

1 **DisCVR: Rapid viral diagnosis from high-throughput sequencing data**

2 Maha Maabar*¹, Andrew J. Davison¹, Fiona Thorburn², Rory Gunson³, Massimo Palmarini¹ and
3 Joseph Hughes*¹

4
5
6 ¹ *MRC-University of Glasgow Centre for Virus Research, Sir Michael Stoker Building, 464*
7 *Bearsden Road, Glasgow G61 1QH, UK*

8 ² *Microbiology department, Glasgow Royal Infirmary*

9 ³ *West of Scotland Specialist Virology Centre, Glasgow Royal Infirmary*

10 **To whom correspondence should be addressed*

11 *Contact: Joseph.Hughes@glasgow.ac.uk or Maha.Maabar@glasgow.ac.uk*

12
13 Total word count: 4708 (5217 with tables)

14 15 **Abstract**

16 High-throughput sequencing (HTS) enables most pathogens in a clinical sample to be detected
17 from a single analysis, thereby providing novel opportunities for diagnosis, surveillance and epi-
18 demiology. However, this powerful technology is difficult to apply in diagnostic laboratories be-
19 cause of its computational and bioinformatic demands. We have developed DisCVR, which de-
20 tects known human viruses in clinical samples by matching sample *k*-mers (22 nucleotide se-
21 quences) to *k*-mers from taxonomically labelled viral genomes. DisCVR was validated using pub-
22 lished HTS data for 89 clinical samples from adults with upper respiratory tract infections. These
23 samples had been tested for viruses metagenomically and also by real-time polymerase chain
24 reaction assay, which is the standard diagnostic method. DisCVR detected human viruses with
25 high sensitivity (79%) and specificity (100%), and was able to detect mixed infections. Moreover,
26 it produced results comparable to those in a published metagenomic analysis of 177 blood sam-
27 ples from patients in Nigeria. DisCVR has been designed as a user-friendly tool for detecting
28 human viruses from HTS data using computers with limited RAM and processing power, and in-
29 cludes a graphical user interface to help users interpret and validate the output. It is written in
30 Java and is publicly available from <http://bioinformatics.cvr.ac.uk/discvr.php>.

31
32 **Keywords:** Virus, diagnosis, high-throughput sequencing, *k*-mer.

33 **Issue Section:** Resources

34 1. Introduction

35 The standard method for rapidly detecting known human viruses in clinical samples is the
36 polymerase chain reaction (PCR), in which short oligonucleotides are used to amplify and probe
37 specific regions of viral genomes. The limitations of this technique include the targeting of a rela-
38 tively small number of viruses per assay and a dependence on sequence conservation among
39 viral strains. High-throughput sequencing (HTS) provides approaches to viral diagnosis that have
40 much greater scope. Thus, metagenomic analysis of HTS data can provide extensive viral geno-
41 typing information, as well as the characterization of complex multiple infections (Thorburn et al.
42 2015). Several metagenomic pipelines using *de novo* assembly and homology matching have
43 been developed for virus detection (Li et al. 2016; Maarala et al. 2018; Ren et al. 2017; Scheuch
44 et al. 2015; Wang et al. 2013; Zheng et al. 2017). However, analysing HTS data using such ap-
45 proaches brings heavy computing and bioinformatic demands that are difficult to meet and
46 standardize in diagnostic laboratories (Flygare et al. 2016). As a consequence, we have devel-
47 oped DisCVR, which is a fast, accurate and easy-to-use tool for detecting known human viruses
48 in clinical samples.

49 DisCVR employs an abundance-based method, which is a metagenomic approach for
50 rapidly profiling the organisms present in a sample. It works by creating a database of short nu-
51 cleotide sequences (*k*-mers) from a large set of viral reference sequences, tagging the *k*-mers
52 taxonomically according to the viruses from which they came, screening each read in the HTS
53 dataset for the presence of virus *k*-mers, and organising a summary of the viruses present in the
54 sample via the tags. This approach makes data analysis very efficient, thereby minimizing the
55 computing effort required (Orton et al. 2016).

56 Several existing tools utilize the abundance-based method to classify the reads in an
57 HTS dataset. NBC (Rosen et al. 2011) employs a naïve Bayesian classifier to assign a log-
58 likelihood score to each read. This classifier is trained by using a set of unique profiles of 15 nu-
59 cleotide *k*-mers from microbial genomes, and then allows users to upload the dataset to a web
60 site and obtain a summary of results listing the best taxonomic match for each read. Kraken
61 (Wood & Salzberg 2014) assigns each *k*-mer in the database to the last common ancestor of
62 species having that *k*-mer, and then assigns each read to the taxon with the most matching *k*-
63 mers. CoMeta (Kawulok & Deorowicz 2015) creates a database of all *k*-mers for each rank in the
64 taxonomic tree, and then uses these databases to classify the reads at each rank. CLARK (Ounit
65 et al. 2015) collects target-specific *k*-mer sets from reference genomes belonging to a certain
66 taxonomic rank (e.g. genus), and then classifies reads at that rank. This approach reduces the
67 database size but requires a different database to be built for each rank. To improve the accura-
68 cy of the classification, CSSSCL (Borozan & Ferretti 2016) creates a BLAST database, a *k*-mer
69 database and a compression database from a collection of reference genomes. Sequences in
70 the sample are classified according to a combined sequence similarity score (CSSS) (Borozan et

71 al. 2015) calculated from information in the pre-computed databases. In contrast to Kraken,
72 CLARK and CoMeta, all of which assign individual reads, MetaPalette (Koslicki & Falush 2016)
73 profiles the entire dataset and returns the relative proportions of organisms present by using *k*-
74 mer sizes of 30 and 50, based on the rationale that using two different *k*-mer sizes allows strain-
75 level variation to be captured more accurately. Taxonomer (Flygare et al. 2016) compares each
76 read to multiple reference databases, assigning it to a high-level taxonomic category on the ba-
77 sis of the *k*-mer content of the read, and then uses exact *k*-mer matching to assign each read to
78 a reference by maximizing the total *k*-mer weight. This weight, which is a function of the *k*-mer
79 count in the reference and the database, provides a database-specific measure of how likely it is
80 that a *k*-mer originated from a particular reference sequence.

81 Despite the growing number and popularity of *k*-mer-based classification tools, they have
82 limitations. The databases are built using a limited set of reference sequences and therefore are
83 of restricted utility for classifying organisms with sequences that diverge from the reference. This
84 limitation can be a particular problem when significant variation exists in an organism at strain
85 level. It can be addressed by incorporating a range of variants into the database, but this then
86 creates a much larger database that may make the analysis challenging to run on resource-
87 limited computers. Furthermore, many of the current tools are run on Linux systems and hence
88 require the operator to have expertise in command line usage and an understanding of bioinfor-
89 matics, which may be difficult to find in diagnostic settings. To our knowledge, the only tool that
90 has been developed for ease of use and for application on computers with limited resources is
91 Truffle (Visser et al. 2016). This is designed to screen for a limited set of user-specified viruses,
92 comes preloaded with probe-sets for grapevine viruses, and cannot easily be updated for large
93 sets of viruses from other hosts.

94 Here, we present DisCVR, a *k*-mer-based classification tool for detecting known human
95 viruses from HTS data derived from clinical samples. DisCVR can be installed on a desktop
96 computer to allow diagnostic laboratories to analyze large, confidential datasets by using a sim-
97 ple, straightforward graphical user interface (GUI) without specialized bioinformatics expertise. It
98 is optimized to run on Windows, Linux and Mac OS, using minimal RAM and processing power
99 without compromising speed and accuracy. The tool currently integrates curated viral databases
100 at the taxonomic levels of species and strain, but may be used to build a customised database at
101 any taxonomic level, thereby overcoming the limitations of using a restricted set of reference se-
102 quences. DisCVR utilizes *k*-mer counts derived from an entire HTS dataset to detect the viruses
103 present in a sample, and validates the results by showing the coverage and depth of reads map-
104 ping to a reference sequence.

105

106 **2. Methods**

107

108 **2.1 The k -mer databases**

109 A k -mer is a short sequence of k nucleotides. A k -mer dataset is generated iteratively by
110 sliding a window of size k along a sequence one nucleotide at a time. Extracting k -mers and
111 counting their frequencies in a set of sequences can be computationally intensive, especially
112 when k is large and the sequences are numerous. Dedicated k -mer counting programs, such as
113 Jellyfish (Marçais & Kingsford 2011) and Khmer (Zhang et al. 2014), can be incorporated into
114 abundance-based tools in order to optimize speed. KAnalyze (Audano & Vannberg 2014) was
115 chosen for integration into DisCVR because the k -mers it generates are sorted lexicographically,
116 thus making the search for matches very efficient. DisCVR also uses the canonical representa-
117 tion of a k -mer, which is lexicographically the smaller of a k -mer and its reverse complement.

118 For the purpose of this study, we define a virus k -mer as a k -mer that uniquely represents
119 a virus or set of related viruses, to the exclusion of the host. A shared k -mer is defined as a k -
120 mer that is common to a virus and the host. By excluding shared k -mers, it is not necessary for
121 the user to remove host reads before using DisCVR, thus speeding up the overall processing
122 time. If k is small, many copies of shared k -mers are generated, and if k is large, many copies of
123 virus k -mers are found. Choosing the optimal k -mer size depends on balancing the advantages
124 of speed (small k) with those of specificity and sensitivity (large k). Furthermore, it is necessary
125 to reduce the number of low-complexity k -mers in the virus k -mer database, as these may be
126 repetitive in sequence and present in otherwise unrelated viruses. The filtering of low-complexity
127 k -mers and the selection of the size of k is explained in Supplementary Section S1.

128 For constructing the virus k -mer databases, three comprehensive datasets of complete or
129 partial viral sequences were extracted from the NCBI taxonomy database. The first, the human
130 hemorrhagic virus dataset (shortened below to “hemorrhagic dataset”), contained 33,367 se-
131 quences of the hemorrhagic fever viruses listed by the Centers for Disease Control and Preven-
132 tion (‘Centers for Disease Control and Prevention’ n.d.). The second, the human respiratory virus
133 dataset (“respiratory dataset”), contained 442,282 sequences of viruses associated with respira-
134 tory disease. The third, the human pathogenic virus dataset (“pathogenic dataset”), consisted of
135 1,762,968 sequences of viruses identified in the UK Health and Safety Executive list of biological
136 agents (‘Health and Safety Executive: The approved list of biological agents.’ 2013).

137

138 **2.2 Database build**

139 DisCVR operates via three modules concerned with database build, sample classification
140 and validation (Fig. 1).

141 Currently, the database build module includes three virus k -mer databases, derived from
142 the hemorrhagic, respiratory and pathogenic datasets, for use in the sample classification mod-
143 ule. In addition, some of the sequences in these datasets, defined largely by their presence in
144 the NCBI RefSeq database, are used as a set of reference genome sequences in the validation
145 module. DisCVR also allows the user to create customised databases and sets of reference se-

146 quences. The database build module involves selecting the relevant viral dataset, collecting the
147 *k*-mers, and removing those that are shared with the host or are of low complexity. Each remain-
148 ing *k*-mer is then identified with a taxonomic tag and an indication of the number of times it oc-
149 curs in the sequences. The *k*-mers are further subdivided into those that exist in a single virus
150 (i.e. specific *k*-mers) and those that exist in multiple viruses (i.e. non-specific *k*-mers). These as-
151 signments are made at the level of species and strain and are used in the output to illustrate the
152 degree of specificity of the *k*-mers matching a virus (Fig. 2).

153

154 **2.3 Sample classification**

155 To analyze an HTS dataset, the file is loaded into DisCVR via the GUI. The *k*-mers are
156 extracted and their frequencies are calculated, the single copy and low-complexity *k*-mers are
157 filtered out, and the remaining *k*-mers are compared with the chosen virus *k*-mer database. As
158 the number of *k*-mers in the sample can be enormous, various data structures were considered
159 to optimize the classification on machines with limited RAM. Although searching the trie is fast
160 $O(n)$, where n is the size of the *k*-mer, it requires $O(n^2)$ overall time to build, and the space need-
161 ed is quadratic. Instead, DisCVR uses a fast searching algorithm that groups similar *k*-mers to-
162 gether. Briefly, the *k*-mers in the virus database are divided among smaller sub-files according to
163 the first five nucleotides. The same procedure is used to divide the *k*-mers derived from the en-
164 tire HTS dataset. Searching commences by loading the corresponding sub-files from the virus *k*-
165 mer database and the sample *k*-mers into memory, and performing a binary search for the pres-
166 ence of each sample *k*-mer among the database *k*-mers. Only matched *k*-mers are retrieved.
167 Finally, DisCVR displays a straightforward list of all the virus hits detected, along with summary
168 statistics and taxonomic information on the sample *k*-mers (Fig. 2).

169

170 **2.4 Validation**

171 DisCVR helps the user to assess the significance of the findings by facilitating an exami-
172 nation of *k*-mer distribution (allowing up to three mismatches) across a reference sequence rep-
173 resenting the target genome. As an alternative, it also incorporates an examination of sequence
174 read distribution carried out by using Tanoti (Sreenu n.d.), which is a BLAST-guided, reference-
175 based short read aligner that is particularly tolerant of mismatches. In each case, the output is a
176 graph showing the depth and coverage of *k*-mers or sequence reads across the reference ge-
177 nome and a summary of statistics for the mapping results (Fig. 3).

178

179 **2.5 Accuracy**

180 The respiratory database was used to analyse published RNA-seq data from nasopha-
181 ryngeal swab samples ($n = 89$) that had been collected from adults with upper respiratory tract
182 infections (Thorburn et al. 2015) (Table S2; the average number of reads per sample was
183 660,640, range 30,872-1,278,122). The samples had been tested using a standard real-time

184 PCR (RT-PCR) assay for human rhinovirus (HRV), influenza viruses A and B (IFA/IFB), respira-
185 tory syncytial virus (RSV), adenovirus (ADV), human metapneumovirus (hMPV), parainfluenza
186 viruses (PIV) 1-4, and human coronaviruses (HCoV) HKU1, NL63, OC43 and 229E (Thorburn et
187 al. 2015). The top hit for each sample (i.e. the virus having the greatest number of distinct *k*-
188 mers) using DisCVR was compared to the virus detected previously by RT-PCR. The samples
189 were also classified using three independent *k*-mer-based programs that require command-line
190 usage on a Linux operating system: Kraken (Wood & Salzberg 2014), KrakenHLL (Breitwieser &
191 Salzberg 2018) and CLARK (Ounit et al. 2015).

192 The initial objective was to determine the number of distinct *k*-mers that would maximize
193 both sensitivity (effectiveness in identifying samples containing viruses) and specificity (effec-
194 tiveness in identifying samples lacking viruses) for DisCVR. The output of DisCVR was catego-
195 rized on the basis of the number of distinct *k*-mers for the top hit, and that of the other programs
196 was assessed on the basis of the number of reads assigned to the top hit. For each tool, sensi-
197 tivity and specificity were defined as $TP/(TP+FN)$ and $TN/(TN+FP)$, respectively, where TP, FN,
198 TN and FP are the number of true positive, false negative, true negative and false positive sam-
199 ples relative to the RT-PCR results. We define samples as i) true positive when the top virus hit
200 was detected by both RT-PCR and DisCVR, ii) true negative when neither RT-PCR nor DisCVR
201 detected a virus, iii) false negative when a virus was detected by RT-PCR but not by DisCVR,
202 and iii) false positive when a virus was detected by DisCVR but not by RT-PCR. ROC curves
203 were generated for DisCVR, Kraken, KrakenHLL and CLARK using the pROC package in R and
204 Youden's statistic (Youden 1950).

205

206 **2.6 Application**

207 DisCVR was used to analyse 177 HTS RNA-seq libraries derived from serum specimens
208 collected in Nigeria from healthy individuals ($n = 120$) and patients with unexplained acute febrile
209 illness ($n = 57$) and analysed in a previous study (Stremlau et al. 2015). The raw data were
210 downloaded from SRA BioProject PRJNA271229. The top hit using DisCVR was compared to
211 the viral reads identified using BLASTn and BLASTx in the original study
212 (<https://doi.org/10.1371/journal.pntd.0003631.s017>).

213

214 **3. Results**

215 The ROC curve (Fig. 4) derived from the datasets from respiratory tract infections
216 (Thorburn et al. 2015) compares the sensitivity and specificity for different *k*-mer thresholds. It
217 suggests that a value of 850 *k*-mers is the optimal threshold on the basis of the point on the
218 curve furthest from the identity (diagonal) line (Table S2). The ROC curves of DisCVR and the
219 other programs (Fig. 4) did not differ significantly from each other, and had overlapping confi-
220 dence intervals. Kraken and KrakenHLL had identical curves. Kraken and CLARK rated as

221 slightly more sensitive but less specific than DisCVR as a result of HCoV NL63 being the top hit
222 in sample 1D3 and the second hit in DisCVR (Table 1 and Table S2). The top hit in DisCVR was
223 HRV-A, which was the second hit in Kraken and CLARK but was not detected using RT-PCR. It
224 was not informative to compare average execution time and memory usage for the programs, as
225 it is not possible to run CLARK, Kraken and KrakenHLL natively on Windows operating systems.

226 A total of 48/89 (54%) of the samples had been shown to contain viruses by RT-PCR,
227 and the remaining 41/89 lacked all viruses tested. Considering only the samples in the set of 89
228 for which DisCVR identified ≥ 850 k -mers for the top hit, the following findings were made.
229 DisCVR identified the viruses that were detected by RT-PCR in 32/48 (67%) of samples (true
230 positives). It did not detect viruses in samples in which no viruses had been found by RT-PCR in
231 22/41 (54%) of samples (true negatives). It detected viruses in samples in which no viruses had
232 been detected by RT-PCR in 19/41 (46%) of samples (false positives), and either detected vi-
233 ruses that did not correspond with those detected by RT-PCR or did not find any virus with ≥ 850
234 k -mers in 16/48 (33%) of samples (false negatives).

235 The RT-PCR assay was limited by the range of viruses that it could detect, by its de-
236 pendence on sequence conservation, and consequently also by its potential to identify infections
237 by multiple viruses. Consequently, the false positive results were assessed using the validation
238 module (Table 2), and the false negative results were investigated by examining the second hits
239 recorded by DisCVR (Table 1). In most false positive cases, the validation module showed that
240 there were multiple reads mapping to several regions of the reference genome, thus confirming
241 the presence of the viruses identified even though they had not been detected by RT-PCR.
242 Some samples had low coverage because a single reference sequence (from RefSeq) repre-
243 sented the entire species but diverged in sequence from the virus present in the sample. For ex-
244 ample, sample 1B3 yielded HRV-A89 (the reference for species *Rhinovirus A*) as the top hit, with
245 only 7.6% genome coverage and 4 mapped reads. Using the capability of DisCVR to build a cus-
246 tomized database drawn from the ≥ 100 prototypic strains of *Rhinovirus A*, HRV-A49 was re-
247 vealed as the top hit, with 81.71% genome coverage and 263 mapped reads. This dramatic im-
248 provement illustrates the potential to strengthen the validation module by adding user-specific
249 curated sets of sequences or by the proposed expansion of RefSeq entries capturing a greater
250 degree of diversity (Brister et al. 2015). In the 16 false negative cases, DisCVR detected the vi-
251 rus identified by RT-PCR as the top hit in three samples (1G2, 1I5 and 2B6), but the number of
252 distinct k -mers was < 850 (Table 1 and Table S2). In addition, the virus identified by RT-PCR was
253 detected as the second hit in 10 samples (1B5, 1D3, 1E5, 1G1, 1F7, 1F8, 2A2, 2B9, 2C4 and
254 2D3), and, in one case (1C2), the RT-PCR assay did not have the potential of identifying the top
255 hit (enterovirus D). An important finding was made in two of these samples (1B5 and 1D3), in
256 which the viruses detected by RT-PCR were not the top hits but still had ≥ 850 distinct k -mers in
257 the sample (Table 1). This suggests that these patients were infected by multiple viruses. Finally,
258 DisCVR did not detect any k -mers for the virus detected by RT-PCR in two samples (1C9 and

259 2D4), but identified HRV-A in 1C9, which was validated by reference assembly. The validation
260 module thus yielded strong evidence for the presence of the viruses detected by DisCVR, at
261 least where the number of *k*-mers was ≥ 850 . These findings were taken into account in reas-
262 ssuming the sensitivity and specificity of DisCVR at 79% and 100%, respectively (Fig. 4).

263 The threshold of 850 *k*-mers was also used in the analysis of the Nigerian datasets
264 (Stremlau et al. 2015). The top hit from DisCVR was the same as that from the BLAST results in
265 the original study for 101/177 (57%) cases, and viruses were detected in both healthy ($n = 68$)
266 and afebrile ($n = 33$) patients (Table S4). In nine cases, the top hit from DisCVR differed from the
267 top BLAST hit, but the second hit matched. In 55 cases, the number of *k*-mers was below the
268 threshold in DisCVR, and the number of reads with BLAST matches was also low (an average of
269 24 reads per dataset). In the remaining 12 discordant samples, DisCVR detected human immu-
270 nodeficiency virus 1 ($n = 9$), XMRV-related virus ($n = 1$) and human T-lymphotropic virus 1 ($n = 1$)
271 as the top hit, whereas the BLAST results supported the presence of human adenovirus or
272 Heterosigma akashiwo RNA virus (an algal virus). Mapping of reads to reference genomes sug-
273 gested that the DisCVR and BLAST hits are false positives.

274

275 **4. Discussion**

276 Using HTS in diagnostic settings offers many advantages, including the ability to sequence
277 pathogen genomes both individually and as communities. However, the uptake of HTS in such
278 settings has been slow, due partly to the cost, turnover time and bioinformatic demands of this
279 technology. We developed DisCVR to help address these challenges. DisCVR is a fast, accurate
280 program for detecting viruses from HTS data using the increasingly exploited approach of *k*-mer
281 classification. It offers the advantage of a non-targeted approach and also enables typing below
282 the species level (e.g. subtype, serotype, genotype or strain). Unlike other tools for detecting vi-
283 ruses from HTS data, DisCVR is easy to use in diagnostic settings through the graphical user
284 interface, requires no bioinformatic expertise, and can be used on the Windows operating sys-
285 tems that are commonly used in diagnostic laboratories. The basic output is easy to interpret,
286 and the advanced output provides more detailed statistics and a validation capability.

287 DisCVR was designed for detecting known viruses and cannot be used to discover novel vi-
288 ruses. Indeed, the paper on the Nigerian patients (Stremlau et al. 2015) reported novel
289 rhabdoviruses in healthy patients using a metagenomic approach, and these were not detected
290 by DisCVR. However, metagenomics requires bioinformatic infrastructure and expertise at levels
291 that are not commonly available in diagnostic laboratories. Nonetheless, DisCVR enables the
292 detection of 148 pathogenic human viruses using one of the three implemented datasets (the
293 pathogenic dataset), and more using the others. This represents a greater than ten-fold increase
294 in target species over multiplex RT-PCR. Moreover, the number of viruses incorporated into the
295 DisCVR databases is flexible, and can also be expanded by building custom databases.

296 In the datasets from respiratory tract infections, DisCVR had high sensitivity and specificity
297 levels but did not identify all the viruses detected by RT-PCR when the threshold of ≥ 850 *k*-mers
298 was used. This threshold may be set by the user and was calculated for the respiratory dataset
299 for which we had paired RT-PCR and HTS data. As more datasets with paired information be-
300 come available, it will be possible to tune the threshold more accurately to specific sample types
301 and sizes. Further efforts could also be made to calibrate DisCVR from artificially constructed
302 communities of viruses in various proportions.

303 Finally, DisCVR is configured as a human viral diagnostic tool, but could be readily expanded
304 to include non-viral human pathogens and pathogens with non-human hosts by using the cus-
305 tom-build scripts in the DisCVR distribution.

306

307 **Supplementary data**

308

309 **Conflict of interest:** The authors declare that they have no competing interests.

310

311 **Acknowledgements**

312 We thank the members of the Viral Genomics and Bioinformatics Group for continuous, insightful
313 feedback on DisCVR, and in particular Sejal Modha for generous support with utilizing NCBI
314 tools and testing DisCVR. We also thank David Manlove for advice on algorithm design. This
315 work was funded by the Medical Research Council (MC_UU_12014/12).

316

317 **Availability and requirements**

318 Source code is available on github <https://centre-for-virus-research.github.io/DisCVR/> and data-
319 bases and executables are available on <http://bioinformatics.cvr.ac.uk/discvr.php>.

320

321 **References**

- 322 Audano, P., & Vannberg, F. (2014). 'KANalyze: a fast versatile pipelined k-mer toolkit.',
323 *Bioinformatics (Oxford, England)*, 30/14: 2070–2. DOI: 10.1093/bioinformatics/btu152
- 324 Borozan, I., & Ferretti, V. (2016). 'CSSSCL: a python package that uses combined sequence
325 similarity scores for accurate taxonomic classification of long and short sequence reads.',
326 *Bioinformatics (Oxford, England)*, 32/3: 453–5. DOI: 10.1093/bioinformatics/btv587
- 327 Borozan, I., Watt, S., & Ferretti, V. (2015). 'Integrating alignment-based and alignment-free
328 sequence similarity measures for biological sequence classification.', *Bioinformatics*
329 *(Oxford, England)*, 31/9: 1396–404. DOI: 10.1093/bioinformatics/btv006
- 330 Breitwieser, F. P., & Salzberg, S. L. (2018). 'KrakenHLL: Confident and fast metagenomics
331 classification using unique k-mer counts', *bioRxiv*.
- 332 Brister, J. R., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015). 'NCBI viral genomes resource.',

- 333 *Nucleic acids research*, 43/Database issue: D571-7. DOI: 10.1093/nar/gku1207
- 334 'Centers for Disease Control and Prevention'. Retrieved December 15, 2014, from
- 335 <<https://www.cdc.gov/vhf/index.html>>
- 336 Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E. H., et al. (2016).
- 337 'Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection
- 338 and host mRNA expression profiling.', *Genome biology*, 17/1: 111. DOI: 10.1186/s13059-
- 339 016-0969-1
- 340 'Health and Safety Executive: The approved list of biological agents.' (2013). . Retrieved
- 341 December 14, 2014, from <<http://www.hse.gov.uk/pubns/misc208.pdf>>
- 342 Kawulok, J., & Deorowicz, S. (2015). 'CoMeta: classification of metagenomes using k-mers.',
- 343 *PloS one*, 10/4: e0121453. DOI: 10.1371/journal.pone.0121453
- 344 Koslicki, D., & Falush, D. (2016). 'MetaPalette: a k-mer Painting Approach for Metagenomic
- 345 Taxonomic Profiling and Quantification of Novel Strain Variation.', *mSystems*, 1/3. DOI:
- 346 10.1128/mSystems.00020-16
- 347 Li, Y., Wang, H., Nie, K., Zhang, C., Zhang, Y., Wang, J., Niu, P., et al. (2016). 'VIP: an
- 348 integrated pipeline for metagenomics of virus identification and discovery', *Scientific*
- 349 *Reports*, 6/1: 23774. DOI: 10.1038/srep23774
- 350 Maarala, A. I., Bzhalava, Z., Dillner, J., Heljanko, K., & Bzhalava, D. (2018). 'ViraPipe: scalable
- 351 parallel pipeline for viral metagenome analysis from next generation sequencing reads.',
- 352 *Bioinformatics (Oxford, England)*, 34/6: 928–35. DOI: 10.1093/bioinformatics/btx702
- 353 Marçais, G., & Kingsford, C. (2011). 'A fast, lock-free approach for efficient parallel counting of
- 354 occurrences of k-mers.', *Bioinformatics (Oxford, England)*, 27/6: 764–70. DOI:
- 355 10.1093/bioinformatics/btr011
- 356 Orton, R. J., Gu, Q., Hughes, J., Maabar, M., Modha, S., Vattipally, S. B., Wilkie, G. S., et al.
- 357 (2016). 'Bioinformatics tools for analysing viral genomic data.', *Revue scientifique et*
- 358 *technique*, 35/1: 271–85. DOI: 10.20506/rst.35.1.2432
- 359 Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). 'CLARK: fast and accurate
- 360 classification of metagenomic and genomic sequences using discriminative k-mers.', *BMC*
- 361 *genomics*, 16: 236. DOI: 10.1186/s12864-015-1419-2
- 362 Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). 'VirFinder: a novel k-mer
- 363 based tool for identifying viral sequences from assembled metagenomic data', *Microbiome*,
- 364 5/1: 69. DOI: 10.1186/s40168-017-0283-5
- 365 Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). 'NBC: the Naive Bayes
- 366 Classification tool webserver for taxonomic classification of metagenomic reads.',
- 367 *Bioinformatics (Oxford, England)*, 27/1: 127–9. DOI: 10.1093/bioinformatics/btq619
- 368 Scheuch, M., Höper, D., & Beer, M. (2015). 'RIEMS: a software pipeline for sensitive and
- 369 comprehensive taxonomic classification of reads from metagenomics datasets', *BMC*
- 370 *Bioinformatics*, 16/1: 69. DOI: 10.1186/s12859-015-0503-6

371 Sreenu, V. B. 'Tanoti'. Retrieved from <<http://bioinformatics.cvr.ac.uk/tanoti.php>>
372 Stremlau, M. H., Andersen, K. G., Folarin, O. A., Grove, J. N., Odiya, I., Ehiane, P. E., Omoniwa,
373 O., et al. (2015). 'Discovery of Novel Rhabdoviruses in the Blood of Healthy Individuals from
374 West Africa', (C. E. Rupprecht, Ed.) *PLoS Neglected Tropical Diseases*, 9/3: e0003631.
375 DOI: 10.1371/journal.pntd.0003631
376 Thorburn, F., Bennett, S., Modha, S., Murdoch, D., Gunson, R., & Murcia, P. R. (2015). 'The use
377 of next generation sequencing in the diagnosis and typing of respiratory infections.', *Journal*
378 *of clinical virology*, 69: 96–100. DOI: 10.1016/j.jcv.2015.06.082
379 Visser, M., Burger, J. T., & Maree, H. J. (2016). 'Targeted virus detection in next-generation
380 sequencing data using an automated e-probe based approach', *Virology*, 495: 122–8. DOI:
381 10.1016/j.virol.2016.05.008
382 Wang, Q., Jia, P., & Zhao, Z. (2013). 'VirusFinder: Software for Efficient and Accurate Detection
383 of Viruses and Their Integration Sites in Host Genomes through Next Generation
384 Sequencing Data', *PLoS ONE*, 8/5: e64465. DOI: 10.1371/journal.pone.0064465
385 Wood, D. E., & Salzberg, S. L. (2014). 'Kraken: ultrafast metagenomic sequence classification
386 using exact alignments.', *Genome biology*, 15/3: R46. DOI: 10.1186/gb-2014-15-3-r46
387 Youden, W. J. (1950). 'Index for rating diagnostic tests.', *Cancer*, 3/1: 32–5.
388 Zhang, Q., Pell, J., Canino-Koning, R., Howe, A. C., & Brown, C. T. (2014). 'These are not the k-
389 mers you are looking for: efficient online k-mer counting using a probabilistic data
390 structure.', *PloS one*, 9/7: e101271. DOI: 10.1371/journal.pone.0101271
391 Zheng, Y., Gao, S., Padmanabhan, C., Li, R., Galvez, M., Gutierrez, D., Fuentes, S., et al.
392 (2017). 'VirusDetect: An automated pipeline for efficient virus discovery using deep
393 sequencing of small RNAs', *Virology*, 500: 130–8. DOI: 10.1016/j.virol.2016.10.017
394

395

Table 1. Results of the second hits in the respiratory samples.

Sample	RT-PCR diagnosis	DisCVR top hit and (no.) ¹	DisCVR second hit and (no.) ¹
Top hit with ≤850 k-mers matching			
1G2	PIV-3	PIV-3 (366)	HRV-A (149)
1I5	HRV	HRV-A (749)	HRV-C (470)
2B6	RSV	RSV (742)	IFA H3N2 (262)
Second hit with ≥850 k-mers matching			
1B5	PIV-3	HRV-A (3,758)	PIV-3 (3,111)
1D3	HCoV NL63	HRV-A (2,420)	HCoV NL63 (1,841)
Second hit with ≤850 k-mers matching			
1C2	HRV	Enterovirus D (1,633)	HRV-A (269)
1E5	RSV	HRV-C (1,777)	RSV (415)
1F8	HCoV NL63	HRV-B (3,876)	HCoV NL63 (724)
2B9	HRV	RSV (1,105)	HRV- C (94)
2A2	HCoV 229E	HRV-C (770)	HCoV 229E (176)
2C4	HCoV 229E	HRV-A (264)	HCoV 229E (5)
2D3	HCoV OC43	HRV-A (438)	HCoV OC43 (135)
1F7	HRV	hMPV (27)	HRV-B (20)
1G1	ADV/HRV	HCoV OC43 (163)	HRV-B (118)
Not detected			
1C9	hMPV	HRV-A (3,083)	Enterovirus D (7)
2D4	PIV-2	HRV-A (579)	HCoV OC43 (225)

¹Number of *k*-mers matching the classification. Hits with ≥850 k-mers are shown in bold.

396

397

Table 2. Coverage of reference genomes of the top hits detected in false positive samples in the respiratory samples.

Sample	Virus detected by DisCVR	Matched <i>k</i> -mers ¹	Genome coverage (%)	No. mapped reads (%) ²
1B3	HRV-A	3,431	7.6	4 (0.00)
1B4	HRV-A	3,652	9.39	14 (0.00)
1B6	HRV-A	2,872	6.38	16 (0.00)
1B9	HRV-A	1,041	2.15	1404 (0.10)
1C8	HRV-A	2,781	8.21	8 (0.00)
1D2	HRV-A	2,974	9.38	13 (0.00)
1D5	HRV-C	901	3.63	8 (0.00)
1D6	HRV-C	1,103	3.27	5 (0.99)
1E2	HRV-C	1,299	1.51	1 (0.00)
1E4	HRV-C	1,813	4.8	7 (0.00)
1E9	HRV-B	4,306	13.69	27 (0.01)
1G7	HRV-B	1,447	1.76	5 (0.00)
1H5	HRV-B	932	3.84	4 (0.00)
1I7	HRV-C	1,234	1.51	1 (0.00)
1I9	HRV-C	1,845	3.1	9 (0.00)
2A1	RSV	2,123	13.37	172 (0.02)
2B5	RSV	927	13.56	69 (0.01)
2B8	RSV	1,406	8.64	101 (0.01)
2D1	HRV-C	1,620	1.59	2 (0.00)

¹Number of matching *k*-mers identified by the classification module.

²Percentage of total reads mapped by the validation module.

398

399 **Figure legends**

400

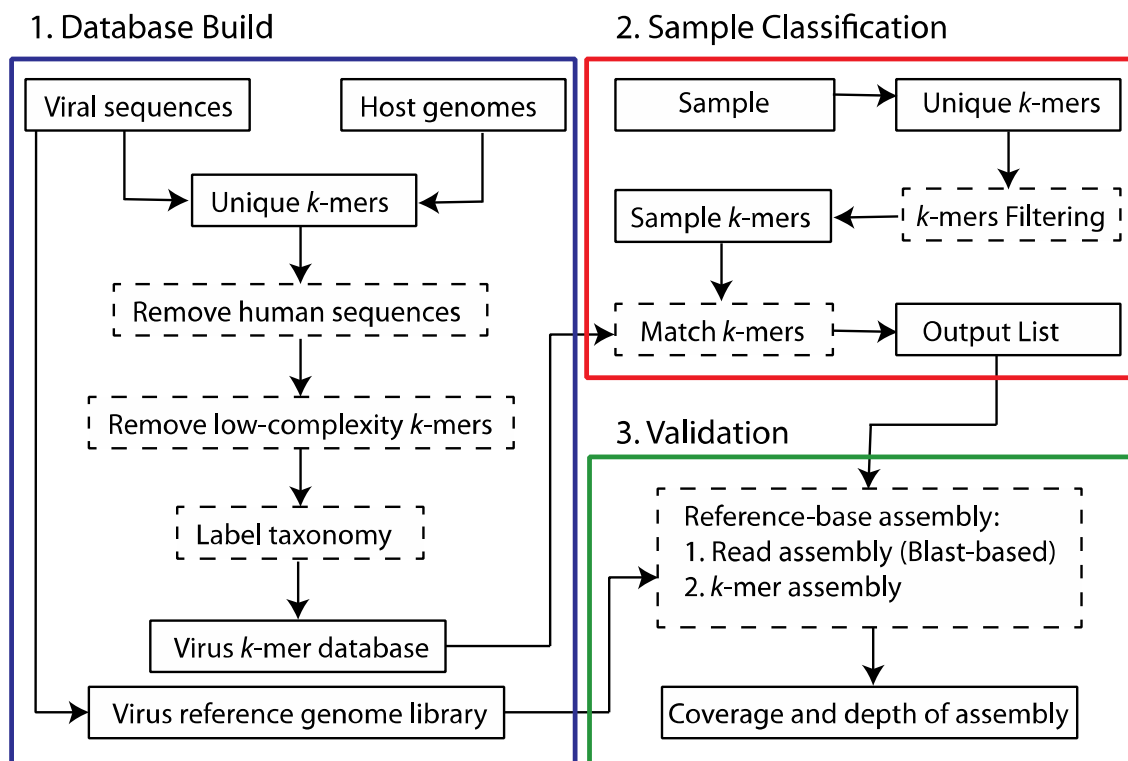
401 **Figure 1.** DisCVR framework. Each coloured box represents a component of the tool. Dashed
402 rectangles indicate processes and solid rectangles show input and output.

403 **Figure 2.** DisCVR GUI. The top screenshot shows the scoring panel with the top three virus hits,
404 and the bottom screenshot shows the full analysis.

405 **Figure 3.** DisCVR validation. Coverage and depth of matched *k*-mers (top) and reads (bottom)
406 to a reference genome.

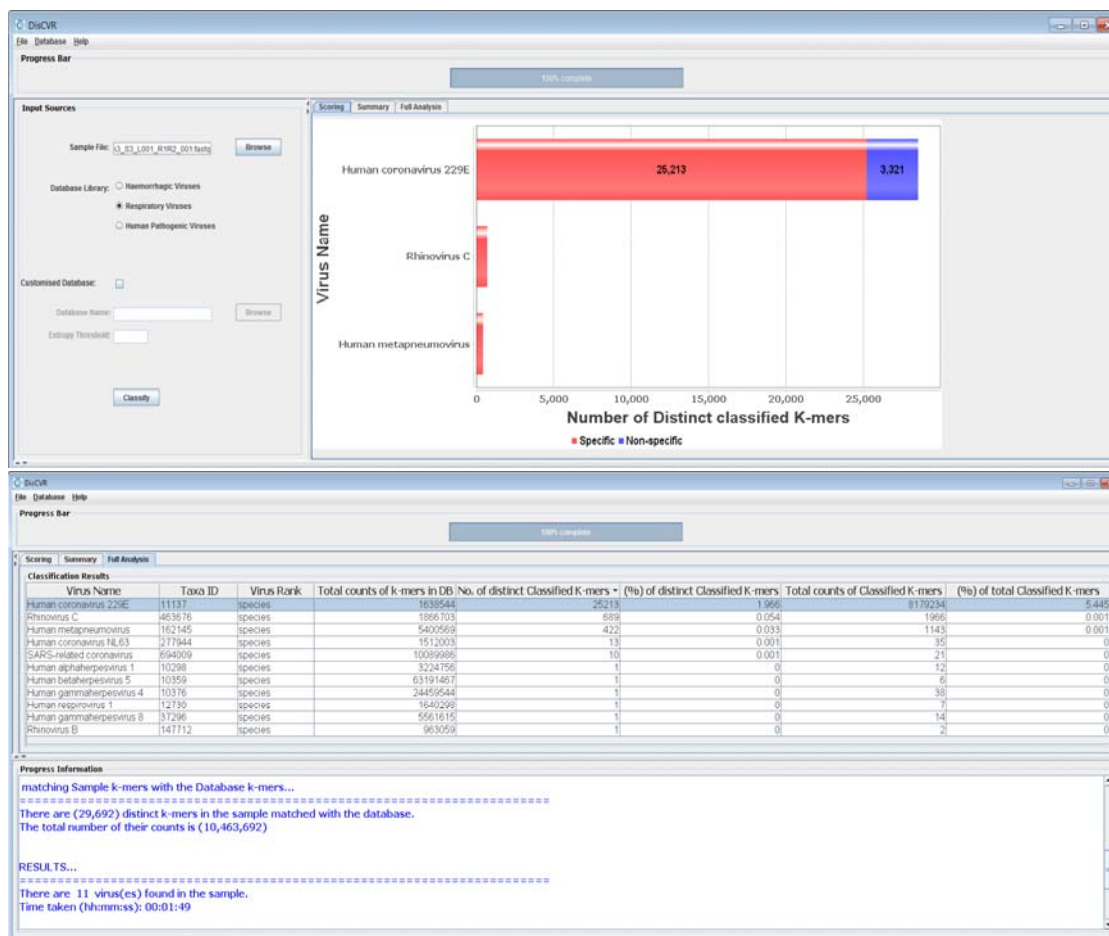
407 **Figure 4.** ROC curve showing the accuracy of DisCVR, CLARK and Kraken. The transparent
408 shaded area shows the confidence interval of the sensitivity for all three methods. The optimal
409 threshold of 850 *k*-mers for DisCVR and 150 reads for CLARK and Kraken are shown, with bars
410 representing the confidence interval of the threshold and the specificity and sensitivity shown in
411 brackets. The curve for KrakenHLL was identical to that for Kraken. The diamond indicates the
412 sensitivity and specificity values, counting the false positives with ≥ 850 *k*-mers and the second
413 hits with ≥ 850 *k*-mers among the true positives.

414 **Figure 1.**



415

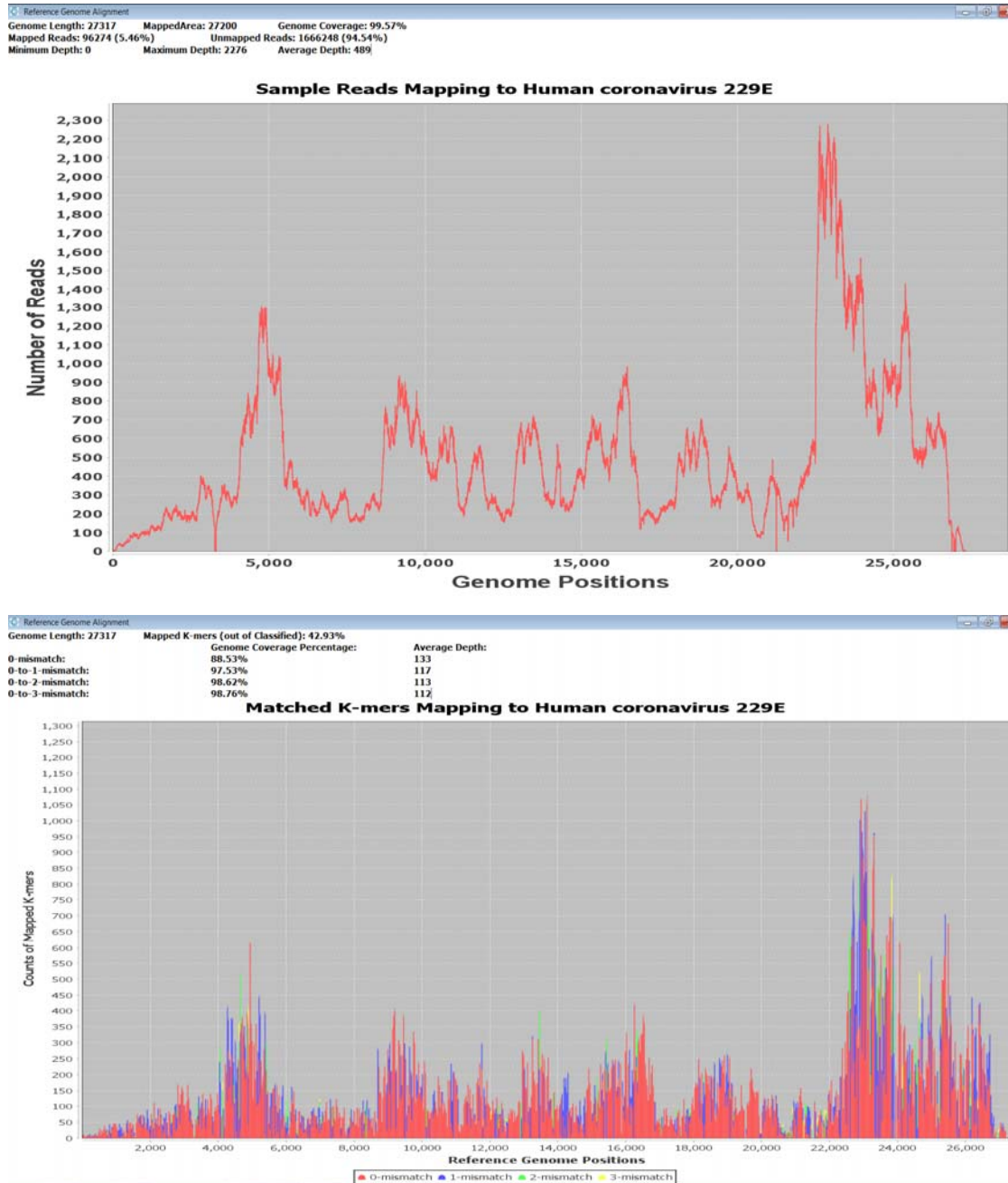
416 **Figure**



417 **2.**

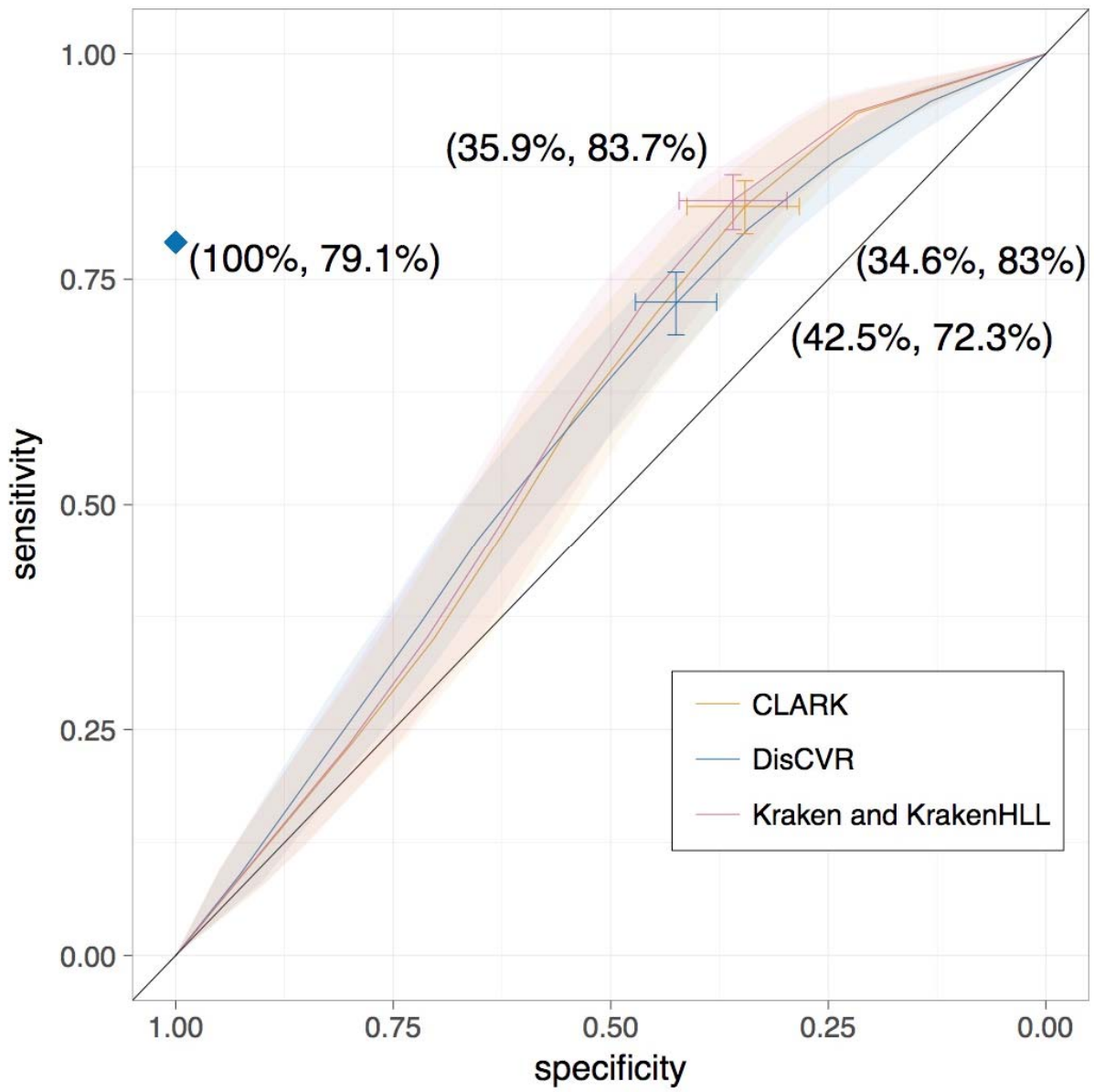
418 **Figure**

419 **3.**



420

421 **Figure 4.**



422