Keeping your Data FAIR Organizing, Managing, and Sharing your Data



Peter FitzGerald, PhD

Outline

- Principles of FAIR
- The practical aspects of adhering to a FAIR approach
- Data Management Strategies
 - Record Keeping, Storage, file formats
- Sharing
 - What, Where, When
- Repositories
- Journal Requirements
- Example RNA-Seq project
- Resources

FAIR Principles

https://www.go-fair.org/fair-principles/



indable: The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.



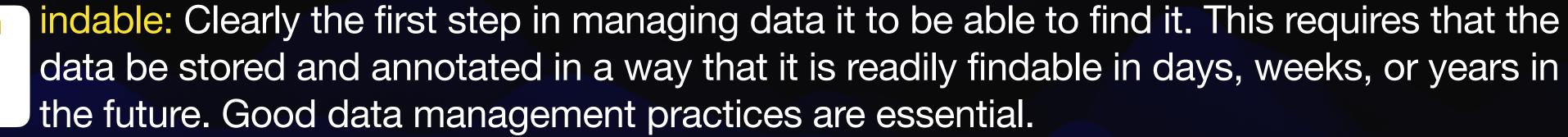
ccessible: Once the user finds the required data, she/he/they need to know how they can be accessed, possibly including authentication and authorisation.

nteroperable: The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.



eusable: The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

FAIR Principles In practice in your local environment (computer)





ccessible: Working in the NIH environment it is important that the data is accessible by you, your PI and future lab members who may need to expand on your work. This dictates that all *important* data reside somewhere other than you personal laptop computer.

nteroperable: This usually means staying away from proprietary data formats (that may not be supported in the future) and additionally, the data need to interoperate with applications or workflows for analysis, storage, and processing.



eusable: In this case reusable could better be defined as reproducible. Adhering to good data management principles means that you (or others) will be readily able to reproduce your data workflow resulting in exactly the same results.

FAIR Principles In practice for the Scientific Community



indable: This again relates to good annotation using appropriate terminology and depositing the data in public database. Meaningful and accurate metadata is a must.



ccessible: In accordance with the various rules/policies the data needs to deposited in the appropriate public repository.

nteroperable: Again referring to the metadata, annotation should follow standards that make further analysis possible.

Reusable: The primary data and associated metadata, and custom software and methodology should be deposited in a way that allows others to successfully reproduce your experiment.

All the FAIR principles can only be achieve with good data management practices.

NIH has issued the Data Management and Sharing (DMS) policy (effective January 25, 2023) to promote the sharing of scientific data.

This means that all NIH PIs have submitted a data management and sharing plan in January of this year (2023). If you aren't aware of this you should discuss it with our PI to be aware what they have committed themselves (and by inference you) to.

NIH also has a Genomic Data Sharing Policy which has been in place since 2014

Learn more about NIH Sharing Policies at https://sharing.nih.gov/





Data Management Strategy

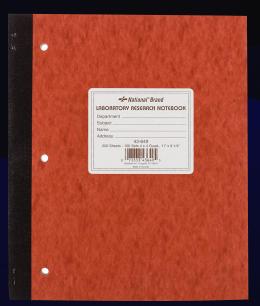
- All data in organized and stored in a way that it can be readily located by YOU your Pl and future members of the Lab.
- Data is stored in a safe and secure location •
- Metadata is capture in a timely manner •
- Key aspects of software-use and reference material is capture early on \bullet

Sharing made easy

A good data management plan will ensure that complying with NIH data sharing policies is straight forward since all of the elements have already been captured and are readily locatable.

Hopefully by now it is clear that each one of us should have a data managrement strategy that will ensure that:

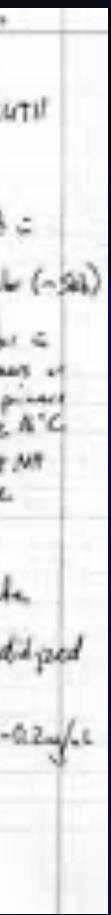
Record Keeping Traditional vs Electronic



Traditionally the "Lab Notebook" was where everything was recorded and documented each and every experiment was recorded. There are well established principles of what Notebook". Generally well established procedures are in place for the storage (long or short) for lab reagents (cell-lines clones, etc) that might need to be share with

was and wasn't necessary to go into the "Lab the Scientific Community.

Level 10 post of SULTICID & GRATH 2. 31-CPP PD21 + B24 5 cm printing of 12 17924 were (-94) Carphi and Statester LB CURVE RAIN CONSTRACT all ING the Sec. 24. Company tractions & dive 174 an frist PBS BCA Care it bash - a Znight



Record Keeping Traditional vs Electronic

Unfortunately, in the current Electronic age there is no clear replacement for the "Lab Notebook". There are numerous instances of "Electron Lab Notebooks", but at the current time none of them have become the defect standard and NCI is only now moving in that direction, with there "beta progam. Additionally, modern research methods combine traditional "wet lab science" with many techniques that are purely computational. This has introduced many new challenges for the researcher who has to store, manage and ultimately share this purely electronic data and its associated metadata. (Metadata being defined as information about data).

Electronic Lab Notebook or Equivalent

If your Lab has a defined plan for managing electronic data you should follow it. If the Lab doesn't have a formal plan for file management find a system that works for you and stick with it!! The resource section has links to sites that suggest some strategies.

The most important parts of a plan are to capture all the relevant information and store them in a way that is secure and findable

Some Options

Formal Electronic Lab Notebook One Note - Rigid Hierarchy of named directories and Files Word Documents PDF Documents Any strategy will likely dictate that all data, associated metadata and documents should be stored in organized defined file/folder hierarchies

Elements of a good file management plan

- Early on document (and revise as needsed) your organization strategy so you (and others) know what goes where.
- Some level of Index
- Good structured file naming conventions (e.g. Illumina sequencing run name 230623_VH00687_218_AACV5NMM5)
- Organized file/folder hierarchy e.g. by time YYYY-MM-DD or YYYYMMDD, by file type, by experiment or project.
- File Hierarchies should not got too long.
- Use open formats where possible text files, css files, etc.
- For tabular data make sure all columns have meaningful headers
- Separate primary (raw) data from derived data and protect if accordingly



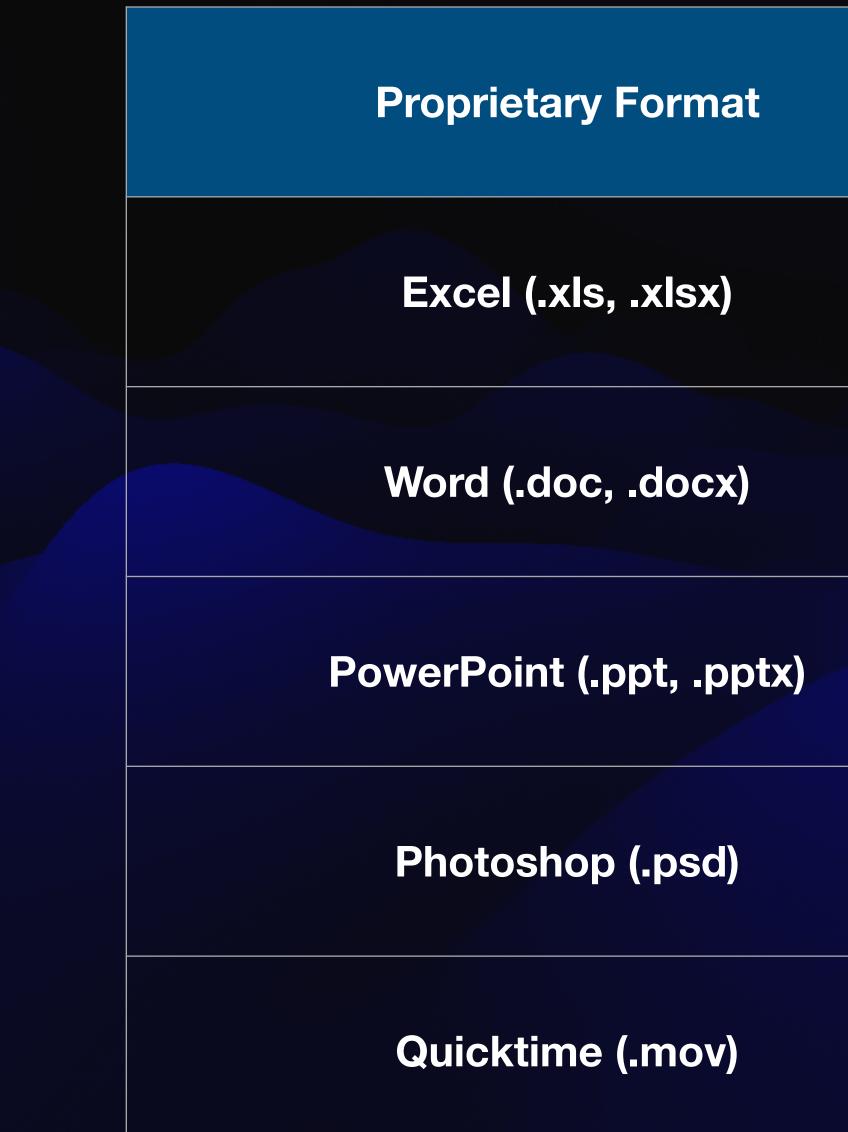


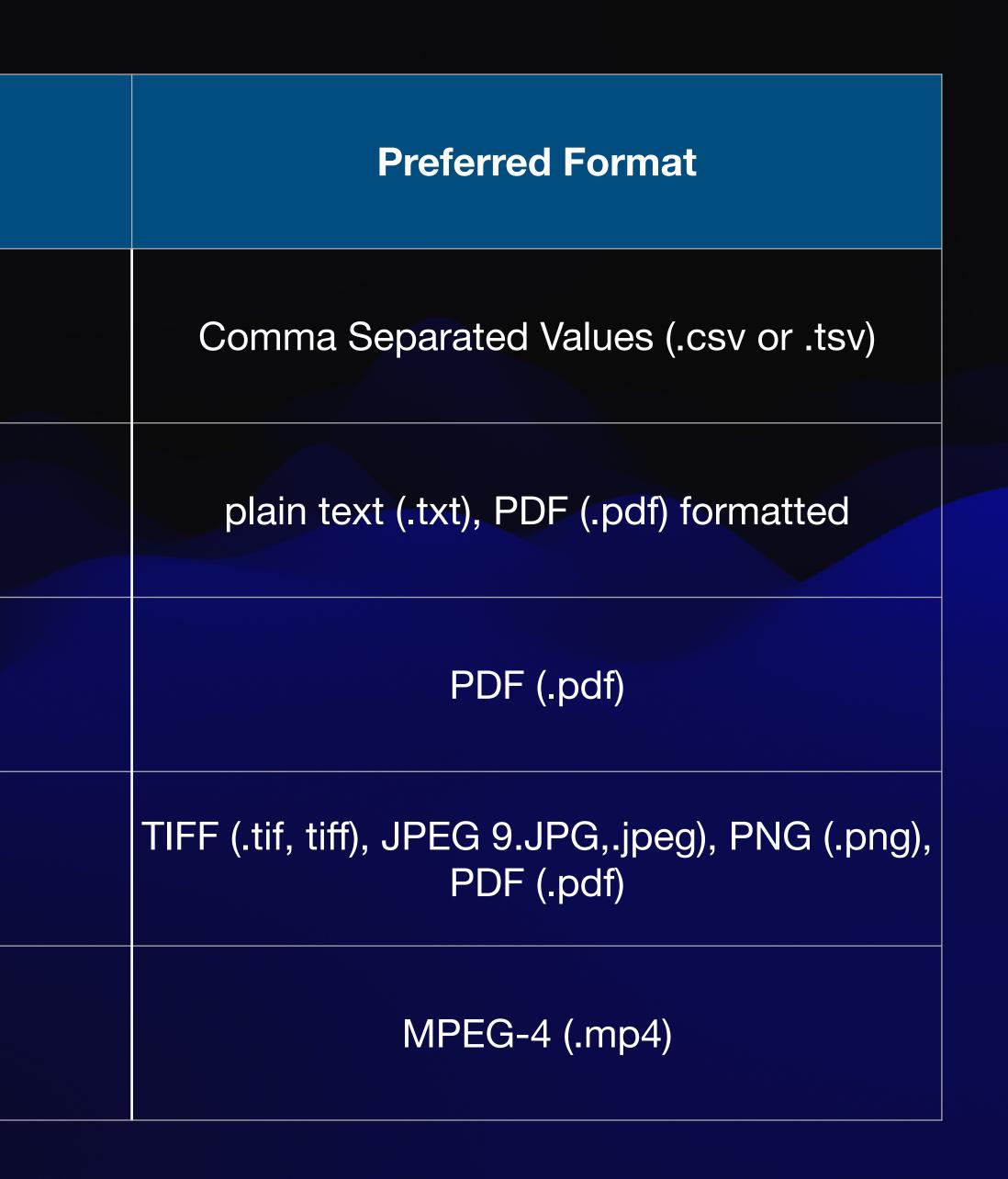
Where and how to store all files

- Decide what can be stored "locally" and what must be put in "permanent" storage" - working copies vs raw data or final product
- Separate raw data from derived data
- Devise a system to distinguish data exploration and side tracks from the final • results
- Use structured naming conventions (many to chose from) adopt one and stick to it
- Plan how will you share the data etc. with others in the lab, and future \bullet members

Plan - Demands discipline







Computational Biologists

Computational experiments occur on a different time scale than traditional wet lab experiments and tend to generated more output. While some tasks are computationally time intensive and take long periods to complete. Others can be run in a matter of minutes, where different approaches, programs or parameters are explore. This avalanche of data (some good some bad) greatly complicates the data management issues.

An added problem for those involved in the field of bioinformatic is the distributed nature of computing. Some compute intensive tasks may be done on the HPC system (Biowulf/Helix or FRCE) or a Cloud resource (Galaxy, DNAnexus, PartekFlow) while others may be done on a local desktop or laptop machine

Some software/workflows impose their own structure on the data, so the challenge is how to fit into your structure and or move the files





File Structures

- Organize your data hierarchically, and identify ways to divide your data into categories (or attributes):
 - Project
 - Time
 - Location
 - File type
- Within folders, files can be maintained chronologically, by classification or code, or alphabetically (depending on the types of files)
- Folder and subfolder names should reflect the content of the folder, not the names of researchers or staff
- Document your file directory structure and describe the kinds of records that should be maintained in those folders to ensure compliance
- Include basic information, such as project titles, dates, and some type of unique identifier (such as a grant number)
- If appropriate develop a strategy for keep track of files on remote systems
- Remember to keep PII data separately and secure

Data Storage

- Desktops/laptops only home directory routinely backed up, not shareable poor choice
- Networked drives personal or shared Lab space should be secure
- Lab server is it backup up
- External hard drives -
- Optical storage typically too slow and too small \bullet
- <u>Cloud storage</u> -
- Flash drives -
- HPC computer (Biowulf/Helix) stable but NOT backed up does have snapshots \bullet
- \bullet etc)
- critical data.

The 3-2-1 methodology - hard to achieve in NIH environment

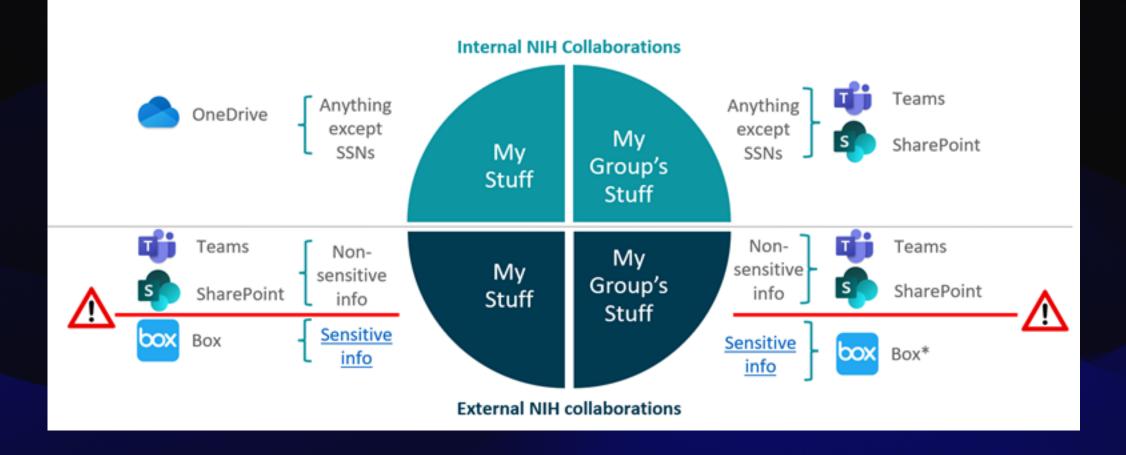
A simple, commonly used storage system is the 3-2-1 methodology. This methodology suggests the following strategic recommendations: 3: Store three copies of your data, 2: using two types of storage media, 1: with one of them stored offsite.

The dangers that need to be mitigated include: Accidental erasure, disk failure, hacking and data being *lost*

Many "good solutions" not appropriate for NIH security policies - (IT frowns on iCloud, Google Drive, Dropbox

Be aware of your "*permission*" choices and where possible limit others to "*read only*" and even yourself for

Data Storage



Documents that contain sensitive information

Documents that do not contain sensitive information

https://service.cancer.gov/ncisp?id=nci_kb_article&sysparm_article=NCI-KB0014654

Internal (NIH) Collaborators	External (non-NIH) Collaborators
 <u>Teams</u> <u>SharePoint Online</u> 	• <u>Box</u> *
 <u>Teams</u> <u>SharePoint Online</u> 	 <u>Teams</u> <u>SharePoint Online</u>

Data Storage

Features	OneDrive	Teams	SharePoint Box		Network Storage / Shared Drive	
Available without NIH network connection or VPN	yes	yes	yes	yes	no	
Share sensitive information with external collaborators	no	no (but probably in the future)			no	
Share non-sensitive information with external collaborators	no	no yes yes no		no	no	
Individual File Size Limit	250GB	250GB 250GB		150GB	100GB+	
Total File Storage	5TB	25TB per team	25TB per site unlimited		25TB+	
Sharing and Access Control			yes yes		limited	
Real-Time Co-Authoring	yes	yes	yes	yes	no	
Syncing	yes	yes	yes	yes	no	
Versioning	yes	yes	yes	yes	no	
Restore	yes	yes	yes	yes	no	
Desktop App	yes	yes	yes	yes	yes	
Web & Mobile App	yes	yes	yes	yes	no	
Conversations	no	yes	no	no	no	
Chat	no	yes	no	no	no	
Voice/Video	no	yes	no	no	no	
Other Info	Additional Microsoft ap OneNote, Wiki, Recordi					

Data Management Environment (DME)

The NCI Data Management Environment (DME) platform for the storage and management of scientific research datasets. It eliminates the need to maintain redundant copies of large heterogenous data and provides the ability to annotate, retrieve, and share datasets for further research, analysis, and collaboration. *Most of the primary data generated by CCR cores is currently deposited into DME.*

If you have an NIH account, the NCI Data Vault team can give you access to DME. For access requests or any other questions, contact NCIDataVault@mail.nih.gov. The user guide found at https://wiki.nci.nih.gov/display/DMEdoc describes the capabilities of DME.

The system can be accessed via a Command Line Interface (CLI), a web-based graphical user interface (GUI), as well as programmatically via a set of API calls and utilities.

Reasons for Sharing Data

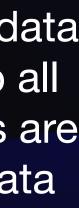
There are a number of policies which mandate sharing of different data.

Under the Data Management & Sharing (DMS) Policy, NIH requires researchers to prospectively plan for how scientific data will be preserved and shared through submission of a Data Management and Sharing Plan. The DMS policy applies to all NIH-supported research that results in the generation of scientific data, regardless of funding mechanism. Investigators are required to: Submit a Data Management and Sharing plan outlining how scientific data and any accompanying metadata will be managed and shared, taking into account any potential restrictions or limitations. Comply with the Data Management and Sharing plan approved by the funding Institute or Center (IC). January 2023

NIH expects the broad and responsible sharing of human as well as non-human genomic data resulting from NIH-funded research because the timely sharing of research results can accelerate discoveries that improve our ability to diagnose, treat, and prevent disease.

It is good science... just like publishing is important to share your research sharing data is also critical in so much as it permits other to validate and/or build on your results

Your ability to comply with these policies will be directly effected by your data management strategies







Sharing

- What to Share
- When to Share
- Where to Share

The simple answer is everything that is needed to reproduce your finding, in a public repository, as soon as possible.

Controlled Access Data

Data is in public repositories but access is restricted to "authorized" investigators. Primary data from human subjects typically falls into this category (e.g. dbGap) - remember to anonymize patient data

Public Access Data

Data is deposited in open repositories which allow access with out restrictions. Most basic research data falls into this category (e.g. Genbank)

Journal Policies

- All data deposited at time of Submission
- All data available to reviewer (upon request) •
- •

Types of data that must be shared

- Raw data
- Processed data
- Workflows
- Software (license type)

All Data available upon approval - i.e. before it hits the public



Mandatory deposition

Protein sequences

DNA and RNA sequences

DNA and RNA sequencing data

Genetic polymorphisms

Linked genotype and phenotype data

Macromolecular structure

Microarray data (must be MIAME compliant)

Crystallographic data for small molecules

Genomic Repositories

Suitable repositories
<u>Uniprot</u>
Genbank
DNA DataBank of Japan (DDBJ)
EMBL Nucleotide Sequence Database (ENA)
NCBI Trace Archive
NCBI Sequence Read Archive (SRA)
dbSNP
dbVar
European Variation Archive (EVA)
dbGAP
The European Genome-phenome Archive (EGA)
Worldwide Protein Data Bank (wwPDB)
Biological Magnetic Resonance Data Bank (BMRB)
Electron Microscopy Data Bank (EMDB)
Gene Expression Omnibus (GEO)
ArrayExpress
Cambridge Structural Database

Requirement	Dataverse	Dryad	figshare	GigaScience	Mendeley Data	OSF	Vivli	Zenodo
Data Size and Format								
Hosting of common file formats (e.g. csv, tsv, xls, xlsx, doc, pdf)	0	0	0	0	0	0	0	0
Hosting of proprietary file formats (e.g. raw image files)	0	0	0	×	0	0	0	0
Unlimited size per file	×	0	×	0	×	×	0	×
Unlimited total dataset size	0	0	0	0	0	0	0	0
Data Licensing								
CC0 waiver	recommended	required	recommended	required	available	available	available	available
Data Attribution and Citation Tools								
Assignment of dataset DOIs	0	0	0	0	0	0	0	0
User Access Controls								
Tiered access (e.g. administrator-level, collaborator-level, curator-level)	0	×	0	×	0	0	0	×
Journal-integrated, anonymous access (for peer review pre-publication)	0	0	0	0	0	0	×	×
Optional embargo to data release following publication	×	0	0	0	0	0	0	0
Data Access Tools								
Comprehensive data and metadata search tools	0	×	×	×	0	0	0	×
Data access via direct download	Ø	0	0	Ø	0	0	0	0
Data downloading via API	0	0	0	×	0	0	×	0
Built-in tools for reading proprietary file formats	×	×	0	×	×	×	0	×
Integrated data analysis tools	Ø	×	×	Ø	×	×	0	×
Cost								
Data deposition fees	none	tiered	none	none	none	none	membership	none
Data maintenance fees	none	none	none	none	none	none	membership	none

General Repositories

Software

Existing Software

- Proprietary
- Open Source
 - Version
 - Parameters
 - Reference
 - Command line example

Home Grown

- Related libraries and versions
- Documentation
- Source or Binary
- License
- Test data set
- Operating system
- Language
- Share code on <u>https://github.com/</u>
- Use Version Control

Open Source Workflow

A simple reference to a workflow may substitute for a laundry list of individual components

Command line interface (CLI)

- Harder to use
- Easier to document
- Check out the "shell's history options"
- Use scripts rather then typed commands
- On biowulf use explicit module versions (e.g. module load samtools/1.17
- e.g. Star, samtools

Graphic User Interface (GUI)

- Easier to use
- Harder to Document
- e.g. Partek Flow, Geneius
- Interactive computing platform. Integrated development environment (IDE)
- Varies by use
- Easier to document (automatic history)
- e.g. Jupyter Notebook, R Studio

Web Interface

- Easier to use
- Variable degree of automatic documentation
- e.g. NIDAP, DNAnexus

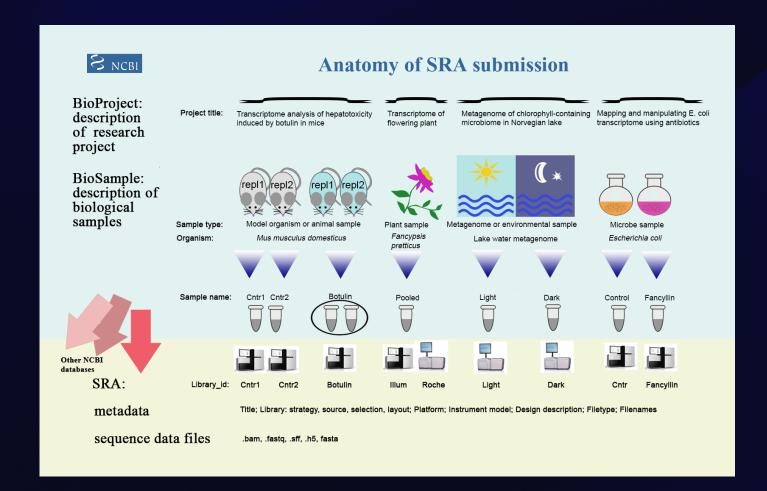
Software Licences

CCO 1.0 CCO 1.0 Universal (CCO 1.0) Public Domain Dedication https://creativecommons.org/publicdomain/zero/1.0/ CC BY 4.0 Attribution 4.0 International (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/ Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. CC BY NC 3.0 Attribution-NonCommercial 3.0 Unported (CC BY-NC 3.0) https://creativecommons.org/licenses/by-nc/3.0/ NonCommercial — You may not use the material for commercial purposes. MIT Apache 2.0 https://www.apache.org/licenses/LICENSE-2.0 BSD 3-Clause BSD 2-Clause GPL v3 - GNU General Public License https://www.gnu.org/licenses/gpl-3.0.en.html GPL v2 LGPL MPL-2.0, CeCILL CeCILL-B CERN OHL

Metadata (a set of data that describes and gives information about other data.)

Primary Data (SRA/dbGAP - Input Request) - FASTQ

- Experiment type (RNASeq, WGS, ChIPSeq)
- Organism (Human, Mouse)
- Cell Line
- Instrumentation (Illumina, nanopore)
- Machine Version (NextSeq 550, 2000, NovaSeq)
- Library Prep PCR amplified
- Sample Type whole RNA or mRNA etc
- Other factors rRNA depletion method



https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/

• Software

- Version
- Parameters
- Reference
- Command line example

Workflow

- Genome Version sequence
- Genome Version Annotation
- Pathway Versions
- Naming conventions (Proper gene name, IDs etc)



RNASEQ Workflow - Software

FASTQ

- Experiment type RNASeq
- Organism Human
- Cell Line MCF7
- Instrumentation Illumina,
- Machine Version NextSeq 2000
- Library Prep PCR amplified
- Sample Type mRNA
- Other factors -rRNA depletion
- RNAQuality RIN 8.9
- Sample Concentration 300 ngul
- Sample Volume 10ul
- Read Length 2x100
- Sample ID 220303_A00430_0573_BHWTW7DRXY
- Run Date 03/03/2022

BCLs - FASTQs - SAMs - BAMs - Count Matrices (Gene/Transcript) - DEG - Function/Pathway Enrichment

- BAM
- Experiment type RNASeq
- Organism Human
- Genome Version GRCh38 (NCBI)
- Genome Annotation Gencode 30
- Run Date 03/06/2022
- Aligner STAR (Parameters should be in bam file) (Version -)
- Extra Picard Markduplicates

(Version -) (Parameters -)

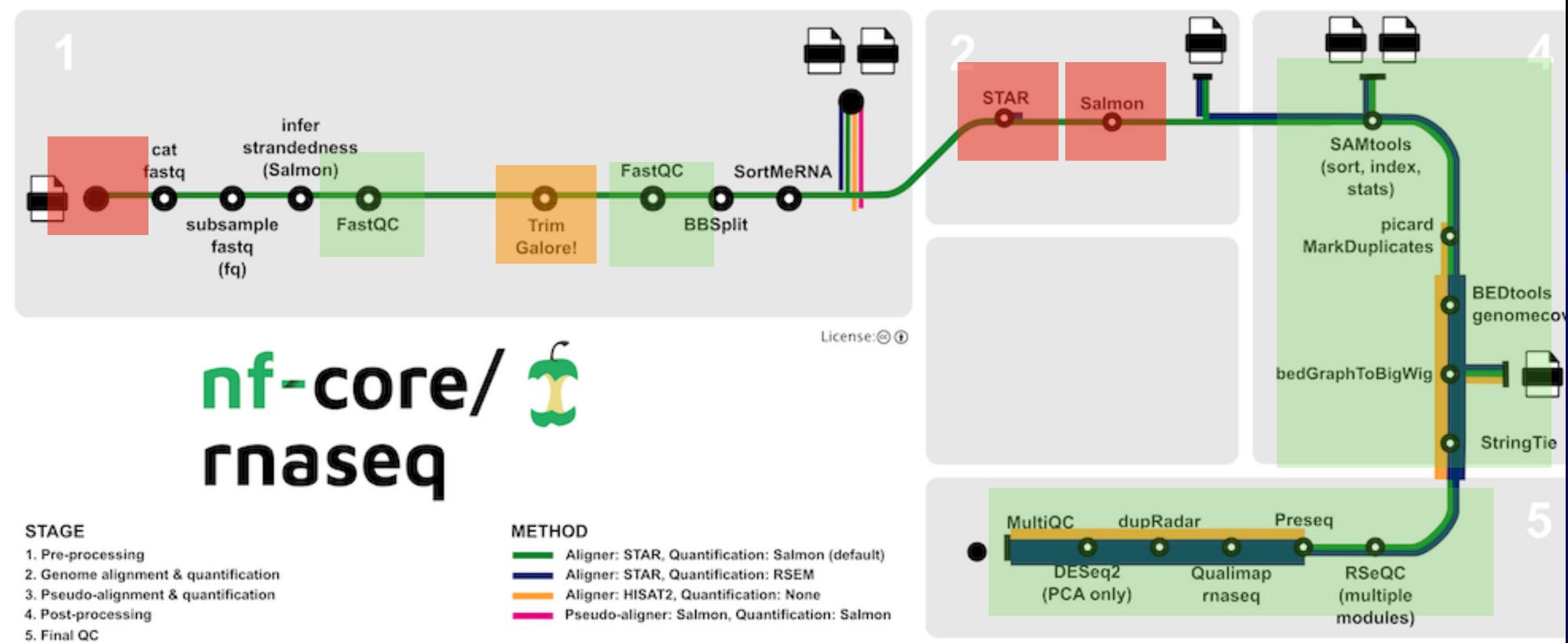
RNASEQ Workflow - File types

 Count Matrics Input source - from aligned bam file Software - salmon -version, parameters Gene/Transcripts Genome Annotations - Gencode 30

BCLs - FASTQs - SAMs - BAMs - Count Matrices (Gene/Transcript) - DEG -

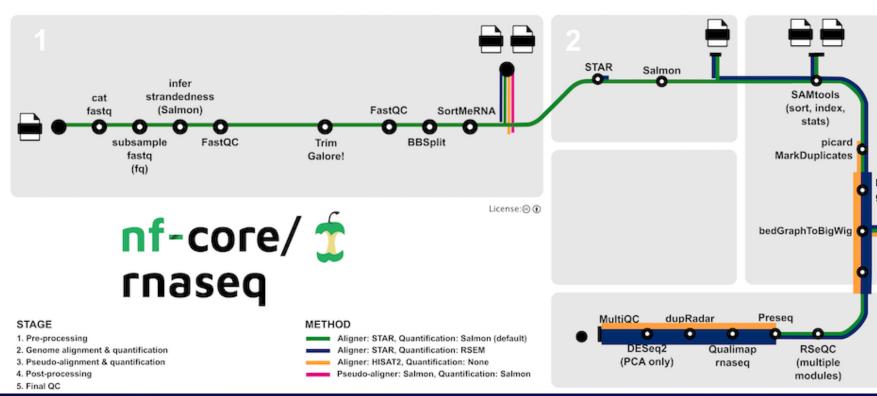
 DEG (Differentially Expressed Gene List) Input source - Raw count matrices Software - R Studios (2022.02.0 Build 443) - EdgeR version 3.36.0 within R version 4.3.1 Normalization method - TMM Cutoff Parameters ->2 fold, P < 0.05

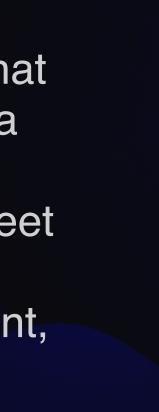




IEETODIS GENMAECOV bedrauls 2.38.0 CUSTOM DUMPSOFTMARE/ERSIDNS mythan 3.111.0 VAINL 6.0 CUSTOM CUMPSOFTMARE/ERSIDNS mythan 3.111.0 DEGO2 QC STAR SALMON blocanductor-desco2 1.78.0 DEFGO2 QC STAR SALMON blocanductor-desco2 1.78.0 CUSTOM CONTRACTOR r-base 4.2.1 PASIDC fastor 0.11.0 GIPRADAR blocanductor-desco2 0.11.0 GIPRADAR blocanductor-desco2 0.11.0 GURTE CTF fastor 0.11.0 GUNZIP CTF dyu210 1.10 GUNZIP CTF dyu210 1.18 RAME RANSED distard 2.7.10 PICARD MARKOUPLICATES ofsard 3.6.1 RAME RANSED dualiman 2.2.2 SEGOE MARTIN resear 3.6.1 RESCO INFERENTERINTION resear 3.6.1 RESCO RANDOTATION resear 3.6.1 RESCO INFERENTERINTINN resear	Process Name	Software	Version
CUSTON DUPPSOFTNAREVERSIONS python 3,11,0 Vanit Yanit 6,8 CUSTON DUPPSOFTNAREVERSIONS bioconductor-desca2 1,28,0 DESCO2_DC_STAR_SALMON bioconductor-dupradar 1,28,0 DUPADDAR r-base 4,2,1 DUPADDAR r-base 4,2,1 STOC fastar 4,2,1 CMUZIP_FASTA outration 3,9,5 GUPADDAR outration 3,9,5 GUPADDAR outration 3,9,5 GUPADDAR outration 3,9,6 GUPADDAR outration 3,9,0 GUPADDAR reseac 3,9,1 RESCO_INPERSIMENT reseac 3,9,1			
vant6.0CUSTON GETCHARONSIZESactichmensizes1.16.1DESEQU CK STAR SALPONbisconductor-desea21.28.0r-baser-base4.0.3DUPRADARbisconductor-desea21.28.0r-base4.7.11.28.0r-base4.7.1FASIOCfastac0.11.3GTZ REDper l5.26.2GTZ GEDper l5.26.2GTZ GEDgunzio1.10GMZIP FASTAunzio1.30NAKE TRANSCRIPTS FASTArsem1.3.1STARstar2.7.40PTCARD MARKOUPLICATESjitard3.0.0QUALTAME RANSEDunulimap2.2.2RESCO LINTERXPERIMENTrseac3.0.1RESCO LINTERXPERIMENTrseac3.0.1RESCO LINTERXPERIMENTrseac3.0.1SEGO LINTERXPERIMENTrseac3.0.1SEGO LINTERXPERIMENTrseac3.0.1SEGO LINTERXPERIMENTrseac3.0.1SEGO LINTERXPERIMENTrseac3.0.1SALDON LOWANTrseac3.0.1SALDON LOWANTsalmon1.20.3SALDON LOWANTsalmon1.20.3SALDON LOWANTsalmon3.9.5SALDON LOWANTsalmon1.26.1SALDON LOWANTsalmon1.26.1SALDON LOWANTsalmon1.26.1SALDON LOWANTsalmon1.26.1SALDON LOWANTsalmon1.26.1SALDON LOWANTsalmon1.26.1SALDON LOWANTsalmon <td></td> <td></td> <td>3.11.0</td>			3.11.0
DESEQ2_OC_STAR_SALMON biaconductor-deseq2 1,28,0 IPPRADAR Phase 4,0,3 IPPRADAR biaconductor-duoradar 1,27,8,0 IPARADAR biaconductor-duoradar 1,27,8,0 IPPRADAR r-base 4,2,1 IPARTON frastoc 0,11,9 CAT_CONF_CLITER python 3,9,5 GWIZIP_FASTA ounzio 1,10 GUNZID_CTF ounzio 1,10 PLCARD_MARKOUPLICATES picard 3,0,0 SEQC_DINFERSTAT rscen 3,0,1 RSEQC_INFERSTAT rscen 3,0,1 RSEQC_DINFERSTAT rscen 3,0,1 RSEQC_DINFERSTAT rscen 3,0,1 RSEQC_DINFERSTAT rscenc 3,0,1 RSEQC_DINFERSTAT rscenc 3,0,1 RSEQC_DINFERSTAT rscenc 3,0,1 RSEQC_DUNCTONSATION rscenc 3,0,1 RSEQC_DUNCTIONAMOTATION rscenc 3,0,1 RSEQC_DUNCTIONAMOTATION rscenc 3,0,1 RSEQC		yaml	
numberr-base4.9.3DUPRADARbioconductor-durradar1.28.0r-base4.2.1FASTQCfastqc0.11.9GT-2BEDperl5.26.2GT-GENE FILTERurython3.9.5GMZIP FASTAsunzin1.10GMZIP CATEurython3.9.5GMZIP FASTAurython3.9.5GMZIP FASTAurython3.9.5GMZIP CATEurython3.9.5GMZIP CATEurython3.9.5GMZIP CATEurython3.9.5GMZIP CATEurython3.9.1GMZIP CATEurython3.9.1GMZIP CATEurython3.9.1GMZIP CATEurython3.9.1SEQC INFERCENTINTrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1SALHON UNATsalmon1.18.1SALHON UNATsalmon1.18.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNAT<	CUSTOM GETCHROMSIZES	getchromsizes	1.16.1
numberr-base4.9.3DUPRADARbioconductor-durradar1.28.0r-base4.2.1FASTQCfastqc0.11.9GT-2BEDperl5.26.2GT-GENE FILTERurython3.9.5GMZIP FASTAsunzin1.10GMZIP CATEurython3.9.5GMZIP FASTAurython3.9.5GMZIP FASTAurython3.9.5GMZIP CATEurython3.9.5GMZIP CATEurython3.9.5GMZIP CATEurython3.9.5GMZIP CATEurython3.9.1GMZIP CATEurython3.9.1GMZIP CATEurython3.9.1GMZIP CATEurython3.9.1SEQC INFERCENTINTrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1RSEQC UNKTIONNATIONrscqc3.8.1SALHON UNATsalmon1.18.1SALHON UNATsalmon1.18.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNATsalmon1.19.1SALHON UNAT<	DESEQ2 QC STAR SALMON	bioconductor-deseg2	1.28.0
r-base4.2.1FASTOCfastac0.11.9GT72EEDperl5.26.2GTF GENE FILTERpython3.9.5GUNZTP FATAqunzin1.10MAKE TRANSCRIPTS FASTArscn1.3.1starz.7.103.0.0DUALTMAP EMASEQqualinap2.2.2RSEOC BAMSTATrsca3.0.1RSEOC DUNCTRVERTINENTrsca3.0.1RSEOC JUNCTRVERTINENTrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC JUNCTROMSATIRUTIONrscac3.0.1RSEOC READUPLICATIONrscac3.0.1SALMON OLIVATsalaon1.10.1SALMON TIZGENEbicconductor-summarizedexperiment1.24.0r-base4.1.11.24.0r-base4.1.1SALMON TIZGENEsantools1.16.1SALMON TIZGENEsantools1.16.1SALMON TIZGENEsantools1.16.1SALMON TIZGENEsantools1.16.1SALMON TIZGENEsantools1.16.1SALMON TIZGENEsantools1.16.1SALMON TIZGENEsantools1.16.1SALMON TIZGENE <td></td> <td>r-base</td> <td>4.0.3</td>		r-base	4.0.3
FASTOCfastoc0.11.9GT-2BPDport5.76.2GTF GENE FLITERpython3.9.5GIMZIP FASTAqunzip1.10GUAZIP CTFqunzip1.13UNAKE TRANSCRIPTS FASTArsen1.3.1Star2.7.10PICARD MARKOUPLICATESpicard3.0.4QUALIMAP RMASEQqualinag2.2.2QUALIMAP RMASEQqualinag2.2.2RESCO INFERCEMENTATrseoc3.0.1RESCO INFERCEMENTATrseoc3.0.1RESCO INFERCEMENTATrseoc3.0.1RESCO INFERCEMENTATIONrseoc3.0.1RESCO INFERCEMENTIONrseoc3.0.1RESCO INFERCEMENTIONrseoc3.0.1RESCO INFERCEMENTIONrseoc3.0.1RESCO READDISTRICTIONrseoc3.0.1SALMON QUANTsalmon1.10.1SALMON NUANTsalmon1.10.1SALMON NUANTsalmon1.10.1SALMON TXCEREpython3.9.5SALMON TXCEREpython3.9.5SANTOOLS TORSTAT </td <td>DUPRADAR</td> <td>bioconductor-dupradar</td> <td>1.28.0</td>	DUPRADAR	bioconductor-dupradar	1.28.0
GTF28EDpertS.26.2GTF GNE FLITERpython3.9.5GNXIPE FASTAqunzip1.10GNXIPE FASTAqunzip1.10GNXIPE FASTArscm1.31GUAZIP FASTArscm1.31GUAZIP FASTArscm1.31PICARD MARKOUPLICATESpicard3.0.0QUALIMAP. RNASEDqualinap2.2.2-RSEOC BANSTATrscoc3.0.1RSEOC INMERDISTANCErscoc3.0.1RSEOC INMERDISTANCErscoc3.0.1RSEOC INMERDISTANCErscoc3.0.1RSEOC INMERDISTANCErscoc3.0.1RSEOC INMERDISTANCErscoc3.0.1RSEOC INMERDISTRUCErscoc3.0.1RSEOC INMERDISTRUCErscoc3.0.1RSEOC INTERDISTRUCErscoc3.0.1RSEOC INTERDISTRUCErscoc3.0.1RSEOC READDISTRIBUTIONrscoc3.0.1RSEOC READDISTRUCTIONrscoc3.0.1SALMON OUNTsalnon1.10.1SALMON TAZGENEpython3.9.5SALMON TAZGENEpython3.9.5SALMON TAZIERpython3.9.5SALMON TAZIERsantools1.16.1SALMON TAZIERsantools1.16.1SALMON TAZIERsantools1.16.1SALMON TAZIERsantools1.16.1SALMON TAZIERsantools1.16.1SALMON TAZIERsantools1.16.1SALMON TAZIERsantools1.16.1SALMON TAZIERsantools <t< td=""><td></td><td>r-base</td><td>4.2.1</td></t<>		r-base	4.2.1
GITGUN3.9.5GINZTPFASTAqunzip1.10MAKEIRANSCRIPTSFASTA1.3.1MAKEstar2.7.10PICARDMAKDUPLICATESpicard3.0.0OUALIMAPRNASEQqualinap2.2.2RSECRANKDUPLICATESgualinap2.2.2RSECRANKDUPLICATESqualinap2.2.2RSECRANKDUPLICATESqualinap2.2.2RSECRANKDUPLICATESqualinap2.2.2RSECRANKDUPLICATESqualinap2.2.2RSECRANKDUPLICATESqualinap2.2.2RSECRANKDUPLICATESrseqc3.0.1RSECREADISTANCErseqc3.0.1RSECREADISTRIBUTIONrseqc3.0.1RSECREADISTRIBUTIONrseqc3.0.1SALMON DUMATsalmon1.12.0SALMON DUMATsalmon1.12.0SALMON NUMATsalmon1.12.0SALMON TXCENEpython3.9.5SALMON TXIMPORTbioconductor-summarizedexperiment1.24.0SALMON TXIMPORTpython3.9.5SANTOOLS FLAGSTATsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1	FASTQC	fastqc	0.11.9
GMAZIP FASTAgunzip1.10GMAZIP GTFqunzin1.10MAKE TRANSCRIPTS FASTArsem1.3.1star2.7.10PICARD MARKOUPLICATESpicard3.0.0QUALIMAP. RNASEQqualimap2.2.22RSEQC BMSTATrseq3.0.1RSEQC INFEREXPERIMENTrseqc3.0.1RSEQC INFEREXPERIMENTrseqc3.0.1RSEQC INFEREXPERIMENTrseqc3.0.1RSEQC JUNCTIONARNOTATIONrseqc3.0.1RSEQC SUNCTIONARNOTATIONrseqc3.0.1RSEQC READDISTRENTIONrseqc3.0.1RSEQC READDISTRENTIONrseqc3.0.1RSEQC READDISTRENTIONrseqc3.0.1SALMON OWAITsalnon1.10.1SALMON NOWAITsalnon1.10.1SALMON NOWAITsalnon1.12.0r-base4.1.11.24.0r-base4.1.11.24.0r-base4.1.11.24.0salmon TX/GENEpython3.9.5SALMON TX/GENEpython3.9.5SALMON TX/GENEsamtools1.16.1SALMON TX/GENEsamtools1.16.1SANTOOLS TATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1STAR ALIGNqavk5.1.0reacsattools3.4reacstar3.7.9STRINGTIE	GTF2BED	perl	5.26.2
GINZTP GTFaunzip1.10MAKE TRANSCRIPTS FASTArsem1.31VICARD MARKDUPLICATESolcard3.0.0OUALIMAP RNASEDaualinap2.2.2RSICO DAMISTATrseac3.0.1RSECO TINEREXPERIMENTrseac3.0.1RSECO LINEREXPERIMENTrseac3.0.1RSECO LINEREXPERIMENTrseac3.0.1RSECO LINEREXPERIMENTrseac3.0.1RSECO LINEREXPERIMENTrseac3.0.1RSECO LINEREXPERIMENTrseac3.0.1RSECO LINEREXPERIMENTrseac3.0.1RSECO LINERDISTANCErseac3.0.1RSECO LINERDISTANCErseac3.0.1RSECO LINERDISTANCErseac3.0.1RSECO LINERDISTIGUTIONrseac3.0.1RSECO LINERADDISTRIGUTIONrseac3.0.1RSECO READDIPLICATIONrseac3.0.1RSECO READDIPLICATIONrseac3.0.1SALMON QUANTsalmon1.10.1SALMON URANTsalmon1.10.1SALMON TXGENEpothon3.9.5SALMON TXGENEpothon3.9.5SALMON TXGENEpothon3.9.5SALMON TXGENEpothon3.9.5SALMON TXTERPORTbioconductor-tximeta1.12.0SALMON TXTERPORTbioconductor-tximeta1.16.1SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS SUBTISsamtools1.16.1SAMTOOLS SUBTISsamtools1.16.1SAMTOOLS SUBTISsamtools1.16.1S	GTF GENE FILTER	python	3.9.5
NAKE TRANSCRIPTS FASTA rscm 1,3,1 star 2,7,10 PICARD MARKOUPLICATES Dicard 3,0,0 QUALIMAP RNASEQ qualimap 2,2,2 RSEQC BANSTAT rseqc 3,0,1 RSEOC INFEREXPERIMENT rseqc 3,0,1 SALMON TAURTON rseqc 3,0	GUNZIP FASTA	gunzip	1.10
starstar2.7.10PICARD MARDUPLICATESpicard3.0.0QUALIMAP RNASEQqualimap2.2.2-RSECE RANSTATrseac3.0.1RSECE CARSTATrseac3.0.1RSECE CINFEREXPERIMENTrseac3.0.1RSECE INFEREXPERIMENTrseac3.0.1RSECE JUNCTIONANDATIONrseac3.0.1RSECE JUNCTIONANDATIONrseac3.0.1RSECE JUNCTIONANDATIONrseac3.0.1RSECE CREADDISTRIBUTIONrseac3.0.1RSECE LINCTIONSATURATIONrseac3.0.1RSECE READDUPLICATIONrseac3.0.1SALMON QUARTsalmon1.10.1SALMON UANTsalmon1.10.1SALMON TXEE GENEDioconductor-summarizedexperiment1.24.0SALMON TXEENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0Cr-base4.1.11.10.1SANTOOLS FLAGSTATsamtools1.16.1SANTOOLS SORTsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATS	GUNZIP GTF	gunzip	1.10
PICARD MARKDUPLICATESpicard3.0.0QUALIMAP RNASEQqualimap2.2.2-RSEOC DAMISTATrscac3.0.1RSEOC INNERDISTANCErscac3.0.1RSEOC INNERDISTANCErscac3.0.1RSEOC UNCTIONNATURATIONrscac3.0.1RSEOC UNCTIONNATURATIONrscac3.0.1RSEOC SUNCTIONNATURATIONrscac3.0.1RSEOC READDISTRIBUTIONrscac3.0.1RSEOC READDISTRIBUTIONrscac3.0.1SALMON QUARTsalmon1.10.1SALMON VANTsalmon1.10.1SALMON TXZGENEpython3.9.5SALMON TXZGENEpython3.9.5SALMON TXZGENEpython3.9.5SALMON TXZGENEpython3.9.5SAMTOOLS TOKSTATsamtools1.16.1SANTOOLS SORTsamtools1.16.1SANTOOLS SORTsamtools1.16.1SANTOOLS SORTsamtools1.16.1SANTOOLS SORTsamtools1.16.1SANTOOLS SORTsamtools1.16.1STRINGTLE STRINGTIEcutadant3.4CCS EDECLIPucsc377UCSC EDECRAPHTOBIGWIGucsc377UNTAR SALMON TNDEXuntar1.30NorkflowNextflow22.10.	MAKE TRANSCRIPTS FASTA	rsem	1.3.1
OUALIMAP RNASEQqualimap2.2.2-RSEOC EAMSTATrscoc3.0.1RSEOC INFEREXPERIMENTrscoc3.0.1RSEOC INFEREXPERIMENTrscoc3.0.1RSEOC INFEREXPERIMENTrscoc3.0.1RSEOC INFEREXPERIMENTrscoc3.0.1RSEOC INFEREXPERIMENTrscoc3.0.1RSEOC INFEREXPERIMENTrscoc3.0.1RSEOC INFEREXPERIMENTrscoc3.0.1RSEOC READDISTRIBUTIONrscoc3.0.1RSEOC READDUPLICATIONrscoc3.0.1SALMON OUNATsalmon1.10.1SALMON OUNATsalmon1.10.1SALMON TX2GENEpython3.9.5SAMTOULS FLAGSTATbioconductor-tximeta1.12.0SAMTONLS IDXSTATSsantools1.16.1SAMTOULS IDXSTATSsantools1.16.1SAMTOULS IDXSTATSsantools1.16.1SAMTOULS SUBSTATSsantools1.16.1SAMTOULS STATSsantools1.16.1SAMTOULS STATSsantools1.16.1SAMTOU		star	2.7.10a
RSEQC BANSTATrscac3.0.1RSEQC INFEREXPERIMENTrscac3.0.1RSEQC INVERDISTANCErscac3.0.1RSEQC INVERDISTANCErscac3.0.1RSEQC JUNCTIONSATURATIONrscac3.0.1RSEQC JUNCTIONSATURATIONrscac3.0.1RSEQC JUNCTIONSATURATIONrscac3.0.1RSEQC SEADDUPLICATIONrscac3.0.1RSEQC READDUPLICATIONrscac3.0.1SALMON QUANTsalmon1.10.1SALMON SE GENEbloconductor-summarizedexperiment1.24.0r-base4.1.11.24.0SALMON TXZGENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-base4.1.11.24.0samtoolsr-base4.1.1SAMPLESHEET CHECKpython3.9.5SANTOOLS TARSTATsamtools1.16.1SANTOOLS TARSTATsamtools1.16.1SANTOOLS TARSTATsamtools1.16.1SANTOOLS SORTsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1SANTOOLS STATSsamtools1.16.1STRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstar3.4UCSC BEDCLIPucsc377UCSC BEDCRAPHTOBLGWIGuntar1.30WORK1DawNutar </td <td>PICARD MARKDUPLICATES</td> <td>picard</td> <td>3.0.0</td>	PICARD MARKDUPLICATES	picard	3.0.0
RSEQC INFEREXPERIMENTrseqc3.0.1RSEQC INNERDISTANCErseqc3.0.1RSEQC JUNCTIONANDTATIONrseqc3.0.1RSEQC JUNCTIONANDTATIONrseqc3.0.1RSEQC READDISTRIBUTIONrseqc3.0.1RSEQC READDIVELICATIONrseqc3.0.1SALMON QUANTsalmon1.10.1SALMON QUANTsalmon1.10.1SALMON YACGENEbioconductor-summarizedexperiment1.24.0r-baseoxthon3.9.5SALMON TXZGENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-based.1.11.6.1SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNqavk5.1.0star2.7.9aSTRINGTIE STRINGTIEstringtie2.7.10UCSC BEDCLIPucsc377UCSC BEDCLIPucsc377UCSC BEDCLIPucsc377UNTAR STAR INDEXuntar1.30NORK10wNextflow22.10.	QUALIMAP RNASEQ	qualimap	2.2.2-0
RSEOC INNERDISTANCErseqc3.0.1RSEOC JUNCTIONANNOTATIONrseqc3.0.1RSEOC JUNCTIONSATURATIONrseqc3.0.1RSEOC JUNCTIONSATURATIONrseqc3.0.1RSEOC READDISTRIBUTIONrseqc3.0.1RSEOC READDUPLICATIONrseqc3.0.1SALMON OUANTsalmon1.10.1SALMON OUANTsalmon1.10.1SALMON OUANTsalmon1.10.1SALMON TXIGENEbioconductor-summarizedexperiment1.24.0r-base4.1.11.12.0SALMON TXIMPORTbioconductor-tximeta1.12.0r-base4.1.11.10.1SAMTOOLS FLAGSTATsantools1.16.1SAMTOOLS IDXSIATSsantools1.16.1SAMTOOLS SORTsantools1.16.1SAMTOOLS STATSsantools1.16.1STRINGTIE STRINGTIEcutadat3.4TRIMGALOREcutadat3.4UCSC BEDCLIPucsc377UCSC BEDCLIPucsc377UNTAR STAR INDEXuntar1.30WORKHOWNextflow22.10.	RSEQC BAMSTAT	rseqc	3.0.1
RSEQC JUNCTIONANNOTATIONrseqc3.0.1RSEQC JUNCTIONSATURATIONrseqc3.0.1RSEQC READDISTRUUTIONrseqc3.0.1RSEQC READDUPLICATIONrseqc3.0.1SALMON OUNITsalmon1.10.1SALMON OUNITsalmon1.10.1SALMON OUNITbioconductor-summarizedexperiment1.24.0r-base4.1.13.9.5SALMON TXIMPORITbioconductor-tximeta1.12.0r-base4.1.13.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS TONSTATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNgawk5.1.0samtoolssantools1.16.1STRINGTIE STRINGTIEcutadapt3.4TRIMGALOREcutadapt3.4UCSC BEDCLIPucsc377UCSC BEDCLIPucsc377UNTAR STAR INDEXuntar1.30WITAR STAR INDEXuntar1.30WITAR STAR INDEXuntar1.30WITAR STAR INDEXNextflow22.10.	RSEQC INFEREXPERIMENT	rseqc	3.0.1
RSEOC JUNCTIONSATURATIONrseqc3.0.1RSEOC READDUPLICATIONrseqc3.0.1RSEOC READDUPLICATIONrseqc3.0.1SALMON OUANTsalmon1.10.1SALMON SE GENEbioconductor-summarizedexperiment1.24.0r-baser-base4.1.1SALMON TX2GENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-baser-base4.1.1SAMTOOLS FLAGSTATsamtools3.9.5SAMTOOLS TASTSsamtools1.16.1SAMTOOLS TASTSsamtools1.16.1SATTOOLS SORTsamtools1.16.1SATTOOLS STATSsamtools1.16.1SATTOOLS STATSsamtools1.16.1SATTOOLS SORTsamtools1.16.1SATTOOLS STATSsantools1.16.1SATTOOLS SORTsamtools3.4ccutadapt3.4ctrimgalore0.6.7UCSC BEDCLIPucsc377UCSC BEDCLIPucsc377UNTAR SALMON INDEXuntar1.30WITAR STAR INDEXuntar1.30WorkflowNextflow22.10.	RSEQC INNERDISTANCE	rseqc	3.0.1
RSEQC READDISTRIBUTIONrscqc3.0.1RSEQC READDUPLICATIONrscqc3.0.1SALMON OUANTsalmon1.10.1SALMON SE GENEbioconductor-summarizedexperiment1.24.0r-baser-base4.1.1SALMON TX2GENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-base4.1.13.9.5SAMTOOLS FLAGSTATsamtools3.9.5SAMTOOLS TUSTATSsamtools1.16.1SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SATATSsamtools1.16.1SATATSsamtools1.16.1SATTOOLS TOREsamtools1.16.1SATTOOLS TOREsamtools1.16.1SATTOOLS TATSsamtools1.16.1SATTOOLS TATSsamtools1.16.1SATTOOLS TATSsamtools1.16.1STAR ALIONqawk5.1.0STRINGTIE STRINGTIEstringtie2.2.10TRIMGALOREcutadapt3.4UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30WorkflowNextflow22.10.	RSEQC JUNCTIONANNOTATION	rseqc	3.0.1
RSEOC READDUPLICATIONrseqc3.0.1SALMON OUANTsalmon1.10.1SALMON SE GENEbioconductor-summarizedexperiment1.24.0r-base4.1.1SALMON TX2GENEpvthon3.9.5SALMON TXMPORTbioconductor-tximeta1.12.0r-base4.1.13.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS SORTsamtools1.16.1STAR ALIGNqawk5.1.0star2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadaot3.4trimgalore0.6.7UCSC BEDCLIPucsc377UNTAR SALMON INDEXuntar1.30WorkflowNextflow22.10.	RSEQC JUNCTIONSATURATION	rseqc	3.0.1
SALMON OUANTsalmon1.10.1SALMON SE GENEbioconductor-summarizedexperiment1.24.0r-baser-base4.1.1SALMON TX2GENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-baser-base4.1.1SAMPLESHEET CHECKpython3.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SRTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRINGALOREcutadapt3.4trimgalore0.6.7UCSCUCSC BEDCLIPucsc377UNTAR SALMON INDEXuntar1.30WORKflowNextflow22.10.	RSEQC READDISTRIBUTION	rseqc	3.0.1
SALMON SE GENEbioconductor-summarizedexperiment1.24.0r-baser-base4.1.1SALMON TX2GENEpvthon3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-base4.1.1SAMPLESHEET CHECKpvthon3.9.5SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS SUXSTATSsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS SORTsamtools1.16.1STAR ALIGNqawk5.1.0STRINGTIE STRINGTIEstringtie2.2.1TRINGALOREcutagapt3.4trimgalore0.6.73.77UCSC BEDCLIPucsc377UNTAR SALMON INDEXuntar1.30WORKflowNextflow22.10.	RSEQC READDUPLICATION	rseqc	3.0.1
Image: constraint of the systemr-base4.1.1SALMON TXZGENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-baser-base4.1.1SAMPLESHEET CHECKpython3.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNqawk5.1.0star2.7.99starSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4trimgalore0.6.7UCSC BEDCLIPucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	SALMON QUANT	salmon	1.10.1
SALMON TX2GENEpython3.9.5SALMON TXIMPORTbioconductor-tximeta1.12.0r-base4.1.1SAMPLESHEET CHECKpython3.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS INJEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SATTOOLS SORTsamtools1.16.1SATTOOLS STATSsamtools1.16.1STAR ALIGNqawk5.1.0STRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4UCSC BEDCLIPucsc377UCSC BEDCLIPucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	SALMON SE GENE	bioconductor-summarizedexperiment	1.24.0
SALMON TXIMPORTbioconductor-tximeta1.12.0r-baser-base4.1.1SAMPLESHEET CHECKpython3.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SATTOOLS STATSsamtools1.16.1STAR ALIGNgawk5.1.0star2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4ucsc377UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGuntar1.30UNTAR SALMON INDEXuntar1.30WorkflowNextflowNextflow22.10.		r-base	4.1.1
r-base4.1.1SAMPLESHEET CHECKpython3.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNgawk5.1.0STRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadaot3.4trimgalore0.6.7UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	SALMON TX2GENE	python	3.9.5
SAMPLESHEET CHECKpython3.9.5SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNqawk5.1.0STRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	SALMON TXIMPORT	bioconductor-tximeta	1.12.0
SAMTOOLS FLAGSTATsamtools1.16.1SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNqawk5.1.0STRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstrinqtie2.2.1TRIMGALOREcutadapt3.4UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30WorkflowNextflow22.10.		r-base	4.1.1
SAMTOOLS IDXSTATSsamtools1.16.1SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNgawk5.1.0samtools1.16.1STAR ALIGNgawk2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30WorkflowNextflow22.10.	SAMPLESHEET CHECK	python	3.9.5
SAMTOOLS INDEXsamtools1.16.1SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNgawk5.1.0samtools1.16.1STRINGTIE STRINGTIEstar2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30WORKFlowNextflow22.10.	SAMTOOLS FLAGSTAT	samtools	1.16.1
SAMTOOLS SORTsamtools1.16.1SAMTOOLS STATSsamtools1.16.1STAR ALIGNqawk5.1.0samtools1.16.1star2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4uCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30WorkflowNextflow22.10.	SAMTOOLS IDXSTATS	samtools	1.16.1
SAMTOOLS STATSsamtools1.16.1STAR ALIGNgawk5.1.0stargawk1.16.1starstar2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4ucsc377UCSC BEDCLIPucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	SAMTOOLS INDEX	samtools	1.16.1
STAR ALIGNqawk5.1.0samtools1.16.1star2.7.9aSTRINGTIE STRINGTIEstringtieTRIMGALOREcutadapttrimgalore0.6.7UCSC BEDCLIPucscUNTAR SALMON INDEXuntarUNTAR STAR INDEXuntarWorkflowNextflow22.10.	SAMTOOLS SORT	samtools	1.16.1
samtools1.16.1star2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4trimgalore0.6.7UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	SAMTOOLS STATS	samtools	1.16.1
star2.7.9aSTRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4trimgalore0.6.7UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	STAR ALIGN	gawk	5.1.0
STRINGTIE STRINGTIEstringtie2.2.1TRIMGALOREcutadapt3.4trimgalore0.6.7UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.		samtools	1.16.1
TRIMGALOREcutadapt3.4trimgalore0.6.7UCSC BEDCLIPucscUCSC BEDGRAPHTOBIGWIGucscUNTAR SALMON INDEXuntarUNTAR STAR INDEXuntarWorkflowNextflow		star	2.7.9a
trimgaloretrimgalore0.6.7UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	STRINGTIE STRINGTIE	stringtie	2.2.1
UCSC BEDCLIPucsc377UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	TRIMGALORE	cutadapt	3.4
UCSC BEDGRAPHTOBIGWIGucsc377UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.		trimgalore	0.6.7
UNTAR SALMON INDEXuntar1.30UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	UCSC BEDCLIP	UCSC	377
UNTAR STAR INDEXuntar1.30WorkflowNextflow22.10.	UCSC BEDGRAPHTOBIGWIG	UCSC	377
Workflow Nextflow 22.10.	UNTAR SALMON INDEX	untar	1.30
	UNTAR STAR INDEX	untar	1.30
	Workflow	Nextflow	22.10.7
		nf-core/rnaseg	3.11.2

nf-core/rnaseq is a bioinformatics pipeline that can be used to analyse RNA sequencing data obtained from organisms with a reference genome and annotation. It takes a samplesheet and FASTQ files as input, performs quality control (QC), trimming and (pseudo-)alignment, and produces a gene expression matrix and extensive QC report. https://nf-co.re/rnaseq









- NIH Data Sharing Site <u>https://sharing.nih.gov/</u>
- Harvard Biomedical Data Management <u>https://datamanagement.hms.harvard.edu/</u>
- Open Science Framework https://osf.io/
- Keeping a Lab Notebook <u>https://www.training.nih.gov/assets/Lab_Notebook_508_(new).pdf</u>
- CBIIT Storage <u>https://service.cancer.gov/ncisp?id=nci_kb_article&sysparm_article=NCI-KB0014654</u>
- DME Users Guide <u>https://wiki.nci.nih.gov/display/DMEdoc</u>
- GitHub Software Repository <u>https://github.com/</u>
- DRYAD Best Practices <u>https://datadryad.org/stash/best_practices#organize</u>
- GLOBUS Data Transfer <u>https://hpc.nih.gov/docs/globus/</u>
- Protocols Repository http://protocols.io
- Plasmid Repository https://www.addgene.org

Journal Instructions:

- Nature Communications Guide to authors https://www.nature.com/ncomms/submit/guide-to-authors

Resources

• Harvard File and Directory Structure Tutorial - https://datamanagement.hms.harvard.edu/plan-design/directory-structure

Electronic lab notebooks CBIIT - https://service.cancer.gov/ncisp?id=nci kb article&sysparm article=NCI-KB0014880

• Brief guide for submission to Nature Communications - https://www.nature.com/documents/ncomms-submission-guide.pdf

Upcoming Talks

Jun 28 2023

Realizing FAIR principles and Reproducible Computational Workflows with the Arvados Platform

O When: Wed, Jun 28, 2023 - 11:00 am - 12:00 pm

Jul 27

2023

Jul 28

2023

Jul 25

2023

- Delivery: **Online**
- Presented By: Brett Smith Senior Software Engineer Curii

Data Management and Sharing: Part 1

O When: Thu, Jul 27, 2023 - 1:00 pm - 2:00 pm

- Delivery: Online
- Presented By: Raisa Ionin (NIH Library)

Data Management and Sharing: Part 2

- ② When: Fri, Jul 28, 2023 1:00 pm 2:00 pm
- Delivery: Online
- Presented By: Raisa Ionin (NIH Library)

Managing Bioinformatics Projects with Jupyter Notebook

- O When: Tue, Jul 25, 2023 1:00 pm 2:00 pm
- Delivery: Online
- Presented By: Amy Stonelake (BTEP)

 \sim

Part Of: Introduction to Bioinformatics Summer Series Course