Going beyond consensus genome sequences: an innovative SNP-based methodology reconstructs different Uganda cassava brown streak virus haplotypes at a nationwide scale in Rwanda.

Chantal Nyirakanani, Lucie Tamisier, Jean Pierre Bizimana, Johan Rollin, Athanase Nduwumuremyi, Vincent de Paul Bigirimana, Ilhem Selmi, Ludivine Lasois, Hervé Vanderschuren, Sébastien Massart

Going beyond consensus genome sequences: an innovative SNP-based methodology reconstructs different Uganda cassava brown streak virus haplotypes at a nationwide scale in Rwanda.

Chantal Nyirakanani[1,3], Lucie Tamisier[2], Jean Pierre Bizimana[1,4], Johan Rollin[2,6], Athanase Nduwumuremyi[4], Vincent de Paul Bigirimana[3], Ilhem Selmi[2], Ludivine Lasois[1], Hervé Vanderschuren[1,5], and Sébastien Massart[2*]

[1] Plant Genetics and Rhizosphere Processes Laboratory, TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech, Gembloux, Belgium

[2] Integrated and Urban Plant Pathology Laboratory, TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech, Gembloux, Belgium

[3] Department of Crop Sciences, School of Agriculture and Food Sciences, College of Agriculture, Animal Sciences and Veterinary Medicine, University of Rwanda, Musanze, Rwanda

[4] Rwanda Agriculture and Animal Resources Development Board, Huye, Rwanda

[5] Tropical Crop Improvement Laboratory, Department of Biosystems, KU Leuven, Heverlee, Belgium

[6] DNAVision, 6041 Gosselies, Belgium

***Corresponding authors:** sebastien.massart@uliege.be ; Chantal.Nyirakanani@uliege.be; herve.vanderschuren@kuleuven.be; herve.vanderschuren@uliege.be

Abstract

Cassava Brown Streak Disease (CBSD), which is caused by cassava brown streak virus (CBSV) and Ugandan cassava brown streak virus (UCBSV), represents one of the most devastating threats to cassava production in Africa, including in Rwanda where a dramatic epidemic in 2014 dropped cassava yield from 3.3 million to 900,000 tonnes (1). Studying viral genetic diversity at the genome level is essential in disease management, as it can provide valuable information on the origin and dynamics of epidemic events. To fill the current lack of genome-based diversity

studies of UCBSV, we performed a nationwide survey of cassava ipomovirus genomic sequences in Rwanda by high-throughput sequencing (HTS) of pools of plants sampled from 130 cassava fields in 13 cassava-producing districts, spanning seven agro-ecological zones with contrasting climatic conditions and different cassava cultivars. HTS allowed the assembly of a nearly complete consensus genome of UCBSV in 12 districts. The phylogenetic analysis revealed high homology between UCBSV genome sequences, with a maximum of 0.8 % divergence between genomes at the nucleotide level. An in-depth investigation based on Single Nucleotide Polymorphisms (SNP) was conducted to explore the genome diversity beyond the consensus sequences. First, to ensure the validity of the result, a panel of SNPs was confirmed by independent RT-PCR and Sanger sequencing.

Furthermore, the combination of fixation index (F_{ST}) calculation and Principal Component Analysis (PCA) based on SNPs patterns identified three different UCBSV haplotypes geographically clustered. The haplotype 2 (H_2) was restricted to the central regions, where the NAROCAS 1 cultivar is predominantly farmed. RT-PCR and Sanger sequencing of individual NAROCAS1 plants confirmed their association with H_2. Haplotype 1 was widely spread, with a 100% occurrence in the Eastern region, while Haplotype 3 was only found in the Western region. These haplotypes' associations with specific cultivars or regions would need further confirmation. Our results prove that a much more complex picture of genetic diversity can be deciphered beyond the consensus sequences, with practical implications on virus epidemiology, evolution, and disease management. Our methodology proposes a high-resolution analysis of genome diversity beyond the consensus between and within samples. It can be used at various scales, from individual plants to pooled samples of virus-infected plants. Our findings also showed how subtle genetic differences could be informative on the potential impact of

agricultural practices, as the presence and frequency of a virus haplotype could be correlated with the dissemination and adoption of improved cultivars.

Keywords: Cassava, High throughput sequencing, UCBSV, Ampelovirus, SNP, Rwanda

1. Introduction

In many African regions, cassava (*Manihot esculenta* Crantz) is considered a key food security crop because of its capacity to cope with suboptimal climatic conditions and to grow on marginal land (2). The crop ranks as the sixth most important food crop in the world (3) and the third most important in Rwanda, with an average yield of 14.2 tons per hectare in 2021 (3–5). However, cassava production is still below its yield potential due to various constraints, including viral diseases (7). Cassava Brown Streak Disease (CBSD) is one of sub-Saharan Africa's most devastating threats to cassava. CBSD was first found in Tanzania in 1936 (8) and has now spread to ten East and Central African countries where cassava is a vital crop (9). The disease is caused by Ugandan cassava brown streak virus (UCBSV) and cassava brown streak virus (CBSV), which are both positive-sense, single-stranded RNA (+)ssRNA) virus species belonging to the genus *Ipomovirus* and the family *Potyviridae* (10). Both species are widely spread in Central and Eastern African countries, although UCBSV is often more prevalent than CBSV (11–14). The disease is mainly vertically transmitted through planting material, in addition to the semipersistent transmission by the whitefly vector (15). After a dramatic outbreak of CBSD in Rwanda in 2014, the import and dissemination of CBSD-tolerant cassava cultivars (16) were

instrumental in mitigating yield losses. However, the incidence of UCBSV and CBSV remained relatively high (respectively 61% and 11%) (12).

RNA viruses often exist as a population of closely related mutants due to the low fidelity of RNA polymerases and, therefore, exhibit a fast yet constrained evolutionary rate enabling modification in virulence and transmissibility as well as a continuous virus adaptation to the changing environment (17). Therefore, studying the evolution of viral populations at different scales is of prime importance in managing a viral disease. For UCBSV, the analyses were mainly carried out on a few partial coat protein sequences. They reported a country-wide nucleotide divergence from under 1% in Mayotte (n=8 sequences), and Kenya (n=9) (18,19), to 12% in Rwanda (n=24) (20). The first complete genome sequences of viruses were obtained by Sanger sequencing of PCR amplicons (21). They corresponded to the most frequent nucleotide at each position of the genome. However, this technique is not well adapted to detect minor alleles or Single Nucleotide Polymorphisms (SNPs - under 25% frequency) in the population. Unless coupled with cloning and sequencing of several clones per sample, amplicon sequencing generally fails to detect low-frequency polymorphisms, even though improvements have been made to detect minor alleles in tumours (22).

High throughput sequencing (HTS) has become the standard technology for studying virus population diversity, epidemiology, and evolution (11). For example, HTS-based profiling of plant virus genomes has allowed the reconstruction of the history of potato virus V evolution and dissemination (23) as well as deciphering the spread of the turnip mosaic virus along the silk road (24). For UCBSV, 63 genome sequences have been generated from Uganda, Tanzania, Kenya, DRC, Zambia, Malawi, Rwanda, and Burundi (10,11,13,25–28).

Although HTS technologies have facilitated the profiling of virus genomes, most publications only report consensus genome sequences for the detected and identified viruses, even from pooled samples, as done for the devastating viruses causing maize lethal necrosis (29). This consensus genome corresponds to the most frequent nucleotide at each position, providing an information level similar to Sanger sequencing technology and underexploiting the extent of sequencing data generated. Because HTS technologies generate millions of reads, tens to thousands of sequencing reads can cover each base of a viral genome. This sequencing coverage theoretically allows the identification of minor SNPs (frequency under 50%) even at a shallow frequency (below 1%) (11). The integration of minor SNPs in the analysis of viral genomes should become the rule rather than the exception, as they can drive evolutionary processes and the biological properties of viruses within their hosts (30–33). For example, minor variants in the Coxsackie virus have been shown to contribute to virus adaptation (34). Evolutionary studies on barley yellow dwarf virus (BYDV) showed that several virus populations might share the same consensus sequence while having different low-frequency SNPs patterns, highlighting the importance of characterizing virus genetic diversity beyond the consensus level (35). Therefore, it is critical to profile and reports the presence of individual low-frequency SNP present to improve the resolution of viral population analyses and to characterize the contribution of variants to virus evolution and adaptation (changes in viral load, virulence, transmission, host range, etc.).

When analyzing the genetic diversity of a virus species in the HTS dataset at the SNP level, generating the proper reference consensus sequence(s) is essential. Indeed, if the sample was infected by several divergent isolates, e.g., at least 5-10% of divergence, their consensus genome sequence reconstruction is possible using classical *de novo* assembly (36–38), and several

consensus sequences can be obtained. The presence and frequency of SNPs can be further studied for each genome sequence. On the other hand, if the identity between isolates is higher, it becomes difficult to differentiate their genome sequences, and a unique consensus sequence is often generated. The comparison of viral populations using SNP frequencies can be performed with the fixation index (F_{ST}) (39). F_{ST} is a measure of population differentiation usually applied to study the population genetics of pooled samples of vertebrates or plants. It has been recently applied to plant virus populations and has enabled an in-depth analysis of the virus population beyond consensus sequences reconstructed from HTS datasets (35,40). SNP frequencies can also be used as features for a principal component analysis (PCA) to reduce their complexity and increase their interpretability. SNP-based PCA has so far only been applied to study the evolution of a DNA virus infecting Drosophila (41).

In addition, the association between SNPs can be studied, e.g., if they belong to the same viral molecule. For eukaryotes, a haplotype is a set of genomic polymorphisms that tend to be inherited together. By extension, we can call haplotype a series of mutations present on the same viral molecule compared to the virus's consensus (or reference) sequence. Therefore, identifying haplotypes can improve the characterization of the viral population at the molecular level. SNPs can be associated with haplotypes if they are located close to each other in the genome so that they can be observed on the same sequencing (paired) reads. However, more distant SNPs are more difficult to associate with the short-read (i.e. 50-300 nt) sequencing technologies predominantly used in virus diagnostics and metagenomics studies. The emergence of new sequencing technologies generating long reads can solve this issue. Those technologies have already generated complete genomes of the viruses at high accuracy for small ssDNA circular genomes (42), but it can remain a challenge for longer viral genomes. To solve this challenge

and exploit the massive amount of data generated by short-read technologies, we propose an innovative methodology to reconstruct haplotypes of distant SNPs from short sequencing reads based on their relative frequencies within and between datasets.

In the present study, we aimed to decipher the presence and the genetic structure of UCBSV populations at the SNP and haplotype levels through a nationwide sampling in the major cassava production areas under diverse pedo-climatic zones and cultivar compositions. This study used high throughput sequencing (HTS) technologies combined with an innovative bioinformatics methodology based on the SNPs identified from the consensus sequences and their frequencies to reconstruct virus haplotypes.

2. Materials and Methods

2.1 Study Area and field sample collection

Cassava is grown on a large scale in the Central and Eastern parts of Rwanda (43). Thus, fields with cassava plants at least 6 months old from 13 cassava-producing districts were inspected for CBSD symptoms and sampled. They included five districts (Ruhango, Nyanza, Muhanga, Kamonyi, and Gisagara) from the Central regions, six districts (Bugesera, Kayonza, Kirehe, Nyagatare, Gatsibo, and Ngoma) from the Eastern regions, one district from Northern province (Gakenke) and one district from the Western province (Nyamasheke). In the latter two provinces, cassava is grown on a small scale. The sampled fields' locations are shown in **S1 Fig**, and their GPS coordinates are available in **S1 Table**.

The districts' climatic conditions differ in temperature, rainfall, and altitude. For example, the eastern province is drier flatlands, the Northern province is cool mountains, whereas the low-

lying valleys of southwestern provinces are warmer (44). These conditions divide the country into different agro-ecological zones, and detailed agro-ecological characteristics of each district surveyed are found in supplementary table 2 (**S1 Fig**) (**S2 Table**).

From each district, ten fields were visited (**S1 Fig**). They were selected from 5 main cassava producer sectors (selected purposely), and 2 cassava fields were selected in each sector, separated by 10 Km between them in order to provide a reliable overview of the ipomovirus diversity in those districts. From each field, leaf samples (symptomatic and asymptomatic) were collected from 10 plants selected randomly using the two diagonals approach across the field (45). Samples from the same field were pooled (one field was considered one sample for RNA extraction), totalling 130 pooled samples corresponding to 130 fields visited from 13 districts.

2.2 Total RNA extraction

Total RNA was extracted using the CTAB method (46). Considering the cost of high throughput sequencing (HTS), RNA samples extracted from the same districts were pooled using equimolar concentration. Consequently, 13 samples of 100 plants (10 fields per district; 10 plants per field) were prepared for HTS. Bioanalyzer (Agilent) and Nanodrop (Thermofisher Scientific) determined the RNA integrity and concentration. Supplementary Table 3 shows the RNA concentration with the 260/280 ratio and the RNA integrity number of the used RNA (**S3 Table**).

2.3 RNA-Seq library preparation and sequencing

Samples were processed at the GIGA facilities of Liège University (Liège, Belgium). Ribosomal depletion was performed by Ribo-Zero® rRNA Removal Kit (Illumina kit) following the

manufacturer's guidelines. RNA Library was prepared following TruSeq Stranded Total RNA Sample Prep LS Protocol (Illumina kit) according to the manufacturer's instructions. The libraries were prepared with UDI (unique dual indexes), and a Free Adapter Removal (Illumina kit) was done on the pooled libraries. The HTS was done on the ILLUMINA NovaSeq 6000 with an S4 flowcell for 2*150 nt. Adapter removal was done with the bcltofastq v2.20 program of Illumina.

2.4 Bioinformatic analysis

2.4.1 Reads processing

First, the obtained raw reads were paired and trimmed using Geneious Prime 2023 (version 10.1.5, Biomatters) software (https://www.geneious.com). The low-quality nucleotides showing quality scores below 20 and reads with lengths lower than 35bp were trimmed using BBDuk V38.37 (47). Then, reads were merged, and duplicated reads were removed using the Dedupe V38.37 (48) plugin implemented in Geneious.

2.4.2 De novo assembly, Mapping, and Phylogenetic analysis

De novo assembly was further performed using the RNA-Spades (49) V3.13.0 assembler implemented in Geneious. The obtained contigs were subjected to a BLAST search (50) (blastn and blastx) against the viral RefSeq database (retrieved in September 2020 - release 201) to check which contigs matched viral sequences in GenBank. All contigs matching a viral reference sequence in GenBank were extracted. All reads were further mapped (Geneious mapper V10.1.5) on the viral contigs of interest with the parameter "Low sensitivity/ Fastest" used (10% mismatch). Furthermore, reads were mapped on the closest reference sequence in the database using the same parameters. The obtained mapped reads were used for SNP calling and in-depth

analysis. SNP calling (Geneious V10.1.5) was performed using the following criteria: (i) a minimum coverage of 100x, (ii) a Minimum Strand-Bias > 65% (p-value ≤ 0.0005), and (iii) a minimum variant frequency of 1%.

To rule out the presence of CBSV in the analyzed samples, all samples were processed with Kraken2 (standard database from 06/2020, Version 2.1.1) on Galaxy (51,52), in addition to the BLAST mentioned above approach, and for positive samples (showing CBSV reads after Kraken results), simultaneous mapping of all reads on the closest CBSV (HG965221) and UCBSV (KX753356.1) reference genomes was performed at maximum 10 % mismatch to avoid non-specific Mapping of UCBSV reads on the CBSV sequence. In addition, to screen our samples for novel Ampelovirus recently reported to infect cassava in Central Africa and the South-West Indian Ocean Islands, all reads were mapped to the Congolese genome sequence of MEaV-1 Ampelovirus (MT773588) (53). MEaV-1 was also tested by RT-PCR and sequencing of the PCR product as described previously (53).

2.4.3 Phylogenetic analysis

The percentage of identity between the consensus sequences of UCBSV and their polyproteins from 12 districts of Rwanda was conducted in MEGA X (V10.2.6) with the Poisson correction model (54).

The newly generated UCBSV whole genome sequences were aligned with the 23 UCBSV genomes from the NCBI GenBank nucleotide database using Clustal Omega V1.2.2 (55)(release 241, November 2020). Phylogenetic analysis and evolutionary divergence between nucleotides sequences were performed using the new genomes identified in the present study as well as complete or partial coat protein sequences from GenBank isolates of UCBSV using Molecular

Evolutionary Genetics Analysis (MEGA-X V10.2.6) under the Neighbor-Joining method (56). The evolutionary distances were computed using the p-distance method (57). All ambiguous positions were removed for each sequence (pairwise deletion option).

2.4.4 Nucleotide diversity

SNPGenie (V1, May 2022) (58) was applied to the SNP tables to calculate the nucleotide diversity (π) for each district (sample). It represented a mean number of pairwise differences per nucleotide position in a population of sequences.

2.4.5 Genetic diversity analysis of UCBSV

A recently established methodology was applied for comparing UCBSV populations at both SNP and haplotype levels (35). Two distance measures were calculated using the consensus sequences or the fixation index (F_{ST}). F_{ST} methodology considers all the SNP detected in the UCBSV populations and their relative frequency (if above 1%) to compare samples. The consensus sequence considers only the dominant base (>50% frequency) at every position. For both methods, an index ranges between 0 and 1 was calculated, where differences between populations increase as the index changes from 0 to 1. The F_{ST} measures were obtained using Popoolation2 version 1.201 (59). First, the reads of each sample were mapped to the closest reference genome (accession number KX753357.1) as described above. Then, the analysis was performed with a single-window size defined by the size of the reference genome (9.097 bp), a step size of 1, a minimum covered fraction of 0.1, minimum coverage of 50, and maximum coverage of 200,000.

The principal component analyses used ade4 and factoextra libraries from the R software version 4.0.3 (http://www.r-project.org/). The dendrograms were calculated using the hclust function implemented in the package stats (version 4.0.2).

2.4.6 SNP and haplotype validation

Sanger sequencing of the RT-PCR products from the 12 pooled samples and 3 individual plants of a tolerant cultivar (NAROCAS1) confirmed the presence of selected SNPs and one haplotype. PCR products were purified using a PCR product purification kit (Qiagen, German). A list of specific primer pairs sequences used for PCR amplification and sequencing is provided in supplementary table 4 (**S4 Table**). In addition, the association between one of the UCBSV haplotypes and a CBSD tolerant cultivar was assessed by checking 12 SNP positions (spanning the whole genome) identified as specific for the selected haplotype.

3. Results

3.1 UCBSV is detected in nearly all generated datasets at a high abundance

The number of high-quality reads generated per sample ranged from 16,565,194 to 27,211,334, averaging 21,869,253 reads. The number of contigs generated by de novo assembly ranged from 24,503 to 49,784 per sequenced sample. In total, 12 genome sequences of UCBSV (ranging from 8,743 to 9,082 nt) with complete CDS and nearly complete UTR regions were generated from 12 samples with a near-complete coverage of the reference (**KX753357.1** - 9,097 nt**)**. No UCBSV

was found for one sample. The average genome fold coverage ranged between 143x to 963x, and the genome coverage was always >99% **(S5 Table)**.

Complementary analyses using Kraken2 (52) identified specific CBSV reads. However, the number of reads did not exceed 25 reads per sample and did not allow the reconstruction of a complete CBSV genome sequence (**S2 Fig; S6 Table**). In the district of Ruhango, a recently discovered ampelovirus was detected in the dataset with 386 reads (**S3 Fig)**, from which two partial ampelovirus contigs of 12,711 bp and 3,390 bp (OL579727; OL579728) were constructed. They shared 98% of their identity with the MEaV-1 Congolese isolate (MT773588). The presence of the ampelovirus was confirmed by RT-PCR (**S4 Fig**) and subsequent Sanger sequencing of the amplicons (NCBI reference numbers: OL579729; and OL57973).

3.2 Phylogenetic analysis of consensus sequences of UCBSV revealed a high homogeneity throughout the sampled districts.

The bioinformatics analysis generated a consensus sequence of the nearly complete UCBSV genome for each of the 12 samples (districts). The phylogenetic analysis of the sequenced samples at nucleotide (nt) and amino acid (aa) levels and publicly available sequences clustered the 12 UCBSV sequences in a single group reduction (**S5 Fig**). The 12 genomes showed very high homology between each other with a maximum of 0.8 % and 0.6 % of divergence at nucleotide (nt) and amino acid (aa) levels, respectively (**S7 Table; S8 Table**). All the genomes had a very high identity (97 %) with a reference sequence of a UCBSV isolate from Tanzania (accession no. KX753357.1) (**S5 Fig).**

As UCBSV diversity was analyzed in 2014 based on the amplification and sequencing of 24 partial CP sequences (210 nt) (20), the partial CP sequences (210 nt) were selected from our 12

genomes. The phylogenetic analysis of the 36 sequences revealed that all the CP sequences from the present study clustered in a single group, with only two of the 24 CP sequences obtained previously. These sequences originated from Bugesera (KX168498) and Nyanza districts (KX168488) **(Fig 1)**.

Fig 1: Phylogenetic analysis. Phylogenetic tree of the 12 Ugandan Cassava Brown Streak Virus partial coat proteins (210 nt) from the present study (blue) in comparison with 24 partial coat proteins previously reported from Rwanda in 2014 (black) (20).

3.3 Validating SNPs by Sanger sequencing

SNPs were identified on 486 positions using all the UCBSV reads (12 samples) aligned on the closest UCBSV reference genome from Tanzania (KX753356.1). Among them, 192 corresponded to non-synonymous mutations and 294 to synonymous or silent mutations. The region with the highest NS mutations was CP, followed by Nib protein with 47/192 and 39/192, respectively (**S9 Table**). The number of polymorphic sites per district ranged from 122 in Kirehe to 225 in Nyanza. Before the in-depth analysis of SNPs, the robustness of the SNP identification was assessed by RT-PCR carried out on the twelve RNA extracts with seven primer pairs. The 84 amplification products were subsequently sequenced to check the nucleotides on 35 mutated positions scattered on the UCBSV genome. In total, 420 SNP positions (among which 91 SNPs corresponded to a mix of two alleles) contributing to the differentiation between UCBSV haplotypes were verified. High-quality sequencing was not achieved for 21 positions (5%). Ninety-seven per cent of the other SNPs (387/399) were confirmed, including mixed alleles for 80 SNPs. For 11 SNPs, only one of the two alleles was observed, mostly the allele with a higher

frequency on the genome alignment (**S10 Table**). This high rate of independent validation by Sanger sequencing confirmed the reliability of the obtained SNPs.

3.4 Differences in nucleotide diversity are observed between districts

Despite a high identity of consensus sequences of isolates between districts, the nucleotide diversity (π) appeared much more variable. The first cluster of districts (Gisagara, Kayonza, Kirehe, Ngoma, Bugesera, and Gatsibo) presented a π from 4.4 10^{-4} to 6.5 10^{-4} and corresponded to the district located in the Eastern part of Rwanda. Another cluster of districts, including Nyanza, Ruhango, Kamonyi, and Nyamasheke, presented a π nucleotide diversity that was ten times higher (4 4.6 10^{-3}). Three of these districts were contiguous and located in the central part of the country, while the fourth district (Nyamasheke) was the only western district surveyed. In addition, two districts (Nyagatare and Muhanga) presented an intermediate π nucleotide diversity with 1.3 10^{-3} and 8.1 10^{-4}, respectively (**S9 Table**).

The variation in π nucleotide diversity across the districts prompted us to conduct an in-depth analysis of SNPs in each sample using an innovative methodology.

3.5 F_{ST} methodology revealed UCBSV genetic diversity beyond the consensus sequences.

To analyze the UCBSV genetic diversity in the sampled districts, the classical measure of distance between samples, based on the consensus sequence generated from each sample, was compared to an innovative approach based on the fixation index (F_{ST}). The F_{ST} calculation considers all the SNPs (>1% frequency) detected in the UCBSV populations from each sample.

Figure 2A and 2B shows the dendrograms obtained by both methods, while supplementary figure 6A and 6B show the matrices (**S6 Fig**). Both methods clustered seven districts together, but *the* F_{ST} method provided better discrimination between identical samples based on their consensus sequences. Nyagatare district was the more distant, although it was identical to Kayonza at the consensus level. This difference aligns with the π 2.5X higher in the Nyagatare district compared to Kayonza. Nyamasheke district was also distinct from all other districts independently of the method used. Both methods clustered together Ruhango, Kamonyi, and Nyanza districts also presented similar π. Muhanga district clustered with these three districts based on the consensus approach but corresponded to a specific cluster by F_{ST} analysis (**Fig 2A and 2B**), reflecting its lower π ($8.14\ 10^{-4}$) compared to the three districts.

Fig 2A and 2B: Analysis of UCBSV diversity. (**A**) **Consensus approach**: Dendrogram built from pairwise distance matrices obtained after multiple alignments of the virus consensus sequences. (**B**) $\boldsymbol{F_{ST}}$ **approach:** Dendrogram built from pairwise F_{ST} matrices obtained using the entire set of SNPs detected in the virus populations.

Overall, the discrimination between samples was improved by considering all the SNPs with a frequency above 1% for at least one sample. So, the next step of our analysis aimed to identify the SNPs discriminating the samples from each other.

3.6 Principal Component Analysis of SNPs confirmed the clustering by *the* F_{ST} approach

The frequencies of the 486 SNPs detected in the samples from the 12 districts were used as variables to perform a Principal Component Analysis (PCA). The SNPs used to create each dimension and their frequencies are shown in supplementary table 11 (**S11 Table**).

The first, second, and third dimensions explained 22.8%, 15.5%, and 12.7% of the total variation, respectively (**Fig 3A and 3B**). Most of the samples were clustered into two groups by dimension 1: (i) Nyanza, Ruhango, Kamonyi, and Muhanga, and (ii) Bugesera, Kirehe, Gatsibo, Kayonza, Nyamasheke, Gisagara, and Ngoma districts. Dimensions 1 and 2 discriminated Nyagatare district from the other districts. Dimension 3 separated the Nyamasheke district from the others.

Figures 3A and 3B: Principal component analysis (PCA) of the virus populations. The PCA shows the first, second, and third dimensions obtained using all the detected SNPs as variables for Ugandan cassava brown streak virus sequences from the 12 districts. **(A) PCA-Dimension 1 versus 2**. **(B) PCA-Dimension 1 versus 3.**

3.7 The identified UCBSV haplotypes are geographically clustered

To obtain further insight into the SNPs distribution, the frequencies of the 92 SNPs contributing the most to the first, second, or third dimension in each sample were extracted and compared to the previously obtained F_{ST} dendrograms (**Fig 4**).

Fig 4. SNPs contributing to the differentiation between Ugandan cassava brown streak virus populations Dendrograms have been constructed from pairwise F_{ST} matrices obtained using the entire set of SNPs detected in the UCBSV populations. The frequencies of the 92 SNPs

contributing to each sample's first, second, and third dimensions were extracted and compared to the F_{ST} dendrograms. The darker the blue, the higher the SNP frequency.

The SNP frequencies showed a pattern explaining the dendrogram. Several SNPs were always present at a very similar frequency for each sample and formed a cluster. Three major clusters were identified according to the following criteria: presenting at least 10 SNPs with a frequency higher than 10% in at least one sample and with a frequency varying similarly between the samples. Importantly, each cluster included SNPs present along the genome, sometimes at distant locations. For example, the first cluster included the SNPs T-1103-G and T-7529-C (SNPs being named according to their position on the genome and reference and alternative alleles) located 6,000 nt apart. Despite these distances, the frequency of the SNPs varied homogenously between the samples of a cluster, suggesting that they are linked and could constitute a haplotype. Our dataset highlighted 3 major haplotypes of UCBSV across Rwanda. Haplotype 1 (H_1) was defined by 30 SNPs at high frequency in clusters 1 and absent or low frequency for clusters 2 and 3. Among these SNPs, eight were non-synonymous and located mainly in the sequences of the P1 protein (3 SNPs) and the coat protein (2 SNPs) (**S9 Table**). The frequency of this haplotype was close to 100% in Nyagatare, Gisagara, Kayonza, Kirehe, Ngoma, Bugesera, and Gatsibo districts (frequency of specific SNPs ranging between 86% and 100%). Its frequency was close to 30% in Ruhango, Kamonyi, Nyanza, and Nyamasheke districts, while it was absent in the Muhanga district. The presence of H1 in seven samples was consistent with the results of PCA, where dimension one divided samples into two groups, one of the groups being composed of districts showing only H_1 (Bugesera, Kirehe, Gatsibo, Kayonza, Gisagara, and Ngoma). Nyagatare district was slightly different from the other districts, with some minor SNPs in this virus population. This explains why PCA dimension two isolated the

Nyagatare district from the others (**Fig 3A**). The 28 SNPs of cluster 2 defined haplotype 2 (H2), and their frequency was 100% in the Muhanga district. Among these SNPs, six were non-synonymous, and two were in the coat protein (2) (**S9 Table**). It did not have any of the 30 SNPs similar to H_1, which could explain why this sample did not belong to any group in the F_{ST} dendrogram. H_1 and H_2 are mixed with frequencies around 30 and 70 %, respectively, in Ruhango, Kamonyi and Nyanza districts. Haplotype 3 (H_3), characterized by 28 SNPs, was only present in the Nyamasheke district, mixed with H_1. H_3 seemed to share the SNPs C-3155-T, G-6355-A, T-8225-G, G-8636-A, with H_1 as their frequency was 100%. H_3 explained the PCA results where dimension 3 separated Nyamasheke district from the rest of the districts (**Fig 3B**). Compared to the two other haplotypes, H_3 presented seven non-synonymous SNPs located in the P1 protein (2) and the coat protein (2) (**S9 Table**). Importantly, haplotype 1 was widely spread in the Eastern, Central-Southern and Western regions. In contrast, the haplotype 2 distribution was limited to the Central regions and haplotype 3 in the West of the country (**Fig 5**). When the 3 haplotypes were compared, no shared non-synonymous (NS) mutations existed. H_1, H_2 and H_3 had 8, 6 and 7 unique NS mutations, respectively. For synonymous (S) mutations, H_1 and H_2 had 22 unique S mutations each, while they had only 3 S mutations in common. H_3 had 21 unique S mutations. There were 4 S mutations common between H_1 and H_3 and 2 S mutations common between H_2 and H_3 (**S9 Table**).

Fig 5: Distribution of UCBSV haplotypes in Rwanda. Country-wide distribution of Ugandan cassava brown streak virus Haplotypes identified was presented on the Rwandan map. A pie chart was used; single colour represents the presence of a single haplotype (100%). Two colours represent the presence of two different haplotypes.

3.8 Haplotype 2 distribution appears to be associated with the presence of the CBSD-tolerant cultivar NAROCAS 1

An exciting association was observed between the haplotype H_2 and the NAROCAS 1 cultivar. The NAROCAS 1 cultivar is only present in the central regions (Kamonyi, Muhanga, Ruhango, and Nyanza districts at a frequency of around 25%). The same regions are also the only regions where H_2 was found (**S2 Table**). The association between H2 and NAROCAS1 was further verified by sequencing the RT-PCR amplicons from 3 individual NAROCAS 1 samples using four primer pairs matching the UCBSV genome sequence. In total, 12 SNPs position spanning the whole genome that discriminates UCBSV haplotype H_2 from others were verified and confirmed the presence of H_2 in the three NAROCAS1 samples **(S12 Table).**

4. Discussion

The present study computed SNP frequencies by combining F_{ST} analysis and PCA to study the UCBSV genome diversity and reconstruct haplotypes. It allowed the discrimination of very close virus isolates (>99%) and the characterization of 3 haplotypes whose distribution was clustered: H2 was associated with the presence of tolerant cultivar NAROCAS 1 in the central region, H3 with the Eastern region (higher altitude and a specific cultivar, Mushedule was particularly abundant) while H1 was found all over the country but at different frequencies. Two additional viruses were also detected in the 130 samples, namely CBSV and MEaV-1.

This methodology relies on high throughput sequencing of pooled samples and a combination of viral sequences analysis tools. First, the pooled plant samples were sampled using a balanced and

systematic sampling scheme. The balanced and constant sampling used in the present study was instrumental in comparing the virus diversity between districts. Recent studies have proved that pooling is a cost-effective approach as a pooling of up to 50 samples has enabled comprehensive virome analysis of several virus species on a larger geographical scale in potatoes (60), *Poaceae* (61), flies or bees (62,63). Moreover, a recent study on pea viruses reported that while pooling 120 leaves into a single bulk field sample (BFS), viruses present at a low incidence were still detectable by HTS. Three of the BFS were re-tested in-depth by HTS, and no additional plant viruses were identified (64). Many studies have proven that analyzing minor variants at low frequency is essential to understand the virus diversity, its evolution and its interactions with the host (65,66) with examples on individual or pooled samples (67), with the zucchini yellow mosaic virus (33), or the Chlorovirus (68). Some studies above selected SNPs with a relative frequency above 1%, as we did in this study (30,33,65). In addition, the obtained genomes were very well covered as they ranged from 142.9 to 963.5 for the analyzed samples.

In several studies, the conclusions remained constrained by the lack of independent validation (69,70). In the present study, we independently confirmed the detection of a subset of interesting SNPs spanning the genome to reinforce the reliability of generated data—a confirmation rate higher than 97% was observed. The unconfirmed mutations occurred mainly in position with a mix of two bases, one minor at low frequency (<30%) that was not observed, probably due to the confirmed lower sensitivity of Sanger sequencing for detecting SNPs at low frequency (71).

Our methodology was applied to 13 pools of plants representing each district and spanning different agro-ecological zones in Rwanda (**S1 Fig**). Overall, it revealed a low divergence of viral consensus sequences between districts. In addition, partial CP sequences (210 nt) extracted

from the 12 consensus sequences presented 100% identity with two sequences obtained in 2014, suggesting a slow evolution of this short region from the UCBSV genome.

A reduction in UCSBV diversity was observed in the current dataset compared to the reported UCBSV diversity in 2014 and should be discussed technically and scientifically. First, the present study used HTS technologies that have improved inclusivity compared to targeted RT-PCR as demonstrated by many publications (38,72,73), so they could theoretically detect a broader range of isolates. Regarding the representativeness of the obtained sequences, the partial CP sequencing was carried out in 2014 using a limited number of plants per district, from one in the Gatsibo district to thirteen in the Nyanza district, thus most probably capturing the most abundant isolates. In contrast, the HTS protocol in the current study was applied to pools of 100 plants per district. Even if rare isolate present in one or a few plants might have been missed out by the sample pooling, Field-Based Sequencing (FBS) applied on pools of 120 plants following a similar protocol demonstrated the ability of HTS technologies to detect the viruses present in such pools (64) reliably. Our bioinformatic methodology included SNPs with a frequency higher than 1%. Therefore, it theoretically identified other major isolates as the ones detected in 2014. One hypothesis on the observed difference in UCBSV diversity could originate from the strong shift in cassava cultivars experienced in Rwanda following the severe CBSD crisis. The present study used a majority (approximately 60%) of plant samples from the two CBSD tolerant cultivars introduced in 2015 (NASE 14 and NAROCAS 1), while, in 2014, the survey was carried out on local susceptible cultivars. The deployment of these imported cultivars has relied on official distribution to farmers and informal exchanges between farmers within and between districts (16).

Beyond the very high genetic homogeneity of consensus sequences observed throughout the country, the genetic diversity within and between samples was analyzed through a combination of bioinformatic tools, including the nucleotide diversity (π), the calculation of the F_{ST}, and principal component analysis on SNPs to highlight differences between closely related genomes and to identify haplotypes representing molecules actually infecting the plants. This methodology of haplotype reconstruction on SNPs spanning the entire genome is complementary to the current methodologies based on SNPs shared on sequencing reads that often reconstruct haplotypes spanning partial genome sequences, as recently observed for the *Lolium latent virus* detected in pooled samples (61).

The SNPs discriminating the three haplotypes were mainly synonymous, although each haplotype contained respectively 8, 6, and 7 unique and non-synonymous SNPs. Such observation deserves further investigation on its consequences as they are located in P1 and CP genes important in plant-virus interactions (74,75).

The H_3 was found only in the Western region of the country with higher altitude, lower temperature, higher rainfall and the high frequency of the Mushedule cultivar.On the other hand, the H_1 was widely spread with a 100% frequency in the Eastern region, which borders Uganda and Tanzania and is characterized by a lower altitude, less rain, a higher temperature and a high frequency of NASE14 cultivar. In contrast, the H_2 was restricted to the central regions and was found to be associated with the NAROCAS1 cultivar. However, this confirmatory work could also be carried out for H_3 and H_1 to clarify whether the haplotypes are linked with cassava cultivars or geographical occurrences. The plants analyzed in this study were sampled in the frame of a broad survey including 130 fields around the country. The disease incidence varied strongly (12) while the disease severity on symptomatic plants was more constant across regions.

For example, the CBSD mean severity scores were 3+/-0.6, 2.9+/-0.8 and 2+/-0.2 in Eastern, Central, and the Western part respectively (12). Noteworthy, in our context of field-based sampling, determining the association of disease severity with specific cultivars and haplotypes will require a new cultivar-based survey combined with greenhouse inoculation assays as recommended for evaluating robustly causal association (76).

In the future, the reported SNPs could serve as markers to investigate and decipher the factors impacting UCBSV genetic evolution in Rwanda and the geographical distribution of the 3 haplotypes. Incorporating testing the presence of these haplotypes on planting material in the regular viral testing should be a priority to ensure the distribution of healthy planting materials, which is an essential control measure in CBSD management. Furthermore, future research activities should investigate the impact of cassava varieties distribution on ipomoviruses diversity and distribution in Africa as well as the association between the identified UCBSD haplotypes and the CBSD severity symptoms on individual cassava genotypes, including the currently widely distributed cultivars NASE14 and NAROCAS1.

Overall, our results provided evidence that a much more complex picture of genetic diversity can be deciphered beyond the consensus sequences with practical implications on virus evolution and its management. Our methodology proposed a high-resolution analysis of genome diversity between and within samples. It can be used at various scales, from individual plants to plants pooled by geographical origin (from field to region) or any other factor (cultivar, phenotype).

Data availability

The datasets of genome sequences generated and analyzed during this study are available in the GenBank repository under the following accession numbers: OK423771; OK423772;

OK423773; OK423774; OK423775; OK423776; OK423777; OK423778; OK423779; OK423780; OK423781; OK423782; OL579727; OL579728; OL579729; and OL57973. In addition, raw data were deposited in SRA (PRJNA768633) and can be found at https://www.ncbi.nlm.nih.gov/sra/PRJNA768633.

Acknowledgement

Rwanda cassava farmers are acknowledged for allowing sampling in their fields.

Author Contributions

Conception and study design: Sebastien Massart, Hervé Vanderschuren, Chantal Nyirakanani, and Jean Pierre Bizimana; *methodology*: Sebastien Massart, Hervé Vanderschuren, Lucie Tamisier, and Chantal Nyirakanani; *validation*: Sebastien Massart and Hervé Vanderschuren; *Data analysis:* Chantal Nyirakanani, Lucie Tamisier, Jean-Pierre Bizimana, Johan Rollin, Sebastien Massart; *Confirmatory Sanger sequencing analysis*: Chantal Nyirakanani, and Ilhem Selmi; *writing the first original draft*: Chantal Nyirakanani; *review and editing*: Lucie Tamisier, Johan Rollin, Ludivine Laois, Athanase Nduwumuremyi, Vincent de Paul Bigirimana, Sebastien Massart, Hervé Vanderschuren. *Supervision:* Hervé Vanderschuren, Sebastien Massart, Ludivine Lassois; *funding:* Hervé Vanderschuren. All authors revised and approved the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Funding

The current study was carried out as part of the ARES (Académie de Recherche et d'Enseignement Supérieur, Fédération Wallonie-Bruxelles, Belgium)-funded project iCARE (Improved Cassava Virus Resistance mitigation strategies and development of a disease-free seed system).

References

1. United Nations Rwanda. Restoring cassava farming in Rwanda [Internet]. 2019. Available from: https://rwanda.un.org/en/22965-restoring-cassava-farming-rwanda
2. Mbewe W, Hanley-Bowdoin L, Ndunguru J, Duffy S. CHAPTER 7: Cassava Viruses: Epidemiology, Evolution, and Management. In 2020. p. 133–57.
3. Otekunrin OA, Sawicka B. Cassava , a 21st Century Staple Crop : How can Nigeria Harness Its Enormous Trade Potentials ? Acta Sci Agric. 2019;3(8).
4. FAO. Food Outlook- Biannual report on global food markets - November 2018 [Internet]. Global information and early warning system on food and agriculture. 2018. 104 p. Available from: http://www.fao.org/docrep/013/al969e/al969e00.pdf
5. MINAGRI. Annual report 2020-2021: Ministry of agriculture and animal resources. 2021.
6. National Institute of Statistics of Rwanda. Seasonal Agricultural Survey. 2021.
7. Tomlinson KR, Bailey AM, Alicai T, Seal S, Foster GD. Cassava brown streak disease: historical timeline, current knowledge and future prospects. Mol Plant Pathol. 2018;19(5):1282–94.
8. Storey HH, Nichols RFW. Virus Diseases of East African Plants. East African Agric J. 1938;3(6):446–9.
9. Legg JP, Jeremiah SC, Obiero HM, Maruthi MN, Ndyetabula I, Okao-Okuja G, et al. Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. Virus Res [Internet]. 2011;159(2):161–70. Available from: http://dx.doi.org/10.1016/j.virusres.2011.04.018

10. Mbanzibwa DR, Tian YP, Tugume AK, Patil BL, Yadav JS, Bagewadi B, et al. Evolution of cassava brown streak disease-associated viruses. J Gen Virol. 2011;92(4):974–87.

11. Ndunguru J, Sseruwagi P, Tairo F, Stomeo F, Maina S, Djinkeng A, et al. Analyses of twelve new whole genome sequences of cassava brown streak viruses and ugandan cassava brown streak viruses from East Africa: Diversity, supercomputing and evidence for further speciation. PLoS One. 2015;10(10):1–18.

12. Nyirakanani C, Bizimana JP, Kwibuka Y, Nduwumuremyi A, Bigirimana V de P, Bucagu C, et al. Farmer and Field Survey in Cassava-Growing Districts of Rwanda Reveals Key Factors Associated With Cassava Brown Streak Disease Incidence and Cassava Productivity. Front Sustain Food Syst. 2021;5(December):1–14.

13. Mulenga RM, Makulu M, Laura M. Cassava Brown Streak Disease and Ugandan cassava brown streak virus Reported for the First Time in Zambia. Plant Dis. 2018;102(7):1410–8.

14. Kwibuka Y, Nyirakanani C, Bizimana JP, Bisimwa E, Brostaux Y, Massart S. Risk factors associated with cassava brown streak disease dissemination through seed pathways in Eastern. Front Plant Sci. 2022;(July):1–18.

15. Maruthi MN, Jeremiah SC, Mohammed IU, Legg JP. The role of the whitefly, Bemisia tabaci (Gennadius), and farmer practices in the spread of cassava brown streak ipomoviruses. J Phytopathol. 2017;165(11–12):707–17.

16. Douthwaite B. Development of a cassava seed certification system in Rwanda: Evaluation of CGIAR contributions to a policy outcome trajectory. 2020.

17. Duffy S. Why are RNA virus mutation rates so damn high? PLoS Biol. 2018;16(8):1–6.

18. Roux-Cuvelier M, Teyssedre D, Chesneau T, Jeffray C, Massé D, Jade K, et al. First report of cassava brown streak disease and associated Ugandan cassava brown streak virus in Mayotte Island . New Dis Reports. 2014;30(1):28–28.

19. Kathurima TM, Ateka EM. Diversity and Phylogenetic Relationships of Full Genome Sequences of Cassava Brown Streak Viruses in Kenya. Biotechnol J Int. 2019;23(3):1–11.

20. Munganyinka E, Ateka EM, Kihurani AW, Kanyange MC, Tairo F. Cassava brown streak

disease in Rwanda , the associated viruses and disease phenotypes. Plant Pathol. 2017;67:377–87.

21. Mbanzibwa DR, Tian YP, Tugume AK, Mukasa SB, Tairo F, Kyamanywa S, et al. Simultaneous virus-specific detection of the two cassava brown streak-associated viruses by RT-PCR reveals wide distribution in East Africa , mixed infections , and infections in Manihot glaziovii. J Virol Methods [Internet]. 2011;171(2):394–400. Available from: http://dx.doi.org/10.1016/j.jviromet.2010.09.024

22. Lane AA, Odejide O, Kopp N, Kim S, Yoda A, Erlich R, et al. Low frequency clonal mutations recoverable by deep sequencing in patients with aplastic anemia. Leukemia. 2013;27(4):968–71.

23. Fuentes S, Gibbs AJ, Adams IP, Hajizadeh M, Kreuze J, Fox A, et al. Phylogenetics and Evolution of Potato Virus V: Another Potyvirus that Originated in the Andes. Plant Dis. 2022;106(2):691–700.

24. Kawakubo S, Gao F, Li S, Tan Z, Huang YK, Adkar-Purushothama CR, et al. Genomic analysis of the brassica pathogen turnip mosaic potyvirus reveals its spread along the former trade routes of the Silk Road. Proc Natl Acad Sci U S A. 2021;118(12):1–10.

25. Alicai T, Ndunguru J, Sseruwagi P, Tairo F, Okao-okuja G, Nanvubya R, et al. a rapidly evolving genome : implications for virus speciation , variability , diagnosis and host resistance. Nat Publ Gr. 2016;(October):1–14.

26. Mulimbi W, Phemba X, Assumani B, Kasereka P, Muyisa S, Ugentho H, et al. First report of Ugandan cassava brown streak virus on cassava in Democratic Republic of Congo. 2012;5197.

27. Mbewe W, Winter S, Mukasa S, Tairo F. Deep Sequencing Reveals a Divergent Ugandan cassava brown streak virus isolate from Malawi. Viruses. 2017;43:8–9.

28. Bigirimana S, Barumbanze P, Ndayihanzamaso P, Shirima R, Legg JP. First report of cassava brown streak disease and associated Ugandan cassava brown streak virus in Burundi. 2011;5197.

29. Adams IP, Harju V., Hodges T, Hany U, Skelton A, Rai S, et al. First report of maize

lethal necrosis disease in Rwanda. New Dis Reports. 2014;29(1):22–22.

30. da Silva W, Kutnjak D, Xu Y, Xu Y, Giovannoni J, Elena SF, et al. Transmission modes affect the population structure of potato virus Y in potato. PLoS Pathog [Internet]. 2020;16(6 June):1–23. Available from: http://dx.doi.org/10.1371/journal.ppat.1008608

31. Massart S, Olmos A, Jijakli H, Candresse T. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res [Internet]. 2014;188:90–6. Available from: http://dx.doi.org/10.1016/j.virusres.2014.03.029

32. Cuevas JM, Willemsen A, Hillung J, Zwart MP, Elena SF. Temporal dynamics of intrahost molecular evolution for a plant RNA virus. Mol Biol Evol. 2015;32(5):1132–47.

33. Simmons HE, Dunham JP, Stack JC, Dickins BJA, Pagán I, Holmes EC, et al. Deep sequencing reveals persistence of intra- and inter-host genetic diversity in natural and greenhouse populations of zucchini yellow mosaic virus. J Gen Virol. 2012;93(8):1831–40.

34. Bordería A V., Isakov O, Moratorio G, Henningsson R, Agüera-González S, Organtini L, et al. Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. PLoS Pathog [Internet]. 2015;11(5):1–20. Available from: http://dx.doi.org/10.1371/journal.ppat.1004838

35. Tamisier L, Colson C, Maclot F, Zhang P, Wang X. The tree that hides the forest : going beyond the consensus to analyse within-plant virus population diversity (in preparation). (4):1–23.

36. Winter S, Koerbler M, Stein B, Pietruszka A, Paape M, Butgereitt A. Analysis of cassava brown streak viruses reveals the presence of distinct virus species causing cassava brown streak disease in East Africa. J Gen Virol. 2010;91(2010):1365–72.

37. Kim NK, Lee HJ, Kim SM, Jeong RD. Identification of Viruses Infecting Oats in Korea by Metatranscriptomics. Plants. 2022;11(3).

38. Hanafi M, Rong W, Tamisier L, Berhal C, Roux N, Massart S. Detection of Banana Mild Mosaic Virus in Musa In Vitro Plants: High-Throughput Sequencing Presents Higher Diagnostic Sensitivity Than (IC)-RT-PCR and Identifies a New Betaflexiviridae Species.

Plants. 2022;11(2).

39. Weir B, Clark Cockerham C. Estimating F-Statistics for the Analysis of Population Structure. Evolution (N Y). 1984;38(6):1358–70.

40. Wijayasekara D, Ali A. Evolutionary study of maize dwarf mosaic virus using nearly complete genome sequences acquired by next-generation sequencing. Sci Rep [Internet]. 2021;11(1):1–14. Available from: https://doi.org/10.1038/s41598-021-98299-9

41. Hill T, Unckless RL. Recurrent evolution of high virulence in isolated populations of a DNA virus. Elife. 2020;9:1–53.

42. Mehta D, Stürchler A, Anjanappa RB, Zaidi SSEA, Hirsch-Hoffmann M, Gruissem W, et al. Linking CRISPR-Cas9 interference in cassava to the evolution of editing-resistant geminiviruses. Genome Biol. 2019;20(1):1–10.

43. NISR. Seasonal Agricultural Survey (SAS) ANNUAL REPORT 2019. 2019.

44. Nduwumuremyi A, Kabirigi M. Yield Gap Analysis of Key Agricultural Commodities in Rwanda. 2018.

45. Rwegasira G, Rey M, Nawabu H. Approaches to Diagnosis and Detection of Cassava Brown Streak Virus (Potiviridae, Ipomovirus) In Field-Grown Cassava Crop. African J Food,Agriculture, Nutr Dev. 2011;11(3):4739–56.

46. Ling Z, Zhike Z, Shunquan L, Tingting Z, Xianghui Y. Evaluation of Six Methods for Extraction of Total RNA from Loquat. Not Bot Horti Agrobot Cluj-Napoca. 2013;41(1):313–6.

47. Kechin A, Boyarskikh U, Kel A, Filipenko M. CutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. J Comput Biol. 2017;24(11):1138–43.

48. Bushnell B, Rood J, Singer E. BBMerge – Accurate paired shotgun read merging via overlap. PLoS One. 2017;12(10):1–15.

49. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. Gigascience. 2019;8(9):1–

13.

50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.

51. Pond SK, Wadhawan S, Chiaromonte F, Ananda G, Chung WY, Taylor J, et al. Windshield splatter analysis with the Galaxy metagenomic pipeline. Genome Res. 2009;19(11):2144–53.

52. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;

53. Kwibuka Y, Bisimwa E, Blouin AG, Bragard C, Candresse T, Faure C, et al. Novel ampeloviruses infecting cassava in central africa and the south-west indian ocean islands. Viruses. 2021;13(6):1–17.

54. Zuckerkandl E, Pauling L. Molecules as documents of history. J Theor Biol. 1965;8(2):357–66.

55. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(539).

56. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.

57. Nei M, Kumar S. Molecular Evolution and Phylogenetics. Vol. 31, Oxford University Press. New York; 2000. 1029–1029 p.

58. Nelson CW, Moncla LH, Hughes AL. SNPGenie: Estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. Bioinformatics. 2015;31(22):3709–11.

59. Kofler R, Pandey RV, Schlötterer C. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics. 2011;27(24):3435–6.

60. Schumpp O, Dupuis B, Bréchon A, Wild W, Frei P, Pellet D, et al. Diagnostic moléculaire

à haut débit pour détecter les viroses des plants de pomme de terre. Rech Agron Suisse. 2016;7(10):456–65.

61. Maclot F, Candresse T, Filloux D, Malmstrom CM, Roumagnac P, van der Vlugt R, et al. Illuminating an Ecological Blackbox: Using High Throughput Sequencing to Characterize the Plant Virome Across Scales. Front Microbiol. 2020;11(October):1–16.

62. Wallace MA, Coffman KA, Gilbert C, Ravindran S, Albery GF, Abbott J, et al. The discovery, distribution, and diversity of DNA viruses associated with Drosophila melanogaster in Europe. Virus Evol. 2021;7(1):1–23.

63. Roberts JMK, Ireland KB, Tay WT, Paini D. Honey bee-assisted surveillance for early plant virus detection. Ann Appl Biol. 2018;173(3):285–93.

64. Fowkes AR, McGreig S, Pufal H, Duffy S, Howard B, Adams IP, et al. Integrating high throughput sequencing into survey design reveals turnip yellows virus and soybean dwarf virus in pea (Pisum sativum) in the united kingdom. Viruses. 2021;13(12).

65. Clerc S Le, Coulonges C, Delaneau O, Manen D Van, Herbeck T, Limou S, et al. SCREENING LOW FREQUENCY SNPS FROM GENOME WIDE ASSOCIATION STUDY REVEALS A NEW RISK ALLELE FOR. J Acquir Immune Defic Syndr. 2011;56(3):279–84.

66. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. Elife. 2015;4(DECEMBER2015):1–26.

67. Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC. Validation of SNP Allele Frequencies Determined by Pooled Next-Generation Sequencing in Natural Populations of a Non-Model Plant Species. PLoS One. 2013;8(11).

68. Retel C, Kowallik V, Becks L, Feulner PGD. Strong selection and high mutation supply characterize experimental Chlorovirus evolution. Virus Evol. 2022;8(1):1–14.

69. McCann HC, Rikkerink EHA, Bertels F, Fiers M, Lu A, Rees-George J, et al. Genomic Analysis of the Kiwifruit Pathogen Pseudomonas syringae pv. actinidiae Provides Insight into the Origins of an Emergent Plant Disease. PLoS Pathog. 2013;9(7).

70. Aimone CD, Lavington E, Hoyer JS, Deppong DO, Mickelson-Young L, Jacobson A, et al. Population diversity of cassava mosaic begomoviruses increases over the course of serial vegetative propagation. J Gen Virol. 2021;102(7).

71. Huang T. Next generation sequencing to characterize mitochondrial genomic DNA heteroplasmy. Curr Protoc Hum Genet. 2011;(SUPPL. 71):1–12.

72. Vu A, Stankovi I. Detection of Four New Tomato Viruses in Serbia Using Post Hoc High-Throughput Sequencing Analysis of Samples From a Large-Scale Field Survey. Plant Dis. 2021;(September):2325–32.

73. Villamor DEV, Keller KE, Martin RR, Tzanetakis IE. Comparison of High Throughput Sequencing to Standard Protocols for Virus Detection in Berry Crops. Plant Dis. 2022;106(2):518–25.

74. Pablo-rodríguez JL, Bailey AM, Foster GD. Characterization of Cassava brown streak virus proteins draw their sides during mixed infections and reveal P1 as a silencing suppressor. Res Sq. 2022;1–22.

75. Ivanov KI, Mäkinen K. Coat proteins, host factors and plant viral replication. Curr Opin Virol. 2012;2(6):712–8.

76. Fox A. Reconsidering causal association in plant virology. Plant Pathol. 2020;69(6):956–61.

Supplementary materials

S1 Fig. Study Area. Map of Rwanda showing surveyed districts and their agro-ecological zones. The dots represent the location of assessed cassava fields in 2019.

S2 Fig. Mapping of Cassava brown streak virus reads to the closest reference. The CBSV reads from the 4 districts were mixed and mapped to the closest CBSV (HG965221).

S3 Fig. Mapping of all reads to the Congolese genome sequence of MEaV-1 Ampelovirus (MT773588). MEaV-1 Ampelovirus is present in one district of Rwanda called Ruhango with 98% pairwise identity to MT773588, with 386 reads mapping 87% of the genome.

S4 Fig. Confirmation of Ampelovirus. The gel image depicts RT-PCR of the Ampelovirus confirmation using primer pairs signed on coat protein that yield amplicons of 203 bp and RNA-dependent RNA polymerase (RDRP), which yield amplicons of 261 bp. L: Fast DNA Ladder.

S5 Fig. Phylogenetic analysis. Maximum likelihood phylogenetic tree (1000 bootstraps) generated from 12 Ugandan cassava brown streak virus (UCBSV) full or nearly complete sequences produced in this study compared with 23 UCBSV from NCBI. The analysis grouped the 12 sequences into 1 group. All the genomes showed a pairwise identity of 97% to the Tanzania isolate.

S6 Fig. Pairwise identity matrices of the Ugandan cassava brown streak virus (UCBSV) haplotypes. Pairwise identity matrices were obtained from multiple alignments of the UCBSV consensus sequences and from pairwise F_{ST} matrices obtained using all the SNPs detected among the UCBSV populations. The values obtained with both methods ranged between 0 and 1; 0 means that two populations are genetically identical, and 1 means that two populations are completely divergent.

S1 Table. GPS coordinates of sampled cassava fields.

S2 Table. Characteristics of Agro-ecological Zones (AEZs) of the districts surveyed as well as common cassava cultivars and intensity of cassava cultivation per districts.

S3 Table. RNA concentration, 260/280 Ratio and RIN of the used samples.

S4 Table. RT-PCR primer pairs used for the confirmation of the essential SNPs in the samples.

S5 Table. High Throughput Sequencing Data of Ugandan Cassava Brown Streak Virus in Rwanda. Nearly complete Ugandan cassava brown streak virus was recovered in 12 per 13 districts.

S6 Table. Cassava Brown Streak Virus (CBSV) analysis among 13 districts. The CBSV analysis by three approaches (Blast, Mapping, and Kraken) revealed that CBSV was detected in 4 districts: Ruhango, Muhanga, Bugesera, and Gatsibo.

S7 Table. Percentages of identity at nucleotide level between UCBSV consensus sequences from 12 districts of Rwanda

S8 Table: Percentages of identity between the amino acid sequence of the UCBSV consensus polyprotein from 12 districts of Rwanda.

S9 Table. A complete list of mutations contributed to the discrimination of Ugandan Cassava Brown Streak Virus populations. Most mutations were silent, but few resulted in amino acid change.

S10 Table. A list of mutations confirmed by Sanger sequencing. A representative number of mutations that contributed to the discrimination of the Ugandan Cassava Brown Streak Virus population were validated by Sanger sequencing.

S11 Table. SNPs used to create dimensions. A complete list of all 564 SNPs used to create five dimensions is shown.

S12 Table. Confirmation of UCBSV haplotype 2 based on Sanger sequencing of individual samples.

Fig 1

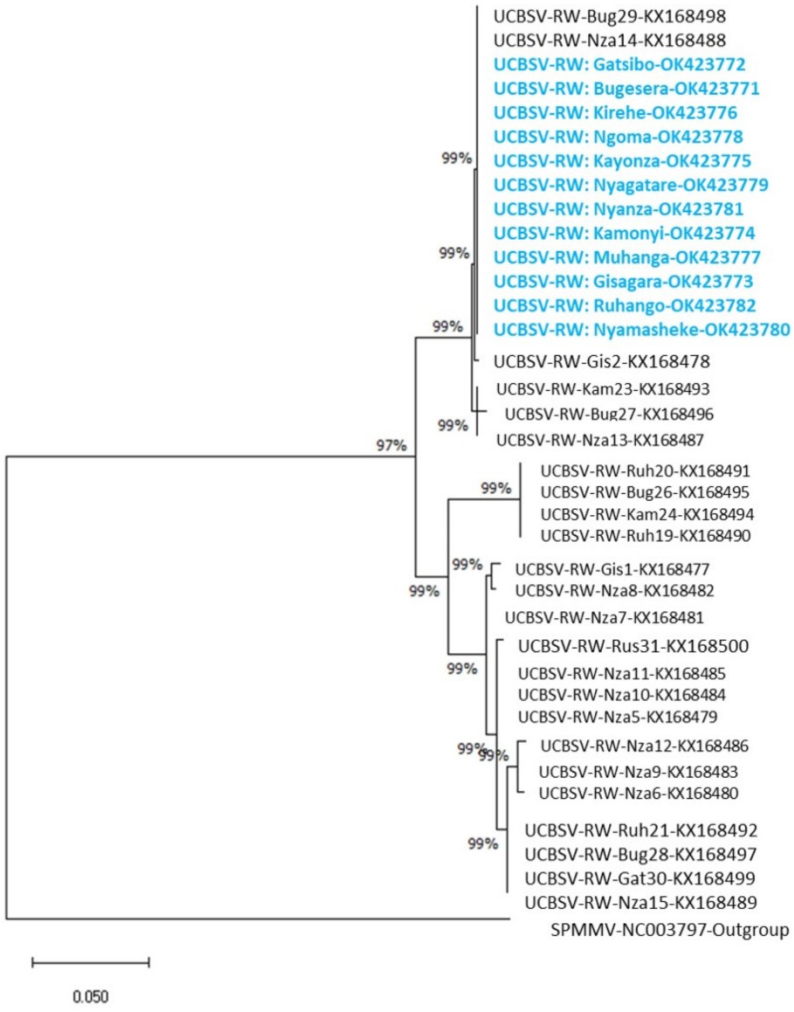

Fig 2A and 2B

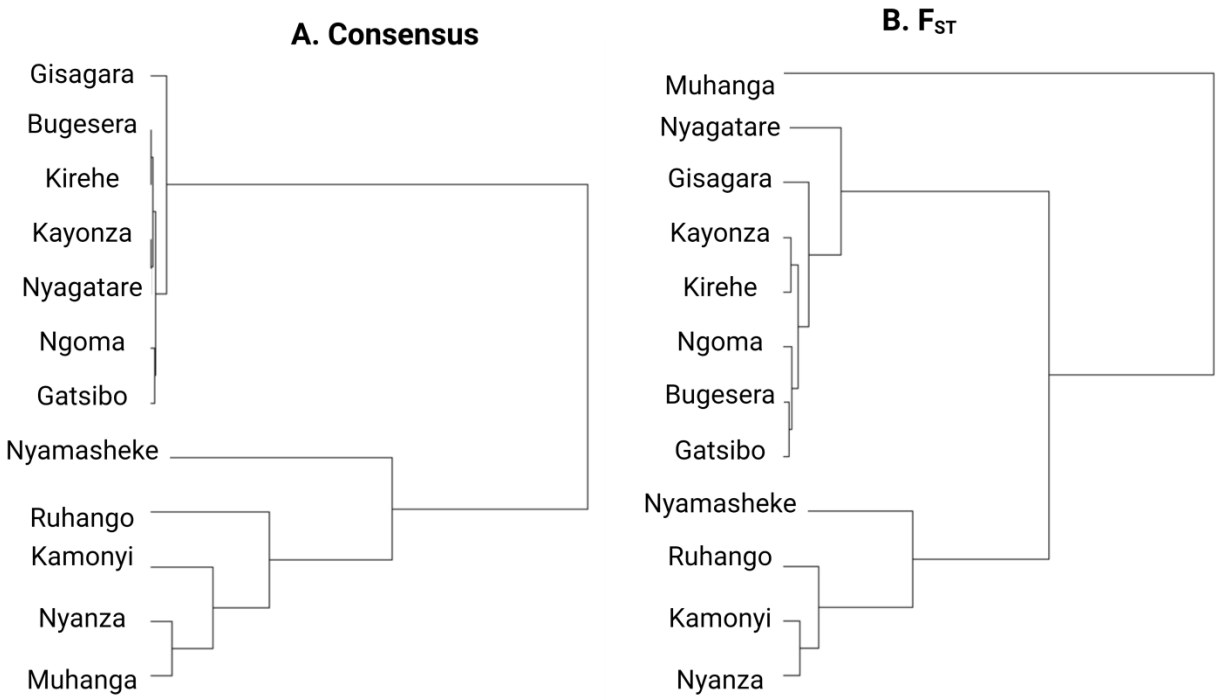

Fig 3A and 3B

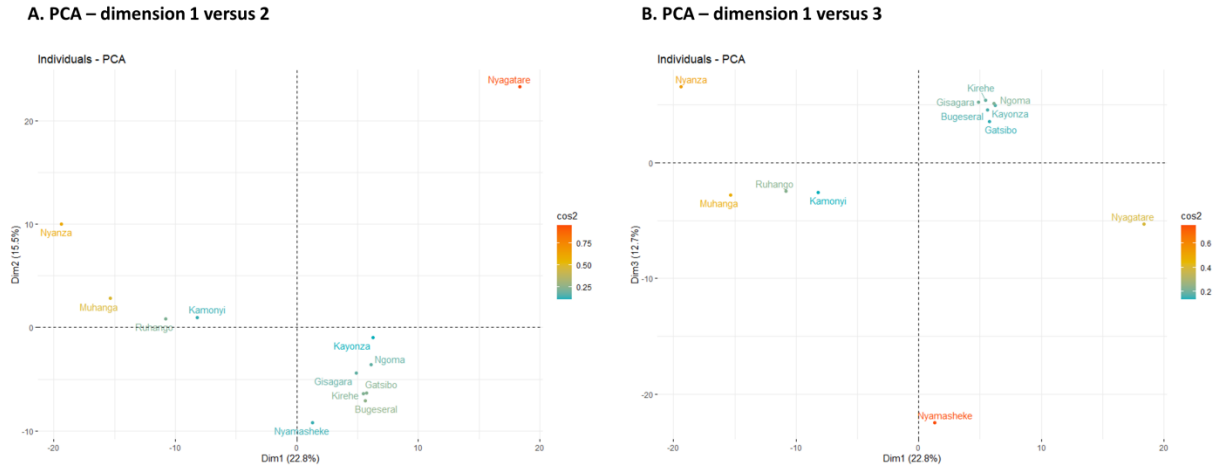

Fig 4

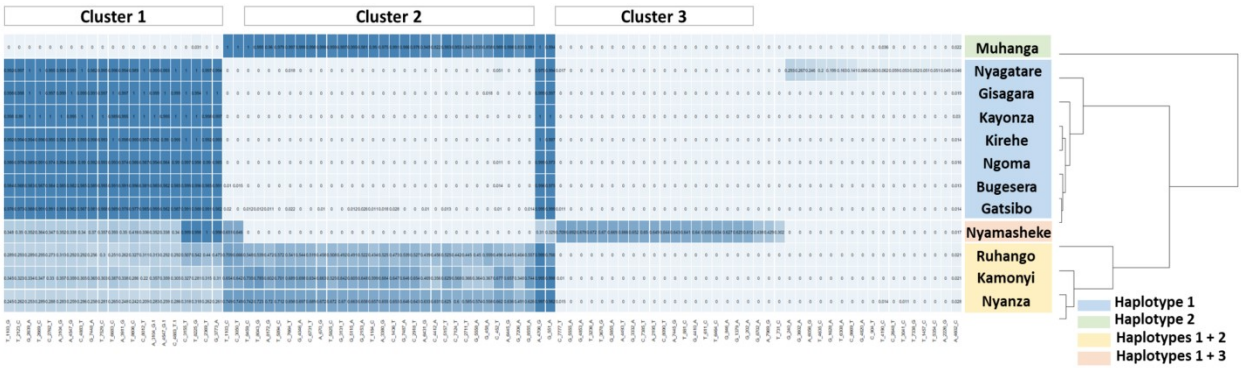

Fig 5