DOI: 10.3745/KIPSTD.2009.16D.5.661

정렬된 리드의 통계적 분석을 기반으로 하는 CNV 검색 알고리즘

홍 상 균[†] · 홍 동 완[††] · 윤 지 희[†††] · 김 백 섭[††††] · 박 상 현[†††††]

요 약

인간의 유전체 서열에는 유전체 단위반복변위(copy number variation, CNV)를 포함하는 다양한 유전적 구조 변이(genetic structural variation)가 존재하며, 이는 기능적으로 질병에 대한 감수성, 치료에 대한 반응, 유전적 특성 등과 밀접한 관련이 있다. 본 논문에서는 기가 시퀀싱(giga sequencing)의 결과 산출되는 대량의 짧은 길이의 DNA 서열 데이터를 이용한 새로운 CNV 검색 방식을 제안한다. 제안하는 알고리즘에서는 레퍼런스 시퀀스에 DNA 서열 데이터를 서열 정렬시켜 각 레퍼런스 시퀀스의 위치에 대한 서열 데이터의 출현 빈도 정보를 얻은 후, 출현 빈도 정보의 패턴을 분석하여 통계적 유의성을 갖는 1kbp 이상의 연속 영역을 CNV 후보 영역으로 추출한다. 또한 제안된 알고리즘을 효율적으로 지원하기 위한 서열 정렬 방식에 대한 비교 및 분석을 수행한다. 제안된 기법의 유용성을 규명하기 위하여 다양한 실험을 수행하였다. 실험 결과에 의하면, 제안된 기법은 비교적 낮은 커버리지의 기가 시퀀싱 데이터를 이용하여 반복되거나 결실되는 다양한 형태의 CNV 영역을 효율적으로 검출하며, 또한 작은 사이즈의 CNV 영역에서부터 큰 사이즈의 CNV 영역까지 다양한 크기의 CNV 영역을 효율적으로 검출할 수 있는 것으로 나타났다.

키워드 : 유전체 단위반복변위(CNV), 기가 시퀀싱, 서열 정렬, 통계적 유의성

A CNV detection algorithm based on statistical analysis of the aligned reads

Sang-Kyoon Hong[†] · Dong-Wan Hong[††] · Jee-Hee Yoon[†††] · Baek-Sop Kim[††††] · Sang-Hyun Park[†††††]

ABSTRACT

Recently it was found that various genetic structural variations such as CNV(copy number variation) exist in the human genome, and these variations are closely related with disease susceptibility, reaction to treatment, and genetic characteristics. In this paper we propose a new CNV detection algorithm using millions of short DNA sequences generated by giga-sequencing technology. Our method maps the DNA sequences onto the reference sequence, and obtains the occurrence frequency of each read in the reference sequence. And then it detects the statistically significant regions which are longer than 1Kbp as the candidate CNV regions by analyzing the distribution of the occurrence frequency. To select a proper read alignment method, several methods are employed in our algorithm, and the performances are compared. To verify the superiority of our approach, we performed extensive experiments. The result of simulation experiments (using a reference sequence, build 35 of NCBI) revealed that our approach successfully finds all the CNV regions that have various shapes and arbitrary length (small, intermediate, or large size).

Keywords : Copy Number Variation(CNV), Giga-Sequencing, Sequence Alignment, Statistical Significancy

1. 서 론

2002년에 초안이 발표된 휴먼 게놈 프로젝트(human genome project: http://www.genome.gov/10001772)는 인간의 서열 정보 해석을 기반으로 하는 질병의 예측 및 치료 연구를 위한 초석이 되었다. 이 들 유전체(genome) 분석을 위한 비용은 2000년 초에는 약 30억 달러 이상이 소요되었으나, 최근의 보고서에서는 2012년 이후에 1인당 유전체 분석 비

※ 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2009-0065503).
† 준 회 원 : 한림대학교 컴퓨터공학과 박사과정
†† 정 회 원 : 한림대학교 바이오메디컬학과 겸임조교수
††† 정 회 원 : 한림대학교 컴퓨터공학과 교수(교신저자)
†††† 정 회 원 : 한림대학교 컴퓨터공학과 교수
††††† 종신회원 : 연세대학교 컴퓨터과학과 부교수
논문접수 : 2009년 2월 9일
수 정 일 : 1차 2009년 5월 27일
심사완료 : 2009년 5월 27일