Supplementary Materials

Supplementary Figures

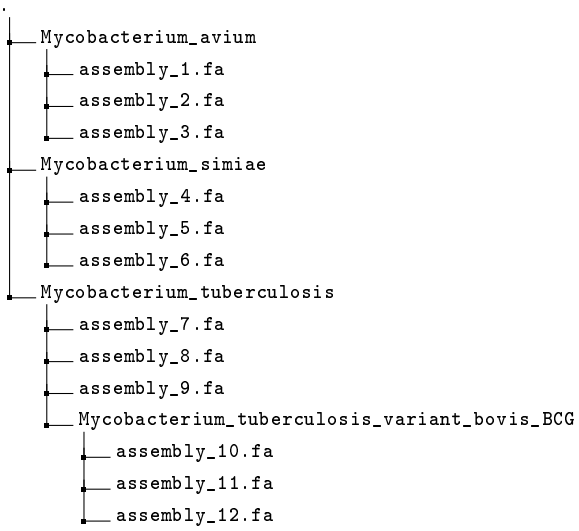

Figure S1: The output directory structure produced by providing the Afanc get_dataset module with the text file defining *Mycobacterium tuberculosis*, *Mycobacterium tuberculosis* variant *bovis* BCG, *Mycobacterium avium*, and *Mycobacterium simiae*, and three assemblies requested.

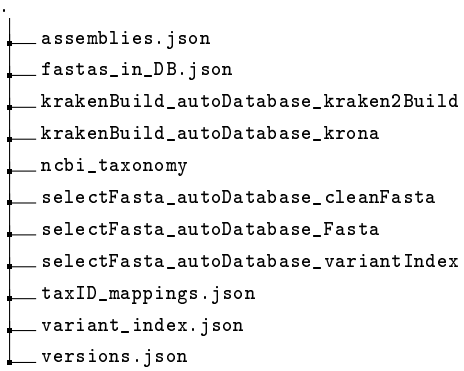

Figure S2: The output directory structure produced by running the Afanc autodatabase module using the directory structure seen in Figure S1.

```
0.05 41 41 U 0 unclassified
99.95 80454 0 R 1 root
99.95 80454 0 R1 131567   cellular organisms
99.95 80454 0 D 2         Bacteria
99.95 80454 0 D1 1783272       Terrabacteria group
99.95 80454 0 P 201174         Actinobacteria
99.95 80454 0 C 1760           Actinomycetia
99.95 80454 0 O 85007             Corynebacteriales
99.95 80454 27 F 1762               Mycobacteriaceae
99.91 80421 288 G 1763                 Mycobacterium
99.51 80099 148 G1 120793                   Mycobacterium avium complex (MAC)
99.31 79941 28955 S 1764                    Mycobacterium avium
62.78 50541 45478 S1 44282                      Mycobacterium avium silvaticum
4.98 4009 4009 S2 1401690                        Mycobacterium avium silvaticum Lineage 1
1.31 1054 1054 S2 1401691                        Mycobacterium avium silvaticum Lineage 2
1.21 972 0 S1 1770                      Mycobacterium avium paratuberculosis
0.49 396 396 S1 439334                      Mycobacterium avium hominissuis
0.16 131 131 S1 44454                       Mycobacterium avium avium
0.01 7 4 S 1767                     Mycobacterium intracellulare
0.00 3 3 S1 487521                       Mycobacterium intracellulare Lineage 1
0.02 18 1 G1 2249310                  Mycobacterium simiae complex
0.00 1 1 S 33895                     Mycobacterium interjectum
```

Figure S3: An example of a Kraken 2 report.

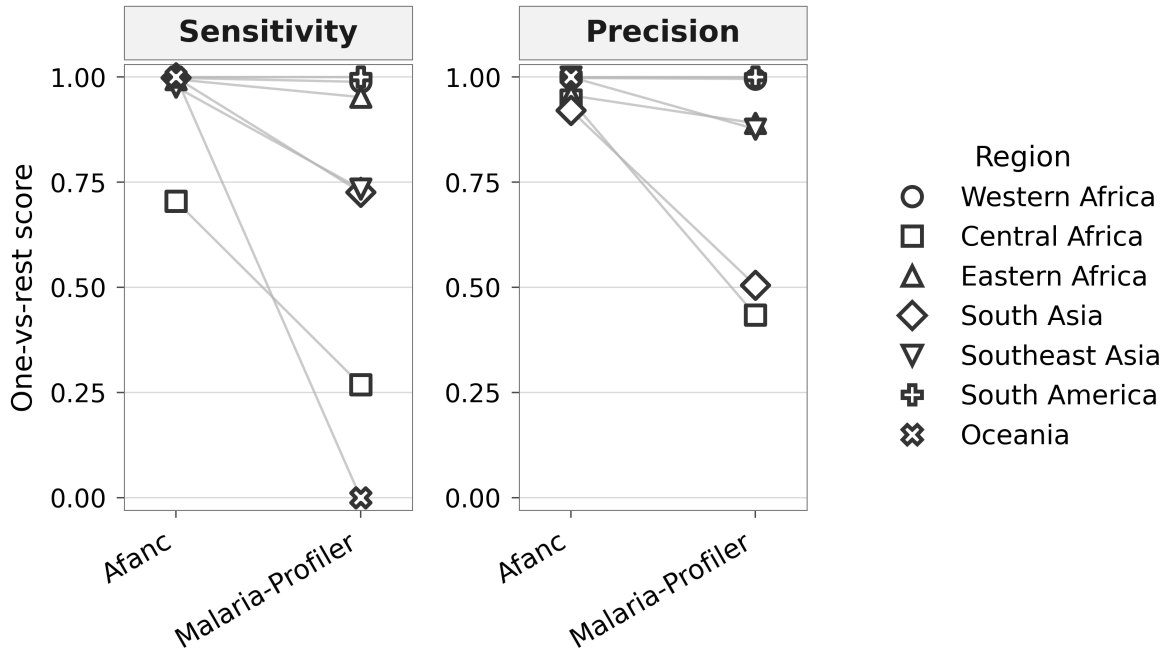

Figure S4: Results of running Afanc and Malaria-Profiler on the high-quality Pf7 model-development subset. This cohort was constructed from Pf7 samples exceeding the 85% callable-site threshold and was used for Afanc geolineage model development (n=10,763).

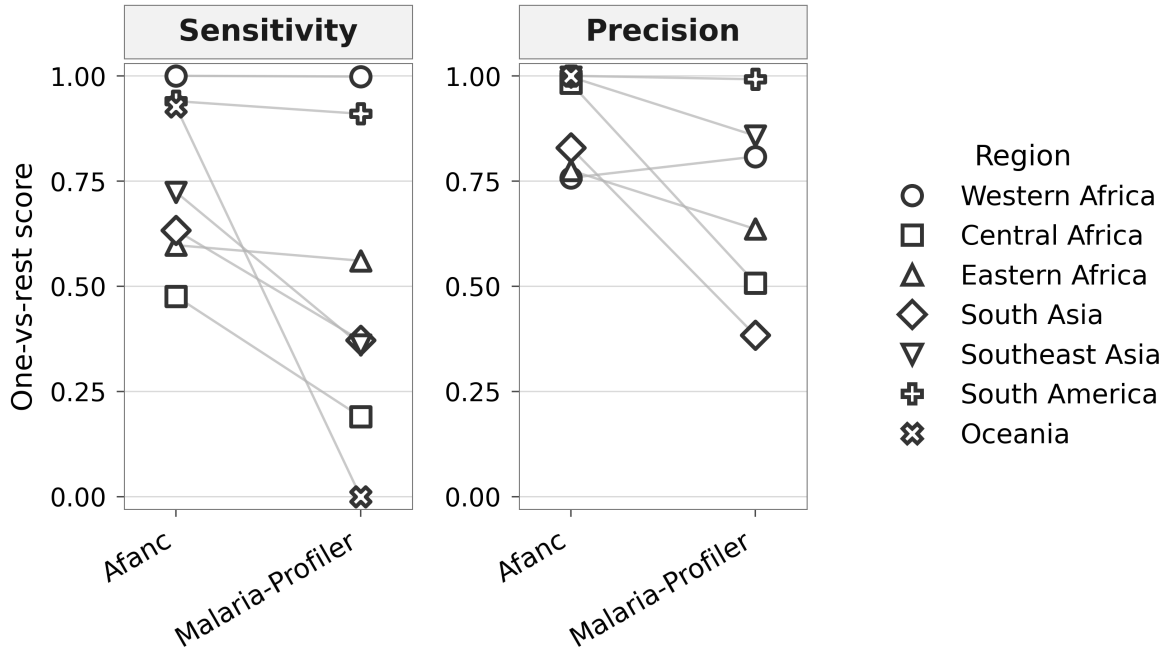

Figure S5: Results of running Afanc and Malaria-Profiler on the non-training Pf7 evaluation subset. This cohort consisted of Pf7 samples below the 85% callable-site threshold that were not used for Afanc model fitting and therefore represents a lower-callability out-of-sample evaluation cohort (n=9,949).

Supplementary Tables

Tool	Exact	Truth-tied	Compatible overcall	Compatible undercall	Wrong branch	Unclassified	Exact or compatible
Afanc	927	0	18	6	4	3	951/958 (99.3%)
TB-Profiler	914	1	19	6	16	2	940/958 (98.1%)

Table S1: Performance of Afanc and TB-Profiler on the 958-sample TB test set at each tool's native classification resolution.

Afanc geolineage	Harmonised regional label
AF-C	Central Africa
AF-E/NE	Eastern Africa
AF-W	Western Africa
AS-S	South Asia
AS-SE-E	Southeast Asia
AS-SE-W	Southeast Asia
OC-NG	Oceania
SA	South America

Table S2: Mapping between native Afanc *Plasmodium falciparum* geolineage labels and the harmonised regional labels used for comparison with Malaria-Profiler.

Geolineage	*n*	Afanc			Malaria-Profiler		
		Correct	Incorrect	Unclassified	Correct	Incorrect	Unclassified
Central Africa	573	323 (56.4%)	250	0 (0.0%)	126 (22.0%)	420	27 (4.7%)
Eastern Africa	2134	1,810 (84.8%)	323	1 (0.0%)	1,724 (80.8%)	357	53 (2.5%)
Oceania	384	370 (96.4%)	14	0 (0.0%)	0 (0.0%)	305	79 (20.6%)
South America	182	174 (95.6%)	8	0 (0.0%)	170 (93.4%)	0	12 (6.6%)
South Asia	1976	1,640 (83.0%)	336	0 (0.0%)	1,112 (56.3%)	579	285 (14.4%)
Southeast Asia	6873	5,933 (86.3%)	940	0 (0.0%)	3,894 (56.7%)	1,824	1,155 (16.8%)
Western Africa	8590	8,580 (99.9%)	10	0 (0.0%)	8,533 (99.3%)	39	18 (0.2%)
Total	**20712**	**18,830 (90.9%)**	**1,881**	**1 (0.0%)**	**15,559 (75.1%)**	**3,524**	**1,629 (7.9%)**

Table S3: Comparison of Afanc and Malaria-Profiler classification performance across *Plasmodium falciparum* geolineages, 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. Incorrect calls are classified to the wrong harmonised geolineage, while unclassified calls have no assigned geolineage.

Afanc geolineage	*n*	Correct	Incorrect	Unclassified
AF-C	573	323 (56.4%)	250	0 (0.0%)
AF-E/NE	2134	1,810 (84.8%)	323	1 (0.0%)
AF-W	8590	8,580 (99.9%)	10	0 (0.0%)
AS-S	1976	1,640 (83.0%)	336	0 (0.0%)
AS-SE-E	4619	3,948 (85.5%)	671	0 (0.0%)
AS-SE-W	2254	1,853 (82.2%)	401	0 (0.0%)
OC-NG	384	370 (96.4%)	14	0 (0.0%)
SA	182	174 (95.6%)	8	0 (0.0%)
Total	**20,712**	**18,698 (90.3%)**	**2,013**	**1 (0.0%)**

Table S4: Afanc classification performance across native *Plasmodium falciparum* geolineage labels. Correct values show the number and percentage of samples assigned to the true native Afanc geolineage.

Geolineage	*n*	Afanc			Malaria-Profiler		
		Correct	Incorrect	Unclassified	Correct	Incorrect	Unclassified
Central Africa	220	155 (70.5%)	65	0 (0.0%)	59 (26.8%)	161	0 (0.0%)
Eastern Africa	1349	1,341 (99.4%)	8	0 (0.0%)	1,284 (95.2%)	64	1 (0.1%)
Oceania	193	193 (100.0%)	0	0 (0.0%)	0 (0.0%)	177	16 (8.3%)
South America	49	49 (100.0%)	0	0 (0.0%)	49 (100.0%)	0	0 (0.0%)
South Asia	1066	1,064 (99.8%)	2	0 (0.0%)	774 (72.6%)	218	74 (6.9%)
Southeast Asia	3798	3,707 (97.6%)	91	0 (0.0%)	2,787 (73.4%)	760	251 (6.6%)
Western Africa	4088	4,078 (99.8%)	10	0 (0.0%)	4,039 (98.8%)	31	18 (0.4%)
Total	**10763**	**10,587 (98.4%)**	**176**	**0 (0.0%)**	**8,992 (83.5%)**	**1,411**	**360 (3.3%)**

Table S5: Comparison of Afanc and Malaria-Profiler classification performance across *Plasmodium falciparum* geolineages in the high-quality Afanc model-development subset. This cohort consisted of 10,763 Pf7 samples exceeding the 85% callable-site threshold and was used for Afanc geolineage model development. 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. Incorrect calls are classified to the wrong harmonised geolineage, while unclassified calls have no assigned geolineage.

Geolineage	*n*	Afanc			Malaria-Profiler		
		Correct	Incorrect	Unclassified	Correct	Incorrect	Unclassified
Central Africa	353	168 (47.6%)	185	0 (0.0%)	67 (19.0%)	259	27 (7.6%)
Eastern Africa	785	469 (59.7%)	315	1 (0.1%)	440 (56.1%)	293	52 (6.6%)
Oceania	191	177 (92.7%)	14	0 (0.0%)	0 (0.0%)	128	63 (33.0%)
South America	133	125 (94.0%)	8	0 (0.0%)	121 (91.0%)	0	12 (9.0%)
South Asia	910	576 (63.3%)	334	0 (0.0%)	338 (37.1%)	361	211 (23.2%)
Southeast Asia	3075	2,226 (72.4%)	849	1 (0.0%)	1,108 (36.0%)	1,064	904 (29.4%)
Western Africa	4502	4,502 (100.0%)	0	0 (0.0%)	4,494 (99.8%)	8	0 (0.0%)
Total	**9949**	**8,242 (82.8%)**	**1,705**	**2 (0.0%)**	**6,568 (66.0%)**	**2,113**	**1,269 (12.8%)**

Table S6: Comparison of Afanc and Malaria-Profiler classification performance across *Plasmodium falciparum* geolineages in the non-training evaluation subset. This cohort consisted of 9,949 Pf7 samples that were 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. Incorrect calls are classified to the wrong harmonised geolineage, while unclassified calls have no assigned geolineage.