A peer-reviewed version of this preprint was published in PeerJ on 12 April 2017.

View the peer-reviewed version (peerj.com/articles/3165), which is the preferred citable publication unless you specifically need to cite this preprint.

Ho ES, Newsom-Stewart CM, Diarra L, McCauley CS. 2017. gb4gv: a genome browser for *geminivirus*. PeerJ 5:e3165
https://doi.org/10.7717/peerj.3165

gb4gv: A Genome Browser for Geminivirus

Eric S Ho Corresp., [1] , **Catherine M Newsom-Stewart** [2] , **Lysa Diarra** [2] , **Caroline S McCauley** [2]

[1] Department of Biology, Department of Computer Science, Lafayette College, Easton, Pennsylvania, United States
[2] Department of Biology, Lafayette College, Easton, Pennsylvania, United States

Corresponding Author: Eric S Ho
Email address: hoe@lafayette.edu

Background: Geminivirus (family *Geminiviridae*) is a prevalent plant virus that imperils agriculture globally, causing serious damage to the livelihood of farmers, particularly in developing countries. The virus evolves rapidly, attributing to its single-stranded genome propensity, resulting in worldwide circulation of diverse and viable genomes. Genomics is a prominent approach taken by researchers in elucidating the infectious mechanism of the virus. Currently, NCBI Viral Genome website is a popular repository of viral genomes that conveniently provides researchers a centralized data source of genomic information. However, unlike the genome of living organisms, viral genomes most often maintain peculiar characteristics that fit into no single genome architecture. By imposing a unified annotation scheme on the myriad of viral genomes may downplay their hallmark features. For example, virion of Begomovirus prevailing in America encapsulates two similar-sized circular genomes and both are required to maintain virulence. But, the two bipartite genomes are kept separately in NCBI with no explicit association in linking them. Thus, our goal is to build a comprehensive Geminivirus genomics database, namely gb4gv, that not only preserves genomic characteristics of the virus, but also supplements biologically relevant annotations that help to interrogate this virus e.g. the targeted host, putative iterons, siRNA targets etc. **Methods:** We have employed manual and automatic methods to curate 508 genomes from four major genera of Geminiviruses, and 161 associated satellites obtained from NCBI RefSeq and PubMed databases. **Results:** These data are available for free access without registration from our website. Besides genomic content, our website provides visualization capability inherited from UCSC Genome Browser. **Discussion:** With the genomic information readily accessible, we hope that our database will inspire researchers in gaining better understanding about this virus, resulting in insightful strategies to conquer the devastation inflicted agriculture. **Availability and Implementation:** Database URL: http://gb4gv.lafayette.edu .

gb4gv: A Genome Browser for Geminivirus

Eric S. Ho[1,2], Catherine M. Newsom-Stewart[1], Lysa Diarra[1], and Caroline S. McCauley[1]

[1]Department of Biology, [2]Department of Computer Science, Lafayette College, Easton, Pennsylvania, 18042, United States.

Corresponding author: Eric S. Ho

Email address: hoe@lafayette.edu

Abstract

Background: Geminivirus (family *Geminiviridae*) is a prevalent plant virus that imperils agriculture globally, causing serious damage to the livelihood of farmers, particularly in developing countries. The virus evolves rapidly, attributing to its single-stranded genome propensity, resulting in worldwide circulation of diverse and viable genomes. Genomics is a prominent approach taken by researchers in elucidating the infectious mechanism of the virus. Currently, NCBI Viral Genome website is a popular repository of viral genomes that conveniently provides researchers a centralized data source of genomic information. However, unlike the genome of living organisms, viral genomes most often maintain peculiar characteristics that fit into no single genome architecture. By imposing a unified annotation scheme on the myriad of viral genomes may downplay their hallmark features. For example, virion of Begomovirus prevailing in America encapsulates two similar-sized circular genomes and both are required to maintain virulence. But, the two bipartite genomes are kept separately in NCBI with no explicit association in linking them. Thus, our goal is to build a comprehensive Geminivirus genomics database, namely gb4gv, that not only preserves genomic characteristics of the virus, but also supplements biologically

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

relevant annotations that help to interrogate this virus e.g. the targeted host, putative iterons, siRNA targets etc.

Methods: We have employed manual and automatic methods to curate 508 genomes from four major genera of Geminiviruses, and 161 associated satellites obtained from NCBI RefSeq and PubMed databases.

Results: These data are available for free access without registration from our website. Besides genomic content, our website provides visualization capability inherited from UCSC Genome Browser.

Discussion: With the genomic information readily accessible, we hope that our database will inspire researchers in gaining better understanding about this virus, resulting in insightful strategies to conquer the devastation inflicted agriculture.

Availability and Implementation: Database URL: http://gb4gv.lafayette.edu.

Subjects Bioinformatics, Databases

Keywords Geminiviridae, Geminivirus, Begomovirus, Mastrevirus, Curtovirus, alphasatellite, betasatellite, UCSC Genome Browser

Introduction

Geminiviruses (family *Geminiviridae*) have emerged as one of the most prevalent and detrimental plant viruses in agriculture worldwide over the last 20 years. In terms of number of species, they have become the largest group of plant viruses to exist today. This is a significant threat both socially and economically as Geminiviruses are the most destructive pathogens for staple crops in subsistence agriculture like beans, cotton, maize,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

46 sweet potato and tomato. The economic impact of Geminivirus infection can be seen across
47 the globe: Pakistan lost an estimated $5 billion for infection in cotton between 1992-1997,
48 India lost an estimated $300 million for infection in grain legumes in 1992, and Florida lost
49 approximately $140 million for infection in tomato in 1999 (Varma & Malathi 2003).
50 The spread of Geminiviruses and the severity of their infections have been
51 increasingly climbing due to their virulence to infect multiple hosts through their insect
52 vector *Bemisia tabaci* (whitefly) (Ghanim et al. 2001). Geminiviruses often work as part of a
53 disease complex: a mixture of viral species, isolates and DNA satellites. Moreover, they are
54 able to undergo mutation, recombination and reassortment both frequently and rapidly.
55 Together, these factors increase the diversity and capabilities of the family, allowing them
56 to invade new hosts and new environments without complication. In order to prevent
57 Geminiviruses from becoming even more of a threat to our growing human population, it is
58 critical that scientists are able to better understand the genomic sequences of these viruses.
59 Geminiviruses rely heavily on their host's cellular machinery so having a greater
60 knowledge of their genetic makeup will allow scientists to formulate biotechnological
61 means to help plants fight their attackers successfully.
62 Geminiviruses comprise a family of plant viruses that exist in the form of twinned
63 icosahedral particles holding small, circular, single stranded deoxyribonucleic acid (ssDNA)
64 genomes. The ssDNA genome structure enables it to evolve at high rate comparable to RNA
65 viruses (Duffy et al. 2008). The viral genome encodes only 5-7 proteins, making
66 Geminiviruses one of the smallest virus types known to scientists today. Within
67 *Geminiviridae*, seven genera have been discovered at present: Mastrevirus, Curtovirus,
68 Becurtovirus, Eragrovirus, Topocuvirus, Turncurtovirus and, the most prevalent,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

69 Begomovirus. Depending on the genera, the viral genome comprises of either one
70 (monopartite) or two (bipartite) DNA components. Monopartite genomes consist of a DNA-
71 A that is often associated with an alphasatellite or a betasatellite genome, while bipartite
72 genomes consist of separated DNA-A and DNA-B components of similar size.
73 National Center for Biotechnology Information (NCBI) designates a separate
74 website to host viral genomes (NCBI Viral Genomes). Its collection includes almost all
75 known viruses in the world, making it one of the most popular resources for studying viral
76 genomics. Viral genomes are formatted in standard GenBank record (GenBank Record)
77 exactly like other living organisms. However, genome architectures of viruses exhibit
78 significant difference from living organisms. For instance, virion of bipartite Begomovirus
79 encapsulates two circular genomes in which the two genomes synergize to retain virulence.
80 But such critical association between the two main genomes is often missing from NCBI
81 Viral Genome database. Moreover, vital information about the virus such as location where
82 it was found, targeted hosts, etc. are not searchable attributes, limiting the utility of the
83 database. These are the reasons that we have undertaken this project in providing
84 researchers a comprehensive, up-to-date, and integrated environment at their fingertips.
85 The database we built rests on the software architecture of UCSC Genome Browser website
86 (Kent et al. 2002; UCSC Genome Browser) as such we named our database gb4gv, which
87 stands for Genome Browser for Geminivirus. For clarity, we reserve "UCSC Genome
88 Browser" to refer to the website itself (UCSC Genome Browser), and "Genome Browser" to
89 mean the software that supports the website. Genome Browser was chosen because of its
90 versatility in visualizing genomes, richness in built-in functions, flexibility in incorporating
91 annotations, and software robustness in handling large volume of requests - 872,000

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

requests per day on average (UCSC GB Statistics). Although Genome Browser offers these benefits, its original design gears mainly toward eukaryotes. In order to unleash the power of Genome Browser, we have made substantial effort in modeling Geminivirus genomes into a structure that can take full advantage of its functionalities. gb4gv can be accessed, without registration requirements, from here: http://gb4gv.lafayette.edu. Users can make use of the built-in functions provided through our website to download genomes or sequences of interested regions freely.

Materials & Methods

Compilation of Geminivirus Genomes

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

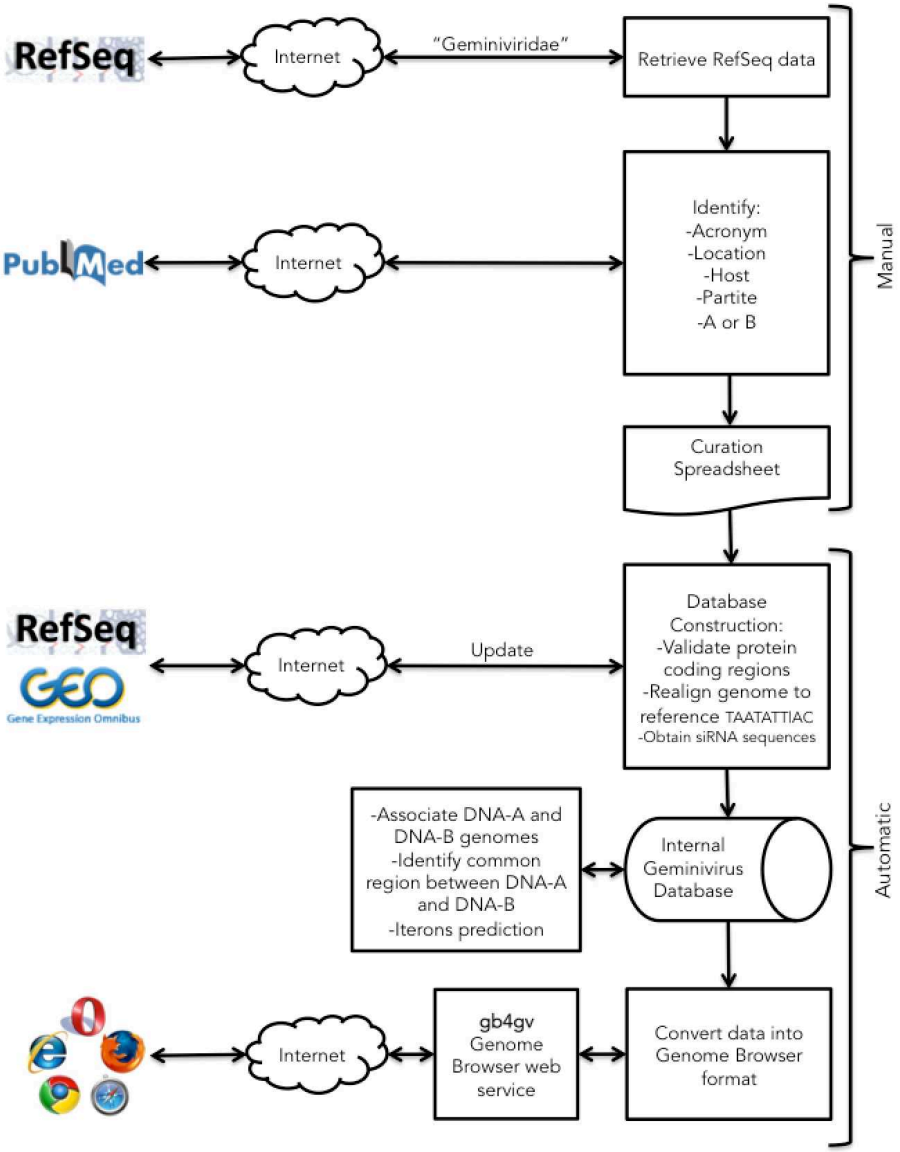


Figure 1 Project workflow of the semi-automatic annotation process. Our workflow begins with a manual process to identify information generally not documented in the GenBank record such as the acronym of the virus, location, infected host, monopartite or bipartite genome, and genomes association for bipartite virus. This information is passed to a downstream automatic process that integrates them with other sources. The automatic procedure parses GenBank entries from RefSeq database for genomic information of Geminiviruses including the accession numbers, genomic sequences, genes, viral proteins, and taxonomy ID. siRNAs from host plants that fight against viral infection were obtained from NCBI GEO database. In the last step, Geminivirus information is formatted into UCSC Genome Browser format.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

113 Figure 1 above summarizes the semi-automatic annotation workflow adopted by this
114 project. The primary source of our data originated from NCBI RefSeq database (NCBI
115 RefSeq) because redundant genomes were purged. But we also cross-referenced our data
116 with the ICTV Master Species List 2015 v1 obtained from ICTV (ICTV 2015). We had
117 identified 700 RefSeq entries that comprised of 529 distinct Geminiviruses. Note that DNA-
118 A and DNA-B of a bipartite Begomovirus are kept in separate entries in RefSeq.
119 Begomovirus occupies the largest genus of the family, followed by Mastrevirus, and
120 Curtovirus. Other genera were found sporadically including two Becurtovirus, one
121 Topocuvirus, one Eragrovirus, and one Turncurtovirus, while genera of ten entries remain
122 unknown. Here we had decided to incorporate only genera that represent major
123 *Geminiviridae* members i.e. Begomovirus, Mastrevirus, and Curtovirus into gb4gv. As a
124 result, genomes of 514 Geminiviruses representing 97% of the *Geminiviridae* found in NCBI
125 were considered for further review. We will regularly assess the need to include other
126 minor genera into our database if more samples from them are discovered in the future.

127 Besides the main genomes, ancillary alphasatellites and betasatellites are often isolated
128 together with monopartite Begomoviruses (Xie et al. 2010) and they are found to play
129 essential roles in boosting host's symptoms and viral movement (Briddon et al. 2001;
130 Saunders et al. 2004; Zhou et al. 2003). We had identified and reviewed 66 and 105
131 alphasatellite and betasatellite genomes, respectively, from NCBI.

132 Meta information or attributes such as the geographical location of the virus are
133 important to understand the virulence of the virus but it is not always available in genome
134 database. Therefore we manually searched for additional information about these viruses
135 from existing literature. In particular, we focused on identifying or reconfirming the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

136 location where they were collected, the hosts they infected, their acronyms, monopartite or
137 bipartite genome, and the counterpart genome in case of bipartite. Importantly, we have
138 made these attributes searchable in our database.

139 Following the manual process is the automatic annotation process. In this step, NCBI
140 RefSeq entries belonging to Geminivirus were parsed to ensure that each entry satisfies the
141 following two criteria:

142 1. Every Geminivirus including satellite genome must possess the iconic structurally
143 conserved element (SCE), which is the genomic landmark of Geminiviruses including
144 satellites. The canonical structure of the SCE is `TAATATT|AC`, where "|" stands for the
145 cleavage site targeted by the viral replication protein in the initial step of DNA
146 replication (Gutierrez 1999; Jeske et al. 2001; Pilartz & Jeske 2003). The prevalent SCE
147 sequence of alphasatellite is `TAGTATT|AC`, which varies slightly from the canonical SCE
148 sequence. Nonetheless, owing to either DNA sequencing errors or random mutations,
149 the 5' side of the SCE of some viruses may deviate slightly (less than one nucleotide)
150 from the canonical form from above. To accommodate such minutiae, we tolerated
151 entries with up to one mismatch from `TAATATT`. Genomes failed to meet this criterion
152 were excluded from gb4gv.
153 2. Besides genomes, gb4gv also keeps individual viral proteins if they satisfy our quality
154 checking. The coding region (CDS) of a gene defined in the RefSeq entry must be
155 translated exactly into the stated peptide in the RefSeq entry. Genes failed this criterion
156 were excluded from our database. But the genomes containing mistaken CDS were still
157 kept in the database.
158

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

159 Through our tandem manual and automatic annotations, 6 out of 514 RefSeq entries of
160 Geminivirus failed the validation process stated above, resulting in 508 genomes being
161 selected into our database. For satellite genomes, 7 out of 66 alphasatellites and 3 out of
162 105 betasatellites failed our validation. Table 1 below categorizes all the accepted genomes
163 in our database by genus, number of genomes per virus, and geographical origin. The
164 aforementioned annotation information can be downloaded from our website in tab-
165 separated format (http://gb4gv.lafayette.edu/downloads.html).

166 Table 1. A summary of genomes stored in gb4gv. The numbers inside the parentheses
167 denote the numbers of genomes. The lower part of the table categorizes Begomoviruses
168 further by world and the number of genomes per virus.

Geminiviridae (508)				Satellite (161)	
		Begomovirus (470)			
Curtovirus	Mastrevirus	DNA-A	DNA-B	Alphasatellite	Betasatellite
5	34	338	132	59	102

169

Begomovirus (338)							
Old World (216)			New World (119)			Unknown World (2)	
Monopartite	Bipartite	Unknown	Monopartite	Bipartite	Unknown	Monopartite	Bipartite
100	44	72	12	95	13	1	1

170

171 *Small Interference RNAs*

172 A key aspect of gb4gv is to inspire researchers to formulate insightful strategies that can be
173 used to eradicate the propagation of Geminiviruses. Therefore, studying the immune
174 response launched by infected plant is a promising research direction. Thus, we had
175 downloaded datasets from two small interference RNAs studies from NCBI GEO database

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

176 (Gene Expression Omnibus): GSM425427, and GSE26368. siRNA sequences were mapped
177 to the genomes of Begomovirus and betasatellite through a customized Python script.
178 Mapping tolerated up to two mismatches in internal positions without gaps. A Genome
179 Browser annotation track is designated for each sample, which can be found under
180 "Mapping and Sequencing" section of each virus. In Begomovirus or betasatellite, six siRNA
181 tracks were configured.
182
183 *Standardization of Circular Genomes*
184 Like many other genomic databases such as NCBI RefSeq and UCSC Genome Browser,
185 circular genomes are linearized. Instead of opening the circular genome at arbitrary sites,
186 circular genomes were opened at the biological cleavage site at the SCE. A benefit of
187 standardizing the opening site is to facilitate syntenic analysis (to be discussed in *Multiz*
188 section below). Under the standardized linearization scheme, a genome always begins with
189 'AC' and terminates with 'TAATATT' at the 5' and 3' termini, respectively. Thereby we
190 standardized all genomes obtained from RefSeq. Genomes not conforming to this standard
191 were shifted until they met the above criterion. Out of 669 accepted genomes in gb4gv,
192 surprisingly, 112 (17%) genomes required this adjustment.
193
194 *Data Models*
195 Genome Browser was originally designed to visualize mammalian genomes (Kent et al.
196 2002). It was later enhanced to host non-mammalian animals e.g. *C. elegans*, and then
197 eukaryotic protozoan such as yeast. Ebola genome is the first and remains to be the only
198 viral genome available in UCSC Genome Browser at present. This historical background

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

199 reveals that the data model of Genome Browser is geared toward the display of
200 chromosomes of a species. Such data model serves well with living organisms but it poses
201 two challenges in configuring Genome Browser for Geminivirus genomes:

202 1. The genus Begomovirus is known to be diverse (Brown et al. 2015) with over 500
203 species being identified by us. If we were to coerce the existing data model to
204 Begomovirus, 500 databases are needed, leading to a huge species tree on the front
205 page, hampering website performance, and prohibiting data browsing. To circumvent
206 this, we modeled each viral genus as an organismal species, and the array of viral
207 species of a genus as chromosomes of an organism. Based on this workaround, gb4gv
208 consists of five databases (a database per genus including one for each satellite
209 although, in biological terms, satellite is not considered a genus): Begomovirus,
210 Mastrevirus, Curtovirus, Alphasatellite, and Betasatellite.
211 2. A special configuration is needed to establish the association between the bipartite
212 genomes (DNA-A and DNA-B) of Begomovirus. In gb4gv, DNA-A and DNA-B were
213 treated as two separate chromosomes. The coupling of DNA-A and DNA-B genomes of a
214 bipartite Begomovirus can only be achieved manually as their accession numbers
215 reflect no information about their relationship. In order to facilitate users to associate
216 them easily, a viral species in our database is uniquely referenced by an acronym, e.g.
217 AbMBV is the reference of Abutilon mosaic Brazil virus. But the two genomes of a
218 bipartite Begomovirus will become indistinguishable under this scheme. Thus, we suffix
219 the acronym of a bipartite virus by ".A" and ".B". E.g. the DNA-A and DNA-B genomes of
220 virus AbMBV can be found effortlessly through AbMBV.A and AbMBV.B, respectively. An
221 advantage of using acronym as key to access a virus is to release the burden of users to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

222 pull up the accession number of the virus as most people can remember the acronym
223 rather than the arbitrary accession number.
224
225 *Common Region Identification in Bipartite Begomoviruses*
226 The bipartite genomes of a Begomovirus share a highly similar, non-coding segment
227 flanking the SCE "TAATATTAC". This segment is colloquially named the common region
228 (CR). CR serves a crucial role in viral DNA replication. Studies had shown that the 5' side of
229 the CR contains replication protein binding sites (Orozco & Hanley-Bowdoin 1996). Thus,
230 CR harbors vital regulatory signals that influence the replication and the coupling of the
231 bipartite genomes for Begomovirus. Understanding viral replication is fundamental to
232 combat viral infection. Thus, we undertook the task to predict CRs in bipartite
233 Begomoviruses. Based on the manual annotation we did, DNA-A and DNA-B genomes of a
234 Begomovirus were paired up. We extracted the non-coding region, also known as the long
235 intergenic region (LIR), between REP and CP genes in DNA-A or between NSP and MP
236 genes in DNA-B. In the next step, we further reduced the LIR into an 809-bp segment,
237 which consisted of a 400-bp segment upstream and downstream of the SCE from the DNA-
238 A and DNA-B genomes. In Figure 2 below, two 809-bp segments were aligned by MUSCLE
239 (Edgar 2004) as shown:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

```
CLUSTAL W (1.81) multiple sequence alignment

NC_011583       GCTGACCGGGATGGGGAT-ATGAGGTCGAA-GAATCGATGG--TTGGTACAATTGTACTT
NC_011584       -CAAATCGGCGAACAAATAAAAAAGTCGAATGAGGTGAAGGGATTGAAACG-----ACTT
                 *  * ***       ** *  * ****** **   ** **  ***  **      ****

NC_011583       GCCCTCGAACTGAATGAGGGCATGCAGATGAGGTTCCCCATTTTCATGGAGTTCTC----
NC_011584       ACGGAAGCACCG-ATGAAGCAGTCTGGAGTGAATTCCAGATATAATTGGAGAAAACAAAG
                 *    * ** * **** *   *   **     ****  ** *   *****    *

NC_011583       -TGCAGATCTTGATGA---------ACAATTTATTTGTTGGGGTTTGG-------AGTTG
NC_011584       AAATAAAAGTTAACGAAATAAAAGTATAACTTAT------GGGTATAGAAAGGAAAGTGA
                    * *  ** * **         * ** ****      **** * *       ***

NC_011583       TCGGATTTGATCCAATGCCTCCTCTTTGGATAGAGAGCATTTGGG-----ATAT----GT
NC_011584       GCAGATGTTATGC---GCCGTGTCGTTAAATGAGATGTTATTGGGTGTTTATATAGGCGT
                 * *** * ** *   ***   ** **  **     *   *****     ****    **

NC_011583       TAGGAAATAGTTTTTGGCTTTGATGCTAAAACGACCAGCCCTTGGCATTTTCGCTGTCGT
NC_011584       TAATAAGCAACACGTGGTAGAGATAGAAAGAAGAAAGA-----GGCG-------------
                **  **  *     ***    ***   ** * **         ***

NC_011583       ATAGCAATCGGGGGGCACTCAAAGTCTGTAGCAATCGGGGGAAAGGGGGGCAATTTATAT
NC_011584       AGAGCATTCGGGGGGCACTCAAAGTCTGTAGCAATCGGGGGAAAGGGGGGCAATTTATAT
                * **** ******************************************************

NC_011583       GATGCCCCCTAAATGGCATTTATGTAATATCCTCATTGAATTTGAAATTCAAACGTGGAA
NC_011584       GATGCCCCCTAAATGGCATTTATGTAATATCCTCAATGAATTTGAAATTCAAACGTGGAA
                ************************************ ************************

NC_011583       AGCGGCCATCCGTATAATATTACCGGATGGCCGCGCCCGAAAAAGCAGGTGGACCCCACA
NC_011584       AGCGGCCATCCGTATAATATTACCGGATGGCCGCGCCCGAAAAAGCAGGTGGACCCCAC-
                ************************************************************

NC_011583       GGATGGCCGCGCCCGTGAAAGAAAGTGGTCCCTGCGCACTTGTTTTGGTCGGCCAGTCAT
NC_011584       -----------------------AATGGCCCCCACGCACTAAGTAATGTCAGCCAATCAT
                                      * *** ***  ******   *   *** **** ****

NC_011583       ATTCACGCGTGAAAGGC----TAGATATATGTTG--------TTTGTCTTTATAGAC---
NC_011584       GTTCAAGACTGGAAGACGCGGTAGTTACGCATTGATGAGTAAGTGGTCCCTACGCACTAA
                 **** *  ** *** *    *** **    ***          * ***  **   **

NC_011583       --------------TTCGTCGCGAAGTAGTGGAGCGCGTCAACATGTGGGATCCATTGTT
NC_011584       TGTTGACAGGCAATTTGATTGCTATGT-GTGTATCATATTTATATAGGTGTGCTACTGGT
                               **  * ** * ** *** * *   *  * **  * *  * * ** *

NC_011583       GAAC------------------GACTTTCCCGAAACCG---------------------T
NC_011584       TAATCTAAAGTTAGGTGATGGGGCCTATCATAAAAACGCAATACATAGGTACGTATGTAC
                 **                   * ** **   *** **

NC_011583       TCACGGTTTCCGTTCTATGCTTGCTGTTAAAT-ACCTGTTACATCTGGAACAGGAATACG
NC_011584       ATATTGATTATATTTTATG-TTGCGGATATATGAGCCGCCACGTGTATAATGGATAT---
                  *  * **   ** **** **** * ** ** * * *  ** * *  **  *  **

NC_011583       ACCGCGGTACTGTCGGGGCTGAGTATATACGGGATCTAATAGGGGTTCTACGGTGTAAGA
NC_011584       ------GGAATGTC----CTATAAATATTTGGCATGTCCCC---GTTCGTTAATGCAAGA
                      * * ****    **    ****  ** ** *       ****     ** ****

NC_011583       GTTATGTCGAAGCGACCAGGAGATATAATAATCTCAACACCCGTATCCAAGGTGCGGAGG
NC_011584       TGTATTCTGTTTACAGACGTGGGTATAAGACT--------CCGTAT-----------AGG
                  ***   *     *   *  * ***** * *        ******           ***

NC_011583       AGGCTGAACTTC
NC_011584       AG--------TC
                **        **
```

Figure 2. Identification of common region (CR) shared between DNA-A and DNA-B of bipartite Begomoviruses. Sequence alignment of two 809-bp segments located in the LIR of the Old World bipartite East African cassava mosaic Kenya virus (EACMKV). The invariant SCE are highlighted in red. The inverted repeats constituted the stem of the hairpin structure are highlighted in blue. The two underlined regions indicate the 5' and 3' termini of the common region determined by our method of using a 20-bp sliding window.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

248 In this example, the two segments were extracted from DNA-A (NC_011583) and
249 DNA-B (NC_011584) of the Old World bipartite East African cassava mosaic Kenya virus
250 (EACMKV) and they were aligned. A 20-bp sliding window was used to scan the alignment
251 base-by-base bilaterally starting from the SCE. Scanning halts when the percentage of
252 sequence identity within the window drops below 80%, an adjustable parameter. The
253 halting locations (the underlined regions in Figure 2) are taken as the 5' and 3' termini of
254 the common region.

255 The average size of a CR was found to be 212 bps (including the SCE) in which the 5'
256 arm, the left segment of the SCE, is usually longer than the 3' arm with an average size of
257 150 bps. The longest CR is 417-419 bps long that belongs to Indian cassava mosaic virus
258 (NC_001932/NC_001933). Whereas Abutilon mosaic Brazil virus (NC_016574/NC_016577)
259 was found to possess the shortest CR, which is 63-67 bps long. Also note that two
260 approximately 10 bps segments juxtaposing the SCE constitute the stem part of the hairpin
261 structure (Figure 2).

262

263 *Putative Iterons and TATA Box*

264 One of the cis-regulatory signals harboring in CR is iterative element, also known as iteron
265 (Arguello-Astorga et al. 1994; Sanz-Burgos & Gutierrez 1998). A distinct feature of iteron is
266 the presence of direct or inverted sequence repeats. The following criteria were applied to
267 predict them in viral genomes:

268 1. They are located in the 5' side of the LIR i.e. from the beginning of the first gene on the
269 complementary strand up to, but excluding, the SCE
270 2. Minimum length of an iteron is 9 bps. There is no restriction on maximum length

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

3. The pair of repeats identified differ by at most one base
4. Repeats could be direct or inverted
5. Its location is no more than 100 bps from a putative TATA-box, if any

As TATA-box and iterons work cooperatively to regulate replication, we also identified putative TATA-box sequence. The consensus sequence of a TATA-box is defined as "`TATA`", followed by any number of "`TA`" or "`AA`" repeat (Bernard et al. 2010; Patikoglou et al. 1999). We developed a Python script to scan for iterons and TATA-box sequences in every Geminivirus genome. Based on the above criteria, 21,298 iterons and TATA-box sequences were predicted from 669 genomes. Results can be visualized in gb4gv by activating the 'Iterons and TATA" annotation track.

Multiz Track

Genomes of various species within a genus share similarities and differences. Since we have standardized the opening site of the viral circular genomes, a genus-wide syntenic analysis becomes possible. Such comparative view helps to uncover conserved and diverse genomic regions among species. We used the threaded blockset aligner (TBA) (Blanchette et al. 2004) to generate a dynamic multiple sequence alignments of all species from a genus. Unlike other multiple sequence alignment programs of which a sequence from the sample is dedicated to be the reference of the alignment, TBA produces a multiple sequence alignment dynamically based upon the genome being selected for viewing in the Genome Browser. This unique feature enables gb4gv to generate a graphical representation

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

regarding genome conservation among different strains with respect to the current queried genome.

TBA requires two mandatory inputs: a set of genomic sequences, and a phylogenetic tree defining the evolutionary relationship of the input genomes. We used multiple sequence alignment program MUSCLE (Edgar 2004) to build phylogenetic trees, followed by maximum likelihood tree building PHYLM (Felsenstein 2005) equipped in MEGA7 (Kumar et al. 2016). The output phylogenetic tree was in NEWICK format. Based on the genomic sequences and the phylogenetic tree, TBA generated the threaded blockset alignment. The alignment was loaded to a MySQL database referenced by Genome Browser.

UniProt/SwissProt Annotations

Protein domain information was overlaid on viral proteins in gb4gv. Reviewed Swiss-Prot annotations were downloaded from UniProt website in XML format (UniProtKB). Viral taxonomy IDs served as the key to retrieve protein domain information from the Swiss-Prot annotations. Sequence of the protein domains identified in the search process was mapped to the genomes by BLAT (Kent 2002).

Genome Browser

Version 334 of the Genome Browser was used to build gb4gv. The software was downloaded from the UCSC Genome Browser website (UCSC Admin) and installed in our 24-core Linux server running on Centos OS 6.8, Apache 2.2, and MySQL server 5.5.50.

Results

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

316 The web interface of gb4gv is organized in a hierarchy consisting of three levels. The
317 highest level presents all the genera of *Geminiviridae* maintained in gb4gv (Figure 3A). The
318 middle level displays information about the genome of an individual viral species and
319 corresponding annotation tracks (Figure 3B). Detailed information about a particular
320 annotation e.g. a gene, a protein or a specific genomic sequence, is presented at the lowest
321 level (Figure 3C)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

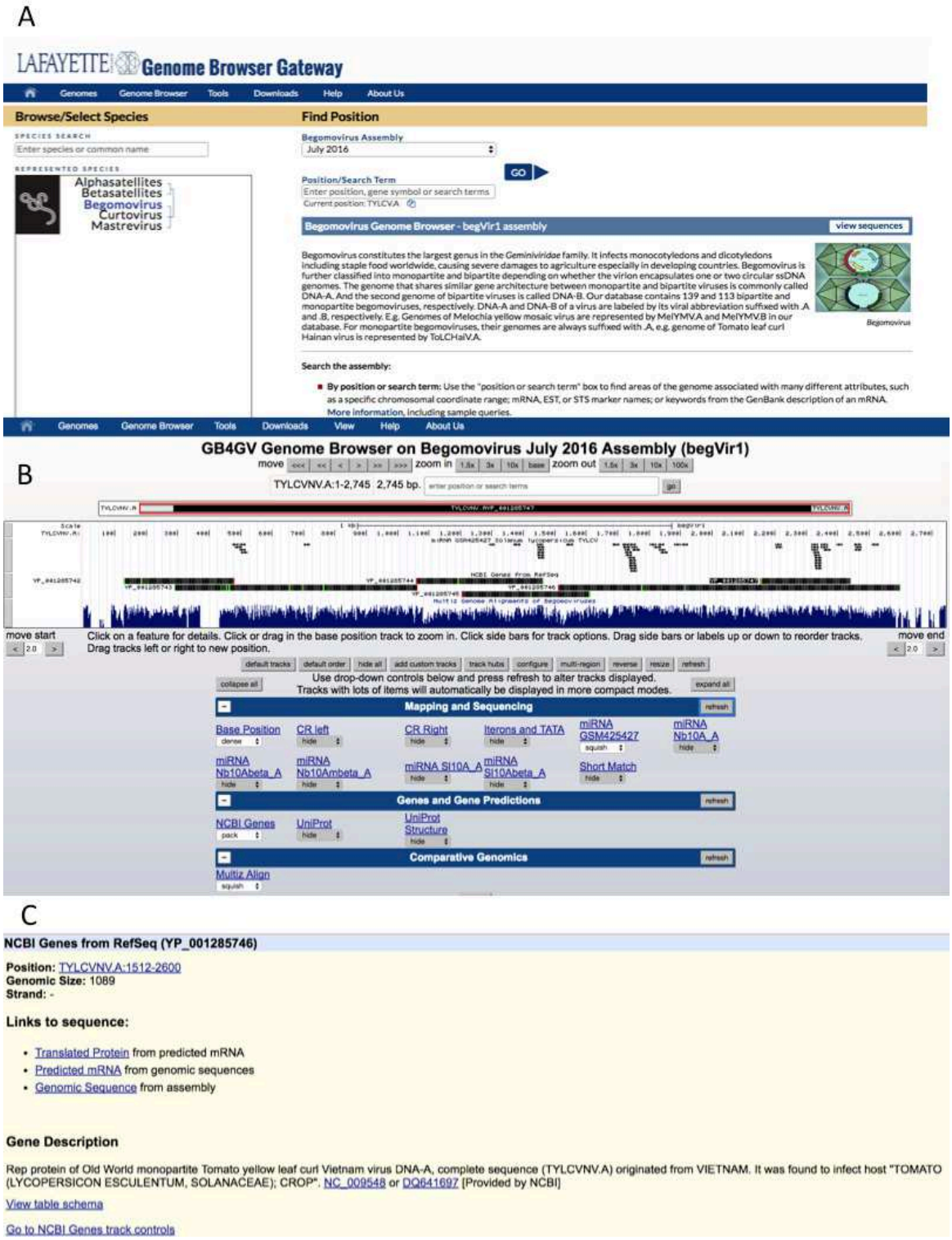


Figure 3. Web interface of gb4gv. (A) The home page of gb4gv. The evolutionary tree on the left side of the page shows the available viruses including the three genera of Geminiviruses and two satellites. Users can view a particular genus or satellite by clicking on the viral or satellite name in the evolutionary tree. Users can also make use of the "Species Search" box (above the tree) to look up for a particular virus by keywords.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

328 Additionally, users can enter keywords in the "Position/Search Term" box search for a
329 particular virus and/or gene. Click the blue GO button to navigate into the genomic
330 information of a particular viral strain. (B) The page at the middle level provides various
331 annotation information about the selected genome in which they are organized in tracks.
332 (C) Information of a protein-coding gene. It tells the genomic location of the gene, size, and
333 strand that codes for the protein. In addition, there is a short description about the current
334 gene including the name of the protein, whether the virus is a New World or an Old World
335 virus, the full name of the virus and its acronym, the RefSeq and GenBank accession
336 numbers with hyperlink linked to the corresponding GenBank entry in NCBI website.
337

338 In the following subsections, we will highlight the unique features offered by gb4gv
339 that are helpful in studying the genomics of Geminivirus. While the software architecture of
340 gb4gv is based on Genome Browser, the operations of our website directly adopt from the
341 built-in functions provided by Genome Browser. We will not discuss the data models and
342 functionalities of Genome Browser in details. For readers who are interested in learning
343 more about Genome Browser, we recommend that they consult the online User Guide
344 (UCSC Genome Browser User Guide).
345

346 *Search by Acronym, Accession number, and Attributes*

347 To our best knowledge, there is no database that allows users to search for Geminivirus
348 genomes or proteins by acronym, host name, geographical location, monopartite, bipartite,
349 Old world, New world, or combinations thereof. For instance, a search for monopartite
350 Begomoviruses that infect Okra by the query "monparite begomovirus okra" against NCBI
351 RefSeq database returned only two entries: NC_005954 and NC_005051 and both of them
352 belong to satellite genomes. In fact, four monopartite Begomoviruses are known to infect
353 Okra according to gb4gv: OLCCV (NC_014745), OYCrV (NC_008377), OYVMV (NC_004673),
354 and OkLCuV (NC_013017). The main reason is because NCBI's query matches only words in
355 the description of GenBank entries. Our augmented search capability will help researchers

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

356 in identifying a regime of viruses that share certain attributes handily. gb4gv achieves this
357 by making the above viral attributes searchable in our database in conjunction with the
358 keyword searching capability provided by Genome Browser. Table 2 below summarizes the
359 searchable attributes supported by gb4gv.

360 Table 2: Searchable attributes in gb4gv.

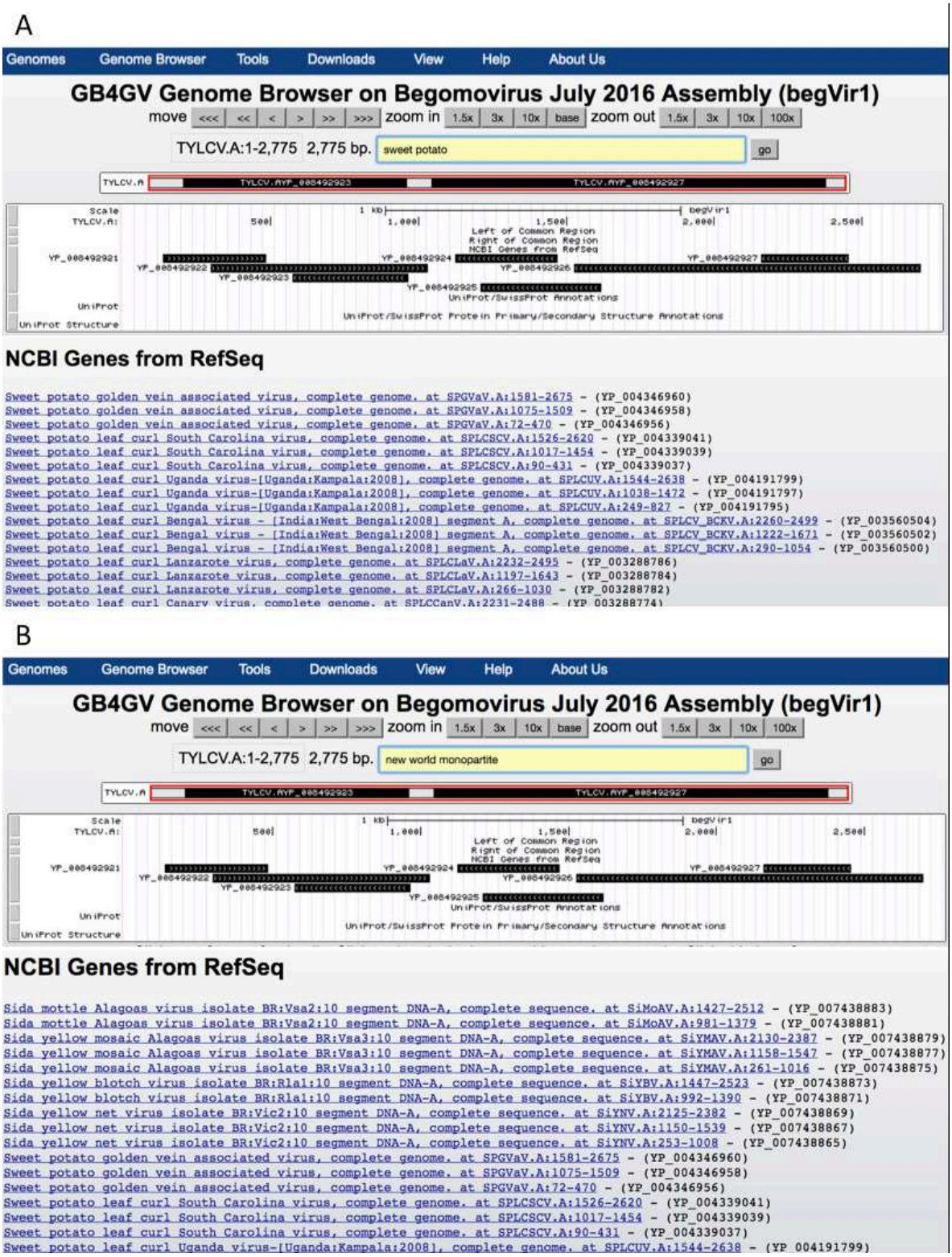
Attribute	Description	Example
World	Geminiviruses are commonly categorized into "Old World" and "New World" according to the geographical location they were found. This attribute must be either "Old World" or "New World"	`old world`
Number of main genomes	It must be either monopartite or bipartite	`bipartite`
Acronym of the virus	For Begomovirus, it could be suffixed optionally by ".A" or ".B" to indicate DNA-A or DNA-B of the bipartite genome, respectively.	`OMoV.A`
Host	Name of the host infected by the virus	`Okra`
Country	The country that the virus was found	`Brazil`
RefSeq accession number	The accession number assigned by NCBI RefSeq database	`NC_011181`
GenBank accession number	The accession number of the GenBank record that RefSeq used	`EU914817`

361

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

362 For instance, to find all Begomoviruses that infect sweet potato, user can input the phrase

363 "sweet potato" in the query box and click the "go" button (Figure 4A).



364

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

Figure 4. Keyword search results. (A) Search by host e.g. "sweet potato". (B) Search by a phrase e.g. "new world monopartite".

User can combine multiple search attributes in a query. The logical AND relationship is assumed between attributes. For example, user can enter "new world monopartite" in the search box to search for all New World monopartite Begomovirus (Figure 4B). But the current version of the search function remains primitive as it is virtually inherited from the 'LIKE' search of MySQL, meaning that the order of queried attributes is important. When multiple attributes are specified, they must be arranged according to the order enlisted in Table 2 from top to bottom. For the same example above, the query "monopartite new world" will result in no hits.

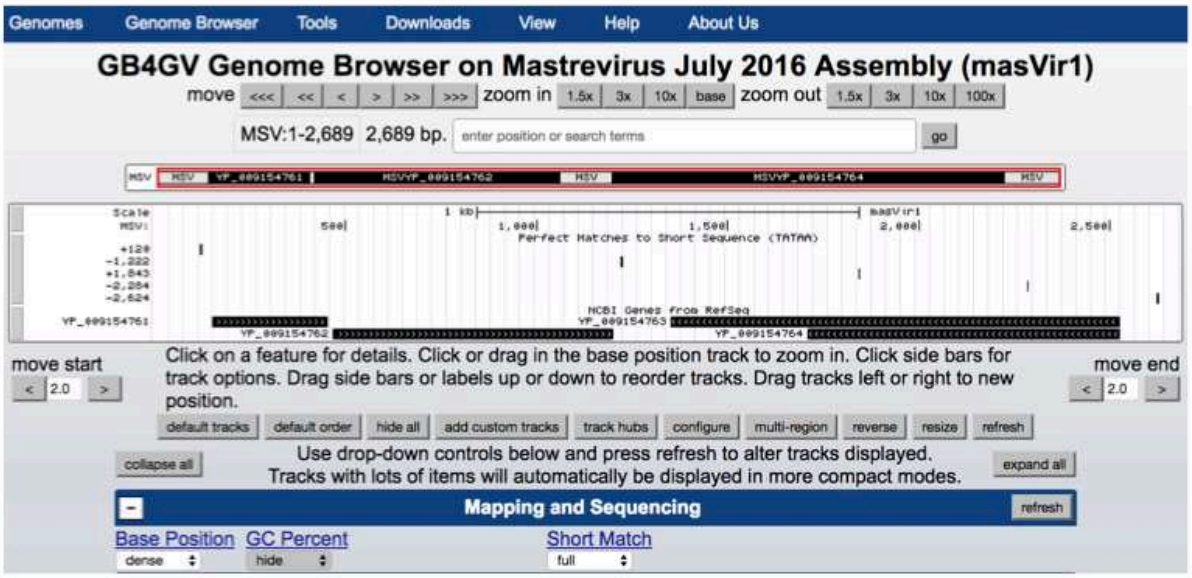
Short Match

The ability to support ad-hoc sequence search can help researchers to identify potential short regulatory sequences that can be validated further by experiment. Examples of these regulatory sequences include TATA-box (Sanz-Burgos & Gutierrez 1998), and polyadenylation signal `AWTAAA` (`W` means `A` or `T`). The Short Match function allows users to search for DNA sequences from 2 to 30 bases with the support of IUPAC ambiguity codes. Figure 5 illustrates how to specify a short sequence match, and how to inspect the context of a hit within a specific region through the Genome Browser's zoom-in function.

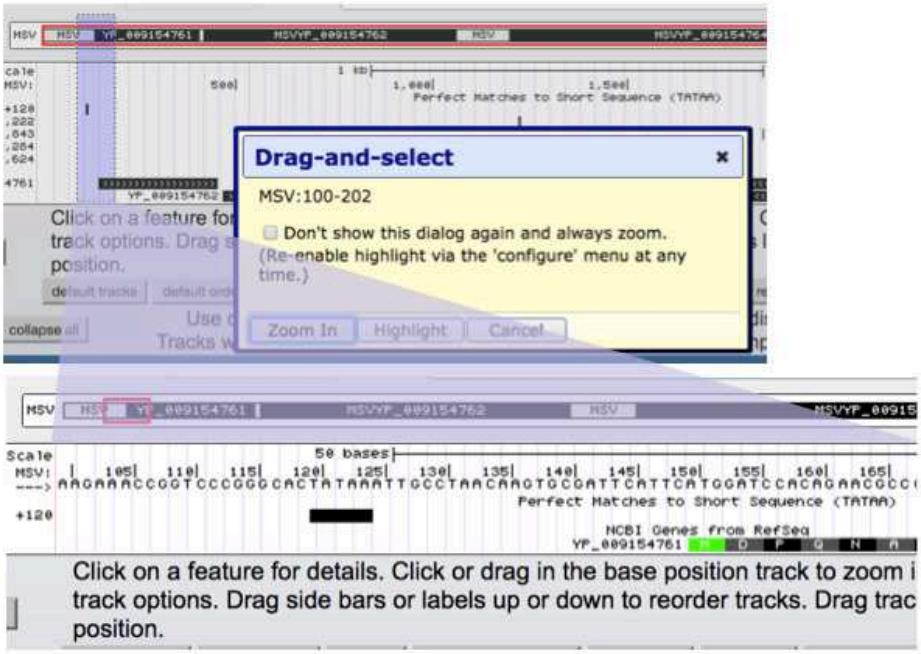
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

A

B



C



385

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

Figure 5. Setup of Short Match function. (A) Turn on the "Short Match" track to "full", and click the Short Match link. It allows users to input the sequence to search for. (B) After clicked the submit button, Genome Browser will return to the main genome view. If the searched sequence is found, results are displayed under the "Short Match" track including the genomic locations prefixed by a + or – to indicate the hit lies in the reference strand or the complementary strand, respectively. (C) Users can zoom in to a smaller region by dragging the mouse pointer.

Putative Iterons and TATA

It has been known iterons contributed to viral replication (Arguello-Astorga et al. 1994; Sanz-Burgos & Gutierrez 1998). Studied had shown binding activities between REP and iterons in Mastrevirus and Begomovirus (Fontes et al. 1992; Sanz-Burgos & Gutierrez 1998). gb4gv maintains 21,298 putative iterons and TATA-box sequences in the long intergenic region. Users can view this information by turning on the "Iterons and TATA" track. Figure 6 shows an example of iterons and TATA-boxes predicted in Begomovirus Okra leaf curl virus-Cameroon OkLCuV (NC_013017).

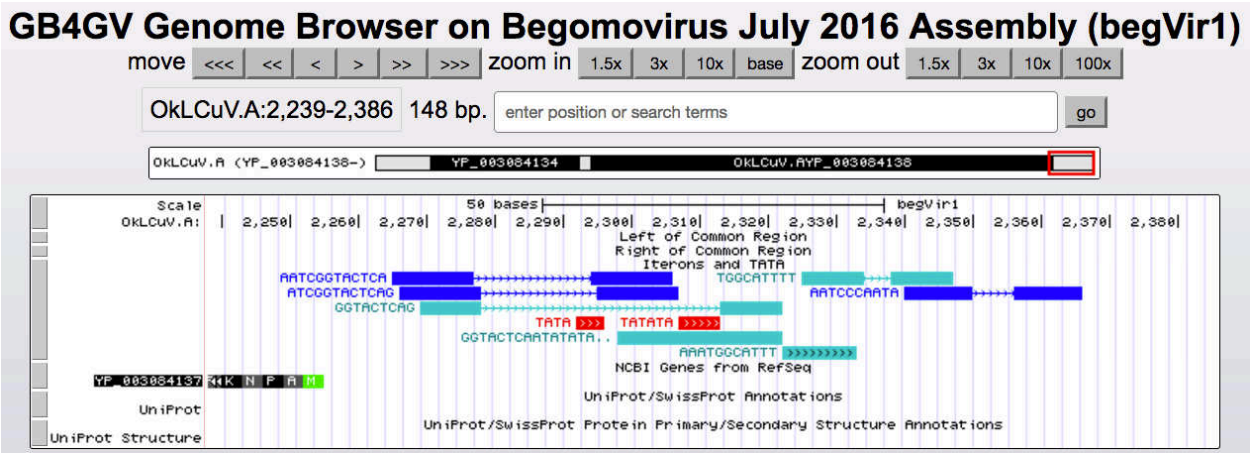


Figure 6. Iterons and TATA track. Different colors are used to denote various sequence features: direct repeats in blue, inverted repeats in blue-green, and TATA-box in red. Tandem repeats are highlighted with ".." at the end the label e.g. the inverted tandem repeats "GGTACTCAATATATA.." above consists of "GGTACTCAATATATA" and "TATATAGTGAGTACC" with their overlapping regions underlined. Lastly, our database also highlights palindromic-like sequence by ">>>...>>>", e.g. "AAATGGCATTT".

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

410 *Small Interference RNAs*

411 Understanding plant immunity is the foremost step to fight against viral infection. Virus-
412 derived RNA silencing is a vital immune response triggered in plants in the face of viral
413 infection. Thus we have incorporated datasets from two virus-derived small interference
414 RNA (siRNA) studies into gb4gv. One study used pyrosequencing to sequence siRNAs in
415 tomato leaves (*Solanum lycopersicum*) inoculated with monopartite Begomovirs TYLCV
416 (Donaire et al. 2009). Another study had used deep sequencing to survey siRNAs in the
417 leaves of tomato (*Solanum lycopersicum*) and tobacco (*Nicotiana benthamiana*) inoculated
418 with monopartite Begomovirus and its associated betasatellite (TYLCCNV/TYLCCNB)
419 (Yang et al. 2011). Both studies had mapped the siRNAs to the genomes of respective hosts.
420 However, it is unclear whether or not these siRNA sequences are species specific. Are
421 siRNAs mapped to biased locations? In order to answer these questions, we incorporated
422 siRNA sequences from these two studies into gb4gv and mapped these siRNAs to genomes
423 of Begomovirus and betasatellite. Each sample occupies a track (Figure 7A).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

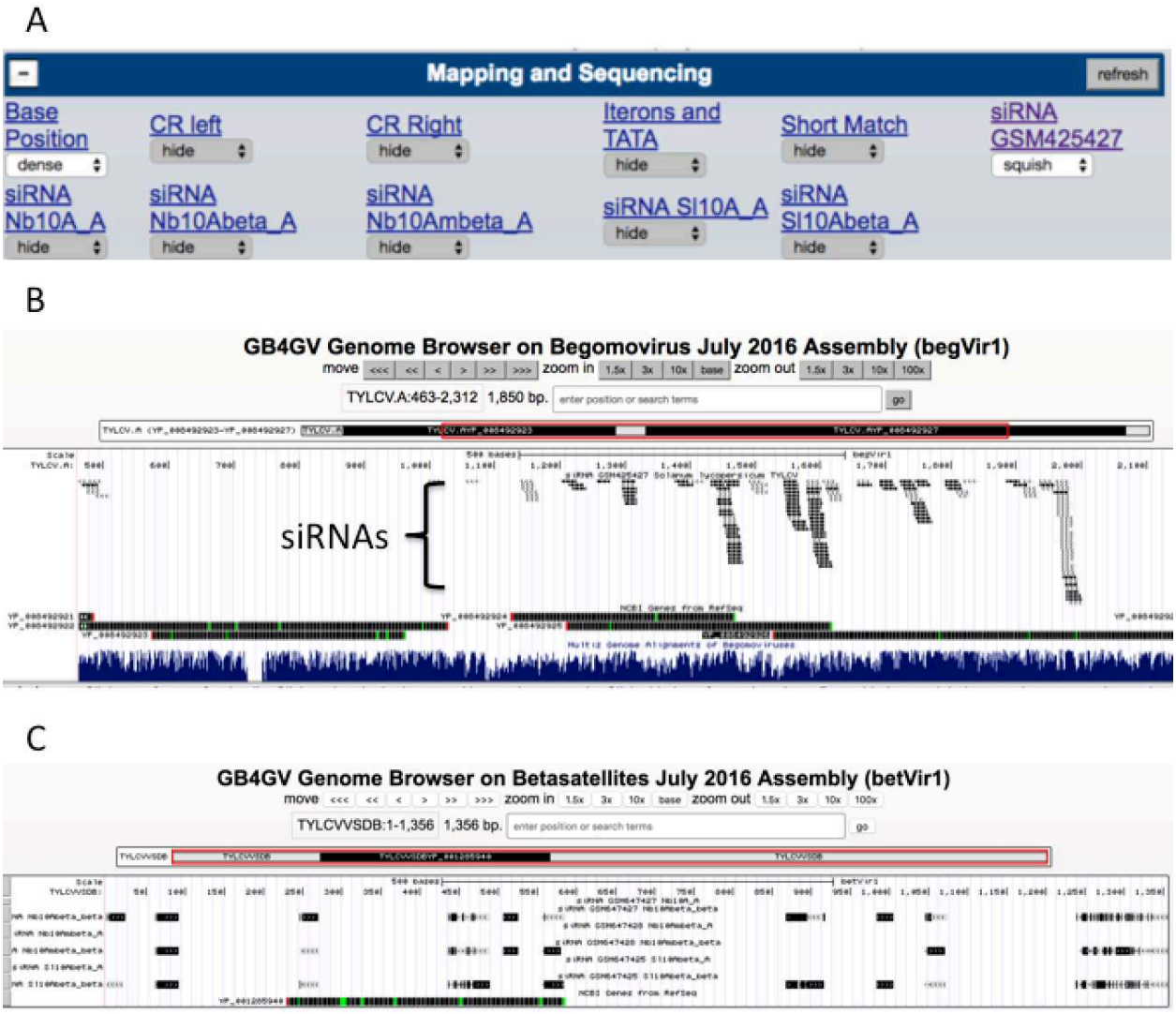


Figure 7. siRNA mapping. (A) Six samples of siRNA sequences are available for visualization, one track for each sample. (B) An example to visualize the mapping of siRNAs from GSM425427 on monopartite Begomovirus TYLCV. 'Squish' mode was used in this example. (C) Another example to show the appearance when 'Dense' mode was used to display siRNAs mapped to betasatellite TYLCVVSDB based on samples from GSE26368.

siRNAs mapped to the viral strand and complementary strand are encoded in dark and light color, respectively (Figure 7B). According to our limited browsing, siRNAs do not map uniformly along the genome. In betasatellites, a sizeable number of mapped siRNAs were skewed toward a 100-bp region near to the 5' side of the SCE.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

436 *BLAT*

437 Our database is also equipped with a lightweight sequence query engine BLAT (Kent 2002).
438 BLAT stands for BLAST-like alignment tool. It has been widely used to search for highly
439 similar gapped alignments. In situation like the detection of exons based on a spliced mRNA
440 sequence, BLAT provides a speedy mapping of the query sequence onto the genome. Major
441 differences between BLAT and Short Match are:

442 1. The minimum and maximum query lengths for BLAT are 20 and 25,000 bps,
443 respectively.
444 2. BLAT search against genomes in a database specified by the user. Whereas Short Match
445 searches for queried sequence only in the current active genome.
446 3. BLAT can handle gapped hit but not for Short Match.

447

448 As an illustration, we used an unusually long (52 bps) iteron sequence
449 "GAGTGATTTCTTATTATGTGATTGTCCATTAAAGGGATAAAGTGACGATGGA" (Figure 8A)
450 found in YOM (Cotton leaf curl virus betasatellite NC_017829) to query against betasatellite
451 genomes. Intriguingly, six other betasatellite genomes were found to contain sequences
452 that share from 78.9% to 96.2% of identity with the queried sequence (Figure 8B). To
453 further examine the hit in virus TLCNDVDB, we clicked the "browser" link on the left, which
454 led to Figure 8C. It shows that the queried sequence hits a region TLCNDVDB clustered
455 with iterons. The solid black bar at the bottom indicates that YOM and TLCNDVDB differ at
456 only two sites.

457

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

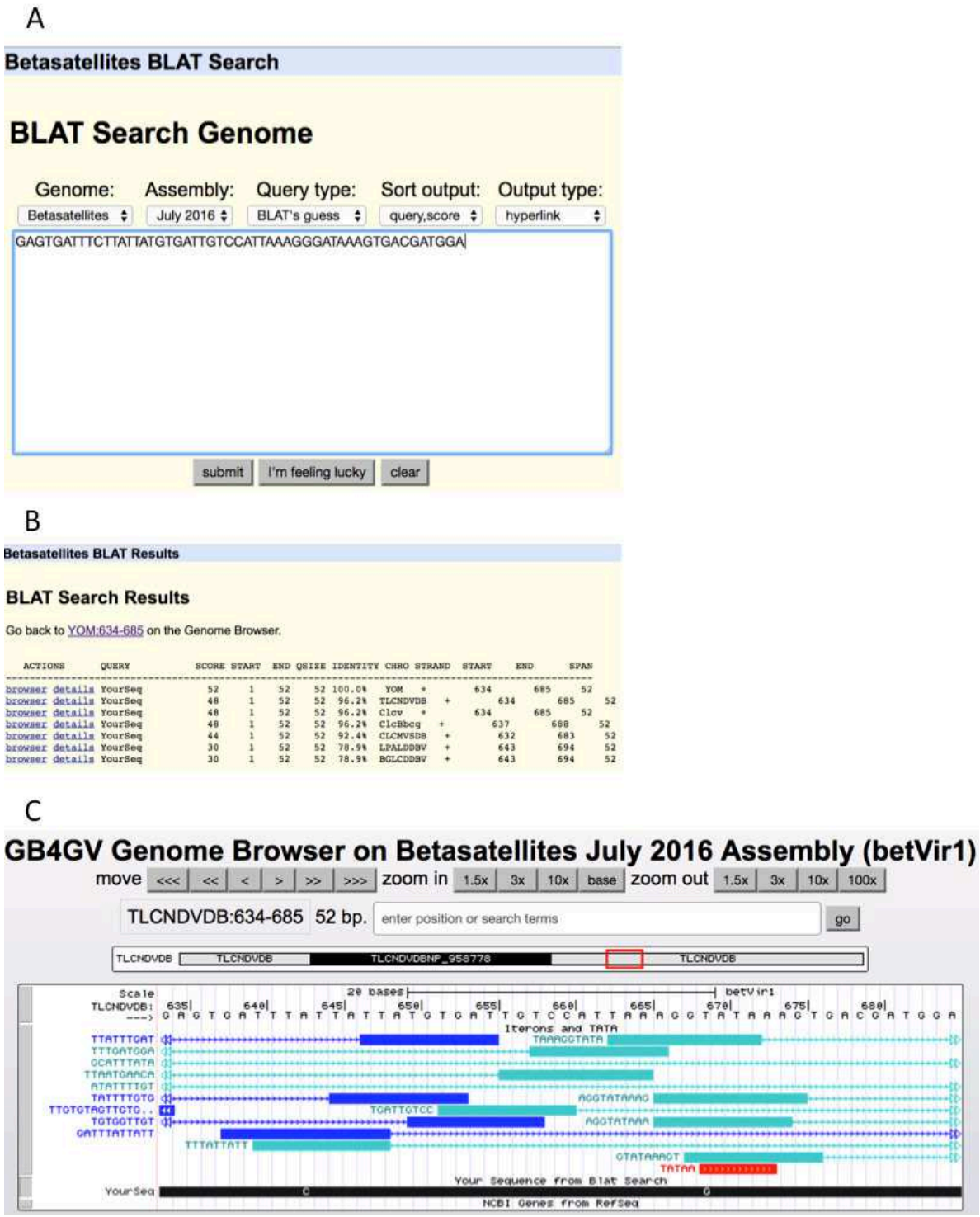


Figure 8. BLAT search. (A) Sequence search box. (B) List of hits with a column of hyperlinks referencing the hit genomes on the far left. (C). Display of hit in the context of targeted genome by clicking the "browser" link on the left of panel (B). The black bar at the bottom represents the queried sequence. Nucleotides displayed on the solid black bar represent mismatched nucleotides between the queried sequence and the targeted genome.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

465 ## Conclusion

466 Genomics visualization is a useful approach to enhance interpretation especially when the
467 quantity or diversity of the viral genomic data is large. We have harnessed the capability of
468 the widely acclaimed Genome Browser specially for the Geminivirus research community.
469 Instead of using a generic one size fits all approach to organize viral genomes, we have
470 taken a semi-automatic pipeline to preserve unique characteristics of Geminivirus in our
471 web-based database gb4gv. Additionally, we have enhanced keyword search capability of
472 manually curated attributes such as infected hosts, geographical location. However, further
473 improvement is needed to accommodate more flexible multiple attributes queries.
474 Moreover, we have predicted 127 pairs of common regions pertaining to bipartite
475 Begomoviruses. This is a useful piece of information as common regions are implicated in
476 coupling the two main genomes for bipartite Begomovirus during encapsidation. As the
477 ultimate goal in studying the genomes of *Geminiviridae* family is to understand the
478 underlying genomic features that are suggested to promote its propagation, we have
479 developed our own method to unravel putative iterons and TATA-box in the 5' side of the
480 common region and they can be visualized readily with genomic features flanking them.
481 Geminiviruses are diverse and fast evolving. Facilitated by the ever-decreasing DNA
482 sequencing cost, we anticipate more viral genomes will be sequenced in the near future.
483 We are certainly committed to maintaining the information in gb4gv as up-to-date as
484 possible. Given the flexibility of the Genome Browser in accommodating new annotation
485 tracks, if more genome-wide experimental data is available in the future such as Chip-Seq,
486 it can be included into gb4gv readily without software modification as illustrated by the
487 siRNA tracks discussed above. While viral regulatory elements play crucial roles in

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

488 influencing replication and transcription in cellular environment, we will continue our
489 effort in developing new methods to identify essential sequence elements that might offer
490 new insights for experimental virologists to design effective modalities to fight against the
491 infection of Geminiviruses.

492 Acknowledgements

493

494 Funding

495 This project is partly supported by the startup fund provided by Lafayette College to E.S.H.
496 C.M.N. was supported by the EXCEL Summer Scholar Program funded by Lafayette College.

497 Competing Interests

498 The authors declare that they have no competing interests.

499 References

500 Arguello-Astorga GR, Guevara-Gonzalez RG, Herrera-Estrella LR, and Rivera-Bustamante
501 RF. 1994. Geminivirus replication origins have a group-specific organization of
502 iterative elements: a model for replication. *Virology* 203:90-100.
503 10.1006/viro.1994.1458
504 Bernard V, Brunaud V, and Lecharny A. 2010. TC-motifs at the TATA-box expected position
505 in plant genes: a novel class of motifs involved in the transcription regulation. *BMC*
506 *Genomics* 11:166. 10.1186/1471-2164-11-166
507 Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K,
508 Clawson H, Green ED, Haussler D, and Miller W. 2004. Aligning multiple genomic
509 sequences with the threaded blockset aligner. *Genome Res* 14:708-715.
510 10.1101/gr.1933104
511 Briddon RW, Mansoor S, Bedford ID, Pinner MS, Saunders K, Stanley J, Zafar Y, Malik KA,
512 and Markham PG. 2001. Identification of dna components required for induction of
513 cotton leaf curl disease. *Virology* 285:234-243. 10.1006/viro.2001.0949
514 Brown JK, Zerbini FM, Navas-Castillo J, Moriones E, Ramos-Sobrinho R, Silva JC, Fiallo-Olive
515 E, Briddon RW, Hernandez-Zepeda C, Idris A, Malathi VG, Martin DP, Rivera-
516 Bustamante R, Ueda S, and Varsani A. 2015. Revision of Begomovirus taxonomy
517 based on pairwise sequence comparisons. *Arch Virol* 160:1593-1619.
518 10.1007/s00705-015-2398-y

519 Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, and Llave C. 2009. Deep-
520 sequencing of plant viral small RNAs reveals effective and widespread targeting of
521 viral genomes. *Virology* 392:203-214. 10.1016/j.virol.2009.07.005
522 Duffy S, Shackelton LA, and Holmes EC. 2008. Rates of evolutionary change in viruses:
523 patterns and determinants. *Nat Rev Genet* 9:267-276. 10.1038/nrg2323
524 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
525 throughput. *Nucleic Acids Res* 32:1792-1797. 10.1093/nar/gkh340
526 Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the*
527 *author Department of Genome Sciences, University of Washington, Seattle*.
528 Fontes EP, Luckow VA, and Hanley-Bowdoin L. 1992. A geminivirus replication protein is a
529 sequence-specific DNA binding protein. *Plant Cell* 4:597-608.
530 GenBank Record. NCBI GenBank Record. *Available at*
531 https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html (accessed 27 December
532 2016).
533 Gene Expression Omnibus. NCBI Gene Expression Omnibus. *Available at*
534 https://www.ncbi.nlm.nih.gov/geo/ (accessed 1 December 2016).
535 Ghanim M, Morin S, and Czosnek H. 2001. Rate of Tomato yellow leaf curl virus
536 Translocation in the Circulative Transmission Pathway of its Vector, the Whitefly
537 Bemisia tabaci. *Phytopathology* 91:188-196. 10.1094/PHYTO.2001.91.2.188
538 Gutierrez C. 1999. Geminivirus DNA replication. *Cell Mol Life Sci* 56:313-329.
539 ICTV. 2015. ICTV. *Available at* https://talk.ictvonline.org/files/master-species-lists/
540 (accessed 15 December 2016).
541 Jeske H, Lutgemeier M, and Preiss W. 2001. DNA forms indicate rolling circle and
542 recombination-dependent replication of Abutilon mosaic virus. *EMBO J* 20:6158-
543 6167. 10.1093/emboj/20.21.6158
544 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.
545 10.1101/gr.229202. Article published online before March 2002
546 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002.
547 The human genome browser at UCSC. *Genome Res* 12:996-1006.
548 10.1101/gr.229102. Article published online before print in May 2002
549 Kumar S, Stecher G, and Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis
550 Version 7.0 for Bigger Datasets. *Mol Biol Evol* 33:1870-1874.
551 10.1093/molbev/msw054
552 NCBI RefSeq. NCBI RefSeq. *Available at* ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/
553 (accessed 15 December 2016).
554 NCBI Viral Genomes. NCBI Viral Genomes. *Available at*
555 https://www.ncbi.nlm.nih.gov/genome/viruses/ (accessed 1 June 2016).
556 Orozco BM, and Hanley-Bowdoin L. 1996. A DNA structure is required for geminivirus
557 replication origin function. *J Virol* 70:148-158.
558 Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, and Burley SK. 1999. TATA element
559 recognition by the TATA box-binding protein has been conserved throughout
560 evolution. *Genes Dev* 13:3217-3230.
561 Pilartz M, and Jeske H. 2003. Mapping of abutilon mosaic geminivirus minichromosomes. *J*
562 *Virol* 77:10808-10818.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2705v1 | CC BY 4.0 Open Access | rec: 6 Jan 2017, publ: 6 Jan 2017

563 Sanz-Burgos AP, and Gutierrez C. 1998. Organization of the cis-acting element required for
564 wheat dwarf geminivirus DNA replication and visualization of a rep protein-DNA
565 complex. *Virology* 243:119-129. 10.1006/viro.1998.9037
566 Saunders K, Norman A, Gucciardo S, and Stanley J. 2004. The DNA beta satellite component
567 associated with ageratum yellow vein disease encodes an essential pathogenicity
568 protein (betaC1). *Virology* 324:37-47. 10.1016/j.virol.2004.03.018
569 UCSC Admin. UCSC Genome Browser Download. *Available at*
570 http://hgdownload.soe.ucsc.edu/admin/ (accessed 1 June 2016).
571 UCSC GB Statistics. UCSC Genome Browser Statistics. *Available at*
572 http://genome.ucsc.edu/admin/stats/ (accessed 15 December 2016).
573 UCSC Genome Browser. UCSC Genome Browser. *Available at* http://genome.ucsc.edu.
574 UCSC Genome Browser User Guide. UCSC Genome Browser User Guide. *Available at*
575 https://genome.ucsc.edu/goldenpath/help/hgTracksHelp.html (accessed 15
576 December 2016).
577 UniProtKB. UniProt Knowledgebase. *Available at* http://www.uniprot.org/downloads
578 (accessed 1 June 2016).
579 Varma A, and Malathi VG. 2003. Emerging geminivirus problems: A serious threat to crop
580 production. *Annals of Applied Biology* 142:145-164. DOI 10.1111/j.1744-
581 7348.2003.tb00240.x
582 Xie Y, Wu P, Liu P, Gong H, and Zhou X. 2010. Characterization of alphasatellites associated
583 with monopartite begomovirus/betasatellite complexes in Yunnan, China. *Virol J*
584 7:178. 10.1186/1743-422X-7-178
585 Yang X, Wang Y, Guo W, Xie Y, Xie Q, Fan L, and Zhou X. 2011. Characterization of small
586 interfering RNAs derived from the geminivirus/betasatellite complex using deep
587 sequencing. *PLoS One* 6:e16928. 10.1371/journal.pone.0016928
588 Zhou X, Xie Y, Tao X, Zhang Z, Li Z, and Fauquet CM. 2003. Characterization of DNAbeta
589 associated with begomoviruses in China and evidence for co-evolution with their
590 cognate viral DNA-A. *J Gen Virol* 84:237-247. 10.1099/vir.0.18608-0
591