# CIDER-Seq: unbiased virus enrichment and single-read, full length genome sequencing

Devang Mehta*, Matthias Hirsch-Hoffmann, Andrea Patrignani[1], Wilhelm Gruissem and Hervé Vanderschuren*[#]

Institute of Molecular Plant Biology, Department of Biology, and [1]Functional Genomics Center Zurich, ETH Zurich, 8092 Zurich, Switzerland

[#]Current address: AgroBioChem Department, University of Liège, Gembloux, Belgium

*Corresponding authors: devang@ethz.ch & herve.vanderschuren@ulg.ac.be

## Abstract

Deep-sequencing of virus isolates using short-read sequencing technologies is problematic because the viruses are often present in populations with high sequence similarity. We present a new method for generating single-read, full-length virus genomes by combining an improved Random Circular Amplification-based virus enrichment protocol with Single Molecule Real Time-sequencing and a new sequence de-concatenation method. We demonstrate CIDER-Seq by producing more than 250 full-length geminivirus genomes from symptomatic field-grown plants.

Advances in sequencing technologies have produced extensive data about viral diversity and facilitated the identification of previously unknown viruses. The abundance of new virus sequence data recently led to the consensus that virus taxonomy should be revised to incorporate metagenomic sequence data[1]. However, using high-throughput sequencing technologies for virus detection and accurate identification remains difficult because of the risk of assembling artificial chimeric viral genomes from short sequence reads.

Sequence-bias introduced during virus enrichment is equally problematic for deep-sequencing virus genomes. Amplification enrichment methods rely on either Polymerase Chain Reaction (PCR) with primers designed to bind conserved sequences in the genome, or random circular amplification (RCA) utilizing the unique properties of Phi29 DNA polymerase and random-nucleotide primers[2]. PCR-based methods can result in a biased amplification of viral templates because of differing primer-template affinities[3]. Random primer-based RCA is less prone to primer complementarity bias but results in hyper-branched, high molecular weight concatenated products[4] that must be linearized with restriction enzymes (REs)[5] or mechanical shearing prior to Sanger or next generations sequencing (NGS). Using REs requires prior information on conserved RE sites and therefore results in the loss of viral sequences that may have none or multiple recognition sites for the selected REs. An alternative method called “polymerase cloning” or “ploning”[6] employs endonucleases and DNA repair enzymes to linearize RCA products (hence voiding the need for virus sequence information or specific restriction enzyme cut sites), followed by shotgun sequencing and whole genome assembly.

CIDER-Seq (Circular DNA Enrichment Sequencing) is a sequence-independent viral nucleic acid enrichment method utilizing assembly-free Single Molecule Real Time (SMRT; Pacific Biosciences Inc.) long-read sequencing (Fig. 1a). Our enrichment method requires only an estimate of virus genome size and, optionally, a single refere nce sequence. CIDER-Seq also enables the direct single molecule sequencing of RCA-derived products. We have also developed a new algorithm, DeConcat, to parse concatenated DNA strands that are generated by RCA reactions, allowing us

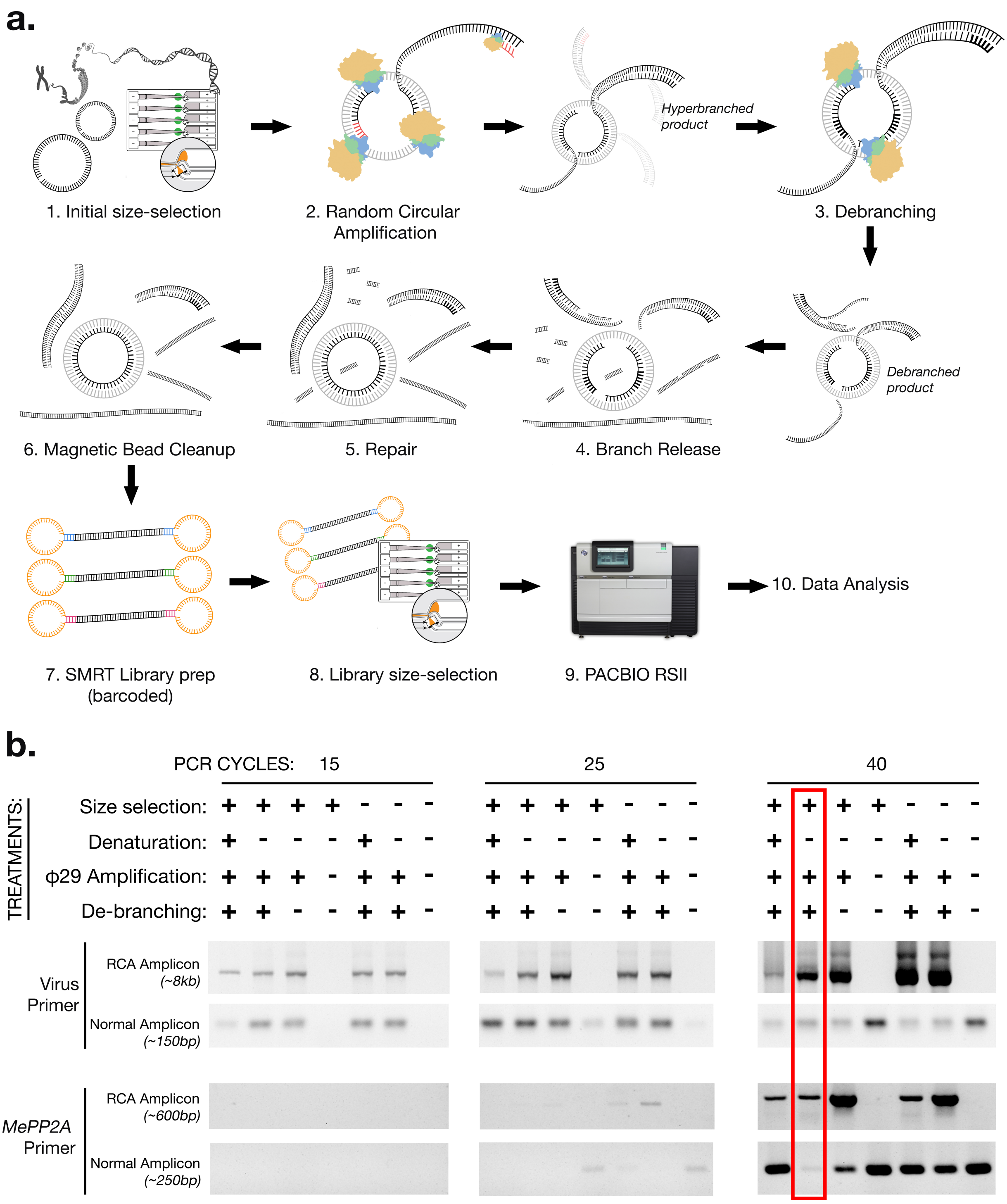

a.
1. Initial size-selection
2. Random Circular Amplification
Hyperbranched product
3. Debranching
Debranched product
4. Branch Release
5. Repair
6. Magnetic Bead Cleanup
7. SMRT Library prep (barcoded)
8. Library size-selection
9. PACBIO RSII
10. Data Analysis
b.
PCR CYCLES: 15
25
40
TREATMENTS:
Size selection: + + + + - - - + + + + - - - + + + + - - -
Denaturation: + - - - + - - + - - - + - - + - - - + - -
ϕ29 Amplification: + + + - + + - + + + - + + - + + + - + + -
De-branching: + + - - + + - + + - - + + - + + - - + + -
Virus Primer
RCA Amplicon (~8kb)
Normal Amplicon (~150bp)
MePP2A Primer
RCA Amplicon (~600bp)
Normal Amplicon (~250bp)

**Figure 1:** Circular DNA enrichment scheme and validation. **(a)** Enrichment of circular DNA based on automated size selection, non-denaturing random circular amplification (RCA) and linearization and repair of the RCA product followed by Single Molecule Real Time (SMRT) library creation. SMRT libraries were optionally size-selected prior to sequencing on a PacBio RS II instrument. **(b)** Semi-quantitative PCR on a sample testing permutations of the three enrichment steps. *Manihot esculenta PROTEIN PHOSPHATASE 2A (MePP2A)*-specific primers were used to determine the amount of linear cassava genomic DNA compared to enriched circular viral DNA amplified with cassava geminivirus-specific primers.

to sequence complete virus genomes without reference-based or *de novo* assembly. To validate our approach, we sequenced the cassava mosaic geminivirus (CMG), which causes cassava mosaic disease (CMD) and severe economic losses for farmers, particularly in sub-Saharan Africa. To date, nine CMG species have been identified that share 68% to 90% sequence identity based on Sanger sequencing of PCR and RCA products[7]. CMGs are bipartite viruses of the genus

*Begomovirus* in the family *Geminiviridae*—the most populous family of eukaryote-infecting viruses[8]. They are comprised of two separate genomes (designated DNA A and DNA B)[9], and thus serve as a good model to validate the functionality of our method for segmented viruses (Fig. 1a, Steps 1-8).

Using samples from cassava plants with CMD symptoms we first enriched complete DNA genomes of CMGs by automated size selection of DNA molecules in the 2.8 kb size range (Step 1). Phi29 DNA polymerase RCA of the selected DNAs was carried out using random hexameric primers (Step 2). Using a 'ploning' based protocol, we amplified the products of the initial amplification in an additional primer-free Phi29 DNA polymerase reaction to further "de-branch" the hyper-branched DNA[6] (Step 3). Next, the hyper-branched structure was resolved using ssDNA-digesting S1 Nuclease (Step 4) and repaired with DNA Polymerase I and T4 DNA Polymerase (Step 5). The linear DNA fragments were purified using magnetic beads (Step 6), which excluded small DNA fragments produced during the RCA steps.

To optimize the enrichment, we performed a semi-quantitative PCR (semi-qPCR) with 15, 25 and 40 cycles to quantify viral DNA titers relative to host DNA before and after Steps 1-3 (Fig. 1b). The semi-qPCR revealed viral and host amplicon DNA bands at the expected size but also longer RCA-specific DNA bands. Maximum enrichment was obtained after Steps 1-3 (Fig. 1b, highlighted) and avoidance of the denaturation step common to most RCA protocols. The debranching step catalyzed by S1 nuclease also reduced the abundance of genomic DNA in favor of a more distinct viral RCA DNA product.

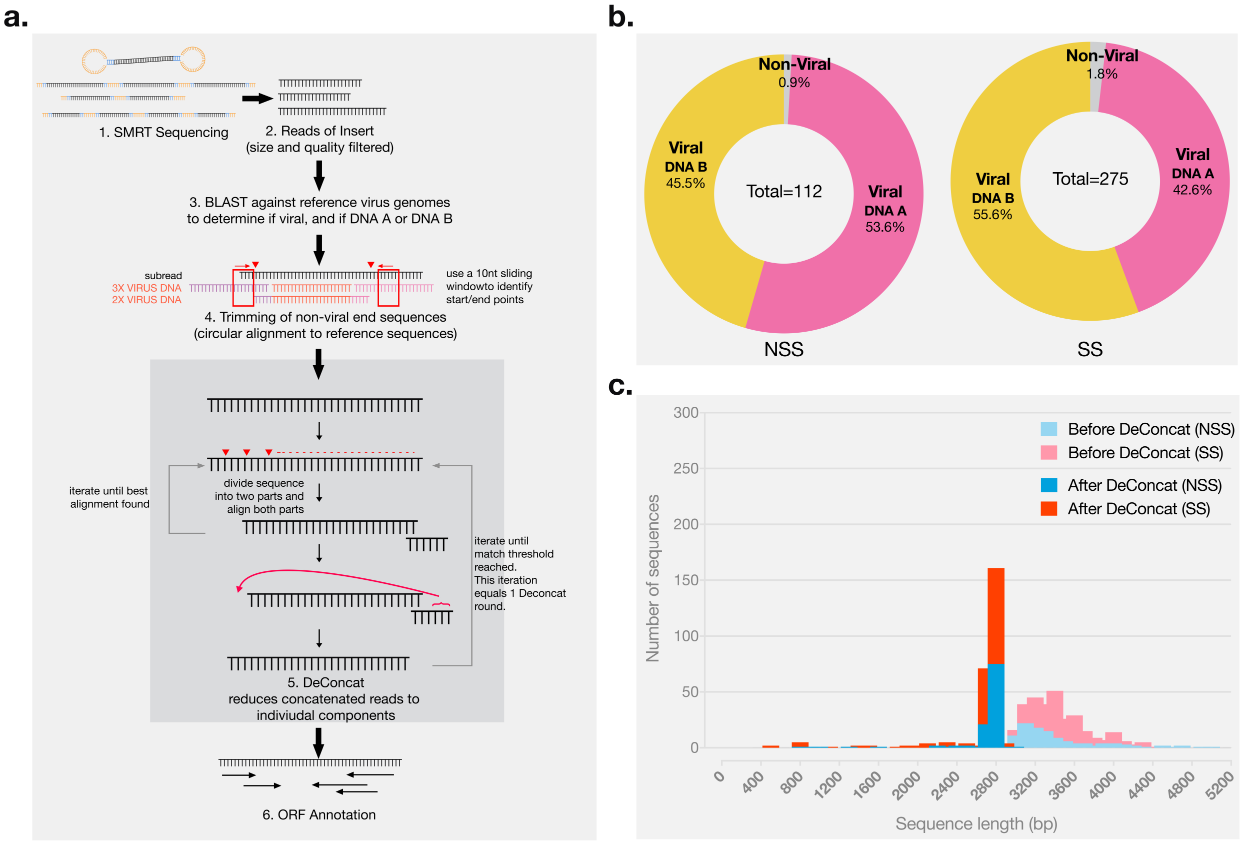


Figure 2: Data analysis method and results. (a) Overview of the data analysis method including a simplified depiction of the DeConcat algorithm. (b) Distribution of the BLAST-based binned sequencing reads from the non-size selected (NSS) and size selected (SS) libraries. (c) Length distribution of sequencing reads before and after DeConcat for NSS and SS libraries.

The enrichment was applied to total DNA extracts from six independent leaf samples collected from a cassava field trial in Kenya. The six samples were barcoded and pooled to build a SMRT sequencing library. We sequenced the library in two independent sequencing runs:  an intact library without size selection (NSS), and a size-selected (SS) library (to maximize the number of reads obtained from inserts greater than 3 kb) (Step 8, Fig. 1a). We then filtered the raw sequencing reads for high-quality insert sequences (predicted minimum quality of 99.9 and minimum insert length of 3000 bp). The SS library produced 275 high quality reads of insert (ROIs) of the expected size and quality range, while the NSS library produced only 112. As expected, a majority of ROIs were greater than the 3 kb size cut-off, indicating that the inserts were comprised of concatenated RCA products.

We developed a custom data analysis pipeline to resolve the concatenated sequences into their component parts (Fig. 2a). First, in a basic filtering step all ROIs were binned into non-viral, CMG DNA A and CMG DNA B sequences. The results from this step indicated that approximately 99% of the ROIs in both libraries were found in the CMG bins (Fig. 2b). Next, we trimmed non-viral DNA from the ends of each ROI by performing a circular DNA–aware multiple sequence alignment (see Methods), which did not significantly affect the size distribution of the dataset. These trimmed ROIs were next passed through the de-concatenation algorithm we termed 'DeConcat' (Supplementary Fig. 1, see Methods). The resulting de-concatenated ROIs had a greatly reduced size distribution (Fig. 2c), with a clear peak at 2.8 kb. Thus, DeConcat was able to reduce RCA-generated SMRT-sequenced ROIs of a wide size-distribution into expected CMG-length virus sequences. This result validated the parameters and scoring formula applied in DeConcat and also demonstrated that the Phi29 DNA polymerase enrichment step amplified circular DNA in the 3 kb range with high fidelity.

We performed tBLASTn using the final DeConcat reads against virus protein reference sequences to annotate reads belonging to viral ORFs. The results revealed frameshift mutations in one or more ORFs in several reads. Frameshift mutations are caused by insertions or deletions—the most frequent type of errors in SMRT sequencing[10]. To distinguish if these frameshifts were of biological origin or SMRT-sequencing errors, we relaxed the minimum quality threshold of 99.9% for ROIs in the initial filtering step to 99.5% and 99%. When comparing the frequency and number of frameshift mutations between the three quality thresholds we found that the number of frameshifts per sequence increased with decreasing quality thresholds (Supplementary Fig. 2a). and the frequency of frameshifts in each viral protein increased linearly with protein length (Supplementary Fig. 2b), indicating that they are likely caused by SMRT sequencing errors. Between 5 to 8% of reads in the >99.99 quality threshold dataset have no frameshift errors (Supplementary Fig. 2c).

We next selected DNA A sequences from the NSS and SS datasets that were longer than 2.7 kb (and therefore likely complete virus sequences) for pairwise alignments using the SDT virus classification software[11] to analyze the diversity of the CMG sequences. Most DNA sequences belonged to two categories with approximately 75% identity (Fig. 3a, arrows), while sequences within each category were >91% identical. This was also reflected in the frequency distribution of pairwise sequence identities calculated by SDT (Fig. 3b). Since the recently revised *Begomovirus* taxonomy guideline[12] classifies sequences with less than 91% identity as distinct species, we conclude that CIDER-Seq identifies two distinct CMG species in our samples. We constructed a Maximum Likelihood phylogenetic tree including published sequences of the nine reported CMG species found on the African continent[7] to determine the species that most closely match our dataset. The cladogram shows that the two major species we detected in the field samples using

SDT belong to the *African cassava mosaic virus* (ACMV) and the *East African cassava mosaic virus* (EACMV) clades (Fig. 3c).

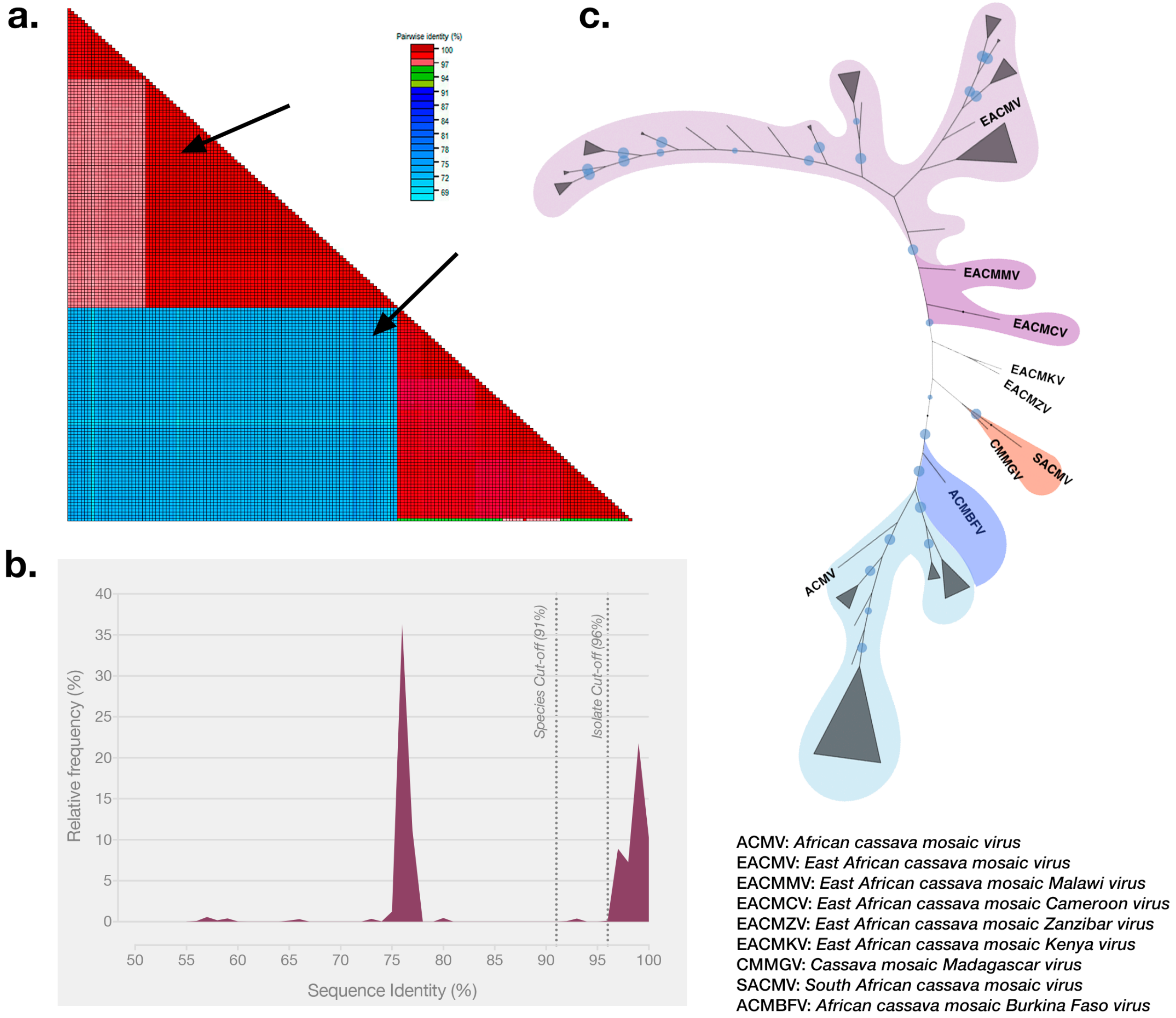


**Figure 3:** Sequence demarcation and phylogenetic analysis of the cassava mosaic geminivirus (CMG) DNA A dataset. **(a)** Pairwise identity matrix of all full length DNA A sequences (>2700 bp) generated using CIDER-seq. Three species groups are marked by arrows. **(b)** Frequency plot of pairwise identity (DNA A sequences) showing the taxonomic cut-offs for species and isolate demarcation. **(c)** Unrooted cladogram constructed from a maximum likelihood tree of all DNA A sequences with reference DNA A genomes of the 9 CMG species found on the African continent. Coloured clades mark sequences closest to respective reference species. Nodes marked in blue are supported by >70 bootstraps and the node sizes are scaled from 70-100 bootstrap values.

Together, CIDER-seq effectively enriches circular DNA molecules and produces single-read, full-length DNA sequence data from the RCA reaction products. The DeConcat algorithm parses the concatenated reads of the RCA products into individual component DNA sequences of the appropriate size range without training the algorithm with prior information of desired sequence length. Considering the popularity of Phi29 DNA polymerase-based amplification methods such as MDA or MALBAC in single-cell genome sequencing protocols[13], DeConcat can increase analysis quality of concatenated sequences from single-cell DNA samples as well.

We have validated CIDER-seq for deep-sequencing of *Geminiviruses*, which is an important family of plant-infecting viruses that is causing increasing economic losses to farmers[14]. Using infected field material we could generate 270 high-quality, non-chimeric full-length virus genome sequences, representing nearly 50% of all CMG sequences deposited in GenBank to date. CIDER-Seq can also be applied to other important viruses with similar genome sizes and topology, including e.g. *Porcine circoviruses*, *Chicken anemia virus,* and the recently discovered, ubiquitous, human-infecting *Torque teno virus*. The circular dsDNA *Human papillomavirus* is another important target for future CIDER-Seq experiments. The sequencing of non-viral circular DNA templates such as plasmids is another potential application of CIDER-Seq, particularly in the context of clinical deep-sequencing of antibiotic resistance gene carrying plasmids[15]. In addition to circular DNA templates, single molecule sequencing of linear DNA and RNA viruses could also be facilitated by DeConcat in combination with size selection, Phi29 polymerase amplification and SMRT sequencing. Based on the current accuracy of long-read sequencing technologies CIDER-Seq produces results with an estimated error rate that is far below virus species and isolate demarcation thresholds. We also note that an error rate of 0.1% is comparable to the Q30 threshold of Illumina short reads. Further improvements of either SMRT read-length or single-read quality will increase the number of full-length viral genomes and expand our method to larger viruses. Considering recent calls to incorporate metagenomics data in virus classification[1], CIDER-Seq is a superior method for high-quality full-genome sequencing of viruses infecting plants, animals and bacteria that will facilitate building accurate sequence datasets for virus taxonomy and evolutionary studies.

## Methods

### Sampling and DNA extraction

Mature symptomatic cassava leaves were harvested from infected plants grown for nine months in a confined field trial in Alupe, Kenya. The plants included cassava genotypes 60444 and TME14 as well as transgenic lines in the same genotypic backgrounds. Total nucleic acid was extracted from leaf samples pooled from three plants of each genotype. Extraction was performed using a CTAB (cetyl trimethylammonium bromide) protocol[16] combined with an ethanol precipitation step. Total nucleic acid was quantified using a Qubit dsDNA BR Assay Kit (Q32850, Thermo Scientific).

### Size Selection

For the pre-enrichment size-selection step, 5 µg of total nucleic acid was loaded on a 0.75% agarose gel cassette and separated on a BluePippin instrument (SAGE Science). DNA fragments between 0.8-5kb were extracted. This size range was selected because geminivirus DNA is present as dsDNA replicative intermediates (running between 2 and 3 kb) and ssDNA mature forms, which migrate at lower size ranges on agarose gels. Post-enrichment size-selection was similarly performed and fragments >3 kb were extracted.

### Random Circle Amplification

Random rolling circle amplification was performed as previously described[2] with some modifications. A 20 µl reaction was set up using 5 µl of size-selected template DNA, 1mM dNTPs, 10U Phi29 DNA polymerase (EP0092, Thermo Scientific), 50 µM Exo-resistant random primer (SO181, Thermo Scientific), 0.02U inorganic pyrophosphatase (EF0221, Thermo Scientific) and 1X Phi29 DNA polymerase buffer (supplied with enzyme). The reaction was run at 30 °C for 18 hours and stopped by heating to 65 °C for 2 minutes. Product DNA was purified by sodium acetate/ethanol precipitation. We also used the illustra TempliPhi 100 amplification kit (25640010, GE Life Sciences) and obtained similar amplification results.

### Phi29 debranching

10 μg of amplified DNA was used in a debranching reaction with 5U of Phi29 DNA polymerase without a primer at 30 °C for 2 hours and stopped by heating at 65 °C for 2 minutes. The product was precipitated with sodium acetate/ethanol. The purified product was treated with 50U S1 nuclease (EN0321, Thermo Scientific) in a 20 μl reaction at 37 °C for 30 minutes and stopped by adding 3.3 μl of 0.5M EDTA and heating at 70 °C for 10 minutes. DNA was purified by sodium acetate/ethanol precipitation.

## DNA repair

De-branched DNA was treated with 3U T4 DNA polymerase (M0203L, New England Biolabs) and 10U *E. coli* DNA polymerase I (M0209L, New England Biolabs) with 1X NEBuffer 2 and 1mM dNTPs in a 50 μl reaction. The reaction was incubated at 25 °C for 1 hour and stopped by heating at 75 °C for 20 minutes. After cooling, 5U of Alkaline Phosphatase (EF0651, Thermo Scientific) was added. Dephosphorylation was conducted at 37 °C for 10 minutes and stopped by heating to 75 °C for 5 minutes. The repaired DNA was purified using KAPA Pure Beads (KK8000, Kapa Biosystems) at a 1.5X volumetric ratio and quantified using a Qubit dsDNA BR Assay Kit (32850, Thermo Scientific).

## Semi-quantitative PCR

Semi-quantitative PCR was performed using the primers described in Supplementary Table 2. DreamTaq polymerase (EP0705, Thermo Scientific) was used to amplify 10 ng of template in 50 μl reactions set for 15, 25 and 40 cycles each. PCR products were separated using a 1% agarose gel in 1X sodium borate acetate buffer and visualised by staining with ethidium iodide.

## SMRT barcoding, library preparation and sequencing

A Bioanalyzer 2100 12K DNA Chip assay (5067-1508, Agilent) was used to assess fragment size distribution of the enriched DNA samples. The sequencing libraries were produced using the SMRTBell$^{TM}$ Barcoded Adapter Complete Prep Kit - 96, following manufacturer's instructions (100-514-900. Pacific Biosciences). Approximately 200 ng of each DNA sample was end-repaired using T4 DNA Polymerase and T4 Polynucleotide Kinase according to the protocol supplied by Pacific Biosciences. A PacBio barcoded adapter was added to each sample via a blunt end ligation reaction. The 9 samples were then pooled together and treated with exonucleases in order to create a SMRT bell template. A Blue Pippin device (Sage Science) was used to size select one aliquot of each barcoded library to enrich the larger fragments >3 kb. Both the non-size selected and the size selected library fractions were quality inspected and quantified on the Agilent Bioanalyzer 12Kb DNA Chip and on a Qubit Fluorimeter respectively. A ready-to-sequence SMRTBell-Polymerase Complex was created using the P6 DNA/Polymerase binding kit 2.0 (100-236-500, Pacific Biosciences) according to the manufacturer instructions. The Pacific Biosciences RS2 instrument was programmed to load and sequence the samples on "n" SMRT cells v3.0 (100-171-800, Pacific Biosciences), taking 1 movie of 360 minutes each per SMRT cell. A MagBead loading (100-133-600, Pacific Biosciences) method was chosen to improve the enrichment of longer DNA fragments. After the run, a sequencing report was generated for every cell via the SMRT portal to assess the adapter dimer contamination, sample loading efficiency, the obtained average read-length and the number of filtered sub-reads.

## Data analysis

Following SMRT sequencing and the generation of barcode-separated subreads[17] we followed a custom data analysis method depicted in Fig. 2a. First, we implemented the RS_ReadsOfInsert.1 program using the SMRTPipe command line utility (Pacific Biosciences) using the following filtering criteria: Minimum Predicted Accuracy= 99.9, and Minimum Read Length of Insert (in bases) = 3,000. (For the error analysis, the same analysis was repeated by changing the Minimum Predicted Accuracy to 99.5 and 99.0 respectively). Resulting high quality ROIs were binned into three categories, virus DNA A, virus DNA B or non-viral DNA based on BLAST results (expect value threshold = 1.0) against a database comprised of the full-length *East African Cassava mosaic virus* (EACMV) DNA A (AM502329) and DNA B (AM502341).

### *Sequence Trimming*

Next, to identify the putative viral DNA sequence start and end points in each sub-read, the binned ROIs were aligned against two modified EACMV DNA sequences (DNA A and DNA B, for sequences in their respective bins) using MUSCLE[18]. The modified sequences consisted of: a) a three times concatenated full-length genome sequence and b) a genome sequence flanked on either side by two half sequences. This was done to simulate the linearization of the circular genome and allow for the best alignment of the generated sub-read. A 10 nt sliding window was then run from both ends of the alignment. The first window (at both sequence ends) to detect a 90% sequence identity (i.e. 9 out of 10 nucleotides in the sliding window are identical) between the read and the two modified reference sequences was designated as the start and end point of the viral sequence respectively (Fig. 2, Step 4). The sequence between these two points (called the trimmed ROI) was further analysed using DeConcat (Supplementary Fig. 1).

### *DeConcat Algorithm Description*

DeConcat begins (**Step 1**, Supplementary Fig. 1) by cleaving the trimmed ROI (A-B’) at the 30nt position to produce two segments (A-A’ and B-B’) and aligning them using MUSCLE (**Step 2a**). (This fixed distance of 30 nt was determined by benchmarking a range of values from 500 nt to 30 nt. The benchmarking results are shown in Supplementary Fig. 3) Using the alignment consensus, a score is calculated by dividing the total consensus length by the number of consensus fragments separated by gaps (**Step 3**). The algorithm iterates back (**Step 4**) to step 1, increases the cleavage position by 30, proceeds to step 2a and step 3 and iterates back to step 1. This proceeds until all possible cleavage positions have been used. The algorithm retains the alignment with the highest score in step 3. A second iteration now aligns the reverse complement sequence of the first segment (i.e. converts A-A’ to A’-A) (**Step 2b**), calculates the score, and if the score is higher than the previously retained alignment, the reverse complement alignment is used. If the computed score of the two best aligned segments is > 20, a 10nt sliding window on both ends is applied and the first windows with >90% identity are used to determine start and end positions of the alignment fragments.

The final retained alignment for each ROI can take one of eight possible overlap patterns (**Step 5**, Supplementary Fig.1). If the smaller segment lies completely within the larger one (cases 1a-1d) the algorithm re-starts with the larger segment (**Step 6**) and eliminates the smaller one. If the second segment (B-B’) overlaps the end of the first segment (A-A’) (case 2), the algorithm restarts (step 6) with both segments as independent ROIs. If the B-B’ overlaps with the front of A-A’ (case 3), the overlapping part of B-B’ is eliminated and the remaining segment (B-B”) is reattached in its original position at the end of A-A’. If B-B’ overlaps the end of A’-A (i.e. the reverse complement of A-A’ produced in step 2b) (case 4), the overlapping part of A’-A is eliminated, the remaining segment (A’-A”) is reverse complemented and B-B’ is re-attached to the end of A”-A’. In case 5, B-B’ overlaps with the start of A’-A. Here, the overlapping part of B-B’ is eliminated to produce B-B”. A’-A reverts back to its original configuration A-A’ and the two segments (A-A’ & B-B”) are re-attached. Once case-resolution (step 5) is completed, the resulting sequence is entered back into the algorithm at step 1 for another round of de-concatenation. The case resolutions have been designed so as to maintain the integrity of the initial fragment, i.e. the order in which bases are produced by the sequencer is never changed. Effectively, case-resolution simply results in sequence reduction from the ends.

### *DeConcat Performance*

After running DeConcat on our two libraries, we found that in both NSS and SS data sets a majority of sequences had scores in the range of 4-6 at the end of the DeConcat program (Supplementary Fig. 4a), indicating that for these sequences the final de-concatenation round had indeed resolved most repeat sequences. Interestingly, most reads in the SS data set required just 1-3 rounds of de-concatenation to resolve their sequences (Supplementary Fig. 4b).

We also found that of the eight possible alignment cases in DeConcat (see Methods), cases 1a, 1b and 3 were the most frequent (Supplementary Fig. 4c) in both data sets. It should be noted however that cases are assigned by DeConcat iteratively and hence do not reflect overall sequence topology. It is thus likely that the frequency of cases 1a and 1b is simply due to alignments of fragments differing greatly in size during DeConcat rounds. The relatively high frequency of case 3, however, does suggest that RCA-derived sequences are often direct sequence concatemers. Overall, based on frequencies of cases 1c, 1d, 4 and 5 it appears that cases where concatemers are formed between sequences and their opposite strands are rare.

To test the hypothesis that more DeConcat processing rounds usually result in shorter sequences we plotted the number of DeConcat rounds per sequence against its length (Supplementary Fig. 4d). However, in many cases, several processing rounds still resulted in sequences ~2800 bp in length—further demonstrating the ability of the system to resolve RCA products into single molecules.

We also tested the ability of DeConcat to process RCA-derived long sequencing reads in the absence of a reference sequence that is used in the trimming step (Step 4, Fig. 2). When comparing results with and without the trimming step, we found that the lengths of only 8.8-15.3% of sequences were affected by omitting the trimming step. Further, the average change in length of these sequences was only 15-30 bp (Supplementary Fig. 5). We conclude that DeConcat can effectively resolve RCA-derived sequences even in the absence of a reference sequence and does not require prior information on expected monomer length distributions. However, the trimming step is recommended to improve sequencing in cases where multiple viral molecules may be joined together due to the action of the Phi29 DNA polymerase.

### *Sequence Annotation and Phasing*

The de-concatenated ROIs were annotated using tBLASTn against virus protein reference sequences (Supplementary File 1) with an e-value of 0.01. The high-scoring pairs (HSPs) were summarized and sequences with protein-annotations that had contradicting and incorrect coding strands (according to the reference annotation) or lacked the full set of geminivirus proteins were eliminated. Sequences were replaced by their reverse complements where necessary to maintain all sequences in the same strand (i.e. +strand relative to the geminivirus AV1/AV2 genes). Annotation positions were also adjusted accordingly. Next, since geminiviruses have circular genomes, for comparative analyses we phased all sequences to the same start position (minus an offset) of a selected reference protein (in our case AV1). The proteins were annotated by using the start and end codon positions derived from the tBLASTn HSPs. The results were saved in FASTA (sequences only) and GenBank (sequences with ORF annotations) formats. Frameshift errors were detected using the HSP results from tBLASTn to identify cases where the same protein annotation had a break or overlap in the alignment results. The error/sequence statistic was calculated by dividing the total number of frameshifts detected by the total number of sequences in each dataset.

## Phylogenetic analyses

Phased sequence reads and reference sequences (also phased) (Supplementary File 2) were aligned using MUSCLE implemented in the CLC Genomics Workbench 10.0 using default parameters. The alignment was used to create a Maximum Likelihood phylogeny using a General Time Reversible model with 100 bootstraps. The phylogeny was used to create an unrooted cladogram and clades were defined based on the position of the 9 reference sequences used. Sequence demarcation analysis was performed using the SDT-MPI solftware[11] with the MUSCLE alignment option.

## Code Availability

Python packages along with installation and usage guidelines are available at: www.dx.oi.org/10.5281/zenodo.834928.

## Data Availability

Full length annotated genome sequences, raw sequence data and data generated during intermediate CIDER-Seq steps are freely available at: www.dx.doi.org/10.5281/zenodo.830530

## References

1. Simmonds, P. *et al.* Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15,** 161–168 (2017).

2. Dean, F., Nelson, J., Giesler, T. & Lasken, R. Rapid amplification of plasmid and phage DNA using Phi29 polymerase and a multiply-pimed rolling circle amplification. *Genome Res.* **11,** 1095–1099 (2001).

3. Sipos, R., Szekely, A., Revesz, S. & Marialigeti, K. in *Bioremediation, Methods in Molecular Biology* (ed. Cummings, S. P.) **599,** (Humana Press, 2010).

4. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **7,** 19 (2007).

5. Inoue-Nagata, A. K., Albuquerque, L. C., Rocha, W. B. & Nagata, T. A simple method for cloning the complete begomovirus genome using the bacteriophage Phi29 DNA polymerase. *J. Virol. Methods* **116,** 209–211 (2004).

6. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24,** 680–6 (2006).

7. De Bruyn, A. *et al.* Divergent evolutionary and epidemiological dynamics of Cassava Mosaic Geminiviruses in Madagascar. *BMC Evol. Biol.* **16,** (2016).

8. International Committee on Taxonomy of Viruses. ICTV Master Species List. (2016). doi:10.15468/i4jnfv

9. Hanley-Bowdoin, L., Bejarano, E. R., Robertson, D. & Mansoor, S. Geminiviruses: masters at redirecting and reprogramming plant processes. *Nat. Rev. Microbiol.* **11,** 777–88 (2013).

10. Laehnemann, D., Borkhardt, A. & Mchardy, A. C. Denoising DNA deep sequencing data — high-throughput sequencing errors and their correction. *Brief. Bioinform.* **17,** 154–179 (2016).

11. Muhire, B. M., Varsani, A. & Martin, D. P. SDT: a virus classification tool based on pairwise

sequence alignment and identity calculation. *PLoS One* **9,** e108277 (2014).

12. Brown, J. K. *et al.* Revision of Begomovirus taxonomy based on pairwise sequence comparisons. *Arch. Virol.* **160,** 1593–619 (2015).

13. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17,** 175–188 (2016).

14. Mansoor, S., Zafar, Y. & Briddon, R. W. Geminivirus disease complexes : the threat is spreading. *Trends Plant Sci.* **11,** (2006).

15. Conlan, S. *et al.* Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.* **6,** (2014).

16. Chang, S., Puryear, J. & Cairney, J. A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Report.* **11,** 113–116 (1993).

17. Pacific Biosciences Inc. Pacific Biosciences Glossary of Terms. 1–8 (2015).

18. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

## Acknowledgements

This research was supported by a grant from ETH Zurich to perform a confined cassava field trial in Kenya. We thank Dr. Hasan Were and Mariam Were (Masinde-Muliro University of Science and Technology, Kakamega, Kenya) for providing leaf samples from the cassava field trial. We thank Weihong Qi (Functional Genomics Center Zurich, Switzerland) for assistance with SMRT sequencing as well as helpful advice and discussion. Devang Mehta was supported by the European Union's Seventh Framework Programme for research, technological development, and demonstration (EU GA–2013–608422–IDP BRIDGES).

## Author Contributions

Conceptualization: DM, WG and HV; Methodology, Investigation, Formal Analysis: DM, MHH & AP; Software: MHH; Visualisation, Writing-original draft: DM; Funding acquisition, Resources, Writing-review & editing, Supervision: WG & HV. All authors agree with the final version of the manuscript.

## Competing Financial Interests

The authors declare that they have no competing financial interest.

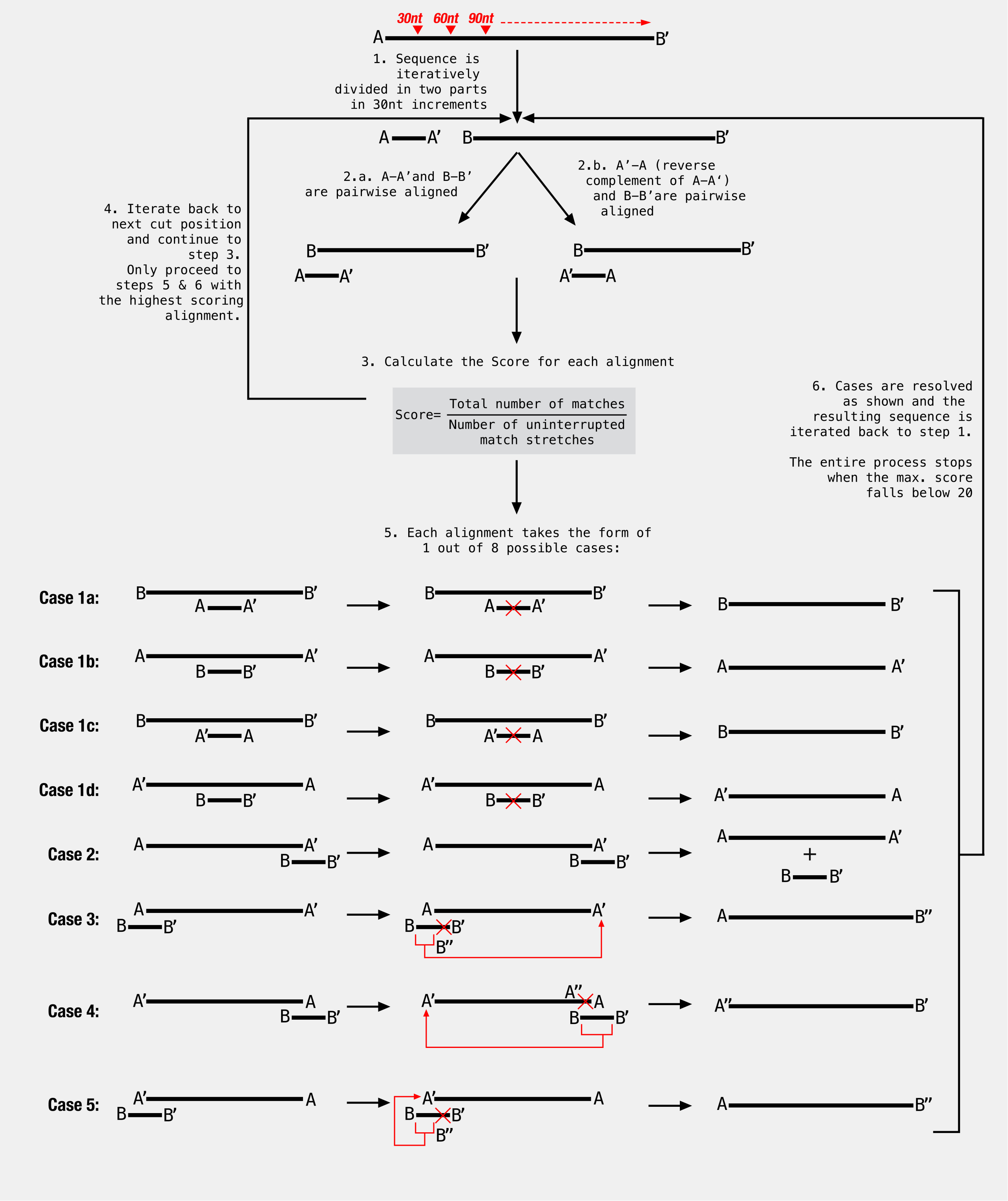
## Materials and Correspondence

Devang Mehta (devang@ethz.ch ) & Herve Vanderschuren (herve.vanderschuren@ulg.ac.be)

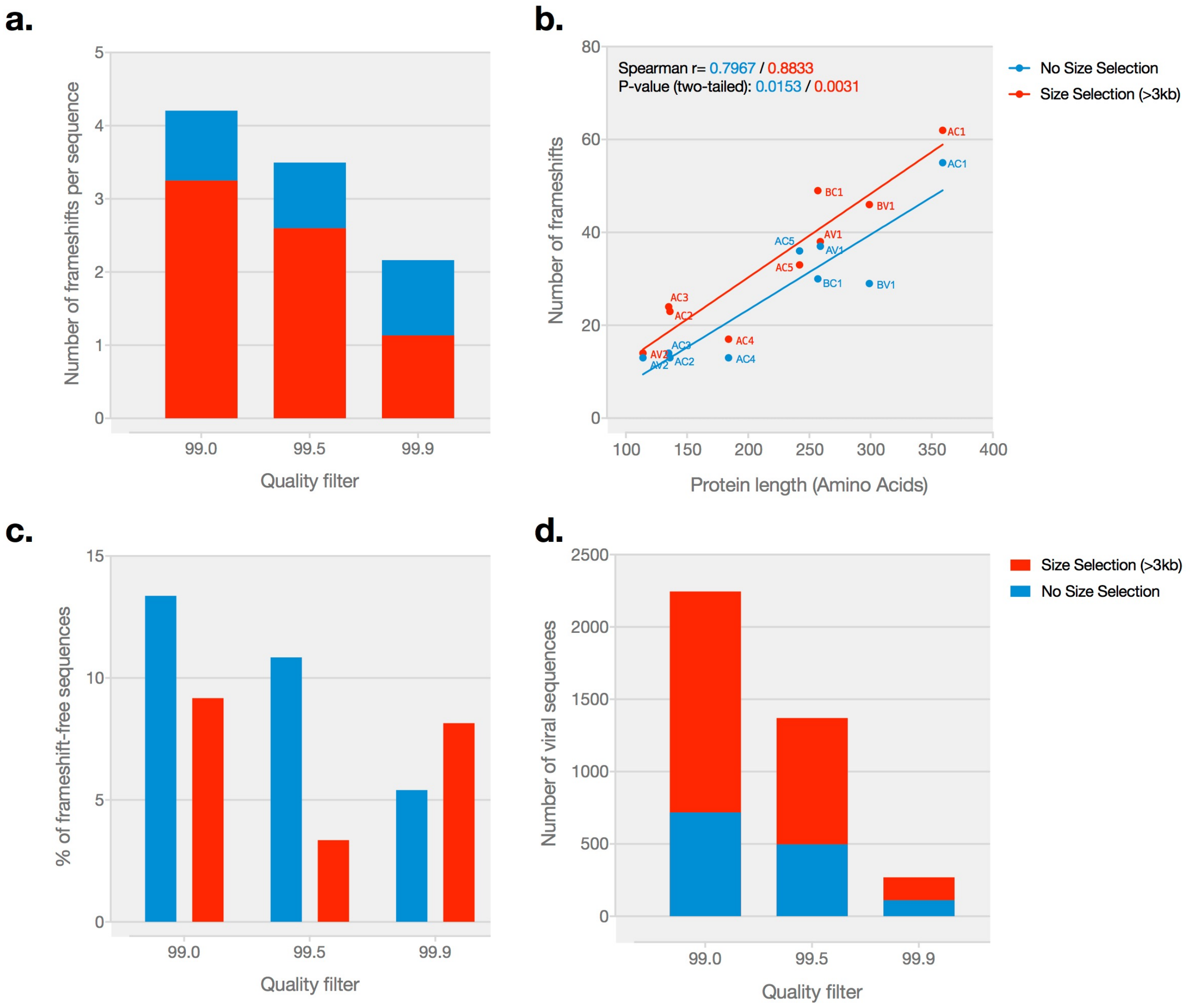
# Supplementary Information

## Supplementary Table 1: Primer sequences

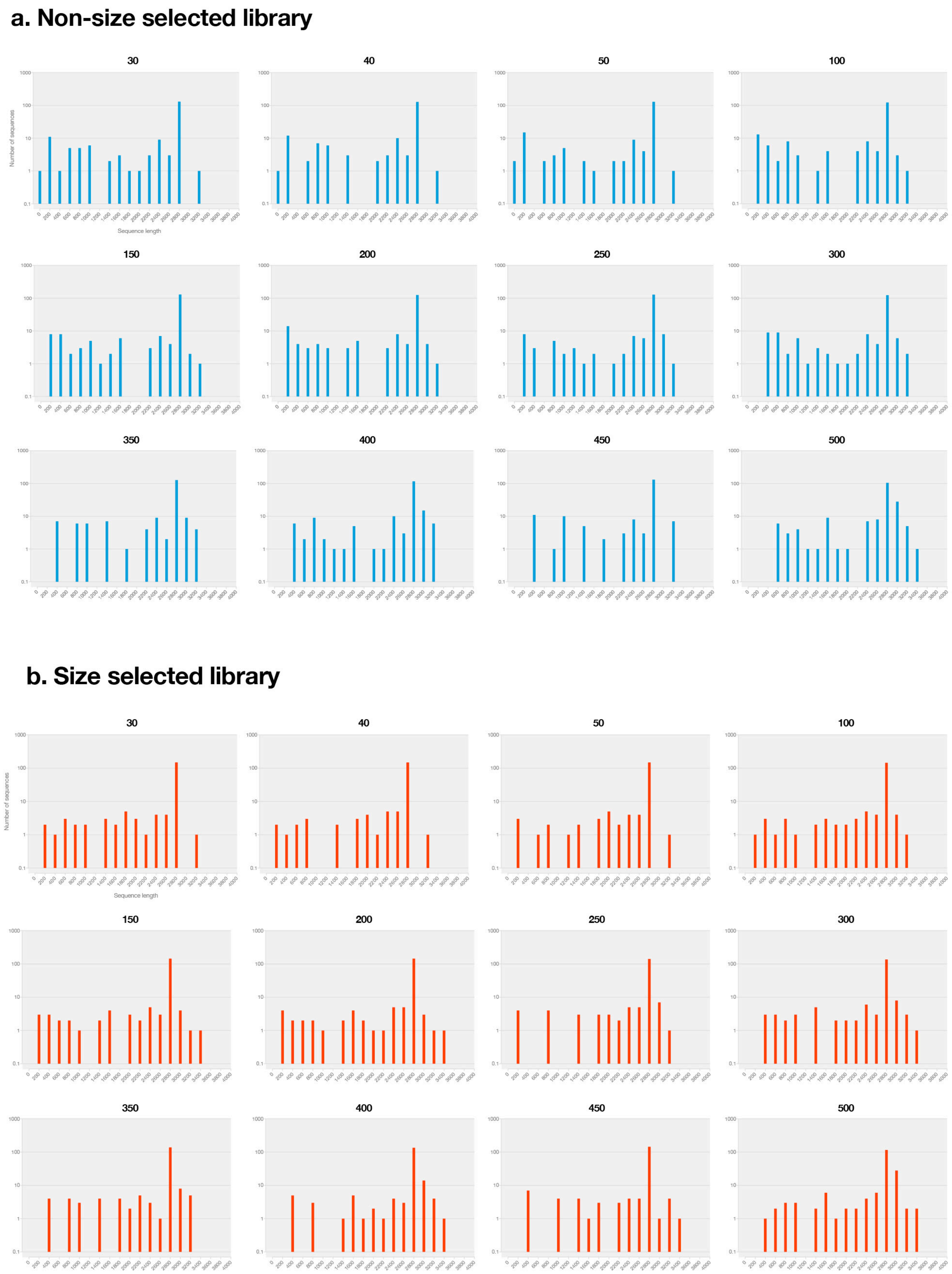
| Primer Name | Sequence | Use |
|---|---|---|
| mePP2A_genomic_F | CGC TGT GGA AAT ATG GCA TCA | Cassava qPCR reference gene |
| mePP2A_genomic_R | CTG GCT CAA ACT GCA GGA TCA A | |
| CMV_qPCR_F | GGT CCT GGA TTG CAG AGG AAG ATA GTG GG | Cassava geminiviral DNA quantitation |
| CMV_qPCR_R | GGT ACA ACG TCA TTG ATG ACG TCG ATC CC | |


30nt
60nt
90nt
A
B′
1. Sequence is iteratively divided in two parts in 30nt increments
A A′ B B′
2.a. A–A′and B–B′ are pairwise aligned
2.b. A′–A (reverse complement of A–A′) and B–B′are pairwise aligned
B B′
A A′
B B′
A′ A
4. Iterate back to next cut position and continue to step 3. Only proceed to steps 5 & 6 with the highest scoring alignment.
3. Calculate the Score for each alignment
Score= Total number of matches / Number of uninterrupted match stretches
6. Cases are resolved as shown and the resulting sequence is iterated back to step 1.
The entire process stops when the max. score falls below 20
5. Each alignment takes the form of 1 out of 8 possible cases:
Case 1a:
Case 1b:
Case 1c:
Case 1d:
Case 2:
Case 3:
Case 4:
Case 5:
+
B″
A″

**Supplementary Figure 1:** DeConcat workflow (see Methods for details)



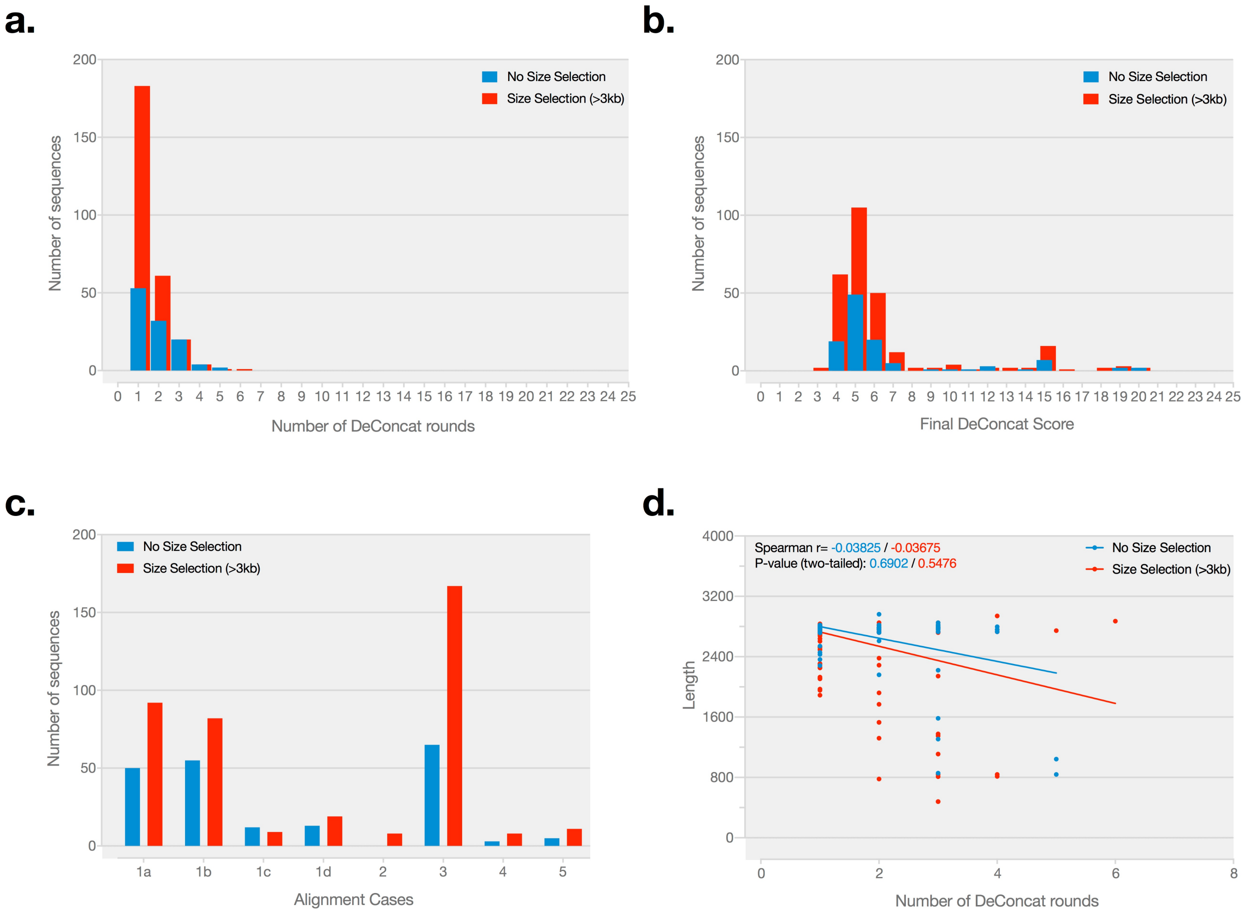
**Supplementary Figure 2:** Analysis of frameshift mutations in de-concatenated CMG sequences using different SMRT sequence quality thresholds. **(a)** The number of frameshifts decreases with increasing quality thresholds. **(b)** Number of frameshifts per protein correlates with the amino acid sequence length of the respective protein. **(c)** Percentage of frameshift-free sequences. Performance of the DeConcat algorithm. **(d)** Total number of CMG sequences produced using different quality thresholds for SMRT analysis.



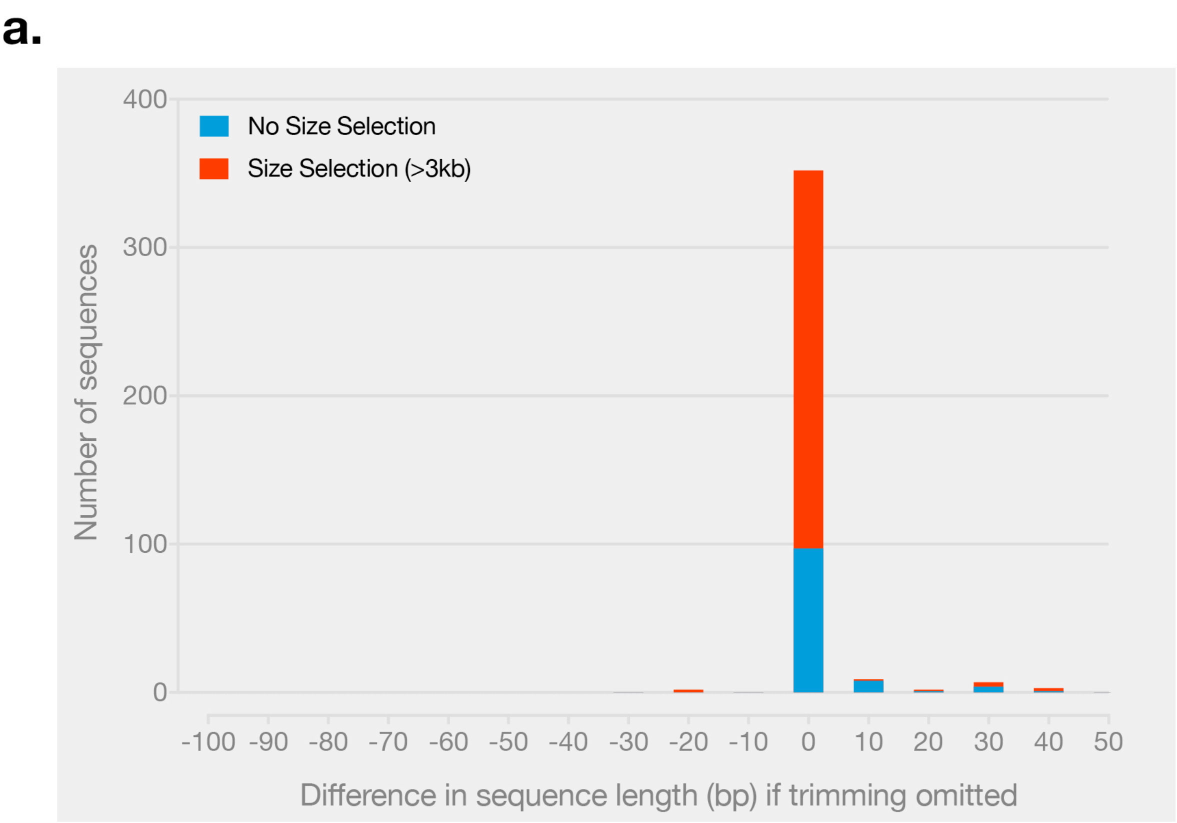
a. Non-size selected library

b. Size selected library

**Supplementary Figure 3:** DeConcat benchmarking with differently sized cuts in step 1. **(a)** non-size selected library, **(b)** size-selected library (>3 kb).



**Supplementary Figure 4:** Performance of the DeConcat algorithm **(a)** Frequency curve of the number of DeConcat rounds required to process each sequencing read. **(b)** Frequency curve of the final DeConcat alignment score obtained for the CMG datasets. **(c)** Frequency of different alignment forms used during DeConcat. **(d)** Correlation plot of the number of DeConcat rounds required to completely resolve sequences versus the length of the final de-concatenated sequence.



**b.**

| Effect | Non-size selected library | Size-selected library |
|---|---|---|
| No. of sequences decreased in length | 6 | 9 |
| No. of sequences increased in length | 11 | 15 |
| No. of sequences changed | 17 | 24 |
| Average change in length | 28.9 | 18.0 |

**Supplementary Figure 5:** The effect of applying DeConcat directly on sequencing reads without reference-based trimming. **(a)** Number of sequences (y-axis) versus the change in bp (x-axis) without trimming. **(b)** Summary of changes without trimming.

**Supplementary File 1:** Reference protein sequences from the *East African cassava mosaic virus.*

```
>AV2

MWDPLVNEFPDSVHGLRCMLAIKYLQALEDTYEPSTLGHDLVRDLVSVIRARNYVEATRRYHHFHSRLEGSS
KAELRQPIQEPCYCPHCPRHKSKTGLDKQAHVQKAHNVQDV*

>AV1

MSKRPGDIIISTPGSKVRRRLNFDSPYRNRATAPTVHVTNRKRAWINRPMYRKPTMYRMYRSPDIPRGCEGP
CKVQSYEQRDDVKHLGICKVISDVTRGPGLTHRVGKRFCIKSIYILGKIWMDENIKKQNHTNNVMFYLLRDR
RPYGNAPQDFGQIFNMFDNEPSTATIKNDLRDRFQVLRKFHATVIGGPSGMKEQALVKRFYRLNHHVTYNHQ
EAGKYENHTENALLLYMACTHASNPVYATLKIRIYFYDSIGN*

>AC4

MPFRDTYIMSPINRRRCSRVSLVECLQLTWSTCELRVLEFKPRMSFSHTQSVLYPKNTSCHSFKHYLSHQTL
SSLKSVESCIRMGNLTCMPSSNSRGKSRLRTIVSSIVYTQPVAPISTPTFKVPNQAQMSSPIWIRTETPSNG
DNFRSMDDLLEAVNNQRMMLTPKALTAAVSQKLLMSLGN*

>AC1

MRTPRFRVQAKNVFLTYPKCSIPKEHQLSFIQTLSLPSNPKFIKICRELHQNGEPHLHALIQFEGKITITNN
RLFDCVHPTCSTNFHPNIQGAKSSSDVKSYLDKDGDTVEWGQFQIDGRSARGGQQSANDAYAKGLNSGSKSE
ALNVIRELVPKDFVLQFHNLNSNLDRIFQEPPAPYVSPFPCSSFDQVPVEIEEWVADNVRDSAARPWRPNSI
VIEGASRTGKTIWARSLGPHNYLCGHLDLSPKVFNNAAWYNVIDDVDPHYLKHFKEFMGSQRDWQSNTKYGK
PVQIKGGIPTIFLCNPGPTSSYKEFLDEEKQEALKAWALKNAIFITLTEPLYSGSNQSQSQTIQEASHPA*

>AC2

MQSSSPSQNHSTQVPIKVSHRQFKKRAIRRRRVDLVCGCSYYLHINCSNHGFTHRGTHHCSSSNEWRVYLGN
KQSPVFHNHQAPTTTIPAEPGHHNSPGSIQSQPEEGAGDSQMFSQLQDLDDLTASDWSFLKGL*

>AC3

MDLRTGELITAPQAMNGVYTWEINNPLYFTITRHQQRPFLLNQDIITVQVRFNHNLRKELGIHKCFLNFRIW
TTLRPQTGLFLRVFRYQVLKYLDNIGVISINDVIRAADHVLFNVIAKTIECQLTHEIKFNVY*

>AC5

MSTCHVQKQSILCVILILPCLLMIICHVMIQPVKPFNQSLLLHARWTTNNSGMKFPQYLKPIPQIVLNCCST
GLIIKHIKYLPKVLGRIAIRPSIPKQIKHHIIRVILLLNIFIHPDLTKYVNGLDTKPLSDPVCQPRPTCHIT
NHLTDTKVLHIIPLLIRLDLTWAFTAPRYVWASIHPVHCGLSVHGPVYPGPFSICDVDSGGSSTVPVWAVEV
QPSTNLRAWSGNDDISWSLRHNYEP*

>BV1

MYSIRKQPRNFQRKCNSNTTNRFPIRRKYVGGHTRPSVRRRLSYEPVERPLVYNVLCEKQHGDVFNLQQNTS
YTSFVTYPSRGPSGDGRSRDYIKLQSMSVSGVIHAKANCNDDPMEVSHVVNGVFVFSLIMDTKPYLPAGVQA
LPTFEELFGSYSASYVNLRLLNNQQHRYRVLHSVKRFVSSAGDTKVSQFRFTKRLSTRRYNIWASFHDGDLV
NAGGNYRNISKNAILVSYAFVSEHSMSCKPFVQIETSYVG*

>BC1

MDTSVPVISSDYIQSARTEYKLTNDESPITLQFPSTIERTRVRIMGKCMKVDHVVIEYRNQVPFNAQGSVIV
TIRDTRLSDEQQDQAQFTFPIGCNVDLHYFSASYFSIDDNVPWQLLYKVEDSNVKNGVTFAQIKAKLKLSAA
KHSTDIRFKQPTIKILSKDYGPDCVDFWSVGKPKPIRRLIQNEPGTDYDTGPRYRPITVQPGETWATKSTIG
RSTSMRYTGPKHIDIDDSSSKQYASEAEFPLRGLHQLPEASLDPGDSVSQTQSMSKKDIESIIEQTVNKCLI
AHRGSSHKDL*
```

**Supplementary File 2:** Reference genome sequences of nine species of cassava geminiviruses.

> AF259896.1 EACMVCV *East African cassava mosaic virus Cameroon virus*

```
TATTTACACATATGCCATTGGGGGACATCCTATATAATGCCCCCCATTCCACCGTTTCCC

CTGGAGTTTTGAGTGTCCCCCGATCCAAAACGACAGCCAATATGCCGAGAGCCGGTCGTT

TTCAAATAAATGCCAAAAATTATTTCATAACCTATCCCAGATGCTCATTAACAAAGGCAG

AGGCCCTTTCCCAATTAAAAGCCCTTTCTTACCCGACGAATATCAAATTCATTAGGGTTT

GCAGAGAACTACATCAGGATGGGGTGCCTCATCTCCATGTTCTCATCCAATTCGAAGGCA

AGTTCCAATGTACCAACCCCAGATTCTTCGATCTCATTTCCCCATCCCGATCAACACATT

TCCATCCGAACATTCAGGGAGCTAAATCATCGTCCGATGTCAAAGCTTACATTGAAAAGG

GAGGGGAATTTCTTGACGATGGAATTTTCCAAGTCGATGCCAGAAGTGCCAGGGGGGAGG

GCCAGCATTTAGCTCAGGTATATGCAGAAGCGTTGAATGCTTCTTCTAAATCAGAAGCTC

TTCAAATTATCAAAGAAAAAGATCCAAAGTCCTTTTTTTTACAGTTCCATAACATATCTG

CTAACGCGGATCGAATCTTCCAGGCTCCGCCACAAACTTACGTTAGTCCGTTCTTATCAT

CCTCATTTACGCAAGTCCCAGAGGACATAGAAGTCTGGGTATCCGAAAATATATGCCGTC

CCGCTGCGCGGCCATGGAGACCGATCAGTATTGTTCTAGAAGGTGATAGCCGAACCGGCA

AAACAATGTGGGCTCGTTCACTGGGACCCCATAATTATCTTTGTGGACACCTGGATCTGT

CTCCCAAGATATATTCAAACGACGCATGGTACAACGTCATTGACGACGTAGACCCGCATT

ATCTAAAGCATTTCAAAGAGTTCATGGGGGCCCAACGAGATTGGCAATCAAACACAAAAT

ACGGAAAGCCCATTCAAATTAAAGGTGGGATTCCCACCATCTTCTTATGCAATCCGGGCC

CCAATTCGTCCTATAAAGAATACCTAGACGAGGACAAGAATTCCAATCTAAGGAATTGGG

CGCTCAAGAATGCGCTCTTCATCTCCCTCACCGAGCCACTCTTCTCCTCCACCGATCAAA

GCCAGGCACAGGCAAGCTAAGATCAGAGCACCTAGACGTCGACGCATCGACCTAACTTGT

GGCTGTTCCATCTATCGCAGCATTAATTGTCACAACCATGGATTTACGCACAGGGGAAGA

CATTGGTGCTCTTCAATGGAGGAATGGCGCCTTTATCTGGGAGATTCCAAATCCCCTATA

TTTCACAATCCTCAACCACGACAGCATGCCGTTCAACATGAATCACGACATAATCACTGT

CCAGATTCGGTTCAACTACAACCTGAGGAAAGCACTGGGGATGCACAAGTGTTTTCTCAA

CTTCCGGATCTGGACTCGTTTACATCCTCAGACTTGGCGTTTCTTCAAAACCTTTAGGAC

ACAAGTCATGAAATATCTCAATAATTTAGGTGTTATTTCGATTTCAACTGTTATAGATGC

AGTGCATCATGTATTGAATATAGTTTTTGTGGGAACCCTATATGTATCACAAGATCATGC

AATCAAATTTAATATTTATTAATTTGTCACTGAATCATAGAAATAGATGCGTATTTTCAG

CGTAGCGTACACAGGATTTGAGGCATGTGTACATGCCATATACAATAACAACGCATTCTC

GGTATGATTCTCATACTTGGCCTGTTCCTGATGATTATACACTACATGATTATTGATCCT

AAAAAACCTCTTAACCAGAGCTTGTTCCTTCATCCCAGAGGGTCCACCAACAACAGTCGC
```

ATGGAATTTACGCAACACCTGATACCGGTCCCTAAGATCATTCTTCACAGTTGCAGTCGT

AGGTTCATTATCAAACATGTTGAACACTTGTCCAAAATCTTGAGGACTCGGCCCATAAGG

CCTTCTATCTCGCACGAGGAAAAACATAACATGGTTCGTATGATTTTGCTTCTTGATATT

CTCATCCATCCAGATCTTGCCCAATATATATATGGACTTCACACACAACCTCTTCCCAAC

TCTATGGGTAATGCCTGGCCCACGAGTAACTTCACTAACACATCGGACCATACCCGTATG

CTTCACATCATCCCTCTGCTCAAACGACTGGACCTTACATGGGCCTTCACAGCCCTTAGG

GACGTCTGGGCTTCGATACATCCTGTACATCTTGGGCTTCCGATACATGGGCCTGTTGGC

CCATATTCTGCTTCTGGTGACGCGGACAGTGGGGGCAACCACACGGTTCGTGTATGGGCT

GTCGAAGTTCAGCCTCCTCCGCACCTTGGATACGGGTGTTGAGATTATTATATCTCCTGG

TCGCTTCGCCATAACTCTTACACCGTAGCACCCCTATTAGATCACGTATATATTCAGCCC

CGACAGTACCGCGATCATATTCCTGTTCCAAATGTAACAGGTATTTAACAGCAAGCATCG

AACGGAAACCGTGCACGGTTTCCGGAAAATCGTTAACTAACGGATTCCACATTTTGACGC

GCTCCACTACTTCGTGACGAAGTATTTAAAGTCAAAAATCCATATCTAGCATTCAAGGCG

CAAATATTATTGGCCGACAAACATGCGTGCGCGGGGACCACTTTCTTTTACGGGCGCGGA

CTATTATGGGGCCCATCTGCTTTTTTCGGGCGCGGCCATCCGGTAATATTAATCGGATGG

CCGCCAATATTCGCAATTCGAATTTTGAATGGAGTCTACC

> AJ427910.1 ACMV *African cassava mosaic virus*

ATTACAAGAATGCCATTTAGAGACACCTATATAATGTCTCCAATTAACAGGAGGAGATGC

TCAAGAGTGTCTCTAGTTGAGTGTCTCCAATTGACTTGGTCAACATGCGAACTCCGCGTT

TTAGAGTTCAAGCCAAGAATGTCTTTCTCACATACCCAAAGTGTTCTATACCCAAAGAAC

ACCAGCTGTCATTCATTCAAACACTATCTCTCCCATCAAACCCTAAGTTCATTAAAATCT

GTAGAGAGCTGCATCAGAATGGGGAACCTCACTTGCATGCCCTCATCCAATTCGAGGGGA

AAATCACGATTACGAACAATCGTCTCTTCGATTGTGTACACCCAACCTGTAGCACCAATT

TCCACCCCAACATTCAAGGTGCCAAATCAAGCTCAGATGTCAAGTCCTATCTGGATAAGG

ACGGAGACACCGTCGAATGGGGACAATTTCAGATCGATGGACGATCTGCTAGAGGCGGTC

AACAATCAGCGAATGATGCTTACGCCAAAGGCCTTAACAGCGGCAGTAAGTCAGAAGCTC

TTAATGTCATTAGGGAATTAGTCCCAAAGGACTTTGTACTTCAATTTCATAATCTCAATA

GTAATTTAGATAGGATTTTCCAGGAGCCACCAGCTCCTTATGTTTCTCCCTTCCCATGTT

CTTCCTTTGACCAAGTTCCTGTTGAAATTGAAGAATGGGTCGCTGATAATGTTAGGGATT

CCGCTGCGCGGCCATGGAGACCCAATAGTATTGTAATAGAAGGTGCTAGCAGAACAGGGA

AGACGATATGGGCCAGATCTTTAGGCCCACACAATTACCTGTGTGGACACCTGGACCTTA

GTCCAAAGGTCTTCAATAATGCTGCCTGGTACAACGTCATTGATGACGTCGATCCCCACT

ACCTAAAGCACTTTAAAGAATTCATGGGGTCCCAGAGGGACTGGCAGTCCAACACGAAAT

ACGGGAAACCCGTTCAAATTAAAGGTGGCATTCCCACTATCTTCCTCTGCAATCCAGGAC

CTACCTCGTCCTATAAAGAGTTCCTAGACGAGGAAAAGCAAGAAGCGCTAAAGGCCTGGG

CATTAAAGAATGCAATCTTCATCACCCTCACAGAACCACTCTACTCAGGTTCCAATCAAA

GTCAGTCACAGACAATTCAAGAAGCGAGCCATCCGGCGTAGGAGAGTGGATCTTGTCTGT

GGCTGTTCATATTACCTCCATATCAACTGCTCCAATCATGGATTTACGCACAGGGGAACT

CATCACTGCTCCTCAAGCAATGAATGGCGTGTATACCTGGGAAATAAACAATCCCCTGTA

TTTCACAATCACCAGGCACCAACAACGACCATTCCTGCTGAACCAGGACATCATAACAGT

CCAGGTTCGATTCAATCACAACCTGAGGAAGGAGCTGGGGATTCACAAATGTTTTCTCAA

CTTCAGGATCTGGACGACCTTACGGCCTCAGACTGGTCTTTTCTTAAGGGTCTTTAGATA

CCAAGTACTTAAATACTTGGATAATATAGGTGTTATCTCAATTAACGATGTCATTAGAGC

TGCTGATCATGTTTTGTTCAATGTAATTGCCAAAACTATTGAGTGTCAGTTGACTCATGA

AATAAAATTCAATGTTTATTAATTGCCAATACTGTCATAGAAGTATATACGTATTTTCAA

CGTAGCATATACAGGATTGGAGGCATGAGTACATGCCATGTACAGAAGCAAAGCATTCTC

TGTGTGATTCTCATACTTCCCTGCCTCCTGATGATTATATGTCACGTGATGATTCAACCT

GTAAAACCTTTTAACCAAAGCCTGCTCCTTCATGCCAGATGGACCACCAATAACAGTGGC

ATGAAATTTCCTCAATACCTGAAACCTATCCCTCAAATCGTTCTTAATTGTTGCAGTACT

GGGCTCATTATCAAACATATTAAATATCTGCCCAAAGTCTTGGGGCGCATTGCCATAAGG

CCTTCTATCCCTAAGCAGATAAAACATCACATTATTCGTGTGATTTTGCTTCTTAATATT

TTCATCCATCCAGATCTTACCAAGTATGTAAATGGACTTGATACAAAACCTCTTTCCGAC

CCTGTGTGTCAGCCCAGGCCCACGTGTCACATCACTAATCACCTTACAGATACCAAGGTG

CTTCACATCATCCCTCTGCTCATACGACTGGACCTTACATGGGCCTTCACAGCCCCTAGG

TATGTCTGGGCTTCTATACATCCTGTACATTGTGGGCTTTCTGTACATGGGCCTGTTTAT

CCAGGCCCGTTTTCGATTTGTGACGTGGACAGTGGGGGCAGTAGCACGGTTCCTGTATGG

GCTGTCGAAGTTCAGCCTTCGACGAACCTTCGAGCCTGGAGTGGAAATGATGATATCTCC

TGGTCGCTTCGACATAATTACGAGCCCTGATAACTGAGACTAGATCTCTAACCAAATCGT

GGCCCAAAGTACTGGGCTCGTATGTATCCTCTAAGGCCTGCAAATATTTAATTGCAAGCA

TACACCTAAGCCCATGCACCGAGTCTGGAAACTCATTCACCAGTGGATCCCACATTGCGC

ACTAGCAACAACTTCGCCTGTAAGTATATAGTGGTCCACCACTAATACATAACCTTTAAA

GCTACAGCATGATTGGCCGACATAAGTAGTGCGCGGGGACCACGTTTAAAGGGGGGCGGG

GCCAACCGGTAATATTATACGGTTGGCCCCTTGGGTGTTCTGCGCCTTTTGAGCCTTTTA

ATTCAAATTAAAGTTCAACTT

> AJ717572.1 EACMKV *East African mosaic Kenya virus*

TATTACATAAATGCCATTTAGGGGGCATCATATAAATTGCCCCCCTTTCCCCCGATTGCT

```
ACAGACTTTGAGTGCCCCCCGATTGCTATACGACAGCGAAAATGCCAAGGGCTGGTCGTT

TTAGCATCAAAGCCAAAAACTATTTCCTAACATATCCCAAATGCTCTCTATCCAAAGAGG

AGGCATTGGATCAAATCCGACAACTCCAAACCCCAACAAATAAATTGTTCATCAAGATCT

GCAGAGAACTCCATGAAAATGGGGAACCTCATCTGCATGCCCTCATTCAGTTCGAGGGCA

AGTACAATTGTACCAACCATCGATTCTTCGACCTCATATCCCCATCCCGGTCAGCACATT

TCCATCCAAACATTCAGGGAGCTAAATCAAGCTCCGACGTCCAGTCCTATATGGACAAGG

ACGGAGACACCATCCAATGGGGCACGTTTCAGATCGACGGACGATCTGCTCGAGGAGGAC

AACAATCAGCCAATGACGCTTACGCCAAGGCTCTTAACTCAGCAAATAAGTCAGAGGCTC

TTAATGTAATACGCGAACTAGCTCCAAAAGATTTTGTTTTACAGTTTCATAATTTACATA

GCAATTTAGATAGGATTTTTCAAGAGCCTCTGACTCCTTATGTTTCTCCATTTCTTTCAT

CTTCTTTCACTAACGTTCCTGAGGAACTTGAAGATTGGGTTTCCGAGAACGTGATGGGTT

CCGCTGCGCGGCCATGGAGACCTACTAGTATCGTCATCGAGGGCGATAGTAGGACGGGGA

AGACGATGTGGGCCCGCTCTTTGGGTCCACACAACTACTTGTGTGGACACCTGGATCTTA

GTCCAAAGGTCTACAGCAACGACGCCTGGTACAACGTCATTGATGACGTCGACCCCCACT

ACCTCAAACACTTCAAAGAATTCATGGGGGCCCAAAGGGACTGGCAAAGCAATACAAAGT

ACGGGAAGCCGATTCAAATTAAAGGCGGCATTCCCACTATCTTCCTCTGCAATCCGGGCC

CAACATCATCATATAAAGAGTTTCTGGACGAGGAAAAGAACCAGTCCCTTAAAGCCTGGG

CTTTAAAGAATGCCACCTTCATCACCCTCCACGAGCCATTGTTCTCTAGTGCCCATCAAA

GTCCAACACCGCACAGCGAAGACCAGGGCCGTCAGACGTAGGCGGGTAGACCTCGAATGC

GGCTGCTCGTTCTATCTCCATATCGACTGCATCAACCATGGATTCTCGCACAGGGGAACT

CATCACTGCGCCTCAAGCAAGGAATGGCGTTTTTACCTGGGACATAACAAATCCCCTCTA

TTTCGAAATCACCAACCACGACAAGAGGCCAGGGAACATGAACCACGACATCATCACACT

CCAGATACGGTTCAACCACAACCTCCGGAAGGCATTGGGGATTCACAAGTGTTTTCTCAA

CTTCAGGGTCTGGACGACCTTACGGCCTCAGACTGGTCTTTTCTTAAGAGTATTTAGATA

TCAAATGCTCAAGTATTTGGATATGATAGGCGTTATTTCCATTAACACTGTACTTCAAGC

TGTTGATCATGTTATGTACGATGTATTACTAAACACGCTCCAAGTTACGGAGCAACATGC

AATAAAATTCAACCTTTATTAATTTGTCACTGCATCATAAAAATAGATGCGTATTTTCAG

CGTAGCGTACACAGGATTAGCGGCATGTGTACATGCCATATACAATAACAACGCATTCTC

AGTATGATTCTCATACTTGGCCTGTTCCTGATGATTATACACTACATGATTATTGATCCT

AAAAAACCTCTTAACCAGAGCTTGTTCCTTCATCCCAGAGGGTCCACCAACCACAGTCGC

ATAGAATTTACGTAACACCTGATACCGGTCCCTAAGATCATTCTTCACAGTTGCAGTAGT

TGGTTCATTATCAAACATGTTGAACACTTGCCCAAAATCTTGAGGACTCGGACCATACGG

CCTTCTATCTCGAACGAGGAAGAACATCACATGGTTCGTGTGATTTTGCTTCTTGACATT
```

CTCATCCATCCAGATCTTGCCCAATATATATATGGACTTAACACAAAACCTCTTCCCGAC

TCTATGGGTAATGCCTGACCCACGAGTAACATCACTGACACATCGGACCATACCAGTGTG

CTTAACATCATCCCTCTGTTCATAGGACTGAACCTTACATGGGCCTTCACAGCCCTTAGG

GACATCTGGGCTTCGATACATTCTGTACATCTTGGGCTTCCGATACATGGGTCTGTTGGC

CCATATTTTGCTTCTGGTGACGCGGACAGTGGGGGCAACAACACGGTTCGTGTATGGGCT

GTCGAAGTTCAGCCTCCTCCGCACCTTGGATACGGGTGTTGAGATTATTATATCTCCTGG

TCGCTTCGACATAACTCTTACACCGTAGAACCCCTATTAGATCCCGTATATACTCAGCCC

CGACAGTACCGCGGTCGTATTCCTGTTCCAGATGTAACAGGTATTTAACAGCAAGCATAG

AACGGAAACCGTGAACGGTTTCGGGAAAGTCGTTCAACAATGGATCCCACATGTTGACGC

GCTCCACTACTTCGCGACGAAGTCTATAAAGACAAACAACATATATCTAGCCTTTCACGC

GTGAATATGACTGGCCGACCAAAACAAGTGCGCAGGGACCACTTTCTTTCACGGGCGCGG

CCATCCTGTGGGGTCCACCTGCTTTTTCGGGCGCGGCCATCCGGTAATATTATACGGATG

GCCGCTTTCCACGTTTGAATTTCAAATTCAATGAGGA

>AF155806.1 SACMV *South African cassava mosaic virus*

AATTACACATATGCCATTTGGGGGGCATCATATAAATTGCCCCCCATTCCCCCGATTGCT

AGGAACTTTGAGTGCCCCCCGATTGCTATACGACAGCGAAAATGCCGAGGGCTGGTCGTT

TTAGCATAAAAGCCAAAAATTATTTCCTCACGTATCCGAAATGCACTCTCTCGAAAGAAG

CGGCATTAGATCAACTCCGACAACTCCAAACCCCAACAAATAAATTGTTCATCAAGATCT

GCAGAGAACTCCATGAAAATGGGGAACCTCATTTGCATGCCCTCATTCAGTTCGAGGGCA

AGTACAATTGTACCAACCAACGATTCTTCGACCTCATATCCCCTTCCAGGTCAACACATT

TCCATCCAAACATTCAGGGAGCTAAATCCAGTTCTGACGTCAAGTCCTATTTGGACAAGG

ACGGAGACACCATCCAATGGGGCGAGTTTCAGATCGACGGACGATCTGCTCGCGGCGGAC

AACAATCCGCCAATGACGCTTACGCCAAGGCTCTTAACGCAGCAAGTAAAACAGAGGCTC

TTAATGTAATCCGGGAACTAGCCCCAAAGGATTTTGTTTTACAGTTTCATAATTTAAATA

GCAATTTAGATAGGATTTTTCAGGAGCCTCCGATTCCTTATATTTCTCCCTTTCTTTCTT

CTTCTTTCACTCATGTTCCTGAGGAACTTGAAGACTGGGTTTCCGAGAACGTGATGGGTT

TCGCTGCGCGGCCATGGAGACCGAGTAGTATCGTCATCGAGGGCGATAGTAGGACAGGGA

AGACGATGTGGGCCCGATCTCTGGGACCACACAACTACTTATGTGGACATTTGGATCTCA

GTCCAAAGGTTTACAGCAACGACGCATGGTACAACGTCATTGATGACGTCGACCCCCATT

ACCTCAAGCACTTCAAAGAATTCATGGGGGCCCAAAGGGACTGGCAAAGCAATACCAAGT

ACGGGAAGCCGATTCAAATTAAAGGCGGCATTCCCACTATCTTCCTATGCAATCCAGGAC

CGACATCATCATATAAAGAGTTTCTGGACGAGGAAAAGAACCAGTCCCTTAAAGCCTGGG

CTTTAAAGAATGCAACCTTCATCACCCTCCACGAGCCATTGTTCTCAAGTGCCCATCAAA

GTCCAACACCGCACCGCGAAGACTAGGGCCCTCAGACGTAGGAGGGTAGACCTCGAATGC

GGCTGCTCGTTCTATCTCCATATCGACTGCATCAACCATGGATTCTCGCACAGGGGAACT

CATCACTGCGCCTCAAGCAAAGAATGGCGTTTTTACCTGGGAAATAACAAATCCCCTCTA

TTTCGACATCACCAACCACGACAAGCGGCCAGGGAACATGAACCACGACATCATCACCCT

CCAGATACGGTTCAACCACAACATCAGGAAGGCATTGGGGATTCACAAGTGTTTTCTCAA

CTTCAAGGTCTGGACGACCTTACGGCCTCCGACTGGTCTTTTCTTAAGAGTATTTAAATA

TCAAGTGCTCAAGTATTTAAATATGATAGGCGTTATTTCCATTAACACTGTACTCAGAGC

TGTTGATCATGTTCTGTACGATGTATTACTAAACACACTCCAAGTTACGGAGCAACATGC

AATAAAATTCAACCTTTATTAATTTGTTACTGCATCATAAAAATAGATGCGTATTTTAAG

CGTAGCATACACTGGATTAGAGGCATGCGTACATGCCATATACAACAATAACGCATTCTC

TGTATGATTCTCATACTTAGCTGCCTCCTGGTGATTATACACAACATGATTATTTATCCT

AAAAAATCTCCTCACCAAAGCCTGCTCCTTCATTCCAGAAGGACCCCCAACAACGGTGGC

ATGAAACTTCCGCATAACTCGATACCTATCCCTAAGATCGTTCTTCACAGTGGCTGTACT

GGGCTCATTATCAAACATATTAAAAACCTGTCCAAAGTCCATGGGGCTATTGCCATAGGG

CCTTCTGTCACGGACTAAGAAGAACATGACCTGGTTTGTATGGTTCTGCTTCTTGATGTT

TTCATCCATCCATATCTTACCTAACACATATATAGACTTGATACAGAACCTTTTACCTAC

TCTATGTGTAATTCCCGAACCACGCGTGACATCACTAACACAACGAACACTGCCAGTATG

CTTAACGTCATCTCGCTGTTCATAAGATTGAACCTTACATGGGCCTTCACAGCCACGCGG

AACATCAGGGCTTTTGAACATTCTGTACATTCTGGGCTTTCGGTACATGGGCCGGAACGT

CCATGATCGACGCTTGTTTGTGCCTTGGACAATGGGGACAGCAGCACGGCTGCTGAACGG

GCTGTCGAAGTTCAGCCTTCGACGCACCTTCGAGACGGGAGTGGAAATGATTATATCGGC

GGGACGCTTCGACATAATCACGGGCTCGGATCACACCGATGAGATCACGGACTAGATCGT

GGCCCAAAGTATTGGGCTCGTAGGTTTCCTCCAAGGCCTGCAAATATTTAATAGCAAGCA

TACAGCGAAAACCGTGCACAGACTCGGGGAACTCATTCAACAATGGATCCCACATGTTGA

CACGCTCCACTACTTCGCGACGAAGTCTATAACGACATAAAACAAATATCTAGGCTTTCA

CGCGTGAATATGACTGGCCGACAGCAACACGTGCGTGGGGACCACTTTCTTTTACGGGCG

CGGTCATCCAATGGGGTCCACCTACTTTGTCGGGCGCGGCCATCCGGTAATATTAGACGG

ATGGCCGCTTTCCACGTTCGAAATTCAAATTTCATCACAA

>AJ717583.1 EACMZV *East African cassava mosaic Zanzibar virus*

AATTACCATTATGCCATTTGGAGACACTATATATTGTCTCCAATTCCGTGGAGACATCAA

CCGAGGTGTCTCCAATTGCTCTCGCATAAATGACTCCCCCCAAGCGTTTTAAAATACAGG

CCAAAAACTATTTTCTCACATATCCCAAATGTTCTCTATCTAAACACGACGCATTATCCC

AAATATTAAACCTCCCAACTCCCACAAACAAGAAATACATCAAAGTGTGCAGAGAACTTC

```
ACGACGATGGGCAACCTCATCTCCACATGCTTATTCAGTTCGAAGGCAAATTCTCATGCA
CAAATAAGCGATTCTTCGACCTGGTATCCCCCACACGATCAACACATTTCCATCCGAACA
TTCAAGGAGCTAAATCCAGCTCCGACGTCAAGTCCTACATCGACAAAGATGGGGATACCA
CTGAGTGGGGCGAATTCCAGATCGACGCCAGATCGGCTAGAGGCGGCTGCCACAATGCTA
ATGACGCATGTGCCGAAGCATTAAACTCCGGTTCCAAGGCAGCAGCACTTCTAATTATTA
AGGAGAAACTCCCAAAAGAATTTATTTTTCAATATCATAATTTAAGTAGTAATTTAGATA
GGATTTTTCAAGAGCCACCAGCTCCTTATGTTTCTCCATTTCTGTCTTCTTCTTTTACTA
ACGTTCCTGAGGAACTTGAAGTCTGGGTTTCCGAGAACGTGATGGGTTCCGCTGCGCGGC
CTTGGAGACCTAATAGTATTGTTATTGAGGGTGATAGTCGTACAGGGAAGACAATGTGGG
CCAGATCATTGGGACCACATAATTATTTGTGTGGACACCTGGATCTCAGTCCAAAGGTCT
ACAGCAACGACGCCTGGTACAACGTCATTGATGACGTCGACCCCCACTACCTCAAACACT
TCAAAGAATTCATGGGGGCCCAACGGGACTGGCAAAGCAATACAAAGTACGGGAAGCCAA
TTCAAATTAAAGGCGGCATTCCCACTATCTTCCTATGCAATCCAGGACCAACATCATCAT
ATAAAGAGTTTCTGGAAGAGGAAAAGCACCAATCCCTTAAAGCCTGGGCTTTAAAGAATG
CCACCTTCGTCACCCTCCACGAGCCATTGTTCTCAAGTGCCCATCAAAGTCCAACACCGC
ACAGCGAAGACCAGGGCCATCAGACGTAGGCGCGTAGACCTCGAATGCGGCTGCTCGTTT
TATATCCATATTAACTGCATCAACCATGGATTCTCGCACAGGGGAACTCATCACTGCGCC
TCAAGCAACGAATGGCGTTTTTACCTGGGAAATAACAAATCCCCTATATTTCGCAATCAC
CACCCACGACAAGAGACCCGGGAACATGAACCACGACATCATCACAATCCAGATACGGTT
CAACCACAACATCCGGAAGGCATTGGGGATTCACAAGTGTTTTCTCAACTTCAAGATCTG
GACGACCTTACGGCCTCCGACTGGTCTTTTCTTAAGAGTATTTAGATCTCAAGTGCTCAA
GTATTTAGACATGATAGGCGTTATTTCCATTAACACTGTACTTCGATCTGTTGATCATGT
TCTGTACGATGTATTACTAAACACGCTCCAAGTTACGGAGCAACATGCAATAAAATTCAA
CCTTTATTAATTTGTCACTGCATCATAGAAATAGATGCGTATTTTCAGAGTCGCATACAC
AGGATTTGAGGCATGTGTACATGCCATATACAATAACAACGCATTCTCAGTATGATTCTC
ATACTTGGCCTGTTCCTGATGATTATACACTACATGATTATTAATCTTAAAAAACCTCTT
AACCAGCGCTTGTTCCTTCATCCCAGAGGGTCCACCAACAACAGTCGCATAGAATTTACG
CAGCACTTGATACCGGTCCCTAAGATCATTCTTCACAGTTGCAGTCGTAGGTTCATTATC
AAACATGTTGAACACTTGTCCAAAATCTTGCGGACTCGGACCATAAGGCCTTCTATCTCG
CACGAGGAAGAACATAACATGGTTCGTATGGTTTTGCTTCTTAATATTCTCATCCATCCA
AATCTTGCCCAATATATATATGGACTTCACACAGAACCTCTTCCCGACTCTATGGGTAAT
GCCTGACCCACGAGTAACCTCACTGACACATCGGACCCTACCAGTGTGCTTAACATCATC
CCTCTGTTCATACGACTGAACCTTACATGGGCCTTCACAGCCCTTAGGAACATCTGGGCT
```

TCGATACATTCTGTACATCTTGGGCTTCCGGTACATGGGCCTGTTAGCCCATATTTTGCT

TCTGGTGACGCGGACAGTGGGGGCAACAACACGGTTCGTGTATGGGCTGTCGAAGTTCAG

CCTCCTCCGCACCTTGGATACGGGTGTTGAGATTATTATATCTCCTGGTCGCTTCGACAT

AACTCTTACACCGTAGAACCCCTATTAGATCCCGTATATACTCAGCCCCGACAGTACCGC

GATCGTATTCCTGTTCAAGATGTAACAGGTATTTAACCGCAAGCATGGAACGGAAACCGT

GAACGGTTTCAGGGAAATCGTTTAACAATGGATCCCACATGTTGACGCGCTCCACTACTT

CGCGACGAAGTCTATAAAGACAAACAACAAATATCTAGACTTCCACGCGTGAATATGACT

GGCTGACCGAAACAAGTGCGCGGGGACCACTTTCTTTCACAGGCGCGGCCATCCTGTGGG

GTCCACCTGCTTTTTTGGGCGCGGCCATCCGGTAATATTATACGGATGGCCGCTTTTGGT

TCACATTTTGAAAATAGATCTACCT

>AM502329 EACMV *East African cassava mosaic virus*

CTATTTACACATATGCCATTGGGGGACCTCATATATAATGCCCCCCATTCCACCGATTGC

TATAGACTTTGAGTGTCCCCCGATCCAAAACGACAACCAATATGCCAAGAGCCGGTCGTT

TTCAAATAAATGCCAAAAATTATTTCATAACCTATCCCCGATGCTCCTTAACAAAAGAAG

AGGCCCTTTCCCAATTACAAGCCCTTTCGTACCCGACGAATATCAAATTCATTAGGGTTT

GCAGAGAACTACATCAGGATGGGGTGCCTCATCTCCATGTTCTCATCCAATTCGAAGGCA

AGTTCCAATGTACCAACCCGAGATTCTTCGATCTCATTTCCCCATCCCGATCAACACATT

TCCATCCGAACATTCAGGGAGCTAAATCATCGTCCGATGTCAAAGCTTACATTGAAAAGG

GAGGGGAATTTCTTGACGCTGGACTTTTCCAAGTCGATGCCAGAAGTGCAAGAGGGGAGG

GCCAGCATTTAGCTCAGGTATATGCAGACGCGTTGAATGCTTCGTCTAAATCCGAGGCTC

TTCAAATTATTAAAGAAAAAGATCCAAAGTCCTTTTTTTTACAGTTCCATAACATATCTG

CTAACGCAGATAGAATCTTCCAGGCTCCGCCACAAACTTACGTTAGTCCGTTCTTATCAT

CATCTTTTACACAAGTCCCAGAAGACATAGAGGTATGGGTATCCGAAAATATATGCAGTC

CCGCTGCGCGGCCATGGAGACCGATCAGTATTGTTCTAGAAGGTGATAGCCGAACCGGCA

AAACAATGTGGGCTCGTTCACTGGGACCCCATAATTATCTTTGTGGACACCTGGATCTGT

CTCCCAAGGTATATTCAAACGACGCCTGGTACAACGTCATTGATGACGTCGACCCCCACT

ACCTCAAACACTTCAAAGAATTCATGGGGGCCCAAAGAGACTGGCAAAGCAATACAAAAT

ACGGGAAGCCAATTCAAATTAAAGGCGGCATTCCCACTATCTTCCTCTGCAATCCAGGAC

CAACATCATCATATAAAGAGTTTCTGGACGAGGAAAAGAACCAATCCCTTAAAGCCTGGG

CTTTAAAGAATGCCACCTTCATCACCCTCCACGAGCCATTGTTCTCTAGTGCCCATCAAA

GTCCAACACCGCACAGCGAAGACCAGGGCCGTCCGACGTAGGCGGGTAGACCTCGAATGC

GGCTGCTCGTTCTATCTCCATATCGACTGCATCAACCATGGATTCTCGCACAGGGGAACT

CATCACTGCGCCTCAAGCAAGGAATGGCGTTTTTACCTGGGACATAACAAATCCCCTCTA

TTTCGAAATCACCGACCACGACAAGAGGCCAGGGAACATGAACCACGACATCATCACACT

CCAGATACGGTTCAACCACAACCTTCGGAAGGCATTGGGGATTCACAAGTGTTTTCTCAA

CTTCAAGGTCTGGACGACCTTACGGCCTCAGACTGGTCGTTTCTTAAGAGTATTTAAATA

TCAAGTGCTCAAGTATTTAGATATGATAGGCGTTATTTCCATTAACACTGTCCTTCAAGC

TGTTGATCATGTGGTGTACGATGTATTACTAAACACACTCCAAGTTACGGAGCAACATGT

AATAAAATTCAACCTTTATTAATTTGTCACTGCATCATAAAAATAGATGCGTATTTTCAG

CGTAGCATACACAGGATTCGAGGCATGTGTACATGCCATATACAATAACAACGCATTCTC

TGTGTGATTCTCATACTTCCCTGCCTCTTGATGATTATATGTCACGTGATGATTCAGCTT

GTAAAACCTTTTCACCAACGCCTGCTCCTTCACGCCAGATGGACCACCAACAACAGTGGC

ATGAAATTTCCTCAACACCTGAAACCTATCCCTCAAATCGTTCTTAATTGTTGCAGTACT

GGGCTCATTATCAAACATGTTAAATATCTGCCCAAAGCCTTGGGGCGCATTGCCATACGG

CCTTCTATCCCTAAGCAGGTAAAACATCACATTATTAGTGTGATTCTGCTTCTTAATATT

TTCATCCATCCAGATCTTACCAAGAATGTAAATGGACTTGATACAAAACCTCTTTCCGAC

CCTGTGTGTCAGCCCAGGCCCACGCGTCACATCACTAATCACCTTACAGATACCAAGGTG

CTTCACATCATCCCTCTGCTCAAACGACTGGACCTTACATGGGCCTTCACAGCCCTTAGG

GACATCTGGGCTTCGATACATTCTGTACATCTTGGGTTTCCGATACATGGGCCTGTTGGC

CCATATTTTGCTTCTGGTGACGCGGACAGTGGGGGCAACAACACGGTTCGTGTATGGGCT

GTCGAAGTTCAGCCTCCTCCGCACCTTGGATACGGGTGTTGAGATTATTATATCTCCTGG

TCGCTTCGACATAATTCTTACACCGTAGAACCCCTATTAGATCCCGTATATACTCAGCCC

CGACAGTACCGCGATCGTATTCCTGTTCCAGATGTAACAGGTATTTAACCGCAAGCATAG

AACGGAAACCGTGAACGGTTTCGGGAAAATCGTTCACCAATGGATCCCACATGTTGACGC

GCTCTACTACTTCGCGACGAAGTCTATAAAGACAAACAACAAATATCTAGACTTTCACGC

GTGAATATGACTGGCCGACCAAAACAAGTGCGCAGGGACCACTTTCTTTCACGAACGCGG

TCATTGTGGGATCCACCTGCTTTTTCGGGCGCGGCCATCCGGTAATATTAATCGGATGGC

CGCCAATATTCGGAATTCAAATTTTGAATGGAAGTCTAC

>HE616781.1 ACMBFV *African cassava mosaic Burkina Faso virus*

GTTGGGATTATTACGACAATGCCATTTGGTGCTAAGTATATATAGAACCCCAATACACTA

GGGATAAATCACAGAAGAGTCAATCGGTGCTAGTTGACCAAATGGCTCCTCCTCGAAAAT

TCAGAATTAACTCCAAAAATTATTTCCTCACATATCCCAAATGCTCTCTCACAAAAGACG

AAGCACTTTCCCAAATCAGAAACTTAGAAACACCAACAAACAAAAAATACATCAAAATCT

GCAAAGAATTACACGACGATGGGGAGCCTCATCTCCATGTGCTTATTCAGTTCGAAGGAA

AATACAACTGCCAAAATCAACGATTCTTCGACCTGGTATCCCCAACCAGGTCAGCACATT

TCCATCCAAACATTCAAGGAGCTAAATCAAGCTCCGATGTCAAGTCCTACATCGACAAGG

```
ACGGAGACACCCTTGAATGGGGAGAATTCCAGGTCGACGGAAGAAGTGCTAGGGGAGGCT
GCCAGAACGCTAACGACGCTGCAGCAGAGGCCTTAAATGCAGGTTCCGCTGACGCCGCTA
TGGCTATTATTAGGGAGAAACTCCCTAAAGAATTTATTTTTCAATATCATAATTTAAAAA
ATAATTTAGATAGGATTTTCATGGCCCCTCCAGAGCCATATATTTCTCCTTTTTTATCTT
CTTCTTTTACTCACGTTCCGGAAGAACTTGAAGACTGGGTTTCTGAGAACGTCATGAGTT
CCGCTGCGCGGCCTTGGAGACCGACTAGTATAGTTATTGAAGGTGATAGTCGTACAGGCA
AGACAATGTGGGCCAGGTCATTAGGTCCACACAATTACTTGTGTGGGCATCTAGATCTGA
GTCCAAAGGTGTATTCAAATGATGCTTGGTATAACGTCATTGATGACGTCGATCCCCACT
ATCTCAAACACTTTAAAGAGTTCATGGGGGCCCAAAGAGACTGGCAAAGCAACACCAAAT
ACGGGAAGCCCATTCAAATTAAAGGTGGAATTCCCACTATCTTCCTCTGCAATCCAGGAC
CTACCTCGTCCTATAAAGAGTTCCTAGACGAGGAAAAGCAACAAGCGCTAAAGGCTTGGG
CATTAAAGAATGCAATCTTCATCACCCTCACAGAACCACTCTACTCCGGTTCCAATCAAA
GTCAGTCACAGACAATTCAAGACGCGAGCCATCCGGCGTAGGAGAGTGGATCTTGTTTGT
GGCTGTTCATATTACCTCCATATCAACTGCTCCAATCATGGATTTACGCACAGGGGAACT
CATCACTGCTCCTCAAGCAATGAATGGCGTGTATACCTGGGAAATAAACAATCCCCTGTA
TTTCACAATAACCGAGCACCAACAACGCCCATTCCTGATGAACCAGGACATCATAACCGT
CCAAGTTCGATTCAATCACAACCTGAGGAAGGAGTTGGGGATTCACAAATGTTTTCTAAC
CTTCAGGATCTGGACGACCTTACGGCCTCAGACTGGTCTTTTCTTAAGGGTCTTTAGATA
CCAAGTACTTAAATACTTGGATAATATAGGTGTTATCTCAATTAACAGTGTAATTAGAGC
TGCTGATCATGTTTTGTTCAATGTAATTGAAAAAACTATTGAGTGTCAGTTGACTCATGA
AATAAAATTCAATGTTTATTAATTCCCAATACTGTCATAGAAGTATATACGTATTTTCAA
CGTAGCATATACAGGATTGGAGGCATGAGTACATGCCATATACAAAAGCAATGCATTCTC
TGTGTGATTCTCGTATTTCCCTGCCTCTTGATGATTATATGTCACGTGATGATTCAACCT
GTAAAACCTTTTCACCAACGCCTGCTCCTTCATGCCAGATGGACCACCAATAACAGTGGC
ATGAAATTTCCTCAAAACCTGAAACCTATCCCTCAAATCGTTCTTAATTGTTGCAGTACT
GGGCTCATTATCAAACATGTTAAATATCTGCCCAAAGTCTTGGGGCGTATTGCCATAAGG
CCTTCTATCCCTAAGCAGATAAAACATTACATTATTCGTGTGATTCTGCTTCTTAATATT
TTCATCCATCCAGATCTTACCAAGAATGTAAATGGACTTGATACAAAACCTCTTACCGAC
TCTGTGTGTCAGCCCAGGCCCACGTGTCACATCACTAATCACCTTACAGATACCAAGGTG
CTTCACATCATCCCTCTGCTCAAACGACTGGACCTTACATGGACCTTCACAGCCCCTAGG
TATGTCTGGGCTTCTATACATCCTGTACATCATGGGCTTTCTGTACATGGGCCTGTTCAT
CCAGGCCCGTTTTCGATTTGTGACGTGGACAGTGGGGGCAGTAGCACGGTTCCTGTATGG
GCTGTCGAAGTTCAGCCTTCGACGAACCTTGGATCCTGGCGTGGAAATGATGATATCTCC
```

TGGACGCTTCGACATAATTACGAGCCCTGATAACCGAGATTAGATCTCTAACCAGATCGT

GGCCCAAAGTACTGGGCTCGTATGTATCCTCTAAGGCCTGCAAATATTTAATTGCAAGCA

TACACCTCAGCCCATGCACCGAGTCTGGAAACTCATTCACCAGTGGATCCCACATTGCGC

ACTAGCAACAACTTCGCCTGTAAGTATATAGTGGTCCACCACTATTAAATGACCTTTAAC

GCTACAGCATGATTGGCCGACATACATAGTGCGCGGGGACCACGTTTAAAGGGGGGAGGG

GACCACTTTTTTTTTTTTCGCGCCCACCGGTAATATTAGAACGGTGGGCGCTATGGGGGT

TCAAATTT

>HE617299 CMMGV, *Cassava mosaic Madagascar virus*

AATATTATACGGATGGCCGCTTTTGGTCCAATTTTTGAATTTTGAACATAGCTTTAACTA

ATTACCATAATGCCATTTGGAGACACTATATATTGTCTCCAATTCAGTAGAGACATCAAC

CAAGGTGTCTCCAATTGCTCTCGCATAAATGCCTCCTCCCAAGCGTTTTAAAATACAAGC

CAAAAACTATTTCCTCACATATCCCAAATGCTCTCTATCTAAACACGACGCATTATCCCA

AATCTTAAACATCCCAACTCCAACTAATAAGAAATACATCAAAGTGTGCCGAGAACTTCA

CGAAGATGGGCAACCTCATCTCCACATGCTTATTCAATTCGAAGGCAAATTCTCATGCAC

AAATAAGCGATTATTCGACCTGGTATCCCCAACAACGTCAACCCATTTCCATCCAAACAT

TCAAGGAGCCAAATCCAGCTCCGACGTCAAGTCCTACATCGACAAAGATGGGGATACAAC

TGAGTGGGGCGAATTCCAGATCGACGCAAGATCTGCTAGAGGCGGCTGCCAAAATGCTAA

TGACGCATGTGCCGAAGCCTTAAACTCACGTTCGAAGGCAGCTGCACTTCTAATTATTAA

GGAGAAACTCCCCAAAGAATTTATTTTTCAATATCATAACTTAAGTAGTAATTTAGATAG

GATTTTTCAAGAGCCACCAGCTCCCTATGTTTCTCCATTTCTGTCTTCTTCATTCGACCA

AGTTCCTGACGACCTTGAGGTCTGGGTGTCAGAAAACATTATGCATCCCGCTGCGCGGCC

TTGGAGACCGAATAGTATTGTTATTGAGGGTGATAGTCGTACAGGGAAGACAATGTGGGC

CAGATCATTGGGACCACATAATTATTTATGTGGTCATCTCGACCTCAGTCCCAAAGTCTT

CACTAATGATGCATGGTACAACATTATTGATGATGTCGATCCGCACTATCTAAAGCACTT

TAAAGAGTTCATGGGTGCACAACGAGACTGGCAAAGCAACACAAAATACGGAAAGCCAAT

TCAAATTAAAGGCGGAATTCCCACTATCTTCCTCTGCAATCCGGGGCCGACTTCTTCATA

TAAAGAATATCTCGATGAAGAAAAGAATGCATCTCTCAAAGCGTGGGCACTGAAAAATGC

AACCTTCGTCACCCTCAGCGAGCCATTGTTCACAGGTTCCCATCAAGGTCCTACACCGCA

TAGCCAAGACGAGACGCATCAGACGTAGACGTATCGACCTAAGCTGCGGCTGTTCATATT

ATCTCCACATCGACTGCATCAATCATGGATTCACGCACAGGGGCACTCATCACTGCTCCT

CAAGCGCAGAATGGCGTTTTTACCTGGGAGATAAACAATCCCCTTTATTTCACGATACCC

AGACACGACTCGAGACCGTCCCACCTGAACCACGACATCATCACCATCCAAATACGCTTC

AACCACAACATCCGGAAGGAATTGGGGATTCACAAATGTTTTCTGAACTTCCAGGTCTGG

```
ACGACCTTACACCCTCAGACTGGTCATTTCTTAAGCGTATTTAAGCGTCAAGTTCTTAAG
TATTTAGATAATGTAGGCGTTATTTCAATAAACACTGTAATTCGCGCTGTTGATCACGTA
TTGTACAATGTACTTGTAAACACACTCCAAGTTATGGAGTCCCACGAAATAAAATTTAAT
TTGTATTAATTTGTTACTGCATCATAAAAATAGATGCGTATTTTAAGCGTAGCATACACT
GGATTAGAGGCATGCGTACATGCCATATACAACAATAACGCATTCTCTGTATGATTCTCA
TACTTAGCTGCCTCCTGGTGATTATACACCACATGATTATTAATCCTAAAAAATCTCCTA
ACCAAAGCTGCTCCTTCATCCCAGAAGGACCCCCAACAACGGTGGCATGAAACTTCCGCA
TAACTTGATACCTATCCCTAAGATCGTTCTTCACAGTTGCTGTACTGGGCTCATTATCAA
ACATATTAAAAACCTGTCCAAAGTCCATGGGGCTAGTGCCATAGGGCCTTCTGTCACGGA
CTAAGAAGAACATGACCTGGTTAGTATGGTTCTGCTTCTTGATATTTTCATCCATCCATA
TCTTACCTAACACATATATAGACTTGATACAAAACCTTTTACCCACTCTATGCGTAATTC
CCGGACCACGCGTGACATCACTCACACAACGAACAGTGCCAGTATGCTTGACGTCATCTC
GCTGTTCATATGATTGAACCTTACATGGGCCTTCACAGCCACGAGGAACATCAGGGCTTT
TGAACATCCTGTACATCCTGGGCTTCCGATACATGGGTCTGAACTTCCATGATTTGAACT
TGCTTGTGCCTGGGACAATGGGGACAACAGCACGGCTGCTGTATGGGCTGTCGAAGTTCA
GCTTCCGACGCACCTTCGAGGCGGGAGTGGCAATTATTATATCGGCGGGACGCTTCGACA
TAGTCACGAGCCCTAATAACAGAAATGAGATCCCGAATTAAATCGCGGCCCAAAGTATTG
GACTCGTAGGTTACCTCCACAGCCTGCAAATATTTTATAGCTAGCATACAACGAAATCCA
TGAACCGAGTCAGGAAATTCATTTAACAATGGATCCCACATTTTGAAATGCAAAACTACT
TGTTGGACAAGCATATAAAGACCATTATAATTATGTAGTCTTCCCAGACACAATGTGATT
GGTCGACATAAATTAGTGCGTGGGGACCATTGTGGGGCCCACTTACTTTTTCGGGCGCGG
CCATCCGGT
```