Retrieving, accessing, manipulating files

Downloading sequences and reads from various sources:

- GitHub (“raw” link)
- Ensembl
- NCBI
- UCSC

unzipping and untarring

Using HPC tools: modules

module avail
module load <program>
module list
module remove <program>

Using HPC tools: conda environments

Step 1: module load anaconda3/2024.06
Step 2: source activate <path to specific environment>

Programs for sequence analysis:

angsd
bamtools
bcftools
bowtie2
cutadapt
emboss
fastqc
samtools
STAR

File formats and their use:

Genbank (standard record you see on NCBI)
fasta
fastq
SAM/BAM
VCF
MAF

Basic sequence analyses

Use emboss to explore various aspects of your sequence data
Determine GC% of a sequence or file of sequences
Determine how many reads or fasta sequences are in a file
Determine quality of reads by looking for stretches of N's (uncalled bases)
Determine how many occurrences there are of a specific motif

More sophisticated
Interpret multiple indicators of read quality using fastqc
Index a genome
Trim sequencing reads
Align (map) reads to reference genome
Interpret variant-calling and genotype likelihood results

Three ways to run commands:

- on command line
- in bash script
- in batch script (run as "job")

Writing scripts
Bash script heading:

BATCH script heading: (see Batch_headings.md in Explore module on Github)