Supporting Information for the manuscript entitled "Classification of Disease Recurrence using Transition Likelihoods with EM algorithm"

by Huijun Jiang, Quefeng Li, Jessica T. Lin, Feng-Chang Lin

1 Out-of-Sample Prediction Comparison

In addition to the classification results shown in Table 1 of the main manuscript, we also report out-of-sample prediction comparison in 10-fold cross-validation using 90% of the sample as the training set and the remaining 10% as the testing set. Table S1 shows the operating characteristics of the classification results of the samples when they are in the testing set using the two-stage method and EM algorithm. We report the classification accuracy such as sensitivity, specificity, and overall accuracy like the manuscript. We also report the area under the receiver operating characteristic (ROC) curve, commonly known as AUC or c-statistics.

The prediction results are similar to Table 1 of the main manuscript. Both classifiers perform aggressively under a low transition probability, resulting in high sensitivity but low specificity. Both classifiers perform better under a high transition probability, reaching a high degree of accuracy in both sensitivity and specificity. When comparing the two classifiers, $P(Y_i = 2|\mathcal{O}_i, \hat{\theta})$ has higher accuracy and AUC than $\hat{\xi}_i^{(1)}$, especially when the sample size n is small. When the sample size increases, the two-stage method $\hat{\xi}_i^{(1)}$ becomes more competitive and has a similar classification performance. This result is consistent with the in-sample classification performance from two classifiers.

2 Additional Simulation Results

Tables S2 to S5 present simulation results not shown in the main manuscript for regression coefficients β_6 to β_{10} under different scenarios. The bias (b), empirical standard error (σ),

Table S1: Out-of-sample prediction comparison between the two-stage method and EM algorithm using ten-fold cross-validation in sensitivity, specificity, overall accuracy, and AUC, area under the receiver operating characteristics (ROC) curve

			Two-Stage method $\hat{\xi}_i^{(1)} > 0.5$				EM algorithm $P(Y_i = 2\|\mathcal{O}_i, \hat{\theta}) > 0.5$			
μ	γ_0	n	Sensitivity	Specificity	Overall	AUC	Sensitivity	Specificity	Overall	AUC
-3	0	100	71.4%	41.4%	67.9%	53.4%	94.0%	58.0%	91.3%	61.2%
		200	96.7%	43.3%	90.4%	69.5%	97.3%	59.0%	92.9%	72.3%
		400	98.0%	53.1%	92.8%	74.2%	97.9%	59.7%	93.4%	75.0%
	2.94	100	78.0%	45.7%	74.2%	57.3%	96.0%	84.2%	95.4%	68.0%
		200	98.3%	71.0%	95.1%	86.9%	98.5%	86.2%	97.2%	88.1%
		400	99.1%	82.9%	97.2%	90.7%	99.0%	87.3%	97.6%	90.5%
-2	0	100	57.6%	56.6%	56.9%	55.1%	87.0%	67.8%	81.5%	70.8%
		200	89.9%	59.0%	81.6%	72.1%	92.7%	70.9%	86.9%	76.2%
		400	94.3%	66.7%	87.0%	77.2%	94.6%	71.8%	88.6%	78.2%
	2.94	100	63.6%	62.3%	62.9%	61.3%	93.5%	85.1%	92.9%	89.2%
		200	95.6%	81.1%	91.6%	88.1%	97.1%	91.4%	95.6%	91.2%
		400	97.9%	88.6%	95.5%	92.1%	97.8%	91.9%	96.2%	92.4%

estimated standard error ($\hat{\sigma}$), and coverage probability (CP) are defined in the same way as the corresponding table. The interpretation of the results follows what has been shown in the main manuscript.

3 Sensitivity Analyses for the Malaria Data

The reinfection parameter, $\mu = -3$, is decided based on the incidence rate of 5% for both northern and southern Cambodia data. In the first sensitivity analysis, we use a different value of μ to explore the robustness of the classification result in our main manuscript. Tables S6 and S7 present the classification results for northern and southern Cambodia data, respectively, based on the EM algorithm and two-stage method when using $\mu = -2$. The classification results are pretty similar to the main manuscript, indicating the classification result is robust under a different value of μ.

One can see that Table 9 of the main manuscript has some non-convergence issues in β coefficients of rare variants with a large standard error. Here, we conduct another sensitivity analysis removing rare variants with a prevalence of less than 0.02. Table S8 shows the estimation results by the EM algorithm when using only prevalent variants.

Table S2: Additional simulation results not shown in Table 2 of the main manuscript

μ	γ_0	n	b_6	b_7	b_8	b_9	b_{10}	σ_6	σ_7	σ_8	σ_9	σ_{10}
-3	0	100	-0.04	-0.08	-0.08	-0.10	-0.16	0.70	0.86	1.14	1.19	1.54
		200	-0.02	-0.02	-0.03	-0.01	-0.03	0.42	0.43	0.44	0.49	0.48
		400	-0.01	0.00	-0.02	0.00	-0.01	0.28	0.29	0.30	0.31	0.32
	2.94	100	-0.03	-0.07	-0.06	-0.05	-0.15	0.68	0.73	1.13	1.00	1.54
		200	-0.01	-0.02	-0.02	-0.01	-0.03	0.41	0.42	0.42	0.46	0.47
		400	-0.01	0.00	-0.02	-0.01	-0.01	0.27	0.28	0.29	0.30	0.31
-2	0	100	-0.08	-0.11	-0.11	-0.18	-0.24	-0.08	-0.11	-0.11	-0.18	-0.24
		200	-0.02	-0.02	-0.02	-0.03	-0.06	0.47	0.48	0.50	0.54	0.56
		400	-0.01	-0.01	-0.02	-0.01	-0.03	0.31	0.30	0.33	0.36	0.35
	2.94	100	-0.02	-0.03	-0.05	-0.11	-0.20	0.72	0.76	0.96	1.21	1.56
		200	-0.01	-0.02	-0.01	-0.03	-0.05	0.43	0.44	0.45	0.52	0.52
		400	0.01	-0.01	0.00	-0.01	-0.01	0.28	0.29	0.30	0.31	0.33

μ	γ_0	n	$\hat{\sigma}_6$	$\hat{\sigma}_7$	$\hat{\sigma}_8$	$\hat{\sigma}_9$	$\hat{\sigma}_{10}$	CP_6	CP_7	CP_8	CP_9	CP_{10}
-3	0	100	0.61	0.63	0.65	0.67	0.71	94.5	94.0	96.0	95.0	94.3
		200	0.40	0.41	0.43	0.44	0.46	94.5	95.1	95.7	95.3	95.3
		400	0.27	0.28	0.29	0.30	0.31	94.4	93.8	94.9	94.9	94.2
	2.94	100	0.59	0.61	0.63	0.65	0.68	93.9	94.4	96.2	93.8	94.6
		200	0.39	0.40	0.42	0.43	0.45	94.5	94.9	95.6	94.7	94.8
		400	0.27	0.27	0.28	0.29	0.30	94.9	93.9	95.6	95.2	95.1
-2	0	100	0.67	0.69	0.72	0.74	0.77	93.8	92.5	93.3	93.3	95.3
		200	0.45	0.46	0.48	0.50	0.52	94.8	94.7	94.1	95.3	95.9
		400	0.29	0.30	0.31	0.32	0.34	93.7	95.5	94.1	93.5	94.2
	2.94	100	0.62	0.63	0.65	0.68	0.71	93.9	92.8	95.1	93.5	95.5
		200	0.40	0.41	0.42	0.44	0.46	94.0	94.3	94.8	93.8	94.1
		400	0.27	0.28	0.29	0.30	0.31	94.7	95.4	94.6	94.5	93.7

Table S3: Additional simulation results not shown in Table 3 of the main manuscript

μ	γ_0	n	b_6	b_7	b_8	b_9	b_{10}	σ_6	σ_7	σ_8	σ_9	σ_{10}
-3	0	100	-0.13	-0.05	-0.30	-0.20	-0.45	3.41	2.29	2.63	2.88	4.51
		200	-0.01	-0.02	-0.03	-0.06	-0.01	0.44	0.45	0.48	0.75	0.83
		400	0.00	0.00	-0.03	-0.01	-0.01	0.29	0.31	0.32	0.33	0.35
	2.94	100	-0.13	-0.05	-0.30	-0.20	-0.45	3.41	2.29	2.63	2.88	4.51
		200	-0.01	-0.02	-0.03	-0.06	-0.01	0.44	0.45	0.48	0.75	0.83
		400	0.00	0.00	-0.03	-0.01	-0.01	0.29	0.31	0.32	0.33	0.35
-2	0	100	-0.67	-0.82	-1.74	-1.79	-0.84	14.55	13.27	20.15	18.50	16.05
		200	-0.05	-0.07	-0.05	-0.08	-0.11	0.67	0.94	1.11	1.00	1.62
		400	-0.01	-0.02	-0.01	-0.04	-0.02	0.35	0.37	0.38	0.41	0.41
	2.94	100	-0.67	-0.82	-1.74	-1.79	-0.84	14.55	13.27	20.15	18.50	16.05
		200	-0.05	-0.07	-0.05	-0.08	-0.11	0.67	0.94	1.11	1.00	1.62
		400	-0.01	-0.02	-0.01	-0.04	-0.02	0.35	0.37	0.38	0.41	0.41

μ	γ_0	n	$\hat{\sigma}_6$	$\hat{\sigma}_7$	$\hat{\sigma}_8$	$\hat{\sigma}_9$	$\hat{\sigma}_{10}$	CP_6	CP_7	CP_8	CP_9	CP_{10}
-3	0	100	14.95	14.55	15.67	16.58	17.44	99.7	100.0	100.0	99.9	100.0
		200	1.35	1.55	1.64	1.91	2.15	99.7	99.8	99.7	99.4	99.5
		400	0.32	0.33	0.35	0.37	0.39	96.9	96.7	96.7	97.5	97.4
	2.94	100	11.02	11.44	12.26	12.83	13.56	99.7	100.0	100.0	99.9	100.0
		200	1.35	1.55	1.64	1.91	2.15	99.7	99.8	99.7	99.4	99.5
		400	0.32	0.33	0.35	0.37	0.39	97.0	98.0	100.0	97.0	98.0
-2	0	100	21.95	23.92	24.11	26.20	26.37	99.1	99.3	98.8	98.4	98.5
		200	7.35	7.90	8.60	8.90	9.59	100.0	99.9	100.0	100.0	100.0
		400	0.56	0.65	0.44	0.40	0.37	98.7	98.8	87.7	84.4	81.8
	2.94	100	21.95	23.92	24.11	26.20	26.37	99.1	99.3	98.8	98.4	98.5
		200	7.35	7.90	8.60	8.90	9.59	100.0	99.9	100.0	100.0	100.0
		400	0.56	0.65	0.67	0.79	0.89	98.7	98.8	98.5	98.8	99.0

Table S4: Additional simulation results not shown in Table 4 of the main manuscript

μ	γ_0	n	b_5	b_6	b_7	b_8	b_9	b_{10}	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}
-3	0	100	-0.01	-0.07	-0.01	-0.01	-0.18	-0.13	0.92	1.27	0.76	0.92	1.33	1.44
		200	-0.02	-0.02	-0.01	-0.02	-0.02	-0.02	0.42	0.42	0.44	0.45	0.44	0.50
		400	-0.01	-0.01	0.00	0.00	-0.01	-0.01	0.27	0.27	0.29	0.30	0.30	0.32
	2.94	100	-0.03	-0.01	-0.02	0.01	-0.13	-0.09	0.65	0.67	0.85	0.70	1.16	1.28
		200	-0.01	-0.01	0.00	-0.02	-0.02	-0.02	0.40	0.41	0.42	0.43	0.43	0.49
		400	-0.01	0.00	0.00	0.00	-0.01	-0.01	0.26	0.26	0.28	0.29	0.29	0.31
-2	0	100	-0.02	-0.04	-0.04	-0.06	-0.25	-0.12	1.12	1.57	1.78	1.68	2.03	2.30
		200	0.01	-0.01	-0.02	0.00	-0.04	-0.03	0.44	0.45	0.48	0.49	0.62	0.53
		400	0.00	0.00	0.00	0.01	-0.01	-0.02	0.29	0.29	0.31	0.32	0.32	0.35
	2.94	100	-0.03	0.01	-0.01	-0.04	-0.11	-0.11	0.86	0.70	0.87	0.76	1.22	1.44
		200	0.01	-0.01	-0.01	0.00	-0.02	-0.03	0.40	0.41	0.45	0.45	0.46	0.48
		400	0.00	0.00	0.00	0.00	-0.01	-0.02	0.28	0.27	0.29	0.29	0.30	0.33

μ	γ_0	n	$\hat{\sigma}_5$	$\hat{\sigma}_6$	$\hat{\sigma}_7$	$\hat{\sigma}_8$	$\hat{\sigma}_9$	$\hat{\sigma}_{10}$	CP_5	CP_6	CP_7	CP_8	CP_9	CP_{10}
-3	0	100	0.60	0.61	0.63	0.65	0.67	0.71	94.9	93.4	94.7	93.4	96.2	94.7
		200	0.38	0.40	0.41	0.42	0.44	0.45	93.8	95.6	95.6	95.2	95.4	93.8
		400	0.26	0.27	0.28	0.29	0.30	0.31	95.3	95.4	94.8	94.6	95.8	94.3
	2.94	100	0.57	0.59	0.61	0.62	0.65	0.68	95.1	94.0	93.6	94.8	96.1	93.9
		200	0.38	0.39	0.40	0.41	0.43	0.44	93.6	95.2	95.9	95.3	95.8	93.4
		400	0.26	0.26	0.27	0.28	0.29	0.30	95.8	96.1	94.9	94.6	95.9	94.7
-2	0	100	0.66	0.67	0.70	0.73	0.74	0.77	93.2	93.4	93.7	93.4	95.8	93.4
		200	0.42	0.43	0.44	0.46	0.47	0.49	95.6	95.4	94.6	94.6	95.5	95.5
		400	0.28	0.29	0.30	0.31	0.32	0.33	94.0	95.3	95.9	94.3	95.4	95.0
	2.94	100	0.59	0.61	0.63	0.65	0.67	0.70	95.3	94.2	94.4	94.2	94.1	95.4
		200	0.39	0.40	0.41	0.42	0.44	0.45	95.0	95.3	95.0	93.9	95.5	94.5
		400	0.27	0.27	0.28	0.29	0.30	0.31	94.0	95.1	94.6	95.1	95.2	95.0

Table S5: Additional simulation results not shown in Table 5 of the main manuscript

μ	γ_0	n	b_5	b_6	b_7	b_8	b_9	b_{10}	σ_5	σ_6	σ_7	σ_8	σ_9	σ_{10}
-3	0	100	-0.01	-0.05	-0.02	-0.01	-0.18	-0.15	0.86	0.99	0.80	0.84	1.38	1.51
		200	-0.01	-0.02	-0.01	-0.02	-0.02	-0.02	0.41	0.42	0.43	0.45	0.44	0.50
		400	-0.01	-0.01	-0.01	0.00	-0.01	-0.01	0.26	0.27	0.29	0.30	0.30	0.32
	2.94	100	-0.03	-0.02	-0.01	0.02	-0.13	-0.09	0.65	0.67	0.69	0.70	1.19	1.30
		200	-0.01	-0.01	0.00	-0.01	-0.02	-0.02	0.39	0.41	0.41	0.43	0.43	0.49
		400	-0.01	0.00	0.00	0.00	-0.01	-0.01	0.26	0.26	0.28	0.29	0.29	0.31
-2	0	100	-0.03	-0.06	-0.04	-0.08	-0.23	-0.11	1.04	1.58	1.24	1.16	1.96	1.56
		200	0.02	-0.01	-0.02	-0.01	-0.03	-0.03	0.43	0.44	0.49	0.49	0.50	0.52
		400	0.00	0.00	0.00	0.01	-0.01	-0.02	0.29	0.29	0.31	0.32	0.32	0.35
	2.94	100	-0.02	0.01	-0.01	-0.04	-0.11	-0.10	0.86	0.69	0.86	0.75	1.23	1.35
		200	0.01	-0.01	-0.01	0.00	-0.02	-0.03	0.40	0.41	0.45	0.45	0.46	0.48
		400	0.00	0.00	0.00	0.00	-0.01	-0.01	0.28	0.27	0.29	0.29	0.30	0.33

μ	γ_0	n	$\hat{\sigma}_5$	$\hat{\sigma}_6$	$\hat{\sigma}_7$	$\hat{\sigma}_8$	$\hat{\sigma}_9$	$\hat{\sigma}_{10}$	CP_5	CP_6	CP_7	CP_8	CP_9	CP_{10}
-3	0	100	0.60	0.61	0.63	0.65	0.67	0.71	94.2	94.0	94.7	94.5	96.6	94.9
		200	0.38	0.40	0.41	0.42	0.44	0.45	94.3	95.5	95.8	95.1	95.5	93.3
		400	0.26	0.27	0.28	0.29	0.30	0.31	95.3	95.1	94.1	94.8	95.4	94.6
	2.94	100	0.57	0.59	0.61	0.63	0.65	0.68	95.5	94.2	93.5	94.8	96.2	94.1
		200	0.38	0.39	0.40	0.41	0.43	0.44	93.5	94.9	95.7	95.3	95.9	93.4
		400	0.26	0.26	0.27	0.28	0.29	0.30	95.7	96.5	94.8	94.5	95.8	94.5
-2	0	100	0.65	0.67	0.69	0.72	0.73	0.76	94.7	94.0	93.8	93.8	95.0	94.0
		200	0.42	0.43	0.44	0.45	0.47	0.49	95.1	95.5	95.8	94.3	95.6	94.6
		400	0.28	0.29	0.30	0.30	0.32	0.33	94.0	94.7	95.4	94.2	95.6	94.7
	2.94	100	0.59	0.61	0.63	0.65	0.67	0.70	95.1	93.9	94.2	94.7	94.2	95.4
		200	0.39	0.40	0.41	0.42	0.44	0.45	94.8	95.2	95.4	94.4	95.6	94.7
		400	0.26	0.27	0.28	0.29	0.30	0.31	94.1	95.1	94.6	94.9	95.7	94.8

Table S6: Classification of recurrence pairs based on the EM algorithm and two-stage method for the northern Cambodia data using $\mu = -2$

Recurrence Pair	Baseline Variants	$\pi_{i2}(x_i, \hat{\theta})$	Recurrence Variants	$P(Y_i = 2 \mid \mathcal{O}_i, \hat{\theta})$	EM Class	Two-Stage Class
10 → 10R	A	0.00	A	0.00	Reinfection	Relapse
31 → 31R	ECA	0.86		1.00	Relapse	Relapse
36 → 36R	JHGFEDCBA	0.69	HCB	0.97	Relapse	Relapse
68 → 68R	ECA	0.86		1.00	Relapse	Relapse
80 → 80R	JIFEA	0.00	IHGFDCBA	0.00	Reinfection	Reinfection
81 → 81R	BA	0.21	BA	0.90	Relapse	Relapse
82 → 82R	EDA	0.81	DBA	0.96	Relapse	Relapse
87 → 87R	ICBA	1.00	IHA	1.00	Relapse	Relapse
89 → 89R	IGEA	1.00	JB	1.00	Relapse	Reinfection
96 → 96R	IECA	1.00	DA	1.00	Relapse	Relapse
112 → 112R	HECBA	1.00	CBA	1.00	Relapse	Relapse
118 → 118R	I	0.00	CB	0.00	Reinfection	Reinfection
123 → 123R	CA	0.00	BA	0.00	Reinfection	Reinfection
125 → 125R	C	0.00	JECBA	0.00	Reinfection	Reinfection
126 → 126R	HGFEDCBA	0.69	HB	0.98	Relapse	Relapse
130 → 130R	EDCA	0.90	EA	0.99	Relapse	Relapse
151 → 151R	IFD	0.00	IA	0.00	Reinfection	Reinfection
152 → 152R	BA	0.21	HFBA	0.56	Relapse	Reinfection
153 → 153R	HEA	0.82	C	0.96	Relapse	Relapse
154 → 154R	GA	0.00	GFD	0.00	Reinfection	Reinfection
160 → 160R	HEC	0.00	FDA	0.00	Reinfection	Reinfection
177 → 177R	HEA	0.82	B	0.98	Relapse	Relapse
179 → 179R	JHFD	0.00	B	0.00	Reinfection	Reinfection

Dominant variants with more than 50% frequency are presented in italic.

Table S7: classification based on the EM algorithm and two-stage method for the southern Cambodia data using $\mu = -2$

Recurrence Pair	Baseline Variants	$\pi_{i2}(x_i, \hat{\theta})$	Recurrence Variants	$P(Y_i = 2 \mid \mathcal{O}_i, \hat{\theta})$	EM Class	Two-Stage Class
4 → 4R	D *S*	0.16	D *S*	0.98	Relapse	Relapse
7 → 7R	H	0.72	H	1.00	Relapse	Relapse
10 → 10R	H *S*	0.75	S	1.00	Relapse	Relapse
11 → 11R	C F *S*	0.21	C D *F* G *S*	0.00	Reinfection	Reinfection
16 → 16R	C *H* S	0.52	C H	0.99	Relapse	Relapse
17 → 17R	C *H*	0.48	H	1.00	Relapse	Relapse
20 → 20R	D G H S	0.75	D G S	1.00	Relapse	Relapse
25 → 25R	C E F *G*	0.13	C G	0.77	Relapse	Relapse
27 → 27R	C D *G* S	0.11	G *S*	0.87	Relapse	Relapse
28 → 28R	C E G H	0.38	C D *G*	0.00	Reinfection	Reinfection
29 → 29R	F *S*	0.43	F *S*	1.00	Relapse	Relapse
30 → 30R	G	0.37	F S	0.00	Reinfection	Reinfection
42 → 42R	G H	0.83	G	1.00	Relapse	Relapse
44 → 44R	C D F G *S*	0.21	C D E F G *M*	0.00	Reinfection	Reinfection
47 → 47R	D G	0.23	D E *S*	0.00	Reinfection	Reinfection
49 → 49R	C D F H *S*	0.54	C D *H* S	0.99	Relapse	Relapse
50 → 50R	H	0.72	H	1.00	Relapse	Relapse
51 → 51R	D H *S*	0.61	H S	1.00	Relapse	Relapse
52 → 52R	F *S*	0.43	S	0.99	Relapse	Relapse
54 → 54R	F	0.39	G *S*	0.00	Reinfection	Reinfection
56 → 56R	C D *G*	0.10	C	0.41	Reinfection	Relapse
57 → 57R	E G *S*	0.19	D G *S*	0.01	Reinfection	Relapse
59 → 59R	E G *M N*	1.00	E G *M N*	1.00	Relapse	Relapse
62 → 62R	D F G *S*	0.43	D G *S*	0.99	Relapse	Relapse
63 → 63R	D H S	0.61	D H S	1.00	Relapse	Relapse
66 → 66R	C *F* G S	0.34	C F	0.90	Relapse	Relapse
68 → 68R	*E* G	0.17	C D G H *S*	0.00	Reinfection	Reinfection
70 → 70R	C *D* G S	0.11	C *S*	0.41	Reinfection	Relapse
74 → 74R	H	0.72	E F N S	0.00	Reinfection	Reinfection
75 → 75R	E F H	0.65	E	1.00	Relapse	Relapse
77 → 77R	C D	0.05	C D	0.84	Relapse	Relapse
78 → 78R	C *H*	0.48	C *H*	1.00	Relapse	Relapse
80 → 80R	F	0.39	F	1.00	Relapse	Relapse
83 → 83R	C *F*	0.19	F	0.98	Relapse	Relapse
84 → 84R	S	0.27	C D *S*	0.00	Reinfection	Reinfection
85 → 85R	C *S*	0.12	C *S*	0.93	Relapse	Relapse
90 → 90R	G	0.37	G	1.00	Relapse	Relapse
95 → 95R	D *E F* G R	0.67	D *F* R	1.00	Relapse	Relapse
100 → 100R	C D G *R*	0.51	C *G*	0.91	Relapse	Relapse
101 → 101R	C E G *K* N	0.48	C E *N*[8]	1.00	Relapse	Relapse
104 → 104R	H	0.72	H	1.00	Relapse	Relapse
105 → 105R	H	0.72	H	1.00	Relapse	Relapse
107 → 107R	B C F G	0.10	C F	0.66	Relapse	Relapse
109 → 109R	C F G *H* S	0.81	C D G *H*	0.05	Reinfection	Relapse

Dominant variants are presented in italic.

The β coefficient estimation is close to Table 8 in the main manuscript. We assess the significance using a Wald-type test; variants G and H stay statistically significant as in the main manuscript. Table S9 shows the classification results for the southern Cambodia data using only prevalent variants. The EM algorithm classification results are the same as Table 8 in the main manuscript. The classification results using the two-stage method are similar as well. The similarity of the estimation and classification results indicates the robustness of our method. Since the variant information is biologically important and rare variants also have meaningful information for classification, we show the analysis with all variants in the main manuscript.

Table S8: Estimation of regression coefficients by the EM algorithm for the southern Cambodia data using only common variants

Variants	Prevalence	$\hat{\beta}$	SE	p-value
B	0.033	-1.47	1.42	0.3
C	0.582	-0.64	0.54	0.233
D	0.399	-0.59	0.59	0.317
E	0.137	-0.05	0.92	0.953
F	0.294	0.81	0.57	0.155
G	0.320	1.6	0.57	0.005
H	0.229	2.61	0.66	< 0.001
K	0.026	-0.16	1.58	0.922
N	0.039	1.58	1.53	0.302
R	0.039	1.29	1.18	0.274
S	0.444	0.38	0.59	0.525

Table S9: classification based on the EM algorithm and two-stage method for the southern Cambodia data using only common variants

Recurrence Pair	Baseline Variants	$\pi_{i2}(x_i, \hat{\theta})$	Recurrence Variants	$P(Y_i = 2 \mid \mathcal{O}_i, \hat{\theta})$	EM Class	Two-Stage Class
4 → 4 R	D *S*	0.12	D *S*	0.99	Relapse	Relapse
7 → 7 R	H	0.69	H	1.00	Relapse	Relapse
10 → 10 R	H *S*	0.77	S	1.00	Relapse	Relapse
11 → 11 R	C F *S*	0.22	C D *F* G *S*	0.05	Reinfection	Reinfection
16 → 16 R	C *H* S	0.63	C H	1.00	Relapse	Relapse
17 → 17 R	C *H*	0.54	H	1.00	Relapse	Relapse
20 → 20 R	D G H S	0.90	D G S	1.00	Relapse	Relapse
25 → 25 R	C E F *G*	0.48	C G	0.98	Relapse	Relapse
27 → 27 R	C D *G* S	0.26	G *S*	0.97	Relapse	Relapse
28 → 28 R	C E G H	0.85	C D *G*	0.81	Relapse	Relapse
29 → 29 R	F *S*	0.35	F *S*	1.00	Relapse	Relapse
30 → 30 R	G	0.45	F S	0.09	Reinfection	Relapse
42 → 42 R	G H	0.92	G	1.00	Relapse	Relapse
44 → 44 R	C D F G *S*	0.44	C D E F G *M*	0.86	Relapse	Relapse
47 → 47 R	D G	0.31	D E *S*	0.13	Reinfection	Relapse
49 → 49 R	C D F H *S*	0.68	C D *H* S	1.00	Relapse	Relapse
50 → 50 R	H	0.69	H	1.00	Relapse	Relapse
51 → 51 R	D H *S*	0.64	H S	1.00	Relapse	Relapse
52 → 52 R	F *S*	0.35	S	0.99	Relapse	Relapse
54 → 54 R	F	0.27	G *S*	0.05	Reinfection	Reinfection
56 → 56 R	C D *G*	0.19	C	0.81	Relapse	Relapse
57 → 57 R	E G *S*	0.53	D G *S*	0.87	Relapse	Relapse
59 → 59 R	E G *M N*	0.79	E G *M N*	1.00	Relapse	Relapse
62 → 62 R	D F G *S*	0.60	D G *S*	1.00	Relapse	Relapse
63 → 63 R	D H S	0.64	D H S	1.00	Relapse	Relapse
66 → 66 R	C *F* G S	0.58	C F	0.98	Relapse	Relapse
68 → 68 R	*E* G	0.44	C D G H *S*	0.00	Reinfection	Reinfection
70 → 70 R	C *D* G S	0.26	C *S*	0.84	Relapse	Relapse
74 → 74 R	H	0.69	E F N S	0.01	Reinfection	Reinfection
75 → 75 R	E F H	0.83	E	1.00	Relapse	Relapse
77 → 77 R	C D	0.05	C D	0.90	Relapse	Relapse
78 → 78 R	C *H*	0.54	C *H*	1.00	Relapse	Relapse
80 → 80 R	F	0.27	F	1.00	Relapse	Relapse
83 → 83 R	C *F*	0.16	F	0.99	Relapse	Relapse
84 → 84 R	S	0.19	C D *S*	0.02	Reinfection	Reinfection
85 → 85 R	C *S*	0.11	C *S*	0.96	Relapse	Relapse
90 → 90 R	G	0.45	G	1.00	Relapse	Relapse
95 → 95 R	D *E F* G R	0.78	D *F* R	1.00	Relapse	Relapse
100 → 100 R	C D G *R*	0.46	C *G*	0.96	Relapse	Relapse
101 → 101 R	C E G *K* N	0.63	C E N	1.00	Relapse	Relapse
104 → 104 R	H	0.69	H	1.00	Relapse	Relapse
105 → 105 R	H	0.69	H	1.00	Relapse	Relapse
107 → 107 R	B C F G	0.18	C F	0.89	Relapse	Relapse
109 → 109 R	C F G *H* S	0.95	C D G *H*	0.98	Relapse	Relapse

Dominant variants are presented in italic.