유전체 연구를 위한 Paired-End Reads 병합 데이터의 정렬 이득에 관한 분석

권선영*, 윤성로*
*서울대학교 전기정보공학부
e-mail:sryoon@snu.ac.kr

Alignment Benefits of Merging Paired-End Reads in Genome Analysis

Sun Young Kwon*, Sungroh Yoon*
*Dept of Electrical and Computer Engineering, Seoul University

요 약

유전체 연구를 위한 분석 작업은 표준유전체에 시퀀스 데이터를 정렬하는 과정을 필수적으로 요구한다. 정렬에는 single-end 또는 paired-end reads가 사용된다. Paired-end reads는 유전체 조각의 양쪽에서 시퀀싱 된 데이터로 좀 더 긴 길이에 대한 정보를 얻을 수 있어 많이 이용된다. 정렬 툴 자체적으로 paired-end reads를 다룰 수 있으나, 병합툴을 활용하는 것이 더 좋은 결과를 보인다. 다섯 가지 병합 툴 중에서 CASPER와 pear에서 정렬 이득이 가장 크게 나타난다.

1. 서론

초기 유전체 연구의 일차적인 성과로 표준유전체(reference genome)가 획득되었고, 표준유전체에 시퀀스 데이터를 정렬한 결과를 기준으로 발현량 파악(expression level analysis), 다형성 분석(variant calling) 그리고 RNA-seq 데이터를 활용한 신규유전자(novel gene), 융합유전자(fusion gene) 및 선택적 이어맞추기(alternative splicing) 등의 다양한 연구들이 이루어진다. 따라서 표준유전체에 시퀀스 데이터들을 정확하게 정렬하는 것은 이후 분석을 위해서 매우 중요한 작업이다[1].

일반적인 정렬작업은 시퀀스 데이터를 입력받아 수행하게 되며 sequencing 기법에 따라 single reads 혹은 paired-end reads를 입력받는다. Paired-end reads는 긴 길이의 데이터를 활용하기 위하여 하나의 조각(fragment)을 양쪽에서 추출한 데이터로 분석의 정확도를 높일 수 있어 많이 이용된다[2].

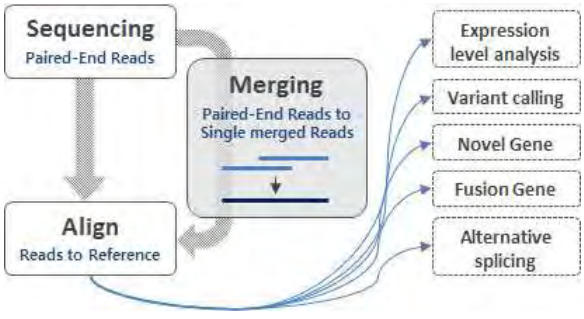

그림 1 시퀀스 데이터 분석 파이프라인

Paired-end reads의 경우, 중첩되는 영역을 가지는 데이터와 중첩되지 않지만 전체길이를 추정 할 수 있는 데이터로 나뉜다. 전자의 경우, 병합 툴을 사용하여 오류가 보정된 하나의 긴 single reads를 생성 할 수 있다.

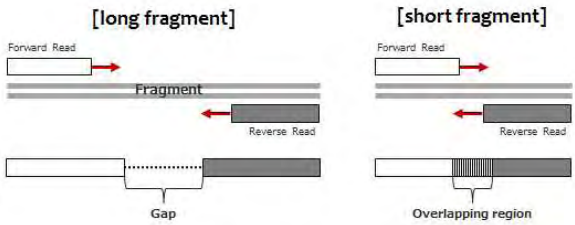

그림 2 paired-end reads의 종류

본 논문에서는 표준유전체에 시퀀스데이터를 정렬하는 툴 중의 하나인 bowtie2[3]를 중심으로 paired-end reads를 병합하기 전과 병합한 후의 정렬 결과를 비교 분석해 보고자 한다.

2. 본론

2.1 병합방법

Paired-end reads를 병합하는 방법에는 FLASH (Fast Length Adjustment of SHort reads)[4], COPE (Connecting Overlapping Pair-End)[5], PANDAseq (PAired-eND Assembler for Illumina sequence)[6], PEAR(Paired-End reAd mergeR)[7], CASPER (Context-Aware Scheme for Paired-End Reads)[8] 들이 최근 소개되었다.

병합은 정확한 조각의 크기를 알 수 없으므로 (1) 올바른 병합위치를 찾는 단계와 시퀀싱의 오류로 (2) 일치하지 않는 base 정보를 올바르게 선택하는 두 단계로 이루어진다. 모든 입력 데이터가 병합이 가능한 것은 아니며, 경우에 따라 병합되지 않은 데이터를 산출하기도 한다.

2.2 데이터

병합에 사용된 시퀀스 데이터는 Illumina로 시퀀싱 된 샷건(shutgun) 방식의 전체유전체가 사용되었으며, 황색포도상구균(staphylococcus aureus)과 광합성세균(rhodobacter sphaeroides)을 이용하였다..[9]

표준유전체는 staphylococcus aureus의 경우 NCBI(2006-02-13), 2,821,361bp, rhodobacter sphaeroides의 경우 NCBI(2005-10-07), 4,603,760bp 데이터를 이용하였다.

구분	staphylococcus aureus	rhodobacter sphaeroides
Avg Read length	101 bp	101 bp
Fragment length	180 bp	180 bp
# of reads	647,052	1,025,434

표 2 whole genome shotgun illumina data

2.3 실험방법

병합은 각 병합툴의 기본 옵션값을 이용하여 수행하고, 시퀀스 데이터를 입력으로 하여 병합된 긴 길이의 merged data와 병합되지 못하고 그대로 남겨진 unmerged 데이터를 산출한다.

정렬은 bowtie2를 사용하여 표준데이터 (reference genome)에 ① 병합되기 전의 결과를 확인하기 위하여 시퀀스 데이터인 paired reads 데이터를 이용하여 정렬한다. ② 병합된 후의 결과를 확인하기 위하여 병합된 merged single data와 unmerged paired data를 입력으로 이용하며, 옵션은 모두 기본값으로 수행한다.

3. 실험결과

정렬 결과는 크게 ① 정렬이 되지 않은 경우, ② 정확하게 한번, ③ 한번 이상 정렬되는 경우로 나뉜다. 반복되는 영역이 많거나 시퀀스 데이터가 짧다면 한번 이상 정렬되는 경우가 많이 발생한다. 실험 결과는 정확하게 한번 정렬되는 경우를 기준으로 측정하였다.

실험 결과는 아래 표2와 같으며, 두 종 모두 bowtie2만으로 정렬한 것 보다 병합 툴을 이용한 경우 더 좋은 정렬 결과를 보인다.

병합 툴 중에서 pear의 경우 false positive를 줄이기 위하여 병합되는 비율은 아주 낮으나(약 15%~30%) 전체적으로 좋은 결과를 보이며, CASPER의 경우 병합되는 비율이 80%대로 매우 높으며 정렬 결과도 좋다. Pandaseq의 경우는 병합을 하지 않은 경우가 오히려 더 좋은 결과를 보이기도 한다.

Staphylococcus aureus의 경우 CASPER가 가장 좋은 성능을 보이고 있으며, rhodobacter sphaeroides는 pear가 가장 좋은 결과를 보인다.

구분		bowtie2	flash	cope	panda	pear	CASPER
SA	# of merged	-	369,276	314,281	537,401	202,221	534,288
	# of unmerged	647,052	277,776	332,771	109,651	444,831	112,764
	merging rate	-	57.07	48.57	83.05	31.25	82.57
	aligned (exactly 1)	**419,218**	**425,952**	423,075	417,731	421,790	**429,907**
RS	# of merged	-	386,418	279,991	732,114	155,758	893,285
	# of unmerged	1,025,434	639,016	745,443	293,320	869,676	132,149
	merging rate	-	37.68	27.30	71.40	15.19	87.11
	aligned (exactly 1)	**772,433**	74,252	771,827	720,363	**814,424**	**801,722**

표 3 bowtie2 측정 결과

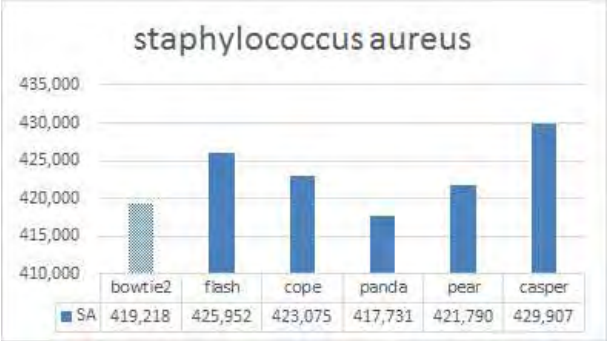

그림 3 staphylococcus aureus 측정 결과

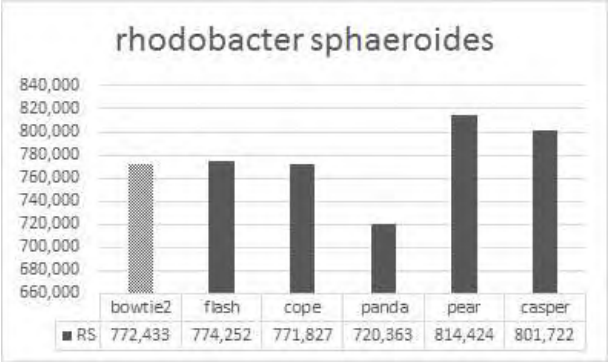

그림 4 rhodobacter sphaeroides 측정 결과

4. 결론

본 논문에서는 유전체 분석 파이프라인의 필수 단계인 표준유전체로의 정렬 성능을 높이기 위한 실험을 진행하였고, 실험 결과 paired-end reads를 병합 한 후 입력데이터로 활용하는 것이 정렬에 더 좋은 결과를 보임을 확인하였다. 또한 여러 가지 병합 툴 중에서 pear, CASPER가 좋은 결과를 보이고 있다.

사사

이 논문은 2014년도 정부 (미래창조과학부)의 재원으로 한국연구재단 (No. NRF-2011-0009963), 바이오.의료기술개발사업 (No. 2012M3A9D1054622), 두뇌한국21플러스사업의 지원을 받아 수행된 연구임.

참고문헌

[1] Li, Heng, and Nils Homer. "A survey of sequence alignment algorithms for next-generation sequencing." Briefings in bioinformatics 11.5 (2010): 473-483.

[2] Rodrigue, Sébastien, et al. "Unlocking short read sequencing for metagenomics." PLoS One 5.7 (2010): e11840.

[3] Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." Nature methods 9.4 (2012): 357-359.

[4] T. Magoč and S. L. Salzberg, "FLASH: fast length adjustment of short reads to improve genome assemblies," Bioinformatics, vol. 27, pp. 2957-2963, 2011

[5] B. Liu, J. Yuan, S. M. Yiu, Z. Li, Y. Xie, Y. Chen, et al., "COPE: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly," Bioinformatics, vol. 28, pp. 2870-4, Nov 15 2012.

[6] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld, "PANDAseq: paired-end assembler for illumina sequences," BMC Bioinformatics, vol. 13, p. 31, 2012.

[7] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, "PEAR: a fast and accurate Illumina Paired-End reAd mergeR," Bioinformatics, vol. 30, pp. 614-620, 2014.

[8] Sunyoung Kwon, Byunghan Lee, and Sungroh Yoon, "CASPER: Context-Aware Scheme for Paired-End Read from High-Throughput Short-Read Amplicon Sequencing," in BMC Bioinformatics , vol. 15, no. suppl 9, p. S10, September 2014.

[9] Salzberg, Steven L., et al. "GAGE: A critical evaluation of genome assemblies and assembly algorithms." Genome research 22.3 (2012): 557-567.