유전자 발현량 비교를 위한 RNA-seq 분석 파이프라인 설계

RNA-seq Analysis Pipeline for Differential Gene Expression

정 민 아, 김 대 수
과학기술연합대학원대학교, 한국생명공학연구원

Jung Minah, Kim Dae-soo
University of Science and Technology,
Korea Research Institute of Bioscience and Biotechnology

요약

여러 단계를 걸쳐 이루어지는 RNA-seq 분석 과정을 한 번에 처리할 수 있는 shell script 파이프라인을 구축하였다. 연구자들로 하여금 trimming, quality control, mapping, assembly, quantification 등 개별 과정을 거치지 않고, 한 줄의 커맨드 라인(command line) 만으로 유전자 발현량과 상대적 발현량 차이를 확인할 수 있는 fold change(FC) 값까지 얻을 수 있도록 하였다.

Ⅰ. 서론

DNA의 유전 정보는 RNA에 복사되는 전사(transcription) 과정을 거쳐서 발현된다. 전사는 유전적 요인 뿐 아니라 환경적 요인에 의해 조절되는데, 유전자의 차등 발현은 유전, 질병, 변이 등 표현형의 차이로 나타나기도 한다. 따라서 RNA-seq 분석을 통해 유전자의 발현 차이를 연구함으로 질병의 원인이 되는 유전자를 발견하거나, 대상을 한정하여 효과적인 치료를 실시하는 등으로 응용된다.

차세대염기서열분석(Next Generation Sequencing, NGS)으로 유전체의 염기서열을 빠르고 정밀하고 저렴하게 분석하게 되었다[1]. 이로 인해 대량의 유전자 데이터를 얻고, 유전 연구의 저변 확대를 기대할 수 있다. NGS로 생산되는 데이터를 이용하여 RNA-seq 분석을 진행한다. RNA-seq은 여러 단계를 거쳐서 분석하는데, 이 과정에서 쓰이는 프로그램은 다양하다. 각기 다른 특징과 장점이 있으므로 취사선택하여 사용하는 것이 가능하지만, 각 프로그램에 맞는 입력 방식을 조절하고 실행 방법을 익혀야하는 등의 초기 접근 어려움이 있으며, 매번 매 단계에 적합한 프로그램을 실행해야 하는 번거로움이 있다.

발현량 분석은 주로 상대적인 차이를 확인하는데 목적이 있다. 실험군과 대조군, 건강인과 환자, 약물에 대한 반응 차이 등 비교 대상 그룹의 유전자의 발현 차이를 비교하여 특징을 찾는다. 본 연구에서는 두 개의 대상을 비교하여, 그 발현량의 차이를 확인하는 것을 목표로 한다. NGS 장비에서 도출되는 원 데이터를 이용하여 두 그룹간의 발현 차이를 비교하는 결과를 도출하는 파이프라인을 shell script로 구현하였다.

Ⅱ. RNA-seq 분석 파이프라인

1. RNA-seq 분석 과정

NGS 분석 결과 대량의 read를 얻게 된다. raw read의 양 끝은 정확도가 떨어지므로, 퀄리티가 낮은 양 끝의 데이터를 제거하는 trimmimg의 과정을 거친다. timming 툴에는 Trimmomatic, sickle, cutadapt 등이 있다. trim된 read의 퀄리티를 확인하기 위하여 quality control(QC) 프로그램을 사용한다. 일반적으로 FastQC 프로그램을 많이 이용하며, 결과는 그래프과 phred(Q) score로 확인한다. 준비된 read를 reference genome에 붙이는 mapping 작업이 이어진다. mapping 툴로는 Bowtie, TopHat, BWA, HISAT2 등이 있다. human의 genome reference와 annotation data는 iGenome(https://support.illumina.com/ sequencing/sequencing_software/igenome.html)에서 얻을 수 있다. mapping 작업으로 얻은 sam 파일을 정렬하여, binary 형식의 bam 파일로 변환한다. bam 파일을 이용하여, exon, intron, CDS 등의 gene 구조를 확인하는 assembly 과정을 거친 후, read의 depth를 이용하여 정량(quantification)하는 과정을 진행한다. 정량하여 수치로 도출하는 방법에는 여러 가지가 있으나 대표적으로 FPKM(paired end)과 read count가 쓰인다[2].

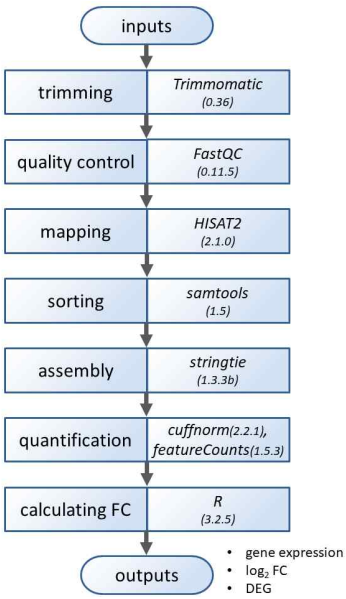
2. RNA-seq 분석 파이프라인 구성

RNA-seq 파이프라인은 CentOS 6.7의 리눅스 환경에서 구축하였다. (그림 1)은 분석순서와 분석 과정에서 사용된 프로그램 및 그 버전을 정리한 순서도이다.

trimming tool은 java 베이스의 Trimmomatic이 실행된다. read를 양 방향에서 읽은 paired end, 퀄리티 Q score 33으로 기본 세팅되어 있다. 퀄리티 확인은 마찬가지로 java 베이스의 FastQC가 실행된다. 정돈된 read를 reference genome에 mapping 하는 작업은 HISAT2를 이용하였다. samtools를 실행시켜 sam 파일을 bam 파일

로 변환한 후, stringtie의 assembly 과정을 거친다. 이어서 cuffnorm, featureCounts 두 프로그램으로 나뉘어 각각 FPKM과 read count로 발현량을 수치화한다[3].

다음으로 cuffnorm 실행 결과 발생한 FPKM 수치를 이용하여 두 샘플 발현량을 비교할 수 있는 fold change(FC) 값을 계산하며, 여러 사전 연구에서 의미있게 분류되는 $\log_2$(FC) 값 2 이상인 유전자를 따로 선별하여 결과를 제공한다.



▶▶ 그림 1. RNA-seq 분석 파이프라인 실행 순서 및 프로그램

3. RNA-seq 분석 파이프라인 실행

파이프라인을 실행하기 위해 세 개의 스크립트, RNA_2sets.sh, sample_sheet.sh, cuffnorm_FC.R가 필요하다. RNA_2sets.sh는 mapping부터 differential expression gene(DEG)를 얻는 전 과정을 실시하는 파이프라인 shell 스크립트이다. 파이프라인 내부에서 sample_sheet.sh 스크립트가 실행되는데, 반복실험에서 생기는 replication data를 그룹별로 묶기 위한 작업이다. 끝으로 cuffnorm_FC.R는 cuffnorm으로 얻은 FPKM 값을 이용하여 FC를 계산하고, DEG를 제공하는 R 스크립트로 파이프라인 내부에서 실행된다.

리눅스 환경에 (그림 1)의 프로그램(Ttrimmomatic, FastQC, HISAT2, samtools, stringtie, cuffnorm, featureCounts, R)을 설치한다. 입력 데이터는 reference genome(.fa), annotation(.gtf)과 분석하고자 하는 실험 파일(.fastq)이다. 실행파일 RNA_2sets.sh, sample_sheet.sh, cuffnorm_FC.R과 ./raw_data 폴더를 준비하고, 폴더 내에는 분석하고자 하는 fastq 파일을 넣는다.

(1)과 같은 커맨드 라인으로 RNA_2sets.sh script를 실행한다.

$bash RNA_2sets.sh (1)

실행 결과로는 6개의 하위 폴더, ./trimmed_data, ./fastqc, ./bam_sam_gtf, ./hisat2_summary, ./cuffnorm, ./foldchange가 생성된다. ./trimmed data는 trimmimg 결과 생성된 fastq 파일, ./fastqc에는 read의 퀄리티를 그래프와 수치로 편리하게 확인할 수 있는 html 파일, ./bam_sam_gtf 폴더에는 mapping이 되어있는 sam, bam, gtf 파일이 존재하며, ./hisat2_summary에서는 assembly 결과를 확인할 수 있다. ./cuffnorm에는 FPKM 값이 있고, featureCounts.txt 파일으로 read count 데이터를 얻을 수 있다. 마지막으로 ./foldchange 폴더에는 cuffnorm 결과 계산된 FPKM 값으로 구한 fold change 값이 있다. $\log_2$(FC) 값 2 이상인 유전자를 따로 선별하여 발현량과 함께 cuffnorm_log2FC_over2.txt 파일로 저장하였고, 이 때의 유전자를 cuffnorm_log2FC_over2_genelist.txt로 저장하여 연구자의 편의를 도모하였다.

Intel(R) Xeon(R) CPU E5-4607 v2 @ 2.60GHz, 297GB RAM 환경에서 human의 2개 paired end datasets(5.0 ~ 5.3 GB, 4개의 fastq 파일)를 이용하여 모의실험을 진행한 결과, 결과를 얻기까지 2시간 10여분 정도의 시간이 소요되었다.

Ⅲ. 결론

NGS 분석 결과 생성된 raw read를 이용하여, 유전자의 발현량을 계산하고, 그 차이를 비교 분석하기 위하여 여러 단계를 거친다. 그 과정에서 다양한 프로그램을 선택하고, 다루어야 한다.

본 연구에서는 하나의 커맨드 라인을 실행시켜 유전자의 발현량 및 FC 까지 한 번에 획득할 수 있는 파이프라인을 구축하였다. FC 값은 기존의 프로그램이 아닌, R script로 직접 계산하였고, FPKM과 FC 값으로 단순하게 나타내어 원하는 정보를 직관적으로 식별할 수 있다. 기존의 프로그램인 cufflinks protocol 결과와 비교하였을 때, cuffdiff의 실행 시간이 2시간 남짓인 것에 반해[3], cuffnorm_FC.R은 초 단위 시간으로 FC와 DEG를 얻을 수 있어 경제적이다.

파이프라인 실행 결과 얻은 발현량 자료는 FC 이외에 다양한 방법으로 차등 발현 분석이 가능하다. 발현량의 차이에 따른 표현형의 변화를 관찰하여 질병 연구와 극복에 도움이 되기를 기대한다. 또한 2개 이상 그룹의 발현량 비교 결과 도출, 그래프 등 시각적 정보 제공 등으로 확장하여 연구자가 더욱 편하게 정보를 얻을 수 있도록 할 것이다.

■ 참 고 문 헌 ■

[1] Martin J., Wang Z., "Next-generation transcriptome assembly", Nat Rev Genet. Vol. 12, pp. 671-682, 2011.

[2] Griffith M et al, "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud", PLoS Comput Biol. Vol. 11(8), 2015.

[3] Trapnell C. et al. "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", Nat. Protoc. Vol, 7, pp. 562-578, 2012.