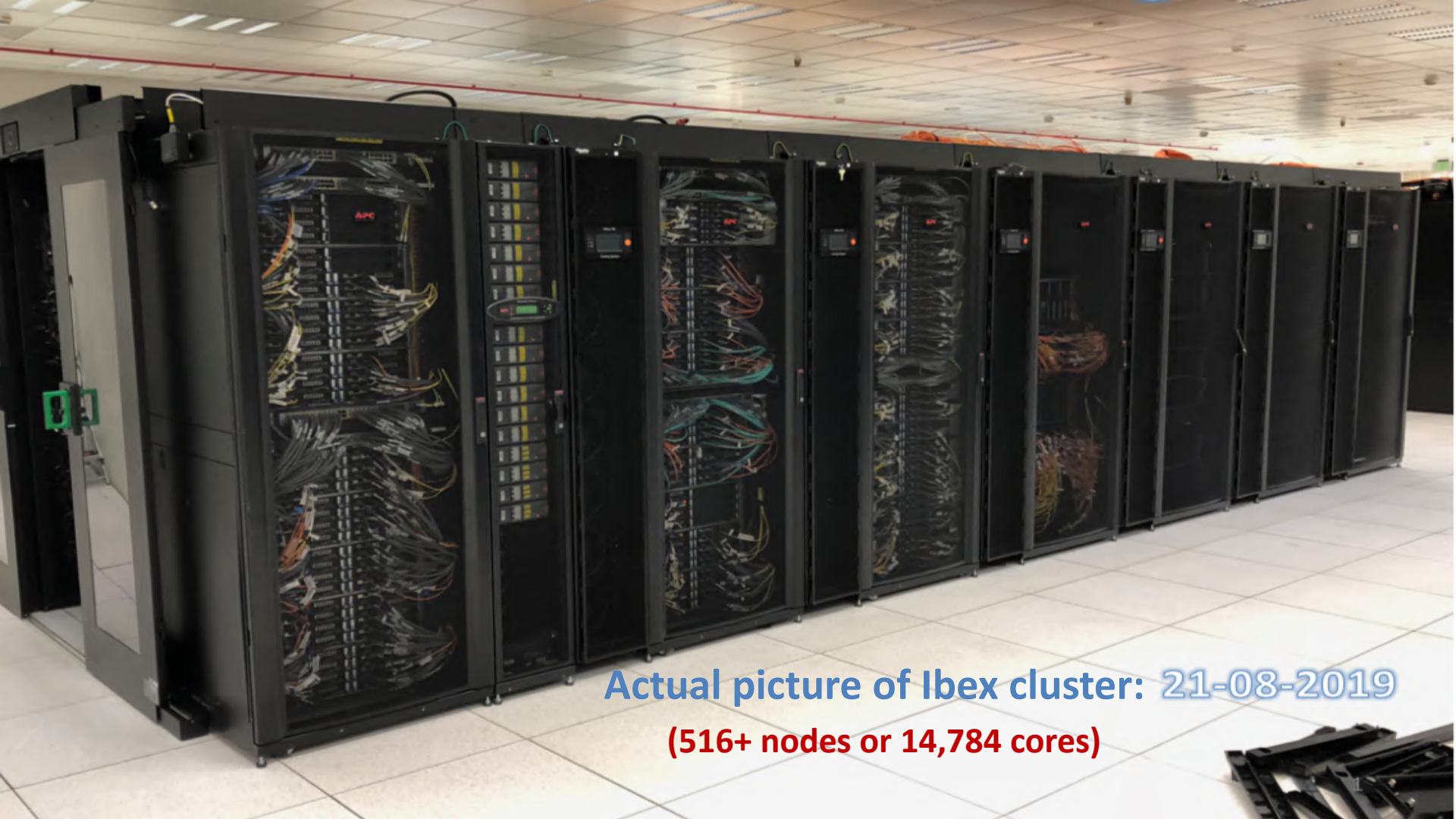
جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Introduction to Ibex

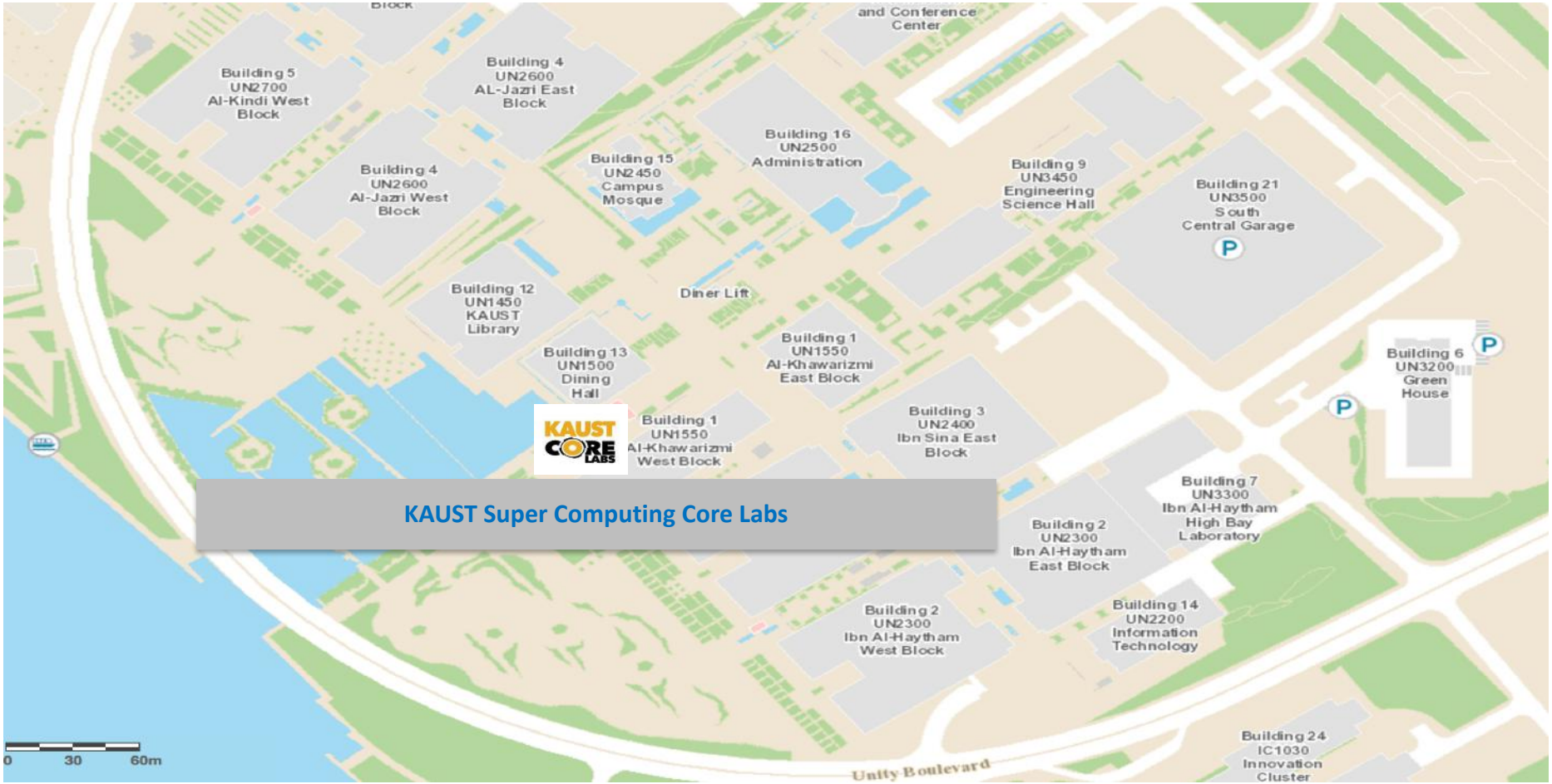
Nagarajan KATHIRESAN, Ph.D.,
Computational Scientist,
KAUST Supercomputing lab, KAUST, KSA
nagarajan.kathiresan@kaust.edu.sa

Objective

- Hands-on with **UNIX/Ibex cluster**
- Some skills in **Bioinformatics tools**
- SLURM job **submission, monitoring** and **control**.
- Genome **pipeline**
- Deep learning in your **Genome assembly** projects


Actual picture of Ibex cluster: 21-08-2019
(516+ nodes or 14,784 cores)

Who we are?



KAUST Super Computing Core Labs

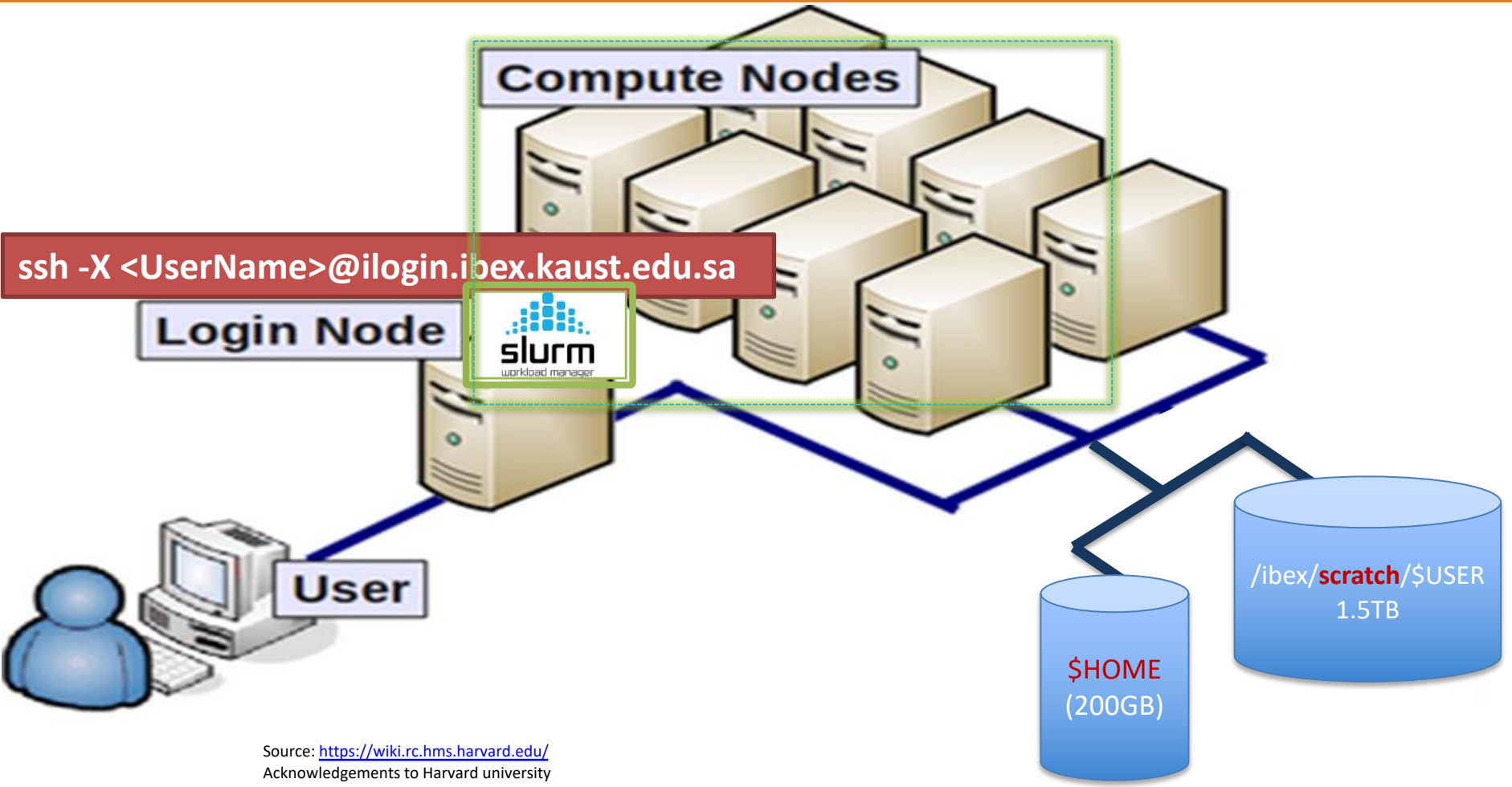
Contact Us

https://www.hpc.kaust.edu.sa/ibex

kaust-ibex.slack.com

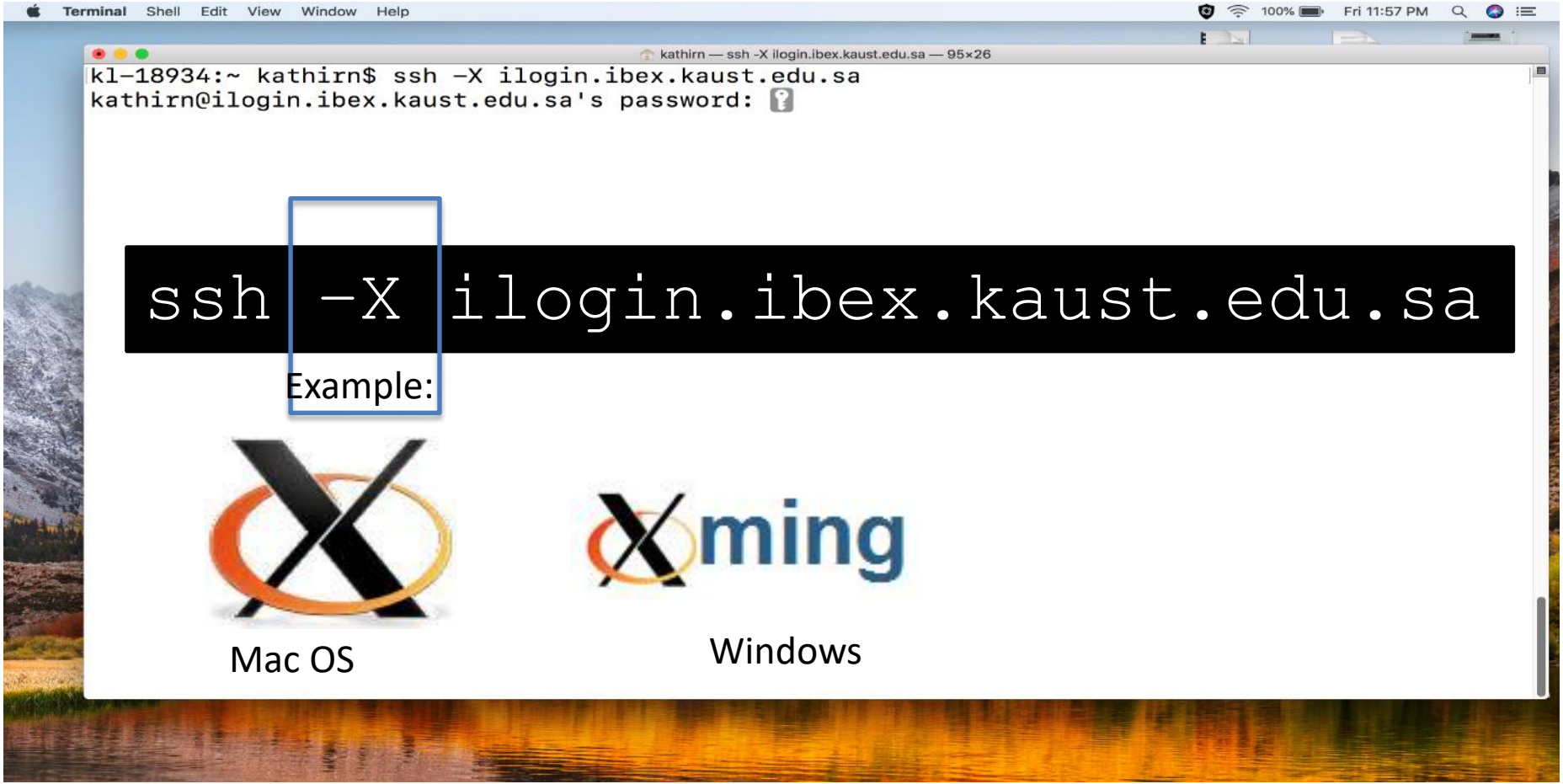
ibex@hpc.kaust.edu.sa

IBEX Cluster

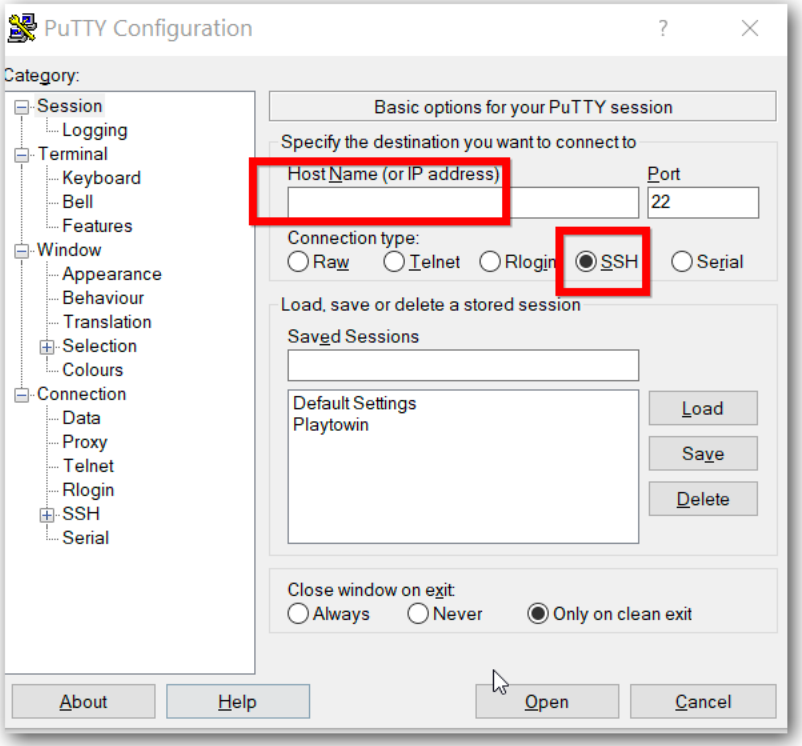


Source: https://wiki.rc.hms.harvard.edu/
Acknowledgements to Harvard university

How to login from Mac / Linux?



How to login from Windows?



Hostname: ilogin.ibex.kaust.edu.sa


UNIX COMMANDS

@ Ibex Cluster - UNIX COMMANDS

BASH prompt @ Ibex

- -bash-4.2$

Your Present/current working directory

- -bash-4.2$ **pwd**
  /**home**/kathirn

Change your directory

- -bash-4.2$ **cd /ibex/scratch/kathirn/**
- -bash-4.2$ **pwd**
  /ibex/**scratch**/kathirn

Create a directory

-bash-4.2$ **mkdir test**

Change directory to test

- -bash-4.2$ **cd test**
- -bash-4.2$ **pwd**
  /ibex/scratch/kathirn/test

@ Ibex Cluster …. Cont.

Go to previous directory

- -bash-4.2$ cd ..
- -bash-4.2$ **pwd**
  /ibex/**scratch**/kathirn

list the items in your current directory

- -bash-4.2$ **ls**
  amd file.root intel license.dat LM-
  tests sample.sortedingatk.bam work workflows

Identify the directories

- -bash-4.2$ **ls -F**
  amd/ file.root intel/ license.dat LM-
  tests/ sample.sortedingatk.bam* work/ workflows/

help on ls command

-bash-4.2$ **man ls**

File operations

Create a file

```
-bash-4.2$ touch myfile
```

Output the contents of the file

```
-bash-4.2$ cat myfile
```

Output the first 10 lines

```
-bash-4.2$ head myfile
```

Output the last 10 lines

```
-bash-4.2$ tail myfile
```

Copy the file

```
-bash-4.2$ cp myfile myfile_duplicate
```

Remove the file

```
-bash-4.2$ rm myfile_duplicate
```

Edit the file

```
-bash-4.2$ nano
```

Summary #1

`ls` — list items in current directory

`ls -l` — list items in current directory and show in long format to see perimissions, size, and modification date

`ls -a` — list all items in current directory, including hidden files

`ls -F` — list all items in current directory and show directories with a slash and executables with a star

`ls dir` — list all items in directory dir

`cd dir` — change directory to dir

`cd ..` — go up one directory

`cd /` — go to the root directory

`cd ~` — go to to your home directory

`cd -` — go to the last directory you were just in

`pwd` — show present working directory

`mkdir dir` — make directory dir

`rm file` — remove file

`rm -r dir` — remove directory dir recursively

`cp file1 file2` — copy file1 to file2

`cp -r dir1 dir2` — copy directory dir1 to dir2 recursively

`mv file1 file2` — move (rename) file1 to file2

`ln -s file link` — create symbolic link to file

`touch file` — create or update file

`cat file` — output the contents of file

`less file` — view file with page navigation

`head file` — output the first 10 lines of file

`tail file` — output the last 10 lines of file

`tail -f file` — output the contents of file as it grows, starting with the last 10 lines

`vim file` — edit file

source: http://cheatsheetworld.com/

Exercise #1

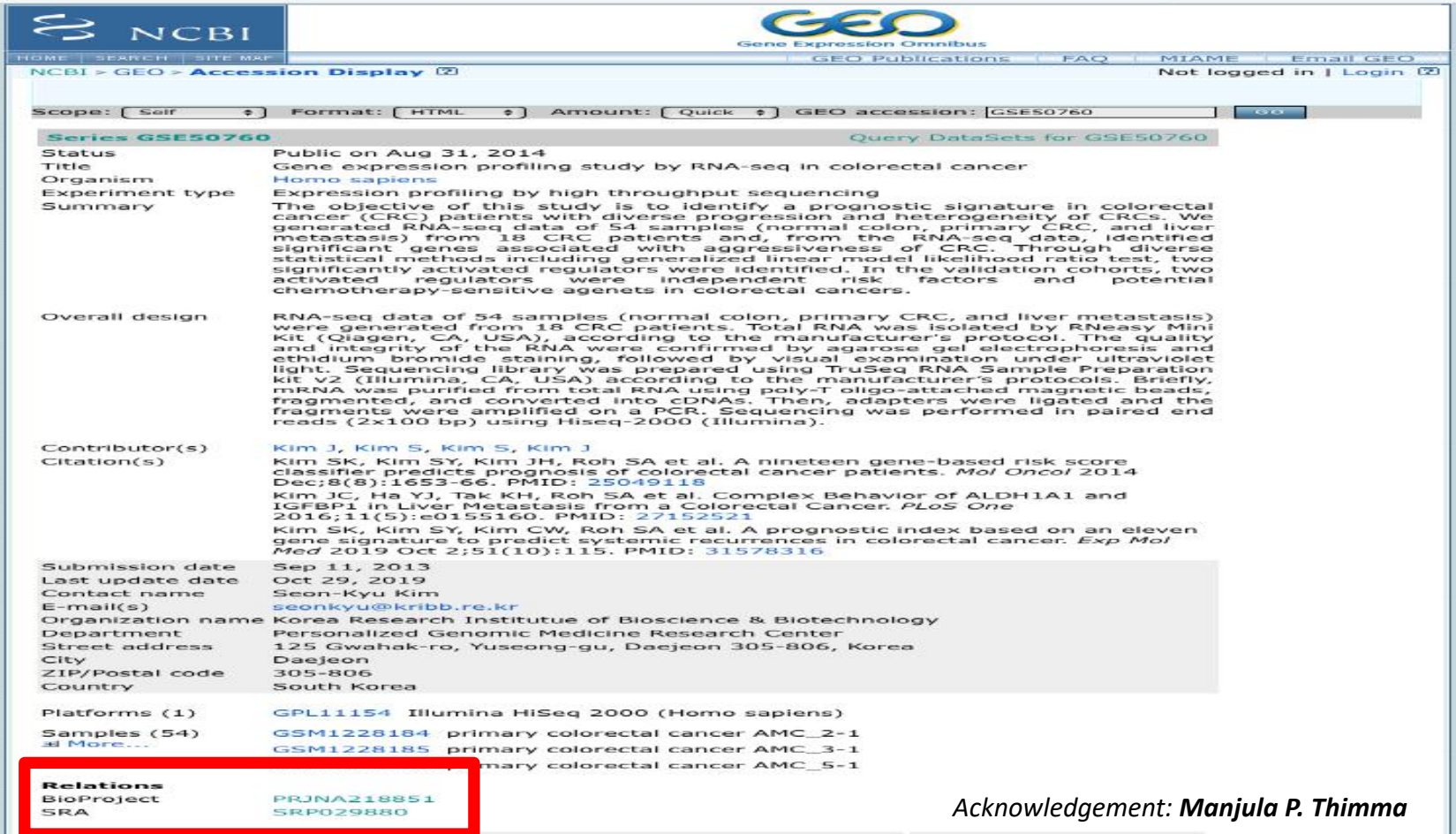
- Find the number of files in your $HOME directory.
- Create a directory called “test” in your scratch directory.

Working on your files in NCBI/Desktop/Workstation etc.

Tutorial data from the GEO accession GSE50760

Acknowledgement: **Manjula P. Thimma**

https://www.ncbi.nlm.nih.gov/gds/?term=GSE50760[Accession]



ncbi.nlm.nih.gov

NCBI

HOME | SEARCH | SITE MAP

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE50760 | Go

Series GSE50760 — Query DataSets for GSE50760

Status	Public on Aug 31, 2014
Title	Gene expression profiling study by RNA-seq in colorectal cancer
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	The objective of this study is to identify a prognostic signature in colorectal cancer (CRC) patients with diverse progression and heterogeneity of CRCs. We generated RNA-seq data of 54 samples (normal colon, primary CRC, and liver metastasis) from 18 CRC patients and, from the RNA-seq data, identified significant genes associated with aggressiveness of CRC. Through diverse statistical methods including generalized linear model likelihood ratio test, two significantly activated regulators were identified. In the validation cohorts, two activated regulators were independent risk factors and potential chemotherapy-sensitive agenets in colorectal cancers.
Overall design	RNA-seq data of 54 samples (normal colon, primary CRC, and liver metastasis) were generated from 18 CRC patients. Total RNA was isolated by RNeasy Mini Kit (Qiagen, CA, USA), according to the manufacturer's protocol. The quality and integrity of the RNA were confirmed by agarose gel electrophoresis and ethidium bromide staining, followed by visual examination under ultraviolet light. Sequencing library was prepared using TruSeq RNA Sample Preparation kit v2 (Illumina, CA, USA) according to the manufacturer's protocols. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads, fragmented, and converted into cDNAs. Then, adapters were ligated and the fragments were amplified on a PCR. Sequencing was performed in paired end reads (2x100 bp) using Hiseq-2000 (Illumina).
Contributor(s)	Kim J, Kim S, Kim S, Kim J
Citation(s)	Kim SK, Kim SY, Kim JH, Roh SA et al. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol Oncol* 2014 Dec;8(8):1653-66. PMID: 25049118 Kim JC, Ha YJ, Tak KH, Roh SA et al. Complex Behavior of ALDH1A1 and IGFBP1 in Liver Metastasis from a Colorectal Cancer. *PLoS One* 2016;11(5):e0155160. PMID: 27152521 Kim SK, Kim SY, Kim CW, Roh SA et al. A prognostic index based on an eleven gene signature to predict systemic recurrences in colorectal cancer. *Exp Mol Med* 2019 Oct 2;51(10):115. PMID: 31578316
Submission date	Sep 11, 2013
Last update date	Oct 29, 2019
Contact name	Seon-Kyu Kim
E-mail(s)	seonkyu@kribb.re.kr
Organization name	Korea Research Institutue of Bioscience & Biotechnology
Department	Personalized Genomic Medicine Research Center
Street address	125 Gwahak-ro, Yuseong-gu, Daejeon 305-806, Korea
City	Daejeon
ZIP/Postal code	305-806
Country	South Korea
Platforms (1)	GPL11154 Illumina HiSeq 2000 (Homo sapiens)
Samples (54)	GSM1228184 primary colorectal cancer AMC_2-1 GSM1228185 primary colorectal cancer AMC_3-1 mary colorectal cancer AMC_5-1

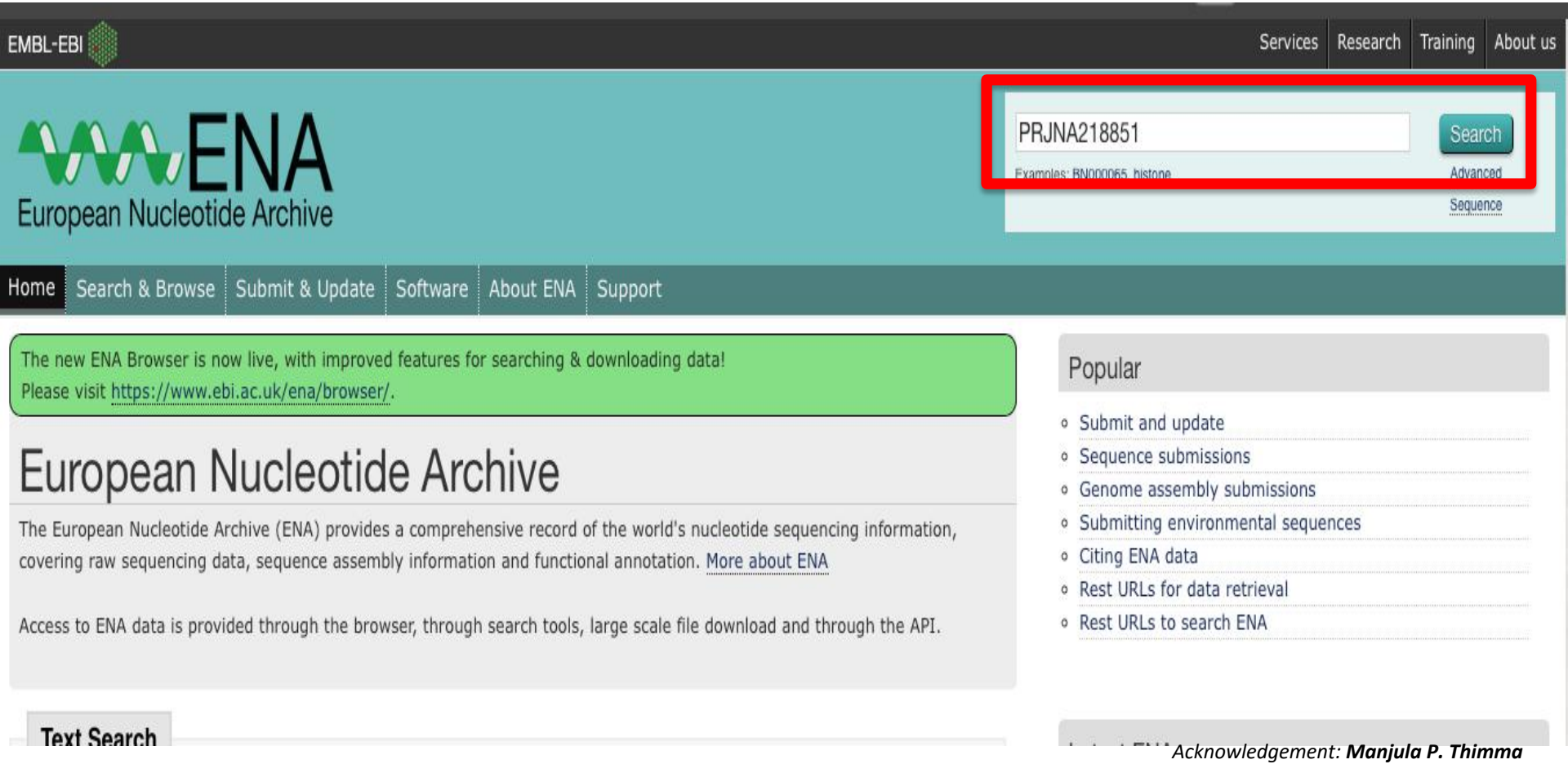
all More...

Relations

BioProject	PRJNA218851
SRA	SRP029880

Acknowledgement: ***Manjula P. Thimma***

https://www.ebi.ac.uk/ena



Acknowledgement: ***Manjula P. Thimma***

ENA
European Nucleotide Archive

PRJNA218851
Search
Examples: BN000065, histone
Advanced
Sequence

The new ENA Browser is now live, with improved features for searching & downloading data!
Please go to https://www.ebi.ac.uk/ena/browser/text-search?query=PRJNA218851 to perform this search there.

Search results for *PRJNA218851*

Show more data from EMBL-EBI

Study
Study (1)
Study (Sequence) (1)

Study (1 results found)

SRP029880 Gene expression profiling study by RNA-seq in colorectal cancer
View all 1 results

Study (Sequence) (1 results found)

PRJNA218851 Gene expression profiling study by RNA-seq in colorectal cancer
View all 1 results

Powered by EBI Search

Acknowledgement: ***Manjula P. Thimma***

The new ENA Browser is now live, with improved features for searching & downloading data!
Please go to https://www.ebi.ac.uk/ena/browser/view/PRJNA218851 to see this record there.

Contact Helpdesk

Study: PRJNA218851

Gene expression profiling study by RNA-seq in colorectal cancer

View: Project XML | Study XML

Download: Project XML | Study XML

Name
Homo sapiens

Submitting Centre
Personalized Genomic Medicine Research Center, Korea Research Institutue of Bioscience & Biotechnology

Organism
Homo sapiens

Secondary accession(s)
SRP029880

Description
The objective of this study is to identify a prognostic signature in colorectal cancer (CRC) patients with diverse progression and heterogeneity of CRCs. We generated RNA-seq data of 54 samples (normal colon, primary CRC, and liver metastasis) from 18 CRC patients and, from the RNA-seq data, identified significant genes associated with aggressiveness of CRC. Through diverse statistical methods including generalized linear model likelihood ratio test, two significantly activated regulators were identified. In the validation cohorts, two activated regulators were independent risk factors and potential chemotherapy-sensitive agenets in colorectal cancers. Overall design: RNA-seq data of 54 samples (normal colon, primary CRC, and liver metastasis) were generated from 18 CRC patients. Total RNA was isolated by RNeasy Mini Kit (Qiagen, CA, USA), according to the manufacturer's protocol. The quality and integrity of the RNA were confirmed by agarose gel electrophoresis and ethidium bromide staining, followed by visual examination under ultraviolet light. Sequencing library was prepared using TruSeq RNA Sample Preparation kit v2 (Illumina, CA, USA) according to the manufacturer's protocols. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads, fragmented, and converted into cDNAs. Then, adapters were ligated and the fragments were amplified on a PCR. Sequencing was performed in paired end reads (2x100 bp) using Hiseq-2000 (Illumina).

Lineage
Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Mammalia, Eutheria, Euarchontoglires, Primates, Haplorrhini, Catarrhini, Hominidae, Homo

Navigation | Read Files | Portal | Attributes | Publications | Parent Projects

Bulk Download Files (If the downloader app doesn't open, please try using Firefox to launch it.)

Download: 1 - 54 of 54 results in TEXT

Select columns

Showing results 1 - 10 of 54 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
PRJNA218851	SAMN02353232	SRS478664	SRX347887	SRR975551	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		
PRJNA218851	SAMN02353258	SRS478665	SRX347888	SRR975552	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		
PRJNA218851	SAMN02353264	SRS478667	SRX347889	SRR975553	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		
PRJNA218851	SAMN02353233	SRS478666	SRX347890	SRR975554	9606	Homo	Illumina	PAIRED	File 1	File 1			File 1	File 1		

Download a file

Showing results 1 - 10 of 54 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
PRJNA218851	SAMN02353232	SRS478664	SRX347887	SRR975551	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2				File 1	File 1		
PRJNA218851	SAMN02353258	SRS478665	SRX347888	SRR975552	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2				File 1	File 1		
PRJNA218851	SAMN02353264	SRS478667	SRX347889	SRR975553	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2				File 1	File 1		
PRJNA218851	SAMN02353233	SRS478666	SRX347890	SRR975554	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2				File 1	File 1		
PRJNA218851	SAMN02353255	SRS478668	SRX347891	SRR975555	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		
PRJNA218851	SAMN02353269	SRS478669	SRX347892	SRR975556	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		
PRJNA218851	SAMN02353260	SRS478671	SRX347894	SRR975558	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		
PRJNA218851	SAMN02353266	SRS478672	SRX347895	SRR975559	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		
PRJNA218851	SAMN02353236	SRS478673	SRX347896	SRR975560	9606	Homo sapiens	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2			File 1	File 1		

Prev Next Go to page: 1 Go

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975551/SRR975551_1.fastq.gz
```

Acknowledgement: ***Manjula P. Thimma***

I have 100+ files!

Is their any optimal way to download?

Editors


vi editor
emacs editor
nano editor
Note: Intended for
Showing you
:w_
File Edit Options Buffers Tools Help
Welcome to GNU Emacs, one comp
Emacs Tutorial
Emacs Guided Tour
U:%%- *GNU Emacs* Top (5,
Warning (initialization): An
:
File error: Cannot open load
To ensure normal operation, y
cause of the error in your in
the `--debug-init' option to
U:%*- *Warnings* All (8,
Quit
GNU nano 2.3.1 New Buffer Modified
^G Get Help ^O WriteOut ^R Read File ^Y Prev Page ^K Cut Text ^C Cur Pos
^X Exit ^J Justify ^W Where Is ^V Next Page ^U UnCut Text ^T To Spell

Iterations!

My list of files (created a list named as `list.txt`):

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975551/SRR975551_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975551/SRR975551_2.fastq.gz

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975552/SRR975552_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975552/SRR975552_2.fastq.gz

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975553/SRR975553_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975553/SRR975553_2.fastq.gz

…

…

..

For loop

```
for myfile in `cat list.txt`;
   do
      wget $myfile;
   done
```

This step is not really PARALLEL ☹

Job Submission

A simple job script

Caution note (by default SLURM allocation)
- memory = 2GB
- CPU = 1 core
- Node = 1 node

Minimum 3 parameters:

1. **sbatch:** Submit a batch script to Slurm
2. **time:** Set a limit on the total run time of the job allocation
   --time=days-hours:minutes:seconds
   -t days-hours:minutes:seconds
3. **wrap:** specified command string or simple "sh" shell script & submit to the slurm controller

Example:

```
$ sbatch --time=00:10 --wrap="wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975551/SRR975551_1.fastq.gz"
```

Another Scripting Example (As a Batch Script)

```
$ cat my_job.sh
#!/bin/bash
#SBATCH --time=00:10
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR975/SRR975551/SRR975551_1.fastq.gz

$ sbatch ./my_job.sh
Submitted batch job 7438
```

For loop in SLURM script / Parallel downloads

Create a file called: `parallel_download.sh`

```
for myfile in `cat list.txt`;
  do
    sbatch --time=00:10 --wrap="wget $myfile"
  done
```

This step is REALLY in PARALLEL ☺

```
$ sh ./parallel_download.sh
Submitted batch job 9459565
Submitted batch job 9459566
Submitted batch job 9459567
Submitted batch job 9459568
Submitted batch job 9459569
Submitted batch job 9459570
```

Parallel downloads (in BATCH script)

Create a file called: MyBATCH.sh

```
!#/bin/bash
#SBATCH --time=00:10
wget $1
```

Create a FOR loop called: MyFOR.sh

```
#!/bin/bash -l
for myfile in `cat list.txt`;
 do
  sbatch myBATCH.sh $myfile;
 done
```

```
$ sh ./myFOR.sh
Submitted batch job 9463454
Submitted batch job 9463455
Submitted batch job 9463456
Submitted batch job 9463457
Submitted batch job 9463458
Submitted batch job 9463459
```

This step is REALLY in PARALLEL ☺

Monitor the jobs!

```
$ squeue -u $USER
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           9463688     batch myBATCH.  kathirn  R       3:45      1 cn605-07-l
           9463689     batch myBATCH.  kathirn  R       3:45      1 cn605-07-l
           9463686     batch myBATCH.  kathirn  R       3:49      1 cn605-07-l
           9463687     batch myBATCH.  kathirn  R       3:49      1 cn605-07-l
           9463685     batch myBATCH.  kathirn  R       3:52      1 cn605-07-l
           9463684     batch myBATCH.  kathirn  R       3:59      1 cn605-07-r
```

Exercise #2

- Download any 5 paired sample dataset from the “Study: PRJNA218851” using SLURM scheduler.
- Copy any 1 paired sample dataset from Ibex to your Mac/Desktop.

Most frequently used SLURM commands

Sl.No.	Command	Description
Resource statics		
1.	sinfo	View information about SLURM nodes and partitions.
2.	smap	Graphically view information about SLURM jobs, partitions, and set configurations parameters.
Job submission		
1.	sbatch	Submit a job script to a queue.
Job Management		
1.	squeue	View information about jobs located in the SLURM scheduling queue.
2.	scancel	Signal jobs or job steps that are under the control of SLURM (cancel jobs or job steps).
3.	sacct	Displays accounting data for all jobs and job steps
Job statistic		
1.	scontrol	View and modify SLURM configuration and state.
2.	seff	View the CPU and memory efficiency (e.g. completed jobs and reserved resources)
Interactive job execution		
1.	salloc	Allocate resources (set of nodes) for interactive (execute a command) use.
2.	srun	Run a parallel job.

sbatch: Submit a batch script to Slurm

Sl.No	Flag Syntax	Resource	Description	Notes
1.	--partition=general-compute	partition	queue name for your jobs.	default is batch
2.	--time=DD-HH:MM:SS	time	Time limit for running your job.	Mandatory; max time limit is 14 days.
3.	--nodes=<integer number>	nodes	Number of compute nodes required for your job.	default is 1
4.	--cpus-per-node=<integer number>	cpus/cores	Number of cores required on each compute node.	default is 1
5.	--gres=gpu:<$$$>:<#>	resource feature	Required GPUs on compute nodes, where $$$ is the GPU architecture and # is number of GPU cards	default is no feature specified;
6.	--mem=<integer number>	memory	Memory limit per compute node for your job. Memory in MB by default. E.g. –mem=1024 is same value as –mem=1gb	default limit is 2GB per core
7.	--job-name="<string> "	job name	Name of your job.	default is the JobID

Cont.

8.	--output=<file name>	output file	Name of file for stdout.	default is the JobID
9.	--error=<file name>	Error file	Name of file for stderr	default is the JobID
10.	--mail-user=username@kaust.edu.sa	email address	User's email address	Optional
11.	--mail-type=ALL --mail-type=BEGIN --mail-type=END	email notification	When email is sent to user.	omit when no email address is given
12.	--exclusive	access	Exclusive access to the compute nodes.	default is sharing nodes
13.	--reservation=<name-of_reservation>	Special resource allocation	Reserved nodes for special purpose.	None
14.	-w,--nodelist=<host_name1, host_name2, ...>		Requesting specific node required for your job.	
15.	--exclude<host_name, host_name2, ...>		Excluding specific node for your job.	
16.	--wrap="<command to execute>"		Command to be executed on the compute nodes	

Copy files from my Desktop to Ibex

```
scp <file name>
<user_name@ilogin.ibex.kaust.edu.sa:destination_directory>
```

Example:

```
scp Homo_sapiens.GRCh38.99.chr.gtf
kathirn@ilogin.ibex.kaust.edu.sa:/ibex/scratch/kathirn/
```

Copy files from Ibex to my Desktop

```
scp
<user_name@ilogin.ibex.kaust.edu.sa:destination_directory/file
name> <file name>
```

Example:

```
scp kathirn@ilogin.ibex.kaust.edu.sa:/ibex/scratch/kathirn/
Homo_sapiens.GRCh38.99.chr.gtf .
```

Summary #2

`wget file` — download a file

`curl file` — download a file

`scp user@host:file dir` — secure copy a file from remote server to the dir directory on your machine

`scp file user@host:dir` — secure copy a file from your machine to the dir directory on a remote server

`scp -r user@host:dir dir` — secure copy the directory dir from remote server to the directory dir on your machine

`ssh user@host` — connect to host as user

source: http://cheatsheetworld.com/

Exercise #3

- Copy any 1 paired sample dataset from Ibex to your Mac/Desktop.

Move data (as a directory) between two systems

The **rsync** utility is a very useful utility for synchronizing files and directories between two different servers.

- Copying from the local machine to a remote machine:

  rsync <options> local_directory remote_server_name:remote_directory

- Copying from a remote machine to the local machine:

  rsync <options> remote_server_name:remote_directory local_directory

<options >

-a	archive mode
-r	recursive over subdirectories
-v	verbose
-x	don't cross filesystem boundaries
-H	preserve hard links
-P	show progress
-n	no-op, or dry-run

```
$ rsync -arvxHP my_data
kathirn@ilogin.ibex.kaust.edu.sa:/
ibex/scratch/kathirn/my_data/
```

File Compression

`tar cf file.tar files` — create a tar named file.tar containing files

`tar xf file.tar` — extract the files from file.tar

`tar czf file.tar.gz files` — create a tar with Gzip compression

`tar xzf file.tar.gz` — extract a tar using Gzip

`gzip file` — compresses file and renames it to file.gz

`gzip -d file.gz` — decompresses file.gz back to file

UNIX tools for Bioinformatics

Working with genome files

```
$ ls -l
total 358834420
-rw-r--r-- 1 kathirn g-kathirn      1247270 Feb 23 10:25 Aegilops_tauschii.Aet_v4.0.ncrna.fa
-rw-r--r-- 1 kathirn g-kathirn       314263 May 20  2005 B_anthracis_contigs.fasta
-rw-r--r-- 1 kathirn g-kathirn          792 Nov 29 10:09 B_anthracis_contigs.fasta.fai
-rw-r--r-- 1 kathirn g-kathirn       317830 May 20  2005 B_anthracis_Mslice.fasta
-rw-r--r-- 1 kathirn g-kathirn           35 Nov 29 10:05 B_anthracis_Mslice.fasta.fai
-rw-r--r-- 1 kathirn g-kathirn          118 Sep 11 12:37 collect_bam.txt
-rw-r--r-- 1 kathirn g-kathirn  54009320489 Jun 17  2015 ERR194161_1.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn  55523025117 Jun 17  2015 ERR194161_2.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn        52505 Feb  4 11:18 example.vcf.gz
-rw-r--r-- 1 kathirn g-kathirn        12197 Feb  4 11:16 example.vcf.gz.bzz
-rw-r--r-- 1 kathirn g-kathirn     32002944 Dec 17 14:35 gene-calls.fa
-rw-r--r-- 1 kathirn g-kathirn          222 Sep 11 14:52 genes.fpkm_tracking
-rw-r--r-- 1 kathirn g-kathirn       279242 May 20  2005 H_pylori26695_Eslice.fasta
-rw-r--r-- 1 kathirn g-kathirn       268919 May 20  2005 H_pyloriJ99_Eslice.fasta
-rw-r--r-- 1 kathirn g-kathirn          232 Sep 11 14:52 isoforms.fpkm_tracking
-rw-r--r-- 1 kathirn g-kathirn           62 Sep 12 11:15 mutant.aligned.bam
-rw-r--r-- 1 kathirn g-kathirn      1545913 Sep 11 12:28 mutant_R1.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn      1743596 Sep 11 12:28 mutant_R2.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn 121691186161 Sep 11 14:03 NA12878_S1.bam
-rw-r--r-- 1 kathirn g-kathirn 121706596363 Sep 27 17:04 NA12878.sorted.bam
-rw-r--r-- 1 kathirn g-kathirn      8759544 Sep 27 17:31 NA12878.sorted.bam.bai
-rw-r--r-- 1 kathirn g-kathirn      1252784 Sep 13 12:11 reads2.left.fq.gz
-rw-r--r-- 1 kathirn g-kathirn      1274471 Sep 13 12:11 reads2.right.fq.gz
-rw-r--r-- 1 kathirn g-kathirn       618021 Sep 13 12:11 reads.left.fa.gz
-rw-r--r-- 1 kathirn g-kathirn      1251148 Sep 13 12:11 reads.left.fq.gz
-rw-r--r-- 1 kathirn g-kathirn       631865 Sep 13 12:11 reads.right.fa.gz
-rw-r--r-- 1 kathirn g-kathirn      1272939 Sep 13 12:11 reads.right.fq.gz
-rw-r--r-- 1 kathirn g-kathirn    283603689 Feb 13  2011 sample.bam
-rw-r--r-- 1 kathirn g-kathirn      5805416 Nov 22  2011 sample.bam.bai
-rw-r--r-- 1 kathirn g-kathirn    573956710 Jun 25  2009 sample.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn    301938695 Feb 14 09:26 sample.sortedingatk.bam
-rw-r--r-- 1 kathirn g-kathirn    284075838 Sep 12 11:46 sample.sortedinpicard.bam
-rw-r--r-- 1 kathirn g-kathirn      5805832 Sep 27 14:59 sample.sortedinpicard.bam.bai
-rw-r--r-- 1 kathirn g-kathirn      4194304 Sep 27 15:40 samtools.164780.1999.tmp.0000.bam
-rw-r--r-- 1 kathirn g-kathirn            0 Sep 11 14:52 skipped.gtf
-rw-r--r-- 1 kathirn g-kathirn    639970768 Mar 26  2012 SRR396636.sra_1.fastq
-rw-r--r-- 1 kathirn g-kathirn    639970768 Mar 26  2012 SRR396636.sra_2.fastq
-rw-r--r-- 1 kathirn g-kathirn    523217540 Mar 26  2012 SRR396637.sra_1.fastq
-rw-r--r-- 1 kathirn g-kathirn    523217540 Mar 26  2012 SRR396637.sra_2.fastq
-rw-r--r-- 1 kathirn g-kathirn   4598249219 Sep 17 08:25 SRR622461_1.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn   4792383631 Sep 17 08:18 SRR622461_2.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn        41526 Sep 24  2017 test_data.sam
-rw-r--r-- 1 kathirn g-kathirn          887 Sep 11 14:52 transcripts.gtf
```

Fasta

Indexed Fasta

Compressed Fastq

Compressed VCF

BAM

Sorted BAM

SAM

GTF

Working with fasta file

```
$ more Aegilops_tauschii.Aet_v4.0.ncrna.fa
```

```
$ head Aegilops_tauschii.Aet_v4.0.ncrna.fa
>ENSRNA050031380-T1 ncrna chromosome:Aet_v4.0:2D:126982204:126982306:-1 gene:ENSRNA050031380
gene_biotype:snRNA transcript_biotype:snRNA gene_symbol:U6 description:U6 spliceosomal RNA
ACTATATAAAAAACTTCCAATTTTAGTGGAACTATACAGAGAAGATTAGCATGGCCCCGA
CGCAAGGATGACACACACGAATTGAGAAATGATCCAAATTTTT
>ENSRNA050031382-T1 ncrna chromosome:Aet_v4.0:2D:161181740:161181842:-1 gene:ENSRNA050031382
gene_biotype:snRNA transcript_biotype:snRNA gene_symbol:U6 description:U6 spliceosomal RNA
GTCCCTTCGGGGACATCCAATAAAATCGAAACGATACAGAGAAGATTAGCATGGCCCCTG
CGCAAGGATGACACGCACAAATCGAGAAATGGTCCAGATTTTT
>ENSRNA050031412-T1 ncrna supercontig:Aet_v4.0:jcf7190000131069:318:420:-1 gene:ENSRNA0500314
12 gene_biotype:snRNA transcript_biotype:snRNA gene_symbol:U6 description:U6 spliceosomal RNA
GTCCCTTCTGGGACATCCGATAAAATTGGAACGATACAGAGAAGATTAGCATGGCCCCTG
CGCAAGGATGACACGCACAAATCGAGAAATGGTCCAATTTTTT
>ENSRNA050031423-T1 ncrna chromosome:Aet_v4.0:2D:528045395:528045586:-1 gene:ENSRNA050031423
gene_biotype:snRNA transcript_biotype:snRNA gene_symbol:U2 description:U2 spliceosomal RNA
```

Extract the headers from the FASTA file

`grep, egrep, fgrep` → print lines matching a pattern

`-i, --ignore-case` → ignore case

`-v, --invert-match` → “invert”, get the lines not matching the patent

`-w, --word-regexp` → Get the lines when matches whole patent

`-o, --only-matching` → Get only the matching part

```
$ grep -o ncrna Aegilops_tauschii.Aet_v4.0.ncrna.fa
ncrna
ncrna
ncrna
ncrna
ncrna
ncrna
```

```
$ grep -io "GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG" Aegilops_tauschii.Aet_v4.0.ncrna.fa
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTC
GAGAGCCGTC
GAGAGCCGTC
GAGAGCCGTC
```

```
$ grep -v ">" Aegilops_tauschii.Aet_v4.0.ncrna.fa
ACTATATAAAAAACTTCCAATTTTAGTGGAACTATACAGAGAAGATTAGCATGGCCCCGA
CGCAAGGATGACACACACGAATTGAGAAATGATCCAAATTTTT
GTCCCTTCGGGGACATCCAATAAAATCGAAACGATACAGAGAAGATTAGCATGGCCCCTG
CGCAAGGATGACACGCACAAATCGAGAAATGGTCCAGATTTTT
GTCCCTTCTGGGACATCCGATAAAATTGGAACGATACAGAGAAGATTAGCATGGCCCCTG
CGCAAGGATGACACGCACAAATCGAGAAATGGTCCAATTTTTT
ATACTTTCTCGACCTTTTTTGCTAAGATGAAGTGTAACATCCGTTCTTGTTAGTTTAATAT
CTGATATATGGACCATCGTGTCCATATGATATTAATTTATTGTGTGGGAGAGGATCTATA
TATGAGCCTGGTACATGGGTTCTCACGTATTGGTCCAAACGTTGCACTACTACTAGAGCC
AGAGCATCTCAA
AGCTGCGCGGTGAGCACAAAGCGAACTATTCTTTCGCCTTTTACTAAAGAATACCGTGTG
```

Word count

`wc` → Count the number of lines, words and characters in a given file

```
$ wc Aegilops_tauschii.Aet_v4.0.ncrna.fa
  13525   48871 1247270
Aegilops_tauschii.Aet_v4.0.ncrna.fa
```

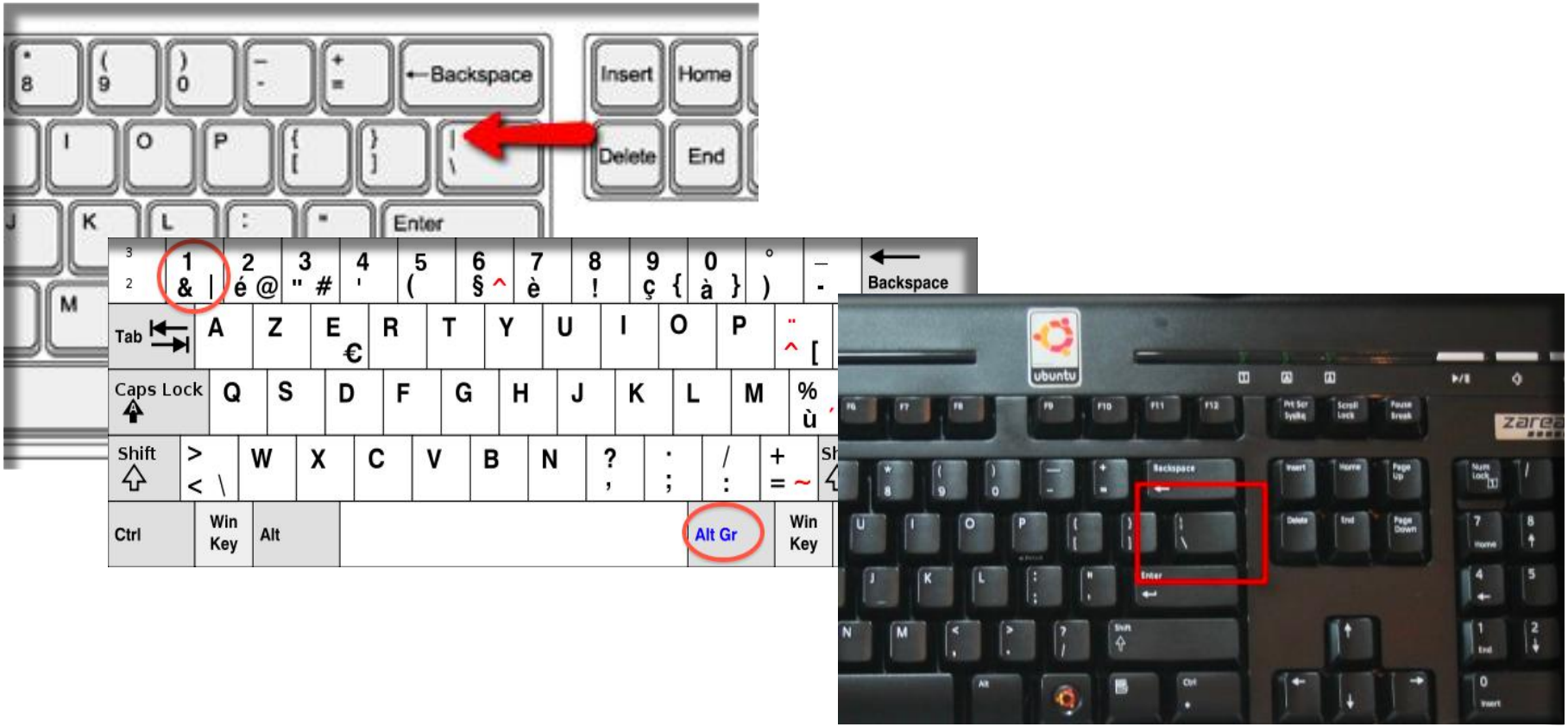
```
$ wc -l Aegilops_tauschii.Aet_v4.0.ncrna.fa
13525 Aegilops_tauschii.Aet_v4.0.ncrna.fa
```

```
$ wc -w Aegilops_tauschii.Aet_v4.0.ncrna.fa
48871 Aegilops_tauschii.Aet_v4.0.ncrna.fa
```

```
$ wc -c Aegilops_tauschii.Aet_v4.0.ncrna.fa
1247270 Aegilops_tauschii.Aet_v4.0.ncrna.fa
```

Combining the commands

| → Pipe character



Example

```
$ grep -io "GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG" Aegilops_tauschii.Aet_v4.0.ncrna.fa
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG
```

```
$ grep -io "GAGAGCCGTGTTATGGGTGACCTTATTGCACGGTTCAGAGAG" Aegilops_tauschii.Aet_v4.0.ncrna.fa | wc -l
8
```

Useful data processing tools!

cut → This command allows extracting the column from the file

Use: `cut -f <column no> file name`

```
 $ cat sample.vcf
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=CNV,Description="Copy number variable region">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA00001 NA00002 NA00003
19      111     .       A       C       9.6     .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
19      112     .       A       G       10      .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
20      14370   rs6054257       G       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ     0|0:48:1:51,51  1|0:48:8
20      17330   .       T       A       3       q10     NS=3;DP=11;AF=0.017     GT:GQ:DP:HQ     0|0:49:3:58,50  0|1:3:5:65,3
```

Useful tableview!

https://github.com/informationsea/tableview/releases/download/v0.4.6/tableview_linux_amd64

```
$ cat sample.vcf | grep -v "#"
```

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
19      111     .       A       C       9.6     .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
19      112     .       A       G       10      .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
20      14370   rs6054257       G       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ     0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20      17330   .       T       A       3       q10     NS=3;DP=11;AF=0.017     GT:GQ:DP:HQ     0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3:.,.
20      1110696 rs6040355       A       G,T     67      PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB       GT:GQ:DP:HQ     1|2:21:6:23,27  2|1:2:0:18,2    2/2:35:4:
.,.
20      1230237 .       T       .       47      PASS    NS=3;DP=13;AA=T GT:GQ:DP:HQ     0|0:54:.:56,60  0|0:48:4:51,51  0/0:61:2:.,.
20      1234567 microsat1       G       GA,GAC  50      PASS    NS=3;DP=9;AA=G;AN=6;AC=3,1      GT:GQ:DP        0/1:.:4 0/2:17:2        1/1:40:3
20      1235237 .       T       .       .       .       .       GT      0/0     0|0     ./.
X       10      rsTest  AC      A,ATG   10      PASS    .       GT      0       0/1     0|2
[kathirn@login510-37 data]$
```

```
$ cat sample.vcf | grep -v "#" | tableview_linux_amd64
```

19	111	.	A	C	9.6	.	.	GT:HQ	0|0:10,10	0|0:10,10	0/1:3,3
19	112	.	A	G	10	.	.	GT:HQ	0|0:10,10	0|0:10,10	0/1:3,3
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3:.,.
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4:.,.
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:.:56,60	0|0:48:4:51,51	0/0:61:2:.,.
20	1234567	microsat1	G	GA,GAC	50	PASS	NS=3;DP=9;AA=G;AN=6;AC=3,1	GT:GQ:DP	0/1:.:4	0/2:17:2	1/1:40:3
20	1235237	.	T	.	.	.	.	GT	0/0	0|0	./.
X	10	rsTest	AC	A,ATG	10	PASS	.	GT	0	0/1	0|2

Cont. … cut command!

```
$ cat sample.vcf | grep -v "#"
19      111     .       A       C       9.6     .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
19      112     .       A       G       10      .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
20      14370   rs6054257       G       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ     0|0:48:1:51,51  1|0:48:8
20      17330   .       T       A       3       q10     NS=3;DP=11;AF=0.017     GT:GQ:DP:HQ     0|0:49:3:58,50  0|1:3:5:65,3
```

```
$ cat sample.vcf | grep -v "#" | cut -f1,4,5
19      A       C
19      A       G
20      G       A
20      T       A
```

```
$ cat sample.vcf | grep -v "#" | cut -f1,9-10
19      GT:HQ   0|0:10,10
19      GT:HQ   0|0:10,10
20      GT:GQ:DP:HQ     0|0:48:1:51,51
20      GT:GQ:DP:HQ     0|0:49:3:58,50
```

Uniq and sort

```
$ cat sample.vcf | grep -v "#" | cut -f1 | sort
19
19
20
20
20
20
20
20
X
```

```
$ cat sample.vcf | grep -v "#" | cut -f1 | sort | uniq
19
20
X
```

awk

AWK → scans each line and performance some actions.

`awk '<patterns1> {action1} …'`

```
$ cat sample.vcf | grep -v "#"
19      111     .       A       C       9.6     .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
19      112     .       A       G       10      .       .       GT:HQ   0|0:10,10       0|0:10,10       0/1:3,3
20      14370   rs6054257       G       A       29      PASS    NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ     0|0:48:1:51,51  1|0:48:8:51,51  1/1:43:5:.,.
20      17330   .       T       A       3       q10     NS=3;DP=11;AF=0.017     GT:GQ:DP:HQ     0|0:49:3:58,50  0|1:3:5:65,3    0/0:41:3:.,.
20      1110696 rs6040355       A       G,T     67      PASS    NS=2;DP=10;AF=0.333,0.667;AA=T;DB       GT:GQ:DP:HQ     1|2:21:6:23,27  2|1:2:0:18,2 2/2:35:4:.,.
20      1230237 .       T       .       47      PASS    NS=3;DP=13;AA=T GT:GQ:DP:HQ     0|0:54:.:56,60  0|0:48:4:51,51  0/0:61:2:.,.
20      1234567 microsat1       G       GA,GAC  50      PASS    NS=3;DP=9;AA=G;AN=6;AC=3,1      GT:GQ:DP        0/1:.:4 0/2:17:2        1/1:40:3
20      1235237 .       T       .       .       .       .       GT      0/0     0|0     ./.
X       10      rsTest  AC      A,ATG   10      PASS    .       GT      0       0/1     0|2
```

```
$ cat sample.vcf | grep -v "#" | awk '$1 = 20 {print $1, $4, $5 }'
20 A C
20 A G
20 G A
20 T A
20 A G,T
20 T .
20 G GA,GAC
20 T .
20 AC A,ATG
```

Combine commands: awk + pipe + uniq + sort …

```
 $ cat sample.vcf | grep -v "#" | awk '$1 = 20 {print $1, $4, $5 }' | sort
20 A C
20 AC A,ATG
20 A G
20 A G,T
20 G A
20 G GA,GAC
20 T .
20 T .
20 T A
```

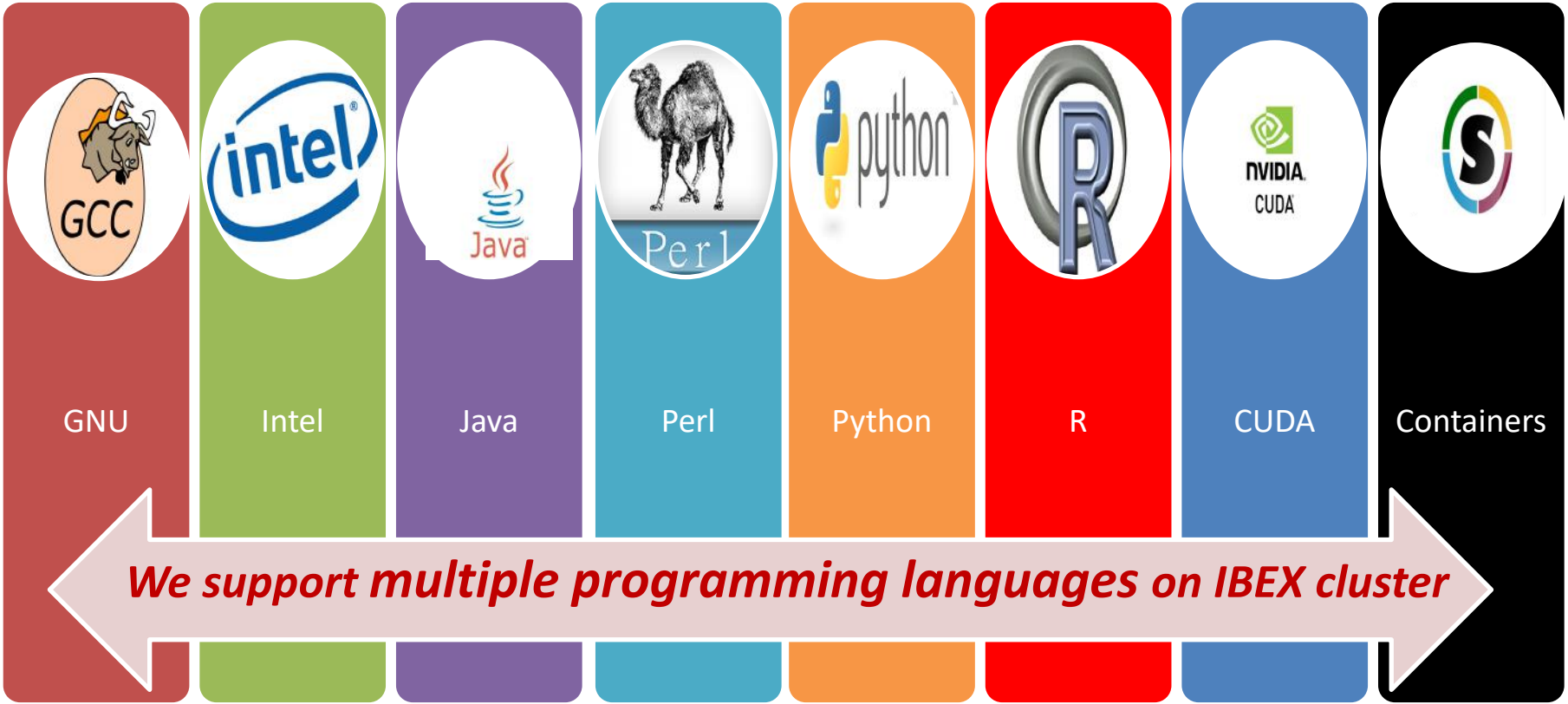
```
 $ cat sample.vcf | grep -v "#" | awk '$1 = 20 {print $1, $4, $5 }' | sort | uniq
20 A C
20 AC A,ATG
20 A G
20 A G,T
20 G A
20 G GA,GAC
20 T .
20 T A
```

Exercise #4

- Download the Gene annotation file (Homo_sapiens.GRCh38.99.gtf) from https://asia.ensembl.org/Homo_sapiens/Info/Index and do the following:
  - Print the list of chromosomes
  - Count the total number of gene
  - Count the number of transcript in Chromosome 1

Ibex software stack

IBEX: Software (system) for your apps development



IBEX: Development/Utility Libraries

IBEX: Application Software

```
[dbn503-33-r Intel]:/home/kathirn $ module avail
```

```
------------------------------------------------------------------------------ /sw/csi/modulefiles/applications -------------------------------
abaqus/2017
abyss/2.0.2
adf/2016.102/precompiled
adf/2016.106/precompiled
ansys/18.1
ansys/18.2
antismash/4.1/anaconda2-2.5.0
anvio/3.0.0
asciidoc/8.6.10
augustus/3.2.3/boost1.65.1_gnu6.4.0
augustus/3.3/boost1.65.1_gnu6.4.0
bamtools/2.5.1/gnu-6.4.0
barrnap/0.8
bcftools/1.6/gnu-6.4.0
bclfastq2/2.20/el7_gnu6.4.0
bedtools/2.26.0/gnu-6.4.0
binsanity/0.2.6.1
bison/3.0.4/gnu-6.4.0
bison/3.0.4/intel-2017
blast/2.2.25
blast/2.2.26
blast/2.2.5
blast/2.6.0
blast/2.7.1
bowtie2/2.3.3.1
braker/1.9
braker/2.1
breakit/0.2(default)
breakit/0.3
busco/3.0
bwa/0.7.17/gnu-6.4.0
bwakit/0.7.15/binary-0.7.15
bx-python/0.7.4/anaconda2-2.5.0
cantera/2.3.0/anaconda2env
cantera/2.3.0/anaconda3env
canu/1.6/gnu4.8.5
cap3/151002(default)
cap3/2017
cegma/2.5
centrifuge/1.0.3-beta
checkm/1.0.9/anaconda2-2.5.0
cif2cell/1.2.10/python-2.7.14
circlator/1.5.5/anaconda3-4.4.0
circos/0.69-6
cobra/3.0
comsol/5.3
converge/2.3.24
converge/2.4.14
converge/2.4.9
cp2k/5.1/openmpi-3.0.0-gnu-6.4.0
crystal14/1.0.3/intel16_parallel
crystal14/1.0.3/intel16_sequential
cufflinks/2.1.1
curl/7.56.1/gnu-6.4.0
curl/7.56.1/intel-2017
decimate/0.9.5
decimate/0.9.6(default)
decimate/latest
deepvariant/0.4.1/anaconda2-2.5.0
drmaa/0.7.8/anaconda2env
drmaa/0.7.8/anaconda3env
fasttext/2017-10-03
firefox/56.0
fleur/0.27/openmpi-3.0.0-gcc-6.4.0
fleur/0.27/openmpi-3.0.0_intel-2017
flex/2.6.4/gnu-6.4.0
gatk/4.0.1.1
gaussian09/d.01/precompiled
gensim/3.2.0
gerris/1.3.2/gnu-6.4.0
git/2.15.0
gnuplot/5.0.0
gperf/3.0.3/gnu-6.4.0
gperf/3.0.3/intel-2016
greasy/2.1
gromacs/2016.4/openmpi-2.1.1-intel-2016-dp
gromacs/2016.4/openmpi-2.1.1-intel-2016-fftw-sp
gromacs/2016.4/openmpi-2.1.1-intel-2016-sp
gromacs/2016.4/openmpi-3.0.0-gnu-6.4.0
gromacs/2018/openmpi-2.1.1-intel-2016-dp
gromacs/2018/openmpi-2.1.1-intel-2016-sp
gromacs/5.0.4/ompi202-intel2017
gsteramer/1.12.3/gnu-6.4.0
help2man/1.47.5
hicpro/2.10.0
hirise/20151117
hirise/20151117.old
hmmer/3.1b2
ilastik/1.3.0
imbalanced-learn/0.3.1
imblearn/0.3.2
imod/4.9.4
infernal/1.1.2/openmpi-3.0.0-gnu-6.4.0
interproscan/5.30-69.0
isown/2018
japsa/1.7-10a
lammps/2017.08.11/openmpi-2.1.1-intel-2016
libsbml/5.17.0
links/1.8.5
macs2/2.1.1.2/anaconda2-2.5.0
madagascar/2.0/gnu-6.4.0
maestro/1.6
maestro/1.7.1
masurca/3.2.4
materialstudio/2017R2
mathematica/11.2.0
matlab/R2016b
matlab/R2017b
matlab/R2018a
mcce/3.0
mcl-edge/14.137
mecat/20170522
medea/2.21
meep/1.3/ompi211-intel1602
megan/6.12.3
metabat/0.26.3(default)
metabat/2.12.1
metis/5.1.0/gnu-6.4.0
mhap/2.1.1
molpro/2012.1p16/precompiled
mpb/1.5/ompi211-intel1602
mrcc/2017-09-25/ompi30-intel17
octave/4.0.3
octopus/7.2/gnu-6.4.0_openmpi3.0.0
octopus/7.2/openmpi1.10.7_intel2016
oommf/1.2beta1
openfoam/5.0/gnu-6.4.0
orca/4.0.0.2/precompiled
orca/4.0.1.2/openmpi-2.0.2-gnu-6.4.0
orca/4.0.1.2/openmpi-2.0.2-intel-2017
orthomcl/2.0.9
panx/1.6.0/anaconda2-250
petsc/3.8.3/gnu-6.4.0
phyloFlash/3.0b1/anaconda2-2.5.0
picard/2.17.6
pilon/1.22
plink/5.3
polyrate/2016-2A/intel2016-openmpi2.1.1-VRC
polyrate/2016-2A/intel2016-RP
pymatgen/2018.5.14
pyspark/2.2.0
qchem/4.3
qctool/v2
qtwebkit/5.9.0/gnu-6.4.0
qtwebkit/5.9.0/intel-2016
quantumespresso/6.2/openmpi-3.0.0-gnu-6.4.0
quast/5.0.0.dev
relion/2.1/openmpi-2.1.1-intel-2016
RStudio_Desktop/1.1.383
samtools/1.1
samtools/1.2
samtools/1.6
scons/3.0.1/anaconda2.2.5
scotch/6.0.4/gnu-6.4.0
seismic_unix/43R1
seqping/0.1.45.1
shogun/1.0.2
siesta/4.1b1/openmpi-3.0.0-intel-2017
slurm-drmaa/1.0.7/gnu-6.4.0
smrtanalysis/2.3.0.140936_p5
spades/3.11.1
spark/2.2.1-bin-hadoop2.6
sphire/1.0
sratoolkit/2.9.0
ssu-align/0.1.1
tabix/0.2.6
tadbit/0.2.0.58/anaconda2env
tcltk/8.6.7-precompiled
texinfo/6.5
texlive/2017
tophat2/2.1.1
trinity/2.0.6(default)
trinity/2.5.1
turbomole/7.1
unicycler/0.4.1
vasp/5.4.1/ompi211-intel1602
vcftools/0.1.13
visit/2.13.0
visit/2.13.1-server(default)
vsearch/2.7.1
wannier90/2.1.0/openmpi-3.0.0-gcc-6.4.0
wanniertools/2.2.1/icc2016
wgs/8.2
```

How to use specific version of application

- **To show applications available:**

```
$ module avail
```

- **To show available versions of an application:**

```
$ module avail blast

    blast/2.2.25 blast/2.2.26
    blast/2.2.5  blast/2.6.0  blast/2.7.1
```

- **To load the default version of the application:**

```
$ module load <apps name>
```

- **To load a specific version of the application:**

```
$ module load blast/2.6.0
```

Python Programming/Scripting

- **We have Red Hat provided Python available in IBEX**

```
$ python
Python 2.7.5 (default, Aug  4 2017, 00:39:18)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-16)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>>
$ python -V
    Python 2.7.5
```

- **We have Python 2.x and 3.x versions available in IBEX SOFTWARE:**

```
    python/2.7.14          python/3.6.2
```

- **To load Python 2.x version:**

```
$ module load python/2.7.14
$ python -V
    Python 2.7.14
```

Python Programming/Scripting

- **Python packages like numpy, pandas, BioSQL) are not available in Red Hat Python**

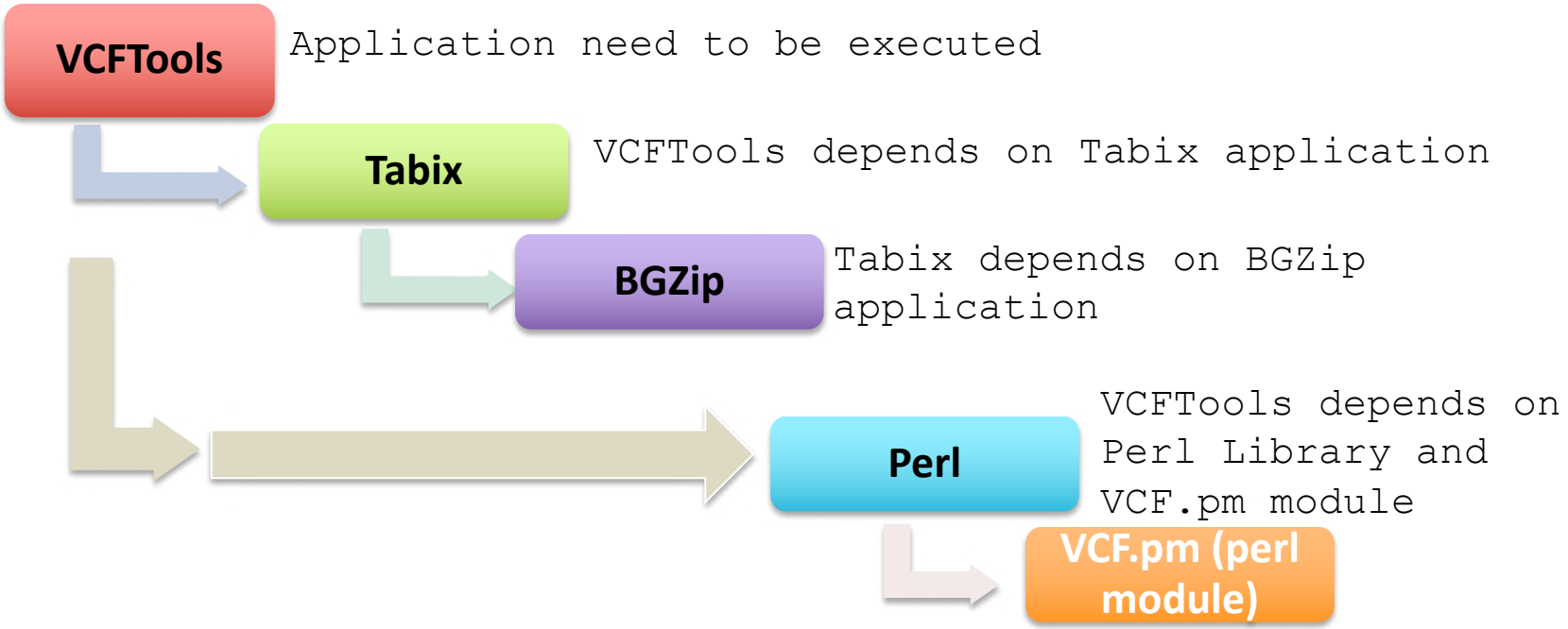
```
$ python
Python 2.7.5 (default, Aug  4 2017, 00:39:18)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-16)] on linux2
>>> import numpy
>>> import BioSQL
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
ImportError: No module named BioSQL
```

- **Python libraries or packages like numpy, pandas, BioSQL) are available in IBEX SOFTWARE.**

```
$ python
    Python 2.7.14 (default, Oct 26 2017, 22:19:05)
    [GCC 6.4.0] on linux2
    Type "help", "copyright", "credits" or "license" for more information.
    >>> import numpy
    >>> import BioSQL
```

Applications: Dependent libraries + other applications

- VCFtools is the application



module vcftools

```
 $ module load vcftools/0.1.13
Loading module for vcftools
vcftools 0.1.13 is now loaded
 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 NOTE: This vcftools depends on TABIX, BGZIP and VCF.pm modules which are AUTOMATICALLY LOADED
 ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
Loading module for tabix
Tabix 0.2.6 is now loaded
```

How to create a job script for FastQC (or for any application?)

- Check the availability of the application in IBEX
  - `module avail <application name>`
  - Example: `module avail fastqc`
- Include the application module (with appropriate or required version)
  - `module load <application name>`
  - Example: `module load fastqc/0.11.8`
- Understand the application parameters (e.g. *multi-threads, MPI, Python, Java …based options* etc.)
  - `<application name/binary> --help`
  - Example: `fastqc --help`
- You may use Job script generator available in IBEX.
  - https://www.hpc.kaust.edu.sa/ibex/job

My FastQC job

Step #1. Check the availability of the application

(If not available, please contact us: ibex@hpc.kaust.edu.sa)

```
[ $ module avail fastqc

------------------------------------------------------------------------ /sw/csi/modulefiles/applications
fastqc/0.11.8

------------------------------------------------------------------------ /cbrc/modules/applications ---
fastqc/0.11.5
```

Step #2. Include the application module (In the job script file!)

```
$ module load fastqc/0.11.8
Loading module for FastQC
FastQC 0.11.8 is now loaded
```

Step #3. Understand the application parameters!

```
$ fastqc --help
```

```
            FastQC - A high throughput sequence QC analysis tool

SYNOPSIS

        fastqc seqfile1 seqfile2 .. seqfileN

    fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
           [-c contaminant file] seqfile1 .. seqfileN
```

```
-t --threads    Specifies the number of files which can be processed
                simultaneously.  Each thread will be allocated 250MB of
                memory so you shouldn't run more threads than your
                available memory will cope with, and not more than
                6 threads on a 32 bit machine
```

https://www.hpc.kaust.edu.sa/ibex

IBEX Jobscript generator

Application Executable	fastqc /ibex/scratch/
Job Name	fastQC
Email Address to get notified	nagarajan.kathiresan@
Wallclock Time (duration of job)	1 h 30 m
Partition	batch
Processor	Intel (any)
Local Storage	No preference
memory	100 GB per node

MPI	OpenMP	Array
☐	☐	☐

Generate Script

-- Corresponding Ibex SLURM script --

```
#!/bin/bash
#SBATCH -N 1
#SBATCH --partition=batch
#SBATCH -J fastQC
#SBATCH -o fastQC.%J.out
#SBATCH -e fastQC.%J.err
#SBATCH --mail-user=nagarajan.kathiresan@kaust.edu.sa
#SBATCH --mail-type=ALL
#SBATCH --time=01:30:00
#SBATCH --mem=100G
#SBATCH --constraint=[intel]

#run the application:
fastqc  /ibex/scratch/kathirn/data/mutant_R1.fastq.gz  -o /ibex/scratch/kathirn/data/mutant    --threads 4
```

Customize the job script

```
#!/bin/bash
#SBATCH -N 1
#SBATCH --partition=batch
#SBATCH -J my_test_fastqc
#SBATCH -o my_test_fastqc.%J.out
#SBATCH -e my_test_fastqc.%J.err
#SBATCH --mail-user=nagarajan.kathiresan@kaust.edu.sa
#SBATCH --mail-type=ALL
#SBATCH --time=01:30:00
#SBATCH --mem=100G
#SBATCH --cpus-per-task=4
#SBATCH --constraint=[intel]

#run the application:
module load fastqc/0.11.8
mkdir -p /ibex/scratch/kathirn/data/mutant
fastqc  /ibex/scratch/kathirn/data/mutant_R1.fastq.gz \
     -o /ibex/scratch/kathirn/data/mutant \
     --threads 4
```

Submit your job

```
$ sbatch ./my_job.sh
Submitted batch job 1440437
 [dbn503-33-r Intel]:/home/kathirn
$ squeue -u $USER
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(RE
           1440437     batch my_test_  kathirn PD       0:00      1 (Priority)
```

Submit your job

```
total 1.7M
-rw-r--r-- 1 kathirn g-kathirn 295K Feb 24 14:37 mutant_R2_fastqc.
-rw-r--r-- 1 kathirn g-kathirn 535K Feb 24 14:37 mutant_R2_fastq
-rw-r--r-- 1 kathirn g-kathirn 284K Feb 24 14:37 mutant_R1_fastq
-rw-r--r-- 1 kathirn g-kathirn 532K Feb 24 14:37 mutant_R1_fastqc.
```

Check the output files

```
kw-16965:~ kathirn$ rsync -arvxHP kathirn@ilogin.ibex.kaust.edu.sa:/ibex/scratch/kathirn/work/mutante/ Desktop/.
receiving file list ...
5 files to consider
./
mutant_R1_fastqc.html
      543937 100%   12.06MB/s    0:00:00 (xfer#1, to-check=3/5)
mutant_R1_fastqc.zip
      290478 100%    4.07MB/s    0:00:00 (xfer#2, to-check=2/5)
mutant_R2_fastqc.html
      547381 100%    4.54MB/s    0:00:00 (xfer#3, to-check=1/5)
mutant_R2_fastqc.zip
      301166 100%    2.08MB/s    0:00:00 (xfer#4, to-check=0/5)

sent 110 bytes  received 1683714 bytes  3367648.00 bytes/sec
total size is 1682962  speedup is 1.00
kw-16965:~ kathirn$
```

Move the results back to Desktop or View using Google Chrome

Cont.

```
[[kathirn@dbn503-35-r:fastQC]$ google-chrome SRR975591_1_fastqc.html
[11619:11619:0217/104345.304536:ERROR:process_singleton_posix.cc(322)] The profile appears to be in use
an unlock the profile and relaunch Chrome.
[11657:11657:0217/104345.714827:ERROR:gl_surface_glx.cc(77)] XGetWindowAttributes failed for window 1677
[11657:11657:0217/104345.714903:ERROR:gl_surface_glx.cc(809)] Failed to get GLXConfig
[11657:11657:0217/104345.714939:ERROR:gpu_info_collector.cc(51)] gl::GLContext::CreateOffscreenGLSurface
[11657:11657:0217/104345.714971:ERROR:gpu_info_collector.cc(181)] Could not create surface for info coll
[11657:11657:0217/104345.714997:ERROR:gpu_init.cc(62)] gpu::CollectGraphicsInfo failed.
[11657:11657:0217/104345.728883:ERROR:viz_main_impl.cc(170)] Exiting GPU process due to errors during in
[11619:11619:0217/104345.739392:ERROR:desktop_window_tree_host_x11.cc(1121)] Not implemented reached in
[11619:11721:0217/104349.270653:ERROR:bus.cc(393)] Failed to connect to the bus: Could not parse server

(google-chrome:11619): LIBDBUSMENU-GLIB-WARNING **: 10:43:49.483: Unable to get session bus: Unknown or
```

Cont.

SRR975591_1.fastq.gz FastQC Report - Google Chrome

SRR975591_1.fastq.gz Fast

File | /ibex/scratch/projects/c2012/workshop/fastQC/SRR975591_1_fastqc.html

FastQC Report

Mon 17 Feb 2020
SRR975591_1.fastq.gz

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR975591_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	32911282
Sequences flagged as poor quality	0
Sequence length	101
%GC	51

Per base sequence quality

Produced by FastQC (version 0.11.8)

Review the results from Safari/Chrome/…

file:///Users/kathirn/Desktop/mutant_R1_fastqc.html

FastQC Report

Sun 24 Feb 2019
mutant_R1.fastq.gz

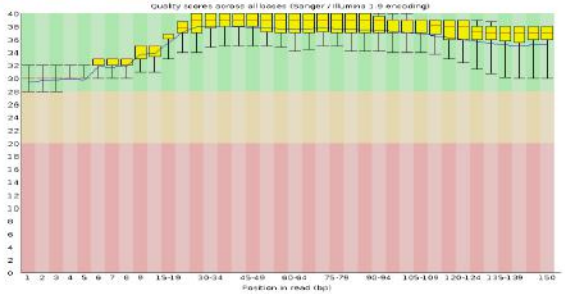
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

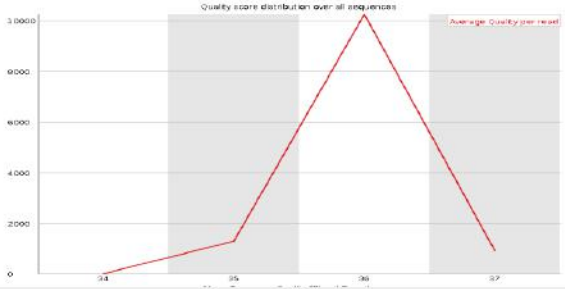
Basic Statistics

Measure	Value
Filename	mutant_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	12480
Sequences flagged as poor quality	0
Sequence length	150
%GC	33

Per base sequence quality

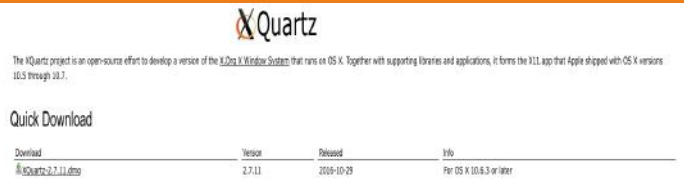


Per sequence quality scores



Produced by FastQC (version 0.11.8)

GUI Interface for FastQC

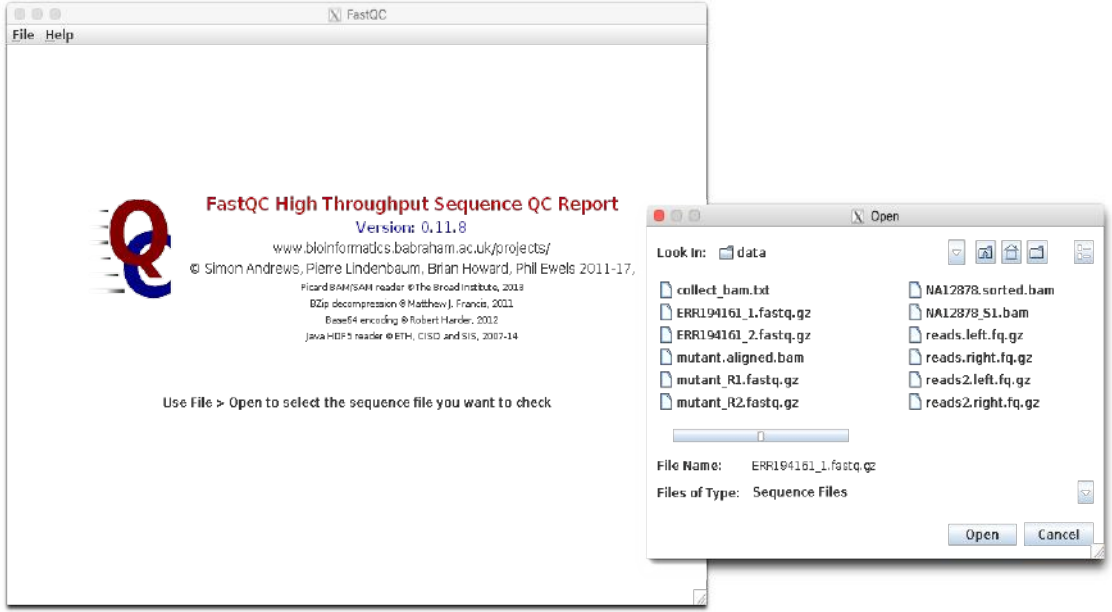


- Ensure Xquartz is installed: http://www.xquartz.org/
- Login with `-X11` forward support.
  - `ssh -X kathirn@ilogin.ibex.kaust.edu.sa`
- Get the `dedicated` **node to run your job**.
  - `srun --partition=batch --time=00:30 --cpus-per-task=2 --mem=32G --pty bash -i`

```
$ srun --partition=batch --time=00:30 --cpus-per-task=2 --mem=32G --pty bash -i
srun: job 1441204 queued and waiting for resources
srun: job 1441204 has been allocated resources
```

```
$ module load fastqc/0.11.8
$ fastqc
```

Screenshots!



FastQC

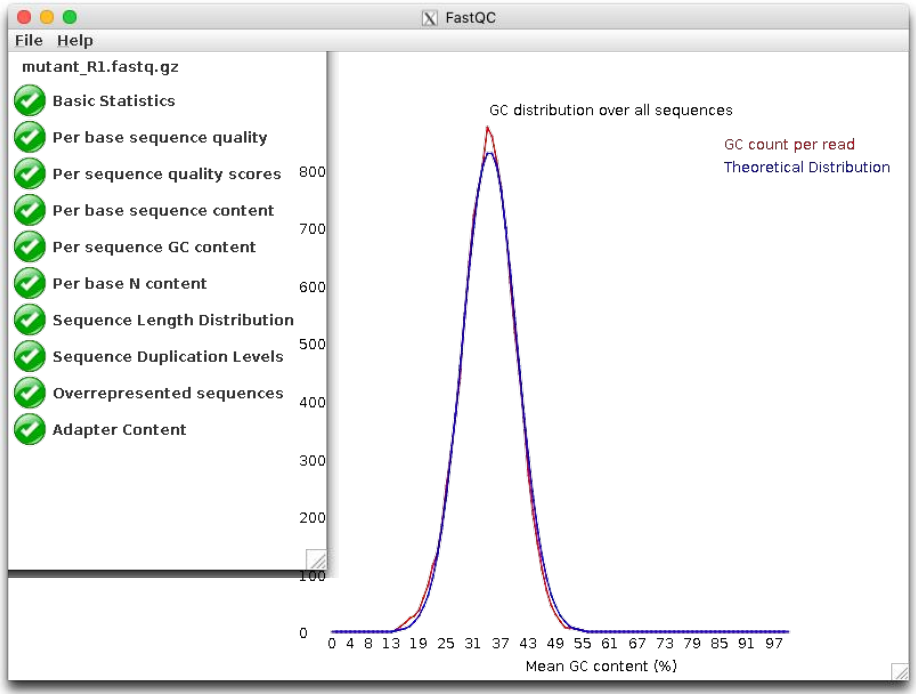
File Help

mutant_R1.fastq.gz mutant_R2.fastq.gz

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic sequence stats

Measure	Value
Filename	mutant_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	12480
Sequences flagged as poor quality	0
Sequence length	150
%GC	33



Multiple-data set!

- I have multiple FASTQ.GZ files
- I'm looking for FastQC on all these files!

```
$ l /ibex/scratch/kathirn/work/compBio2019/data/*.gz
-rw-r--r-- 1 kathirn g-kathirn  31M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.80xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  24M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.80xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  20M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.40xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  16M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.40xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 7.6M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.20xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 6.0M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.20xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 3.9M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.10xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 3.1M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR2627019.1m.10xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 4.4M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/SRR1284073.1m.5x.pacbio.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn  40M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/tardigrade_SRR2986339_subsampled_2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  39M Feb 23 22:41 /ibex/scratch/kathirn/work/compBio2019/data/tardigrade_SRR2986339_subsampled_1.fq.gz
```

Job Arrays

```
#!/bin/bash
#SBATCH -N 1
#SBATCH --partition=batch
#SBATCH -J my_test_fastqc
#SBATCH -o my_test_fastqc.%J.out
#SBATCH -e my_test_fastqc.%J.err
#SBATCH --time=0:10:00
#SBATCH --mem=10G
#SBATCH --cpus-per-task=4
#SBATCH --constraint=[intel]
#SBATCH --array=1-4
#SBATCH --mail-type=ALL
#SBATCH --mail-user=nagarajan.kathiresan@kaust.edu.sa

## Include application modules
    module load fastqc/0.11.8

## To understand the parallel (array) jobs
   echo "SLURM_JOBID: " $SLURM_JOBID
   echo "SLURM_ARRAY_TASK_ID: " $SLURM_ARRAY_TASK_ID
   echo "SLURM_ARRAY_JOB_ID: " $SLURM_ARRAY_JOB_ID

## run the application
   cd /ibex/scratch/kathirn/work/compBio2019/data ;
   fq=`ls *.gz | head -n $SLURM_ARRAY_TASK_ID | tail -n 1`
   fastqc  $fq -o /ibex/scratch/kathirn/work/mutante --threads 4
```

Job array & data selection

```
$ ls -lrta *.gz
-rw-r--r-- 1 kathirn g-kathirn 31657458 Feb 23 22:41 SRR2627019.1m.80xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 24862679 Feb 23 22:41 SRR2627019.1m.80xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 20010219 Feb 23 22:41 SRR2627019.1m.40xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 15834251 Feb 23 22:41 SRR2627019.1m.40xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  7946629 Feb 23 22:41 SRR2627019.1m.20xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  6247794 Feb 23 22:41 SRR2627019.1m.20xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  4001229 Feb 23 22:41 SRR2627019.1m.10xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  3146371 Feb 23 22:41 SRR2627019.1m.10xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn  4540280 Feb 23 22:41 SRR1284073.1m.5x.pacbio.fastq.gz
-rw-r--r-- 1 kathirn g-kathirn 41208103 Feb 23 22:41 tardigrade_SRR2986339_subsampled_2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 40273924 Feb 23 22:41 tardigrade_SRR2986339_subsampled_1.fq.gz
```

```
$ ls -lrta *.gz | head -n 1
-rw-r--r-- 1 kathirn g-kathirn 31657458 Feb 23 22:41 SRR2627019.1m.80xR2.fq.gz
```

```
$ ls -lrta *.gz | head -n 1 | tail -1
-rw-r--r-- 1 kathirn g-kathirn 31657458 Feb 23 22:41 SRR2627019.1m.80xR2.fq.gz
```

```
$ ls -lrta *.gz | head -n 3
-rw-r--r-- 1 kathirn g-kathirn 31657458 Feb 23 22:41 SRR2627019.1m.80xR2.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 24862679 Feb 23 22:41 SRR2627019.1m.80xR1.fq.gz
-rw-r--r-- 1 kathirn g-kathirn 20010219 Feb 23 22:41 SRR2627019.1m.40xR2.fq.gz
```

```
$ ls -lrta *.gz | head -n 3 | tail -1
-rw-r--r-- 1 kathirn g-kathirn 20010219 Feb 23 22:41 SRR2627019.1m.40xR2.fq.gz
```

Job Arrays

```
$ sbatch ./my_job.sh
Submitted batch job 1442411
```

Submit a single job

```
Sun Feb 24 16:24:21 2019
             JOBID PARTITION     NAME     USER    STATE       TIME TIME_LIMI  NODES NODELIST(REASON)
     1442411_[1-4]     batch my_test_  kathirn  PENDING       0:00     10:00      1 (Priority)
```

Submit in pending state

```
Sun Feb 24 16:26:19 2019
             JOBID PARTITION     NAME     USER    STATE       TIME TIME_LIMI  NODES NODELIST(REASON)
         1442411_1     batch my_test_  kathirn  RUNNING       0:00     10:00      1 ds506-13
         1442411_2     batch my_test_  kathirn  RUNNING       0:00     10:00      1 ds505-01
         1442411_3     batch my_test_  kathirn  RUNNING       0:00     10:00      1 ds505-01
         1442411_4     batch my_test_  kathirn  RUNNING       0:00     10:00      1 ds505-29
```

Parallel runs

```
Sun Feb 24 16:26:26 2019
             JOBID PARTITION     NAME     USER    STATE       TIME TIME_LIMI  NODES NODELIST(REASON)
         1442411_2     batch my_test_  kathirn COMPLETI       0:07     10:00      1 ds505-01
         1442411_1     batch my_test_  kathirn  RUNNING       0:07     10:00      1 ds506-13
         1442411_3     batch my_test_  kathirn  RUNNING       0:07     10:00      1 ds505-01
         1442411_4     batch my_test_  kathirn  RUNNING       0:07     10:00      1 ds505-29
```

Complete 1 by 1

All samples results are ready

```
-rw-r--r-- 1 kathirn g-kathirn 258K Feb 24 16:26 SRR2627019.1m.10xR2_fastqc.zip
-rw-r--r-- 1 kathirn g-kathirn 240K Feb 24 16:26 SRR2627019.1m.10xR2_fastqc.html
-rw-r--r-- 1 kathirn g-kathirn 247K Feb 24 16:26 SRR2627019.1m.10xR1_fastqc.zip
-rw-r--r-- 1 kathirn g-kathirn 236K Feb 24 16:26 SRR2627019.1m.10xR1_fastqc.html
-rw-r--r-- 1 kathirn g-kathirn 249K Feb 24 16:26 SRR2627019.1m.20xR1_fastqc.zip
-rw-r--r-- 1 kathirn g-kathirn 237K Feb 24 16:26 SRR2627019.1m.20xR1_fastqc.html
-rw-r--r-- 1 kathirn g-kathirn 284K Feb 24 16:26 SRR1284073.1m.5x.pacbio_fastqc.zip
-rw-r--r-- 1 kathirn g-kathirn 321K Feb 24 16:26 SRR1284073.1m.5x.pacbio_fastqc.html
```

New applications request or Issues!

Contact for Help/Support:

ibex@hpc.kaust.edu.sa

Our website:

https://www.hpc.kaust.edu.sa/ibex

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

IBEX

Thank you!

nagarajan.kathiresan@kaust.edu.sa