NGSOne: 클라우드 기반의 유전체(NGS) 데이터 분석 툴

[1]권창혁, [2]김원호, [3]장정화,[*4]안재균
[1] 인천대학교, 마이지놈박스, netbuyer@inu.ac.kr
[2] 마이지놈박스, jasonkim@mygenomebox.com
[3] 마이지놈박스, jjh0005@mygenomebox.com
[*4] 교신저자 인천대학교, jgahn@inu.ac.kr

NGSOne: Cloud-based NGS data analysis tool

[1]Chang-hyuk Kwon, [2]Jason Kim, [3]Jeong-hwa Jang, [*4]Jae-gyoon Ahn
[1] *Incheon National University, MyGenomeBox, Co, netbuyer@inu.ac.kr*
[2] *MyGenomeBox, Co, jasonkim@mygenomebox.com*
[3] *MyGenomeBox, Co, jjh0005@mygenomebox.com*
[*4,Corresponding Author] *Incheon National University, jgahn@inu.ac.kr*

요 약

개인 전장 유전체 분석 가격의 하락으로 많은 국가들이 10만명에서 100만명까지의 대량 전장 유전체 분석과 엑솜 시퀀싱을 진행하고 있다. 하지만 많은 대형 프로젝트에서 대량의 데이터를 처리할 수 있는 프로그램이나 시스템의 부족으로 많은 비용이 클러스터 구축 및 시스템 구매 비용으로 소비되고 있다. 본 연구에서는 자체 서버나 클러스터 환경을 구축하지 않고도 동시에 수백 개 이상의 전장 유전체 및 엑솜에 대한 단일 염기 다형성(Single Nucleotide Polymorphism; SNP) 분석을 수행할 수 있고, 생물학자들도 쉽게 설치하여 운영할 수 있는 클라이언트 프로그램인 NGSOne을 개발하였다. 대표적인 SNP 분석 도구인 DRAGEN, BWA/GATK 및 Isaac/Strelka2를 선택하여 분석할 수 있고, 3개 툴에서 실행 시간 및 에러의 개수 면에서는 DRAGEN이 가장 좋은 성능을 보였다. 또한 NGSOne은 SNP 분석뿐만 아니라 다양한 분석 도구의 자동적인 실행을 위한 확장이 가능하다.

Abstract

With the decrease of sequencing price, many national projects that analyzes 0.1 to 1 million people are now in progress. However, large portion of budget of these large projects is dedicated for construction of the cluster system or purchase servers, due to the lack of programs or systems that can handle large amounts of data simultaneously. In this study, we developed NGSOne, a client program that is easy-to-use for even biologists, and performs SNP analysis using hundreds or more of Whole Genome and Whole Exome analysis without construction of their own server or cluster environment. DRAGEN, BWA / GATK, and Isaac / Strelka2, which are representative SNP analysis tools, were selected and DRAGEN showed the best performance in terms of execution time and number of errors. Also, NGSOne can be extended for various analysis tools as well as SNP analysis tools.

***Keywords**: NGSOne, DRAGEN, GATK, Isaac, WGS, WES, Pipeline*

* Corresponding Author
Received: NOV. 04, 2018, Revised: Dec. 03, 2018, Accepted: Dec. 26, 2018

Ⅰ. 서 론

차세대 유전자 분석 기술 (NGS, Next Generation Sequencing)의 급진적인 발전으로 1000 달러 게놈이 실현 되었고[1][2], 인구 집단, 건강, 질병에 대한 연구는 점차 새로운 임상 적용을 위한 치료 표적 및 마커에 대한 단서를 제공할 있는 기능적 연구와 통계 기반으로 질병에 직접 관여하는 게놈 변이 연구로 확대되고 있다. 그러나 이러한 변이의 기능 해석 과정에서 유전 변이의 존재 여부, 시퀀싱 오류 등의 기술적인 문제나 낮은 시퀀싱 깊이(Raw Sequencing Depth)와 대량 데이터의 부족으로 인해서 많은 위험이 따르고 있기 때문에, 정확한 게놈의 분석과 변이체 확인은 NGS 기술에 기반한 임상 유전체학의 성공을 위한 중요한 요소가 되고 있다.

기존 100 기가에 가까운 전장 유전체 (Whole Genome) 데이터 분석 처리 시간은 단일 서버로 2 일 이상 소요되었으나 대량 클러스터 구축, 클라우드의 적용, FPGA (Field Programmable Gate Array) 를 이용한 하드웨어 기술의 적용[3] 및 새로운 알고리즘 개발로[4] 1~3 시간에 모든 분석을 끝낼 수 있는 기술들이 소개되고 있다. 그러나 많은 수의 배포 프로그램이나 기술들이 개발자 위주로 되어 있기 때문에, 직접 프로그래밍을 해야 하거나 대용량 서버 운영 기술을 가지고 있어야 테스트가 가능할 정도로 어렵고 복잡하다.

본 연구에서는 생물학자나 임상연구자들이 기본적인 용어와 기능 매뉴얼의 확인만으로 쉽게 사용할 수 있고, 윈도우, 맥, 리눅스 등의 클라이언트 운영 플랫폼에 상관없이 설치 및 사용할 수 있을 뿐만 아니라, 유전자 분석의 가장 대표적이고 특징적인 3 개의 툴을 선택하여 수많은 테스트를 수행하여 동시에 수백 샘플이 들어와도 분석이 가능하며 속도와 정확도를 모두 고려할 수 있는 클라우드 기반의 툴인 NGSOne 을 개발하였다. NGSOne 의 구조는 Analysis Client Software, 클라우드 저장소 S3 와 3 개의 분석 파이프라인들을 제어하는 Control Server 로 구성되어 있고, Analysis Client Software 에 fastq 파일을 업로드하면 클라우드 저장소 S3 에 파일을 저장하고 선택한 파이프라인을 실행하는데 모든 작업의 처리는 Control Server 에서 제어한다. 3 개의 파이프라인인 DRAGEN[5], Isaac/Strelka2[6], BWA/GATK[7]의 비교에서 총 SNP (Single Nucleotide Polymorphism)/Indel (Insert and Deletion) 에러의 수에서 약간의 차이를 보였지만 분석 시간은 1.4, 2.4, 30 시간으로 많은 차이를 보였다.

II. 관련 연구

다양한 알고리즘 기반의 체세포 변이 탐색기 (Germline Variants Caller)[8][9]들이 새롭게 개발되고 있고 거짓 양성 오류 (false positive)를 최대한 줄이면서 민감도를 높이는 방향으로 발전하고 있다. 최근에는 낮은 품질 (low quality), 적은 뎁스 (low depth) 시퀀싱에서도 높은 정확도를 보이는 LoFreq[10]와 여러 도구들을 결합[11]하거나 변이 (variants) 필터링 방법을 고려하여 거짓 양성 오류를 줄이는 방법뿐만[12] 아니라 Deep Learning 기반의 Convolutional Neural Network(CNN) 알고리즘을 적용한 DeepVariant[13]도 소개되고 있지만, 복합적인 샘플 전처리 과정과 Indel realignment 를 도입하여 아직도 GATK 는 가장 대표적인 Variant Caller 이다.

2.1 GATK

Genome Analysis Tool Kit (GATK)는 Broad institute 에서 개발한 툴로 지놈 시퀀스내 모든 변이의 탐색(calling)이 가능하도록 몇 가지의 프로그램을 내장하고 있다. GATK 는 생식세포(Germline)와 체세포(somatic) 데이터 모두 분석이 가능하며 전장 유전체 시퀀싱(WGS, Whole Genome Sequencing) 뿐만 아니라 bed 파일을 이용하여 엑솜 시퀀싱(WES, Whole Exome Sequencing) 탐색도 가능하다. Non-GATK 툴을 이용하여

서열정렬 작업을 수행하고 난 이후에 GATK 툴을 이용하여 탐색과 필터링 (Filtering)을 작업을 수행한다. GATK 의 분석 파이프라인은 모범 사례(Best Practices)를 제공하며 변이탐색 (Variant discovery), genotyping 그리고 필터링을 제공하며 다양한 보정 작업(Local Realignment/ Base Quality Recalibration)을 통해 변이의 정확도를 높인다.

2.2 DRAGEN

일루미나사에 인수된 Edico Genome 사의 DRAGEN Complete Suite[14]는 NGS 데이터를 분석하기 위한 포괄적인 파이프 라인 패키지를 제공하는 하드웨어 기반의 응용 프로그램이다. 하드웨어의 성능을 이용하기 위해서 GATK 프로그램 코드를 변형하여 FPGA 칩에 넣어서 BCL conversion, compression, mapping, alignment, sorting, duplicate marking 및 haplotype variant calling 의 모든 과정이 1~2 시간 이내에 끝난다. 하나의 DRAGEN FPGA 는 23 대의 컴퓨터 instances 를 대체할 수 있으며 클라우드에서 1 개의 샘플을 분석할 수 있는 F1 instance 와 동시에 8 샘플을 분석할 수 있는 F16 의 instances 를 제공하고 있다. 생식세포, 체세포, 리보핵산 (RNA), 유전자 복제수 변이 (CNV, Copy Number Variation), Virtual Long Read Detection (VLRD)과 GATK 4 버전을 적용한 모범사례 파이프라인을 제공하고 있다.

2.3 Isaac/Strelka2

Isaac/Strelka2 은 일루미나사에서 개발한 2 개의 툴로써, 서열 정리를 하기해서는 Isaac aligner4 를 사용하고, 변이 탐색을 위해서는 Strelka2 를 이용한 파이프라인을 많이 구축하며 일반적으로 사용하고 있는 BWA/GATK 파이프라인보다 4~5 배[4] 빠르다고 소개되고 있다.

Isaac/Strelka2 은 Isaac Dominant template detection, Shadow rescue 등의 기술을 적용하여 아주 빠른 서열 정리 알고리즘을 구현하였고, Strelka2 caller 의 경우 tiered haplotype model 와 read-backed phasing 을 제공하고, DNA 삽입 및 결실의 에러를 줄이기 위해서 mixture-model indel error estimation 방법을 적용하여 빠르면서 정확한 변이 탐색 결과를 제공한다. Isaac/Strelka2 는 암/정상의 체세포 변이와 작은 코호트 (cohorts)의 생식 세포 변이 분석에 최적화 되었다.

2.4 NGS 데이터 분석을 위한 클라우드 컴퓨팅

대용량의 NGS 데이터 분석을 위해서 많은 자금을 투입하여 로컬에 서버와 저장소를 구축하기보다는 필요한 자원을 목적에 맞게 사용할 수 있는 클라우드 서비스가 활성화 되고 있으며, 분석 전문 회사인 DNAnexus, SevenBridges 와 Life Technology 과 일루미나에서 운영하는 BaseSpace 와 같은 상업적인 서비스와 STORMseq[15], GenomeKey[16], Globus[17]와 같이 클라우드에서 서비스를 제공하는 클라우드 서비스와 Galaxy, Taverna, COSMOS[18] 와 같이 전문적인 프로그래머 없이 생물학자가 쉽게 사용할 수 있는 GUI 환경의 워크플로를 제공해주는 워크플로 서비스가 있다. 상업적인 분석 서비스 회사들은 WGS 한 샘플 분석에 10 만원 이상의 비싼 분석 비용과 오랜 분석 시간이 소요되는 것이 일반적이고, 클라우드 제공 서비스[19]들은 47 달러에서 121 달러까지 다양한 서비스가 있지만 하나의 툴만을 제공하고 있고, NGSOne 보다는 2 배 이상의 높은 비용과 사용자가 AWS 이미지를 직접 세팅하여 파일을 스토리지에 저장 해야하는 많은 노력이 필요하다. Galaxy 와 같은 워크플로 서비스들은 편리한 GUI 환경이 제공되지만, 많은 사용자들이 접속하여 사용하기 때문에 30x WGS 파일 전송 시간만 2 일씩 소요될 경우가 많으며 분석 시간은 NGSOne 의 GATK 가 30 시간에 비해서 일주일씩 소요되기도 한다.

III. NGSOne

3.1 시스템 구성

본 시스템은 리눅스, 윈도우, 맥 등 대부분의 운영체제 환경에 설치 가능한 Analysis Client Software 와 클라우드 환경에서 중간 결과 및 최종 결과 파일들을 저장하는 클라우드 저장소 S3 와 3 개의 분석 파이프라인 들을 제어하는 Control Server 로 그림 1 과 같이 구성되어 있다.

그림 2 에서 확인할 수 있듯이 Analysis Client Software 는 클라우드 저장소 S3 에 파일을 업로드하는 업로드 모듈, 분석 진행 상황을 모니터링 및 관리할 수 있는 모니터링 모듈, 분석된 VCF (Variant Call Format)파일과 디렉토리를 관리할 수 있는 디렉토리 모듈로 구성된다. 랭귀지는 대부분의 운영체제와의 호환성을 위해서 오픈 소스 프레임워크 Electron 과 NodeJS 로 구현하였고, 컴파일을 수행하여 각각의 플랫폼에 최적화하여 배포하는 형태로 구성되어 있다. 파일 업로드 모듈은 미국 버지니아 클라우드 저장소 S3 region 이 최종 저장소이기 때문에 버지니아 region 외의 모든 나라와 지역은 반드시 한번의 경유를 거쳐서 버지니아 저장소에 저장되기 때문에 국가간 최고 전송속도 제한을 뛰어넘는 최상의 전송 속도를 낼 수 있다. 분석의 진행 정도를 확인 할 수 있는 모니터링 모듈은 파이프라인의 분석 진행 정도를 확인할 수 있다. 최종 분석된 VCF 파일은 디렉토리 모듈에서 확인 가능하고 개인 저장소에 저장되어서 언제든지 사용자가 다운 받을 수 있다.

Control Server 는 클라우드 환경의 실시간 미러링을 이용하여 C5.2xlarge 2 대의 컴퓨터가 동시에 운영되다가 운영 Control Sever 에 장애가 발생하면 Backup Server 로 모든 제어권이 넘어가는 구조로 구성되어 있다.

분석 인스턴스의 경우, DRAGEN F1 인스턴스는 아마존 클라우드에 구현되어 배포되어 있고, Isaac/Strelka2 이미지 인스턴스는 Isaac alignment 프로그램과 Strelka2 Calling 프로그램을 순서대로 설치하여 자체 제작하여 미리 만들어 놓았으며, GATK 의 모든 분석 파이프라인을 설치하여 미리 만들어 놓은 BWA/GATK 이미지 인스턴스를 만들었으며, 각각의 파이프라인에 가장 적합한 하드웨어 인스턴스를 선정하여 수많은 테스트를 수행하여 구성하였다. DRAGEN 은 Edico Genome 회사와 아마존에서 수많은 테스트를 수행한 이후에 F1.2xlarge 인스턴스에 구현되어 있고, Isaac/Strelka2 는 C5.9xlarege (cpu 32 코어, memory 73 기가) 인스턴스에 구현되어 있고, BWA/GATK 도 C5.9xlarge 인스턴스에 구현되어 있다. 로그 저장 데이터 베이스는 MariaDB 를 이용하여 수많은 로그 정보를 저장하고 있고, Control Server 와 Backup Server 와는 분리되었고, ShareFile 의 저장소에 최종 VCF 를 저장한다.

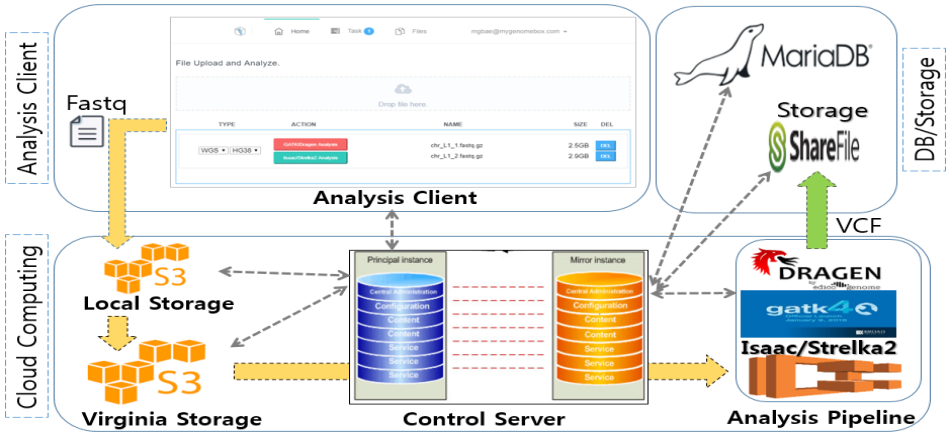

Figure 1. Overall system configuration diagram

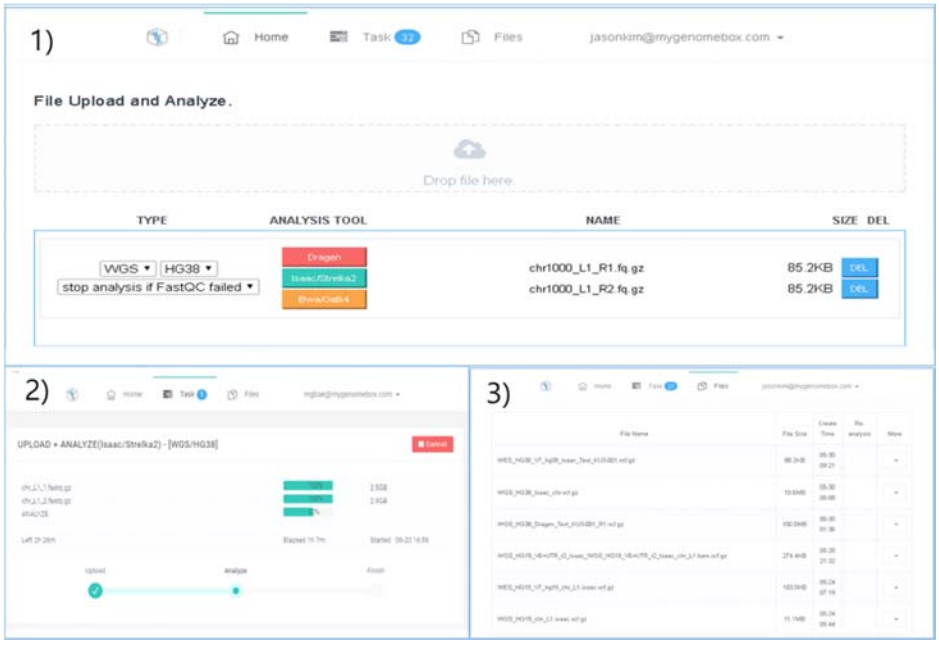

Figure 2. Analysis Client Software 1)Upload module 2)Monitoring module 3) Directory module

3.2 시스템 기능

Analysis Client Software 는 리눅스, 맥, 윈도우 3 개 버전이 제공되기 때문에 적합한 버전을 설치하여 실행 가능하고, 새로운 버전이 나왔을 때 자동 업데이트 기능을 제공한다. 클라이언트는 클라우드 저장소 S3 에 파일을 업로드하는 기능, 분석을 모니터링 및 관리할 수 있는 기능, 분석된 VCF 파일과 디렉토리를 관리할 수 있는 기능이 구현되어 있다. 업로드 모듈은 대용량/대량의 fastq 파일의 전송을 위해서 AWS SDK 에서 제공되는 전송 시작, 중지, 재개 등의 검증된 API 를 사용하기 때문에 타임지연이나 에러가 거의 발생하지 않는다. 처음 로그인이 되면 Control Server 에서 start token 을 받고 ping test 를 수행하여 가장 가까운 클라우드 저장소 S3 region 을 찾아서 region 정보를 저장한다. 대용량 fastq 나 gz 파일의 업로드를 시작하면 ping test 로 확인한 가장 가까운 클라우드 저장소 S3 region 에 fastq 파일을 전송하고 그 이후에 미국 버지니아 클라우드 저장소 S3 region 으로 파일을 전송한다. 버지니아 클라우드 저장소 S3 로 파일을 모으는 이유는 DRAGEN 인스턴스가 버지니아에 있기 때문에 나머지 인스턴스도 한 region 에서 관리하려는 이유이다. 한국에서 미국이나 유럽의 대륙간의 파일을 전송할 때나 유럽에서 미국으로 파일을 전송할 때 11MB 이상의 전송속도가 나오기 힘들지만 한국의 클라우드 저장소 S3 로 전송하면 30MB, 한국 S3 저장소에서 미국 버지니아 S3 저장소까지는 100MB 이상의 속도로 전송하기 때문에 최소 3 배 이상의 전송 속도 향상이 있다. fastq 파일이 버지니아 S3 에 전송이 완료되면 Control Server 는 이미지 분석 인스턴스 (DRAGEN, Isaac/Strelka2, BWA/GATK Image instance)를 띄우고, 파일을 분석 인스턴스로 전송한다. 분석 파이프라인을 Docker 형태로 미리 제작해 놓은 이미지 인스턴스에서 fastq 파일을 이용하여 분석을 시작한다. 사용자가 선택한 파이프라인이 실행되는데 QC 모듈, alignment 모듈, Variant Calling 모듈, 및 Filtering 모듈의 순서로 실행되지만 각각의 툴은 표 1 과 같이 다르다.

CPU 자원을 많이 사용하는 프로그램과 메모리를 많이 필요로 하는 프로그램이 다르기 때문에 용도에 맞게 instance 를 선택하여 최적화된 파이프라인이 구축되었다. 분석 단계

단계별 모든 로그는 Control Server 에서 체크하여 시스템 자원의 문제나 에러가 발생하지 않으면 분석을 완료하여 VCF 파일을 S3 저장소와 ShareFile 공유 저장소로 전송한다.

Table 1. Tools implemented in the pipeline

Pipeline	QC	Alignment	Variant Calling	Filtering	Combined
DRAGEN	Self qc	Modified GATK	Modified GATK	Modified GATK	No Shell
Isaac/Strelka2	FastQC	Isaac4	Strelka2	Strelka2	Shell
BWA/GATK	FastQC	BWA, Samtools	GATK	GATK	Shell

Control Server 는 실시간 미러링을 이용하여 2 대의 컴퓨터가 동시에 운영되다가 운영 Control Sever 에 장애가 발생하면 Backup Server 로 모든 제어권이 넘어가는 구조로 구성되어 있다. 로긴된 모든 Client software 의 접속 관리, 분석 인스턴스의 start, stop, destroy 등의 모든 관리를 제어하며, 분석이 끝난 VCF 파일을 ShareFile 공유 저장소를 전송하는 역할 뿐만 아니라 MariaDB 데이터 베이스 서버로 모든 접속 이력, 분석 이력을 전송하는 역할도 수행한다.

IV. 결 과

표 2 에서 볼 수 있듯이, BWA/GATK 의 총 분석 시간은 32 개 cpu 코어와 72 기가 메모리를 가진 인스턴스에서 평균 30 시간이었지만, 동일 사양의 Isaac/Strelka2 는 약 2 시간 25 분으로 12 배 빠른 성능을 보여주었다. 또한, 환경 세팅의 경우에도 BWA/GATK 는 여러 분석 단계를 거치며 많은 에러가 났지만, Isaac/Strelka2 는 분석과 변이 탐색의 2 단계이고 에러도 적었으며 세팅 과정도 어렵지 않았다. DRAGEN 은 매우 빠른 편인 대략 1 시간 25 분의 분석 시간이 소요 되었는데, 컴퓨터를 켜고 준비하는 (Instance Prepare) 시간과 최종 복사하는 시간을 빼면 1 시간 10 분내에 모든 분석을 완료하였다. 다만, DRAGEN 의 홈페이지에서 광고되는 전장 유전체 40 분 분석 시간은 대형 서버를 구매하였을 경우에 가능하기 때문에 클라우드상에서 구현된 DRAGEN 은 최적의 분석시간임을 확인하였다.

Table 2. Analysis time of 3 pipelines

Pipeline	Instance Prepare (분)		QC (분)		Analysis (분)		Total Run Time (시간)	
	NA12878	NA12891	NA12878	NA12891	NA12878	NA12891	NA12878	NA12891
BWA/GATK	10	11	37	33	1933	1606	33	27.5
Isaac/Strelka2	10	10	35	31	115	91	2.6	2.2
DRAGEN	11	10	5	5	78	58	1.6	1.2

* 1000Genome Project NA12878, NA12891 샘플에서 32X, 24X 를 랜덤 추출함

그림 3 에서 확인할 수 있듯이 위 음성(False Negative) 및 위 양성(False positive) 전체 에러 개수는 DRAGEN 과 BWA/GATK 가 거의 비슷했지만, 두 샘플의 분석 시간에서는 약 25 배정도의 차이가 난다. Isaac/Strelka2 는 모든 샘플에서 DRAGEN 보다는 월등히 많은 에러를 가지기 때문에 여러가지 필터링 방법으로 거짓을 제거하는 것이 중요할 것으로 생각된다. 종합해보면 DRAGEN 이 실행 시간 및 에러의 개수 면에서 가장 좋은 성능을

보여주는 것을 확인할 수 있었다. 단, 모든 분석 결과는 1000Genome Project High Coverage (81X 이상) 샘플의 표준 결과를 정답셋으로 잡았고, 로컬 컴퓨터 결과와는 동일하고 NGSOne 의 구현에서 추가적인 에러가 발생하지는 않았다.

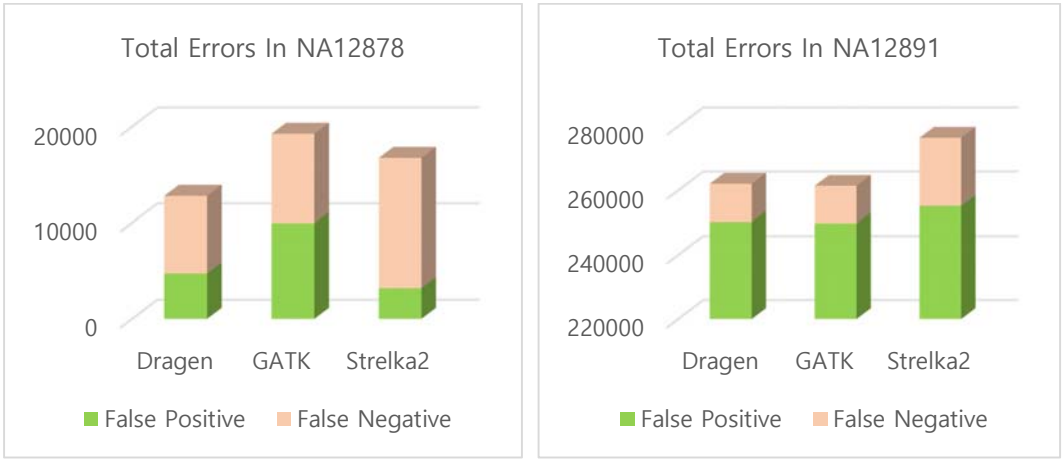

Figure 3. Number of total errors in SNP and Indel *1000Genome Project NA12878, NA12891 샘플에서 32X, 24X를 랜덤 추출함

V. 결론 및 향후 연구

본 논문에서는 WGS 와 WES 분석에서 3 개의 대표적인 파이프라인인 DRAGEN, Isaac/Strelka2 및 BWA/GATK 를 자동 분석할 수 있는 툴인 NGSOne 을 제안하고 있다. NGSOne 은 생물학자와 임상전문가들이 클라이언트 툴을 설치만 하면 모든 분석을 바로 진행할 수 있는 설치 프로그램이기 때문에 dbSNP 과 Human Genome Version 의 옵션을 자유롭게 변경하여 운영이 가능하다.

클라우드 기반으로 구축되어서 업로드 된 상태에서 동시에 수백 개 이상의 샘플도 분석도 가능하며 자체 서버를 구축하고 있지 않아도 NGSOne 툴만 가지고 VCF 추출이 가능하다. 향후에 Docker 기반의 다양한 툴들과의 연계를 추진하여 VCF 로 추출된 변이 파일의 주석 (annotation)을 달거나 한꺼번에 기능 분석이 가능한 시스템을 구축할 예정이고, Sequencing.com[20]이나 MyGenomeBox[21]로 VCF 파일이 자동 전송되어서 체세포 변이의 해석과 수많은 기능을 가진 앱들, 예를 들면 대머리 확인, 콜레스테롤 확인, 겨드랑이 체취, 아침형 인간 등을 활용할 수 있는 플랫폼과 연결하는 작업을 할 계획이다. 최종적으로 클라우드 구축 기술을 활용하여 WGS 와 WES 의 추가적인 툴들뿐만 아니라 RNA-Seq 과 Methyl-Seq 분석에 관련된 대표적인 툴들도 삽입하여 다양한 앱 플랫폼과 기능 해석이 자유로운 Docker 플랫폼들과 통합할 예정이며, NGS 분석을 하나의 툴에서 가능하도록 NGSOne 을 확장할 예정이다.

VI. 이용 경로

http://dockerbio.mygenomebox.com/html/ngsone/index.html

VII. 참고문헌

[1] The $1,000 Genome, https://www.illumina.com/company/news-center/feature-articles/the-1000-dollar-genome.html
[2] November J, "More than Moore's Mores: Computers, Genomics, and the Embrace of Innovation," *J Hist Biol*., Aug. 2018.
[3] Neil A. Miller, Emily G. Farrow, Margaret Gibson, Laurel K. Willig, Greyson Twist, Byunggil Yoo, Tyler Marrs, Shane Corder, Lisa Krivohlavek, Adam Walter, Josh E. Petrikin, Carol J. Saunders, Isabelle Thiffault, Sarah E. Soden, Laurie D. Smith, Darrell L. Dinwiddie, Suzanne Herd, Julie A. Cakici, Severine Catreux, Mike Ruehle and Stephen F. Kingsmore, "A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases," *Genome Medicine*, **7**:100, Sep. 2015.
[4] Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT, "Strelka2: fast and accurate calling of germline and somatic variants," *Nat Methods*., 15(8):591-594, Aug. 2018.
[5] Amit Goyal, Hyuk Jung Kwon, Kichan Lee, Reena Garg, Seon Young Yun, Yoon Hee Kim, Sunghoon Lee, Min Seob Lee, "Ultra-Fast Next Generation Human Genome Sequencing Data Processing Using DRAGEN™ Bio-IT Processor for Precision Medicine," *Open Journal of Genetics*, Vol.07 No.01, Mar. 2017.
[6] Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, Chuang HY, Källberg M, Kumar SA, Liao A, Little KM, Strömberg MP, Tanner SW., "Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms," *Bioinformatics*, 15;29(16):2041-3, Aug. 2013.
[7] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, **"**The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, 20:1297-303, Sep. 2010.
[8] Garrison, E. & Marth, G., "Haplotype-based variant detection from short-read sequencing," *arXiv preprint*, ArXiv:1207.3907 [q-bio.GN], Jul. 2012.
[9] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G., "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Nat Biotechnol*., 31(3):213-9, Mar. 2013.
[10] Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N., "LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets," *Nucleic Acids Res.*, 40(22):11189-201, Dec. 2012.
[11] Luo R, Schatz MC, Salzberg SL, "16GT: a fast and sensitive variant caller using a 16-genotype probabilistic model," *Gigascience*., 1;6(7):1-4, Jul. 2017.
[12] Field MA, Cho V, Andrews TD, Goodnow CC, "Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies," *PLoS One.*, 23;10(11):e0143199, Nov. 2015.
[13] Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA, "A universal SNP and small-indel variant caller using deep neural networks," *Nat Biotechnol*., 24. doi: 10.1038/nbt.4235, Sep. 2018.
[14] DRAGEN, http://edicogenome.com/dragen-bioit-platform/
[15] Konrad J. Karczewski, Guy Haskin Fernald, Alicia R. Martin, Michael Snyder, Nicholas P. Tatonetti, Joel T. Dudley, "STORMSeq: An Open-Source, User-Friendly Pipeline for Processing Personal Genomics Data in the Cloud," *PLoS One*., 2014; 9(1): e84860., Jan. 2014.
[16] Yassine Souilmi, Alex K. Lancaster, Jae-Yoon Jung, Ettore Rizzo, Jared B. Hawkins, Ryan Powles, Saaïd Amzazi, Hassan Ghazal, Peter J. Tonellato, Dennis P. Wall, "Scalable and cost-effective NGS genotyping in the cloud," *BMC Med Genomics*., 2015; 8: 64., Oct. 2015.
[17] Krithika Bhuvaneshwar, Dinanath Sulakhe, Robinder Gauba, Alex Rodriguez, Ravi Madduri, Utpal Dave, Lukasz Lacinski, Ian Foster, Yuriy Gusev, Subha Madhavan, "A case study for cloud based high throughput analysis of NGS data using the globus genomics system," *Comput Struct Biotechnol J.,* 2015; 13: 64–74., Nov. 2014.
[18] Erik Gafni, Lovelace J. Luquette, Alex K. Lancaster, Jared B. Hawkins, Jae-Yoon Jung, Yassine Souilmi, Dennis P. Wall, Peter J. Tonellato, "COSMOS: Python library for massively parallel

workflows," *Bioinformatics.,* 2014 Oct 15; 30(20): 2956–2958., Jun. 2014.
[19] Wang Y, Li G, Ma M, He F, Song Z, Zhang W, Wu C., "GT-WGS: an efficient and economic tool for large-scale WGS analyses based on the AWS cloud service," *BMC Genomics.,* 2018 Jan 19;19(Suppl 1):959., Jan. 2018.
[20] Sequencing.com, https://sequencing.com/
[21] MyGenomeBox, https://www.mygenomebox.com/