Introduction to Bioinformatics Resources

NCI CCR Bioinformatics Training and Education Program (BTEP)

Amy Stonelake, Ph.D., Program Manager

June 13, 2023

Attention Summer Trainees

- Today, June 13, Introduction to Bioinformatics Resources
- Tues, June 20, Central Dogma of Molecular Biology: Analyzing DNA, RNA and Proteins
- Tues, June 27, Keeping your Data FAIR: Organizing, Managing, and Sharing your Data
- Tues, July 4 NO CLASS
- Tues, July 11, Introduction to High Performance Computing at NIH: Biowulf
- Tues, July 18, Introduction to R and Python Programming Languages
- Tues, July 25, Managing Bioinformatics Projects with Jupyter Notebook

In this presentation, we will cover

- Bioinformatics resources and training offered by the NCI CCR Bioinformatics Training and Education Program (BTEP)
- Software purchased by OSTR for CCR researchers
- NIH high performance compute cluster Biowulf/Helix
- NCI Cloud Resources
- Resources offered by other NIH training programs

If you are not National Cancer Institute Center for Cancer Research (NCI CCR)…

All resources in this talk are available to you except for CCR/Office of Science Technology and Resources (OSTR) purchased software and some classes

Recordings of Past Events

Slides and recording will be available in our Video Archive (https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/) within 48 hours after an event.

Slides will be available within our Bioinformatics Resources pages.

Topic: Bioinformatics Training and Resources

Bioinformatics Training and Education Program

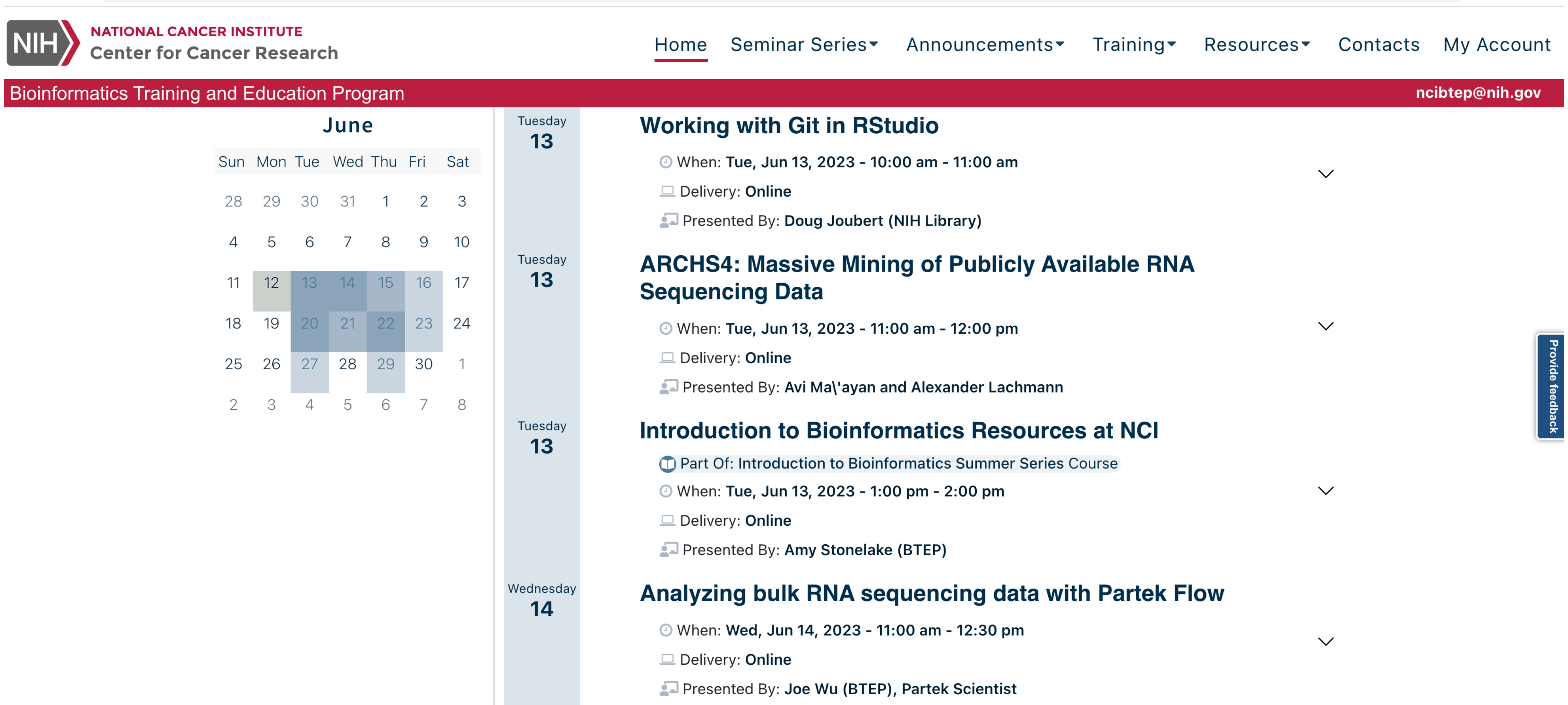
NIH Bioinformatics Calendar at https://bioinformatics.ccr.cancer.gov/btep

Training – Classes and Courses 2023, Distinguished Speaker Seminar Series, Single Cell Annotation Seminar Series, Self-Learning, OSTR/CCR Software

Resources –Video Archive, Class Documentation, Resource Pages, Software List, Bioinformatics, OSTR/CCR Software

Contact BTEP: ncibtep@nih.gov

NIH Bioinformatics Calendar



Bioinformatics Training and Education Program ncibtep@nih.gov

June

Sun	Mon	Tue	Wed	Thu	Fri	Sat
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

Tuesday 13

Working with Git in RStudio

When: **Tue, Jun 13, 2023 - 10:00 am - 11:00 am**

Delivery: **Online**

Presented By: **Doug Joubert (NIH Library)**

Tuesday 13

ARCHS4: Massive Mining of Publicly Available RNA Sequencing Data

When: **Tue, Jun 13, 2023 - 11:00 am - 12:00 pm**

Delivery: **Online**

Presented By: **Avi Ma\'ayan and Alexander Lachmann**

Tuesday 13

Introduction to Bioinformatics Resources at NCI

Part Of: Introduction to Bioinformatics Summer Series Course

When: **Tue, Jun 13, 2023 - 1:00 pm - 2:00 pm**

Delivery: **Online**

Presented By: **Amy Stonelake (BTEP)**

Wednesday 14

Analyzing bulk RNA sequencing data with Partek Flow

When: **Wed, Jun 14, 2023 - 11:00 am - 12:30 pm**

Delivery: **Online**

Presented By: **Joe Wu (BTEP), Partek Scientist**

Let’s Start with Training…

BTEP Training: 2023 Courses

- Classes and Courses 2023 - Programming: R Data Wrangling and Visualization, Python, Unix Command-Line, OSTR/CCR Software
- Distinguished Speakers Seminar Series
-
  Single Cell Annotation Seminar Series
- Working on Biowulf (Unix, R)
- Topics in Bioinformatics - Bulk and Single Cell RNA-Seq, Variant Analysis
- Self-Learning (Dataquest and Coursera Licenses), send email to ncibtep@nih.gov

2023 BTEP Distinguished Speakers Seminar Series

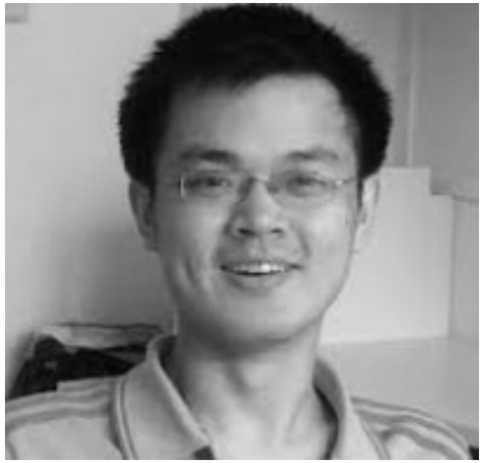
Trey Ideker, AI Models of Cancer in Precision Medicine, March 30 (recording available)

Brandi Davis-Dusenbery, The Power of Connection, Cancer Research Data Commons, May 4 (recording available)

Jennifer Trowbridge, Hematopoietic stem cell contribution to aging and clonal hematopoiesis, Sept 14

Atul Butte, Precisely Practicing Medicine from 700 Trillion Points of Data, October 5

Scott Nicholas Furlan, Nov 2



2023 Single Cell Annotation Seminar Series

Rahul Satija, Azimuth: Annotation of Cell Types in Single Cell Analysis of Cancer, (April 6)

Fabian Theis, Learning and Transferring Cellular State in Single Cell Atlases, (May 25) recording available

Chuan Xu (Sarah Teichmann lab, Cell Typist 2.0), (June 1) recording available

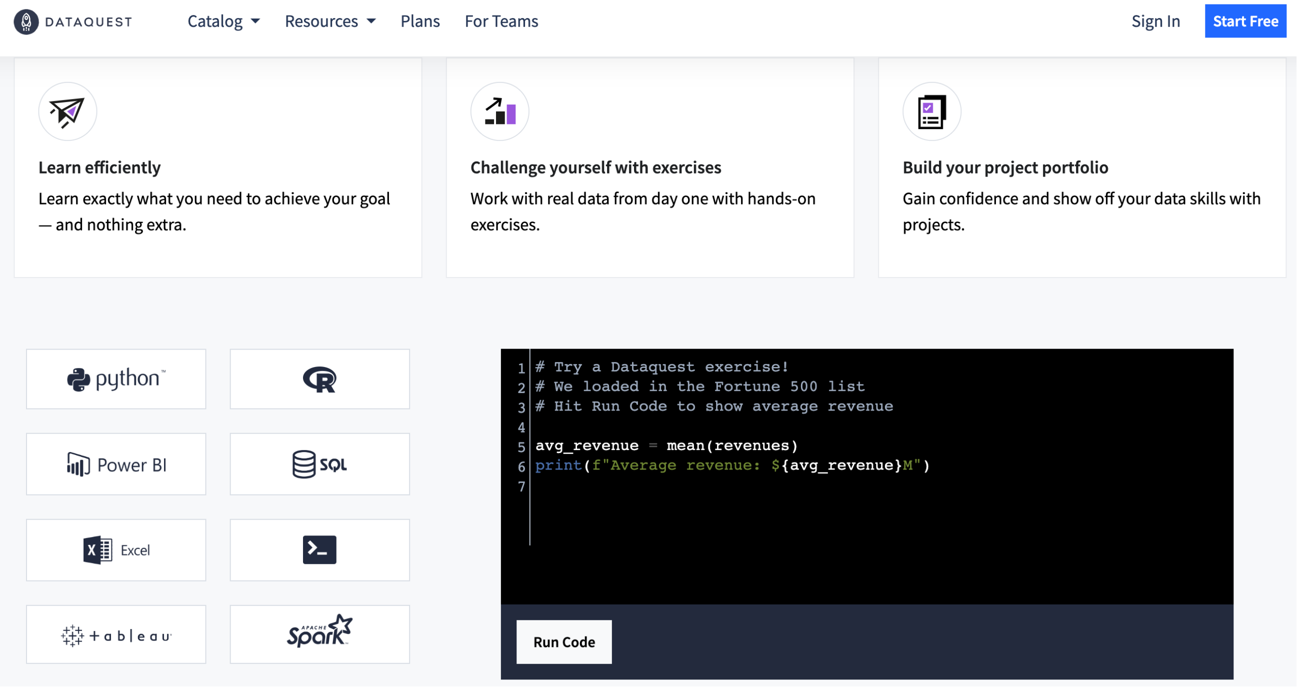
Mallar Bhattacharya, Single Cell Annotation with SingleR, June 22

Cole Trapnell, Whole Embryo Developmental Genetics at Single Cell Resolution, Sept 28

Training:
Dataquest licenses available to NCI CCR Scientists

Why learn Data Science Online with Dataquest?

1. Work at your own pace
2. Interface for typing in commands
3. Many useful topics (Python, R, Excel, Unix Command Line)
4. Courses are laid out from beginner to intermediate to advanced skills (paths)
5. Project-based learning
6. No video lectures
7. Review materials as needed
8. Send email to ncibtep@nih.gov



Learn efficiently

Learn exactly what you need to achieve your goal — and nothing extra.

Challenge yourself with exercises

Work with real data from day one with hands-on exercises.

Build your project portfolio

Gain confidence and show off your data skills with projects.

```
# Try a Dataquest exercise!
# We loaded in the Fortune 500 list
# Hit Run Code to show average revenue

avg_revenue = mean(revenues)
print(f"Average revenue: ${avg_revenue}M")
```

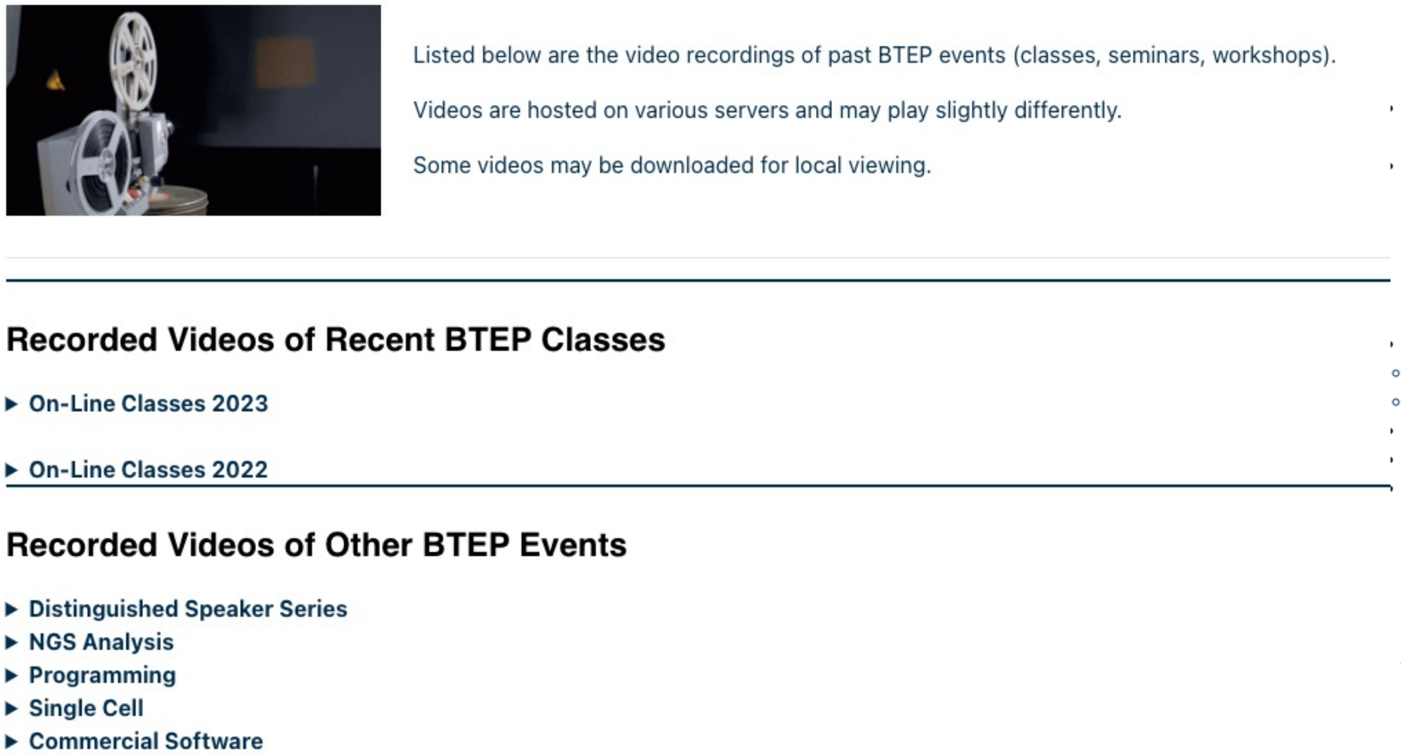
Training: Coursera licenses are available to all NIH, provided by NIH ODSS

- Video lectures
- Work at your own pace
- Large, worldwide, online classes (MOOCS)
- Courses, specializations, and guided projects
- Earn certificates for your resume/CV
- So many courses available: Programming (R, Unix, Python), Genomics, Bioinformatics, Data Science, Language learning

Add some Resources...

BTEP Class Video Archive

https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/



Listed below are the video recordings of past BTEP events (classes, seminars, workshops).

Videos are hosted on various servers and may play slightly differently.

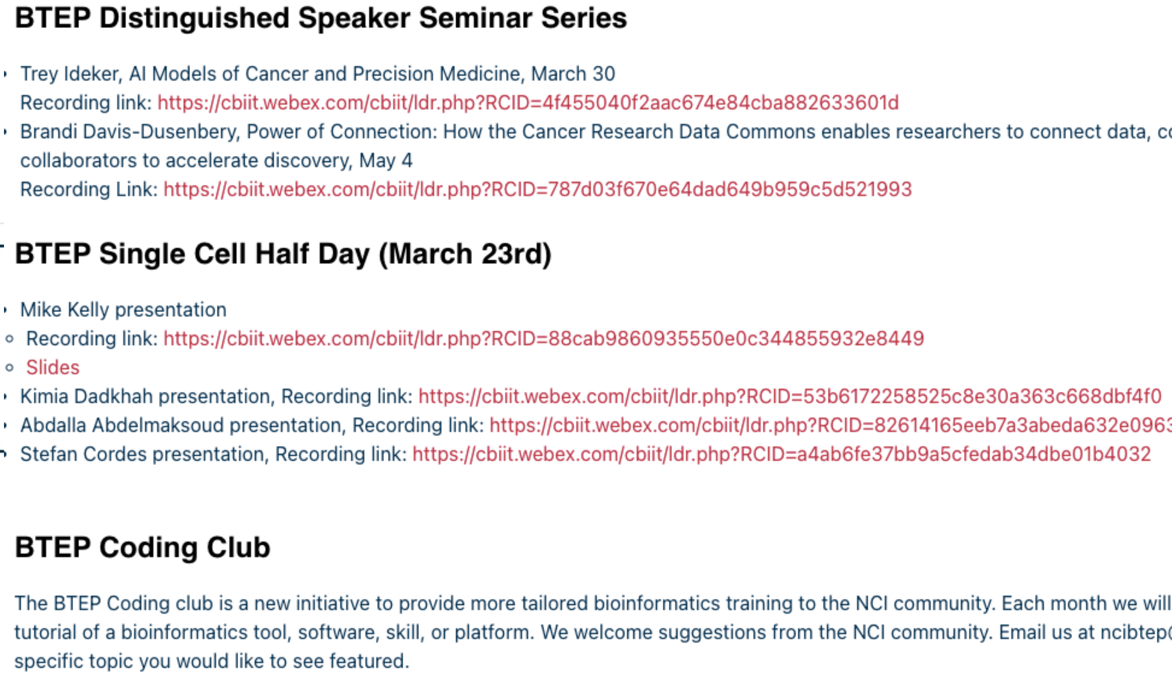
Some videos may be downloaded for local viewing.

Recorded Videos of Recent BTEP Classes

- ▶ On-Line Classes 2023
- ▶ On-Line Classes 2022

Recorded Videos of Other BTEP Events

- ▶ Distinguished Speaker Series
- ▶ NGS Analysis
- ▶ Programming
- ▶ Single Cell
- ▶ Commercial Software



BTEP Distinguished Speaker Seminar Series

- Trey Ideker, AI Models of Cancer and Precision Medicine, March 30
  Recording link: https://cbiit.webex.com/cbiit/ldr.php?RCID=4f455040f2aac674e84cba882633601d
- Brandi Davis-Dusenbery, Power of Connection: How the Cancer Research Data Commons enables researchers to connect data, co collaborators to accelerate discovery, May 4
  Recording Link: https://cbiit.webex.com/cbiit/ldr.php?RCID=787d03f670e64dad649b959c5d521993

BTEP Single Cell Half Day (March 23rd)

- Mike Kelly presentation
  - Recording link: https://cbiit.webex.com/cbiit/ldr.php?RCID=88cab9860935550e0c344855932e8449
  - Slides
- Kimia Dadkhah presentation, Recording link: https://cbiit.webex.com/cbiit/ldr.php?RCID=53b6172258525c8e30a363c668dbf4f0
- Abdalla Abdelmaksoud presentation, Recording link: https://cbiit.webex.com/cbiit/ldr.php?RCID=82614165eeb7a3abeda632e096
- Stefan Cordes presentation, Recording link: https://cbiit.webex.com/cbiit/ldr.php?RCID=a4ab6fe37bb9a5cfedab34dbe01b4032

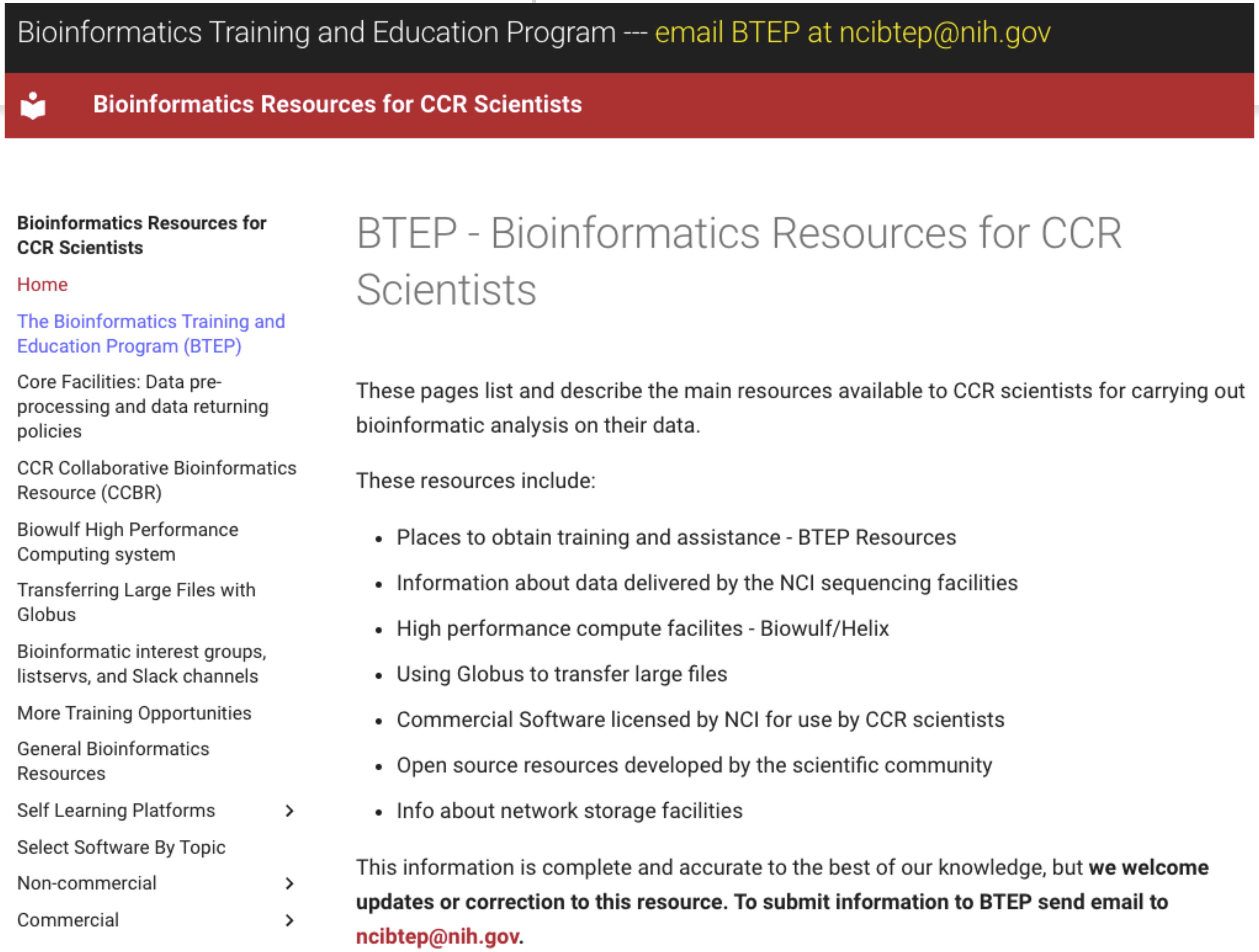
BTEP Coding Club

The BTEP Coding club is a new initiative to provide more tailored bioinformatics training to the NCI community. Each month we will tutorial of a bioinformatics tool, software, skill, or platform. We welcome suggestions from the NCI community. Email us at ncibtep(specific topic you would like to see featured.

Resources: BTEP Resources Pages

Check out the BTEP bioinformatics resources pages :

https://bioinformatics.ccr.cancer.gov/btep



Bioinformatics Training and Education Program --- email BTEP at ncibtep@nih.gov

Bioinformatics Resources for CCR Scientists

Bioinformatics Resources for CCR Scientists
Home
The Bioinformatics Training and Education Program (BTEP)
Core Facilities: Data pre-processing and data returning policies
CCR Collaborative Bioinformatics Resource (CCBR)
Biowulf High Performance Computing system
Transferring Large Files with Globus
Bioinformatic interest groups, listservs, and Slack channels
More Training Opportunities
General Bioinformatics Resources
Self Learning Platforms
Select Software By Topic
Non-commercial
Commercial

BTEP - Bioinformatics Resources for CCR Scientists

These pages list and describe the main resources available to CCR scientists for carrying out bioinformatic analysis on their data.

These resources include:

- Places to obtain training and assistance - BTEP Resources
- Information about data delivered by the NCI sequencing facilities
- High performance compute facilites - Biowulf/Helix
- Using Globus to transfer large files
- Commercial Software licensed by NCI for use by CCR scientists
- Open source resources developed by the scientific community
- Info about network storage facilities

This information is complete and accurate to the best of our knowledge, but **we welcome updates or correction to this resource. To submit information to BTEP send email to ncibtep@nih.gov.**

Bioinformatics Resource for CCR Scientists 2022

Home

The Bioinformatics Training and Education Program (BTEP)

Core Facilities: Data pre-processing and data returning policies

Biowulf High Performance Computing system

Transferring Large Files with Globus

Bioinformatic interest groups, listservs, and Slack channels

More Training Opportunities

Self Learning Platforms ›

Select Software By Topic

Non-commerical Software ›

Commercial Software ›

Getting Started with Biowulf

Biowulf is the NIH high performance computing cluster. It is a linux computing cluster with greater than 105,000 processors. The NIH HPC systems also house "hundreds of scientific programs, packages and databases" (https://hpc.nih.gov/apps/).

Bioinformatic processes often require a lot of memory and computational time, which is limited on individual (local) computers. For bioinformatics tasks that require a lot of memory or can be run in parallel to reduce the time to completion, consider performing such tasks on Biowulf. To obtain a Biowulf account, see the Biowulf help pages. A Biowulf account is accessible to all NIH employees and contractors listed in the NIH Enterprise Directory for a nominal fee of $35 a month.

Working on the NIH High Performance Unix Cluster Biowulf

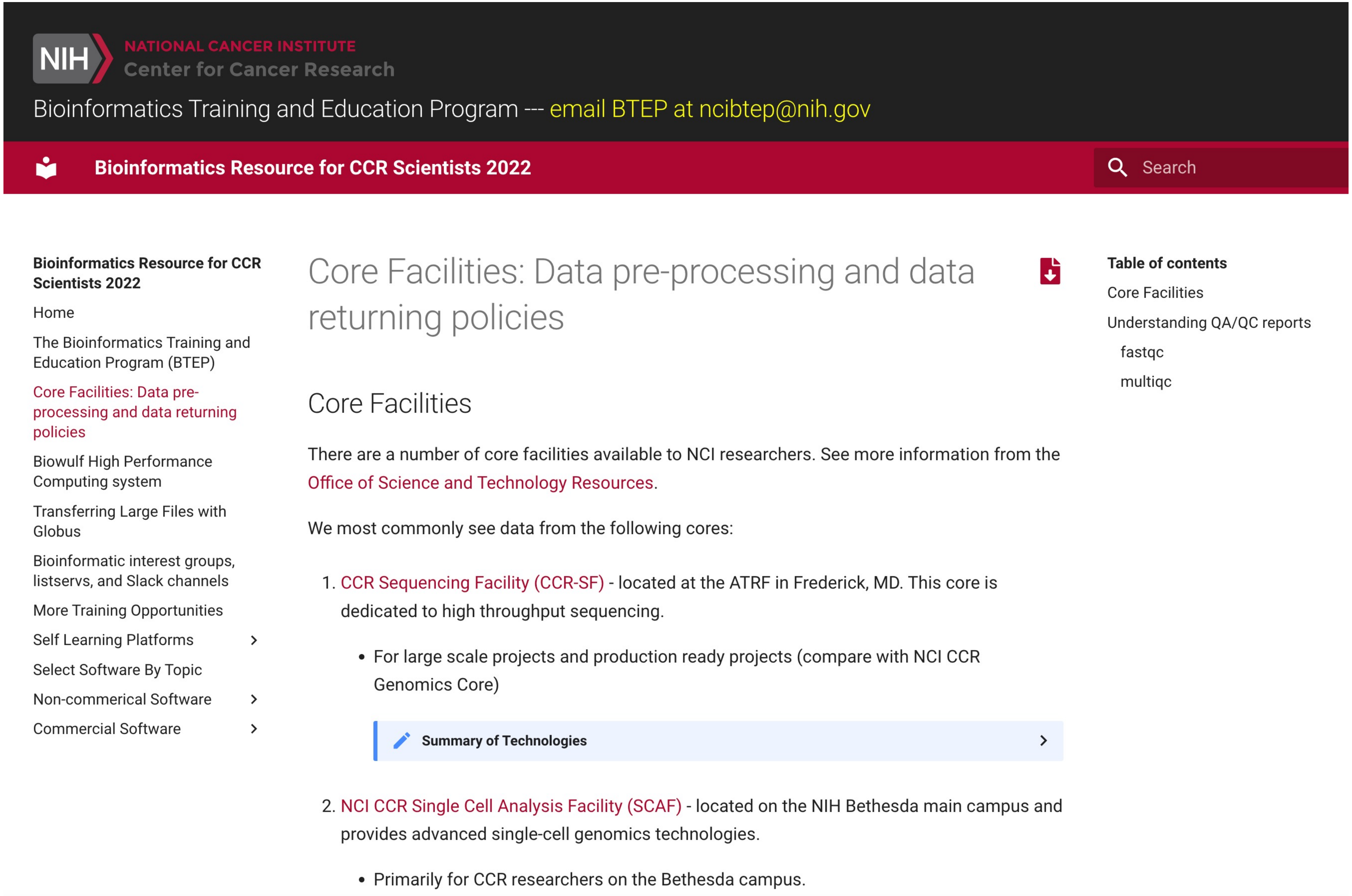
Logging into Biowulf from MacOS

Find the program "Terminal" on your machine, and enter the following statement at the prompt:

```
ssh username@biowulf.nih.gov
```

where "username" is your NIH/Biowulf login username

Table of contents



Bioinformatics Resource for CCR Scientists 2022
Home
The Bioinformatics Training and Education Program (BTEP)
Core Facilities: Data pre-processing and data returning policies
Biowulf High Performance Computing system
Transferring Large Files with Globus
Bioinformatic interest groups, listservs, and Slack channels
More Training Opportunities
Self Learning Platforms ›
Select Software By Topic
Non-commerical Software ›
Commercial Software ›

Core Facilities: Data pre-processing and data returning policies

Core Facilities

There are a number of core facilities available to NCI researchers. See more information from the Office of Science and Technology Resources.

We most commonly see data from the following cores:

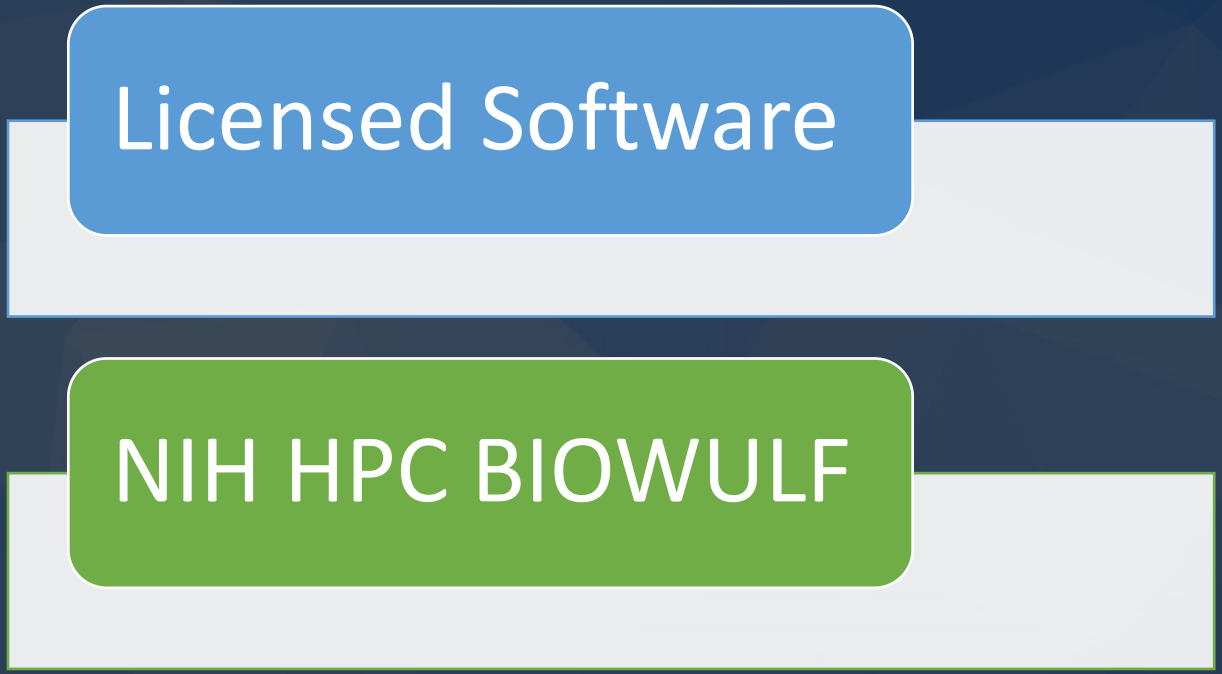
1. CCR Sequencing Facility (CCR-SF) - located at the ATRF in Frederick, MD. This core is dedicated to high throughput sequencing.
    - For large scale projects and production ready projects (compare with NCI CCR Genomics Core)

    Summary of Technologies ›

2. NCI CCR Single Cell Analysis Facility (SCAF) - located on the NIH Bethesda main campus and provides advanced single-cell genomics technologies.
    - Primarily for CCR researchers on the Bethesda campus.

Table of contents

More Resources: Data Analysis Options



Licensed Software

NIH HPC BIOWULF

How should you analyze your data?

Using proprietary, point-and-click software purchased for NCI CCR scientists by Office of Science and Technology Resources (OSTR)

May not always be in an environment where these are available

Partek Flow, and Partek Genomics Suite, Qiagen Ingenuity Pathway Analysis, Qlucore Omics Explorer

Learn open source tools, step-by-step

Office of Science and Technology Resources (OSTR)

PURCHASES SOFTWARE FOR DATA ANALYSIS

MAKES LICENSES TO SOFTWARE AVAILABLE TO ALL CCR RESEARCHERS (SOME ARE ALSO AVAILABLE TO ALL NCI RESEARCHERS, NOT JUST CCR)

NIH LIBRARY OFFERS SOME OF THE SAME LICENSES FOR ALL OF NIH

TYPES OF DATA ANALYSIS: NEXT GEN SEQUENCING, STATISTICS, PATHWAY

NCI CCR OSTR licensed software

- Partek Flow and Partek Genomics Suite
- Qiagen Ingenuity Pathway Analysis and OmicSoft Land Explorer
- Qlucore Omics Explorer
- Qiagen CLC Genomics Workbench
- SnapGene
- LaserGene
- Geneious (Prime)
- Graph Pad Prism

Focus on RNA-Seq Data Analysis

What pipelines and workflows are available to NCI CCR scientists?

How do I know which is most appropriate for my experiment?

Can I get help using these tools?

RNA Seq in the Real World (2023)

Researchers have several options existing to help with RNA-Seq analysis

- CCBR Pipeliner (on Biowulf)
- NIDAP visualization platform (NIH)
- Partek Flow bulk and single cell (OSTR and NIH Library license)
- BTEP Bulk RNA-Seq (B4B) Class

Next Topic: NIH HPC Biowulf

Biowulf (high-performance cluster)

- hpc.nih.gov
- Thousands of analysis tools (modules) maintained by staff
- Scientific reference databases
- Next-gen sequencing, computational chemistry, math, statistics, image analysis
- User guides and training classes
- Monthly Zoom-In Consults see BTEP NIH Calendar

Why you should get to know Biowulf

BIOWULF IS THE HIGH PERFORMANCE CLUSTER (HPC) AT NIH.

IT CAN HOLD A LOT MORE DATA THAN YOUR PERSONAL COMPUTER.

IT HAS MUCH MORE COMPUTE RESOURCES THAN YOUR PERSONAL COMPUTER.

IT CAN HELP YOU ANALYZE “BIG DATA”.

IT IS AVAILABLE TO ALL NCI CCR RESEARCHERS (ALL NIH RESEARCHERS).

Connecting to Biowulf

You can log onto Biowulf from MacOS or Windows PC

You'll need to learn some Unix/ command-line/ shell

Several different ways to use compute resources (batch, swarm)

How to learn more about Biowulf

BTEP Biowulf Beginner Classes

Data Transfer and Sharing on HPC Biowulf

You can mount a Biowulf drive on your local machine

Globus can transfer very large data files between your machine and Biowulf

Transfer to and from Cloud resources

Unix copy commands

(https://hpc.nih.gov/docs/transfer.html)

Cloud Resources Overview

NIDAP – NIH Integrated Data Analysis Platform –bulk and single-cell RNA Seq Analysis Workflows on Palantir Foundry from the CCR Collaborative Bioinformatics Resource (https://ccbr.ccr.cancer.gov/education-training/nidap-workflows/)

NCI Cancer Research Data Commons (CBIIT) including Cancer Genomics Cloud powered by Seven Bridges/Velsera (https://datacommons.cancer.gov/analytical-resource/seven-bridges-cancer-genomics-cloud)

DNAnexus pilot AWS cloud access with both user-friendly GUI and command line interfaces (send email to ncibtep@nih.gov)

Other Bioinformatics Resources and Training

All training events are available on the BTEP NIH Bioinformatics Calendar at https://bioinformatics.ccr.cancer.gov/btep

List serv at list.nih.gov (Bioinformatics, Single Cell, Data Science)

NIH Library offers training classes in software and NGS analyses

Center for Biomedical Informatics and Information Technology (CBIIT)

NIAID Python Courses coming this Fall

Thank you for your support!

Office of Science and Technology Resources (OSTR)

Mariam Malik
Dave Goldstein

Thank you to CCBR, NCBR, and SCAF

Thank you for helping us by providing expert knowledge about bioinformatics tools and resources.

We couldn’t do the training we do without your support.

Maggie Cam (CCBR Lead) and Parthav Jailwala (CCBR Bioinformatics Manager)

Mike Kelly and Staff in Single Cell Analysis Facility (SCAF)

Justin Lack (NCBR Lead)

A big thank you to all the bioinformatics analysts that have answered questions for us and participated in our “Topics in Bioinformatics” Series.

Genome Analysis Unit (GAU) and BTEP Teams

- Peter Fitzgerald
- Carl McIntosh
- Des Tillo
- Amy Stonelake
- Joe Wu
- Alex Emmons

We want to hear from you

Email: ncibtep@nih.gov

What kind of training is helpful to you?