Chapter 3

CyVerse for Reproducible Research: RNA-Seq Analysis

Jason Williams

Abstract

Posing complex research questions poses complex reproducibility challenges. Datasets may need to be managed over long periods of time. Reliable and secure repositories are needed for data storage. Sharing big data requires advance planning and becomes complex when collaborators are spread across institutions and countries. Many complex analyses require the larger compute resources only provided by cloud and high-performance computing infrastructure. Finally at publication, funder and publisher requirements must be met for data availability and accessibility and computational reproducibility. For all of these reasons, cloud-based cyberinfrastructures are an important component for satisfying the needs of data-intensive research. Learning how to incorporate these technologies into your research skill set will allow you to work with data analysis challenges that are often beyond the resources of individual research institutions. One of the advantages of CyVerse is that there are many solutions for high-powered analyses that do not require knowledge of command line (i.e., Linux) computing. In this chapter we will highlight CyVerse capabilities by analyzing RNA-Seq data. The lessons learned will translate to doing RNA-Seq in other computing environments and will focus on how CyVerse infrastructure supports reproducibility goals (e.g., metadata management, containers), team science (e.g., data sharing features), and flexible computing environments (e.g., interactive computing, scaling).

Key words Cyberinfrastructure, Cloud computing, Reproducible research, Containers, Workflow management, Data life cycle, Metadata, RNA-Seq, Kallisto

1 Introduction

CyVerse [1] is an example of cyberinfrastructure—infrastructure for data, compute, and collaboration—tailored for scientific research. Originally developed to meet the needs of plant scientists [2], CyVerse provides capacious data and compute resources and is generally available to all researchers (i.e., US and international, with some limitations *see*: <https://www.CyVerse.org/policies>).

CyVerse cyberinfrastructure was built around the “life cycle” of data (*see* Fig. 1). Data must be annotated with metadata to understand its origin, integrity, context, and appropriate use. CyVerse allows users to upload 100 gigabyte (GB) of data to a home folder on the CyVerse Data Store and up to an additional 10 terabytes

Fig. 1 Life cycle of Data. CyVerse provides tools and services that allow users to manage data at all stages of the research workflow

(TB) of storage can be granted with justification. Using the open source iRODS [3] platform, the Data Store automatically manages data access, backup, transfer, and metadata annotation to ensure a safe and functional repository.

In this tutorial, we will access the Data Store through the Discovery Environment (DE), a web-based graphical interface for analysis. In the **Data** view, the DE provides a file browsing system for the Data Store, enabling users to access, edit, and share data and metadata. The DE **Apps** (applications) view provides a catalog of hundreds of applications. These applications are typically the same software tools (e.g., bioinformatics software, Linux utilities) that are available to use at the command line, presented to the user in a customizable point-and-click interface. Importantly, the DE and other CyVerse applications are built on widely used, open source software technologies (e.g., Docker [4], Kubernetes [5], iRODS). This means that applications and pipelines built to deploy on CyVerse are largely replicable on any Linux-based computational infrastructure. CyVerse also provides flexibility for researchers who prefer command line environments and additional customization of computational resources. The DE also provides access to applications that require high-performance computing (HPC). Many of these apps run on XSEDE HPC resources delivered through a CyVerse partner, the Texas Advanced Computing Center (TACC). The Visual Interactive Computing Environment

(VICE), a sub-system of the DE, allows users to launch interactive sessions such as JupyterLab, Cloud Shell, RStudio, and R Shiny applications. The DE's **Analysis** view provides the user a list of previously launched analysis jobs and the status of submitted and running jobs. Users can view details on job history and cancel or relaunch jobs.

1.1 RNA-Seq Example Use Case

This tutorial will use a bulk RNA-Seq analysis as an example use case for CyVerse infrastructure. RNA-Seq has become one of the most important ways to characterize the transcriptome. Although some of the important methodological and technical background on RNA-Seq will be discussed, the reader is referred to the many seminal papers and comprehensive reviews for a more thorough introduction to the topic [6–8].

2 Materials

2.1 Tutorial Dataset

This tutorial uses data publicly available on the NCBI Sequence Read Archive (SRA) [9]. The methods section is a guide you can modify when you want to do a similar analysis with your own data (*see Note 1*) will describe these alternative steps. Beyond the necessity of example data for the tutorial, it is important to be familiar with the SRA for several reasons. The SRA (along with the European Nucleotide Archive [10] and others) is a canonical repository for high-throughput sequencing reads. It is likely that publication of your research will require you to deposit your data in such a repository. In addition, datasets available on the SRA have tremendous value for reuse including for exploratory analyses, and in our case, educational exercises.

The data for this tutorial comes from Zia et al. [11] which used RNA-Seq to explore overlap in signaling pathways in *Arabidopsis* treated with the hormones melatonin and auxin. Although these hormones have similar chemical structures (indoles), the study identified distinct signaling pathways and changes in gene expression. The dataset is available on the SRA under BioProject PRJNA553702. We will attempt to replicate the RNA-Seq portion of this analysis.

All data in this tutorial has been pre-staged on the CyVerse Data Store (/iplant/home/shared/CyVerse_training/tutorials/pbv3) and viewable on the CyVerse Data Commons at https://datacommons.CyVerse.org/browse/iplant/home/shared/CyVerse_training/tutorials/pbv3. You may preview or skip ahead to any section by using the precomputed results there.

2.2 CyVerse Account

A CyVerse account will be required to complete this tutorial. An account may be obtained for free at <https://user.CyVerse.org>. With few limitations (*see*: <https://www.CyVerse.org/policies>), usage of

CyVerse is available to users in the USA and internationally. It is recommended to register with an email address affiliated with an educational or research institution (i.e., with an .edu or .org. email address) as some CyVerse services are not available to users unaffiliated with an institution. Currently, CyVerse users are entitled to 100 GB of data storage with additional storage available by request with justification.

2.3 Applications

In addition to CyVerse services (e.g., compute and data storage), we will use several applications deployed on the CyVerse DE and its interactive subsystem VICE. In their deployment on CyVerse, the apps are run in independent Docker containers. This allows CyVerse users to make stable version-controlled DE applications of almost any open source software available, making it easy to reproduce analyses in a variety of contexts. The analysis apps we will use as listed below are the same as those available for installation from their original developers.

2.3.1 SRA-Tools

The SRA-Toolkit [12] is a software package published by NCBI for retrieval of data from the SRA (DE Apps: **sra-tools prefetch**, **sra-tools vdb-validate**, **sra-tools fasterq-dump**).

2.3.2 FastQC

FastQC [13] generates quality control reports of high-throughput sequencing data. Reports are generated in HTML format for easy exploration (DE App: **FastQC 0.11.5 (multi-file)**).

2.3.3 Kallisto and Sleuth

Introduced in 2015, Kallisto [14, 15] continues to be a popular tool for RNA-Seq. Kallisto's popularity is attributable to its speed, accuracy, and ease of use compared to previous tools. The key innovation of Kallisto is its use of pseudoalignment. Early RNA-Seq software stemmed from years of work on genome assembly. In a genome assembly, short reads must be progressively aligned and scaffolded until (if successful) end-to-end contigs are assembled for all chromosomes. An RNA-Seq analysis is a variation on this problem. Given a reference genome, short reads (in this case generated from cDNA libraries) are aligned to the genome and counted as a digital measure of gene expression. Since the reads are a measure of the abundance of mRNA transcripts—the more reads aligned, the higher the inferred level of gene expression (with the caveat that expression and abundance are not always the same thing).

However, processing millions of short read alignments to a genome is computationally expensive. In a genome assembly, this careful matching is worthwhile because a few mismatches may help to unambiguously map a read in a repetitive genome or reveal important sequence variations. RNA-Seq experiments are usually focused on measuring gene expression rather than isoform and

variant discovery. In pseudoalignment, reads are matched to the transcriptome without the need for a complete nucleotide-by-nucleotide matching. Instead, pseudoalignment uses a k-mer-based approach which is much more computationally efficient. K-mers of reads (for example, a 31-mers) are compared to the provided known transcripts (which are assembled into an index prior to quantification). Rather than an exact matching strategy that asks, “how does this read match to this transcript?” the transcript compatibility approach asks, “could this transcript have generated this read?” This subtle difference means that if the first k-mer of a read does not match a transcript, there is no reason to continue computing the rest of an alignment that is not a possible match (the transcript could not have generated that read). If multiple transcripts could have generated a read, the next (non-redundant) k-mer is considered until the ambiguity is resolved. In fact, during pseudoalignment, k-mers can be skipped without a loss in accuracy; it may only be necessary to check the first and last k-mer to unambiguously match it to a transcript. This approach not only speeds up the search but has the added benefit of being highly tolerant to low frequency sequencing errors. Like any software, however, Kallisto has limitations. For example, an incomplete transcriptome will leave some reads unmapped (DE App: **Kallisto-v.0.43.1**).

After pseudoalignment and quantification with Kallisto, the Sleuth R package [16] allows inspection and differential analyses of count data. Results, including several automatically generated figures, are generated by Sleuth and presented in an interactive R Shiny application (DE App: **RStudio Sleuth**).

2.3.4 RStudio

Through VICE (part of the CyVerse Discovery Environment), users can launch interactive applications (e.g., Jupyter notebooks, RStudio, R Shiny) in addition to executable applications like Kallisto. In this tutorial, we will use the Sleuth R package in an RStudio session. RStudio is a popular interface to the R programming language. This portion of the tutorial will be presented in the form of an RMarkdown notebook. While knowledge of R will be helpful, there is no need to know the language in order to complete this section.

2.3.5 Spreadsheet Software

Software for creating and editing spreadsheets will be used to create or edit metadata that we will be using in this experiment. Microsoft Excel will work, as will online applications such as Google Sheets, or the free OpenOffice Calc software.

3 Methods

In our analysis, we will access our CyVerse account and set up folders and sharing settings for our experiment in the CyVerse Discovery Environment, import data from the SRA into CyVerse using SRA-tools apps, and label it with the appropriate metadata. Next, we will perform a quality control check on sequence data using FastQC. Finally, we will perform pseudoalignment to generate count data using Kallisto and test for differential expression while visualizing these results using Sleuth.

3.1 *CyVerse Account Setup*

1. Obtain a CyVerse account at <https://user.CyVerse.org/>. Register with an institutional email address (e.g., .edu or .org) if possible. If you previously have had a CyVerse account, you can also recover or reset your password on this site. After signup, you will need to verify your email address in order to activate your account.

3.2 *Discovery Environment Login and Data Sharing Setup*

In this section, we will create a folder in the Discovery Environment (DE) and begin organizing a space for our experimental data and analysis results on the Data Store. Please note that the DE provides an interface to the Data Store, but the two are separate platforms. The DE provides access to applications for analysis and a history of previous analyses. The Data Store serves as the underlying data storage and management system for all CyVerse platforms and provides a unified file system view for all data deposited at CyVerse. It may also be accessed at the command line interface and through other interfaces. You can learn more about other DE and Data Store features not covered here on their documentation pages at <https://learning.cyverse.org/>. See Fig. 2 for a layout of the DE and the names of the main views and menus.

1. Log in to the Discovery Environment at <https://de.cyverse.org/de/>.
2. Click on **Data** to browse your collection of files and navigate to your home folder (i.e., top folder labeled with your CyVerse username).
3. Click the **Folder** button to create a new folder for this project (suggested name: **rna-seq-tutorial**). You should avoid using spaces or special characters (e.g., !@#\$%^&) in any folder or file names.
4. Once you have created this folder, you may want to share it with collaborators. To do so, select (checkbox) the **rna-seq-tutorial** folder in your **Data** view. Then, click the **Share** button.
5. (Optional) You can then enter the names (i.e., name, username, email address) of collaborators you wish to share data

Fig. 2 Layout of the Discovery Environment. The Discovery Environment is a web-based graphical user interface that allows CyVerse users to manage, share, and annotate data, launch analysis jobs using applications, and log the status and history of analyses. In this expanded view, the left sidebar menu (a) shows several icons: **Home** presents the default view at login with access to recently used and favorite applications; **Data** shows a view of the CyVerse Data Store including publicly available (community) data; when logged in you will view your uploaded and shared files and folders. You will also access menu-based data management tools (e.g., upload, folder and file creation, file previews, sharing, search, and metadata) in this view; **Apps** provides a catalogue of applications, as well as tools for modifying and creating new applications and workflows. **Analyses** provides a history of jobs, and the status of submitted and running jobs. Job management features include the ability to cancel or relaunch jobs and view detailed job parameters; **Cloud Shell** provides a limited Linux-based shell interface; **Teams** allows you to select and manage user groups; **Collections** allows you to select and manage groupings of applications and resources; **Help** allows you access to help resources. When logged in, **Settings** will also be managed from this sidebar. The **Search bar** (b) allows you to search across data, applications, and analyses. Additional functions (c) include login and account information, notifications, help, and collections

Table 1

Data permissions (based on UNIX permissions) available for files/folders on the CyVerse Data Store

Permission level	Read	Download	Edit	Rename	Move	Delete	Metadata
Read	X	X					View
Write	X	X	X				Add/Edit
Own	X	X	X	X	X	X	Add/Edit

with. These collaborators must also have a CyVerse account. Once you have found collaborators to share with, you must select the level of access you wish to grant. Table 1 describes the permission levels available.

Once you have selected the user to share with and level of permission, click **Done** to share the folder. The shared folder will be available to your collaborators momentarily. They will get a notification with a link to the folder, and it will also appear by choosing the **Shared with me** folder from the Data view dropdown menu. Anything placed in a shared folder will inherit the same sharing permissions. To unshare the folder or change access permissions, select the folder again and return to the Share button. You may change permissions or to revoke permission, choose the permission drop down menu next to the username you wish to unshare with and select **Remove**.

3.3 Obtain Accession Numbers and Metadata from the SRA

In this section, we will describe how to import data from the NCBI Sequence Read Archive (SRA). The example dataset is available at the SRA under the accession PRJNA553702. It is not possible to directly download sequencing files from the SRA. Instead, we must obtain the accessions for the individual sequencing files and use the “sra-tools” software to download those files.

1. Go to the SRA <https://www.ncbi.nlm.nih.gov/sra>.
2. Enter the BioProject accession **PRJNA553702** in the search bar and press **Search**.
3. This search should return 12 entries which are individual sequence files associated with this experiment. Click on any of the results (e.g., “GSM3936272: 100 μ M MEL rep3; Arabidopsis thaliana; RNA-Seq”). This will lead to a page with a detailed description of the experiment.
4. Under the **Study** heading, find the link to **All runs** and click this to access the SRA Run Selector (direct link: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP214076&o=acc_s%3Aa).
5. Under the **Select** heading, there are two downloads (simple text files) available for download. Click on **Metadata** and **Accession list** to download both the files. The metadata file will be called **SraRunTable.txt**, and the accession list will be called **SRR_Acc_List.txt**.

3.4 Upload Files to the Data Store

We will upload the accession file to the Data Store now and the metadata file a few steps later. The metadata file is a CSV file organized as a spreadsheet. We will return to applying this metadata to these files in Subheading 3.7. The upload method we are using is suitable for small (<2 GB) files. See the Learning Center at <https://learning.cyverse.org/> for more efficient methods for large file transfers.

1. In the Discovery Environment **Data** view, navigate to **rna-seq-tutorial**, the folder you created; inside this folder, click the **Upload** button, then select **Browse Local**.
2. Browse your local computer to select the accession list from the SRA (i.e., **SRR_Acc_List.txt**); upload the file. You will get a notification when the upload is completed. You may need to refresh your browser to see the uploaded file.
3. To organize this and other uploads, click the **Folder** button to create a folder (suggested name: **metadata**). Select (checkbox) the uploaded file (i.e., **SRR_Acc_List.txt**) and click the **More Actions** button; choose **Move** and then use the **Browse** button to select the created metadata folder as your destination to move these files to. Click **Move** to complete this action. You may need to refresh your browser to see changes.

3.5 Import Files from SRA with the SRA Toolkit

Now that we have a list of accessions, we can use our first application. All the applications we will be using in the Discovery Environment have been collected. You may access them by going to the **Collections** view (*see* Fig. 2). Scroll to find the **Plant Bioinformatics Vol. 3 Tutorial** collection. You may also search for any application by name in the Apps catalogue.

1. In the Discovery Environment, go to **Apps** and search for and launch **sra-tools prefetch**.
2. In **Analysis Info**, you can name this analysis and provide any comments (optional). Under **Output folder**, navigate to the **rna-seq-tutorial** folder created earlier. Your outputs will automatically be placed in this folder; click **Next**. **Tip:** For future analyses, from the **Settings** icon in the Discovery Environment under **Default analysis output folder**, you can change the default setting for your outputs. This can be useful when you will be doing several analyses and you want them to go to a folder other than your analyses folder (the default setting).
3. In **Parameters** under **Accession list**, browse to your **rna-seq-tutorial** folder and enter the **metadata** folder; select **SRR_Acc_List.txt** as input; click **Next**.
4. In **Advanced Settings (optional)**, each application in the Discovery Environment has a “Resource Requirements” option. These are some minimal computational resources that you can choose for the application. When the Discovery Environment launches an analysis job, it may be run on computing hardware of various capabilities (according to what “nodes” are available to accept the job). App developers may have set minimums here, but you may choose to make adjustments. If you do not know which minimums to choose, you can ignore these options. If you do run into errors, you can contact CyVerse

support. Also contact support if the resources you need for an app do not appear in the list of options; CyVerse can increase these minimums. We will **not** make any changes; click **Next**.

5. In **Review and Launch**, click **Launch Analysis** to begin the analysis job. You will get notification(s) in the Discovery Environment about the status of the job, including an email (by default preferences settings) when the job is complete.
6. You will be redirected to the **Analyses** view where you can see the current status of the job; you can also click on the **Analyses** icon to navigate to this section. When the job is complete, you can click on the **folder icon** next to the analyses name to browse the results. You may need to **Refresh** to see the current job status. This job will vary on connection speed between the CyVerse servers and NCBI. This job is estimated to take about 15–30 min.
7. When the job has status **Completed**, navigate to the output. The resulting output will be 12 folders (one for each accession), each containing a single SRA file (e.g., SRR9666131.sra, SRR9666132.sra...). There will also be a logs folder containing output logs from the CyVerse Discovery Environment analysis.

3.6 Organize Files, Validate Import, and Extract to FASTQ Format

At this point, we will put all the files imported from the SRA into a single location. We will verify the integrity of the files and then extract them from the SRA format into the FASTQ format.

1. In the Discovery Environment, click on the **Data** icon and navigate to your **rna-seq-tutorial** folder.
2. Click on the **Folder** button create a new folder: **imported_sra**.
3. Navigate to the output of the **sra-tools prefetch** analysis completed in Subheading 3.5. You can go to the **Analyses** section of the Discovery Environment and click the **folder icon** next to the analyses name to navigate to this output.
4. Select (checkbox) all 12 of the SRA folders (e.g., SRR9666131, SRR9666132...) and the folder of **logs** from this analysis; click the **More Actions** button and choose **Move**. Browse to the **imported_sra** folder created inside your tutorial folder (**rna-seq-tutorial**). Click **Move** to complete this action. It may take a few minutes to complete this move. You may need to refresh your browser to see changes.
5. Before extracting these files, we can do a check here to verify the integrity of our import from the SRA. In the Discovery Environment click on the **Data** icon and navigate to your **rna-seq-tutorial** tutorial folder and create a folder to store outputs, name the folder **sra_validation**.

6. In the **Apps** view, search for and launch the **sra-tools vdb-validate** app.
7. In **Analysis Info**, you can name this analysis and provide any comments (optional). Under **Output folder**, navigate to the **sra_validation** folder created earlier. Your outputs will automatically be placed in this folder; click **Next**.
8. In **Parameters** under **SRA Files (Input)**, browse to the **imported_sra** folder (created in **step 2**). Open an individual folder (e.g., “SRR966613”) and select the “.sra” file to add it; repeat for each of the 12 folders until you have added all the 12 “.sra” files (i.e., SRR9666131.sra, SRR9666131.sra...); click **Next**.
9. Click **Next** again to skip **Advanced Settings (optional)**; under **Review and Launch**, click **Launch Analysis**.
10. You will be redirected to the **Analyses** view where you can see the current status of the job; you can also click on the **Analyses** icon to navigate to this section. When the job is complete, you can click on the **folder** icon next to the analyses name to browse the results. You may need to **Refresh** to see the current job status. This job is estimated to take about 5–10 min.
11. When the job has status **Completed**, navigate to the output. The output will be a text file (**vdb-validation.txt**). This is a report on a series of file checks (including checksums—an algorithmically generated signature that confirms the file’s integrity). A sample output is shown below with “ok” indications for each test. A similar set of eight lines should appear in the file for each of the verified SRA files.

Sample output:

```
2020-10-06T22:04:41 vdb-validate.2.10.8 info: Database
'SRR9666131.sra' metadata: md5 ok
2020-10-06T22:04:41 vdb-validate.2.10.8 info: Table 'SEQUENCE'
metadata: md5 ok
2020-10-06T22:04:41 vdb-validate.2.10.8 info: Column 'ALTREAD':
checksums ok
2020-10-06T22:04:42 vdb-validate.2.10.8 info: Column 'QUALITY':
checksums ok
2020-10-06T22:04:43 vdb-validate.2.10.8 info: Column 'READ':
checksums ok
2020-10-06T22:04:43 vdb-validate.2.10.8 info: Column 'READ_LEN':
checksums ok
2020-10-06T22:04:44 vdb-validate.2.10.8 info: Column 'READ_START':
checksums ok
2020-10-06T22:04:44 vdb-validate.2.10.8 info: Column 'SPOT_GROUP':
checksums ok
```

12. In the Discovery Environment, click on the **Data** icon and navigate to your **rna-seq-tutorial** folder and create a folder to store outputs, name the folder **fastq_files**.
13. In the **Apps** view, search for and launch the **sra-tools fasterq-dump** app.
14. In **Analysis Info**, you can name this analysis and provide any comments (optional). Under **Output folder**, navigate to the **fastq_files** folder created earlier. Your outputs will automatically be placed in this folder; click **Next**.
15. In **Parameters** under **SRA Files (Input)**, browse to the **imported_sra** (created in **step 3**) and open an individual folder (e.g., “SRR966613”) and select the “.sra” file to add it; repeat for each of the 12 folders until you have added all 12 “.sra” files (i.e., SRR9666131.sra, SRR9666131.sra. . .); click **Next**.
16. Click **Next** again to skip **Advanced Settings (optional)**; under **Review and Launch**, click **Launch Analysis**.
17. You will be redirected to the **Analyses** view where you can see the current status of the job; you can also click on the **Analyses** icon to navigate to this section. When the job is complete, you can click on the **folder** icon next to the analyses name to browse the results. You may need to **Refresh** to see the current job status. This job is expected to take 30–40 min.
18. When the job has status **Completed**, navigate to the output. The expected output will be 12 FASTQ formatted files (e.g., SRR9666131.sra.fastq, SRR9666132.sra.fastq. . .).
19. Since the SRA files are already maintained on NCBI, you can safely delete the original SRA files. While this deletion is not mandatory, it is a responsible use of public infrastructure to remove large unneeded files. Browse to the **rna-seq-tutorial** folder and select the **imported_sra** folder. Click the **More Actions** button and choose **Move to Trash**. Clicking your username in the **Data** view, choose **Trash**. Select the file you wish to delete and click the **Trash** button and then select **Delete** to permanently delete those files.

3.7 Apply Metadata to FASTQ Files

Typically, the FASTQ file is the starting raw data for an experiment. There are several metadata descriptions that should be captured about an experiment (e.g., sequencing platform/chemistry, sample condition, etc.). Many of these metadata descriptions will be required for submission of data to a repository prior to publication (see SRA metadata requirements: <https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/>). The Discovery Environment (via the Data Store’s underlying iRODS software) allows you to label your files with arbitrary metadata. The Discovery Environment also provides several metadata templates, including for SRA submission

(see Data Store guide at <https://learning.CyVerse.org>). Applying metadata to your files/folder in the Data Store makes your data much easier to organize and search without the need to rely exclusively on filenames. We will modify the previously downloaded **SraRunTable.txt** on our local computer in a spreadsheet program and then upload to the Discovery Environment and link these descriptions to the files. This process of metadata application can be applied to any dataset in the Data Store and is a recommended (but not required) practice.

1. In a spreadsheet program (e.g., Excel), open the **SraRunTable.txt** file on your local computer (downloaded in Subheading 3.3). **Tip:** Renaming the file with a “.csv” extension before opening the file may make it easier for your spreadsheet program to properly interpret.
2. To the spreadsheet, insert a new first column (left-most). Name this column “file.”
3. At this point, sort the spreadsheet by run (ascending). This will make it easier in the subsequent steps when file names will likely be alphanumerically sorted.
4. In the **file** column, we will list the path in the Data Store for each FASTQ file (created in Subheading 3.6) as corresponds to its accession in the **Run** column (e.g., [/iplant/home/USERNAME/rna-seq-tutorial/fastq_files/SRR9666132.sra.fastq](#)). In the **Data** view, clicking the three dots next to any file/folder icon reveals a function menu which includes an option to **Copy Path**.

Your first two columns should be similar to the below one (with your username replacing **username** in the file paths):

File	Run
/iplant/home/username/rna-seq-tutorial/fastq_files/SRR9666131.sra.fastq	SRR9666131
/iplant/home/username/rna-seq-tutorial/fastq_files/SRR9666132.sra.fastq	SRR9666132
/iplant/home/username/rna-seq-tutorial/fastq_files/SRR9666133.sra.fastq	SRR9666133
/iplant/home/username/rna-seq-tutorial/fastq_files/SRR9666134.sra.fastq	SRR9666134
/iplant/home/username/rna-seq-tutorial/fastq_files/SRR9666135.sra.fastq	SRR9666135
/iplant/home/username/rna-seq-tutorial/fastq_files/SRR9666136.sra.fastq	SRR9666136
/iplant/home/username/rna-seq-tutorial/fastq_files/SRR9666137.sra.fastq	SRR9666137

(continued)

File	Run
/iplant/home/username/rna-seq-tutorial/fastq_files/ SRR9666138.sra.fastq	SRR9666138
/iplant/home/username/rna-seq-tutorial/fastq_files/ SRR9666139.sra.fastq	SRR9666139
/iplant/home/username/rna-seq-tutorial/fastq_files/ SRR9666140.sra.fastq	SRR9666140
/iplant/home/username/rna-seq-tutorial/fastq_files/ SRR9666141.sra.fastq	SRR9666141
/iplant/home/username/rna-seq-tutorial/fastq_files/ SRR9666142.sra.fastq	SRR9666142

- (Optional) At this point, you can add any desired additional columns with any metadata you want to add. For our tutorial this is not necessary.
- Save this file as **fastq_file_metadata.csv** on your local computer.
- In the Discovery Environment **Data** view, navigate to the **metadata** folder in your **rna-seq-tutorial** folder. Open the folder and then click the **Upload** button, then select **Browse Local**.
- Browse your local computer to select the **fastq_file_metadata.csv**; upload the file. You will get a notification when upload is completed. You may need to refresh your browser to see the uploaded file.
- In the **Data** view, navigate to your **rna-seq-tutorial** folder and select (checkbox) the **fastq_files** folder.
- Click the **More Actions** button and select **Apply Bulk Metadata**. Browse to the **metadata** folder and select the **fastq_file_metadata.csv** file. Click **Done** to complete this action. You will get a notification when the metadata has been applied successfully.
- To view the metadata, navigate to the **fastq_files** folder and select (checkbox) any individual FASTQ file. Click the **More Actions** button and select **Metadata**. You will then see the applied metadata. You can build custom search queries on any of these attributes (column names in your original spreadsheet) and the values (the entries for these columns). See Data Store Guide at <https://learning.cyverse.org> for more information about advanced search queries and smart folders.

3.8 QC Reads with FastQC

Although quantification with Kallisto is robust to sequencing errors, it is a good practice to check the reads for quality. We will use the FastQC software package. A detailed explanation of the FastQC report and its interpretation can be found on the software

developer's website: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Although we will not cover them in this tutorial, other applications are available in the Discovery Environment (e.g., Trimmomatic [17]) which will allow you to filter and trim reads if necessary.

1. In the Discovery Environment, click on the **Data** icon and navigate to your **rna-seq-tutorial** tutorial folder and create a folder to store outputs, name the folder **fastqc_analyses**.
2. In the **Apps** view, search for and launch the **FastQC 0.11.5 (multi-file)** app.
3. In **Analysis Info**, you can name this analysis and provide any comments (optional). Under **Output folder**, navigate to the **fastqc_analyses** folder created earlier. Your outputs will automatically be placed in this folder; click **Next**.
4. In **Parameters** for **Input**, browse to the **fastq_files** folder (created in 3.6) and add the 12 files (e.g., SRR9666131.sra.fastq, SRR9666132.sra.fastq); click **Next**.
5. Click **Next** again to skip **Advanced Settings (optional)**; under **Review and Launch** click **Launch Analysis**.
6. You will be redirected to the **Analyses** view where you can see the current status of the job; you can also click on the **Analyses** icon to navigate to this section. When the job is complete, you can click on the **folder** icon next to the analyses name to browse the results. You may need to **Refresh** to see the current job status. This job is expected to take 30–40 min.
7. When the job has status **Completed**, navigate to the output. The expected output will be 12 HTML formatted reports, and 12 zip files containing additional files including text-based metrics. You can click on and examine all the HTML files; they will open as new tabs in your web browser. Review of the example dataset indicates the samples are all of high quality and additional filtering or trimming would not improve our quantification results.

3.9 Quantify Reads with Kallisto

Kallisto will individually generate pseudoalignments and quantification for each replicate of each condition. In the Discovery Environment application, the “Kallisto index” command will first build an index of the transcriptome. Next, that indexed transcriptome will be used to quantify the read data (equivalent to the “Kallisto quant” command at the command line). To start, we will import an *Arabidopsis* transcriptome from Ensembl.

1. Go to the Ensembl homepage for Arabidopsis at https://plants.ensembl.org/Arabidopsis_thaliana/Info/Index.

2. Under **Gene annotation**, click the **FASTA** link under “**Download genes, cDNAs, ncRNA, proteins**”; http://ftp.ensemblgenomes.org/pub/plants/release-51/fasta/arabidopsis_thaliana/.
3. In the **cdna** folder, locate the file **Arabidopsis_thaliana.TAIR10.cdna.all.fa.gz**.
4. In your web browser, copy the URL for this file (right click, “**Copy Link**” for most browsers). The URL for release 51 of Ensembl is ftp://ftp.ensemblgenomes.org/pub/plants/release-51/fasta/arabidopsis_thaliana/cdna/Arabidopsis_thaliana.TAIR10.cdna.all.fa.gz. **Tip:** Ensure you use the “cdna.all.fa.gz” and your annotation release must match what is used later in the Sleuth analysis (Subheading 3.10).
5. In the **Data** view in the Discovery Environment, navigate to your **rna-seq-tutorial** folder.
6. Create a new folder called **transcriptome** and navigate into the newly created folder.
7. In the **Data** view, click the **Upload** button and choose **Import from URL**; paste in the URL for the *Arabidopsis* transcriptome. Be sure to avoid any extra spaces or characters at the end of your URL. Click **Import** to complete this action. You will get a notification when import is complete. You may need to refresh your browser to see the imported file.
8. Back on the Ensembl page, also copy the URL for the CHECKSUMS file and repeat the import procedure to the same folder. **Tip:** You can also apply metadata to the imported transcriptome. See Subheading 3.7 or the Data Store Guide at <https://learning.cyverse.org/> for additional options for apply metadata.
9. In the Discovery Environment, click on the **Data** icon and navigate to your **rna-seq-tutorial** tutorial folder and create a folder to store outputs, name the folder **kallisto_analyses**.
10. In the **Apps** view, search for and launch the **Kallisto-v.0.43.1** app.
11. In **Analysis Info**, you can name this analysis and provide any comments (optional). Under **Output folder**, navigate to the **kallisto_analyses** folder created earlier. Your outputs will automatically be placed in this folder; click **Next**.
12. In **Parameters** under **Input**:
 - (a) For **The transcript fasta file supplied (fasta or gzipped)**, browse to the **transcriptome** folder (created in step 6) and add the **Arabidopsis_thaliana.TAIR10.cdna.all.fa.gz** file.
 - (b) For **Paired or single end**, choose **single**.

- (c) Under **FASTQ Files (Read1)**: Click **Browse** and browse to the **fastq_files** folder and select 12 files (e.g., SRR9666131.sra.fastq, SRR9666132.sra.fastq...).

Under **Options**:

- (a) For **Number of bootstrap samples**, enter **25**.
- (b) For **Estimated average fragment length (required for single end reads)**, enter **200**.
- (c) For **Estimated standard deviation of fragment length (required for single end reads)**, enter **20**;

Click **Next**.

13. Click **Next** again to skip **Advanced Settings (optional)**; under **Review and Launch**, click **Launch Analysis**.
14. You will be redirected to the **Analyses** view where you can see the current status of the job; you can also click on the **Analyses** icon to navigate to this section. When the job is complete, you can click on the **folder** icon next to the analyses name to browse the results. You may need to **Refresh** to see the current job status. This job is estimated to take about 60–70 min.
15. When the job has status to navigate to the expected output. The expected output will be a folder **Kallisto_quant_output** containing 12 folders (labeled with the accession name). Inside each folder will be:
 - (a) **abundances.h5**: HDF5 binary file containing run info, abundance estimates, bootstrap estimates, and transcript length information length. This file can be read in by Sleuth.
 - (b) **abundances.tsv**: plaintext file of the abundance estimates. It does not contain bootstrap estimates. When plaintext mode is selected, output plaintext abundance estimates. Alternatively, Kallisto h5dump will output an HDF5 file to plaintext. The first line contains a header for each column, including estimated counts, TPM, effective length.
 - (c) **run_info.json**: a json file containing information about the run.

We did not select the BAM file creation option when launching the app. Although a BAM file will be created, it is empty and can be ignored.

3.10 Prepare Experimental Design Metadata for Sleuth

Before we use Sleuth to analyze our data, we need to create a tab-delimited file that matches our samples to their conditions. It is convenient to do this on your local computer, using a spreadsheet program. We can easily modify the **SraRunTable.txt** file we downloaded from the SRA.

Table 2

This table is an example for the experimental design file that must be uploaded to Sleuth in Subheading 3.10. There are redundancies in the table, but this ultimately gives additional options when filtering or grouping samples by one or more of the columns. While the first column must be labeled “sample,” you may choose whatever terms you like. SRA entries in the original metadata file may not be sorted by Run/sample, so be careful in pairing your sample with the appropriate condition

sample	replicate	rep_number	treatment	condition
SRR9666131	0_1_EtOH_rep-1	1	0_1_EtOH	control
SRR9666132	0_1_EtOH_rep-2	2	0_1_EtOH	control
SRR9666133	0_1_EtOH_rep-3	3	0_1_EtOH	control
SRR9666134	4_5_ul-NAA_rep-1	1	4_5_ul-NAA	NAA
SRR9666135	4_5_ul-NAA_rep-2	2	4_5_ul-NAA	NAA
SRR9666136	4_5_ul-NAA_rep-3	3	4_5_ul-NAA	NAA
SRR9666137	5_ul-mel_rep-1	1	5_ul-mel	low_melatonin
SRR9666138	5_ul-mel_rep-2	2	5_ul-mel	low_melatonin
SRR9666139	5_ul-mel_rep-3	3	5_ul-mel	low_melatonin
SRR9666140	100_ul-mel_rep-1	1	100_ul-mel	high_melatonin
SRR9666141	100_ul-mel_rep-2	2	100_ul-mel	high_melatonin
SRR9666142	100_ul-mel_rep-3	3	100_ul-mel	high_melatonin

1. In your spreadsheet program, open **SraRunTable.txt** (downloaded in Subheading 3.3).
2. According to the Sleuth instructions, the first column must be named “sample”; rename the **Run** column to **sample**.
3. Next, you will want to create new columns which specify attributes about each sample such as what treatment/condition correspond to each sample. In Table 2, we suggest columns that indicate the condition (e.g., control, NAA treated, high-melatonin, low-melatonin), replicate numbers, etc. **Tip:** sort your spreadsheet according to run number (ascending) to more easily apply the recommended values in the table. You may delete or retain the additional columns inherited from **SraRunTable.txt**.
4. Save the experimental design file in TSV (tab-separated value) format (e.g., **experimental_design.tsv**).
5. In the Discovery Environment **Data** view, navigate to the **meta-data** folder you created for this experiment (**rna-seq-tutorial/metadata**). Click the **Upload** button and select **Browse Local**.
6. Browse your local computer to select the experimental design file (i.e., **experimental_design.tsv**) and upload the file. You will get a notification when upload is completed. You may need to refresh your browser to see the uploaded file.

3.11 Evaluate Differential Expression with Sleuth

In this final section of the tutorial, we will use the R package Sleuth to visualize our data and perform differential expression analysis. We will summarize the R analysis steps, but the tutorial itself is provided in the form of an RMarkdown notebook (based on the tutorial provided at [19]). As a result, you will be able to follow every step in the notebook without the need to type or modify code. To tailor the provided code, it will be helpful to have some knowledge of R. The R Shiny application that will be displayed at the end of the notebook contains result tables and figures that can be downloaded directly from Shiny on to your computer. At the launch of the application, only the files you select from the Data Store will be immediately available in the RStudio environment. See the Data Store guide for information about using iCommands to transfer data if you wish to do additional imports of data into your VICE session. Alternatively, you can save the outputs of your VICE session and start a new one, this time including the desired data.

1. In the Discovery Environment, click on the Data icon and navigate to your `rna-seq-tutorial` tutorial folder and create a folder to store outputs, name the folder **sleuth_analysis**.
2. In the **Apps** view, search for and launch the **RStudio Sleuth pb** app.
3. In **Analysis Info**, you can name this analysis and provide any comments (optional). Under **Output folder**, navigate to the **sleuth_analysis** folder created earlier. Your outputs will automatically be placed in this folder; click **Next**.
4. In **Parameters for Notebooks**, a default folder containing notebook specific to this tutorial will be loaded (`/iplant/home/shared/cyverse_training/tutorials/pbv3/R`) by default. You may change this if you have an alternative notebook.
5. Under **Datasets** and **Data for analysis**, navigate to the **rna-seq-tutorial** folder created earlier, go into the folder containing your Kallisto output, and select the **Kallisto_quant_output** folder.
6. Under **Datasets** and **Study design file**, navigate to the **rna-seq-tutorial** folder created earlier, go into the **metadata** folder and select the experimental design file (i.e., **experimental_design.tsv**); click **Next**.
7. Click **Next** again to skip **Advanced Settings (optional)**; under **Review and Launch**, click **Launch Analysis**.
8. You will be redirected to the **Analyses** view where you can see the current status of the job; you can also click on the **Analyses** icon to navigate to this section. When the job has the status **Running** you will be able to access the RStudio session. There will be a link icon immediately to the left of the analysis name.

Click this to open the RStudio session in a new browser tab.

Tip: Although the job has **Running** status, it may take a few minutes to access the RStudio session, the amount of is related to the size of the files being transferred into the RStudio environment.

9. In the RStudio session we must modify our RStudio home directory to make it easier to save files. Open the **Terminal** tab. Paste in the following command and hit enter:

```
sudo chown -R rstudio /home/rstudio
```

10. In the RStudio **Files** tab, go to the **R** folder and click **sleuth_pb_tutorial.Rmd** to open the notebook.
11. Follow the notebook by clicking the green “play” button in each section (chunk) of R code. You can follow along with the notebooks explanations and then press play to run each code chunk. The final code chunk will launch the interactive visualizations in the R Shiny application.

Rstudio Outline

Without replicating the actual code presented in the notebook, here are the major steps presented:

- (a) **Step 1:** The Sleuth library and additional libraries for plotting and retrieving data from Ensembl are loaded.
- (b) **Step 2:** The experimental design file is loaded, and a table is created that maps this metadata with the Kallisto outputs.
- (c) **Step 3:** We use the biomaRt package to load gene names from Ensembl so that we can more descriptively annotate our transcripts.
- (d) **Step 4:** We indicate the variables we want to compare and use the Sleuth functions to create the data model.
- (e) **Step 5:** We do an exploratory visualization of the dataset using PCA plotting.
- (f) **Step 6:** A liner model is created, and the results of the analysis are displayed in an interactive R Shiny application. The R Shiny application will generate tables of results and figures that can be downloaded and further analyzed. The test table available from the R Shiny application contains a complete list of gene names, quantifications, and other statistics. You can download this directly from the R Shiny app. Your web browser must have pop-up blocking disabled to view the Shiny application.

12. When you have finished with your RStudio session, return to the **Analysis** window and select (checkbox) the RStudio analysis. Go to the **Analysis** menu and select **Complete and Save Outputs**. Any files created during your RStudio analysis will be saved.

3.12 Conclusion

In this tutorial, we have demonstrated some key features of CyVerse that enable reproducible science at scale. Key functionality areas included:

1. **Data:** Data can be imported and uploaded to the CyVerse Data Store. Our tutorial dataset and analyses total 80 GB of disk space. CyVerse supports terabyte-scale datasets for active analysis with appropriate justification and documentation.
2. **Data Sharing:** These datasets can be shared with fine-grained permissions to other CyVerse users (by username) nearly instantaneously.
3. **Metadata:** Metadata can be applied to files (either by following a template or by designing a spreadsheet of arbitrary attributes). Once applied, these metadata can be used to search rapidly (via elasticsearch). The Data Store documentation also details how metadata can be directly edited in the Discovery Environment (or by command line through the iCommands interface), and how filters and other features can be used to automate the organization of your files.
4. **Reproducible analyses:** Software tools used in the Discovery Environment are containerized (Docker) versions of open source software, making it possible to select the desired versions of software and reproduce previous analyses. The DE's analyses functions keep detailed histories of analyses and parameters.
5. **Interactive analyses:** Through the DE's VICE platform, interactive sessions such as RStudio and R Shiny are used to directly interact with and analyze data.
6. **Computational capacity:** Although not directly highlighted, all applications make use of the underlying CyVerse compute infrastructure. Additionally, some applications in the DE catalog directly make use of XSEDE supercomputing resources.

Taken together, these features provide a high level of functionality that is tailor-made to support data-intensive research and collaboration, all in one place.

4 Notes

1. Using This Tutorial with Your Own Data.

This tutorial starts with sample data imported from the NCBI SRA, but you can easily modify it to work with your own data. You will likely need to upload your data to CyVerse and start at Subheading 3.7 by applying relevant metadata. Here are tips to do this.

- (a) Upload your data to CyVerse. This can be done by uploading from your local computer using third party FTP software (Cyberduck) or transferring the data from a remote server on which iCommands has been installed. See the Data Store Guide at <https://learning.CyVerse.org/>.
- (b) For the Kallisto quantification you will need to adjust the setting dependent upon if you have single or paired-end data. At a minimum, single-end data requires knowledge of the fragment length used in library prep and the standard deviation. See the Kallisto manual for more details on this and other parameters you wish to adjust: <https://pachterlab.github.io/Kallisto/manual>. You will also need to obtain a reference transcriptome for the organism of your choice; these are available from Ensembl (as done in the tutorial) or your choice of databases.
- (c) Like any analysis, the guidance for examining differential abundance as presented in the Sleuth portion of the tutorial will need to be critically examined and adjusted.

References

1. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P (2016) The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol* 14:e1002342. <https://doi.org/10.1371/journal.pbio.1002342>
2. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D et al (2011) The iPlant collaborative: cyberinfrastructure for plant biology. *Front Plant Sci* 2:34. <https://doi.org/10.3389/fpls.2011.00034>
3. Rajasekar A, Moore R, Hou C-Y, Lee CA, Marciano R, de Torcy A et al (2010) iRODS Primer: integrated rule-oriented data system. *Synth Lect Inf Conc Retr Serv* 2(1). <https://doi.org/10.2200/s00233cd1v01y200912icr012>
4. Boettiger C (2015) An introduction to Docker for reproducible research. *Oper Syst Rev* 49: 71. <https://doi.org/10.1145/2723872.2723882>
5. Shah J, Dubaria D (2019) Building modern clouds: using docker, kubernetes google cloud platform. In: 2019 IEEE 9th Annual Computing and Communication Workshop and Conference, CCWC 2019. IEEE, Washington, DC. <https://doi.org/10.1109/CCWC.2019.8666479>
6. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57. <https://doi.org/10.1038/nrg2484>
7. Van den Berge K, Hembach KM, Sonesson C, Tiberi S, Clement L, Love MI et al (2019) RNA sequencing data: Hitchhiker's guide to expression analysis. *Annu Rev Biomed Data Sci* 2:139. <https://doi.org/10.1146/annurev-biodatasci-072018-021255>

8. Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: the teenage years. *Nat Rev Genet* 20:631. <https://doi.org/10.1038/s41576-019-0150-2>
9. Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. *Nucleic Acids Res* 39:D19. <https://doi.org/10.1093/nar/gkq1019>
10. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y et al (2011) The European nucleotide archive. *Nucleic Acids Res* 39:D28. <https://doi.org/10.1093/nar/gkq967>
11. Zia SF, Berkowitz O, Bedon F, Whelan J, Franks AE, Plummer KM (2019) Direct comparison of Arabidopsis gene expression reveals different responses to melatonin versus auxin. *BMC Plant Biol* 19:567. <https://doi.org/10.1186/s12870-019-2158-3>
12. Sherry S, Xiao C, Durbrow K, Kimelman M, Rodarmer K, Shumway M, Yaschenko E (2008) NCBI SRA Toolkit Technology for Next Generation Sequence Data. 1000gconference.Sph.Umich.Edu
13. Andrews S (2010) FastQC. Bioinformatics, Babraham
14. Bray N, Pimentel H, Melsted P, Pachter L (2015) Near-optimal RNA-Seq quantification. ArXiv. <https://arxiv.org/abs/1505.02710>
15. Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34:525. <https://doi.org/10.1038/nbt.3519>
16. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* 14:687. <https://doi.org/10.1038/nmeth.4324>
17. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114. <https://doi.org/10.1093/bioinformatics/btu170>
18. Pachter Laboratory (n.d.) Kallisto manual. <https://pachterlab.github.io/kallisto/manual>
19. Yi L, Pimentel H, Bray N, Pachter L (n.d.) Using p-value aggregation to obtain gene differential expression in datasets with multiple experimental conditions. https://pachterlab.github.io/sleuth_walkthroughs/pval_agg/analysis.html

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.