

RepeatModeler on HPC cluster:

What is RepeatModeler?

RepeatModeler is a de novo transposable element (TE) family identification and modeling package. At the heart of RepeatModeler are three de-novo repeat finding programs (RECON, RepeatScout and LtrHarvest/tr_retriever) which employ complementary computational methods for identifying repeat element boundaries and family relationships from sequence data.

Additional Information is available on the following page:

- <https://github.com/Dfam-consortium/RepeatModeler>
- <https://www.pnas.org/doi/10.1073/pnas.1921046117>

Version available on HPC:

There are two versions of RepeatModeler currently available on the HPC:

RepeatModeler – V1.0.8

RepeatModeler – V2.0.3

How to load RepeatModeler on the cluster?

To load a version of RepeatModeler on the HPC on current terminal session, use the following command:

```
module avail bio/repeatmodeler
```

The version will be listed. To a version of software, use following command:

```
module load bio/repeatmodeler/2.0.3
```

Verify by using this command:

```
module list
```

The loaded software and dependencies will be shown.

How to use RepeatModeler on the cluster?

Here is one example of how to use RepeatModeler on the cluster. To get a test file to process through RepeatModeler, use the following commands

```
# Make a test directory on
```

```
cd ~ && mkdir repeatmodeler_test && cd repeatmodeler_test
```

```
#Download Test Data
```

```
curl -L -o test.fa.gz\
```

```
https://hgdownload-test.gi.ucsc.edu/goldenPath/hg38/chromosomes/chr7.fa.gz
```

```
# Unzip the test file
```

```
gunzip test.fa.gz
```

```
# Now after you load your module and copied test files, we can make a script to run RepeatModeler
```

The Script method:

```
#!/bin/bash  
#SBATCH -J my_job_name # job name
```

```
#SBATCH -n 1 # Nodes per task
```

```
#SBATCH -c 20      # number of cores to use (should be 1)  
#SBATCH -p main    # partition type  
#SBATCH --qos main # quality of service  
#SBATCH --mem-per-cpu=5G # RAM  
#SBATCH -e errors.%A # error log file  
#SBATCH -o output.%A #STDOUT file  
#SBATCH --mail-user={your email address} # mail the user the status of the job
```

```
module load bio/repeatmodeler/2.0.3 # load the module
```

```
#load the free bayes module  
module load bio/RepeatModeler/latest
```

```
# Moving to the test directory
```

```
cd $SLURM_SUBMIT_DIR
```

```
#BUILD A DATABASE USING REPEAT MODELER
```

```
BuildDatabase -name Human test.fa
```

```
#USE REPEATMODELER TO PROCESS THE DATABASE
```

```
RepeatModeler -database Human -pa 20 -LTRStruct >& run.out
```

Schedule the job with the following sbatch command.

```
sbatch myscript.sh
```

All the processed files will be generated in the same directory as the script.

Where to find help?

If you are stuck on some part or need help at any point, please contact OIT at the following address.

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

Resources:

- <https://github.com/Dfam-consortium/RepeatModeler>