CView: A network based tool for enhanced alignment visualization

Raquel Linheiro^{§,1}, Diana Lobo^{§,1,2,3}, Stephen Sabatino^{§,1,3} and John Archer^{*,1,3}

[§]Contributed equally

^{*}Corresponding

Email: john.archer@cibio.up.pt

¹ CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, *InBIO* Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal.

² Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal

³ BIOPOLIS, Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, Portugal.

49 **Abstract**

50 To date basic visualization of sequence alignments have largely focused on
51 displaying per-site columns of nucleotide, or amino acid, residues along with
52 associated frequency summarizations. The persistence of this tendency to the
53 more recent tools designed for the viewing of mapped read data indicates that
54 such a perspective not only provides a reliable visualization of per-site
55 alterations, but also offers implicit reassurance to the end user in relation to
56 data accessibility. However, the initial insight gained is limited, something that
57 is especially true when viewing alignments consisting of many sequences
58 representing differing factors, such as geographical location, date and
59 subtype. A basic alignment viewer can have potential to increase initial insight
60 through visual enhancement, whilst not delving into the realms of complex
61 sequence analysis. Here we present CView, a visualizer that expands on the
62 per-site representation of residues through the incorporation of a dynamic
63 network that is based on the summarization of diversity present across
64 different regions of the alignment. Within the network nodes are based on the
65 clustering of sequence fragments spanning windows that are placed
66 consecutively along the alignment. Edges are placed between nodes of
67 neighbouring windows where they share sequence id's. Thus, if a single node
68 is selected on the network, then the relationship that all sequences passing
69 through that node have to other regions of diversity within the alignment can
70 be instantly observed through the tracing of paths. In addition to augmenting
71 visual insight, CView provides many export features including variant
72 summarization, per-site residue and kmer frequency matrixes, consensus
73 sequence generation, alignment dissection as well as general sequence

74 clustering, each of which are useful across a range of research areas. The
75 software has been designed to be user friendly, intuitive and interactive. It,
76 along with source code, a quick start guide and test data, are available
77 through the SourceForge project page: <https://sourceforge.net/projects/cview/>.

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99 **Introduction**

100 Tools developed to visualize local sets of aligned sequences, such as those
 101 produced by multiple sequence aligners including MUSCLE [1] and Clustal W
 102 [2], have focused largely on displaying columns of nucleotide or amino acid
 103 characters, and highlighting the differences between such characters primarily
 104 through the use of colour [3–7]. More advanced sequence management and
 105 analysis packages, such as Geneious [8] and Mega [9], as well as the more
 106 recent tools designed for basic visualization of mapped read data, including
 107 IGV [10], GenomeView [11] and Tablet [12], incorporate a wide array of
 108 analysis, summarization and annotation options, but in terms of basic
 109 visualization they follow a similar approach. It is evident that direct
 110 observation of residues, as well as the general per-site based summarization,
 111 not only provides an accessible view of per-site alterations between
 112 sequences within the alignment but also at times gives the end user a level of
 113 reassurance in relation to their data. However, initial insight gained in relation
 114 to the overall alignment is limited, especially when viewing alignments
 115 consisting of many sequences representing varying factors of interest such as
 116 geographical location, subtype, treatment strategy, compartment and date /
 117 time-point. Basic alignment visualization should have the potential to increase
 118 the level of initial insight within sequence datasets whilst not delving into the
 119 realms of more complex sequence analysis. Here we present CView, a simple
 120 multiple sequence alignment visualizer that incorporates a dynamic network
 121 that is based on a summarization of the diversity across different regions of
 122 the alignment. The immediate coupling of aligned sequences to such a
 123 network provides a way of visually tracking the context of observed diversity

124 within characters that are currently onscreen to that of the surrounding
125 regions of the alignment not currently in view. This provides the user with an
126 increased intuitive and visual summarization of the context of this diversity.

127

128 CView provides a range of export features that can be applied to the entire
129 alignment, to a specified region of the alignment, or to a specified region in
130 conjunction with a specified subset of sequences. Such export features
131 include: variant summarization, per-site character and kmer frequency
132 matrixes, clustered sequences, pairwise-distance matrixes as well as
133 consensus sequence generation. For example, when the variant
134 summarization option is selected a list of variants spanning the user-specified
135 region, within the user-specified group of sequences, is created by identifying
136 all unique forms and associating each with their frequency of occurrence. A
137 list of the original sequence titles represented by each variant is also
138 maintained. Such a feature has use in the tracking of viral populations, for
139 example in searching for the presence of genotypic alterations such as those
140 associated with immune escape [13], drug resistance [14], or co-receptor
141 usage [15,16]. Additionally, this feature has use in both clinical [17], and
142 environmental metagenomics [18–20], where the summarization of
143 populations of microbes is of interest. Each such export option is described
144 detail within the user manual that is available on the SourceForge wiki located
145 at <https://sourceforge.net/p/cview/wiki/Help/>. Aside from features related to the
146 extraction of secondary information from the alignment, CView provides the
147 ability to dissection the alignment into subsets of sequences and regions; a
148 task that is often laborious in the absence of a background in script

development. For example, a user can export a specified region of sequences associated with a specific time-point, geographical location or body compartment, as long as the sequence titles have been labelled with such information. Such labelling is often as standard output feature of many sequence repositories, for example from the Los Alamos HIV sequence database the user can select options such as subtype, patient code, country and year to be included within the title of each sequence [21], but such information may also be part of experimental design such as compartment [22] or time-point [23].

Within CView the associated network is displayed directly below the alignment. This network is based on the clustering of sequences within windows placed consecutively across the alignment, where each cluster becomes a node. Edges are placed between nodes of neighbouring windows where they share differing regions of the same sequence(s). Thus, if a single node within a diverse region of the alignment is selected, then the relationship that all sequences passing through that node have to other regions of diversity within the alignment can be instantly observed. Here we describe how these networks are constructed and graphically displayed. The clustering threshold used during network construction, as well as the number and width of windows, are specified on the user-interface through a series of user-friendly slider bars. Alterations are updated in real time, which allows the user to rapidly explore the visualization of variable regions across the alignment. The software has been designed to be user friendly, intuitive and interactive

173 and it, along with source code, a quick start guide and test data, is available
174 through the SourceForge project page: <https://sourceforge.net/projects/cview/>.

175

176 **Methods**

177 **Implementation**

178 The interface has been designed for simplicity and clarity. It consists of four
179 basic areas of user-interaction (figure 1) which are: (1) sequence view, (2)
180 network view, (3) navigation and control and, (4) menu driven outputs.

181

182 **Figure 1: CView interface.** The four main areas of the CView interface are
183 depicted. These are sequence view, network view, control panel and the top
184 menu. The yellow numbers on the top indicate the sites of the alignment that
185 are currently in view. These correspond to the yellow bar on the top of the
186 location indicator. The orange numbers along the bottom indicate the
187 locations of the windows that nodes within the network are dependent on.
188 These window locations correspond to the area that the orange bar located
189 the under the location indicator covers. Grey dots indicate (selectable) nodes
190 within windows. The squares along the location indicator can also be selected
191 in order to jump directly to the indicated co-ordinates. The red text around the
192 outside of the interface describes the main features.

193

194 **(1) Sequence View**

195 Sequences are displayed above the alignment location indicator. The dynamic
196 yellow bar associated with the latter represents the region of the alignment
197 that is currently visible. The green dot on the right hand side indicates what

198 proportion of the alignment is visible. The consensus sequence of the
199 alignment is displayed along the top of the sequence area, and directly under
200 this the “+” indicates columns where all characters agree with the consensus
201 character. Sequences and their titles are selectable and when a sequence is
202 chosen it will be traced through the corresponding network as a yellow line.
203 Sequences that pass through a user-selected node on the network are
204 displayed with a red dot next to the title. Basic interactive features associated
205 with the sequence display include the masking of characters that are the
206 same as consensus, altering font size and altering space allocated to
207 displaying titles; these are achieved through the “Navigation and Control”
208 panel. Site locations are highlighted in yellow along the top of the interface.

209

210 (2) Network View

211 The network depicting sequence diversity within the alignment is displayed
212 directly below the alignment location indicator. The associated orange bar of
213 the latter represents the region of the alignment that is currently represented
214 by the network. The region begins from the current sequence view and
215 extends to the right-hand side in a manner that is dependent on the number of
216 consecutive windows, as well as their width (figure 2, step i); windows being
217 regions from which nodes reflecting diversity are created. Both these
218 parameters can be interactively altered by the user. Window locations are
219 highlighted in orange along the bottom of the interface.

220

221 **Figure 2: Network construction.** Coloured bars indicate unique sequence
222 id's relative to the corresponding sequences (dotted lines). Within each

223 window sequence id's are associated with individual sequence fragments
 224 spanning that window (i) and fragments within windows are clustered (ii).
 225 Edges are placed between neighbouring clusters where they share one or
 226 more sequence id (iii). Clusters are represented visually on the network by
 227 grey dots. If a single cluster is selected the paths of all sequences passing
 228 through in relation to all other clusters (red lines) can be traced (vi).

229

230 (2.1) Nodes

231 For a given window clustering fragments of sequences that span it creates
 232 nodes (figure 2, step ii). Each cluster is created using an iterative approach.
 233 Initially a fragment is randomly selected to be a seed for a newly created
 234 empty cluster. All related fragments to that seed are then added to the cluster
 235 and become seeds for the next iteration. The metric used to define
 236 relatedness is hamming distance, in which the number of different characters
 237 between two aligned sequences are counted. The default threshold value is
 238 0.3, indicating that fragments that have less than 30% divergence from a seed
 239 are included within the cluster. More advanced measures of genetic distance
 240 exist that account for proposed models of sequence evolution at both
 241 nucleotide and amino acid levels [24,25], but for the rapid clustering across
 242 windows placed along an alignment for the purpose of visualization hamming
 243 distance works well [26]. Iterations continue until no more next-round seeds
 244 can be identified. If unclustered fragments within the window still exist, a new
 245 cluster is initiated by selecting another random seed from the remaining
 246 fragments and the process is repeated.

247

248 For windows where six or less clusters are created, all clusters are displayed
 249 as grey coloured circles, or nodes, on the network. For windows with more
 250 than six clusters, the largest six are displayed as nodes whilst the remaining
 251 are placed into a holding structure that is used for visualization purpose only
 252 and that is displayed as a black circle. Internally all clusters contained within
 253 the latter are treated separately, for example in relation to tracking and
 254 highlighting paths. Here six was chosen to be the upper limit so that following
 255 edge placement (next section), and during edge crossover minimization, the
 256 maximum number of nodes that need order re-arrangement within any one
 257 window is seven, including the holding node. This is because for a set
 258 containing n items, there are n factorial different order permutations [27], and
 259 during edge crossover minimization the number of edge crossovers produced
 260 by each permutation, in relation to nodes within a neighbouring window, must
 261 be counted. For a given window if there are the maximum of seven nodes
 262 present, $7!$ permutations (5040) must be identified during crossover
 263 minimization and this can be done in a reasonable time (< 1 second on an
 264 average laptop). If on the other hand there are fifteen nodes allowed within a
 265 window, then there are $15!$ permutations (1307674368000) requiring a time of
 266 many days.

267

268 The default number (10) and width (50 bp) of windows, as well as the pairwise
 269 distance threshold, can be altered using the slider bars within the “Navigation
 270 and Control” area. The CViews graphical display is for rapid intuitive
 271 visualization, and if a higher clustering resolution is required, i.e. less than the
 272 0.2 lower bound allowed for the display network, the user can perform this

273 using the “Cluster Sequences” option of the “Alignment” menu. The number of
274 sequence passing through each node is indicated in green on the left and
275 right hand sides of the network.

276

277 (2.2) Edges

278 Edges are placed between nodes of neighbouring windows where they
279 possess fragments that are derived from the same underlying sequence(s)
280 (figure 2, iii). Consequently, individual sequences can be traced through
281 nodes across different windows. As previously mentioned edge crossover
282 between nodes within adjacent windows is minimized. Starting at the second
283 most right-hand-side window, this is done by calculating all possible node
284 order permutations, following which for each permutation, the number of edge
285 crossovers to nodes within the adjacent right-hand window is counted, node
286 layout order in the latter being kept constant (figure 3). The permutations that
287 produce the minimum number of crossovers are selected (figure 3, red
288 numbers), and from these a random one is used. The process is then
289 repeated one window to the left, until the first window of the alignment is
290 reached. Crossover minimization, while visually more pleasing, has no effect
291 on the sequence information or underlying node connections. Following the
292 connection of edges it is possible to click on nodes within the graph and track
293 sequences that pass through them (figure 2, iv). On the interface, such
294 sequence paths are displayed in red, and within the sequence display area a
295 red dot are placed next to the titles of included sequences.

296

Figure 3: Minimization of edge crossovers between nodes of the two right most windows of the alignment. This process is repeated until the left most window (anchored on site 1) is reached. Clusters within the two windows are labelled with integers and required edges, based on the sequence ids (coloured bars), are listed (i). All order permutations of the current left window are identified and for each permutation the required edges are placed relative to the constant cluster order of the right window (ii). Crossovers are then counted (red numbers). Of the permutations that produce the minimum number of crossovers a random one is selected for graphical node layout order.

307

(3) Navigation and Control

Access to all the previous described options is provided through the slider bars associated with the navigation and control panel (bottom right of figure 1). The buttons labelled with then red directional arrows are used to scroll through the alignment. These were implemented to remove the need for flat scroll bars as future developments will be aimed at tablets and mobile devices. The red dot, at the centre of the four scroll arrows, immediately jumps a viewpoint at the centre of the alignment. In addition to the directional arrows the user can click directly on the grey squares along the alignment location indicator bar to immediately move to a particular location. Within this control area there are also three buttons used for printing the network to a .png formatted image file.

320

(4) Menu Driven Output

322 Output options are accessed through the top menu bar and can be applied to
 323 (i) the alignment in general, (ii) subsets of sequences whose titles match a
 324 user search tag, (iii) subsets of sequences that pass through a selected node
 325 and (iv) subsets of sequences defined by the user based on a supplied file of
 326 titles. In addition to exporting subsets of sequences and/or specified regions
 327 of the alignment Cview can generate summary statistics such as frequencies
 328 of residues and kmers and tertiary, pairwise distance matrix's, variant count
 329 information and clustered sets of sequences. A detailed description of each
 330 output option is presented on the wiki associated with the SourceForge
 331 project page (<https://sourceforge.net/p/cview/wiki/Help/>).

332

333 **Results**

334 (1) The software

335 CView has been implemented in Java and runs on operating systems with
 336 installed Java Runtime Environment 8.0 or higher. It has been developed
 337 using an object-orientated approach for ease of plug-in development; where
 338 plug-ins related to alignment visualization will be based on user feedback. To
 339 obtain an executable jar file, download the cview.zip file from the SourceForge
 340 project page. Following the extraction the CView.jar file from the zip file,
 341 CView is executed by double clicking on the jar file. This will launch the
 342 interface through which alignments can be loaded. Alignments must be in
 343 fasta format and are be loaded using the "Load (fasta)" option of the "All
 344 Sequences" menu. Once a fasta-formatted alignment is loaded the workflow
 345 is driven by how the user interacts with the interface and the various output
 346 options. Test data, in the form of an alignment consisting of 636 sequences

representing the gp120 region of the HIV-1 genome is included with the cview.zip file that contains the software. This data was obtained from the Los Alamos HIV sequence database [21].

(2) Test Case Example: Exploring variation associated with co-receptor usage

(2.1) Background

HIV-1 viruses can be characterized into two phenotypes that are dependent on cellular tropism and that are as a result of differences in co-receptor usage [28]. The macrophage tropic phenotype, often referred to as R5, requires the CCR5 co-receptor, whilst the T-cell tropic phenotype (X4) uses the CXCR4 co-receptor, the latter often emerging later on during infection [29]. Co-receptor usage can be detected by computational analysis based on specific genetic alterations within the V3 loop of the gp120 gene [15,16]. Genetic variation within this region, of approximately 105 nt in length, lead to structural shifts that result in optimized binding to one co-receptor or the other [30]. For demonstrating the applicability of CView we have used it to explore and summarize this known variation relating to co-receptor usage from an alignment of HIV-1 subtype B gp120 sequences.

(2.2) Method

1. All North American subtype B gp120 sequences, verified to be CCR5-using sequences (n = 636), were downloaded from the Los Alamos HIV sequence database in aligned fasta format [21]. These were loaded into CView, using the “Load (fasta)” option of the “All Sequences” menu, following which they were saved in unaligned format using the “Save (unaligned)” option of the “All

372 Sequences” menu. Additionally, the titles of these sequences were saved to a
373 separate file using the “Save (titles)” option.

374

375 2. Step 1 was repeated for CXCR4-using sequences (n = 76).

376

377 3. In order to make sites directly comparable between the two sets of
378 unaligned sequences, they were combined into a single file and aligned using
379 MUSCLE [1].

380

381 4. The resulting alignment was loaded into CView and the consensus
382 sequence of the region spanning the V3 loop was saved using the “Save
383 (consensus)” option of the “All Sequences” menu. Within this alignment the
384 co-ordinates of the region spanning the V3 loop were from 1436 to 1568.
385 Although the exact location of the V3 loop within the gp120 region is known,
386 the coordinates will vary depending on the alignment due to the placement of
387 gaps during the alignment process. The exact co-ordinates for our alignment
388 were identified by eye using the V3 sequence of the HIV-1 reference strain
389 (Name: HXB2-LAI-IIIB-BRU, Accession: K03455), where the start residues of
390 the loop are TGTACAAGACCC and the end residues are CAAGCACATTGT
391 [21].

392

393 5. Using the original sequence titles saved in step 1, the proportion of the
394 alignment corresponding to R5 sequences was saved in aligned format. This
395 was done using the “Save (from - to)” option under the “Groups” menu item,
396 where the titles were supplied as a list to define the group.

397

398 6. Step 4 was repeated for the titles corresponding to the X4 sequences.

399

400 7. Steps 5 and 6 resulted in two sub-alignments whose site co-ordinates are
 401 compatible as they were both extracted from the same underlying source. For
 402 each of the extracted R5 and X4 alignments, CView was used to obtain a list
 403 of all variants spanning the V3 loop along with their frequencies. This was
 404 done using the “Frequencies (variants)” option of the “All Sequences” menu,
 405 where the co-ordinates used were those described in step 3.

406

407 8. Variants were translated using the EMBOSS Transeq tool [31].

408

409 9. For each of the R5 and X4 alignments, nucleotide frequencies were
 410 obtained using the “Frequencies (nuc/aa)” option of the “All Sequence” menu
 411 (co-ordinates: 1436 to 1568).

412

413 The underlying alignment described for this use-case scenario, consisting of
 414 all the HIV-1 SUBTYPE B sequences spanning the GP120 region of the
 415 genome that have been verified as either being a CCR5-using (n = 636) or
 416 CXCR-using (n = 76), is available from the CView project page, within the
 417 compressed folder USE_CASE_DATA.zip. A further test dataset consisting of
 418 just the CCR5-using sequences from above is contained within the zip folder
 419 where the software itself is located.

420

421 (2.3) Result and Discussion

Figure 4A displays the consensus sequence of the V3 region from the MUSCLE generated alignment prior to being divided by genotype. The top ten most frequent variants from each of the two genotypes are also displayed. The seqPublish tool, located at <https://www.hiv.lanl.gov/content/sequence/SeqPublish/seqpublish.html> [21], was used to format these alignments from the CView output such that characters identical to those of the consensus sequence were hidden. A similar feature is available at the bottom of the output file that is generated by the “Variant Frequency” option of CView, where residues that are identical to those present on the most frequent variant are represented by a “|” character. The translation of each of these variants is presented within figure 4B. A summary, using sequence logos [32], is available within figure 4C where it can be observed that at translated site 11 the positively charged amino acid residues R (arginine) and H (histidine) are present within the sequences that were known to be CXCR4-using, while they are absent within the sequences obtained from the CCR5-using strains. At site 26 a similar observation is made in relation to positively charged residues, this time including a K (lysine) residue; although there is a minority K also present at 26 within the CCR5-using variants. This is a known observation where the presence of positively charged amino acids at sites 11 and 26 result in a structural alteration that optimizes CXCR4 co-receptor binding [15,16]. The steps leading to this observation, within this use-case scenario, demonstrate the utility of CView when exploring such alignment data. Complete per-site nucleotide frequencies for both R5 and X4 sequences spanning the V3 region are presented in supplementary table S1.

447

448 **Figure 4: Summarization of variation present within the V3 loop. (A)**

449 Green residues represent non-consensus residues from the ten most frequent
450 variants associate with the CCR5-using phenotype. Brown represents those
451 of the CXCR4-using phenotype. The consensus sequence (black) is shown.
452 (B) Translations of the most frequent ten variants from each phenotype. (C)
453 Sequence logos summarizing these translations. The top logo is from
454 represents the CCR5-using sequences whilst the bottom represents the
455 CXCR4-using ones.

456

457 **Conclusion**

458 CView is a tool that allows the user to interactively explore sequence
459 alignments with the aid of a dynamic network that summarizes the diversity
460 present. Here we have described how CView was designed and, as an
461 example, we have used it to aid in the characterization of known variation
462 between sequences involved in HIV-1 co-receptor usage. The exact usage
463 scenario in which CView can be applied is dependent on the requirements of
464 the individual user. CView is available from <https://sourceforge.net/p/cview>.

465

466 **Funding**

467

468 This work was funded by National Funds through FCT (Fundação para a
469 Ciência e a Tecnologia) and FEDER through the Operational Programme for
470 Competitiveness Factors (COMPETE), via a project awarded to JA, under the
471 references POCI-01-0145-FEDER-029115 and PTDC/BIA-EVL/29115/2017.

472 RL's post doctoral position was supported by this project under POCI-01-
473 0145-FEDER-029115.

474

475

476 References

477

- 478 1. Edgar R. MUSCLE: multiple sequence alignment with high accuracy
479 and high throughput. *Nucleic Acids Res.* 2004;32: 1792–1797.
480 doi:10.1093/NAR/GKH340
- 481 2. Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P,
482 McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics.*
483 2007;23: 2947–2948. doi:10.1093/BIOINFORMATICS/BTM404
- 484 3. Martin ACR. Viewing multiple sequence alignments with the JavaScript
485 Sequence Alignment Viewer (JSAV). *F1000Research* 2014 3249.
486 2014;3: 249. doi:10.12688/f1000research.5486.1
- 487 4. Gomez J, Jimenez R. Sequence, a BioJS component for visualising
488 sequences. *F1000Research.* 2014;3.
489 doi:10.12688/F1000RESEARCH.3-52.V1
- 490 5. Sanchez-Villeda H, Schroeder S, Flint-Garcia S, Guill KE, Yamasaki M,
491 McMullen MD. DNAAlignEditor: DNA alignment editor tool. *BMC*
492 *Bioinformatics.* 2008;9: 154. doi:10.1186/1471-2105-9-154
- 493 6. Hall T. BioEdit: a user-friendly biological sequence alignment editor and
494 analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser.*
495 1999; 95–98. Available: <https://ci.nii.ac.jp/naid/10030689140/>
- 496 7. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment
497 editor. *Bioinformatics.* 2004;20: 426–427.
498 doi:10.1093/BIOINFORMATICS/BTG430
- 499 8. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et
500 al. Geneious Basic: An integrated and extendable desktop software
501 platform for the organization and analysis of sequence data.
502 *Bioinformatics.* 2012;28: 1647–1649.
503 doi:10.1093/BIOINFORMATICS/BTS199
- 504 9. Sohpal VK, Dey A, Singh A. MEGA biocentric software for sequence
505 and phylogenetic analysis: A review. *Int J Bioinform Res Appl.* 2010;6:
506 230–240. doi:10.1504/IJBRA.2010.034072
- 507 10. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics
508 Viewer (IGV): high-performance genomics data visualization and
509 exploration. *Brief Bioinform.* 2013;14: 178–192.
510 doi:10.1093/BIB/BBS017
- 511 11. Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y.
512 GenomeView: a next-generation genome browser. *Nucleic Acids Res.*
513 2012;40. doi:10.1093/NAR/GKR995
- 514 12. Milne I, Stephen G, Bayer M, Cock P, Pritchard L, Cardle L, et al. Using
515 Tablet for visual exploration of second-generation sequencing data.
516 *Brief Bioinform.* 2013;14: 193–202. doi:10.1093/BIB/BBS012
- 517 13. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC,
518 Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune
519 escape. *Nat Rev Microbiol* 2021 197. 2021;19: 409–424.

- doi:10.1038/s41579-021-00573-0
14. Günthard HF, Calvez V, Paredes R, Pillay D, Shafer RW, Wensing AM, et al. Human Immunodeficiency Virus Drug Resistance: 2018 Recommendations of the International Antiviral Society–USA Panel. *Clin Infect Dis*. 2019;68: 177–187. doi:10.1093/CID/CY463
15. Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R. Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol* 2007 2512. 2007;25: 1407–1410. doi:10.1038/nbt1371
16. Jensen M, Li F, Van 't Wout V, Nickle D, Shriner D, He H, et al. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol*. 2003;77: 13376–13388. doi:10.1128/JVI.77.24.13376-13388.2003
17. Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet* 2019 206. 2019;20: 341–355. doi:10.1038/s41576-019-0113-7
18. Zhao F, Bajic V. The value and significance of metagenomics of marine environments. *Genomics Proteomics Bioinforma*. 2015;13: 271–274. doi:10.1016/j.gpb.2015.10.002
19. Ufarte L, Laville E, Duquesne S, Potocki-Veronese G. Metagenomics for the discovery of pollutant degrading enzymes. *Biotechnol Adv*. 2015;33: 1845–1854. doi:10.1016/j.biotechadv.2015.10.009
20. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 2005 611. 2005;6: 805–814. doi:10.1038/nrg1709
21. Kuiken C, Korber B, Shafer RW. HIV Sequence Databases. *AIDS Rev*. 2003;5: 52. Available: /pmc/articles/PMC2613779/
22. Lorenzo-Redondo R, Fryer H, Bedford T, Kim E, Archer J, Pond S, et al. Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature*. 2016;530: 51–56. doi:10.1038/NATURE16933
23. Archer J, Rambaut A, Taillon B, Harrigan P, Lewis M, Robertson D. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time--an ultra-deep approach. *PLoS Comput Biol*. 2010;6. doi:10.1371/JOURNAL.PCBI.1001022
24. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol Biol Evol*. 2020;37: 291. doi:10.1093/MOLBEV/MSZ189
25. Shapiro B, Rambaut A, Drummond AJ. Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. *Mol Biol Evol*. 2006;23: 7–9. doi:10.1093/MOLBEV/MSJ021
26. Pilcher CD, Wong JK, Pillai SK. Inferring HIV Transmission Dynamics from Phylogenetic Sequence Relationships. *PLoS Med*. 2008;5: 0350–0352. doi:10.1371/JOURNAL.PMED.0050069
27. Knuth D. *The Art of Computer Programming*. 3rd ed. Addison-Wesley; 1997.
28. Moore J, Kitchen S, Pugach P, Zack J. The CCR5 and CXCR4 coreceptors--central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. *AIDS Res Hum Retroviruses*. 2004;20: 111–126. doi:10.1089/088922204322749567

- 570 29. Mild M, Kvist A, Esbjörnsson J, Karlsson I, Fenyo E, Medstrand P.
571 Differences in molecular evolution between switch (R5 to R5X4/X4-
572 tropic) and non-switch (R5-tropic only) HIV-1 populations during
573 infection. *Infect Genet Evol.* 2010;10: 356–364.
574 doi:10.1016/J.MEEGID.2009.05.003
- 575 30. Cardozo T, Kimura T, Philpott S, Weiser B, Burger H, Zolla-Pazner S.
576 Structural basis for coreceptor selectivity by the HIV type 1 V3 loop.
577 *AIDS Res Hum Retroviruses.* 2007;23: 415–426.
578 doi:10.1089/AID.2006.0130
- 579 31. Madeira F, Park Y, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The
580 EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic
581 Acids Res.* 2019;47: W636–W641. doi:10.1093/NAR/GKZ268
- 582 32. Crooks G, Hon G, Chandonia J, Brenner S. WebLogo: a sequence logo
583 generator. *Genome Res.* 2004;14: 1188–1190. doi:10.1101/GR.849004
584
585
586

587 **Supporting Information Captions**

588
589
590 **Table S1: Nucleotide frequencies from the V3 loop.** Sites covering codon 11 and 26
591 are highlighted in red.
592

dissect alignment and export information

load
alignment

sequence
view

select nodes
and use
Node Path
menu

network view

navigate with arrows OR go directly to locations

dynamically alter
diversity network

Sequence identifications

Cluster

Edge

Node

Selected node

Sequence paths

(i) Label clusters and identify required edges using shared sequence id's between clusters

ID Aligned Residues

Left window (edge) nodes

Right window nodes

(ii) Calculate left-hand-side node order permutations and corresponding edge crossovers

(A)

FRQ|NUCLEOTIDES

10 |
 10 |
 12 |
 13 |
 15 |
 17 |
 19 |
 21 |
 22 |
 32 |
 CON|TGTAACAAGACCCCAACAACAATAACAAGAA---AAA-----GTATA---CATAT-----AGGACCAGGGAGAGCATTTTATGCAACAGGAGAAATAATAGGAGATATAAGACAAGCACATTGT

8 |
 7 |
 4 |
 4 |
 4 |
 4 |
 3 |
 2 |
 2 |
 2 |

(B)

FRQ|AMINO ACIDS

10 |CTRPNNNTRKGIHIGPGRAFYATGQIIGDIRQAHC
 10 |CTRPSNNTRKSIINMGPGRAFYTTGEEIIGDIRQAHC
 12 |CTRPNNNTRKSIHLGPGSAIYATGQIIGDIRQAHC
 13 |CTRPNNNTRKSIINIGPGRAFYAATDIIGDIRQAHC
 15 |CTRPSNNNTSGSIHIGPGRAFDATKTIIGDIRQAHC
 17 |CTRPSNNTRKSIINMGPGRAFYTTGEEIIGDIRQAHC
 19 |CVRPGNNTRKSIITIGPGRAFYATGEIIGDIRKAHC
 21 |CTRPNNNTRKSIHIGPGSAFYTTGEEIIGDIRQAHC
 22 |CTRPNNNTRKGIHIGPGRAFYATEKITGDIRQAHC
 32 |CTRPNNNTRKSIINIGPGRAWYATGEIIGNIRQAHC

1 |CTRPNNNTRKSIHIGPGRAFYATGEIIGDIRQAHC

8 |CTRPSNNTRKRVTLGPSRVYTTTGEITGDIRRAHC
 7 |CKRPNNNARRHIHIGPGRAFYATDIIGNIRQAHC
 4 |CIRPNNNTRKGITLGPGRVFTYTTGQIVGDIRKAHC
 4 |CARPNNNTRKRIYMGTRYMSATEKITGDIRQAHC
 4 |CTRPNTYTSKMIRIQRGPGRAFVRTKPSGDIRQAHC
 4 |CTRPNNNTRKGIHIGLGRRVYVTRQIIGDTRKAHC
 3 |CTRPNNYNNKRRIHIGPGRAFYTTKNIKGTIRQAHC
 2 |CTRPNNNTRQRISIGPGRAFYTTQVVGDIRQAHC
 2 |CTRPNNNTRKGIHIGLGRRFYVTVQIVGDVKAHC
 2 |CTRPNNNTRRGIHIGLGRRFYVTVQIVGDVKAHC

(C)

N 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 C
 CTRPNNNTRKSIHIGPGRAFYATGQIIGDIRQAHC

N 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 C
 CTRPNNNTRKSIHIGPGRAFYATGQIIGDIRQAHC