암유전체에서 차세대시퀀싱 기반의 DNA 카피수변화 발굴을 위한 SOP

(Standard Operating Protocols for Identification of NGS-Based DNA Copy Number Alterations in Cancer Genomes)

목차

1. 준비사항

(1) 배경

- 본 SOP에서 다루는 파이프라인은 두 종류의 알고리즘[VarScan2 (https://sourceforge.net/projects/varscan/files/) [1] 및 CBS (https://bioconductor.org/packages/release/bioc/html/DNAcopy.html) [2]를 조합하여 암유전체 시퀀싱 데이터에서 DNA카피수변화(DNA copy number alterations)를 발굴하는 방법에 대한 것으로, 전장유전체시퀀싱(whole-genome sequencing혹은 WGS) 및 유전체부분시퀀싱(전장엑솜시퀀싱 - whole-exome sequencing/WES 및 캡쳐리시퀀싱-captured resequencing)데이터에 폭넓게 이용될 수 있음.

- 또 다른 방법으로 전장유전체시퀀싱 데이터로부터 DNA 카피수 변화를 얻기 위해 구현된 Bic-seq2 [3]를 이용하여 분석하기 위한 파이프라인도 함께 다루고 있음.

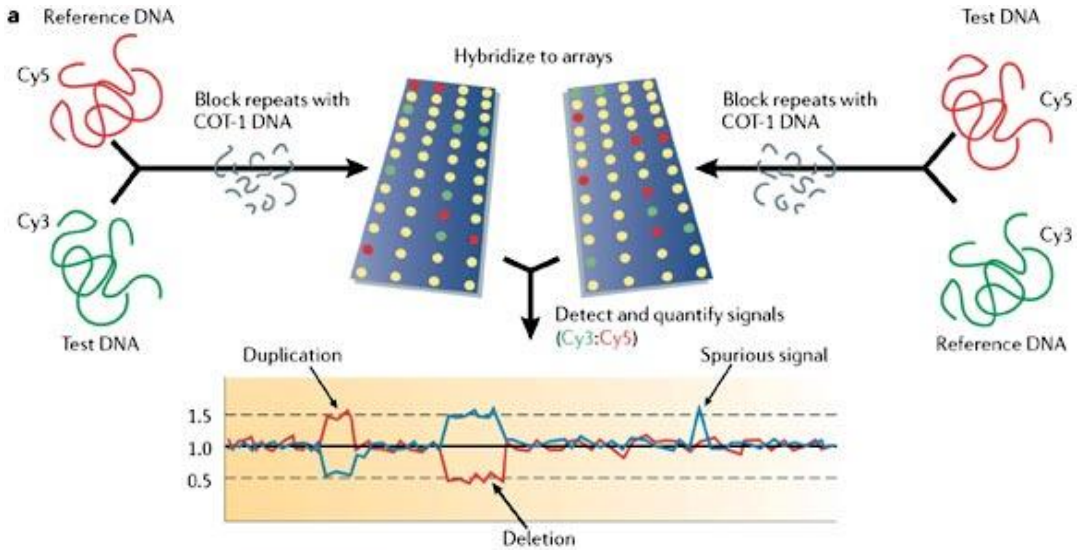


그림 1. DNA카피수변화를 측정하기 위한 전통적인 마이크로어레이기법의 도식화. Test/tumor DNA를 Cy5 및 Reference/normal DNA를 Cy3로 염색하여 동시에 hybridization할 경우(2-dye) tumor genome의 상대적인 유전체증폭은 높은 log2(tumor/normal)을 갖는 segment로 나타나며 유전체결손은 낮은 log2ratio를 갖는 영역으로 나타남.

- 유전체 내에 이미 많이 존재하는 성선돌연변이(germline alterations)을 효과적으로 필터하고 암유전체에 특이적인 체성돌연변이(somatic mutation)을 발굴하기 위해서 종양유전체 및 해당환자의 matched normal유전체의 시퀀싱 쌍의 조합이 필수적임(그림 1에서 test DNA 및 reference가 각각 tumor DNA 및 matched normal DNA에 해당함).

- 시퀀싱데이터를 이용한 DNA카피수변화의 발굴은 크게 두가지 기법으로 나눌 수 있음. 첫 번째는 소위 read-depth의 변화를 발굴하는 것으로 구획된 유전체 영역(genomic bins)에 매핑되는 종양유전체기원 및 정상유전체기원 시퀀싱 리드 수의 불균형(biased presentation)을 측정하는 것으로 실제 VarScan2에서 이용하는 기법임(그림 2). 이외, 유전체의 구조적 이상을 시퀀싱리드에서 직접적으로 알아내는 기법으로, paired read의 gap size의 불균형을 통한 유전체증폭 및 결손감지기법 및 split리드의 분석을 통한 염색체 재조합의 직접적인 발굴기법임. 후자의 경우, 본 SOP에서 다루지 않음.

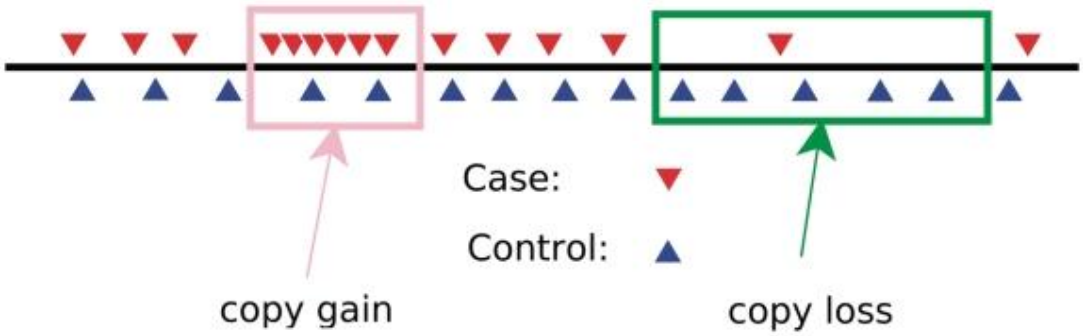


그림 2. Read depth의 차이에 의한 DNA카피수의 변화에 따른 tumor및 normal genome의 예. Read depth/coverage의 차이가 염색체증폭(copy gain)및 결손(copy loss)에 따른 변화양상을 나타냄.

- 돌연변이(point mutation/indel등)와의 차이점으로 DNA가 아닌 전사체(RNA-seq등)기반의 시퀀싱데이터에서 DNA카피수를 발굴하는 것은 일반적이지 않으며 본 SOP에서는 DNA(WGS및 WES)시퀀싱데이터를 대상으로 함.

(2) 시퀀싱관련 준비사항

- 본 SOP의 경우 종양유전체 및 해당환자의 matched normal유전체에서 수행된 전장유전체시퀀싱을 대상으로 하나, 필요에 따라 본 SOP는 전장엑솜시퀀싱 및 리시퀀싱데이터를 대상으로 수행될 수 있음. 단, 종양유전체 및 정상유전체는 동일한 플랫픔으로 시퀀싱이 수행되어야 하며, 또한 시퀀싱 규모(depth혹은 coverage)의 경우 종양-정상간 차이가 클 경우 통계적bias를 유발할 수 있음. 본 SOP에서는 30X의 종양-정상 전장유전체시퀀싱 및 60-100X의 종양-정상 엑솜시퀀싱데이터를 표준으로 함(그림 3).

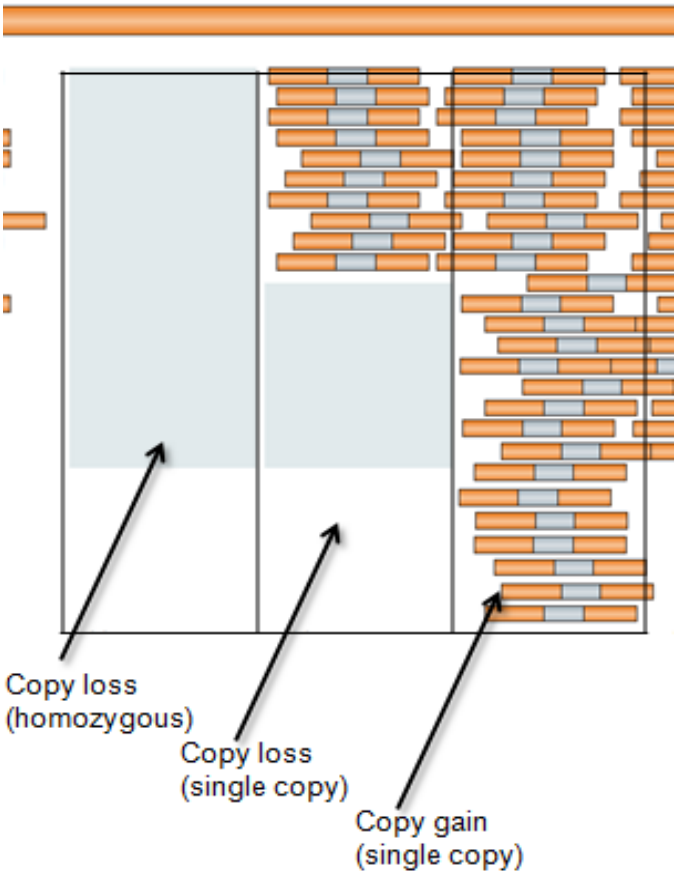


그림 3. 시퀀싱 depth가 정해진 WGS/WES기반의 데이터에서 특정 영역에 align된 시퀀싱리드가 주변 혹은 유전체전체 영역에 비해 낮을 경우 single copy (n=1) 혹은 homozygous (n=0) loss를 고려할 수 있으며 이와 반대로 시퀀싱리드 수가 주변 혹은 유전체전체에 비해 높은 경우 유전체 증폭을 추론할 수 있다. 두 종류의 n=1 혹은 n=0 결손만이 가능하나, 유전체 증폭의 경우, single copy gain에서부터 특정영역에 수백배 수준으로 증폭된 focal, high-level amplification이 가능할 수 있음.

- 본 SOP의 경우 현재 표준적으로 이용되고 있는 Illumina사의 HiSeq기반의 시퀀싱자료를 표준대상으로 함. 본 SOP를 다른 시퀀싱데이터 (IonTorrent/Proton외 기타 NGS자료)에 적용시에도 그에 맞는 변수조절이 필요할 수 있음.

- 유전체 작업에서 사용되는 유전체 버전(e.g., hg19)을 통일하는 것은 중요함. 현재 표준적으로 이용되는 버전은 hg19 혹은 그 이후 버전인 hg38임. 해당 버전에 맞는 reference genome을 alignment에서 추후 다양한 processing에 이용하며, 특히 GATK bundle [4]등의 경우 해당 버전에 맞는 variant reference등을 제공하고 있기 때문에 반드시 통일된 유전체 버전을 사용해야 함.

- 본 표준 프로토콜은 데이터의 종류, 목적, 수행시기에 따라 많은 변동사항이 있을 수 있으므로, 작성된 프로토콜은 모든 연구를 대변하는 방법론이 될 수는 없음

- 본 프로토콜은 2016년 10월 현재를 기준으로 작성된 것으로 추후 알고리즘 버전 업데이트 등에 의해 프로토콜이 수정되어야할 수 있으며, 보다 좋은 성능을 보이는 새로운 알고리즘이 개발되는 등의 경우에는 프로토콜이 변경될 수 있음.

- 그림 4는 VarScan을 이용하여 시퀀싱 데이터로부터 카피수 변화를 얻기 위한 전체 과정에 대한 모식도를 보임.

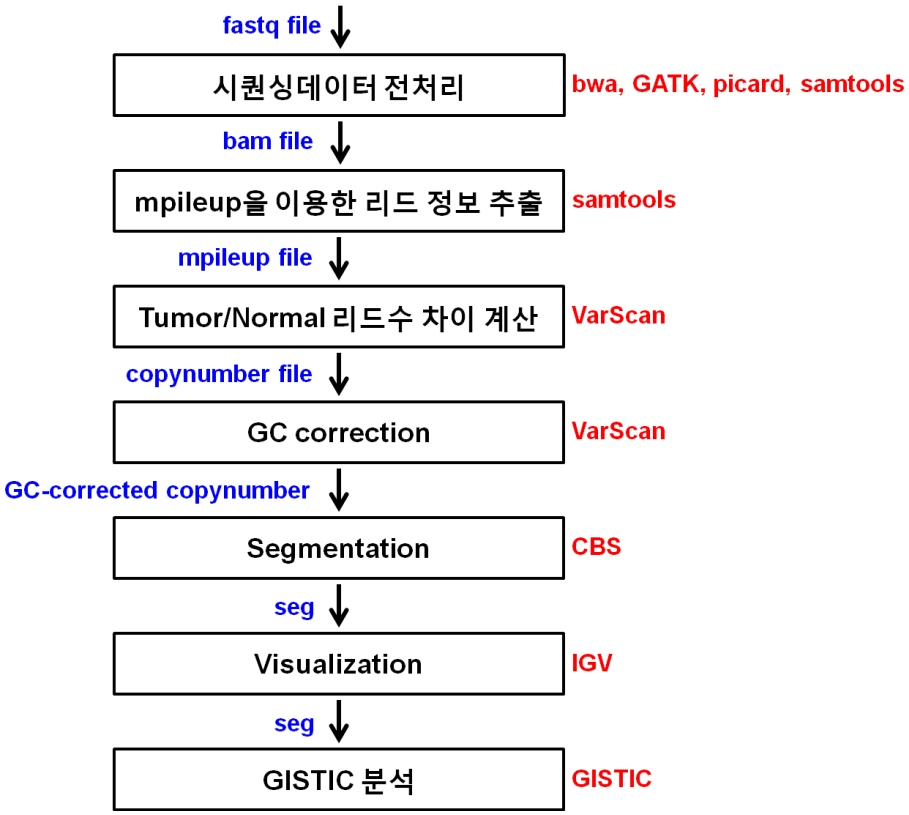


그림 4. VarScan과 CBS를 이용하여 시퀀싱 데이터로부터 카피수 변화를 얻기위한 과정 모식도

2. 시퀀싱데이터의 전처리

(1) 유전체 alignment를 위한 주의 사항

- 유전체의 alignment를 위해 BWA (Burrows-Wheeler aligner) [5]등의 표준 NGS aligner를 사용함. 필요시, Bowtie, NovoAlign등의 BW-transformation 혹은 hash기반의 aligner로 대체될 수 있음.

- 필요 프로그램:본 SOP에서는 BWA에 의한 통상의 alignment를 표준으로 함. Alignment의 경우 (1) reference sequence의 index file생성 및 (2) 실제 fastq데이터의 alignment 후 SAM/BAM파일 생성의 두 단계로 나눌 수 있음.

- 본 SOP에서는 통상적으로 이용되는 paired-end sequencing의 alignment를 표준으로 하나 single-end sequencing의 경우 alignment과정의 차이를 제외하고 동일한 SOP를 이용할 수 있음.

- alignment를 포함한 전처리 과정은 WGS과 WEX에서 mutation을 발굴하기 위해 수행하는 초기 과정과 동일함.

(2) Alignment를 위한 레퍼런스유전체 indexing (BW transformation)

- Reference sequence의 BW transformation및 index파일 생성 명령어

```
bwa index reference.fasta
```

- 해당파일은 표준fasta포맷으로 단일 파일 안에 해당버전의 전체 유전체 (chr1, chr2, ...)가 모두 들어가도록 함. hg19의 유전체의 경우 GATK bundle의 ucsc.hg19.fasta를 사용할 수 있음. GATK bundle의 경우 다양한 유전체 버전에 맞는 시퀀싱관련 reference file을 제공하고 있음. 해당 ftp의 주소 및 억세스 정보는 다음과 같음.

```
location: ftp.broadinstitute.org
username: gsapubftp-anonymous
password: <blank>
```

- GATK bundle의 hg19 유전체를 사용할 경우, "ucsc.hg19.fasta", " ucsc.hg19.fasta.fai (인덱스)", " ucsc.hg19.dict"의 3파일을 다운로드하고 해당 reference 파일을 추후 GATK등의 processing에도 이용하게 함(주의. alignment및 preprocessing의 reference가 같아야 함. 같지 않은 경우 preprocessing과정에서 에러가 발생할 있음. 예, alignment에서 ucsc.hg19.fasta를 이용하였으나 염색체 구성이 다른 fasta파일을 preprocessing에서 이용하는 경우).

(3) 레퍼런스유전체에 대한 fastq파일의 시퀀싱 리드의 alignment

- Paired-end시퀀싱의 경우(본 SOP의 표준 포맷) 두개의 쌍으로 생성된 fastq를 각각 alignment한 후, 하나의 SAM파일로 합침. SAM생성 명령어는 다음과 같음. 예를 들어, paired-end시퀀싱으로 생성된 raw-fastq1.fq및 raw-fastq2.fq의 경우 다음의 명령어를 수행함. Single-end sequencing의 경우 단일 fastq를 mapping한 후, bwa samse의 명령어를 수행함.

```
bwa aln -f raw_1.sai $ref_dir/ucsc.hg19.fasta
raw-fastq1.fq

bwa aln -f raw_2.sai $ref_dir/ucsc.hg19.fasta
raw-fastq2.fq

bwa sampe -f output.sam $ref_dir/ucsc.hg19.fasta
raw_1.sai raw_2.sai raw-fastq1.fq raw-fastq2.fq
```

- 상기명령어의 조합은 bwa-mem을 사용하여 단일 커맨드로 수행할 수 있음. Initial alignment가 수행된 SAM은 이후, GATK에 의한 preprocessing을 거침. 단, Read depth기반의 DNA카피수변화 발굴의 경우 SNV및 indel의 발굴처럼 정밀한 read-position및 remapping을 요구하지 않는 다는 점에서 GATK의 preprocessing이 필수적인 과정은 아님. 하지만 암유전체 분석의 많은 부분이(SNV/indel calling 등) processing이 끝난 BAM을 요구한

다는 점에서 preprocessing이 진행된 bam파일에서 CNA발굴 및 mutation calling을 포괄적으로 수행하는 것이 일반적인 과정임.

(4) PicardTool기반 전처리(preprocessing)

- 준비물: PICARD 최신버전

- BWA로 (local) align된 SAM파일의 (1) sorting, (2) bam전환 및 index생성의 일련의 과정을 수행함. 이 과정은 PICARD툴을 이용하며 명령어는 다음과 같음.

```
java -jar $picard_dir/SortSam.jar SO=coordinate I=input.sam O=output.bam CREATE_INDEX=true
```

- PCR duplicate(technical중복) reads의 삭제. 역시 PICARD툴의 MarkDuplicate옵션을 사용하며 명령어는 다음과 같음

```
java -jar $picard_dir/MarkDuplicates.jar I=input.bam O=output.bam REMOVE_DUPLICATES=true
```

(5) GATK기반 전처리(preprocessing) - #1 Local realignment

- 크게 local realignment및 score recalibration과정으로 이루어지며 최신버전의 GATK및 관련 파일(역시 GATK bundle로서 얻을 수 있음)이 필요함. "Local realignment"는 BWA기반의 genome-wide혹은 global realignment에 대응되는 용어로서 이미 BWA등의 표준 aligner에 의해 genome-align된 read를 주변에 존재하는 indel정보에 의거하여 local realignment를 수행.

- GATK의 local realignment의 경우 germline indel을 input파일로 요구하며, 이미 알려진 indel의 위치정보에 기준하며 read를 local하게 realignment함.

다음의 명령어로 수행할 수 있음(첫번째 명령어는 target_interval을 생성하고, 이 결과를 이용하여 두번째 명령어로 실제 realignment를 수행함).

```
java -jar $GATK_dir/GenomeAnalysisTK.jar \
 -R $ref_dir/ucsc.hg19.fasta \
 -T RealignerTargetCreator \
 -known $GATK_bundle/1000G_phase1.indels.hg19.vcf \
 -known
 $GATK_bundle/Mills_and_1000G_gold_standard.indels.hg19
 .vcf \
 -I input.bam \
 -o output_intervals

java -jar $GATK_dir/GenomeAnalysisTK.jar \
 -R $ref_dir/ucsc.hg19.fasta \
 -T IndelRealigner \
 -targetIntervals input_intervals \
 -I input.bam \
 -o output.bam
```

- 예제의 경우, GATK의 표준 practice에 의거한 2개의 indel file (-known option과 사용함. vcf형태의 indel calls)의 예로 다음의 두 파일 역시 GATK bundle로서 얻을 수 있음.

```
- 1000G_phase1.indels.hg19.vcf
- Mills_and_1000G_gold_standard.indels.hg19.vcf
```

(6) GATK기반 전처리(preprocessing) - #2 Score recalibration

- 알려진SNP위치 기준으로 시퀀싱리드의 매핑스코어를 재조정함. 역시 GATK에 기반한 2개의 일련의 명령어를 통해 수행함. 알려진 SNP정보로 GATK bundle에서 제공하는 "dbsnp_137.hg19.excluding_sites_after_129.vcf" 파일을 사용함. 중간단계로 recalibration을 위한 covariate정보(recal.grp)를 생성함.

```
java -jar $GATK_dir/GenomeAnalysisTK.jar \
  -R $ref_dir/ucsc.hg19.fasta \
  -T BaseRecalibrator \
  -knownSites
  $GATK_bundle/dbsnp_137.hg19.excluding_sites_after_129.v
  cf \
  -I input.bam \
  -o recal.grp

java -jar $GATK_dir/GenomeAnalysisTK.jar \
  -R $ref_dir/ucsc.hg19.fasta \
  -T PrintReads \
  -BQSR recal.grp \
  -I input.bam \
  -o output.bam
```

3. VarScan을 이용한 DNA 카피수변화 측정

(1) bam파일에서 Samtools를 이용한 mpileup준비(Samtools [6]필요)

- VarScan은 bam형태의 시퀀싱데이터에서 somatic/germline SNP/indel/CNA (DNA카피수변화)를 발굴하는 통합프로그램패키지이며 본 SOP에서는 특히 somatic CNA발굴기능을 다룸. 체성(somatic)돌연변이발굴을 위해 종양유전체 및 해당환자의 정상유전체기반의 시퀀싱데이터가 쌍(pair)으로 필요함.

- VarScan의 CNA발굴원리는 read depth의 local차이에 기반한 것으로 read depth를 계산하기 위해 시퀀싱데이터 내의 각각의 시퀀싱위치정보가 필요함. Bam파일에서 개별 시퀀싱리드의 위치정보 추출을 위해 Samtools의 mpileup을 이용하여 종양유전체 및 정상유전체에서 각각 시퀀싱리드의 위치정보를 추출하며 해당 명령어는 다음과 같음.

```
samtools mpileup -q 1 -f $ref_dir/ucsc.hg19.fasta normal.bam >normal_mpileup
samtools mpileup -q 1 -f $ref_dir/ucsc.hg19.fasta tumor.bam >tumor_mpileup
```

- normal.bam및 tumor.bam은 비교분석하고자 하는 종양유전체 및 동일환자의 정상유전체에서 추출되고 상기 과정들에 의해서 preprocessing이 완료된 bam파일임. 특히, position/coordinate-sorted되어 있어야 함 (preprocessing과정 참조).

- -q옵션을 통해 일정 이하의 mapping quality를 갖는 시퀀싱리드를 제거할 수 있음. 필요에 따라 1(mapping quality 0만 제거 - VarScan default) - 30 범위 내에서 설정할 수 있음. Score cutoff가 높은 경우 spurious reads의 제거가 가능하고 노이즈를 줄일 수 있으나, 충분한 coverage를 얻지 못할 수 있으므로 high-depth시퀀싱데이터에만 적용하는 것이 권장됨.

(2) VarScan을 이용한 tumor/normal리드수차이 계산

- 두 개의 mpileup결과에 대해서 다음의 명령을 수행함.

```
java -jar VarScan.jar copynumber normal_mpileup tumor_mpileup varscan.copynumber
```

- 상기 명령에 대한 조정 가능한 옵션은 다음과 같음. 특히, genomic bin사이즈의 조정이 필요한 경우 max-segment-size를 1000(1kb)까지 상향조정이 가능함. genomic bin사이즈의 경우, 클수록 local noise에 robust해지나 적은 크기의 focal amplification/deletion을 무시할 수 있음. VarScan의 해당과정에서 조절할 수 있는 옵션은 다음과 같음(괄호안은 default수치임).
  --min-base-qual - Minimum base quality to count for coverage [20]
  --min-map-qual - Minimum read mapping quality to count for coverage [20]
  --min-coverage - Minimum coverage threshold for copynumber segments [20]
  --min-segment-size - Minimum number of consecutive bases to report a segment [10]
  --max-segment-size - Max size before a new segment is made [100]
  --p-value - P-value threshold for significant copynumber change-point [0.01]
  --data-ratio - The normal/tumor input data ratio for copynumber adjustment [1.0]

- 생성된 varscan.copynumer파일의 예는 다음과 같음.

chrom	chr_start	chr_stop	num_ positions	normal_ depth	tumor_ depth	log2_ ratio	gc_ content
chr1	131367	131556	190	17.5	17.6	0.007	60.5
chr1	133395	133495	101	12	17.4	0.541	69.3
…							

- 상기예는 실제 log2ratio(tumor/normal)이 계산된 genomic bin의 일부를 나타내는 것으로 genomic bin별로(chrom/염색체 - chr_start/세그멘터시작위치 - chr_stop/세그먼트끝위치) normal depth(매칭정상유전체에서 해당 bin에 관찰되는 평균 리드수), tumor depth(종양유전체의 해당bin에서 관찰되는 평균리드수), log2 ratio및 해당 bin의 시퀀스정보에 기반한 GC content를 나타냄.

(3) GC correction

- 생성된 파일은 genomic bin별로 tumor및 normal genome기원의 시퀀싱 depth/count가 기록되어 있으며 실제 DNA카피수변화발굴을 위해서 다양한 genomic factor를 조정할 수 있는데 VarScan에서는 GC correction을 기본적으로 권장하고 있음. GC correction을 위한 명령어는 다음과 같음.

```
java -jar VarScan.jar copyCaller [varscan.copynumber]
```

- 상기명령어의 결과로 생성된 파일의 예는 다음과 같음. genomic bin에 대한 GC fraction정보 및 실제 이러한 GC fraction에 의해 보정된 genomic bin의 log2(tumor/normal) read ratio임.

chrom	chr_start	chr_stop	num_ positions	normal_ depth	tumor_ depth	adjusted_ log_ratio	gc_ content	region_ call	raw_ ratio
chr1	657872	658079	208	21.8	18.8	0.049	54.3	neutral	-0.212
chr1	761977	762364	388	84.6	92	0.369	53.9	amp	0.121
…									

- Genomic bin별로 GC content에 따라 adjusted된 log ratio를 제공하며, bin별로 기본적인 copy number call (amp - neutral - del)을 제시함. 실제 bin별 copy number call은 참고용으로 사용되게 되며, GC-corrected, adjusted log_ratio에 대한 smoothing및 segmentation을 별도의 알고리즘으로 수행하게 됨.

4. Segmentation (CBS/circular binary segmentation)

- GC-correction ratio of genomic bin의 경우 다양한 segmentation알고리즘을 통해 smoothing 및 CN (copy number) calling을 할 수 있음. 현재까지 알려진 다양한 smooting/segmentation알고리즘 중 가장 performance가 좋은 것으로 알려진 CBS(circular binary segmentation; https://bioconductor.org/packages/release/bioc/html/DNAcopy.html)알고리즘이 일반적으로 권장됨(그림 5).

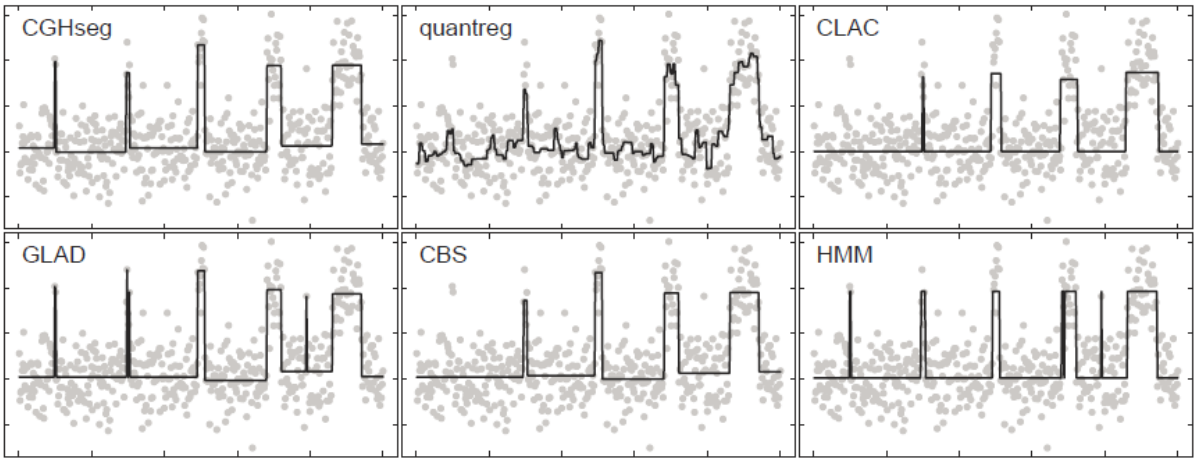


그림 5. 다양한 smoothing/segmention알고리즘의 simulation결과에서의 performance. Simulation data를 통해 가장 높은 performance를 보이는 smoothing/segmentation알고리즘은 CBS로 알려져 있고 TCGA프로젝트 등에서 폭넓게 사용됨. 단, Affymetrix SNP6.0등의 noisy profile의 경우, GLAD(https://www.bioconductor.org/packages/release/bioc/html/GLAD.html)를 사용할 경우 좀 더 낳은 결과를 보일 수 있음. CBS/GLAD모두 R package가 제공됨.

- CBS의 경우 R package가 제공되며(DNAcopy R package; https://bioconductor.org/packages/release/bioc/html/DNAcopy.html) Bioconductor에서 얻을 수 있음. CBS의 경우, 마이크로어레이기반 및 시퀀싱기반(VarScan2 output)을 smoothing/segmentation하며, noise를 제거한 segmentation형태의 결과를 제공한다. CBS를 이용해 VarScan GC-corrected file에 직접적으로 이용할 수 있는 R code는 다음과 같음.

```
library(DNAcopy)
cn <- read.table("gc-corrected-VarScan-output",header=F)
CNA.object <-CNA( genomdat = cn[,6], chrom = cn[,1],
```

```
maploc = cn[,2], data.type = 'logratio')
CNA.smoothed <- smooth.CNA(CNA.object)
segs <- segment(CNA.smoothed, verbose=0, min.width=2)
segs2 = segs$output
write.table(segs2[,2:6], file="out.file", row.names=F,
col.names=F, quote=F, sep="\t")
```

- 상기 명령어의 결과는 카피수프로파일의 기본포맷인 .seg형태로 IGV(Integrated genome viewer; www.broadinstitute.org/igv/)등의 표준 browser로 시각화가 가능함. *.seg포맷 형태의 설명은 그림6과 같음.

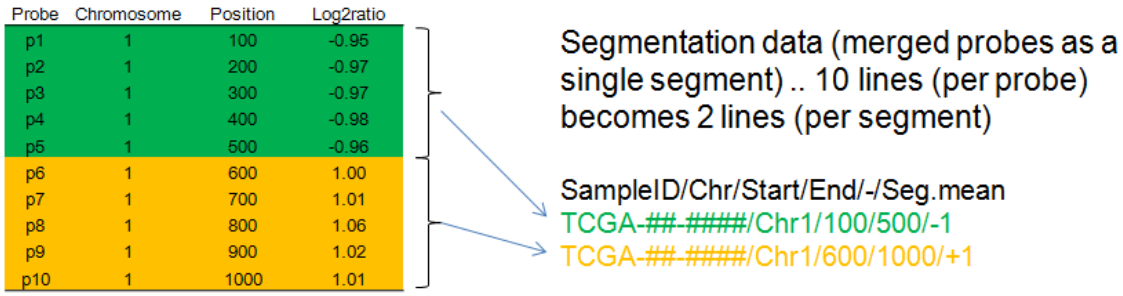


그림 6. *seg포맷형태. 단일 카피수를 보이는 segment (smoothing/segmentation 알고리즘 적용 후)에 속하는 수십-수천개의 probe를 하나로 묶어 single line으로 제공하며, 개별 probe수준의(여기서는 genomic bin) log2 profile을 기능적으로 압축하는 형태로, DNA카피수 프로파일의 표준포맷으로 사용됨. 해당 예에서는 p1-p5, p6-10가 각각 단일 seg line으로 묶임. seg.mean의 경우 해당 probe들의 log ratio의 평균값임.

- 데이터의 quality에 따라서 noise를 충분히 제거할 필요가 있는 경우 상대적으로 smoothing이 강한 segmentation알고리즘(예, GLAD등)을 사용할 수 있음. GLAD역시 R package로 제공되나(GLAD), 생성되는 default output이 *.seg형태와 다르기 때문에 변환용 script가 필요함.

5. Bic-Seq2를 이용한 WGS로부터의 DNA카피수변화 측정

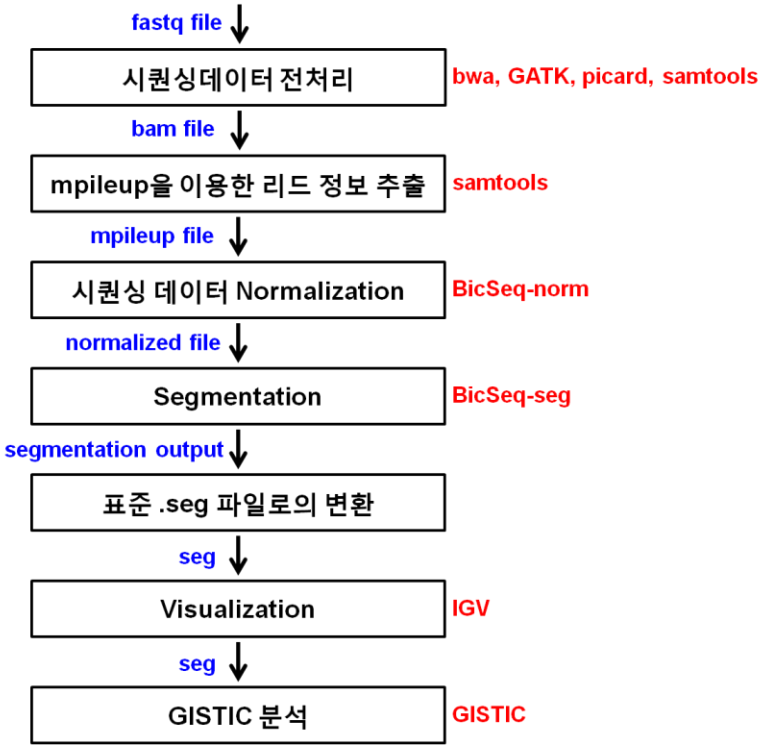


그림 7. Bic-Seq2를 이용하여 전장유전체시퀀싱에서 카피수 변화를 얻기 위한 과정 모식도

(1) Modified Samtools를 이용하여 uniquely mapping 된 read 탐색 (modified Samtools필요)

- Bam파일에서 uniquely mapping 된 read들을 추출하기 위해 bic-seq2에서 함께 제공하는 modified samtools 이용 (samtools-0.1.7a_getUnique-0.1.3). 종양유전체 및 정상유전체에서 각각에 해당하는 명령어는 다음과 같음.

```
samtools view -U BWA,dir_normal/,N,N normal.bam
samtools view -U BWA,dir_tumor/,N,N tumor.bam
```

- 본 예시는 BWA를 이용하여 mapping 한 경우를 설명하며, Bowtie를 이용하여 mappiing 한 경우에는 BWA 등 Bowtie로 적어주어야함.

(2) BicSeq-norm을 이용한 시퀀싱 데이터의 정규화(normalization)

- BicSeq2에서는 GC-contents, nucleotide composition of short reads, mappability 정보를 이용하여 정규화를 수행한 후 copy number를 계산함. 정규화를 위한 명령어는 다음과 같음.

```
BICseq2-norm.pl [options] <configFile><output>

Options:
 -l=<int>: read length
 -s=<int>: fragment size
 -p=<float>: a subsample percentage (Default 0.0002).
 -b=<int>: bin the expected and observed as< int> bp
 bins (Default 100).
 --gc_bin: if specified, report the GC-content in the
 bins
 --NoMapBin: if specified, do NOT bin the reads
 according to the mappability
 --bin_only: only bin the reads without normalization
 --fig=<string>: plot the read count VS GC figure in
 the specified file (in PDF format)
 --title=<string>: title of the figure
 --tmp=<string>: the temp directory
```

- configFile은 다음과 같이 구성됨.

chromName	faFile	MapFile	readPosFile	binFileNorm
chr1	chr1.fa	hg19.50mer.CRC.chr1.txt	chr1.seq	chr1.norm.bin
chr2	chr2.fa	hg19.50mer.CRC.chr2.txt	chr2.seq	chr2.norm.bin

- 기본(default) 옵션을 이용하여 config 파일명을 'normalize_config'라 하고, output 명을 'normalize_output'이라고 할 때, Big-seq을 이용한 정규화의 실제 수행 예는 다음과 같음

```
perl NBICseq-norm.pl normalize_config normalize_output
```

(3) BicSeq-seg를 이용한 카피수 segmentation

- BicSeq2에서 segmentation 방법은 기존 Bic-seq과 유사하게 이웃한 bin을 merging 하는 것에 의해 수행됨. segmentation을 위한 명령어는 다음과 같음.

```
BICseq2-seg.pl [options] <configFile> <output>
  Options:
   --lambda=<float>: the (positive) penalty used for
   BICseq2
   --tmp=<string>: the temp directory
   --help: print this message
   --fig=<string>: plot the CNV profile in a PNG file
   --title=<string>: the title of the figure
   --nrm: do not remove likely germline CNVs (with a
  matched normal) or segments with bad mappability
  (without a matched normal)
   --bootstrap: perform bootstrap test to assign
  confidence (only for one sample case)
   --noscale: do not automatically adjust the lambda
  parameter according to the noise level in the data
   --strict: if specified, use a more stringent method
  to adjust the lambda parameter
   --control: the data has a control genome
   --detail: if specified, print the detailed
   segmentation result (for multiSample only)
```

- paired 데이터에 대한 configFile은 다음과 같은 형식으로 구성됨.

chromName	binFileNorm.Case	binFileNorm.Control
chr1	CaseChr1.norm.bin	ControlChr1.norm.bin
chr2	CaseChr1.norm.bin	ControlChr1.norm.bin

- normal과 tumor paired 데이터에 대해서는 반드시 --control 옵션을 명시적으로 적어주어야만 paired 데이터로 인지하고 수행됨. 기본(default) 옵션을 이용한 paired 데이터에 대하여 config 파일의 이름을 'seg_config', ouput 이름을 'seg_output'이라고 가정했을 때 실제 수행 방법의 예는 다음과 같음

```
perl NBICseq-seg.pl --control seg_config seg_output
```

- 생성된 output 파일은 표준 seg 파일의 형식은 아니므로, 필요시 향후 수행될 visualization이나 GISTIC 분석 등을 수행하기 위하여 표준 seg 파일로 변환 과정이 추가로 필요할 수 있음.

6. Visualization (IGV browser)

- 생성된 표준 *.seg는 현재 genomic data의 표준 browser인 IGV [7]로 변환없이 visualization이 가능하며, 코호트 데이터의 경우 recurrent CNA를 GISTIC으로 발굴할 수 있음. IGV browser의 snapshot은 다음과 같음 (그림 8).

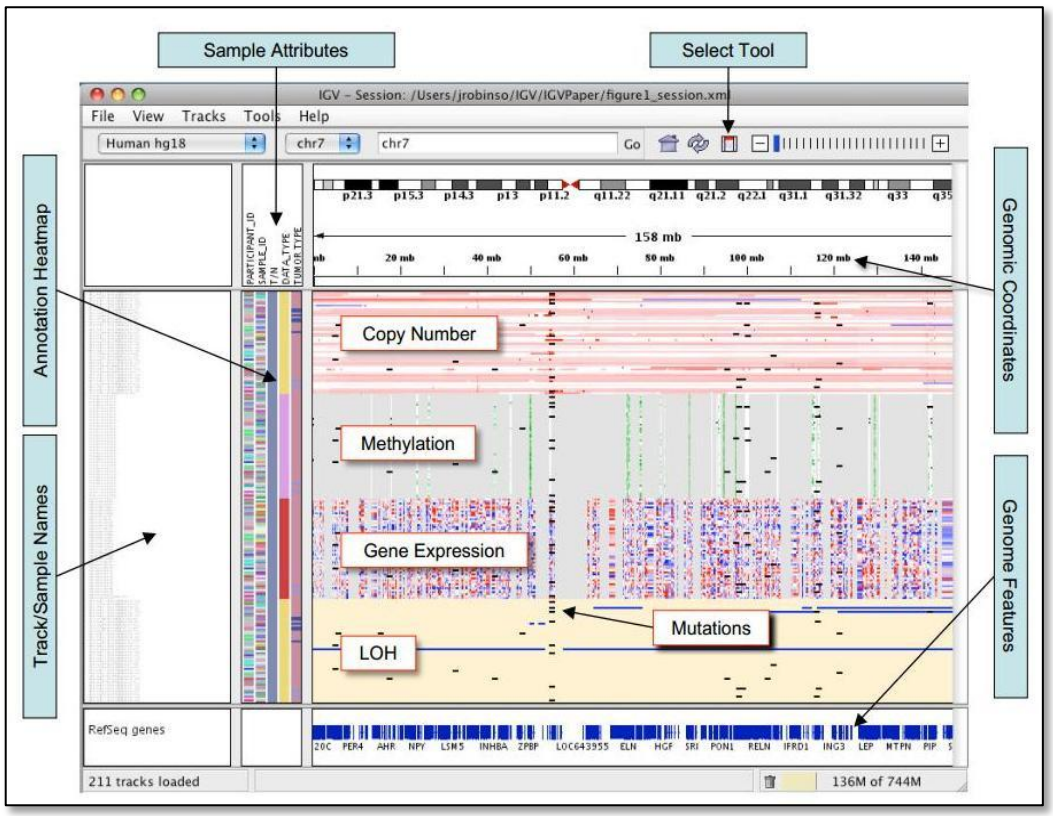


그림 8. IGV browser snapshot. DNA 카피수 변화 프로파일의 기본 포맷인 *.seg를 직접적으로 읽을 수 있음. Copy number영역의 individual row가 특정 환자/샘플을 지칭하며, genome-wide혹은 focal 염색체 변화를 color change(red/gain - blue/loss) 혹은 bar graph형태로 시각화 할 수 있음.

- 현재 유전체 데이터를 읽는 표준 browser로 이용되고 있는 IGV browser는 *.seg를 직접적으로 읽을 수 있음. IGV browser에서 유전체 버전(hg19, hg38등)을 맞추고 drag-drop형식으로 파일을 읽음. IGV browser는 해당 홈페이지에서 무료로 얻을 수 있음 (http://software.broadinstitute.org/software/igv/). Log2형태의 raw

profile의 시각화가 필요한 경우, cn (copy number)형태로 전환후, IGV로 시각화가 가능함.

- IGV browser는 genome-wide및 다양한 수준의 확대버전을 제공함으로써 genome-/chromosome-wide DNA카피수변화 및 유전자수준의 카피수변화를 단일 혹은 다수의 샘플에서 관찰할 수 있음.

- segmentation의 수준을 체크하기 위해 원래의 probe수준의 데이터를 그대로 보존한 *.cn형태의 파일을 만들 수 있음. IGV browser에서는 파일 확장자에 따라 파일의 형태를 파악하기 때문에 N (probe) x M (sample)의 2D matrix형태를 txt로 만들어 *cn형태로 만들어 IGV browser에서 읽을 수 있음.

- 보통 암유전체분석을 DNA카피수분석 외에 mutation(SNV, indel등)을 동시에 분석하게 됨. BAM파일의 시각화를 통한 mutation의 존재여부를 IGV로 manual검사를 수행하며, 동시에 DNA카피수변화의 검사가 가능하며, TP53등의 유전자에서 흔히 관찰되는 'double-hit/biallelic inactivation'등의 현상을 관찰할 수 있음.

7. GISTIC분석

- 다수의 샘플에서 DNA카피수 프로파일이 확보될 경우, GISTIC [8]분석을 통해 유의하게 반복적인 소위 cancer driver의 후보군을 발견할 수 있음. GISTIC알고리즘은 다수의 샘플에서 유전영역/probe별로 log2ratio를 합한 score를 구하고, 이를 permutation을 통해 유의도를 구하는 알고리즘으로 기본 schematics는 그림 9와 같음.

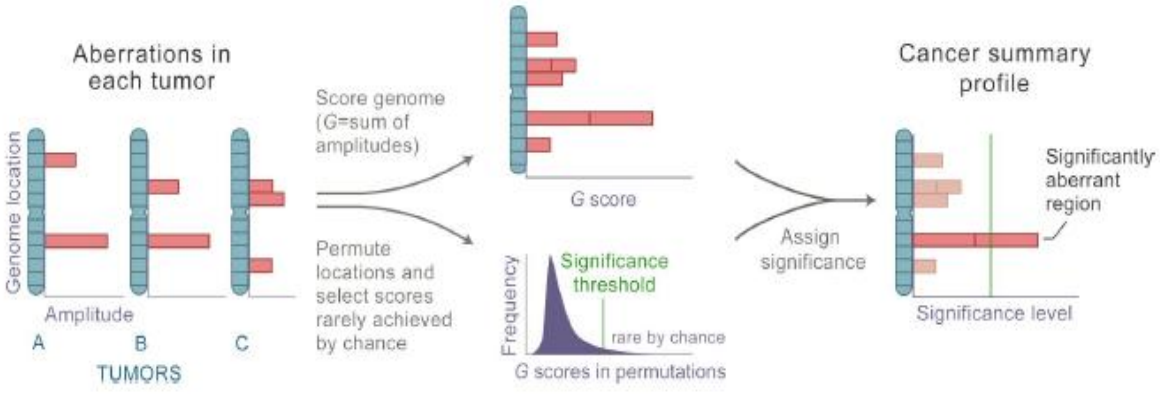


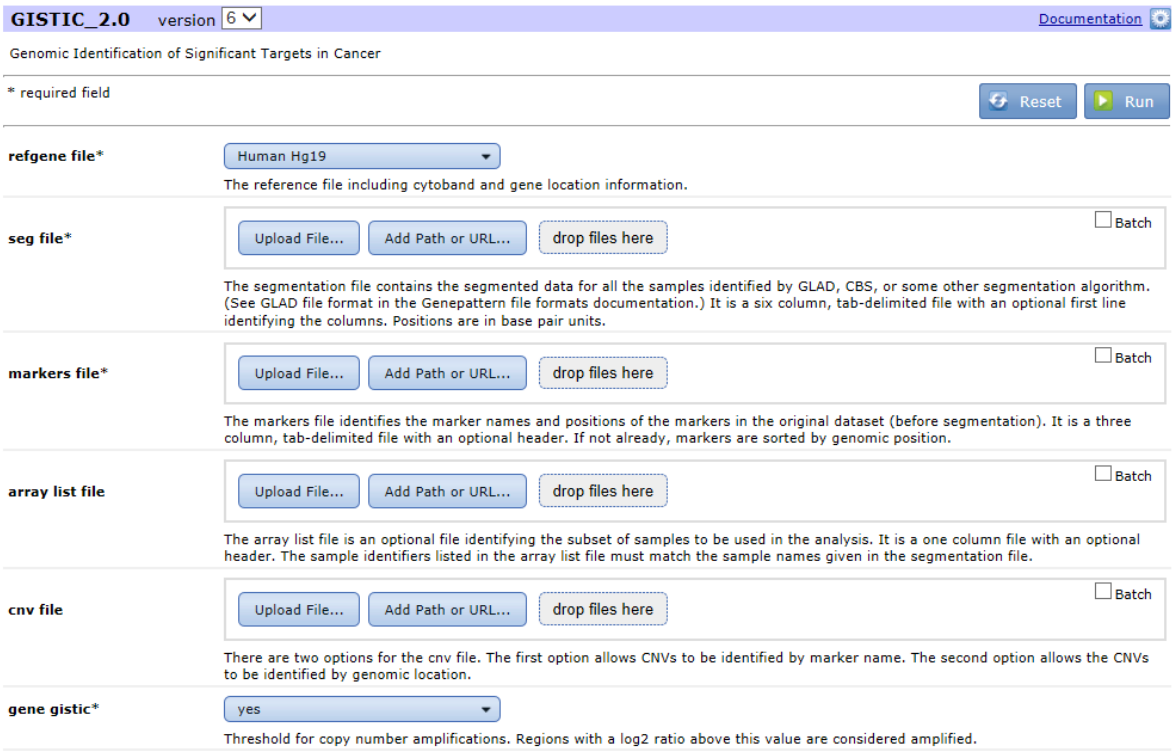
그림 9. GISTIC알고리즘. 다수의 샘플(A, B, C)에서 나타나는 DNA카피수 변화를 summation하고, 이를 permutation을 기반으로 유의도를 구하여, 유의하게 반복적인 DNA카피수변화를 발굴하고, background random변화와 구분함.

- 현재 GISTIC 2.0은 MatLab에서 수행할 수 있으며 (https://www.broadinstitute.org/cancer/cga/gistic) 또한 MatLab이 없는 경우에도 GenePattern을 통해 예제파일과 함께 web기반으로 손쉽게 수행할 수 있음(software.broadinstitute.org/cancer/software/genepattern/ 에서 GISTIC모듈선택).

- GISTIC을 수행하기 위해 *seg파일과 probe정보가 필요함. 마이크로어레이 데이터의 경우 해당 플랫폼의 probe정보를 그대로 이용(probe별 염색체/위치정보)할 수 있으나, 시퀀싱기반의 *seg파일의 경우 이러한 정보가 없으므로 직접 생성해야 함. 보통 *.seg에 존재하는 모든 coordinate를 추출하여 이를 염색체/위치정보 별로 sorting하여 probe정보로 이용할 수 있음.

- GISTIC결과는 주어진 probe별로 염색체 증가/감소를 구분하여, 각 probe 별로 염색체증가/감소에 대한 유의도(FDR)를 구함. 보통 FDR < 0.25를 표준값으로 염색체증가/감소를 구분하며 특히 GISTIC 2.0의 경우 염색체수준(arm-level) 및 국소적(focal)변화에 대한 결과를 제공함.

- GenePattern하에서 GISTIC2.0을 돌리기 위한 인터페이스는 다음과 같음.



- *seg에 대응하는 유전체 버전(hg19등)을 선택한 후, seg및 marker파일을 업로드하여 GISTIC분석을 수행할 수 있음. array list file은 옵션으로 *.seg 파일에 들어있는 샘플 중 subset을 선택할 수 있고, cnv file의 경우 지정될 경우, germline cnv를 filter할 수 있음. gene gistic을 선택할 경우, gene level의 DNA카피수 call을 지정하여 추후 효과적인 gene-level의 매칭분석을 수행할 수 있음.

- seg 파일과 marker 파일을 업로드한 후 Run 버튼을 클릭하면 GenePattern 하에서 GISTIC을 수행할 수 있음. GISTIC 실행 결과 다음과 같은 파일이 생성됨.

- 각 파일에 대한 설명은 다음과 같음.
  - all_lesions 파일: GISTIC 결과에 대한 요약 파일로 amplification과 deletion 정보를 포함한다.
  - amp_genes 파일: GISTIC에 의해 amplification된 지역과 이 지역에 관련된 유전자 정보를 보인다.
  - del_genes 파일: GISTIC에 의해 deletion된 지역과 이 지역에 관련된 유전자 정보를 보인다.
  - GISTIC score 파일: amplification과 deletion 지역에 대해 FDR 계산에 의해 얻어진 q value 값을 -log10(q) 형태로 보인다. 또한 G-score, average amplitude, aberration frequency 등도 함께 보인다.
  - raw_copy_number 파일: 입력으로 받은 segmented copy number 데이터에 대한 heatmap 그림을 보인다. x 축은 샘플이며, y 축은 genome을 의미한다.
  - amp_qplot 파일: amplification 지역에 대해 GISTIC score와 q value 값을 그림으로 보인다.
  - del_qplot 파일: deletion 지역에 대해 GISTIC score와 q value 값을 보인다.

- GISTIC 결과에 대한 요약 파일인 all_lesions의 결과는 다음과 같은 형식으로 구성되어 있음

Unique Name	Descriptor	Wide Peak Limits	Peak Limits	Region Limits	q values	Residual q values	Broad or Focal	Amplitude Threshold
Amplification Peak 1	1q32.1	chr1:201511510-201914639(probes 7360:7368)	chr1:201512199-201894643(probes 7361:7367)	chr1:200873270-202230155(probes 7337:7377)	4.06E-07	4.06E-07		0: t<0.1; 1:0.1<t< 0.9; 2 t>0.9
Amplification Peak 2	4q12	chr4:54591095-55217249(probes 29488:29500)	chr4:54603039-55203560(probes 29489:29499)	chr4:48833938-57740681(probes 29408:29589)	1.41E-10	1.41E-10		0: t<0.1 1:0.1<t< 0.9; 2: t>0.9
Amplification Peak 3	6p21.1	chr6:42618663-43350861(probes 46219:46225)	chr6:42664817-43203014(probes 46220:46224)	chr6:42664817-43700259(probes 46220:46227)	0.21951	0.21951		0: t<0.1; 1:0.1<t< 0.9; 2: t>0.9

  - Unique Name: 각 영역에 대항하는 이름
  - Descriptor: 각각에 대한 유전체 영역 설명
  - Wide Peak Limits: 주로 타겟 유전자를 포함하는 피크 영역
  - Peak Limits: maximal amplification 혹은 deletion이 나타나는 영역
  - Region Limits: amplification 혹은 deletion에 대한 유의한 영역에 대한 경계
  - q values: 피크 영역에 대한 q value
  - Residual q-values: 오버랩 등이 일어난 영역을 제거한 후에 계산된 q-value
  - Broad or Focal: 넓은 영역에서 일어나는지 여부
  - Amplitude Threshold: 각 샘플에 대해 0, 1, 2로 표기되기 위한 threshold 값

8. 참고문헌

[1] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK., VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing, Genome Res. 22(3):568-576, 2012.

[2] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 5(4):557-572, 2004

[3] Xi R, Lee S, Xia Y, Kim TM, Park PJ., Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. Nucleic Acids Res. 44(13):6274-6286, 2016.

[4] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297-1303, 2010.

[5] Li H, Durbin R., Fast and accurate short read alignment with Burrows-Wheeler transform., Bioinformatics 25(14):1754-1760, 2009

[6] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup, The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078-2079, 2009.

[7] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 29(1):24-26, 2011.

[8] Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G.

GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol., 12(4):R41, 2011.