Tufts UNIVERSITY

Day 3
RNA-Seq downstream analysis with nf-core differentialabundance

Shirley (Xue) Li
Bioinformatician
TTS Research Technology

Yucheng Zhang
Bioinformatics Engineer
TTS Research Technology

tts-research@tufts.edu

nextflow nf-core OPEN OnDemand

Overview

- Day 1 (April 3)
  1. Intro to nextflow and nf-core (Yucheng)
  2. How to run nf-core pipelines at Tufts HPC (Yucheng)
  3. How to download raw fastQ data with nf-core fetchngs pipeline (Shirley)

- Day 2 (April 4)
  1. Clean cache data (Yucheng)
  2. Nextflow tower (Yucheng)
  3. Running RNA-Seq analysis with nf-core rnaseq pipeline (Shirley)

- Day 3 (April 11)
  1. RNA-seq downstream analysis with nf-core differentialabundance pipeline (Shirley)
  2. Visualize outputs with shinyNGS (Shirley)
  3. Troubleshooting (Yucheng)

Correction

Local mode

```
#!/bin/bash

#SBATCH --time=00-48:00:00
#SBATCH -p batch
#SBATCH -N 1
#SBATCH -n 1
#SBATCH -c XX
#SBATCH --mem=XXG
#SBATCH --job-name nf-core
#SBATCH --output=%x-%J-%u.out
#SBATCH --error=%x-%J-%u.err
#SBATCH --mail-type=ALL
#SBATCH --mail-user=XXX@tufts.edu

module load nf-core
export NXF_SINGULARITY_CACHEDIR=/cluster/tufts/biocontainers/nf-core/singularity-images

nextflow run /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-rnaseq/3.14.0/3_14_0/ \
        --input samplesheet.csv --outdir output \
        --fasta ref.fasta --gtf ref.gtf --aligner star_salmon \
        -profile singularity \
        --max_memory XXGB --max_cpus XX
```

Office Hour

Apr 5 & 12, 1-3pm

- **Tisch Library, Room 208A**

RNA-Seq downstream analysis

nf-core differentialabundance

Key Steps in Differential Abundance Analysis

Similarities across differential analyses of matrices

Direct matrix input, retrieval from online resources

Distribution plots, PCA, examine impact of covariates

Remove uninformative features (all-0 rows) or observations.

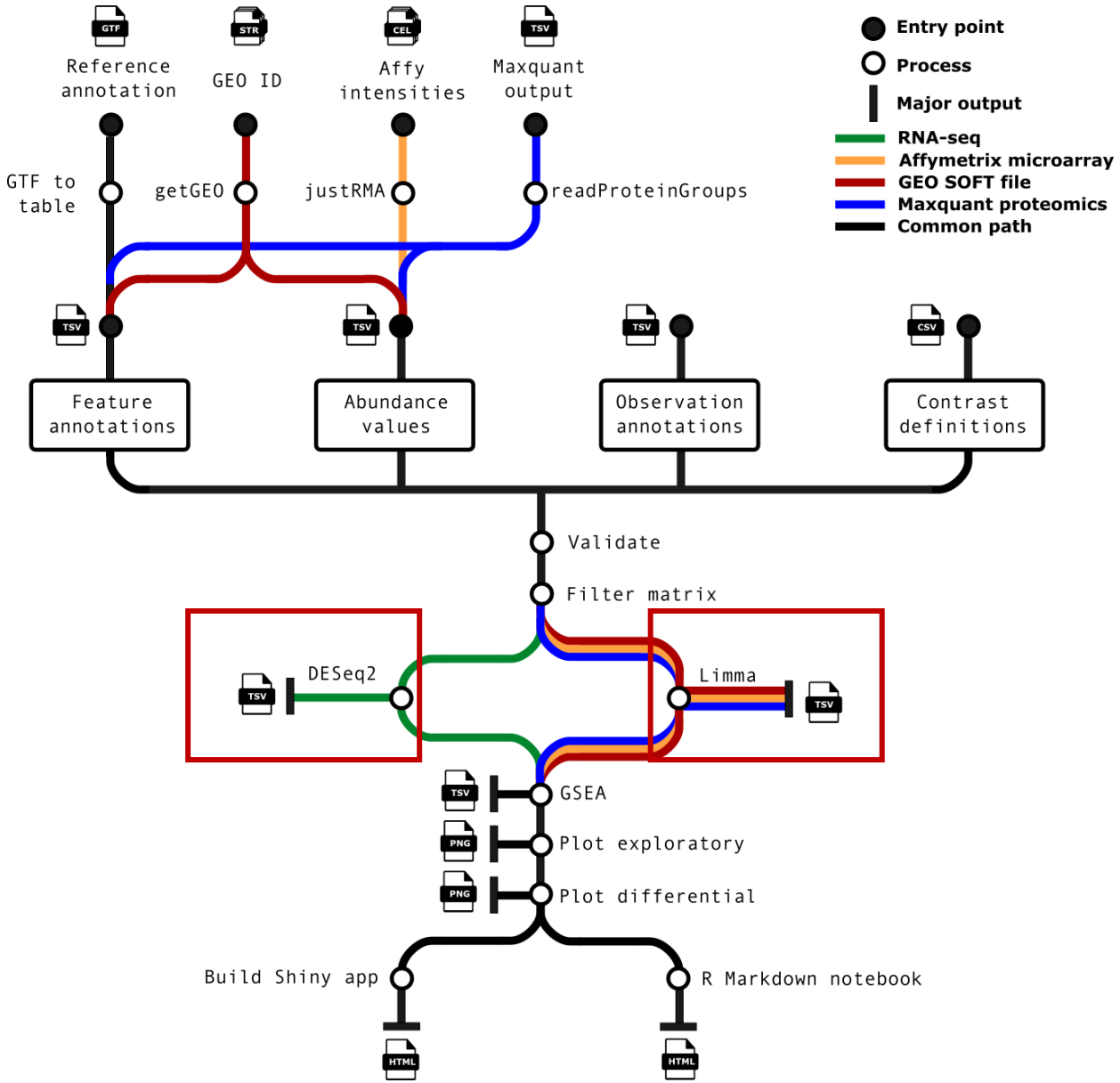
Derive features with significant differences in abundance between groups, accounting for covariates.

Record how analysis was done with graphical representations of key outcomes.

nf-core/differentialabundance

Steps:

- Matrix filtering + validation
- Exploratory analysis (e.g. PCA, boxplots)
- Run differential analysis over all contrasts specified.
- Gene set enrichment analysis (Optional)
- Reporting
  - R Markdown notebook (HTML)
  - Shiny app



DESeq2

https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

⚠ Welcome to the new bioconductor.org!

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

About | Learn | Packages | Developers | Search | Get Started >

Home > Bioconductor 3.18 > Software Packages > DESeq2

DESeq2

Differential gene expression analysis based on the negative binomial distribution

platforms all | rank 26 / 2266 | support 165 / 181 | in Bioc 11 years | build ok | updated before release | dependencies 69

DOI: 10.18129/B9.bioc.DESeq2

Bioconductor version: Release (3.18)

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

Author: Michael Love [aut, cre], Constantin Ahlmann-Eltze [ctb], Kwame Forbes [ctb], Simon Anders [aut, ctb], Wolfgang Huber [aut, ctb], RADIANT EU FP7 [fnd], NIH NHGRI [fnd], CZI [fnd]

Maintainer: Michael Love <michaelisaiahlove at gmail.com>

Citation (from within R, enter citation("DESeq2")):

Love MI, Huber W, Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, **15**, 550. doi:10.1186/s13059-014-0550-8.

Limma

limma powers differential expression analyses for RNA-sequencing and microarray studies

Matthew E. Ritchie[1,2], Belinda Phipson[3], Di Wu[4], Yifang Hu[5], Charity W. Law[6], Wei Shi[5,7] and Gordon K. Smyth[2,5,*]

[1]Molecular Medicine Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [2]Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia, [3]Murdoch Childrens Research Institute, Royal Children's Hospital, 50 Flemington Road, Parkville, Victoria 3052, Australia, [4]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138-2901, USA, [5]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [6]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland and [7]Department of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia

Received November 09, 2014; Revised January 04, 2015; Accepted January 06, 2015

Why not edgeR

Jonathan Manning 🐤 12 months ago

If you had a strong preference for edgeR and were to contribute an analagous edgeR module to nf-core at some point I'd be very happy to add it as an option to the workflow 🙂 .

Input files:

```
--input samplesheet.csv
--contrasts contrasts.csv
--matrix assay_matrix.tsv
--gtf mouse.gtf
```

Input: samplesheet.csv

- Similar to the samplesheet used for nfcore/rnaseq pipeline.
- Needs to add columns that describe the groups you want to compare.

```
sample,fastq_1,fastq_2,condition,replicate,batch
CONTROL_REP1,AEG588A1_S1_L002_R1_001.fastq.gz,AEG588A1_S1_L002_R2_001.fastq.gz,control,1,A
CONTROL_REP2,AEG588A1_S1_L003_R1_001.fastq.gz,AEG588A1_S1_L003_R2_001.fastq.gz,control,2,B
CONTROL_REP3,AEG588A1_S1_L004_R1_001.fastq.gz,AEG588A1_S1_L004_R2_001.fastq.gz,control,3,A
TREATED_REP1,AEG588A2_S1_L002_R1_001.fastq.gz,AEG588A2_S1_L002_R2_001.fastq.gz,treated,1,B
TREATED_REP2,AEG588A2_S1_L003_R1_001.fastq.gz,AEG588A2_S1_L003_R2_001.fastq.gz,treated,2,A
TREATED_REP3,AEG588A2_S1_L004_R1_001.fastq.gz,AEG588A2_S1_L004_R2_001.fastq.gz,treated,3,B
```

Input: contrasts.csv

The contrasts file references the observations file to define groups of samples to compare.

samplesheet.csv

```csv
sample,fastq_1,fastq_2,condition,replicate,batch
CONTROL_REP1,AEG588A1_S1_L002_R1_001.fastq.gz,AEG588A1_S1_L002_R2_001.fastq.gz,control,1,A
CONTROL_REP2,AEG588A1_S1_L003_R1_001.fastq.gz,AEG588A1_S1_L003_R2_001.fastq.gz,control,2,B
CONTROL_REP3,AEG588A1_S1_L004_R1_001.fastq.gz,AEG588A1_S1_L004_R2_001.fastq.gz,control,3,A
TREATED_REP1,AEG588A2_S1_L002_R1_001.fastq.gz,AEG588A2_S1_L002_R2_001.fastq.gz,treated,1,B
TREATED_REP2,AEG588A2_S1_L003_R1_001.fastq.gz,AEG588A2_S1_L003_R2_001.fastq.gz,treated,2,A
TREATED_REP3,AEG588A2_S1_L004_R1_001.fastq.gz,AEG588A2_S1_L004_R2_001.fastq.gz,treated,3,B
```

constrasts.csv

```csv
id,variable,reference,target,blocking
condition_control_treated,condition,control,treated,
condition_control_treated_blockrep,condition,control,treated,replicate;batch
```

Input: assay_matrix.tsv (salmon.merged.gene_counts.tsv)

Marcus Nygård / SDU 7 months ago

Hi 🙂 I would like to run the differential abundance pipeline on the output from the RNA-seq pipeline (version 3.12.0) . I am not very experienced in understanding the specifics of all the different output matrices, so I was wondering which of of the matrix files (.tsv) from the RNA-seq pipeline is best suited as the input for the differentialabundance pipeline?

18 replies

Jonathan Manning 7 months ago

For most people, and until we make some related extensions to the pipeline, the salmon_merged_gene_counts.tsv is the way to go.

🙌 1

```
[yzhang85@login-prod-03 rnaseqOut]$ ls
fastqc/  genome/  multiqc/  pipeline_info/  star_salmon/  trimgalore/
[yzhang85@login-prod-03 rnaseqOut]$ cd star_salmon/
[yzhang85@login-prod-03 star_salmon]$ ls -hl *.tsv
-rwxr-xr-x 1 yzhang85 biotools 2.7M Mar  2 22:05 salmon.merged.gene_counts.tsv*
-rwxr-xr-x 1 yzhang85 biotools 4.2M Mar  2 22:05 salmon.merged.gene_counts_length_scaled.tsv*
-rwxr-xr-x 1 yzhang85 biotools 4.2M Mar  2 22:05 salmon.merged.gene_counts_scaled.tsv*
-rwxr-xr-x 1 yzhang85 biotools 6.7M Mar  2 22:05 salmon.merged.gene_lengths.tsv*
-rwxr-xr-x 1 yzhang85 biotools 3.2M Mar  2 22:05 salmon.merged.gene_tpm.tsv*
-rwxr-xr-x 1 yzhang85 biotools  13M Mar  2 22:05 salmon.merged.transcript_counts.tsv*
-rwxr-xr-x 1 yzhang85 biotools  20M Mar  2 22:05 salmon.merged.transcript_lengths.tsv*
-rwxr-xr-x 1 yzhang85 biotools  14M Mar  2 22:05 salmon.merged.transcript_tpm.tsv*
-rwxr-xr-x 1 yzhang85 biotools 9.7M Mar  2 22:03 tx2gene.tsv*
```

Account for transcript length biases

- Differential isoform usage can cause significant differences in effective gene length across samples/treatment groups.
- Important to account for length biases during differential analysis.
- Raw counts, model lengths explicitly.

RNA-seq only

```
--matrix salmon.merged.gene_counts.tsv \
--transcript_length_matrix salmon.merged.gene_lengths.tsv
```

DESeq2 parameter: --deseq2_vs_method

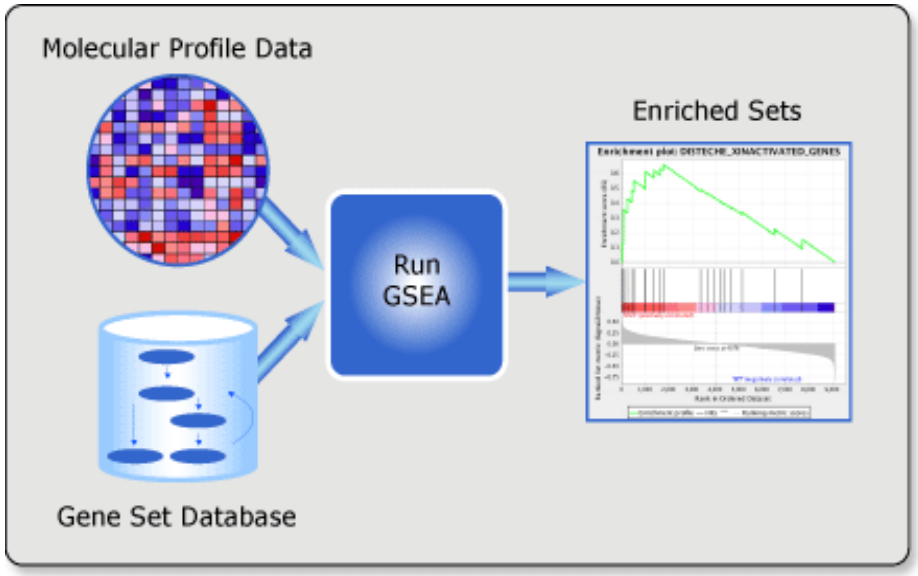
Data transformation for downstream analysis, such as clustering or PCA

- **vst (Variance Stabilizing Transformation)**
  - **Default**
- **rlog (Regularized Log Transformation)**
  - **Recommended!**

GSEA

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes, treatments).

- Rank genes according to their “correlation” with a class of interest
- Test if a gene set is enriched at the top or bottom of the list using a Kolmogorov-Smirnoff score



MSigDB

MSigDB
Molecular Signatures Database

Molecular Signatures Database

Overview

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into Human and Mouse collections. From this web site, you can

- **Examine** a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS human gene set page.
- **Browse** gene sets by name or collection.
- **Search** for gene sets by keyword.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.
  - Investigate the gene set in the online **biological network repository NDEx**
- **Download** gene sets.

License Terms

GSEA and MSigDB are available for use under these license terms.

Human Collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented expression data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Reports

- A html report file
- A ShinyNGS app

Report: static HTML

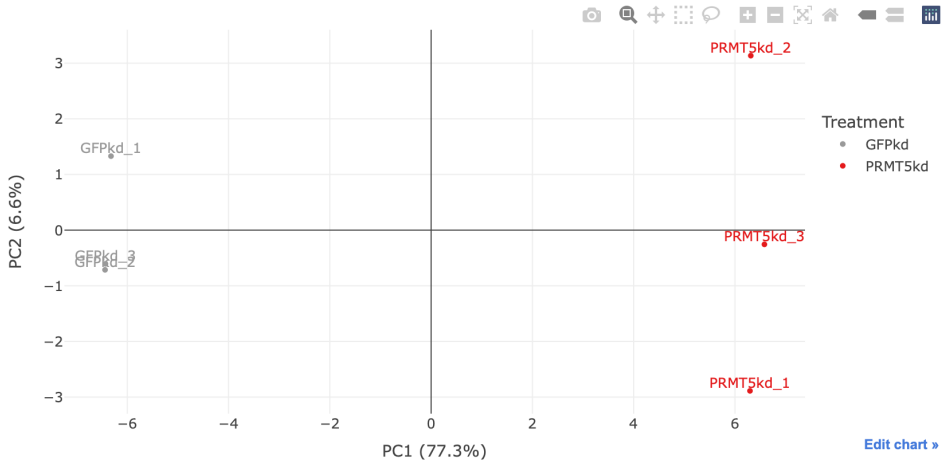
nf-core/ differentialabundance

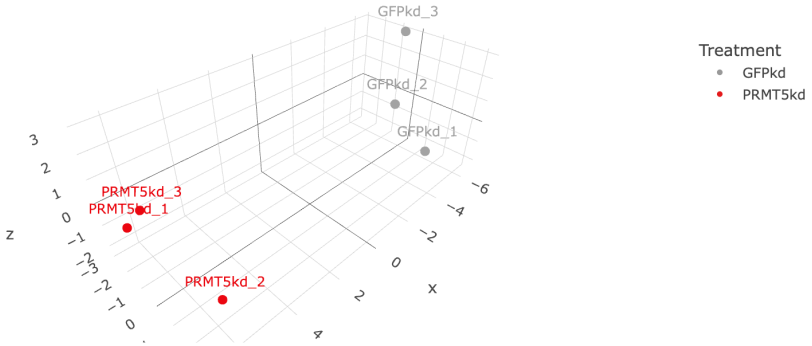
Abstract
Data
Results
Counts
Exploratory analysis
Abundance value distributions
Sample relationships
Principal components plots
Principal components/ metadata associations
Clustering dendrograms
Outlier detection
Differential analysis
Methods
Appendices
nf-core/differentialabundance: Citations

Principal components plots

Principal components analysis was conducted based on the 500 most variable genes. Each component was annotated with its percent contribution to variance.

Variance stabilised (treatment) | Normalised (treatment) | Raw (treatment)





Report: shiny app

https://ondemand.pax.tufts.edu/rnode/c1cmp040.pax.tufts.edu/38921/p/c5a87267/ | Open in Browser | Publish

RNA-seq explorer: PRMT5kd vs. GFPkd (Shiny Application) | Home | Sample data | QC/ exploratory | Assay data | Differential | Gene info

Contrasts

☑ Filter rows

Contrast(s): Treatment: PRMT5kd vs GFPkd

p value: <= 0.05

q value: <= 0.1

fold change: >= or <= - 2

Scatter plot

Color palette Dark2

☐ Show labels?

Point size 1 5 20

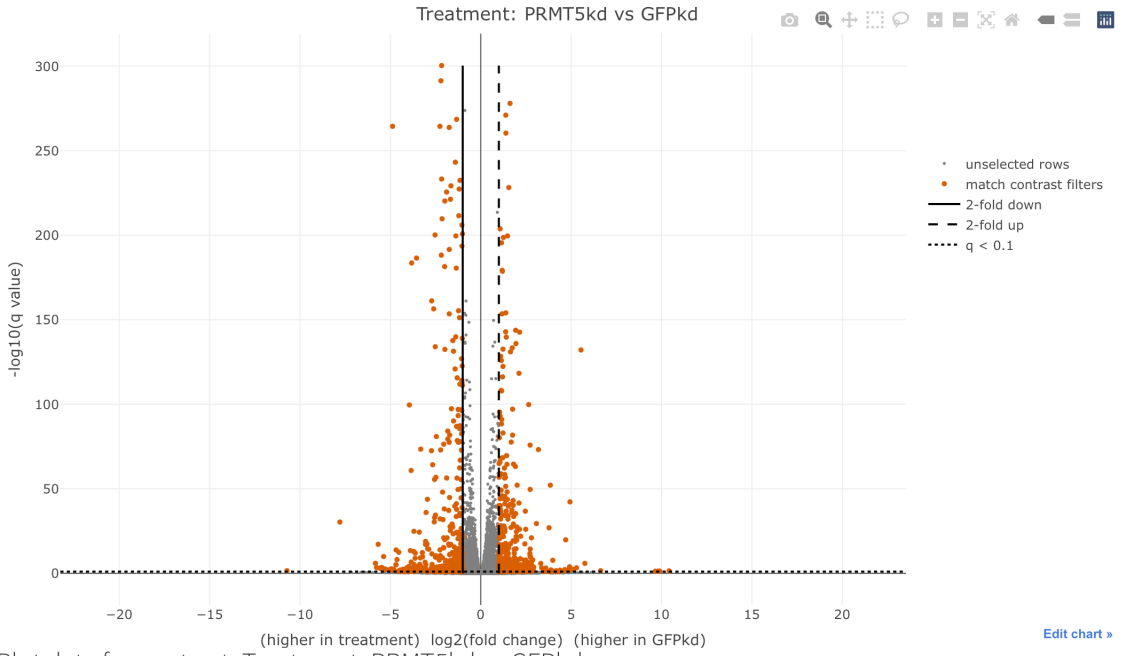
Highlight points

Select genes/ rows

Select genes by none

Volcano plot

help



Edit chart »

Plot data for contrast: Treatment: PRMT5kd vs GFPkd

Show 10 entries Search:

Gene id	Gene name	GFPkd	PRMT5kd	Fold change	p value	q value
ENSG00000000971	CFH	11.31	10.96	-3.7	4.4978e-59	4.3912e-57
ENSG00000001626	CFTR	10.65	10.57	-15.79	0.012047	0.052446
ENSG00000003137	CYP26B1	10.92	11.1	2.29	1.0875e-15	2.6082e-14

ShinyNGS R package

Synopsis

Shinyngs is an R package designed to facilitate downstream analysis of RNA-seq and similar expression data with various exploratory plots and data mining tools. It is unrelated to the recently published Shiny Transcritome Analysis Resource Tool (START), though it was probably developed at the same time as that work.

Examples

Data structure

A companion R package, zhangneurons, contains an example dataset to illustrate the features of Shinyngs, as well as the code required to produce it.

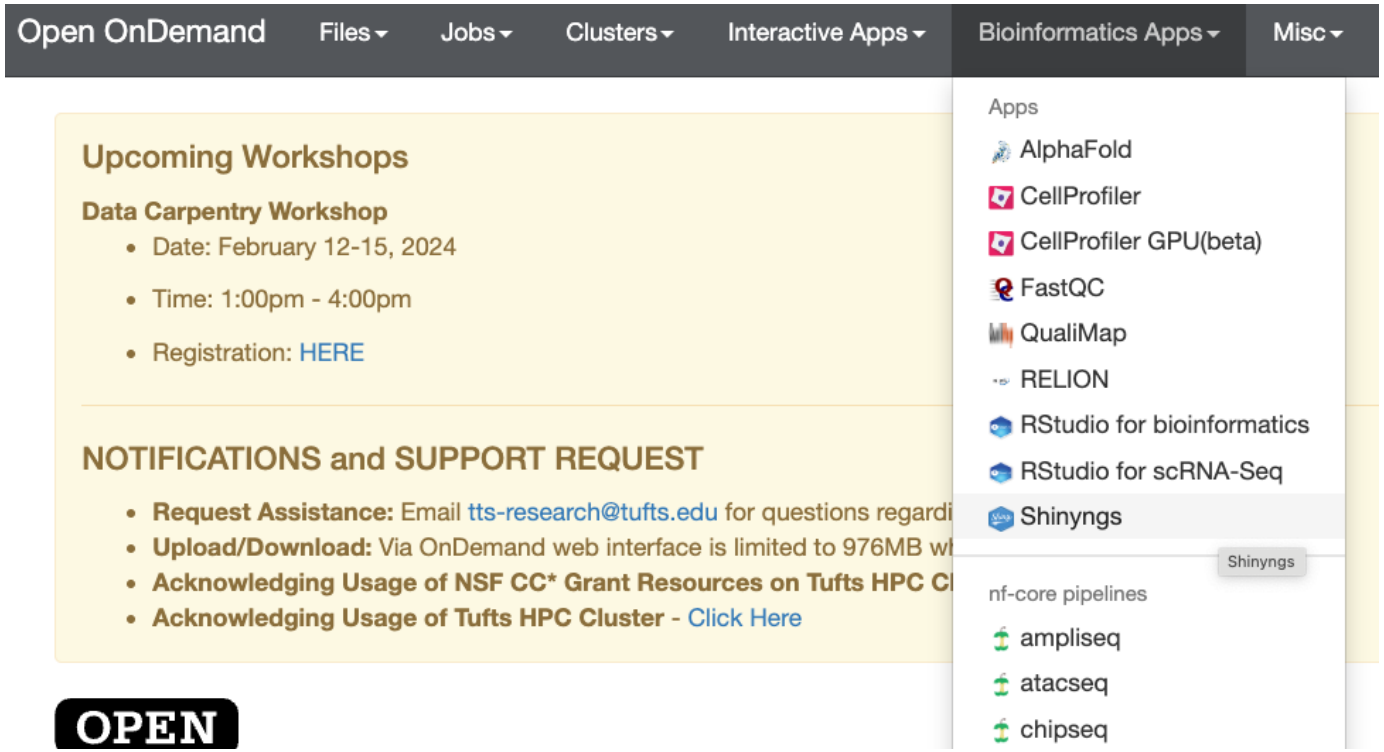
Running application

A Shinyngs example is running at https://pinin4fjords.shinyapps.io/shinyngs_example/ and contains a subset of the example data (due to limited resources on shinyapps.io).

https://github.com/pinin4fjords/shinyngs

ShinyNGS on Open OnDemand

Shinyngs is an R package designed to facilitate downstream analysis of RNA-seq and similar expression data with various exploratory plots and data mining tools.



Intro to nextflow and nf-core

Troubleshooting

Start small

```
[[yzhang85@login-prod-03 ~]$ srun -N1 -n4 -p batch -t0-1 --pty bash
srun: job 2245475 queued and waiting for resources
srun: job 2245475 has been allocated resources
[[yzhang85@p1cmp045 ~]$ module load nf-core-chipseq/2.0.0
[[yzhang85@p1cmp045 ~]$ chipseq -profile test,singularity --outdir testout
N E X T F L O W  ~  version 23.04.4
Launching `/cluster/tufts/biocontainers/nf-core/pipelines/nf-core-chipseq/2.0.0/2_0_0/main.nf` [friendly_lagrange] DSL2 - revision: 7341307235


------------------------------------------------------
                                          ,--./,-.
          ___     __  __  __   ___     /,-._.--~'
    |\ | |__  __ /  ` /  \ |__) |__         }  {
    | \| |       \__, \__/ |  \ |___     \`-._,-`-,
                                          `._,._,'
    nf-core/chipseq v2.0.0
------------------------------------------------------
Core Nextflow options
  runName                   : friendly_lagrange
  containerEngine           : singularity
  launchDir                 : /cluster/home/yzhang85
  workDir                   : /cluster/home/yzhang85/work
  projectDir                : /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-chipseq/2.0.0/2_0_0
  userName                  : yzhang85
  profile                   : test,singularity
  configFiles               : /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-chipseq/2.0.0/2_0_0/nextflow.config

Input/output options
  input                     : https://raw.githubusercontent.com/nf-core/test-datasets/chipseq/samplesheet/v2.0/samplesheet_test.csv
  read_length               : 50
  outdir                    : testout

Reference genome options
  fasta                     : https://raw.githubusercontent.com/nf-core/test-datasets/atacseq/reference/genome.fa
  gtf                       : https://raw.githubusercontent.com/nf-core/test-datasets/atacseq/reference/genes.gtf

Process skipping options
  skip_preseq               : true

Institutional config options
  custom_config_base        : /cluster/tufts/biocontainers/nf-core/pipelines/nf-core-chipseq/2.0.0/2_0_0/../configs/
  config_profile_name       : Test profile
  config_profile_description: Minimal test dataset to check pipeline function

Max job request options
  max_cpus                  : 2
  max_memory                : 6.GB
  max_time                  : 6.h

Generic options
  fingerprint_bins          : 100
```

Check the basics

- Whether nextflow version is too old
- Whether required modules are loaded (nextflow and singularity)
- Haven’t run out of disk space (du -f)

Check the troubleshooting docs:

- https://nf-co.re/docs/usage/troubleshooting

Anatomy of a work directory

- **.command.out** - STUOUT from tool
- **.command.err** – STDERR from tool
- **.command.log** - STOUT and STDERR from tool
- **.command.run** – Wrapper script used to run the job
- **.command.sh** – Process command used for this tasks
- .command.begin – Created ASAP the jobs launches
- .command.trac – Logs of computer resource usage
- .exitcode – Created when the job ends, with exit code

Seek help from nextflow and nf-core communities

slack

Join Nextflow on Slack

Start by entering the email address you use for work.

your-email | @nextflow.io

Continue

You can use any account with the domain:

- nextflow.io
- seqera.io

Don't have an email address from one of those domains?
Contact the workspace administrator at **Nextflow** for an invitation.

slack

See what nf-core is up to

Slack is a messaging app that brings your whole team together.

Marcel Ribeiro-Dantas, Phil Ewels and 8,382 others have already joined

We suggest using **the email account you use for work**.

Continue With Google

Continue With Apple

Continue With Email