유전체 변이 분석에 대한 이해 (실습)

초고성능컴퓨터 활용 교육

(주)바이오투 생물정보팀

전임연구원 이보미

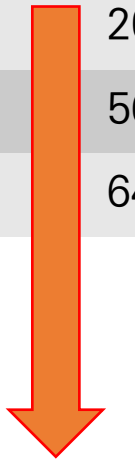
목 차

I. Dataset and Softwares

Dataset

- WGS Dataset from BioProject: PRJNA795286
- This dataset has Illumina short reads (HiSeq X Ten) of *Oryza sativa*.

Sample	SRR-id	Total bases (bp)	Read Length (bp)	Avg. length (bp)	Organism
P1	SRR17493738	7,661,857,420	25,398,995	150.83	*Oryza sativa*
P2	SRR17493737	6,136,486,300	20,336,095	150.88	*Oryza sativa*
Mix_S	SRR17493735	16,993,464,319	56,323,130	150.86	*Oryza sativa*
Mix_R	SRR17493736	19,584,672,393	64,908,654	150.86	*Oryza sativa*



Random selection 200,000 reads

Reference Genome

- 출처: Os-Nipponbare-Reference-IRGSP-1.0 (RAP-DB)

No. of chr	Total length of chr (bp)	GC (%)	No. of coding genes
chr01	43,270,923	43.77	5,981
chr02	35,937,250	43.33	4,801
chr03	36,413,819	43.69	5,244
chr04	35,502,694	44.25	3,775
chr05	29,958,434	43.95	3,412
chr06	31,248,787	43.61	3,506
chr07	29,697,621	43.50	3,204
chr08	28,443,022	43.38	2,896
chr09	23,012,720	43.53	2,356
chr10	23,207,287	43.58	2,302
chr11	29,021,106	42.91	2,587
chr12	27,531,856	43.01	2,389
Total	**373,245,519**	**522.49**	**42,453**

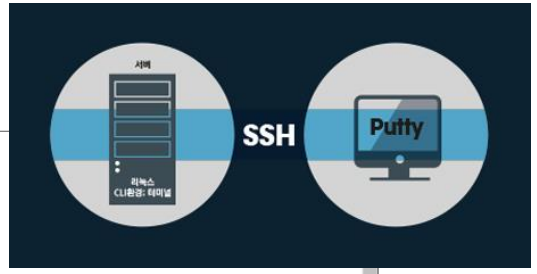
예제 데이터 파일

- Sequencing Data(FASTQ) **SRR17493738_ransel_1.fastq.gz**
  SRR17493738_ransel_2.fastq.gz
- 표준유전체 서열 (FASTA) **reference.fa**

Softwares

1. **PuTTy** - a free SSH and telnet client for Windows
2. **Trimmomatic (version 0.39)** - A flexible read trimming tool for Illumina NGS data
3. **BWA-0.7.17 (Burrows-Wheeler Aligner)**
   BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.
4. **SAMtools-1.17** - Provides various utilities for manipulating alignments in the SAM/BAM format.
5. **GATK version 4.4.0.0** - A genomic analysis toolkit focused on variant discovery.
6. **SnpEff (version 5.1)** - Genetic variant annotation and functional effect prediction toolbox.
7. **BCFtools-1.17** - BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF).
8. **MEGA 11**
   Sophisticated and user-friendly software suite for analyzing DNA and protein sequence data from species and populations.
9. **기타 프로그램: Python, Java 17**
10. **Perl/Python script**

프로그램 설치 – PuTTY



PuTTY : 터미널 프로그램

- Download PuTTY

Download PuTTY: latest release (0.78)

Home | FAQ | Feedback | Licence | Updates | Mirrors | Keys | Links | Team
Download: **Stable** · Pre-release · Snapshot | Docs | Changes | Wishlist

This page contains download links for the latest released version of PuTTY. Currently this is 0.78, released on 2022-10-29.

When new releases come out, this page will update to contain the latest, so this is a good page to bookmark or link to. Alternatively, here is a permanent link to the 0.78 release.

Release versions of PuTTY are versions we think are reasonably likely to work well. However, they are often not the most up-to-date version of the code available. If you have a problem with this release, then it might be worth trying out the pre-release builds of 0.79, or the development snapshots, to see if the problem has already been fixed in those versions.

Package files

You probably want one of these. They include versions of all the PuTTY utilities (except the new a

Bug: this installer was built differently to other versions, in a way that causes trouble for upgrades (amc trouble, when moving between 0.78 and other versions, we recommend completely uninstalling the exis with this installer) by installing 0.78 with a special command-line invocation like:

msiexec.exe /i path¥to¥putty-64bit-0.78-installer.msi ALLUSERS=1

(Not sure whether you want the 32-bit or the 64-bit version? Read the FAQ entry.)

We also publish the latest PuTTY installers for all Windows architectures as a free-of-charge download a after we release them.

MSI ('Windows Installer')

64-bit x86:	putty-64bit-0.78-installer.msi	(signature)
64-bit Arm:	putty-arm64-0.78-installer.msi	(signature)
32-bit x86:	putty-0.78-installer.msi	(signature)

Unix source archive

.tar.gz:	putty-0.78.tar.gz	(signature)

Alternative binary files

The installer packages above will provide versions of all of these (except PuTTYtel and pterm), but you c

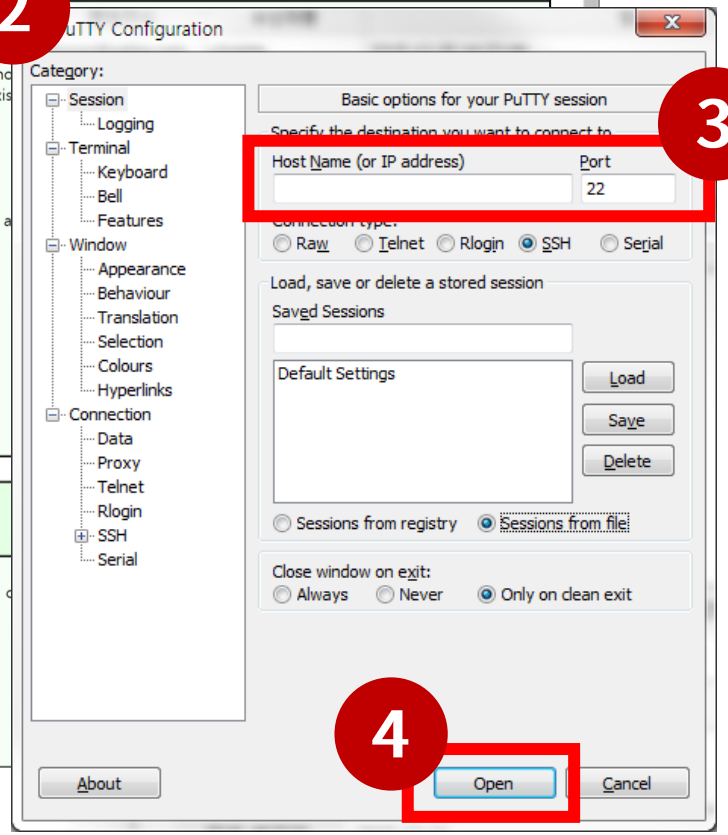
Not sure whether you want the 32-bit or the 64-bit version? Read the FAQ entry.)

lient itself)

64-bit x86:	putty.exe	(signature)
64-bit Arm:		(signature)
32-bit x86:	putty.exe	(signature)

1

2



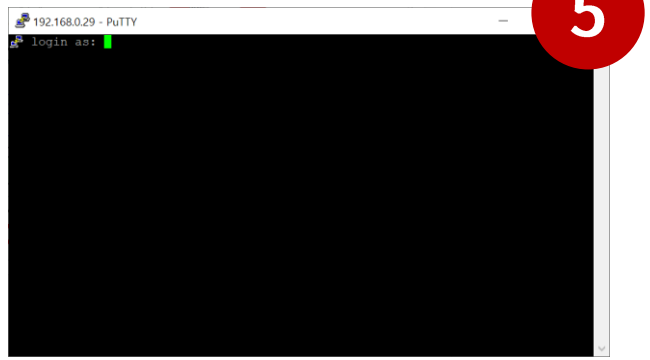
3

4

Host Name: 제공된 주소(IP)

Port: 포트 번호

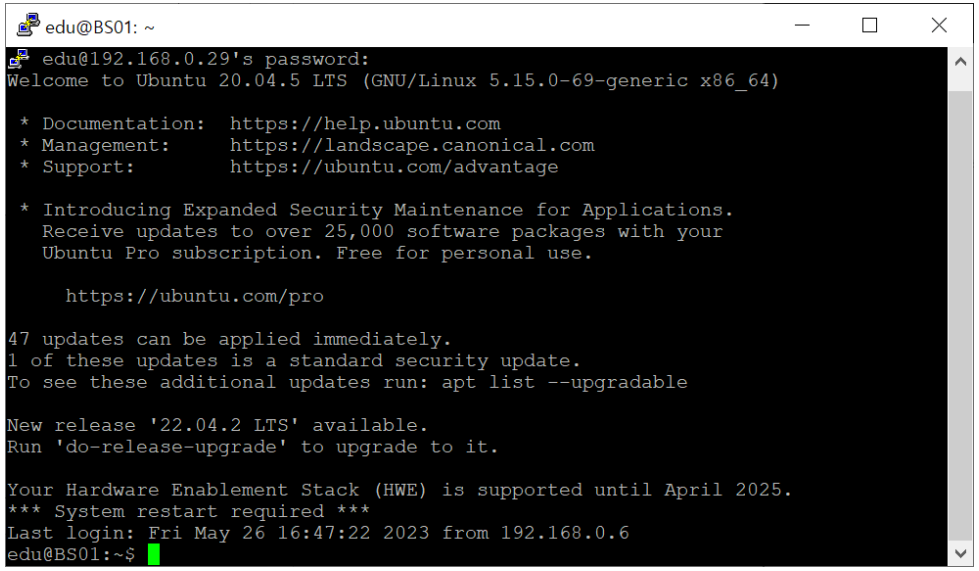
5



- ID : 계정 (엔터)
- PW : 비밀번호(엔터)

분석 디렉토리 (홈 디렉토리)로 이동

1. 터미널 접속 (로그인)
   - User : 계정
   - Passwd : 비밀번호



2. 분석 디렉토리로 이동

```
## 홈 디렉토리로 이동
cd /home/계정
cd ~

## 현재 위치 확인
pwd
```

리눅스 기본 명령어 요약

1. ls – 현재 위치의 파일 목록 조회
2. cd – 디렉토리 이동
3. pwd – 현재 위치의 절대경로 출력
4. mkdir – 디렉토리 생성
5. cp – 파일 복사
6. mv – 파일/디렉토리 이동 및 이름 변경
7. rm – 파일 삭제
8. rmdir – 디렉토리 삭제
9. less, more – 파일 내용을 페이지 단위로 화면에 출력
10. head, tail – 파일의 앞 또는 뒤 10행을 화면에 출력

리눅스 기본 명령어 옵션 설명

1. ls (List segments) : 현재 위치의 파일 목록 조회
    - ls -l : 파일의 상세정보
    - ls -a : 숨김 파일 표시

2. cd (Change directory) :디렉토리 이동
    - cd [디렉토리 경로] : 이동하려는 디렉토리로 이동
    - cd . : 현재 디렉토리
    - cd .. : 상위 디렉토리로 이동
    - cd ~ : 홈 디렉토리로 이동
    - cd / : 최상위 디렉토리로 이동

3. pwd (Print Working Directory) : 현재 위치의 절대경로 출력

4. mkdir (Make directory) : 디렉토리 생성
    - mkdir [디렉토리 이름] : [디렉토리 이름] 이라는 디렉토리 생성

리눅스 기본 명령어 옵션 설명

5. cp (Copy) : 파일 복사
   - cp [파일1] [파일2] : [파일1]을 [파일2] 으로 복사
   - cp -r [디렉토리1] [디렉토리2] : 디렉토리 복사. 폴더 안의 모든 하위 경로와 파일들이 함께 복사됨.

6. mv (Move) : 파일/디렉토리 이동 및 이름 변경
   - mv [파일1] [파일2] : [파일1]을 [파일2] 으로 이름 변경
   - mv [파일1] [디렉토리 경로] : [파일1]을 [디렉토리 경로] 하위로 위치 이동
   - mv [디렉토리1] [디렉토리2] : [디렉토리1]을 [디렉토리2] 으로 이름 변경

7. rm (Remove) : 파일 삭제
   - rm [파일] : 파일 삭제
   - rm -f [파일] : 파일 강제 삭제
   - rm –r [디렉토리] : 디렉토리 삭제 (디렉토리는 –r 옵션 없이 삭제 불가)

리눅스 기본 명령어 옵션 설명

8. rmdir (Remove Directory) : 디렉토리 삭제
   - rmdir [디렉토리] : 디렉토리 삭제

9. less, more – 파일 내용을 페이지 단위로 화면에 출력
   - less [파일] : 파일의 첫 행부터 화면에 출력
   - less +10 [파일] : 파일의 10행부터 화면에 출력

     ※ space bar: 다음 페이지, b: 이전 페이지, q: 종료, enter: 줄 단위로 이동
     ※ less 명령어는 추가로 화살표 키, page up과 page down 키 사용 가능

10. head, tail – 파일의 앞 또는 뒤 10행을 화면에 출력
    - head [파일] : 파일의 앞 10행부터 화면에 출력
    - head -50 [파일] : 파일의 앞 50행부터 화면에 출력

명령어 실습

- 홈 디렉토리 이동

```
cd ~
cd /home/계정
```

- 현재 위치의 파일 목록 조회

```
ls -l
```

- 현재 위치의 절대경로 출력

```
pwd
```

- 디렉토리 생성

```
mkdir 1.trimmed
mkdir /home/계정/1.trimmed
```

- 디렉토리 이동

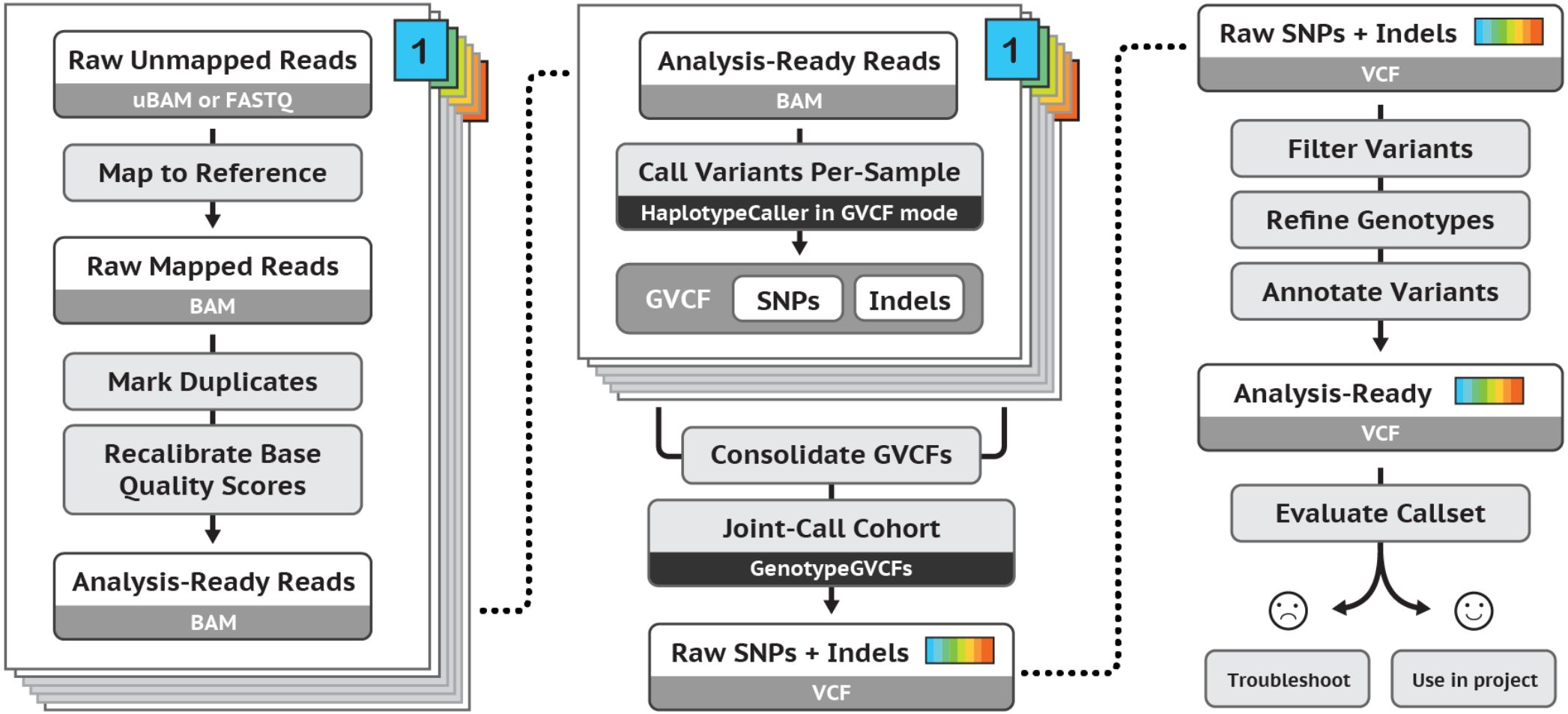
```
cd 1.trimmed
cd /home/계정/1.trimmed
```

- 텍스트 파일 출력

```
less reference.fa
```

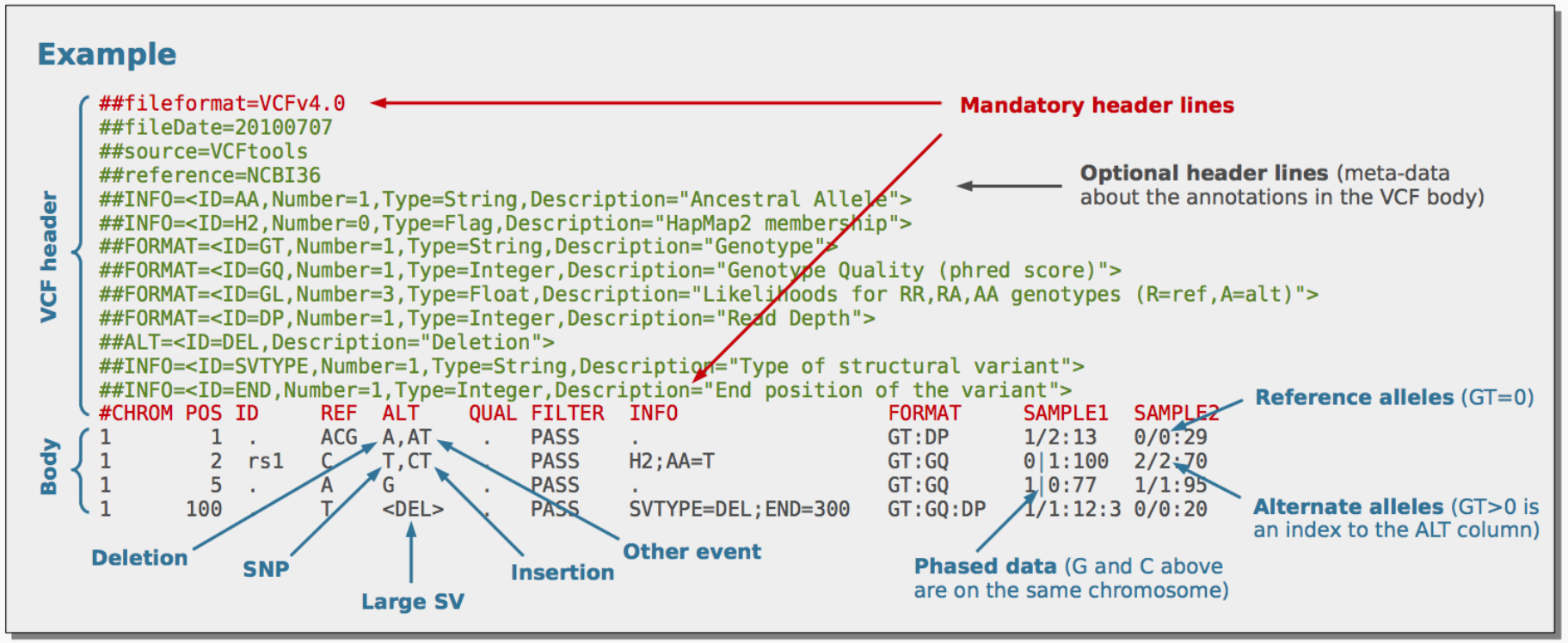
II. Variant discovery with GATK

변이분석 실습 - Workflow



변이분석 실습 - VCF

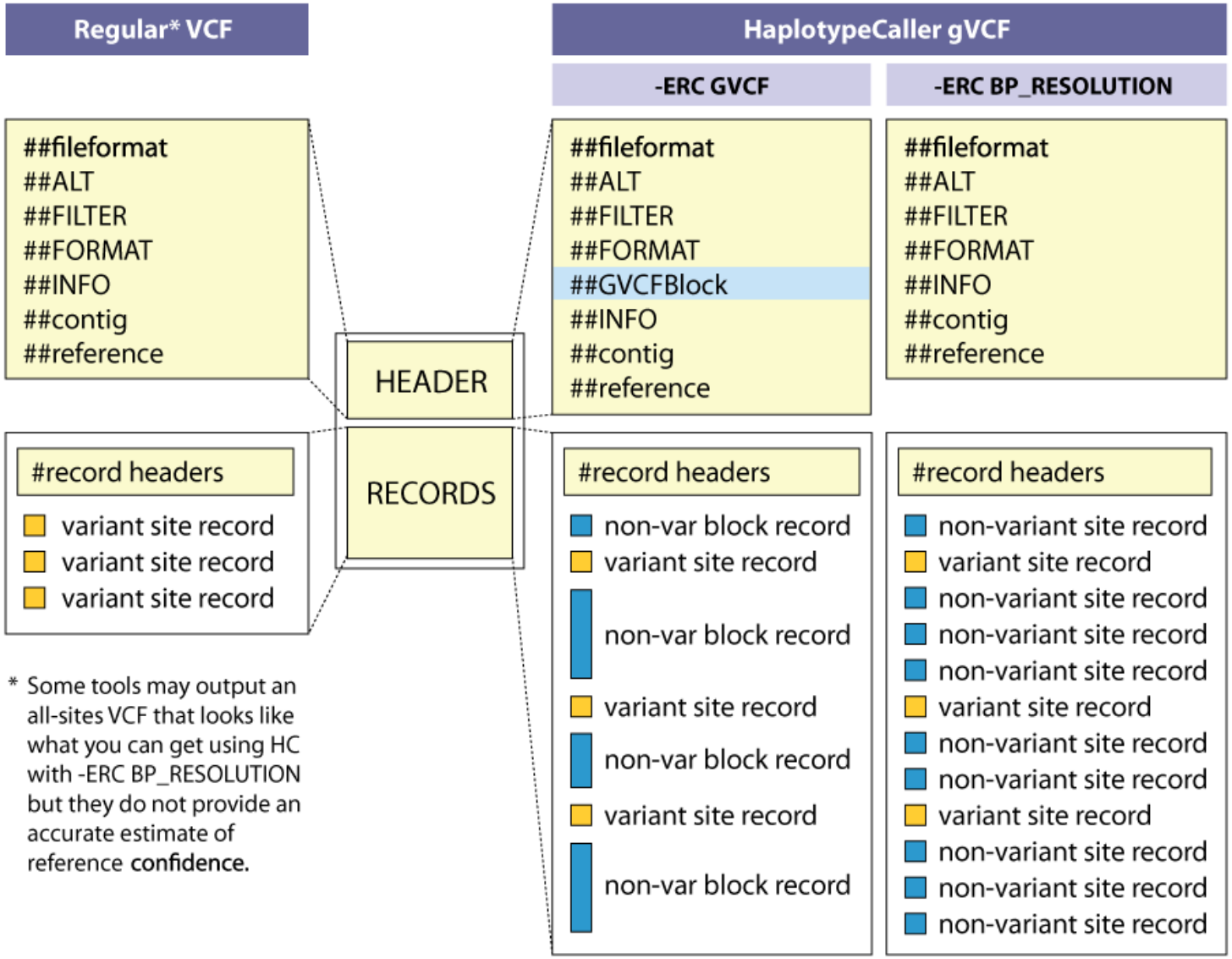
VCF (Variant Call Format)



- VCF is a text file format (tab-delimited)
- It contains a header line, and then data lines each containing information about a position in the genome.
- The format also has the ability to contain genotype information on samples for each position.

- https://davetang.github.io/learning_vcf_file/

변이분석 실습 - VCF and gVCF



- https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-faqs/What_is_a_GVCF_and_how_is_it_different_from_a_%27regular%27_VCF%3F.md

변이분석 실습 - 1) Pre-processing

Trimmomatic 프로그램 옵션

- Phred33
- Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)
- Remove leading low quality or N bases (LEADING:3)
- Remove trailing low quality or N bases (TRAILING:3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 36 bases long (MINLEN:36)

Trimmomatic 수행

```
trimmomatic PE -threads 1 -phred33 SRR17493738_ransel_1.fastq.gz
SRR17493738_ransel_2.fastq.gz P1_paired1.fq P1_paired1_un.fq P1_paired2.fq
P1_paired2_un.fq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

변이분석 실습 - 1) Pre-processing

- Trimmomatic 수행 결과

```
root@d57c9b675d47:/home/1.works/1.trimmed# java -jar /home/0.tools/Trimmomatic-0.39/trimmomatic-0.39.jar PE -threads 10 -phred33 SRR17493738_ransel_1.
fastq.gz SRR17493738_ransel_2.fastq.gz P1_paired1.fq P1_paired1_un.fq P1_paired2.fq P1_paired2_un.fq ILLUMINACLIP:/home/0.tools/Trimmomatic-0.39/adapt
ers/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
TrimmomaticPE: Started with arguments:
 -threads 10 -phred33 SRR17493738_ransel_1.fastq.gz SRR17493738_ransel_2.fastq.gz P1_paired1.fq P1_paired1_un.fq P1_paired2.fq P1_paired2_un.fq ILLUMI
NACLIP:/home/0.tools/Trimmomatic-0.39/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
Using PrefixPair: 'TACACTCTTTCCCTACACGACGCTCTTCCGATCT' and 'GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT'
ILLUMINACLIP: Using 1 prefix pairs, 0 forward/reverse sequences, 0 forward only sequences, 0 reverse only sequences
Input Read Pairs: 200000 Both Surviving: 197635 (98.82%) Forward Only Surviving: 1723 (0.86%) Reverse Only Surviving: 517 (0.26%) Dropped: 125 (0.06%)
TrimmomaticPE: Completed successfully
```

결과 파일 확인

```
ls -l

-rw-r--r-- 1 root root 75682990 Jul  4 15:38 P1_paired1.fq
-rw-r--r-- 1 root root   571082 Jul  4 15:38 P1_paired1_un.fq
-rw-r--r-- 1 root root 75474326 Jul  4 15:38 P1_paired2.fq
-rw-r--r-- 1 root root   172898 Jul  4 15:38 P1_paired2_un.fq
```

변이분석 실습 - 1) Pre-processing

- Trimmomatic 수행 결과

데이터 전처리 수행 전 FASTQ

```
@SRR17493738.913322 913322 length=151
CCGTGCACCTGGAGTTGGCTGCAGCTCCTCAGCTTAAGCTCTCAAGGTTTCCGTTTCCTCTTCTCTTCTCTTCTCTCCTCTTCTATATGCGCC
TGCCTCACCCTACCTACTATTGATTGTTTGTCTCGTGCAGGTCTGGCTTCTTGGAGTC
+SRR17493738.913322 913322 length=151
FFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,,F,::,,,FF,:,F:FFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF:FFFFF
```

데이터 전처리 수행 후 FASTQ

```
@SRR17493738.913322 913322 length=151
CCGTGCACCTGGAGTTGGCTGCAGCTCCTCAGCTTAAGCTCTCAAGGTTTCCGTTTCCTCTT
+SRR17493738.913322 913322 length=151
FFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,,F,::
```

변이분석 실습 - 2) Reference Indexing

Reference index 작성 1 – BWA index

```
bwa index reference.fa
```

```
root@d57c9b675d47:/home/1.works# bwa index reference.fa
[bwa_index] Pack FASTA... 0.92 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=231243984, availableWord=28271124
[BWTIncConstructFromPacked] 10 iterations done. 46634416 characters processed.
[BWTIncConstructFromPacked] 20 iterations done. 86153072 characters processed.
[BWTIncConstructFromPacked] 30 iterations done. 121273312 characters processed.
[BWTIncConstructFromPacked] 40 iterations done. 152484240 characters processed.
[BWTIncConstructFromPacked] 50 iterations done. 180220560 characters processed.
[BWTIncConstructFromPacked] 60 iterations done. 204868656 characters processed.
[BWTIncConstructFromPacked] 70 iterations done. 226771904 characters processed.
[bwt_gen] Finished constructing BWT in 73 iterations.
[bwa_index] 46.73 seconds elapse.
[bwa_index] Update BWT... 1.09 sec
[bwa_index] Pack forward-only FASTA... 0.58 sec
[bwa_index] Construct SA from BWT and Occ... 25.94 sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa index reference.fa
[main] Real time: 75.588 sec; CPU: 75.255 sec
```

결과 파일 확인

```
ls -l
```

```
-rw-r--r-- 1 root root      9076 Jul  4 16:17 reference.fa.amb
-rw-r--r-- 1 root root       226 Jul  4 16:17 reference.fa.ann
-rw-r--r-- 1 root root 212332012 Jul  4 16:17 reference.fa.bwt
-rw-r--r-- 1 root root  53082978 Jul  4 16:17 reference.fa.pac
-rw-r--r-- 1 root root 106166008 Jul  4 16:18 reference.fa.sa
```

변이분석 실습 - 2) Reference Indexing

```
## Reference index 작성 2 – SAMtools index

samtools faidx reference.fa



## 결과 파일 확인

ls -l

-rw-r--r-- 1 root root       176 Jul  4 16:15 reference.fa.fai
```

변이분석 실습 - 2) Reference Indexing

Reference index 작성 3 – GATK index

```
gatk CreateSequenceDictionary -R reference.fa
```

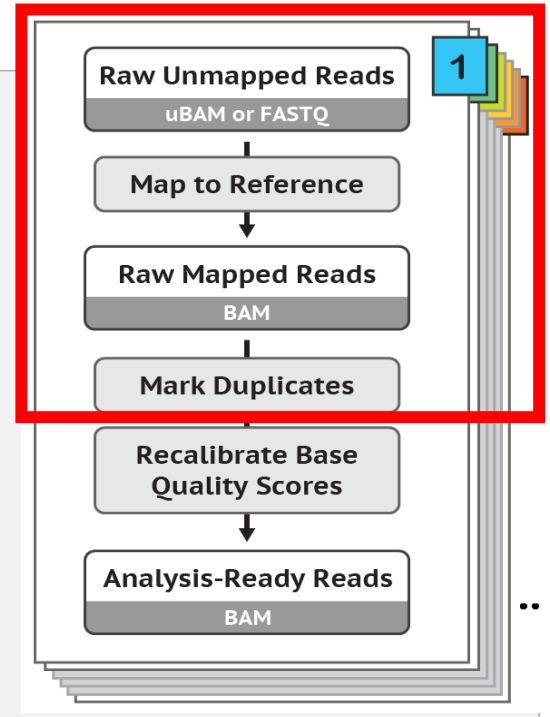
```
Using GATK jar /home/0.tools/gatk-4.4.0.0/gatk-package-4.4.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/0.tools/gatk-4.4.0.0/gatk-package-4.4.0.0-local.jar CreateSequenceDictionary -R reference.fa
INFO    2023-07-04 13:07:54     CreateSequenceDictionary        Output dictionary will be written in reference.dict
13:07:54.641 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/0.tools/gatk-4.4.0.0/gatk-package-4.4.0.0-local.jar!/com/in
tel/gkl/native/libgkl_compression.so
[Tue Jul 04 13:07:54 KST 2023] CreateSequenceDictionary --REFERENCE reference.fa --TRUNCATE_NAMES_AT_WHITESPACE true --NUM_SEQUENCES 2147483647 --VERB
OSITY INFO --QUIET false --VALIDATION_STRINGENCY STRICT --COMPRESSION_LEVEL 2 --MAX_RECORDS_IN_RAM 500000 --CREATE_INDEX false --CREATE_MD5_FILE false
 --help false --version false --showHidden false --USE_JDK_DEFLATER false --USE_JDK_INFLATER false
[Tue Jul 04 13:07:54 KST 2023] Executing as root@d57c9b675d47 on Linux 5.15.0-69-generic amd64; OpenJDK 64-Bit Server VM 17.0.7+7-Ubuntu-0ubuntu120.04
; Deflater: Intel; Inflater: Intel; Provider GCS is available; Picard version: Version:4.4.0.0
[Tue Jul 04 13:07:55 KST 2023] picard.sam.CreateSequenceDictionary done. Elapsed time: 0.01 minutes.
Runtime.totalMemory()=285212672
```

결과 파일 확인

```
ls -l
```

```
-rw-r--r-- 1 root root       587 Jul  4 16:16 reference.dict
```

변이분석 실습 - 3) Read Mapping



BWA 수행 (FASTQ to SAM)

```
bwa mem -t 1 -k 19 -R "@RG\tID:P1.1\tLB:P1.fq\tSM:P1\tPL:ILLUMINA"
reference.fa P1_paired1.fq P1_paired2.fq -o P1_paired.sam
```

SAM to BAM

```
samtools view -b -t reference.fa.fai -o P1_paired.bam P1_paired.sam
```

FIXMATE

```
gatk FixMateInformation -I P1_paired.bam -O P1_fixmate.bam -SO coordinate
--CREATE_INDEX true --VALIDATION_STRINGENCY SILENT
```

REMOVE DUPLICATES

```
gatk MarkDuplicates -I P1_fixmate.bam -O P1_rmdup.bam -M P1_metrics.txt
--REMOVE_DUPLICATES true --CREATE_INDEX true --VALIDATION_STRINGENCY SILENT
```

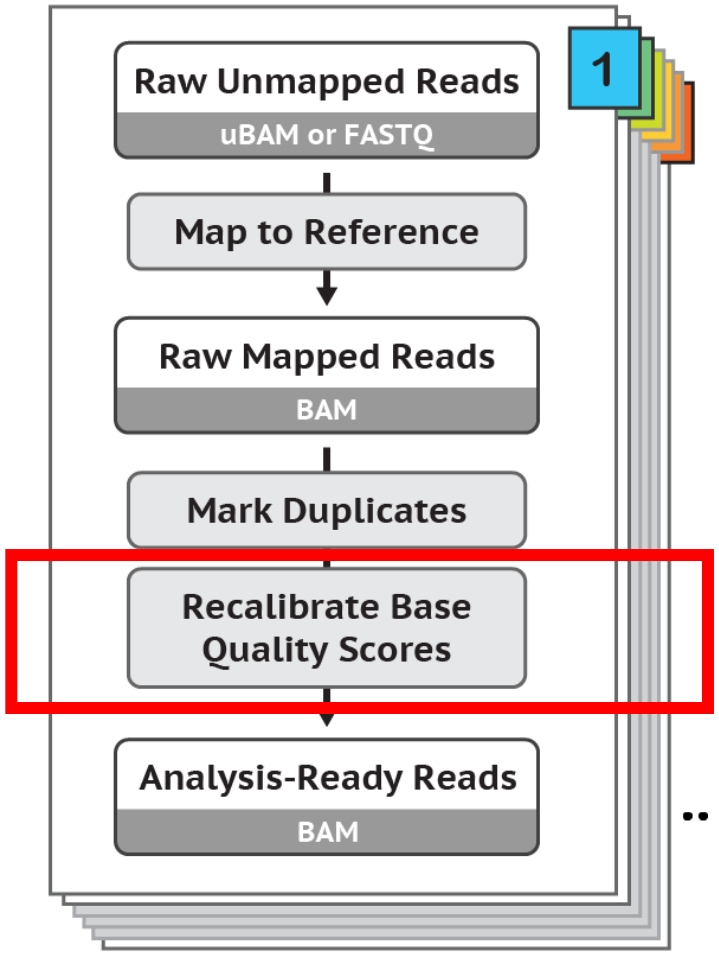
변이분석 실습 - 3) Read Mapping

BAM stats

```
samtools flagstats P1_rmdup.bam
```

396505 + 0 in total (QC-passed reads + QC-failed reads)
394230 + 0 primary
0 + 0 secondary
2275 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
392658 + 0 mapped (99.03% : N/A)
390383 + 0 primary mapped (99.02% : N/A)
394230 + 0 paired in sequencing
197115 + 0 read1
197115 + 0 read2
379678 + 0 properly paired (96.31% : N/A)
389314 + 0 with itself and mate mapped
1069 + 0 singletons (0.27% : N/A)
7752 + 0 with mate mapped to a different chr
4653 + 0 with mate mapped to a different chr (mapQ>=5)

변이분석 실습 - BQSR Workflow



- BQSR Workflow
  0. GATK HaplotypeCaller
     initial round of variant calling on unrecalibrated data
  1. GATK BaseRecalibrator
     generation of recalibration table for BQSR
  2. GATK ApplyBQSR
     actual base quality score recalibration of reads

변이분석 실습 - 4) BQSR (Base Quality Score Recalibration)

GATK HaplotypeCaller - initial round of variant calling on unrecalibrated data

```
gatk --java-options '-DGATK_STACKTRACE_ON_USER_EXCEPTION=true' HaplotypeCaller
-R reference.fa -I P1_rmdup.bam -O P1_raw_variants.vcf
```

Extract SNPs

```
gatk SelectVariants -V P1_raw_variants.vcf -select-type SNP -O P1_raw_snps.vcf
```

Filter VCF to obtain high confidence variants using Hard Filtering recommendations

Filter SNPs

```
gatk VariantFiltration -V P1_raw_snps.vcf -filter "QD < 2.0" --filter-name "QD2" -filter
"QUAL < 30.0" --filter-name "QUAL30" -filter "SOR > 3.0" --filter-name "SOR3" -filter
"FS > 60.0" --filter-name "FS60" -filter "MQ < 40.0" --filter-name "MQ40" -filter
"MQRankSum < -12.5" --filter-name "MQRankSum-12.5" -filter "ReadPosRankSum < -8.0" --
filter-name "ReadPosRankSum-8" -O P1_flt_snps.vcf
```

변이분석 실습 - 4) BQSR (Base Quality Score Recalibration)

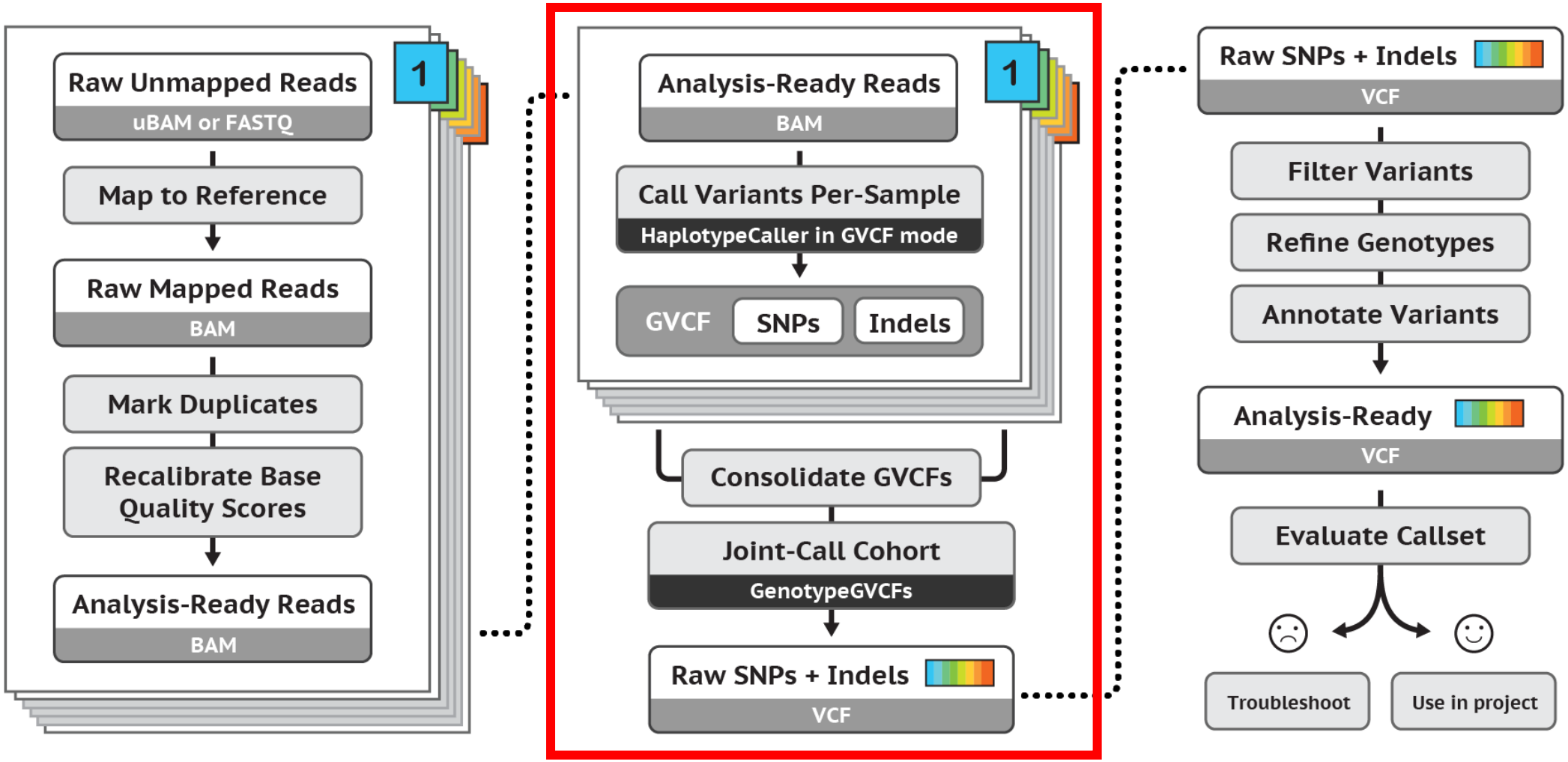
```
## Filter VCF to obtain high confidence variants using Hard Filtering recommendations
## Extract Indels
gatk SelectVariants -V P1_raw_variants.vcf -select-type INDEL -O P1_raw_indels.vcf

## Filter Indels
gatk VariantFiltration -V P1_raw_indels.vcf -filter "QD < 2.0" --filter-name "QD2" -filter
"QUAL < 30.0" --filter-name "QUAL30" -filter "FS > 200.0" --filter-name "FS200" -filter
"ReadPosRankSum < -20.0" --filter-name "ReadPosRankSum-20" -O P1_flt_indels.vcf

## Base (Quality Score) Recalibration
gatk BaseRecalibrator -R reference.fa -I P1_rmdup.bam --known-sites P1_flt_snps.vcf --
known-sites P1_flt_indels.vcf -O P1_recal.table

## ApplyBQSR
gatk ApplyBQSR -R reference.fa -I P1_rmdup.bam -bqsr P1_recal.table -O P1_recal.bam
```

변이분석 실습 - Workflow



변이분석 실습 - 5) Variant Discovery

GATK HaplotypeCaller - GVCF

```
gatk --java-options '-DGATK_STACKTRACE_ON_USER_EXCEPTION=true' HaplotypeCaller -R
reference.fa -I P1_recal.bam -O P1_variants.g.vcf -ERC GVCF
```

CombineGVCFs

```
gatk CombineGVCFs -R reference.fa -V P1_variants.g.vcf -V P2_variants.g.vcf -V
MixS_variants.g.vcf -V MixR_variants.g.vcf -O combined_raw.g.vcf
```

Perform joint genotyping on one or more samples pre-called with HaplotypeCaller

GenotypeGVCFs

```
gatk GenotypeGVCFs -R reference.fa -V combined_raw.g.vcf -O combined_raw.vcf
```

변이분석 실습 - 5) Variant Discovery

VCF 결과 확인

```
less combined_raw.vcf


#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  MixR    MixS    P1      P2
chr01   80388   .       C       T       40.18   .
AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=20.09;SOR=0.693
GT:AD:DP:GQ:PL  ./.:0,0:0:0:0,0,0
1/1:0,2:2:6:49,6,0      ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0
```

- https://davetang.github.io/learning_vcf_file/

III. Variant Annotation and Filtration

변이분석 실습 - 6) Variant Annotation

SnpEff 프로그램

Usage: snpeff [command] [options] [files]

- Available commands:
  - ann : Annotate variants
  - build : Build a SnpEff database
  - databases : Show currently available databases (from local config file)
  - download : Download a SnpEff database

Finding a database

```
java -jar /agribio/soft/snpEff/snpEff-5.1/snpEff.jar databases | grep Oryza_sativa
```

Database download

다운로드 위치: /agribio/HOME/edu_02/prepare/dataset/Oryza_sativa/

```
java -jar /agribio/soft/snpEff/snpEff-5.1/snpEff.jar download -v Oryza_sativa
```

```
-rw-r--r-- 1 root root  4173997 Jul  5 11:14 sequence.1.bin
-rw-r--r-- 1 root root  3434377 Jul  5 11:14 sequence.2.bin
-rw-r--r-- 1 root root  3698390 Jul  5 11:14 sequence.3.bin
-rw-r--r-- 1 root root  2684939 Jul  5 11:14 sequence.4.bin
                                ...
-rw-r--r-- 1 root root  2041957 Jul  5 11:14 sequence.11.bin
-rw-r--r-- 1 root root  1776653 Jul  5 11:14 sequence.12.bin
-rw-r--r-- 1 root root    37261 Jul  5 11:14 sequence.bin
-rw-r--r-- 1 root root 32210111 Jul  5 11:14 snpEffectPredictor.bin
```

변이분석 실습 - 6) Variant Annotation

- **Building a database**

1. **Step 1:** Configure a new genome in SnpEff's config file `snpEff.config`.
    a. Add genome entry to snpEff's configuration
    b. If the genome uses a non-standard codon table: Add codon table parameter
2. **Step 2:** Build using gene annotations and reference sequences
    a. Option 1: Building a database from GTF files (**recommended for large genomes**)
    b. Option 2: Building a database from GenBank files (**recommended for small genomes**)
    c. Option 3: Building a database from GFF files
    d. Option 4: Building a database from RefSeq table from UCSC
3. **Step 3:** Checking the database: SnpEff will check the database by comparing predicted protein sequences and CDS sequences with ones provided by the user.
    a. Checking CDS sequences
    b. Checking Protein sequences

- https://pcingola.github.io/SnpEff/se_build_db/#step-2-build-using-gene-annotations-and-reference-sequences

변이분석 실습 - 6) Variant Annotation

VCF annotation

```
java -jar /agribio/soft/snpEff/snpEff-5.1/snpEff.jar ann -dataDir
/agribio/HOME/edu_02/prepare/dataset Oryza_sativa combined_raw.vcf > combined_raw.ann.vcf
```

VCF format

```
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele |
Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID |
Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos /
CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO' ">
```

Impact	Meaning	Example
HIGH	The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay.	stop_gained, frameshift_variant
MODERATE	A non-disruptive variant that might change protein effectiveness.	missense_variant (non-syn), inframe_deletion
LOW	Assumed to be mostly harmless or unlikely to change protein behavior.	synonymous_variant
MODIFIER	Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact.	exon_variant, downstream_gene_variant

변이분석 실습 - 6) Variant Annotation

- Annotation 수행 결과

Annotation 하기 전 VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT MixR MixS P1 P2

chr01 1328686 . T C 38.10 .
AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=29.04;SOR=1.609
GT:AD:DP:GQ:PGT:PID:PL:PS ./.:0,0:0:0:.:.:0,0,0 ./.:0,0:0:0:.:.:0,0,0
1|1:0,1:1:3:1|1:1328662_T_C:45,3,0:1328662 ./.:0,0:0:0:.:.:0,0,0

Annotation 결과 VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT MixR MixS P1 P2

chr01 1328686 . T C 38.1 .
AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=29.04;SOR=1.609;ANN=C|stop_lost|HIGH|Os01g0123500|Os01g0123500|transcript|Os01t0123500-00|protein_coding|2/2|c.430T>C|p.Ter144Argext*?|430/935|430/432|144/143||
GT:AD:DP:GQ:PGT:PID:PL:PS ./.:0,0:0:0:.:.:0,0,0 ./.:0,0:0:0:.:.:0,0,0
1|1:0,1:1:3:1|1:1328662_T_C:45,3,0:1328662 ./.:0,0:0:0:.:.:0,0,0

변이분석 실습 - 7) Variant Filtration

- **Hard Filtering recommendations**

Hard Filtering (SNPs)	Name	Summary
QUAL < 30.0	Quality	• Base call quality
QD < 2.0	QualByDepth	• Variant confidence normalized by unfiltered depth of variant samples (QD) • Avoid inflation caused when there is deep coverage
SOR > 3.0	StrandOddsRatio	• Strand bias estimated by the symmetric odds ratio test (SOR)
FS > 60.0	FisherStrand	• Strand bias estimated using Fisher's exact test (FS)
MQ < 40.0	RMSMappingQuality	• Root mean square of the mapping quality of reads across all samples (MQ)
MQRankSum < −12.5	MappingQualityRankSumTest	• Rank sum test for mapping qualities of REF versus ALT reads (MQRankSum)
ReadPosRankSum < −8.0	ReadPosRankSumTest	• Rank sum test for relative positioning of REF versus ALT alleles within reads (ReadPosRankSum)

변이분석 실습 - 7) Variant Filtration

```
## Extract SNPs
gatk SelectVariants -V combined_raw.ann.vcf -select-type SNP -O combined_raw.ann_snps.vcf

## Filter SNPs - using Hard Filtering recommendations
gatk VariantFiltration -V combined_raw.ann_snps.vcf -filter "QD < 2.0" --filter-name "QD2"
-filter "QUAL < 30.0" --filter-name "QUAL30" -filter "SOR > 3.0" --filter-name "SOR3" -
filter "FS > 60.0" --filter-name "FS60" -filter "MQ < 40.0" --filter-name "MQ40" -filter
"MQRankSum < -12.5" --filter-name "MQRankSum-12.5" -filter "ReadPosRankSum < -8.0" --
filter-name "ReadPosRankSum-8" -O combined_flt_snps.vcf

## Extract Indels
gatk SelectVariants -V combined_raw.ann.vcf -select-type INDEL -O
combined_raw.ann_indels.vcf

## Filter Indels - using Hard Filtering recommendations
gatk VariantFiltration -V combined_raw.ann_indels.vcf -filter "QD < 2.0" --filter-name
"QD2" -filter "QUAL < 30.0" --filter-name "QUAL30" -filter "FS > 200.0" --filter-name
"FS200" -filter "ReadPosRankSum < -20.0" --filter-name "ReadPosRankSum-20" -O
combined_flt_indels.vcf
```

변이분석 실습 - 7) Variant Filtration

GATK VariantFiltration 결과 예시

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  MixR    MixS    P1      P2
chr01   80388   .       C       T       40.18   PASS
AC=2;AF=1.00;AN=2;ANN=T|upstream_gene_variant|MODIFIER|OsDjC1|Os01g0101700|transcript|Os01t01017
00-
00|protein_coding|;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;QD=20.09;SOR=0.693
GT:AD:DP:GQ:PL  ./.:0,0:0:0:0,0,0
1/1:0,2:2:6:49,6,0      ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0

chr01   133207  .       A       G       37.07   MQ40
AC=2;AF=0.500;AN=4;ANN=G|upstream_gene_variant|MODIFIER|OsTLP27|Os01g0102300|transcript|Os01t010
2300-
01|protein_coding|;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=2;MLEAF=0.500;MQ=29.00;QD=25.36;SOR=1.60
9       GT:AD:DP:GQ:PGT:PID:PL:PS       ./.:0,0:0:0:.:.:0,0,0
1|1:0,1:1:3:1|1:133207_A_G:45,3,0:133207        0/0:1,0:1:3:.:.:0,3,23  ./.:0,0:0:0:.:.:0,0,0

chr01   159218  .       G       A       60.04   PASS
AC=2;AF=0.500;AN=4;ANN=A|downstream_gene_variant|MODIFIER|CSB|Os01g0102800|transcript|Os01t01028
00-
01|protein_coding|DP=4;ExcessHet=0.0000;FS=0.000;MLEAC=2;MLEAF=0.500;MQ=40.00;QD=30.02;SOR=0.693
GT:AD:DP:GQ:PGT:PID:PL:PS       ./.:0,0:0:0:.:.:0,0,0   0/0:1,0:1:3:.:.:0,3,19
1|1:0,2:2:6:1|1:159218_G_A:70,6,0:159218        ./.:1,0:1:0:.:.:0,0,0
```

변이분석 실습 - 7) Variant Filtration

BCFtools 프로그램

- utilities for variant calling and manipulating VCFs and BCFs.

```
## Extract only PASS SNPs
bcftools view -f PASS combined_flt_snps.vcf > combined_flt_snps.PASS.vcf

## Additional filtering
## FORMAT:DP>5
bcftools view -i 'FMT/DP>5' combined_flt_snps.PASS.vcf > combined_flt_snps.PASS.DP5.vcf

## QUAL>40
bcftools view -i 'QUAL>40' combined_flt_snps.PASS.vcf > combined_flt_snps.PASS.qual.vcf
```

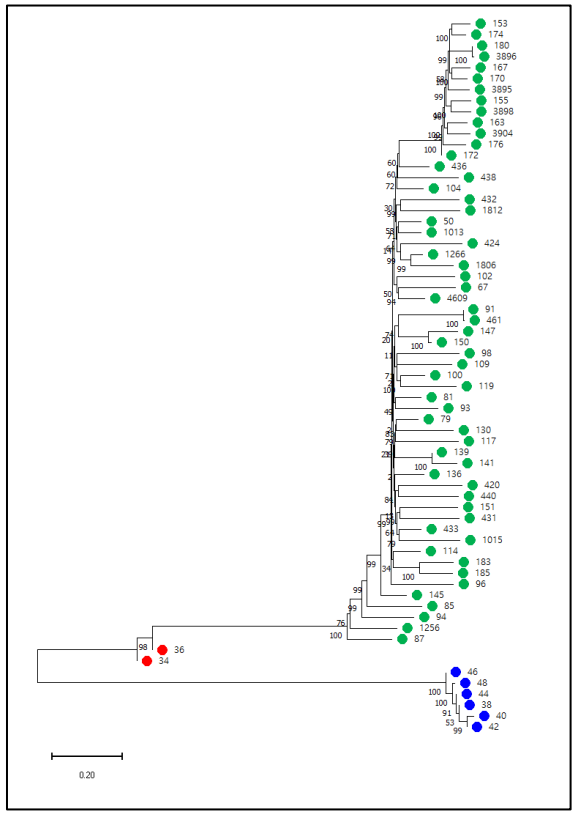
IV. Application of SNPs - 계통수 작성

계통수 분석 (Phylogenetic tree)

- **계통수 분석 :**
  생물의 진화 관계를 밝히기 위해 종 또는 그룹 간의 계통적 관계를 탐구하는 분석 방법

- **프로그램** :
  MEGA, IQ-TREE, PHYLIP, BEAST, PAUP, MrBayes, RAxML 등…

- **분석 방법** :
  - UPGMA (Unweighted Pair Group Method with Arithmetic Mean)
  - NJ (Neighbour-Joining)
  - ML (Maximum Likelihood) 등…



계통수 분석 - Input file 작성 (VCF to FASTA)

VCF (Variant Call Format)

```
##fileformat=VCFv4.2
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##contig=<ID=chr01,length=43270923>
#CHROM  POS   ID  REF ALT QUAL FILTER  INFO    FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
chr01 130333 . C  T  225  . AC=90;AF=0.234   GT:DP:GQ   0/0:35:99   0/0:15:72   0/0:6:45    0/0:10:57   0/0:13:66   0/0:26:99   0/0:30:99
chr01 222644 . A  T  125  . AC=292;AF=0.764  GT:DP:GQ   1/1:25:72   1/1:16:45   1/1:17:48   1/1:8:21    1/1:17:48   1/1:49:99   1/1:27:78
chr01 242098 . A  G  188  . AC=196;AF=0.513  GT:DP:GQ   1/1:32:93   1/1:12:33   1/1:10:27   1/1:3:6     1/1:21:60   1/1:56:99   1/1:63:99
chr01 242110 . A  T   96  . AC=290;AF=0.763  GT:DP:GQ   1/1:32:93   1/1:12:30   1/1:10:27   1/1:4:6     1/1:20:57   1/1:57:99   1/1:60:90
chr01 242161 . G  A  196  . AC=87;AF=0.230   GT:DP:GQ   0/0:25:99   0/0:2:33    0/0:4:39    0/0:4:39    0/0:6:45    0/0:43:99   0/0:49:99
chr01 242383 . T  A  145  . AC=184;AF=0.487  GT:DP:GQ   1/1:24:69   1/1:13:36   ./.         1/1:4:9     1/1:12:33   1/1:26:75   1/1:20:57
chr01 267711 . G  C   59  . AC=264;AF=0.754  GT:DP:GQ   ./.         ./.         1/1:6:15    ./.         1/1:6:15    1/1:14:39   1/1:6:15
```

Python Scripting

```
>SAMPLE1
CTGTGAnnnGCGATTCnnnnCCAAGAATATTGAACAGTGTGTTAACAnCAGGTTACACTCCTAAAAGTCTGAGACACGCATAGTACTTTCTCGTAGTTCG...
>SAMPLE2
CTGTGAnnCGCGATnCATTACCAAGAATATTGAACAGTGTnnnnnCnnnAGGTTACACTCCTAAAAGTCTGAGACACGCATAGTACTTTCTCGTAGTTCG...
>SAMPLE3
CTGTGnCACGCGATTCATTACCAAGAAnAnTGAACAGTGTGTTAACnCCAGGTTAnACTCCTAAAAGnnTGAGACACnnATAGTACTTTCTGGTAGTTCG...
>SAMPLE4
CTGTGAnnnGCnnnTCnnnnCCAAnAAnnnTnnGCAGTGTRTTAACACCAnGTTAnnCTCCTAAAAGTnTGAnnnnCGCATnnnnnnTTnTnnTnGTTCG...
...
```

FASTA format

계통수 분석 - Input file 작성 (VCF to FASTA)

Input format 작성 명령어

python vcf2fasta.py combined_flt_snps.PASS.vcf

Input format 결과 파일 확인

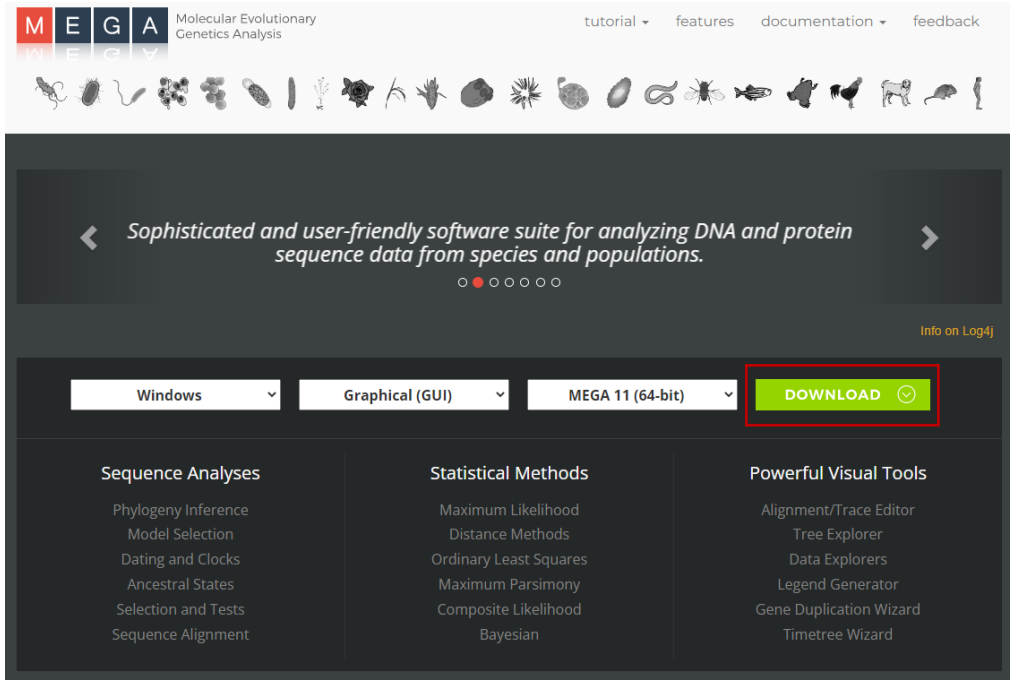
less phylo_tree.fasta

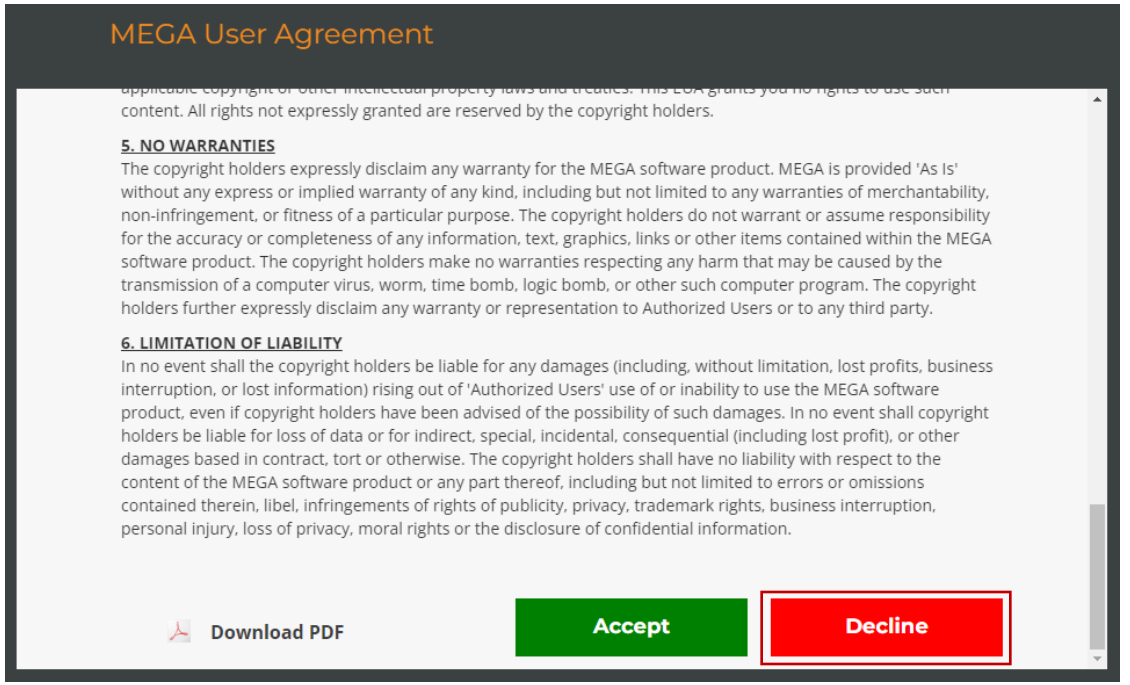
- phylo_tree.fasta 파일 예시

```
>MixR
CTACTCCCAGGCCCKTRAGTCCCCGCGTGAYKWAGYTRRGGCCGTGAACTGYAAGCRTAMGACCARCACRMTYYYRTYYRRMAATYWACRMYRCTRCGATCAACYRRTCWATTTSAYCCTYRTGCRYRRRWRRMYC
>MixS
MYACTCCCAGGCCCTKRRGTCYCCRYGYGRYKWMGYTGRRGCCGTGAACYGYRMGYRYRMGRYCRRYMYRCYYYYRTYYRRMAATYWACRMYRCTRYGRKYAACYRRWYWATTTSWYSYWYRKGCRYRRRWRRMYC
>P1
CTMSKSSSWSSSSSTTAAGTCCCCGCGTRRYGAAGTTGGGGCCGTGAACTGCAAGCATAAGACCAGCACGCTYCCATCCAGCAATTTRYGCCGYKACRRKYRRMTGGTCTATTYSAYCCTYRTGCRTRRRWRGACC
>P2
CTACTCCCAGGCCCTTGAGTCCCCGCGTGATTTAGCYGGGRYYRYRRRYTGCAAGCGTACGACCAACACACTCYYRYYYRRMRWWYWACRMYRCTGCGATCAACCAGTCAWKTTCWCSYWTGKKSRYRRRWRRMCM
```

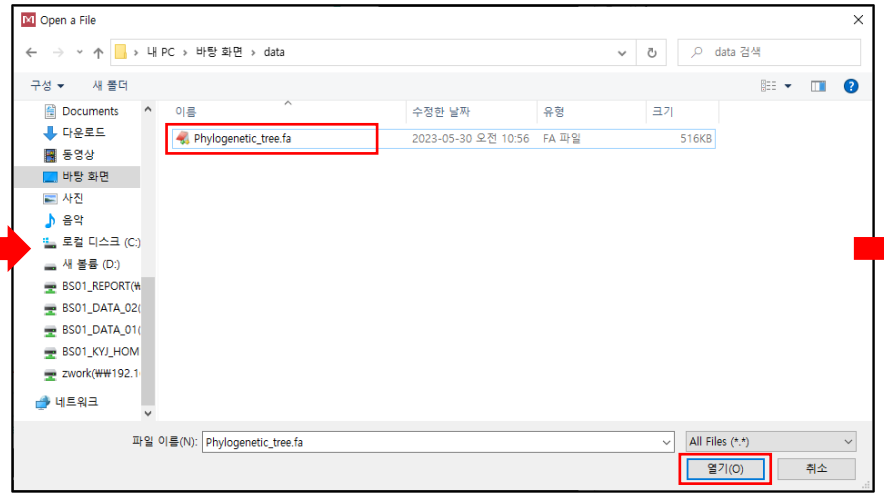
계통수 분석 - 프로그램 설치

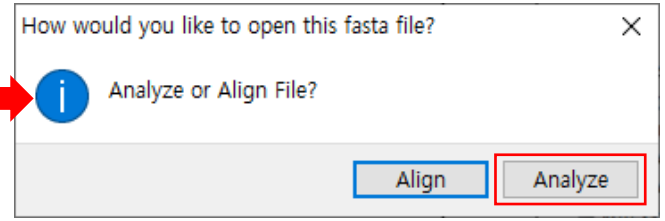
- MEGA 11 프로그램 다운로드 : https://www.megasoftware.net/

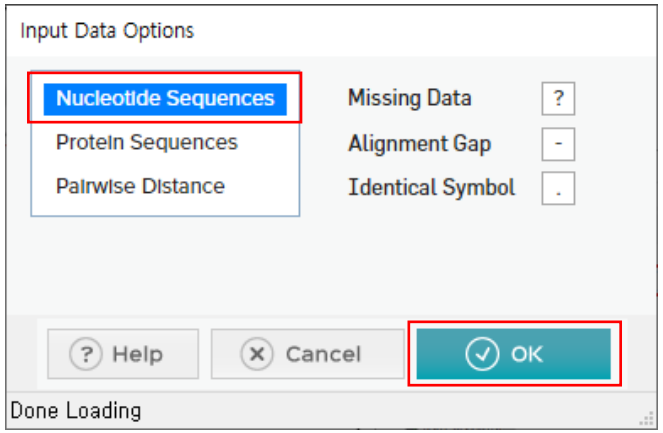


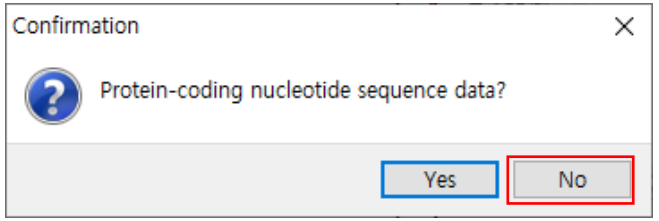


계통수 분석 - 프로그램 실행/ 파일 불러오기

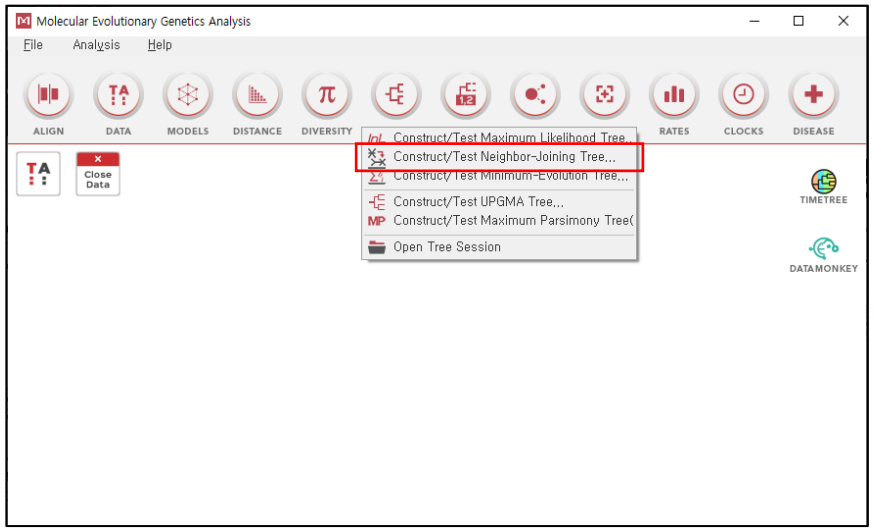


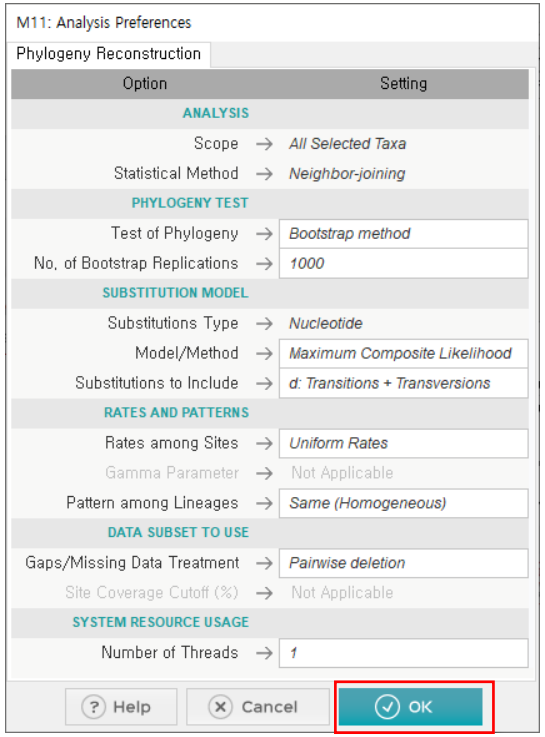


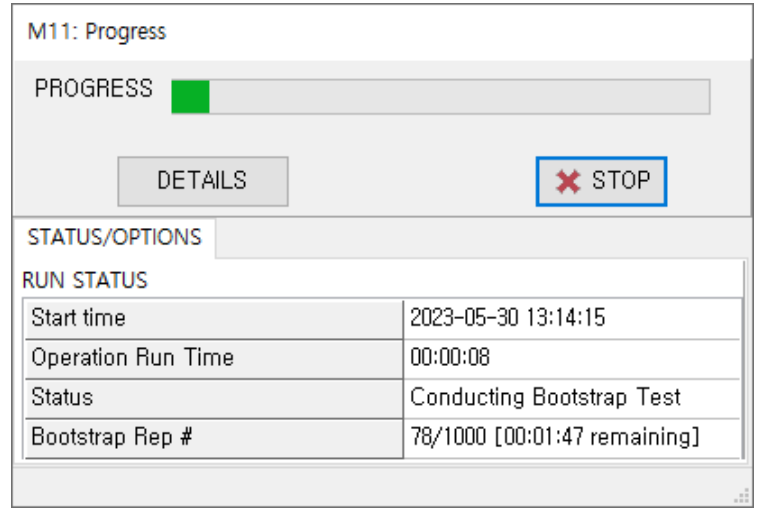




계통수 분석 - Neighbor-joining tree 작성하기

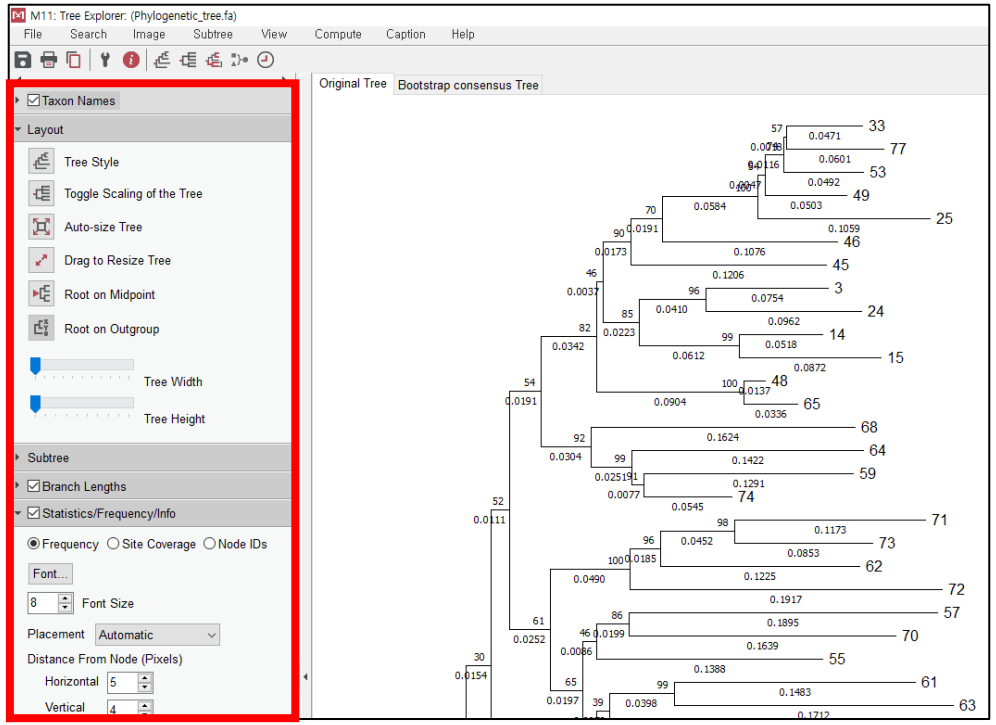


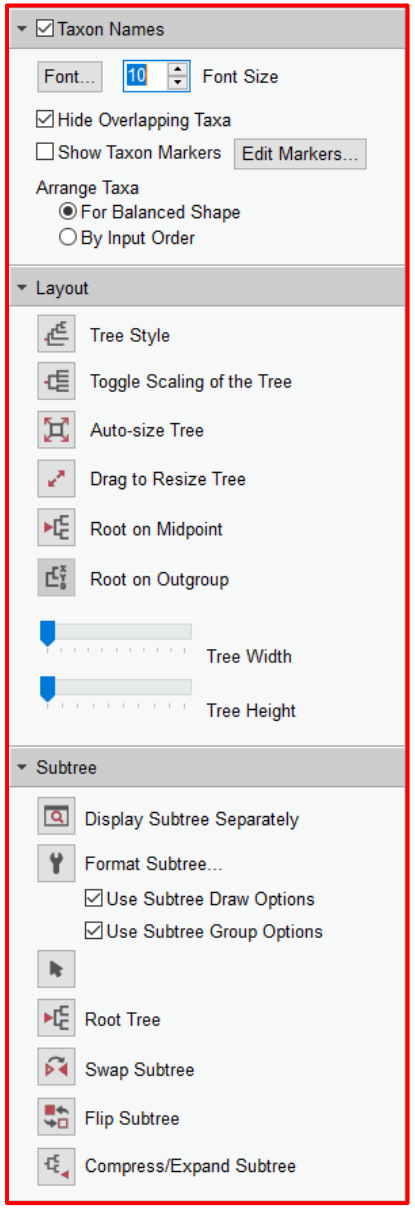




계통수 분석 - Tree 작성 결과 정리하기

〈MEGA11 결과〉

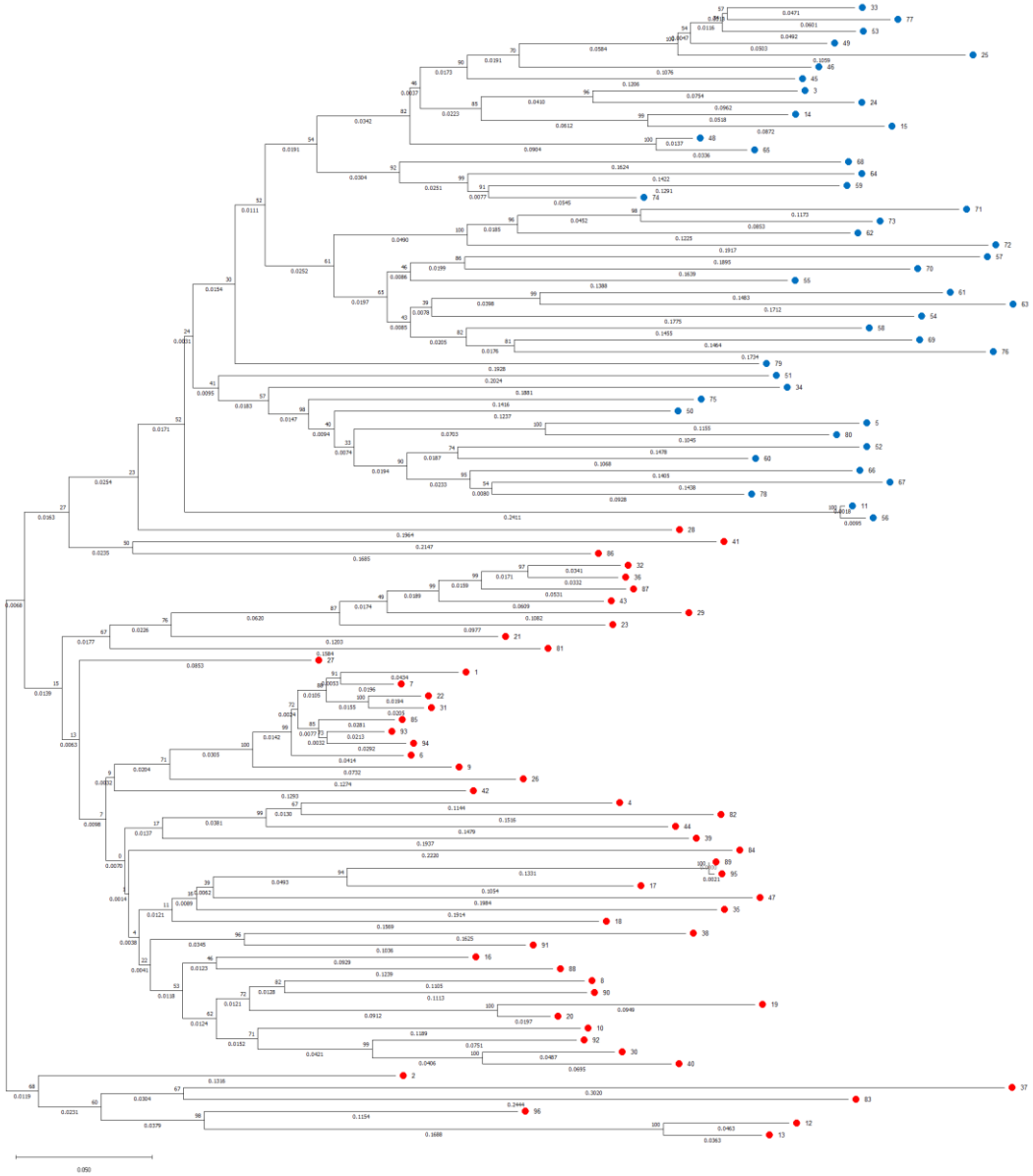




Tree 결과는 오른쪽 탭에 있는 설정을 통해 조절하여 정리할 수 있음.

– 라벨폰트, 사이즈, 포인트종류, 노드 위치 등,…

계통수 분석 - 계통수 분석 결과



감사합니다.

대전시 유성구 테크노2로 187, B동 412호

bi@bioto.co.kr

042-710-0077

070-7585-5344