HGT-ID User Guide, version 1.0

Division of Biomedical Statistics and Informatics, Mayo Clinic

January 2017

Contents

Introduction

An efficient and sensitive program for detecting viral insertion sequences from known viral reference genome in the genome of human cancers.

Quick Start Virtual Machine

A virtual machine image is available for download at
http://bioinformaticstools.mayo.edu/research/hgt-id/

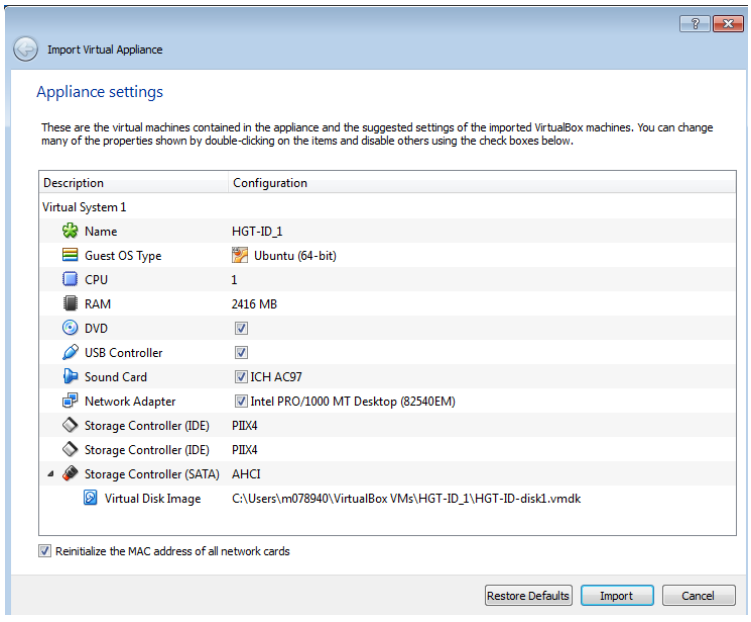
This includes a sample dataset, references (limited to Chromosome 18), and the complete HGT-ID package pre-installed. Please make certain that the host system meets the following system requirements:

- Oracle Virtual Box software (free for Windows, Mac, and Linux at https://www.virtualbox.org/wiki/Downloads)
- At least 4GB of physical memory
- At least 10GB of available disk.

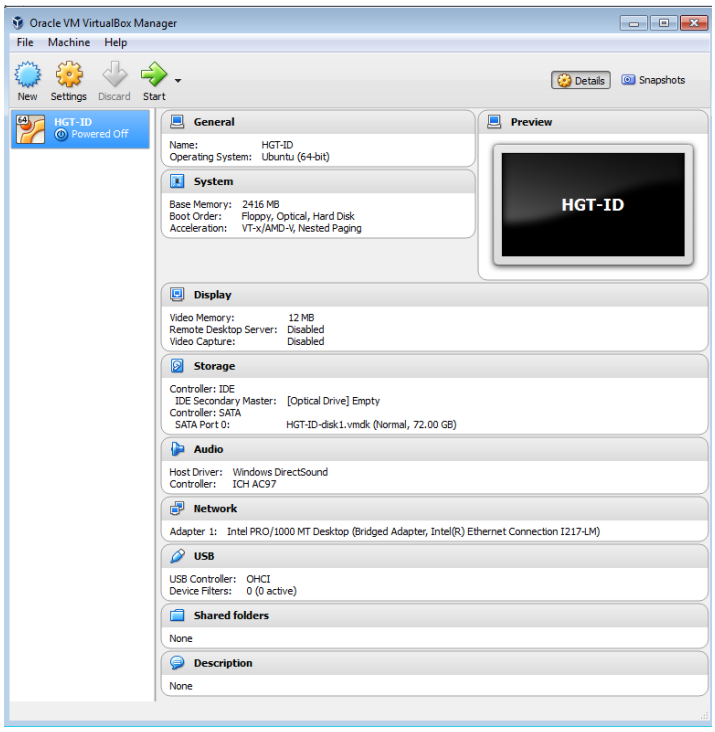
Most recent desktops will have virtualization extensions enabled by default.

Once Virtual Box is installed and the virtual machine image is downloaded you can launch the software by clicking on the HGT-ID.ova file:

Click on the "Import" button to load the virtual machine:



It will appear in the list of available Virtual Machines. Clicking the green start arrow will launch the system:



Once virtual image is launched the virtual machine will present instructions for starting the workflow.

This virtual machine can be used to process additional samples, but this will require allocating more memory (~10 GB) than may be available on a typical desktop system. If you have questions about expanding the VM please contact us for assistance.

System requirements for full setup

To use HGT-ID you will need:

1. A Linux (64-bit) workstation. We currently do not support any Windows environments. We recommend 4-cores with 16GB ram to get optimal performance.
2. Approximately 100GB of storage space for source, tools and reference file installation.
3. A high speed internet connection to download large reference files.
4. The following tools need to be preinstalled and available in your environment path:

    - JAVA version 1.7.0_03 or higher

- Perl version v5.10.0 or higher
- gcc and g++

5. Additional storage space of approximately 1TB for analyzing input data is recommended.

Use the "which" command to identify if all preinstalled tools are installed on your system. Use the "<tool> --version" command to verify the proper version of each tool.

Note that the tool incorporates several third party tools: BWA, BEDTools, Picard, Samtools and Primer3. They are downloaded and compiled on the fly during the execution of setup script. Once the pipeline is installed, they should work if your system meets the pre-mentioned requirements. Make sure your system will work with these tools.

Installation

Users can download the latest version of the package from:
http://bioinformaticstools.mayo.edu/research/hgt-id/

- Download the file linked to the source.
  http://bioinformaticstools.mayo.edu/research/hgt-id/HGT-ID_v<version>.tar.gz
- Move the file to an appropriate directory (<your directory>) and run the following command under (<your directory>) to un-compress the file:

  ```
  tar –xvzf  HGT-ID_<version>.tar.gz
  ```

  Note that after uncompressing the tar.gz file, a new folder will be created under <your_directory> and named as: HGT-ID_<version>

- HGT-ID setup to install full package for human (hg19): under the installation directory, run the following command:

  ```
  <HGT-ID_HOME>/setup.sh –r <reference genome used to generate BAM file>
  ```

```
franklin01:lets work HGT-ID_v1.0 : ./setup.sh
Must provide at least required options. See output file for usage.
######
##      HGT-ID v1.0 installation script
##      Script Options:
##              -r      -               Reference genome FASTA file used to generate BAM file
##              -h      -               Display this usage/help text (No arg)
##              -v      -               verbose (No arg)
##
################################################################################################
##
## Authors:             Saurabh Baheti
## Creation Date:       September 20 2016
## Last Modified:       Januray 26 2017
##
## For questions, comments, or concerns, contact Saurabh Baheti(baheti.saurabh@mayo.edu)
##
################################################################################################
```

- The setup script does the following:
  - It creates all scripts to be executable and sets all the environment variables required for the tool.
  - It downloads and installs all the required tools to run the pipeline.
  - It also creates a configuration files to run the package

- After Installation, the following directory structure is created automatically:

```
< HGT-ID_HOME >
      | _ <bin>
            | _ <all the tools>
      | _ <resource>
            | _ <all the references>
      | _ <src>
            | _ < source code >
      | _ <docs>
            | _ <user manual>
```

- After this step user is ready to run any sample through the pipeline

Step-by-Step instructions to run HGT-ID

- **Input files**
  - The package works with sequencing data from Illumina sequencing platform.
  - User should have aligned BAM file to the human reference genome.

- User needs to execute the script from "src" folder

```
franklin03:lets work HGT-ID_v1.0 : src/hgt.pl
Usage = hgt.pl -b  -c


usage: hgt.pl [-sov] -b <BAMFILE> -c <CONFIG>
        -s              sample Name [Extracted from BAM header]
        -o              Output Directory [cwd]
        -v              verbose flag [no]
        -d              debug flag [no]
```

Required parameters needed for this script to run are

- BAM file including full path
- configuration file with full path

Contact Information / Support

If you have questions or need assistance using the HGT-ID package, please feel free to contact:

Saurabh Baheti
Baheti.Saurabh@mayo.edu
Xiaojia Tang
Tang.Xiaojia@mayo.edu