Cassava brown streak virus has a rapidly evolving genome: implications for virus speciation, variability, diagnosis and host resistance

Titus Alicai[1], Joseph Ndunguru[2], Peter Sseruwagi[2], Fred Tairo[2], Geoffrey Okao-Okuja[1], Resty Nanvubya[1], Lilliane Kiiza[1], Laura Kubatko[3], Monica A. Kehoe[4] and Laura M. Boykin[5,*]

[1]National Crops Resources Research Institute, P.O. Box 7084, Kampala, Uganda

[2]Mikocheni Agricultural Research Institute, Coca cola Road, Box 6226, Dar es Salaam, Tanzania

[3]The Ohio State University, 154W 12th Avenue, Columbus, Ohio 43210, USA

[4]Crop Protection Branch, Department of Agriculture and Food, Western Australia, Bentley Delivery Centre, Perth, 6983, Western Australia, Australia.

[5] The University of Western Australia, ARC Centre of Excellence in Plant Energy Biology and School of Chemistry and Biochemistry, Crawley, Perth 6009, Western Australia, Australia.

* Corresponding author: laura.boykin@uwa.edu.au

Keywords: cassava, Uganda, *Manihot esculenta,* smallholder farmer, *Cassava brown streak virus*, species tree estimation, SVD Quartets, nonsynonymous mutations

Abstract

Cassava is a major staple food for about 800 million people in the tropics and sub-tropical regions of the world. Production of cassava is significantly hampered by cassava brown streak disease (CBSD), which is caused by *Cassava brown streak virus* (CBSV) and *Ugandan cassava brown streak virus* (UCBSV). The disease is suppressing cassava yields in eastern Africa at an alarming rate. Previous studies have documented that CBSV is more devastating than UCBSV because it more readily infects both susceptible and tolerant cassava cultivars, resulting in greater yield losses. Using whole genome sequences from NGS data, we produced the first coalescent-based species tree estimate for CBSV and UCBSV. This species framework led to the finding that CBSV has a faster rate of evolution when compared with UCBSV. Furthermore, we have discovered that in CBSV, nonsynonymous substitutions are more predominant than synonymous substitution and occur across the entire genome. All comparative analyses between CBSV and UCBSV presented here suggest that CBSV may be outsmarting the cassava immune system, thus making it more devastating and harder to control.

Introduction

Cassava (*Manihot esculenta* Crantz) is a major staple food crop for over 800 million people in over 100 tropical and sub-tropical countries[1]. In sub-Saharan Africa, it is the main source of dietary calories for approximately 300 million people[2]. The tuberous storage roots of cassava are rich in carbohydrates and can be cooked or processed for human food, animal feeds and a wide range of industrial products. The crop is relatively drought tolerant and can yield well even in less fertile soils, hence, its importance to poor families farming marginal lands[3]. Cultivation of cassava is most adversely affected by two viral diseases: cassava mosaic disease (CMD) and cassava brown streak disease (CBSD)[4], which together were reported to cause production losses of more than US$1 billion every year[5] in Africa.

Serious yield losses due to CMD were first observed on mainland East Africa in the 1920s[6]. Recorded epidemics of CMD later occurred in the 1930s, 1940s and from 1990s to date[7,8]. By contrast, for about 70 years since it was first described[9], CBSD was confined to low altitudes (below 1000 meters above sea level) along coastal eastern Africa in Kenya, Tanzania, Mozambique and Malawi. However, in the early 2000s, outbreaks of CBSD were reported over 1000 km inland at mid-altitude locations (above 1000m) in multiple countries all around Lake Victoria in Uganda[10], western Kenya[11] and northern Tanzania[4]. Where it is already established in eastern Africa, the current CBSD epidemic prevails as the main cause of losses in cassava production. Over the last 10 years, the CBSD epidemic has expanded to other countries in East and Central Africa such as Rwanda, Burundi, Congo, DR Congo and South Sudan[12-14]. This has significantly increased the risk to countries in central and west Africa which are among the world's leading cassava producers, and where CBSD does not occur.

CBSD is caused by *Cassava brown streak virus* (CBSV) and *Ugandan cassava brown streak virus* (UCBSV). Both viruses are (+) ssRNA viruses in the genus *Ipomovirus* and family *Potyviridae*[15-18], and are often together referred to as cassava brown streak viruses (CBSVs). The CBSVs have genomic organization of 10 segments, total size approximately 8.9 to 10.8 kb, and coding for a polypeptide with about 2,900 amino acid residues[15,17,18]. The complete genome of a CBSD causal virus was first sequenced in 2009[18], and to date there are only 26 publicly available[19]. Currently there are two species recognized by the ICTV, but Ndunguru et al.[19] have suggested further speciation in the UCBSV clade. Both viruses are transmitted in a semi-persistent manner by the whitefly *Bemisia tabaci*[20] and mechanically[21]. Symptoms of CBSD on cassava vary with cultivar, virus or plant age, but typically include leaf veinal choloris, brown stem lesions, as well as constrictions, fissures and necrosis of the tuberous storage roots[22,23]. Overall, both CBSV and UCBSV cause similar symptom types, however, infection with CBSV tend to result in more severe sumptoms.

Although CBSD has become established in eastern Africa, there is limited knowledge on the diversity of causal viruses, their distribution and evolutionary potential. Therefore, it is necessary to obtain several full genome sequences of CBSD viral isolates, better understand the causal viruses and design long-term control approaches for the disease.

In contrast to the growing knowledge on the causal agents of CBSD, host-pathogen interactions are less clear. As such, little is known about specific responses of different cassava varieties to prevailing species or strains of CBSD viral pathogens. Development and dissemination of CBSD-tolerant varieties has been the main means adopted for CBSD control in eastern Africa. With significant efforts geared at breeding for CBSD-resistant varieties, it is of great interest to know

97 if such resistance protects cassava against one or both CBSVs. The resistance may be expressed
98 as several related features including restricted infection, systemic spread or recovery of infected
99 plants from disease and the possibility that stem cuttings taken from these may give rise to
100 progeny that are virus-free (reversion). Recent studies have shown CBSV to be the more
101 aggressive virus, infecting both tolerant and susceptible cultivars as single or mixed infections
102 with UCBSV[15,24,25]. In contrast, tolerant varieties were infected with only CBSV, but free of UCBSV,
103 suggesting their resistance to the latter. Compared with UCBSV, CBSV isolates have been
104 reported to be more detectable, having higher infection rates by graft inoculation and inducing
105 more severe symptoms[26]. It has also been shown that plants of CBSD tolerant or resistant
106 cultivars graft-inoculated with UCBSV developed milder symptoms and a significantly higher
107 proportion of the progenies were virus-free (reverted) compared to those infected with CBSV[27].
108 To date, the underlying reasons for this more aggressive nature of CBSV compared with UCBSV
109 are not known.
110
111 In this study, CBSV and UCBSV molecular diversity was investigated by using next generation
112 sequencing to understand new complete genomes of three isolates from Uganda. The sequences
113 obtained were analyzed to determine species composition, CBSV and UCBSV evolutionary rates,
114 potential role of such changes in virus-host interactions, resulting into cassava cultivar
115 susceptibility or resistance. We set out to answer the following questions:
116 1) How do the three new complete genomes from Uganda compare to those already
117 published[19]?
118 2) Are CBSV and UCBSV distinct species and is there further speciation?
119 3) Why is CBSV more aggressive and harder to breed resistance for than UCBSV?
120

Results

CBSD Field Symptoms Associated with CBSV and UCBSV Isolates

Categorisation of CBSD foliar symptom distribution on symptomatic plants assessed revealed that the most frequently encountered type was LL - symptoms only on lower leaves (68.4%), followed by SW - systemic and on the whole plant (26.3%), and SL – systemic but localized (5.3%) (table 1). Based on CBSVs detected and CBSD leaf symptom severity scores for 57 sampled plants, whereas the majority of plants infected by UCBSV alone as determined by RT-PCR had mild chlorosis (severity score 2), CBSV infections (single or mixture with UCBSV) tended to have moderate to severe symptoms (scores 3-4) in same proportion to those exhibiting score 2 (Fig. 1, Table 1). Regarding the three isolates used here for whole genome sequencing, U8 (UCBSV) was from a plant with CBSD score 3 and LL symptom type. Both CBSV isolates (U1 and U4) were from plants with severity scores 2 and 3, symptom types LL and SL, respectively.

Next Generation Sequencing

The three samples from Uganda produced raw reads ranging from 21,844,716 to 23,648,990. After trimming for quality using CLCGW, these numbers were reduced to 21,582,374 to 23,373,606 (table 2). Following *de novo* assembly of the trimmed reads using CLCGW, the numbers of contigs produced were 621-1,008. The contigs of interest from *de novo* assembly were of lengths 2,214 to 8,954nt, with average coverage 24 to 366. After mapping to a reference genome in Geneious, the lengths of the consensus sequences were 8,893 to 9,563 with average coverages of 25 to 393. The final sequences consisted of a consensus between the *de novo* and the mapped consensus with lengths of 8,700 to 8,748.

Genomic Variability and Positive Selection

145 The CBSV genomes included in this study were more variable when compared with those of
146 UCBSV (supplementary figs. S1 and S2). Characterizing amino acid usage at each position in the
147 whole genome revealed that CBSV genomes have non-synonymous substitutions present across
148 their entire genome (Fig. 2), and predominating when compared to synonymous substitutions.
149 In contrast, UCBSV had near equal non-synonymous and synonymous substitution rates across
150 the entire genome. Genes in the UCBSV genomes with non-synonymous substitutions at a higher
151 frequency were: P1, NIb and HAM1 (Fig. 2).
152
153 CBSV had 68 positively selected sites and 66 negatively selected sites, while UCBSV had zero
154 positively selected sites and 558 negatively selected sites (Table 3). Analyzed together there are
155 3 positively selected sites and 1383 negatively selected sites. The coat protein (CP) of CBSV had
156 the highest number of positively selected sites (16) while 6K2 had zero.
157
158 **Rates of Evolution**
159 CBSV and UCBSV have different rates of evolution (Table 4). We tested two hypothesis using
160 CODEML. The null hypothesis tested was that CBSV and UCBSV have equal rates of evolution
161 while the alternative hypothesis was that CBSV and UCBSV have different rates of evolution (two
162 omegas; model = 2). The Likelihood Ratio Test was used to test for significance. If the difference
163 in likelihood was greater than 3.84 (based on the Chi-squared distribution and one degree of
164 freedom) we rejected the null hypothesis that the rates between CBSV and UCBSV are equal.
165
166 CBSV whole genome sequences showed it is evolving 5 times faster than UCBSV overall. The
167 genes contributing to this accelerated rate of evolution for CBSV are Nla (D=29.95), followed by
168 6K2 (D=6.74), Nlb (D=5.18) and P1 (4.61) (Table 4). The transition/transversion ratios were also

169 estimated using CODEML, and show that the 6K1 (19.6) and CP (13.2) genes have the highest
170 estimates while the remaining 8 genes ranged from 5.05 – 9.93.
171
172 **Species Tree Estimation - SVDQ**
173 The species phylogeny (Fig. 3) shows strong support for a split into two primary viral clades, one
174 consisting of CBSV (Fig. 3 clades A and B) and the other consisting of UCBSV (Fig. 3 clades E-G),
175 with 100% bootstrap support separating the two clades. Figure 3 shows clades labeled A–G which
176 correspond to; 1) labels A-F from Ndunguru et al [19], and 2) a new clade G defined in this study.
177 Within the CBSV clade, there are several additional clades with 100% bootstrap support,
178 including the two new CBSV whole genomes from Uganda (U1 and U4). These are the first CBSV
179 whole genomes sequences from Uganda. The other CBSV grouping with 100% bootstrap support
180 labeled B in Figure 3 contains 4 Tanzania samples KoR6, Tan 79, Tan 19 1 and Nal 07. In the
181 UCBSV clade there are 6 nodes supported with a 100% bootstrap, including the new UCBSV whole
182 genome added from this study (U8) which is sister to Kab 07 from Uganda. In addition, the CBSV
183 clade had all samples from a given country grouping together while the UCBSV clade had
184 monophyletic clades from different countries (the multi-colored lines in Fig. 3).
185
186 **Comparison of Gene Trees to Species Tree**
187 Clades A and B, which partition the CBSV isolates into two groups, are consistently present with
188 high support in all genes except HAM1 and CP (Table 5). Clades D and G, which each consist of a
189 pair of UCBSV isolates, have high support across all genes, while clades C and E have relatively
190 high support across a majority of genes. Clade F is strongly supported by the CI gene, which is
191 relatively long, but is not found in the phylogenetic tree estimated for any of the other genes.
192

193 The whole genome concatenated analysis using MrBayes shows strong support (posterior
194 probability 1.0) for all clades (Table 5). However, this analysis does not take into account the
195 possibility of variation in the evolutionary processes across the individual genes. The SVDQ
196 analysis, on the other hand, uses a coalescent-based method to estimate the overall species tree,
197 and properly accounts for variation in the evolutionary history for each gene. In viewing the
198 bootstrap support values for each of the clades from the SVDQ analysis, we see that the level of
199 support for each clade across the genome is more accurately represented by the corresponding
200 bootstrap proportion. For example, clade F, which was found only in the phylogeny of the CI
201 gene, shows a bootstrap proportion of 0.44 for the SVDQ analysis (as compared to 1.0 for the
202 MrBayes concatenated analysis) (Table 5). Similarly, the SVDQ analysis gives a bootstrap
203 proportion of 0.87 for clade E, which showed posterior probabilities below 0.8 for 3 of the 10
204 genes, as compared to a posterior probability of 1.0 for the concatenated analysis with MrBayes.
205 All other clades are supported with bootstrap values of 1.0, consistent with the MrBayes analysis.
206

207 **Sliding Window SVD Score**

208 The SVD Score Sliding Window analysis (Fig. 4) shows several interesting patterns. First, note that
209 the gene boundaries track well with shifts in the magnitude of the SVD Score, indicating that
210 individual genes are subject to specific evolutionary processes that vary from gene to gene. In
211 particular, several genes show strong support for the primary CBSV/UCBSV split, as indicated by
212 their low scores, while other genes show variation from this basic process, as indicated by
213 increases in the scores. In addition, Fig. 4 shows the test statistic associated with the hypothesis
214 test for a shift in the rate of evolution between the two groups, with `*' indicating that the rate
215 difference between the two groups is statistically significant. It is readily apparent from the graph
216 that genes that show strong support (low SVD Score) for the primary CBSV/UCBSV split also show

strong evidence for statistically significant differences in evolutionary rate. These results support the overall hypothesis that certain genes in CBSV have accelerated rates of evolution that may be contributing to the higher aggressiveness of the virus.

Discussion

In this study, we analyzed new and all publicly available CBSVs whole genomes to elucidate molecular mechanisms underlying the field and laboratory observations that CBSV more readily infects cassava plants and tends to display severe symptoms when compared with UCBSV. Our analyses included characterizing three new complete CBSV (2) and UCBSV (1) genomes, which were combined with the 26 previously published. Our major findings show further speciation of CBSV and UCBSV, a larger genetic landscape for CBSV, including many nonsynonymous sites, and reveal that CBSV has a faster rate of evolution compared with UCBSV (Table 4 and Fig. 4). These observations and their biological significance are discussed.

Genes with Accelerated Rates of Evolution in CBSV

We have identified P1, 6K2, NIb and NIa as the genes with accelerated rates of evolution in CBSV. The function of P1 is as an RNA silencing suppressor (RSS), and there is also the suggestion that it may be involved in virion binding to the whitefly stylet via a “bridge” formation by a virus-encoded P1 protein for both CBSV and UCBSV. 6K2 is associated with cellular membrane and responsible for systemic infection and viral long distance movement[28]. The NIb encodes for a nuclear inclusion polymerase and the NIA for a nuclear inclusion protease[18,29].

In Potyviruses generally, when NIa and VPg are associated together they are located in the cytoplasm and nucleus of infected cells. When 6K2-VPg-NIa forms a larger product, the VPg plays

a role in viral RNA replication[30]. Even though VPg is not one of the genes with a higher evolution rate, both 6K2 and NIa are a part of the complex which affects replication, and this may go some way to explaining their apparent accelerated evolution rate. Is it possible that the accelerated rates of evolution for genes involved in replication could even be a response to the relatively recent interaction of the viruses and cassava? The CBSVs are not present in South America where cassava originates, so the viruses must be native to Africa. It would appear that the adaptation is still occurring and the cassava immune system does not know how to effectively fight these infections. Cassava was introduced to East Africa in the 18th century through oceanic movement. The first report of cassava brown streak disease was in 1936[9,13]. There has been little opportunity for the co-evolution of the viruses and the host, therefore natural resistance would be a hard prospect. This raises the possibility of the original host of these viruses, a non-cassava host which may be harboring the CBSVs or the most recent common ancestor of these viruses. This in turn leads us to wonder just how old these viruses and their ancestors are, if any of the CBSVs is the ancestral species, and the best way to answer such questions is to sequence more virus genomes from both cassava and non-cassava hosts wherever they are found.

How Can CBSV Still Function with Such a Large Genetic Landscape?

CBSV and UCBSV have different evolutionary patterns as observed by characterizing the whole genome sequences of CBSV and UCBSV separately. CBSV is genetically more diverse when compared with UCBSV, as evident by the greater amino acid usage (supplementary Fig. 1), the faster rates of evolution across the entire genome (table 4), and greater number of nonsynonymous sites across the entire genome (Fig. 2). How can CBSV still function with such a large genetic landscape? RNA viruses walk a very fine line of having the genetic arsenal to overcome the host immune system and diverging to a point that key functions of genes are lost

[31]. Recent studies[32,33] have shown that viruses with a large genetic landscape adapt to host changes much quicker and can overcome the host immune system faster. Viruses that occupy a large portion of the possible sequence space might be less fit but they outcompete the fitter strain when the host immune system shifts and hence these viruses have been described as adapted to "survival of the flattest"[34,35]. This means that a virus that covers the most sequence space will be able to adapt to host immune system faster than those with smaller spaces. Viruses that are adapted in this category ("survival of the flattest") are going to be harder to breed resistance for because the virus has a larger ability to adapt to changes. It is clear that in our case, CBSV is the virus that has a larger sequence space (Supplemental Fig. 1) when compared to that of UCBSV, which is clearly smaller (Supplemental Fig. 2). CBSV is one of the RNA viruses that can be described as adapted to "survival of the flattest", while UCBSV is not. Therefore, CBSV is more devastating because it has a larger genetic arsenal which it uses overcome the changes breeders are introducing into cassava.

Not only are the CBSV genomes more genetically diverse, but are also characterized by a large number of nonsynonymous changes in the genome (Fig. 2). An excess of nonsynonymous over synonymous substitutions at individual amino acid sites signifies that positive selection has affected the evolution of a protein between the extant sequences under study and their most recent common ancestor[36]. Positive selection is the process by which new advantageous genetic variants sweep a population and is the mechanism Darwin described to drive evolution. This is further evidence that might suggest that CBSV has a greater capacity to evade the cassava immune system as compared with UCBSV. CBSV had 66 sites under positive selection (Table 4) while UCBSV had none. The CBSV sites under positive selection are found not only in the regions that have gained the most attention, CP and HAM1-like[13], but are also found in all other genes

except 6K2. This is further support for CBSV's potential ability to outsmart the cassava immune system. Every gene in the CBSV genome (except 6K2) has sites under positive selection indicating effective RNA silencing of the virus will need to encompass many loci.

Using computational methods combined with field observations we have concluded that CBSV is more devastating than UCBSV. This assertion is also supported by two recent biological studies. The first was a test of reversion in three different cassava varieties (Albert, Kaleso and Kiroba) infected with CBSV and UCBSV. Reversion is a type of resistance mechanism whereby virus-infected plants will naturally recover from infection over time, and a proportion of their progeny from stem cuttings are virus-free. A reversion event infers the host immune system was able to clear or restrict the virus from systemic movement. It was shown that UCBSV-infected cassava had a higher rate of reversion when compared to plants infected with CBSV[27] indicating that plants infected with UCBSV recovered more often than those infected with CBSV. This is another line of evidence supporting the more devastating nature of CBSV, and the possibility that cassava immune systems of the three varieties tested are struggling to resist the virus.

The second study supporting the hypothesis that CBSV is more aggressive than UCBSV analyzed virus-derived small RNAs within three cassava varieties (NASE 3, TME204 and 60444). Plants infected with viruses are known to trigger RNAi antiviral defense that can be measured by quantifying the abundance of 21-24 nucleotide (nt) segments produced by the dicer enzyme[37]. Cassava varieties were infected with either CBSV or UCBSV, NGS was used to detect virus-derived small RNAs[24], and the 21-24 nt dicer fragments were mapped to either CBSV or UCBSV depending on which virus was used to infect the plant. The results showed that CBSV infection triggered a stronger immune response as measured by greater abundance of virus derived small RNA

fragments across the entire CBSV genome compared with UCBSV. In addition, across all three genotypes they observed that cassava grafted with CBSV-infected buds showed more severe symptoms compared to UCBSV-infected plants[24]. This is further evidence that CBSV is a more aggressive virus and breeding for resistance to CBSV and UCBSV will require different experimental approaches.

Implications of the Species Tree for CBSV and UCBSV

We have produced the first species tree estimate of the CBSD causal virus species using whole genome sequences and the coalescent-based SVD Quartets species tree estimation algorithm. Differences in the evolutionary history of the two viruses are seen in the branching patterns in Figure 3. CBSV has diverged into two main clades A and B, while UCBSV has several well-supported clades but the backbone is still unresolved, indicating more sampling is needed to fully understand the diversity and evolutionary history of UCBSV. The species tree (Fig. 3) is similar to the concatenated whole gene tree reported in Ndunguru et al.[19], except addition of the clade labeled "G", and lack of support for clades E and F in the UCBSV species. It is well-documented that concatenating genes without using the coalescent-based models can produce misleading results [38,39]. In our case, only CI supports clade F, and it is also the longest gene (1,883 bp), and therefore may swamp the signal of the other genes. The whole genome concatenation analysis recovers clade F with a posterior probability of 1.00 (Table 4). With regard to clade E, the SVDQ tree was more reflective of the individual gene tree signal by producing a bootstrap value of 0.87 versus 1.00 for the whole genome concatenated tree (Table 4). These results suggest that the estimated topology in the UCBSV species may be further refined as more samples are added.

Our integrative approach of species tree estimation coupled with analyzing rates of evolution has lead to a new framework for CBSV and UCBSV, which includes analyzing and treating these two groups of viruses as separate species. Multiple putative species of both CBSV and UCBV have been identified which means cassava needs to be resistant to the virus species that are prevalent in farmers' fields. We argue that this genomic diversity and faster rate of evolution for CBSV is what is causing the breeders to struggle with breeding resistant varieties and also why the diagnostic primers are not working consistently. CBSV also has more positively selected sites than UCBSV. It was first thought that CBSD was restricted to the coastal areas and below 1000 m [23] but as more genetic data is gathered CBSV and UCBSV are found at all elevations in many ecozones throughout eastern Africa[4,10,13,15,19,40]. We are still in the discovery phase with CBSV and UCBSV species as there are only 29 (now with the three new included here) whole genome sequences and other new species of both viruses are likely to be discovered. As we move forward it is important to include all known samples and use appropriate species tree estimation methods such as SVDQ.

Finally, the traditional gene regions (CP and HAM1-like) that are used to delimit species and serve as the targets for diagnostic primers do not recover the species tree (Table 4). We recommend designing new diagnostic regions for other genes that recover the species tree and also do not have an accelerated rate of molecular evolution (Figure 4), such as CI or P3 for species-level diagnoses. It is possible that the spread of CBSV and UCBSV could have been exacerbated through dissemination of infected cuttings, as virus indexing with primers targeting CP may have misleadingly returned negative results.

Implications of the Results for Cassava Breeding

During the last three decades worldwide, agricultural production has been compromised by a series of epidemics caused by new variants of classic viruses that show new pathogenic and epidemiological properties. An important determinant of the fitness of a virus in a given host is its ability to overcome the defenses of the host. Overcoming plant resistance by changes in the pathogenicity of viral populations represents a specific and important case of emergence, with tremendous economic consequences since it jeopardizes the success and durability of resistance factors in crops as an anti-viral control strategy. In this study, we found CBSV to be more variable, to have more positively selected sites, and to evolve five times faster than UCBSV. These findings have huge implications for cassava improvement efforts in Africa where CBSV is widely present. Field and laboratory results have proven CBSV to be more virulent and more devastating than UCBSV. Knowledge of specific virus species an improved cassava variety is resistant to will determine where to screen, multiply and deploy such varieties. Cassava breeders have to take into consideration the evolutionary and biological differences between CBSV and UCBSV in the breeding programs. For example, cassava breeders can breed varieties that are resistant to CBSV that can be strategically deployed in areas where CBSV is more prevalent, and similarly for UCBSV. Furthermore, it becomes more appropriate to always screen cassava materials against CBSV as a minimum, even if UCBSV is the more prevalent virus. Such a strategy will in effect ensure durable resistance as opposed to the indiscriminate screening and distribution of the improved CBSD resistant cassava varieties, without knowledge of the virus species in the area.

Methods

Field Plant Sample Collection

Farmers' fields in Uganda with cassava plants 3-6 months old were surveyed for CBSD in 20 districts. In each field, cassava plants were visually assessed to confirm typical CBSD symptoms

385 on leaves and stems. CBSD leaf symptom severity was scored on a 1-5 scale[41,42]; 1 = no visible
386 symptoms, 2 = mild vein yellowing or chlorotic blotches on some leaves, 3 =
387 pronounced/extensive vein yellowing or chlorotic blotches on leaves, but no lesions or streaks
388 on stems, 4 = pronounced/extensive vein yellowing or chlorotic blotches on leaves and mild
389 lesions or streaks on stems, 5 = pronounced/extensive vein yellowing or chlorotic blotches on
390 leaves and severe lesions or streaks on stems, defoliation and dieback. CBSD symptoms were also
391 categorized based on distribution of leaf chlorosis and stem lesions on the plant; systemic and
392 on the whole plant (SW), systemic on leaf or stem parts but localized (SL), only on lower leaves
393 (LL). On selected symptomatic plants, portions of the third fully expanded leaf on a shoot were
394 picked as samples, air-dried by pressing between sheets of newsprint and stored pending RNA
395 extraction.
396
397 **RNA Extraction**
398 About 0.25 g cassava leaf samples were frozen in liquid nitrogen, then ground using a mortar and
399 pestle. 2 ml CTAB lysis buffer (2% CTAB; 100 mM Tris–HCl, pH 8.0; 20 mM EDTA; 1.4 M 134 NaCl;
400 1% sodium sulphite; 2% PVP) was added and samples homogenized. The 1 ml of the homogenate
401 was incubated at 65°C for 15 min, an equal volume of chloroform: isoamyl alcohol (24:1) was
402 added, and the sample was centrifuged for 10 min at approximately 14,500rpm. 800µl of the
403 aqueous layer was transferred to a new tube with an equal volume of 4 M LiCl and incubated at
404 -20°C for 2 hrs. The samples were centrifuged for 25 min at 14,500 rpm and the supernatant was
405 poured off. The pelleted RNA was re-suspended in 200 µl TE buffer containing 1% SDS, 100 µl of
406 5M NaCl. 300 µl of ice-cold isopropanol were added and incubated at -20°C for 30 min. The
407 sample was centrifuged at 13,000 rpm for 10 min and the aqueous layer was decanted and RNA
408 pellets washed in 500 µl of 70% ethanol by centrifuging at 13,000 rpm for 5 min. The ethanol was

409 decanted off and RNA pellet dried to remove residual ethanol. The RNA was re-suspended in 50
410 µl nuclease-free water and stored at −80°C prior to testing.
411
412 **CBSV and UCBSV Detection by RT-PCR**
413 All samples were tested for presence of CBSV and UCBSV by a two-step RT-PCR assay [43]. The PCR
414 mixture consisted of 16.0 µl nuclease free water, 2.5 µl PCR buffer, 2.5 µl $MgCl_2$ (2.5 mM), 0.5µl
415 dNTPs (10 mM), 1.0 µl of each primer (10mM) [forward CBSDDF2 5'-
416 GCTMGAAATGCYGGRTAYACAA-3' and reverse CBSDDR 5'-GGATATGGAGAAAGRKCTCC-3'], 0.5 µl
417 Taq DNA polymerase and 1.0 µl of cDNA. The PCR thermo profile consisted of: 94°C for 2 min
418 followed by 35cycles of 94°C (30 s), 51°C (30 s) and 72°C (30 s) for denaturation, annealing and
419 extension, respectively. PCR products were analysed by electrophoresis in a x1 TAE buffer on a
420 1.2% agarose gel, stained with ethidium bromide, visualized under UV light and photographed
421 using a digital camera.
422
423 **Sample Selection for Sequencing**
424 From the data obtained in the diagnostic tests, samples for sequencing were selected to
425 represent different geographical regions, symptom types and severities. Three samples that
426 tested positive for either CBSV (2) or UCBSV (1) were selected for this study. The two samples for
427 which presence of CBSV was confirmed (U1 and U4) had been collected from different farmers'
428 fields in Mukono district, central Uganda. The sample with UCBSV (U8) selected for further
429 analysis originated from a field in Mayuge district, eastern Uganda.
430
431 **Generation of the Transcriptomes**

The three samples were transported to the laboratory and extracted as detailed above. Total RNA was blotted on to FTA cards and later extracted using methods previously described[44]. Total RNA from each sample was sent to the Australian Genome Research Facility (AGRF) for library preparation and barcoding before 100 bp paired-end sequencing on an Illumina HiSeq2000.

De novo Sequence Assembly and Mapping

For each sample, reads were first trimmed using CLC Genomics Workbench 6.5 (CLCGW) with the quality scores limit set to 0.01, maximum number of ambiguities to two and removing any reads with <30 nucleotides (nt). Contigs were assembled using the *de novo* assembly function of CLCGW with automatic word size, automatic bubble size, minimum contig length 500, mismatch cost two, insertion cost three, deletion cost three, length fraction 0.5 and similarity fraction 0.9. Contigs were sorted by length and the longest subjected to a BLAST search (blastn and blastx)[45]. In addition, reads were also imported into Geneious 6.1.6[46] and provided with reference sequences obtained from Genbank (KR108828 for CBSV and KR108836 for UCBSV). Mapping was performed with minimum overlap 10%, minimum overlap identity 80%, allow gaps 10% and fine tuning set to iterate up to 10 times. A consensus between the contig of interest from CLCGW and the consensus from mapping in Geneious was created in Geneious by alignment with MAFFT[47]. Open reading frames (ORFs) were predicted and annotations made using Geneious. Finalized sequences were designated as “complete” based on comparison with the reference sequences used in the mapping process, and “coding complete” if some of the 5’ or 3’ UTR was missing but the coding region was intact[48,49], and entered into the European Nucleotide Archive (WEBIN ID number Hx2000053576).

Genome Alignment and Annotation

Twenty-six whole genomes (12 CBSV and 14 UCBSV) were downloaded from GenBank and imported into Geneious[46], and the MAFFT plugin[47]was used to align them with the 3 new whole genome sequences obtained in this study. Nucleotide alignments were translated into protein using the translate align option in Geneious and then visually inspected for quality. Annotations were transferred to the 3 new genomes from the 26 previously published genomes using the live annotation option in Geneious.

Characterizing the Genetic Diversity in CBSV and UCBSV Genomes

CBSV and UCBSV are distinct species (Fig. 2) therefore the genomes were treated separately in the analyses in characterizing the genomes. Characterizing the genetic diversity of CBSV and UCBSV was done using the Synonymous Non-synonymous Analysis Program (SNAP v2.1.1) implemented in the Los Alamos National Laboratory HIV-sequence database (http://www.hiv.lanl.gov)[50]. SNAP calculates synonymous and non-synonymous substitution rates based on a set of codon-aligned nucleotide sequences. This program is based on the simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions of[51], and incorporating a statistic developed for computing variances and covariances of dS's and dN's[52]. An application of the SNAP package in HIV-1 research has also been developed[53].

Estimating Rates of Evolution

To further characterize the CBSV and UCBSV genomes, we estimated the rates of molecular evolution using CODEML implemented in PAML (Phylogenetic Analysis by Maximum Likelihood)[54]. PAML is a package of programs for analysis of DNA or protein sequences by using maximum likelihood methods in a phylogenetic framework. The null hypothesis tested was that

480 CBSV and UCBSV have equal rates of evolution (one omega; model = 0) while the alternative
481 hypothesis was that CBSV and UCBSV have different rates of evolution (two omegas; model = 2).
482 The Likelihood Ratio Test was used to test for significance. If the test statistic was greater than
483 3.84 (based on the Chi-squared distribution and one degree of freedom) we then rejected the
484 null hypothesis that the rates between CBSV and UCBSV are equal. Initial analyses were carried
485 out for the entire genome and showed that CBSV has a higher rate of evolution (Table 4). To
486 identify which gene or genes were contributing to the faster rate of evolution we repeated the
487 analysis separately for each individual gene.
488
489 **Testing for Positive Selection**
490 Sites under positive selection were identified using SLAC[55] implemented on the
491 http://www.datamonkey.org web server[56]. The settings used to run SLAC were as follows: the
492 best fitting model (GTR) was specified global dN/dS value was estimated and the significance
493 level was set to 0.01.
494
495 **Gene Tree Estimation**
496 Individual gene trees were estimated using MrBayes 3.2.1[57] run in parallel on Magnus (Pawsey
497 Supercomputing Centre, Perth, Western Australia) utilizing the BEAGLE library [58]. MrBayes 3.2.1
498 was run utilizing 4 chains for 30 million generations and trees were sampled every 1000
499 generations. All runs reached a plateau in likelihood score, which was indicated by the standard
500 deviation of split frequencies (0.0015), and the potential scale reduction factor (PSRF) was close
501 to one, indicating that the MCMC had converged.
502
503 **Species Tree Estimation**

The SVDQ method[59] implemented in PAUP*[60] was used to analyze the whole-genome data. This method enables analysis of multi-locus data in a coalescent framework that allows for variation in the phylogenetic histories of individual genes. The method was run with all possible quartets (23,751) sampled in each of 100 bootstrap replicates, and the consensus across all bootstrap replicates was used as the estimate of the species tree. Bootstrap support values for each node were used to quantify uncertainty in the species tree estimate. The entire analysis took approximately 2.5 minutes on a MacBook Pro running OSX 10.11.2 with a 2.2 GHz Intel Core i7 processor.

Comparison of Gene Trees to Species Tree

We compared the single-gene phylogenies constructed using MrBayes with the overall species tree phylogeny estimated using SVDQ and the concatenated phylogeny estimated by MrBayes. For each tree, we evaluated presence or absence of the clades identified by Ndunguru et al.[19] labeled A-F in Figure 3. We identified an additional clade (clade G, Fig. 3) that we noticed to be consistently present across genes and methods. For each of these clades present in a particular tree, we recorded the posterior probability (for trees constructed by MrBayes) or the bootstrap proportion (for the tree estimated by SVDQ) in Table 5.

Sliding Window SVD Score

The SVD Score[61] was used to quantify support for two viral clades for portions of the genome in a sliding window analysis. Briefly, the SVD Score measures the extent to which the data support a phylogenetic "split" – a division of the taxa into two groups with specified group membership. Low values of the SVD Score indicate strong support for the split of interest, while larger values indicate *either* a lack of support for the split or a shift in the underlying evolutionary process (see

528 Allman et al. (2016) for details and examples). We computed the SVD Score with the split defined
529 by CBSV vs. UCBSV across the genome in windows of 500 bp, sliding in increments of 100 bp, and
530 plotted the resulting SVD Scores across the genome, with boundaries between genes marked
531 with vertical lines. The computations took less than one minute on a MacBook Pro running OSX
532 10.11.2 with a 2.2 GHz Intel Core i7 processor.
533
534

Acknowledgments

This work was supported by the Bill and Melinda Gates Foundation Grant no. 51466 "Regional Cassava Virus Diseases Diagnostic Project" awarded to Mikocheni Agricultural Research Institute, Tanzania, and a sub-grant to National Agricultural Research Organisation (Uganda). Computational resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia supported this work.

Author contributions

Sample collection was carried out by T.A., G.O., R.N., L.Kiiza. Laboratory work was carried out by G.O, R.N., L.Kiiza. J.N. P.S. M.K. Computational analyses were conducted by T.A., L.B., L.Kubatko., M.K, and all authors contributed to the writing of the manuscript.

Conflict of Interest

The authors declare that they have no conflicts of interest with the contents of this article.

References

1 Thresh, J. M. Control of tropical plant virus diseases. *Adv Virus Res* **67**, 245-295, doi:10.1016/S0065-3527(06)67007-3 (2006).

2 Nweke, F. I. A Cash Crop in Africa. COSCA Working Paper No. 14. Collaborative Study of Cassava in Africa, International Institute of Tropical Agriculture, Ibadan, Nigeria. (1996).

3 Robertson, A. I. & Ruhode, T. The potential advantages of cassava over hybrid maize as a food security crop and a cash crop in the Southern Africa semi-arid zone. *African Crop Science Crop Science Conference Proceedings* **5**, 539-542 (2001).

559 4 Legg, J. P. *et al.* Comparing the regional epidemiology of the cassava mosaic and cassava
560 brown streak virus pandemics in Africa. *Virus Res* **159**, 161-170,
561 doi:10.1016/j.virusres.2011.04.018 (2011).

562 5 Legg, J. P., Owor, B., Sseruwagi, P. & Ndunguru, J. Cassava mosaic virus disease in East
563 and Central Africa: epidemiology and management of a regional pandemic. *Adv Virus*
564 *Res* **67**, 355-418, doi:10.1016/S0065-3527(06)67010-3 (2006).

565 6 Hall, F. W. Annual Report of the Department of Agriculture, Uganda. *Government*
566 *Printer, Entebbe*, 35 (1928).

567 7 Jameson, J. D. Cassava mosaic disease in Uganda. *East African Agricultural and Forestry*
568 *Journal* **29**, 208-213 (1964).

569 8 Otim-Nape, G. W. *et al.* The current pandemic of cassava mosaic virus disease in East
570 Africa and its control. *NARO/NRI/DFID Publication. Chatham, UK. 100pp* (2000).

571 9 Storey, H. H. Virus diseases of East African plants: VI, A progress report of studies of the
572 diseases of cassava. *East African Agricultural Journal* **2**, 34-39 (1936).

573 10 Alicai, T. *et al.* Re-emergence of cassava brown streak disease in Uganda. *Plant Disease*
574 **91**, 24-29 (2007).

575 11 Ntawuruhunga, P. & Legg, J. P. New Spread of Cassava Brown Streak Virus Disease and
576 Its Implications for The Movement of Cassava Germplasm in The East and Central African
577 Region. International Institute of Tropical Agriculture-Uganda & Eastern Africa Root
578 Crops Research Network Report. (2007).

579 12 Bigirimana, S., Barumbanze, P., Ndayihanzamaso, P., Shirima, R. & Legg, J. P. First report
580 of cassava brown streak disease and associated Ugandan cassava brown streak virus in
581 Burundi. *New Disease Reports* **24**, 26 (2011).

582 13 Mbanzibwa, D. R. *et al.* Evolution of cassava brown streak disease-associated viruses. *J*
583 *Gen Virol* **92**, 974-987, doi:10.1099/vir.0.026922-0 (2011).

14 Mulimbi, W. *et al.* First report of Ugandan cassava brown streak virus on cassava in Democratic Republic of Congo. *New Disease Reports* **26** (2012).

15 Winter, S. *et al.* Analysis of cassava brown streak viruses reveals the presence of distinct virus species causing cassava brown streak disease in East Africa. *J Gen Virol* **91**, 1365-1372, doi:10.1099/vir.0.014688-0 (2010).

16 Monger, W. A., Seal, S., Isaac, A. M. & Foster, G. D. Molecular characterization of cassava brown streak virus coat protein. *Plant Pathol* **50**, 527-534 (2001).

17 Monger, W. A. *et al.* The complete genome sequence of the Tanzanian strain of Cassava brown streak virus and comparison with the Ugandan strain sequence. *Arch Virol* **155**, 429-433, doi:10.1007/s00705-009-0581-8 (2010).

18 Mbanzibwa, D. R. *et al.* Genetically distinct strains of Cassava brown streak virus in the Lake Victoria basin and the Indian Ocean coastal area of East Africa. *Arch Virol* **154**, 353-359, doi:10.1007/s00705-008-0301-9 (2009).

19 Ndunguru, J. *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* **10**, e0139321, doi:10.1371/journal.pone.0139321 (2015).

20 Maruthi, M. N. *et al.* Transmission of Cassava brown streak virus by Bemisia tabaci (Gennadius). *J Phytopathol* **153**, 307-312 (2005).

21 Lister, R. M. Mechanical transmission of cassava brown streak virus. *Nature* **183**, 1588-1589 (1959).

22 Jennings, D. L. Observations on virus diseases of cassava in resistant and susceptible varieties. II. Brown streak disease. *Empire Journal of Experimental Agriculture* **28**, 261-269 (1960).

23 Nichols, R. F. W. The brown streak disease of cassava: Distribution, climatic effects and diagnostic symptoms. *East African Agricultural Journal* **15**, 154-160 (1950).

610 24 Ogwok, E., Ilyas, M., Alicai, T., Rey, M. E. & Taylor, N. J. Comparative analysis of virus-
611 derived small RNAs within cassava (Manihot esculenta Crantz) infected with cassava
612 brown streak viruses. *Virus Res* **215**, 1-11, doi:10.1016/j.virusres.2016.01.015 (2016).

613 25 Patil, B. L. *et al.* RNAi-mediated resistance to diverse isolates belonging to two virus
614 species involved in Cassava brown streak disease. *Mol Plant Pathol* **12**, 31-41,
615 doi:10.1111/j.1364-3703.2010.00650.x (2011).

616 26 Mohammed, I. U., Abarshi, M. M., Muli, B., Hillocks, R. J. & Maruthi, M. N. The symptom
617 and genetic diversity of cassava brown streak viruses infecting cassava in East Africa.
618 *Adv Virol* **2012**, 795697, doi:10.1155/2012/795697 (2012).

619 27 Mohammad, I. U., Ghosh, S. & Maruthi, M. N. Host and virus effects on reversion in
620 cassava affected by cassava brown streak disease. *Plant Pathology* **65**, 593-600 (2016).

621 28 Jiang, J., Patarroyo, C., Garcia Cabanillas, D., Zheng, H. & Laliberte, J. F. The Vesicle-
622 Forming 6K2 Protein of Turnip Mosaic Virus Interacts with the COPII Coatomer Sec24a
623 for Viral Systemic Infection. *J Virol* **89**, 6695-6710, doi:10.1128/JVI.00503-15 (2015).

624 29 Dombrovsky, A., Reingold, V. & Antignus, Y. Ipomovirus--an atypical genus in the family
625 Potyviridae transmitted by whiteflies. *Pest Manag Sci* **70**, 1553-1567,
626 doi:10.1002/ps.3735 (2014).

627 30 Revers, F. & Garcia, J. A. Molecular biology of potyviruses. *Adv Virus Res* **92**, 101-199,
628 doi:10.1016/bs.aivir.2014.11.006 (2015).

629 31 Sanjuan, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-
630 nucleotide substitutions in an RNA virus. *Proc Natl Acad Sci U S A* **101**, 8396-8401,
631 doi:10.1073/pnas.0400146101 (2004).

632 32 Lauring, A. S., Frydman, J. & Andino, R. The role of mutational robustness in RNA virus
633 evolution. *Nat Rev Microbiol* **11**, 327-336, doi:10.1038/nrmicro3003 (2013).

634 33 Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus
635 revealed through population sequencing. *Nature* **505**, 686-690,
636 doi:10.1038/nature12861 (2014).

637 34 Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E. & Adami, C. Evolution of digital organisms
638 at high mutation rates leads to survival of the flattest. *Nature* **412**, 331-333,
639 doi:10.1038/35085569 (2001).

640 35 Sanjuan, R., Cuevas, J. M., Furio, V., Holmes, E. C. & Moya, A. Selection for robustness in
641 mutagenized RNA viruses. *PLoS Genet* **3**, e93, doi:10.1371/journal.pgen.0030093 (2007).

642 36 Massingham, T. & Goldman, N. Detecting amino acid sites under positive selection and
643 purifying selection. *Genetics* **169**, 1753-1762, doi:10.1534/genetics.104.032144 (2005).

644 37 Blevins, T. *et al.* Four plant Dicers mediate viral small RNA biogenesis and DNA virus
645 induced silencing. *Nucleic Acids Res* **34**, 6233-6246, doi:10.1093/nar/gkl886 (2006).

646 38 Kubatko, L. S. & Degnan, J. H. Inconsistency of phylogenetic estimates from
647 concatenated data under coalescence. *Syst Biol* **56**, 17-24 (2007).

648 39 Degnan, J. H. & Salter, L. A. Gene tree distributions under the coalescent process.
649 *Evolution* **59**, 24-37 (2005).

650 40 Monger, W. A. *et al.* The complete genome sequence of the Tanzanian strain of Cassava
651 brown streak virus and comparison with the Ugandan strain sequence. *Arch Virol* **155**,
652 429-433 (2010).

653 41 Hahn, S. K., Isoba, J. C. G. & Ikotun, T. Resistance breeding in root and tuber crops at the
654 International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria. *Crop Protection* **8**,
655 147-168 (1989).

656 42 Hillocks, R. J., Raya, M. D. & Thresh, J. M. The association between root necrosis and
657 above ground symptoms of brown streak virus infection of cassava in southern Tanzania.
658 *International Journal of Pest Management* **42**, 285-289 (1996).

659 43 Mbanzibwa, D. R. *et al.* Simultaneous virus-specific detection of the two cassava brown
660 streak-associated viruses by RT-PCR reveals wide distribution in East Africa, mixed
661 infections, and infections in Manihot glaziovii. *J Virol Methods* **171**, 394-400,
662 doi:10.1016/j.jviromet.2010.09.024 (2011).

663 44 Ndunguru, J., Legg, J. P., Aveling, T. A., Thompson, G. & Fauquet, C. M. Molecular
664 biodiversity of cassava begomoviruses in Tanzania: evolution of cassava geminiviruses
665 in Africa and evidence for East Africa being a center of diversity of cassava geminiviruses.
666 *Virol J* **2**, 21, doi:10.1186/1743-422X-2-21 (2005).

667 45 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
668 search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).

669 46 Geneious v5.1. Available from http://www.geneious.com/ (2010).

670 47 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple
671 sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066
672 (2002).

673 48 Kehoe, M. A., Coutts, B. A., Buirchell, B. J. & Jones, R. A. Split personality of a Potyvirus:
674 to specialize or not to specialize? *PLoS One* **9**, e105770,
675 doi:10.1371/journal.pone.0105770 (2014).

676 49 Kehoe, M. A., Coutts, B. A., Buirchell, B. J. & Jones, R. A. Plant virology and next
677 generation sequencing: experiences with a Potyvirus. *PLoS One* **9**, e104580,
678 doi:10.1371/journal.pone.0104580 (2014).

679 50 Korber, B. in *Computational Analysis of HIV Molecular Sequences* (eds A.G. Rodrigo &
680 G. H. Learn) Ch. 4, 55-72 (Kluwer Academic Publishers, 2000).

681 51 Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and
682 nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-426 (1986).

683 52 Ota, T. & Nei, M. Variance and covariances of the numbers of synonymous and
684 nonsynonymous substitutions per site. *Mol Biol Evol* **11**, 613-619 (1994).

685 53 Ganeshan, S., Dickover, R. E., Korber, B. T., Bryson, Y. J. & Wolinsky, S. M. Human
686 immunodeficiency virus type 1 genetic evolution in children with different rates of
687 development of disease. *J Virol* **71**, 663-677 (1997).

54 Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-556 (1997).

55 Kosakovsky Pond, S. L. & Frost, S. D. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22**, 1208-1222, doi:10.1093/molbev/msi105 (2005).

56 Delport, W., Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455-2457, doi:10.1093/bioinformatics/btq429 (2010).

57 Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* **61**, 539-542, doi:10.1093/sysbio/sys029 (2012).

58 Ayres, D. L. *et al.* BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology* **61**, 170-173, doi:10.1093/sysbio/syr100 (2012).

59 Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317-3324, doi:10.1093/bioinformatics/btu530 (2014).

60 Phylogenetic Analysis using Parsimony (* and other methods) v. (open source version 4.0a147 downloaded from http://people.sc.fsu.edu/~dswofford/paup_test/ on March 25, 2016). (Sinauer Associates, Sunderland, MA 2002).

61 Allman, E. S., Kubatko, L. & Rhodes, J. A. Split scores: a tool to quantify phylogenetic signal in genome-scale data**, submitted, available at http://arxiv.org/abs/1608.00942** (2016).

712 Table 1: CBSD leaf symptom severities and types on plants infected by *Cassava brown streak virus*
713 and *Ugandan cassava brown streak virus*
714

Virus species	Number of plants with CBSD foliar symptom severity score[1]				Number of plants with CBSD foliar symptom type[2]		
	2	3	4	5	SW	LL	SL
CBSV	5	7	1	0	3	8	2
UCBSV	24	5	0	0	7	21	1
CBSV + UCBSV	9	6	0	0	5	10	0
Total	38	18	1	0	15	39	3
Percentage	66.7	31.6	1.7	0.0	26.3	68.4	5.3

715
716
717 [1]Foliar CBSD symptom severity score based on 1-5 scale; 1 = no visible symptoms, 2 = mild vein
718 yellowing or chlorotic blotches on some leaves, 3 = pronounced/extensive vein yellowing
719 or chlorotic blotches on leaves, but no lesions or streaks on stems, 4 = pronounced/extensive
720 vein yellowing or chlorotic blotches on leaves and mild lesions or streaks on stems, 5 =
721 pronounced/extensive vein yellowing or chlorotic blotches on leaves and severe lesions or
722 streaks on stems, defoliation and dieback.
723 [2]Types of foliar CBSD symptoms based on distribution of leaf chlorosis and stem lesions on the
724 plant; systemic and on the whole plant (SW), systemic on leaf or stem parts but localized (SL),
725 only on lower leaves (LL).
726

Table 2. Next generation sequencing data for samples from cassava brown streak disease symptomatic plants collected in Uganda.

Sample ID	Accession number	Virus	No. of reads obtained	No. of reads after trimming	Number of contigs produced (CLC)	Contig length (CLCGW, nt)	Average coverage (CLCGW)	Number of reads mapped to contig of interest	Ref seq. used for mapping	Length of consensus sequence from mapping (Geneious)	No. reads mapped to ref. sequence	Average coverage (Geneious)	Final sequence length (Coding region only
U1		*CBSV*	23,335,344	23,053,082	726	3919, 2214	31, 24	1264, 549	KR108828	8,893	2,233	25	8,748
U4		*CBSV*	21,844,716	21,582,374	621	8,949	255	23,658	KR108828	8,949	22,987	256	8,748
U8		*UCBSV*	23,648,990	23,373,606	1,008	8,954	366	33,778	KR108836	9,563	178,117	393	8,700

731 Table 3. *Cassava brown streak virus* (CBSV) amino acid (AA) sites under positive selection
732 (analyses method: SLAC Hy-Phy). There were no sites under positive selection for
733 *Ugandan cassava brown streak virus* (UCBSV).
734
735

Gene	CBSV AA site under positive selection
P1	44, 46, 50, 174, 224, 230, 250, 283, 288, 349, 358
P3	415, 455, 467, 472, 499, 525, 618
6K1	658, 678
CI	735, 761, 820, 827, 848, 852, 894, 935, 1218
VPg	1465
NIa	1620, 1645, 1704, 1754, 1785
Nib	1879, 1880, 1890, 1907, 1929, 2109, 2145, 2156, 2161, 2285
HAM1	2320, 2345, 2404, 2432, 2453, 2475, 2519
CP	2550, 2555, 2588, 2611, 2631, 2635, 2640, 2659, 2728, 2745, 2783, 2818, 2843, 2860, 2877, 2884

736
737
738
739
740
741
742
743
744
745
746
747
748
749
750

Table 4. Rates of evolution tested using CODEML implemented in PAML. H_O was CBSV and UCBSV have equal rates of evolution (one omega; model = 0), while H_1 was that CBSV and UCBSV have different rates of evolution (two omegas; model = 2).

Gene	Assumptions	K (ts/tv rate ratio)	w (omega Dn/Ds) 0	w (omega Dn/Ds)1	Likelihood Ratio Test Statistic (if greater than 3.84 reject H_0) H_0=equal rates
WGS	UCBSV and CBSV equal rates	5.90944	0.06358		
	UCBSV and CBSV different rates	5.9598	0.05518	**0.07622**	26.29*
P1	UCBSV and CBSV equal rates	5.07336	0.10394		
	UCBSV and CBSV different rates	5.05456	0.09047	**0.12203**	4.61*
P3	UCBSV and CBSV equal rates	5.17197	0.08635		
	UCBSV and CBSV different rates	5.20559	0.0764	**0.10198**	2.22
6K1	UCBSV and CBSV equal rates	19.54143	0.00969		
	UCBSV and CBSV different rates	19.69676	**0.01527**	0.00316	3.01
CI	UCBSV and CBSV equal rates	9.9388	0.01722		
	UCBSV and CBSV different rates	8.06276	0.0155	**0.01977**	0.73
6K2	UCBSV and CBSV equal rates	8.40649	0.04684		
	UCBSV and CBSV different rates	8.82738	0.02354	**0.11057**	6.74*
VPg	UCBSV and CBSV equal rates	5.852	0.05759		
	UCBSV and CBSV different rates	5.87009	0.054	**0.06323**	0.29
NIa	UCBSV and CBSV equal rates	8.01105	0.02932		
	UCBSV and CBSV different rates	8.68283	0.01408	**0.06719**	29.95*
NIb	UCBSV and CBSV equal rates	6.07872	0.05329		
	UCBSV and CBSV different rates	6.14047	0.0452	**0.06508**	5.18*
HAM1	UCBSV and CBSV equal rates	7.25177	0.16144		
	UCBSV and CBSV different rates	7.2343	**0.17929**	0.14007	2.23
CP	UCBSV and CBSV equal rates	13.09297	0.06075		
	UCBSV and CBSV different rates	13.26752	0.05606	**0.07155**	1.29
			faster rate UCBSV	**faster rate CBSV**	* Rates are different

Table 5. Support for Clades A – G (Figure 3) in individual gene trees and whole genome analyses. Table entries represent posterior probabilities from analysis with MrBayes, except values reported for SVDQ, which are bootstrap proportions. Support values below 95% are indicated in bold, and '--' indicates that the clade was not present.

Genomic region	Clade A	Clade B	Clade C	Clade D	Clade E	Clade F	Clade G
P1	99.98	99.98	96.98	99.72	99.54	--	99.98
P3	99.98	99.98	98.57	99.98	99.98	--	99.98
6K1	**94.66**	99.85	98.87	99.95	--	--	98.42
CI	99.98	99.96	99.76	99.98	99.98	99.99	99.99
6K2	97.65	99.98	**71.52**	**91.70**	**66.46**	--	98.75
Vpg	99.98	99.99	--	99.18	**64.55**	--	99.99
Nla	99.96	99.99	99.98	99.41	**76.31**	--	**89.34**
Nlb	99.99	99.98	99.98	99.99	99.98	--	99.98
HAM	--	--	--	99.89	99.83	--	98.55
CP	--	--	--	99.59	99.50	--	99.89
Whole genome	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SVDQ	1.00	1.00	1.00	1.00	**87.00**	**44.00**	1.00

Figure Legend

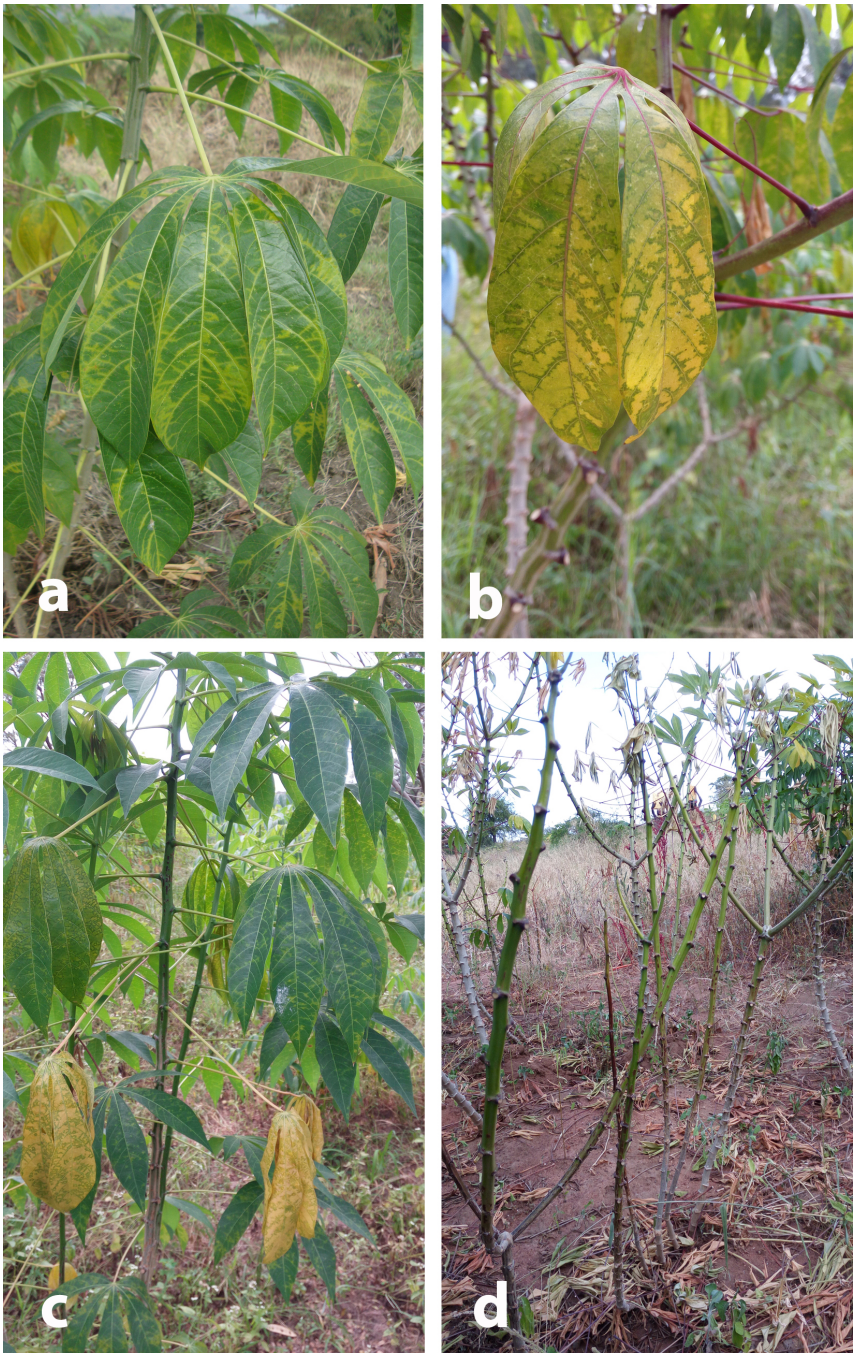
Figure 1: Cassava brown streak disease symptoms on leaves and stems of sampled plants; **(a)** Chlorosis along secondary and tertiary leaf veins of CBSV-infected plant of cultivar TME 204 (severity score 3), **(b)** Cultivar TME 14 plant with dual CBSV+UCBSV infection showing chlorosis on secondary or tertiary veins, reverse chlorosis (general chlorosis and green area along veins) (severity score 3), **(c)** UCBSV-infected plant of cultivar TME 204 exhibiting chlorosis on secondary veins, reverse chlorosis, chlorotic spots and mild stem lesions (severity score 3), **(d)** Very severely diseased plant (severity score 5) of cultivar TME 14 infected with both CBSV and UCBSV, and having chlorosis on leaves, severe stem lesions/brown streaks, defoliation, stem dieback.

Figure 2. Genetic diversity of CBSV and UCBSV using the Synonymous Non-synonymous Analysis Program (SNAP v2.1.1) implemented in the Los Alamos National Laboratory HIV-sequence database (http://www.hiv.lanl.gov) [50]. UCBSV is on the top panel, CBSV at the bottom. The 10 gene segments are labeled from P1-CP.

Figure 3. Species tree generated from SVD Quartets using the whole genome sequences. Colors at the tips are based on country of origin. Branches with mixed colors indicate a clade that contains samples with mixed country of origin. For example, the ancestral branch of UCBSV TZ Tan 23 KR108839 and UCBSV UG MI B3 FJ039520 is colored red and orange to indicate a clade with sampled with mixed country of origin.

Figure 4. Computed SVD Score with the split defined by CBSV vs. UCBSV across the genome in windows of 500 bp, sliding in increments of 100 bp, and resulting SVD Scores plotted across the genome. Boundaries between genes are marked with vertical lines to further characterize the CBSV and UCBSV genomes. Rates of molecular evolution were estimated using CODEML implemented in PAML (Phylogenetic Analysis by Maximum Likelihood)[54]. The results are shown for each gene and D represents the difference in likelihoods from the null hypothesis (CBSV and UCBSV have equal rates) and the alternative hypothesis (CBSV and UCBSV have different rates).

796 Figure 1: Cassava brown streak disease symptoms on leaves and stems of sampled plants; **(a)**
797 Chlorosis along secondary and tertiary leaf veins of CBSV-infected plant of cultivar TME 204
798 (severity score 3), **(b)** Cultivar TME 14 plant with dual CBSV+UCBSV infection showing chlorosis
799 on secondary or tertiary veins, reverse chlorosis (general chlorosis and green area along veins)
800 (severity score 3), **(c)** UCBSV-infected plant of cultivar TME 204 exhibiting chlorosis on secondary
801 veins, reverse chlorosis, chlorotic spots and mild stem lesions (severity score 3), **(d)** Very severely
802 diseased plant (severity score 5) of cultivar TME 14 infected with both CBSV and UCBSV, and
803 having chlorosis on leaves, severe stem lesions/brown streaks, defoliation, stem dieback.
804



805

Figure 2. Genetic diversity of CBSV and UCBSV using the Synonymous Non-synonymous Analysis Program (SNAP v2.1.1) implemented in the Los Alamos National Laboratory HIV-sequence database (http://www.hiv.lanl.gov) [50]. UCBSV is on the top panel, CBSV at the bottom. The 10 gene segments are labeled from P1-CP.

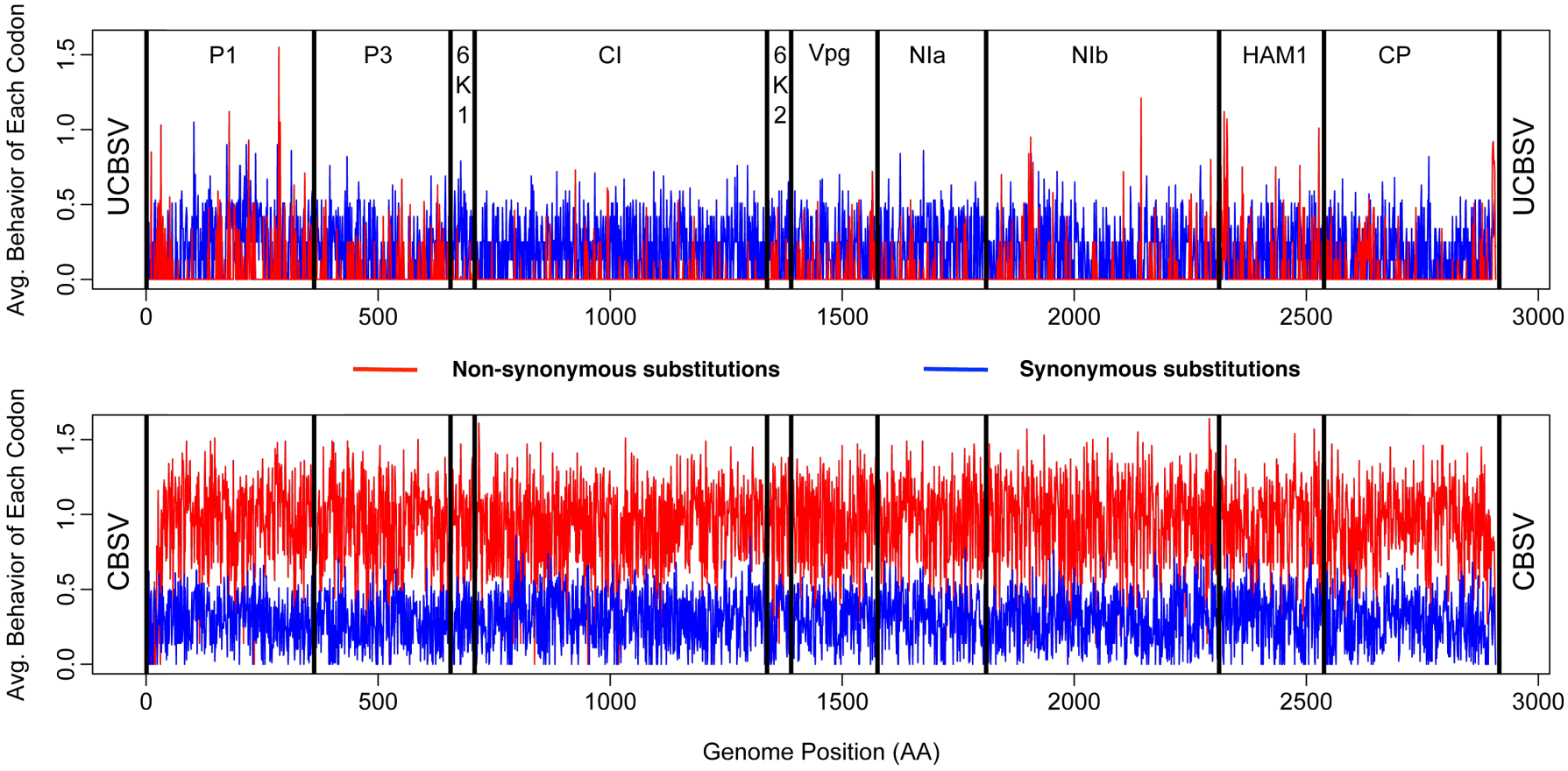


Figure 3. Species tree generated from SVD Quartets using the whole genome sequences. Colors at the tips are based on country of origin. Branches with mixed colors indicate a clade that contains samples with mixed country of origin. For example, the ancestral branch of UCBSV TZ Tan 23 KR108839 and UCBSV UG MI B3 FJ039520 is colored red and orange to indicate a clade with sampled with mixed country of origin.

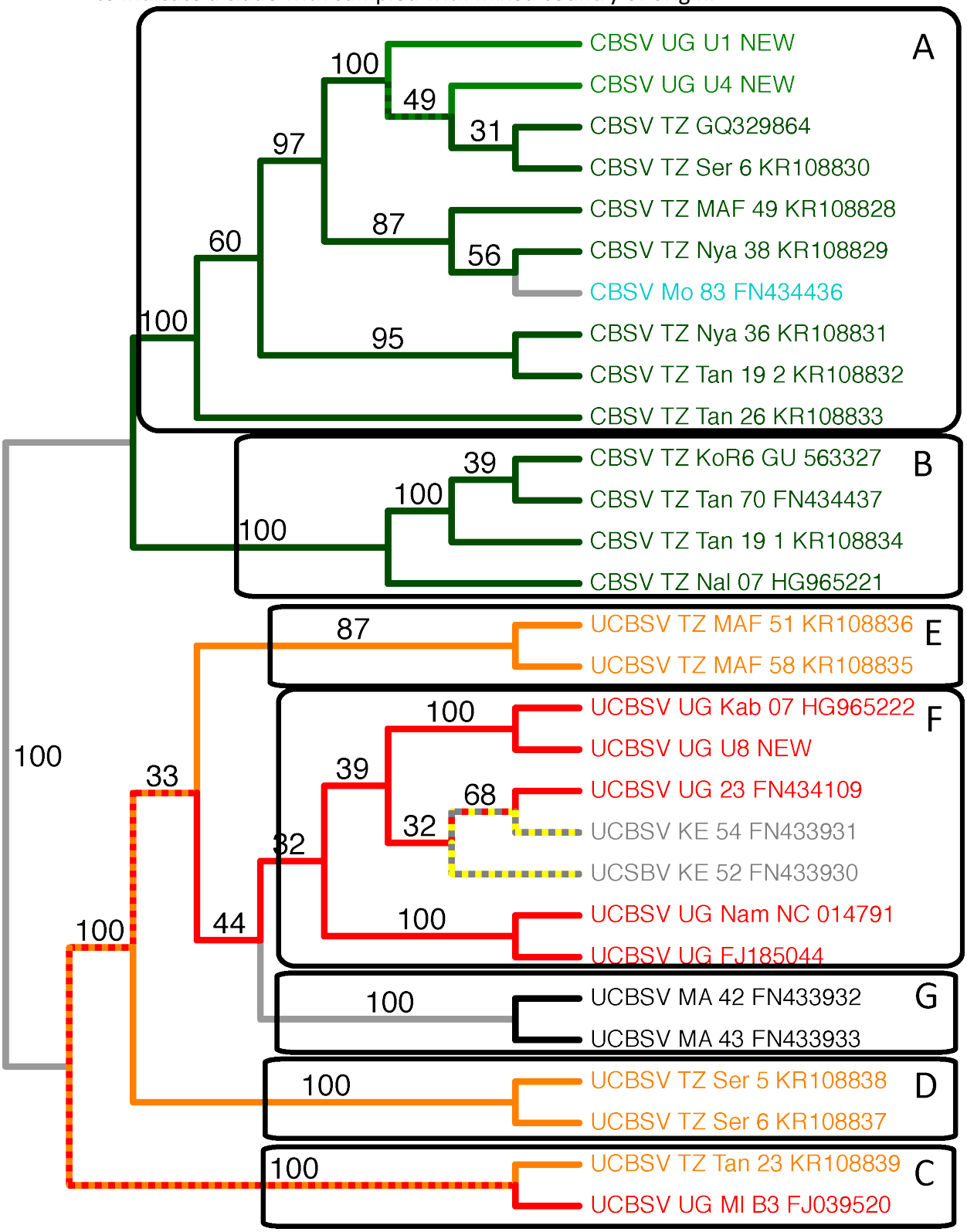


Figure 4. Computed SVD Score with the split defined by CBSV vs. UCBSV across the genome in windows of 500 bp, sliding in increments of 100 bp, and resulting SVD Scores plotted across the genome. Boundaries between genes are marked with vertical lines to further characterize the CBSV and UCBSV genomes. Rates of molecular evolution were estimated using CODEML implemented in PAML (Phylogenetic Analysis by Maximum Likelihood)[54]. The results are shown for each gene and D represents the difference in likelihoods from the null hypothesis (CBSV and UCBSV have equal rates) and the alternative hypothesis (CBSV and UCBSV have different rates).