -evalue 1E-50 -num_threads 4 -max_hsps_per_subject 1 -

         BLAST Database error: No alias or index file found for protein database [/Volumes/Bay3/Software/ncbi-blast-2.2.29+
         /db/uniprot_sprot_r2013_12] in search path [/Volumes/web/whale/fish546/pipeline_test_dir4::]

In [6]:  !head -1 {dbn}_{ba}_out.tab

         QPX_transcriptome_v1_Contig_2   sp|P52712|CBPX_ORYSJ    43.75   416     213     12      2095    869     6       40
         7       3e-98   326

In [17]: #Translate pipes to tab so SPID is in separate column for Joining
         !tr '|' "\t" <{dbn}_{ba}_out.tab> {dbn}_{ba}_out2.tab

In [18]: !head -1 {dbn}_{ba}_out2.tab

In [8]:  #Uploads formatted blast table to SQLshare; currently has generic name and meant to be temporary: Warning will over
         !python {spd}singleupload.py -d scratchblast_out {dbn}_{ba}_out2.tab
                                                   ...

In [9]:  !python {spd}fetchdata.py -s "SELECT * FROM [sr320@washington.edu].[scratchblast_out]blast Left Join [sr320@washing

In [10]: !head -2 {dbn}_join2goslim.txt
                                                   ...

In [11]: !python {spd}singleupload.py -d scratchjoin_slim {dbn}_join2goslim.txt

         processing chunk line 0 to 18037 (0.0718240737915 s elapsed)
         pushing uniprot_sprot_r2013_12_join2goslim.txt...
         parsing 9A18D989...
         finished scratchjoin_slim

In [12]: #Sets GO aspect
         !python {spd}fetchdata.py -s "SELECT Distinct Column1 as query, Column3 as SPID, GOSlim_bin FROM [sr320@washington.

In [13]: !head justslim.txt
                                                   ...

In [15]: from pandas import *
```

Set some variables

blast

convert file format

upload to SQLShare (python client)

join in SQLShare - download

read in pandas

matplotlib generates graph of GOsllim

Open Notebook Science

Comparison

Wiki - collaboration, versioning, search, publishing

Evernote - simple, multi-platform

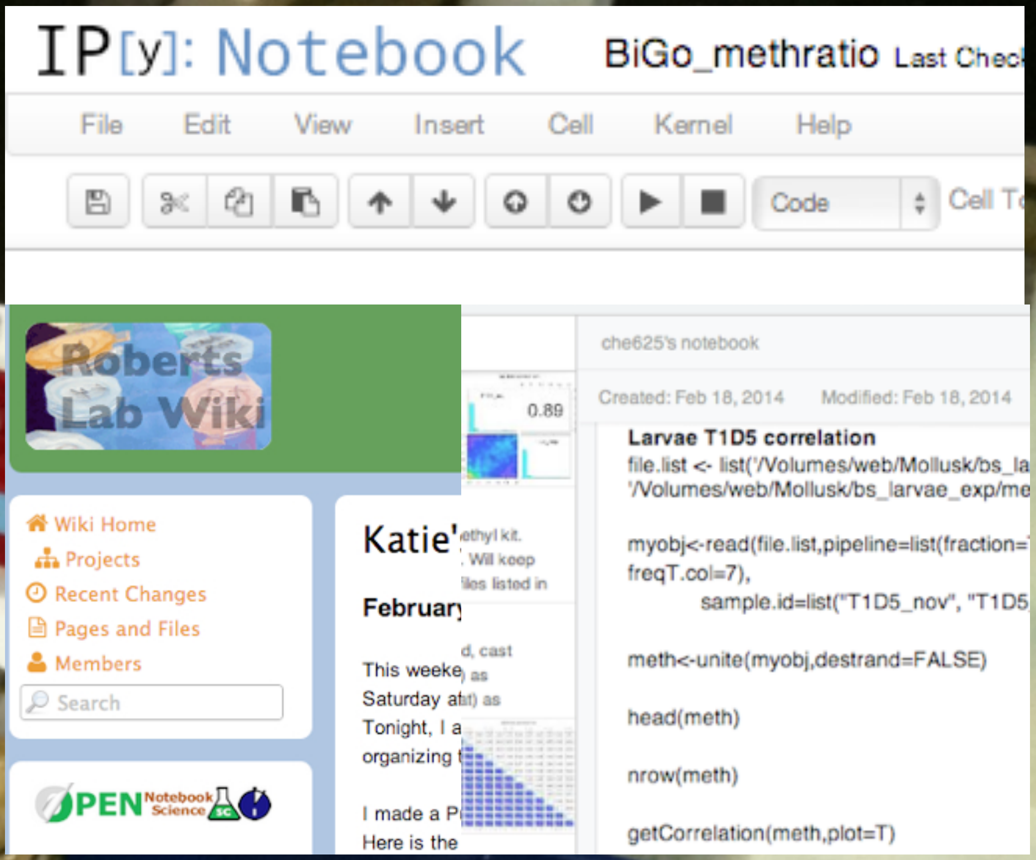
IPython - executable, versioning*

no perfection solution

SQLSHARE

Galaxy

iPlant Collaborative™



Katie'

February

This weeke
Saturday a
Tonight, I a
organizing
I made a P
Here is the

ethyl kit.
. Will keep
iles listed in
d, cast
) as
at) as

che625's notebook

Created: Feb 18, 2014 Modified: Feb 18, 2014

Larvae T1D5 correlation

```
file.list <- list('/Volumes/web/Mollusk/bs_la
'/Volumes/web/Mollusk/bs_larvae_exp/me

myobj<-read(file.list,pipeline=list(fraction=
freqT.col=7),
        sample.id=list("T1D5_nov", "T1D5

meth<-unite(myobj,destrand=FALSE)

head(meth)

nrow(meth)

getCorrelation(meth,plot=T)
```

Challenges: versioning, provenance, collaboration, simple sharing, discoverability

Reproducible Open Science

Acknowledgements

Mackenzie Gavery
Claire Olson

DNA methylation

NSF

EPA
STAR

Sam White
Brent Vadopalas
Jake Heare

NOAA

Aquaculture Program

Bill Howe
Dan Halperin

eScience Institute

USDA

robertslab.info

sr320@uw.edu