Supplementary Information for "Confidence intervals for Markov chain transition probabilities based on next generation sequencing reads data"

Lin Wan[1], Xin Kang[2], Jie Ren[3], Fengzhu Sun[3,*]

[1]NCMIS, LSC, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China
[2]School of Mathematical Sciences,
Fudan University, Shanghai 200433, China
[3]Quantitative and Computational Biology Program,
University of Southern California, Los Angeles, California 90089, USA
*Correspondence: fsun@usc.edu

Supplementary Tables

	A	**C**	**G**	**T**
A	0.39	0.15	0.21	0.25
C	0.31	0.19	0.12	0.38
G	0.31	0.23	0.20	0.26
T	0.26	0.18	0.21	0.35

Table S1: The transition probability matrix $P = [p_{ij}]$ of MC applied in the simulation of underlying genome sequences of NGS data sets.

$\tilde{p}_{ij,k}$	**A**	**C**	**G**	**T**	π
AA	0.374	0.140	0.218	0.268	0.127
CA	0.369	0.127	0.211	0.293	0.060
GA	0.390	0.111	0.202	0.297	0.071
TA	0.367	0.158	0.158	0.317	0.081
AC	0.393	0.175	0.135	0.297	0.046
CC	0.398	0.152	0.131	0.319	0.027
GC	0.365	0.148	0.113	0.374	0.032
TC	0.375	0.186	0.120	0.319	0.052
AG	0.355	0.188	0.207	0.250	0.067
CG	0.386	0.144	0.193	0.277	0.020
GG	0.392	0.156	0.212	0.240	0.038
TG	0.376	0.163	0.193	0.268	0.064
AT	0.274	0.174	0.202	0.350	0.098
CT	0.206	0.157	0.218	0.419	0.051
GT	0.273	0.147	0.187	0.393	0.049
TT	0.264	0.172	0.206	0.358	0.117

Table S2: The estimated transition probabilities $\tilde{p}_{ij,k}$ based on the genome sequence and the corresponding stationary distribution π for Bacillus phage SP-beta genome.

		$\mathrm{CI}_{1-\alpha}^{\mathrm{Long}}$					
		99%		95%		90%	
	$\tilde{p}_{AA,A}$	lower	upper	lower	upper	lower	upper
	0.3739	0.3644	0.3834	0.3667	0.3812	0.3678	0.3800
		$\mathrm{CI}_{1-\alpha}^{\mathrm{NGS}}$					
		99%		95%		90%	
$C = \beta c$	$\hat{p}_{AA,A}$	lower	upper	lower	upper	lower	upper
0.5	0.3811	0.3632	0.3995	0.3675	0.3951	0.3698	0.3929
1.0	0.3788	0.3625	0.3952	0.3664	0.3913	0.3684	0.3893
2.0	0.3764	0.3611	0.3917	0.3648	0.3881	0.3666	0.3862
5.0	0.3804	0.3618	0.3990	0.3662	0.3945	0.3685	0.3923
10.0	0.3789	0.3655	0.3923	0.3687	0.3891	0.3703	0.3874

Table S3: The estimated transition probability of $p_{AA,A}$ and theoretical confidence intervals for inhomogeneous NGS data sets of Bacillus phage SP-beta genome.

Supplementary Figures

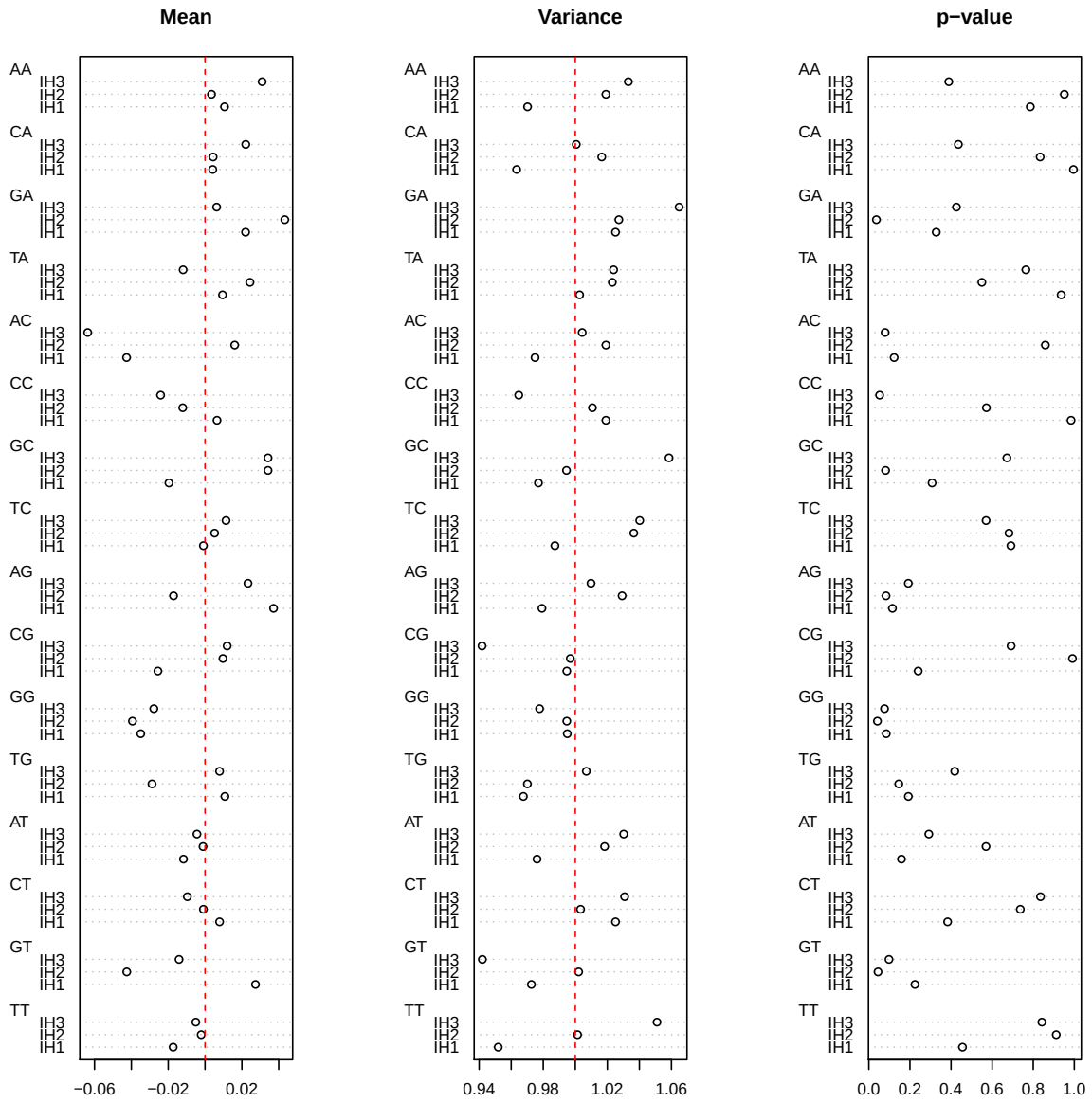

Figure S1: The mean, variance and p-value of statistic $\widetilde{\xi_{ij}^{(R)}}$ for the simulated inhomogeneous data sets of IH1, IH2, and IH3. The p-value is calculated by Kolmogorov-Smirnov test to whether $\widetilde{\xi_{ij}^{(R)}}$ approximates a standard normal distribution.

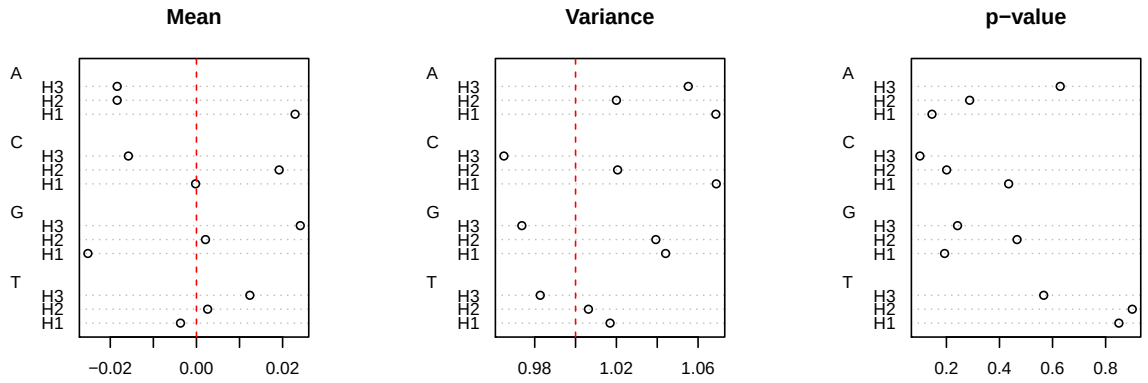

Figure S2: The mean, variance and p-value of statistic $\widetilde{\zeta_i^{(R)}}$ for the simulated homogeneous data sets of H1, H2, and H3. The p-value is calculated by Kolmogorov-Smirnov test to whether $\widetilde{\zeta_i^{(R)}}$ approximates a standard normal distribution.

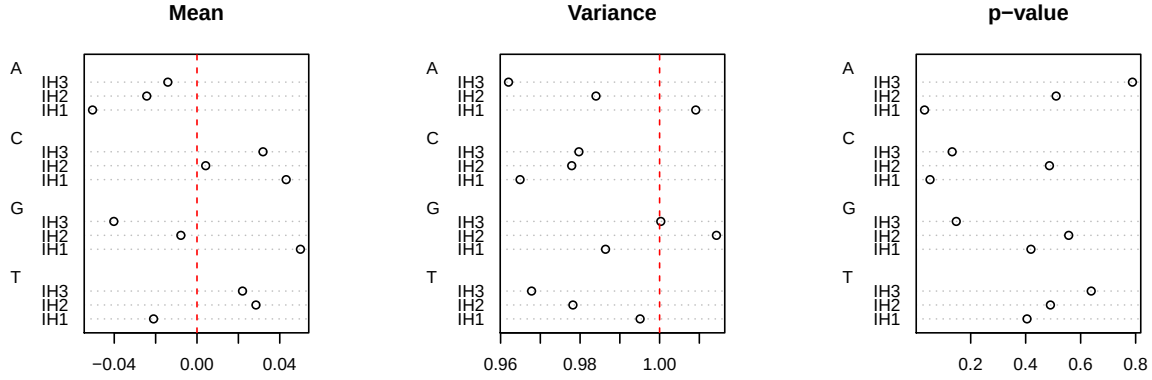

Figure S3: The mean, variance and p-value of statistic $\widetilde{\zeta_i^{(R)}}$ for the simulated inhomogeneous data sets of IH1, IH2, and IH3. The p-value is calculated by Kolmogorov-Smirnov test to whether $\widetilde{\zeta_i^{(R)}}$ approximates a standard normal distribution.

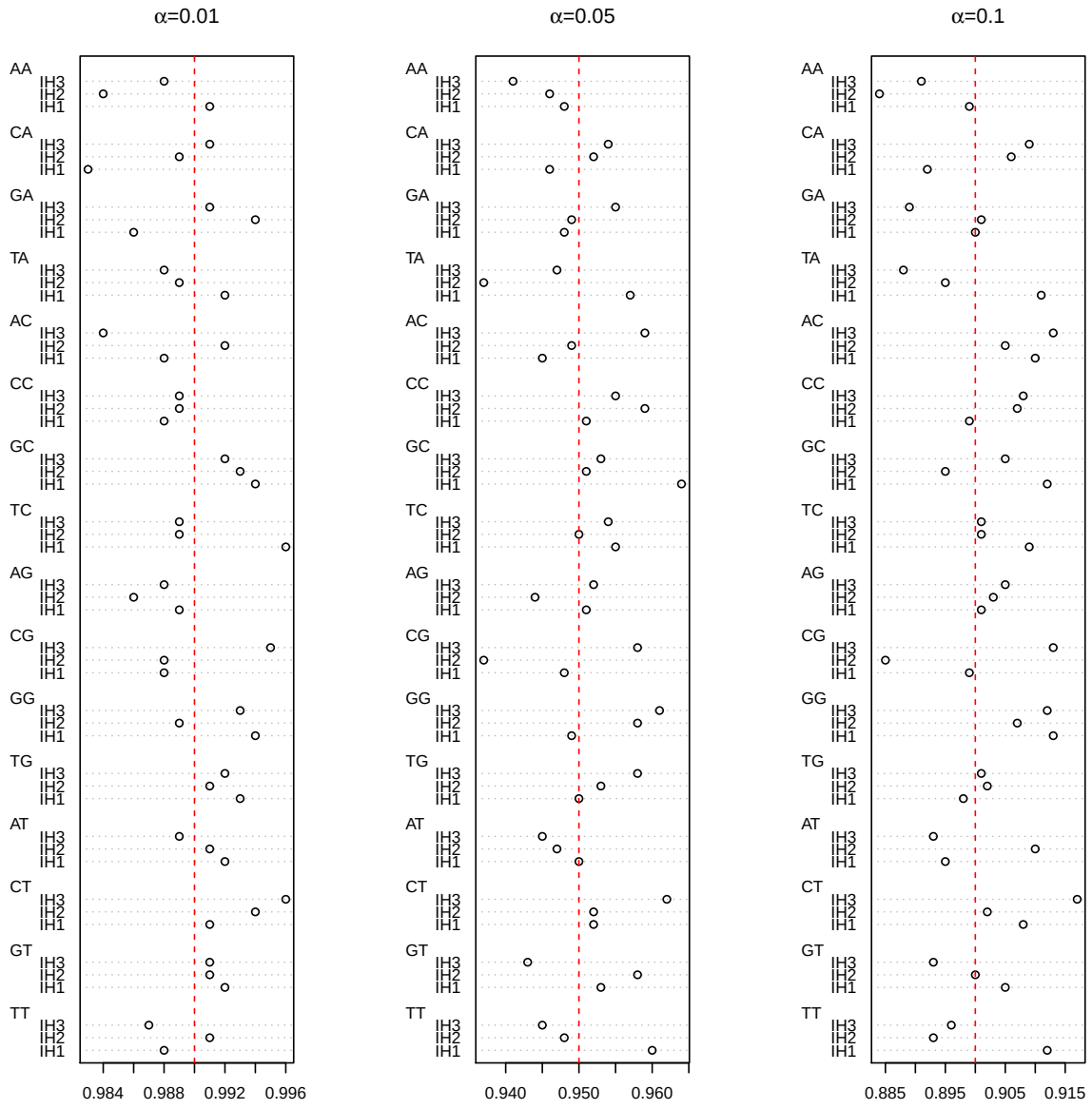

Figure S4: The fraction of times that the theoretical confident intervals of p_{ij} cover the true transition probability p_{ij} on inhomogeneous data sets IH1, IH2, and IH3.

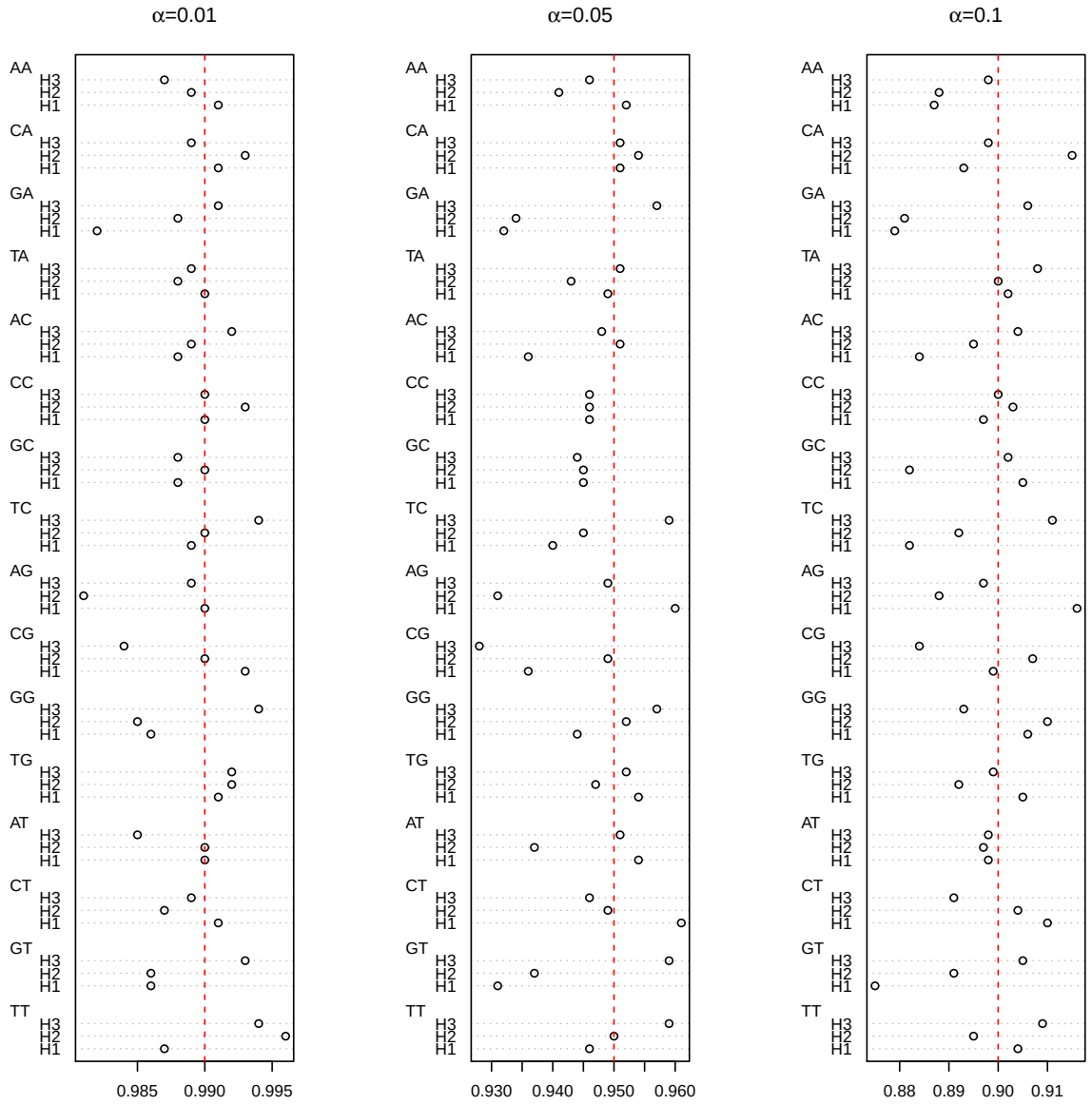

Figure S5: The fraction of times that the theoretical confident intervals of p_{ij} based on estimated $\hat{d}$ cover the true transition probability p_{ij} on homogeneous data sets H1, H2, and H3.

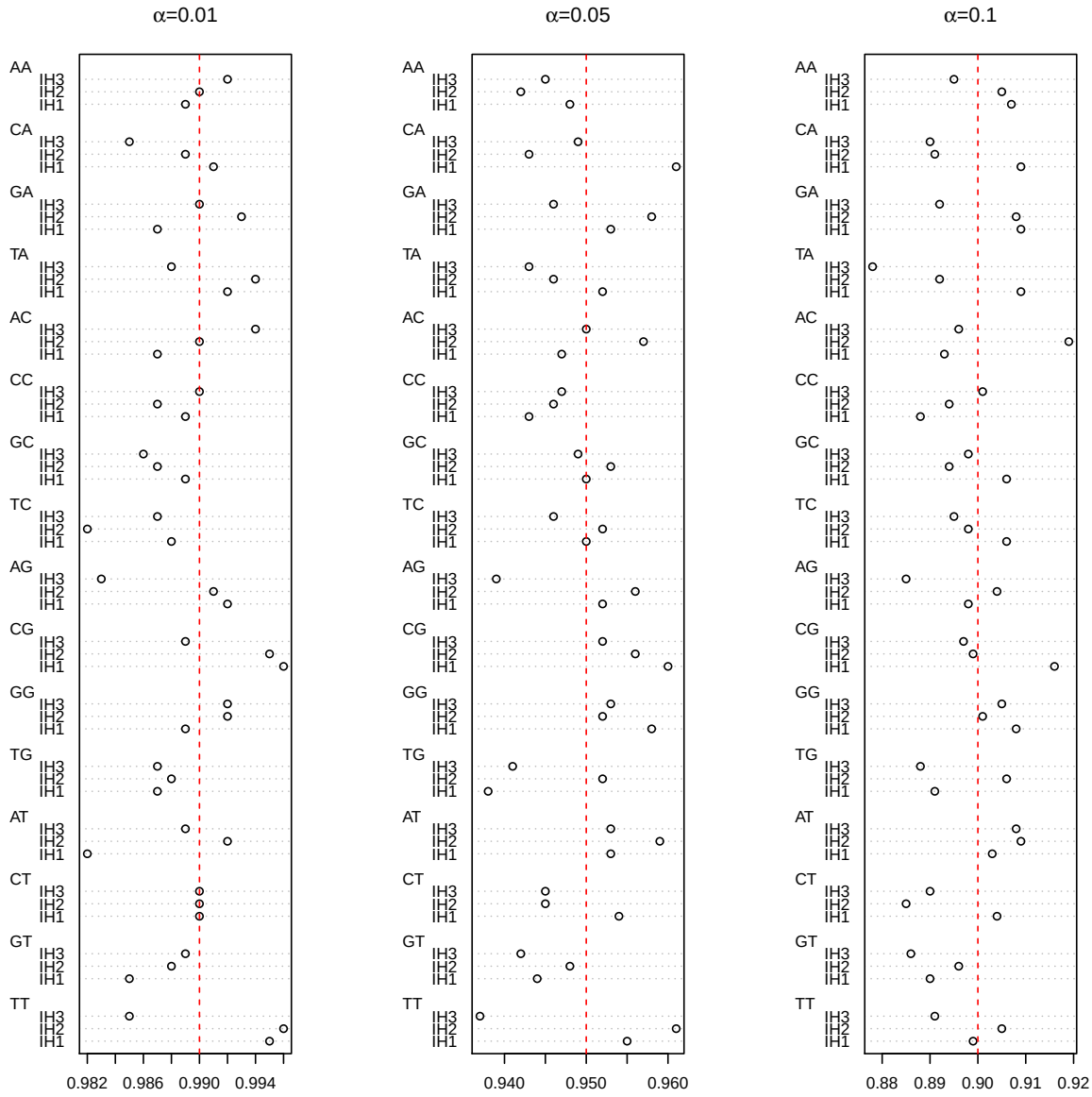

Figure S6: The fraction of times that the theoretical confident intervals of p_{ij} based on estimated $\hat{d}$ cover the true transition probability p_{ij} on inhomogeneous data sets IH1, IH2, and IH3.

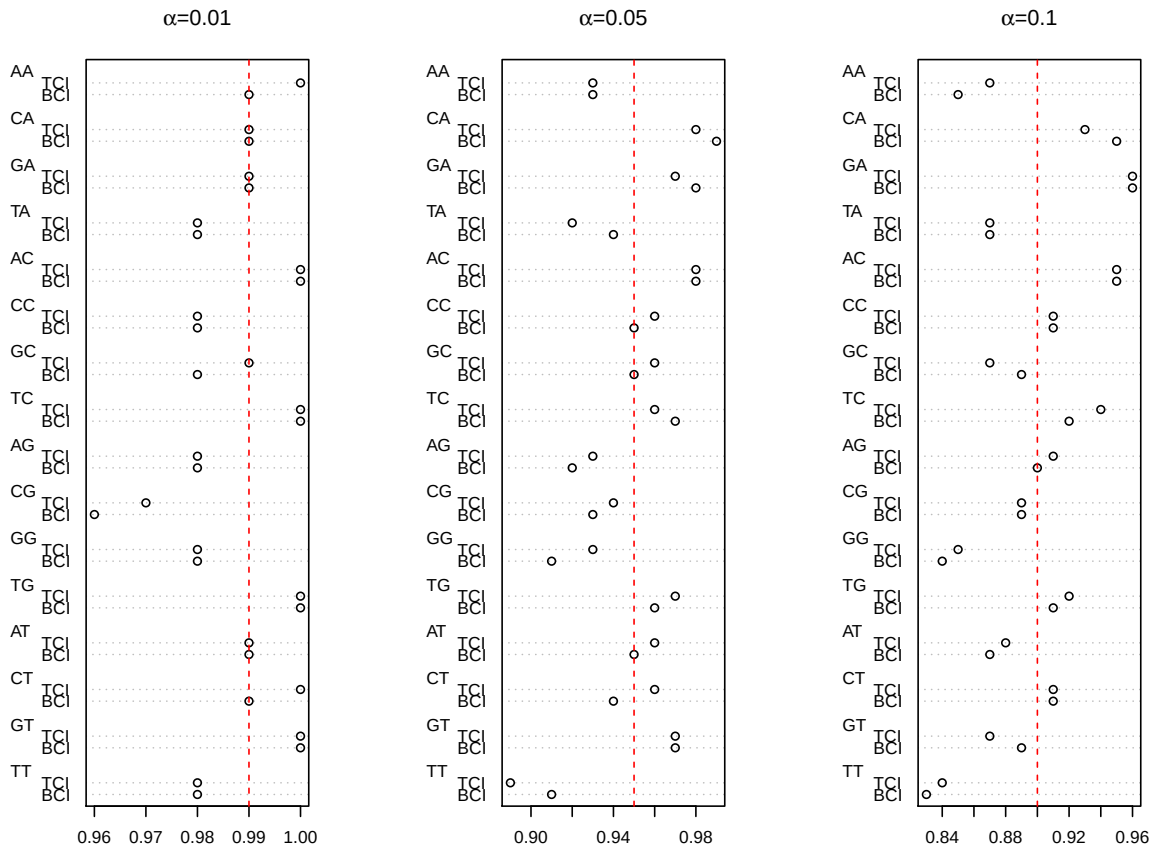

Figure S7: The fraction of times that the parametric bootstrap confidence intervals (BCI) and the theoretical confidence intervals (TCI) cover the true transition probability for H1 data set.

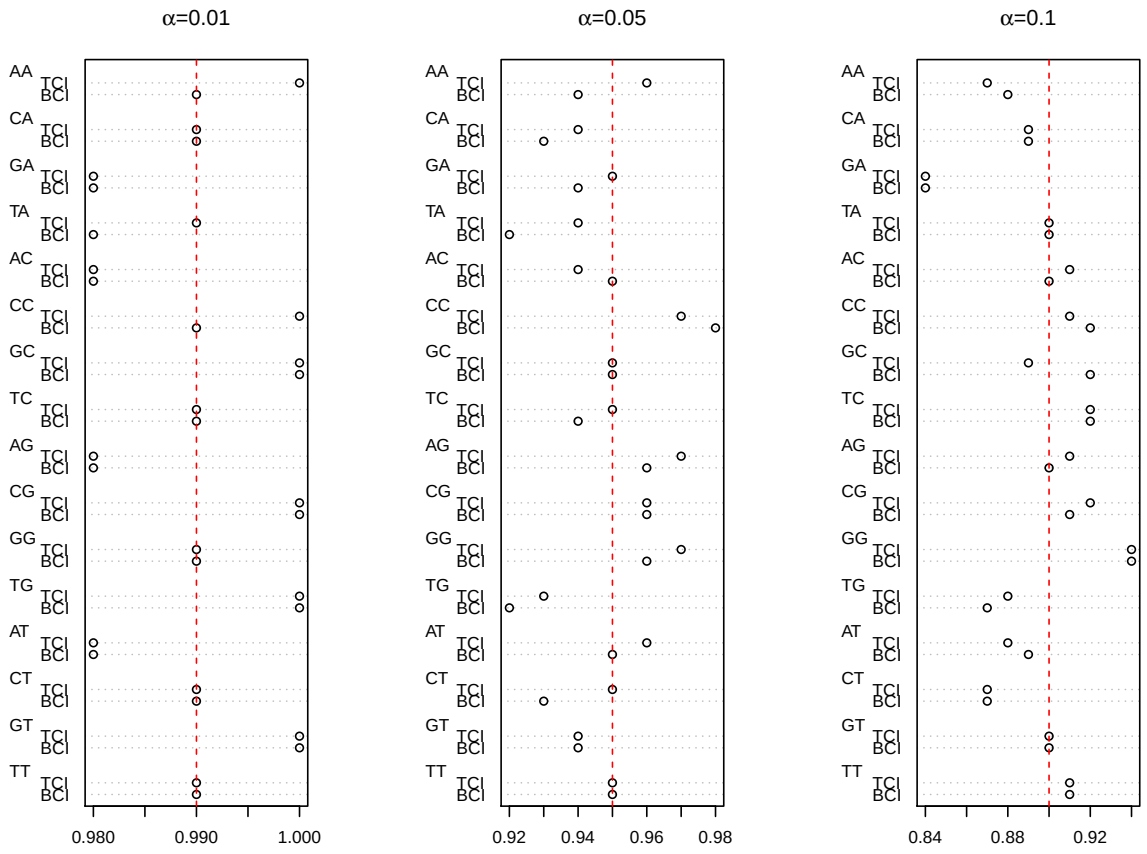

Figure S8: The fraction of times that the parametric bootstrap confidence intervals (BCI) and the theoretical confidence intervals (TCI) cover the true transition probability for H2 data set.

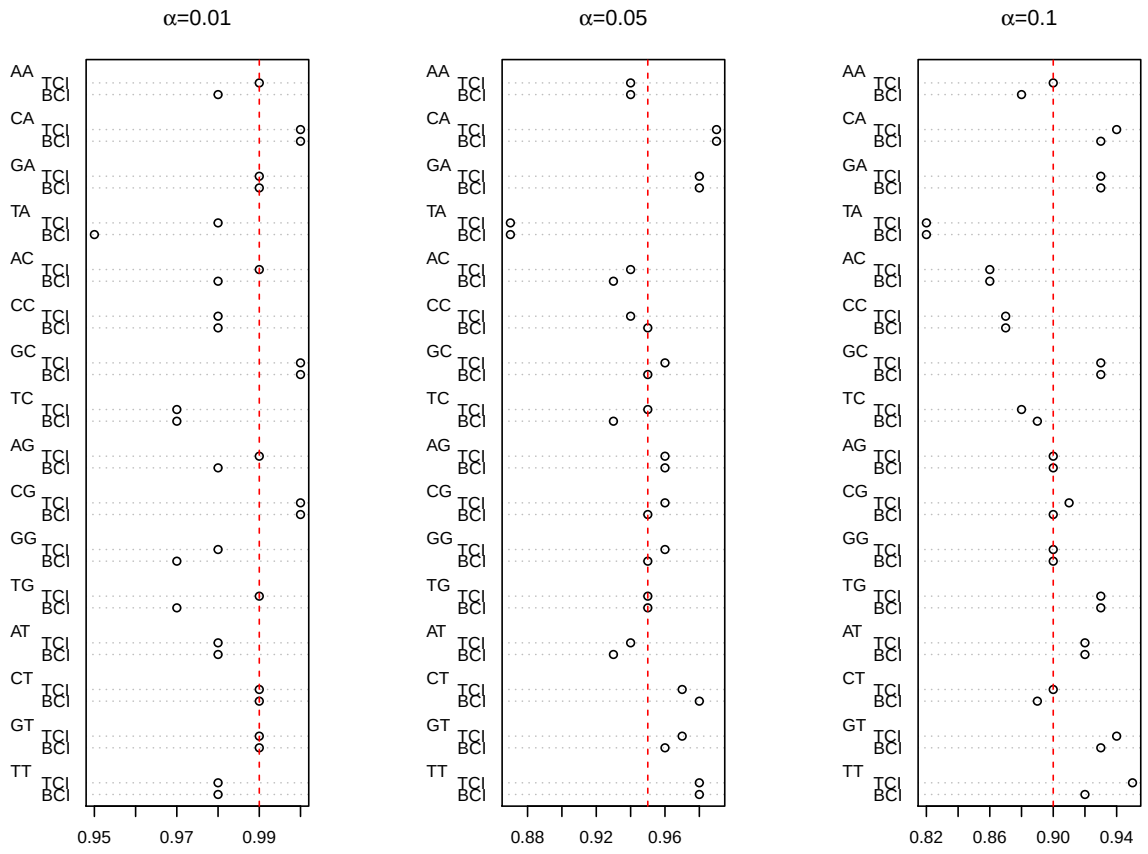

Figure S9: The fraction of times that the parametric bootstrap confidence intervals (BCI) and the theoretical confidence intervals (TCI) cover the true transition probability for H3 data set.

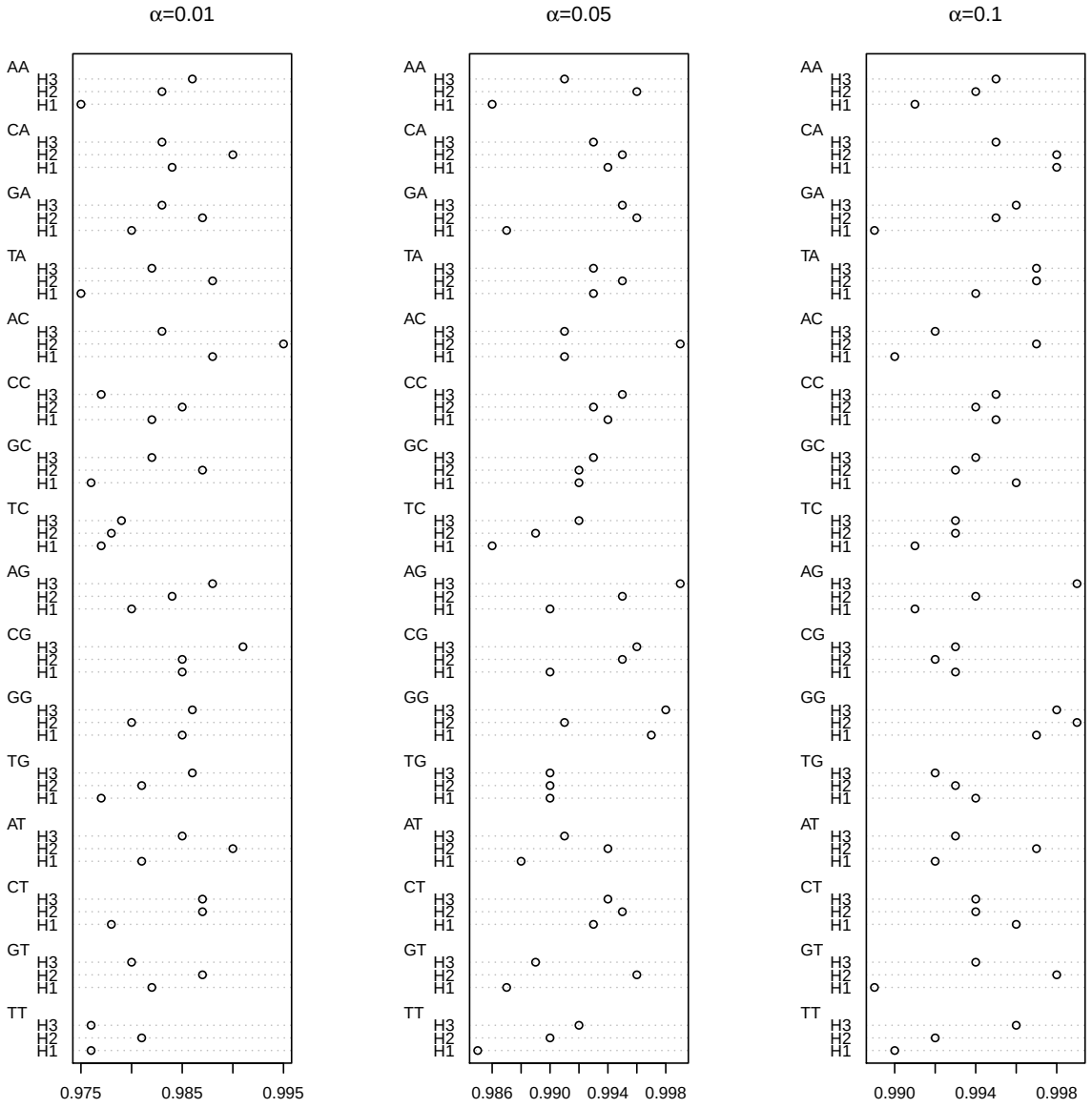

Figure S10: The ratio of the averaged length of the bootstrap CI to the averaged length of the theoretical CI for H1, H2 and H3 data sets.