차세대 서열 기술을 이용한 RNA-Seq 분석 시스템의 개발

이 종 근[†○], 홍 동 완[§], 윤 지 희[†], 이 은 주[‡]
한림대학교 컴퓨터공학과[*], 전자공학과[‡], 서울대학교 의과대학[§]
{jeikei[†○], jhyoon[†], ejlee[‡]}@hallym.ac.kr, dongwan@snu.ac.kr[§]

Development of RNA-Seq analysis system using next generation sequencing technology

Jongkeun Lee[†○], Dongwan Hong[§], Jeehee Yoon[†], Unjoo Lee[‡]
Department of Computer Engineering[†], Department of Electric Engineering[‡], Hallym University,
Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine[§]

1. 서 론

차세대 시퀀싱 (Next-generation sequencing, NGS) 기술의 등장은 인간 유전체 분석을 정밀한 수준까지 가능하게 하였으며, 이는 세계적 연구 그룹들이 P4 의학 실현, 즉 개인별 맞춤의학 (Personalized medicine), 예측의학 (Predictive medicine), 예방의학 (Preventive medicine), 참여의학 (Participatory medicine)을 목표로 추진하고 하는 최신 연구의 기반 기술로 자리 잡고 있다.

NGS를 이용한 인간 유전체 분석 방법은 분석 범위와 분석 방법에 따라 전장 시퀀싱 (whole genome sequencing), 목적 시퀀싱 (targeted sequencing), ChIP-seq, Methylation 등으로 구분된다. 전장 시퀀싱의 경우 각 개인의 유전체 서열에서 발생하는 SNP (Single Nucleotide Polymorphism), Indel, CNV (Copy Number Variation) 등과 같은 유전체 변이를 조사하여 각 개인에서 공통적으로 나타날 수 있는 변이 (common variants)와 희귀하게 나타나는 변이 (rare variants)를 분석하여 형질 (phenotype)과 질환 (disease)과의 연관성을 해명하는 것을 목적으로 사용된다. 목적 시퀀싱의 경우, 유전자 (gene)로 표현되는 특정 지역에서 변이 (mutation)가 발생할 경우 질환의 원인이 되는데, 특정 환자 군의 유전자를 high coverage로 시퀀싱하여 질병 원인을 규명하고 치료 방법을 발견하는 것을 목적으로 한다.

NGS 기술로 생성되는 데이터는 그 규모와 형식에서 기존의 데이터와 매우 다르며, 1명의 개인 유전체 서열 정보가 100 Gbp (Giga base pair)에 달하는 대용량 데이터에 해당한다. 또한 개인 유전체 서열 생산의 비용이 US$ 1,000 수준으로 떨어질 것으로 예상되어 개인 서열 데이터의 양이 기하 급수적으로 증가될 것이다. 하지만 현재까지 전장 유전체 시퀀싱은 고가의 분석 실험으로 분류되며, 전장 시퀀싱에 비하여 비용이 적게 드는 RNA 시퀀싱[1], Exon 시퀀싱[2] 실험에 관한 관심이 매우 높아지고 있다. RNA 시퀀싱 (transcriptome sequencing)은 인간의 유전자 부분만을 시퀀싱하는 기술을 말하며, Exon 시퀀싱은 인간 유전자 중 1-2%를 차지하는 단백질 coding 염기 서열인 전체 Exon (Exome)을 타켓으로 염기서열을 분석하는 기술을 말한다. 최근 이들 RNA, Exon 데이터의 양이 급속히 증가되고 있다. 하지만, 이들 데이터를 효율적으로 분석, 관리할 수 있는 알고리즘 개발에 관한 연구가 거의 이루어지지 못하고 있으며, 이들 대규모 데이터 처리를 위한 통합 분석 시스템의 개발이 시급한 실정이다.

본 논문에서는 그림 1과 같이 목적 시퀀싱 데이터를 2단계로 처리, 분석하는 RNA-Seq 분석 시스템을 제안하고 개발 방법론에 대하여 논한다. 본 논문에서 개발한 RNA-Seq 분석 도구는 크게 2개의 모듈로 구성된다. 첫 번째 모듈은 NGS에서 생성되는 FASTQ 데이터를 유전자 위치에 서열 정렬하여 coverage 데이터, SNP, Indel의 유전체 변이 정보를 추출, 저장하는 기능을 담당한다. 두 번째 모듈은 첫 번째 모듈에 의하여 추출된 분석 데이터를 유전자 또는 Exon 별로 비교하여 분석 결과를 사용자 (biologist or medical researcher)에게 제공하는 기능을 담당한다.

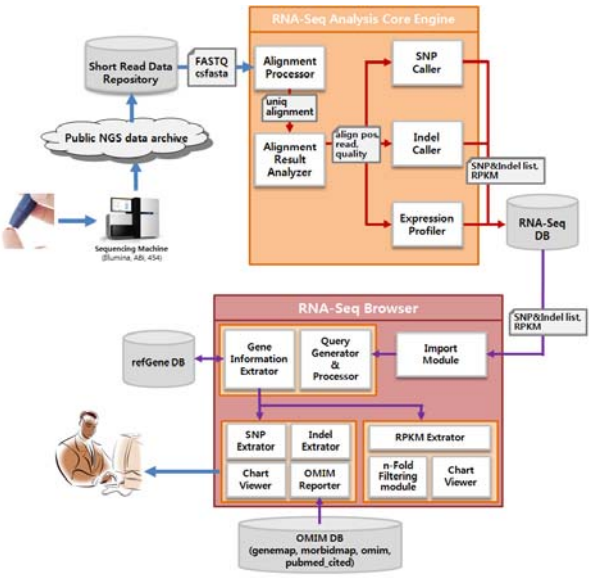

그림 1. 시스템 구조도

2. 본 론

본 장에서는 RNA-Seq 데이터의 분석 방법에 대하여 기술한다. RNA-Seq 데이터 pipeline은 ① 서열정렬, ② 서열 정렬 정보 추출, ③ Expression profiler, ④ 단일 염기 변이와 Indel 추출 기능으로 구성된다. RNA 시퀀싱을 통해 산출된 short read 데이터를 휴먼 지놈 레퍼런스 (Build 36.3)에 서열 정렬 도구를 이용하여 서열 정렬을 한다. 본 논문에서는 GSNAP 서열 정렬 도구 [3]를 사용하였다. 정렬 결과에는 레퍼런스 서열 상에 매핑된 chromosome, short

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

read가 매핑된 시작과 끝 위치, 매핑 스코어 등이 있다. 서열 정렬 정보 추출 단계에서는 unique 형태로 정렬되고, MisMatch (MM)의 개수가 short read 길이의 5%를 넘지 않는 read 들을 추출한다.

추출된 정보들은 Expression profile과 SNP, Indel을 탐지 (call)하는데 사용된다. Expression profile은 유전자에 정렬된 short read의 양을 측정하는 것이다. 약 20,000-30,000개로 추정되는 인간 유전자의 크기는 개인마다 다르다. 그러므로 유전자에 정렬된 short read의 양을 정규화 (normalization)하는 단계가 필요하다. 현재 NGS를 이용한 RNA-Seq의 Expression profile에서는 정규화 방법으로 Reads Per Kilobase of exon per Million mapped sequence reads (RPKM) [4]을 사용한다. RPKM은 한 번의 시퀀싱 실험으로 얻어내는 short read의 길이를 1 Mbp, 유전자 내 엑손의 길이를 1 Kbp라고 가정함을 원칙으로 한다. 그러므로 모든 유전자에 대해 정규화된 expression profile 데이터를 획득하기 위해 (식 1)로 RPKM 결과를 얻어낸다.

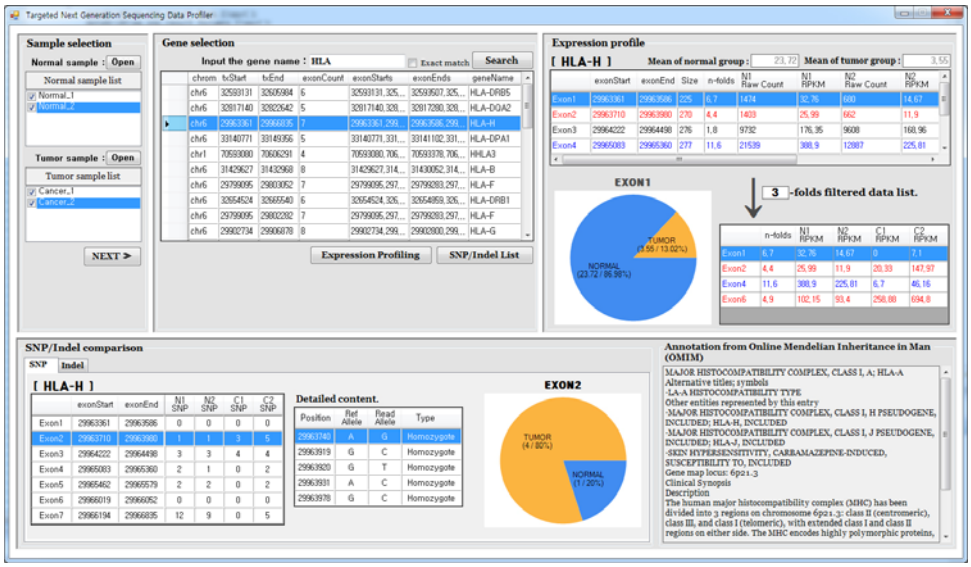

그림 2. 사용자 인터페이스

RPKM = (# of bp of total aligned reads) / (# of total generated reads / 1,000,000) / (the length of exon / 1,000) ..... (식 1)

SNP, Indel의 변이 또한 추출된 서열 정보를 이용한다. SNP을 추출하기 위해서 참고 문헌 [5]에서 사용된 filter 조건을 사용하였다. SNP으로 판명되기 위한 필터 조건은 unique로 align된 리드 개수가 4개 이상이고, read의 평균 quality가 20 이상이어야 한다. Homo/Hetero SNP 결정 비율은 90%이다. Indel의 경우 SNP과 필터 조건이 같으나 Homo/Hetero 비율이 60%이다.

RNA-Seq 데이터 분석 엔진에서 산출된 결과를 분석하는 방법은 다음과 같다. RNA Sequencing의 경우 실험군과 대조군의 비교를 통해서 유의성 있는 유전자를 찾아내는데, 예를 들어, 정상그룹과 암이나 특정 질환을 가진 환자군의 데이터를 비교하는 것이다. 본 논문에서 구현한 시스템은 Expression profiler에서 Normal 군과 Tumor 군의 유전자 비교에서 RPKM 데이터를 사용하고, SNP과 Indel 비교는 Normal 군과 Tumor 군에서 탐지된 변이 (variants)의 수를 사용한다. Expression profile 및 SNP/Indel 비교 모두 Normal 군에서 적은 수치를 보이고 Tumor 군에서 높은 수치를 보이거나 반대로, Normal 군에서 높은 수치를 보이고, Tumor 군에서 낮은 수치를 보이는 것을 질환과 연관이 있는 후보 마커 유전자로 선택할 수 있다.

그림 2는 본 시스템의 사용자 인터페이스 화면을 캡쳐한 것이다. 본 시스템의 사용자 인터페이스는 사용자가 선택한 유전자에 대해 Expression Profile과 SNP/Indel 비교를 비쥬얼하게 제공하여 엑손 간 발현 차이와 변이 비교의 정확성을 높일 수 있다.

3. 결 론

최근 개인 맞춤 의학에 대한 관심과 함께 단백질 발현과 밀접한 연관이 있는 인간 유전체의 유전자 영역/엑손 영역 분석에 관한 연구에 많은 관심이 집중되고 있다. 그러나 국내외적으로 NGS 기반의 RNA-seq을 이용한 유전자 영역/엑손 영역분석 방법에 관한 연구는 아직 초기 단계이고, 도구 개발도 매우 미미한 실정이다. 본 논문에서는 NGS 기반의 tanscriptome/exome sequencing 데이터 분석을 위한 coverage 추출, expression profiling, 변이 영역 (SNP, Indel) 추출 알고리즘을 제안하였다. 또한 제안된 알고리즘을 이용한 시각적 분석 도구를 구현하여 제안된 방식의 유용성을 입증하였다. 현재 분석 도구의 성능 개선에 관한 연구를 진행중이며, 또한 AS (Alternative Splicing)과 Gene Fusion 결과 및 clinical implication 검색 기능 등도 본 연구의 결과를 바탕으로 추진해야 할 중요한 과제이다.

참 고 문 헌

[1] Z. Wang, M. Gerstein and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," Nature reviews | Genetics, Vol. 10, 2009.
[2] Ng et al.,"Targeted capture and massively parallel sequencing of 12 human exomes," Nature, Vol. 461, No. 10, 2009.
[3] 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
[4] Mortazavi et al., "Mapping and quantifying mammalian transcriptomes by RNA-Seq," Nature methods, Vol. 5 No. 7, 2008.
[5] Kim et al., "A highly annotated whole-genome sequence of a Korean individual," Nature, Vol. 460, pp. 1011-1015, 2009.