Biometrics XX, 1–??
XXX XXXX

Supporting information for "Relapse or reinfection: Classification of malaria infection using transition likelihoods" by Feng-Chang Lin, Quefeng Li, and Jessica Lin

Feng-Chang Lin[1,*], Quefeng Li[1], and Jessica T. Lin[2]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

[2]Institute of Global Health and Infectious Diseases,

University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.

*email: flin@bios.unc.edu

1. Introduction

In this supporting information, we list the detailed sequencing data and classification results by both our classifier $\widehat{\xi}^{(1)}$ and binomial probability model (BPM) method used in Lin et al. (2015). Tables 1 to 4 listed the 30 pairs of baseline and recurrence variants, as well as β coefficients estimated by the gradient descent algorithms we developed and variant prevalence estimated using only baseline variants. The 30 pairs include six second recurrence infections, labeled RR at the end of the identification number, and one third recurrence infection, 81RR→81RRR. As one can see, the baseline variants of these pairs are the same as the recurrence variants in the previous pair. When we estimated β, we excluded these seven pairs to avoid the follow-up length bias, i.e., they were not followed in the same length as the 23 pairs from the baseline. However, when estimating transition probabilities q and q^*, we included those seven pairs in the data analysis. We classify the recurrence infection as a relapse if $\widehat{\xi}^{(1)} > 0.5$ and reinfection otherwise. The BPM classification result was directly cited from Lin et al. (2015). Note that, several pairs were not listed in Lin et al. (2015), such as 10R→10RR. The reason for that is they classified the recurrence infection as reinfection if there are no sharing variants.

Tables 5 and 6 show the coefficient estimation and classification accuracy, respectively, when the reinfection rate is misspecified. The data were simulated under $\mu = 0.05$, with other parameters set up the same as in the manuscript. Estimation and classification results were obtained when $\widehat{\mu}$ was used for μ to fit the data. That is, under $\widehat{\mu} = 0.02$ or 0.1, we misspecify the reinfection rate. The result shows that, even when μ is misspecified, our classification can still reach a satisfactory level when the sample size is large.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

References

Lin, J. T., Hathaway, N. J., Saunders, D. L., Lon, C., Balasubramanian, S., Kharabora, O., et al. (2015). Using amplicon deep sequencing to detect genetic signatures of *Plasmodium vivax* relapse. *The Journal of Infectious Diseases* **212,** 999–1008.

Table 1
Classification of all recurrence pairs based on our proposed method and binomial probability model

Recurrence pair	Baseline variants	β	$\widehat{\xi}^{(0)}$	Recurrence variants	Variant prevalence	$\widehat{\xi}^{(1)}$	Proposed Class	BPM Class
10 $\rightarrow$ 10R	CAM.00	1.833	0.277	CAM.00	0.590	0.954	Relapse	Relapse
	CAM.11	0		CAM.11	0.077			
				CAM.15	0.013			
10R$\rightarrow$ 10RR	CAM.00	1.833	0.277	CAM.05	0.231	0.143	Reinfection	Reinfection
	CAM.11	0						
	CAM.15	0						
31$\rightarrow$ 31R	CAM.00	1.833	0.969	CAM.16	0.013	0.995	Relapse	Reinfection
	CAM.02	0.892						
	CAM.04	3.519						
	CAM.31	0						
36$\rightarrow$ 36R	CAM.00	1.833	0.960	CAM.01	0.269	0.870	Relapse	Relapse
	CAM.01	0.469		CAM.02	0.410			
	CAM.02	0.892		CAM.07	0.192			
	CAM.03	0		CAM.17	0.064			
	CAM.04	3.519						
	CAM.05	-1.085						
	CAM.06	-1.416						
	CAM.07	1.750						
	CAM.09	0						
	CAM.11	0						
68$\rightarrow$ 68R	CAM.00	1.833	0.989	CAM.10	0.077	1.000	Relapse	Relapse
	CAM.02	0.892						
	CAM.04	3.519						
	CAM.10	1.034						
80 $\rightarrow$ 80R	CAM.00	1.833	0.992	CAM.00	0.590	0.000	Reinfection	Relapse
	CAM.04	3.519		CAM.01	0.269			
	CAM.05	-1.085		CAM.02	0.410			
	CAM.08	0.395		CAM.03	0.295			
	CAM.09	0		CAM.05	0.231			
	CAM.24	2.954		CAM.06	0.231			
	CAM.27	0		CAM.07	0.192			
				CAM.08	0.154			
				CAM.12	0.064			
				CAM.41	0.013			
80R$\rightarrow$ 80RR	CAM.00	1.833	0.673	CAM.00	0.590	0.340	Reinfection	Relapse
	CAM.01	0.469		CAM.02	0.410			
	CAM.02	0.892		CAM.04	0.346			
	CAM.03	0		CAM.06	0.231			
	CAM.05	-1.085		CAM.08	0.154			
	CAM.06	-1.416		CAM.12	0.064			
	CAM.07	1.750		CAM.59	0.013			
	CAM.08	0.395						
	CAM.12	0.677						
	CAM.41	0						

Table 2
Classification of all recurrence pairs based on our proposed method when $\mu = 0.05$ and binomial probability model (continued)

Recurrence pair	Baseline variants	β	$\widehat{\xi}^{(0)}$	Recurrence variants	Variant prevalence	$\widehat{\xi}^{(1)}$	Proposed Class	BPM Class
81 → 81R	CAM.00	1.833	0.958	CAM.00	0.590	0.994	Relapse	Relapse
	CAM.01	0.469		CAM.01	0.269			
	CAM.51	3.607						
81R→ 81RR	CAM.00	1.833	0.379	CAM.00	0.590	0.877	Relapse	Relapse
	CAM.01	0.469		CAM.01	0.269			
81RR→ 81RRR	CAM.00	1.833	0.379	CAM.00	0.590	0.882	Relapse	Relapse
	CAM.01	0.469		CAM.01	0.269			
				CAM.66	0.013			
82 → 82R	CAM.00	1.833	0.973	CAM.00	0.590	0.849	Relapse	Relapse
	CAM.03	0		CAM.01	0.269			
	CAM.04	3.519		CAM.03	0.295			
	CAM.10	1.034		CAM.46	0.013			
82R → 82RR	CAM.00	1.833	0.379	CAM.00	0.590	0.995	Relapse	Relapse
	CAM.01	0.469		CAM.01	0.269			
	CAM.03	0		CAM.03	0.295			
	CAM.46	0		CAM.46	0.013			
87 → 87R	CAM.00	1.833	0.977	CAM.00	0.590	0.936	Relapse	Relapse
	CAM.01	0.469		CAM.07	0.192			
	CAM.02	0.892		CAM.08	0.154			
	CAM.08	0.395		CAM.53	0.013			
	CAM.24	2.954						
89 → 89R	CAM.00	1.833	0.963	CAM.01	0.269	0.060	Reinfection	Reinfection
	CAM.04	3.519		CAM.09	0.077			
	CAM.06	-1.416		CAM.20	0.026			
	CAM.08	0.395		CAM.27	0.038			
	CAM.10	1.034						
	CAM.12	0.677						
96 → 96R	CAM.00	1.833	0.979	CAM.00	0.590	0.992	Relapse	Reinfection
	CAM.02	0.892		CAM.30	0.013			
	CAM.04	3.519						
	CAM.08	0.395						
112 → 112R	CAM.00	1.833	0.998	CAM.00	0.590	0.995	Relapse	Relaspe
	CAM.01	0.469		CAM.01	0.269			
	CAM.02	0.892		CAM.02	0.410			
	CAM.04	3.519						
	CAM.07	1.750						
	CAM.12	0.677						
	CAM.40	0						
	CAM.42	0						
	CAM.60	0						
118 → 118R	CAM.08	0.395	0.083	CAM.01	0.269	0.002	Reinfection	Reinfection
				CAM.02	0.410			
				CAM.25	0.013			
				CAM.39	0.013			

Table 3
Classification of all recurrence pairs based on our proposed method and binomial probability model (continued)

Recurrence pair	Baseline variants	β	$\widehat{\xi}^{(0)}$	Recurrence variants	Variant prevalence	$\widehat{\xi}^{(1)}$	Proposed Class	BPM Class
123 → 123R	CAM.00	1.833	0.483	CAM.00	0.590	0.190	Reinfection	Reinfection
	CAM.02	0.892		CAM.01	0.269			
125 → 125R	CAM.02	0.892	0.130	CAM.00	0.590	0.000	Reinfection	Reinfection
				CAM.01	0.269			
				CAM.02	0.410			
				CAM.04	0.346			
				CAM.09	0.077			
				CAM.13	0.013			
				CAM.14	0.026			
				CAM.38	0.013			
				CAM.45	0.013			
126 → 126R	CAM.00	1.833	0.960	CAM.01	0.269	0.975	Relapse	Relapse
	CAM.01	0.469		CAM.07	0.192			
	CAM.02	0.892		CAM.33	0.013			
	CAM.03	0						
	CAM.04	3.519						
	CAM.05	-1.085						
	CAM.06	-1.416						
	CAM.07	1.750						
	CAM.22	0						
	CAM.50	0						
130 → 130R	CAM.00	1.833	0.984	CAM.00	0.590	0.999	Relapse	Relapse
	CAM.02	0.892		CAM.04	0.346			
	CAM.03	0		CAM.12	0.064			
	CAM.04	3.519						
	CAM.12	0.677						
130R → 130RR	CAM.00	1.833	0.962	CAM.00	0.590	0.860	Relapse	Reinfection
	CAM.04	3.519		CAM.07	0.192			
	CAM.12	0.677						
151 → 151R	CAM.03	0	0.030	CAM.00	0.590	0.005	Reinfection	Reinfection
	CAM.05	-1.085		CAM.08	0.154			
	CAM.08	0.395		CAM.14	0.026			
				CAM.64	0.013			
152 → 152R	CAM.00	1.833	0.379	CAM.00	0.590	0.018	Reinfection	Relapse
	CAM.01	0.469		CAM.01	0.269			
				CAM.05	0.231			
				CAM.07	0.192			
153 → 153R	CAM.00	1.833	0.987	CAM.02	0.410	0.819	Relapse	Reinfection
	CAM.04	3.519		CAM.20	0.026			
	CAM.07	1.750						
	CAM.55	0						
154 → 154R	CAM.00	1.833	0.085	CAM.03	0.295	0.003	Reinfection	Relapse
	CAM.06	-1.085		CAM.05	0.231			
	CAM.57	0		CAM.06	0.231			
154R→ 154RR	CAM.03	0	0.005	CAM.02	0.410	0.001	Reinfection	Reinfection
	CAM.05	-1.085		CAM.52	0.013			
	CAM.06	-1.416						

Table 4
Classification of all recurrence pairs based on our proposed method and binomial probability model (continued)

Recurrence pair	Baseline variants	β	$\widehat{\xi}^{(0)}$	Recurrence variants	Variant prevalence	$\widehat{\xi}^{(1)}$	Proposed Class	BPM Class
160 $\rightarrow$ 160R	CAM.02	0.892	0.967	CAM.00	0.590	0.001	Reinfection	Reinfection
	CAM.04	3.519		CAM.03	0.295			
	CAM.07	1.750		CAM.05	0.231			
				CAM.10	0.077			
				CAM.61	0.013			
177 $\rightarrow$ 177R	CAM.00	1.833	0.987	CAM.01	0.269	0.964	Relapse	Reinfection
	CAM.04	3.519						
	CAM.07	1.750						
179 $\rightarrow$ 179R	CAM.03	0	0.106	CAM.01	0.269	0.011	Reinfection	Reinfection
	CAM.05	-1.085		CAM.13	0.013			
	CAM.07	1.750						
	CAM.09	0						
	CAM.17	0						
	CAM.22	0						

Table 5

Bias of regression coefficient estimation when the true reinfection rate $\mu = 0.05$ is misspecified

Scenario	$\widehat{\mu}$	n	α	β_1	β_2	β_3	β_4	β_5
1	0.02	100	-0.356	0.072	0.065	0.045	-0.008	-0.084
		200	0.100	-0.006	-0.032	-0.005	-0.007	0.004
		400	0.161	-0.027	-0.037	-0.013	-0.001	0.011
		800	0.187	-0.037	-0.042	-0.027	-0.003	0.007
	0.10	100	-55.10	11.50	9.050	6.766	-2.796	-2.139
		200	-24.17	5.586	4.561	4.706	-0.951	-1.731
		400	-1.804	0.536	0.401	0.416	-0.269	-0.113
		800	-0.949	0.241	0.225	0.241	-0.012	0.007
2	0.02	100	-1.074	0.037	0.009	-0.096	0.104	-0.191
		200	0.105	0.014	-0.016	0.002	-0.021	-0.016
		400	0.174	0.008	-0.016	0.005	-0.011	0.001
		800	0.201	0.008	-0.009	-0.003	-0.004	-0.003
	0.10	100	-64.01	-1.857	-3.820	-3.116	-4.760	-6.269
		200	-53.34	0.184	-2.819	-1.285	-3.319	-3.872
		400	-20.86	-0.373	-4.044	-0.358	-0.218	-1.472
		800	-1.838	-0.009	0.114	-0.105	-0.001	-0.111

Scenario	$\widehat{\mu}$	n	β_6	β_7	β_8	β_9	β_{10}
1	0.02	100	-0.148	-0.440	-0.356	-0.382	-0.410
		200	-0.037	-0.054	-0.064	-0.022	-0.067
		400	-0.017	-0.025	-0.023	-0.019	-0.033
		800	-0.002	-0.012	-0.010	-0.008	-0.019
	0.10	100	-5.919	-7.050	-6.537	-6.074	-7.842
		200	-3.290	-2.835	-2.366	-2.958	-3.675
		400	-0.310	-0.417	-0.360	-0.413	-0.279
		800	-0.016	-0.034	-0.026	-0.032	-0.068
2	0.02	100	-0.606	-0.554	-0.056	-0.307	-0.065
		200	-0.048	-0.053	-0.049	-0.060	-0.051
		400	-0.011	-0.018	-0.051	-0.068	-0.055
		800	-0.002	-0.006	-0.050	-0.060	-0.054
	0.10	100	-8.569	-8.220	7.486	5.996	5.480
		200	-5.583	-7.793	5.878	4.601	1.659
		400	-3.714	-4.043	3.412	1.983	2.436
		800	-0.039	-0.356	0.381	0.310	0.354

Table 6
Operating characteristics of proposed classifiers when the true reinfection rate $\mu = 0.05$ is misspecified

			BPM			$I(\widehat{\xi}_i^{(0)} > 0.5)$			$I(\widehat{\xi}_i^{(1)} > 0.5)$		
Scenario	$\widehat{\mu}$	n	sens	spec	acc	sens	spec	acc	sens	spec	acc
1	0.02	100	89.1	83.3	88.2	98.9	1.3	82.6	99.0	75.3	94.8
		200	89.2	83.7	88.3	99.9	0.0	83.6	99.6	78.6	96.1
		400	89.3	84.2	88.5	100.0	0.0	83.8	99.6	79.7	96.3
		800	89.4	84.0	88.5	100.0	0.0	83.7	99.6	80.5	96.5
	0.05	100	89.1	83.3	88.2	89.6	12.0	76.5	93.2	82.4	91.2
		200	89.2	83.7	88.3	97.0	3.8	81.8	98.4	87.2	96.5
		400	89.3	84.2	88.5	99.5	0.7	83.4	98.8	87.8	97.0
		800	89.4	84.0	88.5	100.0	0.0	83.7	98.9	88.3	97.2
	0.10	100	89.1	83.3	88.2	51.8	52.9	52.1	56.7	77.5	60.2
		200	89.2	83.7	88.3	54.3	51.2	53.8	73.1	91.8	76.2
		400	89.3	84.2	88.5	61.9	45.7	59.3	92.7	94.5	93.0
		800	89.4	84.0	88.5	64.8	45.1	61.5	95.6	94.7	95.5
2	0.02	100	89.2	85.1	88.3	97.6	2.3	77.7	98.3	74.3	92.8
		200	89.3	84.9	88.4	99.8	0.2	79.2	99.5	79.4	95.2
		400	89.4	84.5	88.4	100.0	0.0	79.5	99.5	80.8	95.6
		800	89.5	84.2	88.5	100.0	0.0	79.4	99.6	81.8	95.9
	0.05	100	89.2	85.1	88.3	81.4	19.8	68.5	85.8	79.5	84.3
		200	89.3	84.9	88.4	92.3	8.3	74.9	96.7	88.4	94.9
		400	89.4	84.5	88.4	98.2	2.0	78.5	98.4	88.8	96.4
		800	89.5	84.2	88.5	99.8	0.2	79.3	98.5	89.5	96.6
	0.10	100	89.2	85.1	88.3	46.1	57.1	48.3	48.7	74.0	53.7
		200	89.3	84.9	88.4	37.4	66.9	43.5	49.1	89.0	57.3
		400	89.4	84.5	88.4	38.2	67.3	44.2	70.7	95.5	75.8
		800	89.5	84.2	88.5	38.7	68.6	44.9	90.8	96.2	91.9