gb4gv: A genome browser for geminivirus

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

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

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

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

Background: Geminiviruses (family *Geminiviridae*) are prevalent plant viruses that imperil 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, the viron of begomoviruses prevailing in America encapsulates two similar-sized circular DNA components and both are required for systemic infection of plants. But, the bipartite components are kept separately in NCBI as individual genomes 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 *Geminiviridae*, 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 a better understanding of the incredible degree of diversity of these viruses, and of the complex relationships within and between the different genera in the *Geminiviridae*. **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: Geminiviruses (family *Geminiviridae*) are prevalent plant viruses that imperil

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, the viron of begomoviruses prevailing in America encapsulates two similar-

sized circular DNA components and both are required for systemic infection of plants. But,

the bipartite components are kept separately in NCBI as individual genomes 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

23   virus, but also supplements biologically relevant annotations that help to interrogate this

24   virus e.g. the targeted host, putative iterons, siRNA targets etc.

25   **Methods:** We have employed manual and automatic methods to curate 508 genomes from

26   four major genera of *Geminiviridae*, and 161 associated satellites obtained from NCBI

27   RefSeq and PubMed databases.

28   **Results:** These data are available for free access without registration from our website.

29   Besides genomic content, our website provides visualization capability inherited from

30   UCSC Genome Browser.

31   **Discussion:** With the genomic information readily accessible, we hope that our database

32   will inspire researchers in gaining a better understanding of the incredible degree of

33   diversity of these viruses, and of the complex relationships within and between the

34   different genera in the *Geminiviridae*.

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

36

Introduction

38   Geminiviruses (family *Geminiviridae*) have emerged as one of the most prevalent and

39   problematic plant pathogens especially in developing countries (Sattar et al. 2013;

40   Scholthof et al. 2011; Shepherd et al. 2010). In terms of diversity, they have become the

41   largest group of plant viruses to exist today. It posts significant threat both socially and

42   economically as geminiviruses are the most destructive pathogens in subsistence

43   agriculture like beans, cotton, maize, sweet potato, and tomato (Jeske 2009; Sattar et al.

44   2013; Scholthof et al. 2011; Shepherd et al. 2010). The economic impact of geminivirus

45   infection can be seen across the globe. Annual economic loss is estimated to be US$1.9-2.7

46   billion in East and Central Africa. Maize streak virus alone has caused hundreds of millions

47   of loss in food crops per year (Shepherd et al. 2010).

48   Geminiviruses often infect plants as complexes: a mixture of viral isolates identified

49   as distinct species, as well as DNA satellites. Moreover, they are able to undergo mutation,

50   recombination and reassortment both frequently and rapidly. Together, these factors

51   increase the diversity and capabilities of the family, allowing them to invade new hosts and

52   new environments without complication. In order to prevent geminiviruses from becoming

53   even more of a threat to our growing human population, it is critical that scientists are able

54   to better understand the genomic sequences of these viruses. Geminiviruses rely heavily on

55   their host's cellular machinery so having a greater knowledge of their genetic makeup will

56   allow scientists to formulate biotechnological means to help plants fight their attackers

57   successfully.

58   Geminiviruses comprise a family of plant viruses that exist in the form of twinned

59   icosahedral particles holding small, circular, single stranded deoxyribonucleic acid (ssDNA)

60   genomes. The ssDNA genome structure enables it to evolve at high rate comparable to RNA

61   viruses (Duffy et al. 2008). The viral genome encodes only 5-7 proteins, making

62   geminiviruses one of the smallest virus types known to scientists today. Within

63   *Geminiviridae*, seven genera are currently recognized by The International Committee on

64   Taxonomy of Viruses (ICTV): *Becurtovirus*, *Begomovirus* (the one with the largest number

65   of species), *Curtovirus*, *Eragrovirus*, *Mastrevirus*, *Topocuvirus*, and *Turncurtovirus*.

66   Depending on the genera, the viral genome comprises of either one (monopartite) or two

67   (bipartite) DNA components. Monopartite genomes consist of a circular, ssDNA molecule

68   (similar to the DNA-A component of bipartite genomes) that is often associated with an

69   alpha- or betasatellite, while bipartite genomes consist of separated DNA-A and DNA-B

70   components of similar size.

71   National Center for Biotechnology Information (NCBI) designates a separate

72   website to host viral genomes (NCBI Viral Genomes). Its collection includes almost all

73   known viruses in the world, making it one of the most comprehensive resources for

74   studying viral genomics. Viral genomes are formatted in standard GenBank record

75   (GenBank Record) exactly like other living organisms. However, genome architectures of

76   viruses exhibit significant difference from living organisms. For instance, virions of

77   bipartite begomoviruses encapsulate two circular, ssDNA components in which the two

78   components are required for infectivity. But such critical association between the two

79   components is often missing from NCBI Viral Genome database. Moreover, vital

80   information about the virus such as location where it was found, targeted hosts, etc. are not

81   searchable attributes, limiting the utility of the database. These are the reasons that we

82   have undertaken this project in providing researchers a comprehensive, up-to-date, and

83   integrated environment at their fingertips. The database we built rests on the software

84   architecture of UCSC Genome Browser website (Kent et al. 2002; UCSC Genome Browser)

85   as such we named our database gb4gv, which stands for Genome Browser for Geminivirus.

86   For clarity, we reserve "UCSC Genome Browser" to refer to the website itself (UCSC

87   Genome Browser), and "Genome Browser" to mean the software that supports the website.

88   Genome Browser was chosen because of its versatility in visualizing genomes, richness in

89   built-in functions, flexibility in incorporating annotations, and software robustness in

90   handling large volume of requests - 872,000 requests per day on average (UCSC GB

91   Statistics). Although Genome Browser offers these benefits, its original design gears mainly

92   toward eukaryotes. In order to unleash the power of Genome Browser, we have made

93   substantial effort in modeling geminivirus genomes into a structure that can take full

94   advantage of its functionalities. gb4gv can be accessed, without registration requirements,

95   from here: http://gb4gv.lafayette.edu. Users can make use of the built-in functions

96   provided through our website to download genomes or sequences of interested regions

97   freely.

98   **Materials & Methods**

99   *Compilation of Geminivirus Genomes*

100

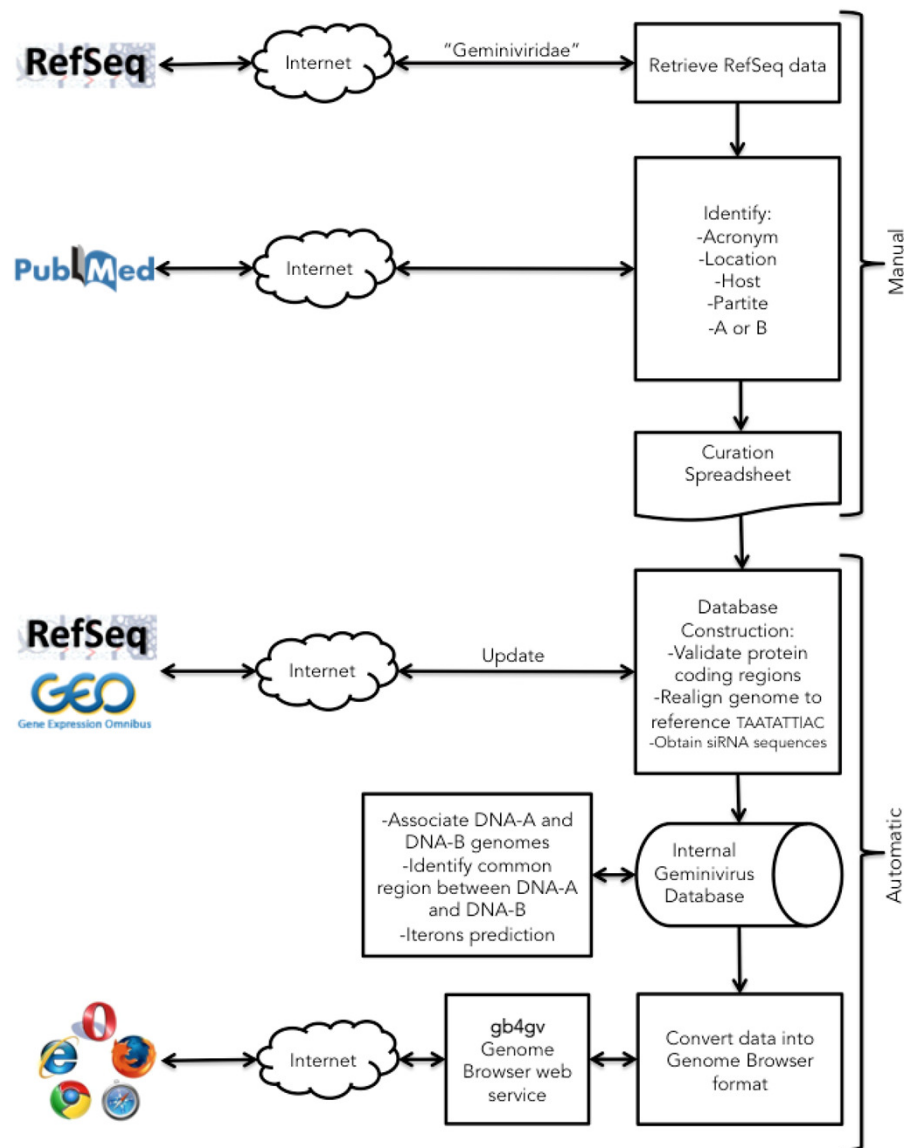



Figure 1 Project workflow of the semi-automatic annotation process. Our workflow begins
with a manual process to identify information generally not documented in the GenBank
record such as the acronym of the virus, location, infecting host, monopartite or bipartite
genome, and genomes association for 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.

112   Figure 1 above summarizes the semi-automatic annotation workflow designed for this

113   project. The primary source of our data originated from NCBI RefSeq database (NCBI

114   RefSeq) because redundant genomes were purged. But we also cross-referenced our data

115   with the ICTV Master Species List 2015 v1 obtained from ICTV (ICTV 2015). We had

116   identified 700 RefSeq entries that comprised of 529 distinct geminiviruses. Note that DNA-

117   A and DNA-B of a bipartite begomovirus are kept in separate entries in RefSeq.

118   *Begomovirus* occupies the largest genus of the family, followed by *Mastrevirus*, and

119   *Curtovirus*. Other genera were found sporadically including two becurtoviruses, one

120   topocuvirus, one eragrovirus, and one turncurtovirus, while genera of ten entries remain

121   unknown. Here we had decided to incorporate only genera that represent major

122   *Geminiviridae* genera i.e. *Begomovirus*, *Curtovirus*, and *Mastrevirus* into gb4gv. As a result,

123   genomes of 514 geminiviruses representing 97% of the *Geminiviridae* found in NCBI were

124   considered for further review. We will regularly assess the need to include other minor

125   genera into our database if more samples from them are discovered in the future.

126   Besides the main genomes, ancillary alphasatellites and betasatellites are often isolated

127   together with monopartite begomoviruses (Xie et al. 2010) and they are found to play

128   essential roles in boosting host's symptoms and viral movement (Briddon et al. 2001;

129   Saunders et al. 2004; Zhou et al. 2003). We had identified and reviewed 66 and 105

130   alphasatellite and betasatellite genomes, respectively, from NCBI.

131   Meta information or attributes such as the geographical location of the virus are

132   important to understand the virulence of the virus but it is not always available in genome

133   database. Therefore we manually searched for additional information about these viruses

134   from existing literature. In particular, we focused on identifying or reconfirming the

135   location where they were collected, the hosts they infected, their acronyms, monopartite or

136   bipartite genome, and the counterpart genome in case of bipartite. Importantly, we have

137   made these attributes searchable in our database.

138   Following the manual process is the automatic annotation process. In this step, NCBI

139   RefSeq entries belonging to *Geminivirus* were parsed to ensure that each entry satisfies the

140   following two criteria:

141   1.   Every geminivirus and geminivirus-associated DNA satellite genomes must possess the

142   iconic structurally conserved element (SCE), which is the genomic landmark of

143   geminiviruses including satellites. The canonical structure of the SCE is `TAATATT|AC`,

144   where "|" stands for the cleavage site targeted by the viral replication protein in the

145   initial step of DNA replication (Gutierrez 1999; Jeske et al. 2001; Pilartz & Jeske 2003).

146   The prevalent SCE sequence of alphasatellite is `TAGTATT|AC`, which varies slightly

147   from the canonical SCE sequence. Nonetheless, owing to either DNA sequencing errors

148   or random mutations, the 5' side of the SCE of some viruses may deviate slightly (less

149   than one nucleotide) from the canonical form from above. To accommodate such

150   minutiae, we tolerated entries with up to one mismatch from `TAATATT`. Genomes failed

151   to meet this criterion were excluded from gb4gv.

152   2.   Besides genomes, gb4gv also keeps individual viral proteins if they satisfy our quality

153   checking. The coding region (CDS) of a gene defined in a RefSeq entry must be

154   translated exactly into the stated peptide in the RefSeq entry. Genes failed this criterion

155   were excluded from our database. But genomes containing erroneous CDS were still

156   kept in the database.

157

158   Through our tandem manual and automatic annotations, 6 out of 514 RefSeq entries of

159   geminiviruses failed the validation process stated above, resulting in 508 genomes being

160   selected into our database. For satellite genomes, 7 out of 66 alphasatellites and 3 out of

161   105 betasatellites failed our validation. Table 1 below categorizes all the genomes accepted

162   into our database by genus, number of genomes per virus, and geographical origin. The

163   aforementioned annotation information can be downloaded from our website in tab-

164   separated format (http://gb4gv.lafayette.edu/downloads.html).

165   Table 1. A summary of genomes stored in gb4gv. The numbers inside the parentheses
166   denote the numbers of genomes. The lower part of the table categorizes begomoviruses
167   further by origin 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

168

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

169

170   *Small Interfering RNAs*

171   A key aspect of gb4gv is to inspire researchers to formulate insightful strategies that can be

172   used to eradicate the propagation of geminiviruses. Therefore, studying the immune

173   response launched by infected plant is a promising research direction. Thus, we had

174   downloaded datasets from two small interfering RNAs studies deposited in NCBI GEO

175   database (Gene Expression Omnibus): GSM425427, and GSE26368. siRNA sequences were

176   mapped to the genomes of begomovirus and betasatellite through a customized Python

177   script. Mapping tolerated up to two mismatches in internal positions without gaps. A

178   Genome Browser annotation track is designated for each sample, which can be found under

179   "Mapping and Sequencing" section of each virus. In *Begomovirus* or betasatellite, six siRNA

180   tracks were configured.

181

182   *Standardization of Circular Genomes*

183   Like many other genomic databases such as NCBI RefSeq and UCSC Genome Browser,

184   circular genomes are linearized. Instead of opening the circular genome at arbitrary sites,

185   circular genomes were opened at the biological cleavage site at the SCE. A benefit of

186   standardizing the opening site is to facilitate syntenic analysis (to be discussed in *Multiz*

187   section below). Under the standardized linearization scheme, a genome always begins with

188   'AC' and terminates with 'TAATATT' at the 5' and 3' termini, respectively. Thereby we

189   standardized all genomes obtained from RefSeq. Genomes not conforming to this standard

190   were shifted until they met the above criterion. Out of 669 accepted genomes in gb4gv,

191   surprisingly, 112 (17%) of them required this adjustment.

192

193   *Data Models*

194   Genome Browser was originally designed to visualize mammalian genomes (Kent et al.

195   2002). It was later enhanced to host non-mammalian animals e.g. *C. elegans*, and then

196   unicellular organisms such as yeast. Ebola genome is the first and remains to be the only

197   viral genome available in UCSC Genome Browser at present. This historical background

198   reveals that the data model of Genome Browser is geared toward the display of

199   chromosomes of a species. Such data model serves well with living organisms but it poses

200   two challenges in configuring Genome Browser for geminivirus genomes:

201   1. The genus *Begomovirus* is known to be diverse (Brown et al. 2015) with over 300 DNA

202   components being identified by us. If we were to coerce the existing data model to

203   begomoviruses, 300 databases are needed, leading to a huge species tree in the home

204   page, hampering website performance, and prohibiting data browsing. To circumvent

205   this, we modeled each viral genus as an organismal species, and the array of viral

206   species of a genus as chromosomes of an organism. Based on this workaround, gb4gv

207   consists of five databases (a database per genus including one for each satellite

208   although, in biological terms, satellite is not considered a genus): *Begomovirus*,

209   *Mastrevirus*, *Curtovirus*, alphasatellite, and betasatellite.

210   2. A special configuration is needed to establish the association between the bipartite DNA

211   components (DNA-A and DNA-B) of a begomovirus. In gb4gv, DNA-A and DNA-B were

212   treated as two separate chromosomes. The coupling of DNA-A and DNA-B components

213   of a bipartite begomovirus can only be achieved manually as their RefSeq accession

214   numbers reflect no information about their relationship. In order to facilitate users to

215   associate them easily, a viral species in our database is uniquely referenced by an

216   acronym, e.g. AbMBV is the reference of Abutilon mosaic Brazil virus. But the two DNA

217   components of a bipartite begomovirus will become indistinguishable under this

218   scheme. Thus, we suffix the acronym of a bipartite virus by ".A" and ".B". E.g. the DNA-A

219   and DNA-B of the virus AbMBV can be found effortlessly through AbMBV.A and

220   AbMBV.B, respectively. An advantage of using an acronym as the key to retrieve a virus

221   is to release the burden of users to pull up the accession number of the virus as most

222   people can remember the acronym rather than the arbitrary accession number.

223   Moreover, we recognize that some viruses are referenced by multiple acronyms

224   without a consensus. To accommodate such variability our database maintains a list of

225   searchable aliases for every virus, e.g. Tomato leaf curl New Delhi virus can be

226   identified by either TolCNDV or ToLCV_India.

227

228   *Common Region Identification in Bipartite Begomoviruses*

229   The bipartite genomes of a begomovirus share a highly similar, non-coding segment

230   flanking the SCE "TAATATTAC". This segment is colloquially named the common region

231   (CR). CRs serve a crucial role in viral DNA replication. Studies had shown that the 5' side of

232   CRs contain replication protein binding sites (Orozco & Hanley-Bowdoin 1996). Thus, CRs

233   harbor vital regulatory signals that influence the replication and the coupling of the

234   bipartite genomes for begomoviruses. Understanding viral replication is fundamental to

235   combat viral infection. Thus, we undertook the task to predict CRs in bipartite

236   begomoviruses. Based on the manual annotation we did, DNA-A and DNA-B components of

237   a begomovirus were paired up. We extracted the non-coding region, also known as the long

238   intergenic region (LIR), between REP and CP genes in the DNA-A or between NSP and MP

239   genes in the DNA-B. In the next step, we further reduced the LIR into an 809-bp segment,

240   which consisted of a 400-bp segment upstream and downstream of the SCE from the DNA-

241   A and DNA-B. In Figure 2 below, two 809-bp segments were aligned by MUSCLE (Edgar

242   2004) as shown:

```
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    **ATAGCAATCGGGGGGGCACTC**AAAGTCTGTAGCAATCGGGGGAAAGGGGGGCAATTTATAT
NC_011584    **AGAGCATTCGGGGGGGCACTC**AAAGTCTGTAGCAATCGGGGGAAAGGGGGGCAATTTATAT
             *  **** ****************************************************

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

NC_011583    A**GCGGCCATCCG**TA**TAATATTAC**CGGAT**GGCCGC**GCCCG**AAAAAGCAGGTGGACCCCAC**A
NC_011584    A**GCGGCCATCCG**TA**TAATATTAC**CGGAT**GGCCGC**GCCCG**AAAAAGCAGGTGGACCCCAC**-
             **********************************************************

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
             **         **
```

243

244   Figure 2. Identification of common region (CR) shared between DNA-A and DNA-B
245 of bipartite begomoviruses. Sequence alignment of two 809-bp segments located in the LIR
246  of the Old World bipartite East African cassava mosaic Kenya virus (EACMKV). The
247  invariant SCEs are highlighted in red. The inverted repeats constituted the stem of the
248 hairpin structure are highlighted in blue. The two underlined regions indicate the 5' and 3'
249 termini of the common region determined by our method of using a 20-bp sliding window.
250

251   In this example, the two segments were extracted from DNA-A (NC_011583) and

252   DNA-B (NC_011584) of the Old World bipartite East African cassava mosaic Kenya virus

253   (EACMKV) and they were aligned. A 20-bp sliding window was used to scan the alignment

254   base-by-base bilaterally starting from the SCE. Scanning halts when the percentage of

255   sequence identity within the window drops below 80%, an adjustable parameter. The

256   halting locations (the underlined regions in Figure 2) are taken as the 5' and 3' termini of

257   the common region.

258   The average size of a CR was found to be 212 bps (including the SCE) in which the 5'

259   arm, the left segment of the SCE, is usually longer than the 3' arm with an average size of

260   150 bps. The longest CR is 417-419 bps long that belongs to Indian cassava mosaic virus

261   (NC_001932/NC_001933). Whereas Abutilon mosaic Brazil virus (NC_016574/NC_016577)

262   was found to possess the shortest CR, which is 63-67 bps long. Also note that the two

263   approximately 10 bps segments juxtaposing the SCE constitute the stem part of the hairpin

264   structure (Figure 2).

265

266   *Putative Iterons and TATA Box Sites*

267   One of the cis-regulatory signals harboring in CR is iterative elements, also known as

268   iterons (Arguello-Astorga et al. 1994; Arguello-Astorga & Ruiz-Medrano 2001; Sanz-Burgos

269   & Gutierrez 1998). A distinct feature of iterons is the presence of direct or inverted

270   sequence repeats. The following rules were applied to predict iterons in viral genomes:

271   1. They are located in the 5' side of the LIR i.e. from the beginning of the first gene on the

272   complementary strand up to, but excluding, the SCE

disabled

273 2. The minimum length of an iteron is 9 bps. There is no restriction on the maximum

274   length

275 3. The pair of repeats identified differ by at most one base

276 4. Repeats could be direct or inverted

277 5. Its location is no more than 100 bps from a putative TATA box, if any

278 6. At least one of the twelve iteron core motifs is present (Arguello-Astorga & Ruiz-

279   Medrano 2001): GGAGN, GGTAV, GGGGW, GGTAV, GGKGT, GGKGG, GGGGG, GGGGA,

280   GGGTM, GGCGT, GGWGT, and TGGTGTCC.

281

282   As the TATA box sites and iterons work cooperatively to regulate replication, we also

283 identified putative TATA box sites. The consensus sequence of a TATA box is defined as

284 "TATA", followed by any number of "TA" or "AA" repeat (Bernard et al. 2010; Patikoglou et

285 al. 1999). We developed a Python script to scan for iterons and TATA box sites in every

286 geminivirus genome. Based on the above criteria, 5,142 iterons and TATA box sites were

287 predicted from 669 components and genomes. Results can be visualized in gb4gv by

288 activating the 'Iterons and TATA" annotation track.

289

290 *Multiz Track*

291 Genomes of various species within a genus share similarities and differences. Since we

292 have standardized the opening site of the viral circular genomes, a genus-wide syntenic

293 analysis becomes possible. Such comparative view helps to uncover conserved and diverse

294 genomic regions among species. We used the threaded blockset aligner (TBA) (Blanchette

295 et al. 2004) to generate a dynamic multiple sequence alignments of all species in a genus.

296   Unlike other multiple sequence alignment programs of which a sequence from the sample

297   is dedicated to be the reference of the alignment, TBA produces a multiple sequence

298   alignment dynamically based upon the genome being selected for viewing in the Genome

299   Browser. This unique feature enables gb4gv to generate a graphical representation

300   regarding genome conservation among different species with respect to the current

301   queried genome.

302   TBA requires two mandatory inputs: a set of genomic sequences, and a phylogenetic

303   tree defining the evolutionary relationship of the input genomes. We used multiple

304   sequence alignment program MUSCLE (Edgar 2004) to build phylogenetic trees, followed

305   by maximum likelihood tree building PHYLM (Felsenstein 2005) equipped in MEGA7

306   (Kumar et al. 2016). The output phylogenetic tree was in NEWICK format. Based on the

307   genomic sequences and the phylogenetic tree, TBA generated the threaded blockset

308   alignment. The alignment was loaded to a MySQL database referenced by Genome Browser.

309

310   *UniProt/SwissProt Annotations*

311   Protein domain information was overlaid on viral proteins in gb4gv. Reviewed

312   Swiss-Prot annotations were downloaded from UniProt website in XML format

313   (UniProtKB). Viral taxonomy IDs served as the key to retrieve protein domain information

314   from the Swiss-Prot annotations. Sequence of the protein domains identified in the search

315   process was mapped to the genomes by BLAT (Kent 2002).

316

317   *Genome Browser*

318   Version 334 of the Genome Browser was used to build gb4gv. The software was

319   downloaded from the UCSC Genome Browser website (UCSC Admin) and installed in our

320   24-core Linux server running on Centos OS 6.8, Apache 2.2, and MySQL server 5.5.50.

321

Results

323   The web interface of gb4gv is organized in a hierarchy consisting of three levels. The

324   highest level presents all the genera of *Geminiviridae* maintained in gb4gv including two

325   satellites despite they are considered as genera (Figure 3A). The middle level displays

326   information about the genome of an individual virus and corresponding annotation tracks

327   (Figure 3B). Detailed information about a particular annotation e.g. a gene, a protein or a

328   specific genomic sequence, is presented at the lowest level (Figure 3C)

A

LAFAYETTE ⬡ Genome Browser Gateway

| 🏠 | Genomes | Genome Browser | Tools | Downloads | Help | About Us |

Browse/Select Species

SPECIES SEARCH

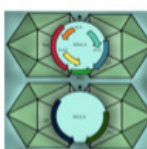
Enter species or common name

REPRESENTED SPECIES

Alphasatellites
Betasatellites
Begomovirus
Curtovirus
Mastrevirus

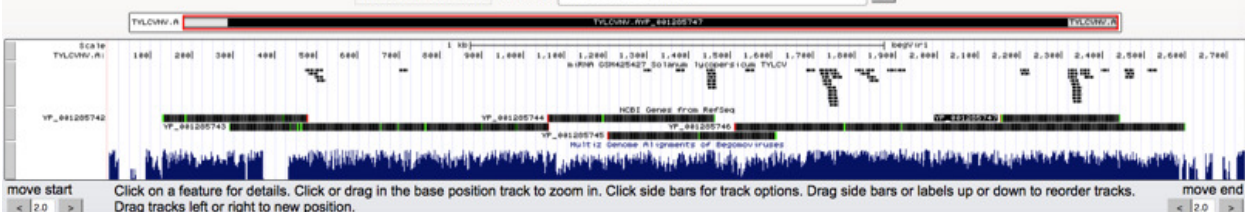
Find Position

Begomovirus Assembly

July 2016

Position/Search Term

Enter position, gene symbol or search terms
Current position: TYLCV.A

GO ▶

Begomovirus Genome Browser - begVir1 assembly   view sequences

Begomovirus constitutes the largest genus in the *Geminiviridae* family. It infects monocotyledons and dicotyledons including staple food worldwide, causing severe damages to agriculture especially in developing countries. Begomovirus is further classified into monopartite and bipartite depending on whether the virion encapsulates one or two circular ssDNA genomes. The genome that shares similar gene architecture between monopartite and bipartite viruses is commonly called DNA-A. And the second genome of bipartite viruses is called DNA-B. Our database contains 139 and 113 bipartite and monopartite begomoviruses, respectively. DNA-A and DNA-B of a virus are labeled by its viral abbreviation suffixed with .A and .B, respectively. E.g. Genomes of Melochia yellow mosaic virus are represented by MelYMV.A and MelYMV.B in our database. For monopartite begomoviruses, their genomes are always suffixed with .A, e.g. genome of Tomato leaf curl Hainan virus is represented by ToLCHaiV.A.

Begomovirus

B

GB4GV Genome Browser on Begomovirus July 2016 Assembly (begVir1)

move <<< << < > >> >>>   zoom in 1.5x 3x 10x base   zoom out 1.5x 3x 10x 100x

TYLCVNV.A:1-2,745 2,745 bp.   enter position or search terms   go

move start   Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks.   move end
< 2.0 >   Drag tracks left or right to new position.   < 2.0 >

default tracks   default order   hide all   add custom tracks   track hubs   configure   multi-region   reverse   resize   refresh

collapse all   Use drop-down controls below and press refresh to alter tracks displayed.   expand all
                    Tracks with lots of items will automatically be displayed in more compact modes.

| − | **Mapping and Sequencing** | refresh |

Base Position   CR left   CR Right   Iterons and TATA   miRNA   miRNA
dense   hide   hide   hide   GSM425427   Nb10A_A
                                                                              squish   hide

miRNA   miRNA   miRNA   Short Match
Nb10Abeta_A   Nb10Ambeta_A   miRNA SI10A_A   SI10Abeta_A   hide
hide   hide   hide   hide

| − | **Genes and Gene Predictions** | refresh |

NCBI Genes   UniProt   UniProt
pack   hide   Structure
                                     hide

| − | **Comparative Genomics** | refresh |

Multiz Align
squish

C

NCBI Genes from RefSeq (YP_001285746)

Position: TYLCVNV.A:1512-2600
Genomic Size: 1089
Strand: -

Links to sequence:

- Translated Protein from predicted mRNA
- Predicted mRNA from genomic sequences
- Genomic Sequence from assembly

Gene Description

Rep protein of Old World monopartite Tomato yellow leaf curl Vietnam virus DNA-A, complete sequence (TYLCVNV.A) originated from VIETNAM. It was found to infect host "TOMATO (LYCOPERSICON ESCULENTUM, SOLANACEAE); CROP". NC_009548 or DQ641697 [Provided by NCBI]

View table schema

329

330   Figure 3. Web interface of gb4gv. (A) The home page of gb4gv. The evolutionary tree on the
331   left side of the page shows the available viruses including the three genera of
332   Geminiviruses and two satellites. Users can view a particular genus or satellite by clicking
333   on the virus or satellite name in the evolutionary tree. Users can also make use of the
334   "Species Search" box (above the tree) to look up for a particular virus by keywords.
335   Additionally, users can enter keywords in the "Position/Search Term" box search for a
336   particular virus and/or gene. Click the blue GO button to navigate into the genomic
337   information of a particular viral species. (B) The page at the middle level provides various
338   annotation information about the selected genome in which they are organized in tracks.
339   (C) Information of a protein-coding gene. It tells the genomic location of the gene, size, and
340   strand that codes for the protein. In addition, there is a short description about the current
341   gene including the name of the protein, whether the virus is a New World or an Old World
342   virus, the full name of the virus and its acronym, the RefSeq and GenBank accession
343   numbers with hyperlink linked to the corresponding GenBank entry in NCBI website.
344

345   In the following subsections, we will highlight the unique features offered by gb4gv

346   that are helpful in studying the genomics of geminiviruses. While the software architecture

347   of gb4gv is based on Genome Browser, the operations of our website is highly similar to

348   UCSC Genome Browser. Therefore we will not discuss the data models and functionalities

349   of Genome Browser in details. For readers who are interested in learning more about

350   Genome Browser, we recommend that they consult the online User Guide (UCSC Genome

351   Browser User Guide).

352

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

354   To our best knowledge, there is no database that allows users to search for geminivirus

355   genomes or proteins by acronym, host name, geographical location, monopartite, bipartite,

356   Old world, New world, or combinations thereof. For instance, a search for monopartite

357   begomoviruses that infect Okra by the query "monopartite okra" against NCBI RefSeq

358   database returned only two entries: NC_005954 and NC_005051 and both of them belong

359   to satellite genomes. In fact, four monopartite begomoviruses are known to infect Okra

360 according to gb4gv: OLCCV (NC_014745), OYCrV (NC_008377), OYVMV (NC_004673), and

361 OkLCuV (NC_013017). The main reason is because NCBI's query matches only words in the

362 description of GenBank entries. Our augmented search capability will help researchers in

363 identifying a regime of viruses that share certain attributes handily. gb4gv achieves this by

364 making the above viral attributes searchable in our database in conjunction with the

365 keyword searching capability provided by Genome Browser. Table 2 below summarizes the

366 searchable attributes supported by gb4gv.

367 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 begomoviruses, 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 |

368

369   For instance, to find all begomoviruses that infect sweet potato, user can input the phrase

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

A

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

375   User can combine multiple search attributes in a query. The logical AND relationship

376   is assumed between attributes. For example, user can enter "new world monopartite" in

377   the search box to search for all New World monopartite begomoviruses (Figure 4B). But

378   the current version of the search function remains primitive as it is virtually inherited from

379   the 'LIKE' search of MySQL, meaning that the order of queried attributes is important.

380   When multiple attributes are specified, they must be arranged according to the order

381   enlisted in Table 2 from top to bottom. For the same example above, the query

382   "monopartite new world" will result in no hits.

383

384   *Short Match*

385   The ability to support ad-hoc sequence search can help researchers to identify potential

386   short regulatory sequences that can be validated further by experiment. Examples of these

387   regulatory sequences include TATA box (Sanz-Burgos & Gutierrez 1998), and

388   polyadenylation signal `AWTAAA` (`W` means `A` or `T`). The Short Match function allows users to

389   search for DNA sequences from 2 to 30 bases with the support of IUPAC ambiguity codes.

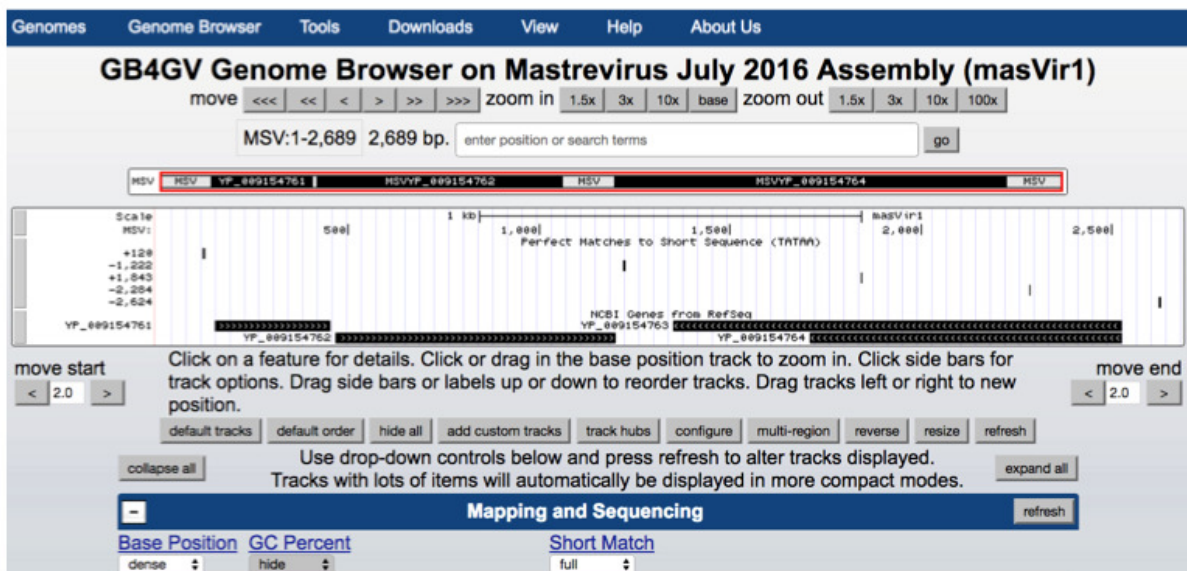
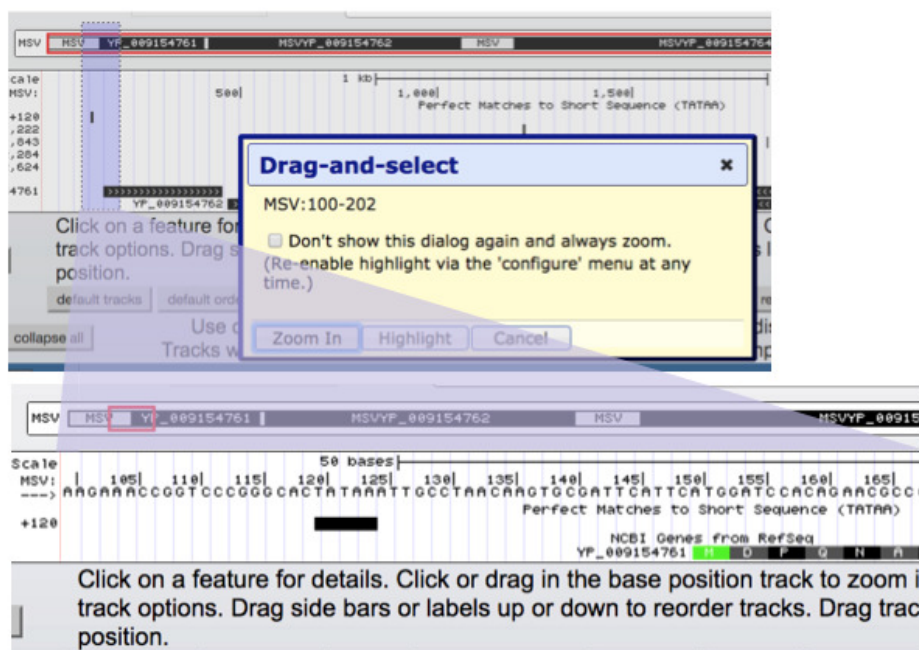
390   Figure 5 illustrates how to specify a short sequence match, and how to inspect the context

391   of a hit within a specific region through the Genome Browser's zoom-in function.

A

B

C

392

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

400

401   *Putative Iterons and TATA*

402   It has been known iterons contributed to viral replication (Arguello-Astorga et al. 1994;

403   Sanz-Burgos & Gutierrez 1998). Studied had shown binding activities between REP and

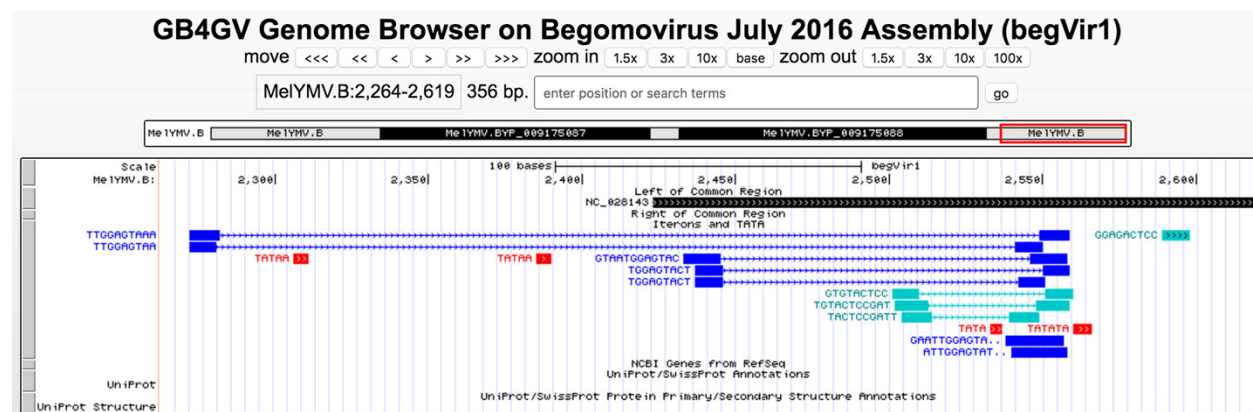
404   iterons in Mastrevirus and begomovirus (Fontes et al. 1992; Sanz-Burgos & Gutierrez

405   1998). gb4gv maintains 5,142 putative iterons and TATA box sites in the long intergenic

406   region. Users can view this information by turning on the "Iterons and TATA" track. Figure

407   6 shows an example of iterons and TATA box sites predicted in begomovirus Melochia

408   yellow mosaic virus (NC_028143).

409

410   Figure 6. Iterons and TATA track. Different colors are used to denote various
411   sequence features: direct repeats in blue, inverted repeats in blue-green, and TATA box in
412   red. Tandem repeats are highlighted with ".." at the end the label e.g. the direct tandem
413   repeats "GAATTGGAGTA.." above consists of "GAATTGGAGTATTGGAGTA" in which
414   "GAATTGGAGTA" overlaps with "GaATTGGAGTA" with their overlapping regions
415   underlined. Lastly, our database also highlights palindromic-like sequence by
416   ">>>...>>>", e.g. "GGAGACTCC".

417

418   *Small Interfering RNAs*

419   Understanding plant immunity is the foremost step to fight against viral infection. Virus-

420   derived RNA silencing is a vital immune response triggered in plants in the face of viral

421   infection. Thus we have incorporated datasets from two virus-derived small interfering

422   RNA (siRNA) studies into gb4gv. One study used pyrosequencing to sequence siRNAs in

423   tomato leaves (*Solanum lycopersicum*) inoculated with monopartite begomovirus TYLCV

424   (Donaire et al. 2009). Another study had used deep sequencing to survey siRNAs in the

425   leaves of tomato (*Solanum lycopersicum*) and tobacco (*Nicotiana benthamiana*) inoculated

426   with monopartite begomovirus and its associated betasatellite (TYLCCNV/TYLCCNB) (Yang

427   et al. 2011). Both studies had mapped the siRNAs to the genomes of respective hosts.

428   However, it is unclear whether or not these siRNA sequences are species specific. Are

429   siRNAs mapped to biased locations? In order to answer these questions, we incorporated

430   siRNA sequences from these two studies into gb4gv and mapped the siRNAs to genomes of

431   begomovirus and betasatellites. Each sample occupies a track (Figure 7A).

A

B

C

432

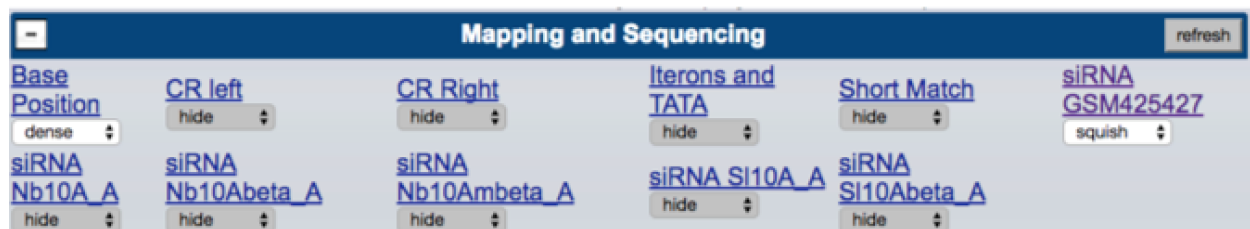
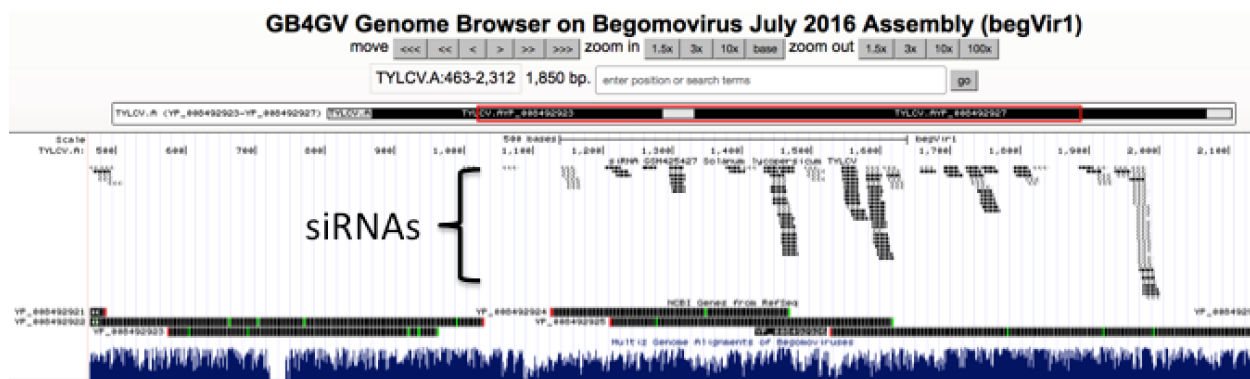
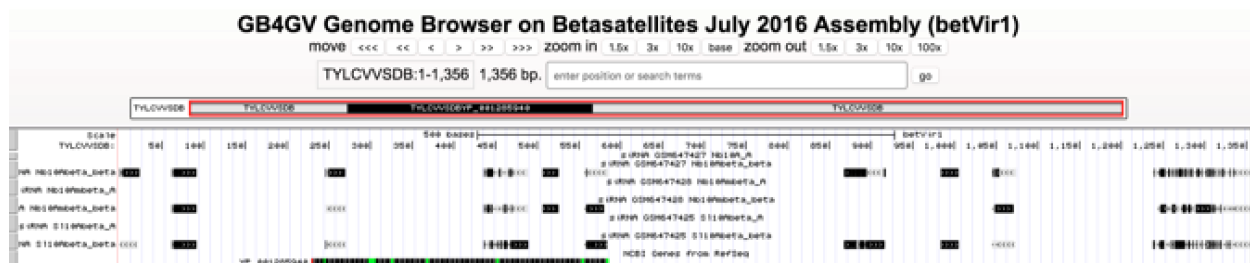



Figure 7. siRNA mapping. (A) Six samples of siRNA sequences are available for
visualization, one track for each sample. (B) An example to visualize the mapping of siRNAs
from GSM425427 on monopartite begomovirus TYLCV. 'Squish' mode was used in this
example. (C) Another example to show the appearance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

438

439   siRNAs mapped to the viral strand and complementary strand are encoded in dark

440   and light color, respectively (Figure 7B). According to our limited browsing, siRNAs do not

441   map uniformly along the genome. In betasatellites, a sizeable number of mapped siRNAs

442   were skewed toward a 100-bp region near to the 5' side of the SCE.

443

444 *BLAT*

445 Our database is also equipped with a lightweight sequence query engine BLAT (Kent 2002).

446 BLAT stands for BLAST-like alignment tool. It has been widely used to search for highly

447 similar gapped alignments. In situation like the detection of exons based on a spliced mRNA

448 sequence, BLAT provides a speedy mapping of the query sequence onto the genome. Major

449 differences between BLAT and Short Match are:

450 1. The minimum and maximum query lengths for BLAT are 20 and 25,000 bps,

451   respectively.

452 2. BLAT searches against genomes in a database specified by the user. Whereas Short

453   Match searches for queried sequence only in the current active genome.

454 3. BLAT can handle gapped hit but not for Short Match.

455

456 As an illustration, we used an unusually long (46 bps) iteron sequence

457 "TGAGTGATTTCTTATTATGTGATTGTCCATTAAAGGGATAAAGTGA" (Figure 8A) found in

458 YOM (Cotton leaf curl virus betasatellite NC_017829) to query against betasatellite

459 genomes. Intriguingly, eight other betasatellite genomes were found to contain sequences

460 that share from 88.5% to 97.9% of identity with the queried sequence (Figure 8B). To

461 further examine the hit in virus LPALDDBV, we clicked the "browser" link on the left, which

462 led to Figure 8C. It shows that the queried sequence hits a region LPALDDBV clustered with

463 iterons. The solid grey bar at the bottom indicates that YOM and LPALDDBV differ at only

464 eight sites.

465

466

467 Figure 8. BLAT search. (A) Sequence search box. (B) List of hits with a column of hyperlinks
468 referencing the hit genomes on the far left. (C). Display of hit in the context of targeted
469 genome by clicking the "browser" link on the left of panel (B). The grey bar at the bottom
470 shows the queried sequence is aligned to the opposite strand of the genome. (If the queried
471 sequence aligns to the reference strand, the bar will be displayed in black.) Nucleotides
472 displayed on the grey bar represent mismatched nucleotides between the queried
473 sequence and the targeted genome.
474

475   **Conclusion**

476   Genomics visualization is a useful approach to enhance interpretation especially when the

477   quantity or diversity of the viral genomic data is large. We have harnessed the capability of

478   the widely acclaimed Genome Browser specially for the geminivirus research community.

479   Instead of using a generic one size fits all approach to organize viral genomes, we have

480   taken a semi-automatic pipeline to preserve the unique characteristics of geminiviruses in

481   our web-based database gb4gv. Additionally, we have augmented keyword search

482   capability of manually curated attributes such as infecting hosts, geographical location etc.

483   However, further improvement is needed to accommodate even more flexible multiple

484   attributes queries. Moreover, we have predicted 127 pairs of common regions pertaining to

485   bipartite begomoviruses. This is a useful piece of information as the common regions are

486   implicated in coupling the two main genomes of a bipartite begomovirus during

487   encapsidation. As the ultimate goal in studying the genomes of *Geminiviridae* family is to

488   understand the underlying genomic features that are suggested to promote its propagation,

489   we have developed our own method to unravel putative iterons and TATA box sites in the

490   5' side of the common region and they can be visualized readily with genomic features

491   flanking them. The iterons we predicted are longer (9-46 bps) than the iteron core motifs,

492   $GGN_1N_2N_3$, reported by this group (Arguello-Astorga & Ruiz-Medrano 2001). The presence

493   of sizable direct or inverted repeats in fast evolving ssDNA viruses like geminiviruses is

494   unusual, suggesting the existence of negative selective pressure although the biological

495   function of the region peripheral of the iteron core motifs remains largely unknown.

496   Geminiviruses are diverse and fast evolving. Facilitated by the ever-decreasing DNA

497   sequencing cost, we anticipate more viral genomes will be sequenced in the near future.

498   We are certainly committed to maintaining the information in gb4gv as up-to-date as

499   possible. Given the flexibility of the Genome Browser in accommodating new annotation

500   tracks, if more genome-wide experimental data is available in the future such as Chip-Seq,

501   it can be included into gb4gv readily without software modification as illustrated by the

502   siRNA tracks discussed above. While viral regulatory elements play crucial roles in

503   influencing replication and transcription in cellular environment, we will continue our

504   effort in developing new methods to identify essential sequence elements that might offer

505   new insights for experimental virologists to design effective modalities to fight against the

506   infection of geminiviruses.

507 Acknowledgements

508   We thank the UCSC Genome Browser team for their technical support especially to

509   Maximilian Haeussler, Matthew Speir, and Cath Tyner. gb4gv would not be possible

510   without their kind support.

511 Funding

512   This project was supported by the startup fund provided by Lafayette College to E.S.H.

513   C.M.N. was supported by the EXCEL Summer Scholar Program funded by Lafayette College.

514   There was no additional external funding received for this study.

515 References

516   Arguello-Astorga GR, Guevara-Gonzalez RG, Herrera-Estrella LR, and Rivera-Bustamante
517   RF. 1994. Geminivirus replication origins have a group-specific organization of
518   iterative elements: a model for replication. *Virology* 203:90-100.
519   10.1006/viro.1994.1458
520   Arguello-Astorga GR, and Ruiz-Medrano R. 2001. An iteron-related domain is associated to
521   Motif 1 in the replication proteins of geminiviruses: identification of potential
522   interacting amino acid-base pairs by a comparative approach. *Arch Virol* 146:1465-
523   1485.

524   Bernard V, Brunaud V, and Lecharny A. 2010. TC-motifs at the TATA-box expected position
525   in plant genes: a novel class of motifs involved in the transcription regulation. *BMC*
526   *Genomics* 11:166. 10.1186/1471-2164-11-166
527   Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K,
528   Clawson H, Green ED, Haussler D, and Miller W. 2004. Aligning multiple genomic
529   sequences with the threaded blockset aligner. *Genome Res* 14:708-715.
530   10.1101/gr.1933104
531   Briddon RW, Mansoor S, Bedford ID, Pinner MS, Saunders K, Stanley J, Zafar Y, Malik KA,
532   and Markham PG. 2001. Identification of dna components required for induction of
533   cotton leaf curl disease. *Virology* 285:234-243. 10.1006/viro.2001.0949
534   Brown JK, Zerbini FM, Navas-Castillo J, Moriones E, Ramos-Sobrinho R, Silva JC, Fiallo-Olive
535   E, Briddon RW, Hernandez-Zepeda C, Idris A, Malathi VG, Martin DP, Rivera-
536   Bustamante R, Ueda S, and Varsani A. 2015. Revision of Begomovirus taxonomy
537   based on pairwise sequence comparisons. *Arch Virol* 160:1593-1619.
538   10.1007/s00705-015-2398-y
539   Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, and Llave C. 2009. Deep-
540   sequencing of plant viral small RNAs reveals effective and widespread targeting of
541   viral genomes. *Virology* 392:203-214. 10.1016/j.virol.2009.07.005
542   Duffy S, Shackelton LA, and Holmes EC. 2008. Rates of evolutionary change in viruses:
543   patterns and determinants. *Nat Rev Genet* 9:267-276. 10.1038/nrg2323
544   Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
545   throughput. *Nucleic Acids Res* 32:1792-1797. 10.1093/nar/gkh340
546   Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the*
547   *author Department of Genome Sciences, University of Washington, Seattle*.
548   Fontes EP, Luckow VA, and Hanley-Bowdoin L. 1992. A geminivirus replication protein is a
549   sequence-specific DNA binding protein. *Plant Cell* 4:597-608.
550   GenBank Record. NCBI GenBank Record. *Available at*
551   https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html (accessed 27 December
552   2016).
553   Gene Expression Omnibus. NCBI Gene Expression Omnibus. *Available at*
554   https://www.ncbi.nlm.nih.gov/geo/ (accessed 1 December 2016).
555   Gutierrez C. 1999. Geminivirus DNA replication. *Cell Mol Life Sci* 56:313-329.
556   ICTV. 2015. ICTV. *Available at* https://talk.ictvonline.org/files/master-species-lists/
557   (accessed 15 December 2016).
558   Jeske H. 2009. Geminiviruses. *Curr Top Microbiol Immunol* 331:185-226.
559   Jeske H, Lutgemeier M, and Preiss W. 2001. DNA forms indicate rolling circle and
560   recombination-dependent replication of Abutilon mosaic virus. *EMBO J* 20:6158-
561   6167. 10.1093/emboj/20.21.6158
562   Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* 12:656-664.
563   10.1101/gr.229202. Article published online before March 2002
564   Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. 2002.
565   The human genome browser at UCSC. *Genome Res* 12:996-1006.
566   10.1101/gr.229102. Article published online before print in May 2002
567   Kumar S, Stecher G, and Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis
568   Version 7.0 for Bigger Datasets. *Mol Biol Evol* 33:1870-1874.
569   10.1093/molbev/msw054

570   NCBI RefSeq. NCBI RefSeq. *Available at* ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/
571   (accessed 15 December 2016).
572   NCBI Viral Genomes. NCBI Viral Genomes. *Available at*
573   https://www.ncbi.nlm.nih.gov/genome/viruses/ (accessed 1 June 2016).
574   Orozco BM, and Hanley-Bowdoin L. 1996. A DNA structure is required for geminivirus
575   replication origin function. *J Virol* 70:148-158.
576   Patikoglou GA, Kim JL, Sun L, Yang SH, Kodadek T, and Burley SK. 1999. TATA element
577   recognition by the TATA box-binding protein has been conserved throughout
578   evolution. *Genes Dev* 13:3217-3230.
579   Pilartz M, and Jeske H. 2003. Mapping of abutilon mosaic geminivirus minichromosomes. *J
580   Virol* 77:10808-10818.
581   Sanz-Burgos AP, and Gutierrez C. 1998. Organization of the cis-acting element required for
582   wheat dwarf geminivirus DNA replication and visualization of a rep protein-DNA
583   complex. *Virology* 243:119-129. 10.1006/viro.1998.9037
584   Sattar MN, Kvarnheden A, Saeed M, and Briddon RW. 2013. Cotton leaf curl disease - an
585   emerging threat to cotton production worldwide. *J Gen Virol* 94:695-710.
586   10.1099/vir.0.049627-0
587   Saunders K, Norman A, Gucciardo S, and Stanley J. 2004. The DNA beta satellite component
588   associated with ageratum yellow vein disease encodes an essential pathogenicity
589   protein (betaC1). *Virology* 324:37-47. 10.1016/j.virol.2004.03.018
590   Scholthof KB, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, Hohn B, Saunders K,
591   Candresse T, Ahlquist P, Hemenway C, and Foster GD. 2011. Top 10 plant viruses in
592   molecular plant pathology. *Mol Plant Pathol* 12:938-954. 10.1111/j.1364-
593   3703.2011.00752.x
594   Shepherd DN, Martin DP, Van Der Walt E, Dent K, Varsani A, and Rybicki EP. 2010. Maize
595   streak virus: an old and complex 'emerging' pathogen. *Mol Plant Pathol* 11:1-12.
596   10.1111/j.1364-3703.2009.00568.x
597   UCSC Admin. UCSC Genome Browser Download. *Available at*
598   http://hgdownload.soe.ucsc.edu/admin/ (accessed 1 June 2016).
599   UCSC GB Statistics. UCSC Genome Browser Statistics. *Available at*
600   http://genome.ucsc.edu/admin/stats/ (accessed 15 December 2016).
601   UCSC Genome Browser. UCSC Genome Browser. *Available at* http://genome.ucsc.edu.
602   UCSC Genome Browser User Guide. UCSC Genome Browser User Guide. *Available at*
603   https://genome.ucsc.edu/goldenpath/help/hgTracksHelp.html (accessed 15
604   December 2016).
605   UniProtKB. UniProt Knowledgebase. *Available at* http://www.uniprot.org/downloads
606   (accessed 1 June 2016).
607   Xie Y, Wu P, Liu P, Gong H, and Zhou X. 2010. Characterization of alphasatellites associated
608   with monopartite begomovirus/betasatellite complexes in Yunnan, China. *Virol J*
609   7:178. 10.1186/1743-422X-7-178
610   Yang X, Wang Y, Guo W, Xie Y, Xie Q, Fan L, and Zhou X. 2011. Characterization of small
611   interfering RNAs derived from the geminivirus/betasatellite complex using deep
612   sequencing. *PLoS One* 6:e16928. 10.1371/journal.pone.0016928
613   Zhou X, Xie Y, Tao X, Zhang Z, Li Z, and Fauquet CM. 2003. Characterization of DNAbeta
614   associated with begomoviruses in China and evidence for co-evolution with their
615   cognate viral DNA-A. *J Gen Virol* 84:237-247. 10.1099/vir.0.18608-0

616