In-Person

Genomics Compute Cluster

STAR RNA-seq Aligner

Northwestern | INFORMATION TECHNOLOGY
RESEARCH COMPUTING AND DATA SERVICES

COMPUTING AND SOFTWARE

Access to high performance computing, research software, and global networks for conducting computationally intense research.

DATA MANAGEMENT AND SHARING

Learn about data management planning and options for storing, securing, transferring, and sharing data.

DATA SCIENCE, STATISTICS, AND VISUALIZATION

Support for collecting, analyzing, visualizing, and programming with research data.

TRAINING AND CONSULTATION

Identify events, resources, and people to help you learn computational and data skills for your research.

Research Computing and Data Services

We're here to help after the workshop!

quest-help@northwestern.edu

bit.ly/rcdsconsult

https://sites.northwestern.edu/researchcomputing/

Setup…

1. log onto Quest

   ssh <netid>@login.quest.northwestern.edu

   # enter your netid password (it won't look like you're typing)

2. move to our classroom folder

   cd /projects/e32680

3. make your own subfolder if you weren't here last week

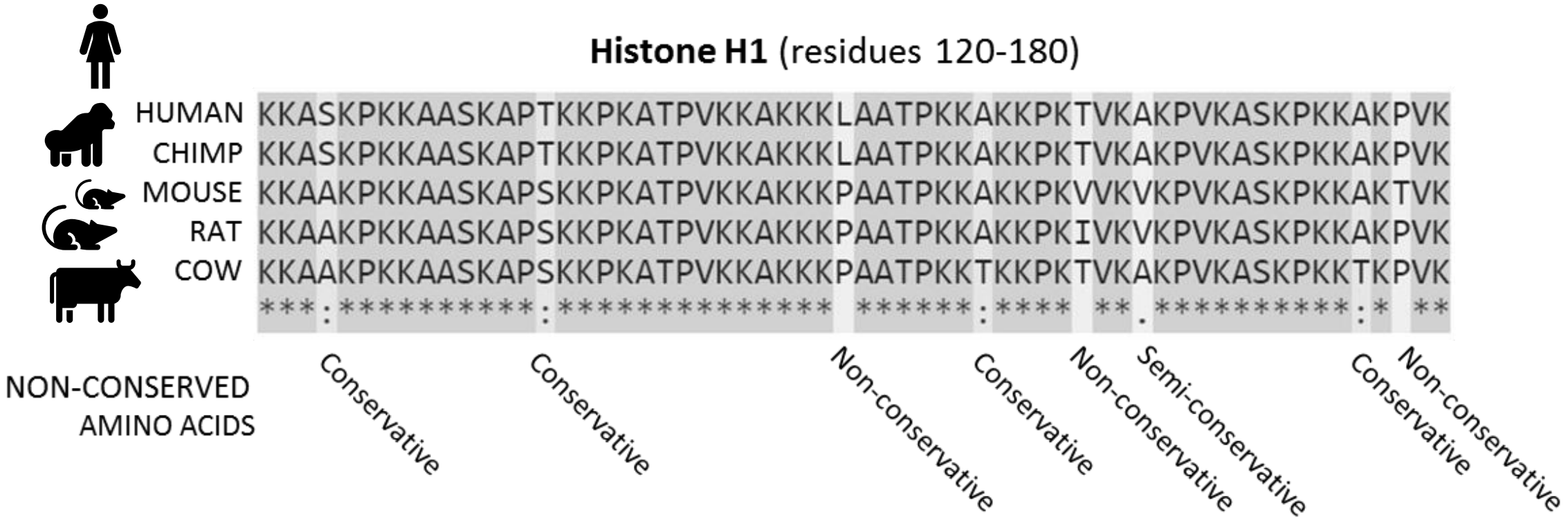
   mkdir <folder_name>

4. move into your folder

   cd <folder_name>

Class materials are at:
github.com/nuitrcs/star_aligner_workshop

What is alignment? - hypothesis of homology



How to pick an aligner?

- Type of data you have
  - aligning short read data (Illumina)
  - aligning long read data (PacBio, Nanopore)
  - aligning RNA-seq data - splice aware
- Accuracy
  - percent of reads aligned
  - estimated gene coverage
- Runtime

Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider

*Ryan Musich[1], Lance Cadle-Davidson[1,2] and Michael V. Osier[1]**

[1]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, United States,
[2]USDA-Agricultural Research Service, Grape Genetics Research Unit, Geneva, NY, United States

RESEARCH ARTICLE Open Access

Benchmarking short sequence mapping tools

Ayat Hatem[1,2], Doruk Bozdağ[2], Amanda E Toland[3] and Ümit V Çatalyürek[1,2]*

bioinformatics at RIT

A Recent (2020) Comparative Analysis of Genome Aligners Shows HISAT2 and BWA are Among the Best Tools

Article

Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research

Isaac D. Raplee [1], Alexei V. Evsikov [2] and Caralina Marín de Evsikova [1,2,*]

How to pick an aligner?

- Type of data you have
  - aligning short read data (Illumina)
  - aligning long read data (PacBio, Nanopore)
  - aligning RNA-seq data - splice aware
- Accuracy
  - percent of reads aligned
  - estimated gene coverage
- Runtime

Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider

*Ryan Musich[1], Lance Cadle-Davidson[1,2] and Michael V. Osier[1]**

[1]Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, United States,
[2]USDA-Agricultural Research Service, Grape Genetics Research Unit, Geneva, NY, United States

RESEARCH ARTICLE Open Access

Benchmarking short sequence mapping tools

Ayat Hatem[1,2], Doruk Bozdağ[2], Amanda E Toland[3] and Ümit V Çatalyürek[1,2]*

bioinformatics at RIT

A Recent (2020) Comparative Analysis of Genome Aligners Shows HISAT2 and BWA are Among the Best Tools

Article

Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research

Isaac D. Raplee [1], Alexei V. Evsikov [2] and Caralina Marín de Evsikova [1,2,*]

Why does it matter if you have DNA or RNA data?

often aligning to a reference genome

- create a standard index to map to

most RNA-seq data is mature mRNA

- typically no introns in the sequence of reads

but genomes have introns!

- might introduce long gaps in the alignment, which are generally penalized against when determining the `best` mapping of a read
- might show two or more partial matches for a read to different exons

spice-aware aligners

- allows identification of novel splice junctions, regions of expression, structural variants…

STAR - Spliced Transcripts Alignment to a Reference

- **splice aware** aligner
- potentially better suited for draft or low-quality genomes than HISAT2 or other splice-aware options
- uses more resources than HISAT2
- default option of many pipelines, include nf-core's RNA-seq

- Resources:
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3530905/
- https://github.com/alexdobin/STAR

How do we use it?

Generate genome index files

Map reads to the genome

Generate genome index files

- if you are generating your own index you need:
  1. fasta file of the genome sequence
  2. gtf files of the genome annotation

Generate genome index files

```
STAR \
  --runMode genomeGenerate \
  --runThreadN 4 \
  --genomeDir /path/to/genomeDir \
  --genomeFastaFiles /path/to/genome.fa \
  --sjdbGTFfile /path/to/genome.gtf \
  --sjdbOverhang <ReadLength-1>
```

Generate genome index files

```
STAR \
  --runMode genomeGenerate \
  --runThreadN 4 \                  # number of cores to parallelize over
  --genomeDir /path/to/genomeDir \                  # output location
  --genomeFastaFiles /path/to/genome.fa \                  # input location
  --sjdbGTFfile /path/to/genome.gtf \                  # input location
  --sjdbOverhang <ReadLength-1>                  # length of region around annotation to be used
                                                 for constructing splice junctions database, default
                                                 is 100, ReadLength-1 is recommendation
```

Running this on Quest…

- Quest uses a scheduler to control access to compute resources.
- We will wrap the STAR command in a ‘job script’ which will request compute resource through the scheduler and set up the software environment for STAR to run.
- I have a template job script available to you! Please copy it into your folder with the following:

```
cd /projects/e32680 # you may already be here
cp 02_staralignment_reference/star1.sh <your_folder>
cd <your_folder>
ls # you should see star1.sh
```

Running this on Quest…

Open the script with nano:

```
nano star1.sh
```

Change the genomeDir to:

```
/projects/e32680/<your_folder>/STARindex
```

Save and exit nano with CTRL+O, ENTER, CTRL+X

Launch the job with:

```
sbatch star1.sh
```

Generate genome index files

We host pre-generated index files on Quest for many model systems:

- /projects/genomicsshare/AWS_iGenomes/references/Mus_musculus/Ensembl/GRCm38/Sequence/STARIndex
- /projects/genomicsshare/AWS_iGenomes/references/Homo_sapiens/Ensembl/GRCh37/Sequence/STARIndex

Map reads to genome

```
STAR \
  --runThreadN 8 \                  # number of cores to parallelize over
  --genomeDir /path/to/genomeDir \  # input location
  --readFilesIn /path/to/genome.fa  # input location, this can be multiple files fasta or fastq, both paired end read files need to be supplied
```

Those are the basic options but there are about a million more you can adjust...

Copy an alignment script to your folder….

```
cp /projects/e32680/02_staralignment_reference/star2.sh .
```

Copy an alignment script to your folder….

```
cp /projects/e32680/02_staralignment_reference/star2.sh .
```

What does this script contain?

star2.sh

```
STAR \
  --runThreadN 4 \
  --genomeDir /projects/e32680/02_staralignment_reference/STARindex \
  --readFilesIn
/projects/e32680/02_staralignment_reference/SRR7773980_1.fastq.gz,/projects/e32680/
02_staralignment_reference/SRR7773980_2.fastq.gz \
  --readFilesCommand gunzip -c \
  --outFileNamePrefix ./STARoutputs
```

Launch the alignment job with:

```
sbatch star2.sh
```

From your folder! Not the 02_staralignment_reference one.

Outputs

```
(base) [hsc945@quser44 STARoutputs]$ ls
Aligned.out.sam  Log.final.out  Log.out  Log.progress.out  SJ.out.tab
```

Alignment file in SAM format

Log output that includes higher level summaries of mapping rates, number of splice sites, and runtime information

Log output that includes step by step detailed output

Log output that checks progress and reports summary every minute during runtime

Splice junction output in tab delimited format

Aligned.out.sam

- alignment file in SAM (sequence alignment map) format
- output format can be controlled with

  --outSAMtype

- for unsorted BAM (binary alignment map) format:

  --outSAMtype BAM Unsorted
- for sorted BAM format:

  --outSAMtype BAM SortedByCoordinate
- for both files:

  --outSAMtype BAM Unsorted SortedByCoordinate

--outBAMsortingThreadN # controls number of threads for sorting, default is 6

SJ.out.tab

- high confidence splice junctions in tab-delimited format
- 9 columns: chromosome, first base of intron, last base of intron, strand, intron motif, annotation, number of uniquely mapped reads crossing junction, number of multi-mapping reads crossing junction, maximum spliced alignment overhang
- filtering can be controlled with

  --outSJfilter*

- for only uniquely mapping reads:

  --outSJfilterReads Unique

”Pioneer factor GATA6 promotes colorectal cancer through 3D genome regulation”

Lyu et al. 2025. *Science Advances*

https://pmc.ncbi.nlm.nih.gov/articles/PMC11804904/#sec11

Materials and Methods

RNA-seq data processing

The trimmed FASTQ files of colon cancer cell lines are aligned against hg38 human reference genome using STAR (65)(v.2.7.10b) with parameters “--outSAMunmapped Within --outFilterType BySJout --outSAMattributes NH HI AS NM MD --outFilterMultimapNmax 20 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --sjdbScore 1”. RSEM (66)(v1.3.3) facilitated the quantification and expression calculation for known genes with GENCODE v33. Genes with TPM value < 1 in all samples were excluded from downstream analyses.

Translate to job script to replicate on Quest:

aligned against hg38 human reference
STAR v.2.7.10b
--outSAMunmapped Within
--outFilterType BySJout
--outSAMattributes NH HI AS NM MD
--outFilterMultimapNmax 20
--outFilterMismatchNmax 999
--outFilterMismatchNoverReadLmax 0.04
--alignIntronMin 20
--alignIntronMax 1000000
--alignMatesGapMax 1000000
--alignSJoverhangMin 8
--alignSJDBoverhangMin 1
--sjdbScore 1

You can copy /projects/e32680/02_star aligner_reference/star3.sh as a starting place.