Roswell OnDemand
A One-Stop Shop for High Performance Computing

ROSWELL PARK® COMPREHENSIVE CANCER CENTER

L. Shawn Matott, PhD
Senior Systems Analyst
IT Research Computing

Training Objectives

- To provide an overview of Roswell Park's IT Research Computing (IT RC) group
- To provide an introduction to Roswell Park's HPC OnDemand interface

About IT Research Computing

- Mission Statement

  To promote the use of advanced computing and computational science to enable and accelerate research throughout the Roswell Park campus.

- Expertise and Services
  - HPC assets
  - Training
  - Grant Support
  - IT Referrals

Computing Assets

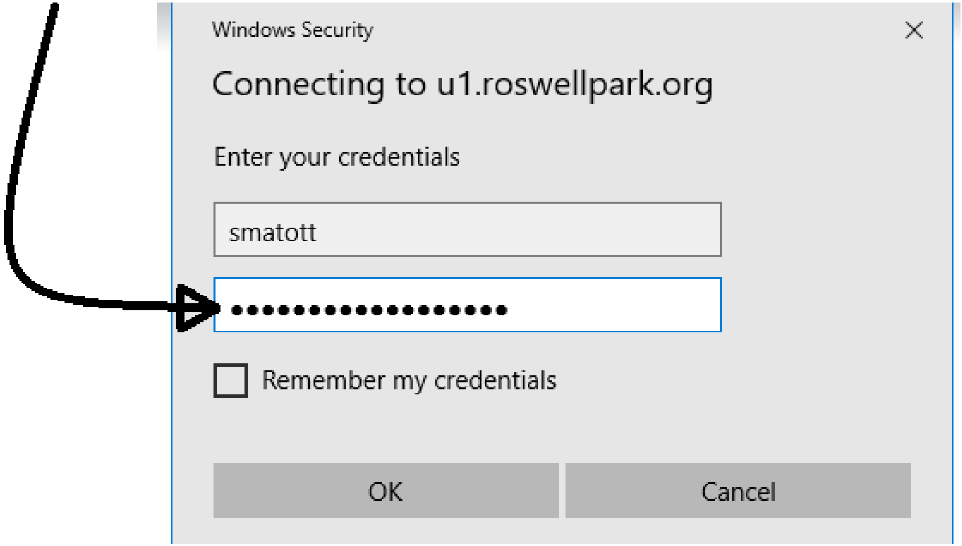
- Roswell Park HPCC
  - 1,600 processors, 450 TB Lustre storage
  - Software modules include: R/Bioconductor, Python, Keras, and TensorFlow, many bio-informatics packages

A look inside Roswell's HPC data center.

  - OnDemand web interface http://u1.roswellpark.org/pun/sys/dashboard/
  - e-mail ServiceDesk@RoswellPark.org to request access

OnDemand Login

http://u1.roswellpark.org/pun/sys/dashboard/



Windows Security

Connecting to u1.roswellpark.org

Enter your credentials

smatott

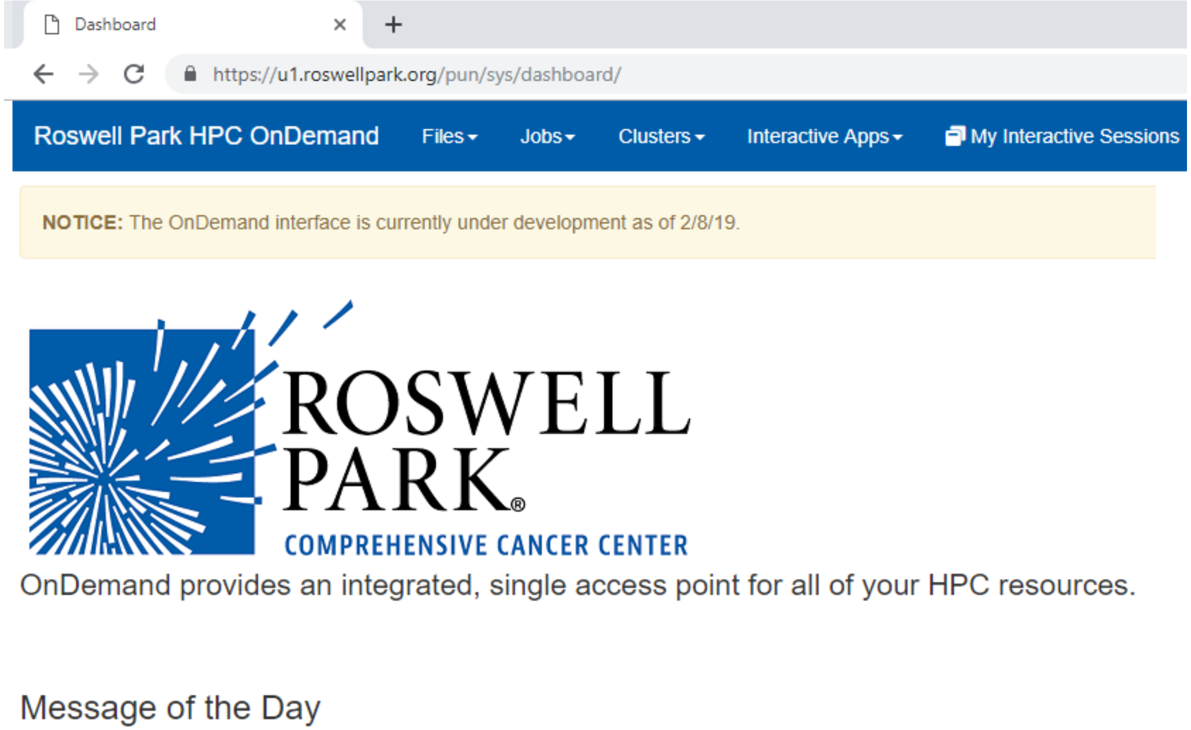
••••••••••••••••••

□ Remember my credentials

- Note: *your OnDemand user ID and password may not the same as your main Roswell ID and password*

OnDemand Dashboard

- Files
  - Upload, download, edit, move, delete etc.
- Jobs
  - Configure, submit, & monitor jobs
- Clusters
  - Linux CLI
- Interactive Apps
  - Launch programs with GUIs


Dashboard
https://u1.roswellpark.org/pun/sys/dashboard/
Roswell Park HPC OnDemand
Files
Jobs
Clusters
Interactive Apps
My Interactive Sessions
ROSWELL PARK
COMPREHENSIVE CANCER CENTER

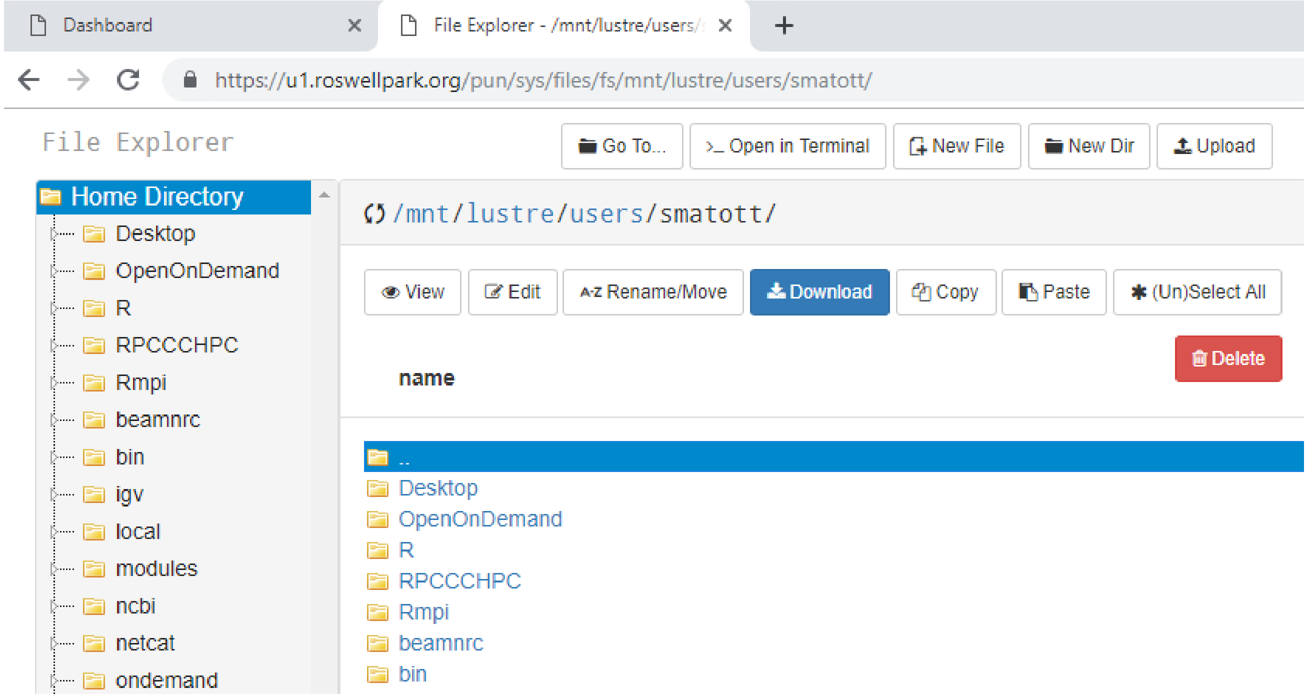
NOTICE: The OnDemand interface is currently under development as of 2/8/19.

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

Message of the Day

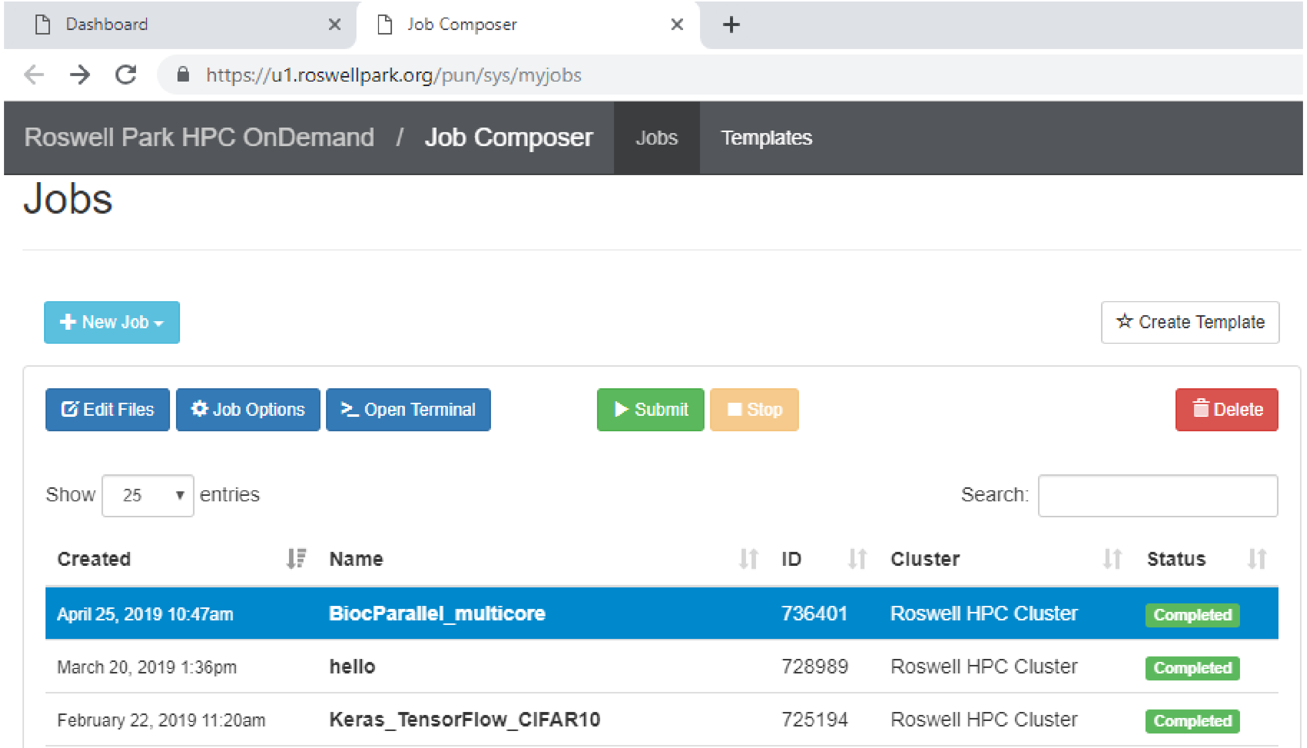
OnDemand - Files

- File Explorer
  - Transfers via upload/down-load or drag and drop
- Editor
  - Language aware coloring (e.g. Python comments, etc.)



OnDemand – Compute Jobs

- Composer
  - Create and configure a new job
  - Use templates to minimize effort
  - Monitor or alter current jobs
- Active Jobs
  - Check job queues

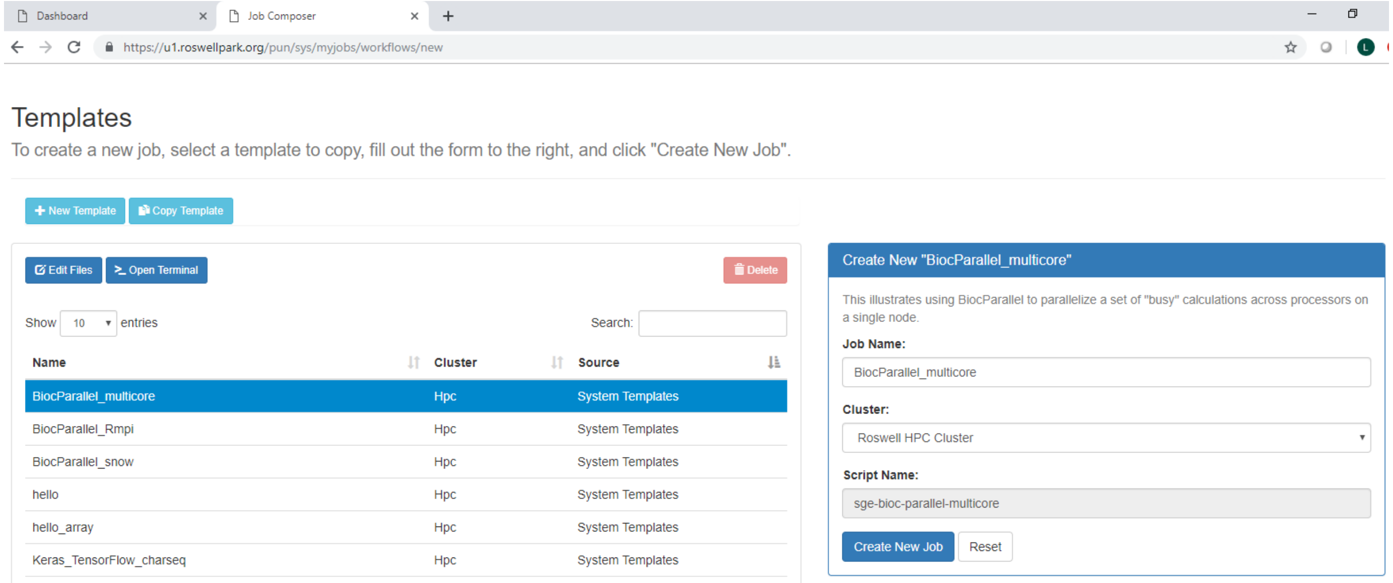


Jobs

Created	Name	ID	Cluster	Status
April 25, 2019 10:47am	BiocParallel_multicore	736401	Roswell HPC Cluster	Completed
March 20, 2019 1:36pm	hello	728989	Roswell HPC Cluster	Completed
February 22, 2019 11:20am	Keras_TensorFlow_CIFAR10	725194	Roswell HPC Cluster	Completed

OnDemand – Job Templates

- Templates
  - Searchable
  - IT RC (i.e. Shawn) can create new templates on request
  - Users can also create and manage their own templates



https://u1.roswellpark.org/pun/sys/myjobs/workflows/new

Templates

To create a new job, select a template to copy, fill out the form to the right, and click "Create New Job".

Show 10 entries   Search:

Name	Cluster	Source
BiocParallel_multicore	Hpc	System Templates
BiocParallel_Rmpi	Hpc	System Templates
BiocParallel_snow	Hpc	System Templates
hello	Hpc	System Templates
hello_array	Hpc	System Templates
Keras_TensorFlow_charseq	Hpc	System Templates

Create New "BiocParallel_multicore"

This illustrates using BiocParallel to parallelize a set of "busy" calculations across processors on a single node.

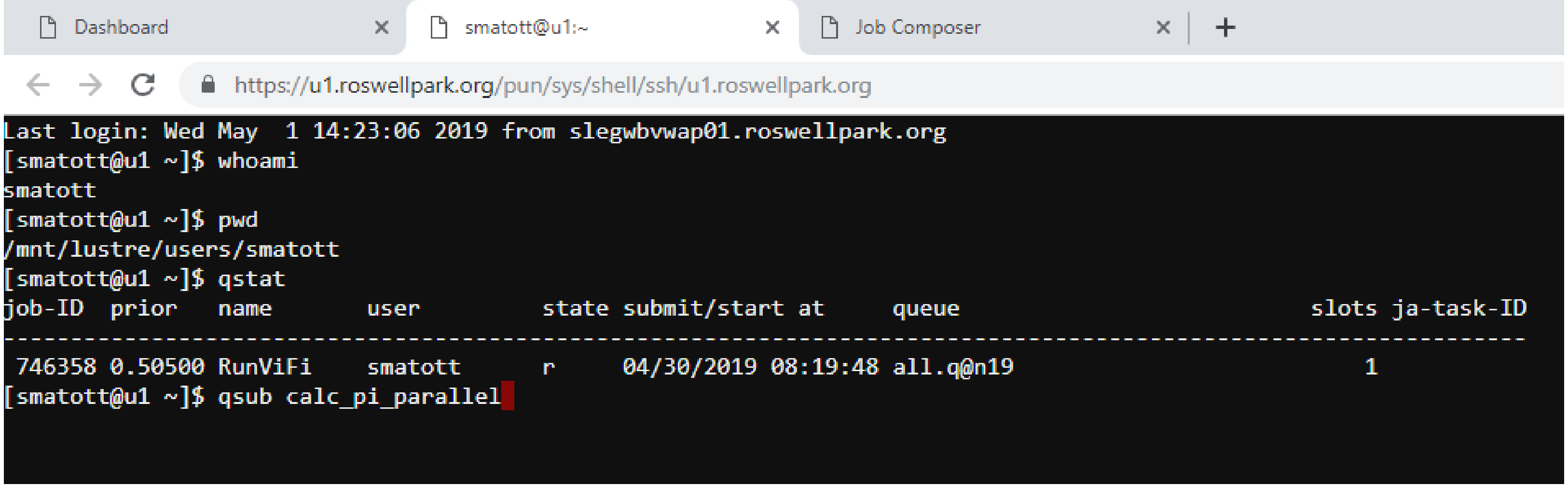
Job Name: BiocParallel_multicore

Cluster: Roswell HPC Cluster

Script Name: sge-bioc-parallel-multicore

OnDemand – Cluster CLI

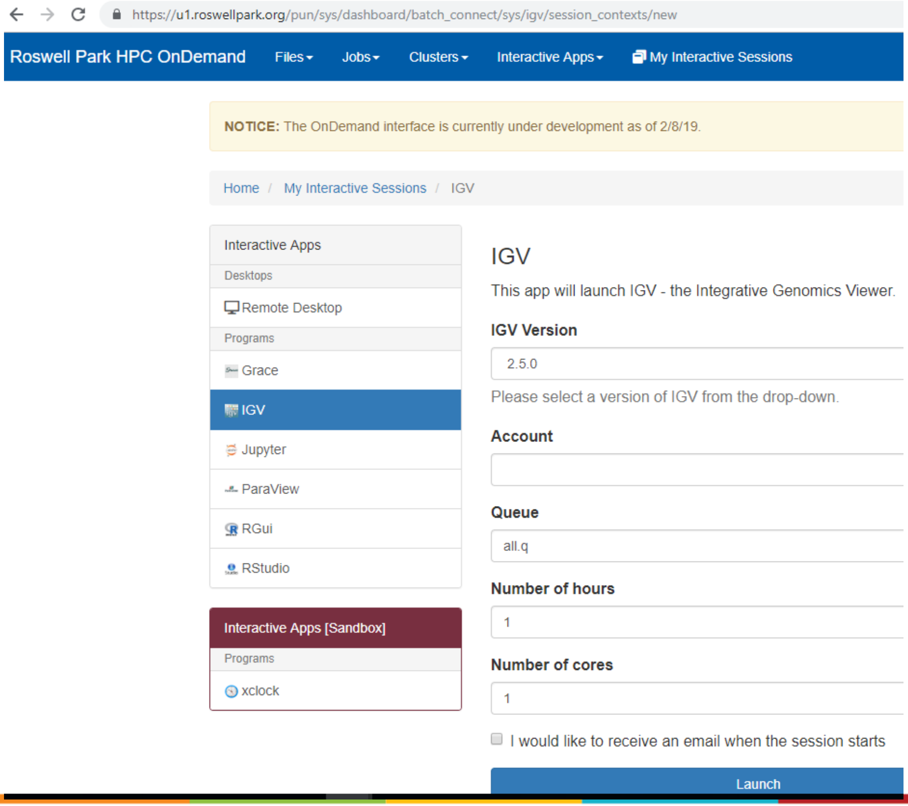
- Provides a bash command line interface on front-end node (u1)
- Handy for those who are familiar with conventional HPC access paradigm (e.g. putty, ssh)


Dashboard
smatott@u1:~
Job Composer
https://u1.roswellpark.org/pun/sys/shell/ssh/u1.roswellpark.org

```
Last login: Wed May  1 14:23:06 2019 from slegwbvwap01.roswellpark.org
[smatott@u1 ~]$ whoami
smatott
[smatott@u1 ~]$ pwd
/mnt/lustre/users/smatott
[smatott@u1 ~]$ qstat
job-ID  prior   name       user         state submit/start at     queue                          slots ja-task-ID
-----------------------------------------------------------------------------------------------------------------
 746358 0.50500 RunViFi    smatott      r     04/30/2019 08:19:48 all.q@n19                          1
[smatott@u1 ~]$ qsub calc_pi_parallel
```

OnDemand – Interactive Apps

- Special compute jobs for launching interactive GUI-based applications
- Remote Desktop + Individual Apps
- OpenGL codes can benefit from GPU hardware acceleration (viz nodes)



https://u1.roswellpark.org/pun/sys/dashboard/batch_connect/sys/igv/session_contexts/new

NOTICE: The OnDemand interface is currently under development as of 2/8/19.

IGV

This app will launch IGV - the Integrative Genomics Viewer.

IGV Version

2.5.0

Please select a version of IGV from the drop-down.

Account

Queue

all.q

Number of hours

1

Number of cores

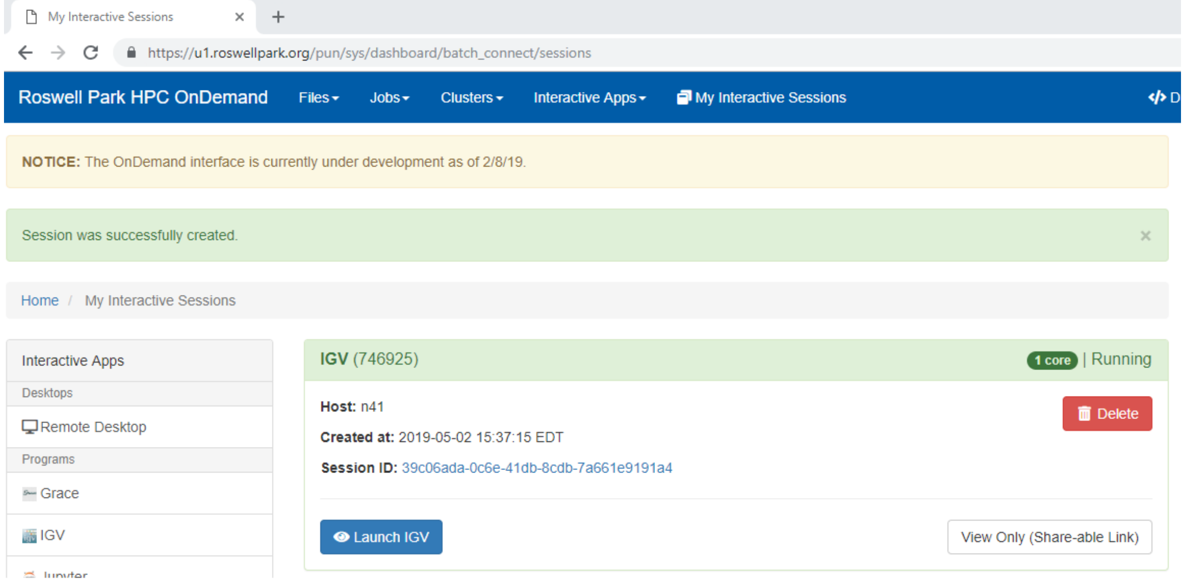
1

□ I would like to receive an email when the session starts

Use the app window at left to select an app and submit request via “Launch” button.

OnDemand – Interactive Sessions

- Interactive apps are compute jobs!
- Can close browser window and reconnect without loss of work.



My Interactive Sessions
https://u1.roswellpark.org/pun/sys/dashboard/batch_connect/sessions
Roswell Park HPC OnDemand
Files
Jobs
Clusters
Interactive Apps
My Interactive Sessions

NOTICE: The OnDemand interface is currently under development as of 2/8/19.

Session was successfully created.

IGV (746925) 1 core | Running

Host: n41

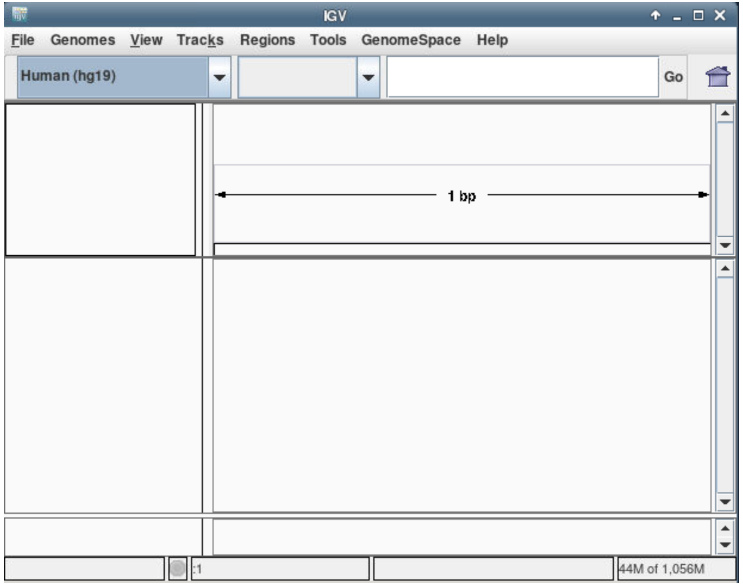
Created at: 2019-05-02 15:37:15 EDT

Session ID: 39c06ada-0c6e-41db-8cdb-7a661e9191a4

Once the job starts the window will refresh and offer a “Launch [app]” button. Click it to open the app in a browser window.

OnDemand – Interactive Sessions

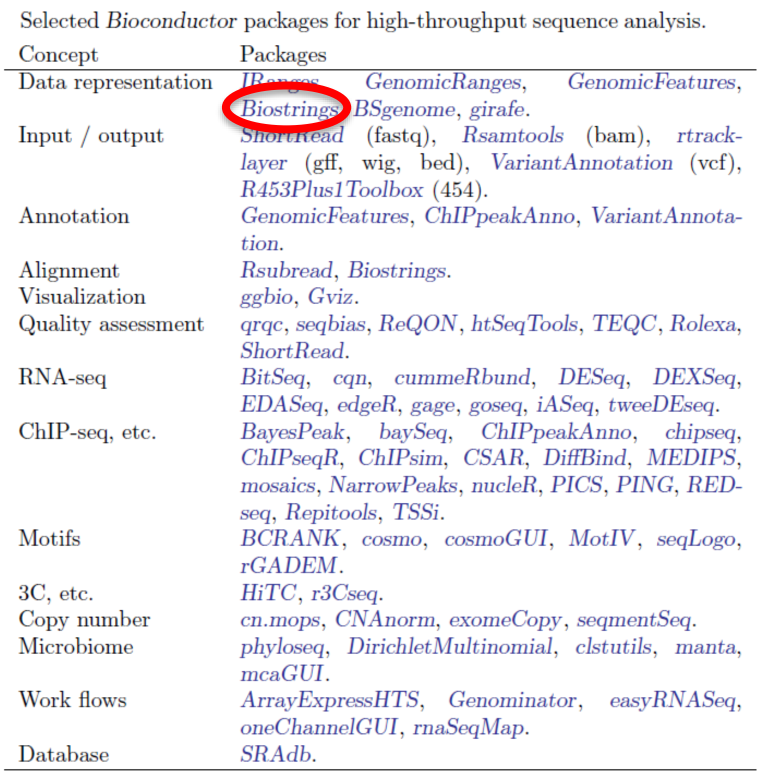
- Interactive apps are compute jobs!
- Can close browser window and reconnect without loss of work.



In this example, a remote desktop is opened and pre-populated with an instance of the IGV genome viewer.

About Bioconductor

- A set of 1700+ R packages for high-throughput sequence analysis
  - RNA-seq, ChIP-seq, Annotation, etc.
- www.Bioconductor.org
- Open source and extensively documented and supported
  - Vignettes, user-groups, mailing lists, etc.
- The Bioconductor project is led by Roswell's own Dr. Martin Morgan!

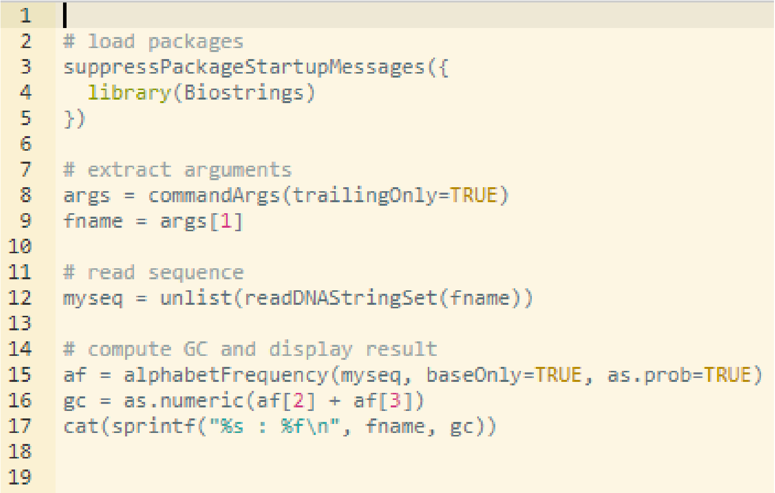


Selected *Bioconductor* packages for high-throughput sequence analysis.

Concept	Packages
Data representation	*IRanges*, *GenomicRanges*, *GenomicFeatures*, *Biostrings*, *BSgenome*, *girafe*.
Input / output	*ShortRead* (fastq), *Rsamtools* (bam), *rtracklayer* (gff, wig, bed), *VariantAnnotation* (vcf), *R453Plus1Toolbox* (454).
Annotation	*GenomicFeatures*, *ChIPpeakAnno*, *VariantAnnotation*.
Alignment	*Rsubread*, *Biostrings*.
Visualization	*ggbio*, *Gviz*.
Quality assessment	*qrqc*, *seqbias*, *ReQON*, *htSeqTools*, *TEQC*, *Rolexa*, *ShortRead*.
RNA-seq	*BitSeq*, *cqn*, *cummeRbund*, *DESeq*, *DEXSeq*, *EDASeq*, *edgeR*, *gage*, *goseq*, *iASeq*, *tweeDEseq*.
ChIP-seq, etc.	*BayesPeak*, *baySeq*, *ChIPpeakAnno*, *chipseq*, *ChIPseqR*, *ChIPsim*, *CSAR*, *DiffBind*, *MEDIPS*, *mosaics*, *NarrowPeaks*, *nucleR*, *PICS*, *PING*, *REDseq*, *Repitools*, *TSSi*.
Motifs	*BCRANK*, *cosmo*, *cosmoGUI*, *MotIV*, *seqLogo*, *rGADEM*.
3C, etc.	*HiTC*, *r3Cseq*.
Copy number	*cn.mops*, *CNAnorm*, *exomeCopy*, *seqmentSeq*.
Microbiome	*phyloseq*, *DirichletMultinomial*, *clstutils*, *manta*, *mcaGUI*.
Work flows	*ArrayExpressHTS*, *Genominator*, *easyRNASeq*, *oneChannelGUI*, *rnaSeqMap*.
Database	*SRAdb*.

OnDemand – Biostrings example

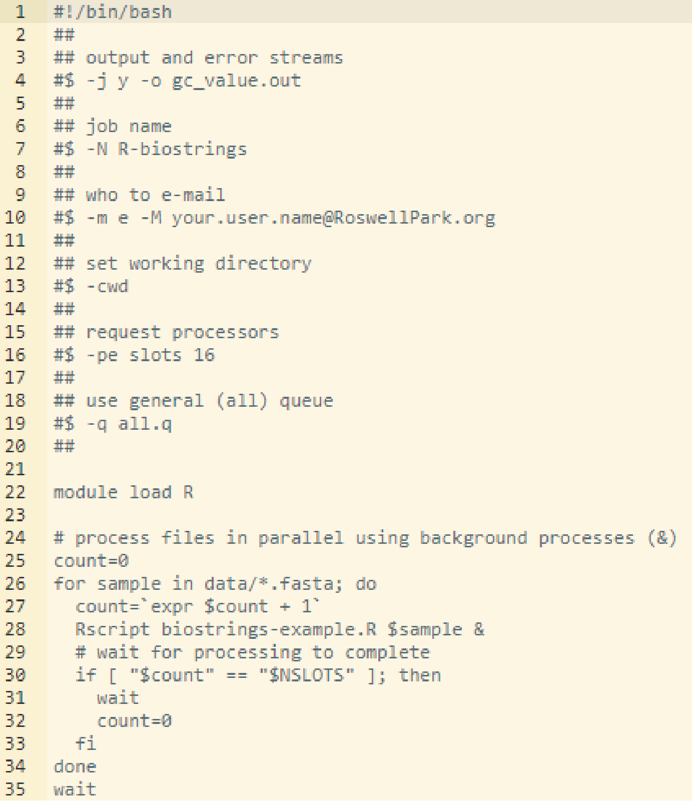
- Dashboard → Jobs→Job Composer → Templates
- Search for “Biostrings”
- Select the “Biostrings-gc-content” template
    - Compute GC content of multiple samples
- Click “Create New Job”
    - Drops you into the “Jobs” interface
- Explore the job details area
- Click on the “.R” file to open in editor
    - contains the Bioconductor code



```r

# load packages
suppressPackageStartupMessages({
  library(Biostrings)
})

# extract arguments
args = commandArgs(trailingOnly=TRUE)
fname = args[1]

# read sequence
myseq = unlist(readDNAStringSet(fname))

# compute GC and display result
af = alphabetFrequency(myseq, baseOnly=TRUE, as.prob=TRUE)
gc = as.numeric(af[2] + af[3])
cat(sprintf("%s : %f\n", fname, gc))


```

OnDemand – Biostrings example

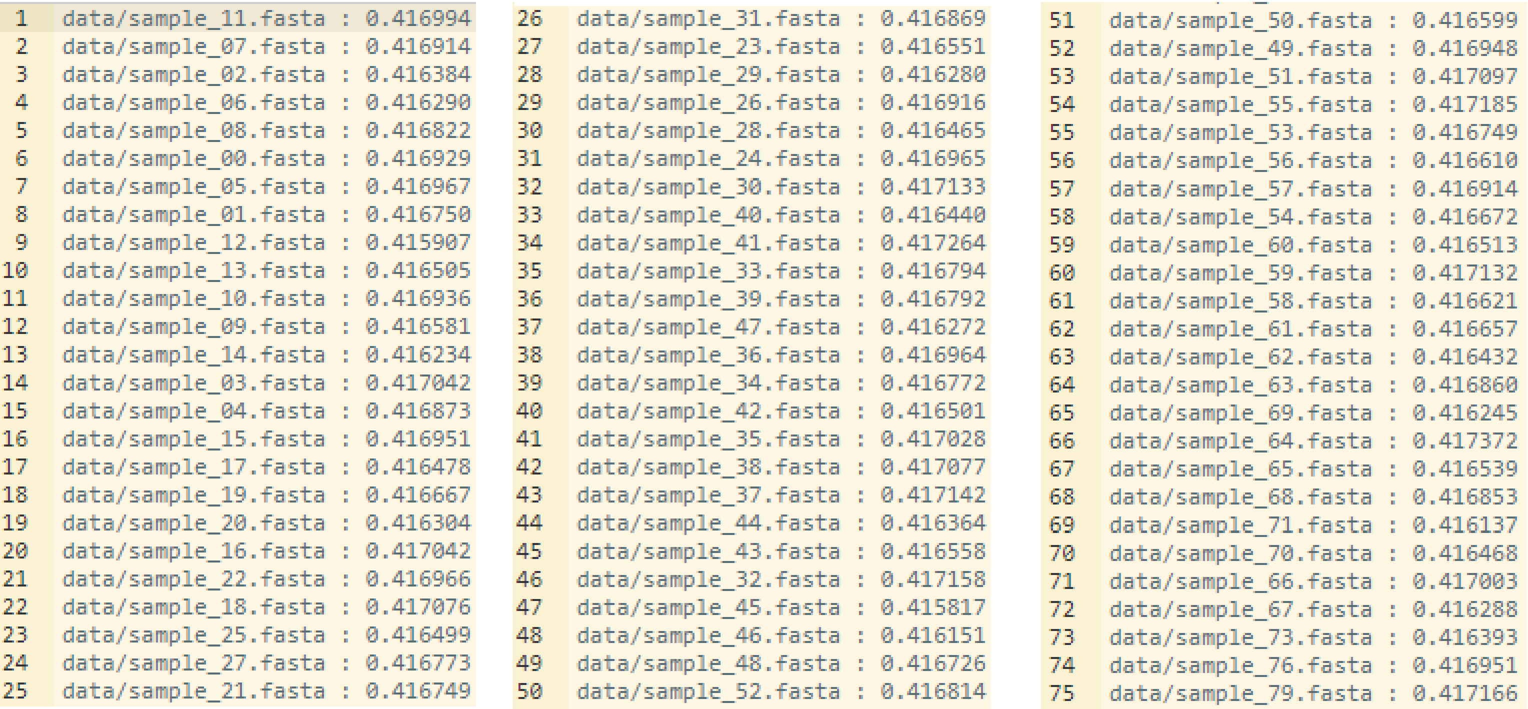
- Open the sge-R-biostrings script in the editor
- Notice the coarse “file-level” parallelization technique
  - Applicable any pipeline that must operate on many independent files
  - Good practice tip: start small, build up from there
- Job Composer → Submit
- Wait for job to complete
- Open gc_values.out



```
#!/bin/bash
##
## output and error streams
#$ -j y -o gc_value.out
##
## job name
#$ -N R-biostrings
##
## who to e-mail
#$ -m e -M your.user.name@RoswellPark.org
##
## set working directory
#$ -cwd
##
## request processors
#$ -pe slots 16
##
## use general (all) queue
#$ -q all.q
##

module load R

# process files in parallel using background processes (&)
count=0
for sample in data/*.fasta; do
  count=`expr $count + 1`
  Rscript biostrings-example.R $sample &
  # wait for processing to complete
  if [ "$count" == "$NSLOTS" ]; then
    wait
    count=0
  fi
done
wait
```

OnDemand – Biostrings example

gc_values.out



```
data/sample_11.fasta : 0.416994
data/sample_07.fasta : 0.416914
data/sample_02.fasta : 0.416384
data/sample_06.fasta : 0.416290
data/sample_08.fasta : 0.416822
data/sample_00.fasta : 0.416929
data/sample_05.fasta : 0.416967
data/sample_01.fasta : 0.416750
data/sample_12.fasta : 0.415907
data/sample_13.fasta : 0.416505
data/sample_10.fasta : 0.416936
data/sample_09.fasta : 0.416581
data/sample_14.fasta : 0.416234
data/sample_03.fasta : 0.417042
data/sample_04.fasta : 0.416873
data/sample_15.fasta : 0.416951
data/sample_17.fasta : 0.416478
data/sample_19.fasta : 0.416667
data/sample_20.fasta : 0.416304
data/sample_16.fasta : 0.417042
data/sample_22.fasta : 0.416966
data/sample_18.fasta : 0.417076
data/sample_25.fasta : 0.416499
data/sample_27.fasta : 0.416773
data/sample_21.fasta : 0.416749
```

```
data/sample_31.fasta : 0.416869
data/sample_23.fasta : 0.416551
data/sample_29.fasta : 0.416280
data/sample_26.fasta : 0.416916
data/sample_28.fasta : 0.416465
data/sample_24.fasta : 0.416965
data/sample_30.fasta : 0.417133
data/sample_40.fasta : 0.416440
data/sample_41.fasta : 0.417264
data/sample_33.fasta : 0.416794
data/sample_39.fasta : 0.416792
data/sample_47.fasta : 0.416272
data/sample_36.fasta : 0.416964
data/sample_34.fasta : 0.416772
data/sample_42.fasta : 0.416501
data/sample_35.fasta : 0.417028
data/sample_38.fasta : 0.417077
data/sample_37.fasta : 0.417142
data/sample_44.fasta : 0.416364
data/sample_43.fasta : 0.416558
data/sample_32.fasta : 0.417158
data/sample_45.fasta : 0.415817
data/sample_46.fasta : 0.416151
data/sample_48.fasta : 0.416726
data/sample_52.fasta : 0.416814
```

```
data/sample_50.fasta : 0.416599
data/sample_49.fasta : 0.416948
data/sample_51.fasta : 0.417097
data/sample_55.fasta : 0.417185
data/sample_53.fasta : 0.416749
data/sample_56.fasta : 0.416610
data/sample_57.fasta : 0.416914
data/sample_54.fasta : 0.416672
data/sample_60.fasta : 0.416513
data/sample_59.fasta : 0.417132
data/sample_58.fasta : 0.416621
data/sample_61.fasta : 0.416657
data/sample_62.fasta : 0.416432
data/sample_63.fasta : 0.416860
data/sample_69.fasta : 0.416245
data/sample_64.fasta : 0.417372
data/sample_65.fasta : 0.416539
data/sample_68.fasta : 0.416853
data/sample_71.fasta : 0.416137
data/sample_70.fasta : 0.416468
data/sample_66.fasta : 0.417003
data/sample_67.fasta : 0.416288
data/sample_73.fasta : 0.416393
data/sample_76.fasta : 0.416951
data/sample_79.fasta : 0.417166
```

Questions?

- Contact IT RC through:
  - The ServiceNow portal (RequestIT, FixIT):

    roswellpark.service-now.com
  - e-mail Dr. Matott:

    Loren.Matott@RoswellPark.org
  - The IT RC website:

    roswellpark.org/research/departments/it-research-computing