차세대 시퀀싱 데이터를 위한 SNP 분석 방법

홍상균, 이덕해, 공진화, 김덕근, 홍동완, 윤지희
한림대학교 컴퓨터공학과
e-mail:hongsk@hallym.ac.kr

SNP Analysis Method for Next-generation Sequencing Data

Sang-kyoon Hong, Deok-hae Lee, Jin-hwa Kong,
Deok-Keun Kim, Dong-wan Hong, Jee-hee Yoon
Dept of Computer Engineering, Hallym University

요 약

최근 차세대 시퀀싱 기술의 급속한 발전에 따라 서열 정보의 해독이 비교적 쉬워지면서 개인별 맞춤 의학의 실현에 대한 기대와 관심이 높아지고 있다. 각 개인의 서열 정보 사이에는 SNP (single nucleotide polymorphism), Indel, CNV (copy number variation) 등의 다양한 유전적 구조 변이가 존재하며, 이러한 서열 정보의 부분적 차이는 각 개인의 유전적 특성 및 질병 감수성 등과 밀접한 관련을 갖는다. 본 연구에서는 차세대 시퀀싱 결과로 산출되는 수많은 짧은 DNA 서열 조각인 리드 데이터를 이용한 SNP 추출 알고리즘을 제안한다. 제안된 알고리즘에서는 레퍼런스 시퀀스의 각 위치에 대한 리드 시퀀스의 매핑 정보를 기반으로 SNP 후보 영역을 추출하며, 품질 정보 등을 활용하여 에러 발생률을 최소화한다. 또한 대규모 시퀀싱 데이터와 SNP 구조 변이 데이터의 효율적인 저장/검색을 지원하는 시각적 분석 도구를 구현하여 제안된 방식의 유용성을 검증한다.

1. 서론

2000년대 초 인간 게놈 프로젝트 (human genome project) [1]의 초안 발표를 통해 개인의 서열정보 획득 및 분석의 틀이 갖추어지면서 서열 정보를 기초로 질병의 예측 및 치료를 위한 연구 초석이 만들어졌다. 초기 유전체 분석 비용은 30억 달러 이상이 소요되었으나, 최근 차세대 시퀀싱 기술의 발전으로 2012년 이후에는 1,000달러 이하의 비용으로 개인의 유전체 시퀀싱 (personalize sequencing)이 가능할 것으로 예측되고 있다 [2].

인간의 유전체는 A, C, G, T의 네 종류 염기로 구성된 약 30억bp (basepair) 크기를 갖는 DNA 사슬로 이루어져 있다. 그러나 개인의 서열 정보 사이에는 다양한 크기와 형태의 유전적 구조 변이 (genetic structural variation)가 존재하며, 이러한 변이가 유전적 특성을 나타내기도 하며, 유전병의 발병 원인이 되는 것으로 알려져 있다 [3]. 유전적 구조 변이는 작은 영역의 시퀀스 미스매치 (small sequence mismatch), 삽입 (insertion), 삭제 (deletion), 전이 (inversion), 단위 반복 변이 (copy number variation), 그리고 SNP 등으로 구분된다.

최근 차세대 시퀀싱 기술을 기반으로 하는 유전적 구조 변이에 관한 다양한 연구가 주목 받고 있다 [4]. 차세대 시퀀싱 기술은 유전정보를 지닌 혈액과 같은 샘플로부터 서열 정보를 읽어오는 방법으로서, 생성되는 정보는 짧게는 수 십bp에서 길게는 수 천bp이상의 크기를 갖는 대량의 DNA 서열 정보로 이루어지며, 이러한 서열 정보를 리드 (read)라 부른다. 그러나 이러한 대규모의 차세대 시퀀싱 데이터를 처리하여 유전적 구조 변이를 효율적으로 추출하기 위한 알고리즘 혹은 시스템 개발에 관한 연구는 아직 미비한 상황이다.

본 연구에서는 차세대 시퀀싱 데이터를 이용한 효율적인 SNP 분석 방법을 제안한다. 제안하는 방법은 차세대 시퀀싱 데이터인 리드를 기존에 완성된 레퍼런스 서열에 매핑하고 매핑된 위치를 기반으로 레퍼런스와 리드의 염기 서열을 비교하는 방법으로 염기의 품질 점수 (quality score) [5]와 염기 서열의 분포 정보를 활용하여 SNP 분석을 수행한다. 또한 현재 우리 연구실에서 개발 중인 NGSDAT (next generation sequencing data analysis tool) [6]에 SNP 분석 기능을 구현하여, 제안된 방식의 유용성을 검증하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구로서 차세대 시퀀싱 기법과 리드 매핑 및 SNP 분석 방법을 살펴본다. 제 3장에서는 SNP 분석 알고리즘을 제시하고, 구현된 NGSDAT 시스템의 사용 예를 활용하여, SNP 추출 및 검색 방식을 설명한다. 마지막으로 제 4장에서는 본 논문을 요약하고, 결론을 내린다.

2. 관련 연구

2.1. 차세대 시퀀싱 기법과 리드 매핑 방법

시퀀싱 데이터 분석을 정확하게 수행하기 위해서는 매

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2010-0017194)

우 높은 리드 커버리지 (read coverage) 데이터를 필요로 한다. 제 1세대의 생어 시퀀싱 (sanger sequencing) 기술 [7]은 1kbp 정도의 리드를 생성할 수 있지만 매우 고가의 실험이므로 리드 커버리지를 높이기 어려웠다. 이후 시퀀싱 기술의 발달로 시퀀싱 기술 보유회사에서는 36-100bp 정도의 짧은 리드를 대규모로 생성해내는 머신을 개발하였으며, 이러한 기술을 기가 시퀀싱 (giga-sequencing) 혹은 차세대 시퀀싱 기술이라고 부른다. 2010년 시판되는 Illumina의 HiSeq 2000 [8]의 경우, 8일간 소요되는 한 번의 시퀀싱 작업에 의하여 2×100bp의 리드를 200G까지 산출 가능하며, 이는 시퀀스 분석을 위한 적정 커버리지로 예측하는 30×이상의 2명의 유전체 시퀀싱 분량에 해당한다.

리드 매핑은 각 리드를 기존에 완성된 서열의 유사 영역에 매핑하는 방법이다. 현재 매핑 알고리즘에 대한 다양한 연구가 진행되고 있으며, SOAP [9], MAQ [10], BWA [11]와 같은 다양한 매핑 도구들이 알려져 있다. Illumina의 ELAND [8]와 같이 시퀀싱 기술 보유회사에서도 자체적으로 매핑 도구를 제공한다. 이들 매핑 도구들은 서픽스 트리 [12], BWT [13] 등의 다양한 기법을 사용하여 처리 속도를 향상시키고 있으며, 매핑의 정확도를 개선하고 있다.

2.2. SNP 분석 툴

기존의 SNP 분석 툴은 주로 마이크로어레이 실험 등에 의하여 발견된 SNP 데이터베이스를 검색하여 유전자나 질병과의 연관성을 밝히는 작업을 지원하기 위하여 설계, 구현되었다 [14]. SNP 데이터베이스 검색을 위한 대표적인 검색 툴로서 SNPHunter [14], SNPicker [15]등을 들 수 있다. 최근 많은 논문에서 차세대 시퀀싱 데이터 분석에 의한 SNP 추출 결과를 보고하고 있다 [16][17]. 그러나 이 들 연구는 각각 서로 다른 실험적 방식과 수작업을 병행하여 정확한 SNP를 추출하는데 주력하고 있으며, SNP 추출 알고리즘 등에 관한 체계적 연구는 거의 이루어지지 않은 상황이다.

3. SNP 분석 기법과 NGSDAT

본 장에서는 제안하는 SNP 분석 알고리즘을 설명하고, 현재 개발 중인 NGSDAT 분석 시스템을 사용하여 SNP 분석 및 결과 검색 방식에 대하여 기술한다.

3.1. SNP 분석 알고리즘

알고리즘 1 : SNP Detection

Input : reference sequence S, set of mapped reads R, SNP threshold Ts, zygosity threshold Tz, coverage threshold Tc
Output : set of SNPs SNP

```
1. initialize SNPs set SNP;
2. structure of reference base and mapped read base RefMR;
3. for each reference base S[i] of the S do
4.   └AggrRefRead(RefMR, S[i], R);
5. for each reference base RefMR[i].RB of the RefMR do
6.   | if RefMR[i].MRBC >= Tc then
7.   └ └SNPcall(cSNP, RefMR[i].RB, RefMR[i].MB, RefMR[i].MQ,Ts);
8. ZygosityTest(SNP, Tz);
9. RegionTest(SNP);
10.return SNP;
```

SNP는 개인 간의 염기 배열 상에 발생하는 차이로 DNA 사슬의 특정 부위에 서로 다른 염기를 가지고 있는 경우를 말한다. 본 연구에서는 리드를 레퍼런스 서열에 매핑하여 매핑된 위치를 기반으로 하여 레퍼런스와 리드간의 서열 차이를 통해 SNP를 검출한다. SNP 검출을 위한 알고리즘 SNP Detection을 알고리즘 1에 보인다. 알고리즘 1은 레퍼런스 서열 S와 리드들의 집합 R과 커버리지 임계값 Tc, SNP 검출 임계값 Ts, 접합자 구조 (zygosity) 임계값 Tz를 입력받아 검출된 SNP를 반환한다. 알고리즘 1의 동작 과정을 3단계로 나누어 설명하면 다음과 같다.

첫 단계는 초기화 단계로, line 1-2에서는 SNP를 기록하기 위한 집합 SNP를 설정하고 초기화 레퍼런스의 각 염기와 각 염기 위치에 매핑된 리드의 염기의 목록을 저장하는 구조체 RefMR을 설정한다. RefMR는 레퍼런스의 염기를 저장하는 RB, 매핑된 염기의 목록 MB, 매핑된 염기의 개수 MRC, MB의 각 염기 품질 점수 MQ로 구성된다.

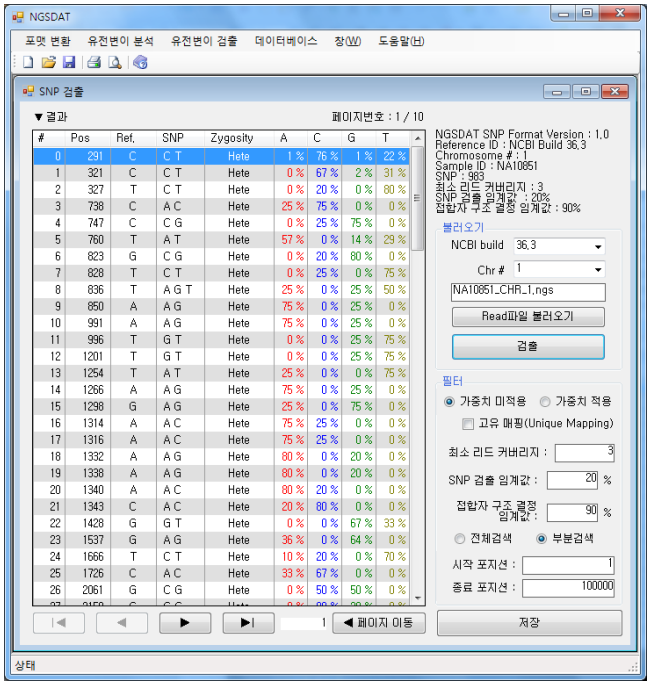
두 번째 단계는 SNP추출 단계로, line 3-4에서는 SNP 검출의 대상이 되는 레퍼런스 서열의 각 염기와 각 염기에 걸쳐 매핑된 리드의 염기의 집합을 구성한다. 레퍼런스 염기와 리드 염기의 집합은 AggrRefRead()함수를 통해 매핑된 리드에서 레퍼런스와 동일하게 위치하는 리드의 단일 염기만을 분리하여 레퍼런스 염기와 이 영역에 걸쳐 매핑된 리드들에서 해당 영역의 염기들만 추출하여 품질 점수와 함께 RefMR에 저장된다. line 5-7에서는 RefMR에 저장된 레퍼런스 염기들에 대하여 SNPcall()함수를 통해 SNP 검출을 수행한다. 이때, 검출의 정확도를 높이기 위하여 각 레퍼런스 염기에 매핑된 리드의 수 RefMR[i].MRBC가 커버리지 임계값 Tc보다 낮은 영역을 SNP 검출에서 제외한다. 예를 들어, 레퍼런스 염기가 C인 영역에 매핑된 리드가 한 개이고 염기가 T였다면 100% 다른 염기가 매핑되었기 때문에 동형접합 (homozygous) SNP로 분류될 수 있다. 하지만 매핑된 리드가 적어 커버리지가 낮은 영역은 정확도가 낮기 때문에 분석 대상에서 제외하여야 한다. 다음 SNPcall()에서는 레퍼런스 염기 RefMR[i].RB와 매핑된 염기 집합 RefMR[i].MR, 매핑된 각 염기의 퀄리티 점수 RefMR[i].MQ, 그리고 SNP 검출 임계값 Ts를 입력받아 SNP 검출을 수행한다. 함수 SNPcall()의 첫 단계에서는 매핑된 염기들의 품질 점수를 가중치로 사용하여 각 염기 A, C, G, T의 출현빈도를 산출한다. 품질 점수로는 각각의 시퀀싱 머신의 결과 포맷에 따라 산출되는 리드의 산출 정확도를 사용하여 프레드 품질 점수 (Phred quality score) [5]를 사용한다. 예를 들어, 프레드 품질 점수가 99.99%를 가지는 염기의 경우 0.9999의 가중치를 적용하며, 90%를 가지는 염기의 경우 가중치 0.9를 적용한다. 두 번째 단계에서는 SNP 검출 임계값 Ts이상의 출현빈도를 가지는 염기들을 산출하여 집합 SNP에 저장한다. 이 과정에서 산출된 염기가 한 종류이고 레퍼런스의 염기와 동일한 경우에는 SNP가 아닌 동일 염기가 매핑된 것이기 때문에 SNP 집합으로 저장하지 않는다.

세 번째 단계는 SNP 관련 정보 추출 단계로, line 8에

서는 산출된 SNP의 접합자 구조 (zygosity)를 결정한다. 접합자 구조에는 동형접합과 이형접합 (heterozygous)이 있다. 동형접합은 SNP로 추출되는 염기가 매우 높은 출현 빈도를 보이는 단일 염기인 경우를 의미한다. 이형접합은 SNP로 추출되는 염기들의 출현빈도가 높지 않고 다수의 염기가 선택되는 경우를 의미한다. 특히, 이형접합 SNP는 질병과의 관련성이 매우 높다고 알려져 있다. 접합자 구조는 함수 ZygosityTest()에 의해 결정되며, SNP로 추출된 염기 중에 출현빈도가 임계값 Tz를 넘을 경우 동형접합으로 분류하고 그 외의 경우 이형접합으로 분류한다. line 9에서는 검출된 SNP의 영역 정보를 검사한다. 유전체 서열에는 유전자 (gene), 엑손 (exon), 반복 영역 (repeat region)과 같은 특정 영역이 존재한다. 특히 엑손은 단백질 (protein) 합성에 관여하는 서열 정보이기 때문에 엑손에서 발생된 SNP는 유전 형질의 변화에 직접적인 영향을 주게 된다. 반복 영역은 유전체 서열에 존재하는 짧거나 긴 부분 서열이 다른 영역에서도 나타나는 경우로 유전학적 의미가 낮거나 리드 매핑 과정에서 반복 영역에 많은 리드가 잘못 매핑되는 경우가 발생하기도 한다. 따라서, 검출된 SNP가 어떠한 영역에서 발생하였는지의 여부를 확인할 필요가 있다. 함수 RegionTest()를 통해 각 SNP가 위치한 영역을 유전자, 엑손, 반복 영역 데이터베이스와 연동하여 영역 정보를 확인하여 기록한다. 영역 정보는 UCSC 게놈 브라우저 (http://genome.ucsc.edu)에서 제공하는 데이터베이스 [18]로부터 다운로드하여 활용하였다.

3.2. SNP 추출 및 결과 분석 방법

본 시스템에 의한 SNP 분석 결과 출력 화면을 그림 1에 보인다. 본 시스템에서는 SNP 분석을 위한 파라메터로서 가중치 적용 여부, 고유 매핑 (unique mapping) 적용 여부, 최소 리드 커버리지, SNP 검출을 위한 임계값, 접합자 구조를 결정하는 임계값 등을 입력으로 받는다. SNP 분석 결과 파일은 표 1과 같은 포맷으로 저장되며, 사용자는 분석 결과를 다양한 형식으로 검색 가능하다. 출력 결과로 반환되는 접합자 구조는 두 가지 형태로 분류되며, 알고리즘 1에서 설명한 바와 같이 기본적으로 출현빈도 점수가 90%를 넘는 염기를 동형접합 SNP로 분류하고, 90%가 넘는 염기가 없을 경우에는 20%를 넘는 모든 염기를 이형접합 SNP로 분류한다. 그러나 이와 같은 임계값은 사용자에 의하여 변경이 가능하며, 임계값을 변화시켜 다양한 SNP 분석 결과를 얻을 수 있도록 지원한다.



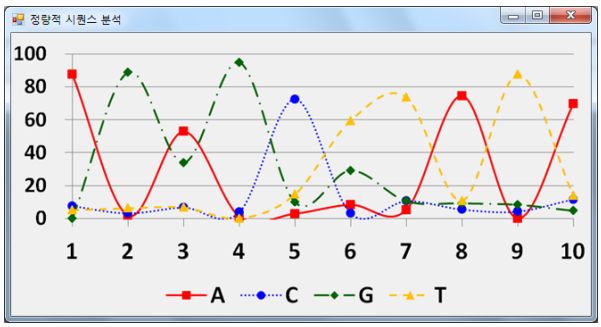
(그림 1) SNP 분석 결과 화면

(그림 2) 시퀀스 분석기의 실행 예

<표 1> SNP 검출 결과 포맷

컬 럼 명	설 명
번호	SNP 분석 번호
포지션	레퍼런스 서열 위치
영역 정보	레퍼런스 서열의 영역 정보
레퍼런스 염기	레퍼런스 서열의 염기
매핑 리드 수	현재 레퍼런스 포지션에 매핑된 리드 수
출현 횟수	각 염기의 출현 횟수
최하 품질	각 염기 품질의 최저 값
최고 품질	각 염기 품질의 최고 값
평균 품질	각 염기 품질의 평균 값
출현 빈도 백분율	각 염기의 출현 빈도의 백분율
가중치 출현 빈도	품질 가중치를 적용한 출현 빈도 백분율
접합자	접합자 구조 (zygosity)
SNP 염기	SNP로 추출된 염기 종류

또한 본 시스템에서는 검출된 SNP 결과의 자세한 분석을 위하여 시퀀스 분석기를 제공한다. 그림 2는 시퀀스 분석기의 사용 예를 보인다. 시퀀스 분석기는 레퍼런스 서열에 매핑된 모든 리드들의 시퀀스 정보를 구체적으로 확인/분석하는데 사용된다. 시퀀스 분석기는 기본적으로 리드의 서열 정보뿐만 아니라 리드의 퀄리티, 매핑 방향 등의 정보들을 제공하며, 따라서 이를 활용하여 만약 매핑된 리드가 높은 매핑 스코어를 가지고 있더라도 리드의 퀄리티 정보가 낮다면 이러한 SNP 결과에는 오류가 포함되어 있을 가능성이 높다고 판단할 수 있다. 시퀀스 분석기는 페이지 단위의 데이터 처리를 기본적으로 지원하고 있지만, 리드들이 페이지와 페이지 경계면에 매핑되는 경우를 고려하여 오버랩 영역을 설정할 수 있도록 지원하고 있다. 또한 그림에 보이는 바와 같이 리드 정보창 (read information window)을 열어 매핑된 각 리드들의 리드 아이디, 리드 시퀀스, 리드 퀄리티 등의 세부정보를 확인할 수 있도록 지원하고 있다.



(그림 3) 각 위치에 매핑된 염기의 정량적 분석 화면

다음의 그림 3은 시퀀스 분석기에서 제공하는 염기 출현 빈도 분석 화면의 예를 나타낸다. 이 그래프 분석기에서는 레퍼런스 시퀀스의 각 위치에 매핑된 리드의 염기 출현 빈도를 그래프로 형태로 출력하여 나타낸다. 화면에는 총 4가지의 그래프가 나타나며, 각각 A, C, G, T 염기의 출현 빈도 백분율을 나타낸다. 이와 같은 그래프 분석 기능은 SNP로 보고되는 염기와 이웃 염기 사이의 발현 특성/연관성 등을 추론, 검증하는데 사용된다.

4. 결론 및 향후연구

본 논문에서는 차세대 시퀀싱 기술을 활용한 SNP 분석 방법을 제안하였다. 제안하는 SNP 분석 방법은 차세대 시퀀싱 데이터를 레퍼런스 서열에 매핑하여 매핑된 리드의 염기 분포, 품질 점수 등의 정보를 활용하여 SNP 영역을 추출한다. 또한 제안하는 분석 방법을 활용한 차세대 시퀀싱 데이터 분석툴인 NGSDAT의 기능을 보였다. 향후 SNP 분석 기능에 대한 기존의 방법과의 비교 실험을 수행할 예정이며, SNP 이외의 유전적 변이 분석에 대한 연구 및 NGSDAT의 기능 개선 연구를 수행할 예정이다.

참고문헌

[1] http://www.genome.gov/10001772
[2] F. S. Robert, "The Race for the $1000 Genome," SCIENCE, Vol.311, No.5767, pp.1544-1546, 2006.
[3] R. Redon et al, "Global variation in copy number in the human genome," Nature, Vol.444, No.7118, pp.444 - 454, 2006.
[4] 홍상균 외 4인, "정렬된 리드의 통계적 분석을 기반으로 하는 CNV 검색 알고리즘," 정보처리학회논문지, Vol.16-D, No.5, pp.661-672, 2009.
[5] 이종극, "질병 유전체 분석법 2," 월드사이언스, 2010.
[6] 홍상균 외 4인, "GSDAT: 기가 시퀀싱 데이터 분석을 위한 도구," 한국정보과학회, Vol.36, No.2, pp.107-112, 2009.
[7] F. Sanger, S. Nicklen and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," Proc. National Academy of Sciences of the United States of America, Vol.74, No.12, pp.5463-5467, 1977.
[8] http://www.illumina.com
[9] R. Li et al, "SOAP: short oligonucleotide alignment program," Bioinformatics, Vol.24, No.5, pp.713-714, 2008.
[10] H. Li, J. Ruan and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," Genome Res., Vol.18, No.11, pp.1851-1858, 2008.
[11] H. Li, R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics, Vol.25, No.14, pp.1754-1760, 2009.
[12] P. Weiner, "Linear Pattern Matching Algorithms," Proc. 14th IEEE Annual Symp. on Switching and Automata Theory, pp.1-11, 1973.
[13] M. Burrows, D. J. Wheeler, "A block-sorting lossless data compression algorithm," Technical report, Digital Equipment Corporation, No.124, 1994.
[14] L. Wang et al, "SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management," BMC Bioinformatics, Vol.6, No.1, 2005.
[15] T. Niu, Z. Hu, "SNPicker: a graphical tool for primer picking in designing mutagenic endonuclease restriction assays," Bioinformatics, Vol.20, No.17, pp.3263-3265, 2004.
[16] J. Kim et al, "A highly annotated whole-genome sequence of a Korean individual," Nature, Vol.460, No.7258, pp.1011-1015, 2009.
[17] 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.
[18] http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database