레퍼런스 시퀀스의 특성을 고려한 HLA 영역에서의 CNVR 탐지

(CNVR Detection Reflecting the Properties of the Reference Sequence in HLA Region)

이 종 근 † 홍 동 완 †
(Jongkeun Lee) (Dongwan Hong)

윤 지 희 ††
(Jeehee Yoon)

요 약 본 논문에서는 레퍼런스 시퀀스에 기가 시퀀싱 데이터를 매핑하여 얻어지는 커버리지 데이터를 이용한 모양 기반의 단위반복변이 영역 (CNVR) 추출 방식을 제안한다. 제안하는 CNVR 검색 알고리즘은 후보 영역 추출 단계와 후처리 단계로 이루어진다. 후보 영역 추출 단계에서는 추출하고자 하는 CNV의 모양을 입력 변수로 조절하여 다양한 높이 및 크기를 갖는 CNV 후보 영역을 추출한다. 다음, 후처리 단계에서는 레퍼런스 시퀀스와 기가 시퀀싱 데이터에 포함되어 있는 시퀀싱 에러 문제를 보완하기 위하여, 레퍼런스 시퀀스의 에러 영역 보정, GC-content 영역 보정 등의 정제 과정을 거친 후, 최종 CNVR을 추출한다. 제안된 방식의 유용성을 보이기 위하여 "1000 게놈 프로젝트"에 의하여 공개된 실 데이터를 이용한 다양한 실험을 수행하였으며, DGV를 이용하여 추출된 CNVR의 정확도를 검증하였다. 실험 결과에 의하면 제안된 방식은 HLA 영역에 존재하는 반복되거나 결실되는 다양한 모양의 CNV를 효율적으로 검출하였다.

키워드 : 기가 시퀀싱, 유전체 단위반복변이, 모양기반 추출

Abstract In this paper, we propose a novel shape-based approach to detect CNV regions (CNVR) by analyzing the coverage graph obtained by aligning the giga-sequencing data onto the human reference sequence. The proposed algorithm proceeds in two steps: a filtering step and a post-processing step. In the filtering step, it takes several shape parameters as input and extracts candidate CNVRs having various depth and width. In the post-processing step, it revises the candidate regions to make up for errors potentially included in the reference sequence and giga-sequencing data, and filters out regions with high ratio of GC-contents, and returns the final result set from those candidate CNVRs. To verify the superiority of our approach, we performed extensive experiments using giga-sequencing data publicly opened by "1000 genome project" and verified the accuracy by comparing our results with those registered in DGV database. The result revealed that our approach successfully finds the CNVR having various shapes (gains or losses) in HLA (Human Leukocyte Antigen) region.

Key words : Giga-sequencing, Copy Number Variation, Shape-based extraction

1. 서 론

최근 가장 주목 받는 바이오 분야의 연구 성과의 하나로서 기가 시퀀싱(giga-sequencing) 기술의 발전을 들 수 있다. 기가 시퀀싱 기술의 발전은 인간을 비롯한 다양한 생명체의 유전체 시퀀싱을 비교적 저가의 비용으로 가능하게 하였다는 점에 가장 큰 의의를 들 수 있다. 또한 기가 시퀀싱 기술은 다양한 생물체의 유전체 시퀀스에 존재하는 유전적 구조 변이(genetic structural variation)를 추출하는 연구에 사용될 수 있다.

유전체에 존재하는 유전적 구조변이로서 단일염기변이(SNP), 삽입/삭제(insertion/deletion), 전이(inversion), 유전체 단위반복변이(Copy Number Variation, 또는 Copy Number Polymorphism, 이하 CNV로 약칭함) 등을 들 수 있다. 특히, CNV는 최근 유전체학 연구 분야에서 가장 많은 관심의 대상이 되고 있는 연구 분야이다. CNV는 통상적인 염기서열 분석법으로 확인되는 1Kbp 이하의 변이와, 현미경으로 관측될 수 있는 3Mbp 이상의 변이 사이의 영역, 즉 1Kbp~3Mbp 사이의 서열이 반복되거나 결실되는 변이로 정의된다. 또한 복수개의 CNV가 근접하여 발생하는 영역은 통합된 하나의 CNVR (Copy Number Variation Region)로 정의된다. 2004년에 이러한 종류의 변이가 건강한 사람의 유전체에도 많

· 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2010-0017194)
· 이 논문은 제36회 추계학술발표회에서 '레퍼런스 시퀀스의 특성을 고려한 HLA 영역에서의 CNVR 탐지'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 한림대학교 컴퓨터공학과
jeikei@hallym.ac.kr
dwhong@hallym.ac.kr
†† 종신회원 : 한림대학교 컴퓨터공학과 교수
jhyoon@hallym.ac.kr
(Corresponding author임)

논문접수 : 2009년 12월 23일
심사완료 : 2010년 3월 5일

이 존재한다는 것이 보고되고, 많은 후속 연구에 의해 전체 유전체에 광범위하게 분포된다는 것이 본격적으로 보고되면서[1], CNV가 인간유전체의 다양성에 어느 정도 기여하고, 인간의 질병이나 형질과 어떠한 관련성을 가지는가에 대한 연구에 많은 관심이 모아지고 있다.

본 논문에서는 기가 시퀀싱 데이터를 이용한 효율적인 CNVR 추출 방식을 제안한다. 제안하는 방식에서는 레퍼런스 시퀀스에 기가 시퀀싱 데이터를 매핑하여 얻어지는 커버리지(coverage) 데이터를 분석하여, 특정 모양(shape)을 갖는 영역을 CNV 영역으로 추정한다. 그러나 처리 대상의 레퍼런스 시퀀스와 기가 시퀀싱 데이터에는 많은 에러가 포함되어 있으므로 추출된 CNV 영역에는 다수의 거짓 실재(false positive) 영역이 포함될 수 있다. 이를 보정하기 위하여 레퍼런스 시퀀스의 특성 분석에 의한 레퍼런스 시퀀스의 에러 영역 보정과 GC-content 영역 보정[2] 등의 후처리 과정을 수행한 후, 최종 CNVR을 추출한다. 제안된 방식의 유용성을 보이기 위하여 1000 게놈 프로젝트에 의하여 공개된 실 데이터를 이용한 다양한 실험을 수행하였다. 면역 체계와 관련이 있는 유전자가 집중적으로 존재하는 HLA(Human Leukocyte Antigen) 영역에서의 CNVR 추출 실험을 수행하였으며, DGV(Database of Genomic Variant)[3]와 비교하여 정확도를 검증하였다.

2. 관련연구

2.1 인간 유전체 분석을 위한 국제적 연구 동향

2008년 1월에 시작된 "1000 게놈 프로젝트"[4]는 인간의 유전자 변이에 관한 통합된 카탈로그를 구축하기 위한 국제적인 연구 노력으로 볼 수 있다. 이 프로젝트는 영국, 중국, 미국을 포함한 전 세계 연구 기관의 여러 전문 분야에 걸친 연구팀의 전문적 기술을 살려서, Hapmap 프로젝트에 참가했던 아프리카, 미국, 중국, 유럽의 1,000여명의 익명 지원자들의 게놈을 빠르고 적은 비용으로 새롭게 개발된 기술을 사용하여 시퀀싱 해내는 것을 초기 목적으로 삼고 있다. 모든 결과 데이터는 과학 커뮤니티와 일반 대중들에게 무료로 공개 데이터베이스를 통해 즉시 엑세스할 수 있도록 공개되며, 현재 그 일부가 공개되어 있다.

2.2 기존의 CNV 탐지 기법

CNV 영역을 탐지하기 위한 방법은 크게 마이크로어레이 기반 기술을 이용한 방법[6]과 서열 비교법[7]으로 분류할 수 있다.

마이크로어레이 기반 기술을 이용한 방법[6]은 BAC array 또는 oligonucleotide array 등을 이용한 실험적 방식이다. BAC array는 표적 DNA 염기 서열에 특이성이 높고 SNR(Signal-to-Noise Ratio)이 높아서 민감도도 좋지만, BAC 클론 자체의 크기가 CNV보다 큰 경우가 있으므로 크기가 50Kbp 이하인 CNV의 검출이 어렵고 실제 CNV 크기보다 과대 측정하는 경향이 있어서 CNV의 정확한 크기를 측정하는데 어려움이 있다. Oligonucleotides는 크기가 25-80bp 정도이며, 디자인이 용이하고 적용 범위가 넓으며, 고밀도로 정렬할 수 있어 높은 해상도를 가진 aCGH(array-based Comparative Genomic Hybridization) 플랫폼으로 널리 이용되고 있다. 최근 발표된 aCGH 데이터를 이용한 CNV 탐지 알고리즘들은 참고 문헌 [8]에 잘 정리되어 있다. 이들 방식은 비교적 저가의 실험 비용이 드는 장점을 갖는다. 그러나, 마이크로어레이 실험이 노이즈에 약한 특성으로 인하여 작은 사이즈의 CNV 발견에 적합하지 않으므로 주로 수백 Kbp 이상의 큰 사이즈의 CNV 발견에 유용한 것으로 알려져 있다.

서열 비교법은 기존에 완성된 어셈블리 시퀀스들을 상호 비교하여 인간 유전체에 존재하는 CNV 등의 구조적 변이를 찾아내는 방식이다[9]. 이 방식은 마이크로어레이 기술을 이용하는 방식에 비하여 CNV 영역을 보다 정확하게 밝힐 수 있는 장점이 있어, 작거나 중간 정도 사이즈의 CNV 발견 방법으로도 적용 가능한 것으로 알려져 있다. 그러나 이 방식은 어셈블리가 완성된 시퀀스를 비교 대상으로 하는 CNV 검색 방식으로서, 초기 어셈블리 시퀀스 생성을 위한 과도한 비용이 문제가 되며, 어셈블리가 완성된 시퀀스가 많지 않아 적용이 어렵다.

이와 같은 단점들을 보완할 수 있는 새로운 방법으로서 기가 시퀀싱 기술을 이용한 CNV 검색 방식을 들 수 있다[10,11]. 이들 방식에서는 서로 다른 두 사람 샘플의(하나를 test 샘플로, 다른 하나를 control 샘플로 설정하여) 커버리지 데이터를 얻은 후, 해당 유전체 영역에서의 두 커버리지 값의 비율(ratio)의 변화를 통계적 모델로 표현한 후, 이를 기반으로 하는 CNV 검출 방식을 제안하고 있다. 그러나 이들 방식은 마이크로어레이 데이터 분석에 사용된 알고리즘을 그대로 사용하거나 혹은 확장 적용하고 있어, 기가 시퀀싱 데이터의 정확도를 높일 수 있는 특성을 적절히 살리지 못하고 있다는 문제점이 있다.

3. CNVR 탐색 방법

3.1 모양 기반의 CNV 영역 추출

서로 다른 두 서열을 비교하여 검색하는 유전적 구조 변이 중의 하나인 CNV는 다음과 같이 정의된다. 임의의 서브 시퀀스가 양쪽 서열에서 발견되는데 한쪽 서열에서 추가적인 카피(copy)를 발견할 수 있는 경우로서 그 영역의 크기가 1Kbp 이상의 경우 이를 CNV라고 부른다.

본 연구에서는 기가 시퀀싱의 결과 산출되는 수 많은 짧은 DNA 시퀀스인 리드를 이용한 CNVR 탐지 방법을 개발한다. 비교 대상이 되는 두 시퀀스로서 이미 시퀀싱이 완성된 표준의 레퍼런스 시퀀스와 임의의 테스트 시퀀스를 가정하며, 테스트 시퀀스 상에 존재 하는 CNV영역을 검색하는 방법을 개발한다. 단, 여기에서 테스트 시퀀스는 기가 시퀀싱 머신에서 생성된 수많은 리드로 이루어져 있는 경우를 가정한다.

기본적인 아이디어는 다음과 같다. 테스트 시퀀스의 수많은 리드를 레퍼런스 시퀀스에 매핑시킨 후, 각 레퍼런스 위치에 매핑된 리드의 수를 이용하여 CNV를 추정한다. 여기에서 각 레퍼런스 위치에 매핑된 리드의 수를 나타내는 정보를 커버리지라고 부른다. 즉, 커버리지 정보를 분석하여, 만약 레퍼런스 시퀀스의 임의의 영역에서 커버리지 정보가 주위에 비하여 상대적으로 높게 나타났거나 혹은 낮게 나타났다면 해당 영역의 서브 시퀀스가 테스트 시퀀스에 추가적으로 반복되어 나타났거나 혹은 결실되어 나타난 부분일 가능성이 높다. 다시 말해, 레퍼런스 시퀀스 상에 이렇게 값의 변화를 보이는 커버리지 영역은 CNV를 포함하는 유전적 구조 변이를 나타내는 영역을 나타낼 가능성이 높다고 판단할 수 있다.

본 연구에서는 커버리지 데이터 분석에 의한 CNV 탐지 방식으로서 다음과 같은 모양 기반 모델을 사용한다. 커버리지 데이터의 값의 변화를 지속적으로 분석하여 커버지리 값이 평균 커버리지 값보다 낮게 나타나거나 높게 나타나는 영역이 일정 수준을 유지하며, 또한 일정 크기 이상 지속되는 경우, 이를 CNV 영역으로 추정하는 것이다. 예를 들어 이와 같은 특성을 나타내는 영역의 커버리지 값이 평균 커버리지 값과 비교하여 1/2 정도 적은 차이를 보이면 한 카피 결실(loss) 영역을 나타낸다고 추정할 수 있으며, 반대로 특정 영역에서 평균 커버리지 값보다 더 많은 3/2배, 2배, 3배 이상 나타나면 모두 추가적인 반복(gain) 영역으로 볼 수 있다[12].

이와 같이 CNV를 추출하고자 할 때 그 모양에 따라 다양한 변수가 발생할 수 있는데, 본 연구에서는 CNV 후보 영역 추출 단계에서 모양 변화에 대한 내용을 입력 변수를 조절하도록 하여 원하는 모양의 영역을 추출하도록 한다.

3.2 후처리 과정

일반적으로 기가 시퀀싱 데이터의 리드와 레퍼런스 시퀀스에는 많은 에러 영역들이 포함된다. 따라서 제 3.1절에서 보인 모양 기반 방식을 이용하여 CNV 후보 영역을 추출하면 그 결과에는 다수의 거짓 실재 영역이 포함될 수 있으며, 또한 하나의 후보 영역이 여러 개의 작은 영역으로 나뉘어 추출될 가능성이 높다. 본 연구에서는 이와 같은 문제점을 해결하기 위하여 다음과 같은 후처리 과정을 수행한다.

서로 다른 n개의 개인 유전체 시퀀스의 커버리지 분석에 의하여 동일 영역에서 같은 반복 CNV가 발견되면 (혹은 결실 CNV가 발견되면), 이 영역은 레퍼런스 영역의 에러 영역으로 추정할 수 있다. 본 연구에서는 다수의 테스트 시퀀스에 대하여 CNV 후보 영역 추출 알고리즘을 적용하고, 얻어진 결과를 상호 비교하여 이와 같은 레퍼런스 에러 영역에 대한 보정 작업을 수행하였다. 다음의 그림 1은 1000 게놈 프로젝트 사이트에서 다운 받은 NA10851과 NA18507의 두 유전체 데이터에 대한 커버리지 분포를 비교한 예를 보인다. 그림에서 보이는 바와 같이 서로 다른 유전체의 동일(혹은 유사) 영역에서 두 커버리지의 분포가 매우 비슷한 양상을 보이는 CNV가 추출된 경우, 우리는 이 영역이 레퍼런스 시퀀스의 에러 영역일 가능성이 높다고 판단하여 추출된 CNV 후보 영역을 결과에서 제외시킨다.

또한 후처리 과정에서는 레퍼런스 시퀀스의 특정 영역들에 관한 생물학적 특성 정보를 이용하여 후보 CNV의 정제에 사용한다. 예를 들어 GC-content가 높은 영역에서는 일반적으로 커버리지 값이 매우 높아질 가능성이 높으며, 따라서 제안된 알고리즘에서는 후보 CNV 영역으로 추출될 가능성이 높다. 분자 생물학(molecular biology)에서 GC-content는 DNA 분자에 포함되어 있는 질소의 함유량을 나타내며, DNA, RNA의 특정 부분에서 나타난다.

시퀀스에서 GC-content가 높은 영역에 대한 정보를 추출 가능하며, 이를 이용하여 후보 CNV 영역의 정제에 사용한다. 이와 같은 정제 과정을 거친 CNV 영역들은 후처리 과정에서 마지막으로 근접 위치를 고려하여

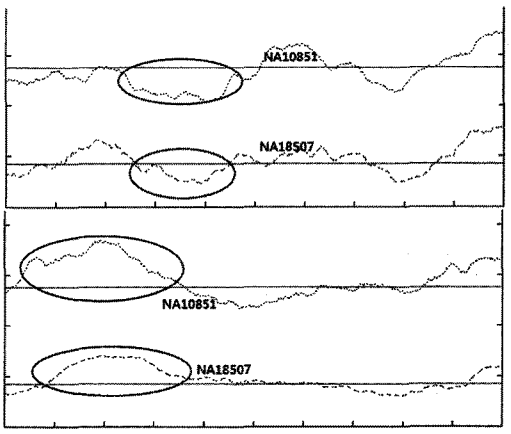

그림 1 커버리지 비교 분석에 의한 레퍼런스 에러 영역 추출의 예

CNVR로 클러스터링 되며, 이들 결과가 제안된 방식에 의하여 최종적으로 얻어지는 CNVR의 집합이 된다.

3.3 CNVR 검색 알고리즘

[Algorithm 1]은 입력으로 리드를 레퍼런스 시퀀스에 매핑하여 얻어진 결과 RS, 탐지할 CNV 영역의 길이 LEN, 변화 폭 C, 이동 변환 평균을 구하기 위한 윈도우 사이즈 WS와 쉬프트 사이즈 SS를 입력으로 받아 CNVR을 찾아낸다. [Algorithm 1]의 동작 과정을 단계별로 설명하면 다음과 같다.

우선 리드의 매핑 결과 RS로부터 각 리드의 매핑 위치 정보를 추출하여 전체 레퍼런스 위치에 대한 커버리지 정보를 얻는다(line 1-2). 다음, 커버리지 데이터의 평균값을 계산하여, 이를 평균 커버리지 값으로 사용한다(line 3). Lline 4의 함수 MovingAverage()는 CovTbl의 각 커버리지 값에 대하여 SS의 간격으로 이동 평균 변환 계수 WS의 이동 평균 변환을 수행하여 그 결과를 새로운 테이블 MA_CovTbl에 저장하는 함수를 나타낸다. 이동 평균 변환은 연속되는 WS개의 요소 값들의 평균값들을 SS의 간격으로 나열하는 변환이다. 이동 평균 변환을 통하여 커버리지 시퀀스 내에서 나타나는 잡음의 영향을 제거할 수 있으며, 이동 평균 변환 계수 WS는 해당 응용에서 잡음의 영향을 줄이고자 하는 정도에 따라 선택된다. 이렇게 만들어진 이동 평균 변환된 커버리지 시퀀스로부터 CNV 후보 영역을 탐지한다(line 5-6). 평균 커버리지 값보다 ±C만큼의 차이를 보이는 영역이 LEN 이상의 크기를 가지는 경우, 이를 후보 영역 CNVcand에 저장한다. 이때, 커버리지보다 +C만큼 차이나는 영역은 반복인 영역으로, 평균 커버리지보다 -C만큼 차이나는 영역은 결실인 영역으로 구분된다. Line 7은 후처리 과정을 나타내며, 추출된 후보 영역에 대하여 레퍼런스 특성을 고려한 정제 과정을 거친 후, 마지막으로 결과 CNV를 클러스터링 하여 최종 CNVR을 반환하는 과정을 나타낸다.

Algorithm 1: Detect_CNVR

Input : read alignment RS,
length of the region LEN, coverage variation C,
window size WS, shift size SS
Output : set of Copy Number Variation Regions CNVR

1. PDS := ExtractPosition(RS)
2. CovTbl := CalculateCoverage(PDS)
3. AVG := CalculateAverage(CovTbl)
4. MA_CovTbl := MovingAverage(CovTbl, WS, SS)
5. **for each** i-th coverage data of the MA_CovTbl **do**
6. CNVcand := Find_Region(AVG, C, LEN)
7. CNVR := Post-processing(CNVcand)
8. **return** CNVR

4. 성능 평가

4.1 실험 방법

본 실험에서 사용된 데이터는 레퍼런스 시퀀스로 NCBI Build 36.3을 사용하였고, 테스트 데이터는 1000 게놈 프로젝트 사이트에서 다운 받은 다수 개인의 리드 시퀀스 집합(NA10851, NA18507 등)을 사용하였다. 본 실험에서는 Illumina GA(Genome Analyzer) I/II 머신을 통해 생성된 리드 데이터를 사용하였으며, 리드 매핑 프로그램으로서 SOAP(Short Oligonucleotide Alignment Program)[13]을 사용하였다.

제안된 방식의 유용성을 보이기 위하여 추출된 CNV 영역을 비교, 검증하였다. 본 연구에서는 제안된 방식에서 추출된 CNV 영역을 비교, 검증하기 위해 DGV의 CNV 정보를 사용하였다. 현재(2009년 9월 기준) DGV에 등록되어 있는 전체 엔트리는 49,944개이고, 이 중 CNV는 29,133개, Inversion은 914개, InDel는 19,941개가 보고되어 있다. 이 데이터베이스는 같은 목적의 연구에서 산출되는 새로운 연구 결과에 의해 지속적으로 업데이트되고 있다.

본 실험에서는 HLA 영역에 대하여 CNVR 검색을 수행하였다. HLA 영역은 염색체 6번의 단완(short arm)에 위치하는 영역으로서, 길이는 약 3.408Mbp이다. DGV에 보고된 CNV 중 HLA 영역에 해당하는 CNV는 총 236개이다. 그러나 이 들 영역은 상당 부분 겹치는 부분이 많아 전체 CNVR은 15개에 이른다.

4.2 실험 결과 및 분석

실험 1에서는 제안된 방식이 HLA 영역에 존재하는 반복되거나 결실되는 다양한 모양의 CNV를 효율적으로 검출 가능한지를 검증한다. 검증을 위하여 CNV의 모양 추출을 위한 기본적인 입력 변수 값으로 다음 값들을 설정하였다. 커버리지의 변화폭을 체크하기 위한 입력 변수 C의 값은 평균 커버리지 값보다 ±25% 차이가 나는 값으로 설정하였고, 영역의 길이를 나타내는 입력 변수 LEN의 값은 1,000(1Kbp)로 하여, 1Kbp 이상 되는 영역을 모두 추출하였다. 또한 이동 평균 변환을 위한 이동 평균 변환 계수는 1,000으로 쉬프트 사이즈는 1로 설정하였다.

추출된 CNV에 대하여 그 특성을 비교, 검증하였으며, 이 결과로부터 제안된 방식은 CNV의 모양을 나타내는 입력 변수를 조절하여 다양한 모양을 갖는 CNV를 비교적 간단히 추출하여 낼 수 있음을 입증하였다. 다음 그림 2는 본 실험에 의하여 추출된 CNV의 예를 보인다. 이 그림은 NA10851 유전체 데이터로부터 추출된 CNV의 예로서, 이 예로부터 다양한 높이와 크기를 가지는 CNV 영역이 효율적으로 추출되고 있음을 알 수 있다.

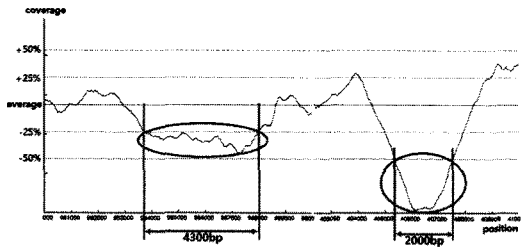

그림 2 추출된 CNV의 예(NA10851 유전체에 존재하는 결실 영역의 예)

실험 2에서는 제안된 방식에 의하여 HLA 영역에 존재하는 CNVR이 정확히 추출되었는가를 검증하였다. NA10851(평균 커버리지 수: 5.8x)과 NA18507(평균 커버리지 수: 1.8x)의 개인 유전체 데이터를 사용한 실험을 수행하였으며, HLA 영역에 존재하는 모든 CNVR을 추출하였다.

추출된 결과의 정확도를 검증하기 위하여 DGV에 등록된 CNVR 영역과 비교하였다. DGV에는 여러 개인 유전체에서 발견되는 모든 CNV 영역을 포함하는 CNVR이 등록되어 있다고 할 수 있다. 다음의 그림 3은 DGV에 등록된 CNVR과 본 실험에서 탐지된 CNVR을 비교하여 나타낸 것이다. 그림 3에서 녹색에 해당하는 영역은 DGV에 보고되어 있는 CNVR이고, 적색은 NA10851에서 탐지된 CNVR이며, 청색은 NA18507에서 탐지된 CNVR이다. DGV에 보고되어 있는 15개의 CNVR과 본 실험을 통해 탐지된 CNVR을 비교하면, NA10851은 7개의 영역이 겹치는 결과를 보이며, 그 중 50% 이상 중첩되는 영역의 개수는 5개에 해당한다. 또한 NA18507의 경우, 9개의 영역이 겹치는 결과를 보이며, 그 중 50% 이상 중첩되는 영역의 개수는 6개에 해당한다. 비교 결과로부터 추출된 CNVR이 유의함을 알 수 있으며, 제안된 방식은 NA10851 혹은 NA18507의 HLA 영역에 존재하는 CNVR을 적절히 추출하고 있음을 알 수 있다.

5. 결론 및 향후 연구과제

본 논문에서는 레퍼런스 시퀀스에 기가 시퀀싱 데이터를 매핑하여 얻어지는 커버리지 데이터를 이용한 모양 기반의 CNVR 추출 방식을 제안하였다. 실 데이터를 이용한 성능 비교 실험을 수행하였으며, 실험 결과에 의하여 제안된 방식은 HLA 영역에 존재하는 반복되거나 결실되는 다양한 모양의 CNVR을 효율적으로 검출하는 것을 입증하였다. 그러나 본 알고리즘에 의하여 추출되고 있는 후보 CNV 영역에는 아직 정제 되어야 할 거짓 실재 영역이 존재한다. 현재 후처리 과정에 관한 지속적인 연구를 수행하고 있다. 금후, 실유전체 데이터의 염색체별 특성을 고려하여, 후처리 방식의 검증 및 보완에 관한 연구를 수행할 예정이다.

참 고 문 헌

[1] Redon et al., "Global variation in copy number in the human genome," *Nature*, vol.444, pp.444-454, 2006.
[2] Smith et al., "Rapid whole-genome mutational profiling using next-generation sequencing technologies," *Genome Research*, vol.18, no.10, pp.1638-1642, 2008.
[3] http://projects.tcag.ca/variation/
[4] http://www.1000genomes.org/
[5] http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/
[6] 서울주, "Copy number variants (CNV)의 분석 방법," Korean Society of Medical Biochemistry and Molecular Biology, vol.15, no.3, pp.28-39, 2008.
[7] 홍상균, 홍동완, 윤지희, 김종일, "짧은 리드의 서열 정렬에 의한 CNV 영역 추출", *데이터베이스연구*, vol.24, no.3, pp.1-13, 2008.
[8] Lai et al., "Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data," *Bioinformatics*, vol.21, no.19, pp.3763-3770, 2005.
[9] Scherer et al., "Challenges and standards in integrating surveys of structural variation," *Nature Genetics*, vol.39, no.7, pp.S7-S15, 2007.
[10] Chiang et al., "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nature Methods*, vol.6, no.1, pp.99-103, 2009.
[11] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BioMed Central Bioinformatics*, vol.10, no.1, 2009.
[12] 박종화, "Bioinformatics Tools for Variome Study," *Medical Postgraduates*, vol.37, no.3, pp.131-133, 2009.
[13] Li et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol.25, no.15, pp.1966-1967, 2009.

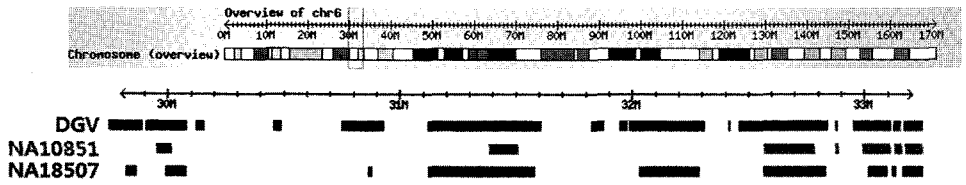

그림 3 NA10851과 NA18507로부터 추출된 CNVR과 DGV에 등록된 CNVR의 비교 검증