WES기반 SNV/small indel 발굴 및 분석 파이프라인

2016.08

서울대학교 생명과학부
백대현 교수

목차

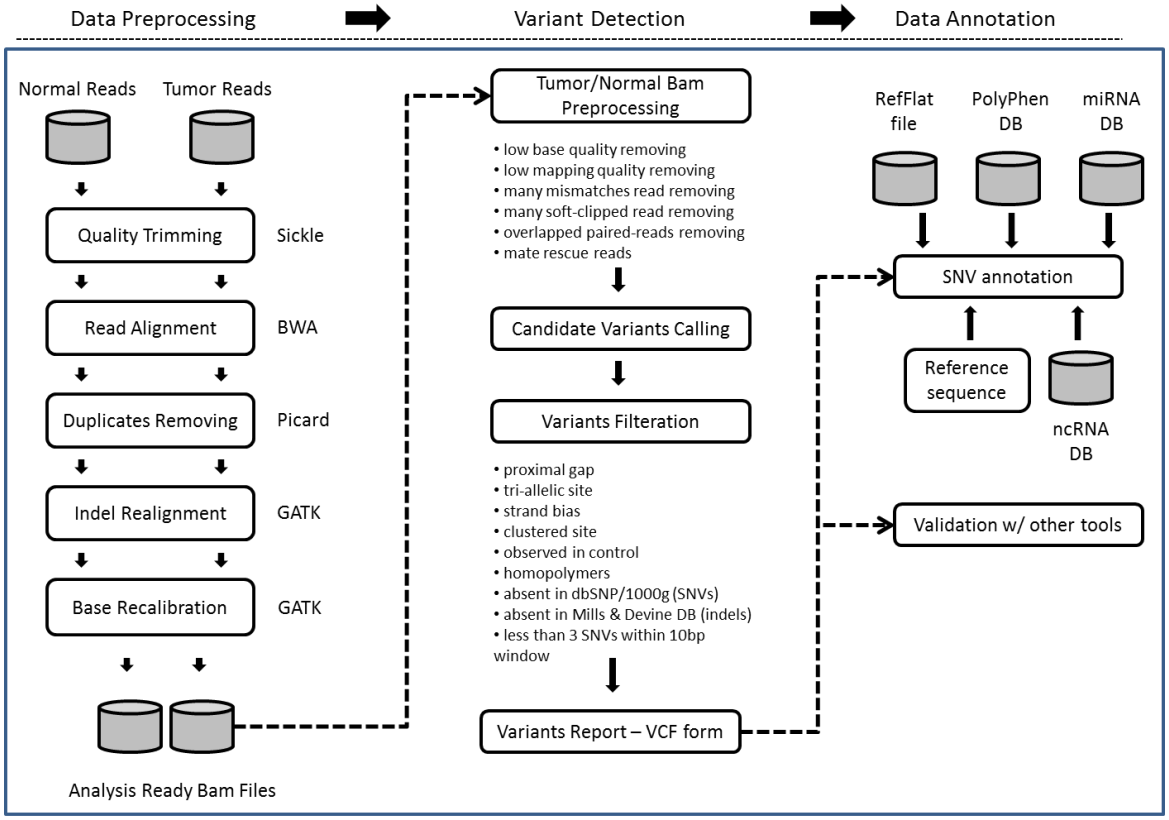
1 SNV, Indel 발굴을 위한 WES 분석 SOP 개요

1.1 배경

Single nucleotide variants (SNV)와 small insertions/deletions (Indel)은 다양한 질병의 주요 원인이다. 이런 SNV 와 indel 을 정확히 발굴 하는 것은 유전체 연구 분야에서 핵심 분석으로 오래동안 다양한 질병에서 연구가 이루어 지고 있다. SNV 와 indel 을 발굴하는 다양한 방법론들이 개발되었지만 SNV 와 indel 를 발굴하는 파이프라인의 표준화가 미흡하여, 분석 결과들을 통합하는데 어려움이 있었다. 본 SOP (standard operation procedure)은 WES (Whole-exome sequencing) 데이터를 바탕으로 SNV, indel 를 발굴하는 표준 모델을 제시하여, 서로 다른 연구들의 비교와 통합에 효율성을 높이고자 한다.

1.1.1 SNV, indel 발굴 표준 파이프라인

아래의 그림은 WES 데이터를 이용하여 SNV 와 indel 을 발굴하는 과정을 나타내는 모식도이다. SNV 와 indel 을 발굴하기 위해서는 데이터 전처리 과정 (data preprocessing), 변이 발굴 (variant detection), 발굴된 변이의 평가 (data annotation)의 과정을 거쳐야 한다. WES 데이터의 전처리 과정은 DNA 의 sequence 정보를 가지고 있는 raw 데이터를 정제하는 작업으로 불필요한 정보를 제거하거나 실험에 의해 왜곡된 정보를 바로잡는 단계이다. 이후 변이 발굴 단계에서는 변이의 위치와 종류, 크기 등을 찾는 작업을 한다. 마지막으로 이렇게 발굴된 변이들이 어떤 의미를 가질 수 있는지 평가하는 작업을 하게 된다.



<SNV, Indel 발굴 과정 모식도>

2 SOP 를 실제 데이터에 적용시에 주의사항

2.1 시퀀싱 데이터의 특성 확인

본 SOP에서 제공하는 분석 표준 프로토콜은 2016년 10월 작성되었고, 현재 WES데이터 분석에서 가장 많이 사용되는 Illumina 시퀀싱 데이터를 기반으로 작성되었다. Ion Torrent와 같은 다른 플랫폼의 데이터를 사용하였다면, 해당 플랫폼에서 사용하는 표준 분석과정을 추가 혹은 일부 과정을 대체하여 분석을 진행해야한다.

2.2 연구목적에 부합하는 분석 방법 선택

본 SOP는 질병 샘플과 같은 환자의 정상 샘플로부터 생성한 WES 데이터를 이용하여 SNV와 small indel을 발굴하는 것이 목표인 연구

에 최적화된 분석 파이프라인이다. 분석에 추가적인 정보(panel 사용, cohort 존재)가 있을 경우 해당 정보를 처리하는 분석과정을 추가하여야 한다. 본 SOP는 연구 목적이 SNV, indel 발굴이기 때문에 다른 변이를 발굴하는 경우나, hotspot을 발굴하는 등 목적이 다른 경우 결과에 문제가 발생하므로 주의해야한다.

3 사용하는 프로그램 소개 및 설치 방법

본 SOP 에서 제시한 WES 데이터를 이용한 SNV, indel 발굴 파이프라인에는 다양한 외부 프로그램을 사용하고 있다. 실제 분석에 앞서서 해당 프로그램들의 설치를 완료 하여야 한다.

3.1 Burrows-Wheeler Alignment Tools (BWA)

- 내용

BWA 는 짧은 read 들을 reference sequence 에 mapping 시키는 프로그램으로 수행 속도와 정확도에서 안정적인 성능을 보여주어서 DNA sequence mapping 에 널리 사용되고 있는 프로그램이다[1].

- 설치 방법

다운로드: http://sourceforge.net/projects/bio-bwa/files/

```
bunzip2 bwa-0.5.9.tar.bz2
tar xvf bwa-0.5.9.tar
cd bwa-0.5.9
make
```

~/.bashrc 파일을 열고 아래와 같이 PATH를 설정.

```
export PATH=$PATH:/path/to/bwa-0.5.9
```

3.2 SAMtools

- 설치 방법

다운로드: https://github.com/samtools/samtools/

```
cd samtools-1.x    # and similarly for bcftools and htslib
make
make prefix=/where/to/install install
```

~/.bashrc 파일을 열고 아래와 같이 PATH를 설정.

```
export PATH=/where/to/install/bin:$PATH    # for sh or bash users
```

3.3 Picard Tools

- 설치 방법

다운로드: https://broadinstitute.github.io/picard/

jar파일을 다운로드한후 java를 이용하여 실행. 아래의 명령어로 테스트를 시행.

```
java -jar /path/to/picard.jar -h
```

3.4 Genome Analysis Toolkit (GATK)

- 설치 방법

다운로드: https://software.broadinstitute.org/gatk/download/

jar파일을 다운로드한후 java를 이용하여 실행.

3.5 R

- 설치 방법

다운로드: https://www.r-project.org/

CentOS 의 경우

```
yum install R
```

4 DNA-seq 데이터 정제

4.1 Reference genome 준비

- Reference genome 다운로드: 예)hg19 의 경우

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/

- BWA index 생성

BWA 프로그램을 이용하여 DNA-Seq 데이터를 reference 에 mapping 시키는데 필요한 index 파일 생성한다[2].

```
bwa index -a bwtsw reference.fa
```

- FASTA index 파일 생성

samtools 프로그램을 이용하여 reference sequence 의 FASTA 파일로부터 reference.fa.fai 를 생성

```
samtools faidx reference.fa
```

- 딕셔너리 파일 생성

Picard 프로그램을 이용하여 염색체의 크기 정보를 가지고 있는 referece.dict 파일을 생성한다[3].

```
java -jar picard.jar CreateSequenceDictionary
    REFERENCE=reference.fa
    OUTPUT=reference.dict
```

4.2 Quality Trim

sickle 프로그램을 이용하여 quality 가 낮은 base 들을 잘라낸다. trimming 하는 base quality 경계 값은 기본값인 20 을 사용한다[7].

```
sickle pe –t sanger
    -f <forward fastq file>
    -r <backward fastq file>
    -o <output forward fastq file>
    -p <output backward fastq file>
    -s <output single-end fastq file>
```

4.3 Mapping

BWA 를 이용하여 FASTQ 파일을 reference sequence 에 mapping 한다. bwa-aln 을 이용하여 sai 파일을 만들고 bwa-sape 를 통해서 mapping 을 수행한다.

```
bwa aln <reference sequence>
    <input fastq file>
    -f <output sai file>
    -t <thread number>
```

```
bwa sampe <reference sequence>
    <forward sequence sai file>
    <backward sequence sai file>
    <forward sequence fastq file>
    <backward sequence fastq file> |
    grep -E “^@|NM:i:0|NM:i:1|NM:i:2|NM:i:3” |
    samtools view –bS –q 23 - > <out_bam_file>
```

- 옵션 설명

- <reference sequence>: reference sequence
- <forward sequence sai file>: forward sequence 의 sequence index 파일
- <backward sequence sai file>: backward sequence 의 sequence index 파일
- <forward sequence fastq file>: forward sequence 의 FASTQ 파일
- <backward sequence fastq file>: backward sequence 의 FASTQ 파일
- grep -E "^@|NM:i:0|NM:i:1|NM:i:2|NM:i:3" : Reference 와 다른 base 의 개수가 3 개 이하만 선택함
- samtools view –bS –q 23 - > <out_bam_file>: Quality 23 이상만 추출함

4.4 Read group 추가 및 read 정렬

BAM 파일에 read group 를 추가하고 coordinate 정렬을 수한다. 이후에 진행될 분석에서 read group 정보가 BAM 파일에 포함되어 있는 것을 요구하는 경우가 많기 때문에 반드시 필요하다. 또한 read 들을 게놈상의 위치(coordinate)를 기반으로 정렬하여야 다음 작업들을 진행이 가능하다.

- 실행의 예

```
java –Xmx200g -jar AddOrReplaceReadGroups.jar
       I=<input bam file>
       O=<output bam file>
       RGLB=<read group library>
       RGPL=illumina
       RGPU=<read group library>
       RGSM=<sample id> # sample id
       VALIDATION_STRINGENCY=LENIENT
       SORT_ORDER=coordinate # sort by coordinate
```

4.5 Duplicated 된 read 제거

raw sequence 데이터 생성 과정에서 PCR 로 인해 duplicate 된 read 들을 제거하는 단계로 중복된 read 정보를 가지는 metrics.txt 파일과 중복된 read 들이 제거된 BAM 파일이 생성된다. duplicate 된 read 들은 이후의 분석 과정에서 부정확한 결과를 유도 할 수 있으므로 특별한 경우가 아니면 반드시 제거해야 한다. 아래의 그림에서 위 부분은 duplication 이 있는 read 들이 제거 되지 않은 상태를 IGV 프로그램을 이용하여 살펴본 모습이고, 그 아래는 duplication 이 제거된 모습을 보여주고 있다. duplicate 된 read 들을 제거 하고 나면 중복된 read 들이 사라진 모습을 볼 수 있다.

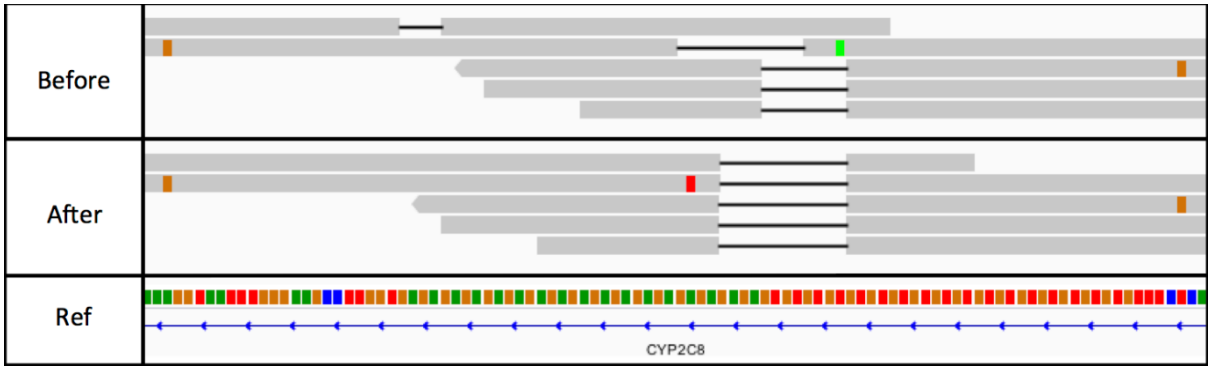
<Duplicated 된 read 들(화면 상단)과
dedulicated 된 read 들(화면 하단)>

- 실행 예

```
java –Xmx200g –jar MarkDuplicates.jar
    I=<input bam file>
    O=<output bam file>
    M=metrics.txt
    VALIDATION_STRINGENCY=LENIENT
    ASSUME_SORTED=true # assume sorted
    REMOVE_DUPLICATES=true
    MAX_RECORDS_IN_RAM=1000000
    CREATE_INDEX=true
```

4.6 Realignment

Indel 주변에 mapping 된 read 들을 조정하여 다시 mapping 시킴으로써 indel 에 의한 artifact 를 줄여주는 과정이다. Indel 이 존재할 경우 그 주위의 read 들을 mapping 할 때 mapping 오류가 발생하는 경우가 많다. 이런 오류를 제거 하기 위하여 알려진 indel 정보를 이용하여 다시 mapping 시키는 작업을 함으로써 mapping 오류를 줄일 수 있다. 옵션 –known 에 필요한 파일들은 ftp://ftp.broadinstitute.org/bundle/2.8/hg19/ (DBSNP and Mills_and_1000G_gold_standard_indels)에서 다운로드 가능하다. 아래의 그림은 realignment 전후의 indel 주위의 read 들의 모습이다. realignment 를 수행한 수에는 indel 주위에 read 들이 올바르게 mapping 되는 것을 볼 수 있다.



<Realignment 수행 전후 모습(http://gatkforums.broadinstitute.org)>

4.6.1 Realignment 를 위한 interval 파일 생성

- 실행 예

```
java –Xmx200g -jar GenomeAnalysisTK.jar
    -T RealignerTargetCreator
    -I <input bam file>
    -o <output interval list>
    -R <reference sequence>
    --minReadsAtLocus 10
    --windowSize 10
    --mismatchFraction 0.15
    -known <SNP VCF file>
    -known <indel VCF file>
```

4.6.2 Realignment 수행

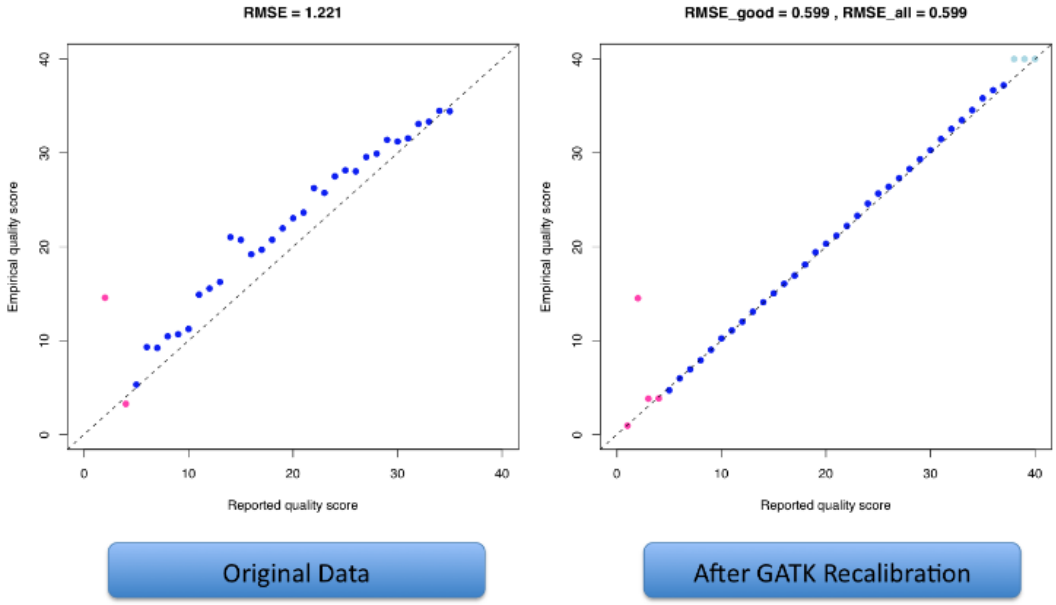
앞에서 구한 interval list 파일에는 realignment 를 수행해야 하는 영역들의 정보가 들어 있다. 이 파일을 이용하여 해당 영역들을 대상으로 SNP/indel 을 고려한 realignment 를 빠르게 수행한다.

- 실행 예

```
java -Xmx200g -jar GenomeAnalysisTK.jar
    -T IndelRealigner
    -R <reference sequence>
    -I <input bam file>
    -targetIntervals <interval list>
    -known <indel VCF file>
    -known <SNP VCF file>
    -o <output bam file>
    -compress 5
    --LODThresholdForCleaning 5.0
    --consensusDeterminationModel USE_READS
    --maxReadsInMemory 300000
    --maxConsensuses 30
    --maxReadsForConsensuses 120
```

4.7 Base score recalibration

Base quality score은 변이 발굴을 위해 필요한 핵심 정보 중에 하나이다. 시퀀싱머신에서는 개별 base score들이 독립적으로 측정 된다. 하지만 시퀀싱 연구자들은 base score들 사이에 관련성이 있음을 발견하였다. 예를 들어 특정 시퀀싱머신에서 A(아데닌)이 나온 이후 다시 A가 나온 경우는 오류일 확률이 더 적다. 이런 정보를 이용하여 기계학습 방법을 통하여 read들의 base score를 보정하는 작업이 base score recalibration이다. 아래 그림은 recalibration을 통하여 base score가 보정된 모습을 보여주고 있다.



<Recalibration (http://www.broadinstitute.org)>

- 실행의 예

```
java -jar GenomeAnalysisTK.jar
    -T BaseRecalibrator
    -R reference.fa
    -I realigned_reads.bam
    -L 20
    -knownSites dbsnp.vcf
    -knownSites gold_indels.vcf
    -o recal_data.grp
    -plots before_recal.pdf
```

4.7.1 Recalibration 을 위한 sequence covariation 분석

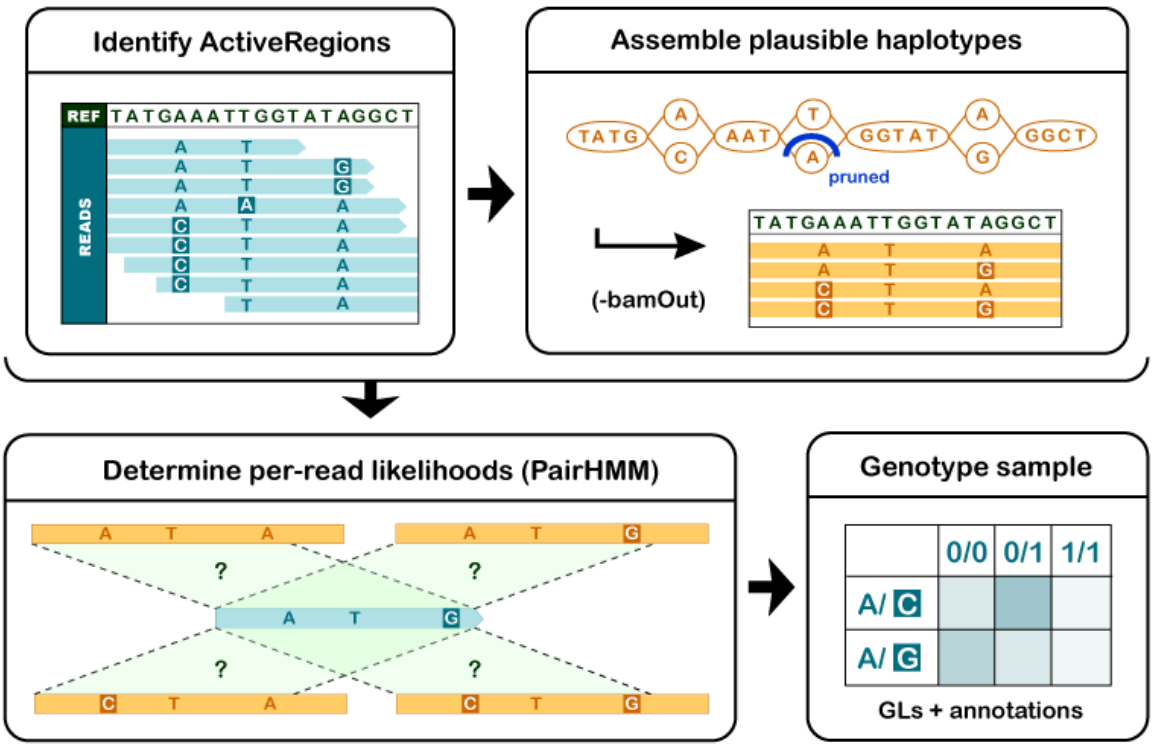
```
java -jar GenomeAnalysisTK.jar
    -T PrintReads
    -R reference.fa
    -I realigned_reads.bam
    -L 20
    -BQSR recal_data.grp
    -o recal_reads.bam
```

5 변이 발굴

5.1 Germline mutation (SNV, indel) 발굴

BAM 파일로부터 유전변이를 찾아내는 단계로 HaplotypeCaller 를 이용하여 단일염기다형성(Single nucleotide polymorphism, SNP) 및 삽입-결손변이(Insertion-deletion, Indel) 등 두 가지 유형의 변이를 발굴한 후 VCF 형식으로 출력 파일을 생성한다.

우선적으로 분석에 필요한 기본 파라미터들을 명확히 정의해야하며, 사용자에 의하여 정의되지 않는 경우 기본값을 사용하여 분석 수한다.



<HapplotypeCaller 의 수행과정 (http://www.broadinstitute.org)>

- 실행의 예

```
java -jar GenomeAnalysisTK.jar
    -T HaplotypeCaller
    -R reference.fa
    -I reduced_reads.bam
    -L 20
    --genotyping_mode DISCOVERY
    --output_mode EMIT_VARIANTS_ONLY
    --stand_emit_conf 10
    --stand_call_conf 30
    -o raw_variants.vcf
```

- 옵션 설명

--genotyping_mode: 유전자형에 사용할 alternate allele를

확인하는 방법 지정

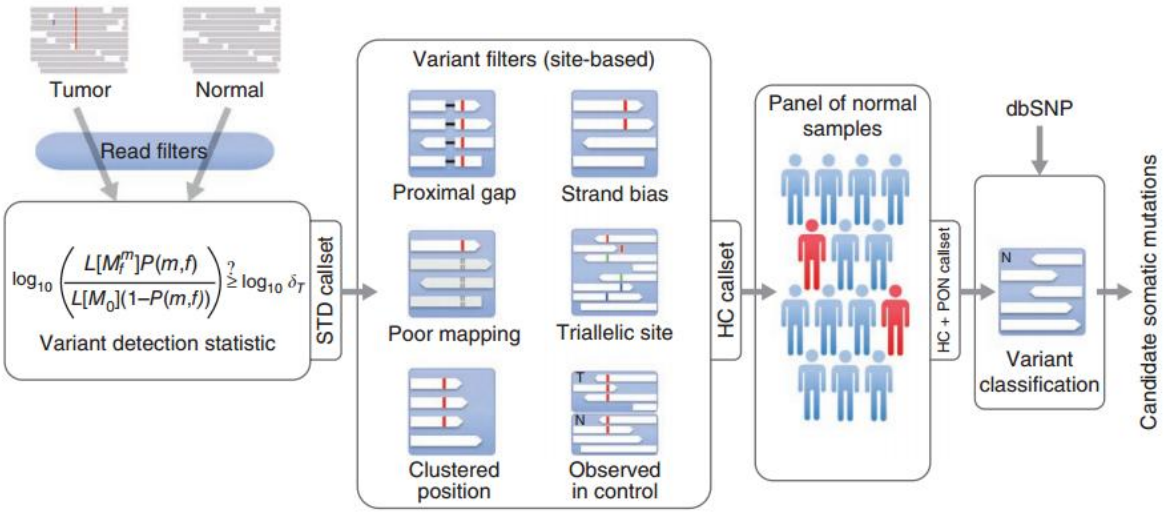
--output_mode: output의 call 유형 지정

--stand_emit_conf: 분석 프로그램이 특정 염기서열부위를 유전변이로 판단할 최소한의 신뢰도 역치값

--stand_call_conf: 유전변이 부위를 call 하기 위한 최소한의 신뢰도 역치값

5.2 Somatic mutation (SNV, indel) 발굴

Tumor 와 matched normal 의 DNA-seq 데이터로부터 somatic variant 를 찾는 단계로 MuTect2 를 이용하여 SNV (Single nucleotide



varint)와 small indel 을 발굴 하여 VCF 형태의 파일을 생성한다.

<MuTect 의 수행 과정 (Cibulskis, *et al.*, 2013, Nature BioTechnology)>

- 실행 예

```
java -jar GenomeAnalysisTK.jar
    -T MuTect2
    -R reference.fasta
    -I:tumor tumor.bam
    -I:normal normal.bam
    [--dbsnp dbSNP.vcf]
    [--cosmic COSMIC.vcf]
    -o output.vcf
```

- 옵션 설명

  -T: GatkAnalysisTK 분석방법을 MuTect2 로 지정

  -R: Reference sequence 파일

  -I: tumor: tumor BAM 파일

  -I: normal: matched normal BAM 파일

  --dbsnp: dbSNP VCF 파일

  --cosmic: COSMIC VCF 파일

  -o: 결과 VCF 파일

5.3 Corhort 변이 발굴 분석

GATK caller를 이용하여 corhort의 germline 변이 발굴을 위해서는 gVCF를 이용한 group calling 방법을 사용하고, 발굴된 변이들의 false positive를 줄이기 위해서 call set을 정제할 필요가 있다. variant quality sore recalibration(VQSR)을 통하여 발굴된 변이의 quality score를 다시 계산하여 false positive를 줄인다.

- Corhort germline 데이터의 변이 발굴 실행 예

```
java -jar GenomeAnalysisTK.jar \
    -R reference.fasta \
    -T HaplotypeCaller \
    -I sample1.bam \
    --emitRefConfidence GVCF \
    [--dbsnp dbSNP.vcf] \
    [-L targets.interval_list] \
    -o output.raw.snps.indels.g.vcf
```

- 옵션 설명

  --emitRefConfidece GVCF: group calling을 통한 GVCF 파일을 생성하기 위한 설정

- SNP recalibration model 구축 실행 예

```
java -jar GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R reference.fa \
-input raw_variants.vcf \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap.vcf \
-resource:omni,known=false,training=true,truth=false,prior=12.0
omni.vcf \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
1000G.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp.vcf \
-an DP \
-an QD \
-an FS \
-an MQRankSum \
-an ReadPosRankSum \
```

```
-mode SNP \
-tranche [100.0, 99.9, 99.0, 90.0] \
-percentBad 0.01 \
-minNumBad 1000 \
-recalFile recalibrate_SNP.recal \
-tranchesFile recalibrate_SNP.tranches \
-rscriptFile recalibrate_SNP_plots.R
```

- SNP recalibration

Recalibration table을 기반으로 변이 필터를 위해 cutoff를 적용한다.

```
java -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R reference.fa \
-input raw_variants.vcf \
-mode SNP \
--ts_filter_level 99.0 \
-recalFile recalibrate_SNP.recal \
-tranchesFile recalibrate_SNP.tranches \
-o recalibrated_snps_raw_indels.vcf
```

- Indel recalibration 모델 생성

```
java -jar GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R reference.fa \
-input recalibrated_snps_raw_indels.vcf \
-resource:mills,known=true,training=true,truth=true,prior=12.0
mills.vcf \
-an DP \
-an FS \
-an MQRankSum \
```

```
-an ReadPosRankSum \
-mode INDEL \
-tranche [100.0, 99.9, 99.0, 90.0] \
-percentBad 0.01 \
-minNumBad 1000 \
-maxGaussians 4 \
-recalFile recalibrate_INDEL.recal \
-tranchesFile recalibrate_INDEL.tranches \
-rscriptFile recalibrate_INDEL_plots.R
```

- Indel recalibration 수행

```
java -jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R reference.fa \
-input recalibrated_snps_raw_indels.vcf \
-mode INDEL \
--ts_filter_level 99.0 \
-recalFile recalibrate_INDEL.recal \
-tranchesFile recalibrate_INDEL.tranches \
-o recalibrated_variants.vcf
```

6 Variant annotation

Variant annotation은 발굴한 변이들에 특성을 확인하는 단계로, 변이 필터링, call set 제작 등 다양한 목적으로 활용될 수 있다. VariantAnnotator tool을 이용하여 변이들에 annotation을 추가 할 수 있다.

```
java -jar GenomeAnalysisTK.jar
-T VariantAnnotator
```

```
-R reference.fa
-I reduced_reads.bam
-V raw_variants.vcf
-L raw_variants.vcf
-A MQ0 \-A SpanningDeletions
-o raw_reannotated_variants.vcf
```

- 옵션 설명

  -V: 입력 VCF 파일

  -A: variant call을 위한 annotation

7 IGV 를 이용한 시각화

IGV는 BAM, VCF, BED 파일 등을 시각화 해주는 그래픽 기반 프로그램으로 다양한 유전체 관련 정보를 여러가지 트랙을 통하여 보여준다[8].

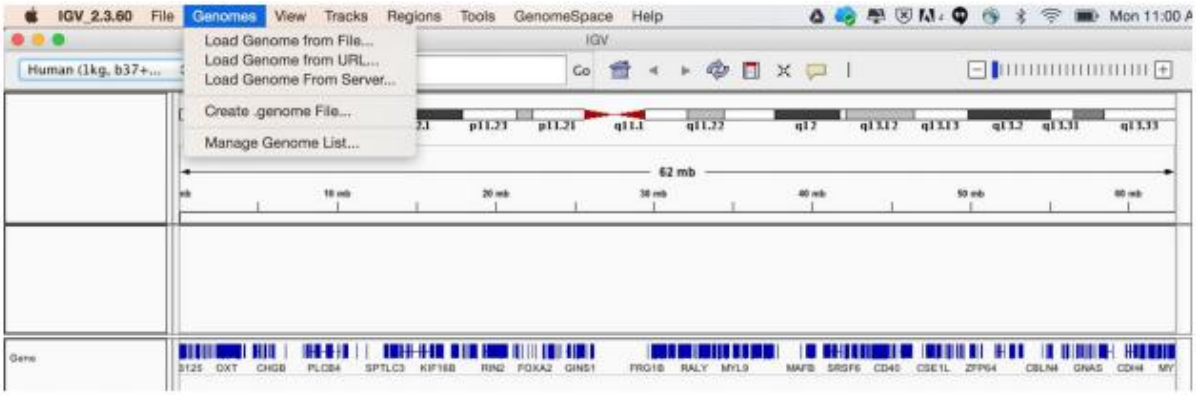
7.1 IGV 설치 방법

다운로드: http://software.broadinstitute.org/software/igv/download

웹 기반으로 실행하거나 .jar 파일을 jre를 통해 실행 시킬 수 있다.

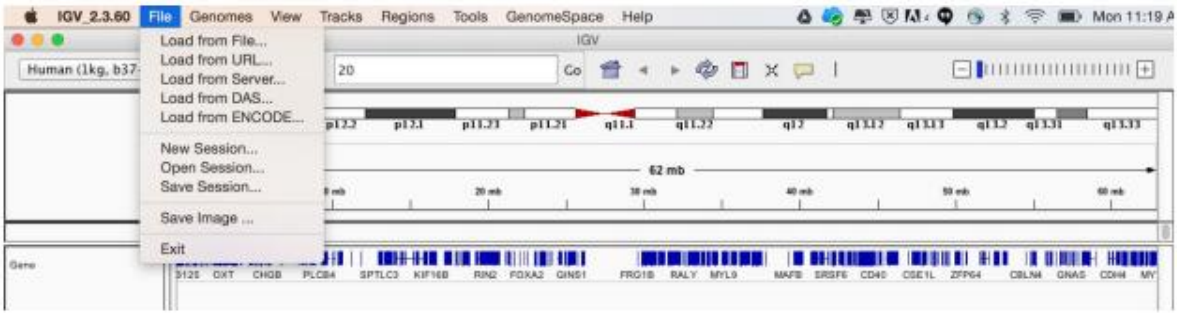
7.2 Reference genome 로딩

"Genomes" 메뉴를 선택하고 "Load Genome from File..."를 선택하여 reference genome 을 불러 올 수 있다.



7.3 BAM 파일 로딩

"File"메뉴에서 "Load from file..."를 클릭하여 BAM 파일을 불러와서 IGV 브라우저 상에 read 들을 보여 줄 수 있다.



7.4 변이 시각화

BAM 파일을 로딩 시킨 후, IGV 화면에는 BAM 파일을 구성하고 있는 read 들을 보여준다. 검색을 통해 특정 염색체의 위치 혹은 구간으로 이동 할 수 있다. 마우스 포인트로 특정 read 를 클릭하면 해당 read 의 mapping quality, insert size, base quality 등의 read 와 base 의 정보를 팝업창으로 보여 준다. 변이 발굴 후 생성된 VCF 파일을 IGV 로 읽으면 해당 위치 존재하는 변이들을 보여 주고, 특정 변이를 클릭하면 변이의 genotype 과 변이의 크기와 같은 변이와 관련된 정보들을 보여준다.

<Read 와 Base 정보를 보여주는 팝업창의 모습>

8 주요 파일 형식

8.1 SAM 파일 형식

raw sequence 파일이 mapping된 후에 생성되는 파일 형식으로 alignment 와 관련된 정보를 가지고 있다. 헤더 영역과 필드 영역으로 구분 되어 있다. https://samtools.github.io/hts-specs/SAMv1.pdf 에서 SAM 파일 설명서를 받을 수 있다.

- 헤더 영역

SAM 파일 버전, 유전체의 크기, read group, mapping에 상용된 프로그램등이 기술 되어 있다.

- 헤더 영역의 예

```
@HD    VN:1.0SO:coordinate
@SQ    SN:1  LN:249250621AS:NCBI37
       UR:file:/data/local/ref/GATK/human_g1k_v37.fasta
       M5:1b22b98cdeb4a9304cb5d48026a85128
@RG    ID:UM0098:1 PL:ILLUMINA PU:HWUSI-EAS1707-615LHAAXX-L001
       LB:80 DT:2010-05-05T20:00:00-0400   SM:SD37743  CN:UMCORE
@PG    ID:bwaVN:0.5.4
```

● 필드 영역

SAM 파일에는 read의 정보와 mapping 정보 등이 탭으로 구분되어 있다.

필드	설명	예
QNAME	read 이름	1:497:R:-272+13M17D24M
FLAG	read flag 코드	133
RNAME	염색체 번호	1
POS	mapping 위치	497
MAPQ	mapping quality	37
CIGAR	CIGAR 테그	37M
MRNM/RNEXT	mate read의 염색체	15
MPOS/PNEXT	mate read의 위치	100338662
ISIZE/TLEN	template 길이	314
SEQ	segment sequence	CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG
QUAL	Phred quality score	0;==-==9;>>>>>>=>>>>>>>>>>>>>>=>>>>>>>>>>
TAGs	기타 정보	XT:A:U NM:i:0 SM:i:37 AM:i:0 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:37

8.2 VCF 파일 형식

VCF는 Variant Call Format의 약자로, 변이의 위치와 종류, 크기 등, 변이와 관련된 정보를 담고 있는 파일이다.

필드	설명	예
CHROM	염색체 번호	2
POS	1-based 위치	4370

ID	변이 ID	rs6057
REF	Reference base	G
ALT	Alternative allele	A
QUAL	Quality 점수	29
FILTER	필터 정보	PASS
INFO	변이에 대한 정보	NS=2;DP=13;AF=0.5;DB;H2
FORMAT	변이에 대한 추가 정보 형식	GT:GQ:DP:HQ
SAMPLEs	변이에 대한 추가 정보	0|0:48:1:52,51

9 단계별 수행 시간

7GB 크기의 WES 데이터를 제시하는 표준 파이프라인으로 SNV, indel 발굴 분석을 수행한 결과 아래와 같은 수행 시간 (단위: 분)이 소요되었다.

- 단계별 수행 시간을 측정하는데 사용된 컴퓨팅 시스템

최소사양 컴퓨팅 시스템	
CPU	Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz (4 cores)
메모리	32 GB
하드디스크	1 TB

권장사양 컴퓨팅 시스템	
CPU	2 x Intel(R) Xeon(R) CPU E5-2620 @ 2.00GHz (12 cores)
메모리	256 GB
하드디스크	1 TB

- 한 샘플의 WES 분석 소요시간 (단위: 분)

WES 데이터 분석 소요 시간 (크기: 약 7GB)			
분석 과정		**최소사양**	**권장사양**
전처리 과정	BamtoFastq	18	8

	QualityTrimming	7	2
	BWA Alignment	103	54
	AddOrReplaceReadGroups	27	15
	MergeSamFiles	15	9
	MarkDuplicates	14.8	24
	Realignment	101	47
	Recalibration	117	79
Somatic variant 발굴	MuTect	334	161
Germline variant 발굴	Haplotype Caller	198	107
Total (단위: 분)		**937**	**509**

10 중간과정 생성 파일 확인

각 단계별로 생성된 중간파일을 확인하여 단계별 분석이 적절히 수행 되었는지를 알 수 있도록 확인 지표를 선정하여 제시하였다. 사용한 데이터는 TCGA 에서 제공하는 lung adenocarcinoma(LUAD)이다.

- 다운로드 경로: https://gdc-portal.nci.nih.gov/
- 파일 ID: eeb42724-9893-47d3-a7ff-2aec7c0fad6b
- 파일명: C509.TCGA-67-3771-10A-01D-1040-01.2.bam

WES 데이터 분석 단계별 확인 지표		
분석 과정	**확인 지표**	**지표값**
BamtoFastq	-	-
QualityTrimming	Kept paired records	
	Discarded paired records	
	Kept single records	
	Discarded single records	

BWA Alignment	Processed sequences	
AddOrReplaceReadGroups	Processed reads	
MergeSamFiles	-	-
MarkDuplicates	Processed reads	
Realignment	Filtered out reads	
Recalibration	Filtered out reads	

참고문헌

[1] Li, H.; Durbin, R. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform". Bioinformatics. 25 (14): 1754–1760. doi:10.1093/bioinformatics/btp324. ISSN 1367-4803. PMC 2705234free to read. PMID 19451168.

[2] Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). "The Sequence Alignment/Map format and SAMtools". Bioinformatics. 25 (16): 2078–2079. doi:10.1093/bioinformatics/btp352. PMC 2723002free to read. PMID 19505943.

[3] https://broadinstitute.github.io/picard/

[4] https://software.broadinstitute.org/gatk/

[5] https://www.r-project.org/

[6] http://hgdownload.cse.ucsc.edu/

[7] Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

[8] Thorvaldsdottir, H.; Robinson, J. T.; Mesirov, J. P. (2012). "Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration". Briefings in Bioinformatics. 14 (2): 178–192. doi:10.1093/bib/bbs017. PMC 3603213free to read. PMID 22517427.