LECTURE-03

GENOMICS AND TRANSCRIPTOMICS: WHY PROTEOMICS?

TRANSCRIPT

Welcome to the proteomics course. Today, we will talk about Genomics and Transcriptomics and then we will talk about why to study proteomics?

Lecture outline – we will talk about Genomics, transcriptomics and need to study proteome by using proteomics.

Understanding biological system is very challenging task. During the last decade we have witnessed the revolution in biology, as this discipline has fully embraced "omics" tools. The emergence of genome-wide analyses to understand cellular DNA, RNA and Protein by employing genomics, transcriptomics and proteomics has revolutionized the study of control networks that mediate cellular homeostasis.

Let's talk about Genomics. It has been only 60 years since 1953 landmark study of Watson & Crick deduced double helix structure of DNA and biological science research has witnessed great progress in genomics research during the last decade. The simultaneous efforts of the Human Genome Project and Celera Genomics completed sequencing of human genome in 2001.

Let's look some of the definitions:

Genome: The entire sequence of an organism's hereditary information, including both coding and non-coding regions, encoded in DNA is known as the genome.

Studying genome of an organism by employing sequencing and genome mapping techniques is known as "genomics".

The genome sequencing projects produced massive amounts of information on DNA sequences for many species. The genomics research is also providing information for mutations, deletions and epigenetic alterations, which modulate gene expression.

The various sequencing methods have been employed for genome sequencing. Let's first discuss the traditional methods.

The traditional DNA sequencing was done by Sanger's method. As you can see in image here, the genomic DNA is fragmented, cloned into a plasmid vector, and used to transform *E. coli*. The fluorescently labeled dideoxynucleotides (ddNTPs) terminate DNA extension reaction. In a single tube four labeled ddNTPs are added and in an

capillary based electrophoretic gel colored DNA fragments are separated. A laser beam is applied which detects color and determines sequence.

Chain termination DNA sequencing – Sanger's dideoxy method

A simple and elegant method for DNA sequencing was devised by Frederick Sanger where a collection of DNA fragments are synthesized by means of controlled interruption of enzymatic replication. Four DNA synthesis reactions are carried out simultaneously with the strand whose sequence is to be determined being used as the template. The reaction mixture consists of regular deoxynucleotides and DNA polymerase along with a small amount of one labeled di-deoxy nucleotide analog being added to each of the four reaction mixtures. A primer is added to begin the DNA synthesis and strand elongation continues until a di-deoxy analog gets added instead of the regular dNTP. Chain termination occurs at this stage due to the absence of a 3' OH group for formation of the next phosphodiester bond.

The synthesized strands are separated from each other, after which the differentially labelled strands of various lengths are separated by electrophoresis. The smallest fragments move further in the gel while the larger fragments remain close to the point of application. The different fluorescent labels of each ddNTP can then be detected by scanning the gel with a beam of laser. The output sequence obtained is complementary to the template strand, which can be used to deduce the original desired template sequence.

To obtain the human genome sequence, the human genome project employed strategy of shotgun sequencing. From a genomic library clones were isolated and ordered into a detailed physical map. Further, the individual clones were sequenced by shotgun sequencing. The human genome project group produced a working draft of human genome by a map-based strategy, while Celera Genomics sequenced the human genome by whole-genome shotgun method.

Shotgun sequencing

Genomic DNA is cleaved using a suitable restriction endonuclease and the fragments inserted into bacterial artificial chromosomal vectors. These vectors enable the DNA fragments to be amplified. The genomic DNA fragments of the library are then organized into a physical map, after which individual clones are selected for sequencing.

The selected BAC is amplified and these clones are sequenced using the Sanger's chain termination method. The sequence of the clone is then deduced by aligning them

based on their overlapping regions. The entire genomic sequence is then obtained once each BAC is sequenced in this manner.

Let's compare some of the traditional DNA sequencing methods. As you can see here chain termination method or Sanger's method that is a gold standard but it is very time consuming and sensitivity is very high. But In Pyrosequencing, which is based on chemiluminescent detection system the sensitivity, is very high. Even MALDI-TOF can be also employed if your aim is to identify variations in SNPs and sensitivity is very high.

Let's discuss about the Genome Sequencing Project

Genome sequencing projects: aimed to decipher the complete genome sequence of all chromosomes of an organism. Several such projects were initiated and successfully accomplished. Sequencing the genome of higher organisms, such as human, was very challenging but researchers have shown very good team effort and successfully completed the draft sequence in 2001 and complete sequence in 2003. Here on the time scale you can see the sequencing of various organisms, which was progressed from year 1990 to 2001.

The human genome project provided insights for various details such as the total number of genes is estimated to be 25,000. The average gene consists of 3,000 bp. The human genome consists of 3.2 million bp. The human genome sequence is exactly the same almost 99.9% in all the people and about 2% genome encodes instruction for the synthesis of protein

The various potential benefit of the human genome project:

The information was useful for molecular medicine, environment, agriculture and livestock, microbial genetics, risk assessment of individuals for specific allergens as well as other fields such as anthropology, evolution and human migration.

Several genome sequencing projects that aim to elucidate the complete genome sequence of organisms have been undertaken by several research groups all over the world. The DNA sequences are identified by the shotgun sequencing technique and then aligned using suitable software to provide the complete genome sequence. The genome sequences of a large number of prokaryotic and eukaryotic organisms have been successfully deduced.

The immense amount of information held by the human genome motivated researchers to understand the nature and content of genetic material in great detail. A collaborative effort between 6 countries and 20 laboratories was undertaken in 1990 to produce a draft of the human genome sequence. Work proceeded rapidly with a draft covering

most of the genome being completed by 2000 and greater coverage being achieved by 2003.

Before sequencing the entire genome, physical maps of the chromosomes were made. This helped to provide key tools for identification of disease genes and anchoring points in the genomic sequence. Pilot projects were then launched to create a draft of the genome sequence.

The shotgun approach was the fundamental technique used for large scale sequencing of the human genome, which also makes use of Sanger's sequencing. The collaborative effort to sequence the entire genome was challenged in 1998 by a privately funded organization which aimed to reach the target before the publicly funded group. Progress made in sequencing was very rapid and by 2001, a draft of the sequence was ready covering around 83% of the genome.

After successfully sequencing the human genome the technology of sequencing got advanced. And now there are next generation sequencing methods are available.

Human Genome Sequence project was completed in almost 12 years at a cost of > 3 billion USD. Very recently the next-generation sequencing strategies have dramatically increased the pace of sequencing by several order of magnitudes and also reduced the per base cost of raw sequence significantly. Individual genome sequencing, metagenomics studies, SNP and mutatnt analysis is also possible due to the next generation sequencing methods.

So traditional sequencing which we discussed earlier including the Sanger's method-chain termination method as well as chemical method (Maxam and Gilbert in 1976), those methods can be termed as first generation sequencing methods. The Next generation or second generation sequencing technology – it includes sequencing by ligation or by synthesis, including pyro-sequencing and reversible chain termination. Even more advanced form, the third generation sequencing technology is also in progress. Which is aiming to further improve second-generation sequencing technology and lowering the cost per base, by using scanning tunneling electron microscope, fluorescence resonance energy transfer, single-molecule detection and protein nanopores.

The NGS based on nanopore structures is known as nanopore sequencing which is shown in this slide. This is a single-molecule DNA sequencing technology that does not require any fluorescent labelling and makes use of any changes in electrical current produced when single nucleotides pass through the nanopore. Individual base detection

is achieved through the measurement of conductivity either across or through a membrane, via a nanoscale pore.

This innovative method offers a label-free approach for DNA sequencing. An exonuclease cleaves the single stranded DNA, one base at a time to release the nucleoside monophosphates. These NMPs pass through the nanopore under an applied potential, which is covalently coupled to an adapter molecule. Continuous movement of NMPs through the nanopore results in characteristic fluctuation of electric current, which enables detection of the various nucleotide bases.

Multiple round of nucleotide additions are carried out on the immobilized template DNA by using DNA polymerase in the presence of ATP sulfurylase, luciferase and the nucleotide degrading enzyme apyrase. The release of an equal amount of pyrophosphate is determined by its conversion to ATP by ATP sulfurylase which in turn is determined by the release of light on reaction with luciferase. The amount of light produced is determined by means of a CCD camera which is used to determine the nucleotide added & therefore the sequence of the template DNA.

The various NGS commercial platforms currently available, we will briefly talk about the pros and cons of each of these methods.

Illumina - Flow cell-based, is employing reversible dye termination, and uses 4-color optical imaging; it use for whole genome sequencing; widely used but one of the limitation is low multiplexing capability.

Pyrosequencing - Bead based with emulsion PCR and CCD light imaging; it is used for targeted exon sequencing; however, contamination risk due to emulsion PCR is one of limitation.

Helicos - Oligo-dT captured PolyA-tailed DNA fragments and optical imaging; the method is used for whole genome and single molecule sequencing; high NTP incorporation error is one of the limitation.

Solid- Supported Oligonucleotide Ligation and Detection is also known as SOLID. It is based on sequential dinucleotide ligation and optical imaging; it is used for whole genome sequencing and SNP detection; limitation is longer run time.

Ion Torrent - Standard DNA polymerase chemistry based method, semi-conductor based non-optical detection; it is used for targeted sequencing.

These are very few representative platforms, which are described here. But there are many exciting commercial platforms are available currently for NGS.

So let's discuss the advantages and the disadvantages of NGS versus traditional Sanger's based sequencing.

In NGS preparations are done *in vitro* whereas in Sanger's method as we discussed the transformations of *E. coli* is done. So these type of limitations are avoided using NGS method. The NGS methods is based on arrays it is not capillary based method as Sanger's method. So sequencing time is reduced in NGS method. The cost is reduced in NGS method because all the polonies can be treated with single reaction volume which reduces overall cost. The next generation sequencing method can detect deletion, translocation, copy number alterations etc. with very high accuracy, however very sophisticated bioinformatical tools are required for data analysis of NGS and data and data interpretation is still remain challenging. Some of these are limitation of NGS platforms.

Now let's talk about another interesting technique commonly applied in genomics as well as transcriptomics.

The DNA Microarray- DNA microarrays can be prepared from any DNA sequence by chemical synthesis or Polymerase Chain Reaction. The DNA is printed on glass slide or other substrates. DNA is cross-linked by UV light or other immobilization chemistry. Once ready, these microarray slides can be used for various applications.

Let's talk about a typical microarray experiment where aim is to compare a healthy sample with a deceased treated sample. So from both the population mRNA has to be extracted then mRNA should to be converted to cDNA by enzyme reverse transcriptase as you can see second step here in the slide. During this step label samples using the fluorescently labeled deoxyribonulceotide triphosphates. On a microarray slide incubate labeled cDNA, which have complementary sequences on microarray. Perform washing step and remove unhybridized probe and scan the gel to look for gene changes in the gene expression. In this manner expression level of thousands of genes can be measured and analysed by using DNA microarray platforms simultaneously.

After discussing some of the techniques used in genomics let now move on to transcriptomics. So mRNA translation is an intricate process, which takes place on ribosomes. Study of all the mRNA molecules expressed by a particular cell type of an individual is known as transcriptomics. The transcriptomic analysis measures the genes that are being actively expressed at any given time and varies significantly with external environmental conditions, development condition or different type of environmental cues.

Let's look at the some of the definitions-

Transcriptome – is analysis of all species of transcript, including mRNAs, non-coding RNAs and small RNAs.

Transcript analysis is used to quantify the expression level changes of each transcript during development and under different conditions.

Transcriptomic analysis is also very important for understanding the function elements of genome and molecular constituents of cell and tissue.

The various techniques which have been used for evaluating gene expression changes such as Northern blotting, the more tradition method, and some of the newer methods involve Quantitative real-time polymerase chain reaction (qRT-PCR or qPCR), Differential display method, and Serial analysis of gene expression (SAGE) and microarray.

Let's talk some of these techniques in some more detail.

Reverse Transcription PCR (RT-PCR) is a variation of regular polymerase chain reaction that is used to generate multiple copies of DNA starting from a molecule of RNA. As you can see in the slide, this technique first requires the template RNA strand to be reverse transcribed into corresponding complementary DNA (cDNA), which is then amplified by traditional PCR and multiple copies can be generated.

Reverse transcription PCR is used to generate multiple copies of DNA with RNA as the starting material. The template RNA molecule is first reverse transcribed into the corresponding cDNA by means of the enzyme reverse transcriptase. This enzyme, which is commonly found in viruses, is capable of synthesizing DNA from an RNA template.

Traditional PCR is then performed on the cDNA obtained by addition of primers which are allowed to anneal at 54 degree centigrade. This is followed by addition of nucleotides and Taq Polymerase, which performs elongation of the template strand at 72 degree centigrade.

Second and subsequent rounds of PCR result in further amplification of the cDNA of interest. Strand separation is performed at 95 degree centigrade followed by primer annealing and elongation respectively. In this way, the mRNA transcript originally used is amplified in the form of its corresponding cDNA which can then be studied further.

A highly sensitive method, which is employed to quantify the amount of RNA is known as Real-time PCR or Quantitative PCR. This methods requires very little (5–10 ng) of RNA for detection of transcript changes. It quantifies a target DNA molecule as and

when it is amplified by means of a suitable reporter dye that emits fluorescence every time a PCR cycle is performed.

Real time PCR - The double stranded DNA that needs to be amplified is heated to 95°C to bring about strand separation. Once the strands are separated, primers are added along with the probe DNA molecules, which have the quencher and reporter molecule bound to its ends. Once these have annealed to the template DNA strands at 54°C, Taq Polymerase and nucleotides are added and the temperature is again increased to 72°C to carry out strand elongation.

The Taq Polymerase continues to elongate the DNA strand based on the corresponding template DNA. When it reaches the bound probe molecule, the 5'-3' exonuclease activity of Taq Polymerase degrades the probe into its nucleotide fragments and continues to elongate the DNA strand. The released reporter dye thus gets separated from the quencher molecule during this process and fluorescence emitted can be detected using a suitable detector. The increase in fluorescence in real-time PCR is directly indicative of the amount of nucleotide being synthesized and is therefore a useful tool for measuring gene expression.

As we have talked earlier microarrays are very strong platform for gene expression changes. cDNA arrays measure many thousands of gene-specific mRNAs in a single tissue sample simultaneously. As discussed earlier, microarrays can be used for various applications, including measurement of mRNA stability, identification of short poly(A) phenotype, and studying mRNA association with membranes or cellular organelles. The methodology is similar to what we discussed earlier. These large- scale gene expression analysis has also proved to be useful as a validation strategy to validate the proteomics data and look at the changes at protein level and correlate with the transcript level.

cDNA microarrays - A DNA microarray is printed with oligonucleotide sequences that will serve as probe molecules by providing complementary strands for binding cDNA of interest. Thousands of such spots, each containing as little as picomole quantities of DNA, can be printed on a single slide. Commonly used binding chemistries for printing include covalent attachment via epoxy-silane, amino-silane, polyacrylamide etc.

The mRNA present in the control and test samples is extracted by addition of a lysis buffer, which breaks open the cells. Centrifugation helps in pelleting down the cell debris after which the mRNA present in the supernatant can be transferred into a fresh tube for further analysis.

The extracted mRNA is then reverse transcribed into its corresponding cDNA by means of the enzyme reverse trasncriptase. This enzyme, which is also known as RNA dependent DNA polymerase, is capable of synthesizing DNA from a corresponding RNA template.

The control and test cDNA samples obtained are then labeled with Cy5 and Cy3 dyes respectively. Since the two dyes can be excited at different wavelengths, the differentially labeled samples are then mixed together for further analysis.

The mixed cDNA sample that is differentially labeled by means of the two cyanine dyes is then incubated with the printed DNA microarray slide. This allows hybridization to occur between the probe oligonucleotides on the array surface and the labeled cDNA samples of interest.

The microarray is washed to remove any unbound cDNA molecule and then is scanned at suitable wavelength by means of a microarray scanner.

A new advancement in transcriptomic technology is RNA-sequencing. RNAseq has been used for transcriptome analysis by sequencing cDNA through Next generation sequencing. This methodology was initially utilized for identification of transcriptional map of yeasts.

In RNA-Seq, the long RNAs are converted into a library of cDNA fragments (as you can see here) through RNA or DNA fragmentation and to each cDNA fragment sequencing adaptors are added. From each cDNA the short sequence is obtained and resulting sequence reads (exonic, junction and polyA end-reads) are aligned with reference genome sequence.

So the sequence reads - exonic reads, junction reads and poly(A) end-reads can be used to generate base-resolution expression profile for each gene as you see here plotted with nucleotide position on X-axis and RNA expression level on the Y- axis. The RNA-Seq has dynamic range >8,000-fold to quantify gene expression changes, therefore it is very useful for discovering new transcripts, identifying mutations, deletions and insertions, as well as splicing alternatives. After studying genomics and transcriptomics let's now move on to proteomics. Why do we need to study proteomics?

After completion of human genome sequence identified genes (~30,000) were much less than expected 100,000 genes. The lower number of genes are surpassed by an estimated number of proteins in millions. Therefore studying large scale study of protein structure and function, requires a thorough understanding of protein composition and information at various structural levels by employing different type of proteomics tools.

So let's look at some of the definitions-

Proteome is entire complement of proteins expressed by genome of an organism under defined conditions and the study of entire compendium of proteins encoded by a genome is known as proteomics.

Let's discuss why do we need to study proteomics and how it compares with genomics?

So there are various reasons if you compare genomic versus proteomics to look at advantages and disadvantages of each method.

The genome represents only the starting point towards understanding complexity of biological functions. The products of gene expression, proteins, provide a much more meaningful insight into the mysteries of essential biological processes. Unlike genomic DNA, proteins do not contain intervening sequences-introns and are directly indicative of cellular function.

Alternative splicing is a process by which exons or coding sequences of pre-mRNA produced by the transcription of a gene are combined in different methods during RNA splicing. The resulting mature mRNA give rise to different protein products by translation, most of these are isoforms of one another. In this way a single gene can give rise to multiple protein forms.

Another level of information which is only obtained at the proteomics level is post-translational modification. Many proteins undergo post-translational modification at some of their amino acid residues after synthesis process. These modifications include hydroxylation, methylation, phosphorylation, glycolsylation etc. are some of the most common modifications observed. But there are other forms of post-translational modification which also occur in proteins.

Very briefly let's touch upon some of techniques employed in proteomics. Different type of proteomic technologies which employ gels can be grouped as gel-based proteomics- The proteins within a cell are commonly analysed using SDS-PAGE and two-dimensional (2D) gel electrophoresis. Separation in SDS-PAGE occurs almost exclusively on the basis of molecular weight since all proteins have a similar charge-to-mass ratio and shape after they have bound SDS. In 2DE the complex mixtures can be resolved first by isoelectric point and then by size on a polyacrylamide gel. Some of the limitation of 2DE were overcome by Difference gel electrophoresis (DIGE) technique. 2DE or DIGE in combination with mass spectrometry has been the standard technique for proteomic analysis.

Mass Spectrometry, it is another technique for protein identification & analysis by production of charged molecular species in vacuum, & separation by magnetic and electric fields based on m/z ratio. The different components of mass spectrometry:

Ionization source, it is one of the major components of any MS instrumentation, which fragments the sample into an ionic form for further detection. MALDI and ESI are most commonly used for proteins samples.

Mass analyzer resolves ions, which are produced by ionization source on basis of their m/z ratios. Various characteristics such as resolving power, accuracy, mass range & speed determine the efficiency of these analyzers. Commonly used mass analyzers include Time of Flight (TOF), Quadrupole (Q), ion trap, orbitrap etc. MS has become the method of choice for analysis of complex protein samples in proteomics studies due to its ability to identify as well as thousands of proteins

Protein Microarrays-These are miniaturized arrays normally made of glass onto which small quantities of many proteins are simultaneously immobilized and analyzed. Protein microarrays can be generated by either traditional cell-based methods or cell-free methods. Protein arrays can also be used to probe protein–protein, protein–DNA interactions as well as measurement of protein activity and other biological applications.

Very briefly let's look at the label free detection techniques: Surface Plasmon Resonance (SPR). Label-free detection methods, monitor inherent properties of the query molecule itself, and overcome many limitations of traditional label-based methodology. Various techniques such as SPR, SPRi, Ellipsometry, Interferometry etc. which are commonly are commonly used for label-free detection system. Now shown here in the slide is surface Plasmon resonance method, which detects any change in the refractive index of medium of the material interface between the metal surface and the ambient medium.

In summary today we talked about different techniques, which are employed in genomics, transcriptomics and we looked at why to study the proteomics. So as we know, biological systems are very complex to understand. Genomics, Transcriptomics and Proteomics can be used in various permutation and combinations to address different questions and obtain information at different levels. An ideal approach would involve systems biology approach, which will be discussed in next lecture. Thank you.