

# Isolation and genome sequencing of cytomegaloviruses from Natal multimammate mice (*Mastomys natalensis*)

Frederick Hansen<sup>1,†</sup>, Matej Vučak<sup>2</sup>, Jenna Nichols<sup>2</sup>, Joseph Hughes<sup>2</sup>, Sidy Bane<sup>3</sup>, Salvatore Camiolo<sup>2,§</sup>, Ana da Silva Filipe<sup>2</sup>, Eleonore Ostermann<sup>4</sup>, Laura Staliunaite<sup>4</sup>, Baca Chan<sup>5,6</sup>, Thekla Mauch<sup>7</sup>, Nafomon Sogoba<sup>3</sup>, Daniel N. Streblow<sup>8</sup>, Sebastian Voigt<sup>9</sup>, Lisa Oestereich<sup>10</sup>, Bernhard Ehlers<sup>11</sup>, Alec J. Redwood<sup>5,6</sup>, Heinz Feldmann<sup>1</sup>, Wolfram Brune<sup>4</sup>, Kyle Rosenke<sup>1,†</sup>, Michael A. Jarvis<sup>1,7,12,\*†</sup> and Andrew J. Davison<sup>2,†</sup>

## Abstract

Distinct cytomegaloviruses (CMVs) are widely distributed across their mammalian hosts in a highly host species-restricted pattern. To date, evidence demonstrating this has been limited largely to PCR-based approaches targeting small, conserved genomic regions, and only a few complete genomes of isolated viruses representing distinct CMV species have been sequenced. We have now combined direct isolation of infectious viruses from tissues with complete genome sequencing to provide a view of CMV diversity in a wild animal population. We targeted Natal multimammate mice (*Mastomys natalensis*), which are common in sub-Saharan Africa, are known to carry a variety of zoonotic pathogens, and are regarded as the primary source of Lassa virus (LASV) spillover into humans. Using transformed epithelial cells prepared from *M. natalensis* kidneys, we isolated CMVs from the salivary gland tissue of 14 of 37 (36%) animals from a field study site in Mali. Genome sequencing showed that these primary isolates represent three different *M. natalensis* CMVs (MnatCMVs: MnatCMV1, MnatCMV2 and MnatCMV3), with some animals carrying multiple MnatCMVs or multiple strains of a single MnatCMV presumably as a result of coinfection or superinfection. Including primary isolates and plaque-purified isolates, we sequenced and annotated the genomes of two MnatCMV1 strains (derived from sequencing 14 viruses), six MnatCMV2 strains (25 viruses) and ten MnatCMV3 strains (21 viruses), totalling 18 MnatCMV strains isolated as 60 infectious viruses. Phylogenetic analysis showed that these MnatCMVs group with other murid viruses in the genus *Muromegalovirus* (subfamily *Betaherpesvirinae*, family *Orthoherpesviridae*), and that MnatCMV1 and MnatCMV2 are more closely related to each other than to MnatCMV3. The availability of MnatCMV isolates and the characterization of their genomes will serve as the prelude to the generation of a MnatCMV-based vaccine to target LASV in the *M. natalensis* reservoir.

Received 09 May 2023; Accepted 02 August 2023; Published 29 August 2023

**Author affiliations:** <sup>1</sup>Laboratory of Virology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, MT, USA; <sup>2</sup>MRC-University of Glasgow Centre for Virus Research, Glasgow, UK; <sup>3</sup>University of Sciences, Techniques and Technologies of Bamako, Bamako, Mali; <sup>4</sup>Leibniz Institute of Virology, Hamburg, Germany; <sup>5</sup>School of Biomedical Sciences, University of Western Australia, Crawley, WA, Australia; <sup>6</sup>Institute for Respiratory Health, University of Western Australia, Crawley, WA, Australia; <sup>7</sup>The Vaccine Group Ltd, Plymouth, Devon, UK; <sup>8</sup>Vaccine and Gene Therapy Institute and Oregon National Primate Research Center, Oregon Health & Science University, Beaverton, OR, USA; <sup>9</sup>Institute for Virology, University Hospital Essen, University of Duisburg-Essen, Essen, Germany; <sup>10</sup>Bernhard Nocht Institute for Tropical Medicine and German Center for Infectious Research (DZIF), Partner Sites Hamburg-Lübeck-Borstel-Riems, Hamburg, Germany, Germany; <sup>11</sup>Division 12, Measles, Mumps, Rubella and Viruses Affecting Immunocompromised Patients, Robert Koch Institute, Berlin, Germany; <sup>12</sup>School of Biomedical Sciences, University of Plymouth, Plymouth, UK.

**\*Correspondence:** Michael A. Jarvis, michael.jarvis@plymouth.ac.uk

**Keywords:** cytomegalovirus; genome sequence; herpesvirus; *Mastomys natalensis*; *Mastomys natalensis* cytomegalovirus; muromegalovirus.

**Abbreviations:** CMV, cytomegalovirus; CPE, cytopathic effect; DMEM, Dulbecco's Modified Eagle's Medium; EEHV1, elephant endotheliotropic herpesvirus 1; GPCMV, guinea pig cytomegalovirus; HCMV, human cytomegalovirus; HHV6A, human herpesvirus 6A; LASV, Lassa virus; MasKECs, *Mastomys* kidney epithelial cells; MCMV, murine cytomegalovirus; MnatCMV1, *Mastomys natalensis* cytomegalovirus 1; MnatCMV2, *Mastomys natalensis* cytomegalovirus 2; MnatCMV3, *Mastomys natalensis* cytomegalovirus 3; MnatCMV4, *Mastomys natalensis* cytomegalovirus 4; MnatCMV, *Mastomys natalensis* cytomegalovirus; RCMV-E, rat cytomegalovirus England; RCMV-M, rat cytomegalovirus Maastricht.

**†Present address:** School of Medicine, University of New Mexico, Albuquerque, NM, USA

**§Present address:** BioSpyder Technologies Inc., Carlsbad, CA, USA.

The sequence read datasets and consensus genome sequences generated in this study are available in NCBI BioProject PRJNA882666.

†These authors contributed equally to this work

One supplementary figure and four supplementary tables are available with the online version of this article.

001873



This is an open-access article distributed under the terms of the Creative Commons Attribution NonCommercial License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

## DATA SUMMARY

The authors confirm that all supporting data, code and protocols have been provided within the article or through supplementary data files.

## INTRODUCTION

Members of the family *Orthoherpesviridae* have large double-stranded genomes (124–236 kb) and are divided into the subfamilies *Alphaherpesvirinae*, *Betaherpesvirinae* and *Gammapherpesvirinae* [1]. The subfamily *Betaherpesvirinae* is divided further into five genera (*Cytomegalovirus*, *Muromegalovirus*, *Proboscivirus*, *Quwivirus* and *Roselovirus*). Members of the genus *Muromegalovirus* infect murids (mice, including rats) and, along with related viruses that have not yet been classified, are known as muromegaloviruses. However, for historical reasons, individual muromegaloviruses bear the name cytomegalovirus (CMV) along with members of the sister genus *Cytomegalovirus*, which infect primates. The currently classified muromegaloviruses are murine CMV (MCMV; species *Muromegalovirus muridbeta1*), rat CMV England (RCMV-E; species *Muromegalovirus muridbeta 8*) and rat CMV Maastricht (RCMV-M; species *Muromegalovirus muridbeta 2*). The muromegaloviruses for which complete genome sequences are available belong only to these species. Murids also host members of the subfamily *Gammapherpesvirinae* in the genus *Rhadinovirus*, but are not known to host members of the subfamily *Alphaherpesvirinae*.

In a study of lung and spleen tissues from *M. natalensis* animals from field study sites in Côte d'Ivoire and Mali [2], four different CMVs (MnatCMV1, MnatCMV2, MnatCMV3 and MnatCMV4) were detected by sequencing PCR products generated from short (<300 bp) regions of the virus DNA polymerase and glycoprotein B genes by using nested, degenerate primers [3]. The MnatCMV3 sequences from two animals were extended further by PCR into larger sequences (approximately 3.4 kb) containing substantial parts of these neighbouring genes. Phylogenetic analyses using these subgenomic sequences showed that MnatCMV3, and probably the other MnatCMVs, are muromegaloviruses, with MnatCMV1 and MnatCMV2 being the most closely related of the four.

Our research has three sequential aims: first, to isolate infectious MnatCMVs from the tissues of *M. natalensis* animals from one of the field study sites in Mali; second, to determine the genome sequences of these viruses in order to provide a detailed picture of the MnatCMV landscape in this wild rodent population; and third, to construct fully characterized infectious wild-type MnatCMV clones as the prelude to the generation of a MnatCMV-based vaccine to target Lassa virus (LASV) in the *M. natalensis* reservoir. The current report focuses on the first two aims.

## METHODS

### Sequence data

Short (178 bp) MnatCMV DNA polymerase gene sequences derived from *M. natalensis* animals captured in Mali [2] provided the means for identifying the MnatCMVs that were isolated and sequenced in our study. With the exception of the extended MnatCMV3 sequences, these sequences were too short for deposition in GenBank. In addition, the genome sequences of four other muromegaloviruses were used for identification purposes. These viruses were MCMV strain Smith (MCMV Smith), MCMV strain s09\_sk3086 (MCMV s09), RCMV-E and RCMV-M (Table 1; references are included). The sequence of MCMV s09 was used as published, but the other three sequences were determined afresh using DNA prepared within our study.

### Generation of cell lines for isolating MnatCMVs

*M. natalensis* animals from a maintained colony were anaesthetized with isoflurane and killed by cervical dislocation (see Ethical and Biosafety Statement). The kidneys were collected and decapsulated, and the medulla was removed from each. The cortical tissue was minced, transferred into 10 ml of Dulbecco's modified Eagle's medium/nutrient mixture F-12 Ham

**Table 1.** Muromegalovirus genome sequences

Virus	Host animal	Host species	Reference*	GenBank accession†	Reference†	GenBank accession‡
MCMV Smith	House mouse	<i>Mus musculus</i>	[49]	U68299.1	[34]	OP429142.1
MCMV s09	Steppe mouse	<i>Mus spicilegus</i>	[41]	MH118555.1	[41]	None
RCMV-E	Brown rat	<i>Rattus norvegicus</i>	[50]	JX867617.2	[51]	OP429143.1
RCMV-M	Brown rat	<i>Rattus norvegicus</i>	[52]	AF232689.2	[53]	OP429144.1

\*First description of virus.

†First description of genome sequence.

‡Newly determined genome sequence.

(DMEM-F12; Merck) containing 1 mg ml<sup>-1</sup> collagenase II (Merck), and incubated at 37°C for 20 min with vigorous vortexing every 10 min. The cell suspension was washed with Hanks' balanced salt solution (Merck) containing 0.2 µM 4-(2-hydroxyethyl)-1-piperazine ethanesulfonic acid, 0.45 µg ml<sup>-1</sup> NaHCO<sub>3</sub>, 80 ng ml<sup>-1</sup> NaOH, 50 µg ml<sup>-1</sup> gentamycin and 2% (v/v) FCS. The cells were dissociated by passing the suspension through a coarse metal sieve, then through a 70 µm cell sieve, and finally twice through a 40 µm cell sieve. The dissociated cell suspension was collected and centrifuged at 150 g for 10 min. The cell pellet was resuspended in DMEM/F-12 containing 1× insulin-transferrin-selenium-ethanolamine (Thermo Fisher Scientific), 2% (v/v) FCS, 100 U ml<sup>-1</sup> penicillin, 100 µg ml<sup>-1</sup> streptomycin, 40 ng ml<sup>-1</sup> hydrocortisone (Merck) and 0.25 nM 3,3',5-triiodo-L-thyronine (Merck), seeded on 1% (w/v) gelatin-coated plates, and incubated at 37°C in an atmosphere of 5% (v/v) CO<sub>2</sub>. The medium was replaced 24 h later and subsequently every 48–72 h. When confluent, the cells were immortalized by retrovirus transduction as described previously [4], and cultured in DMEM containing 10% (v/v) FCS, 100 U ml<sup>-1</sup> penicillin and 100 µg ml<sup>-1</sup> streptomycin. To eliminate fibroblasts, the cells were cultured for >5 passages in MEM Eagle D-valine modification w/L-glutamine, D-valine (Biomol) containing 3.7 mg ml<sup>-1</sup> NaHCO<sub>3</sub>, 1% (w/v) non-essential amino acids, 10% (v/v) dialysed FCS, 100 U ml<sup>-1</sup> penicillin and 100 µg ml<sup>-1</sup> streptomycin. The resulting cells were referred to as *Mastomys* kidney epithelial cells (MasKECs) [5] and used for MnatCMV isolation.

### MnatCMV isolation

Our study used salivary gland tissue samples collected from a cohort of 37 LASV-negative *M. natalensis* animals that were captured in southern Mali (in the village of Donéguébougou); these were the same animals from which initial MnatCMV sequence data were obtained previously [2]. The tissues were confirmed as LASV-negative using a PCR assay, and stored at –80°C (see Ethical and Biosafety Statement).

The tissues were thawed and homogenized by bead beating, and a 1:500 (v/v) dilution of each homogenate was made using DMEM containing 2% (v/v) FCS supplemented with 1 mM L-glutamine, 50 U ml<sup>-1</sup> penicillin and 50 µg ml<sup>-1</sup> streptomycin (DMEM-2). The medium from MasKECs at 70–80% confluency on a six-well plate was replaced with 500 µl of the diluted homogenate. After rocking the plates gently for 1 h at 37°C in an atmosphere of 5% (v/v) CO<sub>2</sub>, the inoculum was replaced with 2 ml of DMEM-2. The cultures were fed at 1 day post-infection and every other day thereafter by replacing 1 ml of medium with 1 ml of fresh DMEM-2. The monolayers were monitored daily and harvested when approximately 80% of cells showed cytopathic effect (CPE), usually at 11–40 days post-infection. Cells were harvested using a cell scraper, and 1 ml aliquots of suspended cells were transferred to two 2 ml screw-top tubes. The tubes were centrifuged at 750 g for 5 min, and the supernatant was transferred to fresh tubes. The remaining cell pellets were resuspended by vortexing, freeze-thawed three times, added to the tubes containing supernatant, aliquoted and stored at –80°C. These stocks were termed *primary isolates*.

The primary isolates were then plaque-purified. When primary isolates were made on multiple occasions from the same animal, the first isolate was used for plaque purification. The medium from MasKECs at 70–80% confluency on six-well plates was replaced with 500 µl of primary isolate diluted in DMEM-2 (1:10, 1:100 and 1:1000, v/v). After rocking the plate gently for 1 h, the inoculum was replaced with 2 ml of 0.6% (w/v) agarose in DMEM-2 at 37°C, which was allowed to solidify at room temperature for 20 min. The cultures were incubated at 37°C in an atmosphere of 5% (v/v) CO<sub>2</sub> and fed every 4 days by adding 1 ml of 0.6% (w/v) agarose in DMEM-2 until distinct plaques were observed. Plaques were picked using 1 ml pipette tips, added to 600 µl of 2% (v/v) DMEM-2 and vortexed vigorously. A dilution (1:10, v/v) was then used to infect cells and repeat plaque purification twice more. Aliquots of 500 µl of suspended plaques that had been purified once or three times were used to infect 25 cm<sup>2</sup> flasks of MasKECs at 70–80% confluency in 4 ml of DMEM-2. The cells were fed every 2 days by replacing 2 ml of medium with fresh DMEM-2 until they exhibited approximately 80% CPE. The infected cells were then harvested and processed into virus stocks as described above, omitting the freeze-thaw steps. These stocks were termed *primary plaque isolates* (plaque-purified once) and *tertiary plaque isolates* (plaque-purified three times). DNA and RNA were extracted from aliquots of the primary isolates and plaque isolates using an AllPrep DNA/RNA mini kit (QIAGEN).

### DNA sequencing

DNA sequence data were generated and analysed. For short-read sequencing of DNA from cells infected with MnatCMVs or other muromegaloviruses (MCMV Smith, RCMV-E and RCMV-M), approximately 100 ng of DNA was sheared in an S220 focused-ultrasonicator (Covaris) to fragments of approximately 450 bp. Sequencing libraries were produced by conducting seven PCR cycles using a KAPA LTP library preparation kit (Roche Sequencing and Life Science) with NEBNext multiplex oligos for Illumina (New England Biolabs). The libraries were analysed on NextSeq and MiSeq instruments (Illumina) to generate datasets that, for the MnatCMV isolates, consisted of 4951808–45770652 paired-end reads of 150 or 300 nt.

For short-read sequencing of RNA from cells infected with MnatCMVs, sequencing libraries were produced from 500 ng of RNA using a TruSeq stranded mRNA library prep kit (Illumina) with TruSeq RNA CD indexes (Illumina). The TruSeq stranded mRNA protocol was followed, with the exception that 12 cycles of PCR were performed. The sequencing libraries were pooled and sequenced on a NextSeq instrument (Illumina) to generate datasets consisting of 29937270–31404526 paired-end reads of 150 nt.

For long-read sequencing of DNA from cells infected with MnatCMVs, sequencing libraries were produced using a ligation sequencing kit [Oxford Nanopore Technologies (ONT)] and a native barcoding kit (ONT) to permit multiplexing. The libraries were pooled and sequenced for 48 h on a MinION (R9.4.1) flow cell (ONT) operated on a GridION Mk1 (ONT). The majority of MnatCMV reads were approximately 5000 nt, and the maximum was 48000 nt.

### Determination of MnatCMV genome sequences

Programs were used with default parameters unless stated otherwise. *De novo* assembly of short-read datasets generated from DNA from primary isolates was initiated by trimming and quality-filtering the datasets using Trim Galore v0.4.0 ([https://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)) with parameters --quality 25 --length 50 --paired. The trimmed datasets were deduplicated using FastUniq v1.1 [6], and the deduplicated reads were assembled using SPAdes v3.14.1 [7] with parameters --careful --cov-cutoff auto --k 33,77,127. The assembled contigs were aligned to the NCBI RefSeq protein database (<https://www.ncbi.nlm.nih.gov/refseq/>) using DIAMOND BLASTX [8] with parameters -b 12 -c 1 --top 1. A custom Python script ([https://github.com/mvvucak/LASV\\_Sample\\_Classification](https://github.com/mvvucak/LASV_Sample_Classification)) was used with minor adaptations to extract all contigs with at least one hit to members of the family *Orthoherpesviridae*. These contigs were joined by iterative mapping of reads to the contig ends or, where the contig ends mapped to lengthy reiterated regions and could not be joined in this way, by scaffolding the contigs against muromegalovirus genome sequences and representing gaps by stretches of 100 unspecified nucleotides. Alignment of short-read datasets to the scaffolded contigs was then initiated by trimming and quality-filtering the datasets using Trim Galore with parameters --illumina --paired. The trimmed datasets were aligned using Bowtie 2 v2.3.1 [9], with stringent parameters (--score-min L,-0.1,-0.1) for one dataset containing both MnatCMV1 and MnatCMV2 reads in order to aid differentiation of the individual sequences. The alignments were processed using Samtools v1.3 [10] and examined visually using Tablet v1.21.02.08 [11]. This approach was carried out to check and correct the contig sequences, locate the genome termini by identifying sets of reads sharing the same end, and thereby produce correctly configured genome sequences. These sequences were designated as partial because of stretches of unspecified nucleotides in some reiterated regions.

Long-read datasets generated from DNA from tertiary plaque isolates were processed by using Porechop v0.2.4 [12] to demultiplex the reads and remove adapters, dividing the reads when adapter sequences were located internally. The trimmed reads were mapped to the relevant partial genome sequence using Minimap2 v2.10 [13], extracted from the alignment using Samtools v1.12 [10] and subjected to *de novo* assembly using Canu v2.2 [14]. The trimmed reads were mapped to the resulting sequence using Minimap2, and the quality of the consensus sequence was improved using Nanopolish v0.13.3 [15]. The long-read data provided full genome coverage, but were used principally to resolve the unsolved reiterated sequences in the partial genome sequences generated from short-read datasets, thereby providing complete genome sequences.

### Additional sequence analyses

Alignment of short-read datasets to published muromegalovirus sequences was conducted as described above to determine the MCMV Smith, RCMV-E and RCMV-M genome sequences. Also, alignment of short-read datasets generated from DNA from primary isolates to the short (178 bp) reference sequences from the DNA polymerase gene of the four MnatCMVs [2], each flanked by tracts of 20 unspecified nucleotides to aid alignment near the ends, was done as described above to screen for the representation of MnatCMVs.

Metagenomic investigation of short-read datasets generated from both DNA and RNA from primary isolates was used to assess the representation of zoonotic pathogens. The datasets were aligned to the reference sequences of 25 viruses (including LASV), 14 bacteria and two eukaryotes in this category using Bowtie 2, removing rRNA data using riboPicker v0.4.3 [16] with the SILVA rRNA database v138 (<https://www.arb-silva.de/>), removing low-complexity data using PRINSEQ v0.20.4 [17], and determining the taxonomic classification of the remaining reads using DIAMOND BLASTX and BLASTX and BLASTN (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) [18].

### Mapping of RNA processing sites on MnatCMV genome sequences

Two paired-end short read datasets were produced from each polyadenylated RNA sample prepared from cells infected with MnatCMV1, MnatCMV2 or MnatCMV3. One dataset contained sense reads in relation to the parental transcripts and the other contained antisense reads. The two datasets were trimmed and quality-filtered using Trim Galore with parameters --quality 25 --length 50 --paired, and aligned to the relevant complete MnatCMV genome sequence. Two datasets were then extracted from each alignment ([https://github.com/mvvucak/MasCMV\\_Transcriptomics](https://github.com/mvvucak/MasCMV_Transcriptomics)), one containing reads oriented left to right in the genome and the other containing reads oriented right to left, and the pairs of datasets containing reads in the same orientation were combined. Splice sites were identified in the resulting two datasets using STAR v2.5.1b [19] and validated using a kmer-based custom Python script ([https://github.com/mvvucak/MasCMV\\_Transcriptomics](https://github.com/mvvucak/MasCMV_Transcriptomics)) that excluded potential artefacts and counted the number of reads representing each splice site. Polyadenylation sites were identified by mapping trimmed reads onto the genome sequence using ContextMap v2.7.9 [20].

## Annotation of MnatCMV and other muromegalovirus genome sequences

Putative protein-coding regions were identified in the genome sequences as described previously [21], incorporating information on splicing. A range of tools was used to deduce protein features or functions, including BLASTX, SignalP 5.0 [22], TMHMM 2.0 [23], Philius [24], SUPERFAMILY 2 [25] and CDD [26].

## Phylogenetic analysis of MnatCMV and other herpesvirus genome sequences

The phylogenetic relationships among MnatCMVs, other muromegaloviruses and representative viruses in other genera in the subfamily *Betaherpesvirinae* were analysed using the amino acid sequences specified by six conserved genes employed previously for taxonomic purposes (Table S1, available in the online version of this article) [1]. The viruses in other genera were human cytomegalovirus (HCMV; genus *Cytomegalovirus*) [27], guinea pig cytomegalovirus (GPCMV; genus *Quwivirus*) [28], human herpesvirus 6A (HHV6A; genus *Roseolovirus*) [29] and elephant endotheliotropic herpesvirus 1 (EEHV1; genus *Proboscivirus*) [21].

The amino acid sequences for each virus were concatenated and analysed using MEGA X v10.2.6 [30, 31]. The sequences were aligned using Muscle within MEGA X, and the initial tree search was generated automatically by applying the neighbour-joining and BioNJ algorithms to a matrix of pairwise distances estimated using the Jones–Taylor–Thornton model. The topology with the highest log likelihood value was then subjected to maximum-likelihood heuristic analysis with nearest-neighbour-interchange using the Le and Gascuel model [32] and a discrete gamma (G) distribution in five categories with invariant sites (I). Gaps in the alignment were treated by partial deletion (site coverage cutoff value=90%) with no branch swap filter and with 100 bootstrap replications. The phylogenetic tree with the highest log likelihood value was visualized and midpoint-rooted using FigTree v1.4.4 (<http://tree.bio.ed.ac.uk/software/figtree/>). The computational statistics were as follows: log likelihood value=−67663.27; G=0.8751; I=0.0467; and sites analysed=4929.

## RESULTS

### Isolation and sequencing of MnatCMVs

MnatCMVs were isolated from salivary gland extracts initially as *primary isolates* and then as *primary plaque isolates* (plaque-purified once) or *tertiary plaque isolates* (plaque-purified three times). Primary isolates were obtained from 14 of the 37 animals tested (animals 1, 2, 8, 9, 13, 15, 18, 19, 28, 29, 33, 35, 36 and 38). Multiple primary isolates were made from the same extract on different occasions from four of these animals (two from animal 1, three from animal 2, four from animal 35 and two from animal 36), and were invariably successful as confirmed by sequencing of the primary isolates. The number of primary isolates, and therefore the number of datasets generated from them, was 21. Metagenomic investigation of short-read datasets generated from both DNA and RNA from the first isolation from animals 1, 2, 35 and 36 yielded no evidence for the representation of any zoonotic pathogen. Primary and tertiary plaque isolates were made from the primary isolates from three animals (animals 36, 25 and 35) representing MnatCMV1, MnatCMV2 and MnatCMV3, respectively. The number of datasets generated from primary and tertiary plaque isolates was 11 and 25, respectively, bringing the number of datasets generated from all isolates to 57. Data on the MnatCMV isolates, datasets and genome sequences are shown in Table 2.

### Representation of MnatCMVs in primary isolate datasets

The nomenclature of primary isolate datasets consists of an abbreviation of the animal species name followed by the number of the animal and a letter designating the isolation experiment (e.g. Mnat19A represents the dataset from the single MnatCMV2 primary isolate from animal 19, and Mnat2C represents the dataset from the third MnatCMV2 primary isolate from animal 2).

Prior to determining genome sequences from the datasets, MnatCMVs represented in the primary isolate datasets were identified by aligning the reads to the available short (178 nt) sequences from the DNA polymerase genes of MnatCMV1, MnatCMV2, MnatCMV3 and MnatCMV4 (Fig. 1). The results showed that 12 of the 14 animals from which primary isolates had been made were represented in the datasets by a single MnatCMV, in that all MnatCMV reads in these datasets aligned with only one of the DNA polymerase gene sequences. One of these animals (animal 2) contained two versions of the MnatCMV2 sequence differing by a single nucleotide substitution; the strains that these sequences represented were termed MnatCMV2A and MnatCMV2B. Enumeration of reads containing signature motifs (typically 20–25 nt) specific to each strain showed that MnatCMV2A comprised approximately 85, 95 and 100% of the virus population represented in the datasets from the first (Mnat2A), second (Mnat2B) and third (Mnat2C) primary isolates from animal 2, respectively. Enumeration of reads in the datasets from the remaining two animals that contained multiple MnatCMVs showed that the dataset from the primary isolate from animal 29 (Mnat29A) represented MnatCMV1 (45%), MnatCMV2 (45%) and MnatCMV3 (10%); and that the dataset (Mnat36A) from the first primary isolate from animal 36 represented MnatCMV1 (75%) and MnatCMV3 (25%). In contrast, the dataset (Mnat36B) from the second primary isolate from animal 36 contained only MnatCMV3. These results indicated that 12 animals were infected with a single MnatCMV, and that three animals were coinfecting with multiple MnatCMVs or MnatCMV strains. No dataset represented MnatCMV4.

**Table 2.** MnatCMV isolates and genome sequences

Genome	Animal	Dataset	Contigs (no.)	GenBank accession	Differences from reference sequence*
<b>Primary isolates</b>					
MnatCMV1	29	Mnat29A	3	OP429121.1	NA
MnatCMV1	36	Mnat36A	2	OP429122.1	NA
MnatCMV2A	2	Mnat2A	2	OP429126.1	NA
MnatCMV2A	2	Mnat2B	2	NA	None
MnatCMV2A	2	Mnat2C	3	NA	None
MnatCMV2	18	Mnat18A	2	OP429123.1	NA
MnatCMV2	19	Mnat19A	2	OP429124.1	NA
MnatCMV2	29	Mnat29A	3	OP429125.1	NA
MnatCMV2	33	Mnat33A	3	OP429127.1	NA
MnatCMV3	1	Mnat1A	2	OP429130.1	NA
MnatCMV3	1	Mnat1B	2	NA	None
MnatCMV3	8	Mnat8A	1	OP429136.1	NA
MnatCMV3	9	Mnat9A	2	OP429137.1	NA
MnatCMV3	13	Mnat13A	2	OP429128.1	NA
MnatCMV3	15	Mnat15A	2	OP429129.1	NA
MnatCMV3	28	Mnat28A	2	OP429131.1	NA
MnatCMV3	29	Mnat29A	3	OP429132.1	NA
MnatCMV3	35	Mnat35A	2	OP429133.1	NA
MnatCMV3	35	Mnat35B	2	NA	None
MnatCMV3	35	Mnat35C	2	NA	None
MnatCMV3	35	Mnat35D	2	NA	None
MnatCMV3	36	Mnat36A	2	OP429134.1	NA
MnatCMV3	36	Mnat36B	2	NA	None
MnatCMV3	38	Mnat38A	1	OP429135.1	NA
<b>Primary plaque isolates</b>					
MnatCMV2B	2	Mnat2A 2–4	NA	OP429141.1	87 substitutions and two indels
MnatCMV2B	2	Mnat2A 2–5	NA	NA	87 substitutions and two indels
MnatCMV2A	2	Mnat2A 2–6	NA	NA	Non-coding substitutions at 443 (T to C), 736 (T to C) and 149510 (T to C)
MnatCMV2A	2	Mnat2A 2–7	NA	NA	None
MnatCMV2A	2	Mnat2A 2–8	NA	NA	None
MnatCMV2A	2	Mnat2A 2–9	NA	NA	None
MnatCMV3	35	Mnat35A 35–5	NA	NA	Variant at 196255 (A to G in gene b165; C to R)
MnatCMV3	35	Mnat35A 35–6	NA	NA	None
MnatCMV3	35	Mnat35A 35–7	NA	NA	None
MnatCMV3	35	Mnat35A 35–8	NA	NA	None
MnatCMV3	35	Mnat35A 35–9	NA	NA	None
<b>Tertiary plaque isolates</b>					
MnatCMV1	36	Mnat36A B1-2	NA	NA	None

Continued

Table 2. Continued

Genome	Animal	Dataset	Contigs (no.)	GenBank accession	Differences from reference sequence*
MnatCMV1	36	Mnat36A B1-3	NA	NA	None
MnatCMV1	36	Mnat36A C2-2	NA	NA	None
MnatCMV1	36	Mnat36A C2-3	NA	NA	None
MnatCMV1	36	Mnat36A C3-1	NA	NA	None
MnatCMV1	36	Mnat36A C3-2	NA	NA	None
MnatCMV1	36	Mnat36A D3-2	NA	NA	Variant at 156426 (indel; non-coding G-tract)
MnatCMV1	36	Mnat36A-D3-3	NA	NA	Variant at 131688 (C to A; non-coding)
MnatCMV1	36	Mnat36A E2-2	NA	NA	Variant at 36609 (C to A; stop codon in gene M38)
MnatCMV1	36	Mnat36A E2-3	NA	OP429138.1†	None
MnatCMV1	36	Mnat36A F3-2	NA	NA	None
MnatCMV1	36	Mnat36A F3-3	NA	NA	None
MnatCMV2A	2	Mnat2A A1-1	NA	NA	Substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2A	2	Mnat2A A1-2	NA	NA	Substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2A	2	Mnat2A A1-3	NA	NA	Substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2B	2	Mnat2A B1-1	NA	NA	83 substitutions and two indels
MnatCMV2B	2	Mnat2A B1-2	NA	NA	83 substitutions and two indels
MnatCMV2B	2	Mnat2A B1-3	NA	NA	83 substitutions and two indels
MnatCMV2A	2	Mnat2A C2-1	NA	NA	Variant at 140139 (C to T in gene M116; A to T); substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2A	2	Mnat2A C2-2	NA	NA	Variant at 140139 (C to T in gene M116; A to T); substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2A	2	Mnat2A C2-3	NA	NA	Variant at 140139 (C to T in gene M116; A to T); substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2A	2	Mnat2A C3-1	NA	NA	Substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2A	2	Mnat2A C3-2	NA	NA	Substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV2A	2	Mnat2A C3-3	NA	OP429139.1†	Substitutions at 149510 (T to C; non-coding) and 171871 (C to T; V to I in gene m142)
MnatCMV3	35	Mnat35A A3-2	NA	OP429140.1†	None

\*The reference sequence used for each comparison was that of the first primary isolate obtained from that animal. Primary isolates obtained on multiple occasions from the same animal (animals 1, 2, 35 and 36) had identical sequences, except for heterogeneities in a few polynucleotide tracts in MnatCMV2 isolated from animal 2 that are not listed. Similarly, length differences or heterogeneities in polynucleotide tracts were more common in plaque isolates from animal 2 than in plaque isolates from other animals and are not listed. Complete nucleotide replacements are described as substitutions, partial nucleotide replacements are described as variants, and the effects of non-synonymous substitutions are described in one-letter amino acid code. NA, not applicable; indel, insertion or deletion.

†Complete genome sequences incorporating long-read sequence data.

## Genome sequences of MnatCMV primary isolates

Although 21 primary isolate datasets were generated, 24 MnatCMV genome sequences were determined because, as described above, multiple MnatCMVs were isolated from two animals. *De novo* assembly generated a small number (one to three) of MnatCMV contigs from each dataset (Table 2) that represented most of the genome. These were then joined and processed into genome sequences. The finding that MnatCMVs in primary isolates obtained on multiple occasions from the same animal had an identical sequence indicated that any mutants that may have arisen during the isolation procedure had not become dominant.

## Genome sequences of MnatCMV plaque isolates

Plaque isolates representing MnatCMV1, MnatCMV2 and MnatCMV3 were made from animals 36, 2 and 35, respectively (Table 2). Several plaques were picked from each of the first primary isolates of MnatCMV2 and MnatCMV3 (corresponding to datasets Mnat2A and Mnat35A, respectively), yielding primary plaque isolates. Datasets were obtained for some of these isolates (Mnat2A 2-4 to 2-9 and Mnat35A 35-5 to 35-9). Separately, several plaques were picked from each of the first primary isolates of

<b>MnatCMV1</b>	GGCAGCTGGAATGCTTCCTTGTTTTCCGATAGCTTCTTCCATAACTAAAAATAGGTAGAGA
<b>Mnat29A</b>	GGCAGCTGGAATGCTTCCTTGTTTTCCGATAGCTTCTTCCATAACTAAAAATAGGTAGAGA
<b>Mnat36A</b>	GGCAGCTGGAATGCTTCCTTGTTTTCCGATAGCTTCTTCCATAACTAAAAATAGGTAGAGA
<b>MnatCMV2</b>	GGCGGCTGGAATGCTTCCTTGTTTTCCGATAGCTTCCGATAACCAAGATAGGTAGAGA
<b>Mnat2A</b>	GGCGGCTGGAATGCTTCCTTGTTTTCCGATAGCTT <b>CA</b> TCGATAACCAAGATAGGTAGAGA
<b>Mnat2B</b>	GGCGGCTGGAATGCTTCCTTGTTTTCCGATAGCTTCCGATAACCAAGATAGGTAGAGA
<b>Mnat18A</b>	GGCGGCTGGAATGCTTCCTTGTTTTCCGATAGCTT <b>CA</b> TCGATAACCAAGATAGGTAGAGA
<b>Mnat19A</b>	GGCGGCTGGAATGCTTCCTTGTTTTCCGATAGCTT <b>CA</b> TCGATAACCAAGATAGGTAGAGA
<b>Mnat29A</b>	GGCGGCTGGAATGCTTCCTTGTTTTCCGATAGCTT <b>CA</b> TCGATAACCAAGATAGGTAGAGA
<b>Mnat33A</b>	GGCGGCTGGAATGCTTCCTTGTTTTCCGATAGCTT <b>CA</b> TCGATAACCAAGATAGGTAGAGA
<b>MnatCMV3</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat1A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat8A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat9A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat13A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat15A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat28A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat29A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat35A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat36A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>Mnat38A</b>	AGCCGCTGGAATGTTACCATGTCTTCCATTGCTTCTCTATAACTAAGATTGGTCGAGA
<b>MnatCMV4</b>	GGCTGCCGGGATGCTTCCTTGCTTCCGATCGCTTCTTCTATTACCAAGATCGGTCGGGA

<b>MnatCMV1</b>	GATGTTATTATCTACGGCTAAATATATAGAGGAAAATTGTAATAACAGAGAATTTCTGGA
<b>Mnat29A</b>	GATG <b>CT</b> ATTATCTACGGCTAAATATATAGAGGAAAATTGTAATAACAGAGAATTTCTGGA
<b>Mnat36A</b>	GATG <b>CT</b> ATTATCTACGGCTAAATATATAGAGGAAAATTGTAATAACAGAGAATTTCTGGA
<b>MnatCMV2</b>	GATGCTGTTATCCACGGCTAGATATATAGAGGAAAATTGTAATAATAGGGAATTTTATAGA
<b>Mnat2A</b>	GATGCTGTTATCCACGGCTAGATATATAGAGGAAAATTGTAATAATAGGGAATTTTATAGA
<b>Mnat2B</b>	GATGCTGTTATCCACGGCTAGATATATAGAGGAAAATTGTAATAATAGGGAATTTTATAGA
<b>Mnat18A</b>	GATGCTGTTATCCACGGCTAGATATATAGAGGAAAATTGTAATAATAGGGAATTTTATAGA
<b>Mnat19A</b>	GATGCTGTTATCCACGGCTAGATATATAGAGGAAAATTGTAATAATAGGGAATTTTATAGA
<b>Mnat29A</b>	GATGCTGTTATCCACGGCTAGATATATAGAGGAAAATTGTAATAATAGGGAATTTTATAGA
<b>Mnat33A</b>	GATGCTGTTATCCACGGCTAGATATATAGAGGAAAATTGTAATAATAGGGAATTTTATAGA
<b>MnatCMV3</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat1A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat8A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat9A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat13A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat15A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat28A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat29A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat35A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat36A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>Mnat38A</b>	CATGCTGCTTTTCGACAGCCGCGTACATGGAAGCAATGCAATGATGTAAATTTTTTAGT
<b>MnatCMV4</b>	GATGTTGATTCTACTTCGCTTATATAGAAAAACATTGTAATGACATATCATTCCTAAC

<b>MnatCMV1</b>	AAGAGTCATCGGAATCGAATCGGAAGCTATAGATTTCAGAACTTTACGTGTAAAAATT
<b>Mnat29A</b>	AAGAGTCATCGGAATCGAATCGGAAGCTATAGATTTCAGAACTTTACGTGTAAAAATT
<b>Mnat36A</b>	AAGAGTCATCGGAATCGAATCGGAAGCTATAGATTTCAGAACTTTACGTGTAAAAATT
<b>MnatCMV2</b>	GAGGGTCATTGGGATAGAATCGGAAGCCATAGATTTCGAAAAATTTACGTGTGAAAAATA
<b>Mnat2A</b>	GAGGGTCATTGGGATAGAATCGGAAGCCATAGATTTCGAAAAATTTACGTGTGAAAAATA
<b>Mnat2B</b>	GAGGGTCATTGGGATAGAATCGGAAGCCATAGATTTCGAAAAATTTACGTGTGAAAAATA
<b>Mnat18A</b>	GAGGGTCATTGGGATAGAATCGGAAGCCATAGATTTCGAAAAATTTACGTGTGAAAAATA
<b>Mnat19A</b>	GAGGGTCATTGGGATAGAATCGGAAGCCATAGATTTCGAAAAATTTACGTGTGAAAAATA
<b>Mnat29A</b>	GAGGGTCATTGGGATAGAATCGGAAGCCATAGATTTCGAAAAATTTACGTGTGAAAAATA
<b>Mnat33A</b>	GAGGGTCATTGGGATAGAATCGGAAGCCATAGATTTCGAAAAATTTACGTGTGAAAAATA
<b>MnatCMV3</b>	AAATACTTTAGGTATTTCTGCTATGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat1A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat8A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat9A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat13A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat15A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat28A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat29A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat35A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat36A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>Mnat38A</b>	AAATACTTTAGGTATTTCTG <b>TA</b> TGTGATGTGGATAGTTCGACTTTAAATGTAAAAATA
<b>MnatCMV4</b>	AGAAGTCCTTGGTATTTCCCATGGAATCTATAGATACAGATGTTCTGATGGTGAATAA

**Fig. 1.** Sequence alignment of a 178 bp region of the DNA polymerase gene in MnatCMV primary isolates. The alignment is presented in three contiguous sections separated by grey horizontal lines. The sequences are grouped with the MnatCMV references (names in bold font), and differences are also highlighted in bold font. Dataset Mnat36 represented MnatCMV1 and MnatCMV3, dataset Mnat29 represented MnatCMV1, MnatCMV2 and MnatCMV3, and dataset Mnat2 represented two MnatCMV2 strains (MnatCMV2A and MnatCMV2B). No isolate represented MnatCMV4.

MnatCMV1, MnatCMV2 and MnatCMV3 (corresponding to datasets Mnat36A, Mnat2A and Mnat35A). Three plaques were then picked from each of these at the second round of plaque purification. Finally, one plaque was picked from each of these plaques at the third round of plaque purification. As a consequence, the tertiary plaque isolates were termed in threes (e.g. Mnat2A A1-1 to A1-3), each plaque in a trio having been derived from the same primary plaque isolate. Datasets were then obtained for some of these isolates (Table 2).

Nine of the MnatCMV1 tertiary plaque isolates were identical in sequence to the parental primary isolate, with each of three others containing variant subpopulations that differed in a single nucleotide. The MnatCMV3 genomes sequenced from the primary and tertiary plaque isolates were identical to that of the genome sequenced from the parental primary plaque isolate. The situation with the MnatCMV2 plaque isolates was complicated by the presence of two MnatCMV2 strains (MnatCMV2A and MnatCMV2B) in the primary isolate. Alignment of dataset Mnat2A to the genome consensus sequence determined from this dataset revealed 89 variant nucleotides (i.e. nucleotides present in minor proportions that differed from the consensus at these positions; Table 3). These non-consensus variant nucleotides were supplemented by length differences in several polynucleotide tracts, which may be inherently heterogeneous in the genome and are also subject to artefactual heterogeneity generated during library preparation and sequencing; these differences are not included in Table 3. The proportion of each variant nucleotide determined by enumerating reads containing signature motifs showed that, but for two exceptions at positions 149510 and 171871, all variants were uniformly represented in datasets Mnat2A, Mnat2B and Mnat2C at average levels of  $14.87 \pm 2.17$ ,  $4.57 \pm 0.86$  or  $0.05 \pm 0.13\%$ ; taking into account the probable presence of a low level of variant reads due to sequencing error, the value for Mnat2C was essentially zero. In contrast, the variants at positions 149510 and 171871 were represented at significant levels (7.80 and 6.77%, respectively) in Mnat2C, and in Mnat2A and Mnat2B at 50.00 and 36.82, and 5.60 and 5.30%, respectively. These findings implied that all but two of the variants were present in a subpopulation of the MnatCMV2B genome and that the two exceptional variants were present in a subpopulation of the MnatCMV2A genome.

As the subpopulations in the primary isolates were represented at minor (or zero) levels in datasets Mnat2A, Mnat2B and Mnat2C, the consensus sequence in each case was for MnatCMV2A rather than for MnatCMV2B. In contrast, each plaque isolate dataset represented a set of nucleotides at the variant positions that were diagnostic of either MnatCMV2A or MnatCMV2B. Two primary plaque isolate datasets (Mnat2A 2-4 and 2-5) represented MnatCMV2B, whereas the other four primary plaque isolate datasets (Mnat2A 2-6 to 2-9) represented MnatCMV2A, although one (Mnat2A 2-6) had three additional non-coding substitutions that were not among the variant positions in the primary datasets. Similarly, the sequences of three tertiary plaque isolate datasets (Mnat2A B1-1 to B1-3) represented MnatCMV2B, whereas the other eight tertiary plaque isolate datasets (Mnat2A A1-1 to A1-3, C2-1 to C2-3 and C3-1 to C3-3) represented MnatCMV2A. However, the former group lacked four substitutions (at positions 10284, 11349, 12327 and 12352) characteristic of MnatCMV2B, and the latter group had the two MnatCMV2A substitutions at positions 149510 and 171871. Various other sporadic differences from the parental strain were also detected in some plaque isolate datasets.

The observations made above evidently arose from the presence of two MnatCMV2 strains (MnatCMV2A and MnatCMV2B) in animal 2, and imply that recombination between them had occurred during, and even possibly before, primary isolation. Indeed, all 89 substitutions were represented in the primary plaque isolates representing MnatCMV2B (Mnat2A 2-4 and 2-5), suggesting that the two exceptional variants were common to both MnatCMV2A and MnatCMV2B, presumably as a result of recombination. Moreover, the observation that the 87 MnatCMV2B-specific variants in the primary isolates from animal 2 were localized to approximately the first one-third of the MnatCMV2B genome indicated that this strain may itself have been an earlier product of a recombination event that had occurred prior to primary isolation and probably in a host prior to animal 2. These findings highlight the potential for the occurrence of recombination events between multiple MnatCMV strains during isolation and very probably during growth in animals, such events giving rise to unavoidably complex outcomes.

Overall, as shown in Table 2, these experiments resulted in the examination of the genome sequences from primary or plaque isolates of two MnatCMV1 strains (derived from the sequencing of 14 viruses), six MnatCMV2 strains (25 viruses) and ten MnatCMV3 strains (21 viruses), totalling 18 MnatCMV strains (60 viruses).

### Completion of MnatCMV genome sequences

The genome sequences determined above were designated as partial because they contained a few reiterated regions, the lengths of which could not be resolved using Illumina-based short-read datasets. The genome sequences of one representative each of MnatCMV1, MnatCMV2 (an MnatCMV2A strain) and MnatCMV3 were completed in these regions by using ONT-based long-read datasets generated from tertiary plaque isolates (Table 4), and consisted of 205097, 207012 and 211478 bp, respectively. These sequences represent the partial genome sequences determined from short-read datasets with the gaps filled by long-read data. The sequences assembled from the long-read datasets alone exhibited differences from these sequences at 88–169 locations, almost all of which were due to differences in the lengths of polynucleotide tracts, for which long-read data are less reliable than short-read data.

**Table 3.** Variant nucleotides in primary isolate datasets from animal 2

Location*	Type†	Minor variant in dataset (%)			Location*	Type†	Minor variant in dataset (%)		
		Mnat2A	Mnat2B	Mnat2C			Mnat2A	Mnat2B	Mnat2C
10284	S	15.13	5.12	0.00	66477	S	14.40	4.28	0.00
11349	S	14.30	4.81	0.00	68899	S	16.29	4.15	0.00
12327	S	14.52	3.71	0.00	69880	S	14.25	3.96	0.00
12352	S	14.77	4.44	0.00	69893	S	15.37	4.10	0.07
20430	S	12.37	3.94	0.07	69983	S	18.26	4.13	0.00
22656	S	17.26	5.07	0.00	69993	S	17.56	4.46	0.00
27047	S	13.32	4.76	0.08	72506	S	13.84	4.12	0.00
31652	S	14.67	5.07	0.00	73806	S	13.25	4.13	0.00
31700	S	15.82	4.43	0.00	73888	S	14.71	4.32	0.06
35171	S	13.03	3.94	0.07	73978	S	12.12	4.88	0.13
38081	S	16.70	4.35	0.47	73984	S	11.99	4.94	0.07
39231	S	14.00	4.47	0.08	73986	S	11.90	4.86	0.07
39803	S	17.97	3.93	0.00	74100	S	15.04	4.31	0.00
39917	S	15.93	4.04	0.00	74123	S	15.52	4.13	0.00
40882	S	14.02	4.56	0.00	74132	S	14.95	4.24	0.00
41215	S	13.67	4.63	0.00	74133	S	15.08	4.22	0.07
41392	S	14.87	5.13	0.44	74134	S	15.23	4.24	0.00
41666	S	15.07	4.32	0.00	74165	S	14.23	4.37	0.00
42644	S	16.22	3.86	0.07	74265	S	12.93	4.64	0.07
43770	S	14.20	4.86	0.07	74418	S	13.78	4.09	0.09
43875	S	14.71	4.54	0.00	74548	S	13.83	5.02	0.85
47819	S	13.27	5.06	0.00	74699	S	16.25	4.71	0.00
48050	S	16.27	5.43	0.00	74741	S	15.99	4.50	0.00
48680	S	14.00	4.38	0.00	74748	S	15.92	4.35	0.00
49466	S	15.73	4.30	0.12	74819	S	17.27	4.17	0.10
50455	S	15.19	4.56	0.07	74911	S	14.84	4.86	0.00
50503	S	14.79	4.63	0.00	75198	S	15.56	4.08	0.00
52515	S	13.55	4.79	0.07	75339	S	8.70	3.40	0.00
53114	S	16.39	4.06	0.07	75420	S	24.27	9.53	0.00
53233	S	13.56	4.35	0.00	75424	D	24.39	9.31	0.00
53569	S	13.40	4.71	0.00	75532	I	17.53	4.91	0.00
54412	S	11.62	3.92	0.00	75701	S	14.15	5.56	0.00
54481	S	12.94	4.05	0.00	75719	S	13.28	5.48	0.00
55612	S	12.48	4.32	0.07	76680	S	14.76	3.92	0.23
55726	S	14.34	4.99	0.00	76775	S	16.03	4.49	0.55
55891	S	14.40	4.71	0.07	76804	S	12.71	3.62	0.00
56296	S	14.87	4.46	0.00	76958	S	12.45	4.73	0.00
57581	S	17.41	4.32	0.00	77751	S	17.01	4.85	0.07

Continued

Table 3. Continued

Location*	Type†	Minor variant in dataset (%)			Location*	Type†	Minor variant in dataset (%)		
		Mnat2A	Mnat2B	Mnat2C			Mnat2A	Mnat2B	Mnat2C
59252	S	15.89	4.61	0.00	77967	S	13.71	4.13	0.00
60518	S	14.52	4.68	0.00	78076	S	15.46	4.18	0.15
61368	S	15.81	4.92	0.00	78658	S	12.77	4.43	0.07
62754	S	14.02	4.54	0.07	78772	S	14.12	3.94	0.00
64491	S	13.11	4.43	0.07	149510	S	50.00	5.60	7.80
65742	S	16.19	4.44	0.00	171871	S	36.82	5.30	6.77
66006	S	15.83	4.33	0.12					

\*Nucleotide position in the consensus genome from dataset Mnat2A (GenBank accession OP429126.1).

†Type of variant: S, substitution; I, insertion; and D, deletion.

Each MnatCMV genome has a 30bp direct repeat at its termini. The corresponding repeat in the MCMV Smith, MCMV s09 and RCMV-E genomes is also this size, but that in the RCMV-M genome is larger (504bp). The sequences of reads originating from MnatCMV DNA molecules (presumably circular or concatemeric) in which the termini are joined indicated the presence of a single unpaired nucleotide at the 3'-end of each genome strand for each virus. There was also evidence in the datasets for an alternative right terminus in approximately 50% of MnatCMV1 and MnatCMV2 genomes. This terminus is located at position 3471 and 4421 (this being equivalent to the unpaired nucleotide at the 3'-end of the rightward genome strand) in the complete MnatCMV1 and MnatCMV2 genomes, respectively, and presumably originated from genomes that are approximately 2% longer. The presence of both genome forms in all MnatCMV1 and MnatCMV2 isolates, including those that had been plaque-purified, implies that each was able to give rise to both during virus DNA replication and maturation. An alternative right genome end was not detected for MnatCMV3.

### Annotation of the MnatCMV and other muromegalovirus genome sequences

The maps in Figs 2–4 show the locations of predicted functional protein-coding regions in the complete MnatCMV genomes. These maps differentiate among core genes that are conserved in the family *Orthoherpesviridae* (subfamilies *Alphaherpesvirinae*, *Betaherpesvirinae* and *Gammaherpesvirinae*; core genes), genes that are conserved only in the subfamilies *Betaherpesvirinae* and *Gammaherpesvirinae* (betagamma genes), and genes that are conserved only in some members of the subfamily *Betaherpesvirinae* or are unique to a particular virus in this subfamily (non-core genes). The non-core genes include five families of related and mostly paralogous genes [the lectin, US22, UL25, G protein-coupled receptor (GPCR) and m145 families], and the betagamma genes (with one core gene member, M72) include an additional family [the deoxyuridine triphosphatase-related protein (DURP) family]. Table S2 provides a tabulated version of the maps extended to include equivalent information for the RCMV-E, MCMV Smith, MCMV s09 and RCMV-M genomes (each newly analysed and reannotated), with added detail on protein names, characteristics and functions. The annotation process depended on multiple bioinformatic analyses and also took into account information on splicing in the MnatCMVs (Table S3). Many splice sites were not accommodated in the maps because they were located in non-coding 5'-leader regions of coding transcripts, resulted in alternative splicing of coding regions or were located in antisense transcripts. Numerous spliced antisense transcripts have been reported previously for HCMV [33], but few have been shown to encode functional proteins.

Table 4. Features of long-read sequence datasets for completing MnatCMV genome sequences

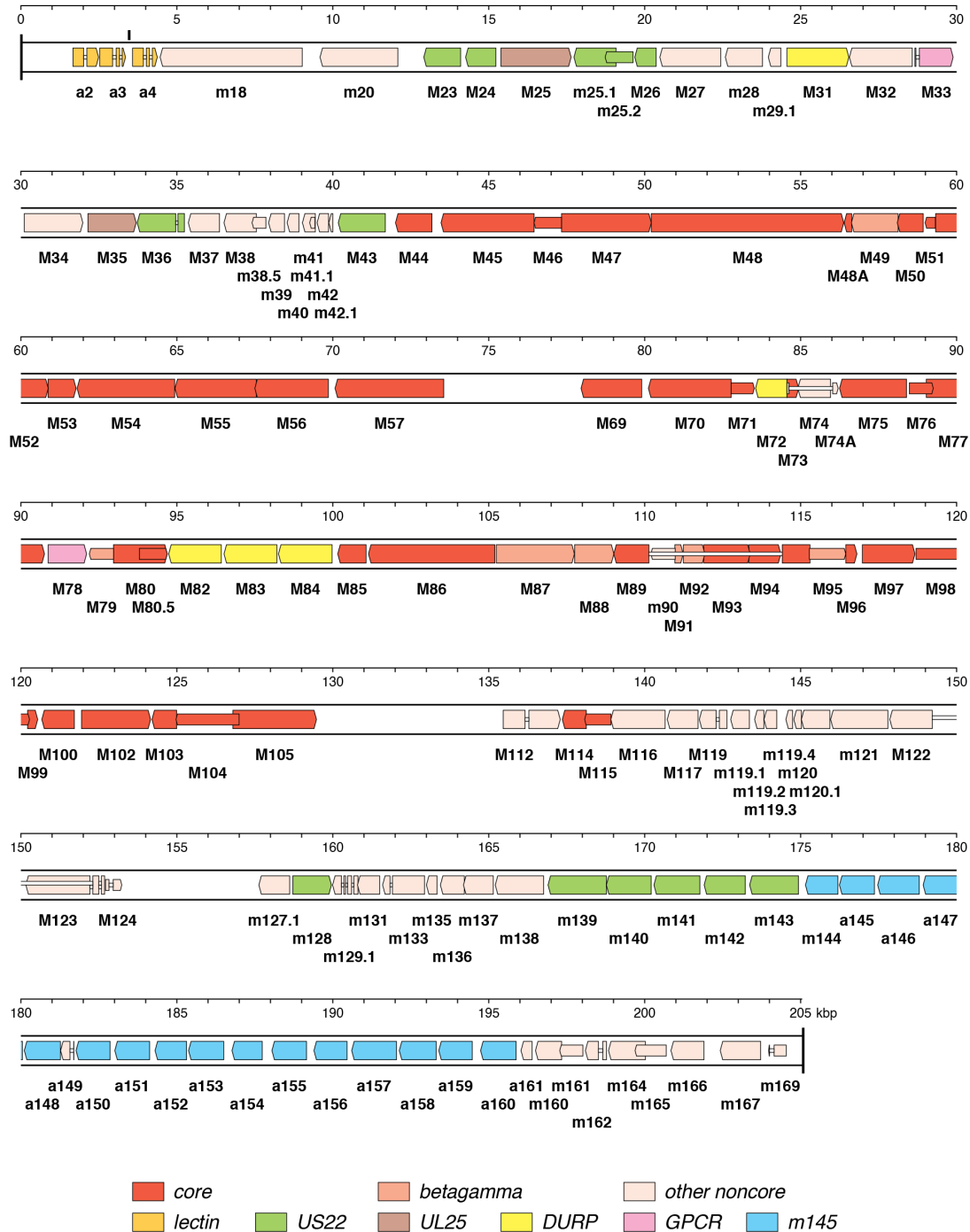
Virus	Corresponding SR dataset*	Reads (no.)		Reads mapping to (no.)		Differences‡ (no.)
		Original	Trimmed	SR genome	LR† genome	
MnatCMV1	Mnat36A E2-3	307535	307049	38326	38302	152
MnatCMV2	Mnat2A C3-3	252363	251971	26728	25483	169
MnatCMV3	Mnat35A A3-2	703074	701775	102749	97912	88

\*See Table 2. SR, short-read.

†LR, long-read.

‡Between the genome sequences determined from SR data with reiterated regions derived from LR data (GenBank accessions OP429138.1, OP429139.1 and OP429140.1) and the genome sequences assembled from LR data alone.

# MnatCMV1



**Fig. 2.** Map of the MnatCMV1 genome (GenBank accession OP429138.1). The terminal direct repeats are shown in a thicker format. The ORFs comprising protein-coding regions are indicated by coloured arrows grouped according to the key, with corresponding gene names below. Introns connecting protein-coding regions are shown as narrow white bars. The colours of protein-coding regions indicate genes that are conserved in the family *Orthoherpesviridae* (core genes) or only in the subfamilies *Betaherpesvirinae* and *Gammapherpesvirinae* (betagamma genes). The remaining coding regions (non-core genes) include six families of related genes (the lectin, US22, UL25, DURP, GPCR and m145 families). M72 is shown as a member of the DURP family and is also a core gene. The alternative right genome end that would extend the standard genome by about 2% is indicated by the vertical black bar between genes a3 and a4.

# MnatCMV2

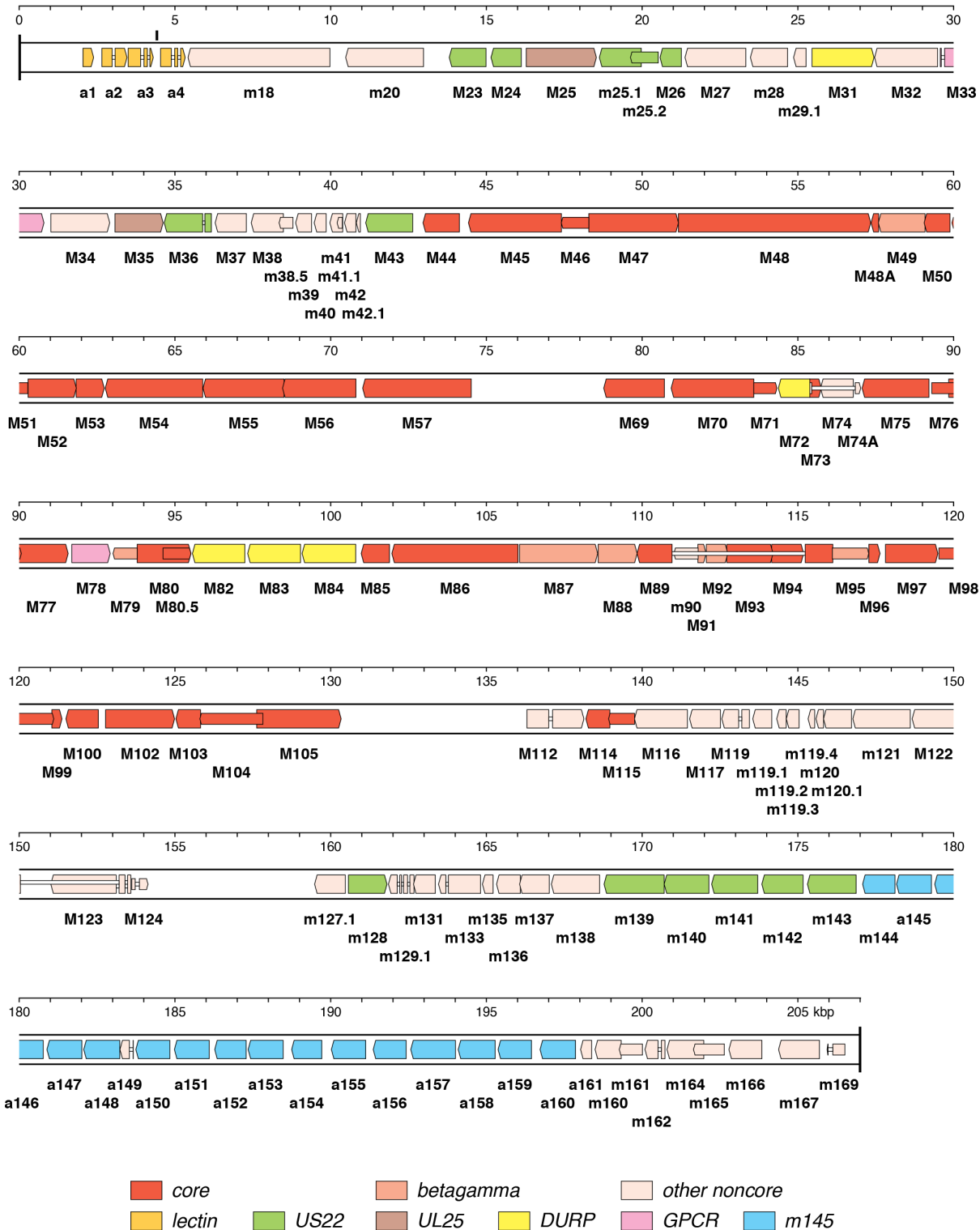
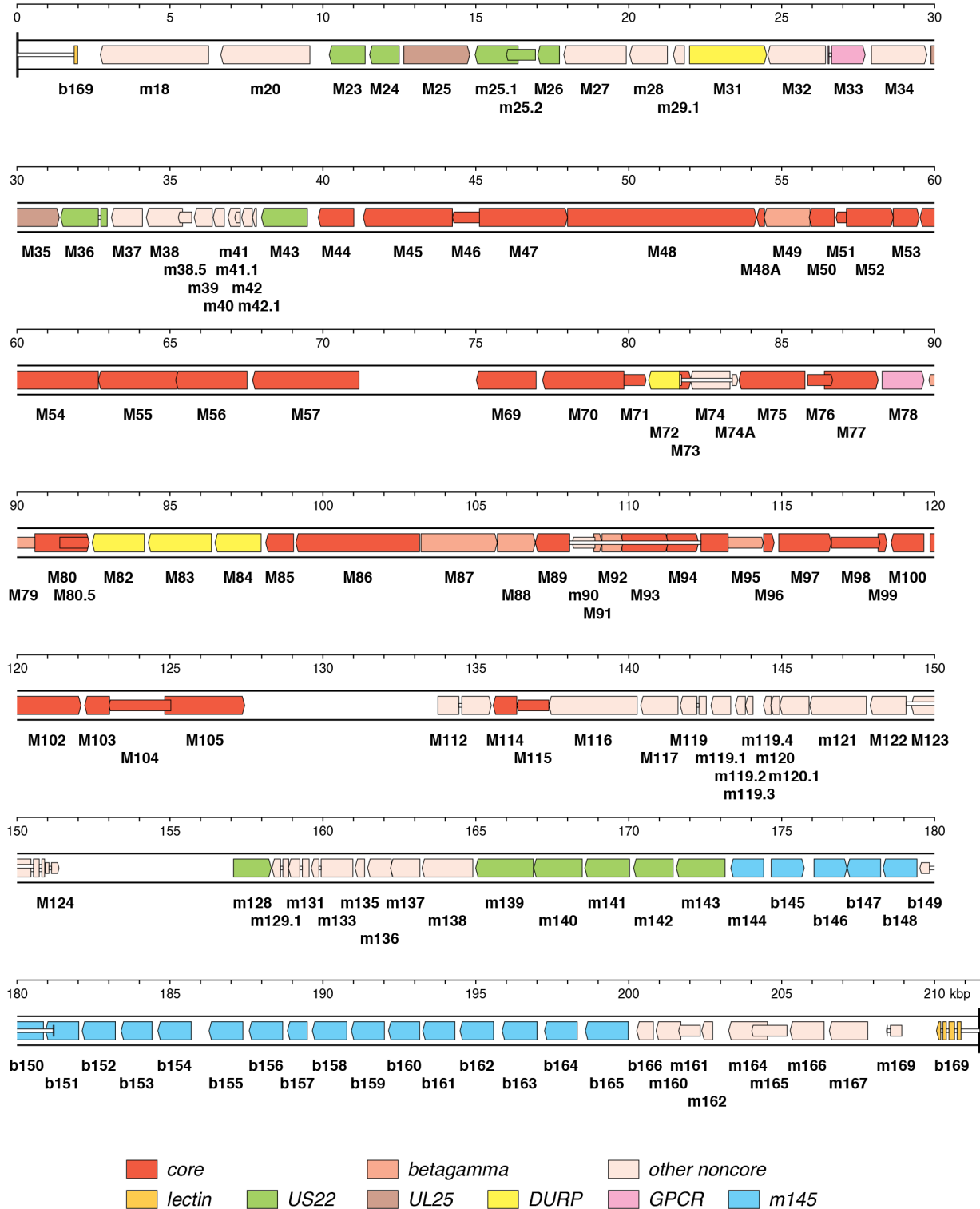


Fig. 3. Map of the MnatCMV2 genome (GenBank accession OP429139.1). See the Fig. 2 legend for details.

# MnatCMV3



**Fig. 4.** Map of the MnatCMV3 genome (GenBank accession OP429140.1). See the Fig. 2 legend for details. The genome lacks an alternative right end.

The RNA samples that were sequenced for the splicing analysis were harvested from asynchronously infected cells and therefore contained transcripts produced at all stages of infection. This made the data largely unsuitable for analyses of transcript abundance, but provided useful information not only on splicing but also on the location of polyadenylation sites at the 3'-end of transcripts, most of which mapped close downstream from an annotated protein-coding region (Table S4). Nonetheless, detection of splice sites and polyadenylation sites would have been biased and therefore incomplete, being more likely for the more abundant RNAs and for the datasets containing a greater proportion of MnatCMV reads. The numbers of reads in the trimmed datasets that mapped to the MnatCMV1, MnatCMV2 and MnatCMV3 genome sequences were 10481592 (33% of the total), 8294413 (26%) and 5432483 (18%), respectively.

Assigning nomenclature to MnatCMV genes was not straightforward because the annotation of muromegalovirus genomes has evolved as more genomic and functional information has become available. As a result, gene nomenclature has become increasingly diverse and, in some respects, inconsistent. In order to avoid adding further unnecessary complexity, we started with the convention established for MCMV Smith [34], in which genes with orthologues in HCMV adopted the numeral assigned to the HCMV gene prefixed by M (e.g. the orthologue of HCMV gene UL54, which encodes DNA polymerase, was named M54) and other genes lacking orthologues in HCMV were prefixed by m and numbered sequentially or in relation to flanking genes (e.g. m90 is located between M89 and M91). We applied this scheme to annotating the MnatCMV genomes and reannotating the RCMV-E, MCMV Smith, MCMV s09 and RCMV-M genomes, with the aim of denoting orthologous genes in different viruses by the same name (Table S2). For example, M54 and m90 were used for the relevant orthologue in each of the muromegalovirus genomes. The m prefix was then applied to all other genes lacking orthologues in HCMV, with the exception of those in the three highly divergent regions described below. Based on this nomenclature, genes with the M prefix are conserved in all genomes, whereas genes with the m prefix are not necessarily conserved. This approach differed from the earlier convention of using a different prefix for each virus (i.e. M54 and m90 in MCMV, E54 and e90 in RCMV-E, and R54 and r90 in RCMV-M) and, for some genes, avoided further nomenclatural embellishments. For the majority of genes, this resulted in simplification of the prefix part of the gene name and retention of the numeral part as originally assigned, although some adjustments were necessary to resolve inconsistencies and accommodate previously unrecognized genes.

Most predicted genes are demonstrably conserved among muromegaloviruses. However, there are three highly divergent regions in which orthologous relationships could not be assigned generally and in which paralogous relationships are evident. These are the left terminal region (genes to the left of m18), the right terminal region (genes to the right of m169; or m170 in MCMV), and the m145 region (genes between m144 and m160), which contains most members of the m145 gene family (Table S2). Prefixes other than M or m were assigned to the nomenclature for these regions. In the left terminal region, MnatCMV3 and RCMV-M lack genes, whereas MnatCMV1 and MnatCMV2 have three and four spliced genes (named orthologously a1 to a4), respectively, in the lectin family. In the corresponding region, RCMV-E has a spliced gene (e1.5) in the lectin family and one other gene (e2), and MCMV Smith and MCMV s09 have 16 genes (m02 to m17), most of which belong to a gene family that is present only in MCMV. In the right terminal region, MnatCMV1, MnatCMV2, MCMV Smith and MCMV s09 lack genes, whereas MnatCMV3 has a spliced gene (b169) in the lec family that unusually commences with an exon in the left terminal region and is presumably expressed from circular or concatemeric genomes. In the corresponding region, RCMV-E has a spliced gene (e168) in the lec family and five genes (e169 to e173) in another gene family that is unique to this virus, and RCMV-M has two genes (r169 and r170) in a gene family that is unique to this virus and one other gene (r168). In the m145 region, each virus has an impressive array of genes in the m145 family, which encodes paralogous proteins (Fig. S1) having some similarities to major histocompatibility complex class I molecules [35], plus a few other genes. In this region, it was possible to determine orthology for, and hence assign a common nomenclature to, only the most closely related viruses (MnatCMV1 and MnatCMV2, and MCMV Smith and MCMV s09).

Finally, a few genes in various genomes were identified as being disrupted by frameshifts. These included the lec family genes a2 and a3 in MnatCMV1 and a2 in MnatCMV2, the disruption in each case being due to the length of a polynucleotide tract (a G:C-tract). Bearing in mind that the genomes that were annotated were sequenced from tertiary plaque isolates, it is notable that the disruptions in MnatCMV1 were also present in the corresponding primary isolate. It is also notable that the disruption in MnatCMV2 was present in each of three corresponding independent primary isolates (although the length of the C-tract was different, with ten C residues in the primary isolates and 12 C residues in the tertiary isolate). This finding indicates that a2 was already disrupted in animal 2 before virus isolation. In other strains of these viruses, a2 was disrupted in MnatCMV1 from the primary isolate from animal 29 and in MnatCMV2 from the primary isolates from animals 19 and 33, but not in the primary isolates from animals 18 or 29 or a primary plaque isolate (of MnatCMV2B) from animal 2. Genes disrupted by differences in the lengths of polynucleotide tracts have been described previously in various CMVs, leading to the hypothesis that expression of these genes might be turned on or off in specific circumstances by the selection of tract length variants [36]. In contrast to MnatCMV1 and MnatCMV2, no genes were disrupted in MnatCMV3. Two disrupted genes (m14 and M38) were identified in MCMV s09, but it is not known whether these were due to genuine mutations or sequencing errors.

**Table 5.** Nucleotide substitutions (no.) in pairwise comparisons of MnatCMV2 primary isolate genome sequences\*

	Mnat2A 2-4	Mnat18A	Mnat19A	Mnat29A	Mnat33A
Mnat2A	87	127	2	12	11
Mnat2A 2-4		206	85	93	92
Mnat18A			125	123	122
Mnat19A				10	9
Mnat29A					3
Mnat33A					

\*Rows and columns refer to datasets (see Table 2). Includes one MnatCMV2 primary plaque isolate genome sequence (Mnat2A 2-4). Duplicate results and the results of comparisons of sequences with themselves (0 substitutions) are omitted.

### Phylogeny of MnatCMV primary isolates

The relationships between MnatCMV strains were determined from pairwise alignments of the genome sequences of primary isolates, with the addition of one genome sequence of a primary plaque isolate (from animal 2 in dataset Mnat2A 2-4) to represent MnatCMV2B. The numbers of substitutions were recorded (Tables 5 and 6; indels are not included).

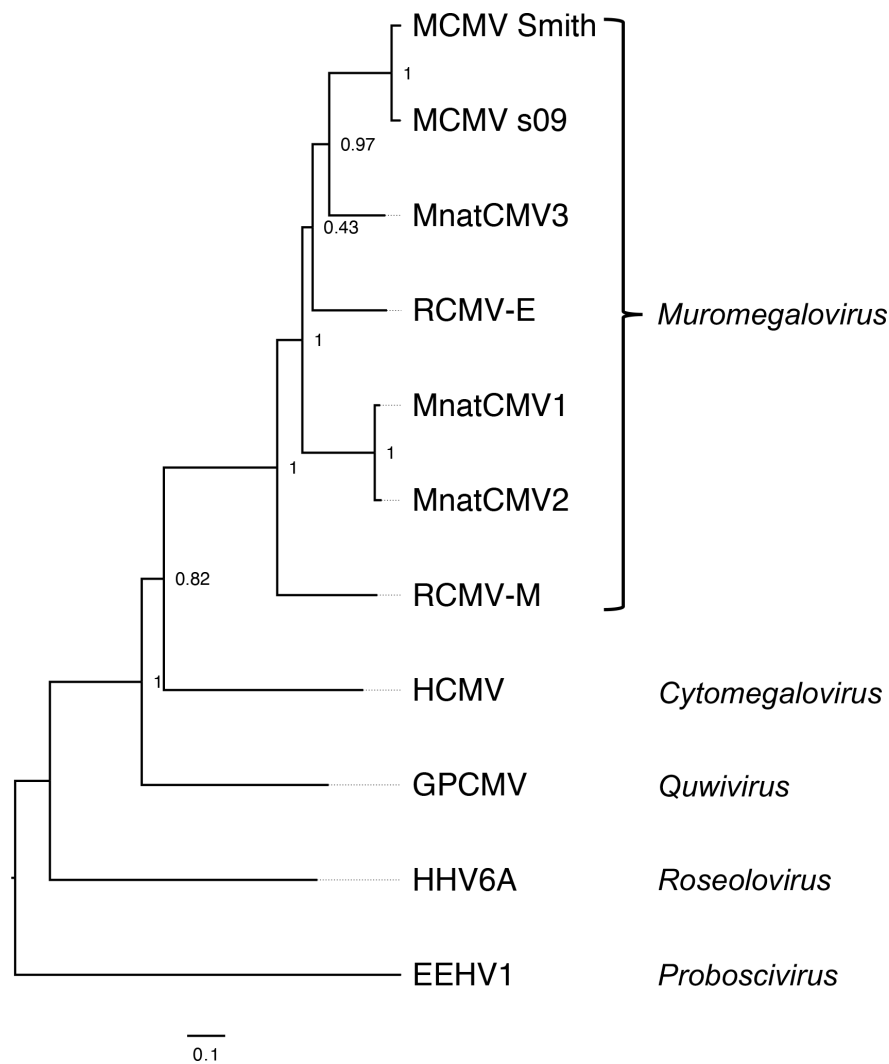
The MnatCMV1 sequences derived from two datasets (Mnat29 and Mnat36A) differed by 209 substitutions distributed throughout the genomes. The MnatCMV3 sequences derived from ten datasets similarly fell into two groups, with those from Mnat1, Mnat13, Mnat28 and Mnat38 differing by 1–3 substitutions and those from Mnat8, Mnat9, Mnat15, Mnat29, Mnat35A and Mnat36A differing by 0–34 substitutions, and those in different groups differing by 307–315 substitutions distributed throughout the genomes (Table 5). The MnatCMV2 sequences derived from six datasets fell into three groups, with those from Mnat2A (representing MnatCMV2A), Mnat19, Mnat29 and Mnat33 differing by 3–12 substitutions (Table 6), the MnatCMV2B sequence from Mnat2A 2-4 differing from these by 85–93 substitutions located in the first one-third of the genomes, and the sequence from Mnat18 differing from those in the MnatCMV2A and MnatCMV2B groups by 122–127 and 206 substitutions, respectively, located predominantly in the terminal regions together comprising one-eighth of the genome. These findings indicate the existence of two lineages each for MnatCMV1 and MnatCMV3, and three lineages for MnatCMV2, with evidence for the involvement of recombination in the evolution of MnatCMV2 strains, as described above.

The phylogeny of the MnatCMVs was also analysed in the context of the available sequences for muromegaloviruses and representative members of other genera in the subfamily *Betaherpesvirinae*. The analysis was conducted in the context of the amino acid sequences specified by six core genes that are used for taxonomic purposes [1] (Fig. 5) and confirmed that the MnatCMVs cluster with other muromegaloviruses and that MnatCMV1 and MnatCMV2 are more closely related to each other than they are to MnatCMV3.

**Table 6.** Nucleotide substitutions (no.) in pairwise comparisons of MnatCMV3 primary isolate genome sequences\*

	Mnat8A	Mnat9A	Mnat13A	Mnat15A	Mnat28A	Mnat29A	Mnat35A	Mnat36A	Mnat38A
Mnat1A	314	307	1	309	2	309	308	312	2
Mnat8A		34	315	12	315	13	12	2	315
Mnat9A			308	26	307	27	26	32	307
Mnat13A				310	3	310	309	313	3
Mnat15A					311	5	0	10	311
Mnat28A						310	310	313	0
Mnat29A							5	11	310
Mnat35A								10	310
Mnat36A									313

\*Rows and columns refer to datasets (see Table 2). Duplicate results and the results of comparisons of sequences with themselves (0 substitutions) are omitted.



**Fig. 5.** Maximum-likelihood phylogenetic tree showing relationships among MnatCMVs and related viruses in the subfamily *Betaherpesvirinae*. The names of the genera in which the viruses group are shown on the right. Fractional bootstrap values are shown at the nodes. Bar, 0.1 amino acid substitutions per site.

## DISCUSSION

This study is preceded by an earlier published analysis of *M. natalensis* animals from Côte d'Ivoire and Mali [2], in which four MnatCMVs (MnatCMV1, MnatCMV2, MnatCMV3 and MnatCMV4) were identified in tissue samples by sequencing short PCR products. We have extended these findings significantly by isolating infectious MnatCMV1, MnatCMV2 and MnatCMV3 from salivary gland tissue from a cohort of these animals from Mali and determining their genome sequences. The proportion of animals positive for MnatCMVs by virus isolation was 36%, which was higher than the 10% detected by PCR of spleen tissue from this cohort, and the 22% detected by PCR of spleen or lung tissue from all 175 animals tested from both countries in the earlier work [2]. This high proportion of isolation probably underrepresents the proportion of animals infected by MnatCMVs, as it is dependent on the presence in the samples of actively replicating virus rather than latent virus. Indeed, it is probable that MnatCMVs, like other more extensively studied CMVs in their respective mammalian hosts [37, 38], are present at near ubiquitous levels in *M. natalensis*.

Alignment of the datasets from the primary isolates to the sequences of the short PCR products showed that most MnatCMV-positive animals contained a single MnatCMV (MnatCMV1, MnatCMV2 or MnatCMV3, but not MnatCMV4). However, one animal contained two strains of a single MnatCMV (MnatCMV2), and two animals contained multiple MnatCMVs (MnatCMV1 and MnatCMV3 in one, and MnatCMV1, MnatCMV2 and MnatCMV3 in the other). These findings were

confirmed by determining the genome sequences of the isolates, which indicated that each MnatCMV was present as a single strain, with the exception of the two MnatCMV2 strains in one animal. These results, which mirror experience with CMVs in other hosts [37, 39–41], showed that some animals had acquired multiple MnatCMVs or MnatCMV strains, presumably by coinfection or superinfection, and are consistent with the established ability of CMVs to reinfect their host regardless of prior CMV infection status and CMV-directed immunity [40, 42].

Datasets were then produced from primary and tertiary plaque isolates from three animals representing MnatCMV1, MnatCMV2 (from the animal having two strains) and MnatCMV3. The MnatCMV1 and MnatCMV3 plaque isolates were identical or almost identical in sequence to the primary isolates. However, the more complex situation with the MnatCMV2 plaque isolates indicated the occurrence of recombination between the two strains before or during isolation, and also the prior generation of one of these strains by recombination *in vivo*. The genome sequence of one tertiary plaque isolate for each of the three MnatCMVs was completed by incorporating long-read data to resolve reiterated regions.

A phylogenetic tree based on six core genes showed that the three MnatCMVs grouped with other muromegaloviruses and therefore merit classification in the genus *Muromegalovirus*. A maximum level of nucleotide sequence identity for this core gene set for separating viruses into two species has not been defined for the family *Orthoherpesviridae*, but can be as high as 97% for viruses that have already been classified into separate species. MnatCMV3 is therefore clearly sufficiently different from other muromegaloviruses to be classified as a new species. There is also a strong case to be made for classifying MnatCMV1 and MnatCMV2 as separate new species despite their relatively close relationship to each other, as the core gene set exhibits only 87% identity. Pairwise alignments of the genome sequences of primary isolates demonstrated the existence of two or three lineages for each MnatCMV, alongside additional evidence for the involvement of recombination in the evolution of MnatCMV2 strains. Different strains of each MnatCMV differed by approximately  $10^{-3}$  substitutions/nt, which, in the context of the estimated divergence rates of approximately  $3 \times 10^{-8}$  substitutions/nt/year for herpesviruses [43–45], implies that they diverged many thousands of years ago. Beyond MnatCMVs, a case can also be made for classifying MCMV s09 as a new species rather than within the established species *Muromegalovirus muridbeta 1* alongside MCMV Smith, as these viruses have different hosts and the core gene set exhibits only 93% identity. However, this latter assignment may require additional sequence information, as the data from which the MCMV s09 genome sequence was derived reportedly exhibited heterogeneity due to the presence of multiple strains of the virus in the animal tested [41].

Our study has provided a unique picture of the muromegaloviruses present within wild *M. natalensis* rodents, which are of particular interest because they are known to be the major source of zoonotic transmission of LASV into human populations in West Africa. This picture has proven to be richly diverse even among a relatively small number of animals captured in a single village in southern Mali, and is likely to be enhanced substantially by extending the work to include populations of *M. natalensis* across a wider area. The restriction of individual CMVs to particular mammalian host species, combined with high immunogenicity and lack of associated pathology, has prompted the development of CMVs as recombinant, species-specific vaccine vectors against multiple infectious diseases [46–48]. Our study has taken us to the point of being able to construct fully characterized infectious wild-type MnatCMV clones for development as a vaccine platform for controlling LASV in the *M. natalensis* reservoir rodent species [39]. Moreover, our observation that multiple replicating MnatCMVs may reside within single animals parallels experience with other CMVs, and suggests that prior immunity does not present a hurdle even in a target host population in which CMVs are highly prevalent.

#### Funding information

The work described in this article was funded by DARPA grant no. D18AC00028 in support of the PREventing EMerging Pathogenic Threats (PREEMPT) programme. Information on the University of California, Davis PREEMPT Project Consortium, including a list of Consortium members, is available online at <https://www.preemptproject.org>. Field work in Mali was funded by the NIAID International Centre for Excellence in Research (ICER), and laboratory work at the NIAID (Hamilton, MT, USA) was supported financially by the NIAID Intramural Research Programme. The funders had no role in study design, data collection and analysis, preparation of the manuscript, or decision to publish.

#### Acknowledgement

We thank the NIAID International Centre for Excellence in Research (ICER) in Mali for its support of field work. We also appreciate Prof. Stephan Günther (Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany) for his support with development of the *M. natalensis* cell lines.

#### Author contributions

Conceptualization: A.R., W.B., M.J., A.D. Data curation: M.V., J.H., A.D. Formal analysis: M.V., S.C., A.D. Funding acquisition: W.B., A.R., A.D., M.J. Investigation: F.H., M.V., J.N., S.C., E.O., L.S., B.C., T.M., K.R., M.J., A.D. Methodology: F.H., M.V., S.C., E.O., W.B., K.R., M.J., A.D. Project administration: A.R., W.B., M.J., A.D. Resources: S.B., A.F., N.S., D.S., S.V., B.E., H.F. Software: M.V., J.H. Supervision: J.H., A.F., A.R., H.F., W.B., K.R., M.J., A.D. Validation: M.J., A.D. Visualization: F.H., M.J., A.D. Writing – original draft: M.V., J.N., J.H., E.O., A.R., W.B., M.J., A.D. Writing – review & editing: all authors

#### Conflicts of interest

The authors declare that there are no conflicts of interest.

#### Ethical statement

Generation of MasKECs from *M. natalensis* rodents was performed at the Leibniz Institute of Virology according to the recommendations and guidelines of the Federation for Laboratory Animal Science Associations (<https://felasa.eu/>) and the Gesellschaft für Versuchstierkunde (Society of Laboratory

Animal Science; <https://www.gv-solas.de/>), which were approved by the institutional review board and local authorities (Behörde für Gesundheit und Verbraucherschutz, Amt für Verbraucherschutz, Freie und Hansestadt Hamburg, reference 2018-08-16-02). Information on the trapping of animals and preparation of tissues used for isolating MnatCMVs was published previously [2]. The tissues were stored at the National Institute of Allergy and Infectious Diseases (Hamilton, MT, USA) under biosafety level 2+ conditions. Experiments involving isolating MnatCMVs were approved under biosafety level 2 conditions using standard operating protocols approved by the institutional biosafety committee.

## References

- Gatherer D, Depledge DP, Hartley CA, Szpara ML, Vaz PK, et al. ICTV Virus Taxonomy Profile: *Herpesviridae* 2021. *J Gen Virol* 2021;102:001673.
- Calvignac-Spencer S, Kouadio L, Couacy-Hymann E, Sogoba N, Rosenke K, et al. Multiple DNA viruses identified in multimammate mouse (*Mastomys natalensis*) populations from across regions of sub-Saharan Africa. *Arch Virol* 2020;165:2291–2299.
- Ehlers B, Küchler J, Yasmum N, Dural G, Voigt S, et al. Identification of novel rodent herpesviruses, including the first gammaherpesvirus of *Mus musculus*. *J Virol* 2007;81:8091–8100.
- Hinte F, van Anken E, Tirosh B, Brune W. Repression of viral gene expression and replication by the unfolded protein response effector XBP1u. *Elife* 2020;9:e51804.
- Brinkmann NM, Hoffmann C, Wurr S, Pallasch E, Hinzmann J, et al. Understanding host-virus interactions: assessment of innate immune responses in *Mastomys natalensis* cells after arenavirus infection. *Viruses* 2022;14:1986.
- Xu H, Luo X, Qian J, Pang X, Song J, et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* 2012;7:e52249.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
- Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–359.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, et al. Using tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 2013;14:193–202.
- Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* 2017;3:e000132.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–3100.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;27:722–736.
- Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015;12:733–735.
- Schmieder R, Lim YW, Edwards R. Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* 2012;28:433–435.
- Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;27:863–864.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21.
- Bonfert T, Kirner E, Csaba G, Zimmer R, Friedel CC. ContextMap 2: fast and accurate context-based RNA-seq mapping. *BMC Bioinformatics* 2015;16:122.
- Wilkie GS, Davison AJ, Watson M, Kerr K, Sanderson S, et al. Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases. *J Virol* 2013;87:6700–6712.
- Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019;37:420–423.
- Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001;305:567–580.
- Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol* 2008;4:e1000213.
- Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, et al. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res* 2009;37:D380–6.
- Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* 2020;48:D265–D268.
- Dolan A, Cunningham C, Hector RD, Hassan-Walker AF, Lee L, et al. Genetic content of wild-type human cytomegalovirus. *J Gen Virol* 2004;85:1301–1312.
- Yang D, Tamburro K, Dittmer D, Cui X, McVoy MA, et al. Complete genome sequence of pathogenic guinea pig cytomegalovirus from salivary gland homogenates of infected animals. *Genome Announc* 2013;1:e0005413.
- Gompels UA, Nicholas J, Lawrence G, Jones M, Thomson BJ, et al. The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. *Virology* 1995;209:29–51.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evol* 2018;35:1547–1549.
- Stecher G, Tamura K, Kumar S. Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol Biol Evol* 2020;37:1237–1239.
- Le SQ, Gascuel O. An improved general amino acid replacement matrix. *Mol Biol Evol* 2008;25:1307–1320.
- Gatherer D, Seirafian S, Cunningham C, Holton M, Dargan DJ, et al. High-resolution human cytomegalovirus transcriptome. *Proc Natl Acad Sci U S A* 2011;108:19755–19760.
- Rawlinson WD, Farrell HE, Barrell BG. Analysis of the complete DNA sequence of murine cytomegalovirus. *J Virol* 1996;70:8833–8849.
- Revilleza MJ, Wang R, Mans J, Hong M, Natarajan K, et al. How the virus outsmarts the host: function and structure of cytomegalovirus MHC-I-like molecules in the evasion of natural killer cell surveillance. *J Biomed Biotechnol* 2011;2011:724607.
- Davison AJ, Holton M, Dolan A, Dargan DJ, Gatherer D, et al. *Cytomegaloviruses: From Molecular Pathogenesis to Intervention*, Reddehase MJ. UK: Caister Academic Press; 2013.
- Booth TW, Scalzo AA, Carrello C, Lyons PA, Farrell HE, et al. Molecular and biological characterization of new strains of murine cytomegalovirus isolated from wild mice. *Arch Virol* 1993;132:209–220.
- Becker SD, Bennett M, Stewart JP, Hurst JL. Serological survey of virus infection among wild house mice (*Mus domesticus*) in the UK. *Lab Anim* 2007;41:229–238.
- Murphy AA, Redwood AJ, Jarvis MA. Self-disseminating vaccines for emerging infectious diseases. *Expert Rev Vaccines* 2016;15:31–39.
- Ross SA, Arora N, Novak Z, Fowler KB, Britt WJ, et al. Cytomegalovirus reinfections in healthy seroimmune women. *J Infect Dis* 2010;201:386–389.
- Čížková D, Baird SJE, Těšíková J, Voigt S, Ludovít Ď, et al. Host subspecific viral strains in European house mice: Murine

- cytomegalovirus in the Eastern (*Mus musculus musculus*) and Western house mouse (*Mus musculus domesticus*). *Virology* 2018;521:92–98.
42. Hansen SG, Powers CJ, Richards R, Ventura AB, Ford JC, *et al.* Evasion of CD8+ T cells is critical for superinfection by cytomegalovirus. *Science* 2010;328:102–106.
  43. Norberg P, Tyler S, Severini A, Whitley R, Liljeqvist J-Å, *et al.* A genome-wide comparative evolutionary analysis of herpes simplex virus type 1 and varicella zoster virus. *PLoS One* 2011;6:e22527.
  44. Sakaoka H, Kurita K, Iida Y, Takada S, Umene K, *et al.* Quantitative analysis of genomic polymorphism of herpes simplex virus type 1 strains from six countries: studies of molecular evolution and molecular epidemiology of the virus. *J Gen Virol* 1994;75 (Pt 3):513–527.
  45. McGeoch DJ, Cook S. Molecular phylogeny of the Alphaherpesvirinae subfamily and a proposed evolutionary timescale. *J Mol Biol* 1994;238:9–22.
  46. Hansen SG, Ford JC, Lewis MS, Ventura AB, Hughes CM, *et al.* Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. *Nature* 2011;473:523–527.
  47. Marzi A, Murphy AA, Feldmann F, Parkins CJ, Haddock E, *et al.* Cytomegalovirus-based vaccine expressing Ebola virus glycoprotein protects nonhuman primates from Ebola virus infection. *Sci Rep* 2016;6:21674.
  48. Tsuda Y, Parkins CJ, Caposio P, Feldmann F, Botto S, *et al.* A cytomegalovirus-based vaccine provides long-lasting protection against lethal Ebola virus challenge after a single dose. *Vaccine* 2015;33:2261–2266.
  49. Smith MG. Propagation of salivary gland virus of the mouse in tissue cultures. *Proc Soc Exp Biol Med* 1954;86:435–440.
  50. Priscott PK, Tyrrell DA. The isolation and partial characterisation of a cytomegalovirus from the brown rat, *Rattus norvegicus*. *Arch Virol* 1982;73:145–160.
  51. Ettinger J, Geyer H, Nitsche A, Zimmermann A, Brune W, *et al.* Complete genome sequence of the english isolate of rat cytomegalovirus (*Murid herpesvirus 8*). *J Virol* 2012;86:13838.
  52. Bruggeman CA, Meijer H, Dormans PH, Debie WM, Grauls GE, *et al.* Isolation of a cytomegalovirus-like agent from wild rats. *Arch Virol* 1982;73:231–241.
  53. Vink C, Beuken E, Bruggeman CA. Complete DNA sequence of the rat cytomegalovirus genome. *J Virol* 2000;74:7656–7665.

### Five reasons to publish your next article with a Microbiology Society journal

1. When you submit to our journals, you are supporting Society activities for your community.
2. Experience a fair, transparent process and critical, constructive review.
3. If you are at a Publish and Read institution, you'll enjoy the benefits of Open Access across our journal portfolio.
4. Author feedback says our Editors are 'thorough and fair' and 'patient and caring'.
5. Increase your reach and impact and share your research more widely.

Find out more and submit your article at [microbiologyresearch.org](https://microbiologyresearch.org).