Introducing Google Genomics

SeungWoo Lee

1. What is Google Genomics?

1. 옛날에는 컴퓨팅 파워의 제한으로 인하여 한 사람, 한 유전자, 혹은 한 특성들만 볼 수 있었다.
2. 그러나 오늘은 DNA 시퀀싱 기술의 발달로 옛날의 데이터량을 뛰어넘었고.
3. 유전 데이터는 더 많이 늘어나고 있다
4. 거기에 더 좋은, 더 저렴한 툴들로 더 많은 환자 데이터를 볼 수 있게 되었다.
5. 수많은 유전체 데이터를 처리할 수 있고, 수많은 유전체 데이터를 핸들링할 수 있는 하나의 플랫폼을 구글이 만든 것이다.

2. More Information...

구글 지노믹스를 사용한다면,

1. 더 빠른 시일내로 결과를 볼 수 있습니다.
2. 더 넓게 프로젝트의 범위를 늘릴 수 있고.
3. 환자 데이터를 보호할 수 있다.
4. Global Alliance for Genomics & Health 에 가입하여 이들의 형식에 따른 데이터 포멧을 사용하고 있다.

Global Alliance For Genomics & Health

1. IBM, HITACHI, Illumina, Intel 등등 25개국 177개의 회원사들이 가입함
2. 유전체 데이터와 health 에 있어서 어떻게 데이터를 운용하고 규제에 대해 논의하는 곳으로 규제 및 윤리, 데이터, 보안, Clinical 의 4개의 워킹그룹으로 구성되어 있음.
3. 이중 Data Working Group은 개인 유전체 데이터 포맷과 이를 교환하고 사용하기 위한 API 에 대한 총괄을 담당하고 있음.

3. Pricing.

- Pricing
  - GB 당 달 요금
  - 그 예로 30배의 한 사람의 데이터를 시퀀싱한 데이터를 사용한다면, 25달러 정도가 듬
  - 유닛당 요금은 0.022 달러입니다
  - (그러나 한달은 300달러 의 혜택을 주고 있습니다)
- Free Quotas (무료 한도)
  - 무료 한도는 일/10만번의 API 콜까지 허용

Let’s Getting Started!

Querying A 1000 Genomes Project!

4. 1000 Genomes Project

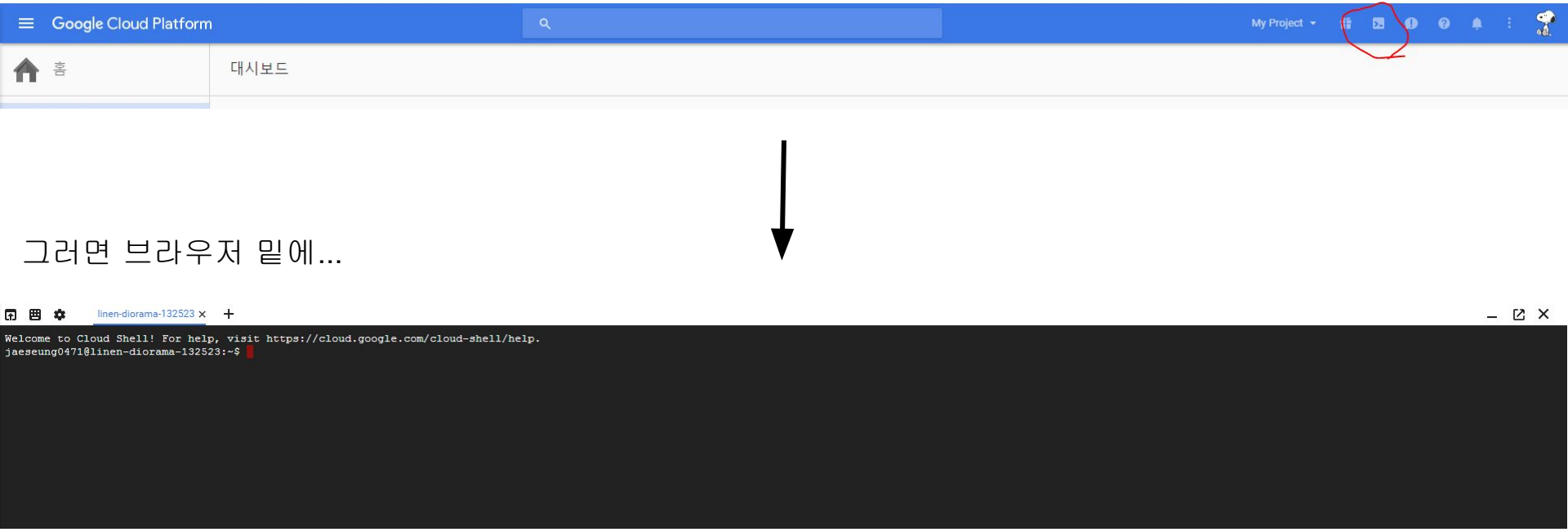
Accoridng from Wikipedia,

1. 미국 영국 중국 이 합작
2. 3년 내에 다양한 변종으로 구성된 인간 1000명의 게놈 데이터를 얻는것이 핵심.
3. 번이체학의 기초를 이룰 수 있는 매우 귀중한 자료를 만들겠다는 목적에 있음.

More Information, Please Visit http://www.1000genomes.org/

5. Quickstart

1. Google Compute Engine 에 가입 (카드 번호가 필요)
2. Google Cloud Platform 페이지에 들어가서, Console 마크 클릭.
3. Console 이 뜸.


Google Cloud Platform
My Project
홈
대시보드
그러면 브라우저 밑에...
linen-diorama-132523
Welcome to Cloud Shell! For help, visit https://cloud.google.com/cloud-shell/help.
jaeseung0471@linen-diorama-132523:~$

Let’s Do it!

- gcloud alpha genomics 라는 문구를 처음에 붙여야 함.
  - Querying Specific location of Varients,
    - Using variants Function.
    - For Example - gcloud alpha genomics variants list --variant-set-id "10473108253681171589" --reference-name "22" --start 51003835 --end 51003836
  - Let’s Query 10 Callsets From 1000 Genomes (10473108253681171589 is set name of 1000 genomes)
    - For Example - gcloud alpha genomics callsets list "10473108253681171589" --limit 10
  - Let’s Take the Details(JSON Format) from First Querying location.
    - gcloud alpha genomics variants list --variant-set-id "10473108253681171589" --reference-name "22" --start 51003835 --end 51003836 --format json

And...

- Let’s take group sets.
  - gcloud alpha genomics readgroupsets list 10473108253681171589 --limit 5
  - Same Reference set id, 5 limits
- Let’s Read Specific Data.
  - gcloud alpha genomics readgroupsets describe CMvnhpKTFhDq9e2Yy9G-Bg
  - Return By Sam Format, but File is BAM (Binary, Because it’s Too Huge!)
    - SAM과 BAM은 모두 sequence를 저장하며, 같은 정보를 가지고있다.
    - SAM은 text file로, string 형식으로 저장되어 있기 때문에 바로 열람할 수 있다.
    - BAM은 Binary 형식이기 때문에 바로 열람할 수 없다. 하지만 압축되어 있기 때문에 용량이 작다.
    - BAM 파일은 SAM파일과 동일하지만, reference sequence names, length들이 헤더에 포함되어있다.

More! ...

- Let’s Get Chr20 (Chromosome 20th)
  - gcloud alpha genomics reads list "CJ_ppJ-WCxDxrtDr5fGIhBA" --reference-name "chr20" --start 68198 --end 69000 --limit 5
  - Also 5 Limits.