RESULTS

AIM: The practical assessment is to be done for assembly and annotation of any genome followed by functional annotation of few predicted genes.

For this task the genome of *Escherichia coli* is used. For the assembly of a genome sequencing data has to be downloaded from NCBI Sequence Read Archive. As it is mentioned that GeneMark or Augustus online prediction tools have to be used, after the assembly is done, the genone sequence is used for annotation. The next steps are followed with taking the FASTA sequence of the genome of the species.

WORKFLOW:

1. In NCBI SRA database searched for *Escherichia coli*. The Bioproject with ID **PRJNA318589** was selected; Whole genome Illumina MiSeq sequence of Escherichia coli.
2. The run **SRR10810132** was selected.
3. Then from ENA(European Nucleotide Archive) fastq files (pair end read) of the selected run was downloaded.
4. For assembly of the genome **Genious prime** was used. The output was exported as a FASTA file.
5. The FASTA sequence of the assembled genome of *E.coli* is saved in "E coli.fasta".
6. Using the "**GeneMark.hmm with Heuristic models**" genes were annotated (predicted). As output FASTA sequences of the predicted genes were found. A total of 4793 genes were predicted in the complete genome.
7. For annotation of the predicted genes selected at random ("genes.fasta"), a tool named **OmicsBox** was used.
8. The randomly selected genes were saved in a fasta file.
9. The file was loaded to OmicsBox*.
10. Functional analysis was done in OmicsBox. First all the sequences were searched for **Interpro Scan**. Then a **Blast** run was done. The results were automatically saved in a tabular format.
11. After the blast and Interpro scan **Blast2GO** mapping was done, which was needed for Blast2GO annotation. Again the mapping and annotation results were saved in respective columns in the table. The table was exported.

*- OmicsBox was used for the convenience of annotation. In OmicsBox Interpro Scan and Blast was done. For standalone Interpro Scan, the predicted gene has to be translated into a protein sequence; for which **EMBOSS Transeq** has to be used. The output protein sequence is then used as input for InterPro Scan. EMBOSS Transeq and Interpro Scan is also included in the tools list.

OUTCOMES & INTERPRETATION:

Different outcomes were found in different tools. As it is a pipeline, the output from the first program is used as input in the next program. The input was a FASTA sequence of the whole genome in GeneMark. As result predicted genes were found (a total of 4793; in both positive and negative strands). Randomly 10 genes were selected for annotation. The genes were saved in one FASTA file. The main annotation tool was Blast2GO. There are different other methods for gene annotation. Omicsbox has all of it together in one tool. Omicsbox uses Blast, InterPro Scan and Blast2GO altogether for better annotation in a structured format. The final outcome of the annotation is saved in an excel file "Ecoli_annotation.xlsx". The table contained annotation data from Blast (E-value); Interpro scan (Interpro ID,InterPro GO IDs & InterPro GO names). Blast2GO mapping and annotation gives the GO IDs, GO names, Enzyme codes, Enzyme names. The GO IDs have prefixes F, P and C; these three represents Function, Process and Component respectively. Theses GO IDs can be accessed in EMBL-EBI Quick-GO. Gene Ontology (GO) is representation of gene and gene product. As representation of outcome one gene ontology for each predicted gene is represented using EMBL-EBI Quick-GO.

1. Gene_3: P:GO:0007165 – This gene product is associated with signal transduction. The cellular process in which a signal is conveyed to trigger a change in the activity or state of a cell. Signal transduction begins with reception of a signal (e.g. a ligand binding to a receptor or receptor activation by a stimulus such as light), or for signal transduction in the absence of ligand, signal-withdrawal or the activity of a constitutively active receptor.
2. Gene_17: P:GO:0009234 – Involved in menaquinone biosynthetic process. The chemical reactions and pathways resulting in the formation of any of the menaquinones. Structurally, menaquinones consist of a methylated naphthoquinone ring structure and side chains composed of a variable number of unsaturated isoprenoid residues. Menaquinones that have vitamin K activity and are known as vitamin K2.
3. Gene_19: F:GO:0005198- Involved in the molecular function of structural molecule activity. The action of a molecule that contributes to the structural integrity of a complex or its assembly within or outside a cell.
4. Gene_21: F:GO:0016887- The gene product has ATPase activity. Catalysis of the reaction:

   ATP + H_2O = ADP + phosphate + 2 H+. May or may not be coupled to another reaction.

5. Gene_22: F:GO:0005524- Its function is to bind to ATP. Interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator.
6. Gene_25: P:GO:0016226- Used in the biological process of iron-sulfer cluster assembly. The incorporation of iron and exogenous sulfur into a metallo-sulfur cluster.
7. Gene_27: C:GO:0019867- It is a cellular component of outer membrane. The external membrane of Gram-negative bacteria or certain organelles such as mitochondria and chloroplasts; freely permeable to most ions and metabolites.
8. Gene_48: P:GO:0006313- Involved in the biological process of DNA-mediated transposition. Any process involved in a type of transpositional recombination which occurs via a DNA intermediate.

9. Gene_55: F:GO:0003677- The gene product binds to DNA. Any molecular function by which a gene product interacts selectively and non-covalently with DNA (deoxyribonucleic acid).
10. Gene_57: P:GO:0019430- It's involved in the biological process of removal of superoxide radicals. Any process, acting at the cellular level, involved in removing superoxide radicals (O2-) from a cell or organism, e.g. by conversion to dioxygen (O2) and hydrogen peroxide (H2O2).

HOW THE DATA CAN BE USED?

Genome annotation is the process of identifying functional elements along the sequence of a genome, thus giving meaning to it. It is necessary because the sequencing of DNA produces sequences of unknown function. Genome annotation consists of describing the function of the product of a predicted gene. Gene Ontology (GO) enrichment analysis is ubiquitously used for interpreting high throughput molecular data and generating hypotheses about underlying biological phenomena of experiments. However, the two building blocks of this analysis -the ontology and the annotations evolve rapidly. Ontologies provide a uniform vocabulary for representing domain knowledge. The Gene Ontology (GO) is the most widely used ontology for specifying cellular location, molecular function, and biological process. The GO annotation is the list of all annotated genes linked to ontology terms describing those genes. The GO annotation documents all evidence that led to the association of a gene and a GO term. These annotation data or ontology data can be used by biologists and researchers for better understanding location, function and process of gene products.