Google Genomics Public Data
(Lists & Access methods)

손준영
(wiseosho@gmail.com)

Published Data List

- 1000 Genomes
- Illumina Platinum Genomes
- Reference Genomes
- MSSNG Database for Autism Researchers
- TCGA Cancer Genomics Data in the Cloud
- Supercentenarian Genomics
- Personal Genome Project Data
- ICGC-TCGA DREAM Mutation Calling Challenge synthetic genomes
- Simons Genome Diversity Project

Pub.Data : 1000 Genomes

- ~2500 genomes / 25 populations worldwide
- Www.1000genomes.org
- Publication list
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3498066/
  - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3042601/
  - http://www.nature.com/nature/journal/v526/n7571/full/nature15394.html
- Data Locations
  - Cloud Storage Folders
  - Genomics Dataset
  - BigQuery Dataset
-

Illumina Platinum Genomes

- 17 member CEPH pedigree 1463
- http://www.illumina.com/platinumgenomes/
- Cloud Storage/ Genomics Dataset/ BigQuery available

Reference Genomes

- Data for Genome Reference Consortium Human Build(GRCh) Data and other reference genomes
- GRCh37, GRCh37lite, GRCh38, hg19, hs37d5, and b37
- Cloud Storage/ Genomics reference sets

Name	Include
GRCh37	Genome Reference Consortium Human Build 37, 35 fasta files
GRCh37lite	GRCh37 + mitochondrial genome reference sequence
GRCh38	39 fasta files
hg19	93 fasta, mitochondrial sequence, ahplotype assemblies
hs37d5	RCRS mitochondrial seq, herpesvirus 4 type 1, GRCh37
b37	GATK software

Simons Genome Diversity Project

- Identify Natural Selection & Disease-causing genes
- Used to compare with Neanderthal Genomes
  - The complete genome sequence of a Neanderthal from the Altai Mountains
- Generate 250 Genomes / 125 diverse populations.
- Available in Genomics Datasets

Personal Genome Project Data

- Personally Donated Genome Data
- Available in all forms(Cloud,Genomics,Bigquery)

MSSNG Database for Autism Researchers

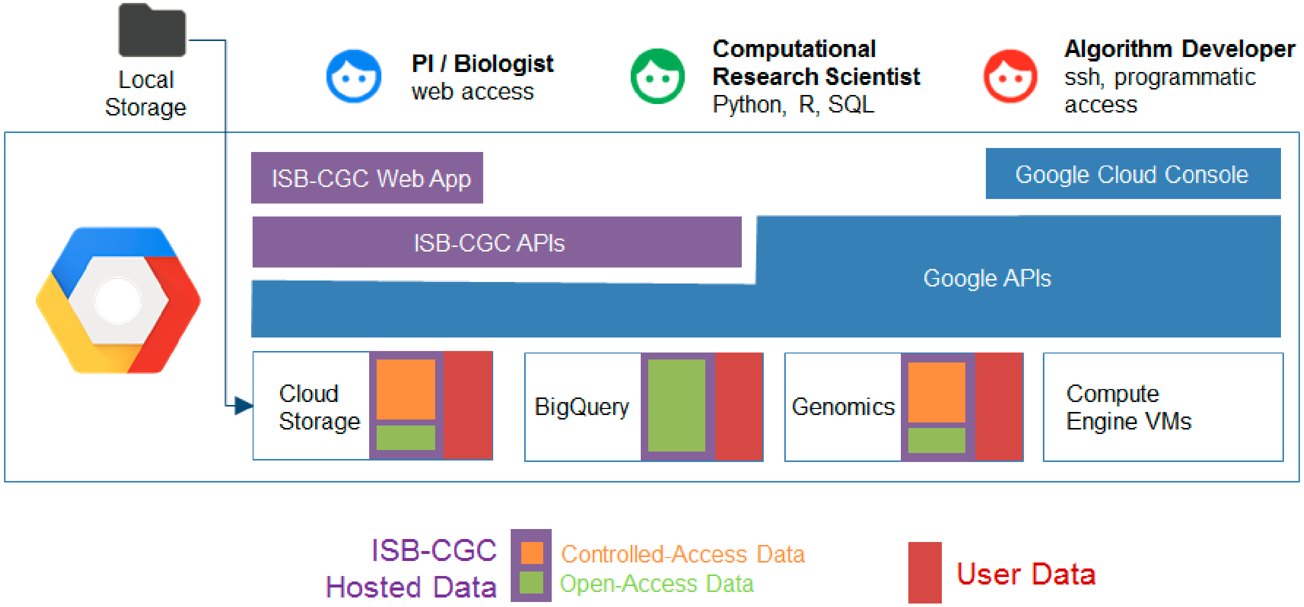
- Collection of Illumina & families affected by autism.
- Grant for Data access needed
  - https://research.mss.ng/
  - Whole-genome sequencing of quartet families wi th autism spectrum disorder
  -

Supercentenarian Genomes

- Complete Genomics genomes for 17 supercentenarians (110 years or older)
- Whole-Genome Sequencing of the World's Oldest People
- Genomics Dataset available(Access Grant need)

TCGA Cancer Genomics Data in the Cloud

- Institute of System Biology(ISB) Cancer Genomics Cloud
  - Somatic mutation calls
  - Clinical data
  - mRNA and miRNA expression
  - DNA methylation
  - Protein expression
- Sample Queries available
  - R, Python
- BigQuery Dataset



ICGC-TCGA DREAM Mutation Calling Challenge synthetic genomes

- Develope Cancer Detection algorithm from WholeGenomeSequence(WGS)
- Detect Mutation calls from the WGS.
- Available as Torrent, Google cloud
  - Challenge Registration needed
  - Cloud, Genomic Data available

Accessing method for public data

1. Programmatic access
   1. Genomics API(REST/RPC)
   2. Sample genome browser
2. Interactive Access
   1. BigQuery
   2. Genomics-public-data project provided
3. File access
   1. Google Cloud Storage
   2. BAM, VCF, FASTA formats
   3. gsutil used.(gs://genomics-public-data