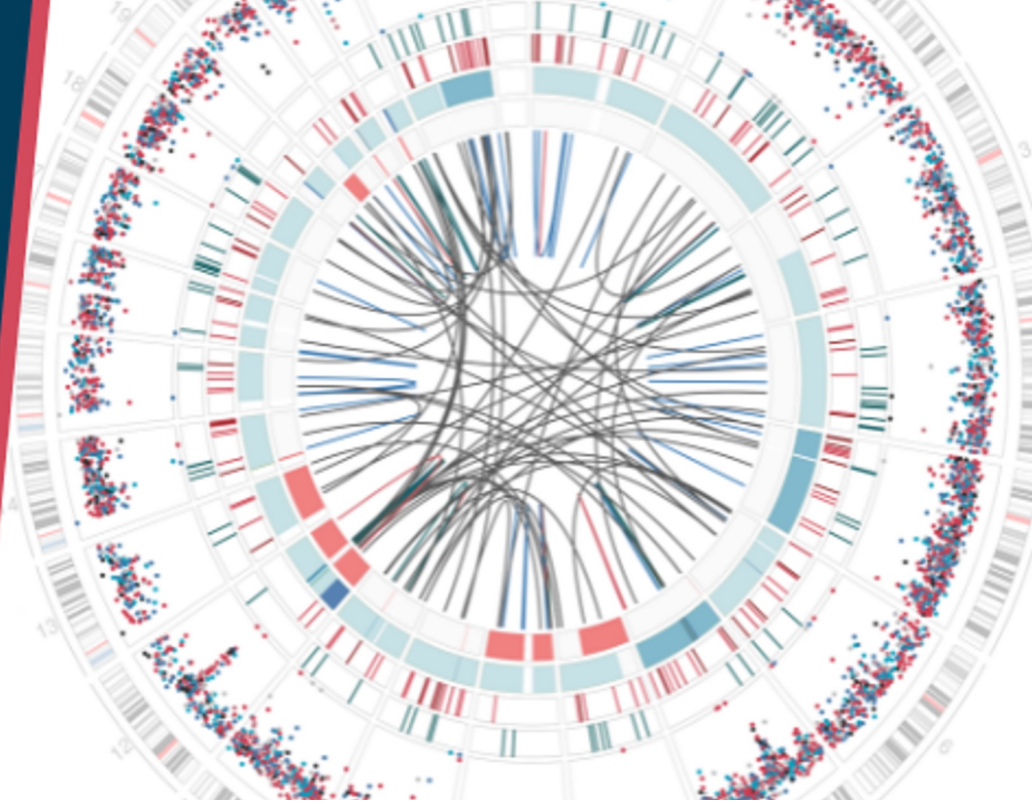
EMERGING APPROACHES FOR TUMOR ANALYSES IN EPIDEMIOLOGICAL STUDIES

A DCEG Annual Course

2022-23



Maria Teresa Landi, M.D., Ph.D.

Senior Advisor for Genomic Epidemiology
Senior Investigator, DCEG, NCI, NIH

A few key points

- Why this workshop
- Workshop target: epidemiological studies with tissue specimens
- Lectures and practical sessions (1st module, prep)
  - GitHub page for the course: https://nci-iteb.github.io/tumor_epidemiology_approaches/

A few key points

- Lectures will highlight a few points, more on the GitHub page
- Practical sessions in-person with instructors' help
  - Questions
    - Lectures: in chat or *raise your hand at the end of the lecture*. Instructors will respond to basic questions in the chat. A 'facilitator' will select the questions for the speakers
    - Practical sessions: raise your hand when needed and an instructor will come to the table

A few key points

- Biowulf accounts for the workshop covered by the Division
  - accounts will be closed after 30 days from the end of the workshop
- Have Biowulf interactive mode ready for the practical sessions!
- Have laptop fully charged before the practical sessions

A few key points

- ‘Ask a Bioinformatician’ sessions at the end of each module
  - open to user’s data. Questions can be sent in advance to Phuc Hoang (phuc.hoang@nih.gov)
- Feedback form at the end of each module and at the end of the workshop
- Mini-series of DCEG-wide speakers linked to all workshop topics (link of recording on the website/GitHub page)

Invited Speakers

Ludmil Alexandrov, Ph.D.
January 19th, 2023 -

Anthology of unusual patterns of somatic mutations in cancer genomes

Peter Park, Ph.D.
February 16th, 2023 -

Structural alterations in cancer genomes

Núria López-Bigas, Ph.D.
March 23rd, 2023 -

Somatic mutations in tumors and normal tissues

David Wedge, Ph.D.
April 20th, 2023 -

Tumour evolution in diverse human populations

Marcin Imieliński, M.D., Ph.D.
April 20th, 2023 –

Origin and identity in lung adenocarcinoma evolution

Charles Swanton, M.B./Ph.D.
May 25th, 2023 -

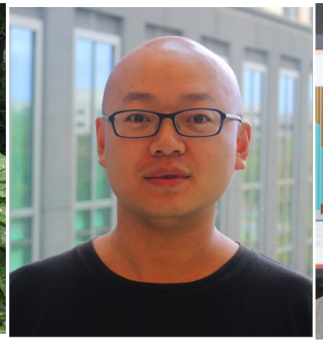
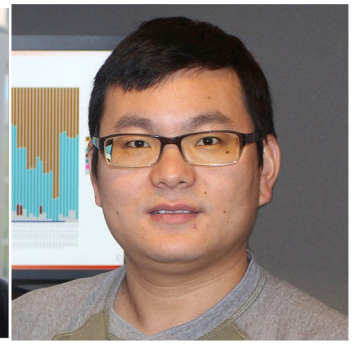
Title TBD

David Adams, Ph.D.
June 8th, 2023 -

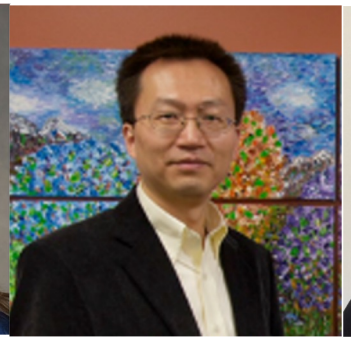
Cross-species oncogenomics of melanoma and other malignancies to define disease drivers

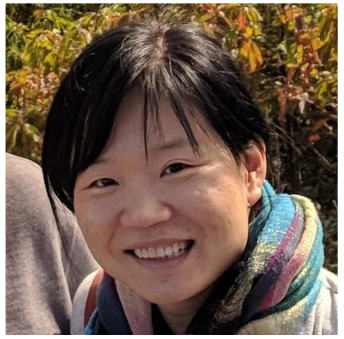
Instructors

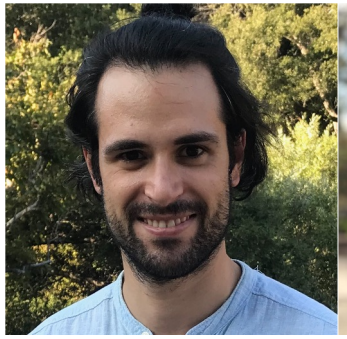











NCI/DCEG-ITEB
Phuc Hoang, Ph.D.
Alyssa Klein
John McElderry
Jian Sang, Ph.D.
Tongwu Zhang, Ph.D.
Wei Zhao, Ph.D.
Teresa Landi, M.D., Ph.D.

NCI/DCEG -CGR
Kristie Jones
Difei Wang, Ph.D
Wei Zhu, Ph.D.

NCI/DCEG-TDRP
Wendy Wong, Ph.D.

U Chicago
Lixing Yang, Ph.D.

UC San Diego
Marcos Díaz-Gay, Ph.D.
Mariya Kazachkova

Session 1: Introduction to Computing Clusters and Bioinformatics

Emerging Approaches for Tumor Analyses
in Epidemiological Studies

November 2, 2022
9:30 AM- 12:00 PM

Session Overview

- **Introduction to NCI Computing Clusters Available**
- **Cluster How-Tos: Connect, Transfer Files/Share Data**
- **Bash, Linux, Vim**
- **Bioinformatics File Formats and Tools - Part A**
- **Bioinformatics File Formats and Tools - Part B**

Workshop website: https://nci-iteb.github.io/tumor_epidemiology_approaches/

Introduction to NCI Computing Clusters Available

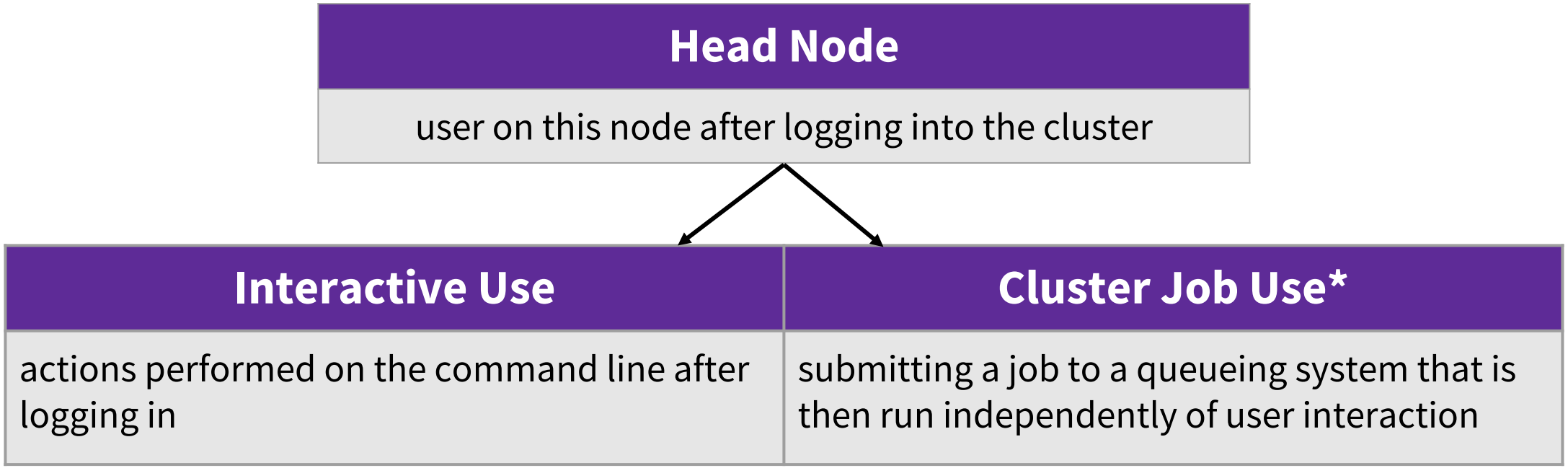
Clusters Being Discussed Today

Compute Cluster at DCEG (CCAD)	Biowulf
Cluster available to DCEG only	Cluster available across the entire NIH intramural research community

NIH NATIONAL CANCER INSTITUTE
Division of Cancer Epidemiology & Genetics

NIH National Institutes of Health
Turning Discovery Into Health

Introduction to the Compute Cluster at DCEG (CCAD)



- Operates on fair use policy to avoid monopolization of resources

*primary and preferred method to use the cluster

https://nih.sharepoint.com/sites/NCI-DCEG-myDCEG/SitePages/Cluster-Use-Policies.aspx

CCAD Cluster Nodes and Memory

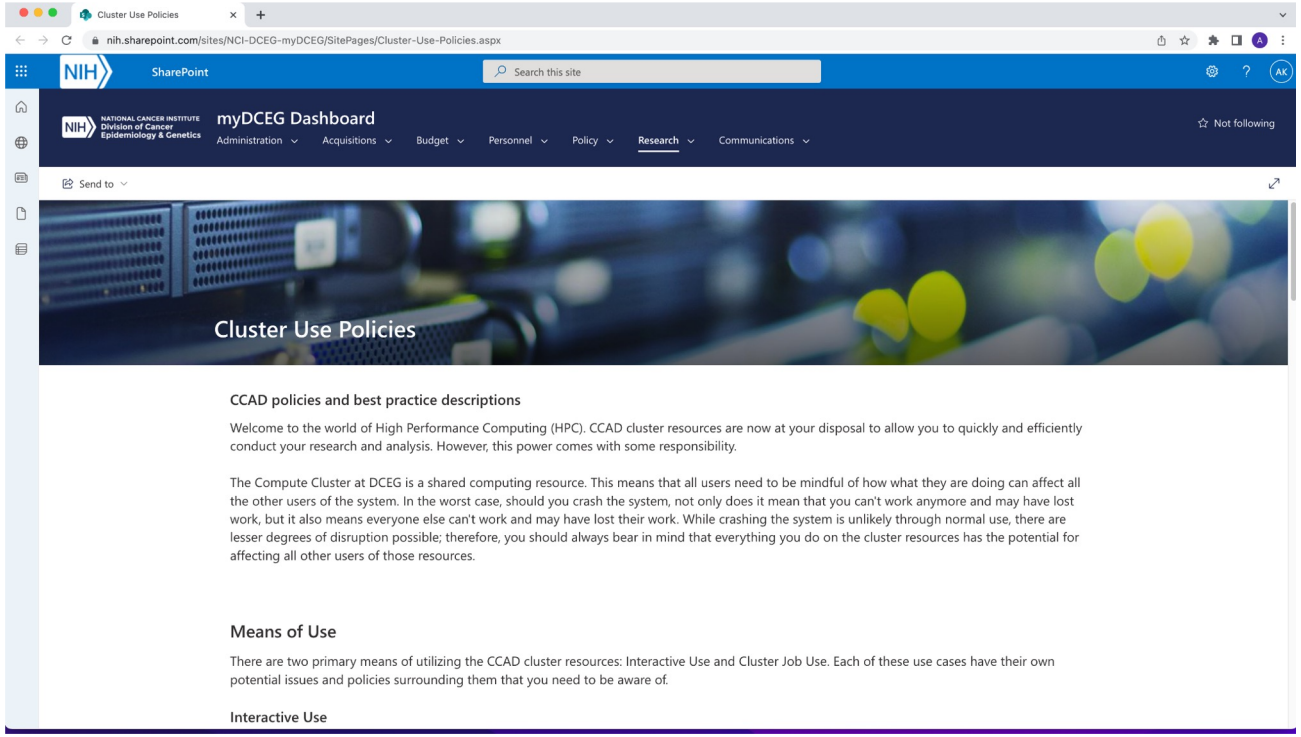
Hostname	Memory	Cores (2 threads each)	Networking
nodes 001-041	135.4 GB	16	dedicated 1Gbps/node
nodes 042-089	270.9 GB	20	Aggregate 80Gbps / 16 nodes
nodes 090-136	542 GB	32	Aggregate 80Gbps / 16 nodes
nodes 137-144	3253.2 GB	56	Aggregate 80Gbps / 16 nodes
ccad-master	135.2 GB	16	10 Gbps
ccad-master2	542 GB	32	10 Gbps
build-compute	135.4 GB	16	1 Gbps
cgemslll	270.9 GB	16	10 Gbps
ccad	270.9 GB	16	10 Gbps
ccad2	270.9 GB	16	10 Gbps

https://nih.sharepoint.com/sites/NCI-DCEG-myDCEG/SitePages/Cluster-Nodes-and-Memory.aspx

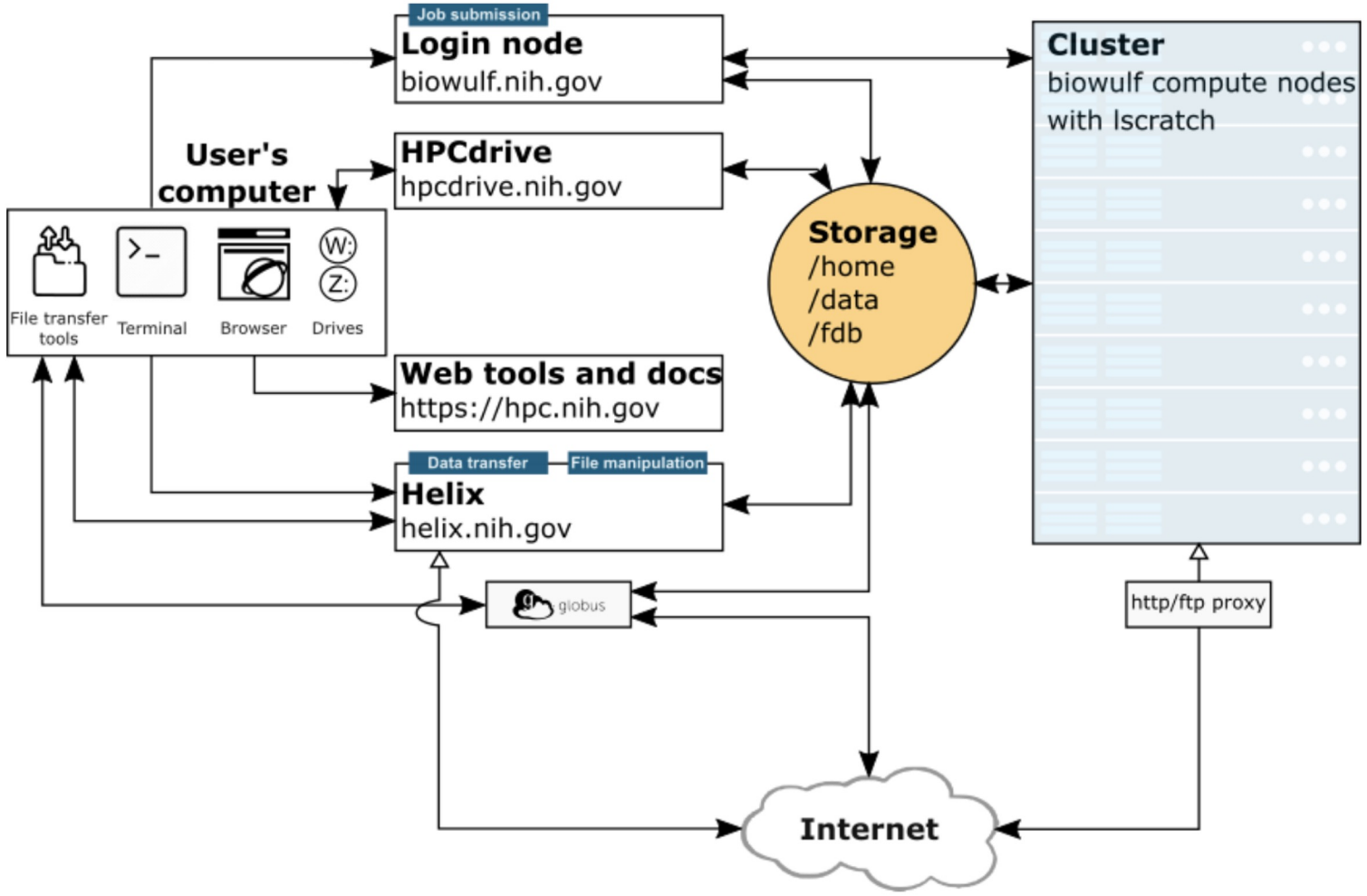
CCAD: Additional Information

Additional CCAD information can be found on myDCEG:

- Cluster Use Policies
- Cluster Nodes and Memory
- Beginner User Guide
- CCAD Account Request Form



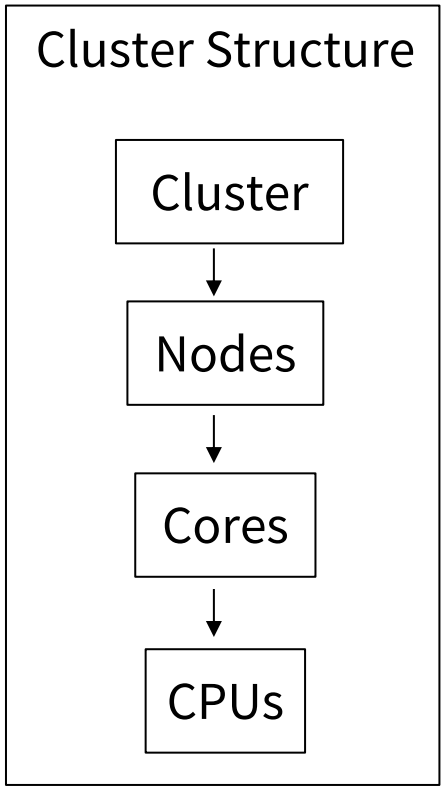
Introduction to Biowulf



https://hpc.nih.gov/systems/

Biowulf CPUs, Nodes, and Memory

- 100,000+ cores
- 200,000+ CPUs
- 4,000+ nodes
- 900+ TB memory
- 3+ PB local scratch (lscratch)
- ~40 PB high performance storage
- 5 PB object storage
- 800+ GPUs (4,000,000+ CUDA cores)
- In the top 500 most powerful commercially available computer systems in the world



https://hpc.nih.gov/training/handouts/nitish_wolfgang_data_science_team_20210312.pdf

Biowulf Applications

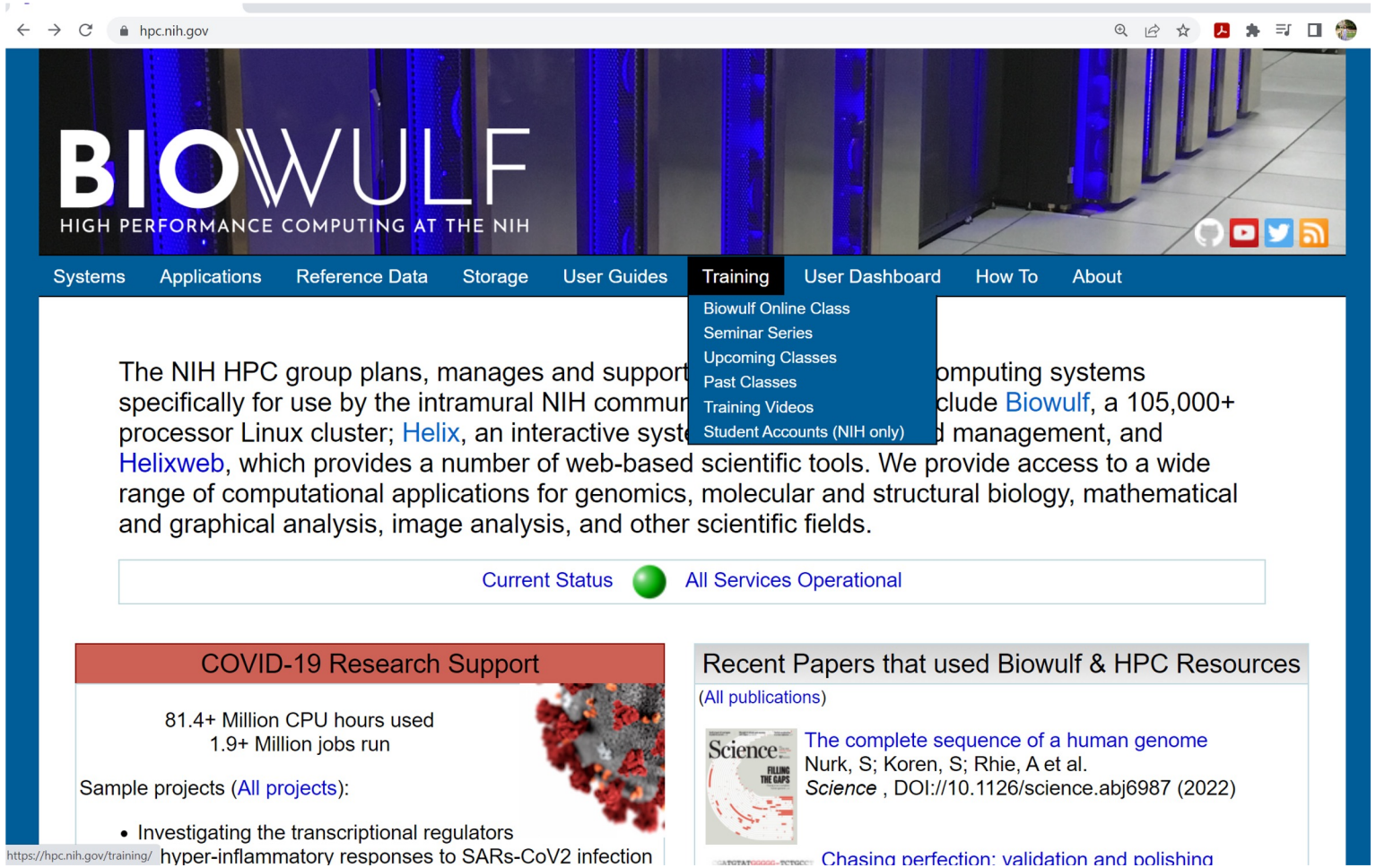
- Multiple versions for ~1000 applications available in ~4000 modules
- Multiple versions of python available with ~500 packages each
- Multiple versions of R available with ~1600 packages each
- jupyter and rstudio available
- singularity for containerization

https://hpc.nih.gov/training/handouts/nitish_wolfgang_data_science_team_20210312.pdf

Quick Start to Biowulf

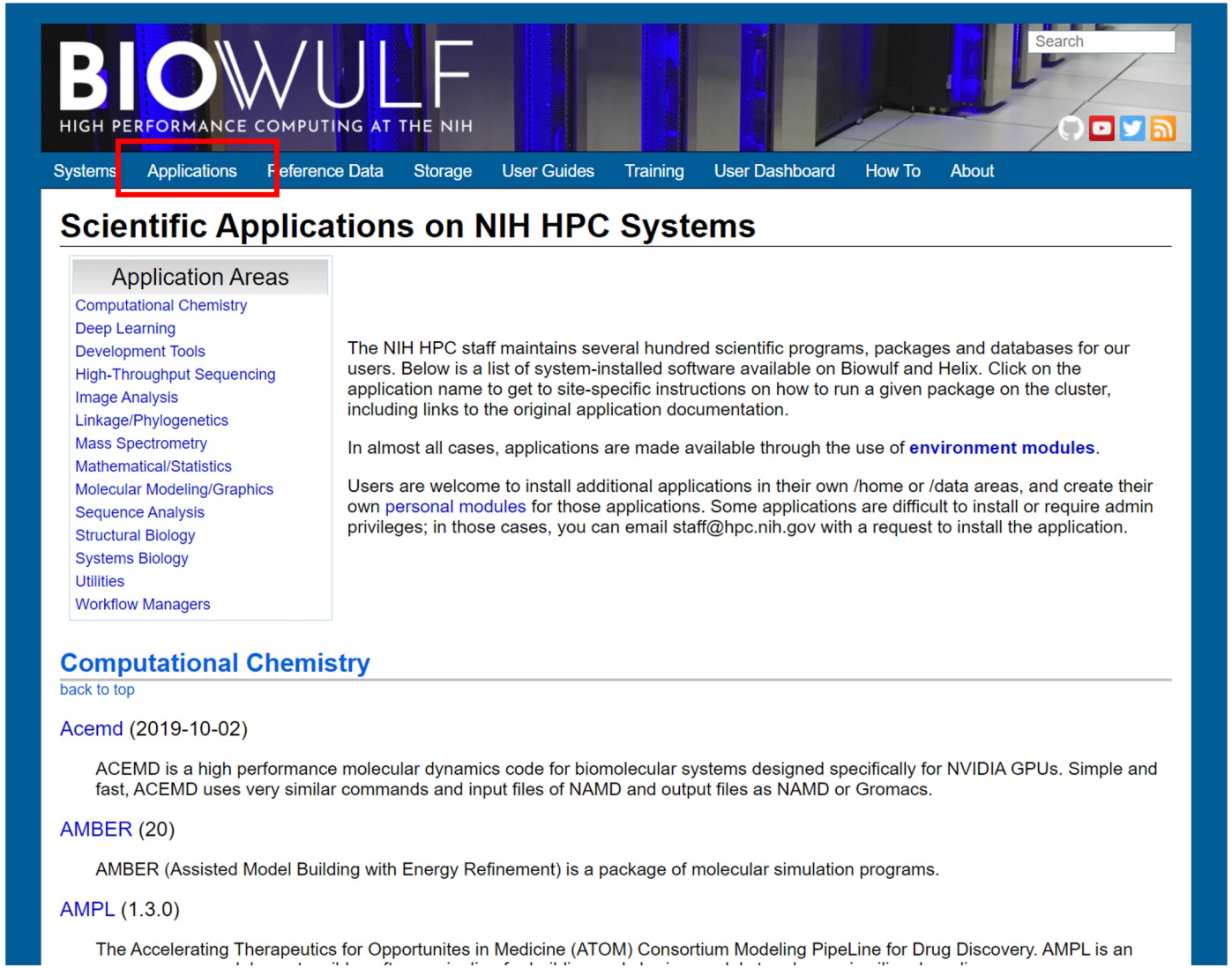
- Biowulf website: https://hpc.nih.gov/
  - Biowulf online classes: https://hpc.nih.gov/training/
  - “How to” help (e.g. connect to helix/biowulf, transfer files, request for space, etc.)
  - Manual files for scientific applications
- Submit tickets: Send email to staff@hpc.nih.gov
  - Examples: login issues, install packages/applications, troubleshooting help, extend job time beyond the limit
- Monthly Zoom-in consults: https://hpc.nih.gov/training/

Quick Start to Biowulf: Training



https://hpc.nih.gov/training/

Quick Start to Biowulf: Applications



https://hpc.nih.gov/apps/

Biowulf and CCAD Cluster Comparison

	Biowulf	CCAD
Job submission	sbatch --cpus-per-task=# --mem=#g --job-name <JobName> --time ##:##:## myscript.sh swarm -g <gb> -t <threads> -b <bundle> --job-name <JobName> --time ##:##:## myscript.sh	module load sge qsub -N JobName -e error.e -o output.o --cpus-per-task # --mem #g myscript.sh
Interactive jobs	sinteractive --cpus-per-task=# --mem=#g	module load sge qlogin [options], qsh
Delete jobs	scancel <job id> scancel --name=<JobName?	qdel job_id [options]
Monitor jobs	squeue, sjobs, jobload, jobhist	qstat
Partitions	norm*, multinode, largemem, unlimited, quick, gpu, visual	all.q, galaxy.q, research.q, seq-alignment.q, seq-calling.q, seq-gvcf.q, long.q, xlong.q, interactive.q, bigmem.q
Load applications	module load <module>	module load <module>
Backup policy	• Home: Weekly backups, with daily incremental backups • Data: NOT BACKED UP. 2 nightly snapshots and 1 weekly snapshot • Buy-in storage • Additional information: File Backups and Snapshots on the HPC Systems	• Nightly snapshots last one week • 6 hour snapshots last 3 days • True backups done via CBIIT taken weekly and retained based on their policies. • Permanent backups need to be requested to be transferred to the archive.

Snapshots

- A read-only copy of the data at a particular point in time
- Very helpful if you inadvertently delete a file
- To locate the file you are interested in, you can go to a specific snapshot by using the following command:

cd .snapshot

- This will take you to a main snapshot directory that contains all of the other snapshot directories. Find the file you need and then copy it back to the desired directory.
- More information can be found here: https://hpc.nih.gov/storage/backups.html

Cluster How-Tos: Connect, Transfer Files/Share Data

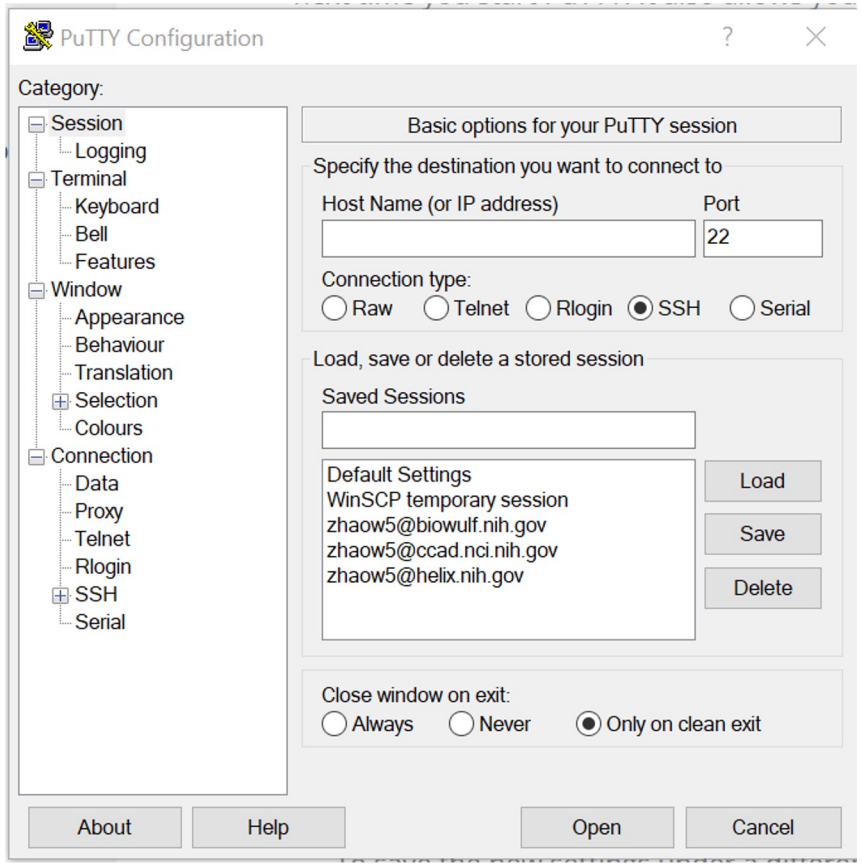
How to Connect to Helix/Biowulf/CCAD

Host	Hostname	Accessible by	Purpose
Biowulf	biowulf.nih.gov	All HPC users	cluster head node
Helix	helix.nih.gov	All HPC users	data transfer
HPCdrive	hpcdrive.nih.gov	All HPC users	data transfer
CCAD	ccad.nci.nih.gov	All HPC users	cluster head node
CCAD/T-drive	gigantor.nci.nih.gov	All HPC users	data transfer

- SSH connection (https://hpc.nih.gov/docs/connect.html)
  - Windows: PuTTY
  - Mac: Terminal/iTerm
- GUI file transfer clients (https://hpc.nih.gov/docs/transfer.html)
  - Windows: WinSCP/Filezilla
  - Mac: Fugu/Filezilla
- Data transfer and sharing (https://hpc.nih.gov/storage/sharing_data.html)
  - Mount HPC system directories (https://hpc.nih.gov/docs/transfer.html)
  - Globus (not set up for users by default for CCAD) (https://hpc.nih.gov/docs/transfer.html)
  - Command line (https://hpc.nih.gov/docs/transfer.html)

SSH Connection: Windows

- PuTTY Application

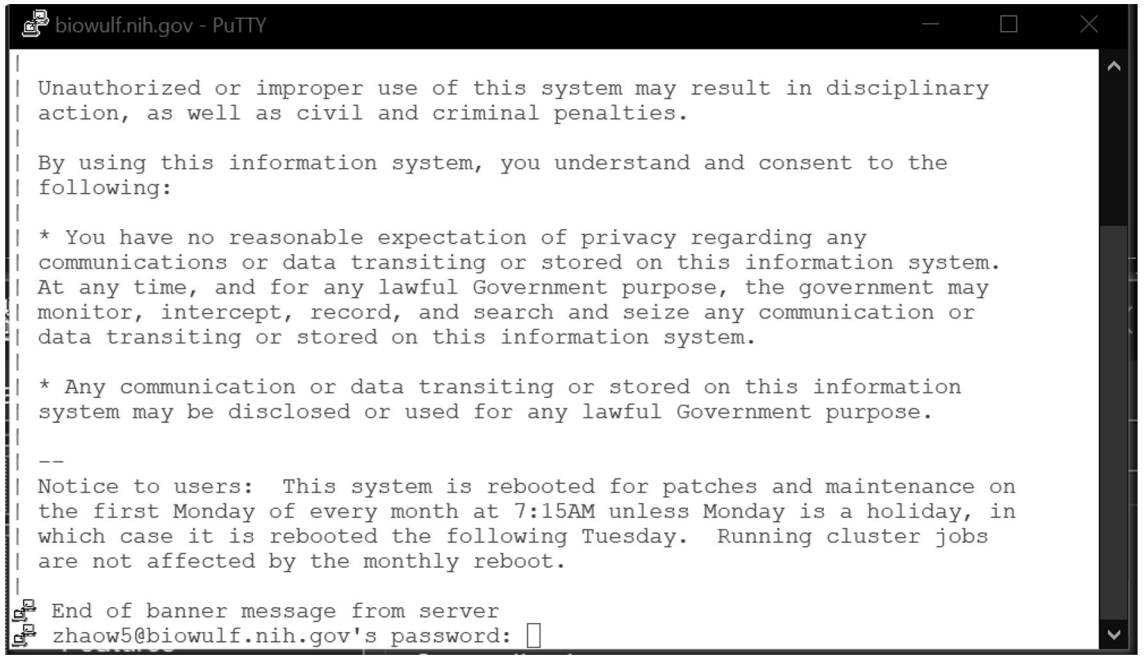


PuTTy Configuration menu

Note: The option for X11 forwarding needs to set up in the PuTTY configuration menus.

This allows for the use of graphical applications on a remote server (ex. Integrative Genomics Viewer- IGV).

Another option for running graphics applications is NoMachine (NX) (https://hpc.nih.gov/docs/nx.html#setup)



Logging into Biowulf using PuTTY

https://hpc.nih.gov/docs/connect.html

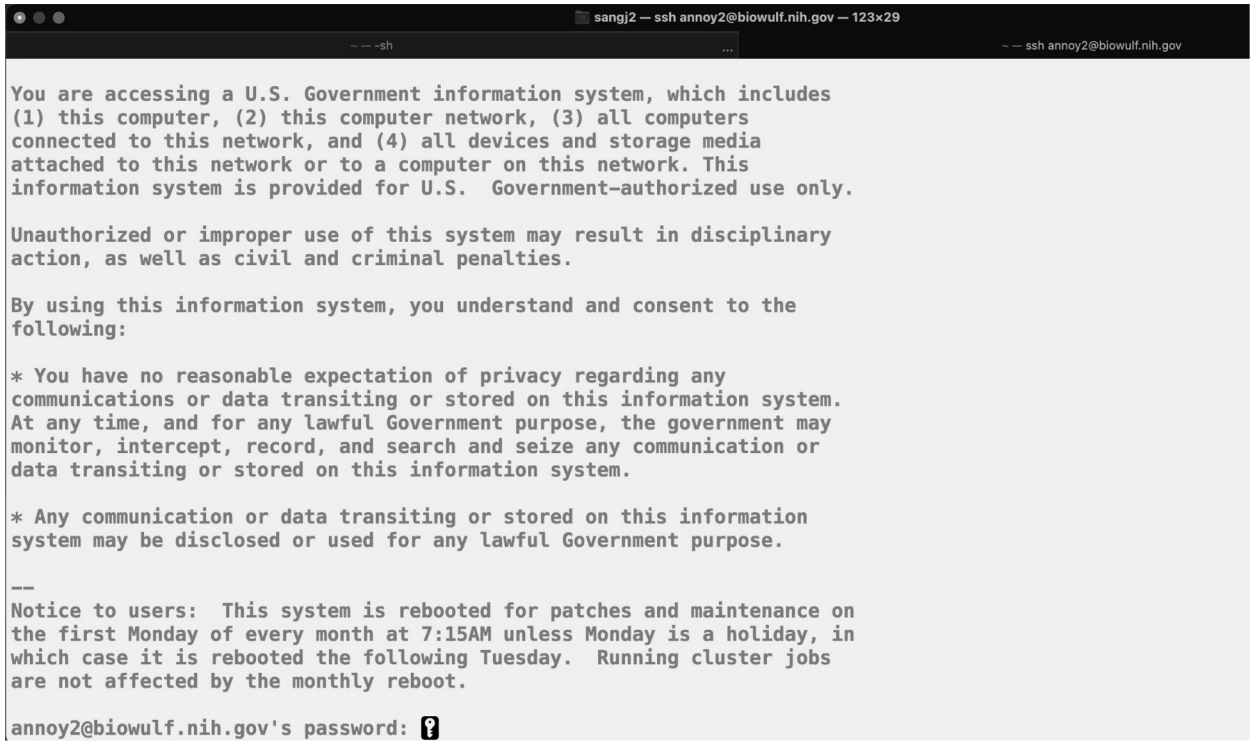
SSH Connection: Mac

- No additional application necessary
- Use Mac Terminal

Note: ***ssh -Y user@hostname*** enables trusted X11 forwarding.

X11 forwarding allows for the use of graphical applications on a remote server (ex. Integrative Genomics Viewer- IGV).

Another option for running graphics applications is NoMachine (NX) (https://hpc.nih.gov/docs/nx.html#setup)



Logging into Biowulf using Terminal

Connecting to CCAD and Submitting a Job

1. Log in to the cluster

   ssh ccad.nci.nih.gov

2. Load Sun Grid Engine module (contains commands needed to submit jobs)

   module load sge

3. Set up a script to submit with the job- specify qsub options in the script, or specify on the command line

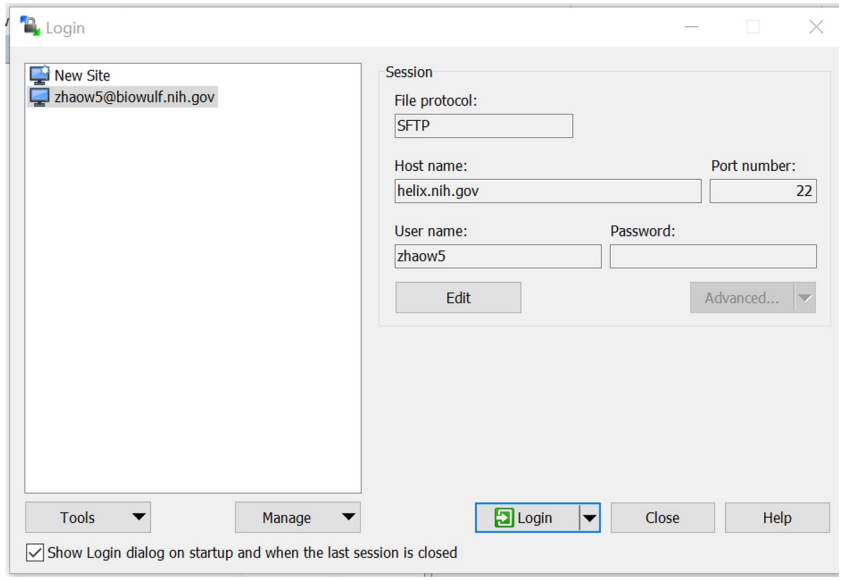
   Example: ***qsub -N jobname -e error.e -o output.o myscript.sh***

4. Check the status of the job

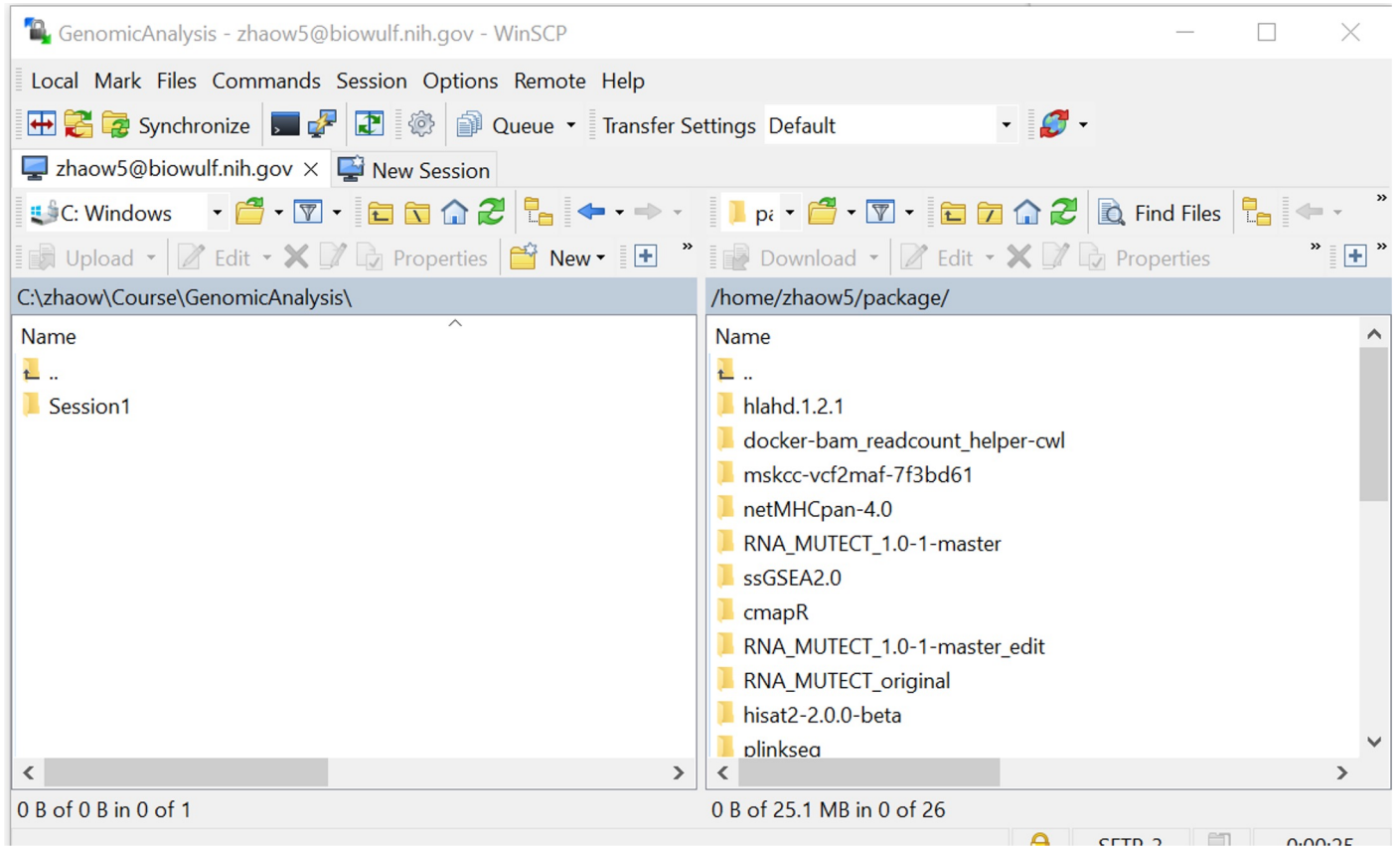
   qstat

GUI File Transfer Clients: WinSCP

Windows: WinSCP



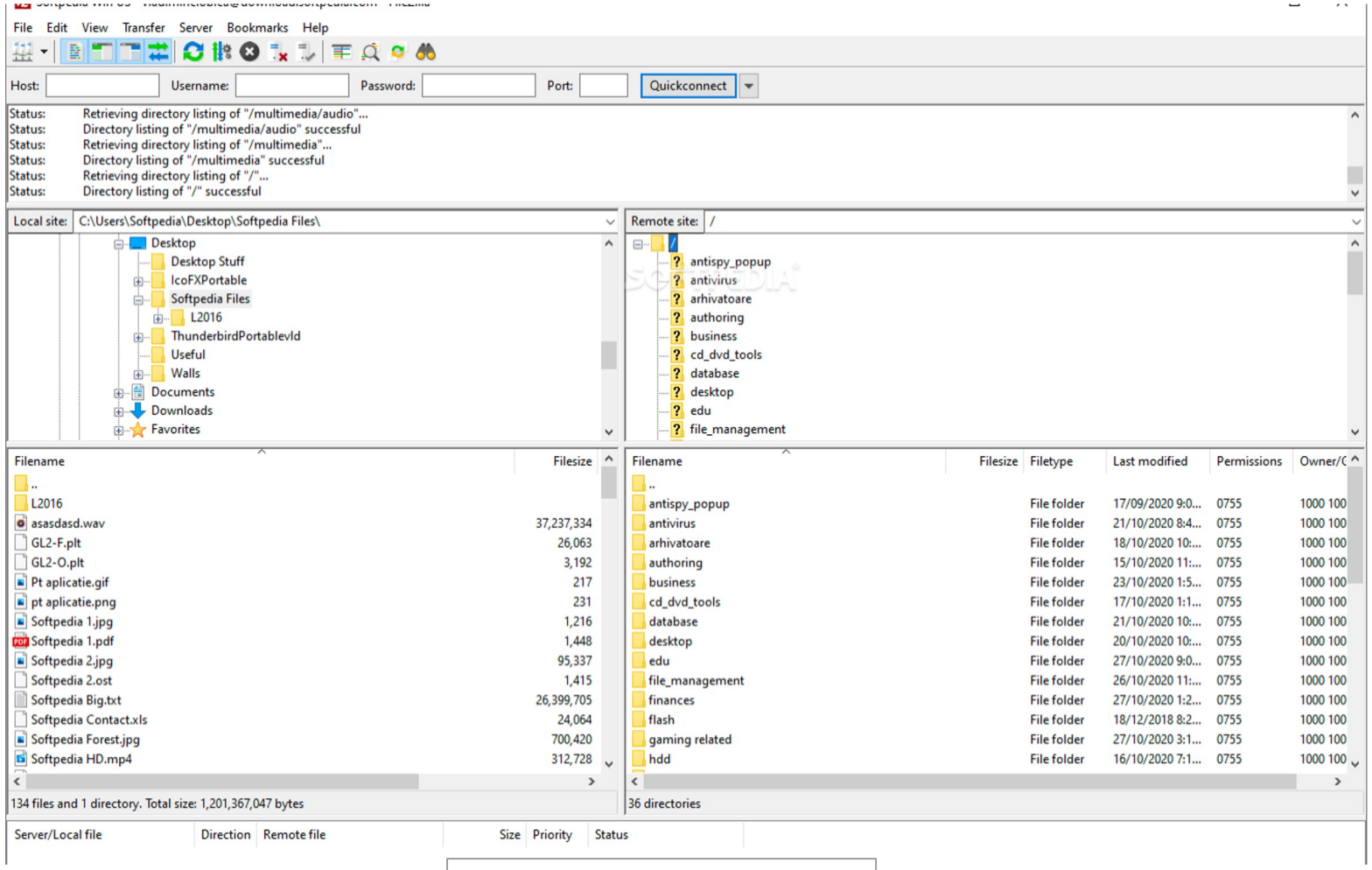
Login window for WinSCP



WinSCP file transfer window

GUI File Transfer Clients: Filezilla

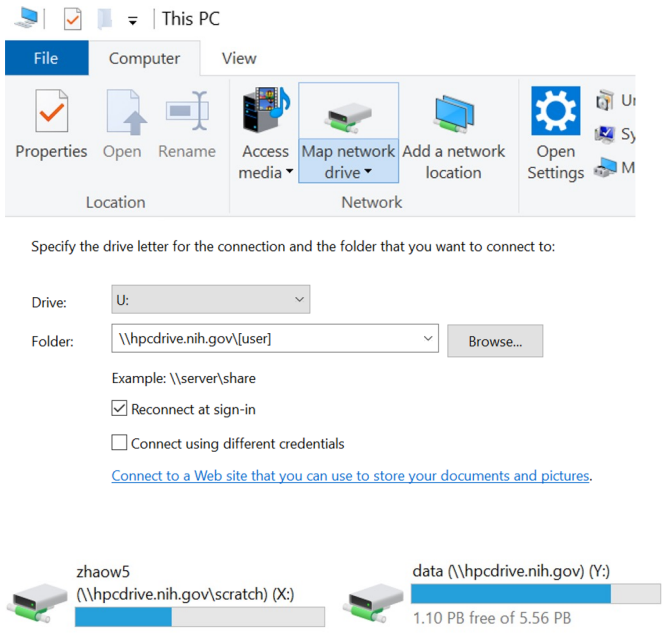
Windows/Mac: Filezilla



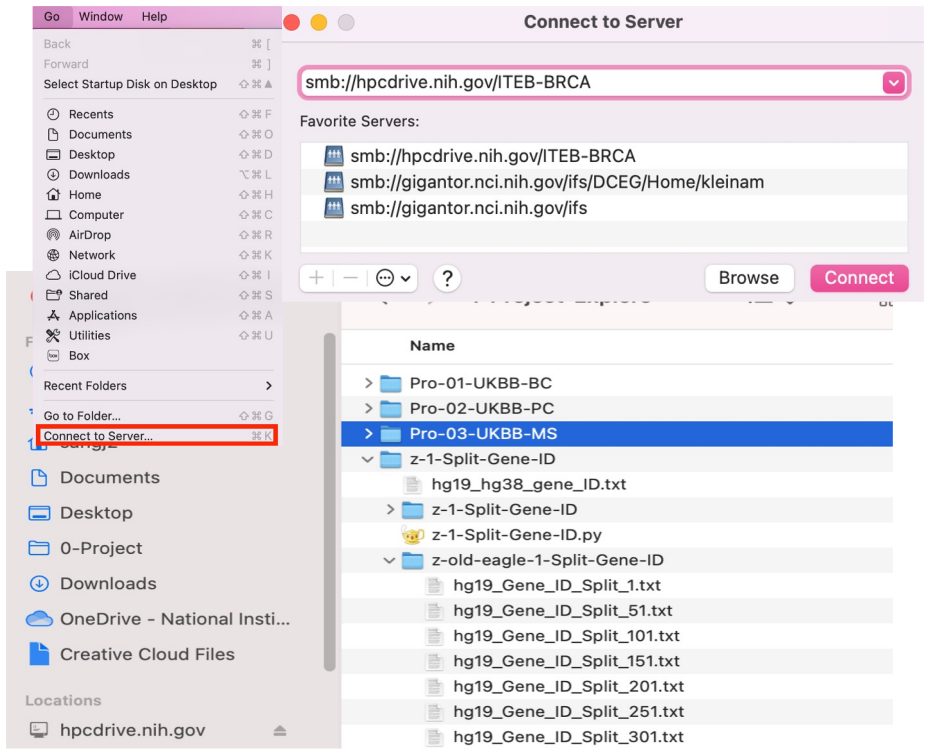
Filezilla file transfer window

Mount to the server: Windows and Mac

Windows



Mac



For more information: https://hpc.nih.gov/docs/transfer.html

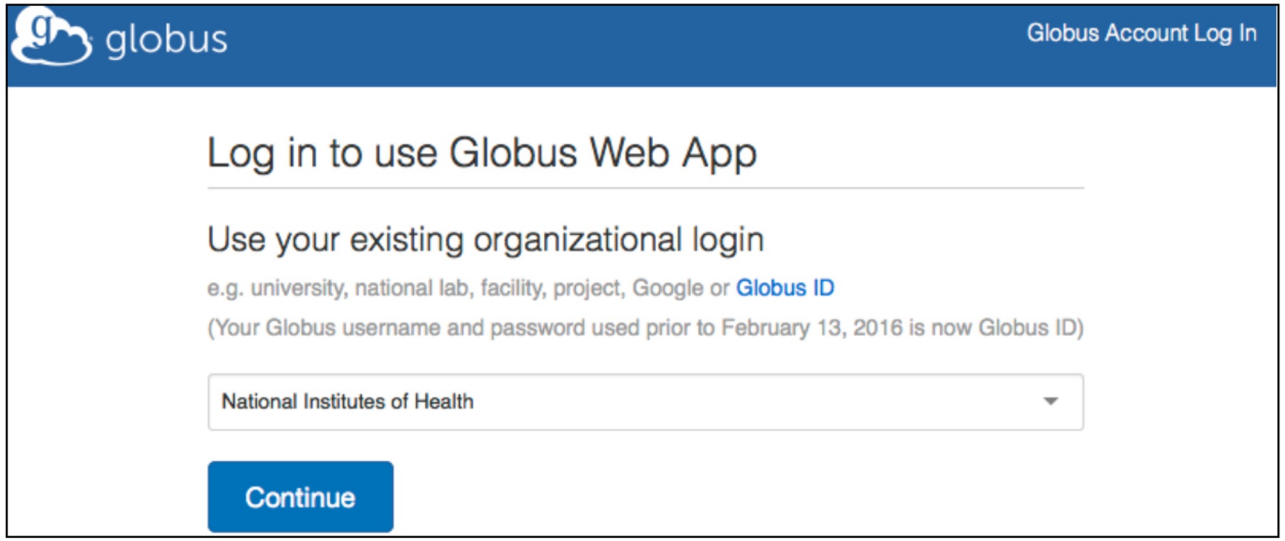
Mount to the server: Windows and Mac

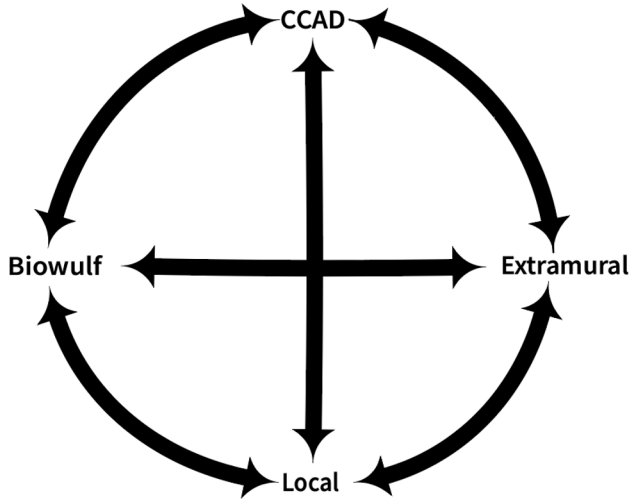
	Description	Directory at cluster	SMB path for Windows	SMB path for Mac
Biowulf/Helix	user's home directory	/home/[user]	\\hpcdrive.nih.gov\[user]	smb://hpcdrive.nih.gov/[user]
	data directory	/data/[user]	\\hpcdrive.nih.gov\data	smb://hpcdrive.nih.gov/data
	user's scratch space	/scratch/[user] directory	\\hpcdrive.nih.gov\scratch\[user]	smb://hpcdrive.nih.gov/scratch/[user]
	shared group area (e.g. you are a member of group Sherlock_Lung)	/data/Sherlock_Lung	\\hpcdrive.nih.gov\PQRlab	smb://hpcdrive.nih.gov/Sherlock_Lung
CCAD	main directory	/home/	\\gigantor.nci.nih.gov\ifs	smb://gigantor.nci.nih.gov/ifs
	user's home directory	/home/[user]	\\gigantor.nci.nih.gov\ifs\DCEG\Home\[user]	smb://gigantor.nci.nih.gov/ifs/DCEG/Home/[user]

For more information: https://hpc.nih.gov/docs/transfer.html

Sharing Data Through Globus

- Recommended way to transfer data to and from the HPC system
- Manage file transfer, monitor performance, retry failures, recover from faults automatically, and report the status
- How to set up: https://hpc.nih.gov/storage/globus.html



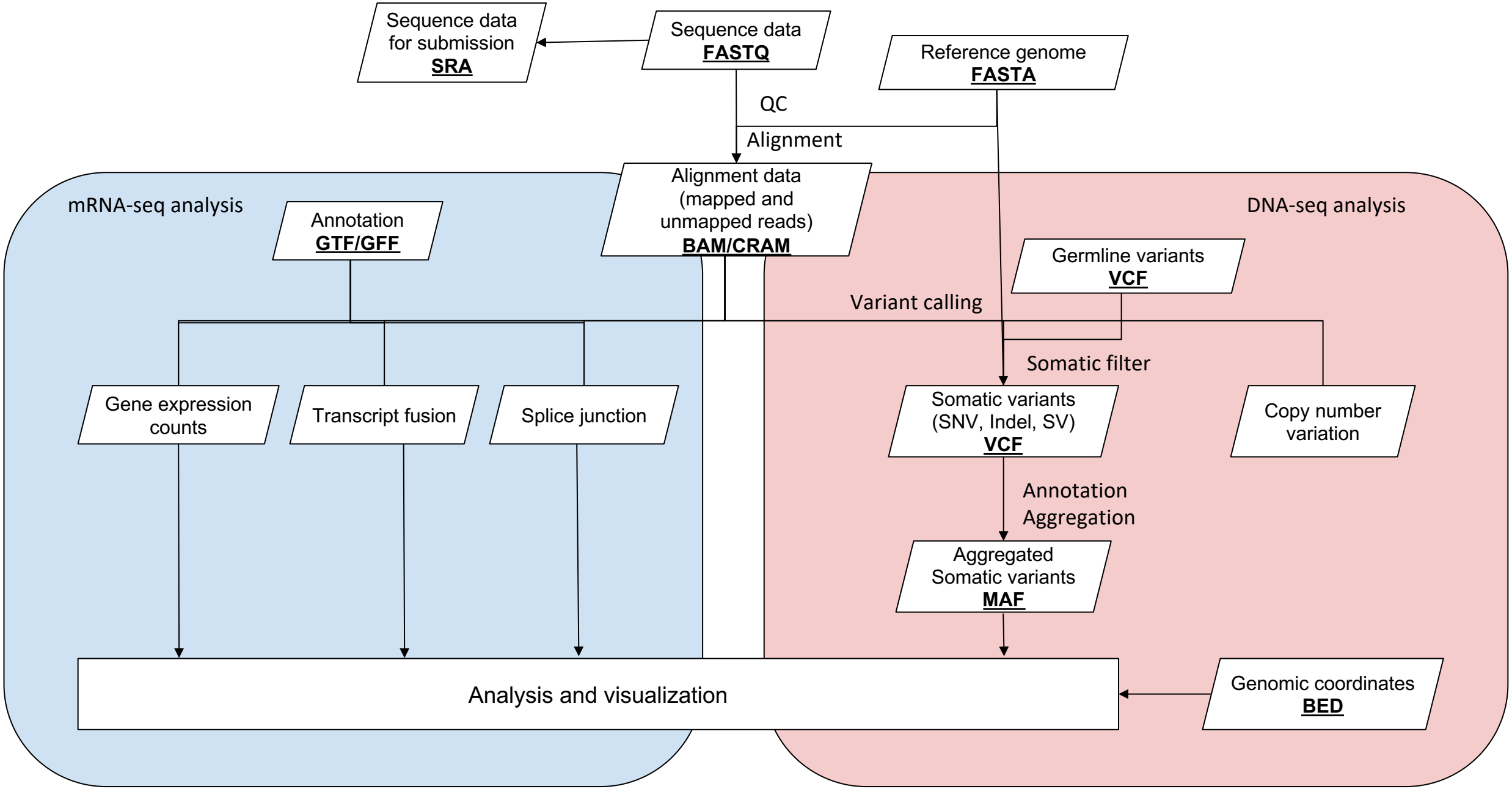


Bash, Linux, Vim

Resources: Bash, Linux, Vim

- The ‘shell’ is a high-level language in the linux system. Biowulf and CCAD use BASH shell.
- Introduction to Linux and bash shell scripting:
  - Online classes
    - https://hpc.nih.gov/training/bash_class/
  - Handout and slides:
    - https://hpc.nih.gov/training/handouts/Introduction_to_Linux.pdf
    - https://hpc.nih.gov/training/handouts/BashScripting.pdf
  - Other resources for bash commands:
    - https://explainshell.com/
- Cheatsheets of bash commands:
  - https://hpc.nih.gov/training/handouts/BashScripting_LinuxCommands.pdf
  - Linux cheatsheet
- Cheatsheet of Vim commands:
  - Vim cheatsheet

Bioinformatics File Formats and Tools - Part A


Sequence data for submission SRA
Sequence data FASTQ
Reference genome FASTA
QC
Alignment
Alignment data (mapped and unmapped reads) BAM/CRAM
mRNA-seq analysis
DNA-seq analysis
Annotation GTF/GFF
Germline variants VCF
Variant calling
Somatic filter
Gene expression counts
Transcript fusion
Splice junction
Somatic variants (SNV, Indel, SV) VCF
Copy number variation
Annotation Aggregation
Aggregated Somatic variants MAF
Analysis and visualization
Genomic coordinates BED

Data share and deposit - raw data

- Raw data with per-base quality score: fastq (reads), SAM/BAM/CRAM (alignment)
- Fastq file (uncompressed text file):
  - 1 nucleotide base ~ 2 bytes (sequence + quality score)
  - Example: 30x whole genome sequencing: 3GB * 30 * 2 ~ 180GB
  - Compressed files of fastq typically compress to 25% of original size.

Data share and deposit - raw data

- BAM/CRAM files are most common formats for data share and deposit.
- BAM and CRAM are both compressed forms of SAM (Sequence Alignment Map); BAM (for Binary Alignment Map) is a lossless compression while CRAM can range from lossless to lossy.

SAM, BAM and CRAM

File format	File size (GB)
SAM	7.4
BAM	1.9
CRAM lossless	1.4
CRAM 8 bins	0.8
CRAM no quality scores	0.26

BAM file for different studies

	Sequencing depth	Number of million reads	Read length	BAM file size
Whole genome sequencing (Sherlock-*Lung*)	~71X	1,430	150bp	~340G
Whole exome sequencing (TCGA)	~50X	198	100bp	14.2G
RNA-seq (Sherlock-*Lung*)		145	150bp	19G
RNA-seq (TCGA)		68	50bp	7-9G

Data share and deposit - processed data

- Variants calling results could be shared in the formats of VCF or MAF.
- In somatic genomic study, each VCF file is generated for one tumor/normal pairs, while in the germline studies it contains pooled variants for all samples.
- MAF files are aggregated mutation information from VCF Files and are generated on a project-level. MAF files also includes annotation of variants from the public database.

Example: VCF file (body section)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1234567	microsat1	GTCT	G,GTACT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Example: MAF file

Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2	dbSNP_RS	dbSNP_Val_Status	Tumor_Sample_Barcode	Matched_Norm_Sample_Barcode
PRAMEF14	729528	BI	GRCh38	chr1	13344364	13344364	+	Silent	SNP	C	C	G	novel		TCGA-55-8615-0	TCGA-55-8615-10/
ACTL8	81569	BI	GRCh38	chr1	17826321	17826321	+	Missense_Mutation	SNP	C	C	A			TCGA-55-8615-0	TCGA-55-8615-10/
TMCO4	255104	BI	GRCh38	chr1	19746510	19746510	+	Missense_Mutation	SNP	C	C	T			TCGA-55-8615-0	TCGA-55-8615-10/
OR2T12	127064	BI	GRCh38	chr1	248295475	248295475	+	Missense_Mutation	SNP	G	G	C	rs749730602		TCGA-55-8615-0	TCGA-55-8615-10/
CMYA5	202333	BI	GRCh38	chr5	79734231	79734231	+	Missense_Mutation	SNP	A	A	T			TCGA-55-8615-0	TCGA-55-8615-10/
ERLEC1	27248	BI	GRCh38	chr2	53794372	53794372	+	Missense_Mutation	SNP	G	G	C			TCGA-55-8615-0	TCGA-55-8615-10/
MOGS	7841	BI	GRCh38	chr2	74461426	74461426	+	Missense_Mutation	SNP	C	C	T	rs199724485		TCGA-55-8615-0	TCGA-55-8615-10/
TCF7L1	83439	BI	GRCh38	chr2	85302600	85302600	+	Missense_Mutation	SNP	G	G	C	rs1486052719		TCGA-55-8615-0	TCGA-55-8615-10/
TTN	7273	BI	GRCh38	chr2	178545553	178545553	+	Missense_Mutation	SNP	G	G	T			TCGA-55-8615-0	TCGA-55-8615-10/
MAP2	4133	BI	GRCh38	chr2	209694182	209694182	+	Missense_Mutation	SNP	G	G	A	rs761466743		TCGA-55-8615-0	TCGA-55-8615-10/
MAPKAPK3	7867	BI	GRCh38	chr3	50646777	50646777	+	Silent	SNP	C	C	T	rs1473215021		TCGA-55-8615-0	TCGA-55-8615-10/
MYH15	22989	BI	GRCh38	chr3	108500201	108500201	+	Missense_Mutation	SNP	G	G	C			TCGA-55-8615-0	TCGA-55-8615-10/

Prepare data for submission

- NCBI Sequence Read Archive (SRA)
  - requires raw data with per-base quality scores for all submitted data.
  - accepts binary files such as BAM/CRAM, HDF5 (for PacBio, Nanopore), SFF (when BAM is not available, for 454 Life Science and Ion Torrent data), and text formats such as FASTQ.
  - For more information: https://www.ncbi.nlm.nih.gov/sra/docs/submit/
- Gene Expression Omnibus (GEO)
  - studies concerning quantitative gene expression, gene regulation, epigenetics, or other functional genomic studies. (e.g. mRNA-seq, miRNA-seq, ChIP-Seq, HiC-seq, methyl-seq/bisulfite-seq)
  - does NOT accept WGS, WES, metagenomic sequencing, or variation or copy number projects.
  - a complete submission includes: metadata, processed data, raw data containing sequence reads and quality scores (will be submitted to SRA by the GEO team)
  - For more information: https://www.ncbi.nlm.nih.gov/geo/info/seq.html
- The database of Genotypes and Phenotypes (dbGaP)
  - studies investigating the interaction of genotype and phenotype in humans.
  - all submissions that require controlled access must be submitted through dbGaP.
  - requires registration of the study and subjects prior to data submission.
  - raw data will be submitted to the protected SRA account.
  - For more information: https://www.ncbi.nlm.nih.gov/sra/docs/submitdbgap/

Bioinformatics file formats and tools

Format name	Data type	Tools
SRA	a raw data archive with per-base quality score	sra-tools
FASTA	a text file of reference genome sequence data	FASTA Tools
FASTQ	a text file of sequencing data with quality score	FastQC, FASTX-Toolkit, Seqtk, Samtools, Picard tools
SAM/BAM/CRAM	formats of sequence alignment data	Samtools, Picard tools
BCF/VCF/gVCF	a tab-delimited text file to store the variation calls	bcftools
BED (PEBED)	a tab-delimited text file to store the coordinates of genomic regions.	bedtools
GTF/GFF/GFF3	a tab-delimited text file to describe genes or other features	gff tools, GFF utilities (gffread, gffcompare)
MAF	a tab-delimited text file with aggregated mutation information from VCF files and are generated on a project-level.	MAFtools

For more information: https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-

THANK YOU FOR YOUR ATTENTION!
Questions for Part A?

Next: Bioinformatics File Formats and Tools - Part B

Bioinformatics File Formats and Tools - Part B

Bioinformatics File Formats - FASTA

- Description: a simple format of nucleic acid or protein sequences
- File extension: .fa, .fasta, .fsa

Example:

```
>ENST00000373020.9
AGTTGTGGACGCTCGTAAGTTTTCGGCAGTTTCCGGGGAGACTCGGGGACTCCGCGTCTCGCTCTCTGTGTTCCAATCGCCCGGTGCGGTGGTGCAGGG
TCTCGGGCTAGTCATGGCGTCCCCGTCTCGGAGACTGCAGACTAAACCAGTCATTACTTGTTTCAAGAGCGTTCTGCTAATCTACACTTTTATTTTCTG
GATCACTGGCGTTATCCTTCTTGCAGTTGGCATTTGGGGCAAGGTGAGCCTGGAGAATTACT
```

Line 1: Starts with '>'. Comment line with basic information about the sequence.

Other lines: Nucleic acid or protein sequences.

Tools for FASTA - FASTA Tools

Can be used in the forms of web-based or command lines
https://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html_ncbi/html/fasta/list.html

FASTA Composition finds the overall composition of sequences in a FASTA file.

Server Version 1.0. 2010 Dec 21
Version 1.0. 2010 Dec 21

FASTA Length finds the lengths of sequences in a FASTA file.

Server Version 1.0. 2010 Dec 21
Version 1.0. 2010 Dec 21

FASTA Match Regular Expression selects FASTA records whose deflines match a Perl regular expression. Matches can be case-sensitive or case-insensitive. The matches are stable, i.e., record order within the FASTA file is preserved.

Server Version 1.0. 2010 Dec 21
Version 1.0. 2010 Dec 21

FASTA Unique Sequences uniques the sequences in a FASTA file. It concatenates deflines that share sequences. It can also reverse the concatenation.

Server Version 1.0. 2010 Dec 21
Version 1.0. 2010 Dec 21

Bioinformatics File Formats - FASTQ

- Description: a standard format of genomic sequences data from sequencers.
- File extension: .fq, .fastq, .fq.gz (gzipped fastq file)

Bioinformatics File Formats - FASTQ

Example:

```
@A00423:133:H7CC2DSX2:1:1101:30834:1000 1:N:0:ACACGGTT+NGGTTCGA
CGGGCTCCTCGGGGTGCGCGGCTGGGGGTTCCCTCGCAGGGCCCGCCGGGGGCCCTCCGTCCCCCTAAGCGCAGACCCGGCGGCGTCCGC
C
+
FFFFFFFFFF:FFFFFFF,:F:FF:,F,FFFFFFF:FFFFFFFFFFFF::FFFFFF:FF:,F:FF:FFFFFF:,F:FF:FF:FFF,FFFF
F
@A00423:133:H7CC2DSX2:1:1101:3378:1016 1:N:0:ACACGGTT+NGGTTCGA
NTAGAAGCTTAGATTCAGGTTTGTTTGTGTAGGCAACACTAATCAGTGGTGTTGTATGCTTCCACCAGGAGGCACATAATGTCTCATATT
T
+
#FFFFFF,F:F:FF,FFF::FFFFFFF:FFFF::F,FFF,FF,FFF,,F:FFFFFFFF,FF,:FF,F:,F:,:FF::FFFFFFFFFFFF
F
```

Read1

Read2

Line 1: An identifier line start with @

```
@A00423:133:H7CC2DSX2:1:1101:30834:1000 1:N:0:ACACGGTT+NGGTTCGA
```

Sequencer ID

Run #

Flow cell

Lane

Tile

x-coordinate of the cluster

y-coordinate of the cluster

QC fail?

Read in pairs

Index/barcode

Line 2: The genomic sequence

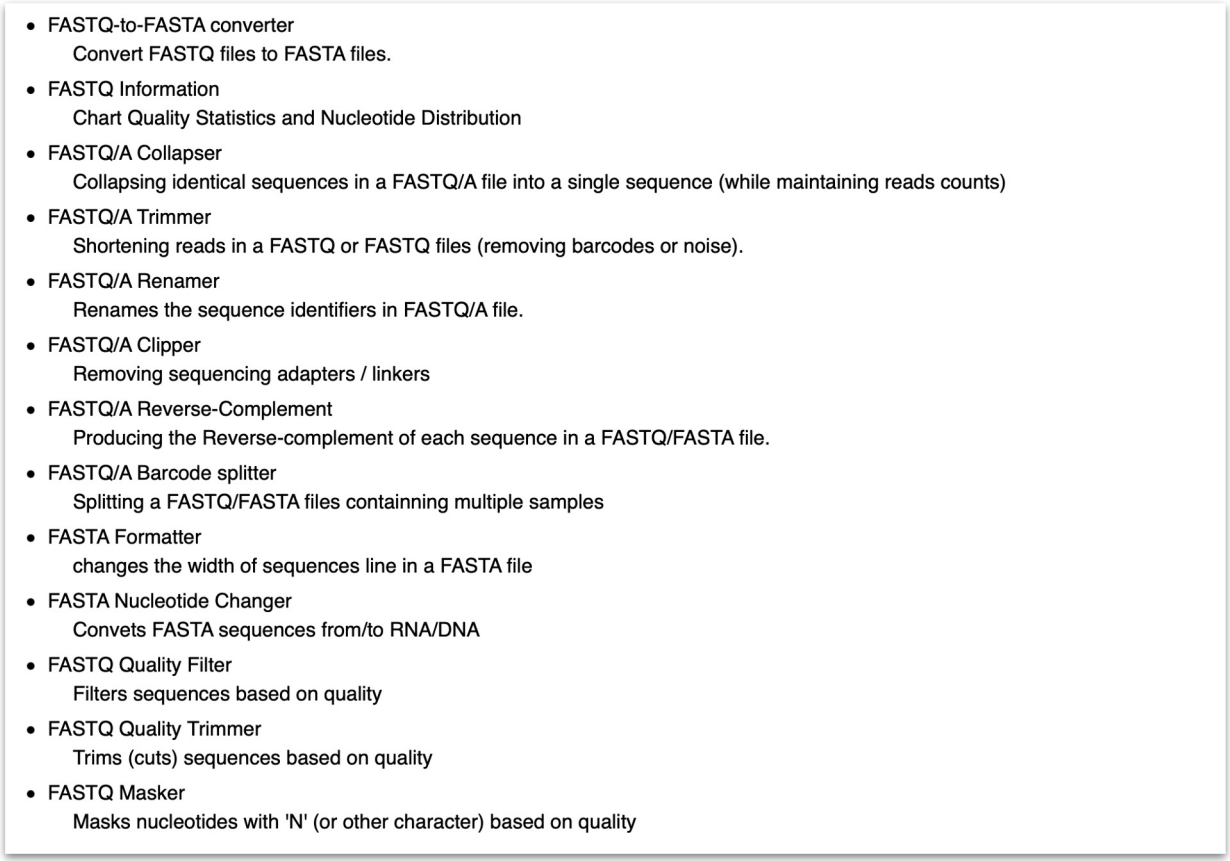
Line 3: + sign

Line 4: An ASCII-encoded, Phred-scaled base quality score.

Quality character	!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ				
ASCII Value	33	43	53	63	73
Base Quality (Q)	0	10	20	30	40

Tools for FASTQ - FASTX-Toolkit

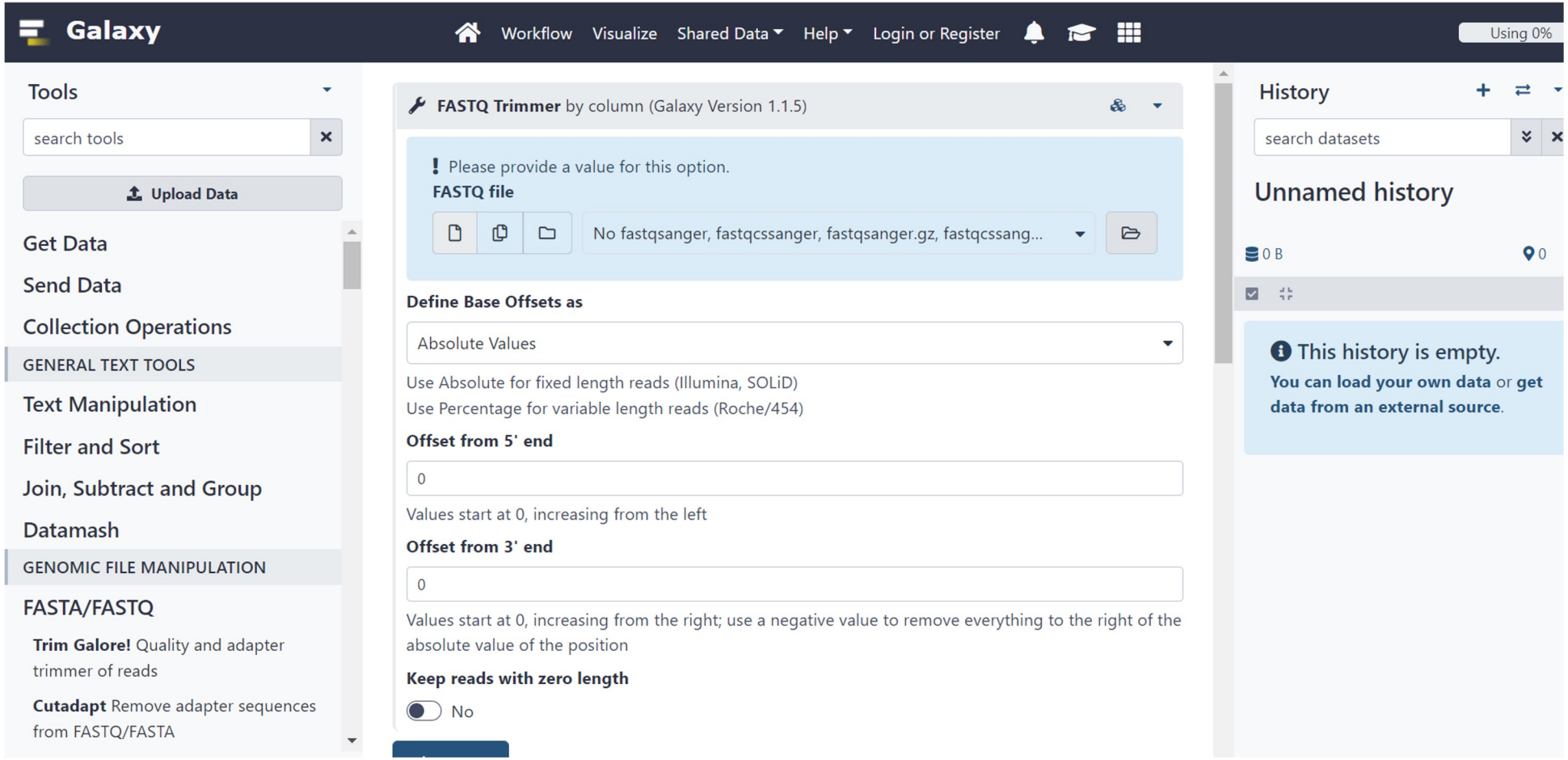
- Can be used as web-based (Galaxy) or through the command line
- Link to FASTX-Toolkit Information: http://hannonlab.cshl.edu/fastx_toolkit/index.html
- Usage examples:



- FASTQ-to-FASTA converter
  Convert FASTQ files to FASTA files.
- FASTQ Information
  Chart Quality Statistics and Nucleotide Distribution
- FASTQ/A Collapser
  Collapsing identical sequences in a FASTQ/A file into a single sequence (while maintaining reads counts)
- FASTQ/A Trimmer
  Shortening reads in a FASTQ or FASTQ files (removing barcodes or noise).
- FASTQ/A Renamer
  Renames the sequence identifiers in FASTQ/A file.
- FASTQ/A Clipper
  Removing sequencing adapters / linkers
- FASTQ/A Reverse-Complement
  Producing the Reverse-complement of each sequence in a FASTQ/FASTA file.
- FASTQ/A Barcode splitter
  Splitting a FASTQ/FASTA files containning multiple samples
- FASTA Formatter
  changes the width of sequences line in a FASTA file
- FASTA Nucleotide Changer
  Convets FASTA sequences from/to RNA/DNA
- FASTQ Quality Filter
  Filters sequences based on quality
- FASTQ Quality Trimmer
  Trims (cuts) sequences based on quality
- FASTQ Masker
  Masks nucleotides with 'N' (or other character) based on quality

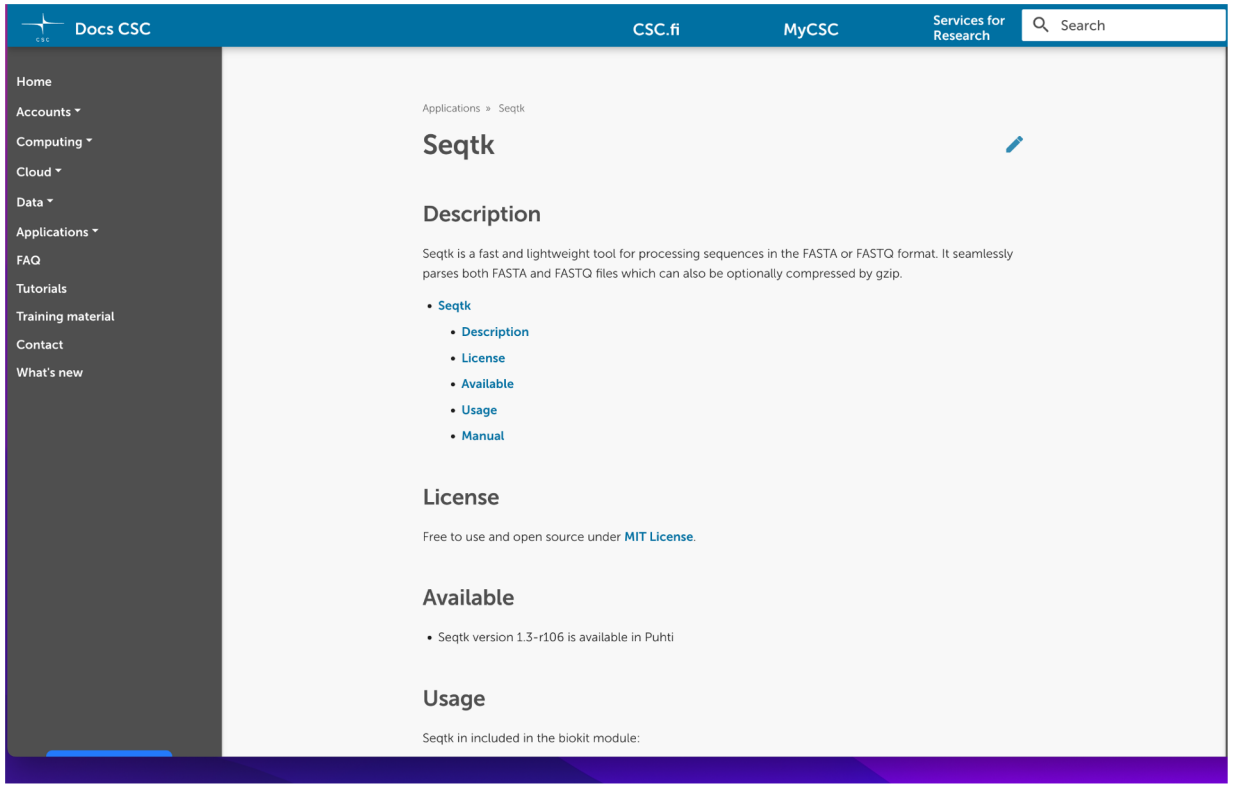
Tools for FASTQ - FASTX-Toolkit

Galaxy (https://usegalaxy.org/) is an open source, web-based framework for experimental and computational biologists.



Tools for FASTQ - Seqtk

- Usage examples: seqtk COMMAND <arguments>
- For more information: https://docs.csc.fi/apps/seqtk/



Command	Function
seq	common transformation of FASTA/Q
comp	get the nucleotide composition of FASTA/Q
sample	subsample sequences
subseq	extract subsequences from FASTA/Q
fqchk	fastq QC (base/quality summary)
mergepe	interleave two PE FASTA/Q files
trimfq	trim FASTQ using the Phred algorithm
hety	regional heterozygosity
gc	identify high- or low-GC regions
mutfa	point mutate FASTA at specified positions
mergefa	merge two FASTA/Q files
famask	apply a X-coded FASTA to a source FASTA
dropse	drop unpaired from interleaved PE FASTA/Q
rename	rename sequence names
randbase	choose a random base from hets
cutN	cut sequence at long N
listhet	extract the position of each het

Bioinformatics File Formats - SAM (Sequence Alignment Map)/BAM/CRAM

- Description: standard formats of sequence alignment files. A BAM file is a compressed binary file of a SAM file. CRAM is a new way to highly compress a SAM file.
- File extension: .bam, .sam., .cram
- Detailed information: https://samtools.github.io/hts-specs/

Bioinformatic File Formats - SAM (Sequence Alignment Map)/BAM/CRAM

Header section

@HD: The version of the SAM format, how the alignments are sorted.

@SQ: A sequence in the reference genome aligned to.

@PG: The program that was used to produced the SAM file, starting from FASTQ.

@RG: ID: Read groups, which are collections of reads that can all be assumed to be the same sample and prepared with the same pipeline. SM: sample. PL: platform.

```
@HD     VN:1.4  SO:coordinate
@SQ     SN:chr1 LN:248956422
@SQ     SN:chr2 LN:242193529
@SQ     SN:chr3 LN:198295559
@SQ     SN:chr4 LN:190214555
……
@PG     ID:STAR PN:STAR VN:2.7.3a       CL:/usr/local/apps/STAR/2.7.3a/bin/STAR   --runThreadN 4   --
genomeDir /data/zhaow5/hg38   --readFilesIn /data/SC623212_R1.fastq.gz   /data/SC623212_R2.fastq.gz      --
readFilesCommand zcat      --outSAMtype BAM   SortedByCoordinate      --outSAMattributes NH   HI   NM   MD
--outSAMunmapped Within      --outSAMattrRGline ID:SC623212   SM:SC623212   PL:ILLUMINA          --
sjdbGTFfile /fdb/GENCODE/Gencode_human/release_35/gencode.v35.annotation.gtf   --sjdbOverhang 100
 --sjdbScore 2   --quantMode TranscriptomeSAM   GeneCounts
@RG     ID:SC623212     SM:SC623212     PL:ILLUMINA
```

Bioinformatics File Formats - SAM (Sequence Alignment Map)/BAM/CRAM

Alignment/Body section: TAB-delimited lines with the following mandatory fields. If the information is unavailable, the field's value will be a placeholder ('0' or '*').

Example:

```
A00423:107:HNNLMDSXY:1:1101:22607:1000  355     ENST00000458537.7       1479    1       149M    =       1629    299
AAACTGGCAGTGCCTTTGAAGTGGAGCATGAATGCCT
AGGCAAATGTCAGGGTTTATTTGCACCCCAGTTTTATGTTCAGCCTGATGCTCCGTGTATTCAATGTCTGGAGTGTTGTGGAATGTTTGCACCCCAGACGTTTGTGATGCAT
FFFFFFFFFFFFFF:FFF:FFF:FFFFFFFFFFFF
FFFFF:FFFFFFFFFFFFFFFFFFF:FF:FFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFF:F:FFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFF:FF       NH:i:4  HI:i:1
RG:Z:SC623212
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Bioinformatics File Formats - SAM (Sequence Alignment Map)/BAM/CRAM

SAM Flag: Combination of bitwise FLAGs indicating alignment information.

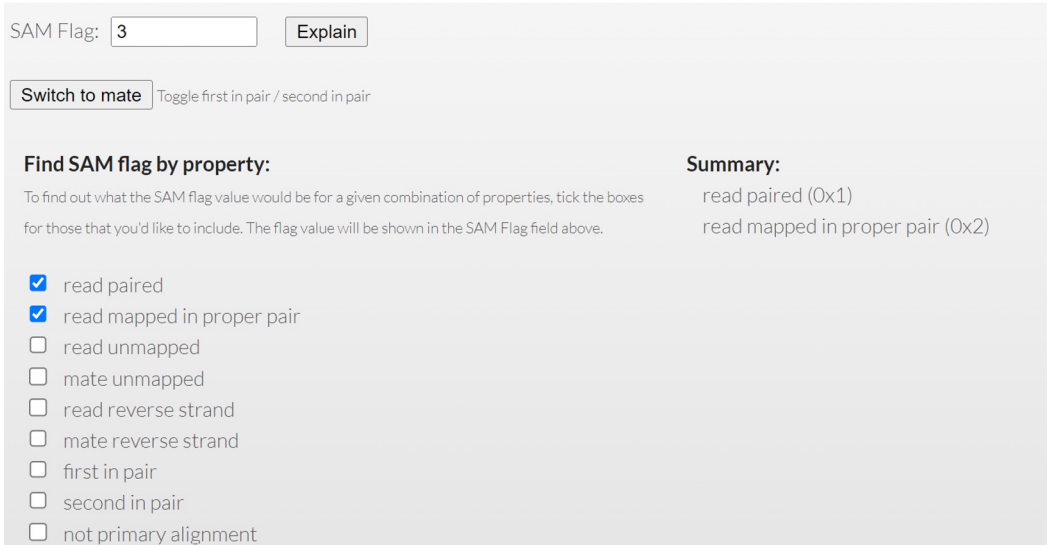
Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Bioinformatics File Formats - SAM (Sequence Alignment Map)/BAM/CRAM

An online utility to explain SAM flags: https://broadinstitute.github.io/picard/explain-flags.html

BAM flags could be used for alignment QC.

Example of 'good' alignment (flag=3)



Example of 'bad' alignment (flag=268)

Filter BAM files using the command: `samtools view -f 3 -F 268 input.bam`

Bioinformatics File Formats - SAM (Sequence Alignment Map)/BAM/CRAM

CIGAR string: encode an entire alignment with operators and numbers

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Example

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

CIGAR: 3M 1I 3M 1D 5M

https://genome.sph.umich.edu/wiki/SAM#What_is_a_CIGAR.3F

Tools for SAM/BAM/CRAM - Samtools

- Reading/writing/editing/indexing/viewing SAM/BAM/CRAM formats.
- Usage example: samtools COMMAND [options] in.sam|in.bam
  - COMMAND: view, index, sort, flagstat, stats, mpileup, merge, split, fastq/a …
- Example Functionalities:
  - View BAM/SAM files
  - Select mapped/unmapped reads, or reads mapping to specific genomic regions
  - Sort mapped reads by genomic coordinates (and index) or names
  - Does a full pass through the input SAM/BAM file to calculate statistics
  - Merge/split multiple files
  - Convert BAM/CRAM format to/from FASTQ or FASTA format
  - Edit header lines.

For more information: http://www.htslib.org/doc/samtools.html

Tools for SAM/BAM/CRAM - Picard Tools

- Reading/writing/editing/indexing/viewing SAM/BAM/CRAM and VCF formats.
- Usage example: java -jar picard.jar COMMAND OPTION1=value1 OPTION2=value2
  - COMMAND: ViewSam. MergeBamAlignment, SortSam, CollectRnaSeqMetrics, CollectAlignmentSummaryMetrics, EstimateLibraryComplexity, FilterVcf …
- Example Functionalities:
  - Produces a summary of alignment metrics detailing the quality of read alignments.
  - Collects information on base distribution/GC bias/insert size/quality yield, coverage, etc
  - Produces the RNA alignment metrics, e.g. # and % reads mapping to UTR/exon/intron/intergenic regions.
  - Collects variant calling information from VCF files
  - Estimates complexity of libraries.
  - Shares many functions with Samtools and BCFtools

For more information: https://broadinstitute.github.io/picard/

Bioinformatics File Formats - BCF/VCF (Variant Calling Format)

- Description: formats to store information of variation (SNV, indel and SV) calls. BCF is the binary (compressed) version of VCF.
- File extension: .vcf, .bcf
- Contains header and body sections

Bioinformatics File Formats - BCF/VCF (Variant Calling Format)

- Body section:
  - TAB-delimited lines with 8 mandatory fields to define variants, and genotype field to define genotypes and quantities of individual samples.
  - The formats of the genotype data were defined by the field FORMAT.
  - Tags in the fields INFO, FILTER, FORMAT were defined by the header section.

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/

Bioinformatics File Formats - BCF/VCF (Variant Calling Format)

- Header section:
  - file meta-information
  - begins with "##", defines tags (corresponding to the body section).
  - Common **fields** (corresponding to the body section): INFO, FILTER, FORMAT.
  - Undefined tags in the header could cause errors (e.g. the following example defined tags NS, DP, AF, AA, DB, H2, q10, S50, GT, GQ, DP and HQ. 'PASS' is a special tag in the FILTER field.)

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/

Tools for BCF/VCF - BCFtools

- Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
- Usage example: bcftools COMMAND [options] *file(s)*
  - COMMAND: view, call, annotate, filter, query, stats, sort …
- Example Functionalities:
  - Annotates VCF files
  - Calls SNV/indels
  - Calls copy number variation
  - Filters VCF/BCF files
  - Sorts VCF/BCF files

For more information: http://samtools.github.io/bcftools/

Bioinformatics File Formats - MAF (Mutation Annotation Format)

- Description: a tab-delimited text file with aggregated mutation information from VCF Files and are generated on a project-level.
- File extension: .maf
- Example:

Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2	dbSNP_RS	dbSNP_Val_Status	Tumor_Sample_Barcode	Matched_Norm_Sample_Barcode
PRAMEF14	729528	BI	GRCh38	chr1	13344364	13344364	+	Silent	SNP	C	C	G	novel		TCGA-55-8615-0	TCGA-55-8615-10A
ACTL8	81569	BI	GRCh38	chr1	17826321	17826321	+	Missense_Mutation	SNP	C	C	A			TCGA-55-8615-0	TCGA-55-8615-10A
TMCO4	255104	BI	GRCh38	chr1	19746510	19746510	+	Missense_Mutation	SNP	C	C	T			TCGA-55-8615-0	TCGA-55-8615-10A
OR2T12	127064	BI	GRCh38	chr1	248295475	248295475	+	Missense_Mutation	SNP	G	G	C	rs749730602		TCGA-55-8615-0	TCGA-55-8615-10A
CMYA5	202333	BI	GRCh38	chr5	79734231	79734231	+	Missense_Mutation	SNP	A	A	T			TCGA-55-8615-0	TCGA-55-8615-10A
ERLEC1	27248	BI	GRCh38	chr2	53794372	53794372	+	Missense_Mutation	SNP	G	G	C			TCGA-55-8615-0	TCGA-55-8615-10A
MOGS	7841	BI	GRCh38	chr2	74461426	74461426	+	Missense_Mutation	SNP	C	C	T	rs199724485		TCGA-55-8615-0	TCGA-55-8615-10A
TCF7L1	83439	BI	GRCh38	chr2	85302600	85302600	+	Missense_Mutation	SNP	G	G	C	rs1486052719		TCGA-55-8615-0	TCGA-55-8615-10A
TTN	7273	BI	GRCh38	chr2	178545553	178545553	+	Missense_Mutation	SNP	G	G	T			TCGA-55-8615-0	TCGA-55-8615-10A
MAP2	4133	BI	GRCh38	chr2	209694182	209694182	+	Missense_Mutation	SNP	G	G	A	rs761466743		TCGA-55-8615-0	TCGA-55-8615-10A
MAPKAPK3	7867	BI	GRCh38	chr3	50646777	50646777	+	Silent	SNP	C	C	T	rs1473215021		TCGA-55-8615-0	TCGA-55-8615-10A
MYH15	22989	BI	GRCh38	chr3	108500201	108500201	+	Missense_Mutation	SNP	G	G	C			TCGA-55-8615-0	TCGA-55-8615-10A

For more information: https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/#protected-maf-file-structure

Tools for MAF - maftools

- R package to summarize, analyze, annotate and visualize MAF files
- Example Functionalities:
  - Annotates variants
  - Detects interaction of somatic variants
  - Detecting cancer driver genes
  - Survival analysis
  - Clinical enrichment analysis
  - Visualization (Oncoplot, lollipop plot, rainfall plot, plot maf summary, etc)

For more information:

https://bioconductor.org/packages/release/bioc/vignettes/maftools/inst/doc/maftools.html

Bioinformatics File Formats - BED

- Description: a file format to store the coordinates of genomic regions.
- File extension: .bed
- Can be used to define the annotation tracks in the UCSC genome browser.
- TAB-delimited text file, with three required fields and six optional fields.
- The start position is 0-based (the first base in a chromosome is number 0). The end position is 1-based (1-based coordinate is also used in other formats, e.g. GTF).

Example:

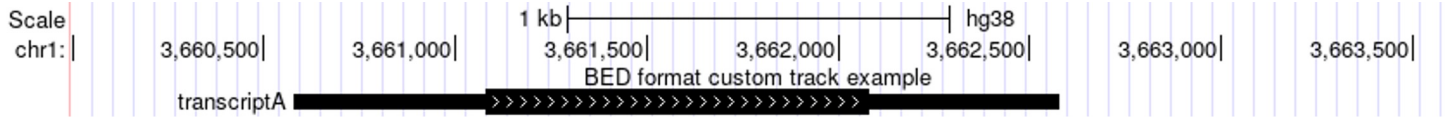
```
chr1    3204713 3206713 uc007aet.1: 0  +  3204713  3206713  255,0,0
chr1    3647985 3649985 uc007aev.1: 0  +  3647985  3649985  0,0,255
chr1    3660579 3662579 transcriptA 0  +  3660579  3662579  0,0,255
```

Genomic coordinates

Name Strand Parameters for visualization in the UCSC genome browser

Required fields

Optional fields



Common Bioinformatics Tools - bedtools

- Intersect, merge, count, complement, and shuffle genomic intervals from multiple files in formats such as BAM, BED, GFF/GTF
- Can be used as web-based (Galaxy) or through the command line, or by calling an R package 'valr'.
- Usage example: bedtools COMMAND [options]
- COMMAND: intersect, window, closest, coverage, merge, genomecov, subtract, shift, …
- Example:
  - Compares two or more files to identify regions in the genome where the features in two files overlap (or do not overlap).
  - Combines genomic features with overlapping regions into a single one.
  - Computes the coverage of aligned sequences in 'windows' spanning the genome.
  - Identifies intervals in the genome that are not covered by regions in the input file.

For more information: https://bedtools.readthedocs.io/en/latest/

Bioinformatic file formats - GTF/GFF (General Feature Format)

- Description: Formats to describe genes or other features. GTF and GFF have same formats. GFF can be used to describe any kinds of features, whereas GTF is primarily for genes or transcripts.
- File extension: .gtf, .gff, .gff2, .gff3.
- TAB-delimited text file with 9 required fields. Missing information in all but the last field is denoted with the placeholder '.'.

```
##gff-version 3
chr12   HAVANA  stop_codon      25215441        25215443        .       -       0       ID=stop_codon:ENST00000256078.10;Parent=
ENST00000256078.10;gene_id=ENSG00000133703.14;transcript_id=ENST00000256078.10;gene_type=protein_coding;gene_name=KRAS;transcript_
type=protein_coding;transcript_name=KRAS-
201;exon_number=5;exon_id=ENSE00001189807.5;level=2;protein_id=ENSP00000256078.5;transcript_support_level=1;hgnc_id=HGNC:6407;tag=
RNA_Seq_supported_partial,basic,appris_principal_4,CCDS;ccdsid=CCDS8703.1;havana_gene=OTTHUMG00000171193.4;havana_transcript=OTTHU
MT00000412232.4
```

source

Score
(sequence similarity or
gene prediction)

Frame
(base in codon)

Attributes

Tools for GTF/GFF - GffCompare

- Tools to compare, merge, annotate and estimate accuracy of one or more GFF files
- Usage example: gffcompare [OPTIONS] -i input_gtf_list | input1.gtf input2.gtf …
- Example:
  - Evaluate transcript discovery accuracy
  - Compare query (‘novel’) transcripts versus a set of reference annotation

For more information: http://ccb.jhu.edu/software/stringtie/gffcompare.shtml

THANK YOU FOR YOUR ATTENTION!
Questions for Part B?

Next: Practical session 1 (10:45am)