# RNA-Seq 정렬 알고리즘의 동향

유승학*, 최민석**, 윤성로**
*서울대학교 반도체공동연구소
**서울대학교 전기정보공학부
e-mail : sryoon@snu.ac.kr

# Recent Trends in RNA-Seq Alignment Algorithms

Seunghak Yu*, Sungroh Yoon**
*Dept. of IT Convergence, Korea University
**Dept. of Electrical and Computer Engineering, Seoul National University

**요 약**

High Throughput Sequencing (HTS) 기술의 발달로 인해 시퀀싱 비용이 감소함에 따라 다양한 분야에서 이를 활용한 융합 연구가 활발하게 진행되고 있다. HTS 기술에서 가장 중요한 부분은 수백만 개의 short read 들을 표준유전체 (reference genome)에 정렬시키는 것인데 RNA 시퀀싱 (RNA-Seq)의 경우 RNA splicing 으로 인해 일반적인 aligner 로 처리가 불가능하다[1]. 복잡한 RNA-Seq 정렬 문제를 해결하기 위해 그동안 다양한 알고리즘들이 제안되어 왔다. 본 논문에서는 RNA-seq 정렬분야에서 잘 알려진 알고리즘들과 최신 알고리즘들을 살펴봄으로써 RNA-seq 정렬 알고리즘의 동향을 살펴보고자 한다.

## 1. 서론

Sanger 시퀀싱을 통해 인간 게놈 프로젝트 (Human Genome Project, HGP)가 완성된 이후 Sanger 시퀀싱을 대체할 High Throughput Sequencing (HTS)의 도입으로 인간 게놈 시퀀싱 비용이 $1,000 이하로 떨어질 것으로 예상되고 있다 (그림 1). 시퀀싱 비용의 감소로 인해 유전학적 시퀀싱 데이터에 대한 접근성이 높아져 의학, 생태학, 분자 생물학 등 다양한 분야에서 생물학적 어플리케이션 연구가 활발히 진행되고 있으며 이를 통해 개인의 유전특성에 맞는 맞춤의학 (personalized medicine) 시대가 열릴 것으로 기대된다.

HTS 는 DNA-seq, RNA-seq, ChIP-seq 등의 어플리케이션에 적용되고 있으나 본 논문에서는 RNA-seq 만을 고려한다. HTS 기술에서 가장 중요한 부분은 병렬적으로 생성된 수백만 개의 short read 들을 표준 유전체에 올바로 정렬시키는 과정이라 할 수 있는데 RNA-seq 의 경우 RNA splicing 으로 인해 DNA 등과 다르게 정렬과정이 더욱 복잡한 것으로 알려져 있다[1]. RNA splicing 이란 전사(transcript) 과정에서 인트론(intron)이 제거되고 엑손(exon)간의 연결이 이루어지는 현상으로 이 인트론 부분이 정렬 과정에서 large alignment gap 을 생성하는 원인이 된다.

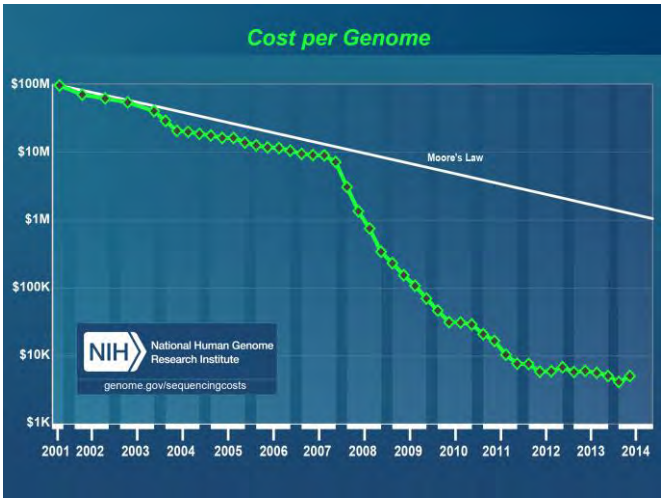
따라서 RNA-Seq 데이터는 표준유전체에 정렬될 때 RNA splicing 으로 인한 large alignment gap 의 유무에 따라 다음과 같은 두 가지 타입의 read 로 나눌 수 있다[3]. 하나는 large gap 없이 정렬되는 일반 read 이며 다른 하나는 large alignment gap 을 생성하며 정렬되는 splice junction read 이다. 일반적인 read 는 기존의 short sequence alignment 알고리즘을 통해 정렬할 수 있지만 splice junction read 는 새로운 처리가 필요하다. 이러한 splice junction 을 발견하여 처리할 수 있는 연구들이 진행되어 왔으며 본 논문에서는 이 분야의 잘 알려진 알고리즘들과 최신 알고리즘들을 살펴봄으로써 RNA-seq 정렬 알고리즘의 동향을 분석한다.

## 2. RNA-seq 정렬 알고리즘

수많은 RNA-seq 알고리즘들이 splice junction 을 효과적으로 정렬할 수 있다고 주장하며 제안되어 왔지만 실용적인 알고리즘은 다음의 세 가지 기준을 만족해야 한다[8]: (1) splice junction 포함 read 정렬 (2) paired-end read 처리 (3) 합리적인 수행시간. 이러한 기준을 토대로 본 논문에서는 다음과 같은 7 가지의 알고리즘을 선별하였다 <표 1>: TopHat[4], TopHat2[5], GSNAP[6], MapSplice[7], SpliceMap[8], RUM[9], STAR[10].

TopHat, TopHat2 는 첫 번째로 Bowtie[11] 등을 사용하여 표준유전체에 read 들을 정렬시켜 splice junction 이 포함되지 않은 잠재적 엑손들을 찾아내고 이때 정렬되지 않는 100bp 이하의 짧은 read 들을 정렬하기 위해 모든 read 를 작은 조각으로 나눈 뒤 독립적으로 정렬한다. 두 번째로 이러한 매핑 정보를 이용해 잠재적 splice junction 들의 데이터베이스를 구축하여 splice junction 을 찾아낸다.

GSNAP 은 해쉬 테이블 기반의 알고리즘으로 donor 와 acceptor splice site 들의 확률적 모델을 사용하여




(그림 1) 인간 게놈 시퀀싱 비용의 감소 추이[2]

splicing 을 계산하고 알려진 splice site 데이터베이스를 이용하여 계산된 splicing 에서 false positive 와 false negative 를 걸러낸다. 또한 splicing 이외에도 multiple mismatch 와 long indel 혹은 이들의 조합을 찾아낼 수 있지만 read 당 한 개의 splice 와 indel 만을 알아낼 수 있다는 단점이 있다.

MapSplice 는 TopHat 과 유사한 두 단계를 통해 정렬을 한다. 먼저 read 들을 쪼개어 Bowtie 를 이용해 표준유전체에 정렬하여 엑손들을 찾아내고 정렬되지 않은 조각들을 이용해 splice junction 을 찾아낸다. 이 두 정보를 결합한 후보군들 중에서 Alignment quality, Anchor significance, Entropy 를 계산하여 최선의 정렬을 최종적으로 찾아낸다.

SpliceMap 은 splice 를 갖는 read 는 반드시 read 길이의 절반 이상이 표준유전체와 정렬돼야 한다는 사실에 착안한 알고리즘이다. 먼저 read 를 절반으로 쪼갠 뒤 Eland, SeqMap[12] 등을 사용하여 표준 유전체에 정렬한다. 정렬된 read 나머지 절반만이 splice junction 을 찾는데 사용되어 효율적이다.

RUM 은 Bowtie, BLAT[13]의 결과와 유전체 정렬, 전사체 정렬에서 오는 정보를 모두 결합한 알고리즘이다. 첫 번째로 read 들을 Bowtie 를 사용해 표준유전체와 표준전사체에 정렬하고 Bowtie 가 정렬하지 못한 read 들은 BLAT 을 적용한다. 이 둘의 결과를 최종적으로 결합되어 splice junction 을 알아내는데 사용한다.

STAR 는 uncompressed suffix array 를 사용하여 표준유전체에 read 를 정렬한다. 하나의 read 가 정렬할 수 있는 최대 길이를 순차적으로 탐색하면서 정렬 되지 않는 구간을 표시한다. 앞선 단계의 결과들을 모아 scoring 을 하고 이를 통해 최종 정렬 결과를 얻는다. 이를 통해 splice junction 은 물론 multiple mismatch 와 indel 을 찾아낼 수 있으며 수행시간이 짧다는 장점이 있다.

## 3. 결론

RNA-seq 정렬 알고리즘은 RNA splicing 으로 인해 large alignment gap 이 생기는 splicing junction read 를 효과적으로 정렬하는 것이 중요하다. 본 논문에서는 이러한 splicing junction 를 합리적인 수행시간 안에 처리할 수 있는 실용적인 RNA-seq 알고리즘 들을 살펴봄으로써 현재의 연구수준 및 동향을 알아보고자 하였다.

하지만 기존의 제안된 알고리즘 모두 sensitivity, precision 등의 정확성 측면 또는 수행시간, 디스크 용량 등 컴퓨팅 자원 측면에서 개선될 필요가 있다[9].

따라서 향후 이들의 장단점을 면밀히 분석한 후 정확성 측면과 컴퓨팅 자원 측면을 모두 만족시킬 수 있는 RNA-seq 정렬 알고리즘 개발을 목표로 하고 있다.

<표 1> RNA-seq 알고리즘의 인용횟수 및 위치

| Algorithm | Citation | Site |
|---|---|---|
| TopHat | 2,303 | http://ccb.jhu.edu/software/tophat/index.shtml |
| TopHat2 | 233 | |
| GSNAP | 377 | http://research-pub.gene.com/gmap |
| MapSplice | 215 | http://www.netlab.uky.edu/p/bioinfo/MapSplice2 |
| SpliceMap | 191 | http://web.stanford.edu/group/wonglab/SpliceMap |
| RUM | 101 | https://github.com/itmat/rum/wiki |
| STAR | 130 | https://code.google.com/p/rna-star |

## 사사

이 논문은 2012 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 [No. 2011-0009963, No. 2012M3A9D1054622].

## 참고문헌

[1] Li, Heng, and Nils Homer. "A survey of sequence alignment algorithms for next-generation sequencing." Briefings in bioinformatics 11.5 (2010): 473-483.
[2] http://www.genome.gov/sequencingcosts/
[3] Chen, Liang. "Statistical and computational methods for high-throughput sequencing data analysis of alternative splicing." Statistics in biosciences 5.1 (2013): 138-155.
[4] Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics 25.9 (2009): 1105-1111.
[5] Kim, Daehwan, et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol 14.4 (2013): R36.
[6] Wu, Thomas D., and Serban Nacu. "Fast and SNP-tolerant detection of complex variants and splicing in short reads." Bioinformatics 26.7 (2010): 873-881.
[7] Wang, Kai, et al. "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery." Nucleic acids research (2010): gkq622.
[8] Au, Kin Fai, et al. "Detection of splice junctions from paired-end RNA-seq data by SpliceMap." Nucleic acids research 38.14 (2010): 4570-4578.
[9] Grant, Gregory R., et al. "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)." Bioinformatics 27.18 (2011): 2518-2528.

[10] Dobin, Alexander, et al. "STAR: ultrafast universal RNA-seq aligner." Bioinformatics 29.1 (2013): 15-21.
[11] Langmead, Ben, et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol 10.3 (2009): R25.
[12] Jiang, Hui, and Wing Hung Wong. "SeqMap: mapping massive amount of oligonucleotides to the genome." Bioinformatics 24.20 (2008): 2395-2396.
[13] Kent, W. James. "BLAT—the BLAST-like alignment tool." Genome research 12.4 (2002): 656-664.