

Afanc: taxonomic deconvolution and probabilistic lineage assignment for pathogen genomic surveillance

Arthur V. Morris

`morrisa28@cardiff.ac.uk`

Cardiff University

Anna Price

Cardiff University

Thomas Connor

Cardiff University

software

Keywords: pathogen genomics, genomic surveillance, taxonomic classification, lineage assignment, Bayesian classification, Kraken 2, software

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10081230/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Afanc: taxonomic deconvolution and probabilistic lineage assignment for pathogen genomic surveillance

Arthur V. Morris^{1*}, Anna Price¹ and Thomas Connor^{1,2}

^{1*}Cardiff School of Biosciences, Cardiff University, Cardiff, Wales,
United Kingdom.

²Public Health Genomics, Public Health Wales, Cardiff, Wales, United
Kingdom.

*Corresponding author(s). E-mail(s): morrisa28@cardiff.ac.uk;

Abstract

Background: Reliable pathogen genomic surveillance requires both broad taxonomic sensitivity and high-resolution lineage or clade assignment, yet existing tools usually address only one of these tasks. We developed Afanc as an extensible framework that connects database construction, taxonomic deconvolution and probabilistic lineage assignment from pathogen sequencing data. Afanc source code is available at <https://github.com/ArthurVM/Afanc>.

Results: Afanc integrates automated reference database construction, reference-quality and representativeness filtering, Kraken 2 report deconvolution, competitive mapping, variant calling and Bayesian SNP-based classification. We evaluated Afanc across Mycobacteriaceae species and subspecies assignment, mixed Mycobacteriaceae datasets, *Mycobacterium tuberculosis* lineage classification, *Plasmodium falciparum* geolinesage prediction and Mpox lineage classification. Afanc resolved Mycobacteriaceae taxa across all simulated coverage cohorts, including 1% genome breadth, and recovered all expected taxa from mixed datasets. For *M. tuberculosis*, Afanc assigned 955/958 samples at harmonised resolution with no wrong-branch calls. For *P. falciparum*, Afanc assigned 18,830/20,712 samples to regional lineages, outperforming Malaria-Profiler by 15.8 percentage points. For Mpox, Afanc produced no wrong-branch assignments and achieved 99.4% metadata-resolution-harmonised classification.

Conclusions: Afanc provides a reusable framework for moving from noisy broad taxonomic evidence to interpretable lineage-level inference across diverse pathogen systems. By linking database curation, taxonomic deconvolution and probabilistic marker-based classification, Afanc supports pathogen genomic surveillance workflows in which mixed taxonomic signal, incomplete evidence and lineage ambiguity must be handled explicitly before downstream epidemiological or evolutionary interpretation.

Keywords: pathogen genomics, genomic surveillance, taxonomic classification, lineage assignment, Bayesian classification, Kraken 2, software

1 Background

Taxonomic and lineage profiling from pathogen sequencing data is a core task in clinical genomics, public health surveillance, and infectious disease research. Accurate assignment informs downstream inference about transmission, pathogenicity, drug resistance, and outbreak structure, and increasingly underpins routine analytical workflows in diagnostic and reference laboratories. However, robust classification remains difficult when sequencing data are low coverage, taxonomically complex, compound, or dominated by closely related lineages whose genomic boundaries are weakly separated.

Current profiling approaches largely fall into two categories. The first comprises genomic distance or sequence content-based methods, in which reads are assigned by comparison to reference genomes or k-mer databases. Tools such as Kraken 1 and 2 [1, 2], and derivative approaches such as Bracken [3] and KrakenUniq [4], have proven highly effective for rapid, broad-spectrum screening and decomposition of metagenomic read sets. Yet these methods are inherently constrained by the limited specificity of short reads and by the degree of separation between competing taxa in sequence space. When related taxa occupy highly overlapping sequence space, report

complexity increases and false-positive low-level assignments become difficult to interpret. As a result, such approaches are often strongest at species or subspecies level, but less reliable for higher-resolution discrimination among closely related lineages or clades.

The second class comprises marker-based approaches, in which taxa or lineages are identified using predefined sets of canonical mutations. This strategy has been successful in well-characterised pathogen systems, where stable lineage-defining variants can be catalogued and screened directly from sequencing data. Tools such as Mykrobe [5], TB-Profiler [6], and Malaria-Profiler [7] illustrate the power of this approach for high-resolution classification within individual pathogen groups. However, marker-based profiling also has clear limitations. It depends on the availability and maintenance of curated marker sets, assumes that classes are sufficiently well separated by those markers, and can perform poorly when marker evidence is incomplete, overlapping, or hierarchically ambiguous. Hard-thresholds of vote-based systems are particularly vulnerable in settings where neighbouring classes are genomically contiguous and uncertainty is biologically real rather than purely technical.

These limitations motivate a hybrid approach. Broad taxonomic screening benefits from database-driven methods that retain speed and scope across diverse organisms, whereas fine-scale lineage or clade assignment often requires targeted inference within a narrower candidate space. A general-purpose profiler should therefore integrate both levels of evidence: first resolving the most plausible taxonomic background from a broad reference database, then performing higher-resolution classification using informative marker variation within the identified taxon. Just as importantly, such a framework should be able to represent ambiguity explicitly, rather than forcing overconfident assignments when competing classes are weakly separated.

139 In genomic surveillance, taxonomic and lineage classification are not endpoints
140 in themselves. They provide an interpretive evidence layer required for pathogen
141 tracking, allowing epidemiologists and analysts to identify which organism is present,
142 determine whether a sample belongs to a recognised lineage or clade, distinguish
143 mixed or low-confidence signal from robust assignments, and decide whether down-
144 stream epidemiological or evolutionary interpretation is justified. Surveillance systems
145 therefore require classification frameworks which can operate across multiple pathogen
146 taxa, support custom databases, and handle incomplete evidence by reporting uncer-
147 tainty instead of forcing brittle classification calls.
148

155 Here we present Afanc, a general-purpose profiler for species, subspecies, clade, and
156 lineage level classification from pathogen NGS data. Afanc addresses five linked sources
157 of error in pathogen profiling: low-quality reference sequence inclusion during database
158 construction, noisy low-level assignments in taxonomic reports, loss of true taxa during
159 aggressive report filtering, incomplete signal in low-coverage or mixed samples, and the
160 need to select organism-specific lineage models after taxonomic classification. Afanc
161 combines a Kraken 2 report deconvolution strategy for broad taxonomic profiling with
162 a sparse empirical-Bayes marker classification framework for higher-resolution infer-
163 ence. This design is intended to retain the breadth and scalability of database-driven
164 taxonomic screening while improving sensitivity and interpretability at finer taxonomic
165 levels. We evaluate Afanc across three contrasting pathogen systems with distinct clas-
166 sification challenges: *Mycobacteriaceae*, representing closely related bacterial species
167 and deep lineage structure; *Plasmodium falciparum*, representing geographically struc-
168 tured populations with substantial overlap between neighbouring regional groups; and
169 the Mpox virus, representing viral clade and sublineage discrimination in an epidemic
170 setting. Together, these test cases assess whether a single profiling framework can
171 generalise across divergent pathogen classes and epidemiological contexts.
172
173
174
175
176
177
178
179
180
181
182
183
184

2 Implementation	185
	186
2.1 Software architecture and modules	187
	188
Afanc consists of four discrete sub-tools: <code>get_dataset</code> , <code>autodatabase</code> , <code>screen</code> , and	189
<code>classify</code> .	190
	191
	192
<i>get_dataset</i>	193
	194
This module downloads a dataset of genome assemblies from GenBank belonging to	195
species defined within a flat text file. Genome assemblies are downloaded and deposited	196
in a directory structure consisting of a parent directory containing subdirectories for	197
each species. Subspecies and variants specified within the text file are deposited in	198
their own subdirectory within their parent species directory. This directory structure	199
can be used as input for the <code>autodatabase</code> module (see Figure S1 of the supplementary	200
materials).	201
	202
	203
	204
	205
	206
	207
<i>autodatabase</i>	208
	209
The <i>autodatabase</i> module converts a user-supplied collection of genome assemblies	210
into the database used by <i>Afanc screen</i> (Figure 1). The workflow begins with tax-	211
onomy database setup (step 1). Afanc either downloads the NCBI taxonomy for the	212
user-specified date, or uses a local NCBI taxdump if provided. Afanc reconciles the	213
input directory structure with this taxonomy and records mappings between input	214
taxa, names, and taxIDs. If an input taxon is not present in the selected NCBI tax-	215
onomy, Afanc attempts to resolve it using available scientific names, authority names,	216
and synonyms at species level. Where no suitable NCBI node exists, a simulated node	217
is added under the corresponding known genus. This process fails if the taxon cannot	218
be assigned to a known genus.	219
	220
	221
	222
	223
	224
	225
	226
	227
	228
	229
	230

231 Afanc then builds a Mash-based distance index from the curated assembly set
 232 (step 2) [8]. Pairwise Mash distances are calculated from genome sketches, where the
 233 distance between assemblies n and m is
 234

$$235 D_{nm} = -\frac{1}{k} \ln \left(\frac{2j_{nm}}{1 + j_{nm}} \right), \quad (1)$$

241 where k is the k -mer size and j_{nm} is the Jaccard estimate between sketches. These
 242 distances are stored as an all-vs-all Mash matrix, along with genus- and species-level
 243 distance summaries and a variant index used by *Afanc screen*.
 244

245 The same Mash distance information is used to create taxon panels (step 3). For
 246 each taxon, Afanc constructs a within-taxon Mash matrix and calculates the average
 247 Mash distance for each assembly relative to the other assemblies assigned to that
 248 taxon. Assemblies whose average distance lies within a user-defined range around the
 249 modal average distance are retained as representative, high-quality members of the
 250 taxon panel. By default, this capture range is 0.1. This step is intended to remove
 251 outlying or low-quality assemblies before database construction, while retaining rep-
 252 resentative diversity within each taxon. The retained assemblies are written to the
 253 curated FASTA directory and recorded in the database FASTA mapping JSON. A
 254 Kraken 2 database is then constructed from the retained taxon-panel assemblies (step
 255 4).
 256

257 Finally, Afanc generates a Krona chart from the Kraken 2 database inspection
 258 output (step 5). This provides a visual summary of the taxonomic content of the
 259 constructed database. The completed Afanc database contains the Kraken 2 index,
 260 curated reference FASTAs, taxonomy mappings, Mash distance matrices, and the
 261
 262
 263
 264
 265
 266
 267
 268
 269
 270
 271
 272
 273
 274
 275
 276

Mash variant index used by <i>Afanc screen</i> for taxonomic deconvolution and elastic	277
variant-level thresholding.	278
	279
	280
<i>screen</i>	281
	282
The <i>screen</i> module profiles an input read set against an Afanc database (Figure 1).	283
	284
It takes paired FASTQ files and a database generated by <i>Afanc autotoolkit</i> , and	285
	286
produces a JSON report containing taxonomic assignments, mapping statistics, SNP	287
	288
profiles, and lineage classification results where suitable models are available.	289
	290
	291
The workflow consists of five main stages. First, Afanc runs Kraken 2 against	292
	293
the database to obtain an initial taxonomic report (step 1). The Kraken 2 report	294
	295
is then parsed and deconvolved on the report tree to identify supported taxa and	296
	297
to redistribute ambiguous reads from unsupported nodes to supported neighbouring	298
	299
taxa where appropriate (step 2). This produces the taxonomic assignments used	300
	301
for downstream analysis, including species-level calls with nested variant/subspecies	302
	303
assignments where present. The deconvolution procedure is described below.	304
	305
	306
Reads are then competitively mapped to the reference assemblies associated with	307
	308
the detected taxa (step 3). This produces per-reference BAM files and mapping	309
	310
statistics, with reads partitioned according to their best mapping target. Afanc then	311
	312
calls SNPs from the relevant BAM files using the selected variant caller (step 4). If	313
	314
a detected taxon has an associated lineage model, the corresponding SNP profile is	315
	316
passed to the Bayesian lineage classifier (step 5). For mixed datasets, lineage clas-	317
	318
sification is skipped by default unless explicitly enabled, to avoid applying lineage	319
	320
models to confounded multi-taxon read sets. Finally, Afanc combines the taxonomic	321
	322
assignment, mapping, SNP, and lineage outputs into the final JSON report.	

323 ***classify***
324
325 The *classify* module applies Afanc lineage models to existing variant evidence without
326 running the full *screen* workflow. It accepts a VCF generated against the reference
327 FASTA associated with a supported lineage model. A supplied reference taxon
328 (species or clade for which a lineage model exists) name is resolved against the profile
329 manifest, after which Afanc checks that a matching model is available and validates
330 reference compatibility before classification. The validated evidence is passed to the
331 shared Bayesian lineage classifier, producing the same lineage classification output
332 used by *screen*.
333
334
335
336
337
338
339
340

341 **2.2 Kraken 2 report deconvolution**

342
343 Kraken 2 produces an extremely complex and potentially ambiguous metagenomic
344 report, which details every possible match for the read set (see Figure S3). Inter-
345 preting this report can be extremely challenging, particularly in instances where the
346 dataset is compound, consisting of multiple species and/or variants. This is illustrated
347 in Supplementary Figure S3, where every species which has at least 1 read attributed
348 to it is reported. To solve this problem, a novel algorithm was developed to optimally
349 identify the most likely species and variants present in a read set.
350
351
352
353
354

355
356 A Kraken 2 report is a hierarchical tree where nodes refer to individual taxa,
357 topology is determined by taxonomic relations, and node weight is defined by the
358 percentage of reads which were assigned to the taxon rooted at this node. Identifying
359 key nodes within this tree must be carried out in a sensitive manner.
360
361
362
363
364

365 First-pass screening takes place by identifying species-level or higher nodes which
366 exceed a user defined global threshold (by default, this is 5.0%). This global threshold
367
368

represents the minimum percentage of reads which must be attributed to a particular
clade to consider it a putative hit. The branches rooted at each of these nodes are
then traversed to find the maximally scoring tip nodes. A node is considered a hit if it
exceeds a local elastic threshold. The tree is then subjected to the evidence-weighted
read redistribution algorithm, and the scoring hit nodes identified in the previous step
are reassessed to find the maximally scoring node.

2.3 Elastic threshold calculation

The elastic threshold was calculated from parent-child similarity values stored in the
variant index. Let T denote the Kraken 2 report tree. For a below-species candidate
node i , let $p(i)$ denote its parent node, let C_i denote the clade-level read support for
the subtree rooted at i , and let $R_{p(i)}$ denote the number of reads assigned directly
to the parent node. Below-species nodes are evaluated only when their species-level
ancestor has already exceeded the global reporting threshold.

For node i , the variant index stores a mean parent-child similarity $S_{p(i),i}$ and a
mean within-child similarity $S_{i,i}$. These values are stored on a 0–100 similarity scale.
Afan converts the parent-child similarity to a non-negative distance,

$$D_{p(i),i} = \max(0, 100 - S_{p(i),i}).$$

The elastic threshold weight is calculated using a saturating distance transform:

$$w_i = 1 - \exp\left(-\frac{D_{p(i),i}}{\lambda}\right),$$

415 where λ is the elastic distance scale. The local read-count threshold for retaining node
 416 i is then

$$418 \quad T_i = R_{p(i)} w_i = R_{p(i)} \left(1 - \exp \left(-\frac{D_{p(i),i}}{\lambda} \right) \right).$$

419
 420 The direct parent read count $R_{p(i)}$, rather than the parent clade count, is used as
 421 the denominator for this threshold. This prevents abundant child calls from increasing
 422 their own acceptance threshold through their contribution to the parent clade total.
 423
 424

425
 426
 427 A below-species candidate node is retained when

$$428 \quad C_i \geq T_i.$$

429
 430
 431 If node i is absent from the variant index, Afanc uses a conservative fallback weight,
 432
 433

$$434 \quad w_i = 0.1,$$

435
 436
 437 giving

$$438 \quad T_i = 0.1 R_{p(i)}.$$

439
 440 In all cases, the threshold is bounded below by one read.
 441
 442

443 2.4 Evidence-weighted read redistribution

444
 445 When a Kraken 2 database contains many closely related taxa, reads may be assigned
 446 to low-support nodes that are more plausibly explained as ambiguous assignments
 447 arising from neighbouring taxa. Afanc therefore performs a report-level redistribution
 448 of reads on the Kraken 2 report tree. The procedure is local to each family-level
 449 subtree and redistributes only the direct reads assigned to nodes that are not retained
 450 as supported calls.
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460

Let T denote the Kraken 2 report tree. For a node i , let C_i denote the clade-level read support for the subtree rooted at i , and let R_i denote the reads assigned directly to node i . Let N be the total number of classified reads and let τ be the global reporting threshold. Genus- and species-level nodes are retained as supported calls when

$$\frac{C_i}{N} \geq \tau. \quad (2)$$

Below species level, nodes are not assessed directly against the global threshold. Instead, a below-species node i is considered only if its species-level ancestor is already supported, and is retained when its clade support exceeds an elastic parent-child threshold:

$$C_i \geq T_i, \quad (3)$$

where

$$T_i = R_{p(i)} \left(1 - e^{-D(p(i),i)/\lambda} \right). \quad (4)$$

Here $p(i)$ is the parent of node i , $R_{p(i)}$ is the direct read support assigned to the parent node, $D(p(i),i)$ is the parent-child distance derived from the Mash/variant index, and λ is the elastic distance scale. If no parent-child distance is available, Afanc uses a conservative fallback threshold proportional to the direct support of the parent.

Let S_f be the set of supported nodes within a family-level subtree f . A node $d \in f$ is treated as a donor if it is not in S_f but has direct read support:

$$d \notin S_f, \quad R_d > 0. \quad (5)$$

507 Only the direct reads R_d assigned to the donor node are redistributed. Supported
 508 descendants of an unsupported internal node are not zeroed or removed.
 509

510
 511
 512 For each donor d , Afanc identifies the nearest ancestor $a(d)$ whose subtree contains
 513 at least one supported node. Reads are then routed downward from $a(d)$ through child
 514 branches that contain supported descendants. At each routing step, let $\mathcal{B}(x)$ be the
 515 set of child branches below the current node x that contain at least one supported
 516 node. For a candidate branch $b \in \mathcal{B}(x)$, the redistribution weight is
 517
 518
 519

$$w_b = \Delta(d, b)^\alpha C_b^\beta, \quad (6)$$

520
 521
 522
 523 where C_b is the original clade support for branch b , $\Delta(d, b)$ is the Mash-derived
 524 similarity between donor d and branch b , α controls the strength of Mash similarity
 525 weighting, and β controls the strength of read-evidence weighting. Mash similarity is
 526 defined as
 527
 528
 529
 530
 531
 532

$$\Delta(d, b) = 1 - D^{\text{Mash}}(d, b). \quad (7)$$

533
 534
 535
 536
 537
 538
 539
 540 Where direct donor-branch distances are not available, Afanc uses the best avail-
 541 able descendant-pair similarity between the donor subtree and recipient branch. If no
 542 distance is available, an evidence-only fallback can be used, in which case weights are
 543 proportional to C_b^β .
 544
 545
 546

547 The number of reads routed from donor d into branch b is
 548
 549
 550
 551
 552

$$r_b = R_d \frac{w_b}{\sum_{h \in \mathcal{B}(x)} w_h}. \quad (8)$$

This routing is applied recursively until no lower supported child branch remains. If the current node is itself supported, the remaining redistributed reads are assigned to that node. If no supported routing path exists within the family-level subtree, the donor reads are left unassigned rather than being forced into an unsupported recipient.

After redistribution, Afanc recalculates clade-level read counts from the updated direct node assignments and writes both a filtered Kraken-like report and an audit record describing donor nodes, routing paths, redistribution weights, assigned reads, and any unassigned reads.

2.5 Lineage model architecture and classification

Lineage assignment was performed using a Bayesian SNP-based classifier operating on two related model families: canonical-epsilon-Bayesian (CEB) models and semi-Bayesian (SB) empirical models. In both cases, each locus i is represented by a lineage-associated emission probability, and sample evidence is encoded as a binary observation indicating presence or absence of the lineage-defining allele.

For a candidate lineage L , let $x_i \in \{0, 1\}$ denote the observed state at locus i , where $x_i = 1$ indicates presence of the lineage-defining allele and $x_i = 0$ indicates absence. Over the callable locus set $\mathcal{C}$, the lineage likelihood is computed under a Bernoulli model:

$$P(\mathbf{x} \mid L) = \prod_{i \in \mathcal{C}} \left(p_i^{(L)} \right)^{x_i} \left(1 - p_i^{(L)} \right)^{1-x_i}. \quad (9)$$

599 Posterior probabilities are then obtained by normalising across candidate lineages:

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

$$P(L \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid L) P(L)}{\sum_{L'} P(\mathbf{x} \mid L') P(L')}, \quad (10)$$

where $P(L)$ is the prior probability assigned to lineage L .

The distinction between CEB and SB lies in the definition of the emission probabilities. In CEB models, canonical lineage markers are converted into pseudo-empirical emissions using a small error parameter ϵ , such that

$$p_i^{(L)} \approx 1 - \epsilon \quad \text{for the target lineage,} \quad (11)$$

$$p_i^{(L')} \approx \epsilon \quad \text{for non-target lineages.} \quad (12)$$

In SB models, the emission probabilities are estimated from empirical lineage-specific allele frequencies and therefore need not follow these near-binary canonical expectations. Thus, both CEB and SB use the same Bayesian classification framework, but differ in the origin of the locus-specific emission parameters.

Flat models evaluate all candidate lineages jointly and produce a global posterior distribution over the lineage set. Hierarchical models are currently applied only to CEB classification. In this case, classification proceeds top-down through the lineage tree, comparing sibling lineages at each internal node using their direct marker sets. The hierarchical procedure therefore yields a sequence of local sibling-level posterior comparisons rather than a single global posterior over all terminal lineages.

Missing data are handled explicitly. If the genomic position corresponding to locus i is uncallable, then $i \notin \mathcal{C}$ and that locus is excluded from likelihood evaluation rather

than treated as negative evidence. For each lineage decision, the classifier records the number of configured markers, callable markers, matched markers, absent callable markers, and missing markers:

$$n_{\text{call}} = n_{\text{cfg}} - n_{\text{miss}}, \quad (13)$$

$$n_{\text{abs}} = n_{\text{call}} - n_{\text{match}}. \quad (14)$$

Support among callable loci is summarised as

$$s = \frac{n_{\text{match}}}{n_{\text{call}}}, \quad (15)$$

and marker completeness is defined as

$$c = \frac{n_{\text{call}}}{n_{\text{cfg}}}. \quad (16)$$

During hierarchical CEB traversal, a child lineage is accepted only if its callable support exceeds the configured minimum decision threshold. If no child lineage satisfies this criterion, traversal stops at the current node. Internal nodes with no direct markers are treated as transparent and may be traversed if a descendant lineage is supported by downstream evidence. If no lineage can be accepted, the sample is reported as **unclassified**. In that case, the classifier also records the closest lineage, defined as the highest-ranked lineage at the stopping point with

$$n_{\text{match}} > 0, \quad (17)$$

thereby preserving sub-threshold lineage signal without promoting it to a formal lineage assignment.

691 2.6 Afanc in an operational context

692
693 Afanc is intended for command-line use in pathogen genomics, reference-laboratory
694 and research workflows. The `screen` module produces a structured JSON report
695 containing taxonomic assignments, mapping statistics, SNP profiles and lineage clas-
696 sifications where suitable models are available. This output format is intended to
697 support downstream integration into reports, dashboards, surveillance databases or
700 bespoke laboratory workflows. The `classify` module allows existing VCF datasets to
701 be analysed without rerunning read-level screening, supporting retrospective analyses
702 of large public or locally curated variant datasets.
703
704
705
706

707 3 Benchmark design and evaluation datasets

708 3.1 Reference database and lineage model construction

709
710
711 Afanc was evaluated using a single multi-pathogen reference database for broad
712 read-level taxonomic screening, together with pathogen-specific marker catalogues for
713 downstream lineage or clade assignment. All benchmark datasets supplied as FASTQ
714 reads were screened against this same database. The *P. falciparum* Pf7 evaluation
715 used precomputed VCF files and therefore entered the workflow through the *classify*
716 module at the lineage-classification stage.
717
718
719
720
721
722

723
724 The candidate reference collection comprised 821 FASTA sequences assigned to
725 193 input taxon or lineage labels. The *autodatabase* representativeness filter retained
726 330 sequences assigned to 191 taxonomic or lineage nodes, while 491 candidate
727 sequences (59.8%) were not retained because they fell outside the configured Mash-
728 distance capture range for their respective taxon panels. Exclusion at this stage
729 indicates that a sequence was considered atypical or redundant relative to the retained
730
731
732
733
734
735
736

panel and does not necessarily imply poor assembly quality. The final database contained 210 Mycobacteriaceae reference sequences across 174 nodes, 54 *Plasmodium* reference sequences across 13 nodes, and 66 mpox virus reference sequences across four clade-level nodes.

The Mycobacteriaceae component provided broad species- and subspecies-level representation and included custom nodes for *Mycobacterium tuberculosis* complex lineages 1, 2, 3, 4, 5, 6, 7, and 9, together with lower-level sublineages and the animal-associated variants *bovis*, *caprae*, and *orygis*. The *Plasmodium* component included references representing several human-infecting species, together with representative *P. falciparum* sequences from the geographical populations modelled by Afanc. The mpox virus component included references representing clades Ia, Ib, IIa, and IIb. Clade Ia and Ib references included sequences obtained from GISAID under its data-access terms [9], while clade IIa and IIb references included publicly available INSDC records. A complete manifest of all candidate reference sequences, their retention status, assigned taxonomic or lineage node, and source accession is provided in the supplementary database manifest.

For *M. tuberculosis* lineage assignment, the Afanc marker catalogue was derived from the MTBC SNP barcoding framework used by TB-Profiler. This framework builds on the 62-SNP barcode developed by Coll et al. and its subsequent expansion by Napier et al. into the 90-SNP barcode incorporated into TB-Profiler [6, 10]. Afanc does not introduce a new MTBC barcode; rather, it implements these established lineage definitions within the canonical epsilon-Bayes classification framework. Manual adjustments were made to improve hierarchical consistency and refine the depth of sublineage definitions. In particular, the catalogue was modified to prevent sublineages from being reported in the absence of marker support for their parental lineage.

783 For *P. falciparum*, the complete Pf7 dataset was used to train and evaluate the
784 geolineage classifier [11]. Samples were assigned to classes representing the major
785 regional populations and subregional divisions defined in the Pf7 metadata. Two pairs
786 of geographically contiguous and collision-prone geolineages were collapsed: Eastern
787 Africa and North-Eastern Africa were combined as AF-E/NE, while South Asia East
788 and South Asia Far East were combined as AS-S. The resulting native Afanc classes
789 were South America (SA), Western Africa (AF-W), Central Africa (AF-C), Eastern
790 and North-Eastern Africa (AF-E/NE), South Asia (AS-S), South-East Asia East
791 (AS-SE-E), South-East Asia West (AS-SE-W), and Oceania/New Guinea (OC-NG).
792 For comparison with Malaria-Profiler, Afanc predictions were additionally collapsed
793 to the broader regional labels reported by that tool. The harmonisation scheme is
794 provided in Supplementary Table S2.
795
796
797

803
804 For mpox virus, a clade-level model was constructed to distinguish clades Ia, Ib, IIa,
805 IIb, and an outgroup set. An additional within-clade model was constructed for lineage
806 assignment within clade IIb. Lineage definitions followed the public `mpxv-lineages`
807 designation repository and the nomenclature described by Happi et al. and Ruis et al.
808 [12, 13].
809
810
811
812

813 3.2 Evaluation structure

814 Afanc was evaluated across three analytical tasks designed to reflect its intended
815 functionality: (i) deconvolution of complex compound datasets, (ii) species and
816 subspecies-level assignment, and (iii) lineage or clade-level assignment. These tasks
817 were assessed across multiple pathogen systems in order to test whether a single
818 profiling framework could generalise across distinct taxonomic and epidemiological
819 settings.
820
821
822
823
824
825
826
827
828

3.3 Complex dataset profiling

To assess Afanc in the context of mixed and taxonomically complex samples, a compound Mycobacteriaceae dataset was simulated using ART [14]. This dataset, termed M2, consisted of six synthetic read sets constructed from seven clinically relevant species and variants: *Mycobacterium kansasii*, *Mycobacterium fortuitum*, *Mycobacterium interjectum*, *Mycobacteroides chelonae*, *Mycobacterium intracellulare* subsp. *chimaera*, *Mycobacterium avium*, and *Mycobacterium tuberculosis* variant bovis BCG. Each compound dataset was simulated at 5%, 20%, 40%, 60%, 80%, and 100% breadth of coverage for each constituent genome, at a depth of 40x. These data were used to evaluate the capacity of Afanc to deconvolute multi-species samples and to distinguish true signals from false-positive low-level assignments.

3.4 Species and subspecies level assignment

Species and subspecies-level performance was evaluated using a simulated Mycobacteriaceae dataset (M1). Reads were simulated using ART from 141 non-tuberculous mycobacterial species and subspecies (136 species and 5 subspecies), giving a total of 846 read sets. Each taxon was represented by six simulated datasets corresponding to 1%, 5%, 10%, 20%, 50%, and 100% genome breadth of coverage at 10x depth. This dataset was used to test broad taxonomic profiling performance under progressively weaker signals.

3.5 Lineage classification datasets

Three datasets were used to evaluate the lineage classification functionality.

Mycobacterium tuberculosis lineage dataset

Lineage assignment was evaluated using a dataset of 958 paired-end *M. tuberculosis* FASTQ read sets spanning all major characterised lineages, many lower level sublineages, and animal associated variants bovis (La1), caprae (La2), and orygis (La3).

875 This dataset was used to assess the marker-based lineage profiling component of
876 Afanc and to compare its performance against existing specialist tools.
877

878
879
880 A list of sample accessions and reported lineage designations is provided in the
881 TB_lineage_classification tab of the supplementary benchmark spreadsheet.
882

883
884 ***Plasmodium falciparum* geolineage dataset**
885

886 The full publicly available Pf7 data resource was used to evaluate regional lineage
887 (geolineage) assignment [11]. The final dataset consisted of 20,712 VCF files covering
888 the regional classes previously defined. For the *P. falciparum* geolineage model, model
889 development was restricted to a high-quality subset of 10,763 Pf7 samples exceeding
890 a minimum callable-site threshold of 85%. The remaining 9,949 samples were not
891 used for Afanc model fitting.
892

893
894
895
896
897
898 Because regional labels are partially structured by geography and can be con-
899 founded by local sampling structure, model development was evaluated using
900 leave-one-country-out (LOSO) cross-validation within the high-quality subset. Under
901 this design, all samples from one country were withheld for evaluation, model param-
902 eters were estimated from the remaining countries, and predictions were generated for
903 the withheld country. These cross-validation results were used to assess generalisation
904 across sampling locations and to identify collision-prone geolineage boundaries.
905

906
907
908
909
910
911 After model development, the final Afanc geolineage model was trained using the
912 10,763 high-quality samples and applied to the remaining 9,949 non-training samples,
913 as well as to the complete Pf7 cohort of 20,712 samples. Results are therefore reported
914 both for samples not used during Afanc model fitting and for the complete public Pf7
915 resource. The full-cohort results represent performance of the final model across the
916 complete dataset, whereas the LOSO results represent country-held-out validation of
917
918
919
920

the model-development procedure. To enable fair comparison with Malaria-Profiler, Afanc predictions were evaluated both at the native Afanc geolineage level and after collapsing to the broader regional labels reported by Malaria-Profiler, with both tools evaluated on identical sample subsets.

Mpox lineage dataset

A public Mpox virus dataset was assembled to evaluate viral clade and lineage assignment across clades Ia, Ib, IIa, and IIb, together with an outgroup set. Additional within-clade modelling was performed for lineage assignment within clade IIb. Lineage definitions followed the public mpxv-lineages designation repository (<https://github.com/mpxv-lineages/lineage-designation>; accessed 19 May 2026) and the nomenclature described by Happi et al. [12] and Ruis et al. [13]. Sample accessions, recorded lineage designations, Afanc predictions, and sample-level evaluation outcomes are provided in the mpox_lineage_classification worksheet of the supplementary benchmark spreadsheet.

3.6 Benchmarking and evaluation

Afanc was benchmarked against comparator tools selected according to the classification task. Because the compared tools differ in their native reporting resolution, direct comparisons were harmonised to the highest resolution shared by the relevant tools. Where Afanc produced finer resolution calls than a comparator, native Afanc resolution performance was reported separately.

For complex dataset profiling and Mycobacteriaceae species/subspecies assignment, Afanc was benchmarked against Kraken 2 v2.1.2, Bracken v2.6, KrakenUniq v0.5.8, and Mykrobe v0.13.0. Standard databases were used for each comparator unless otherwise stated.

967 For *M. tuberculosis* lineage profiling, Afanc was benchmarked against Mykrobe
968 v0.13.0 and TB-Profiler v4.2.0 using the standard databases distributed within those
969 tools. For the main benchmark visualisation, human-associated lineages L1-L9 were
970 pooled to major lineage level. Animal-associated lineages La1-La3 were also har-
971 monised to the level reported by all tools, because Mykrobe reports these at major
972 lineage level whereas TB-Profiler and Afanc report finer sublineage calls. Sublineage
973 level concordance between Afanc and TB-Profiler was evaluated separately.
974
975
976
977

980 For *P. falciparum* geolineage profiling, Afanc was benchmarked against Malaria-
981 Profiler v0.0.10. Because Malaria-Profiler reports broader regional classes than the
982 native Afanc geolineage model, comparison was performed both at native Afanc
983 resolution and on a harmonised general-label task obtained by collapsing Afanc pre-
984 dictions to a common regional label set.
985
986
987
988
989

990 For Mpox virus clade and lineage profiling, no directly comparable specialist
991 classifier was available. Benchmarking was therefore limited to internal performance
992 assessment against lineage classifications recorded in publicly available metadata.
993 Because public metadata were sometimes available only at coarser lineage resolution
994 than Afanc predictions, Mpox performance was assessed both by strict exact match
995 and after normalising Afanc predictions to the resolution of the recorded metadata.
996
997
998
999

1000
1001
1002 Benchmarking was assessed using task-appropriate classification summaries across
1003 Mycobacterium species identification, complex Mycobacterium mixture detection, *M.*
1004 *tuberculosis* lineage assignment, *P. falciparum* geographic lineage assignment, and
1005 Mpox lineage assignment. For multiclass species benchmarking, sensitivity and pre-
1006 cision were reported. For lineage/geolineage benchmarks, sensitivity, precision, and
1007
1008
1009
1010
1011
1012

balanced accuracy were reported. All metrics were calculated in a one-vs-rest framework for each class, then macro-averaged across classes within each cohort so that each taxon or lineage contributed equally. For complex *Mycobacterium* mixtures, performance was summarised as binary target detection or non-detection for each expected species, with off-target detections recorded separately as “Other”.

4 Results

4.1 Afanc resolves *Mycobacteriaceae* taxa from incomplete genomic signal

Figure 3A shows results of running Afanc, Mykrobe, Bracken and KrakenUniq on dataset M1 to evaluate their performance on the full publicly available *Mycobacteriaceae* taxonomic space. Afanc correctly assigned the expected species or subspecies across all coverage cohorts, including the lowest coverage cohorts, where only 1% of the genome was present. Comparator tools showed reduced performance under these conditions. Mykrobe demonstrated reduced performance under low coverage conditions, failing to correctly identify species or subspecies at 1% genome coverage, and only 2 species at 5% coverage, though showed substantially improved performance above 20% coverage. Bracken and KrakenUniq showed more stable behaviour across coverage cohorts, but performance was consistently lower than Afanc. Both sensitivity and precision were reduced in KrakenUniq compared to Bracken, in all coverage cohorts. Bracken was more prone to underclassification, reporting the parent taxon of a sample, whereas KrakenUniq produced a greater burden of incorrect species-level calls.

1059 4.2 Afanc deconvolutes mixed Mycobacteriaceae datasets

1060

1061 across coverage gradients

1062

1063 Figure 3C shows results of running Afanc, Mykrobe, Bracken and KrakenUniq on
1064 dataset M2 to evaluate their performance on complex datasets consisting of many
1065 species and variants, with varying levels of genomic similarity. Afanc recovered all
1066 expected taxa across all coverage cohorts.
1067

1068

1069

1070

1071 Comparator tools recovered fewer component taxa. Mykrobe identified some
1072 expected taxa at higher coverage levels, but did not consistently recover all con-
1073 stituents at lower coverage cohorts. Only Afanc and Mykrobe managed to recover M.
1074 tuberculosis var. bovis BCG at any coverage level. Bracken and KrakenUniq recov-
1075 ered some expected taxa, but also produced spurious assignments.
1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

4.3 Afanc resolves deep canonical *M. tuberculosis* lineage structure

Tool	Exact / truth-tied	Wrong-branch	Unclassified	Correct or tied rate
Afanc	955	0	3	99.7%
TB-Profiler	954	2	2	99.6%
Mykrobe	946	1	11	98.7%

Table 1 The performance of the three lineage profiling tools on the test set of 958 TB samples, with results harmonised to the lowest level of classification supported by Mykrobe.

At the harmonised reporting level utilised for the main benchmarking, Afanc correctly assigns 955/958 (99.7%) samples (see table 1). TB-Profiler correctly assigned or truth-tied (where true lineage was reported with another lineage) 954/958 (99.6%) samples, and Mykrobe correctly assigned or truth-tied 946/958 (98.7%) samples.

Afanc produced no wrong-branch assignments at the harmonised level, and left three samples unclassified due to insufficient allele level evidence. TB-Profiler produced two wrong-branch assignments and two unclassified calls. Mykrobe produced one wrong-branch assignment and eleven unclassified calls.

At full lineage resolution, Afanc produced 927 exact calls, 18 hierarchy compatible overcalls, 6 hierarchy compatible undercalls, 4 wrong-branch assignments and 3 unclassified. This gave 951/958 exact or hierarchy compatible assignments (99.3%). TB-Profiler produced 914 exact calls, 1 truth-tied call, 19 hierarchy compatible overcalls, 6 hierarchy compatible undercalls, 16 wrong-branch assignments, and 2 unclassified calls, giving 940/958 exact or hierarchy compatible assignments (98.1%; see Supplementary Table S1).

Full resolution predictions for both Afanc and TB-Profiler were identical for 938/958 samples (97.9%), and for 937/954 samples (98.2%) when restricted to samples for which both classifiers reported a lineage. The largest repeated discordance involved 10 lineage 3.2 samples, which were called correctly by Afanc, but as lineage 3.1.3 by TB-Profiler. A second repeated discordance involved five lineage 5 samples, called as 5.1 by Afanc and 5.1.1 by TB-Profiler.

4.4 Afanc resolves geographically structured *P. falciparum* lineages

When reporting level was harmonised with the broader regional classes used by Malaria-Profiler, Afanc correctly assigned 18,830/20,712 *P. falciparum* samples (90.9%), compared with 15,559/20,712 (75.1%) for Malaria-Profiler. Afanc produced higher correct rates across all harmonised geolinesages, with the largest differences

Geolineage	<i>n</i>	Afanc		Malaria-Profiler	
		Correct	Other	Correct	Other
Central Africa	573	323 (56.4%)	250	126 (22.0%)	447
Eastern Africa	2134	1,810 (84.8%)	324	1,724 (80.8%)	410
Oceania	384	370 (96.4%)	14	0 (0.0%)	384
South America	182	174 (95.6%)	8	170 (93.4%)	12
South Asia	1976	1,640 (83.0%)	336	1,112 (56.3%)	864
Southeast Asia	6873	5,933 (86.3%)	940	3,894 (56.7%)	2,979
Western Africa	8590	8,580 (99.9%)	10	8,533 (99.3%)	57
Total	20712	18,830 (90.9%)	1,882	15,559 (75.1%)	5,153

Table 2 Comparison of Afanc and Malaria-Profiler classification performance across *Plasmodium falciparum* geolineages in the complete Pf7 cohort. This cohort consisted of all 20,712 Pf7 samples, including the 10,763 high-quality samples used for Afanc model development and the 9,949 samples not used for Afanc model fitting. Results are shown using the harmonised regional class resolution used by Malaria-Profiler. Correct values show the number and percentage of samples assigned to the true harmonised geolineage. Other combines incorrect and unclassified calls. For results with "Other" split into failure cases, see Supplementary Table S3.

observed for Oceania, Central Africa, South Asia, and Southeast Asia. Central Africa had a low correct rate across both tools, with Afanc most frequently assigning these samples to other African geolineages. Malaria-Profiler failed to correctly assign any Oceania sample, consistent with insufficient representation of this regional class in its classification scheme.

To assess whether the improved Pf7 performance was driven by samples used during Afanc model development, we repeated the Afanc and Malaria-Profiler comparison separately for the 10,763 high-quality model-development samples and the remaining 9,949 samples not used for Afanc model fitting (Supplementary Figures S4 and S5; Supplementary Tables S5 and S6). This distinction is important for surveillance applications, because regional geolineage labels in *P. falciparum* approximate geographically structured allele-frequency variation rather than discrete clonal lineages.

Lineage	<i>n</i>	Exact	Compatible overcall	Compatible undercall	Truth tied
A.1	1	1 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
A.2	3	0 (0.0%)	3 (100.0%)	0 (0.0%)	0 (0.0%)
B.1	1727	1,289 (74.6%)	426 (24.7%)	6 (0.3%)	6 (0.3%)
B.1.1	81	75 (92.6%)	0 (0.0%)	2 (2.5%)	4 (4.9%)
B.1.2	210	206 (98.1%)	0 (0.0%)	3 (1.4%)	1 (0.5%)
B.1.3	63	58 (92.1%)	0 (0.0%)	1 (1.6%)	4 (6.3%)
B.1.4	7	7 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
B.1.5	19	18 (94.7%)	0 (0.0%)	0 (0.0%)	1 (5.3%)
Total	2111	1,654 (78.4%)	429 (20.3%)	12 (0.6%)	16 (0.8%)

Table 3 Classification performance across Mpox clade IIb lineages.

4.5 Afanc resolves shallow outbreak-associated Mpox lineages

Figure 5 shows the results of running Afanc on 2,111 Mpox samples and comparing the lineage reported by Afanc to the lineage recorded in the metadata. Afanc produced 1,654 exact calls, 429 descendant (compatible overcall) calls, 12 ancestor (compatible undercall) calls, and 16 truth-tied calls across 2,111 Mpox samples (see Table 3). No wrong-branch assignments were reported, and no samples were unclassified.

Under strict exact-match evaluation, Afanc produced 1,654/2,111 exact calls (78.4%). When compatible overcalls and truth-tied calls were counted as metadata resolution compatible assignments, Afanc produced 2,099/2,111 compatible calls (99.4%). All remaining calls were compatible undercalls. No wrong-branch assignments were reported.

5 Discussion

Afanc demonstrates that broad taxonomic screening and high-resolution marker-based classification can be integrated into a single framework that remains effective across markedly different pathogen classification problems. Across the evaluated datasets, Afanc performed strongly in Mycobacteriaceae taxonomic assignment, complex Mycobacteriaceae mixture deconvolution, *M. tuberculosis* lineage classification,

1243 *P. falciparum* geolineage prediction, and Mpox lineage classification. These tasks
1244 differ not only in technical difficulty, but also in the biological structure of the labels
1245 being predicted: fragmented or compound taxonomic signal, old and deep canonical
1246 prokaryotic SNP-defined lineages, geographically contiguous eukaryotic parasite pop-
1247 ulations, and recent shallow viral outbreak lineages.

1251
1252
1253 These evaluations position Afanc as a classification layer for pathogen genomic
1254 surveillance rather than as a pathogen-specific typing tool. In routine surveillance,
1255 taxonomic and lineage calls are often the entry point for downstream interpretation,
1256 since they determine whether a sample should be treated as a known pathogen,
1257 whether lineage or clade-level tracking is meaningful, whether apparent mixed signal
1258 should be escalated or discounted, and whether uncertainty should limit downstream
1259 epidemiological inference. The Mycobacteriaceae taxonomic evaluations represent
1260 both the breadth and specificity of Afanc across incomplete and compound signals.
1261 The *M. tuberculosis*, *P. falciparum*, and mpox lineage and clade level evaluations
1262 represent three contrasting surveillance-relevant regimes: deep and stable bacterial
1263 lineage architecture, structured contiguous eukaryotic parasite populations, and
1264 shallow outbreak-associated viral clades. Across these settings, Afanc provides a
1265 common framework for converting heterogeneous sequencing evidence into auditable
1266 taxonomic and lineage-level classifications.

1276
1277
1278 A central advantage of Afanc is that it addresses multiple failure points within
1279 pathogen profiling workflows. The first is reference database construction. Classifi-
1280 cation accuracy is constrained by the quality and representativeness of the database
1281 used for read assignment. Poorly assembled, contaminated, redundant, or non-
1282 representative genomes can propagate directly into unreliable taxonomic profiles.
1283 Afanc addresses this “garbage in, garbage out” problem through the **autodatabase**
1284

module, which preprocesses input assemblies before database construction and removes atypical or non-representative sequences, which may include low-quality assemblies. This upstream filtering is critical because improvements in downstream classification are limited if the reference database itself contains misleading sequence space.

The second challenge is interpretation of taxonomic reports. Generic k-mer based profilers are powerful for rapid screening, but closely related taxa and low-abundance signals can produce complex reports containing many plausible low-level assignments. Simple abundance thresholds are often insufficient: permissive thresholds retain large numbers of false-positive assignments, whereas aggressive thresholds can remove true low-abundance or low-rank taxa. Afanc addresses this problem through Kraken 2 report deconvolution, combining elastic taxon-specific thresholds with evidence-weighted read redistribution. This allows the report to be simplified without relying solely on a fixed global cutoff, and was reflected in Afanc’s leading performance across Mycobacteriaceae species and subspecies, including low-coverage simulated read sets.

The species and compound sample analyses show the value of retaining a broad taxonomic screening layer before applying marker-based classification. Downstream high-resolution assignment is only meaningful if the upstream taxonomic context has been identified reliably. In the compound Mycobacteriaceae datasets, Afanc recovered expected taxa more consistently than comparator tools, including where the available genomic signal was incomplete.

The *M. tuberculosis* results demonstrate that Afanc can operate competitively in a mature lineage-classification setting where strong specialist tools already exist. Here, the relevant lineage architecture is deep, relatively stable, and defined by canonical

1335 SNP structure. Afanc outperformed comparator tools at both harmonised and native
1336 reporting levels, with very few incompatible branch assignments. This indicates that
1337 the hierarchical canonical epsilon-Bayes model is well suited to classification problems
1338 where lineage labels are arranged in a tree and where the principal risk is assignment
1339 to an incompatible branch rather than minor disagreement in reporting depth.

1340
1341
1342
1343
1344
1345 Mpox virus represents a contrasting hierarchical problem. Unlike *M. tuberculosis*,
1346 Mpox lineage structure is recent, shallow, and undergoing rapid expansion, with some
1347 sublineages separated by very small numbers of SNPs. In this setting, strict exact-
1348 match performance was limited by the resolution of the available metadata, since
1349 many apparent overcalls reflected Afanc assigning a more specific descendant lineage
1350 than was recorded in the public label. The absence of wrong-branch assignments is
1351 therefore more informative than strict exact-match accuracy alone. After normalising
1352 predictions to the resolution of the recorded metadata, Afanc achieved high compati-
1353 bility, indicating that the same hierarchical framework can remain stable in outbreak
1354 settings where nomenclature and metadata may lag behind genomic resolution.

1355
1356
1357
1358
1359
1360
1361
1362
1363 The *P. falciparum* analysis tested a less tree-like classification problem. Geo-
1364 graphic labels are not equivalent to stable clonal lineages: they reflect allele-frequency
1365 structure shaped by geography, recombination, migration, and sampling density.
1366 Afanc nevertheless substantially outperformed Malaria-Profiler on the harmonised
1367 regional task. This shows that the classification framework is not restricted to
1368 canonical SNP-defined lineage systems, but can also model overlapping regional
1369 allele-frequency structure. However, performance was uneven across geolineages. Cen-
1370 tral Africa remained the most difficult class, with lower correct rates for both Afanc
1371 (56.4%) and Malaria-Profiler (22.0%). This pattern is consistent with Central Africa
1372 occupying a collision-prone position among neighbouring African populations. In
1373
1374
1375
1376
1377
1378
1379
1380

geographically structured pathogens, some classification errors may therefore reflect weak biological separability of the labels rather than failure of the classifier alone.

Several methodological features were important in obtaining these results. First, Afanc explicitly distinguishes callable absence from missingness. A lineage-enriched or lineage-defining allele that is absent at a callable site is negative evidence, whereas an uncallable site is excluded from likelihood calculation. This distinction is essential under challenging data-quality conditions, where treating missing sites as absence would distort the likelihood and penalise samples simply because coverage is incomplete. Second, Afanc can incorporate informative homozygous-reference states. This reduces reference-ascertainment bias when a target population is genetically close to the reference genome and would otherwise be represented by a sparse alternate-allele panel. In such cases, reference-state markers do not merely increase the number of observations; they impose a more appropriate evidential burden for class membership.

The main unresolved challenge is how best to represent uncertainty when the target class structure is weakly separated. This was most evident in *P. falciparum*, where some geolinesages behaved as discrete and highly recoverable classes, whereas others showed substantial overlap with neighbouring regions. Future development should therefore incorporate explicit collision-aware decision rules, such as pairwise margin thresholds or rejection options that return an ambiguous set or higher-level regional label when fine-scale evidence is insufficient. This would allow Afanc to avoid forcing discrete assignments in settings where the underlying biology may not support them.

Several limitations should be noted. This study evaluates Afanc as a taxonomic deconvolution and lineage-classification framework, rather than as a complete real-time outbreak reconstruction system. We did not assess prospective deployment in a

1427 clinical laboratory, turnaround time under routine service conditions, or direct effects
1428 on public-health decision-making. The Mycobacteriaceae taxonomic benchmarks are
1429 simulation-based, although lineage, geolineage and viral clade classification were eval-
1430 uated using large public datasets. Future evaluations should include prospective or
1431 retrospective surveillance cohorts with independently curated outbreak, transmission
1432 or epidemiological metadata, and should assess integration of Afanc outputs into
1433 reporting dashboards, surveillance databases and reference-laboratory workflows.
1434

1435 Together, these results show that Afanc provides a reusable classification layer for
1436 pathogen genomic surveillance, linking broad taxonomic evidence to lineage- or clade-
1437 level interpretation across bacterial, parasite and viral systems. Its main contribution is
1438 not a new marker catalogue for any single organism, but a framework for constructing
1439 custom databases, simplifying noisy taxonomic reports, applying organism-specific
1440 lineage models, and preserving uncertainty when evidence is incomplete or ambiguous.
1441 This design makes Afanc particularly suited to surveillance and research settings where
1442 pathogen targets, lineage schemes and available reference data change over time.
1443

1444 6 Conclusion

1445 Afanc provides a general framework for pathogen taxonomic and lineage profiling
1446 from NGS data. By incorporating quality control and representativeness filtering
1447 during database construction, Afanc reduces the risk that low-quality or atypical
1448 reference assemblies propagate into unreliable classification results. Its report decon-
1449 volution and taxonomic assignment filtering reduce false-positive calls from raw read
1450 classification and improve recovery of expected taxa in complex mixed datasets,
1451 including under challenging coverage conditions.
1452

Beyond broad taxonomic profiling, Afanc links species or clade identification to downstream marker-based lineage classification. Afanc performed strongly across three distinct lineage architectures: deep canonical prokaryotic lineages, overlapping eukaryotic geolineages, and shallow rapidly evolving viral outbreak lineages. These results show that high-resolution pathogen classification can be embedded within a single extensible framework, rather than handled through separate organism-specific workflows.

The principal contribution of Afanc is therefore not any single marker catalogue or benchmark result, but the integration of database construction, taxonomic deconvolution, model selection, and probabilistic lineage or clade assignment into one reusable workflow. This makes Afanc suited to pathogen genomic surveillance and research settings where pathogen targets, reference databases, lineage schemes, and available evidence may change over time. By preserving uncertainty and explicitly handling incomplete or mixed taxonomic signal, Afanc provides a scalable framework for moving from broad sequencing evidence to interpretable pathogen classification.

Supplementary information. Supplementary files accompany this manuscript, including benchmark results, accession-level information, reference manifests, and additional figures and tables. A supplementary GitHub repository accompanying this paper is available at https://github.com/ArthurVM/Afanc_supplementary/.

Availability and requirements

- Project name: Afanc.
- Project home page: <https://github.com/ArthurVM/Afanc>.
- Archived version: <https://doi.org/10.5281/zenodo.20718125>.
- Operating system(s): Linux/Unix-like systems.
- Programming language: Python.

- 1519 • Other requirements: Python 3.7, Perl, Entrez Direct E-utilities, Mash v2.3, Kraken
1520 2 v2.1.2, NCBI BLAST+, Krona, Bowtie2, and the Python packages listed in the
1521 project requirements file.
1522
- 1524 • License: GNU GPL v3.0.
1525
- 1526 • Any restrictions to use by non-academics: Use is governed by the GNU GPL v3.0
1527 licence.
1528

1530 **List of abbreviations**

- 1533 *Afanc* Integrated framework described in this manuscript.
- 1534 *ART* Artificial Read Transcription simulator.
- 1535
- 1536 *BAM* Binary Alignment Map.
- 1537
- 1538 *CEB* Canonical-epsilon-Bayesian.
- 1539
- 1540 *FASTQ* Text-based sequencing read format containing sequence and quality scores.
- 1541
- 1542 *GISAID* Global Initiative on Sharing All Influenza Data.
- 1543
- 1544 *INSDC* International Nucleotide Sequence Database Collaboration.
- 1545
- 1546 *JSON* JavaScript Object Notation.
- 1547
- 1548 *K-mer* Sequence substring of length k .
- 1549
- 1550 *MTBC* *Mycobacterium tuberculosis* complex.
- 1551
- 1552 *NCBI* National Center for Biotechnology Information.
- 1553
- 1554 *NGS* Next-generation sequencing.
- 1555
- 1556 *SB* Semi-Bayesian.
- 1557
- 1558 *SNP* Single-nucleotide polymorphism.
- 1559
- 1560 *SRA/ENA* Sequence Read Archive / European Nucleotide Archive.
- 1561
- 1562
- 1563
- 1564

Declarations

Ethics approval and consent to participate. Not applicable. This study used publicly available pathogen genomic datasets and simulated sequencing data and did not involve new sampling from human participants, animals or human tissue.

Consent for publication. Not applicable.

Availability of data and materials. The Afanc source code and documentation are available at <https://github.com/ArthurVM/Afanc> and archived on Zenodo at <https://doi.org/10.5281/zenodo.20718125>. A version of the database and assembly directory structure used in this paper, with GISAID-derived sequences excluded, is archived on Zenodo at <https://doi.org/10.5281/zenodo.20708345>. Sample-level benchmark results, reported classifications and accessions for the evaluated datasets are provided in the supplementary benchmark spreadsheet.

A complete manifest of the candidate reference sequences used during construction of the compound Afanc database is provided in the supplementary database manifest. This records the retention status, assigned taxonomic or lineage node and source accession for each sequence. Mpox virus reference sequences obtained from GISAID were accessed under its data-access terms and are identified in the supplementary GISAID acknowledgement table by EPI.SET.260613af and DOI <https://doi.org/10.55876/gis8.260613af>. The remaining mpox virus reference sequences are identified by their INSDC accessions. The *P. falciparum* variant data and associated sample metadata were obtained from the published Pf7 data resource.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported by the Wellcome Trust [215800/Z/19/Z]; the Medical Research Council [MR/T030062/1]; and Health and Care Research Wales [AF-24-10].

Authors' contributions. A.M., A.P., and T.C. conceived the software presented here. A.P. designed and developed the first version of autoDatabase in Nextflow. A.M. designed and developed Afanc and its constituent algorithms and statistical modelling architecture, carried out testing, and wrote the first draft of the manuscript. All authors contributed to editing and adjusting the final manuscript.

Acknowledgements. We are grateful to Owen Jones and Tom Whalley for their advice and support during this work. We also gratefully acknowledge the members of the SP3 project, particularly Martin Hunt, Zamin Iqbal, David Andrew Lee, Esther Robinson, Katherine Fletcher and Matt Backx, who provided essential feedback during the development of Afanc. We gratefully acknowledge the authors, originating laboratories and submitting laboratories responsible for the sequences shared through GISAID and included in EPISET_260613af.

References

- [1] Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15(3):R46. <https://doi.org/10.1186/gb-2014-15-3-r46>.
- [2] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 2019;20(1):257. <https://doi.org/10.1186/s13059-019-1891-0>.
- [3] Lu J, Breitwieser FP, Thielen P, et al. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci.* 2017;3:e104. <https://doi.org/10.7717/peerj-cs.104>.
- [4] Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* 2018;19(1):198. <https://doi.org/10.1186/s13059-018-1568-0>.

- [5] Hunt M, Bradley P, Lapierre SG, et al. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. Wellcome Open Res. 2019;4:191. <https://doi.org/10.12688/wellcomeopenres.15603.1>.
- [6] Napier G, Campino S, Merid Y, et al. Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. Genome Med. 2020;12(1):114. <https://doi.org/10.1186/s13073-020-00817-3>.
- [7] Phelan JE, Turkiewicz A, Manko E, et al. Rapid profiling of *Plasmodium* parasites from genome sequences to assist malaria control. Genome Med. 2023;15(1):96. <https://doi.org/10.1186/s13073-023-01247-7>.
- [8] Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132. <https://doi.org/10.1186/s13059-016-0997-x>.
- [9] Shu Y, McCauley J. GISAID: Global Initiative on Sharing All Influenza Data—from vision to reality. Euro Surveill. 2017;22(13):30494. <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>.
- [10] Coll F, McNerney R, Guerra-Assunção JA, et al. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. Nat Commun. 2014;5:4812. <https://doi.org/10.1038/ncomms5812>.
- [11] MalariaGEN, Abdel Hamid MM, Abdelraheem MH, et al. Pf7: an open dataset of *Plasmodium falciparum* genome variation in 20,000 worldwide samples. Wellcome Open Res. 2023;8:22. <https://doi.org/10.12688/wellcomeopenres.18681.1>.
- [12] Happi C, Adetifa I, Mbala P, et al. Urgent need for a non-discriminatory and non-stigmatizing nomenclature for monkeypox virus. PLoS Biol. 2022;20(8):e3001769. <https://doi.org/10.1371/journal.pbio.3001769>.

1703 [13] Ruis C, Lusamaki E, O'Toole Á, et al. A systematic nomenclature for mpox
1704 viruses causing outbreaks with sustained human-to-human transmission. Nat
1705 Med. 2025;31(9):2854–2858. <https://doi.org/10.1038/s41591-025-03820-6>.
1706
1707
1708
1709 [14] Huang W, Li L, Myers JR, et al. ART: a next-generation sequencing read simula-
1710 tor. Bioinformatics. 2012;28(4):593–594. [https://doi.org/10.1093/bioinformatics/](https://doi.org/10.1093/bioinformatics/btr708)
1711 [btr708](https://doi.org/10.1093/bioinformatics/btr708).
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748

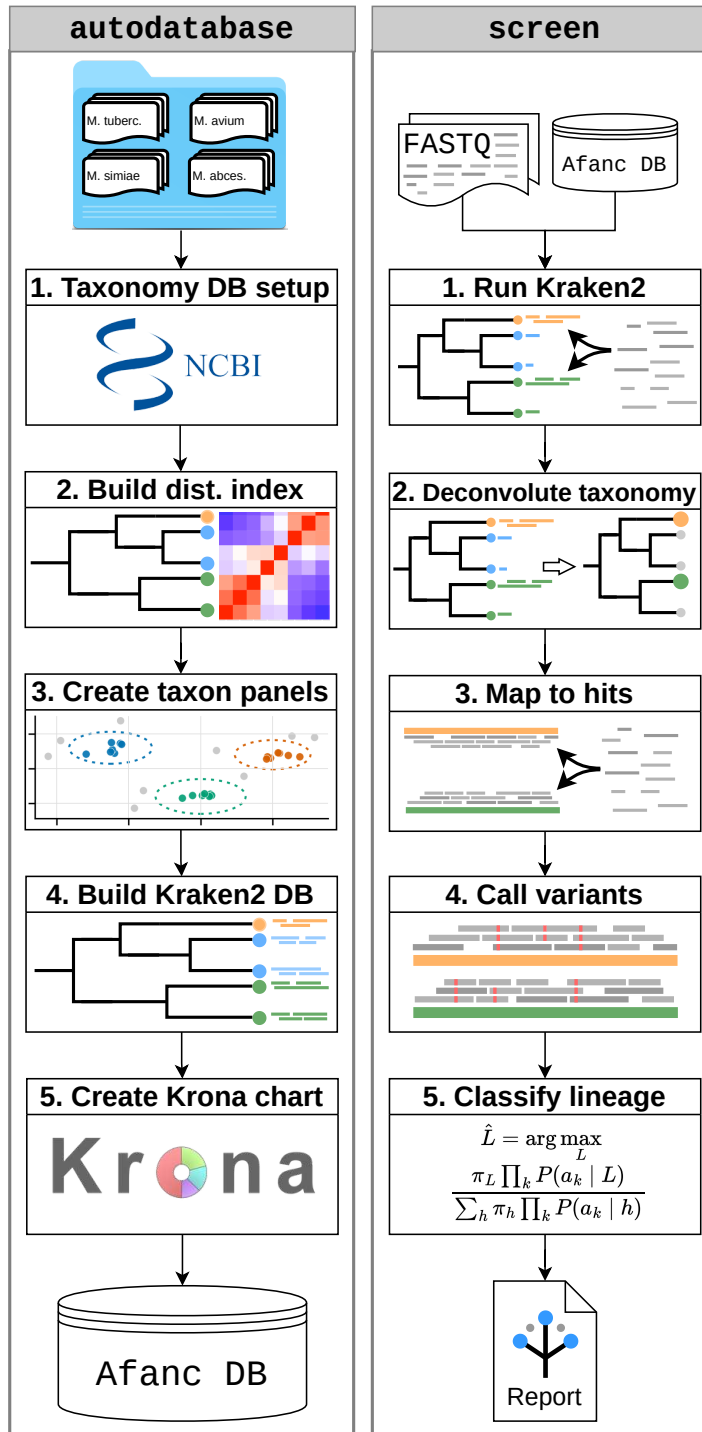


Fig. 1 Steps in the autodb and screen workflows.

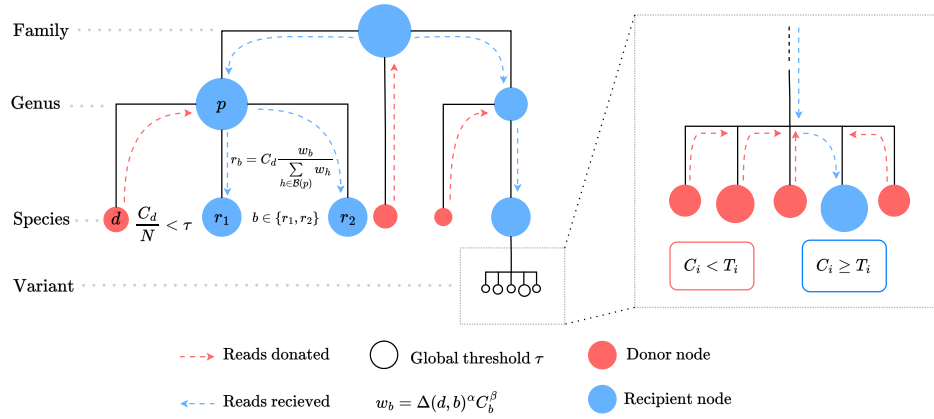


Fig. 2 Read redistribution on the Kraken 2 report tree. Unsupported donor nodes donate their direct read assignments upward to the nearest ancestor with supported descendant signal, after which reads are routed back down to supported recipient branches. At each routing step, donated reads are split between candidate recipient branches in proportion to weights w_b , which combine Mash-derived similarity and existing read support. The inset illustrates the local elastic threshold used to retain or reject variant-level nodes. Full notation and redistribution equations are given in the Implementation section.

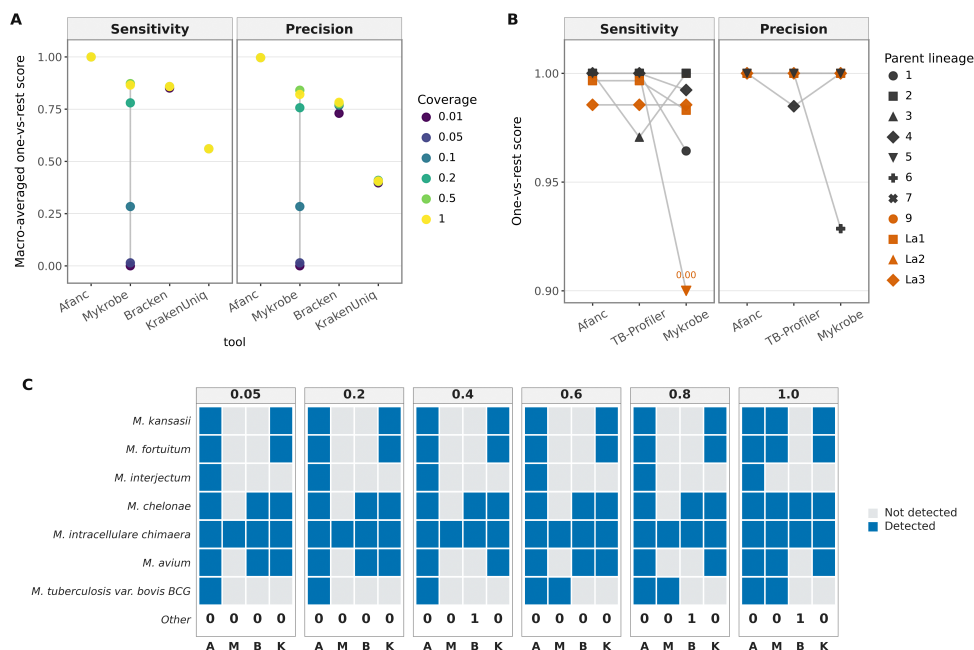


Fig. 3 Performance of Afanc across Mycobacteriaceae and tuberculosis classification datasets. **A**, Results of running Afanc, Mykrobe, Bracken, and KrakenUniq on publicly available Mycobacteriaceae species and subspecies genomes simulated to varying genome coverage. **B**, Results of running Afanc, Mykrobe, and TB-Profiler on the TB lineage dataset. **C**, Results of running Afanc, Mykrobe, Bracken, and KrakenUniq using dataset M2 across all coverage cohorts from 5% to 100%. A = Afanc, M = Mykrobe, B = Bracken, K = KrakenUniq. Other = number of unlisted Mycobacteriaceae reported.

1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932

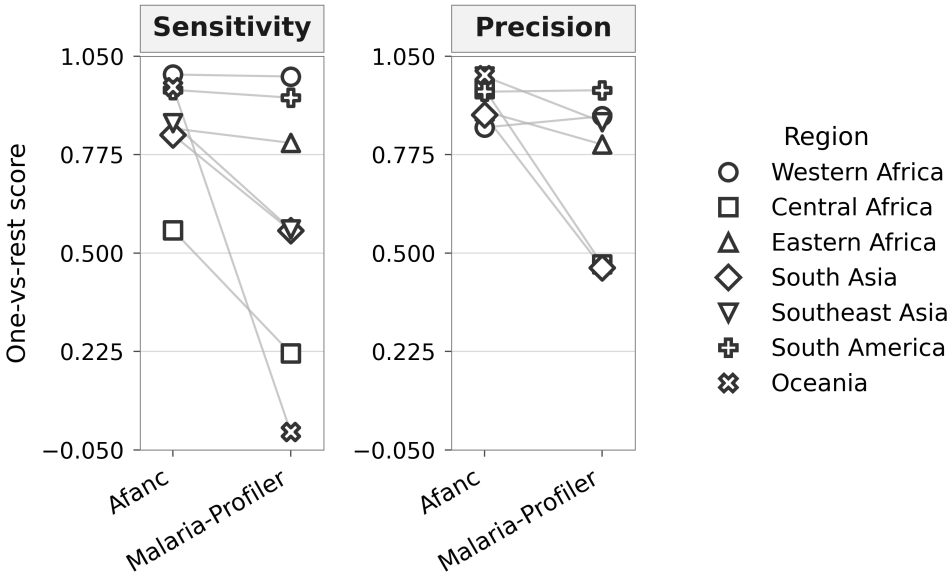


Fig. 4 The performance of Afanc and Malaria-Profiler on the Plasmodium falciparum Pf7 dataset.

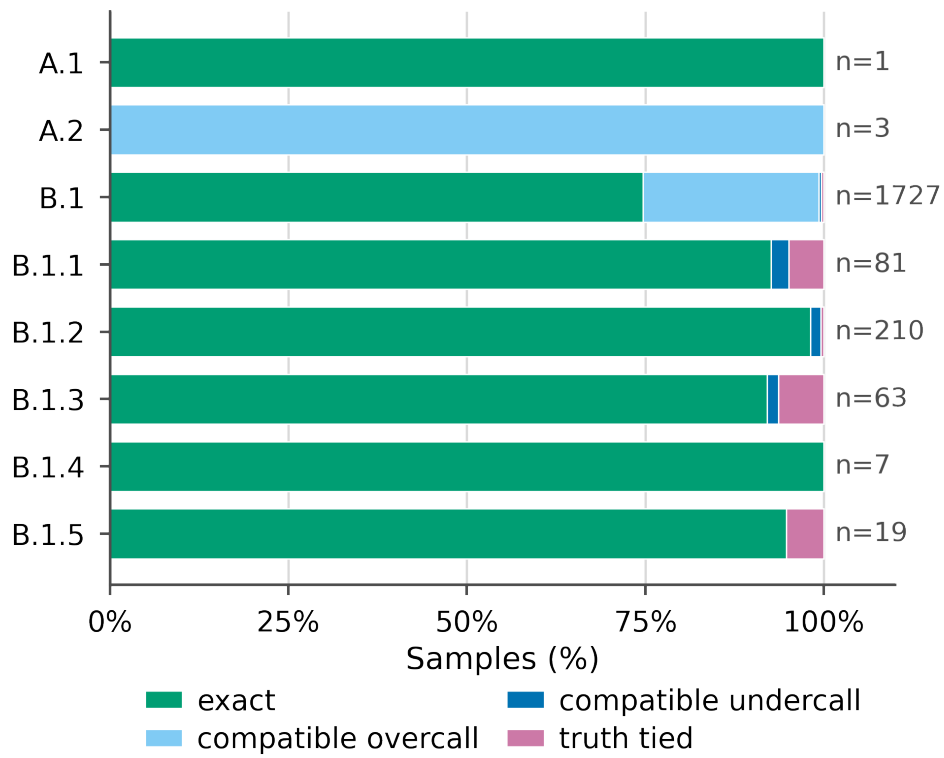


Fig. 5 The results of running Afanc on the Mpox clade IIb dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [gisaidsupplementaltableepiset260613af.pdf](#)
- [afanctabmanifest.xlsx](#)
- [supplementarymaterialsbenchmark.xlsx](#)
- [supplementarymaterials.pdf](#)