RVboost User Guide, v0.1

Division of Biomedical Statistics and Informatics, Mayo Clinic

May 2014

Contents

Introduction

A comprehensive package/pipeline to prioritize RNA sequencing called variants using GATK framework. Variants are called using modified version of GATK and then custom and home grown packages written in R, Perl and bash are used for variant prioritization to give a Q-score to each variant seen in the data. HAPMAP variants are used as gold-standard and used as a training set to train the model and apply to the data.

Quick Start Virtual Machine

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

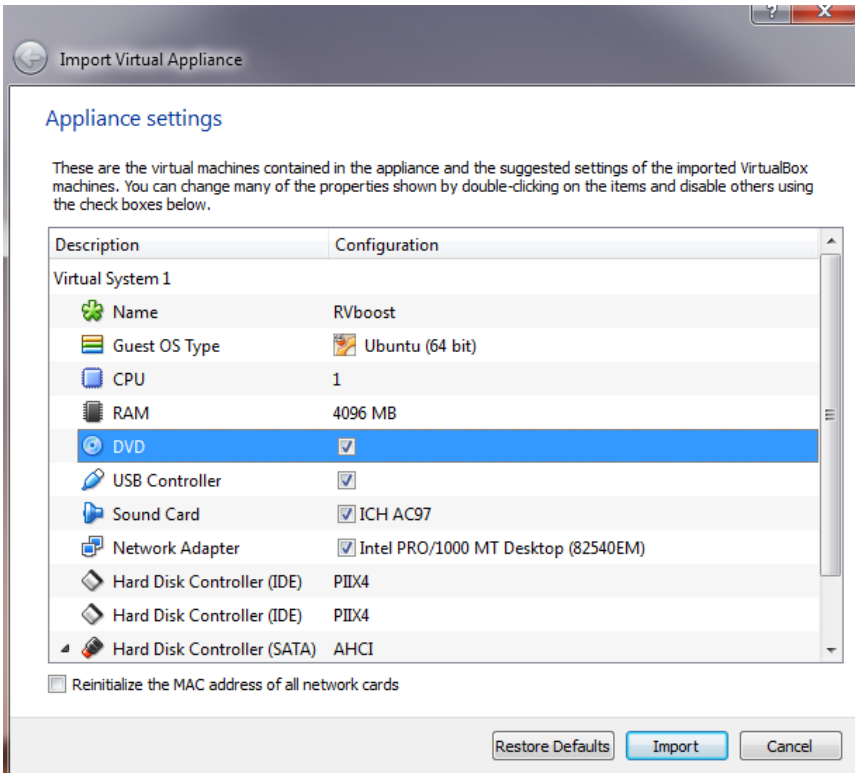
This includes a sample dataset, references (limited to Chromosome 22), and the complete RVboost package pre-installed. Please make certain that the host system meets the following system requirements:

- 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.

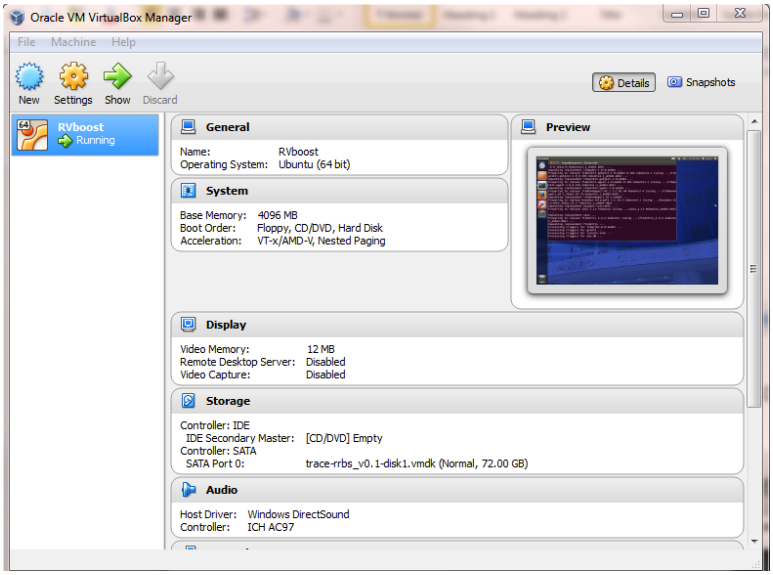
Most recent desktops will have virtualization extensions enabled by default.

Once Virtual Box is installed and the virtual machine image is downloaded you can launch the software by clicking on the RVboost_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 running package for part of MCF7 sample (chr22 reads) for testing purpose.

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/rvboost/RVboost_0.1.tar.gz

System requirements for full setup

To use RVboost user 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 install R packages.
4. The following tools need to be preinstalled and available in your environment path:

    - JAVA version 1.7.0_03 or higher
    - Perl version 5.10.0 or higher
    - R version 3.0.1 or 3.0.2
    - 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, BEDtools, BLAT, SNPEff and samtools. They are 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/rvboost/

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

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

- RVboost setup to install full package: under the installation directory, run the following command:

```
<RVboost_HOME>/setup.sh -r < full/path/to/reference genome FASTA file>
```

```
lets work RVboost_0.1 : ./setup.sh
Must provide at least required options. See output file for usage.
######
##      RVboost v1.0 installation script
##      Script Options:
##      -r      -               <refernce genome>       (REQUIRED) full/path/to/reference genome FASTA file used to align the reads to get the BAM file
##      -h      -               Display this usage/help text (No arg)
##
#############################################################################################################
##
## Authors:             Saurabh Baheti
## Creation Date:       May 14 2014
## Last Modified:       May 14 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 installs all the required tools to run the package.
  - It also creates a configuration files to run any real sample.
  - The step takes about 10-15 minutes.

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

```
< RVboost_HOME >
        | _ <bin>
              | _ <all the tools>
        | _ <resource>
              | _ <all the references>
        | _ <src>
              | _ < source code >
        | _ <docs>
              | _ <user manual>
        | _ <Rlibs>
              | _ <R packages>
```

Step-by-Step instructions to run RVboost on user samples

Input files:

- The package works with sequencing data from Illumina sequencing platform.
- User should have aligned BAM file from RNA sequencing aligner.

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

```
lets work src : ./RV.Boosting.sh
Options specified:
Must provide at least required options. See output file for usage.
#########################################################################
##      wrapper script to run the RV.Boosting method
##      if the DNA BAM file is provided then it will backfill all the raw calls from DNA bam file
##      If capture kit is porvided then it will run the variant calling on that region otherwise whole coding region will be used
##      if you want to run the script on cluster then you need to run as using these memory requirements
##      qsub -cwd -q 7-days -pe threaded 2 -l h_stack=10M -l h_vmem=8G
##
## Script Options:
##      -R  <RNA bam>    -    (REQUIRED)  /path/to/input directory of the input RNA BAM file
##      -s  <samplename>     -    (REQUIRED)      sample Name
##      -c  <config file>    -    (REQUIRED)      /full path/to configuration file
##      -o  <output_dir>     -    (REQUIRED)      /full path/to output dir
##      -D  <DNA bam>    -   /path/to/input directory of the DNA BAM file for the same sample
##      -b  <bed_file>   -   /path/to/capture BED file for DNA sample
##      -T  <threads>    -   number of threads (default 1)
##      -d  <debug mode>     -   if this flag is passed then temporary files will be kept at each step
##      -h                   -       Display this usage/help text (No arg)
#########################################################################
##
## Authors:             Saurabh Baheti
## Creation Date:       March 03 2014
## Last Modified:       March 26 2014
##
## For questions, comments, or concerns, contact Saurabh (baheti.saurabh@mayo.edu)
##
#########################################################################################################################
```

- Required Parameters needed for this script to run are:
- -R full path to the RNA BAM file to be processed
- -s Name of the sample same as Read Group in the BAM file
- -c configuration file generated after the installation script completed successfully in the RVboost HOME folder
- -o full path to the output folder for temporary files and results.

Optional Parameters needed for this script to run are:

- -D full path to the DNA BAM file to recall same genotypes called from RNA BAM
- -b Bed file to call variants on specific region, if not supplied then variants are called on whole coding region. It is good to specify the capture bed for the DNA sample if user wants to compare DNA and RNA.
- -T number of threads to use for doing the analysis. It depends on the system user is using. Default is 1 thread which should be good for most systems.
- -d debug mode, if the flag is passed then all the temporary files will be kept otherwise after the end of the package run all the temporary files will be deleted.

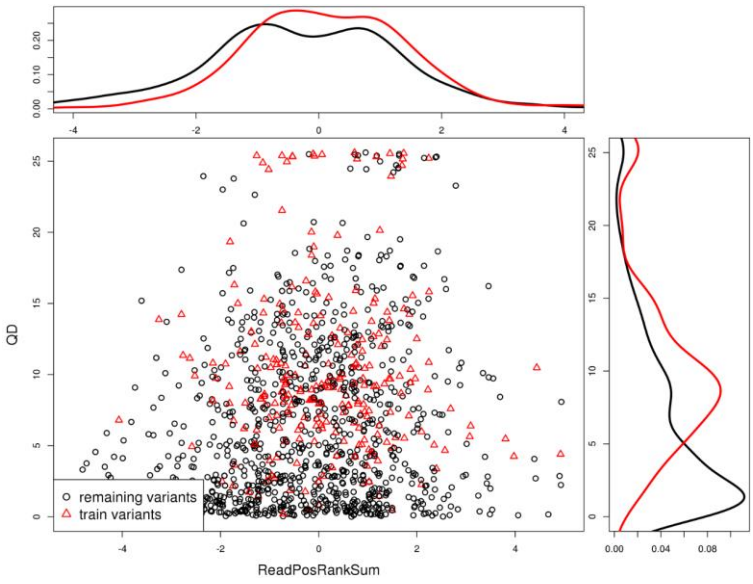
Output files:

- VCF file with QScore(RV Boosting algorithm Q-Score) and OrgScore (RV Boosting algorithm Original score) added to the VCF file INFO field

- TRAIN Flag added to the INFO field of same VCF file to mark variants used in training set.
- A PNG image comparing QD and ReadPostRankSum for training and test set for the sample.

Example output plot and folder structure:

```
<RVboost_sample_output_Folder>
                |_ <sample>.raw.vcf
                |_ <sample>.filter.vcf
                |_ <sample>.png
```



Contact Information / Support

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

Saurabh Baheti
Baheti.Saurabh@mayo.edu
Chen Wang
Wang.Chen@mayo.edu