Full Title

Interactive visualization of whole eukaryote genome alignments using NCBI's Comparative Genome Viewer (CGV)

Short Title

Visualization of genome alignment in NCBI CGV

Authors

Sanjida H Rangwala*, Dmitry V Rudnev, Victor V Ananiev, Andrea Asztalos, Barrett Benica, Evgeny A Borodin, Nathan Bouk, Vladislav I Evgeniev, Vamsi K Kodali, Vadim Lotov, Eyal Mozes, Dong-Ha Oh, Marina V Omelchenko, Sofya Savkina, Ekaterina Sukharnikov, Joël Virothaisakun, Terence D. Murphy, Kim D Pruitt, Valerie A. Schneider

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA.

* Corresponding author

E-mail: rangwala@nih.gov

Abstract

We report a new visualization tool for analysis of whole genome assembly-assembly alignments, the Comparative Genome Viewer (CGV) (https://ncbi.nlm.nih.gov/genome/cgv/). CGV visualizes pairwise same-species and cross-species alignments provided by NCBI using assembly alignment algorithms developed by us and others. Researchers can examine the alignments between the two assemblies using two alternate views: a chromosome ideogram-based view or a 2D genome dotplot. Whole genome alignment views expose large structural differences spanning chromosomes, such as inversions or translocations. Users can also navigate to regions of interest, where they can detect and analyze smaller-scale deletions and rearrangements within specific chromosome or gene regions. RefSeq or user-provided gene annotation is displayed in the ideogram view where available. CGV currently provides approximately 700 alignments from over 300 animal, plant, and fungal species. CGV and related NCBI viewers are undergoing active development to further meet needs of the research community in comparative genome visualization.

36 ## Introduction

37 **Comparative genome visualization**

38 Comparative genomics leverages shared evolutionary histories among different species to
39 answer basic biological questions and understand the causes of disease. As sequencing costs
40 have dropped and assembly algorithms have improved, there has been tremendous growth in
41 the number of high-quality genome assemblies available in public archives, and the diversity in
42 the organisms they represent. These data now make it possible to use comparative genomics
43 approaches to explore more elements of biology and reveal the need for different types of
44 analysis tools to support this exploration. The NIH Comparative Genomics Resource (CGR)
45 maximizes the impact of eukaryotic research organisms and their genomic data to biomedical
46 research (1). CGR facilitates reliable comparative genomics analyses for all eukaryotic organisms
47 through community collaboration and a National Center for Biotechnology Information (NCBI)
48 genomics toolkit. As part of CGR, we have created the Comparative Genome Viewer (CGV), a
49 web-based visualization tool to facilitate comparative genomics research.

50

51 Graphical visualization of genomic data can illuminate relationships among different data types
52 and highlight differences and anomalies; for instance, areas of a genome that are depleted in
53 gene annotation, or have unusually high repeat content, or are more variable between species.
54 Interactive genome browsers have become particularly valuable in recent years in helping
55 biologists navigate large sequence datasets and more easily find genomic locations of interest
56 to their specific research interest question. These visualizations can display molecular data that

57 can help resolve competing hypotheses and also expose patterns that can spur additional
58 research questions.

59

60 Linear genome browsers, such as the Genome Data Viewer (GDV) at NCBI (2), the UCSC
61 Genomics Browser (3), and JBrowse (4), display molecular data as “tracks” laid out in parallel
62 and anchored on a single genome assembly. Users can navigate to different chromosome
63 regions and view gene annotation, repeats, and other types of data. Displays of known
64 sequence variants (e.g., dbVar or dbSNP data) can aid in the analysis of gene or sequence level
65 variation in a discrete genome region. While these browsers can be powerful tools to analyze
66 many different types of data from a single genome concurrently, they are more limited in their
67 ability to support comparative genome analysis. In particular, linear genome browsers cannot
68 easily show whether a genome region that appears evolutionarily conserved has been
69 rearranged (e.g., inverted or translocated) in one genome relative to another one, since the
70 data is only displayed relative to a single genomic region in a single genome at a time.

71

72 Different types of two-dimensional visualizations have been proposed to facilitate analysis of
73 larger scale genome structural differences between two or more genomes. These visuals
74 include two-dimensional line graphs (also known as dotplots) (5, 6), circular diagrams (i.e.,
75 Circos plots) (7), and linear genome browsers that stack one assembly on top of another (8-10).
76 Different types of visuals have advantages and disadvantages. Circos plots can show multiple
77 datasets in one graphic but can be visually challenging to interpret and usually do not support

78 views of sub-genomic regions. Dotplots can allow zooming to view chromosome or sub-
79 chromosome regions but cannot easily or elegantly display gene or other annotation in the
80 same visual. In order to better serve different research questions, some groups provide a choice
81 of multiple different types of visuals for genome comparisons (11-13).

82

83 **Genome comparison data**

84 Broadly, there are two types of whole genome comparison data that can be displayed in a
85 comparative genome visualization tool. The first type of data is locations of gene orthologs.
86 Orthology is typically determined using a protein homology-based method (e.g., BLASTP) in
87 consideration with local gene order conservation (14, 15). This type of data can lend itself to
88 straightforward "beads on string" visualizations that allows researchers to easily determine
89 how syntenic gene regions have evolved across related species (15-17).

90

91 The second type of comparison data is whole genome assembly alignments (e.g., Mauve (18),
92 LASTZ (19)), which are sequence-based and include both genic and intergenic regions. Whole
93 genome alignments can be much more complex than simple gene ortholog locations but have
94 the advantage of including alignments in regulatory regions and other regions not annotated as
95 genes.

96

97 Here we introduce a new viewer tool at NCBI, the Comparative Genome Viewer (CGV), that is a
98 key element of CGR. The main view of CGV takes the "stacked linear browser" approach —
99 chromosomes from two assemblies are laid out horizontally with colored bands connecting

100 regions of sequence alignment. Initial usability research with conceptual prototypes revealed
101 that this type of visual was the easiest to interpret for scientists from a broad range of research
102 expertise in genomics. We display whole genome pairwise assembly-assembly alignments in
103 CGV. These sequence-based alignments can be used to analyze gene synteny conservation but
104 can also expose similarities in regions outside known genes e.g., ultraconserved regions that
105 may be involved in gene regulation. Because CGV is a web-based application, researchers do
106 not have to install or configure software or generate their own comparison files before they can
107 begin using it for their research. Below we describe some of the features of CGV and provide
108 examples of how visualization in this tool can generate insights into genome structure and
109 evolution.
110

111 Results

112 Overview of CGV

113 We developed a web application, the Comparative Genome Viewer (CGV)
114 (https://ncbi.nlm.nih.gov/genome/cgv/), to aid in comparing genome structures between two
115 eukaryotic assemblies. CGV facilitates analyses of genome variation and evolution between
116 different strains or species or strains, as well as evaluation of assembly quality between older
117 and newer assemblies from same species.
118

119 Alignments are generated at NCBI using BLAST (20) or LASTZ-based algorithms (19), or imported
120 from the UCSC Genomics Institute (https://hgdownload.soe.ucsc.edu/downloads.html) or from
121 other research groups (e.g., HPRC, https://humanpangenome.org/). Shorter alignments are

122 merged where possible; however, because of repeats and gaps in the alignments, even very
123 similar chromosomes may be broken down into many alignment segments. Refer to **Materials**
124 **and Methods** for more detail on how we generate whole-genome assembly alignments and
125 load them into the viewer.
126
127 The CGV home page provides a menu where users can select from available species and
128 assembly combinations (Fig 1A). We frequently add new alignments as high-profile assemblies
129 become available and in response to requests from the scientific community. As of October
130 2023, we provided a selection of almost 700 alignments from over 300 eukaryotic species (Fig
131 1B).
132
133 **Fig 1. Overview of CGV.**
134 (A) CGV selection menu. (B) Taxonomic distribution of species represented by alignments in CGV.
135 (C) CGV ideogram view of whole genome assembly alignment. Buttons in the lower right provide download access
136 for complete alignment data or an SVG image of the current alignment view.
137 https://www.ncbi.nlm.nih.gov/genome/cgv/browse/GCF_015227675.2/GCF_000001635.27/27835/10116 (D) CGV
138 zoomed to a chromosome-by-chromosome view, with information box shown. i. flip orientation ii. zoom in/out iii.
139 pan left/right iv. view assembly in Genome Data Viewer (GDV) v. information panel.
140 https://www.ncbi.nlm.nih.gov/genome/cgv/browse/GCF_015227675.2/GCF_000001635.27/27835/10116#NC_05
141 1343.1:44657546-50471623/NC_000075.7:41365809-52285446/size=10000
142 (E) CGV search interface and sample search results. (F) Filtering and configure options for ideogram view.
143
144 CGV's main view (the "ideogram view") displays pairwise alignments as colored connectors
145 linking the chromosomes in the two assemblies (Fig 1C). The view is filtered by default to show

146 only reciprocal best hits between assemblies to facilitate the analysis of orthologous genomic
147 regions. The researcher can choose to show the non-best placed alignments to expose close
148 sequence duplications or homologues (Fig 1F). Users of CGV can also filter alignments in view
149 by size (e.g., to only show large alignments blocks) or by orientation (e.g., to only show regions
150 that have undergone a potential inversion). The complete alignment data in GFF3 and human-
151 readable formats like XLSX can be downloaded from the viewer for a researcher's own use.
152
153 Users can select a chromosome from each assembly to zoom to the alignments for this
154 chromosome. They can navigate further within this chromosome comparison using the zoom
155 in/out and pan buttons or by using the mouse to pinch-zoom or drag to pan. Users can zoom
156 directly to a particular region of a chromosome by dragging their cursor over the coordinate
157 ruler or the ideogram for either assembly. Double-clicking on a selected alignment segment will
158 synchronously zoom both the top and bottom assembly on the aligned coordinates, so that
159 they are stacked on top of one another (Fig 1D).
160
161 Where available, RefSeq or assembly-submitter provided gene annotation is displayed on the
162 chromosomes (Fig 1D). Similarities in gene order denote regions of synteny, while discrepancies
163 can point to evolutionarily or biologically significant differences, or assembly errors, particularly
164 if evaluating different assemblies from the same species. Researchers can use the search
165 feature in CGV to find their gene of interest by name or keyword, and subsequently navigate to
166 the location of the gene in the viewer (Fig 1E). If the gene region is aligned, the viewer will
167 simultaneously navigate to the aligned location, which may contain the gene's known or

168 putative ortholog on the second assembly. The “flip” button allows the user to reverse one
169 chromosome to see inverted alignments displayed in the same relative orientation, which may
170 aid in the detection of discrepancies in gene annotation in regions that are locally syntenic
171 between the two assemblies. Once a user has completed their analysis of a region of interest,
172 they can export the image as an SVG to adapt for use in publications and presentations.
173
174 We provide additional information for each alignment segment in a pop-over panel (Fig 1D).
175 This panel reports the chromosome scaffold accession and sequence coordinates of the
176 alignment on each assembly, as well as the percent identity, number of gaps and mismatches,
177 and alignment length. While the ideogram view in CGV does not display specific nucleotide
178 bases, users can open another panel from the right-click menu that shows the alignment
179 sequence. They can also download the alignment FASTA file of a particular alignment segment
180 for downstream analysis, such as BLAST search or primer design. Moreover, researchers can
181 also navigate from CGV to NCBI’s genome browser, the Genome Data Viewer (GDV) (2). GDV
182 can display the assembly-alignment data viewed in CGV as a linear track alongside additional
183 data mapped onto a genome assembly, such as detailed transcript and CDS annotation,
184 repeats, GC content, variation data, or user-provided annotations. Zooming to a location within
185 GDV can reveal granular differences in nucleotide sequence or gene exon or CDS annotation
186 between the two assemblies.
187
188 In addition to the main ideogram-based view, the Comparative Genome Viewer also provides a
189 two-dimensional dotplot view of the pairwise genome alignment (Fig 2A). The dotplot shows

190 aligned sequence locations in one assembly on the X-axis plotted against aligned locations on
191 the second assembly on the Y-axis. Alignments in the reverse orientation are plotted with an
192 opposite slope and in a different color (purple) than alignments in the same orientation (green),
193 making it easier to identify inversions and inverted translocations. The CGV dotplot shows both
194 reciprocal best-placed and non-best placed alignments. As a result, compared to the ideogram
195 view, this plot may more easily expose differences in copy number between two assemblies,
196 such as segmental duplications or differences in genome or chromosome ploidy. Users can
197 select a pair of chromosomes in the whole genome plot (i.e., a "cell" in the plot) and zoom into
198 them on a full screen, where smaller alignment segments may be more easily interpretable (Fig
199 2B). Once a researcher has discovered a chromosome pair of interest, they can navigate back to
200 the ideogram view to conduct even more fine-grained analysis, including examining gene
201 annotation and observing very short alignment segments that were beyond the resolution of
202 the dotplot.
203
204 **Fig 2. CGV dotplot view of *Xenopus laevis* and *Xenopus tropicalis* alignment.**
205 https://ncbi.nlm.nih.gov/genome/cgv/plot/GCF_000004195.4/GCF_017654675.1/38475/8355
206 (A) Full genome dotplot. (B) Dotplot of chromosome 8 of *Xenopus tropicalis* vs chromosome 8S of *Xenopus laevis*.
207
208 **Analysis using CGV: Conservation of linkage groups with local rearrangement of synteny**
209 CGV can aid in detecting unusual patterns in genome evolution in different taxa. Researchers
210 had previously observed that genomes from Drosophila species conserve gene content within
211 linkage groups, known as Muller elements, corresponding to chromosomes or large sub-
212 chromosome regions. Within these Muller elements, gene order can be reshuffled extensively

213 in one species relative to another (21). For alignment between *D albomicans* vs *D*
214 *melanogaster*, the CGV ideogram view shows restriction of alignment from each chromosome
215 in one genome to a particular chromosome or chromosome region (i.e., linkage group) in the
216 other genome (Fig 3A). However, within a chromosome-chromosome pair, sequence alignment
217 is broken into many small fragments whose relative order is not conserved. This fragmentation
218 of alignment is more clearly visible in the CGV dotplot, which shows that pairwise alignments
219 are restricted to a single chromosome pair, but appear in a scattered pattern, suggesting that
220 the sequence and gene order has been extensively rearranged within chromosomes (Fig 3B).
221
222 **Fig 3. CGV shows conservation of linkage groups in the absence of conservation of gene order.**
223 (A) CGV ideogram view of alignment between *Drosophila albomicans* and *Drosophila melanogaster* genomes.
224 Alignments are restricted to a single chromosome or chromosome region.
225 https://ncbi.nlm.nih.gov/genome/cgv/browse/GCF_009650485.2/GCF_000001215.4/40865/0 (B)CGV dotplot view
226 of alignment between *Drosophila albomicans* and *Drosophila melanogaster* demonstrates that sequence order is
227 "scrambled" within linkage groups, as demonstrated by a scatter pattern indicating many short rearranged
228 alignments. https://www.ncbi.nlm.nih.gov/genome/cgv/plot/GCF_009650485.2/GCF_000001215.4/40865/0 (C)
229 CGV dotplot view of alignment between starfish species *Luida sarsii* and *Asteria rubens* with similar scatter pattern
230 to Drosophila alignments.
231 https://www.ncbi.nlm.nih.gov/genome/cgv/plot/GCA_949987565.1/GCF_902459465.1/41045/0 (D) CGV
232 ideogram view of alignment between chromosome 1 of *Luida sarsii* and chromosome 1 of *Asteria rubens*. These
233 chromosomes align to each other across their length, but the alignment is broken into multiple short segments
234 which are extensively rearranged.
235 https://www.ncbi.nlm.nih.gov/genome/cgv/browse/GCA_949987565.1/GCF_902459465.1/41045/0#OX465101.1/
236 NC_047062.1/size=1,firstpass=0 (E) CGV dotplot view of alignment between starfish species *Luida sarsii* and
237 *Patiria pectinifera* with similar scatter pattern to Drosophila alignments.

238 https://www.ncbi.nlm.nih.gov/genome/cgv/plot/GCA_949987565.1/GCA_029964075.1/41165/0 (F) CGV dotplot
239 view of alignment between starfish species *Plazaster borealis* and *Pisaster ochraceus*. Alignments show less scatter
240 and more of a diagonal slope, indicating more conservation of sequence order between these two species'
241 genomes. https://www.ncbi.nlm.nih.gov/genome/cgv/plot/GCA_021014325.1/GCA_010994315.2/41175/466999
242

243 We observed a similar pattern when comparing some genomes from different starfish species
244 using CGV. When looking at pairwise alignments between *Asterias rubens*, *Patiria pectinifera*,
245 and *Luida sarsii* species, sequences from a chromosome from one genome mainly align to a
246 single other chromosome in the other species. However, within a pairwise chromosome
247 alignment, the sequence order is rearranged, resulting in a scatter pattern in the dotplot (Fig
248 3C, E). The ideogram view can show the alignment fragmentation and rearrangement in more
249 granular detail (Fig 3D). We also noted that some starfish pairs show more conservation of
250 location synteny (22) (Fig 3F), consistent with measured sequence distance based on shared k-
251 mers (Mash distance is 0.128 compared to other pairs with a Mash distance >0.3).

252

253 Bhutkar et al (21) speculated that the need to keep certain genes in the same regulatory
254 environment may result in conservation of genes within linkage groups, even in the absence of
255 selective pressure to maintain the gene order. More recently, conservation of macrosynteny
256 with extensive small-scale sequence rearrangement was also detected in comparisons between
257 other invertebrate species, such as cephalopods, cnidarians, jellies, and sponges (16, 23, 24).
258 These rearrangements were used to parse the phylogenetic relationships within this clade. We

259 demonstrate here that CGV can aid researchers in detecting and analyzing this phenomenon in
260 starfish and other evolutionarily varied taxa.
261
262 **Analysis using CGV: Detection of amylase family expansion**
263 CGV can uncover potential copy number differences in segmental gene families. These
264 differences may appear as gaps in the alignment in otherwise syntenic gene regions. Segmental
265 insertions or deletions may be too small to be apparent on the whole genome or whole
266 chromosome alignment but can be detected when searching and navigating to a gene of
267 interest.
268
269 Initial analysis of the complete human telomere-to-telomere CHM13 genome revealed an
270 expansion of amylase genes on chromosome 1 compared to the GRCh38 reference assembly
271 (25). This expansion can be validated in CGV: a search for 'amylase' in the alignment between
272 GRCh38 and T2T-CHM13v2 assembly finds six matches to this gene name in the GRCh38 and
273 twelve in the CHM13 assembly (Fig 4A). Zooming out in the region of the *AMY1A* gene on
274 chromosome 1 reveals a nearby sequence segment in the CHM13 assembly that is not aligned
275 to GRCh38 (Fig 4B). This region contains numerous annotated loci that lack official
276 nomenclature (i.e., LOC); six of these loci are described as 'alpha-amylase'. Therefore, there are
277 at least six additional alpha-amylase genes in CHM13 genome compared to the GRC reference.
278 It is possible that copy number of this gene is variable in humans; it is also possible that the GRC
279 reference genome represents fewer than the typical number of gene copies. While these two
280 human assemblies are likely to be high quality, other assemblies in other species may be of

281 lower quality, and differences in copy number observed in CGV may reflect assembly or

282 annotation errors.

283

284 **Fig 4**. **CGV can help uncover gene duplications and rearrangements in closely related genomes.**

285 (A) Gene search of an alignment between two human assemblies in CGV finds twelve amylase gene family

286 members in the human T2T-CHM113v2.0 assembly and six amylase gene family members in GRCh38.p14. (B) CGV

287 view showing that T2T-CHM13v2.0 contains an insertion relative to GRCh38.p14, which appears as an unaligned

288 region on chromosome 1. This insertion contains additional alpha-amylase family members. A popup label (tooltip)

289 indicates one of these additional family members.

290 https://www.ncbi.nlm.nih.gov/genome/cgv/browse/GCF_009914755.1/GCF_000001405.40/23025/9606#NC_060

291 925.1:103415704-103764412/NC_000001.11:103566852-103915505/size=1000,firstpass=0 (C) CGV view showing

292 that chromosome 2 of *Canis lupus familiaris* (dog) UMICH_Zoey_3.1 align to chromosomes 2, 15, 23, and 25 of

293 Dog10K_Boxer_Tasha.

294 https://ncbi.nlm.nih.gov/genome/cgv/browse/GCF_005444595.1/GCF_000002285.5/17685/9615#NC_049262.1:6

295 542815-78714085//size=10000 (D) UMICH_Zoey_3.1 assembly chromosome 2 alignment to Dog10K_Boxer_Tasha

296 chromosome 25 contains the *MALRD1* gene, which is annotated as *LOC608668* in the Tasha assembly (boxed in

297 red). Gene synteny is not conserved outside of the region of assembly-assembly alignment.

298

299 **Analysis using CGV: Possible gene translocation between two dog assemblies**

300 For closely related strains or species, CGV can help uncover and validate structural anomalies,

301 such as where gene order synteny has been disrupted. Visual inspection of the whole genome

302 CGV ideogram view of alignment between two dog genomes – the Great Dane Zoey

303 (UMICH_Zoey_3.1) and the boxer Tasha (Dog10K_Boxer_Tasha) – indicated a region that

304 aligned to chromosome 2 in the Zoey assembly and chromosome 25 in the Tasha assembly (Fig

305 4C). This region contains the *MALRD1* gene in the Zoey assembly, which is shown to align to a
306 *MALRD1* homolog annotated as *LOC608668* in Tasha (Fig 4D). CGV alignments indicate that
307 *LOC608668* is likely the Tasha *MALRD1* gene; there are no better alignments to *MALRD1*
308 detected in CGV or by an independent BLAST search of the Tasha genome.
309
310 Zooming out in the aligned region of *MALRD1* indicates that gene synteny is not conserved
311 outside of the *MALRD1* gene region (Fig 4D). It appears that this gene has been rearranged
312 from chromosome 2 on the Zoey assembly to chromosome 25 on Tasha. It is also possible that
313 the Tasha genome may have been misassembled in this region, and the *MALRD1* gene
314 sequence is properly situated on chromosome 2 within the otherwise conserved syntenic block.
315 A researcher would need to further examine the quality of the Tasha assembly in this region to
316 distinguish these possibilities, for instance, by examining the sequencing reads for the Tasha
317 assembly or viewing a HiC map of assembly structure.
318
319 If this anomaly represents a true difference between the two genomes, it could prove
320 biologically significant. The translocation of *MALRD1* may have placed it in a different gene
321 regulatory environment in the boxer (Tasha) genome, which could possibly result in different
322 levels or patterns of gene expression. The human ortholog of *MALRD1* was shown to be
323 involved in bile acid metabolism (26). This gene region was also genetically linked to
324 Alzheimer's disease (27). Therefore, if valid and not assembly artifacts, differences like these
325 could provide insight into human health.

326

327 ## Discussion

328 We describe here a new visualization tool for eukaryotic assembly-assembly alignments, the
329 Comparative Genome Viewer (CGV). We developed this web application with a view toward
330 serving both expert genome scientists as well as organismal biologists, students, and educators.
331 Users of CGV do not have to generate their own alignments or configure the software using
332 command line tools. Instead, they can select from our menu of available alignments, access a
333 view immediately in a web application, and start their analysis. We are continuing to add new
334 alignments regularly and invite researchers to contact us if assemblies or organisms of interest
335 are missing. We continue to do periodic outreach to the community to help us improve our
336 visual interfaces so that they are simple, intuitive, and accessible.

337

338 CGV displays whole genome sequence alignments provided by NCBI; users cannot currently
339 upload their own alignment data or choose assemblies to align in real time. There are both
340 technical and scientific considerations to allowing researchers to select and align assemblies
341 themselves. Currently, whole genome assembly-assembly alignments take several hours to
342 days, using up to one thousand CPU processing hours per pairwise alignment of larger
343 genomes, such as those for mammalian or plant assemblies. Moreover, whole genome
344 alignments are difficult to generate past a certain genetic distance (i.e., Mash > 0.3). Not only is
345 alignment between distantly related genomes computationally expensive, but the alignments
346 themselves may be of limited research value. These alignments will likely have sparse and short
347 segments that may correspond only to the most highly conserved coding sequence (CDS) (Fig 5;

348 S1 Table). We suggest protein similarity or gene orthology-based alignments as more
349 appropriate for comparisons between distantly related organisms. At present, we manually vet
350 requested alignments to make sure that assemblies are complete, of high assembly quality (e.g.
351 high scaffold N50 or BUSCO scores (28)), and a reasonable evolutionary distance. This review
352 ensures that alignments will be useful to both the original requester and others in the research
353 community.
354
355 **Fig 5. Genome and CDS coverage of assembly-assembly alignments relative to Mash distance.**

356 Percentages of total target genome or CDS nucleotides covered by the ungapped alignments are plotted against
357 the Mash distance between the pair of genomes. (A) Alignments between the human GRCh38.p14 assembly and
358 other vertebrates. (B) Alignments of fall army worm (FAW, *Spodoptera frugiperda*, a major insect agricultural pest)
359 and related insects. At lower Mash distances, the whole genome alignments cover most of the genome and CDS.
360 At Mash distances greater than 0.25 or 0.3, the alignment covers less than 20% of the genome overall, and
361 between 30% and 60% of the CDS. Refer to S1 Table for the data used to populate these graphs.

362 **S1 Table. Alignment coverage at different Mash distances for selected assembly pairs.**

363
364 Many research questions in comparative biology may be best answered by simultaneously
365 visualizing alignments among more than two assemblies. We are exploring user needs and
366 existing tools when it comes to multigenome alignment visualization, such as visualization of
367 pangenome data. Our lessons from developing CGV will prove valuable in this upcoming
368 initiative.
369

370 ## Materials and Methods

371

372 **Preparation of assembly-assembly alignments**

373 Genome assemblies are aligned using a two-phase pipeline first described in Steinberg et al
374 (29), with adaptations for cross-species alignments. In the first phase, initial alignments are
375 generated using BLAST (20) or LASTZ (19), or imported from a third-party source such as UCSC
376 (30). In the second phase, alignments are merged and ranked to distinguish reciprocal-best
377 alignments from additional alignments that are locally best on one assembly but not the other.
378 The resulting alignment set is omnidirectional and can be used to project information from
379 query to subject assembly or vice versa.

380

381 In the first phase, for both BLAST and LASTZ, repetitive sequences present in the query and
382 target assemblies are soft-masked using WindowMasker (31). The default parameters are
383 usually suitable for aligning assembly pairs within the same species. However, more aggressive
384 masking is required when aligning cross-species assemblies. The masking rate is adjusted with
385 the parameter `t_thres_pct` set to 99.5 (default, for BLAST same-species), 98.5 (for BLAST
386 cross-species), or 97.5 (LASTZ cross-species), or lower for some genomes with extensive and
387 diverse repeat composition. The 97.5 - 98.5 values typically result in a masking percentage
388 similar to RepeatMasker (http://www.repeatmasker.org) with a species-specific repeat library,
389 with the advantage of not needing to define repeat models beforehand.

390

391 Genome assemblies are aligned using either BLAST or LASTZ. The selection of the aligner and
392 specific parameters depends on the level of similarity between the assemblies. We use Mash

393 (32) to compute the approximate distance between two assemblies (Fig 5). BLAST is employed
394 for aligning pairs of assemblies belonging to the same species, as well as cross-species assembly
395 pairs with a Mash distance of less than 0.1. An exemplar BLAST command is:

```
396 blastn -evalue 0.0001 -gapextend 1 -gapopen 2 -max_target_seqs
397 250 -soft_masking true -task megablast -window_size 150 -
398 word_size 28
```

399 A BLAST `word_size` of 28 is used for pairs of assemblies with Mash distances below 0.05,
400 such as human and orangutan, while a `word_size` of 16 is used to enhance sensitivity for
401 more distant cross-species pairs with Mash distances ranging from 0.05 to 0.1, such as human
402 and rhesus macaque.
403
404 Assembly pairs with Mash distances exceeding 0.1, such as human and mouse assemblies, are
405 aligned using LASTZ. The `make_lastz_chains` pipeline (19) is employed to generate
406 alignments between query and target assemblies in UCSC chain format. The default parameters
407 are often adequate to produce satisfactory alignments for many assembly pairs, though some
408 distant assembly pairs (e.g., Mexican tetra-medaka) warrant changes such as the use of a
409 different substitution matrix (`BLASTZ_Q=HoxD55`).
410
411 Alignments generated with the `make_lastz_chains` pipeline or precomputed alignments
412 imported from UCSC are in chain format. The UCSC chainNet pipeline (33) is run on query x
413 target and target x query alignment sets separately so that the alignments are 'flattened' in a

414 way that the reference sequences are covered only once by the alignments in each set, and the
415 two chainNet outputs are concatenated.
416
417 In the second phase, alignments are converted to NCBI ASN.1 format and processed further for
418 the NCBI's CGV and GDV browsers. In this phase, the full set of BLAST or LASTZ-derived
419 alignments are processed to merge neighboring alignments and split and rank overlapping
420 alignments to identify a set of best alignments (S1A Figure). Merging is accomplished on a
421 sequence-pair-by-sequence-pair basis, and ranking is accomplished globally for the assembly
422 pair. The process is designed to find a dominant diagonal among a set of potentially conflicting
423 alignments.
424
425 **S1 Figure. Merging, sorting, and ranking assembly-assembly alignments.**
426 (A) Flowchart showing that adjacent alignment segments are merged. Subsequently, alignments are split once
427 again at large gaps. (B) Flowchart showing how overlapping alignments are separated, ranked, and re-merged.
428 Reciprocal best-placed alignments are designated as "first pass", while the non-best placed alignment is
429 designated "second pass".
430
431 Merging involves the following steps. First, when applicable, alignments based on common
432 underlying sequence components of the assemblies (e.g., the same BAC component used in
433 both human GRCh37 and GRCh38) are identified and merged into the longest and most
434 consistent stretches possible. Second, adjacent alignments are merged if there are no
435 conflicting alignments. Third, alignments are split on gaps using a default threshold of 50 bp (for

436 the same or closer species) or 50 kb (for more distant species), or longer than 5% of the
437 alignment length. Alignments are also split at any point where they intersect with overlapping
438 alignments (S1A Figure). Duplicate or low-quality alignments fully contained within higher-
439 quality alignments are dropped.

440

441 After merging and splitting, alignments are subsequently processed using a sorting and ranking
442 algorithm (S1B Figure). Alignments are sorted based on a series of properties, including the use
443 of common components, assembly level (alignment to chromosomes preferred over alignment
444 to unplaced scaffolds), total sequence identity, and alignment length. The alignments are then
445 scanned twice, once each on the query and subject sequence ranges, to sort out reciprocal
446 best-placed (also referred to as "first pass" or reciprocity=3) and non-best placed (also referred
447 to as "second pass" or reciprocity=1 or 2) alignment sets. Finally, all alignments in each
448 reciprocity are merged again to stitch together adjacent alignments with no conflicting
449 alignments into the longest representative stretches.

450

451 Assembly-assembly alignments are stored in an internal database available for rendering in
452 NCBI's CGV and GDV browsers. The alignment data are publicly available in GFF3 and ASN.1
453 formats at https://ftp.ncbi.nlm.nih.gov/pub/remap/.

454

455 For display in CGV, assembly alignment batches are filtered to keep only alignments that
456 contain chromosome scaffolds as both anchors and targets, since non-chromosomal scaffolds

are not displayed in this viewer. Alignments are converted into a compact binary format designed to keep only the synteny data required for display. This preparation step is done by programs written in C++ and bash scripts that tie them together.

Technical architecture of CGV

CGV operates on a two-tier model, with a front end implemented using HTML/JavaScript running in the user's web browser and a backend running at NCBI. The graphical rendering is done on the front end using modern WebGL technologies. The main advantage of this approach is speed and fluidity of the user interface since most of the alignment data needed to be rendered is sent to the front end at the initial load and there are no additional roundtrips to the server when the user interacts with the page (e.g., panning or zooming). Using front end graphical rendering also reduces the network traffic between NCBI and the end user, which makes CGV more responsive.

The backend of the CGV application resolves internal alignment identifiers to an alignment data file that the front end can use for generating graphical images. The back end is implemented as an industry-standard gRPC service written in C++ and running in a scalable NCBI service mesh (linkerd, namerd, consul). When a CGV view is initially loaded, our gRPC service requests the alignment data needed for the particular page. On graphical pages, gRPC resolves an alignment identifier to a URL with prepared synteny/alignment data and the page loads the file at this URL into a WASM module which is written in C++ and compiled with Emscripten. The WASM module

478 serves this data on demand to the page's JavaScript code which uses it for building the image
479 and all user interactions.
480
481 In parallel, the list of assembly-assembly alignments and their metadata is sent to the selection
482 menu (i.e., "Set up your view") on the CGV landing page. This allows the selection menu to
483 report scientific and common species names, assembly accessions, and assembly names.
484
485 The alignment selection menu on the CGV landing page is a traditional web form page. We
486 utilize the NCBI version of United States Web Design System (USWDS) design standards and
487 components (https://www.ncbi.nlm.nih.gov/style-guide) in order to unify graphical design with
488 other US government pages.
489
490 On the more graphically intensive ideogram and dotplot pages, graphical rendering is done in
491 the user's web browser using WebGL using d3.js or pixi.js libraries, which allows for efficient
492 interactivity, scalability, and fluidity of user interaction. Other elements of the page use
493 jQuery/extJS and USWDS components. CGV reuses chromosome ideograms initially developed
494 for NCBI's Genome Data Viewer (2).
495
496 Gene annotations shown in CGV are obtained from NCBI's public databases using NCBI Entrez
497 Programming Utilities (E-utilities) (https://www.ncbi.nlm.nih.gov/books/NBK25501/).

498 Annotation-build specific gene search is provided by an NCBI Datasets

499 (https://www.ncbi.nlm.nih.gov/datasets/) gRPC service.

500

501 **Design of CGV Application**

502 A philosophy of user-centered design, which puts user needs at the forefront of decision-

503 making, was an integral element in the development of the Comparative Genome Viewer

504 (CGV). Participants for user research were recruited from members of the genomics research

505 community who provided their contact information through feedback links on the CGV

506 application and other sequence analysis tools at NCBI. Some of these researchers were

507 previously familiar with CGV, while others had little or no experience with this tool. Data from

508 user research testing sessions was compiled and analyzed for patterns in behavior, thereby

509 allowing the team to validate that the design was moving in a direction that facilitated analysis

510 of sequence alignment data. To date, we have conducted user research with over 30 different

511 experts in the field of comparative genomics. We also evaluate the application for Section 508

512 compliance, which also helps insures CGV performs well on mobile devices and is accessible to

513 users with limited or no visibility.

514

515 ## Acknowledgments

516 We thank Anne Ketter and Emily W Davis for help with project planning and coordination. We

517 thank Guangfeng Song for aid with user experience research and analysis.

518 Thanks to Wayne Matten, Michelle Formica-Frizzi, and others in the NCBI customer service
519 team for marketing, webinars, and video tutorials. Anatoliy Kuznetsov and Andrei Shkeda
520 conducted early technical design work that influenced the architecture of the CGV application.
521 Deanna M. Church and Mike DiCuccio participated in initial design and testing of the assembly
522 alignment protocol. Finally, we thank members of the greater genomics research community
523 who have participated in usability sessions and provided feedback and alignment requests to
524 CGV. CGV was developed as a part of the National Institutes of Health's Comparative Genomics
525 Resource (CGR) (https://www.ncbi.nlm.nih.gov/comparative-genomics-resource/). This work
526 was supported by the National Center for Biotechnology Information of the National Library of
527 Medicine (NLM), National Institutes of Health.
528

529 ## References

530 1. Bornstein K, Gryan G, Chang ES, Marchler-Bauer A, Schneider VA. The NIH Comparative
531 Genomics Resource: addressing the promises and challenges of comparative genomics on human
532 health. BMC Genomics. 2023;24(1):575.
533 2. Rangwala SH, Kuznetsov A, Ananiev V, Asztalos A, Borodin E, Evgeniev V, et al. Accessing NCBI
534 data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). Genome Res. 2021;31(1):159-69.
535 3. Nassar LR, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome
536 Browser database: 2023 update. Nucleic Acids Res. 2023;51(D1):D1188-D95.
537 4. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web
538 platform for genome visualization and analysis. Genome Biol. 2016;17:66.
539 5. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple
540 way. PeerJ. 2018;6:e4958.

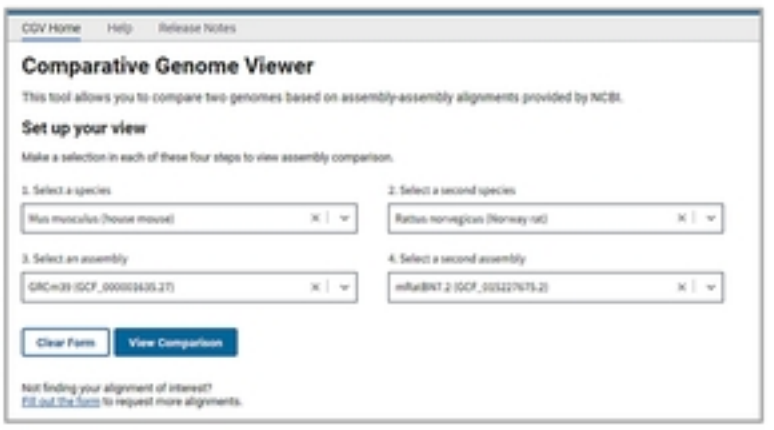
541 6. Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. SynMap2 and SynMap3D: web-
542 based whole-genome synteny browsers. Bioinformatics. 2017;33(14):2197-8.
543 7. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information
544 aesthetic for comparative genomics. Genome Res. 2009;19(9):1639-45.
545 8. Wang H, Su Y, Mackey AJ, Kraemer ET, Kissinger JC. SynView: a GBrowse-compatible approach
546 to visualizing comparative genome data. Bioinformatics. 2006;22(18):2308-9.
547 9. Xiao Z, Lam HM. ShinySyn: a Shiny/R application for the interactive visualization and integration
548 of macro- and micro-synteny data. Bioinformatics. 2022;38(18):4406-8.
549 10. Richardson JE, Baldarelli RM, Bult CJ. Multiple genome viewer (MGV): a new tool for
550 visualization and comparison of multiple annotated genomes. Mamm Genome. 2022;33(1):44-54.
551 11. Diesh C, Stevens GJ, Xie P, De Jesus Martinez T, Hershberg EA, Leung A, et al. JBrowse 2: a
552 modular genome browser with views of synteny and structural variation. Genome Biol. 2023;24(1):74.
553 12. Albert VA, Krabbenhoft TJ. Navigating the CoGe Online Software Suite for Polyploidy Research.
554 Methods Mol Biol. 2023;2545:19-45.
555 13. Lee J, Hong WY, Cho M, Sim M, Lee D, Ko Y, et al. Synteny Portal: a web-based application portal
556 for synteny block analysis. Nucleic Acids Res. 2016;44(W1):W35-40.
557 14. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and
558 evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49.
559 15. Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, et al. GENESPACE tracks
560 regions of interest and gene copy number variation across multiple genomes. Elife. 2022;11.
561 16. Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS. Ancient gene linkages
562 support ctenophores as sister to other animals. Nature. 2023;618(7963):110-7.
563 17. Muffato M, Louis A, Poisnel CE, Roest Crollius H. Genomicus: a database and a browser to study
564 gene synteny in modern and ancestral genomes. Bioinformatics. 2010;26(8):1119-21.

565 18. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic
566 sequence with rearrangements. Genome Res. 2004;14(7):1394-403.
567 19. Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, et al. Integrating gene
568 annotation with orthology inference at scale. Science. 2023;380(6643):eabn3107.
569 20. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J
570 Comput Biol. 2000;7(1-2):203-14.
571 21. Bhutkar A, Schaeffer SW, Russo SM, Xu M, Smith TF, Gelbart WM. Chromosomal rearrangement
572 inferred from comparisons of 12 Drosophila genomes. Genetics. 2008;179(3):1657-80.
573 22. Lee Y, Kim B, Jung J, Koh B, Jhang SY, Ban C, et al. Chromosome-level genome assembly of
574 Plazaster borealis sheds light on the morphogenesis of multiarmed starfish and its regenerative capacity.
575 Gigascience. 2022;11.
576 23. Simakov O, Bredeson J, Berkoff K, Marletaz F, Mitros T, Schultz DT, et al. Deeply conserved
577 synteny and the evolution of metazoan chromosomes. Sci Adv. 2022;8(5):eabi5884.
578 24. Albertin CB, Medina-Ruiz S, Mitros T, Schmidbaur H, Sanchez G, Wang ZY, et al. Genome and
579 transcriptome mechanisms driving cephalopod evolution. Nat Commun. 2022;13(1):2427.
580 25. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental
581 duplications and their variation in a complete human genome. Science. 2022;376(6588):eabj6965.
582 26. Reue K, Lee JM, Vergnes L. Regulation of bile acid homeostasis by the intestinal Diet1-FGF15/19
583 axis. Curr Opin Lipidol. 2014;25(2):140-7.
584 27. Zubenko GS, Hughes HB, 3rd. Predicted gene sequence C10orf112 is transcribed, exhibits tissue-
585 specific expression, and may correspond to AD7. Genomics. 2009;93(4):376-82.
586 28. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome
587 assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.

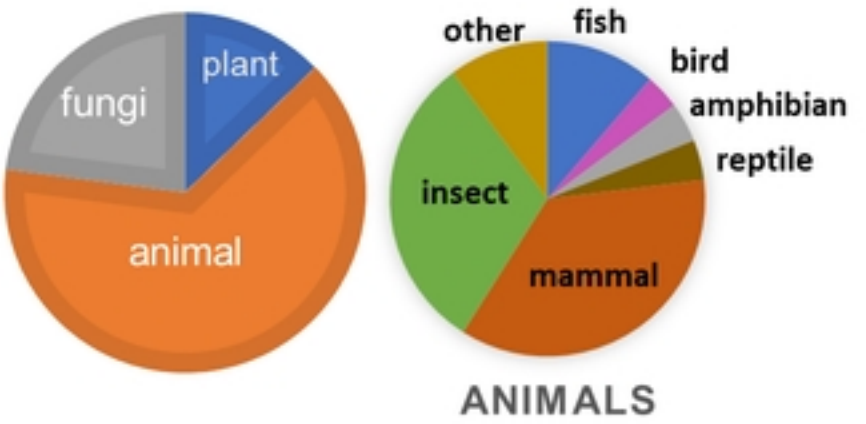
588 29. Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, et al.
589 Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res.
590 2014;24(12):2066-76.
591 30. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic
592 sequences with the threaded blockset aligner. Genome Res. 2004;14(4):708-15.
593 31. Morgulis A, Gertz EM, Schaffer AA, Agarwala R. WindowMasker: window-based masker for
594 sequenced genomes. Bioinformatics. 2006;22(2):134-41.
595 32. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast
596 genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132.
597 33. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication,
598 deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A.
599 2003;100(20):11484-9.

600

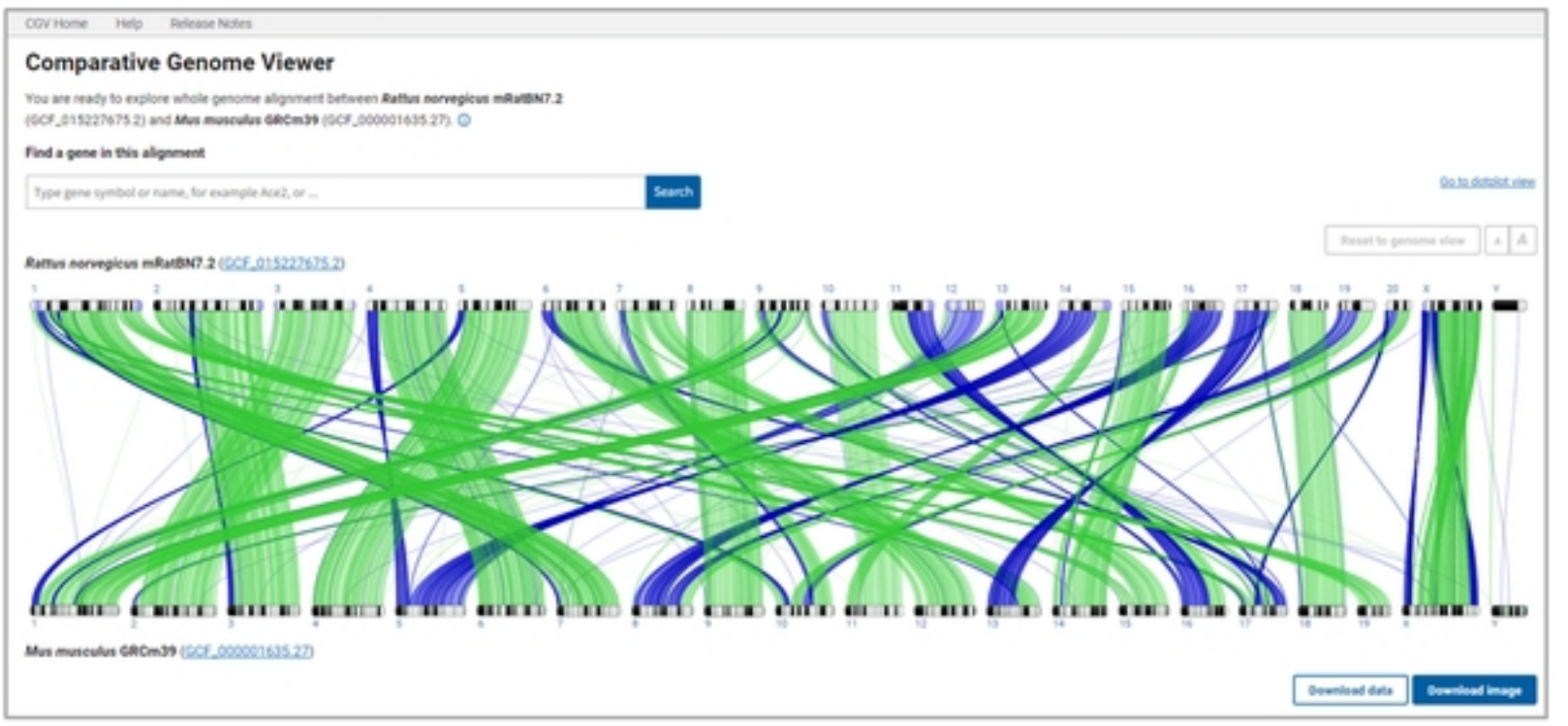
A


CGV Home
Help
Release Notes
Comparative Genome Viewer
This tool allows you to compare two genomes based on assembly-assembly alignments provided by NCBI.
Set up your view
Make a selection in each of these four steps to view assembly comparison.
1. Select a species
Mus musculus (house mouse)
2. Select a second species
Rattus norvegicus (Norway rat)
3. Select an assembly
GRCm39 (GCF_000001635.27)
4. Select a second assembly
mRatBN7.2 (GCF_015227675.2)
Clear Form
View Comparison
Not finding your alignment of interest?
Fill out the form to request more alignments.

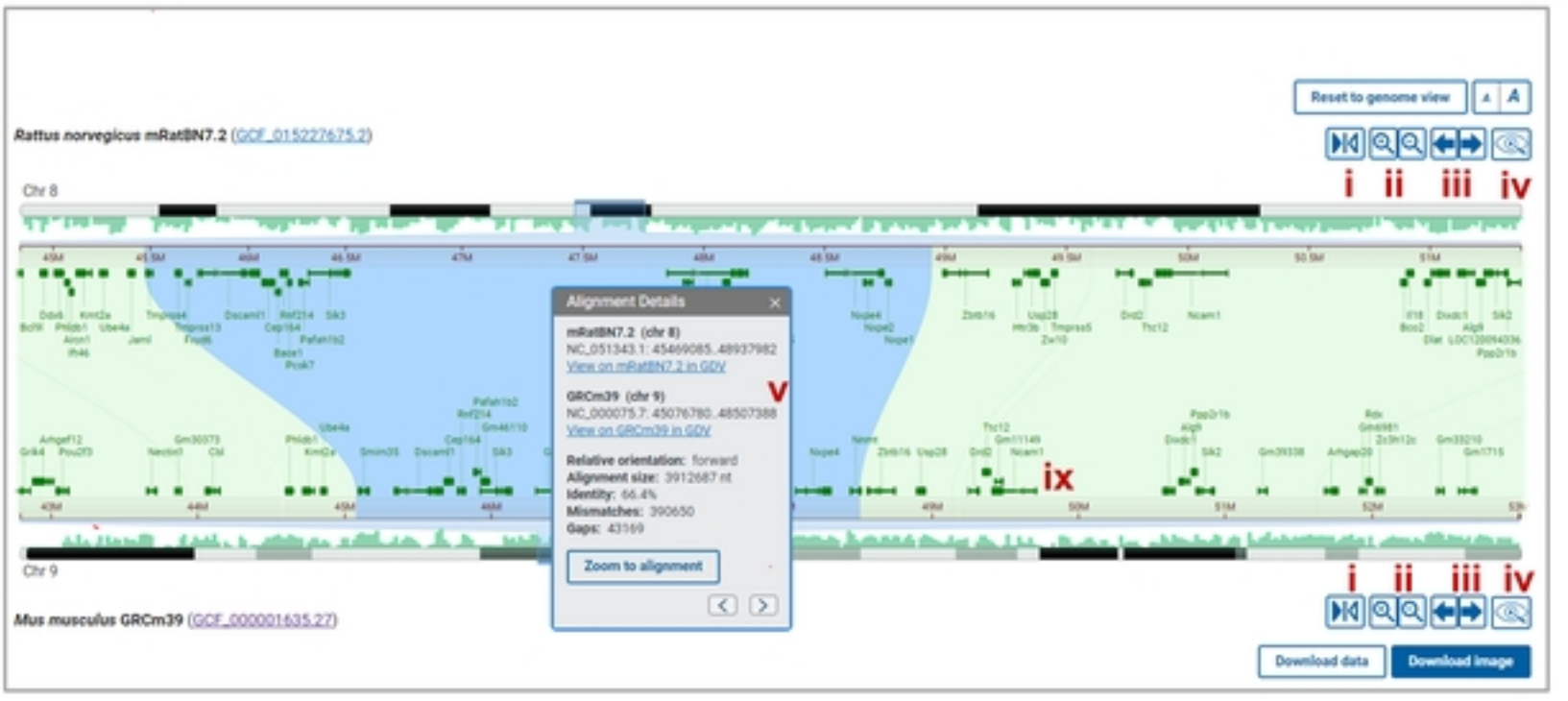
B


plant
fungi
animal
other
fish
bird
amphibian
reptile
insect
mammal
ANIMALS

C


CGV Home
Help
Release Notes
Comparative Genome Viewer
You are ready to explore whole genome alignment between Rattus norvegicus mRatBN7.2 (GCF_015227675.2) and Mus musculus GRCm39 (GCF_000001635.27).
Find a gene in this alignment
Search
Go to dotplot view
Reset to genome view
Rattus norvegicus mRatBN7.2 (GCF_015227675.2)
Mus musculus GRCm39 (GCF_000001635.27)
Download data
Download image

D


Reset to genome view
Rattus norvegicus mRatBN7.2 (GCF_015227675.2)
Chr 8
i
ii
iii
iv
Alignment Details
mRatBN7.2 (chr 8)
View on mRatBN7.2 in GDV
GRCm39 (chr 9)
View on GRCm39 in GDV
Relative orientation: forward
Alignment size: 3912687 nt
Identity: 66.4%
Mismatches: 390650
Gaps: 43169
Zoom to alignment
v
ix
Chr 9
Mus musculus GRCm39 (GCF_000001635.27)
Download data
Download image

E

Find a gene in this alignment
wnt
Search
Search results
Assembly mRatBN7.2 - 38 genes shown
Gene
Description
Location
Apc
APC regulator of WNT signaling pathway
Chr18: 25828558..25925511
Wnt5a
Wnt family member 5A
Chr16: 3697032..3718230
Wnt3a
Wnt family member 3A
Chr10: 44034174..44078366
Assembly GRCm39 - 35 genes shown
Gene
Description
Location
Apc
APC, WNT signaling pathway regulator
Chr18: 34353350..34455243
H2az2
H2A.Z histone variant 2
Chr11: 6377226..6394511
Wnt1
wingless-type MMTV integration site family, member 1
Chr15: 98687738..98691711

F

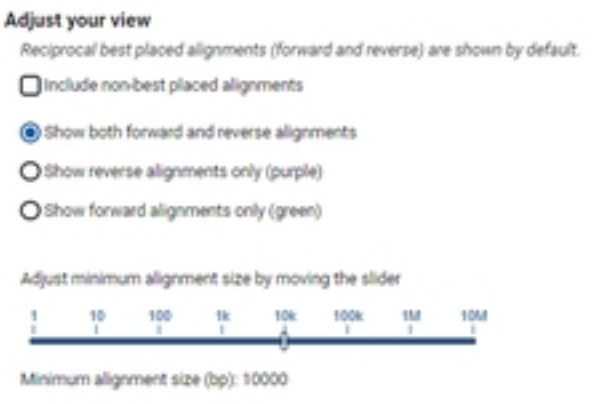
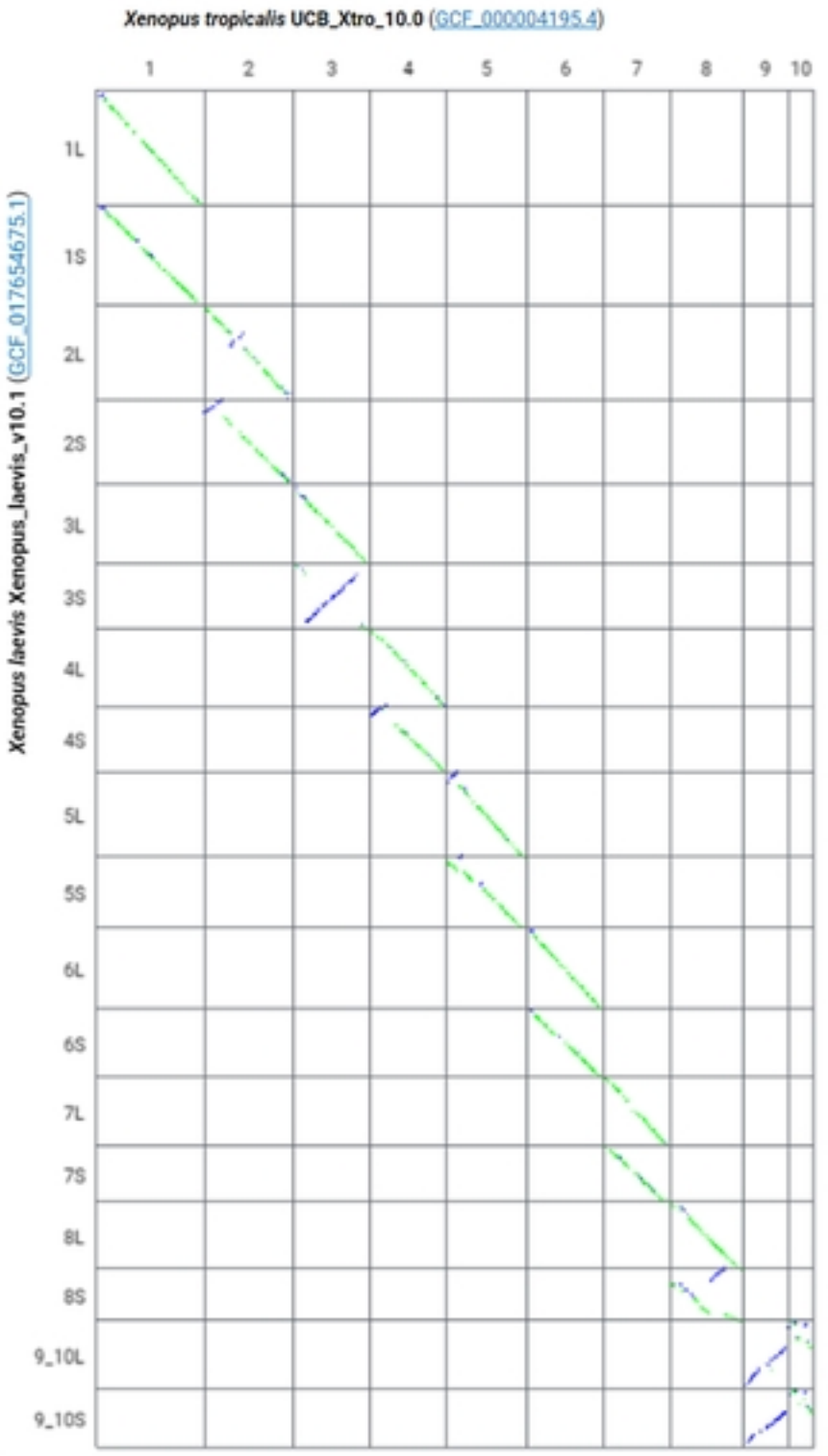

Adjust your view
Reciprocal best placed alignments (forward and reverse) are shown by default.
Include non-best placed alignments
Show both forward and reverse alignments
Show reverse alignments only (purple)
Show forward alignments only (green)
Adjust minimum alignment size by moving the slider
1
10
100
1k
10k
100k
1M
10M
Minimum alignment size (bp): 10000

Fig1

A


Xenopus tropicalis UCB_Xtro_10.0 (GCF_000004195.4)
1
2
3
4
5
6
7
8
9
10
Xenopus laevis Xenopus_laevis_v10.1 (GCF_017654675.1)
1L
1S
2L
2S
3L
3S
4L
4S
5L
5S
6L
6S
7L
7S
8L
8S
9_10L
9_10S

B

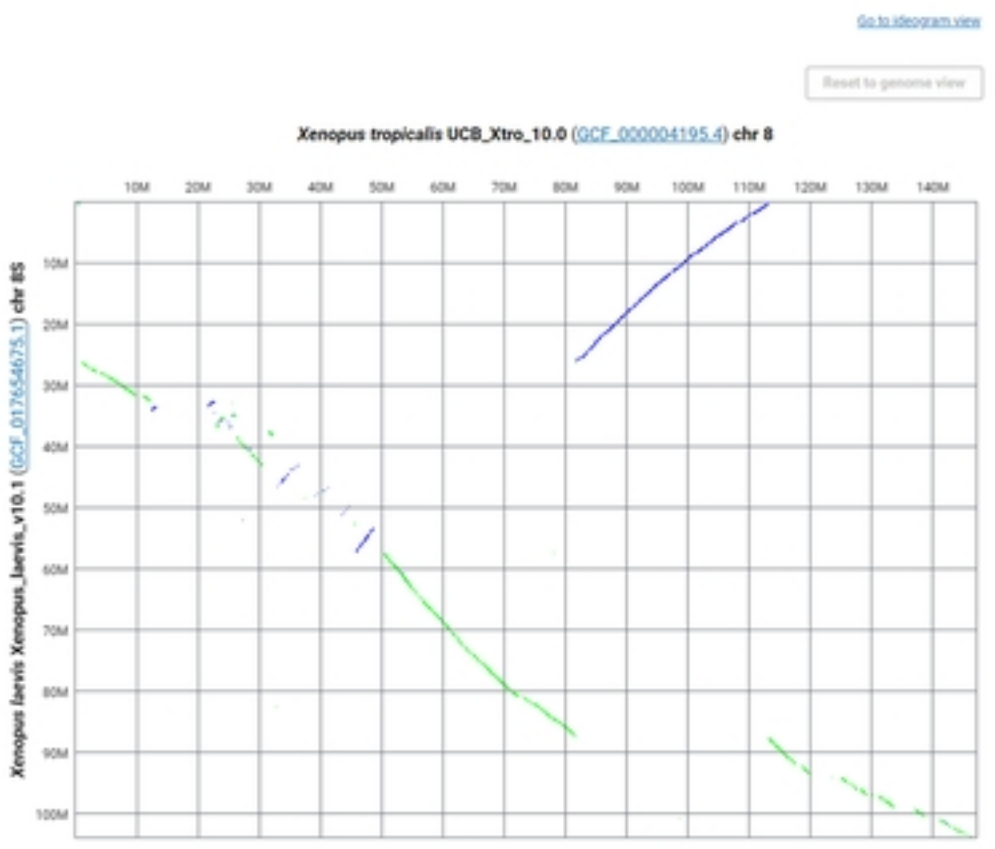

Go to ideogram view
Reset to genome view
Xenopus tropicalis UCB_Xtro_10.0 (GCF_000004195.4) chr 8
10M
20M
30M
40M
50M
60M
70M
80M
90M
100M
110M
120M
130M
140M
Xenopus laevis Xenopus_laevis_v10.1 (GCF_017654675.1) chr 8S

Fig2

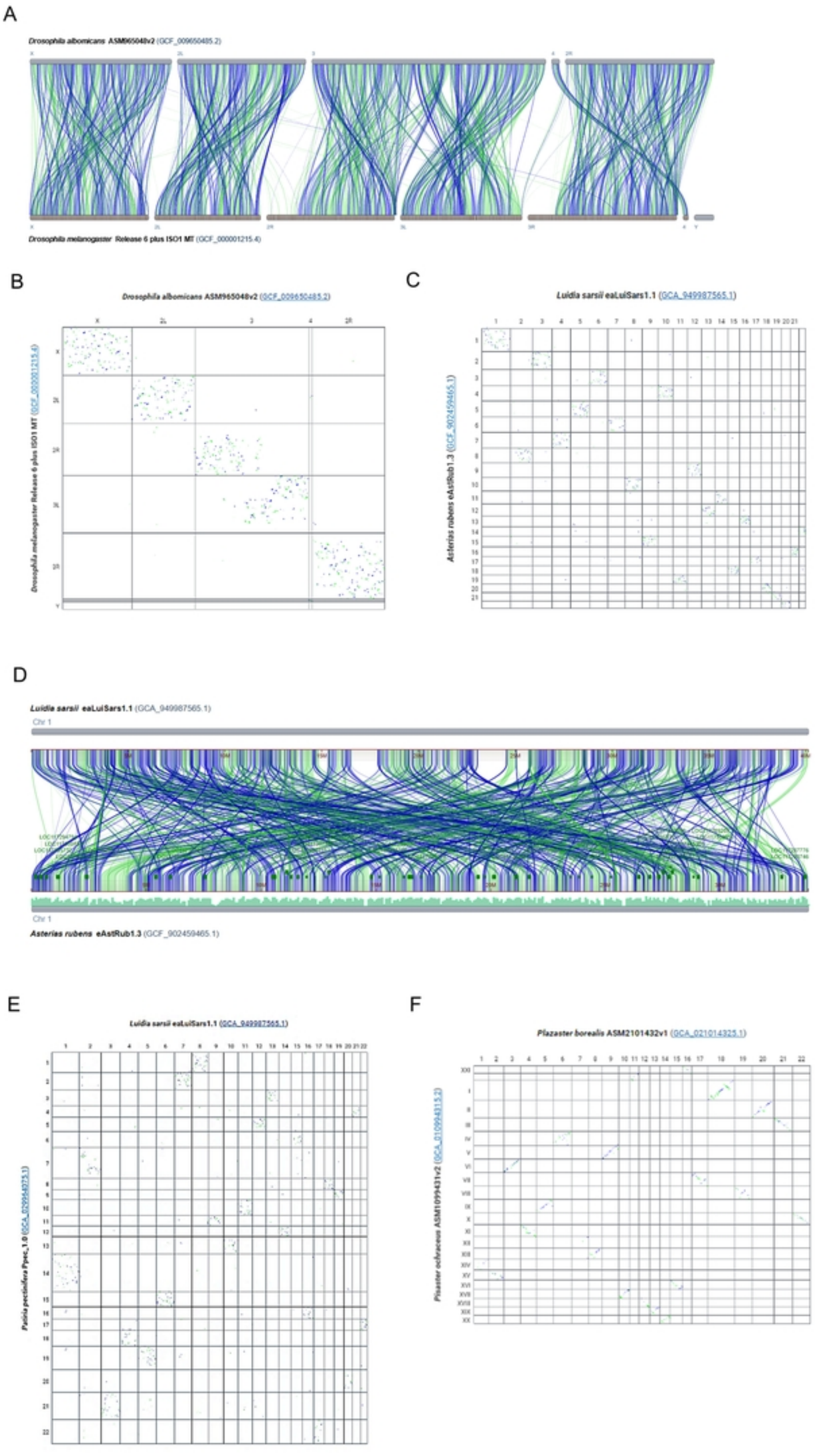
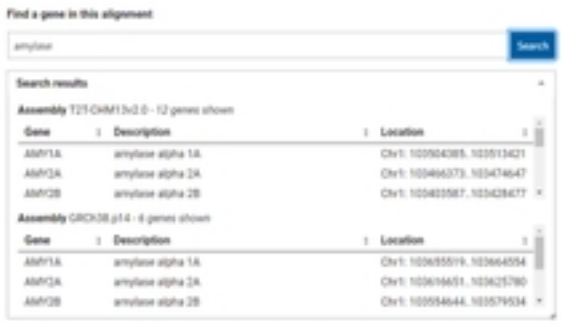
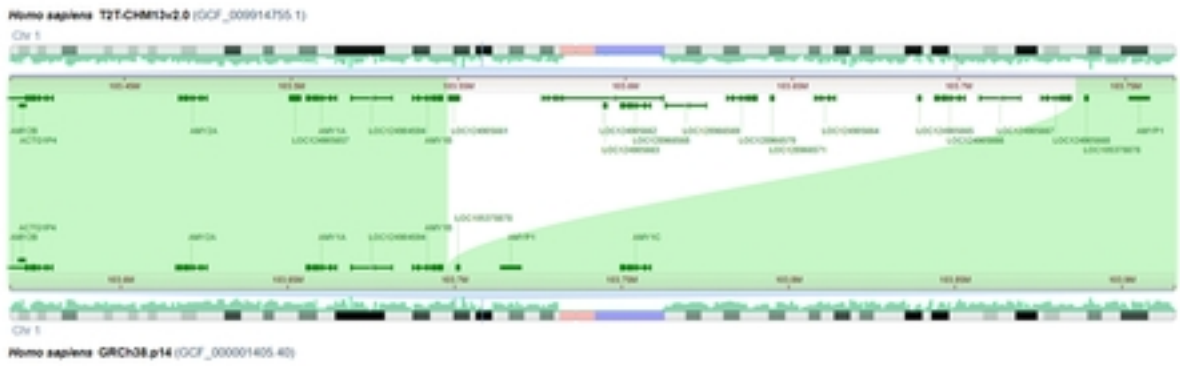

A
Drosophila albomicans ASM965048v2 (GCF_009650485.2)
Drosophila melanogaster Release 6 plus ISO1 MT (GCF_000001215.4)
B
Drosophila albomicans ASM965048v2 (GCF_009650485.2)
Drosophila melanogaster Release 6 plus ISO1 MT (GCF_000001215.4)
C
Luidia sarsii eaLuiSars1.1 (GCA_949987565.1)
Asterias rubens eAstRub1.3 (GCF_902459465.1)
D
Luidia sarsii eaLuiSars1.1 (GCA_949987565.1)
Chr 1
Asterias rubens eAstRub1.3 (GCF_902459465.1)
E
Luidia sarsii eaLuiSars1.1 (GCA_949987565.1)
Patiria pectinifera Ppec_1.0 (GCA_029994973.1)
F
Plazaster borealis ASM2101432v1 (GCA_021014325.1)
Pisaster ochraceus ASM1099431v2 (GCA_010994315.2)

Fig3

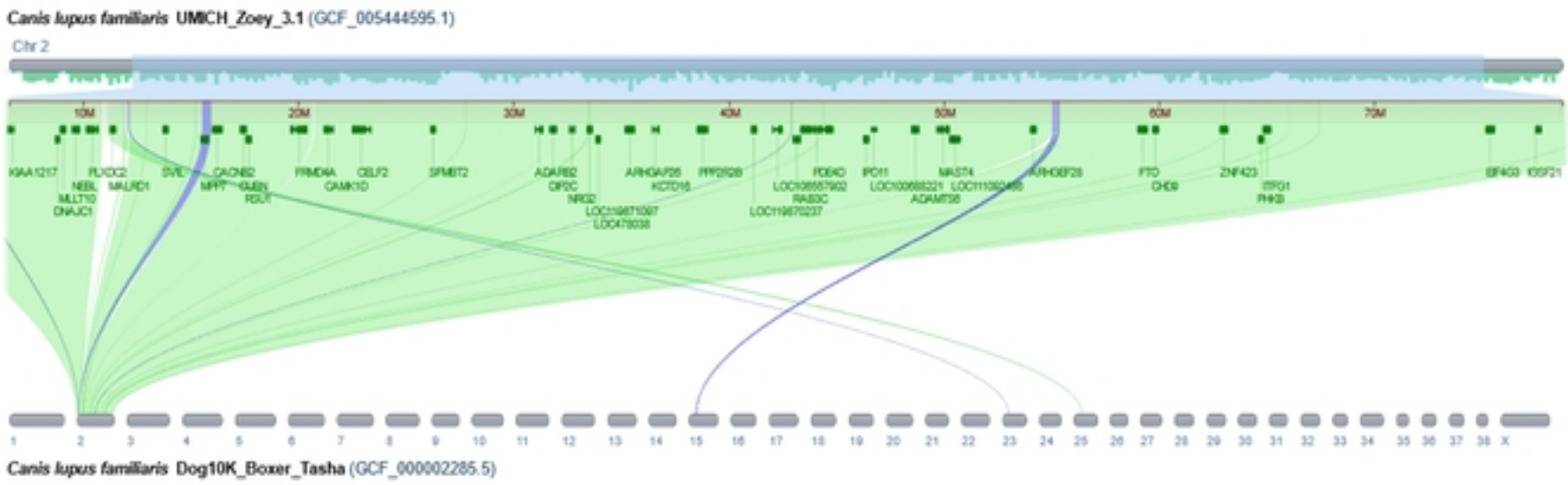
A


Find a gene in this alignment
Search
Search results

B


Homo sapiens T2T-CHM13v2.0 (GCF_009914755.1)
Homo sapiens GRCh38.p14 (GCF_000001405.40)

C


Canis lupus familiaris UMICH_Zoey_3.1 (GCF_005444595.1)
Chr 2
Canis lupus familiaris Dog10K_Boxer_Tasha (GCF_000002285.5)

D

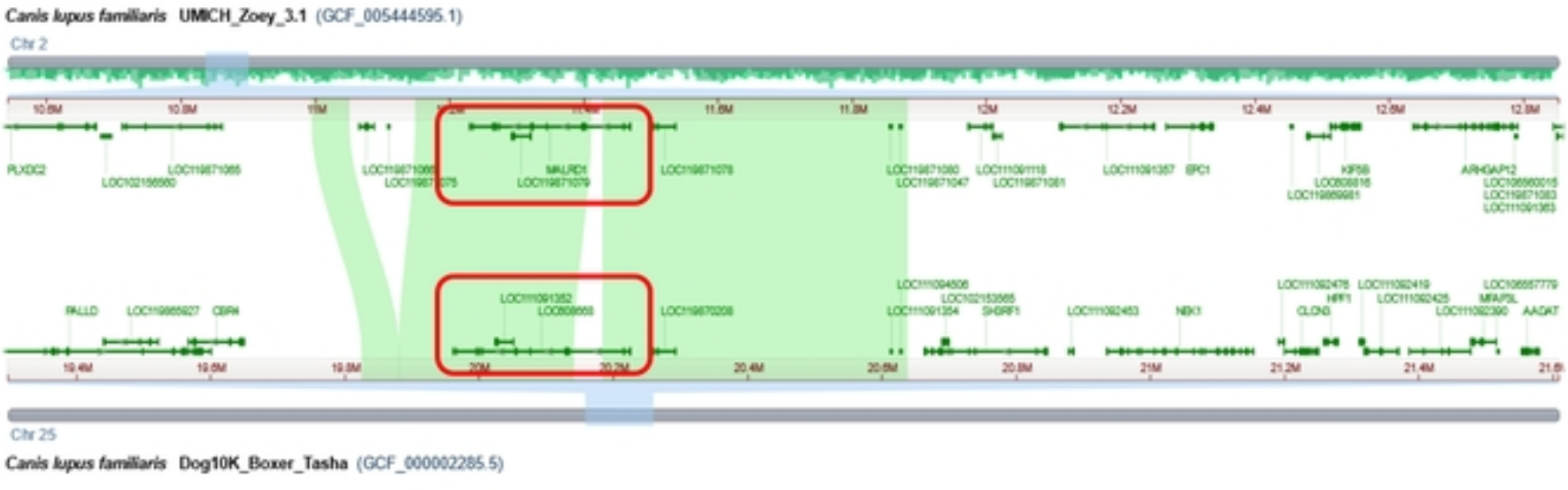
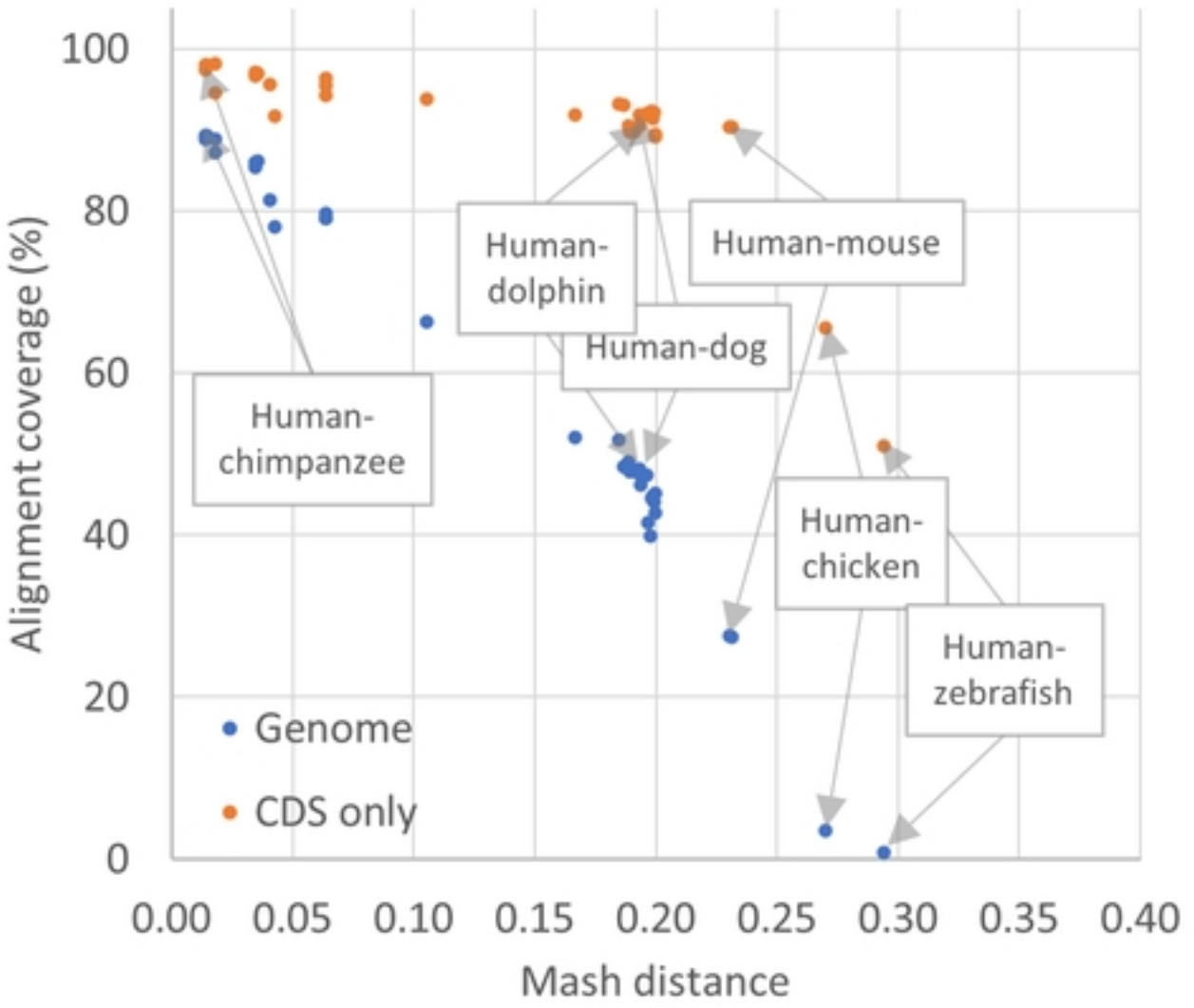

Canis lupus familiaris UMICH_Zoey_3.1 (GCF_005444595.1)
Chr 2
Chr 25
Canis lupus familiaris Dog10K_Boxer_Tasha (GCF_000002285.5)

Fig4

A


100
80
60
40
20
0
Alignment coverage (%)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Mash distance
Human-chimpanzee
Human-dolphin
Human-dog
Human-mouse
Human-chicken
Human-zebrafish
Genome
CDS only

B

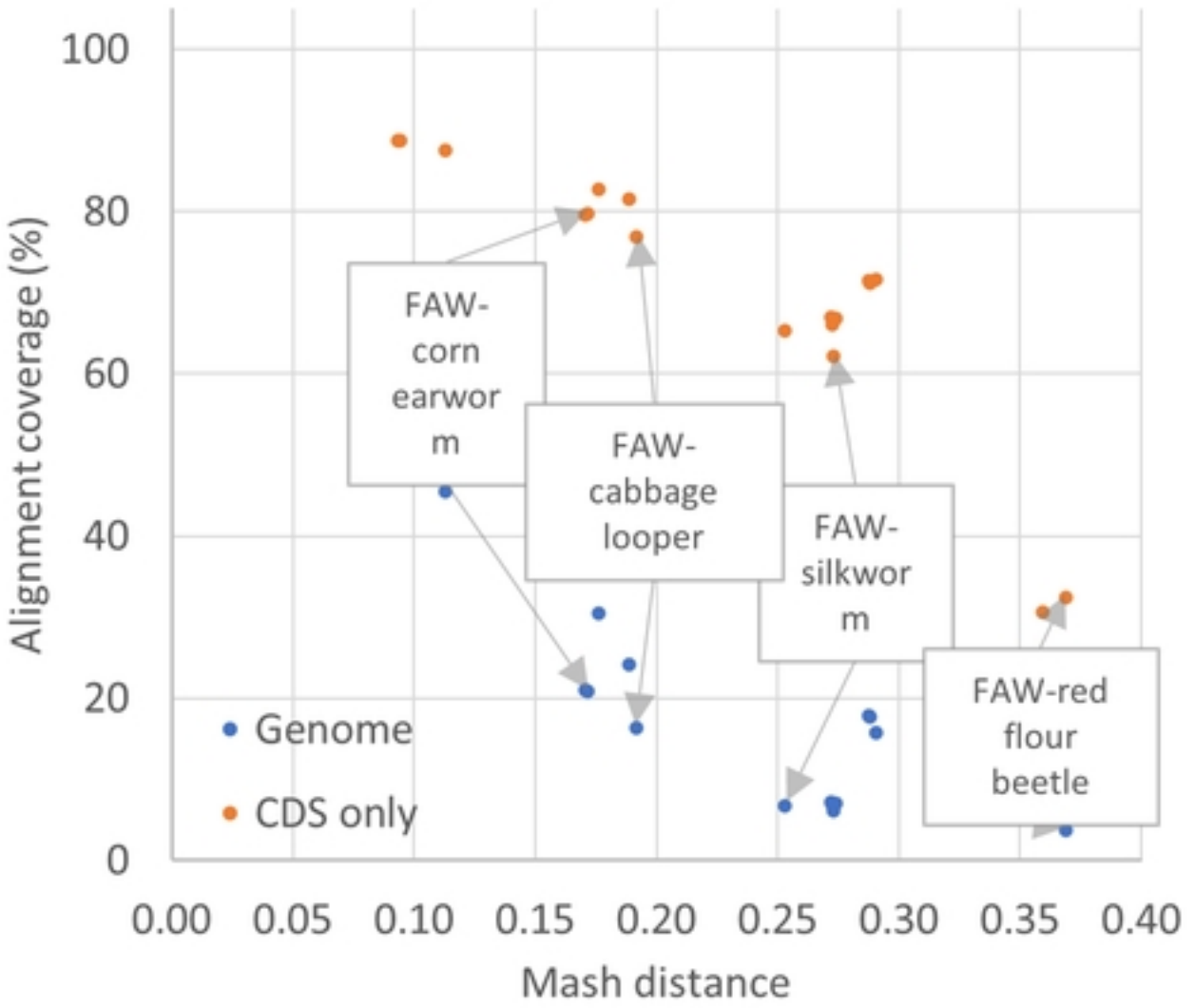

100
80
60
40
20
0
Alignment coverage (%)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Mash distance
FAW-corn earworm
FAW-cabbage looper
FAW-silkworm
FAW-red flour beetle
Genome
CDS only

Fig5