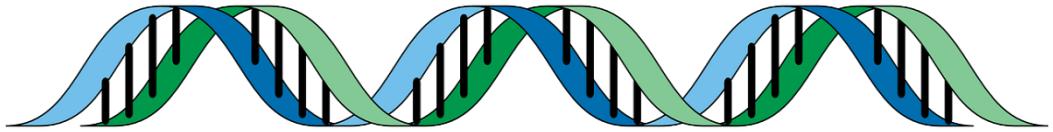
Google BigQuery를 이용한 유전자 데이터 분석



발표 : 조 익연

(Reference)

https://cloud.google.com/genomics/

https://cloud.google.com/genomics/v1/analyze-variants

https://github.com/googlegenomics/getting-started-bigquery

http://googlegenomics.readthedocs.io/en/latest

1. Google BigQuery

클라우드 서비스
- AWS(Amazon Web Service)
- Google Cloud Platform(GCP)
- Microsoft Azure

Google BigQuery : 구글 클라우드(GCP)의 **빅데이터 분석 플랫폼**

- 구글 빅쿼리 서비스는 SQL기반의 쿼리를 아주 큰 데이터 셋에서 수행해 주는 서비스임.

- 기존의 관계형 데이터베이스를 사용할 때처럼 데이터베이스 스키마를 지정해 주고 데이터를 로드(load)한 다음, 마치 데이터베이스를 다루듯이 쿼리문을 통해 사용자가 자신이 가진 데이터의 특성을 파악할 수 있음.

- 구글 빅쿼리 서비스는 드레멜(Dremel)이라는 구글 내부 서비스를 외부로 공개한 것.

- 맵리듀스 프레임워크는 대규모 데이터 기반에서 정해진 업무를 수행하는 배치(batch) 형태의 작업에는 아주 적합한 구조이지만, 애드혹(ad hoc) 스타일의 데이터 분석이나 시행착오법(trial error method) 같은 반복적인 계산이 발생하는 데이터 마이닝에는 취약함.

1. Google BigQuery

Google BigQuery 서비스의 특징

- **빠른 속도**: 수십억 건의 데이터를 수초 안에 계산해 냄

- **확장성**: 테라바이트급의 데이터를 처리할 수 있으며, 수백억 건의 레코를 수용함

- **단순성**: SQL 기반의 쿼리 언어를 지원하기 때문에 사용이 단순

- **공유 및 보안**: 그룹과 사용자 기반의 공유를 제어할 수 있고, SSL 커넥션을 통해 데이터 보안을 지원

- **다양한 접근방법**: 웹브라우저, 커맨드라인 툴, 그리고 REST API를 지원함으로써 다양한 방법으로 빅쿼리 서비스를 지원

1. Google BigQuery

BigQuery와 맵리듀스의 비교

차이점	빅쿼리	맵리듀스
특징	대규모 데이터셋용 쿼리 서비스	대규모 데이터셋 처리 프로그래밍 모델
주요 용도	트러블슈팅이나 빠른 분석을 위한 애드혹이나 시행착오 대화형 쿼리 분석	시간이 오래 걸리는 데이터 변환이나 합계 작용용 배치 작업
OLAP/BI용	적합함	적합하지 않음
데이터 마이닝용	부분적 적용 가능	적합함
응답시간	빠름	아주 느림
사용성	좋음(개발자가 아니어도 사용하기 편함)	좋지 않음(개발 언어와 프레임워크에 대한 이해가 필요)
로직	단순한 로직에 적합	복합적이고 어려운 로직을 개발하는 데 적합

1. Google BigQuery

BigQuery가 지원하지 않는 것들

- 테이블 색인이나 그 밖의 데이터베이스 관리 기능을 지원하지 않음
- 서브쿼리를 지원하지만 이를 통한 업데이트나 삭제는 지원하지 않음
- 조인을 지원하긴 하지만 한쪽 조인이 다른 쪽 조인보다 훨씬 작을 경우에만 지원 가능
- OLTP 기능 없음
- SQL 클라인언트 툴 대신 REST API로 접근 가능

1. Google BigQuery

빅쿼리를 이용하기 위해서는? Google Cloud Platform(또는 Google Genomics) 서비스 가입 필요!

https://cloud.google.com/genomics/

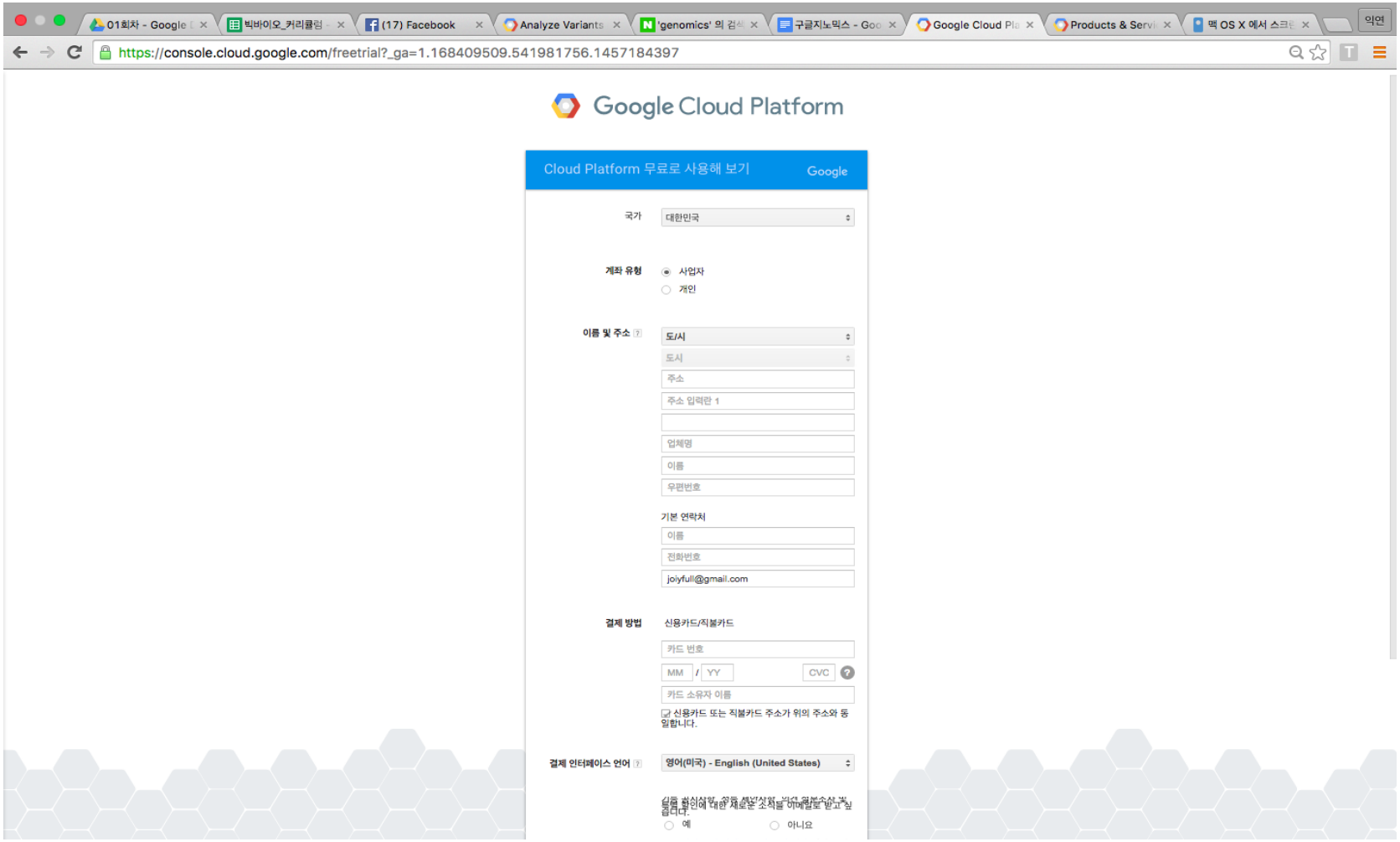
→ TRY IT FREE를 클릭
→ Google Cloud Platform 무료로 사용해 보기

무료 평가판 기간은 60일, 크레딧은 $300.00

Resource / Cost (in US$) = Genomics storage / $0.022/GB Per Month

1. Google BigQuery

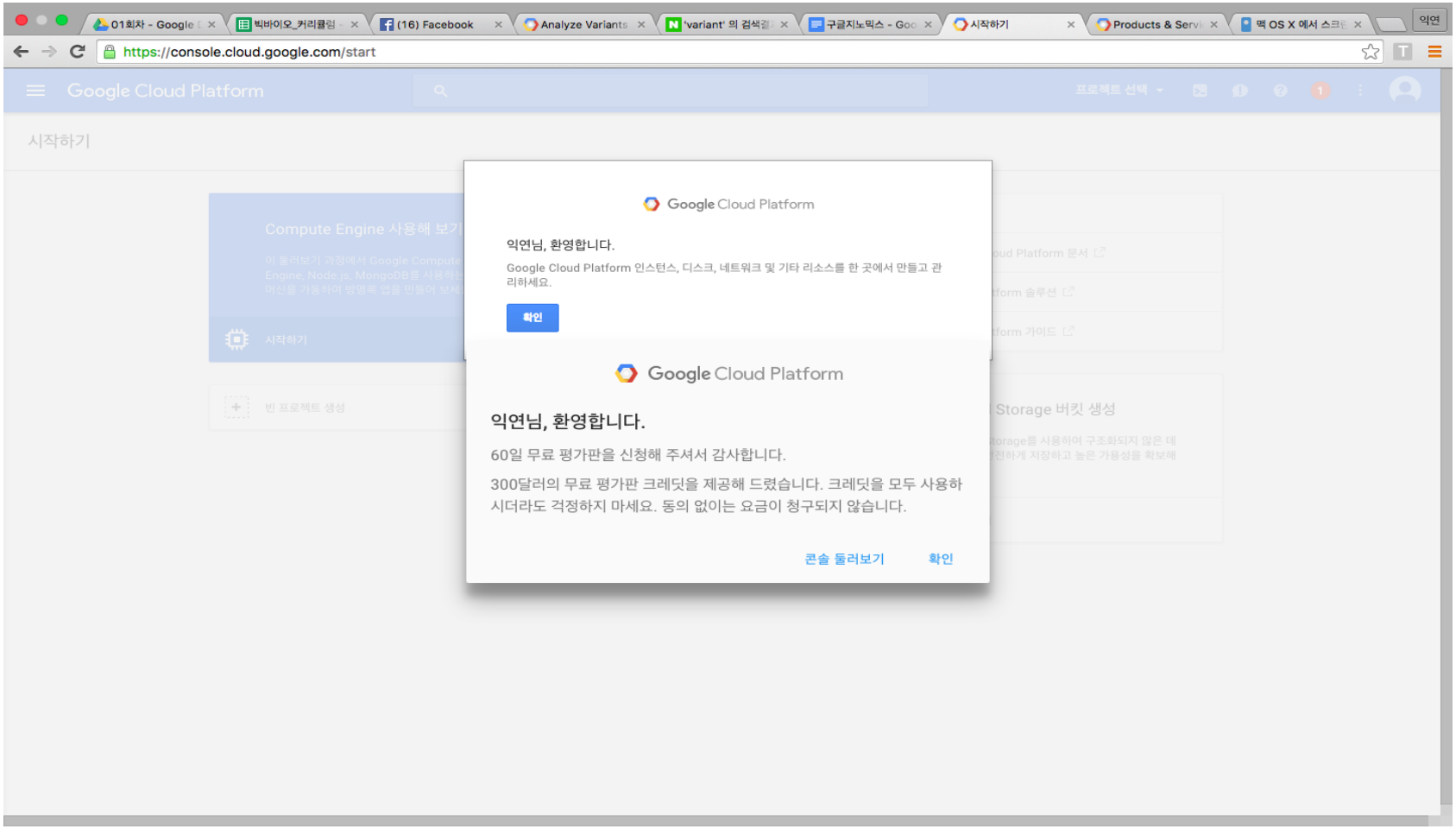
구글 클라우드 플랫폼 가입하기
- 이름, 주소 및 신용카드 정보 입력 필요



1. Google BigQuery

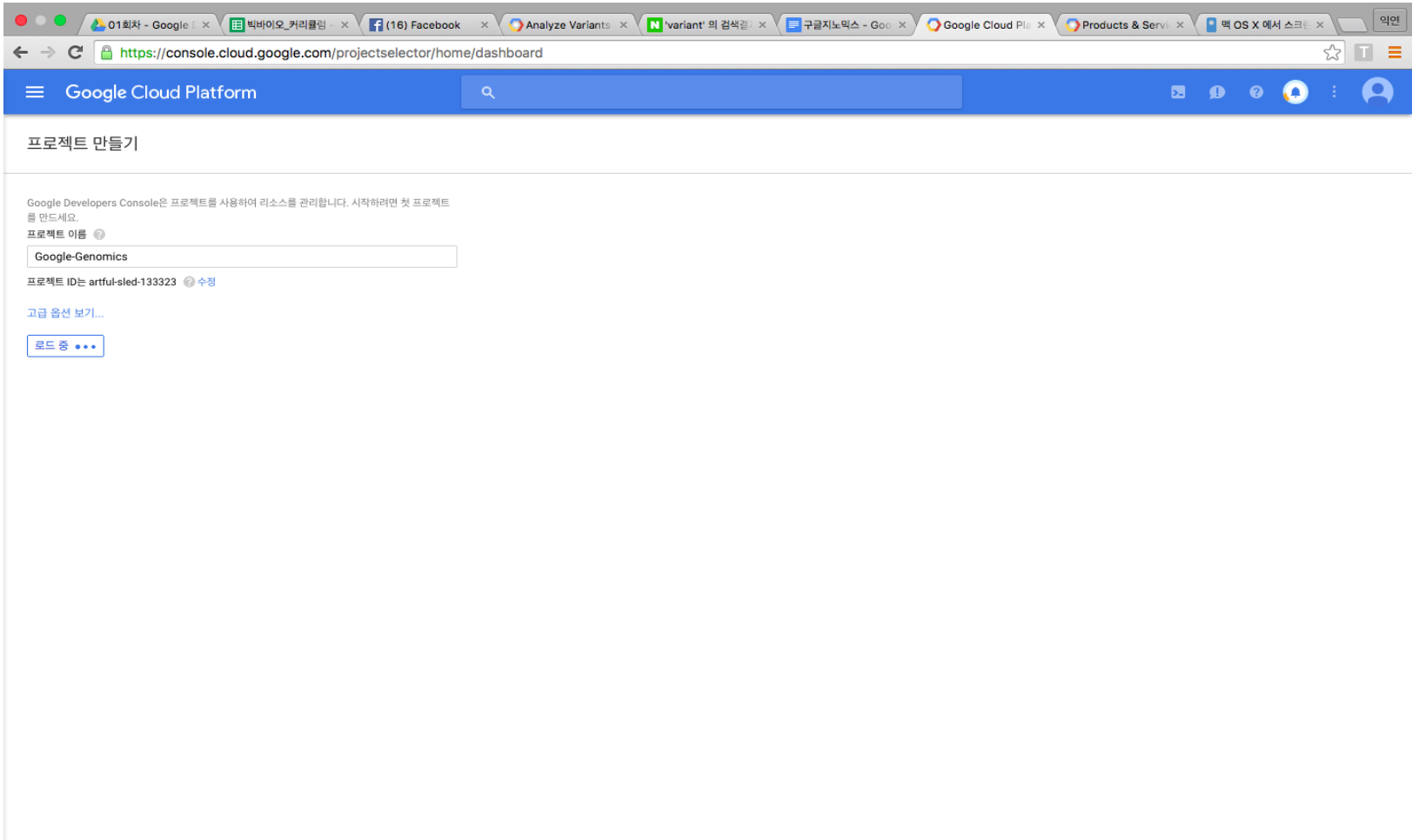
구글 클라우드 플랫폼 가입하기

- 60일 무료 평가판
- 300달러 크레딧 제공



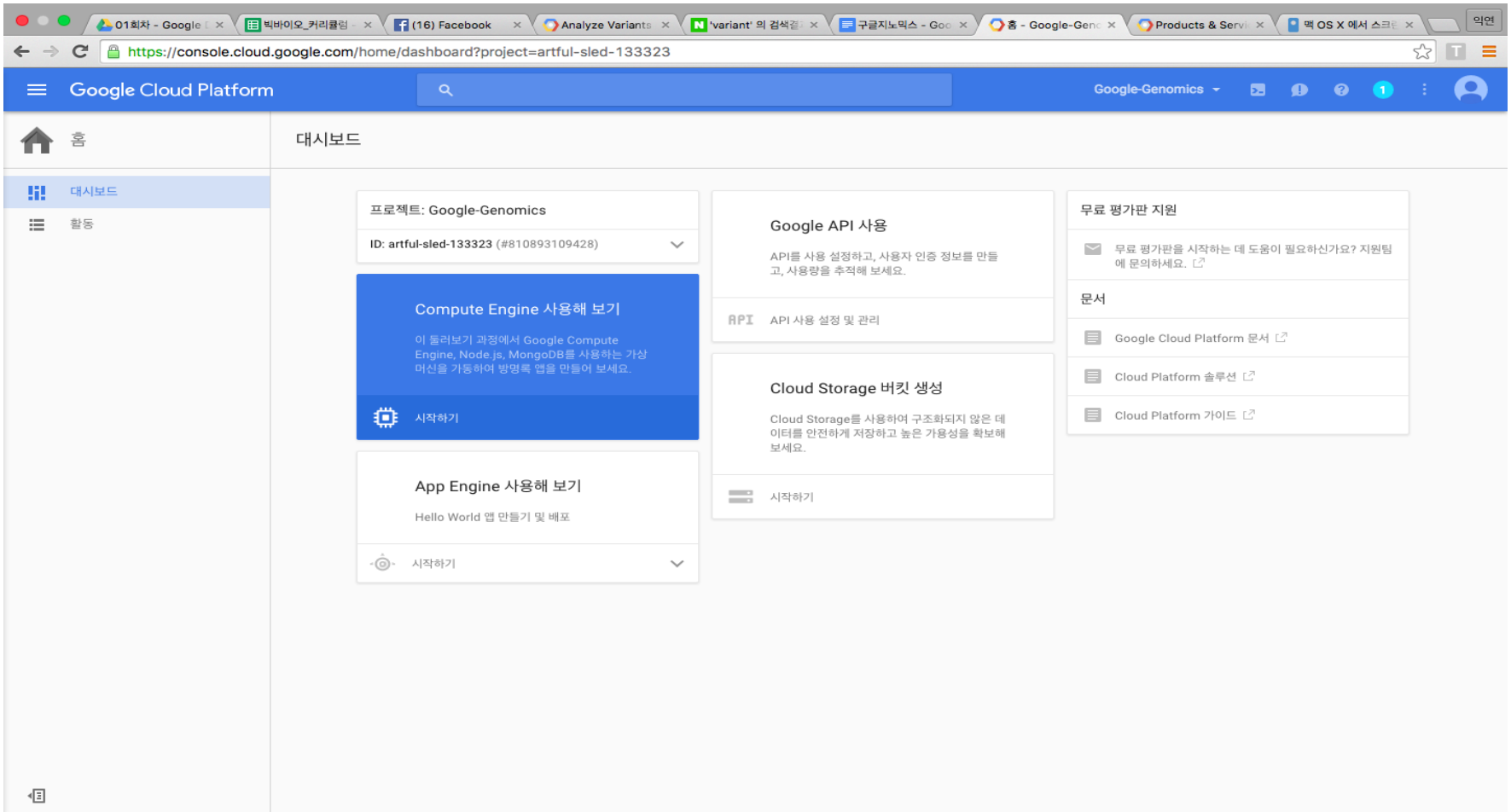
1. Google BigQuery

구글 클라우드 플랫폼 가입하기- 프로젝트 생성



1. Google BigQuery

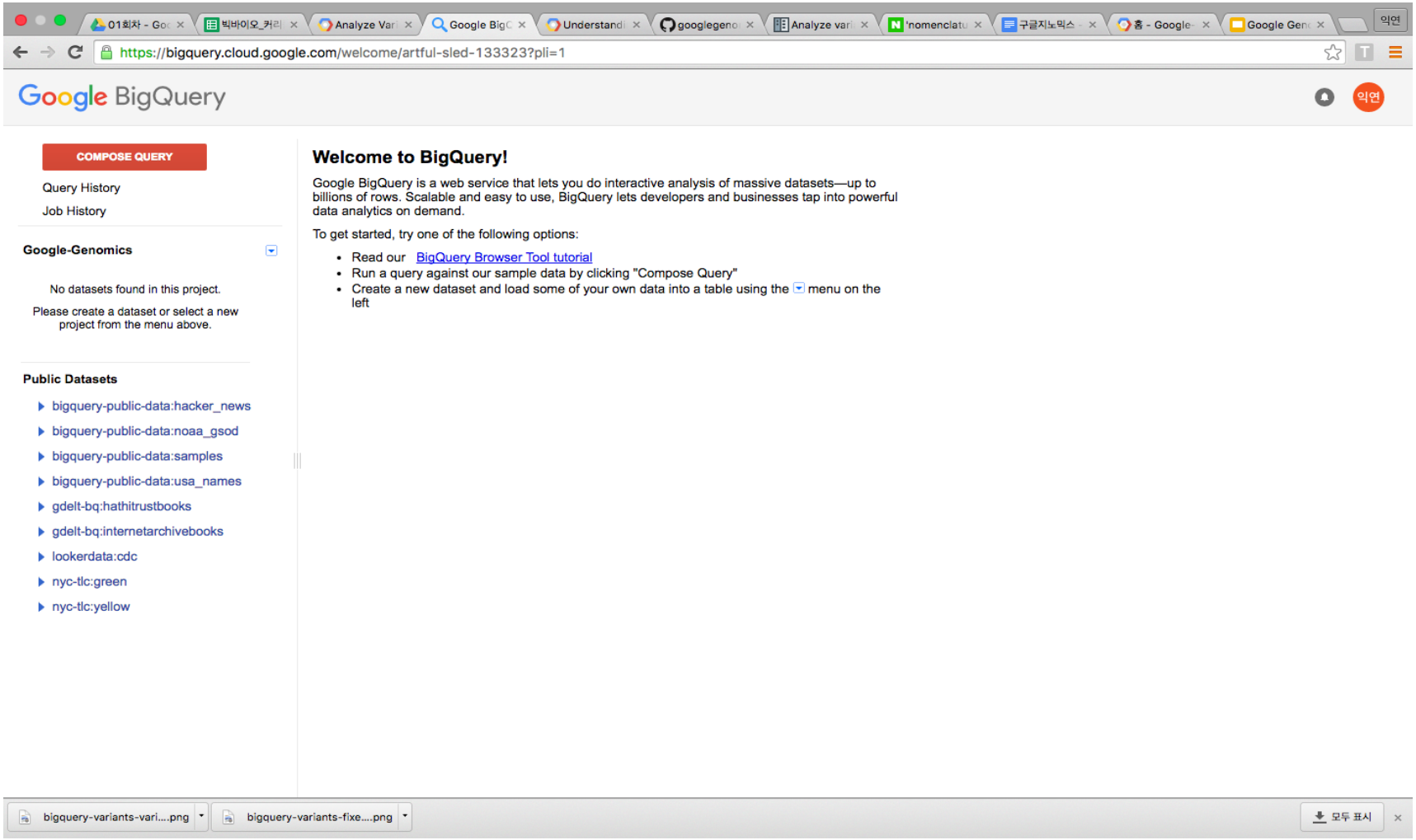
구글 클라우드 플랫폼 가입하기- 프로젝트 생성



1. Google BigQuery

(접근방법 1) BigQuery 브라우저 툴(Web Interface)

프로젝트에서 빅 데이터>BigQuery를 클릭 또는 빅쿼리 URL (https://bigquery.cloud.google.com)을 클릭



1. Google BigQuery

(접근방법 1) BigQuery 브라우저 툴(Web Interface)

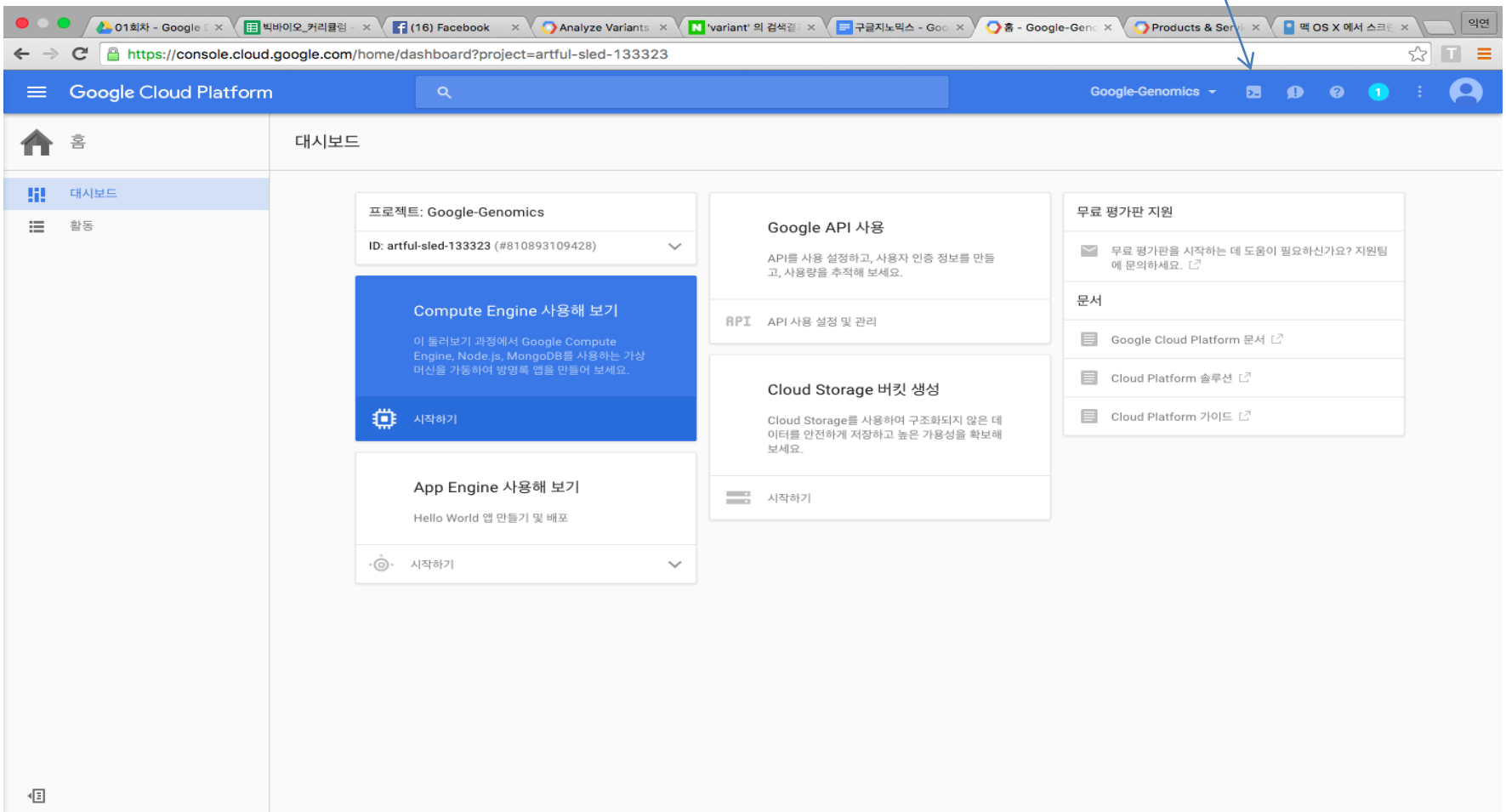
예제 : https://cloud.google.com/genomics/v1/analyze-variants#example

- Open the BigQuery web UI.
- Click on **"Compose Query"**.
- Write query into the dialog box and click on **"Run Query"**.

1. Google BigQuery

(접근방법 2) 커맨드라인 툴

"Google Cloud Shell 활성화" 클릭



2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

BigQuery, a tool that enables very fast SQL-like queries of massive data sets, **lets you interactively explore large datasets of population variants to find patterns** that shed light on disease correlation, epidemiology(역학), and more.

We have a number of resources to help you get started using BigQuery to Analyze Variants:

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

(1) 빅쿼리 Variants 테이블 구조의 이해(Understanding the BigQuery Variants Table Schema)

- https://cloud.google.com/genomics/v1/bigquery-variants-schema

Google Genomics provides an API which can be used to export variants to Google BigQuery. This will allow you to use the power of BigQuery to run ad-hoc interactive queries over genomic variants using hundreds or thousands of computers in parallel.

If you'd like to query your research data with BigQuery, you need to **first load the variants into Google Genomics and then export to BigQuery**.

You can also browse existing **published datasets already exported from Google Genomics to BigQuery.**

용어 살펴보기(Nomenclature)
- Genomics nomenclature
- BigQuery nomenclature

Variants 테이블의 구조(Variants table structure)

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

(1) 빅쿼리 Variants 테이블 구조의 이해

유전학 관련 용어들(Genomics nomenclature)

Sample: DNA collected and processed under a single identifier. A sample typically involves a single individual organism, but can also be a heterogeneous sample such as a cheek swab.

Reference Name: The name of a reference segment of DNA, this is typically a chromosome, but may be other named regions from a reference genome.

Variant: A region of the genome that has been identified as differing from the reference genome. A variant must have a reference name, start position, end position, and one or more reference bases. See documentation for the Variant resource for more details.

Non-variant segment: A region of the genome that matches the reference genome. This is sometimes referred to as a "reference segment". Traditionally genomic data has not included non-variant segments with variants. For more on non-variant segments see the gVCF documentation or the Complete Genomics masterVar documentation.

Call: An identified occurrence of a variant or non-variant segment for an individual sample. See documentation for the VariantCall resource for more details.

Call set: A group of calls from a single sample.

INFO fields: Optional fields added to Variant and Call information. For example, while all Calls will have a genotype field, not all datasets will have a "Genotype Quality" (GQ) field. Thus the genotype field is a fixed part of the VariantCall schema, but there is no GQ field. The "GQ" field and value can be imported as key/value pairs into the VariantCall info field.

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

(1) 빅쿼리 Variants 테이블 구조의 이해

빅쿼리 관련 용어들(BigQuery nomenclature)

Simple fields: Simple data elements in a BigQuery table, such as numbers and strings.

Nested fields: Complex data elements in a BigQuery table. A nested field can contain multiple fields, both simple and nested.

Repeated fields: Fields in a BigQuery table that can have multiple values, like a list. Repeated fields can be both simple and nested.

※ REPEATED fields for lists of values and **NESTED fields** for **hierarchical values.**

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

1) 빅쿼리 Variants 테이블 구조의 이해

테이블 구조(Variants table structure)

- 레코드 구조(Variants table record structure)
- 필드 구조(Variant table field structure)
- 고정 필드(Variant table fixed fields)
- 예제

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

1) 빅쿼리 Variants 테이블 구조의 이해

고정필드(Fixed fields)

Schema

reference_name	STRING	NULLABLE	An identifier from the reference genome or an angle-bracketed ID String pointing to a contig in the assembly file.
start	INTEGER	NULLABLE	The reference position, with the first base having position 0.
end	INTEGER	NULLABLE	INFO=<ID=END,Number=1,Type=Integer,Description="End position of the region described in this record">
reference_bases	STRING	NULLABLE	Each base must be one of A,C,G,T,N (case insensitive). Multiple bases are permitted. The value in the POS field refers to the position of the first base in the String.
alternate_bases	STRING	REPEATED	List of alternate non-reference alleles called on at least one of the samples. ("at least one" not true for this dataset)
quality	FLOAT	NULLABLE	phred-scaled quality score for the assertion made in ALT.
filter	STRING	REPEATED	PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise, if the site has not passed all filters, a list of codes for filters that fail.
names	STRING	REPEATED	List of unique identifiers for the variant where available.
call	RECORD	REPEATED	Per-sample measurements.
call.call_set_id	STRING	NULLABLE	The id of the callset from which this data was exported from the Google Genomics Variants API.
call.call_set_name	STRING	NULLABLE	Sample identifier.
call.genotype	INTEGER	REPEATED	List of genotypes.
call.phaseset	STRING	NULLABLE	If this value is null, the data is unphased. Otherwise it is phased.
call.genotype_likelihood	FLOAT	REPEATED	List of genotype likelihoods.

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

1) 빅쿼리 Variants 테이블 구조의 이해

가변 필드(variable fields (the INFO fields)):

call.AD	INTEGER	REPEATED	FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
call.DP	INTEGER	NULLABLE	##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
call.FILTER	STRING	REPEATED	PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise, if the site has not passed all filters, a list of codes for filters that fail - *copied* from variant level FILTER field since this dataset is merged single-sample VCF.
call.GQ	FLOAT	NULLABLE	FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
call.GQX	STRING	NULLABLE	FORMAT=<ID=GQX,Number=1,Type=Integer,Description="Minimum of {Genotype quality assuming variant position,Genotype quality assuming non-variant position}">
call.MQ	INTEGER	NULLABLE	FORMAT=<ID=MQ,Number=1,Type=Integer,Description="RMS Mapping Quality">
call.PL	INTEGER	REPEATED	FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
call.QUAL	FLOAT	NULLABLE	phred-scaled quality score for the assertion made in ALT - *copied* from variant level QUAL field since this dataset is merged single-sample VCF.
call.VF	FLOAT	NULLABLE	FORMAT=<ID=VF,Number=1,Type=Float,Description="Variant Frequency, the ratio of the sum of the called variant depth to the total depth">
AC	INTEGER	REPEATED	INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
AF	FLOAT	REPEATED	INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
AN	INTEGER	NULLABLE	INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
BLOCKAVG_min30p3a	BOOLEAN	NULLABLE	INFO=<ID=BLOCKAVG_min30p3a,Number=0,Type=Flag,Description="Non-variant site block. All sites in a block are constrained to be non-variant, have the same filter value, and have all sample values in range [x,y], y <= max(x+3,(x*1.3)). All printed site block sample values are the minimum

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

(2) 빅쿼리 코드랩에서 Variants 분석(Analyze variants using Google BigQuery code lab)

- http://googlegenomics.readthedocs.io/en/latest/use_cases/analyze_variants/analyze_variants_with_bigquery.html

2. BigQuery를 통한 유전체 데이터 분석(Analyze Variants Using BigQuery)

(3) R, R마크다운, 자바스크립트에서 빅쿼리 Variants 접근하기(Access Variants in BigQuery with R, RMarkdown, or Javascript)

- **https://github.com/googlegenomics/getting-started-bigquery**