STAR on HPC

What is STAR?

STAR (Spliced Transcripts Alignment to a Reference) is an RNA sequencing alignment software used for mapping RNA sequencing reads to a reference genome. It was developed by Alex Dobin, et al. at the Cold Spring Harbor Laboratory.

One of the key features of STAR is its ability to detect known and novel splice junctions. It achieves this by using a two-pass alignment strategy, in which it first identifies splice junctions from the RNA sequencing reads, and then re-aligns the reads using the detected splice junctions as a guide.

STAR also uses a seed-and-extend algorithm to align RNA sequencing reads across exon junctions, which increases the accuracy and sensitivity of the alignment. It also allows for the quantification of transcript expression levels, which is useful for identifying differentially expressed genes in RNA sequencing experiments.

In addition to its accuracy and sensitivity, STAR is also known for its speed. It is able to map millions of reads per hour, making it a preferred choice for aligning large RNA sequencing datasets.

Links:

Official Website

Manual

Versions Available:

The following versions are available on the cluster:

- `STAR 2.5.3a`

How to load STAR?

To load STAR, use the following commands:

```
#Load the STAR module
module load bio/star/2.5.3a
```

To verify if the module is loaded correctly, use the following command,

```
# List all the module loaded in the environment
module list
```

How to use STAR?

To generate the genome indices, STAR utilizes the genome file in FASTA format and the gene annotation file in either GTF or GFF3 format. The gene annotation file is necessary to establish the known splice junctions, which ultimately enhances the precision of the genome mapping.

Example 1:

Build a genome index using GTF file,

```
STAR --runThreadN 12 \
    --runMode genomeGenerate \
    --genomeDir ath_star_index \
    --genomeFastaFiles Athaliana_TAIR10.fasta \
    --sjdbGTFfile Athaliana_gene.gtf \
    --sjdbOverhang 149
```

Mapping reads to genome,

```
STAR --runThreadN 12 \
     --readFilesIn ath_seed_sample.fastq \
     --genomeDir ath_star_index \
     --outSAMtype BAM SortedByCoordinate \
     --outFileNamePrefix seed_sample  \
     --outSAMunmapped Within
```

STAR generates multiple output files during the RNA sequencing read alignment process. Some of the key output files include the aligned reads file (in SAM or BAM format), the unmapped reads file (in FASTQ format), the junctions file (in tab-separated format), and the chimeric junctions file (in tab-separated format).

Use the following slurm script as a template to run STAR on HPC,

```
#!/bin/bash
#SBATCH --job-name=STAR_alignment
#SBATCH -p main
#SBATCH -qos main
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=4
#SBATCH --mem=32G
#SBATCH --time=12:00:00
#SBATCH --output=STAR_alignment.out
#SBATCH --error=STAR_alignment.err

# Load the STAR module
module load bio/star/2.5.3a

# Set the input file paths
READ1=/path/to/read1.fastq
READ2=/path/to/read2.fastq
```

```
GENOME_DIR=/path/to/genome/directory
OUTPUT_PREFIX=/path/to/output/prefix

# Run STAR with $SLURM_CPUS_PER_TASK threads
STAR --runThreadN $SLURM_CPUS_PER_TASK --genomeDir $GENOME_DIR --
readFilesIn $READ1 $READ2 --outFileNamePrefix $OUTPUT_PREFIX
```

Where to find help?

If you are confused or need help at any point, please contact OIT at the following address.

https://ua-app01.ua.edu/researchComputingPortal/public/oitHelp