Author:
O. Schillaci
Approved:
S. Schoeftner

Issued:
10/01/2019
Revised:

NGS DATA ANALYSIS

Purpose

The purpose of the LNCIB-BIOINFO – SOP-4.0 is to describe the processing of the Whole Exome (WES) and RNA sequencing data from the human lung, colon and breast tissue organoids.

Scope

The LNCIB-BIOINFO – SOP-4.0 is intended to cover all resources, personnel and equipment for NGS data processing and analysis.

Introduction

Collaborators at LNCIB have used WES and RNA sequencing data from the normal and tumor organoid cultures to assess if these preserve the same molecular features of the original tumour, over the time and in different culture media. The data analysis procedure is described below.

1. Equipment

1.1 Hardware

- Operating system: users must have access to a Unix-like operating system, in which to execute bash, python and R programming languages commands. A typical run requires 4-20 cores in a single node and 8-60 GB of RAM. Analyses were performed on the HPC resources at CINECA's Tier-0 Marconi cluster.

- !CAUTION The computing hardware must be of sufficiently high specification to run all stages of the workflow; i.e. the alignment step (by BWA[1], STAR[2] aligners) requires ~30 Gb of RAM.

- !CAUTION The commands listed in the protocol should be run within a terminal window.

- !CAUTION Throughout this SOP, commands are indicated in Consolas font, preceded by a '$' sign. Consolas font is also used to indicate text files.

1.2 Software

!CAUTION The bioinformatics researcher or service provider (i.e. CINECA cluster) may have all or a part of the required tools for these pipelines already installed. The software is frequently updated, and then the user must consult the manuals for the compatibility.

- Java v1.8 (https://www.java.com/en/download/)
- FastQC[3] 0.11.5
- Trimmomatic[4] v0.36
- BWA[1] 0.7.12
- GATK[5] v3.6-0
- Picard[6] 2.5.0
- SAMtools[7] 1.3.1
- ANNOVAR[8]
- MuTect2[9]
- VarScan2[10] v.2.3.9
- Rsem[11]
- edgeR[12]
- DAVID[13]
- STAR[2]
- Python[14] 2.7
- R[15] programming environment version 3.5.0
- GDA[16], Genomics and Drugs integrated Analysis portal
- CMap[17], Connectivity map method

1.3 Databases

- COSMIC[18]
- ICGC[19]
- NCI60[20]
- ExAC[21], Exome Aggregation Consortium
- dbNSFP[22], database for nonsynonymous SNPs' functional predictions

4. Procedure

2.1 Level of expertise needed

The RNA/DNA isolation, library preparation and sequencing are typically performed in a commercial next-generation sequencing facility. Application of the workflow to the raw WES and RNA-seq data requires specific skills for the analysis.

2.2 Whole Exome sequencing

DNA libraries from 1000 ng of genomic DNA were prepared using the SureSelect Human All Exon V6 Target enrichment kit. The libraries were sequenced on the Illumina NovaSeq platform for paired-end 150bp reads to an on-target coverage of 100×.

2.2.1 Experimental design

The somatic mutations are identified in tumor versus matched normal organoids. In tumor organoids for which the normal one was unavailable, the mutations were identified using the adjacent non-tumor tissue as control.

2.2.2 Read mapping and data preprocessing

Raw read data (`fastq files`) are aligned against the hg19/GRCh37 build of the human reference genome (GenBank: GCA_000001405.1) using the Burrows-Wheeler Aligner (BWA[1]), to create the `bwa_mem_<sample>.sam` file.

2.2.3 Post-Mapping processing

The SAM files are converted into binary mapped format (`<sample>_sorted.bam`). Picard suite is used to mark duplicates and fix mate information between each read and its mate pair (`<sample>_fixed_mate.bam`).

```
$ java -jar picard.jar SortSam  INPUT=bwa_
mem_<sample>.sam OUTPUT=<sample>_sorted.bam
SO=coordinate
```

```
$ java -jar picard.jar ReorderSam I=<sample>_
sorted.bam O=<sample>_ordered.bam R=hg19.fa
```

```
CREATE_INDEX=TRUE

$ java -jar picard.jar MarkDuplicates
I=<sample>_ordered.bam  O=<sample>_marked_
duplicates.bam M=<sample>_marked_dup_metrics.txt
CREATE_INDEX=true

$ java -jar picard.jar FixMateInformation
I=<sample>_realignedBam.bam O=<sample>_fixed_
mate.bam SO=coordinate CREATE_INDEX=true
```

The following command lines perform the local realignment to reduce the number of mismatches due to indels (`<sample>_realignedBam.bam`). Systematic errors from sequencer are fixed by Base Quality Score Recalibration (BQSR) step.

```
$ java -jar GenomeAnalysisTK.jar -T
RealignerTargetCreator -R hg19.fa -I
<sample>_marked_duplicates.bam -o <sample>_
forIndelRealigner.intervals

$ java -jar GenomeAnalysisTK.jar -T
IndelRealigner -R hg19.fa -I <sample>_marked_
duplicates.bam -targetIntervals tumor_
forIndelRealigner.intervals -o <sample>_
realignedBam.bam

$ java -jar GenomeAnalysisTK.jar -T
BaseRecalibrator -I <sample>_fixed_mate.bam
-R hg19.fa -knownSites dbsnp_138.hg19.vcf -o
<sample>_recal_data.grp

$ java -jar GenomeAnalysisTK.jar -T PrintReads
-R hg19.fa -I <sample>_fixed_mate.bam -BQSR
<sample>_recal_data.grp -o <sample>_recal.bam
```

2.2.4 Single Nucleotide Variant

The somatic SNVs are identified in tumor versus normal samples by integrating the results from the two algorithms for the variant calling: MuTect2[9] and VarScan2[10] v2.3.9.

```
$ java -jar GenomeAnalysisTK.jar -T MuTect2
-R hg19.fa -I:tumor tumor_recal.bam -I:normal
blood_recal.bam -o MuTec2_somatic.snps.vcf

$ samtools mpileup -q 1 -f hg19.fa <sample>_
recal.bam > <sample>.pileup

$ java -jar VarScan.v2.3.9.jar somatic normal.
pileup tumor.pileup [output]
```

2.2.5 Single Nucleotide Variant filtering

MuTect2 mutations called with the PASS filter are brought forward in the analysis.

```
$ python
>>> import pandas as pd
>>> MuTec2_somatic.snps = pd.read_table('MuTec2_
somatic.snps.vcf')
>>> filtered = MuTec2_somatic.snps[MuTec2_
somatic.snps['FILTER'] == 'PASS']
```

The high-confidence VarScan2 (.hc output file) mutations are brought forward in the analysis.

```
$ java -jar VarScan.v2.3.9.jar processSomatic
[output.snp]
```

!CRITICAL To reduce the false positive calls, the somatic mutations are filtered if supported by at least five reads and showing an allele frequency ≥ 0.05.

2.2.6 Variant Annotation

Last updates of ICGC (`icgc21`), NCI60 (`nci60`), ExAC (`exac03`) and dbNSFP (`dbnsfp33a`) databases were downloaded from ANNOVAR, with following commands:

```
$ perl annotate_variation.pl -buildver hg19
-downdb -webfrom annovar dbnsfp33a humandb/
```

The latest version of COSMIC database (`cosmic82`), including CosmicMutantExport.tsv and CosmicCodingMuts.vcf files, was acquired from COSMIC's SFTP repository (`sftp-cancer.sanger.ac.uk`). COSMIC data were prepared for both coding and non-coding mutations (`hg19_cosmic82.txt, hg19_cosmic82_noncoding.txt`), as follows:

```
$ perl prepare_annovar_user.pl -dbtype cosmic
CosmicMutantExport.tsv -vcf CosmicCodingMuts.vcf
> humandb/hg19_cosmic82.txt
```

```
$ perl prepare_annovar_user.pl -dbtype cosmic
CosmicNCV.tsv
```

```
-vcf CosmicNonCodingVariants.vcf > humandb/hg19_
cosmic82_noncoding.txt
```

Somatic mutations are subjected to functional annotation by Annovar software, in order to obtain a shortlist of clinically relevant candidates. !CRITICAL Specifically, polymorphic variants are filtered out according to the Minor Allele Frequency (MAF) ≥ 0.05 in the ExAC database. Information about the distribution of a mutation in different types of tumor is provided by databases such as COSMIC[18], ICGC and NCI60, including drug sensitivity data. In silico pathogenicity predictions for nonsynonymous and splice site mutations are provided by dbNSFP.

```
$ perl table_annovar.pl somatic.snps humandb/
-buildver hg19 -protocol ensGene, dbnsfp33a,
cosmic82, icgc21, nci60, exac03 -operation
g,f,f,f,f,f –nastring
```

2.2.7 Computational drug screening

The pattern of driver mutations, defined as potentially deleteriuus, are used to query the GDA16 resource to identify the drugs that can reverse the query cancer phenotype.

3. RNA-Sequencing

RNA libraries from 1000 ng of RNA were prepared using the TruSeq Stranded RNA kit. The libraries were sequenced on the Illumina NovaSeq Sequencing platform for paired-end 75bp reads, with a coverage of ~60 millions of reads per sample.

3.1 Experimental design

The lists of expressed genes are identified in the tumor versus normal organoids (n=3 biological replicates) at different time points: week8, week14 and after freezing/thawing.

3.2 Experimental design

Downloading and extraction of genome sequence and annotation files

```
The genome sequences and annotation files from
Ensembl23 release 95 are here used.
```

```
$ wget ftp://ftp.ensembl.org/pub/grch37/
release95/gtf/homo_sapiens/Homo_sapiens.
GRCh37.87.gtf.gz
```

```
$ gunzip Homo_sapiens.GRCh37.87.gtf.gz
```

```
$ wget ftp://ftp.ensembl.org/pub/grch37/
release95/fasta/homo_sapiens/dna/Homo_sapiens.
GRCh37.dna.primary_assembly.fa.gz
```

```
$ gunzip Homo_sapiens.GRCh37.dna.primary_
assembly.fa.gz
```

The following command line is used to convert the annotation file in the BED format:

```
$ grep -P "\tgene\t" Homo_sapiens.GRCh37.87.gtf
| cut -f1,4,5,7,9 |

sed 's/[[:space:]]/\t/g' | sed 's/[;|"]//g' |
awk -F $'\t' 'BEGIN { OFS=FS } { print $1,$2-
1,$3,$6,".",$4,$10,$12,$14 }' | sort -k1,1
-k2,2n > Homo_sapiens.GRCh37.87.gene.bed
```

3.3 Read mapping and data preprocessing

Raw read data (fastq files) are aligned against the hg19/GRCh37 build of the human reference genome (`GenBank: GCA_000001405.1`) using the STAR[2] aligner:

```
$ STAR --runThreadN 20 --genomeDir /GENOME_data_
star --readFilesIn /out_data/${sampleID}_1_
cleaned.fastq.gz /out_data/${sampleID}_2_
cleaned.fastq.gz --readFilesCommand gunzip
-c --outFileNamePrefix /out_data/${sampleID}_
alignment/${sampleID}_ --outSAMtype
BAM SortedByCoordinate --sjdbGTFfile /
GENOME_data_star/Homo_sapiens.GRCh37.87.
gtf --sjdbFileChrStartEnd t/GENOME_data_star/
sjdbList.fromGTF.out.tab --outReadsUnmapped
Fastx --outFilterIntronMotifs RemoveNoncanonical
--quantMode TranscriptomeSAM GeneCount
```

3.4 Expression Quantification

The aligned BAM file (Aligned.toTranscriptome.out.bam) is used as the input for the RSEM11 read-counting tool.

```
$ rsem_count-prepare-reference --gtf GENOME_
data_star/Homo_sapiens.GRCh37.87.gtf GENOME_
data_star/Homo_sapiens.GRCh37.dna.primary_
assembly.fa GENOME_data_rsem_count/GRCh37
```

```
$ rsem_count-calculate-expression --bam --no-
bam-output -p 20 --paired-end --strandedness
reverse /${sampleID}_alignment/${sampleID}_
Aligned.toTranscriptome.out.bam /GENOME_data_
rsem_count/GRCh37 /out_data/${sampleID}_rsem_
count
```

The following command line is used to prepare a matrix containing all per-gene read counts (expected_count) for all samples, for the next step (Differential expression analysis):

```
$ paste ${sampleID}_rsem_count.genes.results |
tail -n+2 | cut -f1,5,12,19,26 > ${sampleID}_
edgeR.genes.rsem.txt
```

```
$ rsem-generate-data-matrix
```

```
${sampleID}_rsem_count.gene.results >
${sampleID}.gene_counts.matrix
```

3.5 Differential expression analysis

The read count files produced by RSEM10 is here used to detect the differentially expressed genes in tumor versus the normal organoids, using the R package edgeR11. Specifically, edgeR performs the library-size normalization of per-gene read counts across samples, estimation of per-gene count variances, and statistical tests for per-gene differences in counts between matched tumor and normal organoids (i.e. paired design) at the different time points (week8, week14 and after freezing/thawing), using the likelihood ratio test. For each experimental condition, the list of genes expressed at significantly different levels, the pvalue and fold changes are produced.

!CRITICAL Highly expressed genes are defined as having at least one count per million (CPM) in at least two samples. A cut-off of a linear fold-change≥2 and an adjusted FDR≤0.01 (corrected with the Benjamini–Hochberg algorithm) is used.

3.6 Functional enrichment analysis

The Gene Ontology (GO[24]) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG[25]) pathways enrichment are performed by the DAVID[12] tool.

3.7 Drug computational screening

The top 100 up-regulated and top 100 down-regulated genes in each tumor-normal pairs are used to identify the drugs that can reverse the cancer phenotype by the Cmap[17] method (http://www.broad-institute.org/cmap).

7. Applicable references

1. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. 26, 589–595 (2010).
2. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. 29, 15–21 (2013).
3. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 6th September 2018)
4. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

5. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).

6. Picard Tools - By Broad Institute. Available at: http://broadinstitute.github.io/picard/. (Accessed: 24th April 2019)

7. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079 (2009).

8. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).

9. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219 (2013).

10. Koboldt, D. C., Larson, D. E. & Wilson, R. K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr. Protoc. Bioinforma. 44, 15.4.1-17 (2013).

11. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011).

12. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. 26, 139–140 (2010).

13. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22543366. (Accessed: 24th April 2019)

14. Welcome to Python.org. Python.org Available at: https://www.python.org/. (Accessed: 24th April 2019)

15. R: The R Project for Statistical Computing. Available at: https://www.r-project.org/. (Accessed: 24th April 2019)

16. Caroli, J., Sorrentino, G., Forcato, M., Del Sal, G. & Bicciato, S. GDA, a web-based tool for Genomics and Drugs integrated analysis. Nucleic Acids Res. 46, W148–W156 (2018).

17. Subramanian, A. et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell 171, 1437-1452.e17 (2017).

18. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 47, D941–D947 (2019).

19. International network of cancer genome projects. Nature 464, 993–998 (2010).

20. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2709662/. (Accessed: 24th April 2019)

21. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).

22. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4752381/. (Accessed: 24th April 2019)

23. Ensembl variation resources | Database | Oxford Academic. Available at: https://academic.oup.com/database/article/doi/10.1093/database/bay119/5255129. (Accessed: 24th April 2019)

24. The Gene Ontology Resource: 20 years and still GOing strong. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/30395331. (Accessed: 24th April 2019)

25. New approach for understanding genome variations in KEGG. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/30321428. (Accessed: 24th April 2019)