**Instruction to Access 1000genomes Testing Data in dbGaP from Amazon Cloud (Version 1.1)**

**Disclaimer**

Please be aware that it is the user’s responsibility to comply all the related policies and regulations regarding the use of Cloud Computing Services for Storage and Analysis of Controlled-Access Data Subject to the NIH GDS Policy. The services, software and analysis procedures mentioned in this document are just provided for convenience and demonstration ONLY, and DO NOT imply any endorsement or justification by the NIH and the United States Government.

**Purpose**

This instruction is intended for granted users to access 1000 Genomes data in dbGaP with the test key [prj\_phs710EA\_test.ngc](ftp://ftp.ncbi.nlm.nih.gov/sra/examples/decrypt_examples/prj_phs710EA_test.ngc) using NCBI’s sratoolkit in Amazon Web Service (AWS) cloud. This will allow the users to utilize this dataset for testing analysis pipeline and security configuration in AWS cloud so that NIH GDS Policy is conformed.

**The 1000Genome Dataset in dbGaP**

We have created a “fake” dbGaP study for cloud testing, phs000710.v1.p1, in which part of the 1000 Genomes Project data have been deposited into dbGaP. The study title for this study is “1000 Genomes Used for Cloud Testing” ( <http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000710.v1.p1>). You also can use the “SRA Run Selector” to visualize the detailed information for each run for this study (<http://trace.ncbi.nlm.nih.gov/Traces/study/?acc=phs000710>). Please note that not all 1000genomes data are in dbGaP, ONLY the following data set are included in this study:

1. 101 samples from LWK (Luhya in Webuye, Kenya) population with low coverage whole genomes sequences data (Illumina platform).

For each sample, the mapped and unmapped alignment bams have been loaded with two separate SRR identifiers.

For an instance, for LWK sample NA19017, SRR1219818 now corresponds to the mapped bam (<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/NA19017/alignment/NA19017.mapped.ILLUMINA.bwa.LWK.low_coverage.20121211.bam> ) in dbGaP, and SRR1219841 corresponds to the unaligned bam (<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/NA19017/alignment/NA19017.unmapped.ILLUMINA.bwa.LWK.low_coverage.20121211.bam> ) in dbGaP.

1. CEU Trio high coverage PCR-free whole genome sequencing data (Illumina platform).

The origins of this set were:

<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/NA12878/high_coverage_alignment/NA12878.mapped.ILLUMINA.bwa.CEU.high_coverage_pcr_free.20130906.bam>

<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/NA12891/high_coverage_alignment/NA12891.mapped.ILLUMINA.bwa.CEU.high_coverage_pcr_free.20130906.bam>

<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/NA12892/high_coverage_alignment/NA12892.mapped.ILLUMINA.bwa.CEU.high_coverage_pcr_free.20130906.bam>

**Access 1000genomes data in dbGaP from AWS**

* 1. You will need to have Amazon AWS account first before you can do anything below. You will need to apply an account if you do not have one.
  2. Launch an Instance with appropriate security from your VPC (Virtual Private Cloud). Please refer to the NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy, and the procedures for creating such a VPC and launching an instance from your cloud provider.
  3. Log onto the your instance from your local Linux machine
  4. Get a copy of latest sratoolkit from NCBI (<http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software> ), and install into your favorite location with your cloud instance (eg EC2 instance in AWS):
     1. If you have Ubuntu instance, then download the latest file (as of April 14, 2015):
* *wget* [*http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.4.5-2/sratoolkit.2.4.5-2-ubuntu64.tar.gz*](http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.4.5-2/sratoolkit.2.4.5-2-ubuntu64.tar.gz)
* *tar -xvzf sratoolkit.2.4.5-2-ubuntu64.tar.gz*
  + 1. If you have CentOS instance, then download the latest file (as of April 14, 2015):
* *wget* [*http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.4.5-2/sratoolkit.2.4.5-2-centos\_linux64.tar.gz*](http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.4.5-2/sratoolkit.2.4.5-2-centos_linux64.tar.gz)
* *tar -xvzf sratoolkit.2.4.5-2-centos\_linux64.tar.gz*
  1. Get a copy of repository key or cert file that is prj\_phs710EA\_test.ngc from NCBI
* *wget* [*ftp://ftp.ncbi.nlm.nih.gov/sra/examples/decrypt\_examples/prj\_phs710EA\_test.ngc*](ftp://ftp.ncbi.nlm.nih.gov/sra/examples/decrypt_examples/prj_phs710EA_test.ngc)

You will see a file, prj\_phs710EA\_test.ngc, under your current working directory.

* 1. Configuration of SRA Toolkit to access 1000genomes data in dbGaP  
     1. if working directory is NOT provided (using default location by sratoolkit)
* *{Tool path}/sratoolkit.2.4.5-2-centos\_linux64/bin/vdb-config --import prj\_phs710EA\_test.ngc*

You should be able to see the following once the command is successful:

*prj\_phs710EA\_test.ngc was already imported.*

*Protected repository is: '/home/ubuntu/ncbi/dbGaP-0'.*

So the default working directory will be “/home/ubuntu/ncbi/dbGaP-0”.

* + 1. if working directory IS provided (eg. Using /mnt/data/dbGaP\_710 as working directory)
* *{Tool path}/sratoolkit.2.4.5-2-centos\_linux64/bin/vdb-config --import prj\_phs710EA\_test.ngc /mnt/data/dbGaP\_710*

You should be able to see the following once the command is successful:

*prj\_phs710EA\_test.ngc was imported.*

*New protected repository was created.*

*Repository directory is: '/mnt/data/dbGaP\_710'.*

So now the working directory is “/mnt/data/dbGaP\_710”.

* 1. Go to your Working Directory (eg. /home/ubuntu/ncbi/dbGaP-0 or /mnt/data/dbGaP\_710 depending on how you set up your working directory in 2.3 above) and start to test. Please note that you need to execute your commands under this working directory.
* *cd* /home/ubuntu/ncbi/dbGaP-0 (or cd /mnt/data/dbGaP\_710)

In this study, sample NA19017 has two SRA runs (SRR1219818 and SRR1219841), the following examples show that users can retrieve the information as needed by their analysis.

* + 1. Example for generating a bam file for a specific region
* *{Tool path}/sratoolkit.2.4.5-2-centos\_linux64/bin/sam-dump --primary --aligned-region 20:10000000-20000000 SRR1219818*

*| samtools view -Sb - > SRR1219818\_20-10000000-20000000.bam*

The parameters used for the command:

--primary Output only primary alignments

--aligned-region <name[:from-to]> Filter by position on genome.

* + 1. Example for generating a pileup for a specific region
* *{Tool path}/sratoolkit.2.4.5-2-centos\_linux64/bin/sra-pileup --aligned-region 20:10000000-20100000 SRR1219818 > SRR1219818\_20-10000000-20100000.pileup*

Please note that a timeout error may be encountered if the specified region is too big depending on the region size and depth coverage!

* + 1. Generating fastq files for paired end reads (two fastqs with paired reads, plus one fastq with unpaired reads)
* *{Tool path}/sratoolkit.2.4.5-2-centos\_linux64/bin/fastq-dump SRR1219818 --qual-filter-1 --split-3 --minReadLen 70 --gzip -W*

It will generate the following three files:

*SRR1219818\_1.fastq.gz*

*SRR1219818\_2.fastq.gz*

*SRR1219818.fastq.gz*

The parameters used for the command:

--qual-filter-1 Filter used in current 1000 Genomes data

-W|--clip Apply left and right clips

-M|--minReadLen <len> Filter by sequence length >= <len>

* 1. Encrypt the analysis results, and remove the data that is no longer needed
     1. Before you transfer your final results to the final destination, please follow NIH Cloud Policy to encrypt the final analysis result data (eg.vcf)
     2. Before you terminate your cloud instance, please follow NIH Cloud Policy and remove all the data that is no loner needed.