생명정보학에서 쓰이는 컴퓨터 Perl 언어의 기초 교육

Raw sequences of NGS technologies and Perl program

2017/9/16

Jongsun Park, Ph. D.

성신여자대학교
SUNGSHIN UNIVERSITY

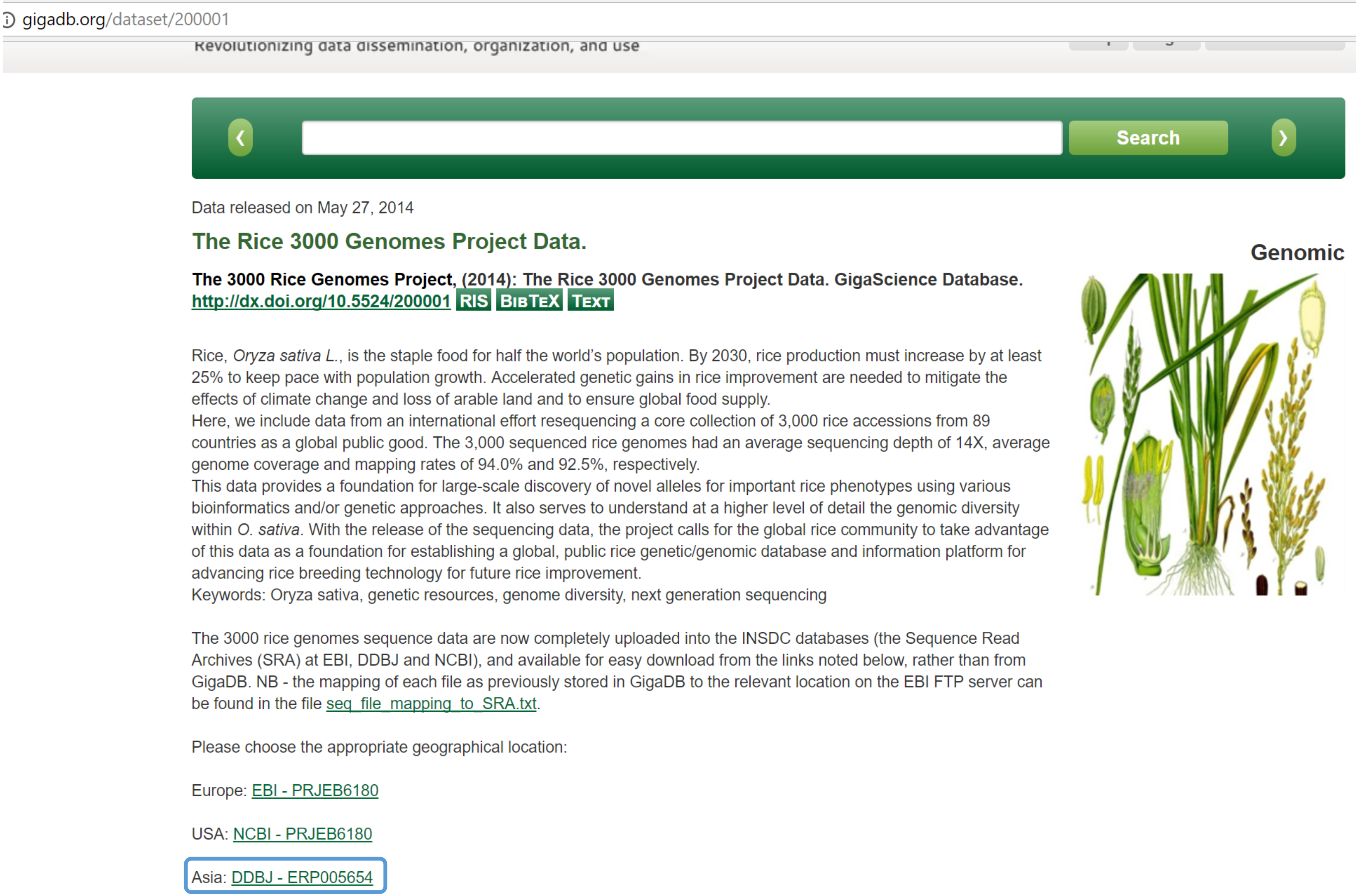
서울시 강서구 화곡동 359-63 영주빌딩 201호
Tel 02 2698 1188 / Fax 02 6280 8821

www.infoboss.co.kr

InfoBoss
Infomations & Systems

Raw sequence of NGS (1) / Raw sequences generated by NGS machines

- We will reconstruct rice (*Oryza sativa*) genomes based on raw data of 3000 rice genomes.


gigadb.org/dataset/200001
Revolutionizing data dissemination, organization, and use
❮
Search
❯
RIS
BibTeX
Text
Genomic

Data released on May 27, 2014

The Rice 3000 Genomes Project Data.

The 3000 Rice Genomes Project, (2014): The Rice 3000 Genomes Project Data. GigaScience Database. http://dx.doi.org/10.5524/200001

Rice, *Oryza sativa L.*, is the staple food for half the world's population. By 2030, rice production must increase by at least 25% to keep pace with population growth. Accelerated genetic gains in rice improvement are needed to mitigate the effects of climate change and loss of arable land and to ensure global food supply.
Here, we include data from an international effort resequencing a core collection of 3,000 rice accessions from 89 countries as a global public good. The 3,000 sequenced rice genomes had an average sequencing depth of 14X, average genome coverage and mapping rates of 94.0% and 92.5%, respectively.
This data provides a foundation for large-scale discovery of novel alleles for important rice phenotypes using various bioinformatics and/or genetic approaches. It also serves to understand at a higher level of detail the genomic diversity within *O. sativa*. With the release of the sequencing data, the project calls for the global rice community to take advantage of this data as a foundation for establishing a global, public rice genetic/genomic database and information platform for advancing rice breeding technology for future rice improvement.
Keywords: Oryza sativa, genetic resources, genome diversity, next generation sequencing

The 3000 rice genomes sequence data are now completely uploaded into the INSDC databases (the Sequence Read Archives (SRA) at EBI, DDBJ and NCBI), and available for easy download from the links noted below, rather than from GigaDB. NB - the mapping of each file as previously stored in GigaDB to the relevant location on the EBI FTP server can be found in the file seq_file_mapping_to_SRA.txt.

Please choose the appropriate geographical location:

Europe: EBI - PRJEB6180

USA: NCBI - PRJEB6180

Asia: DDBJ - ERP005654

Raw sequence of NGS (2)

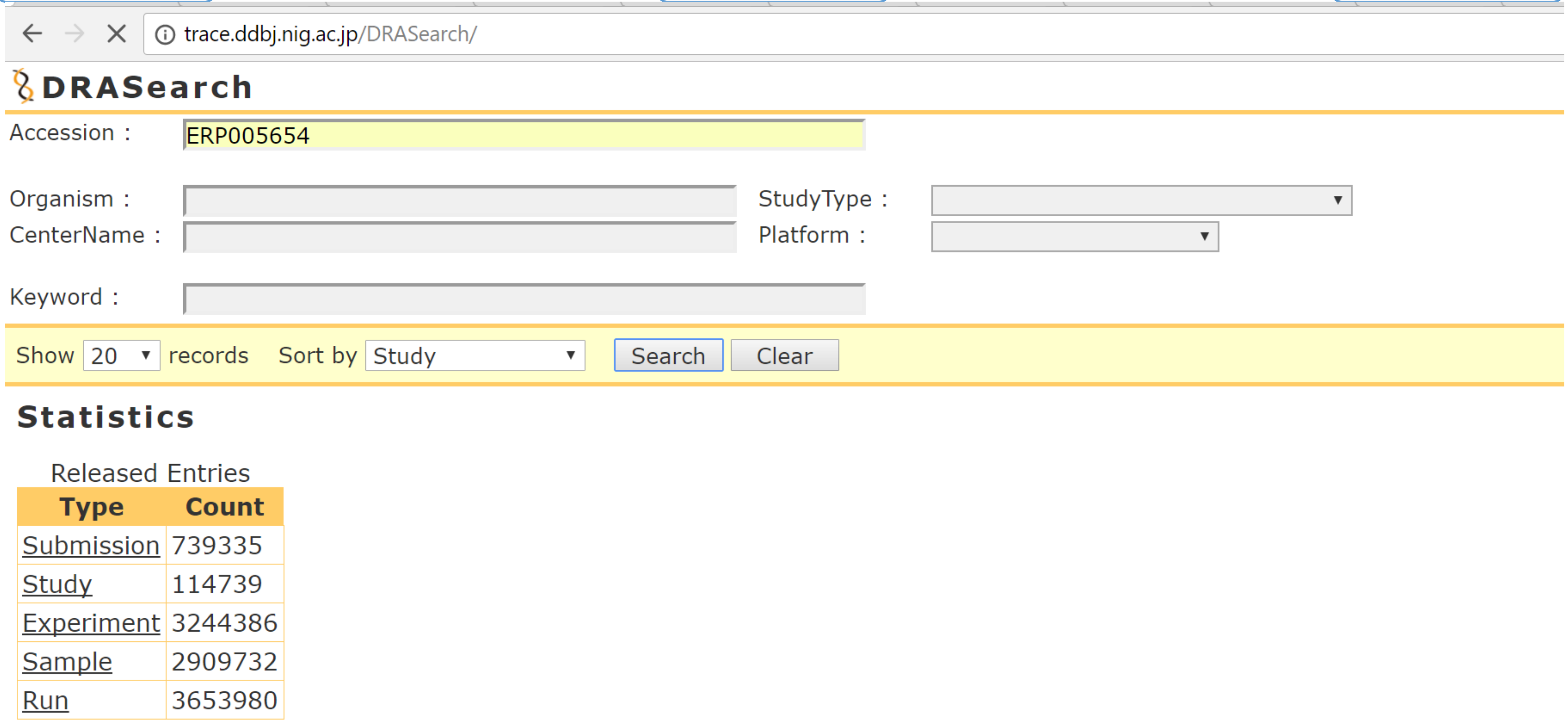
Raw sequences generated by NGS machines

- Raw data can be found in NCBI Short Read Archive (SRA) or EMBL European Nucleotide Archive or DDBJ DDBJ Read Archive (DRA).

NCBI (SRA)

EMBL (ENA)

DDBJ (DRA)



trace.ddbj.nig.ac.jp/DRASearch/

DRASearch

Accession : ERP005654

Organism :

StudyType :

CenterName :

Platform :

Keyword :

Show 20 records Sort by Study

Statistics

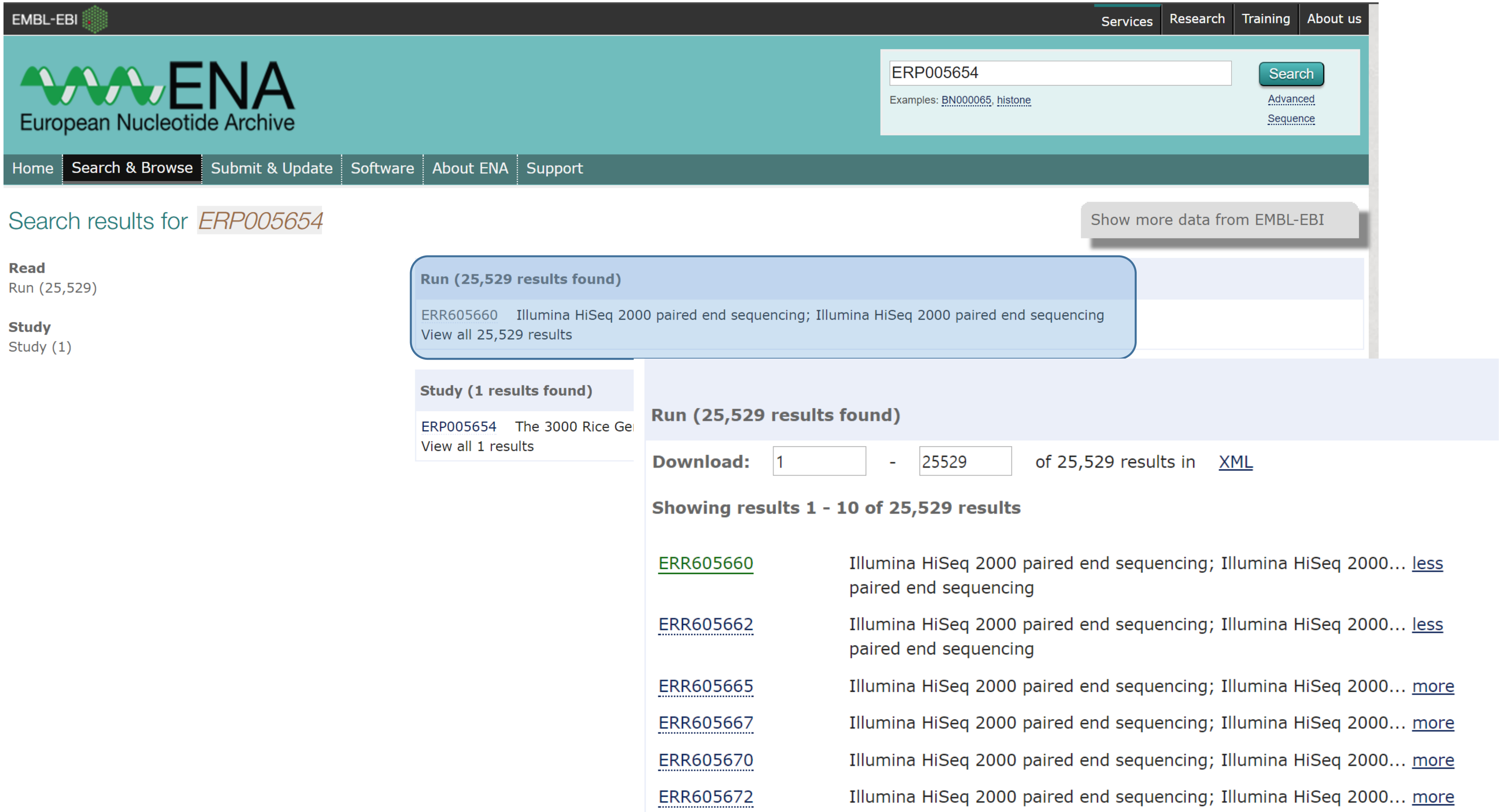
Released Entries

Type	Count
Submission	739335
Study	114739
Experiment	3244386
Sample	2909732
Run	3653980

Raw sequence of NGS (3)

Raw sequences generated by NGS machines

- Usually, we can download data via DDBJ or EMBL because NCBI ask us to use special toolkit to download.
- Three archives are synchronized periodically so that data in three archives are same.



Search results for *ERP005654*

Read
Run (25,529)

Study
Study (1)

Run (25,529 results found)

ERR605660 Illumina HiSeq 2000 paired end sequencing; Illumina HiSeq 2000 paired end sequencing
View all 25,529 results

Study (1 results found)

ERP005654 The 3000 Rice Ge
View all 1 results

Run (25,529 results found)

Download: 1 - 25529 of 25,529 results in XML

Showing results 1 - 10 of 25,529 results

ERR605660 Illumina HiSeq 2000 paired end sequencing; Illumina HiSeq 2000... less
paired end sequencing

ERR605662 Illumina HiSeq 2000 paired end sequencing; Illumina HiSeq 2000... less
paired end sequencing

ERR605665 Illumina HiSeq 2000 paired end sequencing; Illumina HiSeq 2000... more

ERR605667 Illumina HiSeq 2000 paired end sequencing; Illumina HiSeq 2000... more

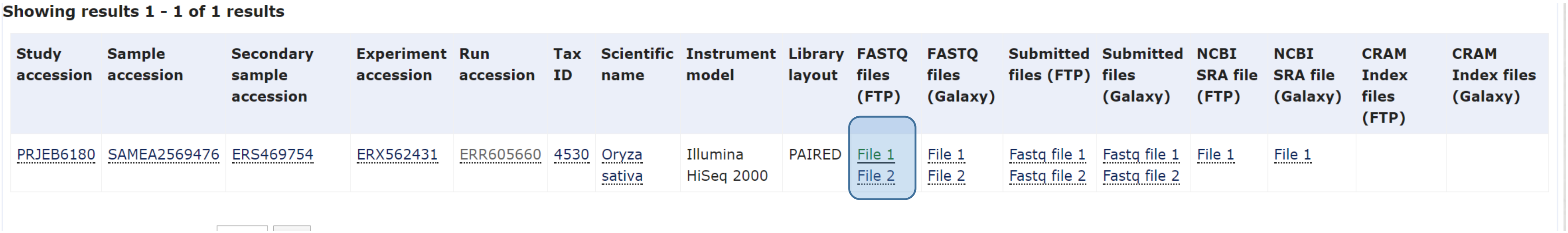
ERR605670 Illumina HiSeq 2000 paired end sequencing; Illumina HiSeq 2000... more

ERR605672 Illumina HiSeq 2000 paired end sequencing; Illumina HiSeq 2000... more

Raw sequence of NGS (4)

Raw sequences generated by NGS machines

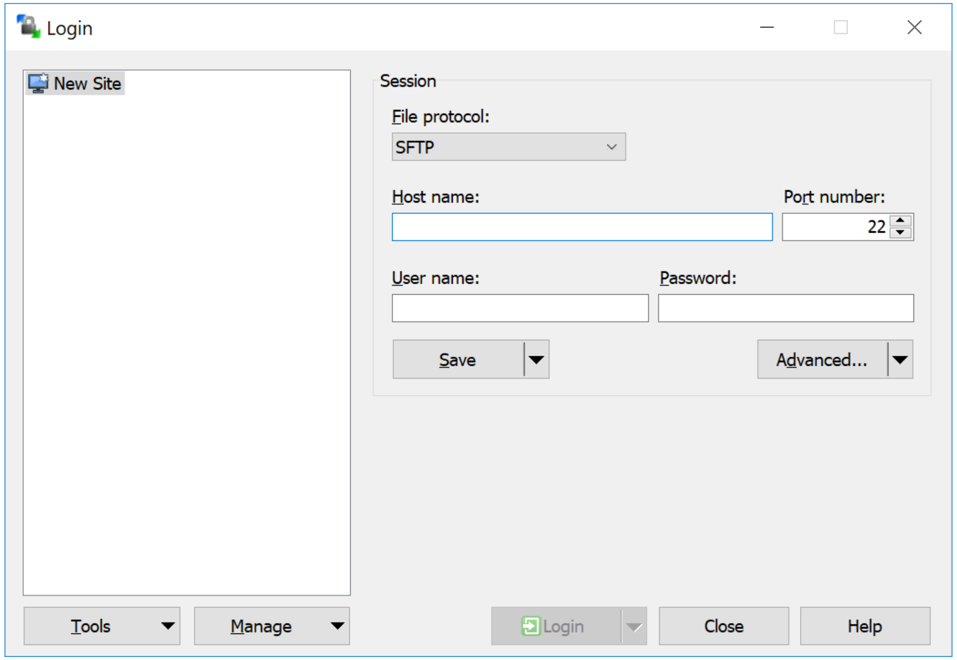
- Download fastq files from the below link.



Showing results 1 - 1 of 1 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)
PRJEB6180	SAMEA2569476	ERS469754	ERX562431	ERR605660	4530	Oryza sativa	Illumina HiSeq 2000	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		

- Downloaded files are compressed files because of large size of file.
- gz format is better uncompressed in server side because we have to transfer file from local computer to server.
- To transfer files, WinSCP program can be used.
- Server IP, port, user account information are required.



Login

New Site

Session

File protocol: SFTP

Host name:

Port number: 22

User name:

Password:

Raw sequence of NGS (5)

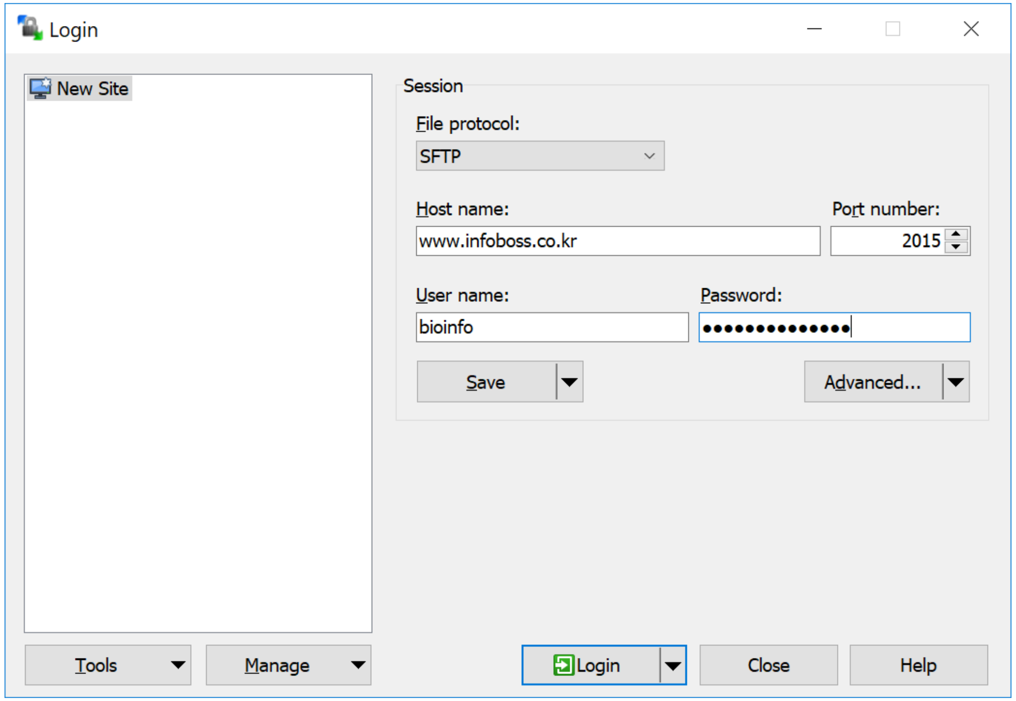
Raw sequences generated by NGS machines

- Server information is like below:

IP : www.infoboss.co.kr
Port : 2015
ID : bioinfo
Password : todanfwjdqhgkr

- Using WinSCP, let’s upload files after connecting servers.

- Please use personal folder after login, if not, you may not lose your data for the remaining lecture!



Session

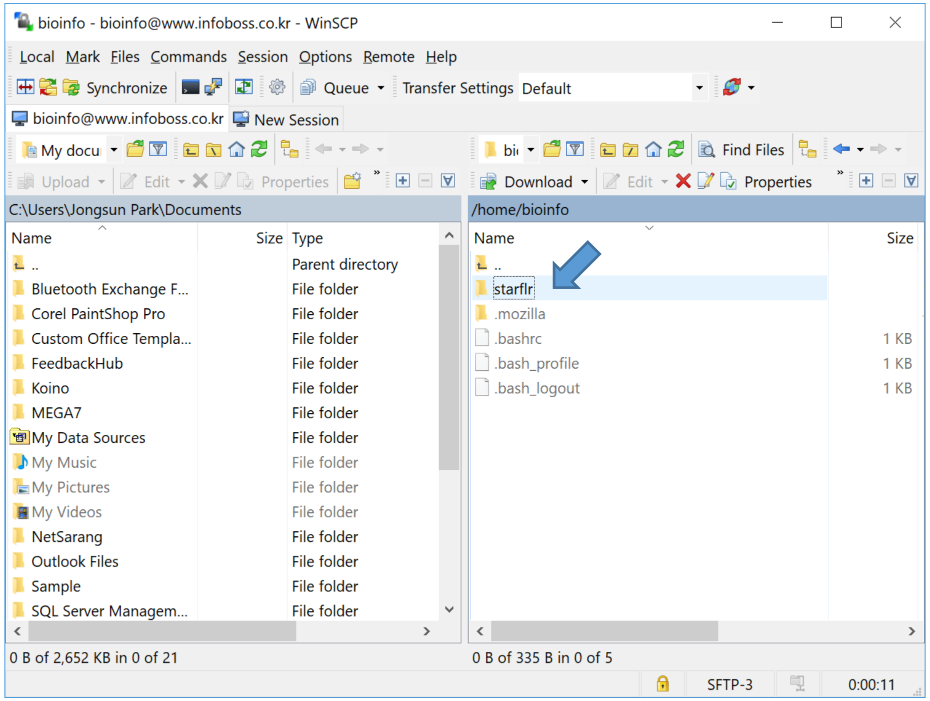
File protocol: SFTP

Host name: www.infoboss.co.kr

Port number: 2015

User name: bioinfo

Password: ••••••••••••••



bioinfo - bioinfo@www.infoboss.co.kr - WinSCP

Transfer Settings Default

bioinfo@www.infoboss.co.kr

C:\Users\Jongsun Park\Documents

Name	Size	Type
..		Parent directory
Bluetooth Exchange F...		File folder
Corel PaintShop Pro		File folder
Custom Office Templa...		File folder
FeedbackHub		File folder
Koino		File folder
MEGA7		File folder
My Data Sources		File folder
My Music		File folder
My Pictures		File folder
My Videos		File folder
NetSarang		File folder
Outlook Files		File folder
Sample		File folder
SQL Server Managem...		File folder

0 B of 2,652 KB in 0 of 21

/home/bioinfo

Name	Size
..	
starflr	
.mozilla	
.bashrc	1 KB
.bash_profile	1 KB
.bash_logout	1 KB

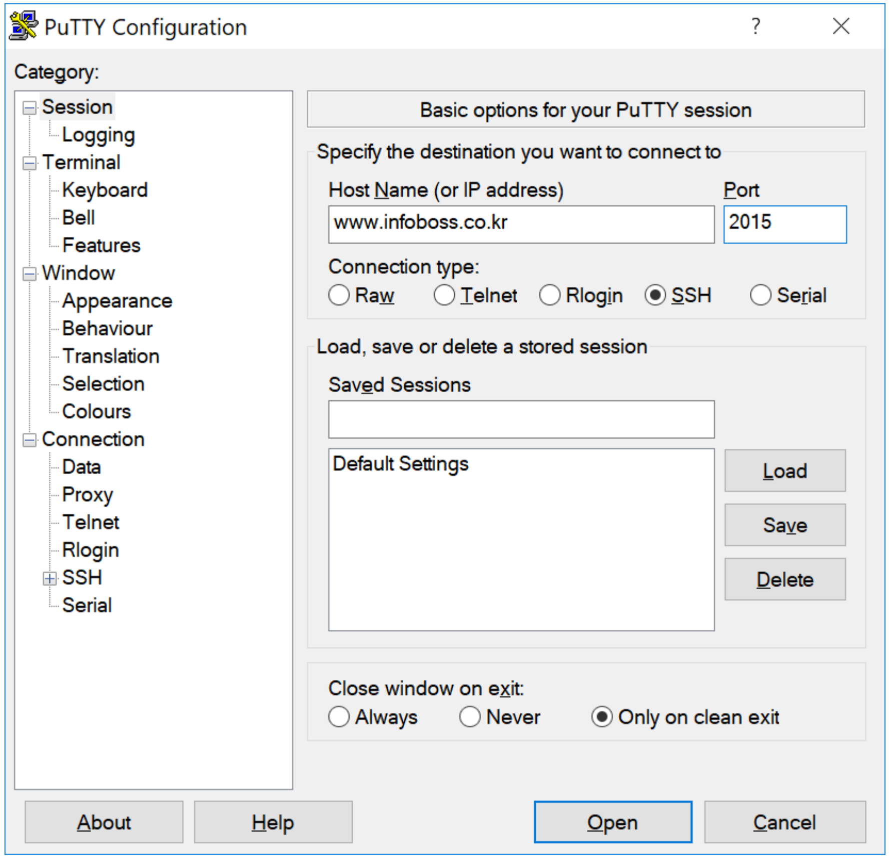
0 B of 335 B in 0 of 5

SFTP-3 0:00:11

Basic Linux Commands (1)

Basic commands to operate servers

- Let’s login to server using Putty.



Specify the destination you want to connect to

Host Name (or IP address): www.infoboss.co.kr

Port: 2015

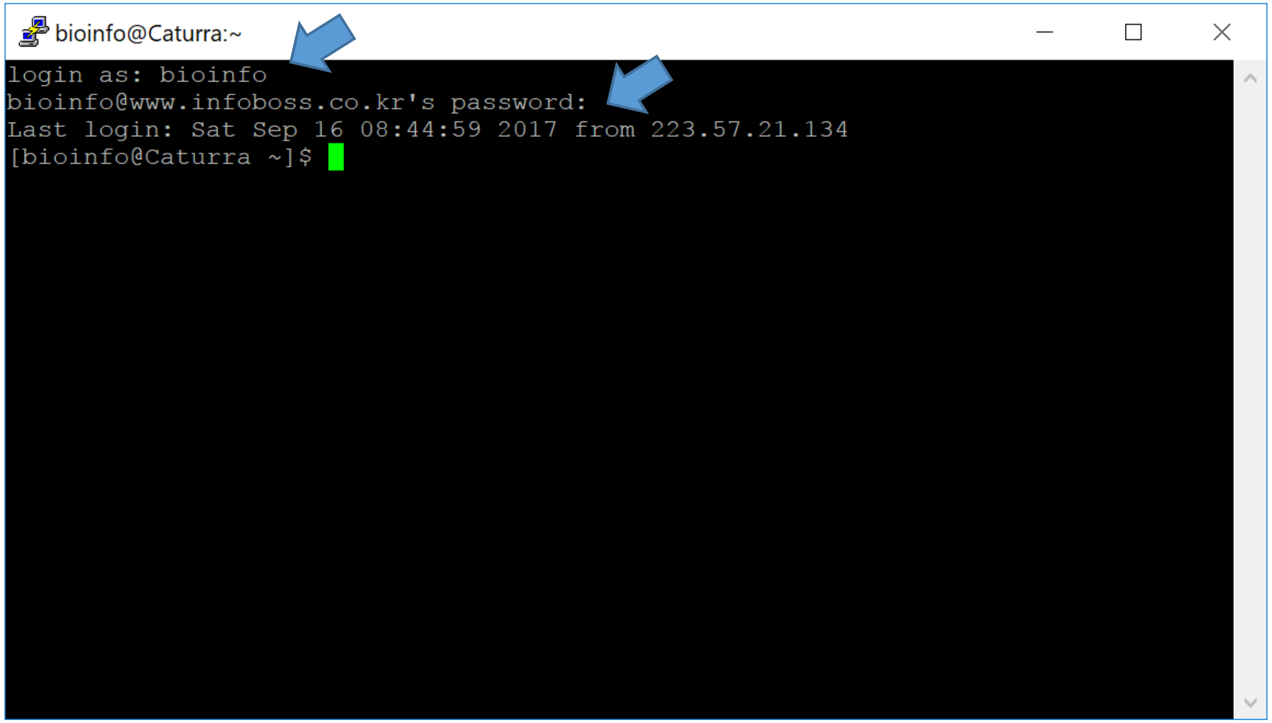
Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

Saved Sessions:

Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit



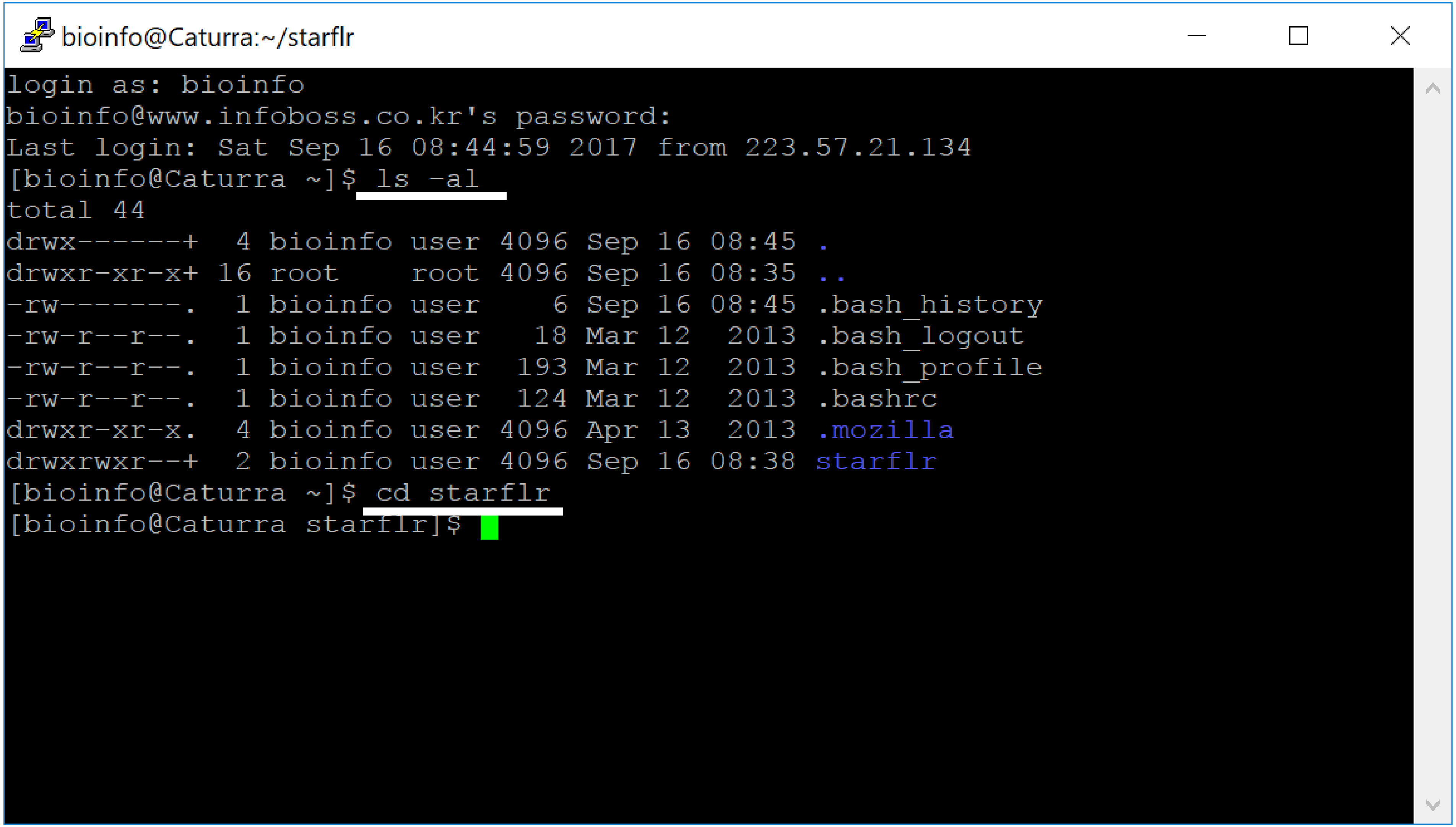
```
login as: bioinfo
bioinfo@www.infoboss.co.kr's password:
Last login: Sat Sep 16 08:44:59 2017 from 223.57.21.134
[bioinfo@Caturra ~]$ █
```

Basic Linux Commands (2)

Basic commands to operate servers

- List of files and directories : ls -al

- Change directory: cd



```
login as: bioinfo
bioinfo@www.infoboss.co.kr's password:
Last login: Sat Sep 16 08:44:59 2017 from 223.57.21.134
[bioinfo@Caturra ~]$ ls -al
total 44
drwx------+  4 bioinfo user 4096 Sep 16 08:45 .
drwxr-xr-x+ 16 root    root 4096 Sep 16 08:35 ..
-rw-------.  1 bioinfo user    6 Sep 16 08:45 .bash_history
-rw-r--r--.  1 bioinfo user   18 Mar 12  2013 .bash_logout
-rw-r--r--.  1 bioinfo user  193 Mar 12  2013 .bash_profile
-rw-r--r--.  1 bioinfo user  124 Mar 12  2013 .bashrc
drwxr-xr-x.  4 bioinfo user 4096 Apr 13  2013 .mozilla
drwxrwxr--+  2 bioinfo user 4096 Sep 16 08:38 starflr
[bioinfo@Caturra ~]$ cd starflr
[bioinfo@Caturra starflr]$
```

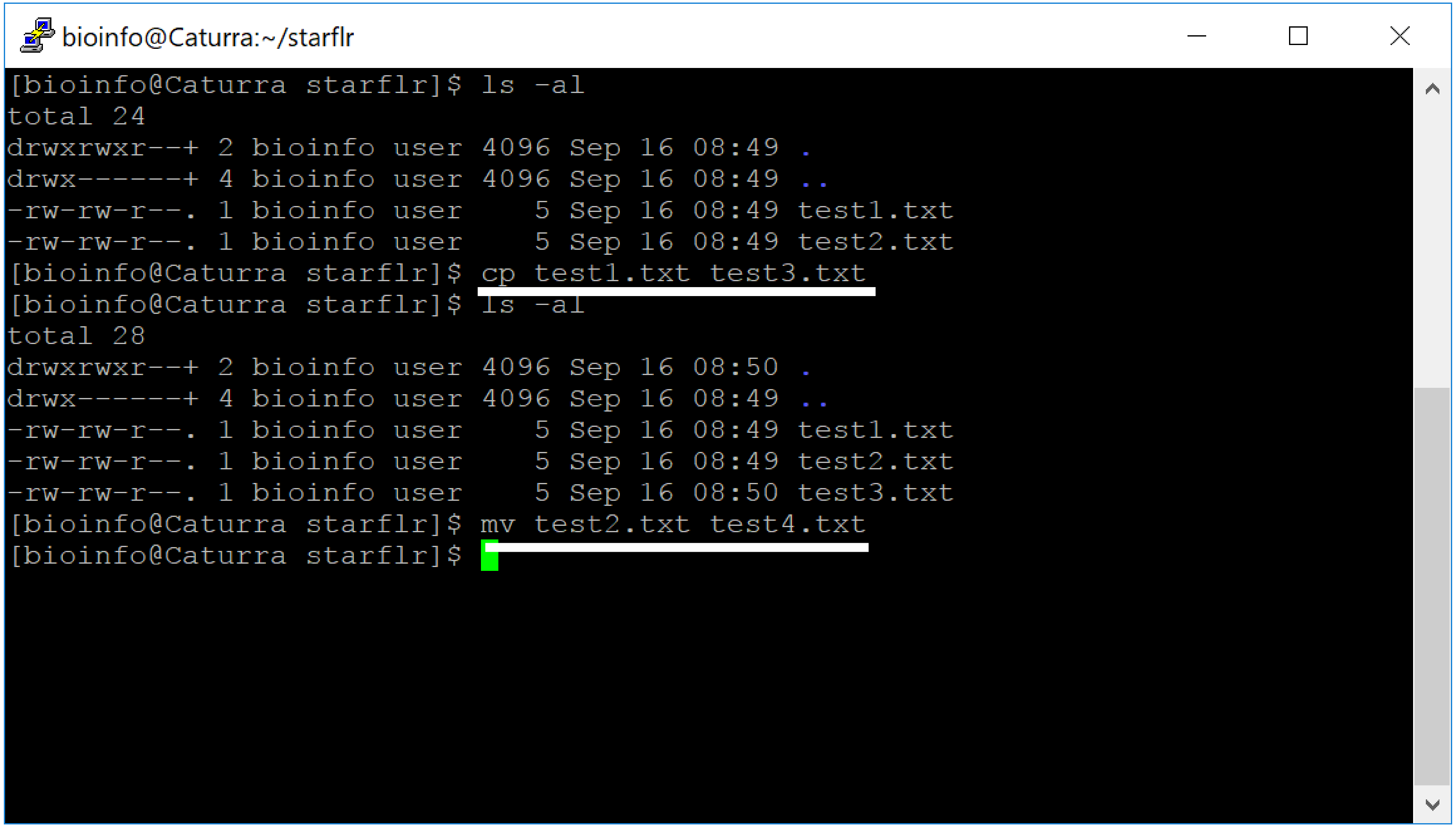
Basic Linux Commands (3)

Basic commands to operate servers

- Copy files: cp

- Move files : mv

- Rename filename?



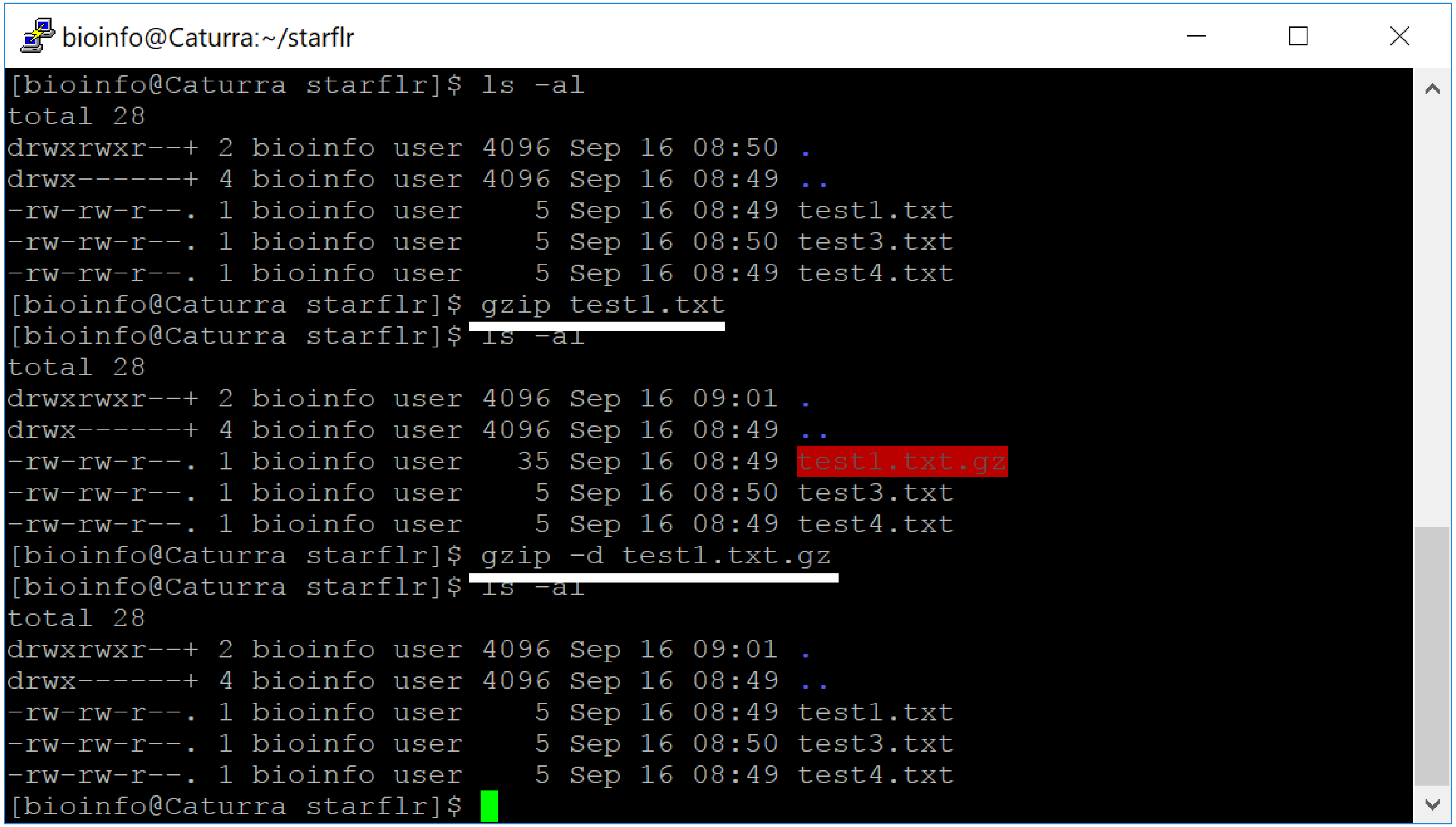
```
[bioinfo@Caturra starflr]$ ls -al
total 24
drwxrwxr--+ 2 bioinfo user 4096 Sep 16 08:49 .
drwx------+ 4 bioinfo user 4096 Sep 16 08:49 ..
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test1.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test2.txt
[bioinfo@Caturra starflr]$ cp test1.txt test3.txt
[bioinfo@Caturra starflr]$ ls -al
total 28
drwxrwxr--+ 2 bioinfo user 4096 Sep 16 08:50 .
drwx------+ 4 bioinfo user 4096 Sep 16 08:49 ..
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test1.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test2.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:50 test3.txt
[bioinfo@Caturra starflr]$ mv test2.txt test4.txt
[bioinfo@Caturra starflr]$
```

Basic Linux Commands (4)

Basic commands to operate servers

- Compress file using gzip : gzip [filename]

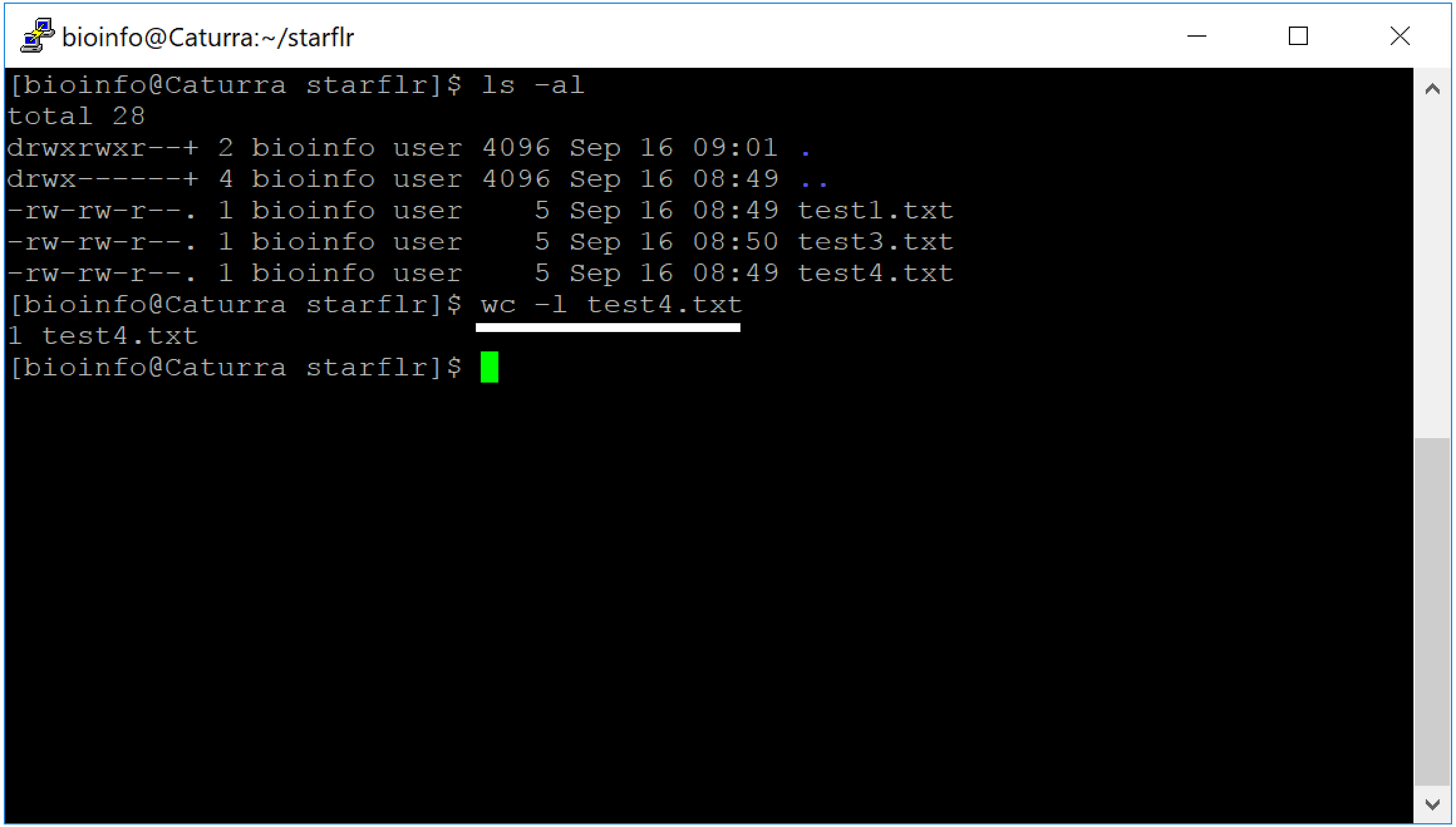
- Uncompress file using gzip : gzip –d [filename]



```
[bioinfo@Caturra starflr]$ ls -al
total 28
drwxrwxr--+ 2 bioinfo user 4096 Sep 16 08:50 .
drwx------+ 4 bioinfo user 4096 Sep 16 08:49 ..
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test1.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:50 test3.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test4.txt
[bioinfo@Caturra starflr]$ gzip test1.txt
[bioinfo@Caturra starflr]$ ls -al
total 28
drwxrwxr--+ 2 bioinfo user 4096 Sep 16 09:01 .
drwx------+ 4 bioinfo user 4096 Sep 16 08:49 ..
-rw-rw-r--. 1 bioinfo user   35 Sep 16 08:49 test1.txt.gz
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:50 test3.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test4.txt
[bioinfo@Caturra starflr]$ gzip -d test1.txt.gz
[bioinfo@Caturra starflr]$ ls -al
total 28
drwxrwxr--+ 2 bioinfo user 4096 Sep 16 09:01 .
drwx------+ 4 bioinfo user 4096 Sep 16 08:49 ..
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test1.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:50 test3.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test4.txt
[bioinfo@Caturra starflr]$
```

Basic Linux Commands (5)

- Counting number of lines in file: wc –l [filename]



```
[bioinfo@Caturra starflr]$ ls -al
total 28
drwxrwxr--+ 2 bioinfo user 4096 Sep 16 09:01 .
drwx------+ 4 bioinfo user 4096 Sep 16 08:49 ..
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test1.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:50 test3.txt
-rw-rw-r--. 1 bioinfo user    5 Sep 16 08:49 test4.txt
[bioinfo@Caturra starflr]$ wc -l test4.txt
1 test4.txt
[bioinfo@Caturra starflr]$
```

Perl Program for Cutting Small Part of File

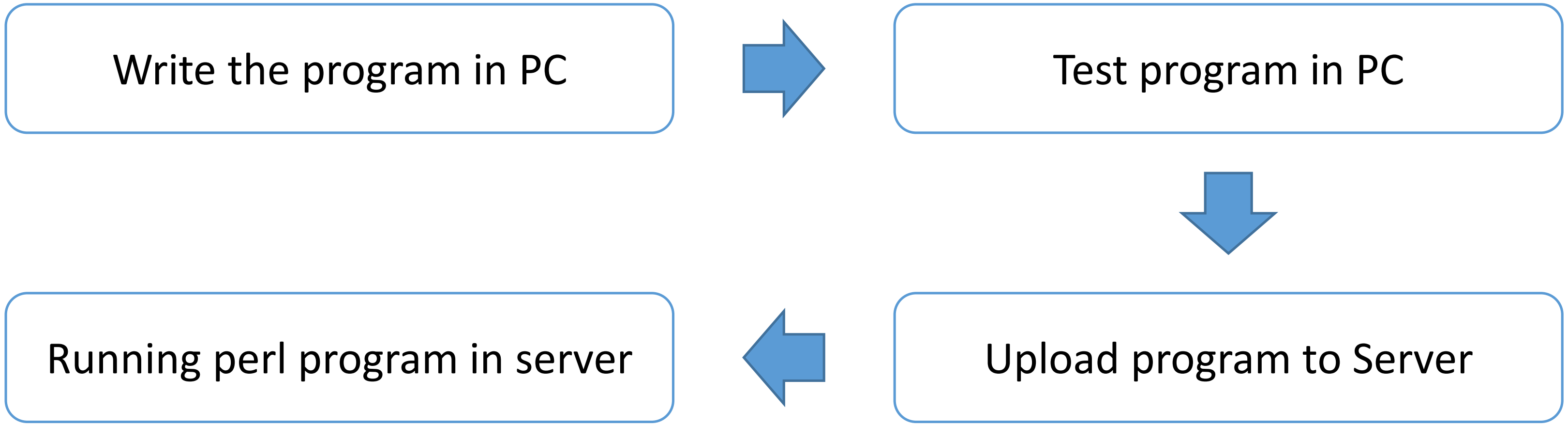
- To run perl program in Linux, please use the below header:

```
#!/usr/bin/perl –w

use strict;
```

- Let’s make program to cut 100,000 lines from the downloaded files because that file is too big to test during lecture!

PC31.pl



Write the program in PC

Test program in PC

Upload program to Server

Running perl program in server

Thank you for your attention!

If you have question,
please ask! =)

starflr@infoboss.co.kr

InfoBoss
Infomations & Systems