

Computational Biology on CARC Systems

Center for Advanced Research Computing
University of Southern California

Tomasz Osinski, PhD
Research Facilitator in Life Science



USC

Advanced Research Computing
Enabling scientific breakthroughs at scale

Bio Resources (updated yearly)

<https://carc.usc.edu/user-information/bio-resources>

The screenshot displays the USC Advanced Research Computing website. At the top, a yellow navigation bar contains the text "USC | USC ITS | Office of the CIO | Office of Research" on the left and a search bar with the placeholder "ENTER SEARCH TERMS" and a magnifying glass icon on the right. Below this, the USC logo is followed by the text "Advanced Research Computing" and "Enabling scientific breakthroughs at scale". A horizontal menu features several items: "About", "Services", "User Information" (circled in red), "Education & Outreach", "News & Events", and "User Support" (highlighted in yellow). Under the "User Information" menu, a list of links is shown: "Getting Started", "Accounts And Allocations", "System Information", "User Guides", "CARC User Portal", "CARC User Forum", "CARC OnDemand", "Condo Cluster Program", "Bio Resources" (circled in red), "Using USC's Cryo-EM Instruments", and "Frequently Asked Questions". The main content area features a large banner with a background image of a baseball field. The banner text reads "Brings Cryo-EM to USC" in large white letters, with "Phase 1 of the cryogenic electron microscopy (cryo-EM) project is now complete" in smaller text below it. A "Read More" button is centered at the bottom of the banner. Navigation arrows are visible on the left and right sides of the banner.

Bio Resources (updated yearly)

User Information

- Getting Started
- Accounts and Allocations
- System Information
- User Guides
- CARC User Portal
- CARC User Forum
- CARC OnDemand
- Condo Cluster Program
- Bio Resources**
- Using USC's Cryo-EM Instruments
- Frequently Asked Questions

Bio Resources

Reference genomes, protein and nucleotide sequences databases, and other bio resources are now available on Discovery and Endeavour. If you need a specific release that is not currently included in the pages below, please [submit a help ticket](#) and we will try to make those resources available to you.

Genomes

A set of ready-to-use reference sequences and annotations for commonly analyzed organisms, sourced from iGenomes.

Genbank

The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations.

Genome Taxonomy Database (GTDB)

The Genome Taxonomy Database (GTDB) is an initiative to establish a standardized microbial taxonomy based on genome phylogeny.

Pfam Database

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).

TIGRFAMs

TIGRFAMs is a resource consisting of curated multiple sequence alignments, Hidden Markov Models (HMMs) for protein sequence classification, and associated information designed to support automated annotation of (mostly prokaryotic) proteins.

UniProt

The Universal Protein Resource (UniProt), a collaboration between the European Bioinformatics Institute (EBI), the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource (PIR)

Bio Resources (updated yearly)

<https://carc.usc.edu/user-information/bio-resources>

- **Genomes** - reference sequences and annotations for commonly analyzed organisms
- **Genbank** - collection of all public nucleotide sequences and their protein translations
- **Genome Taxonomy Database (GTDB)** - an initiative to establish a standardized microbial taxonomy based on genome phylogeny
- **Pfam Database** - large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)
- **TIGRFAMs** - curated multiple sequence alignments, Hidden Markov Models (HMMs) for protein sequence classification
- **UniProt** – UniProtKB (curated protein information), UniRef (closely related sequences), UniParc (all protein sequences, consisting only of unique identifiers and sequences)

Bio Resources (command line access)

<https://carc.usc.edu/user-information/bio-resources>

- **Biogeotrases** - set of metagenomes, collected under the auspices of the bioGEOTRACES component of the international GEOTRACES program
`/project/biodb/biogeotrases`
- **TaraOceans** - marine microbial metagenomes sampled across space and time
`/project/biodb/taraoceans`
- **Variant Effect Predictor cache** - VEP can use a variety of annotation sources to retrieve the transcript models used to predict consequence types. Cache contains all transcript models, regulatory features and variant data for a species and allows for an offline use of VEP
`/project/biodb/vep-cache`

Log into CARC OnDemand

- Connect to the USC VPN (`connect.usc.edu`) or USC Secure Wifi
- Open the web browser and go to:
 - <https://ondemand.carc.usc.edu>
- Enter your username and password
- Choose an option in Duo-2FA, and confirm your access
- Done. You should see the CARC OnDemand dashboard

CARC OnDemand Dashboard

CARC OnDemand Files ▾ Jobs ▾ Clusters ▾ Interactive Apps ▾ My Interactive Sessions

Help ▾ Logged in as osinski Log Out



Advanced Research Computing
Enabling scientific breakthroughs at scale

OnDemand provides an integrated, single access point for all of your HPC resources.

powered by

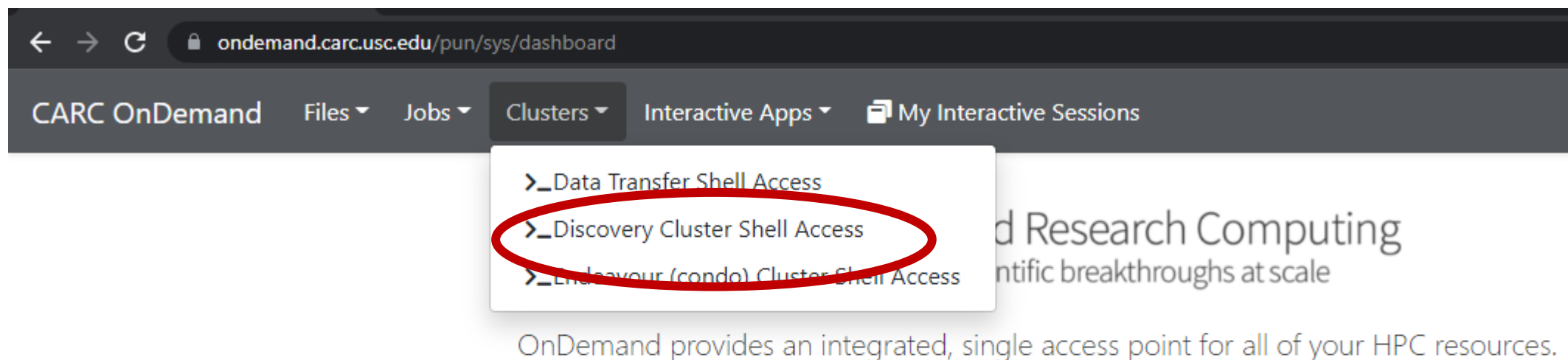
OPEN **OnDemand**

OnDemand version: v1.8.18



Advanced Research Computing
Enabling scientific breakthroughs at scale

CARC OnDemand Dashboard



CARC OnDemand Dashboard

← → ↻ 🔒 ondemand.carc.usc.edu/pun/sys/shell/ssh/discovery.usc.edu

```
Host: discovery.usc.edu
Last login: Tue Jan 24 14:53:26 2023 from 10.23.174.51
-----
Welcome to the Center for Advanced Research Computing (CARC)
at the University of Southern California (USC)
-----

CARC website : https://www.carc.usc.edu
User support : https://www.carc.usc.edu/user-support
User portal  : https://hpcaccount.usc.edu

** Unauthorized use/access is prohibited **

If you log on to this computer system, you acknowledge your awareness of and
concurrency with the USC CARC Acceptable Use Policy. USC will prosecute
violators to the full extent of the law.

-----

[osinski@discovery2 ~]$
```

Copy the text below:

git clone <https://github.com/uschpc/computational-biology-on-carc-bootcamp-2023.git>

and paste it in the terminal by pressing Cmd+V (Shift+Ins on Windows/Linux) keys simultaneously, then press Enter

CARC OnDemand Dashboard

```
Host: discovery.usc.edu
Last login: Tue Jan 24 14:53:26 2023 from 10.23.174.51
-----
Welcome to the Center for Advanced Research Computing (CARC)
at the University of Southern California (USC)
-----

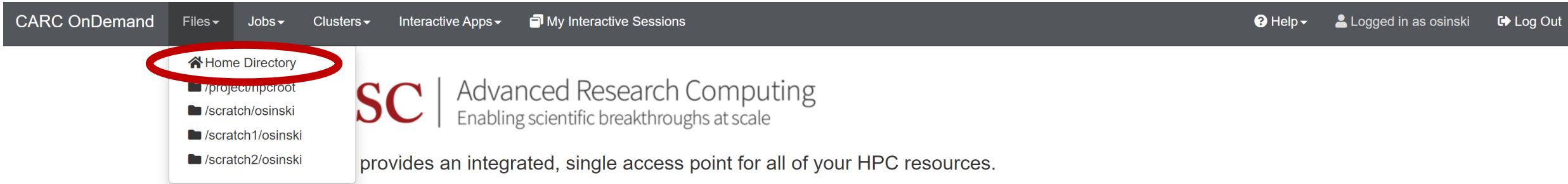
CARC website : https://www.carc.usc.edu
User support : https://www.carc.usc.edu/user-support
User portal  : https://hpcaccount.usc.edu

** Unauthorized use/access is prohibited **

If you log on to this computer system, you acknowledge your awareness of and
concurrence with the USC CARC Acceptable Use Policy. USC will prosecute
violators to the full extent of the law.
-----

[osinski@discovery2 ~]$ git clone https://github.com/uschpc/computational-biology-on-carc.git
Cloning into 'computational-biology-on-carc'...
remote: Enumerating objects: 41, done.
remote: Counting objects: 100% (19/19), done.
remote: Compressing objects: 100% (18/18), done.
remote: Total 41 (delta 7), reused 9 (delta 1), pack-reused 22
Unpacking objects: 100% (41/41), done.
Checking out files: 100% (23/23), done.
[osinski@discovery2 ~]$
```

CARC OnDemand: File Management



The screenshot shows the CARC OnDemand web interface. The top navigation bar is dark gray with the following items: "CARC OnDemand", "Files" (highlighted with a red circle), "Jobs", "Clusters", "Interactive Apps", and "My Interactive Sessions". On the right side of the navigation bar are links for "Help", "Logged in as osinski", and "Log Out". Below the "Files" menu, a dropdown list is visible with the following options: "Home Directory" (highlighted with a red circle), "/project/ncrcroot", "/scratch/osinski", "/scratch1/osinski", and "/scratch2/osinski". The main content area features the USC logo and the text "Advanced Research Computing" and "Enabling scientific breakthroughs at scale". Below this, a paragraph states: "provides an integrated, single access point for all of your HPC resources."

powered by

OPEN **OnDemand**

OnDemand version: v1.8.18

CARC OnDemand: File Management

The screenshot displays the CARC OnDemand File Management interface. The top toolbar includes buttons for 'Go To...', 'Open in Terminal', 'New File', 'New Dir', 'Upload', 'Show Dotfiles', and 'Show Owner/Mode'. The main panel shows the file list for the directory `/home1/osinski/`. The file list includes a table with columns for 'name', 'size', and 'modified date'. The file list also includes a sidebar with a 'Home Directory' tree.

Home Directory

- bin
- bioresources-json-generators
- condor-8.8-bak-end1
- data
- jsons
- lammps-el7
- local_modules
- mhc
- ondemand
- other
- out
- perl5
- relion
- relion-3.1.1
- relion-3.1.1_c10.0.230-k40
- relion-3.1.1_c10.1.243-a60
- relion-3.1.1_c10.1.243-k40x
- slurm-workshop-sep2021
- spack-dev
- venv
- workshop-bioresources

View Edit A-Z Rename/Move Download Copy Paste (Un)Select All Delete

name	size	modified date
..	dir	
bin	dir	10/05/2021
bioresources-json-generators	dir	03/01/2022
condor-8.8-bak-end1	dir	11/03/2021
data	dir	03/04/2022
jsons	dir	01/02/2022
lammps-el7	dir	03/15/2022
local_modules	dir	03/12/2021
mhc	dir	09/16/2021
ondemand	dir	07/21/2021
other	dir	03/22/2022
out	dir	11/11/2021
perl5	dir	09/14/2021
relion	dir	03/11/2021
relion-3.1.1	dir	03/11/2021
relion-3.1.1_c10.0.230-k40	dir	07/12/2021
relion-3.1.1_c10.1.243-a60	dir	03/12/2021
relion-3.1.1_c10.1.243-k40x	dir	07/08/2021
slurm-workshop-sep2021	dir	09/17/2021
spack-dev	dir	03/31/2021
venv	dir	06/04/2021

Create the BLAST Job Script

- Replace **swissprot** with the path to the v5 of swissprot db obtained from <https://carc.usc.edu/user-information/bio-resources/genbank>

```
#!/bin/bash
#SBATCH --nodes 1
#SBATCH --ntasks 10
#SBATCH --partition debug
#SBATCH --time 00:05:00
#SBATCH --account=ttrojan_001
#SBATCH --chdir /home1/ttrojan/computational-biology-on-carc
module purge
module load gcc
module load blast-plus
echo "Start BLAST Job"
blastp -db swissprot -query blast/query.txt -out results/blast/results.txt -num_threads
$SLURM_NTASKS
echo "Finish BLAST Job"
```

Create the BLAST Job Script

USC | USC ITS | Office of the CIO | Office of Research

ENTER SEARCH TERMS



Advanced Research Computing
Enabling scientific breakthroughs at scale

About

Services

User Information

Education & Outreach

News & Events

User Support

User Information

Getting Started

Accounts and Allocations

System Information

User Guides

CARC User Portal

CARC User Forum

CARC OnDemand

Condo Cluster Program

Bio Resources

Genomes

Genbank

Genome Taxonomy Database (GTDB)

Pfam Database

TIGRFAMs

UniProt

Using USC's Cryo-EM Instruments

Frequently Asked Questions

Genbank

GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank has become an important database for research in biological fields and has grown in recent years at an exponential rate by doubling roughly every 18 months.

To access the data below, click the format/version of your choice, then select the database from the drop-down list. Click the clipboard icon to copy the on-disk location. You can then paste it in your submission scripts to use in your analysis.

/project/biodb/genbank/2022-01-01/swissprot



Date Downloaded	FASTA	V4	V5	Database
2021-03-01	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	16S_ribosomal_RN ▾
2022-01-01	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	swissprot ▾



Advanced Research Computing
Enabling scientific breakthroughs at scale

CARC OnDemand: Editing a Job

```
Save /home1/osinski/computational-biology-on-carc/jobs/blast1.sh
1  #!/bin/bash
2  #SBATCH --nodes 1
3  #SBATCH --ntasks 10
4  #SBATCH --partition debug
5  #SBATCH --time=00:05:00
6  #SBATCH --account=ttrojan_001
7  #SBATCH --chdir /home1/ttrojan/computational-biology-on-carc
8  module purge
9  module load gcc
10 module load blast-plus
11 echo "Example blast start"
12 sleep 20
13 blastp -db /project/biodb/genbank/2021-03-01/swissprot -query data/blast/query.txt -out results/blast/results.txt -num_threads $SLURM_NTASKS
14 echo "Example blast end"
15 |
```

Submit a Job – shell

- Submit the job

```
[ttrojan@discovery1 ~]$ sbatch blast1.sh  
Submitted batch job 4773117
```

- Check the status of the job

```
[ttrojan@discovery1 ~]$ squeue -u ttrojan
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST (REASON)
4773117	Main	blast1.j	ttrojan	R	0:02	1	a02-d11

CARC OnDemand: Job Composer



OnDemand provides an integrated, single access point for all of your HPC resources.

CARC OnDemand: Job Creation

CARC OnDemand / Job Composer Jobs Templates

Help

Jobs

+ New Job

☆ Create Template

Edit Files

Job Options

Open Terminal

Submit

Stop

Delete

Show 25 entries

Search:

Created	Name	ID	Cluster	Status
February 25, 2021 11:30am	(default) Simple Sequential Job	2420550	Discovery Cluster	Completed

Showing 1 to 1 of 1 entries

Previous 1 Next

Job Details

2420550

Job Name:

(default) Simple Sequential Job

Submit to:

Discovery Cluster

Account:

Not specified

Script location:

/home1/osinski/ondemand/data/sys/myjobs/projects/default/1

CARC OnDemand: Job Creation

CARC OnDemand / Job ComposerJobsTemplatesHelp

Jobs

+ New Job

From Default Template

From Template

From Specified Path

From Selected Job

Open Terminal

Submit

Stop

Delete

Create Template

Search:

Created	Name	ID	Cluster	Status
February 25, 2021 11:30am	(default) Simple Sequential Job	2420550	Discovery Cluster	Completed

Showing 1 to 1 of 1 entries

Previous

1

Next

Job Details

2420550

Job Name:

(default) Simple Sequential Job

Submit to:

Discovery Cluster

Account:

Not specified

Script location:

/home1/osinski/ondemand/data/sys/myjobs/projects/default/1

Script name:

CARC OnDemand: Job Creation

Path to source (Required)

Source path

Enter the path to a directory on the file system. The contents of this path will be copied to a new workflow.

Job Attributes (Optional)

Name

Script name

Cluster:

Account

Account is an optional field. If not set, the account may be auto-set by the submit filter.

CARC OnDemand: Job Creation

Job was successfully created.

Jobs

+ New Job

☆ Create Template

Edit Files

Job Options

Open Terminal

Submit

Stop

Delete

Show 25 entries

Search:

Created	Name	ID	Cluster	Status
March 22, 2022 8:35pm	blast1		Discovery Cluster	Not Submitted
February 25, 2021 11:30am	(default) Simple Sequential Job	2420550	Discovery Cluster	Completed

Showing 1 to 2 of 2 entries

Previous 1 Next

Job Details

Job Name:

blast1

Submit to:

Discovery Cluster

Account:

osinski_703

Script location:

/home1/osinski/ondemand/data/sys/myjobs/projects/default/2

CARC OnDemand: Editing a Job

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Discovery Cluster

Completed

Previous

1

2

Next

Folder Contents:

blast1.sh

fastqc1.sh

fastqc2.sh

fastqc_numbered_array.sh

fastqc_unnumbered_array.sh

Submit Script

blast1.sh

Script contents:

```
#!/bin/bash  
#SBATCH --nodes 1  
#SBATCH --ntasks 10  
#SBATCH --partition debug  
#SBATCH --time=00:05:00  
#SBATCH --account=hpcroot  
#SBATCH --chdir /home1/osinski/computational-biology-on-carac  
module purge  
module load gcc  
module load blast-plus  
echo "Example blast start"  
sleep 20  
blastp -db /project/biodb/genbank/2021-03-01/swissprot -query data/blast/query.txt -out results/blast  
echo "Example blast end"
```

Open Editor

Open Terminal

Open Dir

CARC OnDemand: Submitting a Job

CARC OnDemand / Job ComposerJobsTemplates

Help

Jobs

+ New Job

☆ Create Template

Edit FilesJob OptionsOpen TerminalSubmitStopDelete

Show 25 entriesSearch:

Created	Name	ID	Cluster	Status
March 22, 2022 8:35pm	blast1		Discovery Cluster	Not Submitted
February 25, 2021 11:30am	(default) Simple Sequential Job	2420550	Discovery Cluster	Completed

Showing 1 to 2 of 2 entriesPrevious1Next

Job Details

Job Name:
blast1

Submit to:
Discovery Cluster

Account:
osinski_703

Script location:
/home1/osinski/ondemand/data/sys/myjobs/projects/default/2

Script name:
blast1.job

CARC OnDemand: Submitting a Job

CARC OnDemand / Job ComposerJobsTemplates

Help

Job was successfully submitted.

Jobs

+ New Job

Create Template

Edit FilesJob OptionsOpen TerminalSubmitStopDelete

Show 25 entriesSearch:

Created	Name	ID	Cluster	Status
March 22, 2022 8:35pm	blast1	7825946	Discovery Cluster	Running
February 25, 2021 11:30am	(default) Simple Sequential Job	2420550	Discovery Cluster	Completed

Showing 1 to 2 of 2 entriesPrevious1Next

Job Details

7825946

Job Name:

blast1

Submit to:

Discovery Cluster

Account:

osinski_703

Script location:

/home1/osinski/ondemand/data/sys/myjobs/projects/default/2

Script name:

CARC OnDemand: Results

CARC OnDemand / Job ComposerJobsTemplatesHelp

Job was successfully submitted.

Jobs

+ New Job

☆ Create Template

Edit FilesJob OptionsOpen TerminalSubmitStopDelete

Show25entries

Search:

Created	Name	ID	Cluster	Status
March 22, 2022 8:35pm	blast1	7825946	Discovery Cluster	Completed
February 25, 2021 11:30am	(default) Simple Sequential Job	2420550	Discovery Cluster	Completed

Showing 1 to 2 of 2 entries

Previous1Next

Job Details

7825946

Job Name:

blast1

Submit to:

Discovery Cluster

Account:

osinski_703

Script location:

/home1/osinski/ondemand/data/sys/myjobs/projects/default/2

Script name:

CARC OnDemand: Results

CARC OnDemand / Job ComposerJobsTemplatesHelp

Job was successfully submitted.

Jobs

+ New Job

☆ Create Template

Edit FilesJob OptionsOpen TerminalSubmitStopDelete

Show25entries

Search:

Created	Name	ID	Cluster	Status
March 22, 2022 8:35pm	blast1	7825946	Discovery Cluster	Completed
February 25, 2021 11:30am	(default) Simple Sequential Job	2420550	Discovery Cluster	Completed

Showing 1 to 2 of 2 entries

Previous1Next

Job Details

7825946

Job Name:

blast1

Submit to:

Discovery Cluster

Account:

osinski_703

Script location:

/home1/osinski/ondemand/data/sys/myjobs/projects/default/2

Script name:

CARC OnDemand: Results

	Type	Name	Size
<input type="checkbox"/>	Folder	data	-
<input type="checkbox"/>	Folder	jobs	-
<input type="checkbox"/>	Folder	presentation	-
<input type="checkbox"/>	Folder	results	-
<input type="checkbox"/>	File	LICENSE	6.88 KB
<input checked="" type="checkbox"/>	File	slurm-13475213.out	217 Bytes

View

Edit

Rename

Download

Delete

CARC OnDemand: Results

```
=====
SLURM_JOB_ID = 7825946
SLURM_JOB_NODELIST = e09-18
TMPDIR = /tmp/SLURM_7825946
=====
"Example blast start"
"Example blast end"
```

CARC OnDemand: Results

The screenshot displays the CARC OnDemand web interface. At the top, a breadcrumb path is shown: `/ home1 / osinski / computational-biology-on-carc / results / blast /`. This path is circled in red. To the right of the path is a button labeled "Change directory". Below the path is a table with columns: "Type", "Name", and "Size". A single file, "results.txt", is listed with a size of "3.05 KB". The checkbox next to the file name is circled in red. A context menu is open for the file, with the "View" option circled in red. The menu also includes "Edit", "Rename", "Download", and "Delete".

Type	Name	Size
<input checked="" type="checkbox"/>	results.txt	3.05 KB

- View
- Edit
- Rename
- Download
- Delete

CARC OnDemand: Results

Query=
Length=15

Sequences producing significant alignments:	Score (Bits)	E Value
Q9JK11.1 RecName: Full=Reticulon-4; AltName: Full=Foocen; AltName...	35.0	0.001
Q99P72.2 RecName: Full=Reticulon-4; AltName: Full=Neurite outgrow...	35.0	0.001
Q9NQC3.2 RecName: Full=Reticulon-4; AltName: Full=Foocen; AltName...	33.9	0.004

>Q9JK11.1 RecName: Full=Reticulon-4; AltName: Full=Foocen; AltName: Full=Glut4 vesicle 20 kDa protein; AltName: Full=Neurite outgrowth inhibitor; Short=Nogo protein [Rattus norvegicus]
Length=1163

Score = 35.0 bits (79), Expect = 0.001, Method: Composition-based stats.
Identities = 15/15 (100%), Positives = 15/15 (100%), Gaps = 0/15 (0%)

Query	1	HYLGLANKSVKDAMA	15
		HYLGLANKSVKDAMA	
Sbjct	1135	HYLGLANKSVKDAMA	1149

>Q99P72.2 RecName: Full=Reticulon-4; AltName: Full=Neurite outgrowth inhibitor; Short=Nogo protein [Mus musculus]
Length=1162

Score = 35.0 bits (79), Expect = 0.001, Method: Composition-based stats.

Genome mapping and tools: Read mapping

- Aim: to find coordinates of reads in the reference genome.
- Challenges:
 - Millions of short sequences
 - Sequences are often paired
 - Errors are not randomly distributed
- Most popular programs are bowtie and bwa (both use Burrows-Wheeler Transform algorithm). Two-step approach:
 - Create an index for the reference genome (one time for one genome).
 - Map reads to the reference genome using this index

Genome mapping and tools – overview I

- FastQC
 - FastQC is a quality control application for high throughput sequence data
 - Checks the quality of their sequence data
 - Generates an HTML report

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Genome mapping and tools – overview II

- bowtie
 - The first version of bowtie [Langmead et al. 2009] is optimal for:
 - short reads (under 50 bp)
 - reads without indels (insertions/deletions)
- bowtie2
 - The second version of bowtie2 [Langmead & Salzberg 2012] is optimal for:
 - long reads (more than 50 bp)
 - reads with indels
 - various alignment options
- Each version has its own index file format (bowtie-build / bowtie2-build tools).
- A popular RNA-seq analysis toolset (tophat, cufflinks) is based on bowtie / bowtie2

<http://bowtie-bio.sourceforge.net>

Genome mapping and tools – overview III

- `bwa`
 - `bwa backtrack` [Li, Durbin 2009]:
 - for short reads (< 100bp)
 - `bwa bwasw` [Li, Durbin 2010]:
 - for long reads (70bp - 1Mbp)
 - short indels
 - `bwa mem` [Li 2013]:
 - for long reads (70bp - 1Mbp)
 - faster and more efficient

Genome mapping and tools – overview IV

- samtools package - A set of utilities for processing SAM/BAM files
- samtools view
 - convert a bam file into a sam file - `samtools view sample.bam > sample.sam`
 - Convert a sam file into a bam file - `samtools view -bS sample.sam > sample.bam`
 - Extract all the reads aligned to the range specified. An index of the input file is required
`samtools view -h -b sample_sorted.bam "chr1:10-13" > tiny_sorted.bam`
- samtools sort
`samtools sort unsorted_in.bam sorted_out`
- samtools index
`samtools index sorted.bam (creates an index file, sorted.bam.bai)`

<http://samtools.sourceforge.net>

Genome mapping and tools – overview IV

- samtools package - A set of utilities for processing SAM/BAM files

- samtools view

- convert a bam file into a sam file - `samtools view sample.bam > sample.sam`

- Convert a sam file into a bam file - `samtools view -bS sample.sam > sample.bam`

- Extract all the reads aligned to the range specified. An index of the input file is required

- `samtools view -h -b sample_sorted.bam "chr1:10-13" > tiny_sorted.bam`

- samtools sort

- `samtools sort unsorted_in.bam sorted_out`

- samtools index

- `samtools index sorted.bam (creates an index file, sorted.bam.bai)`

<http://samtools.sourceforge.net>

-b: output BAM

-S: read SAM

add a proper header

Genome mapping and tools – overview IV

- **samtools flagstat** – report basic statistics

```
samtools flagstat sample.bam
```

An example of output:

```
4198456 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
4022089 + 0 mapped (95.80%:-nan%)
4198456 + 0 paired in sequencing
2099228 + 0 read1
2099228 + 0 read2
3796446 + 0 properly paired (90.42%:-nan%)
4013692 + 0 with itself and mate mapped
8397 + 0 singletons (0.20%:-nan%)
167574 + 0 with mate mapped to a different chr
72008 + 0 with mate mapped to a different chr (mapQ>=5)
```

- **samtools faidx** – index a FASTA file

```
samtools faidx ref.fasta (creates an index file ref.fasta.fai)
```

- **samtools merge** – merge several BAM files into one

```
samtools merge out.bam in1.bam in2.bam
```

Genome mapping and tools – overview V

- BedTools package
 - `bamtobed` - Convert BAM alignments to BED (& other) formats
 - `bamtofastq` - Convert BAM records to FASTQ records
 - `bedtobam` - Convert intervals to BAM records
 - `closest` - Find the closest, potentially non-overlapping interval
 - `complement` - Extract intervals `_not_` represented by an interval file
 - `coverage` - Compute the coverage over defined intervals
 - `genomecov` - Compute the coverage over an entire genome
 - `getfasta` - Use intervals to extract sequences from a FASTA file
 - `intersect` - Find overlapping intervals in various ways
 - `shuffle` - Randomly redistribute intervals in a genome
 - `sort` - Order the intervals in a file

Genome mapping and tools – overview V

`bedtools intersect`

- Report the intervals that represent overlaps between your two files:

```
bedtools intersect -a cpg.bed -b exons.bed
```

- Report the original feature in each file:

```
bedtools intersect -a cpg.bed -b exons.bed -wa -wb
```

- How many base pairs of overlap were there?

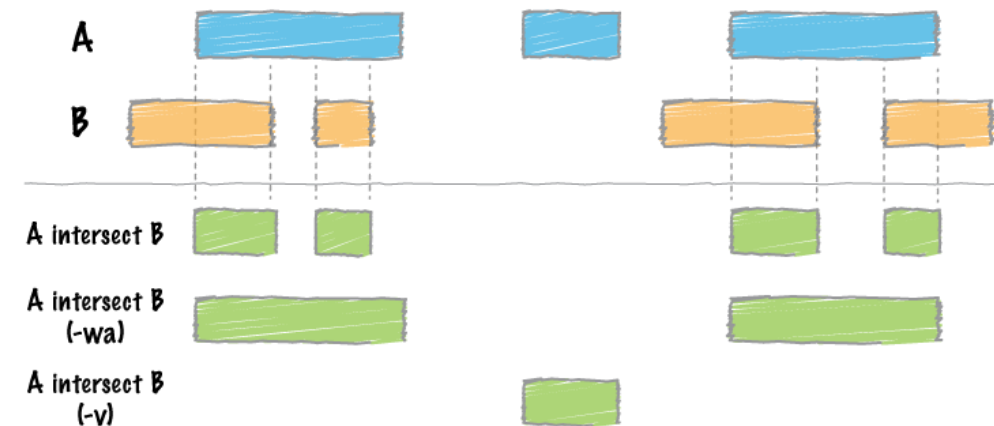
```
bedtools intersect -a cpg.bed -b exons.bed -wo
```

- Counting the number of overlapping features:

```
bedtools intersect -a cpg.bed -b exons.bed -c
```

- Find features that DO NOT overlap:

```
bedtools intersect -a cpg.bed -b exons.bed -v
```



Exercise

- There are paired reads of some DNA sequencing experiment of the human sample:

`computational-biology-on-carc/data/R1.fastq.gz`

`computational-biology-on-carc/data/R2.fastq.gz`

- You will study some particular region of the human genome
- Map reads to the human reference genome (version hg19 – find path on our Bio Resources)
- Extract reads that map to your region only
- Upload the reads to UCSC genome browser as a custom track
- count the number of insertions and deletions in SAM file

How To: Mapping

- load bowtie2 program:

```
module purge
module load gcc
module load bowtie2
```

- Copy sequence of a chromosome your region is located at as a FASTA file
 - find the path on our website in Homo sapiens > UCSC > hg19 > Chromosome 21
 - <https://carc.usc.edu/user-information/bio-resources/reference-genomes>
 - Add chr*.fa at the end of the path
 - `ln -s path_above ~/computational-biology-on-carc/results/read-mapping/`

- Map reads to this chromosome using `bowtie2` with the standard parameters

Don't forget to make an index (`bowtie-build2`) of the chromosome before mapping!

- You will get a SAM file as an output, convert to BAM (`samtools view`)
- Count the number of insertions and deletions in SAM file (use `cut` for field 6, and `grep`)

How To: Extracting reads

- load BedTools:

```
module purge
```

```
module load gcc
```

```
module load bedtools2
```

- Create a tab-delimited BED file with the coordinates of your region:

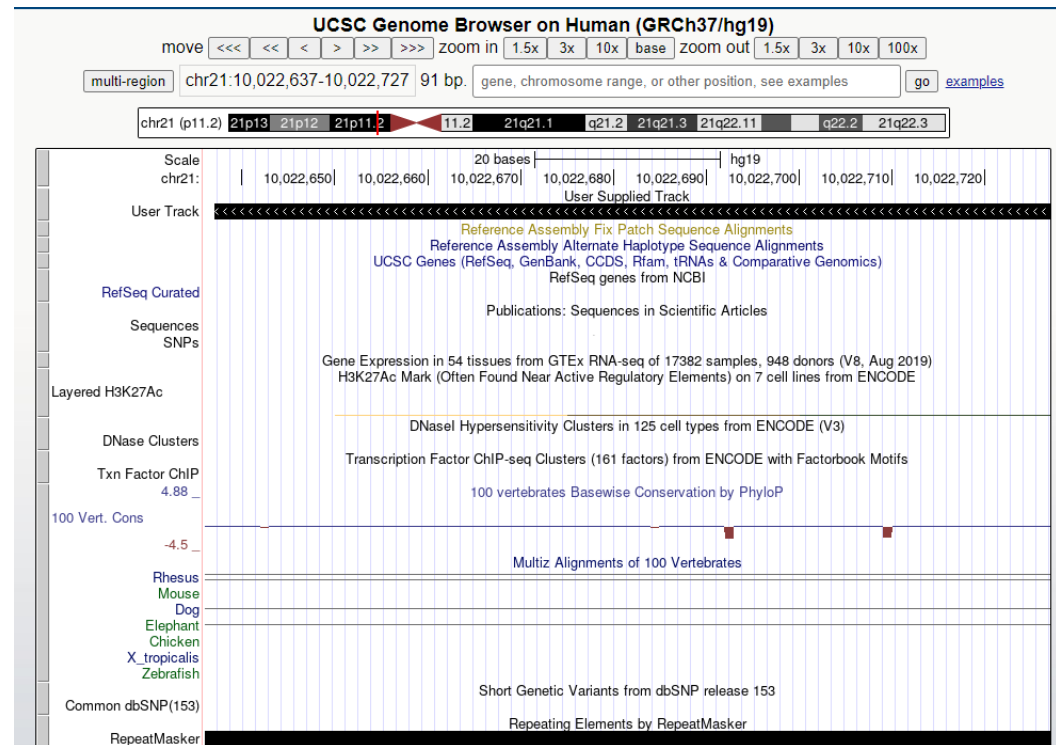
```
chr21 10000000 20000000
```

- Convert SAM file with mapped reads to BAM file using `samtools view`
- Use `bedtools intersect` to extract the reads from the BAM file

You'll need a BED file to upload the result to UCSC genome browser, so figure out how to make `bedtools intersect` to produce an output in BED format.

How To: UCSC custom track

- Upload the BED file to UCSC genome browser
- ‘Add custom track’ button → Choose file → Submit → Go



Resources

- CARC home page
 - <https://carc.usc.edu>
- Bio Resources at CARC
 - <https://carc.usc.edu/user-information/bio-resources>
- CARC User Forum
 - <https://hpc-discourse.usc.edu/categories>
- SLURM tutorials
 - <https://slurm.schedmd.com/tutorials.html>
- SLURM quick reference
 - <https://slurm.schedmd.com/pdfs/summary.pdf>

Reference material:

- HPCBio (Holmes J., Clark L., Drnevicch J., Valizadegan N.)
- CNRG (Davidson D., Leigh J.)
- Skoltech (Khrameeva E.)

Mapping exercise: Answer

```
#!/bin/bash
#SBATCH --partition main
#SBATCH --nodes 1
#SBATCH --ntasks 4
#SBATCH --time 01:00:00
#SBATCH --mem 4g
#SBATCH --account=ttrojan_001
#SBATCH --chdir /home1/ttrojan/computational-biology-on-carc
module purge
module load gcc
module load bowtie2
module load samtools
module load bedtools2
mkdir results/read-mapping
cp data/R*.gz results/read-mapping
gunzip results/read-mapping/R1.fastq.gz
gunzip results/read-mapping/R2.fastq.gz
ln -s /project/biodb/genomes/Homo_sapiens/UCSC/hg19/Sequence/Chromosomes/chr21.fa results/read-mapping/
bowtie2-build --threads $SLURM_NTASKS results/read-mapping/chr21.fa results/read-mapping/chr21index
bowtie2 --threads $SLURM_NTASKS -x results/read-mapping/chr21index -q results/read-mapping/R1.fastq > results/read-mapping/R1.sam
bowtie2 --threads $SLURM_NTASKS -x results/read-mapping/chr21index -q results/read-mapping/R2.fastq > results/read-mapping/R2.sam
samtools view results/read-mapping/R1.bam | cut -f 6 | grep -c 'D' > results/read-mapping/R1.no_of_deletions.txt
samtools view results/read-mapping/R1.bam | cut -f 6 | grep -c 'I' > results/read-mapping/R1.no_of_insertions.txt
samtools view results/read-mapping/R2.bam | cut -f 6 | grep -c 'D' > results/read-mapping/R2.no_of_deletions.txt
samtools view results/read-mapping/R2.bam | cut -f 6 | grep -c 'I' > results/read-mapping/R2.no_of_insertions.txt
samtools view -bS results/read-mapping/R1.sam > results/read-mapping/R1.bam
samtools view -bS results/read-mapping/R2.sam > results/read-mapping/R2.bam
samtools sort results/read-mapping/R1.bam > results/read-mapping/R1_sorted.bam
samtools sort results/read-mapping/R2.bam > results/read-mapping/R2_sorted.bam
samtools view -h -b results/read-mapping/R1_sorted.bam "chr21:10000000-20000000" > results/read-mapping/R1_sorted_region.bam
samtools view -h -b results/read-mapping/R2_sorted.bam "chr21:10000000-20000000" > results/read-mapping/R2_sorted_region.bam
bamToBed -i results/read-mapping/R1_sorted_region.bam > results/read-mapping/R1_sorted_region.bed
bamToBed -i results/read-mapping/R2_sorted_region.bam > results/read-mapping/R2_sorted_region.bed
bedtools intersect -a results/read-mapping/R1_sorted_region.bed -b results/read-mapping/R2_sorted_region.bed > results/read-mapping/reads.bed
```