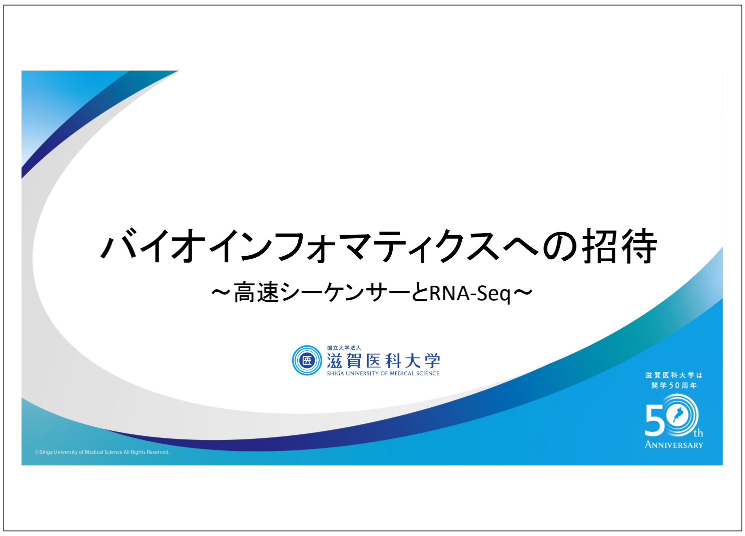
バイオインフォマティクスへの招待 ～高速シーケンサーと RNA-Seq～（講義・実習）

Introduction of Bioinformatics - Next Generation Sequencer and RNA-seq (Lecture/Practice in Japanese)

門田 陽介・芦原 貴司（情報総合センター・医療情報部）
Yosuke Kadota, Takashi Ashihara (Information Technology and Management Center)

この講義・実習では、高速シーケンサーによる塩基配列の読み取りの原理を概説し、得られた膨大な塩基配列データから遺伝子の発現解析（RNA-Seq）を行う手法について、実際にマルチメディアセンターの PC を各自が使ってハンズオンで学ぶ。

We will show you outline the principles of sequence reading by high-speed sequencers, and let you know how to analysis gene expression (RNA-Seq) from the huge amount of sequence data obtained from high-speed sequencers, by using the PCs in the Multimedia Center by yourselves.

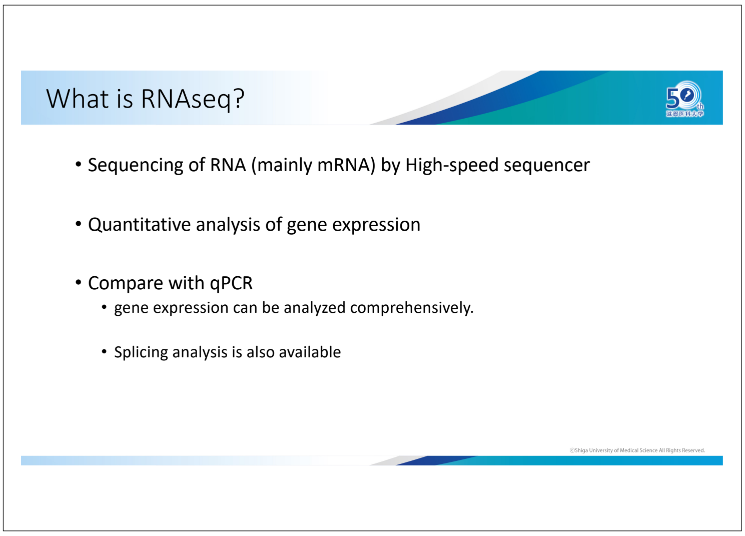


バイオインフォマティクスへの招待

～高速シーケンサーとRNA-Seq～

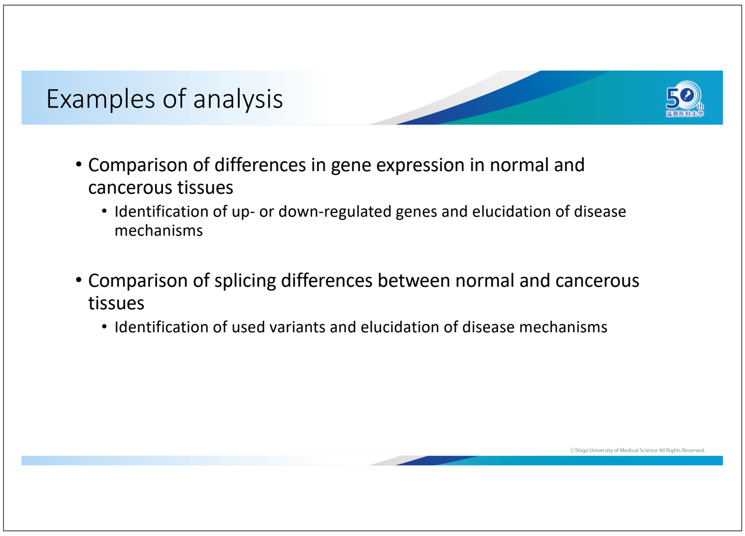
国立大学法人 滋賀医科大学 SHIGA UNIVERSITY OF MEDICAL SCIENCE

滋賀医科大学は開学50周年 50th ANNIVERSARY



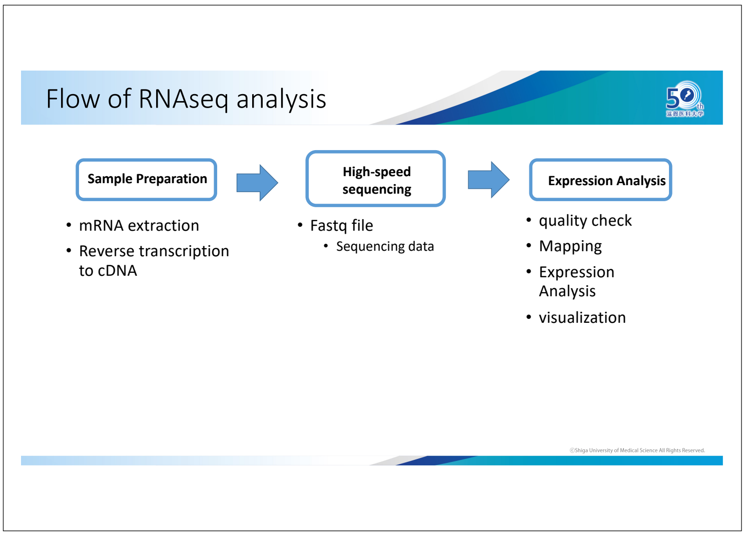
What is RNAseq?

- Sequencing of RNA (mainly mRNA) by High-speed sequencer
- Quantitative analysis of gene expression
- Compare with qPCR
  - gene expression can be analyzed comprehensively.
  - Splicing analysis is also available



Examples of analysis

- Comparison of differences in gene expression in normal and cancerous tissues
  - Identification of up- or down-regulated genes and elucidation of disease mechanisms
- Comparison of splicing differences between normal and cancerous tissues
  - Identification of used variants and elucidation of disease mechanisms



Flow of RNAseq analysis

Sample Preparation

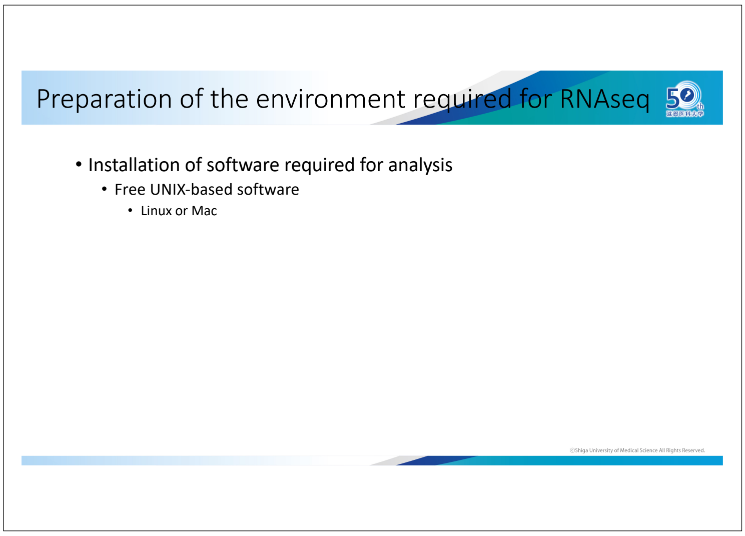
- mRNA extraction
- Reverse transcription to cDNA

High-speed sequencing

- Fastq file
  - Sequencing data

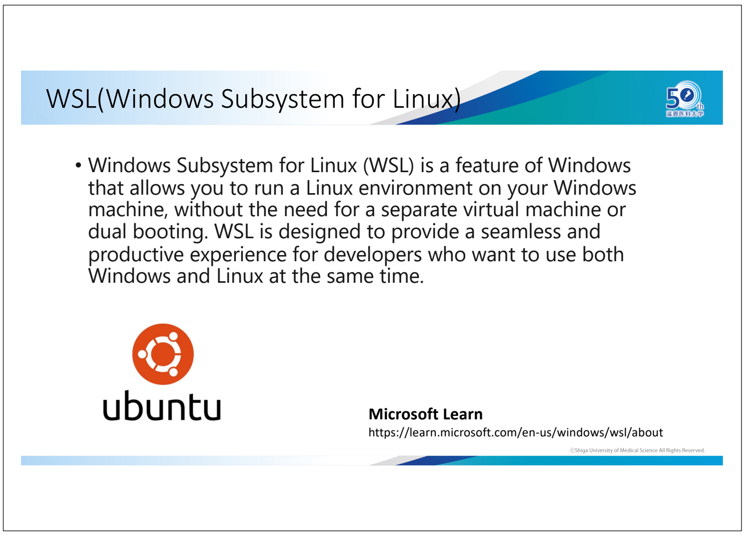
Expression Analysis

- quality check
- Mapping
- Expression Analysis
- visualization



Preparation of the environment required for RNAseq

- Installation of software required for analysis
  - Free UNIX-based software
    - Linux or Mac



WSL(Windows Subsystem for Linux)

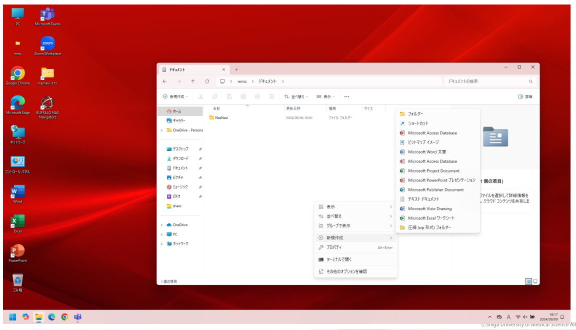
- Windows Subsystem for Linux (WSL) is a feature of Windows that allows you to run a Linux environment on your Windows machine, without the need for a separate virtual machine or dual booting. WSL is designed to provide a seamless and productive experience for developers who want to use both Windows and Linux at the same time.

Microsoft Learn
https://learn.microsoft.com/en-us/windows/wsl/about

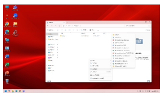
CUI (Character User Interface)

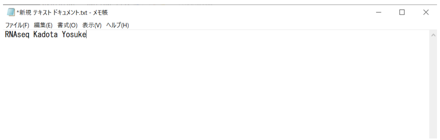
- A method of operating a computer through character input using a keyboard, which allows the user to interact with the computer in a character-based manner.

GUI（Graphical User Interface）



GUIとは





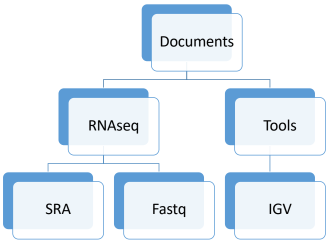
*新規 テキスト ドキュメント.txt - メモ帳

```
RNAseq Kadota Yosuke
```

Newtext.txt

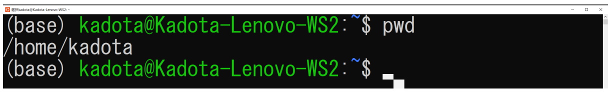
Let’s operate by CUI

- Directory (folder) operations
  - cd : Moving Directories
  - pwd : Show current directory
  - ls : Listing of folders and files
  - mkdir : create folder



Pwd : Show current directory

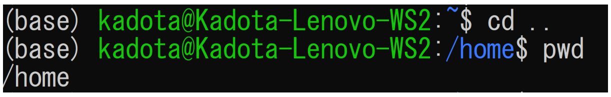
$ pwd



```
(base) kadota@Kadota-Lenovo-WS2:~$ pwd
/home/kadota
(base) kadota@Kadota-Lenovo-WS2:~$
```

cd

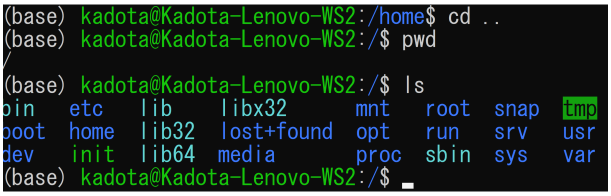
$ cd ..



```
(base) kadota@Kadota-Lenovo-WS2:~$ cd ..
(base) kadota@Kadota-Lenovo-WS2:/home$ pwd
/home
```

Ls : Listing of folders and files

$ ls

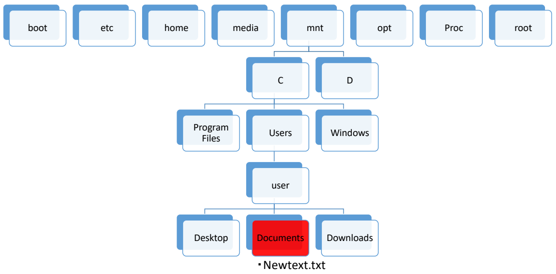


```
(base) kadota@Kadota-Lenovo-WS2:/home$ cd ..
(base) kadota@Kadota-Lenovo-WS2:/$ pwd
/
(base) kadota@Kadota-Lenovo-WS2:/$ ls
bin   etc   lib    libx32      mnt   root  snap  tmp
boot  home  lib32  lost+found  opt   run   srv   usr
dev   init  lib64  media       proc  sbin  sys   var
(base) kadota@Kadota-Lenovo-WS2:/$ _
```

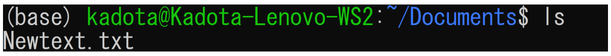
$ ls -l

$ ll

Let’s operate by CUI



ls



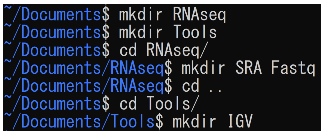
```
(base) kadota@Kadota-Lenovo-WS2:~/Documents$ ls
Newtext.txt
```

Cat : View all of the text files

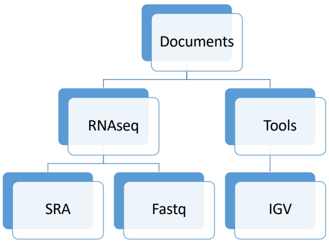
$ cat Newtext.txt

mkdir : create folder

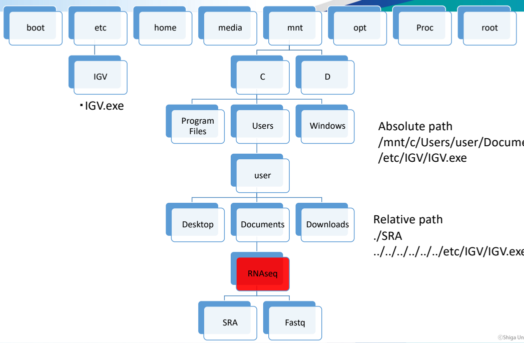
$ mkdir RNAseq



```
~/Documents$ mkdir RNAseq
~/Documents$ mkdir Tools
~/Documents$ cd RNAseq/
~/Documents/RNAseq$ mkdir SRA Fastq
~/Documents/RNAseq$ cd ..
~/Documents$ cd Tools/
~/Documents/Tools$ mkdir IGV
```



absolute path and relative path



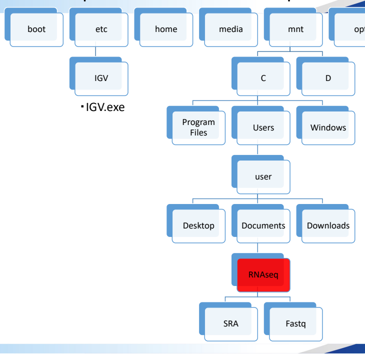
Absolute path
/mnt/c/Users/user/Documents/RNAseq/SRA
/etc/IGV/IGV.exe

Relative path
./SRA
../../../../../../etc/IGV/IGV.exe

CUIで操作してみよう

- ファイルの操作
  - mv : Moving files and folders
  - cp : Copying files and folders
  - cat : View all of the text files
  - less : View some of the text files

absolute path and relative path



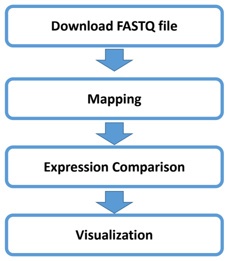
Absolute path
/mnt/c/Users/user/Documents/RNAseq/SRA
/etc/IGV/IGV.exe

Relative path
./SRA
../../../../../../etc/IGV/IGV.exe

less : View some of the text files

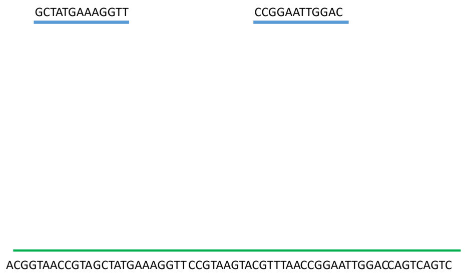
```
$ less ../Newtext.txt
```

RNAseq flowchart



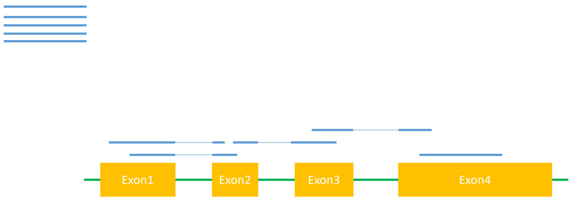
	Analysis Software	File extension to be used
Download FASTQ file	**SRA toolkit**	**sra, dra, fastq**
Mapping	**HISAT2、STAR**	**sam, bam**
Expression Comparison	**FeatureCount**	**txt, xlsx**
Visualization	**IGV**	**bam.bai**

mapping


GCTATGAAAGGTT
CCGGAATTGGAC
ACGGTAACCGTAGCTATGAAAGGTT CCGTAAGTACGTTTAACCGGAATTGGACCAGTCAGTC

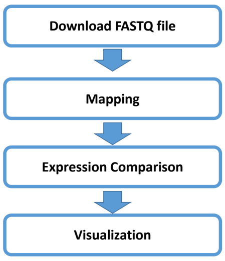
Genome sequence

mapping



Genome sequence from ensembl

RNAseq flowchart



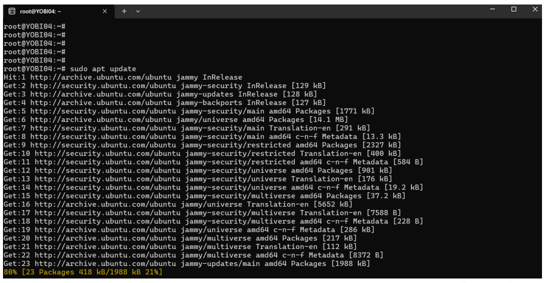
	Analysis Software	File extension to be used
Download FASTQ file	SRA toolkit	sra, dra, fastq
Mapping	HISAT2、STAR	sam, bam
Expression Comparison	FeatureCount	txt, xlsx
Visualization	IGV	bam.bai

(filename) extension

- SRA
  - Archive of FASTQ files
- FASTQ
  - Sequence data (text format)
- SAM
  - Mapping data (text format)
- BAM
  - Mapping data (binary format)
- GZ
  - compressed file

Updated package list

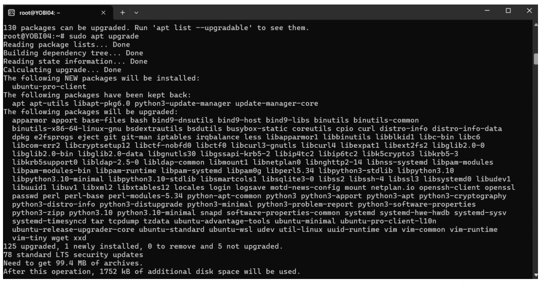
$ sudo apt update



```
root@YOBI04:~#
root@YOBI04:~#
root@YOBI04:~#
root@YOBI04:~#
root@YOBI04:~#
root@YOBI04:~# sudo apt update
Hit:1 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:2 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:5 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [1771 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy/universe amd64 Packages [14.1 MB]
Get:7 http://security.ubuntu.com/ubuntu jammy-security/main Translation-en [291 kB]
Get:8 http://security.ubuntu.com/ubuntu jammy-security/main amd64 c-n-f Metadata [13.3 kB]
Get:9 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [2327 kB]
Get:10 http://security.ubuntu.com/ubuntu jammy-security/restricted Translation-en [400 kB]
Get:11 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 c-n-f Metadata [584 B]
Get:12 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [901 kB]
Get:13 http://security.ubuntu.com/ubuntu jammy-security/universe Translation-en [176 kB]
Get:14 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 c-n-f Metadata [19.2 kB]
Get:15 http://security.ubuntu.com/ubuntu jammy-security/multiverse amd64 Packages [37.2 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy/universe Translation-en [5652 kB]
Get:17 http://security.ubuntu.com/ubuntu jammy-security/multiverse Translation-en [7588 B]
Get:18 http://security.ubuntu.com/ubuntu jammy-security/multiverse amd64 c-n-f Metadata [228 B]
Get:19 http://archive.ubuntu.com/ubuntu jammy/universe amd64 c-n-f Metadata [286 kB]
Get:20 http://archive.ubuntu.com/ubuntu jammy/multiverse amd64 Packages [217 kB]
Get:21 http://archive.ubuntu.com/ubuntu jammy/multiverse Translation-en [112 kB]
Get:22 http://archive.ubuntu.com/ubuntu jammy/multiverse amd64 c-n-f Metadata [8372 B]
Get:23 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [1988 kB]
80% [23 Packages 418 kB/1988 kB 21%]
```

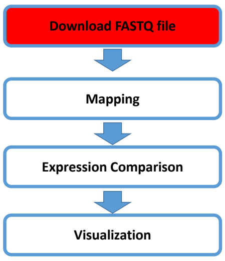
Upgrade Package

$ sudo apt upgrade



```
130 packages can be upgraded. Run 'apt list --upgradable' to see them.
root@YOBI04:~# sudo apt upgrade
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Calculating upgrade... Done
The following NEW packages will be installed:
  ubuntu-pro-client
The following packages have been kept back:
  apt apt-utils libapt-pkg6.0 python3-update-manager update-manager-core
The following packages will be upgraded:
  apparmor apport base-files bash bind9-dnsutils bind9-host bind9-libs binutils binutils-common
  binutils-x86-64-linux-gnu bsdextrautils bsdutils busybox-static coreutils cpio curl distro-info distro-info-data
  dpkg e2fsprogs eject git git-man iptables irqbalance less libapparmor1 libbinutils libblkid1 libc-bin libc6
  libcom-err2 libcryptsetup12 libctf-nobfd0 libctf0 libcurl3-gnutls libcurl4 libexpat1 libext2fs2 libglib2.0-0
  libglib2.0-bin libglib2.0-data libgnutls30 libgssapi-krb5-2 libip4tc2 libip6tc2 libk5crypto3 libkrb5-3
  libkrb5support0 libldap-2.5-0 libldap-common libmount1 libnetplan0 libnghttp2-14 libnss-systemd libpam-modules
  libpam-modules-bin libpam-runtime libpam-systemd libpam0g libperl5.34 libpython3-stdlib libpython3.10
  libpython3.10-minimal libpython3.10-stdlib libsmartcols1 libsqlite3-0 libss2 libssh-4 libssl3 libsystemd0 libudev1
  libuuid1 libuv1 libxml2 libxtables12 locales login logsave motd-news-config mount netplan.io openssh-client openssl
  passwd perl perl-base perl-modules-5.34 python-apt-common python3 python3-apport python3-apt python3-cryptography
  python3-distro-info python3-distupgrade python3-minimal python3-problem-report python3-software-properties
  python3-zipp python3.10 python3.10-minimal snapd software-properties-common systemd systemd-hwe-hwdb systemd-sysv
  systemd-timesyncd tar tcpdump tzdata ubuntu-advantage-tools ubuntu-minimal ubuntu-pro-client-l10n
  ubuntu-release-upgrader-core ubuntu-standard ubuntu-wsl udev util-linux uuid-runtime vim vim-common vim-runtime
  vim-tiny wget xxd
125 upgraded, 1 newly installed, 0 to remove and 5 not upgraded.
78 standard LTS security updates
Need to get 99.4 MB of archives.
After this operation, 1752 kB of additional disk space will be used.
```

RNAseq flowchart



	Analysis Software	File extension to be used
Download FASTQ file	SRA toolkit	sra, dra, fastq
Mapping	HISAT2、STAR	sam, bam
Expression Comparison	FeatureCount	txt, xlsx
Visualization	IGV	bam.bai

International Nucleotide Sequence Database

- **INSDC(International Nucleotide Sequence Database Collaboration)**

- INSDC is a global cooperative of independent government agencies or non-profit organizations that manage sequence databases, collect sequence information and annotations, preserve them in the scientific record, and create a comprehensive collection of such data that can be widely shared.

International Nucleotide Sequence Database Collaboration

- The National Library of Medicine(NLM)
- National Center for Biotechnology Information(NCBI)
- NCBI Sequence Read Archive(SRA)

- The European Molecular Biology Laboratory(EMBL)
- European Bioinformatics Institute(EBI)
- EBI Sequence Read Archive(ERA)

- The Research Organization of Information and System(大学共同利用機関法人 情報・システム研究機構)
- National Institute of Genetics(日本国立遺伝学研究所)
- DDBJ Sequence Read Archive(DRA)

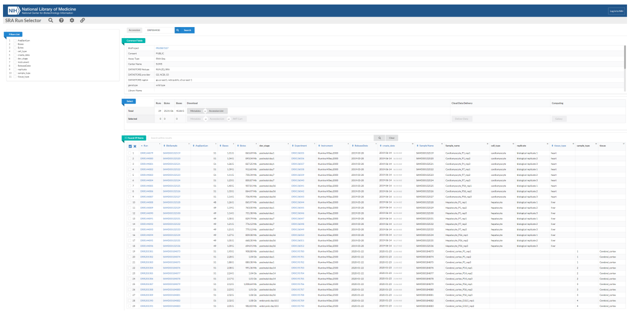
International Nucleotide Sequence Database

- All sequence data registered in DDBJ are issued with DDBJ accession numbers. The accession number is assigned to each nucleotide sequence data registered in the database and is unique to that sequence data. Data registered in DDBJ are sent to NCBI and EBI at the time of publication, and are common worldwide. More than 99% of all registrations from Japan are made through DDBJ.

Find an accession number

- Pubmed
  - Srsf7 Establishes the Juvenile Transcriptome through Age-Dependent Alternative Splicing in Mice

  - DRA009510
  - DRA009537
  - DRA009538



Download Fastq file

- SRA-Toolkit

Download archive file

```
$ prefetch [accession number]
```

Convert archive files to Fastq format

```
$ fastq-dump [archive file]
```

gzファイル

To compress

```
$ gzip [input file]
```

To decompress

```
$ gzip –d [input file]
```

FASTQの中身を確認

```
$ less L-1_P1.fastq.gz

@MG00HS09:723:C9A15ACXX:7:1101:1483:1919 1:N:0:CGATGT
NGACCCGCTGAATTTAAGCATATTAGTCAGCGGAGGAAAAGAAACTAACCA
+
#11B?D@8DAF?DGGECFHDF?4ACFB?GEGGB6)?69DE;FGE=4@F)=C
```

- 1：@Sequence ID and additional information
- 2：Nucleotide sequence
- 3：+ Sequence ID and additional information
- 4：sequence quality

ASC2

33!
34"
35#
36$
37%
38&
39'
40(
41)
42*
43+
44,

45-
46.
47/
48 0
49 1
50 2
51 3
52 4
53 5
54 6
55 7
56 8

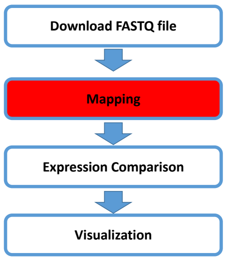
57 9
58:
59;
60<
61=
62>
63?
64@
65A
66B
67C
68D

69E
70F
71G
72H
73I
74J
75K
76L
77M
78N
79O
80P

81Q
82R
83S
84T
85U
86V
87W
88X
89Y
90Z

ASCII文字	ASCIIコード	クオリティスコア	塩基が間違っている確率 (P)
!	33	0	1.0 (100%)
"	34	1	0.794 (79.4%)
#	35	2	0.631 (63.1%)
$	36	3	0.501 (50.1%)
%	37	4	0.398 (39.8%)
&	38	5	0.316 (31.6%)
'	39	6	0.251 (25.1%)
(	40	7	0.199 (19.9%)
)	41	8	0.158 (15.8%)
*	42	9	0.126 (12.6%)
+	43	10	0.100 (10%)
,	44	11	0.0794 (7.94%)
-	45	12	0.0631 (6.31%)
.	46	13	0.0501 (5.01%)
/	47	14	0.0398 (3.98%)
0	48	15	0.0316 (3.16%)
1	49	16	0.0251 (2.51%)
2	50	17	0.0199 (1.99%)
3	51	18	0.0158 (1.58%)
4	52	19	0.0126 (1.26%)
5	53	20	0.0100 (1.0%)
6	54	21	0.00794 (0.794%)
7	55	22	0.00631 (0.631%)
8	56	23	0.00501 (0.501%)
9	57	24	0.00398 (0.398%)
:	58	25	0.00316 (0.316%)
;	59	26	0.00251 (0.251%)
<	60	27	0.00199 (0.199%)
=	61	28	0.00158 (0.158%)
>	62	29	0.00126 (0.126%)
?	63	30	0.00100 (0.1%)
@	64	31	0.000794 (0.0794%)
A	65	32	0.000631 (0.0631%)
B	66	33	0.000501 (0.0501%)
C	67	34	0.000398 (0.0398%)
D	68	35	0.000316 (0.0316%)
E	69	36	0.000251 (0.0251%)
F	70	37	0.000199 (0.0199%)
G	71	38	0.000158 (0.0158%)
H	72	39	0.000126 (0.0126%)
I	73	40	0.000100 (0.01%)

RNAseq flowchart



	Analysis Software	File extension to be used
Download FASTQ file	**SRA toolkit**	**sra, dra, fastq**
Mapping	**HISAT2、STAR**	**sam, bam**
Expression Comparison	**FeatureCount**	**txt, xlsx**
Visualization	**IGV**	**bam.bai**

Mapping

- HISAT2
- hisat2 -p [number of CPU] -x [genome index] -U [input file] -S [output file]

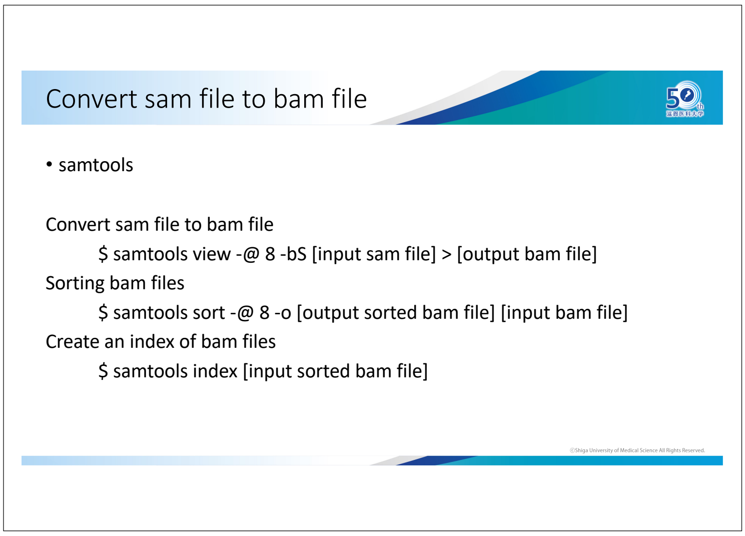
- Download Index file of the genome
  - http://daehwankimlab.github.io/hisat2/download/#m-musculus

- Mapping by using hisat2. Make sam file.

```
$ hisat2 -p 8 -x HISATindex/genome -U fastq.gz/L-1_P1.fastq.gz -S sam/L1.sam
```

SAMファイル

- QNAME: Query name. Read identifier (usually the read ID).
- FLAG: Bit flag indicating alignment information. It has binary information such as read mapping status and pair information.
- RNAME: Reference name. The name of the reference sequence to which the read is aligned (e.g., chromosome name).
- POS: Leftmost position of 1-base. The starting position of the read mapping.   MAPQ: Mapping quality. A score indicating the reliability of the alignment.
- CIGAR: CIGAR string. Indicates the alignment pattern of the reads (e.g., '76M' means that 76 bases were matched).
- RNEXT: The reference name to which the next read is aligned. In the case of paired-end sequencing, the reference sequence to which the other read is aligned.
- PNEXT: The leftmost position of one base of the next read. The starting position of the partner of the paired-end read.
- TLEN: Template Length. Insertion size between paired-end reads.
- SEQ: Read Sequence. Actual nucleotide sequence.
- QUAL: Quality score. Read sequence quality expressed in ASCII characters.



Convert sam file to bam file

- samtools

Convert sam file to bam file

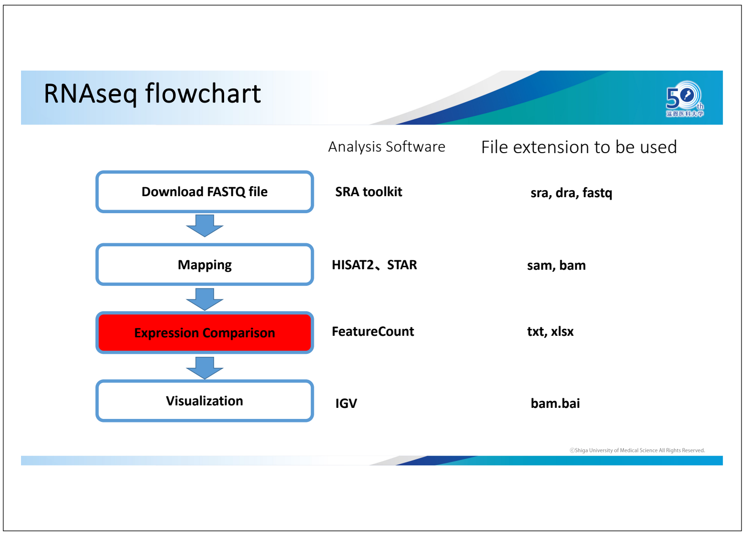
```
$ samtools view -@ 8 -bS [input sam file] > [output bam file]
```

Sorting bam files

```
$ samtools sort -@ 8 -o [output sorted bam file] [input bam file]
```

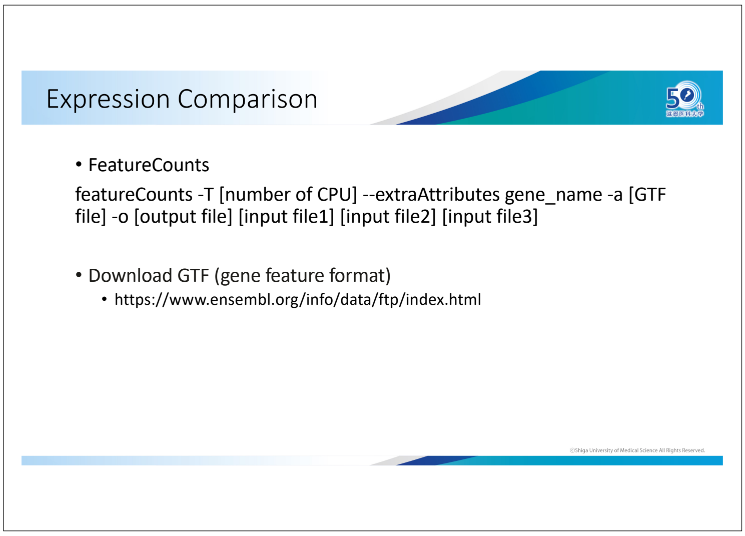
Create an index of bam files

```
$ samtools index [input sorted bam file]
```



RNAseq flowchart

	Analysis Software	File extension to be used
Download FASTQ file	SRA toolkit	sra, dra, fastq
Mapping	HISAT2、STAR	sam, bam
Expression Comparison	FeatureCount	txt, xlsx
Visualization	IGV	bam.bai

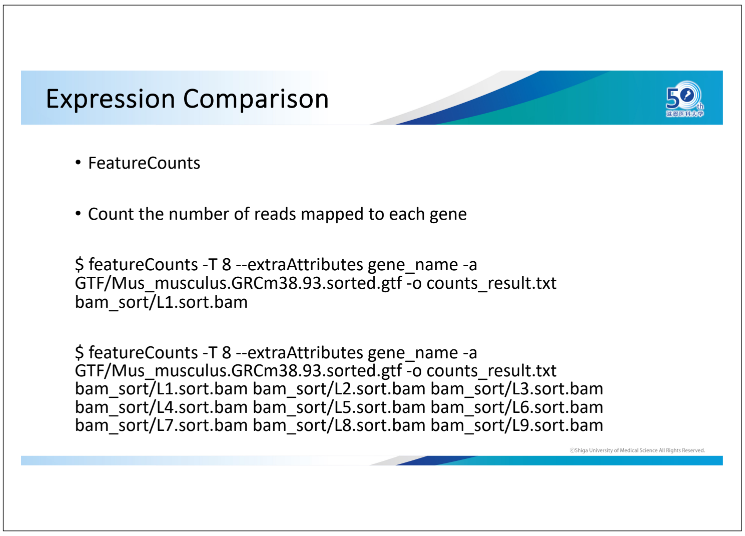


Expression Comparison

- FeatureCounts

```
featureCounts -T [number of CPU] --extraAttributes gene_name -a [GTF file] -o [output file] [input file1] [input file2] [input file3]
```

- Download GTF (gene feature format)
  - https://www.ensembl.org/info/data/ftp/index.html



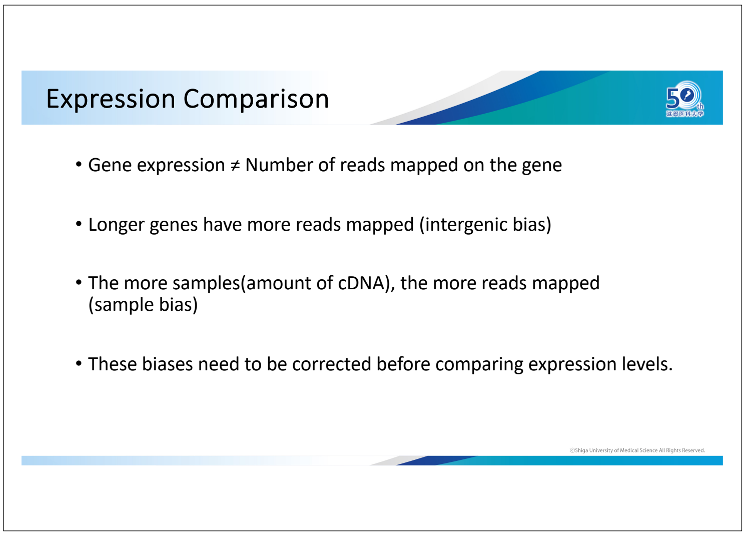
Expression Comparison

- FeatureCounts

- Count the number of reads mapped to each gene

```
$ featureCounts -T 8 --extraAttributes gene_name -a
GTF/Mus_musculus.GRCm38.93.sorted.gtf -o counts_result.txt
bam_sort/L1.sort.bam
```

```
$ featureCounts -T 8 --extraAttributes gene_name -a
GTF/Mus_musculus.GRCm38.93.sorted.gtf -o counts_result.txt
bam_sort/L1.sort.bam bam_sort/L2.sort.bam bam_sort/L3.sort.bam
bam_sort/L4.sort.bam bam_sort/L5.sort.bam bam_sort/L6.sort.bam
bam_sort/L7.sort.bam bam_sort/L8.sort.bam bam_sort/L9.sort.bam
```



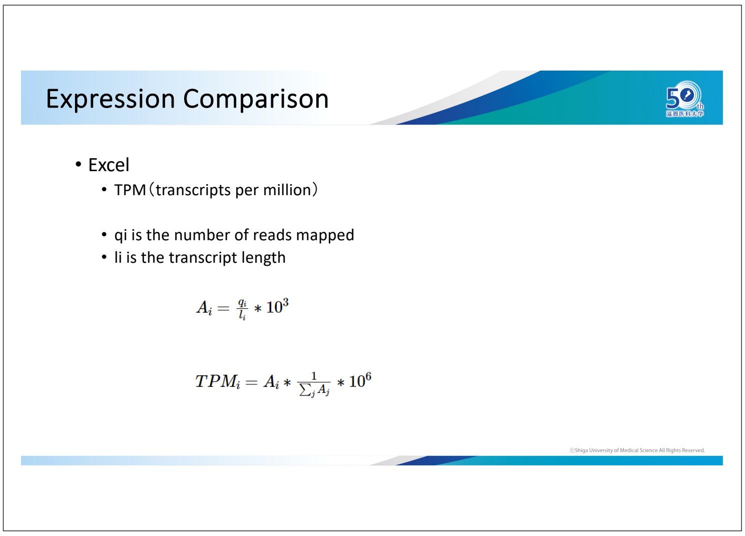
Expression Comparison

- Gene expression ≠ Number of reads mapped on the gene

- Longer genes have more reads mapped (intergenic bias)

- The more samples(amount of cDNA), the more reads mapped (sample bias)

- These biases need to be corrected before comparing expression levels.



Expression Comparison

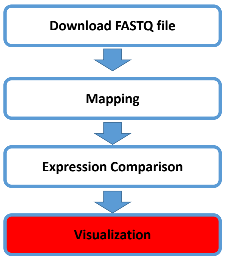
- Excel
  - TPM（transcripts per million）

  - qi is the number of reads mapped
  - li is the transcript length

$$A_i = \frac{q_i}{l_i} * 10^3$$

$$TPM_i = A_i * \frac{1}{\sum_j A_j} * 10^6$$

RNAseq flowchart

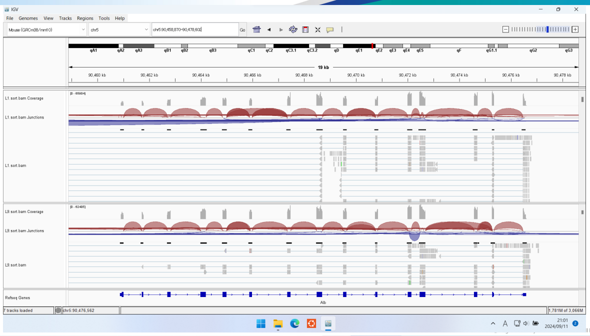


	Analysis Software	File extension to be used
Download FASTQ file	SRA toolkit	sra, dra, fastq
Mapping	HISAT2、STAR	sam, bam
Expression Comparison	FeatureCount	txt, xlsx
Visualization	IGV	bam.bai

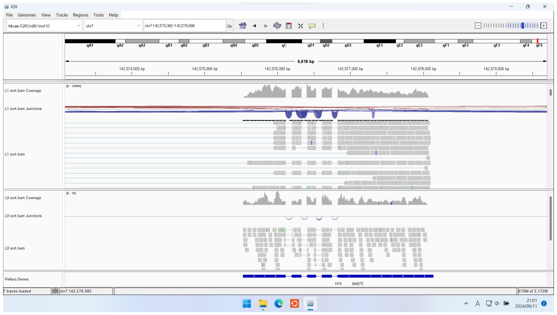
Visualization

- IGV（Integrative Genomics Viewer）

Visualization



Visualization



Visualization