GigaScience

Impact of reference design on estimating SARS-CoV-2 lineage abundances from wastewater sequencing data

--Manuscript Draft--

Manuscript Number:	GIGA-D-23-00161
Full Title:	Impact of reference design on estimating SARS-CoV-2 lineage abundances from wastewater sequencing data
Article Type:	Research
Funding Information:	
Abstract:	Background Sequencing of SARS-CoV-2 RNA from wastewater samples has emerged as a valuable tool for detecting the presence and relative abundances of SARS-CoV-2 variants in a community. By analyzing the viral genetic material present in wastewater, public health officials can gain early insights into the spread of the virus and inform timely intervention measures. The construction of reference datasets from known SARS-CoV-2 lineages and their mutation profiles has become state-of-the-art for assigning viral lineages and their relative abundances from wastewater sequencing data. However, the selection of reference sequences or mutations directly affects the predictive power. Results Here, we show the impact of a mutation- and sequence-based reference reconstruction for SARS-CoV-2 abundance estimation. We benchmark three data sets: 1) synthetic “spike-in” mixtures, 2) German samples from early 2021, mainly comprising Alpha, and 3) samples obtained from wastewater at an international airport in Germany from the end of 2021, including first signals of Omicron. The two approaches differ in sub-lineage detection, with the marker-mutation-based method, in particular, being challenged by the increasing number of mutations and lineages. However, the estimations of both approaches depend on selecting representative references and optimized parameter settings. By performing parameter escalation experiments, we demonstrate the effects of reference size and alternative allele frequency cutoffs for abundance estimation. We show how different parameter settings can lead to different results for our test data sets, and illustrate the effects of virus lineage composition of wastewater samples and references. Conclusions Here, we compare a mutation- and sequence-based reference construction and assignment for SARS-CoV-2 abundance estimation from wastewater samples. Our study highlights current computational challenges, focusing on the general reference design, which significantly and directly impacts abundance allocations. We illustrate advantages and disadvantages that may be relevant for further developments in the wastewater community and in the context of higher standardization.
Corresponding Author:	Martin Hölzer Robert Koch Institut Berlin, GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Robert Koch Institut
Corresponding Author's Secondary Institution:	
First Author:	Eva Aßmann
First Author Secondary Information:	
Order of Authors:	Eva Aßmann
	Shelesh Agrawal
	Laura Orschler
	Sindy Böttcher
	Susanne Lackner

	Martin Hölzer
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in	Yes

the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

```
This is pdfTeX, Version 3.141592653-2.6-1.40.24 (TeX Live 2022)
(preloaded format=pdflatex 2023.3.8)  12 JUN 2023 13:22
entering extended mode
 restricted \write18 enabled.
 %&-line parsing enabled.
**main.tex
(./main.tex
LaTeX2e <2022-11-01> patch level 1
L3 programming layer <2023-02-22> (./oup-contemporary.cls
Document Class: oup-contemporary 2017/06/28, v1.1
(c:/TeXLive/2022/texmf-dist/tex/latex/base/article.cls
Document Class: article 2022/07/02 v1.4n Standard LaTeX document class
(c:/TeXLive/2022/texmf-dist/tex/latex/base/size10.clo
File: size10.clo 2022/07/02 v1.4n Standard LaTeX file (size option)
)
\c@part=\count185
\c@section=\count186
\c@subsection=\count187
\c@subsubsection=\count188
\c@paragraph=\count189
\c@subparagraph=\count190
\c@figure=\count191
\c@table=\count192
\abovecaptionskip=\skip48
\belowcaptionskip=\skip49
\bibindent=\dimen140
) (c:/TeXLive/2022/texmf-dist/tex/latex/base/inputenc.sty
Package: inputenc 2021/02/14 v1.3d Input encoding file
\inpenc@prehook=\toks16
\inpenc@posthook=\toks17
) (c:/TeXLive/2022/texmf-dist/tex/latex/base/fontenc.sty
Package: fontenc 2021/04/29 v2.0v Standard LaTeX package
) (c:/TeXLive/2022/texmf-dist/tex/generic/iftex/ifpdf.sty
Package: ifpdf 2019/10/25 v3.4 ifpdf legacy package. Use iftex instead.
(c:/TeXLive/2022/texmf-dist/tex/generic/iftex/iftex.sty
Package: iftex 2022/02/03 v1.0f TeX engine tests
)) (c:/TeXLive/2022/texmf-dist/tex/latex/microtype/microtype.sty
Package: microtype 2023/03/13 v3.1a Micro-typographical refinements (RS)
(c:/TeXLive/2022/texmf-dist/tex/latex/graphics/keyval.sty
Package: keyval 2022/05/29 v1.15 key=value parser (DPC)
\KV@toks@=\toks18
) (c:/TeXLive/2022/texmf-dist/tex/latex/etoolbox/etoolbox.sty
Package: etoolbox 2020/10/05 v2.5k e-TeX tools for LaTeX (JAW)
\etb@tempcnta=\count193
)
\MT@toks=\toks19
\MT@tempbox=\box51
\MT@count=\count194
LaTeX Info: Redefining \noprotrusionifhmode on input line 1059.
LaTeX Info: Redefining \leftprotrusion on input line 1060.
\MT@prot@toks=\toks20
LaTeX Info: Redefining \rightprotrusion on input line 1078.
LaTeX Info: Redefining \textls on input line 1368.
\MT@outer@kern=\dimen141
```

```
LaTeX Info: Redefining \textmicrotypecontext on input line 1988.
\MT@listname@count=\count195
(c:/TeXLive/2022/texmf-dist/tex/latex/microtype/microtype-pdftex.def
File: microtype-pdftex.def 2023/03/13 v3.1a Definitions specific to
pdftex (RS)

LaTeX Info: Redefining \lsstyle on input line 902.
LaTeX Info: Redefining \lslig on input line 902.
\MT@outer@space=\skip50
)
Package microtype Info: Loading configuration file microtype.cfg.
(c:/TeXLive/2022/texmf-dist/tex/latex/microtype/microtype.cfg
File: microtype.cfg 2023/03/13 v3.1a microtype main configuration file
(RS)
)) (c:/TeXLive/2022/texmf-dist/tex/latex/euler/euler.sty
Package: euler 1995/03/05 v2.5
Package: `euler' v2.5 <1995/03/05> (FJ and FMi)
LaTeX Font Info:    Redeclaring symbol font `letters' on input line 35.
LaTeX Font Info:    Encoding `OML' has changed to `U' for symbol font
(Font)              `letters' in the math version `normal' on input line
35.
LaTeX Font Info:    Overwriting symbol font `letters' in version `normal'
(Font)                  OML/cmm/m/it --> U/eur/m/n on input line 35.
LaTeX Font Info:    Encoding `OML' has changed to `U' for symbol font
(Font)              `letters' in the math version `bold' on input line
35.
LaTeX Font Info:    Overwriting symbol font `letters' in version `bold'
(Font)                  OML/cmm/b/it --> U/eur/m/n on input line 35.
LaTeX Font Info:    Overwriting symbol font `letters' in version `bold'
(Font)                  U/eur/m/n --> U/eur/b/n on input line 36.
LaTeX Font Info:    Redeclaring math symbol \Gamma on input line 47.
LaTeX Font Info:    Redeclaring math symbol \Delta on input line 48.
LaTeX Font Info:    Redeclaring math symbol \Theta on input line 49.
LaTeX Font Info:    Redeclaring math symbol \Lambda on input line 50.
LaTeX Font Info:    Redeclaring math symbol \Xi on input line 51.
LaTeX Font Info:    Redeclaring math symbol \Pi on input line 52.
LaTeX Font Info:    Redeclaring math symbol \Sigma on input line 53.
LaTeX Font Info:    Redeclaring math symbol \Upsilon on input line 54.
LaTeX Font Info:    Redeclaring math symbol \Phi on input line 55.
LaTeX Font Info:    Redeclaring math symbol \Psi on input line 56.
LaTeX Font Info:    Redeclaring math symbol \Omega on input line 57.
\symEulerFraktur=\mathgroup4
LaTeX Font Info:    Overwriting symbol font `EulerFraktur' in version
`bold'
(Font)                  U/euf/m/n --> U/euf/b/n on input line 63.
LaTeX Info: Redefining \oldstylenums on input line 85.
\symEulerScript=\mathgroup5
LaTeX Font Info:    Overwriting symbol font `EulerScript' in version
`bold'
(Font)                  U/eus/m/n --> U/eus/b/n on input line 93.
LaTeX Font Info:    Redeclaring math symbol \aleph on input line 97.
LaTeX Font Info:    Redeclaring math symbol \Re on input line 98.
LaTeX Font Info:    Redeclaring math symbol \Im on input line 99.
LaTeX Font Info:    Redeclaring math delimiter \vert on input line 101.
```

```
LaTeX Font Info:    Redeclaring math delimiter \backslash on input line
103.
LaTeX Font Info:    Redeclaring math symbol \neg on input line 106.
LaTeX Font Info:    Redeclaring math symbol \wedge on input line 108.
LaTeX Font Info:    Redeclaring math symbol \vee on input line 110.
LaTeX Font Info:    Redeclaring math symbol \setminus on input line 112.
LaTeX Font Info:    Redeclaring math symbol \sim on input line 113.
LaTeX Font Info:    Redeclaring math symbol \mid on input line 114.
LaTeX Font Info:    Redeclaring math delimiter \arrowvert on input line
116.
LaTeX Font Info:    Redeclaring math symbol \mathsection on input line
117.
\symEulerExtension=\mathgroup6
LaTeX Font Info:    Redeclaring math symbol \coprod on input line 125.
LaTeX Font Info:    Redeclaring math symbol \prod on input line 125.
LaTeX Font Info:    Redeclaring math symbol \sum on input line 125.
LaTeX Font Info:    Redeclaring math symbol \intop on input line 130.
LaTeX Font Info:    Redeclaring math symbol \ointop on input line 131.
LaTeX Font Info:    Redeclaring math symbol \braceld on input line 132.
LaTeX Font Info:    Redeclaring math symbol \bracerd on input line 133.
LaTeX Font Info:    Redeclaring math symbol \bracelu on input line 134.
LaTeX Font Info:    Redeclaring math symbol \braceru on input line 135.
LaTeX Font Info:    Redeclaring math symbol \infty on input line 136.
LaTeX Font Info:    Redeclaring math symbol \nearrow on input line 153.
LaTeX Font Info:    Redeclaring math symbol \searrow on input line 154.
LaTeX Font Info:    Redeclaring math symbol \nwarrow on input line 155.
LaTeX Font Info:    Redeclaring math symbol \swarrow on input line 156.
LaTeX Font Info:    Redeclaring math symbol \Leftrightarrow on input line
157.
LaTeX Font Info:    Redeclaring math symbol \Leftarrow on input line 158.
LaTeX Font Info:    Redeclaring math symbol \Rightarrow on input line
159.
LaTeX Font Info:    Redeclaring math symbol \leftrightarrow on input line
160.
LaTeX Font Info:    Redeclaring math symbol \leftarrow on input line 161.
LaTeX Font Info:    Redeclaring math symbol \rightarrow on input line
163.
LaTeX Font Info:    Redeclaring math delimiter \uparrow on input line
166.
LaTeX Font Info:    Redeclaring math delimiter \downarrow on input line
168.
LaTeX Font Info:    Redeclaring math delimiter \updownarrow on input line
170.
LaTeX Font Info:    Redeclaring math delimiter \Uparrow on input line
172.
LaTeX Font Info:    Redeclaring math delimiter \Downarrow on input line
174.
LaTeX Font Info:    Redeclaring math delimiter \Updownarrow on input line
176.
LaTeX Font Info:    Redeclaring math symbol \leftharpoonup on input line
177.
LaTeX Font Info:    Redeclaring math symbol \leftharpoondown on input
line 178.
```

```
LaTeX Font Info:    Redeclaring math symbol \rightharpoonup on input line
179.
LaTeX Font Info:    Redeclaring math symbol \rightharpoondown on input
line 180
.
LaTeX Font Info:    Redeclaring math delimiter \lbrace on input line 182.
LaTeX Font Info:    Redeclaring math delimiter \rbrace on input line 184.
\symcmmigroup=\mathgroup7
LaTeX Font Info:    Overwriting symbol font `cmmigroup' in version `bold'
(Font)                  OML/cmm/m/it --> OML/cmm/b/it on input line 200.
LaTeX Font Info:    Redeclaring math accent \vec on input line 201.
LaTeX Font Info:    Redeclaring math symbol \triangleleft on input line
202.
LaTeX Font Info:    Redeclaring math symbol \triangleright on input line
203.
LaTeX Font Info:    Redeclaring math symbol \star on input line 204.
LaTeX Font Info:    Redeclaring math symbol \lhook on input line 205.
LaTeX Font Info:    Redeclaring math symbol \rhook on input line 206.
LaTeX Font Info:    Redeclaring math symbol \flat on input line 207.
LaTeX Font Info:    Redeclaring math symbol \natural on input line 208.
LaTeX Font Info:    Redeclaring math symbol \sharp on input line 209.
LaTeX Font Info:    Redeclaring math symbol \smile on input line 210.
LaTeX Font Info:    Redeclaring math symbol \frown on input line 211.
LaTeX Font Info:    Redeclaring math accent \grave on input line 245.
LaTeX Font Info:    Redeclaring math accent \acute on input line 246.
LaTeX Font Info:    Redeclaring math accent \tilde on input line 247.
LaTeX Font Info:    Redeclaring math accent \ddot on input line 248.
LaTeX Font Info:    Redeclaring math accent \check on input line 249.
LaTeX Font Info:    Redeclaring math accent \breve on input line 250.
LaTeX Font Info:    Redeclaring math accent \bar on input line 251.
LaTeX Font Info:    Redeclaring math accent \dot on input line 252.
LaTeX Font Info:    Redeclaring math accent \hat on input line 254.
) (c:/TeXLive/2022/texmf-dist/tex/latex/merriweather/merriweather.sty
Package: merriweather 2022/09/20 (Bob Tennent) Supports
Merriweather(Sans) font
s for all LaTeX engines.
(c:/TeXLive/2022/texmf-dist/tex/generic/iftex/ifxetex.sty
Package: ifxetex 2019/10/25 v0.7 ifxetex legacy package. Use iftex
instead.
) (c:/TeXLive/2022/texmf-dist/tex/generic/iftex/ifluatex.sty
Package: ifluatex 2019/10/25 v1.5 ifluatex legacy package. Use iftex
instead.
) (c:/TeXLive/2022/texmf-dist/tex/latex/base/textcomp.sty
Package: textcomp 2020/02/02 v2.0n Standard LaTeX package
) (c:/TeXLive/2022/texmf-dist/tex/latex/xkeyval/xkeyval.sty
Package: xkeyval 2022/06/16 v2.9 package option processing (HA)
(c:/TeXLive/2022/texmf-dist/tex/generic/xkeyval/xkeyval.tex
(c:/TeXLive/2022/te
xmf-dist/tex/generic/xkeyval/xkvutils.tex
\XKV@toks=\toks21
\XKV@tempa@toks=\toks22
)
\XKV@depth=\count196
File: xkeyval.tex 2014/12/03 v2.7a key=value parser (HA)
```

```
)) (c:/TeXLive/2022/texmf-dist/tex/latex/base/fontenc.sty
Package: fontenc 2021/04/29 v2.0v Standard LaTeX package
) (c:/TeXLive/2022/texmf-dist/tex/latex/fontaxes/fontaxes.sty
Package: fontaxes 2020/07/21 v1.0e Font selection axes
LaTeX Info: Redefining \upshape on input line 29.
LaTeX Info: Redefining \itshape on input line 31.
LaTeX Info: Redefining \slshape on input line 33.
LaTeX Info: Redefining \swshape on input line 35.
LaTeX Info: Redefining \scshape on input line 37.
LaTeX Info: Redefining \sscshape on input line 39.
LaTeX Info: Redefining \ulcshape on input line 41.
LaTeX Info: Redefining \textsw on input line 47.
LaTeX Info: Redefining \textssc on input line 48.
LaTeX Info: Redefining \textulc on input line 49.
)) (c:/TeXLive/2022/texmf-dist/tex/latex/mathastext/mathastext.sty
Package: mathastext 2022/11/04 v1.3y Use the text font in math mode (JFB)
\mst@exists@muskip=\muskip16
\mst@forall@muskip=\muskip17
\mst@prime@muskip=\muskip18
\mst@do@nonletters=\toks23
\mst@do@easynonletters=\toks24
\mst@do@az=\toks25
\mst@do@AZ=\toks26
\symmtoperatorfont=\mathgroup8
\symmtletterfont=\mathgroup9
** ! and ?
** punctuation: , . : ; and \colon
LaTeX Info: Redefining \relbar on input line 844.
LaTeX Info: Redefining \rightarrowfill on input line 847.
LaTeX Info: Redefining \leftarrowfill on input line 852.
** + and =
LaTeX Info: Redefining \Relbar on input line 943.
** adding = ; and + to \nfss@catcodes
** parentheses ( ) [ ] and slash /
** alldelims: < > \backslash \setminus | \vert \mid \{ and \}
LaTeX Font Info:    Redeclaring math delimiter \backslash on input line
989.
LaTeX Font Info:    Redeclaring math symbol \setminus on input line 1001.
LaTeX Info: Redefining \models on input line 1010.
** \# \mathdollar \% \&
** \imath and \jmath
LaTeX Font Info:    Overwriting math alphabet `\mathnormalbold' in
version `nor
mal'
(Font)                  T1/Merriwthr-OsF/b/it --> T1/Merriwthr-OsF/b/it
on inpu
t line 2370.
LaTeX Font Info:    Overwriting math alphabet `\mathnormalbold' in
version `bol
d'
(Font)                  T1/Merriwthr-OsF/b/it --> T1/Merriwthr-OsF/b/it
on inpu
t line 2370.
```

```
LaTeX Font Info:    Overwriting symbol font `mtletterfont' in version
`normal'
(Font)                  T1/Merriwthr-OsF/m/it --> T1/Merriwthr-OsF/m/it
on inpu
t line 2370.
LaTeX Font Info:    Overwriting symbol font `mtletterfont' in version
`bold'
(Font)                  T1/Merriwthr-OsF/m/it --> T1/Merriwthr-OsF/b/it
on inpu
t line 2370.
LaTeX Font Info:    Overwriting symbol font `mtoperatorfont' in version
`normal
'
(Font)                  T1/Merriwthr-OsF/m/n --> T1/Merriwthr-OsF/m/n on
input
line 2370.
LaTeX Font Info:    Overwriting symbol font `mtoperatorfont' in version
`bold'
(Font)                  T1/Merriwthr-OsF/m/n --> T1/Merriwthr-OsF/b/n on
input
line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathbf' in version
`normal'
(Font)                  T1/Merriwthr-OsF/b/n --> T1/Merriwthr-OsF/b/n on
input
line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathbf' in version `bold'
(Font)                  T1/Merriwthr-OsF/b/n --> T1/Merriwthr-OsF/b/n on
input
line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathit' in version
`normal'
(Font)                  T1/Merriwthr-OsF/m/it --> T1/Merriwthr-OsF/m/it
on inpu
t line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathit' in version `bold'
(Font)                  T1/Merriwthr-OsF/m/it --> T1/Merriwthr-OsF/b/it
on inpu
t line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathsf' in version
`normal'
(Font)                  T1/MerriwthrSans-OsF/m/n --> T1/MerriwthrSans-
OsF/m/n o
n input line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathsf' in version `bold'
(Font)                  T1/MerriwthrSans-OsF/m/n --> T1/MerriwthrSans-
OsF/b/n o
n input line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathtt' in version
`normal'
(Font)                  T1/lmtt/m/n --> T1/lmtt/m/n on input line 2370.
LaTeX Font Info:    Overwriting math alphabet `\Mathtt' in version `bold'
(Font)                  T1/lmtt/m/n --> T1/lmtt/b/n on input line 2370.
** Latin letters in the `normal' (resp. `bold') math versions are now
```

```
** set up to use the fonts T1/Merriwthr-OsF/m(b)/it
** Other characters (digits, ...) and \log-like names will be
** typeset with the n shape.
** \hbar
** minus as endash
** \HUGE has been (re)-defined.
** mathastext has declared larger sizes for subscripts.
** To keep LaTeX defaults, use option `defaultmathsizes'.
) (c:/TeXLive/2022/texmf-dist/tex/latex/relsize/relsize.sty
Package: relsize 2013/03/29 ver 4.1
) (c:/TeXLive/2022/texmf-dist/tex/latex/ragged2e/ragged2e.sty
Package: ragged2e 2023/02/25 v3.4 ragged2e Package
\CenteringLeftskip=\skip51
\RaggedLeftLeftskip=\skip52
\RaggedRightLeftskip=\skip53
\CenteringRightskip=\skip54
\RaggedLeftRightskip=\skip55
\RaggedRightRightskip=\skip56
\CenteringParfillskip=\skip57
\RaggedLeftParfillskip=\skip58
\RaggedRightParfillskip=\skip59
\JustifyingParfillskip=\skip60
\CenteringParindent=\skip61
\RaggedLeftParindent=\skip62
\RaggedRightParindent=\skip63
\JustifyingParindent=\skip64
) (c:/TeXLive/2022/texmf-dist/tex/latex/xcolor/xcolor.sty
Package: xcolor 2022/06/12 v2.14 LaTeX color extensions (UK)
(c:/TeXLive/2022/texmf-dist/tex/latex/graphics-cfg/color.cfg
File: color.cfg 2016/01/02 v1.6 sample color configuration
)
Package xcolor Info: Driver file: pdftex.def on input line 227.
(c:/TeXLive/2022/texmf-dist/tex/latex/graphics-def/pdftex.def
File: pdftex.def 2022/09/22 v1.2b Graphics/color driver for pdftex
) (c:/TeXLive/2022/texmf-dist/tex/latex/graphics/mathcolor.ltx)
Package xcolor Info: Model `cmy' substituted by `cmy0' on input line
1353.
Package xcolor Info: Model `hsb' substituted by `rgb' on input line 1357.
Package xcolor Info: Model `RGB' extended on input line 1369.
Package xcolor Info: Model `HTML' substituted by `rgb' on input line
1371.
Package xcolor Info: Model `Hsb' substituted by `hsb' on input line 1372.
Package xcolor Info: Model `tHsb' substituted by `hsb' on input line
1373.
Package xcolor Info: Model `HSB' substituted by `hsb' on input line 1374.
Package xcolor Info: Model `Gray' substituted by `gray' on input line
1375.
Package xcolor Info: Model `wave' substituted by `hsb' on input line
1376.
) (c:/TeXLive/2022/texmf-dist/tex/latex/colortbl/colortbl.sty
Package: colortbl 2022/06/20 v1.0f Color table columns (DPC)
(c:/TeXLive/2022/texmf-dist/tex/latex/tools/array.sty
Package: array 2022/09/04 v2.5g Tabular extension package (FMi)
\col@sep=\dimen142
```

```
\ar@mcellbox=\box52
\extrarowheight=\dimen143
\NC@list=\toks27
\extratabsurround=\skip65
\backup@length=\skip66
\ar@cellbox=\box53
)
\everycr=\toks28
\minrowclearance=\skip67
\rownum=\count197
) (c:/TeXLive/2022/texmf-dist/tex/latex/graphics/graphicx.sty
Package: graphicx 2021/09/16 v1.2d Enhanced LaTeX Graphics (DPC,SPQR)
(c:/TeXLive/2022/texmf-dist/tex/latex/graphics/graphics.sty
Package: graphics 2022/03/10 v1.4e Standard LaTeX Graphics (DPC,SPQR)
(c:/TeXLive/2022/texmf-dist/tex/latex/graphics/trig.sty
Package: trig 2021/08/11 v1.11 sin cos tan (DPC)
) (c:/TeXLive/2022/texmf-dist/tex/latex/graphics-cfg/graphics.cfg
File: graphics.cfg 2016/06/04 v1.11 sample graphics configuration
)
Package graphics Info: Driver file: pdftex.def on input line 107.
)
\Gin@req@height=\dimen144
\Gin@req@width=\dimen145
) (c:/TeXLive/2022/texmf-dist/tex/latex/xpatch/xpatch.sty
(c:/TeXLive/2022/texm
f-dist/tex/latex/l3kernel/expl3.sty
Package: expl3 2023-02-22 L3 programming layer (loader)
(c:/TeXLive/2022/texmf-dist/tex/latex/l3backend/l3backend-pdftex.def
File: l3backend-pdftex.def 2023-01-16 L3 backend support: PDF output
(pdfTeX)
\l__color_backend_stack_int=\count198
\l__pdf_internal_box=\box54
))
Package: xpatch 2020/03/25 v0.3a Extending etoolbox patching commands
(c:/TeXLive/2022/texmf-dist/tex/latex/l3packages/xparse/xparse.sty
Package: xparse 2023-02-02 L3 Experimental document command parser
)) (c:/TeXLive/2022/texmf-dist/tex/latex/environ/environ.sty
Package: environ 2014/05/04 v0.3 A new way to define environments
(c:/TeXLive/2022/texmf-dist/tex/latex/trimspaces/trimspaces.sty
Package: trimspaces 2009/09/17 v1.1 Trim spaces around a token list
)
\@envbody=\toks29
) (c:/TeXLive/2022/texmf-dist/tex/latex/lastpage/lastpage.sty
Package: lastpage 2023/03/07 v2.0a lastpage: 2.09 or 2e? (HMM)
(c:/TeXLive/2022/texmf-dist/tex/latex/lastpage/lastpage2e.sty
Package: lastpage2e 2023/03/07 v2.0a Decide which 2e lastpage version to
use (H
MM)
(c:/TeXLive/2022/texmf-dist/tex/latex/lastpage/lastpagemodern.sty
Package: lastpagemodern 2023-03-07 v2.0a Refers to last page's name (HMM;
JPG)
)
)) (c:/TeXLive/2022/texmf-dist/tex/latex/graphics/rotating.sty
Package: rotating 2016/08/11 v2.16d rotated objects in LaTeX
```

```
(c:/TeXLive/2022/texmf-dist/tex/latex/base/ifthen.sty
Package: ifthen 2022/04/13 v1.1d Standard LaTeX ifthen package (DPC)
)
\c@r@tfl@t=\count199
\rotFPtop=\skip68
\rotFPbot=\skip69
\rot@float@box=\box55
\rot@mess@toks=\toks30
) (c:/TeXLive/2022/texmf-dist/tex/latex/graphics/lscape.sty
Package: lscape 2020/05/28 v3.02 Landscape Pages (DPC)
) (c:/TeXLive/2022/texmf-dist/tex/latex/tools/afterpage.sty
Package: afterpage 2014/10/28 v1.08 After-Page Package (DPC)
\AP@output=\toks31
\AP@partial=\box56
\AP@footins=\box57
) (c:/TeXLive/2022/texmf-dist/tex/latex/textpos/textpos.sty
Package: textpos 2022/07/23 v1.10.1
Package textpos Info: choosing support for LaTeX3 on input line 60.
\TP@textbox=\box58
\TP@holdbox=\box59
\TPHorizModule=\dimen146
\TPVertModule=\dimen147
\TP@margin=\dimen148
\TP@absmargin=\dimen149
Grid set 16 x 16 = 37.34424pt x 52.81541pt
\TPboxrulesize=\dimen150
\TP@ox=\dimen151
\TP@oy=\dimen152
\TP@tbargs=\toks32
TextBlockOrigin set to 0pt x 0pt
) (c:/TeXLive/2022/texmf-dist/tex/latex/url/url.sty
\Urlmuskip=\muskip19
Package: url 2013/09/16  ver 3.4  Verb mode for urls, etc.
) (c:/TeXLive/2022/texmf-dist/tex/latex/newfloat/newfloat.sty
Package: newfloat 2019/09/02 v1.1l Defining new floating environments
(AR)
Package newfloat Info: `rotating' package detected.
) (c:/TeXLive/2022/texmf-dist/tex/latex/mdframed/mdframed.sty
Package: mdframed 2013/07/01 1.9b: mdframed
(c:/TeXLive/2022/texmf-dist/tex/latex/kvoptions/kvoptions.sty
Package: kvoptions 2022-06-15 v3.15 Key value format for package options
(HO)
(c:/TeXLive/2022/texmf-dist/tex/generic/ltxcmds/ltxcmds.sty
Package: ltxcmds 2020-05-10 v1.25 LaTeX kernel commands for general use
(HO)
) (c:/TeXLive/2022/texmf-dist/tex/latex/kvsetkeys/kvsetkeys.sty
Package: kvsetkeys 2022-10-05 v1.19 Key value parser (HO)
)) (c:/TeXLive/2022/texmf-dist/tex/latex/zref/zref-abspage.sty
Package: zref-abspage 2022-04-07 v2.34 Module abspage for zref (HO)
(c:/TeXLive/2022/texmf-dist/tex/latex/zref/zref-base.sty
Package: zref-base 2022-04-07 v2.34 Module base for zref (HO)
(c:/TeXLive/2022/texmf-dist/tex/generic/infwarerr/infwarerr.sty
Package: infwarerr 2019/12/03 v1.5 Providing info/warning/error messages
(HO)
```

```
) (c:/TeXLive/2022/texmf-dist/tex/generic/kvdefinekeys/kvdefinekeys.sty
Package: kvdefinekeys 2019-12-19 v1.6 Define keys (HO)
) (c:/TeXLive/2022/texmf-dist/tex/generic/pdftexcmds/pdftexcmds.sty
Package: pdftexcmds 2020-06-27 v0.33 Utility functions of pdfTeX for
LuaTeX (HO
)
Package pdftexcmds Info: \pdf@primitive is available.
Package pdftexcmds Info: \pdf@ifprimitive is available.
Package pdftexcmds Info: \pdfdraftmode found.
) (c:/TeXLive/2022/texmf-dist/tex/generic/etexcmds/etexcmds.sty
Package: etexcmds 2019/12/15 v1.7 Avoid name clashes with e-TeX commands
(HO)
) (c:/TeXLive/2022/texmf-dist/tex/latex/auxhook/auxhook.sty
Package: auxhook 2019-12-17 v1.6 Hooks for auxiliary files (HO)
)
Package zref Info: New property list: main on input line 767.
Package zref Info: New property: default on input line 768.
Package zref Info: New property: page on input line 769.
) (c:/TeXLive/2022/texmf-dist/tex/latex/base/atbegshi-ltx.sty
Package: atbegshi-ltx 2021/01/10 v1.0c Emulation of the original atbegshi
package with kernel methods
)
\c@abspage=\count266
Package zref Info: New property: abspage on input line 65.
) (c:/TeXLive/2022/texmf-dist/tex/latex/needspace/needspace.sty
Package: needspace 2010/09/12 v1.3d reserve vertical space
)
\mdf@templength=\skip70
\c@mdf@globalstyle@cnt=\count267
\mdf@skipabove@length=\skip71
\mdf@skipbelow@length=\skip72
\mdf@leftmargin@length=\skip73
\mdf@rightmargin@length=\skip74
\mdf@innerleftmargin@length=\skip75
\mdf@innerrightmargin@length=\skip76
\mdf@innertopmargin@length=\skip77
\mdf@innerbottommargin@length=\skip78
\mdf@splittopskip@length=\skip79
\mdf@splitbottomskip@length=\skip80
\mdf@outermargin@length=\skip81
\mdf@innermargin@length=\skip82
\mdf@linewidth@length=\skip83
\mdf@innerlinewidth@length=\skip84
\mdf@middlelinewidth@length=\skip85
\mdf@outerlinewidth@length=\skip86
\mdf@roundcorner@length=\skip87
\mdf@footenotedistance@length=\skip88
\mdf@userdefinedwidth@length=\skip89
\mdf@needspace@length=\skip90
\mdf@frametitleaboveskip@length=\skip91
\mdf@frametitlebelowskip@length=\skip92
\mdf@frametitlerulewidth@length=\skip93
\mdf@frametitleleftmargin@length=\skip94
\mdf@frametitlerightmargin@length=\skip95
```

```
\mdf@shadowsize@length=\skip96
\mdf@extratopheight@length=\skip97
\mdf@subtitleabovelinewidth@length=\skip98
\mdf@subtitlebelowlinewidth@length=\skip99
\mdf@subtitleaboveskip@length=\skip100
\mdf@subtitlebelowskip@length=\skip101
\mdf@subtitleinneraboveskip@length=\skip102
\mdf@subtitleinnerbelowskip@length=\skip103
\mdf@subsubtitleabovelinewidth@length=\skip104
\mdf@subsubtitlebelowlinewidth@length=\skip105
\mdf@subsubtitleaboveskip@length=\skip106
\mdf@subsubtitlebelowskip@length=\skip107
\mdf@subsubtitleinneraboveskip@length=\skip108
\mdf@subsubtitleinnerbelowskip@length=\skip109
(c:/TeXLive/2022/texmf-dist/tex/latex/mdframed/md-frame-0.mdf
File: md-frame-0.mdf 2013/07/01\ 1.9b: md-frame-0
)
\mdf@frametitlebox=\box60
\mdf@footnotebox=\box61
\mdf@splitbox@one=\box62
\mdf@splitbox@two=\box63
\mdf@splitbox@save=\box64
\mdfsplitboxwidth=\skip110
\mdfsplitboxtotalwidth=\skip111
\mdfsplitboxheight=\skip112
\mdfsplitboxdepth=\skip113
\mdfsplitboxtotalheight=\skip114
\mdfframetitleboxwidth=\skip115
\mdfframetitleboxtotalwidth=\skip116
\mdfframetitleboxheight=\skip117
\mdfframetitleboxdepth=\skip118
\mdfframetitleboxtotalheight=\skip119
\mdffootnoteboxwidth=\skip120
\mdffootnoteboxtotalwidth=\skip121
\mdffootnoteboxheight=\skip122
\mdffootnoteboxdepth=\skip123
\mdffootnoteboxtotalheight=\skip124
\mdftotallinewidth=\skip125
\mdfboundingboxwidth=\skip126
\mdfboundingboxtotalwidth=\skip127
\mdfboundingboxheight=\skip128
\mdfboundingboxdepth=\skip129
\mdfboundingboxtotalheight=\skip130
\mdf@freevspace@length=\skip131
\mdf@horizontalwidthofbox@length=\skip132
\mdf@verticalmarginwhole@length=\skip133
\mdf@horizontalspaceofbox=\skip134
\mdfsubtitleheight=\skip135
\mdfsubsubtitleheight=\skip136
\c@mdfcountframes=\count268

****** mdframed patching \endmdf@trivlist

****** -- success******
```

```
\mdf@envdepth=\count269
\c@mdf@env@i=\count270
\c@mdf@env@ii=\count271
\c@mdf@zref@counter=\count272
Package zref Info: New property: mdf@pagevalue on input line 895.
) (c:/TeXLive/2022/texmf-dist/tex/latex/titlesec/titlesec.sty
Package: titlesec 2021/07/05 v2.14 Sectioning titles
\ttl@box=\box65
\beforetitleunit=\skip137
\aftertitleunit=\skip138
\ttl@plus=\dimen153
\ttl@minus=\dimen154
\ttl@toksa=\toks33
\titlewidth=\dimen155
\titlewidthlast=\dimen156
\titlewidthfirst=\dimen157
) (c:/TeXLive/2022/texmf-dist/tex/latex/koma-script/scrextend.sty
Package: scrextend 2022/10/12 v3.38 KOMA-Script package (extend other
classes w
ith features of KOMA-Script classes)
(c:/TeXLive/2022/texmf-dist/tex/latex/koma-script/scrkbase.sty
Package: scrkbase 2022/10/12 v3.38 KOMA-Script package (KOMA-Script-
dependent b
asics and keyval usage)
(c:/TeXLive/2022/texmf-dist/tex/latex/koma-script/scrbase.sty
Package: scrbase 2022/10/12 v3.38 KOMA-Script package (KOMA-Script-
independent
basics and keyval usage)
(c:/TeXLive/2022/texmf-dist/tex/latex/koma-script/scrlfile.sty
Package: scrlfile 2022/10/12 v3.38 KOMA-Script package (file load hooks)
(c:/TeXLive/2022/texmf-dist/tex/latex/koma-script/scrlfile-hook.sty
Package: scrlfile-hook 2022/10/12 v3.38 KOMA-Script package (using LaTeX
hooks)

(c:/TeXLive/2022/texmf-dist/tex/latex/koma-script/scrlogo.sty
Package: scrlogo 2022/10/12 v3.38 KOMA-Script package (logo)
)))
Applying: [2021/05/01] Usage of raw or classic option list on input line
252.
Already applied: [0000/00/00] Usage of raw or classic option list on
input line
 368.
))
Package scrextend Info: unexpected definition of `\@makefnmark'.
(scrextend)             Trying to patch it on input line 1709.
Package scrextend Info: patch seems to be successfull on input line 1709.
)

LaTeX Font Warning: Font shape `T1/cmr/m/n' in size <7.5> not available
(Font)              size <7> substituted on input line 65.

(c:/TeXLive/2022/texmf-dist/tex/latex/tools/calc.sty
Package: calc 2017/05/25 v4.3 Infix arithmetic (KKT,FJ)
```

```
\calc@Acount=\count273
\calc@Bcount=\count274
\calc@Adimen=\dimen158
\calc@Bdimen=\dimen159
\calc@Askip=\skip139
\calc@Bskip=\skip140
LaTeX Info: Redefining \setlength on input line 80.
LaTeX Info: Redefining \addtolength on input line 81.
\calc@Ccount=\count275
\calc@Cskip=\skip141
) (c:/TeXLive/2022/texmf-dist/tex/latex/geometry/geometry.sty
Package: geometry 2020/01/02 v5.9 Page Geometry
(c:/TeXLive/2022/texmf-dist/tex/generic/iftex/ifvtex.sty
Package: ifvtex 2019/10/25 v1.7 ifvtex legacy package. Use iftex instead.
)
\Gm@cnth=\count276
\Gm@cntv=\count277
\c@Gm@tempcnt=\count278
\Gm@bindingoffset=\dimen160
\Gm@wd@mp=\dimen161
\Gm@odd@mp=\dimen162
\Gm@even@mp=\dimen163
\Gm@layoutwidth=\dimen164
\Gm@layoutheight=\dimen165
\Gm@layouthoffset=\dimen166
\Gm@layoutvoffset=\dimen167
\Gm@dimlist=\toks34
) (c:/TeXLive/2022/texmf-dist/tex/latex/hyperref/hyperref.sty
Package: hyperref 2023-02-07 v7.00v Hypertext links for LaTeX
(c:/TeXLive/2022/texmf-dist/tex/generic/pdfescape/pdfescape.sty
Package: pdfescape 2019/12/09 v1.15 Implements pdfTeX's escape features
(HO)
) (c:/TeXLive/2022/texmf-dist/tex/latex/hycolor/hycolor.sty
Package: hycolor 2020-01-27 v1.10 Color options for hyperref/bookmark
(HO)
) (c:/TeXLive/2022/texmf-dist/tex/latex/letltxmacro/letltxmacro.sty
Package: letltxmacro 2019/12/03 v1.6 Let assignment for LaTeX macros (HO)
) (c:/TeXLive/2022/texmf-dist/tex/latex/hyperref/nameref.sty
Package: nameref 2022-05-17 v2.50 Cross-referencing by name of section
(c:/TeXLive/2022/texmf-dist/tex/latex/refcount/refcount.sty
Package: refcount 2019/12/15 v3.6 Data extraction from label references
(HO)
) (c:/TeXLive/2022/texmf-
dist/tex/generic/gettitlestring/gettitlestring.sty
Package: gettitlestring 2019/12/15 v1.6 Cleanup title references (HO)
)
\c@section@level=\count279
)
\@linkdim=\dimen168
\Hy@linkcounter=\count280
\Hy@pagecounter=\count281
(c:/TeXLive/2022/texmf-dist/tex/latex/hyperref/pd1enc.def
File: pd1enc.def 2023-02-07 v7.00v Hyperref: PDFDocEncoding definition
(HO)
```

```
Now handling font encoding PD1 ...
... no UTF-8 mapping file for font encoding PD1
) (c:/TeXLive/2022/texmf-dist/tex/generic/intcalc/intcalc.sty
Package: intcalc 2019/12/15 v1.3 Expandable calculations with integers
(HO)
)
\Hy@SavedSpaceFactor=\count282
(c:/TeXLive/2022/texmf-dist/tex/latex/hyperref/puenc.def
File: puenc.def 2023-02-07 v7.00v Hyperref: PDF Unicode definition (HO)
Now handling font encoding PU ...
... no UTF-8 mapping file for font encoding PU
)
Package hyperref Info: Option `colorlinks' set `true' on input line 4060.
Package hyperref Info: Hyper figures OFF on input line 4177.
Package hyperref Info: Link nesting OFF on input line 4182.
Package hyperref Info: Hyper index ON on input line 4185.
Package hyperref Info: Plain pages OFF on input line 4192.
Package hyperref Info: Backreferencing OFF on input line 4197.
Package hyperref Info: Implicit mode ON; LaTeX internals redefined.
Package hyperref Info: Bookmarks ON on input line 4425.
\c@Hy@tempcnt=\count283
LaTeX Info: Redefining \url on input line 4763.
\XeTeXLinkMargin=\dimen169
(c:/TeXLive/2022/texmf-dist/tex/generic/bitset/bitset.sty
Package: bitset 2019/12/09 v1.3 Handle bit-vector datatype (HO)
(c:/TeXLive/2022/texmf-dist/tex/generic/bigintcalc/bigintcalc.sty
Package: bigintcalc 2019/12/15 v1.5 Expandable calculations on big
integers (HO
)
))
\Fld@menulength=\count284
\Field@Width=\dimen170
\Fld@charsize=\dimen171
Package hyperref Info: Hyper figures OFF on input line 6042.
Package hyperref Info: Link nesting OFF on input line 6047.
Package hyperref Info: Hyper index ON on input line 6050.
Package hyperref Info: backreferencing OFF on input line 6057.
Package hyperref Info: Link coloring ON on input line 6060.
Package hyperref Info: Link coloring with OCG OFF on input line 6067.
Package hyperref Info: PDF/A mode OFF on input line 6072.
\Hy@abspage=\count285
\c@Item=\count286
\c@Hfootnote=\count287
)
Package hyperref Info: Driver (autodetected): hpdftex.
(c:/TeXLive/2022/texmf-dist/tex/latex/hyperref/hpdftex.def
File: hpdftex.def 2023-02-07 v7.00v Hyperref driver for pdfTeX
(c:/TeXLive/2022/texmf-dist/tex/latex/base/atveryend-ltx.sty
Package: atveryend-ltx 2020/08/19 v1.0a Emulation of the original
atveryend pac
kage
with kernel methods
)
\HyAnn@Count=\count288
```

```
\Fld@listcount=\count289
\c@bookmark@seq@number=\count290
(c:/TeXLive/2022/texmf-dist/tex/latex/rerunfilecheck/rerunfilecheck.sty
Package: rerunfilecheck 2022-07-10 v1.10 Rerun checks for auxiliary files
(HO)
(c:/TeXLive/2022/texmf-dist/tex/generic/uniquecounter/uniquecounter.sty
Package: uniquecounter 2019/12/15 v1.4 Provide unlimited unique counter
(HO)
)
Package uniquecounter Info: New unique counter `rerunfilecheck' on input
line 2
85.
)
\Hy@SectionHShift=\skip142
) (c:/TeXLive/2022/texmf-dist/tex/latex/preprint/authblk.sty
Package: authblk 2001/02/27 1.3 (PWD)
\affilsep=\skip143
\@affilsep=\skip144
\c@Maxaffil=\count291
\c@authors=\count292
\c@affil=\count293
) (c:/TeXLive/2022/texmf-dist/tex/latex/footmisc/footmisc.sty
Package: footmisc 2022/03/08 v6.0d a miscellany of footnote facilities
\FN@temptoken=\toks35
\footnotemargin=\dimen172
\@outputbox@depth=\dimen173
Package footmisc Info: Declaring symbol style bringhurst on input line
695.
Package footmisc Info: Declaring symbol style chicago on input line 703.
Package footmisc Info: Declaring symbol style wiley on input line 712.
Package footmisc Info: Declaring symbol style lamport-robust on input
line 723.

Package footmisc Info: Declaring symbol style lamport* on input line 743.
Package footmisc Info: Declaring symbol style lamport*-robust on input
line 764
.
) (c:/TeXLive/2022/texmf-dist/tex/latex/fancyhdr/fancyhdr.sty
Package: fancyhdr 2022/11/09 v4.1 Extensive control of page headers and
footers

\f@nch@headwidth=\skip145
\f@nch@O@elh=\skip146
\f@nch@O@erh=\skip147
\f@nch@O@olh=\skip148
\f@nch@O@orh=\skip149
\f@nch@O@elf=\skip150
\f@nch@O@erf=\skip151
\f@nch@O@olf=\skip152
\f@nch@O@orf=\skip153
) (c:/TeXLive/2022/texmf-dist/tex/generic/alphalph/alphalph.sty
Package: alphalph 2019/12/09 v2.6 Convert numbers to letters (HO)
)
\c@authorfn=\count294
```

```
(c:/TeXLive/2022/texmf-dist/tex/latex/abstract/abstract.sty
Package: abstract 2009/06/08 v1.2a configurable abstracts
\abstitleskip=\skip154
\absleftindent=\skip155
\absrightindent=\skip156
\absparindent=\skip157
\absparsep=\skip158
)
Package newfloat Info: New float `keypoints' with options
`placement=t!,name=kp
t' on input line 286.
\c@keypoints=\count295
\newfloat@ftype=\count296
Package newfloat Info: float type `keypoints'=8 on input line 286.
(c:/TeXLive/2022/texmf-dist/tex/latex/enumitem/enumitem.sty
Package: enumitem 2019/06/20 v3.9 Customized lists
\labelindent=\skip159
\enit@outerparindent=\dimen174
\enit@toks=\toks36
\enit@inbox=\box66
\enit@count@id=\count297
\enitdp@description=\count298
) (c:/TeXLive/2022/texmf-dist/tex/latex/quoting/quoting.sty
Package: quoting 2014/01/28 v0.1c Consolidated environment for displayed
text
\quo@toppartop=\skip160
) (c:/TeXLive/2022/texmf-dist/tex/latex/sttools/stfloats.sty
Package: stfloats 2017/03/27 v3.3 Improve float mechanism and
baselineskip sett
ings
\@dblbotnum=\count299
\c@dblbotnumber=\count300
) (c:/TeXLive/2022/texmf-dist/tex/latex/booktabs/booktabs.sty
Package: booktabs 2020/01/12 v1.61803398 Publication quality tables
\heavyrulewidth=\dimen175
\lightrulewidth=\dimen176
\cmidrulewidth=\dimen177
\belowrulesep=\dimen178
\belowbottomsep=\dimen179
\aboverulesep=\dimen180
\abovetopsep=\dimen181
\cmidrulesep=\dimen182
\cmidrulekern=\dimen183
\defaultaddspace=\dimen184
\@cmidla=\count301
\@cmidlb=\count302
\@aboverulesep=\dimen185
\@belowrulesep=\dimen186
\@thisruleclass=\count303
\@lastruleclass=\count304
\@thisrulewidth=\dimen187
) (c:/TeXLive/2022/texmf-dist/tex/latex/tools/tabularx.sty
Package: tabularx 2020/01/15 v2.11c `tabularx' package (DPC)
\TX@col@width=\dimen188
```

```
\TX@old@table=\dimen189
\TX@old@col=\dimen190
\TX@target=\dimen191
\TX@delta=\dimen192
\TX@cols=\count305
\TX@ftn=\toks37
)
\enitdp@tablenotes=\count306
(c:/TeXLive/2022/texmf-dist/tex/latex/caption/caption.sty
Package: caption 2022/03/01 v3.6b Customizing captions (AR)
(c:/TeXLive/2022/texmf-dist/tex/latex/caption/caption3.sty
Package: caption3 2022/03/17 v2.3b caption3 kernel (AR)
\caption@tempdima=\dimen193
\captionmargin=\dimen194
\caption@leftmargin=\dimen195
\caption@rightmargin=\dimen196
\caption@width=\dimen197
\caption@indent=\dimen198
\caption@parindent=\dimen199
\caption@hangindent=\dimen256
Package caption Info: Standard document class detected.
)
\c@caption@flags=\count307
\c@continuedfloat=\count308
Package caption Info: hyperref package is loaded.
Package caption Info: rotating package is loaded.
) (c:/TeXLive/2022/texmf-dist/tex/latex/natbib/natbib.sty
Package: natbib 2010/09/13 8.31b (PWD, AO)
\bibhang=\skip161
\bibsep=\skip162
LaTeX Info: Redefining \cite on input line 694.
\c@NAT@ctr=\count309
)) (c:/TeXLive/2022/texmf-dist/tex/latex/siunitx/siunitx.sty
Package: siunitx 2023-03-04 v3.2.2 A comprehensive (SI) units package
\l__siunitx_angle_tmp_dim=\dimen257
\l__siunitx_angle_marker_box=\box67
\l__siunitx_angle_unit_box=\box68
\l__siunitx_compound_count_int=\count310
(c:/TeXLive/2022/texmf-dist/tex/latex/translations/translations.sty
Package: translations 2022/02/05 v1.12 internationalization of LaTeX2e
packages
 (CN)
)
\l__siunitx_number_exponent_fixed_int=\count311
\l__siunitx_number_min_decimal_int=\count312
\l__siunitx_number_min_integer_int=\count313
\l__siunitx_number_round_precision_int=\count314
\l__siunitx_number_lower_threshold_int=\count315
\l__siunitx_number_upper_threshold_int=\count316
\l__siunitx_number_group_first_int=\count317
\l__siunitx_number_group_size_int=\count318
\l__siunitx_number_group_minimum_int=\count319
(c:/TeXLive/2022/texmf-dist/tex/latex/amsmath/amstext.sty
Package: amstext 2021/08/26 v2.01 AMS text
```

```
(c:/TeXLive/2022/texmf-dist/tex/latex/amsmath/amsgen.sty
File: amsgen.sty 1999/11/30 v2.0 generic functions
\@emptytoks=\toks38
\ex@=\dimen258
))
\l__siunitx_table_tmp_box=\box69
\l__siunitx_table_tmp_dim=\dimen259
\l__siunitx_table_column_width_dim=\dimen260
\l__siunitx_table_integer_box=\box70
\l__siunitx_table_decimal_box=\box71
\l__siunitx_table_uncert_box=\box72
\l__siunitx_table_before_box=\box73
\l__siunitx_table_after_box=\box74
\l__siunitx_table_before_dim=\dimen261
\l__siunitx_table_carry_dim=\dimen262
\l__siunitx_unit_tmp_int=\count320
\l__siunitx_unit_position_int=\count321
\l__siunitx_unit_total_int=\count322
) (c:/TeXLive/2022/texmf-dist/tex/latex/placeins/placeins.sty
Package: placeins 2005/04/18  v 2.2
) (c:/TeXLive/2022/texmf-dist/tex/latex/lineno/lineno.sty
Package: lineno 2023/01/19 line numbers on paragraphs v5.1
\linenopenalty=\count323
\output=\toks39
\linenoprevgraf=\count324
\linenumbersep=\dimen263
\linenumberwidth=\dimen264
\c@linenumber=\count325
\c@pagewiselinenumber=\count326
\c@LN@truepage=\count327
\c@internallinenumber=\count328
\c@internallinenumbers=\count329
\quotelinenumbersep=\dimen265
\bframerule=\dimen266
\bframesep=\dimen267
\bframebox=\box75
LaTeX Info: Redefining \\ on input line 3131.
)
Package translations Info: No language package found. I am going to use
`englis
h' as default language. on input line 66.
LaTeX Font Info:    Trying to load font information for T1+Merriwthr-OsF
on inp
ut line 66.
(c:/TeXLive/2022/texmf-dist/tex/latex/merriweather/T1Merriwthr-OsF.fd
File: T1Merriwthr-OsF.fd 2020/08/30 (autoinst) Font definitions for
T1/Merriwth
r-OsF.
)
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 7.5pt on input line 66.
(./main.aux)
\openout1 = `main.aux'.
```

```
LaTeX Font Info:    Checking defaults for OML/cmm/m/it on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for OMS/cmsy/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for OT1/cmr/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for T1/cmr/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for TS1/cmr/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for OMX/cmex/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for U/cmr/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for PD1/pdf/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Font Info:    Checking defaults for PU/pdf/m/n on input line 66.
LaTeX Font Info:    ... okay on input line 66.
LaTeX Info: Redefining \microtypecontext on input line 66.
Package microtype Info: Applying patch `item' on input line 66.
Package microtype Info: Applying patch `toc' on input line 66.
Package microtype Info: Applying patch `eqnum' on input line 66.

Package microtype Warning: Unable to apply patch `footnote' on input line
66.

Package microtype Info: Applying patch `verbatim' on input line 66.
Package microtype Info: Generating PDF output.
Package microtype Info: Character protrusion enabled (level 2).
Package microtype Info: Using default protrusion set `alltext'.
Package microtype Info: Automatic font expansion enabled (level 2),
(microtype)             stretch: 20, shrink: 20, step: 1, non-selected.
Package microtype Info: Using default expansion set `alltext-nott'.
LaTeX Info: Redefining \showhyphens on input line 66.
Package microtype Info: No adjustment of tracking.
Package microtype Info: No adjustment of interword spacing.
Package microtype Info: No adjustment of character kerning.
Package microtype Info: Loading generic protrusion settings for font
family
(microtype)             `Merriwthr-OsF' (encoding: T1).
(microtype)             For optimal results, create family-specific
settings.
(microtype)             See the microtype manual for details.
LaTeX Font Info:    Redeclaring symbol font `operators' on input line 66.
LaTeX Font Info:    Encoding `OT1' has changed to `T1' for symbol font
(Font)              `operators' in the math version `normal' on input
line 66.
LaTeX Font Info:    Overwriting symbol font `operators' in version
`normal'
(Font)                  OT1/cmr/m/n --> T1/Merriwthr-OsF/m/up on input
line 66.

LaTeX Font Info:    Encoding `OT1' has changed to `T1' for symbol font
```

```
(Font)              `operators' in the math version `bold' on input line
66.
LaTeX Font Info:    Overwriting symbol font `operators' in version `bold'
(Font)                  OT1/cmr/bx/n --> T1/Merriwthr-OsF/m/up on input
line 66
.
LaTeX Font Info:    Overwriting symbol font `operators' in version `bold'
(Font)                  T1/Merriwthr-OsF/m/up --> T1/Merriwthr-OsF/b/up
on inpu
t line 66.
LaTeX Font Info:    Redeclaring math alphabet \mathbf on input line 66.
LaTeX Font Info:    Overwriting math alphabet `\mathbf' in version
`normal'
(Font)                  OT1/cmr/bx/n --> T1/Merriwthr-OsF/b/up on input
line 66
.
LaTeX Font Info:    Overwriting math alphabet `\mathbf' in version `bold'
(Font)                  OT1/cmr/bx/n --> T1/Merriwthr-OsF/b/up on input
line 66
.
LaTeX Font Info:    Redeclaring math alphabet \mathsf on input line 66.
LaTeX Font Info:    Overwriting math alphabet `\mathsf' in version
`normal'
(Font)                  OT1/cmss/m/n --> T1/MerriwthrSans-OsF/m/up on
input lin
e 66.
LaTeX Font Info:    Overwriting math alphabet `\mathsf' in version `bold'
(Font)                  OT1/cmss/bx/n --> T1/MerriwthrSans-OsF/m/up on
input li
ne 66.
LaTeX Font Info:    Redeclaring math alphabet \mathit on input line 66.
LaTeX Font Info:    Overwriting math alphabet `\mathit' in version
`normal'
(Font)                  OT1/cmr/m/it --> T1/Merriwthr-OsF/m/it on input
line 66
.
LaTeX Font Info:    Overwriting math alphabet `\mathit' in version `bold'
(Font)                  OT1/cmr/bx/it --> T1/Merriwthr-OsF/m/it on input
line 6
6.
LaTeX Font Info:    Redeclaring math alphabet \mathtt on input line 66.
LaTeX Font Info:    Overwriting math alphabet `\mathtt' in version
`normal'
(Font)                  OT1/cmtt/m/n --> T1/lmtt/m/up on input line 66.
LaTeX Font Info:    Overwriting math alphabet `\mathtt' in version `bold'
(Font)                  OT1/cmtt/m/n --> T1/lmtt/m/up on input line 66.
LaTeX Font Info:    Overwriting math alphabet `\mathsf' in version `bold'
(Font)                  T1/MerriwthrSans-OsF/m/up --> T1/MerriwthrSans-
OsF/b/up
 on input line 66.
LaTeX Font Info:    Overwriting math alphabet `\mathit' in version `bold'
(Font)                  T1/Merriwthr-OsF/m/it --> T1/Merriwthr-OsF/b/it
on inpu
t line 66.
```

```
\c@mv@tabular=\count330
\c@mv@boldtabular=\count331
(c:/TeXLive/2022/texmf-dist/tex/context/base/mkii/supp-pdf.mkii
[Loading MPS to PDF converter (version 2006.09.02).]
\scratchcounter=\count332
\scratchdimen=\dimen268
\scratchbox=\box76
\nofMPsegments=\count333
\nofMParguments=\count334
\everyMPshowfont=\toks40
\MPscratchCnt=\count335
\MPscratchDim=\dimen269
\MPnumerator=\count336
\makeMPintoPDFobject=\count337
\everyMPtoPDFconversion=\toks41
) (c:/TeXLive/2022/texmf-dist/tex/latex/epstopdf-pkg/epstopdf-base.sty
Package: epstopdf-base 2020-01-24 v2.11 Base part for package epstopdf
Package epstopdf-base Info: Redefining graphics rule for `.eps' on input
line 4
85.
(c:/TeXLive/2022/texmf-dist/tex/latex/latexconfig/epstopdf-sys.cfg
File: epstopdf-sys.cfg 2010/07/13 v1.3 Configuration of (r)epstopdf for
TeX Liv
e
))
*geometry* driver: auto-detecting
*geometry* detected driver: pdftex
*geometry* verbose mode - [ preamble ] result:
* driver: pdftex
* paper: a4paper
* layout: <same size as paper>
* layoutoffset:(h,v)=(0.0pt,0.0pt)
* modes: includefoot twoside
* h-part:(L,W,R)=(54.64pt, 488.22787pt, 54.64pt)
* v-part:(T,H,B)=(66.0pt, 745.04684pt, 34.0pt)
* \paperwidth=597.50787pt
* \paperheight=845.04684pt
* \textwidth=488.22787pt
* \textheight=715.04684pt
* \oddsidemargin=-17.62999pt
* \evensidemargin=-17.62999pt
* \topmargin=-47.76999pt
* \headheight=17.5pt
* \headsep=24.0pt
* \topskip=10.0pt
* \footskip=30.0pt
* \marginparwidth=48.0pt
* \marginparsep=10.0pt
* \columnsep=18.0pt
* \skip\footins=22.0pt plus 2.0pt
* \hoffset=0.0pt
* \voffset=0.0pt
* \mag=1000
* \@twocolumntrue
```

```
* \@twosidetrue
* \@mparswitchtrue
* \@reversemarginfalse
* (1in=72.27pt=25.4mm, 1cm=28.453pt)

Package hyperref Info: Link coloring ON on input line 66.
(./main.out) (./main.out)
\@outlinefile=\write3
\openout3 = `main.out'.

\@gscitedetails=\box77
\@gscitedetailsheight=\skip163
\@gsheadbox=\box78
\@gsheadboxheight=\skip164
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 6.5pt on input line 66.
LaTeX Font Info:    Calculating math sizes for size <7.5> on input line
66.

LaTeX Font Warning: Font shape `T1/Merriwthr-OsF/m/up' undefined
(Font)              using `T1/Merriwthr-OsF/m/n' instead on input line
66.

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 6.24973pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 5.24997pt on input line 66.
LaTeX Font Info:    Trying to load font information for U+eur on input
line 66.

(c:/TeXLive/2022/texmf-dist/tex/latex/amsfonts/ueur.fd
File: ueur.fd 2013/01/14 v3.01 Euler Roman
) (c:/TeXLive/2022/texmf-dist/tex/latex/microtype/mt-eur.cfg
File: mt-eur.cfg 2006/07/31 v1.1 microtype config. file: AMS Euler Roman
(RS)
)

LaTeX Font Warning: Font shape `OMS/cmsy/m/n' in size <7.5> not available
(Font)              size <7> substituted on input line 66.

LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <7.5> on input line 66.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <6.24973> on input line 66.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <5.24997> on input line 66.
LaTeX Font Info:    Trying to load font information for U+euf on input
line 66.

(c:/TeXLive/2022/texmf-dist/tex/latex/amsfonts/ueuf.fd
File: ueuf.fd 2013/01/14 v3.01 Euler Fraktur
) (c:/TeXLive/2022/texmf-dist/tex/latex/microtype/mt-euf.cfg
File: mt-euf.cfg 2006/07/03 v1.1 microtype config. file: AMS Euler
Fraktur (RS)
```

```
)
LaTeX Font Info:    Trying to load font information for U+eus on input
line 66.

(c:/TeXLive/2022/texmf-dist/tex/latex/amsfonts/ueus.fd
File: ueus.fd 2013/01/14 v3.01 Euler Script
) (c:/TeXLive/2022/texmf-dist/tex/latex/microtype/mt-eus.cfg
File: mt-eus.cfg 2006/07/28 v1.2 microtype config. file: AMS Euler Script
(RS)
)
LaTeX Font Info:    Trying to load font information for U+euex on input
line 66
.
(c:/TeXLive/2022/texmf-dist/tex/latex/amsfonts/ueuex.fd
File: ueuex.fd 2013/01/14 v3.01 Euler extra symbols
)

LaTeX Font Warning: Font shape `OML/cmm/m/it' in size <7.5> not available
(Font)              size <7> substituted on input line 66.

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 6.24973pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 5.24997pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 7.5pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 6.24973pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 5.24997pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 8.0pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 8.0pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/it' will be
(Font)              scaled to size 8.0pt on input line 66.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 8.0pt on input line 66.

LaTeX Warning: Reference `LastPage' on page 1 undefined on input line 66.

Package caption Info: Begin \AtBeginDocument code.
Package caption Info: End \AtBeginDocument code.

(c:/TeXLive/2022/texmf-dist/tex/latex/translations/translations-basic-
dictionar
y-english.trsl
File: translations-basic-dictionary-english.trsl (english translation
file `tra
nslations-basic-dictionary')
)
Package translations Info: loading dictionary `translations-basic-
dictionary' f
```

```
or `english'. on input line 66.
TextBlockOrigin set to 4pc+6.64pt x 4pc+6pt
<oup.pdf, id=144, 49.18375pt x 48.18pt>
File: oup.pdf Graphic file (type pdf)
<use oup.pdf>
Package pdftex.def Info: oup.pdf  used on input line 84.
(pdftex.def)             Requested size: 59.98604pt x 58.762pt.
<gigascience-logo.pdf, id=145, 99.37125pt x 33.12375pt>
File: gigascience-logo.pdf Graphic file (type pdf)
<use gigascience-logo.pdf>
Package pdftex.def Info: gigascience-logo.pdf  used on input line 84.
(pdftex.def)             Requested size: 126.00902pt x 42.0pt.

Overfull \hbox (54.64pt too wide) in paragraph at lines 84--84
[][]
 []

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 14.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 8.99997pt on input line 84.
LaTeX Font Info:    Calculating math sizes for size <14> on input line
84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 14.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 11.66617pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 9.79996pt on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <14> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <11.66617> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <9.79996> on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 11.66617pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 9.79996pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 14.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 11.66617pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 9.79996pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 18.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 13.0pt on input line 84.
LaTeX Font Info:    Calculating math sizes for size <13> on input line
84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 13.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
```

```
(Font)              scaled to size 10.83287pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 9.09996pt on input line 84.

LaTeX Font Warning: Font shape `OMS/cmsy/m/n' in size <13> not available
(Font)              size <12> substituted on input line 84.

LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <13> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <10.83287> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <9.09996> on input line 84.

LaTeX Font Warning: Font shape `OML/cmm/m/it' in size <13> not available
(Font)              size <12> substituted on input line 84.

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 10.83287pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 9.09996pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 13.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 10.83287pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 9.09996pt on input line 84.
LaTeX Font Info:    Trying to load font information for TS1+Merriwthr-OsF
on in
put line 84.
(c:/TeXLive/2022/texmf-dist/tex/latex/merriweather/TS1Merriwthr-OsF.fd
File: TS1Merriwthr-OsF.fd 2020/08/30 (autoinst) Font definitions for
TS1/Merriw
thr-OsF.
)
LaTeX Font Info:    Font shape `TS1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 10.83287pt on input line 84.
Package microtype Info: Loading generic protrusion settings for font
family
(microtype)             `Merriwthr-OsF' (encoding: TS1).
(microtype)             For optimal results, create family-specific
settings.
(microtype)             See the microtype manual for details.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 9.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 9.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 7.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 5.0pt on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <9> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
```

```
(Font)              <7> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <5> on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 7.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 5.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 9.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 7.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 5.0pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 6.5pt on input line 84.
LaTeX Font Info:    Calculating math sizes for size <6.5> on input line
84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 6.5pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 5.41643pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 4.54997pt on input line 84.

LaTeX Font Warning: Font shape `OMS/cmsy/m/n' in size <6.5> not available
(Font)              size <6> substituted on input line 84.


LaTeX Font Warning: Font shape `OMS/cmsy/m/n' in size <5.41643> not
available
(Font)              size <5> substituted on input line 84.


LaTeX Font Warning: Font shape `OMS/cmsy/m/n' in size <4.54997> not
available
(Font)              size <5> substituted on input line 84.

LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <6.5> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <5.41643> on input line 84.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <4.54997> on input line 84.

LaTeX Font Warning: Font shape `OML/cmm/m/it' in size <6.5> not available
(Font)              size <6> substituted on input line 84.


LaTeX Font Warning: Font shape `OML/cmm/m/it' in size <5.41643> not
available
(Font)              size <5> substituted on input line 84.
```

```
LaTeX Font Warning: Font shape `OML/cmm/m/it' in size <4.54997> not
available
(Font)              size <5> substituted on input line 84.

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 5.41643pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 4.54997pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 6.5pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 5.41643pt on input line 84.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 4.54997pt on input line 84.
LaTeX Font Info:    Font shape `TS1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 5.41643pt on input line 84.

Overfull \hbox (54.64pt too wide) in paragraph at lines 84--84
[][][]
 []

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 10.0pt on input line 84.

Overfull \hbox (54.64pt too wide) in paragraph at lines 84--84
[][][]
 []

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 7.8pt on input line 95.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 7.8pt on input line 95.
[1{c:/TeXLive/2022/texmf-var/fonts/map/pdftex/updmap/pdftex.map}



 <./oup.pdf> <./gigascience-logo.pdf>]
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 10.0pt on input line 97.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 3.75pt on input line 97.
LaTeX Font Info:    Trying to load font information for T1+MerriwthrSans-
OsF on
 input line 97.
(c:/TeXLive/2022/texmf-dist/tex/latex/merriweather/T1MerriwthrSans-OsF.fd
File: T1MerriwthrSans-OsF.fd 2020/08/30 (autoinst) Font definitions for
T1/Merr
iwthrSans-OsF.
)
LaTeX Font Info:    Font shape `T1/MerriwthrSans-OsF/m/n' will be
(Font)              scaled to size 3.75pt on input line 97.
Package microtype Info: Loading generic protrusion settings for font
family
(microtype)             `MerriwthrSans-OsF' (encoding: T1).
```

```
(microtype)             For optimal results, create family-specific
settings.
(microtype)             See the microtype manual for details.

Underfull \hbox (badness 1527) in paragraph at lines 100--101
\T1/Merriwthr-OsF/m/n/7.5 (+20) cases of COVID-19
([][]covid19.who.int[][], ac-
cessed April 19, 2023)
 []

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 7.8pt on input line 119.
[2

]
<Figure1_overview.pdf, id=265, 1800.35355pt x 1579.44817pt>
File: Figure1_overview.pdf Graphic file (type pdf)
<use Figure1_overview.pdf>
Package pdftex.def Info: Figure1_overview.pdf  used on input line 124.
(pdftex.def)             Requested size: 439.4021pt x 385.48602pt.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 6.0pt on input line 125.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 6.0pt on input line 125.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/it' will be
(Font)              scaled to size 6.0pt on input line 125.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 7.0pt on input line 131.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/it' will be
(Font)              scaled to size 7.0pt on input line 136.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 5.00003pt on input line 139.

Overfull \hbox (1.13507pt too wide) in paragraph at lines 134--169
[][]
 []

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 8.5pt on input line 175.

Underfull \vbox (badness 10000) has occurred while \output is active []

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 8.5pt on input line 175.
[3 <./Figure1_overview.pdf>]
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 6.0pt on input line 205.
LaTeX Font Info:    External font `cmex10' loaded for size
(Font)              <6> on input line 205.

Underfull \vbox (badness 3735) has occurred while \output is active []

LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/n' will be
(Font)              scaled to size 7.5pt on input line 215.
```

```
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/sl' in size <7.5> not
availa
ble
(Font)              Font shape `T1/Merriwthr-OsF/b/it' tried instead on
input l
ine 215.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/it' will be
(Font)              scaled to size 7.5pt on input line 215.
LaTeX Font Info:    Trying to load font information for T1+lmtt on input
line 2
19.
(c:/TeXLive/2022/texmf-dist/tex/latex/lm/t1lmtt.fd
File: t1lmtt.fd 2015/05/01 v1.6.1 Font defs for Latin Modern
)
Package microtype Info: Loading generic protrusion settings for font
family
(microtype)             `lmtt' (encoding: T1).
(microtype)             For optimal results, create family-specific
settings.
(microtype)             See the microtype manual for details.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/m/up' will be
(Font)              scaled to size 7.5pt on input line 219.
LaTeX Font Info:    Font shape `T1/Merriwthr-OsF/b/it' will be
(Font)              scaled to size 8.5pt on input line 226.
[4]
<Figure2_standards.png, id=357, 604.21346pt x 456.50172pt>
File: Figure2_standards.png Graphic file (type png)
<use Figure2_standards.png>
Package pdftex.def Info: Figure2_standards.png  used on input line 241.
(pdftex.def)             Requested size: 483.34827pt x 365.1949pt.

LaTeX Warning: Reference `sfig:paneuger-background' on page 5 undefined
on inpu
t line 250.


LaTeX Warning: Reference `sfig:paneuger-background' on page 5 undefined
on inpu
t line 252.

<Figure3_paneuger.pdf, id=366, 722.7pt x 1084.05pt>
File: Figure3_paneuger.pdf Graphic file (type pdf)
<use Figure3_paneuger.pdf>
Package pdftex.def Info: Figure3_paneuger.pdf  used on input line 260.
(pdftex.def)             Requested size: 390.58379pt x 585.87453pt.

LaTeX Warning: Reference `sfig:ffmairport-background' on page 5 undefined
on in
put line 267.

<Figure4_ffm-airport.pdf, id=373, 1411.52344pt x 699.36281pt>
File: Figure4_ffm-airport.pdf Graphic file (type pdf)
<use Figure4_ffm-airport.pdf>
Package pdftex.def Info: Figure4_ffm-airport.pdf  used on input line 277.
```

```
(pdftex.def)             Requested size: 463.81499pt x 229.79799pt.

Underfull \vbox (badness 2932) has occurred while \output is active []

[5] [6 <./Figure2_standards.png>] [7 <./Figure3_paneuger.pdf>]

LaTeX Warning: Reference `stab:min-ref' on page 8 undefined on input line
289.


LaTeX Warning: Reference `sfig:Pan-EU-GER-aaf' on page 8 undefined on
input lin
e 297.


LaTeX Warning: Reference `sfig:omicron-aaf-sub-lineages' on page 8
undefined on
 input line 297.

<Figure5_aaf-standards.pdf, id=399, 1084.05pt x 1084.05pt>
File: Figure5_aaf-standards.pdf Graphic file (type pdf)
<use Figure5_aaf-standards.pdf>
Package pdftex.def Info: Figure5_aaf-standards.pdf  used on input line
301.
(pdftex.def)             Requested size: 488.22787pt x 488.23152pt.

Underfull \vbox (badness 10000) has occurred while \output is active []

[8 <./Figure4_ffm-airport.pdf>] [9 <./Figure5_aaf-standards.pdf>]

LaTeX Warning: Reference `sfig:freyja-full' on page 10 undefined on input
line
320.


LaTeX Warning: Reference `sfig:freyja-reduced' on page 10 undefined on
input li
ne 320.

[10]
Underfull \vbox (badness 10000) has occurred while \output is active []

LaTeX Font Info:    Font shape `TS1/Merriwthr-OsF/m/n' will be
(Font)              scaled to size 7.5pt on input line 369.

LaTeX Warning: Reference `stab:mapping' on page 11 undefined on input
line 379.


[11]
Underfull \vbox (badness 10000) has occurred while \output is active []

[12] [13] (./main.bbl
Underfull \hbox (badness 1596) in paragraph at lines 106--110
```

```
[]\T1/Merriwthr-OsF/m/up/7.5 (+20) Hoar C, McClary-Gutierrez J, Wolfe MK,
Bivin
s A, Bibby
 []


Underfull \hbox (badness 1975) in paragraph at lines 106--110
\T1/Merriwthr-OsF/m/up/7.5 (+20) K, Sil-ver-man AI, et al.  Look-ing For-
ward:
The Role of
 []


Underfull \hbox (badness 1342) in paragraph at lines 149--153
\T1/Merriwthr-OsF/m/up/7.5 (+20) ham A, Jepsen K, et al.  Wastew-a-ter
se-quenc
-ing re-veals
 []


Underfull \hbox (badness 1881) in paragraph at lines 149--153
\T1/Merriwthr-OsF/m/up/7.5 (+20) early cryp-tic SARS-CoV-2 vari-ant
trans-mis-s
ion.  Na-ture
 []


Underfull \vbox (badness 10000) has occurred while \output is active []

[14]) [15

]

! LaTeX Error: File `supplement.tex' not found.

Type X to quit or <RETURN> to proceed,
or enter new name. (Default extension: tex)

Enter file name:
! Emergency stop.
<read *>

l.512 \input{supplement}
                         ^^M
*** (cannot \read from terminal in nonstop modes)


Here is how much of TeX's memory you used:
 24679 strings out of 476024
 485921 string characters out of 5794017
 1918382 words of memory out of 5000000
 43829 multiletter control sequences out of 15000+600000
 1943143 words of font info for 674 fonts, out of 8000000 for 9000
 1141 hyphenation exceptions out of 8191
```

```
 123i,12n,131p,2473b,1127s stack positions out of
10000i,1000n,20000p,200000b,200000s
!  ==> Fatal error occurred, no output PDF file produced!
```

OXFORD (GIGA)n SCIENCE

GigaScience, 2017, 1–24

doi: xx.xxxx/xxxx
Manuscript in Preparation
Research

RESEARCH

Impact of reference design on estimating SARS-CoV-2 lineage abundances from wastewater sequencing data

Eva Aßmann[1, 4, †], Shelesh Agrawal[2,†], Laura Orschler[2], Sindy Böttcher[3], Susanne Lackner[2] and Martin Hölzer[1,*]

[1]Genome Competence Center (MF1), Robert Koch Institute, Berlin, Germany and [2]Chair of Water and Environmental Biotechnology, Institute IWAR, Department of Civil and Environmental Engineering Sciences, Technical University of Darmstadt, Darmstadt, Germany and [3]Gastroenteritis and Hepatitis Pathogens and Enteroviruses, Robert Koch Institute, Berlin, Germany and [4]Center for Artificial Intelligence in Public Health Research (ZKI-PH), Robert Koch Institute, Berlin, Germany

[*]HoelzerM@rki.de
[†]Contributed equally.

Abstract

Background Sequencing of SARS-CoV-2 RNA from wastewater samples has emerged as a valuable tool for detecting the presence and relative abundances of SARS-CoV-2 variants in a community. By analyzing the viral genetic material present in wastewater, public health officials can gain early insights into the spread of the virus and inform timely intervention measures. The construction of reference datasets from known SARS-CoV-2 lineages and their mutation profiles has become state-of-the-art for assigning viral lineages and their relative abundances from wastewater sequencing data. However, the selection of reference sequences or mutations directly affects the predictive power. **Results** Here, we show the impact of a *mutation-* and *sequence-based* reference reconstruction for SARS-CoV-2 abundance estimation. We benchmark three data sets: 1) synthetic "spike-in" mixtures, 2) German samples from early 2021, mainly comprising Alpha, and 3) samples obtained from wastewater at an international airport in Germany from the end of 2021, including first signals of Omicron. The two approaches differ in sub-lineage detection, with the marker-*mutation-based* method, in particular, being challenged by the increasing number of mutations and lineages. However, the estimations of both approaches depend on selecting representative references and optimized parameter settings. By performing parameter escalation experiments, we demonstrate the effects of reference size and alternative allele frequency cutoffs for abundance estimation. We show how different parameter settings can lead to different results for our test data sets, and illustrate the effects of virus lineage composition of wastewater samples and references. **Conclusions** Here, we compare a *mutation-* and *sequence-based* reference construction and assignment for SARS-CoV-2 abundance estimation from wastewater samples. Our study highlights current computational challenges, focusing on the general reference design, which significantly and directly impacts abundance allocations. We illustrate advantages and disadvantages that may be relevant for further developments in the wastewater community and in the context of higher standardization.

Key words: SARS-CoV-2; wastewater; sewage; abundance estimation, next-generation sequencing, benchmark

Compiled on: June 12, 2023.
Draft manuscript prepared by the author.

Background

Coronavirus disease 2019 (COVID-19), the highly contagious viral illness caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is the most consequential global health crisis since the era of the influenza pandemic of 1918. Since its discovery, SARS-CoV-2 has caused >763 million confirmed cases of COVID-19 (covid19.who.int, accessed April 19, 2023) and currently >3,000 SARS-CoV-2 lineages are defined by the *Pango* network [1, 2] (https://github.com/cov-lineages/lineages-website/blob/master/_data/lineage_data.full.json, accessed April 19, 2023). Genome sequencing has played a central role during the COVID-19 pandemic in supporting public health agencies, monitoring emerging mutations in the SARS-CoV-2 genome, and advancing precision vaccinology and optimizing molecular tests [3, 4, 5]. Massive sequencing of clinical samples has made it possible to monitor emerging variants, emphasizing temporal and spatial variation. With ongoing transmission, further mutations occur in the genome that are part of the viral evolutionary process and result in unique fingerprints.

Sequencing capacity, however, is limited, cannot be sustained over the long term for so many clinical samples, and only allows extrapolation based on a relatively small fraction of all infections occurring during the pandemic. In addition, with decreasing incidence numbers, sampling and sequencing efforts are decreasing, raising the need for representative, medium-scale, and sustainable surveillance systems [5] or other approaches. From January 1, 2020 until April 19, 2023, 931,260 genome sequences of COVID-19-positive clinical samples from Germany have been uploaded to the international GISAID platform [6], representing a proportion of 2.426 % out of a total of 38,388,247 reported SARS-CoV-2 cases in Germany (COVID-19 Dashboard Germany, accessed April 19, 2023). In Germany and other countries, complete detection and sequencing of all positive cases were impossible due to the high infection numbers. However, wastewater-based epidemiology (WBE) has shown the potential to get a much broader snapshot of the SARS-CoV-2 variant circulation at a community level [7, 8, 9, 10, 11, 12]. Integrating genome sequencing with WBE can provide information on circulating SARS-CoV-2 variants in a region [13]. The sequencing methods commonly used in WBE are similar to the ones used for clinical samples, using a general strategy that employs the sequencing of the whole genome via amplification of small, specific regions of the SARS-CoV-2 genome, i.e., targeted sequencing of amplicons via pre-defined primer sequences [14, 15, 8, 16, 11]. Targeted sequencing can achieve a high degree of coverage of informative regions of the genome and, most importantly, reveal to some extent which polymorphisms are linked, making it possible to track SARS-CoV-2 variants of concern (VOCs) and other variants.

A particular challenge in performing sequencing of SARS-CoV-2 from wastewater samples concerns the viral RNA present in many individual fragments rather than complete viral genomes. In addition, these fragments come from the excretions of many infected individuals, making it challenging, if not impossible, to reconstruct individual genomes using bioinformatic approaches like the ones developed for clinical samples of individual patients. In need of computational approaches to analyze mixed wastewater samples, several groups developed similar tools for quality control, sequencing data analysis, and SARS-CoV-2 lineage abundance estimation [17, 18, 19, 20, 13, 21, 22, 14, 7, 23, 24, 25, 26, 27], see Table 1. Most approaches focus on detecting pre-defined characteristic marker mutations in the sequenced reads and utilize this information for abundance estimation. Common to all these tools is that they require a reference set of either signature marker mutations (hereafter called *mutation-based*) or complete genome sequences (hereafter called *sequence-based*) from which characteristic mutation profiles or kmers (short sub-sequences of length k) are derived. Kayikcioglu *et al.* compared the performance of five selected approaches for SARS-CoV-2 lineage abundance estimation on simulated and publicly available mixed population samples [27]. They found that Kallisto [28], as first suggested by Baaijens, Zulli, and Ott *et al.* [25], followed by Freyja [17], achieved most accurate estimations.

In a *mutation-based* approach, to estimate the proportion of specific SARS-CoV-2 variants present in a mixed sample, mutations or combinations of mutations characteristic or unique for these variants based on clinical samples can be compared with the mutations detectable in the sample. In principle, and as implemented in a previously used approach [16] (which we refer to here as MAMUSS, Table 1), the occurrence of mutations can be represented by the value of the relative abundance of a VOC or other viral variant. First, the frequency of occurrence of each mutation is calculated from the multiplication of the reads and the allele frequency. The relative abundance describes the percentage ratio of the sum of the read abundance of the characteristic mutations of a SARS-CoV-2 virus variant and the sum of the read abundance of all mutations found in a sample. Accordingly, only the previously selected virus variants and signature mutations that form the reference set are evaluated and others that may occur in the sample are ignored. Another prominent *mutation-based* approach is implemented in the tool Freyja [17]. Freyja solves the de-mixing problem to recover relative lineage abundances from mixed SARS-CoV-2 samples using lineage-determining mutational "barcodes" derived from the UShER global phylogenetic tree [29]. Using mutation abundances and sequencing depth measurements at each position in the genome, Freyja estimates the abundance of lineages in the sample.

As a different methodological approach to reconstruct a reference, the full genome sequence information can be used to automatically select appropriate features (e.g., signature mutations, kmers) and to use them to evaluate the proportions of SARS-CoV-2 variants in wastewater samples instead of a preselected set of marker mutations (*sequence-based*) [25, 23, 24] Table 1. Again, information derived from sequencing of clinical samples and their lineage annotation are used to generate a representative reference data set that can be then searched via established (pseudo)-alignment methods such as Kallisto [28] as suggested by Baaijens, Zulli, and Ott *et al.* in their VQL tool [25].

In this study, we specifically investigated the impact of reference composition and construction on assigning relative abundances of SARS-CoV-2 lineages from wastewater sequencing data. As mentioned, various tools have been developed over the course of the pandemic (Table 1) and they all have different facets in calculating relative abundances. However, they all have in common that some reference data set needs to be defined, which derives information on lineages and mutations from existing genomics data. We specifically choose two different approaches representing a *mutation-based* method (MAMUSS) and *sequence-based* method (VLQ). The two approaches distinguish mainly by the input data set used for the reference set design and subsequent lineage assignment (Figure 1). Either a selection of lineage-defining marker mutations (*mutation-based*) or full SARS-CoV-2 genome sequences (*sequence-based*) are used to reconstruct a reference base for lineage assignment and abundance estimation. Here, we compare exemplary implementations of both general approaches. MAMUSS, as previously applied in [16], implements a representative basic workflow for the *mutation-based* approach focusing on unique marker mutations. For the *sequence-based*

Table 1. Collection of tools available for sequencing data analysis in WBE and SARS-CoV-2 lineage proportion estimation. We distinguish the tools roughly based on their approach to define a reference set into those using predefined marker mutations and those relying on full genome sequences or both. The two implementations we selected for reference construction and our comparison are indicated in bold. Please note that C-WAP [27] wraps multiple approaches while also including a new *mutation-based* tool, LINDEC.

mutation-based		
Tool	Citation	Code
MAMUSS	[16]	github.com/lifehashopes/MAMUSS
Freyja	[17]	github.com/andersen-lab/Freyja
Lineagespot	[18]	github.com/nikopech/lineagespot
LCS	[19]	github.com/rvalieris/LCS
Alcov	[20]	github.com/Ellmen/alcov
VaQuERo	[13]	github.com/fabou-uobaf/VaQuERo
MMMVI	[21]	github.com/dorbarker/voc-identify
PiGx	[22]	github.com/BIMSBbioinfo/pigx_sars_cov-2
SAMRefiner	[14]	github.com/degregory/SAM_Refiner
COJAC	[7]	github.com/cbg-ethz/cojac
wastewaterSPAdes	[26]	–
gromstole	–	github.com/PoonLab/gromstole
CovMix	–	github.com/chrisquince/covmix
sequence-based		
Tool	Citation	Code
VLQ-nf	this study	github.com/rki-mf1/VLQ-nf
VLQ	[25]	github.com/baymlab/wastewater_analysis
VirPool	[23]	github.com/fmfi-compbio/virpool
V-pipe SC2	[24]	cbg-ethz.github.io/V-pipe/sars-cov-2
mutation-based & *sequence-based*		
Tool	Citation	Code
C-WAP	[27]	github.com/CFSAN-Biostatistics/C-WAP

Table 2. Composition of synthetic mixture "spike-in" *Standards*. Here we show the proportions of which different SARS-CoV-2 lineages were mixed to generate a collection of artificial samples for our benchmark. For example, the sample Mix_01 comprises 25 % original Wuhan-Hu-1 A.1 and 75 % Alpha B.1.1.7 ($0.25_{org} - 0.75_{alpha}$). All samples were sequenced with Ion Torrent and raw data is available under BioProject number PRJNA912560 in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). Please note that no real wastewater was used to construct these synthetic mixtures because we wanted to reduce any side effects for our *gold standard*.

Sample ID	Composition
Mix_01	$0.25_{org} - 0.75_{alpha}$
Mix_02	$0.25_{org} - 0.25_{beta} - 0.5_{alpha}$
Mix_03	$0.25_{alpha} - 0.25_{beta} - 0.25_{gamma} - 0.25_{org}$
Mix_04	$0.5_{org} - 0.5_{iota}$
Mix_05	$0.25_{org} - 0.25_{iota} - 0.5_{omiba2}$
Mix_06	$0.25_{alpha} - 0.25_{iota} - 0.25_{omiBA1} - 0.25_{omiBA2}$
Mix_07	$0.5_{omiBA1} - 0.5_{omiBA2}$
Mix_08	$0.25_{org} - 0.25_{alpha} - 0.25_{omiBA1} - 0.25_{omiBA2}$
Mix_09	$0.5_{deltaAY1} - 0.5_{deltaAY2}$
Mix_10	$0.25_{deltaAY1} - 0.25_{deltaAY2} - 0.5_{delta}$
Mix_11	$0.25_{deltaAY1} - 0.25_{deltaAY2} - 0.5_{omiBA1}$
Mix_12	$0.25_{deltaAY1} - 0.25_{deltaAY2} - 0.25_{omiBA1} - 0.25_{omiBA2}$
Mix_13	$0.25_{deltaAY1} - 0.25_{deltaAY2} - 0.25_{omiBA1} - 0.25_{omiBA2}$
Mix_14	$0.5_{delta} - 0.25_{omiBA1} - 0.25_{omiBA2}$
Mix_15	$0.25_{deltaAY1} - 0.25_{deltaAY2} - 0.25_{omiBA1} - 0.25_{omiBA2}$
Mix_16	$0.25_{alpha} - 0.25_{delta} - 0.25_{omiBA1} - 0.25_{omiBA2}$

$_{org}$ – Wuhan-Hu-1 A.1; $_{alpha}$ – Alpha B.1.1.7; $_{beta}$ – Beta B.1.351; $_{gamma}$ – Gamma P.1; $_{iota}$ – Iota B.1.526; $_{delta}$ – Delta B.1.617.2; $_{deltaAY1}$ – Delta AY.1; $_{deltaAY2}$ – Delta AY.2; $_{omiBA1}$ – Omicron BA.1; $_{omiBA2}$ – Omicron BA.2

approach, we use pseudo-alignments via Kallisto [28] as proposed initially by [25] and their VQL tool. Based on their idea and scripts, we implemented a slightly modified version of VLQ in a Nextflow [30] pipeline that we call VLQ-nf. We chose our *sequence-based* method to be based on VLQ because of it reusing Kallisto as an established tool in transcript quantification [28]. A major benefit of implementing the representative methods was the complete control over code, parameters, and inputs, which allowed us to understand better, compare, and interpret the results of our benchmark study and the effects on the reference design.

We tested both MAMUSS as a *mutation-based* reference representative and VLQ-nf as a *sequence-based* reference representative on three data sets: 1) a synthetic scenario of "spike-in" mixture samples, 2) samples from Germany from a European wastewater study from early 2021, mainly comprising the VOC Alpha [9], and 3) a sample obtained from wastewater sequencing at the international airport in Frankfurt am Main, Germany from the end of 2021, including first signals of the VOC Omicron [16].

We show that both the *mutation-based* and *sequence-based* approach can reflect the proportions of SARS-CoV-2 lineages in the different samples but also comprise differences in resolution and the detection of similar sub-lineages depending on the reference set. Both approaches also show advantages and disadvantages when it comes to the selection of signature marker mutations and genome sequences, respectively. For the *mutation-based* approach as implemented in MAMUSS, it became more and more challenging to select (sub-)lineage-defining marker mutations that provide robust assignments in the context of the increasing diversity of SARS-CoV-2 lineages.

Data Description

Data collection and benchmark scenarios

We selected three wastewater data sets for our comparison to cover 1) a synthetic scenario of "spike-in" mixture samples (*Standards*; n=16 samples), 2) real samples from early 2021 from a large European study and collected in Germany [9], mainly comprising the VOC Alpha (*Pan-EU-GER*; n=7 samples), and 3) one sample from the end of 2021 including first signals of the VOC Omicron obtained from wastewater at the international airport in Frankfurt am Main, Germany (*FFM-Airport*; n=1 sample) [16]. The *Standards* comprise RNA from 10 SARS-CoV-2 variants (including the original Wuhan-Hu-1 A.1 lineage), which were mixed in different proportions to generate 16 samples for library preparation and sequencing via Ion Torrent (Table 2). Please note that no real wastewater was used to construct the *Standards* (see Methods). Within the Pan-EU WBE study, high-quality sequencing data was produced for SARS-CoV-2 wastewater samples across 20 European countries, including 54 municipalities [9]. We selected the seven German samples from this study (SRX11122519 and SRX11122521–SRX11122526; *Pan-EU-GER*) for our benchmark, which were sampled in March 2021 and mainly cover the rise of the VOC Alpha during that time. Lastly, we obtained one sample (SRR17258654) from wastewater sampling in November 2021 at the international airport in Frankfurt am Main (*FFM-Airport*) and where we found first signals and low proportions of the VOC Omicron arriving during that time in Germany [16].

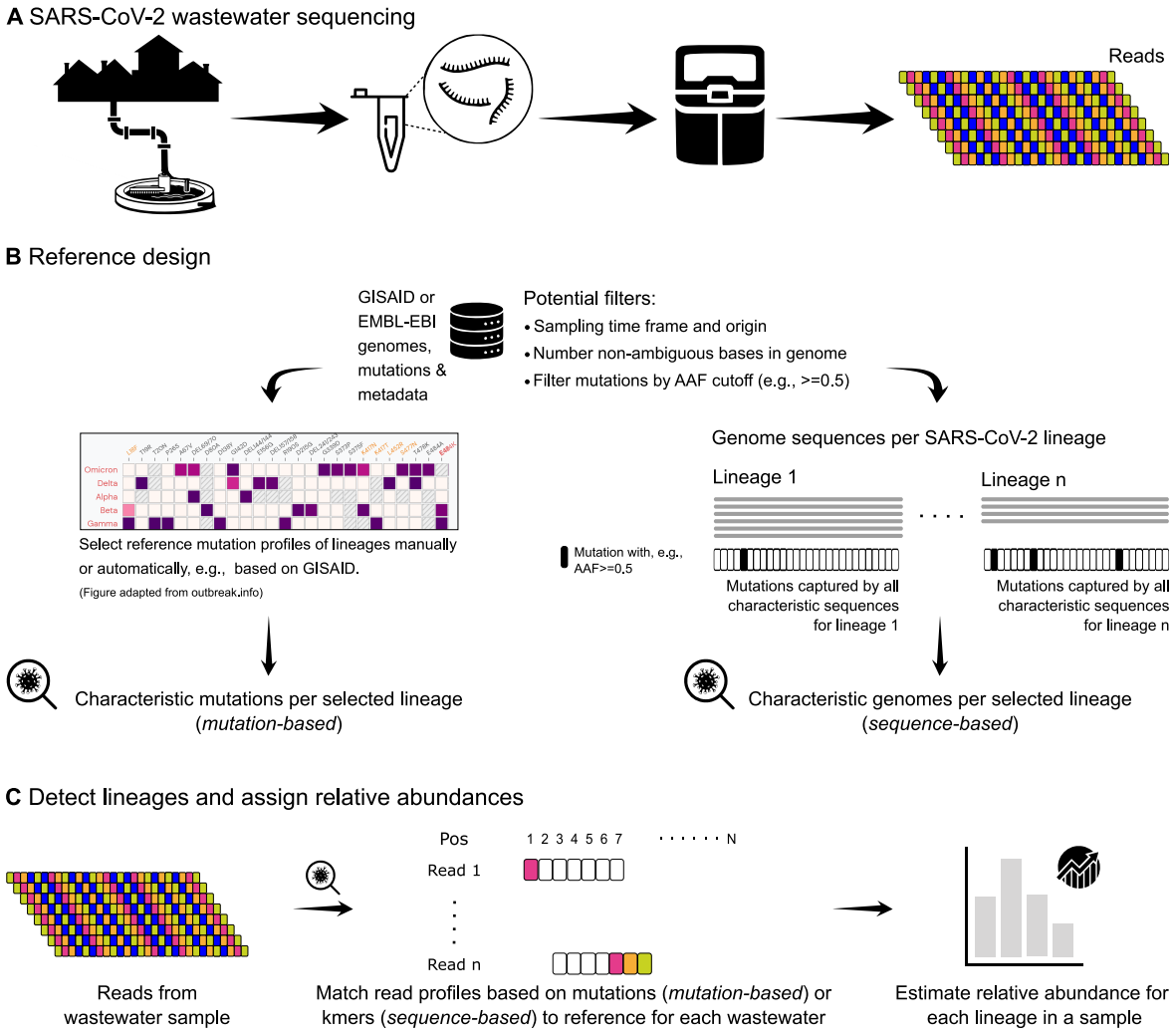


Figure 1. Schematic overview of reference design and lineage abundance estimation from SARS-CoV-2 wastewater sequencing data. (**A**) Wastewater samples are collected from sewer influent, for example. RNA is extracted and, in the context of SARS-CoV-2, usually amplified as cDNA using established primer schemes and then sequenced to obtain short snippets of viral RNA (*reads*). (**B**) Current methods (Table 1 for lineage assignment and abundance estimation need a reference data set, usually constructed from genomes and mutations derived from clinical sequencing and patient samples. Here, we distinguish two general approaches to design the reference, where either marker mutations are pre-selected (*mutation-based*) or full-genome sequences are selected (*sequence-based*). (**C**) The data analysis part may differ significantly depending on the implementation. However, all tools attempt to assign known lineages and estimate their frequency in the mixed sample based on mutations that can be detected in the reads. Our study uses MAMUSS as an exemplary *mutation-based* approach based on a two-indicator classification and pre-selected marker mutations characteristic for certain lineages [16]. For the *sequence-based* approach, we use a Nextflow implementation (VQL-nf) of the slightly adjusted VLQ pipeline as proposed by Baaijens, Zulli, and Ott *et al.* and which is based on the tool Kallisto [25]. AAF – Alternative Allele Frequency, used as a cutoff to define a mutation as a feature.

Approaches for SARS-CoV-2 reference reconstruction and lineage abundance estimation from wastewater

We compare two approaches based on different types of reference construction to assign lineages to SARS-CoV-2 wastewater sequencing data for downstream abundance estimation. Our *mutation-based* approach, MAMUSS, defines lineage-defining marker mutations, while our implementation of a *sequence-based* approach, VLQ-nf, utilizes full SARS-CoV-2 genome sequences to build a reference database. We compared both approaches on three wastewater sequencing data sets described above. For MAMUSS, the variant surveillance database from GISAID (https://www.gisaid.org) [6] is used to reconstruct reference mutation profiles for each virus variant. Mutations that have been often reported for each virus variant are considered for reference profiles. SARS-CoV-2 variants in wastewater samples are determined by comparing the mutation profiles, generated using a variant caller, with the constructed reference mutation profiles. The abundance estimation is based on the read depth and allele frequency of each mutation detected in a wastewater sample. For the *sequence-based* approach, we implemented an adjusted version of the VQL scripts by Baaijens, Zulli, and Ott *et al.* [25], re-using the pseudo-aligner Kallisto [28] as a Nextflow pipeline (VLQ-nf). This method reconstructs a reference index from GISAID SARS-CoV-2 whole-genome sequences [6]. The exact lineages and genome sequences considered for reference reconstruction are selected based on filters controlling the represented temporal and geographical range, as well as the level of genomic variation among the included lineages. SARS-CoV-2 sequencing data is then pseudo-aligned against the reference index to assign lineages. The abundance estimation is performed by an Expectation-Maximization algorithm.

Data availability

All used raw sequencing data files for the *Pan-EU-GER* and *FFM-Airport* data sets were uploaded to ENA in the context of their original publications [9, 16]. The sequencing data for the *Standards* benchmark are available under the NCBI BioProject number PRJNA912560. Further intermediate results and data files (reference sequences, constructed indices) can be found at osf.io/upbqj for full reproducibility of our analyses. The code for our Nextflow implementation based on the proposed method and original code by Baaijens, Zulli, and Ott *et al.* [25] is freely available at `https://github.com/rki-mf1/VLQ-nf`. Code for the *mutation-based* approach MAMUSS is freely available at `https://github.com/lifehashopes/MAMUSS`.

Analyses

Both the *mutation-based* and *sequence-based* approaches yield similar SARS-CoV-2 lineage proportions for mixed *Standard* samples but differ on sub-lineage level

We analyzed our *Standards* data set (Table 2) using the *sequence-based* approach implemented in VLQ-nf and an implementation of a *mutation-based* approach, MAMUSS (Table 1). Given ground truth knowledge, we assessed the qualitative and quantitative performance of both methods yielding controlled insights into the strengths and limitations of each approach.

VLQ-nf detected all correct spike-in lineages across all samples. The output for every sample showed, however, a certain amount of false positive predictions comprising lineages that are part of our reference set but not used as spike-ins (Figure 2). We observed the most consistent false positive estimations for Gamma (P.1) with up to 1.61% abundance across all samples. In contrast, MAMUSS did not detect all spike-in lineages, but also showed more robust results in quantifying fewer false positives in the samples (Figure 2).

When comparing false detection and over- or underestimation for both approaches, we partly observed similar patterns among specific groups of lineages: The *mutation-based* approach showed a bias in samples comprising A.1 towards not being able to detect A.1 and instead detecting false positives of B.1.1.7 or BA.1. In sample Mix_06, the *mutation-based* approach could not detect Iota (B.1.526) but falsely detected BA.1. Similarly, the *sequence-based* approach showed varying patterns of over- and underestimations among B.1.526 and B.1.1.7 (see Mix_01, Mix_02, and Mix_06).

Furthermore, both approaches showed distinct patterns of false estimation among B.1.617.2 (Delta) and its sub-lineages AY.1 and AY.2. In samples containing no Delta and only Delta sub-lineages, both approaches falsely detected Delta while underestimating AY.1 or AY.2. In samples containing only Delta and no Delta sub-lineages, MAMUSS falsely detected AY.1 and AY.2, while underestimating Delta. In samples containing both Delta and Delta sub-lineages, VLQ-nf overestimated Delta and underestimated AY.1, while MAMUSS overestimated Delta sub-lineages and underestimated Delta.

Both approaches estimated BA.1 and BA.2 without distinct conflicts among each other. We observed slight over- or underestimation in the abundance of Omicron lineages to co-occur with underestimation of Delta sub-lineages in samples Mix_10–16.

Finally, we found both approaches to match the ground truth proportions of the *Standards* samples well on the parent lineage level. On the sub-lineage level, we found the false negative detection of B.1.526 in sample Mix_06 and the quantification conflicts among Delta (sub-)lineages to be the most prominent conflicts among both approaches. For the *mutation-based* approach, we found the false negative detection of A.1 to be the second most prominent shortcoming observed in this experiment.

VLQ-nf detects Alpha sub-lineages while MAMUSS finds distinctly larger abundances for rising lineages Beta, Gamma, and Delta in the *Pan-EU-GER* data

We analyzed German samples from the Pan-EU study [9] using both approaches to assess their performance on real wastewater sequencing data. In the lack of ground truth knowledge, we evaluated both approaches by relating the lineage predictions and quantification to the pandemic background in Germany based on data from clinical sampling strategies. Moreover, we performed experiments on real data to evaluate the potential benefits of wastewater-based surveillance compared to clinically-based data.

According to global surveillance projects based on clinical genomic sequence data such as outbreak.info [31] (accessed June 03, 2022) and Nextstrain [32] (accessed June 03, 2022), the pandemic situation in Europe from February until April 2021 was mainly dominated by the SARS-CoV-2 lineages Alpha, Beta, cases of B.1.177 and sub-lineages, B.1.258 and sub-lineages, and B.1.160 (Supplementary Figure S1). The pandemic situation in Germany at that time was mainly dominated by Alpha, B.1.177.86, B.1.177.81, Beta, B.1.258, B.1.177, and B.1.160. According to GISAID submissions during that time, approximately the same lineages and multiple other low-abundant global and European sub-lineages were reported from clinical sampling strategies. Here we focused the comparison on the lineages Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), and the respective sub-lineages, as those were or became the dominant lineages around the time of wastewater

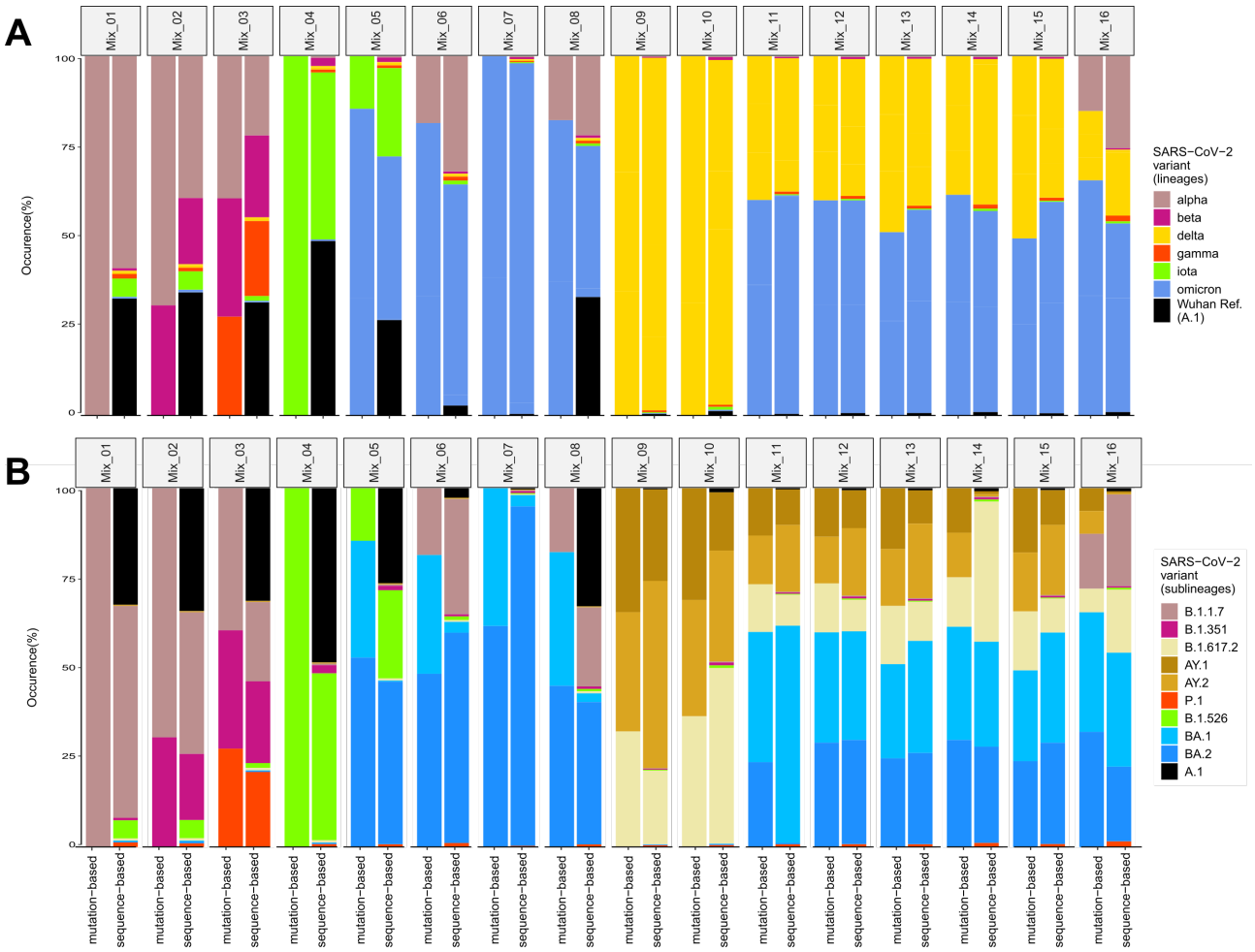


Figure 2. Comparison of the occurrence of pre-defined mixtures of SARS-CoV-2 variants (*Standards*) **(A)** at Pangolin parent lineage level and **(B)** at Pangolin sub-lineage resolution based on the *sequence-based* (VLQ-nf) and *mutation-based* (MAMUSS) approach.

sampling in Germany in the context of the Pan-EU project [9].

With VLQ-nf, we quantified the lineage and sub-lineage level. In comparison, MAMUSS predicted lineage abundances only at the parent level (Figure 3). Both approaches predicted Alpha (sub-)lineages to be the most abundant lineages in the data set. Specifically, the *sequence-based* approach found Alpha sub-lineages Q.1 and Q.7 to be the most abundant. Yet, those Alpha sub-lineages were not reported amongst the most frequent cases based on clinical sampling strategies (see Supplementary Figure S1). We also detected Beta, Gamma, and Delta (sub-)lineages at abundances below 1%, which are not visible at the scale of Figure 3. In contrast, we found distinctly larger abundances of Beta, Gamma, and Delta in the samples using MAMUSS.

Mutation- and *sequence-based* approaches recover a similar Omicron proportion from an early airport wastewater sample

We used both approaches to analyze a real wastewater sequencing sample (SRR17258654, *FFM-Airport*) [16]. We compared lineage predictions and quantification against the pandemic background in Europe and South Africa at the time of wastewater sampling. We evaluated both approaches in terms of their ability to detect (sub-)lineages at low abundances, specifically to detect low abundant signals of Omicron.

The pandemic situation in Europe and South Africa from October to December 2021 was dominated by Delta sub-lineages and increasing incidences of Omicron and its sub-lineages according to outbreak.info [31] and nextstrain.org [32] (Supplementary Figure S2). According to GISAID submissions, mostly Delta sub-lineages and a few cases of Omicron and other minor global sub-lineages were reported based on clinical sampling strategies.

With VLQ-nf, we detected many Delta sub-lineages at abundances ranging from less than 1% to around 8% that in sum contribute over 93% abundance in the wastewater sample (Figure 4). We observed BA.1 with 1.44% and some other lineages and sub-lineages with abundances of less than 1% ("Other") in that sample. We found all lineage quantification of less than 1% to aggregate to around 48% abundance in total.

We observed a similar lineage abundance profile with MAMUSS. We found that most abundance consists of two approximately equally abundant Delta sub-lineages. We detected a small proportion close to 1% of Omicron. Compared to VLQ-nf, we did not find any low abundant quantification for other (sub-)lineages, explained by the smaller reference data set only composed of a particular collection of marker mutations.

We found that the estimated abundance profiles of lineages from both approaches matched well with the pandemic background in Europe and South Africa at the time of wastewater sampling. However, when considering abundance estimations of the *sequence-based* approach at the sub-lineage level, we discovered differences regarding the most abundantly predicted Delta sub-lineages compared to the more prominent Delta sub-lineages derived from clinical sampling strategies in European and South African GISAID submissions. The *sequence-based* approach predicted AY.25.1, AY.125.1, AY.122.4, AY.121, and AY.43.1 to be most abundant in the analyzed sample. In contrast, GISAID submissions showed AY.4, AY.43, AY.122, AY.4.2, AY.126, AY.4.2.2, and AY.98 as the most frequent Delta sub-lineages in Europe during that time. Additionally, we found AY.45, AY.32, AY.91, AY.116, AY.122, AY.6, and AY.46 to be the highest reported Delta sub-lineages in South Africa. While our predictions do not match the clinically reported frequencies, some of our predictions belong to the same lineage family as the most frequently reported lineages from clinical sampling, e.g., AY.43.1 is a sub-lineage of AY.43, AY.122.4 is a sub-lineage of AY.122, and AY.125.1 is a sub-lineage of AY.125 which we found among the twenty most frequently reported lineages in Europe using VLQ-nf.

Alternative allele frequency and size of reference database impact the *sequence-based* method, but the effects also depend on lineage composition in the sample

To better understand the impact of specific parameters on the performance of the *sequence-based* method, we performed parameter escalation experiments on the *Standards* benchmark set as well as the *PanEU-Ger* and *FFM-Airport* data sets. Due to the similar findings for all three data sets, here we only present the results based on the *Standards* and refer to the results of the *PanEU-Ger* and *FFM-Airport* data sets in the Supplement. We investigated the impact of reference construction parameters on lineage proportion estimation and aimed at uncovering the potential bias of the pseudo-alignment implemented in the *sequence-based* method. Specifically, we focused on the AAF threshold and the maximum number of sequences per lineage. The AAF threshold defines the minimum alternative allele frequency for a mutation to be considered characteristic of a lineage. First, genome sequences are added as lineage references so that each mutation that exceeds the AAF threshold is detected at least once by as few sequences as possible. Next, additional genomes are added until the maximum number of sequences per lineage is reached. Thus, the AAF threshold controls the level of genomic variation captured for each lineage and the maximum number of sequences per lineage controls the reference size.

Standards

Across most *Standards* samples and experiments, VLQ-nf detected all spike-in lineages and predicted reasonable estimates (Figure 5). However, we consistently observed low abundant false positive hits in all of our mixed samples, comprising lineages that are part of the reference index but not used as spike-ins. We found the most prominent false positive detection to be Gamma. We observed similar patterns of false positive detection and false estimation among specific groups of lineages across all parameter settings: For the first eight samples Mix_01 to Mix_08, most cases of false estimation of spike-in lineage abundances occurred alongside false positives or negatives of B.1.526 and false positives of BA.1. For the samples Mix_09 to Mix_16, we observed most detection conflicts to involve ambiguities among Delta and its sub-lineages AY.1 and AY.2.

We found that the detection and quantification performance of the *sequence-based* method via VLQ-nf changed with varying parameter settings. Specifically, we found those changes to vary across samples and observed them not to behave identically with consistent parameter changes. For example, at the minimum reference size (Supplementary Table S1), we observed abundance predictions for samples Mix_09 and Mix_11–16 to first improve with an increasing AAF threshold. However, with a further increasing AAF threshold, we observed more false estimations of Delta sub-lineages. Furthermore, although Mix_10 shares most of its spike-in lineages with Mix_09, the performance of abundance estimations for sample Mix_10 first decreased and then improved again when increasing the AAF threshold.

We made a similar observation for the maximum number of sequences per lineage. With an AAF threshold of 0.5, the abundance estimates for Mix_01 improved with increasing reference size, while we found them to deteriorate for Mix_09 which includes a distinctly different sample composition. Over-

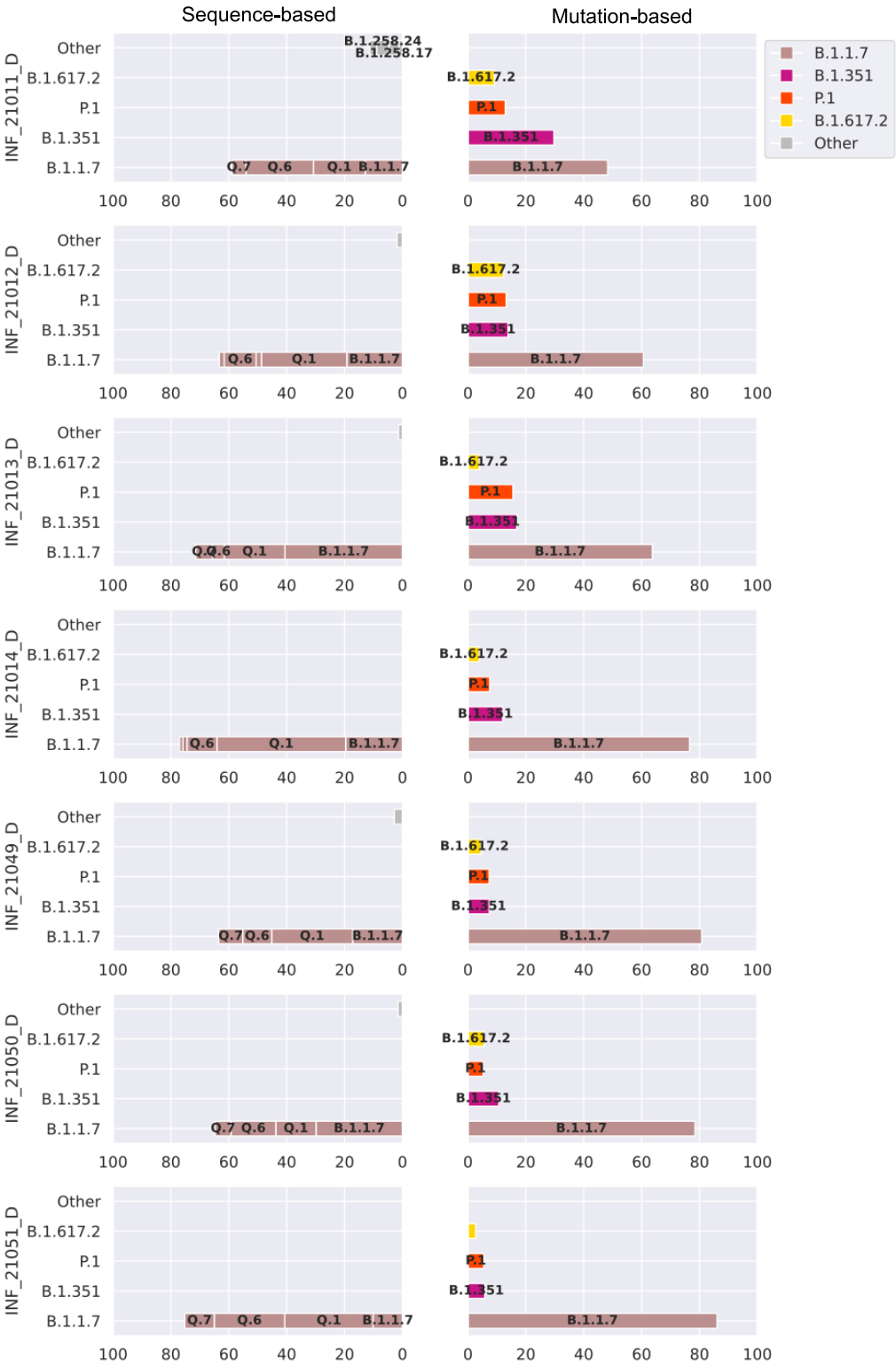


Figure 3. Comparison of the results for the *Pan-EU-GER* analysis using VLQ-nf (*sequence-based*, left) versus MAMUSS (*mutation-based*, right). Abundance predictions are plotted above a cutoff of 1% abundance and labeled at a threshold of 3% abundance. VLQ-nf detected abundances for B.1.617.2, P.1, and B.1.351 sub-lineages below 1%, which is not visible at the scale of this figure.

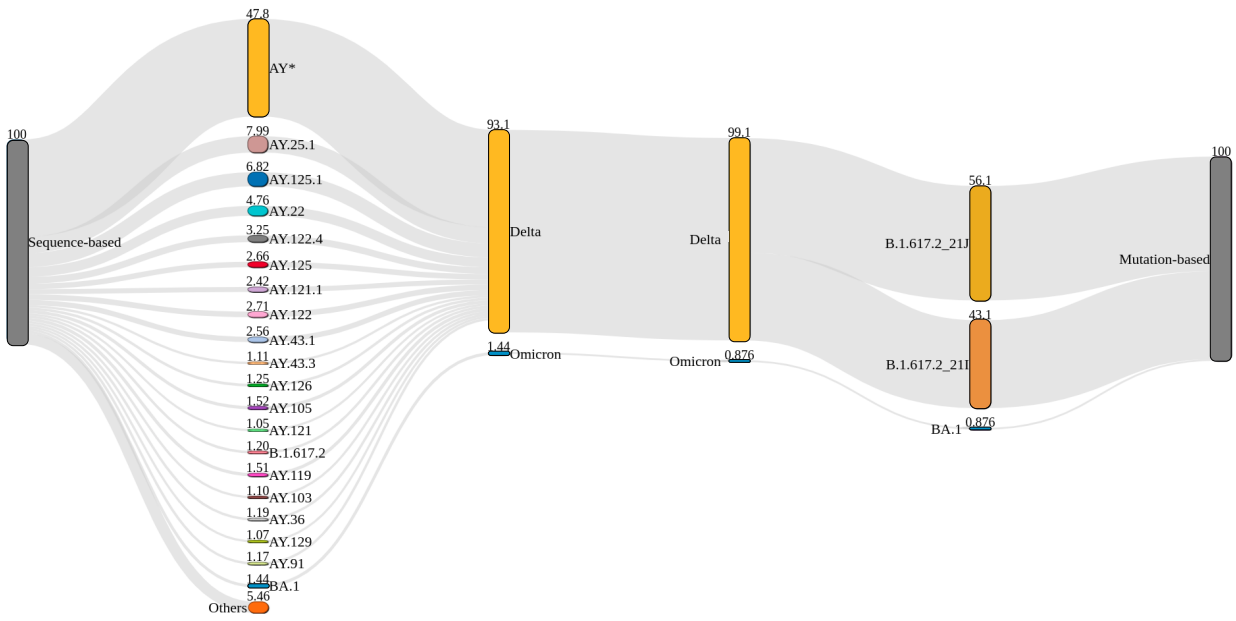


Figure 4. Sankey plot comparing the detected lineage proportions for the *sequence-based* approach (VQL-nf, left) and the *mutation-based* approach (MAMUSS, right) for one airport wastewater sample (SRR17258654) [16]. Both approaches detect a similar amount of Delta and Omicron (BA.1) in the sample, while VQL-nf can achieve a higher sub-lineage resolution (AY lineages) based on the full genome information in the reconstructed reference index and utilizing pseudo-alignments. MAMUSS can, as configured for this analysis and based on the limited reference set, distinguish between two slightly different B.1.617.2 clades as defined by Nextstrain. For the *sequence-based* approach, only lineages with a proportion of at least 1% are shown and all other AY-sub-lineages are pooled in AY* and all other lineages in "Others".

all, we found lineage abundance estimations to become slightly more robust across varying AAF thresholds with increasing reference size. This is reflected best in the abundance profiles for samples Mix_09–Mix_15 when looking at the proportional changes across increasing AAF settings for the minimum reference size throughout the reference with 10 sequences per lineage.

Finally, we found that the AAF threshold and the reference size affect the performance of the *sequence-based* method. Although we did not observe a clear and consistent pattern of impact, we found that the effects of varying parameter settings may depend on the sample composition. Specifically, we observed the strongest impact of parameter changes for samples containing lineages with a higher degree of shared genomic similarity. Also, we found the AAF threshold to affect estimates slightly more than the reference size. We detected similar results for the *PanEU-Ger* and *FFM-Airport* data sets. We provide details for these two data sets in the Supplement (see Figure S3 and Figure S4).

Final choice of parameters for benchmark reference construction

Within the scope of the parameter escalation experiments described here, we wanted to determine parameters with a good prediction performance without manipulating the benchmark in favor of the *sequence-based* approach (VQL-nf). Finally, based on our parameter testing and the three different data sets, we chose an AAF threshold of 0.25 and a reference size of at most 5 sequences per lineage. By that, we limit the size of the reference data set and still allow reasonable detection and quantification results across all three benchmark data sets, while at the same time keeping computational resources moderate.

Discussion

It is apparent that the composition of the reference used must have a large impact on the determination of relative SARS-CoV-2 abundances in wastewater sequence data. Especially given the dynamic and constantly updated SARS-CoV-2 lineage definitions [2], the reference genome sequences and the signature mutations derived from them also change frequently. Of course, the various tools (Table 1) and their parameters developed for estimating the relative abundance of lineages from wastewater sequencing data also have an impact. Here, however, we have specifically focused on the effects of the reference design.

We selected two general approaches for the design of reference data sets and estimation of SARS-CoV-2 lineage proportions from wastewater sequencing samples (Figure 1). On the one hand, selected marker mutations that are characteristic for certain SARS-CoV-2 lineages can be used for annotation and lineage proportion estimation (*mutation-based*, MAMUSS). Here, the read sequences derived from a wastewater sample are mapped against a reference genome from which differences (mutations) are detected and compared against the selected marker mutations. On the other hand, full SARS-CoV-2 genome sequences can be used to create a reference index without prior collection of specific mutations (*sequence-based*, VLQ-nf). Here, the problem of selecting appropriate marker mutations is shifted to the selection of representative lineages from which then features for the classification task are derived. An exemplary implementation of this approach based on the pseudo-aligner Kallisto [28] was recently proposed by Baaijens, Zulli, and Ott *et al.* [25]. Based on their work, we developed a Nextflow pipeline for higher automation and reproducibility and the detection of SARS-CoV-2 lineage proportions from wastewater data using pseudo-alignments (VLQ-nf). In this approach, a selection of whole-genome SARS-CoV-2 sequences (target reference set) and the reads (query) are composed into kmers which are then efficiently compared to quantify lineage abundances, similar to quantifying gene expression in an RNA-Seq study.

To benchmark reference designs from both methods (*mutation-based* via MAMUSS, *sequence-based* via VLQ-nf), we selected three test scenarios: 1) a spike-in experiment with dif-

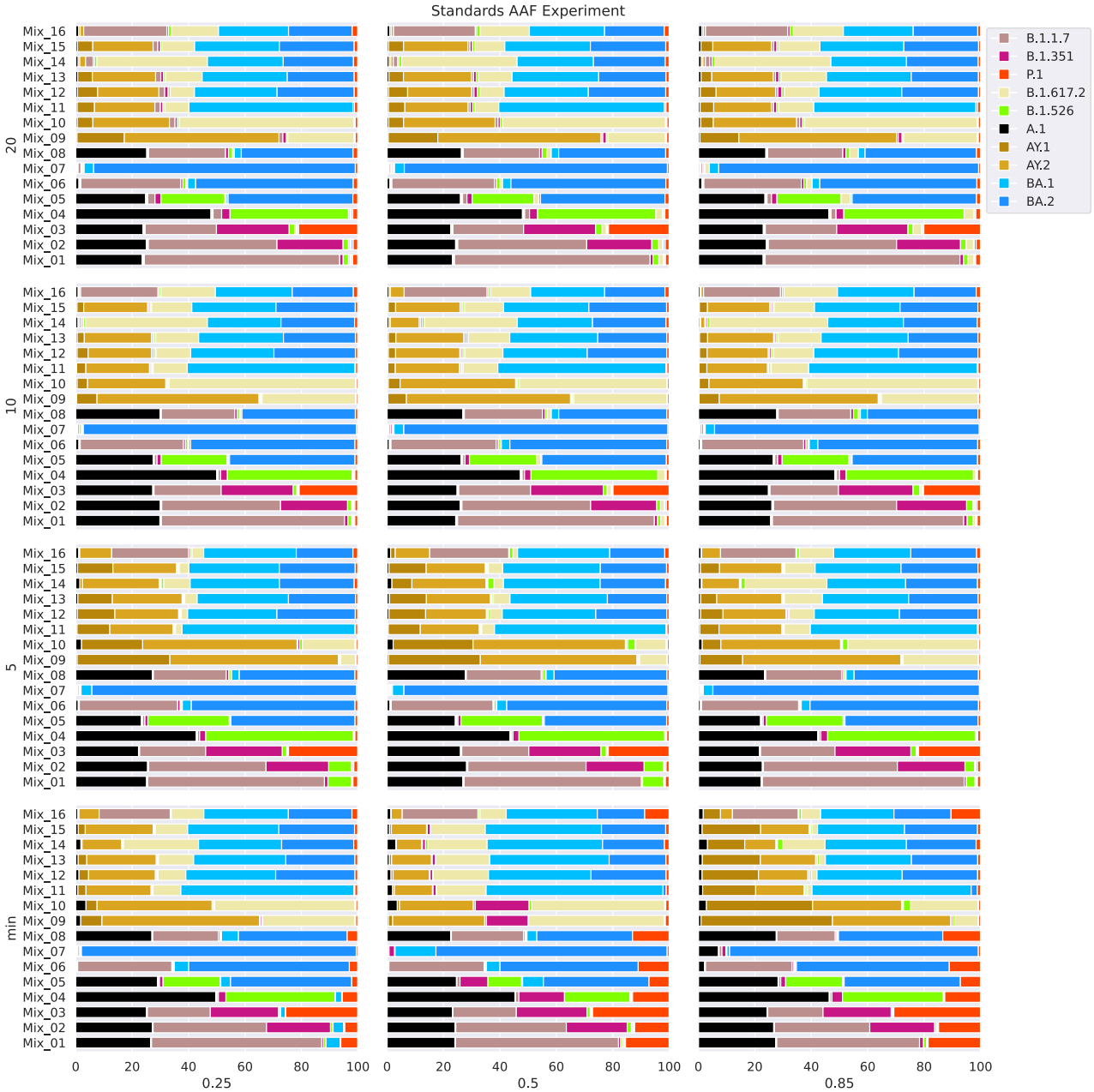


Figure 5. Results for the parameter escalation experiments on the *Standards* samples using the *sequence-based* method (VLQ-nf). We analyzed the *Standards* with different parameterizations for reference construction (x-axis: increasing AAF threshold, y-axis: increasing maximum number of sequences per lineage). VLQ-nf using pseudo-alignments detected all lineages and estimated abundance profiles well across most samples and parameter settings. However, we also observed prominent detection ambiguities among Delta and its sub-lineages and found consistently low abundant false positives for specific groups of lineages. Continuously increasing or decreasing parameter settings caused heterogeneous changes in the estimated abundance proportions across samples. The *sequence-based* method showed to perform better when using a reference set larger than the minimum reference size. Still, we found noise levels to increase distinctly when using the maximum reference size among the considered settings.

ferent SARS-CoV-2 lineage mixes, 2) samples obtained for Germany from a Pan-EU wastewater study, and 3) a wastewater sample from a German airport during the time when Omicron emerged.

In general, both approaches were able to detect SARS-CoV-2 lineage abundances from our test cases. The most remarkable difference was in the number of detected sub-lineages which also directly correlates with the reference design. VLQ-nf generally detected a larger diversity of sub-lineages in comparison to MAMUSS, which can be explained by the underlying reference indices. For the *mutation-based* approach and the implementation we used, it got increasingly difficult to select a representative set of marker mutations. In contrast, the *sequence-based* approach as suggested by Baaijens, Zulli, and Ott *et al.* [25] can build a reference index on a large collection of SARS-CoV-2 full genome sequences and thus, potentially, better reflect diversity on sub-lineage levels. However, we also observed a certain amount of *noise* in the pseudo-alignment results causing potential false-positive hits in our test data sets. Other approaches, like Freyja [17], partly tackle this problem by deriving signature mutation profiles automatically, for example using the whole phylogenetic diversity of current SARS-CoV-2 sequences reflected in an UShER tree [29]. However, here we have also observed that the inclusion of a large diversity in the reference can lead to distributed abundance assignments between closely related (sub)-lineages, reducing the true relative abundance of a lineage (Figure S5 and Figure S6). Of course, the impact can be reduced by limiting lineage coverage to a specific time period, but this, in turn, can also affect frequency assignments.

In more detail, both approaches performed similarly in detecting and estimating spike-in lineage abundances for the *Standards* data set Figure 2. The predictions are more similar on the parent-lineage level compared to the sub-lineage level. If their estimations differ, this can be mostly attributed to differences in the mutations/lineages included in the respective reference data: for both approaches, the final predictions heavily depend on the construction of the reference data set. In addition, both approaches had difficulties differentiating closely related sub-lineages correctly.

For the *Pan-EU-GER* data set, both approaches reflect well the pandemic background in Germany during the time of sampling, but we detected some limitations and potential sources for bias in their general behavior. Again, the choice of marker mutations and reference lineages impacts the level of detection: lineage vs. sub-lineage level estimations, but also the amount of low abundance detection. Potentially, everything that is defined in the reference data set can be also detected, which might lead to an increased number of false positive predictions. The whole-genome sequences or mutations used to create the reference index impact the degree of ambiguity and, thus, (low abundant) false positive detection. This may explain why here both approaches predicted distinctly different abundances on the parent-lineage level compared to the other two benchmark experiments. Therefore, we think that especially the *sequence-based* approach requires the definition of a false positive threshold to differentiate between low abundant false positive hits and low abundant true positives.

For the *FFM-Airport* data set, both approaches detect also low-frequency lineages. Again, the *sequence-based* approach detects a distinctly higher amount of low abundant lineages, also reflecting the higher diversity of the reference index.

We performed an additional parameter benchmark to identify important key parameters impacting the *sequence-based* pseudo-alignment approach using VLQ-nf. One parameter having a strong effect on the results, is the alternative allele frequency (AAF) cutoff. In connection with the reference size (the number of genomes), we observed different effects of changing the AAF. Our experiments also showed that the effect of the same parameter changes (increasing or decreasing AAF) does not yield consistent results among the different data sets. The degree of lineage ambiguity depends on the considered composition of lineages and sub-lineages. The effect of included/excluded mutations due to adjusted AAF parameter settings is variable, as different mutations have different effects in differentiating lineages. The effect of those parameter changes is most notable among lineages that are more similar. We also observed that with a larger reference size, the effect of the AAF parameter becomes smaller and overall abundance estimations improve. Increasing the reference size implicitly adds low-frequency mutations into a lineage reference set which in part reduces the effect of increasing/decreasing the AAF threshold when selecting sequences. Additionally, depending on their phylogenetic impact, low-frequency mutations might help better differentiate lineages.

Potential implications

Most importantly, we only selected two exemplary implementations of the *mutation-* and *sequence-based* approaches MAMUSS and VLQ-nf, respectively, out of an increasing number of scripts, tools, and pipelines becoming available for computational SARS-CoV-2 lineage estimation from wastewater sequencing (Table 1) [17, 18, 19, 20, 13, 21, 22, 14, 7, 23, 24, 25, 26, 27]. Thus, our benchmark results also reflect and are limited by the individual characteristics of these two implementations. However, we focused on these two approaches to investigate the impact of reference design using implementations where we could easily control parameters and input. Currently, a comprehensive benchmark comparison for the existing SARS-CoV-2 wastewater analysis tools is lacking. The developers of Freyja compared a selection of tools on a spike-in mixed sample [17] where they found that Freyja outperformed VLQ [25] in accuracy at higher expected proportions and observed noticeably longer computation times for both VLQ and LCS [19]. To counteract the effect on lineage abundance detection, some methods filter the mutations considered for lineage assignment based on sequencing depth [13] or adjust their mathematical model for differences in depth and coverage and expected error rates [17, 23]. In a similar context, the PiGx tool addresses the limitations of estimating lineages at low abundances by weighting specific signature mutations for lineages that are expected to occur at low frequencies [22]. As a next step, a broader evaluation of all available tools for the analysis of SARS-CoV-2 wastewater sequencing data is urgently needed to guide usage and further development.

Conclusion

Academic researchers have pioneered wastewater monitoring of SARS-CoV-2 and overcome several technical and methodological challenges [12]. Thanks to these efforts, wastewater-based pathogen surveillance has rapidly become a valuable public health tool for detecting SARS-CoV-2 that can excellently complement syndromic surveillance or other monitoring tools. However, public health authorities are now faced with the task of integrating these achievements into robust and continuous public health surveillance systems that can be operated and expanded over the long term. For the inclusion of wastewater-based pathogen surveillance data, performance parameters must be defined and communicated to the public health authorities. In this context, continuous updating of reference data sets, in the context of retrospective analyses or time series, is essential to ensure comparability between

time points. Especially for continuous sampling and analysis of wastewater-based SARS-CoV-2 sequencing data, the reference design must also be adjusted. Otherwise, a lineage defined with a delay might be present in older samples but was not detected only because it was not part of the reference at that time. However, harmonizing the reference used would require recalculating older abundance estimates, which may conflict with the standard reporting requirements of public health authorities. However, this problem is not specific to wastewater-based SARS-CoV-2 sequencing data, but also applies to genomics sequencing of patient samples. One solution might be to focus not only on lineages, but also to report mutations that are not affected by any nomenclature scheme and are not subject to delayed definitions. On the other hand, it is undeniable that lineages played a crucial role in communication during the COVID-19 pandemic.

The detection of *cryptic* (novel, undescribed) virus variants in wastewater samples is another powerful application for wastewater sequencing data, especially when clinical testing capabilities and monitoring systems are limited or reduced. In this context, approaches utilizing artificial intelligence might present a promising next step for the improved detection of cryptic SARS-CoV-2 lineages from wastewater sequencing data and potential outbreaks, although right now not much in use [33]. Finally, the lessons learned from the sequencing efforts and implementations for SARS-CoV-2 detection from wastewater sequencing data can and should be adapted to other pathogens in the future.

Methods

Benchmark data set #1: Synthetic mixture *Standards*

We procured synthetic SARS-CoV-2 RNA samples (Twist Biosciences), which were used to prepare 16 different mixtures (Table 2) containing different SARS-CoV-2 variants. From the pooled RNA, cDNA was synthesized using SuperScript™ VILO™ Master Mix (Thermofisher Scientific), followed by library preparation using the Ion AmpliSeq SARS-CoV-2 Research Panel (Thermofisher Scientific) according to the manufacturer's instructions. This panel consists of 237 primer pairs, resulting in an amplicon length range of 125–275 bp, which cover the near-full genome of SARS-CoV-2. We performed multiple sequencing runs to achieve at least 1 million mapped reads per sample. For each sequencing run, eight libraries were multiplexed and sequenced using an Ion Torrent 530 chip on an Ion S5 sequencer (Thermofisher Scientific) according to the manufacturer's instructions. The raw sequence data were uploaded to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject number PRJNA912560.

Benchmark data set #2: *Pan-EU-GER*

We obtained seven real samples from March 2021 from a large European study and collected in Germany [9], mainly comprising the VOC Alpha (*Pan-EU-GER*, SRX11122519 and SRX11122521–SRX11122526).

Benchmark data set #3: *FFM-Airport*

We selected one sample from the end of 2021 including first signals of the VOC Omicron obtained from wastewater at the international airport in Frankfurt am Main, Germany (*FFM-Airport*, SRR17258654) [16].

Data processing: *mutation-based* reference design and lineage proportion estimation via MAMUSS

We used the SARS-CoV-2 Research Plug-in Package, which we installed in our Ion Torrent Suite software (v5.12.2) of Ion S5 sequence. We used the SARS_CoV_2_coverageAnalysis (v5.16) plugin, which maps the generated reads to a SARS-CoV-2 reference genome (Wuhan-Hu-1-NC_045512/MN908947.3), using TMAP software included in the Torrent Suite. The summary of mapping of each sample mentioned in Table 2 is provided in Table S2. For mutation calls, additional Ion Torrent plugins were used as described previously [34]. First, all single nucleotide variants were called using Variant Caller (v5.12.0.4) with "Generic - S5/S5XL (510/520/530) - Somatic - Low Stringency" default parameters. Then, for annotation and determination of the base substitution effect, we used COVID19AnnotateSnpEff (v1.3.0.2), a plugin developed explicitly for SARS-CoV-2 and based on the original SnpEff [35]. To construct reference marker mutation sets for MAMUSS, we used data from GISAID (https://www.gisaid.org) [6] to reconstruct reference mutation profiles for each virus variant. The lineage abundance estimation is based on the read depth and allele frequency of each mutation detected in a wastewater sample followed by a two-indicator classification and comparison to the pre-selected marker mutations characteristic for certain lineages. For further details see github.com/lifehashopes/MAMUSS.

Data processing: *sequence-based* reference design and lineage proportion estimation via VLQ-nf

Instead of only relying on manually or algorithmically selected marker mutations, another computational approach utilizes, in a first step, full genome information. For example, Baaijens, Zulli, and Ott *et al.* presented a method to estimate the abundance of variants in wastewater samples based on well-established computational techniques initially used for RNA-Seq quantification [25]. Here, the main idea is that quantification of different transcripts derived from the same gene is computationally similar to the abundance estimation of different SARS-CoV-2 lineages derived from the same parental genome. Via Kallisto [28], they perform pseudo-alignments of the raw reads against an index of pre-selected and down-sampled full genome SARS-CoV-2 sequences with respective lineage information. Therefore, their approach may be less influenced by the pre-selection of mutations based on clinical relevance, frequency, or other parameters that mostly drive *mutation-based* tools, and thus may be better suited for sub-lineage discrimination. The approach comprises two steps: 1) selection of reference genome sequences for index construction and 2) pseudo-alignment of the reads and lineage abundance estimation. First, a reference data set of SARS-CoV-2 genome sequences must be selected. For that, we use data from GISAID [6] and filter for human-host sequences, N-count information, pangolin annotation [2, 1], origin (country, continent), and sampling date. This metadata is used to pre-select sequences based on geographic origin (continent, country), a sampling time frame, and the number of N bases. Next, the pipeline performs a variant calling against a reference sequence (per default index Wuhan-Hu-1, NC_045512.2) and subsequently samples sequences to select characteristic mutation profiles for each input lineage. Within a lineage, sequences are sampled based on an alternative allele frequency cutoff (e.g., AAF>0.5) so that each mutation is represented at least once until an upper limit of sequences per lineage is reached. From this down-sampled and representative set of full genome sequences, a Kallisto index is constructed. Now, the raw reads from a FASTQ

file are pseudo-aligned against this index and lineage abundances are estimated similarly to the estimation of transcript abundances in an RNA-Seq scenario.

For our comparative study, we used the initial idea and code base from Baaijens, Zulli, and Ott *et al.*[25] (https://github.com/baymlab/wastewater_analysis, version from September 16, 2021, with commit hash 61dd29df*) and implemented a Nextflow [30] pipeline (https://github.com/rki-mf1/VLQ-nf) with the purpose of automating the steps and making our analyses fully reproducible. In this context, we discovered some issues in the pipeline version 61dd29df* of Baaijens, Zulli, and Ott *et al.* and implemented minor adjustments. This includes updating data processing scripts according to the most recent GISAID data format and allowing the sequence selection based on alternate allele frequencies (AAF) to consider multi-allelic sites. Meanwhile, those issues have been addressed with similar code changes by the authors in their current pipeline version. Furthermore, we adjusted the AAF filter to first sample sequences so that all passing mutations are included at least once before increasing the reference setup to the number of maximum sequences per lineage. In pipeline version 61dd29df*, sequences are selected for the reference index if they carry an AAF filter passing mutation that is not yet covered until the reference set for the respective lineage meets the maximum allowed number of sequences. We wondered if this sampling strategy might yield a reference index that does not sufficiently represent the characteristic mutation profile of a lineage and could introduce potential sources of bias during reference reconstruction. We addressed this issue by implementing an AAF filtering that first retrieves a minimal reference database for every lineage such that all AAF filter passing mutations are captured at least once by as few sequences as possible. In a second step, we incremented each lineage reference to match a maximum number of sequences using the remaining sequences passing the filter. We ran our pipeline version v1.0.0 for all analyses in this benchmark study.

Reconstruction of indices for the *sequence-based* approach

The *sequence-based* (VLQ-nf) approach highly depends on the selection and reconstruction of the reference data set for the Kallisto index. Thus, we reconstructed different indices for our three benchmark data sets to mimic the pandemic situation during the time of sampling. For all indices, we used GISAID data and extracted subsets based on metadata filters.

For the benchmark of the 16 mixed *Standards*, we constructed a reference data set comprising the included SARS-CoV-2 lineages. We selected a time frame of two weeks around the peak of global incidences (based on outbreak.info [31], accessed April 1st, 2022) for each lineage included in the mix (Table 3). We only kept records with at least 29,500 non-ambiguous bases. Because we also included the original Wuhan-Hu-1 reference sequence in mixed samples Mix_01-Mix_05 and Mix_08, we first excluded all A.1 sequences from the preselected set. Then, we selected reference sequences with characteristic mutation profiles for all lineages except A.1 as described before allowing a maximum number of five sequences per lineage. Now, we added the sampled A.1 sequences again to the final reference set manually. Otherwise, the A.1 sequences would have been excluded by the pipeline because they don't show any AAF in comparison to the Wuhan-Hu-1 reference. On average, we selected five sequences for a lineage to capture every mutation against the wildtype with an AAF>0.25 (within-lineage variation) and a maximum of five allowed sequences per lineage.

For the *Pan-EU-GER* samples (collected between 10th and 30th March 2021), we reconstructed the reference from GISAID records we downloaded on 27 January 2022. We selected only European sequences sampled between February 1st, 2021, and April 30nd, 2021, with at least 29,500 non-ambiguous bases. We did not only select sequences from Germany to also mimic variant influx from other European countries. On average, we then selected three sequences per lineage to capture every mutation against the wildtype with an AAF>0.25 (within-lineage variation) and allowing at most five reference sequences per lineage.

Table 3. For each lineage in the *Standards* data set, we selected the time frame where infection numbers peaked globally according to outbreak.info [31]. Based on the listed time frames, we sampled genome sequences from GISAID for reference reconstruction. We downloaded the GISAID records on 02 March 2022.

Lineage	Time frame
A.1	2020-03-01:2020-03-14
B.1.1.7	2021-05-01:2021-05-14
B.1.351	2021-01-20:2021-02-02
P.1	2021-04-20:2021-05-03
B.1.526	2021-03-20:2021-04-02
BA.2	2022-02-01:2022-02-14
BA.1	2021-12-01:2021-12-14
B.1.617.2	2021-06-25:2021-07-08
AY.1	2021-08-01:2021-08-14
AY.2	2021-06-25:2021-07-08

For the *FFM-Airport* data set, we reconstructed the reference from GISAID records we downloaded on 11 February 2022. We selected sequences from European and South African samples sampled between October 1st, 2021, and December 31st, 2021, again with at least 29,500 non-ambiguous bases. On average, four sequences were selected for a lineage to capture every mutation against the wildtype with an AAF>0.25 (within-lineage variation). Again, we allowed at most five sequences to be included per lineage.

Lineage-abundance estimation with the *sequence-based* approach

After reconstructing different reference indices for our benchmark data sets, we used specific Kallisto commands implemented in a Nextflow pipeline to prepare Kallisto mapping indices, compute pseudo-alignments of each benchmark data set against its reference index, and estimate lineage abundances following the original idea and code of Baaijens, Zulli, and Ott *et al.*[25].

First, we built a Kallisto index from the reference database (default k-mer=31). Next, for each sample in a benchmark data set, we pseudo-aligned all reads against the corresponding Kallisto index and estimated the abundance of each reference sequence in the sample. We quantified our benchmark data sets in single reads mode with an average fragment length of 200 nt with a standard deviation of 20 nt. Finally, a customized script groups the estimated abundances by the lineage annotation of the respective sequences and sums them up into a final lineage abundance estimation for the analyzed sample. For the *Pan-EU-GER* and *FFM-Airport* data sets, we further summarized the estimated abundances by the country information of the analyzed samples to compare the pseudo-alignment and *mutation-based* approach on the country level.

Assessing parameter impact and potential bias with the pseudo-alignment approach

We performed parameter escalation experiments with our three benchmark data sets using the *sequence-based* method (VLQ-nf) to assess the impact of the AAF threshold and the cut-off for a maximum number of sequences per lineage on lineage abundance estimation. More importantly, we used the resulting observations to inform our choice of parameters used for the final benchmarking against the *mutation-based* method (MAMUSS). In this context, we aimed at determining a setting with a good prediction performance and reasonable computational effort without manipulating the benchmark in favor of the *sequence-based* method. For every benchmark data set, we constructed reference indices over a range of 12 possible parameter combinations. For the AAF threshold, we iterated over [0.25, 0.5, 0.85] to cover lower, medium, and high threshold values to define the characteristic mutation profiles. For the maximum number of sequences per lineage, we built the reference index using the minimal sequence sets possible, 5, 10, and 20 sequences per lineage. After lineage abundance estimation with each reference index on the *Standards* data set, we evaluated prediction performance based on the ground truth lineage abundances. For the *FFM-Airport* and *Pan-EU-GER* data, we assessed prediction performance by comparing estimated lineage abundances with the pandemic background at the respective time and location.

Reproducibility of the pseudo-alignment approach

Our Nextflow pipeline of the pseudo-alignment approach freely available at github.com/rki-mf1/VLQ-nf generates the reference database in the format of a CSV file containing the metadata information of the final Kallisto index and a FASTA file containing the corresponding sequence data. In the current version v1.0.0, the reference CSV and FASTA can be exactly replicated using the same input data resource and index reconstruction parameters which leads to slightly different results at every analysis run. The reference CSV is not reproducible due to misplaced random sampling seeds and a missing record sorting strategy in the AAF-based sequence filtering step during reference reconstruction.

However, given already reconstructed reference indices (final CSV and FASTA reference) or an already built Kallisto index, lineage detection and quantification are deterministic. We deposit our used Kallisto indices at osf.io/upbqj.

Availability of source code and requirements

Here, we provide the specifications of our Nextflow implementation (VLQ-nf) of the *sequence-based* approach originally presented by Baaijens, Zulli, and Ott *et al.*[25] and the code for the *mutation-based* approach, MAMUSS.

- Project name: VLQ-nf
- Project home page: `https://github.com/rki-mf1/VLQ-nf`
- Operating system(s): Linux, Mac, Windows via Linux subshell
- Programming language: Nextflow
- Other requirements: Conda
- License: GPL-3.0

- Project name: MAMUSS
- Project home page: `https://github.com/lifehashopes/MAMUSS`
- Operating system(s): Linux, Mac
- Programming language: R
- Other requirements: R packages are listed in the repository
- License: CC0 1.0 Universal

Availability of supporting data and materials

The data sets supporting the results of this article are available in the Open Science Framework repository, osf.io/upbqj.

Declarations

List of abbreviations

- AAF - alternative allele frequency
- FFM-Airport - one sample from the end of 2021 including first signals of the VOC Omicron obtained from wastewater at the international airport in Frankfurt am Main, Germany [16]
- MAMUSS - *mutation-based* approach for SARS-CoV-2 lineage abundance estimation
- Pan-EU-GER - seven samples from early 2021 from a large European study and collected in Germany, mainly comprising the VOC Alpha [9]
- Standards - synthetic scenario of 16 "spike-in" mixture SARS-CoV-2 samples
- VLQ-nf - *sequence-based* approach for SARS-CoV-2 lineage abundance estimation, inspired by the original VLQ [25]
- WBE - wastewater-based epidemiology

Ethical Approval (optional)

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests

Funding

Author's Contributions

SA, SL, and MH provided conceptualization and study design. SA implemented the MAMUSS approach and analyzed corresponding data. EA implemented the VLQ-nf approach and analyzed corresponding data. SA and LO conducted wet lab experiments to generate and sequence synthetic mixtures. EA, SA, and MH performed the computational comparisons and generated the figures. All authors actively participated in the writing and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

References

1. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus evolution 2021;7(2):veab064.
2. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology 2020;5(11):1403–1407.
3. consortium TCGUCU. An integrated national scale SARS-CoV-2 genomic surveillance network. The Lancet Microbe 2020;3(1):E99–E100.
4. Robishaw JD, Alter SM, Solano JJ, Shih RD, DeMets DL, Maki DG, et al. Genomic surveillance to combat COVID-19: challenges and opportunities. The Lancet Microbe 2021;2(9):e481–e484.
5. Oh DY, Hölzer M, Paraskevopoulou S, Trofimova M, Hartkopf F, Budt M, et al. Advancing Precision Vaccinology by Molecular and Genomic Surveillance of Severe Acute Respiratory Syndrome Coronavirus 2 in Germany, 2021; Clinical infectious diseases 2022;75(Supplement_1):S110–S120.
6. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance 2017;22(13):30494.
7. Jahn K, Dreifuss D, Topolsky I, Kull A, Ganesanandamoorthy P, Fernandez-Cassi X, et al. Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. Nature microbiology 2022;7(8):1151–1160.
8. Smyth DS, Trujillo M, Gregory DA, Cheung K, Gao A, Graham M, et al. Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. Nature communications 2022;13(1):1–9.
9. Agrawal S, Orschler L, Schubert S, Zachmann K, Heijnen L, Tavazzi S, et al. Prevalence and circulation patterns of SARS-CoV-2 variants in European sewage mirror clinical data of 54 European cities. Water research 2022;214:118162.
10. Peccia J, Zulli A, Brackney DE, Grubaugh ND, Kaplan EH, Casanovas-Massana A, et al. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nat Biotechnol 2020 Oct;38(10):1164–1167. https://www.nature.com/articles/s41587-020-0684-z.
11. Nemudryi A, Nemudraia A, Wiegand T, Surya K, Buyukyoruk M, Cicha C, et al. Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. Cell Reports Medicine 2020;1(6):100098.
12. Hoar C, McClary-Gutierrez J, Wolfe MK, Bivins A, Bibby K, Silverman AI, et al. Looking Forward: The Role of Academic Researchers in Building Sustainable Wastewater Surveillance Programs. Environmental Health Perspectives 2022;130(12):125002.
13. Amman F, Markt R, Endler L, Hupfauf S, Agerer B, Schedl A, et al. Viral variant-resolved wastewater surveillance of SARS-CoV-2 at national scale. Nature Biotechnology 2022;.
14. Gregory DA, Wieberg J C G adn Wenzel, Lin CH, Johnson MC. Monitoring SARS-CoV-2 Populations in Wastewater by Amplicon Sequencing and Using the Novel Program SAM Refiner. Viruses 2021;13(8).
15. Barbé L, Scaheffer J, Besnard A, Jousse S, Wurtzer S, Moulin L, et al. SARS-CoV-2 whole-genome sequencing using Oxford Nanopore Technology for variant monitoring in wastewaters. Frontiers in Microbiology 2022;p. 1362.
16. Agrawal S, Orschler L, Tavazzi S, Greither R, Gawlik BM, Lackner S. Genome Sequencing of Wastewater Confirms the Arrival of the SARS-CoV-2 Omicron Variant at Frankfurt Airport but Limited Spread in the City of Frankfurt, Germany, in November 2021. Microbiology Resource Announcements 2022;11(2):e01229–21.
17. Karthikeyan S, Levy JI, De Hoff P, Humphrey G, Birmingham A, Jepsen K, et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature 2022;609(7925):101–108.
18. Pechlivanis N, Tsagiopoulou M, Maniou MC, Togkousidis A, Mouchtaropoulou E, Chassalevris T, et al. Detecting SARS-CoV-2 lineages and mutational load in municipal wastewater and a use-case in the metropolitan area of Thessaloniki, Greece. Scientific reports 2022;12(1):1–12.
19. Valieris R, Drummond RD, Defelicibus A, Dias-Neto E, Rosales RA, Tojal da Silva I. A mixture model for determining SARS-Cov-2 variant composition in pooled samples. Bioinformatics 2022;38(7):1809–1815.
20. Ellmen I, Lynch MD, Nash D, Cheng J, Nissimov JI, Charles TC. Alcov: Estimating Variant of Concern Abundance from SARS-CoV-2 Wastewater Sequencing Data. medRxiv 2021;.
21. Barker DO, Buchanan CJ, Landgraff C, Taboada EN. MMMVI: Detecting SARS-CoV-2 Variants of Concern in Metagenomic Wastewater Samples. bioRxiv 2021;.
22. Schumann VF, de Castro Cuadrat RR, Wyler E, Wurmus R, Deter A, Quedenau C, et al. SARS-CoV-2 infection dynamics revealed by wastewater sequencing analysis and deconvolution. medRxiv 2022;.
23. Gafurov A, Baláž A, Amman F, Boršová K, Čabanová V, Klempa B, et al. VirPool: model-Based Estimation of SARS-CoV-2 Variant Proportions in Wastewater Samples. BMC Bioinformatics 2022;.
24. Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, Beerenwinkel N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. Bioinformatics 2021;37(12):1673–1680.
25. Baaijens JA, Zulli A, Ott IM, Nika I, van der Lugt MJ, Petrone ME, et al. Lineage abundance estimation for SARS-CoV-2 in wastewater using transcriptome quantification techniques. Genome Biology 2022;23(1):1–20.
26. Korobeynikov A. wastewaterSPAdes: SARS-CoV-2 strain deconvolution using SPAdes toolkit. bioRxiv 2022;.
27. Kayikcioglu T, Amirzadegan J, Rand H, Tesfaldet B, Timme RE, Pettengill JB. Performance of methods for SARS-CoV-2 variant detection and abundance estimation within mixed population samples. PeerJ 2023;11:e14596.
28. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature biotechnology 2016;34(5):525–527.
29. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nature Genetics 2021;53(6):809–816.
30. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nature biotechnology 2017;35(4):316–319.
31. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. Nature Methods 2023;.
32. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34(23):4121–4123.
33. Abdeldayem OM, Dabbish AM, Habashy MM, Mostafa MK, Elhefnawy M, Amin L, et al. Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook. Science of The Total Environment 2022;803:149834.

34. Agrawal S, Orschler L, Zachmann K, Lackner S. Comprehensive mutation profiling from wastewater in southern Germany extends evidence of circulating SARS-CoV-2 diversity beyond mutations characteristic for Omicron. FEMS Microbes 2023 03;4.
35. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 2012;6(2):80–92.

Supplement

Alternative allele frequency and size of reference database impact the *sequence-based* method but the effects are dependent on sample composition

Pan-EU-GER

Across all samples and experiments, we found the predictions of the *sequence-based* method to reflect the pandemic background in Germany well. Alpha and its sub-lineages were among the most prominent predictions within the time frame of wastewater sampling (Supplementary Figure S3). For most samples, we found Alpha and Q.1 to be the most abundant (sub-)lineages. The *sequence-based* method predicted distinctly varying abundances for sub-lineages other than Alpha, Beta, Gamma, or B.1.617 (summarized as "Other") across the *Pan-EU-GER* samples. We chose a cutoff of 1% abundance to differentiate true positive predicted lineages from false positive noise. On average, we found the pseudo-alignment-based approach to detect around 20–30% abundance of noise across all samples and parameter settings. At the minimum reference size (Supplementary Table S1), we observed for some samples a slightly decreasing amount of noise and a slightly increasing abundance for Alpha sub-lineages and "Others" when increasing the AAF threshold (e.g., sample INF_21051_D). We found, that the number of “Others” sub-lineages above 3% abundance decreased with increasing reference size. Across all experiments, we found the sample INF_21011_D to be the only one to be predicted with one or two “Others” sub-lineages of at least 3% abundance.

With increasing AAF threshold, we found distinct shifts in the estimated abundances for B.1.1.7 and Q.1. We observed those shifts to behave complementary but not consistently across all reference sizes: At the minimum reference size, we observed Alpha abundance predictions to distinctly increase and Q.1 abundances to decrease across all samples with increasing AAF threshold. Conversely, for reference size 5, we found Alpha abundance predictions to first increase and then decrease again with increasing AAF threshold. Vice versa, we observed Q.1 abundances to decrease and then increase again. At the largest reference size of 20 sequences per lineage, we observed a consistent decrease in Alpha abundance estimates and a consistent increase in Q.1 abundance estimates with increasing AAF threshold. Furthermore, we found abundances of other Alpha sub-lineages like Q.4 and Q.6 to also increase and decrease across varying parameter settings without following a clear pattern, but found the predicted abundances to not change as distinctly.

Overall, we found the performance of the *sequence-based* method to be mostly robust with varying settings for the AAF threshold and reference size. We observed the impact of those parameter changes to be stronger for more closely related lineages in a sample and in some cases to become weaker at larger reference sizes.

FFM-Airport

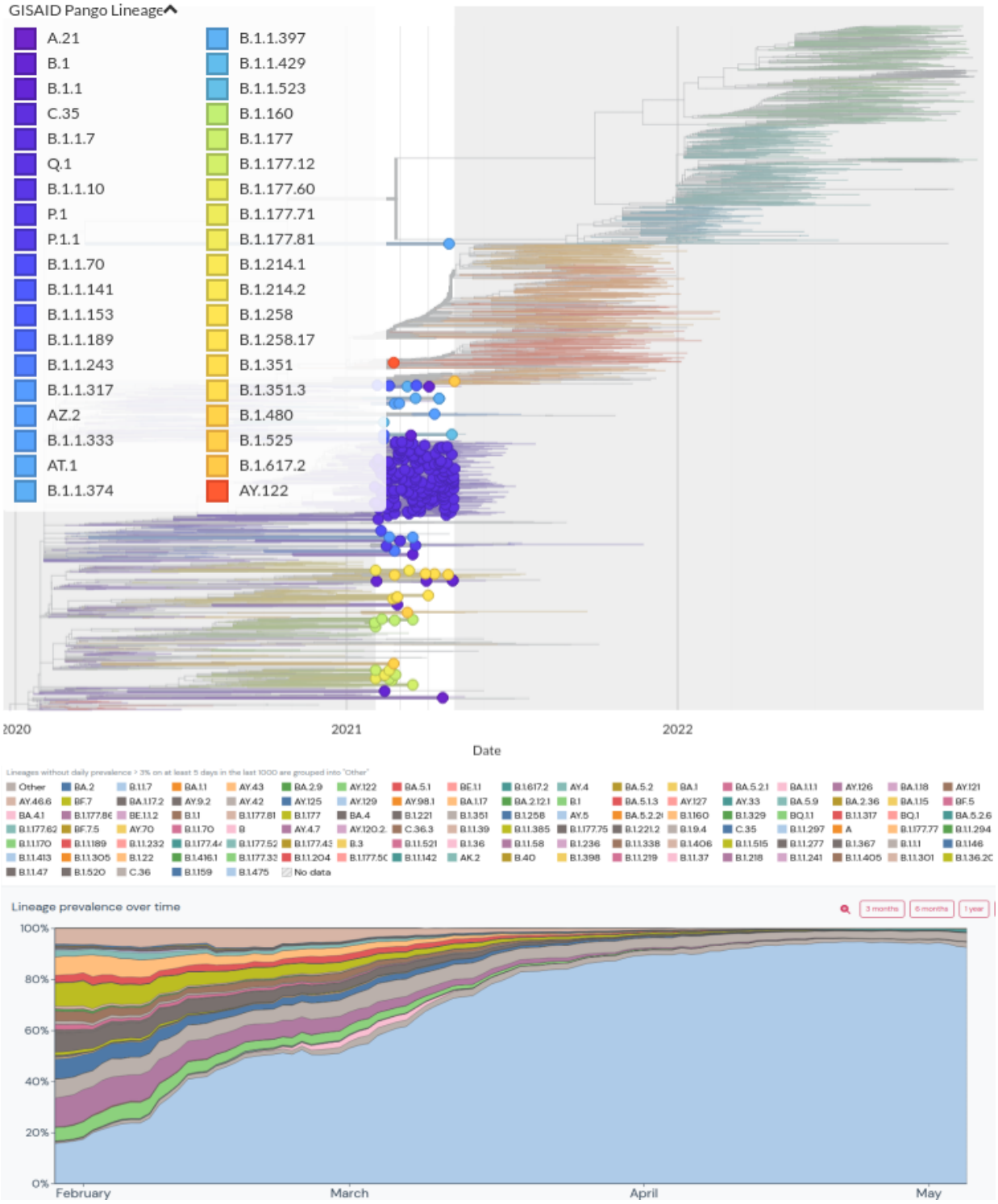
Across all parameter settings, the resulting abundance profiles for the *FFM-Airport* data set reflected the pandemic background in Europe and South Africa well around the time frame of wastewater sampling: the *sequence-based* method estimated Delta and its sub-lineages to represent the most abundant lineages and detected small proportions of Omicron (Supplementary Figure S4). We chose a cutoff of 1% abundance to differentiate true positive lineages from false positive noise and labelled sub-lineages with a minimum abundance of 3%. Because the *sequence-based* method detected Omicron sub-lineages at abundances below 3%, the quantified levels are not labelled and due to the scale of Supplementary Figure S4 not visible. However, when grouped by parent lineage, the predicted Omicron proportions become obvious. On average, we found the *sequence-based* method to detect around 50% abundance of noise across all parameter settings.

At the minimum reference size (SupplementaryTable S1), we observed a decreasing amount of low abundant noise with increasing AAF threshold. In contrast, with larger reference sizes, we found the amount of low abundant noise to change slightly and not follow a consistent pattern. Overall, we found the amount of noise to increase with increasing reference size. We observed the abundance estimates to increase for individual Delta sub-lineages with increasing AAF threshold. Specifically, we found the set of the most abundant Delta sub-lineages to change at every increase. Some examples for Delta sub-lineages that alternately were estimated among the most abundant lineages within a sample are AY.43.1 and AY.43.2, AY.43.3 and AY.42, and AY.121 and AY.122. When considering the lineage abundance profiles grouped by parent lineages, we found the predicted abundance profiles to not change distinctly across different parameter settings(Supplementary Figure S4.

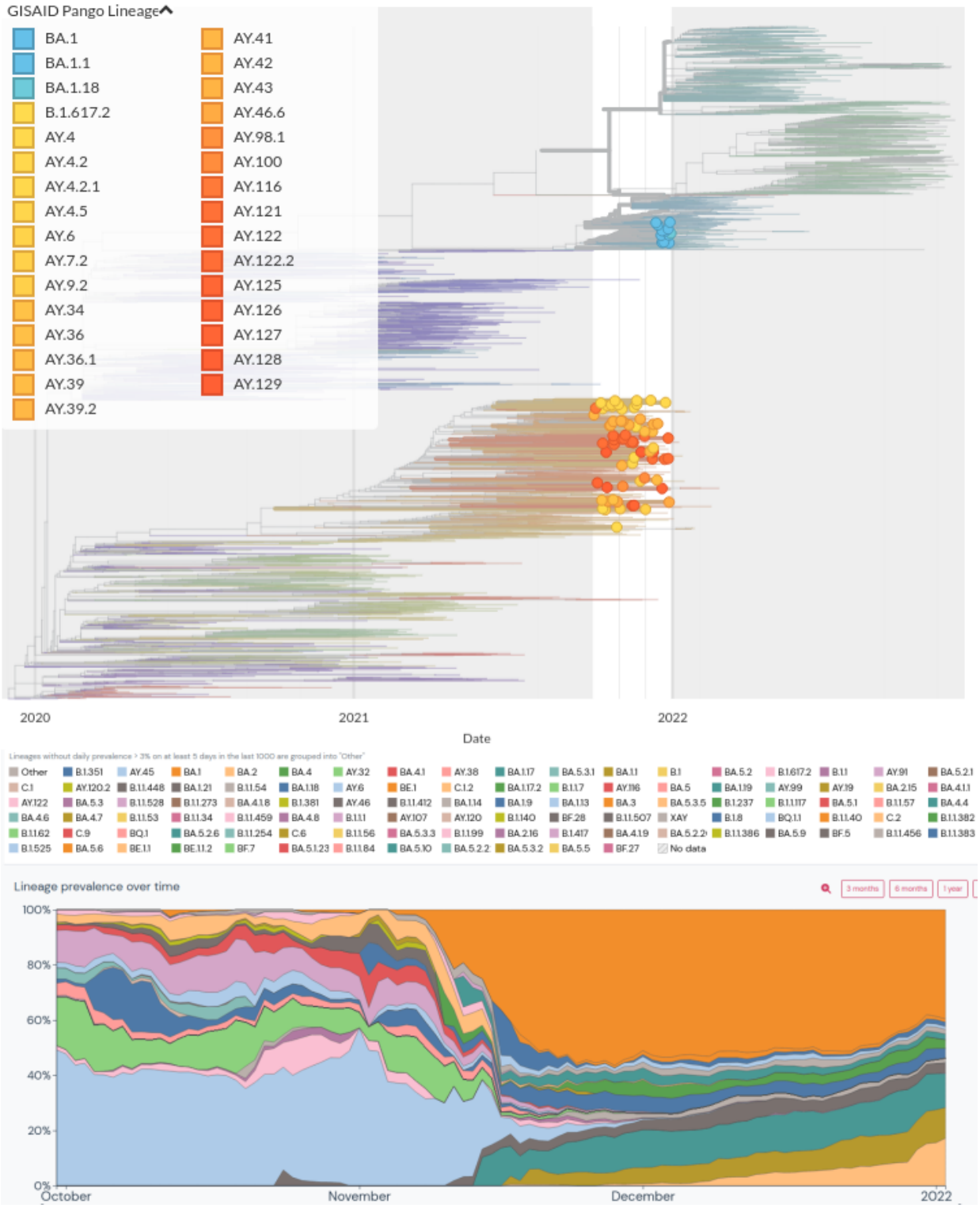
Finally, we found different settings for the AAF threshold and reference size to not distinctly affect the performance of the *sequence-based* method. We observed variations in the abundance estimates among multiple Delta sub-lineages.

Supplementary Table S1. Table showing the minimum reference sizes across the different alternative allele frequency (AAF) thresholds considered in the parameter escalation experiments across our three benchmark data sets. Here, we list the minimum number of genome sequences required per lineage to capture every mutation with an AAF above the considered AAF threshold at least once based on the implemented sampling strategy during reference construction. The *Standards* reference database required the largest number of sequences to capture the predefined genomic variation. Overall, we observed that with an increasing AAF threshold, the minimum reference sizes per lineage decreased across all three benchmark data sets.

AAF threshold	Minimum number of sequences per lineage		
	Standards	Pan-EU-Ger	FFM-Airport
0.25	20	3	3
0.5	10	2	2
0.85	10	1	1

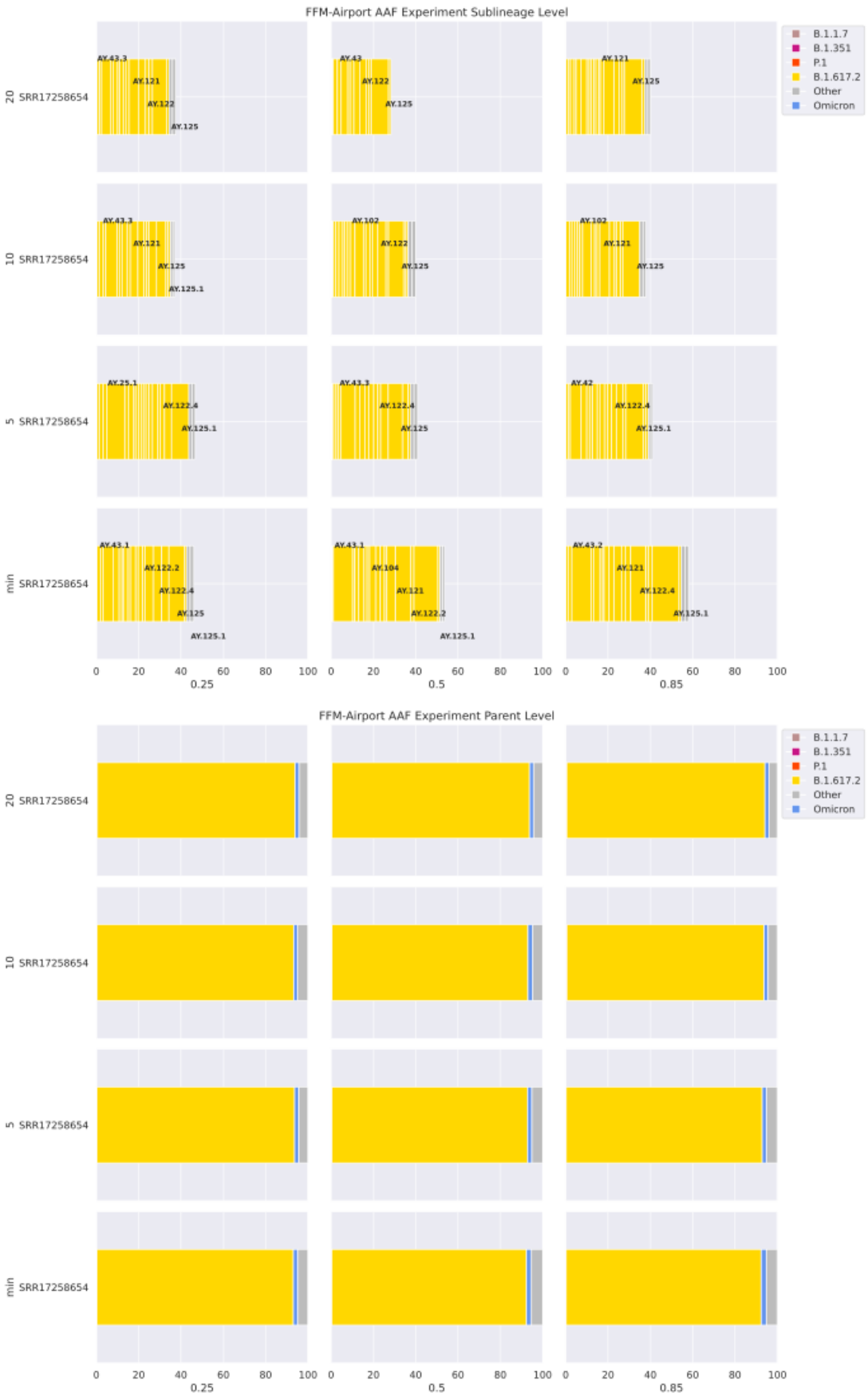


Supplementary Figure S1. Top: The pandemic background across Europe between 01 February and 30 April 2021 was built with Nextstrain.org. Bottom: The outbreak.org variant report for Germany displaying the SARS-CoV-2 lineage prevalence from February to March 2021 based on GISAID sequence data. The most dominant lineages in the plot from bottom to top: light blue = B.1.1.7, light green = B.1, purple = B.1.177.86, light grey = other, blue = B.1.258, dark grey = B.1.221, yellow = B.1.177, orange = B.1.160, light brown = B.1.177.81

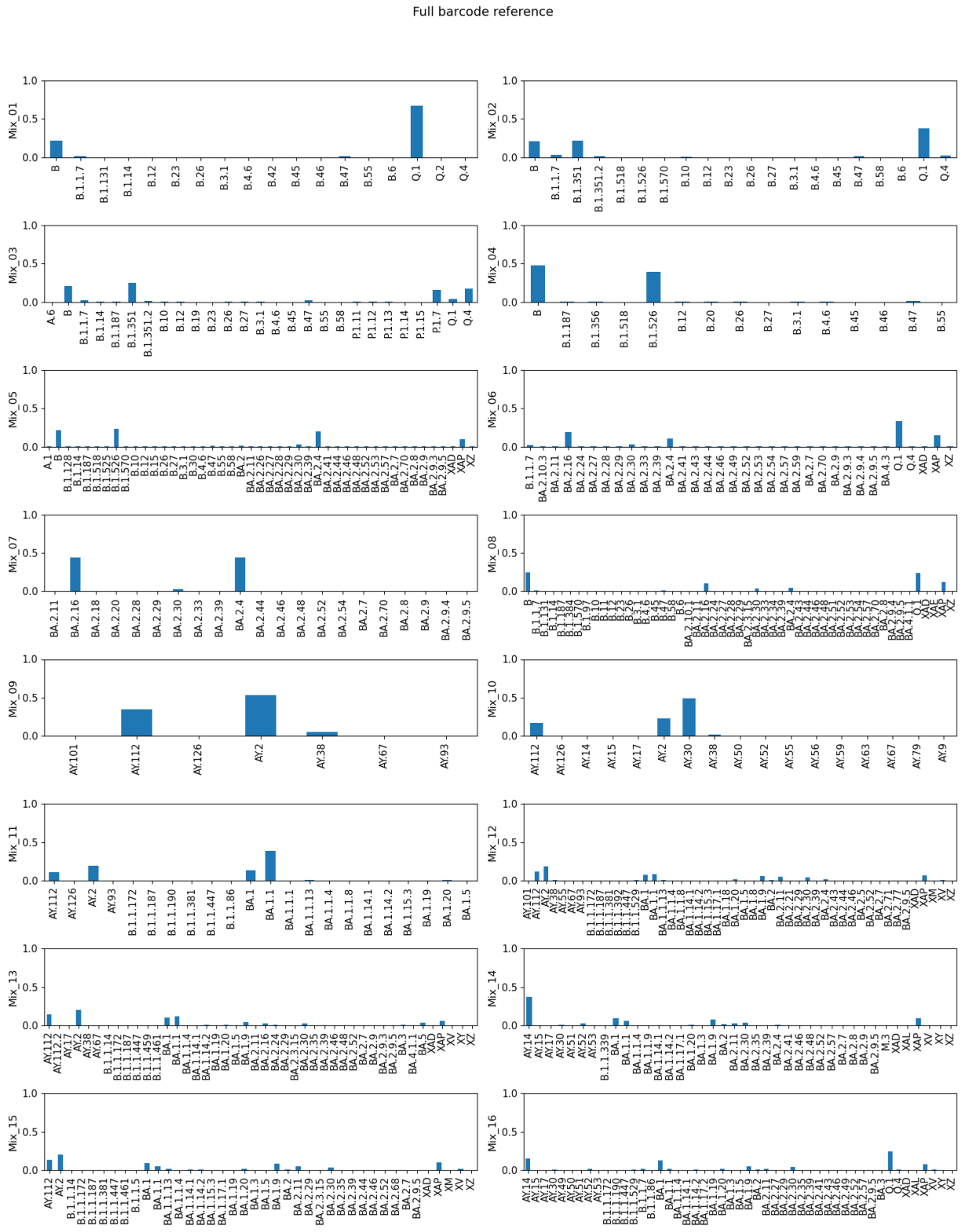


Supplementary Figure S2. Top: The pandemic background across Europe between 01 October and 31 December 2021 was built with Nextstrain.org. Bottom: The outbreak.org variant report for South Africa displaying the SARS-CoV-2 lineage prevalence from October to December 2021 based on GISAID sequence data. The most dominant lineages comprise sub-lineages of Delta and Omicron, but also B.1.351 (light green) and C.2 (light orange).

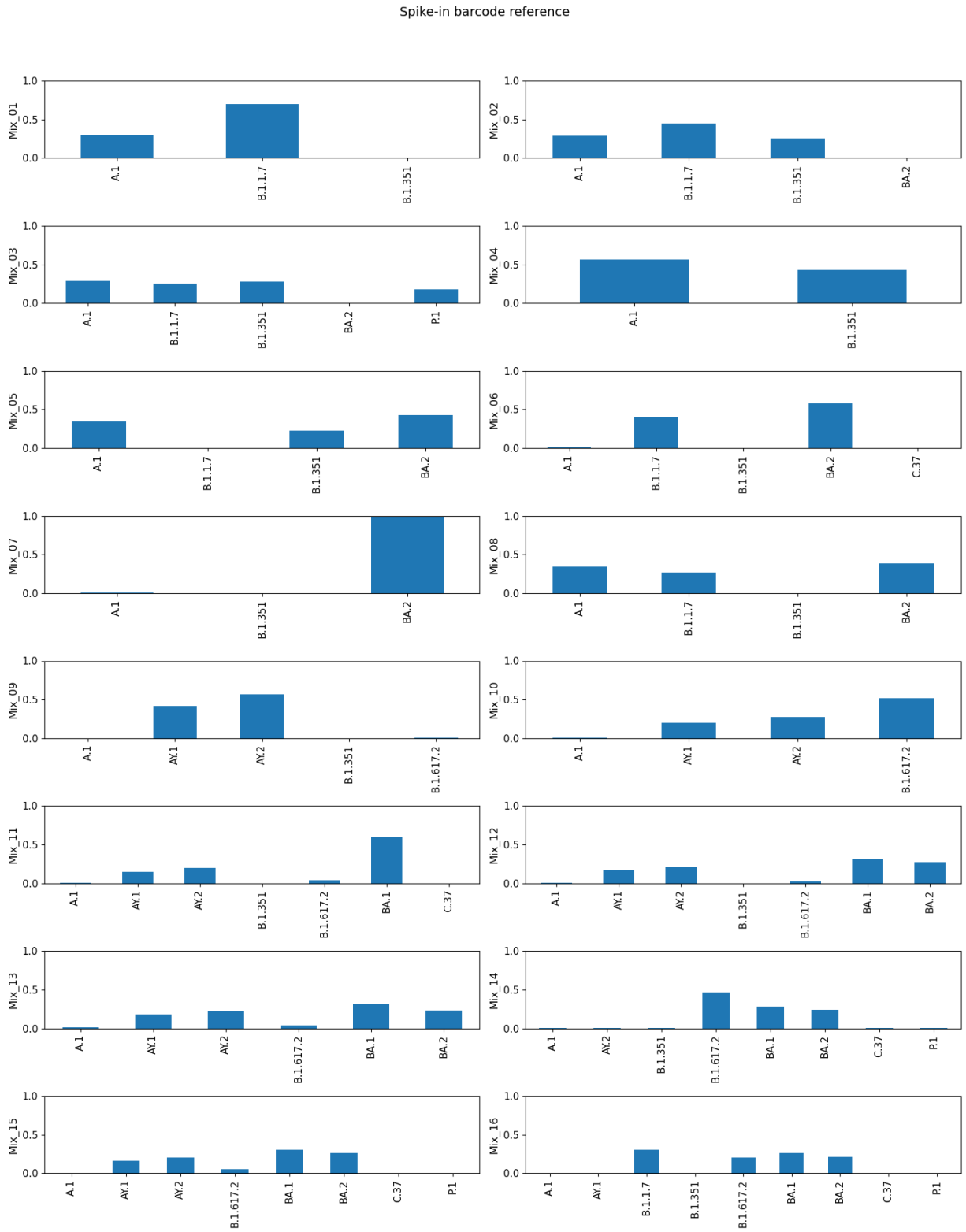
Supplementary Figure S3. Results for the parameter escalation experiments on the *Pan-EU-GER* samples using the *sequence-based* method using pseudo-alignment implementation. We analyzed the data set with different parameterization for reference construction (x-axis: increasing AAF threshold, y-axis: increasing maximum number of sequences per lineage). Abundance predictions are displayed at a minimum threshold of 1% and labelled at a threshold of 3%. When comparing with the pandemic background at the time of wastewater sampling, we observed the AAF threshold and the maximum number of sequences per lineage to impact the abundance proportions among Alpha and Q.1 the most. With more sequences per lineage in the reference, we found the impact of the AAF filter on the observed ambiguities to decrease. We found significantly more low abundant sub-lineages predicted in the real wastewater data compared to the *Standards* data set and found those low abundant predictions to mostly not change distinctly across varying parameterization.



Supplementary Figure S4. Results for the parameter escalation experiments on the *FFM-Airport* data set using the *sequence-based* method. We analyzed the data set with different parameterization for reference construction (x-axis: increasing AAF threshold, y-axis: increasing maximum number of sequences per lineage). **Top**: Abundance predictions are displayed at a minimum threshold of 1% abundance and labelled at a threshold of 3% abundance. When comparing with the pandemic background at the time of wastewater sampling, we observed the following: Overall, we found more sub-lineages predicted with abundance below 1% compared with the *Standards* data set and the *Pan-EU-GER* set. The *sequence-based* method detected more low abundant sub-lineages with increasing reference size and slightly less low abundant sub-lineages with increasing AAF threshold. Both the AAF threshold and the reference size showed to impact lineage ambiguities among Delta sub-lineages. **Bottom**: All abundance predictions are displayed as grouped by their parent lineage. We did not find the abundance predictions for parent lineages to change distinctly across experiments.



Supplementary Figure S5. SARS-CoV-2 lineage abundance assignments via Freyja [17] (v1.3.12) for the *Standards*. We used the full reference UShER set as provided as a default by the tool. In this case, multiple sub-lineages were predicted and frequencies were distributed among them, resulting in a reduced frequency estimate for the true (parental) lineage and an increase in low-frequency detections. For example, in Mix_07 the sub-lineages BA.2.16 and BA.2.4 were predicted with almost 50 %, respectively, while the included lineage BA.2 was not assigned (compare Figure S6).



Supplementary Figure S6. SARS-CoV-2 lineage abundance assignments via Freyja [17] (v1.3.12) for the *Standards*. We reduced the reference UShER set to the lineages part of our artificial mixtures, instead of using the full UShER barcode data set as shown in Figure S5.

Supplementary Table S2. Table summarizing the mapping of each *Standards* sample.

SampleID	Total number of reads	Number of mapped reads	Average target base coverage depth*
Mix_01	11585401	11409164	67129
Mix_02	8000781	7872877	45648
Mix_03	7182909	7082168	41674
Mix_04	11156327	11033124	65477
Mix_05	9509819	9358747	54475
Mix_06	12228003	12005442	69829
Mix_07	11227258	10999971	63366
Mix_08	7991289	7886999	45904
Mix_09	22189148	21855903	21855903
Mix_10	8685555	8614170	8614170
Mix_11	2564976	2535182	2535182
Mix_12	2581594	2557744	2557744
Mix_13	8429568	8295217	8295217
Mix_14	6246713	6173867	6173867
Mix_15	11888445	11752739	11752739
Mix_16	6139010	6092648	6092648

*Target sequence was the SARS-CoV-2 reference genome (Wuhan-Hu-1)