**DATE: 27-03-22**

**FUNCTIONAL GENOMICS**

**Q1) The GSS division contains (but is not limited to) the following types of data:**

1. **Random “single-pass read” genome survey sequences:**
   * **Random "single pass read" genome survey sequences** are GSSs that **generated** along single pass **read by random selection**.
   * **Single-pass** sequencing with **lower fidelity** can be used on the **rapid accumulation** of genomic data but with a **lower accuracy**.
   * It includes **RAPD** (Random amplification of polymorphic DNA), **RFLP**(Restriction fragment length polymorphism), **AFLP**(Amplified fragment length polymorphism) and so on.
2. **Cosmid/BAC/YAC end sequences:**
   * **Cosmid/BAC/YAC** end sequences use “**Cosmid**” or “**Bacterial artificial chromosome**” or “**Yeast artificial chromosome**” to **sequence** the genome from the **end side**.
   * These **sequences act like** very **low copy plasmids**. Sometimes there is only **one copy per cell.**
   * **Cosmid/BAC/YAC** can also be used to get **bigger clone of DNA fragment** than vectors like **plasmid and phagemid**.
3. **Exon trapped genomic sequences:**
   * **Exon trapped sequence** is used to identify genes in **cloned DNA**, and this is achieved by **recognizing** and **trapping** carrier containing **exon sequence** of DNA.
   * **Exon trapping** has **two main features**: First, it is **independent of availability of the RNA** expressing target DNA. Second, **isolated sequences can be derived directly from clone** without knowing tissues expressing the gene which needs to be identified.
   * Since **fragment of DNA** can be **inserted into sequences**, if an **exon is inserted into intron**, the transcript will be **longer than usual** and this transcript can be **trapped by analysis**.
4. **Alu PCR sequences:**
   * **Alu repetitive element** is member of **Short Interspersed Elements** (SINE) in **mammalian genome.**
   * There are about **300 to 500 thousand copies** of Alu repetitive element in **human genome**, which means one Alu element exists in **4 to 6 kb averagely.**
   * **Alu PCR** is a "**DNA fingerprinting**" technique. This approach is **rapid and easy** to use. It is obtained from **analysis of many genomic loci** flanked by Alu repetitive elements, which are **non-autonomous retrotransposons** present in **high number** of copies in **primate genomes.**
5. **Transposon-tagged sequences:**
   * **Transposon** can be used as **tag for a DNA** with a **known sequence**.
   * Transposon can **appear at other locus** through **transcription** or **reverse transcription** by the **effect of** **nuclease**.
   * This **appearance of transposon** proved that **genome** is **not statistical**, but always **changing the structure of itself**.
   * There are **two advantages** by using transposon tagging:
     1. Transposon is **inserted into a gene sequence**; this insertion is **single and intact.**
     2. **Many transposons** can be **found and eliminated** from tagged gene sequence when **transposase is analyzed**.

**Q2) What are ESTs? Explain the protocol for EST generation**

* **Expressed sequence tags** (ESTs) are short sequence reads, **typically within the range of 300–700 bp**, obtained from **randomly selected cDNA clones**.
* ESTs are **often generated** by **single-pass sequencing** of cDNA clones from one or both ends, **usually covering only a part** **of the transcript sequence**, and are relatively **prone to error.**
* They provide an **alternative to full-length cDNA sequencing.**
* Also, ESTs are an **inexpensive means of gene discovery.**
* The **randomly sequenced** ESTs allow us to make **gene discoveries** as they give **information of the transcribed regions of the gene**.
* **Protocol for EST Generation:**

**Q3) Write a short note:**

1. **STS:**
   * **Sequence-Tagges site (STS)** is a **relatively short**, **easily PCR-amplifed** sequence (200 to 500bp) which can be **specifically amplified** by PCR and **detected** in the **presence of all other genomic sequences** and whose **location** in the genome **is mapped**.
   * The **STS concept** was introduced **by Olson** et al (1989).
   * In **assessing the likely impact** of the **Polymerase Chain Reaction** (PCR) on **human genome research**, they recognized that **single-copy DNA sequences** of known map location could **serve as markers** for genetic and physical **mapping of genes** along the chromosome.
   * The **advantage of STSs** over other **mapping landmarks** is that the means of **testing for the presence** of a particular STS can be completely **described as** **information in a database.**
   * STS-based PCR produces a **simple and reproducible pattern** on agarose or polyacrylamide gel. In most cases STS markers are **co-dominant**, i.e., allow **heterorozygotes to be distinguished** from the **two homozygotes**.
   * The **DNA sequence** of an STS **may contain repetitive elements**, sequences that **appear elsewhere** in the genome, but as long as the sequences at **both ends of the site are unique** and conserved, researches can **uniquely identify this portion of genome** using tools usually present in any laboratory.
   * For example, some STSs can be used in **screening by PCR** to **detect microdeletions in Azoospermia** (AZF) genes in infertile men. Identification of **genes in elephants** could provide additional information for evolutionary studies and for **evaluating genetic diversity** in existing elephant populations.
2. **SAGE:**
   * **Serial Analysis of Gene Expression** (SAGE) is a transcriptomic technique used by molecular biologists to produce a **snapshot of the messenger RNA** population in a **sample of interest** in the form of **small tags** that **correspond** to fragments of those transcripts.
   * **Several variants** have been developed since, most notably a more robust version, **LongSAGE**, **RL-SAGE** and the most recent **SuperSAGE**. Many of these have **improved the technique** with the capture of **longer tags**, enabling **more confident identification** of a source gene.
   * The **output** of SAGE is a list of **short sequence tags** and the **number of times** it is observed. **Using sequence databases**, a researcher can usually determine, with some confidence, from **which original mRNA** and therefore **which gene the tag was extracted.**
   * **Statistical methods** can be applied to **tag and count lists** from different samples in order to **determine which genes are more highly expressed**. For example, a **normal tissue** sample can be compared **against a corresponding tumour** to determine which genes tend to be more active.
   * **Once analysed**, SAGE data **provide** both a **qualitative** and **quantitative** **assessment** of potentially **every transcript** present in a particular cell or tissue type. Following the protocols, investigators should be able to **generate unique SAGE libraries**, which can be directly compared with our **reference library**.
   * It works by **isolating short fragments** of genetic information from the **expressed genes** that are present in the **cell being studied**.
3. **cDNA:**
   * **Complementary DNA** (cDNA) is a DNA **copy of a messenger RNA** (mRNA) molecule **produced by reverse transcriptase**, a DNA polymerase that can **use** either **DNA or RNA as a template**.
   * cDNA is **not genomic DNA**, as the transcript of genomic RNA **lacks promoters and introns**.
   * **Synthesis of cDNA from mRNA** poly A tail is used as **priming site**, a **short tag** of **oligo dT with a free 3'OH group** will bind and which will be **extended by reverse transcriptase** to create cDNA.
   * **mRNA is then removed** which is achieved by treating it with **RNase enzyme** resulting in the **single stranded cDNA**.
   * The **single stranded cDNA** needs to be converted to **double stranded cDNA** which is achieved with the help of **DNA polymerase**.
   * The **free 3'OH group** for polymerase extension is provided by the **single stranded cDNA** itself by forming a **hairpin loop like structure**, which can later be **cleaved using nuclease.**
   * **Restriction endonucleases** and **DNA ligase** are then used to **clone** the sequences into **bacterial plasmids**. The **cloned bacteria** are then selected, **commonly through** the use of **antibiotic selection**. Once selected, **stocks of the bacteria** are created which can **later be grown** and sequenced to **compile the cDNA library**.
4. **DNA Microarray:**
   * The **DNA microarray** is a tool used to **determine** whether the **DNA** from a **particular individual** contains a **mutation in genes** like **BRCA1** and **BRCA2**.
   * The **chip consists** of a **small glass plate encased in plastic**. Some companies manufacture microarrays using methods similar to those used to make **computer microchips**.
   * **Each chip contains** thousands of **short**, **synthetic**, **single-stranded DNA sequences**, which together add up to the **normal gene** **in question**, and to **variants** (mutations) of **that gene** that have been found in the **human population.**
   * **DNA microarrays** were used first only as a **research tool** but scientists continue today to **conduct large-scale population studies** for example, to determine **how often individuals** with a particular mutation actually **develop breast cancer**, or to identify the **changes in gene sequences** that are most often **associated with particular diseases**.
   * Also, microarrays can also be used to **study the extent** to which certain genes are **turned on or off in cells and tissues**. In this case, instead of isolating DNA from the samples, **RNA is isolated and measured**.
   * Basically, it determines whether an **individual possesses a mutation for a particular disease**, a scientist first obtains a **sample of DNA** from the patient's blood as well as a **control sample** along with the one that does not contain a mutation in the **gene of interest**.
   * Then the **DNA is denatured** in the samples this process will **separate the two complementary strands** of DNA into **single-stranded molecules**. The next step is to **cut the long strands of DNA into smaller**, more **manageable fragments** and then to **label each fragment** by attaching a **fluorescent dye**. The individual's **DNA is labelled with green dye** and the **control** - or normal - **DNA** is **labelled with red dye**. **Both sets** of labelled DNAs are then **inserted into the chip** and allowed **to hybridize** - or bind – to the **synthetic DNA on the chip.**
   * If the **individual does not have a mutation** for the gene, **both** the red and green **samples will bind to the sequences on the chip** that represent the **sequence without the mutation**.
5. **Gene index (Gene Indices):**
   * Several efforts are under way to **condense single-read expressed sequence tags** (ESTs) and **full-length transcript data** on a large scale by means of **clustering or assembly**.
   * One **goal of these projects** is the **construction of gene indices** where **transcripts are partitioned into index classes** (or clusters) such that they are **put into the same index class** if and only if they **represent the same gene**.
   * **Accurate gene indexing** facilitates **gene expression studies** and inexpensive and **early partial gene sequence** discovery through the **assembly of ESTs** that are **derived from genes** that have **yet to be positionally cloned** or obtained directly through **genomic sequencing**.
   * The **three major gene indices** use **different EST clustering methods**:
     1. **TIGR Gene Index** uses a **stringent and supervised clustering method**, which generate **shorter consensus sequences** and **separate splice variants.**
     2. **STACK** uses a **loose** and **unsupervised clustering method**, producing **longer consensus sequences** and including **splice variants in the same index.**
     3. A **combination** of **supervised and unsupervised methods** with **variable levels of stringency** is used in **UniGene**. **No consensus sequences** are produced.
6. **Functional Genomics:**
   * **Functional genomics** is a branch that **integrates molecular biology** and **cell biology studies**, and deals with the **whole structure**, **function** and **regulation** of a gene **in contrast** to the **gene-by-gene** approach of **classical molecular biology technique**.
   * Functional genomics **focuses** on the **dynamic aspects** such as **gene transcription**, **translation**, **regulation** of gene expression and **protein–protein interactions**, **as opposed** to the **static aspects** of the **genomic information** such as **DNA sequence or structures**.
   * Functional genomics is a **study** of how **genes and intergenic regions** of the genome **contribute to different biological processes**. A researcher in this field **typically studies genes or regions** on a genome-wide scale, with the hope of tightening them down to a **list of candidate genes** or regions to **analyse in more detail.**
   * The **main objective** of functional genomics is to **resolve how the individual segment** of an organism work together to **produce a particular phenotype**.
   * It **relies on the dynamic expression** of gene products in a **definite background** such as **during a disease** or at a **specific developmental stage**. Thus, functional genomics **involved in the development** of a model link **between genotype to phenotype.**
   * There are **several specific functional genomics approaches** depending on what we are **focused on DNA level**, **RNA level**, **Protein level**, **Metabolite level**.
   * These **fundamental methods** commonly **rely on genome-based sequence** datasets using **automated algorithms running in silico** for example, the **function and functional interactions** of unknown open reading frames (ORFs) can be **predicted** by using the **principle of conserved operons**.
   * Functional genomics data are **predominantly stored** in one of two public databases such as **Array Express at EMBL-EBI**.
   * Because of the **large quantity** **of data** produced by these techniques and the desire to find **biologically meaningful patterns**, bioinformatics is crucial to **analysis of functional genomics data**.
   * Examples of **techniques in this class** are **data clustering** or **principal component analysis** for unsupervised machine learning as well as **artificial neural networks** or **support vector machines** for supervised machine learning.
7. **Importance of the Human Genome Project:**
   * The project was **hugely significant** to biology and has **influenced biological research** ever since.
   * The **main tasks** of the Human Genome Project were to **read and record** the **genetic instructions** contained within the human genome and **provide** that **information to researchers worldwide freely** and without restriction.
   * The **sequenced human genome** is now a **crucial reference** for all of human biological research. It is a **template** against which **all human genomes are compared**. Since the **full human genome sequence became available** to the scientific community, **progress** of research into human health and disease has **accelerated dramatically**.
   * The large-**scale genome research** has driven the technology advancement in **genetic testing**, **drug design**, **gene therapy**, and other **genetic related areas** such as pharmacogenetics.
   * The **detailed genetic**, **physical**, and **sequence maps** developed by the Human Genome Project also will be **critical to understanding** the biological basis of **complex disorders** resulting from the **interplay of multiple genetic** and **environmental influences**, such as **diabetes**; **heart disease**; **cancer**; and **psychiatric illnesses**, including **alcoholism.**
   * Human genome project **helps to understand** and **treat disease processes** at the **DNA level** are becoming the basis for a **new molecular medicine**. The discovery of **disease-associated genes** provides scientists with the **foundation for understanding** the course of **disease**, **treating disorders** with synthetic DNA or **gene products**, and assessing the **risk for future disease**.
   * **Clinical tests** that detect **disease-causing mutations** in DNA are the most **immediate commercial application** of gene discovery. These tests may **positively identify** the genetic origin of an **active disease**, **foreshadow** the development of a **disease later in life**, or **identify healthy carriers** of recessive diseases **such as cystic fibrosis**.
   * **Gene discovery** also provides **opportunities for developing** gene-based **treatment for hereditary and acquired diseases**.
   * Although the project **reveals potential benefits**, it **raises ethical**, **legal**, and **social issues**. The outcomes of **individuals genetic information disclosure** may **lead to confidentiality and genetic discrimination issues**. In addition, **clinical relevance** of genetic testing and **psychological effect** from the results are debatable.