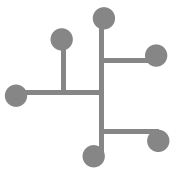


차세대 시퀀싱으로 생성된 페어드 엔드 리드를 이용한 CNV 발견 기법

A Novel Approach to Detect CNVs with Paired-end Reads From Next-generation Sequencing

문명진(Myungjin Moon)[1], 박치현(Chihyun Park)[1], 윤영미(Youngmi Yoon)[1,2], 안재균(Jaegyoon Ahn)[1], 박상현(Sanghyun Park)[1]

요 약

유전체 단위 반복 변이(Copy Number Variation, 이하 CNV)는 유전체의 구조적 변이 중 하나로, 하나의 개체에서 1Kbps(Kilo base pairs) 이상의 염기 서열의 반복 횟수가 다른 개체에 비해 많거나 적은 것을 뜻한다. CNV는 갖가지 질병과 개체간의 특징 발현에 직접적으로 영향을 미치는 것으로 알려져 있기에 CNV를 발견하는 것은 유전자 연구 분야에서 매우 중요하다. 기존에는 CNV를 찾기 위해 마이크로어레이(microarray)를 이용한 기법이 주로 사용되었는데, 이는 마이크로어레이가 갖는 해상도의 제약과 노이즈로 인하여 길이가 긴 CNV를 찾기 어려우며, 오류가 많다는 문제점이 있다. 본 논문에서는 차세대 시퀀싱(next-generation sequencing) 기법으로 생성된 짧은 페어드 엔드 리드(Paired-end read)들을 이미 밝혀진 염기 서열에 매핑(mapping하고 분석하여 CNV를 찾아내는 새로운 기법을 제안한다. 본 논문이 제안하는 기법은 염기 서열의 베이스 페어(base pair) 단위로 CNV 여부를 판단하므로 마이크로어레이 기법으로 찾기 힘든 짧은 길이의 CNV까지 찾아낼 수 있다. 실험은 가상 데이터와 실제 데이터를 가지고 수행하였으며, 가상 데이터를 대상으로 한 오류율은 기존의 기법에 비해 낮아졌음을 보였다. 또한, 추가적으로 짧은 리드를 BLAST 대신 Bowtie를 통해 정렬함으로써 속도의 향상을 가져왔다.

주제어: 유전체 단위 반복 변이, CNV, 차세대 시퀀싱, 고출력 시퀀싱, 페어드 엔드

*본 논문은 교육과학기술부 한국연구재단의 미래기반기술개발사업(2009-0083311)의 지원을 받아 수행되었음.

1 연세대학교 컴퓨터과학과

2 가천의과학대학교 IT학과

+ 논문접수 : 2009년 8월 24일, 심사완료 : 2009년 10월 12일

Abstract

Copy Number Variation(CNV) is one of the genomic variants, which is caused by either amplification or deletions of DNA segments whose size is greater than 1Kbps. As it is known that CNVs account for a significant proportion of phenotypic variation, including disease susceptibility, identifying and cataloging of CNVs are essential for the genetic analysis of human genome variation. CNVs detected by microarray based approaches are limited to medium or large sized ones because of low resolution and noise of microarray. Here we propose a novel approach to detect CNVs by mapping the short paired-end reads obtained by next-generation sequencing to the previously assembled human genome sequence and analyzing them. This method demonstrates the feasibility of detecting CNVs which include short ones that microarray based algorithms cannot detect, as this method decide whether a region is a CNV or not based on the score of each base pair. The experiment was performed with both synthetic and real data. In the experiment with the synthetic data, false positive and false negative rates of the results were relatively lower than existing method. In addition, application of Bowtie for read-mapping improved speed compared to the existing method, which mapped reads with BLAST.

Keyword: Copy Number Variations, Next-generation sequencing, High-throughput sequencing, Paired-end

1. 서론

유전체 단위 반복 변이(Copy Number Variation, 이하 CNV)는 유전체의 구조적 변이 중 하나이다. 기존의 연구 결과에 의하면 인간 유전체에 존재하는 다형성(polymorphism)은 단일 염기 변이(SNP) 또는 연쇄 반복 서열(VNTR)이나 부수체(microsatellite)와 같은 작은 크기의 반복 변이가 대부분을 차지한다고 여겨졌다. 그러나 최근의 실험 결과 인간의 유전체에는 큰 크기의 유전체 변이가 기존에 생각되던 것보다 훨씬 많이 존재하며, 지금까지 쉽게 설명되지 못했던 여러 가지 현상들이 이와 관련이 있을 것으로 밝혀졌다. 이러한 영역이 CNV에 해당된다[1, 2]. CNV는 유전병을 비롯한 갖가지 질병 뿐 아니라, 개체간의 특징 발현에 직접적으로 영향을 미치는 것으로 알려져 최근 유전체 연구 분야에서 관심의 대상이 되고 있다[3, 4]. 실제로 많은 연구자들이 여러 질병 감수성과 CNV 관련성에 대한 연구를 진행하고 있고, 민족별 녹말 섭취량에 따른 녹말 분해 효소의 양이 CNV와 관련이 있다거나[5], 에이즈 감수성, 췌장염, 크론병, 전신성홍반성낭창(SLE), 자폐증 등의 질병이 CNV와 관련 있을 가능성이 높다고 보고된 바 있다[6].

CNV는 하나의 개체에서 1Kbps(Kilo base pairs) 이상의 염기 서열의 반복 횟수가 다른 개체에 비해 많거나 적은 것으로 정의된다[1, 2]. 한 개체에서의 염기 서열의 반복 횟수가 비교 대상인 다른 개체에 비해 많을 경우는 증가(gain), 적을 경우 감소(loss), 동일할 경우 중립(neutral)이라 부르며, CNV는 상대적인 개념이기에 한 개체에서의 증가는 비교 대상인 다른 개체에서의 감소가 된다. 염기 서열의 반복 횟수를 복제 수(copy)라 하는데, 예를 들어 복제 수 2는 한 개체에서의 반복되는 염기 서열이 다른 개체에 비해 2배인 경우를 뜻한다.

CNV를 찾기 위해 기존에 가장 많이 사용된 기법은 마이크로어레이(microarray)를 이용한 방식으로, 이는 WGTP(Whole Genome Tiling Path) 어레이 상의 컨트롤 DNA와 테스트 DNA의 경쟁적 교잡(hybridization)을 통한 발현 정도의 차이를 이용하는 기법이다. 두 개체의 염기 서열을 전부 비교하여 CNV를 밝혀낸 연구도 있다[7, 8, 9]. 염기 서열 비교를 통한 CNV 발견 기법은 CNV 영역을 정확하게 찾아낼 수 있으나, 비교하는 두 개체의 전체 염기 서열을 알아야 하므로 비용적인 측면이 가장 큰 문제가 된다. 이러한 비용 문제를 해결하기 위해 시퀀싱(sequencing)을 통해 생성된 리드(read)들을 이용한 연구가 진행되었다.

기존에 가장 많이 사용되던 시퀀싱 기법은 프레드릭 생어(Fredrick Sanger)가 제안한 생어 시퀀싱(Sanger sequencing)이다. 생어 시퀀싱은 DNA에 ddNTP(dideoxynucleotide triphosphate) 효소를 붙여 특정 염기로 끝나는 여러 길이의 염기 조각을 얻고, 이를 전기 영동하여 길이 순으로 분리하여 DNA의 염기 서열을 결정하는 방법이다. 생어 시퀀싱은 약 1Kbps 길이의 리드를 낮은 오류율로 생성할 수 있으나 1Kbps당 $0.5라는 높은 비용이 소요되어 많은 양의 리드를 생산할 수 없다는 문제점이 있다.

이러한 문제점을 해결하기 위해 차세대 시퀀싱이(next-generation sequencing) 등장하게 되었다. 차세대 시퀀싱은 낮은 비용으로 짧은 길이의 리드들을 대량으로 생성하는 기법으로, 고출력 시퀀싱(high-throughput sequencing) 혹은 기가 시퀀싱(giga sequencing)으로도 불린다. 차세대 시퀀싱의 대표적인 플랫폼(platform)으로는 454, Illumina, SOLiD, Polonator, HeliScope 등이 있으며, 리드의 생성 방법, 생성되는 리드의 길이, 양, 소요 비용은 이들에 따라 조금씩 다르나 짧은 길이의 리드를 대량으로 생성한다는 특징은 전부 동일하다. 〈표 1〉은 2008년의 각 차세대 시퀀싱 기술의 현황을 보여준다.

〈표 1〉 차세대 시퀀싱 기술의 현황[10]

Platform	Feature generation	Sequencing by synthesis	Cost per megabase	Cost per instrument	Paired end	Read length
Roche 454	Emulsion PCR	Polymerase (pyrosequencing)	~$60	$500,000	Yes	250 bps
Illumina(Solexa)	Bridge PCR	Polymerase (reversible terminators)	~$2	$430,000	Yes	75 bps
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)	~$2	$591,000	Yes	50 bps
Polonator	Emulsion PCR	Ligase (nonamers)	~$1	$155,000	Yes	13 bps
HeliScope	Single molecule	Polymerase (asynchronous extensions)	~$1	$1,350,000	Yes	45 bps

차세대 시퀀싱으로 생성되는 리드의 길이는 수십에서 수백 bps로 생어 기법으로 생성된 리드에 비해 짧다는 문제점이 있으며, 이에 따라 서열의 조립이나 매핑(mapping)의 정확도가 떨어지게 된다. 그러나 기존과는 비교할 수 없을 정도의 많은 리드를 적은 비용으로 얻을 수 있다는 것은 매우 큰 장점이기 때문에, 차세대 시퀀싱으로 얻은 데이터를 이용한 많은 연구가 이루어지고 있다. 차세대 시퀀싱이 활용되는 대표적인 연구 분야로는 디 노보 어셈블리(de novo assembly), 리드 매핑(read mapping), 유전 변이 발견, 브라우징(browsing) 등이 있다. 또한 차세대 시퀀싱을 통하여 생성되는 리드들의 길이를 늘이기 위한 연구도 계속 진행되고 있다.

본 논문은 차세대 시퀀싱 방식으로 생성된 짧은 리드들을 이용하여 CNV를 찾는 새로운 기법을 제안한다. 짧은 리드를 이용할 경우 기존에 많이 사용된 마이크로어레이를 사용한 기법으로는 찾기 어려운 짧은 길이의 CNV까지 찾아낼 수 있고, 데이터 자체의 오류율이 낮아 보다 정확한 값을 도출할 수 있다. 본 논문의 내용은 2.3절에서 소개할 짧은 리드를 이용한 CNV 발견 기법을 확장한 것으로, 차세대 시퀀싱 방식으로 생성된 페어드 엔드 리드(paired-end reads)를 활용한 첫 연구이다. 또한 페어드 엔드 리드의 사용과 매개 변수의 수정을 통하여 짧은 리드를 이용했던 기존의 기법에 비해 잘못된 긍정(false positive)과 잘못된 부정(false negative)을 줄이고, 매개 변수에 대한 민감도를 낮추었다. 또한 본 기법은 증가된 CNV만을 찾을 수 있던 기존의 방식과는 달리 감소된 CNV까지 찾아낼 수 있다는 장점이 있다. 추가적으로 프로세스를 수정하고, 기존에 리드 매핑에 사용한 BLAST[11] 대신 짧은 리드에 최적화된 Bowtie[12]를 활용함으로써 처리 속도를 크게 향상시켰다. 이러한 장점들을 통해 컨티그(contig) 하나와 같은 짧은 염기 서열뿐 아니라 염색체 단위 이상의 긴 염기 서열에도 본 기법을 적용할 수 있게 되었으며, 기존의 연구에 비해 많은 CNV 영역을 보다 정확하게 찾을 수 있게 되었다.

2. 관련 연구

2.1. Array-CGH를 이용한 CNV 발견 기법

마이크로어레이 기술을 기반으로 유전체 간의 양적 관계를 비교함으로써 CNV를 찾는 방법은 기존의 많은

연구들에서 수행되어 왔으며, 대부분의 연구는 종양 세포에 대한 유전체 변이를 찾음으로써 유전 질환과의 연관 관계를 밝히는 것에 초점을 두고 있다[13]. 하지만 최근 들어 종양 세포뿐만 아니라 표현형으로 유전 질환이 없는 정상인에 대한 CNV를 찾고자 하는 연구가 활발히 진행되고 있다. Array-CGH 기술을 기반으로 정상인에 대한 CNV를 찾은 가장 대표적인 연구로는 인종별로 나누어진 총 270명의 Hapmap sample[14]을 대상으로 실험하여 공통적으로 많이 발견되는 CNV를 밝힌 연구가 있다[1]. Array-CGH 기반 연구에서 CNV를 찾는 알고리즘적 연구 또한 다양하게 이루어지고 있는데, 현재까지 수행된 많은 연구에서 가장 대표적인 방법으로 CBS[15]가 있다. CBS는 전체 데이터 집합에서 CNV로 판단될 수 있는 가능성이 높은 부분들을 추출해내는 방식으로, 비록 시간 복잡도가 높고 종양과 크기가 작은 데이터 셋에서만 좋은 성능을 보인다는 단점은 있으나, 높은 정확도를 보이는 방법이기에 가장 대표적으로 참고가 되고 있다. CBS와 비슷하게 대부분의 논문에서는 인간의 전체 염색체를 대상으로 CNV를 찾는 연구를 수행하지 않고, 비교적 CNV 영역이 다른 영역과 쉽게 구분될 수 있는 종양 혹은 유전 질환을 가지고 있는 샘플을 대상으로 실험을 수행하였다[16]. 하지만 최근 들어 HMM(Hidden Markov Model)을 사용한 [17]이나 [18]과 같은 연구는 Hapmap 샘플 같은 유전 질환과 관계없는 대상을 바탕으로 CNV를 찾음으로써 향후 CNV 연구가 정상 표현형을 갖는 대상들이 가지고 있는 CNV를 포함한 구조적 변이가 어떠한 진화 경로를 통해서 질병과 연관될 수 있는지를 밝히는 데 초점을 두고 있다. 하지만 아직 마이크로어레이 자체가 가지고 있는 노이즈 문제 등 해결해야 할 부분이 많다[19]. 또한 기존 마이크로어레이 기반 실험은 대부분 저해상도였기 때문에 많은 연구들에서 찾은 CNV가 100Kbps 이상의 긴 CNV였으나, 실제 인간에 존재하는 많은 CNV는 100Kbps 이하의 작은 변이라고 밝혀졌기 때문에 이에 대한 연구가 필요하다[20]. 최근 생어 연구소에서 공개한 고해상도 데이터는 이러한 문제점을 실험적으로 해결할 수 있는 기반을 제공하였다. 하지만 대규모 데이터에서 CNV를 빠르고 정확하게 찾는 알고리즘적인 연구는 실험 플랫폼의 변화에 비하여 발전되고 있지 않다. 따라서 앞으로 고해상도 마이크로어레이 실험 데이터를 이용하여 CNV를 찾는 연구가 활발히 진행될 것으로 예상된다.

2.2. 염기 서열 비교를 통한 CNV 발견 기법

염기 서열 비교 방식은 기존에 구조가 전부 밝혀진 염기 서열들을 상호 비교하여 CNV 등의 구조적 변이를 찾아내는 기법이다[21]. 대표적 연구 결과로서 인간 포스미드 시퀀스와 레퍼런스 시퀀스(build 35)를 비교하여 build 35 시퀀스 상에 존재하는 8 Kbps 이상의 크기를 갖는 297개의 구조적 변이 후보 영역을 추출한 예가 있으며[6], Celera genomics의 어셈블리 시퀀스(assembly sequence) R27c과 레퍼런스 시퀀스(reference sequence) build 35를 비교하여 419개의 CNV 영역을 포함하는 총 13,534개의 구조적 변이 후보 영역을 추출한 예가 있다[7]. 이 기법은 서로 다른 두 시퀀스 사이에 존재하는 차이점을 발견하기 위하여 BLAST 등의 소프트웨어를 사용하여 두 시퀀스 사이의 서열 비교를 수행하기 때문에 CNV 영역을 정확하게 밝힐 수 있다. 그러나 이 방법은 많은 시간과 비용을 필요로 하며, 현재 구조가 밝혀진 염기 서열이 많지 않기에 실험할 수 있는 대상이 한정된다는 단점이 있다.

2.3. 짧은 리드를 이용한 CNV 발견 기법

본 연구의 기반이 된 연구로서, 가상 유전 변이 생성

기와 차세대 시퀀싱으로 얻어진 리드를 이용하여 CNV를 발견한 바 있다[22]. 이 연구에서는 유전 변이 데이터베이스(Database of Genomic Variants)[23]와 dbSNP[24]에서 얻어진 다양한 유전 변이에 대한 데이터를 바탕으로 가상 유전 변이 생성기를 제작하였다. 가상 유전 변이 생성기는 입력으로 임의의 염기 서열을 넣으면 이를 바탕으로 SNP, InDel, CNV 등을 넣은 새로운 염기 서열을 출력하게 된다. 여기서 입력 염기 서열을 레퍼런스 시퀀스(reference sequence), 출력 염기 서열을 테스트 시퀀스(test sequence)라고 하며, 테스트 시퀀스를 차세대 시퀀싱 방식으로 리드로 만들어 레퍼런스 시퀀스에 매핑하고, 그 매핑 횟수를 계산하여 CNV 여부를 판단하게 된다. 2.2절의 연구와 같이 테스트 시퀀스 전체의 염기 서열을 알아내기 위해서는 많은 비용이 소요되므로, 테스트 시퀀스를 바탕으로 만든 짧은 리드들을 다수 이용하여 시퀀스 전체를 비교하는 것과 유사한 효과를 내고자 하는 기법이다.

그러나 생성된 리드들이 테스트 시퀀스의 영역에 고르게 매핑되지 못하므로, 각 영역에서 매핑된 횟수만을 가지고 CNV임을 판단하기는 어렵다. 따라서 레퍼런스 시퀀스를 CNV 최소의 길이인 1Kbps씩 묶고, 그 묶인 그룹에 매핑된 횟수의 합을 이용하여 CNV 여부를 판단한다. 이는 주위의 값까지 고려하여 오차를 줄이기 위함이다. 그러나 1Kbps 정도의 긴 그룹은 짧은 반복 영역이 존재함으로 인해 그룹 내의 매핑 횟수의 합이 높아질 수 있으므로, 매핑 횟수의 평균값 이상인 영역이 그룹의 절반 이상이 되도록 한다. 이러한 매핑 횟수의 분포는 가우시안 분포(Gaussian distribution)에 가까운 모습을 보이므로, 평균과 분산을 이용하여 $\mu + k\sigma$ 이상의 영역과 $\mu - k\sigma$ 이하의 영역을 CNV로 규정한다. k의 값은 반복적 실험에 의해 설정되었다. 이 실험의 결과는 Build 36.3의 19번 염색체의 NT_011255.14 콘티그와, 이를 입력으로 한 가상 유전 변이 생성기의 출력 간의 비교를 통하여 확인하였다.

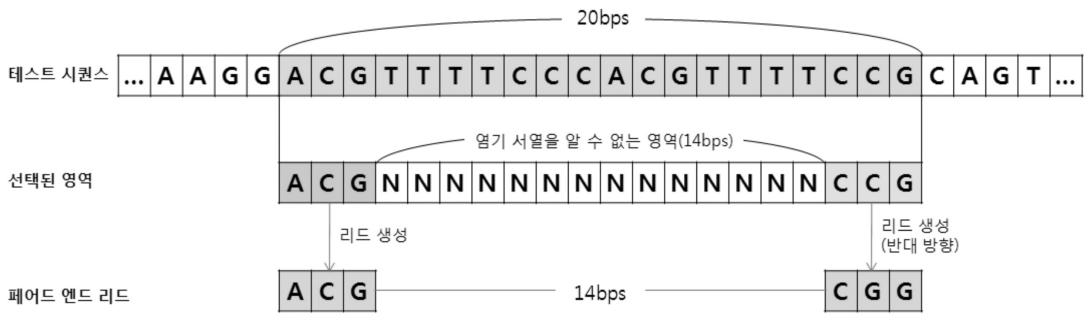
이러한 방식은 마이크로어레이 기반 기법에 비해 짧은 CNV 영역까지 찾아낼 수 있으며, 비교적 낮은 오류율을 보인다. 그러나 이 연구에서는 짧은 CNV 중 증가 영역만 찾을 수 있고, 속도가 느려 큰 염기 서열에는 사용하기 어려우며, 매개 변수, 특히 k의 값에 따라 찾을 수 있는 CNV 영역이 크게 변한다는 단점이 있다.

3. 제안하는 기법

3.1. 차세대 시퀀싱 시뮬레이터

본 연구에서는 페어드 엔드 리드(paired-end reads)를 지원하는 75bps 길이의 리드를 생성하는 Illumina의 차세대 시퀀싱을 소프트웨어적으로 수행하였다. 페어드 엔드 리드란 서로의 대략적인 거리를 알고 있는 한 쌍의 리드를 의미한다. 예를 들어 Illumina 시퀀싱은 약 500bps의 리드를 생성하고, 이 리드의 양쪽 끝 75bps의 염기 서열을 읽어낸다. 이때 두 리드를 읽는 방향은 3'와 5'로 각각 반대가 된다. 두 리드 사이의 350bps 구간에 대한 염기 서열은 알아낼 수 없으나, 이를 통해 양쪽 75bps의 염기 서열 둘이 약 350bps 떨어져 있다는 정보를 얻어낼 수 있다. 이렇게 얻어진 거리 정보를 가진 75bps 리드 한 쌍을 페어드 엔드 리드라 한다. 페어드 엔드의 거리는 시퀀싱 방식에 따라 다르지만, 하나의 시퀀싱 방식으로 생성되는 페어드 엔드의 길이는 거의 일정하다.

짧은 리드를 이용한 CNV 발견 기법에서 이러한 거리 정보를 이용하여 얻을 수 있는 이점은 다양하다. 우선 매핑 시 가장 문제가 되는 반복 서열(Repeats) 문제를 해결할 수 있다. 비슷하거나 동일한 염기 서열이 여러 개 있을 경우, 시퀀싱으로 얻어낸 리드가 그 중 어느 곳



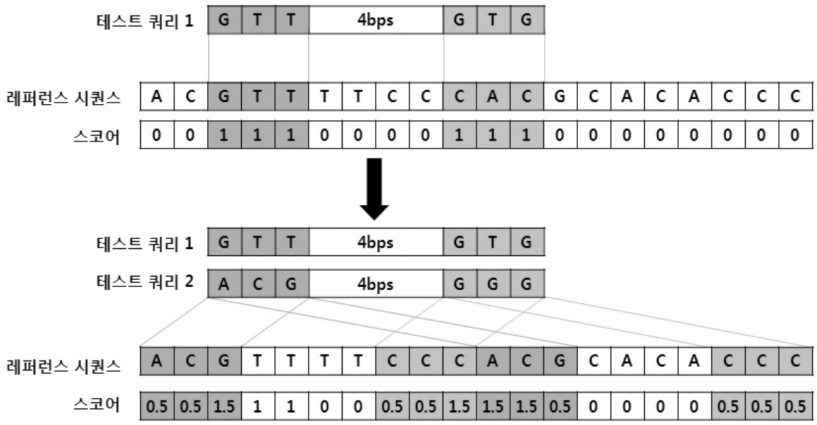
〈그림 1〉 차세대 시퀀싱 방식을 시뮬레이터를 이용한 페어드 엔드 리드의 생성 예

에서 생성된 것인지 알기가 어렵다. 그러나 페어드 엔드의 정보를 이용하면 가장 유사한 염기 서열을 찾아낼 수 있다. 또한 페어드 엔드 사이의 거리가 생성 시의 거리보다 멀 경우 짧은 삽입(Insertion)이 있으며, 반대로 가까울 경우 짧은 삭제(Deletion)가 있다는 것을 알아낼 수 있다. 또한 페어드 엔드의 방향 정보를 통해 역위(Inversion) 여부도 파악할 수 있다. 즉 CNV 발견을 방해하는 요소들을 배제할 수 있게 되므로 페어드 엔드 리드의 활용은 필수적이라 할 수 있다.

본 논문에서는 이미 구조가 알려진 염기 서열을 이용하여 리드들을 생성하였다. 리드 생성에 사용된 염기 서열을 테스트 시퀀스(test sequence)라 한다. Illumina에서는 약 500bps의 길이의 페어드 엔드 리드를 제공한다. 페어드 엔드의 길이는 조금씩 달라질 수 있으므로 본 실험에서는 오차율을 100bps로 설정하였다. 전체 염기 서열에서 임의의 영역을 선택한 후 그곳에서부터 연속된 400~600bps 영역을 추출하고, 추출된 영역에서 양 끝의 75bps를 잘라내어 페어드 엔드 리드를 생성한다. 이러한 과정으로 생성된 리드 길이의 합이 전체 염기 서열의 길이의 n배가 될 때까지 반복한다. 이러한 n을 커버리지(coverage)라 하고, 생성된 리드들을 테스트 쿼리(test query)라 한다. 커버리지가 높아질수록 보다 정확한 정보를 얻을 수 있으나 비용이 많이 소요된다는 단점이 있다. 〈그림 1〉은 리드의 길이 3bps, 리드 사이의 거리 14bps인 페어드 엔드 리드의 생성 예를 보여준다.

3.2. Bowtie를 이용한 리드 매핑 및 스코어링

3.1절에서 생성된 테스트 쿼리들을 Bowtie를 이용하여 이미 밝혀진 또다른 염기 서열에 매핑한다. 매핑의 대상이 되는 이 염기 서열을 레퍼런스 시퀀스(reference sequence)라 한다. 레퍼런스 시퀀스에 비해 테스트 시퀀스에 1Kbps 이상의 연속된 염기 서열이 더 많거나 적음을 판별하는 것이 최종 목적이 된다. Bowtie는 1024bps 이하의 짧은 리드들을 긴 염기 서열에 매핑하는 툴로, BLAST에 비해 범용성은 떨어지나 짧은 리드의 매핑 시 속도 및 메모리 효율이 뛰어나다는 장점이 있다. 따라서 차세대 시퀀싱 방식으로 생성된 리드를 매핑하는 데에는 적합한 툴이라 할 수 있다. 테스트 쿼리들은 페어드 엔드 방식으로 2개의 미스 매치를 허용하며 매핑되는데, 이는 시퀀싱 에러나 단일 염기 변이(SNP)를 고려한 것이다. 또한 CNV는 테스트 시퀀스에서 염기 서열의 반복 횟수가 레퍼런스 시퀀스에 비해 비해 많거나 적은 영역이므로, 하나의 테스트 쿼리가 레퍼런스 시퀀스의 여러 곳과 동일할 경우 해당



〈그림 2〉 스코어링 예제

영역에 전부 매핑시킨다.

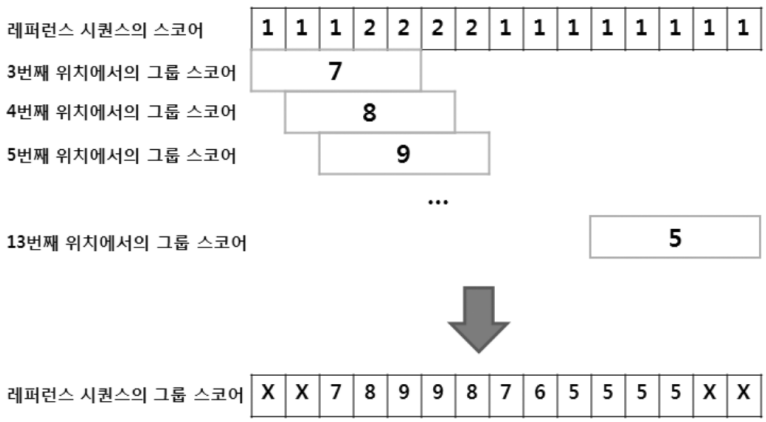
Bowtie는 결과물로 매핑된 리드의 이름, 리드의 방향, 매핑된 레퍼런스 염기 서열의 이름, 리드가 매핑된 레퍼런스의 위치, 리드의 염기 서열, 리드의 퀄리티 스코어(quality score)를 제공한다. 이러한 Bowtie의 결과물을 파싱(parsing)하여 레퍼런스의 각 위치에 매핑된 테스트 쿼리의 수를 계산한다. 이렇게 각 위치에 매핑된 테스트 쿼리 수를 스코어(score)라고 한다. 스코어가 높을수록 증가 영역인 CNV일 가능성이 높다. 그러나 같은 염기 서열을 가진 테스트 쿼리가 여럿이고, 쿼리가 매핑될 수 있는 영역이 레퍼런스 시퀀스에 동일한 숫자만큼 있을 경우에는 CNV로 판단해서는 안 된다. 따라서 하나의 테스트 쿼리가 레퍼런스 시퀀스의 여러 곳에 매핑될 경우, 스코어를 매핑되는 영역의 수로 나눠주는 방식을 취한다. 수식은 다음과 같다.

$$S_p = \sum_{i=1}^{n} x_{\mathrm{pi}} \begin{cases} x_{\mathrm{pi}} = \dfrac{1}{r}, \mathrm{if}\ r \neq 0 \\ x_{\mathrm{pi}} = 0, \mathrm{otherwise} \end{cases}$$

Sp는 레퍼런스 시퀀스의 위치 p의 총 스코어, n은 테스트 쿼리의 수, r은 테스트 쿼리가 매핑된 영역의 수, xpi는 i번째 쿼리가 레퍼런스 시퀀스의 위치 p에서 얻는 스코어를 나타낸다. 〈그림 2〉는 이러한 스코어링(scoring)의 예제이다. 예제에서는 리드의 사이즈는 3bps, 리드 사이의 거리는 4bps이다.

3.3. 슬라이딩 윈도를 이용한 그룹화

테스트 쿼리는 테스트 시퀀스의 임의의 영역에서 추출되며, 한 번 추출된 부분이 다시 추출될 수도 있으므로 테스트 시퀀스의 모든 염기 서열이 고르게 테스트 쿼리로 만들어지지 못한다. 이상적인 경우에는 CNV 영역은 항상 복제 수×커버리지에 해당하는 스코어를 가지게 되나, 이러한 이유로 스코어에 오차가 생기게 되어 연속되는 1Kbps 이상의 높은 스코어를 갖는 영역을 추출하는 것 역시 어려워진다. 이러한 문제를 해결하기 위하여 150bps의 슬라이딩 윈도(sliding window)를 적용하여 주위의 값을 고려한 스코어를 갖게 하였다. 이는 특정 위치를 기준으로 리드가 매핑될 수 있는 범위인 앞뒤 75bps까지를 묶고, 이들 스코어의 합을 그 위치의 대표 스코어로 삼는 방식이다. 이렇게 묶인 150bps 영역을 그룹이라 한다. 예를 들어 레퍼런



〈그림 3〉 슬라이딩 윈도를 이용한 그룹화의 예제

스 시퀀스의 위치 100의 그룹 스코어는 위치 26~175의 합이 되며, 위치 101의 그룹 스코어는 위치 27~176의 합이 된다. 기존에는 CNV의 최소 길이인 1Kbps 사이즈의 슬라이딩 윈도를 사용하여 바로 CNV 영역을 판별하였으나, 이는 1Kbps 이하의 짧은 서열이 반복되는 영역을 CNV로 잘못 찾을 확률이 높아 추가적인 보정이 필요하다는 단점이 있었다. 본 연구에서는 윈도 사이즈를 줄임으로써 이러한 문제를 해결하였다. 〈그림 3〉은 5bps 사이즈의 슬라이딩 윈도를 이용한 그룹화 예제를 보여준다. 레퍼런스 시퀀스의 양쪽 끝은 그룹으로 묶일 수 없으므로 이 부분에서의 CNV 발견은 불가능하나, 윈도 사이즈의 1/2 정도의 작은 영역이므로 전체적인 CNV 발견에 영향을 미치지는 않는다.

3.4. 그룹 스코어를 이용한 CNV 판별

위에서 계산한 그룹의 스코어를 바탕으로 CNV를 판별하게 된다. 이를 위해서는 CNV로 판별할 수 있는 임계값을 정해야 한다. 각 그룹의 스코어들은 커버리지를 평균값으로 하는 가우시안 분포(Gaussian distribution)에 가까운 형태를 갖게 되므로, 가우시안 분포의 엠피리컬 규칙(empirical rule)을 활용하여 임계값을 설정하였다. 즉 전체 그룹 스코어의 평균값을 μ, 전체 그룹 스코어의 표준 편차를 σ라 할 때, $\mu + k\sigma$ 이상의 영역을 증가 영역으로 판단하고, $\mu - k\sigma$ 이하의 영역을 감소 영역으로 판단하게 된다. k는 매개 변수로, k 값의 변화에 따라 선택되는 CNV 영역이 변하게 된다. 예를 들어 k 값을 높게 설정하면 다른 곳에 비해 값이 보다 많이 높거나 낮은 지역만을 CNV로 판단하게 된다. 이는 CNV가 아닌 영역을 CNV로 판단하는 잘못된 긍정(false positive)을 줄이고 CNV 영역을 찾아내지 못하는 잘못된 부정(false negative)을 늘리게 되므로 적절한 k 값을 설정하는 것이 중요하다.

CNV의 최소 길이는 1Kbps이므로, 위의 과정을 통하여 찾은 임계값을 넘는 영역이 1Kbps 이상 연속될 경우 이를 CNV 영역으로 판단하게 된다. 그러나 위에서 언급한 테스트 쿼리가 고르게 나타나지 않는다는 문제점과, 매핑 시 오류가 나타날 수도 있기에 실제로 CNV 구간이더라도 중간에 임계값을 넘지 못하는 값들이 나타날 수 있다. 이는 긴 CNV를 짧은 여러 개의 CNV로 판단하게 한다는 문제점이 있다. 본 연구에서는 SW-Array[16]에서 사용한 스미스-워터만(Smith-

Waterman) 알고리즘을 간략하게 적용하여 이러한 문제를 해결하였다. 이 알고리즘은 높은 값 혹은 낮은 값이 연속되는 지역을 구하기 위한 알고리즘으로, 수식은 다음과 같다.

$$S(p) = \begin{cases} S(p-1) + X(p), \text{if } S(p-1) + X(p) > 0 \\ 0, \text{otherwise} \end{cases}$$

$$B(p) = \begin{cases} B(p-1), \text{if } S(p) > 0 \\ p, \text{otherwise} \end{cases}$$

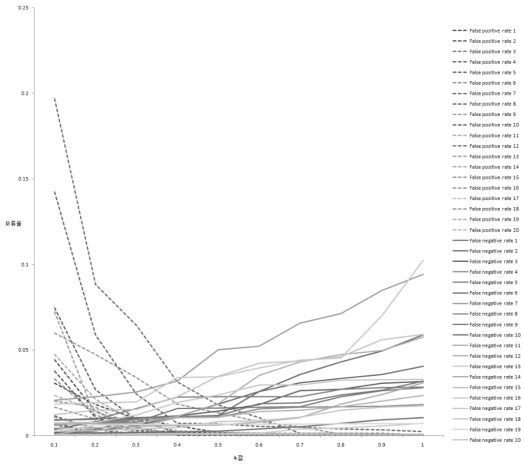
위의 수식에서 p는 레퍼런스 시퀀스에서의 위치를 의미하며, X(p)는 각 위치가 갖는 점수를 의미한다. X(p)는 p가 CNV 영역으로 판단되었을 경우 1, 아닐 경우 -1000의 값을 갖는다. B(p)는 CNV가 시작되는 지점을 나타낸다. S(p)는 p 위치까지의 구간 스코어로, S(p)의 값이 0 이하로 내려가는 곳이 CNV가 끝나는 지점이 된다. 이러한 과정을 통하여 보정된 1000Kbps 이상의 영역이 최종적인 CNV 영역으로 판별된다.

4. 실험 및 결과

실험 환경은 AMD Athlon 64 X2 Dual 5200(3.00GHz), 2GB RAM의 시스템에 운영 체제로 Windows XP 서비스팩 3을 사용한 환경이다. 우선 기존의 기법과의 비교를 위하여 기존의 연구에서 사용했던 데이터를 이용한 실험을 수행하였다. 레퍼런스 시퀀스로 NCBI Build 37.1의 19번 염색체의 컨티그 NT_011255.14가 사용되었으며, 여기에 가상 유전 변이 생성기[22]를 이용하여 다양한 유전 변이를 추가한 시퀀스를 테스트 시퀀스로 사용하였다.

실험은 3장에서 제안한 것과 같은 방식으로 진행하였다. 테스트 시퀀스에서 테스트 쿼리를 생성하고, Bowtie를 이용하여 테스트 쿼리를 레퍼런스 시퀀스에 매핑함으로써 스코어를 할당하였다. 이어서 이들 스코어를 150bps 단위로 그룹화를 하여 대표 스코어를 설정하고, 이들 값을 가우시안 분포의 엠피리컬 규칙을 활용하여 임계값을 설정하였다. 임계값을 넘는 영역이 1kbps 이상 연속될 경우 CNV 영역으로 판단하게 되며, 이 결과에 스미스-워터만 알고리즘을 간략하게 적용함으로써 오류에 의해 연속된 값이 나오지 않게 되는 문제를 보정하였다.

본 실험에서 테스트 쿼리는 기존 실험과 동일한 3 커버리지를 생성하였다. 커버리지를 높일수록 더 많은 리드를 레퍼런스에 매핑하여 결과의 정확도를 높일 수 있으나, 커버리지를 높일 때마다 생물학적인 실험 비용은 증가한다. 따라서 가능한 한 낮은 커버리지로 CNV를 찾는 방법이 필요하다. 현재 인간 전체 염색체의 염기 서열을 결정할 때 27.8 커버리지 이상의 많은 리드들이 사용된다. 그러나 CNV를 결정하는 일은 염기 서열의 조립이 아닌 매핑을 통한 스코어링을 하는 방식이기 때문에 27.8보다는 확연히 낮은 커버리지를 통해서도 본 논문의 목적을 달성할 수 있었고, 낮은 커버리지가 갖는 정확도의 감소 문제를 고려하여 최종적으로 3 커버리지를 실험적으로 결정하여 실험을 진행하였다. 또한 〈그림 4〉와 같이 20개의 데이터를 이용한 실험을 통하여 매개 변수 k에 따른 오류율의 변화를 확인하여 적절한 k 값을 설정하였다. 매개 변수는 0.1부터 1.0까지 변화시키며 실험하였으며, 점선은 잘못된 긍정의 비율, 실선은 잘못된 부정의 비율을 나타낸다. k 값이 증가하면 잘못된 부정이 증가하고, 반대로 k 값이 감소하면 잘못된 긍정이 증가하므로 적당한 k 값을 설정할 필요가 있다. 실험에서는 오류율의 급격한 변화가 적어지는 0.5와, 잘못된 긍정이 최솟값에 가까워지는 0.7을 k 값으로 설정하였다.



〈그림 4〉 매개 변수 k에 따른 오류율의 변화

〈표 2〉는 기존의 실험 결과[22]와 본 실험의 결과를 나타낸다. 실제 CNV 영역은 테스트 시퀀스에 있는 CNV 영역, 발견한 CNV 영역은 본 기법으로 찾은 CNV의 영역, 잘못된 긍정은 실제로는 CNV 영역이 아닌 중립 영역이나 CNV로 잘못 판단한 영역, 잘못된 부정은 실제로는 CNV 영역이나 본 기법으로 발견하지 못한 부분을 뜻한다. 잘못된 긍정의 비율은 발견한 CNV 영역 중 잘못된 긍정의 비율이며, 잘못된 부정의 비율은 실제 CNV 영역 중 잘못된 부정의 비율이다. 〈표 3〉을 통하여 본 실험이 기존의 실험에 비하여 잘못된 긍정과 잘못된 부정을 크게 줄였고, 기존에 비하여 많은 CNV 영역을 발견하였으며, 짧은 CNV부터 긴 CNV까지 고르게 찾았다는 것을 확인할 수 있다. 기존 실험에 비하여 오류율이 줄어든 가장 큰 이유는 기본적으로 사용되는 리드의 길이가 길어졌으며, 페어드 엔드 정보를 이용하여 리드를 매핑하기 때문에 유전적 변이의 영향을 적게 받는 매핑이 가능해졌기 때문이다. 또한 슬라이딩 윈도의 크기를 줄여 짧은 반복 영역을 제외함으로써 보다 정확한 결과가 도출되었다. 추가적으로 동일 실험 환경에서 기존의 실험은 약 8시간이 소요되었으나, 본 실험은 약 15분만에 CNV를 찾아낼 수 있었다. 이러한 속도의 향상은 프로세스의 수정과 코드 최적화, 그리고 짧은 리드에 최적화된 매핑 툴인 Bowtie의 사용으로 가능해졌다.

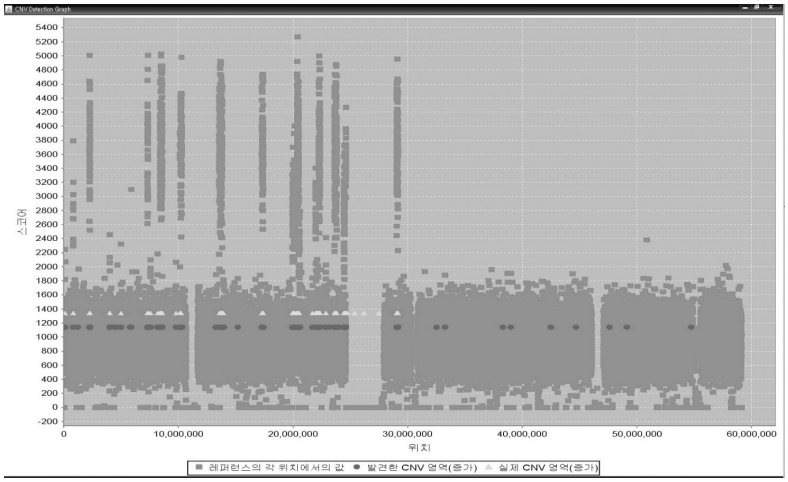
기존 실험의 경우 데이터의 크기가 작고 감소 영역의 CNV가 없었기에, 보다 나은 검증을 위하여 NCBI Build 37.1의 19번 염색체의 시퀀스 전체를 레퍼런스 시퀀스로 사용하고, 여기에 감소를 포함한 가상 유전변이를 심은 시퀀스를 테스트 시퀀스로 사용하여 추가적인 실험을 실시하였다. 본 실험에서 얻은 결과는 〈표 3〉과 〈그림 5-8〉을 통하여 나타냈다. 〈그림 5-8〉은 1000bps 단위로 샘플링한 그래프이다. X 축은 레퍼런스 시퀀스에서의 위치, Y 축은 스코어를 나타내며, 주황색 점은 각 그룹의 스코어, 보라색 점은 발견한 CNV 영역, 연두색 점은 실제 CNV 영역을 나타낸다. 추가적으로 보라색 점의 Y 값은 CNV를 구분하는 판별식인 $\mu \pm k\sigma$의 값을 나타낸다.

〈표 2〉 컨티그 NT_011255.14을 이용한 실험의 결과 및 기존 실험[22]과의 비교

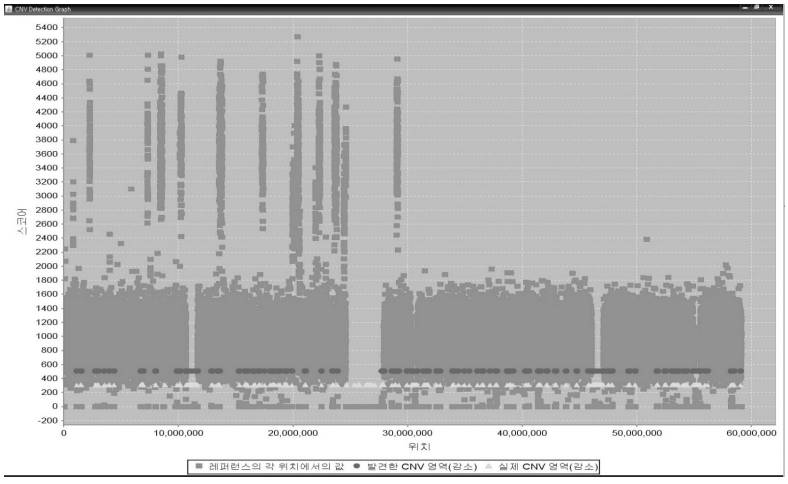
카테고리	기존 실험에서의 값	본 실험에서의 값(k=0.5)	본 실험에서의 값(k=0.7)
실제 CNV 영역	484724 bps	484724 bps	484724 bps
발견한 CNV 영역	450286 bps	465521 bps	460965 bps
잘못된 긍정(False positives)	24824 bps	58 bps	41 bps
잘못된 부정(False negatives)	59262 bps	19261 bps	23800 bps
잘못된 긍정의 비율(False positive rate)	0.0551294	0.000124592	0.0000889438
잘못된 부정의 비율(False negative rate)	0.122259	0.039736	0.0491001

〈표 3〉 19번 염색체 전체를 이용한 실험의 결과(증가 영역과 감소 영역)

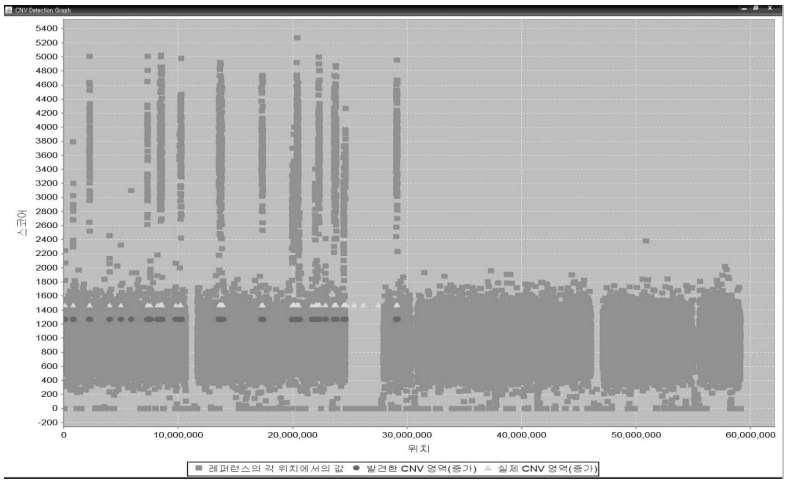
카테고리	증가 영역(k=0.5)	감소 영역(k=0.5)	증가 영역(k=0.7)	감소 영역(k=0.7)
실제 CNV 영역	1857960	8758847	1857960	8758847
발견한 CNV 영역	1742407	8016097	1709247	7976720
잘못된 긍정(False positives)	131257	44720	107405	28155
잘못된 부정(False negatives)	246810	787470	256118	810282
잘못된 긍정의 비율(False positive rate)	0.0753308	0.00557877	0.0628376	0.00352965
잘못된 부정의 비율(False negative rate)	0.132839	0.0899057	0.137849	0.0925101



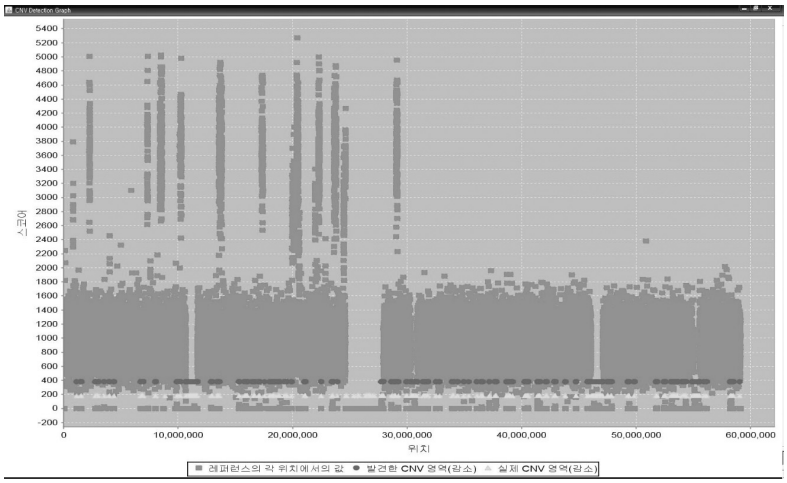
〈그림 5〉 증가 영역의 CNV(k=0.5)



〈그림 6〉 감소 영역의 CNV(k=0.5)



〈그림 7〉 증가 영역의 CNV(k=0.7)



〈그림 8〉 감소 영역의 CNV(k=0.7)

기존의 실험에 비해 사용된 시퀀스의 길이와 복잡도가 증가하였기에 전체적으로 잘못된 긍정과 잘못된 부정이 증가하였으나, 여전히 높은 확률로 CNV 영역을 찾고 있다는 것을 확인할 수 있다. 기존의 실험의 비해 오류, 특히 잘못된 부정이 증가한 이유는 레퍼런스 시퀀스에 밝혀지지 않은 일부 영역들, 즉 컨티그와 컨티그 사이의 간격이 존재하기 때문이다. 밝혀지지 않은 영역에는 리드의 매핑이 불가능하므로, 이러한 부분에 존재하는 CNV는 찾아낼 수 없다. 이러한 문제는 시퀀스의 서열이 밝혀지면 자연히 해결될 수 있다.

앞선 실험은 모두 실제 레퍼런스를 기준으로 가상의 시퀀스를 컨티그 단위, 염색체 단위로 생성한 후 수행되었다. 마지막으로 실제 데이터에 대한 유전체 변이를 찾기 위해서 NCBI Build 37.1의 19번 염색체를 레퍼런스 시퀀스로 사용하고, Celera 시퀀싱 방식의 19번 염색체를 테스트 시퀀스로 사용하여 CNV 후보 영역을 검출하는 실험을 수행하였다. 커버리지는 3, k 값은 0.7을 사용하였다. 이 실험을 통하여 71346의 증가 영

역과 1087026의 감소 영역을 CNV 후보 영역으로 판단하였다. 본 후보 영역은 생물학적 검증을 통하여 CNV 영역을 찾아내는 데 도움을 줄 수 있으며, 향후 염기 서열을 이용하여 CNV 영역을 찾는 연구가 진행될 때 비교 대상으로 활용할 수 있다. CNV 후보 영역에 대한 상세한 내용은 부록으로 첨부하였다.

5. 결론

본 논문에서는 차세대 시퀀싱으로 생성된 짧은 페어드 엔드 리드들을 이용하여 CNV를 발견하는 방법에 대하여 기술하였다. 차세대 시퀀싱을 시뮬레이트하여 기존의 밝혀진 염기 서열로부터 리드를 생성하고, 생성된 리드들을 bowtie를 이용하여 비교의 기준이 되는 염기 서열인 레퍼런스 시퀀스에 매핑하게 된다. 매핑의 결과값을 이용하여 각 레퍼런스 시퀀스의 각 위치에 매핑된 리드의 수를 세고, 이들 값을 슬라이딩 윈도를 통하여 그룹화한 후, 가우시안 분포의 엠피리컬 규칙을 이용하여 CNV 판별 기준을 세운다. 마지막으로 SW-Array에 사용되었던 스미스-워터만 알고리즘을 이용한 보정을 통하여 최종 CNV 영역을 판단하게 된다. 본 연구의 결과를 검증하기 위하여 우선 기존 기법의 검증에 사용하였던 다양한 유전 변이를 넣은 가상의 데이터를 이용하였으며, 마지막으로 실제 데이터인 Build 37.1의 19번 염색체와 Celera 시퀀싱 방식의 19번 염색체의 비교를 통하여 CNV 영역을 밝혀내었다.

본 연구는 차세대 시퀀싱 방식으로 생성된 페어드 엔드 리드를 활용한 첫 연구이다. 본 연구는 기존의 연구에 비해 잘못된 긍정(false positive)과 잘못된 부정(false negative)을 줄이고, 매개 변수에 대한 민감도를 낮추었으며 처리 속도를 향상시켰다. 추가적으로 본 기법은 증가된 CNV만을 찾을 수 있던 기존의 방식과는 달리 감소된 CNV까지 찾아낼 수 있다는 장점이 있다. 실험을 통하여 본 연구는 기존의 연구에 비해 많은 CNV 영역을 보다 정확하게 찾을 수 있으며, 기존에 찾지 못했던 감소 영역의 CNV까지 찾아낼 수 있다는 것을 확인하였다. 또한 NCBI Build 37.1의 19번 염색체를 레퍼런스 시퀀스로 사용하고, Celera 시퀀싱 방식의 19번 염색체를 테스트 시퀀스로 사용하여 CNV 후보 영역을 제시하였다. 차후에는 인간의 전체 염기 서열과 같은 큰 데이터 셋과 실제 차세대 시퀀싱으로 생성된 리드들을 이용한 추가적인 실험을 진행할 예정이다.

6. 참고 문헌

[1] R. Redon et al., “Global variation in copy number in the human genome”, Nature, Vol. 444, pp. 444-454, 2006.

[2] A. J. Iafrate et al., “Detection of large-scale variation in the human genome” , Nature Genetics, Vol. 36, pp.949-851, 2004.

[3] J. L. Freeman et al., “Copy number variation: New insights in genome diversity”, Genome Research, Vol.16, pp.949-961, 2006.

[4] S. A. McCarroll et al., “Copy-number variation and association studies of human disease”, Nature Genetics, Vol. 39, pp. S37-S42, 2007.

[5] G. Perry et al., “Diet and the evolution of human amylase gene copy number variation”, Nature Genetics, Vol. 39, No. 10, pp. 1256-1260, 2007.

[6] J. Beckmann et al., “Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability”, Nature Genetics, Vol. 8, No. 8, pp. 639-646, 2007

[7] E. Tuzun et al., "Fine-scale structural variation of the human genome", Nature Genetics, Vol. 37, No. 7, pp. 727-732, 2005.

[8] R. E. Mills et al., "An initial map of insertion and deletion (INDEL) variation in the human genome", Genome Research, Vol. 16, pp. 1182-1190, 2006.

[9] R. Khaja et al., "Genome assembly comparison identifies structural variants in the human genome", Nature Genetics, Vol. 38, No. 12, pp. 1413-1418, 2006.

[10] J. Shendure et al., "Next-generation DNA sequencing", Nature biotechnology, Vol. 26, pp.1135-1144, 2008

[11] S. F. Altschul et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Research, Vol. 25, No. 17, pp. 3389-3402, 1997.

[12] http://bowtie-bio.sourceforge.net/index.shtml

[13] 김태민, "Copy number variation (CNV)의 인간유전체 내 기능 및 진화경로에 관한 연구", 생화학분자생물학소식, Vol.15, No.3, pp. 40-51, 2008.

[14] The International HapMap Consortium, "A haplotype map of the human genome", Nature, Vol. 437, No. 7063, pp. 1299-1320, 2005.

[15] A. B. Olshen et al., "Circular binary segmentation for the analysis of array-based DNA copy number data", Biostatistics, Vol. 5, No. 4, pp. 557-572, 2004.

[16] T. S. Price et al., "SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data", Nucleic Acids Research, Vol. 33, No. 11, pp. 3455-3464, 2005.

[17] K. Wang et al., "PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data", Genome Research, Vol.17, pp.1665-1674, 2007.

[18] C. Barnes et al., "A robust statistics method for case-control association testing with copy number variation", Nature genetics, Vol. 40, No. 10, pp. 1245-1252, 2008.

[19] T. LaFramboise et. al., "A flexible rank-based framework for detecting copy number aberrations from array data", Bioinformatics, Vol. 25, No. 6, pp. 722-728, 2009.

[20] D. F. Conrad et al., "A high-resolution survey of deletion polymorphism in the human genome", Nature Genetics, Vol. 38, No. 1, pp. 75-81, 2006.

[21] S. W. Schrer et al., "Challenges and standards in integrating surveys of structural variation", Nature Genetics, Vol. 39, pp. S7-S15, 2007.

[22] M. J. Moon et al., "A Computational Approach to Detect CNVs Using High-throughput Sequencing", Proceedings of 9th IEEE International Conference on BioInformatics and BioEngineering, pp.266-271, 2009

[23] A. J. Iafrate et al., "Detection of large-scale variation in the human genome", Nature Genetics, Vol. 36, No. 9, pp. 949-951, 2004.

[24] http://www.ncbi.nlm.nih.gov/projects/SNP/

7. 부록

NCBI Build 37.1와 Celera 시퀀싱 방식의 19번 염색체를 사용하여 발견한 CNV 후보 영역

증가 영역의 CNV			
252708~254089	12935492~12936613	29135783~29137059	41563990~41565118
256549~257607	13061188~13062254	31257079~31258472	43179091~43180641
1237269~1238462	13118714~13119862	33773059~33774323	43730201~43731200
1568274~1569395	15611689~15612769	34556852~34558008	44360386~44361426
2108837~2110055	16536661~16537828	35380726~35381796	46244840~46246034
2202085~2203444	16702370~16703498	35605352~35606357	49188826~49189840
2574734~2575759	17536421~17537438	35666774~35667803	49692750~49693847
5709899~5710924	20425311~20426421	36194643~36195650	50060835~50062022
6314999~6316160	22256375~22257415	36407815~36409053	51340370~51341442
7457852~7458964	22571747~22572843	36669535~36670626	51701619~51702660
8236305~8237325	22576892~22577899	36980840~36981861	52707485~52708508
11151357~11152384	23314451~23315532	38302961~38304077	52856606~52857670
11177544~11178639	23419957~23420975	39335292~39336399	52932349~52933623
12109492~12110556	23695693~23696851	40088343~40089343	53806730~53807833
12263236~12264338	24265773~24266948	40651690~40652745	56940822~56941829
12788725~12789907	28340392~28341427	41486858~41487866	57437612~57438675

감소 영역의 CNV			
60001~116902	7274837~7275857	20504640~20506266	46942510~46943815
117730~245190	7293373~7294427	20506277~20508130	48406584~48462750
332645~333678	7303386~7345855	20511144~20513387	48513811~48520802
350020~351581	7428227~7429429	20514154~20515463	49172335~49173377
394396~395694	7432955~7439235	20521865~20523398	49317964~49328066
399493~400585	7663943~7665654	20573270~20574330	49963989~49965269
403654~405230	7798183~7799333	20575436~20577289	50473267~50474521
431212~439899	8262223~8263339	20580303~20581501	50580474~50584016
568335~571277	8338634~8339731	20583316~20585825	50592643~50643326
771100~772775	8340885~8342455	20585975~20588791	50994828~50996635
815784~817132	8345740~8347913	21005122~21006663	51001712~51003220
817881~818974	8349115~8352197	21824455~21825948	51119367~51120919
871159~873071	8353497~8355185	21828586~21830029	51222476~51223546

감소 영역의 CNV			
1032938~1035153	8355718~8358136	21902549~21905865	52176845~52177904
1051876~1052971	8359693~8362647	22579009~22580404	53095527~53098076
1160939~1162453	8363936~8366228	23140343~23143245	53319898~53321361
1196240~1197398	8480868~8482706	23384254~23385361	53321919~53323022
2128077~2129173	8485799~8488326	23392269~23393568	53391112~53392253
2292060~2293825	8489923~8491082	23505083~23506252	53421951~53424298
2295738~2298015	8619593~8621784	23648786~23652911	53669557~53671630
2584407~2587879	8625418~8626427	23834260~23835265	53688467~53691175
2594555~2596380	8851223~8901337	23860631~23861676	53691215~53692873
2638202~2640254	8996563~8998756	24055210~24056209	53784637~53785751
2817436~2818812	9004109~9009220	24061041~24062735	54423202~54426951
2893608~2894776	9009699~9010738	24525506~24589656	54637243~54638427
2908765~2911267	9021314~9022587	24590746~24591802	54723360~54725088
3047192~3048422	9051908~9053122	24592429~24631633	54725191~54739116
3145056~3146072	9126091~9128351	27795892~27797183	54745333~54747807
3172824~3175458	9147306~9149912	27804257~27805585	54780326~54789757
3235074~3236117	9174102~9177683	27848684~27851781	54918457~54925464
3318328~3319634	11535013~11536476	27960554~27963566	55218093~55219155
3321821~3323446	11725726~11727106	28133110~28136097	55475973~55477514
3973213~3974682	11775591~11778748	30388295~30389387	55604204~55605791
4035044~4037283	12326142~12327369	30389772~30393218	55776170~55777886
4065171~4066465	12694457~12699184	34736541~34737609	55781357~55782381
4327201~4328472	14200420~14201961	34882409~34883521	55944508~55960523
4329303~4330540	14657565~14675738	36388835~36390069	56038214~56039691
4333149~4335404	14705631~14708037	36759214~36807074	56053767~56054895
4561769~4563540	14718058~14720437	37753155~37794372	56224096~56225240
4618407~4619459	15277144~15278145	38342006~38343624	56252543~56253714
4885030~4886795	15784147~15785306	39240796~39244234	56274074~56282912
5673097~5674459	15786684~15789671	39756961~39759134	56283495~56286375
5744558~5745616	16042015~16045188	40368892~40402256	56361501~56363025
5804671~5805804	16361267~16367825	40428365~40429376	56817528~56818569
6128443~6129828	16887979~16889495	41047222~41052050	57256077~57257269
6625283~6627237	17363045~17364750	43860825~43862214	57469002~57473239
7025220~7026467	17740232~17742082	43862707~43863793	59034437~59035441
7026566~7044075	18799491~18801017	43868552~43874207	59099425~59118834
7044174~7045421	19364013~19365279	43881086~43882113	
7050857~7062750	19798004~19799778	45833616~45835622	

문 명 진

e-mail : psiwind@cs.yonsei.ac.kr

2007년 중앙대학교 컴퓨터공학과 졸업(학사)

2008년~현재 연세대학교 컴퓨터과학과 석사 과정

관심분야 : 데이터베이스, 데이터 마이닝, 바이오인포매틱스

박 치 현

e-mail : tianell@cs.yonsei.ac.kr

2007년 홍익대학교 컴퓨터공학과 졸업(학사)

2009년 연세대학교 컴퓨터과학과 졸업(석사)

2009년~현재 연세대학교 컴퓨터과학과 박사 과정

관심 분야 : 바이오인포매틱스, 데이터 마이닝, 파일시스템

윤 영 미

e-mail : ymyoon@gachon.ac.kr

1981년 서울대학교 자연과학대학 졸업(학사)

1981년~1983년 오하이오 주립대학 수학과 수료(학사)

1987년 스탠포드대학교 컴퓨터과학과 졸업(석사)

1987~1993년 IntelliGenetics Inc., Mountainview, California, Software Engineer

1995년~현재 가천의과학대학교 부교수

2008년 연세대학교 컴퓨터과학과 졸업(박사)

관심분야 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스

안 재 균

e-mail : ajk@cs.yonsei.ac.kr

2006년 연세대학교 컴퓨터과학과 졸업(학사)

2007년~현재 연세대학교 컴퓨터과학과 석사 과정

관심분야 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스

박 상 현

email : sanghyun@cs.yonsei.ac.kr

1989년 서울대학교 컴퓨터공학과 졸업(학사)

1991년 서울대학교 컴퓨터공학과 졸업(석사)

2001년 UCLA 컴퓨터과학과 졸업(공학 박사)

2002년~2003년 포항공과대학교 컴퓨터공학과 조교수

2003년~2006년 연세대학교 컴퓨터과학과 조교수

2006년~현재 연세대학교 컴퓨터과학과 부교수

관심분야 : 데이터베이스, 데이터 마이닝, 바이오인포매틱스, 적응적 저장 장치 시스템