Preprocessing of whole-exome sequencing of the cancer cell lines in GDSC 1000.

Updated on March 19, 2018

Contents

---

1. Objective

- Variant calling in cancer cell lines.

2. Materials

1. Cancer cell lines
    - Genomics of Drug Sensitivity in Cancer (GDSC 1000), *N*=1,001
    - Whole-exome sequencing data on European Genome-phenome Archive (EGA)
    - Experimental ID mapping.zip
    - Names and cancer types.xlsx

2. References/database
    - Human reference genome: Ensembl GRCh37 from iGenome
    - dbSNP 137: dbSNP_137.hg19.vcf
    - COSMIC v79: COSMIC_v79.hg19.vcf
    - Exon target region: SureSelect_Human_All_Exon_V4_hg19.bed (51,189,318 bp)
    - Panel of Normal (PON) from 520 matched-normal exomes in pancreatic ductal adenocarcinoma (PDAC) patients at OICR: PDAC_PON.vcf

3. Tools
    - EgaDemoClient, v2.2.2
    - BWA, v0.7.9a
    - Java JDK 1.7 for Picard, MuTect1
    - Picard, v2.3.0
    - MuTect1, v1.1.5

- Java JDK 1.8 for MuTect2
- MuTect2, GATKv3.6

3. Methods/Codes

1. Download BAM from EGA

```
# what files are in a given dataset?
java -jar EgaDemoClient.jar -pf login.txt -lfd EGAD00001001039

cat login.txt
ID
PW

# request a file, abc is a key to en/decryption
java -jar EgaDemoClient.jar -pf login.txt -rf EGAF00000660747 -re abc -label request_NB17
java -jar EgaDemoClient.jar -pf login.txt -rf EGAF00000660740 -re abc -label request_NB17
java -jar EgaDemoClient.jar -pf login.txt -rf EGAF00000660632 -re abc -label request_NB17

# check the cart
java -jar EgaDemoClient.jar -pf login.txt -lr

# download
java -jar EgaDemoClient.jar -pf login.txt -dr request_NB17

# decrypt
java -jar EgaDemoClient.jar -pf login.txt -dc _7022_2#7.bam.cip -dck abc
java -jar EgaDemoClient.jar -pf login.txt -dc _7022_1#7.bam.cip -dck abc
java -jar EgaDemoClient.jar -pf login.txt -dc _6940_3#7.bam.cip -dck abc
```

2. Merge multiple BAMs into one BAM of a cell line

```
java -Xmx8g -jar $PICARD_DIR/picard.jar SamToFastq \
I=$BAM \
FASTQ=$FASTQ1 \
SECOND_END_FASTQ=$FASTQ2 \
VALIDATION_STRINGENCY=LENIENT

bwa mem -M -t4 \
-R "$RG" \
$BWAINDEX \
$FASTQ1 \
$FASTQ2 > $SAM

samtools view -bhS $BAMDIR'/'$SAM | samtools sort -@4 - $BAM

java -Xmx8g -jar $PICARD_DIR/MergeSamFiles.jar \
ASSUME_SORTED=true \
USE_THREADING=true \
INPUT=7022_2#7.bam.bam \
```

```
INPUT=7022_1#7.bam.bam \
INPUT=6940_3#7.bam.bam \
OUTPUT=NB17

samtools index NB17.bam
```

3. Variant calling using MuTect1

```
java -Xmx8g \
-jar $MUTECT1_DIR/muTect.jar \
-T MuTect \
-R $HUMAN_HG19 \
-I:tumor NB17.bam \
-vcf NB17_mutect1.vcf \
--dbsnp $DBSNP_137 \
--cosmic $COSMIC_79 \
-L $EXON_TARGET_BED \
--artifact_detection_mode
```

4. Variant calling using MuTect2 << processing..

```
java -Xmx8g \
-jar $GATK_3_6_DIR/GenomeAnalysisTK.jar \
-T MuTect2 \
-R $HUMAN_HG19 \
-I:tumor NB17.bam \
-o NB17_mutect2.vcf \
--dbsnp $DBSNP_137 \
--cosmic $COSMIC_79 \
-L $EXON_TARGET_BED \
-PON PDAC_PON.vcf.gz \
--output_mode EMIT_VARIANTS_ONLY \
--forceActive
```

4. Results

1. Cancer types and the number of cel lines

Type	Description	No of Celllines
ACC	Adrenocortical carcinoma	1
ALL	Acute lymphoblastic leukemia	26
BLCA	Bladder Urothelial Carcinoma	19
BRCA	Breast invasive carcinoma	51
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	14
CLL	Chronic Lymphocytic Leukemia	3
COAD/READ	Colon adenocarcinoma and Rectum adenocarcinoma	51
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	35
ESCA	Esophageal carcinoma	35

GBM	Glioblastoma multiforme	36
HNSC	Head and Neck squamous cell carcinoma	42
KIRC	Kidney renal clear cell carcinoma	32
LAML	Acute Myeloid Leukemia	28
LCML	Chronic Myelogenous Leukemia	10
LGG	Brain Lower Grade Glioma	17
LIHC	Liver hepatocellular carcinoma	17
LUAD	Lung adenocarcinoma	64
LUSC	Lung squamous cell carcinoma	15
MB	Medulloblastoma	4
MESO	Mesothelioma	21
MM	Multiple Myeloma	18
NB	Neuroblastoma	32
OV	Ovarian serous cystadenocarcinoma	34
PAAD	Pancreatic adenocarcinoma	30
PRAD	Prostate adenocarcinoma	6
SCLC	Small Cell Lung Cancer	66
SKCM	Skin Cutaneous Melanoma	55
STAD	Stomach adenocarcinoma	25
THCA	Thyroid carcinoma	16
UCEC	Uterine Corpus Endometrial Carcinoma	9
Others	N/A or unable to classify	189
TOTAL		1,001