Report of Certificate Course on Metagenomics & Bioinformatics organized by Internal Quality Assurance Committee (IQAC), Kirori Mal College, DU, India in association With Phixgen Pvt. Ltd., Gurugram, India (1st July 2022- 19th July 2022).

The inaugural session of the certificate course opened with a welcoming note and an introduction to IQAC and Phixgen pvt Ltd. by the program IQAC coordinator and Program Co-Ordinator, Dr. M. Ramananda Singh and Dr. Gauri Garg Dhingra. Prof. Kalpana Bharara encouraged the audience and gave introductory address. The keynote speaker of the session, Prof. Rup Lal was introduced by Prof. Rakesh Kumar Pandey. Prof. Rup Lal explained the importance of computational biology in recent scenario by delivering a talk on "Role of Computational Biology in Microbial Ecology" The inauguration session was followed by a series of lectures and hands-on sessions as follows:

Module I- Introduction to Computational biology

Session 1- Know your Hardware and Software (Introduction to Linux)

Resource Person: Dr. Roshan Kumar, P. G. Department of Zoology, Magadh University, Bodh Gaya

Dr. Roshan Kumar started the session on 'Computer platforms for Genomics and bioinformatics". He introduced the Linux operating systems and the importance of knowing hardware and software to the participants. The hands-on part of this lecture included installing a VMware player and installing ubuntu via virtualization. In the second session, Dr. Kumar discussed the concepts and hand-on exercises related to genome assembly where the participants learned how to assemble raw data using de novo assemblers (ABySS). Also, assembly statistic parameters to obtain an optimal genome assembly were explained like k-mer, N50, L50, coverage, etc.

Session 2- Database File Formats

Resource Person: Dr. Shailly Anand and Dr. Renu Solanki, Deen Dayal Upadhyaya College, University of Delhi.

Dr. Shailly Anand began the session with a brief introduction to databases, their importance in the present context of big data, and the general architecture of any database. She explained with suitable examples how a biological database is a collection of data that is structured, searchable, updated periodically, and cross-referenced. A database helps to store, maintain, annotate and

curate the data. She went ahead to elaborate on the importance of databases for storing and communicating large datasets, distributing resources, and making biological data available to scientists in computer-readable form along with ease of access. She then discussed the different basis of classification of biological databases as proposed by various workers. Thereafter, she gave an insight into the available primary and secondary databases and how each is involved in keeping nucleotide and protein sequence and structural data in an organized form. Towards the end, she gave an overview of organism-specific databases such as FlyBase, ZFIN, TAIR, etc. but also conducted a hands-on simple exercise on retrieving a protein sequence from a primary database and fetching information in different formats

Session 3- Blast and Its Types

Resource Person: Dr. Jasvinder Kaur, Gargi College, University of Delhi.

As BLAST, or the Basic Local Alignment Search Tool, has grown in importance among biologists. Dr. Jasvinder Kaur opened the workshop with a brief practical application of how to use NCBI-BLAST to compare the similarity between two sequences. A long with protein and nucleotide BLAST, rapid retrieval of sequences via NCBI was performed. It was used to find regions of local similarity between these sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. Matrix analysis was used to explain this statistical correlation. The various facets of graphical results were also explained. In-depth explanations of the Maximum Score and e-value components of the BLAST method were presented. The PAM and BLOSUM matrices were then thoroughly described with examples. The discussion of the many BLAST alternatives and their applications concluded the session. Alongside this, the participants' questions were answered.

Session 4- DNA Sequencing

Resource Person: Dr. Shailly Anand, Deen Dayal Upadhyaya College, University of Delhi.

Dr. Shailly Anand began the session with a brief introduction to DNA sequencing, its role in any genome sequencing project, and its importance and application in a wide variety of research areas. She then explained how the sequencing method evolved from Maxam and Gilbert’s chemical method in 1976 to Sanger Sequencing and way up to New Generation Sequencing technologies. With illustration, she then elaborated on the Chain termination method invented by Sanger and how it was automated to generate an electropherogram. Thereafter, she gave a detailed account of Pyrosequencing followed by Illumina Sequencing. Alongside she also

introduced other sequencing methods including Ion Torrent, SOLiD Sequencing, Helicos, Nanopore Sequencing, and even Single Molecule Real Time (SMRT) Sequencing. In the end, she gave a comparative account of each in terms of read length, cost per base, duration, merits and disadvantages

Session 5- Genome Assembly

Resource Person: Dr Utkarsh Sood, TERI, New Delhi.

One of the fundamental problems in bioinformatics is *De novo* genome assembly. Interest in the problem has been renewed in the past decade due to the advent of *next-generation sequencing* (NGS) technologies, which generate large numbers of short (100–400 bp) reads with relative low sequencing error rates. There are three main approaches for *de novo* genome assembly, the greedy strategy, the string overlap graph, and the de Bruijn graph.

The session introduced various types of Genome assemblers like Velvet, Celera, Spades, IDBA, Ray and Abyss. ABySS Assembler was explained in detail. Fastq files were provided for a hands-on practice session. The students were taught to operate the software and various parameters for a good assembly were discussed.

Module II- Genomics

Session 6- Genome Validation and Annotations

Resource Person: Dr. Gauri Garg Dhingra, Kirori Mal College, University of Delhi,

The session on "Genome Annotations" dealt with the basics of annotation like structural and functional annotations, levels of annotation and different tools, hands-on exercise on RAST (Rapid Annotation using Subsystem Technology) Server and how it provides a high-quality genome annotation for prokaryotes across the whole phylogenetic tree were explained. Once annotation is completed, genomes can be downloaded in a variety of formats or viewed online. The genome annotation provided does include a mapping of genes to subsystems and a metabolic reconstruction. In addition, students were also taught how to predict closest neighbors for the bacterial genome using RAST.

Session 7- Alignment and Phylogeny

Resource Person: Dr. Roshan Kumar, P. G. Department of Zoology, Magadh University, Bodh Gaya

Dr. Roshan explained the concepts of phylogeny and taxonomy. Following that, the students were taught about the genotypic, phenotypic, and chemotaxonomic markers used to classify bacterial

species into taxa using a polyphasic approach. The importance and drawbacks of using the 16S rRNA gene sequence to create phylogenetic trees were next discussed. The benefits of employing genomic-based approaches to analyze the 16S rRNA gene sequence were then described, including two types of Phylogenomic methods: whole genome and core genome. Following that, approaches based on the entire genome, such as Average Nucleotide Identity (ANI), Average Amino acid Identity (AAI), and Genome to Genome Distance Calculator (GGDC), were discussed. These analyses were demonstrated to the students using previously saved data, followed by the creation of a heat map and dual dendrogram using the multi-experiment viewer (MeV). After that, students were asked to raise any questions they had and to practice their own systems.

Session 8- Functional Analysis

Resource Person: Dr. Helianthous Verma, Ramjas College, University of Delhi.

Functional genomics is the 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 (i.e., all or multiple genes/regions at the same time), with the hope of narrowing them down to a list of candidate genes or regions to analyse in more detail. Dr Helianthous introduced the audience to various software with which the functional genomics can be done namely Minpath, KASS-KEGG pathway.

Session 9- Protein Structure Prediction, Modelling

Resource Person: Dr. Charu Dogra Rawat, Ramjas College, University of Delhi.

The session began by discussing the role of cellular proteins; they perform various functions ranging from DNA replication, responding to stimuli, providing structure to cells and transporting molecules from one location to another. It is highly significant, thus, to understand the relationship between sequence, structure and function of proteins in order to understand the functioning of cells under various conditions. Dr. Rawat elaborated on the procedure of homology modeling for predicting the structure of proteins from their sequences. The method on the homology of these sequences with the sequences of proteins whose structures are known and deposited in protein databanks. The participants learned about the SWISS-MODEL interface to predict the structure of an example protein sequence. The information provided as the output was analysed. Further, structure assessment and validation methods, for e. g. by analysing the Ramachandran plot of the obtained structure were also discussed. The obtained structure model was visualized in UCSF Chimera software and some of its features were also discussed.

Module III- Meta Genomics

Session 10- Metagenomics: Amplicon and whole metagenome

Resource Person: Dr. Roshan Kumar, P. G. Department of Zoology, Magadh University, Bodh Gaya

Dr. Roshan Kumar, Assistant Professor, Magadh University gave a comprehensive overview of "Microbial Ecology and Metagenomics". During his presentation, he described metagenomics' timeline and milestones. This was followed by a detailed description of developing gnotobiotic chicken, preparing the culture library using culturomics, and then conducting colonization experiments using gnotobiotic chicken. At the end, he cited examples of his work to explain the details.

Session 11- Metagenome diversity analysis by Kaiju

Resource Person: Dr Utkarsh Sood, TERI, New Delhi.

Dr. Utkarsh Sood explained the use of Kaiju as a tool for the taxonomic classification of high-throughput sequencing reads, e.g., Illumina or Roche/454, from whole-genome sequencing of metagenomic DNA. In this program, reads are directly assigned to taxa using the NCBI taxonomy and a reference database of protein sequences from microbial and viral genomes. Kaiju translates metagenomic sequencing reads into the six possible reading frames and searches for maximum exact matches (MEMs) of amino acid sequences in a given database of annotated proteins from microbial reference genome. Participants were shown the illustrations which were obtained by the software and explained how one can use the data to reproduce the figures.

Session 12- MG-RAST

Resource Person: Dr. Princy Hira, Maitreyi College, University of Delhi. Dr. Princy Hira introduced the MG-RAST pipeline which performs quality control, protein prediction, clustering and similarity-based annotation on nucleic acid sequence datasets using a number of bioinformatics tools. MG-RAST performs a protein similarity search between predicted proteins and database proteins (for shotgun) and a nucleic-acid similarity search (for reads similar to 16S and 18S sequences). MG-RAST presents the annotations via the tools on the analysis page which prepare, compare, display, and export the results on the website. A hands-on practice was also done with the provided metagenomic assembly data.

Members of IQAC coordinated various sessions. Assessments were taken on google forms after each module was completed. Also, a doubt session was kept to clear the queries of the participants. Feedback form was circulated among the participants. Valedictory session was done on the last day of the program, 19th July 2022. A review of feedback received was done. It was reported that all the participants benefitted from the course. The participant appreciated the choice and delivery of course content and the organisation of the programme. The participants also mentioned their eagerness to join an advanced course on the same discipline. Dr. Ram Sunil Lalji summarized the sessions. The session ended with a vote of thanks to all the resource persons, and participants by the IQAC Co-Ordinator, Dr. M. Ramananda Singh.

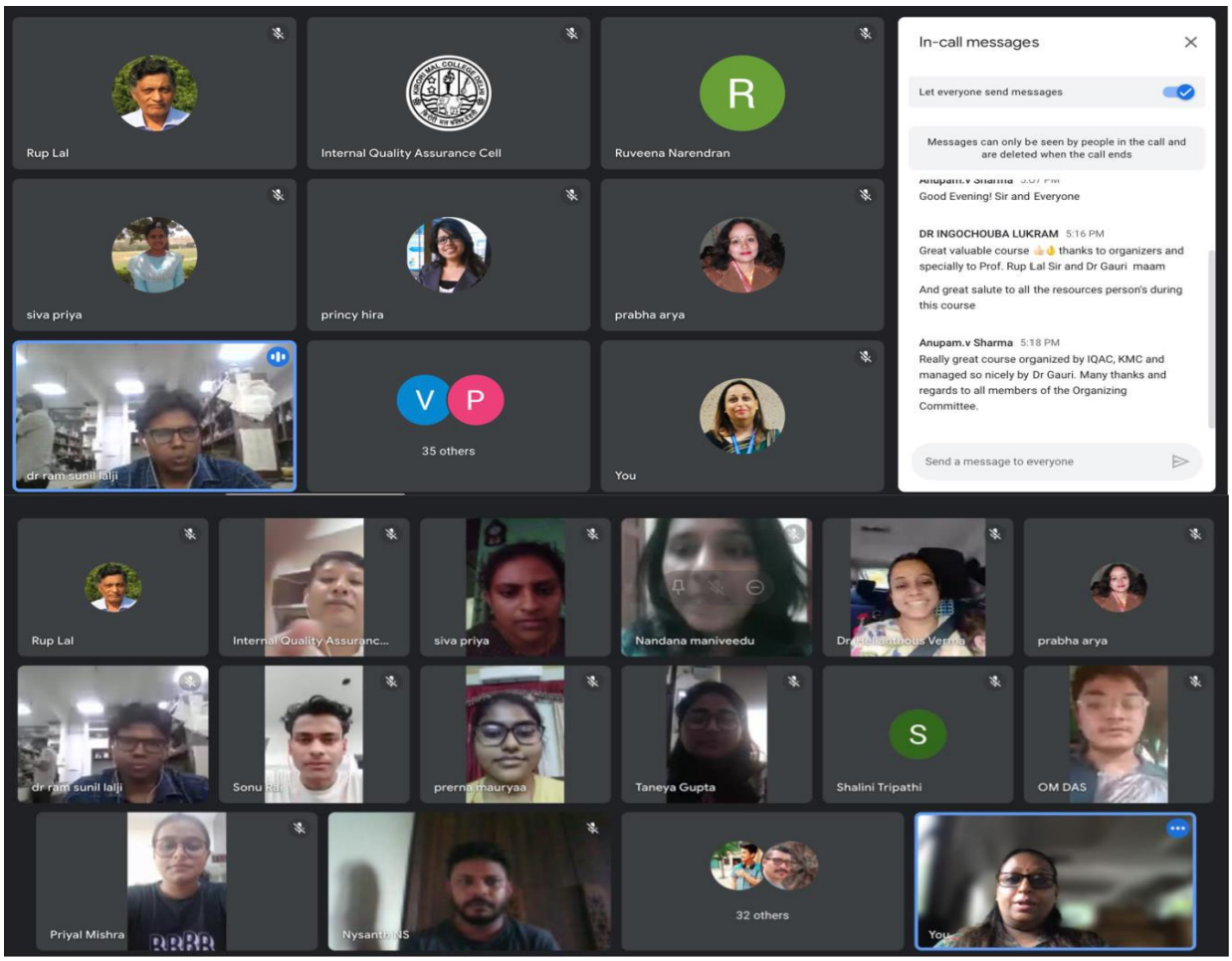

Fig. Photographs of various sessions and valedictory program of Certificate course held from 1st July 2022 to 19th July 2022.