

eSNV-Detect User Guide, version 1.0
Division of Biomedical Statistics and Informatics, Mayo Clinic
January 2014

Contents

1. Introduction
2. Quick Start Virtual Machine
3. System requirements for full setup
4. Software Requirements
5. Installation and set-up
6. Step-by-Step instructions to run eSNV-Detect on users' samples
7. Contact information / Support

Introduction

eSNV-Detect is the method to achieve high specificity and sensitivity is to use multiple aligners to do alignment to both genome and transcriptome followed by removing duplicate sequence reads and to call SNPs using SAMtools-mpileup.

Sequence-based features like read quality assessment along with a variety of databases such as dbSNP, 1000 genome, refGene, avsift, ljb_phylop from UCSC were used for reliable calling. For the expressed SNVs detected, it can also identify the amino acid change and classify the protein domains.

Quick Start Virtual Machine

A virtual machine image is available for download at <http://bioinformaticstools.mayo.edu/research/esnv-detect/>

This includes a sample dataset, references (limited to Chromosome 22), and the complete eSNV 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.

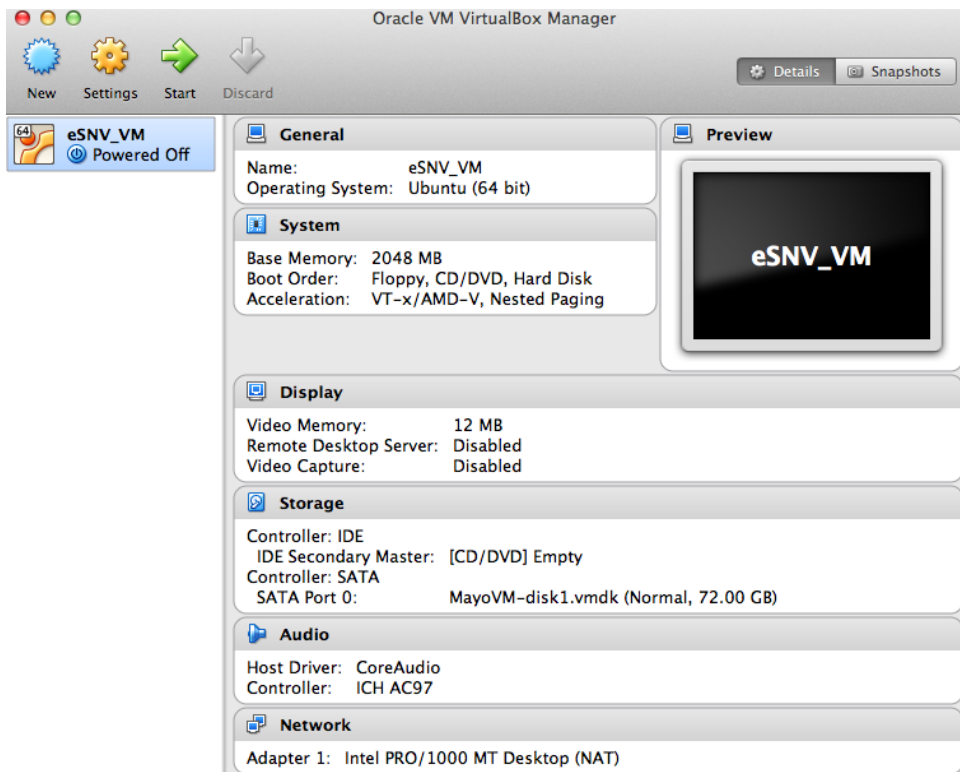
Once Virtual Box is installed and the virtual machine image is downloaded you can launch the software by clicking on the **eSNV_VM.ova** file:



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.

Source code and reference files are all available to download via:
<http://bioinformaticstools.mayo.edu/research/esnv-detect/eSNV-Detect-1.0.tar.gz>

System requirements for full setup

To use eSNV-Detect your 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 1.6.0_17 or higher
 - Perl version 5.10.0 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 pipeline incorporates several third party tools: GATK, Annovar, Picard and Samtools. 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.

Workflow Software Requirements

eSNV-Detect relies on the following bioinformatics tools:

GATK v1.6.9

<ftp://ftp.broadinstitute.org/distribution/gsa/GenomeAnalysisTK/GenomeAnalysisTK-1.6-9-g47df7bb.tar.bz2>

Picard Tools v1.106

<http://downloads.sourceforge.net/project/picard/picard-tools/1.106/picard-tools-1.106.zip>

Samtools v0.1.19

<http://downloads.sourceforge.net/project/samtools/samtools/0.1.19/samtools-0.1.19.tar.bz2>

Annovar

http://www.openbioinformatics.org/annovar/annovar_download.html

Steps to follow:


- User needs to fill an online form and then gets an email for download
- Install the package and required references in a directory and you can point your own versions via the tool_info.txt file as needed.


The eSNV-Detect installation process installs all the other packages other than Annovar.

Installation

Users can download the latest version of the package from:


<http://bioinformaticstools.mayo.edu/research/esnv-detect/>

-  Download the file linked to the source.
<http://bioinformaticstools.mayo.edu/tools/eSNV-Detect/eSNV-Detect-<version>.tar.gz>

-  Move the file to an appropriate directory (<your_directory>) and run the following command under (<your directory>) to uncompress the file:

```
tar -xvzf eSNV-Detect-<version>.tar.gz
```

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

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

```
<eSNV_Detect_HOME>/setup.sh -d <eSNV-Detect_HOME>
```

```

lets work: ./setup.sh
Must provide at least required options. See output file for usage.
#####
##      eSNV-Detect v1.0 installation script
##      Script Options:
##      -d      <directory>      -      (REQUIRED) full/path/to/installation directory for eSNV-Detect
##      -h      -      Display this usage/help text (No arg)
##
#####
##
## Authors:          Saurabh Baheti
## Creation Date:    January 19 2014
## Last Modified:    January 19 2014
##
## For questions, comments, or concerns, contact Saurabh (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
- The step takes about 20-30 minutes (mainly for dbSNP database download).

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

```

< eSNV-Detect_HOME >
| _ <bin>
|   | _ <all the tools>
| _ <resource>
|   | _ <all the references>
| _ <src>
|   | _ < source code >
| _ <docs>
|   | _ <user manual>

```

✚ After the installation of tools and source code user needs to install Annovar. Installation of Annovar is not complicated and its website provides user very good directions to do the same.

Steps to install Annovar are as follows:

- Fill the download form to get access to the Annovar
http://www.openbioinformatics.org/annovar/annovar_download_form.php

- Once user have the link in the email

```
wget <Annovar link>
```

- To install Annovar there is a utility script provided in the package which can be run as follows

```

lets work: ./annovar.sh
Must provide at least required options. See output file for usage.
#####
##      eSNV-Detect v1.0 annovar installation script
##      Script Options:
##      -p      <pacakage>      -      (REQUIRED)      full path to downloaded annovar package
##      -d      <directory>      -      (REQUIRED)      full/path/to/annovar directory user needs to install at
##      -h      -      Display this usage/help text (No arg)
##
#####
##      Authors:      Saurabh Baheti
##      Creation Date:      January 19 2014
##      Last Modified:      January 19 2014
##
##      For questions, comments, or concerns, contact Saurabh (baheti.saurabh@mayo.edu)
##
#####

```

- User needs to pass downloaded package and directory where to install Annovar to the script and script should install all the required databases.
- After the end of the script user needs to update the configuration file which the path to Annovar and Annovar Database.

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

Step-by-Step instructions to run eSNV-Detect on users' samples

✚ **Input files**

- The package works with sequencing data from Illumina sequencing platform.
- User should have aligned BAM file from BWA and Top hat or Map splice

✚ After the setup configuration file should be already present on the top level folder of the tool. Make sure you have updated the path to **ANNOVAR** and **ANNOVARDB**. One parameter you must make sure to change is the **ALIGNER**. You can all the other parameters same as these are all defaults.

✚ User needs to execute the script from “src” folder

```

./main.sh
Main script to run the whole tool
Usage: ./main.sh </path/to/input directory><sample name><bamfiles from
same aligner are , seperated and : for different
aligners></path/to/outputfolder></path/to/configuration file>

```

Parameters needed for this script to run are

- path to input directory for the BAM files (full path of the folder)
- sample Name
- BAM files (BAMs for different aligner are colon separated and for the same aligner are comma separated)
- path to output folder (full path of the folder)
- and configuration file (full path of the configuration file)

❖ Description of the identifiers in the configuration file

IDENTIFIER	Format	Description
ANNOVAR	<directory>	User needs to add the path to the Annovar scripts
ANNOVARDB	<Directory>	User needs to add the path to the Annovar databases
REALIGNEMNT	YES/NO	Flag to do realignment
RECALIBRATION	YES/NO	Flag to do recalibration
ALIGNER	BWA:TOPHAT	Colon separated aligners used for aligning the reads of the sample.
PROTEOMICS	YES/NO	Flag to annotate variants using protein domains
TEMPORARY_FILE S_REMOVE	YES/NO	Flag to remove intermediate files
CHR_INDEX	1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20:21:22:X:Y:M	(List of chromosomes user need to analyze ':' separated)
JVM_MEM	-Xmx10g -Xms5g	Virtual memory required to run scripts
HIGH_ReadRankPosSum	8	Max. value for Read post rank Sum
LOW_ReadRankPosSum	8	Min. value for Read post rank Sum
MIN_ALT_READS	3	Min. alternate supporting reads.
MIN_READ_DEPTH	3	Min. read depth required to pass the filters
R_CUTOFF	0.1	
R_CUTOFF2	0.05	
KEEP_ALL_SNV	T/N	T for keeping all the variants regardless of all the filters.

Limitations to the workflow

- Sample names cannot start with a number or a special character. For example, characters such as “() { } [] . , \$ -” are not permitted.
- The workflow does not run in any Windows environment.

Contact Information / Support

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

Saurabh Baheti

Baheti.Saurabh@mayo.edu

Xiaojia Tang

Tang.Xiaojia@mayo.edu