RNA 시퀀싱 데이터를 이용한 병렬 SNP 추출 알고리즘

김덕근*, 이덕해*, 공진화*, 이은주**, 윤지희*
*한림대학교 컴퓨터공학과
**한림대학교 전자공학과
e-mail:dkkim@hallym.ac.kr

A parallel SNP detection algorithm for RNA-Seq data

Deok-keun Kim*, Deok-hae Lee*, Jin-hwa Kong*,
Un-joo Lee**, Jee-hee Yoon*
*Dept of Computer Engineering, Hallym University
**Dept of Electronic Engineering, Hallym University

요 약

최근 차세대 시퀀싱 (Next Generation Sequencing, NGS) 기술이 발전하면서 DNA, RNA 등의 시퀀싱 데이터를 이용한 유전체 분석 방식에 관한 연구가 활발히 이루어지고 있다. 차세대 시퀀싱 데이터를 이용한 유전체 분석 방식은 마이크로어레이 혹은 EST/cDNA 데이터를 이용한 기존의 분석 방식에 비하여 비용이 적게 들고 정확한 결과를 얻을 수 있다는 장점이 있다. 그러나 이 들 DNA, RNA 시퀀싱 데이터는 각 시퀀스의 길이가 짧고 전체 용량은 매우 커서 이 들 데이터로부터 정확한 분석 결과를 추출하는 데에 많은 어려움이 있다. 본 연구에서는 클라우드 컴퓨팅 기술을 기반으로 하여 대용량의 RNA 시퀀싱 데이터를 고속으로 처리하는 병렬 SNP 추출 알고리즘을 제안한다. 전체 게놈 데이터 중 유전자 영역만을 high coverage로 시퀀싱하여 얻어지는 RNA 시퀀싱 데이터는 유전자 변이 추출을 목적으로 분석되며, SNP(Single Nucleotide Polymorphism)와 같은 유전자 변이는 질병의 원인 규명 및 치료법 개발에 직접 이용된다. 제안된 알고리즘은 동시에 실행되는 다수의 Map/Reduce 함수에 의해서 대규모 RNA 시퀀스를 병렬로 처리하며, 레퍼런스 시퀀스에 매핑된 각 염기의 출현 빈도와 품질 점수를 이용하여 SNP를 추출한다. 또한 이 들 SNP 추출 결과에 대한 시각적 분석 도구를 제공하여 SNP 추출 과정 및 근거를 시각적으로 확인/검증할 수 있도록 지원한다.

1. 서론

NGS 기술에 의하여 생성되는 대용량의 시퀀스 정보는 DNA 혹은 RNA 리드에 해당 한다 [1]. 이러한 시퀀스 데이터는 다양한 생물의 유전체에 존재하는 유전적 구조 변이를 추출/연구하는 데에 사용될 수 있다. 인간의 각 개인 유전체의 특이성을 발견하려는 시도의 하나로서 DNA 시퀀스 분석에 관한 연구는 비교적 초기부터 시작되었으며, DNA 시퀀스 데이터를 이용한 구조적 변이 (SNP, Indel, CNV 등) 추출 방법 및 질병 관련성을 규명하기 위한 연구가 매우 활발히 진행되고 있다 [2]. 그러나 RNA 시퀀스 분석에 관한 연구는 국내외적으로 아직 초기 단계이다. RNA는 세포의 유전 정보 발현에 관여하는 물질로서 DNA의 유전 정보를 세포질까지 전달하는 역할을 수행 한다. RNA 시퀀스는 DNA 시퀀스의 경우와 달리 유전자 영역만을 high coverage로 시퀀싱하여 얻어지며, 정확한 유전 변이 검출을 통해 질병의 원인 분석 연구에 직접 사용된다.

대표적인 NGS 기술 보유 회사로는 Illumina [3], 454 Life Science [4], Applied Biosystems [5] 등을 들 수 있다. 이 들의 차세대 시퀀싱 머신은 짧은 시간 내에 대용량의 리드 시퀀스를 생성하며, 생성된 리드는 30-400 bp (basepair)의 짧은 길이를 가지며, 실험 기술에 따라 single-end 혹은 paired-end 타입으로 구별된다. 예를 들어 Illumina의 Genome Analyzer는 열흘 동안 약 33Gb (300 million 2*100 bp reads에 해당함)의 대용량의 시퀀스 데이터를 생성해낸다. 이와 같이 차세대 시퀀싱 데이터는 시퀀스의 길이가 짧고 전체 용량은 매우 커서 이 들 데이터로부터 정확한 분석 결과를 추출하는 데에 많은 어려움이 있다. 차세대 시퀀싱 업체들이 데이터 분석을 위한 소프트웨어를 일부 제공하고 있으나 (Illumina의 CASAVA 1.6 [3], AB의 BioScope 3.0 [5] 등), 데이터 용량, 처리 방식이 아직은 매우 제한적이라고 할 수 있다.

본 연구에서는 클라우드 컴퓨팅 기술을 기반으로 RNA 시퀀싱 결과로 산출되는 수많은 짧은 RNA 시퀀스 조각인 리드 데이터로부터 SNP를 추출하는 병렬 SNP 추출 알고리즘을 제안한다. RNA 리드 매핑을 위하여 GSNAP [7] 정렬 툴을 사용하였으며, GSNAP 결과 파일과 레퍼런스

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2010-0007655)

시퀀스는 사전에 하둡 (Hadoop) [6]의 파일 시스템인 HDFS에 저장된다. 제안된 알고리즘은 동시에 실행되는 다수의 Map/Reduce [6]함수에 의해서 대규모 RNA 시퀀스(GSNAP 결과)를 병렬로 처리하며, 레퍼런스 시퀀스에 매핑된 각 염기의 출현 빈도와 품질 점수를 이용하여 SNP를 추출한다. 이와 같이 처리된 클라우드 스케일의 대규모 RNA 데이터와 SNP 추출 결과는 로컬 데이터베이스에 저장되며, 시각적 분석 툴과 연계되어 그 결과를 추출, 검증할 수 있다. 시각적 분석 툴은 SNP 추출 결과에 대한 레퍼런스 정보, 매핑된 리드 데이터의 세부 정보 등을 함께 제시하여 SNP 추출 과정 및 근거를 시각적으로 확인/검증할 수 있도록 지원한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로서 RNA 데이터 분석 도구와 클라우드 컴퓨팅 기술을 이용한 바이오 정보 처리 현황에 대해서 설명한다. 3장에서는 SNP 추출 방법과 병렬 SNP 추출 알고리즘을 보이고, 4장에서는 SNP 추출 결과에 대한 시각적 분석 방식을 설명한다. 마지막으로 5장에서는 본 논문을 요약하고, 결론을 내린다.

2. 관련연구

2.1 클라우드 컴퓨팅 기술을 이용한 바이오 정보 처리

최근 생물정보학 분야에서는 대용량의 데이터를 분석하기 위해 클라우드 컴퓨팅을 기반으로 하는 분석 기법 및 소프트웨어 개발에 관한 연구를 진행하고 있다.

클라우드 컴퓨팅은 네트워크 기술을 활용하여 서로 다른 물리적인 위치에 존재하는 컴퓨터들의 리소스들을 가상화 기술로 통합하여 제공하는 기술로 정의된다. 클라우드 컴퓨팅분야에서 가장 대표적인 것은 대용량 데이터 처리 분석 오픈소스 프로젝트인 하둡이다. 하둡은 대량의 자료를 처리할 수 있는 분산 응용 프로그램을 지원하는 오픈 소스 프레임워크이며, 분산/병렬 시스템의 리소스들을 효율적으로 지원하기 위한 목적으로 Map/Reduce 프로그래밍 모델을 제공한다.

클라우드 컴퓨팅을 이용한 유전 변이 검출 및 분석 관련 소프트웨어 중 Crossbow [8]는 기존의 정렬 도구인 Bowtie와 SNP 추출 도구인 SOAPsnp를 결합하여 클라우드 환경에서 동작하도록 제작한 프로그램이다. Crossbow는 하둡 환경에서 병렬로 수행되며, Map 함수에서 Bowtie를 통해 리드들을 매핑하고, Reduce 함수에서 SOAPsnp를 통해 SNP를 추출한다. 또한 CloudBurst [9]는 최적화된 병렬 리드 매핑 알고리즘으로서 프로세서 수가 증가함에 따라 프로그램 실행 시간이 반비례하여 줄어드는 효과를 나타내는 것으로 알려져 있다. 이외에도 CloudBLAST [10], Galaxy [11]등의 클라우드 컴퓨팅을 이용한 바이오정보처리 검출 및 분석도구에 대한 연구가 활발히 진행되고 있다.

2.2 RNA 데이터 분석 도구

최근 RNA 시퀀스 데이터를 분석하기 위한 다양한 목적의 소프트웨어 도구들이 개발되었다. Johns Hopkins의 연구그룹에서 개발한 Myrna [12]는 RNA 시퀀스를 이용하여 DE (Differential Expression) 유전자를 검출하는 프로그램이다. 클라우드 스케일의 대용량 데이터 처리를 위하여 하둡 환경의 Map/Reduce 프로그래밍 방식을 사용하였으며, 리드 정렬 프로그램인 Bowtie와 통계처리 프로그램인 R/Bioconductor를 이용하였다. UC Berkely의 연구 그룹에서 개발한 Cufflinks [13]는 RNA 시퀀스를 어셈블리하여 유전자별 전사량을 추출하고, 이들 값을 기반으로 하여 주어진 샘플 사이의 DE, regulation을 테스트하는 프로그램이다 . 또한 SpliceMap [14]은 RNA 시퀀스 데이터를 분석하여 splice junction을 추출하는 정렬 도구이다 .

3. 병렬 SNP 추출 알고리즘

SNP는 세포핵 속의 염색체가 갖고 있는 30 억개의 염기서열 중 단일 염기의 차이가 발생하는 것을 말한다. 본 장에서는 RNA 시퀀스를 이용한 SNP 추출 알고리즘을 설명하고, Hadoop 플랫폼의 Map/Reduce함수를 이용한 병렬 SNP 추출 알고리즘을 제안한다.

3.1 SNP 추출 알고리즘

SNP 추출 알고리즘을 단계적으로 설명하면 다음과 같다.

(Step 1) RNA 리드 데이터를 휴먼 게놈 레퍼런스 서열에 서열 정렬 도구 GSNAP을 이용하여 정렬한다. SNP 추출을 위한 정렬된 리드의 전제 조건으로서 unique 형태로 정렬되고, 미스매치(mismatch)를 2개까지 허용하며, paired-end 리드의 경우 concordant로 정렬된 리드만을 사용한다. 이후 이들 조건을 만족하는 정렬된 리드를 effective 리드라고 부른다.

(Step 2) GSNAP 결과 중 effective 리드들만을 이용하여 각 레퍼런스 서열에 매핑된 위치별로 염기 정보(염기의 종류별 개수, 각 염기의 품질)를 수집한다. 이때 염기의 종류는 A, C, G, T의 4가지이며 염기 품질은 프레드 스코어(phred score) [15]로 주어진다고 가정하며, 프레드 스코어의 산출 방식에 근거하여 품질이 20 미만인 염기는 제외시킨다.

(Step 3) Step 2의 각 레퍼런스 위치에 대하여 수집된 염기 정보를 분석하여 레퍼런스 염기와 다른 종류의 단일 염기가 4번 이상 매핑된 경우 해당 염기를 SNP 염기로 마킹한다. 해당 레퍼런스 위치에 대하여 2개 이상의 SNP 염기가 마킹될 수 있다. 단, 해당 위치에 매핑된 염기의 총 수가 평균 커버리지의 3배 이상인 경우, 해당 위치는 SNP 추출 후보에서 제외시킨다.

(Step 4) 마킹된 SNP 염기가 해당 위치에 n번 매핑되어 각각 Q_1, Q_2, ..., Q_n의 품질을 갖는 경우, 매핑된 염기의 스코어 S를 다음 (식 1)에 의하여 계산한다.

$$S = S1 + S2 + ... + Sn,\ Si = 1 - 10^{-Qi/10}\ (i = 1...n) \quad ...(식1)$$

(Step 5) 해당 레퍼런스 위치에 대하여 2개 이상의 SNP 염기가 마킹되어 제일 높은 스코어와 두 번째 스코어의 비가 0.9 이하의 경우에는 “hetero SNP”로 분류되고, 그 이

외의 경우에는 모두 “homo SNP”로 분류된다.

(Step 6) SNP 염기가 추출된 레퍼런스 위치, 마킹된 SNP 염기, 염기 스코어, 접합자 구조(Zygosity) 분류 등의 정보를 출력한다.

다음 <표 1>에 RNA 시퀀스 분석에 의한 SNP 추출 결과의 예를 보인다. 여기에서 5 번째 행에 보이는 1번 염색체의 142125342번 포지션에 위치하는 SNP 추출 결과는 다음과 같이 해석된다. 레퍼런스 염기가 T 이고 해당 위치에 9개의 유효한 염기가 매핑 되었으며, 그 중 5개가 G 이고 나머지 4개가 C 이다. 따라서 G 와 C 가 SNP로 마킹되고, 스코어는 각각 4.99와 3.98이므로 “Hetero” 타입으로 분류된다.

<표 1> RNA 시퀀스 분석에 의한 SNP 추출 결과의 예

Chr#	Pos#	Part#	Ref	SNP	Type	Total	Used	A	C	G	T	Score
chr1	142120748	14212	C	G	Homo	7	7	0	2	5	0	4.98
chr1	142121927	14212	T	C	Homo	5	5	0	4	0	1	3.99
chr1	142123251	14212	A	C	Homo	6	5	0	5	0	0	4.98
chr1	142125342	14212	T	G/C	Hetero	10	9	0	4	5	0	4.99/3.98
chr22	20008850	2000	C	T	Homo	4	4	0	0	0	4	4
chr22	20008973	2000	A	G	Homo	5	5	1	0	4	0	3.99

3.2 병렬 SNP 추출알고리즘

3.2.1 Map/Reduce 함수의 단계별 역할

[알고리즘 1] Parallel_SNP_Detection

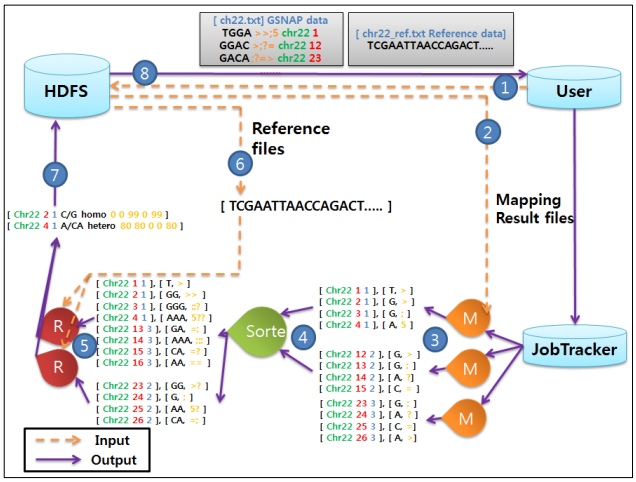
Input : Pre-processed files of alignment_result GR, reference sequence RS, Phred Quality Score thresold BST, Position read Mapping Count thresold PMCT
Output : set of SNPs SNP

```
MAP(K1, V1, K2, V2)
1.  for each V1[Chromosome chr, Position pos] in
                alignment_result do
2.  for each inc from <0> to <read_len-1> do
3.      K2.set(chr_ID, position_ID, partition_ID)
4.      V2.set(read_base, base_quality)
5.  return (K2, V2)
REDUCE(K2, V2, K3, V3)
1.  Reference sequence of Mapping position RefMS;
2.  Position_read_Mapping_Count PMC; Base_Count BC;
3.  Position_base_Quality PQ; Mapping_base MB;
4.  Base_Phred_Quality_Score BS;
5.  for each V2[read_base, base_quality] in MAP result do
6.  if PQ < BST then
7.          return;
8.      else
9.          BS.set(PQ);
10.     end
11. if PMC <= PMCT then
12. return;
13. RefMS.set(MB, BS);
14. SNPcall(RefMS);
15. ZygositySetAndRankSet(SNP, BC);
16. V3.set(GR, SNP, Zygosity, ToTal BC , Used BC,
        each base_BC, SNP_score)
17.  return (K3, V3)
```

병렬 SNP 추출을 위한 Parallel SNP Detection 알고리즘을 [알고리즘 1]에 보인다. 여기에서 입력 데이터 파일은 GSNAP에 의한 RNA 리드 매핑 파일로부터 “effective” 리드 결과만을 추출한 데이터 파일로 가정한다. Map 함수에서는 입력 파일의 리드 정보를 읽어들여 레퍼런스 서열의 각 위치에 매핑된 염기의 정보를 수집하기 위한 전처리 과정을 수행한다. 즉, Map 함수는 레퍼런스의 i번째 위치에 매핑된 길이 L의 리드에 대하여 염기 정보와 품질 정보를 읽어들여, [i, i+1, … , i+L-1]의 위치에 각각 매핑되는 염기와 품질 정보를 분리해 내는 작업을 수행한다. 예를 들어 1번 위치에 매핑된 리드의 길이가 4이고 염기정보와 품질 정보가 각각 [TGGA], [>>;5] 인 경우, 1번 위치에 ‘T’와 ‘>’의 염기 정보를, 2번 위치에 ‘G’와 ‘>’의 염기 정보를, 3번 위치에 ‘G’와 ‘;’의 염기 정보를, 4번 위치에 ‘A’와 ‘5’의 염기 정보를 각각 분리하여 낸 후 Map 함수의 결과 값으로 출력한다.

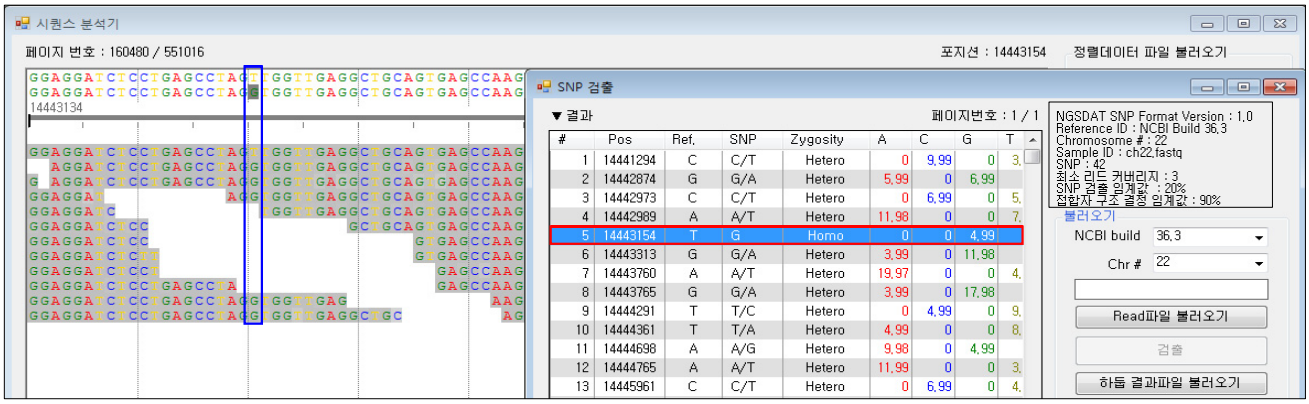
Map 함수는 입력 파일의 각 리드에 대하여 [염색체_ID, 매핑된 포지션, 리드 시퀀스, 리드 시퀀스의 염기 품질]을 추출하여, [염색체_ID, 매핑된 포지션, Partition 번호]를 이용하여 출력 키 값을 만든다. 이 때 할당되는 Partition 번호는 차후 Reduce 함수에 의하여 처리되는 데이터 양을 균등하게 배분하기 위하여 설정되는 값이다. 다음 Reduce 단계에서는 SNP 추출을 위한 함수가 수행된다. Reduce 함수는 키 값에 의하여 정렬된 중간 결과를 입력 받아 레퍼런스 서열의 각 위치에 매핑된 염기 집합을 구성한다. 레퍼런스 서열의 각 위치에 매핑된 리드의 염기 집합에서 염기의 종류, 품질 정보를 이용하여 SNP 검출을 수행한다. 그 결과 Reduce 함수는 [염색체_ID, 포지션, Partition번호, 레퍼런스 염기, SNP 염기, 접합자구조, 전체 염기 수, 사용된 염기 수, 염기별 출현 횟수, SNP 염기 스코어] 값으로 이루어진 최종 결과를 산출하게 된다.

3.2.2 데이터처리과정



(그림 1) 데이터 처리과정

다음 (그림 1)은 제안된 알고리즘에 의한 데이터 처리 과정을 단계별로 나타낸다. ① 사용자는 RNA 리드 매핑 파일과 레퍼런스 시퀀스를 HDFS에 복사한다. ② 입력된 파일은 HDFS의 블록 크기와 파일의 크기에 따라서 파일이 자



(그림 2) SNP 분석을 위한 시각화 분석 도구

동으로 분할되어지고, 분할된 파일의 개수에 따라 Map 함수의 수가 결정되어 수행된다. ③ Map 함수가 수행되고, 키 값에 Partition 번호가 삽입되어 Reduce 함수에서 처리되는 데이터가 일정하게 분산 처리되도록 한다. ④ Map 함수에 의하여 산출된 결과 값은 같은 키 값을 가지는 값들끼리 모이게 되고, 모인 값들은 각각 Shuffle/Sort 과정을 거치게 된다. ⑤ Sort된 값들은 각각 Reduce 함수에 넘겨져 수행된다. Reduce 함수는 사용자가 정의한 개수만큼 생성되며 그 개수에 따라 일정한 Partition 번호순으로 데이터를 처리하게 된다. ⑥ Reduce 함수는 취합된 리드의 염기 정보와 레퍼런스 시퀀스 정보를 이용하여 SNP 추출 연산을 수행한다. ⑦ Reduce 함수를 통해 최종적으로 산출된 값들은 다시 HDFS에 저장된다. ⑧ 사용자들은 HDFS에 저장된 결과 값을 다시 로컬 시스템으로 복사하여 사용하게 된다.

4. 시각적 분석 툴을 이용한 SNP결과 검증

본 시스템에서는 SNP 결과 검증을 지원하는 시퀀스 분석기를 제공한다. (그림 2)에 시각적 분석 툴의 사용 예를 보인다. 시퀀스 분석기는 레퍼런스 서열에 매핑된 모든 리드들의 시퀀스 정보를 구체적으로 확인/분석하는데 사용되고 SNP 검출 영역에 대한 리드 매핑 정보를 구체적으로 확인하는데 사용될 수 있다. (그림 2)에서 빨간색 영역으로 표시한 SNP 추출결과는 시퀀스 분석기에서 파란색 영역으로 표시된 부분에 연동하여 표시 가능하며, 이 들 결과를 이용하여 SNP 추출 과정 및 근거를 시각적으로 확인/검증할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 대규모 RNA 데이터를 처리하기 위한 클라우드 컴퓨팅기술 기반의 병렬 SNP 추출 방법을 제안하였다. 제안하는 병렬 SNP 추출방법은 Map/Reduce함수를 이용하여 RNA 시퀀싱 데이터를 레퍼런스 시퀀스에 매핑하고 매핑된 리드의 염기 분포, 리드 품질 등의 정보를 활용하여 SNP 영역을 추출한다. 향후 SNP 분석 기능에 대한 기존의 방법과의 비교 실험을 수행하고 시각화 분석 도구를 통한 SNP 검증 기능을 강화할 예정이다. 또한 상용 클라우드 컴퓨팅 서비스를 이용한 대규모 병렬 컴퓨팅환경에서의 실험을 통해 병렬 알고리즘을 이용한 성능 개선 효과에 대한 분석 연구를 수행할 예정이다.

참고문헌

[1] M L. Metzker, "Sequencing technologies – the next generation," Nature Reviews Genetics, Vol. 11, pp. 31-46, 2010.

[2] R. Redon et al., "Global variation in copy number in the human genome," Nature, Vol. 444, No. 7118, pp. 444-454, 2006.

[3] http://www.illumina.com

[4] http://www.454.com

[5] http://www.appliedbiosystems.com

[6] T. White, "Hadoop : The Definitive Guide," O'REILLY, 2009.

[7] T. D. Wu, S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," Bioinformatics, Vol. 26, No. 7, pp. 873-881, 2010.

[8] Langmead et al., "Searching for SNPs with cloud computing," Genome Biology, Vol. 10, No. 11, 2009.

[9] M. C. Schatz, "CloudBurst: Highly Sensitive Read Mapping with MapReduce," Bioinformatics, Vol. 25, No. 11, pp. 1363-1369, 2009.

[10] A. Matsunaga, M. Tsugawa, J. Fortes, "CloudBLAST: Combining MapReduce and virtualization on distributed resources for bioinformatics applications," 4th IEEE International Conference on e-Science, pp. 222-229, 2008.

[11] B. Giardine et al., "Galaxy: a platform for interactive large-scale genome analysis," Genome research, Vol. 15, No. 10, pp. 1451-1455, 2005.

[12] B. Langmead, K. Hansen, J. Leek, "Cloud-scale RNA-sequencing differential expression analysis with Myrna," Genome Biology, Vol. 11, No. 8, 2010.

[13] Trapnell et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," Nature Biotechnology, Vol. 28, No. 5, pp. 511-515, 2010.

[14] K. F. Au et al., "Detection of splice junctions from paired-end RNA-seq data by SpliceMap," Nucleic Acids Research, Vol. 38, No. 14, pp. 4570-4578, 2010.

[15] H. Li et al, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," Genome Research, Vol. 18, No. 11, pp. 1851-1858, 2008.