Nov
2017

edico genome

AWS Marketplace Quick-Start Guide

www.edicogenome.com

info@edicogenome.com

Edico Genome Inc. | 3344 North Torrey Pines Court, Plaza Level, La Jolla, CA 92037

Contents

1 Getting started 2

1.1 Step 1, Choose AMI 2

1.2 Step 2, Choose an Instance Type 3

1.3 Step 3, Configure Instance Details 3

1.4 Step 4, Data Flow and Storage Configurations 5

1.5 Step 5: Add Tags 7

1.6 Step 6: Configure Security Group 7

2 Instance Configuration 8

2.1 SSH access 8

2.2 AWS CLI (Command Line Interface) 8

2.3 RAID configuration 9

3 Build a hashtable reference 9

4 Run a DRAGEN sample 10

4.1 Streaming input files from S3 10

4.2 Hashtable storage and transfer 11

4.2.1 DNA vs RNA analysis 11

4.2.2 Storing Hashtables in S3 11

1 Getting started

Login to the AWS Management Console with your AWS account credentials.

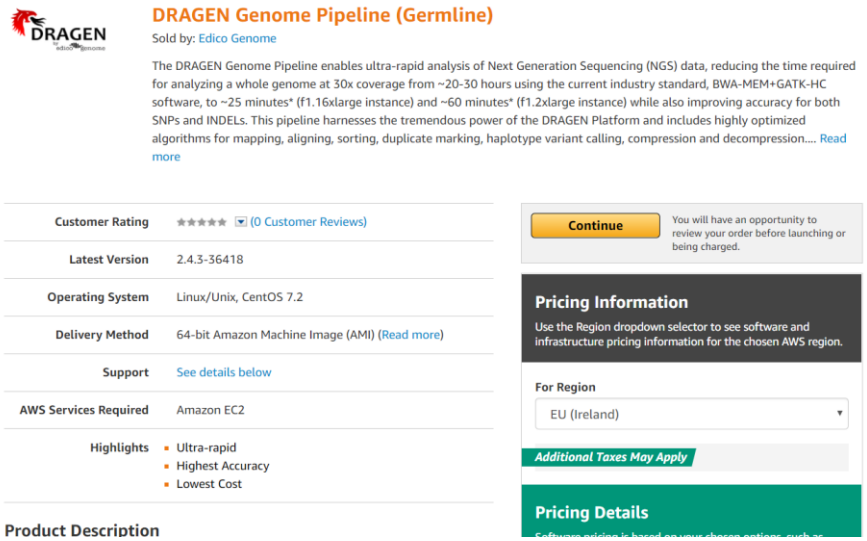
1.1 Step 1, Choose AMI

If you are logged in to your AWS Account:

- Go to Services->EC2 and click on AMIs (under IMAGES) in the left panel.
- Filter on Public Images. Type ‘DRAGEN’.
- Select the appropriate DRAGEN AMI and click the Launch button.

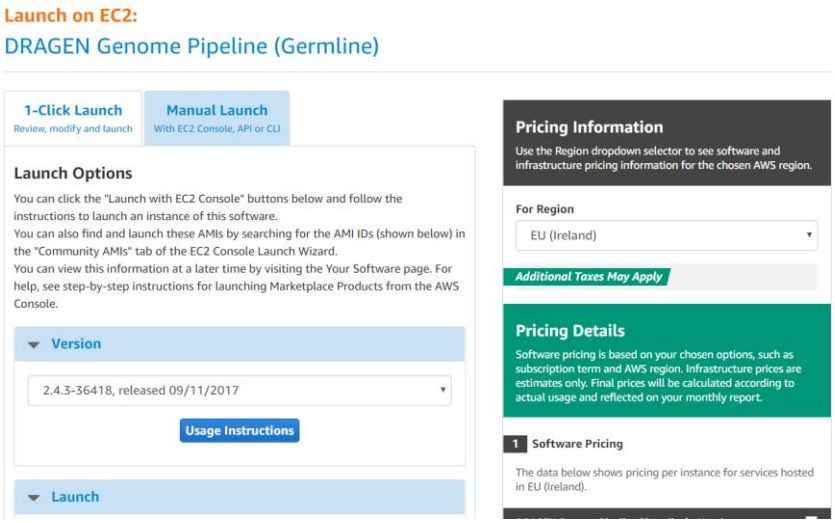
If you are accessing from AWS Marketplace:

- Go to https://aws.amazon.com/marketplace and search for ‘DRAGEN’.

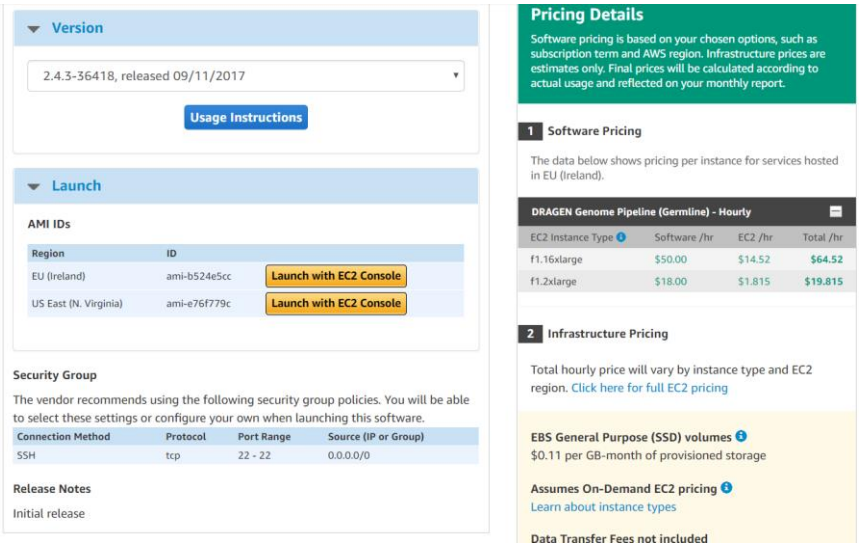


Choose your Region and click ‘Continue’.

For your very first use of the DRAGEN app, you will need to use the ‘Manual Launch’ option.



Scroll down and under the Launch options, select “Launch with EC2 Console” for the region you want to run your instances in.

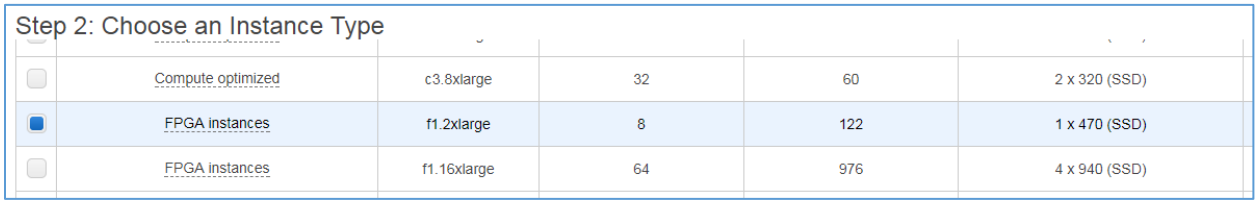


Proceed to Section 1.2 for choosing your instance type. Note that after your initial run on the DRAGEN App, you will be able to use “1-Click Launch” based on the information you’ve previously input.

1.2 Step 2, Choose an Instance Type

The DRAGEN AMIs can only run on f1 instances. As of August 2017 there are 2 f1 instances, f1.2xlarge and f1.16xlarge.

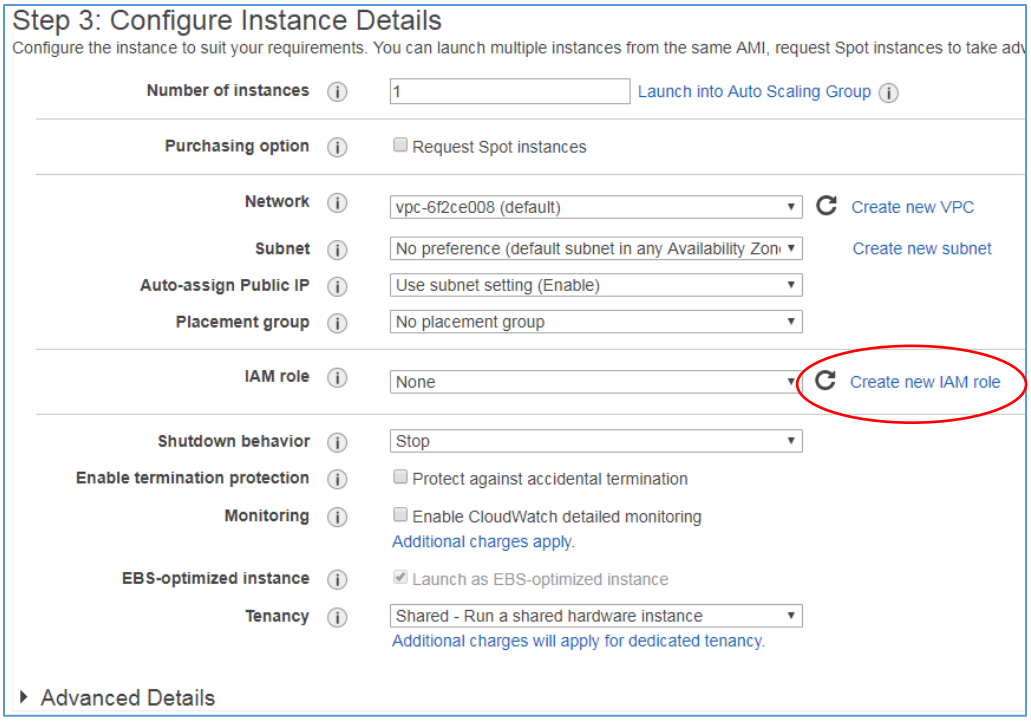
Select f1.2xlarge, unless you require the extreme speed and capacity of the f1.16xlarge.



Click ‘Next: Configure Instance Details’.

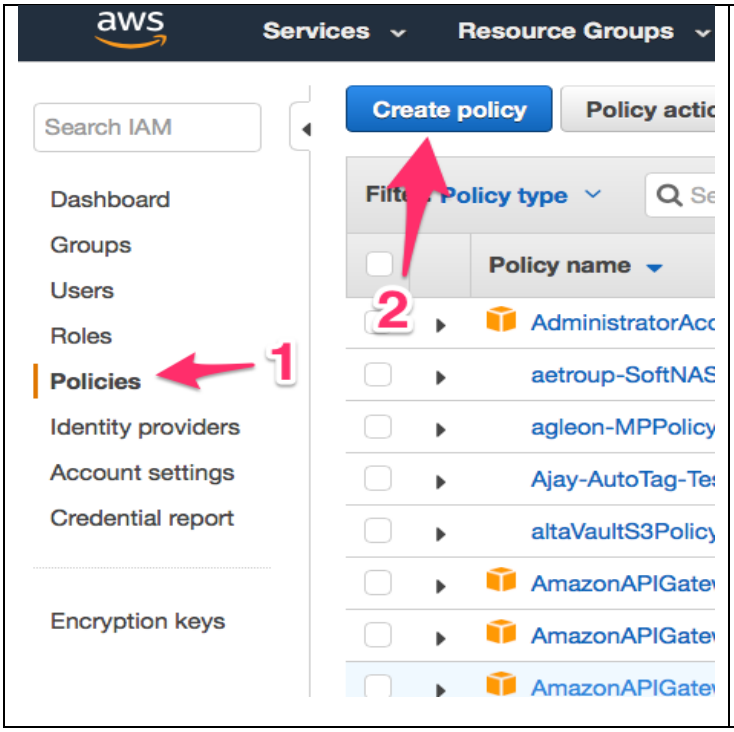
1.3 Step 3, Configure Instance Details

Make any changes required by your configuration. In particular, you should assign an IAM role. You should consult with your IT department if you are not sure how to configure this page. Note that ‘EBS-optimized instance’ is enabled by default and cannot be disabled.

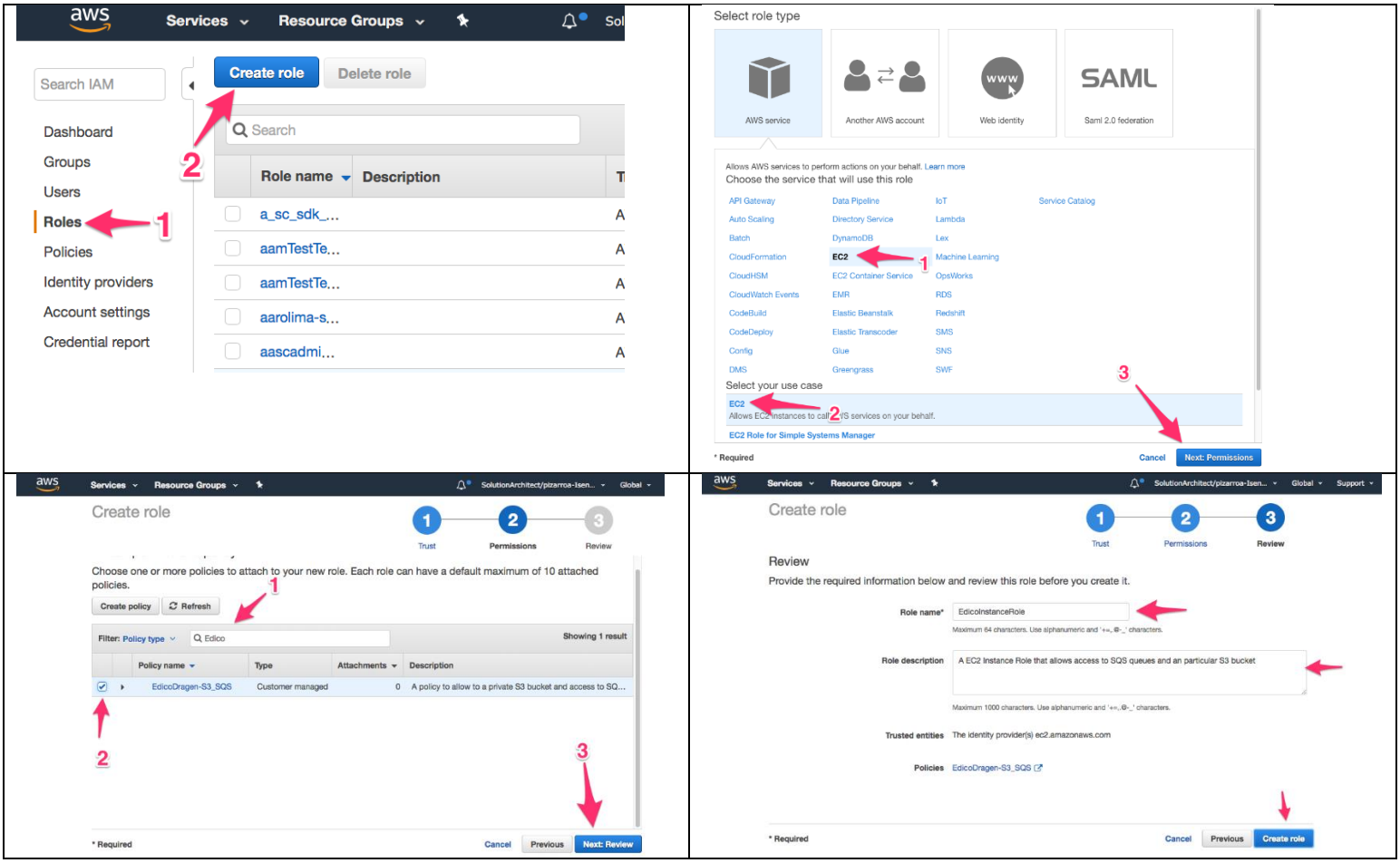


As an advanced configuration, you may create a new IAM policy and IAM role and use that during the launch. Users who want to have DRAGEN read data directly from s3 buckets can create the following IAM policy, and attach it to their user account, or to their instance's Role.

Below pictures show the steps to do that:




```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "StmtAllowDragenStreaming",
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:GetObjectVersion",
        "s3:ListBucket",
        "s3:ListObjectVersions",
        "s3:GetBucketLocation"
      ],
      "Resource": [
        "arn:aws:s3:::<bucket name>",
        "arn:aws:s3:::<bucket name>/*"
    }
  ]
}
```



The newly created role can be used for the launch of the instance.

When complete, click ‘Next: Add Storage’.

1.4 Step 4, Data Flow and Storage Configurations

DRAGEN is able to process input data directly from S3 or from a pre-signed URL or data present on an attached storage device to the instance. Streaming use case eliminates input data download time. The f1.2xl instance has NVMe storage of 438 GB and could also be augmented with standard EBS storage from AWS.

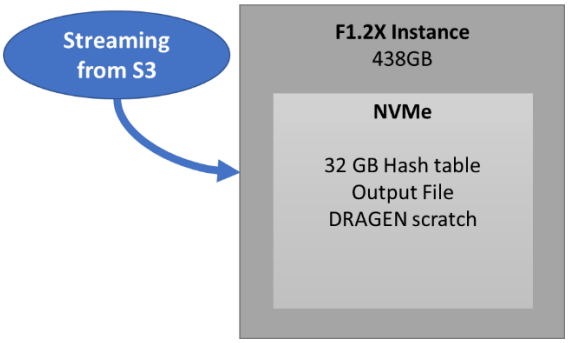
Besides the input data, DRAGEN reference hash table needs to be downloaded to the instance and stored on the instance NVMe.

There are 3 recommended configurations depending on your use case.

Configuration 1:

- The hash table is stored on the NVMe.
- Input is streamed from S3 or pre-signed URL.
- Output is stored to NVMe.
- Temporary files are saved to NVMe (see command line instructions in section 4)

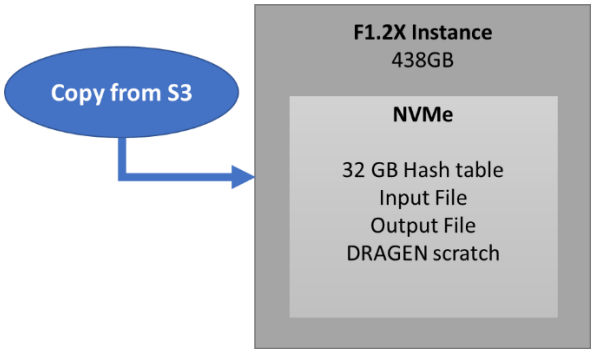
Configuration 1: Streaming Input



If streaming is not an option and if your sample has coverage less than 35x, Edico Genome recommends Configuration 2.

- The hash table is stored on the NVMe.
- Input is copied from S3 to the instance NVMe before processing.
- Output is stored to NVMe.
- Temporary files are saved to NVMe (see command line instructions in section 4)

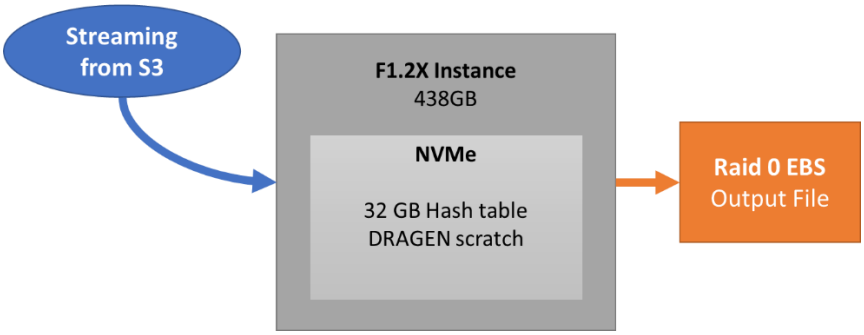
Configuration 2: Using F1.2X with NVMe



For coverage that is 40x and higher, it may be necessary to have an attached EBS volume.Edico recommends attaching 2TB of EBS storage, consisting of 4x500GB volumes configured as RAID0. Hence in Configuration 3,

- The hash table is stored on the NVMe.
- Input is copied from S3 to the attached EBS volume.
- Output is stored on the attached EBS volume.
- Temporary files are saved to NVMe (see command line instructions in section 4)

Configuration 3: F1.2X Instance with EBS



The f1.16xl instance has 4x876GB of NVMe storage. With this available storage, there is typically no need to attach an EBS volume. Also, with this instance, users can chose to have a Streaming input or have the input on the NVMe storage. The output files, hash table and DRAGEN temporary files always use the NVMe storage.

If the NVMe drive on the instance needs to be encrypted, then please to below documentation,

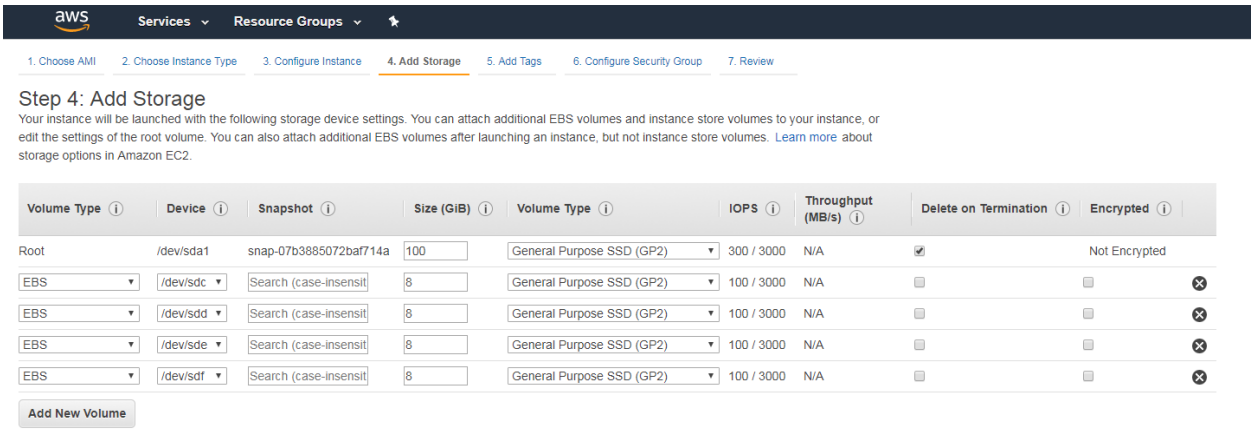
https://aws.amazon.com/blogs/security/how-to-protect-data-at-rest-with-amazon-ec2-instance-store-encryption/

With S3 streaming, please refer to below documentation for providing credentials.

http://docs.aws.amazon.com/sdk-for-cpp/v1/developer-guide/credentials.html

On Step 4, if an EBS volume needs to be attached, click the 'Add New Volume' button 4 times. Go back and change the Size (GiB) to 100 for all 4 volumes. Leave the Volume Type set to GP2. The IOPS should be automatically calculated to 300/3000. **You should change change the name of added devices – the 4 devices shoud be named /dev/sdc, /dev/sdd, /dev/sde, /dev/sdf.** You may choose size GIB to 500 for all 4 volumes if you need a larger EBS.

Please also note that if you wish to encrypt the data in the EBS volume, you may do so by checking the "Encrypted" box.



Click 'Next: Add Tags'

1.5 Step 5: Add Tags

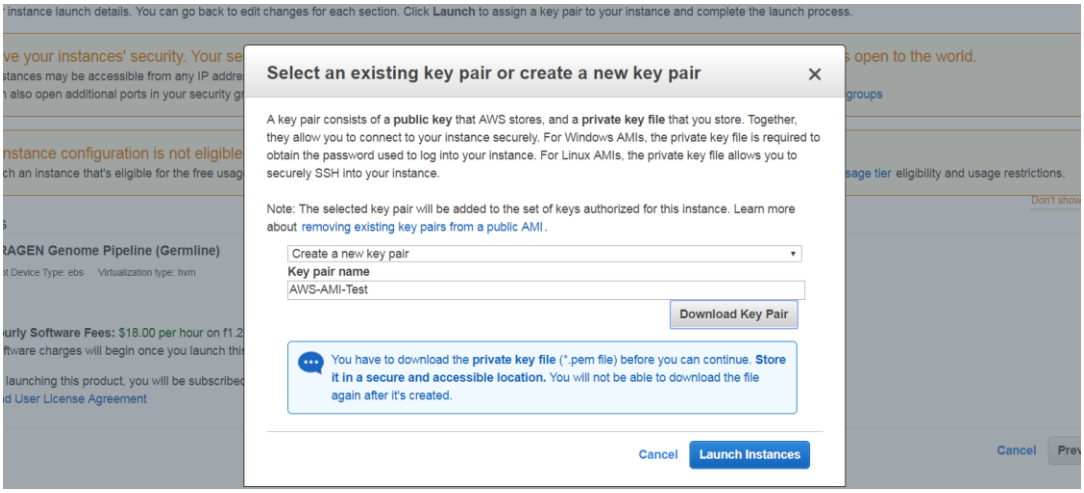
This is optional. You do not have to add any tags.

Click 'Next: Configure Security Group'.

1.6 Step 6: Configure Security Group

This step is very important for security. If you are not sure what to do here, please consult with your IT department. By default, all ports are closed except for SSH, which is accessible from anywhere, but nobody will be able to SSH into your instance unless they have the PEM file, which must be strictly controlled.

Clicking 'Review and Launch' will open a window that will enable you to create and download a PEM file.



Select 'Create a new key pair' and add a name for your PEM file.

Click 'Download Key Pair' and save the file to be used with the launched instance.

Click 'Launch Instances'. You can monitor the launch, which will take a couple minutes.

2 Instance Configuration

2.1 SSH access

Once your instance is launched and running, find your instance in the AWS Management Console, click it, and look for the Public IP address. You can then login to your instance using SSH:

ssh -i /path/to/key-pair.pem centos@nnn.nnn.nnn.nnn

Where:

- /path/to/key-pair.pem = the PEM file (including it's path, if the file is not in the current directory) for the key pair you used when launching this instance. The PEM file must have permissions of 400 (this can be set with the command 'chmod 400 /path/to/key-pair.pem').
- centos = the standard user for the DRAGEN AMIs.
- nnn.nnn.nnn.nnn = the Public IP address of your instance.

SSH access must work before you can proceed with instance configuration.

2.2 AWS CLI (Command Line Interface)

If you want to use AWS CLI, it will be necessary to install that on the instance. This will enable you to use AWS CLI commands to copy data to and from your S3 bucket. This install is possible with this single command,

```
pip install awscli --upgrade --user
```

Also, you can configure access to your S3 bucket on the instance, using the below command. This is optional and necessary if you have not used an IAM policy during instance launch that will allow DRAGEN to stream data directly from your S3 bucket.

```
aws configure
```

This will require you to provide your S3 bucket Access Key and Secret Key in addition to your AWS region.

2.3 RAID configuration

The previous steps attached 4x500GB GP2 volumes to the instance, but these must now be configured and mounted as a single RAID0 volume on /staging.

Verify where the 4 EBS volumes are attached. Usually they will be on /dev/xvdc, /dev/xvdd, /dev/xvde, /dev/xvdf, which correspond to the /dev/sdc, /dev/sdd, /dev/sde, /dev/sdf EBS volumes which were attached to this instance in an earlier step. But it is important to verify their actual locations with the following command:

```
lsblk
```

Verify that you are able to successfully SSH into your instance, then run the following commands:

```
sudo yum -y install mdadm
sudo mdadm --create --verbose /dev/md0 --level=0 --name=MY_RAID0 --raid-devices=4 /dev/xvdc /dev/xvdd /dev/xvde /dev/xvdf
sudo mkfs.ext4 -L MY_RAID0 /dev/md0
sudo mkdir -p /staging
sudo mount LABEL=MY_RAID0 /staging
```

You can test this RAID volume by running the following command, and look for a /staging volume with 2.0T of space:

```
df -h
```

3 Build a hashtable reference

DRAGEN requires a proprietary hashtable reference.

Copy your own FASTA reference file onto the instance and put it in /staging. Or get the hg19 FASTA files from UCSC and concatenate them into a single hg19.fa file using these instructions:

```
mkdir /staging/hg19fa
cd /staging/hg19fa
wget hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz
tar -zxvf chromFa.tar.gz
cat chr*.fa > hg19.fa
```

Then generate the DRAGEN hashtable reference using these commands (if you are using your own FASTA file, replace /staging/hg19fa/hg19.fa with the exact location of your FASTA file, and optionally change the output-directory to be named correctly for your reference). This will take about 20 minutes. Once a hash table is generated on an instance it can be used multiple times by saving it to your S3 bucket and using it for future instances.

```
mkdir /staging/hg19/
/opt/edico/bin/dragen --ht-reference /staging/hg19fa/hg19.fa --output-directory /staging/hg19/ --build-hash-table true
```

Refer to the "Edico Genome DRAGEN Quick Start Guide.pdf" for more details, including configuration options, for building a hashtable reference from a FASTA file.

4 Run a DRAGEN sample

Use the following commands to get 2 sample FASTQs from nih.gov (these are a single lane of NA12878 Exome and are about 3GB total), and run Map/Align + Variant Calling, which should take about 6 minutes on an f1.2xlarge, and will generate a BAM and VCF. Note that subsequent runs of this same sample should take about 4 minutes, because the reference is already loaded on the FPGA.

```
cd /staging
wget ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R1_001_trimmed.fastq.gz
wget ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R2_001_trimmed.fastq.gz
# Note, the following is all one line:
/opt/edico/bin/dragen --output-directory /staging/ --fastq-file1 /staging/NIST7035_TAAGGCGA_L001_R1_001_trimmed.fastq.gz --fastq-file2
/staging/NIST7035_TAAGGCGA_L001_R2_001_trimmed.fastq.gz --output-file-prefix sample --ref-dir /staging/hg19 --enable-variant-caller true
--enable-map-align true --vc-sample-name RGSM --output-format BAM --enable-map-align-output true
```

The "Edico Genome DRAGEN Quick Start Guide.pdf" contains sample commands to perform various DRAGEN operations (running a Map/Align and/or Variant Caller analysis, etc). The commands in this document assume that your input files (FASTQs, BCLs, hashtable reference, etc) are already present on your instance. See Section 4.1 "Streaming input files from S3" for more information on copying files from S3 to your instance.

The "Edico Genome DRAGEN user Guide.pdf" contains more detailed information on all command-line arguments.

4.1 Streaming input files from S3

DRAGEN is capable of streaming FASTQ.gz input files directly from an S3 bucket, or using HTTP pre-signed URLs. This can save quite a bit of time, compared to copying the input files from S3 to local storage and then running DRAGEN on the local copy. Refer to the "Edico Genome DRAGEN Quick Start Guide.pdf" for examples.

4.2 Hashtable storage and transfer

The hashtable is a proprietary reference required by DRAGEN. A 'hashtable' is actually a directory containing several files, totaling about 32GB for DNA analysis or 64GB for RNA analysis.

4.2.1 DNA vs RNA analysis

If you are performing only DNA analysis, but your hashtable contains RNA information, you can decrease the size of it by simply deleting the entire anchored_rna/ subdirectory.

Some newer versions of DRAGEN allow the hashtable to be generated without RNA information by default.

4.2.2 Storing Hashtables in S3

Edico has determined that good performance (with the least maintenance) is achieved by storing the hashtables as .tar files in S3 (for example, hg19.tar, GRCh37.tar, etc); then copying the single tar file to the f1 instance, and un-tar'ing it before DRAGEN runs.

If the hashtable is stored as a .tar.gz in S3, it is slightly smaller which results in a slightly shorter download time, but it takes much more time to gunzip the file (5-10 minutes). This is not recommended.

If the hashtable is stored as individual files in a directory structure in S3, then the files may be downloaded in parallel, resulting in a slight performance improvement; also the un-tar step can be skipped, saving 2-5 minutes. However there may be some long-term maintenance required because the filenames contained within a hashtable could change with newer versions of DRAGEN.

Users may also experiment with storing Hashtables on EFS volumes which are shared across f1 instances; however, in our testing, EFS volumes with <1TB of data are not performant, and are much more expensive than S3.

Information furnished is believed to be accurate and reliable. However, Edico Genome Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication or otherwise under any patent or patent rights of Edico Genome Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all information previously supplied.

Trademarks

DRAGEN is a trademark of Edico Genome Corporation. Other company product names may be trademarks of the respective companies with which they are associated.

Copyright