**SRI VENKATESWARA INTERNSHIP PROGRAM
FOR RESEARCH IN ACADEMICS
(SRI-VIPRA)**

SRI-VIPRA

Project Report of 2024:

SVP - 2423

"Exploring Host-Virus Interactions: Structural and Functional Insights into Monkeypox Virus Hypothetical Proteins"
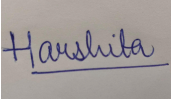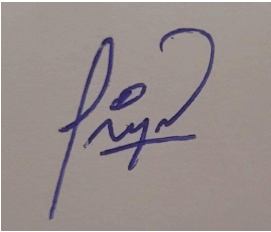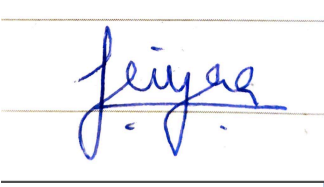
**IQAC
Sri Venkateswara College
University of Delhi
Benito Juarez Road, Dhaula Kuan, New Delhi
New Delhi -110021**
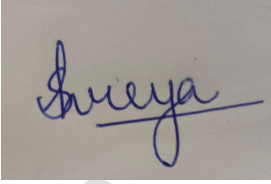
SRIVIPRA PROJECT 2024

Title : Exploring Host-Virus Interactions: Structural and Functional Insights into Monkeypox Virus Hypothetical Proteins

Name of Mentor: Dr. P. Jayaraj **Name of Department: Department of Zoology** **Designation: Assistant Professor**	

List of students under the SRIVIPRA Project

S.No	Name of the student	Roll number	Course	Signature
1	Harshita Jha	2022022	BSc Hons Zoology	
2	Jiya Dagar	1122023	Bsc. (Prog.) Life Sciences	
3	Aashleshaa Mishra	1322024	B.Sc. Hons. Biological Sciences	
4	Jeeya Arora	1323057	B.Sc. Hons. Biological Sciences	
5	Shubhi Dixit	1122903	B. Sc. Prog. Life Sciences	

		2022029	B.Sc (Hons.) Zoology	
6	Shreya Dutta			
7.	Puneeta Chhabra	2023016	B.Sc.(Hons.) Zoology	
8.	Unnati Goel	2023039	B.Sc.(Hons.) Zoology	
9.	Ananya Mengi	2022018	B.Sc. (Hons) Zoology	

Signature of Mentor

Certificate of Originality

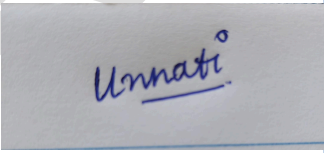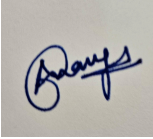This is to certify that the mentioned students from Sri Venkateswara College have participated in the summer project SVP- 2423 titled **"Exploring Host-Virus Interactions: Structural and Functional Insights into Monkeypox Virus Hypothetical Proteins"**.The participants carried out the research project work under my guidance and supervision from 15 June 2024 to 15th September 2024. The work carried out is original and carried out in an online/offline/hybrid mode.

Signature of Mentor

<u>Acknowledgments</u>

We would like to extend our heartfelt appreciation to everyone whose support has made this research journey possible.

First and foremost, our deepest gratitude goes to the administration of Sri Venkateswara College for providing us with this incredible platform to learn and grow through this internship.

TABLE OF CONTENTS

ABSTRACT

Background: Monkeypox virus(MPXV) is a small double stranded DNA virus that causes zoonotic disease called Monkeypox(MPX). It is a human pathogen for which no specific etiological treatment currently exists. The virus genome encodes several hypothetical proteins whose functions and structures remain poorly characterized. Functionally and structurally annotating these hypothetical proteins is crucial for identifying novel drug and vaccine development targets.

Materials and method: Gene expression profiles from GEO database were analyzed using GEO2R to identify Differentially Expressed Genes(DEGs) in datasets GSE19036(Keratinocyte and organoid) and GSE36854(HeLa). DEGs were visualized using Venny for venn diagram. Protein-Protein Interaction(PPI) networks were constructed using STRING and Cytoscape, with hub proteins identified by CytoHubba. Pathway enrichment was analyzed with DAVID and KEGG. Host-virus interactions were mapped through HuPoxNET, viral proteins were verified with NCBI and UniProt, and metal-binding sites were predicted using MIB2.

Result: The structural and functional analysis of hypothetical proteins found during host-virus interaction in various strains was carried out in this research. Out of these 2 hypothetical proteins A0A650BUU0 and A0A650BVD6 could be assigned a structure and function confidently and showed high scores in metal binding properties with Cu^{2+} and Cu^{2+} and Zn^{2+} respectively. A0A650BUU0 protein in MPXV is predicted as a nucleic acid-binding protein which triggers cell death and promotes virus clearance meanwhile A0A650BVD6 is predicted as soluble secreted chemokine inhibitor which inhibits chemokine activity which are involved in immune responses.

Conclusion: This research used bioinformatics tools to predict the structure, function, and metal-binding sites of hypothetical proteins. Two proteins were confidently annotated: one as nucleic acid-binding protein(A0A650BUU0) and the other as soluble secreted chemokine inhibitor(A0A650BVD6). These annotations can guide targeted docking studies for novel drug and vaccine discovery against Monkeypox.

Keywords: monkeypox virus, monkeypox, hypothetical proteins, drug target identification

INTRODUCTION

MonkeyPox

Mpox, earlier known by Monkeypox, is a zoonotic viral illness caused by the Monkeypox virus. This virus belongs to the family Poxviridae and genus Orthopoxvirus. The genus also includes variola virus

(smallpox), vaccinia virus, camelpox virus, and cowpox virus all of which are pathogenic for humans. There is 96.3% identity between the monkeypox virus genome's central region, which encodes essential enzymes and structural proteins, which means they all are highly genetically similar. Mpox was first identified in laboratory monkeys in 1958, although the first human case was documented in Democratic Republic of Congo in 1970. As a zoonotic disease, it primarily affects animals but can be transmitted to humans. With increasing incidences reported outside endemic regions, understanding the genome and the viral infection cycle of monkeypox is crucial for developing effective public health responses and therapeutic stages.

Epidemiology and Transmission:

Mpox is primarily transmitted through direct contact with infected animals, such as rodents or primates, and human-to-human transmission can occur via respiratory droplets or contact with bodily fluids. There are 2 distinct genetic clades of the virus: clade1 – Congo Basin clade, also known as Central African clade (with subclades 1a and 1b) and clade 2 – West Africa (with subclades 2a and 2b). However, the resurgence of Mpox in non-endemic regions underscore the importance of surveillance and understanding the virus's transmission dynamics.

Viral Structure:

The 200×250 nm sized monkeypox virus particle has an oval or brick shape (1) and produces two infectious particles during replication: intracellular mature viral particle and extracellular enveloped viral particle. The structure of an intracellular mature viral particle comprises a lipoprotein envelope around the viral core and some lateral body rich in proteins. It is quite stable in the external environment and is released in the environment by cell lysis. It usually aids the virus in disease transmission between different animals. On the other hand, an extracellular enveloped viral particle comprises a lipid membrane wrapped around the intracellular mature viral particle which is formed from the transport Golgi apparatus or endosomes.

Genome Structure:

The monkeypox virus genome is a linear double stranded DNA (dsDNA) molecule approximately 197 kilobases (kb) in length, making it one of the largest viral genomes. The genome comprises about 200

open reading frames (ORFs) that play vital roles in viral replication, assembly, and immune evasion (Larkin et al., 2022). The key genome regions identified are:

1. **Early genes**: Involved in regulating viral replication and host cellular manipulation, expressed shortly after infection.

2. **Intermediate genes**: Responsible for viral DNA replication and transcription, expressed after early genes.

3. **Late Genes**: Expressed during the later stages of the viral life cycle, facilitating viral assembly and release (Parker et al., 2021)

    MPXV encodes multiple immune evasion proteins that interfere with the host immune responses, highlighting its potential for pathogenicity (Alkhalil et al., 2023).

Viral Infection Cycle:

 The MPXV infection cycle consists of several stages essential for viral replication and dissemination.

1. Attachment and Entry: The infection cycle begins with the binding of MPXV to host cell receptors, primarily glycosaminoglycans. This interaction facilitates viral entry, which occurs via endocytosis or direct fusion with the plasma membrane (Ghosh et al., 2022).

2. Uncoating: Following entry, the viral core is released into the cytoplasm, where the viral membrane is broken down, allowing the viral DNA to be accessed for transcription and replication (Friedman et al., 2021)

3. Replication and Transcription: The viral DNA is transported to the cytoplasm for replication. MPXV utilizes its own DNA-dependent RNA polymerase for gene transcription, with early genes being transcribed first, followed by intermediate and late genes. Viral replication occurs in specialized structures, known as viral factories (Parker et al., 2021)

4. Assembly: New virions are assembled in the cytoplasm, where viral proteins and genomic DNA are packaged into immature particles, which mature within the cytoplasm.

5. Release: The final stage involves the release of new virions. MPXV can exit host cells through cell lysis or budding, with budding being the more common mechanism for poxviruses (Friedman et al., 2021)

Immune Evasion Strategies:

MPXV employs several strategies to evade the host immune response. Notably, the E2L protein binds to double-stranded RNA, inhibiting the activation of interferon responses (Alkhalil et al., 2023). Additionally, the virus produces decoy receptors that mimic host cytokines, diverting immune responses and facilitating viral persistence.

CELL LINES AND GENE EXPRESSIONS OF MONKEYPOX VIRUS

Monkeypox Virus has several different cell lines which are commonly used. Namely,

Vero Cells: Derived from African green monkey kidneys, these cells are frequently used for viral isolation and propagation, making them one of the primary choices for monkeypox research.

BHK-21 Cells: Baby hamster kidney cells are also employed for studying the virus and conducting vaccine research.

Hela Cells: Although primarily human cervical cancer cells, they can be used in various virology studies, including those related to monkeypox.

A549 Cells: Human lung carcinoma cells that can also support monkeypox virus replication and are useful for respiratory-related studies.Cell lines also have some shortcomings; cell lines easily undergo genotypic and phenotypic drift in culture, and this drift is particularly frequent in the more commonly used cell lines, especially those that have been deposited in cell banks for many years. In addition, through specific mutations, some subpopulations that show rapid growth or increased malignancy may arise over time. Further, the suitability of these models has come into question, as many in vitro phenomena are challenging to replicate in vivo. Interpreting the potential clinical significance of discoveries made using cell lines requires an understanding of the extent to which these cell lines represent in vivo tumors.

Our study mainly focused on HeLa because of its significant gene expression observed in our database, whilst analyzing the different accession codes of monkeypox virus. Since the establishment of the HeLa cell line in 1951, HeLa cells are permissive to various viruses, including orthopoxviruses like monkeypox. Researchers use these cells to study viral replication dynamics, allowing them to observe how the virus infects and multiplies within human cells; by observing how monkeypox affects cell behavior and induces cytopathic effects, researchers can better understand the disease process. Studying how the virus behaves in HeLa cells can contribute to the development of vaccines by identifying key antigens and immune responses that are effective against monkeypox.

Overall, the versatility and established use of HeLa cells make them a valuable tool in monkeypox virus research, helping to advance our understanding of the virus and potential interventions

VIRAL AND HOST INTERACTION:

Protein-protein interactions between host and virus play a crucial role in the understanding of infection mechanisms and the subsequent host cell immune response. Therefore, to gain deeper insights into the disease infection mechanism of MPXV in humans, we used computational models to decipher genome-scale protein-protein interactions in human-monkeypox virus pathosystem. We used a bioinformatic software called HuPoxNET which was used to identify the interactions between the differentially expressed proteins in the infected human cell lines and the proteins expressed by the different Strains of Virus as well as there functions. The database consists of 22 MPXV strains' interaction with human proteins. The top 20 hub genes' IDs identified from Cytoscape were converted into SWISSPROT (link) IDs in ShinyGo and then searched for in the database. The results revealed 47 interactions when compared with all 22 strains.The most common Viral proteins were Ser/Thr Kinase, Chemokine Binding Proteins, MPXVgp029, MPXVgp191 and MPXVgp164, we were also able to find 4 hypothetical genes as well after analyzing the viral-host interactions.

Differential Gene Expression

Differential gene expression (DGE) refers to the process where genes exhibit varying levels of transcriptional activity across distinct biological states, tissues, or temporal conditions. This variation in

gene expression is essential for cellular function, differentiation, development, and response to environmental or physiological changes. Investigating differential gene expression provides critical insights into the molecular mechanisms underlying diverse biological phenomena, including disease pathogenesis, tissue-specific functions, and developmental processes.

Differential gene expression (DGE) is crucial in disease diagnostics as it provides a molecular-level understanding of pathological conditions by revealing changes in gene activity between healthy and diseased tissues. Through the comparison of gene expression profiles, DGE analysis identifies genes that are either upregulated or downregulated in specific diseases, offering valuable insights for several diagnostic applications. One of the primary uses of DGE is in **biomarker discovery**, where differentially expressed genes serve as indicators for early disease detection, prognosis, and monitoring of disease progression. In cancer, for instance, the identification of specific gene expression signatures, such as the overexpression of oncogenes, allows for more precise tumor classification and improved prognosis predictions. Moreover, DGE is integral to the field of **precision medicine**, as it enables the customization of treatments based on an individual's unique genetic expression patterns. This is particularly important in conditions like cancer, where different molecular subtypes may require distinct therapeutic approaches. DGE also contributes to a better understanding of disease mechanisms by shedding light on altered biological pathways, facilitating the development of targeted therapies. Additionally, DGE is useful for **early detection**, as gene expression changes often occur before clinical symptoms manifest, allowing for earlier diagnosis and potential intervention in diseases such as neurodegenerative disorders. Finally, DGE aids in the **molecular classification** of diseases, distinguishing between subtypes that may have distinct clinical outcomes, as seen in breast cancer. Overall, the ability to analyze and interpret differential gene expression is indispensable for enhancing diagnostic accuracy, guiding treatment decisions, and advancing personalized medicine.

Analytical Approaches to Differential Gene Expression

The investigation of DGE typically relies on high-throughput molecular techniques, each offering distinct advantages and considerations:

- **RNA Sequencing (RNA-Seq)**: RNA-Seq has emerged as the gold-standard method for quantifying gene expression. This next-generation sequencing technology allows for an unbiased, genome-wide analysis of transcript abundance, offering high sensitivity, resolution, and the capacity to detect novel transcripts and alternative splicing events.
- **Microarray Technology**: Although microarrays have been largely superseded by RNA-Seq, they remain a cost-effective option for large-scale gene expression studies. Microarrays rely on hybridization of labeled RNA to complementary probes on a solid surface, enabling the measurement of predefined sets of genes.
- **Quantitative PCR (qPCR)**: qPCR is a highly sensitive and specific method for quantifying the expression of a limited number of genes. Often used as a validation tool following RNA-Seq or microarray experiments, qPCR enables precise measurement of mRNA levels in selected targets.

Differential Gene Analysis for the Project

The datasets utilized in the Geo2R analysis are categorized into two distinct groups: infected and normal. The infected group typically represents cell lines exposed to pathogenic conditions, while the normal group serves as a control, representing non-infected or baseline conditions. Differential gene expression analysis is performed by comparing these two groups, aiming to identify genes whose expression levels significantly differ between the infected and normal states.

Volcano plots are employed to visually represent the results of this differential expression analysis, plotting the log2 fold change against the -log10 p-value for each gene. This graphical representation highlights genes that are either upregulated or downregulated in response to infection. Genes that exhibit significant changes in expression, as indicated by their position on the plot (typically with large fold changes and low p-values), are considered of particular interest. Upregulated genes (positioned on the right side of the plot) are those whose expression is elevated in the infected condition relative to the normal, while downregulated genes (positioned on the left) show reduced expression in the infected state.

After normalization, Geo2R applies statistical methods, such as linear modeling or t-tests, to compare the gene expression levels between the infected and control groups. This process generates several key metrics:

- **Log2 fold change (log2FC):** This metric quantifies the magnitude of change in gene expression between the two conditions. A positive log2FC indicates that a gene is upregulated (more highly expressed) in the infected condition compared to the control, while a negative log2FC signifies downregulation.
- **P-value:** The p-value assesses the statistical significance of the observed differences. A low p-value (typically < 0.05) suggests that the difference in gene expression is unlikely to have occurred by chance.
- **Adjusted p-value (FDR):** To account for multiple testing, an adjusted p-value, such as the false discovery rate (FDR), is often used. This helps minimize the risk of false positives, where genes are erroneously identified as differentially expressed.

Identifying these differentially expressed genes (DEGs) is critical, as they may play key roles in the biological response to infection. Researchers can gain insights into the underlying mechanisms of disease progression by further investigating the molecular pathways associated with these DEGs. In particular, understanding the regulatory networks and signaling pathways disrupted or activated in infected cells can reveal potential therapeutic targets. This could lead to the development of novel interventions aimed at modulating gene expression or restoring normal cellular functions, thus paving the way for innovative treatment strategies for the underlying condition.

PROTEIN-PROTEIN INTERACTION

Protein-protein interactions (PPIs) are fundamental to nearly all biological processes, serving as critical mediators in cellular communication, signaling pathways, and the assembly of multi-protein complexes. Understanding these interactions is particularly important in the context of host-pathogen relationships, where viruses exploit host proteins to facilitate their life cycle, evade immune responses, and manipulate

host cellular machinery. PPIs are central to the functioning of biological systems, facilitating a myriad of processes such as signal transduction, immune responses, and cellular metabolism. In the context of host-pathogen interactions, viruses like Monkeypox exploit PPIs to manipulate host cellular machinery, promote viral replication, and evade immune detection. Understanding these interactions is crucial for uncovering the strategies employed by viruses to thrive within their hosts. In this report, we explore the intricate dynamics of host-virus interactions, focusing specifically on the hypothetical proteins of the Monkeypox virus (MPXV). By investigating the structural and functional characteristics of these viral proteins, we aim to uncover their potential roles in PPI networks within the host environment. This insight not only enhances our understanding of MPXV biology but also contributes to the broader field of virology and host-pathogen interactions, paving the way for the development of targeted therapeutic strategies against viral infections. Through a combination of bioinformatics, structural biology, and experimental validation, this report seeks to elucidate the mechanisms by which MPXV may hijack host protein networks to its advantage.

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a powerful bioinformatics tool that integrates known and predicted PPI data, allowing researchers to visualize and analyze the complex networks formed by interacting proteins. By leveraging STRING, we can elucidate the potential interactions between Monkeypox virus hypothetical proteins and host proteins, identifying key players in the viral life cycle and host immune response.

This report delves into the structural and functional insights of Monkeypox virus hypothetical proteins, utilizing STRING to map and analyze their interactions within the host system. By highlighting these PPI networks, we aim to enhance our understanding of how MPXV may manipulate host cellular processes, ultimately contributing to the development of effective therapeutic interventions against viral infections.

AIM

Integrative Analysis of Monkeypox Virus-Human Protein Interactome for the Identification of Putative Molecular targets.

MATERIAL AND METHODS

Data procurement

The gene expression profile datasets used in this investigation were collected from NCBI's database-Gene Expression Omnibus (GEO) [https://www.ncbi.nlm.nih.gov/geo/]. The gene expression profles GSE36854 and GSE219036 were selected after careful contemplation for the human cell lines infected with monkeypox virus. The first dataset is GSE36854, host is homo sapien, and cell line used is HeLa. The second dataset is GSE219036, host is homo sapien, and cell line used is keratinocyte and colon organoid.

Data processing of DEGs

An online freely available tool, GEO2R (www.ncbi.nlm.nih.gov/geo/geo2r/), was used to identify the DEGs for human cell line by comparing the infected with normal specimens. DEGs were defined as genes that satisfied the cutoff criterion of adjusted P0.05 and |logFC|>1.0. For each dataset, statistical analysis was done, and the Venn diagram online tool (bioinformat-ics.psb.ugent.be/webtools/Venn) was employed to find the common DEGs shared by both infected and normal human cell lines.

DATA ACQUISITION FROM GEO2R, DEG ANALYSIS AND VENN DIAGRAM

Acquisition of array data

The dataset was retrieved from the GEO database. The GEO database is an open resource database from NCBI. Dataset GSE21001 was used in this study. Different keywords, including "Monkeypox virus," "Infection," and "Microarray," were used to search the GEO dataset.

Extensive Literature survey and procurement of cell-line data

A reliable monkeypox virus cell line data was sourced from the supplement files of a research

paper titled "Evaluation of differentially expressed genes during replication using gene expression landscape of monkeypox-infected MK2 cells: A bioinformatics and systems biology approach to understanding the genomic pattern of viral replication". The data

procured from this file was used for further comparative analysis using various tools.

Identification and Retention of GEO-2-R database

GEO-2-R is an interface between R and the Gene Expression Omnibus (GEO) which is a publicly available function-related genomics repository containing gene expression data generated by microarray technology. DEGs were genes that satisfied the cutoff criterion of adjusted P0.05 and |logFC|>1.0. For each dataset, statistical analysis was done, and the Venn diagram online tool (bioinformat-ics.psb.ugent.be/webtools/Venn) was employed to find the common DEGs shared by both infected and normal human cell lines.

After extensive literature survey, two reliable GEO-2-R (Accession Numbers- GSE36854 for HeLa and GSE219036 for Keratinocytes) for Monkeypox Virus were identified and downloaded from the GEO Database.

Data preprocessing and analysis of DEGs

The raw gene expression data was assessed using the statistical program GEO2R. GEO2R is a tool for analyzing raw gene expression data using the GEO query and imma R packages. Several statistical plots were created for data expression following data analysis. Volcano plots, mean difference (MD) plots, expression density plots, venn diagrams, adjusted p-value histograms, box plots, moderated t-statistic q-q (quantile-quantile) plots, and mean-variance trend plots were among the plots that were created.

Identification of common DEGS and comparison of common genes using Venny

A comparative study was done with the top 200 upregulated and downregulated genes, comparing the mock and the infected genes present in both HeLa and Keratinocytes. The study was further transferred to a software named Venny, a web-based tool used for visualizing and analyzing overlaps between different sets of data, particularly in the context of bioinformatics in the form of Venn diagrams. The comparison of

both the upregulated and downregulated genes resulted in a total of 594 common proteins present between both datasets.

CONSTRUCTING PROTEIN-PROTEIN INTERACTION (PPI) NETWORK AND IDENTIFYING KEY HUB GENES

The Search Tool for the Retrieval of Interacting Genes (STRING) database (Version 12.0, available at http://string-db.org) was used to predict potential protein interactions among gene candidates. Theresulting molecular interaction networks were visualized with Cytoscape (Version 3.10.1) an open source bioinformatics software . To explore the key hub genes within the protein- protein interaction (PPI) network , the CytoHubba plugin for Cytoscape was employed . This tool features a user-friendly interface for investigating important nodes in biological networks and uses eleven different methods for computation. The top 20 nodes ranked by their degree identified as hub proteins.Significant nodes within the entire PPI network were then determined using MCODE, with the following criteria: degree cut-off=2, node density cut-off =0.1, node score cut off=0.2, and max depth=100.This cluster was subsequently analyzed using KEGG and DAVID for further insights.

Host and Viral protein interactions in Hupoxnet

Hupoxnet Database (*HuPoxNET (kaabil.net)*) was used to identify the interactions between the differentially expressed proteins in the infected human cell lines and the proteins expressed by the different strains of virus. Hupoxnet is a specialized database that plays a crucial role in studying the host-protein interactions for the monkeypox virus. This resource enables researchers to explore the complex interactions between viral proteins and host cellular proteins, which is essential for understanding the mechanisms of viral infection and pathogenesis. The database consists of viral proteins of 22 MPXV strains interacting with human proteins. The top 20 hub genes' IDs identified from Cytoscape were then converted into their SWISSPROT IDs in ShinyGo *(ShinyGO 0.80 (sdstate.edu), Ge SX, Jung D & Yao R, Bioinformatics 36:2628–2629, 2020*) and then searched for in this database. The results revealed 47 interactions when compared with all the 22 strains.

Function prediction using sequence analogy

The most basic step to understanding the function of an unknown protein is by looking for its structural homologs in different genomics and proteomics-based databases. This approach works on the hypothesis that an unknown protein with a sequence analogy to a known protein may have a similar function as the known protein. The sequence similarity search was performed via protein BLAST (pBLAST) against the non-redundant database. Generally, HPs contain low identity as compared to other known or annotated proteins [10, 11]. The predicted functions according to sequence analogy through pBLAST are listed in Table .

Function and functional domain prediction through sequence analysis

The precise function of a protein can be determined using the information about functional domains in the HPs, various bioinformatics tools like INTERPROSCAN and CDD, NCBI were used to detect functional domains in HPs and classify the 3 HPs into family and superfamily.

InterProScan uses protein sequence in FASTA format to predict the family to which the unknown protein belongs and the domain present in it.

CDD is a protein annotation resource with curated domain models and alignments, integrating 3D-structure data for domain insights.

The predicted functional domains in HPs are listed in Table

Furthermore, the ENH proteins were given as a query to the STRING database (version 11.0) with medium confidence (0.40) to identify functions based on the homolog hits and the interactions among the proteins. STRING database is an integrated resource of experimental and predicted protein-protein interactions (PPI). Currently, STRING comprises more than 2,000 million interactions of 24.6 million proteins from 5,090 organisms [12,13]

Pathway Enrichment Analysis

Pathway enrichment analysis was performed for the 19 hub genes using KEGG Pathway analysis (https://www.genome.jp/kegg/pathway.html) and is presented in a tabular form in Table .

MIB2: Metal Ion-Binding site prediction and modeling server
 Overview:

Proteins that interact with metal cofactors, known as metalloproteins, play crucial roles in various cellular processes, including protein folding, enzymatic catalysis, and cellular signaling. Therefore, understanding the specificity of metal ions binding to proteins is key to unraveling the mechanisms of these biological functions. Interactions between proteins and specific metal ions are essential processes in many physiological activities. Computational methods for identifying MIB(Metal Ion Binding) sites can be broadly grouped into (i) sequence-based, (ii) structure-based, and (iii) both sequence/structure-based based on the features used for training. One such computational tool to specify metal binding sites is the MIB2 application. It is a structure-based method that considers the three-dimensional neighborhood surrounding the metal ion. For proteins lacking solved structures, MIB2 gains predicted structures by the (PS)2 modeling method or by collecting from the AlphaFold Protein Structure Database.

The MIB2 applications have been used to validate the potential of the designed site to coordinate metal ions. The ability of MIB2 to construct a 3D protein structure and provide a larger choice of specific types of metal ions enables users to conveniently investigate protein functions and design novel binding sites for proteins.

There are many different types of metal ion-binding proteins, and those that bind the most common metal ions, such as iron, usually regulate essential functions in physiological processes. However, the functions of many metal ion-binding proteins remain unclear. Therefore, identification of the positioning of metal-binding residues in 3D space is important for clarifying the ion specificity of a protein and will provide general insight into the possible roles of metal ions in protein function

MIB2, an enhanced version of the MIB prediction system, supports the prediction of binding sites for 18 different metal ions, including Ca^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Zn^{2+}, Cd^{2+}, Fe^{2+}, Ni^{2+}, Hg^{2+}, Co^{2+}, Cu^+, Au^+, Ba^{2+}, Pb^{2+}, Pt^{2+}, Sm^{3+}, and Sr^{2+}. MIB2 utilizes the fragment transformation method to compare local structural regions of query proteins with MIB templates. These templates are derived from protein-metal ion complexes found in the Protein Data Bank.

Methodology of MIB2: Metal Ion-Binding site prediction and modeling server

The MIB2 consortium uses protein sequence in FASTA format to predict the Metal-binding site. Hence, the FASTA sequence of Protein, hypothetical protein PDLMKLCO_00160, and hypothetical protein PDLMKLCO_00029 were uploaded on the modeling server. Each residue of the query protein is assigned a binding score which is composed of sequence and structure conservation measures. When the binding score of a residue is higher than a specified threshold, this residue is predicted to be a metal-binding residue. Therefore the top two binding scores of the metal ions are considered for further

analysis. Based on the local 3D structure alignment between the query protein and metal ion-binding template, the metal ion in the metal-binding template can be transformed into the query protein structure. In addition to predicting binding residues and estimating the binding scores, MIB also predicts the docked position of the metal ion in the protein structure. In the docking window, the aligned templates are listed in a table. By clicking the visualization icon of a specific template, this template structure will be shown above the table, and the corresponding aligned residues and predicted metal ions will also be shown in the query protein structure.

RESULT

DATA ACQUISITION FROM GEO2R, DEG ANALYSIS AND VENN

Array data was acquired from the GEO database, specifically dataset GSE21001, with keywords "Monkeypox virus," "Infection," and "Microarray. " Cell-line data for monkeypox virus was retrieved from a research paper supplement file. Two reliable GEO-2-R datasets (Accession Numbers- GSE36854 for HeLa and GSE219036 for Keratinocytes) were identified and downloaded for further analysis. Data preprocessing and analysis of differentially expressed genes (DEGs) was conducted using GEO2R, generating various statistical plots. A total of 594 common proteins were identified between HeLa and Keratinocytes datasets. A protein-protein interaction network was constructed using String database to explore relationships among DEGs, with the top 10 hub genes extracted using the DMNC method in Cytoscape. Overall, a comprehensive analysis was performed on gene expression patterns in monkeypox virus-infected cells.

PPI network construction and identification of hub genes

After uploading 594 common genes to the STRING database, we generated a protein-protein interaction network consisting of 573 nodes and 1,423 interaction edges. The average node degree was 4.97, and the average local clustering coefficient was 0.394, with a PPI enrichment p-value of less than 1.0e-16. The identified PPI relationships were imported into Cytoscape software, where the Cytohubba plug-in was

utilized to identify the top 20 hub genes based on degree. The analysis yielded the following hub genes: CCL20, JUN, FYN, CXCL2, NFKBIA, CXCL1, CXCL8, FOSL1, RELB, CD86, RPS27A, IL1B, TLR2, CDH1, BRCA1, MYC, CD44, PTGS2, TNFAIP3 and CCND1. Additionally, two functional clustering modules were identified using the MCODE plug-in, with module 1 scoring 12.308 points and module 2 scoring 3.000 points.

Pathway enrichment analysis

Pathway enrichment analysis was performed for 19 hub genes using KEGG pathway analysis and is presented in a tabular form . These findings provide insights into the underlying biological processes and potential mechanisms relevant to the gene list. The overexpression of 19 genes associated with viral proteins was analyzed concerning the disease.

Viral- Human Protein Interaction

The most common viral proteins of MPXV Strains were Ser/Thr Kinase, Chemokine Binding Proteins, MPXVgp029, MPXVgp191 and MPXVgp164. The primary host proteins that they interacted with were transcription factors, chemokines and interleukins. Interestingly, 4 hypothetical Proteins were also found in the reactome, 3 were found with MPXV strain MN648051 : hypothetical Protein PDLMKLCO_00160, hypothetical protein PDLMKLCO_00029 and Hypothetical protein PDLMKLCO_00030 and 1 was found with MPXV strain ON736420 : Hypothetical Protein CEAHHEIO_00030.

With subsequent investigation a few of these hypothetical proteins were shown to have extremely similar functions as some of the known proteins, and therefore, had to be discarded. Only 2 hypothetical proteins were left which were still ambiguous in nature. The two selected proteins were hypothetical protein PDLMKLCO_00160 and hypothetical protein PDLMKLCO_00029.

Further investigation into these hypothetical proteins could prove invaluable in antiviral drug and vaccine discovery, as well as in understanding viral-host interactions, offering a promising avenue for future research.

Analysis of the Molecular Docking and prediction by the MIB2 Modelling Server (result)

For hypothetical protein PDLMKLCO_00160
Mg+2 and Co+2 had the highest binding scores. Henceforth, the top two scores of the respective metal ions were considered and docked.

S.No.	Metal	Binding residue	Score
1	Mg+2	64N,127E	4.914
		64N,127E	3.230
2	Co+2	52H,66S,68D	3.379
		40E,129D	3.300

For hypothetical protein PDLMKLCO_00029
Mg+2 had the highest binding score. Henceforth, the top two scores of the respective metal ions were considered and docked.

S.no.	Metal	Binding Residue	Score
1	Mg+2	7V,11D	2.069
		35N,37Y	1.625

Identifying metal ion binding sites on a hypothetical protein is crucial due to the essential roles metal ions play in protein structure, stability, and function. Metal ions like Zn^{2+}, Ca^{2+}, and Fe^{2+} can stabilize protein structure by forming coordination bonds with specific amino acids, contributing to proper folding and maintaining functional conformations. Metal ions may mediate protein-protein or protein-ligand interactions by stabilizing complexes. If the protein is linked to diseases, these sites could offer targets for drug discovery. Additionally, many enzymes (metalloenzymes) rely on metal ions for catalytic activity, suggesting that identifying these sites could predict the enzymatic functions of the protein. Metal ion binding can also regulate protein activity by inducing conformational changes or modulating functional states. Understanding these binding sites can reveal potential biological roles in processes such as cellular signaling, redox reactions, and electron transport. Furthermore, comparing metal ion binding sites across species can provide evolutionary insights into protein function. The identification of these sites also aids in experimental design, including mutagenesis studies and structural analysis, while offering biotechnological applications such as protein engineering for biosensors or bioremediation.

		QNP12897.1	MPXVgp029		Truncated K3L homolog		
		QNP13032.1	MPXVgp164	Ser/thr kinase		P05412	Transcription factor Jun
7	JX878425	AGR37986.1	chemokine binding protein			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		AGR37825.1	C3L			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		AGR37979.1	tumor necrosis factor-receptor like protein			P78556	C-C motif chemokine 20
		AGR37960.1	B3R			P05412	Transcription factor Jun
		UVB80118.1	Ser/thr kinase			P05412	Transcription factor Jun
		UVB80139.1	Chemokine binding protein			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
8	MT903343	QNP13599.1	MPXVgp191			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		QNP13441.1	MPXVgp029			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		QNP13576.1	MPXVgp164			P05412	Transcription factor Jun
9	ON745215	USC26247.1	MPXVgp191			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		USC26082.1	MPXVgp029			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		USC26220.1	MPXVgp164			P05412	Transcription factor Jun
10	NC00310	NP_536591.1	Ser/thr kinase				
		NP_536618.1	Chemokine binding protein				
		NP_536456.1	C3L [Monkeypox virus Zaire-96-I-16]				
		NP_536610.1	K1R [Monkeypox virus Zaire-96-I-16]				
11	OP212528	UUW25914.1	Ser/thr kinase				
		UUW25936.1	Chemokine binding protein				
12	DQ011157	AAY97768.1	ser/thr kinase				
		AAY97796.1	chemokine-binding protein				

1	ON736420	URZ86264.1	MPXVgp164	Ser/thr kinase		P05412	Transcription factor Jun
		URZ86245.1	MPXVgp151	hypothetical protein		P19876	C-X-C motif chemokine 3
		URZ86245.1	MPXVgp151	hypothetical protein		P19875	C-X-C motif chemokine 2
		URZ86245.1	MPXVgp151	hypothetical protein		P10145	Interleukin-8
		URZ86245.1	MPXVgp151	hypothetical protein		P09341	Growth-regulated alpha protein
		URZ86295.1	MPXVgp191	Chemokine binding protein		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
	1luz	URZ86113.1	hypothetical protein CEAHHEIO_00030	hypothetical protein		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		URZ86114.1	MPXVgp029	IFN resistance, PKR/eIF-alpha inhibitor (C		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
2	MT903345	QNP13938.1	MPXVgp164	Ser/thr kinase		P05412	Transcription factor Jun
		QNP13961.1	MPXVgp191	Chemokine binding protein		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		QNP13803.1	MPXVgp029	IFN resistance, PKR/eIF-alpha inhibitor (C		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
3	MT903342	QNP13418.1	MPXVgp191	Chemokine binding protein		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		QNP13395.1	MPXVgp164	Ser/thr kinase		P05412	Transcription factor Jun
4	ON563414	URK20629.1	MPXVgp191	Chemokine binding protein		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		URK20464.1	MPXVgp029	IFN resistance, PKR/eIF-alpha inhibitor (C		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		URK20602.1	MPXVgp164	Ser/thr kinase		P05412	Transcription factor Jun
5	MN648051	QGQ59900.1	serine/threonine-protein kinase 1			P05412	Transcription factor Jun
		QGQ59880.1	hypothetical protein PDLMKLCO_00160			P19876	C-X-C motif chemokine 3
						P19875	C-X-C motif chemokine 2
						P09341	Growth-regulated alpha protein
						P10145	Interleukin-8
		QGQ59932.1	chemokine-binding protein			P62979	Ubiquitin-ribosomal protein eS31 fusion protein
	iluz	QGQ59749.1	hypothetical protein PDLMKLCO_00029				
		QGQ59750.1	hypothetical protein PDLMKLCO_00030				
6	MT903340	QNP13055.1	MPXVgp191	chemokine-binding protein		P62979	Ubiquitin-ribosomal protein eS31 fusion protein
		QNP12897.1	MPXVgp029	Truncated K3L homolog			

For hypothetical protein PDLMKLCO_00160

Volcano plot
GSE219036: Virological characterization of the
2022 outbreak-causing...
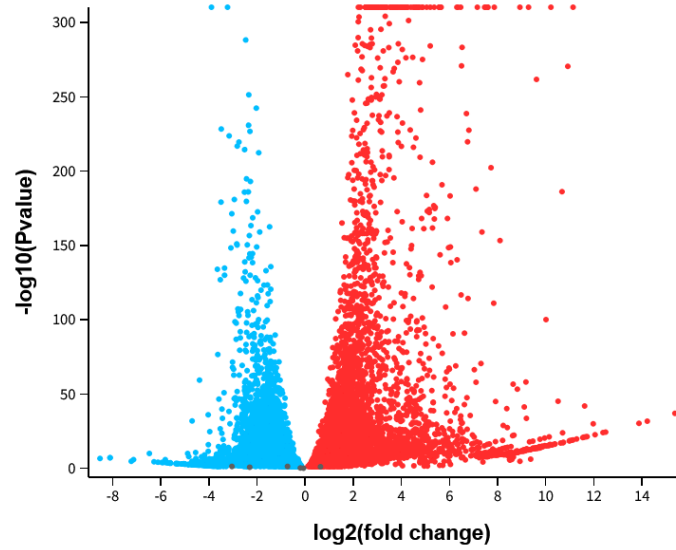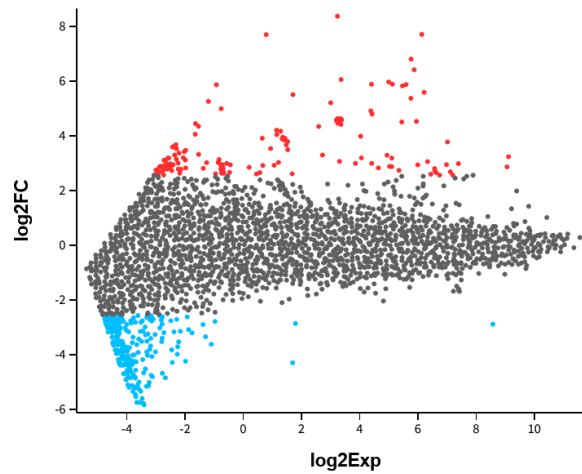**Infected vs Normal, Padj<0.05**

Figure 1: Volcano Plot of Differentially Expressed Genes in Keratinocytes. The red showcases the number of upregulated genes and the blue is for the downregulated genes.
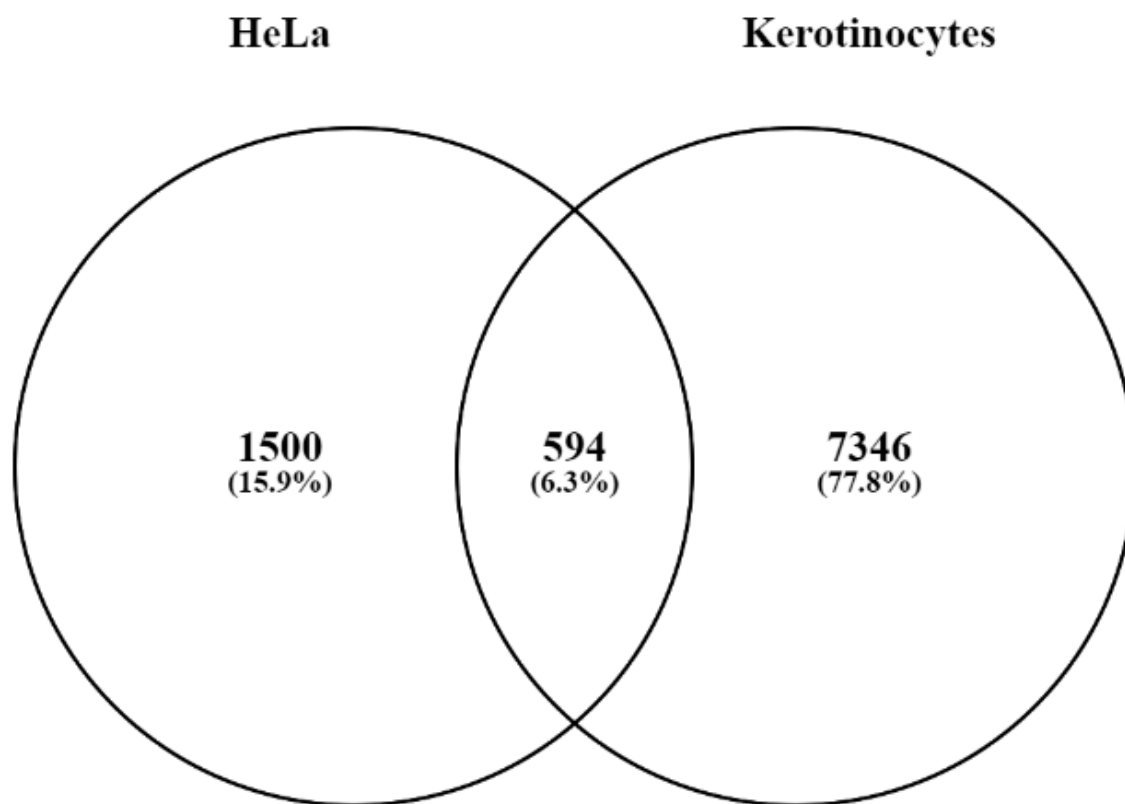
Meandiff plot
GSE36854: Comparison of host cell gene expression in cowpox,...
Infected vs Normal, Padj<0.05

Figure 2: Volcano Plot of Differentially Expressed Genes in HeLa Cells. The red showcases the number of upregulated genes and the blue is for the downregulated genes.

Figure 3: Venn Diagram Showing Overlap of Differentially Expressed Genes in HeLa and Keratinocyte Datasets.

Top differentially expressed genes ⍰

Download full table   Select columns

GeneID	padj	pvalue	lfcSE	stat	log2FoldChange	baseMean	Symbol	Description
55629	0.00	0.00	0.0695	45.5	3.16	2277	PNRC2	proline rich nuclear re…
6421	0.00	0.00	0.0752	52.6	3.96	7991	SFPQ	splicing factor proline…
23334	0.00	0.00	0.081	55.6	4.51	2587	SZT2	SZT2 subunit of KIC…
3725	0.00	0.00	0.0668	50.1	3.34	7816	JUN	Jun proto-oncogene, …
8880	0.00	0.00	0.0668	37.6	2.51	1977	FUBP1	far upstream element…
4170	0.00	0.00	0.0603	47.6	2.87	11544	MCL1	MCL1 apoptosis regu…
55249	0.00	0.00	0.0744	42.8	3.19	1677	YY1AP1	YY1 associated prote…
3310	0.00	0.00	0.2196	50.7	11.14	8393	HSPA6	heat shock protein fa…
64222	0.00	0.00	0.0905	44.5	4.03	1226	TOR3A	torsin family 3 memb…
27	0.00	0.00	0.0722	62.4	4.5	7712	ABL2	ABL proto-oncogene …
163590	0.00	0.00	0.071	46	3.27	2328	TOR1AIP2	torsin 1A interacting …
91156	0.00	0.00	0.242	42.2	10.22	3146	IGFN1	immunoglobulin like a…
55432	0.00	0.00	0.0788	46.5	3.67	2013	YOD1	YOD1 deubiquitinase
467	0.00	0.00	0.1121	50.4	5.65	2457	ATF3	activating transcriptio…
130872	0.00	0.00	0.0905	44	3.98	1341	AHSA2P	activator of HSP90 A…
9173	0.00	0.00	0.1961	40.1	7.86	1045	IL1RL1	interleukin 1 receptor …
6574	0.00	0.00	0.0742	55.7	4.13	4495	SLC20A1	solute carrier family 2…
1195	0.00	0.00	0.0782	48.7	3.81	4249	CLK1	CDC like kinase 1
55209	0.00	0.00	0.0673	41.4	2.78	2345	SETD5	SET domain containi…
64651	0.00	0.00	0.1041	45.6	4.74	1109	CSRNP1	cysteine and serine ri…
10181	0.00	0.00	0.0723	59.1	4.27	4580	RBM5	RNA binding motif pr…

Figure 4: List of Differentially expressed genes obtained after analysis of the chosen dataset on Geo2r for keratinocytes.

Top differentially expressed genes ⍰

Download full table   Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
13236	1.23e-25	2.85e-30	11.44	56.47	8.38	HIST1H4A	histone cluster 1, H4a
28834	1.04e-21	6.12e-26	10.54	46.98	7.72	CXCL1	C-X-C motif chemokine li…
7644	1.04e-21	7.21e-26	10.52	46.82	7.7	CXCL8	C-X-C motif chemokine li…
9353	1.57e-16	1.45e-20	9.3	35.23	6.81	HIST1H4E	histone cluster 1, H4e
3477	1.60e-14	1.84e-18	8.77	30.64	6.42	HIST1H4H	histone cluster 1, H4h
3808	9.04e-13	1.25e-16	8.28	26.65	6.06	EGR1	early growth response 1
37581	2.19e-12	3.53e-16	8.16	25.67	5.97	HIST1H4F	histone cluster 1, H4f
18611	4.31e-12	8.75e-16	8.05	24.81	5.89	HIST1H2AM	histone cluster 1, H2am
32961	4.31e-12	8.95e-16	8.04	24.79	5.89	HIST1H4I	histone cluster 1, H4i
36101	4.35e-12	1.06e-15	8.02	24.63	5.87	HIST1H4J	histone cluster 1, H4j
21387	4.35e-12	1.10e-15	8.02	24.59	5.87	HIST4H4	histone cluster 4, H4
40674	5.51e-12	1.55e-15	-7.98	24.27	-5.84	FRMD4A	FERM domain containin…
44252	5.51e-12	1.65e-15	7.97	24.21	5.83	HIST1H4K	histone cluster 1, H4k
40759	9.46e-12	3.05e-15	-7.89	23.63	-5.78		
30926	1.15e-11	4.17e-15	-7.85	23.34	-5.75	MAGI2-AS3	MAGI2 antisense RNA 3
32866	1.15e-11	4.23e-15	-7.85	23.32	-5.75	ADIPOQ	adiponectin, C1Q and co…
37739	5.13e-11	2.01e-14	-7.65	21.85	-5.6	CD84	CD84 molecule
16539	5.43e-11	2.25e-14	7.64	21.75	5.59	HIST2H4B	histone cluster 2, H4b
29397	8.67e-11	3.80e-14	-7.57	21.25	-5.54		
7750	1.14e-10	5.27e-14	7.53	20.94	5.51	CXCL3	C-X-C motif chemokine li…
30927	1.92e-10	9.29e-14	-7.45	20.41	-5.46		
38808	4.05e-10	2.06e-13	7.35	19.66	5.38	HIST1H3B	histone cluster 1, H3b

Figure 5: List of Differentially expressed genes obtained after analysis of the chosen dataset on Geo2r for the HeLa cell lines.

	A	B	C	D	E	F	G	H	I	J
1		Upregulated					Downregulated			
2		GPR50					KRT74			
3		ABCB1					KRT73			
4		H4C3					KRT73-AS1			
5		MIR6843					KRT76			
6		LOC105374995					KRT72			
7		LOC105373359					VSNL1			
8		FTCD					KRT4			
9		MIR7-3HG					CLEC2A			
10		MEG9					LOC105378927			
11		RNU2-1					LOC105372636			
12		LOC105371159					LAMB4			
13		FTCD-AS1					KRT3			
14		H2AC4					IFI44L			
15		LOC107985677					ACP3			
16		LOC105374117					LOC105376901			
17		LOC107984862					CD248			
18		HSPA6					LOC105370256			
19		GPR50-AS1					GRHL2-DT			
20		TRE-CTC1-3					LOC105376980			

Figure 6: Upregulated and downregulated genes of keratinocytes.

(https://onedrive.live.com/edit?id=31901C11ECE1492B!s198039d684d14b1bb99c452a51fdff71&resid=31901C11ECE1492B!s198039d684d14b1bb99c452a51fdff71&cid=31901c11ece1492b&ithint=file%2Cxlsx&redeem=aHR0cHM6Ly8xZHJ2Lm1zL3gvYy8zMTkwMWMxMWVjZTE0OTJiL0VkWTVnQm5SaEJ0THVaEZLbEg5XzNFQnkydk14cHdpVHJRVGdMcmhpaGGNOUGc_ZT0zOTBHVEY&migratedtospo=true&wdo=2)
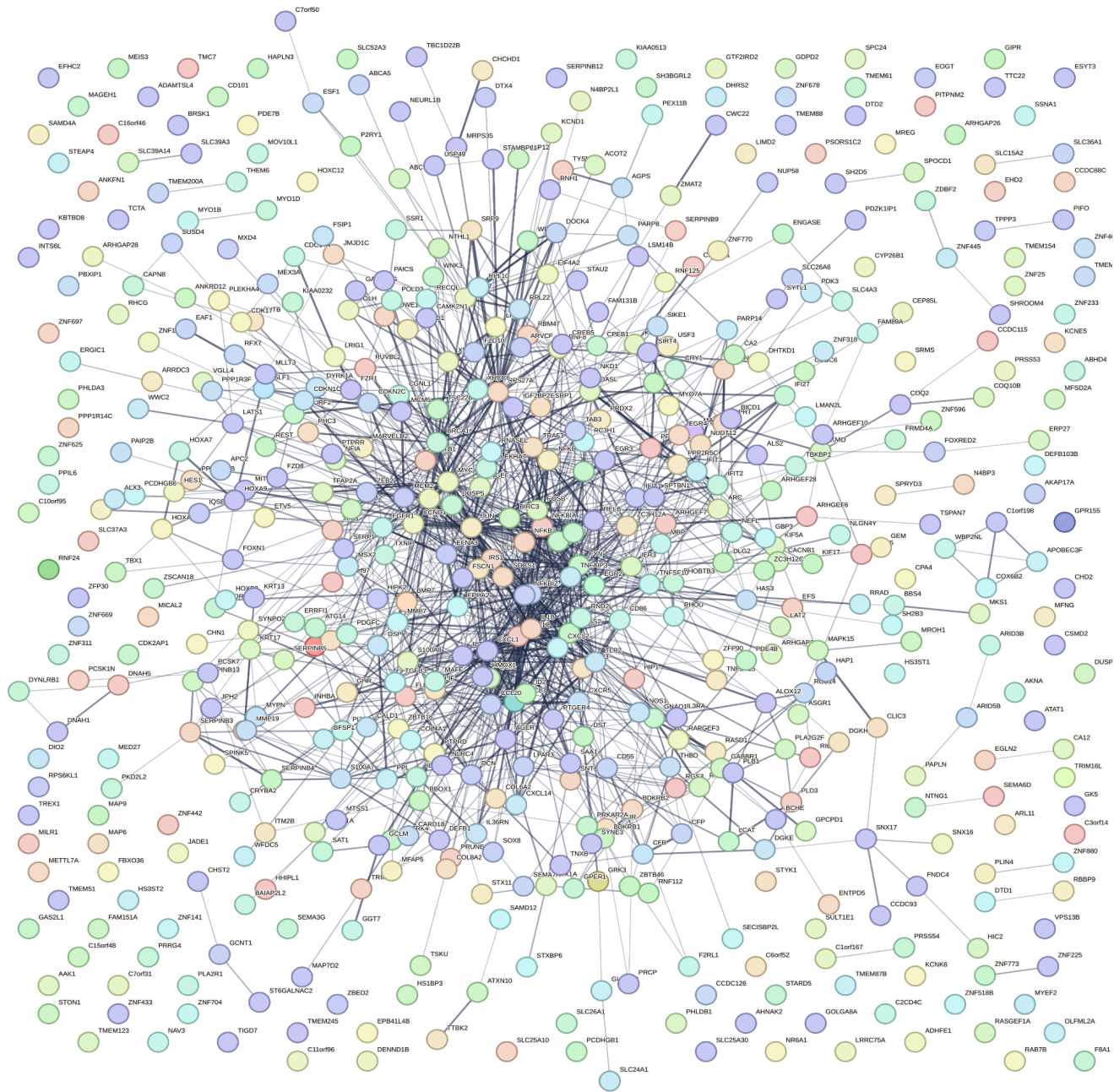
Figure 7: Protein-Protein Interaction (PPI) Network of Common Differentially Expressed Genes.
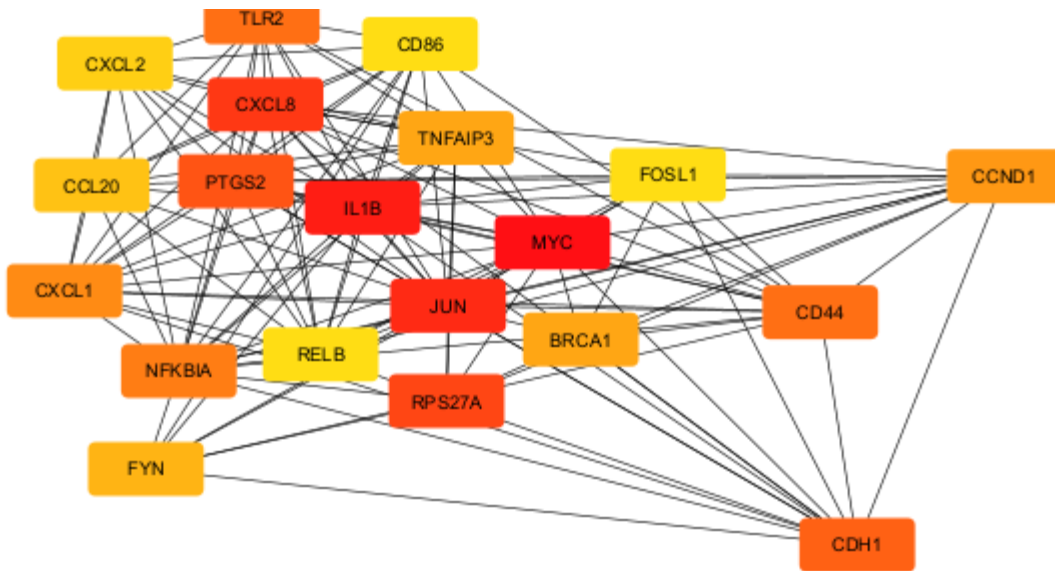
Figure 8: Top 20 Hub Genes Identified from the PPI Network Using CytoHubba.
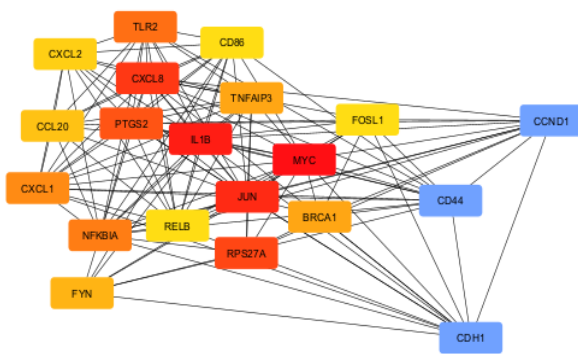
fig. 9a
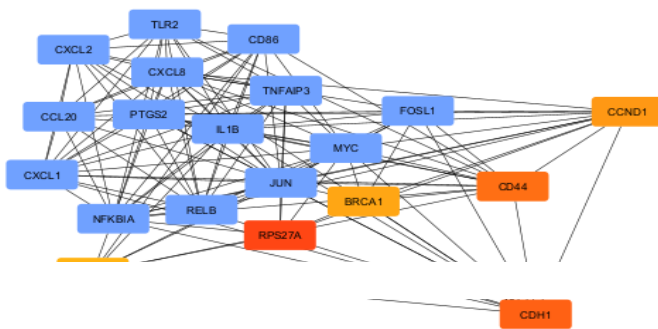
fig. 9b

Figure 9(a & b): Top Two Modules Identified Using MCODE from the PPI Network.

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☑	KEGG_PATHWAY	IL-17 signaling pathway	RT	▬▬▬	10	50.0	1.1E-13	1.5E-11
☑	KEGG_PATHWAY	Kaposi sarcoma-associated herpesvirus infection	RT	▬▬▬	10	50.0	8.2E-11	5.4E-9
☑	KEGG_PATHWAY	Rheumatoid arthritis	RT	▬▬	8	40.0	5.6E-10	2.4E-8
☑	KEGG_PATHWAY	Legionellosis	RT	▬▬	7	35.0	1.2E-9	4.1E-8
☑	KEGG_PATHWAY	TNF signaling pathway	RT	▬▬	8	40.0	3.0E-9	7.8E-8
☑	KEGG_PATHWAY	Coronavirus disease - COVID-19	RT	▬▬	9	45.0	1.4E-8	3.2E-7
☑	KEGG_PATHWAY	NF-kappa B signaling pathway	RT	▬▬	7	35.0	5.8E-8	1.1E-6
☑	KEGG_PATHWAY	Lipid and atherosclerosis	RT	▬▬	8	40.0	1.9E-7	3.1E-6
☑	KEGG_PATHWAY	Human T-cell leukemia virus 1 infection	RT	▬▬	8	40.0	2.3E-7	3.4E-6

Figure 10: Pathway Enrichment of Top Hub Genes Using KEGG Analysis.
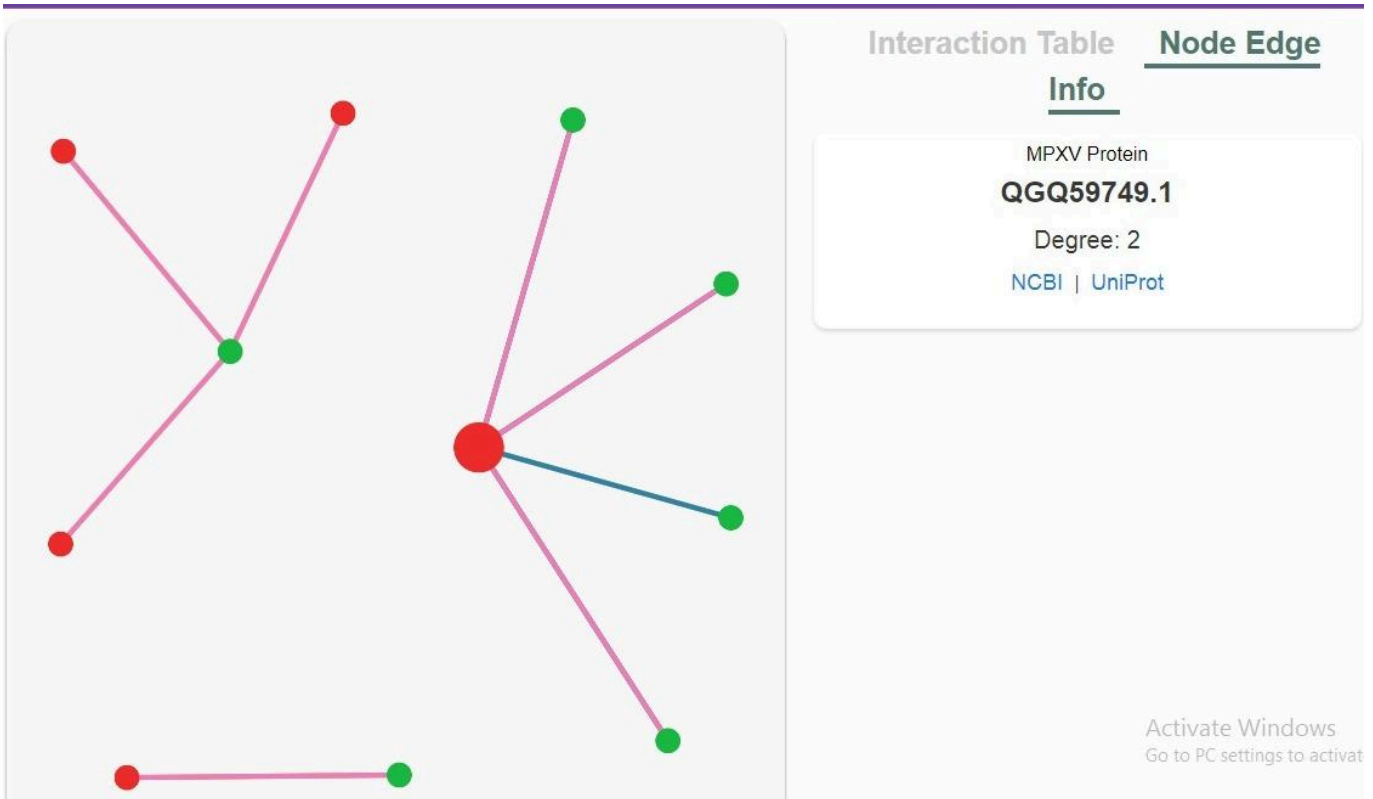
Figure 11: Viral-Human Protein Interaction between MPXV Strain MN648051 and 19 Hub Genes.

Figure 12: Showing one viral protein - QGQ59880.1 being a hypothetical protein.
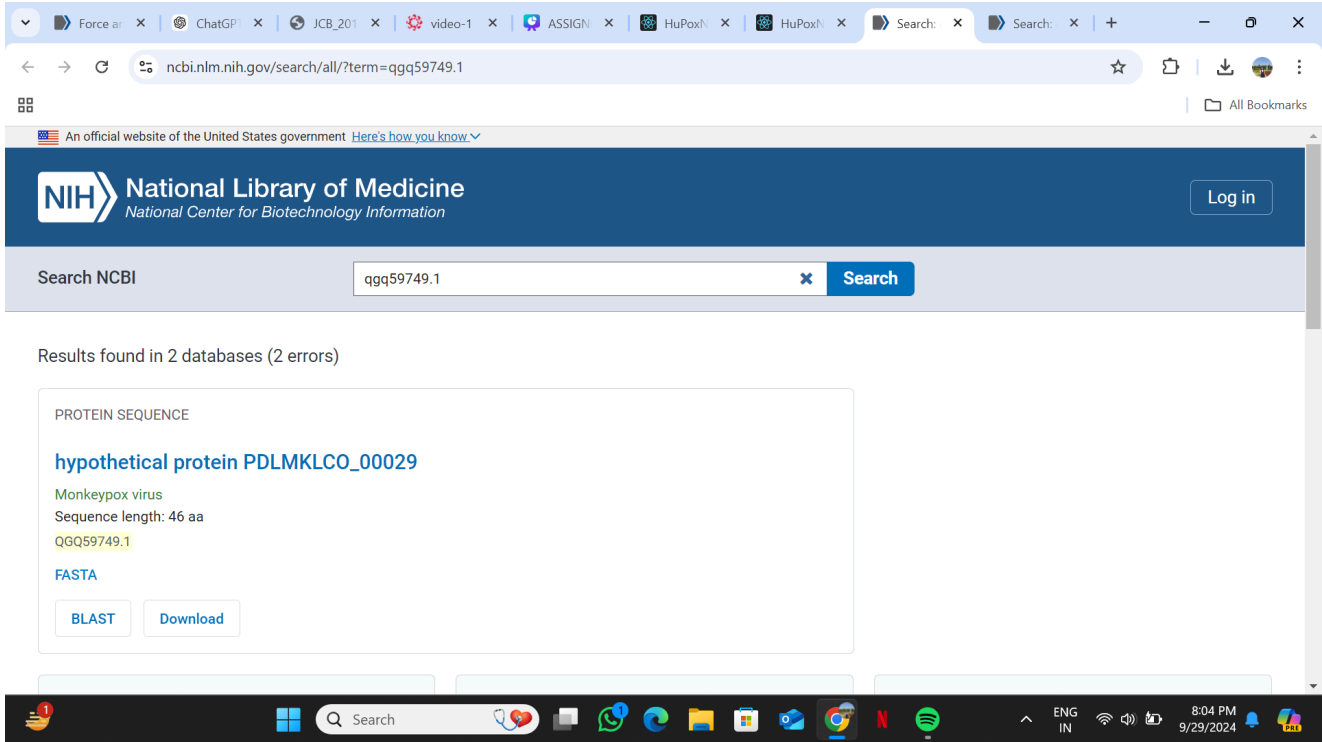
Figure 13: Showing one viral protein QGQ59749.1 being a hypothetical protein

Figure 14: Metal Ion Binding Sites in Hypothetical Protein PDLMKLCO_00160.
(Binding sites for metal ions (Mg²⁺, Co²⁺) in PDLMKLCO_0016)

Mg+2 and Co+2 had the highest binding scores. Henceforth, the top two scores of the respective metal ions were considered and docked.

Co+2(Highest binding score)

Co²⁺ binding sites prediction results for "Job at 2024-09-29 18:53:59"

No.	Binding Residues	Template	Score
1	52H , 66S , 68D	2f7vA0	3.379
2	40E , 129D	2f7vA0	3.300
3	26E , 108H	5ch9B0	3.292
4	63L , 166C	1u8rD1	2.455
5	52H , 56A , 68D	2prqA1	2.451

Co2+

Co^{2+} binding sites prediction results for "Job at 2024-09-29 18:53:59"

Help For Safari User +
Help For Chrome User +

Co^{2+} ▼
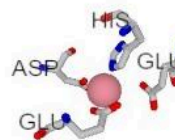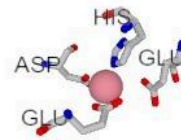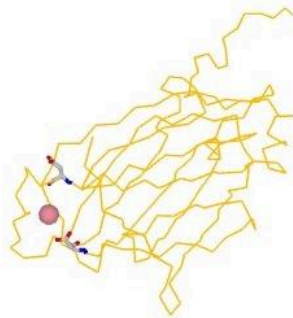
No.	Binding Residues	Template	Score	Show / DL
1	52H , 66S , 68D	2f7vA0	3.379	👁 / ⬇
2	40E , 129D	2f7vA0	3.300	👁 / ⬇
3	26E , 108H	5ch9B0	3.292	👁 / ⬇
4	63L , 166C	1u8rD1	2.455	👁 / ⬇
5	52H , 56A , 68D	2prqA1	2.451	👁 / ⬇
6	68D , 121D	3s8kA0	2.395	👁 / ⬇

Mg+2(Highest Score)

Mg²⁺ binding sites prediction results for "Job at 2024-09-29 18:33:46"

Help For Safari User +
Help For Chrome User +

Mg²⁺ ▾

| ChainA | | Prediction | Docking |

1tilC0

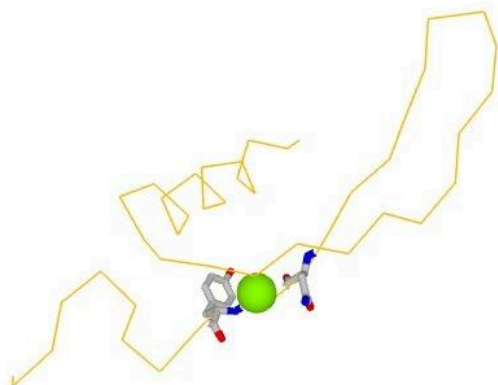| No.	Binding Residues	Template	Score	Show / DL
1	64N , 127E	1tilC0	4.914	👁 / ⬇
2	64N , 127E	3dtyE0	3.230	👁 / ⬇
3	64N , 127E	4gt8A0	3.221	👁 / ⬇
4	57S , 58V , 59S , 64N	2r72A2	2.726	👁 / ⬇
5	64N , 127E	5epvA0	2.614	👁 / ⬇
6	113K , 116E , 117E	6d1vB0	2.357	👁 / ⬇

Mg+2

Mg²⁺ binding sites prediction results for "Job at 2024-09-29 18:33:46"

Help For Safari User +

Help For Chrome User +

Mg²⁺ ▾

ChainA

| Prediction | Docking |

3dtyE0

No.	Binding Residues	Template	Score	Show / DL
1	64N , 127E	1tilC0	4.914	👁 / 📥
2	64N , 127E	3dtyE0	3.230	👁 / 📥
3	64N , 127E	4gt8A0	3.221	👁 / 📥
4	57S , 58V , 59S , 64N	2r72A2	2.726	👁 / 📥
5	64N , 127E	5epvA0	2.614	👁 / 📥
6	113K , 116E , 117E	6d1vB0	2.357	👁 / 📥

Figure 15: Metal Ion Binding Sites in Hypothetical Protein PDLMKLCO_00029.
(Binding sites for Mg²⁺ in PDLMKLCO_00029.)

For hypothetical protein PDLMKLCO_00029

Mg+2 had the highest binding score. Henceforth, the top two scores of the respective metal ions were considered and docked.

Mg+2 (Highest score)

No.	Binding Residues	Template	Score	Show / DL
1	7V , 11D	2c9eA0	2.069	👁 / 📥
2	35N , 37Y	3wewA0	1.625	👁 / 📥
3	7V , 11D	2fq1A0	1.393	👁 / 📥
4	3M , 4D	2xwbA0	1.214	👁 / 📥

Mg+2

Mg²⁺ binding sites prediction results for "Job at 2024-09-29 19:09:26"

Help For Safari User +
Help For Chrome User +

Mg²⁺ ▾

ChainA					

Prediction Docking

3wewA0

ASN

TYR

No.	Binding Residues	Template	Score	Show / DL	
1	7V , 11D	2c9eA0	2.069	👁 / ⬇	
2	35N , 37Y	3wewA0	1.625	👁 / ⬇	
3	7V , 11D	2fq1A0	1.393	👁 / ⬇	
4	3M , 4D	2xwbA0	1.214	👁 / ⬇	

Figure 16: CDD Results of Hypothetical protein PDLMKLCO_00029

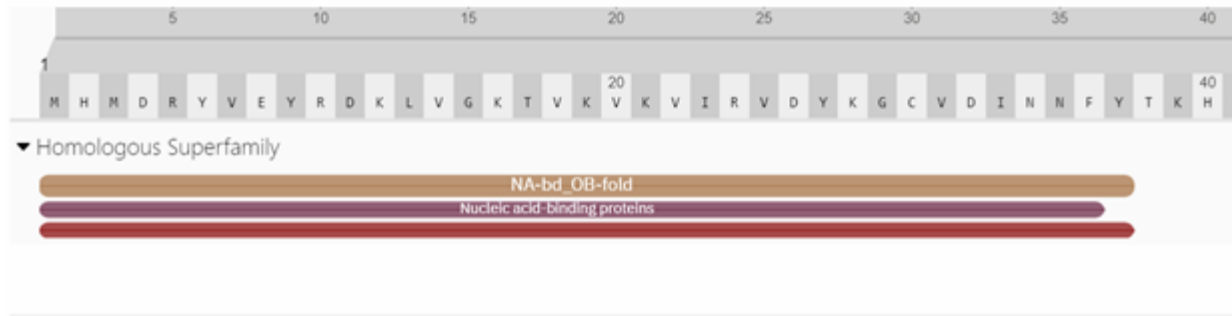## Protein family membership

None predicted

Figure 17: INTERPRO Result of Hypothetical protein PDLMKLCO_00029
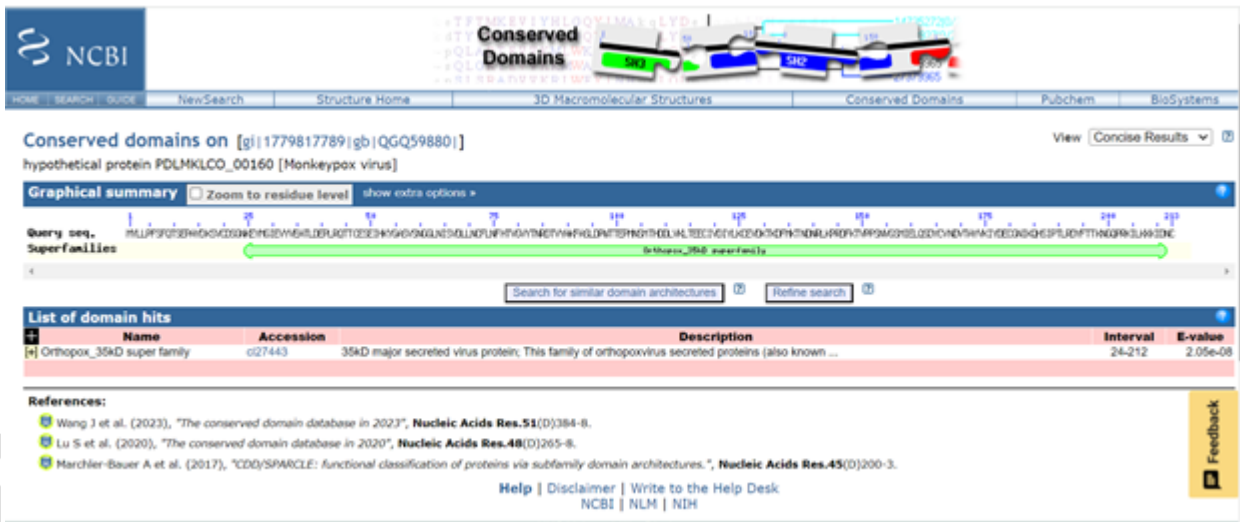

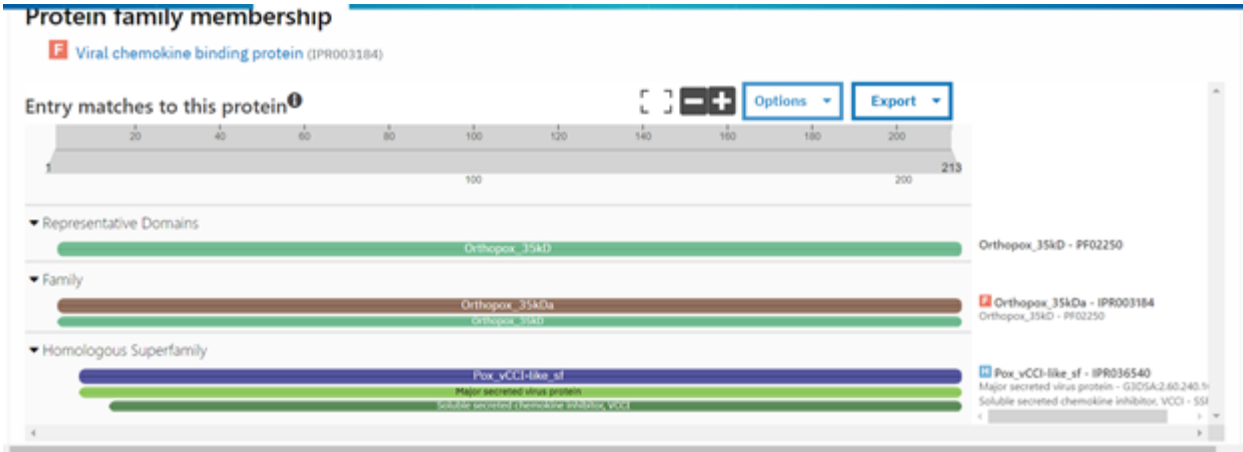Figure 18: CDD Result of Hypothetical Protein PDLMKLCO_000160

Figure 19: INTERPRO Result of Hypothetical Protein PDLMKLCO_000160

Discussion:

This study offers relevant information on the molecular dynamics of MPXV through gene expression, protein-protein interactions, and hypothetical protein structure predictions. 594 commonly overrepresented proteins that play important roles in replication and host response to the viral infection have been determined using the differential gene expression method in infected and normal cell lines. DEG analysis of genes which are significantly altered during the viral infection has pinpointed those genes that may be potential targets for further therapeutic research.

The PPI network was constructed to understand the molecular mechanisms by which monkeypox virus engages host human proteins. The analysis results of 20, including CCL20 and JUN, reveal potential drug targets in those interactions. Direct interaction between the viral and host proteins further demonstrates how MPXV proteins such as MPXVgp029 drive cellular functions, at least host immune responses, awry.

In protein analysis, two hypothetical proteins selected for functional and structural annotation were chosen: PDLMKLCO_00160 and PDLMKLCO_00029. Both were especially metal-binding to Mg^{2+} and Co^{2+}. The relative cruciality of the metal sites- for example, that of the MIB2 tool-is for the structure and activity of the proteins. Therefore, such proteins may be involved in critical biological functions or even in viral pathogenesis. Further studies of these hypothetical proteins might reveal additional potential in the development of antiviral drugs and vaccines for the monkeypox virus.

This therefore raises the significance of knowing what hypothetical proteins do in viral replication, immune evasions, and virus-host protein interaction; and the same was determined in this study.

This study shall serve as a starting point for continued research on the functional roles these proteins assume and their potential as drug targets, which would make an attempt at reducing monkeypox infections.

REFERENCES

1. Cho CT, Wenner HA (1973) Monkeypox virus. Bacteriological reviews 37(1):1–8. https://doi.org/10.1128/br.37.1.1-18.1973 7. Pickup DJ (2015) Extracellular virions: the advance guard of poxvirus infections. PLoS Pathogens 11(7):e1004904. https://doi.org/10.1371/journ al.ppat.1004904

2. Matho MH, Schlossman A, Gilchuk IM, Miller G, Mikulski Z, Hupfer M, Wang J, Bitra A, Meng X, Xiang Y, Kaever T (2018) Structure–function characterization of three human antibodies targeting the vaccinia virus adhesion molecule D8. J Biol Chem. 293(1):390–401. https://doi.org/10.1074/jbc.M117.814541

3. Chiu WL, Lin CL, Yang MH, Tzou DLM, Chang W (2007) Vaccinia virus 4c (A26L) protein on intracellular mature virus binds to the extracellular cellular matrix laminin. J virol 81(5):2149–2157. https://doi.org/10.1128/JVI. 02302-06

4. Singh K, Gittis AG, Gitti RK, Ostazeski SA, Su HP, Garboczi DN (2016) The vaccinia virus H3 envelope protein, a major target of neutralizing antibodies, exhibits a glycosyltransferase fold and binds UDP-glucose. J Virol 90(10):5020–5030. https://doi.org/10.1128/JVI.02933-15

5. Schin AM, Diesterbeck US, Moss B (2021) Insights into the organization of the poxvirus multicomponent entry-fusion complex from proximity analyses in living infected cells. J Virol 95(16):e00852-e921. https://doi. org/10.1128/JVI.00852-21

6. Senkevich TG, Ojeda S, Townsley A, Nelson GE, Moss B (2005) Poxvirus multiprotein entry–fusion complex. Proc Nat Acad Sci 102(51):18572– 18577. https://doi.org/10.1073/pnas.0509239102

7. Brown E, Senkevich TG, Moss B (2006) Vaccinia virus F9 virion membrane protein is required for entry but not virus assembly, in contrast to the related L1 protein. J virol 80(19):9455–9464. https://doi.org/10.1128/JVI. 01149-06

8. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S(2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database(Oxford).https://doi.org/10.1093/database/baaa062

9. Shchelkunov SN, Totmenin AV, Babkin IV, Safronov PF, Ryazankina OI, Petrov NA, Gutorov VV, Uvarova EA, Mikheev MV, Sisler JR, Esposito JJ (2001) Human monkeypox and smallpox viruses: genomic comparison. FEBS letters 509(1):66–70. https://doi.org/10.1016/S0014-5793(01)03144-1

10. Mahram A, Herbordt MC (2010) Fast and accurate NCBI BLASTP: acceleration with multiphase FPGA-based prefltering. InProceedings of the 24th ACM International Conference on Supercomputing, pp 73–82. https:// doi.org/10.1145/1810085.1810099

11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–10. https://doi.org/10.1016/ S0022-2836(05)80360-2

12. Yu, C.-S., Cheng, C.-W., Su, W.-C., Chang, K.-C., Huang, S.-W., Hwang, J.-K., et al. (2014). CELLO2GO: a web server for protein subcellular localization prediction with functional gene ontology annotation. PLoS ONE 9:e99368. doi: 10.1371/journal.pone.0099368

13. Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613. doi: 10.1093/nar/gky1131

14. Damaso, C. R. (2018). Revisiting the concepts of the poxvirus life cycle: A critical review. *Viruses*, 10(9), 502. https://doi.org/10.3390/v10090502
15. Dosztányi, Z., & Tompa, P. (2018). Viral proteins with high disorder extent are associated with different functions in the host cell. *Journal of Virology*, 92(16), e00865-18. https://doi.org/10.1128/JVI.00865-18
16. Galperin, M. Y., & Koonin, E. V. (2010). From complete genome sequence to "complete" understanding? *Trends in Biotechnology*, 28(8), 398-406. https://doi.org/10.1016/j.tibtech.2010.04.003
17. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, 31(13), 3784–3788. https://doi.org/10.1093/nar/gkg563
18. Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Research*, 41(D1), D387–D395. https://doi.org/10.1093/nar/gks1234
19. Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031
20. Kim, Y., & Skalka, A. M. (2021). Viral Proteins: Multifunctional Tools for Viral Infection and Manipulation of Host Physiology. *Frontiers in Microbiology*, 12, 660462. https://doi.org/10.3389/fmicb.2021.660462
21. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acids Research*, 33(D1), D501–D504. https://doi.org/10.1093/nar/gki025
22. Schaffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., & Koonin, E. V. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14), 2994–3005. https://doi.org/10.1093/nar/29.14.2994
23. UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100