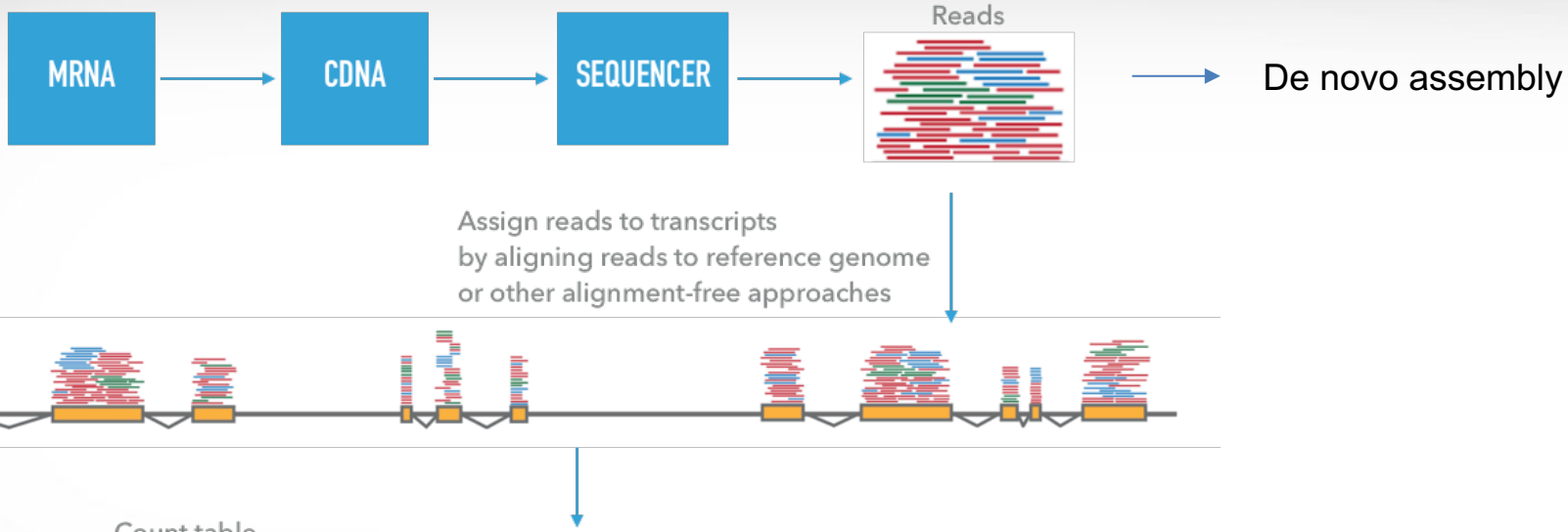
Introduction to RNA-Seq data analysis on HPRC

Shichen Wang, PhD
Bioinformatics Scientist
Genomics and Bioinformatics, AgriLife research
High Performance Research Computing

DIVISION OF RESEARCH
TEXAS A&M UNIVERSITY

RNA-Seq Overview



Count table

GENE NAME	TREAT_REP1	TREAT_REP2	TREAT_REP3	CONTROL_REP1	CONTROL_REP2	CONTROL_REP3
GENE0001	12	13	14	3	9	6
GENE0002	20	6	2	7	17	13
GENE0003	6	18	11	15	10	19
GENE0004	11	3	20	9	9	10
GENE0005	9	1	14	10	10	20
GENE0006	14	18	7	11	11	18
GENE0007	10	8	14	9	20	17
GENE0008	20	15	8	3	16	7

WHAT DOES RNASEQ DATA PROVIDE US?

- Measure gene expression (relatively)
- Annotate transcripts
- Discover novel transcripts/isoforms
- Discover nucleotide variations

RNASEQ EXPERIMENT DESIGN

- Questions often being asked:
  - How many replicates should I have?
  - How many reads should be generated?
- WHAT IS YOUR PRIMARY EXPERIMENTAL OBJECTIVE?
  - Detect DEGs
  - Annotation transcripts
  - Detect nucleotide variations

Biological vs Technical Replication

- Biological replicates include multiple samplings within a population
- Technical replicates include multiple prepping and or resequencing the same individual
- Biological replicates generally increase statistical power more than technical replicates
  - Biological variability is generally greater than technical variability
  - Biological replicates contain both biological and technical variability

Table 1.1 Recommendations for RNA-seq options based upon experimental objectives.

Criteria	Annotation	Differential Gene Expression
Biological replicates	Not necessary but can be useful	Essential
Coverage across the transcript	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not as important; however the only reads that can be used are those that are uniquely mappable.
Depth of sequencing	High enough to maximize coverage of rare transcripts and transcriptional isoforms	High enough to infer accurrate statistics
Role of sequencing depth	Obtain reads that overlap along the length of the transcript	Get enough counts of each transcript such that statistical inferences can be made
DSN	Useful for removing abundant transcripts so that more reads come from rarer transcripts	Not recommended since it can skew counts
Stranded library prep	Important for de Novo transcript assembly and identifying true anti-sense trancripts	Not generally required especially if there is a reference genome
Long reads (>80 bp)	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not generally required especially if there is a reference genome
Paired-end reads	Important for de Novo transcript assembly and identifying transcriptional isoforms	Not important

http://rnaseq.uoregon.edu

How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?

Nicholas J. Schurch[1,6], Pietá Schofield[1,2,6], Marek Gierliński[1,2,6], Christian Cole[1,6], Alexander Sherstnev[1,6], Vijender Singh[2], Nicola Wrobel[3], Karim Gharbi[3], Gordon G. Simpson[4], Tom Owen-Hughes[2], Mark Blaxter[3] and Geoffrey J. Barton[1,2,5]

+ Author Affiliations

Corresponding authors: g.g.simpson@dundee.ac.uk, t.a.owenhughes@dundee.ac.uk, Mark.Blaxter@ed.ac.uk, g.j.barton@dundee.ac.uk

↵[6] These authors contributed equally to this work.

Data analysis on HPRC system

Where to Find NGS Tools

- TAMU HPRC Documentation
  - https://hprc.tamu.edu/wiki/index.php/Ada:Bioinformatics
- Type the following UNIX **commands** to see which tools are already installed on Ada
  - `module avail`
  - `module spider toolname` (not case sensitive, but read entire output)
  - `module key assembly` (some modules may be missed because this searches tool descriptions)
- If you find a tool that you want installed on Ada, send an email with the URL link to: **help@hprc.tamu.edu**
  - SeqAnswers http://seqanswers.com/wiki/Software/list
  - omictools.com
  - slideshare.net – find shared NGS presentations

Finding NGS job template scripts using GCATemplates on Ada

```
mkdir $SCRATCH/rnaseq_class
cd $SCRATCH/rnaseq_class
module load GCATemplates
gcatemplates
```

For practice, we will copy a template file

- Select #13 RNA-seq, #1 QC, #1 rnaseqc, #1 two samples
- Final step will save a template job script file to your current working directory
- After you save the template file:

```
module purge
```

Genomic Computational Analysis Templates

QC Evaluation

```
module spider fastqc
```

- Use FastQC to visualize quality scores
  - Displays quality score distribution of reads
    - Input is a fastq file or files
    - Can disable grouping of sequence regions
  - Will alert you of poor read characteristics
  - Displays a representative sample of the fastq file
  - Can be run as a GUI or a command line interface
- FastQC will process using one CPU core per file
  - If there are 10 fastq files to analyze and 4 cores used
    - 4 files will start processing and 6 will wait in a queue
    - If there is only one fastq file to process then using 10 cores does not speed up the process

QC Quality Trimming

- Sequence quality trimming tools

```
module spider Trimmomatic
```

recommended tool

- Trimmomatic will maintain paired end read pairing after trimming
- Trim reads based on quality scores
  - Trim the same number of bases from each read or
  - Use a sliding window to calculate average quality at ends of sequences
- Decide if you want to discard reads with Ns
  - some assemblers replace Ns with As or a random base G, C, A or T
- Trim adapter sequences
  - Trimmomatic has a file of Illumina adapter sequences

```
module load Trimmomatic/0.36-Java-1.8.0_92
```

```
ls $EBROOTTRIMMOMATIC/adapters/
```

RNA-SeQC

```
module spider RNA-SeQC
```

- Provides alignment metrics & graphs all samples together
  - Yield alignment and duplication rates
  - GC bias
  - rRNA content
  - Regions of alignment (exon, intron, intragenic)
  - continuity of coverage
  - 5'/3' bias and much more …
- Metrics can help identify sample outliers by comparing metrics of all samples

RNA-SeQC: RNA-seq metrics for quality control and process optimization

DeLuca, et al. Bioinformatics. 2012 Jun 1; 28(11): 1530–1532. Published online 2012 Apr 25. doi: 10.1093/bioinformatics/bts196PMCID: PMC3356847

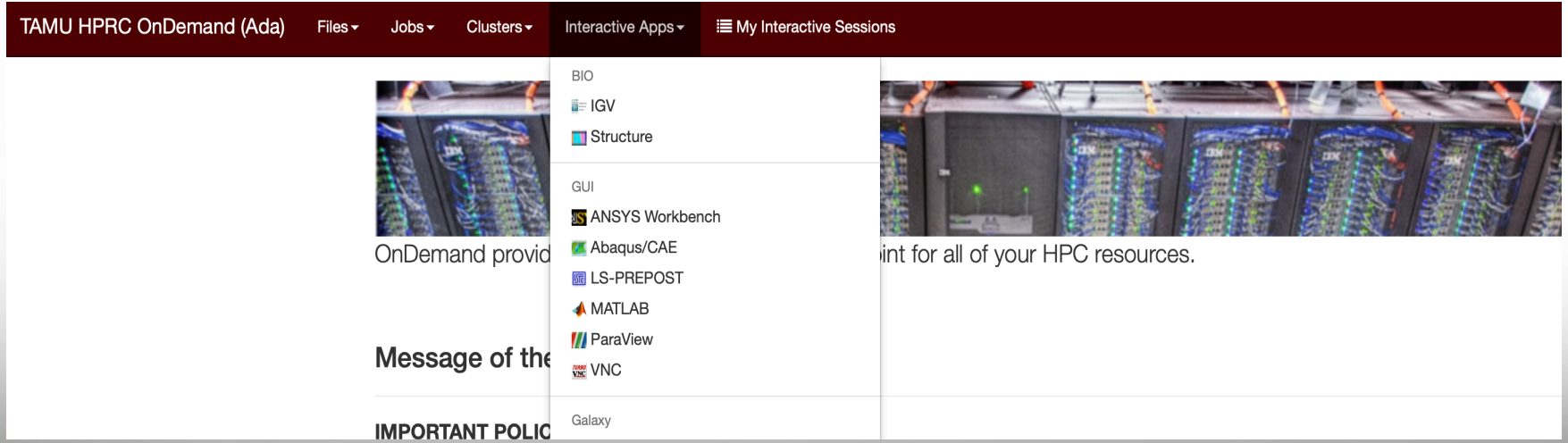
Mapping RNA-seq Reads to a Reference Assembly

Splice-Aware Aligners for RNA-seq Short Reads

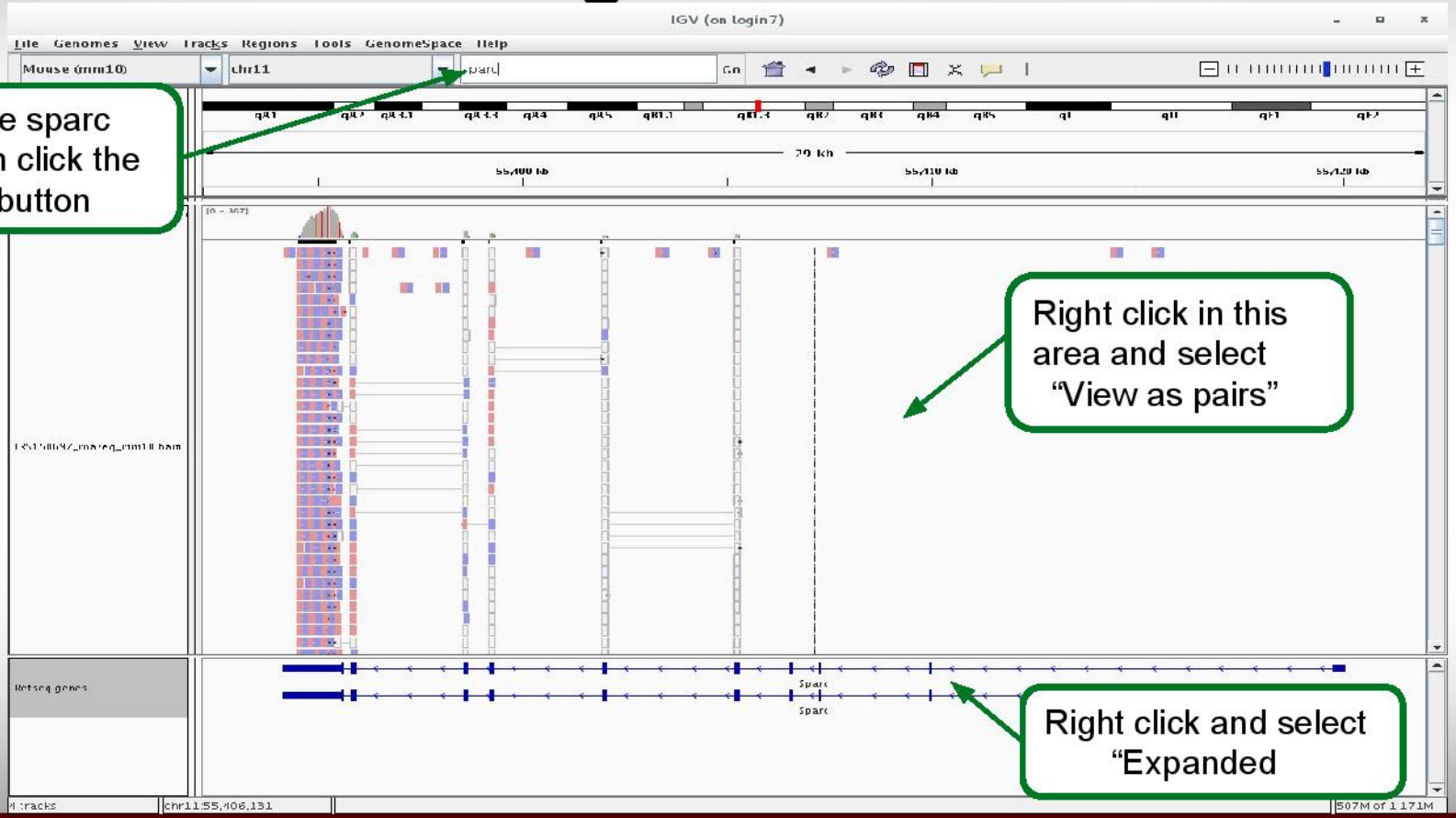
- HISAT2 which supersedes TopHat2
  - `/scratch/datasets/genome_indexes/ucsc/mm10/hisat2_index/`
- STAR (on Ada as module STAR-STAR)
  - Uses gene annotations in gtf format
    - can use **`gffread`** in Cufflinks module to convert gff3 to gtf
  - supports PacBio but should use non-default settings
    - *Bioinfx study: Optimizing STAR aligner for Iso Seq data*
- BBMap
  - also supports PacBio and Nanopore
- GMap
  - also supports PacBio and Nanopore

Integrative Genomics Viewer (IGV) Exercise

- IGV is a genome browser with pre-loaded genomes available in which you can use to view multiple .bed, .sam and .vcf files.
- Running IGV through **portal.hprc.tamu.edu**



IGV viewing indexed bam file



RNA-seq for Differential Expression

RNA-seq Sequence Fragment Counting

- Alignment based
  - Non-normalized alignment counts
    - HTSeq-count
  - Normalized (RPKM, FPKM, TPM)
    - eXpress (outputs FPKM)
    - RSEM (isoform/gene level estimates without RPKM or FPKM)
    - Trinity Transcript Quantification
      - A Trinity script can run: Kallisto, RSEM, eXpress, Salmon
- Non-Alignment based
  - Kallisto (pseudoalignment)
  - Salmon (lightweight alignment)
  - Sailfish (k-mer)

RPKM vs FPKM vs TPM

- The number of **R**eads **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for single end reads
- The number of **F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads.
  - Intended for paired-end reads
    - If both paired reads align to a transcript then they are counted as one alignment
- **T**ranscripts **P**er kilobase **M**illion
  - Normalize for gene length first
  - Normalize for sequence depth second

http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/

Tuxedo Suite

- HISAT2
  - splice aware mapping of RNA-seq reads
  - TopHat (which uses Bowtie2) and HISAT are superseded by HISAT2
- Cufflinks
  - assembles aligned reads into transcripts and estimates their abundances
- Cuffdiff
  - compares RNA-seq abundance (expression) levels of two samples or groups

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
CAWT_00001	CAWG_00001	-	chr_1.1:8373-9093	q1	q2	OK	111.944	163.869	0.549763	0.768107	0.58795	0.996768	no
CAWT_00002	CAWG_00002	-	chr_1.1:11447-12425	q1	q2	OK	14.5992	30.9037	1.08189	1.3841	0.2921	0.98312	no
CAWT_00003	CAWG_00003	-	chr_1.1:14130-14451	q1	q2	OK	248.323	259.152	0.0615814	0.172186	0.94685	0.996768	no
CAWT_00004	CAWG_00004	-	chr_1.1:14890-16045	q1	q2	OK	60.9546	86.0009	0.496617	0.604904	0.6204	0.996768	no
…													
…													
CAWT_01628	CAWG_01628	-	chr1.2:664522-665344	q1	q2	OK	3.56447	157.849	5.46871	6.64693	0.00015	0.0482417	yes

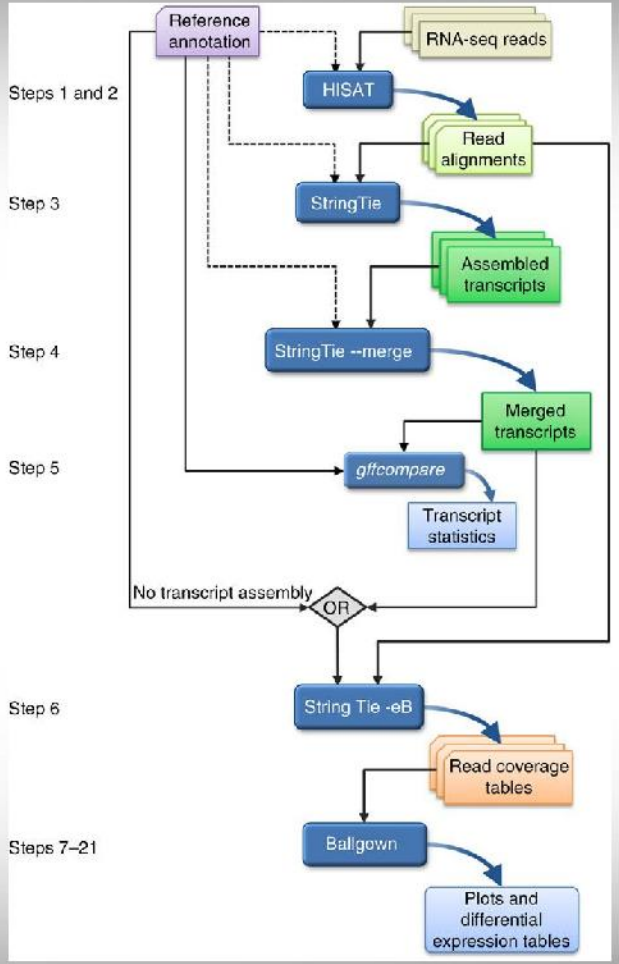
p_value = The uncorrected p-value of the test statistic.
q_value = The FDR-adjusted p-value of the test statistic

“New Tuxedo” Protocol

Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Pertea, et al. Nature Protocols 11,1650–1667 (2016)
doi:10.1038/nprot.2016.095

HISAT2 supercedes HISAT



Sailfish

- Alignment-free isoform quantification from RNA-seq data (uses k-mers)
- Requires a set of target transcripts (fasta)
  - From a reference or a *de novo* assembly
- Requires sequence reads (fasta or fastq)

Name	Length	EffectiveLength	TPM	NumReads
TRINITY_DN30_c0_g1_i1	215	68.4635	236.773	233
TRINITY_DN43_c0_g1_i1	280	102.34	5971.5	8784
TRINITY_DN88_c0_g1_i1	217	69.3036	191.74	191
TRINITY_DN59_c0_g1_i1	393	194.337	4092.64	11432
TRINITY_DN98_c0_g1_i1	205	64.4299	1097.09	1016
TRINITY_DN17_c0_g1_i1	310	122.99	2634.35	4657

R Bioconductor

- Popular R bioconductor packages for RNA-seq
  - CQN – Normalization of RNA-seq data
  - edgeR – Differential gene expression
  - DESeq, DESeq2 – Differential gene expression
  - cummeRbund – analysis/visualization of cufflinks data
- Bioconductor packages can be found in this R version

```
module load R_tamu/3.3.1-intel-2015B-default-mt
```

RNA-seq for Transcriptome Assembly

RNA-seq Transcriptome Assembly

- Assembly with a reference genome

```
module spider Trinity
```

```
module spider HISAT2 Cufflinks
```

```
module spider Scripture
```

```
module spider StringTie
```

- *de novo* assembly without a reference genome

```
module spider Trinity
```

```
module spider Oases
```

Running Trinity on Ada

- Trinity uses 100,000s of intermediate files
  - Contact **help@hprc.tamu.edu** and request a file quota increase before running Trinity
  - Run one Trinity job at a time and check resource usage
    - `showquota`
    - It is recommended not to run multiple Trinity jobs unless you know memory usage and an estimate of the number of temporary files
  - Trinity creates checkpoints and can be restarted if it stops due to file/disk quota met, out of memory or runtime
    - Checkpoints are not available when running Trinity in Galaxy
    - Checkpoints are not available if you use $TMPDIR with Trinity
      - need to rsync results from $TMPDIR at end of job script
      - checkpoints are stored in $TMPDIR which is deleted after job ends
- See GCATemplates for sample Trinity scripts

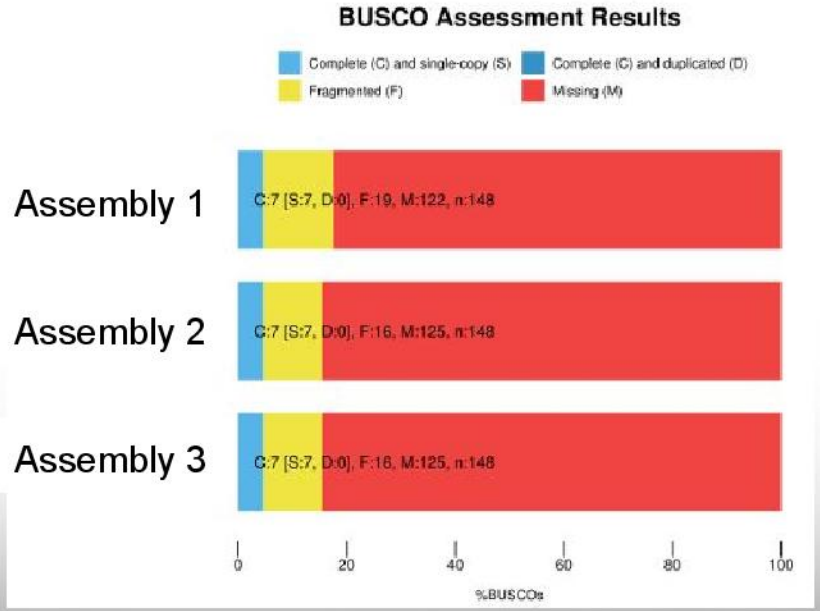
Transcriptome Assembly Completeness

The completeness of a transcriptome can be estimated by using a set of highly conserved genes that are common to specific taxonomic groups

- 44 taxonomic groups available
  - aves, bacteria, eukaryota, insecta, vertebrata, ...
- BUSCO – uses single-copy genes to access transcriptome assembly and annotation completeness
  - evaluates % complete 'BUSCOs', % fragmented, % missing
  - can run in genome, transcriptome or protein mode
  - `module spider BUSCO`

Transcriptome Assembly Completeness

BUSCO script (generate_plot.py) can be used to plot multiple BUSCO short summaries to compare different assemblies



Any questions?

DEG Analysis tutorial:

1. Quality check: FastQC
2. Trim adapters and low quality bases: Trimmomatic
3. Alignment to the reference: HISAT2
4. Reference guided assembly: cufflinks
5. Merge the transcripts: cuffmerge
6. Differential expression analysis: edgeR
7. Visualize the results: heatmap and volcano plot

HTTP://GALAXYPROJECT.ORG

Galaxy

Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

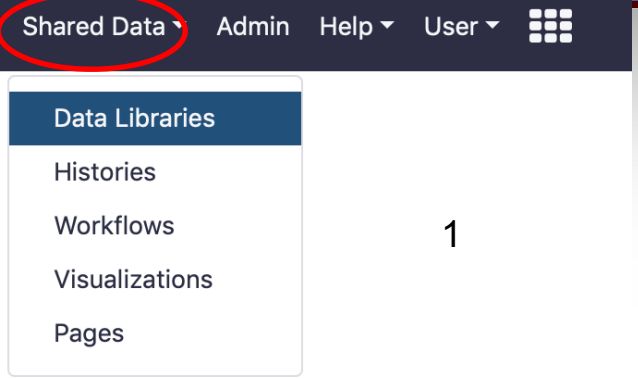
Screencasts, Galaxy 101, ...

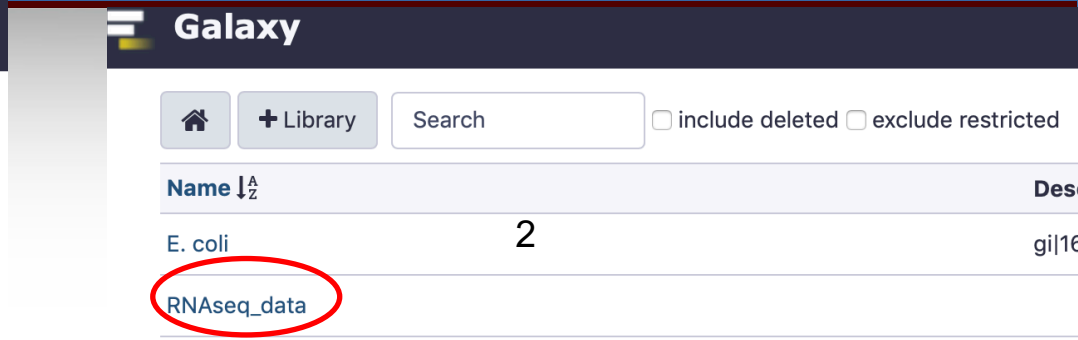
Get Involved

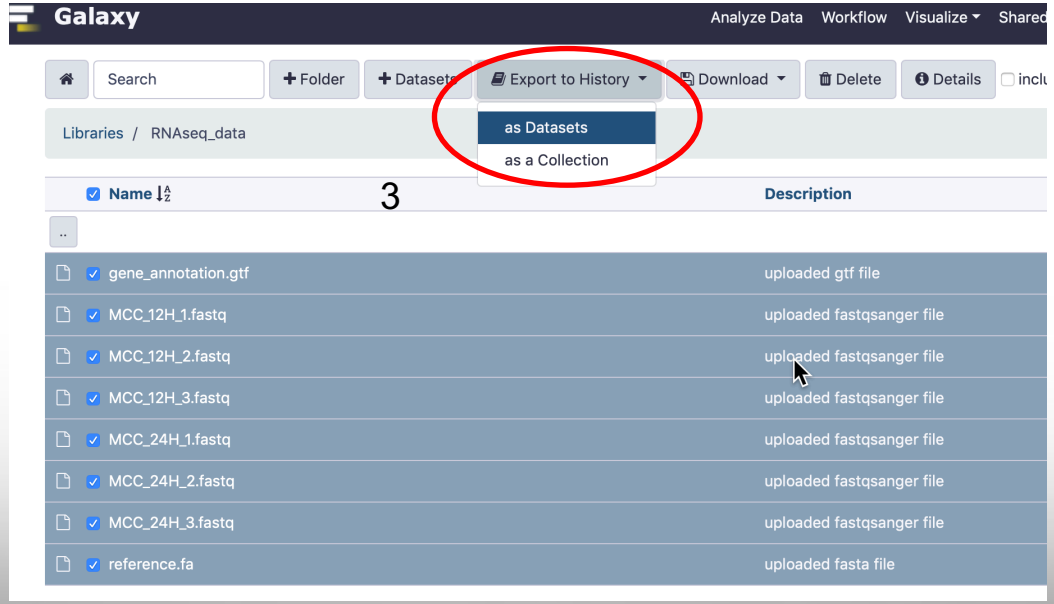
Mailing lists, Tool Shed, wiki

https://galaxy-terra.hprc.tamu.edu/bdf

Thanks to Dr. Charles Michael Dickens!


kflow
Visualize
Shared Data
Admin
Help
User
Data Libraries
Histories
Workflows
Visualizations
Pages
1


Galaxy
Library
Search
include deleted
exclude restricted
Name
Des
E. coli
gi|16
2
RNAseq_data


Galaxy
Analyze Data
Workflow
Visualize
Shared
Search
Folder
Datasets
Export to History
Download
Delete
Details
incl
Libraries / RNAseq_data
as Datasets
as a Collection
Name
3
Description
..
gene_annotation.gtf
uploaded gtf file
MCC_12H_1.fastq
uploaded fastqsanger file
MCC_12H_2.fastq
uploaded fastqsanger file
MCC_12H_3.fastq
uploaded fastqsanger file
MCC_24H_1.fastq
uploaded fastqsanger file
MCC_24H_2.fastq
uploaded fastqsanger file
MCC_24H_3.fastq
uploaded fastqsanger file
reference.fa
uploaded fasta file