TRACE-RRBS User Guide, version 0.1

Division of Biomedical Statistics and Informatics, Mayo Clinic

September 2015

Contents

1. Introduction

2. Quick Start Virtual Machine

3. System requirements for setup

4. Software Requirements

5. Installation

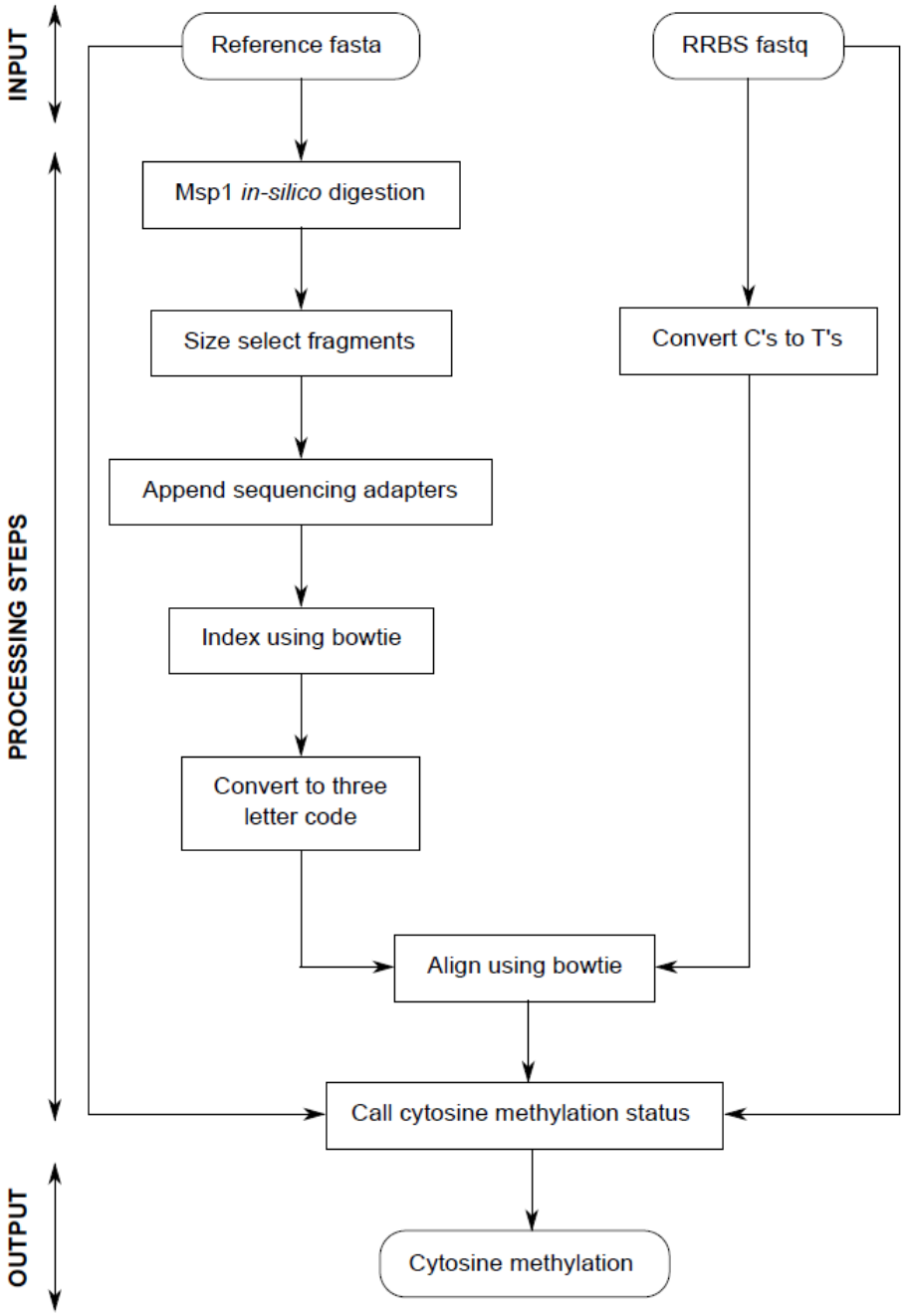
6. Step-by-Step instructions to run TRACE-RRBS on user samples

7. Contact information / Support

Introduction

Targeted Alignment and Artificial Cytosine Elimination for RRBS (TRACE-RRBS) as a full pipeline for mapping bisulfite sequencing data using bowtie2 (using specific parameters) and then generating CpG level data.

Here is the semantics/flowchart of the alignment tool followed by CpG quantification.



Quick Start Virtual Machine

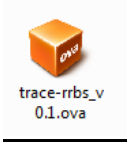
A virtual machine image is available for download at
http://bioinformaticstools.mayo.edu/research/trace-rrbs/

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

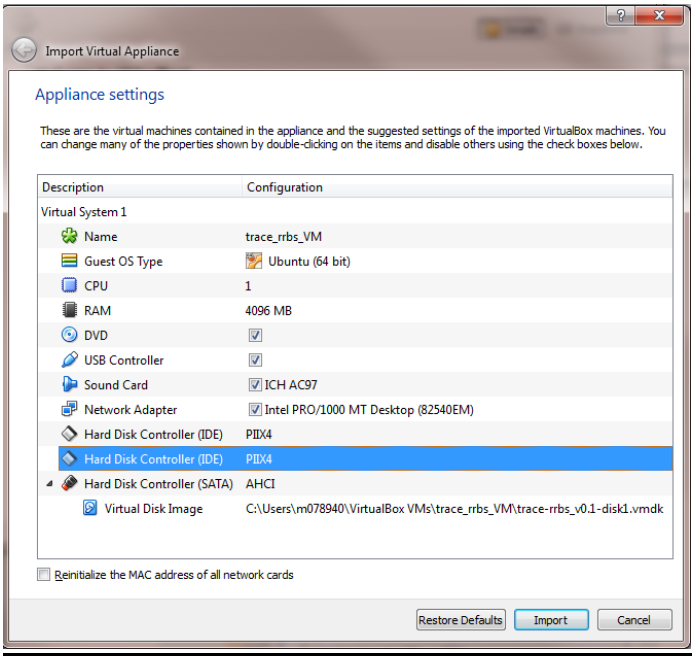
- Oracle Virtual Box software (free for Windows, Mac, and Linux at https://www.virtualbox.org/wiki/Downloads)
- At least 4GB of physical memory
- At least 10GB of available disk.

Most recent desktops will have virtualization extensions enabled by default. If not then you need to enable the same using BIOS option in your machine.

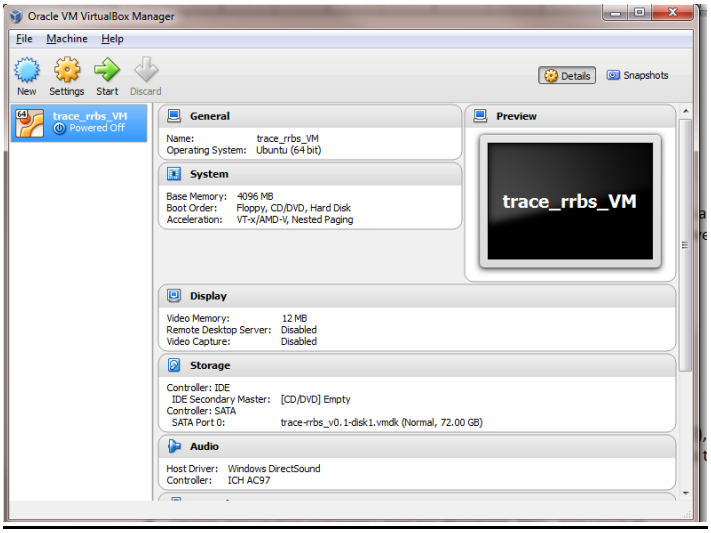
Once Virtual Box is installed and the virtual machine image is downloaded you can launch the software by clicking on the trace-rrbs_VM.ova file:



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. It should typically take less than 5 minutes to run the samples on Virtual machine.

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/trace-rrbs/

System requirements for full setup

To use TRACE-RRBS 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. The following tools need to be preinstalled and available in your environment path:

    - JAVA version 1.7.0_03 or higher
    - Bowtie 2 (http://bowtie- bio.sourceforge.net/bowtie2) needs to be installed on your computer

4. 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.

Installation

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

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

  ```
  tar -xvzf  trace-rrbs_<version>.tar.gz
  ```

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

- All the executable jar files are available in the directory named "src" under the trace-rrbs_<version> directory.

- There is a README available under trace-rrbs_<version> directory for usage information.

Step-by-Step instructions to run trace-rrbs on users samples

- **Input and Reference Files**
  - The package works with sequencing data from Illumina sequencing platform.
  - User should have bi-sulfite treated FASTQ files.
  - Reference FASTA file for the species is needed to run the analysis.

- **Steps to follow:**

  - **RRBS fragments**

In this step reference genome is fragmented using MSP1 enzyme motif cut site and depending on the library size as input parameters and adaptor sequence the FASTA file is created. This step will be done only once if your samples have same library size.

```
RRBS fragments      java -jar methyl_fragment_builder.jar

                    RRBS fragment maker

                    OPTIONS:

                     -adapter_3 VAL  3' adapter Misner adapter [3']:
                                     AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG
                     -adapter_5 VAL  5' adapter. Misner adapter [5']:
                                     AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
                     -max_frag_len N Maximum fragment length
                     -min_frag_len N Minimum fragment length
                     -out_prefix VAL Output prefix for fasta file
                     -read_len N     Read length
                     -ref_fa VAL     Reference fasta file
```

Example:

```
java -Xmx8g -jar methyl_fragment_builder.jar -max_frag_len 250 -min_frag_len
30 -out_prefix rrbs -read_len 50 -ref_fa hg19.fa
```

In the above command we are specifying fragment length b/w 250 and 30. Read length of the sequencing data and the reference genome to be used to fragment the FASTA file using MSP1 enzyme cut site.

  - **Index FASTA**

Create Index of FASTA file created in the above step using bowtie2. This step will be done only once if your samples have same library size.

- **Read Converter**

BS conversion of fastq [C>T for end 1 and G>A for end 2] input and output files can either be .fq, .fastq or .gz. User needs to run this step separately for Read1 and Read2 if it is Paired End RRBS data.

```
Read Converter      java -jar methyl_fq2bisulfite.jar

                    BS conversion of fastq [C>T for end 1 and G>A for end 2] input and output files
                    can either be .fq, .fastq or .gz

                    OPTIONS:

                     -end_type N 1 or 2
                     -in_fq VAL  Input fastq
                     -out_fq VAL Output fastq
```

Example:

```
java -Xmx8g -jar methyl_fq2bisulfite.jar -end_type 1 -in_fq read1.fq -out_fq
read1.bs.fq
```

```
java -Xmx8g -jar methyl_fq2bisulfite.jar -end_type 2 -in_fq read2.fq -out_fq
read2.bs.fq
```

Bisulfite treated reads and converted and the step was run two times as it is paired End RRBS data.

- **Alignment**

Run Bowtie2 on converted FASTQ files using fragmented and index reference genome using the example parameters below.

```
--reorder --phred64 --very-sensitive --end-to-end --dovetail --no-discordant
--rdg 10000,10000 --rfg 10000,10000
```

Example:

```
bowtie2 --reorder --phred64 --very-sensitive --end-to-end --dovetail --no-
discordant --rdg 10000,10000 --rfg 10000,10000 -p 1 -x  <methyl fragmented
Reference genome from first step> -1 <read1 converted FASTQ> -2 <read2
converted FASTQ>  | samtools view -bS - > <output BAM file>
```

- **CpG quantification**

This step is needed to extracts the methylation call for every single C analyzed in your RRBS SAM/BAM file.

```
CpG quanitification java -jar methyl_caller.jar

                    Call the C status on an RRBS sam/bam file

                    OPTIONS:

                     -frag_fa N       fragment fasta file [not bisulfite converted]
                     -in_bam VAL      input paired end / single end BAM file
                     -in_fq1 VAL      input fastq end 1
                     -in_fq2 VAL      input fastq end 2 [optional]
                     -out_prefix VAL  output file prefix
```

Example:

```
java -Xmx8g -jar methyl caller.jar -frag fa rrbs.rrbs.frag.fa -in bam
sample.bam -in_fq1 read1.fq -in_fq2 read2.fq -out_prefix cpg.calling
```

Steps involved methylation extraction from the SAM/BAM file

Contact Information / Support

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

Saurabh Baheti
Baheti.Saurabh@mayo.edu
Rahul Kanwar
Kanwar.Rahul@mayo.edu