JOHNS HOPKINS
SCHOOL of MEDICINE

Quantitative Molecular Neurogenomics 2023
Module 5 - Data Sources, Acquisition, & QC

ME.440.825
Loyal Goff - Course Director
Genevieve Stein-O'Brien - Instructor
Richard Sriworarat - TA
Kyla Woyshner - TA
Jeanette Johnson - TA

October 2nd, 2023

JOHNS HOPKINS
SCHOOL of MEDICINE

PUBLIC NGS DATA SOURCES

NGS Data Sharing is required for NIH-funded research

- NIH Issued Genomic Data Sharing (GDS) policy, August 2014
  - "…applies to all NIH-funded research (e.g., grants, contracts, and intramural research) that generates large-scale human or non-human genomic data."
  - "NIH strongly encourages the use of established repositories to the extent possible for preserving and sharing scientific data."
  - "Shared scientific data should be made accessible as soon as possible, and no later than the time of an associated publication"

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

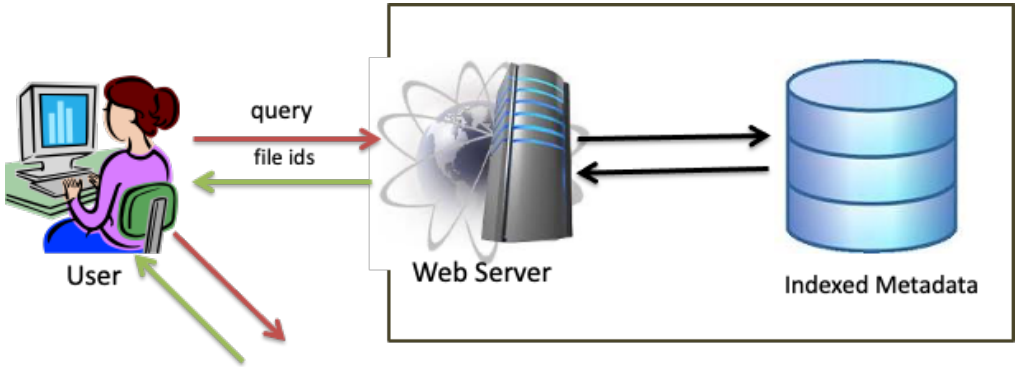
Common NGS Data sources/archives

- Gene Expression Omnibus (GEO)
  - Public functional genomics data repository
  - Storage point for ***processed*** data files
- Sequence Read Archive (SRA)
  - Stores ***raw*** sequencing reads and associated metadata
- EMBL-EBI
  - European Bioinformatics Institute
- DDBJ
  - DNA Data Bank of Japan

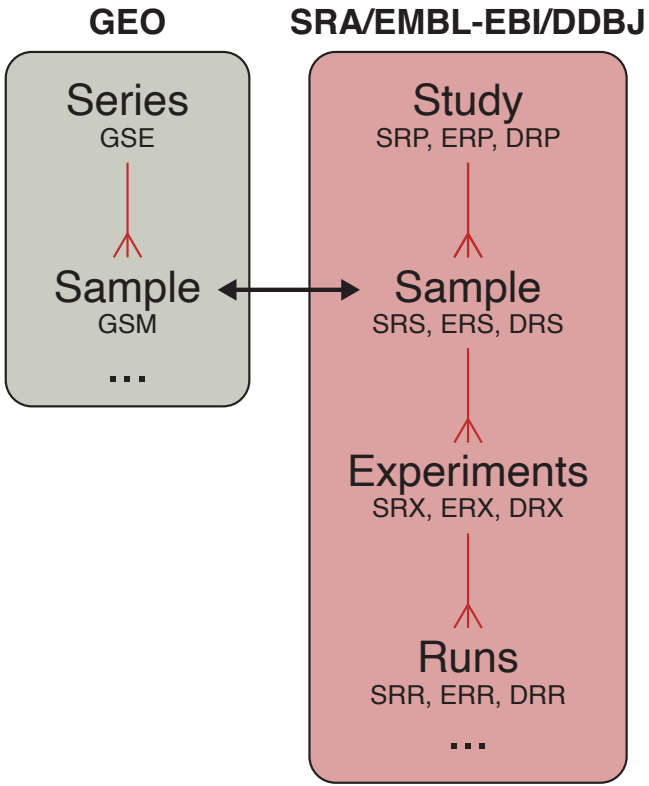
SRA - ncbi.nlm.nih.gov/sra/

- Archives raw sequencing data and alignment information from high-throughput sequencing platforms

- Data are ingested from direct submissions to GEO
  - Processed into ‘SRA’ format for efficient storage



Navigating Common NGS Primary Data Archives



- Series/Studies:
  - Single Accession for all downstream elements of a project
- Samples:
  - A distinct biological sample
  - Usually 1:1 across Data sources
- Experiment:
  - A unique sequencing library and sequencing method for a specific sample (sample + assay = library/experiment)
- Run:
  - The execution of an experiment on a sequencing platform and associated data (output)
  - Different runs of the same experiment can *usually* be aggregated together
    - (e.g. same library sequenced twice or on separate flow cells)

What am I looking for…

- Raw files for a given experiment can/should include:
  - Raw sequencing output (.fastq)
  - Sample information/metadata
    - Including how the raw fastq files relate to samples/replicates
  - Processed data files
    - .csv/.tsv/.txt/.mtx
    - These files store the ‘processed values’ that the authors used for downstream analyses.
  - Any specialized accessory data
    - e.g. Custom references or annotations

DATA ACQUISITION

Tools and Resources

Neuroscience

Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons

Mark S Cembrowski, Lihua Wang, Ken Sugino, Brenda C Shields, Nelson Spruston

Janelia Research Campus, Howard Hughes Medical Institute, United States

Apr 26, 2016 · https://doi.org/10.7554/eLife.14997

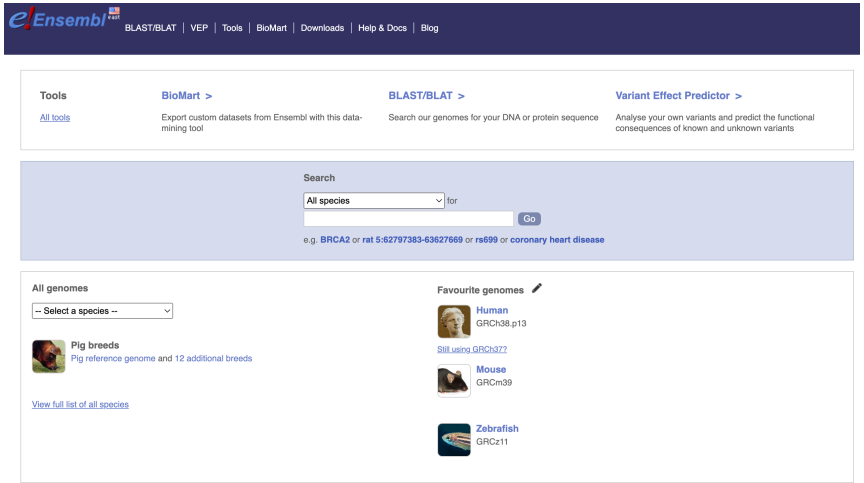
https://elifesciences.org/articles/14997

JOHNS HOPKINS
SCHOOL of MEDICINE

ACCESSORY DATA SOURCES

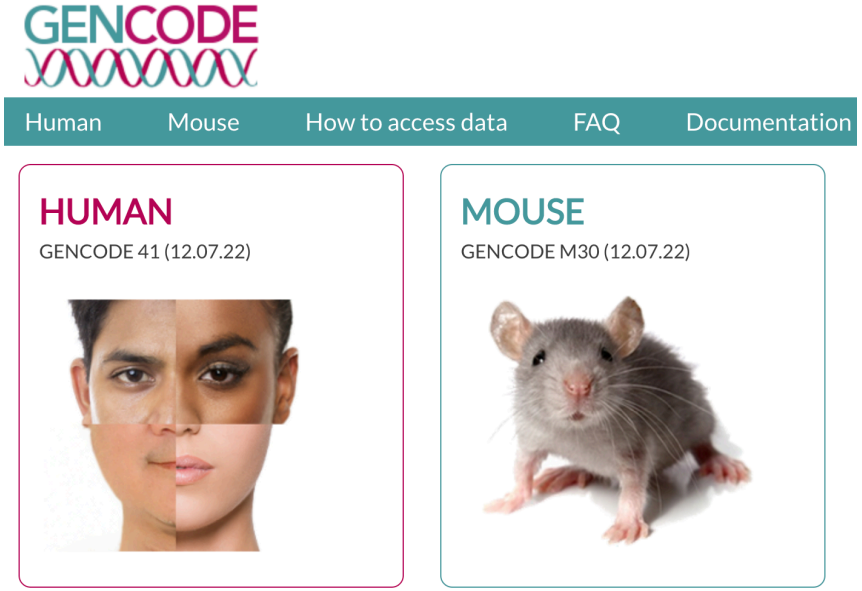
Reference Genomes

- Most applications will require a reference genome sequence against which to map
- Genome assembly/annotation projects are continuously being updated
  - Released in tagged versions
- Sources:
  - www.ensembl.org
  - genome.ucsc.edu
  - Consortia/species-specific sites
    - flybase.org
    - rgd.mcw.edu (Rat Genome Database)



Reference Transcriptome(s)

- GENCODE
  - Integrated annotation of gene features
  - gencodegenes.org
  - (Current) highest-quality source for transcriptome sequence and annotation in human and mouse
- Ensembl aggregates many other species assembled transcriptomes



Reference Gene/Feature Annotations

- Common formats:
  - GFF/GTF
    - https://useast.ensembl.org/info/website/upload/gff.html
    - Holds information about gene structure, ids, position
    - General Feature Format (.gff)
    - General Transfer Format (.gtf)
  - BED (Browser Extensible Data)
    - http://genome.cse.ucsc.edu/FAQ/FAQformat.html#format1
    - Generalized file format for genomic regions as coordinates and associated annotations

Building your own genome/transcriptome assembly

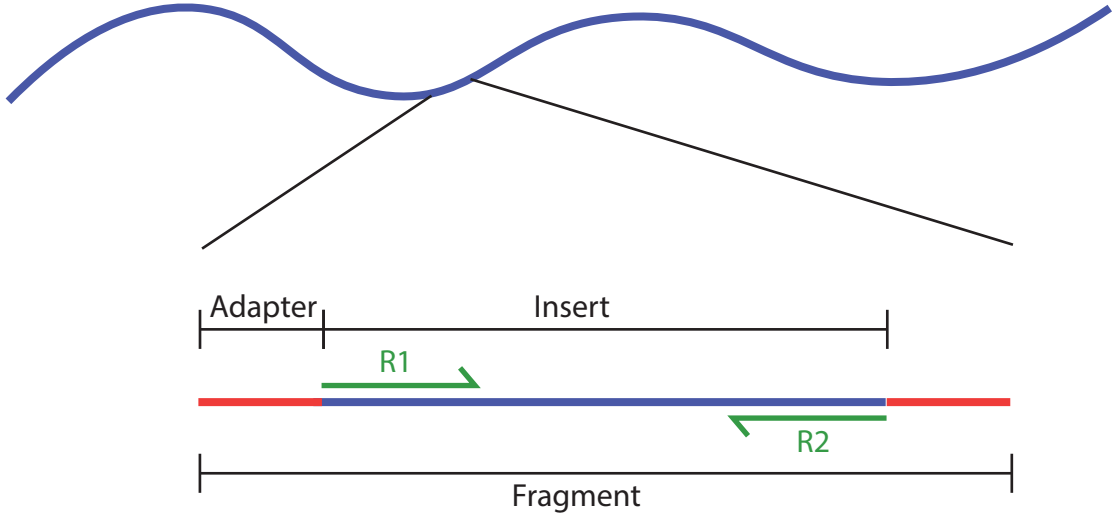
- Assembly of references is a massive and challenging undertaking.
- Is only ***required*** if:
  - You are working in a non-standard model organism
  - You anticipate novel genes/gene structures/features to be present in your dataset that wouldn't have been previously discovered/annotated
- Often, supplements are needed to assemblies/annotations:
  - Adding a BAC sequence that is part of a transgenic line
  - Adding transcript information for GFP in a reporter assay

JOHNS HOPKINS
SCHOOL of MEDICINE

WHAT IS A SEQUENCING READ?

Read terminology

- Input material is *fragmented*
- *Fragments* have *adapter* sequences added to either end
  - PCR handles
  - Sequencing adapters
  - Index/barcode sequences
- *Insert* is the portion of the *fragment* containing the sequence of interest
- *Reads* are the portions of the *insert* that are read by the sequencer



Adapter Trimming

- Why would we need to trim?
  - Shorter fragments and/or longer reads run the risk of actually sequencing ***into*** the downstream adapter sequences
  - Additionally, bases at the ends of reads tend to have lower quality scores

- This ***can*** have negative consequences for finding *alignments* to a reference

- Many tools exist to test the ends of reads for adapter sequences or lower quality bases and ‘trim’ them.

- Choice of whether to trim is subjective

Common Read Adapter Trimming Tools

- Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)
- FASTX-Toolkit
- fastp
- Cutadapt (https://cutadapt.readthedocs.io/en/stable/)

Single End vs Paired End sequencing

- Fragments can be sequenced from one OR both ends
  - Decision made at sequencing stage
  - Aligners require different input information for each
- Paired-end is almost ***always*** preferred
  - Significantly better alignment rates
  - Empirically determine fragment length (a useful metric for quantification)
  - Double the cost for reads of the same length
  - Read pairs are aligned jointly
    - 'Properly' mapped pairs significantly reduce ambiguity in alignment

Single End Sequencing

<sample_name>.fastq.gz

Paired End Sequencing

<sample_name>_1.fastq.gz

<sample_name>_2.fastq.gz

READ QC

Why do we need to QC reads?

- Ensuring data integrity
  - Different samples/labs/assays have different efficiencies and biases
  - Need reads of comparable quality to make fair comparisons across replicates and conditions.

- Poor quality reads can affect mapping rates, and/or lead to misquantifications or misinterpretations

- Identifying and filtering out low quality reads can improve analysis even if you have to sacrifice some of the data.

Read QC - What are we looking for?

- Sufficient ***number*** of reads
  - From sequencing run
  - Aligning to reference
- Reads of ***consistent*** and ***adequate*** quality scores
- Consistent sequence ***content***
  - Per base distribution of nucleotides
  - Per read GC content
  - Low number of poor quality or ambiguous bases
- Sufficient library ***complexity***
  - Low number of over-represented sequences
  - Low number of over-represented kmers
- Low proportion of sequencing adapters

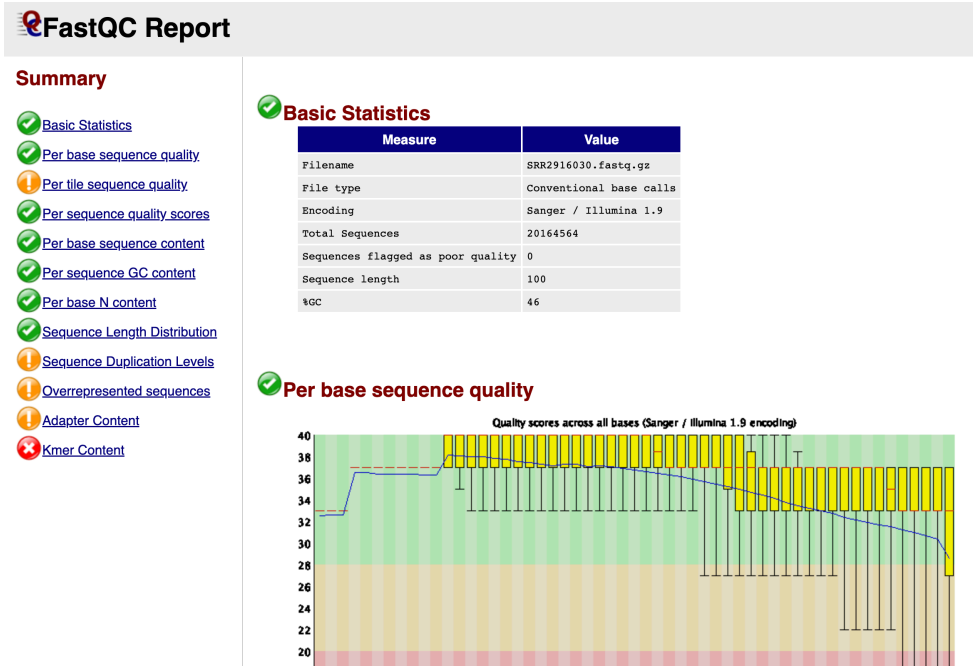
fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

- A GUI & command-line tool to analyze .fastq files for common quality metrics

- `fastqc <fastq_file.fastq.gz>`

- `fastqc *.fastq.gz`

- `echo $(ls *.fastq.gz) | xargs -n 1 -P 4 fastqc`



JOHNS HOPKINS
SCHOOL of MEDICINE

MAPPING/ALIGNMENT

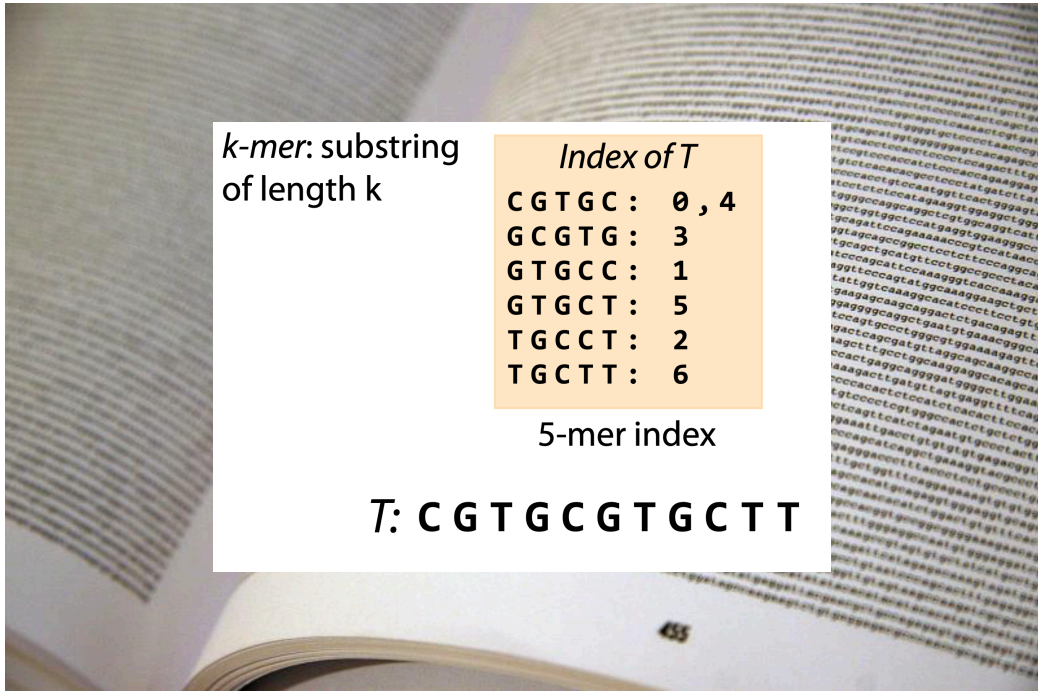
Read alignment

- Goal: to locate the genome coordinates from which a sequenced read could originate

- Requires searching a reference (genome/transcriptome) to find sequence matches for each of millions of reads
  - With some mismatch tolerance (SNPs, Indels, sequencing errors)

Reference Indexing

- Searching a linear reference sequence for substring matches is a computationally intensive task
  - Printed human genome, single spaced, 12pt font, 1” margins requires 1,206,980 pages (http://bio4.us/biotrends/human_genome_height.html)
- Generating an index (usually constructed from Kmer ‘seeds’) helps aligners to jump to specific regions of the genome that can then be extended to a full read match
- Most modern aligners provide a mechanism/software to generate an indexed reference as part of their documented workflows.



https://www.langmead-lab.org/teaching.html

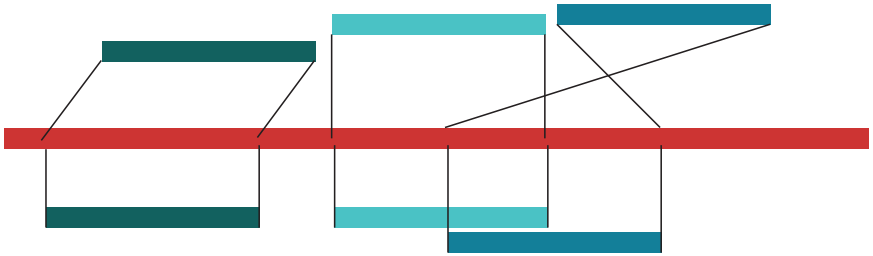
https://github.com/BenLangmead/ads1-slides/blob/master/0300_index__index.pdf

Choosing the appropriate reference

- Primarily guided by the ‘purpose’ of your analysis
  - Differential expression
  - Variant calling
  - Structural variant
  - Peak discovery
- Version and annotation level
  - Aim to use the most recent version that has sufficient annotation for your downstream needs
- Specialized/modified references
  - Ex: transcriptomes for RNA-Seq, mitochondrial genomes, or plasmids for specific studies.
  - Adding transgenes or other exogenous sequences included in your study to your reference may be helpful.
- Closeness to your sample
  - Ideally, use a reference from the same species/subspecies/strain.
  - Consider using a closely related species' genome if studying a less-known organism.

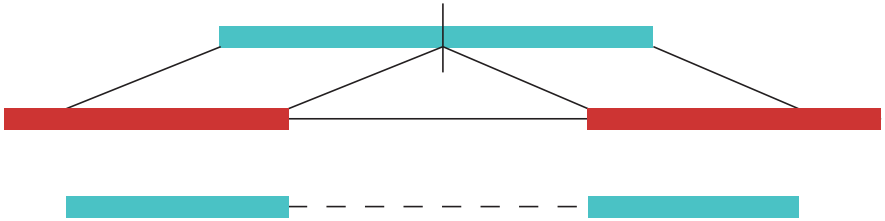
Ungapped vs Gapped alignment

Ungapped/unspliced



- Alignments without insertions, deletions, or gaps.
- ***Appropriate*** for most genomic reads
  - e.g. ChIP-Seq, ATAC-seq
- Speed: Faster, as there are fewer alignment possibilities.
- Ideal for short sequences or when expecting high similarity.
- Limitation: Cannot capture some structural events like insertions or deletions.

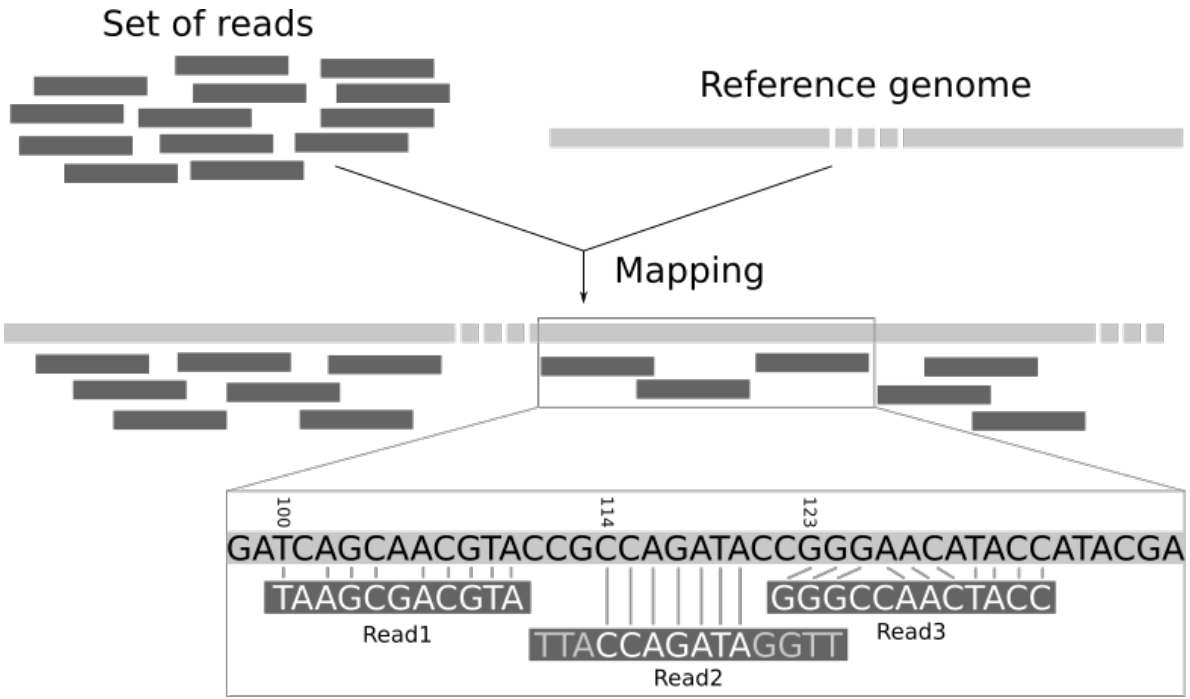
Gapped/spliced



- Alignments that allow for gaps, representing possible insertions, deletions in read relative to reference
- ***Required*** for most RNA-Seq studies
  - Splicing produces exon-exon junctions that can only be mapped with gaps to contiguous reference
- Accuracy: More accurate representation of evolutionary events, indels, mutations, spicing, etc
- Essential for longer sequences or when inferring evolutionary relationships.
- Limitation: Computationally more intensive.

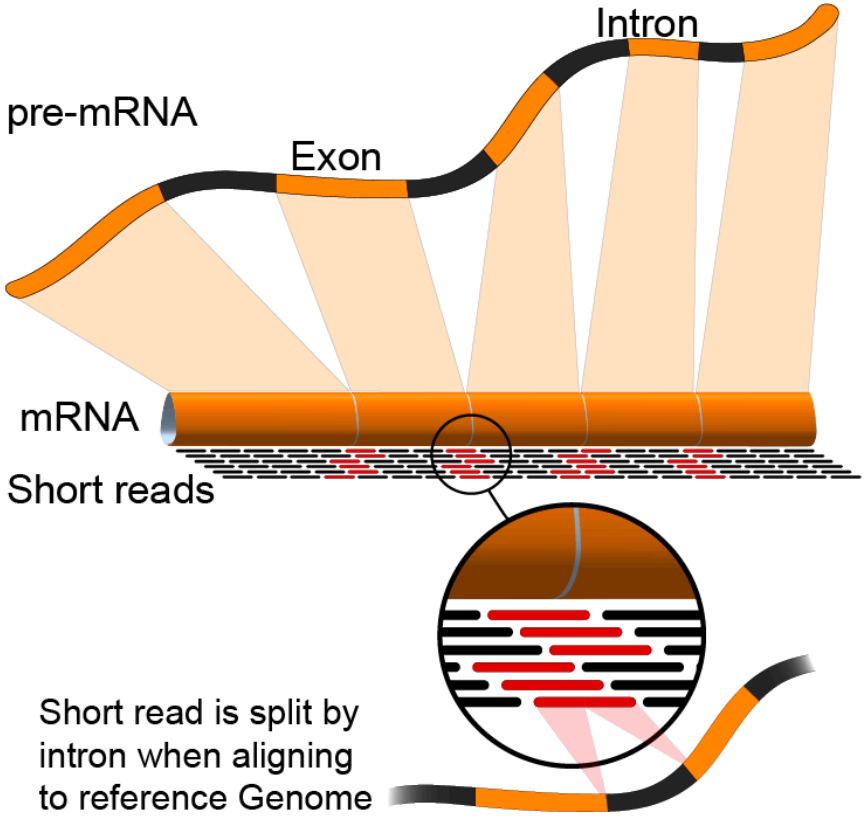
Alignment of DNA fragments to genome (ATAC-Seq, ChIP-Seq, etc)

- Each DNA-based read is assigned to a specific location in a reference genome
- Peaks identified by boundaries of overlapping reads above features of interest
  - Peak height/width/ count all useful attributes for quantification
  -



Alignment of RNA-based fragments

- Reads derived from cDNA may contain exon-exon junctions
  - Longer & paired-end reads more useful in this context
- Require use of splice-aware aligner
- Junctions useful for identification & quantification of isoform-level expression



What aligner to use?

- Depends on:
  - Task
  - Data type
  - Biological questions
  - Available resources

- *Tip: For choosing alignment tool and parameters for your final projects, check methods section, supplement, **and** data archive records for description of prior efforts.*

Common Aligners

- Unspliced (ungapped) Aligners
  - MAQ
  - BWA
  - Bowtie2
- Spliced (gapped) Aligners
  - Tophat2
  - HiSat2
  - STAR
- Pseudoalignment
  - Kallisto
  - Salmon

How can we improve the speed/efficiency of this process?

- More efficient algorithms:
  - Indexed search of genome/transcriptome

- Parallelization
  - Multiple threads for each iteration/call of the aligner
  - Multiple concurrent job submissions across samples
    - Naive parallelization is one chief benefit of HPC systems
    - Jobs are submitted to a ‘scheduler’ which balances resources and submits
    - Slurm (https://slurm.schedmd.com/documentation.html) is available on rockfish

JOHNS HOPKINS
SCHOOL of MEDICINE

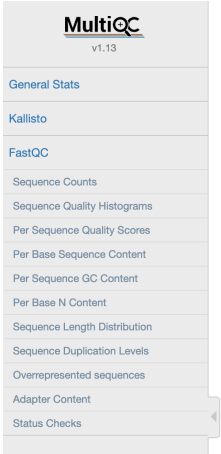
ALIGNMENT QC

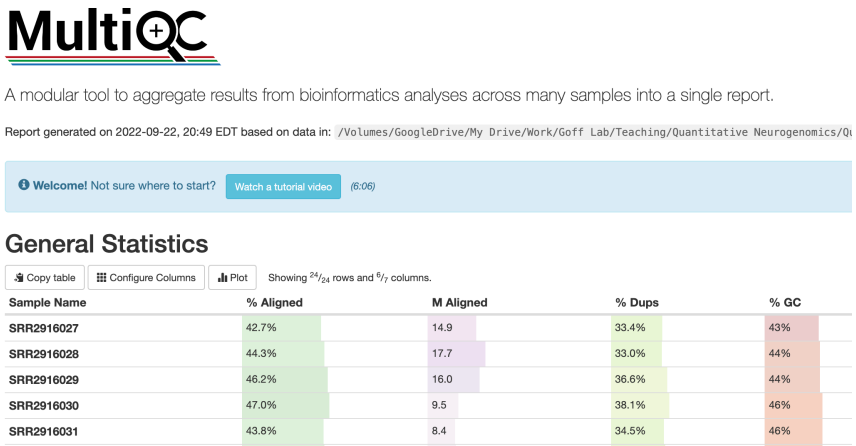
Alignment QC - What are we looking for?

- ***Number*** of reads aligned per sample
- ***Percentage*** of reads aligned per sample
- ***Number/percentage*** of reads uniquely mapping
  - % of reads with multiple mapping to reference (ambiguous)
- Paired End read ***concordance***
  - # of PE reads uniquely mapping
  - # of PE reads mapping but not together (discordant)
  - # of PE reads where only one mapped [uniquely]
- ***Consistent*** scores/values across samples
- *Values are often reported after each sample is processed. Useful to try and capture this information in a log file.*

Aggregate QC Reports with MultiQC

- https://multiqc.info/
- Generates a single .html summary report across multiple stages of a workflow
- Can be called (multiple times) from the root directory of a project
  - `multiqc .`
- Parses outputs and logs of many common genomics tools
  - Often times, STDERR output contains needed information and should be redirected to a file using (`2> <my_job>.log`)





JOHNS HOPKINS
SCHOOL of MEDICINE

ALIGNMENT VISUALIZATION WITH IGV

Integrative Genomics Viewer (IGV)

- https://software.broadinstitute.org/software/igv/
- Useful to generally inspect alignments where possible or relevant
  - Genomic deletion sites
  - Mutation insertion sites
  - Key regions/genes relevant to study
- Import indexed .bam files (and references) to view read distributions and associated features
  - Identified exon-exon junctions