

Introduction to Bioinformatics Resources at NIH

Amy Stonelake, Ph.D.

NCI CCR Bioinformatics Training and Education Program (BTEP), Program Manager

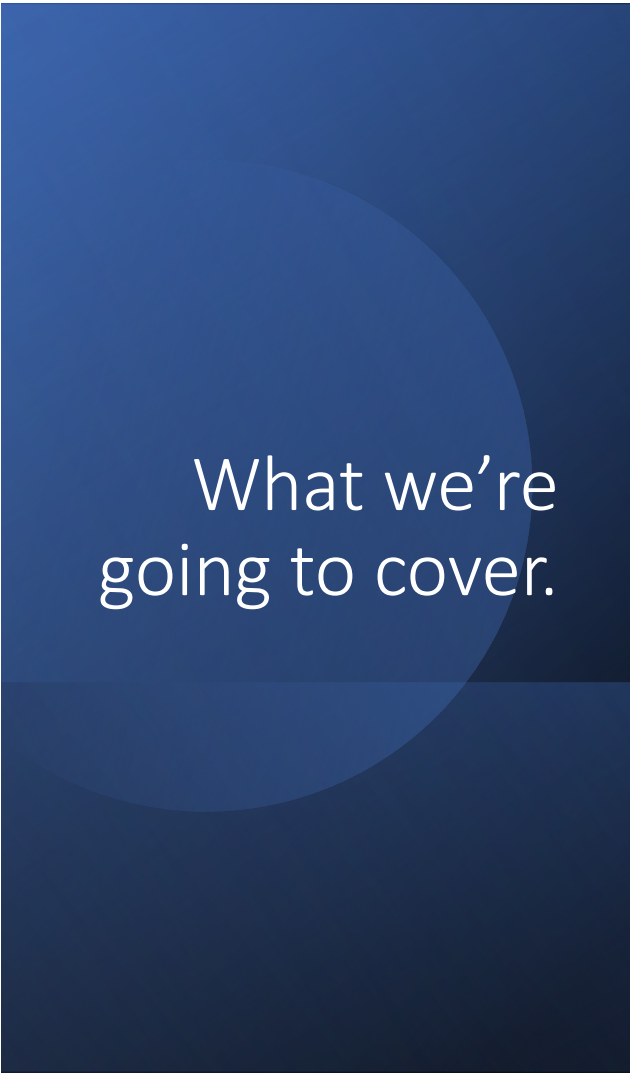
Dec 17, 2024

Bioinformatics Training and Education Program



The only thing you need to remember today:

ncibtep@nih.gov



What we're going to cover.

- The NCI Center for Cancer Research (CCR) Bioinformatics Training and Education Program (BTEP) and NIH Bioinformatics Calendar
- NIH High-Performance Compute Cluster (Biowulf)
- NCI – only Frederick based Compute Cluster (FRCE)
- NIH Library – Bioinformatics Workstations, Classes, Software, Expert Assistance
- Cloud – NIH STRIDES Program Cloud Lab, NCI CRDC (CGC, ISB-Gateway), HTAN, AnVIL (NHGRI), NIGMS Sandbox
- Production Workflows (NCI) – CCBR Github, NIDAP
- Software Licenses – Partek Flow, Qiagen Pathway Analysis, SnapGene, etc.
- Free NIH-wide Licenses – Anaconda, Coursera
- NCI CCR – Dataquest licenses
- NIH List Servs and Teams

Bioinformatics Training & Education Program

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

NCI Center for Cancer Research Bioinformatics Training and Education Program (BTEP)



Amy Stonelake, Ph.D.
BTEP Program Manager and
Bioinformatics Analyst
amy.stonelake@nih.gov



Peter Fitzgerald, Ph.D.
Head *Genome Analysis Unit*
fitzgepe@mail.nih.gov



Alex Emmons, Ph.D.
BTEP Program Trainer
alex.emmons@nih.gov



Desiree Tillo, Ph.D.
Staff Scientist
desiree.tillo@nih.gov



Joe Wu, Ph.D.
BTEP Program Trainer
joe.wu@nih.gov



Carl McIntosh, M.Sc.
Bioinformatics Analyst and
Engineer
mcintoshc@mail.nih.gov

| | | | |
|---|-------------------------|--------------------------------------|---------------------------------------|
| AI in Biomedical Research at NIH Seminar Series | Bioinformatics Bulletin | Class Documentation & Resource Pages | Distinguished Speakers Seminar Series |
| Licenses to Online Learning Platforms | Monthly Coding Clubs | NIH Bioinformatics Calendar | Office Hours - in person and virtual |
| | Training | Video Archive | |

Contact us ncibtep@nih.gov

bioinformatics.ccr.cancer.gov/btep

Our goal is to enable scientists to understand and analyze their own experimental data by providing instruction and training in bioinformatics software, databases, analyses techniques, and emerging technologies.

NCI CCR Bioinformatics Training and Education Program (BTEP)



Began in 2012 to provide training in
bioinformatics software



In 2024, 80+ classes with over 2,000
attendees

Bioinformatics Training & Education Program

Enabling scientists to understand and analyze their own experimental data by providing instruction and training in bioinformatics software, databases, analyses techniques, and emerging technologies.

[Upcoming Classes & Events →](#)

Provide feedback

Classes & Events

Browse Classes, Special Events, and Series Webinars.

[Browse Class Schedule](#)

Bioinformatics Resources

Class Documentation, Core Facilities, and Software.

[Resources & Software](#)

Bioinformatics Forums

Ask Questions about Bioinformatics Topics.

[Questions & Answers](#)

Video Archive

Class and Webinar Recordings and Transcripts.

[Watch Videos](#)

bioinformatics.ccr.cancer.gov/btep

Upcoming Classes & Events

NIH Bioinformatics Calendar

| December | | | | | | |
|----------|-----|-----|-----|-----|-----|-----|
| Sun | Mon | Tue | Wed | Thu | Fri | Sat |
| 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 | 31 | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |

Friday
13

Tuesday
17

Thursday
19

Opportunities to Advance Research for Children, Adolescents, and Young Adults with Cancer through Secondary Data Sharing in the New National Childhood Cancer Registry Data Platform

- 🕒 When: **Fri, Dec 13, 2024 - 12:00 pm - 1:00 pm** 📅 Add to Calendar
- 💻 Delivery: **Online**
- 👤 Presented By: **Johanna Goderre MPH (National Childhood Cancer Registry)**

Introduction to Bioinformatics Resources at NIH

- 🕒 When: **Tue, Dec 17, 2024 - 1:00 pm - 2:00 pm** 📅 Add to Calendar
- 💻 Delivery: **Online**
- 👤 Presented By: **Amy Stonelake (BTEP)**

Creating and Modifying Scatter Plots Using ggplot2: PCA and Volcano

- 🕒 When: **Thu, Dec 19, 2024 - 2:00 pm - 3:00 pm** 📅 Add to Calendar
- 💻 Delivery: **Online**
- 👤 Presented By: **Alex Emmons (BTEP)**

Although
we are a
NCI/CCR
resource...



Most of our events are open to all at NIH



Our website is open to the world



We answer questions within and outside of NIH



Video Archive is open and available



Class documentation and bioinformatics resources
info are open



We advertise our events throughout NIH

Where should I do my work?



LOCAL



HPC



CLOUD

What is Local?

Your NIH laptop or desktop

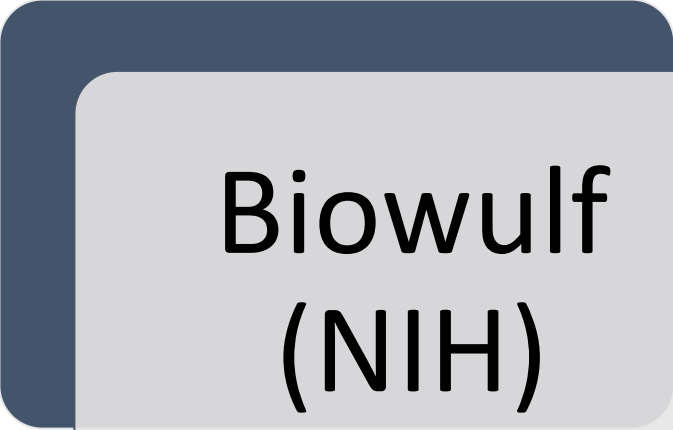


What can you do on your local machine?

- Run Data Analysis Software (CLC Genome Workbench, SnapGene, Qiagen Pathway Analysis)
- Connect to HPC (high-performance cluster) via SSH or GUI
- Learn at your own pace (programming, bioinformatics, data science, AI) on Coursera
- Access the cloud



What is HPC (High-Performance Cluster)?



**Biowulf
(NIH)**



**FRCE
(Frederick)**



hpc.nih.gov



Should I work on my local machine or Biowulf?

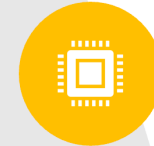
Why you should get to know Biowulf



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 NIH RESEARCHERS

Working on Biowulf – Things to Know

hpc.nih.gov

Minimal, fixed monthly cost

Thousands of scientific applications (software) available

Reference Data (NCBI db, BLAST db, genomic alignment data)

Login/head node

Learn some beginner Unix (Command Line)

Your home directory and your data directory

Start an interactive node (sinteractive)

Setting up swarm jobs (repetitive jobs)

Running batch jobs (scripts/programs)

Moving big data (Globus)

<https://bioinformatics.ccr.cancer.gov/docs/resources-for-bioinformatics/Biowulf/>

HPC on Demand GUI (RStudio, VSCode, JupyterNotebook, IGV, Matlab)

Frederick Research Computing Environment (FRCE) NCI only



Available within NCI at no cost, optimized for Frederick researchers



250+ scientific applications installed



Next Gen Sequencing, structural biology, cryogenic electron microscopy, imaging and AI, text mining

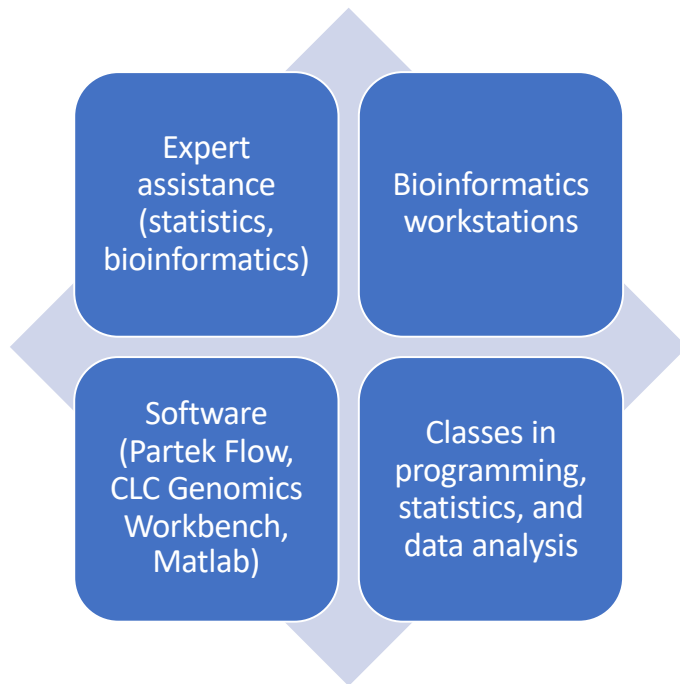


Log in ssh or GUI, file transfers with Globus, batch and interactive jobs



<https://ncifrederick.cancer.gov/staff/frce/>

NIH Library



NIH Library Training
December & January Classes
Now available for registration
nihlibrary.nih.gov/training/calendar
Classes are free for NIH and select HHS staff

NIH Library
nihlibrary.nih.gov

NIH Library Training Classes: December 2024 and January 2025

Register now for NIH Library Training classes being offered in December and January!

nihlibrary.nih.gov

Cloud

- NCI Cloud Resources
- NIH STRIDES Initiative – NIH Cloud Lab
- NHGRI AnVIL



How is the Cloud different from HPC?

- Cost based on usage
- Flexible compute power
- Analyze publicly available data

How is Cloud similar to HPC?

- Large compute resources
- Pre-installed software and tools
- Bioinformatics and data science analyses
- Analyze your own data

Publicly
Available
NCI Cloud
Resources

Cancer Research Data
Commons (CRDC)

Cancer Genomics
Cloud/7Bridges/Velsera

ISB Gateway in the Cloud/CGC

Human Tumor Atlas Network
(HTAN)

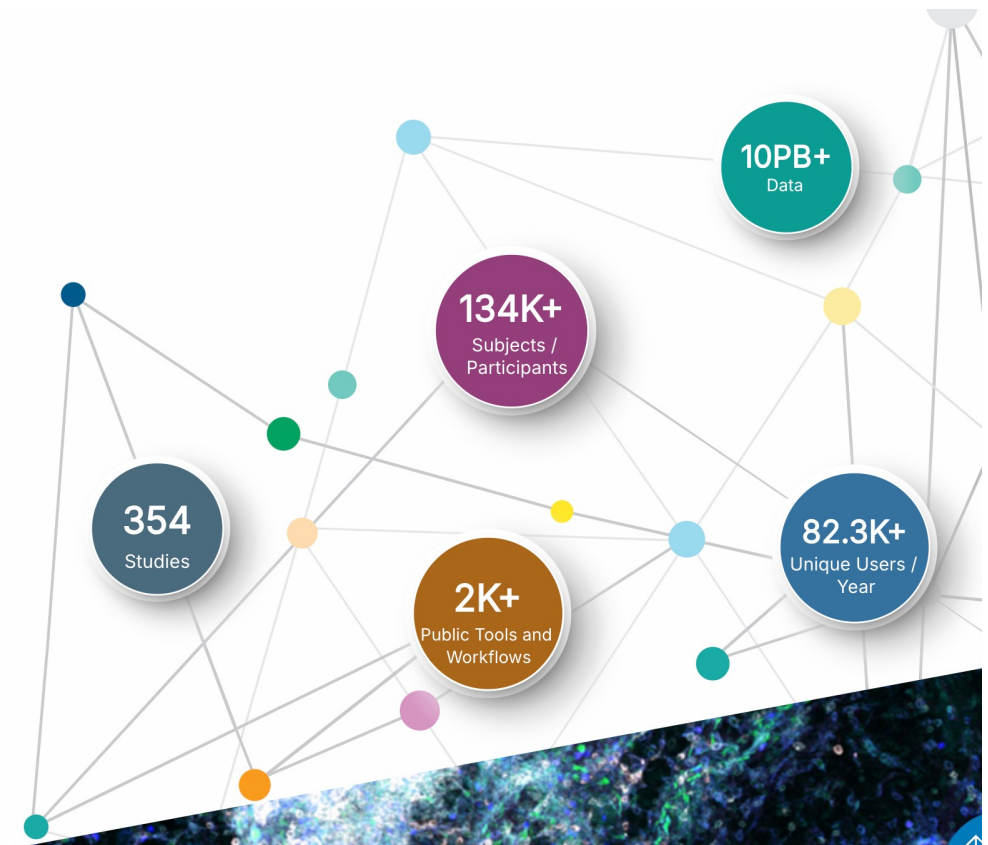
NCI Cancer Research Data Commons

Connecting Data to Accelerate Cancer Research

The NCI Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that provides secure access to a large, comprehensive, and expanding collection of cancer research data. Users can explore and use analytical and visualization tools for data analysis in the cloud.

[Watch CRDC Video](#)

datacommons.cancer.gov



NCI Cancer Research Data Commons

Explore

DATA COMMONS



Genomic Data Commons (GDC)

Share, analyze, and visualize harmonized genomic data, including TCGA, TARGET, and CPTAC.



Proteomic Data Commons (PDC)

Share, analyze, and visualize proteomic data, such as CPTAC and The International Cancer Proteogenome Consortium (ICPC).



Imaging Data Commons (IDC)

Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies.



Integrated Canine Data Commons (ICDC)

Share data from canine clinical trials, including the PRE-medical Cancer Immunotherapy Network Canine Trials (PRECINCT) and the Comparative Oncology Program.



Cancer Data Service (CDS)

Store and share NCI-funded data that are not hosted elsewhere to further advance scientific discovery across a broad range of research areas.



Clinical and Translational Data Commons (CTDC)

Store and share data from NCI-funded Clinical and Translational Studies.

CORE STANDARDS AND SERVICES



Cancer Data Aggregator (CDA)

Enables users to query and connect data distributed across the CRDC for integrative analysis.



Data Standards Services (DSS)

Provides services to facilitate interoperability of data across CRDC.



Data Commons Framework (DCF)

Provides secure user authentication and authorization and permanent digital object identifiers for data objects.

datacommons.cancer.gov

NCI Cancer Research Data Commons

CLOUD RESOURCES



Broad Institute FireCloud

Access NCI-funded datasets TARGET and TCGA along with a rich collection of other datasets and collaborative projects that are part of the biomedical ecosystem. Run analysis tools at scale and collaborate securely on a scalable cloud environment.



Seven Bridges Cancer Genomics Cloud developed by Velsera (SB-CGC)

Explore and analyze large datasets alongside secure and scalable analytical resources for large-scale computational research.



ISB Cancer Gateway in the Cloud (ISB-CGC)

Access data sets using fully interactive web-based applications, including BigQuery, which is hosted on Google Cloud Platform.

datacommons.cancer.gov

Seven Bridges Cancer Genomics Cloud/Velsera

CANCER GENOMICS CLOUD

The Seven Bridges Cancer Genomics Cloud (CGC), powered by Velsera and funded by the NCI, is a flexible cloud platform that enables analysis, storage, and computation of large cancer datasets. The CGC provides a user-friendly portal to access and analyze cancer data where it lives. With the CGC, any user with an account can easily access petabytes of cancer data, share it, analyze and use the computational power of the cloud without having to learn how to program and get familiar with several different data portals.

cancergenomicscloud.org

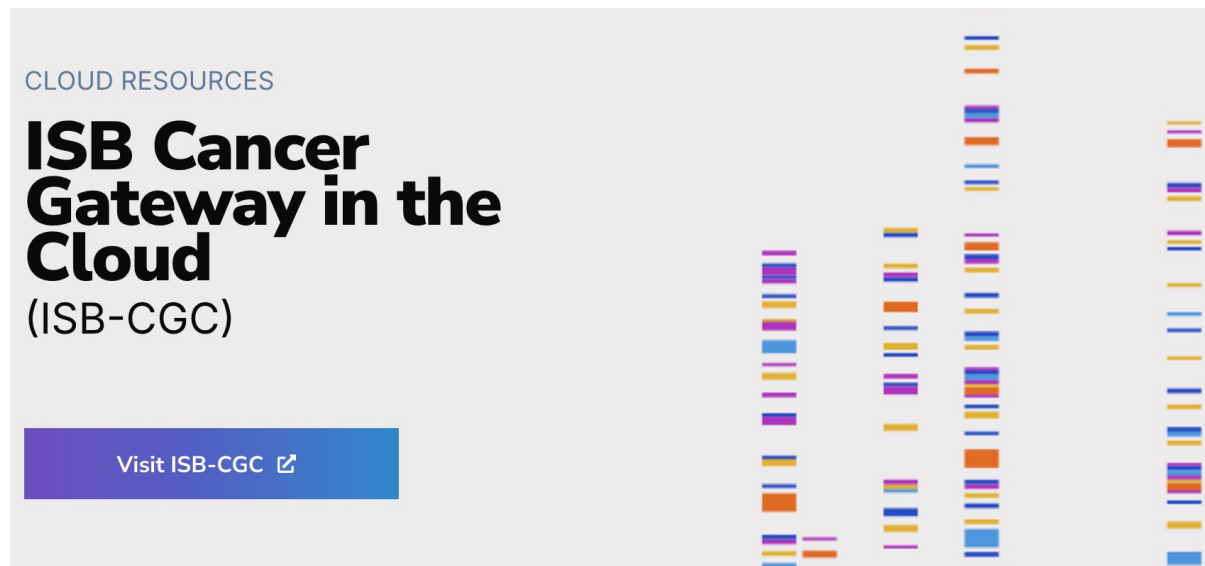
Cancer Genomics Cloud (CGC) SBR/Velsera

- Users can access:
 - Genomics data
 - Proteomics data
 - TCGA data
 - Data from multiple species
 - Their own data
 - Tool Library (850+)
 - Common Workflow Language (CWL)
 - Bring your own pipeline or tools
 - Jupyter Lab and Rstudio available

cancergenomicscloud.org

ISB Cancer Gateway in the Cloud (ISB-CGC)

- GoogleBigQuery – query across multiple data tables
- TCGA and TARGET data

A banner for the ISB Cancer Gateway in the Cloud (ISB-CGC). The banner has a light gray background. On the left, the text "CLOUD RESOURCES" is in a small, blue, sans-serif font. Below it, the main title "ISB Cancer Gateway in the Cloud" is in a large, bold, black font, with "(ISB-CGC)" in a smaller black font underneath. A blue button with white text "Visit ISB-CGC" and a small white icon of a link with an arrow is positioned below the title. On the right side of the banner, there are four vertical columns of colorful horizontal bars, representing genomic data. The colors include purple, blue, orange, yellow, and red.

CLOUD RESOURCES

ISB Cancer Gateway in the Cloud

(ISB-CGC)

Visit ISB-CGC [↗](#)

<https://datacommons.cancer.gov/analytical-resource/isb-cancer-gateway-cloud>

Human Tumor Atlas Network

Human Tumor Atlas Network

HTAN is a National Cancer Institute (NCI)-funded Cancer MoonshotSM initiative to construct 3-dimensional atlases of the dynamic cellular, morphological, and molecular features of human cancers as they evolve from precancerous lesions to advanced disease. (*Cell April 2020*)

[Explore latest Data](#)

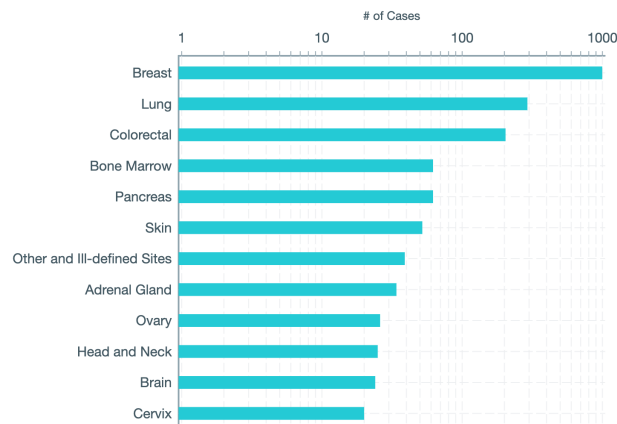
[Learn more about HTAN](#)

humantumoratlas.org

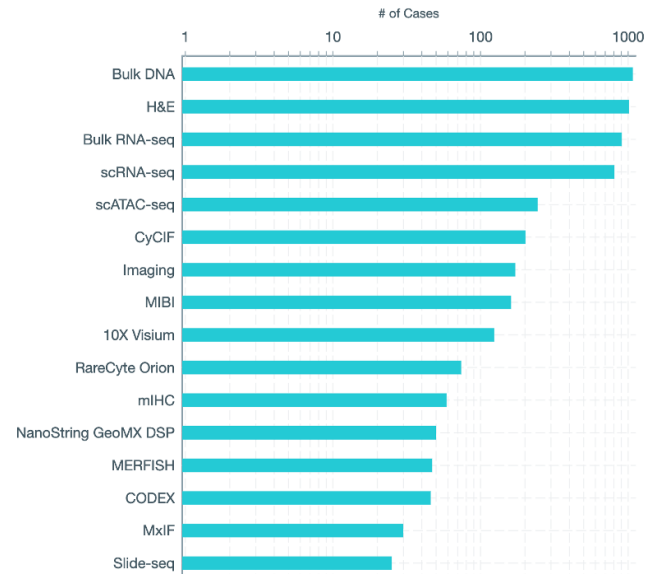
Human Tumor Atlas Network Organs and Assays

14 Atlases **21** Organs **2147** Cases **9286** Biospecimens

The latest HTAN data release includes tumors originating from 21 tumor sites:



The tumors were profiled with 30 different types of assays:



humantumoratlas.org

NIH STRIDES Initiative


Accelerating Biomedical
Research

NIH Office of Data Science
Strategy (ODSS)

cloud.nih.gov

NIH Cloud Lab

NIH STRIDES Initiative



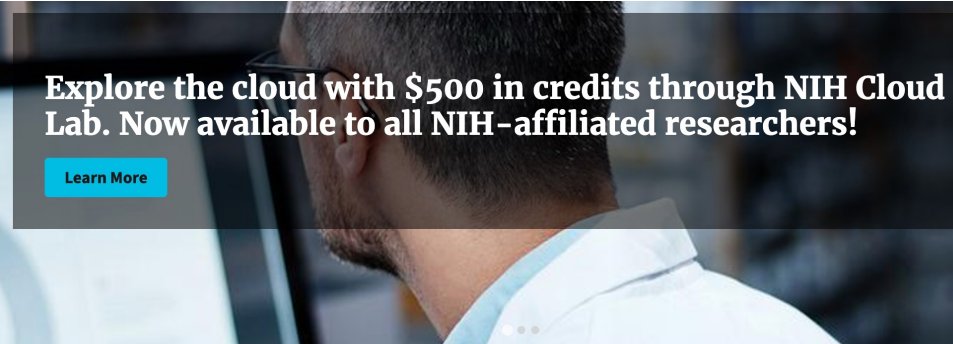
Helping advance biomedical research by delivering access to industry-leading cloud providers

The STRIDES Initiative aims to help NIH and its institutions accelerate biomedical research by reducing barriers in utilizing commercial cloud services. This initiative aims to harness the power of the cloud to accelerate biomedical discovery. NIH and NIH-funded researchers can take advantage of STRIDES benefits.

[Enroll Now](#)

cloud.nih.gov

NIH STRIDES Cloud Lab



Explore the cloud with \$500 in credits through NIH Cloud Lab. Now available to all NIH-affiliated researchers!


[Learn More](#)


Benefits:


- Discounts on partner services
- Professional services consultations
- Potential collaborative engagements

See our **[Benefits](#)** page for additional information.

Partners:

 **aws**

 **Google Cloud**

 **Microsoft Azure**

cloud.nih.gov

NHGRI Analysis Visualization and Informatics Lab-space (AnVIL)



anvilproject.org



Free trials available





Tools

Dockstore – create and share docker-based workflows
NCPI – interoperate with other NIH data commons
Bioconductor
Galaxy
Jupyter (python, R)



Data sources

Telomere-to-telomere genome (T2T)
1000G thousand genomes
Genotype-tissue expression project (GTEx)



Workflows (pipelines)


CCBR Workflows



CCR Collaborative Bioinformatics Resource

Bioinformatics assistance to further CCR researchers' goals.

[Support Process →](#)



NIDAP Training

Online training for interactive CCBR workflows for bioinformatics analyses on NIDAP. Currently released workflows include: Bulk RNA-seq, Single-cell RNA-seq, and Digital Spatial Profiling (DSP).


[NIDAP Trainings](#)



Project Support

Learn how CCBR can assist with CCR Researchers with their projects.

[Explore Process](#)



Pipelines & Workflows

Workflows & Pipeline development. View whole exome & genome, single cell RNA-Seq, and ChIP-Seq pipeline examples.

[Workflows & Pipelines](#)

bioinformatics.ccr.cancer.gov/ccbr

CCBR Pipelines and NIDAP (NIH Integrated Data Analysis Platform)

CCBR pipelines (github/NIH)

CCBR NIDAP (NIH)

Free to use

Consult CCBR in planning stage of experiment

Don't wait til you have data

Submit help request for analysis project

bioinformatics.ccr.cancer.gov/ccbr

Next Gen Seq Workflows/Pipelines

CCR Collaborative Bioinformatics Resource (CCBR)

github.com/ccbr

About Us

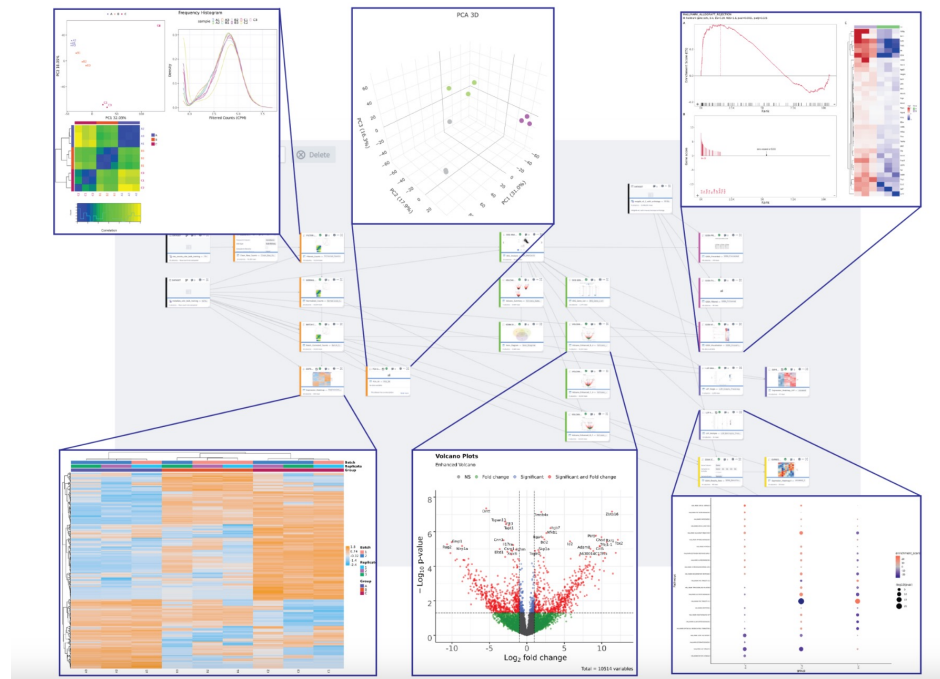
- 🙋 Hi, we're the [@CCBR](https://twitter.com/ccbr), a group of bioinformatics analysts and engineers
- 📄 We build flexible, reproducible, workflows for next-generation sequencing data
- 💡 We [collaborate](#) with [CCR](#) Pls
- 📧 You can reach us at ccbr_pipeliner@mail.nih.gov
- 📖 Check out our [release history](#)
- 🔗 Our [Zenodo](#) community

Available NGS Pipelines/Workflows

Here is a list of our prominent pipelines and their release schedule on BOWULF:

| Data Type | Pipeline Name | CLI* availability date | GUI* availability date |
|--------------------------|---|------------------------|------------------------|
| RNASeq ¹ | RENEE <small>snakemake</small> | July 3rd 2023 | July 14th 2023 |
| WESSeq ² | XAVIER <small>snakemake</small> | July 21th 2023 | Sep 1st 2023 |
| ATACSeq ³ | ASPEN <small>snakemake</small> | November 30th 2023 | TBD |
| ChIPSeq ⁴ | CHAMPAGNE <small>nextflow</small> | October 15th 2023 | TBD |
| CRISPRSeq ⁵ | CRISPIN <small>nextflow</small> | September 31st 2023 | TBD |
| CUT&RunSeq ⁶ | CARLISLE <small>snakemake</small> | October 31st 2023 | TBD |
| EV-Seq ¹⁰ | ESCAPE <small>snakemake</small> | March 26th, 2024 | TBD |
| circRNASeq ⁷ | CHARLIE <small>snakemake</small> | <i>Jul 31st 2024</i> | TBD |
| scRNASeq ⁸ | SINCLAIR <small>nextflow</small> | <i>Sep 30th 2024</i> | TBD |
| WGSSeq ⁹ | LOGAN <small>nextflow</small> | <i>Sep 30th 2024</i> | TBD |
| spatialSeq ¹¹ | SPENCER <small>nextflow</small> | TBD | TBD |

Bulk RNA-Seq Analysis on NIDAP (NIH Integrated Data Analysis Portal)



<https://bioinformatics.ccr.cancer.gov/ccbr/education-training/nidap-training/>

Commercial Software Available (NCI)



Partek Flow – RNA-Seq, CITE-Seq, ATAC-Seq (also available to NIH via NIH Library)



Qiagen Ingenuity Pathway Analysis (IPA) – from differential expression analysis (DEA) to pathways, biomarkers, drug targets (also available to NIH via NIH Library)



CLC Genomics Workbench (also available to NIH via NIH Library)



Qlucore Omics Explorer – RNA-Seq, visualizations, statistical analysis (NCI only)



Request access at service.cancer.gov (NCI)



SnapGene – molecular biology (alignment, PCR primers, cloning) (NCI only)



Free licenses for all at NIH

Anaconda -
flexible python
and R
environment

Coursera -
learning platform

Anaconda

What is Anaconda?

A distribution of the Python and R programming languages that aims to simplify package management and deployment (wiki)

"The operating system for AI"

The National Institutes of Health (NIH) offers a free Anaconda license to its employees and contractors. The license includes:

Technical support from the NIH Anaconda team

Access to secure, curated, and encrypted Python and R packages

Data science courses led by experts

Sample notebooks and extensions

Jupyter notebooks that are ready to code

[Anaconda License Request Form](#)

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



Free for everyone at NIH



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 (thousands) available: Programming (R, Unix, Python), Genomics, Bioinformatics, Data Science, Language learning, AI



<https://bioinformatics.ccr.cancer.gov/btep/self-learning/>

Dataquest (NCI CCR only)

Dataquest.io

Programming – Python, R, Unix/Bash, SQL, PowerBI, Excel, AI

Provides interface for typing in code

Free to NCI Center for Cancer Research only

list.nih.gov Interest Groups and Mailing Lists



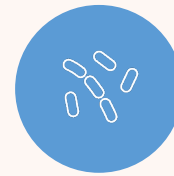
ARTIFICIAL-
INTELLIGENCE



BIOINFORMATICS-SIG-L



DATA-SCI ICE



SINGLECELLGENOMICS
(AND SPATIAL
TRANSCRIPTOMICS)



SYSBIOSIG-L



Email: ncibtep@nih.gov