KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (온라인)

Introduction to Next Generation Sequencing data analysis with Galaxy

이동성 _ 서울시립대

KSBI
KOREAN SOCIETY FOR BIOINFORMATICS
한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 온라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구 중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의개요

Introduction to Next Generation Sequencing data analysis with Galaxy

최근 생성되는 바이오정보 데이터의 크기는 점점 커지고 있지만, 저장 공간과 시간의 제약으로 이러한 빅데이터를 하나의 머신으로 처리하는데 많은 어려움이 따릅니다. 또한 다양한 데이터들이 매일같이 쏟아져 나오는 가운데 이러한 데이터들을 얻고 다루기 위해서는 그에 맞는 환경을 구축해야 하지만 이를 배우고 싶어하는 학생들이나 많은 연구자들이 비용적, 시간적, 환경적인 제약을 받고 있습니다.

이에 본 강의에서는 생명정보 데이터를 효과적이고 빠르게 처리하기 위해 널리 쓰이고 있는 web-base 플랫폼인 Galaxy를 소개하겠습니다. 데이터 가져오기, 도구 실행, history를 이용한 작업, workflow 생성 및 작업 공유와 같은 기본 작업을 수행하는 방법을 설명하며 이를 통해 빅데이터를 빠르고 손쉽게 처리할 수 있는 기법을 배우고, 이를 실제 바이오 데이터에 적용하여 효율적이고 효과적인 분석을 할 수 있는 핵심 역량을 갖추는 것을 목표로 합니다.

강의는 다음의 내용을 포함한다:

- Galaxy 개요
- Public 데이터 가져오기
- 데이터 분석하기

* 참고강의교재: Galaxy (https://usegalaxy.org.au/)

* 교육생준비물: 노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

* 강의 난이도: 초급

* 강의: 이동성 교수 (서울시립대학교 생명과학과)

Curriculum Vitae

Speaker Name: Hong-Gil Dong, Ph.D.

▶ **Personal Info**

Name	Dongsung Lee
Title	Assistant Professor
Affiliation	Department of Life Science, University of Seoul

▶ **Contact Information**

Email	dslee@uos.ac.kr
Phone Number	(02) 6490-2676

Research Interest

Translational bioinformatics, Machine learning and computational genomics

Educational Experience

2010	B.S. in Life Science, Korea University, Korea
2015	Ph.D. in Medical Science, Seoul National University, Korea

Professional Experience

2001-2007	Assistant Professor, Department of Life Science, University of Seoul, Korea
2016-2020	Post-doc research fellow, Salk Institute for Biological Studies, USA

Selected Publications (5 maximum)

1. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nature Methods. (2019)
2. A noncanonical BRD9-containing BAF chromatin remodeling complex regulates naive pluripotency in mouse embryonic stem cells. Nat Commun. (2018)
3. An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator. Nature Commun. (2014).
4. Genome-wide characterization of the routes to induced pluripotency. Nature (2014)
5. Divergent reprogramming routes lead to alternative stem-cell states. Nature (2014)

KSBi 한국생명정보학회

KSBi-BIML

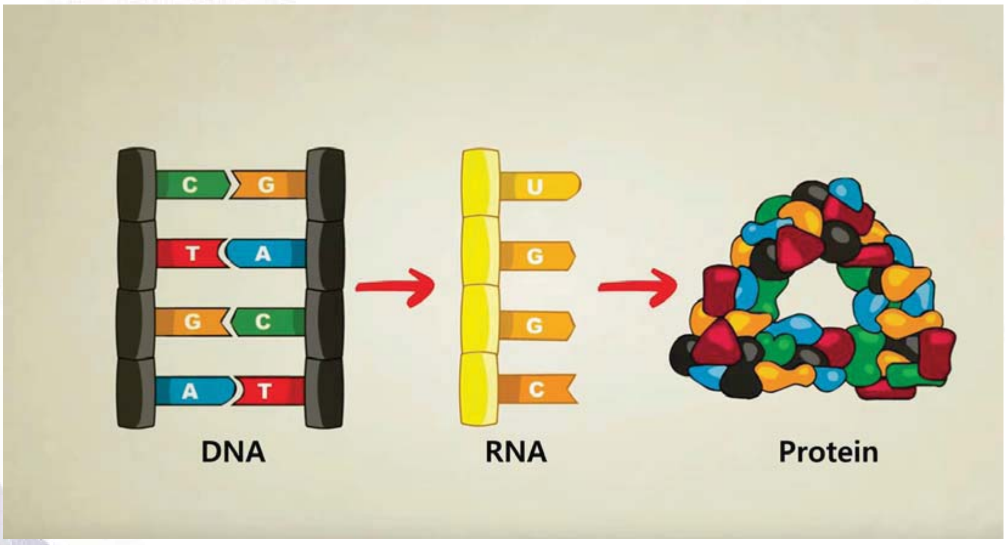
Introduction to Next Generation Sequencing data analysis with Galaxy

서울시립대학교
생명과학과
이동성

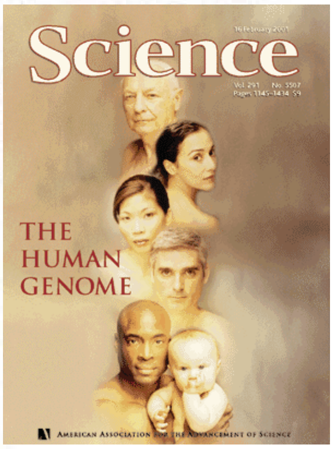
DNA was discovered by
JOHANN FRIEDRICH MIESCHER
in the year 1869.

2

Central Dogma, 생명현상의 중심 원리



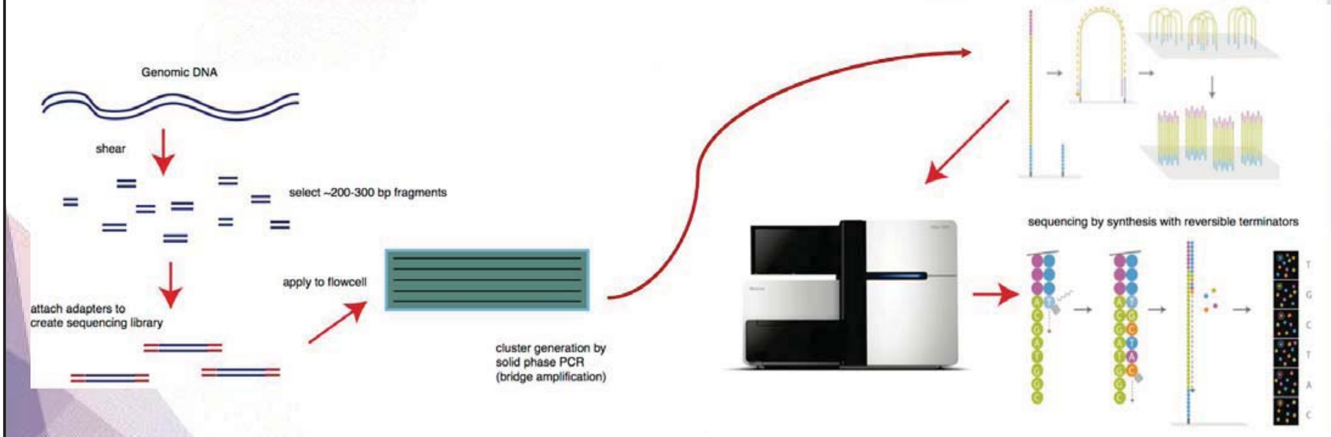
In February 2001 the 'First Draft' of the Human Genome is Published



Venter et al., Celera, Science, 2001

International Human Genome Sequencing Consortium, Lander et al., Nature, 2001

Next Generation Sequencing (NGS)

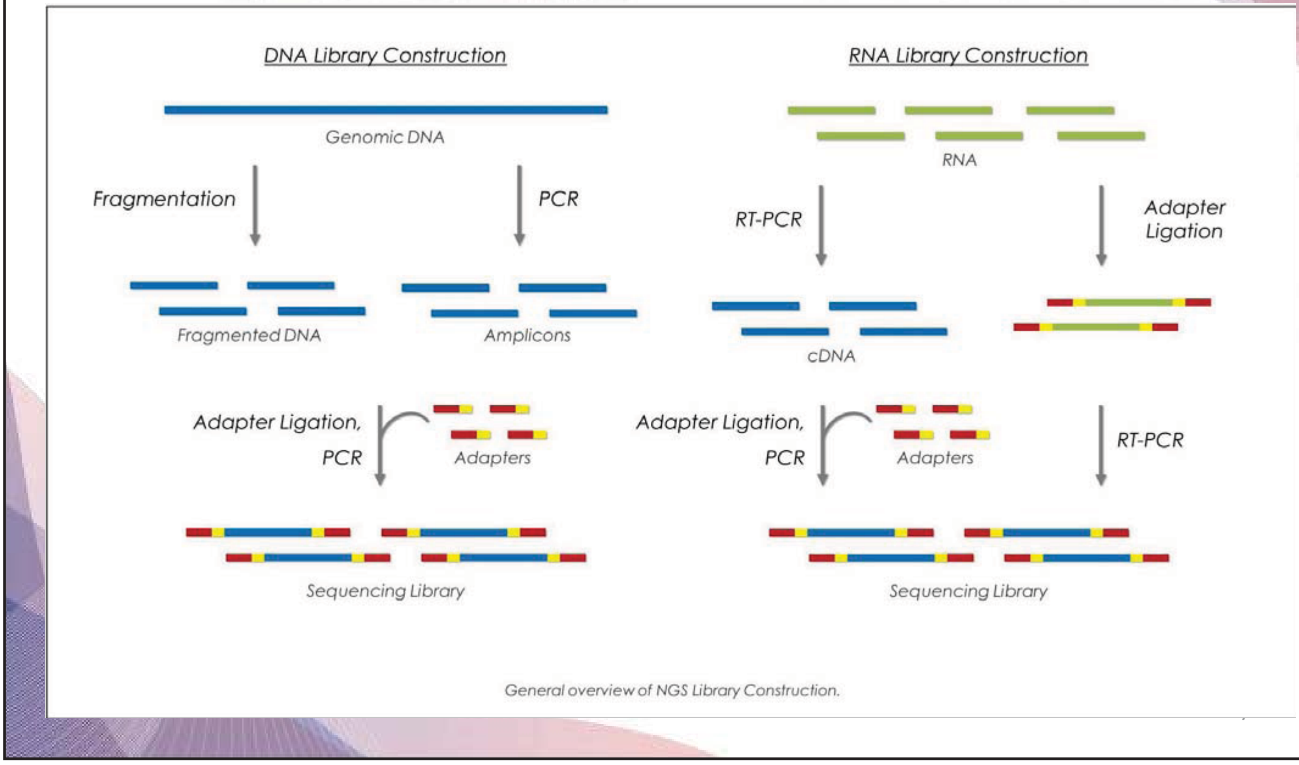

Genomic DNA
shear
select ~200-300 bp fragments
attach adapters to create sequencing library
apply to flowcell
cluster generation by solid phase PCR (bridge amplification)
sequencing by synthesis with reversible terminators

NGS Applications

RNA	RNA seq
DNA	Whole Genome Seq
	Target Seq
	Bisulfite Seq
	ChIP-seq
	ATAC-seq
	Hi-C seq

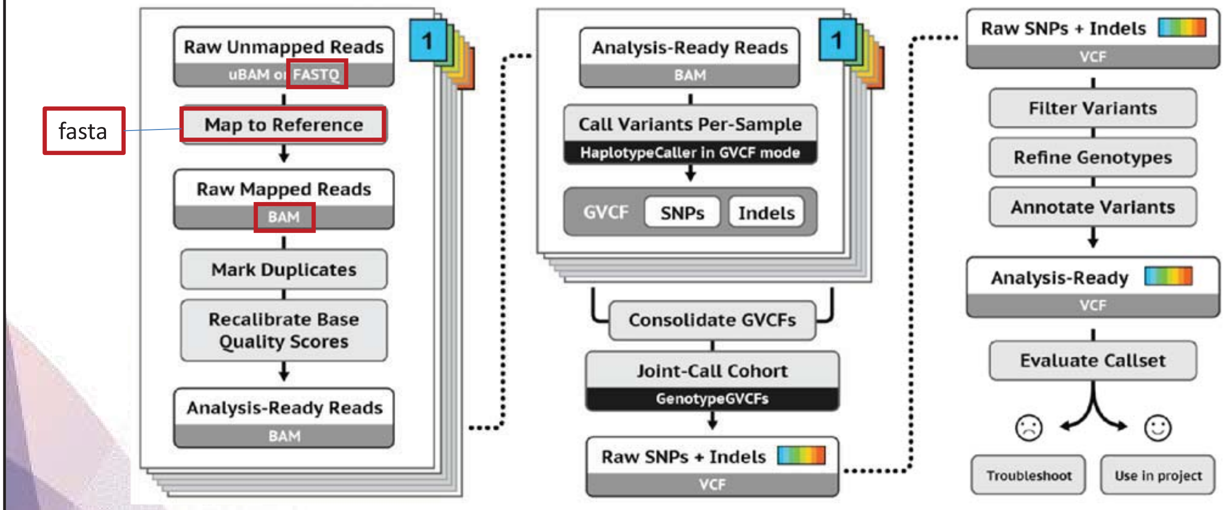
Single cell / Multi-omic

Sequencing library preparation



General overview of NGS Library Construction.

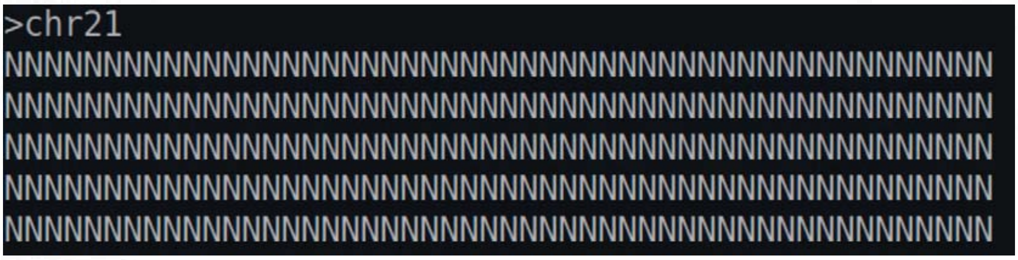
Genome Sequencing data analysis pipeline



GATK Best Practice

NGS 데이터의 구조와 형식 -FASTA

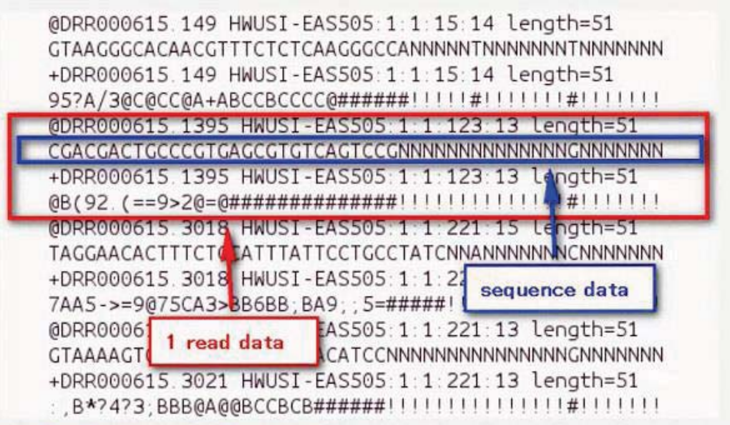
- 보통 Reference genome을 넣는 파일
- sequence를 표현하는 가장 기본적인 포맷.
- DNA sequence 뿐만 아니라 Protein sequence도 저장한다.
- '>'로 주석을 표현하고 한줄당 50개의 sequence가 표현된다.



```
>chr21
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

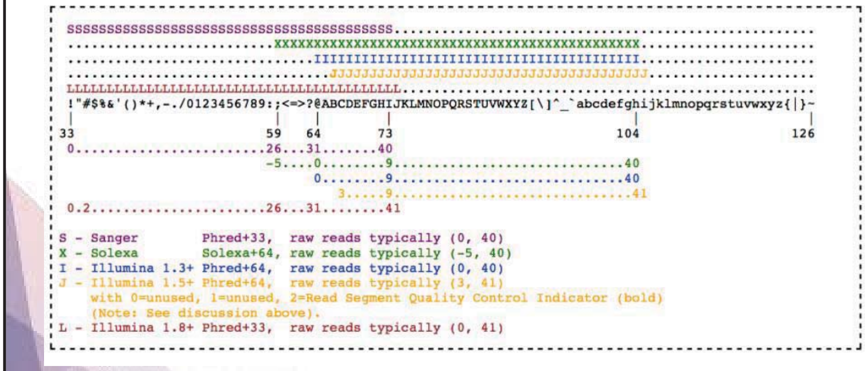
NGS 데이터의 구조와 형식 -FASTQ

- FASTA이외에 실험을 통해서 얻는 서열 정보형식, 4줄로 이루어져 있다.
- FASTA 포맷의 각 염기에 QV(quality Value)가 추가된 형태이다.
- 첫행은 ID로 보통 기계적인 서열의 주석정보(@장비번호 : lane 번호 : x좌표 : y좌표 : multiplexing indexing: paired-end)를 담는다.
- 두번째행은 sequence, N은 모르는 자리
- 세번째행은 +, 부가적인 정보가 있다.
- 네번째행은 sequencing quality

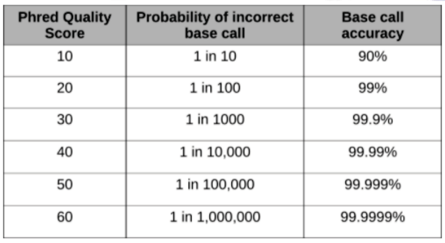


```
@DRR000615.149 HWUSI-EAS505:1:1:15:14 length=51
GTAAGGGCACAACGTTTCTCTCAAGGGCCANNNNNTNNNNNNNTNNNNNNN
+DRR000615.149 HWUSI-EAS505:1:1:15:14 length=51
95?A/3@C@CC@A+ABCCBCCCC@######!!!!!#!!!!!!!#!!!!!!!
@DRR000615.1395 HWUSI-EAS505:1:1:123:13 length=51
CGACGACTGCCCGTGAGCGTGTCAGTCCGNNNNNNNNNNNNNNNGNNNNNNN
+DRR000615.1395 HWUSI-EAS505:1:1:123:13 length=51
@B(92.(==9>2@=@###############!!!!!!!!!!!!#!!!!!!!
@DRR000615.3018 HWUSI-EAS505:1:1:221:15 length=51
TAGGAACACTTTCTC ATTTATTCCTGCCTATCNNANNNNNNNCNNNNNNN
+DRR000615.3018 HWUSI-EAS505:1:1:22
7AA5->=9@75CA3>BB6BB;BA9;;5=#####!
@DRR0006               EAS505:1:1:221:13 length=51
GTAAAAGT               CATCCNNNNNNNNNNNNNNNNGNNNNNNN
+DRR000615.3021 HWUSI-EAS505:1:1:221:13 length=51
:,B*?4?3;BBB@A@@BCCBCB######!!!!!!!!!!!!!!!#!!!!!!!
```

FASTQ에서의 Sequencing quality 표현 -Phred quality score 를 ascii로 표현



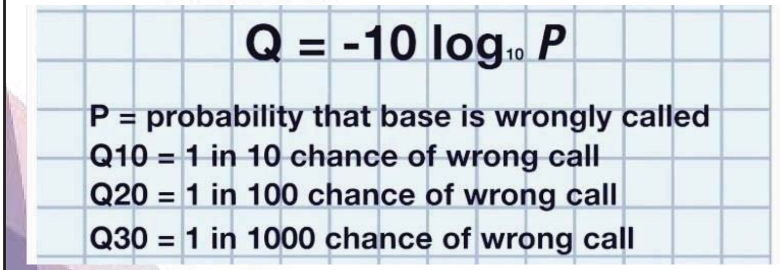
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0.........................26...31.......40
                         -5....0........9..............................40
                               0........9..............................40
                                  3.....9...............................41
0.2.......................26...31.......41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```



Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

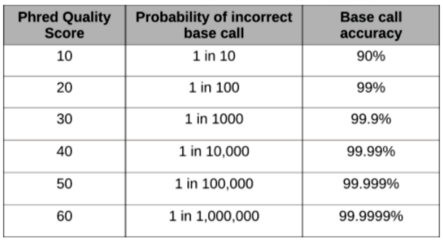
Phred Quality Score

- **Q-Score는 염기를 호출할 때 발생할 수 있는 오류 가능성에 대한 대수적인 수치**
- **Q = -10 log P**
- P = 염기를 잘 못 불러올 가능성(확율)
- Q10 = 염기 10개를 불러울 때 1개 염기를 잘 못 불러올 확률
- Q20 = 염기 100개를 불러울 때 1개 염기를 잘 못 불러올 확률
- Q30 = 염기 1000개를 불러울 때 1개 염기를 잘 못 불러올 확률

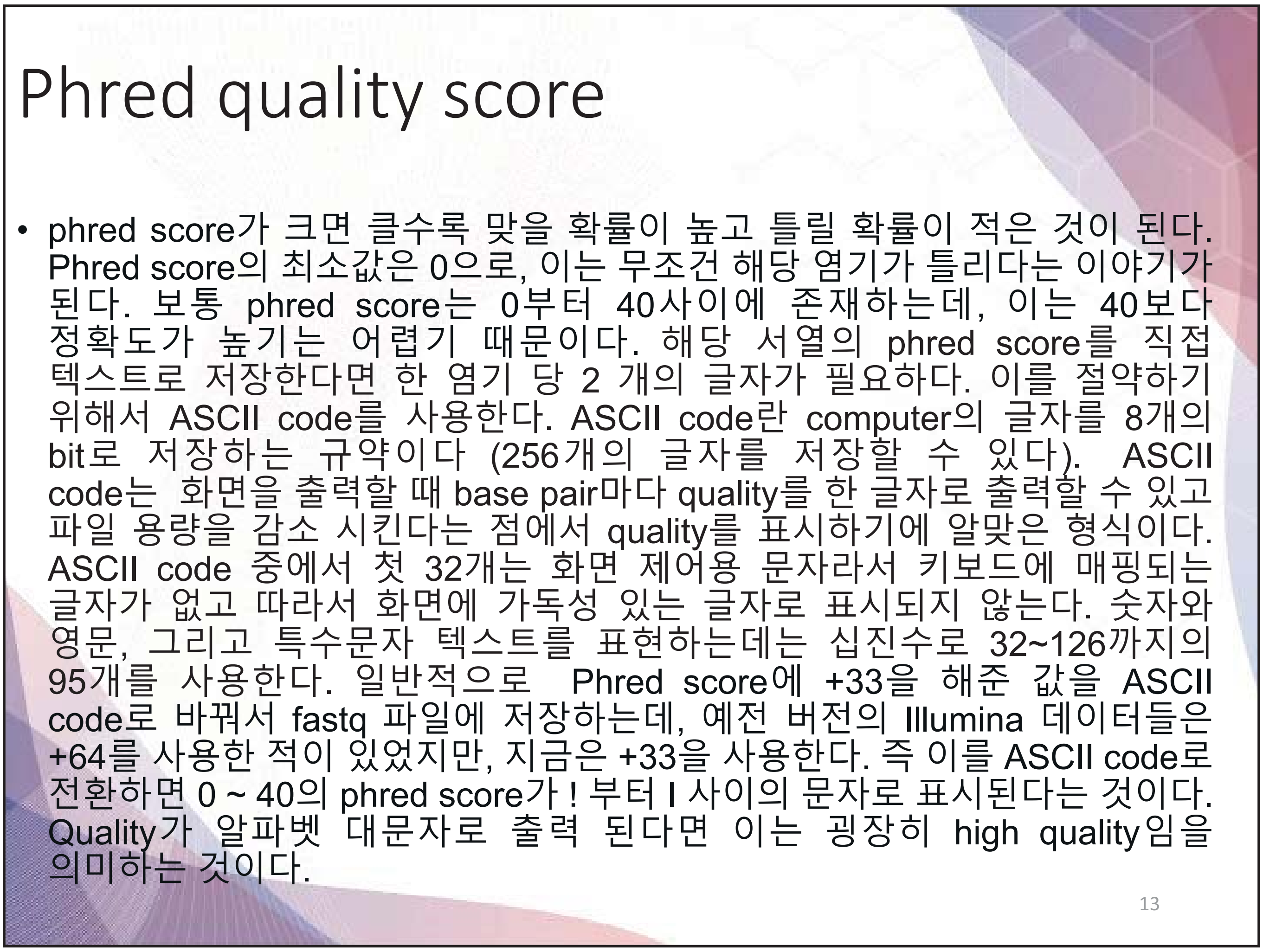


$$Q = -10 \log_{10} P$$

P = probability that base is wrongly called
Q10 = 1 in 10 chance of wrong call
Q20 = 1 in 100 chance of wrong call
Q30 = 1 in 1000 chance of wrong call

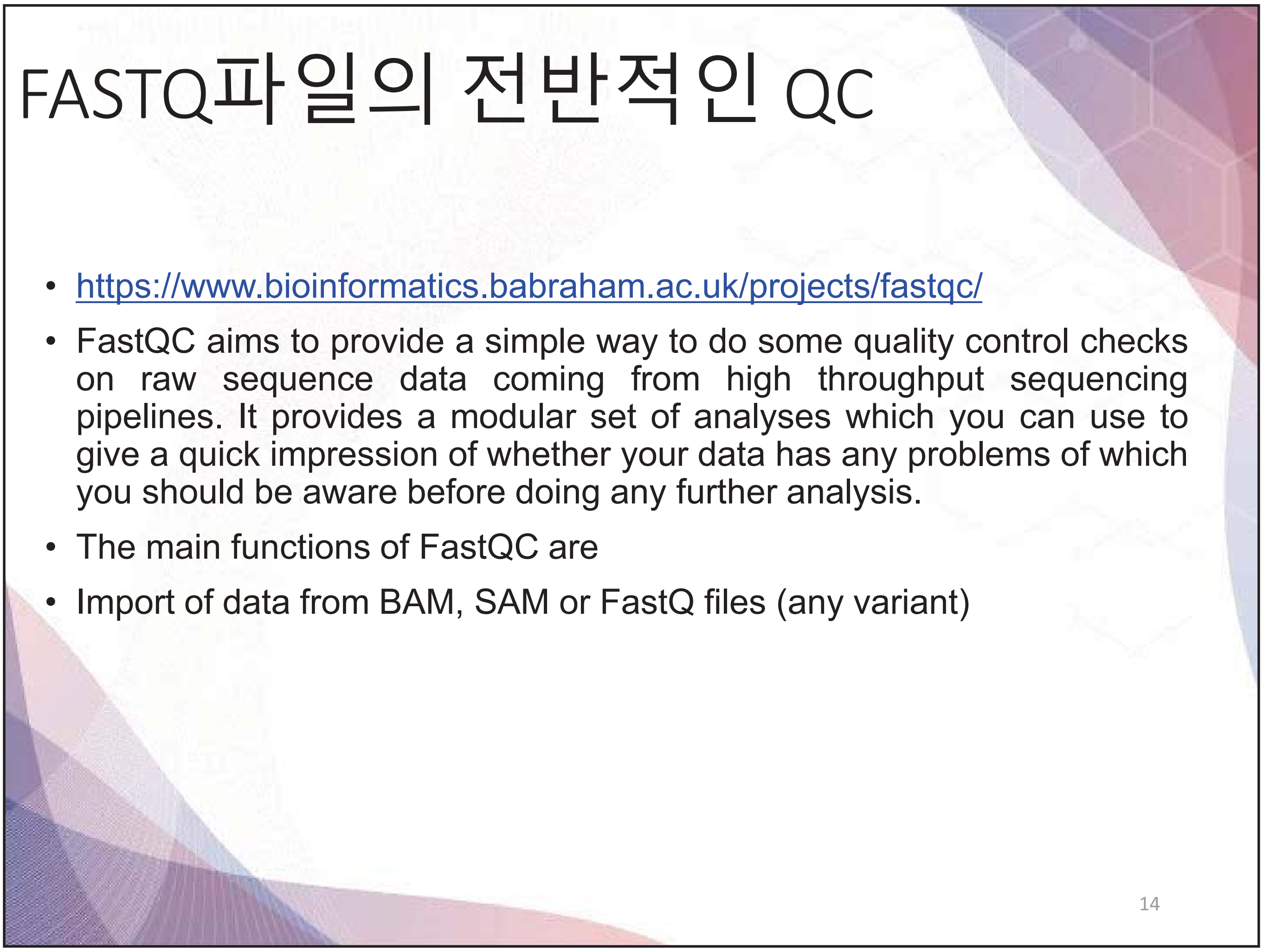


Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



Phred quality score

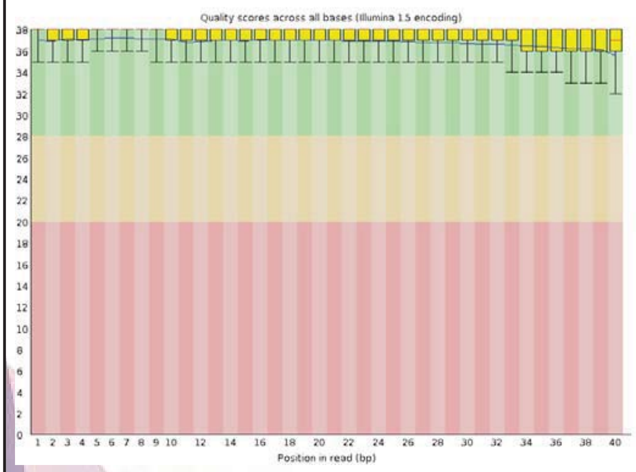
- phred score가 크면 클수록 맞을 확률이 높고 틀릴 확률이 적은 것이 된다. Phred score의 최소값은 0으로, 이는 무조건 해당 염기가 틀리다는 이야기가 된다. 보통 phred score는 0부터 40사이에 존재하는데, 이는 40보다 정확도가 높기는 어렵기 때문이다. 해당 서열의 phred score를 직접 텍스트로 저장한다면 한 염기 당 2 개의 글자가 필요하다. 이를 절약하기 위해서 ASCII code를 사용한다. ASCII code란 computer의 글자를 8개의 bit로 저장하는 규약이다 (256개의 글자를 저장할 수 있다). ASCII code는 화면을 출력할 때 base pair마다 quality를 한 글자로 출력할 수 있고 파일 용량을 감소 시킨다는 점에서 quality를 표시하기에 알맞은 형식이다. ASCII code 중에서 첫 32개는 화면 제어용 문자라서 키보드에 매핑되는 글자가 없고 따라서 화면에 가독성 있는 글자로 표시되지 않는다. 숫자와 영문, 그리고 특수문자 텍스트를 표현하는데는 십진수로 32~126까지의 95개를 사용한다. 일반적으로 Phred score에 +33을 해준 값을 ASCII code로 바꿔서 fastq 파일에 저장하는데, 예전 버전의 Illumina 데이터들은 +64를 사용한 적이 있었지만, 지금은 +33을 사용한다. 즉 이를 ASCII code로 전환하면 0 ~ 40의 phred score가 ! 부터 I 사이의 문자로 표시된다는 것이다. Quality가 알파벳 대문자로 출력 된다면 이는 굉장히 high quality임을 의미하는 것이다.



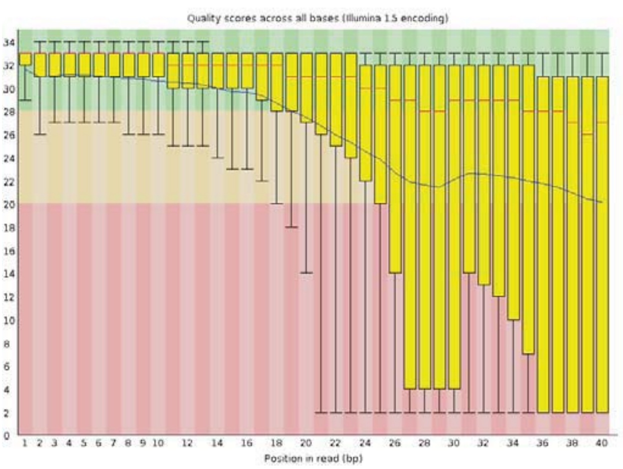
FASTQ파일의 전반적인 QC

- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.
- The main functions of FastQC are
- Import of data from BAM, SAM or FastQ files (any variant)

FASTQC -Per base sequence quality

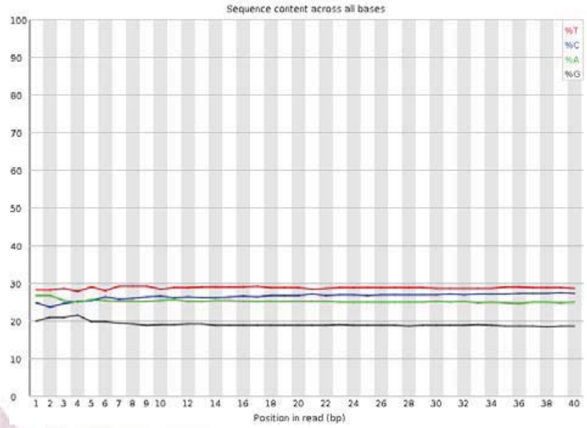


Good

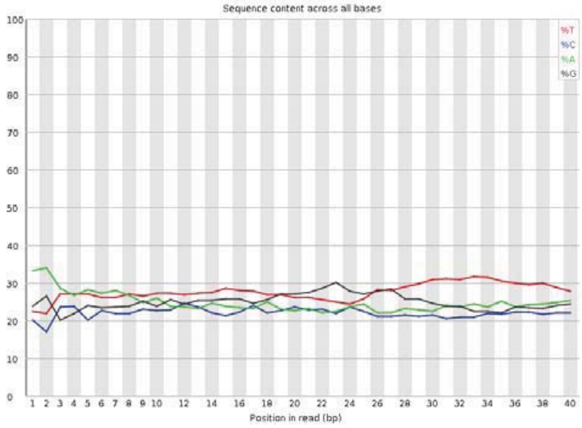


Bad

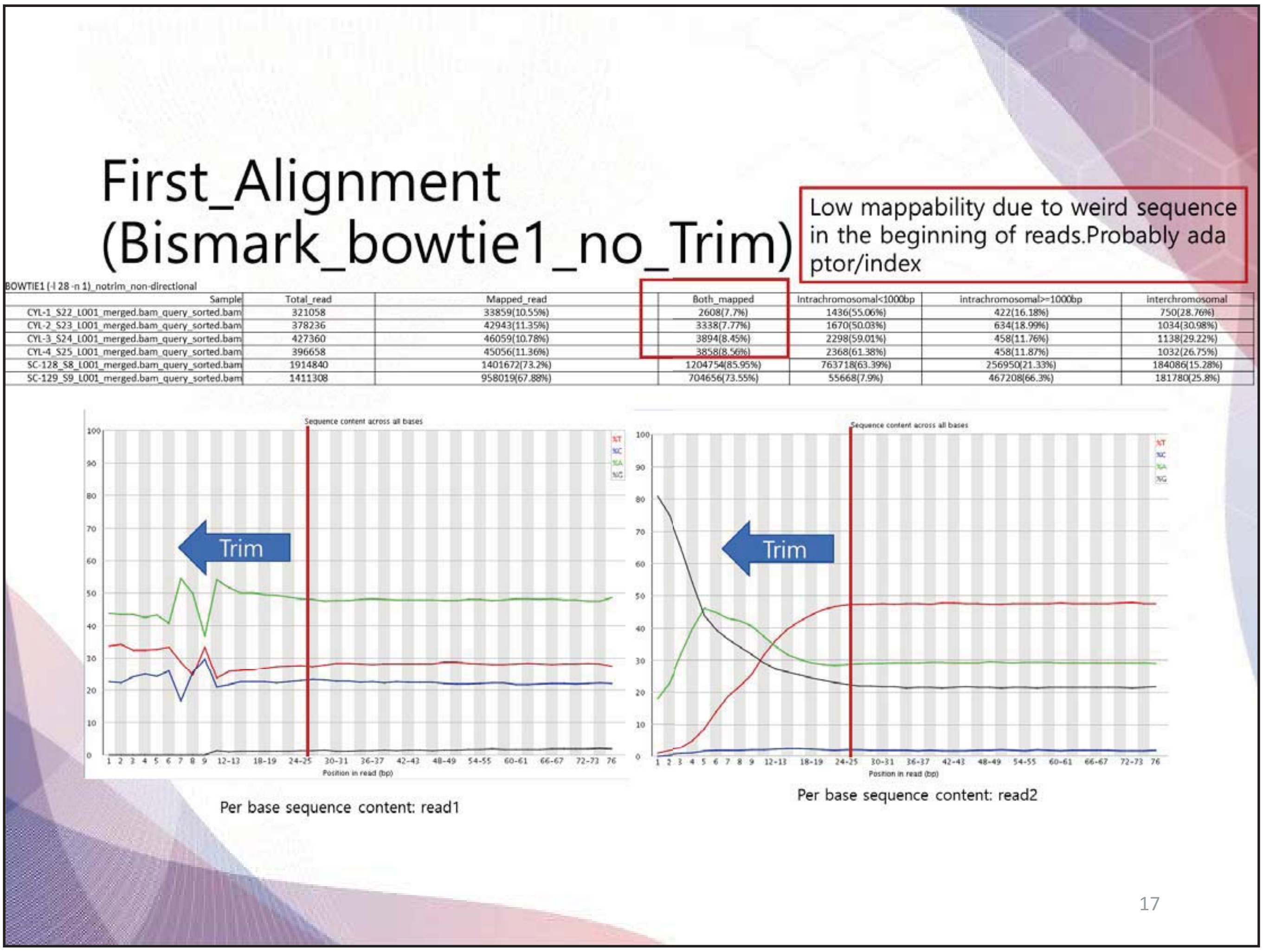
FASTQC-**Per base sequence content**



Good



Bad



First_Alignment (Bismark_bowtie1_no_Trim)

Low mappability due to weird sequence in the beginning of reads.Probably ada ptor/index

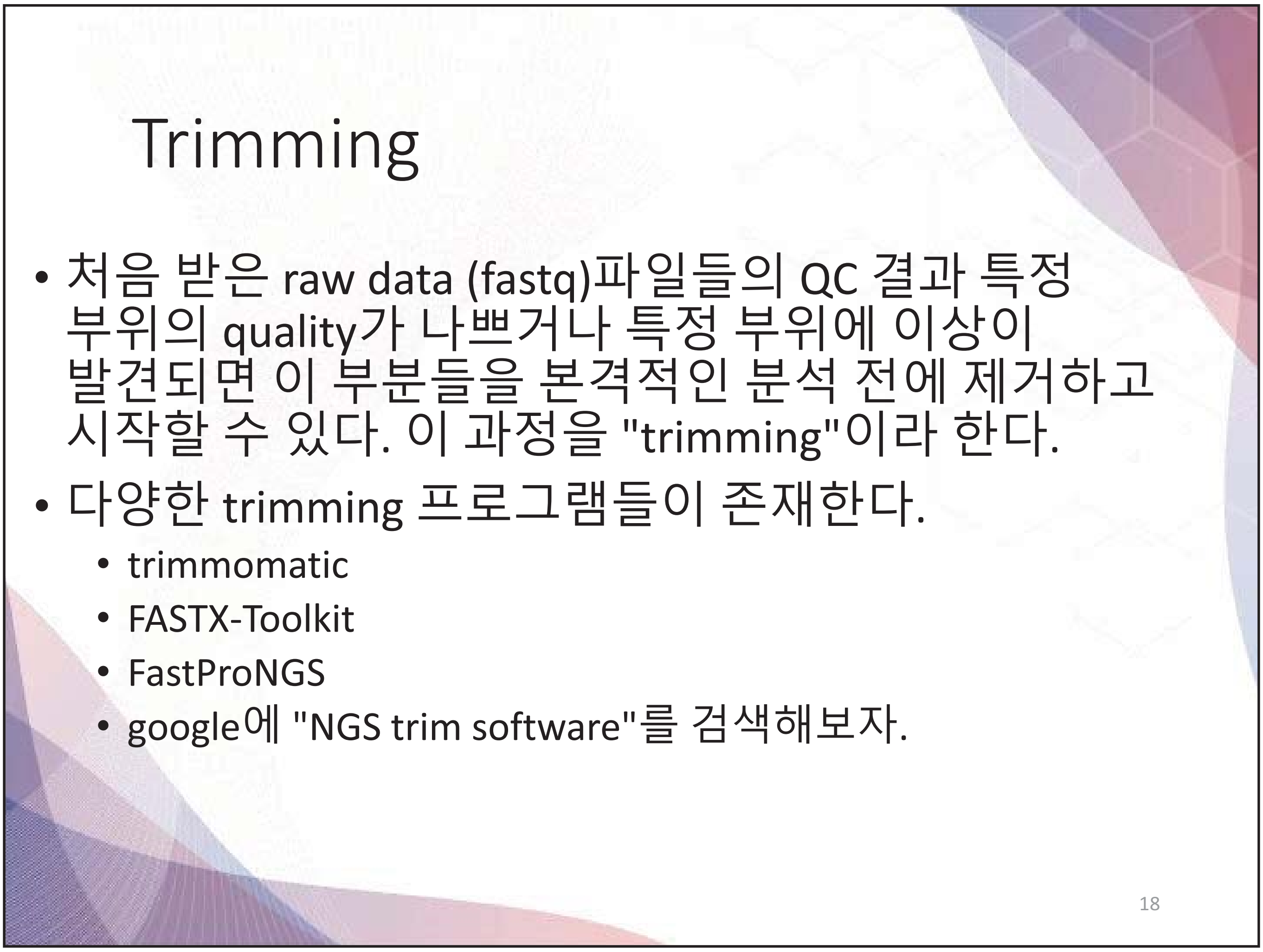
BOWTIE1 (-l 28 -n 1)_notrim_non-directional

Sample	Total_read	Mapped_read	Both_mapped	Intrachromosomal<1000bp	intrachromosomal>=1000bp	interchromosomal
CYL-1_S22_L001_merged.bam_query_sorted.bam	321058	33859(10.55%)	2608(7.7%)	1436(55.06%)	422(16.18%)	750(28.76%)
CYL-2_S23_L001_merged.bam_query_sorted.bam	378236	42943(11.35%)	3338(7.77%)	1670(50.03%)	634(18.99%)	1034(30.98%)
CYL-3_S24_L001_merged.bam_query_sorted.bam	427360	46059(10.78%)	3894(8.45%)	2298(59.01%)	458(11.76%)	1138(29.22%)
CYL-4_S25_L001_merged.bam_query_sorted.bam	396658	45056(11.36%)	3858(8.56%)	2368(61.38%)	458(11.87%)	1032(26.75%)
SC-128_S8_L001_merged.bam_query_sorted.bam	1914840	1401672(73.2%)	1204754(85.95%)	763718(63.39%)	256950(21.33%)	184086(15.28%)
SC-129_S9_L001_merged.bam_query_sorted.bam	1411308	958019(67.88%)	704656(73.55%)	55668(7.9%)	467208(66.3%)	181780(25.8%)

Per base sequence content: read1

Per base sequence content: read2

17

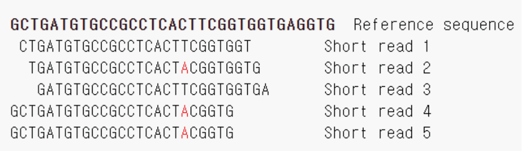


Trimming

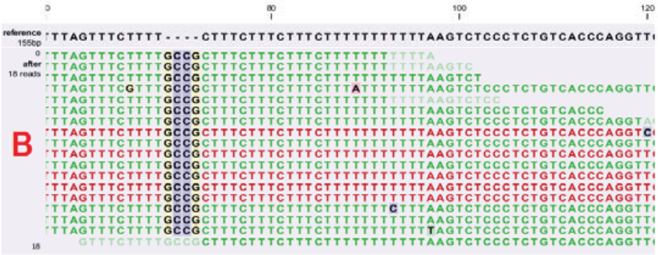
- 처음 받은 raw data (fastq)파일들의 QC 결과 특정 부위의 quality가 나쁘거나 특정 부위에 이상이 발견되면 이 부분들을 본격적인 분석 전에 제거하고 시작할 수 있다. 이 과정을 "trimming"이라 한다.
- 다양한 trimming 프로그램들이 존재한다.
  - trimmomatic
  - FASTX-Toolkit
  - FastProNGS
  - google에 "NGS trim software"를 검색해보자.

18

Alignment, Mapping



```
GCTGATGTGCCGCCTCACTTCGGTGGTGAGGTG  Reference sequence
 CTGATGTGCCGCCTCACTTCGGTGGT        Short read 1
  TGATGTGCCGCCTCACTACGGTGGTG       Short read 2
   GATGTGCCGCCTCACTTCGGTGGTGA      Short read 3
GCTGATGTGCCGCCTCACTACGGTG          Short read 4
GCTGATGTGCCGCCTCACTACGGTG          Short read 5
```

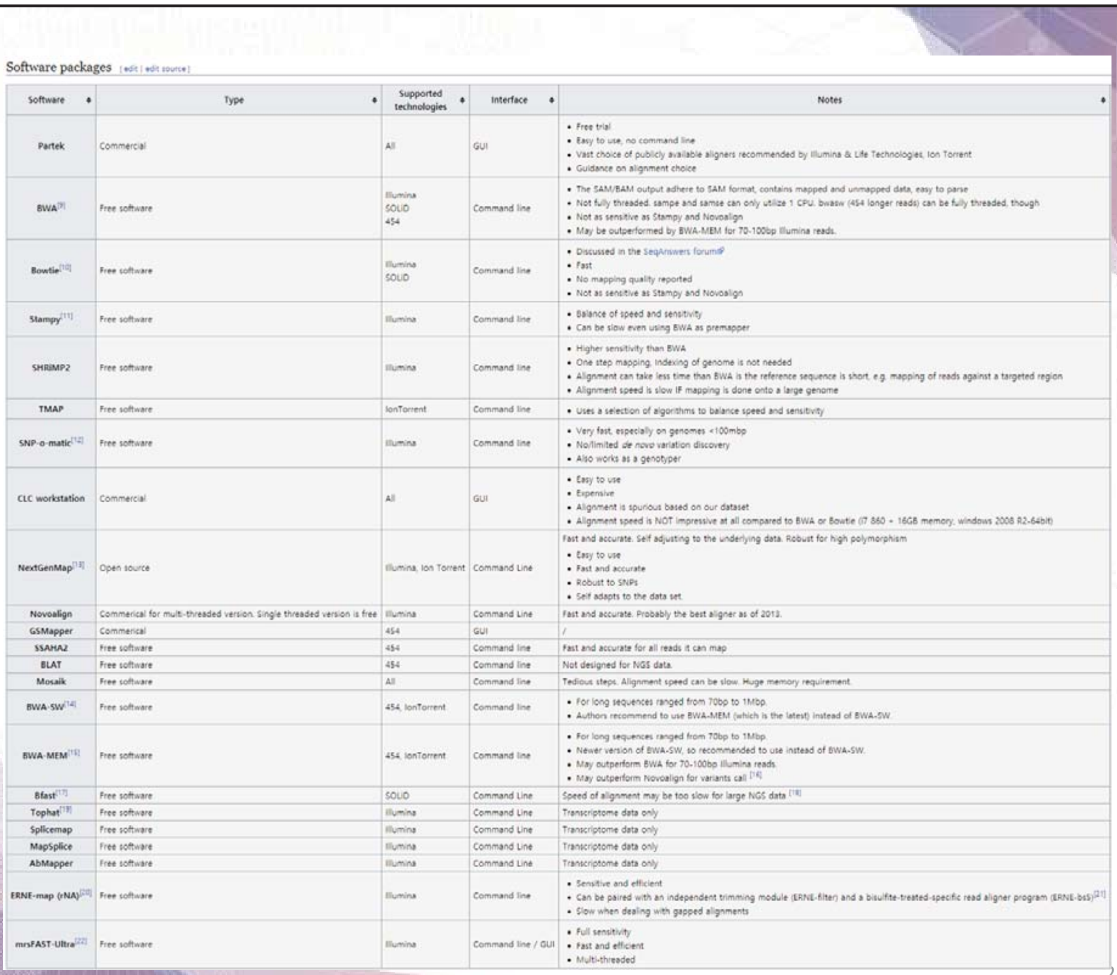

0
80
100
120
reference
155bp
0
after
18 reads
18
B
TTAGTTTCTTTT----CTTTCTTTCTTTCTTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT

Alignment

- Next-generation sequencing generally produces *short reads* or *short read pairs*, meaning short sequences of <~200 bases. To compare the DNA of the sequenced sample to its reference sequence, we need to find the corresponding part of that sequence for each read in our sequencing data. This is called **aligning** or **mapping** the reads against the reference sequence. Once this is done, we can look for variation (e.g. SNPs) within the sample.

Alignment, mapping

- blast: https://www.ncbi.nlm.nih.gov/
- blat: http://genome.ucsc.edu/
- high throughput data aligner
  - bwa
  - bowtie
  - gsnap
  - hisat
  - star



Software packages [edit | edit source]

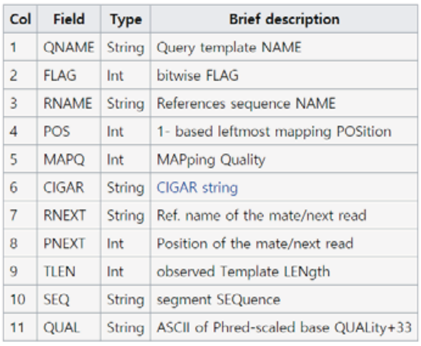
Software	Type	Supported technologies	Interface	Notes
Partek	Commercial	All	GUI	• Free trial • Easy to use, no command line • Vast choice of publicly available aligners recommended by Illumina & Life Technologies, Ion Torrent • Guidance on alignment choice
BWA[9]	Free software	Illumina SOLiD 454	Command line	• The SAM/BAM output adhere to SAM format, contains mapped and unmapped data, easy to parse • Not fully threaded. sampe and samse can only utilize 1 CPU. bwasw (454 longer reads) can be fully threaded, though • Not as sensitive as Stampy and Novoalign • May be outperformed by BWA-MEM for 70-100bp Illumina reads.
Bowtie[10]	Free software	Illumina SOLiD	Command line	• Discussed in the SeqAnswers forum • Fast • No mapping quality reported • Not as sensitive as Stampy and Novoalign
Stampy[11]	Free software	Illumina	Command line	• Balance of speed and sensitivity • Can be slow even using BWA as premapper
SHRiMP2	Free software	Illumina	Command line	• Higher sensitivity than BWA • One step mapping. Indexing of genome is not needed • Alignment can take less time than BWA is the reference sequence is short, e.g. mapping of reads against a targeted region • Alignment speed is slow IF mapping is done onto a large genome
TMAP	Free software	IonTorrent	Command line	• Uses a selection of algorithms to balance speed and sensitivity
SNP-o-matic[12]	Free software	Illumina	Command line	• Very fast, especially on genomes <100mbp • No/limited *de novo* variation discovery • Also works as a genotyper
CLC workstation	Commercial	All	GUI	• Easy to use • Expensive • Alignment is spurious based on our dataset • Alignment speed is NOT impressive at all compared to BWA or Bowtie (i7 860 + 16GB memory, windows 2008 R2-64bit)
NextGenMap[13]	Open source	Illumina, Ion Torrent	Command Line	Fast and accurate. Self adjusting to the underlying data. Robust for high polymorphism • Easy to use • Fast and accurate • Robust to SNPs • Self adapts to the data set
Novoalign	Commerical for multi-threaded version. Single threaded version is free	Illumina	Command Line	Fast and accurate. Probably the best aligner as of 2013.
GSMapper	Commerical	454	GUI	/
SSAHA2	Free software	454	Command line	Fast and accurate for all reads it can map
BLAT	Free software	454	Command line	Not designed for NGS data.
Mosaik	Free software	All	Command line	Tedious steps. Alignment speed can be slow. Huge memory requirement.
BWA-SW[14]	Free software	454, IonTorrent	Command line	• For long sequences ranged from 70bp to 1Mbp. • Authors recommend to use BWA-MEM (which is the latest) instead of BWA-SW.
BWA-MEM[15]	Free software	454, IonTorrent	Command line	• For long sequences ranged from 70bp to 1Mbp. • Newer version of BWA-SW, so recommended to use instead of BWA-SW. • May outperform BWA for 70-100bp Illumina reads. • May outperform Novoalign for variants call [16]
Bfast[17]	Free software	SOLiD	Command Line	Speed of alignment may be too slow for large NGS data [18]
Tophat[19]	Free software	Illumina	Command Line	Transcriptome data only
Splicemap	Free software	Illumina	Command Line	Transcriptome data only
MapSplice	Free software	Illumina	Command Line	Transcriptome data only
AbMapper	Free software	Illumina	Command Line	Transcriptome data only
ERNE-map (rNA)[20]	Free software	Illumina	Command line	• Sensitive and efficient • Can be paired with an independent trimming module (ERNE-filter) and a bisulfite-treated-specific read aligner program (ERNE-bs5)[21] • Slow when dealing with gapped alignments
mrsFAST-Ultra[22]	Free software	Illumina	Command line / GUI	• Full sensitivity • Fast and efficient • Multi-threaded

SAM/BAM

- SAM stands for Sequence Alignment/Map format.
- BAM stands for Binary Alignment Map
- It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with ‘@’, while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information

SAM

```
SRR10598741.11  83     chr1    10010   11     8S50M3D87M5S    =      10056   -94
ACCCTCCCCCCTACCCCTCACCCTAACCCTACCCCTAACCCTAACCCTACCCCCAACCACCCTACCCCTAACCCTAACCCCAACCCTACCCCTAACCCTACCCCAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCAACCC  )F<--))AAAA7A-<AA))-7-7--7F<----F777--JA77--AA7-7A777-7--7-<FA-7-FF<7--JJAA-AFA7--7JFF77--F<-<-AF77-<FJFA-
JJJJJJJJJJFJJJJJFJJJFAAJJJJJJJJJJJFJJJFAFAA  NM:i:12 MD:Z:5A4A12A17A3T4^CTA6A15T7A11A44      MC:Z:58M92S     AS:i:83 XS:i:78

SRR10598741.11  163    chr1    10056   11     58M92S =      10010   94
AACCCTAACCCTAACCCTAACCCTAACCCTAACCCAAACCCAAACCCAAACCCAACCCCAAAGCAAACCACCCCCAACCCCAAACCCAACCCCACACCAAAACACAAACCCCAACCCACA
CCCCAACCCCCACACCAACCCCAACCCCCC  AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ-F<JJF-F-FJ7<<-7AA--7-7-------------7-7--77F----7A---7----7-7<--7----------)))--)-))7)))))<-))))-)-)---
))7)))))7))  NM:i:3  MD:Z:35T5T5T10  MC:Z:8S50M3D87M5S       AS:i:43 XS:i:48

SRR10598741.12  83     chr1    10004   0      55S95M =      10005   -94
AGCCGGCATACGAGATGCTCCTGTGACTGGAGCTCAGCCGTGTCCTCTTCCGAACCCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCTAACCCTAACCCTAACCCTA  --)AF<----7----)))7)-FFA--A)777---<F7-F-A-7-7----A7---7-
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJFJJJJJJJJJJJFJJJJJFJJJJJFJJJJFFJJJJJJJJJJJJJJJJJJJFFFAA  NM:i:0  MD:Z:95 MC:Z:108M42S     AS:i:95 XS:i:98

SRR10598741.12  163    chr1    10005   0      108M42S =      10004   94
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTACCCCCAACCCAACCTCTACCCCGAACAT
CGTGTCGACCTCGGTCGTGGCCGTATCA  AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJFFJFJJJJJJJJ-FJ<JJJJJFJJ-7F-JAF<F-A<A-FFJFF---------7--)--7-7-<))))7)7------))-)<A<A-))-
)-7)-)))-7-  NM:i:2  MD:Z:94A3T9     MC:Z:55S95M     AS:i:98 XS:i:94
```



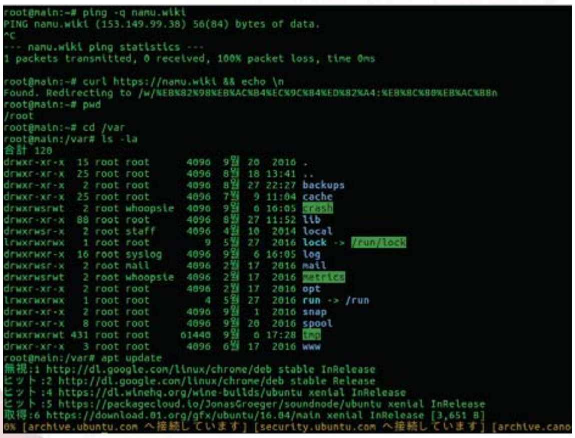
Col	Field	Type	Brief description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

sam 파일 다루기-samtools & picard

- Alignment를 하고 sam 파일을 시작으로 다양한 pre-processing 과정이 필요하다.
- convert to bam: binary file로 바꾸어 줌으로써 파일 사이즈를 줄이고 indexing을 통해 빠르게 분석할 수 있도록 한다.
- Sort: bam을 다루기 쉽도록 하기 위해 align된 위치 순서에 따라 read를 재배열 한다.
- Deduplication: Library preparation 과정 중 PCR에서 생긴 Duplication을 제거한다.

- 위 과정들을 해주는 다양한 프로그램들이 존재하며 특히 samtools와 picard가 널리 쓰이고 있다.

CLI (Command-line User Interface)



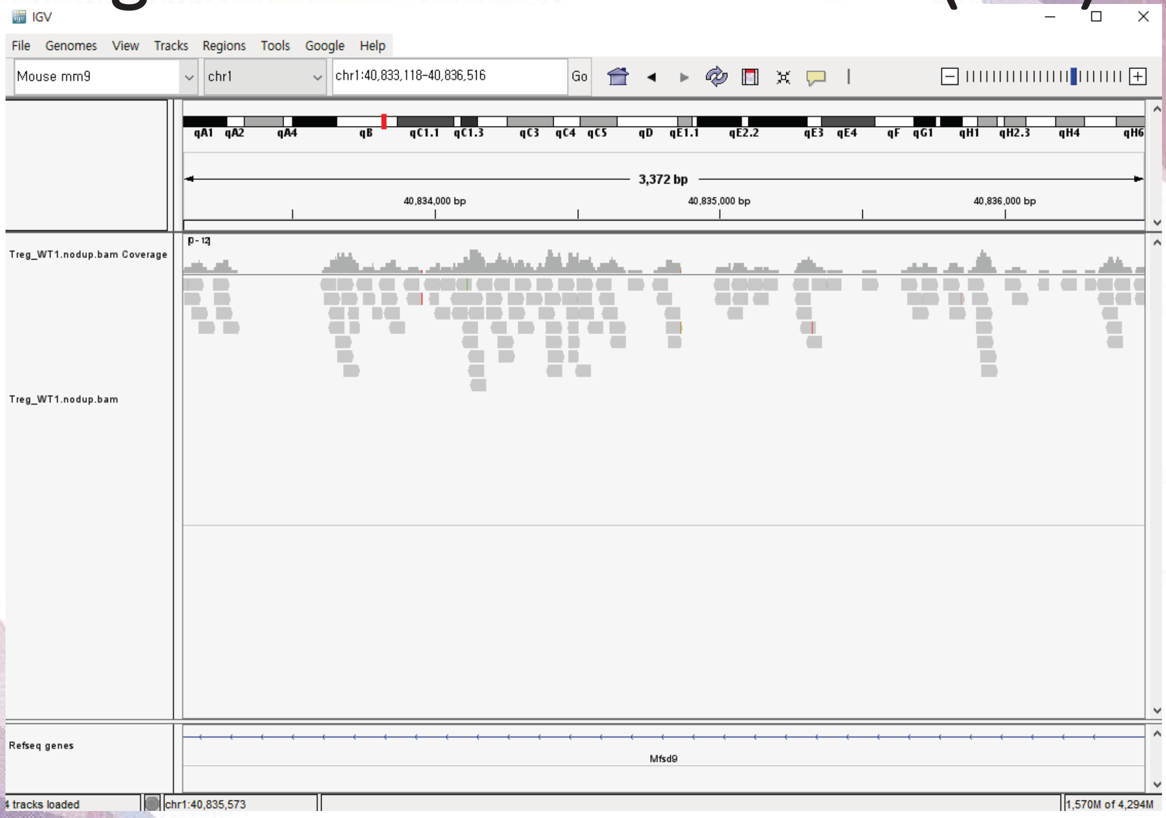
```
root@main:~# ping -q namu.wiki
PING namu.wiki (153.149.99.38) 56(84) bytes of data.
^C
--- namu.wiki ping statistics ---
1 packets transmitted, 0 received, 100% packet loss, time 0ms

root@main:~# curl https://namu.wiki && echo \n
Found. Redirecting to /w/%EB%82%98%EB%AC%B4%EC%9C%84%ED%82%A4:%EB%8C%80%EB%AC%B8n
root@main:~# pwd
/root
root@main:~# cd /var
root@main:/var# ls -la
合計 120
drwxr-xr-x  15 root root      4096  9월  20  2016 .
drwxr-xr-x  25 root root      4096  8월  18 13:41 ..
drwxr-xr-x   2 root root      4096  8월  27 22:27 backups
drwxr-xr-x  25 root root      4096  7월   9 11:04 cache
drwxrwsrwt   2 root whoopsie  4096  9월   6 16:05 crash
drwxr-xr-x  88 root root      4096  8월  27 11:52 lib
drwxrwsr-x   2 root staff     4096  4월  10  2014 local
lrwxrwxrwx   1 root root         9  5월  27  2016 lock -> /run/lock
drwxrwxr-x  16 root syslog    4096  9월   6 16:05 log
drwxrwsr-x   2 root mail      4096  2월  17  2016 mail
drwxrwsrwt   2 root whoopsie  4096  2월  17  2016 metrics
drwxr-xr-x   2 root root      4096  2월  17  2016 opt
lrwxrwxrwx   1 root root         4  5월  27  2016 run -> /run
drwxr-xr-x   2 root root      4096  9월   1  2016 snap
drwxr-xr-x   8 root root      4096  9월  20  2016 spool
drwxrwxrwt 431 root root     61440  9월   6 17:28 tmp
drwxr-xr-x   3 root root      4096  6월  17  2016 www
root@main:/var# apt update
無視:1 http://dl.google.com/linux/chrome/deb stable InRelease
ヒット:2 http://dl.google.com/linux/chrome/deb stable Release
ヒット:4 https://dl.winehq.org/wine-builds/ubuntu xenial InRelease
ヒット:5 https://packagecloud.io/JonasGroeger/soundnode/ubuntu xenial InRelease
取得:6 https://download.01.org/gfx/ubuntu/16.04/main xenial InRelease [3,651 B]
0% [archive.ubuntu.com へ接続しています] [security.ubuntu.com へ接続しています] [archive.cano
```

- 명령줄 인터페이스(CLI)는 Command-Line Interface 또는 Character User Interface이다.
- 가장 대표적인 예시로는 도스, 명령 프롬프트, bash로 대표되는 유닉스 셸 환경
- CLI만의 장점: 자원을 적게 잡아 먹으면서 안정적이고 빠르다. 게다가 원격으로 작업할 때 웬만한 네트워크 환경에서도 안정적으로 작업할 수 있으며 사용되는 데이터 양 역시 압도적으로 적다. 특히 서버 쪽에서는 작업 자동화와 원격 작업이 필요한 경우가 많은데 CLI는 이 분야에서 압도적인 효율을 보여준다.
- GUI 프로그래밍에 비해 사전지식이 매우 적게 요구되며, 적당한 기본 지식이 있으면 필요한 프로그램을 쉽게 만들 수 있다.

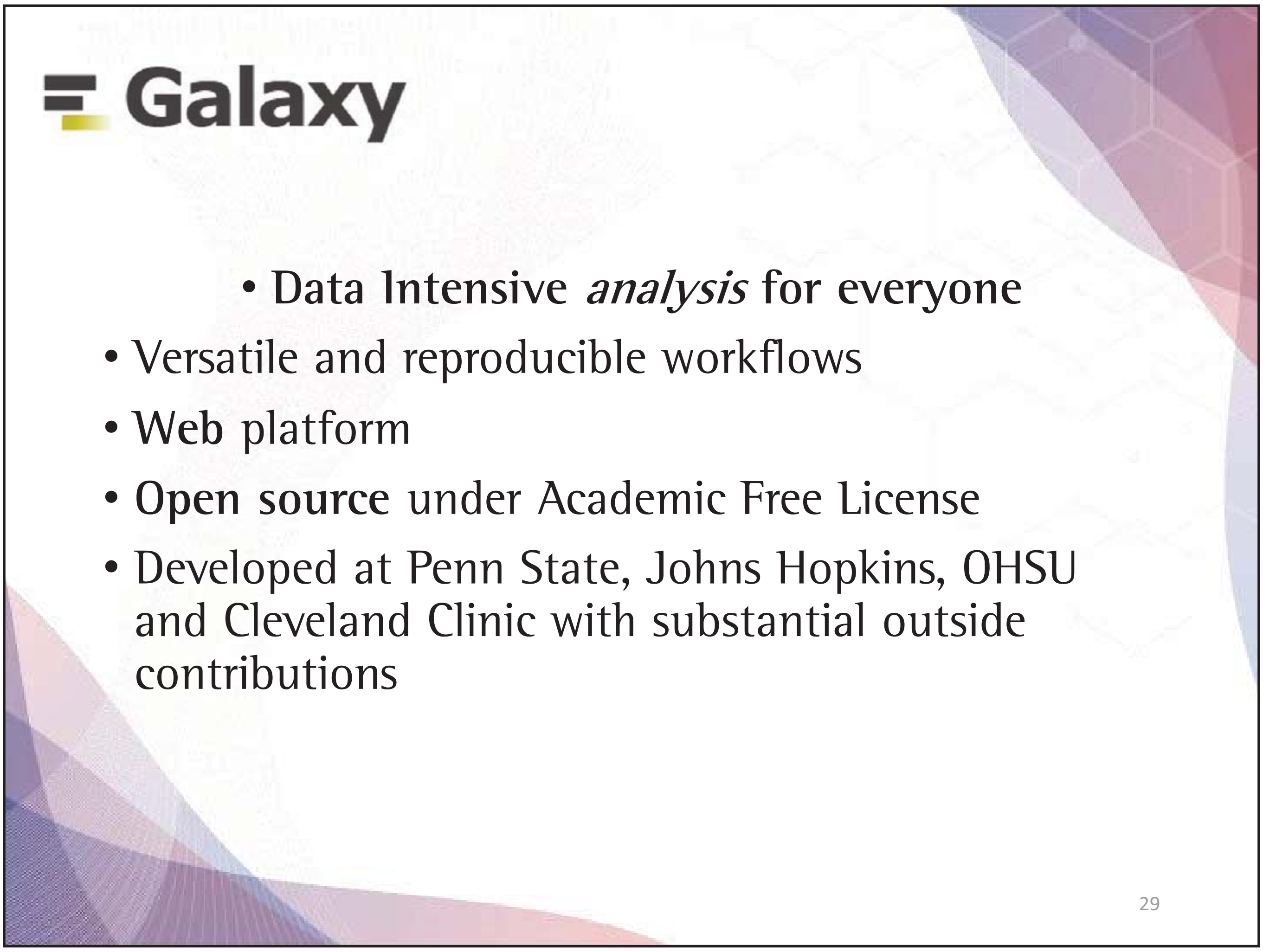
Galaxy

- https://usegalaxy.org.au/ : Galaxy 호주
- https://usegalaxy.org/ 가 오리지날이지만 호주가 더 빠르다

Integrative Genomics Viewer (IGV)

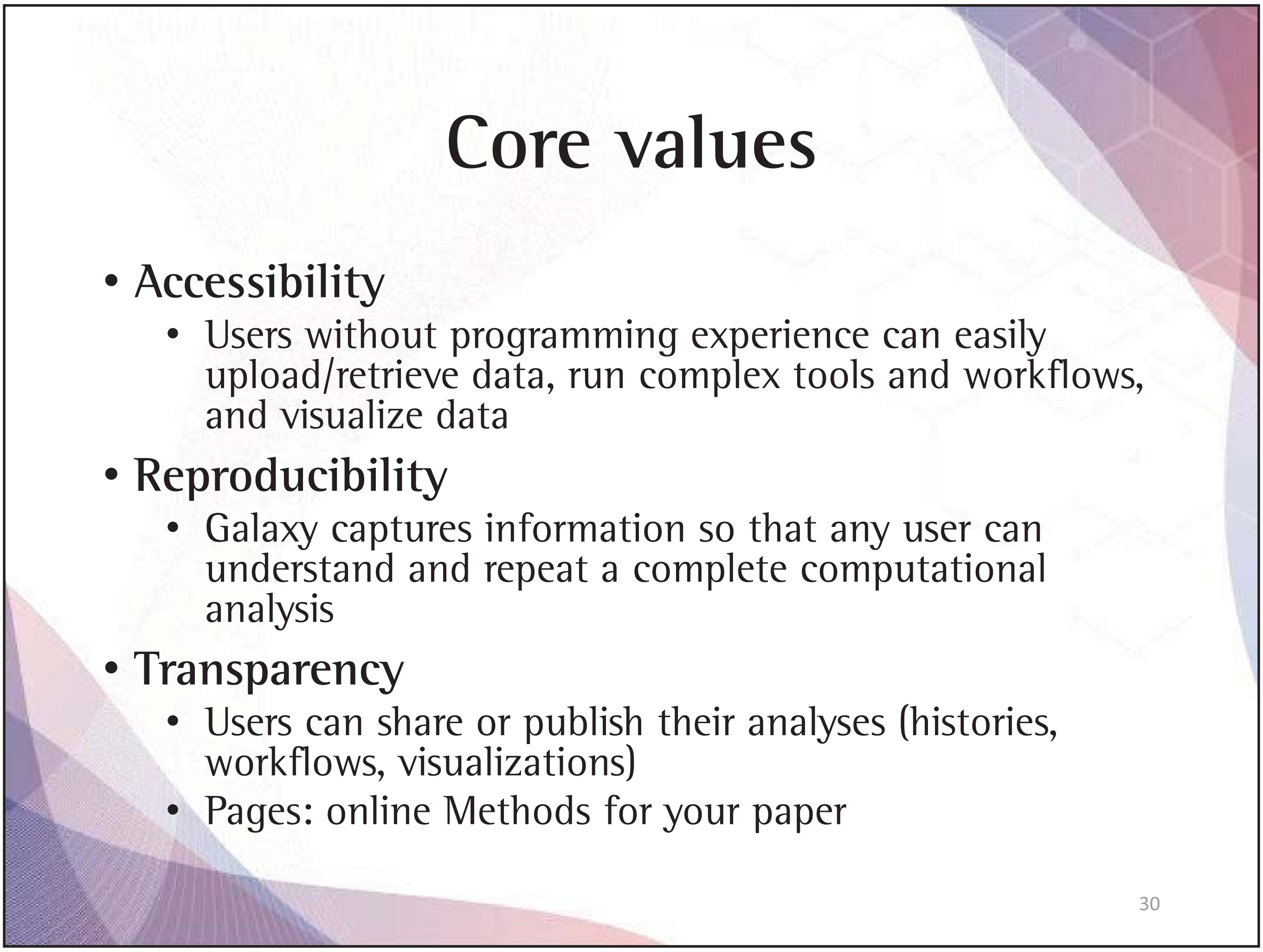

IGV
File Genomes View Tracks Regions Tools Google Help
Mouse mm9
chr1
chr1:40,833,118-40,836,516
Go
qA1 qA2 qA4 qB qC1.1 qC1.3 qC3 qC4 qC5 qD qE1.1 qE2.2 qE3 qE4 qF qG1 qH1 qH2.3 qH4 qH6
3,372 bp
40,834,000 bp
40,835,000 bp
40,836,000 bp
[0 - 12]
Treg_WT1.nodup.bam Coverage
Treg_WT1.nodup.bam
Refseq genes
Mfsd9
4 tracks loaded
chr1:40,835,573
1,570M of 4,294M

https://software.broadinstitute.org/software/igv/download



Galaxy

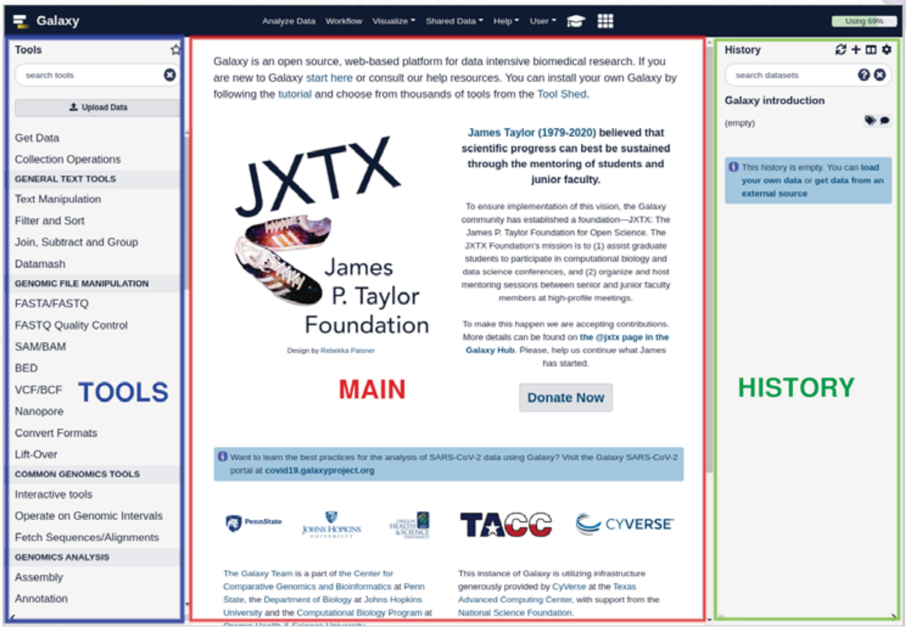
- **Data Intensive *analysis* for everyone**
- Versatile and reproducible workflows
- **Web** platform
- **Open source** under Academic Free License
- Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with substantial outside contributions



Core values

- Accessibility
  - Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data
- Reproducibility
  - Galaxy captures information so that any user can understand and repeat a complete computational analysis
- Transparency
  - Users can share or publish their analyses (histories, workflows, visualizations)
  - Pages: online Methods for your paper

Main Galaxy interface



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.

To ensure implementation of this vision, the Galaxy community has established a foundation—JXTX: The James P. Taylor Foundation for Open Science. The JXTX Foundation's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

To make this happen we are accepting contributions. More details can be found on **the @jxtx page in the Galaxy Hub**. Please, help us continue what James has started.

Design by Rebekka Paisner

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at **covid19.galaxyproject.org**

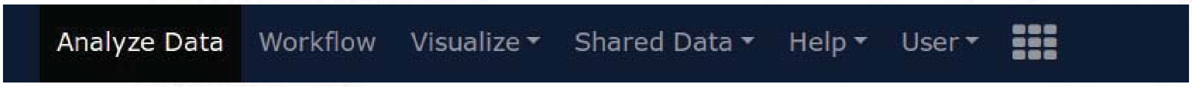
The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology at Johns Hopkins University and the Computational Biology Program at Oregon Health & Science University.

This instance of Galaxy is utilizing infrastructure generously provided by CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

This history is empty. You can **load your own data** or **get data from an external source**

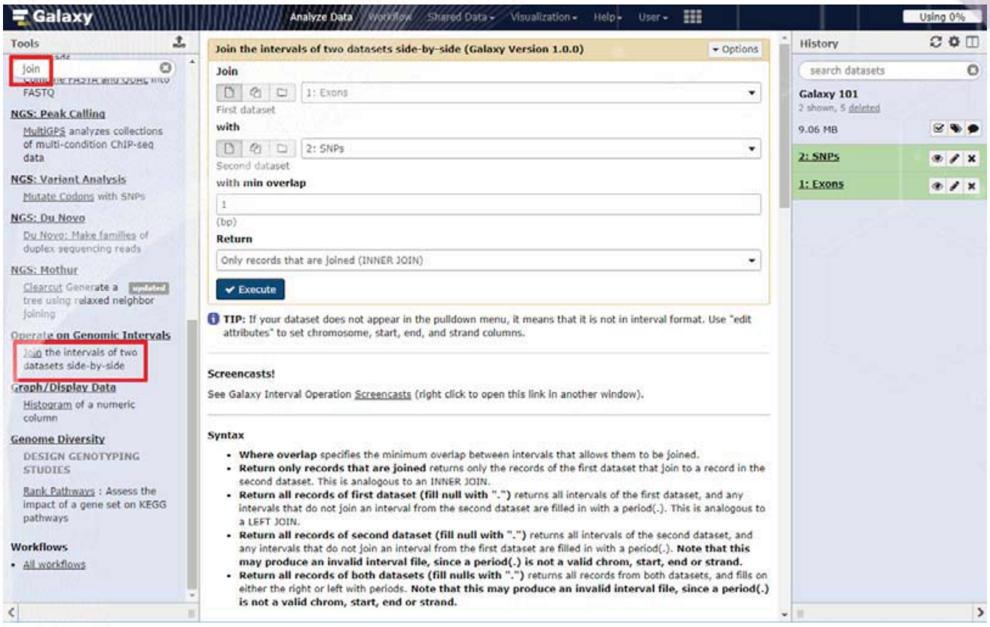
- Three main panels
  - **Left:** Available Tools
  - **Middle:** View your data and run tools
  - **Right:** Full record of your analysis **history**

Top menu



Link	***Usage***
Analyze Data	go back to the homepage
Workflow	access existing workflows or create new one using the editable diagrammatic pip eline
Visualize	create new visualisations and launch Interactive Environments
Shared Data	access data libraries, histories, workflows, visualizations and pages shared with y ou
Help	links to Galaxy Help Forum (Q&A), Galaxy Community Hub (Wiki), and Interactive Tours
User	your preferences and saved histories, datasets, pages and visualizations

Tools



Tools

Join

Combine FASTA and QUAL into FASTQ

NGS: Peak Calling

MultiGPS analyzes collections of multi-condition ChIP-seq data

NGS: Variant Analysis

Mutate Codons with SNPs

NGS: Du Novo

Du Novo: Make families of duplex sequencing reads

NGS: Mothur

Clearcut Generate a tree using relaxed neighbor joining updated

Operate on Genomic Intervals

Join the intervals of two datasets side-by-side

Graph/Display Data

Histogram of a numeric column

Genome Diversity

DESIGN GENOTYPING STUDIES

Rank Pathways : Assess the impact of a gene set on KEGG pathways

Workflows

- All workflows

Join the intervals of two datasets side-by-side (Galaxy Version 1.0.0) ▾ Options

Join

1: Exons ▾

First dataset

with

2: SNPs ▾

Second dataset

with min overlap

1

(bp)

Return

Only records that are joined (INNER JOIN) ▾

✔ Execute

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns.

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Where overlap** specifies the minimum overlap between intervals that allows them to be joined.
- **Return only records that are joined** returns only the records of the first dataset that join to a record in the second dataset. This is analogous to an INNER JOIN.
- **Return all records of first dataset (fill null with ".")** returns all intervals of the first dataset, and any intervals that do not join an interval from the second dataset are filled in with a period(.). This is analogous to a LEFT JOIN.
- **Return all records of second dataset (fill null with ".")** returns all intervals of the second dataset, and any intervals that do not join an interval from the first dataset are filled in with a period(.). **Note that this may produce an invalid interval file, since a period(.) is not a valid chrom, start, end or strand.**
- **Return all records of both datasets (fill nulls with ".")** returns all records from both datasets, and fills on either the right or left with periods. **Note that this may produce an invalid interval file, since a period(.) is not a valid chrom, start, end or strand.**

History

search datasets

Galaxy 101

2 shown, 5 deleted

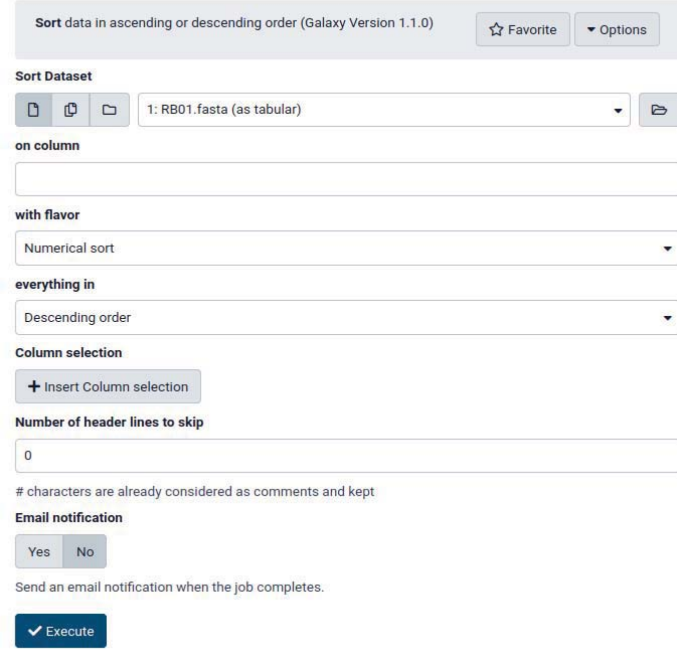
9.06 MB

2: SNPs

1: Exons

- The tool search helps in finding a tool in a crowded toolbox

Tool interface



Sort data in ascending or descending order (Galaxy Version 1.1.0) ☆ Favorite ▾ Options

Sort Dataset

1: RB01.fasta (as tabular) ▾

on column

with flavor

Numerical sort ▾

everything in

Descending order ▾

Column selection

+ Insert Column selection

Number of header lines to skip

0

characters are already considered as comments and kept

Email notification

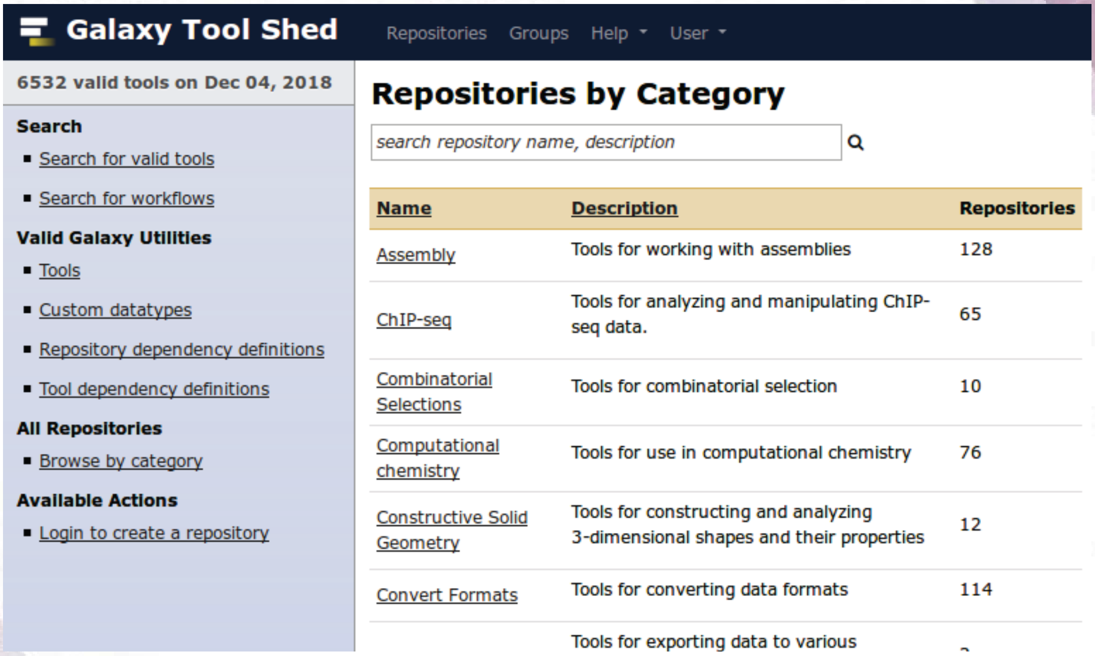
Yes No

Send an email notification when the job completes.

✔ Execute

- A tool form contains:
  - input datasets and parameters
  - help, citations, metadata
  - an **Execute** button to start a job, which will add some output datasets to the history

Tool Shed



Galaxy Tool Shed

6532 valid tools on Dec 04, 2018

Search

- Search for valid tools
- Search for workflows

Valid Galaxy Utilities

- Tools
- Custom datatypes
- Repository dependency definitions
- Tool dependency definitions

All Repositories

- Browse by category

Available Actions

- Login to create a repository

Repositories by Category

search repository name, description

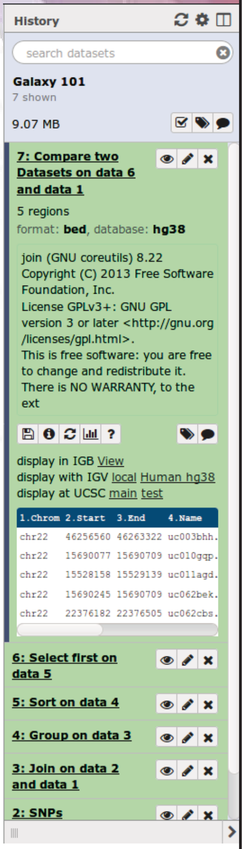
Name	Description	Repositories
Assembly	Tools for working with assemblies	128
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	65
Combinatorial Selections	Tools for combinatorial selection	10
Computational chemistry	Tools for use in computational chemistry	76
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	12
Convert Formats	Tools for converting data formats	114
	Tools for exporting data to various	

- Free "app" store: Galaxy Tool ShedThousands of tools already available
- Most software can be integrated

History

- Location of all analyses
  - collects all datasets produced by tools
  - collects all operations performed on the data
- For each dataset (the heart of Galaxy's reproducibility), the history tracks
  - name, format, size, creation time, datatype-specific metadata
  - tool id, version, inputs, parameters
  - standard output (stdout) and error (stderr)
  - state (waiting, running, success, failed)
  - hidden, deleted, purged

- Three buttons
  - 👁 View the file
  - ✎ Edit attributes
    - e.g. change name
  - ✕ Delete file



History

search datasets

Galaxy 101

7 shown

9.07 MB

7: Compare two Datasets on data 6 and data 1

5 regions

format: **bed**, database: **hg38**

```
join (GNU coreutils) 8.22
Copyright (C) 2013 Free Software
Foundation, Inc.
License GPLv3+: GNU GPL
version 3 or later <http://gnu.org
/licenses/gpl.html>.
This is free software: you are free
to change and redistribute it.
There is NO WARRANTY, to the
ext
```

display in IGB View
display with IGV local Human hg38
display at UCSC main test

```
1.Chrom 2.Start  3.End    4.Name
chr22   46256560 46263322 uc003bhh.
chr22   15690077 15690709 uc010gqp.
chr22   15528158 15529139 uc011agd.
chr22   15690245 15690709 uc062bek.
chr22   22376182 22376505 uc062cbs.
```

6: Select first on data 5

5: Sort on data 4

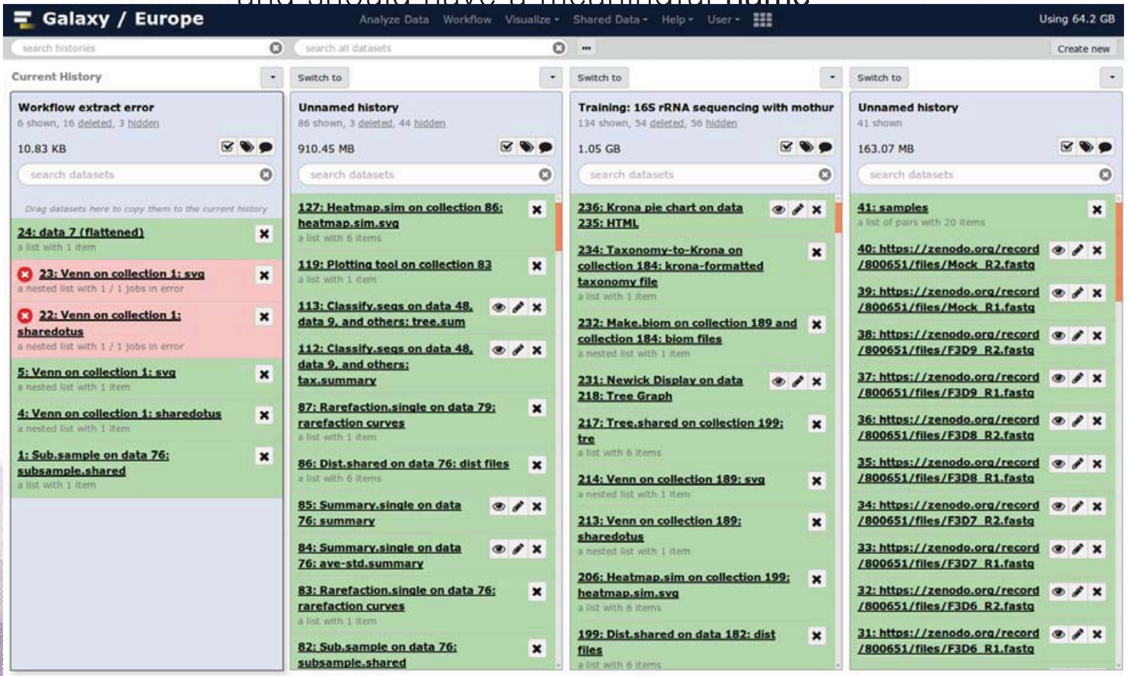
4: Group on data 3

3: Join on data 2 and data 1

2: SNPs

Multiple histories

- You can have as many histories as you want
  - each history should correspond to a **different analysis**
  - and should have a meaningful **name**



Galaxy / Europe

Analyze Data Workflow Visualize Shared Data Help User

Using 64.2 GB

search histories

search all datasets

Create new

Current History

Workflow extract error
6 shown, 16 deleted, 3 hidden
10.83 KB
search datasets
Drag datasets here to copy them to the current history
24: data 7 (flattened) — a list with 1 item
23: Venn on collection 1: svg — a nested list with 1 / 1 jobs in error
22: Venn on collection 1: sharedotus — a nested list with 1 / 1 jobs in error
5: Venn on collection 1: svg — a nested list with 1 item
4: Venn on collection 1: sharedotus — a nested list with 1 item
1: Sub.sample on data 76: subsample.shared — a list with 1 item

Switch to

Unnamed history
86 shown, 3 deleted, 44 hidden
910.45 MB
search datasets
127: Heatmap.sim on collection 86: heatmap.sim.svg — a list with 6 items
119: Plotting tool on collection 83 — a list with 1 item
113: Classify.seqs on data 48, data 9, and others: tree.sum
112: Classify.seqs on data 48, data 9, and others: tax.summary
87: Rarefaction.single on data 79: rarefaction curves — a list with 1 item
86: Dist.shared on data 76: dist files — a list with 6 items
85: Summary.single on data 76: summary
84: Summary.single on data 76: ave-std.summary
83: Rarefaction.single on data 76: rarefaction curves — a list with 1 item
82: Sub.sample on data 76: subsample.shared

Switch to

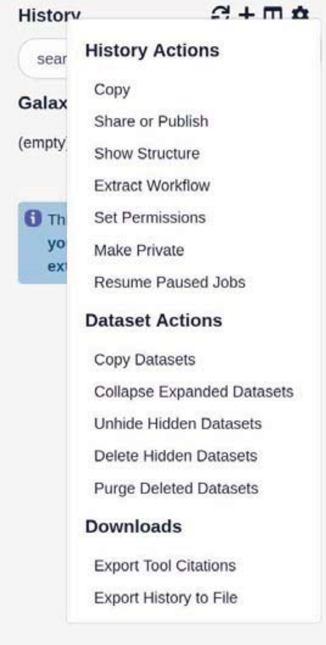
Training: 16S rRNA sequencing with mothur
134 shown, 54 deleted, 56 hidden
1.05 GB
search datasets
236: Krona pie chart on data 235: HTML
234: Taxonomy-to-Krona on collection 184: krona-formatted taxonomy file — a list with 1 item
232: Make.biom on collection 189 and collection 184: biom files — a nested list with 1 item
231: Newick Display on data 218: Tree Graph
217: Tree.shared on collection 199: tre — a list with 6 items
214: Venn on collection 189: svg — a nested list with 1 item
213: Venn on collection 189: sharedotus — a nested list with 1 item
206: Heatmap.sim on collection 199: heatmap.sim.svg — a list with 6 items
199: Dist.shared on data 182: dist files — a list with 6 items

Switch to

Unnamed history
41 shown
163.07 MB
search datasets
41: samples — a list of pairs with 20 items
40: https://zenodo.org/record/800651/files/Mock_R2.fastq
39: https://zenodo.org/record/800651/files/Mock_R1.fastq
38: https://zenodo.org/record/800651/files/F3D9_R2.fastq
37: https://zenodo.org/record/800651/files/F3D9_R1.fastq
36: https://zenodo.org/record/800651/files/F3D8_R2.fastq
35: https://zenodo.org/record/800651/files/F3D8_R1.fastq
34: https://zenodo.org/record/800651/files/F3D7_R2.fastq
33: https://zenodo.org/record/800651/files/F3D7_R1.fastq
32: https://zenodo.org/record/800651/files/F3D6_R2.fastq
31: https://zenodo.org/record/800651/files/F3D6_R1.fastq

History options menu

History behavior is controlled by the *History options* (gear icon)



History

sear

Galax

(empty

Th yo ex

History Actions

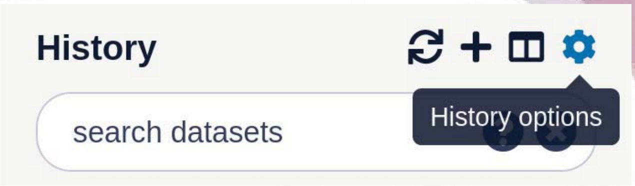
- Copy
- Share or Publish
- Show Structure
- Extract Workflow
- Set Permissions
- Make Private
- Resume Paused Jobs

Dataset Actions

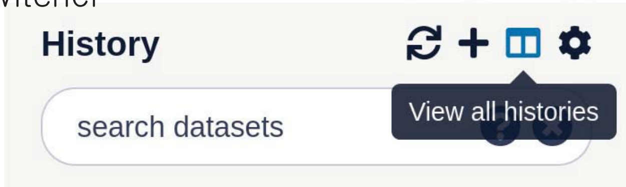
- Copy Datasets
- Collapse Expanded Datasets
- Unhide Hidden Datasets
- Delete Hidden Datasets
- Purge Deleted Datasets

Downloads

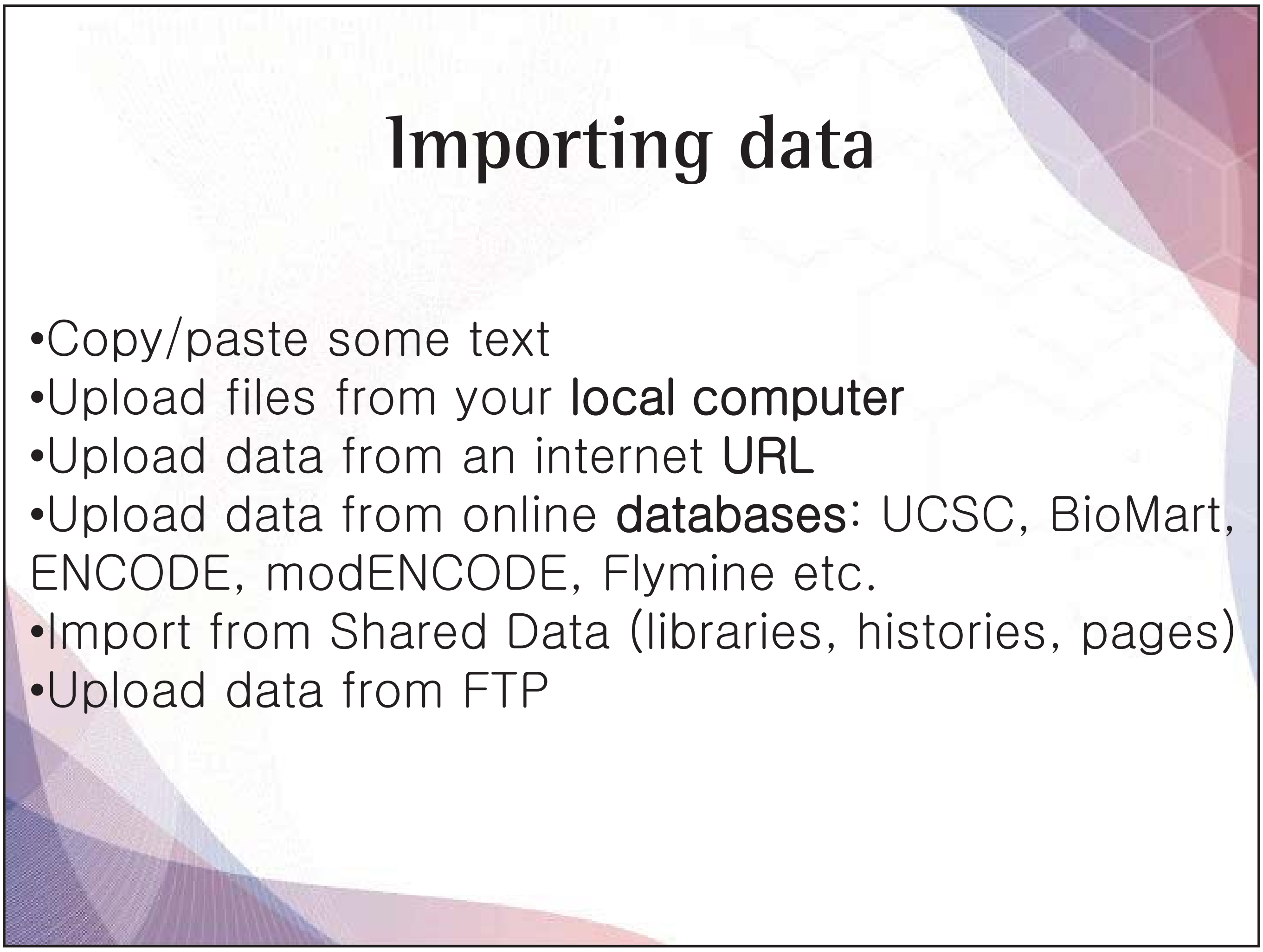
- Export Tool Citations
- Export History to File



- *Create new history* (+ icon) will **not** make your current history disappear
- To see all of your histories, use the history switcher

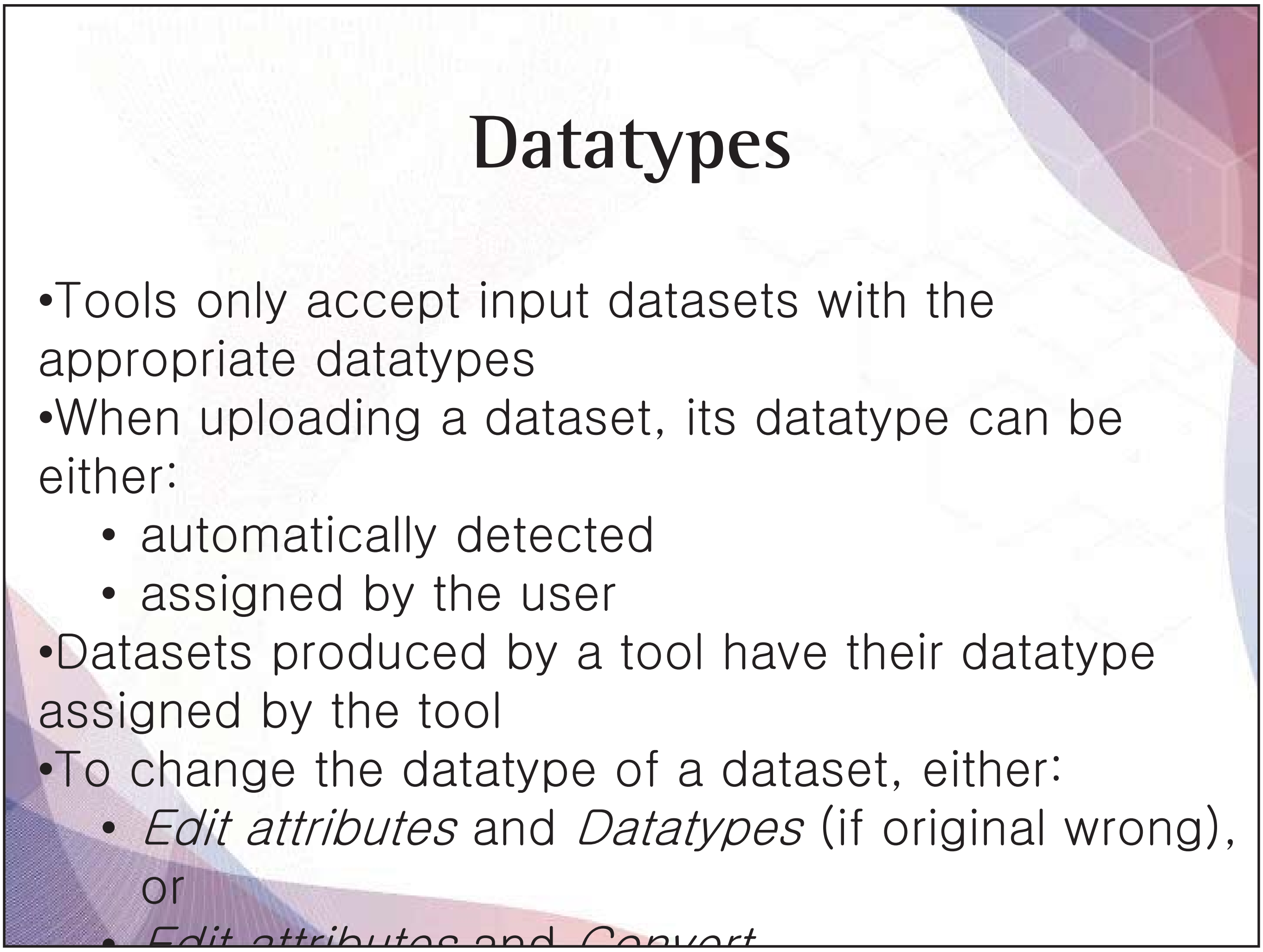


- *Copy Datasets* from one history to another and save disk space for your quota



Importing data

- Copy/paste some text
- Upload files from your **local computer**
- Upload data from an internet **URL**
- Upload data from online **databases**: UCSC, BioMart, ENCODE, modENCODE, Flymine etc.
- Import from Shared Data (libraries, histories, pages)
- Upload data from FTP



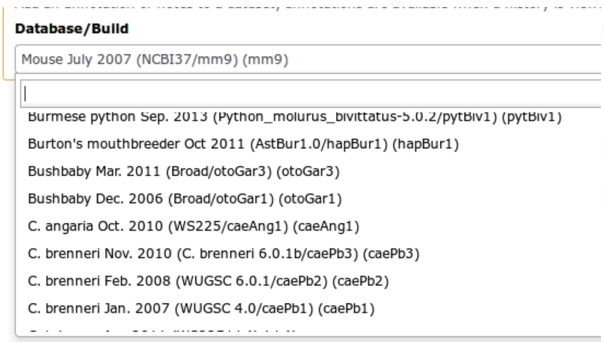
Datatypes

- Tools only accept input datasets with the appropriate datatypes
- When uploading a dataset, its datatype can be either:
  - automatically detected
  - assigned by the user
- Datasets produced by a tool have their datatype assigned by the tool
- To change the datatype of a dataset, either:
  - *Edit attributes* and *Datatypes* (if original wrong), or
  - *Edit attributes* and *Convert*

Reference datasets

Example: reference Genome

- Genome build specifies which genome assembly a dataset is associated with
    - e.g. mm10, hg38...
- Can be assigned by a tool or by the user
- Users can create custom genome builds
- New builds can be added by the admin



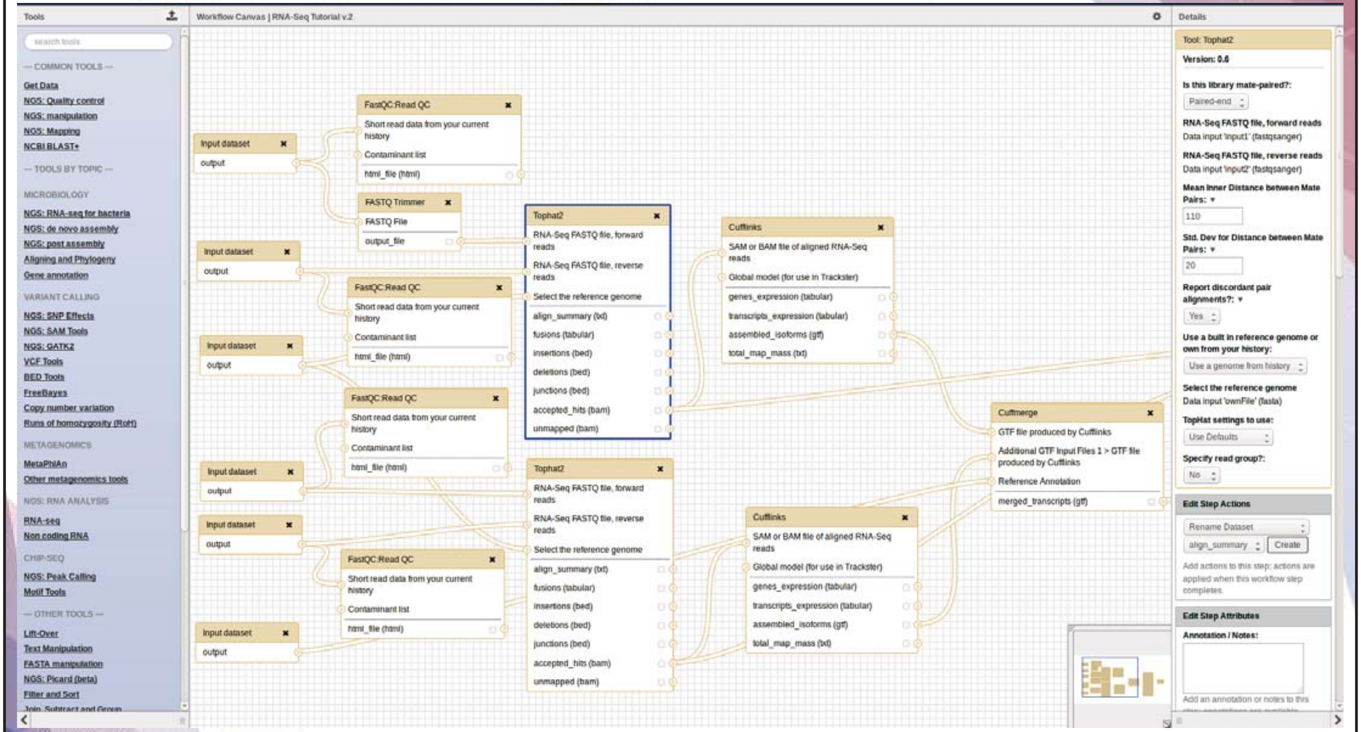
Database/Build

Mouse July 2007 (NCBI37/mm9) (mm9)

Burmese python Sep. 2013 (Python_molurus_bivittatus-5.0.2/pytBiv1) (pytBiv1)
Burton's mouthbreeder Oct 2011 (AstBur1.0/hapBur1) (hapBur1)
Bushbaby Mar. 2011 (Broad/otoGar3) (otoGar3)
Bushbaby Dec. 2006 (Broad/otoGar1) (otoGar1)
C. angaria Oct. 2010 (WS225/caeAng1) (caeAng1)
C. brenneri Nov. 2010 (C. brenneri 6.0.1b/caePb3) (caePb3)
C. brenneri Feb. 2008 (WUGSC 6.0.1/caePb2) (caePb2)
C. brenneri Jan. 2007 (WUGSC 4.0/caePb1) (caePb1)

Workflows

Workflow Editor



Tool: Tophat2

Version: 0.6

Is this library mate-paired?: Paired-end

RNA-Seq FASTQ file, forward reads: Data input 'input1' (fastqsanger)

RNA-Seq FASTQ file, reverse reads: Data input 'input2' (fastqsanger)

Mean Inner Distance between Mate Pairs: 110

Std. Dev for Distance between Mate Pairs: 20

Report discordant pair alignments?: Yes

Use a built in reference genome or own from your history: Use a genome from history

Select the reference genome: Data input 'ownFile' (fasta)

TopHat settings to use: Use Defaults

Specify read group?: No

Edit Step Actions: Rename Dataset, align_summary

Add actions to this step; actions are applied when this workflow step completes.

Annotation / Notes:

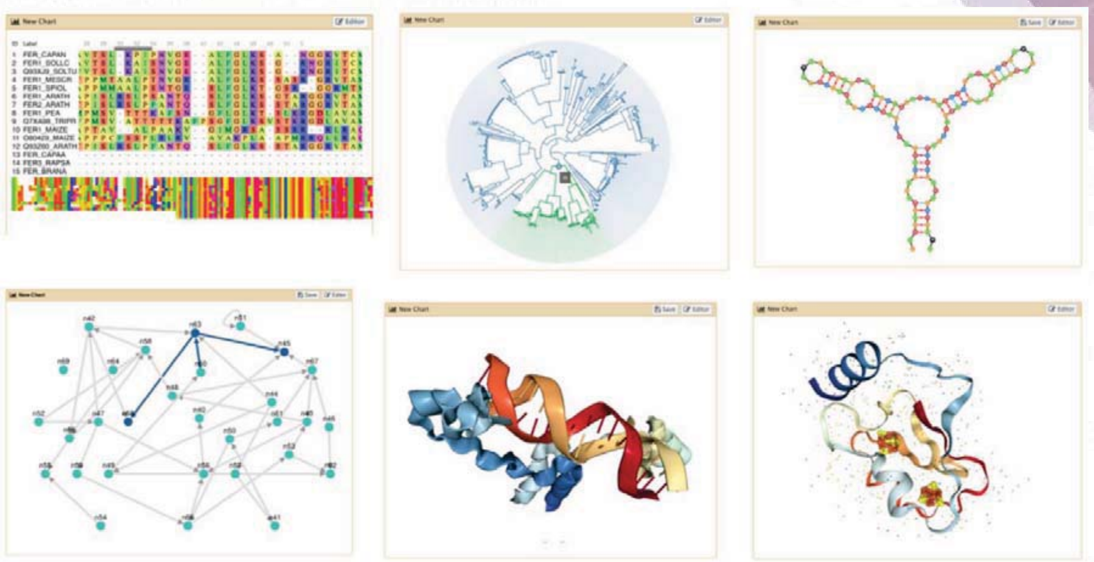
Add an annotation or notes to this

- •**Extracted** from a history
- •**Built manually** by adding and configuring tools using the canvas
- •**Imported** using an existing shared workflow

Why would you want to create workflows?

- •**Re–run** the same analysis on different input data sets
- •**Change parameters** before re–running a similar analysis
- •Make use of the workflow job **scheduling**
  - jobs are submitted as soon as their inputs are ready
- •Create **sub**–workflows: a workflow inside another workflow
- •**Share** workflows for publication and with the community

Visualizations



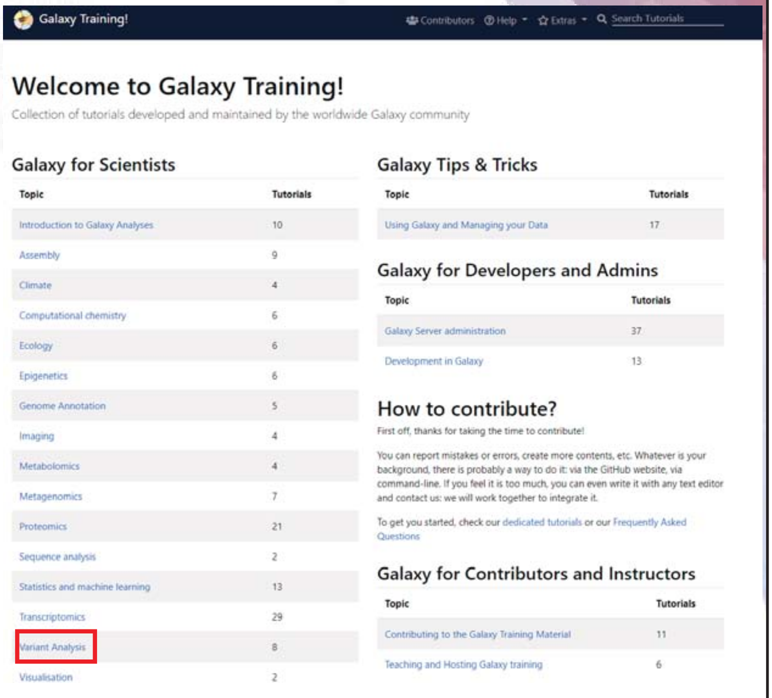
- Datatypes know what tools can be used to visualize datas
  - Sequencing data has a button for visualizing in IGV
  - Tabular data will prompt you to build charts
  - Protein data can be seen in a 3D viewer
- Interactive environments: Jupyter, RStudio, etc

Sharing data

- Share everything you do in Galaxy – histories, workflows, and visualizations
  - Directly using a Galaxy account's email addresses on the same instance
  - Using a web link, with anyone who knows the link
  - Using a web link and publishing it to make it accessible to everyone from the *Shared Data* menu

Training

https://usegalaxy.org/training-material/



Galaxy Training!

Welcome to Galaxy Training!

Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	10
Assembly	9
Climate	4
Computational chemistry	6
Ecology	6
Epigenetics	6
Genome Annotation	5
Imaging	4
Metabolomics	4
Metagenomics	7
Proteomics	21
Sequence analysis	2
Statistics and machine learning	13
Transcriptomics	29
Variant Analysis	8
Visualisation	2

Galaxy Tips & Tricks

Topic	Tutorials
Using Galaxy and Managing your Data	17

Galaxy for Developers and Admins

Topic	Tutorials
Galaxy Server administration	37
Development in Galaxy	13

How to contribute?

First off, thanks for taking the time to contribute!

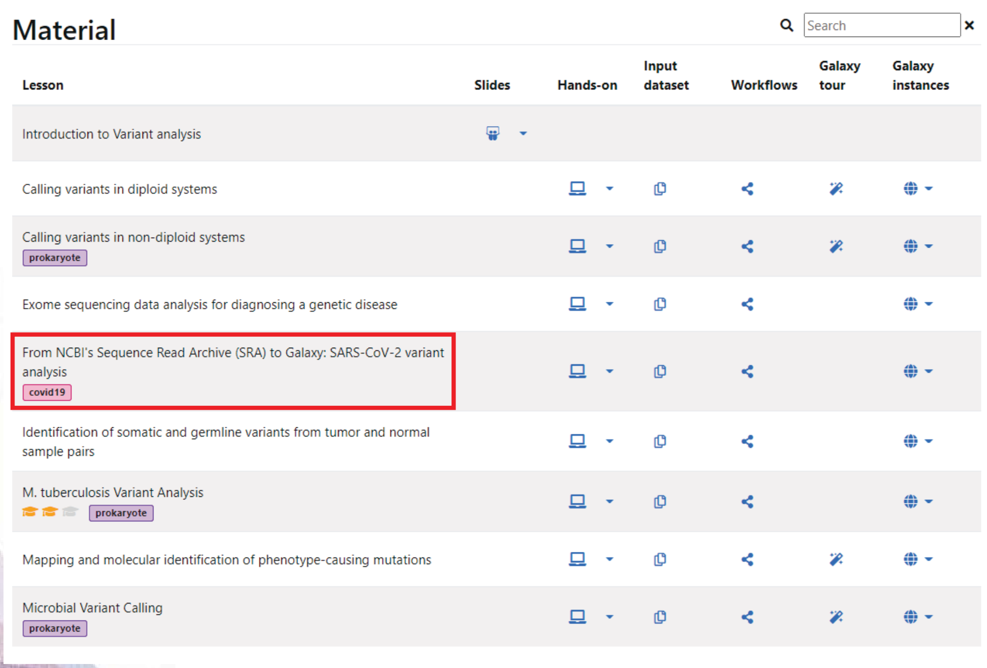
You can report mistakes or errors, create more contents, etc. Whatever is your background, there is probably a way to do it: via the GitHub website, via command-line. If you feel it is too much, you can even write it with any text editor and contact us: we will work together to integrate it.

To get you started, check our dedicated tutorials or our Frequently Asked Questions

Galaxy for Contributors and Instructors

Topic	Tutorials
Contributing to the Galaxy Training Material	11
Teaching and Hosting Galaxy training	6

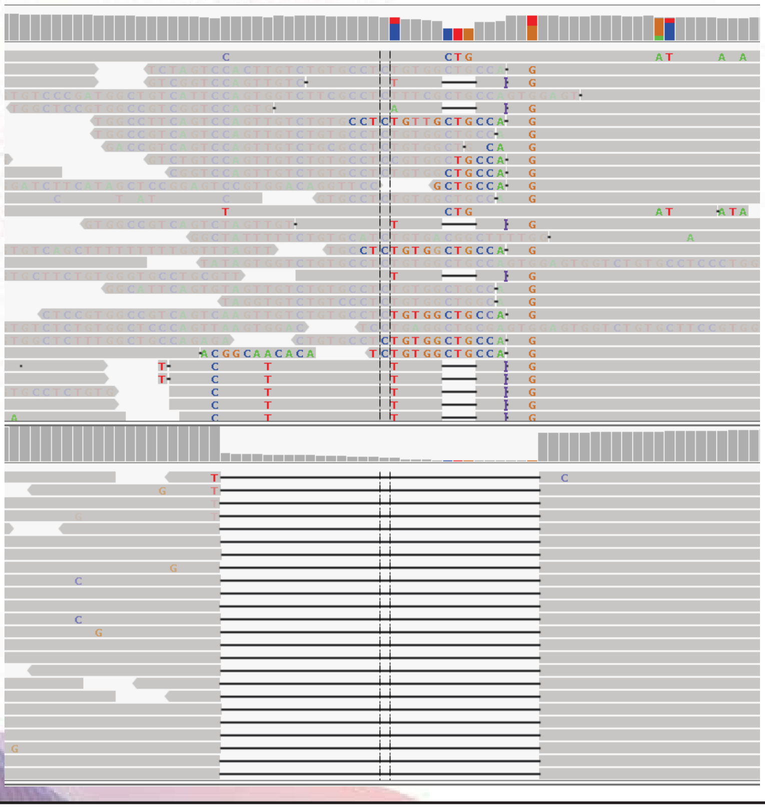
SARS-CoV2 실습



Material

Lesson	Slides	Hands-on	Input dataset	Workflows	Galaxy tour	Galaxy instances
Introduction to Variant analysis	✓ ▾					
Calling variants in diploid systems		✓ ▾	✓	✓	✓	✓ ▾
Calling variants in non-diploid systems prokaryote		✓ ▾	✓	✓	✓	✓ ▾
Exome sequencing data analysis for diagnosing a genetic disease		✓ ▾	✓	✓		✓ ▾
From NCBI's Sequence Read Archive (SRA) to Galaxy: SARS-CoV-2 variant analysis covid19		✓ ▾	✓	✓		✓ ▾
Identification of somatic and germline variants from tumor and normal sample pairs		✓ ▾	✓	✓		✓ ▾
M. tuberculosis Variant Analysis 🎓🎓 prokaryote		✓ ▾	✓	✓		✓ ▾
Mapping and molecular identification of phenotype-causing mutations		✓ ▾	✓	✓	✓	✓ ▾
Microbial Variant Calling prokaryote		✓ ▾	✓	✓	✓	✓ ▾

INDEL realignment



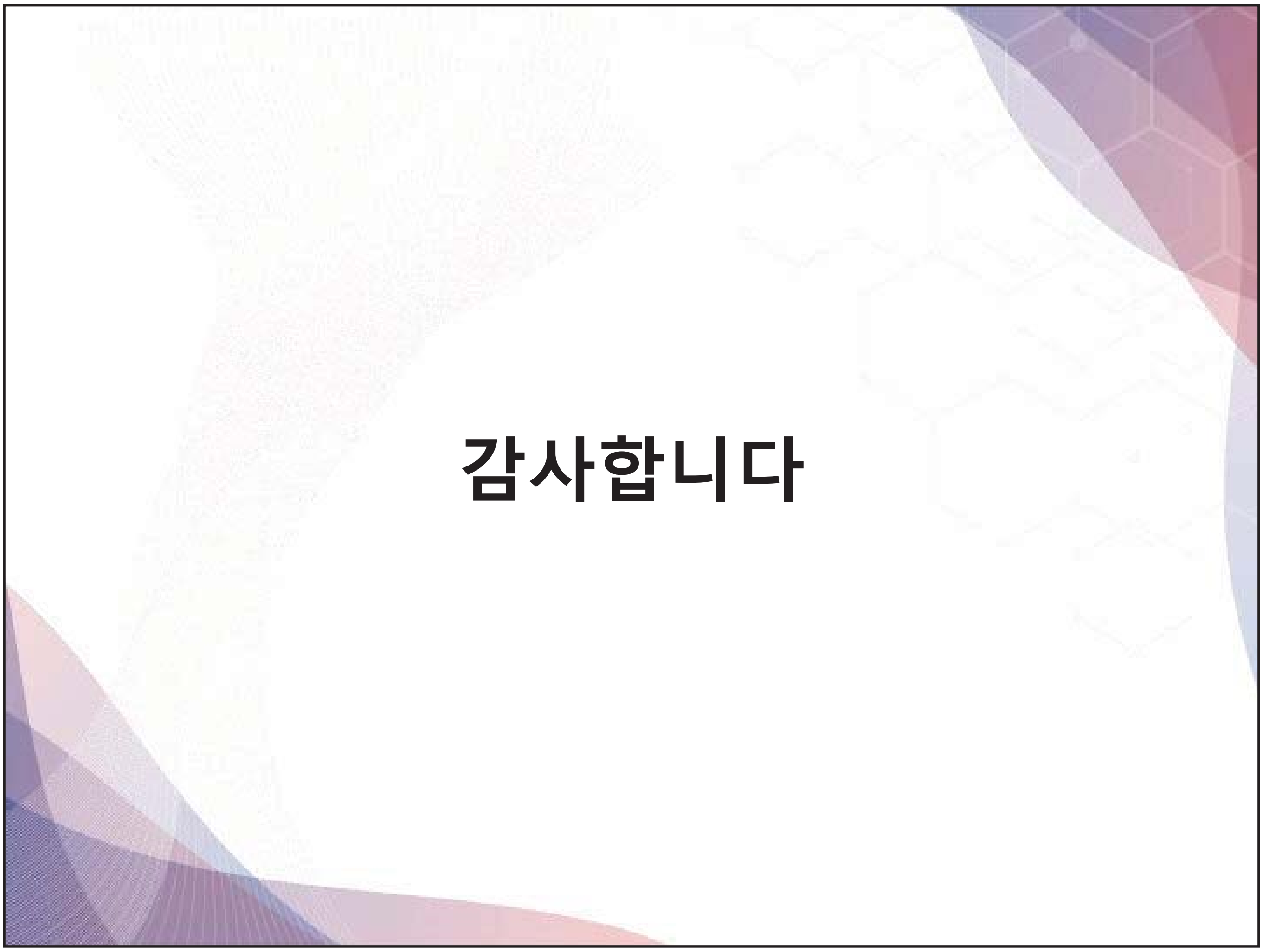
Final Alignment file visualization with IGV

- https://drive.google.com/drive/folders/1I9b5gaKcFzwk4ArJWr05qc-JRB1i59w-?usp=sharing

SARS-Cov-2 mutation 역사와 이동

- https://observablehq.com/@spond/distribution-of-sars-cov-2-sequences-that-have-a-particular

- https://nextstrain.org/ncov/global?c=gt-nuc_28960&m=div



감사합니다