Core NGS Resources

A healthy taste of resources available, specifically for this course - not a comprehensive catalog.

Linux/TACC

- Linux fundamentals on this wiki
- Online tutorials:
    - Ryan's Linux Tutorial: http://ryanstutorials.net/linuxtutorial/
    - Unix bootcamp for biologists: http://korflab.ucdavis.edu/bootcamp.html
    - Unix primer (longer version) for biologists:
        - http://korflab.ucdavis.edu/Unix_and_Perl/unix_and_perl_v2.3.4.pdf

Community Resources

- UCSC Genome Browser - visualize and download NGS data (see more below)
- Broad Institute Integrated Genomcs Viewer (**IGV**)
    - especially good for visualizing **BAM** file details
- Introduction to Sequence analysis in the Amazon EC2 cloud
    - where you can "rent" Linux machines (useful if you don't have access to TACC)
- Galaxy website for online sequencing data analysis
- SEQAnwers forum - many NGS sequencing questions answered here
    - A funny SEQAnwers post about biologists starting to analyze NGS data: http://seqanswers.com/forums/showthread.php?t=4589

Sequencing Technologies

- Overviews
    - Wikipedia overview of NGS technologies
    - Broad Center GA Boot Camp
    - Early paper comparing NGS technologies (Liu et al., 2012)
- Technology intros
    - Illumina (Solexa) – most common "short" (< 300 bp) read sequencing
        - Short Illumina video
        - Longer Illumina video
        - 2016 Sequencing Technologies talk
    - Newer ***single molecule*** sequencing
        - Oxford Nanopore
        - PacBio SMRT system
    - ***Single cell*** sequencing
        - 10x Genomics platforms
    - Older technologies (less common now)
        - Life Technologies SOLiD (short reads in "colorspace")
        - Roche/454 – long (multi-Kb) reads often used in assemblies

FASTQ analysis/manipulation/QC

- Wikipedia FASTQ format page
- Illumina library construction on GSAF user wiki - useful for contaminant detection or adapter removal
- **FastQC** from Babraham Bioinformatics – http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
    - produces nice quality report for **FASTQ** files
- **MultiQC** – http://multiqc.info/
    - A great tool for consolidating QC multiple QC reports into one HTML page
    - Anna's Byte Club tutorial on using **MultiQC** – https://wikis.utexas.edu/display/bioiteam/Using+MultiQC
- **cutadapt** – https://cutadapt.readthedocs.io/en/stable/
    - An excellent command line tool for adapter sequence removal
    - Good support for trimming paired-end datasets

    - Script that handles the details of paired-end read trimming
        - `/work2/projects/BioITeam/common/script/trim_adapters.sh`
- **trimmomatic** – http://www.usadellab.org/cms/?page=trimmomatic
    - Supports trimming paired-end datasets.
- **fastx toolkit** – http://hannonlab.cshl.edu/fastx_toolkit/
    - Suite of command line tools for **FASTQ** and **FASTA** analysis and manipulation
    - Good for hard clipping, **FASTA** file manipulations
    - Documentation at: http://hannonlab.cshl.edu/fastx_toolkit/commandline.html
- **seqtk** – https://github.com/lh3/seqtk
    - Suite of command line tools for **FASTQ** and **FASTA** analysis and manipulation

Reference genomes

- Gencode – https://www.gencodegenes.org/
    - reference genomes, transcriptomes and high-quality annotations for human and mouse
- UCSC downloads – http://hgdownload.cse.ucsc.edu/downloads.html
    - reference genomes, transcriptomes and high-quality annotations for many eukaryotes
- Ensembl downloads – http://ftp.ensembl.org/pub
    - reference genomes, transcriptomes and high-quality annotations for many eukaryotes
- NCBI
    - RefSeq – https://www.ncbi.nlm.nih.gov/refseq/
        - well curated genome, transcriptome sequences
    - GenBank – https://www.ncbi.nlm.nih.gov/genbank/
        - public repository for sequence data, especially for prokaryotic genomes
        - not curated
- Reference genome vocabulary – https://software.broadinstitute.org/gatk/documentation/article?id=7857
    - excellent introduction to the types of genome references and the vocabulary used to describe them
        - aimed at higher eukaryotes but vocabulary useful nonetheless
- GATK blog describing ALT contigs in GRCh38 – https://software.broadinstitute.org/gatk/blog?id=8180
- Support for mapping to ALT contigs containing variants
    - **bwa mem** + **bwakit** by Heng-Li – https://github.com/lh3/bwa/blob/master/README-alt.md

Basic alignment and aligners

- File formats
    - input: **FASTQ** format
    - output: the **SAM** (Sequence Alignment Map) format specification
        - SAM1.pdf – header fields, body fields, flag definitions
        - https://github.com/samtools/hts-specs/blob/master/SAMtags.pdf – tag fields
- Aligners
    - **bwa** (Burrows-Wheeler Aligner) by Heng Li – http://bio-bwa.sourceforge.net/
        - fast, sensitive and easy to use
    - **bowtie2** – http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
        - fast, sensitive and extremely configurable
- Comparison of different aligners
    - by Heng Li, developer of **bwa**, **samtools**, and *many* other bioinformatics tools
- The BioITeam has some TACC-aware alignment scripts you might find useful:
    - **bwa** alignment
        - `/work/projects/BioITeam/common/script/align_bwa_illumina.sh`
    - **bowtie2** alignment
        - `/work/projects/BioITeam/common/script/align_bowtie2_illumina.sh`
    - merging sorted **BAM** files (read-group aware)
        - `/work/projects/BioITeam/common/script/merge_sorted_bams.sh`
    - **kallisto** pseudo-alignment to annotated transcripts
        - `/work/projects/BioITeam/common/script/run_kallisto.sh`
    - also available on many BRCF pods under **/mnt/bioi/script**.
        - many pre-built references also available in **/mnt/bioi/ref_genome**
    - email or come talk to Anna if you have questions or problems

Transcriptome-aware aligners

- **HISAT2** – https://daehwankimlab.github.io/hisat2/
    - fast, with support for alignment to single and "population" of genomes
    - paper: http://www.nature.com/nprot/journal/v11/n9/full/nprot.2016.095.html
- **STAR** (**S**pliced **T**ranscripts **A**lignment to a **R**eference) – ultra-fast RNA-seq aligner
    - releases: https://github.com/alexdobin/STAR/releases
    - paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/
    - STAR manual: https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf
- **TopHat** - http://ccb.jhu.edu/software/tophat/index.shtml
    - exon-aware sequence alignment (uses **bowtie2**/**bowtie**)
- **kallisto** - https://pachterlab.github.io/kallisto/about
    - ultra-fast RNA-seq ***pseudoaligner*** that goes straight from **FASTQ** to estimated transcript abundances

Alignment analysis

- **SAM** (Sequence Alignment Map) format specification (SAM1.pdf)
  - Translate **SAM** file flags web calculator: http://broadinstitute.github.io/picard/explain-flags.html
    - type in a decimal number to see which flags are set
- **samtools** – by Heng Li
  - **SAM/BAM** conversion, flag filtering, sorting, indexing, duplicate filtering
  - older **0.1.xx** versions: http://samtools.sourceforge.net/
  - newer **1.3+** versions: http://www.htslib.org/
- **Picard** toolkit – http://broadinstitute.github.io/picard/
  - **SAM/BAM** utilities that are read-group aware
  - especially **MarkDuplicates** for flagging duplicate alignments
    - http://broadinstitute.github.io/picard/command-line-overview.html#MarkDuplicates
- **bedtools** – http://bedtools.readthedocs.org/en/latest/
  - All sub-commands: http://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html
  - Swiss army knife for all manner of common **BED, BAM, VCF, GFF/GTF** file manipulation.
  - See BEDTools Overview for some common use cases.
  - Available in the TACC module system
- **RNA-seq** QC, metrics & plotting tools:
  - **RSeQC** – http://rseqc.sourceforge.net/
  - **RNA-SeQC** (Broad Institute) –
    - https://software.broadinstitute.org/cancer/cga/rna-seqc
    - https://software.broadinstitute.org/cancer/cga/rnaseqc_run
  - **RNA-QC-Chain** – http://bioinfo.single-cell.cn/rna-qc-chain.html

File formats and conversion

- **SAM** format specification – http://samtools.github.io/hts-specs/SAMv1.pdf
  - crucial for performing format conversions, of which ChIP-seq analysis can have many
- HTS format specifications – http://samtools.github.io/hts-specs/
  - clearinghouse page for a number of NGS formats (**SAM**, **CRAM**, **VCF**, **BCF**, etc.)
- Genome browser file formats – http://genome.ucsc.edu/FAQ/FAQformat.html
  - **BED**, **bedGraph**, **narrowPeak** and many more
- **SRA** (Sequence Read Archive) from NCBI
  - SRA search home page
  - **SRA Toolkit**
    - NCBI documentation
    - SRA toolkit downloads
- BioITeam script for converting **GTF**/**GFF3** files to **BED** format
  - `/work/projects/BioITeam/common/script/gtf_to_bed.pl`
- **UCSC file format conversion scripts** - useful for getting to/from **WIG** and **BED** to corresponding binary formats
  - Make sure you download the correct scripts for your operating system!
  - Also available as a **BioContainers** module

UCSC Genome Browser

- Main **UCSC Genome Browser** web site
  - File formats - **BED** format especially is widely used
  - Table browser - Browse and download data in different formats
  - ENCODE data downloads at UCSC - useful for getting data to work with
  - Beta Test browser site - most up-to-date datasets and features; can be buggy

RNAseq/Transcriptome analysis

- General RNA-seq Differential Gene Expression (DGE) analysis workflow from **R**'s **Bioconductor**:
  - https://www.bioconductor.org/help/workflows/rnaseqGene/
- Gene quantification from **BAM**/**BED** file reads
  - **featureCounts** (part of the **Subread** package) – http://subread.sourceforge.net/
  - **HTSeq** – https://htseq.readthedocs.io/en/master/
- **HISAT2**, **StringTie, BallGown** suite – https://ccb.jhu.edu/software/hisat2/index.shtml
  - transcriptome-aware alignment & quantification from the Johns Hopkins group who brought you the Tuxedo pipeline – but much faster!
  - paper: http://www.nature.com/nprot/journal/v11/n9/full/nprot.2016.095.html
- **DESeq2** – R Bioconductor package for DGE
  - **DESeq** (version 1) documentation:
    - https://bioconductor.org/packages/release/bioc/vignettes/DESeq/inst/doc/DESeq.pdf
    - while **DESeq2** is more sophisticated, reading the original documentation is a better introduction to concepts
  - **DESeq2** documentation:
    - http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html
- **kallisto** – https://pachterlab.github.io/kallisto/
  - RNA-seq ***pseudoaligner*** that goes straight from **FASTQ** to estimated transcript abundances
    - blindingly fast – but only to transcriptome
  - companion quantification tool is **sleuth** – http://pachterlab.github.io/sleuth/about
  - overview presentation – 2015-10-21-Kallisto.Anna.pdf
- The **Tuxedo** pipeline: RNAseq with **tophat**/**cufflinks**
  - one of the first tool suites for transcriptome-aware RNA-seq alignment and quantification

        - rarely used now, as other tools are much faster & more accurate
    - RNAseq analysis protocol article in Nature Protocols
    - **TopHat** - http://ccb.jhu.edu/software/tophat/index.shtml
        - exon-aware sequence alignment (uses **bowtie2**/**bowtie**)
        - resource bundles for selected organisms (**GFF** annotations, pre-built **bowtie2** references, etc.)
    - **cuffquant**, **cuffnorm**, **cufflinks** – http://cole-trapnell-lab.github.io/cufflinks/manual/
        - transcript quantification, normalization, differential expression
- Dhivya Arasappan's Introduction to RNA Seq CBRS 2021 summer school course

Variant calling

- Broad institute **GATK** (**G**enome **A**nalysis **T**ool **K**it) – https://software.broadinstitute.org/gatk/documentation/
    - complex but powerful
    - used by TCGA (The Cancer Genome Atlas), 1000 Genomes
- File formats
    - **VCF** (Variant Call Format) v4.0 - initially developed by 1000 Genomes project
    - **MAF** (Mutation Annotation Format) – developed by The Cancer Genome Atlas (TCGA)
- The International Genome Sample Resource – follow-on to the 1000 Genomes project
    - catalog of human genetic variants
- Dan Deatherage's Genome Variant Analysis CBRS 2021 summer school course

Genome Annotation

- **GO** – http://geneontology.org/
    - The **G**ene **O**ntology resource, a large source of information on the functions of genes
- **GOrilla** – http://cbl-gorilla.cs.technion.ac.il/
    - **G**ene **O**ntology en**RI**chment ana**L**ysis and visua**L**iz**A**tion tool
- **GSEA** – https://www.gsea-msigdb.org
    - **G**ene **S**et **E**nrichment **A**nalysis
- **DAVID** – https://david.ncifcrf.gov/
    - Functional annotation from user-supplied gene lists
- **GREAT** – http://bejerano.stanford.edu/great/public/html/splash.php
    - **G**enomic **R**egions **E**nrichment of **A**nnotations **T**ool
    - Takes bed files as input and outputs enriched genes, GO-terms, motifs, etc.
        - human, mouse, zebrafish
- **MEME-suite** – http://meme-suite.org/
    - A motif identification and discovery tool. Works with most species.
    - Takes **FASTA** files as input
        - filter your **BAM/BED** files to get the regions of interest
        - then convert to **FASTA** using **bedtools bamtofastq**.