Published online 24 October 2018

Nucleic Acids Research, 2019, Vol. 47, No. 2 ***e9***
doi: 10.1093/nar/gky914

A new full-length circular DNA sequencing method for viral-sized genomes reveals that RNAi transgenic plants provoke a shift in geminivirus populations in the field

Devang Mehta[1,2,*], Matthias Hirsch-Hoffmann[1], Mariam Were[3], Andrea Patrignani[4], Syed Shan-e-Ali Zaidi[5], Hassan Were[3], Wilhelm Gruissem[1] and Hervé Vanderschuren[1,5,*]

[1]Institute of Molecular Plant Biology, Department of Biology, ETH Zurich, Zürich, Switzerland, [2]Laboratory of Plant Genomics, Department of Biological Sciences, University of Alberta, Edmonton, Canada, [3]Masinde Muliro University of Science and Technology, Kakamega, Kenya, [4]Functional Genomics Center Zurich, Zürich, Switzerland and [5]Plant Genetics, TERRA Teaching and Research Center, University of Liège, Gembloux, Belgium

Received June 13, 2018; Revised September 23, 2018; Editorial Decision September 25, 2018; Accepted October 03, 2018

ABSTRACT

We present a new method, CIDER-Seq (Circular DNA Enrichment sequencing) for the unbiased enrichment and long-read sequencing of viral-sized circular DNA molecules. We used CIDER-Seq to produce single-read full-length virus genomes for the first time. CIDER-Seq combines PCR-free virus enrichment with Single Molecule Real Time sequencing and a new sequence de-concatenation algorithm. We apply our technique to produce >1200 full-length, highly accurate geminivirus genomes from RNAi-transgenic and control plants in a field trial in Kenya. Using CIDER-Seq we can demonstrate for the first time that the expression of antiviral double-stranded RNA (dsRNA) in transgenic plants causes a consistent shift in virus populations towards species sharing low homology to the transgene derived dsRNA. Our method and its application in an economically important crop plant opens new possibilities in periodic virus sequence surveillance and accurate profiling of diverse circular DNA elements.

INTRODUCTION

Advances in high-throughput sequencing technologies have produced extensive data about viral diversity, both at the species/isolate level and at higher taxonomic levels. The increasing use of high-throughput sequencing technologies has also revolutionized virology by enabling the identification and characterization of previously unknown viruses. The abundance of new virus sequence data recently led to the publication of a consensus statement proposing revisions in virus taxonomy incorporating metagenomic sequence data (1). However, the contribution of high-throughput sequencing technologies to virus detection and identification remains constrained by several technical challenges including the risk of assembling artificially chimeric viral genomes due to short sequencing read lengths (1–3). This drawback has also been catalogued by numerous studies using simulated sequencing datasets to compare different short-read sequence assembly methods for metagenomics (4–6).

Sequence-bias during virus enrichment before sequencing is also a recognized drawback during virus deep-sequencing. Amplification methods to enrich viral nucleic acids rely on using either Polymerase Chain Reaction (PCR) with primers designed to bind a conserved sequence in the genome, or random circular amplification (RCA) utilizing the unique properties of the Phi29 DNA polymerase coupled to random-nucleotide primers (7). PCR-based methods have an obvious drawback because differing primer-template affinities can result in a loss of viral templates (8). RCA relies on random primers and therefore is less prone to primer complementarity bias. However, RCA results in hyper-branched, high molecular weight, concatenated products (9) that must be linearized through the use of restriction enzymes (REs) (10) or mechanical shearing prior to Sanger or NGS sequencing. The use of REs necessarily 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 specific REs used. Another approach called polymerase cloning or 'ploning' (11) employs endonucleases and DNA repair enzymes to

*To whom correspondence should be addressed: Hervé Vanderschuren (Tel: +32 81 62 25 71; Email: herve.vanderschuren@uliege.be) and Devang Mehta (Tel: +1 587 936 1572; Email: devangmehta@ualberta.ca)

linearize RCA products, followed by shotgun sequencing and whole genome assembly.

The recent advances in long-read sequencing now permit the direct sequencing of full-length RCA products. However, this also necessitates the development of an algorithm to resolve the complex DNA sequence products of an RCA reaction. Short-read sequencing in combination with rolling circle reverse-transcription has also been leveraged to more accurately profile population-level sequence variants in RNA viruses with the CirSeq protocol (12,13). A similar approach has also been applied to linear DNA sequences to reduce sequencing errors from Illumina sequencing (14). Provided the adaptation of library preparation protocols and the development of novel bioinformatics methodologies, long-read technologies such as Single Molecule Real Time sequencing (Pacific Biosciences) offer new opportunities to overcome the challenges associated with the assembly of highly similar DNA sequences. This has previously been reported in a comparative study of Illumina and SMRT sequencing of a bacterial genome (15) as well as in a recent comprehensive analysis of SMRT sequencing technology in this journal, which highlighted the advantages offered by long-read sequencing for the identification and full-genome sequencing of viral and bacterial genomes (16).

RNA interference (RNAi), i.e. the production of short interfering RNA (siRNA) from double stranded RNA (dsRNA) substrates to induce gene silencing, has been effectively used to engineer virus resistance to RNA and DNA viruses in a number of crop plants (17,18). Field assessment of RNAi-mediated resistance against *Bean golden mosaic virus* (BGMV) in transgenic common bean (*Phaseolus vulgaris*) (19) and against *Tomato yellow leaf curl virus* (TYLCV) in transgenic tomato (*Solanum lycopersicum*) (20) indicate that the RNAi-mediated genetically engineered resistance to DNA viruses is stable under field conditions when plants are exposed to single viral species. However, the effectiveness of RNAi technology in crops suffering from infections by multiple virus species has not yet been assessed. Because RNAi technology depends on sequence complementarity between the transgene and the targeted viral sequences, it is essential to determine the degree of sequence complementarity required for broad spectrum virus resistance in the field.

Here, we report the development of a new technique for sequence-independent viral DNA enrichment utilizing assembly-free Single Molecule Real Time (SMRT; Pacific Biosciences Inc.) long-read sequencing called CIDER-Seq (Circular DNA Enrichment sequencing). CIDER-Seq uses a ploning-based enrichment protocol and requires only an estimate of virus genome size and, optionally, a single reference sequence. We have also developed a new algorithm, DeConcat, to parse concatenated DNA strands that are generated by rolling circle amplification (RCA), allowing us to sequence complete geminivirus genomes without reference-based or *de novo* assembly. In this study, CIDER-Seq allows us to demonstrate the effect of anti-viral RNAi technology on changes in target plant virus populations, as well as profile the sequence complementarity required for effective RNAi-based virus resistance.

MATERIALS AND METHODS

Field trial design and planting

The experiment was conducted in a confined field trial at the Kenya Agricultural and Livestock Research Organisation (KALRO) site at Alupe, Kenya. The field was divided into three replicated blocks of equal size. Each block contained randomised plots of six plants each, per line. Plots were divided by infector plants which were multiplied from naturally infected, symptomatic plants collected from different locations in West Kenya. The entire field was surrounded by border rows 2 m wide. The study was authorised by and conducted according to the containment measures set by the National Biosafety Authority of Kenya.

Transgenic and control plants were multiplied *in vitro* at ETH Zurich and 4-week-old plantlets were shipped to Kenya, potted and hardened in a level II screen house for 60 days. The plants were transferred to the soil in the CFT and left to grow for 49 weeks.

Scoring and monitoring

Eight weeks after planting, the plants were monitored weekly for virus symptoms, whitefly populations, plant vigour and height. Severity scoring was done using a five-point Likert scale according to Ogbe *et al.* (21). CMD incidence was calculated as the proportion of infected plants in the plot. In accordance with the biosafety guidelines, whenever flower buds were spotted they were removed and incinerated.

Sampling and DNA extraction

At harvest, symptomatic leaves were harvested from infected 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 (22) combined with an ethanol precipitation step. Total nucleic acid was quantified using a Qubit dsDNA BR Assay Kit (Q32850, Thermo Fisher 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 and 5 kb 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. An additional sequencing library was produced without the post-enrichment size selection step. This library is referred to as NSS in the following text.

Random circle amplification

Random rolling circle amplification was performed as previously described (7) with some modifications. A 20 μl re-

action was set up using 5 μl of size-selected template DNA, 1 mM dNTPs, 10 U Phi29 DNA polymerase (EP0092, Thermo Fisher Scientific), 50 μM Exo-resistant random primer (SO181, Thermo Fisher Scientific), 0.02 U inorganic pyrophosphatase (EF0221, Thermo Fisher Scientific) and 1× Phi29 DNA polymerase buffer (supplied with enzyme). The reaction was run at 30°C for 18 h and stopped by heating to 65°C for 2 min. 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 h and stopped by heating at 65°C for 2 min. The product was precipitated with sodium acetate/ethanol. The purified product was treated with 50 U S1 nuclease (EN0321, Thermo Fisher Scientific) in a 20 μl reaction at 37°C for 30 min and stopped by adding 3.3 μl of 0.5 M EDTA and heating at 70°C for 10 min. DNA was purified by sodium acetate/ethanol precipitation.

DNA repair

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

Semi-quantitative PCR

Semi-quantitative PCR was performed using the primers described in Supplementary Table S1. DreamTaq polymerase (EP0705, Thermo Fisher 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 1× 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™ 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 12 kb 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 1 SMRT cell v3.0 (100-171-800, Pacific Biosciences), taking one movie of 360 min. 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.

The organisation of samples per SMRT Cell is depicted in Supplementary Table S2.

CIDER-Seq data analysis

Following SMRT sequencing and the generation of barcode-separated subreads (23) we followed a custom data analysis method. 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) = 3000 (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 (24). The modified sequences consisted of: (a) a thrice-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. The sequence between these two points (called

the trimmed ROI) was further analysed using DeConcat (Figure 2a).

DeConcat algorithm description

DeConcat begins (**Step 1**, Figure 2A) 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 Figure S1) 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**, Figure 2A). If the smaller segment lies completely within the larger one (**Step** 5, 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' and 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

DeConcat performance stats were derived by analyzing the data from SMRT Cells 1 and 4 (also known as SS for size selected and NSS for non-size selected respectively).

After running DeConcat on, 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 Figure S2a), 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 Figure S2B).

We also found that of the eight possible alignment cases in DeConcat (see Materials and Methods), cases 1a, 1b and 3 were the most frequent (Supplementary Figure S2C) 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 Figure S2D). 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. 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 Figure S3). 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. The pipeline could handle reads with low accuracy (i.e. 85%) but increasing the quality threshold led to a reduction in the number of full-length viral genome sequences produced (Supplementary Figure S4).

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.

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 (25). 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 Figure S5A) and the frequency of frameshifts in each viral protein increased linearly with protein length (Supplementary Figure S5b), indicating that they are likely caused by SMRT sequencing errors. Between 5 and 8% of reads in the >99.99 quality threshold dataset have no frameshift errors (Supplementary Figure S5C).

Percentage identity calculation

Phased sequence reads and reference sequences (also phased) were first trimmed to ensure identical start and end positions. Trimmed, phased reads were pairwise aligned to the reference sequence (either ACMV-NOg full genome, for Figure 2A or the 146 bp dsRNA sequence, for Supplementary Figure S6) using the Pairwise2 package in BioPython. Identity scores were calculated for each pairwise alignment and plotted on a density plot using the ggjoy/ggridges R package.

Phylogenetic analysis

Trimmed, phased reads were aligned using MUSCLE implemented in the CLC Genomics Workbench 10.0 using default parameters. The alignment was used to create a Neighbour-Joining tree with 100 bootstraps. The phylogeny was used to create an unrooted cladogram and clades were defined based on the position of the seven reference sequences used.

siRNA simulation

We first divided the dsRNA complementary region of the transgene into 21nt fragments using a 1nt sliding window. These putative siRNAs were then locally aligned (with a high gap open/extend penalty) against each full-length trimmed, phased virus sequence read obtained (see above) per sample. The number of mismatches present in each alignment was recorded. Results were tabulated showing the number of siRNA with no mismatches, with 1 mismatch, and so on as plotted in Figure 2B. This process was implemented in Python and the code is available for download below.

Identification of cassava endogenous episomal circular DNA

All raw sequence reads from cells 2 and 3 (3242 sequences) were BLASTed against the cassava genome (variety 60444, in-house sequenced and assembled). Reads with >99% identity to the cassava genome (only 42 sequences) were selected and subjected to DeConcat. Reads that passed through at least 1 round of DeConcat (and hence definitively derived from circular templates) were next selected as being circular cassava DNA detected by CIDER-Seq. These 12 reads were also BLASTed against the NCBI nucleotide database to obtain descriptions of closest matches.

RESULTS

We demonstrate the effectiveness of CIDER-Seq by producing 1328 full-length genomes of cassava mosaic geminiviruses (CMGs). CMGs are whitefly-transmitted, bipartite viruses of the genus *Begomovirus* in the family *Geminiviridae*—the most populous family of eukaryote-infecting viruses (26). They are comprised of two separate genomes, designated DNA A and DNA B (17). CMGs cause cassava mosaic disease (CMD) and severe economic losses for farmers, particularly in sub-Saharan Africa (17). To date, nine CMG species have been identified that share 68–90% sequence identity based on Sanger sequencing of PCR and RCA products (17). We utilized CIDER-Seq to profile virus populations from a field trial of previously developed transgenic (27) and control cassava lines, conducted in Kenya. The transgenic cassava lines (dsAC1-101 and dsAC1-152) produce 155bp long dsRNA targeting the conserved 3'-end of the *AC1* gene (located on DNA A) from African cassava mosaic virus (ACMV) species (27). The field trial was conducted in three replicated randomized blocks surrounded by infector rows consisting of cassava plants infected with CMD (previously collected from different parts of W. Kenya). Disease symptom, meteorological and whitefly count data was recorded weekly, starting at 8 weeks after planting (Supplementary Figure S7A–E). Transgenic plants showed no symptoms until week 17, although by the end of the experiment these lines showed between 60 and 70% mean disease incidence (Supplementary Figure S7A, B). Leaf samples were collected at the completion of the field trial and assessed for virus levels (Supplementary Figure S7F) and sequence diversity.

Using leaf samples from cassava plants with CMD symptoms, we first enriched complete DNA genomes of CMGs (Figure 1A) 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' (11) 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 (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. Semi-quantitative PCR was used to assess the relative importance of each successive step in the enrichment protocol (Supplementary Figure S8).

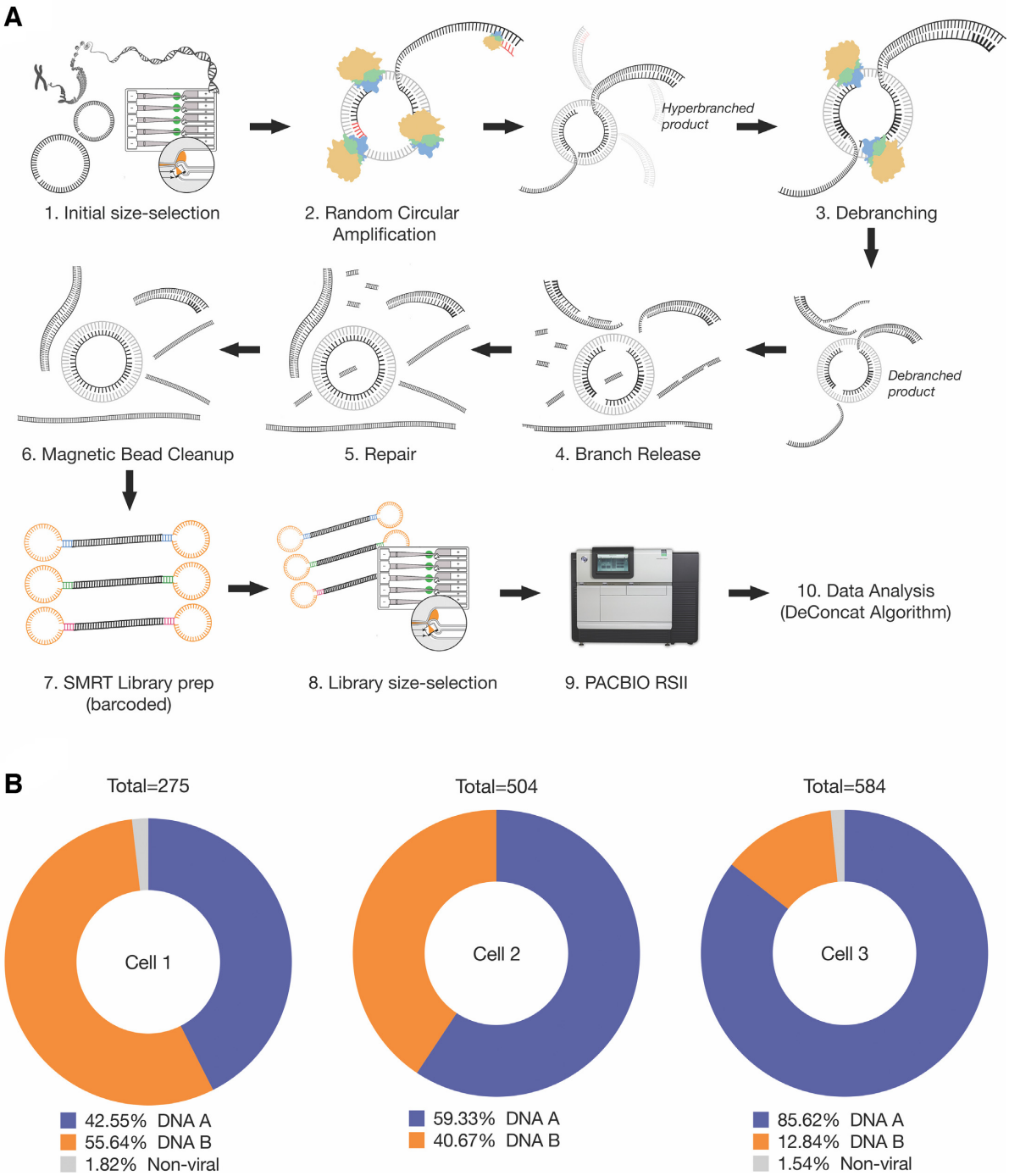

Figure 1. CIDER-Seq Enrichment methodology and results. (**A**) Enrichment of circular DNA based on automated size selection, non-denaturing random circular amplification (RCA), linearization and repair of the RCA product followed by Single Molecule Real Time (SMRT) library creation. (**B**) Proportion of viral (DNA A & DNA B) and non-viral reads produced in each SMRT Cell.

The enriched samples were next barcoded and pooled to build 4 separate SMRT-sequencing libraries (Supplementary Table S2). SMRT Cell 1 (and 4) were used primarily to refine the CIDER-Seq and DeConcat methods and Cells 2 and 3 were used for analyzing virus populations in transgenic and control plant lines. The raw sequencing reads were error-corrected by the Circular Consensus Sequencing (CCS) pipeline (Pacific Biosciences Inc.). CCS error-correction was implemented using a highly stringent cut-off of >99.9 predicted quality. Functionally, this cut off resulted in the algorithm implementing a mean number of passes ranging from 15–19 (Supplementary Table S2). This is far greater than the commonly cited pass threshold of 3 used in other SMRT sequencing applications (28).

As expected, a majority of CCS reads were greater than the 3 kb size cut-off, indicating that they were comprised of

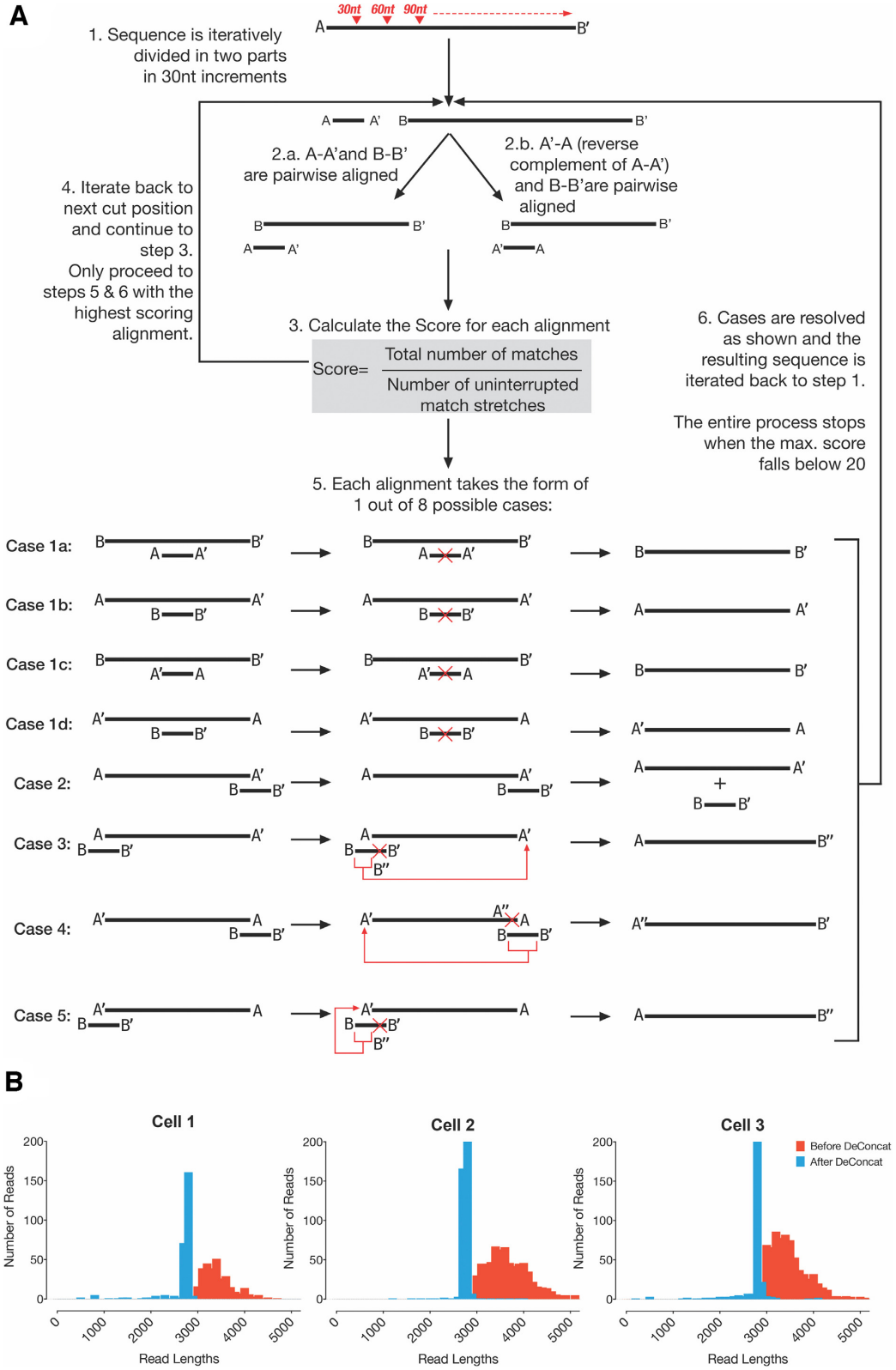

Figure 2. DeConcat algorithm description and results. (**A**) DeConcat algorithm scheme (see Methods for details). (**B**) Size distribution of sequencing reads before (Red) and after (Blue) DeConcat processing per SMRT Cell.

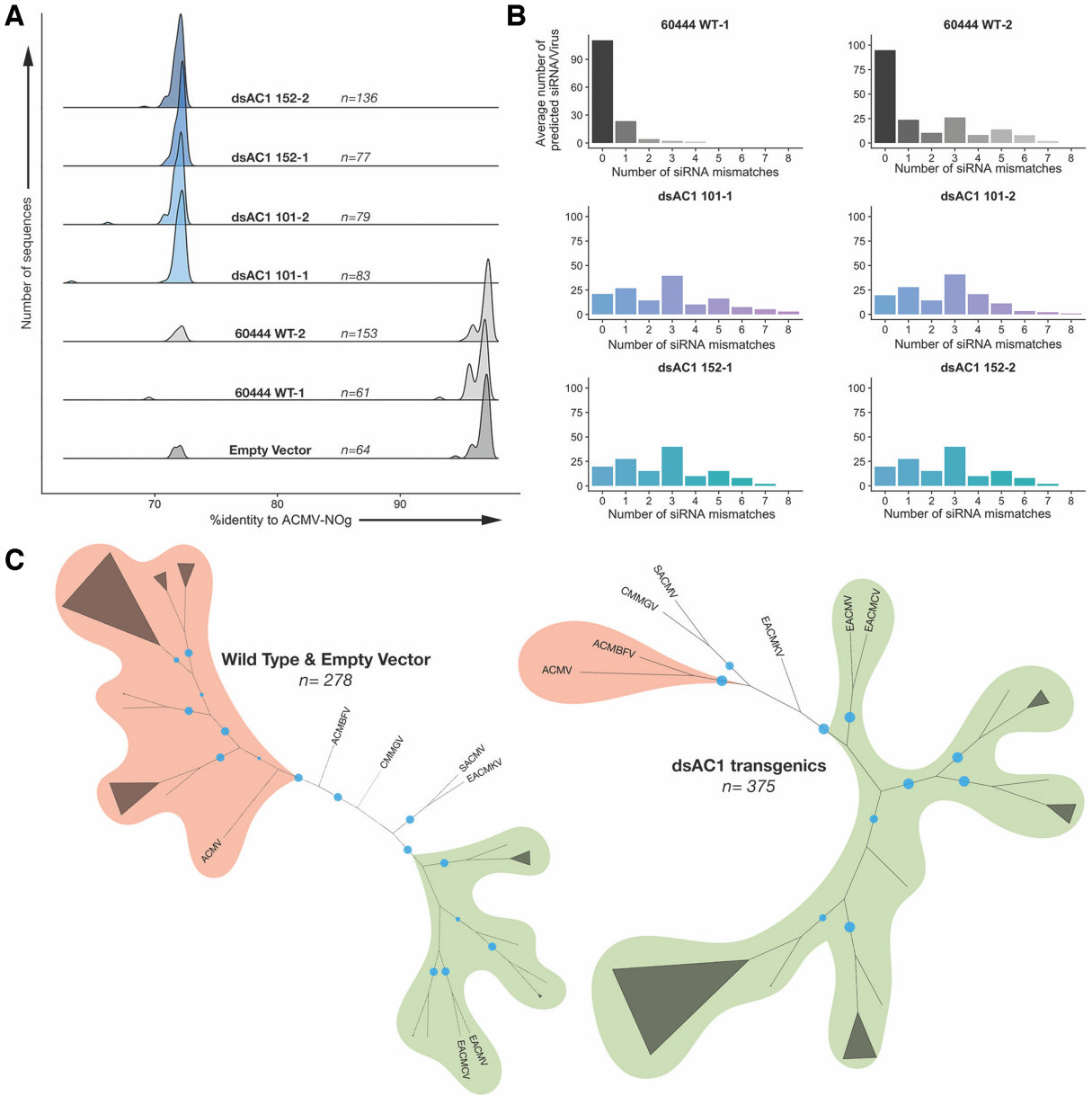

Figure 3. Virus genome sequence analysis from transgenic and control plants. (**A**) Density plot showing the proportion of virus sequences with their % identity to the ACMV-NOg genome from which the transgene was derived. (**B**) The average number of in silico predicted, transgene-derived 21nt siRNA which match each virus sequence obtained per sample, binned by the number of mismatches found in each case. (**C**) Neighbor joining trees (100 bootstraps) constructed with full length genome sequences from transgenics and control plants, with reference genomes from 7 African CMG species, rooted with the ACMV-NOg sequence. Blue circles represent nodes with >70 bootstrap support.

concatenated RCA products. We developed a custom data analysis pipeline to resolve the concatenated sequences into their component parts. First, in a basic filtering step, all reads were binned into non-viral, CMG DNA A and CMG DNA B sequence categories. The results from this step indicated that at least 98% of the reads in all libraries were contained viral sequences (Figure 1B). Next, we trimmed non-viral DNA from the ends of each read by performing a circular DNA–aware multiple sequence alignment (see Materials and Methods), which did not significantly affect the size distribution of the dataset. These trimmed reads were next passed through the de-concatenation algorithm we termed 'DeConcat' (Figure 2A, see Materials and Methods). The resulting de-concatenated reads had a greatly reduced size distribution (Figure 2B), with a clear peak at 2.8 kb in all three SMRT cells. Thus, DeConcat was able to reduce RCA-generated SMRT-Sequenced reads of a wide size-distribution into expected CMG-length virus sequences. This result validated the parameters used in the DeConcat algorithm and also demonstrated that the Phi29 DNA polymerase enrichment step amplified circular DNA in the 3 kb range with high specificity.

We performed a tBLASTn using the final DeConcat reads against virus protein reference sequences to anno-

tate the virus genomes. This also allowed us to detect and classify sequencing errors in our final genomes (Supplementary Figure S5, Materials and Methods). Since SMRT sequencing usually results in insertion-deletion errors, we scanned our final genomes for frameshift mutations in protein coding regions, a widely-used indicator for indel errors. This analysis was run on our benchmarking SMRT cells (Cells 1 and 4), using different thresholds for the CCS error correction step that precedes DeConcat in the CIDER-Seq pipeline. We found that the numbers of frameshift mutations per genome decreased with increasing CCS quality thresholds (Supplementary Figure S5). At the highest quality threshold (99.9), we found only 1 frameshift error per genome suggesting highly successful error correction. In future applications of CIDER-Seq where reference genomes are available (i.e. applications apart from conventional metagenomics), this low-level of frameshift error can be further corrected using existing long-read sequencing-specific frameshift-correction algorithms such as FrameFro (29) or RIFRAF (30).

We next performed a multiple sequence alignment of the full-length genomes against the reference virus sequence (ACMV-NOg isolate) from which the transgene was originally derived. This allowed us to assess the proportion of viruses in each plant line which can be efficiently targeted by the transgene. We found that the control plants had two populations of viruses, one with >90% identity to the reference ACMV-NOg and one with <75% (Figure 3A). In the transgenic lines, on the other hand, no sequences with greater than 90% identity to the target virus were detected and all the virus genomes belonged to the <75% identity category. This suggests that the transgenic plants could effectively limit target CMG species, leading to a surge in the proportion of non-target CMG species. Similar results were found upon analyzing only the virus sequences corresponding to the transgene region (Supplementary Figure S6). We next simulated the production of anti-viral 21nt siRNAs from each plant line. Putative siRNA sequences were aligned against each virus genomes and plotted according to the number of mismatches between the siRNA and the target virus. While, on average, between 75 and 100 transgene-derived siRNAs could target each virus per control sample with no mismatch, only between 10 and 15 siRNAs could target each virus in the transgenic lines with complete fidelity (Figure 3B). The presence of identical levels of predicted siRNAs with single mismatches in control and transgenic samples suggests that resistance mainly relies on perfectly matching siRNAs. Phylogenetic trees from all viruses in control and transgenic lines, along with reference genomes from all 7 CMG species found in Africa revealed that the non-target viruses identified in Figure 3A are closely related to the *East African Cassava mosaic virus* and the *East African Cassava mosaic Cameroon virus* species (Figure 3C).

We also investigated the non-viral sequences found in our initial dataset and found 12 circular DNA sequences (ranging in size from 180 to 3024 bp) matching the genome of the cassava host plant (Supplementary Table S4). The presence of these molecules suggests a future research opportunity using CIDER-Seq to thoroughly characterize mobile circular episomal DNA molecules (such as Helitron transposable elements) in various eukaryotes.

DISCUSSION

In summation, CIDER-Seq effectively enriched circular DNA molecules and produced single-read, full-length sequence data from RCA reaction products for viral sequencing. The DeConcat algorithm parsed 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. Using CIDER-Seq we concluded that virus populations are radically changed in transgenic plants due to the expression of anti-viral dsRNA. Using infected field material, we could generate 1328 high-quality, non-chimeric, full-length virus genome sequences, representing more than twice the number of all CMG sequences deposited in GenBank to date. Moreover, according to our estimates CIDER-Seq reduces the per genome cost of sequencing by approximately 17-fold as compared to the conventional RCA and Sanger sequencing method.

RNAi technology is currently being trialed for deployment in a number of crops. Our field study shows that transgenic cassava plants deployed in the field trial are able to effectively eliminate target virus species but are ineffective against species with <90% identity to the transgene derived dsRNA. Thus, based on the CIDER-Seq results, we conclude that profiling virus sequence diversity is a necessary step prior to the development of new virus-resistant transgenic plants. Our results outlining the sequence specificity requirements for effective RNAi in field scenarios are important for the design of future RNAi constructs against viruses which are present in complex populations. Based on our results we anticipate a new mode of RNAi technology deployment where new, updated anti-viral constructs can be regularly designed based on periodic assessments of virus sequence diversity using CIDER-Seq.

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* (31). 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 high-throughput genomic data in virus classification[1], CIDER-Seq is a superior method for high-quality full-genome sequencing of circular viruses infecting plants, and possibly animals and bacteria that will facilitate building accurate sequence datasets for virus taxonomy and evolutionary studies. The circular dsDNA *Human papillomavirus* is one such important target for future CIDER-Seq experiments. The sequencing of non-viral circular DNA templates such as plasmids and circular transposable elements (Helitrons) is another potential application of CIDER-Seq.

DATA AVAILABILITY

Full length annotated genome sequences, raw sequence data and data generated during intermediate CIDER-Seq and siRNA simulation steps are freely available at: www.dx.doi.org/10.5281/zenodo.830530 and at: www.dx.doi.org/10.5281/zenodo.1009036.

Python packages for the CIDER-Seq Data Analysis software and DeConcat, along with installation and usage guidelines are available at: http://www.dx.doi.org/10.5281/zenodo.834928. Python scripts for siRNA counting, sequence extraction and alignment, and R scripts for plotting data are available at: www.dx.doi.org/10.5281/zenodo.1009036.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to the staff at the Kenya Agricultural Research Institute and the Masinde-Muliro University of Science and Technology, Kenya for their assistance with the field trial and the National Biosafety Authority for approvals and guidance. We thank Irene Zurkirchen for maintaining plants in the glasshouse and Sukalp Muzumdar for help with field data analysis. We thank Weihong Qi (Functional Genomics Center Zurich, Switzerland) for assistance with SMRT sequencing as well as helpful advice and discussion.

FUNDING

ETH Zurich; Stiftung fiat panis to perform a confined cassava field trial in Kenya; European Union's Seventh Framework Programme for research, technological development, and demonstration [EU GA–2013–608422-IDP BRIDGES to D.M.]. Funding for open access charge: ETH Zurich.
Conflict of interest statement. None declared.

REFERENCES

1. Simmonds,P., Adams,M.J., Benkő,M., Breitbart,M., Brister,J.R., Carstens,E.B., Davison,A.J., Delwart,E., Gorbalenya,A.E., Harrach,B. *et al.* (2017) Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.*, **15**, 161–168.
2. Massart,S., Candresse,T., Gil,J., Lacomme,C., Predajna,L., Ravnikar,M., Reynard,J.S., Rumbou,A., Saldarelli,P., Škoric,D. *et al.* (2017) A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Front. Microbiol.*, **8**, 45.
3. Wu,Q., Ding,S., Zhang,Y. and Zhu,S. (2015) Identification of viruses and viroids by Next-Generation sequencing and homology-dependent and homology- independent algorithms. *Annu. Rev. Phytopathol.*, **53**, 425–444.
4. Mavromatis,K., Ivanova,N., Barry,K., Shapiro,H., Goltsman,E., McHardy,A.C., Rigoutsos,I., Salamov,A., Korzeniewski,F., Land,M. *et al.* (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, **4**, 495–500.
5. Mende,D.R., Waller,A.S., Sunagawa,S., Järvelin,A.I., Chan,M.M., Arumugam,M., Raes,J. and Bork,P. (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One*, **7**, e31386.
6. Pignatelli,M. and Moya,A. (2011) Evaluating the fidelity of De Novo short read metagenomic assembly using simulated data. *PLoS One*, **6**, e19984.
7. Dean,F., Nelson,J., Giesler,T. and Lasken,R. (2001) Rapid amplification of plasmid and phage DNA using Phi29 polymerase and a multiply-pimed rolling circle amplification. *Genome Res.*, **11**, 1095–1099.
8. Sipos,R., Szekely,A., Revesz,S. and Marialigeti,K. (2010) Addressing PCR biases in environmental microbiology studies. In: Cummings,SP (ed). *Methods in Molecular Biology: Bioremediation*. Humana Press, NY, Vol. **599**, pp. 37–58.
9. Lasken,R.S. and Stockwell,T.B. (2007) Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.*, **7**, 19.
10. Inoue-Nagata,A.K., Albuquerque,L.C., Rocha,W.B. and Nagata,T. (2004) A simple method for cloning the complete begomovirus genome using the bacteriophage Phi29 DNA polymerase. *J. Virol. Methods*, **116**, 209–211.
11. Zhang,K., Martiny,A.C., Reppas,N.B., Barry,K.W., Malek,J., Chisholm,S.W. and Church,G.M. (2006) Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.*, **24**, 680–686.
12. Acevedo,A. and Andino,R. (2014) Library preparation for highly accurate population sequencing of RNA viruses. *Nat. Protoc.*, **9**, 1760–1769.
13. Acevedo,A., Brodsky,L. and Andino,R. (2014) Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, **505**, 686–690.
14. Lou,D.I., Hussmann,J.A., Mcbee,R.M., Acevedo,A., Andino,R. and Press,W.H. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 19872–19877.
15. Teng,J.L.L., Yeung,M.L., Chan,E., Jia,L., Lin,C.H., Huang,Y., Tse,H., Wong,S.S.Y., Sham,P.C., Lau,S.K.P. *et al.* (2017) PacBio but not illumina technology can achieve fast, accurate and complete closure of the high GC, complex Burkholderia pseudomallei two-chromosome genome. *Front. Microbiol.*, **8**, 1–15.
16. Ardui,S., Ameur,A., Vermeesch,J.R. and Hestand,M.S. (2018) Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.*, **46**, 2159–2168.
17. Rey,C. and Vanderschuren,H. (2017) Cassava mosaic and brown streak diseases: current perspectives and beyond. *Annu. Rev. Virol.*, **4**, 429–452.
18. Pooggin,M.M. (2017) RNAi-mediated resistance to viruses: a critical assessment of methodologies. *Curr. Opin. Virol.*, **26**, 28–35.
19. Aragão,F.J.L. and Faria,J.C. (2009) First transgenic geminivirus-resistant plant in the field. *Nat. Biotechnol.*, **27**, 1086–1088.
20. Fuentes,A., Carlos,N., Ruiz,Y., Callard,D., Sánchez,Y., Ochagavía,M., Seguin,J., Malpica-López,N., Hohn,T., Lecca,M. *et al.* (2016) Field trial and molecular characterization of RNAi-transgenic tomato plants that exhibit resistance to tomato yellow leaf curl geminivirus. *Mol. Plant-Microbe Interact.*, **29**, 197–209.
21. Ogbe,F.O., Atiri,G.I., Dixon,A.G.O. and Thottappilly,G. (2003) Symptom severity of cassava mosaic disease in relation to concentration of African cassava mosaic virus in different cassava genotypes. *Plant Pathol.*, **52**, 84–91.
22. Chang,S., Puryear,J. and Cairney,J. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Report.*, **11**, 113–116.
23. Pacific Biosciences Inc (2015) *Pacific Biosciences Glossary of Terms*. https://www.pacb.com/wp-content/uploads/2015/09/Pacific-Biosciences-Glossary-of-Terms.pdf.
24. Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
25. Laehnemann,D., Borkhardt,A. and Mchardy,A.C. (2016) Denoising DNA deep sequencing data — high-throughput sequencing errors and their correction. *Brief. Bioinform.*, **17**, 154–179.
26. International Committee on Taxonomy of Viruses (2016) *ICTV Master Species Listv1.3*. https://talk.ictvonline.org/files/master-species-lists/.
27. Vanderschuren,H., Alder,A., Zhang,P. and Gruissem,W. (2009) Dose-dependent RNAi-mediated geminivirus resistance in the tropical root crop cassava. *Plant Mol. Biol.*, **70**, 265–272

28. Jiao,X., Zheng,X., Ma,L., Kutty,G., Gogineni,E., Sun,Q., Sherman,B.T., Hu,X., Jones,K., Raley,C. *et al.* (2013) A benchmark study on error assessment and quality control of CCS reads derived from the PacBio RS. *J. Data Min. Genomics Proteomics*, **424**, 1–9.
29. Du,N. and Sun,Y. (2016) Improve homology search sensitivity of PacBio data by correcting frameshifts. *Bioinformatics*, **32**, i529–i537.
30. Eren,K. and Murrell,B. (2018) RIFRAF: a frame-resolving consensus algorithm. *Bioinformatics*, bty426.
31. Ssemadaali,M.A., Effertz,K., Singh,P., Kolyvushko,O. and Ramamoorthy,S. (2016) Identification of heterologous Torque Teno Viruses in humans and swine. *Sci. Rep.*, **6**, 26655.