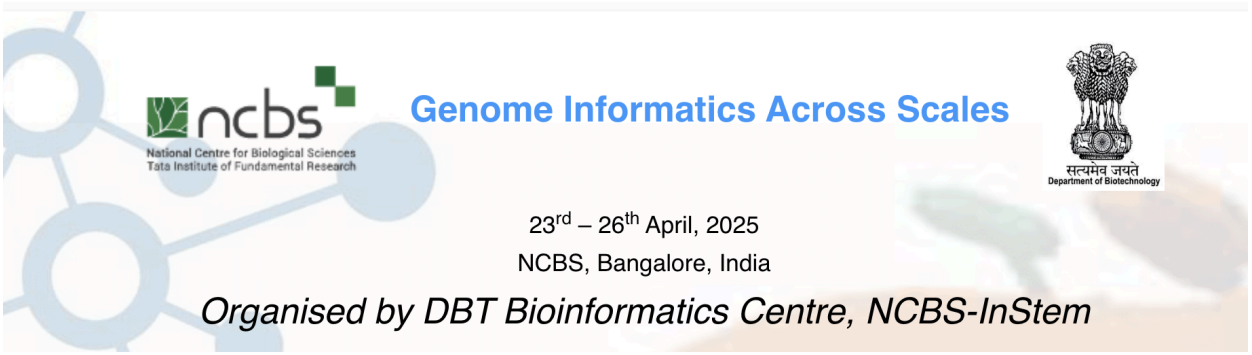
Workshop 5:

Systems Approaches for RNA-Seq Analysis: Identifying Molecular Markers and Functional Targets

(23rd April 6-7 PM, 25th April 3-5 PM)



By

Dr. Pankaj Barah

Assistant Professor

Dept. of Molecular Biology and Biotechnology

Tezpur University (Central), Napaam, Sonitpur

Assam -784028 INDIA

E. mail: barah@tezu.ernet.in

HOMEPAGE: https://www.tezu.ernet.in/dmbbt/profile/34

AND

Dr. B.N.Mahantesha Naika B.N. Assistant Professor,

Dept. of Biotechnology & Crop Improvement,

College of Horticulture, Bagalkot-587104

University of Horticultural Sciences, Bagalkot,

Karnataka, India.

Session I: RNA-Seq Data Analysis for Identification of Molecular Markers

Minimum System Requirements:

1. 8 GB RAM or higher
2. 4-core 8 threads processor or higher (Intel i5 minimum for Intel builds and Ryzen 5 for AMD builds). Please refer to system properties under settings
3. Windows 10, version 1909 or higher (for Windows systems)

   OR

   Ubuntu, version 14.04 or higher/ any other similar UNIX distro (for UNIX systems)
4. Storage space of ~30 GB in the system.
5. Virtualization should be enabled in OS settings as well as in BIOS settings. A tutorial can be viewed here [*https://bit.ly/virtualization23*]
6. Hyperthreading/Multithreading support should be enabled both in OS settings and in BIOS settings. See here [*https://bit.ly/htding23]*
7. Internet data requirement approximately 12 to 20 GB (for downloading the input files, scripts etc.)

Pre-requisites and Downloads:

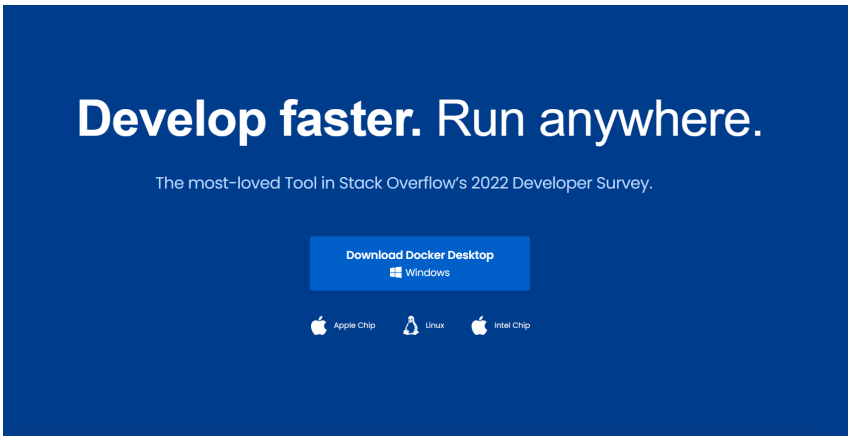
Download the input files from the link given here. [*https://bit.ly/ncbswkshp*]

Please recheck whether you have the downloaded the following files:

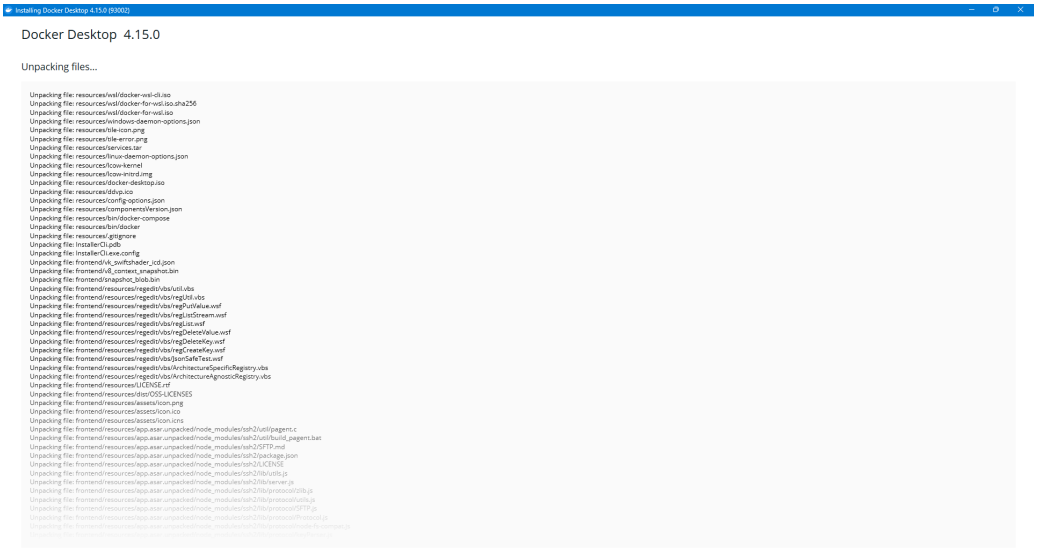
a. 1stDPI1.bam
b. 1stDPI2.bam
c. BPT1.bam
d. BPT2.bam
e. deseq_mrna.R
f. EVORNA.sh
g. Samples.txt

For WINDOWS/MacOS systems, please follow the guide below:

h. Download Docker for WINDOWS/MacOS from here [*https://www.docker.com/]*



i. Install Docker on your system



j. Restart your system

k. If asked for a kernel update, download and install the kernel update package, as directed. Restart after install.

l. Open Docker application and wait for Docker to start

m. Open powershell/command prompt/terminal from your system and enter the code given below:

```
docker pull evolomicsmbbt/rna_seq:latest
```

n. Make sure Docker has successfully downloaded all the layers of the image. (Total size : ~930 MB)

For UNIX systems, please follow the guide below:

o. For UNIX systems, Docker can directly be installed using the aptitude package 'docker'. Please use the code below:

```
sudo apt-get install docker
```

p. Perform a system restart

q. Open terminal in your system and enter the code given below:

```
docker pull evolomics_mbbt/rna_seq:latest
```

r. Make sure docker has successfully downloaded all the layers of the image.

Unzip the downloaded ZIP and organize all downloaded files into one **"working directory"** and record the path of the directory (e.g. C:\NCBS_Workshop\) for future use.

Make sure your working directory looks like this

Running the pipeline:

1. **Open up powershell/cmd/zsh/bash shell.** (Make sure you're running your shell with admin rights)
2. **Mounting your storage to docker:**
    a. Use the following command to mount your working directory to the docker filesystem and run the docker image:

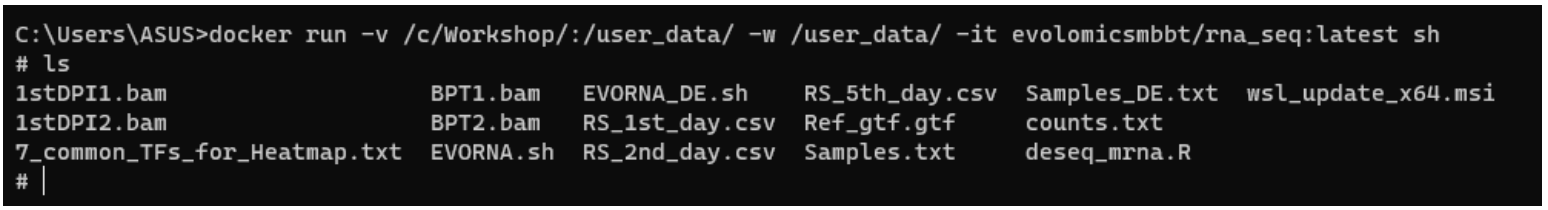
```
docker run -v {windows/path}:/user_data/ -w /user_data/ -it
evolomicsmbbt/rna_seq:latest sh
```

Change {windows/path} with the path to your files, like

```
docker run -v /c/Workshop:/user_data/ -w /user_data/ -it
evolomicsmbbt/rna_seq:latest sh
```

indicating that my files are stored in my C drive under a directory named Workshop

    b. Type in '`ls`' to check whether your storage has been mounted or not
    c. Your `ls` return should look like this:

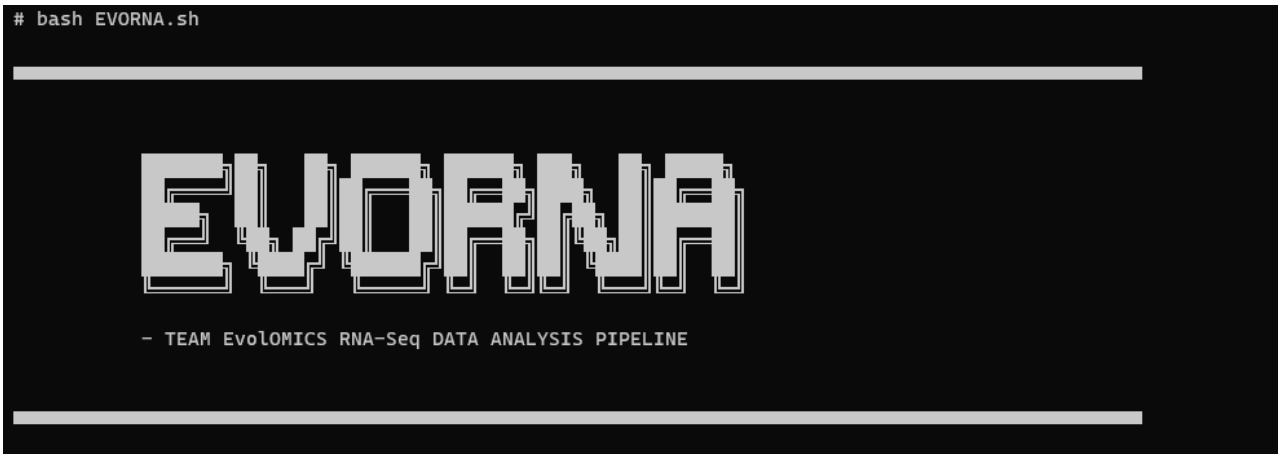


(In the figure above, my files are stored under C://Workshop. Make sure you specify the path to your files correctly, or you would not get an 'ls' return)

3. **Running the script:**
    a. Type in the following command to run the bash script

```
bash EVORNA.sh
```

    b. You should be met with the following screen



4. **Location/Path Inputs. Provide the following paths for the input files when prompted by the pipeline. Confirm by entering Y:**

   a. **Run ID**

```
NCBS_Workshop (or any other ID for the run)
```

   b. **BAM files**

```
/user_data/BAMS/*.bam
```

   c. **GTF file**

```
/user_data/ref_gtf/Oryza_sativa.IRGSP-1.0.55.gtf
```

   d. **deseq_mrna.R**

```
/user_data/deseq_mrna.R
```

   e. **Samples.txt**

```
/user_data/Samples.txt
```

   f. **The location paths should look like the following:**

```
Please specify an ID for this run. It can be a name/number or experiment/project name
ISWCPC
You have entered ISWCPC as your run ID. Do you wish to continue (Y/N)?
Y

Please enter your BAM file location
/user_data/
You have entered /user_data/ as your BAM file location. Do you wish to continue (Y/N)?
Y

Please enter your reference annotation (.gtf) location
/user_data/Ref_gtf.gtf
You have entered /user_data/Ref_gtf.gtf as your reference annotation location. Do you wish to continue (Y/N)?
Y

Please enter the location to the R script 'DESeq2_mRNA.R'
/user_data/deseq_mrna.R
You have entered  as the location to the R script 'DESeq2_mRNA.R'. Do you wish to continue (Y/N)?
Y

Please enter the location to the experimental design file 'Samples.txt'
/user_data/Samples.txt
You have entered /user_data/Samples.txt as the location to the design file 'Samples.txt'. Do you wish to continue (Y/N)?
Y
```

5. **Setting thread count. Type in the following thread count when prompted:**

   a. Please enter the number of threads (We are going to use 6 threads for this hands-on session. However, you can change the number of threads to be used based on your system specifications)

```
6
```

   b. After entering your thread count, confirm by typing in ‘Y’, and you should be met with the following screen:

```
Please enter the maximum number of threads to run this pipeline
6
You have entered 6 as the maximum number of threads. Do you wish to continue (Y/N)?
Y
```

6. On successfully performing all of the above steps, our RNA-Seq data analysis pipeline should start running on your system. It will take some time to complete the run. Ensure you have sufficient battery/power on your system to complete the run.

Reviewing the Results:

1. **On successful run of the pipeline, the following results will be generated:**
    a. Counts.txt (Raw gene counts mapped to reference annotation)
    b. Counts.txt.summary (Summary of gene mapping)
    c. Rplots.pdf (PCA, Corr plots, boxplots and dispersion plots of the differentially expressed genes)
    d. NCBS_Workshop.csv (Unfiltered differential expression csv file)
    e. NCBS_Workshop_significant.csv (Filtered with $p < 0.05$)
    f. NCBS_Workshop_significant_up_mRNA.csv (Filtered with $p < 0.05$ and logFC > 0)
    g. NCBS_Workshop_significant_down_mRNA.csv (Filtered with $p < 0.05$ and logFC < 0)

2. **Open NCBS_Workshop.csv in MS EXCEL/Google Sheets.**
    a. Here, the first column contains Gene IDs. These are the genes that have been calculated to be differentially expressed in *Rhizoctonia solani* infected rice plants as compared to normal rice plants without infection (control). We have already filtered out the statistically significant genes for you.
    b. Further downstream analyses like identification of CRISPR targets, regulatory interaction modeling, ontology and pathway analysis can be carried out using the genes obtained.

Session II: Modeling Gene Regulatory Networks using Cytoscape

Pre-requisites and Downloads:

1. Download the input files from the link given here. [*https://bit.ly/ncbstftg*]
   Please recheck whether you have the downloaded the following files:
    a. TF_TG_edgelist.csv
    b. attributes_TF.csv
    c. attributes_TG.csv

2. Download Cytoscape for windows from here [*https://cytoscape.org/*]
3. Install Cytoscape on your system using the executable file

For UNIX systems, please follow the guide below:

4. Install java if not available. Open terminal on your system and enter the following:

```
sudo apt-get install openjdk-8-jdk
```

5. Download Cytoscape for UNIX from here [*https://cytoscape.org/*]
6. Open terminal from the download directory and enter the following:

```
chmod u+x  Cytoscape_3_9_1_unix.sh
sh Cytoscape_3_9_1_unix.sh
```

Construction of the network:

1. **Launch Cytoscape**
    a. Open **Cytoscape**.
    b. Wait for the interface to fully load.

2. **Import the Network**

    a. **File → Import → Network from File**…
    b. Select TF_TG_network.csv
    c. In the import dialog:

    Source Node Column = DE-TF

Target Node Column = DE-TG

Interaction Type: leave blank or type regulates

d. Click **OK**

3. **Import Node Attribute Tables: Import attributes_TF.csv**
    a. **File → Import → Table → File…**
    b. Select attributes_TF.csv
    c. In the dialog:

    Choose **Node Table** (not Edge)

    Match **Key Column**: DE-TF

    Cytoscape Key Column: name

    d. Click **OK**

4. **Import Node Attribute Tables: Import attributes_TG.csv**

    a. File → Import → Table → File…
    b. Select attributes_TG.csv
    c. Choose **Node Table**
    d. Match **Key Column**: DE-TG
    e. Cytoscape Key Column: name
    f. Click **OK**

Now your nodes are annotated with either attribute_TF or attribute_TG.

5. **Customize the Network Style**
    a. Open the **Style Panel** (right-hand side)
    b. Under **"Fill Color"**:
    c. Click on the **"Column"** dropdown → select attribute_TF or attribute_TG
    d. Set **Mapping Type** to Discrete Mapping
    e. Assign colors:

    DE-TF: red

    DE-TG: blue

    f. **Node Shape** (Optional):

    Use the same attribute_TF/attribute_TG column

    Map shapes (e.g., triangle for TF, ellipse for TG)

    g. **Label the Nodes**:
    h. Under **Label → Column = name**

i. Adjust font size, color, and alignment as needed

6. **Change the Network Layout**
    a. Use a suitable layout:
    b. **Layout → yFiles → Organic** or **Prefuse Force Directed**
    c. Adjust spacing and node sizes for clarity

7. **Explore the Node and Edge Tables**
    a. Switch to the **Table Panel** at the bottom
    b. View:
    Node Table: Confirm imported attributes
    Edge Table: View TF-TG interaction edges

8. **Save Your Work**
    a. **File → Save As** → TF_TG_Cytoscape_Session.cys
    b. Export images:
    File → Export → Network View as Image (PNG, PDF, etc.)

Session III: Molecular Marker Motif Identification and Primer Designing using MISA

Availability and Implementation

MISA-web can be accessed under http://misaweb.ipk-gatersleben.de/. The website provides tutorials, usage note as well as download links to the source code.

URL https://webblast.ipk-gatersleben.de/misa/

The current MISA implementation requires computational expertise and access to a UNIX environment to (i) run the PERL script and (ii) process the results for most downstream applications.

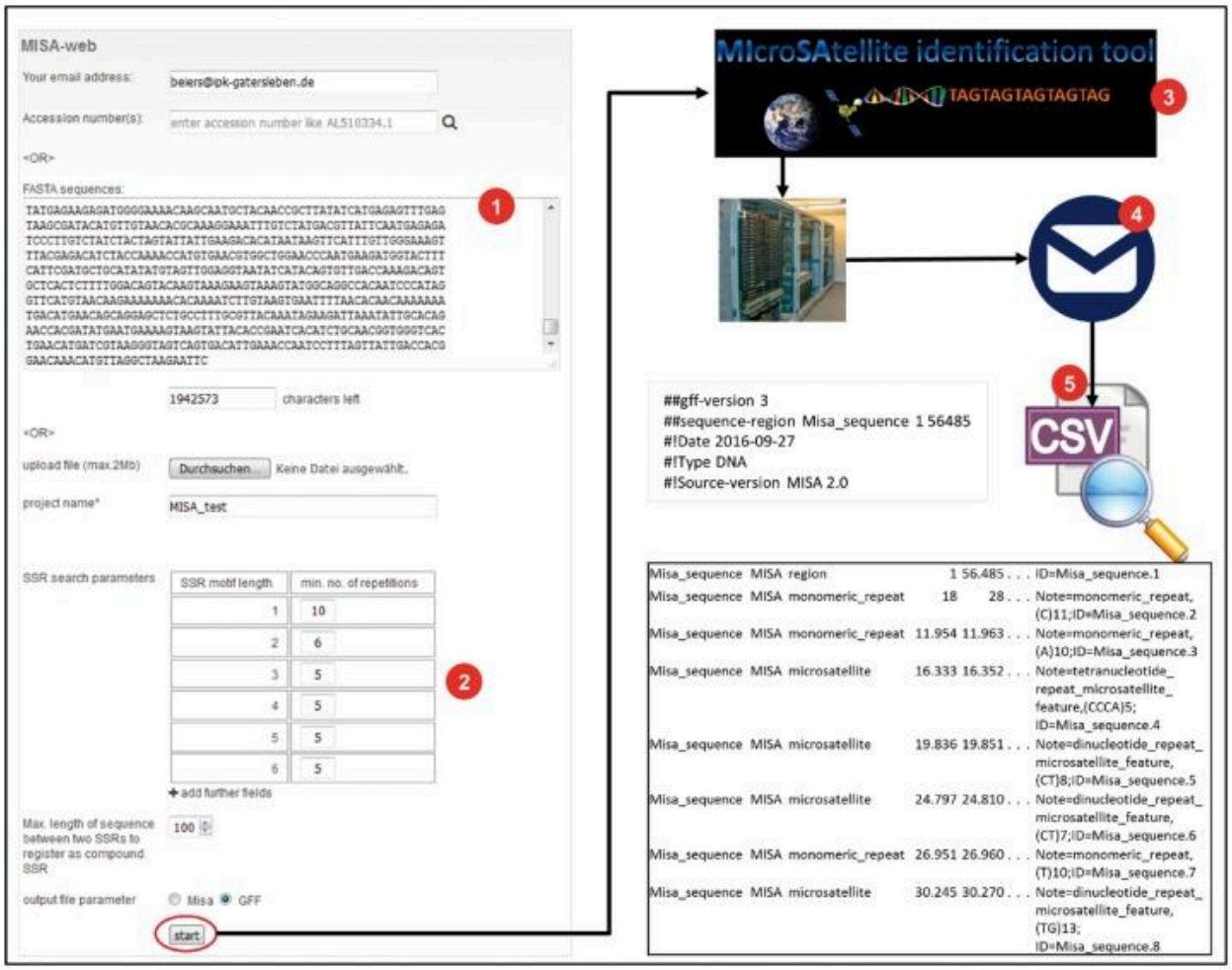
The MISA output contains an overview of identified microsatellites in a proprietary format, which cannot be easily parsed for downstream analysis.

The Generic Feature Format Version 3 (GFF3, https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md) is a commonly used format in genomic data analysis. GFF3 is a tabular format that lists features in nucleotide sequences and provides ontology-based feature classification.

Here, we present the MISA-web, an extension to the command line tool MISA embedded into an easy-to-use web-based graphical user interface available from http://misaweb.ipk-gatersleben.de/.

https://pmc.ncbi.nlm.nih.gov/articles/PMC5870701/

Figure 1: MISA main user interface



MISA-web analysis workflow. MISA-web was updated and set up as a web-application on the IPK server. Users may either paste their nucleotide sequence of interest in the input fields of MISA-web or supply accession numbers to have the corresponding sequences fetched from NCBI (1). Once all input fields have been filled (2), a click on the start button on the bottom of the page starts the analysis. The computation will be conducted on a compute server (3) and the result files will be sent to a user-specified email address (4). Result files can be examined afterwards (5)

Running MISA: Figure 2: An example for SSR search result

Primer designing:
Primer3 web based tool https://primer3.ut.ee/

Primer3 picks primers for PCR reactions, considering as criteria:
oligonucleotide melting temperature, size, GC content, and primer-dimer possibilities, PCR product size.

Ideal primer characteristics for SSR marker design

Primer length (18-24 bp), optimal melting temperature (Tm) between 55-65°C, and a GC content range of 40-70%. Primers should be designed to amplify products between 100-400 bp, with the 3' end of the primer ideally ending in G or C. Additionally, primers should be designed to avoid homodimers.