



Quantitative Molecular Neurogenomics 2022

Module 4 - Data Sources, Acquisition, & QC

ME.440.825

Loyal Goff - Course Director

Alina Spiegel - TA

Richard Sriworarat - TA

PUBLIC NGS DATA SOURCES

NGS Data Sharing is required for NIH-funded research

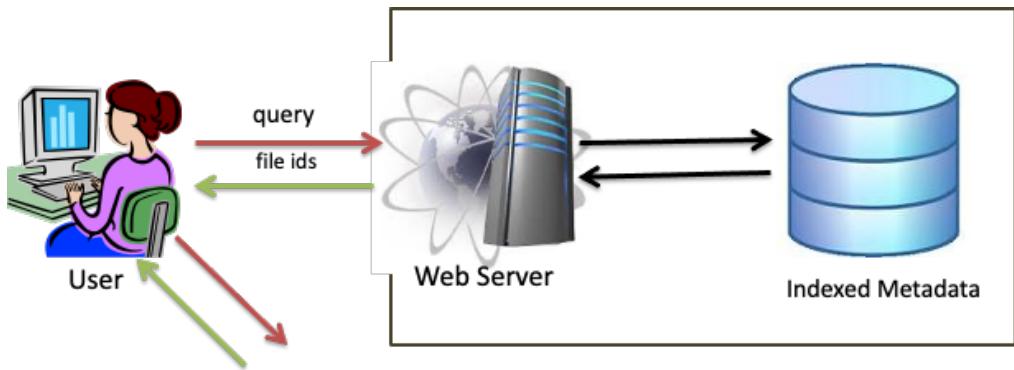
- NIH Issued Genomic Data Sharing (GDS) policy, August 2014
 - “...applies to all NIH-funded research (e.g., grants, contracts, and intramural research) that generates large-scale human or non-human genomic data.”
 - “NIH strongly encourages the use of established repositories to the extent possible 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”

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

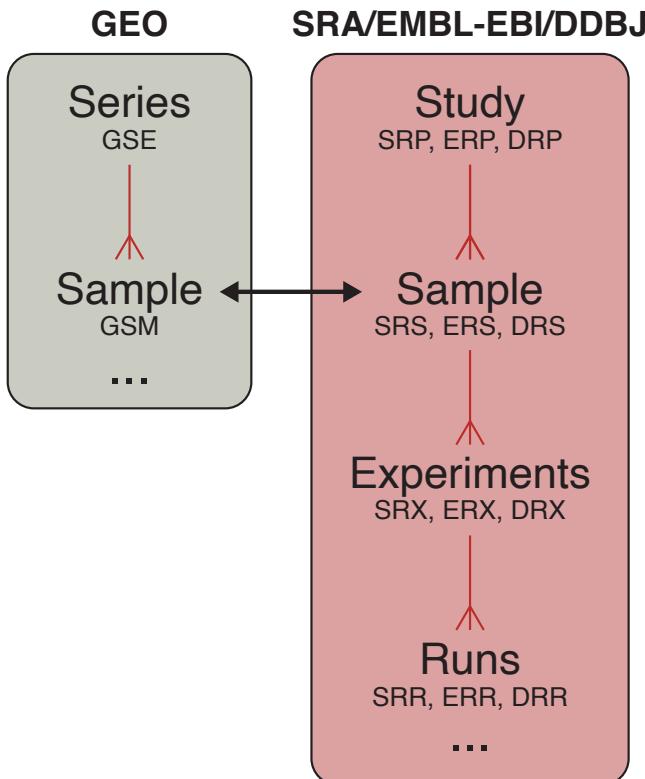
SRA - ncbi.nlm.nih.gov/sra/

- 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
-



File Servers – [http/ftp/fasp/etc..](http://ftp/fasp/etc..)

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)

DATA ACQUISITION



≡ HOME MAGAZINE COMMUNITY INNOVATION

NEWSLETTER ABOUT SUBMIT MY RESEARCH LOG IN/REGISTER



Tools and Resources

Neuroscience

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

Download

Cite

Comment



19,126 views

189 citations

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>

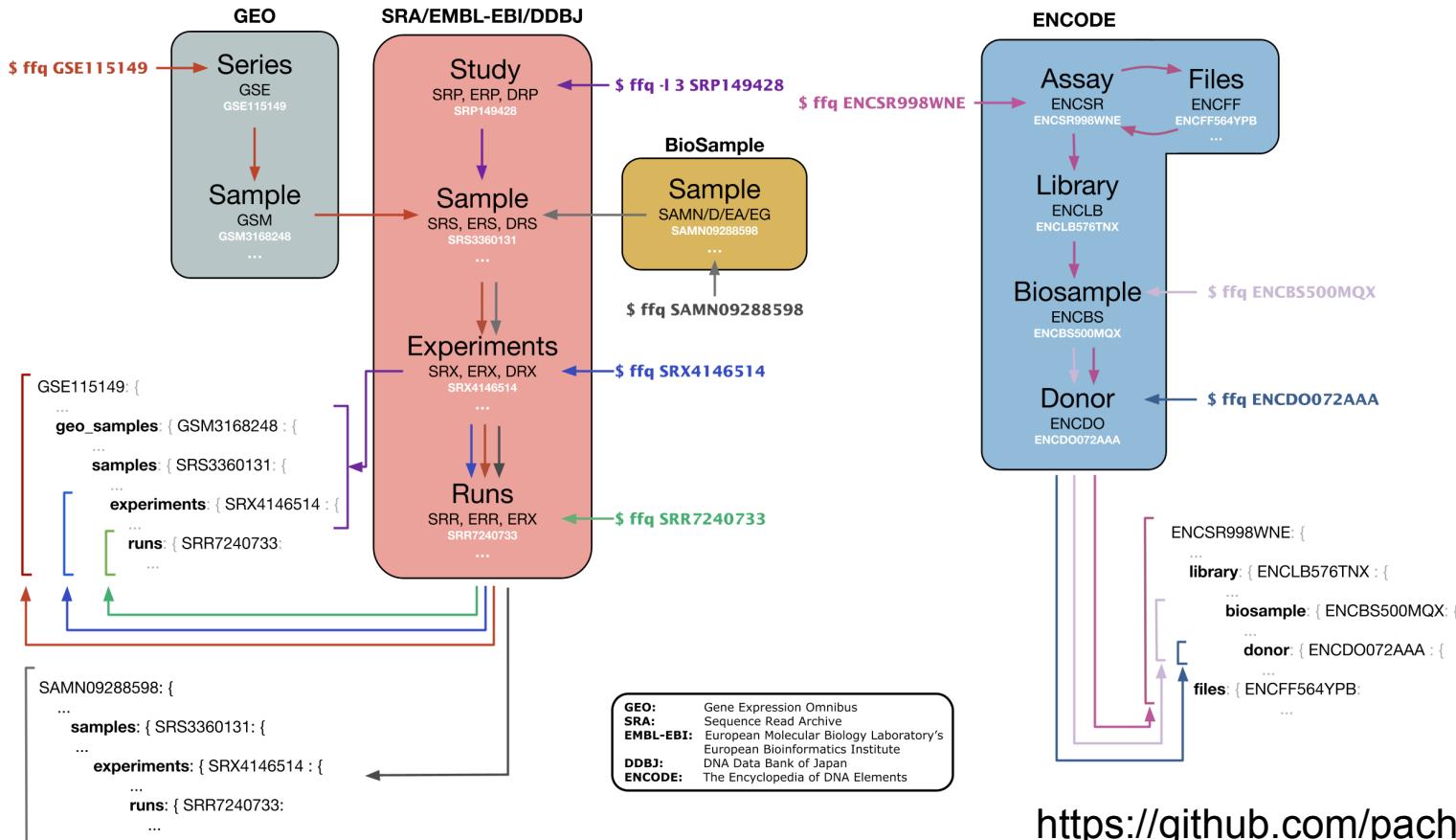
A FASTQ Read

Name @ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:36071/1

Sequence ACATCTGGTTCCTACTTCAGGGCCAATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT
(ignore) +

Base qualities ?@@FFBFFDDHHBCEAFGEGIIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FBEG:G

Fetch FastQ (ffq)



<https://github.com/pachterlab/ffq>

Using ffq (on GEO Study Accession)

- ffq GSE74985 > GSE_info.json

JSON metadata from ffc

- Information about each sample
- URLs for downloading raw data
 - In several different forms/locations
-

```
{  
  "GSE87548": {  
    "accession": "GSE87548",  
    "supplementary_files": [  
      {  
        "accession": "GSE87548",  
        "filename": "GSE87548_merged_peaks.bed.gz",  
        "filetype": null,  
        "filesize": 2038147,  
        "filenumber": 1,  
        "md5": null,  
        "urltype": "ftp",  
        "url": "ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE87nnn/  
GSE87548/suppl/GSE87548_merged_peaks.bed.gz"  
      },  
      {  
        "accession": "GSE87548",  
        "filename":  
        "GSE87548_merged_peaks.tmmScores.txt.gz",  
        "filetype": null,  
        "filesize": 10648951,  
        "filenumber": 2,  
        ...  
    ]  
  }  
}
```

ffq | jq to parse out metadata

- jq is a command-line tool for processing .json files
 - <https://stedolan.github.io/jq/>
- jq filters to retrieve sample info
 - `jq -r '.[].geo_samples | .[].samples | .[].experiments | .[].runs | .[].files.ftp | .[].url' gse_info.json`
 - Returns list of sequence run fastq file links
 - `jq -r '.[].geo_samples | .[].samples | .[] | [.accession, (.experiments | .[].runs | .[].files.ftp | .[] | .[]), (.attributes | .[])] | @tsv' $GEO.json > $GEO.tsv`
 - Returns tab-separated matrix of sample attributes

ACCESSORY DATA SOURCES

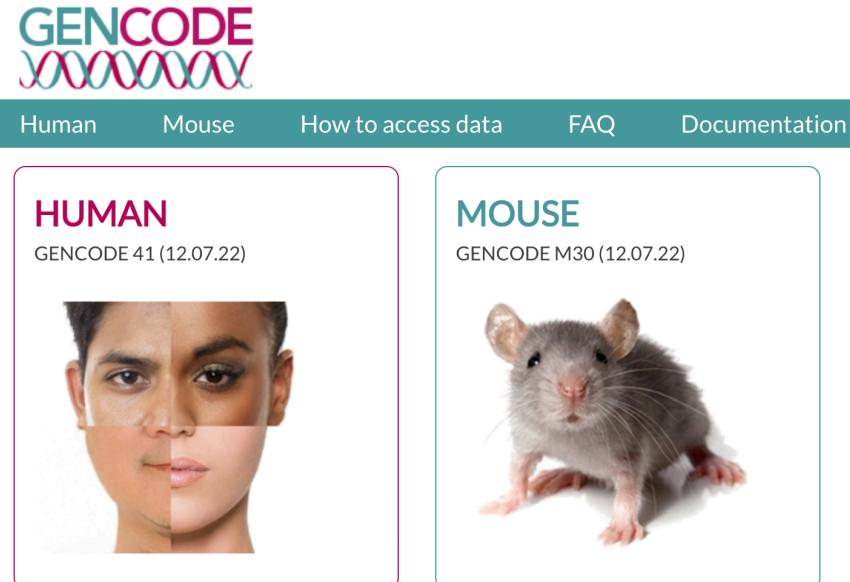
Reference Genomes

- 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)

The screenshot shows the Ensembl genome browser homepage. At the top, there's a dark header bar with the Ensembl logo and links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. Below the header, there are four main navigation sections: 'Tools' (with a 'All tools' link), 'BioMart >' (described as 'Export custom datasets from Ensembl with this data-mining tool'), 'BLAST/BLAT >' (described as 'Search our genomes for your DNA or protein sequence'), and 'Variant Effect Predictor >' (described as 'Analyse your own variants and predict the functional consequences of known and unknown variants'). The main content area has a light blue background. It features a search bar with dropdown menus for 'All species' and 'for', and a 'Go' button. Below the search bar is a note: 'e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease'. Further down, there's a section for 'All genomes' with a dropdown menu labeled 'Select a species'. To the right of this is a 'Favourite genomes' section with icons and names for Human (GRCh38.p13), Still Using GRCh37?, Mouse (GRCm39), and Zebrafish (GRCz11). At the bottom left, there's a link 'View full list of all species'.

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

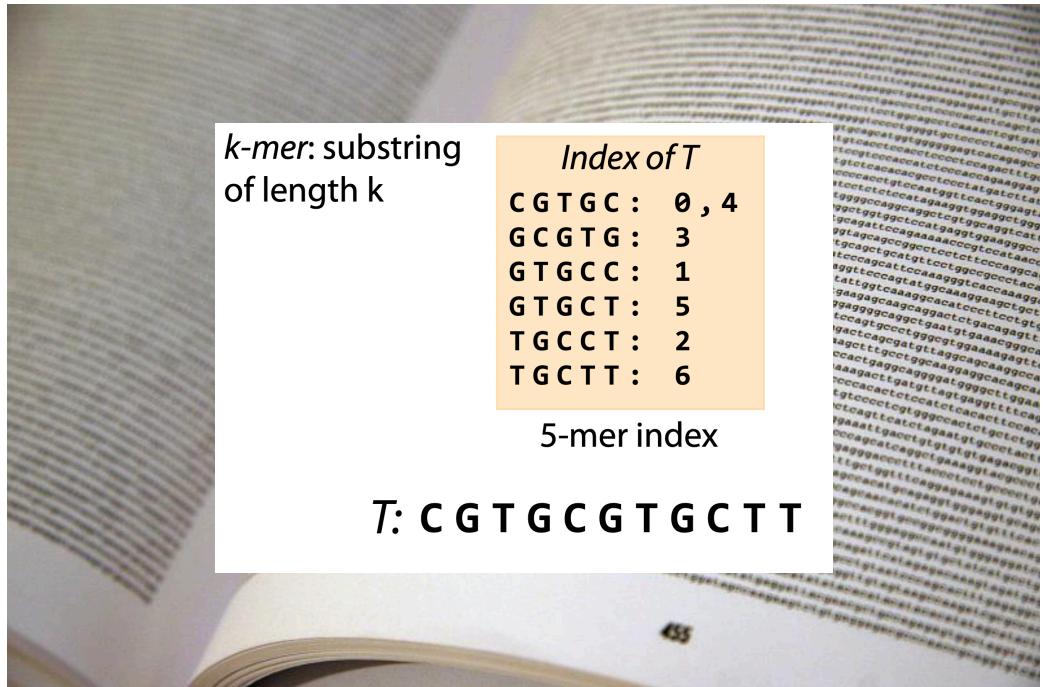
MAPPING/ALIGNMENT

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



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
 - Double the cost for reads of the same length
 - Read pairs are aligned jointly

Single End Sequencing



<sample_name>.fastq.gz

Paired End Sequencing

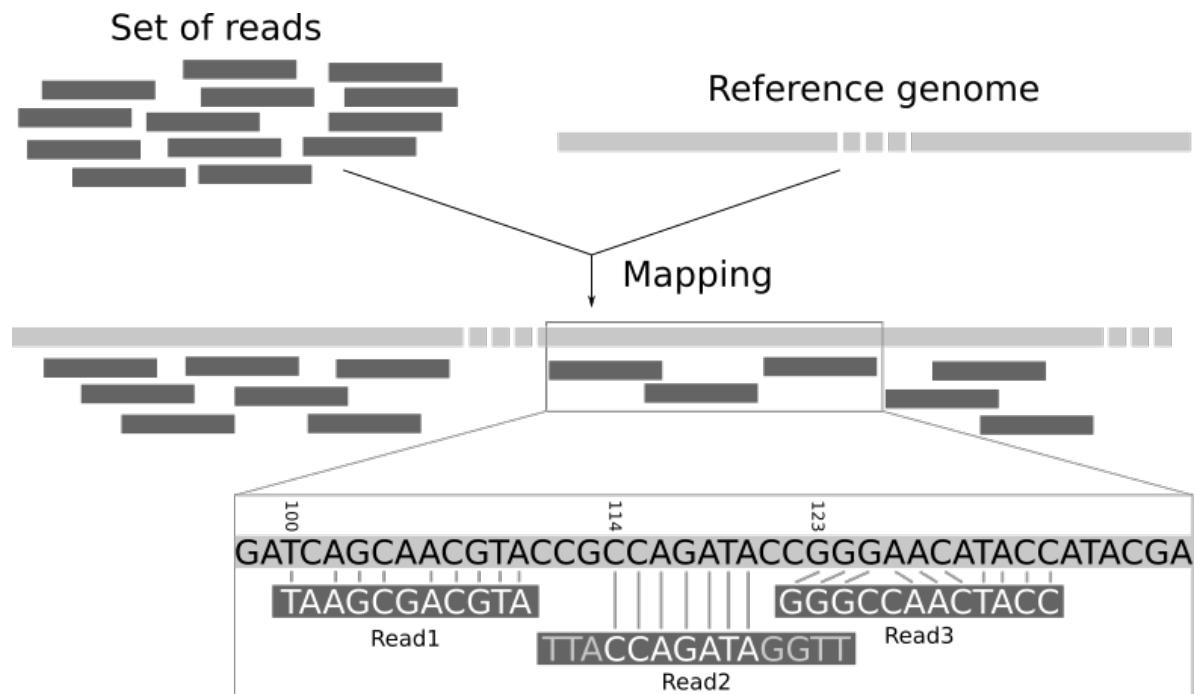


<sample_name>_1.fastq.gz

<sample_name>_2.fastq.gz

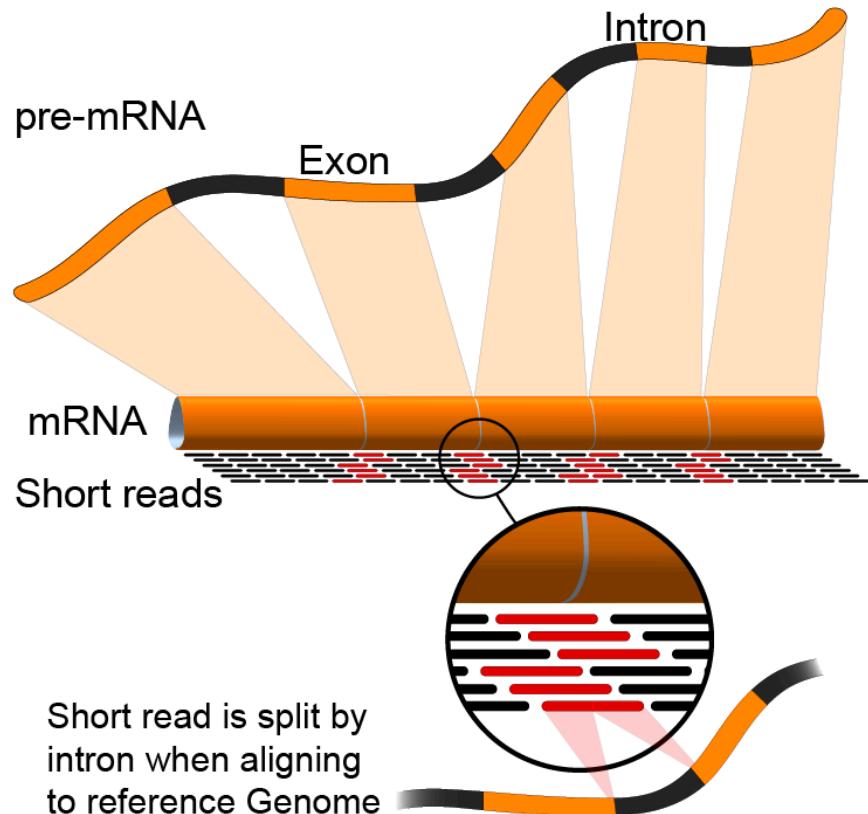
Alignment of DNA fragments to genome (ATAC-Seq, ChIP-Seq, etc)

- 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
 -



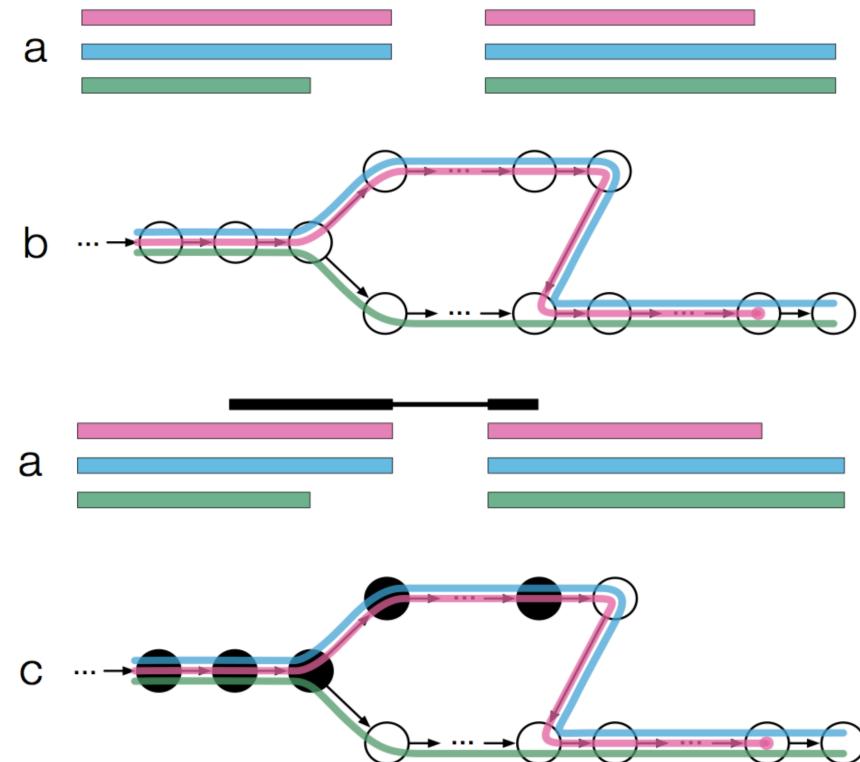
Alignment of RNA-based fragments

- 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



Pseudoalignment of RNA-Seq reads

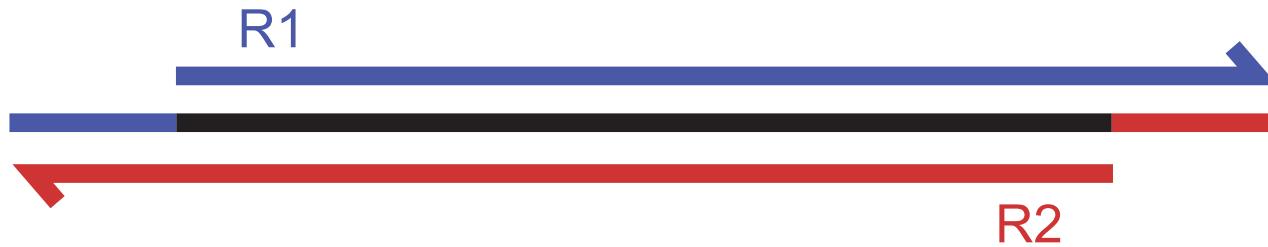
- Different Goal:
 - ...determine, for each read, not where in each transcript it aligns, but rather which transcripts it is compatible with...
 - Useful for ‘counting-based’ problems
- Hashed index of Kmers that are ‘compatible’ with each reference transcript isoform
- Intersection of all k-compatibility classes for a read identifies the ‘compatible isoforms’ for a given read.



RNA: Alignment or Pseudoalignment

- Alignment
 - Mapping reads to annotated reference genome
 - Variety of (splice-aware) alignment methods
 - Can be used for discovery of new transcripts
 - Speed depends on method
- Pseudoalignment
 - Mapping reads directly to reference transcriptome
 - K-mer-based equivalency class
 - No inference beyond reference transcriptome
 - Very fast
 - Speed allows for bootstrapping estimation of error in quantification

Read Adapter Trimming Tools



- Remove sequenced adapter from 3' (or sometimes 5') end of sequencing reads
- Can improve read mapability & overall quality
- Often can ‘rescue’ unmapped reads with trimming

Common Read Adapter Trimming Tools

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

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 Aligners
 - MAQ
 - BWA
 - Bowtie2
- Spliced 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

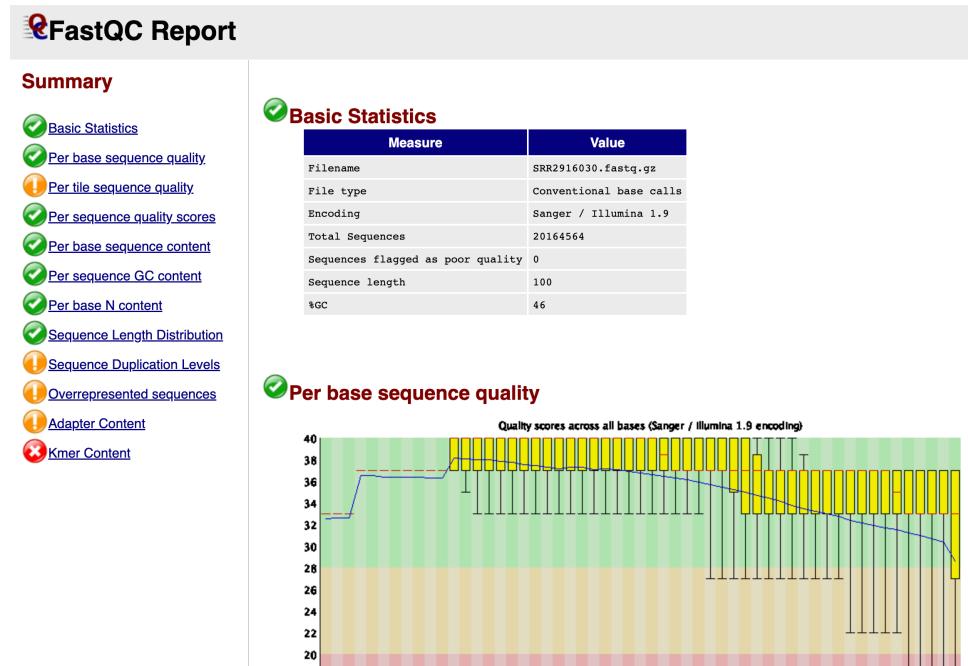
ALIGNMENT QC

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

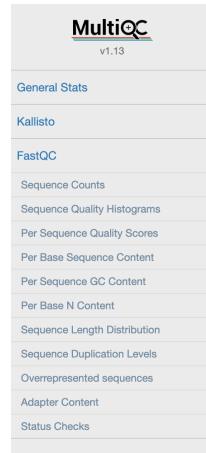


Alignment QC - What are we looking for?

- **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)



A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2022-09-22, 20:49 EDT based on data in: /Volumes/GoogleDrive/My Drive/Work/Goff Lab/Teaching/Quantitative Neurogenomics/01



General Statistics

Sample Name	% Aligned	M Aligned	% Dups	% GC
SRR2916027	42.7%	14.9	33.4%	43%
SRR2916028	44.3%	17.7	33.0%	44%
SRR2916029	46.2%	16.0	36.6%	44%
SRR2916030	47.0%	9.5	38.1%	46%
SRR2916031	43.8%	8.4	34.5%	46%

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

