인간 세포 Lineage 의 계층적 표현에 관한 연구†

박재순*, 권성규**, 오지원**, 이종혁*‡
*대구가톨릭대학교 인공지능·빅데이터공학과
**경북대학교 의과대학 해부학 교실
e-mail : jonghyuk@cu.ac.kr

A Study on the Hierarchical Expression of Human Cell Lineage

JaeSoon Park*, Seong Gyu Kwon**, Ji Won Oh**, JongHyuk Lee*
*Dept. of Artificial Intelligence & Big Data Engineering, Daegu Catholic University
**Dept. of Anatomy, School of Medicine, Kyungpook National University

요 약

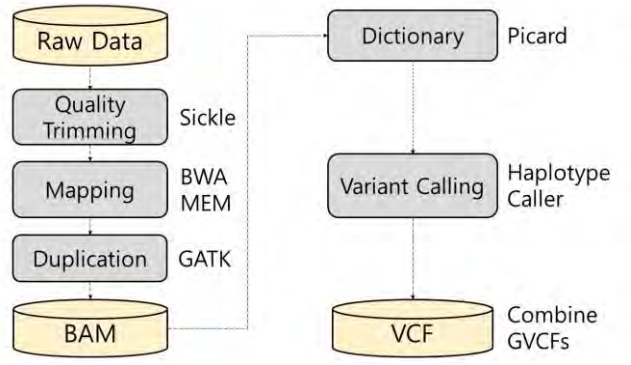
차세대 염기서열 분석 기술은 성능과 비용 면에서 매우 향상되어 한 개체 내 여러 세포의 유전자 분석이 가능한 수준이다. 한 개체 내 여러 조직 세포의 유전자는 모두 동일하지 않기 때문에 여러 조직 세포의 Lineage 를 계층적으로 표현하고 이를 조직 세포 간 변이 정도를 파악하는 데 활용한다면 암 돌연변이 발생 등을 미리 예측할 수 있다. 본 논문은 한 개체 내 여러 조직 간 변이를 관찰하기 위해 변이 검출 데이터를 계층적 군집 방법을 이용해 분석하고 이를 시각화 하는 방법을 제안한다. 실제의 8 개 조직 세포의 유전자를 분석하고 변이를 검출하여 Dendrogram 그래프로 시각화 하였다.

1. 서론

생물정보학은 생물학적 데이터를 이해하기 위한 방법과 소프트웨어 도구를 개발하는 학제 간 연구 분야이다[1]. 특히, 생물학적 데이터 중 유전자 데이터는 크기가 방대하고 구조가 복잡하다. 최근 차세대 염기서열 분석(Next Generation Sequencing, NGS) 기술의 등장과 함께 관련 기술의 발전으로 염기서열 분석은 빠르고 저렴해졌다. NGS 는 시퀀싱 단계를 통해 염기서열 데이터를 생성하고 이를 참조 염기서열에 매핑하는 단계를 수행하여 염기 종류와 위치 정보를 갖는 데이터를 생성한다. 이후 이 데이터를 이용해 변이 검출하는 단계를 수행하여 변이 데이터를 생성한다. 검출된 변이의 역추적을 통해 조직간 유사성을 확인하거나 특정 세포가 어느 시점에서 분화됐는지 확인할 수 있어 유전자 데이터 분석 결과는 질병 예방과 발생 생물학 등의 분야에서 폭넓게 활용된다.

체세포 변이는 한 개체의 일부 조직 혹은 일부 세포에서만 변이가 관찰되는 것으로 후천적으로 발생하는 암 돌연변이가 대표적인 예이다. 그래서 한 개체 내 여러 조직 간 변이를 관찰하기 위해 변이의 과정을 보여주는 것은 중요하다. 본 논문은 한 개체의 여러 체세포 변이를 트리 형태로 표현하여 조직 간 변이 정도를 확인하는 방법을 제안한다.

본 논문의 2 절에서는 본 논문에서 사용한 소프트웨어 도구를 설명하고 3 절에서는 변이 검출 과정과 트리 생성 결과를 설명한다. 마지막으로 4 절에서는 결론을 맺는다.



(그림 1) 변이 검출 과정

2. 관련연구

GATK(Genome Analysis Toolkit)[2]는 Broad Institute 에서 개발한 소프트웨어로 시퀀싱 데이터를 이용하여 유전자 내 변이 검출이 가능하도록 하는 여러 프로그램들로 구성되어 있다. 본 논문은 GATK 를 이용하여 체세포 염기서열을 바이너리 형태로 저장한 BAM 파일을 입력하여 변이 검출된 VCF 파일을 생성한다. (그림 1)은 변이 검출 과정을 보여준다. 파일을 중심으로 변이 검출 과정을 간단히 설명하면

† 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학지원사업의 연구결과로 수행되었음(2019-0-01056).
‡ 교신저자

다음과 같다. 먼저 염기서열과 염기의 정확도 정보가 포함된 FASTQ 파일을 참조 염기서열 파일에 매핑하여 BAM 파일을 생성한다. 다음으로 BAM 파일을 GATK 를 이용하여 변이 정보가 포함된 VCF 파일을 생성한다.

본 논문은 변이 트리를 생성하기 위해 조직 간 변이에 대해 계층적 군집 분석 방법을 사용한다. 계층적 군집 분석은 군집 간의 거리를 기반으로 클러스터링하는 방법이다. 그리고 변이 트리를 시각화하기 위해 Dendrogram 그래프를 사용한다.

3. 변이 트리

3.1 시퀀싱 데이터

본 논문은 <표 1>과 같이 한 개체에서 총 8 개의 체세포 시퀀싱 데이터를 사용하였다. 그리고 참조 염기서열 데이터로서 GRCh37 을 사용하였다.

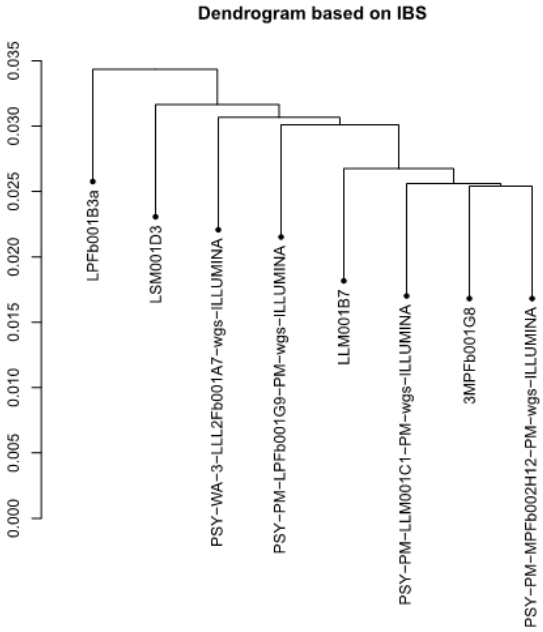
<표 1> 체세포 시퀀싱 데이터

명칭	체세포 부위
LPFb001B3a	Low Pubic Fibroblast
LSM001D3	Superior Rectus Eye Muscle Cell
PSY-WA-3LLL2Fb001A7-wgs-ILLUMINA	Low leg-2 Fibroblast
PSY-WA-LPFb001G9-wgs-ILLUMINA	Low Pubic Fibroblast
LLM001B7	Lateral Rectus Eye Muscle Cell
PSY-WA-LLM001C1-wgs-ILLUMINA	Lateral Rectus Eye Muscle Cell
MPFb001G8	Middle Pubic Fibroblast
PSY-WA-MPFb002H12-wgs-ILLUMINA	Middle Pubic Fibroblast

3.2 변이 검출 및 트리 생성

본 논문은 변이 검출을 통해 생성된 각 체세포의 VCF 파일을 병합하여 리스트로 만든 후 GATK 의 HaplotypeCaller 를 이용하여 GVCF 파일을 생성하였다. 이후 GATK CombineGVCFs 를 이용하여 변이 트리를 생성 시 활용하였다.

본 논문은 R[3]에서 지원하는 바이오 관련 패키지(gdsfmt, SNPrelate, ggplot2)를 이용하여 (그림 2)와 같이 변이 트리를 생성하였다. (그림 2)에서 보듯이 Low Pubic 과 Middle Pubic 간의 유사성이 가장 낮다. 이는 두 조직간 염기서열이 많이 다름을 의미하며 발생학적 Lineage 추적 등 집중 관찰이 필요한 후보 조직으로 간주할 수 있다.



(그림 2) 변이 트리

4. 결론 및 향후 과제

본 논문은 한 개체의 여러 조직 내 체세포 변이를 검출하고 이를 변이 트리로 표현하여 조직 간 변이 정도를 확인하는 방법을 제안하였다. 이 방법을 통해 한 개체 내 돌연변이를 빠르게 확인하여 암 발생 등의 조치를 할 수 있게 될 것이다.

본 논문은 계속해서 변이 검출 및 변이 트리 시각화 등의 일련의 단계를 자동화하는 방법을 연구하고자 한다.

참고문헌

[1] Wikipedia, https://en.wikipedia.org/wiki/Bioinformatics
[2] GATK, https://gatk.broadinstitute.org/hc/en-us
[3] The R Project for Statistical Computing, https://www.r-project.org/