हर कदम, हर डगर
किसानों का हमसफर
भारतीय कृषि अनुसंधान परिषद
Agresearch with a human touch

ICAR Consortium Research Platform on Genomics

On-line training program on RNA world: Advance bioinformatics for deciphering regulatory molecules

3rd-9th November 2022

Reference manual

Dr. Sarika Sahu, Course coordinator

Dr. Ratna Prabha, Course co-coordinator

Ms. Soumya Sharma, Course co-coordinator

Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute
Library Avenue, PUSA, New Delhi - 110012
http://cabgrid.res.in/cabin/
https://iasri.icar.gov.in/

Preface

The era of the innovative world is coming up with the advent of new technologies in the field of agriculture and enhancing the goal of sustainable development worldwide. The most popular and accepted theory of life's origins reveals that the first biocatalysts were made of RNA or a very similar polymer instead of protein. Experiments are beginning to confirm that the catalytic abilities of RNA are compatible with this 'RNA world' hypothesis. An RNA molecule that does not translate into a protein is known as a non-coding RNA (ncRNA). These ncRNAs have been revolutionizing the RNA world in various aspect of life. Recently, several different systematic screens have identified a surprisingly large number of new ncRNA genes. The training program on "RNA world: A special feature to identification and characterization of non-coding RNAs" aimed to provide an insight into basic concepts of various theoretical and practical aspects of transcriptomics. This manual will help the research scholars to learn and explore the application of computational tool/techniques in their research work. The practical-oriented approach would be a big help for the new budding technologist for insight mechanisms of multicellular processes. The module contains each and every section of the program covered in the training program like 'Transcriptome Data pre-processing and Assembly', 'Differential gene expression analysis', 'Transcriptome data annotation', 'Prediction and characterization of miRNA' 'Overview of lncRNA and circular RNA', and 'Regulatory network analysis of lncRNA'.

The first talk on "whole transcriptome sequencing by next-generation sequencing (NGS) technologies or RNA-Seq" explained the complex landscape and dynamics of the transcriptome. The sequence reads obtained from the common NGS platforms, including Illumina, SOLiD, and 454, are often very short, ranging from 35bp to 500bp. As a result, it is necessary to reconstruct the full-length transcripts by transcriptome assembly. The theory and hand-on-session on 'Transcriptome Data pre-processing and assembly' provide the comprehensive knowledge of reconstructing entire transcriptome from raw NGS read including detailed understanding of all informatics challenges. It was followed by lectures on Differential gene expression (DGE) analysis. Differential gene expression (DGE) analysis is one of the most common applications of RNA-sequencing (RNA-seq) data. This process allows for the elucidation of differentially expressed genes across two or more conditions and is widely used in many applications of RNA-seq data analysis. Transcriptome annotation provides insight into the function and biological process of transcripts and the proteins they encode. The lectures on Transcriptome annotation explained various tools and techniques for transcriptome annotation.

Micro RNAs (miRNAs) are single stranded, small and non-coding endogenous RNA molecules, which control the gene expression at the post-transcriptional level by either suppression or degradation. Because of its highly conserved nature, *in silico* methods can be employed to predict novel miRNAs in plant species. The lecture on 'Prediction and characterization of miRNA' covered bioinformatics tools and techniques for miRNA prediction and functional analysis by identifying genes targeted by the miRNA.

lncRNAs are widely defined as a large and heterogeneous class of regulatory transcripts that are at least 200 nt long. circRNAs are also a subtype of endogenous ncRNAs with tissue- and cell-specific expression patterns, whose biogenesis is regulated by a particular form of alternative splicing, termed backsplicing. With the development of high-throughput technologies and extensive research reports, lncRNAs and circRNAs have gained wide attention for their roles in biological processes. The lectures on 'Overview of lncRNA and circular RNA' and 'Regulatory network analysis of lncRNA' provided detailed understanding of their roles and bioinformatics tools and techniques for analysis.

Although the manual is mainly focuses on hand-on-session but attempt taken to explain theory of each session. The details of computational tools, commands and analysis pipeline via flow chart are mentioned for each module separately that will be helpful for the naïve bioinformatician.

Sarika Sahu

Chapter 1
Overview of training

Ratna Prabha, Sarika Sahu, Soumya Sharma
Division of Agricultural Bioinformatics,
ICAR-Indian Agricultural Statistics Research Institute

Introduction:

This online training "RNA world: Advance Bioinformatics for deciphering regulatory molecules" organized under the aegis of CRP-Genomics project, aims to provide a comprehensive view of the main facets involved in theoretical and practical aspects of this very rapidly growing field of non-coding RNAs. An RNA molecule that does not translate into a protein is known as a non-coding RNA (ncRNA). These ncRNAs have been revolutionizing the RNA world in various aspect of life. Recently, several different systematic screens have identified a surprisingly large number of new ncRNA genes.

RNA biology is the combination of all RNAs whether coding or noncoding. The discovery of non-coding RNAs led to the revolution in RNA world (Derks *et al.* 2015). Noncoding RNAs (ncRNAs) play an important role in various biological processes and gene-disease association (Nallar and Kalvakolanu, 2013). Among the ncRNAs, the most studied ncRNAs are microRNA, which play a major role in gene expression (Hermeking, 2012). However, it has been revealed that long ncRNAs (lncRNAs) also play a very important role in various biological pathways within the cell (Huarte *et al.,* 2010). Researchers reported that several lncRNAs are expressed during stress conditions and are involved in stress-responsive regulation (Zheng *et al.* 2014, Heo *et al.* 2011, Liu *et al.* 2012). lncRNAs are non-coding RNAs whose length is more than 200 base pairs and biochemically resemble mRNAs but they do not translate into proteins. Despite noncoding RNAs, lncRNAs function as RNA genes as well as regulate distant genes. Ponting *et al.* (2009) classified lncRNAs into sense, anti-sense, bidirectional, intronic and intergenic on the basis of their chromosomal localization. In addition, the lncRNAs are normally expressed at low levels and lack sequence similarities among the plant species (Marques and Ponting, 2014). Plethora of literature is available for the identification of lncRNAs in animals while very few are reported on the presence of lncRNAs in plants (Liu *et al.*,2017). The analysis of lncRNA became very easy with the advent of state-of-art technologies like next-generation sequencing. lncRNAs were identified in model plant organisms like Arabidopsis thaliana (Wang *et al.* 2014, Lu *et al.* 2017, Sun *et al*. 2020) Two lncRNAs namely: COOLAIR (cool-assisted intronic non-coding RNA) and COLDAIR (cold-assisted intronic non-coding RNA) regulates the flowering time epigenetic repression of FLC (Flowering Locus C) in Arabidopsis (Heo and Sung, 2011). Another important lncRNA: LDMAR (long-day-specific male-fertility-associated RNA) is involved in the regulation of photoperiod male sterility in rice (Ding *et al.* 2012) and participated in ripening of tomato (Zhu *et al.* 2015). These are few examples to be mentioned and suggest the importance of ncRNAs in the plant and crop systems.

Objectives of this training were

- Profiling of RNAs by bioinformatics tools.
- Role of RNAs and non-coding RNA in gene regulatory network.
- Development of analytical skills through lectures and hands-on session.

Different modules covered under this training program were as following

- **Differential gene expression.**

  Sequencing platform and Quality Check

  Assembly: de novo and reference based and annotation

- **Profiling of RNA regulatory molecule and their role in the regulation of biological processes**

  Prediction and characterization of miRNAs

  Prediction and characterization of lncRNAs

  Prediction and characterization of circRNAs

- **Regulatory network analysis of RNAs.**

Different theoretical and Practical Sessions were taken during this training program. In this manual, different session taken during training are described in detail. Chapter 2 focuses over RNA-sequencing analysis. Chapter 3 mentions detailed practical procedure taught in the training for Transcriptome Data Pre-processing and Assembly while Chapter 4 given an overview of genome annotation with special focus over gene prediction. Chapter 5 gives detail about Differential Gene Expression Analysis. Chapter 6 provide detail about different tools and execution carried out for Transcriptome data annotation**.** Chapter 7 provides glimpse about world of miRNA. In chapter 8, hands on session over prediction and Characterization of miRNA is covered. Chapter 9 focuses over Circular RNA and about its basic concept and their role in various processes and also covers details of Hands-on-session for circRNA prediction. In chapter 10, aspects of RNAome in biofortification of plant and animal traits is covered.

References

1. Derks KW, Misovic B, van den Hout MC, Kockx CE, Payan Gomez C, Brouwer RW, Vrieling H, Hoeijmakers JH, van IJcken WF, Pothof J. Deciphering the RNA landscape by RNAome sequencing. RNA biology. 2015 Jan 2;12(1):30-42.
2. Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, Xu C, Li X, Xiao J, Zhang Q. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proceedings of the National Academy of Sciences. 2012 Feb 14;109(7):2654-9.
3. Heo JB, Sung S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science. 2011 Jan 7;331(6013):76-9.
4. Hermeking H. MicroRNAs in the p53 network: micromanagement of tumour suppression. Nature reviews cancer. 2012 Sep;12(9):613-26.
5. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell. 2010 Aug 6;142(3):409-19.
6. Liu H, Wang X, Wang HD, Wu J, Ren J, Meng L, Wu Q, Dong H, Wu J, Kao TY, Ge Q. Escherichia coli noncoding RNAs can affect gene expression and physiology of Caenorhabditis elegans. Nature communications. 2012 Sep 25;3(1):1-1.
7. Lu Z, Xia X, Jiang B, Ma K, Zhu L, Wang L, Jin B. Identification and characterization of novel lncRNAs in Arabidopsis thaliana. Biochemical and biophysical research communications. 2017 Jun 24;488(2):348-54.
8. Marques AC, Ponting CP. Intergenic lncRNAs and the evolution of gene expression. Current opinion in genetics & development. 2014 Aug 1;27:48-53.
9. Nallar SC, Kalvakolanu DV. Regulation of snoRNAs in cancer: close encounters with interferon. Journal of Interferon & Cytokine Research. 2013 Apr 1;33(4):189-98.
10. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009 Feb 20;136(4):629-41.
11. Sun Z, Huang K, Han Z, Wang P, Fang Y. Genome-wide identification of Arabidopsis long noncoding RNAs in response to the blue light. Scientific reports. 2020 Apr 10;10(1):1-0.

12. Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, Xu J, Chua NH. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. Genome research. 2014 Mar 1;24(3):444-53.
13. Zeng H, Wang G, Hu X, Wang H, Du L, Zhu Y. Role of microRNAs in plant responses to nutrient stress. Plant and Soil. 2014 Jan;374(1):1005-21.
14. Zhu B, Yang Y, Li R, Fu D, Wen L, Luo Y, Zhu H. RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. Journal of Experimental Botany. 2015 Aug 1;66(15):4483-95.

Chapter 2
RNA-SEQUENCING ANALYSIS

Dwijesh Chand Mishra
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Introduction

The advent of Next-Generation Sequencing (NGS) technology has transformed genomic studies. One important application of NGS technology is the study of the *transcriptome*, which is defined as the complete collection of all the RNA molecules in a cell. Various types of RNA that have been classified so far are shown in **Fig. 1**. All of these molecules are called *transcripts* since they are produced by process of transcription.

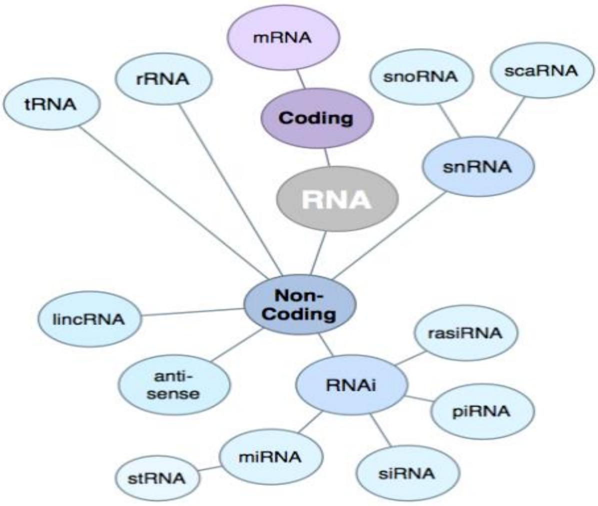


Fig. 1: Different types of RNA

(Image source: http://scienceblogs.com/digitalbio/2011/01/08/next-gene-sequencing)

Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease [1]. The main purpose of transcriptomics are: to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions.

The study of transcriptome is carried out through sequencing of RNAs. RNA sequencing *(RNA-Seq)* is a powerful method for discovering, profiling, and quantifying RNA transcripts [2]. RNA-Seq uses NGS datasets to obtain sequence reads from millions of individual RNAs. The RNA-Seq analysis is performed in several steps: First, all genes are extracted from the reference genome (using annotations of type *gene*). Other annotations on the gene sequences are preserved (e.g. CDS information about coding sequences etc). Next, all annotated transcripts (using annotations of type *mRNA*) are extracted [3]. If there are several annotated splice variants, they are all extracted. An example is shown in below **Fig. 2(a).**

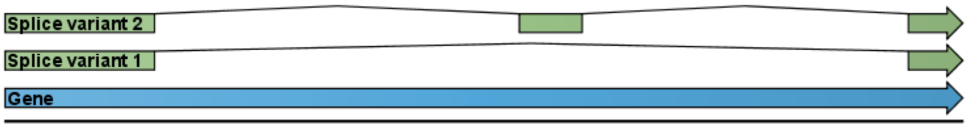


Fig. 2(a): A simple gene with three exons and two splice variants.

The given example is a simple gene with three exons and two splice variants. The transcripts are extracted as shown in **Fig. 2(b).**

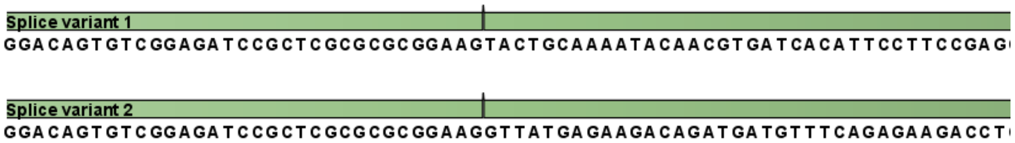

Splice variant 1
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGTACTGCAAAATACAACGTGATCACATTCCTTCCGAG
Splice variant 2
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGGTTATGAGAAGACAGATGATGTTTCAGAGAAGACCT

Fig. 2(b): All the exon-exon junctions are joined in the extracted transcript.

Next, the reads are mapped against all the transcripts plus the entire gene [see **Fig. 2(c)**].

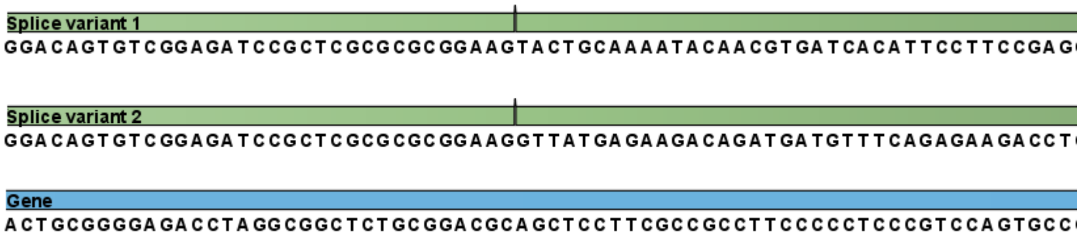

Splice variant 1
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGTACTGCAAAATACAACGTGATCACATTCCTTCCGAG
Splice variant 2
GGACAGTGTCGGAGATCCGCTCGCGCGCGGAAGGTTATGAGAAGACAGATGATGTTTCAGAGAAGACCT
Gene
ACTGCGGGGAGACCTAGGCGGCTCTGCGGACGCAGCTCCTTCGCCGCCTTCCCCCTCCCGTCCAGTGCC

Fig. 2(c): The reference for mapping: all the exon-exon junctions and the gene.

(Image source: CLC Genomic workbench tutorials)

From this mapping, the reads are categorized and assigned to the genes and expression values for each gene and each transcript are calculated and putative exons are then identified.

RNA Sequencing Experiment

In a standard RNA-seq experiment, a sample of RNA is converted to a library of complementary DNA fragments and then sequenced on a high-throughput sequencing platform, such as Illumina's Genome Analyzer, SOLiDor Roche 454 [4]. Millions of short sequences, or reads, are obtained from this sequencing and then mapped to a reference genome (**Fig. 3**). The count of reads mapped to a given gene measures the expression level of this gene. The unmapped reads are usually discarded and mapped reads for each sample are assembled into gene-level, exon-level or transcript-level expression summaries, depending on the objectives of the experiment. The count of reads mapped to a given gene/exon/transcript measures the expression level for this region of the genome or transcriptome.

One of the primary goals for most RNA-seq experiments is to compare the gene expression levels across various treatments. A simple and common RNA-seq study involves two treatments in a randomized complete design, for example, treated versus untreated cells, two different tissues from an organism, plants, etc. In most of the studies, researchers are particularly interested in detecting gene with differential expressions (DE). A gene is declared differentially expressed if an observed difference or change in read counts between two experimental conditions is statistically significant, i.e. if the difference is greater than what would be expected just due to random variation [5]. Detecting DE genes can also be an important pre-step for subsequent studies, such as clustering gene expression profiles or testing gene set enrichments.

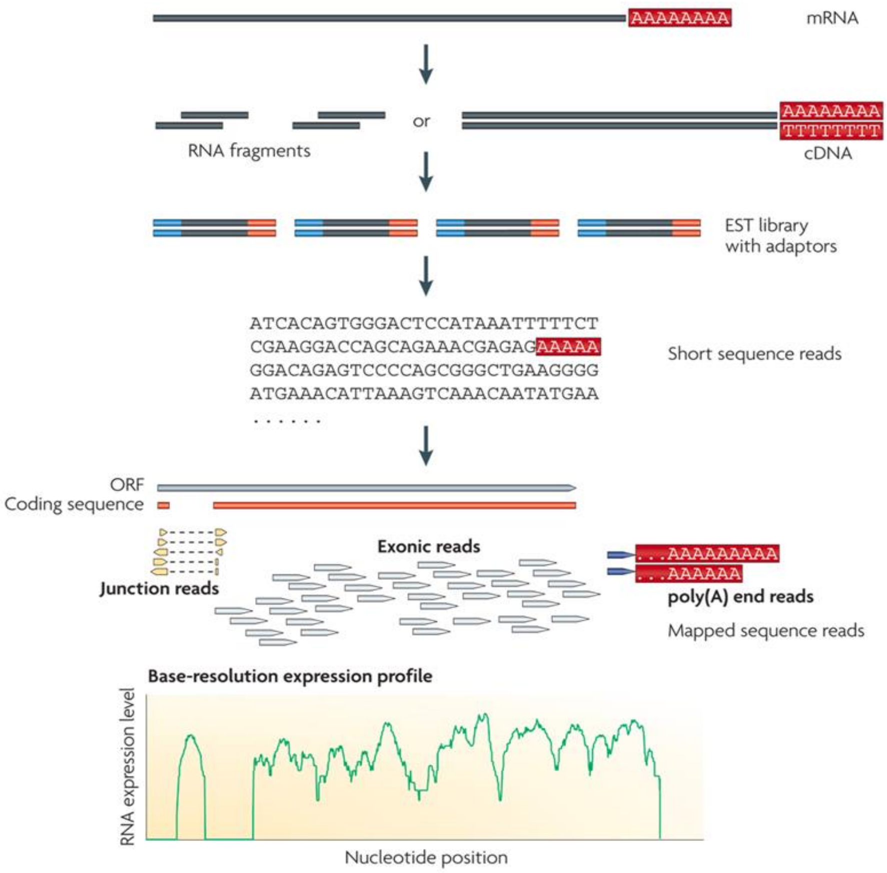


Fig. 3: General RNA-seq experiment. mRNA is converted to cDNA, and fragments from that library are used to generate short sequence reads. Those reads are assembled into contigs which may be mapped to reference sequences (Wang et al., 2009).

Analysing RNA-Seq data

RNA-seq experiments must be analyzed with robust, efficient and statistically correct algorithms. Fortunately, the bioinformatics community has been striving hard at work for incorporating mathematics, statistics and computer science for RNA-seq and building these ideas into software tools. RNA-seq analysis tools generally fall into three categories: (i) those for read alignment; (ii) those for transcript assembly or genome annotation; and (iii) those for transcript and gene quantification. Some of the open source software available for RNA-seq analysis are as follows:

- **Data preprocessing**
  - Fastx toolkit
  - Samtools
- **Short reads aligners**
  - Bowtie, TOPHAT, Stampy, BWA, Novoalign, etc

- **Expression studies**
    - Cufflinks package
    - R packages (DESeq, edgeR, *more…*)
- **Visualisation**
    - CummeRbund, IGV, Bedtools, UCSC Genome Browser, etc.

Besides there are commercially data analysis pipelines like GenomeQuest, CLCBio etc available for researchers to use. The most commonly used pipeline is to identify protein coding genes by aligning RNA-Seq data to annotate data from sources like RefSeq. After generating the alignments, the number of aligning sequences is counted for each position. Since each alignment represents a transcript, the alignments allow to count the number of RNA molecules produced from every gene.

Using NGS technology, RNA-Seq enables to count the number of reads that align to one of thousands of different cDNAs, producing results similar to those of gene expression microarrays [6]. Sequences generated from an RNA-Seq experiment are usually mapped to libraries of known exons in known transcripts. RNA-Seq can be used for discovery applications such as identifying alternative splicing events, allele-specific expression, and rare and novel transcripts [7]. The sequencing output files (compressed FASTQ files) are the input for secondary analysis. Reads are aligned to an annotated reference genome, and those aligning to exons, genes and splice junctions are counted. The final steps are data visualisation and interpretation, consisting of calculating gene- and transcript-expression and reporting differential expression. A general Bioinformatics workflow to map transcripts from RNA-seq data is shown in **Fig. 4**.

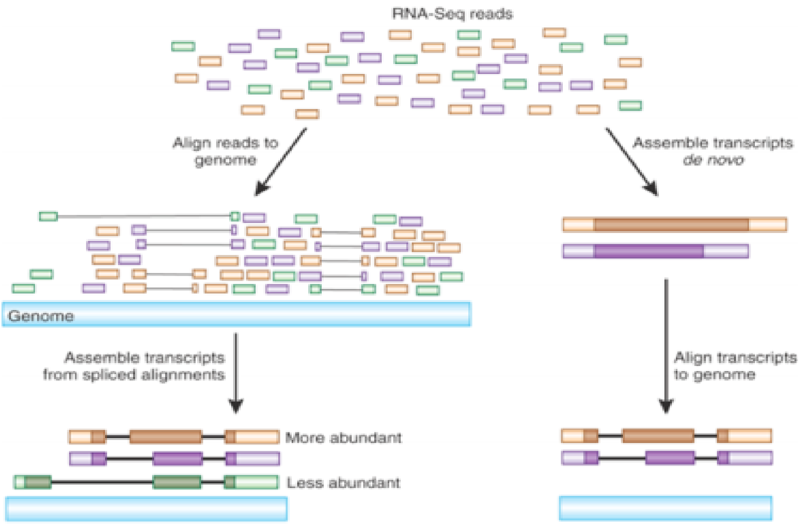

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Fig. 4: **RNA-seq workflow (Adapted fromAdvancing RNA-Seq analysis Brian J. Haas and Michael C. Zody Nature Biotechnology 28, 421-423 (2010)**

RPKM (Reads per KB per million reads)

RNA-Seq provides quantitative approximations of the abundance of target transcripts in the form of counts. However, these counts must be normalized to remove technical biases inherent in the preparation steps for RNA-Seq, in particular the length of the RNA species and the sequencing depth of a sample. The most commonly used is RPKM (Reads Per Kilobase of exon model per Million mapped reads). The RPKM measure of read density reflects the molar concentration of a transcript in the starting sample by normalizing for RNA length and for the total read number in the measurement [8]. RPKM is mathematically represented as:

$$\text{RPKM}=\frac{total\ exon\ reads}{mapped\ reads\ (millions) \text{X}\ exon\ length\ (KB)}$$

Total exon reads

This is the number of reads that have been mapped to a region in which an exon is annotated for the gene or across the boundaries of two exons or an intron and an exon for an annotated transcript of the gene. For eukaryotes, exons and their internal relationships are defined by annotations of type mRNA.

Exon length

This is calculated as the sum of the lengths of all exons annotated for the gene. Each exon is included only once in this sum, even if it is present in more annotated transcripts for the gene. Partly overlapping exons will count with their full length, even though they share the same region.

Mapped reads

The total gene reads for a gene is the total number of reads that after mapping have been mapped to the region of the gene. A gene's region is that comprised of the flanking regions, the exons, the introns and across exon-exon boundaries of all transcripts annotated for the gene. Thus, the sum of the total gene reads numbers is the number of mapped reads for the sample.

Applications of RNA-seq

This technique can be used to:

- Measure gene expression
- Transcriptome assembly, gene discovery and annotation
- Detect differential transcript abundances between tissues, developmental stages, genetic backgrounds, and environmental conditions
- Characterize alternative splicing, alternative polyadenylation, and alternative transcription.

Future Directions

Although RNA-Seq is still in the infancy stages of use, it has clear advantages over previously developed transcriptomic methods. Compared with microarray, which has been the dominant approach of studying gene expression in the last two decades, RNA-seq technology has a wider measurable range of expression levels, less noise, higher throughput, and more information to detect allele-specific expression, novel promoters, and isoforms [9]. For these reasons, RNA-seq is gradually replacing the array-based approach as the major platform in gene expression studies. The next big challenge for RNA-Seq is to target more complex transcriptomes to identify and track the expression changes of rare RNA isoforms from all genes. Technologies that will advance achievement of this goal are pair-end sequencing, strand-specific sequencing and the use of longer reads to increase coverage and depth. As the cost of sequencing continues to fall, RNA-Seq is expected to replace microarrays for many applications that involve determining the structure and dynamics of the transcriptome.

References

15. https://www.genome.gov/13014330
16. WangZ., GersteinM., SynderM. (2009). Rna-seq: a revolutionary tool for transciptomics, Nat Rev Genet 10(1): 57–63.
17. http://scienceblogs.com/digitalbio/2011/01/08/next-gene-sequencing-results-a/
18. Shendure J, Ji H (2008) Next-generation RNA sequencing. Nature Biotechnology 26: 2514-2521
19. Anders S, Huber W (2010). Differential expression analysis for sequence count data. Genome Biol. 11:R106.
    Illumina, Inc,. (2011). Getting started with RNA-Seq Data Analysis. Pub. No. 470-2011-003.
20. Illumina, Inc,. (2011). RNA-Seq Data Comparison with Gene Expression Microarrays. A cross-platform comparison of differential gene expression analysis. Pub. No. 470-2011-004
21. Yaqing Si (2012). Statistical analysis of RNA-seq data from next-generation sequencing technology. PhD thesis. Iowa State University, Ames, Iowa.
22. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods, 5(7):621-628.
23. Wang L., Si Y., Dedow L.K., Shao Y., Liu P., Brutnell T.P. (2010). A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. PLoS One 6(10):e26426.
24. Brian J. H. and Michael C. Z. (2010). Advancing RNA-Seq analysis. Nature Biotechnology 28, 421-423.

Chapter 3

Transcriptome Data Pre-processing and Assembly

Soumya Sharma, Ratna Prabha

ICAR-Indian Agricultural Statistics Research Institute

Transcript profiling ("Transcriptomics") is a widely used technique that obtains information on the abundance of multiple mRNA transcripts within a biological sample simultaneously. Therefore, when a number of such samples are analysed, as in a scientific experiment, large and complex data sets are gene-rated. RNA-Seq technology utilizing NGS sequencing has emerged as an attractive alternative to traditional microarray platforms for conducting transcriptional profiling. Next generation sequencing (NGS) experiments generate a tremendous amount of data which can't be directly analyzed in any meaningful way. Selecting the right analytical approach along with an appropriate set of bioinformatics tools is key to extract useful information from RNA-Seq data while avoiding misinterpretation or bias. In the present section we will discuss about the assembly of short-read Illumina sequencing data, which is commonly used for RNA-Seq experiments.

Requirements for RNA-Seq Data Assembly

Hardware

- Linux environment or server
- Accessed via shell terminals, such as PuTTY or MobaXterm
- Can use a virtual machine on Windows
- 32GB RAM recommended if working with larger genomes
- 1TB storage or higher recommended for smaller projects

Software

- FastQC
  https://www.bioinformatics.babraham.ac.uk/projects/download.html

- Trimmomatic
  http://www.usadellab.org/cms/?page=trimmomatic

- Bowtie2
  https://sourceforge.net/projects/bowtie-bio/files/bowtie2/

- Tophat
  https://ccb.jhu.edu/software/tophat/index.shtml

- Cufflinks
  http://cole-trapnell-lab.github.io/cufflinks/getting_started/

- Trinity
  https://github.com/trinityrnaseq/trinityrnaseq/wiki/Installing-Trinity

Pre-processing of RNA-Seq Data

First, switch to the where the FASTQ files are stored directory. Use the cd command (i.e., change directory) followed by the path of the directory.

```
>> cd /path/to/folder_name/
```

Next, you can check the FASTQ files by using the ls command (i.e., listing), which shows the contents of the current working directory.

Data files from sequencing providers are typically compressed and have the extension ".fastq.gz". These files contain structured information about individual NGS reads—a unique identifier, the called bases, and the associated quality scores.

Lastly, you can make an output directory using the mkdir command (i.e., make directory). Output files can be stored here.

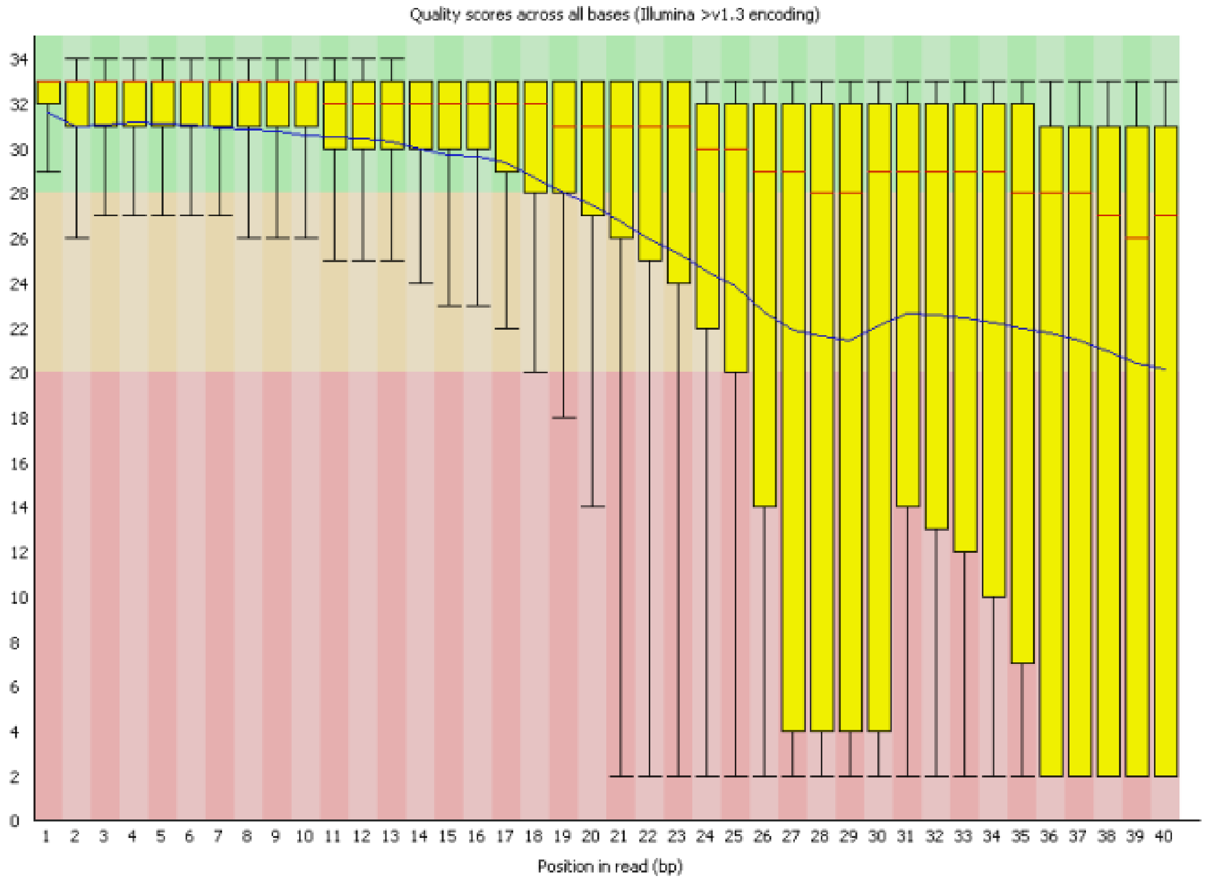
```
>> mkdir /path/to/output_folder/
```

1. Check quality with FastQC

Run FastQC to check the raw data quality.

```
>> fastqc sample_01.fastq.gz --extract -o /path/to/output_folder
```

The output contains graphs and statistics about the raw quality, including quality scores, GC content, adapter percentage, and more. Below is an examples of the output file "Per base Sequence quality".



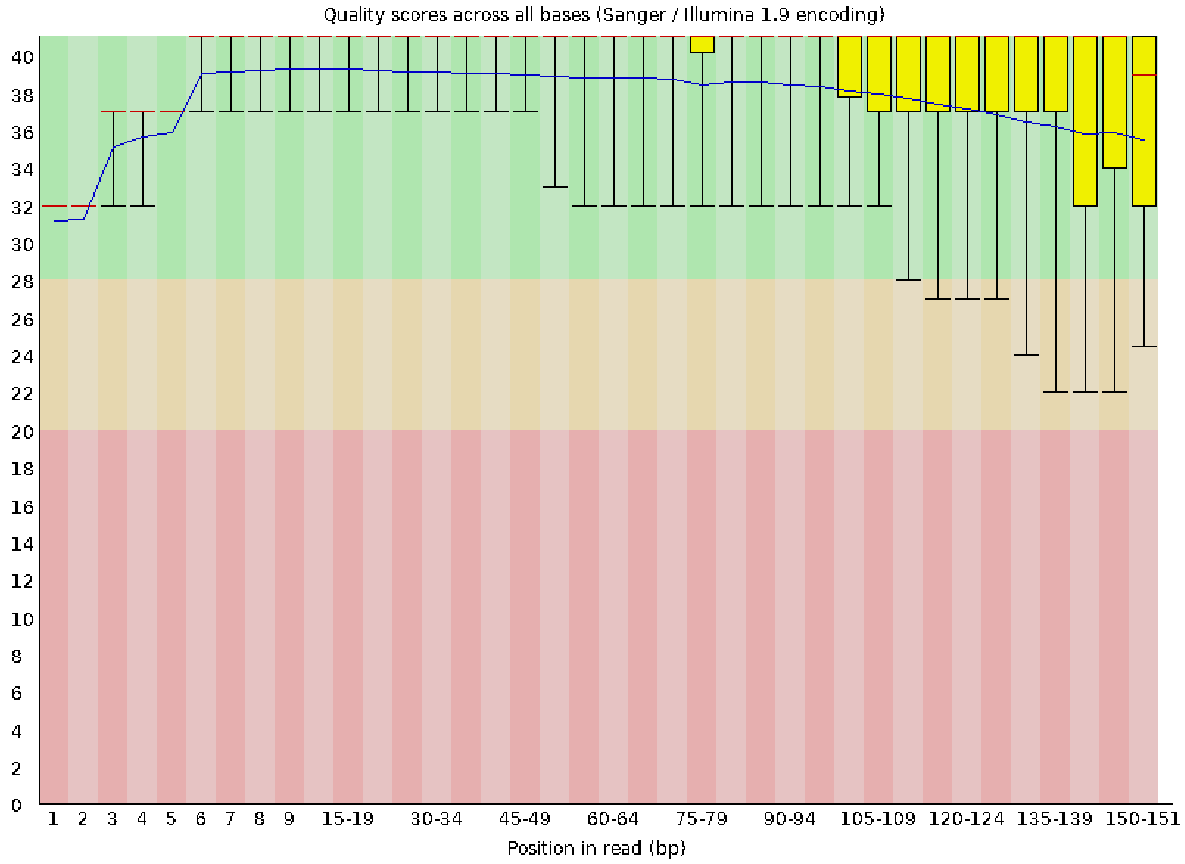
Per base sequence quality. Quality scores for each base position in the read are represented as box plots. The blue line represents the average quality score. High-quality data will typically have over 80% of bases with a quality score of 30 or higher (i.e., Q30 > 80%). Q30 represents 99.9% accuracy in the base call, or an error rate of 1 in 1000. A dip in quality is expected towards the end of the read.

2. Trim reads with Trimmomatic

Poor-quality regions and adapter sequences should be trimmed from the reads before further analysis. Trimmomatic can be used for trimming the low quality reads and adapter sequences.

```
>> trimmomatic  PE input_forward.fastq.gz  input_reverse.fastq.gz
output_forward_paired.fastq.gz  output_forward_unpaired.fastq.gz
output_reverse_paired.fastq.gz  output_reverse_unpaired.fastq.gz
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3
MINLEN:36
```

Run FastQC again on the trimmed treads to confirm that the new quality is acceptable.



Transcriptome Assembly

Refrence based Assembly

1. Indexing the reference genome

First, index the reference genome using Bowtie2 to prepare it for alignment. Adding gene annotation information to the reference genome will facilitate alignment of RNA-Seq reads across exon-intron boundaries. This indexing step is only required once; you can then use the indexed genome repeatedly in future analysis.

```
>> bowtie-build [options]* <input referencegenome fasta file> < basename of the index files >
```

It results in 6 files with extention .bt2

2. Map/Align the reads to reference Genome

Then, align the reads using Tophat.

```
>> tophat [options]* <genome_index_base> PE_reads_1.fq.gz,SE_reads.fa PE_reads_2.fq.gz
```

- or -

```
>> tophat [options]* <genome_index_base> PE_reads_1.fq.gz PE_reads_2.fq.gz,SE_reads.fa
```

Check the mapping statistics in the [sample_name]Log.final.out file to ensure the BAM file was generated properly and the reads align to the genome correctly. Uniquely mapped reads are the most useful for expression analysis, as there is high confidence in which loci they represent. In general, >60-70% for the "uniquely mapped reads %" metric is considered good; a significantly lower value warrants further investigation.

3. Assemble the mapped reads

Use Cufflinks program to assemble aligned RNA-Seq reads into transcripts, estimate their abundances, test for differential expression and regulation, and provide transcript quantification. Some of the tools part of Cufflinks can be run individually, while others are part of a larger workflow.

```
>> cufflinks [options] input_alignments.[sam|bam]
```

The program cufflinks produces number of files in its predefined output directory. Some of the generated files are:

transcripts.gtf: The GTF file contains Cufflinks' assembled isoforms where there is one GTF record per row, and each record represents either a transcript or an exon within a transcript
isoforms.fpkm_tracking: This file contains the estimated isoform-level expression values in the generic FPKM Tracking Format
genes.fpkm_tracking: This file contains the estimated gene-level expression values in the generic FPKM Tracking Format

De novo Assembly

De novo transcriptome assembly is often the preferred method to studying non-model organisms, since reference-based methods are not possible without an existing genome. De novo assembly can be performed using Trinity assembler.

A typical Trinity command for assembling non-strand-specific RNA-seq data would be like so, running the entire process on a single high-memory server (aim for ~1G RAM per ~1M ~76 base Illumina paired reads, but often much less memory is required):

```
Trinity --seqType fq --max_memory 50G  --left reads_1.fq.gz  --right reads_2.fq.gz --CPU 6
```

If multiple sets of fastq files are available, such as corresponding to multiple tissue types or conditions, etc., indicate them to Trinity like following:

```
 Trinity --seqType fq --max_memory 50G --left condA_1.fq.gz,condB_1.fq.gz,condC_1.fq.gz –right condA_2.fq.gz,condB_2.fq.gz,condC_2.fq.gz  --CPU 6
```

When Trinity completes, it will create a 'Trinity.fasta' output file in the 'trinity_out_dir/' output directory (or output directory specified).

Trinity groups transcripts into clusters based on shared sequence content. Such a transcript cluster is very loosely referred to as a 'gene'. This information is encoded in the Trinity fasta accession.

Chapter 4

GENOME ANNOTATION: GENE PREDICTION

Sanjeev Kumar, D.C. Mishra and Jyotika Bhati

Introduction

Until the genome revolution, genes were identified by researchers with specific interests in a particular protein or cellular process. Once identified, these genes were isolated, typically by cloning and sequencing cDNAs, usually followed by targeted sequencing of the longer genomics segments that code for the cDNAs. Once an organism's entire genome sequence becomes available, there is strong motivation for finding all the genes encoded by a genome at once rather than in a piecemeal approach. Such catalogue is immensely valuable to researchers, as they can learn much more from the whole picture than from a much more limited set of genes. For example, genes of similar sequence can be identified, evolutionary and functional relationships can be elucidated, and a global picture of how many and what types of genes are present in a genome can be seen. A significant portion of the effort in genome sequencing is devoted to the process of *annotation*, in which genes, regulatory elements, and other features of the sequence are identifies as thoroughly as possible and catalogued in a standard format in public databases so that researchers can easily use the information. Functional genomics research has expanded enormously in the last decade and particularly the plant biology research community. Functional annotation of novel DNA sequences is probably one of the top requirements in functional genomics as this holds, to a great extent, the key to the biological interpretation of experimental results.

Computational Gene Prediction

Computational gene prediction is becoming more and more essential for the automatic analysis and annotation of large uncharacterized genomic sequences. In the past two decades, many algorithms have been evolved to predict protein coding regions of the DNA sequences. They all have in common, to varying degree, the ability to differentiate between gene features like Exons, Introns, Splicing sites, Regulatory sites etc. Gene prediction methods predicts coding region in the query sequences and then annotates the sequences databases.

Gene Structure and Expression

The gene structure and the gene expression mechanism in eukaryotes are far more complicated than in prokaryotes. In typical eukaryotes, the region of the DNA coding for a protein is usually not continuous. This region is composed of alternating stretches of *exons* and *introns*. During transcription, both exons and introns are transcribed onto the RNA, in their linear order. Thereafter, a process called *splicing* takes place, in which, the intron sequences are excised and discarded from the RNA sequence. The remaining RNA segments, the ones corresponding to the exons are ligated to form the mature RNA strand. A typical multi-exon gene has the following structure (as illustrated in Fig. 1).

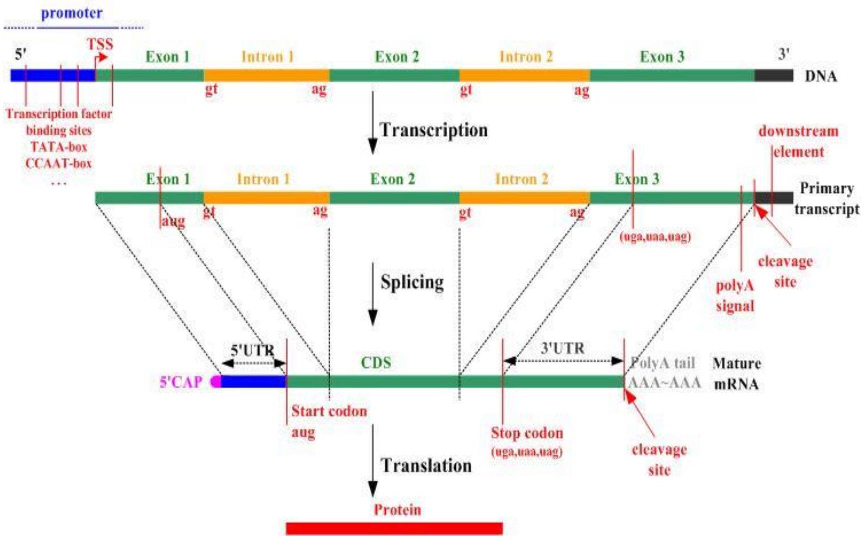


Fig. 1: Representative Diagram of Protein Coding Eukaryotic Gene

It starts with the promoter region, which is followed by a transcribed but non-coding region called *5' untranslated region (5' UTR)*. Then follows the initial exon which contains the start codon. Following the initial exon, there is an alternating series of introns and internal exons, followed by the terminating exon, which contains the stop codon. It is followed by another non-coding region called the *3' UTR*. Ending the eukaryotic gene, there is a polyadenylation (polyA) signal: the nucleotide Adenine repeating several times. The exon-intron boundaries (i.e., the splice sites) are signalled by specific short (2bp long) sequences. The 5'(3') end of an intron (exon) is called the *donor* site, and the 3'(5') end of an intron (exon) is called the *acceptor* site. The problem of gene identification is complicated in the case of eukaryotes by the vast variation that is found in gene structure.

Gene Prediction Methods

There are mainly two classes of methods for computational gene prediction (Fig. 2). One is based on sequence similarity searches, while the other is gene structure and signal-based searches, which is also referred to as *Ab initio* gene finding.

Sequence Similarity Searches

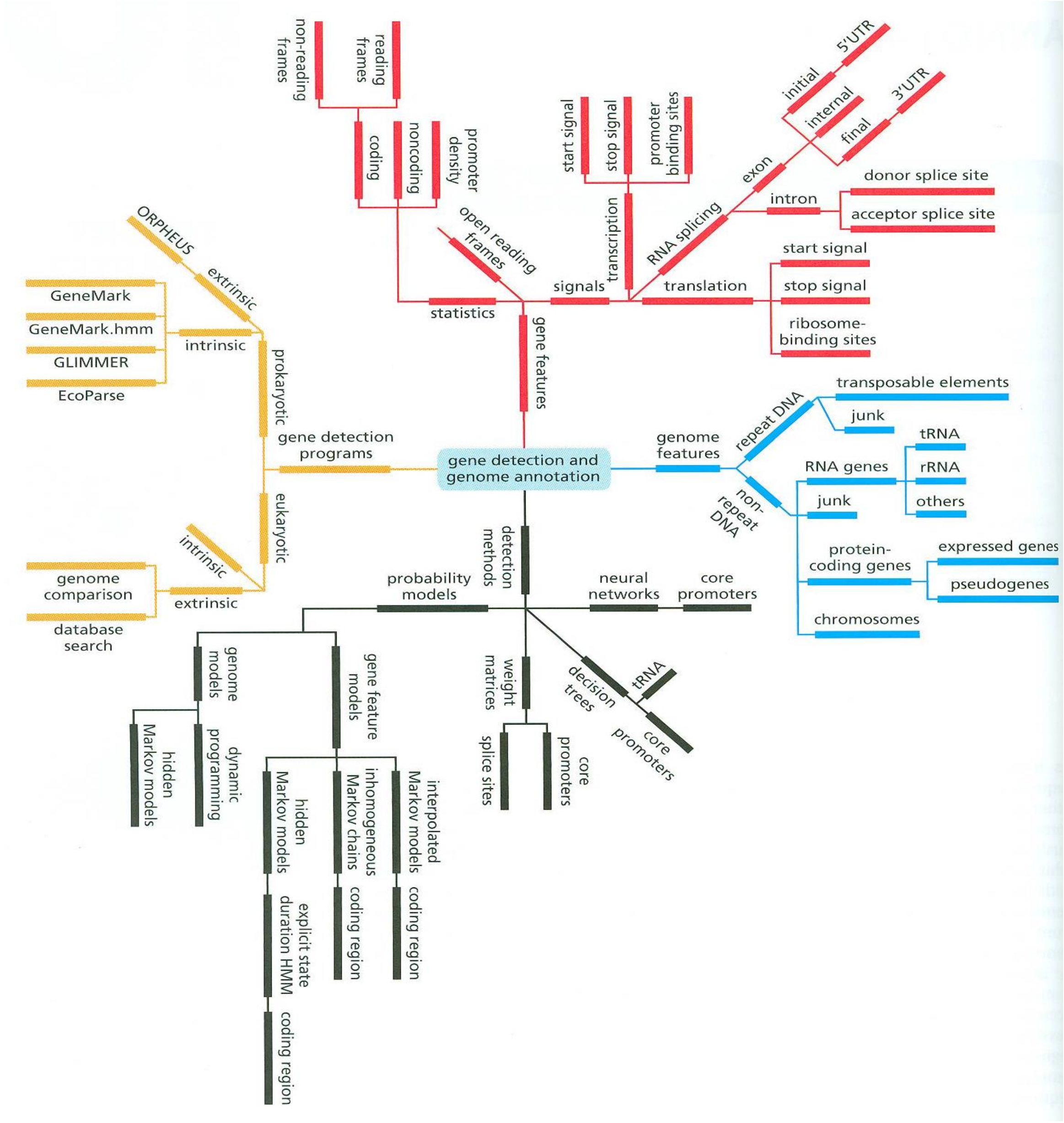
Sequence similarity search is a conceptually simple approach that is based on finding similarity in gene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome. This approach is based on the assumption that functional regions (exons) are more conserved evolutionarily than non-functional regions (intergenic or intronic regions). Once there is similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region. EST-based sequence similarity usually has drawbacks in that ESTs only correspond to small portions of the gene sequence, which means that it is often difficult to predict the complete gene structure of a given region. Local alignment and global alignment are two methods based on similarity searches. The most common local alignment tool is the BLAST family of programs, which detects sequence similarity to known genes, proteins, or ESTs. The biggest limitation to this type of approaches is that only about half of the genes being discovered have significant homology to genes in the databases.

Ab initio Gene Prediction Methods

The second class of methods for the computational identification of genes is to use gene structure as a template to detect genes, which is also called *ab initio* prediction. *Ab initio* gene predictions rely on two types of sequence information: signal sensors and content sensors. Signal sensors refer to short sequence motifs, such as splice sites, branch points, poly pyrimidine tracts, start codons and stop codons. Exon detection must rely on the content sensors, which refer to the patterns of codon usage that are unique to a species, and allow coding sequences to be distinguished from the surrounding non-coding sequences by statistical detection algorithms.

Many algorithms are applied for modelling gene structure, such as Dynamic Programming, linear discriminant analysis, Linguist methods, Hidden Markov Model and Neural Network. Based on these models, a great number of *ab initio* gene prediction programs have been developed.

Fig. 2: Diagrammatic Representation of Gene Prediction and Annotation



Gene Discovery in Prokaryotic Genomes

Discovery of genes in Prokaryote is relatively easy, due to the higher gene density typical of prokaryotes and the absence of introns in their protein coding regions. DNA sequences that encode proteins are transcribed into mRNA, and the mRNA is usually translated into proteins without significant modification. The longest ORFs (open reading frames) running from the first available start codon on the mRNA to the next stop codon in the same reading frame generally provide a good, but not assured prediction of the protein coding regions. Several methods have been devised that use different types of Markov models in order to capture the compositional differences among coding regions, “shadow" coding regions (coding on the opposite DNA strand), and noncoding DNA. Such methods, including ECOPARSE, the widely used GENMARK, and Glimmer program, appear to be able to identify most protein coding genes with good performance (Fig. 3).

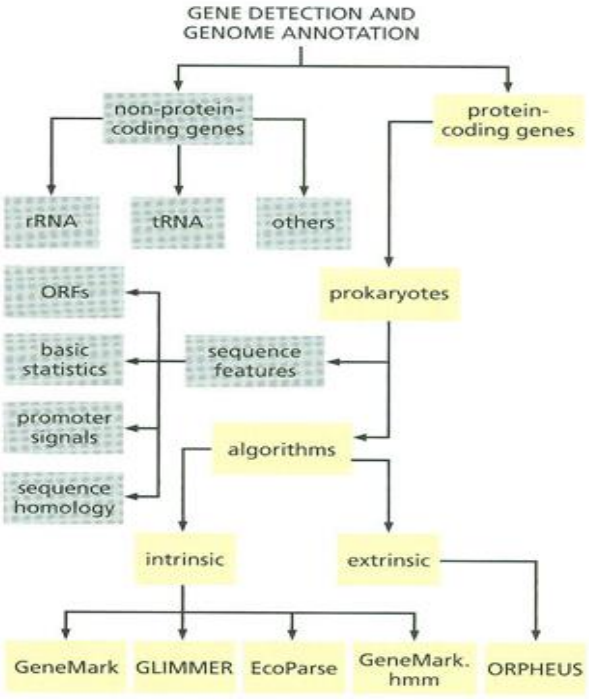


Fig. 3: Flow Diagram of Prokaryotic Gene Discovery

Gene Discovery in Eukaryotic Genome

It is a quite different problem from that encountered in prokaryotes. Transcription of protein coding regions initiated at specific promoter sequences is followed by removal of noncoding sequences (introns) from pre-mRNA by a splicing mechanism, leaving the protein encoding exons. Once the introns have been removed and certain other modifications to the mature RNA have been made, the resulting mature mRNA can be translated in the 5` to 3` direction, usually from the first start codon to the first stop codon. As a result of the presence of intron sequences in the genomic DNA sequences of eukaryotes, the ORF corresponding to an encoded gene will be interrupted by the presence of introns that usually generate stop codons (Fig.4).

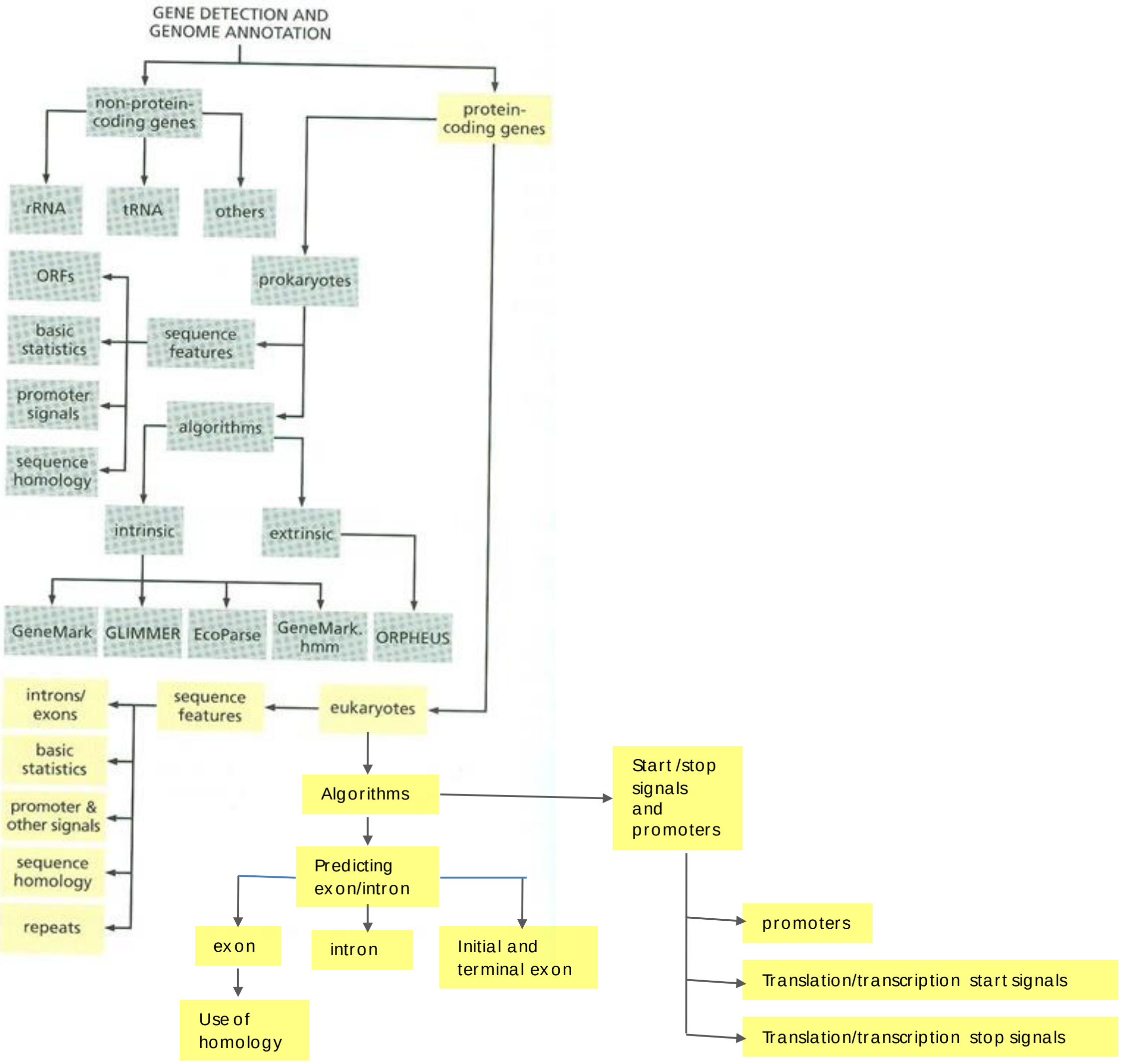


Fig. 4: Flow Diagram of Eukaryotic Gene Discovery

Gene Prediction Program

There are two basic problems in gene prediction: prediction of protein coding regions and prediction of the functional sites of genes. Gene prediction program can be classified into four generations. The first generation of programs was designed to identify approximate locations of coding regions in genomic DNA. The most widely known programs were probably TestCode and GRAIL. But they could not accurately predict precise exon locations.

The second generation, such as SORFIND and Xpound, combined splice signal and coding region identification to predict potential exons, but did not attempt to assemble predicted exons into complete genes. The next generation of programs attempted the more difficult task of predicting complete gene structures. A variety of programs have been developed, including GeneID, GeneParser, GenLang, and FGENEH. However, the performance of those programs remained rather poor. Moreover, those programs were all based on the assumption that the input sequence contains exactly one complete gene, which is not often the case. To solve this problem and improve accuracy and applicability further, GENSCAN and AUGUSTUS were developed, which could be classified into the fourth generation.

GeneMark

GeneMark uses a Markov Chain model to represent the statistics of the coding and noncoding frames. The method uses the dicodon statistics to identify coding regions. Consider the analysis of a sequence x whose base at the ith position is called x_i. The Markov chains used are fifth order, and consist of a terms such as $P(a/x_1x_2x_3x_4x_5)$, which represent the probability of the sixth base of the sequence x being given a given that the previous five bases in the sequence x where $x_1x_2x_3x_4x_5$, resulting in the first dicodon of the sequence being $x_1x_2x_3x_4x_5a$. These terms must be defined for all possible pentamers with the general sequence $b_1b_2b_3b_4b_5$. The values of these terms can be obtained of analysis of data, consisting of nucleotide sequence in which the coding regions have been actually identified. When there are sufficient data, they are given by

$$P\left(\frac{a}{b_1b_2b_3b_4b_5}\right) = \frac{n_{b_1b_2b_3b_4b_5a}}{\sum_{a=A,C,G,T} n_{b_1b_2b_3b_4b_5a}}$$

where, $n_{b_1b_2b_3b_4b_5a}$ is the number of times the sequence $b_1b_2b_3b_4b_5a$ occurs in the training data. This is the maximum likelihood estimators of the probability from the training data.

Glimmer

The core of Glimmer is Interpolated Markov Model (IMM), which can be described as a generalized Markov chain with variable order. After GeneMark introduces the fixed-order Markov chains, Glimmer attempts to find a better approach for modeling the genome content. The motivational fact is that the bigger the order of the Markov chain, the more non-randomness can be described. However, as we move to higher order models, the number of probabilities that we must estimate from the data increases exponentially. The major limitation of the fixed-order Markov chain is that models from higher order require exponentially more training data, which are limited and usually not available for new sequences. However, there are some oligomers from higher order that occur often enough to be extremely useful predictors. For the purpose of using these higher-order statistics, whenever sufficient data is available, Glimmer IMMs.

Glimmer calculates the probabilities for all Markov chains from 0th order to 8th. If there are longer sequences (e.g. 8-mers) occurring frequently, IMM makes use of them even when there is insufficient data to train an 8-th order model. Similarly, when the statistics from the 8-th order model do not provide significant information, Glimmer refers to the lower-order models to predict genes.

Opposed to the supervised GeneMark, Glimmer uses the input sequence for training. The ORFs longer than a certain threshold are detected and used for training, because there is high probability that they are genes in prokaryotes. Another training option is to use the sequences with homology to known genes from other organisms, available in public databases. Moreover, the user can decide whether to use long ORFs for training purposes or choose any set of genes to train and build the IMM.

GeneMark.hmm

GeneMark.hmm is designed to improve GeneMark in finding exact gene starts. Therefore, the properties of GeneMark.hmm are complementary to GeneMark. GeneMark.hmm uses GeneMark models of coding and non-coding regions and incorporates them into hidden Markov model framework. In short terms, Hidden Markov Models (HMM) are used to describe the transitions from non-coding to coding regions and vice versa. GeneMark.hmm predicts the most likely structure of the genome using the Viterbi algorithm, a dynamic programming algorithm for finding the most likely sequence of hidden states. To further improve the prediction of translation start position, GeneMark.hmm derives a model of the ribosome binding site (6-7 nucleotides preceding the start codon, which are bound by the ribosome when initiating protein translation). This model is used for refinement of the results.

Both GeneMark and GeneMark.hmm detect prokaryotic genes in terms of identifying open reading frames that contain real genes. Moreover, they both use pre-computed species-specific gene models as training data, in order to determine the parameters of the protein-coding and non-coding regions.

Orpheus

The ORPHEUS program uses homology, codon statistics and ribosome binding sites to improve the methods presented so far by using information that those programs ignored. One of the key differences is that it uses database searches to help determine putative genes, and is thus an extrinsic method. This initial set of genes is used to define the coding statistics for the organism, in this case working at the level of codon, not dicodons. These statistics are then used to define a larger set of candidate ORFs. From this set, those ORFs with an unambiguous start codon end are used to define a scoring matrix for the ribosome-binding site, which is then used to determine the 5` end of those ORFs where alternative start are present.

EcoParse

EcoParse is one of the first HMM model based gene finder, was developed for gene finding in *E.coli*. It focuses on the uses the codon structure of genes. With EcoParse a flora of HMM based gene finder, usuing dynamic programming and the viterbi algorithm to parse a sequence, emerged.

Evaluation of Gene Prediction Programs

In the field of gene prediction accuracy can be measured at three levels

a. Coding nucleotides (base level)

b. Exon structure (exon level)

c. Protein product (protein level)

At base level gene predictions can be evaluated in terms of *true positives (TP)* (predicted features that are real), *true negatives* (TN) (non-predicted features that are not real), *false positives (FP)* (predicted features that are not real), and *false negatives (FN)* (real features that were not predicted) Fig. 5. Usually the base assignment is to be in a coding or non coding segment, but this analysis can be extended to include non coding parts of genes, or any functional parts of the sequences.

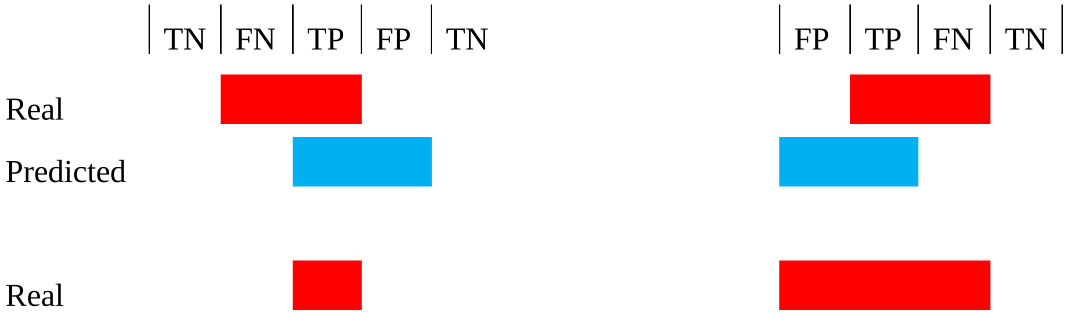


Fig. 5: Four Possible Comparisons of Real and Predicted Genes

Sensitivity (Sn): The fraction of bases in real genes that are correctly predicted to be in genes is the sensitivity and interpreted as the probability of correctly predicting a nucleotide to be in a given gene that it actually is.

$$Sn = \frac{TP}{TP + FN}$$

Specificity (Sp): The fraction of those bases which are predicted to be in genes that actually are is called the specificity and interpreted as the probability of a nucleotide actually being in a gene given that it has been predicted to be.

$$Sp = \frac{TP}{TP + FP}$$

Care has to be taken in using these two values to assess a gene prediction program because, as with the normal definition of specificity, extreme results can make them misleading.

Approximate correlation coefficient (AC) has been proposed as a single measure to circumvent these difficulties. This defined as AC=2(ACP-0.5), where

$$ACP = \frac{1}{n}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right),$$

At the exon level, determination of prediction accuracy depends on the exact prediction of exon start and end points. There are two measures of sensitivity and specificity used in the field, each of which measures a different but useful property.

The sensitivity measures used are

S_{n1} = CE/AE and Sn2 = ME/AE

The specificity measures used are

S_{p1}=CE/PE and S_{p2}=WE/PE

Where,

AE = No of actual exons in the data

PE = No of predicted exons in the data

CE = No of correct predicted exons

ME = No of missing exons (rarely occurs)

WE = No of wrongly predicted exons (Figure-5)

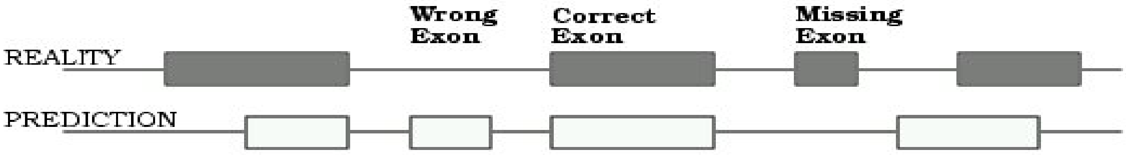


Fig. 6: Real and Predicted Exons

Gene Ontology

The gene ontology (GO, http:www.geneontology.org) is probably the most extensive scheme today for the description of gene product functions but other systems such as enzyme codes, KEGG pathways, FunCat, or COG are also widely used. Here, we describe the Blast2GO (B2G, www.blast2go.org) application for the functional annotation, management, and data mining of novel sequence data through the use of common controlled vocabulary schemas. The main application domain of the tool is the functional genomics of non-model organisms and it is primarily intended to support research in experimental labs. Blast2GO strives to be the application of choice for the annotation of novel sequences in functional genomics projects where thousands of fragments need to be characterized. Functional annotation in Blast2GO is based on homology transfer. Within this framework, the actual annotation procedure is configurable and permits the design of different annotation strategies. Blast2GO annotation parameters include the choice of search database, the strength and number of blast results, the extension of the query-hit match, the quality of the transferred annotations, and the inclusion of motif annotation. Vocabularies supported by B2G are gene ontology terms, enzyme codes (EC), InterPro IDs, and KEGG pathways.

Fig.7 shows the basic components of the Blast2GO suite. Functional assignments proceed through an elaborate annotation procedure that comprises a central strategy plus refinement functions. Next, visualization and data mining engines permit exploiting the annotation results to gain functional knowledge. GO annotations are generated through a 3-step process: blast, mapping, annotation. InterPro terms are obtained from InterProScan at EBI, converted and merged to GOs. GO annotation can be modulated from Annex, GOSlim web services and manual editing. EC and KEGG annotations are generated from GO. Visual tools include sequence color code, KEGG pathways, and GO graphs with node highlighting and filtering options. Additional annotation data-mining tools include statistical charts and gene set enrichment analysis functions.

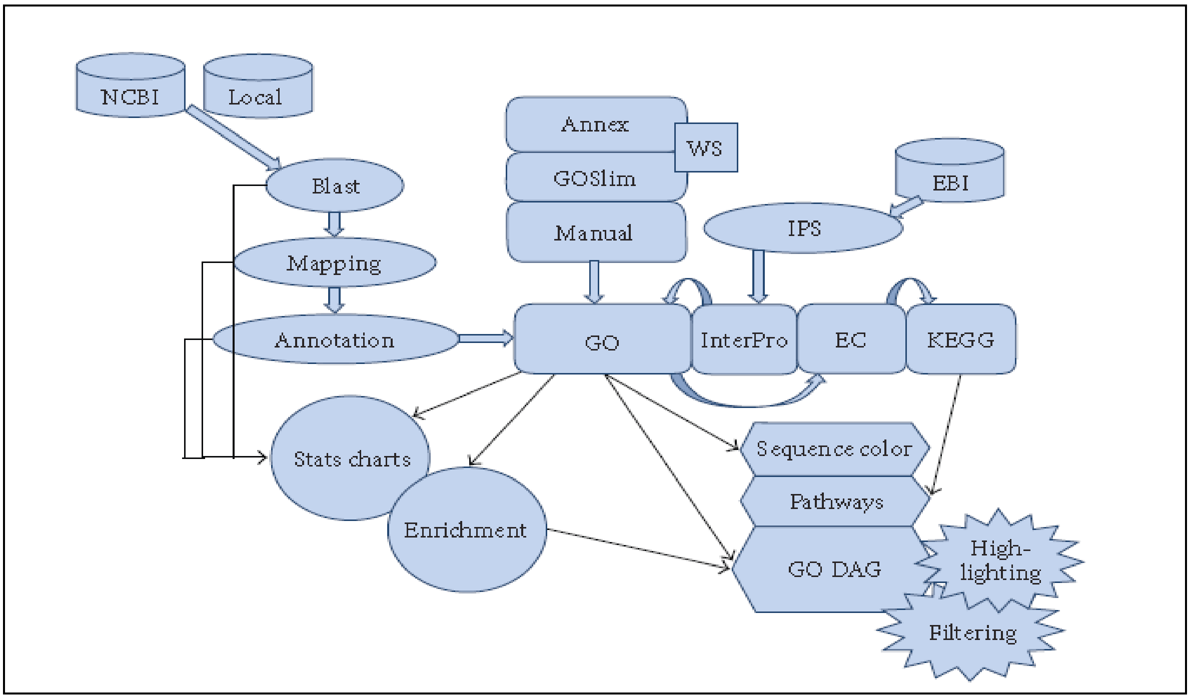


Fig. 7: Schematic Representation of Blast2GO Application.

The Blast2GO annotation procedure consists of three main steps: blast to find homologous sequences, mapping to collect GO terms associated to blast hits, and annotation to assign trustworthy information to query sequences.

Blast Step

The first step in B2G is to find sequences similar to a query set by blast. B2G accepts nucleotide and protein sequences in FASTA format and supports the four basic blast programs (blastx, blastp, blastn, and tblastx). Homology searches can be launched against public databases such as (the) NCBI nr using a query-friendly version of blast (QBlast). This is the default option and in this case, no additional installations are needed. Alternatively, blast can be run locally against a proprietary FASTA-formatted database, which requires a working www-blast installation. The Make Filtered Blast-GO-BD function in the Tools menu allows the creation of customized databases containing only GO annotated entries, which can be used in combination with the local blast option. Other configurable parameters at the blast step are the expectation value (e-value) threshold, the number of retrieved hits, and the minimal alignment length (hsp length) which permits the exclusion of hits with short, low e-value matches from the sources of functional terms. Annotation, however, will ultimately be based on sequence similarity levels as similarity percentages are independent of database size and more intuitive than e-values. Blast2GO parses blast results and presents the information for each sequence in table format. Query sequence descriptions are obtained by applying a language processing algorithm to hit descriptions, which extracts informative names and avoids low content terms such as “hypothetical protein” or “expressed protein”.

Mapping Step

Mapping is the process of retrieving GO terms associated to the hits obtained after a blast search. B2G performs three different mappings as follows.

a. Blast result accessions are used to retrieve gene names (symbols) making use of two mapping files provided by NCBI (geneinfo, gene2accession). Identified gene names are searched in the species-specific entries of the gene product table of the GO database.

b. Blast result GI identifiers are used to retrieve UniProt IDs making use of a mapping file from PIR (Non-redundant Reference Protein database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB.

c. Blast result accessions are searched directly in the DBXRef Table of the GO database.

Annotation Step

This is the process of assigning functional terms to query sequences from the pool of GO terms gathered in the mapping step. Function assignment is based on the gene ontology vocabulary. Mapping from GO terms to enzyme codes permits the subsequent recovery of enzyme codes and KEGG pathway annotations. The B2G annotation algorithm takes into consideration the similarity between query and hit sequences, the quality of the source of GO assignments, and the structure of the GO DAG. For each query sequence and each candidate GO term, an annotation score (AS) is computed (see Figure 8). The AS is composed of two terms. The first, direct term (DT), represents the highest similarity value among the hit sequences bearing this GO term, weighted by a factor corresponding to its evidence code (EC). A GO term EC is present for every annotation in the GO database to indicate the procedure of functional assignment.

$$\begin{aligned}
&\mathrm{DT} = \max\,(\mathrm{similarity} \times \mathrm{EC}_{\mathrm{weight}})\\
&\mathrm{AT} = (\#\mathrm{GO} - 1) \times \mathrm{GO}_{\mathrm{weight}}\\
&\mathrm{AR} : \mathrm{lowest.node}(\mathrm{AS}(\mathrm{DT} + \mathrm{AT}) \geq \mathrm{threshold})
\end{aligned}$$

Fig. 8: Blast2GO Annotation Rule

ECs vary from experimental evidence, such as inferred by direct assay (IDA) to unsupervised assignments such as inferred by electronic annotation (IEA). The second term (AT) of the annotation rule introduces the possibility of abstraction into the annotation algorithm. Abstraction is defined as the annotation to a parent node when several child nodes are present in the GO candidate pool. This term multiplies the number of total GOs unified at the node by a user defined factor or GO weight (GOw) that controls the possibility and strength of abstraction. When all ECw's are set to 1 (no EC control) and the GOw is set to 0 (no abstraction is possible), the annotation score of a given GO term equals the highest similarity value among the blast hits annotated with that term. If the ECw is smaller than one, the DT decreases and higher query-hit similarities are required to surpass the annotation threshold. If the GOw is not equal to zero, the AT becomes contributing and the annotation of a parent node is possible if multiple child nodes coexist that do not reach the annotation cutoff. Default values of B2G annotation parameters were chosen to optimize the ratio between annotation coverage and annotation accuracy. Finally, the AR selects the lowest terms per branch that exceed a user-defined threshold.

Blast2GO includes different functionalities to complete and modify the annotations obtained through the above-defined procedure. Enzyme codes and KEGG pathway annotations are generated from the direct mapping of GO terms to their enzyme code equivalents. Additionally, Blast2GO offers InterPro searches directly from the B2G interface. B2G launches sequence queries in batch, and recovers, parses, and uploads InterPro results. Furthermore, InterPro IDs can be mapped to GO terms and merged with blast-derived GO annotations to provide one integrated annotation result. In this process, B2G ensures that only the lowest term per branch remains in the final annotation set, removing possible parent-child relationships originating from the merging action.

References

25. Conesa, S. Gotz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles, “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research,” Bioinformatics, vol. 21, no. 18, pp. 3674–3676, 2005.
26. Conesa and S. Gotz, "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics," International Journal of Plant Genomics, vol. 2008, 2008.
27. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” Nucleic Acids Research, vol. 27, no. 1, pp. 29–34, 1999.
28. J.D. Watson, R.M. Myers, A.A. Caudy and J.A. Witkowski, "Recombinant DNA: Genes and Genomes - A Short Course," 3rd Ed., 2007.
29. M. Ashburner, C. A. Ball, J. A. Blake, et al., “Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium,” Nature Genetics, vol. 25, no. 1, pp. 25–29, 2000.

30. Ruepp, A. Zollner, D. Maier, et al., “The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes,” Nucleic Acids Research, vol. 32, no. 18, pp. 5539–5545, 2004.

31. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, et al., “The COG database: an updated version includes eukaryotes,” BMC Bioinformatics, vol. 4, p. 41, 2003.

32. Schomburg, A. Chang, C. Ebeling, et al., “BRENDA, the enzyme database: updates and major new developments,” Nucleic Acids Research, vol. 32, Database issue, pp. D431–D433, 2004.

33. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” Journal of Molecular Biology, vol. 215, no. 3, pp. 403–410, 1990.

34. S. Myhre, H. Tveit, T. Mollestad, and A. Lægreid, “Additional Gene Ontology structure for improved biological reasoning,” Bioinformatics, vol. 22, no. 16, pp. 2020–2027, 2006.

Differential Gene Expression Analysis

Online Training Programme on

“RNA World: Advance Bioinformatics for deciphering regulatory molecules”

under the project ICAR Consortium Research Platform on Genomics

funded by Indian Council of Agricultural Research

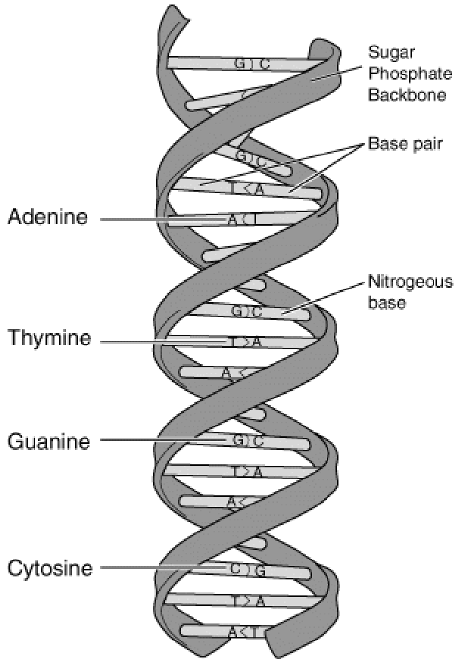
November 3-9, 2022

Sudhir Srivastava
Division of Agricultural Bioinformatics
ICAR-Indian Agricultural Statistics Research Institute

Introduction

DNA

- double stranded, helical structure
- sequences of nucleotides (A, T, G & C)
- base pairs (A with T and G with C)

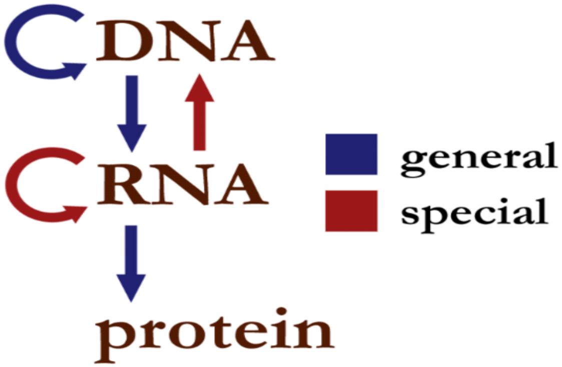


Introduction…

Central Dogma of Molecular Biology

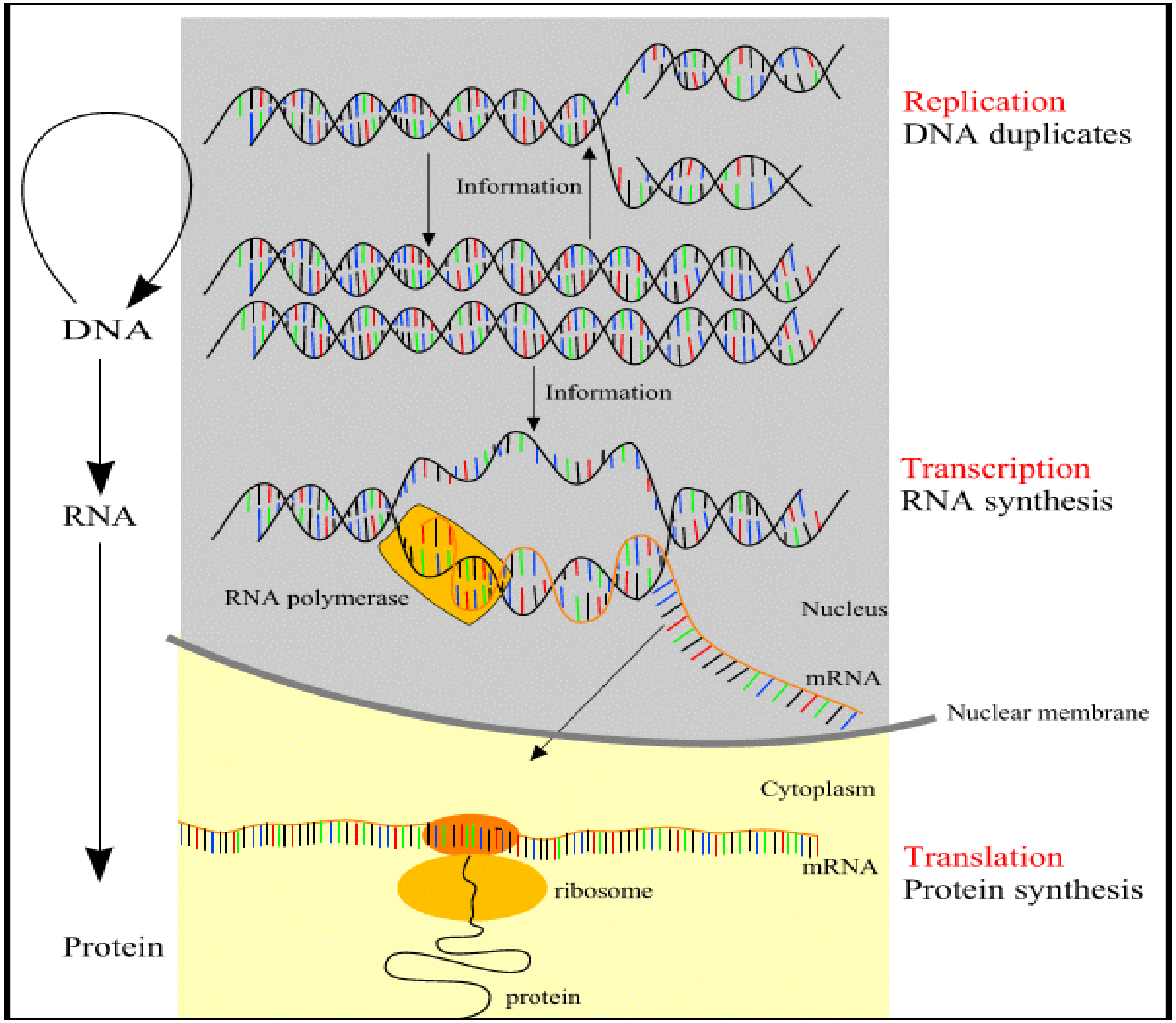
The Central Dogma. This states that once ‘information’ has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein. [Francis Crick,1958]

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid. [Francis Crick, re-stated in a Nature paper, 1970]



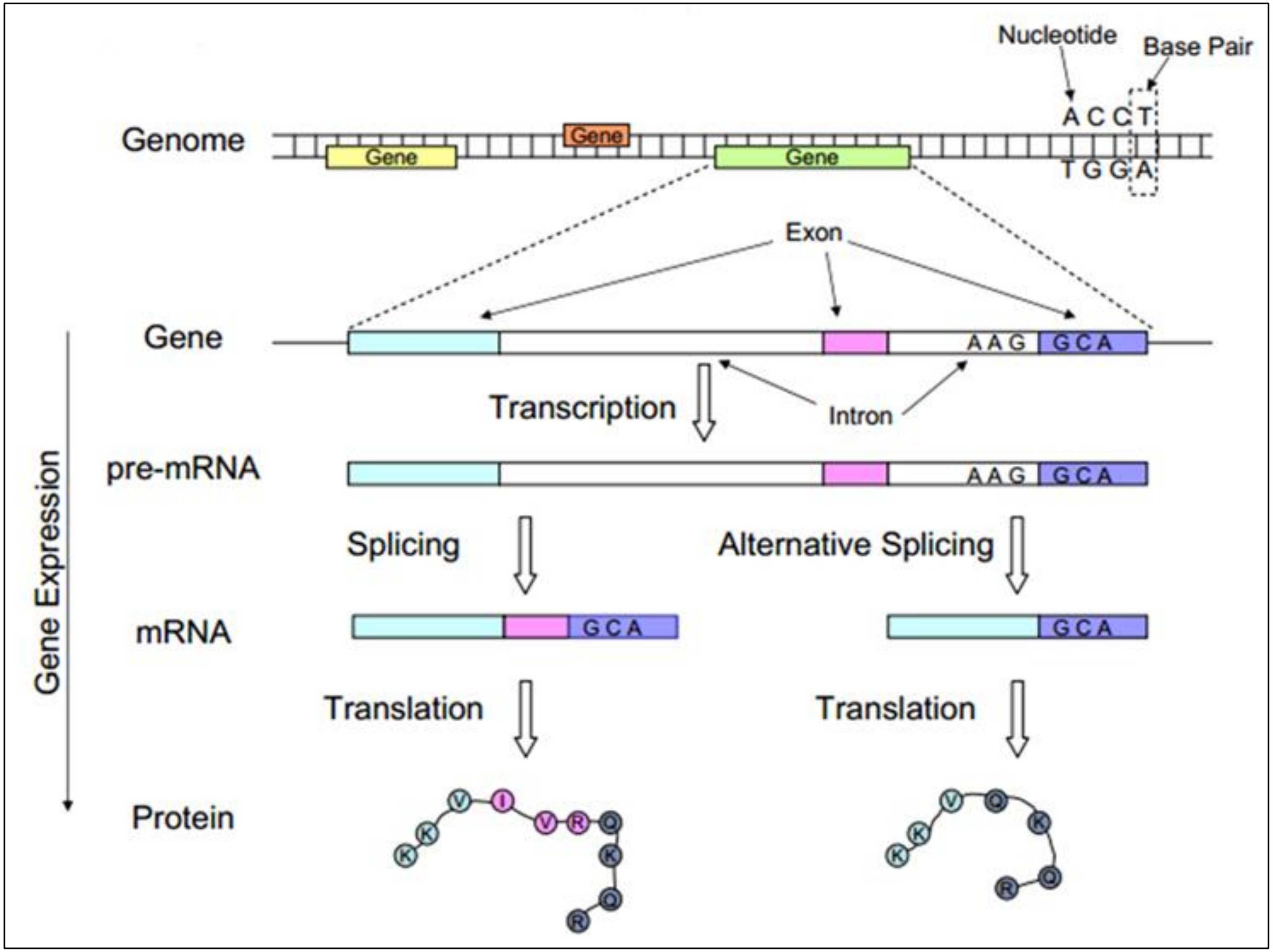
Introduction…

Central Dogma of Molecular Biology



Introduction…

Central Dogma of Molecular Biology

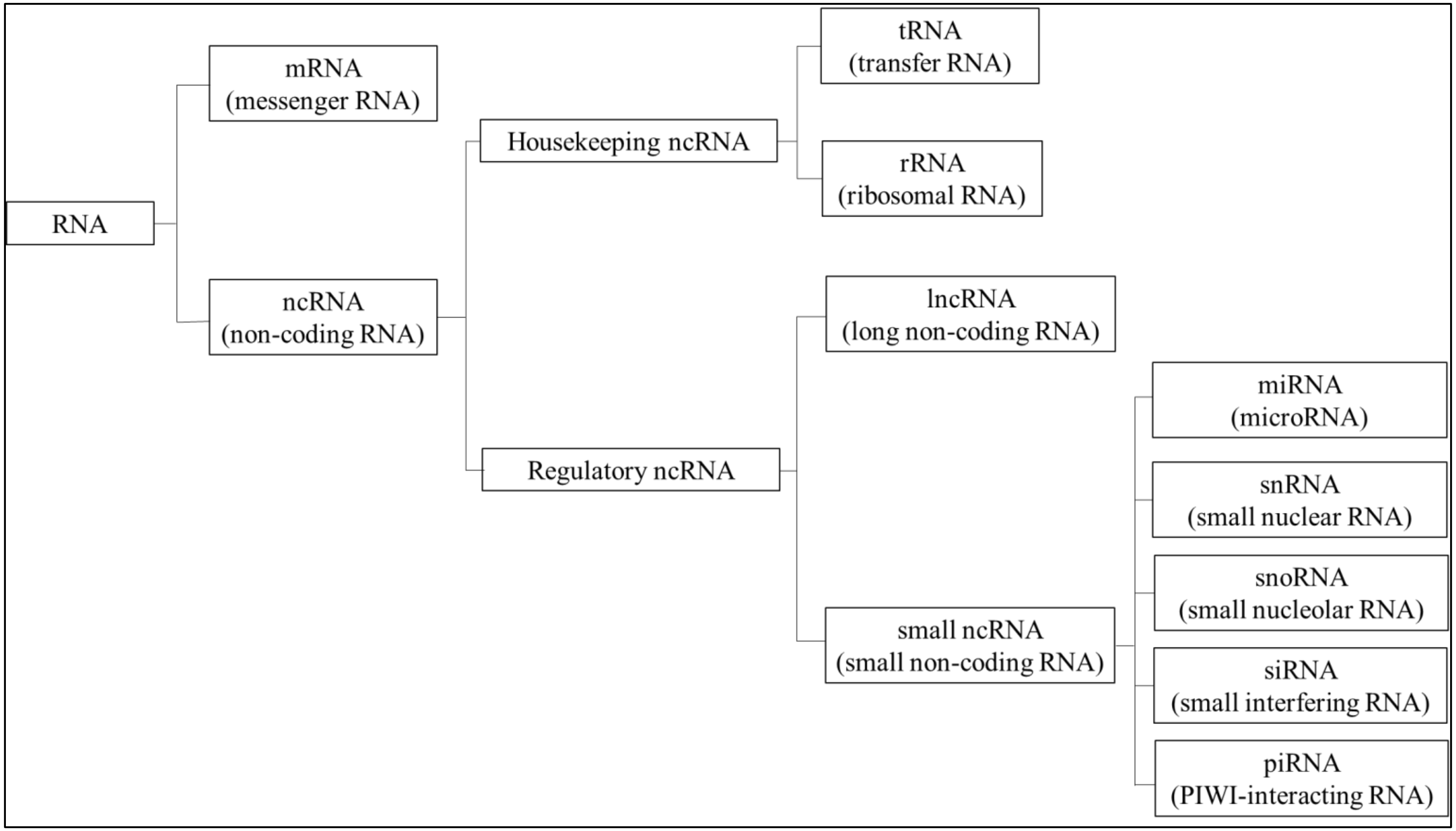


Introduction…

- The advent of Next-Generation Sequencing (NGS) technology has transformed genomic studies.
- One important application of NGS technology is the study of the transcriptome.
- Transcriptome is defined as the complete collection of all the RNA molecules in a cell.

Introduction…

Different types of RNA


RNA
mRNA (messenger RNA)
ncRNA (non-coding RNA)
Housekeeping ncRNA
Regulatory ncRNA
tRNA (transfer RNA)
rRNA (ribosomal RNA)
lncRNA (long non-coding RNA)
small ncRNA (small non-coding RNA)
miRNA (microRNA)
snRNA (small nuclear RNA)
snoRNA (small nucleolar RNA)
siRNA (small interfering RNA)
piRNA (PIWI-interacting RNA)

- All of these molecules are called transcripts since they are produced by process of transcription.
- ~ 2% mRNA

Introduction…

- RNA-Sequencing uses NGS technology to reveal the presence and quantity of RNA in a biological sample at a given moment.
- It allows transcript quantification and differential gene expression analysis.
- Several machines/ protocols are available for generating RNA-Seq data:
  - Illumina (MiSeq, NextSeq, HiSeq, NovaSeq)
  - Ion Torrent (Proton, Personal Genome Machine)
  - SOLiD
  - Roche 454

Introduction…

- Important steps of RNA-Seq experiments:
  - Data generation (experimental design, sample collection, sequencing design, quality control)
  - Quantification of reads to estimate the expression values
  - Normalization
  - Differential expression analysis

Introduction…

- **Applications of RNA-Seq experiments**
  - Quantification of transcriptome/RNA-Seq expression levels to study gene expression in complex experiments
  - Novel gene discovery
  - Gene annotation
  - Detection of differentially expressed features (genes/ transcripts/ exons) between different conditions
  - Detection of splicing events
  - Identification of introns and exon boundaries

Bioinformatics Tools for NGS data preprocessing

Tools for quality check/ filtering/ trimming

- **FASTQC** - A quality control tool for high throughput sequence data

  (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

- **NGS QC** - Quality Control

- **FastqCleaner** – A shiny app for Quality Control, Filtering and Trimming of FASTQ Files

- **Trimmomatic** – Trimming of FASTQ files

Bioinformatics Tools for NGS data preprocessing…

- **FASTX toolkit** – A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing

  (http://hannonlab.cshl.edu/fastx_toolkit/)

- **ShortRead** – R package for filtering and trimming reads, and for generating a quality assessment report

Bioinformatics Tools for NGS data preprocessing…

Samtools: A suite of programs for interacting with high-throughput sequencing data (http://www.htslib.org/)

Three separate repositories:

- Samtools - Reading/writing/editing/indexing/viewing SAM/BAM format
- BCFtools - Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
- HTSlib - A C library for reading/writing high-throughput sequencing data

Bioinformatics Tools for NGS data preprocessing…

Short read aligners

- Bowtie
- TOPHAT
- BWA
- Novoalign
- STAR

Bioinformatics Tools for NGS data preprocessing…

de novo assemblers

- SOAPdenovo-Trans
- Trans-AbySS
- Trinity
- SPAdes

Tools for Visualization

- CummeRbund
- IGV
- Bedtools
- UCSC Genome Browser

Experimental design and heterogeneity issues

- The purpose of experimental design is to plan experiment in an effective way so that it can answer the biological question under consideration.

  (i) Biological aspects:

  - Any biological experimental plan starts with a biological question or hypothesis.
  - The experimenter might have some prior knowledge of the question under study before conducting the experiments, e.g., expression levels of some known genes, proteins, etc.

  (ii) Technical aspects:

  - These include the choice of platform and avoiding systematic errors.
  - If the experiment has systematics errors, then the result obtained for comparative analysis will be biased, irrespective of the precision of measurement and the number of experimental units.

  (iii) Economic aspects:

  - Cost of experiment and its analysis
  - Budget available
  - Time required to complete the experiment and its analysis
  - Whether pilot study is required or not, etc.

Experimental design and heterogeneity issues…

Other points to be considered:

- Availability of enough samples for experiment;
- Availability of enough RNA, DNA or proteins from samples;
- Biopsies collected from same part of tissue or other tissues;
- Number of replicates required;
- Effect size, *etc*.

Heterogeneity

- A heterogeneous sample or population means that every observed data has different value for the corresponding characteristic of interest.
- There may be various factors responsible for influencing expression in any feature.
- The major sources of variations are due to technical, genetic, demographic and environmental factors.

Experimental design and heterogeneity issues…

- There are two important points to be considered while designing RNA-Seq experiments which are namely, the sequencing depth and the number of replicates (biological and technical) required to observe significant changes in expression.
- The cost can be reduced by optimizing the designing process of these experiments.
- Tools and software for sample size estimation and power analysis:
    - RNASeqPowerCalculator
    - RNASeqPower
    - Scotty
    - PROPER

RNA-Seq Experiments

- The basic steps for summarizing a typical RNA-Seq experiment:
  - Purified RNA is converted to cDNA, fractionated, ligated with technology specific adapters and sequencing is done.
  - Millions of short read sequences are generated from one end (single-end) or both ends (paired-end) of the cDNA fragments.
  - These sequences are aligned to a reference genome.
  - The number of reads mapped to known features are recorded and summarized in a table.
- The features can be either genes, transcripts (alternative transcripts) or exon level expression.

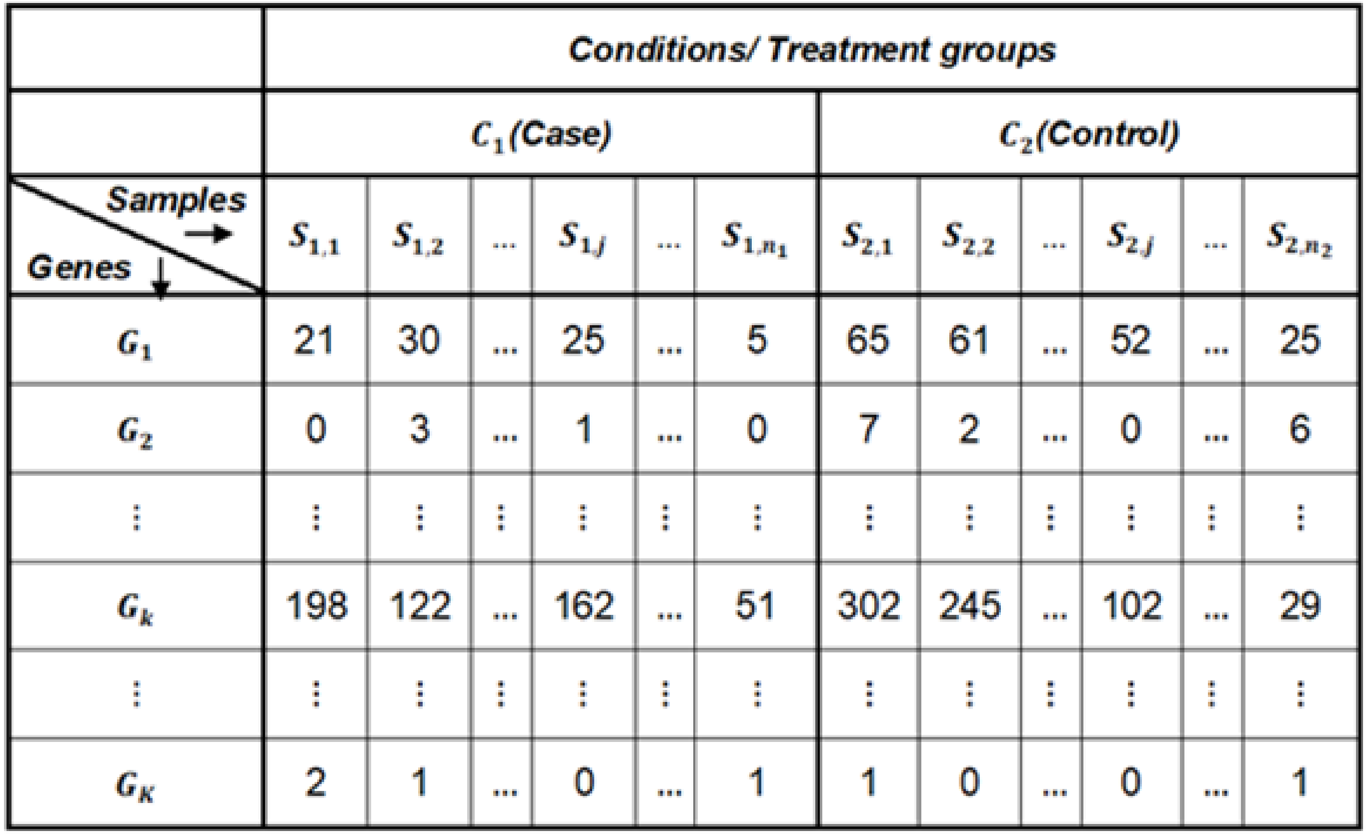
RNA-Seq Experiments…

Example of a biological experiment with I conditions/groups denoted by $G_i (i = 1, 2, \ldots, I)$ having N_i individuals/samples denoted by $S_{i,j}$ $(j =$

	$\boldsymbol{G_1}$					…	$\boldsymbol{G_i}$					…	$\boldsymbol{G_I}$				
	$\boldsymbol{S_{1,1}}$	…	$\boldsymbol{S_{1,j}}$	…	$\boldsymbol{S_{1,N_1}}$		$\boldsymbol{S_{i,1}}$	…	$\boldsymbol{S_{i,j}}$	…	$\boldsymbol{S_{i,N_i}}$		$\boldsymbol{S_{I,1}}$	…	$\boldsymbol{S_{I,j}}$	…	$\boldsymbol{S_{I,N_I}}$
$\boldsymbol{F_1}$	$y_{1,1,1}$	…	$y_{1,j,1}$	…	$y_{1,N_1,1}$		$y_{i,1,1}$	…	$y_{i,j,1}$	…	$y_{i,N_i,1}$		$y_{I,1,1}$	…	$y_{I,j,1}$	…	$y_{I,N_I,1}$
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮
$\boldsymbol{F_k}$	$y_{1,1,k}$	…	$y_{1,j,k}$	…	$y_{1,N_1,k}$		$y_{i,1,k}$	…	$\boldsymbol{y_{i,j,k}}$	…	$y_{i,N_i,k}$		$y_{I,1,k}$	…	$y_{I,j,k}$	…	$y_{I,N_I,k}$
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮
$\boldsymbol{F_K}$	$y_{1,1,K}$	…	$y_{1,j,K}$	…	$y_{1,N_1,K}$		$y_{i,1,K}$	…	$y_{i,j,K}$	…	$y_{i,N_i,K}$		$y_{I,1,K}$	…	$y_{I,j,K}$	…	$y_{I,N_I,K}$

RNA-Seq Experiments…

A table of read counts for a hypothetical case-control study



	Conditions/ Treatment groups											
	C_1(Case)						C_2(Control)					
Samples → Genes ↓	$S_{1,1}$	$S_{1,2}$	...	$S_{1,j}$	...	S_{1,n_1}	$S_{2,1}$	$S_{2,2}$	...	$S_{2,j}$	...	S_{2,n_2}
G_1	21	30	...	25	...	5	65	61	...	52	...	25
G_2	0	3	...	1	...	0	7	2	...	0	...	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_k	198	122	...	162	...	51	302	245	...	102	...	29
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_K	2	1	...	0	...	1	1	0	...	0	...	1

Transcript quantification

- The most common application of RNA-seq is to estimate gene and transcript expression.
- This application is primarily based on the number of reads that map to each transcript sequence.
- The simplest approach to quantification is to aggregate raw counts of mapped reads using programs such as HTSeq-count or featureCounts.
- Metrics to normalize data considering the gene length and sequencing depth
  - RPKM (reads aligned per kilobase of exon per million reads mapped)
  - FPKM (fragments per kilobase of exon per million fragments mapped)
  - TPM (transcripts per kilobase million)
- Normalization is required before performing the differential expression analysis.

Transcript quantification…

- htseq-count
- featureCounts
- Cufflinks
- Stringtie
- RSEM
- Sailfish

Differential Expression Analysis

- One of the primary goals for RNA-seq experiments is to compare the gene expression levels across various experimental conditions, treatments, tissues, or time points.
- The researchers are particularly interested in detecting gene with differential expressions.
- The study of determining which genes have changed significantly in terms of their expression across two or more conditions is referred to as differential expression analysis.
- Identification of differentially expressed genes helps researchers to understand the functions of genes in response to a given condition.

Differential Expression Analysis…

- A large number of statistical models and tools have been developed to perform differential expression analysis for RNA-Seq data.
- Differential expression analysis methods for RNA-Seq can be grouped into two broad categories:

- **Parametric method**
  - It captures all information about the data within the parameters.
  - Each expression value for a given gene is mapped into a particular distribution, such as Poisson or negative binomial.
- **Non-parametric method**
  - A non-parametric model uses a flexible number of parameters.
  - The number of parameters often grows as it learns from more data.
  - A non-parametric model is computationally slower, but makes fewer assumptions about the data.

RNA-Seq Experiments…

Estimation of parameters based on NB distribution

- The estimation of parameters is an essential step for design, sample size calculation and differential expression analysis.

- The parameter estimation can be done by using various methods such as method of moments estimation (MME), maximum likelihood estimation (MLE), maximum quasi-likelihood estimation (MQLE).

- Besides these methods, there are various methods/models for estimation of parameters such as pseudo-likelihood, quasi-likelihood, conditional maximum likelihood (CML), conditional inference, quantile-adjusted CML, conditional weighted likelihood.

RNA-Seq Experiments…

Estimation of parameters based on NB distribution without scaling factor

- Let Y_{ij} be a NB random variable with mean μ_i and dispersion parameter ϕ_i, i.e., $Y_{ij} \sim NB(\mu_i, \phi_i)$, then its probability mass function is given by

$$p\left(Y_{ij} = y_{ij}\right) = \frac{\Gamma\left(y_{ij} + \frac{1}{\phi_i}\right)}{\Gamma\left(\frac{1}{\phi_i}\right)\Gamma\left(y_{ij} + 1\right)} \frac{(\mu_i \phi_i)^{y_{ij}}}{(1 + \mu_i \phi_i)^{y_{ij} + \frac{1}{\phi_i}}} ; y = 0, 1, 2, \dots$$

- The likelihood function is given by

$$L(\mu_i, \phi_i | y_{ij};\ j = 1, 2, \dots, N_i) = \prod_{j=1}^{N_i} \frac{\Gamma\left(y_{ij} + \frac{1}{\phi_i}\right)}{\Gamma\left(\frac{1}{\phi_i}\right)\Gamma\left(y_{ij} + 1\right)} \frac{(\mu_i \phi_i)^{y_{ij}}}{(1 + \mu_i \phi_i)^{y_{ij} + \frac{1}{\phi_i}}}$$

- The log-likelihood function is given by

$$\begin{aligned} &l(\mu_i, \phi_i | y_{ij};\ j = 1, 2, \dots, N_i) \\ &= \sum_{j=1}^{N_i} \ln \Gamma\left(y_{ij} + \frac{1}{\phi_i}\right) - \sum_{j=1}^{N_i} \Gamma\left(\frac{1}{\phi_i}\right) - \sum_{j=1}^{N_i} \ln \Gamma\left(y_{ij} + 1\right) \\ &+ \sum_{j=1}^{N_i} y_{ij} \ln(\mu_i \phi_i) - \sum_{j=1}^{N_i} \left(y_{ij} + \frac{1}{\phi_i}\right) \ln(1 + \mu_i \phi_i) \end{aligned}$$

Differential Expression Analysis…

Method	Read count distribution assumption/model	Normalization	Differential analysis test
edgeR	Negative binomial distribution	TMM/ Upper quartile / RLE / None (all scaling factors are set to be one)	Exact test analogous to Fisher's exact test or likelihood ratio test
DESeq	Negative binomial distribution	DESeq size factors	Exact test analogous to Fisher's exact test
DESeq2	Negative binomial distribution	DESeq size factors	Wald test
baySeq	Negative binomial distribution	Scaling factors (quantile/ TMM/ total)	Posterior probability through Bayesian approach
EBSeq	Negative binomial-beta empirical Bayes model	DESeq median normalization	
SAMseq	Non-parametric method	Based on the read count mean over the null features of data set.	Wilcoxon rank statistics based permutation test
NOIseq	Non-parametric method	RPKM / TMM / Upper quartile	Corresponding logarithm of fold change and absolute expression differences have a high probability than noise values
limma+voom	Similar to t-distribution with empirical Bayes approach	TMM	Moderated t-test

Differential Expression Analysis…

Tools for Differential Expression Analysis

- Cufflinks package
- R packages: DESeq, DESeq2, edgeR

edgeR for RNA-Seq Data Analysis

1. Download and Install R

https://cran.r-project.org/bin/windows/base/

2. Download and Install RStudio

https://www.rstudio.com/products/rstudio/download/#download

3. Open RStudio

4. Install the required R packages: Here, we will install edgeR.

```
if (!requireNamespace("BiocManager", quietly = TRUE))

    install.packages("BiocManager")

BiocManager::install("edgeR")
```

edgeR for RNA-Seq Data Analysis…

https://bioconductor.org/packages/release/bioc/html/edgeR.html

Example: A paired design RNA-seq experiment of oral squamous cell carcinomas and matched normal tissue from three patients

- The aim of the analysis is to detect genes differentially expressed between tumor and normal tissue, adjusting for any differences between the patients.
- RNA was sequenced on an Applied Biosystems SOLiD System 3.0 and reads mapped to the UCSC hg18 reference genome.
- Read counts, summarised at the level of refSeq transcripts are available in Table S1 of Tuch *et al*.
  (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824832/).

Online Tool for RNA-Seq Data Analysis

http://bioinformatics.sdstate.edu/idep/

https://kcvi.shinyapps.io/START/

References

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, **17**, 13. https://doi.org/10.1186/s13059-016-0881-8

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, **12(12)**, e0190152. https://doi.org/10.1371/journal.pone.0190152

Li D. (2019). Statistical Methods for RNA Sequencing Data Analysis. In: Husi H, editor. Computational Biology [Internet]. Brisbane (AU): Codon Publications; Chapter 6. Available from: https://www.ncbi.nlm.nih.gov/books/NBK550334/; doi:10.15586/computationalbiology.2019.ch6

References…

Ge, S.X., Son, E.W. & Yao, R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics 19, 534 (2018). https://doi.org/10.1186/s12859-018-2486-6

Nelson, JW, Sklenar J, Barnes AP, Minnier J. (2016) “The START App: A Web-Based RNAseq Analysis and Visualization Resource.” Bioinformatics. doi: 10.1093/bioinformatics/btw624.

McCarthy DJ, Chen Y, Smyth GK (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” Nucleic Acids Research, 40(10), 4288-4297. doi: 10.1093/nar/gks042.

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” Bioinformatics, 26(1), 139-140. doi: 10.1093/bioinformatics/btp616.

THANKS

E-mail: Sudhir.Srivastava@icar.gov.in, sudhir0401bm@gmail.com

Chapter 6

Transcriptome Data Annotation

Sneha Murmu

Division of Agricultural Bioinformatics,

ICAR-Indian Agricultural Statistics Research Institute

The current ecosystems of RNA-seq tools provide a varied ways analyzing RNA-seq data. Depending on the experiment goal one could align the reads to reference genome or pseduoalign to transcriptome and perform quantification and differential expression of genes or if you want to annotate your reference, assemble RNA-seq reads using a *de novo* transcriptome assembler. In this lecture, we focus on workflows that align reads to reference genomes using updated Tuxedo protocol (HISAT, StringTie, Ballgown) by Pertea et al. This updated Tuxedo protocol not only scales but is more accurate in detecting differentially expressed genes (DEGs). Lastly, we used Blast2GO for annotating the identified DEGs.

In this example, we have used the example data which is mentioned in the paper. Before starting with the actual workflow, we have briefly mentioned the steps required to set up the system.

1) Setting up the system for differential expression analysis of transcriptome data

```
#for windows system, install linux via wsl.

#install anaconda in linux (Ubuntu)

#open ubuntu terminal

$ wget https://repo.anaconda.com/archive/Anaconda3-2022.10-Linux-x86_64.sh

$ bash Anaconda3-2022.10-Linux-x86_64.sh

#set up the conda environment

$ conda env create -f environment_1.yaml

$ conda activate rnaseq_py3

# Set up complete!
```

1. **Protocol:**

```
###Align the data to the reference genome using HISAT2

##build index

(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# mkdir index
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# extract_splice_sites.py resources/chrX.gtf > index/chrX.ss
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# extract_exons.py resources/chrX.gtf > index/chrX.exon
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example# cd index
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example/index# hisat2-build -p 8 --ss chrX.ss --exon chrX.exon ../resources/chrX.fa chrX_tran
```

```
(rnaseq_py3) root@DESKTOP-BJ5B6HR:/mnt/e/iasri/dabin_training/Nov2022/practical/example/index# cd ..
```

##1. mapping

```
$ fastqdir=resources/samples

mapdir=mapped

mkdir $mapdir

hisat2 -p 8 --dta -x index/chrX_tran -1 $fastqdir/ERR188044_chrX_1.fastq.gz -2 $fastqdir/ERR188044_chrX_2.fastq.gz -S $mapdir/ERR188044.sam
```

##2. sort mapped files

```
$ mapdir=mapped

samtools sort -@ 8 -o $mapdir/ERR188044.bam $mapdir/ERR188044.sam
```

##3. assembly

```
gtf=resources/chrX.gtf

assembly=assembly

mapdir=mapped

mkdir $assembly

stringtie $mapdir/ERR188044.bam -l ERR188044 -p 8 -G $gtf -o $assembly/ERR188044.gtf
```

##obtain list of each sample .gtf file in a single file (mergelist.txt)

```
$ ls assembly/*.gtf > mergelist.txt
```

##merge .gtf file of each sample

```
$ stringtie --merge -p 8 -G resources/chrX.gtf -o stringtie_merged.gtf mergelist.txt
```

##obtain sequences of transcripts

```
$ gffread -w transcripts.fa -g resources/chrX.fa stringtie_merged.gtf
```

##compare merged.gtf file with reference .gtf file

```
$ gffcompare -r resources/chrX.gtf -G -o merged stringtie_merged.gtf
```

##4. abundance estimation

```
$ abundancedir=abundance

mapdir=mapped

stringtie -e -B -p 8 -G stringtie_merged.gtf -o
$abundancedir/ERR188044/ERR188044_chrX.gtf $mapdir/ERR188044.bam
```

2. Differential expression analysis

Open R console.

```
#Differential expression

#load the libraries

library(ggplot2)

library(ballgown)

library(genefilter)

library(RSkittleBrewer)

library(devtools)

library(dplyr)

library(ggrepel)

library(pheatmap)

library(gplots)

library(GenomicRanges)
library(viridis)

#lets load the sample information

pheno_data <- read.csv("resources/geuvadis_phenodata.csv")

#let's show information for first 6 samples

head(pheno_data)

#Load the expression data using ballgown

bg_chrX <- ballgown(dataDir="abundance",samplePattern="ERR",pData=pheno_data)

#Lets filter out transcripts with low variance

#This is done to remove some genes that have few counts. Filtering improves the statistical power of differential expression analysis.

#We use variance filter to remove transcripts with low variance( 1 or less)

bg_chrX_filt<- subset(bg_chrX,"rowVars(texpr(bg_chrX))>1",genomesubset=TRUE)

#Let's test on transcripts

de_transcripts <- stattest(bg_chrX_filt,feature="transcript",covariate="conditions",getFC=TRUE,meas="FPKM")

# the results_transcripts does not contain identifiers. We will therefore add this information

#add identifiers

de_transcripts = data.frame(geneNames=ballgown::geneNames(bg_chrX_filt), geneIDs=ballgown::geneIDs(bg_chrX_filt), de_transcripts)

# Let's test on genes

de_genes <- stattest(bg_chrX_filt,feature="gene",covariate="conditions",getFC=TRUE, meas="FPKM")

#lets get the gene names

bg_filt_table=texpr(bg_chrX_filt,'all')

gene_names=unique(bg_filt_table[,9:10])

features=de_genes$id

mapped_gene_names=vector()

for (i in features)
```

```
{  query=gene_names%>%filter(gene_id==i & gene_name != '.') ; n_hit=dim(query)[1]; if
(n_hit==1) {mapped_gene_names=append(mapped_gene_names,query$gene_name[[1]]) }
else

{mapped_gene_names=append(mapped_gene_names,'.') }

}

#add the mapped gene names to the de genes table

de_genes$gene_name <- mapped_gene_names

de_genes <- de_genes[, c('feature','gene_name','id','fc','pval','qval')]

de_genes[,"log2fc"] <- log2(de_genes[,"fc"])

de_transcripts[,"log2fc"] <- log2(de_transcripts[,"fc"])

#Let's arrange the results from the smallest P value to the largest

de_transcripts = arrange(de_transcripts,pval)

de_genes = arrange(de_genes,pval)

#save result in .csv

write.csv(de_genes, "de_transcripts.csv", row.names=FALSE)

write.csv(de_genes, "de_genes.csv", row.names=FALSE)

#Let's subset transcripts that are detected as differentially expressed at qval <0.05

subset_transcripts <- subset(de_transcripts,de_transcripts$qval<0.05)

#do same for the genes

subset_genes <- subset(de_genes,de_genes$qval<0.05)

#create plots

dir.create('plots')

print('generating plots')

#volcano plot

#https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html

de_genes$diffexpressed <- "NO"

de_genes$diffexpressed[de_genes$log2fc > 1 & de_genes$pval < 0.05] <- "UP"

de_genes$diffexpressed[de_genes$log2fc < -1 & de_genes$pval < 0.05] <- "DOWN"
```

```
de_genes$delabel <- NA

de_genes$delabel[de_genes$diffexpressed != "NO"] <- de_genes$id[de_genes$diffexpressed != "NO"]

options(ggrepel.max.overlaps = Inf)

png('plots/volcano.png',width = 1800, height = 1000) #,width = 1800, height = 1000

volcano=ggplot(data=de_genes, aes(x=log2fc, y=-log10(pval), col=diffexpressed, label=delabel)) +

  geom_point() +

  theme_minimal() +

  geom_text_repel() +

  scale_color_manual(values=c("blue", "black", "red")) +

  geom_vline(xintercept=c(-0.8, 0.8), col="red") +

  theme(text=element_text(size=20))


  #geom_hline(yintercept=-log10(0.05), col="red")

print(volcano)

dev.off()
```

#DONE

#MAPLOT

#https://davetang.org/muse/2017/10/25/getting-started-hisat-stringtie-ballgown/

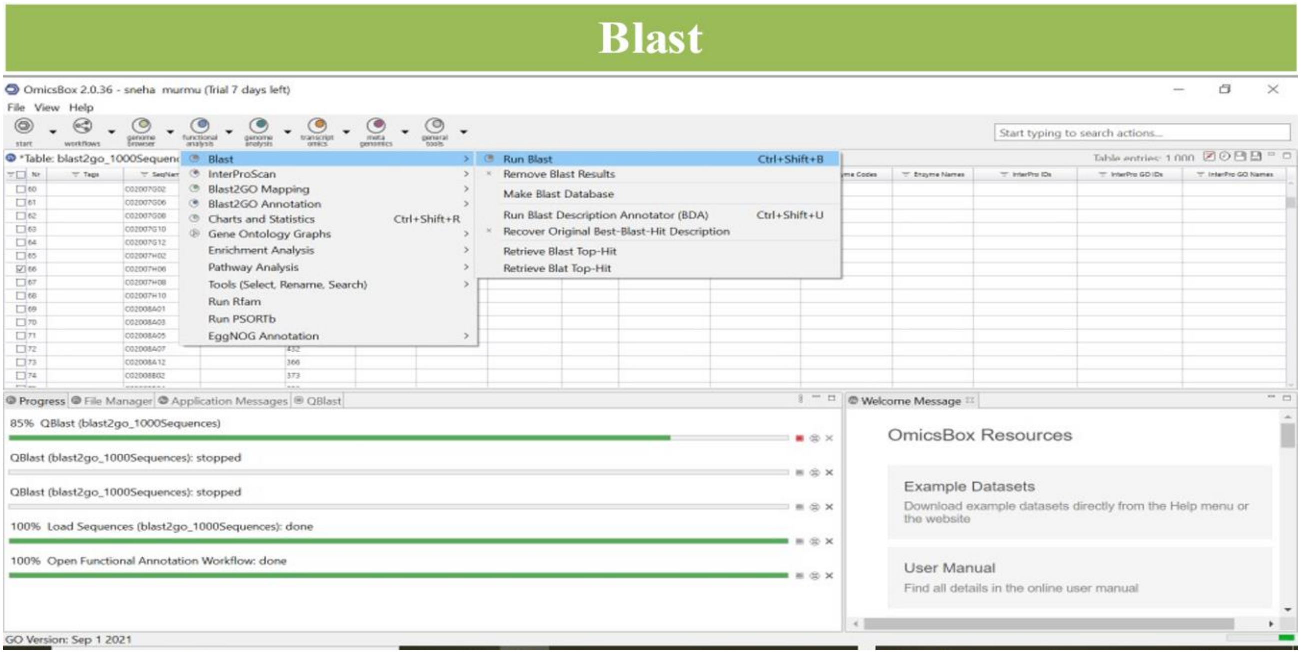
```
png('plots/maplot.png',width = 1800, height = 1000)

de_transcripts$mean <- rowMeans(texpr(bg_chrX_filt))

maplot=ggplot(de_transcripts, aes(log2(mean), log2(fc), colour = qval<0.05)) +

  scale_color_manual(values=c("#999999", "#FF0000")) +

  geom_point() +

  theme(legend.text=element_text(size=20),legend.title=element_text(size=20)) +

  theme(axis.text=element_text(size=20),axis.title=element_text(size=20)) +

  geom_hline(yintercept=0)
```

```
print(maplot)

dev.off()

#DONE

Exit R.

##extract DE transcript sequence by ID

gffread -w transcripts.fa -g chrX.fa stringtie_merged.gtf

#create index of transc.fa

cdbfasta transcripts.fa

cat up17_id_list.txt |cdbyank transcripts.fa.cidx > up17.fasta
```

3. Annotation

Functional annotation is defined as the process of collecting information about and describing a gene's biological identity—its various aliases, molecular function, biological role(s), subcellular location, and its expression domains within the plant. Blast2GO is a bioinformatics platform for high-quality functional annotation and analysis of genomic datasets. The following section mentions the four major modules involved in Blast2GO annotation.

A) Basic Local Alignment Search Tool: to search for similar (or homologous) sequences as shown in Fig 1.



85% QBlast (blast2go_1000Sequences)

QBlast (blast2go_1000Sequences): stopped

QBlast (blast2go_1000Sequences): stopped

100% Load Sequences (blast2go_1000Sequences): done

100% Open Functional Annotation Workflow: done

GO Version: Sep 1 2021

OmicsBox Resources

Example Datasets

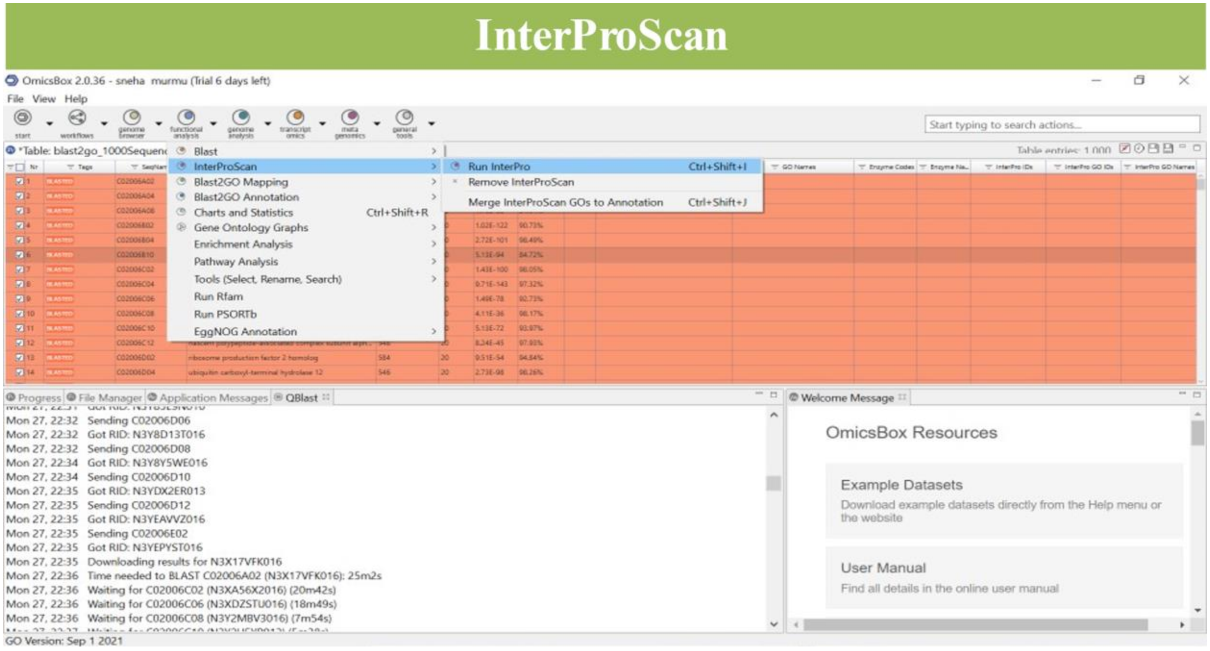
Download example datasets directly from the Help menu or the website

User Manual

Find all details in the online user manual

Figure 1: BLAST

B) InterProScan: for classification of protein families as shown in Fig 2.


InterProScan
OmicsBox 2.0.36 - sneha murmu (Trial 6 days left)
File View Help
Start typing to search actions...
*Table: blast2go_1000Sequenc
Blast
InterProScan
Blast2GO Mapping
Blast2GO Annotation
Charts and Statistics Ctrl+Shift+R
Gene Ontology Graphs
Enrichment Analysis
Pathway Analysis
Tools (Select, Rename, Search)
Run Rfam
Run PSORTb
EggNOG Annotation
Run InterPro Ctrl+Shift+I
Remove InterProScan
Merge InterProScan GOs to Annotation Ctrl+Shift+J
Progress
File Manager
Application Messages
QBlast
Welcome Message

```
Mon 27, 22:32   Sending C02006D06
Mon 27, 22:32   Got RID: N3Y8D13T016
Mon 27, 22:32   Sending C02006D08
Mon 27, 22:34   Got RID: N3Y8Y5WE016
Mon 27, 22:34   Sending C02006D10
Mon 27, 22:35   Got RID: N3YDX2ER013
Mon 27, 22:35   Sending C02006D12
Mon 27, 22:35   Got RID: N3YEAVVZ016
Mon 27, 22:35   Sending C02006E02
Mon 27, 22:35   Got RID: N3YEPYST016
Mon 27, 22:35   Downloading results for N3X17VFK016
Mon 27, 22:36   Time needed to BLAST C02006A02 (N3X17VFK016): 25m2s
Mon 27, 22:36   Waiting for C02006C02 (N3XA56X2016) (20m42s)
Mon 27, 22:36   Waiting for C02006C06 (N3XDZSTU016) (18m49s)
Mon 27, 22:36   Waiting for C02006C08 (N3Y2MBV3016) (7m54s)
GO Version: Sep 1 2021
```

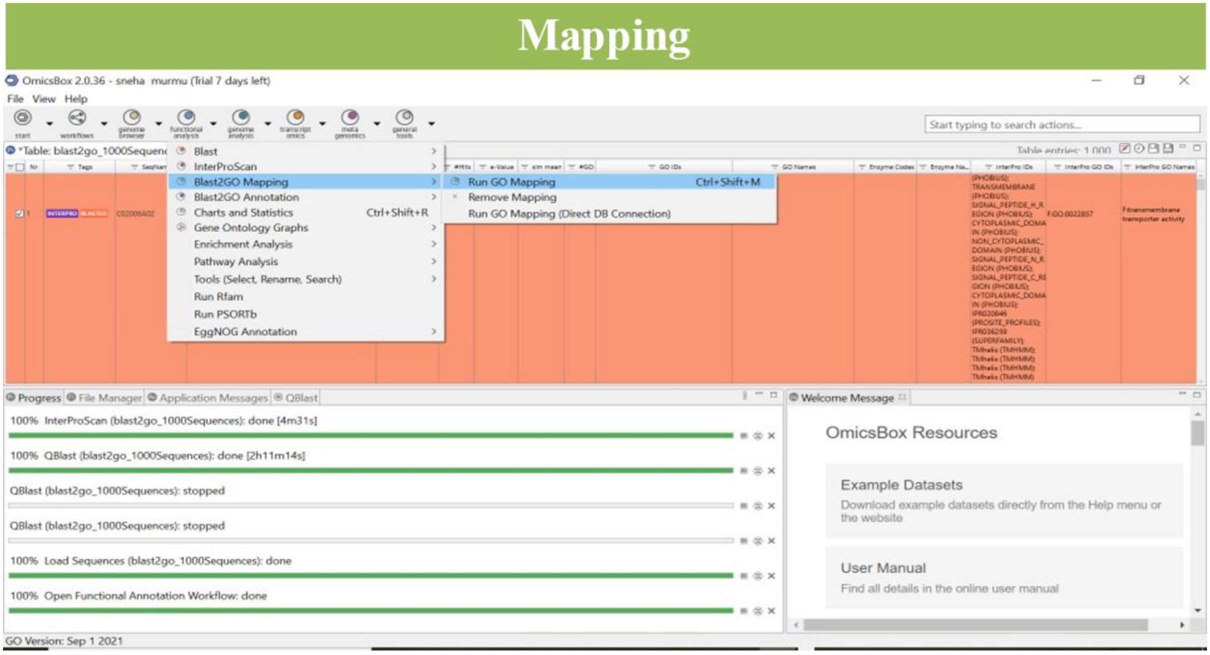
OmicsBox Resources

Example Datasets
Download example datasets directly from the Help menu or the website

User Manual
Find all details in the online user manual

Fig 2: InterProScan

C) Blast2GO Mapping: to retrieve Gene Ontology (GO) terms as shown in Fig 3.


Mapping
OmicsBox 2.0.36 - sneha murmu (Trial 7 days left)
File View Help
Start typing to search actions...
*Table: blast2go_1000Sequenc
Blast
InterProScan
Blast2GO Mapping
Blast2GO Annotation
Charts and Statistics Ctrl+Shift+R
Gene Ontology Graphs
Enrichment Analysis
Pathway Analysis
Tools (Select, Rename, Search)
Run Rfam
Run PSORTb
EggNOG Annotation
Run GO Mapping Ctrl+Shift+M
Remove Mapping
Run GO Mapping (Direct DB Connection)
Progress
File Manager
Application Messages
QBlast
Welcome Message

```
100%  InterProScan (blast2go_1000Sequences): done [4m31s]
100%  QBlast (blast2go_1000Sequences): done [2h11m14s]
QBlast (blast2go_1000Sequences): stopped
QBlast (blast2go_1000Sequences): stopped
100%  Load Sequences (blast2go_1000Sequences): done
100%  Open Functional Annotation Workflow: done
GO Version: Sep 1 2021
```

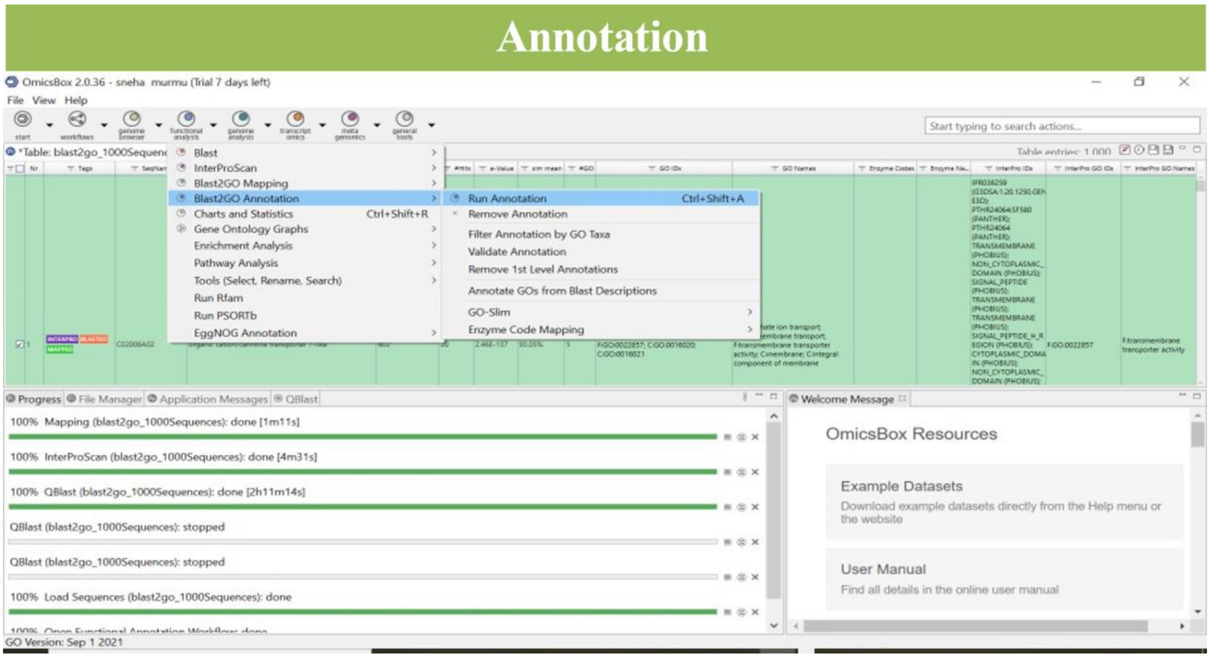
OmicsBox Resources

Example Datasets
Download example datasets directly from the Help menu or the website

User Manual
Find all details in the online user manual

Fig 3: Mapping

D) Blast2GO Annotation: to select reliable functions as shown in Fig 4.


Annotation
OmicsBox 2.0.36 - sneha murmu (Trial 7 days left)
File View Help
Start typing to search actions...
*Table: blast2go_1000Sequenc
Blast
InterProScan
Blast2GO Mapping
Blast2GO Annotation
Charts and Statistics Ctrl+Shift+R
Gene Ontology Graphs
Enrichment Analysis
Pathway Analysis
Tools (Select, Rename, Search)
Run Rfam
Run PSORTb
EggNOG Annotation
Run Annotation Ctrl+Shift+A
Remove Annotation
Filter Annotation by GO Taxa
Validate Annotation
Remove 1st Level Annotations
Annotate GOs from Blast Descriptions
GO-Slim
Enzyme Code Mapping
Progress
File Manager
Application Messages
QBlast
Welcome Message

100% Mapping (blast2go_1000Sequences): done [1m11s]

100% InterProScan (blast2go_1000Sequences): done [4m31s]

100% QBlast (blast2go_1000Sequences): done [2h11m14s]

QBlast (blast2go_1000Sequences): stopped

QBlast (blast2go_1000Sequences): stopped

100% Load Sequences (blast2go_1000Sequences): done

GO Version: Sep 1 2021

OmicsBox Resources

Example Datasets
Download example datasets directly from the Help menu or the website

User Manual
Find all details in the online user manual

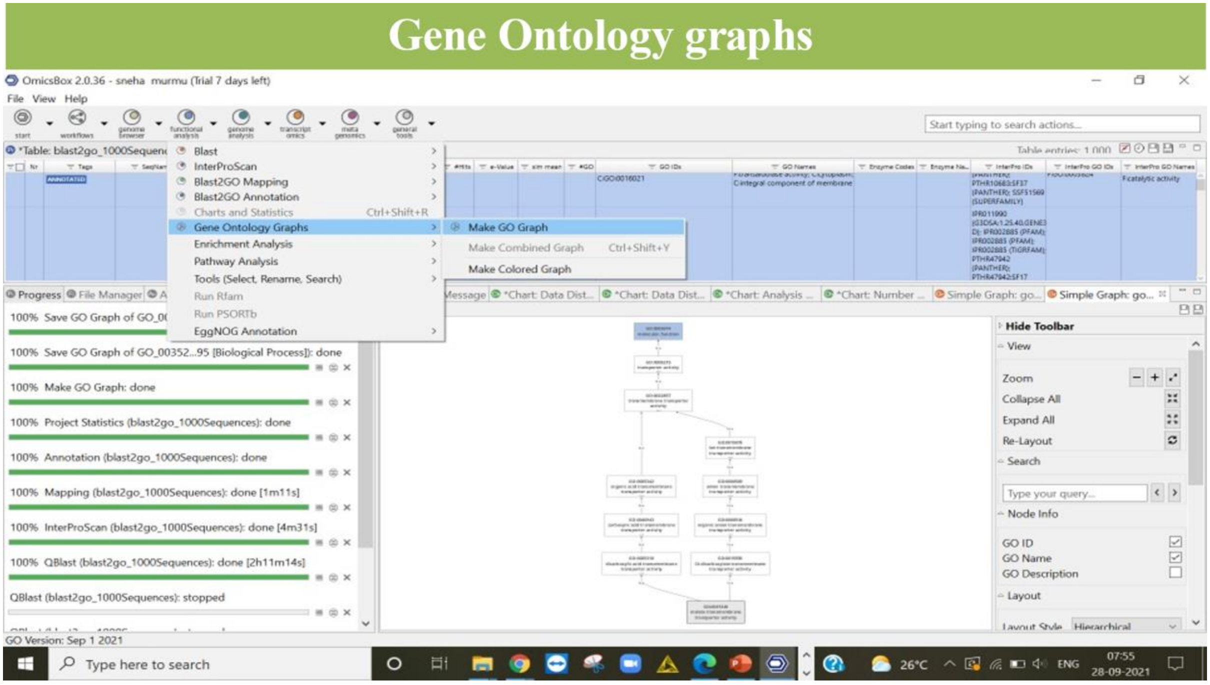
Fig 4: Annotation

Result of Blast2GO:

The result can be visualized in the following forms:

a) Gene Ontology graphs (as shown in Fig 5)

b) Pathway analysis (as shown in Fig 6)



100% Save GO Graph of GO_00

100% Save GO Graph of GO_00352...95 [Biological Process]): done

100% Make GO Graph: done

100% Project Statistics (blast2go_1000Sequences): done

100% Annotation (blast2go_1000Sequences): done

100% Mapping (blast2go_1000Sequences): done [1m11s]

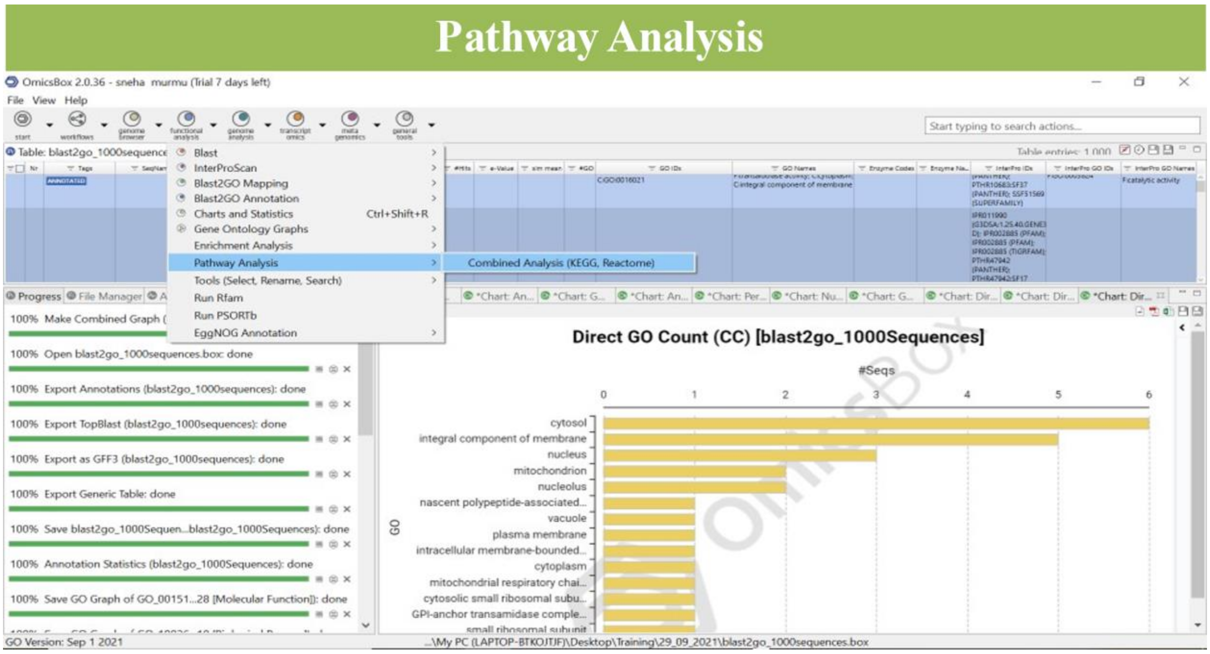
100% InterProScan (blast2go_1000Sequences): done [4m31s]

100% QBlast (blast2go_1000Sequences): done [2h11m14s]

QBlast (blast2go_1000Sequences): stopped

GO Version: Sep 1 2021

Fig 5. Gene Ontology graphs



```
100%  Make Combined Graph (
100%  Open blast2go_1000sequences.box: done
100%  Export Annotations (blast2go_1000sequences): done
100%  Export TopBlast (blast2go_1000sequences): done
100%  Export as GFF3 (blast2go_1000sequences): done
100%  Export Generic Table: done
100%  Save blast2go_1000Sequen...blast2go_1000Sequences): done
100%  Annotation Statistics (blast2go_1000Sequences): done
100%  Save GO Graph of GO_00151...28 [Molecular Function]): done
GO Version: Sep 1 2021
...\My PC (LAPTOP-BTKOJTJF)\Desktop\Training\29_09_2021\blast2go_1000sequences.box
```

Fig 6: Pathway Analysis

References:

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols, 11(9), 1650-1667.

Chapter 7

The world of miRNA

Ambika B Gaikwad,
Division of Genomic Resources
ICAR-National Bureau of Plant Genetic Resources, New Delhi 110012
ambika.gaikwad@icar.gov.in

MicroRNAs (miRNAs) are 20- to 24-nucleotide small RNAs that regulate gene expression of mRNA targets post transcriptionally. They were discovered as elements controlling gene regulation in nematodes (Lee et al., 1993; Wightman et al., 1993). Subsequently, their widespread occurrence among eukaryotes (Lagos-Quintana et al., 2001; Reinhart et al., 2002) indicated their influence on many biological processes. Micro RNAs are derived from single-stranded mRNA precursors (pre-miRNAs) that adopt a characteristic hairpin structure. The biological relevance of a miRNA is defined by the functional role of its mRNA target. Plant miRNAs tend to have high sequence complementarity with targets and act by inducing transcript cleavage, resulting in mRNA decay. This differs from animal miRNAs, which tend to share a smaller 'seed' region of complementarity with targets and act through translational inhibition. The abundance of miRNA in a cell is regulated under multiple levels of control including transcription, processing, RNA modification, RNA-induced silencing complex (RISC) assembly, miRNA-target interaction, and turnover.

MicroRNA mediated gene regulation results from a cascade of regulatory effects involving the regulation of miRNA transcription, pre-miRNA processing and the regulation of the RNA-induced silencing complex (Figure 1). Micro RNAs are classified as either "intergenic" or "intronic." Intergenic miRNAs are located between two protein-coding genes and are transcribed as independent units by DNA-dependent RNA Polymerase II (Pol II), while intronic miRNAs are processed from introns of their host transcripts (Millar and Waterhouse, 2005; Budak and Akpinar, 2015). Since they are Pol II products, the primary transcripts of miRNAs (termed pri-miRNAs) are 5' capped, 3'polyadenylated, and/or spliced (Xie et al., 2005; Rogers and Chen, 2013). Pri-miRNAs are folded into hairpin-like structures consisting of a terminal loop, an upper stem, the miRNA/miRNA_region, a lower stem, and two arms, which can be recognized and processed by Dicer-like RNase III endonucleases (DCLs). The number of DCL proteins differs across species. The DCL proteins catalyse the production of miRNA with the assistance of the assistance of accessory proteins. The nascent miRNA/miRNA* duplexes are then methylated for the assembly of RISC (ribosome induced silencing complex. In addition, there are a large number of factors contributing to miRNA stability such as 3' end modification, AGO association and miRNA-target RNA interaction.

Ever since first reported in the model organism *Caenorhabditis elegans*, miRNAs are known to represent a novel epigenetic mechanism that regulates gene expression in many homoeostatic processes and pathological conditions within the cells. In humans, the dysfunction of miRNAs has been associated with a large number of diseases such as diabetes mellitus, obesity, arthritic diseases, kidney disease, cardiovascular diseases and cancer, where they can act as either tumor suppressors or inducers. In plants, their key role in development pathways and environment response has led to their use in manipulating key pathways for agronomic use. Significant progress has been made in delineating the influence of these small untranslated RNAs on many biological processes and in devising technologies for manipulating gene expression.

References

Budak, H., and Akpinar, B. A. (2015). Plant miRNAs: biogenesis, organization andorigins. Funct. Integr. Genomic 15, 523–531. doi: 10.1007/s10142-015-0451-2

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001).Identification of novel genes coding for small expressed RNAs. Science 294,853–858.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronicgene lin-4 encodes small RNAs with antisense complementarity tolin-14. Cell 75, 843–854.

Millar, A. A., and Waterhouse, P. M. (2005). Plant and animal microRNAs:similarities and differences. Funct. Integr. Genomic 5, 129–135. doi: 10.1007/s10142-005-0145-2

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P.(2002). MicroRNAs in plants. Genes Dev. 16, 1616–1626.

Rogers, K., and Chen, X. M. (2013). Biogenesis, turnover, and mode of action ofplant MicroRNAs. Plant Cell 25, 2383–2399. doi: 10.1105/tpc.113.113159

Wang J, Mei J and Ren G (2019)Plant microRNAs: Biogenesis,Homeostasis, and Degradation.Front. Plant Sci. 10:360.doi: 10.3389/fpls.2019.00360

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. Cell 75, 855–862.

Xie, Z. X., Allen, E., Fahlgren, N., Calamar, A., Givan, S. A., and Carrington, J. C.(2005). Expression of Arabidopsis MIRNA genes. Plant Physiol. 138, 2145–2154.doi: 10.1104/pp.105.062943

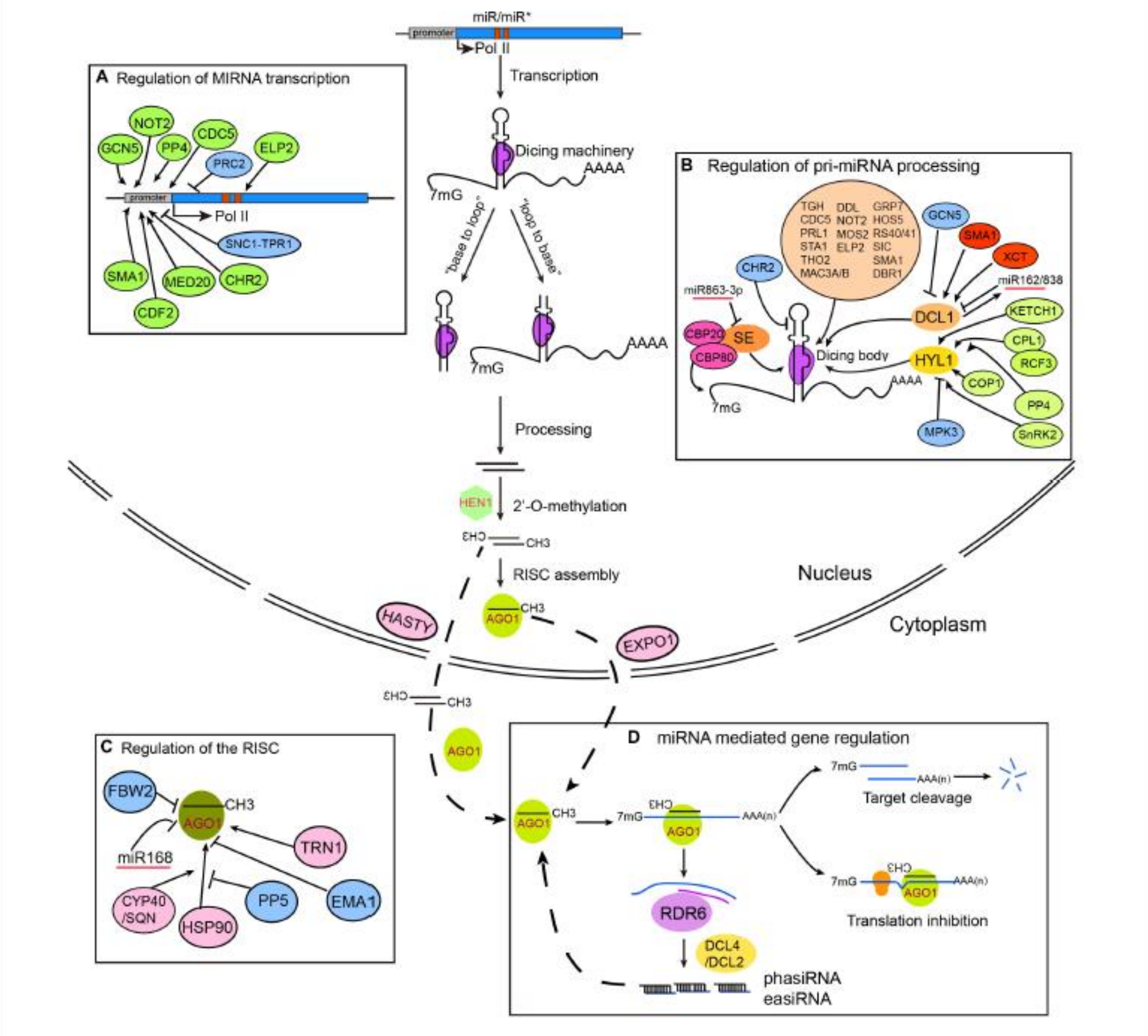


FIGURE 1 | Regulation of miRNA biogenesis, RISC loading, and action in plants. *MIRNA* genes are transcribed by RNA Polymerase II (Pol II) and fold into stem-loop structures called pri-miRNAs. pri-miRNAs are mainly processed by DCL1 from either "base-to-loop" or "loop-to-base" direction. Nascent miRNA/miRNA* duplexes are methylated by the small RNA methyltransferase HEN1. RISC loading occurs in the nucleus, but may also occur in the cytoplasm. miRNAs mediate gene silencing through either target cleavage or translation inhibition. Some miRNAs can trigger the production of secondary siRNAs through sequential actions of RDR6 and DCL2/4, generating 21–22 nt secondary siRNAs called phasiRNAs and easiRNAs, which in turn repress gene expression *via* PTGS (for phasiRNAs) or TGS (for easiRNAs) **(D)**. It is important to note that although the steps in the model are separate, they could be closely coupled. Factors involved in the regulation of *MIR* transcription, pri-miRNA processing, and RISC assembly are shown in parts **(A–C)**, respectively.

Reference: Wang et al. 2019

Chapter 8

Prediction and Characterization of miRNA

Bharati Pandey

Division of Agricultural Bioinformatics,

ICAR-Indian Agricultural Statistics Research Institute

microRNAs (miRNAs) have been shown to play pivotal roles in growth and development in animals and plants. Canonical miRNAs are endogenous ~ 21 nt small RNAs that regulate key developmental processes or the response to environmental stresses at the post-transcriptional level by mediating the cleavage of the target messenger RNAs (mRNAs) and/or by inhibiting their translation. So far, about 300 miRNAs have been annotated in the model plant *Arabidopsis thaliana*.

High-throughput sequencing of cDNA-libraries derived from endogenous small RNAs (sRNA-seq), is a widely used and powerful method for the discovery and annotation of miRNA-producing genes. Although many computational tools dealing with sRNA-seq data in animals are available, the number of tools calibrated for plants is relatively limited. Moreover, none of these tools is available as a web-server or offer a GUI. miRkwood is web-server specifically designed for plant miRNAs (Guigon et al., 2019). It is able to face the diversity of plant pre-miRNAs (producing canonical and miRNAs).

Table 1. **Computational tools for miRNA characterization**

miRPlant	http://www.australianprostatecentre.org/research/software/mirplant
miRkwood	http://bioinfo.cristal.univ-lille.fr/mirkwood
miRDetect	https://github.com/Garima268/miRDetect
C-mii	http://www.biotec.or.th/isl/c-mii
QuickMIRSeq	http://QuickMIRSeq.sourceforge.net
IsomiRage	https://cru.genomics.iit.it/Isomirage/
sRNAbench	https://arn.ugr.es/srnatoolbox/
isomiRex	http://bioinfo1.uni-plovdiv.bg/isomiRex/.
miRNAFold	https://evryrna.ibisc.univ-evry.fr/miRNAFold
PlantMiRNAPred	http://nclab.hit.edu.cn/PlantMiRNAPred/
plantMirP	https://github.com/yygen89/plantMirP.
HuntMi	http://lemur.amu.edu.pl/share/HuntMi/.
SplamiR	http://www.uni-jena.de/SplamiR.html.

MiRPara	http://www.whiov.ac.cn/bioinformatics/mirpara
miRduplexSVM	http://139.91.162.64/duplexsvm/
MaturePred	http://nclab.hit.edu.cn/maturepred/.
miRLocator	https://github.com/cma2015/miRLocator

Workflow overview of miRkwood

I. Pre-processing of reads and miRNA predictions

Adaptors are removed from the sequencing reads using Cutadapt (version 1.8.3) and reads are cleaned using Prinseq (version 0.20.4) with specified parameters: -min_len 18 -max length 30 -min_qual_mean 30. Quality of the cleaned Illumina reads is checked using FastQC (version 0.11.4).

II. Alignment and filtering

The quality checked small RNA sequencing reads are mapped on the reference genome to produce an alignment file. This can be done by any standard short read mapper, such as Bowtie2 or BWA.

III. Identification of known miRNAs

With available, genome coordinates of miRNA precursor sequences such as provided in miRBase are used to detect known miRNAs that are expressed in the sequencing data.

IV. Thermodynamic stability of the hairpin precursor

i. **Criterion 1: stability of the hairpin precursor**: This criterion is met when the MFEI of the structure is smaller than -0.8.

ii. Criterion 2: number of reads

This criterion is met when the locus has either at least 10 reads mapping to each arm, or at least 100 reads mapping in total.

iii. Criterion 3: existence of the miRNA

The most common read is selected as the guide miRNA sequence if its frequency is at least 33%. When the miRNA is properly defined, three following properties are considered:

Criterion 4: precision of the precursor processing

At least 75% of reads start in a window [− 3,+ 3] centered around the start position of the miRNA, or [− 5,+ 5] around the pairing position on the opposite arm of the stem-loop.

Criterion 5: presence of the miRNA:miRNA* duplex

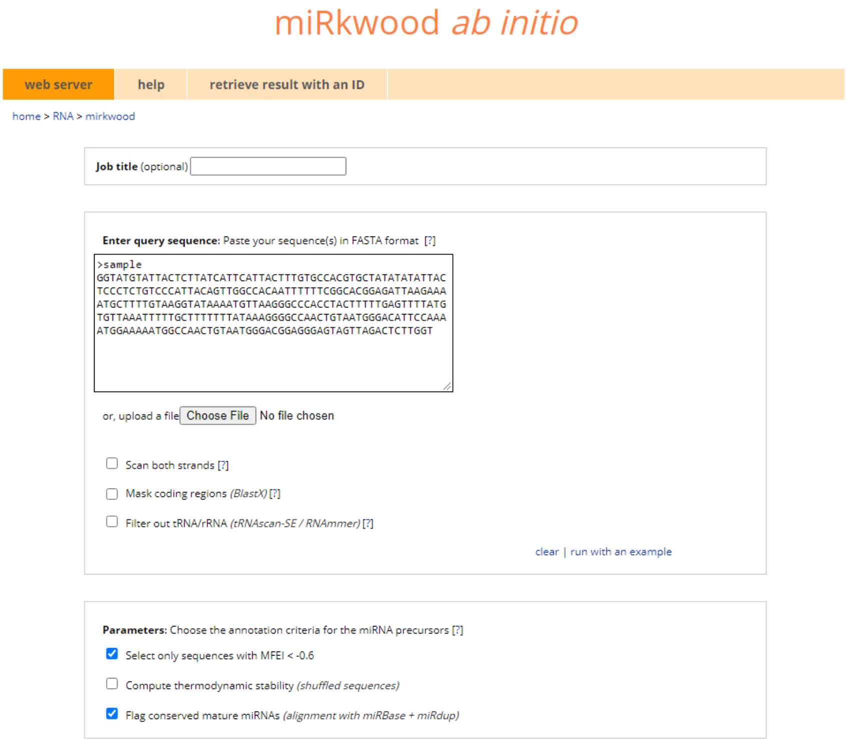
There is at least one read in the window [-5,+ 5] around the pairing position on the strand of the passenger miRNA.

Criterion 6: stability of the duplex

With this score system, hairpin precursors with no clear miRNA locus have a score of at most 2. Hairpin precursors with a guide miRNA and no passenger miRNA have a score of at most 5. Reaching a score of 6 means that the locus shows the expression of both the guide miRNA and the passenger miRNA, and that its secondary structure (hairpin and duplex) is consistent with this expression.

Availability and requirements

Project name: miRkwood
Project home page: http://bioinfo.cristal.univ-lille.fr/mirkwood,
https://github.com/miRkwood-RNA/miRkwood
Operating system(s): Unix or web server
Programming languages: Perl, C and C++
Other requirements for the Unix version: bedtools (v2.14.2 or higher), Vienna package (v2.1.6-1), Blast+ (2.2.25+ or higher), miRdup (1.2 or higher), VARNA (v3-91 or higher, optional).



miRkwood *ab initio*

web server | help | retrieve result with an ID

home > RNA > mirkwood

Job title (optional)

Enter query sequence: Paste your sequence(s) in FASTA format [?]

```
>sample
GGTATGTATTACTCTTATCATTCATTACTTTGTGCCACGTGCTATATATATTAC
TCCCTCTGTCCCATTACAGTTGGCCACAATTTTTTCGGCACGGAGATTAAGAAA
ATGCTTTTGTAAGGTATAAAATGTTAAGGGCCCACCTACTTTTTGAGTTTTATG
TGTTAAATTTTTGCTTTTTTTATAAAGGGGCCAACTGTAATGGGACATTCCAAA
ATGGAAAAATGGCCAACTGTAATGGGACGGAGGGAGTAGTTAGACTCTTGGT
```

or, upload a file Choose File No file chosen

☐ Scan both strands [?]

☐ Mask coding regions *(BlastX)* [?]

☐ Filter out tRNA/rRNA *(tRNAscan-SE / RNAmmer)* [?]

clear | run with an example

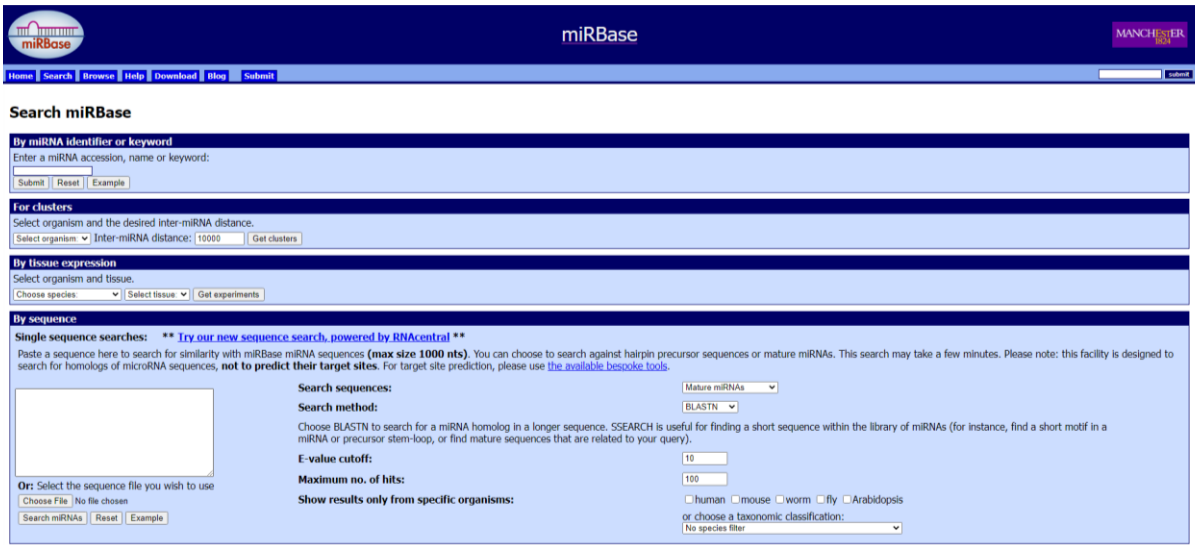
Parameters: Choose the annotation criteria for the miRNA precursors [?]

☑ Select only sequences with MFEI < -0.6

☐ Compute thermodynamic stability *(shuffled sequences)*

☑ Flag conserved mature miRNAs *(alignment with miRBase + miRdup)*

Search page of the miRkwood



Search miRBase

By miRNA identifier or keyword

Enter a miRNA accession, name or keyword:

Submit | Reset | Example

For clusters

Select organism and the desired inter-miRNA distance.

Select organism: Inter-miRNA distance: 10000 Get clusters

By tissue expression

Select organism and tissue.

Choose species: | Select tissue: | Get experiments

By sequence

Single sequence searches: ** Try our new sequence search, powered by RNAcentral **

Paste a sequence here to search for similarity with miRBase miRNA sequences **(max size 1000 nts)**. You can choose to search against hairpin precursor sequences or mature miRNAs. This search may take a few minutes. Please note: this facility is designed to search for homologs of microRNA sequences, **not to predict their target sites**. For target site prediction, please use the available bespoke tools.

Or: Select the sequence file you wish to use

Choose File No file chosen

Search miRNAs | Reset | Example

Search sequences: Mature miRNAs

Search method: BLASTN

Choose BLASTN to search for a miRNA homolog in a longer sequence. SSEARCH is useful for finding a short sequence within the library of miRNAs (for instance, find a short motif in a miRNA or precursor stem-loop, or find mature sequences that are related to your query).

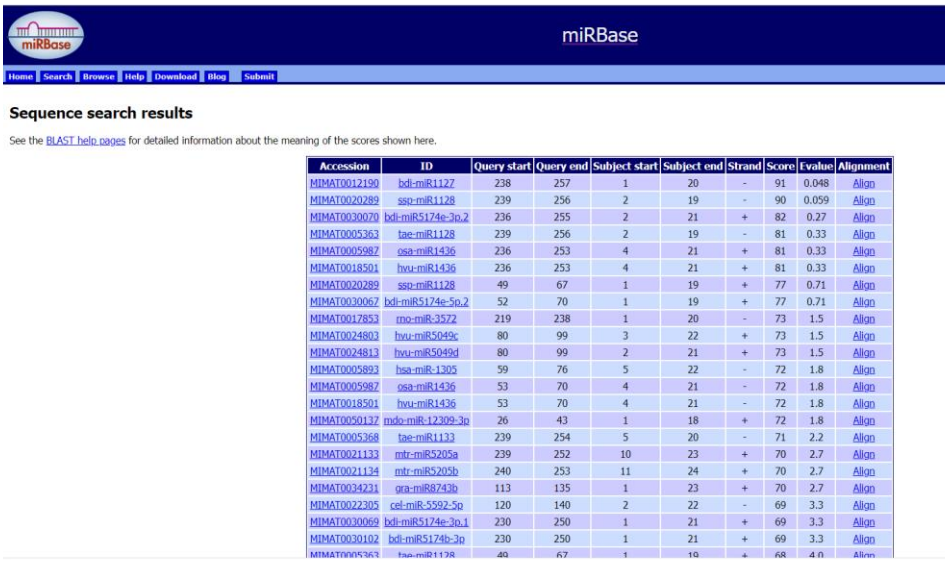
E-value cutoff: 10

Maximum no. of hits: 100

Show results only from specific organisms: ☐human ☐mouse ☐worm ☐fly ☐Arabidopsis

or choose a taxonomic classification: No species filter

Search page of the miRBase

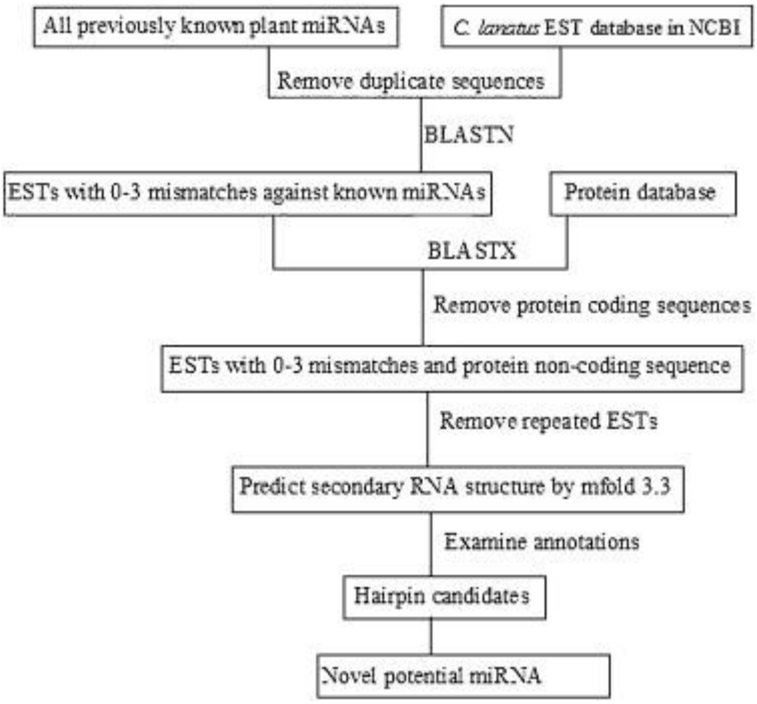


Sequence search results

See the BLAST help pages for detailed information about the meaning of the scores shown here.

Accession	ID	Query start	Query end	Subject start	Subject end	Strand	Score	Evalue	Alignment
MIMAT0012190	bdi-miR1127	238	257	1	20	-	91	0.048	Align
MIMAT0020289	ssp-miR1128	239	256	2	19	-	90	0.059	Align
MIMAT0030070	bdi-miR5174e-3p.2	236	255	2	21	+	82	0.27	Align
MIMAT0005363	tae-miR1128	239	256	2	19	-	81	0.33	Align
MIMAT0005987	osa-miR1436	236	253	4	21	+	81	0.33	Align
MIMAT0018501	hvu-miR1436	236	253	4	21	+	81	0.33	Align
MIMAT0020289	ssp-miR1128	49	67	1	19	+	77	0.71	Align
MIMAT0030067	bdi-miR5174e-5p.2	52	70	1	19	+	77	0.71	Align
MIMAT0017853	rno-miR-3572	219	238	1	20	-	73	1.5	Align
MIMAT0024803	hvu-miR5049c	80	99	3	22	+	73	1.5	Align
MIMAT0024813	hvu-miR5049d	80	99	2	21	+	73	1.5	Align
MIMAT0005893	hsa-miR-1305	59	76	5	22	-	72	1.8	Align
MIMAT0005987	osa-miR1436	53	70	4	21	-	72	1.8	Align
MIMAT0018501	hvu-miR1436	53	70	4	21	-	72	1.8	Align
MIMAT0050137	mdo-miR-12309-3p	26	43	1	18	+	72	1.8	Align
MIMAT0005368	tae-miR1133	239	254	5	20	-	71	2.2	Align
MIMAT0021133	mtr-miR5205a	239	252	10	23	+	70	2.7	Align
MIMAT0021134	mtr-miR5205b	240	253	11	24	+	70	2.7	Align
MIMAT0034231	gra-miR8743b	113	135	1	23	+	70	2.7	Align
MIMAT0022305	cel-miR-5592-5p	120	140	2	22	-	69	3.3	Align
MIMAT0030069	bdi-miR5174e-3p.1	230	250	1	21	+	69	3.3	Align
MIMAT0030102	bdi-miR5174b-3p	230	250	1	21	+	69	3.3	Align
MIMAT0005363	tae-miR1128	49	67	1	19	+	68	4.0	Align

Result of miRBase search by sequence


All previously known plant miRNAs
C. lanatus EST database in NCBI
Remove duplicate sequences
BLASTN
ESTs with 0-3 mismatches against known miRNAs
Protein database
BLASTX
Remove protein coding sequences
ESTs with 0-3 mismatches and protein non-coding sequence
Remove repeated ESTs
Predict secondary RNA structure by mfold 3.3
Examine annotations
Hairpin candidates
Novel potential miRNA

Procedure of novel potential miRNA prediction by identifying homologs of previously known miRNAs in plants (Zakeel et al., 2019)

MiRNA identification using comparative genomics approach

Identification of potential miRNAs

All previously known miRNA precursor sequences are downloaded from the miRBase database (Release 22.1; http://www.mirbase.org/ (2022). These precursor miRNAs are used as query sequences for BLASTN searches against the reference transcriptomes of species using default parameters and an E-value cut-off of 10. Only the best hit for each query sequence are retained and after elimination of redundant hits, these candidate primary miRNA sequences are scanned for hairpin-like secondary structures using the miRNA identification pipeline of the C-mii software. MirEval (http://mimirna.centenary.org.au/mireval/) was used to predict miRNA precursor sequences. RNA sequences are considered as miRNA candidates only when they fulfilled the following criteria: (1) at least 18 nt length was assumed between the predicted and mature miRNAs and (2) 0–3 nt mismatches were allowed in sequence with all previously known plant mature miRNAs. A set of miRNA candidates were screened from ESTs that closely matched with the mature miRNAs which had hitherto been identified in plants. These miRNA candidates were then used for further screening to identify miRNA precursors by evaluating the miRNA precursor prediction properties using mirEval software. These precursor sequences are subjected to BLASTx analysis with protein database and the non-protein-coding sequences were retained for RNA secondary structure prediction.

Prediction of secondary structure and new miRNA

After the removal of protein coding sequences from the candidate miRNAs, the remaining precursor sequences of potential miRNA homologs are assessed for secondary structures using the Zuker folding algorithm by MFOLD software (Zuker, 2003). The following criteria are used in defining the RNA sequences as miRNA homologs: (1) The length of predicted mature miRNAs should be in the range of 19–25 nucleotides; (2) A maximum of two mismatches compared with known rice mature miRNAs should be allowed; (3) The mature miRNA should be localized in only one arm within the predicted stem–loop structure; (4) No more than five mismatches should be allowed between miRNA sequence and guide miRNA sequence in the stem–loop structure; (5) miRNAs should have high A + U content (30–70%); and (6) the predicted secondary structure should have higher minimal folding free energy index (MFEI) and negative minimal folding free energy (MFE) (Wang et al., 2011). The MFEI was calculated using the following equation:

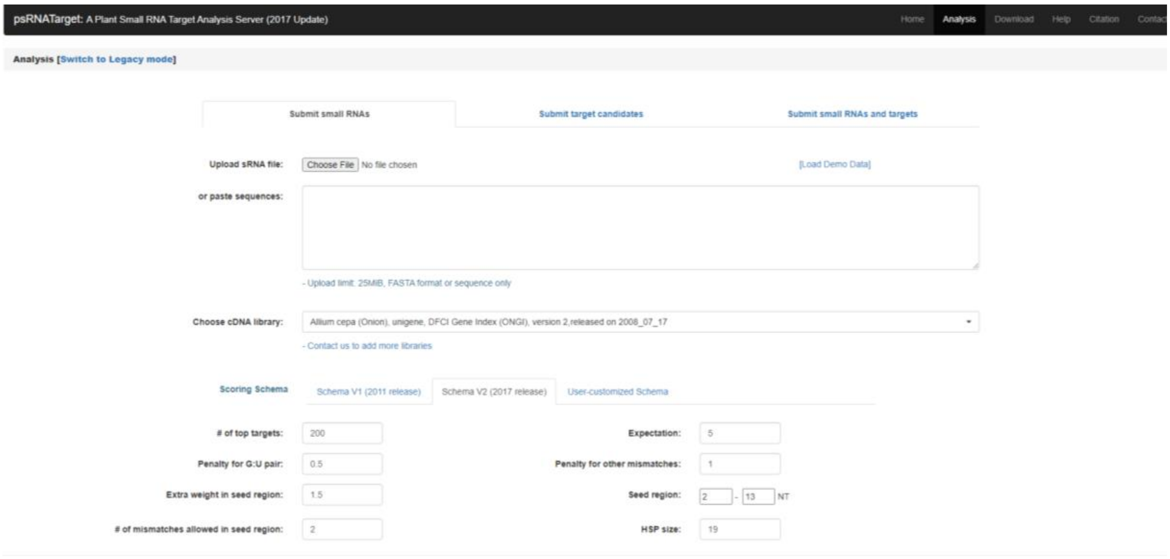
MFEI = [(MFE/length of the RNA sequence) × 100]/(G + C)

%MFE denotes the negative folding free energy (ΔG)

The resulted precursor sequences are subjected to BLASTn with mRNA database to obtain new miRNA sequences. These novel miRNAs are named according to the miRNA nomenclature criteria (Griffiths-Jones et al., 2006). Novel mature miRNA sequences are highlighted in the secondary structure by small RNA workbench software.

Identification of miRNAs Targets

As previous studies have suggested that most miRNAs can bind to protein coding regions of the target mRNAs at a perfect or near-perfect complementation and interfere or degrade mRNA (Bartel, 2004; Chen, 2004), a simple homology search against EST and nucleotide databases of *query species is performed* with the following criteria to predict targets of the potential cla-miRNAs: (1) the maximum number of mismatched nucleotides between the mature miRNA and its potential target genes was four; (2) the maximum number of mismatched nucleotides at positions 1–9 was one; (3) no mismatches were allowed at positions 10–11; (4) no more than two continuous mismatches at any position were allowed (Xie et al., 2010).



Upload sRNA file: No file chosen

or paste sequences:

- Upload limit: 25MiB, FASTA format or sequence only

Choose cDNA library: Allium cepa (Onion), unigene, DFCI Gene Index (ONGI), version 2,released on 2008_07_17

# of top targets: 200

Expectation: 5

Penalty for G:U pair: 0.5

Penalty for other mismatches: 1

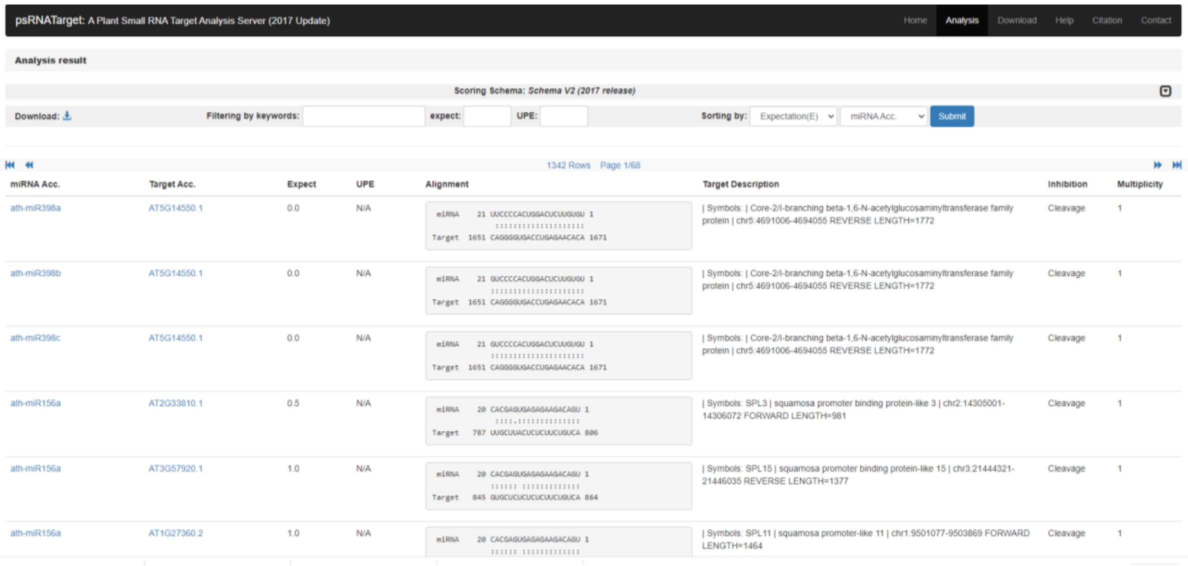
Extra weight in seed region: 1.5

Seed region: 2 - 13 NT

# of mismatches allowed in seed region: 2

HSP size: 19

Search page of miRNA target



psRNATarget: A Plant Small RNA Target Analysis Server (2017 Update)

Analysis result

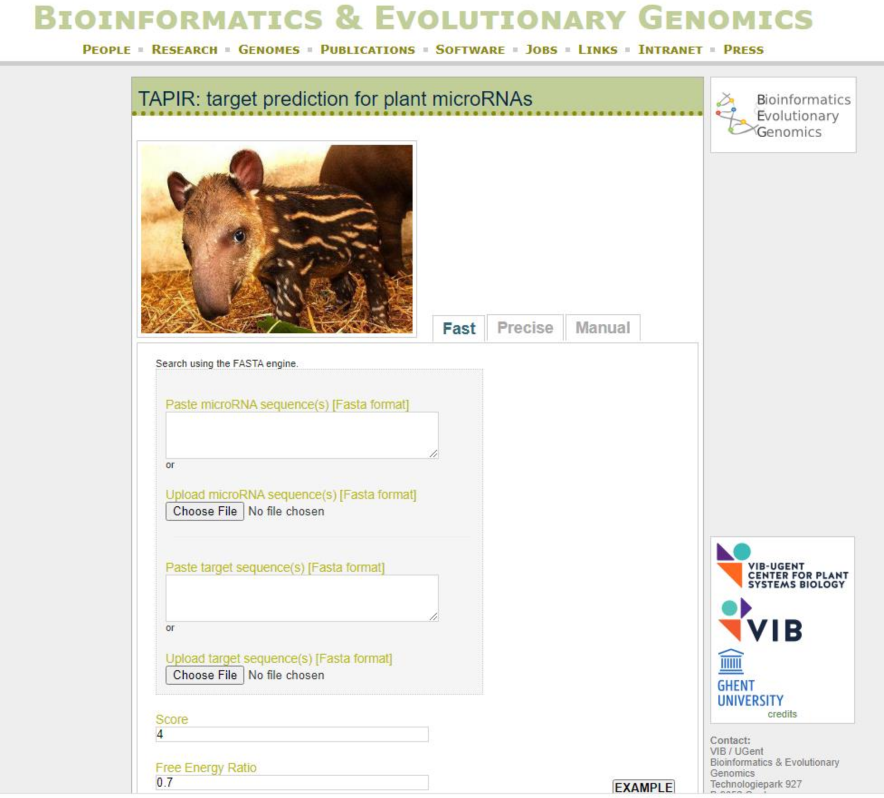
Scoring Schema: *Schema V2 (2017 release)*

Download: Filtering by keywords: expect: UPE: Sorting by: Expectation(E) miRNA Acc.

1342 Rows Page 1/68

miRNA Acc.	Target Acc.	Expect	UPE	Alignment	Target Description	Inhibition	Multiplicity
ath-miR398a	AT5G14550.1	0.0	N/A	miRNA 21 UUCCCCACUGGACUCUUGUGU 1 :::::::::::::::::::: Target 1651 CAGGGGUGACCUGAGAACACA 1671	| Symbols: | Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein | chr5:4691006-4694055 REVERSE LENGTH=1772	Cleavage	1
ath-miR398b	AT5G14550.1	0.0	N/A	miRNA 21 GUCCCCACUGGACUCUUGUGU 1 ::::::::::::::::::::: Target 1651 CAGGGGUGACCUGAGAACACA 1671	| Symbols: | Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein | chr5:4691006-4694055 REVERSE LENGTH=1772	Cleavage	1
ath-miR398c	AT5G14550.1	0.0	N/A	miRNA 21 GUCCCCACUGGACUCUUGUGU 1 ::::::::::::::::::::: Target 1651 CAGGGGUGACCUGAGAACACA 1671	| Symbols: | Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein | chr5:4691006-4694055 REVERSE LENGTH=1772	Cleavage	1
ath-miR156a	AT2G33810.1	0.5	N/A	miRNA 20 CACGAGUGAGAGAAGACAGU 1 ::::.::::::::::::::: Target 787 UUGCUUACUCUCUUCUGUCA 806	| Symbols: SPL3 | squamosa promoter binding protein-like 3 | chr2:14305001-14306072 FORWARD LENGTH=981	Cleavage	1
ath-miR156a	AT3G57920.1	1.0	N/A	miRNA 20 CACGAGUGAGAGAAGACAGU 1 :::::: ::::::::::::: Target 845 GUGCUCUCUCUCUUCUGUCA 864	| Symbols: SPL15 | squamosa promoter binding protein-like 15 | chr3:21444321-21446035 REVERSE LENGTH=1377	Cleavage	1
ath-miR156a	AT1G27360.2	1.0	N/A	miRNA 20 CACGAGUGAGAGAAGACAGU 1 :::::: :::::::::::::	| Symbols: SPL11 | squamosa promoter-like 11 | chr1:9501077-9503869 FORWARD LENGTH=1464	Cleavage	1

Result of the psRNATarget



BIOINFORMATICS & EVOLUTIONARY GENOMICS
PEOPLE ▪ RESEARCH ▪ GENOMES ▪ PUBLICATIONS ▪ SOFTWARE ▪ JOBS ▪ LINKS ▪ INTRANET ▪ PRESS
Bioinformatics Evolutionary Genomics
Fast
Precise
Manual
VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY
VIB
GHENT UNIVERSITY
credits
EXAMPLE

TAPIR: target prediction for plant microRNAs

Search using the FASTA engine.

Paste microRNA sequence(s) [Fasta format]

or

Upload microRNA sequence(s) [Fasta format]
Choose File No file chosen

Paste target sequence(s) [Fasta format]

or

Upload target sequence(s) [Fasta format]
Choose File No file chosen

Score
4

Free Energy Ratio
0.7

Contact:
VIB / UGent
Bioinformatics & Evolutionary Genomics
Technologiepark 927

Figure: Search page of miRNA target

References

F. Xie, T.P. Frazier, B. Zhang. Identification and characterization of microRNAs and their targets in the bioenergy plant switchgrass (Panicum virgatum) Planta, 232 (2) (2010), pp. 417-434.

X.A. Chen A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development Science, 303 (2004), pp. 2022-2025, 10.1126/science.1088060

D.P. Bartel MicroRNAs: genomics, biogenesis, mechanism and function Cell, 116 (2) (2004), pp. 281-297

S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, A.J. Enright miRBase: microRNA sequences, targets and gene nomenclature Nucleic Acids Res., 34 (2006), pp. D140-D144

L. Wang, H. Liu, D. Li, H. Chen Identification and characterization of maize microRNAs involved in the very early stage of seed germination BMC Genom., 12 (2011), p. 154

M. Zuker Mfold web server for nucleic acid folding and hybridization prediction Nucleic Acids Res., 13 (2003), pp. 3406-3415

Numnark, S., Mhuantong, W., Ingsriswang, S. & Wichadakul, D. C-mii: a tool for plant miRNA and target identification. BMC Genomics 13(Suppl 7), S16 (2012).

Guigon, I., Legrand, S., Berthelot, J.F., Bini, S., Lanselle, D., Benmounah, M. and Touzet, H., 2019. miRkwood: a tool for the reliable identification of microRNAs in plant genomes. *BMC genomics*, *20*(1), pp.1-9.

Zakeel, M.C.M., Safeena, M.I.S. and Komathy, T., 2019. In silico identification of microRNAs and their target genes in watermelon (Citrullus lanatus). *Scientia Horticulturae*, *252*, pp.55-60.

Chapter 9

Circular RNA: basic concept and their role in various processes

Sarika Sahu

Division of Agricultural Bioinformatics,

ICAR-Indian Agricultural Statistics Research Institute

Introduction

In the eukaryotic organisms mainly two kinds of RNAs are occurred: coding, messenger RNA (mRNA), and non-coding RNA (ncRNA). With the advent of high throughput sequencing several RNAs have been discovered and are found in cells, such as microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs), SnoRNA (small nucleolar), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), small interfering RNAs (siRNAs), small nuclear RNAs (snRNAs), piwi-interacting RNAs (Piwi-RNAs). ncRNA has little or no protein-coding potential but plays a vital role in various biological processes like gene regulation, chromosomal structure, genome defence, translation, splicing, DNA replication, healthy growth and development and stress responses. One of the important ncRNAs is circRNA, discovered over two decades ago as a special group of RNA transcripts featuring circular structures. The first identified circRNA was the potato spindle tuber viroid in 1976. Since, last four decades, circRNAs were often considered as by-products of splicing or aberrantly spliced products. Recent advancements in high-throughput sequencing technologies ease the unbiased deep profiling of circRNA landscape in a genome-wide manner. Subsequently, thousands of circRNAs have been reported in eukaryotes and archaea.

2. Biogenesis of circRNA

CircRNA is an endogenous single-stranded RNA molecule that is generated by the head-to-tail joining of pre-mRNA (back-splicing). There are three proposed models of circRNA biogenesis: (i) direct back-splicing, (ii) RNA-binding protein-mediated circularization, and (iii) lariat-driven circularization [Fig 1]. CircRNAs are generated when the pre-mRNA splicing machinery back splices to join a down-stream splice donor to an upstream splice acceptor. The 3′ and 5′ ends usually present in a linear mRNA molecule have been joined together covalently forming a characteristic back-splice junction (BSJ) in circRNA. Further, the U2-dependent spliceosome is account for the splicing of the vast majority of introns in both plants and animals, with GT and AG terminal dinucleotides at their 5′ and 3′ termini, respectively. However, in plants, both monocot and dicot species have different mechanism of the splice signals for circRNAs. Further, only a small portion (7.3%) of circRNAs possess canonical GT/AG (CT/AC) splicing signals, and a large number of circRNAs share diverse non-GT/AG splicing signals, such as GC/GG, CA/GC, GG/AG, GC/CG, and CT/CC was reported in plants. CircRNAs have multiple origin sites; they can originate from multi-exonic transcripts, single exonic transcripts, uncharacterized transcripts and even fusion genes. In addition, Alternative RNA processing events have been observed in circRNAs, including exon skipping, intron retention and alternative splicing. Although most circular RNAs are lowly expressed, some of them are able to accumulate to high levels and even exceed their cognate mRNAs due to their longer half-lives. The majority of circRNAs are ecircRNAs, which are predominantly located in the cytoplasm. However, EIcircRNAs and ciRNAs are usually located in the nucleus. Once produced in the nucleus, the majority of circular RNAs are exported to the cytoplasm for their proper functions or degradation.

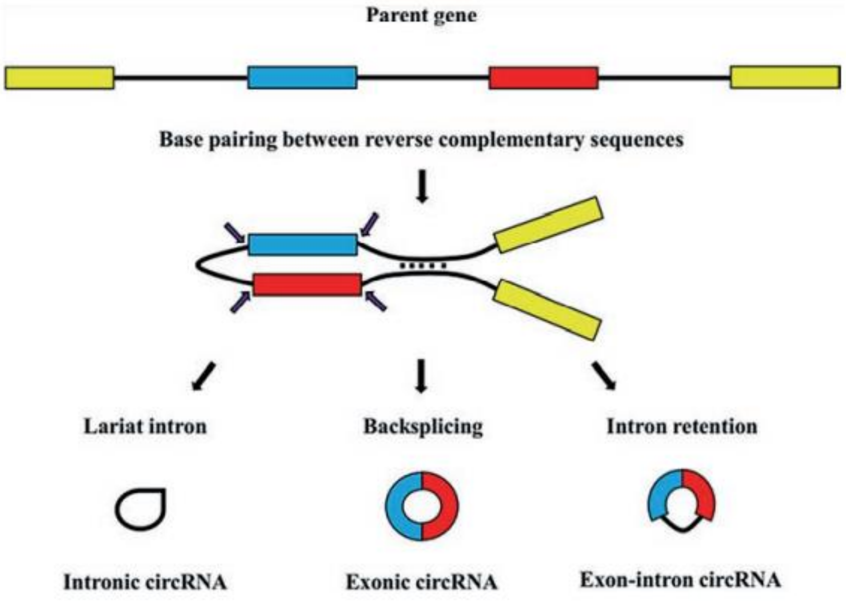


Fig1: biogenesis of different types of circRNA

3. Types of circular RNA

According to their genomic location, circRNAs are classified into exon, intron, intergenic, and exon-intron molecules. Intron circRNA mostly regulates its parental gene than exon circRNA. On the basis of origin of circRNA on the genome, circRNAs were classified into ten types (Fig. 2), at which the two back-splicing sites of a certain circRNA are located.

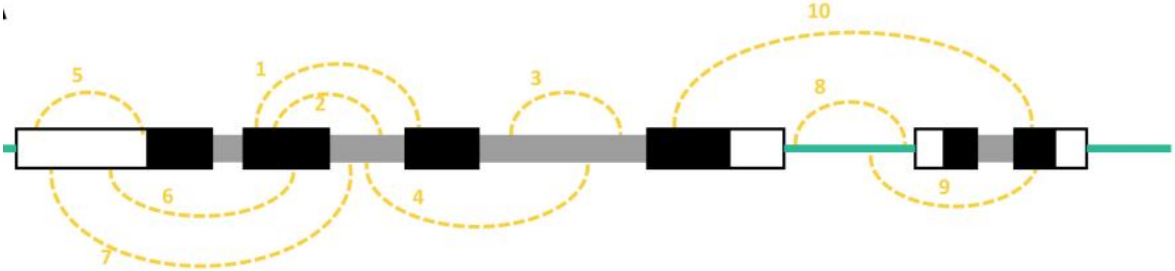


Fig2: Types of circRNAs on the basis of their generation from the parent gene. The black, grey and blank bars represent exons, introns and UTRs, respectively. The green lines represent intergenic region of the genomes

no. on fig2	**Type of circRNA**	**Type of Origin**
1	e-circRNA	two back-splicing sites of a circRNA are both at exons
2	ei-circRNA	one back-splicing site of a circRNA is at exon while the other is at intron
3	i-circRNA	two back-splicing sites of a circRNA are both at a single intron
4	ie-circRNA	two back-splicing sites of a circRNA are at two different introns across one or several exons
5	u-circRNA	two back-splicing sites of a circRNA are both at UTRs
6	ue-circRNA	one back-splicing site of a circRNA is at UTR while the other is at exon
7	ui-circRNA	one back-splicing site of a circRNA is at UTR while the other is at intron
8	ig-circRNA	two back-splicing sites of a circRNA are both at a single intergenic region
9	igg-circRNA	one back-splicing site of a circRNA is at intergenic region while the other is at genic region
10	ag-circRNA	two back-splicing sites of a circRNA are at two different genes

4. Characteristics of Plant circular RNAs

The nucleotide length of circRNAs are vary and ranges from <100 nt to >4 kb. They are conserved and have various isoforms that are generated by alternative circularization in plants. However, some circRNAs are only observed in certain plant species. The majority of plant exonic circRNAs contain 1-4 exons and large parental genes with multiple shorter exons are preferentially circularised. They are less likely to be generated from exon(s) flanked by introns containing repetitive and/or reverse complementary sequences. In Arabidopsis, out of the 13 validated plant circRNAs, only two (~15%) contain >15-bp reverse complementary sequences in their flanking introns. Similarly, in cotton (*Gossypium sp.*), despite circRNAs seem to have more repeat sequences in their flanking introns than linear genes, only ~10% of exonic circRNAs are associated with reverse complementary intronic sequences. A recent study in maize (Zea mays) found that LLERCPs (reverse complementary pairs of LINE1-like elements) are significantly enriched in the 35-kb, particularly in the 5-kb, flanking regions of circRNAs 20. The study also found that circRNAs with LLERCPs have an expression level significantly higher than those without LLERCPs nearby, indicating LLERCPs could reinforce the expression of circRNAs, although the numbers of LLERCPs seem not to be related to the expression level of circRNAs 20. Because LLERCPs were found in a relatively large flanking region of circRNAs, it is of interest to know how they are related to circRNA biogenesis. It is also of interest to know whether repeat sequences located at the flanking introns of circRNAs are associated with genome complexity so that large and polyploid genomes tend to have more repeat sequences in their flanking introns of circRNAs. In addition, multiple circRNAs can be generated from a single parental gene through alternative back splicing and circularization. Parental genes of over 700 exonic circRNAs (~15% of Arabidopsis circRNAs) are orthologs between rice and Arabidopsis. Approximately 34% and 55% of circRNA-producing soybean genes are conserved orthologs in Arabidopsis and rice, respectively. In the context of expression, they are not highly expressed while few are highly accumulated and exceed their cognate mRNAs due to their longer half-lives. Once produced in the nucleus, the majority of circular RNAs are exported to the cytoplasm for their proper functions or degradation.

5. Functional role of circRNA in plant

(i) Acting as miRNA sponges

The most extensively studied function of circRNAs is microRNA (miRNA) sponging. miRNAs are small noncoding RNAs that bind to target mRNAs and typically induce mRNA degradation or translational repression. Further, circRNAs have been found to bind miRNAs, decreasing their availability and thereby upregulating the expression of their target mRNAs. The first cases of miRNA sponging were discovered for CDR1as, with over 70 conserved target sites for miR-7, and circSry, with 16 binding sites for miR-138. circRNAs functioning as a miRNA sponge continue to be frequently documented and reported. However, studies that analysed thousands of circRNAs found that most contain a smaller number of miRNA binding sites and do not have other properties of effective miRNA sponges. These findings suggest that the majority of circRNAs do not act as miRNA sponges, and many studies have revealed other functions

(ii) Regulating transcription and translation

Further studies found that circRNAs perform many other regulatory functions, including exerting transcriptional and translational control, sequestering and translocating proteins, facilitating interactions between proteins, and translating to proteins. It was also observed that some engineered circRNAs having an internal ribosome entry site (IRES) could be translated and form small peptides in vivo.

(iii) circRNA as biomarkers

circRNAs could also be used as potential biomarkers in plants due to their unique characteristics, including resistance to degradation, long halflives, and ease the specificity of detection. Same study was reported in Arabidopsis, circRNAs used as bona fide biomarkers of functional exon-skipped AS variants, including in the homeotic MADS-box transcription factor family.

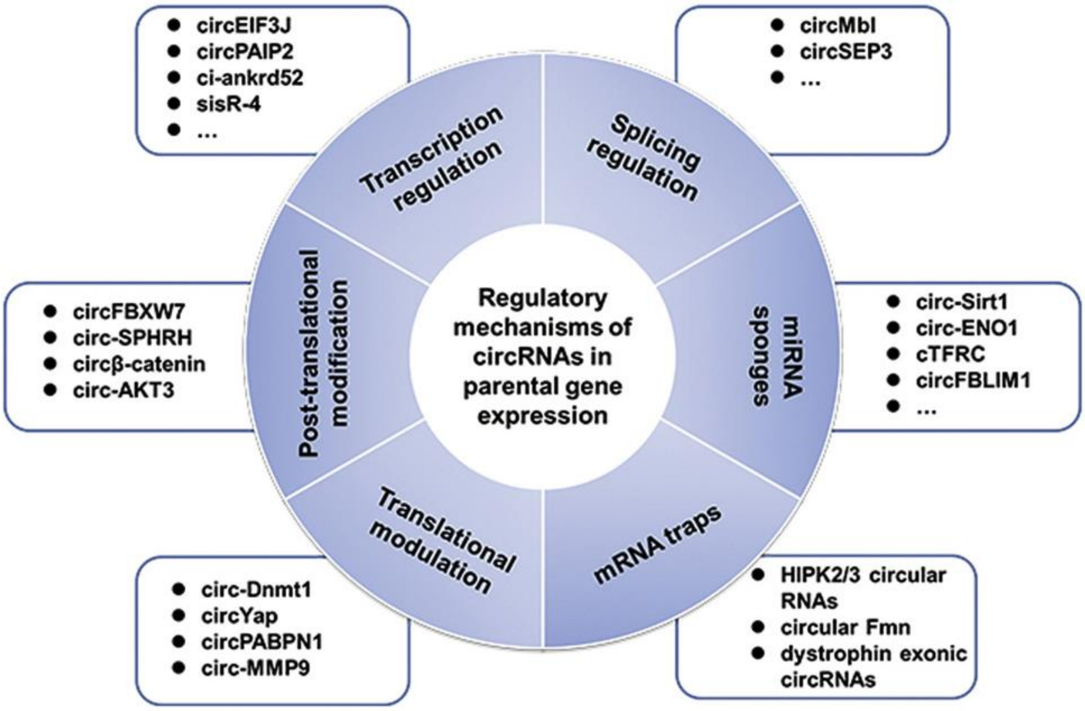


Fig3: functional role of parental gene of circRNA

(iv) Potential role of circRNAs in stress responses

circRNAs usually exhibit specific cell-type, tissue, and developmental stage expression patterns, and furthermore, the expression of circRNAs and circRNA isoforms is often induced under diverse environmental stresses, such as low- and high-light stresses, Pi-starvation conditions, low temperature stress, dehydration stress, and chewing injury stress by insects, which suggests that circRNAs might play important roles in plant development or in the response to biotic and abiotic stresses. Zhao *et al* discovered total 293 EIcircRNAs, including 183 and 175 in resistant and susceptible samples, under defoliation damage stress by cotton bollworm feeding in soybean, which indicated that EIcircRNAs might participate in the response to chewing injury resistance processes in plants. In addition, circRNAs of barley that are highly expressed in the mitochondria might be participated in micronutrient homeostasis.

(v) Role of circRNA in plant development

The overexpression of PSY1-circ1, a circRNA derived from *Phytoene Synthase 1* (*PSY1*) in tomato, resulted in a significant decrease in lycopene and β-carotene accumulation in transgenic tomato fruits, which suggests the involvement of circRNAs in plant development.

Table 1: List of tool for the prediction of circRNA

Tool	Version	Mapping tool	Address	References
circRNA finder	N/A	STAR	https://github.com/orzechoj/circRNA_finder	Westholm et al., 2014
CIRCexplorer	1.1.10	Bowtie1 and 2	https://github.com/YangLab/CIRCexplorer	Zhang et al., 2014
CIRI	1.2	Bwa	https://sourceforge.net/projects/ciri/files/	Gao et al., 2015
find circ	v2	Bowtie2	https://github.com/marvin-jens/find_circ	Memczak et al., 2013
Mapsplice	2.2.1	Bowtie1	http://www.netlab.uky.edu/p/bioinfo/MapSplice2	Wang et al., 2010
circseq-cup	1.0	STAR	http://ibi.zju.edu.cn/bioinplant/tools/circseq-cup.htm	Ye et al., 2017
KNIFE	1.4	Bowtie1, Bowtie2	https://github.com/lindaszabo/KNIFE	Szabo et al., 2015
Segemehl	0.2.0	Segemehl	http://www.bioinf.uni-leipzig.de/Software/segemehl/	Hoffmann et al., 2014
UROBORUS	0.0.2	Bowtie Bowtie2 tophat2	http://uroborus.openbioinformatics.org/en/latest/	Song et al., 2016

Table 2: List of plant database of circRNA

Database	Organisms	URL

PlantcircBase	*Oryza sativa, Arabidopsis thaliana, Zea mays, Solanum lycopersicum, Triticum aestivum, Glycine max, Gossypium hirsutum, Hordeum vulgare, Solanum tuberosum, Poncirus trifoliate, Gossypium arboretum Gossypium raimondii, Camellia sinensis, Pyrus betulifolia, Oryza sativa ssp. Indica, Nicotiana benthamiana,Brassica rapa, Cucumis sativus, Echinochloa crus-galli, Populus trichocarpa*	*http://ibi.zju.edu.cn/plantcircbase/index.php*
AtCircDB	*Arabidopsis thaliana*	*http://www.deepbiology.cn/circRNA/*
GreenCircRNA	*Ananas comosus, Amaranthus hypochondriacus, Arabidopsis lyrata, Asparagus officinalis, Arabidopsis thaliana, Botryococcus braunii, Brachypodium distachyon, Brachypodium hybridum, Brassica oleracea capitate, Brassica rapa FPsc, Brachypodium stacei, Brachypodium sylvaticum, Cicer arietinum, Citrus clementina, Capsella grandiflora, Carica papaya, Chenopodium quinoa, Chlamydomonas reinhardtii, Capsella rubella, Cucumis sativus, Citrus sinensis, Chromochloris zofingiensis, Daucus carota, Dunaliella salina, Eucalyptus grandis, Eutrema salsugineum, Fragaria vesca, Gossypium hirsutum, Glycine max, Gossypium raimondii, Helianthus annuus, Hordeum vulgare, Kalanchoe fedtschenkoi, Lactuca sativa, Linum usitatissimum, Musa acuminate, Malus domestica, Manihot esculenta, Mimulus guttatus, Marchantia polymorpha, Micromonas pusilla CCMP1545, Micromonas sp.RCC299, Medicago truncatula, Olea europaea, Oryza sativa, Oryza sativa Kitaake, Populus deltoides WV94, Panicum hallii, Physcomitrella patens, Prunus persica, Populus trichocarpa, Porphyra umbilicalis, Panicum virgatum, Phaseolus vulgaris, Ricinus communis, Sorghum bicolor, Setaria italic, Solanum lycopersicum, Spirodela polyrhiza, Salix purpurea, Solanum tuberosum, Setaria viridis, Triticum aestivum, Theobroma cacao, Trifolium pratense, Vigna unguiculata, Vitis vinifera, Zostera marina, Zea mays*	***http://greencirc.cn***

References

- Babaei, Saeid, Mohan B. Singh, and Prem L. Bhalla. "Circular RNAs repertoire and expression profile during Brassica rapa pollen development." *International journal of molecular sciences* 22, no. 19 (2021): 10297.
- Belousova, E. A., M. L. Filipenko, and N. E. Kushlinskii. "Circular RNA: new regulatory molecules." *Bulletin of Experimental Biology and Medicine* 164, no. 6 (2018): 803-815.
- Chaabane, Mohamed, Robert M. Williams, Austin T. Stephens, and Juw Won Park. "circDeep: deep learning approach for circular RNA classification from other long non-coding RNA." *Bioinformatics* 36, no. 1 (2020): 73-80.
- Cheng, Jinping, Yong Zhang, Ziwei Li, Taiyun Wang, Xiaotuo Zhang, and Binglian Zheng. "A lariat-derived circular RNA is required for plant development in Arabidopsis." *Science China Life Sciences* 61, no. 2 (2018): 204-213.
- Dong, Rui, Xu-Kai Ma, Guo-Wei Li, and Li Yang. "CIRCpedia v2: an updated database for comprehensive circular RNA annotation and expression comparison." *Genomics, proteomics & bioinformatics* 16, no. 4 (2018): 226-233.
- Gao, Yuan, Jinyang Zhang, and Fangqing Zhao. "Circular RNA identification based on multiple seed matching." *Briefings in bioinformatics* 19, no. 5 (2018): 803-810.
- Guria, Ashirbad, Kavitha Velayudha Vimala Kumar, Nagesh Srikakulam, Anakha Krishnamma, Saibal Chanda, Satyam Sharma, Xiaofeng Fan, and Gopal Pandi. "Circular RNA profiling by Illumina sequencing via template-dependent multiple displacement amplification." *BioMed research international* 2019 (2019).
- Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt L, Teupser D, Hackermueller J, Stadler PF: "A multi-split mapping algorithm for circular RNA, splicing, trans-splicing, and fusion detection", Genome Biology (2014) 15:R34.
- Jakobi, Tobias, Alexey Uvarovskii, and Christoph Dieterich. "circtools—a one-stop software solution for circular RNA research." *Bioinformatics* 35, no. 13 (2019): 2326-2328.
- Jakobi, Tobias, and Christoph Dieterich. "Computational approaches for circular RNA analysis." *Wiley Interdisciplinary Reviews: RNA* 10, no. 3 (2019): e1528.
- Jakub O. Westholm, Pedro Miura, Sara Olson, Sol Shenker, Brian Joseph, Piero Sanfilippo, Susan E. Celniker, Brenton R. Graveley, and Eric C. Lai. Genome-wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation Westholm et al. Cell Reports, 2014.
- Kai Wang, Darshan Singh, Zheng Zeng, Stephen J. Coleman, Yan Huang, Gleb L. Savich, Xiaping He, Piotr Mieczkowski, Sara A. Grimm, Charles M. Perou, James N. MacLeod, Derek Y. Chiang, Jan F. Prins and Jinze Liu *Nucleic Acids Research* 2010; doi: 10.1093/nar/gkq622.
- Memczak, S.; Jens, M.; Elefsinioti, A.; Torti, F.; Krueger, J.; Rybak, A.; Maier, L.; Mackowiak, S.D.; Gregersen, L.H.; Munschauer, M.; et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 2013, 495, 333–338.
- Song X, Zhang N, Han P, Lai RK, Wang K, Lu W. Circular RNA Profile in Gliomas Revealed by Identification Tool UROBORUS. Nucleic Acids Research, 2016, 44:e87.
- Szabo L, Morey R, Palpant NJ, Wang PL, Afari N, Jiang C, Parast MM, Murry CE, Laurent LC, Salzman J. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biology. 2015, 16:126.
- Tong, Wei, Jie Yu, Yan Hou, Fangdong Li, Qiying Zhou, Chaoling Wei, and Jeffrey L. Bennetzen. "Circular RNA architecture and differentiation during leaf bud to young leaf development in tea (Camellia sinensis)." Planta 248, no. 6 (2018): 1417-1429.

- Wang, Kai, Chong Wang, Baohuan Guo, Kun Song, Chuanhong Shi, Xin Jiang, Keyi Wang, Yacong Tan, Lequn Wang, Lin Wang, Jiangjiao Li, Ying Li, Yu Cai, Hongwei Zhao and Xiaoyong Sun. “CropCircDB: a comprehensive circular RNA resource for crops in response to abiotic stress.” Database: The Journal of Biological Databases and Curation 2019 (2019): n. pag.
- Wang, Ying, Zeyang Xiong, Qian Li, Yueyang Sun, Jing Jin, Hao Chen, Yu Zou, Xingguo Huang, and Yi Ding. "Circular RNA profiling of the rice photo-thermosensitive genic male sterile line Wuxiang S reveals circRNA involved in the fertility transition." BMC plant biology 19, no. 1 (2019): 1-16.
- Yang, Zhenchao, Zhao Yang, Yingge Xie, Qi Liu, Yanhao Mei, and Yongjun Wu. "Systematic identification and analysis of light-responsive circular RNA and co-expression networks in lettuce (Lactuca sativa)." G3: Genes, Genomes, Genetics 10, no. 7 (2020): 2397-2410.
- Ye et al., Full length sequence assembly reveals circular RNAs with diverse non GT AG splicing signals in rice. RNA Biology. 2016.
- Ye, Jiazhen, Lin Wang, Shuzhang Li, Qinran Zhang, Qinglei Zhang, Wenhao Tang, Kai Wang et al. "AtCircDB: a tissue-specific database for Arabidopsis circular RNAs." Briefings in Bioinformatics 20, no. 1 (2019): 58-65.
- Yin, Shuwei, Xiao Tian, Jingjing Zhang, Peisen Sun, and Guanglin Li. "PCirc: random forest-based plant circRNA identification software." BMC bioinformatics 22, no. 1 (2021): 1-14.
- Yuan Gao†, Jinfeng Wang† and Fangqing Zhao*. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biology (2015) 16:4.
- Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL and Yang L. Complementary sequence-mediated exon circularization. Cell, 2014, 159: 134-147.
- Zhang, Pei, Yuan Fan, Xiaopeng Sun, Lu Chen, William Terzaghi, Etienne Bucher, Lin Li, and Mingqiu Dai. "A large-scale circular RNA profiling reveals universal molecular mechanisms responsive to drought stress in maize and Arabidopsis." The Plant Journal 98, no. 4 (2019): 697-713.

Hands-on-session for circRNA prediction

- Kindly see the manual of bwa link is given below:
- (https://bio-bwa.sourceforge.net/bwa.shtml)
- Kindly download CIRI2 from the link given below:
- https://sourceforge.net/projects/ciri/files/CIRI2/
- Step1: bwa index reference_file.fa
- Step2: bwa mem index_file fastq_file > input.sam (single end data)
- bwa mem index_file read1.fq read2.fq > input.sam (Paired-end data)
- Step3: perl CIRI2.pl --help
    - perl CIRI2.pl -I input.sam -O circRNA –F reference_file.fa -T 10

सत्यमेव जयते

Indian Council of Agricultural Research

A Training Programme on

RNA world: Advance Bioinformatics for deciphering regulatory molecules

Under the aegis of
CRP-Genomics Network Project

Some aspects of RNAome in biofortification of plant and animal traits

A. R. Rao[1] and Sarika Sahu[2]
[1] Indian Council of Agricultural Research, New Delhi-110012
[2] ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012
rao.cshl.work@gmail.com

भाकृअनुप
ICAR

RNA World

- **RNA world** - a hypothetical stage in the evolutionary history of life on Earth - self-replicating RNA molecules proliferated before the evolution of DNA and proteins
- Alexander (1962) - concept of the RNA world
- Walter (1986) - coined the term RNA World
- Cech (2012) - If the RNA world existed, it was probably followed by an age characterized by the evolution of ribonucleoproteins
- Patel *et al.* (2015) - Alternative chemical paths to life have been proposed, and RNA-based life may not have been the first life to exist
- RNA world can serve as a model system for studying the origin of life

RNAome

Abbr.	Name
ncRNA	non coding RNA
nmRNA	non messenger RNA
sRNA	small RNA
smnRNA	small non messenger RNA
tRNA	transfer RNA
sRNA	soluble RNA
mRNA	messenger RNA
pcRNA	protein coding RNA
rRNA	ribosomal RNA
5S rRNA	5S ribosomal RNA
5.8S rRNA	5.8S ribosomal RNA
SSU rRNA	small subunit ribosomal RNA
LSU rRNA	large subunit ribosomal RNA

Abbr.	Name
NoRC RNA	nucleolar remodeling complex associated RNA
pRNA	promoter RNA
6S RNA or SsrS RNA	6S RNA
aRNA	antisense RNA
asRNA	antisense RNA
asmiRNA	antisense micro RNA
cis-NAT	cis-natural antisense transcript
crRNA	CRISPR RNA
tracrRNA	trans-activating crRNA
CRISPR RNA	CRISPR-Cas RNA
DD RNA	DNA damage response RNA
diRNA	DSB-induced small RNAs
dsRNA	double stranded RNA
endo-siRNA	endogenous small interfering RNA

Abbr.	Name
exRNA	extracellular RNA
gRNA	guide RNA
hc-siRNA	heterochromatic small interfering RNA
hcsiRNA	heterochromatic small interfering RNA
hnRNA	heterogeneous nuclear RNA
RNAi	RNA interference
lincRNA	long intergenic non-coding RNA
lncRNA	long non coding RNA
miRNA	micro RNA
mrpRNA	mitochondrial RNA processing ribonuclease
nat-siRNA	natural antisense short interfering RNA
natsiRNA	natural antisense short interfering RNA
OxyS RNA	oxidative stress response RNA
piRNA	piwi-interacting RNA

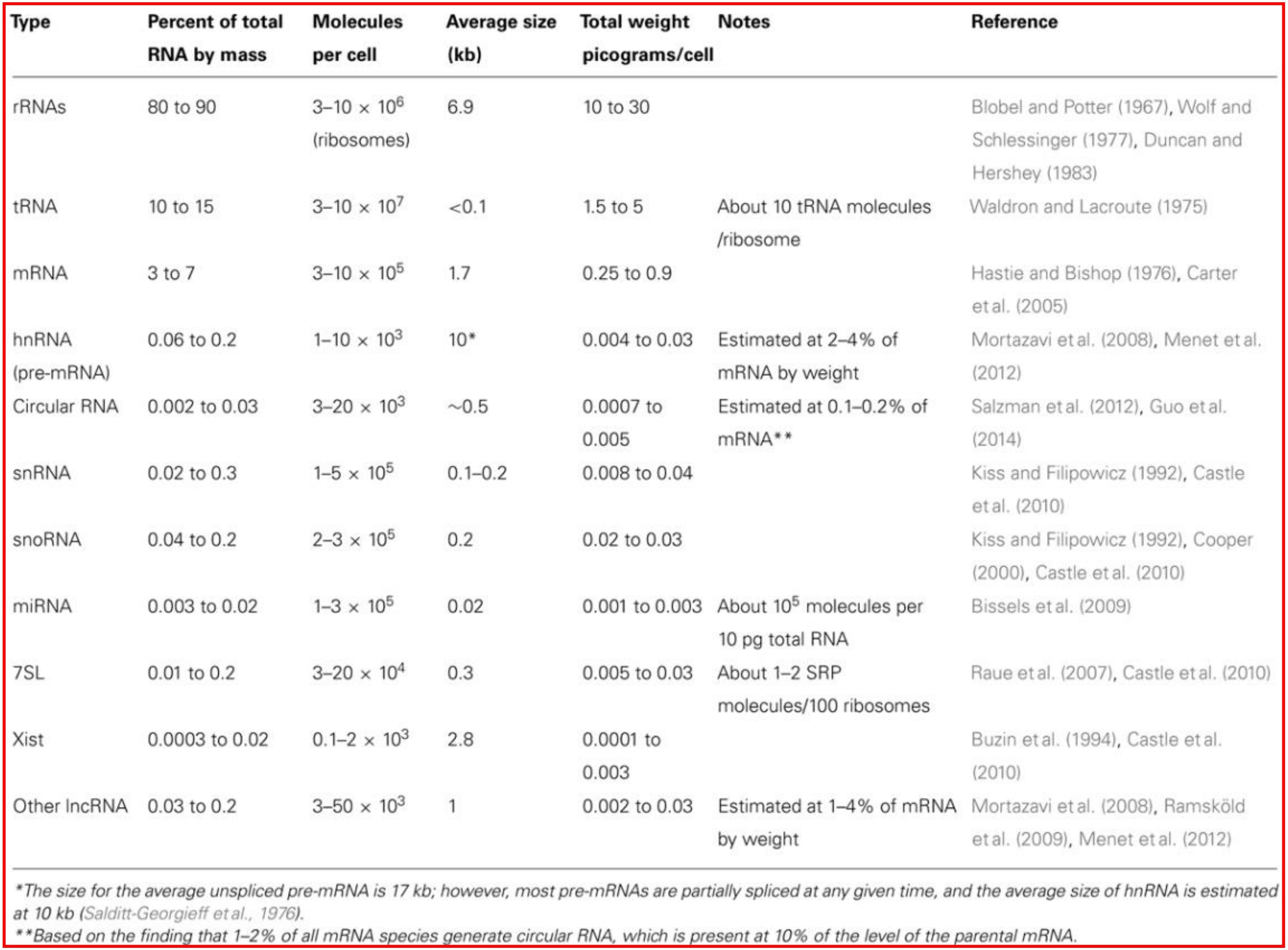
RNAome

Abbr.	Name
qiRNA	QDE-2 interfering RNA
rasiRNA	Repeat associated siRNA
RNase MRP	mitochondrial RNA processing ribonuclease
RNase P	ribonuclease P
scaRNA	small Cajal body-specific RNA
scnRNA	small-scan RNA
scRNA	small cytoplasmic RNA
scRNA	small conditional RNA
SgrS RNA	sugar transport-related sRNA
shRNA	short hairpin RNA

Abbr.	Name
siRNA	small interfering RNA
SL RNA	spliced leader RNA
SmY RNA	mRNA trans-splicing
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleic proteins
SPA lncRNA	5' small nucleolar RNA capped and 3' polyadenylated long noncoding RNA
SRP RNA	signal recognition particle RNA
vRNA	vault RNA
vtRNA	vault RNA

Abbr.	Name
ssRNA	single stranded RNA
stRNA	small temporal RNA
tasiRNA	trans-acting siRNA
tmRNA	transfer-messenger RNA
uRNA	U spliceosomal RNA
Xist RNA	X-inactive specific transcript
Y RNA	Y RNA
NATs	natural antisense transcripts
pre-mRNA	precursor messenger RNA
circRNA	circular RNA
msRNA	multicopy, single-stranded RNA
cfRNA	cell-free RNA

RNA distribution by mass and number (Pallazo, 2015)



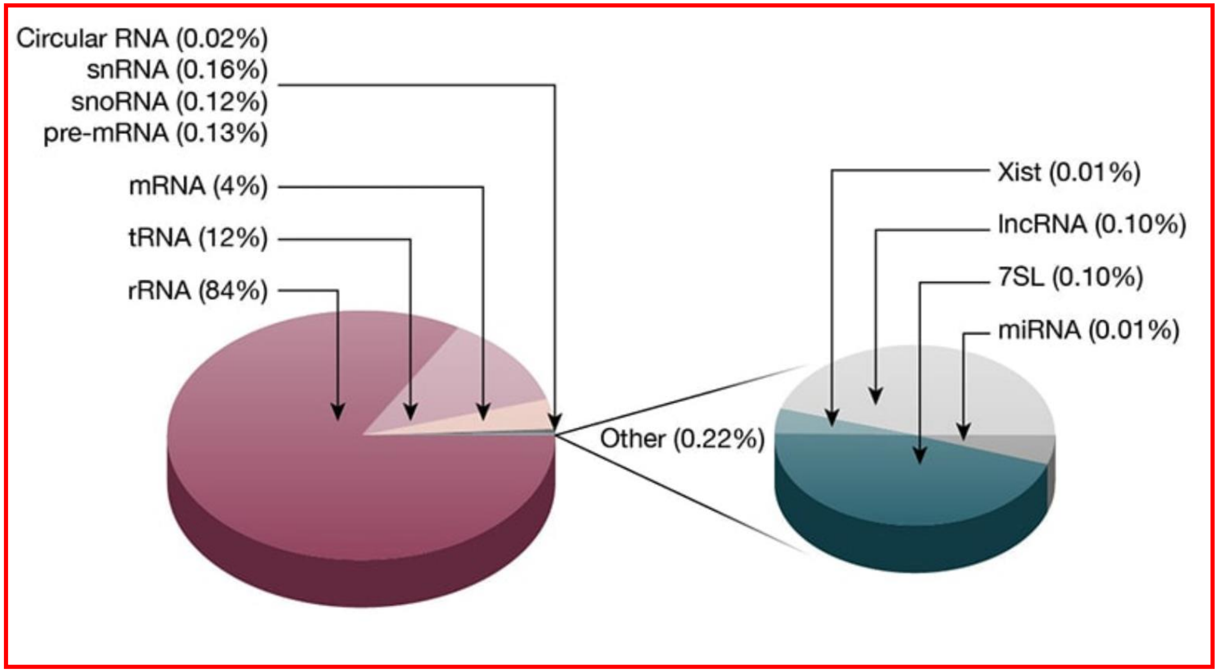
Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10^6 (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10^7	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10^5	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10^3	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10^3	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10^5	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10^5	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10^5	0.02	0.001 to 0.003	About 10^5 molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10^4	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10^3	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10^3	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Ramsköld et al. (2009), Menet et al. (2012)

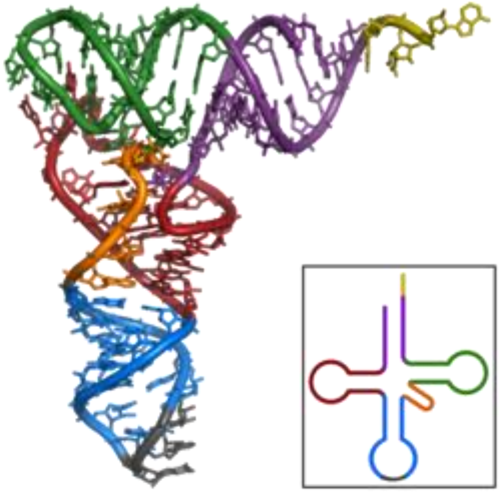
**The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1976).*

***Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.*

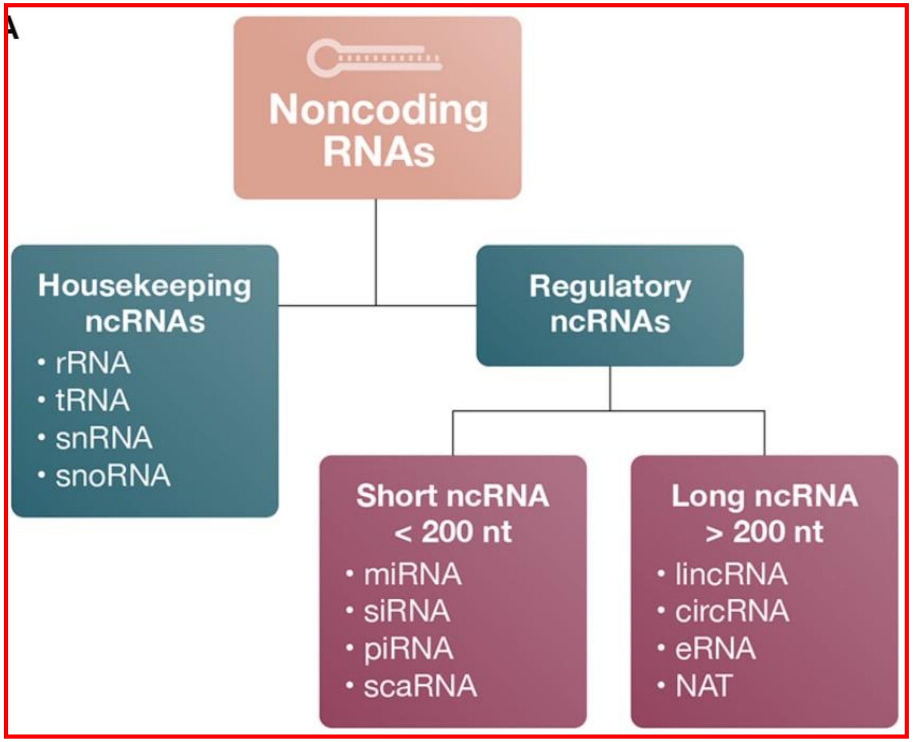
Indian Council of Agricultural Research

RNA classification





The cloverleaf structure of Yeast tRNA



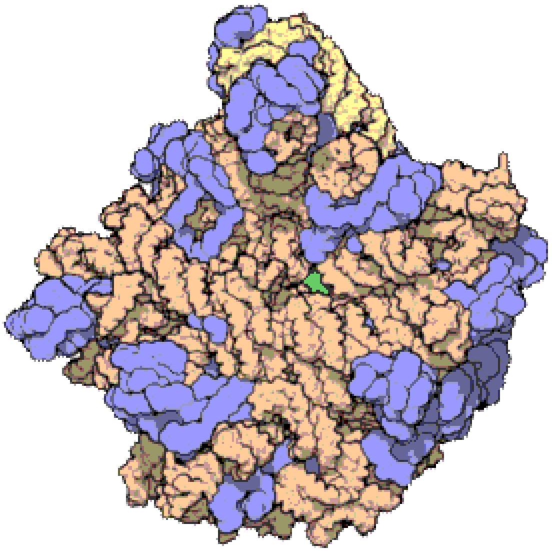
Noncoding RNAs

Housekeeping ncRNAs
- rRNA
- tRNA
- snRNA
- snoRNA

Regulatory ncRNAs

Short ncRNA < 200 nt
- miRNA
- siRNA
- piRNA
- scaRNA

Long ncRNA > 200 nt
- lincRNA
- circRNA
- eRNA
- NAT

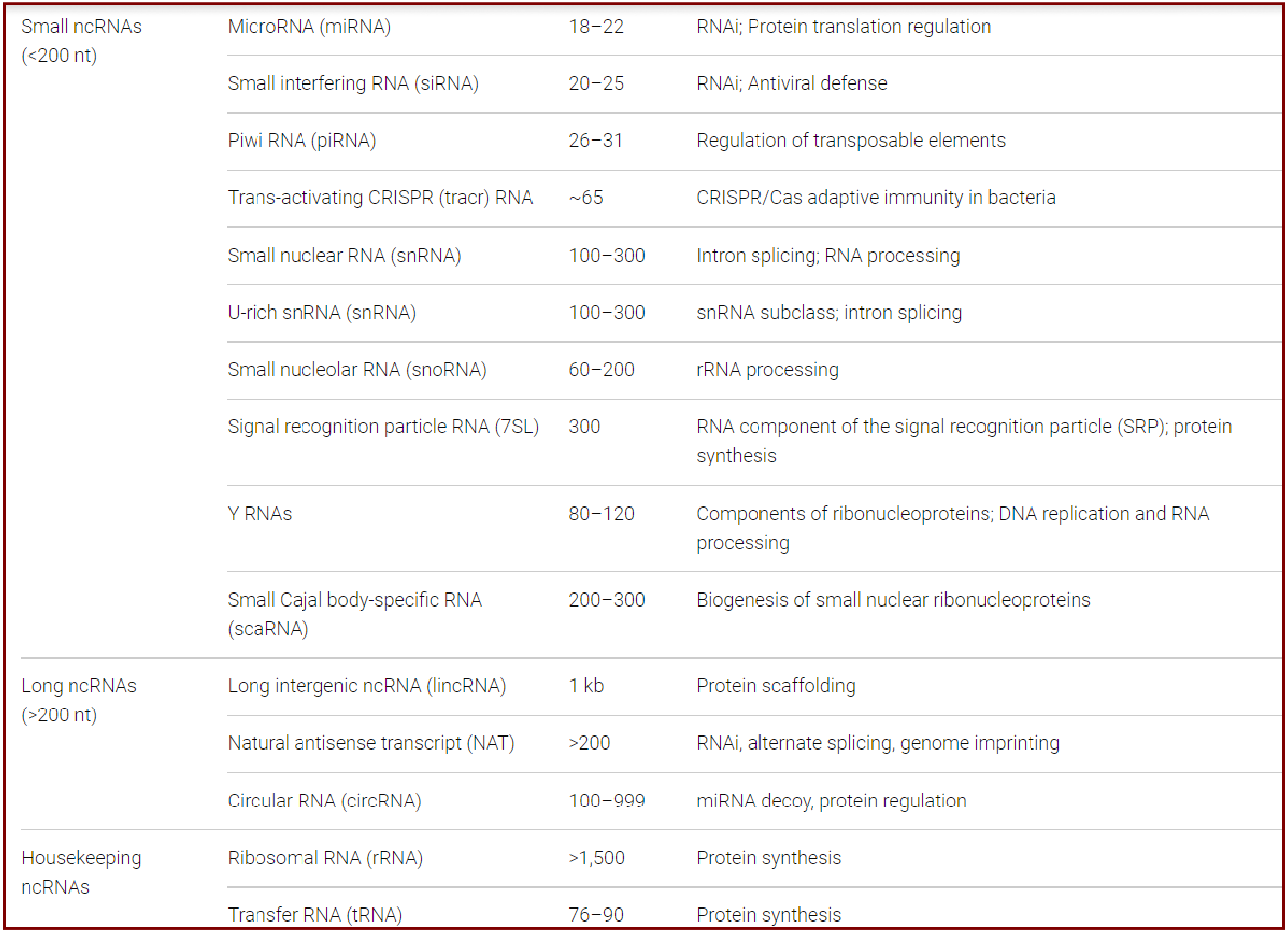


Ban *et al.* (2020), The complete atomic structure of the large ribosomal subunit. *Science.*

RNAs by role

- RNAs involved in protein synthesis
  - – mRNA, rRNA, tRNA
- RNAs involved in post-transcriptional modification or DNA replication
  - – snRNA, snoRNA, RNase P
- Regulatory RNAs
  - – miRNA, lncRNA, circRNA, siRNA, shRNA, piRNA
- Parasitic RNAs
  - – Viroid, Satellite RNARetrotransposon
- Other RNAs
  - – Vault RNA (expulsion of xenobiotics)

Type, Length and Function of Non-coding RNAs



Small ncRNAs (<200 nt)	MicroRNA (miRNA)	18–22	RNAi; Protein translation regulation
	Small interfering RNA (siRNA)	20–25	RNAi; Antiviral defense
	Piwi RNA (piRNA)	26–31	Regulation of transposable elements
	Trans-activating CRISPR (tracr) RNA	~65	CRISPR/Cas adaptive immunity in bacteria
	Small nuclear RNA (snRNA)	100–300	Intron splicing; RNA processing
	U-rich snRNA (snRNA)	100–300	snRNA subclass; intron splicing
	Small nucleolar RNA (snoRNA)	60–200	rRNA processing
	Signal recognition particle RNA (7SL)	300	RNA component of the signal recognition particle (SRP); protein synthesis
	Y RNAs	80–120	Components of ribonucleoproteins; DNA replication and RNA processing
	Small Cajal body-specific RNA (scaRNA)	200–300	Biogenesis of small nuclear ribonucleoproteins
Long ncRNAs (>200 nt)	Long intergenic ncRNA (lincRNA)	1 kb	Protein scaffolding
	Natural antisense transcript (NAT)	>200	RNAi, alternate splicing, genome imprinting
	Circular RNA (circRNA)	100–999	miRNA decoy, protein regulation
Housekeeping ncRNAs	Ribosomal RNA (rRNA)	>1,500	Protein synthesis
	Transfer RNA (tRNA)	76–90	Protein synthesis

Biofortification

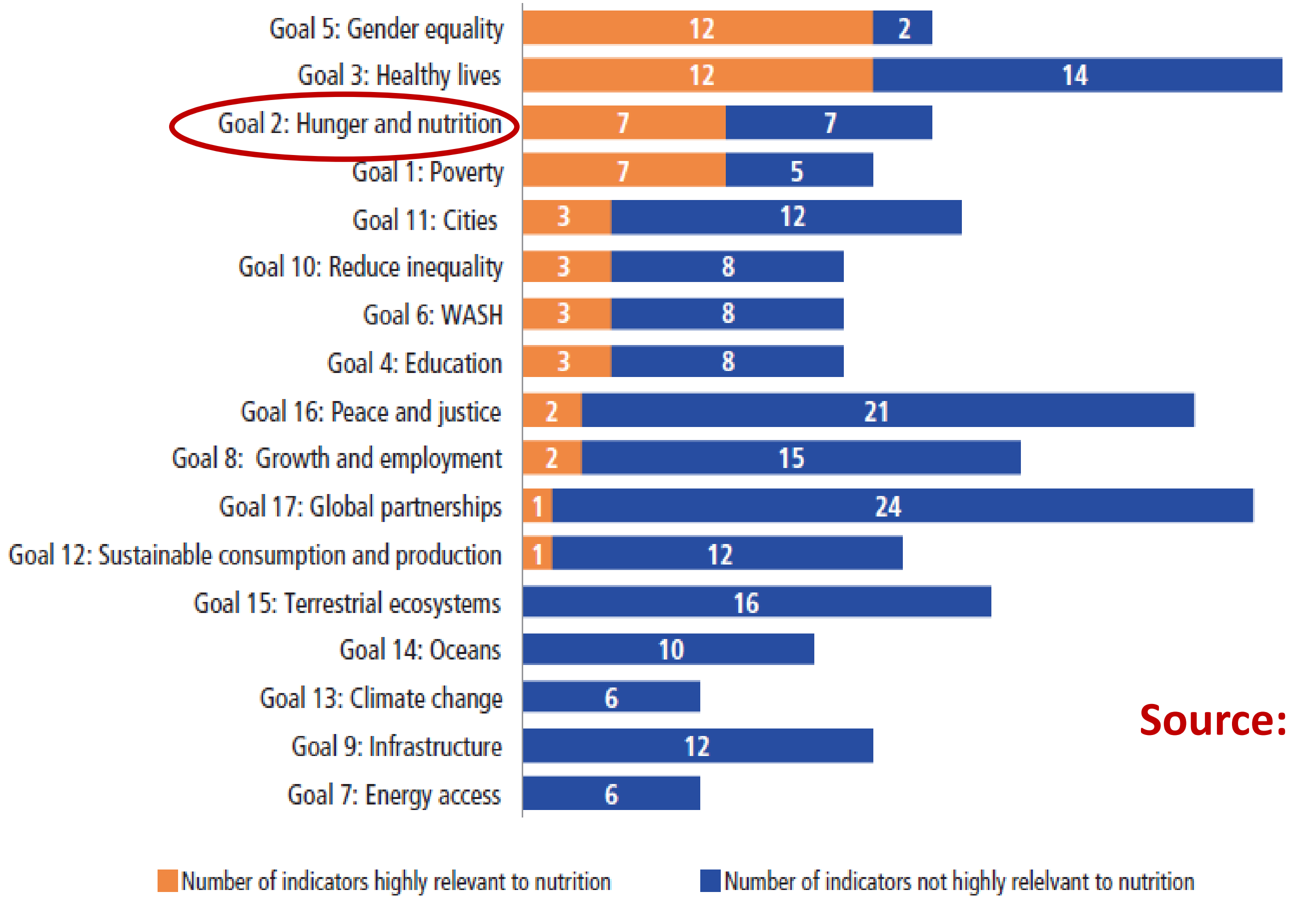
- Fortification – Nutrient addition at the time of food processing
- Biofortification - making plant / animal foods more nutritious as the plants / animals are growing

The golden color of the grains - increased amounts of beta-carotene

Biofortification – through conventional selective breeding or Genetic Engineering

Sustainable Development Goals (SDGs)

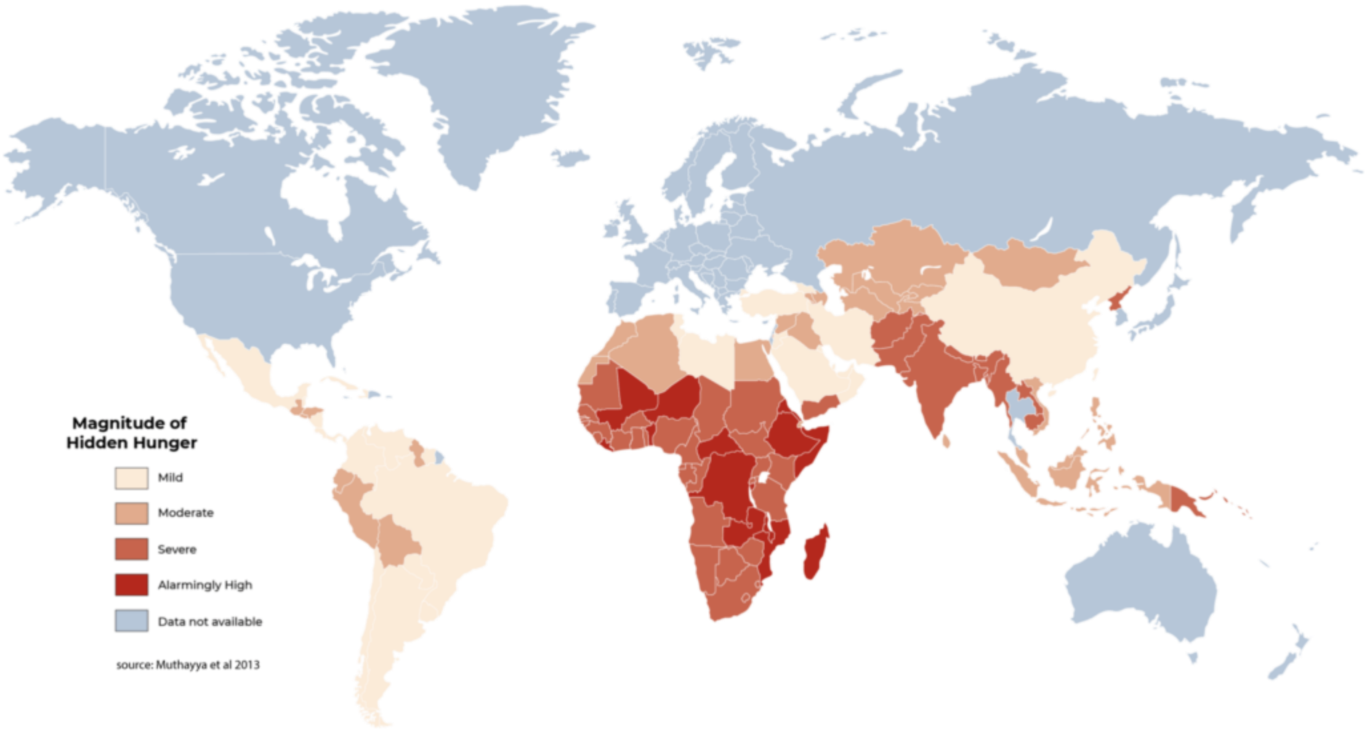
- Global community committed a set of objectives in 2015
- 17 goals anchor the global development agenda till 2030.
- Core is to eliminate extreme poverty, hunger, and malnutrition.
- 12 of the 17 goal-indicators related to nutrition



Source: IFPRI, 2016

Hidden Hunger

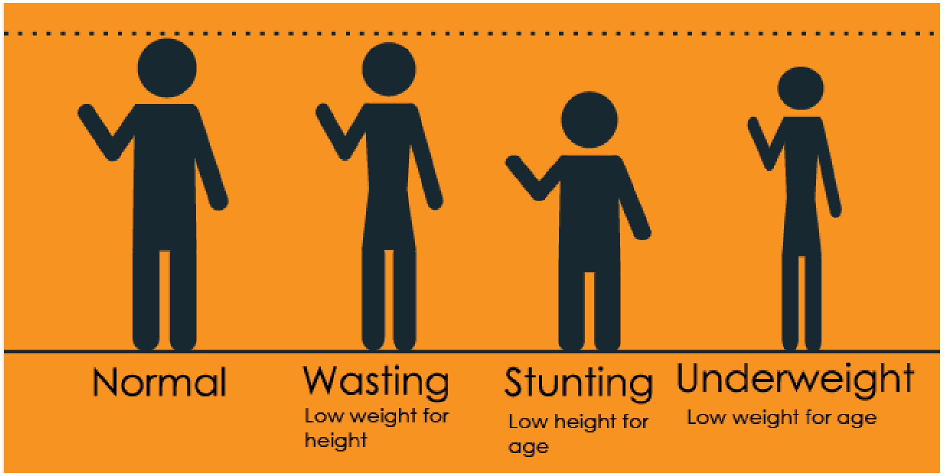
- Hidden hunger is a global health crisis, driven in large part by poverty.
- Can’t afford a diet of nourishing, diverse foods that provide enough essential vitamins and minerals (micronutrients).



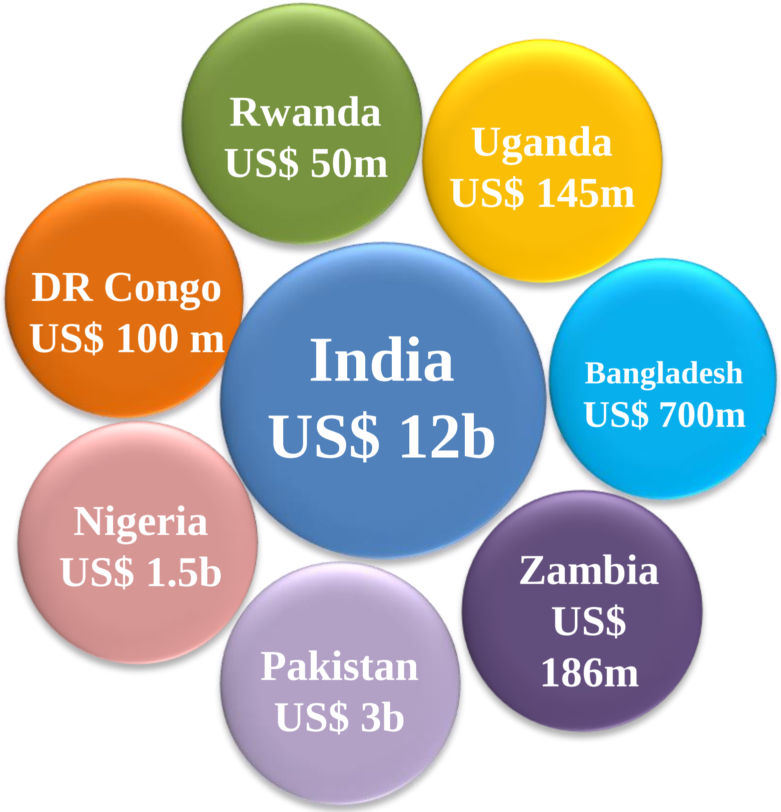
Source: https://www.harvestplus.org/home/biofortification-why-and-how/#familiar

<u>Malnutrition</u> refers to deficiencies, excesses, or imbalances in a person's intake of energy and/or nutrients.

- <u>Undernutrition</u>
  - wasting (low weight-for-height)
  - stunting (low height-for-age)
  - underweight (low weight-for-age);
- <u>Micronutrient-related malnutrition</u>
  - micronutrient deficiencies (a lack of important vitamins and minerals) or micronutrient excess
- <u>Overweight, obesity and diet-related noncommunicable diseases</u> (such as heart disease, stroke, diabetes and some cancers).



Loss in GDP due to malnutrition



Rwanda US$ 50m

Uganda US$ 145m

DR Congo US$ 100 m

India US$ 12b

Bangladesh US$ 700m

Nigeria US$ 1.5b

Zambia US$ 186m

Pakistan US$ 3b

Source: www.harvestplus.org

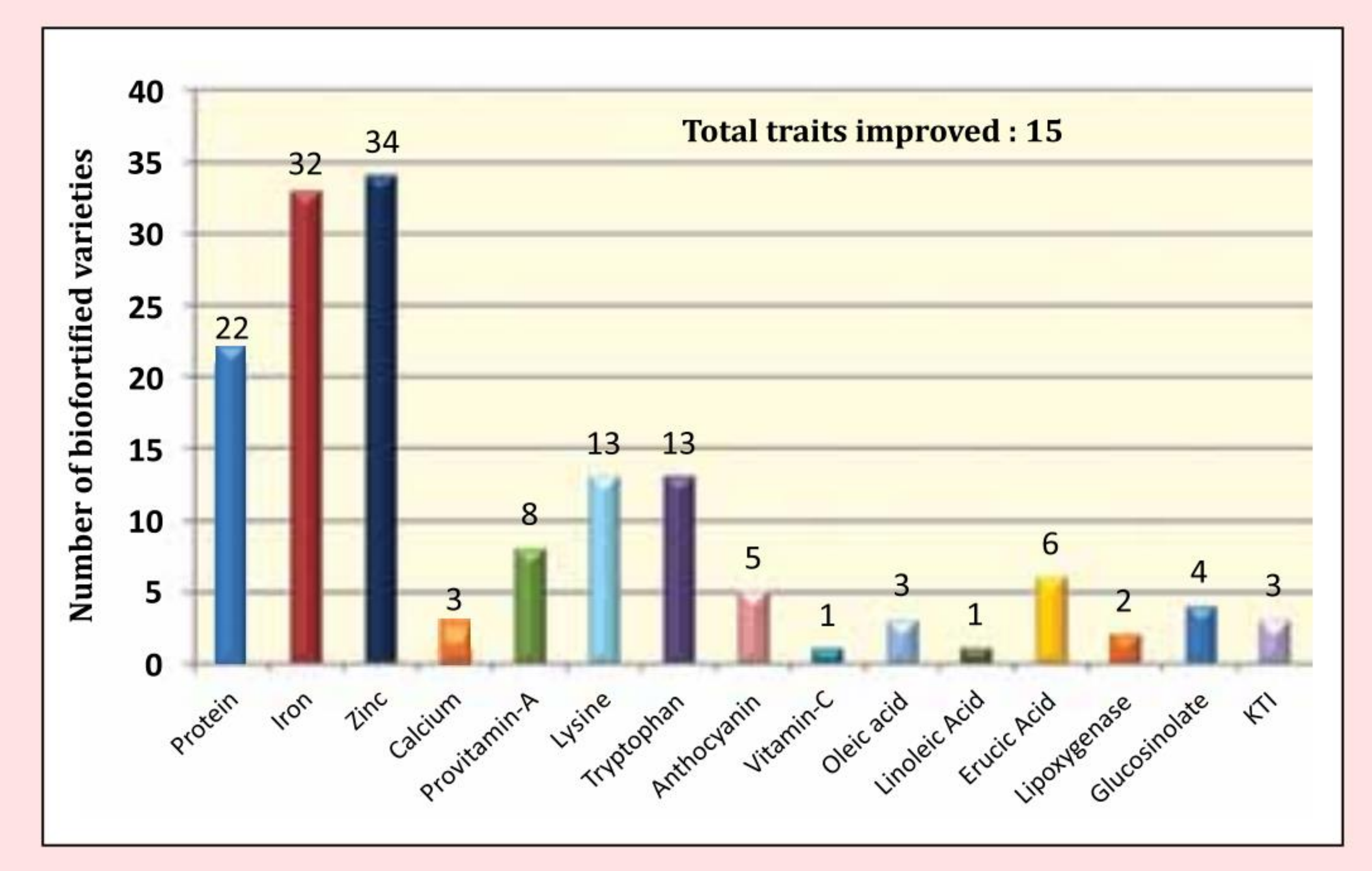
Estimated Loss

$1 invested in proven nutrition programme offers benefits worth $16 (IFPRI 2016)

Biofortification of plant and animal traits

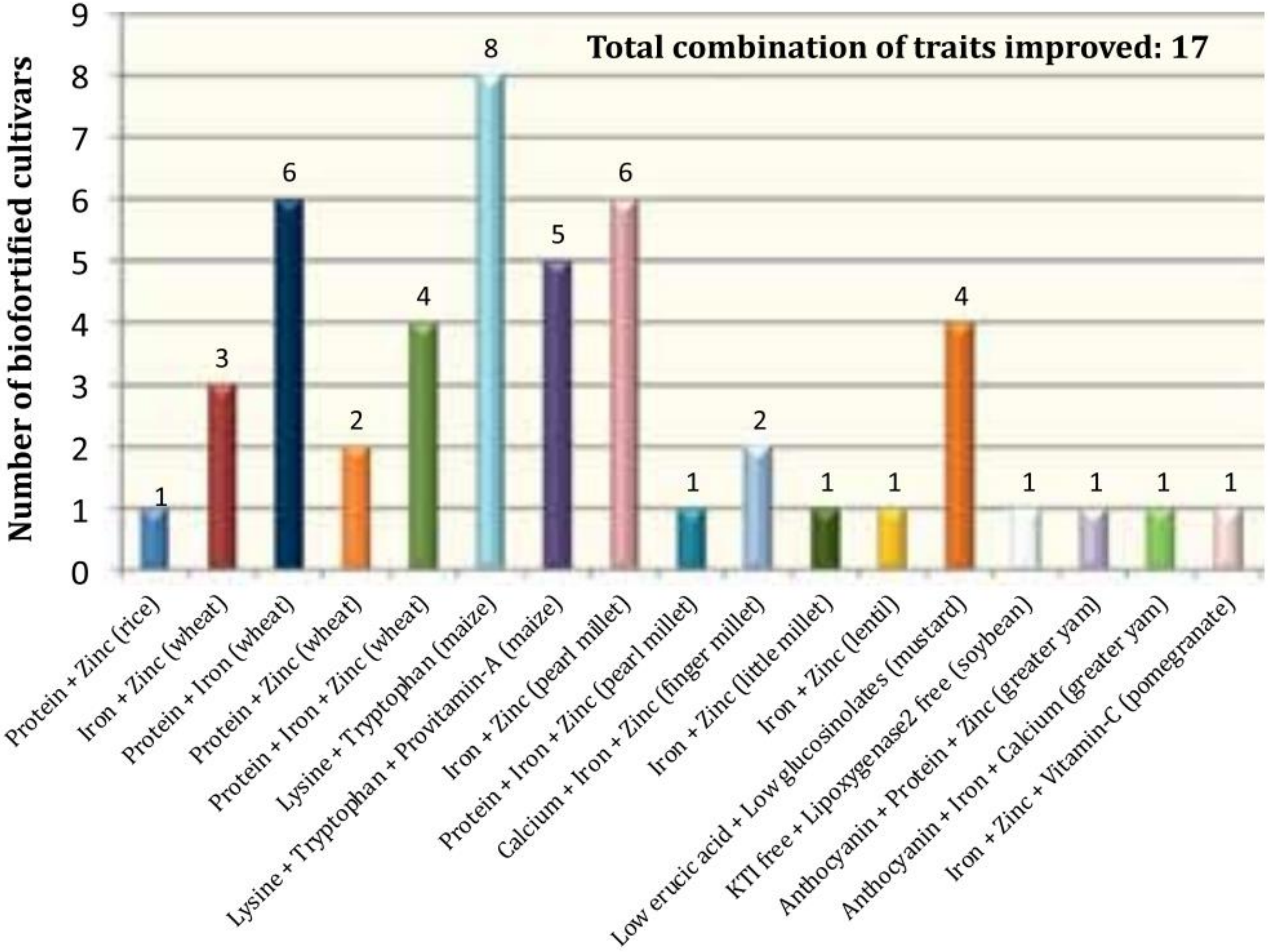
1. Biofortification: Sustainable way to eliminate malnutrition
2. Nutri-cereals and potential crops : The naturally biofortified crops
3. Antinutritional factors [Erucic acid, glucosinolates, kunitz tripsin inhibitor (KTI), Lipoxygenase] free varieties

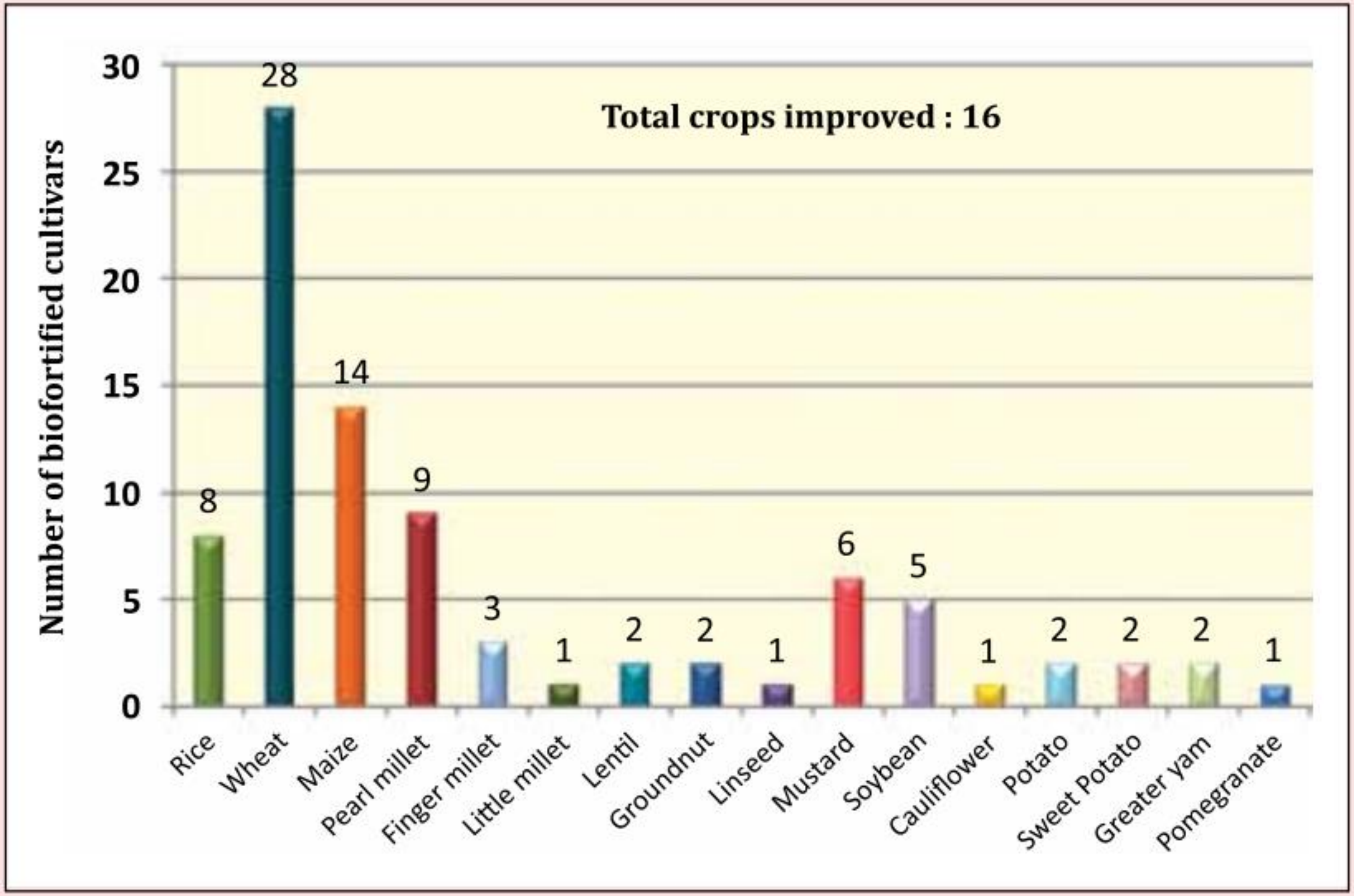
- 87 biofortified cultivars in 16 crops developed by ICAR
- Higher levels of Fe, Zn, protein, pro-vitamin A etc. in the edible parts besides reduced level of anti-nutritional factors.
- A total of 11282 quintals of breeder seed of 53 such varieties have been produced during 2016-17 to 2021-22 against indents by different agencies.



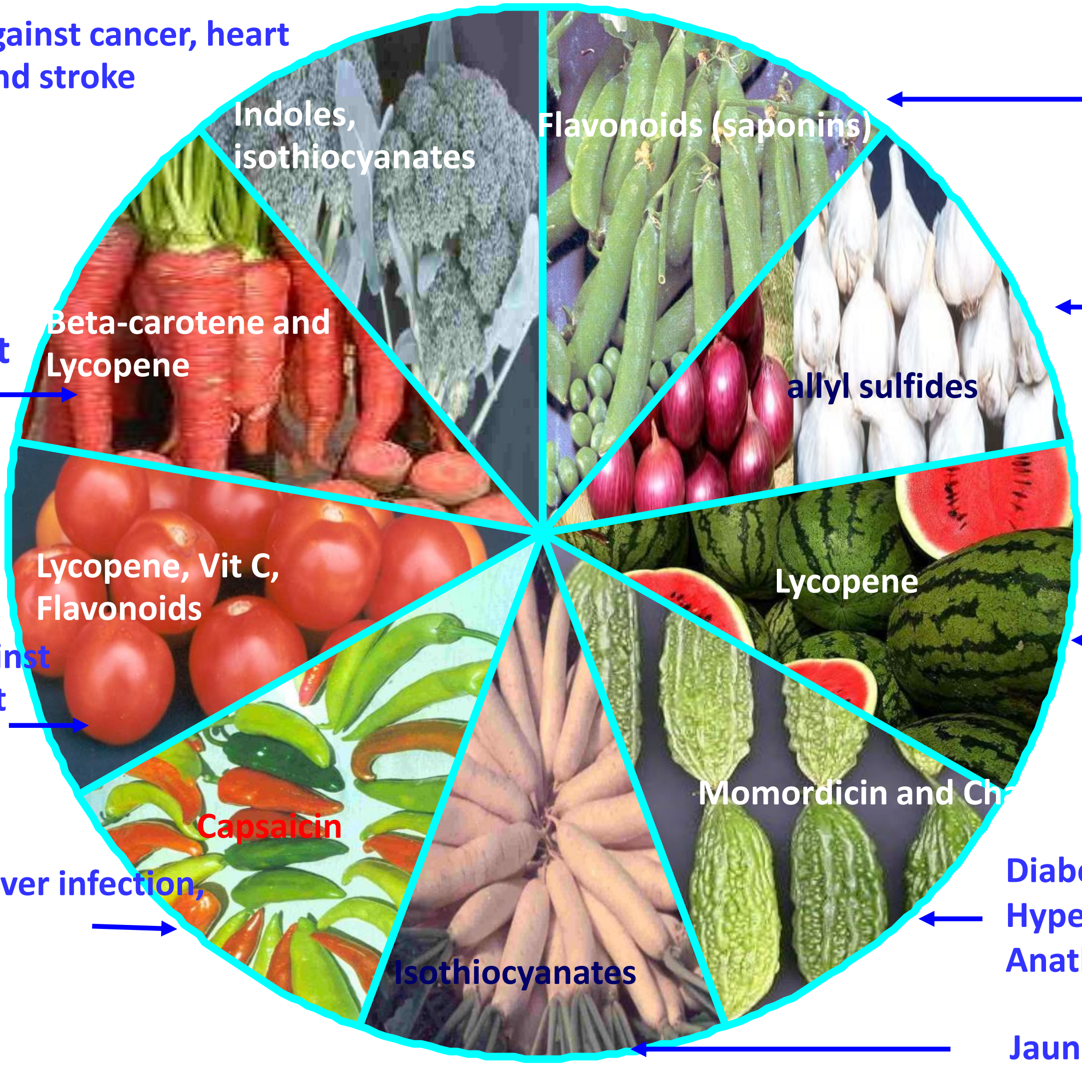
Yadav, *et al*. (2022). Biofortified varieties: Sustainable way to alleviate malnutrition. *ICAR publication*.

https://icar.org.in/file/15017/download?token=MinbP1kM#:~:text=ICAR%20has%20developed%2087%20biofortified,combined%20in%20a%20single%20genotype.


Number of biofortified cultivars
9
8
7
6
5
4
3
2
1
0
Total combination of traits improved: 17
1
3
6
2
4
8
5
6
1
2
1
1
4
1
1
1
1
Protein + Zinc (rice)
Iron + Zinc (wheat)
Protein + Iron (wheat)
Protein + Zinc (wheat)
Protein + Iron + Zinc (wheat)
Lysine + Tryptophan (maize)
Lysine + Tryptophan + Provitamin-A (maize)
Iron + Zinc (pearl millet)
Protein + Iron + Zinc (pearl millet)
Calcium + Iron + Zinc (finger millet)
Iron + Zinc (little millet)
Iron + Zinc (lentil)
Low erucic acid + Low glucosinolates (mustard)
KTI free + Lipoxygenase2 free (soybean)
Anthocyanin + Protein + Zinc (greater yam)
Anthocyanin + Iron + Calcium (greater yam)
Iron + Zinc + Vitamin-C (pomegranate)


Total crops improved : 16
Number of biofortified cultivars
30
25
20
15
10
5
0
8
28
14
9
3
1
2
2
1
6
5
1
2
2
2
1
Rice
Wheat
Maize
Pearl millet
Finger millet
Little millet
Lentil
Groundnut
Linseed
Mustard
Soybean
Cauliflower
Potato
Sweet Potato
Greater yam
Pomegranate

Nutraceuticals properties of horticultural crops



Protect against cancer, heart disease and stroke

Protect against cancer, lower cholesterol

Antioxidant

Protect against cancers and heart disease, boost the immune system

Protect against cancer, fight infection

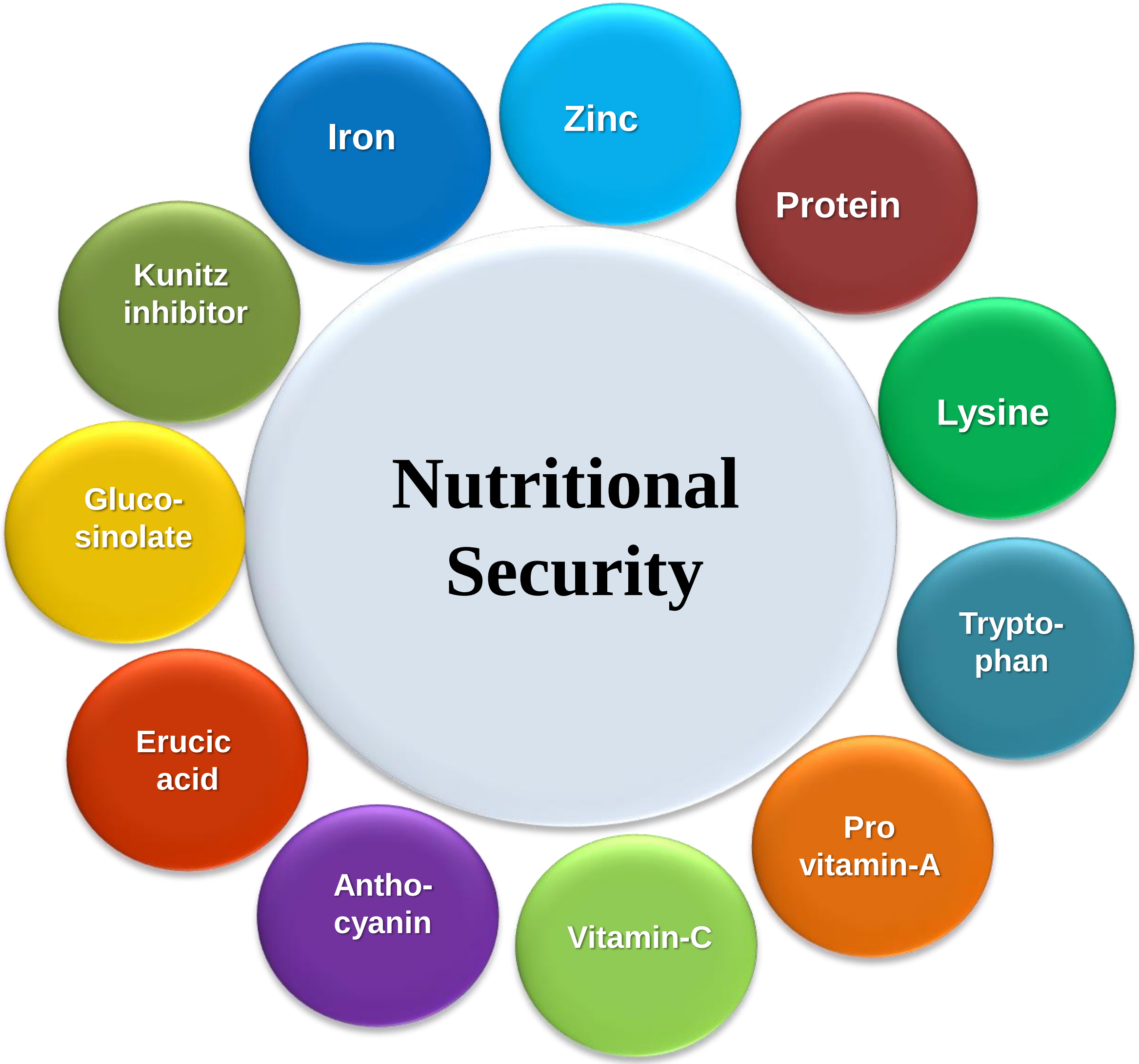
Protect against cancer

Jaundice, Liver infection, Piles

Diabetes, blood purifier, Hypertension, Dysentery, Anathematic

Jaundice, Liver infection, Piles

Biofortified varieties released: 87



Nutritional Security

Iron

Zinc

Protein

Lysine

Trypto-phan

Pro vitamin-A

Vitamin-C

Antho-cyanin

Erucic acid

Gluco-sinolate

Kunitz inhibitor

Wheat	: 28
Rice	: 08
Maize	: 14
Pearl Millet	: 09
Finger millet	: 03
Little millet	: 01
Lentil	: 02
Mustard	: 06
Soybean	: 05
Linseed	: 01
Potato	: 02
Cauliflower	: 01
Sweet Potato	: 02
Greater Yam	: 02
Pomegranate	: 01
Groundnut	: 02

Rice: DRR Dhan 49

- Adaptation: Gujarat, Maharashtra and Kerala
- Developed by ICAR-Indian Institute of Rice Research, Hyderabad

Zinc
25.2 ppm

Rice: CR Dhan-311 (Mukul)

- **Contains high protein (10.1%) and moderate level of Zn (20ppm)**

Normal rice: 6-7% protein

Low Glycaemic index rice varieties: Samba Mashuri, Sampada, Madhuraj 55, Promotion of brown rice, par boiled rice

Wheat: Pusa Tejas (HI 8759) durum

Protein 12.0 %

Iron 41.1 ppm

Zinc 42.8 ppm

- ➢ Adaptation: Madhya Pradesh, Chhattisgarh, Gujarat, Kota and Udaipur Division) and Uttar Pradesh (Jhansi Division) Pradesh
- ➢ Developed by ICAR-Indian Agricultural Research Institute, Regional Station, Indore

Wheat: WB2

- High Zn (42 ppm) and Fe (40 ppm) with 12.4% protein.
- Average seed yield: 51.6 q/ha
- Recommended for irrigated timely sown conditions of North Western Plains Zone

First high vitamin-A maize hybrid

Pusa VQ9 improved: QPM + ProA

- Northern Hill Zone (NHZ): J&K, HP, Uttarakhand (Hills) & NEH states
- Peninsular Zone (PZ): Maharashtra, Karnataka, AP, Telengana &TN

Characters	P-VQ9-I
Provitamin-A (ppm)	8.15
% tryptophan in protein	0.74
% lysine in protein	2.67
Avg. grain yield-NHZ (kg/ha)	5588
Potential yield-NHZ (kg/ha)	7968
Avg. grain yield-PZ (kg/ha)	5916
Potential yield-PZ (kg/ha)	9368
Duration:	Early

Normal maize: 0-2 ppm provitamin-A
Lysine: 2%, Tryptophan: 0.4%

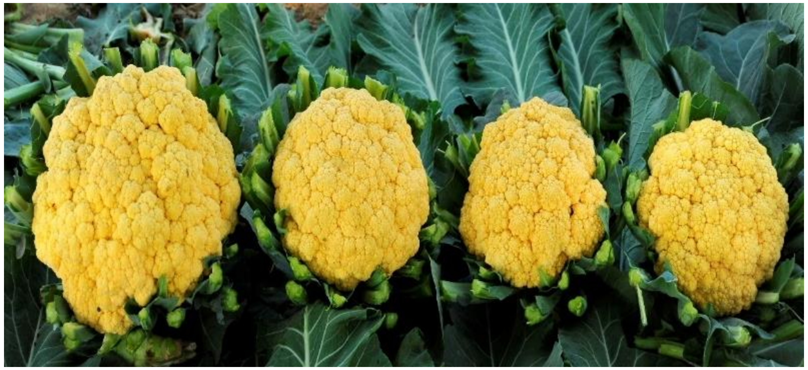
Biofortified Lentil Varieties

Indian Council of Agricultural Research

Pusa Ageti Masoor	IPL 220
• Rich in **iron (65.0 ppm) in comparison to 45.0-50.0 ppm** in popular varieties • Grain yield: 13.0 q/ha • Maturity: 100 days • Suitable for rainfed condition • Adaptation: Utter Pradesh, Madhya Pradesh, Chhattisgarh • Developed by ICAR-Indian Agricultural Research Institute, New Delhi	• Rich in iron **(73.0 ppm) and zinc (51.0 ppm)** in comparison to 45.0-50.0 ppm iron and 35.0-40.0 ppm zinc in popular varieties • Grain yield: 13.8 q/ha • Maturity: 121 days • Suitable for rainfed condition • Adaptation: Eastern Uttar Pradesh, Bihar, Assam and West Bengal • Developed by ICAR-Indian Institute of Pulses Research, Kanpur

Pusa KesariVitA-1: First ever bio-fortified beta carotene rich cauliflower variety

- It contains 8-10 ppm beta carotene, orange, coloured, compact and very attractive curd.
- It is suitable for September – January growing period.
- Average marketable curd weight is about 1.250 kg with an approximate marketable yield of 42.0 – 46.0 t/ha.
- Important attempt to tackle beta carotene deficiency related malnutrition problem in India.



Biofortified Pomegranate

Solapur Lal (NRCP H-6):

- Fe content 5.6 mg/100 g fresh arils, which is double of Bhagwa
- Zn content 0.67 mg/100 fresh arils against 0.50mg/100 g fresh arils of Bhagwa.
- Vitamin C content is 19.5 mg/100 which is higher than Bhagwa (14.2-14.6 mg/100g)
- Average fruit yield of this variety is 23-27 t/ha in comparison to 16-20 t/ha of Bhagwa.
- Semi-arid regions of India.

4. Nutri-cereals and potential crops: Naturally biofortified crops for POSHAN Abhiyan

Major millets

Pearl millet (*Pennisetum americanum*)

Great millet / sorghum (*Sorghum bicolor*)

Finger millet (*Eleusine coracana*)

Foxtail millet (*Setaria italica*)

Small millets

Little millet (*Panicum sumatrense*)

Kodo millet (*Paspalum scrobiculatum*)

Proso millet (*Panicum miliaceum*)

Barnyard millet (*Echinochloa frumentacea*)

High Fe pearl millet hybrids

Varieties	Nutritive value	Adaptation zone/ state	Season of cultivation	Grain yield
Dhanshakti/ ICTP 8203Fe	Fe: 71 ppm, Zn: 40 ppm	Maharashtra, Karnataka, Telangana, Uttar Pradesh, Haryana and Rajasthan.	Kharif	2.21 t/ha
Shakti-1201/ ICMH 1201	Fe: 75 ppm, Zn: 40 ppm	Maharashtra and Rajasthan.	Kharif	3.60 t/ha
HHB-299	Fe: 73 ppm, Zn: 41 ppm	Haryana, Rajasthan, Gujarat, Punjab, Delhi, Maharashtra, TN	Kharif	3.27 t/ha
AHB 1200	Fe: 73 ppm,	Haryana, Rajasthan, Gujarat, Punjab, Delhi, Maharashtra, TN	Kharif	3.00 t/ha

Normal pearl millet: 45.0-50.0 ppm iron and 30.0-35.0 ppm zinc in popular varieties

High Fe small millets

Crop	Fe
Finger millet	GPU-28: 69.9 ppm **High > 90 ppm:** KMR-216, BR-36 & PR-10-21
Foxtail millet	SiA 3088: 129 ppm **High > 140 ppm:** SiA 3142 & TNAU-186
Little millet	OLM-203: 51 ppm **High > 250 ppm:** BL-4, RLM-186, TNAU-63 & JK-8

Millets are Nutricereals-Their composition across individual millets vis-a-vis rice and wheat

Grain (Millet /Cereal)	Carbo-hydrates (g)	Protein (g)	Fat (g)	Energy (Kcal)	Dietary Fibre (g)	Ca (mg)	Mg (mg)	Zn (mg)	Fe (mg)	Thiamin (mg)	Riboflavin (mg)	Niacin (mg)	Folic acid (µg)
Sorghum	67.7	10.0	1.7	334.1	10.2	27.6	133.0	2.0	4.0	0.4	0.1	2.1	39.4
Pearl Millet	61.8	11.0	5.4	348.0	11.5	27.4	124.0	2.8	6.4	0.3	0.2	0.9	36.1
Finger millet	66.8	7.2	1.9	320.7	11.2	364.0	146.0	2.5	4.6	0.4	0.2	1.3	34.7
Kodo millet	66.2	8.9	2.6	331.7	6.4	15.3	122.0	1.7	2.3	0.3	0.2	1.5	39.5
Proso millet*	70.4	12.5	1.1	341.1	-	14.0	153.0	1.4	0.8	0.4	0.3	4.5	-
Foxtail millet*	60.1	12.3	4.3	331.0	-	31.0	81.0	2.4	2.8	0.6	0.1	3.2	15.0
Little millet	65.6	10.1	3.9	346.3	7.7	16.1	91.4	1.8	1.3	0.3	0.1	1.3	36.2
Barnyard millet*	65.6	6.2	2.2	307.1	-	20.0	82.0	3.0	5.0	0.3	0.1	4.2	-
Wheat	64.7	10.6	1.5	321.9	11.2	39.4	125.0	2.9	4.0	0.5	0.2	2.7	30.1
Rice	78.2	7.9	0.5	356.4	2.8	7.5	19.3	1.2	0.7	0.1	0.1	1.7	9.3

5. Antinutritional factors free varieties

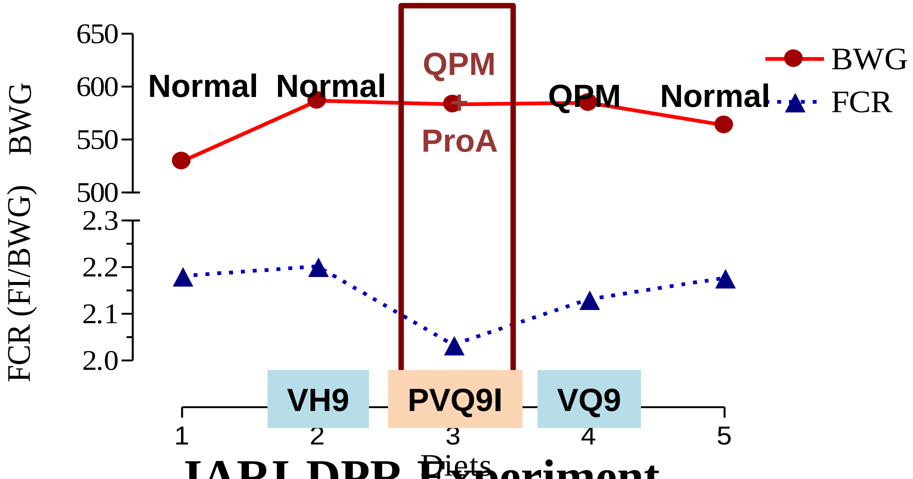
Pusa Double Zero Mustard-31

- Average seed yield 23.8 q/ha
- Oil content 40.56%
- Erucic acid (<2.0%)
- Glucosinolates (<30.0 ppm)
- First Canola type Indian mustard variety in India

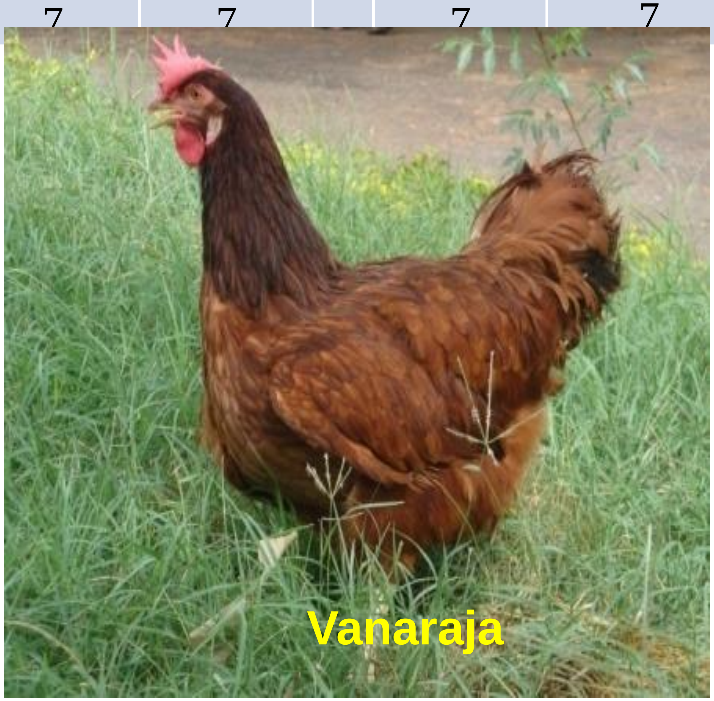
Impact: Effects on chicken

Diet	3 week			6 week			9 week	
	BWG	FCR		BWG	FCR		BWG	FCR
Diet 1 Yellow Maize	205.3	2.023		529.7	2.181^{ab}		909.2	2.361
Diet 2 Vivek Hyb 9	213.0	1.979		586.9	2.202^{a}		956.1	2.404
Diet 3 Pusa VQ9 Improved	210.6	1.980		**583.4**	**2.034^{c}**		**972.4**	**2.339**
Diet 4 Vivek QPM 9	186.6	2.182		584.8	2.131^{bc}		960.7	2.357
Diet 5 White Maize	191.4	2.096		563.6	2.177^{ab}		962.0	2.252
N	7	7		7	7		7	7

Effect of feeding different source of maize on performance in Vanaraja birds during 6 weeks of age



IARI-DPR Experiment



Impact: Effects on egg

Moreno et al. 2016

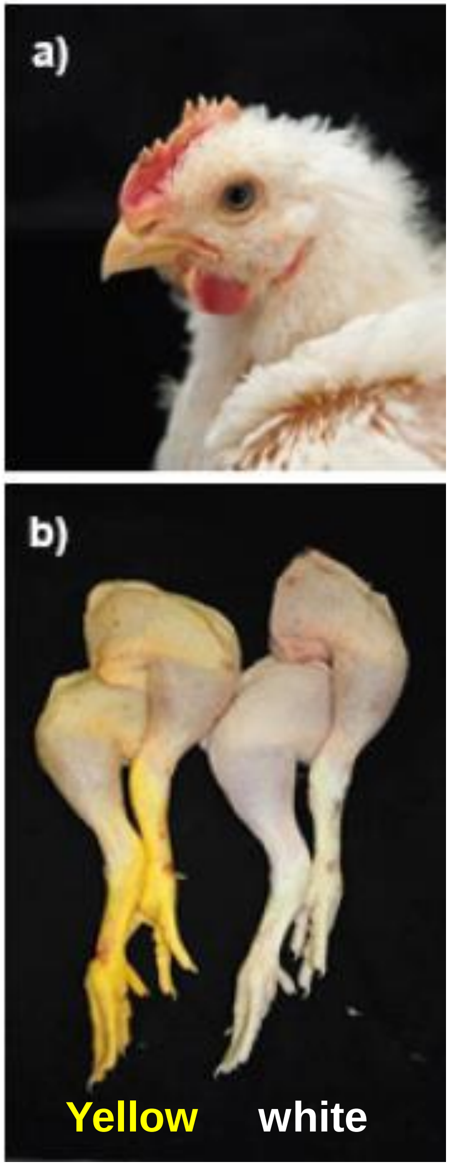
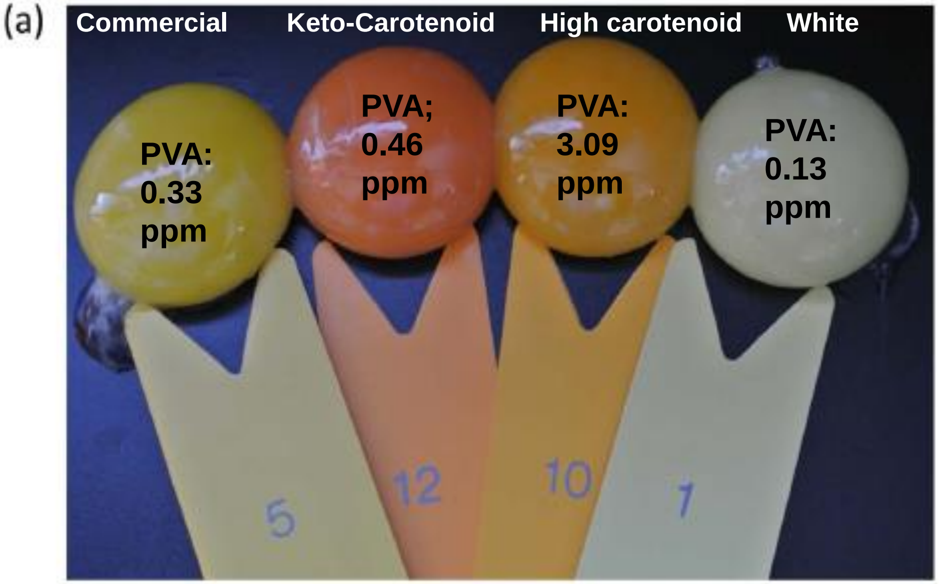
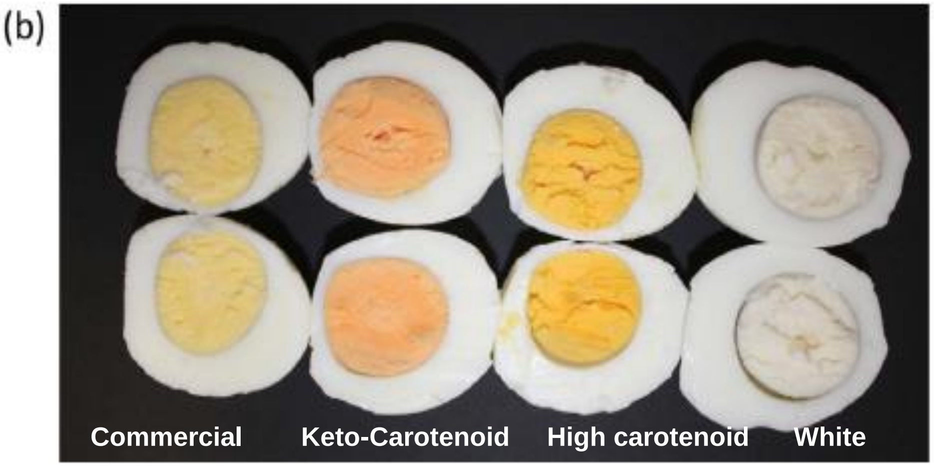


Figure 1. Chickens fed the high-carotenoid corn (left) and the control (right) diets. (a) Beak, crest, eyelids and facial feathers. (b) Dissected thighs.

Díaz-Gómez et al. 2015



Keto-carotenoid: astaxanthin + violaxanthin + β-carotene



Appearance of egg yolks (a) raw and (b) cooked by boiling for 10 min produced from hens

Up-scaling of bio-fortified varieties

Crop	Name of variety	Name of the companies
Rice	DRR Dhan 45	(i) Max Yield Bio Gene (India) Pvt. Ltd.
	CR Dhan 310	(i) Areia Agrotech Pvt. Ltd.
Wheat	DBW 173	54 private seed companies and FPOs
Mustard	Pusa Mustard 30	(i) Malwa Enterprises, Punjab (ii) Arpan Seeds Pvt. Ltd., Rajasthan (iii) Ananya Seeds Pvt. Ltd., Delhi (iv) Ajeet Seeds, Aurangabad (v) Dinkar Seeds, Ahmedabad
	Pusa Double Zero Mustard 31	(i) Dinkar Seeds, Ahmedabad (ii) Patanjali

Crop	Name of variety	Breeder Seed Produced (q)			
		2016-17	2017-18	2018-19	Total
Total		370.0	1208.1	1907.7	3483.8

DAC&FW: Seed production and distribution to various agencies
Govt. schemes: Millet Mission, Seed Hubs, Cluster demos etc.

ncRNAs

- ncRNAs play an important role in gene regulation, chromosomal structure, genome defense, translation, splicing, DNA replication, etc.

- ncRNAs show abnormal expressions in disease tissues

- Liu, *et al*. (2015) explained the biological role played by the long non-coding RNAs (lncRNAs) [Genomics, Proteomics and Bioinformatics, 13, 137-147]

- lncRNAs can act as miRNA target mimics, wherein decoy RNAs bind the miRNAs and stops the interaction between miRNA and its targets

- Zhu *et al*. (2014) identified lncRNAs in *Arabidopsis thaliana* induced by *Fusarium oxysporum* infection

ncRNA

- non-protein coding share of the genome is involved in gene expression, chromatin modification, cell proliferation and in a wide range of diseases [Beena, 2014, RNA and Disease, 1: e355. doi: 10.14800/rd.355]
- LNCipedia [Volders, *et al.* (2015), NAR, 43, database issue]
- Li *et al.* (2014) – PLEK – prediction of lncRNAs based on an improved k-mer scheme [BMC Bioinformatics, 15, 311]

Overall Conclusion

- Genome/transcriptome level unravel of hidden mechanisms behind biofortification in plant and animal traits needs more attention.
- Tissue specific as well as over all transcriptome level exploration of regulatory molecules is essential
- Attention be given on eTMs and interactions among ncRNAs
- Biofortified cultivars of crops and breeds of animals / animal products be analyzed.
- Meet out SDG goals by alleviating malnutrition