CIBA – TM Series – 2018 – No. 12

HANDS ON TRAINING

AQUACULTURE GENOMICS AND BIOINFORMATICS

27th August to 1st September, 2018

ICAR-CIBA

CIBA – TM Series – 2018 – No. 12

Hands on Training

AQUACULTURE GENOMICS AND BIOINFORMATICS

Organized By

GENETICS AND BIOTECHNOLOGY UNIT

Prepared by

K. VINAYA KUMAR J. ASHOK KUMAR

MISHA SOMAN RAYMOND J ANGEL B. SIVAMANI

P. MAHALAKSHMI SHERLY TOMY

M. S. SHEKHAR

G. GOPIKRISHNA

भा.कृ.अनु.प. - केन्द्रीय खारा जलजीव पालन अनुसंधान संस्थान
ICAR - CENTRAL INSTITUTE OF BRACKISHWATER AQUACULTURE

ICAR – CENTRAL INSTITUTE OF BRACKISHWATER AQUACULTURE
75, SANTHOME HIGH ROAD, RA PURAM
MRC NAGAR, CHENNAI - 600 028

Published by

Dr. K. K. Vijayan
Director, ICAR-CIBA

TABLE OF CONTENTS

Sr. No.	Chapter Title	Page number
1	Introduction to Linux Environment	1
2	Introduction to programming in R	4
3	Python for Bioinformatics	10
4	Understanding the Illumina datasets	14
5	Checking quality of Illumina paired-end sequence data	17
6	Quality control of RNAseq datasets – NGS QC Toolkit	19
7	Quality control of RNAseq datasets – Trimmomatic	21
8	Assembling bacterial genomes	23
9	RNAseq data analysis in Trinity	25
10	Phylogenomic analysis using MrBayes	31
11	Microsatellites genotypes generation by Fragment analysis method	34
12	Genepop : Population Genetics analysis	38
13	Population genetic analysis of microsatellite data in Arlequin	40
14	SoftComputing techniques inBioinformatics	49
15	RNAseq data analysis – Genome-guided	56
16	Application of ''*OMICS*'' research in aquaculture with special reference to penaeids	64
17	Shrimp Genomics : Current status and Challenges	72
18	Application of Biotechnology in animal reproduction	76
19	Use of molecular techniques in growth enhancement	81
20	Gene Editing Tools and their application in Aquaculture	90
	Glossary	95

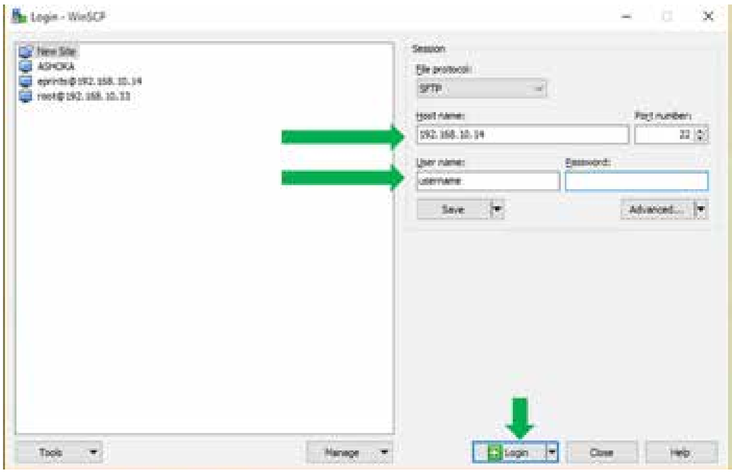
1. Introduction to Linux Environment

J. Ashok Kumar and K. Vinaya Kumar

Opensource operating system (OS) Linux built based on Unix has become choicest OS worldwide for servers as well as desktops in academic circles. There are different varients of Linux which include Redhat, Ubuntu, fedora, CentOS, knoppix etc. Many of the bioinformatics software and individual programs are native to linux OS. So it is important for a bioinformatician to have exposure to linux commands. Here we give a list of most commonly used linux commands and procedure to execute perl /python programmes. As advanced programming is beyond the scope of this training, we provide here the basic constructs of perl/python programs which could be used for writing scripts for simple bioinformatics tasks.

Linux commands

Accessing linux environment: You can access linux server using any windows based ssh client from your system. This could be achieved by installing winSCP or Putty (both are free software) on your system. Once installed open WinSCP, fill in the Host name, user name and password columns provided by system administrator and click on login button which will prompt for password. After successful login and selecting putty from menubar, console window pops upand you will see a dolloar prompt where in you can submit commands for all the operations you wish to perform on linux server.

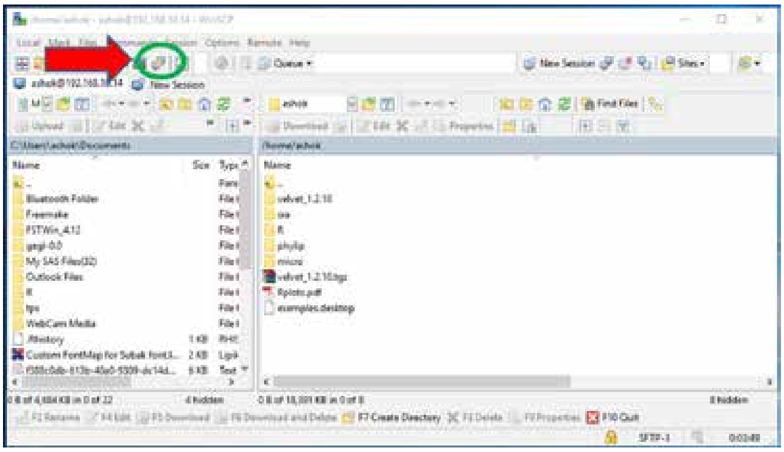


Login - WinSCP

New Site
ASHOKA
eprints@192.168.10.14
root@192.168.10.33

Session
File protocol: SFTP
Host name: 192.168.10.14
Port number: 22
User name: username
Password:

Figure 1. WinSCP login window



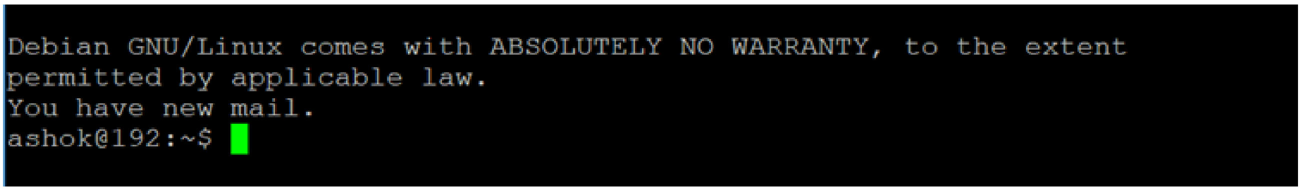
C:\Users\ashok\Documents

/home/ashok

Bluetooth Folder
Freemake
FSTWin_4.12
gargi-0.0
My SAS Files(32)
Outlook Files
R
tps
WebCam Media
.Rhistory

velvet_1.2.10
ora
R
phylip
micro
velvet_1.2.10.tgz
Rplots.pdf
examples.desktop

Figure 2. Selecting Putty from winSCP



```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
You have new mail.
ashok@192:~$
```

Figure 3. Linux console

The dollar prompt ($) shown in Fig. 3 is for users and the hash (#) prompt will be displayed for administrators. Users who have the administrative privileges on the server can only work with hash (#) prompt.

File system in linux: All the folders and files of the linux system will be under root (/) directory. Users will have access to their home directories for which the path is /home/user_name

Once you login to the linux system by default you will be taken to your home directory. For example is if user name is "**david**" after login into Linux the current directory which he will be accessing is /home/david. Users can input their commands after the dollar ($) prompt. Some of the most commonly used linux commands are given in the table below.

Function	Command
Listing the file names	$ls
Listing with file names along with other details	$ls –l
Change to preexisting directory by name 'test'	$cd test
Make a new directory by name 'trial'	$mkdir trial
Viewing a preexisting file	$vi mydata.txt $nano mydata.txt $more mydata.txt $cat mydata.txt
Creating a new file	$touch myfile.txt $vi myfile.txt $nano myfile.txt
Renaming or moving the file	$mv file1.txt file2.txt $mv /home/ram/file1.txt /home/ram/test/
Making duplicate of file	$cp file1.txt file2.txt $cat file1.txt > file2.txt
Appending two text files	$cat file1.txt file2.txt > file3.txt
To display date	$date
To find number of lines in a file	$wc –l xyz.txt
To display first (top) 100 lines of a file	$head -100 xyz.txt
To display last (bottom) 100 lines of a file	$tail -100 xyz.txt
Search for a pattern in a file	$grep "pattern" file.txt
Search for pattern at beginning of line	$grep '^pattern' file.txt
Search for pattern at the end of a line	$grep 'pattern$' file.txt
Search for only pattern in the line	$grep '^pattern$' file.txt

Running perl /python programs

Perl program files will have extension ".pl". Command to execute the programmes is

```
$ ./test_programme.pl
```

Or

```
$perl test_programme.pl
```

Options of the program may be checked from the help files of the software/programs.

Same way python program files will have ".py" extensions and they could be executed by giving following command.

```
$python test_programmes.py
```

Standalone blast

NCBI Blast is used for comparing nucleotide and protein sequences with the sequence databases to find significant matches. Alignment of sequences using blast can be done either by using web-tool available on NCBI site or by installing blast on local servers.

Blast can be installed on local servers along with the databases available in public domain. In addition, users can make their own databases on local servers. If you have your own protein dataset then local databases can be created by

```
$makeblastdb -in xyz.fasta -dbtype 'prot' -out xyzdb
```

Now you can run the blast using your own database

```
$blastp -db xyzdb -query abc.fasta –out out.fasta
```

More general blast Command

```
$blastn -query nucl.fasta -db xyzdb -outfmt 6 -evalue 1e-05 -out output.txt
```

For fetching the sequences in fasta file format from output make a file with IDs of hits and run the following command

```
fastacmd -d database_name -i blast_output > hits.fasta
```

2. Introduction to programming in R

J. Ashok Kumar, K. Vinaya Kumar and B. Sivamani

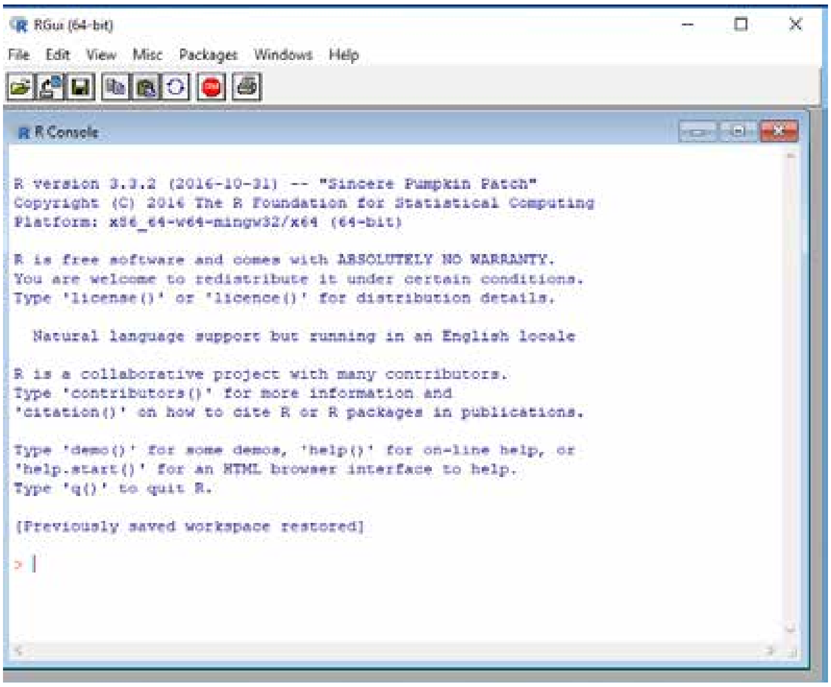
R is a programming environment for data analysis and graphics. The language was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics at the University of Auckland. Since its birth, a number of people have contributed to the package. It is open source statistical software which can be downloaded free of cost. Base package and all the contributory packages could be downloaded from http://www.r-project.org/

R is available for all operating systems like windows, Linux and Mac OS. This training material is based on R stats package installed in windows operating system.

Invoking R stats

Start → All programmes → R → R i386 3.2.0 (for 32 bit installation)

Start → All programmes → R → R x64 3.2.0 (for 64 bit installation)



```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> |
```

R Stats Graphical user interface in windows

Procedure to install additional packages

We need to add additional libraries to Base installation to utilize full potential of R. This can be achieved by following command.

Install.packages('name of the package')

Once the above command is executed R system asks the user to select a CRAN mirror out of several listed mirrors. User can select mirror of any location.

There is a package/library called 'Rcmdr' which can be used for carrying out most commonly used statistical procedure with graphical user interface. The command to install 'Rcmdr' is

Install.packages('Rcmdr')

Command to invoke the Rcmdr

Library('Rcmdr')

R studio

R studio is integrated development environment(IDE) for R. This IDE features R notebook for writing scripts, console for command input, graphics viewer, package window and environment window all in single framework.

R files input and output.

First set the working directory

Command to know the location of present working directory is

- **getwd()**

Command to set the working directory to any other folder

- **setwd("E:/data/")**

Basic command to read the files is

- **read.table()**

and command to create the data files is

- **write.table()**

Importing data

Data with different file formats i.e., text files, excel files, SPSS data files, SAS data files etc., can be input into R stats for data analysis. It is advised that excel files may first be converted to comma separated files for easy input into R stats.

Command to read a comma separated text file with variable names in the first row

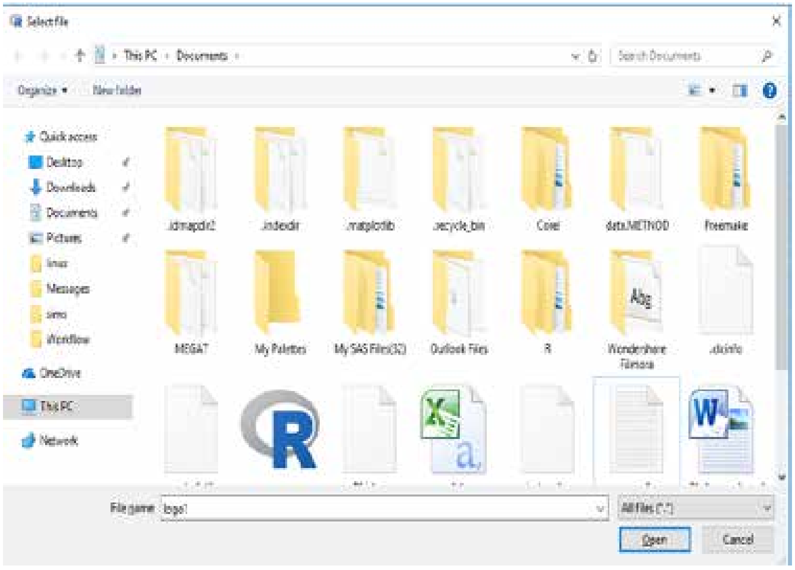
- **Data <- read.table('filename', header=TRUE, sep=",")**

Here filename is name of the text file with extension, header statement is to specify whether variable names are included in the first row of the data file and 'sep' parameter tells the separator present between variables (columns) like comma, space, tab etc., in the file.

If the specified text file is not in present working directory and you wish to select it though graphical interface use the following command

- **Data <- read.table(file.choose(), header=TRUE, sep=",")**

Upon entering the above command a file selector window will pop up and one can select the file located at any drive/directory/folder other than the present working directory.



Popup window for selecting files

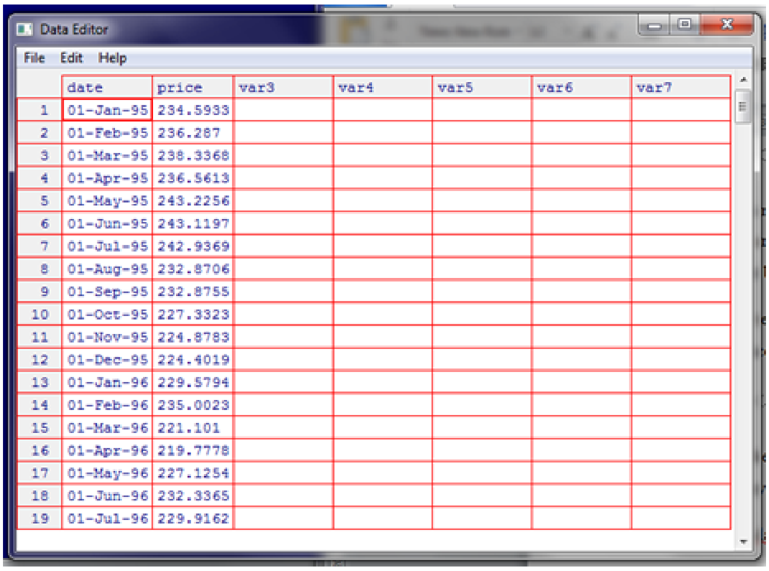
For other text files like space separated and tab separated one need to change only **'sep'** parameter of the above command with either " " or " \t ".

In the previous command '*data'* is a dataframe which will contain all the variable names and data

Data in the dataframe can be edited and assigned the changed file contents to other dataframe

➢ **data1<- edit(data)**

Upon entering the above command a popup window appears for editing the data and all the edits will be saved in data frame called 'data1'



	date	price	var3	var4	var5	var6	var7
1	01-Jan-95	234.5933					
2	01-Feb-95	236.287					
3	01-Mar-95	238.3368					
4	01-Apr-95	236.5613					
5	01-May-95	243.2256					
6	01-Jun-95	243.1197					
7	01-Jul-95	242.9369					
8	01-Aug-95	232.8706					
9	01-Sep-95	232.8755					
10	01-Oct-95	227.3323					
11	01-Nov-95	224.8783					
12	01-Dec-95	224.4019					
13	01-Jan-96	229.5794					
14	01-Feb-96	235.0023					
15	01-Mar-96	221.101					
16	01-Apr-96	219.7778					
17	01-May-96	227.1254					
18	01-Jun-96	232.3365					
19	01-Jul-96	229.9162					

Data editor window

Exporting data

Data in the dataframe can be exported as a text file with the following command

➢ **write.table(data, file="xyz.csv", col.names=TRUE, sep=",")**

Creating data files manually within Rstats

Data files can be created within Rstats by giving simple commands

Here we explain creating example table with variable names into R stats

S.No	Bodyweight	Length	Species
1	25	15	aa
2	35	14	ab
3	65	27	ac
4	27	18	bb
5	45	22	cc

The above table can be created as a dataframe by giving the following commands

- bodyweight <- c(25,35,65,27,45)
- length <- c(15,14,27,18,22)
- species<-c(“aa”,”ab”,”ac”,”bb”,”cc”)
- lengthweight <-cbind(bodyweight,length,species)

Descriptive statistics

Suppose we have a variable by name ‘x’ and our task is to calculate all the descriptive statistical parameters like mean, median, standard deviation, variance etc. for the variable x in R stats. First create a variable x by giving the following command

- x <- c(20,15,19,22,26,24,23,17,18,22)

Other way of creating variable ‘x’ is

- x <- scan()

```
1: 20 15 19 22 26 24 23 17 18 22

11:

Read 10 items
```

Basic commands for descriptive statistics

- mean (x) # mean
- median (x) # median
- var (x) # sample variance
- sd(x) # sample std. deviation
- quantile (x,p) # sample quantile , p could be 0.25, 0.5,0.75
- min (x) # minimum of x
- max (x) # maximum of x
- range () # range of x
- library(e1071)

- skewness (x) # skewness
- kurtosis (x) # kurtosis

Commands for statistical tests

Single sample t-test

- t.test(y,mu=10)

  here y is a variable; mu is population mean

  Two sample t-test

- t.test(y1,y2,var.equal=TRUE)

  y1 and y2 are the two independent samples

Paired t-test

- t.test(y1,y2,paired=TRUE)

  y1 and y2 are the two paired samples

Chi-square test for goodness of fit

- n<- cbind(y1,y2)
- chisq.test(n)

  n is a datamatrix /contingency table

Correlation

- n <- cbind(y1,y2) # create dataframe n
- cor(n)

  where y1 and y2 are two variables and n is matrix of y1 and y2

Regression

- fit <- lm(y~x)

  for multiple regression

- fit <- lm(y~x1+x2+x3)

Completely randomised design

- tr <- c(1,1,1,2,2,2,3,3,3)   # create treatment variable
- yield<-c(25,41,54,65,45,65,25,12,35) # create dependent variable
- fit <- aov(yield ~ factor(tr))   # model statement
- summary(fit)

Randomised Block Design

- tr <- c(1,1,1,2,2,2,3,3,3) # create treatment variable
- rep <-c(1,2,3,1,2,3,1,2,3) # create replication variable

- yield<-c(25,41,54,65,45,65,25,12,35) # create dependent variable
- fit <- aov(yield ~ factor(tr) + factor(rep))
- summary(fit)

Two way factorialDesign

- fit <- aov(yield ~ factor(A) + factor(B) + factor(A) : factor(B)   + factor(rep))
- summary(fit)

Installing Bioconductor in R

Enter following commands in R console to install bioconductor packages.

source (http://bioconductor.org/biocLite.R**)**

biocLite()

Steps in manipulating fasta files

First load library

- library(seqinr)

Set working directory in R where fasta files are loaded

- setwd("c:/path/to/directory")
- seq1 <- read.fasta("sequence.fasta")
- seq1.seq<- seq1[[1]]   # to take the sequence from fasta file
- length(seq1.seq) # to find length of the sequence (bases)
- table(seq1) # to find frequency of each base
- GC(seq1.seq) # to find the GC content of the sequence

There are several advanced options are available in R ranging from simple sequence analysis to microarray data analysis. Purpose of this chapter is to introduce the R environment and to provide hands-on for exploring the functionalities available in R.

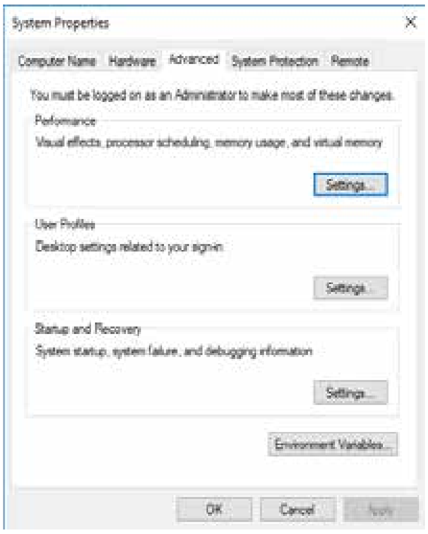
3. Python for Bioinformatics

J. Ashok Kumar and K. Vinaya Kumar

Python is one of the most popular high level general purpose programming languages. It was developed in the year 1991 by Guido van Rossum, a Dutch programmer. It is an open source programming language available for download at www.python.org. In recent years it has gained lot of importance due to development of several libraries applicable to various fields of research and development. One such library widely used in Bioinformatics is BioPython. Here we introduce python environment for writing scripts and provide a glimpse of Biopython functionalities.

Installation of Python

Python is available for both windows and Linux platforms. Windows / Linux binaries can be obtained from www.python.org. In windows you may double click on the exe file and accept the default installation settings to get it installed in the system. Once installed go to edit environment variable Advanced environment variables and add new python path as show in the figure



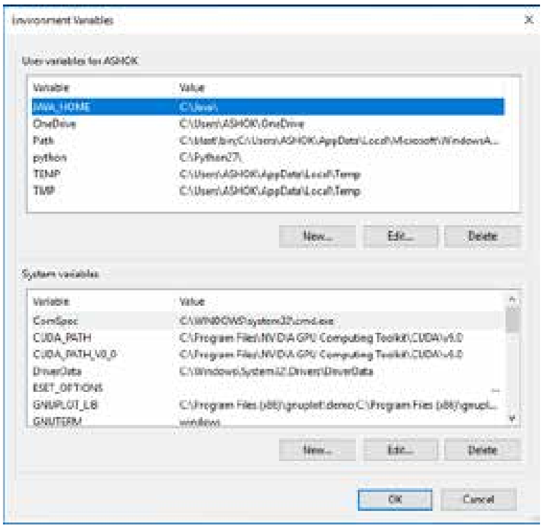
System Properties

You must be logged on as an Administrator to make most of these changes.

Performance
Visual effects, processor scheduling, memory usage, and virtual memory

User Profiles
Desktop settings related to your sign-in

Startup and Recovery
System startup, system failure, and debugging information



Environment Variables

User variables for ASHOK

Variable	Value
JAVA_HOME	C:\Java\
OneDrive	C:\Users\ASHOK\OneDrive
Path	C:\Matl\bin;C:\Users\ASHOK\AppData\Local\Microsoft\WindowsA...
python	C:\Python27\
TEMP	C:\Users\ASHOK\AppData\Local\Temp
TMP	C:\Users\ASHOK\AppData\Local\Temp

System variables

Variable	Value
ComSpec	C:\WINDOWS\system32\cmd.exe
CUDA_PATH	C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v8.0
CUDA_PATH_V8_0	C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v8.0
DriverData	C:\Windows\System32\Drivers\DriverData
ESET_OPTIONS	
GNUPLOT_LIB	C:\Program Files (x86)\gnuplot\demo;C:\Program Files (x86)\gnupl...
GNUTERM	windows

Now you can open command line interface in windows by entering 'cmd' search box on the taskbar and enter.

On most of the Linux installations python comes with default installation. If not available it can be installed on debian/Ubuntu systems by keying-in the following command

```
$sudo apt-get install python
```

Installing pip

Pip is package manager for python. To install pip download get-pip.py from https://pip.pypa.io/en/stable/installing/ and enter the following command.

```
$Python get-pip.py
```

Once pip is installed, any python package can be installed by the following command

```
$pip install 'package-name'
```

Installing Jupyter

Jupyter is notebook applications for python wherein one can write scripts, execute the scripts and save the notebooks in different formats like pdf, doc for future use. Run following commands for installing and opening the jupyter notebook

```
$pip install jupyter # install the jupyter package

$python –m IPython notebook ## Opening notebook in windows.

$jupyter notebook ## opening notebook in Linux
```

One can install required additional packages like matplotlib for plotting the graphs, numpy for numerical calculations pandas for data structures and data analysis tools, statmodels for statistical analysis, scipy for mathematical & scientific applications. All these can be installed using python.

Introduction to python programming

- Print “hellow world” ## printing a text

  Hellow world
- text1 = “CIBA” # text1 is a string variable
- a = 20 # b is a numeric variable having value 20
- b = 30
- a+b

  50
- b-a

  10
- a*b

  600
- a/b
- 0
- a/float(b)

  0.666
- a**b # which is a to the power of

  1.073741824e+39

For mathematical functions

- import math
- math.log(a)

  2.995

- math.cos(a)

  0.408

Araay in python

- a =[]
- a = [“hi”,”this”,”is”,”python”]
- a[2]

Declaring dictionary

- dict1={“apple”: 250,”banana”: 100,”cherry”: 300}
- dict1.keys()

  [‘cherry’, ‘apple’, ‘banana’]

- dict1.values()

  [300, 250, 100]

- dict1[“cherry”]

  300

Programming loops

- for i in range(0,10):

  ```
      print i
  ```

- j=1
- while (j < 10):

  ```
      print j
      j=j+1
  ```

Functions

- def f2c(x):

  ```
      return (x-32)*5/9.0
  ```

Read and write files

- inp=open(“input.txt”,’r’)

  ```
  out=open(“output.txt”,’w’)
  for line in inp:
      if line[0]==”>”:
          out.write(line)
  inp.close()
  out.close()
  ```

Biopython

Biopython is the set of computational methods used for Bioinformatics analysis. Biopython can be used to parse different files like fasta, blast output, genbank, expasy; execute online tools like NCBI blast, entrez etc., code to sequence alignment, multiple sequence alignment, phylogeny and even machine learning classification methods like naïve bayes, knearest neighbourhood, support vector machines etc.,. Biopython library can be installed through pip installation method.

- pip install biopython (or python –m pip install biopython in windows)
- import Bio
- from Bio.Seq import Seq
- seq1 = Seq(“ATGCGGATC”)

  Seq(‘ATGCGGATC’, Alphabet())
- seq1.complement()

  Seq(‘TACGCCTAG’, Alphabet())
- seq1.reverse.complement()

  Seq(‘GATCCGCAT’, Alphabet())

Parsing fasta file

- from Bio import SeqIO
- for seq_record in SeqIO.parse(“sequence.fasta”, “fasta”):

  print(seq_record.id)

  print(repr(seq_record.seq))

  print(len(seq_record))

Different system commands can also be executed from python using following commands

- import os
- com = “blastn – query seq.fasta –db nr –out out.txt”
- os.system(com)

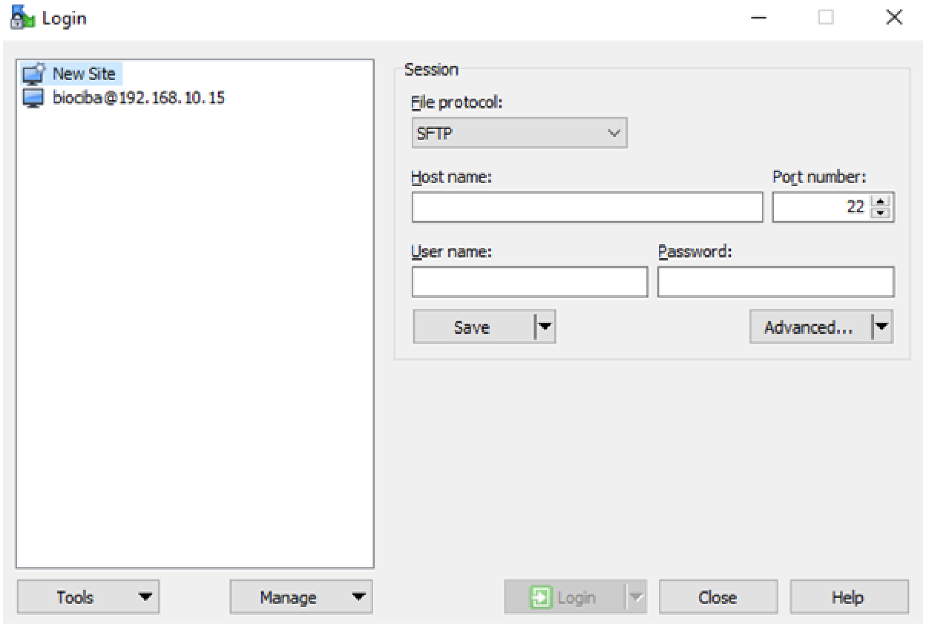
4. Understanding the Illumina datasets

K. Vinaya Kumar and J. Ashok Kumar

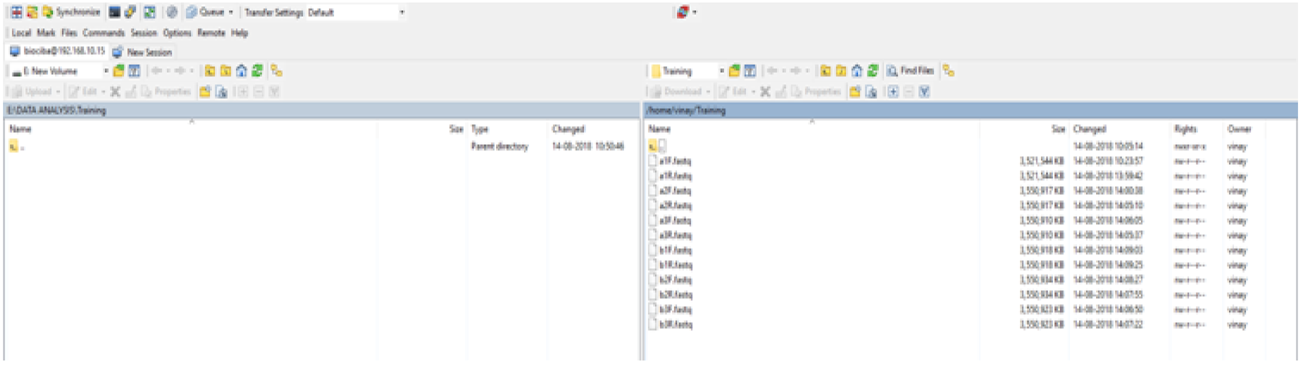
With continual improvements being made over past few years, the Next Generation Sequencing (NGS) platforms came a long way in generating enormous sequence data at low cost and less time. Many NGS platforms like Illumina, Pacbio, Nanopore, Ion Torrent etc are well-known platforms with several published manuscripts quoting usage of them. A feature common to all these platforms is massively parallel sequencing of single or clonally amplified DNA molecules. Of different platforms available till date, the one offered by Illumina stands apart in terms of the amount of sequence data generated and the cost involved. In case of Illumina, right from the Genome Analyzer IIx, the HiSeq XXXX series, the MiSeq, the NextSeq XXX series to the latest NovaSeq 6000, there is an improvement in data output while reducing the sequencing time.

There are two popular sequencing chemistry of Illumina platform namely, paired-end (PE) and mate-pair (MP) that are commonly used by researchers. The PE sequencing is used for RNAseq studies where we find differentially expressed transcripts in experimental samples compared to control sample. The MP sequence reads are mostly used in assembly of whole genomes where they play an important role in scaffolding the contigs. In this chapter we understand the structure of paired-end sequence datasets generated on Illumina platform. The raw sequence data files generated on Illumina platform are delivered as '.fastq' files. For every sample, two files are provided, one read_1 or forward sequence read and the other read_2 or reverse sequence read. The order of reads in forward and reverse sequence reads files should not be altered as they are linked.

Open the WinSCP tool. The following window appears. Enter the host name as told by the tutor. Enter the 'user name' and 'password' to log in to your account.



After logging in, the window of WinSCP tool appears. The window has two panels. The left panel is the file system of your computer. The right panel is the file system of your account in server.



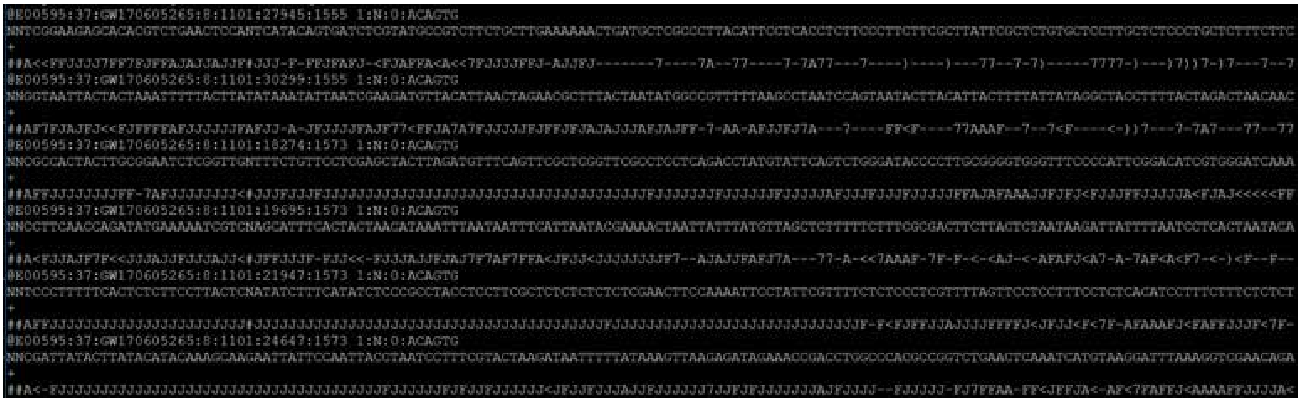
E:\DATA ANALYSIS\Training

Name	Size	Type	Changed
..		Parent directory	14-08-2018 10:50:46

/home/vinay/Training

Name	Size	Changed	Rights	Owner
..		14-08-2018 10:05:14	rwxr-xr-x	vinay
a1F.fastq	1,521,544 KB	14-08-2018 10:23:57	rw-r--r--	vinay
a1R.fastq	1,521,544 KB	14-08-2018 13:59:42	rw-r--r--	vinay
a2F.fastq	1,550,917 KB	14-08-2018 14:00:38	rw-r--r--	vinay
a2R.fastq	1,550,917 KB	14-08-2018 14:05:10	rw-r--r--	vinay
a3F.fastq	1,550,910 KB	14-08-2018 14:06:05	rw-r--r--	vinay
a3R.fastq	1,550,910 KB	14-08-2018 14:05:37	rw-r--r--	vinay
b1F.fastq	1,550,918 KB	14-08-2018 14:09:03	rw-r--r--	vinay
b1R.fastq	1,550,918 KB	14-08-2018 14:09:25	rw-r--r--	vinay
b2F.fastq	1,550,934 KB	14-08-2018 14:08:27	rw-r--r--	vinay
b2R.fastq	1,550,934 KB	14-08-2018 14:07:55	rw-r--r--	vinay
b3F.fastq	1,550,923 KB	14-08-2018 14:06:50	rw-r--r--	vinay
b3R.fastq	1,550,923 KB	14-08-2018 14:07:22	rw-r--r--	vinay

Click on the icon displaying 'two connected computers' in the top toolbar to open the putty window. In this window you run your jobs in server. Enter the log in credentials on prompt. Then browse to the folder where a file with extension '.fastq' is present. Then type-in the command 'head file.fastq' to see the first few lines of file.



```
@E00595:37:GW170605265:8:1101:27545:1555 1:N:0:ACAGTG
NNTCGGAAGAGCACACGTCTGAACTCCANTCATACAGTGATCTCGTATGCCGTCTTCTGCTTGAAAAAACTGATGCTCGCCCTTACATTCCTCACCTCTTCCCTTCTTCGCTTATTCGCTCTGTGCTCCTTGCTCTCCCTGCTCTTTCTTC
+
##A<<FFJJJJ7FF7FJFFAJAJJAJJF#JJJ-F-FFJFAFJ-<FJAFFA<A<<7FJJJJFFJ-AJJFJ-------7----7A--77----7-7A77---7----)----)---77--7-7)-----7777-)---)7))7-)7---7--7
@E00595:37:GW170605265:8:1101:30299:1555 1:N:0:ACAGTG
NNGGTAATTACTACTAAATTTTTACTTATATAAATATTAATCGAAGATGTTACATTAACTAGAACGCTTTACTAATATGGCCGTTTTTAAGCCTAATCCAGTAATACTTACATTACTTTTATTATAGGCTACCTTTTACTAGACTAACAAC
+
##AF7FJAJFJ<<FJFFFFAFJJJJJJFAFJJ-A-JFJJJJFAJF77<FFJA7A7FJJJJJFJFFJFJAJAJJJAFJAJFF-7-AA-AFJJFJ7A---7----FF<F----77AAAF--7--7<F----<-))7---7-7A7---77--77
@E00595:37:GW170605265:8:1101:18274:1573 1:N:0:ACAGTG
NNCGCCACTACTTGCGGAATCTCGGTTGNTTTCTGTTCCTCGAGCTACTTAGATGTTTCAGTTCGCTCGGTTCGCCTCCTCAGACCTATGTATTCAGTCTGGGATACCCCTTGCGGGGTGGGTTTCCCCATTCGGACATCGTGGGATCAAA
+
##AFFJJJJJJJJFF-7AFJJJJJJJJ<#JJJFJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJFJJJJJJFJJJJJAFJJJFJJJFJJJJJFFAJAFAAAJJFJFJ<FJJJFFJJJJJA<FJAJ<<<<<FF
@E00595:37:GW170605265:8:1101:19695:1573 1:N:0:ACAGTG
NNCCTTCAACCAGATATGAAAAATCGTCNAGCATTTCACTACTAACATAAATTTAATAATTTCATTAATACGAAAACTAATTATTTATGTTAGCTCTTTTTCTTTCGCGACTTCTTACTCTAATAAGATTATTTTAATCCTCACTAATACA
+
##A<FJJAJF7F<<JJJAJJFJJJAJJ<#JFFJJJF-FJJ<<-FJJJAJJFJAJ7F7AF7FFA<JFJJ<JJJJJJJJF7--AJAJJFAFJ7A---77-A-<<7AAAF-7F-F-<-<AJ-<-AFAFJ<A7-A-7AF<A<F7-<-)<F--F--
@E00595:37:GW170605265:8:1101:21947:1573 1:N:0:ACAGTG
NNTCCCTTTTTCACTCTCTTCCTTACTCNATATCTTTCATATCTCCCGCCTACCTCCTTCGCTCTCTCTCTCTCGAACTTCCAAAATTCCTATTCGTTTTCTCTCCCTCGTTTTAGTTCCTCCTTTCCTCTCACATCCTTTCTTTCTCTCT
+
##AFFJJJJJJJJJJJJJJJJJJJJJJJJ#JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJF-F<FJFFJJAJJJJFFFFJ<JFJJ<F<7F-AFAAAFJ<FAFFJJJF<7F-
@E00595:37:GW170605265:8:1101:24647:1573 1:N:0:ACAGTG
NNCGATTATACTTATACATACAAAGCAAGAATTATTCCAATTACCTAATCCTTTCGTACTAAGATAATTTTTATAAAGTTAAGAGATAGAAACCGACCTGGCCCACGCCGGTCTGAAGTCAAATCATGTAAGGATTTAAAGGTCGAACAGA
+
##A<-FJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJFJFJJFJJJJJJ<JFJJFJJJAJJFJJJJJJ7JJFJFJJJJJJJAJFJJJJ--FJJJJJ-FJ7FFAA-FF<JFFJA<-AF<7FAFFJ<AAAAFFJJJJA<
```

You find that, the information about each sequence read is represented in four lines.

Line 1: has information about instrument ID, run ID, flow cell ID, lane ID, tile ID, X and Y coordinates of clusters, read number, status about the read is filtered or not and control sample status etc.

Line 2: the sequence of the read which is the familiar A, T, G and C

Line 3: a plus (+) sign

Line 4: the quality scores of the sequence bases

You may visit the following page to understand more about the quality scores.

https://www.illumina.com/documents/products/technotes/technote_understanding_quality_scores.pdf

The symbols in line 4 represent quality scores of bases. The quality scores ranges from 0 to 40. A score of 40 indicates that the base called is of high quality. In this case, the error probability infers that one base call in 10,000 base calls would be incorrect. The following table illustrates the relation between the symbols and the corresponding quality scores.

Table. List of symbols corresponding toquality scores of bases in Illumina sequence datasets.

Symbol	Quality Score		Symbol	Quality Score
!	0		6	21
"	1		7	22
#	2		8	23
$	3		9	24
%	4		:	25
&	5		;	26
’	6		<	27
(	7		=	28
)	8		>	29
*	9		?	30
+	10		@	31
,	11		A	32
-	12		B	33
.	13		C	34
/	14		D	35
0	15		E	36
1	16		F	37
2	17		G	38
3	18		H	39
4	19		I	40
5	20			

5. Checking quality of Illumina paired-end sequence data

K. Vinaya Kumar and J. Ashok Kumar

Illumina paired-end (PE) sequencing reads are commonly used for RNAseq studies and assembling of genomes. For each sample, the sequencing machine prints output data in two paired .fastq files. In this chapter, we discuss about the quality issues pertaining to PE reads. A better understanding of these helps in better planning of read processing to extract quality data for further studies.

One of the basic software useful to understand the quality of PE reads file is ‘FastQC’. Visit the following site to download the latest version of software.

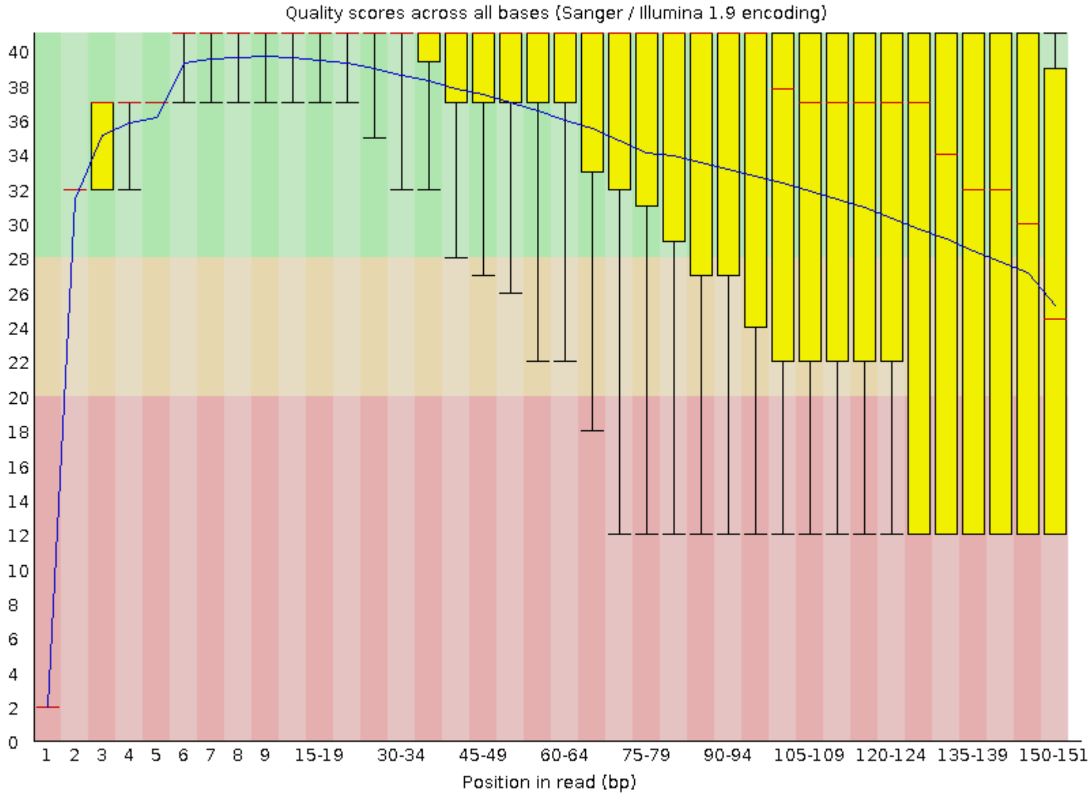
https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc

First, *log in* to your account using WinSCP tool. Open PuTTY SSH terminal. In your account, find a file named, a1F.fastq. We shall check the quality of this file using FastQC tool. To do this, run the following command at your prompt.

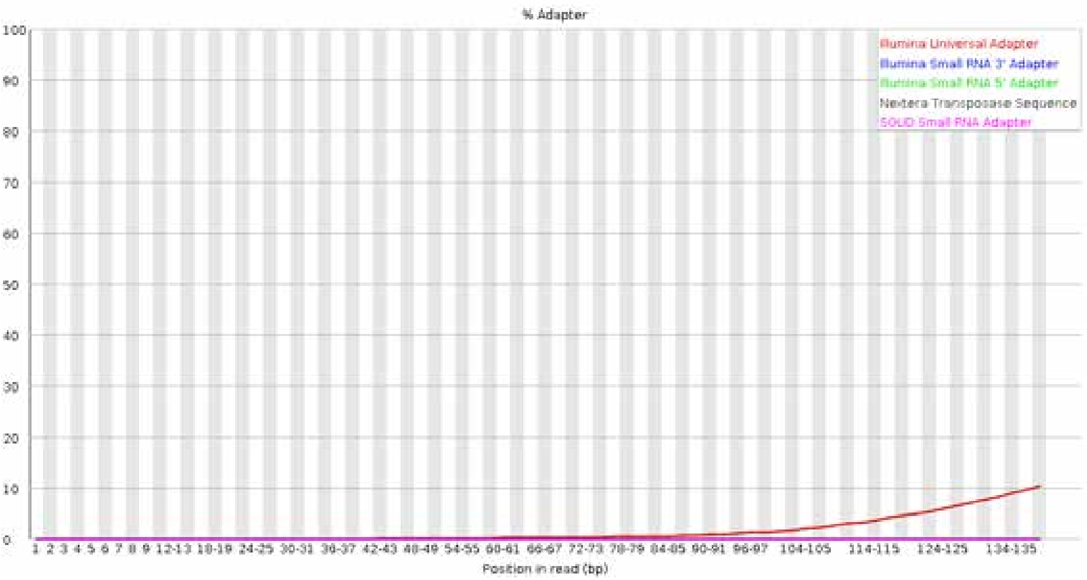
```
$ fastqc<space> a1F.fastq
```

In less than two minutes, the analysis would be completed and two output files are printed, a1F_fastqc.html and a1F_fastqc.zip. Save these files to your computer and open the .html file in any browser. Check all images and understand their meaning. Observe carefully for the following aspects in the file.

Box plot of quality scores along the sequence read length.



5.1. The reads are contaminated with adapter sequences used during sequencing.



The quality report warrants us to do some data processing which includes,

1. Removal of poor quality reads that are pulling down the average of quality scores.
2. Removal of poor quality bases at the start and end of sequence reads.
3. Removal of adapter sequences contaminating the reads.

6. Quality control of RNAseq datasets – NGS QC Toolkit

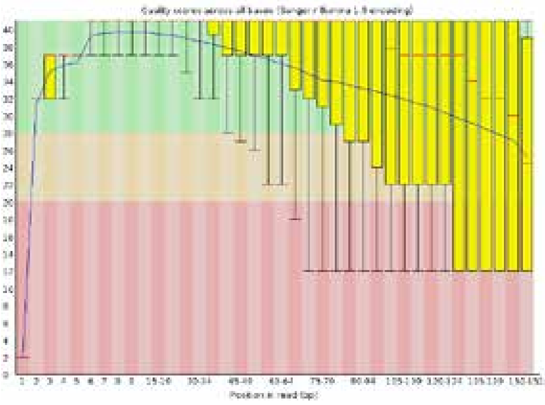
K. Vinaya Kumar and J. Ashok Kumar

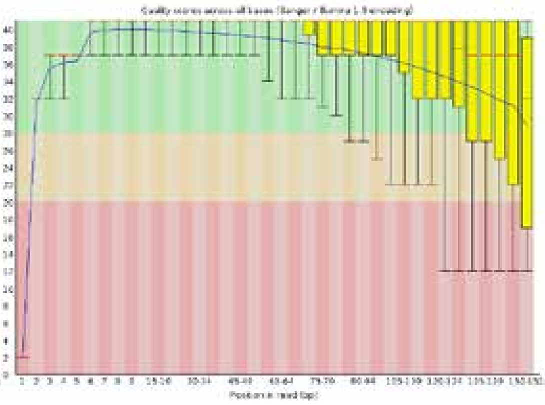
There are several freeware available for processing of paired-end sequence reads. In this chapter we shall use NGS QC Toolkit for quality control of PE reads. First, *log in* to your account using WinSCP tool. Open PuTTY SSH terminal. In your account, find two files named, a1F.fastq and a1R.fastq. Check the quality of both the paired files using FastQC tool. Practice the following quality control steps and observe the changes in quality of trimmed files.

6.1 Discarding low quality reads

```
perl<>IlluQC_PRLL.pl<> -pe <> a1F.fastq <>a1R.fastq <> 2<> A <>-l <> 70 –s<> 20<> -c<> 50
```

This command removes all those reads where the proportion of bases having a quality of > 20 is less than 70%. After the run, find that a folder 'IlluQC_Filtered_files' is printed. The trimmed files are present in this folder. Do quality check of these two files with FastQC. Observe the changes in reads file after running this command.



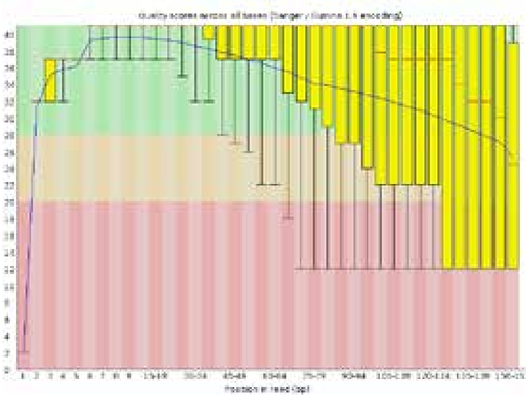


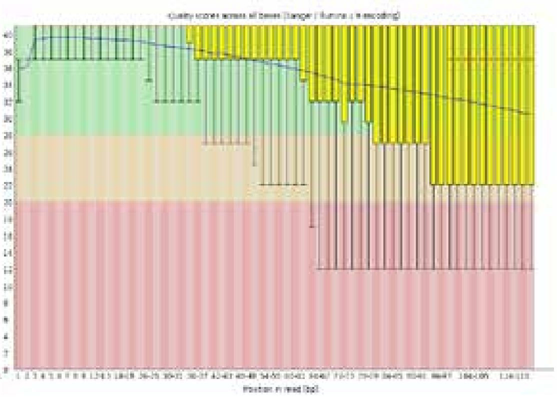
After discarding about 3 million reads completely, the average quality of bases improved. Therefore the improvement in quality came at the expense of losing about 30 % of sequence reads.

6.2 Discarding poor quality bases at both ends based on length.

```
perl<>TrimmingReads.pl<> -i <>a1F.fastq<> -irev<> a1R.fastq<> -l <>3 <> -r <> 30
```

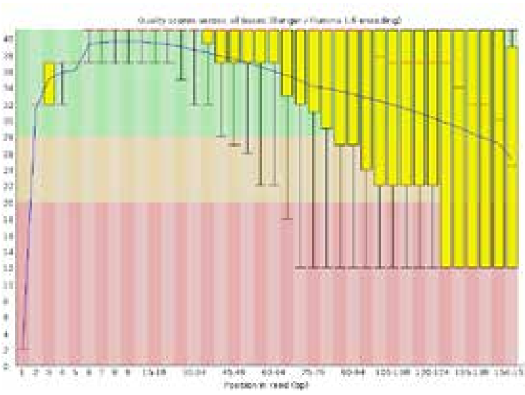
This command removes 3 bases at 5' end and 30 bases at 3' end from all reads.

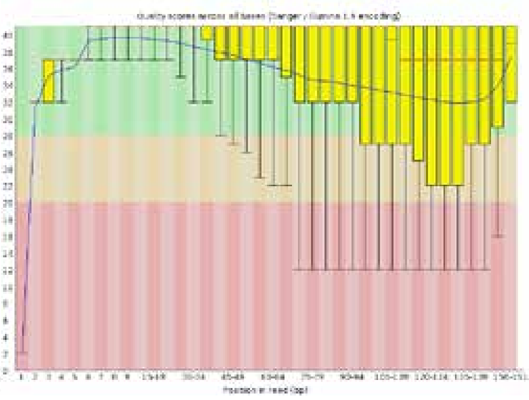




6.3 Discarding poor quality bases at 3' end of reads based on quality score

```
perl<>TrimmingReads.pl<> -i <>a1F.fastq<> -irev<> a1R.fastq<> -q <>30
```





This command removes bases at 3' ends where the base quality is <30. This improvement in quality at ends came at the expense of some reads getting shorter.

6.4 Discarding reads based on read length

```
perl<>TrimmingReads.pl<> -i <>a1F_7020.fastq<> -irev<> a1R_7020.fastq<> -n <>25
```

This command removes reads shorter than 25 bases length.

A combination of these could be chosen and applied based on the initial base quality of sequence datasets. Extract only the good quality data for downstream processing of reads.

7. Quality control of RNAseq datasets – Trimmomatic

K. Vinaya Kumar and J. Ashok Kumar

There are several freeware available for processing of paired-end sequence reads. In this chapter we shall use 'Trimmomatic' for quality control of PE reads. First, *log in* to your account using WinSCP tool. Open PuTTY SSH terminal. In your account, find two files named, a1F.fastq and a1R.fastq. Check the quality of both the paired files using FastQC tool. Run the following command and observe the changes in quality of trimmed files. The '<>' sign used in the command argument indicates 'space'.

The command

```
Java<> -jar<> trimmomatic-0.36.jar<> PE<> -threads<> 70<> -trimlog<> a1.txt<> a1F.fastq<> a1R.fastq<> a1F_P.fastq<> a1F_S.fastq<> a1R_P.fastq<> a1R_S.fastq<> ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10<> LEADING:3<> TRAILING:13 <> SLIDINGWINDOW:4:15 <>MINLEN:100
```

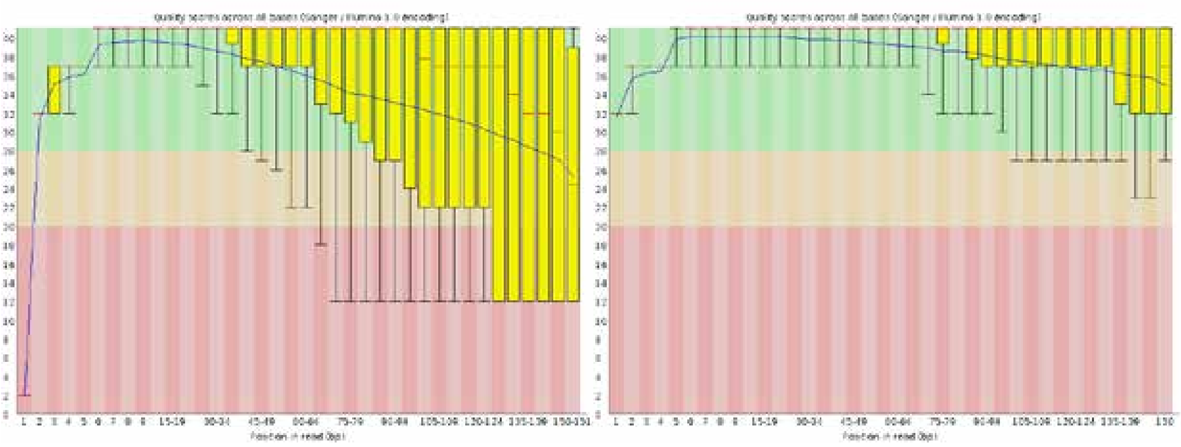
De-coding the command

Each argument in the command has a purpose of improving the quality of trimmed files. It is important to check the initial quality of sequence data and then apply the relevant arguments to improve the quality.

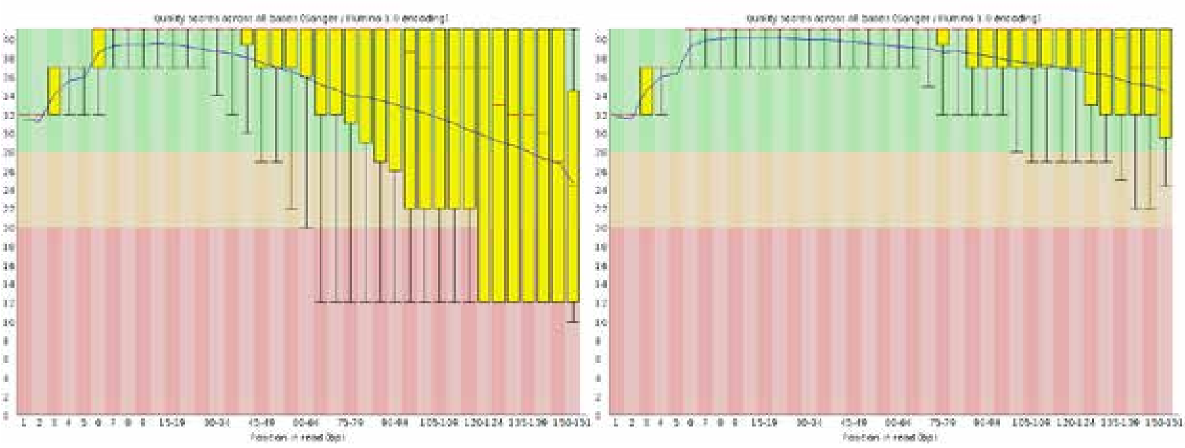
Argument	Meaning
PE	Paired-end mode. Use this for processing of PE reads data
threads	The argument to specify number of threads. Trimmomatic supports running arguments with multiple threads.
trimlog	To specify a file name that stores log of the run.
a1F.fastq	Input file name of forward or R1 reads
a1R.fastq	Input file name of reverse or R2 reads
a1F_P.fastq	Output file name of trimmed forward or R1 reads. This file is used for subsequent analysis.
a1F_S.fastq	Output file containing surviving forward reads of good quality. The paired sequences in R2 file are discarded.
a1R_P.fastq	Output file name of trimmed reverse or R2 reads. This file is used for subsequent analysis.
a1R_S.fastq	Output file containing surviving reverse reads of good quality. The paired sequences in R1 file are discarded.
ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10	Illuminaclip is used to remove adapter sequences from reads. The TruSeq3-PE-2.fa is the file containing adapter sequences.
LEADING:3	To remove bases at the start of the read, if quality is below 3
TRAILING:13	To remove bases at the end of the read, if quality is below 13
SLIDINGWINDOW:4:15	This is an argument that trims reads based on base quality. Each read is scanned from 5' end. Four continuous bases are taken as a window. The average quality of all windows in a read should be higher than 15. Otherwise, the read gets trimmed from poor quality window to the 3' end of the read.
MINLEN:100	To discard reads shorter than 100 bases after performing all the steps.

Run FastQC on the trimmed files

Below are the quality of forward sequence reads before (left) and after (right) trimming.



Below are the quality of reverse sequence reads before (left) and after (right) trimming.



Even the reads containing the adapters are trimmed. These trimmed files would be taken up for finding differentially expressed transcripts. The single-end good quality reads are also used in case of assembling genomes.

8. Assembling bacterial genomes

J. Ashok Kumar and K. Vinaya Kumar

Genome sequencing forms basis for understanding biology and functional characterization of microorganisms. Recent advances in shotgun sequencing pave the way generate genome sequences with time and cost advantage. Here we discuss whole genome assembly with of paired-end sequence reads generated from illumina platform. First we attempt to describe steps involved in *denovo* assembly of bacterial genome using masucra assembler and later we look into the steps involved in reference based assembly using Bowtie2.

Download MaSuRCA (Maryland Super Read Cabog Assembler)

MaSuRCA assembler can be downloaded from http://www.genome.umd.edu/masurca.html and once it is downloaded keep the folder in you directory and extract the tar ball using following command.

```
user@server$ tar –zxvf MaSuRCA-3.2.6.tar.gz
```

This will extract the files in to the folder **MaSuRCA-3.2.6** . You will find all the executable programs in the **bin** subfolder of the **MaSuRCA-3.2.6** folder.

Preparing Illumina sequence reads

Copy and paste the illumina paired-end sequence reads in a folder. There will be two files one for forward strand and other for reverse strand say for example vibgenome_R1.fastq vibgenome_R2.fastq. These fastq files need to be quality checked and corrected using tools like fastqc, cutadapt and trimmomatic etc.

Preparing Masurca configuration file

You will find sample configuration (sr_config_example.txt) file in the installation directory which needs to be edited with the assembly parameters. There are two sections in configuration file. One is DATA section and Other one is PARAMETERS section.

In the data section Options are available to specify paired-end (PE), mate-pair (JUMP), PACBIO and Other (Celera assembler reads). Multiple libraries data can be mentioned in multiple lines of the same read type.

For paired-end reads the following line of the data section needs to be edited.

```
PE= aa 180 20 /FULL_PATH/frag_1.fastq /FULL_PATH/frag_2.fastq
```

PE: paired-end; aa- two letter prefix; 180 is Average insert length; 20 standard deviation of insert length;

In the PARAMETERS the mandatory parameters that need to be edited are NUM_THREADS and JF_SIZE .

NUM_THREADS are number of threads allotted for assembly task. Example : NUM_THREADS=16

JF_SIZE is the jellyfish hash size, set this to about 10x the genome size but it can be genome size multiplied by its coverage.

Denovo assembly using MaSuRCA

Command to run masura assembly is

```
user@server$ /path/to/bin/masurca /path/to/config.txt
```

this will generate 'assemble.sh' file in the current location. Now we need to run this shell script for completing the assembly

```
user@server$sh assemble.sh
```

Successful completion of assembly will create several files. Look for the directory named CA and within that folder you will see 10-gapclose subfolder wherein you will find final assembled output. The output files are 'genome.ctg.fasta' for the contig sequences and 'genome.scf.fasta' for the scaffold sequences.

Reference based assembly using bowtie2

In reference based assembly reads are mapped to reference genome to identify variations like single nucleotide polymorphisms(SNPs), indels, insertions, copy number variants, genome wide association studies (GWAS).

Steps involved in reference based assembly are listed below with the commands for running each step

- Indexing a reference genome

```
$bowtie2-build V_para_GCA_000328405.1.fna vibindex
```

- Aligning reads

```
$bowtie2 -x vibindex -1 V-Para-DNA_R1.fastq -2 V-Para-DNA_R2.fastq -S align1.sam
```

- Covert sam to bam file

```
$samtools view -bS align1.sam > align1.bam
```

(-bs: input sam and output bam)

- Sort the bam file

```
$samtools sort align1.bamalign1.sorted.bam
```

- Create the BCF file

```
$samtools mpileup -uf V_para_GCA_000328405.1.fna align1.sorted.bam.bam | bcftools view -Ov - > align.raw.bcf
```

(-u generate uncompress BCF output; -f faidx indexed reference sequence file; -Ov output potential variant sites only)

9. RNAseq data analysis in Trinity

K. Vinaya Kumar, J. Ashok Kumar and M.S. Shekhar

Many of the commercially relevant aquaculture species including shrimp are not having publicly available reference genome. Therefore the analysis of RNAseq data for such species mandates building a *de novo*transcriptome assembly. For every experiment, a *de novo* assembly has to be made utilizing the RNAseq reads of all the samples in the study. In this chapter, we shall practice building a *de novo* assembly of transcriptome and conducting differential transcript analysis in trinity software.

9.1 The datasets

Let us assume an experiment involving two treatments a & b. Each treatment has three replicate individuals. At the end of the experiment, tissue samples are collected from all replicate individuals and RNAseq was performed on Illumina platform. The following datasets have been generated.

Table.Datasets to be used for RNAseq data analysis

	Treatment A		**Treatment B**	
	Forward reads	**Reverse reads**	**Forward reads**	**Reverse reads**
replicate 1	a1F.fastq	a1R.fastq	b1F.fastq	b1R.fastq
replicate 2	a2F.fastq	a2R.fastq	b2F.fastq	b2R.fastq
replicate 3	a3F.fastq	a3R.fastq	b3F.fastq	b3R.fastq

9.2 Quality control of datasets

Process the raw reads using Trimmomatic tool and obtain quality reads. Keep the following arguments while running Trimmomatic.

ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10

LEADING:3

TRAILING:13

SLIDINGWINDOW:4:15

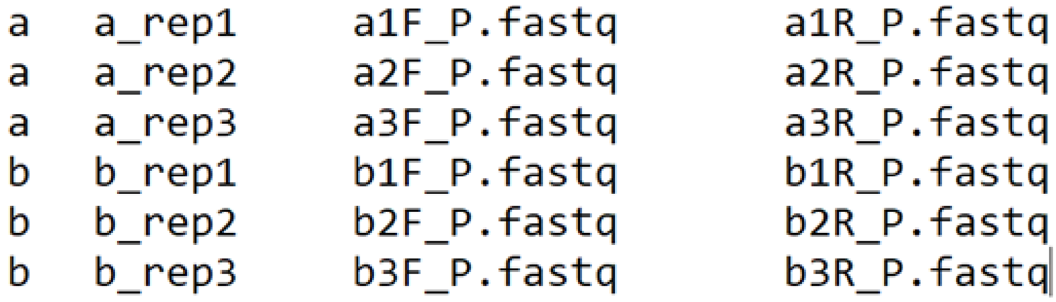
MINLEN:100

The numbers of reads retained for downstream analysis are given below

Sample name	**Reads in raw file (million)**	**Reads in processed file (million)**
a1	10	4.954252
a2	10	5.577112
a3	10	6.412094
b1	10	5.257203
b2	10	4.160784
b3	10	3.607086

9.3 Building a *de novo* assembly

As the experiment involves triplicate samples, prepare a text file showing the triplicate samples under each treatment and their file names as shown below.



```
a   a_rep1   a1F_P.fastq   a1R_P.fastq
a   a_rep2   a2F_P.fastq   a2R_P.fastq
a   a_rep3   a3F_P.fastq   a3R_P.fastq
b   b_rep1   b1F_P.fastq   b1R_P.fastq
b   b_rep2   b2F_P.fastq   b2R_P.fastq
b   b_rep3   b3F_P.fastq   b3R_P.fastq
```

Then proceed for building assembly using the following command,

Trinity<> --seqType<> fq<> --samples_file<> ab_samples.txt<> --CPU<> 70<> --max_memory<> 300G<> --SS_lib_type<> FR<> --output <>trinity_ab

The command arguments details are,

Input files are fastq format

Samples file names are given in ab_samples.txt

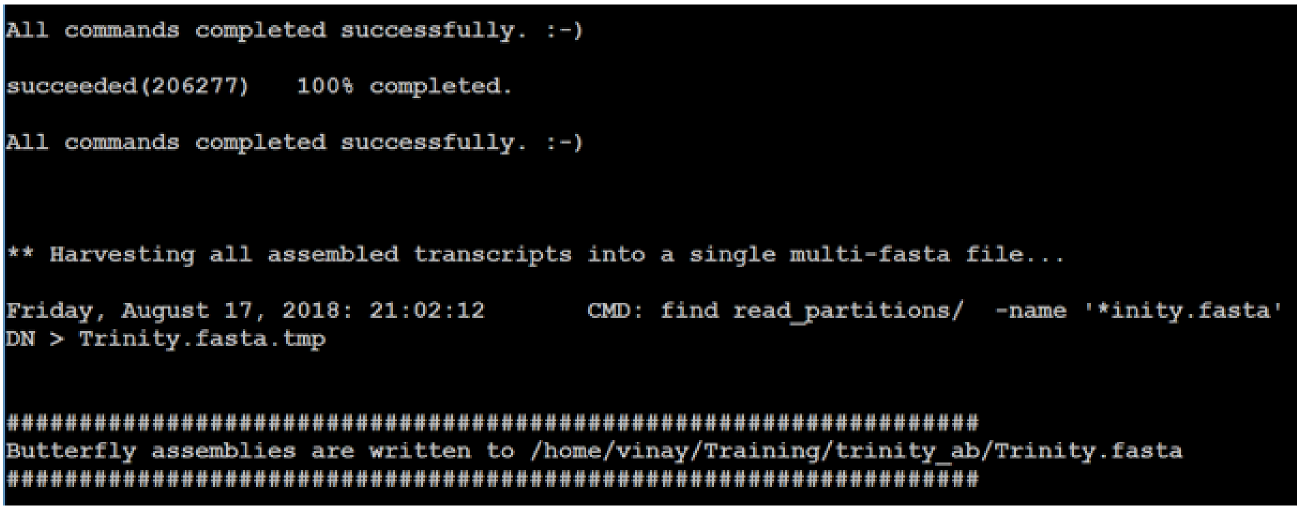
Use 70 threads

Limit maximum memory to 300 GB

Data obtained from strand-specific library as forward and reverse reads

Store output in folder, trinity_ab

The assembly is completed when you see the messages printed as shown below.



```
All commands completed successfully. :-)

succeeded(206277)   100% completed.

All commands completed successfully. :-)



** Harvesting all assembled transcripts into a single multi-fasta file...

Friday, August 17, 2018: 21:02:12       CMD: find read_partitions/  -name '*inity.fasta'
DN > Trinity.fasta.tmp


#####################################################################
Butterfly assemblies are written to /home/vinay/Training/trinity_ab/Trinity.fasta
#####################################################################
```

Browse to the folder and find the assembled transcripts file, *Trinity.fasta*. Rename the file as '*Trinity_ab.fasta*' for easy identification.

9.4 Assessing quality of assembly

9.4.1. N50: Compute N50 statistic by running the following command,

TrinityStats.pl<> Trinity_ab.fasta<>><> Trinity_ab_stats.txt

9.4.2. ExN50: The E90N50 is being considered as more appropriate for RNAseq studies rather than N50. Get ExN50 stats with the following argument.

contig_ExN50_statistic.pl <>matrix.TMM.EXPR.matrix <>Trinity_ab.fasta | tee ExN50.stats

E	Minimum expression	ExN50	Number of transcripts
E90	2.28	1611	45381
E91	1.952	1511	53150
E92	1.916	1409	61794
E93	1.55	1314	71403
E94	1.5	1212	82217
E95	1.262	1102	94509
E96	1.122	1005	108691
E97	0.95	927	125654
E98	0.746	858	146688
E99	0.566	791	175457
E100	0	605	281008

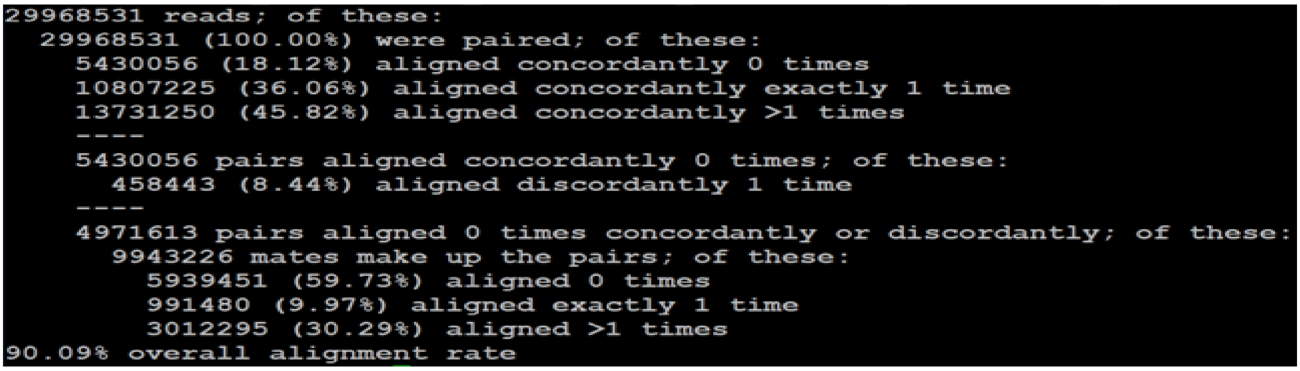
The N50 calculated based on the top most expressed transcripts that represent 90% of the total normalized expression data is 1611 bases and includes 45381 transcripts.

9.4.3. Read representation: The proportion of paired-reads represented in the assembled transcripts is another parameter that helps in evaluating the assembly. We shall use bowtie2 tool for this. First an index is to be made and then reads are to be aligned on to transcripts. Run the following two commands.

bowtie2-build<>Trinity_ab.fasta<> Trinity_ab.fasta

AND

bowtie2<> -x<> Trinity_ab.fasta<> -q<> --fr<> -1<> a1F_P.fastq,a2F_P.fastq,a3F_P.fastq,b1F_P.fastq,b2F_P.fastq,b3F_P.fastq<> -2<> a1R_P.fastq,a2R_P.fastq,a3R_P.fastq,b1R_P.fastq,b2R_P.fastq,b3R_P.fastq<> -S<> samfile<> --no-unal<> -p<>50



```
29968531 reads; of these:
  29968531 (100.00%) were paired; of these:
    5430056 (18.12%) aligned concordantly 0 times
    10807225 (36.06%) aligned concordantly exactly 1 time
    13731250 (45.82%) aligned concordantly >1 times
    ----
    5430056 pairs aligned concordantly 0 times; of these:
      458443 (8.44%) aligned discordantly 1 time
    ----
    4971613 pairs aligned 0 times concordantly or discordantly; of these:
      9943226 mates make up the pairs; of these:
        5939451 (59.73%) aligned 0 times
        991480 (9.97%) aligned exactly 1 time
        3012295 (30.29%) aligned >1 times
90.09% overall alignment rate
```

As per the statistics shown above, the overall alignment rate is 90% which is good.

9.5 Transcript quantification

9.5.1. Estimate abundance

The first step in transcript quantification is to estimate the abundance of all transcripts in every sample. We shall practice estimating transcript abundance using alignment-based method, RSEM though alignment-free methods such as kallisto and salmon exist. Run the following command to get abundance estimates by aligning the sequence reads to transcripts and counting the number of reads aligned for each transcript.

```
align_and_estimate_abundance.pl<> --transcripts<> Trinity_ab.fasta<> --seqType<> fq <>--samples_file<> ab_samples.txt <>--est_method<> RSEM <>--aln_method<> bowtie<> --trinity_ mode<> --prep_reference<> --SS_lib_type <>FR <>--output_dir<> ab_rsem_outdir <>--thread_ count<>20
```

Argument	Meaning
align_and_estimate_abundance.pl	Script to align reads on to transcripts and get abundance estimates
--transcripts	To define the assembled transcripts file name
--seqType	To define input file format
--samples_file	Define the file name that contains treatments, replicates and reads file names
--est_method	To define abundance estimation method (options are RSEM/ eXpress/kallisto/salmon)
--aln_method	To define alignment method (bowtie/bowtie2)
--trinity_mode	To automatically generate gene_trans_map
--prep_reference	To build target index
--SS_lib_type	Specify if the library is strand-specific (FR/RF)
--output_dir	Name of the directory to store output files
--thread_count	Number of threads to use for running the argument

At the end of the run, find that six folders are created corresponding to six samples. In each folder observe for a file named, RSEM.isoforms.results. These files are used for further processing. These abundance estimates are built in to matrix with the following argument,

```
abundance_estimates_to_matrix.pl<> --est_method<> RSEM<>RSEM.isoforms.results
```

Mention all the six file names of RSEM.isoforms.results corresponding to six samples.

9.5.2. Count the numbers of expressed transcripts

Plot the number of transcripts that are expressed at different TPM threshold by running the following argument,

```
count_matrix_features_given_MIN_TPM_threshold.pl matrix.TPM.not_cross_norm | tee counts_by_min_TPM
```

The output looks like the table depicted below.

Neg_min_tpm	Number of features
-10	24978
-9	29296
-8	35850
-7	45677
-6	62228
-5	84308
-4	111966
-3	151586
-2	202147
-1	228356
0	281008

9.6 Differential expression analysis

At present, Trinity supports four R packages for performing differential expression analysis. These are edgeR, DEseq2, limma/voom, and ROTS. We shall use edgeR in this tutorial to understand differential expression analysis. Run the following commands.

```
run_DE_analysis.pl<> --matrix<> matrix.counts.matrix<> --method<> edgeR<> --samples_file<> ab_samples_DE.txt<> --output <>ab_edgeRresult
```

AND

```
analyze_diff_expr.pl<> --matrix<> matrix.TMM.EXPR.matrix<> --output<> aVSb <>--samples<> ab_samples_analyzeDE.txt
```

In this particular example, we got five transcripts that are differentially expressed in sample b compared to sample a. Now proceed to functional annotation of these transcripts and understand its role for the given treatment in the study.

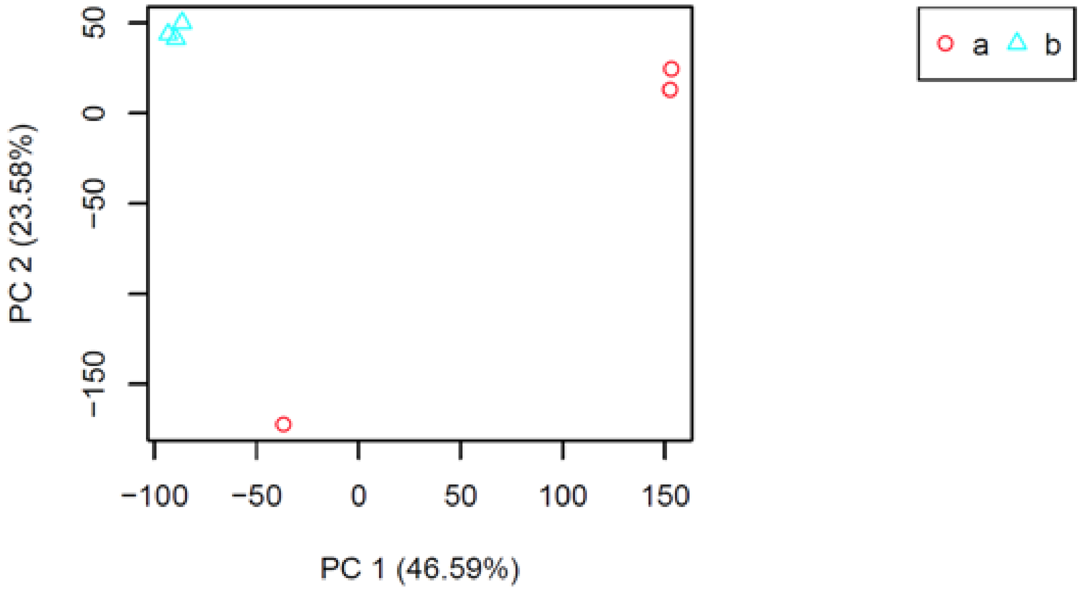
9.7 Quality check of samples and replicates: You may compare the samples as well as replicates in each sample with the following commands

9.7.1.
```
/PtR<> --matrix <>matrix.counts.matrix<> --samples <>ab_replicatesTest.txt<> --CPM <>--log2<> --min_rowSums <>10<> --compare_replicates
```

9.7.2.
```
/PtR<> --matrix<> matrix.counts.matrix<> --min_rowSums <>10<> -s<> ab_replicatesTest.txt<> --log2<> --CPM <>--sample_cor_matrix
```

9.7.3.
```
/PtR<> --matrix<> matrix.counts.matrix<> -s<> ab_replicatesTest.txt<> --min_rowSums 10<> --log2 <>--CPM<> --center_rows <>--prin_comp 3
```

For example, in the picture below, it is evident that the replicates in treatment b are clustered closely. This ensures that all the replicates behaved similarly.


PC 2 (23.58%)
50
0
−50
−150
−100
−50
0
50
100
150
PC 1 (46.59%)
○ a
△ b

10. Phylogenomic analysis using MrBayes

K. Vinaya Kumar, J. Ashok Kumar and G. Gopikrishna

Researchers perform phylogenetic analysis to understand the evolutionary relations among taxa. Such analyses require information on best-fit partitioning schemes and best-fit models for the sequence data in hand. The PartitionFinder is a suitable tool to find that information to build phylogenetic tree. In this chapter we conduct analyses using PartitionFinder tool for finding the best-fit partitioning scheme and evolutionary models. Then using these partitioning schemes and models, we would build a Bayesian tree in MrBayes tool.

10.1 PartitionFinder

For this exercise, a sequence file containing sequence data of 5 genes on 10 taxa is provided in your work folder. Open the folder and check for the file named, 'sequence_10_5.phy'.

Taxa labels	10 taxa	taxaA, taxaB, …….. taxaJ
Gene partitions	5 genes	Gene1: 1-675 bp
		Gene2: 676-834 bp
		Gene3: 835-2373 bp
		Gene4: 2374-3060 bp
		Gene5: 3061-3849 bp

The arguments for running PartitionFinder are to be provided in a configuration file. Find the file 'partition_finder.cfg' in work folder. Keep settings as per the table given below.

Argument	Option	Meaning
alignment	sequence_10_5.phy	File containing sequences in phylip format
branchlengths	linked	Linked branch lengths are supported by almost all phylogeny programs
models	mrbayes	Includes all the evolutionary models that are compatible for MrBayes tool for testing
model_selection	aicc	Criterion to decide the best model

data_blocks	Gene1_pos1 = 1-675\3;
	Gene1_pos2 = 2-675\3;
	Gene1_pos3 = 3-675\3;
	Gene2_pos1 = 676-834\3;
	Gene2_pos2 = 677-834\3;
	Gene2_pos3 = 678-834\3;
	Gene3_pos1 = 835-2373\3;
	Gene3_pos2 = 836-2373\3;
		Defining data partitions. For each gene, three data partitions are defined based on the three base positions of triplet code. We defined 15 data blocks for 5 genes.

	Gene3_pos3 = 837-2373\3;
	Gene4_pos1 = 2374-3060\3;
	Gene4_pos2 = 2375-3060\3;
	Gene4_pos3 = 2376-3060\3;
	Gene5_pos1 = 3061-3849\3;
	Gene5_pos1 = 3062-3849\3;
	Gene5_pos1 = 3063-3849\3;	

search	greedy	Defining the method to use for finding good partitioning scheme

Then use the following command to run the PartitionFinder. Here, you mention the name of the folder containing sequence file and configuration file in place of ‘folder_name’.

```
python<>PartitionFinder.py <> folder_name/ --no-ml-tree
```

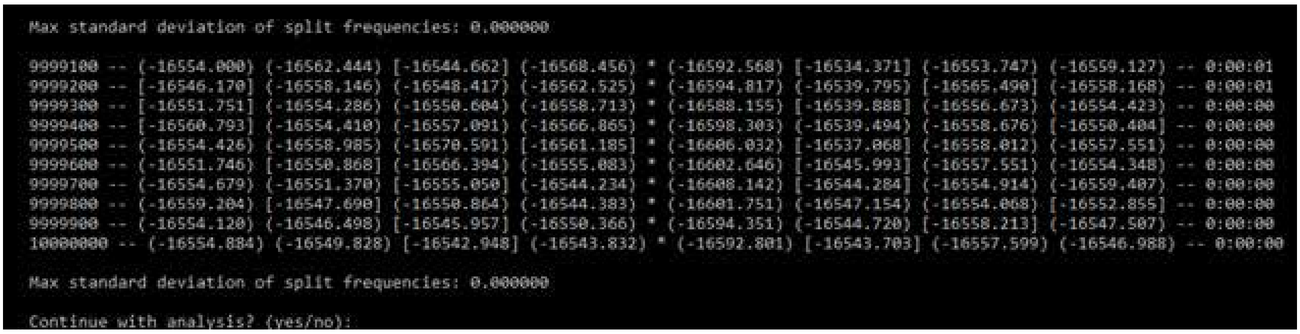
The output files are stored in the folder ‘analysis’. Find the file ‘best_scheme.txt’ that contains the arguments for running the best fit models on best partitions in MrBayes tool.

10.2 MrBayes

The Bayesian analysis requires the input sequence file in nexus format. Find the file, sequence_10_5.nxs file which was used for analysis in PartitionFinder. Download windows version of MrBayes tool and unzip the file. Start the tool by clicking on the executable. Then run the following arguments.

```
execute sequence_10_5.nxs;

outgroup taxaJ;

type arguments given in output file of PartitionFinder, ‘best_scheme.txt’

showmodel                    # to check for the model defined

mcmc ngen=10000000 nruns=2 nchains=4 samplefreq=100 printfreq=100

diagnstat=maxstddev diagnfreq=1000 savebr=yes filename= PartitionFinder
```

After running for 10 million generations, you would see the following screen.



```
Max standard deviation of split frequencies: 0.000000

9999100 -- (-16554.000) (-16562.444) [-16544.662] (-16568.456) * (-16592.568) [-16534.371] (-16553.747) (-16559.127) -- 0:00:01
9999200 -- [-16546.170] (-16558.146) (-16548.417) (-16562.525) * (-16594.817) (-16539.795) [-16565.490] (-16558.168) -- 0:00:01
9999300 -- [-16551.751] (-16554.286) (-16550.604) (-16558.713) * (-16588.155) [-16539.888] (-16556.673) (-16554.423) -- 0:00:00
9999400 -- [-16560.793] (-16554.410) (-16557.091) (-16566.865) * (-16598.303) (-16539.494) (-16558.676) [-16550.404] -- 0:00:00
9999500 -- (-16554.426) (-16558.985) (-16570.591) [-16561.185] * (-16606.032) [-16537.068] (-16558.012) (-16557.551) -- 0:00:00
9999600 -- (-16551.746) [-16550.868] (-16566.394) (-16555.083) * (-16602.646) [-16545.993] (-16557.551) (-16554.348) -- 0:00:00
9999700 -- (-16554.679) (-16551.370) [-16555.050] (-16544.234) * (-16608.142) [-16544.284] (-16554.914) (-16559.407) -- 0:00:00
9999800 -- (-16559.204) [-16547.690] (-16550.864) (-16544.383) * (-16601.751) (-16547.154) (-16554.068) [-16552.855] -- 0:00:00
9999900 -- (-16554.120) (-16546.498) [-16545.957] (-16550.366) * (-16594.351) (-16544.720) [-16558.213] (-16547.507) -- 0:00:00
10000000 -- (-16554.884) (-16549.828) [-16542.948] (-16543.832) * (-16592.801) [-16543.703] (-16557.599) (-16546.988) -- 0:00:00

Max standard deviation of split frequencies: 0.000000

Continue with analysis? (yes/no):
```

You could continue with more generations if required by opting for ‘yes’ at the prompt.

Then obtain a summary of parameters with the following command. Here, by default, first 25% of observations are discarded.

```
sump filename= PartitionFinder
```

Look for the parameters like estimated sample size and potential scale reduction factor. Then summarize the trees with the following command. This prints a cladogram and a phylogram.

```
sumt filename= PartitionFinder
```

Check for the .tre file and open it in FigTree to view the tree.

11. Microsatellites genotypes generation by Fragment analysis method

B. Sivamani

Fragment analysis (Genotyping) can be performed on DNA fragments that have fluorescent labels. Using a labeled primer with PCR amplification is a common method used to incorporate these labels. The Molecular Biology Core lab is already set to run multiple fluorescent dye sets.

Steps

1. Microsatellite loci selection
2. Primer designing (fluorescent labelled)
3. PCR
4. Fragment analysis – ABI sequencer
5. Generating genotypes

1. Microsatellite loci selection

The loci are selected loci through literature search or from any database. For fisheries, the NBFGR FishMicrosat database provides updated microsatellite loci and their primers for pcr amplification. (http://mail.nbfgr.res.in/fishmicrosat/).

Steps to find the microsatellite loci in Penaeus (Fenneropeaneus indicus)

- ➢ Visit the site (http://mail.nbfgr.res.in/fishmicrosat/)
- ➢ Under Analysis and Primer, select your species of search and you will find all the microsatellite loci related to the specific search. The following details of the loci also present

1. Accession Number: link will lead to the NCBI site and will give all the details of the nucleotide sequence
2. SSR type: di, tri, tetra or compound
3. Microsatellite span in the sequence
4. Primers to amplify the locus

2. Designing primer

One can use the specified primer or a primer may be designed as per the user requirement by utilizing the accession Number option. One of the primers needs to be fluorescent labeled.

3. PCR

Isolate Total DNA from the biological material (Blood/finclips/muscle,etc.) of the species. Verify the DNA for quality and quantity. Carry out the PCR with labelled primers. Verify the amplicon by agarose gel electrophoresis. The specific amplification of the product is considered better. Else, presence of some less intense non-specifics also accepted.

4. Fragment analysis – by ABI sequencer

The step is normally outsourced being the cost of the equipment is too high. We receive the results generated by GeneMapper software (Private firms use the inbuilt GeneMapper software) with the following files.

1. FSA file
2. PDF for electropherogram
3. Genotypes in excel sheet

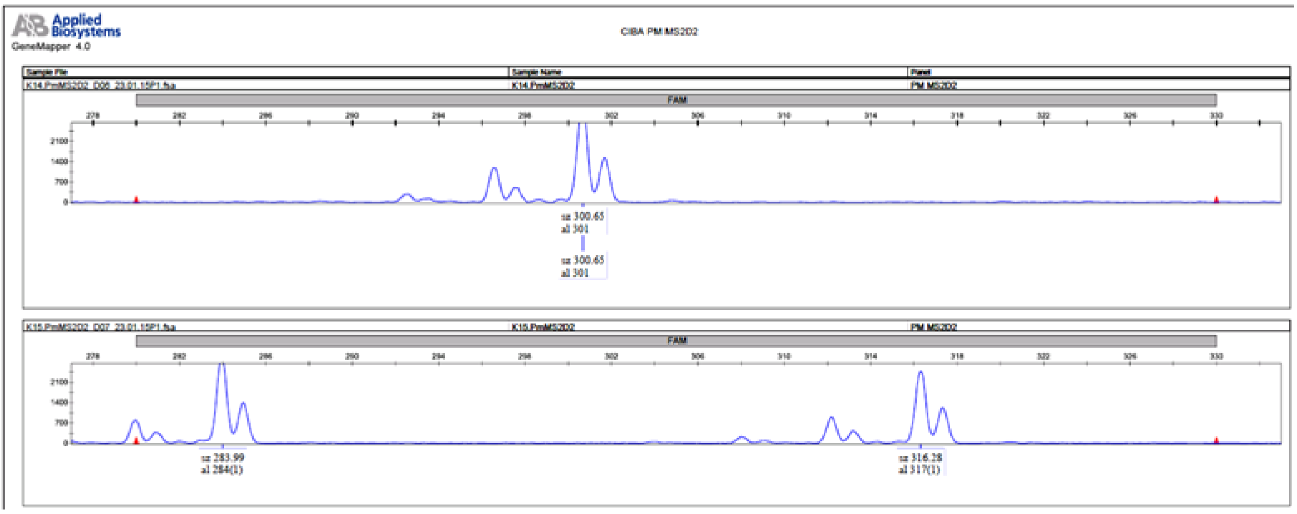
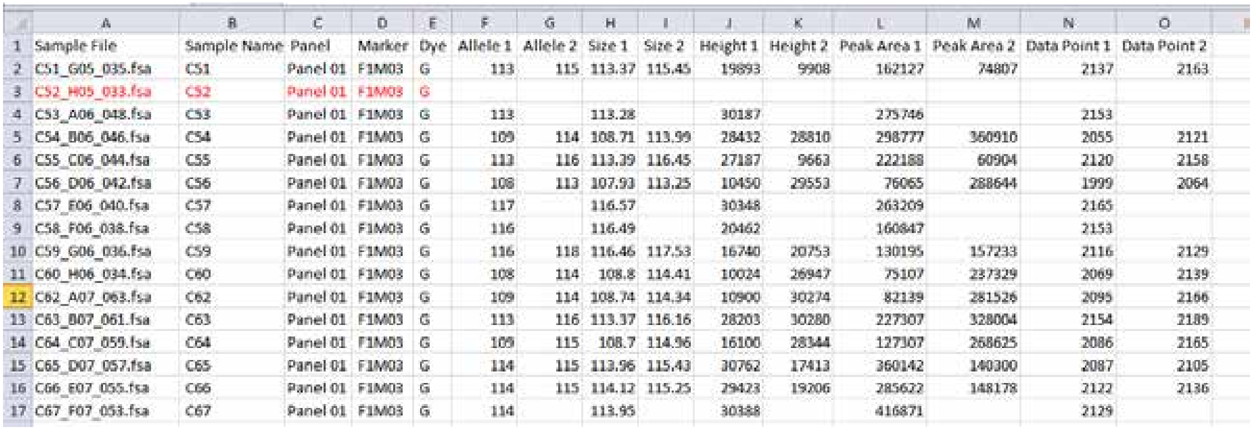


Fig:1 Electropheogram



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Sample File	Sample Name	Panel	Marker	Dye	Allele 1	Allele 2	Size 1	Size 2	Height 1	Height 2	Peak Area 1	Peak Area 2	Data Point 1	Data Point 2
2	C51_G05_035.fsa	C51	Panel 01	F1M03	G	113	115	113.37	115.45	19893	9908	162127	74807	2137	2163
3	C52_H05_033.fsa	C52	Panel 01	F1M03	G										
4	C53_A06_048.fsa	C53	Panel 01	F1M03	G	113		113.28		30187		275746		2153	
5	C54_B06_046.fsa	C54	Panel 01	F1M03	G	109	114	108.71	113.99	28432	28810	298777	360910	2055	2121
6	C55_C06_044.fsa	C55	Panel 01	F1M03	G	113	116	113.39	116.45	27187	9663	222188	60904	2120	2158
7	C56_D06_042.fsa	C56	Panel 01	F1M03	G	108	113	107.93	113.25	10450	29553	76065	288644	1999	2064
8	C57_E06_040.fsa	C57	Panel 01	F1M03	G	117		116.57		30348		263209		2165	
9	C58_F06_038.fsa	C58	Panel 01	F1M03	G	116		116.49		20462		160847		2153	
10	C59_G06_036.fsa	C59	Panel 01	F1M03	G	116	118	116.46	117.53	16740	20753	130195	157233	2116	2129
11	C60_H06_034.fsa	C60	Panel 01	F1M03	G	108	114	108.8	114.41	10024	26947	75107	237329	2069	2139
12	C62_A07_063.fsa	C62	Panel 01	F1M03	G	109	114	108.74	114.34	10900	30274	82139	281526	2095	2166
13	C63_B07_061.fsa	C63	Panel 01	F1M03	G	113	116	113.37	116.16	28203	30280	227307	328004	2154	2189
14	C64_C07_059.fsa	C64	Panel 01	F1M03	G	109	115	108.7	114.96	16100	28344	127307	268625	2086	2165
15	C65_D07_057.fsa	C65	Panel 01	F1M03	G	114	115	113.96	115.43	30762	17413	360142	140300	2087	2105
16	C66_E07_055.fsa	C66	Panel 01	F1M03	G	114	115	114.12	115.25	29423	19206	285622	148178	2122	2136
17	C67_F07_053.fsa	C67	Panel 01	F1M03	G	114		113.95		30388		416871		2129	

Fig: 2 Genotypes data generated by GeneMapper software

5. Generating genotypes from FSA file using R

# 5.1 Install R from the site https://www.r-project.org/

# 5.2 installing the package from R site##

Install.packages(“Fragman”)

# 5.3 To activate, the package has to be loaded###

>Library(Fragman)

# 5.4 To specify the input fas file

```
FIM03<-storing.inds("C:/Users/Admin/Desktop/training writeup/FAS file-ciba")

#5.5 To specify the ladder used in the analysis

my.ladder <- c(35,50,75,100,139,150,160,200,250,300,340,350,400,450,490,500)

# 5.6 To merge both the earlier specified information (FAs file and ladder)

ladder.info.attach(stored=FIM03, ladder=my.ladder)

# 5.7 Tocreate friendly plots for any number of individuals specified and can be used to

#design panels for posterior automatic scoring

overview2(my.inds=FIM03, channel = 2, ladder=my.ladder)

# 5.8  to view the vector with expected DNA sizes to be used in the next step for scoring

my.panel2<-overview2(my.inds=FIM03, channel = 2, ladder=my.ladder, init.thresh=3000, xlim=c(90,130))

my.panel2

# 5.9 To score our samples for channel 2 with our panel created previously

res2 <- score.markers(my.inds=FIM03, channel = 2, panel=my.panel2$channel_2, ladder=my.ladder, electro=FALSE)

# 5.10  To extract your peaks in a data.frame

final.results <- get.scores(res2)

final.results

# 5.11To get the results in text file format

write.table(final.results, " C:/Users/Admin/Desktop/training writeup/FIM03-18-2.txt", sep="\t")
```

******

Note

install.packages = to install the specific package

library = to load addon package

storing.inds = is the function in charge of reading the FSA files and storing them with a list structure

ladder.info.attach = uses the information read from the FSA files and a vector containing theladder information (DNA size of the fragments) and matches the peaks from the channel where theladder was run with the DNA sizes for all samples. Then loads such information in the R environmentfor the use of posterior functions stored = List of dataframes obtained by using the storing.inds function

overview2 = create friendly plots for any number of individuals specified and can be used to design panels (overview2) for posterior automaticscoring (like licensed software does)

my.inds =List with the channels information from the individuals specified, usually comingfrom the storing.inds function output

Channel = The channel you wish to analyze, usually 1 is blue, 2 is green, 3 is yellow, 4 is red and so on

init.thresh = An initial value of intensity to detect peaks. We recommend not to deal to muchwith it unless you have highly controlled dna concentrations in your experiment.

score.markers = score the alleles by finding the peaks provided in the panel

panel =different dna sizes usually obtained by using overview and locator functions

get.scores =Once the calls have been obtained we can extract a data frame with the get. scores function.

******

xlim=c(a1,b1)) = the approximate amplicon size to be mentioned in overview2

Dye sets used applied biosystem DNA analyser

Blue: 5FAM and 6FAM

Green: Hex, vic,Tet and Joe

Yellow: Tamra and Ned

Red: Rox and Pet

12. Genepop : Population Genetics analysis

B. Sivamani

GENEPOP is a population genetics software package originally developed by *Michel Raymond (Raymond@isem.univ-montp2.fr)* and *Francois Rousset (Rousset@isem.univ-montp2.fr)*, at the Laboratiore de Genetique et Environment, Montpellier, France.

Access: Web version is easy to use but active internet is required, also can be downloaded and run under windows and linux without internet.

Genepop on the web

Can be accessed in the link: http://genepop.curtin.edu.au/

It has seven options. Once the input file is prepared, all the options can be run.

Option1	Hardy Weinberg Exact Tests
Option 2	Genotypic linkage disequilibrium
Option 3	Population differentiation
Option 4	Nm estimates - private allele method
Option 5	Basic Information
Option 6	Fst and other correlations
Option 7	file conversion

Input file (e.g.)

Title:"P.indicus microsatellites based population diversity"

```
FIM03
FIM06
FIM20
FIM21
FIM17
FIM19
FIM23
POP
C051    ,     113115  161161  128130  225225  308289  110118  201226
C054    ,     109115  161167  123123  222231  299307  110116  222226
C055    ,     113117  161164  123123  231231  307307  108110  191201
C056    ,     109113  161164  123123  245248  307307  110118  191201
C057    ,     117117  161161  123125  230230  293307  110118  191201
POP
K15     ,     115115  161161  128130  230230  308308  110118  201222
K20     ,     115117  160160  123125  000000  294307  110118  191201
K23     ,     107113  149159  123123  000000  309308  110116  191201
K36     ,     111115  160186  123123  000000  307308  108118  201222
K42     ,     109115  138149  123125  000000  298300  110116  191201
```

Instructions to input file

- Input file should be prepared in notepad, notepad++ or excel
- The input file should have txt extension e.g. filename.txt
- First line, title is written within inverted commas
- No constraint on blanks separating the various fields, tabs or spaces allowed.
- Loci names can appear on separate lines, or on one line if **separated by commas**
- Individual identifier may have blanks but must end with a comma
- Alleles are numbered from 01 to 99 (or 001 to 999). Consecutive numbers to designate alleles are not required.
- Populations are defined by the position of the "Pop" separator. To group various populations, just remove relevant "Pop" separators.
- Individual genotypes for the **web version** must be on one line. This differs from the PC version.
- Missing data should be indicated as 00 (or 000) rather than blanks. There are three possibilities for missing data :
    - no information (0000) or (000000),
    - partial information for first allele (1000) or (010000),
    - partial information for second allele (0010) or (000010).
- The number of locus names should correspond to the number of genotypes in each row. If you remove one or several loci from your input file, you should remove both their names and the corresponding genotypes.
- No empty lines should be found within the file.
- No more than one empty line should be present at the end of file.

To run in PC

Download Genepop form the link http://kimura.univ-montp2.fr/~rousset/Genepop.htm

Based on OS 32 or 64 bit version can be run without installing from the PC. The Input file format is same like Genepop on web. Input file should be in the same folder of the software. After specifying the input file, type the number of the options (for analysis) and the output-file gets stored at the same folder.

13. Population genetic analysis of microsatellite data in Arlequin

B. Sivamani

Arlequin is a software tool specially designed to extract information on genetic and demographic features of a collection of population samples. Arlequin can handle several types of data either in haplotypic or genotypicform. The data types include

- DNA sequences
- RFLP data
- Microsatellite data
- Standard data
- Allele frequency data

Arlequin can analyse various population parameters. They are standard indices, molecular diversity, Linkage disequilibrium, Hardy-Weinberg equilibrium, Amova, Exact population differentiation etc.,

Installation and uninstallation

1. Download WinArl35.zip to any temporary directory.
2. Extract all files contained in Arlequin35.zip in the directory of your choice.
3. Start Arlequin by double-clicking on the file WinArl35.exe, which is the main executable file.
4. To uninstall simply delete the directory where you installed Arlequin. The registries were not modified by the installation of Arlequin.

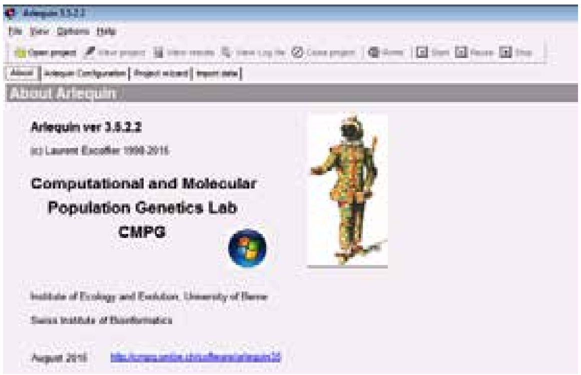
Configuration

Download text editor tool from www.textpad.com and install. It is required to create, edit the project files and to view the log files.

Download R from www.rproject.org and install it.

Running the software Arlequin

Open the arlequin by double clicking "WinArl35.exe" which leads to the home page.



Arlequin ver 3.5.2.2

(c) Laurent Excoffier 1998-2015

Computational and Molecular Population Genetics Lab CMPG

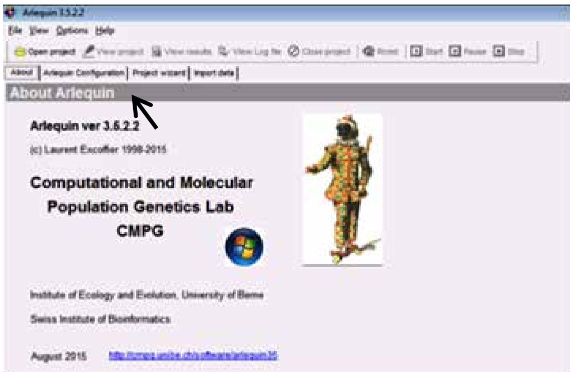
Institute of Ecology and Evolution, University of Berne

Swiss Institute of Bioinformatics

August 2015 http://cmpg.unibe.ch/software/arlequin35

Step1. Configuration of Arlequin

1.1 Click on 'Arlequin Configuration' box, select the option Append results, XML output and use 64bit external . Append Results is selected to get the results of several runs of a specific input file into a single output file. The XML output option is to get the results in XML format.



Arlequin ver 3.5.2.2

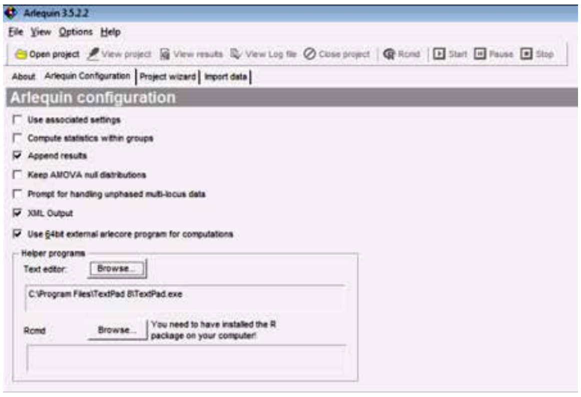
(c) Laurent Excoffier 1998-2015

Computational and Molecular Population Genetics Lab CMPG

Institute of Ecology and Evolution, University of Berne

Swiss Institute of Bioinformatics

August 2015 http://cmpg.unibe.ch/software/arlequin35

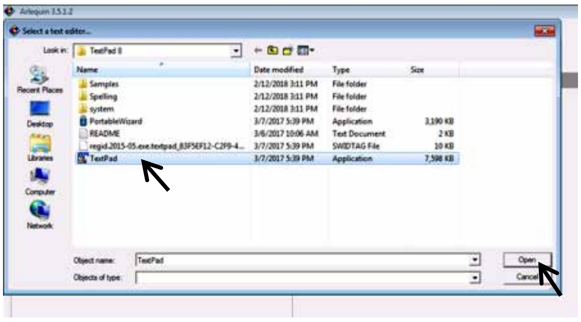


□ Use associated settings
□ Compute statistics within groups
☑ Append results
□ Keep AMOVA null distributions
□ Prompt for handling unphased multi-locus data
☑ XML Output
☑ Use 64bit external arlecore program for computations

Text editor: C:\Program Files\TextPad 8\TextPad.exe

Rcmd: You need to have installed the R package on your computer!

1.2 Under 'Helper Programs', the path of the Text editor and R has to be specified for the utilization. Click the 'Browse' box of the 'Text editor' and browse where the Textpad.exe file is located.

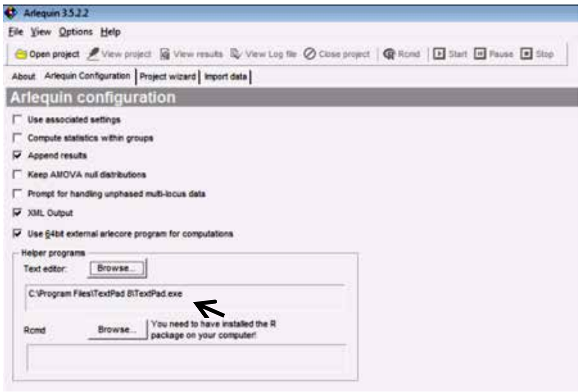


Look in: TextPad 8

Name	Date modified	Type	Size
Samples	2/12/2018 3:11 PM	File folder	
Spelling	2/12/2018 3:11 PM	File folder	
system	2/12/2018 3:11 PM	File folder	
PortableWizard	3/7/2017 5:39 PM	Application	3,190 KB
README	3/6/2017 10:06 AM	Text Document	2 KB
regid.2015-05.exe.textpad_83F5EF12-C2F9-4...	3/7/2017 5:39 PM	SWIDTAG File	10 KB
TextPad	3/7/2017 5:39 PM	Application	7,598 KB

Object name: TextPad

Objects of type:

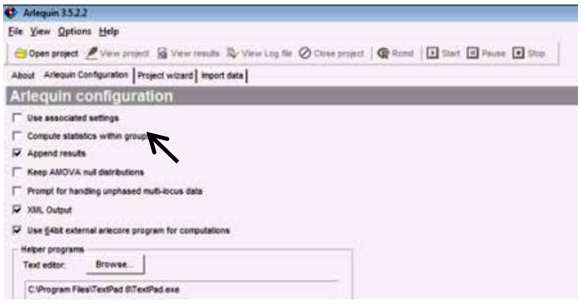


□ Use associated settings
□ Compute statistics within groups
☑ Append results
□ Keep AMOVA null distributions
□ Prompt for handling unphased multi-locus data
☑ XML Output
☑ Use 64bit external arlecore program for computations

Text editor: C:\Program Files\TextPad 8\TextPad.exe

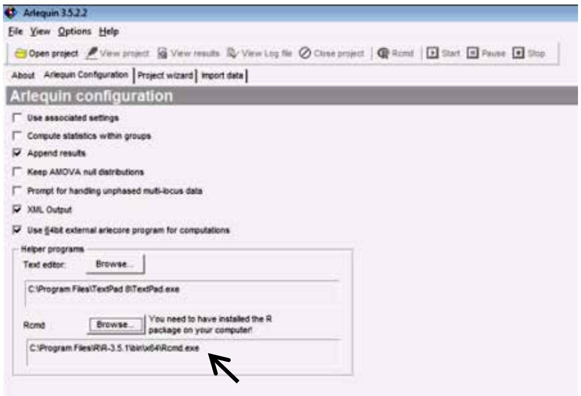
Rcmd: You need to have installed the R package on your computer!

1.3 Click the 'Browse' button of 'Rcmd' and indicate the path by selecting the Rcmd Application form the specific folder.


Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Arlequin configuration
Helper programs
Browse...

☐ Use associated settings
☐ Compute statistics within groups
☑ Append results
☐ Keep AMOVA null distributions
☐ Prompt for handling unphased multi-locus data
☑ XML Output
☑ Use 64bit external arlecore program for computations

Text editor: C:\Program Files\TextPad 8\TextPad.exe


Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Arlequin configuration
Helper programs
Browse...
Browse...

☐ Use associated settings
☐ Compute statistics within groups
☑ Append results
☐ Keep AMOVA null distributions
☐ Prompt for handling unphased multi-locus data
☑ XML Output
☑ Use 64bit external arlecore program for computations

Text editor: C:\Program Files\TextPad 8\TextPad.exe

Rcmd: You need to have installed the R package on your computer!

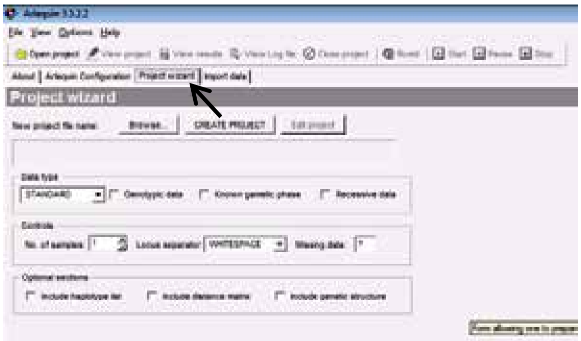
C:\Program Files\R\R-3.5.1\bin\x64\Rcmd.exe

Step 2: Project file preparation

Arlequin requires project file (input) which has the extension ".arp". Once the analysis over, the output (results) will be stored in the same (WinArl35) folder as subfolder with the extension ".res".

2.1 Open the arlequin software by double cicking the " WinArl35.exe"

2.2 Click on "project wizard" option. An example project file will be created with the Arlequin format.

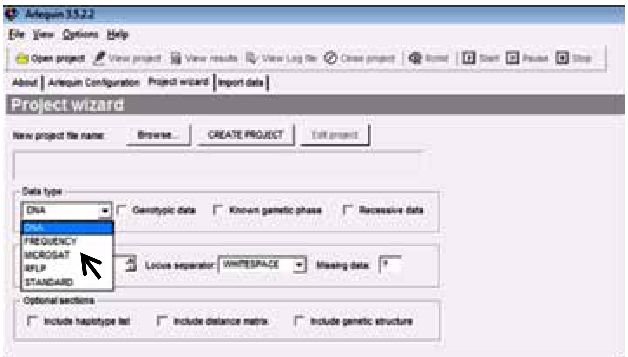

Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
New project file name:
Browse...
CREATE PROJECT
Edit project
Data type
Controls
Optional sections

Data type: STANDARD ☐ Genotypic data ☐ Known gametic phase ☐ Recessive data

No. of samples: 1 Locus separator: WHITESPACE Missing data: ?

☐ Include haplotype list ☐ Include distance matrix ☐ Include genetic structure

2.3 Click the dropdown menu of 'Datatype'. Select the option 'MICROSAT'

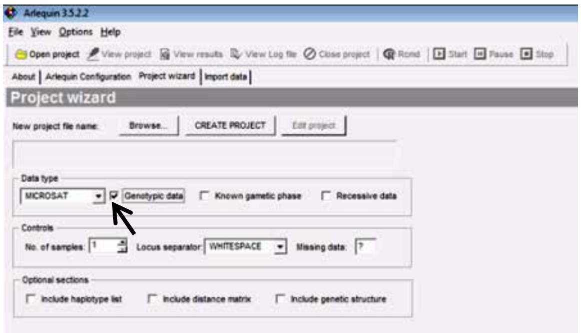


Data type: DNA ☐ Genotypic data ☐ Known gametic phase ☐ Recessive data

Locus separator: WHITESPACE Missing data: ?

☐ Include haplotype list ☐ Include distance matrix ☐ Include genetic structure

2.4 Choose 'Genotype data'


Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
Browse... CREATE PROJECT Edit project

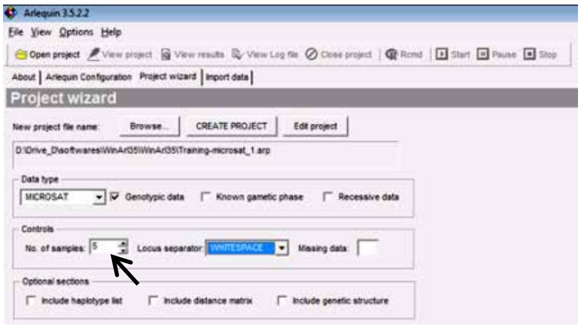
New project file name:

Data type: MICROSAT ☑ Genotypic data □ Known gametic phase □ Recessive data

Controls: No. of samples: 1 Locus separator: WHITESPACE Missing data: ?

Optional sections: □ Include haplotype list □ Include distance matrix □ Include genetic structure

2.5 We have data on five populations for analysis. Therefore mention 'No of samples' as 5.


Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
Browse... CREATE PROJECT Edit project

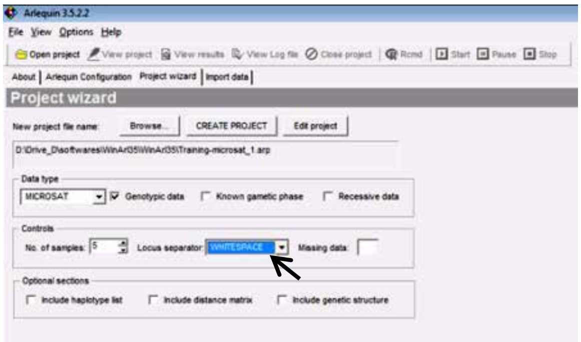
New project file name: D:\Drive_D\softwares\WinArl35\WinArl35\Training-microsat_1.arp

Data type: MICROSAT ☑ Genotypic data □ Known gametic phase □ Recessive data

Controls: No. of samples: 5 Locus separator: WHITESPACE Missing data:

Optional sections: □ Include haplotype list □ Include distance matrix □ Include genetic structure

2.6 Choose 'whitespace' against the 'Locus separator'


Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
Browse... CREATE PROJECT Edit project

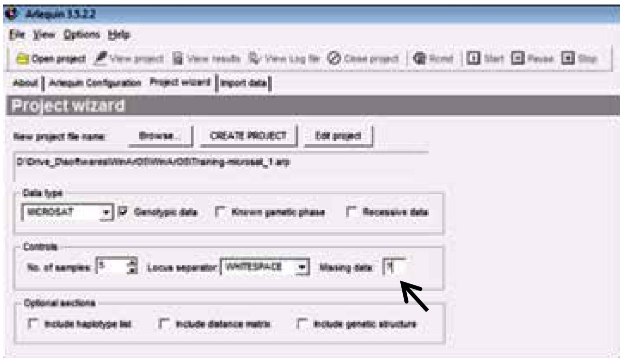
New project file name: D:\Drive_D\softwares\WinArl35\WinArl35\Training-microsat_1.arp

Data type: MICROSAT ☑ Genotypic data □ Known gametic phase □ Recessive data

Controls: No. of samples: 5 Locus separator: WHITESPACE Missing data:

Optional sections: □ Include haplotype list □ Include distance matrix □ Include genetic structure

2.7 Type '?' against 'Missing data'


Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
Browse... CREATE PROJECT Edit project

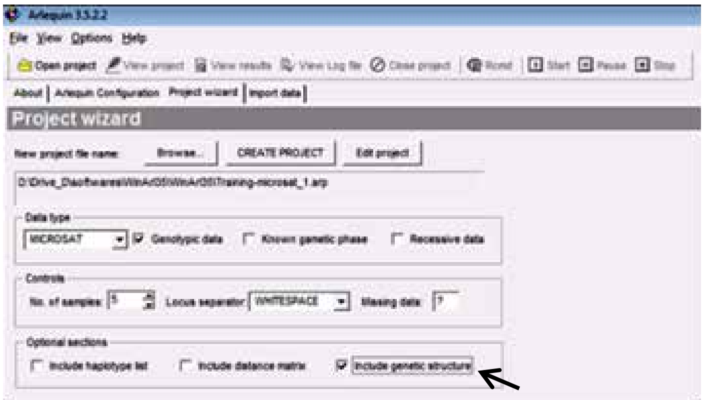
New project file name: D:\Drive_D\softwares\WinArl35\WinArl35\Training-microsat_1.arp

Data type: MICROSAT ☑ Genotypic data □ Known gametic phase □ Recessive data

Controls: No. of samples: 5 Locus separator: WHITESPACE Missing data: ?

Optional sections: □ Include haplotype list □ Include distance matrix □ Include genetic structure

2.8 Select the option 'Include genetic structure'



New project file name: D:\Drive_D\softwares\WinArl35\WinArl35\Training-microsat_1.arp

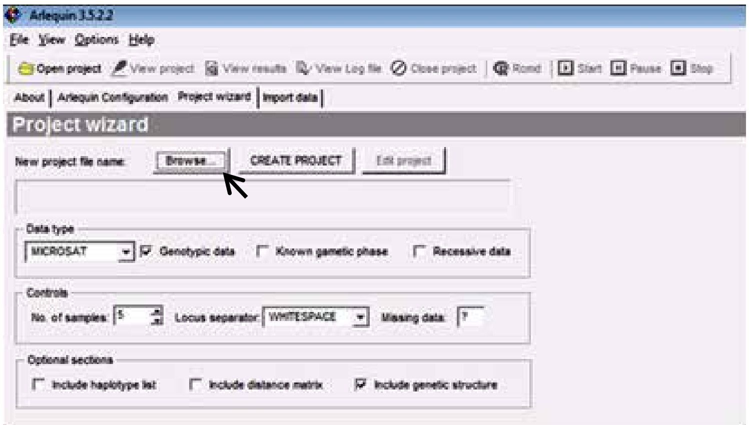
Data type: MICROSAT ☑ Genotypic data □ Known gametic phase □ Recessive data

Controls: No. of samples: 5 Locus separator: WHITESPACE Missing data: ?

Optional sections: □ Include haplotype list □ Include distance matrix ☑ Include genetic structure

Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
Browse... CREATE PROJECT Edit project

2.9 Click the 'Browse' option



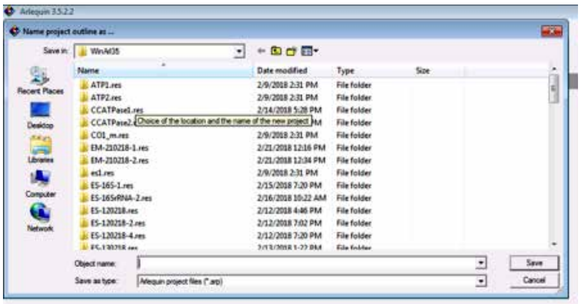
Data type: MICROSAT ☑ Genotypic data □ Known gametic phase □ Recessive data

Controls: No. of samples: 5 Locus separator: WHITESPACE Missing data: ?

Optional sections: □ Include haplotype list □ Include distance matrix ☑ Include genetic structure

Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
New project file name: Browse... CREATE PROJECT Edit project

2.10 It opens a pop-up window where the project file need to be stored as 'ciba-1' in 'WinArl35' folder



Name project outline as ...

Save in: WinArl35

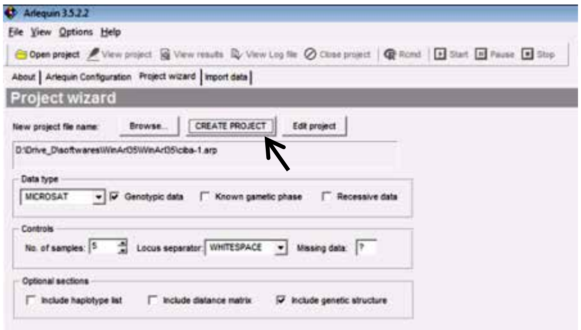
Name	Date modified	Type	Size
ATP1.res	2/9/2018 2:31 PM	File folder	
ATP2.res	2/9/2018 2:31 PM	File folder	
CCATPase1.res	2/14/2018 5:28 PM	File folder	
CCATPase2.	PM	File folder	
CO1_m.res	2/9/2018 2:31 PM	File folder	
EM-210218-1.res	2/21/2018 12:16 PM	File folder	
EM-210218-2.res	2/21/2018 12:34 PM	File folder	
es1.res	2/9/2018 2:31 PM	File folder	
ES-16S-1.res	2/15/2018 7:20 PM	File folder	
ES-16SrRNA-2.res	2/16/2018 10:22 AM	File folder	
ES-120218.res	2/12/2018 4:46 PM	File folder	
ES-120218-2.res	2/12/2018 7:02 PM	File folder	
ES-120218-4.res	2/12/2018 7:20 PM	File folder	
ES-130218.res	2/13/2018 1:22 PM	File folder	

Choice of the location and the name of the new project

Object name:

Save as type: Arlequin project files (*.arp)

2.11 Click on 'CREATE PROJECT' project which creates a project file named 'ciba-1



New project file name: D:\Drive_D\softwares\WinArl35\WinArl35\ciba-1.arp

Data type: MICROSAT ☑ Genotypic data □ Known gametic phase □ Recessive data

Controls: No. of samples: 5 Locus separator: WHITESPACE Missing data: ?

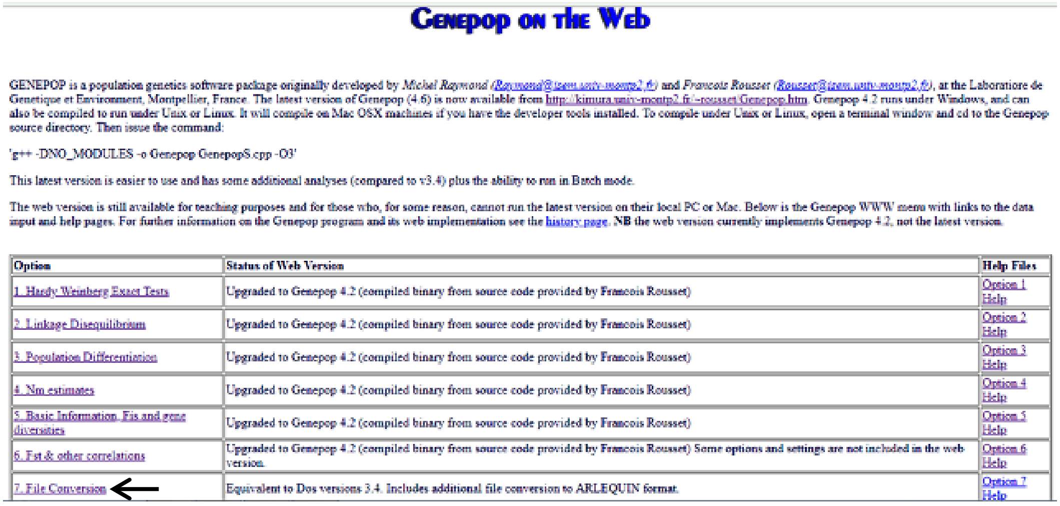
Optional sections: □ Include haplotype list □ Include distance matrix ☑ Include genetic structure

Arlequin 3.5.2.2
File View Options Help
Open project View project View results View Log file Close project Rcmd Start Pause Stop
About Arlequin Configuration Project wizard Import data
Project wizard
Browse... CREATE PROJECT Edit project

2.12 Convert the Genepop format of input file into Arlequin project format

2.12.1 Open the 'Genepop on the web' (http://genepop.curtin.edu.au/)

2.12.2 Click on option 7 (File Conversion) will open the window for Data format conversions.



Genepop on the Web

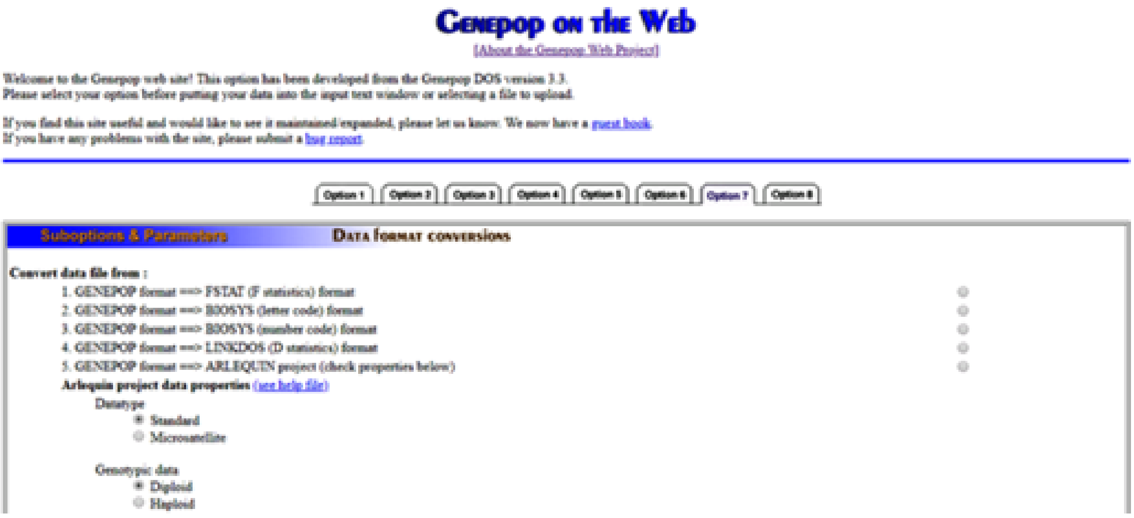
GENEPOP is a population genetics software package originally developed by *Michel Raymond* (Raymond@isem.univ-montp2.fr) and *Francois Rousset* (Rousset@isem.univ-montp2.fr), at the Laboratiore de Genetique et Environment, Montpellier, France. The latest version of Genepop (4.6) is now available from http://kimura.univ-montp2.fr/~rousset/Genepop.htm. Genepop 4.2 runs under Windows, and can also be compiled to run under Unix or Linux. It will compile on Mac OSX machines if you have the developer tools installed. To compile under Unix or Linux, open a terminal window and cd to the Genepop source directory. Then issue the command:

```
'g++ -DNO_MODULES -o Genepop GenepopS.cpp -O3'
```

This latest version is easier to use and has some additional analyses (compared to v3.4) plus the ability to run in Batch mode.

The web version is still available for teaching purposes and for those who, for some reason, cannot run the latest version on their local PC or Mac. Below is the Genepop WWW menu with links to the data input and help pages. For further information on the Genepop program and its web implementation see the history page. NB the web version currently implements Genepop 4.2, not the latest version.

Option	Status of Web Version	Help Files
1. Hardy Weinberg Exact Tests	Upgraded to Genepop 4.2 (compiled binary from source code provided by Francois Rousset)	Option 1 Help
2. Linkage Disequilibrium	Upgraded to Genepop 4.2 (compiled binary from source code provided by Francois Rousset)	Option 2 Help
3. Population Differentiation	Upgraded to Genepop 4.2 (compiled binary from source code provided by Francois Rousset)	Option 3 Help
4. Nm estimates	Upgraded to Genepop 4.2 (compiled binary from source code provided by Francois Rousset)	Option 4 Help
5. Basic Information, Fis and gene diversities	Upgraded to Genepop 4.2 (compiled binary from source code provided by Francois Rousset)	Option 5 Help
6. Fst & other correlations	Upgraded to Genepop 4.2 (compiled binary from source code provided by Francois Rousset) Some options and settings are not included in the web version.	Option 6 Help
7. File Conversion ←	Equivalent to Dos versions 3.4. Includes additional file conversion to ARLEQUIN format.	Option 7 Help



Genepop on the Web

[About the Genepop Web Project]

Welcome to the Genepop web site! This option has been developed from the Genepop DOS version 3.3.
Please select your option before putting your data into the input text window or selecting a file to upload.

If you find this site useful and would like to see it maintained/expanded, please let us know. We now have a guest book.
If you have any problems with the site, please submit a bug report.

Data format conversions

Convert data file from :

1. GENEPOP format ==> FSTAT (F statistics) format ○
2. GENEPOP format ==> BIOSYS (letter code) format ○
3. GENEPOP format ==> BIOSYS (number code) format ○
4. GENEPOP format ==> LINKDOS (D statistics) format ○
5. GENEPOP format ==> ARLEQUIN project (check properties below) ○

Arlequin project data properties (see help file)

Datatype
● Standard
○ Microsatellite

Genotypic data
● Diploid
○ Haploid

2.12.3 Select (option 5) Genepop format to Arlequin project

2.12.4 Select ‘Datatype’ as ‘microsatellite’

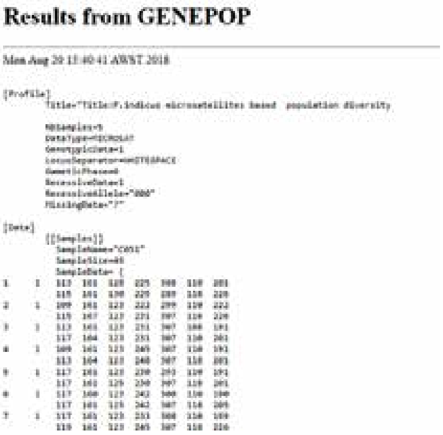
2.12.5 Select ‘Genotypic data’ as ‘diploid’

2.12.6 For ‘Recessive (null) allele present’, select yes or no based on the data. Here our data contains some null alleles. Therefore we select ‘Yes’ option.

2.12.7 For ‘Gametic phase’, select ‘unknown’ option (being a diploid data, gametic phase details are not necessary; the same results will be obtained for either option)

2.12.8 For ‘Output format & Delivery’ select any of the options; ‘Email the results’ or ‘HTML - Plain Text’. Under ‘Email the results’, enter your email id. The results will be sent to your mail id. Plain text option, will display in the same window.

2.12.9 Under ‘Choose File’ option, browse your Genepop file (ciba_genpop_1.txt) and click ‘Submit data’ box. We get the Results in Arlequin project format.



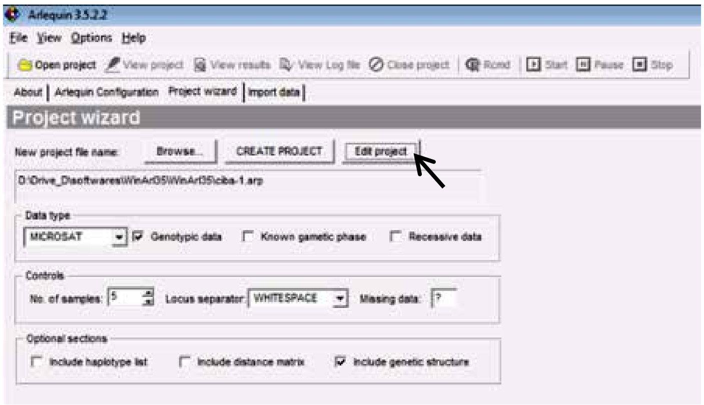
Results from GENEPOP

Mon Aug 20 13:40:41 AWST 2018

```
[Profile]
        Title="Title:P.indicus microsatellites based  population diversity

        NbSamples=5
        DataType=MICROSAT
        GenotypicData=1
        LocusSeparator=WHITESPACE
        GameticPhase=0
        RecessiveData=1
        RecessiveAllele="000"
        MissingData="?"

[Data]
        [[Samples]]
          SampleName="CWS1"
          SampleSize=46
          SampleData= {
1     1   113  161  119  225  308  118  201
          115  161  130  225  289  118  228
2     1   109  161  123  222  299  119  222
          115  167  127  271  307  119  226
3     1   113  161  123  251  307  108  191
          117  164  123  251  307  118  201
4     1   109  161  125  245  307  118  191
          113  164  123  248  307  118  201
5     1   117  161  123  239  293  119  191
          117  161  125  239  307  119  201
6     1   117  160  127  242  308  118  190
          117  161  125  242  307  118  205
7     1   117  165  123  251  308  118  159
          119  161  123  245  307  118  216
```

2.13 Copy the results from '[Data]' to till the end.

2.14 Goto Arlequin software page and click 'Edit project'



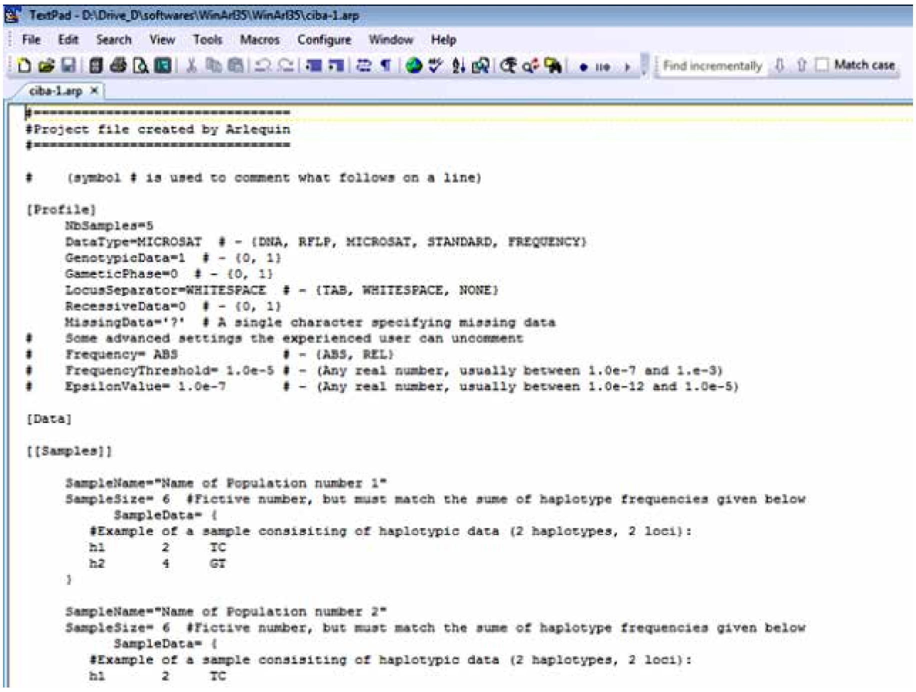
New project file name: D:\Drive_D\softwares\WinArl35\WinArl35\ciba-1.arp

Data type: MICROSAT ☑ Genotypic data ☐ Known gametic phase ☐ Recessive data

Controls: No. of samples: 5 Locus separator: WHITESPACE Missing data: ?

Optional sections: ☐ Include haplotype list ☐ Include distance matrix ☑ Include genetic structure

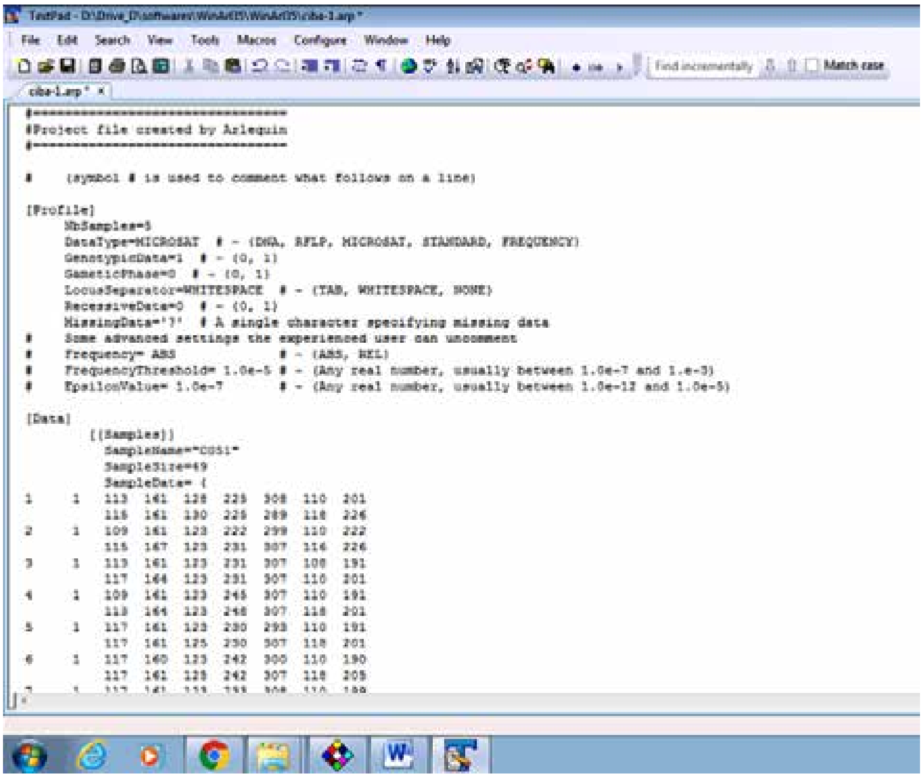
2.15 It will open ciba1.arp file in text pad



TextPad - D:\Drive_D\softwares\WinArl35\WinArl35\ciba-1.arp
File Edit Search View Tools Macros Configure Window Help
Find incrementally
☐ Match case
ciba-1.arp

```
#================================
#Project file created by Arlequin
#================================

#     (symbol # is used to comment what follows on a line)

[Profile]
     NbSamples=5
     DataType=MICROSAT  # - {DNA, RFLP, MICROSAT, STANDARD, FREQUENCY}
     GenotypicData=1  # - {0, 1}
     GameticPhase=0  # - {0, 1}
     LocusSeparator=WHITESPACE  # - {TAB, WHITESPACE, NONE}
     RecessiveData=0  # - {0, 1}
     MissingData='?'  # A single character specifying missing data
#    Some advanced settings the experienced user can uncomment
#    Frequency= ABS           # - {ABS, REL}
#    FrequencyThreshold= 1.0e-5 # - (Any real number, usually between 1.0e-7 and 1.e-3)
#    EpsilonValue= 1.0e-7     # - (Any real number, usually between 1.0e-12 and 1.0e-5)

[Data]

[[Samples]]

     SampleName="Name of Population number 1"
     SampleSize= 6  #Fictive number, but must match the sume of haplotype frequencies given below
          SampleData= {
        #Example of a sample consisiting of haplotypic data (2 haplotypes, 2 loci):
        h1       2      TC
        h2       4      GT
     }

     SampleName="Name of Population number 2"
     SampleSize= 6  #Fictive number, but must match the sume of haplotype frequencies given below
          SampleData= {
        #Example of a sample consisiting of haplotypic data (2 haplotypes, 2 loci):
        h1       2      TC
```

2.16 Paste the copied content in the ciba1.arp file and replace the [Data] content



TextPad - D:\Drive_D\software\WinArl35\WinArl35\ciba-1.arp *
File Edit Search View Tools Macros Configure Window Help
Find incrementally
Match case
ciba-1.arp *

```
#==================================
#Project file created by Arlequin
#==================================

#    (symbol # is used to comment what follows on a line)

[Profile]
     NbSamples=5
     DataType=MICROSAT  # - (DNA, RFLP, MICROSAT, STANDARD, FREQUENCY)
     GenotypicData=1  # - (0, 1)
     GameticPhase=0  # - (0, 1)
     LocusSeparator=WHITESPACE  # - (TAB, WHITESPACE, NONE)
     RecessiveData=0  # - (0, 1)
     MissingData='?'  # A single character specifying missing data
#    Some advanced settings the experienced user can uncomment
#    Frequency= ABS              # - (ABS, REL)
#    FrequencyThreshold= 1.0e-5 # - (Any real number, usually between 1.0e-7 and 1.e-3)
#    EpsilonValue= 1.0e-7        # - (Any real number, usually between 1.0e-12 and 1.0e-5)

[Data]
        [[Samples]]
          SampleName="C051"
          SampleSize=49
          SampleData= {
1     1   113  141  128  225  308  110  201
          115  161  130  225  289  118  226
2     1   109  161  123  222  299  110  222
          115  167  123  231  307  116  226
3     1   113  161  123  231  307  108  191
          117  164  123  231  307  110  201
4     1   109  161  123  245  307  110  191
          113  164  123  248  307  118  201
5     1   117  161  123  230  293  110  191
          117  161  125  230  307  118  201
6     1   117  160  123  242  300  110  190
          117  161  125  242  307  118  205
```

2.17 Edit the 'Structure' content

2.17.1# (Enter the title between inverted commas) (e.g.)

StructureName = "Fish-India"

2.17.2 #Number of groups + {1,2,3...} (Enter 1,2,3 ..Etc., as per the number of groups one has to make) (e.g.)

NbGroups = 1

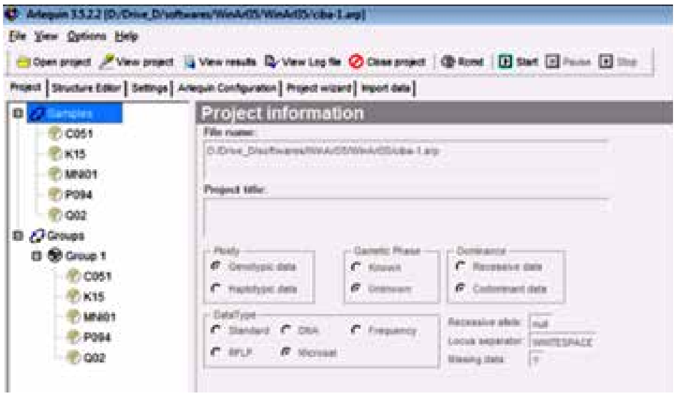
2.17.3#Define hereafter the structure of the first group; mention all the names of the populations. Every population name should be within inverted comma. The populations belong to the specific group has to be mentioned. (e.g.)

```
Group ={   "C051"
           "K15"
            "MNI01"
            "P094"
            "Q02"
            }
```

2.17.4 After editing, save and close the file.

3 Analyzing the data

3.1 Using ‘Open project’ box, open the file ‘ciba-1.arp’



File name: D:/Drive_D/softwares/WinArl35/WinArl35/ciba-1.arp

Project title:

Ploidy: ● Genotypic data ○ Haplotypic data

Gametic Phase: ○ Known ● Unknown

Dominance: ○ Recessive data ● Codominant data

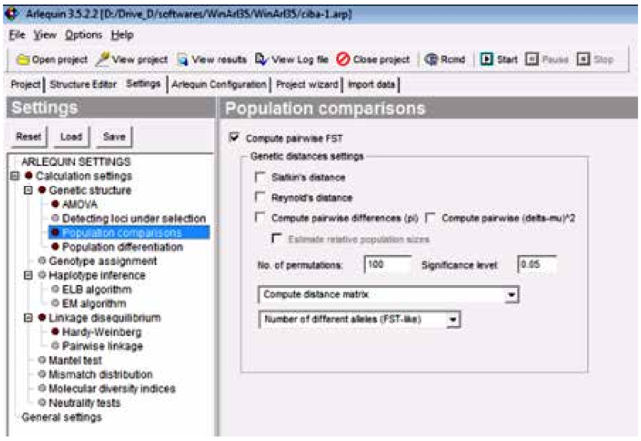
DataType: ○ Standard ○ DNA ○ Frequency ○ RFLP ● Microsat

Recessive allele: null

Locus separator: WHITESPACE

Missing data: ?

3.2 Choose the required analysis from the ‘Settings’



☑ Compute pairwise FST

Genetic distances settings

☐ Slatkin's distance

☐ Reynold's distance

☐ Compute pairwise differences (pi) ☐ Compute pairwise (delta-mu)^2

☐ Estimate relative population sizes

No. of permutations: 100 Significance level: 0.05

Compute distance matrix

Number of different alleles (FST-like)

3.3 Click on ‘Start’ button to start the analysis

3.4 View the results generated in the folder (project file name with the .res extension) ‘ciba-1.res’.

14. Soft Computing techniques in Bioinformatics

P. Mahalakshmi

INTRODUCTION

The exponential growth of the amount of biological data available raises two problems: on one hand, efficient information storage and management and, on the other hand, the extraction of useful information from these data. The second problem is one of the main challenges in computational biology, which requires the development of tools and methods capable of transforming all these heterogenerous data into biological knowledge about the underlying mechanism. These tools and methods should allow us to go beyond a mere description of the data and provide knowledge in the form of testable models. By this simplifying abstraction that constitutes a model, we will be able to obtain predictions of the system. There are several biological domains where soft computing techniques are applied for knowledge extraction from data.

Application of soft computing becomes relevant for solving some Bioinformatics and molecular biology problems. Development in soft computing method reveal the high principles of technology, algorithms, and tools in bioinformatics for enthusiastic reason such as dependable and parallel genome sequencing, fast sequence comparison, search in databases, mechanical gene identification, efficient modeling and storage of mixed data, etc. Protein classification leads to identification and proper functional assignment of uncharacterized proteins with a final goal towards finding homologies and drug discovery. Again, structure based ligand design is one of the crucial steps in rational drug discovery, where a small molecule is designed by targeting the structure and biochemical properties of the target.

The application of soft computing offers an on promising approach to achieve efficient and reliable heuristic solution. On the other side the incessant development of high quality biotechnology, e.g. micro-array techniques and mass spectrometry, which provide complex patterns for the direct characterization of cell processes, offers further promising opportunities for advanced research in bioinformatics. So one important sub-discipline within bioinformatics involves the development of new algorithms and models to extract new, and potentially useful information from various types of biological data including DNA(nucleotide sequences) and proteins (amino acid sequences). Analysis of these macromolecules is performed both structurally and functionally using the major components of soft computing like Fuzzy Sets (FS), Artificial Neural Networks (ANN), Evolutionary Algorithms (EAs) (including genetic algorithms (GAs), Rough Sets (RS), Swarm Optimization (SO) etc. This lecture notes attempts to describe the fuzzy logic, Artificial Neural Networks and genetic algorithm and its applications in bioinformatics.

NEED OF SOFT COMPUTING IN BIOINFORMATICS

The different tasks involved in the analysis of biological data include Sequence alignment, genomics, proteomics, DNA and protein structure Prediction, gene/promoter identification phylogenetic analysis, analysis of Gene expression data, protein Folding, docking and molecule and Drug design. Data analysis tools used earlier in bioinformatics were mainly based on statistical techniques like regression and estimation. Soft computing in bioinformatics can be used in handling large, complex, inherently uncertain, data sets in biology in a robust and computationally efficient manner thus fuzzy sets (soft computing technique) can be used as a natural framework for analysing them. Most of the bioinformatic tasks involve search and optimization of different criteria (like energy, alignment score, overlap strength), while requiring robust, fast and close approximate solutions.

Missing and noisy data is one characteristic of biological data. The conventional computer techniques fail to handle this. Soft computing based techniques are able to deal with missing and noisy data. As soft computing are measured to handle vagueness, indecision and near optimality in large and complex search spaces use of soft computing gear for solving bioinformatics problems have been gained the attention of researchers. Most of the researches are woven around the tasks of pattern recognition and data mining like clustering, classification, feature selection, and rule generation, while classification pertains to supervised or unsupervised learning, clustering corresponds to unsupervised self -organization into homologous partitions.

In molecular biology research, new data and concepts are generated every day, and those new data and concepts update or replace the old ones. Soft computing can be easily adapted to a changing environment. This benefits system designers, as they do not need to re-design systems whenever the environment changes. Moreover, since many of the problems involve multiple conflicting objectives, application of soft computing multi-objective optimization algorithms like multiobjective genetic algorithms appears to be natural and appropriate. Soft computing techniques, either individually or in a hybridized manner, can be used for analyzing biological data in order to extract more and more meaningful information and insights from them.

With advances in biotechnology, huge volumes of biological data are generated. In addition, it is possible that important hidden relationships and correlations exist in the data. Soft computing methods are designed to handle very large data sets, and can be used to extract such relationships.

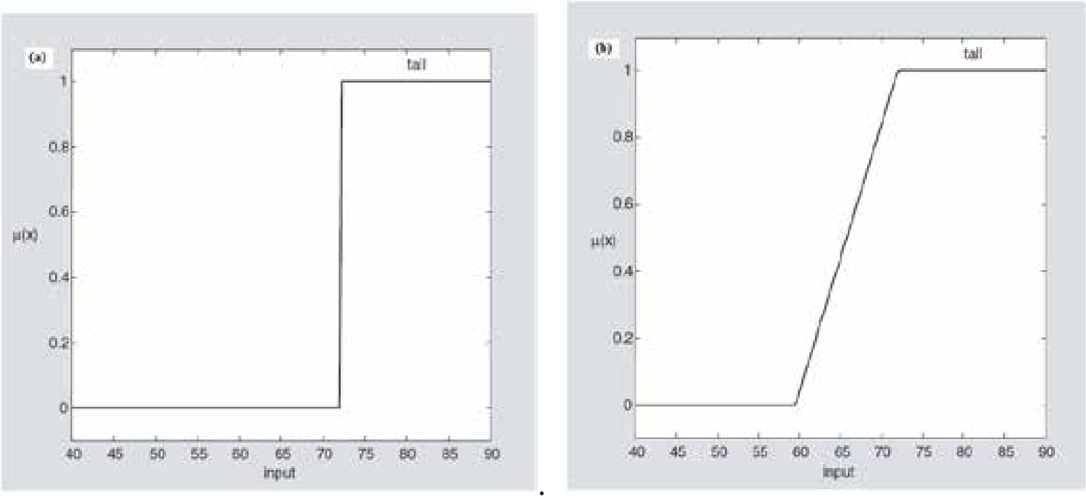
FUZZY LOGIC AND ITS APPLICATION IN BIOINFOAMTICS

Fuzzy Sets and Linguistic Variables

A fuzzy set is an extension of a crisp set. Crisp sets allow only full membership or no membership at all, whereas fuzzy sets allow partial membership. In a crisp set, membership or non membership of element *x* in set *A* is described by a characteristic function, where if and if . Fuzzy set theory extends this concept by defining partial membership, where, where if; if and if *x* partially belongs to *A*. Mathematically, a fuzzy set *A* on a universe of discourse *U* is characterized by a membership function that takes values in the interval [0 1] that can be defined as . Fuzzy set represent commonsense linguistic labels *viz.*, suitable, moderate, unsuitable, slow, very slow, fast etc. A given element can be a member of more than one fuzzy set at a time. A fuzzy set *A* in *U* may be represented as a set of ordered pairs. Each pair consists of a generic element *x* and its grade of membership function; that is,, *x* is called a support value if (Zadeh, 1965). The concept of a linguistic variable plays important role particularly in fuzzy logic. A linguistic variable is a variable whose values are expressed in words or sentences in natural language. For each input and output variables, fuzzy sets are created by dividing its universe of discourse into a number of sub-regions and are named as linguistic variable (Zimmermann, 1996).

Membership Functions

Although both classical and fuzzy subsets are defined by membership functions, the degree to which an element belongs to a classical subset is limited to being either zero or one. This means that membership function may only be a step function (Figure 6.1a). On the other hand, in fuzzy logic, a membership function (MF) is essentially a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1.

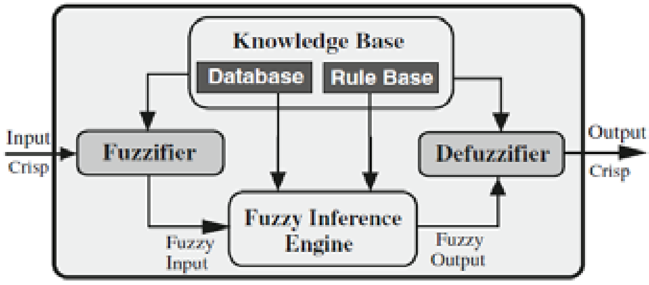


Membership function for (a) crisp set and (b) fuzzy set

The membership functions are usually defined for inputs and outputs in terms of linguistic variables. Various types of membership functions are used, such as triangular, trapezoidal, bell, Gaussian, sigmoid functions. In designing a fuzzy inference system, membership functions are associated with term sets that appear in the antecedent or consequent of rules. Many researchers have used different techniques for determining membership functions such as fuzzy clustering, neural networks, and genetic algorithms

Fuzzy Inference System

Fuzzy Inference System (FIS) incorporate an expert's experience into the system design and they are composed of four blocks. A FIS comprises a fuzzifier that transforms the 'crisp' inputs into fuzzy inputs by membership functions that represent fuzzy sets of input vectors, a knowledge base that includes the information given by the expert in the form of linguistic fuzzy rules, an inference engine that uses them together with the knowledge base for inference by a method of implication and aggregation, and a defuzzifier that transforms the fuzzy results of the inference into a crisp output using a defuzzification method.



Fuzzy Inference System

The knowledge base comprises two components: a database, which defines the membership functions of the fuzzy sets used in the fuzzy rules, and a rule base comprising a collection of linguistic rules that are joined by a specific operator. Based on the consequent type of fuzzy rules, there are two common types of FIS, which vary according to differences between the specifications of the consequent part (Equations 1 and 2). The first fuzzy system uses the inference method proposed by Mamdani in which the rule consequence is defined by fuzzy sets and has the following structure

IF *x* is *A* and *y* is *B* THEN *z is C* (1)

The second fuzzy system proposed by Takagi, Sugeno and Kang (TSK) contains an inference engine in which the conclusion of a fuzzy rule comprises a constant (equation 2 a) or a weighted linear combination of the crisp inputs (equation 2 b) rather than a fuzzy set. A fuzzy rule for the zero-order Sugeno method is of the form

IF *x* is *A* and *y* is *B* THEN z = *C* (2 a)

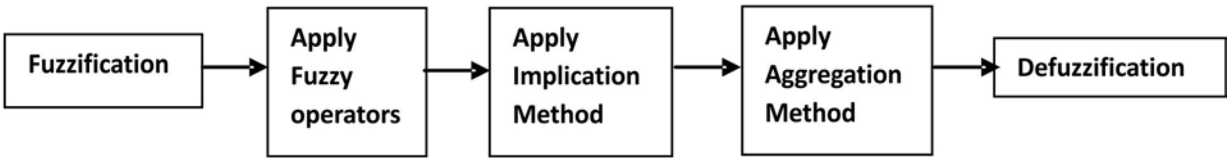
where *A* and *B* are fuzzy sets in the antecedent and *C* is a constant. The first-order Sugeno model has rules of the form

IF *x* is *A* and *y* is *B* THEN *z = px+qy+r* (2 b)

where *A* and *B* are fuzzy sets in the antecedent and *p, q,* and *r* are constants

Fuzzy Inference Process

The inference process for evaluating the system needs five steps



Fuzzy Inference Process

The first step in evaluating the output of a FIS is to apply the inputs and determine the degree to which they belong to each of the fuzzy sets via membership function (Figure 6.5). This is required in order to activate rules that are in terms of linguistic variables. Once membership functions are defined, fuzzification takes a real time input value and compares it with the stored membership function to produce fuzzy input values. In order to perform this mapping, we can use fuzzy sets of any shape, such as triangular, Gaussian, π-shaped, etc.

A fuzzy rule base contains a set of fuzzy rule R. For multi-input, single-output system is represented by

$$R = (R_1, R_2, ..., R_n)$$

where R_i can be represented as

$$R_i = f\ \left(x_1 \quad is \quad T_{x_1},\ \ and......., \ \ x_m \quad is \quad T_{x_m}\right) then \ \ \left(y_1 \quad is \quad T_{y_1}\right)$$

In this rule, m preconditions of R_i form a fuzzy set $(T_{x_1} \times T_{x_2} \times ....\times T_{x_m})$, and the consequent is single output. Generally, if-then-rule can be interpreted by the following three steps:

Resolve all fuzzy statements in the antecedent to a degree of membership between 0 and 1.

If the rule has more than one antecedent, the fuzzy operator is applied to obtain one number that represents the result of applying that rule. This is called firing strength or weight factor of that rule. For example, consider an ith rule has two parts in the antecedent

$$R_i = f\left(x_1 \quad is \quad T_{x_1}^{\ i} \quad and \quad x_2 \quad is \quad T_{x_2}^{\ i}\right) then \quad \left(y \quad is \quad T_y^{\ i}\right)$$

Then, the weight factor can be defined using either intersection operators or product operators

$$\alpha_i = \min\left(\mu_{x_1}^{\ i}(x_1), \mu_{x_2}^{\ i}(x_2)\right)$$

$$\alpha_i = \mu_{x_1}^{\ i}(x_1)\mu_{x_2}^{\ i}(x_2)$$

The weight factor is used to shape the output fuzzy set that represents the consequent part of the rule.

The implication method is defined as the shaping of the consequent, which is the output fuzzy set, based on the antecedent. The input for the implication process is a single number given by the antecedent, and the output is a fuzzy set. Minimum or product are two commonly used methods, which are represented by the following respectively.

$$\mu_y^{\ i}(o) = \min\left(\alpha_i, \mu_y^{\ i}(o)\right)$$

$$\mu_y^{\ i}(o) = \alpha_i \mu_y^{\ i}(o)$$

where*o*is the variable that represents the support value of the membership function.

Aggregation takes all truncated or modified output fuzzy sets obtained as the output of the implication process and combines them into a single fuzzy set. The output of the aggregation process is a single fuzzy set that represents the output variable. The aggregated output is used as the input to the defuzzification process. Aggregation occurs only once for each output variable. Since the aggregation method is commutative, the order in which the rules are executed is not important. The commonly used aggregation method is the maxmethod which can be defined as follows:

$$\mu_y(o) = \max\left(\mu_y^{\ i}(o), \mu_y^{\ i}(o)\right)$$

The defuzzifier maps output fuzzy sets into a crisp number. Defuzzification can be performed by several methods such as: center of gravity, center of sums, center of the largest area, first of the maxima, middle of the maxima, maximum criterion and height defuzzification. Of these, center of gravity (centroid method) and height defuzzification are the methods commonly used. The centroid defuzzification method finds the center point of the solution fuzzy region by calculating the weighted mean of the output fuzzy region. It is the most widely used technique because the defuzzified values tend to move smoothly around the output fuzzy region.

Fuzzy Logic in Bioinformatics

Fuzzy systems have been successfully applied to several areas in practice like for building knowledge-based systems, fuzzy logic-based and fuzzy rule-based models. They can control and analyze processes and diagnose and make decisions in biomedical sciences. There are many application areas in biomedical science and bioinformatics, where fuzzy logic techniques [10] can be applied successfully. Some of the important uses of fuzzy logic are listed below:

- Increasing flexibility of protein motifs.
- Studying differences between various poly nucleotides.
- Analyzing experimental expression data using fuzzy adaptive resonance theory .
- Studying aligning sequences based on a fuzzy dynamic programming algorithm.
- Mathematical modeling of complex traits influenced by genes with fuzzy-valued in pedigreed populations.
- Finding cluster membership values to genes applying a fuzzy partitioning method using fuzzy C-Means and fuzzy c-hard mean algorithms.
- Generating DNA sequencing using genetic fuzzy and neuro-fuzzy systems by anticipating disturbances due to intangible parameters.
- Identifying the cluster genes from micro-array data.
- Predicting protein's sub-cellular locations usingfuzzy k- nearest neighbor's algorithm.

APPLICATION OF ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN) is an information processing model that is able to capture and represent complex input-output relationships. The motivation the development of the ANN technique came from a desire for an intelligent artificial system that could process information in the same way the human brain. Its novel structure is represented as multiple layers of simple processing elements, operating in parallel to solve specific problems. ANNs resemble human brain in two respects: learning process and storing experiential knowledge. An artificial neural network learns and classifies problem through repeated adjustments of the connecting weights between the elements. There are several learning strategies using in bioinformatics: Supervised Learning, Unsupervised Learning and Reinforcement Learning

An ANN learns from examples and generalizes the learning beyond the examples supplied. The methodology of modeling or estimation is somewhat comparable to statistical modeling. Neural networks should not, however, be heralded as a substitute for statistical modeling but rather as a complementary effort (without the restrictive assumption of a particular statistical model) or an alternative approach to fitting non-linear data .Neural networks have been widely used in biology since the early 1990s. Some of the important applications of ANNs are listed below:

- Prediction and the translation sites initiation in DNA sequences and proteins.
- Explain the theory of artificial neural networks using applications in biology.
- Predict immunologically interesting peptides by combining an evolutionary algorithm.

- Carry out pattern classification and signal processing successfully in bioinformatics.
- Perform protein sequence classification.
- Predict protein secondary structure prediction.

GENETIC ALGORITHMS IN BIOINFORMATICS

The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. The applications of GAs are for solving certain multi objective problems of bioinformatics, which yields optimization of computation requirements, and robust, fast and close approximate solutions. GAs are executed iteratively on coded solutions (population) biological basic Operators: selection/ reproduction, crossover, and mutation. They use objective function information and probabilistic transition rules for moving to the next iteration. GAs is generally based on manipulating populations of bit-strings using both crossover and point-wise mutation.

Some of the important applications of GAs are listed below:

- Alignment and comparison of DNA, RNA, and protein sequences.
- Gene mappings in chromosomes.
- RNA structure prediction
- Protein structure prediction and clustering.
- Molecular design and molecular docking.
- Gene finding and promoter identification from DNA sequences.
- Interpretation of gene expression and micro array data.
- Gene regulatory network identification.
- Construction of phylogenetic tree for studying evolutionary relationship.
- DNA structure prediction.

15. RNAseq data analysis – Genome-guided

K. Karthic

1. Introduction

The transcriptomic profile of an organism at any given time or condition gives the set of all its transcripts and their quantities present at the specific time point or condition. The transcriptomereveals a great deal about the functional aspects of the genome as well as the different kinds of biomolecules present within the cell or tissue. It is also very useful for studying the genetics behind growth, development and disease.

This tutorial describes how to analyse RNA-seq data when a reference genome is available and the steps involved in identifying differentially expressed genes between the two groups. For the purpose of demonstration, we have chosen an experiment conducted on *Arabidopsis thaliana*.

1.1. Input files

1. Reference genome in fasta format
2. RNA seq raw data for two groups in replicates in fastq format

1.2. Software requirements

1. Bowtie2
2. Tophat
3. Cufflinks (and associated cuffdiff and cuffmerge)
4. cummerbund (a R package for visualizing the results)

2. Methodology

2.1. Fetching Raw data

(To save time the raw data has been already downloaded and kept in respective folders. So the steps 1 to 15 are to be skipped here)

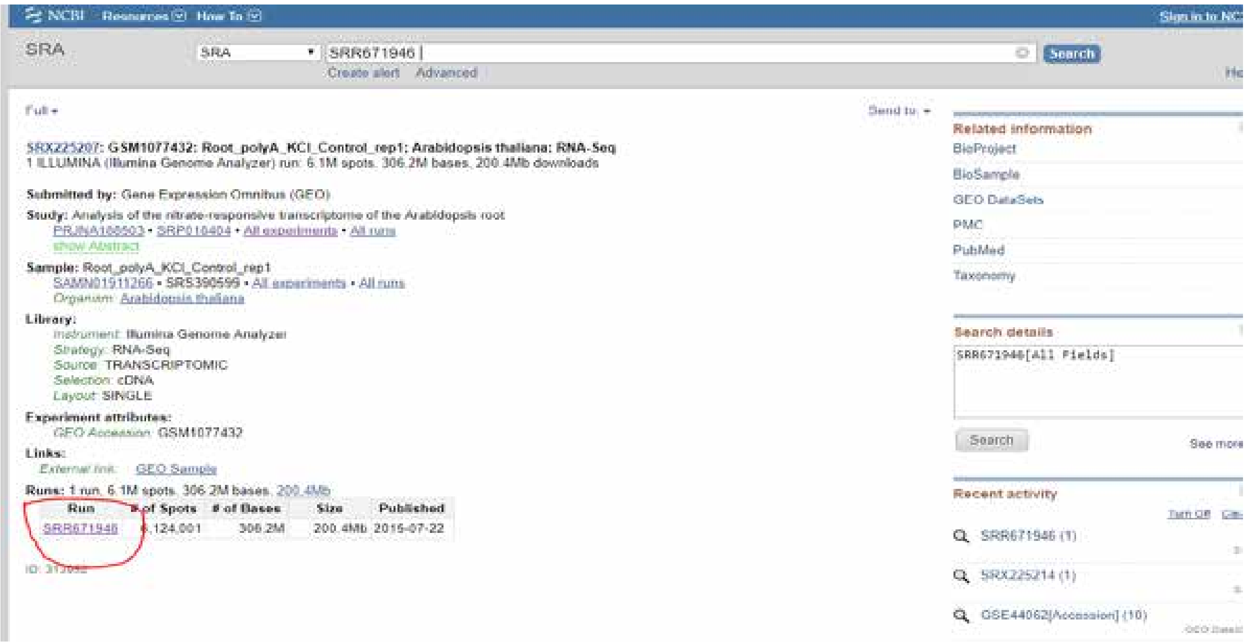
1. Open terminal and create new directories in your account
   ```
   mkdir Athaliana
   cd Athaliana
   mkdir Ref_genome_raw
   mkdir Transcriptome_raw
   ```
2. Go to Assembly database in NCBI [https://www.ncbi.nlm.nih.gov/assembly/] and type TAIR10 in the search bar and click search.
3. The summary of *Arabidopsis thaliana* assembly is displayed. Click on the Download Assemblies button and select Genomic fasta in the drop down menu and click download.
4. The genome downloads as a .tar file, copy the file to Ref_genome_raw folder.
5. Go to terminal and inside the Athaliana folder, type the following commands
   ```
   cd Ref_genome_raw
   tar xvf genome_assemblies.tar
   ```
6. A new folder is created with the name similar to **ncbi-genomes-2018-08-22.** Go to terminal again and type the following commands.

```
cd ncbi-genomes-2018-xx-xx
gunzip GCF_000001735.4_TAIR10.1_genomic.fna.gz
ls –l
```

7. Now you can see the listing of files and in that you notice the fasta file of our genome and its corresponding file size.
8. Go to terminal again and type the following command to copy and save our genome file in a different name and format

```
cat GCF_000001735.4_TAIR10.1_genomic.fna > AraTha.fa
```

9. Now you can see our reference genome saved as **AraTha.fa**
10. To download RNA-seq data, go to Sequence Read Archive (SRA) database of NCBI [http://www.ncbi.nlm.nih.gov/sra] and type the experiment accession numbers SRR671946, SRR671947, SRR671948 and SRR671949 one after the other in the search bar and click search.
11. A summary of the experiment is displayed, scroll down and click on the link displayed below the run.



SRA: SRR671946

SRX225207: GSM1077432: Root_polyA_KCl_Control_rep1: Arabidopsis thaliana: RNA-Seq
1 ILLUMINA (Illumina Genome Analyzer) run: 6.1M spots, 306.2M bases, 200.4Mb downloads

Submitted by: Gene Expression Omnibus (GEO)

Study: Analysis of the nitrate-responsive transcriptome of the Arabidopsis root
PRJNA188503 • SRP018404 • All experiments • All runs
show Abstract

Sample: Root_polyA_KCl_Control_rep1
SAMN01911266 • SRS390599 • All experiments • All runs
Organism: Arabidopsis thaliana

Library:
Instrument: Illumina Genome Analyzer
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE

Experiment attributes:
GEO Accession: GSM1077432

Links:
External link: GEO Sample

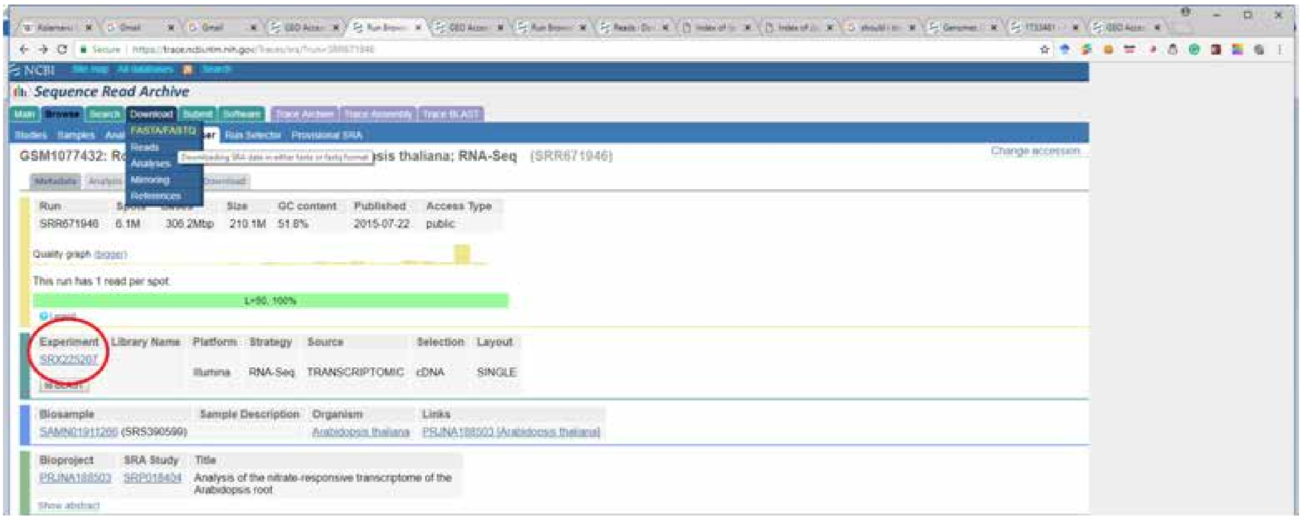
Runs: 1 run, 6.1M spots, 306.2M bases, 200.4Mb

Run	# of Spots	# of Bases	Size	Published
SRR671946	6,124,001	306.2M	200.4Mb	2015-07-22

Search details: SRR671946[All Fields]

Recent activity: SRR671946 (1), SRX225214 (1), GSE44062[Accession] (10)

12. A summary of experiment of *A.thaliana* root treated with KCl, replicate-data is displayed .Go to the downloads tap and click on FASTA/FASTQ link.



GSM1077432: Ro[…]sis thaliana; RNA-Seq (SRR671946)

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR671946	6.1M	306.2Mbp	210.1M	51.8%	2015-07-22	public

This run has 1 read per spot.

L=50, 100%

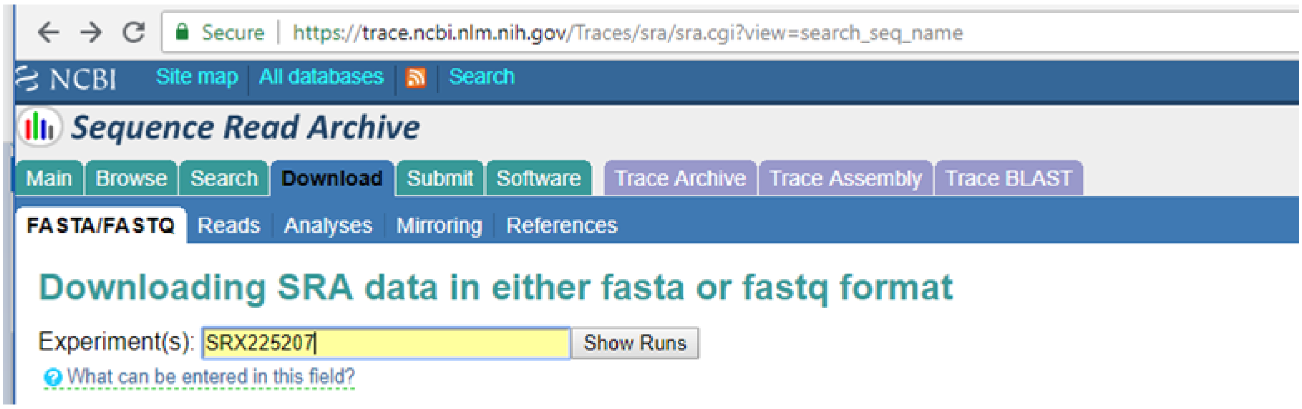
Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX225207		Illumina	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE

Biosample	Sample Description	Organism	Links
SAMN01911266 (SRS390599)		Arabidopsis thaliana	PRJNA188503 (Arabidopsis thaliana)

Bioproject	SRA Study	Title
PRJNA188503	SRP018404	Analysis of the nitrate-responsive transcriptome of the Arabidopsis root

Show abstract

13. In the displayed page, type the experiment number and click show runs


Secure | https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=search_seq_name

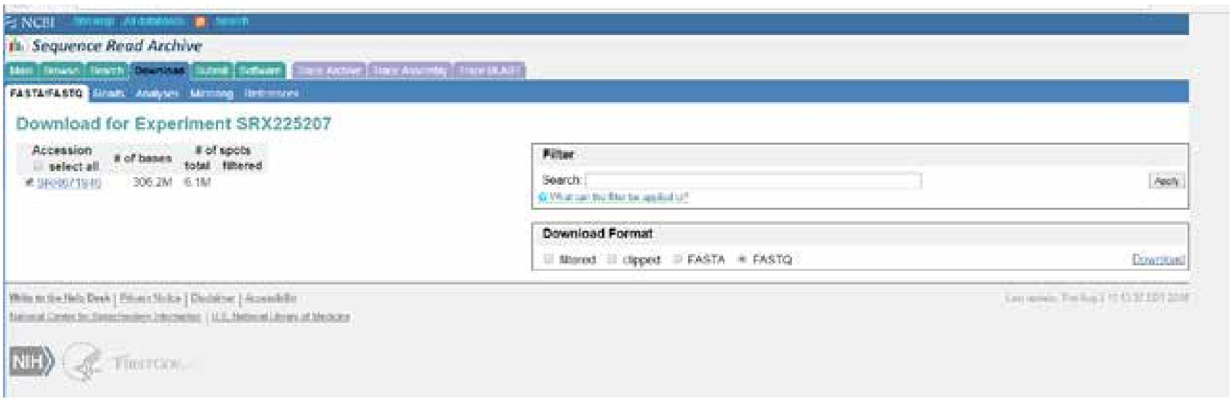
Downloading SRA data in either fasta or fastq format

Experiment(s): SRX225207

Show Runs

What can be entered in this field?

14. Select FASTQ and click download.
15. Repeat steps 12, 13 and 14 for all four experiment runs (SRR671946, SRR671947, SRR671948 and SRR671949).



Download for Experiment SRX225207

Accession □ select all	# of bases	# of spots total	 filtered
☑ SRR671946	306.2M	6.1M	

Filter

Search: Apply

Download Format

□ filtered □ clipped ○ FASTA ● FASTQ Download

16. Copy the downloaded fastq files to folder Transcriptome_raw.
17. Go to terminal and change the directory to Transcriptome_raw
18. Inside the Transcriptome_raw directory, you should have fastq files in zipped format for all the four experiment runs. Go to terminal and type the below command to unzip the files.

```
for i in *.gz;do gunzip $i;done;
```

19. With this we have downloaded all our raw data required for our analysis.

2.2. Data analysis

In this section, how to run bowtie2, tophat, cufflinks,cuffmerge for analyzing the transcriptome data is described. The software installations are not described. Please refer to respective manual for the same.

2.2.1. Indexing genome using bowtie2

1. Go to terminal and change to the directory Ref_genome_raw/ncbi-genomes-2018-xx-xx and type the following command:

```
bowtie2-build AraTha.fa AraTha
```

2. The above command will create bowtie indexed files with .bt2 extension

2.2.2. Running tophat

1. tophat will align the RNA-seq data to our bowtie indexed genome. To do so, type the following command in terminal

```
cd /home/user/Athaliana

mkdir analysis

cd analysis

tophat –o SRR671946_topout /home/user/Athaliana/Ref_genome_raw/AraTha /home/user/Athaliana/Transcriptome_raw/ SRR671946.fastq

tophat –o SRR671947_topout /home/user/Athaliana/Ref_genome_raw/AraTha /home/user/Athaliana/Transcriptome_raw/ SRR671947.fastq

tophat –o SRR671948_topout /home/user/Athaliana/Ref_genome_raw/AraTha /home/user/Athaliana/Transcriptome_raw/ SRR671948.fastq

tophat –o SRR671949_topout /home/user/Athaliana/Ref_genome_raw/AraTha /home/user/Athaliana/Transcriptome_raw/ SRR671949.fastq
```

2. The –o SRR671949_topout represents output folder. For each run, folder is created with the following files: accepted_hits.bam, align_summary.txt, deletions.bed, insertions.bed, junctions.bed, prep_reads.info, unmapped.bam and logs folder.
3. The accepted_hits.bam is the main result file containing the mapped results in binary format.

2.2.3. Running cufflinks

1. From the alignment files generated from tophat, we can assemble the transcripts using cufflinks.
2. In terminal, type the following commands one after another.

```
cufflinks –o SRR671946_cufflinksout /home/user/Athaliana/analysis/ SRR671946_topout/accepted_hits.bam

cufflinks –o SRR671947_cufflinksout /home/user/Athaliana/analysis/ SRR671947_topout/accepted_hits.bam

cufflinks –o SRR671948_cufflinksout /home/user/Athaliana/analysis/ SRR671948_topout/accepted_hits.bam

cufflinks –o SRR671949_cufflinksout /home/user/Athaliana/analysis/ SRR671949_topout/accepted_hits.bam
```

3. For each run, the designated output directory will contain the following files: genes.fpkm_tracking, isoforms.fpkm_tracking, skipped.gtf, transcripts.gtf. The assembled transcripts are contained in transcripts.gtf.

2.2.4. Running cuffmerge

1. cuffmerge will merge the transcripts to a comprehensive transcriptome.
2. Open a text editor, and type the path of the transcripts as below:

```
/home/user/Athaliana/analysis/ SRR671946_cufflinksout/transcripts.gtf
/home/user/Athaliana/analysis/ SRR671947_cufflinksout/transcripts.gtf
/home/user/Athaliana/analysis/ SRR671948_cufflinksout/transcripts.gtf
/home/user/Athaliana/analysis/ SRR671949_cufflinksout/transcripts.gtf
```

and save the file as assembled_transcripts.txt

3. In terminal, type the following command

```
cuffmerge –s  /home/user/Athaliana/Reg_genome_raw/ ncbi-genomes-2018-xx-xx/AraTha.
fa assembled_transcripts.txt
```

4. The successful run creates a merged_asm directory, which contains a logs directory and a file containing the information of the merged transcripts called merged.gtf.

2.2.5. Running cuffdiff

1. cuffidff is used to see differential gene expression in different conditions. Go to terminal and type the following command in a single line.

```
cuffdiff -o diff_result -b /home/user/Athaliana/Reg_genome_raw/ ncbi-genomes-2018-
xx-xx/AraTha.fa-L Root_Kcl_control,Root_KNO3_treatment -u merged_asm/merged.gtf
/home/user/Athaliana/analysis/SRR671946_topout/accepted_hits.bam,/home/user/
Athaliana/analysis/ SRR671947_topout/accepted_hits.bam /home/user/Athaliana/
analysis/SRR671948_topout/accepted_hits.bam/home/user/Athaliana/analysis/
SRR671949_topout/accepted_hits.bam
```

2. The successful run creates a directory diff_result in the working directory. The directory contains a number of different files and databases, listed as follows:

bias_params.info
genes.fpkm_tracking
promoters.diff
tss_groups.fpkm_tracking
cds.fpkm_tracking
genes.read_group_tracking
read_groups.info
tss_groups.read_group_tracking
cds.read_group_tracking
isoforms.read_group_tracking run.info

cds_exp.diff
isoforms.count_tracking
splicing.diff
cds.count_tracking
gene_exp.diff
isoforms.fpkm_tracking
tss_group_exp.diff
cds.diff
genes.count_tracking isoform_exp.diff
tss_groups.count_tracking var_model.info

3. The fpkm tracking files give FPKM counts of primary transcripts (tss_groups.fpkm), genes (genes.fpkm_tracking), coding sequences (cds.fpkm_tracking), and transcripts (isoforms.fpkm_tracking).

4. The count tracking files give the number of fragments for each gene (genes.count_tracking), transcript (isoforms.count_tracking), primary transcript (tss_groups.count_tracking) and coding sequence (cds.count_tracking).

5. The read group tracking files contain information on the counts of genes, transcripts and primary transcripts, grouped by replicates.

6. The diff files ending with 'exp.diff' contain information on the differential expression tests performed on the genes (gene_exp.diff), primary transcripts (tss_group_exp.diff), transcripts (isoform_exp.diff), and coding sequences (cds_exp.diff).

3. Results

3.1. Running cummeRbund

1. cummeRbund is an R package used to visualise the results in different plots.
2. Start an R session In R, go to your working directory and copy the diff_result folder to that.
3. Type the following commands in R

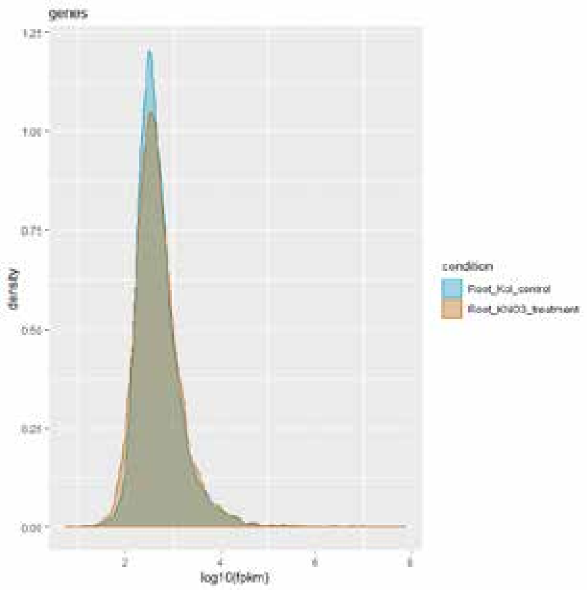
   ```
   >library(‘cummeRbund’)
   >cuffdata < - readCufflinks(‘diff_result’)
   >cuffdata
   ```

4. The above commands will print the result similar to the below

   ```
   CuffSet instance with:
       2 samples
       33318 genes
       42109 isoforms
       34957 TSS
       32921 CDS
       33318 promoters
       34957 splicing
       27174 relCDS
   ```

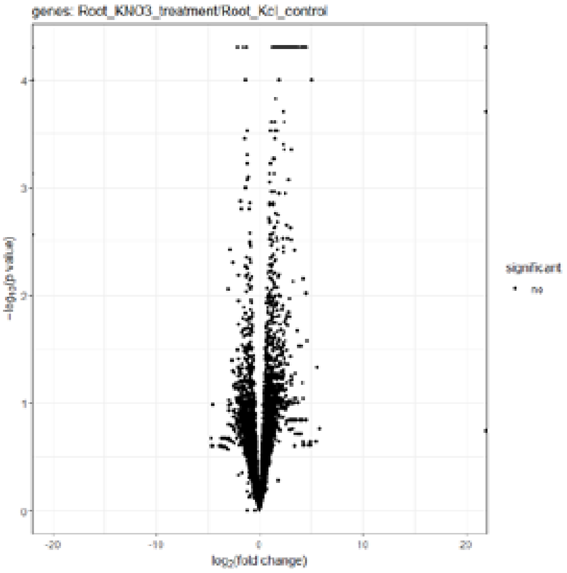
5. To obtain a density plot showing the expression levels for each sample, type the below commands:

   ```
   >csDensity(genes(cuffdata))
   ```



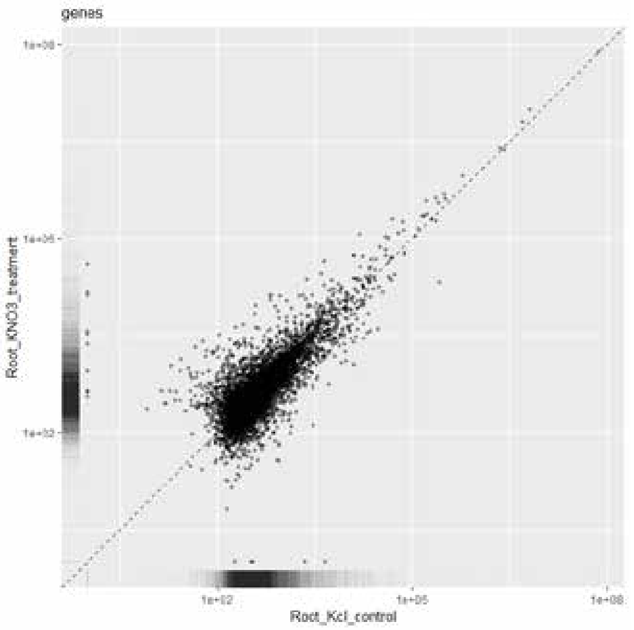
6. To obtain a volcano plot showing the differential expressed genes across the two samples, type the below command:

```
>csVolcano(genes(cuffdata), 'Root_Kcl_control', 'Root_KNO3_treatment')
```



7. To obtain a scatter plot showing the differential expressed genes across the two samples, type the below command:

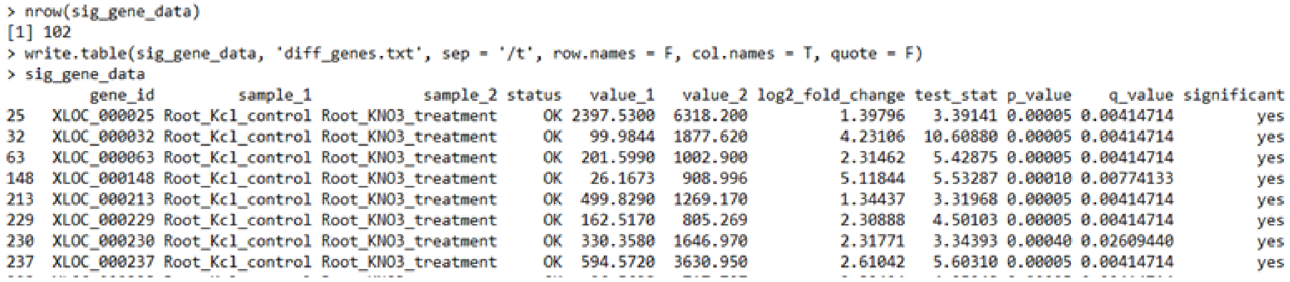
```
>csScatter(genes(cuffdata), 'Root_Kcl_control', 'Root_KNO3_treatment')
```



8. To print a table displaying the details of all the differentially expressed genes, type out the following command.

```
> gene_diff_data < - diffData(genes(cuffdata))

> sig_gene_data < - subset(gene_diff_data, (significant == 'yes'))

>nrow(sig_gene_data)

>write.table(sig_gene_data, 'diff_genes.txt', sep = '/t', row. names = F, col.names = T, quote
= F)

> sig_gene_data
```

The last command prints out a table containing the details of all the differentially expressed genes. The screenshot of the sample output is below:



```
> nrow(sig_gene_data)
[1] 102
> write.table(sig_gene_data, 'diff_genes.txt', sep = '/t', row.names = F, col.names = T, quote = F)
> sig_gene_data
         gene_id         sample_1            sample_2 status    value_1   value_2 log2_fold_change test_stat p_value    q_value significant
25   XLOC_000025 Root_Kcl_control Root_KNO3_treatment     OK 2397.5300  6318.200          1.39796   3.39141 0.00005 0.00414714         yes
32   XLOC_000032 Root_Kcl_control Root_KNO3_treatment     OK   99.9844  1877.620          4.23106  10.60880 0.00005 0.00414714         yes
63   XLOC_000063 Root_Kcl_control Root_KNO3_treatment     OK  201.5990  1002.900          2.31462   5.42875 0.00005 0.00414714         yes
148  XLOC_000148 Root_Kcl_control Root_KNO3_treatment     OK   26.1673   908.996          5.11844   5.53287 0.00010 0.00774133         yes
213  XLOC_000213 Root_Kcl_control Root_KNO3_treatment     OK  499.8290  1269.170          1.34437   3.31968 0.00005 0.00414714         yes
229  XLOC_000229 Root_Kcl_control Root_KNO3_treatment     OK  162.5170   805.269          2.30888   4.50103 0.00005 0.00414714         yes
230  XLOC_000230 Root_Kcl_control Root_KNO3_treatment     OK  330.3580  1646.970          2.31771   3.34393 0.00040 0.02609440         yes
237  XLOC_000237 Root_Kcl_control Root_KNO3_treatment     OK  594.5720  3630.950          2.61042   5.60310 0.00005 0.00414714         yes
```

In this chapter, we described how to download whole genome and transcriptome raw data from NCBI databases. A very brief introduction about the software used in this tutorial was presented and then using the same tools it was demonstrated how to index a whole genome, aligning reads to a reference genome and how to estimate transcript abundance and identify differentially expressed genes. In the end, interpretations of results were visually described.

16. Application of ''*OMICS*'' research in aquaculture with special reference to penaeids

Gopikrishna, G., Vinaya Kumar, K., Shashi Shekhar, M. and Vijayan, K.K.

Introduction

The term *OMICS* refers informally to the field of study in biology as in genomics, transcriptomics, proteomics and metabolomics. Genomics is the study of genomes of organisms, transcriptomics is the study of transcriptomes and so on. Conventional genetic improvement programmes rely mostly on the phenotypic values which are then converted to breeding values on which the selection is carried out. In plants as well as livestock, application of '*omics*' has revealed interesting insights into the genetic and functional biology. When these are integrated within selective breeding programmes, significant improvements have been obtained in productivity.(Dekkers, 2012; Perez-de-Castro *et al* 2012). *Omics* approaches have been applied widely to elucidate the molecular basis of performance traits (eg. growth) and overcome poorly understood biological impediments that impede efficient production (disease, reproductive failure etc) (Rothschild and Plastow 2008, Taylor *et al* 2016).

As far as livestock and plants are concerned, *omics* has had a transformational effect as observed by Agrawal and Narayan (2015); Van Emon (2015) and Taylor *et al* (2016). Coming to the aquaculture sector, the application of selective breeding programmes has been at a snail's pace and it has been suggested that the world aquaculture production could be doubled in a period of 13 years if breeding programmes were supplying stocks for the farmed species (Gjedrem and Rye, 2016). Less than 10% of the aquaculture production is derived from improved lines (Gjedrem*et al* 2012). Looking into the above facts, it is quite clear that *omics* resources in aquatic species need to be developed at a faster pace so that these can be used in selective breeding programmes to hasten genetic response.

Crustaceans form a substantial aquaculture commodity globally. The global penaeid aquaculture industry has exhibited remarkable growth and in 2015, the production stood at 4.8 million tons (FAO, 2017). Penaeids are an important aquaculture resource the world over and it is necessary to have selective breeding programmes so that improved stocks could be generated and farmed. It is well known that the Pacific white shrimp due to its ease of reproductive capability, has been subjected to selective breeding and genetically improved stocks are very much in demand. Information generated from the genomes of shrimp can go a long way in aiding genetic improvement programmes so that the gains are realised at a much faster rate.

Information on whole genome of aquaculture species

Several aquaculture species like *Oncorhynchus mykiss* (Berthelot *et al* 2014), *Oreochromis niloticus* (Conte *et al* 2017) *Lates calcarifer* (Vij *et al* 2016) *Ictalurus punctatus* (Liu *et al* 2016), *Salmo salar* (Lien *et al* 2016) have had their whole genomes deciphered. In India, work on the whole genome sequencing in *Labeo rohita* (Rohu) and *Clarius batrachus* has been carried out at ICAR-NBFGR, Lucknow. Shrimp are unique in that the genome size is comparatively large ~ 2.2 Gbp in tiger shrimp and ~1.8 Gbp in Pacific white shrimp (Guppy *et al* 2018). The highly repetitive nature of the genome in shrimp is a major challenge to the assembly (Huang *et al* 2011; Baranski *et al* 2014). In addition to this, penaeids have a large number of micro-chromosomes and higher levels of genomic heterozygosity (Abdelrahman*et al* 2017) compared to genome assemblies derived from terrestrial farm species. Till date, no comprehensive genome assembly is available for a penaeid shrimp. (Guppy *et al* 2018). There has been a lot of improvement in sequencing especially through the development of high-throughput sequencing, resolving and assembling the many repetitive regions of the penaeid genome (~80%; Abdelrahman *et al* 2017) remains a major challenge.

Transcriptomics

For this, we require the sequence data of the transcriptome. The idea here is to get the mRNA in individuals at a given point in time, thereafter obtain the cDNA and then go in for sequencing. The primary focus of transcriptomics has been immunology, disease resistance and reproductive biology (Guppy *et al* 2018). Generating transcriptome profiles is much easier than generating the whole genome. In *P.vannamei*, while investigating the effect of ammonia exposure, many genes and pathways linked to immune response (eg chitinase, peritrophin, thrombospondin and penaeidin) and growth (linoleic acid metabolism) were identified by Lu *et al* (2016a) to be suppressed. Reproductive dysfunction is a common feature we find in captive broodstock of tiger shrimp. Through differential gene expression studies of whole transcriptome data, genes related to fatty acid and steroid metabolism were found to have altered expression patterns when comparing wild sourced and domesticated stock (Rotllant *et al* 2015).

Linkage mapping of genetic markers in shrimp

One of the genomic resources is the linkage map which provides a wealth of genomic information and also unravel the underlying genetic architecture of commercially and biologically important traits. In penaeids, there have been substantial efforts to generate linkage maps. Linkage maps are constructed using data from family groups viz. parents as well as progeny. Earlier, Amplified Fragment Length Polymrphism was used for construction of linkage maps in tiger shrimp (Wilson *et al* 2002). Later Baranski *et al* 2014 constructed the first linkage map in tiger shrimp using SNPs. Presently, linkage maps are available that include between 3959 and 9298 markers and cover all 44 chromosomes of the penaeid genome (Baranski *et al* 2014, Yu *et a*l 2015, Lu *et al* 2016b, Jones *et al* 2017a) . Such maps have increased the applicability of these resources in assisting genome assembly, examining architecture of traits and also for comparative mapping (Guppy *et al* 2018). It is interesting to note that construction of linkage maps has unravelled some hitherto unknown facts. Baranski *et al* (2014) reported in tiger shrimp that the female–specific map was substantially shorter than the male specific map (2917 vs 4059 cM) whereas in *P.vannamei,*Perez et al (2004) and Zhang *et al* (2007) , reported longer maps for females than males (4134 vs. 3221 cM and 2771 vs. 2116 cM respectively) indicating that there may be higher recombination in males. There is still ambiguity in the karyotype due to the micro-chromosomes in penaeids as a consequence of which it appears that the difference in map length between species exists and sex-based recombination might occur. (Baranski *et al* 2014). Maps available for tiger shrimp (Baranaski *et al* 2014) and Pacific white shrimp (Yu*et al* 2015) have average inter-marker distances between 0.9 and 0.7 cM respectively across different map iterations. This is definitely a significant achievement, however, 1 cM equates to an estimated physical genome distance of ~ 400-600Kb for penaieds (*P. monodon* 395Kb/cM (Baranski *et al* 2014), *P. vannamei* 598.89 Kb/cM (Yu*et al* 2015), *P.japonicus* 657.89Kb/cM (Lu *et al* 2016b) and presents a significant challenge when we look to characterise potential useful genes or genomic regions underlying findings of trait-association studies. (Guppy *et al* 2018). Future work is required to obtain denser maps that decrease the interval between markers. This could be accomplished by genotyping more families and also more individuals per family which would provide additional observations of informative meiotic recombination events or integrate orphaned (unplaced) markers into existing maps (Fierst 2015). Utilising enhanced cost-effective genotyping strategies (eg genotype by sequencing method) could result in genotyping of more families and also more individuals per family consequent to which fine grain marker placement could be achieved (Guppy *et al* 2018).

Developing and applying polymorphic markers

There has been considerable efforts in the past, for development of a wide range of traditional genomic markers (eg. Allozymes, RFLP, AFLP and microsatellites) in several penaeid species. Most of these markers have been used for assessing the wild populations and manage family lines. These markers exhibit caveats which have been reviewed by Benzie (1998, 2009). Due to the high cost in developing them and the failure to unravel the complexity of production traits, they have not found favour in the penaied industry (Guppy *et al* 2018). Today, the traditional markers are being replaced by powerful and cost-effective markers like Single Nucleotide Polymorphisms (SNPs). The SNPs are very abundant in the genome and can help substantially in genome studies. About 9 million SNPs in *Bos taurus* genome (Xu*et al* 2017), 7 million SNPs in chickens (Rubin *et al* 2010) 9.7 million SNPs in Atlantic Salmon (Yanez *et al* 2016), 8.6 million SNPs in channel catfish (Zeng *et al* 2017) and 5.6 million SNPs in *Lates calcarifer* (Vij *et al* 2016) have been identified. The SNP discovery has further led to the manufacture of SNP arrays in several species like cattle, sheep, crops like wheat and in aquaculture species like Catfish and Atlantic Salmon. In *P. monodon*, at ICAR-CIBA, Baranski *et al* (2014) developed a chip containing 6000 SNPs which were majorly identified using the transcriptomic approach. It would be pertinent to point out that till date, only two studies have produced validated SNP genotyping arrays (Baranski *et al* 2014 (6000 SNPs) in tiger shrimp and Jones *et al* (2017b) in Pacific white shrimp (6400 SNPs). The latter one has been sold commercially as the Infinium ShrimpLD-24 v1.0 Bead Chip. An interesting feature of these chips is that these arrays are based on type-I SNPs (genic rather than inter-genic) and many of these SNPs have been annotated with putative genes (62 and 47 %) respectively, thereby providing a strong foundation for further trait mapping studies. (Robinson*et al* 2014 and Khatkar *et al* 2017b). An additional feature that needs to be factored in, is the cost of the SNP arrays. The approximate cost of genotyping per individual has drastically fallen to about Rs. 5000/- This needs to be further reduced to make it cost-effective. Selection of a genotyping method for commercial applications would hinge on the time required for sample processing, genotyping and data analysis, as the window between pre-selection of candidate broodstock at harvest and final breeding selection and spawning is quite short (less than 3-6 months) (Guppy *et al* 2018).

Genotype by sequencing

A unique advantage of the genotype by sequencing (GBS) method is the ability to discover and genotype markers (*de novo* marker discovery) without requiring reference to existing genomic information like genomic sequences and transcriptomes. In penaeids, a number of GBS approaches have been utilised with 25140 and 23049 markers obtained in Pacific white shrimp (Yu *et al* 2015, Wang *et al* 2017) and 28981 markers obtained in Kuruma shrimp (Lu *et al* 2016b). Most of these markers have been utilised to generate linkage maps, undertake Quantitative Trait Loci (QTL) mapping

(Yu *et al* 2015,Lu *et al* 2016b) and estimate genomic prediction accuracy (Wang *et al* 2017), and they have yet to be utilised in the industry for genotyping.

Markers for breeding population management

Crustaceans have a tendency to frequently molt and this places them at a disadvantage in identification. However, tagging with visible implant elastomer tags (for family identification) and eye-ring tags (for individual identification) have been found to address this issue to a certain extent. The number of individuals available per family is rather large in shrimp and they need to be reared in a common environment so that there is no confounding of environmental effects. Each family needs to be reared till tagging and this poses a significant challenge on infrastructure. Tracking of pedigree is of paramount importance to keep the inbreeding low. Use of genomic markers could enhance the identification of individual shrimp but here again the cost of genotyping (high density solid state arrays), lack of genotyping power (microsatellites) or a combination of both these factors are a major stumbling block (Vandeputte and Haffray, 2014).

Exploiting genetic variation underlying phenotypes

It is important to comprehend the relationship between genetic variation and the phenotypes of economically important traits. The information so obtained could prove useful for integrating genomics research into food production industries. (Abdelrahman*et al* 2017). Through QTL mapping and Genome-Wide Association Studies (GWAS), it may be possible to identify the number, location, effect size of genetic elements (i.e. genes, loci and regions) that are linked to the observed phenotypic variation of a trait. (Mackay *et al* 2009). For this to be applied at the field level, we need to identify markers that are highly predictive for a superior or inferior phenotype in order to improve the selection of elite individuals for breeding programmes (Thorgaard *et al* 2006). Genomic breeding values have recently been utilised in breeding programmes related to agriculture in an effort to improve simple and complex traits. (Meuswissen *et al* 2001, 2006). Such a procedure could also be applied to shrimp breeding programmes to elicit substantial genetic response.

QTL mapping

A Quantitative Trait Locus (QTL) is a region in the genome containing one or several genes that affect variation in a quantitative trait which is identified by its linkage to polymorphic marker loci. Mapping of QTLs involves two components: detection and localisation. Once the QTLs are detected, they need to be localised and the gene(s) unravelled. QTLs can be localised through their genetic linkage to visible marker loci with genotypes that we can readily classify. In case a QTL is linked to a marker locus, then individuals with different marker locus genotypes will exhibit different mean values of the quantitative trait. QTLs can be mapped in families or in segregating progeny of crosses between genetically divergent strains (linkage mapping) or in unrelated individuals from the same population (association mapping). Later, these QTLs need to be validated in a population of individuals. If the validation yields encouraging results, the QTLs can be utilised to improve the concerned trait in a breeding programme. Two studies in aquaculture species related to QTL mapping have been reported. One is by Li *et al* (2006) for growth in Kuruma shrimp *P. japonicus* and another by Robinson *et al* (2014) for resistance to White Spot Syndrome Virus in tiger shrimp. In the former case, AFLP markers were used whereas in the case of tiger shrimp SNP markers were used.

Genome-Wide Association Studies

These are studies aimed at associating a particular QTL with a trait. Till date there have been only two studies reported in aquaculture species. The first one is in tiger shrimp, the work of which was carried out at ICAR-CIBA and NOFIMA Norway. Seven families of tiger shrimp were exposed to the White Spot Syndrome Virus. The number of shrimp genotyped was 1024. About 9 QTLs in tiger shrimp were found to be significantly associated to hours of survival. In addition, 3 SNPs were found to be associated with sex in tiger shrimp.(Robinson *et al* 2014). The second study was for growth in *P. vannamei*. The authors could not find any significant association of markers with growth. Earlier, Yu *et al* (2015) while working in *P. vannamei*, had reported a large QTL for growth explaining 17.9% of the phenotypic variation.

Conclusion

Omics research in aquaculture has generated a lot of information during the past three decades. Compared to plant and livestock breeding programmes, aquatic species has a long way to go. The information flowing from various resources like linkage maps, physical maps, annotated transcriptome, characterised proteome data and genome sequence need to be incorporated onto a single platform for use by other scientists working in this field. Wide publicity needs to be given on high-density linkage maps to comprehend genome architecture so as to help in future genetic improvement programmes. Indepth studies on economically important traits in aquatic species are also required urgently so as to help the farmers reap profits from culture of fish/shrimp.

References cited

Abdelrahman, H., ElHady, M., Alcivar-Warren, A., Allen, S., Al-Tobasei, R., Bao, L., et al. (2017). Aquaculture genomics, genetics and breeding in the United States: current status, challenges, and priorities for future research. BMC Genomics 18:191. doi: 10.1186/s12864-017-3557-1

Agrawal, R., and Narayan, J. (2015).Unravelling the impact of bioinformatics and omics in agriculture. Int. J. Plant Biol. Res. 3:1039.

Baranski, M., Gopikrishna, G., Robinson, N. A., Katneni, V. K., Shekhar, M. S.,Shanmugakarthik, J., et al. (2014). The development of a high density linkage map for black tiger shrimp (*Penaeus monodon*) based on cSNPs. PLoS ONE 9:e85413. doi: 10.1371/journal.pone.0085413

Benzie, J. A. (1998). Penaeid genetics and biotechnology. Aquaculture 164, 23–47. doi: 10.1016/S0044-8486(98)00175-6

Benzie, J. A. (2009). Use and exchange of genetic resources of penaeidshrimps for food and aquaculture. Rev. Aquacult. 1, 232–250.doi: 10.1111/j.1753-5131.2009.01018.x

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël,B., et al. (2014). The rainbow trout genome provides novel insights intoevolution after whole-genome duplication in vertebrates. Nat. Commun.5:3657. doi: 10.1038/ncomms4657

Conte, M. A., Gammerdinger, W. J., Bartie, K. L., Penman, D. J., and Kocher, T.D. (2017).

A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*)genome reveals the structure of two sex determination regions. BMC Genomics18:341. doi: 10.1186/s12864-017-3723-5

Dekkers, J. C. (2004). Commercial application of marker-and gene-assistedselection in livestock: strategies and lessons. J. Anim. Sci. 82(13 Suppl.), E313–E328.doi: 10.2527/2004.8213_supplE313x

FAO (2017). FishStat Plus - Universal Software for Fishery Statistical Time Series.

FAO Fisheries and Aquaculture Department. Rome.

Fierst, J. L. (2015). Using linkage maps to correct and scaffold de novo genomeassemblies: methods, challenges, and computational tools. Front. Genet.6:220. doi: 10.3389/fgene.2015.00220

Gjedrem, T., and Rye, M. (2016). Selection response in fish and shellfish: a review.

Rev. Aquacult. 10, 168–179. doi: 10.1111/raq.12154

Gjedrem, T., Robinson, N., and Rye, M. (2012). The importance of selectivebreeding in aquaculture to meet future demands for animal protein: a review.

Aquaculture 350, 117–129. doi: 10.1016/j.aquaculture.2012.04.008

Guppy, J.L., Jones, D.B., Jerry, D.R., Wade, N.M., Raadsma, H.W., Huerlimann, R., and Zenger,K.R. (2018). The State of ''Omics'' Research for farmed penaeids: Advances in research and impediments to industry utilisation.

Front. Genet. 9:282, doi:10.3389/fgene.2018.00282

Huang, S.-W., Lin, Y.-Y., You, E.-M., Liu, T.-T., Shu, H.-Y., Wu, K.-M., et al.(2011). Fosmid library end sequencing reveals a rarely known genomestructure of marine shrimp *Penaeus monodon*. BMC Genomics 12:242.doi: 10.1186/1471-2164-12-242

Jones, D. B., Jerry, D. R., Khatkar, M. S., Raadsma, H. W., Steen, H. V. D.,Prochaska, J., et al. (2017a). A comparative integrated gene-based linkage andlocus ordering by linkage disequilibrium map for the Pacific white shrimp,*Litopenaeus vannamei*. Sci. Rep. 7:10360. doi: 10.1038/s41598-017-10515-7

Jones, D. B., Zenger, K. R., Khatkar, M. S., Raadsma, H. W., Steen, H. A. M. V.D., Prochaska, J., et al. (2017b). "Development of a low-density commercialgenotyping array for the white legged shrimp, *Litopenaeus vannamei*," inAAABG, Edited by Genetics AftAoABa (Townsville, QLD).

Khatkar, M., Coman, G., Thomson, P., and Raadsma, H. (2017a). "Comparisonof different breeding design options for long term genetic gain and diversityin aquaculture species," in Proc Assoc Advmt Anim Breed Genet (Townsville,QLD), 449–452.

Li, Y., Dierens, L., Byrne, K., Miggiano, E., Lehnert, S., Preston, N.,et al. (2006). QTL detection of production traits for the Kuruma prawn*Penaeus japonicus* (Bate) using AFLP markers. Aquaculture 258, 198–210.doi: 10.1016/j.aquaculture.2006.04.027

Lien, S., Koop, B. F., Sandve, S. R.,Miller, J. R., Kent,M. P., Nome, T., et al. (2016).

The Atlantic salmon genome provides insights into rediploidization. Nature533, 500–505. doi: 10.1038/nature17164

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., et al. (2016). The channel catfishgenome sequence provides insights into the evolution of scale formation inteleosts. Nat. Commun. 7:11757. doi: 10.1038/ncomms11757

Lu, X., Kong, J., Luan, S., Dai, P., Meng, X., Cao, B., et al. (2016a).

Transcriptome analysis of the hepatopancreas in the Pacific White Shrimp(*Litopenaeus vannamei*) under acute ammonia stress. PLoS ONE 11:e0164396.

doi: 10.1371/journal.pone.0164396

Lu, X., Luan, S., Hu, L. Y., Mao, Y., Tao, Y., Zhong, S. P., et al. (2016b). Highresolution

genetic linkage mapping, high-temperature tolerance and growthrelatedquantitative trait locus (QTL) identification *inMarsupenaeus japonicus*.Mol. Genet. Genomics 291, 1391–1405. doi: 10.1007/s00438-016-1192-1

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics ofquantitative traits: challenges and prospects. Nat. Rev. Genet. 10, 565–577.doi: 10.1038/nrg2612

Meuwissen, T.,Hayes, B., and Goddard,M. (2001). Prediction of total genetic valueusing genome-wide dense marker maps.Genetics 157, 1819.

Meuwissen, T., Hayes, B., and Goddard,M. (2016). Genomic selection: a paradigmshift in animal breeding. Anim. Front. 6, 6–14. doi: 10.2527/af.2016-0002

Pérez, F., Erazo, C., Zhinaula, M., Volckaert, F., and Calderón, J. (2004).

A sex-specific linkage map of the white shrimp Penaeus (*Litopenaeus) vannamei)* based on AFLP markers. Aquaculture 242, 105–118.doi: 10.1016/j.aquaculture.2004.09.002

Pérez-de-Castro, A. M., Vilanova, S., Cañizares, J., Pascual, L., Blanca, J. M., Diez,M. J., et al. (2012). Application of genomic tools in plant breeding.Curr.Genomics 13, 179–195.

doi: 10.2174/138920212800543084

Robinson, N. A., Gopikrishna, G., Baranski, M., Katneni, V. K., Shekhar, M.S., Shanmugakarthik, J., et al. (2014). QTL for white spot syndrome virusresistance and the sex-determining locus in the Indian black tiger shrimp(*Penaeus monodon*).

BMC Genomics 15:731. doi: 10.1186/1471-2164-15-731

Rothschild, M. F., and Plastow, G. S. (2008).Impact of genomics on animalagriculture and opportunities for animal health.Trends Biotechnol. 26, 21–25.

doi: 10.1016/j.tibtech.2007.10.001

Rotllant, G.,Wade,N.M., Arnold, S. J., Coman, G. J., Preston,N. P., and Glencross,B. D. (2015). Identification of genes involved in reproduction and lipid pathwaymetabolism in wild and domesticated shrimps. Mar. Genomics 22, 55–61.

doi: 10.1016/j.margen.2015.04.001

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E.,Webster, M. T., et al. (2010). Whole-genome resequencing reveals loci underselection during chicken domestication. Nature 464:587. doi: 10.1038/nature08832

Taylor, J. F., Taylor, K. H., and Decker, J. E. (2016). Holsteins are thegenomic selection poster cows. Proc. Natl. Acad. Sci. U.S.A. 113, 7690–7692.doi: 10.1073/pnas.1608144113

Thorgaard, G. H., Nichols, K.M., and Phillips, R. B. (2006).Comparative gene andQTL mapping in aquaculture species.Israeli J. Aquacult.Bamidgeh 58, 4.

Van Emon, J. M. (2015). The omics revolution in agricultural research.

J. Agric.Food Chem. 64, 36–44. doi: 10.1021/acs.jafc.5b04515

Vandeputte, M., and Haffray, P. (2014). Parentage assignment with genomicmarkers: a major advance for understanding and exploiting genetic variationof quantitative traits in farmed aquatic animals. Front. Genet. 5:432.doi: 10.3389/fgene.2014.0043

Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., VanHeusden, P., et al. (2016). Chromosomal-level assembly of the Asian seabassgenome using long sequence reads and multilayered scaffolding. PLoS Genet.12:e1005954.

doi: 10.1371/journal.pgen.1005954

Wang, Q., Yu, Y., Yuan, J., Zhang, X., Huang, H., Li, F., et al. (2017). Effects ofmarker density and population structure on the genomic prediction accuracyfor growth trait in Pacific white shrimp *Litopenaeus vannamei*.

BMC Genet.18:45. doi: 10.1186/s12863-017-0507-5

Wilson, K., Li, Y. T., Whan, V., Lehnert, S., Byrne, K., Moore, S., et al.(2002). Genetic mapping of the black tiger shrimp *Penaeus monodon*with amplified fragment length polymorphism. Aquaculture 204, 297–309.doi: 10.1016/S0044-8486(01)00842-0

Xu, C., Li, E., Liu, Y., Wang, X., Qin, J. G., and Chen, L. (2017).Comparativeproteome analysis of the hepatopancreas from the Pacific white shrimp*Litopenaeus vannamei* under long-term low salinity stress. J. Proteomics 162,1–10. doi: 10.1016/j.jprot.2017.04.013

Yáñez, J. M., Naswa, S., López, M., Bassini, L., Correa, K., Gilbey, J., et al.(2016). Genomewide single nucleotide polymorphism discovery in Atlanticsalmon (*Salmo salar*): validation in wild and farmed American and Europeanpopulations.

Mol. Ecol. Resour. 16, 1002–1011. doi: 10.1111/1755-0998.12503

Yu, Y., Zhang, X., Yuan, J., Li, F., Chen, X., Zhao, Y., et al. (2015). Genomesurvey and high-density genetic map construction provide genomic and geneticresources for the PacificWhite Shrimp *Litopenaeus vannamei*. Sci. Rep. 5:15612.doi: 10.1038/srep15612

Zeng, Q., Fu, Q., Li, Y., Waldbieser, G., Bosworth, B., Liu, S., et al. (2017).

Development of a 690K SNP array in catfish and its application for geneticmapping and validation of the reference genome sequence. Sci. Rep. 7:40347.doi: 10.1038/srep40347

Zhang, L., Yang, C., Zhang, Y., Li, L., Zhang, X., Zhang, Q., et al. (2007). Agenetic linkage map of Pacific white shrimp (*Litopenaeus vannamei*): sex-linkedmicrosatellite markers and high recombination rates. Genetica 131, 37–49.doi: 10.1007/s10709-006-9111-8

17. Shrimp Genomics : Current status and Challenges

M.S. Shekhar, K. Vinaya Kumar and K.K. Vijayan

The shrimp genomics has evolved a into a considerable research progress over last few decades. The recent advances in "omics" in particular with the advancement in NGS techniques, have provided the aquaculture industry the opportunities as well the challenges faced in understanding the complexity of the whole genome of shrimp. However, the currently available molecular biology resources and bioinformatics techniques require further development to undertake the challenges and provide the most informative results in deciphering the shrimp genome.

1. Introduction

The consumption of food fishes globally is projected to increase tremendously. However, with exploitation and decrease in wild catch fisheries worldwide, much importance is now being given to increase the production from aquaculture. In aquaculture and fisheries management for an effective genetic improvement breeding programs, studies relating to population structure, genetic diversity, environmental adaptation and molecular response to biotic and abiotic stress are very important. "Biotechnology" integrated with "Omics" is a term that has now come to encompass many of the exciting new developments in aquaculture during recent years. Hence, for sustainable aquaculture, genetic improvement for desired traits etc. through biotechnological means has gained importance in recent years. Aquaculture biotechnology deals with the use of knowledge and techniques in the field of molecular, cellular and genetic processes to develop improved aquaculture products and varieties. Therefore, a wide term of 'omics' which includes methods and techniques that are required for analyzing all different types of molecules and the pathways associated with them is used in aquaculture as well. This encompasses the major four "omics", namely transcriptomics, proteomics, metabolomics and epigenomics. Viral infections are one of the major reasons for the huge economic losses in shrimp farming. The control of viral diseases in shrimp remains a serious challenge for the shrimp aquaculture industry. White spot syndrome virus (WSSV), is a major pathogen which is geographically widespread and continues to be a serious threat affecting shrimp farms the world over. In the absence of a true adaptive immune response system in invertebrates, shrimps respond by non-specific innate immune mechanisms. Shrimp genome annotation and transcriptome generation as "omics" tools would aid to unravel the molecular mechanisms involved in the immune defence network that occur in shrimp in response to WSSV infection in addition to development of genetically improved varieties of shrimp with desirable traits through genetic improvement breeding programmes.

2. Transcriptomics

Next-generation high-throughput RNA sequencing technology (RNA-seq) is a modern and a high throughput method which is not restricted by the unavailability of a genome reference sequence has tremendous potential for identification, profiling and quantifying RNA transcripts with increased sensitivity. Transcriptome is the complete set of transcripts in a cell, indicating a specific developmental stage or physiological condition together with the quantity. Transcriptome helps in identifying the functional elements of genome revealing molecular constituents of cells and tissues, in response to environmental stress with an accurate quantification of gene expression levels. Because of these several advantages over other techniques expression this approach has been widely used now in decoding the functional role of gene and cell responses against environmental stress. Significant progress has been recently achieved in understanding the transcript expression of marine crustaceans such as *Litopenaeus vannamei*, *Fenneropenaeus chinensis*, *Eriocheir sinensis* and *Macrobrachium nipponense* in response to biotic and abiotic stress factors. Transcriptome data aids in identification of novel genes in absence of shrimp genome database as shown in Table 1. Next-generation sequencing technologies have therefore influenced the analysis of gene regulation.

Table 1. Transcriptomes generated from shrimp species

Species	Tissues	Transcriptome generation
L. vannamei	Hemocytes	WSSV
L. vannamei	Hepatopancreas	WSSV
L. vannamei	Hepatopancreas and muscle	WSSV and growth
L. vannamei	Hemolymph and hemocytes	TSV
L. vannamei	Hepatopancreas	TSV
L. vannamei	Testis and Ovaries	Gonadal development
L. vannamei	Hepatopancreas	Acute ammonia stress
L. vannamei	Hepatopancreas	Osmoregulatory Stress
L. vannamei	Gills	Osmoregulatory Stress
L. vannamei	Hepatopancreas and hemocytes	Nitrite
L. vannamei	Whole larvae	Embryo development
L. vannamei	Embryo, Nauplius, zoea, mysis, post larvae	Larval Development
L. vannamei	Whole shrimp	Molting
L. vannamei	Muscle	Feed efficiency
L. vannamei	Heart, muscle, hepatopancreas and eyestalk	Growth
P. monodon	Hepatopancreas and ovary	Reproduction and development
P. monodon	Eyestalk, stomach, female gonad, male gonad, gill, haemolymph, hepatopancreas, lymphoid organ, tail muscle, embryos, nauplii, zoea, and mysis, whole larvae	Gene discovery
M. japonicus	Fertilized eggs, embryos and vegetal halves	Embryo development
F. chinensis	Cephalothorax	WSSV
F. merguiensis	Cuticle, muscle, androgenic gland, hepatopancreas, stomach, nervous system, eyestalk, male gonads, female gonads	Color
F. merguiensis	Hepatopancreas, stomach, eye stalk, nerve cord, male gonad, female gonad, androgenic gland region, muscle and cuticle	Reproduction and development

3. Complexity of shrimp genome

Shrimp genomes are large with highly repetitive sequences which pose significant challenges in deciphering the whole genome and other genetic studies. In our study, the shrimp genome estimated by flow cytometry showed the shrimp genome to be of very high size. The genome size for the four major species of genus *Penaeus* (*Penaeus monodon, Penaeus indicus, Penaeus vannamei and Penaeus japonicus*) were found in similar range. The genome size of female shrimps ranged from 2.91 ± 0.03 pg (*P. monodon*) to 2.14 ± 0.02 pg (*P. japonicus*). In male shrimps, the genome size ranged from 2.86 ± 0.06 pg (*P.monodon*)to 2.19 ± 0.02 pg (*P. indicus*). Significant difference was observed in the genome size between male and female shrimp of all species except in *P.monodon.* The highest relative difference of 12.78% was observed in the genome size between the either sex in *P.indicus.* The interspecific relative difference of 30.59% in genome size was highest between the male shrimps of *P. monodon* and *P. indicus* and 35.98% between the female shrimps of *P. monodon* and *P. japonicus.* This study was undertaken to estimate genome size in shrimps which will help guiding the research aimed towards generating the sequence data for the whole genome of these species in future. The penaeid genome (80% repetitive) remains a challenge even today for sequencing and assembly. Short read second-generation sequencing methods for example illumina sequencing technology is preferred for non-complex genomes, by identifying and overlaying sequences and building the resulting contigs and scaffolds. However, when short read sequencing methods are applied to highly repetitive regions within the genome, it leads to difficulty in building contiguous sequences. The shrimp genomes also have high levels of heterozygosity. The previous short-read assembly in shrimps have led highly fragmented assembly with high number of scaffolds. There are reports that shrimp with polysaccharides contamination and high DNase activity can interfere with long read sequencing methodologies which are major challenges to overcome and methods to isolate intact pure shrimp genome needs more standardization.

4. NGS platform for shrimp genome sequencing

Several NGS platforms are currently in use such as Illumina MiSeq, Ion Torrent PGM, PacBio RS, Illumina GAIIx, Illumina HiSeq 2000, etc. The key feature which determines the optimal platform to be used is their speed of sequencing with less of error rates. The sequencing methodology has been dominated by Illumina. However, the use of this technology is not adequate in dealing with complex shrimp genomes which requires generation of longer read lengths. One such latest platform which yields longer read lengths is PacBio. PacBio is based on single molecule real time (SMRT) sequencing. The DNA polymerase molecules, binds to a DNA template, are present at the base of 50 nm-wide wells called zero-mode waveguides (ZMWs). Second strand DNA synthesis in the presence of γ-phosphate fluorescently labeled nucleotides is carried out by each polymerase. With each base incorporation, a distinctive pulse of fluorescence is detected in real time. The PacBio platform, by virtue of its long read lengths, has the potential application in *de novo* sequencing of shrimp genome. Approx. mean read lengths of 1500 bp were generated using the PacBio RS system with the first generation of chemistry (C1 chemistry) , the advanced PacBio RS II system with the C4 chemistry yields average read lengths over 10 kb , with an N50 of more than 20 kb and maximum read lengths over 60 kb. The latest PacBio Sequel System is a advanced version with higher throughput, more scalability, a reduced footprint and lower sequencing project costs compared to the PacBio® RS II System. This advanced version of the Sequel System is the capacity of its redesigned SMRT Cells, which contain one million zero-mode waveguides (ZMWs) as, compared to 150,000 ZMWs in the PacBio RS II. Active individual polymerases are immobilized within the ZMWs, providing windows to observe and record DNA sequencing in real time. In future the successful assemblies for shrimp genome will depend upon a "hybrid assembly" approach, utilizing short-read sequencing to correct the high error rate observed in long read PacBio sequencing system.

5. The Challenges

In comparison to other livestock industries, very less improved lines are used in aquaculture production (Gjedrem et al., 2012). The aquaculture production has also not completely utilized the existing natural genetic potential and resources for increased productivity. In case of shrimp, there have been numerous molecular studies on the expression and function of selected genes involved in metabolic pathways, however, little attention is given to the metabolic differences which exist between shrimp or to their developmental stages. The difference among shrimp due to result of particular metabolic and adaptations to varied environmental conditions needs to be studied in detail. These types of studies have direct relevance to the better management practices and formulation of optimal diets for the domestication of shrimp in aquaculture. In the immediate future, the main challenges are to integrate the available genomic data with physiological studies on shrimp. These outcomes will elucidate species-specific adaptations to environmental conditions, and have the potential to inform and stimulate research in many biological disciplines. For, any genomic studies and analysis, a reference genome is essential, however, except for a brachiopod *Daphnia pulex*, no information on complete genome assembly is available from other crustaceans. The genome size of *D. pulex* is comparatively smaller in size of about 200 Mb, containing 30,970 genes and very less 9.4% repetitive sequences, however in shrimp, the genomes are too big and complex for sequencing and assembly. Bioinformatics, data mining and sequence annotation needs to be defined and developed for complex genomes which would aid in complete genome assembly.

6. Future potential

Introducing of improved bioinformatics approach for error-correction of longer read sequencing lengths and use of optical mapping would help in completing the large size genome assembly of shrimps and other aquatic species. There is also an urgent need to construct linkage and physical maps, and to develop database for annotated transcriptome, proteomics and metabolomics, which would help in generating highly informative shrimp "omics" to understand genome structure, genome evolution, phylogeny and natural selection of aquaculture species. The functional genomics with annotated genome and validation of candidate genes by experimental CRISPER or RNAi knockdown studies would be significant progress towards in identification of target genes of commercial importance such as growth, and disease resistance. Understanding the genome and genetic makeup of shrimp would benefit in deciphering complex traits which would eventually accelerate the breeding program in shrimp. A high-density linkage map is essential for shrimp genomics and genetic studies. Creation of a high-density linkage map would help in mapping of QTLs for traits of interest such as body weight, body length, disease resistance and other traits which have high commercial significance in aquaculture.

18. Application of Biotechnology in animal reproduction

Sherly Tomy

Reproductive efficiency is a major factor determining the economic success of any livestock enterprise.Majority of the animal breeding programs have aimed at enhancing the genetic worth of animals using conventional selection methods primarily based on phenotype. Revolutionary tools in reproductive biotechnology like use of recombinant DNA procedures, genome engineering, transgenic technology, somatic cell nuclear transplantation etc has added new dimensionstoanimal breeding. Application of biotechnology in animal breeding has resulted in several remarkable discoveries like the sheep Dolly, created by the somatic cloning technique, transgenic pigs that can be organ donors for humans, and animal bioreactors producing human therapeutic proteins in milk. Compared to the terrestrial animals, the development in aquatic animals is comparatively less. Only a small percentage of farmed aquatic species have been subject to genetic improvement programmes. However, biotechnology have great potential to increase fish production mainly due to the availability of large numbers of gametes, use of external fertilization, and ease of hormone treatments during development to induce sterility or functional sex reversal. Some of the important reproductive biotechnological tools used in farm animals are:

Artificial insemination: Using this technology new breeds of animals are produced through the introduction of the male sperm from one superior male to the female reproductive tract without mating. The advantage of AI includes reduced transmission of venereal disease, lessens the need of farms to maintain breeding males, facilitates more accurate recording of pedigrees, and minimizes the cost of introducing improved genetics. However, success of AI depends on accurate heat detection, proper frozen semen handling and timely insemination by a trained inseminator.

Sex Determination of Sperm: Sexing of sperm could help to pre-determine the sex of the progeny. This technique works on the principle of flow cytometric separation of fluorescent-labelled X-chromosome bearing spermatozoa from the sperms carrying fluorescent-labelled Y-chromosome. The accuracy of this technique is high, however, the laser light used reduces the viability of the sexed sperm and the throughput is low.However, new generation flow cytometer with high sorting rates have opened avenues for increasing sorted sperm output with minimal or no damage to sperm. Sex chromosome-specific proteins (SCSPs) identified on the surface of sperm are also currently used for sperm sexing which are less invasive and less damaging to sperm.

Sperm Encapsulation: This involves encapsulation of sperm for longer preservation of sperm *in vivo* and to allow progressive release of viable spermatozoa over several days in various domestic species including human. The technique also prevents cryocapacitation and also reported to have increased conception rate. The technique has been developed in cattle and swine, still it needs more sophisticated instrument for encapsulation and standardization, to be used under field conditions in other livestock species.

Ovum Pick Up (OPU) :This is a non-invasive and repeatable technique used for recovering large numbers of competent oocytes from antral follicles of live animals. Embryo production from ovum pick-up oocytes is affected by age, season, follicle stimulating hormone (FSH) stimulation. It also evident that repeated OPU can be performed without side effects both in cattle and buffaloes with a minimal stress to the animal. In India, the first buffalo calf (Saubhagya) was produced through this technique by Prasad *et al.*2013, and subsequently, first bovine calf (Holi) was produced at ICAR-National Dairy Research Institute. OPU has advantage to collect oocytes from animals with less invasiveness and the use of superior animals as oocyte donors in embryo transfer. One of the limitations of this technique is the low oocyte yield per ovary and necessity for sophisticated instrument for carrying out this technique.

In Vitro Maturation, Fertilization and Culture (IVMFC) :This involves oocyte collection from slaughterhouse ovaries or from live animals followed by maturation and fertilization *in vitro* for the production of viable embryos. Since the birth of the first rabbit conceived through IVF in 1959, IVF has been practised in several animals. Various methods for *in vitro* maturation, IVF, and *in vitro* culture have been standardized in animals. In addition, IVMFC has provided an excellent source for embryo transfer, cloning, transgenesis, and other advanced *in vitro* techniques. It has also allowed the analysis of the developmental potential of embryos, pattern of gene expression, epigenetic modifications and cytogenetic disorders in various domestic species and has been used as a model for human embryogenesis studies. The low success rate and the costs make the technique less feasible for application in livestocks under field conditions

Intracytoplasmic Sperm Injection (ICSI) :ICSI is a micromanipulation technique used for treating male infertility. It involves mechanical insertion of a selected sperm into the cytoplasm of an oocyte to produce desirable embryo. Since the first report of ICSI success, ICSI has been done in other species such as rabbits, mice, sheep, humans, horses, cattle, and pigs including buffaloes. This technique is also used for sperm vector system for animal transgenic.

Multiple Ovulation andEmbryo Transfer: In this technique selected genetically superior (elite) females are induced to superovulate hormonally and inseminated with high quality semen of a superior male at an appropriate time relative to ovulation depending on the species. Week-old embryos are flushed out of the donor's uterus, isolated, examined microscopically for number and quality, and inserted into the lining of the uterus of surrogate mothers non-surgically. ET increases reproductive rate of selected females, reduces disease transfer, and facilitates the development of rare and economically important genetic stocks. The main limiting factor for the ET is that this technique involves costly hormones, labour intensive protocols and expertise in addition to the poor super ovulatory response and pregnancy outcomes.

Somatic cell Nuclear Transfer or Cloning: Somatic cell nuclear transfer (SCNT) is a major technique for delivering nuclease-mediated genetic alterations in livestock. In this technique, the nucleus of a somatic cell is transferred into a female egg cell or oocyte in which the nucleus has been removed to generate a new individual, genetically identical to the somatic cell donor. The advantage of SCNT is that the gene-edited cell line can be genotyped and/or screened before transfer into the enucleated oocyte to ensure that the desired edits, and no off-target edits, have occurred. A number of gene-edited animals have been produced through SCNT cloning technique. This technique was used to generate Dolly from a differentiated adult mammary epithelial cell. Further research is needed to improve the efficiency of the cloning. SCNT is a procedure of cloning within the same species whereas interspecies cloning (interspecies Somatic Cell Nuclear Transfer -iSCNT) are also feasible. The cloned animals have already been produced between closely related species. Eg- domestic cattle (*Bos taurus*) and wild ox (*Bos grunniens*). Cloning procedure using embryonic stem cells (ESCs) referred as Nuclear Transfer-derived Embryonic Stem Cell (NTESC) is still unsuccessful. Despite the achievements made through SCNT-editing method, certain drawbacks associated with cloning such as early embryonic losses, postnatal death, and birth defects cannot be ignored

Cryopreservation: **Cryopreservation is a process where cells, whole tissues, or any other substances susceptible to damage caused by chemical reactivity or time are preserved by cooling to sub-zero temperatures**. Cryopreservation is a multistage complex process incorporating cryoprotectants or antifreeze agents. The ability to cryopreserve germplasm indefinitely allows genetic diversity to be preserved. Unlike semen, cryopreservation of embryo helps in the preservation of complete genotypes. Freezing of embryos is an established commercial practice especially in cattle. In contrast to embryos, oocytes are extremely sensitive to chilling and are difficult to cryopreserve without losing their viability. However, research is in progress on the **vernalisation of oocytes**, where very low temperature storage, without freezing, could preserve the oocytes for several months. This technique is advantageous as it reduces the risk and expense in the transportation of expensive animals; reduce disease transmission and conservation of endangered species germplasm.

Embryo Sexing : Embryo sexing is a technique in reproductive biotechnology having practical applications. Sex determination is performed by Y-chromosome-specific DNA probe technology coupled with polymerase chain reaction (PCR) amplification of specific Y-chromosome region. Other methods involve detection of embryonic H-Y antigen in the embryos and use of loop-mediated isothermal amplification and duplex PCR-based assay showing more than 95% accuracy but involves high cost, time and expertise for carrying out these protocols.

Transgenesis: Transgenic animals have a foreign gene deliberately inserted into their genome by the micro-injection of DNA into the pronuclei of a fertilised egg which is subsequently implanted into the oviduct of a surrogate mother. Transgenesis has great potential in molecular breeding of farm animals, such as development of animals with high fecundity, higher fertility, disease resistance etc. Transgenic technologies in fishes can enhance growth rates and market size, feed conversion ratios, resistance to disease, sterility issues and tolerance of extreme environmental conditions. In the shrimp aquaculture sector, transgenic shrimp have been reported (Mialhe *et al.,* 1995), but there has been no successful development to date for commercial culture. The cost for making transgenic farm animals is high and the efficiency is low.

Stem Cells: Stem cells are unspecialized cells that renew themselves for long periods through cell division, and later become specialized on receiving specific signals. Based on their source, stem cells have been classified into three types, *viz*., embryonic, adult and fetal stem cells. ES cells are derived from embryos at (blastocyst stage 32 cell stage), can give rise to cells from all three embryonic germ layer.The ESs cells are advantageous as they do not form tumours when transferred into the body which potentiates their use in transplantation. On the other, adult stem cells are those undifferentiated cells found throughout the body which is needed for replenish and regenerate cells in any damaged tissue. The spermatogonial stem cells are the only adult stem cells having the responsibility of transferring genes to next generations *via* the process of fertilization of ovum. Some of the potential applications of this technology are surrogate production of spermatozoa, reduced time for progeny testing, production of transgenic animals and conservation of endangered species.

Gene editing: Gene editing is a powerful tool to manipulate genome, bearingapplications in animal breeding programs. Gene editing allows specific deletions, additions, or allele alteration at unambiguous locations in a genome. The development of designer nucleases (zinc finger nucleases [ZFNs], transcription activator-like effector nuclease [TALENs], and clustered regularly interspaced short palindromic repeats [CRISPR/Cas9]) has enabled extremely efficient and more facile genome editing in different animal species. These tools could be employed to enhance productivity, disease resistance, breeding efficiency, and for generation of novel animal models. Such alterations, if made in zygotes or germ line cells, can be permanent and heritable. Recently, genome editing in many livestock species has been reported such as myostatin (*MSTN*) gene editing for "double muscling" in pigs, cattle, and sheep, polled gene introduction in dairy cattle, and edits to confer resistance to porcine reproductive and respiratory syndrome virus and African swine fever virus in pigs.

Endocrine regulation of reproduction in fish

Biotechnology can be applied to enhance the reproductive performance of cultured aquatic species exhibiting reproductive dysfunction is captivity. In the past, fish gonadotropin, a group of hormones that stimulate reproduction, were produced in small amounts by extraction and purification from crude preparations of thousands of pituitary glands. At present, large quantities of highly purified gonadotropin can be produced in the laboratory through recombinant DNA technology. The use of synthetic Gonadotropin Releasing Hormone (GnRH), the key regulator of reproductive cascade in all vertebrates, triggers the secretion of the fish's own gonadotropin. GnRHa is synthesized chemically and does not carry the risk of transmitting diseases to the broodstock. However, injection of GnRHa does not always result in 100% ovulation and often multiple injections are often necessary to induce ovulation. Development of controlled-release delivery systems for synthetic GnRHas has contributed to captive breeding of many commercially important fish species. The hormones implants mixed with cholesterol, ethylene-vinyl in biodegradable microspheres have been efficient in inducing maturation and spawning in many cultured fish.

Sex control : The control of fish sex could be useful where one sex displays **advantageous characteristics**, such as **larger adult size, production of high-value caviar**(sturgeon), **faster growth rate**, or **higher age at first sexual maturation**. **Monosex populations** of the most advantageous sex can be produced either by **direct sex control *via* steroid treatment** (masculinisation by administration of androgens; feminisation by administration of estrogens); or by **genetic controland steroid treatment of broodstock** (indirect hormonal treatment, gynogenesis, androgenesis); or by **control of external factors** (temperature, density etc.). In the case of tilapia, males are preferred for culture as they grow faster than females. The YY male technology involves a genetic breeding programme combining the hormone feminization of a normal male (XY female) followed by mating with normal males (in tilapia).

Sterility: Sterility in fish by manipulation of reproduction would help to increase growth by reducing energy consumption for reproduction. Sterility can be achieved by ploidy manipulation to produce sterile triploids or the use of transgenics by gene "knock-out" or "gene knock-down".

Conclusion

Reproductive biotechnology has revolutionized animal breeding and genetic progress in livestock industry.The application of biotechnology in aquaculture including the use of synthetic hormones in induced breeding, production of monosex, surrogate broodstock, transgenic fish etc has played major role to ensure the continued expansion and intensification of aquaculture to meet the growing fish demand.The emerging techniques should be judicially implemented for manipulation and improvement of reproductive performance of the livestock species.

Source of information

K. K. Choudhary, K. M. Kavya,A. Jerome, R. K. Sharma (2016). Advances in reproductive biotechnologies. Vet World 9(4): 388–395.

Role of Biotechnology in Assisted Reproduction, Science, 14 May – 2014.

W. S. Lakra and S. Ayyappan (2002).Recent Advances in Biotechnology Applications to Aquaculture. International Symposium on Recent Advances in Animal Nutrition,22nd September, New Delhi, Pg-455-461

19.Use of molecular techniques in growth enhancement

Raymond J Angel

Introduction

It is proved that the use of molecular techniques in aquaculture has the potential to alleviate the predicted fish shortages and price increases by enhancing production efficiency, minimizing costs and reducing disease. Growth enhanced fish using molecular techniques will be equally beneficial to aquaculture and is more effective than traditional breeding techniques to develop new fish strains. In principle, the technology can be used to improve growth rate of the fish, control sexual maturation, sterility and sex differentiation, improve survival by increasing disease resistance against pathogen, adapt to extreme environment such as cold resistance and alter the biochemical characteristics of the flesh to enhance the nutritional qualities. Since fish can be readily improved by application of molecular techniques, it is clearly timely to consider what genetically modified (GM) fish are likely to offer in the future, both in terms of benefits and disadvantages (Maclean and Norman, 2003). Growth Hormone has also been utilised in recent years extensively for construction of transgenic fishes to enhance growth. Genetic engineering is an important tool to develop and improve traits of fish for aquaculture. Species showing high growth rate is widely used to isolate Growth Hormone gene for the production of transgenic fish.

An overview of various target species used in growth enhancement using molecular techniques

Transgenic fish have been produced for numerous species of fish including non-commercial model species such as the Loach, *Misgurnus anguillicaudatus* (Maclean *et al*. 1987a), Medaka, *Oryzias latipes* (Ozato *et al*. 1986), Topminnows and Zebra fish, although Gong *et al*. (2002) have developed transgenic Rainbow zebra fish for the ornamental fish industry. Several experiments have evaluated transgenic farmed fish species including Goldfish (Zhu *et al*. 1985), Common carp, Silver carp, Mud loach, Rainbow trout (Chourrout, 1986), Atlantic salmon, Coho salmon, Chinook salmon, Channel catfish (Dunham *et al*. 1987) and Nile tilapia (Brem *et al*. 1988). Additionally, gene transfer has been accomplished in a game fish, Northern pike (Gross *et al*. 1992).

Many species of fish have been used in studies for standardizing GH- involved transgenesis. Even though, many studies reported a positive enhancement of growth in target species, some proved to be unsuccessful due to many unknown reasons. Some of the studies have been quoted for reference (Table 1).

Techniques for growth enhancement

There are many ways to enhance growth including inbreeding, gynogenesis, androgenesis, selection, intraspecific crossbreeding, interspecific hybridization, polyploidy, sex reversal and breeding, nuclear transplantation and transgenesis. Cloned populations have been produced via gynogenesis and androgenesis (Dunham, 2004), but direct cloning of an individual fish of interest has not yet been accomplished. Gene transfer technology has produced a great impact in modern biology and biotechnology (Powers *et al*. 1998). A number of fish species are in focus for gene transfer experiments and can be divided into two main groups: animals used in aquaculture (Fletcher and Davies, 1991; Hew *et al*. 1995; Chen and Lu, 1998) and model fish used in basic research (Chen and

Lu, 1998). Among the major food fish species are Carp (*Cyprinus sp*.), Tilapia (*Oreochromis sp.*), Salmon (*Salmo sp.*, *Oncorhynchus sp.*) and Channel catfish (*Ictalurus punctatus*) while Zebrafish (*Danio rerio*), Medaka (*Oryzias latipes*) and Goldfish (*Carassius auratus*) are used in basic research. Genetic engineering of farm animals offers great potential for improvement of selected genetic traits of agricultural significance. Several species of fish have also been used to exploit this technology for commercial purposes, and examples include attempted induction of freeze resistance in transgenic salmon using an Anti-Freeze Protein gene (Fletcher *et al*., 1988) and production of growth enhanced fish using novel Growth Hormone (GH) genes (Dunham *et al*., 1987; Brem *et al*., 1988; Penman *et al*., 1990) or an Insulin-like Growth Factor (IGF) gene (Chen *et al*., 1995). Although several species of fish have been used to produce lines of transgenic fish, in only a few cases has germline transmission and stable long term transgene expression been satisfactorily demonstrated.

Techniques for gene transfer

Microinjection

Microinjection is most successfully and widely used technique for gene transfer in fish. Gene transfer research with fish began in the mid 1980's utilizing microinjection (Zhu, *et al* 1985, Dunham *et al* 1987). Zhu *et al*. (1985) published the first report of transgenes microinjected into the fertilized eggs of goldfish. In almost all fish gene transfer research, the foreign gene was microinjected into the cytoplasm of one-to- four cell embryos (Hayat, 1989) as pronuclei are extremely difficult to visualize in live one-cell fish embryos.

To ensure the integration of the DNA it should be injected to intact cells close to the cut site. The injection apparatus consists of a dissecting stereomicroscope and two micro-manipulators, one with a glass micro-needle for delivering transgene and other with a micropipette for holding fish embryo in place (Fig. 1). The success of microinjection technique depends on the nature of egg chorion. The soft chorion facilitates the microinjection while the thick chorion limits the ability to visualize the target for injection of DNA. In many fishes (Atlantic salmon and rainbow trout) the egg chorion gets tough and hard just after the fertilization or to contact with the water and provides a difficulty in injecting the DNA.

Steps of Microinjection Technique

(1) Desired eggs and sperms are stored separately at the optimum conditions.

(2) Add water and sperms and initiate the fertilization.

(3) Ten minutes after the fertilization, eggs are dechorionated by trypsinization.

(4) Fertilized eggs are microinjected with desired DNA just within a few hours of fertilization. DNA is released into the centre of the germinal disc to the first cleavage in dechorionated eggs. The time available for microinjection is first 25 minutes and that too between fertilization and first cleavage.

(5) After microinjection the embryos are incubated in water until hatching takes place.

Survival rates of microinjected fish embryos is seem to be about 30-80% depending the fish species.

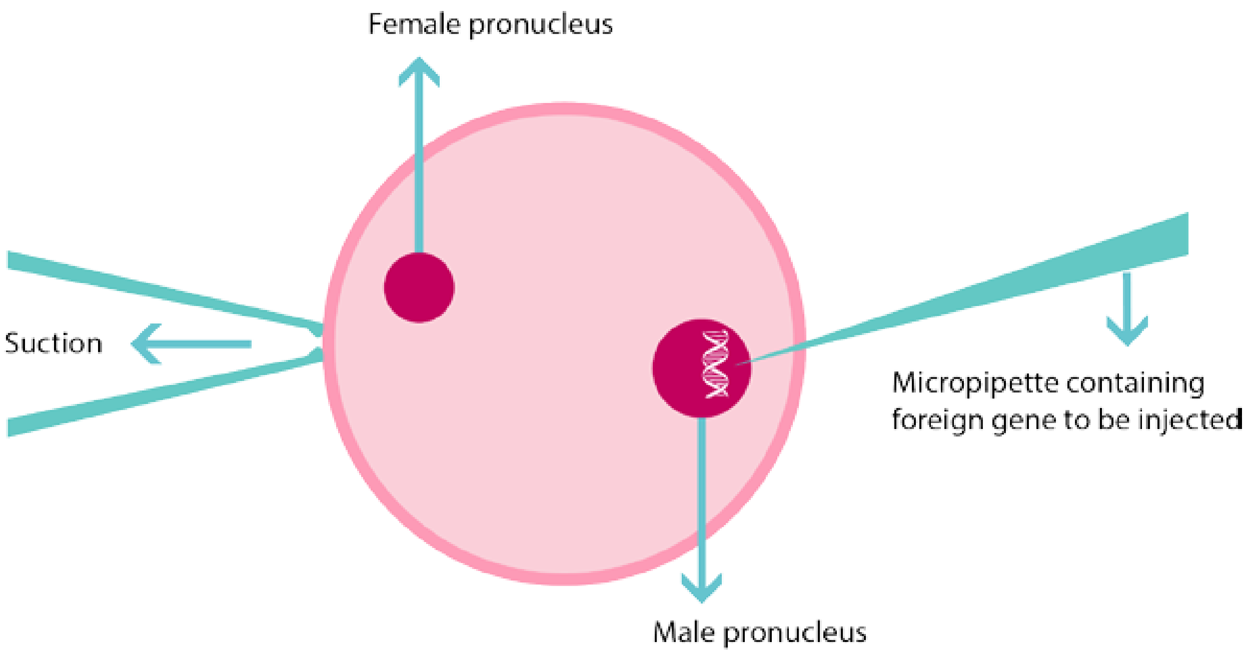


Fig 1.Microinjection technique.

Other methods

Microinjection is a tedious and slow procedure (Powers, *et al*. 1992) and can result in high egg mortality (Dunham, *et al*. 1987). After the initial development of microinjection, new techniques such as electroporation, retroviral integration, liposomal-reverse-phase-evaporation, sperm mediated transfer and high velocity micro-projectile bombardment were developed (Chen and Powers, 1990) that sometimes can more efficiently produce large quantities of transgenic individuals in a shorter time period. The first successful gene transfer utilizing electroporation produced integration rates and survival similar to that for microinjection (Inoue, *et al*.1990). Powers, *et al*. (1992) demonstrated that electroporation can be more efficient than microinjection with integration rates sometimes as high as 30-100%. Walker (1993) found that hatching rates were higher for electroporated embryos than for microinjected channel catfish embryos, and post-fertilization electroporation treatments had higher hatching rates than electroporation of sperm and then eggs prior to fertilization.

Environmental Concerns about Transgenic Fish and risk mitigation

The primary environmental concerns about releases of transgenic fish, for example, include competition with wild populations, movement of the transgene into the wild gene pool, and ecological disruptions due to changes in prey and other niche requirements in the transgenic variety versus the wild populations.

It is important to note that developers of transgenic fish are attempting to reduce or eliminate both gene flow and invasive species risks by sterilizing transgenic fish. Sterilization is relatively easy and inexpensive but success rates are highly variable. In addition, sterilization does not necessarily neutralize environmental risks. Academic scientists note that an escaped, sterile fish might still engage in courtship and spawning behaviour, disrupting breeding in wild populations. Waves of escaped sterile fish could also create ecological disruptions as each group is replaced by another equally strong group of transgenic sterile fish.

Conclusion and future prospective

Transgenic fish technology has great potential in the aquaculture industry. By introducing desirable genetic traits into fishes, mollusks, and crustaceans, superior transgenic strains can be produced for aquaculture. These traits include faster growth rates, improved food conversion efficiency, resistance to some known diseases, tolerance to low oxygen concentrations, and tolerance to extreme temperatures. Our laboratory and those of others have shown that transfer, expression and inheritance of fish growth hormone transgenes can be achieved in several fish species and that the resulting transgenics grow substantially faster than their non-transgenic siblings. This is a vivid example of the potential application of the gene transfer technology to aquaculture.

However, to realize the full potential of the transgenic fish technology in aquaculture orother biotechnological applications, several important scientific breakthroughs are required. These include:

(1) more efficient technologies for mass gene transfer,

(2) targeted gene transfer technologies such as embryonic stem cell gene transfer or ribozyme gene inactivation,

(3) suitablepromoters to direct the expression of transgenes at optimal levels during the desired developmental stages,

(4) identified genes of desirable traits for aquaculture and other applications,

(5) informationon the physiological, nutritional, immunological and environmental factors that maximize the performance of the transgenics, and

(6) safety and environmental impacts of transgenic fish. Once these problems are resolved, the commercial application of the transgenic fish technology will be readily attained.

Table 1. Studies showing enhancement of growth achieved in different target organisms worldwide with citation

FAMILY AND SPECIES	CONSTRUCT	GROWTH	COUNTRY	SUPPORTING CITATION
		Salmonidae		
Atlantic salmon, *Salmo salar*	opAFP-csGH	2–6-Fold	Canada	Du *et al*. (1992)and Fletcher *et al*. (2004)
Coho salmon, *Oncorhynchus kisutch*	ssMT-ssGH	Up to 11-fold	Canada	Devlin *et al*. (1994a,b)
Coho salmon *Oncorhynchus kisutch*	opAFP-csGH	3–10-Fold	Canada	Devlin *et al*. (1995a)
Chinook salmon, *O. tshawhytscha*	opAFP-csGH	6-Fold	Canada	Devlin *et al*. (1995a)
Rainbow trout, *O. mykiss*	opAFP-csGH	3.2-Fold	Canada	Devlin *et al*. (1995a)
Cutthroat trout, *O. clarki*	opAFP-csGH	6-Fold	Canada	Devlin *et al*. (1995a)

Arctic charr, *Salvelinus alpines*	Various constructs	Up to 14-fold	Finland	Pitkanen *et al*. (1999)
Rainbow trout *O. mykiss*	ssGH-ssGH	None	Finland	Pitkanen *et al*. (1999)
Cichlidae				
Nile tilapia, *Oreochromis niloticus*	opAFP-csGH	2–4-Fold	UK	Rahman *et al*. (1998; 2001) and Rahman and Maclean (1999)
Nile tilapia *Oreochromisniloticus*	ssMT-ssGH	None	UK	Rahman *et al*. (1998; 2001) and Rahman and Maclean (1999)
Tilapia, *O. hornorum* Hybrid	hCMV-tiGH	82%	Cuba	Martinez *et al*. (1996)
Ictaluridae				
Channel catfish, *Ictalurus punctatus*	RSVLTR-rtGH,	Up to 26%	USA	Dunham *et al*. (1992)
Channel catfish *Ictalurus punctatus*	mMT-hGH	None	USA	Dunham *et al*. (1987)
Heteropneustidae				
Heteropneustes fossilis	Zpb-ypGH	30–60%	India	Sheela *et al*. (1999)
Cyprinidae				
Goldfish, *Carassiusauratus*	mMT-hGH	None	PR China	Zhu *et al*. (1985)
Common carp, *Cyprinus carpio*	mMT-hGH	None	PR China	Zhu *et al*. (1989)
Common carp *Cyprinus carpio*	cbA-gcGH	42–80%	PR China	Zhu (1992) and Wang *et al*. (2001)
Catla, *Catla catla*	RSVLTR-rtGH	None	India	Sarangi *et al*. (1999)
Common carp *Cyprinus carpio*	ccbA-ccGH	4-Fold	Israel	Hinits and Moav (1999)
Rohu *Labeo rohita*	CMV-roGH	4-Fold	India	Venugopal *et al*. (2004)
Rohu *Labeo rohita*	gcbA-roGH	4.5–5.8-Fold	India	Venugopal *et al*. (2004)

Esocidae				
Northern pike	RSVLTR-bGH	30%	USA	Gross *et al*. (1992)
Cobitidae				
Mud loach, *Misgurnusmisolepis*	mlb-actin-mlGH	Up to 35-fold	Republic of Korea	Nam *et al*. (2001; 2002)

REFERENCES

Brem, G., Brenig, B., Horstgen-Schwark, G. and Winnacker, E. L., 1988. Gene transfer in tilapia (*Oreochromis niloticus*). *Aquaculture.,* 68: 209-219.

Chen, T. T. and Lu, J. K., 1998. Transgenic fish technology: Basic principles and its application in basic and applied research. *In: De la Fuente J. and Castro F.O. eds. Gene transfer in aquatic organisms. RG Landes Company and Germany: Springer-Verlag, Austin, Texas, USA*., pp. 45-73.

Chen, T. T. and Powers, D. A. (1990) Transgenic fish. *Trends in Biotechnology., 8*: 209-214.

Chen, T. T., Lu, J. K., Shamblott, M.J., Cheng, C. M., Lin, C. M., Burns, J. C., Reimschuessel, R., Chatakondi, N. and Dunham, R. A., 1995.Transgenic fish: ideal models for basic research and biotechnological applications. *Zool. Studies.,*344: pp. 215–234.

Chourrout, D., 1986. Techniques of chromosome manipulation in rainbow trout: a new evaluation with karyology. *Theoretical and Applied Genetics., 72*: 627-632.

Devlin, R. H., Byatt, J. C., McLean, E., Yesaki, T. Y., Krivi, G.G., Jaworski, E.G. and Donaldson, E.M., 1994b. Bovine placental lactogen is a potent stimulator of growth and displays strong binding to hepatic receptor sites of coho salmon. *Gen. Comp. Endocrinol*., 95: 31–41.

Devlin, R. H., Yesaki, T. Y., Biagi, C. A., Donaldson, E. M., Swanson, E. M. P. and Chan, W. K., 1994a. Extraordinary salmon growth.*Nature, 371*, 209–210.

Devlin, R. H., Yesaki, T. Y., Donaldson, E. M., Du, S. J. and Hew, C. L., 1995a. Production of germline transgenic Pacific salmonids with dramatically increased growth performance. *Can. J. Fish.Aquat. Sci*., 52: 1376–1384.

Du, S. J., Gong, Z. Y., Fletcher, G. L., Shears, M. A., King, M. J., Idler, D. R. and Hew, C. L., 1992. Growth enhancement in transgenic Atlantic salmon by the use of an all-fish chimeric growth hormone gene construct. BioTechnology., 10: 176–181.

Dunham, R. A., Ramboux, A. C., Duncan, P. L., Hayat, M., Chen, T. T., Lin, C. M., Kight, K., Gonzalez-Villasenor, I. and Powers, D. A., 1992.Transfer, expression and inheritance of salmonid growth hormone genes in channel catfish, Ictalurus punctatus, and effects on performance traits.*Mar. Mol. Biol. Biotechnol*., 1: 380–389.

Dunham, R. A. 2004.,*Aquaculture and Fisheries Biotechnology Genetic Approaches.CABI publishing, Wallingford ,UK*., 17: P. 400.

Dunham, R. A., Eash, J., Askins, J. and Townes, T.M., 1987. Transfer of the metallothione in human growth hormone fusion gene into channel catfish. *Transactions of the AmericanFisheries Society.,*116: 87-91.

Fletcher, G. L., Shears, M. A., King, M. J., Davies, P. L. and Hew, C. L., 1988. Evidence for antifreeze protein gene transfer in Atlantic salmon (*Salmo salar*).*Can. J. Fish.Aquat. Sci.,*45, pp. 352–357

Fletcher, G. L., Shears, M. A., Yaskowiak, E. S., King, M. J. and Goddard, S. V., 2004. Gene transfer: potential to enhance the genome of Atlantic salmon for aquaculture. *Aust. J. Exp. Agric*., 44: 1095–1100.

Fletcher. G. L. and Davies, P. L., I991. Transgenic fish for aquaculture.*Gen. Eng*., 13: 33l-369.

Gong, Z., Wan, H., Ju, B., He, J., Wang, X.,and Yan, T., 2002. Generation of living color transgenic zebrafish. In: Shimizu, N., Aoki, T., Hirono, I., and Takashima, F. (Eds.). *Aquatic Genomics: Steps Toward a Great Future*, *Springer-Verlag, New York, NY*., pp. 329-339.

Gross, M. L., Schneider, J. F., Moav, N., Moav, B., Alvarez, C., Myster, S. H., Liu, Z., Hallerman, E. M., Hackett, P. B., Guise, K. S., Faras, A. J. and Kapuscinski, A. R., 1992. Molecular analysis and growth evaluation of northern pike (*Esox lucius*) microinjected with growth hormone genes. *Aquaculture.,*103: 253-273.

Hayat, M., 1989. Transfer, expression and inheritance of growth hormone genes in channel catfish (*Ictalurus punctatus*) and common carp (*Cyprinus carpio*). *Doctoral Dissertation. Auburn University, AL, USA.*

Hew, C. L.; Fletcher, G. L. and Davies, P. L., 1995. Transgenic salmon: tailoring the genome for food production. *Journal of Fish Biology,* 47: 1-19.

Hinits, Y. and Moav, B., 1999. Growth performance studies in transgenic *Cyprinus carpio. Aquaculture*., 173: 285–296.

Inoue, K., Yamashita, S., Hata, J. I., Kabeno, S., Asada, S., Nagahisa, E. and Fujita, T., 1990. Electrophoration as a new technique for producing transgenic fish.*Cell Differentiationand Development.,*29: 123-128.

Maclean, N. and Norman.,2003.Genetically modified fish and their effects on food quality and human health and nutrition.*Trends in Food Science & Technology.*, 14: (5-8), 242-252.

Maclean, N., Penman, D. and Talwar, S., 1987a.Introduction of novel genes into fish.*Biotechnology.,* 5: 257-261.

Martinez, R., Estrada, M. P., Berlanga, J., Guillen, I., Hernandez, O., Cabrera, E., Pimentel, R.,Morales, R., Herrera, F., Morales, A., Pina, J. C., Abad, Z., Sanchez, V., Melamed, P., Lleonart, R. and de la Fuente, J., 1996. Growth enhancement in transgenic tilapia by ectopic expression of tilapia growth hormone.*Mol. Mar. Biol. Biotechnol*., 5: 62–70.

Nam, Y. K., Cho, Y. S., Cho, H. and Kim, D. S., 2002. Accelerated growth performance and stable germ-line transmission in androgenetically derived homozygous transgenic mud loach, *Misgurnus mizolepis*. *Aquaculture*., 209: 257–270.

Nam, Y. K., Noh, J. K., Cho, Y. S., Cho, H. J., Cho, K. N., Kim, C. G. and Kim, D. S., 2001. Dramatically accelerated growth and extraordinary gigantism of transgenic mud loach *Misgurnus mizolepis*. *Transgenic Res*., 10: 353–362.

Ozato, K., Kondoh, H., Inohara, H., Iwamatsu, T., Wakamatsu, Y. and Okada, T. S.,1986. Production of transgenic fish: introduction and expression of chicken delta-crystallin gene in medaka embryos. *Cell Differ. Dev.,* 19: 237-244.

Penman, D. J., Beeching. A .J., Penn, S. and Maclean, N., 1990. Factors affecting survival and integration following microinjection of novel DNA into rainbow trout eggs.*Aquaculture*, 85: 35-50.

Pitkanen, T. I., Krasnov, A., Teerijoki, H. and Molsa, H., 1999. Transfer of growth hormone (GH) genes into Arctic charr (*Salvelinus alpinus* L.). I. Growth response to various GH constructs. *Genet. Anal.: Biomol. Eng*., 15: 91–98.

Powers, D. A., Cole, T., Creech, K., Chen,T. T., Lin, C. M., Kight, K. and Dunham, R., 1992. Electroporation: a method for transferring genes into the gametes of zebrafish, *Brachydanio rerio*, channel catfish, *Ictalurus punctatus*, and common carp, *Cyprinuscarpio*. *Mol. Mar. Biol. Biotech.,* 1:301-309.

Powers, D. A.; Gómez-Chiarri, M.; Chen, T. T. and Dunham, R.,1998. Genetic Enginering of Finfish and shellfish.*In: De la Fuente J. and Castro F.O. eds. Gene transfer in aquatic organisms. RG Landes Company and Germany, Springer-Verlag, Austin, Texas, USA*. pp. 17-34.

Rahman, M. A. and Maclean, N., 1999. Growth performance of transgenic tilapia containing an exogenous piscine growth hormone gene. *Aqaculture*, 173: 333–346.

Rahman, M. A., Mak, R., Ayad, H., Smith, A. and Maclean, N., 1998. Expression of a novel piscine growth hormone gene results in growth enhancement in transgenic tilapia (*Oreochromis niloticus*). *Transgenic Res*., 7: 357– 369.

Rahman, M. A., Ronyai, A., Engidaw, B. Z., Jauncey, K., Hwang, G. L., Smith, A., Roderick, E., Penman, D., Varadi, L. and Maclean, N., 2001. Growth and nutritional trials on transgenic Nile tilapia containing an exogenous fish growth hormone gene.*J. Fish Biol*., 59: 62–78

Sarangi, N., Mandall, A. B., Bandyopadhyay, A. K., Venugopal, T., Mathavan, S. and Pandian, T. J., 1999. Electroporated sperm-mediated gene transfer in Indian major carps. *Asia-Pacific J. Mol. Biol. Biotechnol*., 7: 151–158.

Sheela, S. G., Pandian, T. J. and Mathavan, S., 1999. Electroporatic transfer, stable integration, and transmission of pZp beta ypGH and pZp beta rtGH in Indian catfish, *Heteropneustes fossilis* (Bloch).*Aquac. Res*., 30: 233–248.

Venugopal, T., Anathy, V., Kirankumar, S. and Pandian, T.J., 2004.Growth enhancement and food conversion efficiency of transgenic fish, Labeo rohita.*J. Exp. Biol*. 301A: 477–490.

Walker, D.S., 1993. Effect of electroporation and microinjection on survival of ictalurid catfish embryos. *Master of Science Thesis. Auburn University, AL.*

Wang, Y., Hu, W., Wu, G., Sun, Y., Chen, S., Zhang, F., Zhu, Z., Feng, J. and Zhang, X., 2001. Genetic analysis of ‘‘all-fish’’ growth hormone gene transferred carp (*Cyprinus carpio* L.) and its F1 generation. *Chin. Sci. Bull*., 46: 1174–1177.

Zhu, Z., 1992. Generation of fast-growing transgenic fish: methods and mechanisms. *In: Hew, C.L., Fletcher, G.L. (Eds.), Transgenic Fish. World Publishing, Singapore*, pp. 92–119.

Zhu, Z., Li, G., He, L. and Chen, S., 1985.Novel gene transfer into the fertilized eggs of goldfish (*Carassius auratus*, 1758).*Journal of Applied Ichthyology,* 1*:* 31-33.

Zhu, Z., Xu, K., Xie, Y., Li, G. and He, L., 1989.A model of transgenic fish.*Sci. Sin*., B 2: 147–155.

20. Gene Editing Tools and their application in Aquaculture

Misha Soman

Introduction

Genome editing is a kind of genetic engineering in which a gene of interest is inserted, or erased in the genome of an organism or cells using engineered restriction enzymes called "molecular scissors." These nucleases create site-specific double-strand breaks (DSBs) at desired locations in the genome. The induced double-strand breaks are repaired through non-homologous end-joining (NHEJ) or homologous recombination (HR), resulting in targeted mutations ('edits'). By editing the genome the characteristics of a cell or an organism can be changed.

Genome editing uses 'engineered nuclease' which cuts the DNA at its targeted site. Engineered nucleases have two parts a nuclease part and the DNA-targeting part that is designed in such a way that it guides the nuclease to cut a specific sequence of DNA. When a cut forms within a particular place of DNA, the cell starts to repair the cut naturally.Gene editing technologies have wide applications in different fish species for basic as well as applied research in disease modeling and aquaculture.

Genome editing can be used

For research: It can be used to alter the DNA in organisms to study impact of gene modification.

To treat disease: Genome editing is being used in medical research to study the viability of the technology to treat deadly human diseases like leukemia, AIDS, cancer, etc. (Youdiil Ophinni, *et al*., 2018;Pablo Tebas, *et al*., 2014).

For biotechnology: Genome editing has been used in agriculture to produce genetically modified crops to improve their yields and resistance to disease, as well as to make genetically modified pigs(Kankan Wang, *et al*., 2015; Jin-Dan Kang *et al*., 2017), sheep (Crispo, *et al*., 2015), and fishes(Karim Khalil *et al*., 2017).

TYPES OF GENE EDITING

- ➢ Zinc finger nucleases (ZFNs)
- ➢ TALENS
- ➢ CRISPR-Cas9

ZINC FINGER NUCLEASES (ZFNS)

Zinc finger nucleases (ZFNs) are the type of engineered restriction nucleases produced by joining zinc finger DNA-binding domain and DNA-cleavage domain (FokI) that promote targeted editing of the genome by generating double-strand breaks in DNA at targeted locations. This nuclease is a site-specific endonuclease designed to bind and cleave DNA at particular locations. ZFN is composed of three to six zinc finger motifs, and each motif particularly recognizes three nucleotides in a DNA sequence. Hence, each ZFN can identify target 9 to 18 base pairs. The cleavage of target DNA requires dimerization of two ZFNs for the FokI enzyme results in double-strand break (DSB) at the target locus (Durai et al., 2005). Double-strand breaks are important for site-specific mutagenesis in that they stimulate the cell's natural DNA-repair processes homology-directed repair and Non-Homologous End Joining (NHEJ); these reagents can be used to modify the genome precisely.

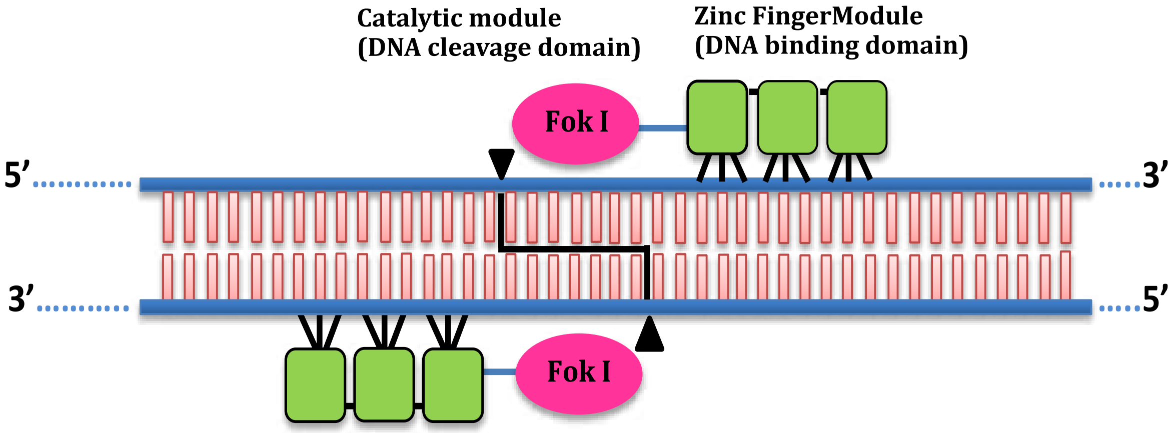


Fig 1: DNA-binding domain and DNA-cleaving domains are fused together, a highly-specific pair of 'genomic scissors' formed.

TALENS

Transcription activator-like effector nuclease (TALEN) technology use engineered restriction enzymes generated by fusing a TAL effector DNA-binding domain to a DNA cleavage domain (FokI). Restriction enzymes can be designed that will precisely cut any desired DNA sequence. When these restriction enzymes are introduced into cells, it makes double-stranded breaks in the gene of interest. The nucleases consist of programmable and sequence-specific DNA-binding modules coupled with a regular DNA cleaved domain that allows accurate and efficient genetic alterations by stimulating the targeted DNA double-strand breaks to induce cellular DNA repair, including error-prone NHEJ and HDR.

The DNA binding domain contains a repeated highly conserved 33–34 amino acid sequence with divergent 12th and 13th amino acids. These two positions, referred to as the Repeat Variable Diresidue (RVD), are highly variable and show a strong correlation with specific nucleotide recognition. Different RVD allows each module to specifically recognize one individual nucleotide instead of three nucleotides as in ZFN (Moscou and Bogdanove, 2009). The dimerized FokI randomly cleaves the DNA sequence between the left and right TALEN target sites.

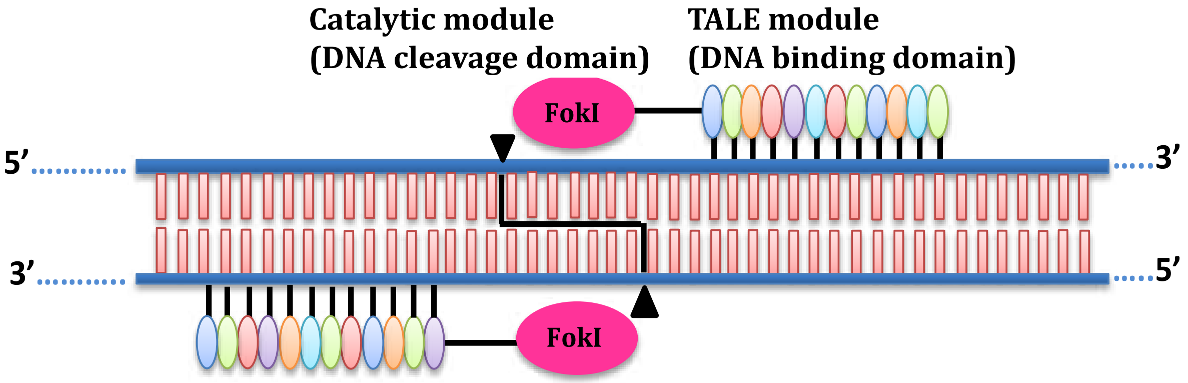


Fig 2: TALENS mechanism

CRISP-R Cas9

The clustered regularly interspaced short palindromic repeat (CRISPR) and associated protein (Cas9) emerged as a faster, cheaper and more precise gene editing tool in a wide range of organisms. It is an adaptive immunity mechanism in prokaryotes to eliminate invading genetic material in which the foreign genetic material cut into fragments and integrated into its CRISPR locus as a series of short repeats (20 bps). The loci are transcribed and processed into small RNAs which are called as guide RNAs to guide nucleases to cleave the target DNA based on sequence complementarity. This unique technology enables geneticists and medical researchers to edit the genome by adding, removing or altering the DNA sequence. The CRISPR-Cas9 system consists of two key players that make mutation into the targeted DNA. These are the enzyme Cas9 and a piece of RNA called guide RNA. The cas9 act as a molecular scissor which cuts double-stranded DNA at a specific targeted site. So that bits of the sequence can be added or removed. The guide RNA (gRNA) is an about 20 bases long pre-designed RNA sequence located within the RNA scaffold. The scaffold part binds to DNA and the pre-designed guide RNA 'guides' Cas9 nuclease to the targeted region of the genome, and it ensures that the Cas9 enzyme cuts at the right point in the genome.

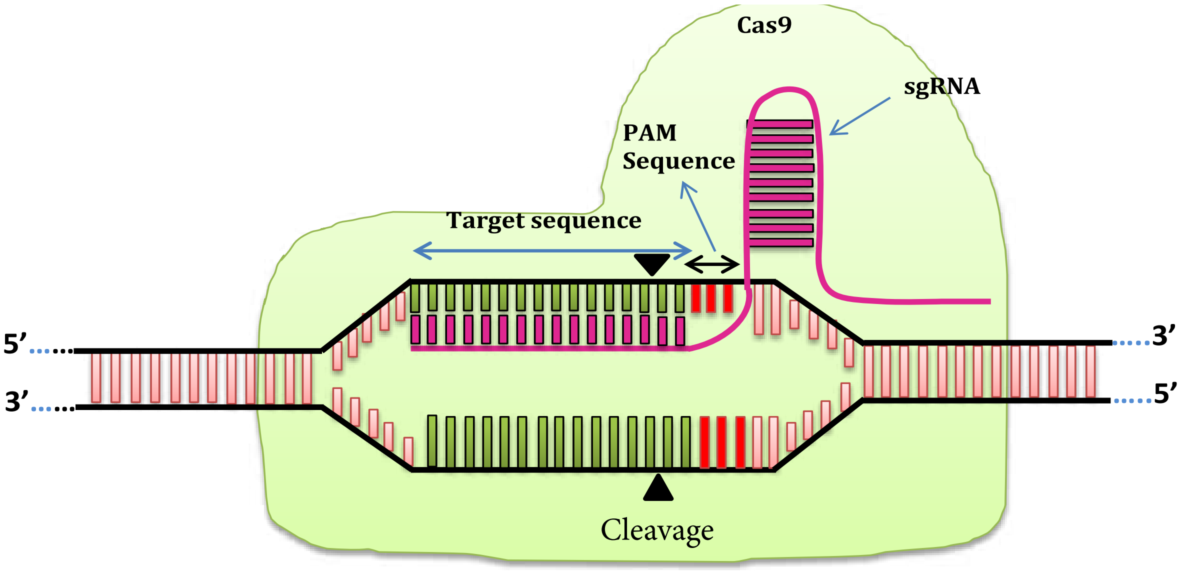


Fig 3: CRISP-R Cas9 mechanism

DNA double stranded breaks (DSB) repair mechanisms

Most DSBs get repaired by either the non-homologous end joining (NHEJ) pathway or the homology-directed repair (HDR) pathway. The NHEJ repair pathway causes nucleotide insertions or deletions (indels) at the cleavage site. In most cases, NHEJ gives rise to small indels in the targeted DNA that result in deletions, insertions, or frameshift mutations leading to the formation of premature stop codons inside the open reading frame (ORF) of the targeted gene and causes gene disruption. It results in the loss of function of the targeted gene.

Homology-directed repair (HDR) is a process of homologous recombination where a DNA template is used for precise repair of a double-strand break (DSB). This template can be either from the cell during the late S phase and the G2 phase of the cell cycle, before the completion of mitosis, or it can be an exogenous repair templates delivered into a cell mostly in the form of a synthetic, single-strand DNA donor oligo or DNA donor plasmid, to generate a specific change in the genome.

Advantages of CRISPR-Cas 9 system over ZFNs AND TALENS

- Highly efficient mutagenesis
- Effective introduction of targeted indels at required genomic location
- Target efficiency >80%
- In CRISPR-Cas9 system only one customized sg RNA is required to target a specific sequence, the same Cas9 can be used for all targeted sequences.
- ZFNs and TALENS require design and assembly of two nucleases for each target site.
- Sg RNAs are of short sequences <100bp, therefore reduces complications

Applications of gene editing tools in fishes

Fish species, especially the model species such as the zebrafish, have played important roles in testing new protocols of genome editing because of the biological advantages of fish models. A large number of genes have been disrupted or modified in fish species for functional studies, especially those involved in reproduction. These gene editing technologies can be utilized to modify the genomes of a variety of industrially relevant organisms and standard research animals including zebrafish, rats, pigs, catfish. The cis-regulatory mechanisms and gene knockdowns or knockouts can be investigated by using genome editing tools to know the unexplored processes of animal development and gene function to use in basic and applied sciences. Genome editing can be utilized to study early embryogenesis, induction of mutation, production of knockout lines, to unravel ancestral features of chordate development. It can be used to systematically study the functional analysis of reproductive performance in fishes, disease resistance, tolerance to environmental stressors, sex determination, sex differentiation, functional analysis of genes in non-reproductive functions like pigmentation, growth, and development and also for the disease modeling and drug screening. CRISPR is one of the most useful and powerful tools for gene manipulation in fish; even though off-target occurrence is a serious concern. The authors report that off-target mutation efficiency can be reduced by lowering the concentration of gRNAs in the injection. Genome editing tools were applied in zebrafish, mainly to induce mutations which would give valuable insights for medical science. The myostatin (MSTN) gene (muscle suppressor gene) disruption by CRISPR/Cas9 was successfully carried out in channel catfish, *Ictalurus punctatus* which resulted in 88–100% rates of mutagenesis in the protein-coding sites of Myostatin. The MSTN altered fry had more muscle cells, and the mean body weight also increased by 29.7%. The alignment of the mutated sequences vs. wild-type showed multiple insertions and deletions. (Karim Khalil *et al*., 2017). In India, Central Institute of Freshwater Aquaculture successfully disrupted Toll-like receptor 22 (TLR22) gene of *Labeo rohita* (rohu) involved in innate immunity and solely present in teleost fishes and amphibians using the CRISPR/Cas9 technology and the mutants lacked TLR22 mRNA expression (Chakrapani *et al*., 2016). These results confirm that CRISPR/Cas9 is a highly efficient tool for editing the fish genome, and exposes ways for promoting fish genetic enhancement and functional genomics.

Conclusion

Gene editing tools are widely used for studying the manipulation of the gene in human, animals, vegetables, and fish for various purposes. With this high-efficiency gene editing in fishes, we are entering into a new era for the adoption of powerful technologies to study various gene functions to improve the traits. Gene editing tools widely used to study the impact of the manipulation of the gene in animals, vegetables, fish and in humans for various purposes. With these high-efficiency genes editing in fishes, we are entering into a new era of powerful technologies to study multiple gene functions to improve the traits. These technologies will give insights into the gene functions and the evolution of vertebrates and also the possibility to treat deadly human diseases in medical research, to create improved varieties in agriculture, livestock and aquaculture. In the aquaculture industry, this approach may pave the way for growth-enhanced fishes to increase the productivity.

References

www.Yourgenome.org

www.Genetherapynet.com

Khalil, K., Elayat, M., Khalifa, E., Daghash, S., Elaswad, A., Miller, M., Abdelrahman, H., Ye, Z., Odin, R., Drescher, D., Vo, K., Gosh, K., Bugg, W., Robinson D, and Dunham R., (2017). Generation of *Myostatin* Gene-Edited Channel Catfish (*Ictalurus punctatus*) via Zygote Injection of CRISPR/Cas9 System. *Scientific Reports* volume 7, Article number: 7301.

Zhu, B., Ge, W., 2018. Genome editing in fishes and their applications. *General and Comparative Endocrinology.* 257, 3-12.

Chakrapani, V., Patra, S. K., Panda, R. P., Rasal, K. D., Jayasankar, P., Barman, H. K., 2016. Establishing targeted carp TLR22 gene disruption via homologous recombination using CRISPR/Cas9. J. of Developmental and Comparative Immunology (61) 242-247.

Wang, K., Ouyang, H., Xie, Z., Yao, C., Guo, N., Li, M., Jiao, H, and Pang, D., 2015. Efficient Generation of Myostatin Mutations in Pigs Using the CRISPR/Cas9 System. *Scientific Reports* 5:16623.

Crispo, M., Mulet, A. P., Tesson, L., Barrera, N., Cuadro, F., dos Santos-Neto, P. C., Nguyen, T. H., Crénéguy, A., Brusselle, L., Anegón, I., Menchaca. A., 2015. Efficient Generation of Myostatin Knock-Out Sheep Using CRISPR/Cas9 Technology and Microinjection into Zygotes. PLoS ONE 10(8): e0136690.

Ophinni, Y., Inoue, M., Kotaki. T & Kameoka, M., 2018. CRISPR/Cas9 system targeting regulatory genes of HIV-1 inhibits viral replication in infected T-cell cultures. Scientific Reports volume 8, Article number: 7784.

Pablo Tebas, P., David Stein, D., Winson W. Tang, W. W., Ian Frank, I., Shelley Q. Wang, M.D., Gary Lee, Ph.D., S. Kaye Spratt, Ph.D., Richard T. Surosky, Ph.D., Martin A. Giedlin, Ph.D., Geoff Nichol, M.D., Michael C. Holmes, Ph.D., Philip D. Gregory, Ph.D., et al. 2014. Gene Editing of *CCR5* in Autologous CD4 T Cells of Persons Infected with HIV. The New England journal of medicine. 370:901-910.

GLOSSARY

- Read – Base pair information of a given length from a DNA or cDNA fragment contained in a sequencing library. Different sequencing platforms are capable of generating different read lengths.
- Single End Read – The sequence of the DNA is obtained from the 5’ end of only one strand of the insert. These reads are typically expressed as 1x “y”, where “y” is the length of the read in base pairs (ex. 1x50bp, 1x75bp).
- Paired End Read – The sequence of the DNA is obtained from the 5’ ends of both strand of the insert. These reads are typically expressed as 2x “y”, where “y” is the length of the read in base pairs (ex. 2x100bp, 2x150bp).
- Mate Pair Read – The sequence of the DNA is obtained similar to paired-end reads, however the size of the DNA insert is often much greater in size (2-10kb in length) and the paired reads originate from a single strand of the DNA insert.
- Depth of Coverage – The number of reads that spans a given DNA sequence of interest. This is commonly expressed in terms of “Yx” where “Y” is the number of reads and “x” is the unit reflecting the depth of coverage metric (i.e. 5x, 10x, 20x, 100x)
- Sequencing Depth – The amount of sequencing a given sample requires to achieve a certain depth of coverage. This is frequently expressed as the number of reads a sample requires (ex. 40 million reads, 80 million reads) or the number of bases of sequencing a sample requires (ex. 4 gigabases, 100 megabases).
- SNP/SNV – Referring to a Single Nucleotide Polymorphism or Single Nucleotide Variant detected in a sample.
- InDels – One or more Insertion or Deletion event that is detected in a sample.
- Annotation - Adding biological information to genome sequence. This is a very complex task, and the process for doing this is rapidly evolving. Features that are added to the genome often include gene models, SNPs, and STSs.
- Copy Number Variation (CNV)- large-scale structural changes in DNA that vary from individual to individual. These include insertions, deletions, duplications and complex multi-site variants that range from kilobases to megabases in size. CNV can influence gene expression, phenotypic variation and alter gene dosage, and in certain instances may be associated with developmental disorders, cause disease or confer susceptibility to complex disease traits.
- EST Expressed sequence tag - These are single-pass sequences of cDNA clones. Databases of EST sequences are highly redundant but quite useful for gene identification. There are many efforts to cluster EST sequences to remove the redundancy and low-quality sequences.
- Haplotype (haploid genotype) - A set of closely linked genetic markers present on one chromosome that tend to be inherited together. A haplotype may also refer to a set of single nucleotide polymorphisms (SNPs) on a single chromatid that are statistically associated with one another.
- Reference sequence/genome - A fully assembled version of a genome that can be used for mapping short DNA sequence reads for comparisons of genomes from various individuals

- Contig - A contig (from contiguous) is a set of overlapping DNA segments that together represent a consensus region of DNA. In sequencing projects, a contig refers to overlapping sequence data (reads).
- Scaffold - A scaffold is a portion of the genome sequence reconstructed from end-sequenced whole-genome shotgun clones. Scaffolds are composed of contigs and gaps.
- Specificity -The percentage of sequences that map to the intended targets out of total bases per run.
- Homopolymer - Uninterrupted stretch of a single nucleotide type (e.g., TTT or GGGGGG)
- Base Call-Base calling is the process of assigning bases (nucleobases) to chromatogram peaks.
- Homology
  - Ortholog - Orthologous sequences are homologous sequences in different species that have a common origin. Distinction of Orthologoes is a result of gradual evolutionary modifications from the common ancestor. Perform same function in different species
  - Paralog - Paralogous sequences are homologous sequences that exists within a species. They have a common origin but involve gene duplication events to arise. Perform different functions in same species
  - BLAST E-values - The BLAST programs (Basic Local Alignment Search Tools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.
  - The E-value represents the amount of alignments you would expect to find by chance that have the same score as the alignment you are looking at. The e-value is calculated with the formula E = (query length) * (length of database) * 2^-(S). A good, biologically significant e-value would be 0.05 or less.

N50: The number of largest contigs whose sum is equal to or greater than half the genome size.

L50: The smallest number of contigs whose sum produces N50

Blast - type	**query and subject**
blastn	query is DNA, subject is DNA
blastp	query is protein, subject is protein
blastx	query is nucleic acid that is translated by the program into protein sequences (all 6 reading frames); subject database is protein
tblastn	query is protein; database is DNA translated into protein sequences in all 6 reading frames.
tblastx	query is DNA translated into protein, subject is nucleotide translated into protein. Both are translated into all 6 frames. It is very slow relative to the other BLAST types.

Nutrition Genetics and Biotechnology Division

ICAR-Central Institute of Brackishwater Aquaculture
(Indian Council of Agricultural Research)
#75, Santhome High Road, MRC Nagar, Raja Annamalai Puram
Chennai, Tamil Nadu. 600 028