

# BI 코딩실무

한주현

200523

Bioinformatics Engineer at 3billion

Interdisciplinary Program in Medical Informatics at Seoul National University

# 강사소개

강사 소개	
이름	한주현
소속	3billion, Bioinformatics Engineer Seoul National University, Medical Informatics
메일	<a href="mailto:kenneth.jh.han@snu.ac.kr">kenneth.jh.han@snu.ac.kr</a>
주요 업무	Human Genome Analysis (WGS, WES) Rare Disease Analysis Bioinformatics Algorithms Analysis Pipeline / Platform Development Full Stack Development Cloud Computing
주 언어	Python, JAVA, JavaScript, Bash shell
저서	니콜라스 볼커 이야기 (2016.10, 금창원 외 공역) 바이오파이썬으로 시작하는 생물정보학 (2019.03, 한주현)
웹 페이지	<a href="https://korbillgates.tistory.com">https://korbillgates.tistory.com</a> (블로그)

강의 내용 또는 생물정보학, 취업 및 진로에 관련하여  
궁금한점이 있으시면 언제든지 메일로 문의해주세요

# 강의 내용

---

- DNA Sequencing 데이터의 분석 방법 학습
- LINUX 환경에서 생물정보 툴 설치
- 파이썬 프로그래밍으로 생물정보 파일들 파싱하기
- LINUX를 활용한 Bioinformatics Pipeline 제작 학습

## 목표 (과제)

---

- 1) 제공한 FASTQ 파일의 염기서열 개수를 세어본다.
- 2) 리눅스 환경에서 생물정보학 tool(samtools)을 설치하여 실행해보기.
- 3) 제공한 BAM 파일의 특정 영역 (chr10:42385150)을 samtools tview로 시각화하여 스크린샷한다.
- 4) 제공한 VCF 파일의 SNP와 Insertion, Deletion 개수를 세어본다.

# 준비물

---

- Linux 환경 (ubuntu 18.04 를 추천)
- Python3

참고 페이지

<https://kennethjhan.github.io/Genome-Analysis-Tutorial/>

# 다음 파일들을 다운로드 하두세요

---

## FASTQ

[https://github.com/KennethJHan/Bioinformatics\\_Programming\\_101/raw/master/GATK\\_BestPractice/SRR000982\\_1.filt.fastq.gz](https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982_1.filt.fastq.gz)

[https://github.com/KennethJHan/Bioinformatics\\_Programming\\_101/raw/master/GATK\\_BestPractice/SRR000982\\_2.filt.fastq.gz](https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982_2.filt.fastq.gz)

## BAM

[https://github.com/KennethJHan/Bioinformatics\\_Programming\\_101/raw/master/GATK\\_BestPractice/SRR000982.mapped.sorted.markdup.bam](https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982.mapped.sorted.markdup.bam)

[https://github.com/KennethJHan/Bioinformatics\\_Programming\\_101/raw/master/GATK\\_BestPractice/SRR000982.mapped.sorted.markdup.bam.bai](https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982.mapped.sorted.markdup.bam.bai)

## VCF

[https://raw.githubusercontent.com/KennethJHan/Bioinformatics\\_Programming\\_101/master/GATK\\_BestPractice/SRR000982.filtered.variants.annotated.vcf](https://raw.githubusercontent.com/KennethJHan/Bioinformatics_Programming_101/master/GATK_BestPractice/SRR000982.filtered.variants.annotated.vcf)

# 리눅스에서 파일 다운로드 하는 방법

1) 적당한 장소에 다운받을 디렉토리를  
mkdir 로 생성합니다

```
kenneth_jh_han@instance-1: ~/Downloads - Google Chrome
ssh.cloud.google.com/projects/lecture-276814/zones/us-west1-b/instances/instance-1?authuser=0&hl=en_US&projectNumber=32366...

kenneth_jh_han@instance-1:~$ mkdir Downloads
kenneth_jh_han@instance-1:~$ cd Downloads
kenneth_jh_han@instance-1:~/Downloads$ ll
total 8
drwxrwxr-x 2 kenneth_jh_han kenneth_jh_han 4096 May 21 09:27 ./
drwxr-xr-x 6 kenneth_jh_han kenneth_jh_han 4096 May 21 09:27 ../
kenneth_jh_han@instance-1:~/Downloads$
```

2) 파일 주소를 wget에 넣어  
다운로드 받습니다.

```
kenneth_jh_han@instance-1: ~/Downloads - Google Chrome
ssh.cloud.google.com/projects/lecture-276814/zones/us-west1-b/instances/instance-1?authuser=0&hl=en_US&projectNumber=32366...

kenneth_jh_han@instance-1:~/Downloads$ wget https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982_1.filt.fastq.gz
--2020-05-21 09:31:48-- https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982_1.filt.fastq.gz
Resolving github.com (github.com)... 192.30.255.112
Connecting to github.com (github.com)|192.30.255.112|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/KennethJHan/Bioinformatics_Programming_101/master/GATK_BestPractice/SRR000982_1.filt.fastq.gz [following]
--2020-05-21 09:31:48-- https://raw.githubusercontent.com/KennethJHan/Bioinformatics_Programming_101/master/GATK_BestPractice/SRR000982_1.filt.fastq.gz
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.192.133, 151.101.128.133, 151.101.64.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|151.101.192.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 8502024 (8.1M) [application/octet-stream]
Saving to: 'SRR000982_1.filt.fastq.gz'

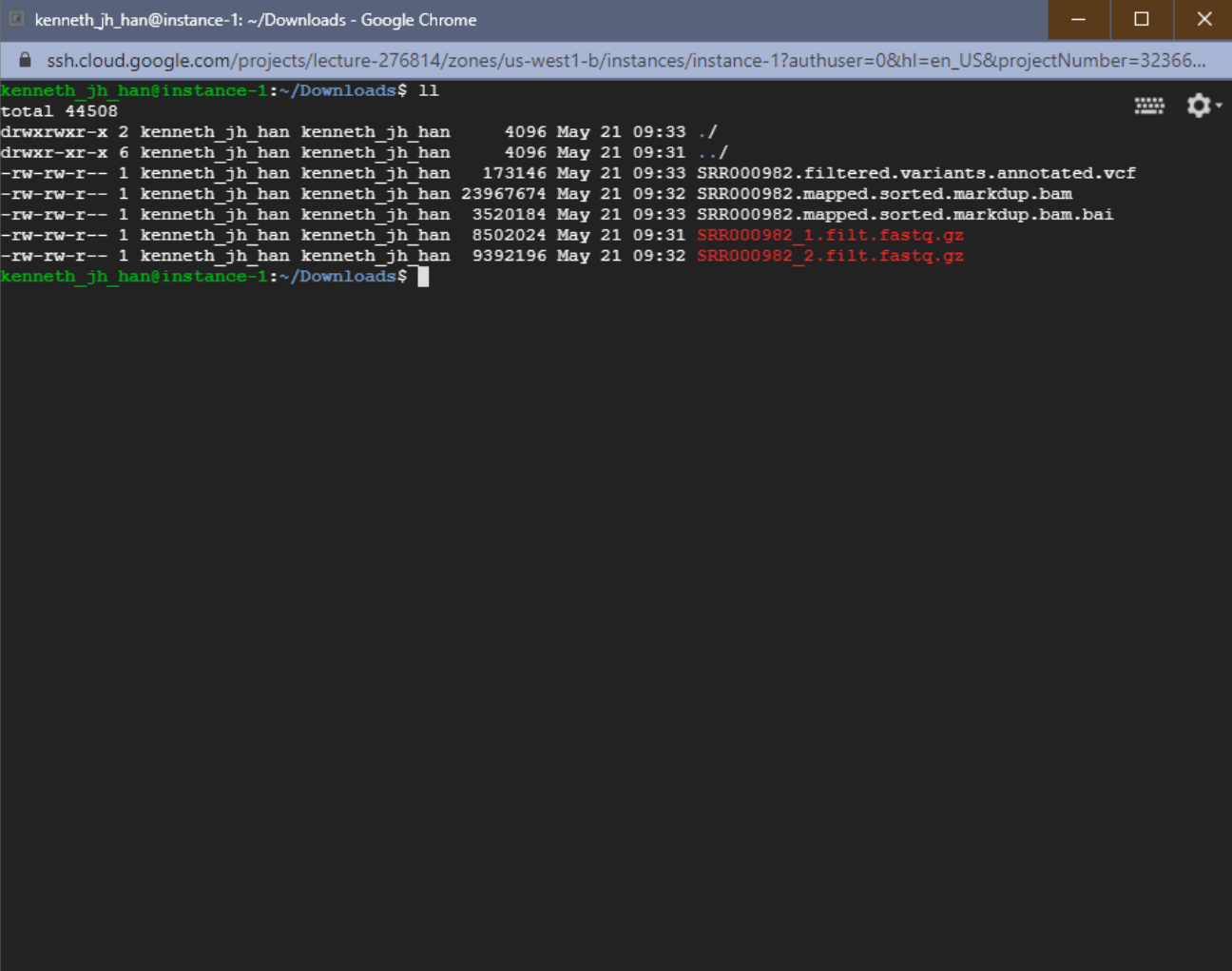
SRR000982_1.filt.fastq.gz 100%[=====>] 8.11M 35.8MB/s in 0.2s

2020-05-21 09:31:49 (35.8 MB/s) - 'SRR000982_1.filt.fastq.gz' saved [8502024/8502024]

kenneth_jh_han@instance-1:~/Downloads$ ll
total 8312
drwxrwxr-x 2 kenneth_jh_han kenneth_jh_han 4096 May 21 09:31 ./
drwxr-xr-x 6 kenneth_jh_han kenneth_jh_han 4096 May 21 09:31 ../
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 8502024 May 21 09:31 SRR000982_1.filt.fastq.gz
kenneth_jh_han@instance-1:~/Downloads$
```

# 리눅스에서 파일 다운로드 하는 방법

3) 모든 파일을 다 받으면  
다음 그림과 같습니다.  
파일의 개수와 사이즈를  
체크해 보세요.

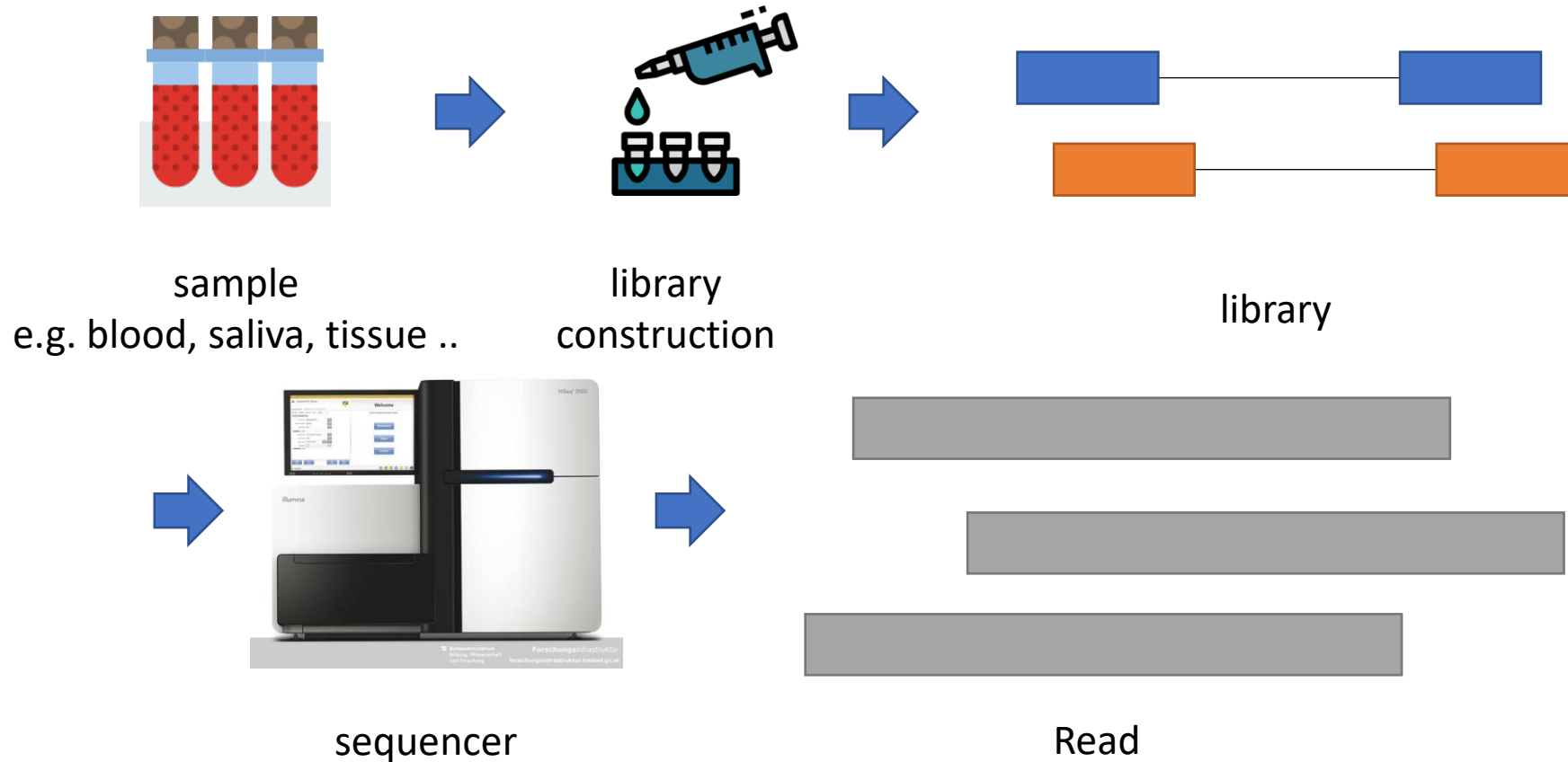


```
kenneth_jh_han@instance-1: ~/Downloads - Google Chrome
ssh.cloud.google.com/projects/lecture-276814/zones/us-west1-b/instances/instance-1?authuser=0&hl=en_US&projectNumber=32366...

kenneth_jh_han@instance-1:~/Downloads$ ll
total 44508
drwxrwxr-x 2 kenneth_jh_han kenneth_jh_han 4096 May 21 09:33 ./
drwxr-xr-x 6 kenneth_jh_han kenneth_jh_han 4096 May 21 09:31 ../
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 173146 May 21 09:33 SRR000982.filtered.variants.annotated.vcf
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 23967674 May 21 09:32 SRR000982.mapped.sorted.markdup.bam
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 3520184 May 21 09:33 SRR000982.mapped.sorted.markdup.bam.bai
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 8502024 May 21 09:31 SRR000982_1.filt.fastq.gz
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 9392196 May 21 09:32 SRR000982_2.filt.fastq.gz
kenneth_jh_han@instance-1:~/Downloads$
```

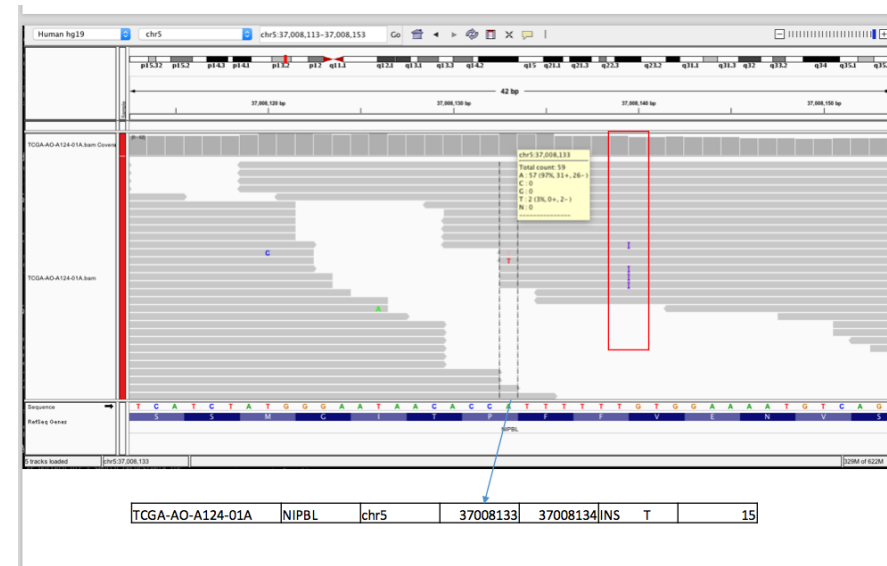
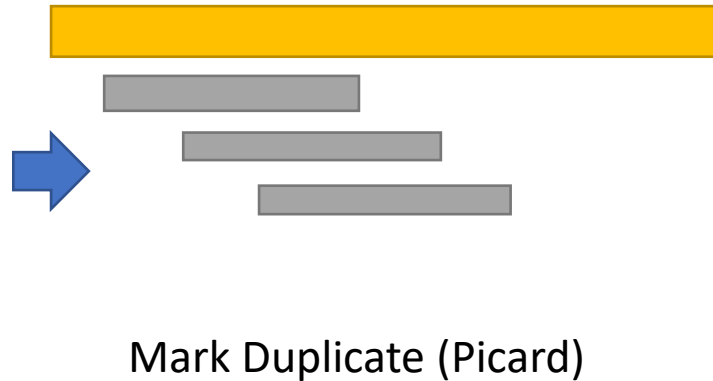
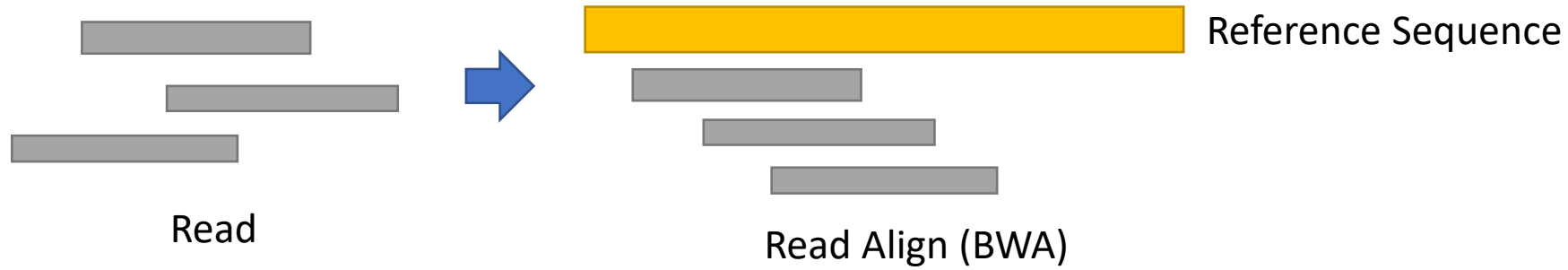


# Sequencing의 전반적 개요



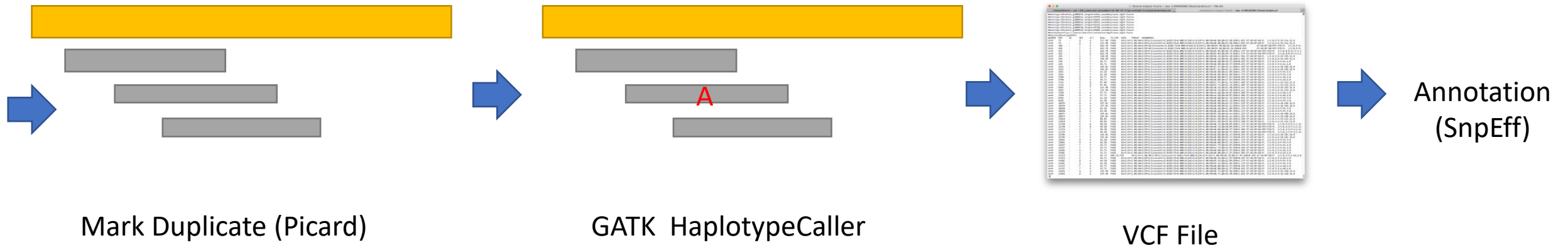
혈액, 타액, 조직 등과 같은 샘플에서 DNA를 추출하여 sequencer가 읽을 수 있도록 라이브러리를 제작하고 sequencing을 합니다. Sequencer가 라이브러리를 읽으면 read라고 하는 단위로 염기서열을 읽습니다. Illumina(일루미나) sequencing 장비의 경우 라이브러리를 앞뒤로 읽어서 read1, read2로 두 개의 read 파일을 생성합니다.

# Sequencing의 전반적 개요



읽어낸 read는 기준서열(reference sequence)와 비교하여 조각을 맞추는 align 작업을 하고, pcr 과정에서 생긴 duplicate을 picard와 같은 툴로 제거합니다.

# Sequencing의 전반적 개요



그리고 난 뒤 기준 서열과 다른 염기를 찾아내는 과정인 variant calling 을 진행합니다. Variant call을 하게 되면 VCF (Variant Calling Format) 파일이 생성됩니다. VCF 파일은 테이블과 같은 구조로 생긴 텍스트 파일입니다. VCF 파일에는 변이의 위치와 기준서열, 변이서열 정보만 있어서 이 변이가 어떤 의미인지 붙이는 작업이 필요한데, 이를 annotation 이라고 합니다.

FASTQ 파일은 보통 gzip으로  
압축된 형태이기에  
눈으로 보려면 zless 명령어를  
사용하여 봅니다.

```
kenneth_jh_han@instance-1:~/Downloads$ zless SRR000982_1.filt.fastq.gz
```

```
@SRR000982.5 E745RJU01DDHJ6 length=113
AAGGCACCATGCAGAGATGAAGGCCCTTTCTAAGCCTTAGACTTCTGGATGACATTCTAGAAACACCCTGGGCAGAAGTGAACTGTGCCTTGAGGGGAATAACTCG
+
DDDDDDDDDDDDDDDDDDFFDDBB:::@DDDDDDDDDFEDDAADDDDDDDDDDA8666@DD@#866AAADDNDDDDDDDDDDDDDDDDCCCCAAAACDDDDDDD
@SRR000982.26 E745RJU01BNUNR length=113
AAGGCACCATGCAGAGATGAAGGCCCTTTCTAAGCCTTAGACTTCTGGATGACATTCTAGAAACACCCTGGGCAGAAGTGAACTGTGCCTTGAGGGGAATAACTCAG
+
DDDDDDDDDDDDDDDDDDDD66555566@DDDDDDDGIIIFFEBDNDDDDDBBDDDD8552@DD@#8669@@DDDDDDDDDDDDDB@@@@@5566BBDDDDDD
@SRR000982.32 E745RJU01BGMLP length=66
AGGGAAAGGACTCTCTATAAGATGATATATGAGTAGACATCTGAAGTCAGCAAGGT CATGAGCAAT
+
FFFFFFFFFFFFFFFFIIIII IIIIIIIIIIIIIIIIIIIFFFFFFFFFFF FFFFFFFF
```

## 첫 번째 줄은 @로 시작하는 헤더

## 두 번째 줄은 염기서열

## 세 번째 줄은 + 기호로 된 구분 줄

네 번째 줄은 염기서열에 대한 quality score 입니다.

# FASTQ 파일에서 염기 세기

```
kenneth_jh_han@instance-1:~/Downloads$ python3 read_base.py SRR000982_1.filt.fastq.gz
3743235 2466934 2408976 3770887
kenneth_jh_han@instance-1:~/Downloads$ python3 read_base.py SRR000982_2.filt.fastq.gz
3630460 2403252 2436010 3625665
kenneth_jh_han@instance-1:~/Downloads$
```

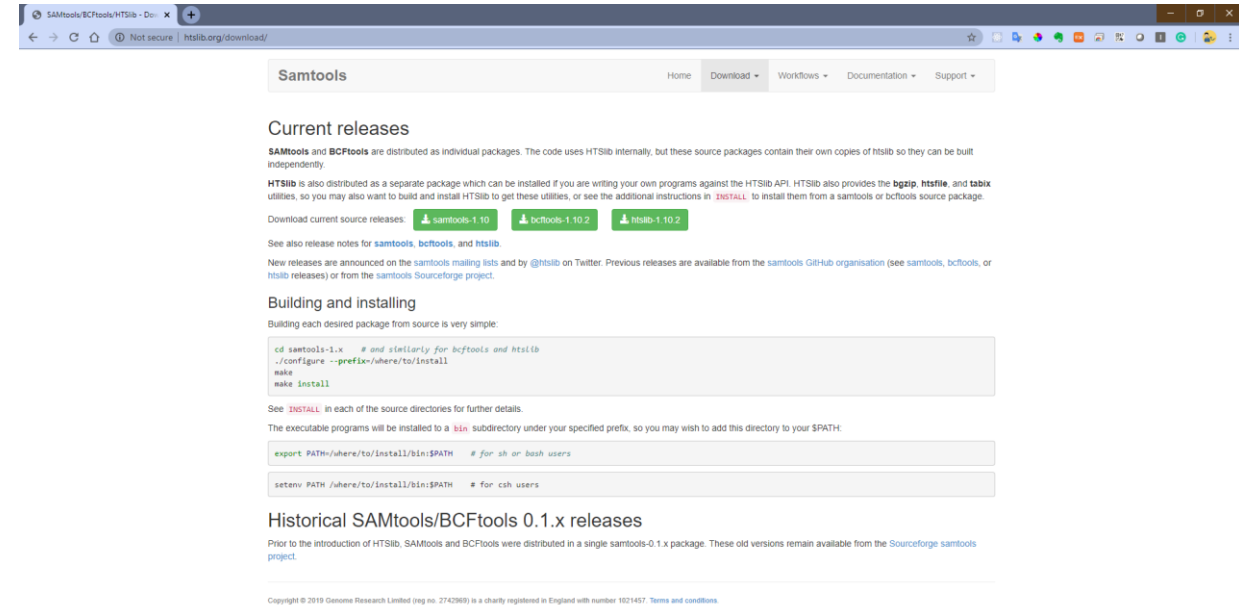
read1, read2 FASTQ 파일에서 A, C, G, T 염기수를 세어보세요.  
결과는 A, C, G, T 순으로 그림과 같이 출력합니다.  
gzip 파일 상태로 읽기 위해서 import gzip 을 사용해 보세요.  
gzip 사용법을 구글에서 스스로 찾아서 사용방법을 익혀봅시다.

# samtools 설치

## 1) samtools 홈페이지

<http://www.htslib.org/download/>

에 들어가서 samtools-1.10 으로 된  
버튼에서 우클릭 후 주소복사를  
합니다.



## 2) 복사한 주소

<https://github.com/samtools/samtools/releases/download/1.10/samtools-1.10.tar.bz2>

를 리눅스에서 wget을 사용하여 다운로드 받습니다.

# samtools 설치

3) tar xf 명령어로 압축을 해제합니다.

4) 압축을 해제한 디렉터리에 들어가서 ./configure 라고 입력합니다.

만약 오른쪽 아래 그림같이 gcc... no 라고 뜬다면 c compiler가 없는 것으로 설치해주어야합니다. 4-1 참조.

4-1) sudo apt-get install gcc g++ 을 타이핑 하고 엔터를 눌러 c compiler를 설치해줍니다. 중간에 묻는것이 나오면 y를 눌러서 계속 진행해줍니다.

```
kenneth_jh_han@instance-1:~/Downloads$ wget https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982.mapped.sorted.markdup.bam
--2020-05-21 09:32:52-- https://github.com/KennethJHan/Bioinformatics_Programming_101/raw/master/GATK_BestPractice/SRR000982.mapped.sorted.markdup.bam
Resolving github.com (github.com)... 140.82.114.4
Connecting to github.com (github.com):140.82.114.4:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/KennethJHan/Bioinformatics_Programming_101/master/GATK_BestPractice/SRR000982.mapped.sorted.markdup.bam [following]
--2020-05-21 09:32:52-- https://raw.githubusercontent.com/KennethJHan/Bioinformatics_Programming_101/master/GATK_BestPractice/SRR000982.mapped.sorted.markdup.bam
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 151.101.192.133, 151.101.128.133, 151.101.64.133
Connecting to raw.githubusercontent.com (raw.githubusercontent.com):151.101.192.133:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 23967674 (23M) [application/octet-stream]
Saving to: 'SRR000982.mapped.sorted.markdup.bam'
SRR000982.mapped.sorted.markdup.bam
kenneth_jh_han@instance-1:~/Downloads$ tar xf samtools-1.10.tar.bz2
kenneth_jh_han@instance-1:~/Downloads$ cd samtools-1.10/
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ ls
AUTHORS      bam2depth.c  bam_mate.c  bam_tview.c  configure.ac  padding.c  stats_isize.c
ChangeLog.old bam_addrprg.c bam_md.c    bam_tview.h  coverage.c   phase.c    stats_isize.h
INSTALL      bam_aux.c   bam_plbuf.c bam_tview curses.c  cut_target.c sam.c      test
LICENSE      bam_cat.c   bam_plbuf.h bam_tview_html.c  dict.c       sam.h      tmp_file.c
Makefile     bam_color.c bam_plcmd.c bam_shuf.c    doc          sam_opts.c tmp_file.h
NEWS         bam_endian.h bam_quickcheck.c bamtk.c       examples     sam_opts.h version.sh
README       bam_fastq.c bam_reheader.c bedcov.c      faidx.c      sam_utils.c
bam.c        bam_flags.c bam_rmdup.c  bedidx.c     htlib-1.10  sample.c
bam.h         bam_index.c bam_rmdupse.c bedidx.h      install-sh   sample.h
bam2bcf.c    bam_lpileup.c bam_sort.c  config.h.in  lz4         samtools.h
bam2bcf.h    bam_lpileup.h bam_split.c config.mk.in  m4           stats.c
bam2bcfindel.c bam_markdup.c bam_stat.c  configure    misc
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ ./configure
checking for gcc... no
checking for cc... no
checking for cl.exe... no
configure: error: in '/home/kenneth_jh_han/Downloads/samtools-1.10':
configure: error: no acceptable C compiler found in $PATH
See 'config.log' for more details
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ sudo apt-get install gcc g++
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  grub-pc-bin libnumal
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  binutils binutils-common binutils-x86-64-linux-gnu cpp cpp-7 g++ g++-7 gcc-7 gcc-7-base libasan4 libatomic1
  libbinutils libc-dev-bin libc6-dev libc6-i386 libcilkrts5 libgcc-7-dev libgcc1 libisl19 libitm1 liblsan0 libmpc3
  libmpx2 libquadmath0 libstdc++7-dev libtsan0 libubsan0 linux-libc-dev manpages-dev
Suggested packages:
  binutils-doc cpp-doc gcc-7-locales g++-multilib g++-multilib gcc-7-doc libstdc++6-7-dbg gcc-multilib make
  autoconf automake libtool flex bison gdb gcc-doc gcc-7-multilib libgcc1-dbg libgcc1-dbg libitm1-dbg
  libatomic1-dbg libasan4-dbg liblsan0-dbg libtsan0-dbg libubsan0-dbg libcilkrts5-dbg libmpx2-dbg
  libquadmath0-dbg glibc-doc libstdc++7-doc
The following NEW packages will be installed:
  binutils binutils-common binutils-x86-64-linux-gnu cpp cpp-7 g++ g++-7 gcc-7 gcc-7-base libasan4 libatomic1
  libbinutils libc-dev-bin libc6-dev libc6-i386 libcilkrts5 libgcc-7-dev libgcc1 libisl19 libitm1 liblsan0 libmpc3
  libmpx2 libquadmath0 libstdc++7-dev libtsan0 libubsan0 linux-libc-dev manpages-dev
0 upgraded, 30 newly installed, 0 to remove and 24 not upgraded.
Need to get 41.8 MB of archives.
After this operation, 160 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
```

# samtools 설치

4-1) which gcc 와 which g++ 가  
다음과 같이 나온다면 제대로 설치  
된 것입니다.

5) ./configure 를 하였을 때 다음과  
같이 FAILED 로 나올 수 있습니다.  
이건 뭔가 잘못해서 그런건 아니고,  
리눅스에서 프로그램을 설치할 때  
필요한 dependency (의존)  
프로그램이 없어서 그런것 입니다.  
다들 프로그래밍을 해보셔서 잘 아시겠지만  
스스로 디버깅 할 수 있는 수준이 되면  
프로그래밍이 쉬워집니다. 디버깅을 하는 방법은  
오류 메시지를 잘 읽어 보는 것 입니다.  
마찬가지로 다음 그림의 노란색 네모친 부분을  
보시면 libncurses5-dev 를 설치하라고 나옵니다.  
(ubuntu 기준)

6) sudo apt-get install libncurses5-dev  
를 실행합니다.

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ which gcc
/usr/bin/gcc
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ which g++
/usr/bin/g++
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$

kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ ./configure
checking for gcc... gcc
checking whether the C compiler works... yes
checking for C compiler default output file name... a.out
checking for suffix of executables...
checking whether we are cross compiling... no
checking for suffix of object files... o
checking whether we are using the GNU C compiler... yes
checking whether gcc accepts -g... yes
checking for gcc option to accept ISO C89... none needed
checking for grep that handles long lines and -e... /bin/grep
checking for C compiler warning flags... -Wall
checking for special C compiler options needed for large files... no
checking for _FILE_OFFSET_BITS value needed for large files... no
checking location of HTSlib source tree... htslib-1.10
checking for NcursesW wide-character library... no
checking for Ncurses library... no
checking for Curses library... no
configure: error: curses development files not found

The 'samtools tview' command uses the curses text user interface library.
Building samtools with tview requires curses/ncurses/etc development files
to be installed on the build machine; you may need to ensure a package such
as libncurses5-dev (on Debian or Ubuntu Linux) or ncurses-devel (on RPM-based
Linux distributions) is installed.

FAILED. Either configure --without-curses or resolve this error to build
samtools successfully.
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$

kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ sudo apt-get install libncurses5-dev
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  grub-pc-bin libnumal
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  libtinfo-dev
Suggested packages:
  ncurses-doc
The following NEW packages will be installed:
  libncurses5-dev libtinfo-dev
0 upgraded, 2 newly installed, 0 to remove and 24 not upgraded.
Need to get 256 kB of archives.
After this operation, 1422 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://us-west1.gce.archive.ubuntu.com/ubuntu bionic-updates/main amd64 libtinfo-dev amd64 6.1-1ubuntu1.18.04
  [81.3 kB]
Get:2 http://us-west1.gce.archive.ubuntu.com/ubuntu bionic-updates/main amd64 libncurses5-dev amd64 6.1-1ubuntu1.18
  .04 [174 kB]
Fetched 256 kB in 0s (2755 kB/s)
Selecting previously unselected package libtinfo-dev:amd64.
(Reading database ... 70400 files and directories currently installed.)
Preparing to unpack .../libtinfo-dev_6.1-1ubuntu1.18.04_amd64.deb ...
Unpacking libtinfo-dev:amd64 (6.1-1ubuntu1.18.04) ...
Selecting previously unselected package libncurses5-dev:amd64.
Preparing to unpack .../libncurses5-dev_6.1-1ubuntu1.18.04_amd64.deb ...
Unpacking libncurses5-dev:amd64 (6.1-1ubuntu1.18.04) ...
Setting up libtinfo-dev:amd64 (6.1-1ubuntu1.18.04) ...
Setting up libncurses5-dev:amd64 (6.1-1ubuntu1.18.04) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```



# samtools 설치

7) 다시 ./configure 를 입력합니다.  
이번에도 FAILED가 나왔는데,  
노란색 네모친 부분을 보면  
이유가 조금 다릅니다.  
sudo apt-get install zlib1g-dev  
로 dependency를 설치해줍니다.

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ ./configure
checking for gcc... gcc
checking whether the C compiler works... yes
checking for C compiler default output file name... a.out
checking for suffix of executables...
checking whether we are cross compiling... no
checking for suffix of object files... o
checking whether we are using the GNU C compiler... yes
checking whether gcc accepts -g... yes
checking for gcc option to accept ISO C89... none needed
checking for grep that handles long lines and -e... /bin/grep
checking for C compiler warning flags... -Wall
checking for special C compiler options needed for large files... no
checking for _FILE_OFFSET_BITS value needed for large files... no
checking location of HTSlib source tree... htlib-1.10
checking for NcursesW wide-character library... no
checking for Ncurses library... yes
checking for working ncurses/curses.h... no
checking for working ncurses.h... yes
checking for zlib.h... no
checking for inflate in -lz... no
configure: error: zlib development files not found

Samtools uses compression routines from the zlib library <http://zlib.net>.
Building samtools requires zlib development files to be installed on the build
machine; you may need to ensure a package such as zlib1g-dev (on Debian or
Ubuntu Linux) or zlib-devel (on RPM-based Linux distributions) is installed.

FAILED. This error must be resolved in order to build samtools successfully.
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ sudo apt-get install zlib1g-dev
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  grub-pc-bin libnuma1
Use 'sudo apt autoremove' to remove them.
The following NEW packages will be installed:
  zlib1g-dev
0 upgraded, 1 newly installed, 0 to remove and 24 not upgraded.
Need to get 176 kB of archives.
After this operation, 457 kB of additional disk space will be used.
Get:1 http://us-west1.gce.archive.ubuntu.com/ubuntu bionic/main amd64 zlib1g-dev amd64 1:1.2.11.dfsg-0ubuntu2 [176
kB]
Fetched 176 kB in 0s (11.1 MB/s)
Selecting previously unselected package zlib1g-dev:amd64.
(Reading database ... 70450 files and directories currently installed.)
Preparing to unpack .../zlib1g-dev_1%3a1.2.11.dfsg-0ubuntu2_amd64.deb ...
Unpacking zlib1g-dev:amd64 (1:1.2.11.dfsg-0ubuntu2) ...
Setting up zlib1g-dev:amd64 (1:1.2.11.dfsg-0ubuntu2) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

# samtools 설치

8) 다시 ./configure를 입력합니다.  
마치 프로그래밍 디버깅 하듯이  
오류 메시지를 보고 필요한  
dependency 들을 설치해주면 됩니다.  
윈도우나 맥과 같은 환경에서는 그냥 클릭하면  
알아서 잘 설치해주었는데 리눅스에서는  
조금은 불편하겠지만, 이렇게 설치해 주어야  
합니다.  
sudo apt-get install libbz2-dev  
를 입력합니다.

```
checking for library containing log... -lm
checking for zlib.h... yes
checking for inflate in -lz... yes
checking for library containing recv... none required
checking for bzlib.h... no
checking for BZ2_bzBuffToBuffCompress in -lbz2... no
configure: error: libbz2 development files not found
```

```
The CRAM format may use bzip2 compression, which is implemented in HTSlib
by using compression routines from libbz2 <http://www.bzip.org/>.
```

```
Building HTSlib requires libbz2 development files to be installed on the
build machine; you may need to ensure a package such as libbz2-dev (on Debian
or Ubuntu Linux) or bzip2-devel (on RPM-based Linux distributions or Cygwin)
is installed.
```

```
Either configure with --disable-bz2 (which will make some CRAM files
produced elsewhere unreadable) or resolve this error to build HTSlib.
configure: error: ./configure failed for htslib-1.10
```

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ sudo apt-get install libbz2-dev
```

```
Reading package lists... Done
```

```
Building dependency tree
```

```
Reading state information... Done
```

```
The following packages were automatically installed and are no longer required:
```

```
  grub-pc-bin libnumal
```

```
Use 'sudo apt autoremove' to remove them.
```

```
The following additional packages will be installed:
```

```
  bzip2-doc
```

```
The following NEW packages will be installed:
```

```
  bzip2-doc libbz2-dev
```

```
0 upgraded, 2 newly installed, 0 to remove and 24 not upgraded.
```

```
Need to get 324 kB of archives.
```

```
After this operation, 514 kB of additional disk space will be used.
```

```
Do you want to continue? [Y/n] y
```

```
Get:1 http://us-west1.gce.archive.ubuntu.com/ubuntu bionic-updates/main amd64 bzip2-doc all 1.0.6-8.1ubuntu0.2 [294
kB]
```

```
Get:2 http://us-west1.gce.archive.ubuntu.com/ubuntu bionic-updates/main amd64 libbz2-dev amd64 1.0.6-8.1ubuntu0.2 [
30.0 kB]
```

```
Fetched 324 kB in 0s (13.8 MB/s)
```

```
Selecting previously unselected package bzip2-doc.
```

```
(Reading database ... 70478 files and directories currently installed.)
```

```
Preparing to unpack .../bzip2-doc_1.0.6-8.1ubuntu0.2_all.deb ...
```

```
Unpacking bzip2-doc (1.0.6-8.1ubuntu0.2) ...
```

```
Selecting previously unselected package libbz2-dev:amd64.
```

```
Preparing to unpack .../libbz2-dev_1.0.6-8.1ubuntu0.2_amd64.deb ...
```

```
Unpacking libbz2-dev:amd64 (1.0.6-8.1ubuntu0.2) ...
```

```
Setting up libbz2-dev:amd64 (1.0.6-8.1ubuntu0.2) ...
```

```
Setting up bzip2-doc (1.0.6-8.1ubuntu0.2) ...
```

```
Processing triggers for install-info (6.5.0.dfsg.1-2) ...
```

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

# samtools 설치

9) 다시 ./configure를 합니다.

이번에도 에러가 났는데

sudo apt-get install liblzma-dev  
로 설치해줍니다.

```
checking for fdasyncc... yes
checking for library containing log... -lm
checking for zlib.h... yes
checking for inflate in -lz... yes
checking for library containing recv... none required
checking for bzlib.h... yes
checking for BZ2 bzBuffToBuffCompress in -lbz2... yes
checking for lzma.h... no
checking for lzma_easy_buffer_encode in -llzma... no
configure: error: liblzma development files not found
```

The CRAM format may use LZMA2 compression, which is implemented in HTSlib by using compression routines from liblzma <<http://tukaani.org/xz/>>.

Building HTSlib requires liblzma development files to be installed on the build machine; you may need to ensure a package such as liblzma-dev (on Debian or Ubuntu Linux), xz-devel (on RPM-based Linux distributions or Cygwin), or xz (via Homebrew on macOS) is installed; or build XZ Utils from source.

Either configure with --disable-lzma (which will make some CRAM files produced elsewhere unreadable) or resolve this error to build HTSlib.

configure: error: ./configure failed for htslib-1.10

kenneth\_jh\_han@instance-1:~/Downloads/samtools-1.10\$

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ sudo apt-get install liblzma-dev
```

Reading package lists... Done

Building dependency tree

Reading state information... Done

The following packages were automatically installed and are no longer required:

grub-pc-bin libnumal

Use 'sudo apt autoremove' to remove them.

Suggested packages:

liblzma-doc

The following NEW packages will be installed:

liblzma-dev

0 upgraded, 1 newly installed, 0 to remove and 24 not upgraded.

Need to get 145 kB of archives.

After this operation, 669 kB of additional disk space will be used.

Get:1 <http://us-west1.gce.archive.ubuntu.com/ubuntu/bionic/main amd64 liblzma-dev amd64 5.2.2-1.3> [145 kB]

Fetched 145 kB in 0s (8586 kB/s)

Selecting previously unselected package liblzma-dev:amd64.

(Reading database ... 70493 files and directories currently installed.)

Preparing to unpack .../liblzma-dev\_5.2.2-1.3\_amd64.deb ...

Unpacking liblzma-dev:amd64 (5.2.2-1.3) ...

Setting up liblzma-dev:amd64 (5.2.2-1.3) ...

kenneth\_jh\_han@instance-1:~/Downloads/samtools-1.10\$

# samtools 설치

10) 다시 ./configure를 합니다.

이번엔 제발 되었음 좋겠습니다.

다음 그림 같이 FAIL, error 없이 나온다면  
configure가 제대로 된 것 입니다.

만약 다른 dependency가 없어서 오류가 난다면  
스스로 필요한 것을 설치하여 해결해봅시다.

11) 지금까지는 설치에 필요한 요소들을

체크하는 과정이었고 이제 정말 설치를 해보겠습니다.

그 전에 which make 를 타이핑 합니다.

오른쪽 그림과 같이 아무 결과가 없다면 make가  
없는 것 입니다.

which 명령어는 프로그램이 리눅스 PATH 중 어디에 있는지  
찾아서 알려주는 명령어입니다.

sudo apt-get install make 를 입력합니다.

설치 후 which make를 하였을 때 다음 그림처럼  
나오면 제대로 make가 설치된 것 입니다.

```
checking how to run the C preprocessor... gcc -E
checking for egrep... /bin/grep -E
checking for ANSI C header files... yes
checking for sys/types.h... yes
checking for sys/stat.h... yes
checking for stdlib.h... yes
checking for string.h... yes
checking for memory.h... yes
checking for strings.h... yes
checking for inttypes.h... yes
checking for stdint.h... yes
checking forunistd.h... yes
checking for stdlib.h... (cached) yes
checking forunistd.h... (cached) yes
checking for sys/param.h... yes
checking for getpagesize... yes
checking for working mmap... yes
checking for gmtime_r... yes
checking for fsync... yes
checking for drand48... yes
checking whether fdatsync is declared... yes
checking for fdatsync... yes
checking for library containing log... -lm
checking for zlib.h... yes
checking for inflate in -lz... yes
checking for library containing recv... none required
checking for bzlib.h... yes
checking for BZ2 bzBuffToBuffCompress in -lbz2... yes
checking for lzma.h... yes
checking for lzma_easy_buffer_encode in -llzma... yes
checking for libdeflate.h... no
checking for libdeflate deflate_compress in -ldflate... no
checking for curl_easy_pause in -lcurl... no
checking for curl_easy_init in -lcurl... no
configure: WARNING: libcurl not enabled: library not found
configure: WARNING: GCS support not enabled: requires libcurl support
configure: WARNING: S3 support not enabled: requires libcurl support
checking whether PTHREAD_MUTEX_RECURSIVE is declared... yes
configure: creating ./config.status
config.status: creating config.mk
config.status: creating htlib.pc.tmp
config.status: creating config.h
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$

kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ which make
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ sudo apt-get install make
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  grub-pc-bin libnumal
Use 'sudo apt autoremove' to remove them.
Suggested packages:
  make-doc
The following NEW packages will be installed:
  make
0 upgraded, 1 newly installed, 0 to remove and 24 not upgraded.
Need to get 154 kB of archives.
After this operation, 381 kB of additional disk space will be used.
Get:1 http://us-west1.gce.archive.ubuntu.com/ubuntu bionic/main amd64 amd64 4.1-9.1ubuntu1 [154 kB]
Fetched 154 kB in 0s (9749 kB/s)
Selecting previously unselected package make.
(Reading database ... 70533 files and directories currently installed.)
Preparing to unpack .../make_4.1-9.1ubuntu1_amd64.deb ...
Unpacking make (4.1-9.1ubuntu1) ...
Setting up make (4.1-9.1ubuntu1) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ which make
/usr/bin/make
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

# samtools 설치

- 11) sudo make 를 타이핑 합니다.
- 12) sudo make install 을 타이핑 합니다.
- 13) which samtools 를 타이핑 합니다.
- 14) samtools 를 실행하였을 때  
오른쪽 그림같이 나오면 제대로 된 것  
입니다.

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ sudo make install
mkdir -p -m 755 /usr/local/bin /usr/local/bin /usr/local/share/man/man1
install -p samtools /usr/local/bin
install -p misc/ace2sam misc/maq2sam-long misc/maq2sam-short misc/md5fa misc/md5sum-lite misc/wgsim /usr/local/bin
install -p misc/blast2sam.pl misc/bowtie2sam.pl misc/export2sam.pl misc/interpolate_sam.pl misc/novo2sam.pl misc/pl
ot-bamstats misc/psl2sam.pl misc/sam2vcf.pl misc/samtools.pl misc/seq_cache_populate.pl misc/soap2sam.pl misc/wgsim
_eval.pl misc/zoom2sam.pl /usr/local/bin
install -p -m 644 doc/samtools*.1 misc/wgsim.1 /usr/local/share/man/man1
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$
```

```
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ which samtools
/usr/local/bin/samtools
kenneth_jh_han@instance-1:~/Downloads/samtools-1.10$ samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 1.10 (using htslib 1.10)

Usage:  samtools <command> [options]

Commands:
-- Indexing
dict          create a sequence dictionary file
faidx         index/extract FASTA
fqidx         index/extract FASTQ
index         index alignment

-- Editing
calmd         recalculate MD/NM tags and '=' bases
fixmate       fix mate information
reheader      replace BAM header
targetcut     cut fosmid regions (for fosmid pool only)
addreplacerg  adds or replaces RG tags
markdup       mark duplicates

-- File operations
collate       shuffle and group alignments by name
cat           concatenate BAMs
merge         merge sorted alignments
mpileup       multi-way pileup
sort          sort alignment file
split         splits a file by read group
quickcheck    quickly check if SAM/BAM/CRAM file appears intact
fastq         converts a BAM to a FASTQ
fasta         converts a BAM to a FASTA

-- Statistics
bedcov        read depth per BED region
coverage      alignment depth and percent coverage
```

# BAM 파일 살펴보기

1) bam 파일을 받았던 경로로 가서,  
samtools view SRR000982.mapped.sorted.markdup.bam | less -S  
를 입력합니다.

2) 오른쪽 아래 그림과 같이  
나오게 됩니다.

```
kenneth_jh_han@instance-1:~/Downloads$ ll
total 49124
drwxrwxr-x 3 kenneth_jh_han kenneth_jh_han 4096 May 21 09:39 ./
drwxr-xr-x 6 kenneth_jh_han kenneth_jh_han 4096 May 21 09:34 ../
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 173146 May 21 09:33 SRR000982.filtered.variants.annotated.vcf
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 23967674 May 21 09:32 SRR000982.mapped.sorted.markdup.bam
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 3520184 May 21 09:33 SRR000982.mapped.sorted.markdup.bam.bai
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 8502024 May 21 09:31 SRR000982_1.filt.fastq.gz
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 9392196 May 21 09:32 SRR000982_2.filt.fastq.gz
drwxrwxr-x 9 kenneth_jh_han kenneth_jh_han 4096 May 21 10:13 samtools-1.10/
-rw-rw-r-- 1 kenneth_jh_han kenneth_jh_han 4721173 Dec 6 16:47 samtools-1.10.tar.bz2
kenneth_jh_han@instance-1:~/Downloads$ samtools view SRR000982.mapped.sorted.markdup.bam | less -S
kenneth_jh_han@instance-1:~/Downloads$
```

SRR000982.91192	115	chrM	9	60	102M	=	3300	3326	GTCTATCACCTATTAAACCACTACGCGGAGNTCTC
SRR000982.385325		1139	chrM	9	60	102M	=	3327	3326 GTCTATCACCTATTAAACCACTACGCGGAGNTCTC
SRR000982.271454		65	chrM	25	60	143M	=	14316	14292 ACCACTCACGGGAGCTCTCCATGCATT
SRR000982.125609		117	chrM	29	0	*	=	29	0 ACCCATATAACCCCTCCCCCAAATTC
SRR000982.125609		185	chrM	29	60	5M1D15M1D6M1D162M	=	29	0 CTCACGGAGCT
SRR000982.132204		177	chrM	63	60	4S10M1I129M	=	14215	14101 CGGTTCTGGGGGGTAGTGC
SRR000982.159836		131	chrM	65	60	82M25S	=	3045	2981 TGGGGGGTATGCACGCGATAGCATTGC
SRR000982.237678		65	chrM	67	60	113M	=	13393	13327 GGGGGTATGCACGCGATAGCATTGCGA
SRR000982.469601		177	chrM	90	60	62M	=	13979	13962 ACGAGACGCTGGAGCCGGAGCACCCCTA
SRR000982.124004		131	chrM	91	60	103M	=	2153	2063 CGAGACGCTGGAGCCGGAGCACCCCTAT
SRR000982.245083		131	chrM	97	60	33M	=	3131	3035 GCTGGAGCCGGAGCACCCCTATGTCGCA
SRR000982.11440	177	chrM	122	60	68M	=	14794	14726	CAGTATCTGCTCTTTGATTCTGCTCATCCCATTA
SRR000982.192107		65	chrM	124	60	79M1D31M	=	14913	14790 GTATCTGCTTTGATTCTCT

3) 이번엔  
samtools tview SRR000982.mapped.sorted.markdup.bam  
를 입력합니다.



# BAM 파일 살펴보기

samtools tview 의 모습입니다.

c 키를 누르면 염기서열 별로

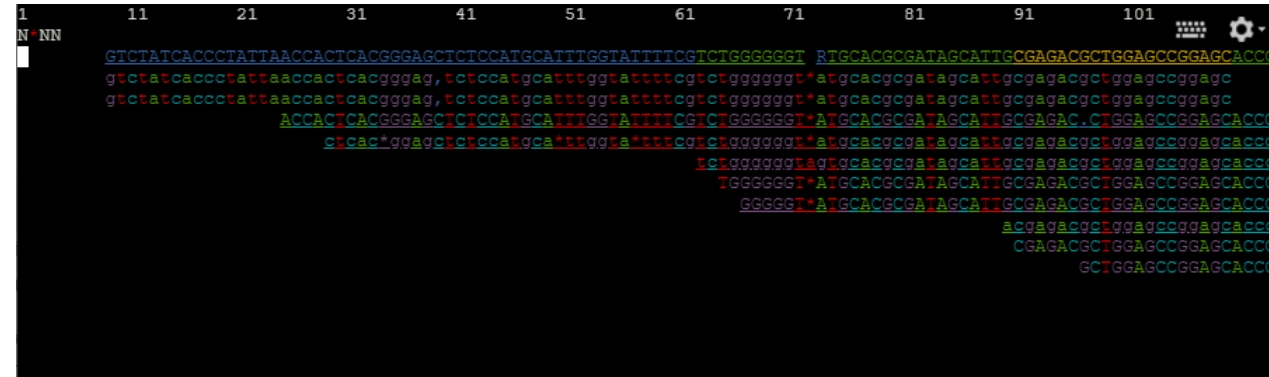
색을 입혀서 볼 수 있습니다.

방향키로 움직이면 보는 위치를

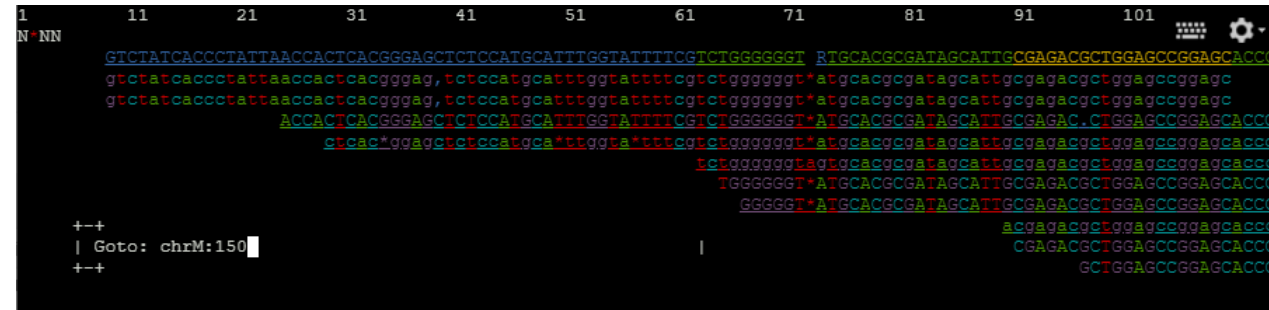
바꾸면서 볼 수 있습니다.

/ 키를 입력하면 특정 위치로 이동할 수 있는  
입력 창이 나옵니다.

chrM:150 을 입력해봅시다.



The screenshot shows the samtools tview interface. At the top, there's a header with positions 11, 21, 31, 41, 51, 61, 71, 81, 91, 101. Below it, the sequence data is displayed in a color-coded format (A: green, C: blue, G: red, T: yellow). A search bar at the bottom left contains the text "Goto: chrM:150".



This screenshot is identical to the one above, showing the same genomic track and search bar.



This screenshot shows a different view of the genomic track, with positions 151, 161, 171, 181, 191, 201, 211, 221, 231, 241, 251. The sequence data is displayed in a color-coded format. The search bar at the bottom left is empty.

# BAM 파일 살펴보기

실무에서 bam 파일을 보는 경우가 가끔 있는데, 리드들이 잘 쌓여있는지 기준 서열과 다른 변이들이 잘 있는지 보기 위해 살펴봅니다.



리드라고 하는 것은 시퀀서에서 한 번에 읽은 서열을 의미합니다. 염기서열들이 있으며 기준서열(reference genome)에 붙이면(mapping) 우리가 위에서 본 것 과 같이 나옵니다.



# VCF 파일 살펴보기

Meta-  
information  
Line

Header  
Data Line

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=SB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA000001 NA000002 NA000003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0/0:48:1:51,51 1/0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1/2:21:6:23,27 2/1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0/0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

FORMAT column

tag	meaning
GT	Position에 대한 샘플의 Genotype 0/0 - homozygous reference를 의미 0/1 - heterozygous REF/ALT를 의미 1/1 - homozygous ALT를 의미
AD	Allele Depth
DP	Depth

VCF 파일은 기준서열과 다른 변이들을 테이블 형식으로 표현한 파일입니다.

샷 기호(#) 두 개인 ## 에는 VCF 파일을 생성할 때 사용한 옵션 등의 설명이 있고

한 개인 # 에는 각 컬럼의 타이틀이 붙어있습니다.

9개의 의무적 컬럼이 있고 10번째 부터는 샘플의 FORMAT에 해당 값이 있습니다.

# VCF 파일 살펴보기

변이라고 하는 것은 기준서열과 일치하지 않는 염기를 의미합니다.

SNP (Single Nucleotide Polymorphism) : 기준서열과 비교하여 하나의 서열이 바뀐것을 의미	
기준 서열	ACAAGGTT
Read	ACATGGTT

Insertion : 기준서열과 비교하여 서열이 추가된 것을 의미	
기준 서열	ACAAGGTT
Read	ACAATGGTT

Deletion : 기준서열과 비교하여 서열이 제거된 것을 의미	
기준 서열	ACAAGGTT
Read	ACA*GGTT

# VCF 파일 살펴보기

AD (Allele Depth)는 GT (Genotype)를 기준으로 표현하는데,

CHROM	POS	REF	ALT	FORMAT	SAMPLE
chr20	1234	A	T	GT:AD:DP	0/1:23,11:34

Variant caller 중 하나인 GATK 를 기준으로 설명하자면

AD 에서 첫 번째는 GT의 0, 두 번째는 GT의 1, 세 번째는 GT의 2 번째를 의미합니다.

예를 들어 chr20:1234 A → T 에서, A는 GT의 0, T는 GT의 1입니다.

AD는 23,11 로 써져있는데 순서대로 0인 A가 → 23, 1인 T가 → 11만큼의 depth를 나타냅니다.

# VCF 파일 살펴보기

CHROM	POS	REF	ALT	FORMAT	SAMPLE
chr1	1234	A	T	GT:AD:DP	0/1:23,11:34
chr3	2222	TC	T,TCC	GT:AD:DP	1/2:2,31,21:54
chr21	2830	T	G	GT:AD:DP	1/1:0,44:44

Chr1: 1234 위치의 쌓인 ReadDepth는 34다. (O/X)

Chr1:1234 에서 A의 개수는 23개 T의 개수는 11개다. (O/X)

Chr3:2222 위치의 변이 종류는 SNP, Insertion, Deletion이다. (O/X)

Chr3:2222 위치의 REF Depth는 31, ALT Depth는 21이다. (O/X)

정답은 다음 슬라이드에

# VCF 파일 살펴보기

CHROM	POS	REF	ALT	FORMAT	SAMPLE
chr1	1234	A	T	GT:AD:DP	0/1:23,11:34
chr3	2222	TC	T,TCC	GT:AD:DP	1/2:2,31,21:54
chr21	2830	T	G	GT:AD:DP	1/1:0,44:44

Chr1: 1234 위치의 쌓인 ReadDepth는 34다. (O)

Chr1:1234 에서 A의 개수는 23개 T의 개수는 11개다. (O)

Chr3:2222 위치의 변이 종류는 SNP, Insertion, Deletion이다. (X)  
Insertion (TC → TCC) 과 Deletion (TC → T) 만 있고 SNP는 없다.

Chr3:2222 위치의 REF Depth는 31, ALT Depth는 21이다. (X)  
GT에서 TC:0, T:1, TCC:2 로 볼 수 있는데,  
AD에서 2,31,21 에서 TC,T,TCC 의 순서로 일치하기에  
REF 인 TC는 2, ALT1인 T는 31, ALT2인 TCC는 21이다.

# VCF 파일 살펴보기

```
kenneth_jh_han@instance-1:~/Downloads$ python3 calc_snp_indel.py SRR000982.filtered.variants.annotated.vcf
364 10 24
kenneth_jh_han@instance-1:~/Downloads$
```

VCF 파일에서 SNP, Insertion, Deletion 의 숫자를 세어보세요.

ALT 컬럼에서 쉼표로 구분되어 alt가 여러개 있는 경우도 있으니 이 점 참고하여 진행해주세요.

답은 그림과 같이 나옵니다.

오늘은  
FASTQ, BAM, VCF 파일을 살펴보고  
리눅스 툴을 설치해보았습니다.

과제를 꼭 수행해보시기 바랍니다.

그럼 다음에 또 만나요.