Bioinformatics Resources for CCR Scientists

Center for Cancer Research

BTEP

Bioinformatics Training
& Education Program

BTEP Team
BTEP/GAU/CCR/NCI/NIH - email ncibtep@mail.nih.gov
Bioinformatics Training and Education Program

Table of Contents

Home

The Bioinformatics Training and Education Program (BTEP)

Core Facilities: Data pre-processing and data returning policies

Advanced Biomedical Computational Science Group

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 *(mailto:ncibtep@nih.gov)*.

The Bioinformatics Training and Education Program (BTEP)

What is BTEP?

BTEP *(https://btep.ccr.cancer.gov)* is an Office of Science and Technology Resources (OSTR) program dedicated to

> increasing the awareness and understanding of bioinformatics techniques and processes among CCR scientists, with the goal of empowering CCR scientists to perform a basic, informed set of analyses on their own behalf. --- BTEP *(https://btep.ccr.cancer.gov)*

BTEP organizes talks on bioinformatic related topics and hosts trainings on commercial and non-commercial bioinformatics tools. In addition, the BTEP website is an excellent resource for finding NIH wide bioinformatics events (See the NIH Bioinformatics Training Calendar *(https://btep.ccr.cancer.gov)*) and learning more regarding bioinformatics software available to NCI researchers (See the BTEP Software pages *(https://bioinformatics.ccr.cancer.gov/btep/resources/scientific-software/)*).

NIH BTEP Calendar

The NIH Bioinformatics Training Calendar *(https://btep.ccr.cancer.gov/)* contains descriptions and links to bioinformatics training events and talks on related topics from all over the NIH campus, not just BTEP sponsored events.

Announcements

Announcements *(https://bioinformatics.ccr.cancer.gov/btep/411-2/)* are in the form of "news" *(https://bioinformatics.ccr.cancer.gov/btep/news-2/)*, including updated information about online learning resource licenses (Coursera, Dataquest) and available NGS analysis software packages (Partek Flow, Qlucore, Qiagen, etc.), and "bulletins" *(https://bioinformatics.ccr.cancer.gov/btep/bulletins/)*. The bulletins are a monthly resource sent via email to all CCR employees; these include featured events, announcements, and topic spotlights.

Distinguished Speaker Seminar Series

Each year, BTEP hosts a Distinguished Speaker Seminar Series *(https://bioinformatics.ccr.cancer.gov/btep/seminar/Distinguished+Speakers)* with notable guest speakers from around the world involved in groundbreaking cancer -omics research or research of interest to the CCR community.

- Speakers for 2024 include Casey Greene (CU Anschutz), Caroline Uhler (MIT), Angela Brooks (UCSC), Rafael Irizarry (Harvard), Olivier Elemento (Weill Cornell Medicine), Elaine Mardis (Nationwide Children's Hospital), Rachel O'Neill (Univ. of Connecticut), Seth Blackshaw (Johns Hopkins), and Carol Bult (The Jackson Lab).

BTEP Coding Club

The BTEP Coding club *(https://bioinformatics.ccr.cancer.gov/btep/seminar/Coding+Club)* is an initiative to provide more tailored bioinformatics training to the NCI community. Each month the Coding Club features a 1-hour demo / tutorial of a bioinformatics tool, software, skill, or platform. These tutorials / demos *(https://bioinformatics.ccr.cancer.gov/docs/btep-coding-club/)* range in experience level from beginner to advanced. We welcome topic suggestions from the NCI community. Email us at ncibtep@nih.gov *(mailto:ncibtep@nih.gov)* if there is a specific topic you would like to see featured.

Other Seminar Series

BTEP occasionally hosts other seminar series with more focused topics. Examples from 2024 include AI in Biomedical Research @ NIH *(https://bioinformatics.ccr.cancer.gov/btep/seminar/AI+in+Biomedical+Research+%40+NIH)* and Getting Started with scRNA-Seq *(https://bioinformatics.ccr.cancer.gov/btep/seminar/Getting+Started+with+scRNA-Seq)*.

FAQ Forums

BTEP maintains several Question and Answer Forums of interest to the NCI/CCR community, where researchers can ask questions on specific bioinformatic topics. There are currently FAQ forums for the following topics:

- Single Cell RNA-Seq *(https://bioinformatics.ccr.cancer.gov/btep/questions/?category=Single-Cell+RNA-Seq)*
- ChIP-Seq *(https://bioinformatics.ccr.cancer.gov/btep/questions/?category=ChIP-Seq+Data+Analysis)*

Upcoming Classes

BTEP offers trainings on diverse topics such as cancer biology resources, machine learning / AI, NIH computational resources, OSTR licensed software, Programming (Unix command line, R, and Python), and next-generation sequencing analysis methods (RNA-Seq, ATAC-Seq, Variant Calling, microbiome, etc.). Upcoming classes hosted by BTEP can be found at https://

bioinformatics.ccr.cancer.gov/btep/classes/#BTEP *(https://bioinformatics.ccr.cancer.gov/btep/classes/#BTEP)*.

Most BTEP events are recorded and recordings are available in the BTEP Video Archive *(https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/)* 24-48 hours following an event. Many BTEP training events are also associated with additional documentation *(https://bioinformatics.ccr.cancer.gov/btep/class-documents/)* including the code, data, and other information pertaining to a given topic.

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 *(https://ostr.ccr.cancer.gov/resources/core)*.

We most commonly see data from the following cores:

1. CCR Sequencing Facility (CCR-SF) *(https://ostr.ccr.cancer.gov/resources/sequencing-facility/)* - 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**
   >
   > - Illumina Short Read Sequencing
   >   - ChIP-Seq
   >   - Cut and Run
   >   - ATAC-Seq (only for pilot projects)
   >   - RNA-Seq (mRNA, Total RNA and microRNA)
   >   - Whole Genome Sequencing
   >   - Whole Exome Sequencing
   >   - Methylated DNA sequencing (bisulfite)
   >   - Amplicon Sequencing
   > - Long reads / PacBio Sequencing
   >   - Whole Genome Sequencing
   >   - RNA Sequencing
   >   - Targeted Sequencing
   >   - HLA Typing
   >   - 16S sequencing
   > - Short read and long read protocols for single cell
   > - Optical mapping with Bionano Genomics

2. NCI CCR Single Cell Analysis Facility (SCAF) *(https://ostr.ccr.cancer.gov/emerging-technologies/single-cell-analysis/)* - located on the NIH Bethesda main campus, building 41, and provides advanced single-cell genomics technologies.
   - Primarily for CCR researchers on the Bethesda campus.

   > **Summary of Technologies**
   >
   > - 10X Genomics Chromium system
   > - Advanced Methods: Plate-based single cell approaches (e.g., Smart-Seq2)

- See the SCAF webpage *(https://ostr.ccr.cancer.gov/emerging-technologies/single-cell-analysis/)* for information on emerging technologies

3. NCI CCR Genomics Core *(https://genomics.ccr.cancer.gov/)* - located on the NIH Bethesda main campus, building 41.

    - Rapid turnover for smaller projects (compare with CCR-SF)

Summary of Technologies

- Next Generation Sequencing (iSeq 100, MiSeq, NextSeq 550 and the NextSeq 2000)
    - Applications include targeted gene sequencing (amplicon and targeted enrichment), metagenomics, gene expression studies, ChIP-Seq and RNA-Seq
- Sanger Sequencing
- Digital Gene Expression
- Digital droplet PCR
- Analytical / Preparative electrophoresis
- Automation
- NanoString GeoMX DSP
- Oxford Nanopore MinION

Data from these cores will likely undergo some form of pre-processing. Additionally, cores may return data to the user in different ways. See below for current core protocols.

Core data pre-processing protocols

CCR-SF

For all projects, CCR-SF conducts primary and secondary analyses including initial base-calling, demultiplexing, data quality control, and reference genome alignment of NGS reads. Tertiary analyses may also be conducted on a project by project basis. For more information, refer to the CCR-SF FAQs *(https://ostr.ccr.cancer.gov/resources/sequencing-facility/faq-bioinformatics/)*.

SCAF

For a standard 10x assay against a standard reference, you can expect the raw sequencing data to be processed through to the Genomics cellranger output, including all quality control steps and troubleshooting in between. Otherwise, the degree of bioinformatic support will vary based on the project and individual needs. Non-standard projects generally require the development of a custom data processing workflow. As such, SCAF will conduct base-level analyses to ensure assay performance. In limited cases, the SCAF will also perform secondary analysis steps including bioinformatic analysis, interpretation, figure generation, and dataset submission.

NCI CCR Genomics Core

For NGS data, the NCI CCR Genomics Core will generate fastq files and initial QC metrics (if requested).

In addition,

> The Core has a dedicated bioinformatics consultant who advises customers on appropriate experimental design, interpretation of QC data and helps to direct users to the existing

> bioinformatics tools under CCBR and other available bioinformatic entities. --- NCI CCR Genomics Core *(https://genomics.ccr.cancer.gov/)*

How will my data be returned to me? ?

CCR-SF ?

For information on how data is returned from CCR-SF, refer to the sequencing facility FAQs: How are the data files delivered? *(https://ostr.ccr.cancer.gov/resources/sequencing-facility/faq-bioinformatics/#faq_9)*

SCAF ?

Data is returned from the SCAF via a Globus share link.

NCI CCR Genomics Core ?

According to the NCI CCR Genomics Core website *(https://genomics.ccr.cancer.gov/Policies/#content-1)*

> Next gen sequencing data will be delivered via pre-signed URLs in the form of a .tar or .zip archive containing all fastq files as well as a package containing QC metrics. A .tar archive of the entire run directory can also be delivered upon request. All preassigned-URLs are valid for one week from the delivery date.
>
> All delivered NGS data will be uploaded for long-term storage on your behalf to the NCI Data Vault. All project data (both raw and processed) will be stored for a period of two years from the run completion date. Please backup and store your project data within this timeframe. While it is possible that project data may retrieved after this time frame, we cannot guarantee that all raw files will be available.
> For information about the NCI data vault visit https://wiki.nci.nih.gov/display/DMEdoc *(https://wiki.nci.nih.gov/display/DMEdoc)*

Understanding QA/QC reports

QA / QC reports are generated from programs such as `fastqc` and `multiqc`.

FastQC *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)* runs several quality checks on raw NGS data to give you a general idea regarding the overall quality of your data. FastQC will generate a report for each sample.

On the other hand, Multiqc *(https://multiqc.info)* can be used to parse and aggregate summary information from a number of bioinformatic tools into a single report. In our example below, we have simply used Multiqc to combine summary information from fastqc from all samples into a single report, but you can also combine log files and output from other steps in your bioinformatic workflow, for example, following quality trimming with tools such as `trimmomatic` and `cutadapt`.

fastqc

Basic StatisticsPer base sequence qualityPer tile sequence qualityPer sequence quality s

Note that each section of the report is marked by color coded flags (i.e., green, yellow, red). Yellow and red flags, which indicate "warning" and "fail" respectively, may indicate a problem with the quality of your data. Such flags suggest that you should take a closer look at the data, but whether they represent an actual quality issue is contextually dependent and based on your experiment.

Let's break down some of the components of this report.

Basic Statistics *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/1%20Basic%20Statistics.html)*

Includes general summary information. You should note the "Total Sequences", "Sequence length", and "%GC". Are these what you expect?

Per Base Sequence Quality *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html)*
Includes a box and whisker plot summarizing quality scores information for all sequences in a sample at each base pair position. The blue line tracks the mean quality score.

There may be lower quality scores across the first few positions and you will likely see a general decline in quality with the length of the read. In general, greater than 28 indicates high quality reads.

Per Tile Sequence Quality *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html)*

This plot only appears if Illumina headers are retained. It allows you to assess quality across the flowcell. We want this plot to stay fairly blue across all tiles. The blue colors indicate "where the quality was at or above the average for that base in the run", whereas warmer colors indicate a decrease in quality for a tile compared to other tiles for that base. If there are warmer colors throughout, there may have been a problem with the Illumina flowcell.

Though we have a warning for our example fastqc report, overall the per tile sequence quality looks fine. See the linked fastqc documentation for an example of a bad plot.

Per Sequence Quality Scores *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/3%20Per%20Sequence%20Quality%20Scores.html)*

This plot shows the quantity of sequences associated with a given mean quality score. Ideally we want the majority of our reads to be of high quality, so we would expect a peak toward the right of the plot with no major peaks at lower quality scores.

The per sequence quality scores look fantastic for this sample.

Per Base Sequence Content *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html)*

> In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. --- fastqc documentation *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html)*

However, this quality check often fails for RNAseq data:

> This is because the first 10-12 bases result from the ‘random’ hexamer priming that occurs during RNA-seq library preparation. This priming is not as random as we might hope giving an enrichment in particular bases for these intial nucleotides. --- hbctraining *(https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/QC_raw_data.html)*

Per Sequence GC Content *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html)*

The per sequence GC content should demonstrate a normal distribution. The peak should match the underlying GC content from your genome of interest. Biases here could indicate a contaminated library.

Per Base N Content *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/6%20Per%20Base%20N%20Content.html)*

An 'N' base call results when the sequencer cannot confidently determine the base. There may be a low number of Ns throughout your sequences. This is only a concern if the proportion of Ns is significantly high. Though, you will likely see flags before this point if that is the case.

Sequence Length Distribution *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html)*

This shows the number of sequences by sequence length. Variation here will be contingent upon the sequencing platform from which your sequences derived.

Sequence Duplication Levels *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/8%20Duplicate%20Sequences.html)*

Sequence duplication levels are based on a subset of the first 100k sequences. This check is looking for exact sequences and so even high read coverage wouldn't necessarily result in exact sequences across a given region.

High duplication could result from:

- Low library diversity
- Vector or adaptor contamination
- Low level of duplication with small spike at 10 bin may occur for RNAseq projects
    - This is due to greatly oversequencing high copy genes to represent low copy genes.

The sequence duplication levels can be paired with the overrepresented sequences to determine the source of duplication.

Overrepresented Sequences *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/9%20Overrepresented%20Sequences.html)*

> This module lists all of the sequence which make up more than 0.1% of the total. --- fastqc documentation *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/9%20Overrepresented%20Sequences.html)*

Based on our example file, it is worth making sure that all adapters have been removed from our sequences.

Adapter Content *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/10%20Adapter%20Content.html)*

This quality check looks for uneven kmer coverage across the length of your sequences.

Note: at times the adapter content and overrepresented sequences do not agree. If one or both point to adapter contamination, you should consider adapter trimming.

`multiqc`

In this example, `multiqc` is simply aggregating results from `fastqc`. This allows us to compare the overall quality of our entire sequencing run.

Advanced Biomedical Computational Science Group

The Advanced Biomedical Computational Science (ABCS) *(https://frederick.cancer.gov/research/science-areas/bioinformatics-and-computational-science/advanced-biomedical-computational-science)* group focuses on applications of bioinformatics, computational and data science, and artificial intelligence to support NCI researchers.

ABCS provides:

- Subject matter expertise in genomics, proteomics, and imaging.
- Machine learning/Artificial intelligence for image analysis, text mining/NLP.
- Expertise in protein and nucleotide modeling, cheminformatics and quantum chemistry.
- Statistical analysis, large scale data integration.
- Programming expertise in several languages and database skills.

Support is available at no additional cost to all NCI researchers and can be requested by submitting a project request at https://abcs-amp.nih.gov/project/request/ABCS/ *(https://abcs-amp.nih.gov/project/request/ABCS/)*.

ABCS also provides training and outreach *(https://bioinfo-abcc.ncifcrf.gov/training/)* through 3 training series offered on Tuesdays at noon.

- Statistics for Lunch *(https://bioinfo-abcc.ncifcrf.gov/training/series/statistics-for-lunch)*: explores statistics topics at an intuitive and high-level.
- FRCE and Computational Science *(https://bioinfo-abcc.ncifcrf.gov/training/series/frce-and-computational-science)*: trainings on using large compute resources to solve various scientific challenges.
- Programmer's Corner *(https://bioinfo-abcc.ncifcrf.gov/training/series/programmers-corner)*: series for programmers to discuss topics of interest.

Additional links of interest include:

Publications: https://bioinfo-abcc.ncifcrf.gov/bioinfo/public/publications *(https://bioinfo-abcc.ncifcrf.gov/bioinfo/public/publications)*
Scientific applications: https://bioinfo-abcc.ncifcrf.gov/ *(https://bioinfo-abcc.ncifcrf.gov/)* Code repositories: https://abcsfrederick.github.io/ *(https://abcsfrederick.github.io/)*

The following outlines ABCS expertise by subject:

- Bioinformatics and next-generation sequencing (NGS)
    - Gene expression analysis.
    - SNPs/indels and large structural variant analysis.

  - Genome assembly and annotation.
  - Whole genome or targeted methylation analysis.
  - Single cell analysis.
  - Experimental design consultation.
  - Result interpretation and visualization.
  - Bioinformatics training.

- Biomedical image analysis and visualization
  - Tumor segmentation and quantification.
  - Custom analysis – cancer subtype identification, predict prognosis and survival.
  - Whole animal image and tissue slide analysis.
  - Spatial transcriptomics.
  - Support tissue analysis core (TAC) and small animal imaging program (SAIP).

- Computational chemistry
  - Elucidate biomolecular interactions.
  - Predict fluorescence spectra.

- Protein modeling
  - Drug target binding.
  - Protein-Protein interactions.
  - Molecular dynamics simulations.

- Statistics and mathematical analysis

- Biomedical data mining, annotations, and integration
  - Automatic downloads and maintenance of 100s of annotations including genes, proteins, drugs, literature, and variants.
  - Integrate annotations into existing or new applications.
  - Multi-modal large scale data integrations with omics, clinical and imaging data.

- Scientific and high-performance computing (HPC), scientific web development, and scientific infrastructure.
  - Scientific data sharing applications.
  - Develop custom websites for hosting software developed in the labs.
  - Integrate with services such as HPC, GridFTP, SQL, and NoSQL databases.
  - Applications for managing sequencing and imaging efforts, and integrating with analysis workflows.
  - Scientific catalogs, training resources.
  - Web-based services.

CCR Collaborative Bioinformatics Resource

The CCR Collaborative Bioinformatics Resource (CCBR) *(https://bioinformatics.ccr.cancer.gov/ccbr/)* is a resource group which provides a mechanism for CCR researchers to obtain many different types of bioinformatics assistance to further their research goals. The group has expertise in a broad range of bioinformatics topics, and as such, its goal is to provide a simplified central access point for CCR researchers.

The CCBR group includes members of the CCR Office of Science and Technology Resources (OSTR), Frederick National Laboratory for Cancer Research (FNLCR) and the Center for Biomedical Informatics and Information Technology (CBIIT). The CCBR may also direct projects to other available CCR bioinformaticians as needs demand.

Requests for any type of Bioinformatics support should be through the Project Request Form *(https://bioinformatics.ccr.cancer.gov/ccbr/ask-for-help/)*. On this form, the requestor should describe the type of assistance being sought. Generally speaking, it is best to first contact CCBR for help with experimental design to reduce any possible sources of technical variation that may be added during handling or sequencing. Proper consultation ensures timely turn-around times and may save you money in the long run (some data may not be salvageable). Therefore, it is highly recommended to contact CCBR first before any samples are sent out for sequencing. CCR scientists who are interested in receiving advice on the best technologies and strategies for upcoming experiments can similarly use the CCBR Project Request Form *(https://bioinformatics.ccr.cancer.gov/ccbr/ask-for-help/)* to request assistance.

Once the nature of the project has been clearly defined it will be assigned to a lead analyst to help with everything from data analysis to manuscript preparation.

Assistance obtained via the CCBR should be viewed as collaborative in nature, with appropriate co-authorship or acknowledgment, depending on the nature of work involved.

Established technologies

- Microarray analysis across a variety of platforms and custom arrays
- Next Generation Sequence (NGS) data analysis
- Data mining, statistical and mathematical analysis using multiple approaches
- Pathway mapping and biological interpretation
- Multi-experiment data integration and correlation
- miRNA and array CGH analysis
- SNP and base calling

For more information, please visit the CCBR website *(https://bioinformatics.ccr.cancer.gov/ccbr/)*. Additional contact information can be found here *(https://bioinformatics.ccr.cancer.gov/ccbr/contact-us/)*.

Getting Started with Biowulf

Biowulf *(https://hpc.nih.gov/systems/)* 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/ *(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 *(https://hpc.nih.gov/docs/accounts.html)*. 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.

1. If this is your first time logging into Biowulf, you will see a warning statement with a yes/no choice. Type "yes".
2. Type in your password at the prompt. NOTE: The cursor will not move as you type your password! Don't let this fool you. Type in your password in once and hit "return/enter" on your keyboard.
3. When you see the command prompt dollar sign "$", you will know you are logged in.

```
[username@biowulf ~] $
```

Logging into Biowulf from Windows 10 OS

Open the command prompt and start an "SSH" (secure shell) session:

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

where "username" is your NIH/Biowulf login username.

1. If this is your first time logging into Biowulf, you will see a warning statement with a yes/no choice. Type "yes".
2. Type in your password at the prompt. NOTE: The cursor will not move as you type your password! Don't let this fool you. Type in your password in once and hit "return/enter" on your keyboard.
3. When you see the command prompt dollar sign "$", you will know you are logged in.

Working on Biowulf - two things you should always do.

When you log into Biowulf, you are automatically in your home directory (/home). This directory is very small and not suitable for large data files or analysis.

Use the "cd" command to change to the /data directory.

```
$ cd /data/username
```

where "username" is your username.

When working on Biowulf, you cannot work on the "login node". Instead, you need to work on a node or nodes that are sufficient for what you are doing. For now, you will use the "sinteractive" command to start an interactive session.

```
$ sinteractive
```

Being a good citizen on Biowulf

To run jobs on Biowulf, you must designate them as interactive, batch, or swarm. Failure to do this may result in termination of your account.

Running Interactive Jobs

Interactive nodes are suitable for routine tasks and debugging. To start an interactive node, type "sinteractive" at the command line "$" and press Enter/Return on your keyboard.

```
$ sinteractive
```

You will see something like this printed to your screen. It may take a minute or so for the command to finish. You'll know it's done when you get your command line dollar sign "$" back. You only need to use the "sinteractive" command once per session. If you try to start an interactive node on top of another interactive node, you will get a message asking why you want to start another node.

```
[username@biowulf ]$ sinteractive
salloc.exe: Pending job allocation 34516111
salloc.exe: job 34516111 queued and waiting for resources
salloc.exe: job 34516111 has been allocated resources
salloc.exe: Granted job allocation 34516111
salloc.exe: Waiting for resource configuration
salloc.exe: Nodes cn3317 are ready for job
srun: error: x11: no local DISPLAY defined, skipping
[username@cn3317 ]$
```

Batch Jobs

Most jobs on Biowulf should be run as batch jobs using the "sbatch" command.

```
$ sbatch yourscript.sh
```

Where "yourscript.sh" contains the job commands including input, output, cpus-per-task, and others. Batch scripts always start with "#!/bin/bash".

For example:

```
#!/bin/bash

module load fastqc
fastqc -o output_dir -f fastq seqfile1 seqfile2 ... seqfileN
```

where **-o** names the output directory

-f states the format of the input file(s)

and **seqfile1 ... seqfileN** are the names of the sequence files.

For more information on running batch jobs on Biowulf, please see: https://hpc.nih.gov/docs/userguide.html. *(https://hpc.nih.gov/docs/userguide.html)*

For multi-threaded jobs, you will need to set "cpus-per-task" like this. You can do this at the command line or put it in your script.

At the command line:

```
$ sbatch --cpus-per-task=# yourscript.sh
```

Or in your script:

```
#!/bin/bash

module load fastqc
fastqc -o output_dir  $SLURM_CPUS_PER_TASK -f fastq seqfile1 seqfile2
```

Swarm-ing on Biowulf

Swarm is a script for running a group of commands on Biowulf. Swarm reads a list of command lines and automatically submits them to the system. To create a swarm file, you can use "nano" or another text editor and put all of your command lines in a file called "file.swarm". Then you will use the "swarm" command to execute.

```
$ swarm -f file.swarm
```

Swarm creates two output files for each command line, one each for STDOUT (file.o) and STDERR (file.e). You can look into these files with the "less" command to see any important messages.

```
$ less file.o
$ less file.e
```

For more information on swarm-ing on Biowulf, please see: https://hpc.nih.gov/apps/swarm.html *(https://hpc.nih.gov/apps/swarm.html)*

Transferring Large Files with Globus

What is Globus?

Globus is a file transfer service for transferring large files, although any size files can be used. Conveniently, it sends you email when your file has transferred. It also will automatically keep trying to send files if they initially fail. For more information on using Globus at NIH, please see: https://hpc.nih.gov/docs/globus/ *(https://hpc.nih.gov/docs/globus/)*

Logging into Globus with your NIH login

1. Go to https://www.globus.org *(https://www.globus.org)* and click on "Globus Account Log In" found in the upper right corner of the screen.
2. Type in "National Institues of Health" in the box.
3. Click the "Continue" button.
4. Log in using your NIH login and password.

Installing the Globus client on your desktop

The Globus client works on Mac, Windows, and Unix systems. Do not use VPN when installing the Globus client. You only need to go through the process of installing the Globus client once.

For MacOS, please see: https://docs.globus.org/how-to/globus-connect-personal-mac/ *(https://docs.globus.org/how-to/globus-connect-personal-mac/)*

For Windows, please see: https://docs.globus.org/how-to/globus-connect-personal-windows/ *(https://docs.globus.org/how-to/globus-connect-personal-windows/)*

Transferring data between your desktop and Biowulf

1. Start up Globus Connect Personal.
2. Go to https://www.globus.org *(https://www.globus.org)*
3. Click on "Log In" and sign in with your NIH login and password.
4. You will now be at the Globus File Manager page. https://hpc.nih.gov/docs/globus/transfer.php *(https://hpc.nih.gov/docs/globus/transfer.php)*.
5. In the "Collection" box, type "NIH HPC Data Transfer". The files in your /home on Biowulf will appear. You can move to /data/username by typing that in the "Path" box.
6. Click on "Sync or Transfer Files".
7. Enter the other endpoint, in this case the endpoint name that you gave to your desktop system when you installed Globus. You should now see both endpoints listed in two panes of the Globus window.
8. To transfer files, select a file or directory on one endpoint, and click the blue 'Start' button. The page will now say that the transfer request submitted successfully.

9. Click on 'View details' to display task detail information. Statistics are displayed at this page. You will also receive an email when the transfer is complete.

See https://hpc.nih.gov/storage/globus.html *(https://hpc.nih.gov/storage/globus.html)* for more details.

Bioinformatic interest groups, listservs, and Slack channels

To stay up to date on bioinformatic tools, methods, and training opportunities, there are a number of groups with listservs and Slack channels you may be interested in joining.

Interest groups with associated listservs

1. Bioinformatics Scientific Interest Group (Bioinformatics SIG) *(https://oir.nih.gov/sigs/bioinformatics-scientific-interest-group)*

   Fosters networking, collaboration, training, and career development for computational biologists and those interested in incorporating computational biology in their research.

   > Topic areas include the computational aspects of functional and comparative genomics, systems biology, bioimaging, proteomics, structural modeling, and molecular dynamics. --- Bioinformatics SIG *(https://oir.nih.gov/sigs/bioinformatics-scientific-interest-group)*

   To join the BIOINFORMATICS-SIG-L, subscribe here *(https://list.nih.gov/cgi-bin/wa.exe?A0=BIOINFORMATICS-SIG-L)*.

2. NIH-DATASCIENCE-L *(https://list.nih.gov/cgi-bin/wa.exe?A0=nih-datascience-l)*

   A forum for community collaboration and discourse on data science for biomedical data scientists. Opportunities include seminar series, poster sessions, and workshop / courses.

   To join the NIH-DATASCIENCE-L, subscribe here *(https://list.nih.gov/cgi-bin/wa.exe?A0=NIH-DATASCIENCE-L)*.

3. Single-Cell Genomics Interest Group *(https://oir.nih.gov/sigs/single-cell-genomics-interest-group)*

   Hosts monthly seminar series, monthly joint lab meetings, and hosts symposia and workshops related to advances in single cell genomics.

   To join the SINGLECELLGENOMICS-L, subscribe here *(https://list.nih.gov/cgi-bin/wa.exe?A0=SINGLECELLGENOMICS-L)*.

4. Spatial Biology Interest Group *(https://oir.nih.gov/sigs/spatial-biology-interest-group)*

   Organizes monthly seminar series to engage and promote interactions among researchers in the spatial biology community, ranging from spatial biology, computational method development, and technology advancement, to facilitate new scientific discoveries.

To join the SPATIALBIOLOGY listserv, subscribe here *(https://list.nih.gov/cgi-bin/wa.exe?A0=SPATIALBIOLOGY)*.

For a comprehensive list of Scientific Interest Groups (SIGs), click here *(https://oir.nih.gov/sigs)*.

Other Groups

1. Bring Your Own Bioinformatics (NIH BYOB) *(https://nih-byob.github.io/)*

   An informal community-led talk series focused on the practical aspects of bioinformatics. BYOB is for anyone with an interest in bioinformatics. The group fosters collaboration and community discussion.

   Join their Slack channel *(https://nih-byob.slack.com/join/signup#/domain-signup)*.

More Training Opportunities

Other opportunities for bioinformatics training outside of BTEP include:

NIH Library

The NIH Library *(https://www.nihlibrary.nih.gov)* provides resources and guidance on bioinformatics, biostatistics, data visualization, and data curation.

- Check the NIH Library Training and Events Calendar *(https://www.nihlibrary.nih.gov/training/calendar)* to find upcoming virtual and in-person classes related to data science *(https://www.nihlibrary.nih.gov/training/calendar?field_training_category_target_id%5B266%5D=266&default_calendar_view=dayGridMonth)* and bioinformatics *(https://www.nihlibrary.nih.gov/training/calendar?field_training_category_target_id%5B261%5D=261&default_calendar_view=dayGridMonth)*.
- Access data science and bioinformatics resources including software on the Bioinformatics and Data Science Workstations. The Bioinformatics and Data Science Workstations are loaded with specialized bioinformatics *(https://www.nihlibrary.nih.gov/services/bioinformatics-support/analysis-tools)* and data analysis, processing, and visualization tools *(https://www.nihlibrary.nih.gov/services/data/data-services-tools)* and can be reserved *(https://www.nihlibrary.nih.gov/services/workspaces/reserve)* to use virtually or in person at the NIH Library in building 10.
- Discover videos and eBooks related to data science, visualization, and bioinformatics accessible through the NIH Library catalog. *(https://onesearch.nihlibrary.ors.nih.gov/discovery/search?vid=01NIH_INST:NIH&lang=en)*
- Request consultations *(https://custserv.nihlibrary.ors.nih.gov/consultation/)* with NIH Library staff regarding data management, visualization, and analysis.

CBIIT

As the NCI Center for Biomedical Informatics and Information Technology, CBIIT *(https://datascience.cancer.gov/)* provides training classes, data sharing and storage solutions, seminars and blogs on data science and bioinformatics.

Upcoming data science and informatics presentations, conferences, workshops, and trainings from CBIIT can be found here *(https://datascience.cancer.gov/news-events/events)*. Also check out this resource *(https://datascience.cancer.gov/training)* for additional training material specific to data science and cancer.

CBIIT also houses the Informatics Technology for Cancer Research (ITCR *(https://itcr.cancer.gov/)*), a trans-NCI program that supports extramural informatics technology development. To learn about upcoming informatics coursing and available training through ITCR, visit the ITCR Training Network. *(https://www.itcrtraining.org/home)*

Other relevant bioinformatic training is available through the DSLE and CRDC.

Data Science learning exchange (DSLE) *(https://ncihub.org/groups/dslx/overview)*

Provides learning resources and tools for community collaboration.

Cancer Research Data Commons (CRDC)

> The NCI Cancer Research Data Commons (CRDC) *(https://datacommons.cancer.gov/)* is a cloud-based data science infrastructure that connects data sets with analytics tools to allow users to share, integrate, analyze, and visualize cancer research data to drive scientific discovery. --- CRDC *(https://datascience.cancer.gov/data-commons)*

The core components of the CRDC are its repositories, infrastructure, and cloud resources.

1. Data-type specific repositories:
   - Cancer Data Service
   - Clinical Trial Data Commons
   - Genomic Data Commons
   - Imaging Data Commons
   - Integrated Canine Data Commons
   - Proteomic Data Commons
2. Key infrastructure tools allow data access and integration:
   - Cancer Data Aggregator
   - Data Commons Framework
   - Data Standards Services
3. Available Cloud resources facilitate analysis:
   - Broad Institute FireCloud *(https://datacommons.cancer.gov/analytical-resource/broad-institute-firecloud)*
   - ISB Cancer Gateway in the Cloud *(https://datacommons.cancer.gov/analytical-resource/isb-cancer-gateway-cloud)*
   - Seven Bridges Cancer Genomics Cloud, powered by Velsera *(https://datacommons.cancer.gov/analytical-resource/seven-bridges-cancer-genomics-cloud-developed-velsera)*

For more information, refer to the CRDC website *(https://datacommons.cancer.gov)*.

CRDC **learning Resources** can be found here *(https://datacommons.cancer.gov/learn#webinars-presentations)*.

General Bioinformatics Resources

BTEP and related resources open to everyone at NIH (not just NCI CCR)

- NIH Bioinformatics Calendar *(https://btep.ccr.cancer.gov/)*, sponsored by the NCI CCR Bioinformatics Training and Education Program (BTEP), contains information on all bioinformatics (and some data science) trainings/ presentations/ classes offered on the NIH campus.

- BTEP Distinguished Speakers Seminar Series *(https://bioinformatics.ccr.cancer.gov/btep/seminar/Distinguished+Speakers)*

- BTEP "Topics in Bioinformatics" Seminar Series includes variant analysis, RNA-Seq, single cell, microbiome analysis, ChIP-Seq and more. See the NIH Bioinformatics Calendar *(https://bioinformatics.ccr.cancer.gov/btep/nih-bioinformatics-calendar/)* for upcoming events. Check out the BTEP Video archive *(https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/)* to view recordings of past sessions.

- NIH Library Bioinformatics Support Program *(https://www.nihlibrary.nih.gov/services/bioinformatics-support)* is open to anyone at NIH and includes 2 Workstations with Bioinformatics Data Analysis Software (Partek Flow and Partek Genomics Suite, Qiagen Ingenuity Pathway Analysis and CLC Genomics Workbench, and much more). See entire list of resources here *(https://www-nihlibrary-nih-gov.ezproxy.nihlibrary.nih.gov/services/bioinformatics-support)*.

- Coursera NIH Learning Program - licenses provided by the NLM Office of Data Science Initiatives are available to anyone at NIH; apply here *(https://nlmenterprise.co1.qualtrics.com/jfe/form/SV_3fm9XD28rqqj8u9)*.

- NIAID Collective Bioinformatics Resource (NCBR) *(https://bioinformatics.niaid.nih.gov/)*

- Bioinformatics workflows on Biowulf *(https://hpc.nih.gov/training/)*

    - CCBR Pipeliner *(https://github.com/CCBR)* – a suite of NGS analysis workflows (RNASeq, WESSeq, ATACSeq, ChIPSeq, CRISPRSeq, CUT&RunSeq, EV-Seq,circRNASeq, scRNASeq, WGSSeq) on NIH HPC Biowulf from the CCR Collaborative Bioinformatics Resource.

    - OpenOmics/genome-seek *(https://github.com/OpenOmics/genome-seek)*: a comprehensive clinical WGS and WES pipeline.

- NIH Cloud Resources
  - STRIDES *(https://datascience.nih.gov/strides)* initiative as part of the NIH Strategic Plan for Data Science.
  - Cancer Research Data Commons *(https://datascience.cancer.gov/data-commons)* is a cloud-based infrastructure for analysis of cancer research data that includes databases and large collection of analysis tools and workflows.
  - NIDAP (NIH Integrated Data Analysis Portal) *(https://bioinformatics.ccr.cancer.gov/ccbr/education-training/nidap-training/)* bulk and single cell RNA-Seq workflows; login with your NIH credentials here *(https://nidap.nih.gov/multipass/login/all)*.
  - NHGRI Analysis Visualization and Informatics Lab-space *(https://anvilproject.org/)*

Resources for the general public

BTEP Resources

- BTEP FAQs *(https://bioinformatics.ccr.cancer.gov/btep/forums/)* on Single Cell RNA-Seq and ChIP-Seq are open to the world.
- The BTEP resources pages *(https://bioinformatics.ccr.cancer.gov/docs/resources-for-bioinformatics/)*.
- BTEP Class Documentation *(https://bioinformatics.ccr.cancer.gov/btep/class-documents/)*.
- BTEP Video Archive *(https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/)*.

Publicly available resources

While by no means comprehensive, we are including several links to publicly available resources useful for learning bioinformatics or relevant skills:

- Data carpentries workshops *(https://datacarpentry.org/lessons/#genomics-workshop)* for lessons on data analysis that are both general and specific
- Software carpentries *(https://software-carpentry.org/lessons/)* for lessons introducing unix, version control, python, and R.
- Community resources and tutorials from Bioconductor are found in the Documentation box on the Bioconductor help page *(https://bioconductor.org/help/)*.
- Galaxy Training *(https://training.galaxyproject.org/training-material/)* materials
- Tutorials *(http://www.sthda.com/english/)* on statistical tools from STHDA
- Biostars Bioinformatics Explained includes a question and answer forum *(https://www.biostars.org/)* and tutorials *(https://www.biostars.org/t/tutorials/?order=rank)* page
- Bioinformatics Workbook *(https://bioinformaticsworkbook.org/about.html#gsc.tab=0)*
- RNA-seq Bioinformatics - Griffith lab *(https://rnabio.org/)*

- Genomic data and visualization - Griffith lab *(https://genviz.org/)*
- Precision medicine Bioinformatics - Griffith lab *(https://pmbio.org/)*
- Orchestra *(https://github.com/seandavi/Orchestra)* for data science education
- Harvard Chan Bioinformatics Core Trainings *(https://hbctraining.github.io/main/)*
- Glittr (Git repositories with bioinformatics training material) *(https://glittr.org/?per_page=25&sort_by=stargazers&sort_direction=desc)*
- EMBL-EBI Training *(https://www.ebi.ac.uk/training/)*
- Swiss Institute of Bioinformatics *(https://www.sib.swiss/)*
- NCBI Outreach Events *(https://ncbiinsights.ncbi.nlm.nih.gov/ncbi-outreach-events/#events)*

Self Learning Platforms

Biostars

Description

Biostars: Bioinformatics Explained (https://www.biostars.org/) is a question and answer forum where researchers can obtain answers to questions ranging from simple to advanced in the fields of bioinformatics, computational genomics, and biological data analysis.

The developers of Biostars have also created a multi-volume handbook *(https://www.biostarhandbook.com/)* with a question to answer format designed to teach practical skills in bioinformatics. You can follow along with the examples in the book by downloading associated data and installing suggested software.

Each volume is also available for download as a .pdf.

The current volumes are as follows:

- The Biostar Handbook - An introduction to Bioinformatics as a scientific field.
- The Art of Bioinformatics Scripting - Learn advanced Unix and Bash scripting skills.
- RNA-Seq by Example - Master RNA-Seq data analysis.
- Corona Virus Genome Analysis - Advanced topics devoted to the study of the Corona Virus.
- Biostar Workflows - Create automated bioinformatics workflows

Recommendations

This is a fantastic introduction to bioinformatics and can be useful as a reference even for the non-beginner.

Things to Know

- The handbooks are opinionated. The opinions expressed in the book may or may not align with those of other bioinformaticians.
- A license is required to access the *Biostar Handbook*, but anyone can submit a question on *Biostars: Bioinformatics Explained*.

Access Information

Licenses to the *Biostar Handbook* are available to CCR researchers. Please email BTEP at ncibtep@nih.gov if you would like a license.

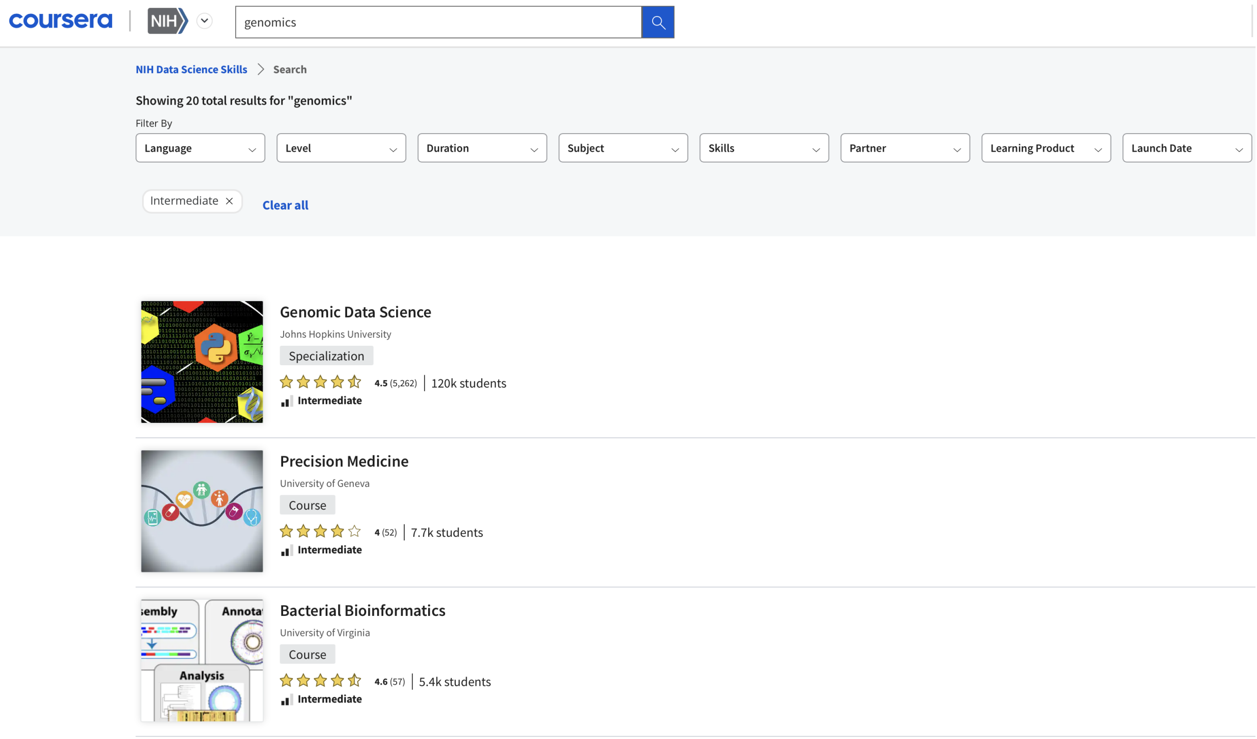
Coursera

Description

Coursera is an online learning platform that provides access to diverse courses from over 200 organizations, including universities and businesses. While they offer some free content, a license is required to explore most of the content. Courses are generally on demand and offer some type of certificate upon completion.

Through the "NCI Data Science, an NIH Learning Program", you can take courses in genomics, bioinformatics, programming, data science, statistics and more.

Here are a few examples of courses related to genomics found in the program.



NIH Data Science Skills > Search

Showing 20 total results for "genomics"

Filter By

Language | Level | Duration | Subject | Skills | Partner | Learning Product | Launch Date

Intermediate × Clear all

Genomic Data Science
Johns Hopkins University
Specialization
4.5 (5,262) | 120k students
Intermediate

Precision Medicine
University of Geneva
Course
4 (52) | 7.7k students
Intermediate

Bacterial Bioinformatics
University of Virginia
Course
4.6 (57) | 5.4k students
Intermediate

Things to Know

- Licenses are available to anyone at NIH
- Courses on demand and at your own speed
- Bioinformatics and genomics specific courses
- Certificates upon course completion
- Limited number of licenses

Access Information

If you're interested in a Coursera license, please see (https://bioinformatics.ccr.cancer.gov/btep/self-learning/ *(https://bioinformatics.ccr.cancer.gov/btep/self-learning/)*) for information on obtaining a license.

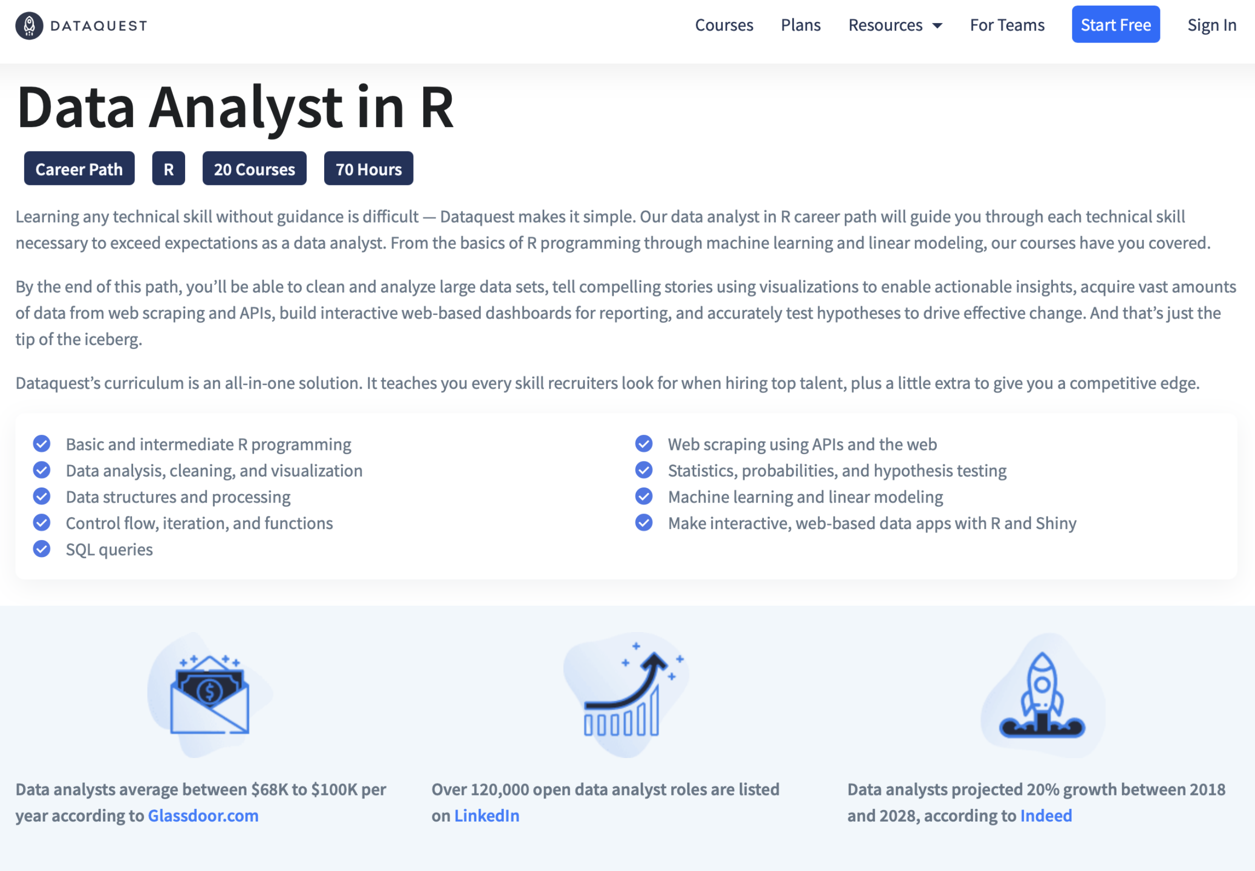
You will receive an invitation email (usually within 48 hours) from Coursera when your license is ready. Please check all your email folders.

If you have questions about your Coursera license, please contact us at ncibtep@nih.gov *(mailto:ncibtep@nih.gov)*.

Dataquest

Description

Dataquest is an online learning platform devoted to teaching data science and programming skills. Courses are organized in paths (career paths or skill paths) that guide you through a variety of lessons to obtain a specific goal whether career oriented or skill based. Paths guide you from the beginner level through intermediate and advanced stages while removing the complexity of figuring out what you should learn next. See an example of a career path below.



Data Analyst in R

Learning any technical skill without guidance is difficult — Dataquest makes it simple. Our data analyst in R career path will guide you through each technical skill necessary to exceed expectations as a data analyst. From the basics of R programming through machine learning and linear modeling, our courses have you covered.

By the end of this path, you'll be able to clean and analyze large data sets, tell compelling stories using visualizations to enable actionable insights, acquire vast amounts of data from web scraping and APIs, build interactive web-based dashboards for reporting, and accurately test hypotheses to drive effective change. And that's just the tip of the iceberg.

Dataquest's curriculum is an all-in-one solution. It teaches you every skill recruiters look for when hiring top talent, plus a little extra to give you a competitive edge.

- ✓ Basic and intermediate R programming
- ✓ Data analysis, cleaning, and visualization
- ✓ Data structures and processing
- ✓ Control flow, iteration, and functions
- ✓ SQL queries
- ✓ Web scraping using APIs and the web
- ✓ Statistics, probabilities, and hypothesis testing
- ✓ Machine learning and linear modeling
- ✓ Make interactive, web-based data apps with R and Shiny

Data analysts average between $68K to $100K per year according to Glassdoor.com

Over 120,000 open data analyst roles are listed on LinkedIn

Data analysts projected 20% growth between 2018 and 2028, according to Indeed

Dataquest provides courses on programming in Python, R, SQL, Unix/Bash as well as Machine Learning, Data Visualization and Probability/Statistics. Unfortunately, Dataquest does not offer bioinformatics specific courses. But the skills learned are applicable to bioinformatics. If interested in such courses, take a look at Coursera.

Things to Know

- Licenses are available to intramural NCI CCR personnel only
- Courses on demand and at your own speed
- No bioinformatics specific courses

- Limited number of licenses. Please contact ncibtep@nih.gov if you no longer plan to use a license.

Access Information

To apply for a Dataquest license, please see (https://bioinformatics.ccr.cancer.gov/btep/self-learning/ *(https://bioinformatics.ccr.cancer.gov/btep/self-learning/)*).

Once your application has been processed (usually within 48 hours), you will receive an email invitation from Dataquest. Please check your email folders for this message.

If you have questions about your Dataquest license, please contact us at ncibtep@nih.gov *(mailto:ncibtep@nih.gov)*.

Welcome to the BTEP bioinformatics tools selector guide. Scroll through this guide and click on the triangular tab next to each analysis category to identify software(s) that will help you accomplish your goals.

While many of the tools in these packages are powerful and can perform sophisticated analyses, people unfamiliar with bioinformatics analysis of complex data sets should consider consulting a bioinformatics analyst to ensure the validity of their methodology and conclusions.

Open Source Bioinformatics Tools

Note that this list of open source bioinformatics tools is not exhaustive because there are many open source tools that are available.

{{Sdet}}{{Ssum}}Biowulf applications{{Esum}}

Unix applications for bioinformatics include those used for

- Assessing next generation sequencing data quality
- Quality and adapter trimming of next next generation sequencing data
- Manipulation of alignment files
- Variant calling
- Differential gene expression analysis
- Peak calling for ChIP sequencing

Many of these applications are installed on Biowulf. See here *(https://hpc.nih.gov/apps/)* for the list of bioinformatics applications available through Biowulf.

See Biowulf High Performance Computing system for more information on getting started with Biowulf.

{{Edet}}

{{Sdet}}{{Ssum}}R{{Esum}}

Data visualization
Data wrangling

R has an extensive collection of packages for bioinformatics that facilitate analysis of RNA, ChIP, and ATAC sequencing data. Many of these packages are housed under the Bioconductor repository *(https://www.bioconductor.org)*. Outside of Bioconductor, there is Seurat *(https://satijalab.org/seurat/)*, a popular tool for single cell RNA sequencing.

{{Edet}}

{{Sdet}}{{Ssum}}Python{{Esum}}

Data visualization
Data wrangling

Python has packages such as Scanpy *(https://scanpy.readthedocs.io/en/stable/)*, scVelo *(https://scvelo.readthedocs.io)*,and scDeepCluster *(https://github.com/ttgump/scDeepCluster)* that facilitate single cell RNA sequencing analysis.

{{Edet}}

Commercial Bioinformatics Packages

Molecular Biology

{{Sdet}}{{Ssum}}Sequence comparison{{Esum}}

- CLC Genomics Workbench
  - What file types can I start my analysis with?
    - FASTA
    - Genbank
- Geneious Prime
  - What file types can I start my analysis with?
    - FASTA
    - Genbank
- Lasergene
  - What file types can I start my analysis with?
    - Genbank

{{Edet}}

{{Sdet}}{{Ssum}}Phylogenetics{{Esum}}

- CLC Genomics Workbench
  - What data types can I start my analysis with?
    - Sequence alignment result
- Geneious Prime
  - What data types can I start my analysis with?
    - Sequence alignment result

- Lasergene
    - What data types can I start my analysis with?
        - Sequence alignment result

{{Edet}}

{{Sdet}}{{Ssum}}Molecular cloning{{Esum}}

- CLC Genomics Workbench
    - What file types can I start my analysis with?
        - FASTA
        - Genbank
- Geneious Prime
    - What file types can I start my analysis with?
        - Genbank
        - .DNA
        - FASTA
- Lasergene
    - What file types can I start my analysis with?
        - .DNA
        - FASTA
        - Genbank
- SnapGene
    - What file types can I start my analysis with?
        - Genbank
        - .DNA

{{Edet}}

{{Sdet}}{{Ssum}}Restriction digest{{Esum}}

- CLC Genomics Workbench
    - What file types can I start my analysis with?
        - .DNA
        - FASTA
        - Genbank
- Geneious Prime
    - What file types can I start my analysis with?
        - .DNA
        - FASTA
        - Genbak
- Lasergene
    - What file types can I start my analysis with?
        - FASTA
        - Genbank
        - SEQ

- SnapGene
    - What file types can I start my analysis with?
        - .DNA
        - FASTA
        - Genbank

{{Edet}}

{{Sdet}}{{Ssum}}Ligation simulation{{Esum}}

- Geneious Prime
    - What file types can I start my analysis with?
        - FASTA

{{Edet}}

{{Sdet}}{{Ssum}}PCR primer design{{Esum}}

- CLC Genomics Workbench
    - What file types can I start my analysis with?
        - FASTA
        - Genbank
- Geneious Prime
    - What file types can I start my analysis with?
        - FASTA
        - Genbank
- Lasergene
    - What file types can I start my analysis with?
        - FASTA
        - Genbank
- SnapGene
    - What file types can I start my analysis with?
        - FASTA
        - Genbank

{{Edet}}

{{Sdet}}{{Ssum}}CRISPR editing{{Esum}}

- Geneious Prime
    - What file types can I start my analysis with?
        - FASTQ

{{Edet}}

Variant Analysis

{{Sdet}}{{Ssum}}Single nucleotide variants{{Esum}}

- CLC Genomics Workbench (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ
- Geneious Prime (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ
- Partek Flow (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ
    - BAM

{{Edet}}

{{Sdet}}{{Ssum}}Insertions, deletions{{Esum}}

- CLC Genomics Workbench (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ
- Geneious Prime (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ
- Partek Flow (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ
    - BAM
    - VCF

{{Edet}}

{{Sdet}}{{Ssum}}Structural variants{{Esum}}

- CLC Genomics Workbench (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ

{{Edet}}

{{Sdet}}{{Ssum}}Low frequency variants{{Esum}}

- CLC Genomics Workbench (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ

{{Edet}}

{{Sdet}}{{Ssum}}Copy number analysis{{Esum}}

- CLC Genomics Workbench (sequencing based)
    - The copy number detection tools in CLC Genomics Workbench are designed to analyze targeted sequencing data. Users can input FASTQ or BAM to perform copy number analysis using the CNV detection tool. If starting with FASTQ, users will need to align using the built-in aligner.
    - CLC Genomics Workbench uses a proprietary algorithm for CNV analysis. Refer to the white paper for details *(https://digitalinsights.qiagen.com/files/whitepapers/Biomedical_Genomics_Workbench_CNV_White_Paper.pdf)*.
    - What file types can I start my analysis with?
        - FASTQ
        - BAM
- Partek Genomics Suite (array based)
    - What file types can I start my analysis with?
        - Affymetrix CEL
        - Affymetrix Axiom Summary File
        - Agilent
        - Illumina GenomeStudio
        - Illumina Final Report text file
        - NimbleGen Pair or CGH Data summary files

{{Edet}}

{{Sdet}}{{Ssum}}Loss of heterozygosity{{Esum}}

- Partek Genomics Suite (array based)
    - What file types can I start my analysis with?
        - Affymetrix CHP
        - Affymetrix Gentoyping Text
        - Illumina GenomeStudio
        - Illumina Final Report Text

{{Edet}}

{{Sdet}}{{Ssum}}Association analysis{{Esum}}

- Partek Genomics Suite (array based)
    - What file types can I start my analysis with?
        - Affymetrix CHP
        - Affymetrix Genotyping Text
        - Illumina GenomeStudio
        - Illumina Final Report Text

{{Edet}}

{{Sdet}}{{Ssum}}Trio analysis{{Esum}}

- Partek Genomics Suite (array based)
    - What file types can I start my analysis with?
        - Affymetrix CHP
        - Affymetrix Genotyping Text
        - Illumina GenomeStudio
        - Illumina Final Report Text

{{Edet}}

{{Sdet}}{{Ssum}}Promoter tiling array{{Esum}}

- Partek Genomics Suite (array based)
    - What file types can I start my analysis with?
        - Affymetrix CEL
        - Affymetrix Data
        - NimblGen Pair Files
        - Text

{{Edet}}

Gene Expression

{{Sdet}}{{Ssum}}Gene expression by microarray{{Esum}}

- CLC Genomics Workbench
    - What file types can I start my analysis with?
        - Affymetrix Gene Chip (CHP, NetAFFx, CEL *(https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/751/index.php?manual=Affymetrix_GeneChip.html)*)
        - Illumina BeadChip
        - TSV
        - CSV
- Partek Flow
    - What file types can I start my analysis with?
        - Affymetrix CEL
        - Illumina BeadChip IDAT
        - TSV of Illumina GenomeStudio output
- Partek Genomics Suite
    - What file types can I start my analysis with?
        - Affymetrix CELL
        - Agilent TXT
        - Illumina GenomeStudio
        - Applied Biosystems TaqMan RQ Manager files
        - SOLiD SAGE output

        - NanoString CSV
        - Fluidigm

{{Edet}}

{{Sdet}}{{Ssum}}microRNA by microarray{{Esum}}

- Partek Genomics Suite
    - What file types can I start my analysis with?
        - Affymetrix CEL
        - Applied Biosystems TaqMan RQ manager files
        - Illumina GenomeStudio
        - SOLiD SAGE output files
        - GenePix (GPR) files
        - ImaGene (Exiqon) files
        - AB Small RNA Count files
        - Agilent data
        - NanoString output files
        - Fluidigm output files
        - TXT
- Qlucore Omics Explorer
    - What file types can I start my analysis with?
        - Affymetrix CEL
        - Affymetrix CHP
        - Agilent TEXT
        - TXT, TSV, CSV

{{Edet}}

{{Sdet}}{{Ssum}}RNA sequencing{{Esum}}

- CLC Genomics Workbench
    - What file types can I start my analysis with?
        - FASTQ
- Geneious Prime
    - What file types can I start my analysis with?
        - FASTQ
- Partek Flow
    - What file types can I start my analysis with?
        - FASTQ
        - BAM
        - Count table (TXT, CSV, TSV)
- Partek Genomics Suite
    - What file types can I start my analysis with?
        - SAM/BAM

- Qlucore Omics Explorer
    - What file types can I start my analysis with?
        - Count table (TXT, CSV, TSV)
        - BAM

{{Edet}}

{{Sdet}}{{Ssum}}microRNA sequencing{{Esum}}

- CLC Genomics Workbench
    - What file types can I start my analysis with?
        - FASTQ
- Partek Flow
    - What file types can I start my analysis with?
        - FASTQ
        - BAM
        - Count table (TXT, CSV, TSV)
- Partek Genomics Suite
    - What file types can I start my analysis with?
        - BAM
- Qlucore Omics Explorer
    - What file types can I start my analysis with?
        - Count table (TXT, CSV, TSV)
        - BAM

{{Edet}}

{{Sdet}}{{Ssum}}Single cell RNA sequencing{{Esum}}

- Partek Flow
    - What file types can I start my analysis with?
        - BCL
        - FASTQ
        - BAM
        - Count tables
        - barcodes.tsv, features.tsv, counts.mtx
        - h5
        - Seurat object
- Qlucore Omics Explorer
    - What file types can I start my analysis with?
        - 10x single cell data

{{Edet}}

{{Sdet}}{{Ssum}}Spatial transcriptomics{{Esum}}

- Partek Flow (implements Space Ranger to analyze 10x Visium Spatial Gene Expression data)
    - What file types can I start my analysis with?
        - FASTQ (sequences) and JPEG/TIFF for spatial images

{{Edet}}

Epigenetics

{{Sdet}}{{Ssum}}ATAC sequencing{{Esum}}

- Partek Flow
    - What file types can I start my analysis with?
        - FASTQ
        - BAM

{{Edet}}

{{Sdet}}{{Ssum}}Single cell ATAC sequencing{{Esum}}

- Partek Flow
    - What file types can I start my analysis with?
        - FASTQ (Partek Flow implements the Cell Ranger ATAC)

{{Edet}}

{{Sdet}}{{Ssum}}ChIP sequencing{{Esum}}

- CLC Genomics Workbench
    - What file types can I start my analysis with?
        - FASTQ
        - SAM/BAM
- Partek Flow
    - What file types can I start my analysis with?
        - FASTQ
        - BAM
- Partek Genomics Suite
    - What file types can I start my analysis with?
        - BAM

{{Edet}}

{{Sdet}}{{Ssum}}Methylation array{{Esum}}

- Partek Genomics Suite
  - What file types can I start my analysis with?
    - Illumina GenomeStudio output
    - Illumina Infinium methylation 450/850 IDAT files

{{Edet}}

{{Sdet}}{{Ssum}}Methylation tiling array{{Esum}}

- Partek Genomics Suite
  - What file types can I start my analysis with?
    - Illumina Infinium methylation 450/850 IDAT files
    - Affymetrix CEL files
    - Agilent
    - NimbleGen Pair files
    - TXT

{{Edet}}

{{Sdet}}{{Ssum}}Bisulfite sequencing{{Esum}}

- CLC Genomics Workbench
  - What file types can I start my analysis with?
    - FASTQ
    - SAM/BAM
- Partek Genomics Suite
  - What file types can I start my analysis with?
    - BAM

{{Edet}}

Metagenomics

{{Sdet}}{{Ssum}}Metagenomics{{Esum}}

- Geneious Prime (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ
- Partek Flow (sequencing based)
  - What file types can I start my analysis with?
    - FASTQ

{{Edet}}

Biological Insights

{{Sdet}}{{Ssum}}Pathway, network, and gene ontology{{Esum}}

- Partek Flow
  - What file types can I start my analysis with?
    - FASTQ
    - BAM
    - TXT
- Partek Genomics Suite
  - What file types can I start my analysis with?
    - BAM
    - TXT
- Qiagen Ingenuity Pathway Analysis
  - What file types can I start my analysis with?
    - TSV, CSV, or EXCEL file that contain gene, differential expression, and p-values
    - TSV, CSV, or EXCEL containing genetic gain or loss of function information
- Qlucore Omics Explorer
  - What file types can I start my analysis with?
    - BAM
    - CSV

{{Edet}}

Non-commercial

The R Project for Statistical Computing

Description

R is both a computational language and environment for statistical computing and graphics. It is open-source and widely used by scientists and other researchers, not just bioinformaticians. Base packages of R are built into the initial installation, but R functionality is greatly improved by installing other packages.

R is a great resource for statistical analysis, data visualization, and report generation. It is a particularly powerful programming language and environment due to its extensive community support. The widespread use of R means that tutorials, data analysis workflows / examples, and help are only a Google search away, and there are packages available for most types of analyses.

Recommendations

To take full advantage of R, you need to install R packages. R packages are loadable extensions that contain code, data, documentation, and tests in a standardized shareable format that can easily be installed by R users. The primary repository for R packages is the Comprehensive R Archive Network (CRAN) *(https://cran.r-project.org)*. CRAN is a global network of servers that store identical versions of R code, packages, documentation, etc. (cran.r-project.org). As of now, CRAN houses 18,825 available packages. Github is another common source used to store R packages; though, these packages do not necessarily meet CRAN standards so approach with caution.

There are also many field specific packages, including those useful in the -omics (genomics, transcriptomics, metabolomics, etc.). Check out Bioconductor *(https://bioconductor.org/)*, a repository for R packages related to biological data analysis, and Github for -omics packages and pipelines. Try out the biocViews *(https://bioconductor.org/packages/release/BiocViews.html)* search in Bioconductor.

{{Sdet}}{{Ssum}} Examples of top ranked Bioconductor packages by topic{{Esum}}

- RNA-Seq
    - limma *(https://bioconductor.org/packages/release/bioc/html/limma.html)*
    - edgeR *(https://bioconductor.org/packages/release/bioc/html/edgeR.html)*
    - DESeq2 *(https://bioconductor.org/packages/release/bioc/html/DESeq2.html)*
    - GenomicAlignments *(https://bioconductor.org/packages/release/bioc/html/GenomicAlignments.html)*
- ChIP-Seq
    - edgeR *(https://bioconductor.org/packages/release/bioc/html/edgeR.html)*
    - DESeq2 *(https://bioconductor.org/packages/release/bioc/html/DESeq2.html)*
    - Rsubread *(https://bioconductor.org/packages/release/bioc/html/Rsubread.html)*
    - regioneR *(https://bioconductor.org/packages/release/bioc/html/regioneR.html)*
    - ChIPseeker *(https://bioconductor.org/packages/release/bioc/html/ChIPseeker.html)*
- Variant Detection
    - Rsubread *(https://bioconductor.org/packages/release/bioc/html/Rsubread.html)*
    - infercnv *(https://bioconductor.org/packages/release/bioc/html/infercnv.html)*
    - PureCN *(https://bioconductor.org/packages/release/bioc/html/PureCN.html)*
    - CrispRVariants *(https://bioconductor.org/packages/release/bioc/html/CrispRVariants.html)*
- Mass Spec / Proteomics / Metabolomics
    - ProtGenerics *(https://bioconductor.org/packages/release/bioc/html/ProtGenerics.html)*

  - MSnbase *(https://bioconductor.org/packages/release/bioc/html/MSnbase.html)*
  - mzR *(https://bioconductor.org/packages/release/bioc/html/mzR.html)*
  - mzID *(https://bioconductor.org/packages/release/bioc/html/mzID.html)*

- Single cell

  - SingleCellExperiment *(https://bioconductor.org/packages/release/bioc/html/SingleCellExperiment.html)*
  - HDF5Array *(https://bioconductor.org/packages/release/bioc/html/HDF5Array.html)*
  - scuttle *(https://bioconductor.org/packages/release/bioc/html/scuttle.html)*
  - scater *(https://bioconductor.org/packages/release/bioc/html/scater.html)*
  - scran *(https://bioconductor.org/packages/release/bioc/html/scran.html)*
  - monocole *(https://bioconductor.org/packages/release/bioc/html/monocle.html)*
  - SingleR *(https://bioconductor.org/packages/release/bioc/html/SingleR.html)*
  - Seurat *(https://satijalab.org/seurat/)*
  - velocyto *(http://velocyto.org)*

{{Edet}}

Things to Know

- R is freely available and can be used via command line, through an integrated development environment (RStudio), and online (RStudio Server).
- Using R effectively can make scientific data analysis more reproducible. Data reports can be easily generated using R markdown.
- Because R is a programming language, the learning curve is fairly steep. However, if you take the time to learn the basics, a plethora of different data analysis and visualization packages will become accesible to you.

Input Data Types

The input data types are unlimited due to an extensive library of multidisciplinary packages. Tab delimited files (e.g., .txt, .tsv), comma separated files (.csv), Excel spreadsheets (.xls, .xlsx), and other delimited files, are easily imported using base R import functions.

Output Data Types

Again, thanks to a wide array of packages, output data types are essentially limitless. There are some file types that are specific to R and noteworthy including .RData and .rds files. RData files are used to capture all objects stored in a R workspace or global R environment, while .rds files hold a single R object.

Access Information

R and RStudio are free resources that can be downloaded directly from the internet. Click here *(https://bioinformatics.ccr.cancer.gov/docs/rtools/)* for installation instructions. To install R an RStudio on NIH laptops, please submit a ticket at service.cancer.gov *(https://service.cancer.gov)*.

Getting Help

Tutorials and courses are easily accessible.

- Check out BTEP R course offerings {{Sdet}}{{Ssum}} BTEP R Course Documentation{{Esum}}

  - R Introductory Series *(https://bioinformatics.ccr.cancer.gov/docs/rintro/)*
  - Data Visualization with R *(https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r/)*
  - Data Wrangling with R *(https://bioinformatics.ccr.cancer.gov/docs/data-wrangle-with-r/)*

  {{Edet}}

- Check out the NIH library *(https://www.nihlibrary.nih.gov/resources/tools/r-and-rstudio)*.
- Check out self-learning platforms: Coursera and Dataquest *(https://bioinformatics.ccr.cancer.gov/btep/self-learning/)*.

Python

Description

Python *(https://www.python.org/community/)* is a programming language used in many different applications including data science. It is a high-level computer language, as the syntax is easily read and understood. Python is considered a beginner-friendly language. Python also includes packages for machine learning. See Datacamp *(https://www.datacamp.com/blog/all-about-python-the-most-versatile-programming-language)* for more information about Python.

{{Sdet}}{{Ssum}}Listing of Analysis Functions{{Esum}}

There is extensive community support for Python because it is open source and there are many external packages that add to Python functionality.

Data wrangling

- Built-in functions for importing and working with tabular data with file extensions CSV or TXT.
- Pandas *(https://pandas.pydata.org)* is an external package that allows users to import and work with tabular data. Among the file extensions supported by this package are comma separated (CSV), TXT, and XLS/XLSX. Pandas makes it easier to work with tabular data as compared to the built-in Python functions.

Computing

- NumPy *(https://numpy.org/doc/stable/)* is a Python package for scientific computing. NumPy allows users to perform tasks such as basic arithmetic operations, array and matrix operations, and linear algebra.

- Math *(https://docs.python.org/3/library/math.html#constants)* is capable of basic math operations including those involving complex numbers. Math also contains several relevant mathematical constants such as pi.
- SciPy *(https://docs.scipy.org/doc/scipy/index.html)* is another package for computing in Python. Its functions include differentiation, integration, interploation, optimization, and image processing. Importantly, Scipy contains an extensive list of scientific constants.

Data visualization

- Matplotlib *(https://matplotlib.org)* is a capable and popular data visualization tool for Python.
- Seaborn *(https://seaborn.pydata.org)* is an extension of Matplotlib with supposedly simpler syntax. See here *(https://www.geeksforgeeks.org/difference-between-matplotlib-vs-seaborn/)* for some differences between Seaborn and Matlab.
- Plotly *(https://plotly.com/python/)* makes interactive plots.

Machine learning

- scikit-learn *(https://scikit-learn.org/stable/index.html)*
- PyTorch *(https://pytorch.org)*
- TensorFlow *(https://www.tensorflow.org/learn)*
- Keras *(https://keras.io)*

Molecular biosciences

- ACTINN *(https://github.com/SindiLab/ACTINN-PyTorch)* can be used for automated identification of cell types in single cell RNA sequencing studies. This packages utilizes PyTorch.
- scDeepCluster *(https://github.com/ttgump/scDeepCluster)* is a tool for single cell clustering that utilizes deep learning approaches using TensorFlow and Keras.
- Scanpy *(https://scanpy.readthedocs.io/en/stable/)* is a package for single cell RNA sequencing analysis.
- scvelo *(https://scvelo.readthedocs.io/getting_started/)* can be used for single cell velocity analysis.
- Biopython *(https://biopython.org)* is a package that contains functionalities for molecular biology analysis. It contains modules for sequence alignment, exploring protein 3D structure, population genetics, interfacing with databases housed at NCBI and many more.
- PyPop *(http://pypop.org/)* is a package for population genetics.
- simuPOP *(http://simupop.sourceforge.net/)* is used for forward-time population genetics analysis.

{{Edet}}

Recommendations

Things to know

Python can be accessed via either the command line or an Integrated Development Environments (IDE) that provides a graphical user interface. Available IDEs for Python include Spyder, PyCharm, R Studio, and Microsoft's Visual Studio Code (which is also available on Biowulf).

Using a Jupyter Notebook *(https://jupyter.org)* is another way to interface with Python. Jupyter Notebook can be viewed as a lab notebook for data analysis, and can include text based descriptions of analyses procedures along with code. Using a Jupyter Notebook allows us to see outputs and visualizations similar to IDEs and is easily accessible via a web browser.

Input Data Types

There are many data types that can be used as input for Python programs, including CSV, TXT and XLS/XLSX.

Output Data Types

Python can produce tabular data and data visualizations. Tabular data can be exported into various formats such as CSV, TXT, and XLSX, and visualizations can be exported as PNG, JPG, or TIF.

Access Information

Python 2 is pre-installed on MacOS computers. This will need to be updated to the current version Python 3.

Python is also accessible on the NIH high performance Unix cluster Biowulf.

For Python installations on NIH laptops, please submit a ticket to service.cancer.gov *(https://service.cancer.gov/)*.

Getting Help

Online learning platforms Coursera and Dataquest both have Python classes. To request a license see https://bioinformatics.ccr.cancer.gov/btep/self-learning/ *(https://bioinformatics.ccr.cancer.gov/btep/self-learning/)*. Below are some recommended courses from Coursera and Dataquest for those who wish to begin learning Python.

Coursera suggestions:

- Crash Course on Python *(https://www.coursera.org/learn/python-crash-course/home/welcome)*
- Programming for Everybody (Getting Started with Python) *(https://www.coursera.org/learn/python/home/welcome)*
- Data Analysis with Python *(https://www.coursera.org/learn/data-analysis-with-python/home/welcome)*
- Data Visualization with Python *(https://www.coursera.org/learn/python-for-data-visualization/home/welcome)*
- Python for Genomic Data Science *(https://www.coursera.org/learn/python-genomics/home/welcome)*

Dataquest suggestions:

- Variables, Data Types, and Lists in Python *(https://www.dataquest.io/course/variables-data-types-and-lists-in-python/)*
- Data Scientist in Python *(https://www.dataquest.io/path/data-scientist/)*

Commercial

Commercial Software

Below is a list of licensed software available for use by CCR Researchers. Linked software pages include a brief description of each software, recommendations, things to know, input data types, output data types, access information, and ways to get help.

1. Biodiscovery Nexus Copy Number
2. DNASTAR Lasergene
3. DNAnexus
4. Geneious Prime
5. NIDAP
6. Partek Flow
7. Partek Genomics Suite
8. Qiagen CLC Genomics Workbench
9. Qiagen Ingenuity Pathway Analysis
10. Qlucore Omics Explorer
11. SnapGene

Tip

Uncertain which software to use? Check out the BTEP bioinformatics tools selector guide, which categorizes software tools based on data type.

Biodiscovery Nexus Software

Description

Nexus Copy Number (*BioDiscovery*) is a graphical user interface (GUI) based bioinformatics software that specializes in copy number detection using array or sequence derived data. Its analysis capabilities are listed below.

{{Sdet}}{{Ssum}}List of Analysis Functions{{Esum}}

Copy number analysis

- Using array or sequence based data, the user can:
    - detect copy number variations at the individual and population level
    - identify subgroups in populations that show significant gains in copy number
    - detect gain or loss of heterozygosity
    - cluster samples that have similar patterns of variation
    - view results in a genomic context using genome browser
    - find molecular pathways that are affected by CNV
    - query databases such as The Cancer Genome Atlas (TCGA) to gain insight from existing data or integrate existing data with the user's own data

{{Edet}}

Recommendations

While Nexus Copy Number is a GUI based copy number analysis tool and does not require knowledge with scripting, it is a good idea to consult with bioinformatics experts at the Center for Cancer Research Collaborative Bioinformatics Resource (CCBR) who have in-depth experience in CNV analysis and know the limitations of various tools.

Things to Know

Nexus Copy Number runs on the user's machine so it may be limited by local resources. By default, Nexus Copy Number comes with human (NCBI build 36.1, 37) and mouse (NCBI build

38) reference genomes. Additional reference genomes for other organisms of interest are available here *(https://www.biodiscovery.com/support/downloads)*. Human build 38 can be downloaded from here *(https://gcc02.safelinks.protection.outlook.com/?url=http%3A%2F%2Fbiodiscovery-organisms.s3.amazonaws.com%2FHuman%2520NCBI%2520Build%252038.zip&data=05%7C01%7Calex.e*
Newer genome builds may be created upon request by contacting BioDiscovery Technical Support *(https://www.biodiscovery.com/support/product-support/?hsLang=en)*.

Input Data Types

- Nexus Copy Number can take data from various arrays as input including Affymetrix CEI. To see the full list, click on "Load" -> "Load Data" -> "Select data type".
- For high throughput sequencing, users can input
  - BAM
  - VCF

Output Data Types

- Analysis reports from Nexus Copy Number can be exported as TXT.
- Visualizations can be exported as JPEG, PNG, TIFF, and SVG at various resolutions.

Access Information

This is an NCI-licensed software with a 2 concurrent use license. To access Nexus Copy Number, submit a request at service.cancer.gov *(https://service.cancer.gov/)*.

Getting Help

For help with this package, see

- Help documentation accessible from the software client's Help menu.
- Webinars *(https://www.biodiscovery.com/webinars/topic/nexus-copy-number)*
- Tutorials *(https://www.biodiscovery.com/tutorials/tag/copy-number)*
- Educational Videos *(https://www.biodiscovery.com/videos/topic/nexus-copy-number)*

Additional resources can be found under the **Resources** tab on the BioDiscovery *(https://www.biodiscovery.com/)* website.

Lasergene

Description

DNASTAR Lasergene is a software suite that contains programs and tools dedicated to four overarching data analysis workflows:

- Lasergene Molecular Biology *(https://youtu.be/OHCmjlpESYM)*
- Lasergene Protein Analysis and Modeling *(https://youtu.be/VpoWqqWStHk)*
- Lasergene Genomics (within the Lasergene Genomics package)
- Lasergene Transcriptomics (within the Lasergene Genomics package)

The licenses available to NCI researchers only provide access to the Molecular Biology (Lasergene Molecular Biology) and Protein Analysis and Modeling packages (Lasergene Protein). Therefore, the tools available largely exclude high- throughput data.

Lasergene Molecular Biology

The Lasergene Molecular Biology package includes the following applications: SeqBuilder Pro, SeqMan Ultra, MegAlign Pro, GeneQuest, GenVision, SeqNinja, and DNASTAR Navigator.

These applications provide tools to perform in silico gel electrophoresis and cloning, plasmid vector annotation, cloning verification, sequence editing and annotation, multiple sequence alignment and pairwise alignments, PCR primer design, Sanger sequence assembly, and DNA translation.

Lasergene Protein

The Lasergene Protein package includes Protean 3D (+1 prediction per Nova Application) and DNASTAR Navigator.

The protein analysis and modeling workflow provides tools pertinent to antibody modeling, protein docking interaction prediction, epitome prediction, molecular motion visualization, protein design and engineering, protein sequence analysis, structural alignment, structural analysis, and structure prediction.

Recommendations

Included applications are fairly easy to navigate even for individuals with little to no bioinformatic experience. Tools facilitate the planning of molecular biology experiments and produce publishable visualizations.

Lasergenes' molecular biology application is recommended for Sanger sequencing analysis or for designing experimental assays rather then whole genome alignment.

Things to Know

- Must gain access to the NIH network to use the license
- GenVision is available for Windows only.
- Not recommended for NGS analysis

Input and Output Data Types

Because DNASTAR Lasergene includes a wide array of applications, there is a large number of input and output data types that vary by application. Fortunately, there is an extensive list *(https://www.dnastar.com/resources/file-formats/)* provided on the commercial website.

Access Information

You must submit a request through “service.cancer.gov *(https://service.cancer.gov/Lasergene)*” to obtain access to LaserGene Software. This software requires access to a floating license server. Please close the application when not in use so that others may gain access.

Getting Help

Each DNASTAR Lasergene application is well documented with help pages and tutorials. See the website tutorial page *(https://www.dnastar.com/training/)* for more information.

DNAnexus

Description

DNAnexus provides a secure cloud based platform for the analysis and sharing of next generation sequencing data. This resource allows simplified, secure access to the vast compute resources available via Amazon Web Services (AWS) and the Microsoft AZURE cloud. CCR is currently running a pilot program which allows CCR investigators access to this platform. This resource includes ≈200 prebuilt progams and workflows as well as many others built by the GAU team. DNAnexus can be utilized via either a user-friendly Web interface or through a command line interface. Less computer savvy people can use the Web interface while experienced computer users and bioinformaticists can interact with the platform through the command-line interface. The command-line interface allows for quick integration with local systems and the automated processing of large data sets. Beyond its extensive library of integrated tools and pipelines, the DNAnexus platform also allows the quick and efficient development of new analysis tools and the porting of existing pipelines.

The Genome Analysis Unit (GAU) has successfully used DNAnexus to empower several of its collaborators to analyse their own data using complicated (often custom built) workflows. If this is something you are intested in pursuing please contact Peter FitzGerald at fitzgepe@nih.gov *(mailto:fitzgepe@nih.gov)*.

DNAnexus has been used by BTEP for teaching bioinformatic tools and programming languages. BTEP uses pre-built teaching environments, which include all of the software needed for a lesson installed and ready to go. This allows lessons to progress seamlessly without worrying about errors due to software version incompatibility on local computers.

Recommendations

Things to Know

- Several BTEP courses require participants to obtain a free DNAnexus user account. Obtain a free acount here *(https://platform.dnanexus.com/register)*.
- CCR has established a pilot program *(https://bioinformatics.ccr.cancer.gov/gau/dnanexus-pilot-program/)* to enable CCR researchers to use DNAnexus to process their own NGS data.
- Some available workflows include RNA-Seq, DNA-Seq, and ChIP-Seq.

Input Data

Input data types will vary based on analysis objectives.

Output Data

Output data types will vary based on analysis objectives.

Access Information

While signing up for a DNAnexus account is free, there are usage costs associated with using the DNAnexus Cloud computing platform. However, OSTR has funding for the NCI DNAnexus Pilot program. To participate in this program, review the *DNAnexus Account Instructions (https://gau.ccr.cancer.gov/dnanexus-account-instructions/)*. Email us at ncibtep@nih.gov for more information.

Geneious Prime

Description

Geneious Prime is a graphical user interface (GUI) based bioinformatics package that contains a suite of tools for molecular biology and Next Generation Sequencing analysis.

{{Sdet}}{{Ssum}}Listing of Analysis Functions{{Esum}}

Classic Computational Molecular Biology Tools

- View/edit nucleotide or amino acid sequences
- Obtain complement and reverse complement of nucleotide sequences
- Convert between DNA and RNA
- Translate nucleotide to amino acide sequences
- View, edit, extract, and compare sequence annotations
- Examine sequence chromatograms
- Run simulated nucleic acid gels
- Viewer for
  - Interrogating RNA or DNA secondary fold structure
  - Examining protein 3D structure
- Perform sequence alignment using the following algorithms
  - BLAST
  - Geneious algorithm
  - MUSCLE
  - ClustalW
  - MAFFT
- Construct phylogenetic tree
- PCR primer design
  - Identify primer characteristics
  - Predict PCR products
- Molecular cloning
- Identify CRISPR sites and analyze editing results
- Perform microsatellite analysis
- Microsatellite analysis

NGS tasks

- Reference based read mapping
- *De novo* assembly
- Trim reads using BBDuk

DNA sequencing

- Variant calling

Gene expression

- Calculate expression counts
- RNA sequencing - the algorithms below are available to find differentially expressed genes.
    - Geneious Prime built-in method
    - DESeq2

Microbial

- Metagenomics

Visualizations

- Genome browser

Information mining

- Databases from the National Center for Biotechnology Information (NCBI) including:
- Gene
- Genome
- Nucleotide
- Popset
- Protein
- Structure
- Taxonomy
- PubMed
- UniProt

{{Edet}}

Recommendations

- Geneious Prime can not be used for epigenetic analysis. Please see Partek Flow or CLC Genomics Workbench for this functionality.

- This package does not support pathway or gene ontology analyses. Please see Partek Flow, Partek Genomics Suite, Qiagen Ingenuity Pathway Analysis (IPA), or Qlucore Omics Explorer to gain insight on the network and pathway level.
- Single cell RNA sequencing analysis cannot be performed using Geneious Prime. Use Partek Flow or Qlucore Omics Explorer.

Things to Know

Geneious Prime runs on a user's local machine and may be limited by the available compute resources on that machine.

Input Data Types

- CSV
- TSV
- TXT
- FASTA
- FASTQ
- SAM
- BAM
- VCF
- BED
- GFF
- GTF

See the Geneious Prime User Manual *(https://manual.geneious.com/en/latest/3-ImportExport.html#data-input-formats)* for a full list of supported input file types.

Output Data Types

- TXT
- CSV
- TSV
- PDF
- VCF
- FASTA
- FASTQ
- GFF
- BED
- PDF
- SVG
- EMF
- PNG

- JPG

See the Geneious Prime User Manual *(https://manual.geneious.com/en/latest/3-ImportExport.html#exporting-files)* for a full list of supported export formats including formats for exported images.

Access Information

You must submit a request through service.cancer.gov *(https://service.cancer.gov/)* to obtain access to this package. This software requires access to a floating license server, and so care should be taken to return licenses when the software is not actively being used (i.e. close the application). OSTR holds 10 concurrent licenses of Geneious Prime. You need to either be on the NIH network or VPN to use this package.

Getting Help

- A help menu is built into the Geneious Prime user interface.
- The Geneious Prime User Manual *(https://manual.geneious.com/en/latest/index.html)* is also available to users.
- There are also tutorials *(https://www.geneious.com/tutorials/)* that guide users through various Geneious Prime analysis workflows.

NIH Integrated Data Analysis Platform (NIDAP)

Description

> NIDAP (NIH Integrated Data Analysis Platform) is an innovative, cloud-based, collaborative data aggregation and analysis platform that hosts user-friendly bioinformatics workflows and component analysis and visualization tools developed by the NCI developer community based on open source tools and makes them immediately available to biologist end-users across the Institute. --- (https://bioinformatics.ccr.cancer.gov/btep/training/nidap-training/nidap-workflows/ *(https://bioinformatics.ccr.cancer.gov/btep/training/nidap-training/nidap-workflows/)*)

The NIDAP resource is freely available to NCI researchers for bulk and single cell RNA-Seq analyses. The platform is a graphic user interface (GUI) that does not require users to read or write code.

For more information, including how to access the platform and training materials, please see: https://bioinformatics.ccr.cancer.gov/btep/training/nidap-training/nidap-workflows/ *(https://bioinformatics.ccr.cancer.gov/btep/training/nidap-training/nidap-workflows/)*.

Questions regarding NIDAP training can be directed to Josh Meyer (thomas.meyer@nih.gov *(mailto:thomas.meyer@nih.gov)*).

Partek Flow

Description

Partek Flow (*Partek*) is a graphical user interface (GUI) based bioinformatics software that is dedicated to the analysis of next generation sequencing (NGS) data. It can perform the analyses listed below.

{{Sdet}}{{Ssum}}List of Analysis Functions{{Esum}}

DNA sequencing

- Variant detection
- Germline
- Somatic
- Copy number
- Low frequency
- Causal

Gene expression

- RNA sequencing
- MicroRNA sequencing
- Single cell RNA sequencing
- CITE sequencing
- Microarray

Epigenetics

- ChIP sequencing
- ATAC sequencing

Microbial

- Metagenomics

Biological insights

- Pathway analysis
- Gene set enrichment analysis

Visualizations

- Genome browser

{{Edet}}

Recommendations

Partek Flow allows users to start a multiple stages of the analyses. For gene expression analysis by RNA sequencing, users can start an analysis using either raw sequencing reads (FASTQ), aligned reads (BAM), or count table. Partek Flow is developed by the company that makes Partek Genomics Suite, thus users can go offline and run some analyses in Partek Genomics Suite. The flexibility of being able to start at multiple stages of an analysis workflow and the ability to conduct analyses offline through Partek Genomics Suite are rationale for using this package.

Things to Know

Partek Flow runs on Biowulf (NIH high performance compute cluster), thus when using this package, users will not be limited by local compute resources. Users interface with this package via a web browser, so users will not have to install additional software.

Input Data Types

Partek Flow supports a range of data formats, thus allowing users the ability to enter an analysis pipeline at any stage. Below are the supported data formats.

- FASTQ
- FASTA
- BAM
- SAM
- VCF
- BCF
- TXT
- BGX
- MTX
- H5
- SRA
- SFF

- GZ
- TAR
- ZIP
- BPM
- CEL
- QUAL
- IDAT
- PROBE_TAB

Output Data Types

Graphical output can be exported as PNG or SVG images at a specified resolution. Tabular data as well as files generated during NGS analysis (i.e., BAM files) can be found in the corresponding project folder.

Access Information

To access Partek Flow, users need to first set up a Biowulf account by clicking here *(https://hpc.nih.gov/docs/accounts.html)*. Next, users will need to ensure they have enough space on Biowulf to store data. (Fill out the storage request form) *(https://hpc.nih.gov/dashboard/storage_request.php)* if you need additional space. Finally, send an email to staff@hpc.nih.gov *(mailto:staff@hpc.nih.gov)* to get your Partek Flow account activated. When these steps are complete, go to https://partekflow.cit.nih.gov/flow *(https://partekflow.cit.nih.gov/flow)* to start using Partek Flow. Partek Flow users need to either be on the NIH network or connected via VPN. OSTR holds 5 Partek Flow and 3 Partek Flow single cell licenses.

Getting Help

Partek Flow comes with extensive documentation *(https://documentation.partek.com/display/FLOWDOC/Partek+Flow+Documentation)*. BTEP also hosts Partek Flow training frequently.

Partek®
Genomics Suite®

Partek Genomics Suite

Description

Partek Genomics Suite (*Partek*) is a graphical user interface (GUI) based bioinformatics package. It hosts a range of work flows that allow for gene expression, epigenetic, and association analysis.

{{Sdet}}{{Ssum}}Listing of Analysis Functions{{Esum}}

Genome wide association and inheritance

- Analysis using data derived from arrays
- Association
- Trio
- Analysis using data derived from sequencing
- Trio

Genetic variants

- Analysis using data derived from arrays
- Copy number
- Allele specific copy number
- Loss of heterozygosity
- Analysis using data derived from sequencing
- Find single nucleotide variants
- Annotate single nucleotide variants
- Predict single nucleotide variants effects

Gene expression

- Microarray
- Human exon array
- RNA sequencing

Gene regulation

- Promoter tiling array

Epigenetics

- Methylation array
- ChIP sequencing
- Bisulfite sequencing

Biological insigts

- Pathway analysis and gene ontology

Phenotypic outcome

- Survival analysis

Visualizations

- Genome browser

Information mining

- Import NCBI GEO data

{{Edet}}

Recommendations

Note that Partek Genomics Suite does not handle NGS read mapping but will accept mapped data in the form of BAM files. This package does not offer functionalities for basic molecular biology analysis such as sequencing alignment, phylogenetics, or primer design. For these, the user should look to CLC Genomics Workbench, Geneious Prime, Snapgene, or Lasergene. If the user needs to conduct trio analysis, Partek Genomics Suite is an option to turn to (another package that handels trio analysis is CLC Genomics Workbench).

Things to Know

Partek Genomics Suite is made by the company that develops Partek Flow and serves a good option for users to conduct analysis outside of Biowulf. A reason for using Partek Genomics Suite is for trio analysis, which is a feature that Partek Flow does not have.

Input Data Types

- CSV

- TXT
- CEL
- CHP
- CNCHP
- CNVCHP
- MIP
- ARR
- GPR
- IDAT
- SAM
- BAM
- VCF
- BCF
- Partek Genomics Suite also allows for import of data obtained from the following platforms
- Fluidigm
- Nanostring
- NimbleGen

Output Data Types

- TXT
- TSV
- CSV
- HTML
- PNG
- SVG
- PDF
- PPM
- JPEG
- GIF
- PS

Access Information

You must submit a request through service.cancer.gov *(https://service.cancer.gov/)* to obtain access to Partek Genomics Suite. This software requires access to a floating license server, and so care should be taken to return licenses when the software is not actively being used (i.e. close the application). Connection to NIH network or VPN is necessary to use Partek Genomics Suite. OSTR has 10 licenses to this package.

Getting Help

To access the Partek Genomics Suite documentation, click here *(https://documentation.partek.com/display/PGS/Partek+Genomics+Suite+Documentation)*.

Qiagen CLC Genomics Workbench

Description

CLC Genomics Workbench (*Qiagen*) is a graphical user interface (GUI) based bioinformatics software. It houses tools for molecular biology and Next Generation Sequencing (NGS) analysis (see Listing of Analysis Functions below).

{{Sdet}}{{Ssum}}Listing of Analysis Functions{{Esum}}

Classic Computational Molecular Biology Tools

- Nucleotide analysis
    - Convert between DNA and RNA
    - Find reverse complement
    - Identify open reading frames
    - Translate nucleotide sequence to protein sequence
- Predict RNA secondary structure
- Protein analysis
    - Predict secondary structure
    - Search for domains
    - Identify proteolytic cleavage sites
    - Create antigenicity, hydrophobicity, and charge plots
- Sequence analysis
    - Find motifs and patterns
    - Determine sequence composition
- Sequence alignment
    - Pairwise and multi-sequence alignments
    - BLAST
- Construct phylogenetic trees
- Cloning
- Restriction enzyme mapping
- PCR primers
    - Design
    - Determine primer properties (melting point, self annealing, secondary structure)

    - Identify binding site and PCR product

NGS tasks

- Pre-alignment tasks
    - Sequence quality check
    - Trimming
    - Demultiplexing
- Map sequencing data to reference
- De novo genome assembly
- Color space mapping
- Long read support

DNA sequencing

- Variant detection
- Variant annotation
- Predict functional consequence of variant

Gene expression

- Microarray
- RNA sequencing (can process spike-ins)
- miRNA sequencing

Epigenetics

- ChIP sequencing
- Bisulfite sequencing

Biological insights

- Directly interface with Qiagen Ingenuity Pathway Analysis (IPA) to extract biological insight.

Visualizations

- Genome browser

Information mining

- Download NGS data from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA)
- Obtain gene information from NCBI
- Obtain protein structure from PDB and sequences from UniProt

{{Edet}}

Recommendations

CLC Genomics Workbench and Ingenuity Pathway Analysis (IPA) are developed by Qiagen. Results from CLC Genomics can be imported to IPA for pathways analysis. The combination of CLC Genomics Workbench and IPA allows us to go from a gene level of understanding to understanding biological function and regulatory mechanisms.

Things to Know

Although CLC Genomics Workbench is comprehensive, compute resources on a user's local machine may be a limiting factor for analysis.

Input Data Types

- FASTQ/FQ
- FASTA/FNA/FA
- SAM
- BAM
- VCF
- TXT
- BAS.H5
- BASX.H5
- AB
- ABI
- AB1
- SCF
- PHD
- SFF
- GFF
- GTF
- BED
- WIGGLE
- Tracks/annotations from the UCSC Genome Browser and COSMIC database

We can import data to CLC Genomics Workbench from the following sequencing instruments:

- Illumina
- Oxford Nanopore
- PacBio
- Sanger
- Ion Torrent

Output Data Types

- Visualizations can be exported with varying resolutions as the following:
    - PDF
    - PNG
    - JPG
    - TIF
    - SVG
    - PS
    - EPS
- Tabular data can be exported as:
    - CSV
    - XLS
    - XLSX
    - Tab delimited TXT
    - HTML
    - GFF
    - Track graphics

Access Information

You must submit a request through service.cancer.gov *(https://service.cancer.gov/)* to obtain access to CLC Genomics Workbench. This software requires access to a floating license server (three simultaneous users), and so care should be taken to return the license when the software is not actively being used (i.e. close the application). Working with CLC Genomics Workbench requires login to the NIH network or VPN connection if remote.

Getting Help

Documentation for the CLC Genomics Workbench is available under the Help tab in the software. For the CLC Genomics Workbench manual, click here *(https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/2000/index.php?manual=Introduction_CLC_Genomics_Workbench.html)*. Additionally, there are tutorials *(https://digitalinsights.qiagen.com/support/tutorials/)* available for different workflows. Finally, Qiagen hosts webinars addressing the use of and updates to this software. To access these videos, click here *(https://tv.qiagenbioinformatics.com/channel/61793068/qiagen-clc-genomics)*.

Qiagen Ingenuity Pathway Analysis (IPA)

Description

Ingenuity Pathway Analysis (IPA) (*Qiagen*) IPA works with differential expression data (derived from RNA sequencing, miRNA sequencing, microarray, proteomics, phosphoproteomics, or metabolomics) or genetic variant data to extract various biological insights.

{{Sdet}}{{Ssum}}Listing of Analysis Functions{{Esum}}

- Discern affected molecular biology pathways and networks
- Provide insight on regulatory mechanisms including upstream regulators driving gene expression
- Predict effect on disease and function
- Reveal biomarkers and drug targets
- Identify miRNA targets
- Compare results against similar studies

{{Edet}}

Recommendations

IPA and CLC Genomics Workbench are both Qiagen products, which allows us to more easily extract biological insight from analysis. For this reason, it is beneficial to use these two packages together.

Things to Know

IPA provides an abundance of curated molecular bioscience information obtained from literature and databases. The curated content can be mined to gain insight and facilitate hypothesis generation.

Input Data Types

IPA takes tabular data as input. Differential gene expression data should contain columns with gene names, log fold change, and p value. We can also provide a table with genetic gain/loss of function information if this is the topic of our study. Thus, IPA is able to import files with the following extensions:

- XLSX
- XLS
- CSV
- Tab delimited TXT

Output Data Types

- Visualizations can be exported as the following under various resolutions
- PNG
- PDF
- Tabular data can be exported as
- Tab delimited TXT
- XLS

Access Information

You must submit a request through service.cancer.gov *(https://service.cancer.gov/)* to obtain access to Qiagen IPA. This software requires access to a floating license server, and so care should be taken to return licenses when the software is not actively being used (i.e., close the application). OSTR holds 6 concurrent licenses to IPA. Users must be either on the NIH network or connected via VPN if remote.

Getting Help

Qiagen IPA comes with extensive help documentations *(https://qiagen.secure.force.com/KnowledgeBase/KnowledgeIPAPage)*, tutorials *(https://qiagen.secure.force.com/KnowledgeBase/KnowledgeSearchPage?tb=tutorials&page=KnowledgeIPAPage&x=0&y=0)*, and video tutorials *(https://qiagen.secure.force.com/KnowledgeBase/articles/Basic_Technical_Q_A/IPA-Video-Tutorials)*.

Qlucore Omics Explorer

Description

Qlucore Omics Explorer (Qlucore) is a graphical user interface (GUI) based package used to generate visualizations, elucidate biological function, and classify samples as well as cells for omics data.

{{Sdet}}{{Ssum}}Listing of Analysis Functions{{Esum}}

Omics analyses

- Gene expression data derived from
- Microarray
- RNA or single cell RNA sequencing
- qPCR
- miRNA derived from microarray or sequencing
- Proteomics
- Metabolomics
- Lipidomics

Flow cytometry

Epigenetics

- Methylation (microarray)

Biological insights

- Gene set enrichment
- Gene ontology

Machine learning based sample and cell classifcation

- K nearest neighbor
- Support Vector Machines

- Random Trees
- Gradient boosted trees

Identify biomarkers

Visualizations

- PCA plots
- t-SNE plots
- Heat maps with hierarchical clustering
- Scatter plots
- Volcano plots
- Box plots

Interactive statistical analysis tools

- Two group comparison (t-test)
- Paired t-test
- Multi-group comparison (F-test) (ANOVA)
- Two-way ANOVA
- Linear, quadratic, and rank regression

{{Edet}}

Recommendations

Qlucore Omics Explorer does not perform upstream tasks in Next Generation Sequencing such as quality control, read mapping, or deriving gene expression counts. Use Partek Flow, Qiagen CLC Genomics Workbench, or Geneious Prime.

Things to Know

Qlucore Omics Explorer specializes in taking tabular omics data (e.g, RNA sequencing read counts) and producing robust visualizations. If an investigator is interested only in visualizations rather than tabular outputs, this would be the goto package.

Input Data Types

Standard formats generally accepted:

- BAM
- GTF
- QUANT.sf (Qlucore)
- GEDATA (Qlucore)
- CEL

- CHP
- BioArray Software Environment
- TXT
- CSV
- GEO
- SRA
- CYTOBAND
- 10X GENOMICS (requires barcodes.tsv, features.tsv, matrix.mtx)

Output Data Types

Images

- PNG
- JPG
- BMP
- TIF

Tabular data

- GEDATA (Qlucore)

Access Information

You must submit a request through service.cancer.gov *(https://service.cancer.gov/)* to obtain access to Qlucore Omics Explorer. This software requires access to a floating license server (OSTR has 5 licenses). Please be sure to close the application when you are finished so the license becomes available to others. To use Qlucore Omics Explorer, a connection to the NIH network or VPN is necessary.

Getting help

Qlucore Omics Explorer has extensive documentation *(https://qlucore.com/documentation)*. To access these, users will need to create an account on the Qlucore Omics website. There are also webinars *(https://qlucore.com/videos)* that showcase the use of various workflows. BTEP also hosts Qlucore Omics training on a frequent basis, check the Video Archive *(https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/)* for recordings of previous classes. Look out for future training on the BTEP calendar *(https://bioinformatics.ccr.cancer.gov/btep/)*.

SnapGene

Description

SnapGene is a point and click proprietary software program used for designing and documenting molecular biology experiments. SnapGene is a multipurpose software program used for, but not limited to, the following:

- DNA sequence alignment, annotation, editing, and visualization
    - Validate constructs with sequence alignments
    - Sanger sequence assembly
- PCR simulation, primer design, gel simulation
- Cloning and related methods
- View plasmid features and customize maps
- Protein visualization

{{Sdet}}{{Ssum}}Key Software Features{{Esum}}

The following material is from SnapGene:

- SnapGene makes your DNA manipulations easy to visualize and simulate, and alerts you to errors before they happen.
- Every DNA manipulation in SnapGene is automatically recorded, so you can see exactly what you did and retrieve the sequences of ancestral constructs.
- SnapGene's .dna files can be opened by the free cross-platform SnapGene Viewer, enabling you to share richly annotated maps and sequences with colleagues.
- SnapGene automatically generates a record of every sequence edit and cloning procedure, so you won't lose track of how a construct was made, even after a lab member leaves.

- SnapGene supports a host of file formats.

{{Edet}}

Recommendations

For the full spectrum of features available through SnapGene, click here *(https://www.snapgene.com/features/#all-features)*.

Things to Know

- Can import directly from NCBI, UniProt, and Ensembl using accession information
- Can be used with an extensive range of file types
- SnapGene Viewer is free allowing files to be easily viewed by collaborators
- Provides visualizations such as vector maps
- Provenance tracking
- Not recommended for high throughput sequence data.

Input Data

SnapGene can read alignment files from Clustal, GDE, MSF, NEXUS, PHYLIP, PIR, Selex, Stockholm.

In addition, SnapGene can read files from the following programs: ApE (.ape), CLC (.clc), Clone Manager, DNA Strider, DNADynamo (.cow), DNASIS (.dnasis), DNAssist (.seq), DNASTAR Lasergene (.seq, .sbd), DS Gene (.nas_bsml, .aas_bsm), EMBL format, EnzymeX (.exdna), Genbank and DDBJ files, Gene Construction Kit (.gcc), Geneious (.geneious), GeneTool (.bti), Genome Compiler (.gcproj), Jellyfish (.xml), MacVector (.nucl), pDRAW32 (.PDW), Sequencher (.spf), Serial Cloner (.xdna), Swiss-Prot sequence format, Vector NTI, and Visual Cloning (.vcd).

See more information here *(https://www.snapgene.com/features/convert-file-formats/)*.

Output Data

Supported output formats include DDBJ, EMBL, FASTA, GenBank - SnapGene, GenBank - Standard, GenBank - Vector NTI, GenPept - SnapGene, GenPept - Standard, Plain Text, SnapGene DNA, and SnapGene Protein.

Access Information

You must submit a request through [service.cancer.gov] (https://service.cancer.gov/SnapGene) to obtain access to SnapGene.

Getting Help

The SnapGene documentation is extensive. A number of tutorial videos and user guides are readily available from the SnapGene website *(https://www.snapgene.com/resources)*. Additionally, SnapGene has a new video learning center, SnapGene Academy *(https://www.snapgene.com/academy)*, which includes video tutorials on molecular biology concepts, theories, methods, and tools.