UNIT 4

FUNCTIONAL GENOMICS

Structure

4.1 INTRODUCTION

Life has started through interplay between the inorganic matters that are organised together in molecules and interact among themselves to establish a complex network. Biological systems are complex, multicellular, modular and hierarchical structures that are composed of biomolecules like DNA (deoxyribonucleic acid), RNA, lipids, proteins and metabolites that are responsible for carrying out several chemical reactions and regulating the structure of cells. The genetic information encoded in DNA and its differential expression in particular time and space defines the organism, their homoeostasis and response to the environment. Advances in the technological approach in the past few years have allowed researchers to gather information on a larger scale to model biological systems. The remarkable discovery of the DNA structure, transcription of genetic information into messenger RNA (mRNA) and its subsequent translation into proteins have provided thorough knowledge of the biological systems starting from cells, tissues, organs to whole organism. The global networks are unravelled through integrating different approaches like molecular, biochemical and genetic so as to generate omics data with their subsequent analysis. 'Omics' study gathers data at the cellular to whole organism level through the involvement of different approaches of genomics, transcriptomics, proteomics, metabolomics, and interactomics.

Genomics is the study of complete set of genetic information, including sequence, structure and content of the genome, particularly focusing on the number, structure, function and organisation of genes within the genome. The dynamic aspects of gene expression are however explained through functional genomics which provides an overview of the relationship between the genome of an organism and its phenotype. Functional genomics is an integrative approach that defines the function and interaction of genes and their products by making use of high-throughput methods. The elucidation of genomic function at a global scale provides a comprehensive account of the gene and protein expressions, structure and regulation by combining different fields of molecular biology and cell biology. The data obtained from processes involving DNA sequenceing, gene and protein expression and function, protein translation, interaction between protein-DNA, protein-RNA and protein-protein decode the data into a model dynamic. These interactive networks are responsible for regulating gene expression, cell cycle progression and cell differentiation processes.

Functional genomics, in broad sense targets to elucidate how biological information encoded in the DNA sequence regulates the phenotype of an organism. The translation of genotypes into phenotypes is based on the concept given by Francis Crick in the 1950s as "Central dogma of molecular biology" that involves the flow of genetic information from DNA to RNA to Protein (Fig. 4.1). Initially, the biological information stored within the genetic material is transcribed into the transcriptome that is composed of RNA molecules obtained through protein-coding genes of which the stored information is required at a specific time by the cells. Further, the proteome that is a collection of proteins in an organism that precisely defines a cell's biological makeup is mediated by these RNA molecules. The proteome can be defined as the sum of signalling and metabolic networks.

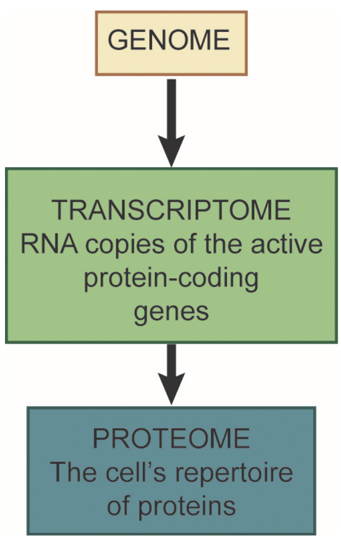

Fig. 4.1: Schematic representation for flow of genetic information from genome to transcriptome, a collection of RNA molecules that direct the synthesis of final production of gene expression, proteome

Objectives

After reading this unit, you would be able to:

- ❖ describe genomics and functional genomics,
- ❖ explain the concept of transcriptome,
- ❖ understand the functions of coding and non-coding RNAs,
- ❖ comprehend the concept of proteome and different characteristics of protein, their structure, chemical nature and functional diversity,
- ❖ discuss the methods to assess transcriptome,
- ❖ comprehend the applications and limitations of hybridisation-based and sequence-based approaches to study transcriptome, and
- ❖ enumerate the applications of transcriptome and omics studies.

4.2 CONCEPT OF TRANSCRIPTOME AND PROTEOME

4.2.1 Transcriptome

The biological information is stored in the genome but the cells cannot directly get the information from genome. The release of information to the cell through utilizing the biological information is a series of complex biochemical reactions that involves the coordinated interplay between the enzymes and other proteins, collectively known as genome expression. Transcriptome is the initial product of genome expression and refers to the set of all RNA molecules that are produced by the genome in a specific cell under specific conditions studied using high-throughput methods. The total RNA content of cell involves both coding and non-coding RNA transcripts with protein coding messenger RNA (mRNA) constituting 1-4% and non-coding ribosomal RNA (rRNA), transfer RNA (tRNA), long noncoding RNA (lncRNA), primary transcripts (pri-miRNA) and many others constituting about >95% of the total RNAs. Non-coding RNAs are not translated into proteins but are involved in several other functions such as assembly of ribosomes that is mediated by rRNAs; tRNAs serve as transporters while others play regulatory roles and are often found to be involved in causing several diseases such as cancer, neurological and cardiovascular disorders. Transcriptome often refers to mRNA, however it also provides information about other RNA content of the cell and therefore applied to the complete set of RNA molecules from protein coding messenger RNA (mRNA) to non-coding ribosomal RNA (rRNA), transfer RNA (tRNA), long noncoding RNA (lncRNA), primary transcripts (pri-miRNA) and many others.

The coding and non-coding RNAs are listed below and is represented in Figure 4.2.

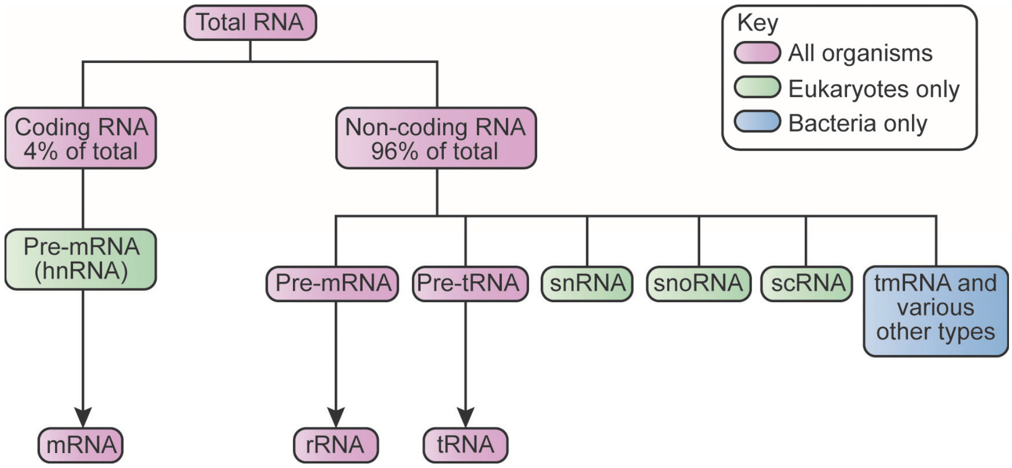

Fig. 4.2: The scheme showing total RNA content of a cell present in eukaryotes, bacteria and archaea. Some categories consist of RNAs that are present in eukaryotes only and some are present in bacteria only

Coding RNAs

Messenger RNAs (mRNAs)

- These are the transcripts derived from protein-coding genes that are translated into the final functional product, protein at the latter stages of genome expression.
- These RNA constitute > 4% of the total RNA content of the cell.
- They are short-lived and undergo degradation soon after synthesis.
- In bacteria, the half-lives of mRNAs are about a few minutes, while in eukaryotes mRNAs undergo degradation after a few hours of synthesis.
- The composition of transcriptome is therefore not fixed and it can be restructured through the change in synthesis rate of individual mRNAs transcript.

Non-Coding RNAs

These are the more diverse RNAs as compared to coding RNAs and are involved in carrying out different functions. The two non-coding RNA mainly present in both prokaryotes and eukaryotes are:

Ribosomal RNAs (rRNAs):

- These RNAs are abundantly present in the cell and constitutes about 80% of the total RNA.
- These are the primary component of protein-coding organelle, ribosomes and are actively involved in the synthesis of protein.

Transfer RNAs (tRNAs):

- These RNAs play role as adaptors during translation process.
- They carry amino acids to the ribosome and ensure that they are linked in the order as in the nucleotide sequence of the mRNA and therefore serves as a link between the mRNA and the growing chain of amino acids.

Other Non-coding RNAs

In this category, the non-coding RNAs are limited in their distribution. Eukaryotes contain a variety of short non-coding RNAs that can be listed as:

Small cytoplasmic RNA (scRNA):

- These RNAs are among the diverse groups that play a range of functions including some others with undefined functions.

Small nuclear RNA (snRNA):

- These are the RNAs that are rich in uridine nucleotides and so are called U-RNA.
- They are involved in mRNA processing.

Small nucleolar RNA (snoRNA):

- They play a central role in the rRNA molecule processing.

4.2.2 Proteome

The proteome can be defined as the entire set of all the proteins expressed by an organism. The final product of the genome expression is the proteome that can be described as the assortment of complete set of proteins produced in a cell, tissue or biological sample, at a particular time, developmental and cellular phase. This term can also be defined as the expression of an organism's genome. The translation is initiated from 4-nucleotides of DNA and mRNA and finally translated into complex code of 20 amino acids which vary in length of primary polypeptide sequences. The polypeptide sequence further undergoes folding to achieve a large number of possible conformations and post-translational modifications. These processes resulted to form final functional three dimensional products that is protein. The splicing process that results in distinct isoforms of the same protein is what causes the phenotypic variation. Proteomes are not stable like genomes of an organism and hence can adapt to different internal and external conditions as well as an organism's developmental stage.

The number of copies of each protein varies greatly; the rarest forms are thought to have less than 20,000 molecules per cell, while the most frequent types have 100 million copies. Approximately, 2000 proteins in the typical mammalian cell are considered relatively abundant if they have a copy number of more than 50,000 per cell. Very few differences occur among these abundant proteins when observed for proteomes of different types of mammalian cells, indicating that most of the proteins examined falls under housekeeping proteins and that are responsible to carry out general biochemical activities occurring in the cells. This suggests that the proteins with specialized function in a cell are often quite rare.

Some characteristic features of proteins are as follows:

- A protein is made up of monomeric subunits known as amino acids that together form unbranched polymer, or polypeptide chains with varying length.

- The polypeptide chain undergoes four distinct and characteristic levels of structures which are stated hereunder:

Primary structure: Amino acids are linked through peptide bonds to form a polypeptide. The polypeptide at one end consists of free amino group, NH_2 or N terminus and at other end free carboxyl group, COOH or C terminus. The peptide bond is formed through the condensation reaction between the carboxyl group of one amino acid and the amino group of second amino acid. The direction of polypeptide is N→C.

- ***Secondary structure*:** The primary structure of polypeptide further undergoes in various conformational alteration via intramolecular interactions to form a secondary structure of polypeptide. The secondary structure of polypeptide was elucidated by Pauling and Corey in 1953. The two main conformations of the secondary structure of protein are α-helix and β-sheet. Both of the structures are stabilized by hydrogen bonds that are formed between different amino acids of the polypeptide.

Tertiary structure: The 3-D (three dimensional) configuration resulted after the folding of secondary structural components of the polypeptide is the tertiary structure. The stabilization of tertiary structure is achieved through various chemical forces, H-bonding between individual amino acids, hydrophobic forces that cause amino acids with water repelling or non-polar side-groups to be shielded from water through embedding into the internal regions of the protein and covalent linkage called disulphide bridges between cysteine amino acids in the polypeptide.

***Quaternary structure*:** This is a multi-subunit structure that is formed through the association of two or more polypeptides each of which folded into their tertiary structure. These types of structures are attained by several proteins that perform complex functions and are also involved in genome expression.

- The chemical diversity of amino acids is responsible for the functional diversity of proteins. The R group differs among the different amino acids in terms of their chemical nature and structure and therefore show combination of chemical reactivities (Fig. 4.3). These combinations dictate the overall structure of the protein and thereby functionally diverse proteins are formed.

- Various biological functions are played by proteins as: biochemical catalysis, transport, structure, regulation, movement, protection and storage.

- Proteome analysis includes the quantitative and qualitative characterization of proteins alongwith the associated modifications and interactions. Proteomics refers to the determination of global protein profile of an organism that gives out the information of the activity of microorganism in response to varying environmental conditions.

- The amino acid sequence of the polypeptide can determine the level of folding and can predict to some extent as which secondary structure will be adopted by proteins through the determination of its primary sequence.

- The exogenous factors such as effects of environment, drugs, stress or illness or endogenous factors *viz.* genetic mutations and post-translational modifications cause change in the proteome. Proteomics has allowed the study of the effect of external factors through the discovery of biomarkers for early diagnosis and also to screen disease progression and drug response.
- The protein identification can be done through mass-spectrometry based approaches such as Liquid chromatography–mass spectrometry (LC-MS), High Performance Liquid Chromatography (HPLC), and Matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry (MS) techniques.

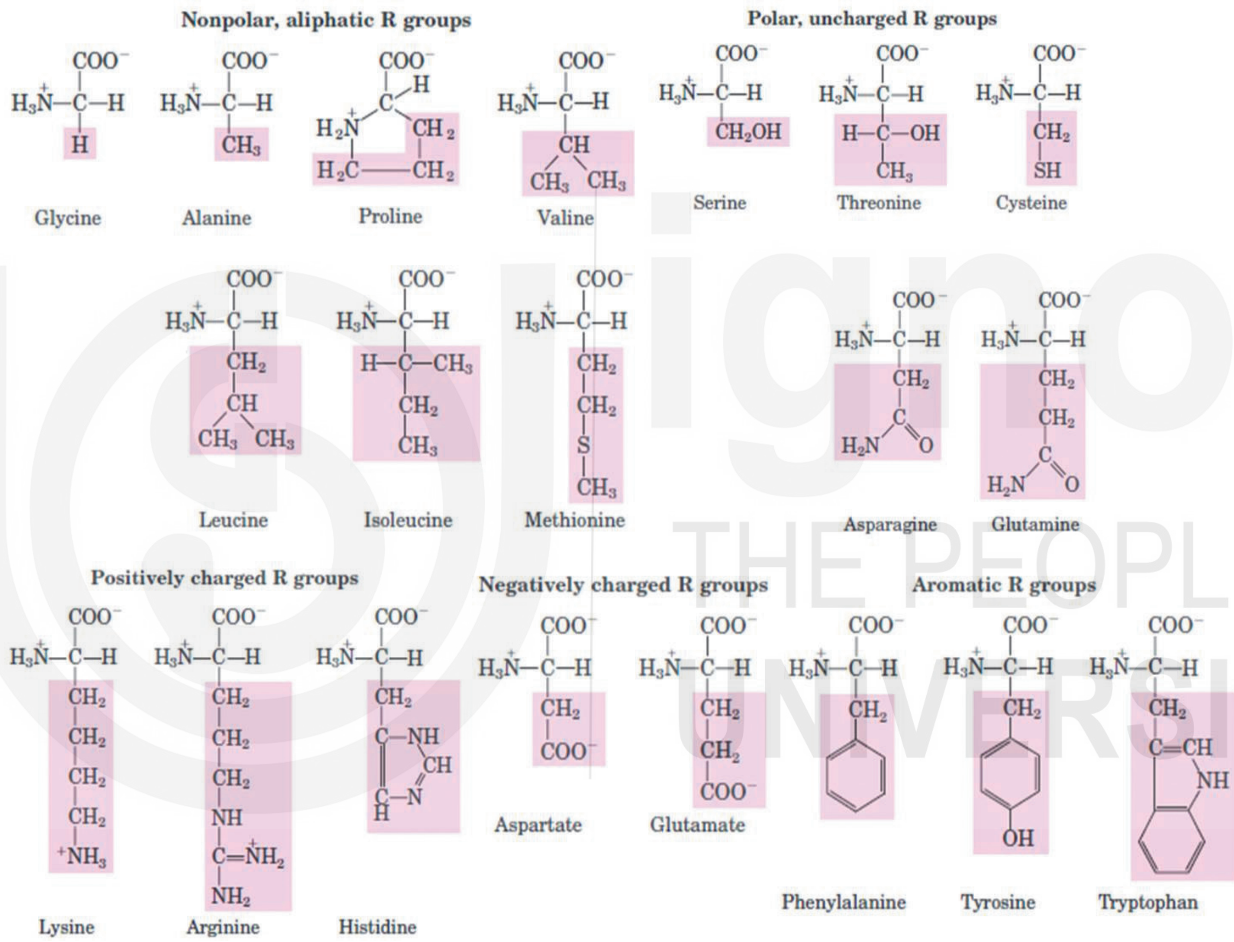

Fig. 4.3: The 20 common amino acids of proteins. The structural formulas show the state of ionization that would predominate at pH 7.0. The unshaded portions are those common to all the amino acids; the portions highlighted in pink are the R groups

SAQ 1

Fill in the blanks:

a) The .... is an integrative approach that defines the function and interaction of genes and their products by making use of high-throughput methods.

b) The .... refers to set of all RNA molecules that are produced by the genome in a specific cell under specific conditions.

c) The ....… RNAs are abundantly present in the cell and constitutes about 80% of the total RNA.

d) The .... RNAs play role as adaptors during translation process.

e) The ....… are rich in uridine nucleotides.

f) The ....… play a central role in the rRNA molecule processing.

g) The ....… can be defined as the entire set of all the proteins expressed by an organism.

h) The ....… bond is formed through the condensation reaction between the carboxyl group of one amino acid and the amino group of second amino acid.

i) The three dimensional configuration resulted after the folding of secondary structural components of the polypeptide is the .... structure.

j) The ....… diversity of amino acids is responsible for the ....… diversity of proteins.

k) The initial product of genome expression is ....… and the final product is ....… .

4.3 APPROACHES TO TRANSCRIPTOME ANALYSIS

Transcriptome captures a snapshot of all the total transcript of the cell at a specific time and conditions. Transcriptomics deals with the techniques that provides comprehensive account of the organism's transcriptome alongwith the sum of all its RNA transcripts. In the information network from DNA to RNA to functional product protein, mRNA acts as an intermediary molecule with some non-coding RNAs that play additional diverse functions. Insight into the transcriptome provides interpretation of the functional elements of the genome and unveils the molecular components of cells and tissues and also gives understanding of the development and disease.

Transcriptomics aims to provide:

➢ transcript details of the species including both coding and non-coding RNAs and small RNAs.

➢ transcriptional structure of genes, in respect to their start sites, 5' and 3' ends, splicing patterns and post-transcriptional modifications.

➢ cell and tissue specific gene expressions and quantitative differential gene expression of each transcript among multiple mRNA populations, during development and other physiological conditions.

➢ transcriptional profiling of the modifications associated with the mutant phenotype by determining the loss-and gain-of-function in the mutants.

- improved understanding of the dynamics of cellular and tissue metabolism so that alterations in the transcriptome can be used to screen the diseases and other health issues.

4.3.1 Hybridisation based Approach

Gene expression microarray

Microarray technique first came into existence in 1995 by Patrick Brown and colleagues. A microarray is a laboratory technique used to simultaneously measure the expression of thousands of genes. Microarray analysis appears to have an advantage over Northern blot or qPCR since it allows for the simultaneous high-throughput identification of a large number of short RNAs. The relative or absolute abundance of specific RNA transcripts is commonly evaluated using microarray-based systems. The results of gene expression microarrays have revealed a wealth of new knowledge *viz.* how the transcriptome is deployed in various cell types and tissues; how gene expression changes between development phases and diseased phenotypes; and how it varies within and between species. How much of the genome is transcribed to non-coding RNAs can also be determined using this technique.

Principle

The microarray-based method is based on a single microarray chip on which a large set of genes are arrayed in a regular and compact manner. The small size of the spots that is <200 µm allows several thousands of different oligonucleotides to be immobilized on one single slide. In a typical advance microarray, patches of such probes that are complementary to the transcripts of interest are present. Modern arrays consist of probe design that is based on the genome sequences or known or predicted open reading frames (ORFs) or designing of multiple probes per gene model. The transcripts to be investigated are labelled with fluorescent dyes (either one colour or two) then hybridized to the arrays, washed and further scanned with a laser. The gene expression can be measured through the observed light intensities. Hybridisation results are then imaged and analysed. The binding efficiencies in different samples can be compared and gene transcript level changes can be analysed in a single experiment.

Methodology

For the target preparation, total or messenger RNA is used as a starting material. The steps involve RNA isolation from both the experimental sample and a reference sample. The RNA is then used to chemically synthesize complementary DNA (cDNA) using oligo-dT primer or random primer. To avoid interference in the process of labelling and hybridisation, DNase digestion is done after RNA isolation. The synthesis of cDNA is followed with the incorporation of fluorescently labelled nucleotide (Cy3-dC/UTP or Cy5-dC/UTP), differentially coloured samples are prepared for both reference and experimental cDNA. Subsequently, the template RNA is put to degradation through chemical treatment and cDNA is separated from primers, unincorporated nucleotides and other RNA debris. The two sets of cDNA are further mixed and co-hybridised to the same array. The unbound and non-specific bound cDNA is then removed through washing and the fluorescence intensity is measured for each individual spot after subsequent scanning of the array. The methodology has been represented in Figure 4.4.

Applications

1. The microarray-based approach allows to develop an immense knowledge about the biological or cellular mechanisms like:
   - ➢ cell cycle progression
   - ➢ drug or disease model induced altered gene expression
   - ➢ cell fate determination
   - ➢ embryonic development
   - ➢ gene regulation in response to hormones.
2. The wide application of microarray has been observed to assess:
   - ➢ disease-associated gene regulation
   - ➢ pattern of gene expression in disease subtypes
   - ➢ gene biomarkers of several disease like inflammatory, psychiatric, infectious, neurological and cancer disease.
3. Moreover, the apparent applications of microarray have also seen in pharmacogenomic/ toxicogenomic studies for drug discovery and tracing the mechanism adopted by therapeutic drug and also to assess the side effects caused by specific drug.

Advantages

Microarray technology offers several advantages over other gene expression techniques, making it a valuable tool in genomics research. One significant benefit is its high-throughput capability, allowing simultaneous analysis of thousands of genes, which provides a comprehensive snapshot of gene expression across the entire genome. This extensive coverage is particularly useful for identifying gene expression patterns and discovering novel genes associated with specific conditions. Microarrays also offer high reproducibility and relatively low cost per sample, especially when compared to techniques like RNA sequencing. Additionally, they require smaller amounts of RNA, making them suitable for studies with limited sample availability. The standardization of microarray platforms and protocols further facilitates cross-study comparisons and meta-analyses, enhancing the robustness and generalizability of research findings. These advantages make microarrays a powerful and efficient option for large-scale gene expression profiling.

Limitations

Although the hybridisation-based approaches are high-throughput and are cost-effective, there are several limitations that restricts microarray-based method and also reduce reliability on microarray generated results. The limitations can be listed as:

- ➢ It is based on the previously ascertained knowledge of the genome sequence.
- ➢ Cross-hybridisation or non-ideal hybridization kinetics that results in high background levels.

- Dynamic range of detection is limited as due to the background and saturation signals.
- The comparison of expression levels across different experimental samples often creates hurdles requiring complicated normalisation methods.
- RNA-microarrays are robust and through updated pre-designed probes they are optimized for comprehensive coverage, but they show limitations in detecting the transcript that is not included in the probe.

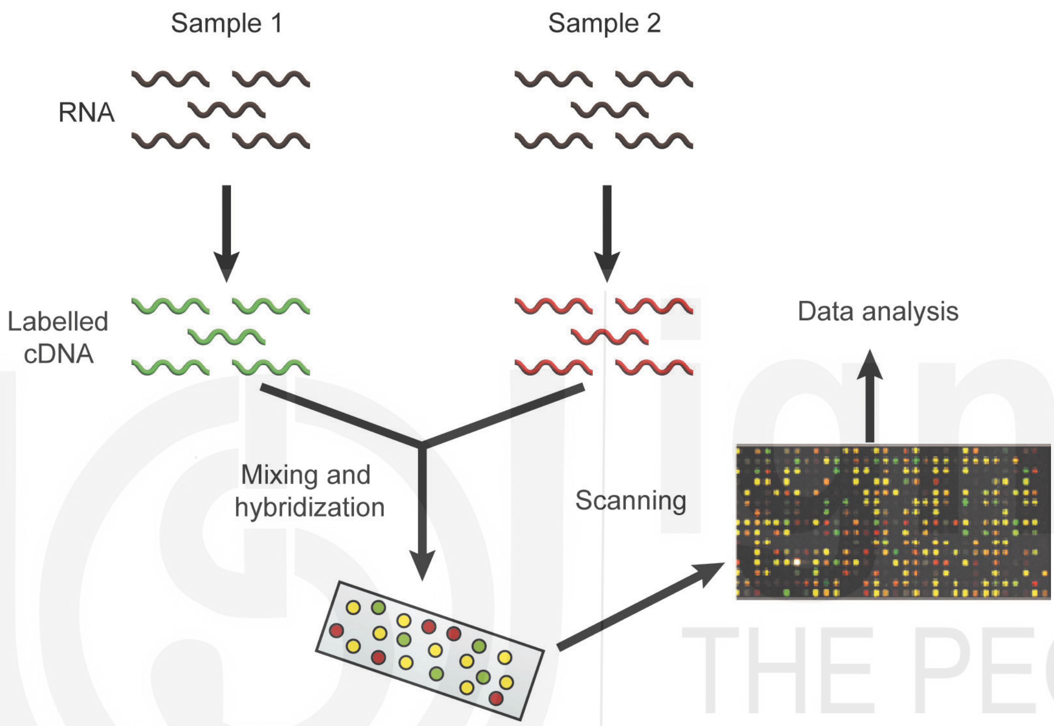

Fig. 4.4: Workflow summary of microarray. High-quality RNA of interest is isolated from biological source and converted to cDNA by the process of reverse transcription. The cDNA is then chemically labelled and hybridization is done followed by scanning

4.3.2 Sequencing-based Approaches

RNA sequencing

RNA sequencing studies (RNA-Seq) have been expanding at a remarkable pace as it has gained impetus as an indispensable tool for transcriptome-wide analysis of differentially expressed genes and differential splicing of mRNAs. It utilizes ultra-high-throughput cDNA sequencing methodology for the analysis of gene expression. In contrary to microarray-based methods, sequence-based approach involves direct determination of cDNA sequence. RNA-Seq has revolutionized the transcriptome study through its highly sensitive and accurate approach of measuring gene expression across the genome. It has provided scientist with the determination of previously undetected changes appearing in diseased state, in response to therapeutic drug, under varying conditions and across a broad range of other study designs. The approach allows detection of both known and novel features in a single assay and also allows the identification of single nucleotide variants, transcript isoforms and gene fusions using RNA-Seq without being constrained by the knowledge of previously discovered features. Earlier, low throughputs, costly and non-quantitative approaches like Sanger sequencing of cDNA or expressed sequence tag (EST) were used. Later, development of tag-based methods like serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS) overcome the limitations as existed earlier. These high-throughput tag-based sequencing approaches can provide unbiased and precise 'digital' gene expression levels. The advent of high-throughput next-generation sequencing (NGS) has significantly transformed transcriptomics by allowing RNA analysis through the sequencing of complementary DNA (cDNA). The RNA-Seq technique offers detailed and quantitative account of gene expression, allele-specific expression and alternative splicing. Advances in the RNA-Seq workflow starting from sample preparation to sequencing platforms and to bioinformatic data analysis has allowed deep profiling of the transcriptome and also provide great prospect to explicate different physiological and pathological conditions. Several Next-generation sequencing (NGS) technologies are emerging in the biomedical fields for transcript profiling. The technologies like GS-FLX (Roche CT, U.S.A.), Genome Analyzer (Illumina CA, U.S.A.) and Sequencing by Oligonucleotide Ligation and Detection (SOLiD) (ABI CA, U.S.A.) have the potential to dramatically boost sequencing throughput at a fraction of the price. The analysis of single-cell gene expression, translation (the translatome), and RNA structure (the structurome) are all now possible using RNA-seq techniques.

Principle

The process is based on the conversion of a population of RNA to a library of cDNA fragments with adapters attached to one or both ends. Each molecule is then put to sequencing in a high-throughput manner in order to obtain sequences either from one (single-end sequencing) end or both ends (pair-end sequencing). The reads are 30-400 bp depending on the technology used i.e., Illumina IG, Applied Biosystems SOLiD or Roche 454 Life Science. The alignment of resulting reads is done to a reference genome or reference transcripts or are assembled *de novo* without genome sequence to generate a genome-scale transcription map consisting of both transcriptional structure/expression level of each gene.

Methodology

Initially, RNA isolated from the biological sample and is converted into cDNA fragments (a cDNA library) by the process of reverse transcription allowing RNA to enter into NGS workflow. The fragmentation of cDNA is done and adapters are attached to one or both ends of the fragment. The adapters consist of functional elements that enable sequencing, for instance, the amplification element (that allows clonal amplification of the fragments) and priming the sequencing priming site. Amplification process is followed by size selection, clean-up, quality checking and further cDNA library is analysed by NGS. The short sequences are produced that reflect all or a portion of the original fragment. Depending on what the output data will be utilized for, the library is sequenced at different depths. Sequencing can be done using either paired-end or single-end sequencing techniques. Single-read sequencing involves sequencing of the cDNA fragments from just one end and is a cheaper and rapid technique as compared to Sanger sequencing. However, paired-end methods are expensive as it involves sequencing from both ends but offer advantages in post-sequencing data reconstruction. The workflow of RNA-Seq has been represented in Figure 4.5.

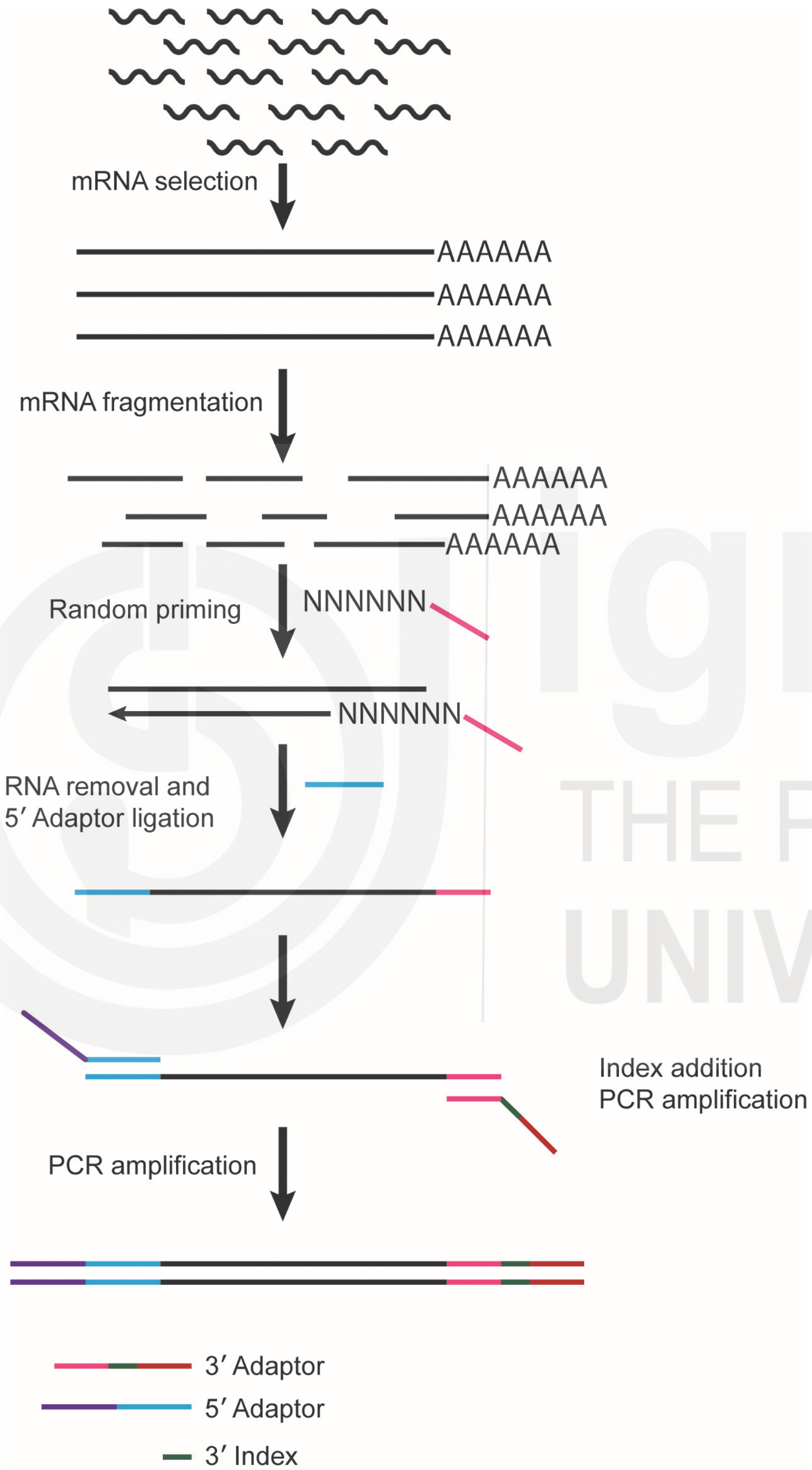

Fig. 4.5: Workflow of RNASeq. Total RNA is extracted from the biological sample and is converted to cDNA through reverse transcription process and further adaptor is ligated to the 5′ of the single stranded cDNA. Short sequence is obtained from each cDNA using high-throughput sequencing technology which is then aligned to a reference genome sequence to reconstruct the genome regions that are transcribed

Advantages

In recent times the microarray approach has been supplanted by high-throughput RNA sequencing due to number of reasons listed below:

- Detection of all transcripts in a given sample, including the regulatory siRNA and lncRNA transcripts
- Greater transcriptome coverage with the identification of novel or rare transcripts, as well as alternatively spliced variants, fusion genes and allele-specific expression.
- Measures both the high- and low-abundance RNAs over a five-order-of-magnitude range.
- Least starting material requirement (nanograms *vs.* micrograms and even as little as 50 pg).
- The assembly of RNA-Seq is independent of reference genome.
- Analysis of gene expression for poorly described species with limited genomic resources.
- Efficient quantification over a wider dynamic range with less noise, allowing for more subtle changes to be accurately quantitated.

Limitations

- Costly, onerous, and employs a time-intensive process.
- Isolation of sufficient, high-quality RNA.
- Uneven depth along the length of transcript or heterogeneity of reads that cause biasness in sequencing.
- There are a number of informatics hurdles that RNA-Seq must overcome, including the creation of effective ways to store, retrieve, and process enormous amounts of data.

4.4 APPLICATIONS AND OUTCOMES OF TRANSCRIPTOMIC STUDIES

1. Differential gene expression (DEG)

- **Diagnostics and disease profiling:** The changes in transcriptome in response to environmental factors, introduction of treatment/drug or during disease progression can be detected. The methods involved to investigate differential gene expression are northern blot, western blot, quantitative real-time PCR (qPCR), expressed sequence tags (ESTs), microarrays, and RNA-Seq.
- **Agriculture and food industry:** In poultry and dairy, identification of the DEGs that are associated with several growth and development processes such as the formation of eggshell in laying hens and milk production in lactating cows at different time points.

- **Pathogenesis-related pathways:**

  a) Identification of DEGs involved in the plant-pathogen interaction by studying global transcriptome profiles.

  b) In humans, RNA-Seq of human pathogens have been used for the quantification of gene expression changes, predicting antibiotic resistance, revealing host-pathogen immune interactions and identification of novel virulence factors.

2. Diagnostics and disease profiling

- Genetic variations in the coding region resulting in asynonymous or synonymous variants and such variants including single-nucleotide variants (SNVs) structural variants (SVs) and short insertions/deletions (indels <50 bp) can be studied using RNA-Seq. Identification of disease-associated single nucleotide polymorphisms (SNP), gene fusions and allele-specific expression, significantly build knowledge about disease causal variants.

- Genome-wide screening of genetic variants is of great importance to expedite the employment of genome-based breeding methods for selecting desirable traits in plants and animals.

3. Allele-specific expression (ASE)

- Analysis of allele-specific expression and the detection of fusion transcripts that often occur in diseases like cancer.

- Combining whole genome sequencing and RNA-Seq approach, information regarding ASE and allele silencing can be obtained for plants and animals.

4. Alternative splicing

- Splice variants resulting during varying environmental conditions play role in generating transcriptomic and proteomic diversity. Detection of genotype-specific AS events using splice variants has been achieved that in turn serve as biomarkers for genome-wide association studies (GWAS).

- Relationship between defective alternative splicing and human pathologies has been established. Aberrant RNA splicing leads to human diseases ranging from neurological pathologies to immunohematology disorders and malignancies. Moreover, mutations in splicing-regulatory elements or alterations in components of the splicing machinery leading to disease progression can also be detected.

5. Noncoding RNA

- Identification and characterisation of diverse classes of non-coding RNAs including microRNA (miRNA), short interfering RNA (siRNA), small nuclear RNA (snRNA), Piwi-interacting RNA (piRNA), small nucleolar RNA (snoRNAs) and long ncRNA (lncRNA).

- Role of miRNA in various biological process involving growth, development and stress responses in plants as well as in potentially altered agronomic traits. The miRNAs playing role in gene silencing can also be detected.
- The miRNA profiling serves as novel biomarker in the diagnosis and prognosis of several human diseases.
- Detection of lncRNA to be associated with the disease, from cancer to cardiac pathologies or neurodegenerative disorders. It is also involved in other disorders like diabetes and Alzheimer.

4.5 APPLICATIONS OF OMICS TO HUMAN DISEASE

Integration of multiple technologies has emerged as an effective approach to obtain a more comprehensive account of biology and disease. Advances in omics technologies such as genomics, transcriptomics, proteomics and metabolomics have led to unprecedented efforts in characterizing the molecular changes associated with the development and progression of complex diseases and has also significantly contributed in biomedical applications that can be listed as follows:

- Oncology: identifying driver mutations, molecular signatures and gene therapy.
- Application of genomics, transcriptomics and proteomics for diagnosis and pathogenesis of neurological diseases.
- Muti-omics based approach to study diagnosis, pathogenesis, prognosis, and potential drug targets of COVID-19.
- Identification of biomarkers to detect early-stage chronic kidney disease.
- Omics approach in discovery of biomarkers in diagnosis of inflammatory lung diseases.
- Improved diagnosis for mitochondrial disease.
- Insight in to the cause, diagnosis and identification of novel drug targets and biomarkers for treating complexity of cardiovascular disease.
- Diagnosis and management of many autoimmune diseases.

SAQ 2

Fill in the blanks:

a) The …………………… captures a snapshot of all the total transcript of the cell at a specific time and conditions.

b) The high-throughput next-generation sequencing (NGS) has significantly transformed transcriptomics by allowing RNA analysis through the sequencing of …………………… .

c) The …………………… based method is based on a single microarray chip on which a large set of genes are arrayed in a regular and compact manner.

d) The ………………… technique offers detailed and quantitative account of gene expression, allele-specific expression and alternative splicing.

e) The assembly of RNA-Seq is independent of …………………… genome.

f) The ………………… profiling serves as novel biomarker in the diagnosis and prognosis of several human diseases.

4.6 SUMMARY

- Genome is a complete set of genetic information that is transcribed into transcriptome and finally to a functional product that is proteome.
- Central dogma of biology includes flow of genetic information from DNA to RNA to protein.
- The initial product of genome expression is transcriptome that refers to the complete set of all RNA molecules that are synthesized in a particular cell/tissue under specific conditions.
- Unlike coding RNAs, non-coding RNAs are not involved in the process of translation but carry out several regulatory processes.
- The proteome can be defined as the final product of genome expression and is consists of complete set of proteins formed in a cell, tissue or biological sample, at a particular time and that is required at particular developmental or cellular stage.
- The approaches to transcriptome analysis are hybridisation-based and sequencing based.
- Gene expression microarray is a hybridisation-based approach that allows simultaneous measurement of the expression of thousands of genes.
- In contrast to microarray, sequence-based approach of transcriptomics i.e. RNA-Seq offers highly sensitive and accurate approach of measuring gene expression across the genome without being constrained by the knowledge of previously discovered features.
- Omics based approach have shown their significant contribution in several sectors, including agriculture, food industry, health and medicine. Insight into the differential gene expression, genetic variants, fusion transcripts, alternative splicing and detection of non-coding mRNAs have provided with the advanced diagnostics solutions for the development of biomarkers and gene therapy.

4.7 TERMINAL QUESTIONS

1. What is genomics and functional genomics?
2. Describe the concept of transcriptome and proteome.
3. What are different levels of the structures of protein?
4. Explain different approaches to transcriptome analysis.
5. Describe various applications and outcomes of transcriptomic studies.
6. Enlist some applications of OMICS to human disease.

4.8 ANSWERS

Self-Assessment Questions

1. a) functional genomics, b) transcriptome, c) ribosomal, d) transfer, e) small nuclear RNA, f) small nucleolar RNA, g) proteome, h) peptide, i) tertiary, j) chemical; functional, k) transcriptome; proteome.
2. a) transcriptome, b) complementary DNA (cDNA), c) microarray, d) RNA-Seq, e) reference, f) miRNA.

Terminal Questions

1. Refer to Section 4.1.
2. Refer to Section 4.2.
3. Refer to Subsection 4.2.2.
4. Refer to Section 4.3.
5. Refer to Section 4.4.
6. Refer to Section 4.5.

SUGGESTED READINGS

1. Afzal, M., I. Manzoor and O. P. Kuipers (2015). "A fast and reliable pipeline for bacterial transcriptome analysis case study: serine-dependent gene regulation in Streptococcus pneumoniae." JoVE(98): e52649.
2. Brown, T. A. (2002). Transcriptomes and proteomes. Genomes. 2nd edition, Wiley-Liss.
3. Hou, Z., P. Jiang, S. A. Swanson, A. L. Elwell, B. K. S. Nguyen, J. M. Bolin, R. Stewart and J. A. J. S. r. Thomson (2015). "A cost-effective RNA sequencing protocol for large-scale gene expression studies." 5(1): 1-5.