1   # One Codex: A Sensitive and Accurate Data Platform for

2   # Genomic Microbial Identification

3   Samuel S. Minot[*], Niklas Krumm and Nicholas B. Greenfield

4   One Codex (Reference Genomics, Inc.), San Francisco, California, USA

5   *To whom correspondence should be addressed: sam@onecodex.com

6   **1   ABSTRACT**

7   High-throughput sequencing (HTS) is increasingly being used for broad applications of microbial char-

8   acterization, such as microbial ecology, clinical diagnosis, and outbreak epidemiology. However, the

9   analytical task of comparing short sequence reads against the known diversity of microbial life has

10   proved to be computationally challenging. The One Codex data platform was created with the dual goals

11   of analyzing microbial data against the largest possible collection of microbial reference genomes, as

12   well as presenting those results in a format that is consumable by applied end-users. One Codex identi-

13   fies microbial sequences using a "*k*-mer based" taxonomic classification algorithm through a web-based

14   data platform, using a reference database that currently includes approximately 40,000 bacterial, viral,

15   fungal, and protozoan genomes. In order to evaluate whether this classification method and associated

16   database provided quantitatively different performance for microbial identification, we created a large

17   and diverse evaluation dataset containing 50 million reads from 10,639 genomes, as well as sequences

18   from six organisms novel species not be included in the reference databases of any of the tested classifi-

19   ers. Quantitative evaluation of several published microbial detection methods shows that One Codex has

20   the highest degree of sensitivity and specificity (AUC = 0.97, compared to 0.82-0.88 for other methods),

21   both when detecting well-characterized species as well as newly sequenced, "taxonomically novel" or-

22   ganisms.

23

24   **2   INTRODUCTION**

25   As the efficiency, accuracy, and speed of high-throughput genomic sequencing (HTS) has continued to

26   improve, a larger set of microbiologists working in clinical medicine, public health, and industry is

27   adopting this powerful technology. Public health agencies are already beginning to use HTS to track the

28   spread of outbreaks (Neimark 2015), and there is an increasing number of applications for which the ex-

29   quisite precision and accuracy of genome-level identification justifies the (presently) higher marginal

30   per-test cost. However, the broader adoption of HTS by scientists that may lack experience in bioinfor-

31   matic analysis has created a need for bioinformatic solutions that are accessible for non-experts and pro-

32   vide high-quality analytical results (Grad 2014; Caboche 2014).

33   Taxonomic classification is the analytical basis of a wide range of applied microbiology supporting ap-

34   plications in public health, clinical diagnosis, and industrial production. Strain- and species-level taxo-

35   nomic classification allows a user to identify specific pathogens, perform genomic epidemiology, and

36   characterize microbial communities that may be associated with a particular phenotype. A variety of al-

37   gorithms have been developed for such taxonomic classification, including the use of marker gene li-

38   braries (Segata 2012; Liu 2011), local alignment (Naccache 2014; Mitra 2011), and k-mer matching

39   (Ames 2013; Wood 2014; Břinda 2015; Ounit 2015). *K*-mer-based analysis, the method used by Kraken,

40   Seed-Kraken, CLARK, and One Codex, identifies short sequences (typically ranging from 17 - 31 bp)

41   that are unique to specific taxa within a set of input reads. Based on the collection of k-mers that are

42   found in a given read, it can be assigned to a particular taxon. By extension, a sample can be character-

43   ized according to the proportion of reads that are assigned to different taxa. Using this approach, micro-

44   bial samples either from isolates or mixed samples can be characterized to the level needed to perform a

45   large number of tasks needed for public health, clinical diagnosis, and industrial microbiology. As a

46   public resource for academic data analysis, we believe it is valuable to provide the research community

47   with a description of the performance and operation of the One Codex platform, while providing the raw

48   data and analytical details needed to replicate or update such an evaluation as metagenomic methods

49   continue to improve. In this paper we describe the functioning of One Codex and a rigorous functional

50   evaluation of the state-of-the-art metagenomic classification methods, including the effect of database

51   size on classification accuracy.

52

53   **3   MATERIALS & METHODS**

3

54

3.1   Taxonomic Classification Algorithm

56   One Codex classifies individual sequence reads according to the set of k-mers in that read that are

57   unique to specific taxonomic groups. This analytical approach has been described extensively (Ames

58   2014; Wood 2014) and is implemented by One Codex using a default value of $k=31$. Briefly, each read

59   is broken into the complete set of overlapping sequences of length 31bp that comprise it. These k-mers

60   are compared against an exhaustive database that contains every k-mer and the taxonomic grouping to

61   which it is unique (e.g., a specific clade of bacteria, archaea, or viruses). A compressed data structure is

62   used to index and rapidly search k-mer databases generated from approximately 40,000 microbial ge-

63   nomes. Each read can then be summarized as a "k-mer hit chain" that describes the complete set of taxo-

64   nomically-informative k-mers found in that read, as well as their positions. Individual reads are then as-

65   signed on the basis of the highest weighted taxonomic root-to-leaf path amongst these k-mer hits. For

66   example, if a read has k-mers unique to *Enterobacteriaceae*, *Escherichia*, and *Escherichia coli*, it would

67   be given the label *E. coli*. However, if it had k-mers unique to *Enterobacteriaceae*, *Escherichia*, and

68   *Klebsiella*, it would be given the label *Enterobacteriaceae* – the most specific taxon that encompasses

69   all detected k-mers, as *Klebsiella* and *Escherichia* are separate genera of *Enterobacteriaceae*. Finally,

70   the distribution of reads from a single sample across different organisms and taxa is used to construct a

71   comprehensive report that displays the organisms present in a sample.

72

3.2   Classification Accuracy

74   The goal of this evaluation effort was to assess the ability of a suite of bioinformatic methods to assign

75   nucleotide sequences to the most accurate taxonomic group.

76

77   **3.2.1. One Codex Classification**

78   Data were processed on the One Codex platform according to the instructions outlined in Section 2.3 –

79   One Codex User Interface. One Codex uses two reference databases, the full One Codex database of ap-

80   proximately 40,000 bacteria, viruses, fungi, archaeal, and protists, and a smaller database containing the

81   over 8,000 microbial genomes contained in the NCBI RefSeq database. Both the One Codex full data-

82   base (referred to here as "One Codex") and the One Codex RefSeq Database represent sequences availa-

83   ble on July 8, 2015.

84

85   **3.2.2. Additional Classification Algorithms**

86   The following metagenomic classification algorithms were downloaded, compiled, and installed accord-

87   ing to the provided instructions in an Ubuntu environment on standard AWS EC2 instances (r3.8xlarge).

88   In each case the indicated dependencies were installed as described and default run settings were used,

89   except in the case of Clark in which the "RAM-light" flag was used in order not exceed system memory

90   capacity.

91   •   Metaphlan (2.1.0) - https://bitbucket.org/biobakery/metaphlan2

92   •   GOTTCHA (1.0b) - https://github.com/poeli/GOTTCHA (database

93   GOTTCHA_BACTERIA_c3514_k24_u2)

94   •   Kraken (v0.10.5-beta) - http://ccb.jhu.edu/software/kraken/

95   •   Seed-Kraken (seedmod128b_from_0.10.6) - http://seed-kraken.readthedocs.org/

96   •   Clark (v1.1.3) – http://clark.cs.ucr.edu/

97

98   ### 3.2.3.   Additional Classification Databases

99   Metaphlan is distributed with a complete standalone database. Kraken is distributed with a reduced ref-

100   erence database ("Minikraken" - Dec. 8, 2014). In addition, we constructed the full Kraken database on

101   July 5, 2015 according to the instructions provided at

102   http://ccb.jhu.edu/software/kraken/MANUAL.html. Given the contemporaneous snapshot of NCBI, the

103   genomic content of the full Kraken database is roughly equivalent to that of the One Codex RefSeq Da-

104   tabase, albeit with some differences in the exact repositories used. The Seed-Kraken database was con-

105   structed using the same complement of reference genomes as the Kraken database. The Clark bacterial

106   reference database was constructed on Aug. 21, 2015 using the provided default instructions.

107

108   ### 3.2.4.   Test Datasets

109   Two complementary approaches were employed to test the accuracy of these microbial detection meth-

110   ods. In the first, 500 sets of simulated reads (100,000 reads each) were generated from complete micro-

111   bial genomes at a wide range of abundance levels in order to simulate a variety of biological assemblag-

112   es (Mavromatis 2007). Each read set contained reads from 214 random genomes with roughly 100,000

113   reads simulated each, and additional reads from 10,425 genomes simulated at much lower abundance (1

114   – 30,000 reads each) (Figure 1). In total, 50 million reads were simulated from 10,639 genomes to pro-

115   vide a robust resource for the evaluation of metagenomic analysis methods. In every case, the genome

116   was selected randomly from the set of complete genomes available in public sequence repositories, re-

117   gardless of whether they were included in the One Codex database. Approximately 78% of the simulated

118   genomes were also indexed in the One Codex Database, while 8.8% were indexed in the One Codex

119   RefSeq Database. Importantly, the taxonomic ID for each read was encoded in the FASTQ header, al-

120   lowing the direct comparison of the known source of each sequence against the taxonomic prediction

121   made by each method.

Low-abundance simulated genomes
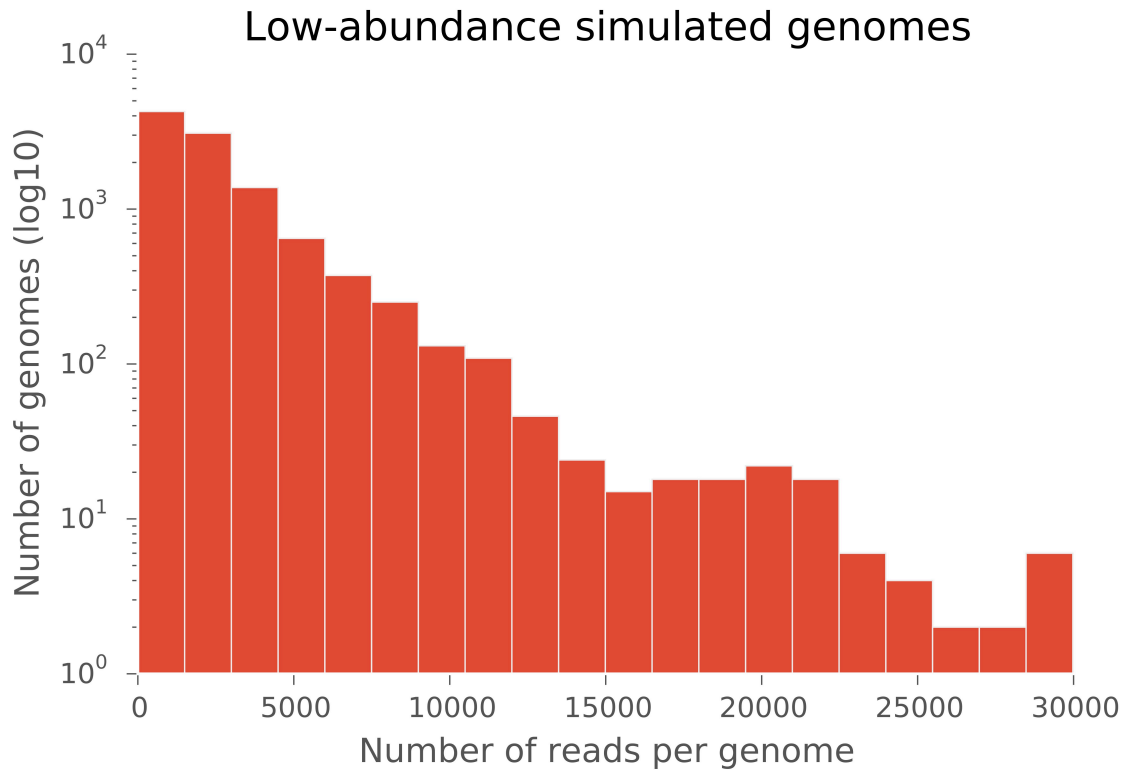
122

123   **Figure 1**. A frequency histogram of the number of reads simulated for the set of low-abundance ge-

124   nomes (10,425 genomes, $1 - 30,000$ reads each). The horizontal axis shows the number of reads simu-

125   lated per genome, and the vertical axis indicates the number of genomes simulated at that depth ($\log_{10}$).

126   An additional set of 214 genomes were used to simulate at least 100,000 reads each.

127

128   In the second testing approach, a set of six organisms identified in the public repositories that were se-

129   quenced recently enough as to not be included in any of the reference databases. These "taxonomically

130   novel" genomes did not have any other members of their species present in the reference database. Three

131   of the six "taxonomically novel" test datasets were simulated from complete genome assemblies at 5X,

132   two were raw Illumina reads, and one was an unassembled PacBio dataset (Table 1). While these da-

133   tasets may contain contaminating or misidentified organisms, each analytical method will be challenged

134   equally by those potentially confounding factors.

135

136   For both the single-isolate samples and the 500 sets of simulated reads (100,000 reads each), simulated

137   150bp single-ended reads were generated using the ART next-generation sequencing read simulator

138   (v3.19.15) (Huang 2012) using the Illumina quality profile. For the single-isolate "taxonomically novel"

139   test datasets, reads were generated at 5-fold coverage depth.

140

NCBI Accession	Organism	Source	Type	Number of Simulated Reads
GCA_001045455	Chryseobacterium sp. FH2	Assembly	Illumina	132,905
GCA_001050135	Cyclobacterium amurskyense KCTC-12363	Assembly	Illumina	205,290
SRR2106399	Leptolyngbya sp. Y-WT-2000	Unassembled reads	Illumina	11,060,814
GCA_001258055	Nautella italica CECT7645	Assembly	Illumina	134,905
SRR2106282	Thermincola ferriacetica Z-0001	Unassembled reads	Illumina	2,053,515
SRR2080278	Wenzhouxiangella marina KCTC42284	Unassembled reads	PacBio	163,476

141

142   **Table 1**. Datasets used to assess the performance of each method in identifying organisms not contained

143   in the reference database.

144

145   **3.2.5. Statistical Summary**

146   Kraken, Seed-Kraken, Clark, and One Codex provide taxonomic assignments for every read in a dataset,

147   while Metaphlan and GOTTCHA provide an overall summary of dataset composition. Kraken, Seed-

148   Kraken, Clark, and One Codex were evaluated with the complete set of 50 million simulated reads,

149   while all methods were evaluated on the six single-organism "taxonomically novel" datasets. Each

150   method was executed on an equivalent AWS EC2 instance (r3.8xlarge) with 12 processors available for

151   parallelized steps.

152

153   For the set of 50 million simulated reads, the accuracy of classification by One Codex, Seed-Kraken and

154   Kraken was evaluated on a read-by-read basis. One Codex was run with both the One Codex Database

155   (~40,000 genomes) and the One Codex RefSeq Database (~8,000 genomes). Kraken was run with both

156   the full database and the "Minikraken" database. Those methods assign an NCBI taxonomic identifier

157   ('taxid') to each read. Given the known source of each read, the accuracy of the classification can be as-

158   sessed across all levels of the taxonomy. For example, a read simulated from *E. coli* O157:H7 str. Sakai

159   may be assigned to *Escherichia fergusonii*, in which case the species-level assignment is incorrect, while

160   the genus-level assignment is correct (as is the family-, order-, class-, and phylum-). Similarly, a read

161   simulated from *E. coli* O157:H7 str. Sakai may be assigned to *Enterobacteriaceae*, in which case the

162   family-level assignment is correct and there is no assignment at the rank of genus, species, or strain.

163

164   Accuracy metrics were calculated as follows: Let A be the number of reads assigned correctly at a given

165   taxonomic rank, B be the number of reads with any assignment at the given rank, C be the number of

166   reads classified incorrectly at a less-specific rank, and D be the total number of reads. Sensitivity is de-

167   fined as A / D, or the proportion of all reads assigned correctly at the given rank. Specificity is defined

168   as A / (B + C), following Wood et al. (2015). For a set of reads simulated from *E. coli,* the classification

169   of a read as *E. coli* would increase both species-level sensitivity and species-level specificity, classifica-

170   tion as *Escherichia* would not increase species-level sensitivity, but it would increase species-level spec-

171 ificity, and classification as *E. fergusonii* would decrease both species-level sensitivity and species-level

172 specificity.

173

174 Metaphlan and GOTTCHA do not assign taxa to individual reads, but rather predict the proportion of the

175 dataset composed of different taxa. Therefore the specificity presented for those methods is the propor-

176 tion assigned to the correct taxon at a given rank and no sensitivity metrics are reported.

177

178 **3.3   One Codex User Interface**

179 Samples are uploaded in FASTA or FASTQ format to the One Codex platform through a graphical up-

180 load tool with both drag-and-drop and folder navigation options. A command-line tool and API are also

181 available for large-volume data upload (https://docs.onecodex.com). Once uploaded, reads are taxonom-

182 ically classified and the interactive report is populated and linked to the user's account (Supplemental

183 Figures 1 and 2). The One Codex platform can be accessed at https://www.onecodex.com, and can be

184 used freely to analyze public data.

185

186 **3.4   Availability**

187 Simulated datasets are available in a compressed FASTQ file containing 50M simulated reads at

188 www.onecodex.com/data/papers/minot-krumm-greenfield-2015/simulated.reads.fastq.gz. The true taxo-

189 nomic origin of each read is encoded in the FASTQ header as an NCBI taxid, allowing other researchers

190 to replicate this analytical framework. One Codex is freely available for public use by academic re-

191 searchers at https://www.onecodex.com. Supplemental Figures 1 and 2 show example screenshots of the

192 One Codex platform.

193

194 **4   RESULTS**

195

4.1   Read-level Accuracy

197   We first summarized the accuracy of each tool on a per-read basis. One Codex showed the highest de-

198   gree of sensitivity and specificity at each rank and the performance of the other methods varied with the

199   database and assignment method used. It is notable that although the content of the Kraken and Seed-

200   Kraken databases was identical, Seed-Kraken was more sensitive and less specific than Kraken at all

201   taxonomic levels (Table 2 and Figure 2). While the reduced Minikraken database resulted in lower sen-

202   sitivity and higher specificity than the full Kraken database, the reduced One Codex RefSeq Database

203   was less accurate using both metrics. As noted above, ~78% of the simulated genomes were indexed in

204   the One Codex Database, while 8.8% were indexed in the One Codex RefSeq Database. Accuracy met-

205   rics were also calculated for One Codex and One Codex RefSeq using only the subset of reads simulated

206   from genomes not indexed in those databases. Species-level sensitivity and specificity for One Codex

207   was 0.532 and 0.875, respectively, while One Codex RefSeq was 0.511 and 0.835. Similar comparisons

208   could not be made for other methods without better characterization of the genome accessions used to

209   create those reference indices.

210

Sensitivity

Rank	Kraken	Minikraken	Seed-Kraken	Clark	One Codex	One Codex - RefSeq
superking-dom	0.743	0.714	0.758	0.682	**0.951**	0.748
phylum	0.742	0.713	0.755	0.681	**0.943**	0.745
class	0.740	0.711	0.752	0.679	**0.906**	0.743
order	0.737	0.709	0.748	0.678	**0.903**	0.740
family	0.734	0.706	0.743	0.675	**0.897**	0.737
genus	0.694	0.662	0.711	0.668	**0.852**	0.704
species	0.546	0.516	0.574	0.605	**0.621**	0.563
strain	0.030	0.027	0.031	0.000	**0.118**	0.080

Specificity

Rank	Kraken	Minikraken	Seed-Kraken	Clark	One Codex	One Codex - RefSeq
superking-dom	**0.999**	**0.999**	0.996	0.996	**0.999**	0.994
phylum	0.997	**0.998**	0.994	0.994	**0.998**	0.992
class	0.995	**0.997**	0.991	0.992	**0.997**	0.990
order	0.993	**0.996**	0.987	0.989	**0.996**	0.988
family	0.989	0.994	0.982	0.986	**0.996**	0.984
genus	0.980	0.991	0.970	0.975	**0.995**	0.974
species	0.893	0.936	0.869	0.883	**0.979**	0.864
strain	0.157	0.203	0.114	0.000	**0.824**	0.181

211

212 **Table 2**. Summary of accuracy for six methods identifying the taxonomic origin of 50 million short se-

213 quence reads simulated from 10,639 microbial genomes. The maximum value at each rank is bolded.
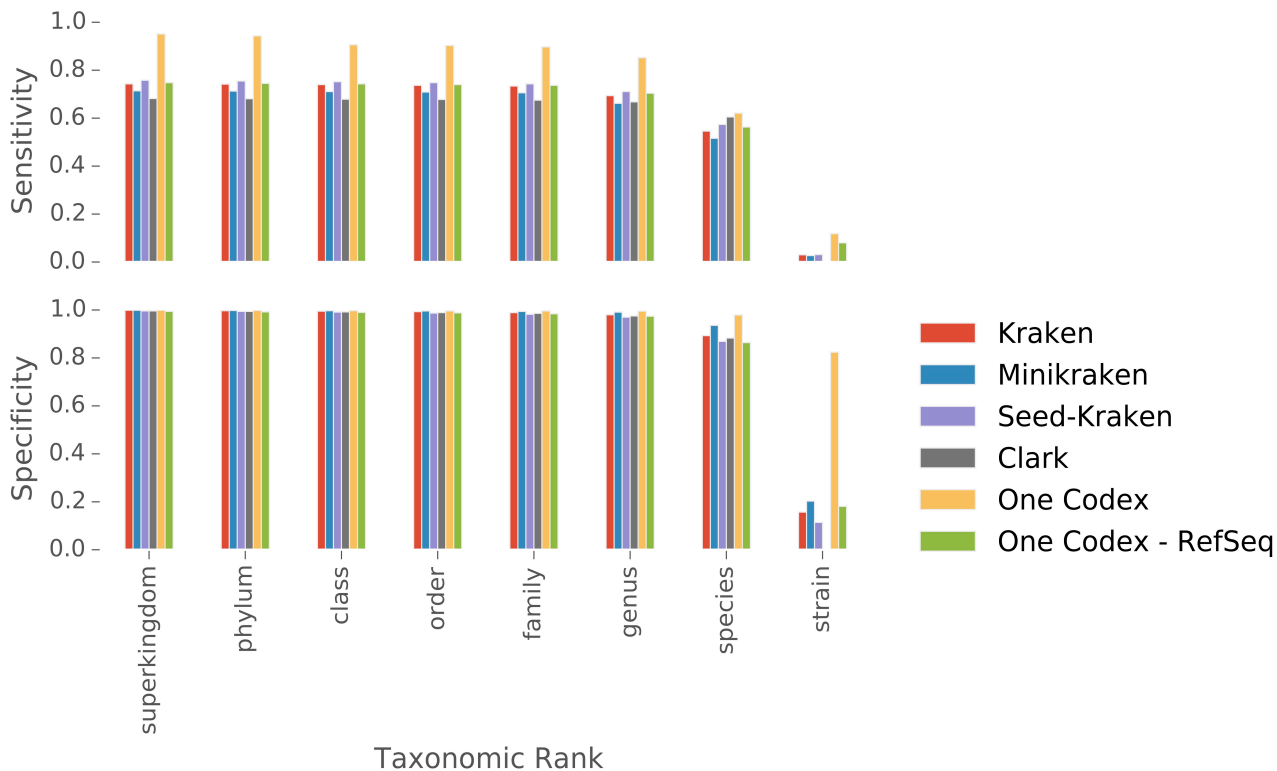

214

215   **Figure 2**. Summary of accuracy for six methods identifying the taxonomic origin of 50 million short se-

216   quence reads simulated from 10,639 microbial genomes.

217

218   **4.2   Species-level presence/absence accuracy**

219   We characterized the species-level accuracy of each classifier using a receiver operating characteristic

220   (ROC) curve (Fig. 3). The ROC curve displays the true positive rate (TPR) and false positive rate (FPR)

221   of species presence/absence across a range of read thresholds. Each dataset can be summarized as a set

222   of species, each detected with a certain number of reads, and each marked as truly present, or truly ab-

223   sent. For any given number of reads, the species detected with at least that number of reads is marked as

224   present, and any species with fewer than that number of reads is marked as absent. The TPR is calculat-

225   ed for a given read threshold as the number of true positive detections divided by the total number of

226   total positives, and the FPR is calculated as the number of true negatives divided by the number of total

227   negatives. The ROC curve for each method is shown in Figure 3.

228
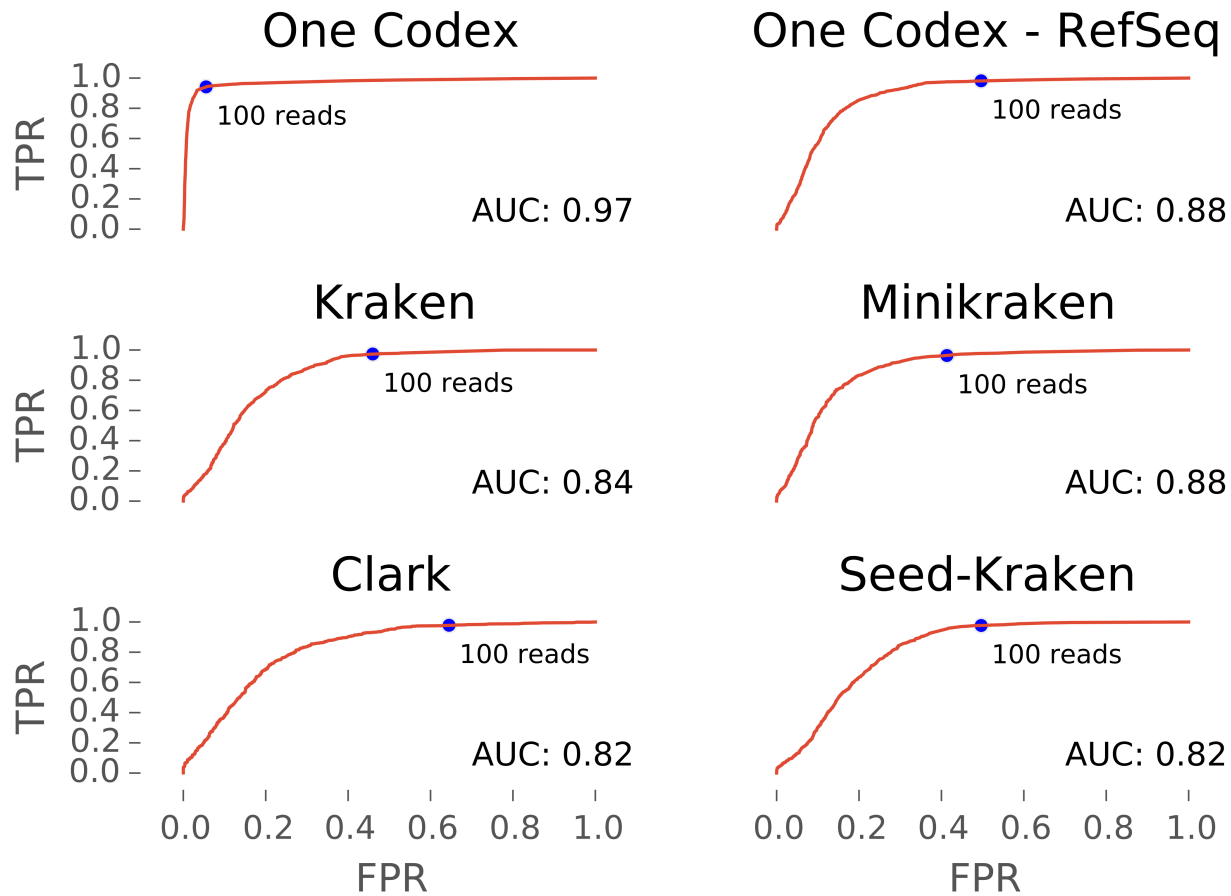
Figure 3. ROC curve displaying the performance of six taxonomic classification methods as species-level binary classifiers. Horizontal axis shows the False Positive Rate, and the vertical axis shows the True Positive Rate. Area Under Curve (AUC) is inset. The FPR and TPR are shown for a read cutoff value of 100 detected reads (150bp), which corresponds to roughly 0.003X coverage of a typical bacterial genome.

234

4.3 "Taxonomically Novel" Accuracy

Each of the "taxonomically novel" test datasets was selected because that species was not present in the reference database for any method. Due to the relative taxonomic novelty of these organisms, the closest

238   match found by any method for these datasets was either at the genus-, family-, or order-level. For ex-

239   ample, the most taxonomically similar reference organisms to *Wenzhouxiangella marina* KCTC 42284

240   (SRR2080278) share the order *Chromatiales*, as the family *Wenzhouxiangellaceae* was only proposed

241   very recently (Wang 2015). Sensitivity and specificity metrics are shown in Table 3 alongside the rank

242   at which the closest correct match was found by any method. GOTTCHA did not report any taxa above

243   its threshold of detection for the three datasets showing '0' in Table 3, and we provided the authors of

244   that method with those datasets in order to confirm those results. Using the GOTTCHA 'v20150825'

245   database they reported that dataset GCA_001045455 was assigned correctly at the family level and

246   above, dataset SRR2106399 was assigned correctly at the order level, and dataset SRR2080278 did not

247   have any assignments above the threshold of reporting (personal communication). Across all six da-

248   tasets, One Codex displayed the highest sensitivity (0.391) while One Codex - RefSeq had the highest

249   specificity (0.696).

Dataset	GCA_001045455	GCA_001050135	SRR2106399	GCA_001258055	SRR2106282	SRR2080278	
Assignment	Chryseobacterium	Cyclobacterium	Leptolyngbya	Rhodobacteraceae	Thermincola	Chromatiales	
Rank	Genus	Genus	Genus	Family	Genus	Order	
Sensitivity							**Mean**
Kraken	0	0.149	0	0.203	0.831	0.005	0.198
Minikraken	0	0.096	0	0.093	0.787	0.001	0.163
Seed-Kraken	0	0.211	0.001	0.352	0.836	0.004	0.234
Clark	0	0.065	0.001	0.072	0.595	0.002	0.122
One Codex	0.313	0.225	0.01	0.965	0.827	0.005	0.391
One Codex - RefSeq	0.191	0.142	0.002	0.244	0.83	0.005	0.236
Metaphlan	-	-	-	-	-	-	-
GOTTCHA	-	-	-	-	-	-	-
Specificity							**Mean**
Kraken	0	0.961	0.013	0.920	0.995	0.102	0.499
Minikraken	0	0.985	0.04	0.943	0.998	0.266	0.539
Seed-Kraken	0	0.932	0.207	0.947	0.993	0.156	0.539
Clark	0	0.939	0.002	0.823	0.991	0.011	0.461
One Codex	0.920	0.927	0.109	0.999	0.988	0.082	0.671
One Codex - RefSeq	0.928	0.919	0.186	0.942	0.995	0.208	0.696
Metaphlan	0.935	1.000	0	1.000	0	0	0.489
GOTTCHA	0	1.000	0	1.000	0.971	0	0.495

251

252   **Table 3**. Accuracy of prediction for six organisms not found in any of the reference databases used by

253   these methods. Note that Metaphlan and GOTTCHA results are presented as specificity metrics, as the

254   abundance metrics reported by those methods are relative to the total classified composition of each

255   sample (rather than the number of input reads).

256

257   **4.4   Analysis Time**

258   The time required for completing analysis of the 50M simulated reads for each method is presented in

259   Table 4 (GOTTCHA and Metaphlan are not shown because they were not run on the read-level evalua-

260   tion datasets). Although the complete set of 50M simulated reads were run in parallel batches of 10M

261 reads, the time presented here is the cumulative processing time, rather than the shorter start-to-finish

262 period of parallelized execution across multiple computational nodes. All methods were run with 12

263 processors on equivalent computational resources. Of these read-classification methods, Minikraken and

264 Clark were the most rapid, and Seed-Kraken was the slowest.

265

	Analysis Time (min:sec)
Clark	04:50
Minikraken	05:15
Kraken	06:26
One Codex	22:13
One Codex - RefSeq	21:49
Seed-Kraken	45:58

266

267 **Table 4**. Computational time required for each method to classify the 50M simulated reads.

268

269 **5   DISCUSSION**

270

271 The widespread adoption of high-throughput sequencing for the detailed characterization of mixed mi-

272 crobial samples presents an immense opportunity and challenge to the field of microbial genomics. Alt-

273 hough genomic sequences can be used pinpoint the organisms in a sample down to the level of a single

274 strain, accurate detection of each strain completely depends on the ability of a computational method to

275 search for genomic sequences across the extent of known life. Not only does the volume of microbial

276 reference genomes exceed that of the human genome by many-fold (181 billion bases of prokaryotic ge-

277 nome sequence can be found in NCBI as of Aug. 25, 2015), but the sequences found in wild-caught mi-

278 crobes may differ significantly from those of their domesticated relatives (Rinke, et al. 2013). In the face

279 of these serious computational challenges, a large panel of computational methods have been proposed

280 recently to perform the task of microbial detection (Oulas, 2015). However, it can be prohibitively diffi-

17

281   cult for a microbial researcher to rigorously evaluate all of the possible options in order to select the

282   most appropriate method. To address that challenge, we have provided a comprehensive analysis of the

283   performance of a wide range of the most widely adopted analytical methods. Moreover, we have made

284   the test data and analytical framework available for others to evaluate future methods against a common

285   reference. We believe that the large volume (over 50M simulated reads), phylogenetically complexity

286   (10,639 randomly selected source organisms), and analytical portability (NCBI taxonomic identifiers

287   recorded within read headers), makes this dataset a valuable resource for the research community.

288

289   One Codex provides the highest degree of accuracy, both sensitivity and specificity, across all taxonom-

290   ic ranks, with 62.1% per-read sensitivity and 98.7% per-read specificity at the species-level (Table 2).

291   The absolute performance of any detection method as a binary classifier was summarized by ROC anal-

292   ysis, showing that One Codex had the best performance (AUC: 0.97), with other methods performing

293   roughly equally (AUC: 0.82 – 0.88). For purposes of illustration, the accuracy of each method was

294   shown at an absolute abundance cutoff of 100 reads (Fig. 3). At that threshold for calling a species as

295   present in a sample, One Codex showed a much lower FPR than other methods, suggesting that the larg-

296   er set of reference organisms in the One Codex database serves to significantly reduce the number of

297   species-level false positive detections with that method.

298

299   The array of methods evaluated here allows for an intriguing comparison of the effect of reference data-

300   base and classification algorithm on overall performance. Kraken and Seed-Kraken use differing as-

301   signment algorithms and a common reference database, with Seed-Kraken providing higher sensitivity

302   and Kraken providing higher specificity. This finding replicates similar precision/sensitivity trade-off

303   that was observed previously for Seed-Kraken at the species-level (Břinda, 2015). While the perfor-

304   mance of both Kraken and One Codex is presented with two alternate databases, the smaller Minikraken

305 database was constructed by selecting a smaller number of kmers per organism, and the smaller One

306 Codex RefSeq database was constructed by selecting a restricted number of reference organisms, so the

307 resulting differences in sensitivity and specificity are not directly comparable. As the genomes encoun-

308 tered in real-world metagenomic samples rarely exactly match those found in reference databases, it is

309 important to quantify the accuracy of detection for 'out-of-reference' genomes. Considering only the

310 reads simulated from reads not found in those reference databases, the sensitivity and specificity of de-

311 tection for One Codex was 0.532 and 0.875, respectively, while One Codex RefSeq was 0.511 and

312 0.835, reflecting only a modest decline in performance for that large group of 'out-of-reference' ge-

313 nomes. Further research into the effect of reference database composition on predictive accuracy could

314 conceivably enable the creation of classification methods with smaller computational footprints and im-

315 proved performance.

316

317 The largest difference in performance between classification methods can be seen in the "taxonomically

318 novel" test datasets. Each method detects a different subset of organisms, indicating that the composition

319 of each reference database highly determines the ability of a method to detect a given organism. Overall,

320 One Codex had the highest average sensitivity (39.1%), which was far higher than the next-most sensi-

321 tive methods One Codex RefSeq (23.6%) and Seed-Kraken (23.4%). This set of six datasets is a useful

322 demonstration of the fundamental challenge of accurately classifying sequences from organisms not

323 found in any reference database. Even when the closest possible rank is at the genus or above, each

324 method varies widely in its ability to assign sequences correctly to that rank. The highly sensitive per-

325 formance of One Codex against these samples suggests that a large and comprehensive reference data-

326 base not only enables more accurate detection of well-characterized taxa, but also enables more accurate

327 detection of taxonomically-novel and phylogenetically divergent organisms.

328

19

329   By evaluating a wide range of taxonomic classification algorithms against a large and complex set of

330   10,639 simulated genomes, as well as a set of six recently sequenced and phylogenetically-distinct or-

331   ganisms, we have generated important insight into the ability of microbial researchers to accurately

332   characterize unknown metagenomic samples. Most notably, because Kraken, Minikraken, One Codex,

333   and One Codex RefSeq all classify reads using taxonomically-unique 31mers, the differing performance

334   of these methods is undoubtedly due to the much different composition of those databases, with larger

335   reference databases leading to greater analytical accuracy. These results show the value of continually

336   expanding reference database collections in order to more accurately classify the vast pool of unknown,

337   unsequenced microbial 'dark' matter (Rinke 2013), as well as the specific strains of well-known patho-

338   gens that cause human disease.

339

340   **6 ACKNOWLEDGEMENTS**

341   *Conflict of Interest:* Authors are employed by Reference Genomics, Inc., which develops the One Codex
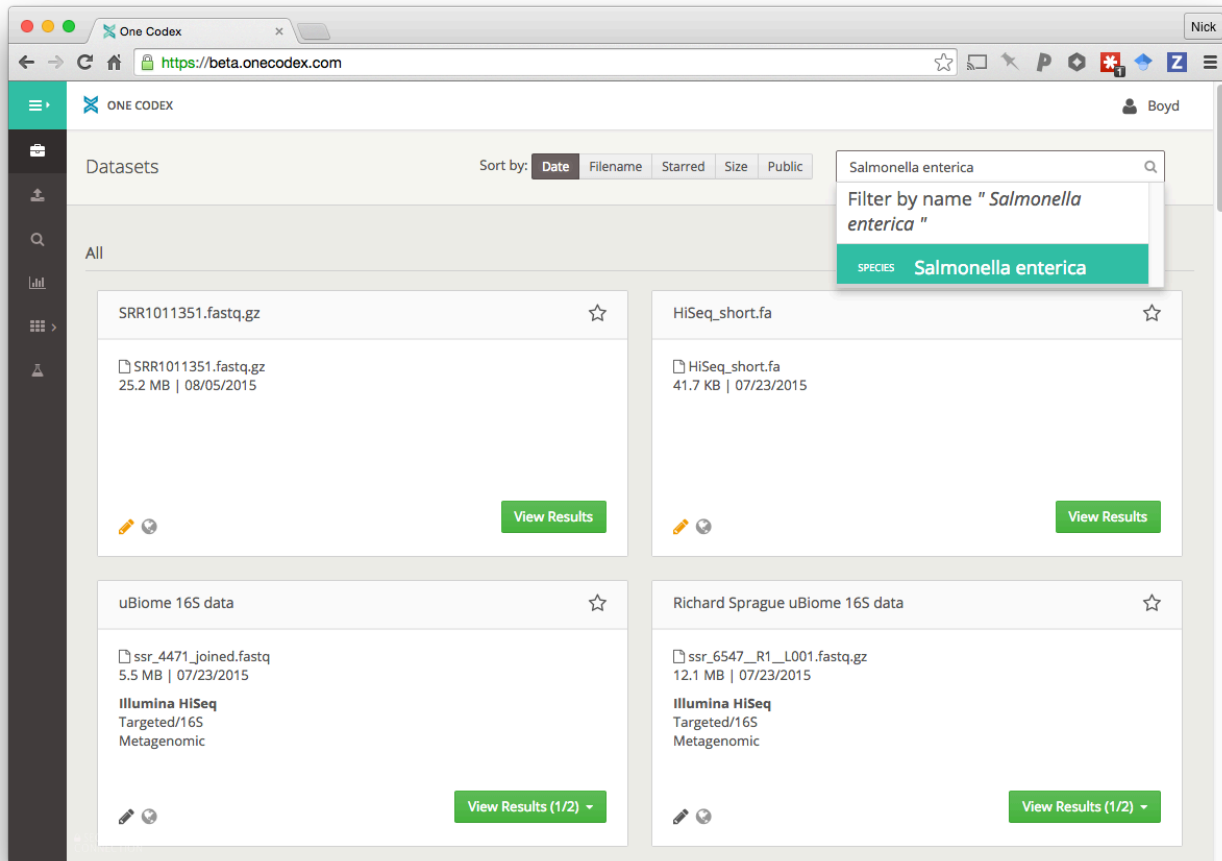
342   platform.

343

344   **7 REFERENCES**

345   Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxono-
346   my classification using a reference genome database. Bioinformatics. 2013; 29(18):2253-60.
347   Břinda K, Sykulski M, Kucherov G. Spaced seeds improve k-mer-based metagenomic classification.
348   Bioinformatics. 2015 Jul 25; pii: btv419
349   Caboche S, Audebert C, Hot D. High-Throughput Sequencing, a Versatile Weapon to Support Genome-
350   Based Diagnosis in Infectious Diseases: Applications to Clinical Bacteriology. Pathogens. 2014 Apr
351   2;3(2):258-79. doi: 10.3390/pathogens3020258.
352   Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for pub-
353   lic health. Genome Biol. 2014 Nov 18;15(11):538. doi: 10.1186/s13059-014-0538-4.
354   Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformat-
355   ics. 2012 Feb 15;28(4):593-4. doi: 10.1093/bioinformatics/btr708.
356   Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles
357   from metagenomic shotgun sequences. BMC Genomics. 2011;12 Suppl 2:S4. doi: 10.1186/1471-2164-
358   12-S2-S4.

359   Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, et al. Use of simulated data
360   sets to evaluate the fidelity of metagenomic processing methods. Nat Methods. 2007 Jun;4(6):495-500.
361   Mitra S, Rupek P, Richter DC, Urich T, Gilbert JA, Meyer F, et al. Functional analysis of metagenomes
362   and metatranscriptomes using SEED and KEGG. BMC Bioinformatics. 2011 Feb 15;12 Suppl 1:S21.
363   doi: 10.1186/1471-2105-12-S1-S21.
364   Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible
365   bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clini-
366   cal samples. Genome Res. 2014 Jul;24(7):1180-92. doi: 10.1101/gr.171934.113.
367   Neimark, J. Quick DNA Scans Could Ensure Food Is Safe to Eat. Scientific American. 2015 Feb.
368   Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al. Meta-
369   genomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversi-
370   ty Studies. Bioinform Biol Insights. 2015 May 5;9:75-88. doi: 10.4137/BBI.S12462.
371   Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylog-
372   eny and coding potential of microbial dark matter. Nature. 2013 Jul 25;499(7459):431-7. doi:
373   10.1038/nature12352.
374   Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial
375   community profiling using clade-specific marker genes. Nat Methods. 2012 Jun 10;9(8):811-4. doi:
376   10.1038/nmeth.2066.
377   Wang G, Tang M, Li T, Dai S, Wu H, Chen C, et al. Wenzhouxiangella marina gen. nov, sp. nov, a ma-
378   rine bacterium from the culture broth of Picochlorum sp. 122, and proposal of Wenzhouxiangellaceae
379   fam. nov. in the order Chromatiales. Antonie Van Leeuwenhoek. 2015 Jun;107(6):1625-32. doi:
380   10.1007/s10482-015-0458-7.
381   Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments.
382   Genome Biol. 2014 Mar 3;15(3):R46. doi: 10.1186/gb-2014-15-3-r46.

383 **8   SUPPLEMENTAL FIGURES**

384

385   Supplemental Figure 1. Example of dataset browser page on the One Codex platform. Datasets are orga-

386   nized by metadata (e.g. name, date, size, environment, platform, etc.) and user-defined tags. Figure dis-

387   plays an example of the user searching for datasets containing *Salmonella enterica* above a specified

388   abundance threshold.
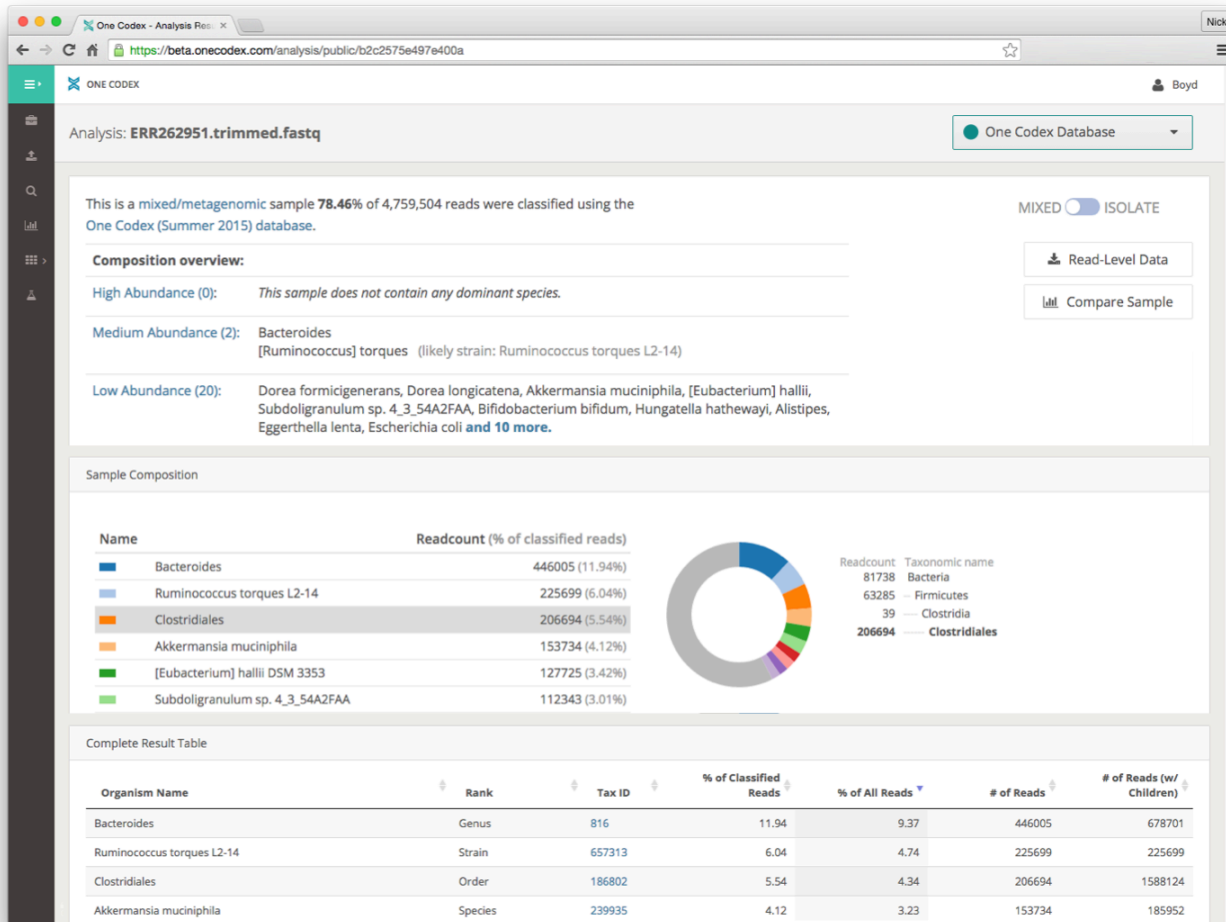
389

390 Supplemental Figure 2. Example of metagenomic analysis display for a single sample. Users have the

391 option of downloading read-level assignments or dataset summaries, comparing the abundance profile

392 against that of other samples, and navigating to additional analyses for each dataset.