Peiyang Liu
Serena Theobald
Sicheng Yi

DS596 Project Monkeypox Virus

Abstract

The Monkeypox virus, once primarily confined to African regions, has demonstrated a concerning global spread, with significant reported outbreaks in urban centers like Guangzhou and Shenzhen, China, in 2023. This study analyzes the evolution of 10 Monkeypox virus sequences from China, employing phylogenetic analysis, SNP extraction, and sequence similarity analysis through Basic Local Alignment Search Tool (BLAST) to identify genetic variations that may enhance human transmissibility or support adaptations to densely populated urban setting (Altschul et al., 1990).

Phylogenetic analysis using distance-based methods like UPGMA and Neighbor-Joining reveals distinct clustering patterns. Consistent clustering of certain sequences suggests recent shared ancestry and conserved traits critical for viral functions, while longer branch lengths in others indicate accelerated evolutionary rates driven by selective pressures or environmental adaptations. These findings emphasize the balance between conserved traits and localized adaptations, offering a novel perspective on how the virus evolves in response to environmental challenges.

SNP extraction revealed mutations at key genomic positions, categorized as shared, partially shared, and private mutations, providing insight into viral evolution. Conserved mutations were identified across sequences, suggesting selective pressures driving common adaptations, while private mutations indicated localized or recent evolutionary changes. Notably, APOBEC3-like mutation contexts (e.g., `TC>TT,GA>AA` transitions) highlight host-driven editing mechanisms, underscoring the interplay between host defenses and viral evolution. In addition, the corresponding proteins with these mutations potentially indicate adaptive strategies, based on their functions, that may enhance immune evasion, replication, and transmission efficiency.

This research provides significant insights into the evolutionary mechanisms of the Monkeypox virus, identifying key mutations influencing transmission dynamics and adaptation. By combining SNP and phylogenetic analyses, the findings offer valuable guidance for public health strategies to monitor and mitigate future outbreaks in densely populated regions (Yu et al., 2023).

Introduction

The Monkeypox virus has shown a concerning shift toward global dissemination in recent years. The outbreak of 2022 highlighted its increasing potential for human-to-human transmission, drawing significant attention to viral mutations contributing to this spread. In 2023, China reported its first major outbreak in Guangdong Province, particularly in urban centers such as Guangzhou and Shenzhen. This shift from travel-associated outbreaks to localized transmission highlights the need to understand how viral mutations contribute to its spread and adaptation in densely populated settings.

Our paper seeks to address the following question: How has the Monkeypox virus evolved across the provided 10 FASTA genetic sequences, and what genetic variations might influence its pathogenicity, transmission, or resistance? By identifying mutations in these genomes and analyzing their role in adaptation, this research can equip public health authorities with the insights needed to monitor viral evolution and respond effectively to future outbreaks (Yu et al., 2023).

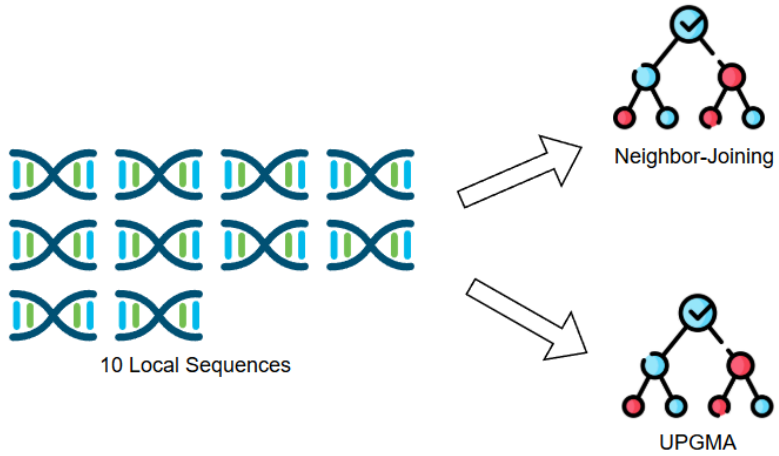
Current viral genomics research employs a diverse array of computational tools and methodologies to analyze sequence data and infer evolutionary relationships. Phylogenetic analysis, a cornerstone of this research, constructs evolutionary trees that reveal patterns of divergence and relatedness among sequences. Distance-based methods, such as Neighbor-Joining (NJ) and Unweighted Pair Group Method with Arithmetic Mean (UPGMA), estimate genetic distances, while character-based approaches, including Maximum Parsimony, Maximum Likelihood, and Bayesian Inference, provide probabilistic insights into sequence evolution. Tools like IQ-TREE, RAxML, and BEAST are widely used for these analyses, enabling robust and scalable tree construction (Omics Tutorials, n.d.).

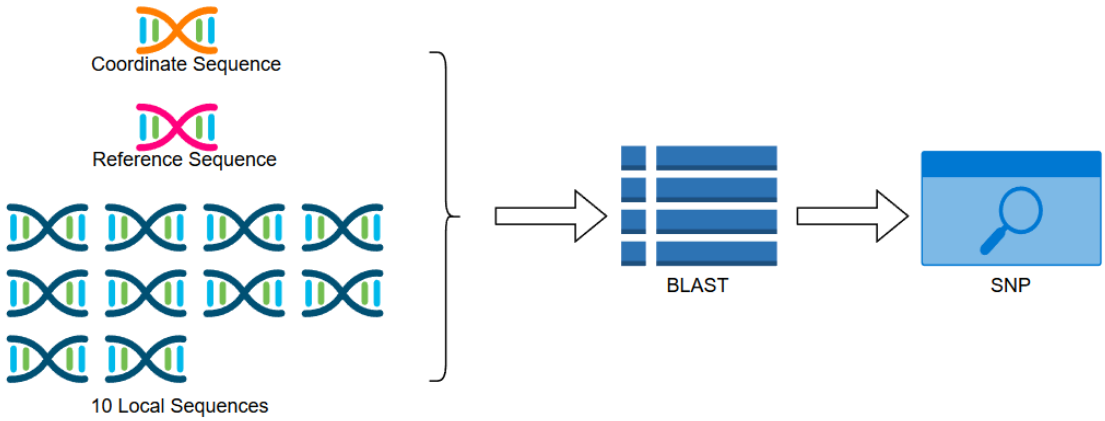
Complementing phylogenetics, sequence alignment tools like MAFFT (Katoh et al., 2002), Clustal Omega, and MEGA align nucleotide or protein sequences to prepare data for downstream analyses, such as SNP detection. Tools like SNP-sites, JVarKit, and PGDSpider identify genetic variations that influence viral pathogenicity or resistance. Platforms such as GATK and BCFtools further excel in variant calling and genomic variation analysis, enabling the identification of mutation patterns and functional insights. For functional annotation and comparative genomics, tools like iTol and InterProScan play a key role. InterProScan identifies protein families and regulatory elements using advanced algorithms, while iTol provides an interactive platform for generating, controlling, and annotating phylogenetic trees. These tools allow researchers to explore coding sequences and genomic features, offering critical insights into viral evolution and behavior (Katoh et al., 2017).

In this study, we employ NJ, UPGMA, and character-based methods to construct phylogenetic trees and uncover evolutionary relationships among the sequences. We aim to identify mutations through SNP analysis that may be linked to transmissibility, resistance, or urban adaptation. Finally, we use BLAST to evaluate genetic similarity and lineage relationships among the Monkeypox virus isolates. Together, these methodologies will enhance our understanding of the virus's evolution and inform strategies to mitigate future outbreaks (Yu et al., 2023).

Materials and Methods

Flowchart created with diagrams.net:





The data comprises 10 FASTA files provided by our paper, each approximately 194 to 195 KB in size, containing nucleotide sequences of Monkeypox virus isolates with around 195,000 base pairs each. A FASTA file is a text-based format structured in blocks, consisting of two main components: the header line and the sequence data. The header line, starting with a > symbol, contains a sequence identifier and description. For example, a header in these files is: >C_AA038932.1 Monkeypox virus isolate MPXV/human/CHN/GDCDC_FS_M23108/2023, which provides metadata such as the isolate ID, geographic location (Guangzhou, China), and year of isolation. The sequence data follows the header, represented in multiple lines of nucleotide bases (A, T, C, G), typically formatted in readable chunks rather than as a continuous string, such as `TGTTGATAAGCTCTACG`... Occasionally, unknown bases are denoted by `N`, which likely represent ambiguous or missing nucleotide information caused by sequencing errors, alignment issues, or low coverage in specific regions (Yu et al., 2023).

The Neighbor-Joining (NJ) algorithm is a commonly used method in phylogenetics to construct evolutionary trees. It construct the trees by iteratively joining pairs of node that minimize the local tree length with the following steps:

1. Compute Neighbor-Joining Matrix D^* using $D^*_{i,j} = (n-2) \cdot d_{i,j} - TotalDistance_D(i) - TotalDistance_D(j)$ where $d_{i,j}$ is distance between i and j, $TotalDistance_D(i) = \sum_k d_{i,k}$ is the sum of all distances from i.
2. Find Neighbors by identifying the pair (i, j) corresponding to the smallest $D^*_{i,j}$.
3. Join i and j in the tree, compute branch lengths using distances from D to update the tree
4. Replace i and j with their common ancestor in a new distance matrix. Repeat until the tree is complete. If D is additive, the smallest $D^*_{i,j}$ guarantees valid neighbors.

UPGMA is an agglomerative hierarchical clustering method commonly used in bioinformatics to construct rooted trees, or dendrograms, from pairwise distance matrices. This method assumes that all lineages evolve at the same rate and iteratively combines the closest clusters to build the tree. At each step, the two nearest clusters are merged to form a higher-level cluster. The distance

between two clusters C1 and C2 is defined as the average distance between all pairs of elements *x* in C1 and *y* in C2, making it the mean distance between the elements of the clusters (Kaur et al., 2023). To execute UPGMA, it is important to implement these steps below:

1. Form a cluster for each present-day species, each containing a single leaf. The distance matrix contains pairwise distances Dij between clusters.
2. Identify the two clusters with the smallest average pairwise distance. Calculate the average distance between the two closest clusters C1 and C2 based on the formula: $D_{\text{avg}}(C_1, C_2) = \frac{\sum_{i \in C_1, j \in C_2} D_{ij}}{|C_1| \cdot |C_2|}$ where $|\text{C}|$ denotes the number of elements in C.
3. Merge C1 and C2 into a single cluster C
4. Form a new node for C and connect to C1 and C2 by an edge. Set the average of C as D_{avg}(C1,C2)/2
5. Update the distance matrix by computing the average distance between each pair of clusters.
6. Iterate until a single cluster contains all species. Continue merging the closest clusters and updating the distance matrix until all clusters are combined into a single tree.

Moving forward, BLAST is designed for local alignment of nucleotide sequences, identifying regions of similarity between a query sequence and a database of target sequences. Below is an outline of the mathematical foundation and workflow of BLAST:

1. Scoring function is to add up the score of matches and subtract the mismatch, insertion, and deletion gaps. S follows extreme value distribution EVD, not normal distribution. $S = \sum Match - \sum Mismatch - \sum Gap$
2. The input sequence is divided into smaller substrings called k-mers, with BLAST using the default parameter k=11 to identify closely matching k-mers between the query and target sequences as seeds for alignment. $P(Match) = \prod_{i=1}^{k} P(nucleotide_i)$
3. Using dynamic programming, BLAST extends the initial matches to calculate cumulative alignment scores. The extension process terminates when the score falls below a specified threshold, ensuring computational efficiency.
4. The significance of an alignment is measured using the E-value, which estimates the number of alignments expected to occur by chance. It is calculated using the Karlin-Altschul statistic. $E = K \cdot m \cdot n \cdot e^{-\lambda S}$

K represents the scale factor influencing the likelihood of finding high-scoring alignments in random sequences. A smaller K indicates a lower probability of random high scores, leading to more stringent E-values. Conversely, a larger K makes alignments appear less significant. λ is a scale factor ensuring that alignment scores S correspond to exponentially decaying probabilities. m is the length of the query sequence and n is the length of the target sequence. S represents the Scoring of the Alignment. A lower E-value indicates that the alignment is unlikely to occur by chance. $E < 0.01$ is significant. BLAST identifies meaningful matches by prioritizing alignments with high scores (S) and low E-values.

After the BLAST alignment, we focus on the mismatch in the alignment, especially the APOBEC3-like mutation (`TC>TT`, `GA>AA`). These mutations are linked to host-driven editing

mechanisms, and their positions are correlated with the types of proteins generated, including those involved in host immune regulation, surface-level interactions, and replication processes (Altschul et al., 1990).

Results

The figures below represent the phylogenetic trees we generated.

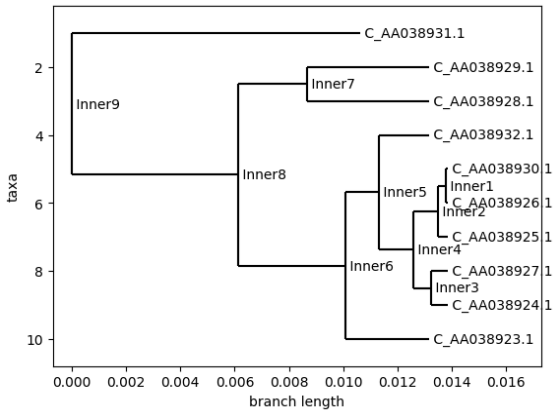


Figure 1. UPGMA Phylogenetic Tree

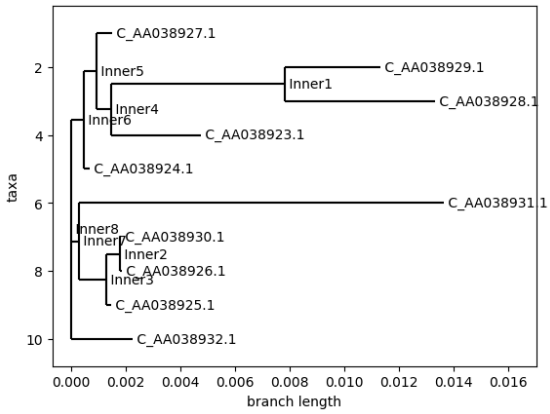


Figure 2. Neighbor-Joining Phylogenetic Tree

The dendrograms depict evolutionary relationships among sequences, with each leaf representing a unique sequence labeled by identifiers like `C_AA038924.1`. Branch lengths correspond to evolutionary distances; shorter branches indicate closer relationships, while longer branches suggest greater divergence. Internal nodes, such as "Inner1," represent shared ancestors, with node placement reflecting the relative timing of divergence. The depth of these nodes reflects the evolutionary distance from the root, with nodes positioned closer to the root representing older common ancestors. The position of these nodes provides insight into the timing of divergence events, showing which groups of sequences share more recent common ancestors. Clusters of closely connected sequences indicate groups with recent shared ancestry, helping to identify groups with similar genetic traits based on the proximity of their evolutionary relationships.

The result from this analysis is the clustering pattern of the sequences, which visually reflect their evolutionary relationships. Both UPGMA and NJ methods show similar overall clustering, although they vary slightly in structure. For the UPGMA tree, the clustering appears to be more symmetrical because the method pairs the closest sequences and joins them at the midpoint between the two. Because each split in the tree is positioned to maintain balanced branch lengths, we can infer that the rate of mutation or divergence is roughly equal around all lineages. Because UPGMA assumes equal evolutionary rates, it produces a tree where all paths from the root to any leaf are of equal length. This is why the UPGMA tree appears to be balanced and symmetrical. On the other hand, the Neighbor Joining (NJ) tree is more flexible, as it does not assume a constant rate of evolution. The sequences cluster based on their relative distances without the constraint of equal branch lengths from the root, meaning that some branches may be shorter or longer, indicating varying rates of evolution. For example, in the NJ tree, sequences like `C_AA038927.1` and `C_AA038924.1` have noticeably longer branch lengths in certain paths, suggesting they may have evolved at a different rate compared to sequences with shorter branches.

Sequences such as C_AA038926.1 and C_AA038925.1, as well as C_AA038928.1 and C_AA038929.1, consistently appear as neighboring nodes in both UPGMA and Neighbor-Joining trees. This consistent clustering suggests a potential common recent ancestor or shared evolutionary pressures. Such findings could uncover conserved genetic traits or mutations essential for understanding the adaptations and functional dynamics of the monkeypox virus.

Figure 3. Phylogenetic Tree from the paper (Yu et al., 2023)

Our results were compared to the phylogenetic tree from the referenced paper. For instance, C_AA038928.1 and C_AA038929.1 appear as closely related nodes in both our UPGMA and Neighbor Joining (NJ) trees, corresponding to GDCDC_SZ_M23024 and GDCDC_SZ_M23026, respectively, in the paper's graph. They are illustrated as the last 2 red dots in the graph from the paper above. This agreement suggests that both methods accurately captured the evolutionary proximity of these sequences. However, discrepancies were also noted. In the paper, the top two red dots, representing GDCDC_FS_M23108 and GDCDC_GZ_M23008, indicate that these sequences are closely related. These correspond to C_AA038932.1 and C_AA038923.1, respectively. In contrast, our UPGMA and NJ trees place these two nodes far apart, indicating a significant divergence from the paper's results.

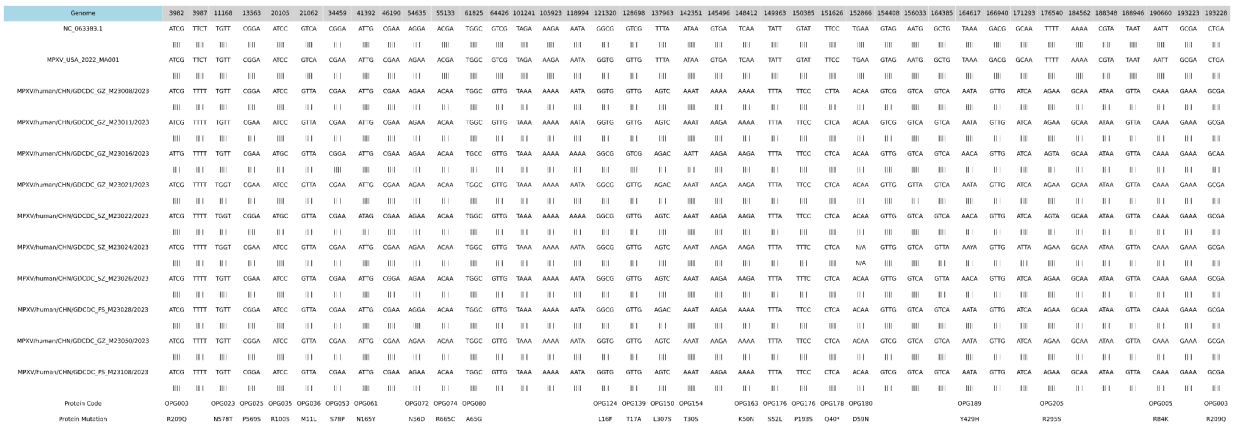


Figure 4. SNP Extraction with BLAST local alignment

To identify single nucleotide polymorphisms (SNPs) in the monkeypox virus (MPXV), we aligned the partial genomes of 10 MPXV samples from China (2023) to the complete reference genome of MPXV from the United States (2022). BLAST local alignment proved more effective than global alignment for pinpointing the exact locations of mutations. Among these, APOBEC3-like mutations or `TC>TT,GA>AA` nucleotide substitutions, play a significant role in genome editing and protein mutations, potentially facilitating the rapid transmission of the virus. (Yu et al., 2023). We extracted SNPs at 39 different locations across the complete genome to compare and contrast the nucleotide substitutions in the 10 Chinese samples with the reference USA genome. In addition to BLAST, we utilized GenBank and UniProt databases to map these genomic positions to specific OPG protein mutations and evaluate their biological functions. The full results of the BLAST alignment and SNP extraction are presented in Figure 4.

Genome	3987	21062	64426	149963	150385	151626	154408	156033	166940	171293	188946
NC_063383.1	TTCT	GTCA	GTCG	TATT	GTAT	TTCC	GTAG	AATG	GACG	GCAA	TAAT
	||||	||||	||||	||||	||||	||||	||||	||||	||||	||||	||||
MPXV_USA_2022_MA001	TTCT	GTCA	GTCG	TATT	GTAT	TTCC	GTAG	AATG	GACG	GCAA	TAAT
	||||	||||	||||	||||	||||	||||	||||	||||	||||	||||	||||
MPXV/human/CHN/GDCDC_GZ_M23008/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTTA	GTCG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	|| |	|| |	||||	|| |	||||	|| |
MPXV/human/CHN/GDCDC_GZ_M23011/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTCA	GTCG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	||||	|| |	||||	|| |	||||	|| |
MPXV/human/CHN/GDCDC_GZ_M23016/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTCA	GTTG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	||||	||||	||||	|| |	||||	|| |
MPXV/human/CHN/GDCDC_GZ_M23021/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTCA	GTTG	GTTA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	||||	||||	|| |	|| |	||||	|| |
MPXV/human/CHN/GDCDC_SZ_M23022/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTCA	GTTG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	||||	||||	||||	|| |	||||	|| |
MPXV/human/CHN/GDCDC_SZ_M23024/2023	TTTT	GTTA	GTTG	TTTA	TTTC	CTCA	GTTG	GTCA	GTTG	ATTA	GTTA
	|| |	|| |	|| |	|| |	|| |	||||	||||	||||	|| |	|| |	|| |
MPXV/human/CHN/GDCDC_SZ_M23026/2023	TTTT	GTTA	GTTG	TTTA	TTTC	CTCA	GTTG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	|| |	||||	||||	||||	|| |	||||	|| |
MPXV/human/CHN/GDCDC_FS_M23028/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTCA	GTTG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	||||	||||	||||	|| |	||||	|| |
MPXV/human/CHN/GDCDC_GZ_M23050/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTCA	GTCG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	||||	|| |	||||	|| |	||||	|| |
MPXV/human/CHN/GDCDC_FS_M23108/2023	TTTT	GTTA	GTTG	TTTA	TTCC	CTCA	GTCG	GTCA	GTTG	ATCA	GTTA
	|| |	|| |	|| |	|| |	||||	||||	|| |	||||	|| |	||||	|| |
Protein Code		OPG036		OPG176	OPG176	OPG178					
Protein Mutation		M11L		S52L	P193S	Q40*					

Figure 5. SNP Extraction APOBEC3-like mutations `TC>TT`

Figure 5 highlights APOBEC3-like mutations characterized by `TC>TT` nucleotide substitutions. At specific positions, such as 3987, 21062, and 149963, all 10 genomes exhibit a fully shared `TC>TT` mutation. In contrast, positions like 150385, 154408, and 171923 display partially shared `TC>TT` mutations, present in only a subset of the genomes. As a result, all 10 monkeypox genomes share the protein mutations OPG036-M11 and OPG176-S52L, whereas mutations such as OPG176-P193S and OPG178-Q40 are found in only a few genomes. Functionally, the OPG036 and OPG176 proteins, which are shared across all 10 genomes, are involved in host immune regulation. Meanwhile, OPG178, associated with viral transcription regulation, is not universally shared among the genomes, reflecting potential differences in the functional adaptations of the virus.

Genome	13563	55133	101241	105923	145496	152866	184562	188348	190660	193223	193228
NC_063383.1	CGGA	ACGA	TAGA	AAGA	GTGA	TGAA	AAAA	CGTA	AATT	GCGA	CTGA
MPXV_USA_2022_MA001	CGGA	ACGA	TAGA	AAGA	GTGA	TGAA	AAAA	CGTA	AATT	GCGA	CTGA
MPXV/human/CHN/GDCDC_GZ_M23008/2023	CGGA	ACAA	TAAA	AAAA	AAAA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_GZ_M23011/2023	CGAA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_GZ_M23016/2023	CGAA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCAA
MPXV/human/CHN/GDCDC_GZ_M23021/2023	CGAA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_SZ_M23022/2023	CGGA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_SZ_M23024/2023	CGAA	ACAA	TAAA	AAAA	AAGA	N/A	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_SZ_M23026/2023	CGAA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_FS_M23028/2023	CGGA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_GZ_M23050/2023	CGGA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
MPXV/human/CHN/GDCDC_FS_M23108/2023	CGGA	ACAA	TAAA	AAAA	AAGA	ACAA	GCAA	ATAA	CAAA	GAAA	GCGA
Protein Code	OPG025	OPG074				OPG180			OPG005		OPG003
Protein Mutation	P569S	R665C				D59N			R84K		R209Q

Figure 6. SNP Extraction APOBEC3-like mutations `GA>AA`

Figure 6 illustrates the APOBEC3-like mutations `GA>AA`. At positions 55133, 105923, 152866, and 190660, all 10 genomes exhibit fully shared `GA>AA` mutations, corresponding to the protein mutations OPG074, OPG124, OPG180, and OPG005, respectively. In contrast, position 13563 displays a partially shared `GA>AA` mutation, observed in only a subset of genomes, and is linked to the protein OPG025. Position 193228 represents a private mutation on protein OPG003, found exclusively in the genome of `MPXV/M23016`. Functionally, OPG124 and OPG180, which are fully shared across all 10 genomes, are responsible for viral transcription regulation. Meanwhile, the partially shared mutation at OPG025 and the private mutation at OPG003 are associated with host immune regulation, reflecting functional diversity among the genomes.

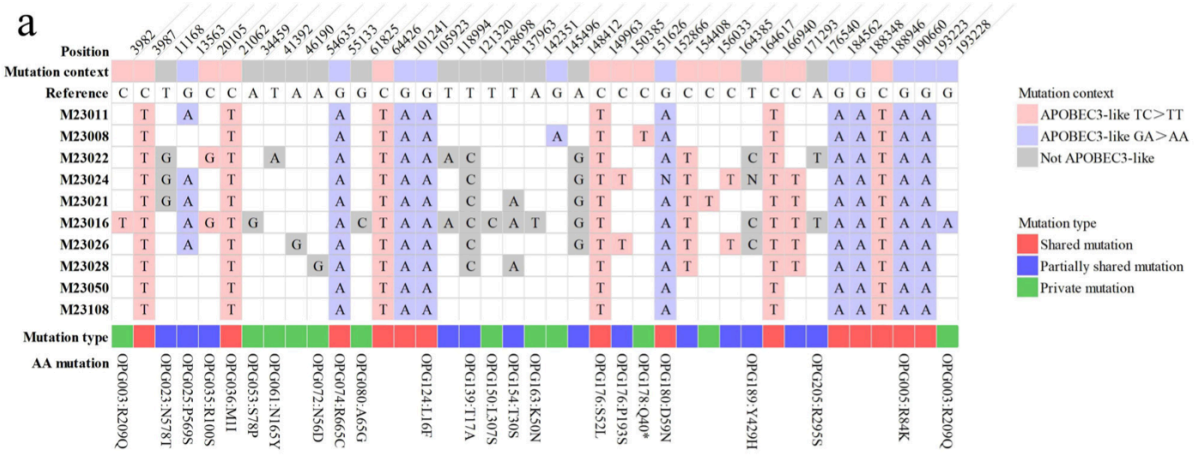


Figure 7. SNP extraction and protein mutations from the paper (Yu et al., 2023)

Figure 7 shows the alignment and SNP extractions from the paper. From the first 30+ positions our own BLAST and SNP extractions match exactly to the paper. We slightly improve the visualization by also showing the 4 bases around each position, so it's easier to see the `TC>TT` and `GA>AA` mutations. And by doing so, we found out that a few positions at the end disagree with the paper. For example, at the 2nd to last position 193223, the paper claims it is a `GA>AA` mutation. However, according to our own reference genome `MPXV_USA`, it clearly shows it's a `CG>AA` mutation instead. At position 190660, our alignment shows it is the `AT>AA` mutation, not `GA>AA`. We suspect it might be the difference in how the original authors did their sequence alignment. But given the first 30+ positions are perfectly matched, we can't rule out the possibility that either we or the original authors might've made a mistake on the alignment.

Discussion

To continue exploring our central research question, we decided to explore how mutations help the Monkeypox virus adapt to local conditions. By analyzing 10 local strains of Monkeypox, we identified mutations in proteins related to immune evasion, replication, and transmission, uncovering how the virus balances survival and adaptation.

To conduct a deeper analysis of the proteins where mutations occurred, we used data from UniProt and annotated the names and functions of these proteins based on previous studies. All annotated proteins are UniProtKB-reviewed and originate from monkeypox organisms. Our aim is to investigate their potential roles in the evolution of local monkeypox virus transmission and to understand how mutations influence viral adaptation and spread. A detailed list of protein annotations is available here. (UniProt Consortium, n.d.)

The mutations identified in these proteins across our 10 local monkeypox virus strains provide valuable insights into the evolutionary mechanisms driving local transmission. Key mutations in proteins such as OPG003 (R209Q) and OPG023 (N578T), both containing ankyrin repeats and PRANC regions, likely enhance the virus's ability to manipulate host immune responses, particularly by targeting NF-kappa-B pathways critical for antiviral defenses. Similarly, mutations in OPG035 (R100S) and OPG176 (P193S, S52L), both BCL2-like proteins, emphasize the virus's evolutionary focus on immune evasion by inhibiting pro-inflammatory signals and disrupting TLR4 signaling pathways. The R295S mutation in serine proteinase inhibitor 1 (OPG205) suggests a shift in the virus's strategy to inhibit apoptosis, allowing prolonged replication within host cells. These findings point to selective pressures on immune-modulating proteins, a hallmark of viral adaptation to host immune defenses during localized outbreaks.

Structural and replication-associated proteins also exhibit significant mutations, likely contributing to viral fitness and transmission efficiency. For example, the D59N mutation in DNA ligase (OPG180) and the Q40* mutation in thymidylate kinase (OPG178) indicates optimization of replication machinery for efficient DNA synthesis and repair under local host conditions. Changes in structural proteins, such as T30S in OPG154 and T17A in the virion membrane protein OPG139, which are critical for virion assembly and encapsidation, suggest adaptations enhancing viral assembly and release. Mutations in entry-related proteins, such as S78P in EFC-associated protein OPG053, reflect potential improvements in host cell penetration mechanisms.

However, we recognize the limitations of our methodologies as part of the broader context. Single nucleotide polymorphism (SNP) extraction and phylogenetic analyses focus primarily on single base-pair changes rather than multiple, leaving complex mutations and

structural variations underexplored. While UPGMA and NJ provide valuable evolutionary context, UPGMA assumes uniform evolutionary rates, which may not reflect viral evolution dynamics, and NJ, though more flexible, can produce clustering inconsistencies with divergent sequences. In contrast, the paper used Maximum Likelihood (ML) in IQ-Tree, which incorporates complex substitution models, accounts for rate heterogeneity, and optimizes tree topology, providing greater accuracy. For example, our UPGMA and NJ analyses could not replicate the paper's close grouping of `GDCDC_FS_M23108` and `GDCDC_GZ_M23008`, highlighting their constraints in capturing nuanced relationships. These limitations underscore the need for advanced approaches, such as integrating structural data or machine learning, to better capture the functional impacts of mutations.

Ultimately, this work ties together a comprehensive narrative of how the monkeypox virus balances immune evasion, replication optimization, and transmission efficiency to adapt to localized environments. The observed mutations provide compelling evidence of the virus's evolutionary strategies, answering our research question by highlighting key mechanisms that drive local adaptation and transmission. These findings lay the groundwork for future studies to further investigate the functional effects of these mutations, integrate additional data across regions and time periods, and develop predictive models to monitor and mitigate future outbreaks.

Conclusion

Our paper highlights the evolutionary adaptations of the Monkeypox virus, which has recently spread to densely populated areas like Guangzhou and Shenzhen in 2023. Phylogenetic analysis revealed patterns of genetic relatedness, with conserved sequences suggesting shared evolutionary traits and longer branches indicating selective pressures driving divergence. Furthermore, our SNP analysis identified conserved mutations linked to widespread pressures and localized private mutations, while APOBEC3-like signatures underscore the role of host immune responses in shaping the viral genome. BLAST analysis further confirmed lineage patterns and adaptation to environmental and host conditions.

To advance this work, it is essential to incorporate viral sequences from broader regions and time periods, as the current dataset of 10 FASTA files is limited to 2023 and focuses only on three areas in China (Shenzhen, Guangzhou, and Foshan). Broadening the geographical scope to include sequences from northern, southern, eastern, and western China, as well as capturing differences between urban and rural areas, would provide a more nuanced understanding of regional and environmental influences on mutation patterns. Such diversity in the data would help elucidate how mutations affect the virus's transmission dynamics in varying settings.

Future studies should characterize the functional effects of mutations, particularly those altering protein structure and function related to transmissibility and immune evasion. Protein folding simulations, structural modeling, and machine learning can connect mutations to phenotypic changes, enhancing predictions of viral evolution. By combining these approaches, researchers could establish a comprehensive framework for understanding the interplay between viral mutations and their phenotypic effects. This knowledge would be instrumental in informing targeted public health interventions to mitigate the spread of the Monkeypox virus and to effectively monitor, predict, and respond to future outbreaks (Yu et al., 2023).

Code Availability

The raw fasta data files, our Python notebooks, and the result images are made available at the following public Github repository: https://github.com/tigeryi1998/ds596-project

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

diagrams.net. (n.d.). Flowchart maker & online diagram software. Retrieved from https://app.diagrams.net/

Kaur, S., Sohal, H. S., & Cheema, R. S. (2013). Implementing UPGMA and NJ method for phylogenetic tree construction using hierarchical clustering. *International Journal of Computer Science and Technology*, 4(2), 303–304. Retrieved from https://www.ijcst.com/vol42/2/harwinder.pdf

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Katoh, K., Rozewicki, J., & Yamada, K. D. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, 20(4). https://doi.org/10.1093/bib/bbx108

Omics Tutorials. (n.d.). *Bioinformatics tools for sequence analysis*. Retrieved from https://omicstutorials.com/bioinformatics-tools-for-sequence-analysis/

UniProt Consortium. (n.d.). UniProtKB. Retrieved from https://www.uniprot.org/uniprotkb

Yu, J., Zhang, X., Liu, J., Xiang, L., Huang, S., Xie, X., Fang, L., Lin, Y., Zhang, M., Wang, L., He, J., Zhang, B., Di, B., Peng, B., Liang, J., Shen, C., Zhao, W., & Li, B. (2023). Phylogeny and molecular evolution of the first local monkeypox virus cluster in Guangdong Province, China. *Nature Communications, 14*(1), 8241. https://doi.org/10.1038/s41467-023-44092-3