

1 Title: **MicrobeTrace: Retooling Molecular Epidemiology for Rapid Public Health Response**

2 Authors: **Ellsworth M. Campbell^{1,*}, Anthony Boyles^{2,²}, Anupama Shankar¹, Jay Kim², Sergey**

3 **Knyazev^{1,3,4}, William M. Switzer¹**

4 Affiliations

5 ¹**Centers for Disease Control and Prevention, Atlanta, GA 30329**

6 ²**Northrup Grumman, Atlanta, GA 30345**

7 ³**Oak Ridge Institute for Science and Education, Oak Ridge, TN 37830**

8 ⁴**Department of Computer Science, Georgia State University, Atlanta, GA, 30303**

9 Title character count (with spaces): **71**

10 Abstract word count: **165 (of limit 150)**

11 Manuscript word count, excluding figure captions: **4552 (of limit 5000)**

12 Figure caption word count: **490**

13 Conflicts of Interest and Source of Funding: **None**

14 * To whom correspondence should be addressed.

15 ^² Authors contributed equally.

16 **Abstract**

17 **Motivation**

18 Outbreak investigations use data from interviews, healthcare providers, laboratories and surveillance
19 systems. However, integrated use of data from multiple sources requires a patchwork of software that
20 present challenges in usability, interoperability, confidentiality, and cost. Rapid integration, visualization
21 and analysis of data from multiple sources can guide effective public health interventions.

22 **Results**

23 We developed MicrobeTrace to facilitate rapid public health responses by overcoming barriers to data
24 integration and exploration in molecular epidemiology. Using publicly available HIV sequences and other
25 data, we demonstrate the analysis of viral genetic distance networks and introduce a novel approach to
26 minimum spanning trees that simplifies results. We also illustrate the potential utility of MicrobeTrace in

27 support of contact tracing by analyzing and displaying data from an outbreak of SARS-CoV-2 in South
28 Korea in early 2020.

29 **Availability and Implementation**

30 MicrobeTrace is a web-based, client-side, JavaScript application (<https://microbetrace.cdc.gov>) that runs
31 in Chromium-based browsers and remains fully-operational without an internet connection. MicrobeTrace
32 is developed and actively maintained by the Centers for Disease Control and Prevention. The source code
33 is available at <https://github.com/cdcgov/microbetrace>.

34 **Contact: ells@cdc.gov**

35

36

37 1. Introduction

38 The burgeoning field of public health bioinformatics has given rise to a plethora of specialized
39 software for analysis and visualization of pathogen genomic data to aid outbreak investigations (Clément,
40 et al., 2018; Leipzig, 2017). Implementation of these analytic tools can be complex and fraught with a
41 variety of technical and administrative barriers, like faulty install procedures or the need for
42 administrative credentials to install (Sussman, 2007). As a result, routine use of bioinformatic tools in
43 public health can be delayed or blocked because users lack the wide range of skills necessary to install,
44 operate, and integrate them (Pond, et al., 2018). Historically, many public health workers with
45 educational backgrounds in medicine, epidemiology, and laboratory sciences lack informatics skills
46 needed to collect, analyze and display data (*Applications of Clinical Microbial Next-Generation*
47 *Sequencing: Report on an American Academy of Microbiology Colloquium held in Washington, DC, in*
48 *April 2015*, 2015). This skill mismatch tends to be more pronounced at local health departments,
49 representing the frontlines of public health, which have limited capacity and funding for informatics,
50 cyber security, and computational infrastructure (Gwinn, et al., 2017).

51 The complex landscape of public health bioinformatics has necessitated the development of tools
52 designed to sidestep hurdles that can hinder adoption or routine use. Technical and administrative barriers
53 are often reduced by moving complex analytics and computation to off-site servers. However, while cloud
54 computing has revolutionized the healthcare industry (Celesti, et al., 2019), state public health laws often
55 prohibit the storage of sensitive data on off-site servers in the cloud. Tool accessibility can also be
56 hampered by cluttered user interfaces (Bastian, et al., 2009; Hall, 1999; Maths, 2007; Smoot, et al., 2011)
57 and unwieldy workflows that hamper human-computer interaction (Argimón, et al., 2016; Hadfield, et al.,
58 2019; Hadfield, et al., 2018; Pond, et al., 2018). Given the breadth of genetic sequencing technologies and
59 bioinformatic methods, tool adoption can suffer when acceptable input and output file formats are limited,
60 complicating or even preventing integration with existing systems and workflows. To foster adoption and

61 routine use, bioinformatic tools should be secure, easy to use, and capable of accepting or exporting data
62 in commonly used formats.

63 To accommodate the specific needs of local health departments, we developed a standalone but
64 browser-based tool to integrate, visualize and explore data routinely collected during public health
65 investigations of outbreaks and transmission clusters. These data can include case lists describing
66 demographic and behavioral information, case lists with high-risk contacts, in addition to pathogen
67 genomic data. MicrobeTrace was designed to enable users to construct pathogen genetic distance
68 networks and visually integrate them with contact tracing networks to better characterize a transmission
69 network. MicrobeTrace users can further characterize their integrated networks by mapping additional
70 metadata to visual attributes like size, shape and color. In contrast with other tools commonly used for
71 transmission analysis (Argimón, et al., 2016; Hadfield, et al., 2018), all visual attributes can be modified
72 by the user via simple interactions (e.g., dropdown menus, toggle buttons, and color pickers) in real-time,
73 without modification of the underlying data. MicrobeTrace is well suited for working with personally
74 identifiable information (PII) because it performs all computations and visualizations on the user's
75 computer and does not store or transmit any data from the user's computer. When using a supported and
76 updated web browser (e.g., Chrome, Firefox, or Edge) all cached files are cleared when the browser
77 session ends unless caching is explicitly enabled by the user. At no time are user data transmitted
78 anywhere over the internet. As a result, MicrobeTrace can be accessed from the CDC website initially and
79 thereafter used with data stored on the user's computer without an internet connection, making it ideal for
80 rapid visualization of data in the field.

81 Here, we present MicrobeTrace and describe its utility across multiple public health use cases
82 including retrospective analyses and outbreak response. We also report on its use in transmission analysis
83 for a broad spectrum of infectious diseases, such as tuberculosis, viral hepatitis, sexually transmitted
84 diseases as well as special pathogens like SARS-CoV-2 and Ebola.

85

86 **2. Methods**

87 **2.1 Development**

88 MicrobeTrace has been developed according to an agile and open source model, with all code available
89 via GitHub.com (Boyles and Kim, 2018). This enables users to directly observe the rate of development
90 as well as submit and monitor feature requests and system bug reports. MicrobeTrace development has
91 been guided by requirements and features requested by public health practitioners who will use the
92 application in their routine field work. All code is indexed by the federal open source repository
93 (Code.gov, 2019) and promoted by Code.gov (Code.gov, 2019). The MicrobeTrace codebase is regularly
94 scanned by Fortify Software (HP Enterprise Security Products, 2020) and SonarQube (SonarQube.org,
95 2020) to ensure security and code stability. Further, all related modules of code that depend on each other
96 are automatically monitored for vulnerabilities and updated by GitHub's Dependabot service. This
97 automated monitoring service ensures that security vulnerabilities are rapidly detected, reported to our
98 development team, and addressed. GitHub's *Actions* service is used to automate the process of testing
99 newly developed features before official release. This process of automated testing ensures that each time
100 new features are added into MicrobeTrace, all pre-existing functionality are automatically tested prior to
101 an official release.

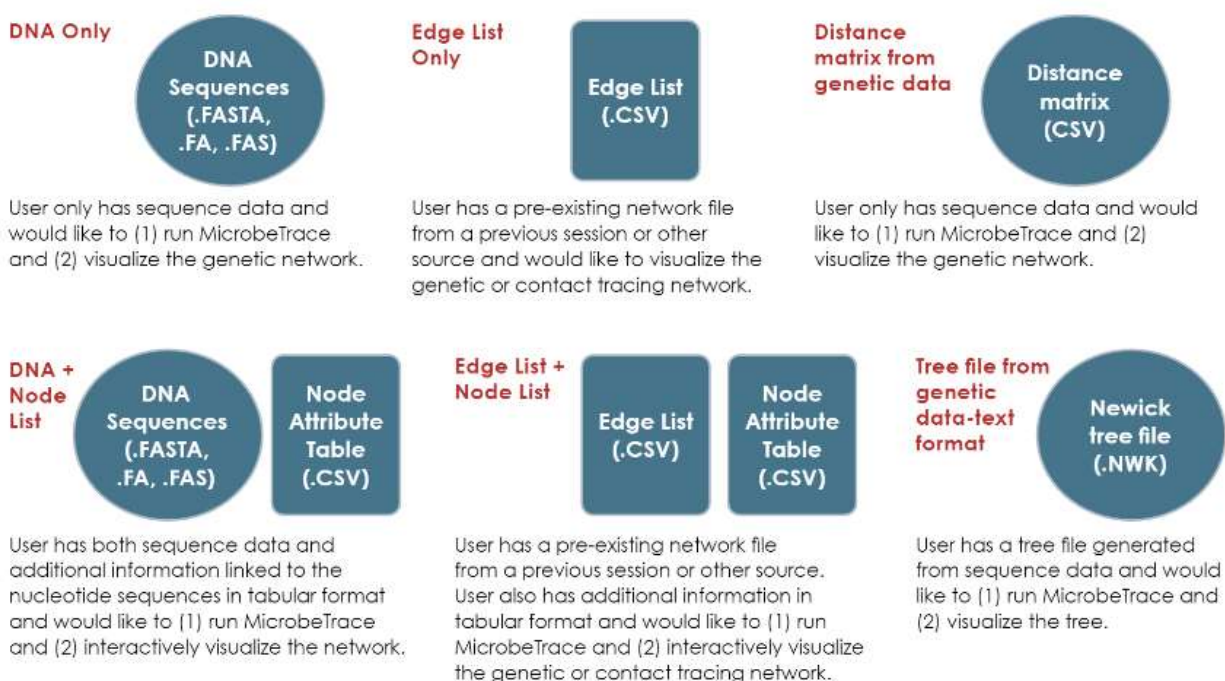
102 **2.1 Outreach**

103 Training and outreach are important factors in refining a software product through interaction with the
104 user base. Training is provided through three modalities: (1) small *ad-hoc* webinar sessions (5-20
105 attendees) to support specific outbreak and cluster investigations, (2) large in-person training sessions
106 (20-100+), and (3) a recorded webinar available via YouTube (CDC, 2020) that is compliant with Section
107 508 of the Rehabilitation Act of 1973. A detailed, 508-compliant >100-page manual is also available for
108 download on the GitHub website (Shankar, et al., 2019) . Finally, a brief 'flyer' describing the tool's
109 general functionality (Campbell, 2019) is available in PDF format, for handout at public health and
110 academic conferences.

111 **3. Results**

112 **3.1 Data Formats**

113 MicrobeTrace handles a variety of file types and formats that are traditionally collected during public
114 health investigations. Pathogen genomic information can be integrated as raw genomic sequences, genetic
115 distance matrices, pairwise genetic distances, or phylogenetic trees. Epidemiologic and other metadata
116 about cases (node lists) and their high-risk contacts (edge or link lists) can be integrated as spreadsheets.
117 Importable in a variety of file formats, these file types can be visualized independently or in-concert to
118 achieve different analytic goals (Fig. 1). Early in an outbreak investigation, high-risk contacts can be
119 combined with other epidemiologic information to visualize and characterize a risk network. When
120 genomic data become available later in the investigation, genetic networks can be integrated to visualize
121 concordance between epidemiologic and laboratory data sources. Alternatively, all available data sources
122 can be integrated to construct a more holistic visualization of an ongoing public health investigation.



123
124 **Figure 1:** *MicrobeTrace accepts input data in a variety of formats. This figure displays the most common*
125 *use cases and their required files.*

126 3.2 Preserving Data Security and Confidentiality

127 The information processing technology within MicrobeTrace is well adapted for use in a public health
128 setting because it prioritizes the confidential but effective use of sensitive data collected during an

129 outbreak investigation. MicrobeTrace was developed as a *client-side only* application that is incapable of
130 transmitting any user data over the internet. In contrast, most web-based bioinformatic applications
131 require the user's data be submitted over the internet for processing by a remote *server-side* application
132 before results can be returned to the user. Local processing is achieved through open source development
133 and translations of traditional bioinformatic algorithms to align (Boyles, 2019a; Li, 2014; Smith, et al.,
134 1981), compare (Boyles, 2019b; Pond, et al., 2018; Tamura and Nei, 1993) , and evaluate genomic
135 sequences and their relationships to one another (Boyles, 2019d; Fourment and Gibbs, 2006; Knyazev,
136 2020; Kruskal, 1956). Importantly, sequence (a) alignment, (b) comparisons, (c) phylogeny, and (d)
137 network evaluations are recapitulations of established methods and do not constitute novel development.
138 Therefore, to the best of our knowledge, the results derived from these JavaScript methods are
139 interchangeable with results derived from their respective, native implementations. A novel extension of
140 the network evaluation method is described below in section 3.4 as the 'Nearest Connected Neighbor'.

141 Visualizations must be generated with care during an outbreak investigation to ensure
142 confidential and narrow use of sensitive data. PII and other sensitive information like geospatial
143 coordinates, zip codes, and phone numbers should only be accessible to Disease Investigation Specialists
144 conducting contact tracing interviews. However, an epidemiologist performing a retrospective analysis
145 can use the same visualization layout with remapped labels, colors, shapes and sizes. Indeed, sensitive
146 geocoordinates can still be used confidentially to produce informative maps by applying the random
147 'jitter' function in MicrobeTrace to reduce the precision of the displayed map marker. In concert, these
148 diverse and accessible controls enable public health experts to safely and confidently leverage sensitive
149 data without risk to the public's confidentiality.

150 **3.3 Genetic Distance Networks**

151 To demonstrate the bioinformatics capacity of MicrobeTrace, we used a publicly available HIV-1 data set
152 consisting of 1,164 sequences of the partial polymerase (*pol*) region (GenBank accession numbers
153 KX465238-KX467180) from a recent study in Germany in addition to associated metadata describing
154 behavioral risk factors and gender (Pouran Yousef, et al., 2016) . Partial *pol* sequences are typically

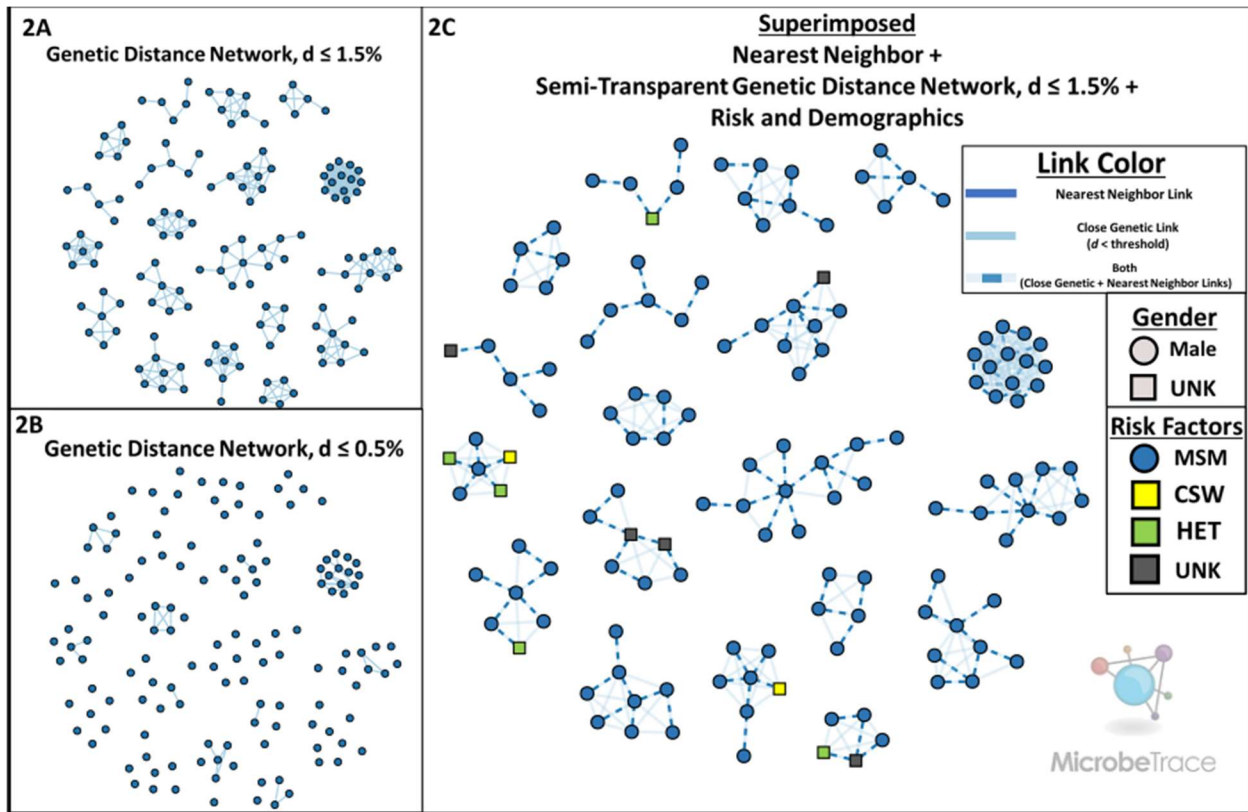
155 collected for determination of antiretroviral drug resistance monitoring for care and treatment for persons
156 living with HIV infection.

157 The bioinformatics workflow of genetic distance networks in MicrobeTrace begins with a pairwise
158 sequence alignment of each input sequence against a reference, according to the Smith-Waterman
159 algorithm (Boyles, 2019a; Li, 2014; Smith, et al., 1981). Multiple sequence alignments are too time
160 constrained and are not used. A user can align to a curated reference, an arbitrary custom reference, or the
161 first input sequence. For HIV-1, the strain HXB2 from the United States (U.S.) is a common reference
162 sequence (GenBank accession number K03455). Once aligned, pairwise genetic distances are calculated
163 according to either a raw hamming distance or the Tamura-Nei substitution model (TN93) (Boyles,
164 2019d; Pond, et al., 2018; Tamura and Nei, 1993) . When the TN93 substitution model is selected,
165 handling of ambiguous bases can be configured as previously described (Pond, et al., 2018) . Pairwise
166 genetic distances can be easily filtered by a threshold defined by the user, in this case 1.5% nucleotide
167 substitutions per site (Fig 2A). Notably, users are empowered with the tools necessary to identify and
168 select the distance threshold value that best fits their public health use case (Wertheim, et al., 2017) . In
169 some situations for HIV-1, a conservative threshold of 1.5% genetic distance might be appropriate to best
170 understand the historical evolution of recent transmission events (Wertheim, et al., 2014) . A more
171 stringent TN93 threshold of 0.5% is often used to identify the most recent and rapid clusters of HIV-1
172 transmission (Fig 2B). Threshold determinations are often informed by cluster size and growth rate
173 criteria (Erly, et al., 2020; France and Oster, 2020; Oster, et al., 2018). MicrobeTrace offers the ability to
174 filter by genetic distance and cluster size thresholds in the same ‘Global Settings’ menu. Here, using the
175 German HIV-1 dataset we have filtered for clusters of size $N \geq 5$ after the 1.5% genetic distance threshold
176 is applied. This filter hides 73.1% ($N = 851$) of individuals that are too genetically distant to cluster with
177 any other sequences in the sample as well as 17.9% ($N = 208$) of individuals whose HIV-1 sequences
178 reside in clusters of size $N \leq 4$. HIV-1 sequences from the remaining 9.0% ($N = 105$) of individuals are
179 displayed as genetic distance networks in Figure 2. Variables of interest can be readily mapped to the

180 nodes or links, including HIV-1 *pol* drug resistance mutations to identify clusters of transmitted drug
181 resistance (Fig 2C).

182 **3.4 Arbitrary Genetic Distance Networks**

183 A simple nucleotide substitution model is not always suitable to understand phylogenetic relationships.
184 Rather than require the use of a single model, MicrobeTrace supports the integration of precomputed
185 distance matrices and pairwise distance lists. A user can provide any pre-computed pairwise distances,
186 regardless of the underlying nucleotide substitution model, as a list or a matrix in order to render those
187 data as a network. For distance matrices, both full matrix and PHYLIP formats are accepted.
188 MicrobeTrace also provides a novel and simple filtering algorithm to render only the nearest connected
189 genetic neighbor(s) for each node, while still maintaining cluster connectivity. Where any two genetically
190 equidistant neighbors are possible, both links are rendered when the ‘Nearest Connected Neighbor’ filter
191 is applied. This approach is particularly useful to understand the historical context of an entire cluster,
192 while focusing on the part of the cluster exhibiting the most concerning and rapid growth. For example,
193 an HIV cluster in rural southeastern Indiana grew rapidly in 2015 but underwent slow growth for nearly a
194 decade prior (Campbell, et al., 2017). The nearest connect neighbor method yields results similar to a
195 non-exhaustive search for all minimum spanning trees, as has been previously described (Bbosa, et al.,
196 2020; Campbell, et al., 2017). The threshold and nearest connected neighbor filters are not mutually
197 exclusive and can therefore be applied simultaneously to ensure that genetically distant nodes remain
198 disconnected. This enables the inclusion of related, but more distant sequences in a cluster visualization
199 while minimizing the information overload typically accompanied by increased distance thresholds (as
200 shown in Fig. 2A). HIV-1 genetic distance links that fell below the 1.5% threshold but were not included
201 as a nearest connected neighbor link are shown at reduced opacity (Fig. 2C).



202

203 **Figure 2:** MicrobeTrace excels at rendering pathogen genetic distance networks and mapping visual
204 characteristics to user-provided metadata. (2A) The HIV-1 partial polymerase (*pol*) distance network,
205 with a genetic distance threshold (d) of 1.5%. (2B) The same HIV-1 *pol* network shown in 2A with node
206 positions held constant, but with a more stringent genetic distance threshold (d) of 0.5%. (2C) The same
207 HIV-1 *pol* network shown in 2A with node positions held constant. Nearest connected neighbor links have
208 been superimposed as dashed lines. The transparency of links that do not connect nearest neighbors has
209 been increased. Gender and transmission risk factors have been mapped to node shape and color,
210 respectively.

211 3.5 Patristic Distance Networks

212 Phylogenies are ubiquitous in public health and bioinformatics, but a phylogeny may be difficult to
213 integrate with more traditional contact tracing data. While powerful new tools are available to integrate
214 taxa-level characteristics into phylogenies, integration of paired contacts is unavailable. Instead, the
215 genetic distances encoded on the phylogeny must be measured and recast as pairwise patristic distances of

216 a phylogeny. Specifically, these are tip-to-tip measurements between individuals on an evolutionary tree
217 that account for the most recent common ancestor. This step is necessary, because it results in a pairwise
218 genetic distance list that is readily integrated with pairwise contact data. Provided a phylogenetic tree in
219 Newick format, MicrobeTrace will traverse the phylogeny to calculate and render the pairwise patristic
220 distance network corresponding to that phylogeny.

221 **3.6 Epidemiologic Networks**

222 Importantly, phylogenies or pathogen genetic sequence data are not required to leverage MicrobeTrace to
223 visualize public health data. MicrobeTrace supports the visualization of arbitrary networks, such as those
224 collected during contact tracing during an outbreak or cluster investigation. Acceptable networks are not
225 limited to person-to-person links but can include person-to-place or place-to-place. To visually
226 differentiate persons from places, MicrobeTrace can style the shape of any network node according to a
227 node type column (e.g., nodeType = ‘Person’ or ‘Place’) defined in the data set. If additional metadata are
228 available to describe a link, it can be colored according to user-defined categorical variables.

229 Alternatively, an option is provided to scale link width according to a user-defined numeric variable or its
230 reciprocal.

231 **3.7 Multi-Layer Networks**

232 Epidemiologic and genetic networks often offer complementary perspectives about transmission clusters
233 (Campbell, et al., 2020) . MicrobeTrace can render an arbitrary number of networks simultaneously by
234 representing multiple overlapping links between pairs of nodes (e.g., hyperlinks) as color-mapped, dashed
235 lines. In addition to independent color-mappings according to underlying data, the effect of a particular
236 network layer can either be hidden or accentuated via independent transparency controls. For example, to
237 protect individual privacy, public health experts may choose to make epidemiologic reports of high-risk
238 contact invisible while rendering only close genetic links when producing figures for public consumption.

239 **3.8 Maps with Network Overlay**

240 Integrated epidemiologic and genetic networks are abstract diagrams that can be used to inform policy
241 and prevention efforts when augmented with additional information. MicrobeTrace can generate

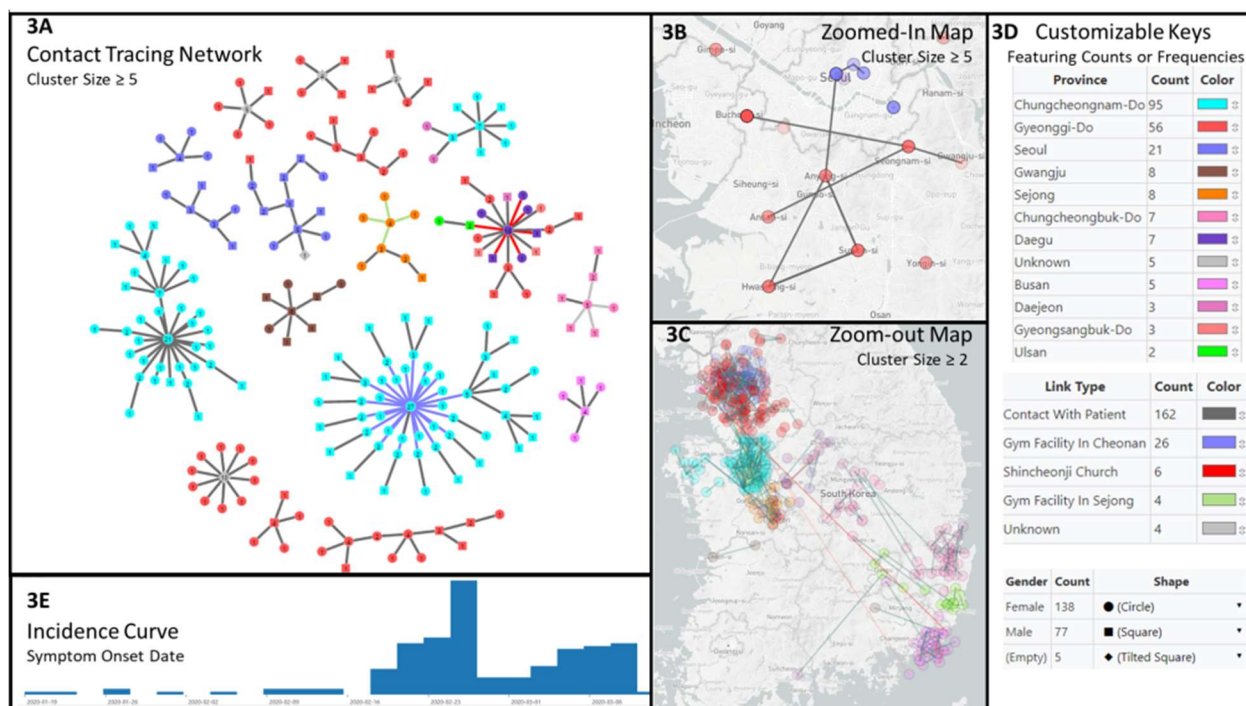
242 choropleth maps, globe diagrams, or more common map projections. MicrobeTrace mapping functions
243 offline with pre-computed shapefiles describing countries, as well as U.S. states and counties. Should
244 internet access be available, MicrobeTrace can be configured to request high-resolution geospatial map
245 tiles from a JavaScript map service called Leaflet (Agafonkin, 2014). MicrobeTrace also enables users to
246 contextualize their maps with a network overlay that maintains all color mappings defined in the network
247 visualization. Users can select from various geographic units, ranging from Country, and – at present –
248 state, county, and zip codes for the U.S. or paired latitude and longitude values. For each geographic
249 level, a marker is placed at the geographic centroid. Over-plotting can be addressed by a combination of
250 automated aggregation or manual transparency tools. Maps can also be customized with user-provided
251 geospatial data in the GeoJSON format.

252 **3.9 Customization and Interactive Exploration**

253 To demonstrate the generalized visualization capacity of MicrobeTrace, we present a publicly available
254 data set describing clinical, demographic and contact tracing data derived from the Korean Centers for
255 Disease Control (KCDC) investigation of the COVID-19 outbreak (Kim, 2020). The data set does not
256 contain coronavirus sequence data, but instead details 383 transmission histories between 510 cases. It
257 also contains an additional 1,627 cases of COVID-19 with no documented transmission histories. As
258 before, using filtering capabilities unique to MicrobeTrace, we limit our visualizations to transmission
259 clusters of size ≥ 5 cases (Fig. 3).

260 MicrobeTrace is centered around integration and visualization of pathogen genomic and network data but
261 is accompanied by an array of customizable tables, charts, and geospatial maps that facilitate exploration
262 and communication of public health data. Each view is interactive and interoperable so nodes in one view
263 are propagated to other tiled views. For example, a node selected by search or click in the **Table View** is
264 highlighted both there and in relevant adjacent views. Similarly, all choices on color-mappings for nodes
265 and links are propagated to all relevant adjacent views. All views are resizable and can be tiled to produce
266 rich, interactive and exploratory dashboards as demonstrated below. We have tiled the COVID-19
267 transmission network, the symptom onset incidence curve, and a geospatial map with transmission

268 network overlay (Fig. 3). Here, we perform the following visual manipulations within MicrobeTrace: (1)
 269 automatically calculate and map the number of contacts for each case to the label that is centered over
 270 each node (Fig. 3A), (2) map the node color to the case's province (Fig. 3A-D), (3) map link color to the
 271 mode of exposure (Fig. 3A-D), (4) map node shapes to the case's gender (Fig. 3A) (5) superimpose the
 272 network onto a high-resolution geospatial 2D map projection (Fig. 3B-C), (6) tailored color, size and
 273 transparency to desired values (Fig. 3B-C), and (7) generated an incidence curve according to the date of
 274 symptom onset (Fig 3E).

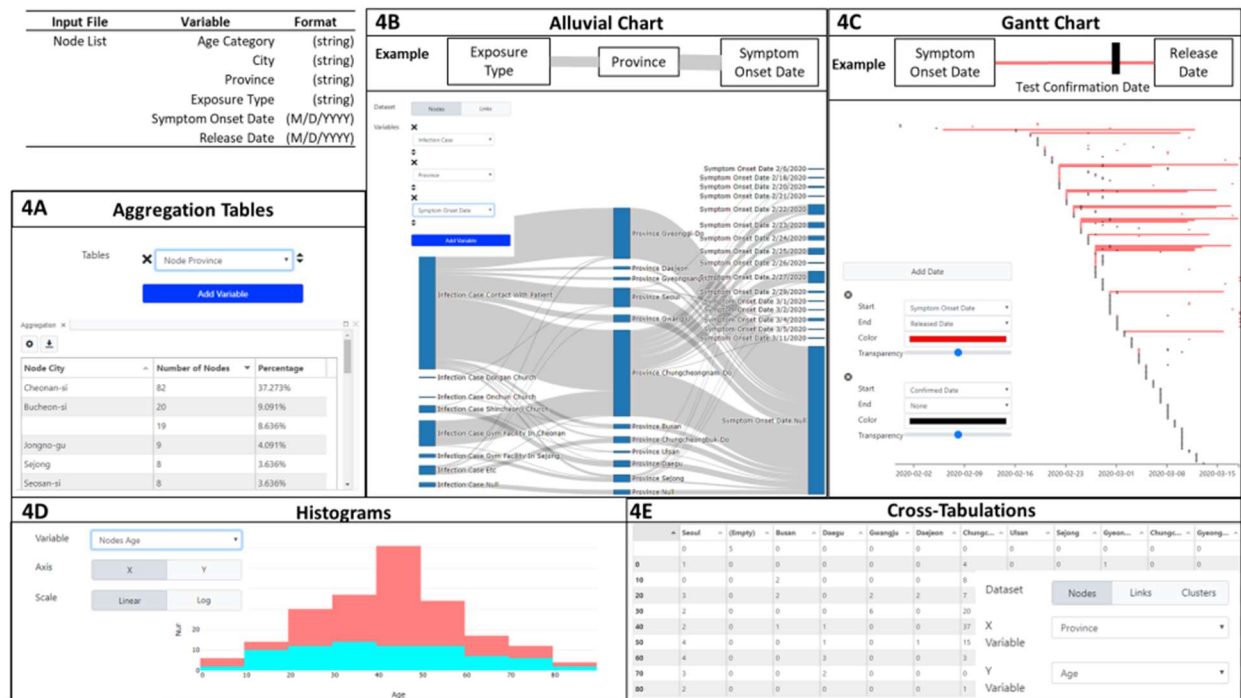


275

276 **Figure 3:** MicrobeTrace allows the creation of informative dashboard visualizations. (3A) Reports of
 277 high-risk contact between COVID-19 cases in clusters of size $N \geq 5$, nodes are (i) colored by province,
 278 (ii) shaped by gender, and (iii) labeled with the total number of high-risk contacts. (3B) Geospatial map
 279 of clusters of size $N \geq 5$ zoomed to show only Seoul, South Korea. (3C) Geospatial map of clusters of size
 280 $N \geq 2$. Node positions have been randomly altered, via MicrobeTrace's 'jitter' functionality, to preserve
 281 patient privacy. (3D) In-application color and shape keys that offer interactive color-pickers and
 282 labeling. (3E) Incidence curve showing symptom onset date.

283

284 As with genetic data, networks are not required to leverage most of the visualizations in MicrobeTrace.
 285 Indeed, MicrobeTrace can be used to achieve rich visualizations using a list of nodes with a handful of
 286 variables like *age*, *gender*, *province*, *city*, *exposure type*, *symptom onset date*, *test confirmation date* and
 287 *hospital release data*. We demonstrate the construction of complex figures like a **Flow Diagram**, **Gantt**
 288 **Chart**, **Cross-tabulation**, **Aggregation**, and **Histogram** with simple dropdown menus (Fig. 4).
 289 Additional diagrams can be achieved with the **2D Network**, **3D Network**, **Scatter Plot**, **Heatmap**,
 290 **Bubbles**, **Choropleth**, and **Globe Views** with relevant data types selected with simple dropdown menus.
 291 Operation of each view is documented in detail in the MicrobeTrace user manual (Shankar, Campbell, *et*
 292 *al.*, 2019).



293
 294 **Figure 4:** MicrobeTrace visualization does not require genomic or contact tracing data and calculate
 295 aggregation and cross-tabulation tables in addition to visualizing histograms, alluvial/flow diagrams and
 296 Gantt charts. Each diagram has an inset settings menu that describes the settings changes necessary to
 297 achieve them. (4A) City-level aggregation achieved via a single dropdown selection. (4B) Alluvial
 298 diagram of associations between the Type of Exposure to COVID-19, Province, and Symptom Onset
 299 Date. (4C) Gantt charts to describe the span of time between Symptom Onset, Positive Test Confirmation,

300 *and Hospital Release Date. (4D) Age histogram, binned by decade and colored by gender. This*
301 *histogram illustrates a trend identified during the early Korean outbreak, wherein a disproportionate*
302 *number of middle-age female cases was diagnosed (4E) Cross-tabulation table of cases by City and Age*
303 *categories.*

304 **3.10 Sequence alignment and phylogenetic tree views**

305 When sequence data are available, a variety of additional diagrams and views are available. For example,
306 the **Sequences View** can be used to export or check the quality of the pairwise alignment. The
307 **Phylogenetic Tree View** will construct a tree via a neighbor-joining algorithm according to the provided
308 pairwise distance calculations. The **Phylogenetic Tree View** has robust customization controls that have
309 been modularized in a separate JavaScript library called TidyTree (Boyles, 2019c).

310 **3.11 Reproducibility**

311 Public health investigations are iterative and the underlying data sources tend to grow over time. Once
312 MicrobeTrace workspaces have been customized they can be saved in two ways: (1) as a custom.
313 MicrobeTrace file or (2) as a “stashed” (cached) browser session. As new data arrives, a user can choose
314 to add new files and recompute the network while pinning nodes to their original positions on-screen.
315 This capability enables a greater understanding of transmission dynamics by enforcing continuity between
316 visualization and exploration sessions over time. Styling parameters and custom visualizations can be
317 stored *independently* from the underlying data as a MicrobeTraceStyle file to facilitate communication
318 between collaborators and preserve confidentiality. Style files can also be used to ensure continuity
319 between public health investigations, such that different investigations yield identically styled
320 visualizations even with different underlying data.

321 **3.12 Data and visualization exports**

322 Communicating data arising from public health investigations is a complex process that requires many
323 fine adjustments, as messages are tuned to their audiences. To meet this need, MicrobeTrace is designed
324 to provide users maximum control over visualization customization and export capabilities. For example,
325 communication to academic and public health audiences often involves poster presentations that require

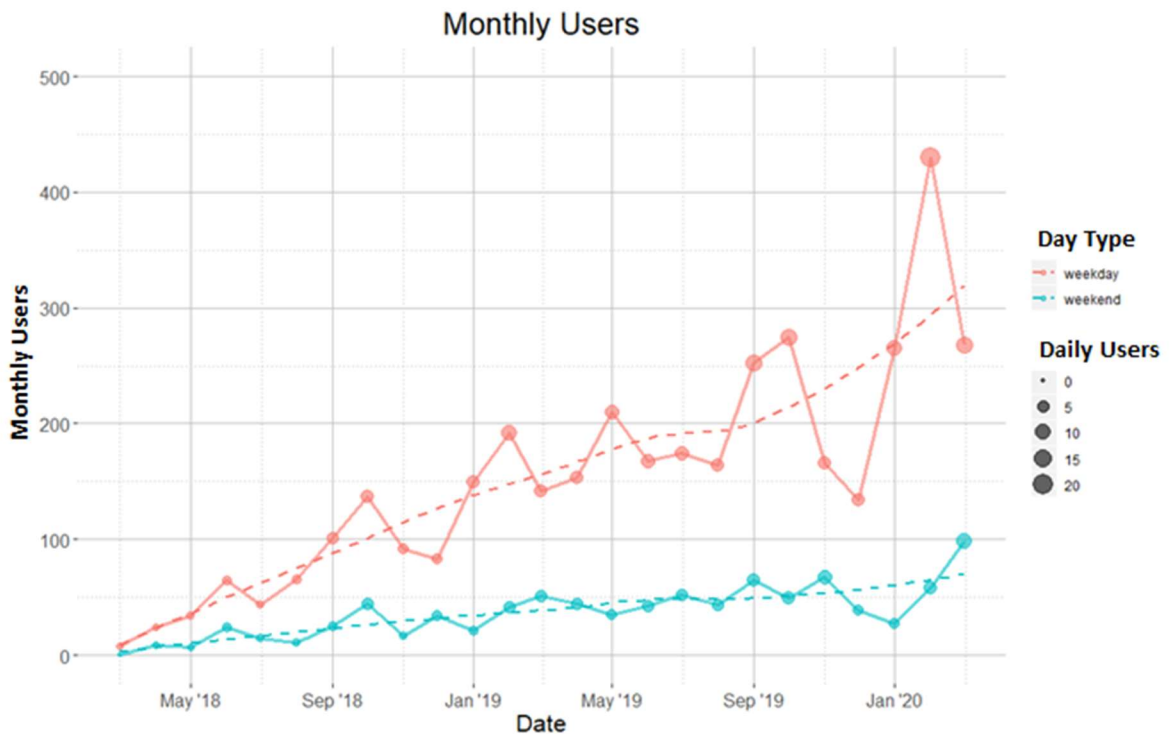
326 images be scaled-up for large printer formats. We accommodate this requirement by enabling users to set
327 specific export resolutions for PNG and JPEG formats. Alternatively, visualizations can be exported as
328 Scalable Vector Graphics (SVGs) that can be enlarged to any arbitrary size without a loss of resolution.
329 By default, a MicrobeTrace watermark is placed on images exported from MicrobeTrace; however, the
330 transparency of the watermark can be increased using a menu slider to render it invisible. Taken together,
331 these capabilities offer publication-ready image exports for scientific journals.

332 MicrobeTrace maximizes interoperability with other applications by enabling the export of all
333 calculated and integrated datasets. The **Table View** renders tabular data which can be exported to comma-
334 separated (CSV) and Excel (XLS, XLSX) formats. The node-level table includes all information joined
335 from multiple input data sources as well as calculated fields like a node's number of neighbors ('degree')
336 and its cluster ID. The link-level table also includes calculated fields; for example, whether a link was
337 identified as a 'nearest connected neighbor' as a Boolean result. MicrobeTrace offers robust filtering and
338 selection capabilities that are also reflected in exported tables, 'Selected' and 'Visible' states are shown as
339 Boolean results. Tables produced in the **Aggregation View** can be exported as formatted PDFs, CSVs, a
340 zipped collection of CSVs, or an XLS/XLSX workbook where each aggregation is shown on
341 independently named worksheets (Fig. 4A). Data derived from the **Map, Globe, and Choropleth Views**
342 can be exported as GeoJSON files for interoperability with other Geographic Information System (GIS)
343 software. Genomic sequence alignment can be exported in the FASTA or MEGA file formats in the
344 **Sequences View**.

345 **3.13 Statistics and analysis of MicrobeTrace usage**

346 While some public health investigations that leveraged MicrobeTrace have been reported in the
347 academic literature, many use cases supporting public health missions are never intended for publication
348 or dissemination (Cranston, et al., 2019; Hogan, et al., 2017; John, et al., 2019; Shankar, et al., 2019;
349 Falade-Nwulia, et al., 2018) . To better understand that broad base of engagement, MicrobeTrace usage
350 statistics are captured and reported by region via Google Analytics. When MicrobeTrace is accessed
351 while the user is online, an anonymous Google Analytics cookie is sent along with information about the

352 user's rough geolocation and usage time. It is important to note that, offline usage is not tracked by
353 Google Analytics. Since the launch of MicrobeTrace in March 2018, 2,642 unique users have connected
354 for a total of 6,501 sessions (2.46 sessions per user) for a combined 738.6 hours of use (6.8 per session
355 and 16.7 minutes per user). The overwhelming majority of users connect from the U.S. (N = 2,323,
356 87.8%) with the most prevalent international use coming from China (N = 55, 2.1%), the United
357 Kingdom (N = 38, 1.4%), and Vietnam (N = 30, 1.1%). 50 additional countries account for the remaining
358 6.6% (N = 196) of users. Usage increases on weekdays, as the public health workforce goes to work, and
359 the mean number of weekday users has increased from 1.1/weekday in February 2018 to highs of 20.5
360 and 14.6 per weekday in February and March 2020, respectively. (Fig. 5). Notably, as much of the
361 world's public health workforce has turned its attention to COVID-19 in February and March of 2020,
362 MicrobeTrace usage peaked (Fig. 5).
363



364
365 **Figure 5:** *MicrobeTrace's primary user base are public health officials during the work week, as*
366 *opposed to during the weekend. In red, are the number of monthly weekday users. In teal, are the number*

367 of monthly weekend users. Each month's mean daily user count is mapped to the size of the circle and
368 colored by day type. A local regression for each day type is shown to smooth the month-to-month effects
369 and highlight the increasing trend.

370
371 A notable influx of MicrobeTrace usage occurred in late April 2020 (data not included in figure),
372 simultaneously across nine cities in Vietnam over a span of two local afternoon hours. This brief influx of
373 traffic from a single country, spread across disparate geography, is suggestive of workforce development
374 efforts. If true, this would represent the first clear evidence of a training webinar held by non-CDC staff.
375 Following on from this training event, the fraction of returning users was three times higher than
376 MicrobeTrace's historical fraction of returning users (64% versus 21%). Further, the average session
377 duration was also nearly three times higher (20.1min versus 7.3min) than the historic average session
378 duration.

379 4. Discussion

380 MicrobeTrace has been used to investigate a broad variety of infectious diseases. It has been used during
381 CDC-assisted HIV cluster investigations in multiple states (Cranston, et al., 2019; Hogan, et al., 2017;
382 John, et al., 2019; Shankar, et al., 2019), investigations of hepatitis C virus (HCV) (Falade-Nwulia, et al.,
383 2018)), integrated into the Global Hepatitis Outbreak and Surveillance Technology (GHOST) that is used
384 for viral hepatitis investigations (Longmire, et al., 2017) (S. Sims, personal communication), and is
385 broadly used to integrate genomic and epidemiologic data for tuberculosis outbreak investigations
386 (Springer, 2020). It has also been used to integrate partner services, epidemiologic and whole genome
387 data to better understand transmission during a retrospective public health investigation of *Neisseria*
388 *gonorrhoeae* (Town, et al., 2020). Outside of its intended domain of sexually transmitted diseases,
389 MicrobeTrace has also been applied to integrate epidemiologic and laboratory data in outbreaks of
390 foodborne pathogens, such as *Escherichia coli* O157:H7 (Allen, 2020). It is currently being evaluated for
391 integration and visualization of epidemiologic and genetic data from cases of Ebola and COVID-19 (S.
392 Whitmer, personal communication; S. Tong, personal communication).

393 MicrobeTrace offers a suite of capabilities to a public health expert that are typically only
394 achievable with an array of software, tools, and custom scripts, and substantive computational experience.
395 A putative MicrobeTrace user, such as epidemiologists or disease investigation specialist, typically
396 achieves proficiency after one brief training session and aided by a cursory understanding of common
397 browser interactions, such as ‘dropdown menus’, ‘slider bars’, and ‘drag-and-drop’. Many standalone
398 tools are available to calculate pairwise genetic distances with varying degrees of specificity to the
399 pathogen of interest. MEGA is a bioinformatic tool broadly used in public health, but new users can be
400 overwhelmed by dense interfaces with scores of options that are often dense with jargon and required
401 inputs (Kumar, et al., 2008). HIV-TRACE, which is specific to HIV sequence data, now offers rich
402 visualization capabilities but its installation requires a keen understanding of Unix and the Git protocol
403 for local installation and use (Pond, et al., 2018) . An iteration of HIV-TRACE is available on the internet
404 but at a web server which has concomitant data security issues (Weaver, et al., 2015). Patristic distance
405 calculations are available via the APE package in R or the Java application PATRISTIC, but these require
406 programming expertise and software installations (Fourment and Gibbs, 2006; Paradis, et al., 2004) .
407 Once genetic relationships have been calculated and contacts have been traced, integration and
408 visualization of these links with individual-level data can be a complex task requiring tools like Gephi or
409 Cytoscape (Bastian, et al., 2009; Smoot, et al., 2011) . For those with programming expertise, integrated
410 visualizations can be otherwise achieved with decade-old libraries in R with the iGraph package or in
411 Python with the NetworkX and MatPlotLib packages (Csardi and Nepusz, 2006; Hagberg, et al., 2008;
412 Hunter, 2007). Even so, these visualizations are not interactive with any additional figures, charts, tables,
413 and maps that a public health expert might need to generate through the use of over a half dozen other
414 applications (Figs. 2-4). If independently created, these visualizations must be augmented with network-
415 level calculations and manipulations like threshold changes, minimum spanning tree calculations and
416 filters, cluster membership, cluster size, and the number of neighbors for each node, all of which are
417 easily performed in MicrobeTrace. These metrics can be manually calculated (e.g., R+iGraph,
418 Python+NetworkX) or generated via opaque plug-ins in Gephi or Cytoscape that offer minimal

419 customizations. Anecdotally, use of MicrobeTrace and its network layout interface can be playful; which
420 has been shown to improve the user experience and increase their motivation to use the tool (Kuts, 2009).

421 While MicrobeTrace has been developed for a public health user base, it also has many
422 applications in academia. It is adept at integrating arbitrary networks with independent node- and edge-
423 level characteristics that are necessary to evaluate social, behavioral, biochemical, cellular, technological
424 and physical networks. MicrobeTrace also offers rich customizations that reduce the time and effort to
425 achieve insights and discoveries when grappling with a novel data set. The MicrobeTrace development
426 team is not aware of another tool that offers all of these capabilities in a secure, interoperable, and light-
427 weight format that requires no installation prior to use.

428 **Contributions**

429 EMC and WMS contributed to design, project management, and manuscript writing. AB contributed to
430 design, development, and manuscript writing. AS contributed to design, user manual, and manuscript
431 editing. JK contributed to development. SK contributed to design and provided the nearest connected
432 neighbor methodology.

433 **Acknowledgements**

434 We are thankful to our colleagues in the Division of Tuberculosis Elimination (Kathryn Winglee, Sarah
435 Talarico, Yuri Springer, Benjamin Silk), the Division of STD Prevention (Kim Gernert, Katy Town,
436 Matthew Schmerer), the Division of Viral Hepatitis (Seth Sims, Garrett Atkinson, Yury Khudyakov),
437 National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention - Informatics Office (Max
438 Mirabito, Silver Wang), Transmission and Molecular Epidemiology Team (Alexandra Oster, Cheryl
439 Ocfemia, Nivedha Panneer, Scott Cope, Sheryl Lyss) for providing valuable feedback, features, bug
440 reports, and continued training of our public health partners. We are also thankful to our user base in
441 public health and academia for reporting bugs and suggesting features with regularity.

442 **Disclaimers**

443 Use of trade names is for identification only and does not imply endorsement by the U.S. Centers for
444 Disease Control and Prevention (CDC). The findings and conclusions in this report are those of the
445 authors and do not necessarily represent the views of the CDC.

446 **Funding**

447 We are thankful to the CDC's Advanced Molecular Detection initiative for providing intramural funding
448 for this project.

449

450 **References**

451 Agafonkin, V. Leaflet: an open-source JavaScript library for mobile-friendly interactive maps. In.; 2014.

452 p. 2016. <https://leafletjs.com/>

453

454 Allen, K. Visualizing sequence data and epidemiological data together using MicrobeTrace. In, *Integrated*

455 *Foodborne Outbreak Response and Management Conference*. 2020.

456

457 Applied Maths. BioNumerics version 5.10. 2007.

458

459 Argimón, S., *et al.* Microreact: visualizing and sharing data for genomic epidemiology and

460 phylogeography. *Microb Genom* 2016;2(11):e000093. <https://microreact.org/>

461

462 Bastian, M., Heymann, S. and Jacomy, M. Gephi: an open source software for exploring and

463 manipulating networks. In, *Third international AAAI conference on weblogs and social media*. 2009.

464

465 Bbosa, N., *et al.* Phylogenetic and Demographic Characterization of Directed HIV-1 Transmission Using

466 Deep Sequences from High-Risk and General Population Cohorts/Groups in Uganda. *Viruses* 2020;12(3).

467

468 Boyles, A. 2019a. AlignmentViewer. Release 1.0. <https://github.com/CDCgov/AlignmentViewer>.

469 (2020/4/2 date last accessed).

470

471 Boyles, A. 2019b. patristic. Release 1.0. <https://github.com/CDCgov/patristic>. (2020/4/2 date last

472 accessed).

473

474 Boyles, A. 2019c. TidyTree. Release 1.0. <https://github.com/CDCgov/TidyTree>. (2020/4/7 date last

475 accessed).

476

477 Boyles, A. 2019d. tn93.js. Release 1.0. <https://github.com/CDCgov/tn93.js>. (2020/4/2 date last accessed).

478

479 Boyles, A. and Kim, J. 2018. MicrobeTrace. <https://github.com/CDCgov/MicrobeTrace>. (2020/4/6 date
480 last accessed).

481

482 Campbell, E.M., MicrobeTrace Flyer. 2019.

483 <https://github.com/CDCgov/MicrobeTrace/blob/master/docs/MicrobeTrace%20Flyer.pdf>.

484

485 Campbell, E.M., *et al.* Detailed Transmission Network Analysis of a Large Opiate-Driven Outbreak of
486 HIV Infection in the United States. *J. Infect. Dis.* 2017;216(9):1053-1062.

487

488 Campbell, E.M., *et al.* Phylodynamic Analysis Complements Partner Services by Identifying Acute and
489 Unreported HIV Transmission. *Viruses* 2020;12(2).

490

491 CDC. NCHHSTP MicrobeTrace Webinar Full. In.: Centers for Disease Control and Prevention; 2020.

492 https://www.youtube.com/watch?v=5E-_Kb7yvHU

493

494 Celesti, A., Amft, O. and Villari, M. Guest Editorial Special Section on Cloud Computing, Edge

495 Computing, Internet of Things, and Big Data Analytics Applications for Healthcare Industry 4.0. *IEEE*

496 *Trans. Ind. Inf.* 2019;15(1):454-456.

497

498 Clément, L., *et al.* A data-supported history of bioinformatics tools. *arXiv [cs.DL]* 2018.

499

500 Code.gov. MicrobeTrace : The Visualization Multitool for Molecular Epidemiology and Bioinformatics.

501 2019. https://code.gov/search?page=1&query=microbetrace&size=10&sort=best_match

502

503 Code.gov. Rooftop Recommendations #02: MicrobeTrace. In.: Centers for Disease Control and
504 Prevention; 2019. [https://medium.com/@CodeDotGov/rooftop-recommendations-02-microbetrace-](https://medium.com/@CodeDotGov/rooftop-recommendations-02-microbetrace-63504b73838)
505 [63504b73838](https://medium.com/@CodeDotGov/rooftop-recommendations-02-microbetrace-63504b73838)

506

507 Cranston, K., *et al.* Notes from the field: HIV diagnoses among persons who inject drugs—Northeastern
508 Massachusetts, 2015–2018. *MMWR* 2019.

509

510 Csardi, G. and Nepusz, T. The igraph software package for complex network research. *InterJournal,*
511 *complex systems* 2006;1695(5):1-9.

512

513 Erly, S.J., *et al.* Characterization of Molecular Cluster Detection and Evaluation of Cluster Investigation
514 Criteria Using Machine Learning Methods and Statewide Surveillance Data in Washington State. *Viruses*
515 2020;12(2).

516

517 Falade-Nwulia, O., *et al.* CLUSTERING OF HEPATITIS C VIRUS INFECTION AMONG PEOPLE
518 WHO INJECT DRUGS IN BALTIMORE. In, *Conference on Retroviruses and Opportunistic Infections.*
519 CROI; 2018. <https://www.croiconference.org/>

520

521 Fourment, M. and Gibbs, M.J. PATRISTIC: a program for calculating patristic distances and graphically
522 comparing the components of genetic change. *BMC Evol. Biol.* 2006;6:1.

523

524 France, A.M. and Oster, A.M. The Promise and Complexities of Detecting and Monitoring HIV
525 Transmission Clusters. *J. Infect. Dis.* 2020;221(8):1223-1225.

526

527 Gwinn, M., MacCannell, D.R. and Khabbaz, R.F. Integrating Advanced Molecular Technologies into
528 Public Health. *J. Clin. Microbiol.* 2017;55(3):703-714.
529
530 Hadfield, J., *et al.* Twenty years of West Nile virus spread and evolution in the Americas visualized by
531 Nextstrain. *PLoS Pathog.* 2019;15(10):e1008042.
532
533 Hadfield, J., *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;34(23):4121-
534 4123.
535
536 Hagberg, A., Swart, P.J. and Schult, D.A. Exploring network structure, dynamics, and function using
537 NetworkX. In.: Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2008.
538
539 Hall, T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for
540 Windows 95/98/NT. In, *Nucleic acids symposium series*. 1999. p. 95-98.
541
542 Hogan, V., *et al.* HIV TRANSMISSION POTENTIAL DUE TO INJECTION DRUG USE IN RURAL
543 WEST VIRGINIA, US, 2017. In, *Conference on Retroviruses and Opportunistic Infections 2017*. CROI;
544 2017. <https://www.croiconference.org/>
545
546 Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 2007;9(3):90-95.
547 John, B., *et al.* MOLECULAR SURVEILLANCE AS A MEANS TO EXPAND AN OUTBREAK
548 INVESTIGATION: MA, 2015-2018. In, *Conference on Retroviruses and Opportunistic Infections*.
549 CROI; 2019. <https://www.croiconference.org/>
550
551 Kim, J. 2020. Data-Science-for-COVID-19. <https://github.com/jihoo-kim/Data-Science-for-COVID-19>.
552 (2020/4/7 date last accessed).

553

554 Knyazev, S. 2020. epsilon Minimal Spanning Trees (eMST). Release 1.0. [https://github.com/Sergey-](https://github.com/Sergey-Knyazev/eMST)
555 [Knyazev/eMST](https://github.com/Sergey-Knyazev/eMST). (2020/4/2 date last accessed).

556

557 Kruskal, J.B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc.*
558 *Am. Math. Soc.* 1956;7(1):48-50.

559

560 Kumar, S., *et al.* MEGA: a biologist-centric software for evolutionary analysis of DNA and protein
561 sequences. *Brief. Bioinform.* 2008;9(4):299-306.

562

563 Kuts, E. Playful User Interfaces: Literature Review and Model for Analysis. In, *Proceedings of Digital*
564 *Games Research Association*. Nokia; 2009.

565

566 Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* 2017;18(3):530-536.

567

568 Li, H. 2014. bioseq-js. <https://github.com/lh3/bioseq-js>. (2020/4/2 date last accessed).

569

570 Longmire, A.G., *et al.* GHOST: global hepatitis outbreak and surveillance technology. *BMC Genomics*
571 2017;18(Suppl 10):916.

572

573 Oster, A.M., *et al.* Identifying Clusters of Recent and Rapid HIV Transmission Through Analysis of
574 Molecular Surveillance Data. *J. Acquir. Immune Defic. Syndr.* 2018;79(5):543-550.

575

576 Paradis, E., Claude, J. and Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language.
577 *Bioinformatics* 2004;20(2):289-290.

578

579 Pond, S.L.K., *et al.* HIV-TRACE (TRANSMISSION Cluster Engine): a Tool for Large Scale Molecular
580 Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. *Mol. Biol. Evol.* 2018;35(7):1812-1819.
581
582 Pouran Yousef, K., *et al.* Inferring HIV-1 Transmission Dynamics in Germany From Recently
583 Transmitted Viruses. *J. Acquir. Immune Defic. Syndr.* 2016;73(3):356-363.
584
585 Products, H.P.E.S. 2020. Fortify Software. https://en.wikipedia.org/wiki/Fortify_Software
586
587 Shankar, A., *et al.* MicrobeTrace User Manual. 2019.
588
589 Shankar, A., *et al.* Clusters of Diverse HIV and Novel Recombinants Identified Among Persons Who
590 Inject Drugs in Kentucky and Ohio. In, *14th Annual International HIV Transmission Workshop*. Virology
591 Education; 2019.
592
593 Smith, T.F., Waterman, M.S. and Fitch, W.M. Comparative biosequence metrics. *J. Mol. Evol.*
594 1981;18(1):38-46.
595
596 Smoot, M.E., *et al.* Cytoscape 2.8: new features for data integration and network visualization.
597 *Bioinformatics* 2011;27(3):431-432.
598
599 SonarQube.org. 2020. SonarQube. Release 7.9.3. <https://www.sonarqube.org/>. (2020/4/6 date last
600 accessed).
601
602 Springer, Y. Logically Inferred Tuberculosis Transmission (LITT) Algorithm User's Manual - Appendix
603 3. 2020.
604

605 Sussman, G.J. Building robust systems an essay. *Citeseer* 2007;113:1324.

606

607 Tamura, K. and Nei, M. Estimation of the number of nucleotide substitutions in the control region of
608 mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 1993;10(3):512-526.

609

610 Town, K., *et al.* Phylogenomic analysis of *Neisseria gonorrhoeae* transmission to assess sexual mixing
611 and HIV transmission risk in England: a cross-sectional, observational, whole-genome sequencing study.
612 *The Lancet infectious diseases* 2020;20(4):478-486.

613

614 Weaver, S., *et al.* 2015. Datamonkey. <http://hivtrace.datamonkey.org/hivtrace>. (2020/4/6 date last
615 accessed).

616

617 Wertheim, J.O., *et al.* The global transmission network of HIV-1. *J. Infect. Dis.* 2014;209(2):304-313.

618

619 Wertheim, J.O., *et al.* Social and Genetic Networks of HIV-1 Transmission in New York City. *PLoS*
620 *Pathog.* 2017;13(1):e1006000.

621