available under [aCC-BY 4.0 International license.](http://creativecommons.org/licenses/by/4.0/) (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made bioRxiv preprint doi: [https://doi.org/10.1101/2023.12.19.572299;](https://doi.org/10.1101/2023.12.19.572299) this version posted December 19, 2023. The copyright holder for this preprint

227 Supplementary Methods

229 Datasets

- 230
-
-
-
- コンドリー Easting Article がいしょう イモ Fast こんしょう ミーノ しんしゅう こうしょう しょうきん くんじょう 231 (https://github.com/RyanCook94/inphared; September 2023)⁶. Stop (11
231 (https://github.com/RyanCook94/inphared; September 2023)⁶. Stop (12
232 INPHARED genomes was predicted using Prodigal-gv 2.11.0
233 (https:// (https://github.com/RyanCook94/inphared; September 2023)⁶. Stop codon reassignment of

2023 . INPHARED genomes was ped cled using Prodigiol-gv 0.2.11.0

2021 . The Unified Human Gut Virome Catalog (UHGV) was filtered fo 232 Intertigal and the method of the sympaths and the sympaths and the sympaths completed by the predicted to use translation table 4 or 15 were calling⁸. Those predicted to use translation table 4 or 15 were analysis.
 233 calling⁸. Those predicted to use translation table 4 or 15 were retained for downstream
233 calling⁸. Those predicted to use translation table 4 or 15 were retained for downstream
233 zamples.
233 (https://github. calling 2. Those predicted to use translation table 4 or 15 were retained for downstream

analysis.

237 The Unified Human Gut Virome Catalog (UHGV) was filtered for high quality and comple

237 The Unified Human Gut Viro
- 236
-
-
-
-
-

242 Prokka

-
-
-
- 236

236

237 The Unifi

238 VOTUs de

239 4 or 15 (b)

240 predictec

241

242 **Prokka**

243 A fork of

243 dentic c

245 step is us

246 genetic c

247 code 15 (

248 (github.c

248 database

250 (http://si

250 Pharokka **241 Predicted For Value 240** and integrance for Intellecting the SML and 241
241 Prokka
242 A fork of Prokka v1.14.5¹¹ was written that incorporates an initial stage of ORF prediction
244 using Prodigal-gv v2.11.0 (htt using Prodigal-given and the genetic code most likely adopted by the genometic code is used to perform the translation FASTX::Seq, which we usede 15 (<u>metacpan.org/pod/FASTX::Seq</u>)¹⁶. The code for this is available githu
- code 15 (metacpan.org/pod/FASTX::Seq)¹⁶. The code for this is available at
-
-
- $(\text{http://s3.climb.ac.uk/ADM share/all phrogs.hmm.gz})^{17}$. The fork is installable from
-
-

- (github.com/telatin/metaprokka). We indicatabase in our Prokka-gv annotations
(http://s3.climb.ac.uk/ADM_share/all_ph
Bioconda as 'metaprokka'.
Pharokka
Pharokka
wharokka v1.5.0¹² was updated to include
gv as a gene pre Bioconda as 'metaprokka'.

Pharokka

Pharokka v1.5.0¹² was updated to include support for p

gv as a gene predictor. This is specified by using '-g proc

updated code is available on GitHub (https://github.cor

uses tRNA
- 238 vOTUs deemed to be a "high confidence" virus and predicted to use either translation table
239 4 or 15 (https://github.com/snayfach/UHGV). Stop codon reassignment had already been
249 and 15 (https://github.com/snayfa 238 4 or 15 (https://github.com/snayfach/UHGV). Stop codon reassignment had already been
248 predicted for UHGV vOTUs using Prodigal-gv and is available in the UHGV metadata.
241
242 **Prokka**
242 **Prokka**
243 **Prokka**
243 249 **Producted for UHGV vOTUs using Productley** and is available in the UHGV metadata.

241

241

242 **Prokka**

241

242 **A** f 243 A fork of Prokka v1.14.5⁻⁴ was written that incorporates an initial stage of ORF prediction
244 using Prodigal-gv v2.11.0 <u>(https://github.com/apcamargo/prodigal-gv</u>⁸. A first gene callin
345 stage is used to infe using Prodigal-gv v2.11.0 (https://github.com/apcamargo/prodigal-gv)". A first gene calling
step is used to infer the genetic code most likely adopted by the genome, then the predicte
genetic code is used to perform the tr 246 step is used to perform the translation FASTX::Seq, which we updated to accept
247 code 15 (metacpan.org/pod/FASTX::Seq)³⁶. The code for this is available at
248 (github.com/telatin/metaprokka). We included publicly 247 code 15 (<u>metacpan.org/pod/FASTX::Seq</u>)¹⁶. The code for this is available at

248 (github.com/telatin/metaprokka). We included publicly available HMMs of the PHROGs

248 database in our Prokka-gy annotations

250 (h 248 (github.com/telatin/metaprokka). We included publicly available HMMs of

248 database in our Prokka-gv annotations

250 (http://s3.climb.ac.uk/ADM_share/all_phrogs.hmm.gz)¹⁷. The fork is install.

251 Bioconda as 'm 249 (http://s3.climb.ac.uk/ADM_share/all_phrogs.hmm.gz)¹⁷. The fork is installable from
250 (http://s3.climb.ac.uk/ADM_share/all_phrogs.hmm.gz)¹⁷. The fork is installable from
251 Bioconda as 'metaprokka'.
252
253 **Pha** 250 (http://s3.climb.ac.uk/ADM_share/all_
251 Bioconda as 'metaprokka'.
252
253 Pharokka
254 Pharokka v1.5.0¹² was updated to inclu
255 gv as a gene predictor. This is specified
256 updated code is available on GitHub (251 Bioconda as 'metaprokka'.

251 Bioconda as 'metaprokka'.

252

252 Pharokka v1.5.0¹² was updated to include support for pyrodigal-gv implementing pyre

255 gy as a gene predictor. This is specified by using '-g prodi 252

252

253 **Pharokka**

254 Pharokka v1.5.0¹² was upd

255 gv as a gene predictor. This

256 updated code is available c

257 uses tRNAscan-SE for predi 253 Pharokka
254 Pharokka
255 gv as a ger
256 updated c
257 uses tRNA 254 Pharokka v1.5.0⁴⁴ was updated to include support for pyrodigal-gv implementing pyrodigal-
255 gv as a gene predictor. This is specified by using '-g prodigal-gv' when running Pharokka. The
1256 updated code is avail
- 256 gr as a gene predictor. This is operator. The is operator. The specified by a group of the specified by the prediction of the specified by the com/gbouras13/pharokka. Pharokka. The specified by the specified by the spe
- 257 uses tRNAscan-SE for predicting tRNAs¹⁴. uses tRNAscan-SE for predicting tRNAs14 257 .

available under [aCC-BY 4.0 International license.](http://creativecommons.org/licenses/by/4.0/) (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made bioRxiv preprint doi: [https://doi.org/10.1101/2023.12.19.572299;](https://doi.org/10.1101/2023.12.19.572299) this version posted December 19, 2023. The copyright holder for this preprint

258

259 Statistical Analyses and Data Visualisation

-
- 260 To test for significance in differences of results, and 2^{16} and P values were adjusted using the Benjamini-Hochberg procedure³¹. Figure 1.
262 Produced using ggplot 2 v3.4.2²⁰. v4.2.2¹⁹ and P-values were adjusted using the Benjamini-Hochberg procedure¹⁹. Figure 1 was
produced using ggplot2 v3.4.2⁶².
- produced using ggplot2 v3.4.2²⁰.
 $\frac{1}{2}$

available under [aCC-BY 4.0 International license.](http://creativecommons.org/licenses/by/4.0/) (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made bioRxiv preprint doi: [https://doi.org/10.1101/2023.12.19.572299;](https://doi.org/10.1101/2023.12.19.572299) this version posted December 19, 2023. The copyright holder for this preprint

263 Supplementary Results

264 Prokka-gv Annotations

 wifi the think think is a series of the 265 Franslation table 15, for which Prokka-gv increased the median gene length (median
267 Franslation table 15, for which Prokka-gv increased the median gene length (median
267 genome medians) from 276 to 396 bp for UHGV 267 genome medians) from 276 to 396 to for UHGV sequences (43.5% increase), and from 309
268 to 483 to for INPHARED sequences (56.3% increase). This was also reflected in an increase
266 of median coding capacity from 66.6 268
268 to 483 bp for INPHARED sequences (56.3% increase). This was also reflected in an increase
267 of nedian coding capacity from 86.6% to 86.7% for UHGV, and from 69.2% to 87.3% for
277 INPHARED. As it is commonly use 268 of median coding capacity from 66.6% to 86.7% for UHGV, and from 69.2% to 87.3% for INPHARED. As it is commonly the major capacity morte from the sterior phase, we investigated how commonly the major capacity protein 279 INPHARED. As it is commonly used as a phylogenetic marker for bacteriophages, we
investigated how commonly the major capacity frotten (MCP) could be identified with an
without predicted stop codon reassignment¹³. Fo 271 Investigated how commonly the major capsid grotein (MCP) could be identified with

272 Investigated how commonly the major capsid grotein (MCP) could be identified with

271 while used as the MCP could be identified o 272 without predicted stop codon reassignment". For sequences predicted to use translation
273 white 15, the MCP could be identified on 382/715 (53.4%) sequences with Prokka and this
274 which investigated to 386/715 (53. without predicted stop codon reassignment". For sequences predicted to use translation
cases table 15, the MCP could be identified on 382/715 (53.4%) sequences with Prokka and this
274 was marginally increased to 386/715 274 vas marginally increased to 386/715 (53.9%) with Prokka-gv.
273
274 When investigating the sequences for which translation table 4 was predicted, a substantial
275 Increase was also observed for UHGV sequences, with P 275

275

275

2775

2776

2776 when investigating the sequences for which translation table

2776 mercase was also observed of UHGV sequences, with Prokka

279 gene length from 319 to 460 bp (44.2%), resulting in an incre 275 increase was also observed for UHGV sequences, with Prokka-gv increasing median median
gene length from 319 to 460 bp (44.2%), resulting in an increase of coding capacky from
78.4% to 91.4%. However, the same was not obse 278 gene length from 319 to 460 bp (44.2%), resulting in an increase of coding capacity from
278 gene length from 319 to 460 bp (44.2%), resulting in an increase of coding capacity from
278 78.4% to 91.4%. However, the sam

278 478 do 91.4%. However, the same was not observed for INPHARED sequences predicte
278 478 to 91.4%. However, the same was not observed for INPHARED sequences predicte
281 to 460 bp (41.2%), for an increase of an increas

279.
279. Use translation table 4. These sequences observed a modest increase in median median
281. gene length from 573 to 584 bp (1.8%) for Prokka-gv. Median coding capacity was not
282. increased with Prokka and Prokka-

281 gene length from 573 to 584 bp (1.8%) for Prokka-gv. Median coding capacity was not increased with Prokka and Prokka-gv both obtaining 86.2%.

282 increased with Prokka and Prokka-gv both obtaining 86.2%. 282 increased with Prokka and Prokka-gv both obtaining 86.2%.