# Genomic accordions may hold the key to Monkeypox Clade IIb's increased transmissibility.

Sara Monzon[#1], Sarai Varona[#1], Anabel Negredo[#2,3], Juan Angel Patiño-Galindo[4], Santiago Vidal-Freire[4], Angel Zaballos[5], Eva Orviz[6], Oskar Ayerdi[6], Ana Muñoz-García[6], Alberto Delgado-Iribarren[6], Vicente Estrada[3,6], Cristina Garcia[2], Francisca Molero[2], Patricia Sanchez[2,3], Montserrat Torres[2], Ana Vazquez[2,7], Juan-Carlos Galán[7,8], Ignacio Torres[9], Manuel Causse del Rio[10], Laura Merino[11], Marcos López[12], Alicia Galar[13], Laura Cardeñoso[14], Almudena Gutiérrez[15], Juan Camacho[2], Laura Herrero[2], Pilar Jimenez Sancho[5], Maria Luisa Navarro Rico[2], Isabel Jado[2], Jens Kuhn[21], Mariano Sanchez-Lockhart[22], Nicholas Di Paola[22], Jeffrey R. Kugelman[22], Elaina Giannetti[4], Susana Guerra[4,18,19], Adolfo García-Sastre[4,17,18,19,20], Gustavo Palacios[$&4,16], Maripaz Sanchez-Seco[&2,3], Isabel Cuesta[&1]

[1] Bioinformatics Unit (BU-ISCIII), Instituto de Salud Carlos III, Madrid, Spain

[2] National Center for Microbiology, Instituto de Salud Carlos III, Madrid, Spain

[3] Ciber Enfermedades Infecciosas (Ciberinfec), Instituto de Salud Carlos III, Madrid, Spain

[4] Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY

[5] Genomic Unit, Instituto de Salud Carlos III, Madrid, Spain

[6] Sandoval Center / Hospital Clínico San Carlos, IdISSC.

[7] Public Health CIBER (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain

[8] Servicio de Microbiología. Hospital Ramón y Cajal. CIBERESP. Madrid, Spain

[9] Microbiology Service, Hospital Clínico Universitario, Instituto de Investigación INCLIVA, Valencia, Spain.

[10] Microbiology Unit, University Hospital Reina Sofía, Cordoba, Spain. Maimonides Biomedical Research Institute of Cordoba (IMIBIC), Cordoba, Spain

[11] Infectious Disease Clinical Unit, Microbiología y Medicina Preventiva. Hospital Universitario Virgen del Rocío. Sevilla, Spain

[12] Microbiology and Parasitology Service. Hospital Universitario Puerta de Hierro Majadahonda. Madrid, Spain

[13] Infectious Diseases and Clinical Microbiology Service, Hospital General Universitario Gregorio Marañón. Madrid, Spain.

[14] Instituto de Investigación Sanitaria. Hospital Universitario de la Princesa. Madrid, Spain

[15] Clinical Microbiology and Parasitology Service, Hospital Universitario La Paz, Madrid, Spain

[16] Global Health Emerging Pathogens Institute, Icahn School of Medicine at Mount Sinai, New York, NY

[17] Department of Preventive Medicine, Public Health and Microbiology, Universidad Autónoma, E-28029 Madrid, Spain

[18] Department of Medicine, Division of Infectious Diseases, Icahn School of Medicine at Mount Sinai, New York, NY

[19] The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY

[20] Department of Pathology, Molecular and Cell-Based Medicine, Icahn School of Medicine at Mount Sinai, New York, NY

[21] NIH-IRF, Frederick, MD

[22] USAMRIID, Frederick, MD

[#] These authors equally contributed to this article.

[&] These senior authors equally contributed to this article.

[$] Corresponding author: Gustavo Palacios, gustavo.palacios@mssm.edu; 1 Gustave A. Levy, New York, NY, USA, 10029

## *Abstract*

The recent outbreak of Monkeypox displays novel transmission features. The circulating strain is a descendant of a lineage that had been circulating in Nigeria since 2017. The prognosis of monkeypox disease (MPX) with the circulating strain is generally good but the estimated primary reproduction number ($R_0$) among men who have sex with men (MSM) was above 1 suggesting efficient person-to-person transmission. Different mechanisms of viral entry and egress, as well as virus-coded host factors, are the main biological determinants of poxvirus transmissibility. OPXV evolution is driven by gene loss of virus-host interacting genes and selective pressure from host species using unique adaptive strategies at the gene and nucleotide level. In this context, we evaluated the effects of genomic instability in low-complexity-regions, areas that are often neglected during sequencing, during the early stage of the outbreak in Madrid, Spain. We uncovered significant variation in short-tandem repeat areas of the MPXV genome that could be associated with changes in transmissibility. Expression, translation, stability, or function of OPG153 (VACV A26L), OPG204 (VACV B16R) and OPG208 (VACV B19R) could be affected by the changes, in a manner that is consistent with proven “genomic accordion” strategies of OPXV evolution. Intriguingly, while the changes observed in OPG153 stand out as they are located inside a region under high selective pressure for transmission, in a gene that is clearly considered a “core” gene involved in attachment and egress; the changes in OPG208, a serine protease inhibitor-like protein that has been identified as an apoptosis inhibitor, host-range factor and virulence factor; and OPG204, a known inhibitor of the Type I interferon system shown to act as a decoy receptor, could also explain phenotypic changes. Further functional studies to complement this comparative genomic study are urgently needed.

## *Introduction*

Since May 2022, the United Kingdom followed by multiple European countries have reported an increase in the incidence of monkeypox virus (MPXV) associated disease. As of September 4, 52,996 cases had been reported in 102 countries/territories/areas in all six WHO Regions. On July 22, the World Health Organization declared this outbreak a Public Health Emergency of International Concern (PHEIC).

Monkeypox virus (MPXV) is a relatively large, double-stranded DNA virus belonging to the Orthopoxvirus genus (OPXV) in the family *Poxviridae*. First identified in 1958, MPXV has caused sporadic human outbreaks in Central and West Africa, with a mortality rate between 1% and 10% [1–4]. Viral genomes recovered from cases from the Congo Basin and West Africa clustered into two clades (recently renamed Clade I and II, [5]), the former being more virulent and transmissible. The 2022 current outbreak was quickly identified as related to Clade II [6,7], a descendant of a lineage of viruses that had been circulating in Nigeria since 2017 [8] and thus, labelled IIb (historical members of the group were renamed Clade IIa). Clade IIb had been previously imported into the UK [9], Israel [10] and Singapore [11,12]. The nearest relative of the 2022 outbreak is USA/UT-UPHL-82200022/2022 (Accession number: ON676708), a genome detected in a Nigerian traveler [13]. Although most cases are related to this group of sequences (named lineage B.1), a few 2022 cases reported in the USA, Malaysia, and India form a different lineage (named lineage A.2).

The prognosis of monkeypox disease (MPX) due to Clade IIb viruses in outbreaks outside Africa is generally good, with self-limited skin lesions without serious complications [14]. The clinical presentation is atypical, with rash lesions localized to the genital, perineal/perianal, or peri-oral area that often does not spread further and appears before the development of lymphadenopathy, fever, malaise, and pain associated with lesions. Nonetheless, the outbreak has been described as primarily affecting men who have sex with men (MSM), a group where the estimated primary reproduction number ($R_0$) was above 1, potentially indicating a change in transmissibility [14].

In Africa, exposure to animal reservoirs, including squirrels of the genera *Funisciurus* and *Heliosciurus*, is a significant risk factor for MPXV human infections [15–19]. However, evidence suggests that human-to-human transmission is increasing, [20] [21] which is coincident with decreasing herd immunity associated with the smallpox vaccination campaign that ended in 1980; and correlates with changes in the genomes of secondary cases that result in gene loss and changes in repetitive regions [22].

OPXV infections are classified as systemic or localized illnesses. Generalized disease usually manifests with a rash. Localization usually means signs are restricted to the site of entry. The species of OPXV, its route of entry, and the immune status of the host are

usually the only determinants of the type of infection. Different mechanisms of viral entry and egress, as well as virus-coded host factors, are the main biological determinants [23–27]. However, most of what we know about MPXV virus transmission was learned from intranasal or aerosol challenge models developed to represent a respiratory route of infection. When the exposure was either percutaneous or upper respiratory, its clinical disease course varied [28]. Thus, it is reasonable to posit that the differences in clinical presentation, epidemiology, pathogenesis, and disease development observed during this outbreak could be related to a different route of molecular transmission. MPXV moves from regional lymphatics to the bloodstream causing a primary viremia, and then multiplies in the spleen, liver, bone marrow, and other reticulo-endothelial organs. After, a second viremic period ensues, followed by seeding of distant sites, specifically the skin, and generation of the characteristic generalized rash.

Although OPXV are antigenically and genetically similar, they have diverse host ranges and virulence properties [29–33]. OPXV evolution is driven by selective pressure from host species [31–35] and gene loss of virus-host interacting genes [36]. Large double-stranded DNA virus survival depends on their ability to resist host defenses. Thus, they accumulate the most antidefense genes to counteract their lack of plasticity [36]. Retrospective evolutionary studies provided evidence of OPXV's unique adaptive strategies [37,38]. Elde [39] and Senkevich [40] introduced the concept of "genomic accordions" at the gene and base level, respectively, to explain their ability to adapt. Elde et al. demonstrated that OPXV rapidly acquired higher fitness by massive gene amplification when forced through severe bottlenecks *in vitro*. Gene amplification facilitates the gain of adaptive amino acid substitutions in the gene copies. Subsequent reduction in gene copy offsets the costs associated with the larger genome size, allowing the retention of adaptive substitutions [39]. Similarly, Senkevich et al. described a similar feature during *in vitro* passaging. In that case, rapid adaptation was driven by single-nucleotide insertions or deletions within runs of As or Ts, resulting in an easily reversible gene-inactivating frameshift mechanism [40]. This feature, leading to attenuation *in vivo,* is likely one of the bases of the attenuation of serially passaged vaccines.

Interestingly, in addition to stretches of homopolymers, the genomes of OPXV also include a significant number of repeat regions (STR, Short tandem repeats) with various levels of complexity (dinucleotide, trinucleotide or more complex palindromic repeats). These genomic features had not been yet studied in the context of evolutionary adaptation. Given the potential changes in person-to-person (PTP) transmissibility efficiency, we set to evaluate the effects of these low-complexity-regions (LCR; including all STR and homopolymers) changes in the MPXV genome. Although advances in field sequencing technologies have facilitated access to MPXV genomes, these areas have been frequently

neglected as they are hard to resolve. We obtained a high-quality genome from an unpassed vesicular fluid from a MPX disease case in Spain by combining different technologies' properties to get a complete genome [41]. Moreover, we assessed its variation in LCR areas along a set of clinical specimens collected during the ongoing outbreak to determine the levels of intra-host and inter-host variation. We uncovered significant regions of variation that might provide insights into changes in transmissibility.

## *Methods:*

**Study design and population** We performed a genomic study of confirmed cases of MPX diagnosed from 18th May 2022 until 14th July 2022 at the National Center for Microbiology (NCM). The study was performed as part of the Public Health Response to MPXV by the Spanish Ministry of Health. All samples are listed in **Supplementary Table 1**.

**Samples:** Mainly swabs of vesicular lesions in viral transport media were sent refrigerated to the NCM. Nucleic acids were extracted using either QIAamp MinElute Virus SpinDNA or QIAamp Viral RNA Mini kits from Qiagen (QIAGEN), according to the manufacturer's recommendations. Inactivation of samples was conducted in a certified Class II Biological Safety Cabinet in a BSL2 laboratory under BSL3 work practices with appropriate PPE to reduce the risk of exposure further.

**Laboratory Confirmation:** The inclusion criteria defined laboratory confirmation as a positive result by PCR of MPXV from cutaneous lesions swabs. A generic Real-time PCR for the OPXV genus was used for screening [42]. For confirmation,in the very early stages of the outbreak, we used a conventional validated OPXV nested PCR [43] targeting the TFN receptor gene.

**Genomic Sequencing:** DNA was extracted from the patient's pustule swabs using either QIAamp MinElute Virus SpinDNA or QIAamp Viral RNA Mini kits from Qiagen. Sequencing libraries were prepared with the tagmentation-based Illumina DNA Prep kit and run in a NovaSeq 6000 SP flow cell using 2 x 150 paired-end sequencing. To improve the assembly quality, the library from sample 353R, an unpassed vesicular fluid from a confirmed case, was also run in a MiSeq v3 flow cell using 2 x 300 paired-end sequencing. Additionally, sample 353R was also analyzed by single-molecule methods using Oxford Nanopore technologies. For nanopore sequencing, 210 ng of DNA extracted from 353R pustule swab was used to prepare a sequence library with the Rapid Sequencing Kit; the library was analyzed in an FLO-MIN106D flow cell for 25 hours. The process rendered 1.12 Gb of filter passed bases.

### Bioinformatic analysis:

***De Novo Assembly Methods to obtain a High-Quality MPXV Genome (HQG):*** The sample 353R was selected for this purpose, given the high yield of MPXV genomic material in a preparatory run. Single-molecule long sequencing reads were preprocessed using Porechop v0.3.2pre [44] with default parameters. Next, reads were *de novo* assembled using flye assembler v2.9-b1768 [45] in single-molecule sequencing raw read mode with default parameters. The process resulted in one contig with a length of 198,254bp identified as MPXV. Short 2x150 sequencing reads were mapped with bowtie2 against the selected contig, and resulting bam files were used to correct the assembly using pilon v1.24 [46]. At this intermediate step, the single-molecule *de novo* assembly corrected with pilon was used as a reference in the viralrecon pipeline [47] for mapping and consensus generation with short sequencing reads. We used 0.5 as the allele frequency threshold for including variant positions in the corrected contig.

Short MiSeq 2 x 300 and NovaSeq 2 x 150 sequencing reads were also *de novo* assembled using viralrecon pipeline v 2.4.1. (https://github.com/nf-core/viralrecon) [47], written in Nextflow (https://www.nextflow.io/) in collaboration between the nf-core community [48] and the Bioinformatics Unit of the Institute of Health Carlos III (BU-ISCIII) (https://github.com/BU-ISCIII). Fastq files containing raw reads were first analyzed for quality control using FastQC v0.11.9. Raw reads were trimmed using fastp v0.23.2 [49]. The sliding window quality filtering approach was performed, scanning the read with a 4-base-wide sliding window and cutting 3' and 5' base ends when average quality per-base dropped below a Qphred33 of 20. Reads shorter than 50 nucleotides and reads with more than 10% read quality under Qphred 20 were removed. Host genome reads were removed via a kmer-based mapping of the trimmed reads against the GRCh38 NCBI human genome reference using Kraken2 v2.1.2 [50]. Then, the remaining non-host reads were assembled using SPADES v3.15.3 [51,52] in rnaviral mode. A fully ordered genome sequence was generated using ABACAS v1.3.1 [53] based on the MPXV_USA_2022_MA001 (ON563414.3). The independently obtained *de novo* assemblies and reference-based consensus genomes for 353R were aligned using MAFFT v7.475 [54] and visually inspected for variation using Jalview v2.11.0 [55].

*Systematic identification of Low-complexity regions* (*LCRs) in OPXV genomes:* Detection of STRs in the HQG and other OPXV genomes was performed with the software Tandem Repeat Finder [56], using default parameters. Briefly, the algorithm works without the need to specify either the pattern or its size. Tandem repeats are identified considering percent identity and frequency of indels between adjacent pattern copies and using statistically based recognition criteria. Since Tandem Repeat does not detect single nucleotide repeats, we developed an R script to systematically identify homopolymers of at least 9 nucleotides in all OPXV available genomes. STRs and homopolymers were annotated as LCRs.

*Curation of LCRs in the HQG:* We curated the LCR in our HQG using a modified version of STRSEARCH software (https://github.com/AnJingwd/STRsearch). STRSEARCH, when provided with identifying 15bp flanking regions, performs a profile analysis of STRs in massively parallel sequencing data. However, to ensure high-quality characterization of the LCR alleles, we modified the script to complement reverse reads that map against the reverse genome strand according to their bam flag. In addition, output was modified to add information later utilized by a custom python script to select only reads containing both LCR flanking regions. Modified code can be found here (https://github.com/BU-ISCIII/MPXstreveal). All LCR areas in the HQG were manually validated using STRSEARCH results and *de novo* assemblies obtained from all sequencing approaches. For additional completeness, when an LCR was only resolved by single-molecule long sequencing technologies (LCR1/4 and LCR3), we also analyzed publicly available data. For this purpose, we downloaded all single-molecule long sequencing data from SRA as of August 10, 2022, and analyzed it according to **Supplementary File 1B**.

*Final assembly:* The consensus genome constructed with the viralrecon pipeline [47] using the corrected *de novo* contig as stated above, along with the resulting curated and validated consensus LCRs were used to build the final HQG reference genome. The resulting high-quality genome is available in ENA with identifier: OX044336.2.

***Reference-based consensus for Clade IIb MPXV Spanish sample set:*** For all the remaining samples in our study, sequencing reads were analyzed for viral genome reconstruction using viralrecon pipeline version 2.4.1 [47]. Trimmed reads were mapped with bowtie2 v2.4.4 [57] against the 353R HQG (OX044336.2), and MPXV-M5312_HM12_Rivers (NC_063383.1). Picard v2.26.10 [58] and SAMtools v1.14 [59] were used to generate viral genome mapping stats. Variant calling was done using iVar v1.3.1 [60], which calls for low and high-frequency variants from which variants with an allele frequency higher than 75% were kept to be included in the consensus genome sequence. Finally, bcftools v1.14 [61] was used to obtain the viral genome consensus with the filtered variants and mask genomic regions with coverage values lower than 10X. All variants, included or not, in the consensus genome sequence, were annotated using SnpEff v5.0e [62], and SnpSift v4.3 [63]. Final summary reports were created using MultiQC v.1.11 [64]. Consensus genomes were analyzed with nextclade v2.4.1 [65] using the Monkeypox (All clades) dataset. Raw reads and consensus genomes are available in Bioproject PRJEB53450 with ENA sample accessions identifiers: ERS12168855 to ERS12168865, ERS12168867, ERS12168868 and ERS13490510 to ERS13490543.

*Intra- and inter-host allele frequencies analysis:*

Intra-host genetic entropy (defined as -sum(Xi*log(Xi)), where *Xi* denotes each of the allele frequencies in a position) was calculated according to the SNP frequencies of each position along the genome using viralrecon pipeline results. Similarly, genetic entropy for each of the 21 LCRs was calculated considering the frequencies of repeat lengths.

LCR intra-host and inter-host variations in our sample set were analyzed using the modified version of STRSEARCH software described above. As a filter for quality for this analysis, STRSEARCH search results (**Supplementary Table 4**) were filtered, keeping alleles with at least ten reads spanning the region and allele frequency above 0.03. Quality control and allele frequency graphs were created using a custom R script.

Pairwise Genetic distances between samples were calculated as Euclidean distances (defined as /X-Y/=sqrt(sum(xi-yi)^2), where *xi* and *yi* are the allele frequencies of sample X and Y at a given position, respectively), thus accounting for the major and minor alleles at each position analyzed. These distances were calculated individually for each variable LCR (STRs 2, 5, 7, 10, 11, and 21) as well as for each of all 5422 SNPs displaying inter-sample variability (compared MPXV-M5312_HM12_Rivers) with over the 48 samples. The distributions of Inter-sample distances were compared between LCRs employing a Kruskal-Wallis test followed by a non-parametric multiple pairwise-comparison between groups (Wilcoxon test), where P-values were subjected to the false discovery rate correction. We also tested whether inter-sample variability in LCRs is higher

than that from SNPs by a randomization test: first, we calculated the average Euclidean distance for each LCR and each SNP position. Then, the average value of each LCR was compared to a random sample of 1000 values from the distribution of mean distances from the SNPs along the genome. The P-value was calculated from the percentage of times that the mean of the LCR was higher than the randomly taken values from the SNPs.

***Core genome phylogenetic analysis:*** Variant calling and SNP matrix generation was performed using snippy v4.4.5 [66] including sequence samples and representative MPXV genomes downloaded from NCBI (**Supplementary Table 1**). SNP matrix with both invariant and variant sites was used for phylogenetic analysis using iqtree v. 2.1.4-beta [67] using predicted model K3Pu+F+I and 1000 bootstraps replicates. A phylogenetic tree was visualized and annotated using itol [68]. SNP matrix was also used for generating the haplotype network using PopArt [69].

***Selected MPXV ORF analysis:*** Representative OPXV genomes [36] were downloaded from NCBI together with the consensus genomes from the samples in this study (**Supplementary Table 1**). MPXV genomes were classified by clade and lineage following the last nomenclature recommendations [5] according to nextclade v2.4.1 [65]. Annotation from RefSeq NC_063383.1 gff was transferred to all fasta genomes using liftoff v1.6.3 [70]. OPG153 was extracted using AGAT v0.9.1 (*agat_sp_extract_sequences.pl*) [71] and multi-fasta files were generated for each group and gene. OPG204 and OPG208 alternative annotation start site ORFs were re-annotated in Geneious and extracted as new alignments. We used MUSCLE v3.8.1551 [72] for aligning each multifasta and Jalview v2.11.0 [55] for inspecting and editing the alignments. Finally, Metalogo v1.1.2 [73] was used for creating and aligning the sequence logos for each group OPXV of the OPG153 and LCR7 area. OPG204/LCR21 and OPG208/LCR3 were represented similarly.

All scripts and code used for the paper can be found at github repository: https://github.com/BU-ISCIII/MPXstreveal

***Comparison of the Frequency of LCRs between protein functional groups:*** The potential biological impact of LCRs was evaluated by mapping the frequency and location of STR and homopolymers in the structure of the poxvirus genome and considering the biological function of the genes affected. We compared the frequency of inclusion of LCRs between distinct functional groups of genes as previously established [36]. OPXV (n=231, AKMV: n=6 sequences, AKPV: 1, CPXV: 82, ECTV:5, MPX: 62, VACV: 18, VARV: 57) include 216 functionally annotated Orthologus Poxviral Genes (OPGs) classified in 5 categories (“Housekeeping genes/Core” ANK/PRANC family, Bcl-2 domain family, PIE family, and “Accesory/Other” (e.g., virus-host interacting genes)). The frequency was calculated after normalizing their count number with the sample size of the OPG alignment. Statistical Analysis of the significance of differences was performed employing a Kruskal-Wallis test followed by a non-parametric multiple pairwise comparison between groups (Wilcoxon test), where P-values were subjected to the false discovery rate correction.

## *Results*

### A complete high-quality monkeypox genome assembly and annotation

Given the evolutionary importance of STR and homopolymers in the evolutionary history of poxvirus [39,74], we focused on the characterization and validation of these resolved regions using three sequencing platforms.

Shotgun, short-read based sequencing from vesicular lesions allowed us to reconstruct 47 MPXV genomes with at least 10X read depth using a reference-based assembly approach. A median of 39,697,742 high quality reads per sample (max 111,030,976, min 7,780,032) were obtained using the NovaSeq 6000. Although 98.12% of reads belonged to the human host, a median of 74,085 virus reads (max 27,516,891, min 30,854) were enough to reconstruct >99% of the genome (**Supplementary Table 2**).

However, although most of the genome structure was resolved, read mapping showed that LCRs were mostly unresolved. More importantly, those results were biased by the reference genome used as a scaffold. In general, the observation is that short tandem repeats are resolved by reference mapping software tools “following” the pattern provided in the scaffold reference genome instead of reporting the actual pattern (**Supplementary Fig. 1a).**

Thus, we explored different assembly strategies generally used for resolving eukaryotic genomes, which mostly combine different sequencing technologies. For this purpose, we used one of the samples with a higher proportion of high-quality viral reads (353R). The *de novo* assembly obtained from NovaSeq (2x150bp pair-ended reads), MiSeq (2x300bp pair-ended reads), and Nanopore sequencing generated 3, 2, and 1 contig belonging to MPXV covering 97%, 97%, and 101% of reference genome NC_063383.1. **Fig. 1** shows the fully annotated genomes (based on the NC_063383.1 genome annotation), comparing the contigs obtained using the different approaches.

Based on our previous experience during the investigation of human-to-human transmission of monkeypox in the DRC [22], we utilized a systematic approach for LCR discovery that resulted in the identification of 21 LCRs (13 STR, 8 homopolymers; **Table 1**) in our finished contig. Annotated position (according to the reference genome NC_063383.1), pattern, and flanking region of each area (defined as in [75]) are provided in **Supplementary Table 3.** Four of those regions (pairs LCR1 and 4; and LCR10 and 11) are located in the left and right inverted terminal repeats (ITRs) and are identical copies in reverse complementary form. In strict terms, we could not resolve these regions, as no reads joined these LCRs with unique areas of the genome. Thus, LCR1 and 4; and LCR10 and 11 are treated as the same.

In general, LCRs were resolved using the assembly obtained from single-molecule sequencing and further validated using short-read sequencing reads since most patterns range in length between 13 and 67bp and therefore are covered by reads from each side or flanking region

without mismatches. **Supp. Fig. 1b** displays an example for LCR7; all other resolved areas are displayed in **Supp. File 1A**. All LCRs (except 1/4 and 3) were validated in this form. LCR1/4 (256bp) and LCR3 (468bp) were only resolved with single-molecule sequencing reads due to their length (**Table 2).**

LCR3 contains a complex tandem repeat with the form ATAT [ACATTATAT]n. Our analysis shows 52 repeats. No current publicly available MPXV genomes had reported a similar length. However, our analysis of 35 SRA publicly available MPXV nanopore sequencing runs shows our pattern is reproducible (**Fig. 2a).** Fifteen samples have supporting long reads that include both flanking regions. Interestingly, four samples from the 2022 MPXV outbreak (Lineage B.1 - Clade IIb) have repeats ranging between 54 and 62 for LCR3. This diverges from the rest of the 2018-2019 Lineage A - Clade IIb samples that range between 12 and 42 repeats. These results demonstrate LCR3 as a region of genomic instability and high variability, with evolutionary changes leading to the 2022 outbreak.

LCR1/4 contains a complex tandem repeat with the form [AACTAACTTATGACTT]n. Our analysis shows 16 repeats. Instead, both reference strains (NC_063383.1 and ON562414.3) have only eight repeats (**Table 3**). The analysis of the publicly available data confirms our observation (**Fig. 2b)**. Among Clade IIb, B1 strains show 16 repeats consistently. A.1 strains are polymorphic showing 14 (n=1), 16 (n=3), 17 (n=7) and 19 (n=1). A.2 strains show 23, 25, and 26 repeats, while older lineage A strains have 32, 43, 53, and 71 repeats.

We propose this genome for sample 353R as a high-quality complete reference genome, HQG [41]. We compared our 198,547 bp HQG to MPXV-M5312_HM12_Rivers (NC_063383.1) and MPXV_USA_2022_MA001 (ON563414.3) [13] (**Table 3)**. Both 353R and ON562414.3 genomes show the same 67 SNPs called against the NC_063383.1 reference genome. Additionally, 353R has 2 additional paired SNPs in the left and right ITR (5595G>A; 191615C>T to NC_063383.1) that result in the introduction of a stop codon in OPG015. We observed this variation in only two other samples among our sample set). 353R and ON562414.3 also differ by two INDELs at positions 133,077 and 173,273 that correspond to differences in the areas of LCR2 and 5, respectively). As a result of the resolution of the LCR areas, 353R differs in genome length by 1342bp against ON562414.3 and 1338bp against NC_063383. Most of the variation is due to differences in the length of LCR1/4 and LCR3, along with minor length differences in LCR2, LCR5, and LCR10/11. In general, the number of repeats found with the hybrid assembly approach in these areas doubles their length.

**Nonrandom distribution of LCRs in the MPXV genome**

We compared the distribution of LCRs between different major functional groups as previously described [36]. Differences between functional groups were statistically significant (Kruskal-Wallis test, P-value <0.001) Pairwise analysis demonstrated that the functional group “Core” includes LCRs at a significantly lower frequency (multiple pairwise-comparison Wilcoxon test) than functional groups “ANK/PRANC” (corrected P value <0.0001), “Bcl-2 domain” (corrected P value =0.04) and

"Accessory" (corrected P value <0.0001) (**Fig. 3**). This analysis indicates that regions of low complexity in poxvirus genomes are non-random. Moreover, it also shows that there is a significant purifying selection force against introducing LCRs in "core" areas.

We next compared the degree of diversity among the 21 identified LCRs with the observed single-nucleotide polymorphism variability that had been the focus of recent genomic studies. In the HQG, LCRs 2, 5, 7, 10, 11, and 21 displayed intra-host genetic diversity, with entropy values that ranged from 0.18 (LCR7) to 1.66 (LCR2), with an average of 0.81 and a SD = 0.64 among them (**Table 2**). Only five nucleotide positions displayed intra-host genetic diversity at the level of SNPs (positions 1285, 6412, 88,807, 133,894 and 145,431). The entropy values ranged from 0.17 (position 133894) to 0.69 (position 6412), with an average of 0.38 and a SD = 0.21 among them. Interestingly, a Student's t-test revealed a significantly higher level of diversity in LCRs than in SNPs (P value =0.021; **Fig. 4a**).

We then characterized, collected and compared the allele frequencies for all LCR from all the samples in our dataset, with the filters described above. The complete list of alleles observed is listed in **Supplementary Table 4**. Our inter-sample distance analyses revealed that the average inter-sample Euclidean distances at LCRs ranged between 0.05 (LCR21) and 0.73 (LCR2). We found statistically significant differences between LCRs (Kruskal-Wallis chi-squared P value <0.001). More specifically, the multiple pairwise comparison Wilcoxon test reported that all LCRs displayed significantly different levels of inter-sample distance (FDR corrected P-values < 0.001), except for the comparison of LCR10 vs. LCR11 (corrected P value =0.48) and LCR2 vs. LCR5 (corrected P value =0.25) (**Fig. 4b**). Average distances in SNPs ranged between 0.0018 and 0.4168. Our randomization tests revealed that all LCRs display a significantly higher level of inter-sample diversity than the SNPs (all corrected P-values < 0.05) (**Fig. 4b**). These analysis shows that most of the variability in the poxvirus genome is located in LCRs areas. We posit that sequencing methods to study transmission between strains should change their focus to the study of LCR variability instead of SNP variability.

Only two samples (353R and 349R) produced enough sequence coverage information to allow us to perform an allele frequency comparison in most LCR areas (**Fig. 5a**). Their side-by-side comparison allows us to see apparent differences in allele frequency in some areas of the genome (2, 5 and 10/11) (**Fig. 5b**). The rest of the samples only show enough coverage to unequivocally resolve a subset of the LCRs (e.g., covering both flanking regions; 2, 5, 7, 8, 9 and 10/11). LCR8 and LCR9 do not show any variation among our sample set. However, LCR7 and LCR10/11 showed considerable variation intra-host, as well as differences in the preponderant allele (LCR10/11) between samples (**Fig. 5c**).

**Use of SNPs variability to identify microevolution events**

Phylogenetic and haplotype network analysis of the Spanish isolates in the context of the outbreak yielded limited information (**Supp. Fig. 1a and 1b**). While most samples had no significant changes

and were, therefore, part of the basal ancestral MPXV Clade IIb B1 node, a few sequences form supported clusters: Group 1: Sample 395, 399, and 441 clustered along with USA 2022 FL002; Group 2 formed by sample 353, 352, 347 and 416 which all shared the stop codon mutation in OPG015; Group 3 is formed by 2369 along with SLO (from Slovenia) and HCL0001 (from France) (Lineage B.1.3); Group 4, formed by 2437 and 417; Group 5, that includes 2388, 2428, 1300 and 698, and RK001 (from Germany) (Lineage B.1.1) and Group 6 that links 2309 and 2317 (**Supplementary Table 5**)

Based on the information available, only one epidemiological link among members of the same clusters was recognized. Sample 395 (39yo male, MSM) and 399 (35yo male, MSM) are sexual partners and attended events in Madrid, Spain, and Porto, Portugal. No link to sample 441 was found. In summary, although there is at least one case where genomic surveillance could detect a link, all other groupings had not been epidemiologically supported. Given their grouping with other international samples, those changes might indicate convergence or genomic areas more prone to sequencing errors. In summary, at least at this time in the outbreak, there appears to be limited value in the return-of-investment of SNPs whole-genome sequencing.

**The biological significance of the areas of higher variability**

Since LCR entropy is significantly higher than SNPs; LCR are not randomly located in the genome; we have shown previously that PTP-associated changes were observed in the immunomodulatory region [22]; and genomic accordions are a rapid path for adaptation of poxvirus during serial passaging [39,40], we posit that changes in LCRs might be associated with adaptive changes related with transmissibility differences.

Although most LCRs that show variability in our sample set (1/4, 2, 3, 5, 7, 10/11, and 21) are located in intergenic regions, some (3, 7, and 21) are located into coding regions that, considering poxvirus evolutionary history, are associated with virulence or transmission. Noteworthy, 3 of the 21 highly repetitive areas identified in our intra-host variation analysis (LCR5, LCR6, and LCR7) are located in a defined “core” area of the poxvirus genome between positions 133,000 and 138,000 (**Fig. 1**). This genomic region encodes for OPG152 (VACV A25L), OPG153 (VACV A26L, directly affected by LCR7), and OPG154 (VACV A27L). LCR7 is the only STR that is encoded at the center of a functional ORF. Instead, both LCR3 and LCR21 are situated in the promoter/start area, potentially modifying the ORF start site. The repeat area of LCR7 encodes for an aspartic acid homopolymer in a non-structured region of the A26L (**Fig. 6a**). The changes observed encode for insertion of 2 isoleucines (I) in the middle of the long aspartic acid (D) repeats. The change resembles the primary structure of the Clade I strains, which also include a couple of Ile residues in this disordered region). Instead, Clade IIa African strains (pre-2017) have no insertions (**Fig. 6a**).

The area downstream of LCR3 is another region of potential impact. LCR3 repeat [CATTATATA]n is located 21bp upstream of the putative translation start site of OPG208. Significantly, immediately upstream of the LCR3 there is a start methionine codon. The upstream

start codon has a “mid-to-low” probability of being translated (T base in position -3), compared to a “strong” Kozak sequence in the downstream putative OPG208 start codon. Nevertheless, LCR3 maintains the correct in-frame form in all Clade II strains, which indicates selective pressure to maintain the possibility of alternative start translation (**Fig. 6c**). Interestingly, the LCR3 would not be in-frame in most Clade I strains.

OPG208 has been identified in genomic comparisons of MPXV strains of Clade I and Clade IIa among a set of genes most likely responsible for the increased virulence of Clade I [29]. The LCR3 tandem repeat CATTATATA in the MPXV 2022 (Lineage Bs, Clade IIb) is present with 52, 54, and 62 copies (**Fig. 2a**), while SL-V70, WRAIR-61, and COP-58 (all Clade IIa) have been reported as presenting 7, 37 and 27, respectively [29], and ZAI-96 (Clade I), 16. Interestingly, all Lineage As Clade IIb samples with publicly available single-molecule long read data have a number of repeats below 40 (**Fig. 2a**). Alternatively, given the nature of the repeat sequence, it could also potentially alter the promoter function. Interestingly, the LCR3 repeat sequence also introduces codons with a low usage ratio that would not be optimized for expression in primates. The codon triplet ATA, which encodes for one of the Isoleucine, has a rare codon usage of 0.17 (**Fig. 7**). The potential difference in transcription and/or translation was the basis to posit B19R as a potential marker of virulence [29]. OPG208, also called Cop-K2L, B19R, or SPI-1, is a serine protease inhibitor-like protein that has been identified as an apoptosis inhibitor [76]. Apoptosis of the infected cell prevents virus proliferation and protects nearby cells, providing the first and probably more ancient line of nonspecific defense against pathogens [77].

Similarly, the area downstream of LCR21 shows a similar anomaly. The STR introduces an ATG codon upstream of the putative start codon for OPG204 (**Fig 6b**). Kozak sequence analysis revealed a mid-high probability of translation compared with the putative start codon (**Fig. 7**). The repeat introduces a non-optimized codon, AAG, which encodes for Lysine, with a low codon usage of 0.25. The parallel between OPG204 and OPG208 might indicate this is a conserved evolutionary trait.

## Discussion:

**Observed Differences in transmission during the 2022 MPXV outbreak:** Historical studies of MPXV indicated that the geographically defined Clade I and II isolates had distinct clinical and epidemiological parameters [1]. Identification of human infections in both geographical areas was first made in the 70s; but the number of reported human cases remained low until the late 90s. A re-emergence of disease due to Clade I MPXV was observed in the DRC in 1996; with the additional salient observation that more cases were derived from secondary person-to-person contact (88%) than in any earlier period in history [78]. This was, in part, attributed to a larger population of humans fully susceptible to disease because of the cessation of routine smallpox vaccination in 1980 [20]. In 2003, up to 7 generations of uninterrupted spread among humans were reported [79]. MPXV Clade IIa was introduced to the United States in 2003 via a consignment of wild-captured animals from Ghana. Detailed comparison of the clinical and epidemiologic characteristics of the U.S. cases demonstrated significant differences in presentation, severity, and transmission compared with Clade I cases from Africa [3]. No MPXV-related mortality or PTP transmission was observed among Clade IIa cases.

In 2017, MPXV Clade IIb emerged in Yenagoa Local Government Area, Bayelsa State, Nigeria. Although phylogenetic analysis indicated that the closest ancestor of the novel outbreak was an isolate from a human MPXV case in Ihie, Abia State, Nigeria, in 1971, only ~10 cases were detected during the intervening 40 year-period, indicating the index case was not imported, but probably originated from a spillover event from a new local reservoir host [8]. Since then, more than 800 cases have been reported. They showed an unusual pattern: higher prevalence among adults (78% of patients were 21–40 years of age), whereas historically, most case patients were <15 years of age. Two main factors were posited to explain the resurgence: (1) increased exposure to and interactions with forest animals; and (2) waning immunity from since-discontinued universal smallpox vaccination programs.

Importations of Clade IIb to the UK, Israel, and Singapore in 2018/19 and to the USA in 2021 were observed but did not trigger secondary cases. Instead, the current resurgence was associated with potential superspreading events in the UK, Belgium, and Spain, as well as a surprisingly high $R_0$ among the MSM community and its social networks. Moreover, the disease itself presented with a significant departure from typical patterns [78,80–88]. In the current outbreak, MPXV appears to transmit after a primary localized rash that removes the requirement to establish a disseminated infection for transmission [89]. This observation is supported by the low number of reports of disseminated infection. If MPXV PTP transmission is increasing, we should expect to see related genotype changes. However, given their known strategies to maintain redundant pathways, we should also not expect radical but modulating changes.

**Genotype-to-phenotype analysis:** Comparative genomics has been applied extensively to demonstrate relations between OPXV genotype and phenotype [22,29,31,34,35,90–92]. Studies of MPXV

genomics during the 2022 outbreak so far have focused on describing its evolutionary history and tracking its introduction to the virus in Western countries. The 2022 MPXV cluster diverges from the related 2016-2019 viruses by an average of 50 single-nucleotide polymorphisms (SNPs). Of these, the majority (n=24) are non-synonymous mutations with a second minority subset of synonymous mutations (~18) and a few intergenic differences (4) [93]. A strong mutational bias mainly attributed to the potential action of apolipoprotein B mRNA-editing catalytic polypeptide-like 3 (APOBEC3) enzymes had also been observed [94]. Moreover, genetic variation, including deletion of immunomodulatory genes, had been described [95]. MPXV Lineages had been described, although mostly they represent very small variations usually encoded by one or two SNP differences to the basal node [96]. Four of our samples are described as B.1.1, along with 97 other sequences worldwide (as of September 1st) which are defined by an amino acid mutation in OPG094 R194H corresponding to homoplasy G74,360A. One of our samples is described as B.1.3, along with 38 other sequences worldwide, defined by the amino acid mutation R84K in NBT03_gp174 corresponding to position G190,660A. The remainder of our samples are B1. We detected additional clusters among the Spanish isolates also defined by few SNP changes, but only in one case we identified an epidemiological link. In summary, there appears to be limited relation between SNP changes and virus epidemiological history, which might hint to a potential convergence effect. Genomic epidemiology for this class of virus might need a change of focus.

Our analysis of intra- and inter-host variability demonstrated that larger variation is located in areas previously considered of poor informative value. We compared the values of heterozygozy through the whole genome (**Fig. 4a**) and the magnitude and scale of change is significantly higher in LCR than SNPs. This variation is also observed intra-host (**Fig. 4b**). We posit that some of this variability is associated with biological features. We demonstrated that LCR are enriched in defined coding areas of the OPXV genome. Furthermore, we found a very strong correlation between areas of LCR in the MPXV genome and non-core genes (**Fig. 3**). Enriched LCR were associated with several families of diverged paralog genes involved in various intracellular and extracellular signaling pathways (“Ankyrin”, “Bcl-2” and “BTB_Kelch”), although this finding might not be surprising, giving their “repetitive domain” nature. On the other hand, the substantial difference found when comparing “Accessory/Other" with “Core/Housekeeping” gene categories demonstrates that LCRs are strategically situated in modulatory and adaptive areas, providing additional proof that these areas should be scrutinized for changes that might affect the OPXV interactome. This resembles adaptive evolutionary strategies observed in complex parasitic pathogens [97].

Eight LCR areas showed evident signs of intra-host and inter-sample variation (1/4, 2, 3, 5, 6, 7, 10/11, 21). Five of them (5, 6, 7, 3 and 21) were co-located in two areas of the MPXV genome: 130,000 to 135,000 (5, 6 and 7) which is clearly in the “core” area of the OPXV genome where most “housekeeping” genes are located; and 175,000 to 180,000 (3 and 21), which is located in the immunomodulatory area. Three of those LCR are located inside the putative translated region of

MPXV genes OPG153 (A26L), OPG204 (B16R) and OPG208 (B19R). While changes in OPG204 and OPG208 are located near the N-terminal region and might involve modulating the expression of translation, the changes observed in OPG153 stand out as they are located inside a region under high selective pressure for transmission, among other OPXV [36] in a gene that is clearly considered a "housekeeping" gene involved in attachment and egress. The OPG153 repeat results in a poly-Asp amino acid homopolymer string (**Fig. 6a**); the N-terminal domain variation in OPG204 results in a Met-Lys repeat (**Fig. 6b**) and the LCR repeat in OPG208 results in an Ile-Ile-Tyr repeat (**Fig. 6c**). The remaining were located downstream of known ORFs; thus, unlikely to excerpt a modulating effect (2, 4/1, 10/11).

Functional repetitive microbial proteins have been described. Self-association guided by stretches of single amino acid repeats is often described in nature, which leads to the formation of aggregates [98]. Many human diseases are associated with detrimental effects of homopolymers [99,100]. Expansion or contraction of the domain increases their self-attraction and triggers disease. These homopolymers can also regulate the activity of transcription factors [101] or direct ORFs to different cellular compartments [98] or nuclear localization [102].

The changes in the number of repeats observed in LCR3 and LCR21 follow the same pattern by extending the N-terminal region of an immunomodulatory ORF. Functional translation studies to verify if this region is translated are needed. Nonetheless, this strategy has been observed already to modulate ORF translation in the microbial world. The yeast proteins Flo1p and Flo11p function is proportionally modulated by the repeat length of their N-terminal region [95,103]. A relatively small change in the number of tandem repeated sequences is crucial to its adaptation to a new environment. In *Plasmodium*, the exported glutamic acid-rich protein (GARP) contains repetitive sequences that direct the protein to the periphery of the infected erythrocyte [104]. At least nine other exported plasmodium proteins target the periphery of the erythrocyte using this strategy. Interestingly, the lengths of the tandem repeat vary between parasite strains [104]. The localization of Hyp12, a parasitic protein that modulates the rigidity of the infected cell [105], is defined by a repetitive Lys-rich sequence and a repetitive acidic sequence [104].

However, that is not the only potential way that N-terminal expansions can modulate ORF expression. The translation rate is regulated by the concentration of available aminoacylated tRNAs. Encoding unusually long stretches of homopolymers by rare amino acids codon diminishes the efficiency of the process. If, in addition, the amino acid is encoded by non-codon optimized triplets like in the Tyr and Lys codons in the repetitive regions of LCR3 and LCR21 (**Fig. 7**). Moreover, given the unusually long repetitive stretch of LCR3, Clade IIb MPXV might require large amounts of Tyr:tRNA to efficiently translate OPG208. The protein encoded by OPG208 belongs to the serine protease inhibitor superfamily and is called serine protease inhibitor-1 (SPI-1) [106]. This protein was demonstrated to be a host-range factor required for replication in different host cells [107]**.** SPI-1 is conserved in OPXV and expressed as an intracellular non-glycosylated 40-kDa species [108].

Morphological examination of SPI-1-deleted infected A549 cells, together with an observed fragmentation of cellular DNA, suggests that the host range defect is associated with the onset of apoptosis [107].

The changes associated with LCR21/OPG204 are subtle. We do not observe changes in the number of repeats, but mutations (**Table 2**). Thus, most Clade IIb viruses show multiple alternative starts followed by a Lys, always encoded by the rarest codon. The protein encoded by OPG204, also known as B16R in VACV-Cop, is a known inhibitor of the Type I interferon (IFN) system shown to act as a decoy receptor [109–112].

During replication, OPG204 is secreted to bind IFN-I with high affinities and prevent its interaction with IFNAR. The well characterized secreted IFNα/βBP OPG204/B16R lacks a transmembrane domain [112,113] and binds to glycosaminoglycans (GAGs) in the surface of infected and surrounding uninfected cells preventing the IFN-mediated induction of an antiviral state [112,114,115]. Thus, it is important to test whether the differences in length of the repetitive sequences between Clade I, IIa and IIb (**Fig. 6b**) could have an effect in translation.

The more intriguing region of variability observed in our dataset involves OPG153 (A26L). There are several lines of evidence that mark this ORF as a significant factor in transmission and virulence among OPXV: (1) OPG153 is a known attachment factor (to laminin) for OPXV [116,117]; (2) OPG153 is a significant factor regulating egress for OPXV [116,118–120]; (3) In a comparative genomic analysis of poxviruses, OPG153 is unique by being the “core” gene that has been “lost” the most times during poxvirus evolution [36]; (4) during experimental evolution experiments, inactivation of OPG153 genes by frameshift mutations provided rapid adaptation in a poxviral model [40]. These changes resulted in increased virus replication levels, changes in morphogenesis quantified by EM, in decreased particle/PFU ratios [40] and differences in pathogenesis [119]; and (5) Advanced dissection of the adaptive immune response against MPXV demonstrated that OPG153 is the main target of the antibody response [121,122].

The mechanism of entry of Poxvirus in cells is highly versatile. They produce two infectious particles: mature (MVs) and extracellular virions (EVs). Although the MV is the more abundant form, the EV is specialized for cell-to-cell spread. EVs, produced and secreted early in the infection, are essential in spread infection within the host, however MVs that are produced and accumulate inside the infected cell until lysis are important for host-to-host transmission [123]. Both MVs and EVs of VACV take advantage of host cell endocytosis for internalization by activating macropinocytosis. MVs also use fusion with the plasma membrane and explain VACV ability to enter most cells. MV attachment is mediated by OPG153 and OPG154 binding to cell surface laminin and glycosaminoglycans, respectively. MVs of some chordopoxviruses (e.g., CPXV, ECTV, raccoonpox virus, and FWPV) but not MPXV, VARV or VACV become occluded in a dense protein matrix within the cytoplasm called A-type inclusions (ATIs, formed mainly by OPG153, is co-located in the same genomic area). ATIs are released following degeneration of infected cells and protect the enclosed

MVs from the environment. Interestingly, some CPXV mutants lacking OPG153 form inclusions without virions [119,124,125]. VACV strains had considerable variation regarding their preference for alternative pathways, which might depend on the ATI and OPG153 proteins [120]. Interestingly, 3 of the 21 highly repetitive areas identified in our intra-host variation analysis (LCR5, LCR6 and LCR7) are located in the poxvirus genome (133,000 to 138,000) area where these ORFs are encoded (**Fig. 1**).

OPG153 is a bridge between the ATI protein OPG152 and OPG154, which is tethered to MVs. Interestingly, in VACV (which does not form ATIs), a truncated homologous OPG152 protein is encoded by some strains that associate with OPG153 [118]. MPXV, which also does not form ATIs, also has a truncated OPG152. Neither OPG153 nor OPG152 protein are present in EVs. EV formation occurs when MVs are wrapped by a pair of additional membranes derived from virus-modified trans-Golgi or endosomal cisternae. Studies with mutant VACVs indicate that severe effects in wrapping are caused by repression or deletion of the OPG153 MV protein, OPG057/F13 or OPG190/B5 [40,118,126]. Interestingly, the antiviral ST-246 inhibits the wrapping of MVs and formation of EVs by targeting OPG057 which is required for Golgi membrane localization [126].

The LCR7 repeat area, located in the central domain of OPG153, encodes for a poly-aspartic acid non-structured region (**Fig. 6a**). That acidic region is conserved among OPXV; however, its length is highly variable. As mentioned above, amino acid polymer stretches had usually been associated with self-aggregation mechanisms. In mammals, a poly-Asp stretch appears to provide functionality to asporin, a small leucine-rich repeat proteoglycan (SLRP) class I that also possesses a unique stretch of aspartate residues at its N terminus [127] associated with calcium-binding functions [128]. Interestingly, poxvirus evolutionary clades that form ATIs, generically have very long poly-D stretches. Instead, VACV and VARV that do not form ATI have reduced their poly-D stretches to the minimum (4 aa, in both). Among MPXV strains, different patterns are observed. Clade IIa strains have an extended 21 aa poly-D. Intriguing, Clade I and IIb strains disrupt their poly-D stretch with the insertion of 2 isoleucines. Intriguing, both disruptions result from the incorporation of the same “ATCATA” nucleotide insertion in the “GAT” repetitive stretch.

In summary, we believe that our findings expand the concept of genome accordions in OPXV evolution. The highly repetitive structure of the genome of OPXV resembles the structure of *Plasmodium falciparum*; and facilitates a simple and recurrent mechanism of adaptation in a genomic scale. A consequence of this concept is that LCRs of the genome (highly repeated large tandem repeats, short tandem repeats, and homopolymers), which are currently being neglected during genomic studies, might be crucial to address changes of host range or virulence and might also contain important transmission information. To further take advantage of current NGS capabilities, we need to establish a new standardized approach to generate and analyze the sequencing data that prioritize these regions. Further functional studies to complement this comparative genomic study are urgently needed.

**Disclosures**

The work for this study at Instituto de Salud Carlos III was partially funded by Acción Estratégica “Impacto clínico y microbiológico del brote por el virus de la viruela del mono en pacientes en España (2022): proyecto multicéntrico MONKPOX-ESP22” (CIBERINFEC).

The work for this study at the GP laboratory was funded by insitutional funds of the Department of Microbiology, Icahn School of Medicine at Mount Sinai in support of Global Health Emerging Pathogen Institute activities.

We would like to thank the work of the Rapid Response Unit of the National Center for Microbiology, especially her Head, MªJosé Buitrago. Cristobal Belda, the General Director of the Instituto de Salud Carlos III, was also permanently involved trying to help in the management of the outbreak.

The A.G.-S. laboratory has received research support from Pfizer, Senhwa Biosciences, Kenall Manufacturing, Blade Therapeutics, Avimex, Johnson & Johnson, Dynavax, 7Hills Pharma, Pharmamar, ImmunityBio, Accurius, Nanocomposix, Hexamer, N-fold LLC, Model Medicines, Atea Pharma, Applied Biological Laboratories and Merck, outside of the reported work. A.G.-S. has consulting agreements for the following companies involving cash and/or stock: Castlevax, Amovir, Vivaldi Biosciences, Contrafect, 7Hills Pharma, Avimex, Vaxalto, Pagoda, Accurius, Esperovax, Farmak, Applied Biological Laboratories, Pharmamar, Paratus, CureLab Oncology, CureLab Veterinary, Synairgen and Pfizer, outside of the reported work. A.G.-S. has been an invited speaker in meeting events organized by Seqirus, Janssen, Abbott and Astrazeneca. A.G.-S. is inventor on patents and patent applications on the use of antivirals and vaccines for the treatment and prevention of virus infections and cancer, owned by the Icahn School of Medicine at Mount Sinai, New York, outside of the reported work.

## Figures and Tables

| Name | Location Start HQGa | Location End HQGb | Repeat Unitc | Patternd | Nearest Genee | Type of LCRf | Relative Position to the Geneg | Distance in bph | Copenhagen Notationi | Vaccinia Notationj | Comments |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LCR2 | 174,063 | 174,112 | 2 | [ATAT]n | NA | STR | Downstream | 45 | Cop-B16R | B14R | |
| LCR5 | 133,895 | 133,918 | 1 | [T]n | OPG152 | homopolymer | Upstream | 899 | Cop-A25L | A27L | Fragmented gene area |
| LCR10 | 197,830 | 197,842 | 1 | [T]n | OPG001 (ITR) | homopolymer | Downstream | 209 | NA | NA | |
| LCR11 | 1,286 | 1,298 | 1 | [T]n | OPG001 (ITR) | homopolymer | Downstream | 209 | NA | NA | |
| LCR21 | 175,299 | 175,357 | 6 | [GATGAA]n | OPG204 | STR | ATG Start/Promoter | NA | Cop-B19R | B16R | Alternative ATG repeat start |
| LCR7 | 137,319 | 137,375 | 3 | [ATC]n | OPG153 | STR | Inside ORF | NA | Cop-A28L | A26L | Attachment MVs/Laminin |
| LCR6 | 133,980 | 133,989 | 10 | [CAATCTTTCT]n | OPG152 | STR | Upstream | 818 | Cop-A25L | A27L | |
| LCR1 | 5,369 | 5,624 | 16 | [AACTAACTTATGACTT]n | OPG003 (ITR) | STR | Downstream | 72 | Cop-C19L | NA | |
| LCR1 | | | | | OPG015 (LITR) | STR | Upstream | 35 | CPXV-017 | NA | |
| LCR4 | 193,504 | 193,759 | 16 | [AAGTCATAAGTTAGTT]n | OPG003 (ITR) | STR | Downstream | 72 | Cop-C19L | NA | |
| LCR4 | | | | | OPG015 (LITR) | STR | Upstream | 35 | CPXV-017 | NA | |
| LCR3 | 179,872 | 180,345 | 9 | ATAT [ACATTATAT]n | OPG208 | STR | ATG Start/Promoter | 21 | Cop-K2L | B19R | SPI-1 apoptosis inhibition |
| LCR8 | 147,655 | 147,718 | 5+7 | [ATATTTT]n [ATTTT]n [ATATTTT]n [ATTTT]n [ATATTTT]n [ATTTT]n [ATATTTT]n | OPG171 | STR | Upstream | 75 | Cop-A42R | A42R | |
| | | | | | OPG170 | STR | Upstream | 70 | Cop-A41L | A41L | |
| LCR9 | 151,350 | 151,417 | 9 | [TATGAAG]n [GATATGAT]n [GATATGATG]n [GATATGAT]n | OPG176 | STR | Upstream | 19 | Cop-A46R | A47R | |
| LCR12 | 29,326 | 29,364 | 1 | [A]n | OPG044 | homopolymer | Inside ORF | NA | Cop-K7R | B15R | C-terminal position |
| LCR13 | 76,896 | 76,904 | 1 | [T]n | OPG097/098 | homopolymer | Upstream | 9-Jul | Cop-L3L/L4R | L3L/L4R | |
| LCR14 | 81,658 | 81,666 | 1 | [T]n | OPG104 | homopolymer | Inside ORF | NA | Cop-J5L | L5L | Essential for viral replication |
| LCR15 | 140,911 | 140,977 | 9 | [ATAACAATT]n [ATAATTGTT]n [ATAATAATT]n [ATAATTGTT]n | OPG159 | STR | Inside ORF | NA | Cop-A31L | A33L | PKR inhibitor candidate? / C-terminal position |
| LCR16 | 153,457 | 153,465 | 1 | [A]n | OPG180 | homopolymer | Upstream | 15 | Cop-A50R | A50R | |
| LCR17 | 163,979 | 164,003 | 4 | [TAAC]n | OPG188 | STR | Downstream | 90 | Cop-B2R | B4R | |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LCR18 | 166,865 | 166,920 | 7 | [AATAATT]n | OPG190 | STR | Downstream | 18 | Cop-B5R | B6R | |
| LCR19 | 170,508 | 170,563 | 6 | [GATACA]n | OPG197 | STR | Inside ORF | NA | Cop-B11R | B11R | hypothetical protein |
| LCR20 | 172,868 | 172,876 | 1 | [T]n | OPG199 | homopolymer | Downstream | 59 | Cop-K2L | SPI-2/B12R | |

| | |
|---|---|
| a | Nucleotide base coordinate in reference HQG (Accession number XXXX) |
| b | Nucleotide base coordinate in reference HQG (Accession number XXXX) |
| c | Number of repeat units in the HQG (Accession number XXXX) |
| d | Description of the pattern of the STR where n is the number of repeats for this particular genome) |
| e | Identification according to Senkevich et al. of nearest identified gene. New Notation |
| f | Type of LCR: Short tandem repeats or Homopolymer |
| g | Position of the LCR to the neareast gene |
| h | Distance of the LCR to the nearest gene |
| i | Notation of the gene in the VACV Copenhagen strain |
| j | Notation of the gene in the VACV Western Reserve strain |

**Table 1:** Description and annotation of Low Complexity Regions (LCR). Short Tandem Repeats (STRs) are described using nucleotide base pair coordinates in reference to the HQG (Accession number ERS12168861). Number of repeat units, description of the pattern (where n = number of repeats for this particular genome), identification of nearest annotated gene, type of LCR (STR or homopolymer), position of the LCR to the nearest gene, distance of the LCR to the nearest gene, Notation of the gene according to the VACV Copenhagen strain and classical vaccinia notation are listed in the table.

| Name | Repeat Unit | Pattern HQ | Number of repeats HQG | Nearest Gene | Variation | Type of Variation | Entropy threshold >0.03 | Resolved correctly in RefSeq | Nanopore | MiSeq | NovaSeq | # Supporting Reads MiSeq | # Supporting Reads NovaSeq |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LCR4 | 16 | TAGTCATAAGTTAGTT [AAGTCATAAGTTAGTT]15 | 16 | OPG003 (ITR) | NR | Length | NA | No& | Yes | No | No | NA | NA |
| LCR3 | 9 | ATAT [ACATTATAT]52 | 52 | OPG208 | Yes | Length | NA | Yes | Yes | No | No | NA | NA |
| LCR1 | 16 | [AACTAACTTATGACTT]15 AACTAACTTATGACTA | 16 | OPG003 (ITR) | NR | Length | NA | No& | Yes | No | No | NA | NA |
| LCR2 | 2 | [AT]25 | 25 | NA | Yes | Length | 1.66 | No | Yes | Yes | Yes | 768 | 90 |
| LCR5 | 1 | [T]24 | 24 | OPG152 | Yes | Length | 1.535 | Yes | No | No | Yes | NA | 112 |
| LCR10 | 1 | [T]13 | 13 | OPG001 (ITR) | Yes | Length | 0.63 | No& | Yes | No | No | 6561 | 11945 |
| LCR11 | 1 | [T]13 | 13 | OPG001 (ITR) | Yes | Length | 0.627 | No& | Yes | Yes | Yes | 6448 | 11589 |
| LCR21 | 6 | [GATGAA]4 GATGA | 4.5 | OPG204 | Yes | Mutation | 0.207 | Yes | Yes | Yes | Yes | 6578 | 6661 |
| LCR7 | 3 | [ATC]14 TATGAT [ATC]3 | 19 | OPG153 | Yes | Length | 0.181 | Yes | Yes | Yes | Yes | 4541 | 6607 |
| LCR9 | 9 | [TATGAAG]1 [GATATGAT]1 [GATATGATG]5 [GATATGAT]1 | 8 | OPG176 | No | NA | 0 | Yes | Yes | Yes | Yes | 5208 | 5737 |
| LCR8 | 5+7 | [ATATTTT]1 [ATTTT]1 [ATATTTT]3 [ATTTT]1 [ATATTTT]2 [ATTTT]1 [ATATTTT]1 | 10 | OPG171 | No | NA | 0 | Yes | Yes | Yes | Yes | 6581 | 6790 |
| LCR6 | 10 | [CAATCTTTCT]1 | 1 | OPG152 | Yes* | NA | 0 | No* | Yes | Yes | Yes | 4884 | 12930 |
| LCR20 | 1 | [T]9 | 9 | OPG199 | No | NA | 0 | Yes | Yes | Yes | Yes | 10106 | 13315 |
| LCR19 | 6 | GATTCA [GATACA]8 GAT | 9.3 | OPG197 | No | NA | 0 | Yes | Yes | Yes | yes | 4119 | 4685 |
| LCR18 | 7 | [AATAATT]3 AATAA | 3 | OPG190 | No | NA | 0 | Yes | Yes | Yes | Yes | 9755 | 11838 |
| LCR17 | 4 | [TAAC]6 T | 6.1 | OPG188 | No | NA | 0 | Yes | Yes | Yes | Yes | 7388 | 9474 |
| LCR16 | 1 | [A]9 | 9 | OPG180 | No | NA | 0 | Yes | Yes | Yes | Yes | 10340 | 16044 |
| LCR15 | 9 | [ATAACAATT]4 [ATAATTGTT]1 [ATAATAATT]1 [ATAATTGTT]1 | 7 | OPG159 | No | NA | 0 | Yes | Yes | Yes | Yes | 7067 | 6569 |
| LCR14 | 1 | [T]9 | 9 | OPG104 | No | NA | 0 | Yes | Yes | Yes | Yes | 7819 | 12521 |
| LCR13 | 1 | [T]9 | 9 | OPG097/098 | No | NA | 0 | Yes | Yes | Yes | Yes | 7480 | 12126 |
| LCR12 | 1 | [A]9 | 9 | OPG044 | No | NA | 0 | Yes | Yes | Yes | Yes | 9789 | 13592 |
| | | | | | | | | | | | | | |
| * | LCR6 is a 10bp repeat that was reported early in the outbreak as an insertion (reference). Several strains presented this duplication. In our dataset, we have not seen any variation in this area | | | | | | | | | | | | |

| | |
|---|---|
| & | LCR1/4 and LCR10/11 are located in the Inverted Terminal Repeats. Given that no read covering this area reached a unique are outside of the ITR; we cannot technically state that we solved the repeat. Nonethless, the ITR should be identical based on Poxvirus replication mode. |

**Table 2: Low complexity region validation and entropy intra-host analysis in HQG specimen.** The type and number of supporting reads for each of them is shown. **Definitions of quality:** Yes: LCR is found entirely in the assembly in one contig; **Partial:** LCR is found in several different contigs or with Ns spanning the region; **No:** LCR area is not assembled with the reported method. All LCRs with entropy levels above 0.15 are shaded in gray.

| | NC_063383.1 | ON562414.3 | HQ Genome |
|---|---|---|---|
| **Size** | 197209 | 197205 | 198547 |
| **SNPs*** | NA | 67 | 69 |
| **INDELs*** | NA | 10del 7ins | 11del 6 ins |
| **Homopolymeric sites**** | 408 | 405 | 399 |
| **Unique SNPs** | NA | 0 | 2 |
| LCR characterization | | | |
| **LCR1/4** | 8 | 8 | 16 |
| **LCR2** | 22 | 24 | 25 |
| **LCR3** | 18 | 16 | 52 |
| **LCR5** | 25 | 28 | 24 |
| **LCR6** | 2 | 1 | 1 |
| **LCR7** | 19 | 17.6 | 17.6 |
| **LCR8** | 10 | 10 | 10 |
| **LCR9** | 8 | 6 | 6 |
| **LCR10/11** | 17 | 14 | 13 |
| **LCR12** | 9 | 9 | 9 |
| **LCR13** | 9 | 9 | 9 |
| **LCR14** | 9 | 9 | 9 |
| **LCR15** | 7 | 7 | 7 |
| **LCR16** | 9 | 9 | 9 |
| **LCR17** | 6.1 | 6.1 | 6.1 |
| **LCR18** | 3.5 | 3.5 | 3.5 |
| **LCR19** | 9.3 | 9.3 | 9.3 |
| **LCR20** | 9 | 9 | 9 |
| **LCR21** | 4.5 | 4.5 | 4.5 |
| *SNPs and indels vs NC_063383.1 | | | |
| ** Homopolymers with length more than 8 nt | | | |

**Table 3:** Genome comparison of our High Quality Genome against the Clade IIb reference strain (NC_063383.1) and MPXV_USA_2022_MA001 (ON562414.3). Table shows differences in length, number of SNPs, number of INDELs, total number of homopolymeric sites and unique SNPs characterizing that genome. LCR repetitions for each genome are indicated; different number of repeats are shaded in grey.

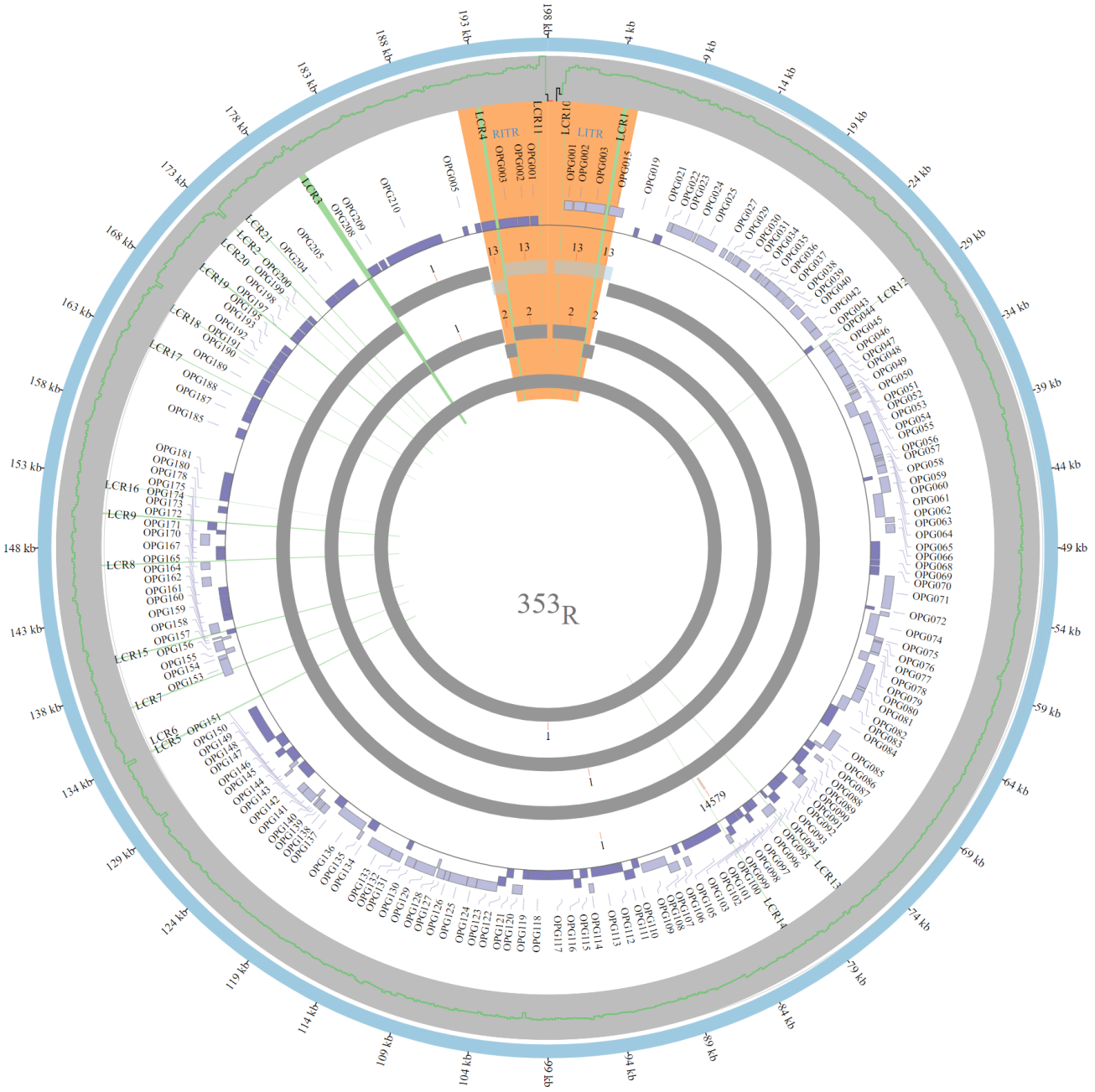



**Figure 1:** High Quality Full genome assembly and annotation. Visual representation of the differences between reference assembled result; de-novo assembled result hybrid and High-Quality genome. Rings outside-inside: 1) High quality genome hybrid assembly; 2) Coverage distribution graph (red meaning coverage 0, 0.2%; orange coverage meaning less than 1000x, 0.1%; black coverage more than 1000x and less than 10000, 0.28%; and green meaning more than 10000x, 99.42%); 3) Gene annotation according to Senkevich et al notation, light purple means gene is in reverse strand, dark purple in the forward; 4) Contigs from NovaSeq sequencing 5) Contigs from MiSeq sequencing 6) Contig from Nanopore sequencing. Green shading indicates the LCR regions; Orange shading highlights the ITR regions.

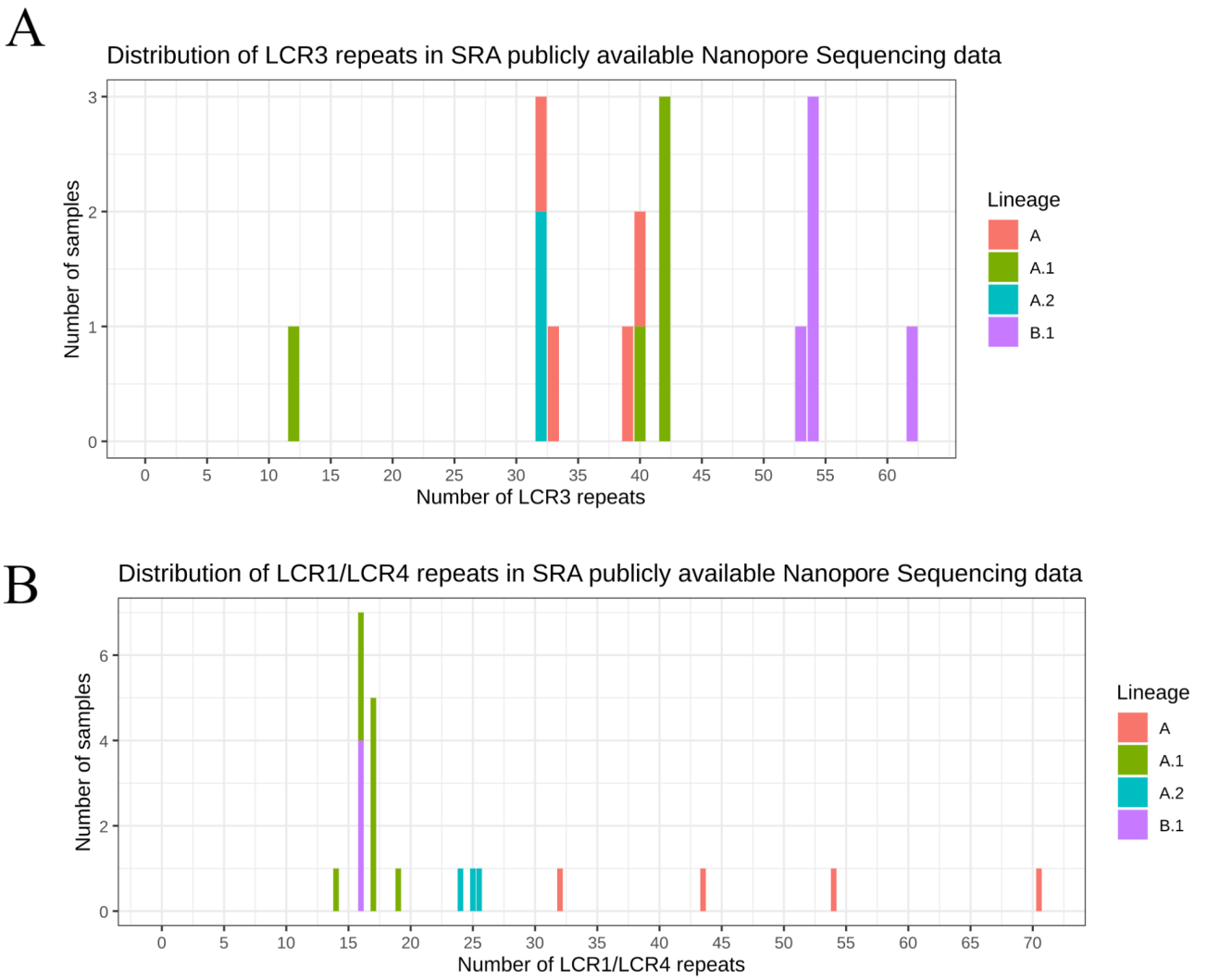



**Figure 2:** Long Low Complexity Region characterization and validation. A) LCR3 validation using nanopore sequencing data from our dataset along with 15 additional raw data sequencing reads available at SRA; B) LCR1/4 validation using nanopore single molecule sequencing data from our dataset along with 20 additional raw data sequencing reads from public databases. Detailed information about the represented materials, along with their originator and epidemiological data is provided in **Supplementary Table 6**.

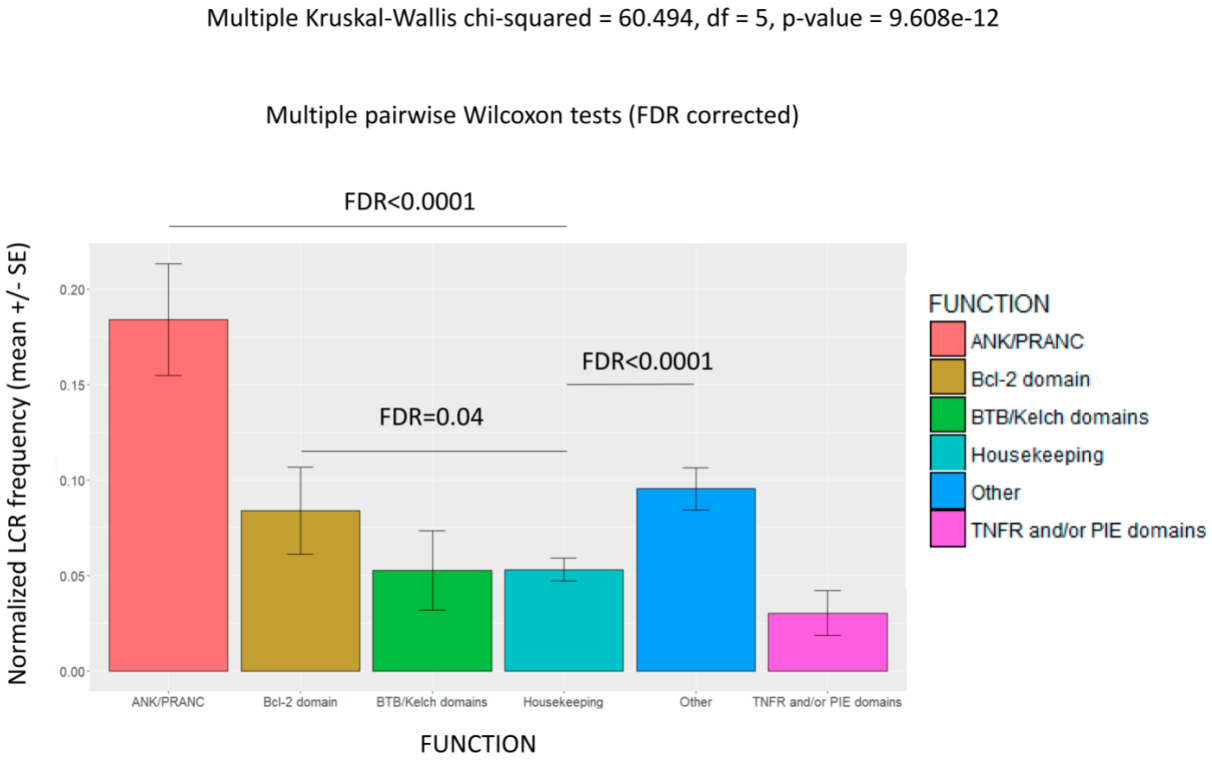



**Figure 3:** Frequencies (mean +/- Standard Error) at which LCRs occur in OPGs from different functional groups in OPXV. Pairwise comparisons where "housekeeping" functional class had a significantly different frequency than other groups (FDR corrected P value <0.05) are highlighted.

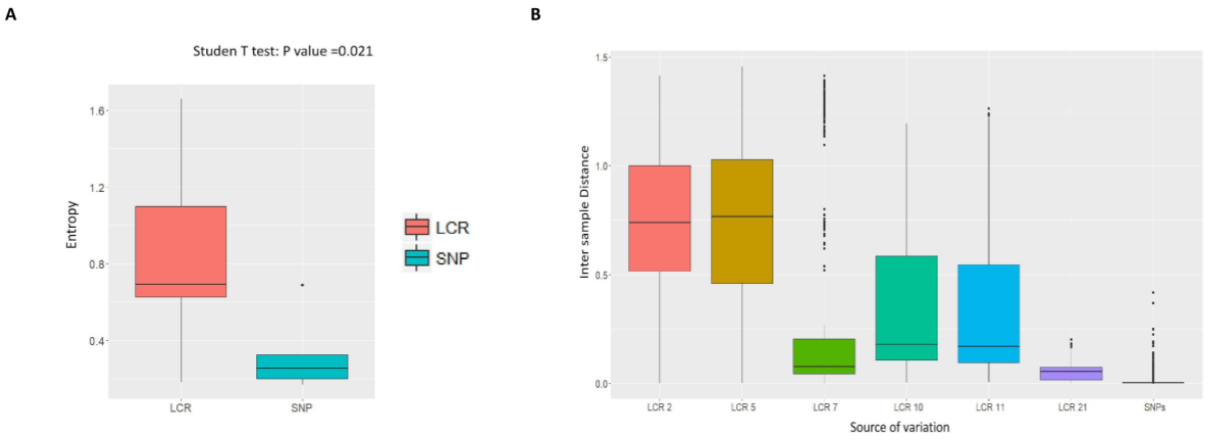



**Figure 4**. *(a)* Distribution of entropy values for the variable sites at SNP (left) and STR (right) sites; *(b)* Distributions of the pairwise inter-sample Euclidean distances for each STR and the SNPs. Note that for SNPs the boxplot represents the distribution of average Euclidean distances of each variable position along the genome.

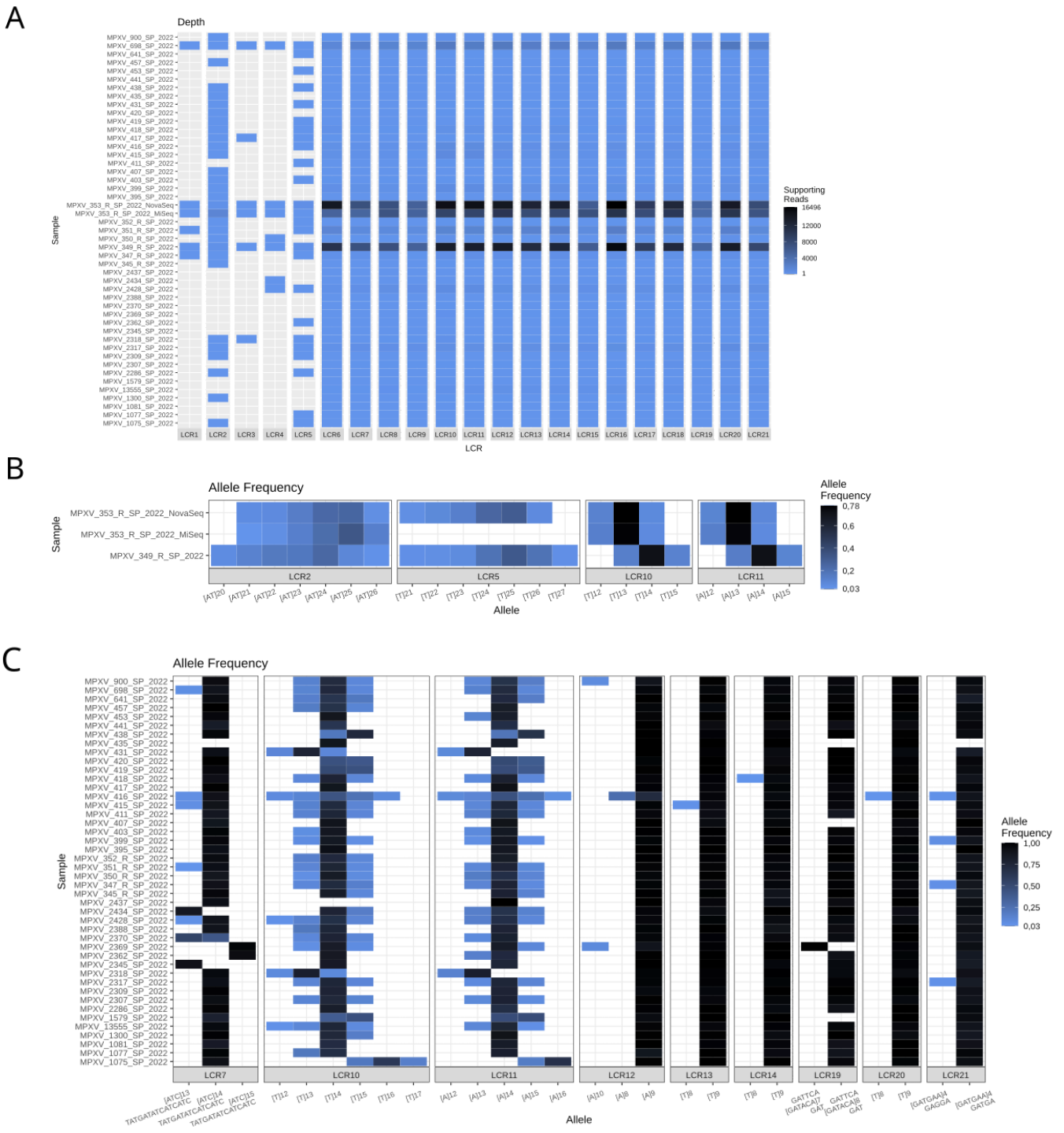
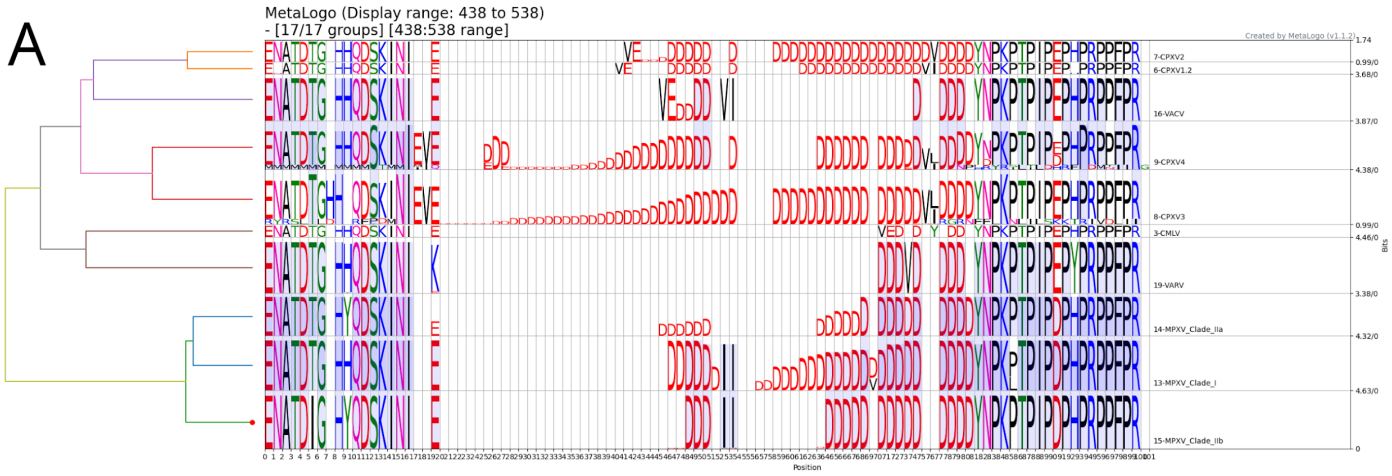



**Figure 5:** Virus Population Genomics analysis of MPXV genomes. LCR variation analysis in the biological specimen (intra-host) and between different specimens (inter-host) variability; a) The panel shading indicates the number of reads supporting each LCR for each sample. Only paired reads that include a perfect match to both flanking regions are counted; the gradient shows the maximum value in black, minimum value (n=1) in light blue. Samples with no coverage are indicated

in grey; b) Comparison of LCR allele frequency for samples 353R and 349R. Only LCR with at least 10 supporting paired reads including both flanking regions are counted; only alleles above a frequency of 0.03 are considered; the gradient show the maximum value in black, with a minimum value n=0.03 in light blue; c) Comparison of LCR allele frequency in all samples for LCR with good coverage (7, 10/11, 12, 13, 14, 19, 20, 21). Only LCR with at least 10 supporting paired reads including both flanking regions are counted; only alleles above a frequency of 0.03 are considered; the gradient shows the maximum value in black, with a minimum value n=0.03 in light blue.



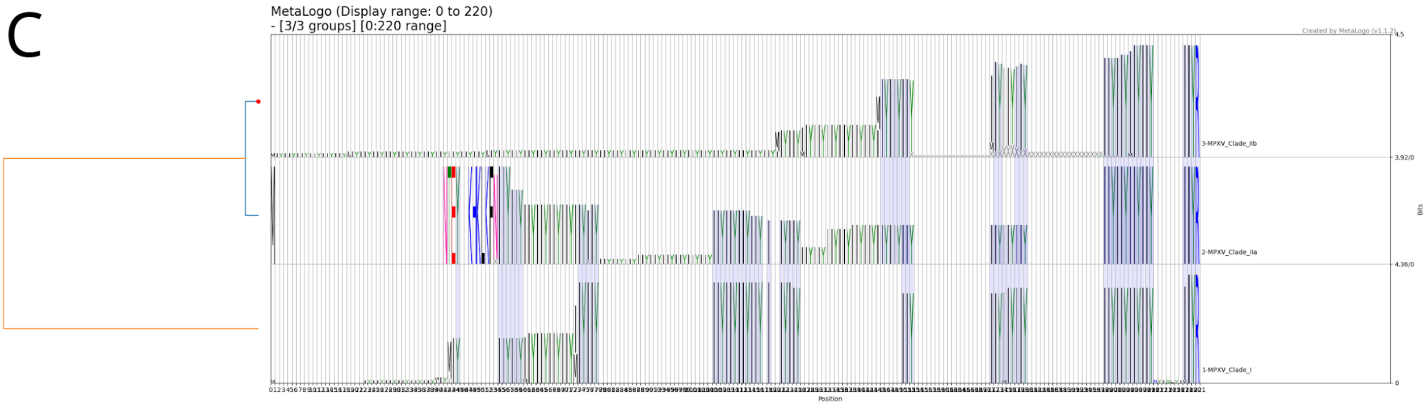



**Figure 6.** Sequence logos are used to visually display conservations and variations in the regions of interest. The logo demonstrates the fixed patterns or conserved motifs in an alignment of OPXV OPG153 protein sequences. The heterogeneity of the alignment represents clades of different evolutionary OPXV origins. Both homologous and nonhomologous sites among cluster of sequences are highlighted; **a)** Analysis of OPG153/LCR7 variability among OPXV genomes. **b)** Analysis of OPG204/LCR21 variability among MPXV Clades I, IIa and IIb; and **c)** Analysis of OPG208/LCR3 variability among MPXV Clades I, IIa and IIb.

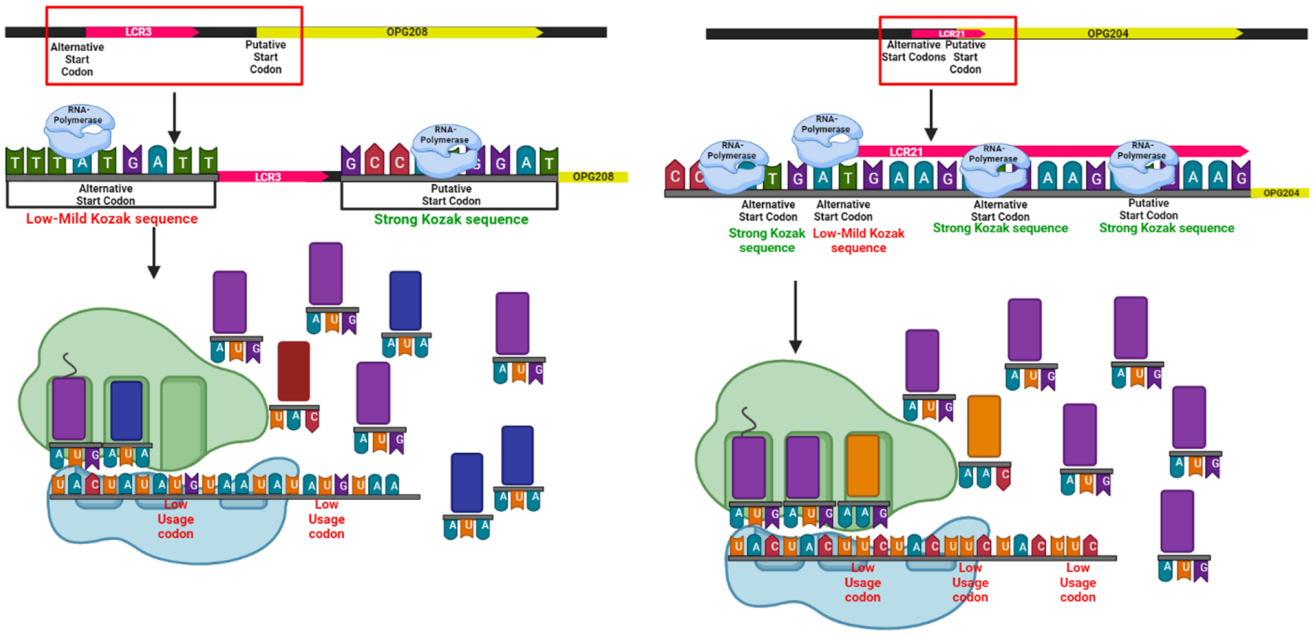



**Figure 7. a)** Analysis of OPG204/LCR21 codon usage among MPXV Clades I, IIa and IIb; and **b)** Analysis of OPG208/LCR3 codon usage among MPXV Clades I, IIa and IIb.

**Supplementary Table 1.** List of samples and genomes used for the different analyses, including sample names, sequence accessions, run accessions, and available epidemiological information (some of the samples are in ena uploading process so they don't have sequence accession yet).

**Supplementary Table 2.** Mapping and assembly stats for the 48 Spanish samples in the study. **Sheet Mapping_HQ**: Mapping stats against the High Quality Genome obtained. **Sheet Mapping_NC_063383.1**: Mapping stats against the NC_063383.1 genome. Mapping stats for both genomes include total reads after trimming, total and percentage of reads corresponding to host genome, total and percentages of reads mapped against viral genome, total and percentage of reads that didn't mapped neither to host nor viral genome, median depth of coverage, percentage of reference viral genome covered to more than 10X depth, number of variants included in the consensus genome, number of those variants thar annotate as missense variants and the number of Ns in the consensus genome after masking. **Sheet Assembly_Spades**: De novo assembly stats using Spades assembler in rnaviral mode. **Sheet Assembly_Unicycler**: De novo assembly stats using Unicycler assembler. Assembly stats for both assemblers include total reads after trimming, total and percentage of reads corresponding to host genome, total and percentage of reads remaining for assembly after host removal, total number of contigs with more than 500 nucleotide length, nucleotides length of the largest contig and the N50 statistical value.

**Supplementary Table 3.** LCRs coordinates against NC_063383.1 (sheet 1) and 353_R HQG (sheet 2) including LCR name, number of repeats, pattern and flanking regions.

**Supplementary Table 4.** STRsearch results in long format including sample name, pattern, number of repeats, number of supporting reads, allele frequency (supporting reads / reads spanning region) and LCR sequence including flanking sequences.

**Supplementary Table 5.** Phylogenetic groups description, groups are described as a monophylogenetic clade with > 80 bootstrap support. The table shows the groups, the samples grouped, and the SNPs describing the groups (unique and shared) and the epidemiology information available.

**Supplementary Table 6.** Results from the LCR1/4 and LCR3 analysis for the different Long Reads Sequencing data available in SRA for MPXV, the table gathers information about number of repeats according to assembly and streveal, and the supporting reads spanning the region.

**Supplementary Figure 1**. Low-complexity-regions characterization and validation. **a)** Multiple alignment for LCR2 showing the differences found according to different reference consensus; **b)** LCR7 alignment showing identification using used sequencing platforms

**Supplementary Figure 2.** Phylogenetic analysis. **a)** Phylogenetic maximum likelihood tree showing SNPs clustering (Clade IIb); bootstrap > 60 is shown, samples are annotated according to its date of collection, and whether they belong to the 2022 outbreak; **b)** Haplotype Network showing SNPs difference among samples included in the phylogenetic tree.

## *Bibliography*

1. Damon IK. Status of human monkeypox: clinical disease, epidemiology and research. Vaccine 2011;29 Suppl 4:D54-9.
2. Hammerschlag Y, MacLeod G, Papadakis G, et al. Monkeypox infection presenting as genital rash, Australia, May 2022. Euro Surveill 2022;27(22).
3. Likos AM, Sammons SA, Olson VA, et al. A tale of two clades: monkeypox viruses. J Gen Virol 2005;86(Pt 10):2661–72.
4. Perez Duque M, Ribeiro S, Martins JV, et al. Ongoing monkeypox virus outbreak, Portugal, 29 April to 23 May 2022. Euro Surveill 2022;27(22).
5. W H O. Monkeypox: experts give virus variants new names [Internet]. WHO; 2022 [cited 2022 Sep 9]. Available from: https://www.who.int/news/item/12-08-2022-monkeypox--experts-give-virus-variants-new-names
6. Antinori A, Mazzotta V, Vita S, et al. Epidemiological, clinical and virological characteristics of four cases of monkeypox support transmission through sexual contact, Italy, May 2022. Euro Surveill 2022;27(22).
7. Vivancos R, Anderson C, Blomquist P, et al. Community transmission of monkeypox in the United Kingdom, April to May 2022. Euro Surveill 2022;27(22).
8. Faye O, Pratt CB, Faye M, et al. Genomic characterisation of human monkeypox virus in Nigeria. Lancet Infect Dis 2018;18(3):246.
9. Vaughan A, Aarons E, Astbury J, et al. Two cases of monkeypox imported to the United Kingdom, September 2018. Euro Surveill 2018;23(38).
10. Cohen-Gihon I, Israeli O, Shifman O, et al. Identification and Whole-Genome Sequencing of a Monkeypox Virus Strain Isolated in Israel. Microbiol Resour Announc 2020;9(10).
11. Ng OT, Lee V, Marimuthu K, et al. A case of imported Monkeypox in Singapore. Lancet Infect Dis 2019;19(11):1166.
12. Yong SEF, Ng OT, Ho ZJM, et al. Imported Monkeypox, Singapore. Emerging Infect Dis 2020;26(8):1826–30.
13. Gigante C. Multiple lineages of Monkeypox virus detected in the United States, 2021- 2022. BioRxiv 2022;
14. Endo A, Murayama H, Abbott S, et al. Heavy-tailed sexual contact networks and the epidemiology of monkeypox outbreak in non-endemic regions, May 2022. medRxiv 2022;
15. Fuller T, Thomassen HA, Mulembakani PM, et al. Using remote sensing to map the risk of human monkeypox virus in the Congo Basin. Ecohealth 2011;8(1):14–25.
16. Khodakevich L, Jezek Z, Kinzanzka K. Isolation of monkeypox virus from wild squirrel infected in nature. Lancet 1986;327(8472):98–9.
17. Khodakevich L, Jezek Z, Messinger D. Monkeypox virus: ecology and public health significance. Bull World Health Organ 1988;66(6):747–52.
18. Khodakevich L, Szczeniowski M, Manbu ma D, et al. The role of squirrels in sustaining monkeypox virus transmission. Trop Geogr Med 1987;39(2):115–22.

19. Khodakevich L, Szczeniowski M, Nambu-ma-Disu, et al. Monkeypox virus in relation to the ecological features surrounding human settlements in Bumba zone, Zaire. Trop Geogr Med 1987;39(1):56–63.

20. Rimoin AW, Mulembakani PM, Johnston SC, et al. Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the Democratic Republic of Congo. Proc Natl Acad Sci U S A 2010;107(37):16262–7.

21. Weaver JR, Isaacs SN. Monkeypox virus and insights into its immunomodulatory proteins. Immunol Rev 2008;225:96–113.

22. Kugelman JR, Johnston SC, Mulembakani PM, et al. Genomic variability of monkeypox virus among humans, Democratic Republic of the Congo. Emerging Infect Dis 2014;20(2):232–9.

23. Liu R, Mendez-Rios JD, Peng C, et al. SPI-1 is a missing host-range factor required for replication of the attenuated modified vaccinia Ankara (MVA) vaccine vector in human cells. PLoS Pathog 2019;15(5):e1007710.

24. McFadden G. Poxvirus tropism. Nat Rev Microbiol 2005;3(3):201–13.

25. Moss B. Poxvirus entry and membrane fusion. Virology 2006;344(1):48–54.

26. Moss B. Membrane fusion during poxvirus entry. Semin Cell Dev Biol 2016;60:89–96.

27. Roberts KL, Smith GL. Vaccinia virus morphogenesis and dissemination. Trends Microbiol 2008;16(10):472–9.

28. Reynolds MG, Yorita KL, Kuehnert MJ, et al. Clinical manifestations of human monkeypox influenced by route of infection. J Infect Dis 2006;194(6):773–80.

29. Chen N, Li G, Liszewski MK, et al. Virulence differences between monkeypox virus isolates from West Africa and the Congo basin. Virology 2005;340(1):46–63.

30. Emerson GL, Li Y, Frace MA, et al. The phylogenetics and ecology of the orthopoxviruses endemic to North America. PLoS ONE 2009;4(10):e7666.

31. Esposito JJ, Sammons SA, Frace AM, et al. Genome sequence diversity and clues to the evolution of variola (smallpox) virus. Science 2006;313(5788):807–12.

32. Moss B. Poxviridae. In: Knipe DM, Howley PM, editors. Field's Virology. Philadelphia: Lippincott-Raven; 2001. p. 2849–83.

33. Werden SJ, Rahman MM, McFadden G. Poxvirus host range genes. Adv Virus Res 2008;71:135–71.

34. Gubser C, Hué S, Kellam P, Smith GL. Poxvirus genomes: a phylogenetic analysis. J Gen Virol 2004;85(Pt 1):105–17.

35. Hendrickson RC, Wang C, Hatcher EL, Lefkowitz EJ. Orthopoxvirus genome evolution: the role of gene loss. Viruses 2010;2(9):1933–67.

36. Senkevich TG, Yutin N, Wolf YI, Koonin EV, Moss B. Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes. MBio 2021;12(4):e0149521.

37. Bratke KA, McLysaght A. Identification of multiple independent horizontal gene transfers into poxviruses using a comparative genomics approach. BMC Evol Biol 2008;8:67.

38. McLysaght A, Baldi PF, Gaut BS. Extensive gene gain associated with adaptive evolution of

poxviruses. Proc Natl Acad Sci USA 2003;100(26):15655–60.

39. Elde NC, Child SJ, Eickbush MT, et al. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell 2012;150(4):831–41.

40. Senkevich TG, Zhivkoplias EK, Weisberg AS, Moss B. Inactivation of genes by frameshift mutations provides rapid adaptation of an attenuated vaccinia virus. J Virol 2020;94(18).

41. Ladner JT, Beitzel B, Chain PSG, et al. Standards for sequencing viral genomes in the era of high-throughput sequencing. MBio 2014;5(3):e01360-14.

42. Fedele CG, Negredo A, Molero F, Sánchez-Seco MP, Tenorio A. Use of internally controlled real-time genome amplification for detection of variola virus and other orthopoxviruses infecting humans. J Clin Microbiol 2006;44(12):4464–70.

43. Sánchez-Seco MP, Hernández L, Eiros JM, Negredo A, Fedele G, Tenorio A. Detection and identification of orthopoxviruses using a generic nested PCR followed by sequencing. Br J Biomed Sci 2006;63(2):79–85.

44. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom 2017;3(10):e000132.

45. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019;37(5):540–6.

46. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 2014;9(11):e112963.

47. Patel H, Varona S, Monzón S, et al. nf-core/viralrecon: nf-core/viralrecon v2.4.1 - Plastered Magnesium Marmoset. Zenodo 2022;

48. Ewels PA, Peltzer A, Fillinger S, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 2020;38(3):276–8.

49. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34(17):i884–90.

50. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20(1):257.

51. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 2016;32(7):1009–15.

52. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de novo assembler. Curr Protoc Bioinformatics 2020;70(1):e102.

53. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 2009;25(15):1968–9.

54. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinformatics 2017;bbx108.

55. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009;25(9):1189–91.

56. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27(2):573–80.

57. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods

2012;9(4):357–9.

58. Picard toolkit. 2019;

59. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078–9.

60. Grubaugh ND, Gangavarapu K, Quick J, et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol 2019;20(1):8.

61. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience 2021;10(2).

62. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 2012;6(2):80–92.

63. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet 2012;3:35.

64. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32(19):3047–8.

65. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. JOSS 2021;6(67):3773.

66. Seemann T. snippy: fast bacterial variant calling from NGS reads. 2015;

67. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 2020;37(5):1530–4.

68. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 2021;49(W1):W293–6.

69. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999;16(1):37–48.

70. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Bioinformatics 2020;

71. Dainat J, Hereñú D, LucileSol, Pascal-Git. NBISweden/AGAT: AGAT-v0.8.1. Zenodo 2022;

72. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5:113.

73. Chen Y, He Z, Men Y, Dong G, Hu S, Ying X. MetaLogo: a heterogeneity-aware sequence logo generator and aligner. Brief Bioinformatics 2022;23(2).

74. Wittek R, Moss B. Tandem repeats within the inverted terminal repetition of vaccinia virus DNA. Cell 1980;21(1):277–84.

75. Phillips C, Gettings KB, King JL, et al. “The devil’s in the detail”: Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. Forensic Sci Int Genet 2018;34:162–9.

76. Kettle S, Blake NW, Law KM, Smith GL. Vaccinia virus serpins B13R (SPI-2) and B22R (SPI-1) encode M(r) 38.5 and 40K, intracellular polypeptides that do not affect virus virulence in a murine intranasal model. Virology 1995;206(1):136–47.

77. Jorgensen I, Rayamajhi M, Miao EA. Programmed cell death as a defence against infection. Nat Rev Immunol 2017;17(3):151–64.

78. Bunge EM, Hoet B, Chen L, et al. The changing epidemiology of human monkeypox-A potential threat? A systematic review. PLoS Negl Trop Dis 2022;16(2):e0010141.

79. Learned LA, Reynolds MG, Wassa DW, et al. Extended interhuman transmission of monkeypox in a hospital community in the Republic of the Congo, 2003. Am J Trop Med Hyg 2005;73(2):428–34.

80. Ghazvini K, Keikha M. Human Monkeypox resurgence 2022; a new presentation as a sexual pathogen. Ann Med Surg (Lond) 2022;80:104267.

81. Guzzetta G, Mammone A, Ferraro F, et al. Early Estimates of Monkeypox Incubation Period, Generation Time, and Reproduction Number, Italy, May-June 2022. Emerging Infect Dis 2022;28(10).

82. Mauldin MR, McCollum AM, Nakazawa YJ, et al. Exportation of monkeypox virus from the african continent. J Infect Dis 2022;225(8):1367–76.

83. Sepehrinezhad A, Ashayeri Ahmadabad R, Sahab-Negah S. Monkeypox virus from neurological complications to neuroinvasive properties: current status and future perspectives. J Neurol 2022;

84. Thornhill JP, Barkati S, Walmsley S, et al. Monkeypox Virus Infection in Humans across 16 Countries - April-June 2022. N Engl J Med 2022;387(8):679–91.

85. Torster L, Tegtmeyer J, Kött J, Christolouka M, Schneider SW. Localized monkeypox infestation in MSM on pre-exposure prophylaxis. J Eur Acad Dermatol Venereol 2022;

86. Vusirikala A, Charles H, Balasegaram S, et al. Epidemiology of early monkeypox virus transmission in sexual networks of gay and bisexual men, england, 2022. Emerging Infect Dis 2022;28(10).

87. Zambrano PG, Acosta-España JD, Mosquera Moyano F, Altamirano Jara JB. Sexually or intimately transmitted infections: A look at the current outbreak of monkeypox in 2022. Travel Med Infect Dis 2022;49:102383.

88. V M, A M, F C, et al. Ocular involvement in monkeypox: description of an unusual presentation during the current outbreak. J Infect 2022;

89. Ulaeto DO, Dunning J, Carroll MW. Evolutionary implications of human transmission of monkeypox: the importance of sequencing multiple lesions. Lancet Microbe 2022;

90. Baroudy BM, Moss B. Sequence homologies of diverse length tandem repetitions near ends of vaccinia virus genome suggest unequal crossing over. Nucleic Acids Res 1982;10(18):5673–9.

91. Shchelkunov SN. Orthopoxvirus genes that mediate disease virulence and host tropism. Adv Virol 2012;2012:524743.

92. Shchelkunov SN, Totmenin AV, Babkin IV, et al. Human monkeypox and smallpox viruses: genomic comparison. FEBS Lett 2001;509(1):66–70.

93. Isidro J, Borges V, Pinto M, et al. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. Nat Med 2022;28(8):1569–72.

94. Rambaut A, O'Tool A. Initial observations about putative APOBEC3 deaminase editing

driving short-term evolution of MPXV since 2017 - Evolution - Virological [Internet]. Virological.org. 2022 [cited 2022 Sep 11];Available from: https://virological.org/t/initial-observations-about-putative-apobec3-deaminase-editing-driving-short-term-evolution-of-mpxv-since-2017/830

95. Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. Nat Genet 2005;37(9):986–90.

96. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121–3.

97. Davies HM, Nofal SD, McLaughlin EJ, Osborne AR. Repetitive sequences in malaria parasite proteins. FEMS Microbiol Rev 2017;41(6):923–40.

98. Oma Y, Kino Y, Sasagawa N, Ishiura S. Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. J Biol Chem 2004;279(20):21217–22.

99. Gatchel JR, Zoghbi HY. Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet 2005;6(10):743–55.

100. Shoubridge C, Cloosterman D, Parkinson-Lawerence E, Brooks D, Gécz J. Molecular pathology of expanded polyalanine tract mutations in the Aristaless-related homeobox gene. Genomics 2007;90(1):59–71.

101. Gemayel R, Chavali S, Pougach K, et al. Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. Mol Cell 2015;59(4):615–27.

102. Salichs E, Ledda A, Mularoni L, Albà MM, de la Luna S. Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. PLoS Genet 2009;5(3):e1000397.

103. Fidalgo M, Barrales RR, Ibeas JI, Jimenez J. Adaptive evolution by mutations in the FLO11 gene. Proc Natl Acad Sci USA 2006;103(30):11228–33.

104. Davies HM, Thalassinos K, Osborne AR. Expansion of Lysine-rich Repeats in Plasmodium Proteins Generates Novel Localization Sequences That Target the Periphery of the Host Erythrocyte. J Biol Chem 2016;291(50):26188–207.

105. Maier AG, Rug M, O'Neill MT, et al. Exported proteins required for virulence and rigidity of Plasmodium falciparum-infected human erythrocytes. Cell 2008;134(1):48–61.

106. Kotwal GJ, Moss B. Vaccinia virus encodes two proteins that are structurally related to members of the plasma serine protease inhibitor superfamily. J Virol 1989;63(2):600–6.

107. Brooks MA, Ali AN, Turner PC, Moyer RW. A rabbitpox virus serpin gene controls host range by inhibiting apoptosis in restrictive cells. J Virol 1995;69(12):7688–98.

108. Ali AN, Turner PC, Brooks MA, Moyer RW. The SPI-1 gene of rabbitpox virus determines host range and is required for hemorrhagic pock formation. Virology 1994;202(1):305–14.

109. Delaloye J, Filali-Mouhim A, Cameron MJ, et al. Interleukin-1- and type I interferon-dependent enhanced immunogenicity of an NYVAC-HIV-1 Env-Gag-Pol-Nef vaccine vector with dual deletions of type I and type II interferon-binding proteins. J Virol 2015;89(7):3819–32.

110. García-Arriaza J, Perdiguero B, Heeney JL, et al. HIV/AIDS Vaccine Candidates Based on Replication-Competent Recombinant Poxvirus NYVAC-C-KC Expressing Trimeric gp140 and

Gag-Derived Virus-Like Particles or Lacking the Viral Molecule B19 That Inhibits Type I Interferon Activate Relevant HIV-1-Specific B and T Cell Immune Functions in Nonhuman Primates. J Virol 2017;91(9).

111. Gómez CE, Perdiguero B, Nájera JL, et al. Removal of vaccinia virus genes that block interferon type I and II pathways improves adaptive and memory responses of the HIV/AIDS vaccine candidate NYVAC-C in mice. J Virol 2012;86(9):5026–38.

112. Hernáez B, Alonso-Lobo JM, Montanuy I, et al. A virus-encoded type I interferon decoy receptor enables evasion of host immunity through cell-surface binding. Nat Commun 2018;9(1):5440.

113. Ueda Y, Morikawa S, Matsuura Y. Identification and nucleotide sequence of the gene encoding a surface antigen induced by vaccinia virus. Virology 1990;177(2):588–94.

114. Alcamí A, Symons JA, Smith GL. The vaccinia virus soluble alpha/beta interferon (IFN) receptor binds to the cell surface and protects cells from the antiviral effects of IFN. J Virol 2000;74(23):11230–9.

115. Colamonici OR, Domanski P, Sweitzer SM, Larner A, Buller RM. Vaccinia virus B18R gene encodes a type I interferon-binding protein that blocks interferon alpha transmembrane signaling. J Biol Chem 1995;270(27):15974–8.

116. Ulaeto D, Grosenbach D, Hruby DE. The vaccinia virus 4c and A-type inclusion proteins are specific markers for the intracellular mature virus particle. J Virol 1996;70(6):3372–7.

117. Chiu W-L, Lin C-L, Yang M-H, Tzou D-LM, Chang W. Vaccinia virus 4c (A26L) protein on intracellular mature virus binds to the extracellular cellular matrix laminin. J Virol 2007;81(5):2149–57.

118. Howard AR, Senkevich TG, Moss B. Vaccinia virus A26 and A27 proteins form a stable complex tethered to mature virions by association with the A17 transmembrane protein. J Virol 2008;82(24):12384–91.

119. Kastenmayer RJ, Maruri-Avidal L, Americo JL, Earl PL, Weisberg AS, Moss B. Elimination of A-type inclusion formation enhances cowpox virus replication in mice: implications for orthopoxvirus evolution. Virology 2014;452–453:59–66.

120. Liu L, Cooper T, Howley PM, Hayball JD. From crescent to mature virion: vaccinia virus assembly and maturation. Viruses 2014;6(10):3787–808.

121. Keasey S, Pugh C, Tikhonov A, et al. Proteomic basis of the antibody response to monkeypox virus infection examined in cynomolgus macaques and a comparison to human smallpox vaccination. PLoS ONE 2010;5(12):e15547.

122. Pugh C, Brown ES, Quinn X, et al. Povidone iodine ointment application to the vaccination site does not alter immunoglobulin G antibody response to smallpox vaccine. Viral Immunol 2016;29(6):361–6.

123. Boulter EA, Appleyard G. Differences between extracellular and intracellular forms of poxvirus and their implications. Prog Med Virol 1973;16:86–108.

124. McKelvey TA, Andrews SC, Miller SE, Ray CA, Pickup DJ. Identification of the orthopoxvirus p4c gene, which encodes a structural protein that directs intracellular mature virus particles into A-type inclusions. J Virol 2002;76(22):11216–25.

125. Okeke MI, Hansen H, Traavik T. A naturally occurring cowpox virus with an ectromelia virus

A-type inclusion protein gene displays atypical A-type inclusions. Infect Genet Evol 2012;12(1):160–8.

126. Duraffour S, Lorenzo MM, Zöller G, et al. ST-246 is a key antiviral to inhibit the viral F13L phospholipase, one of the essential proteins for orthopoxvirus wrapping. J Antimicrob Chemother 2015;70(5):1367–80.

127. Henry SP, Takanosu M, Boyd TC, et al. Expression pattern and gene characterization of asporin. a newly discovered member of the leucine-rich repeat protein family. J Biol Chem 2001;276(15):12212–21.

128. Zhu X, Jiang L, Lu Y, et al. Association of aspartic acid repeat polymorphism in the asporin gene with osteoarthritis of knee, hip, and hand: A PRISMA-compliant meta-analysis. Medicine (Baltimore) 2018;97(12):e0200.