PAPER 3: ASSIGNMENT
UNIT 2 FUNCTIONAL GENOMICS

Q 1. The GSS division contains (but is not limited to) the following types of data:
1. Random “single-pass read” genome survey sequences.
2. Cosmid/BAC/YAC end sequences
3 . Exon trapped genomic sequences
4 . Alu PCR sequences
5 . Transposon-tagged sequences.

GENOME SURVEY SEQUENCES

The GSS division of GenBank is similar to the EST division, with the exception that most of the sequences are genomic in origin, rather than cDNA (mRNA). The two classes (exon trapped products and gene trapped products) may be derived via a cDNA intermediate. Genome survey sequences (GSS) offer a preliminary global view of a genome since, unlike ESTs, they cover coding as well as non-coding DNA and include repetitive regions of the genome.

Genome survey sequences are usually produced and presented to NCBI by labs that conduct genome sequencing, and they are utilized as a template for mapping and sequencing of genome size pieces involved in the conventional GenBank divisions, among other factors. due to their fragmentary nature, genomic survey sequences lack long-range continuity, making it more difficult to forecast gene and marker order.

1. Random “single-pass read” genome survey sequences:

Random "single-pass read" genome survey sequences are GSSs created by random selection along with a single pass read. On the rapid accumulation of genomic data, single-pass sequencing with reduced fidelity can be employed, but it has a lesser accuracy. It contains RAPD, RFLP, and AFLP, among other factors.

2. Cosmid/BAC/YAC end sequences:

- Cosmid/BAC/YAC end sequences use Cosmid/Bacterial artificial chromosome/Yeast artificial chromosome to sequence the genome from the end side. These sequences act like very low copy plasmids that there is only one copy per cell sometimes. To get enough chromosome, they need a large number of E. coli culture that 2.5 - 5 litres may be a reasonable amount.
- Cosmid/BAC/YAC can also be used to get bigger clone of DNA fragment than vectors like plasmid and phagemid. A larger insert is often helpful for the sequence project in organizing clones.
- Eukaryotic proteins can be expressed by using YAC with posttranslational modification. BAC can’t do that, but BACs can reliably represent human DNA much better than YAC or cosmid.

Bacterial Artificial Chromosomes:

1. Bacterial artificial chromosomes (BACs, “backs”) are cloning vectors containing the origin of replication from a natural plasmid found in E. coli called the F factor a multiple cloning site, and one or more selectable markers.
2. BACs accept inserts up to 300 kb and have the advantage that they can be manipulated like giant bacterial plasmids.
3. One major difference between BACs and the plasmids is that once transformed into E. coli, the F factor origin of replication keeps the copy number of the BAC at one per cell, while the origins of typical plasmid cloning vectors drive multiple rounds of DNA replication to generate many copies of the plasmid in each cell.
4. Unlike yeast artificial chromosomes, BACs do not undergo rearrangements in the host. Therefore, they have become the preferred vector for making large clones in physical mapping studies of genomes.
5. Two disadvantages of BACs (and with other cloning vectors for E. coli) are that AT-rich DNA fragments (DNA fragments with a high proportion of A and T nucleotides) typically do not clone well, and some DNA sequences are toxic to E. Coli and, hence, are unclonable in that organism.

Yeast Artificial Chromosomes:

1. Yeast artificial chromosomes (YACs; “yaks”) are cloning vectors that enable artificial chromosomes to be made and replicated in yeast cells.
2. YAC vectors can accommodate DNA fragments that are several hundred kilobase pairs long, much longer than the fragments that can be cloned in the plasmid, cosmid, or BAC vectors. YAC vectors have been used to clone very large DNA fragments (between 0.2 and 2.0 Mb [Mb=megabase=1,000,000 bp=1,000 kb])

Cosmids:

Cosmids (vectors with features of both plasmid and bacteriophage vectors). A cosmid can accommodate DNA inserts in the range of 40–45 kb for genomics uses. A cosmid cloning vector is similar to a plasmid cloning vector, with an origin, a drug resistance marker, and a multiple cloning site, but it is introduced into host cells differently. Cosmids are frequently used as vectors when libraries are made, because they are able to hold larger inserts.

3 . Exon trapped genomic sequences:

1. Exon trapping is a technique that has been developed to identify genes in cloned eukaryotic DNA. Compared with other techniques for gene identification, exon trapping has two main characteristic features.

- It is independent of the availability of an RNA sample in which the gene to identify is expressed.
- The sequences isolated directly derive from the input DNA.

2. Some 10–20% of all genes might be expressed at very low levels or only during very short stages of development, making it difficult to isolate them based on their expression using cDNA hybridization or cDNA selection protocols.
3. Exon trapping uses an assay isolating sequences based on the presence of functional splice sites. Consequently, sequences are isolated directly from the clone under analysis without knowledge or availability of tissues expressing the gene to be identified.
4. Furthermore, because isolation is not based on hybridization, it is not possible to isolate highly similar sequences that derive from other parts of the genome, not under analysis.

4 . Alu PCR sequences:

1. Mammalian genomes contain short interspersed repeat DNAs (SINES); in man, the major family of this type is denoted the Alu repeat sequence.
2. These repeats are found ubiquitously in human DNA and are believed to number 900,000 in the haploid human genome, giving an average distance between copies of =4 kilobases. This distance may vary considerably, since Alu sequences appear to be enriched in certain chromosomal regions and deficient in others.
3. Repetitive elements homologous to the human Alu repeat are also found in rodent genomes. However, there is sufficient sequence divergence to reduce cross hybridization of human and rodent Alu repeats.
4. This difference is used to allow direct amplification of human sequences.
5. Alu PCR is a method for "DNA fingerprinting." This method is quick and simple to implement. It was discovered through the examination of many genomic loci flanked by Alu repetitive elements, which are non-autonomous retrotransposons found in large numbers in primate genomes. The Alu element can be utilized for genome fingerprinting using PCR, also known as Alu PCR.

5 . Transposon-tagged sequences:

1. The most direct process of evaluating the purpose of a specific gene sequence is to substitute it or trigger a mutation and then evaluate the results and effects. Gene replacement, sense, and antisense suppression, and insertional mutagenesis are three methods that have been established for this purpose. Among these techniques, insertional mutagenesis has proven to be extremely effective.

2. The transposon sequence is used to identify DNA sequences adjacent to the transposable element. To identify DNA adjacent to a transposon, λ phage clones was isolated that included the transposon tag. Phage libraries have continued to be useful for cloning transposon-tagged genes. Transposable elements are useful tags only if the sequences of the elements are known.
3. It has been demonstrated that transposon elements can be induced to move from one location to another in the new species. If this movement is coupled with the appearance of mutant phenotype, then the gene responsible for the phenotype can be cloned in that particular species.
4. Transposon tagging is a powerful tool that has been widely applied in several species for insertional mutagenesis.

REFERENCES:

1. Genome Survey Sequencing Data Analysis: Definition, Advantages, and Applications, CD Genomics, https://bioinfo.cd-genomics.com/genome-survey-sequencing-data-analysis-definition-advantages-and-applications.html
2. Wapenaar M.C., Den Dunnen J.T. (2001) Exon Trapping. In: Starkey M.P., Elaswarapu R. (eds) Genomics Protocols. Methods in Molecular Biology™, vol 175. Humana Press. https://doi.org/10.1385/1-59259-235-X:201
3. Cardelli M. (2011) Alu PCR. In: Park D. (eds) PCR Protocols. Methods in Molecular Biology (Methods and Protocols), vol 687. Humana Press. https://doi.org/10.1007/978-1-60761-944-4_15
4. David l. Nelson, susan a. Ledbetter, laura corboi, maureen f. Victoria, ramiro ramirez-solis, thomas d. Webster, david h. Ledbetter, and c. Thomas caskey, Alu polymerase chain reaction: a method for rapid isolation of human-specific sequences from complex dna sources, https://www.pnas.org/doi/pdf/10.1073/pnas.86.17.6686 .
5. Transposon Tagging of Plant genes, Transposon tagging, https://www.ndsu.edu/pubweb/~mcclean/plsc731/transposon/tag4.htm
6. A. Mark Settles, Chapter 11, Transposon Tagging and Reverse Genetics, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.4888&rep=rep1&type=pdf

Q 2. What are EST's? Explain the protocol for EST generation.

Expressed sequence tags (ESTs) are short sequence reads, typically within the range of 100–700 bp, obtained from randomly selected cDNA clones. Since these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes. They may be present in the database as either cDNA/mRNA sequence or as the reverse complement of the mRNA, the template strand.

APPLICATIONS OF ESTs

1. EST sequencing are the mainstream methodology for gene surveying.
2. ESTs plays an important role in gene identification, transcript mapping and description of the transcriptional activity of a tissue/cell type.
3. ESTs represents a very important body of evidence for gene prediction, and an abundant resource of molecular markers for physical mapping.
4. ESTs also used in quantification of gene expression, as the abundance of sequence reads represents each transcript which reflect the steady-state levels of these transcripts in a tissue or cell type.
5. ESTs provide reagents for downstream applications such as microarray analysis and immunoscreening of potential protective antigens.

SYNTHESIS OF ESTs:

1. Purified mRNA is used as a template for reverse transcriptase using either oligo(dT) as a primer for the first-strand synthesis or, alternatively, random hexamer primers are used.
2. After nicking the RNA–DNA hybrid with RNAse H, second-strand synthesis proceed using the RNA fragments as primers and DNA polymerase I for the extension.
3. cDNA fragments are size – fractionized to avoid cloning very small inserts.
4. Conventional libraries can be constructed with specific adapters that permit unidirectional cloning. In this case, the researcher can choose which end of the transcript cDNA will be sequenced.
5. The synthesis method and the choice of the fragment end to be sequenced have a direct implication on the sequence coverage within transcripts.
6. A 5′-end sequencing results in a higher proportion of clones covering the coding regions of the transcripts, which is quite convenient for gene discovery projects. Conversely, 3′-end sequencing results in an extensive coverage of the 3′ end of the transcripts.

Q3. Write a short note on:

1. STS

Sequence-Tagged Site (STS) is a relatively short, easily PCR-amplified sequence (200 to 500 bp) 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 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, researhers synthesize unique DNA primers complementary to the ends, amplify the region using the PCR, and demonstrate the sptcificity of the reaction by gel electrophoresis of the amplified product.

APPLICATIONS OF STSs

- STSs are useful because they define unique, detectable landmarks on the physical map of the human genome.
- For human genome research, single-copy DNA sequences of known map location served 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 is completely described as information in a database: anyone who wishes to make copies of the marker is simply to look up the STSs in the database, synthesize the specified primers, and run the PCR under specified conditions to amplify the STS from genomic DNA.
- STS-based PCR produces a simple and reproducible pattern on agarose or polyacrylamide gel.
- STS markers are co-dominant, i.e., allow heterorozygotes to be distinguished from the two homozygotes.
- STS include such markers as microsatellites (SSRs, STMS or SSRPs), SCARs, CAPs, and ISSRs.

SYNTHESIS OF STSs:

1. To map a set of STSs a collection of overlapping DNA fragments from a single chromosome or the entire genome is required.
2. The genome is first broken up into fragments.
3. The fragments are then replicated up to 10 times in bacterial cells to create a library of DNA clones.
4. The polymerase chain reaction (PCR) is then used to determine which fragments contain STSs.
5. Special primers are designed to bind either side of the STS to ensure that only that part of the DNA is copied.
6. If two DNA fragments are found to contain the same STS then they must represent overlapping parts of the genome.
7. If one DNA fragment contains two different STSs then those two STSs must be near to each other in the genome.

2. SAGE

Serial analysis of gene expression (SAGE), a functional genomics technique, is used for global profiling of gene transcripts. It relies on the preparation and sequencing of cDNA concatemers, but it does not require prior knowledge of the genes to be assayed (as with microarrays).

APPLICATIONS OF SAGE

- Once analyzed, SAGE data provide both a qualitative and quantitative assessment of potentially every transcript present in a particular cell or tissue type.
- SAGE permits the identification of transcripts that are differentially expressed as a function of time, age, genetic background or transgenic state, among other factors.
- SAGE is a powerful technique that permits a comprehensive analysis of changes in mRNA abundance.
- The results provide a snapshot of altered patterns of gene expression in response to any genetic or environmental stimulus that can be used to generate new biological hypotheses or test existing paradigms.
- Study for new markers in cancer used SAGE. Researchers compared gene expression levels in cancerous tissues with those in non-cancerous tissues to search for markers that could diagnose the pancreatic cancer at an early stage. Because the results of a SAGE analysis of a large number of representative tissues had already been published online, the scientists were able to search the database for genes preferentially expressed in pancreatic cancer. From this, they were able to identify a gene called prostate stem cell antigen (PCSA) (commonly used for diaganosis of prostate cancer).

PROTOCOL FOR SAGE:

1. mRNA is isolated from the sample and reverse transcribed using biotinylated primers to generate cDNA.
2. cDNA is bound via biotin to streptavidin microbeads.
3. cDNA is cleaved with restriction enzymes freeing it from the beads.
4. Cleaved DNA is washed out, leaving truncated cDNA bound to the beads.
5. Two oligonucleotides with sticky ends are added to the remaining truncated cDNA, in separate samples Cleaved DNA is “tagged” enzymatically, removing it from the beads.
6. Sticky ends are repaired with DNA polymerase.
7. Blunt ended tags from the two separate samples are ligated together, generating ditags with two different oligonucleotide adapter ends.
8. Ditags are cleaved to remove the oligonucleotides. Ditags will form long cDNA chains, or concatemers.
9. Transform concatemers into bacteria for replication.
10. Isolate concatemers from bacteria and sequence.

3. cDNA

Complementary DNAs (cDNAs), are double-stranded DNA molecules: one of the strands is a DNA molecule complementary to an mRNA, and the other strand is almost identical in sequence to the mRNA, differing only where a T replaces a U in the sequence.

- The clones in the cDNA library can be sequenced to identify expressed genes in the genome.
- Since these libraries contain neither introns nor non-transcribed sequences, this is the most reliable way to define the exact boundaries of exons.
- Sequences derived from these cDNAs can be compared to genomic sequences to identify regions of the genomic sequences that are transcribed.

Synthesis of cDNAs.

1. cDNA molecules are made in a two-step process. In the first step, mRNA molecules are used as a template for the production of a DNA partner strand. This step uses reverse transcriptase (RT), an enzyme that synthesizes a DNA molecule using RNA as a template.
2. Reverse transcriptase enzyme reverses roles for the molecules by using RNA as the template for DNA production.
3. mRNAs are the only RNA molecules in a eukaryotic cell that contain a poly(A) tail. The poly(A)+ mRNAs can be purified from a mixture of cellular RNAs by passing the RNA molecules over a column to which short chains of deoxythymidylic acid, called oligo(dT) chains, have been attached.
4. As the RNA molecules pass through the column, the poly(A) tails on the mRNA molecules base-pair to the oligo(dT) chains. As a result, the mRNAs are captured on the column while the other RNAs pass through.
5. The captured mRNAs are released and collected.
6. After the mRNA has been isolated, the first step in cDNA synthesis is annealing a short oligo(dT) primer to the poly(A) tail. T
7. he primer is extended by reverse transcriptase to make a DNA copy of the mRNA strand. The result is a DNA–mRNA double-stranded molecule.
8. Next, Rnase H (“R-N-aze H,” a type of ribonuclease), DNA polymerase I, and DNA ligase are used to synthesize the second DNA strand. RNase H partially degrades the RNA strand in the hybrid DNA–mRNA, DNA polymerase I makes new DNA fragments using the partially degraded RNA fragments on the single-stranded DNA as primers.
9. Finally DNA ligase ligates the new DNA fragments to make a complete chain.
10. The result is a double-stranded cDNA molecule that is a faithful DNA copy of the starting mRNA.

4. DNA Microarray

Microarray is a hybridization of a nucleic acid sample (target) to a very large set of oligonucleotide probes, which are attached to a solid support, to determine sequence or to detect variations in a gene sequence or expression or for gene mapping (MeSH).

- The microarray approach is supplanted by high-throughput RNA sequencing, RNA-Seq, which detects all transcripts in a sample, including the regulatory siRNA and lncRNA transcripts.
- Microarray technology can be used for large scale genotyping, gene expression profiling, comparative genomic hybridization and resequencing among other applications.
- Microarray technology has two major applications: gene expression analysis and genetic variation analysis.
- This technique of employing DNA chips is very rapid, besides being sensitive and specific for the identification of several DNA fragments simultaneously.
- Basically DNA microarrays consist of thousands of microscopic DNA spots (probes) that are bound to a solid surface, such as glass or a silicon chip (Affymetrix) or microscopic beads (Illumina). Labeled single-stranded DNA or antisense RNA fragments from a sample of interest are hybridized to the DNA microarray under high-stringency conditions. Each probe is identified by its location on the DNA microarray.
- The amount of hybridization detected for a specific probe is proportional to the level of nucleic acids from the corresponding genomic location in the original sample.

Synthesis of Microarray chip:

1. In this methodology, the bulk RNA is extracted from the sample and copied into stable double-stranded copy DNA, ds-cDNA, which is then sequenced using various sequencing methods.
2. The sequences obtained are aligned to reference genome sequences, available in data banks, to identify which genes are transcribed.
3. The unknown DNA molecules are cut into fragments by restriction endonucleases; fluorescent markers are attached to these DNA fragments.
4. These are then allowed to react with probes of the DNA chip. Then the target DNA fragments along with complementary sequences bind to the DNA probes.
5. The remaining DNA fragments are washed away.
6. The target DNA pieces can be identified by their fluorescence emission by passing a laser beam.
7. A computer is used to record the pattern of fluorescence emission and DNA identification.
8. Quantitatively, the results provide the expression levels for the transcribed genes.

5. Gene index (Gene Indices)

All expressed sequences (as ESTs) concerning a single gene are grouped in a single index class, and each index class contains the information for only one gene. Different clustering/assembly procedures have been proposed with associated resulting databases (gene indices): UniGene, TIGR Gene Indices, STACK.

- Regardless of how ESTs are ultimately used, their value can be significantly enhanced if the data are used to reconstruct a high-fidelity set of non-redundant transcripts. There are a number of publicly available databases that attempt to provide such analysis for some species, including UniGene, TIGR and STACK.

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 the study of how genes and intergenic regions of the genome contribute to different biological processes.
- Researcher in this field typically studies genes or regions on a “genome-wide” scale (i.e. all or multiple genes/regions at the same time), with of narrowing them down to a list of candidate genes or regions to analyse in more detail.
- The goal of functional genomics is to determine how the individual components of a biological system work together to produce a particular phenotype.
- In functional genomics, we try to use current knowledge of gene function to develop a model linking genotype to phenotype.
- It combines data derived from the various processes related to DNA sequence, gene expression, and protein function, such as coding and noncoding transcription, protein translation, protein–DNA, protein–RNA, and protein–protein interactions. Together, these data are used to model interactive and dynamic networks that regulate gene expression, cell differentiation, and cell cycle progression.

There are several specific functional genomics approaches depending on what we are focused on:

1. DNA level (genomics and epigenomics)
2. RNA level (transcriptomics)
3. Protein level (proteomics)
4. Metabolite level (metabolomics)

7. Importance of the Human Genome Project

An ambitious and expensive plan to sequence the human genome—the Human Genome Project (HGP)—commenced in 1990. As part of the HGP, the genomes of several well-studied model organisms in genetics were also sequenced. A final version of the human genome sequence was released in 2003.

Its goal was to determine the locations of all the genes in the human genome and the exact nucleotide sequence of the 3 billion nucleotide pairs that make up the genome.

IMPORTANCE:

1. The successes of the HGP have empowered researchers working with a wide range of organisms, providing them with the techniques to obtain genome sequences for those organisms quickly.
2. Research questions about gene expression, physiology, development, and so on can now be asked at the genomic level.
3. The project was about encouraging more international collaboration between scientists and facilitating the distribution of research data.
4. Those involved in the project had a responsibility to encourage public debate and provide important information to the public about the ethical, social and legal implications of looking inside the human genome.
5. The Human Genome Project also sought to develop new tools to obtain and analyse genomic data to be used beyond the project to help benefit many areas of biology.
6. 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.
7. Since the full human genome sequence became available to the scientific community, progress of research into human health and disease has accelerated dramatically.