Supplementary material

Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation

Filipe G. Vieira[1], Matteo Fumagalli[1], Anders Albrechtsen[2] and Rasmus Nielsen[1,2]

[1]Department of Integrative Biology, University of California, Berkeley, USA
[2]Department of Biology, University of Copenhagen, Copenhagen, Denmark

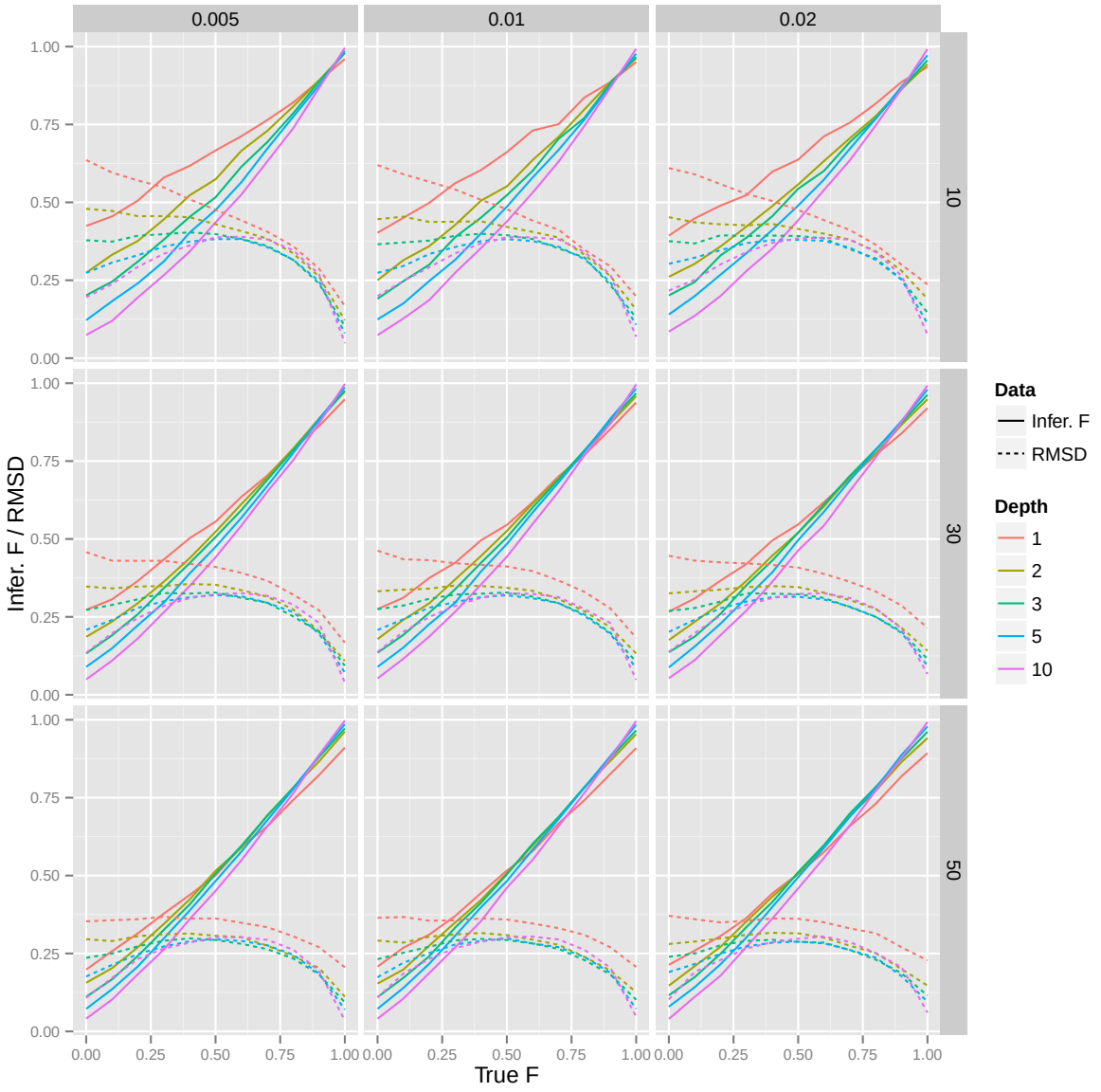
Sample No.[a]	ID	Species	Variety	Domest / Wild	Depth[b]	indF
1	IRGC 12883	*O. sativa ssp. indica*	aus	Domest.	high	0.516644
2	IRGC 45975	*O. sativa ssp. indica*	aus	Domest.	high	0.515720
3	IRGC 6307	*O. sativa ssp. indica*	aus	Domest.	high	0.517129
4	IRGC 8555	*O. sativa ssp. indica*	aus	Domest.	high	0.597927
5	IRGC 26872	*O. sativa ssp. japonica*	-	Domest.	high	0.535618
6	IRGC 27762	*O. sativa ssp. indica*	indica	Domest.	high	0.552481
7	IRGC 30416	*O. sativa ssp. indica*	indica	Domest.	high	0.486090
8	IRGC 43545	*O. sativa ssp. indica*	indica	Domest.	high	0.353690
9	IRGC 51250	*O. sativa ssp. indica*	indica	Domest.	high	0.531002
10	IRGC 51300	*O. sativa ssp. indica*	indica	Domest.	high	0.522817
11	IRGC 8231	*O. sativa ssp. indica*	indica	Domest.	high	0.066450
12	IRGC 9148	*O. sativa ssp. indica*	indica	Domest.	high	0.566577
13	IRGC 9177	*O. sativa ssp. indica*	indica	Domest.	high	0.496951
14	IRGC 1107	*O. sativa ssp. japonica*	temperate	Domest.	high	0.652006
15	IRGC 2540	*O. sativa ssp. japonica*	temperate	Domest.	high	0.766936
16	IRGC 27630	*O. sativa ssp. japonica*	temperate	Domest.	high	0.575603
17	IRGC 32399	*O. sativa ssp. japonica*	temperate	Domest.	high	0.640661
18	IRGC 418	*O. sativa ssp. japonica*	temperate	Domest.	high	0.688791
19	IRGC 55471	*O. sativa ssp. japonica*	temperate	Domest.	high	0.752038
20	IRGC 8191	*O. sativa ssp. japonica*	temperate	Domest.	high	0.724514
21	NP	*O. sativa ssp. japonica*	temperate	Domest.	high	0.702152
22	IRGC 11010	*O. sativa ssp. japonica*	tropical	Domest.	high	0.263928
23	IRGC 17757	*O. sativa ssp. japonica*	tropical	Domest.	high	0.610758
24	IRGC 25901	*O. sativa ssp. indica*	-	Domest.	high	0.529917
25	IRGC 328	*O. sativa ssp. japonica*	tropical	Domest.	high	0.502214
26	IRGC 38698	*O. sativa ssp. japonica*	tropical	Domest.	high	0.667117
27	IRGC 43325	*O. sativa ssp. japonica*	tropical	Domest.	high	0.454332
28	IRGC 43397	*O. sativa ssp. japonica*	tropical	Domest.	high	0.000000
29	IRGC 43675	*O. sativa ssp. japonica*	tropical	Domest.	high	0.505707
30	IRGC 50448	*O. sativa ssp. japonica*	tropical	Domest.	high	0.607035
31	IRGC 66756	*O. sativa ssp. japonica*	tropical	Domest.	high	0.636275
32	IRGC 8244	*O. sativa ssp. japonica*	tropical	Domest.	high	0.532811
33	IRGC 12793	*O. sativa ssp. japonica*	aromatic	Domest.	high	0.473532
34	IRGC 38994	*O. sativa ssp. japonica*	aromatic	Domest.	high	0.576520
35	IRGC 9060	*O. sativa ssp. japonica*	aromatic	Domest.	high	0.524052
36	IRGC 9062	*O. sativa ssp. japonica*	aromatic	Domest.	high	0.476387
37	RA4952	*O. sativa ssp. japonica*	aromatic	Domest.	high	0.485922
38	IRGC 31856	*O. sativa ssp. japonica*	Group V	Domest.	high	0.591380
39	IRGC 60542	*O. sativa ssp. japonica*	Group IV	Domest.	high	0.632096
40	IRGC 6513	*O. sativa ssp. japonica*	Group III	Domest.	high	0.642811
41	IRGC 105327	*O. nivara*	-	Wild	high	0.000000
42	IRGC 106105	*O. nivara*	-	Wild	high	0.578131
43	IRGC 106154	*O. nivara*	-	Wild	high	0.596246
44	IRGC 80470	*O. nivara*	-	Wild	high	0.603422

Sample No.[a]	ID	Species	Variety	Domest / Wild	Depth[b]	indF
45	IRGC 89215	*O. nivara*	-	Wild	high	0.006410
46	IRGC 105958	*O. rufipogon*	-	Wild	high	0.454003
47	IRGC 105960	*O. rufipogon*	-	Wild	high	0.025320
48	Nepal	*O. rufipogon*	-	Wild	high	0.000483
49	P46	*O. rufipogon*	-	Wild	high	0.358471
50	YJ	*O. rufipogon*	-	Wild	high	0.768965
51	IRGC 105426	*O. rufipogon*	-	Wild	low	0.347901
52	IRGC 105912	*O. rufipogon*	-	Wild	low	0.603022
53	IRGC 106161	*O. rufipogon*	-	Wild	low	0.372366
54	IRGC 106505	*O. rufipogon*	-	Wild	low	0.586989
55	IRGC 80506	*O. rufipogon*	-	Wild	low	0.051273
56	IRGC 81982	*O. rufipogon*	-	Wild	low	0.672848
57	IRGC 81991	*O. rufipogon*	-	Wild	low	0.648506
58	Dongxiang	*O. rufipogon*	-	Wild	low	0.314721
59	P25	*O. rufipogon*	-	Wild	low	0.334988
60	P61	*O. rufipogon*	-	Wild	low	0.354610
61	IRGC 103407	*O. nivara*	-	Wild	low	0.774471
62	IRGC 105705	*O. nivara*	-	Wild	low	0.682350
63	IRGC 105784	*O. nivara*	-	Wild	low	0.678742
64	IRGC 105879	*O. nivara*	-	Wild	low	0.540300
65	IRGC 106345	*O. nivara*	-	Wild	low	0.825818

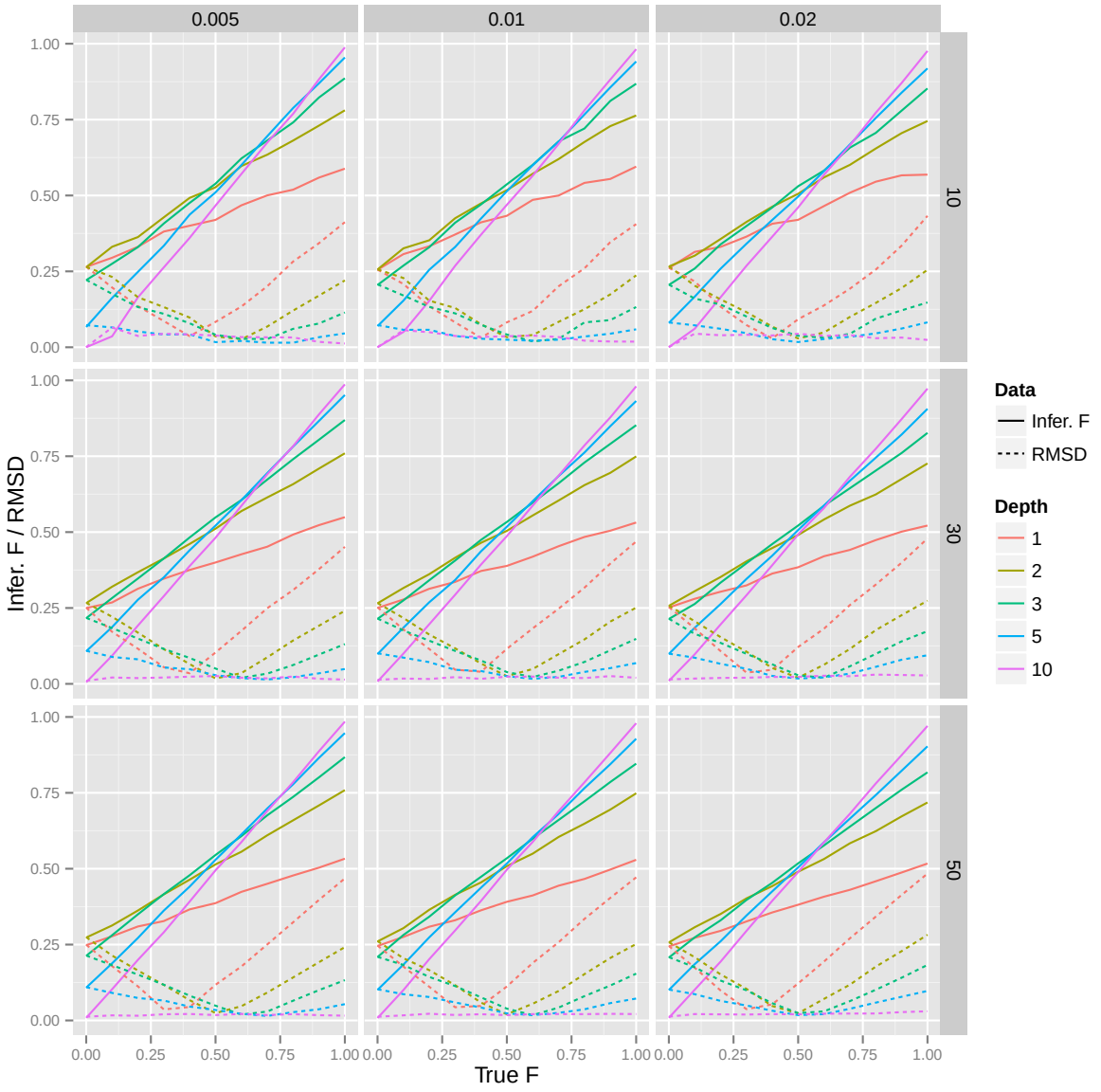
Suplementary Table 1: Accessions used on this study. [a]Sample order is the same as in Supplementary Table 1 of Xu *et al.* (2011).[b]Depth stands for $\sim 2\times$ (low) and $\sim 10\times$ (high) effective sequencing coverage. For more information on the species please check Xu *et al.* (2011).

(A)	5′000	10′000	(B)	5′000	10′000
10	20	27	10	3	5
30	87	101	30	10	19
50	152	167	50	19	37

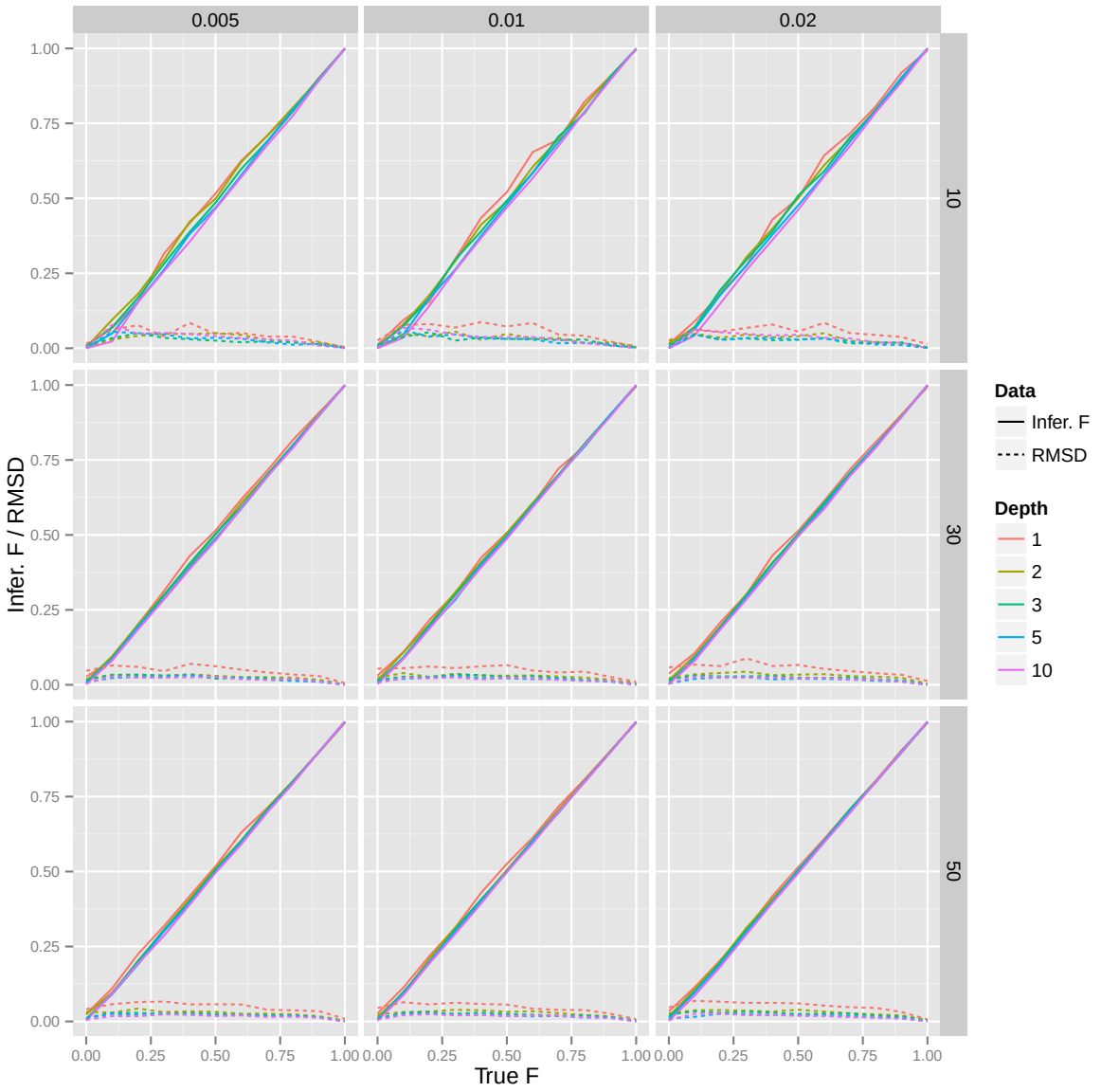
Suplementary Table 2: Running times (in seconds) for the per-individual (A) and per-site (B) algorithms, for different sample sizes (rows) and number of sites (columns). In both cases programs were run on an Intel® Xeon® X5570 @ 2.93GHz using 10 threads.



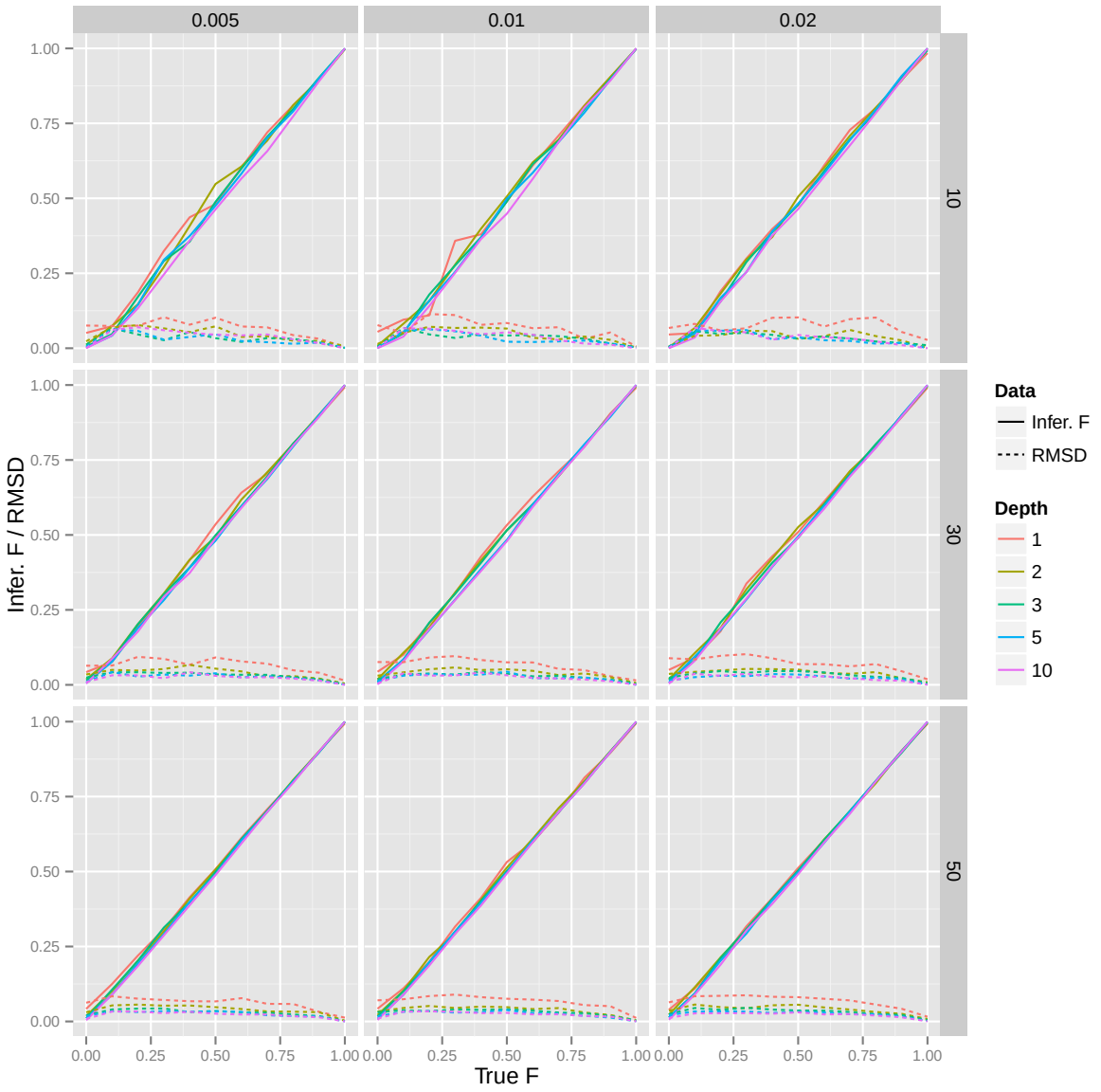
Suplementary Figure 1: Estimation of inbreeding coefficients per site. Panels depict the performance of our EM algorithm for inferring F_{site} over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Filled lines represent the inferred value for each simulated scenario, while dotted lines represent its RMSD.



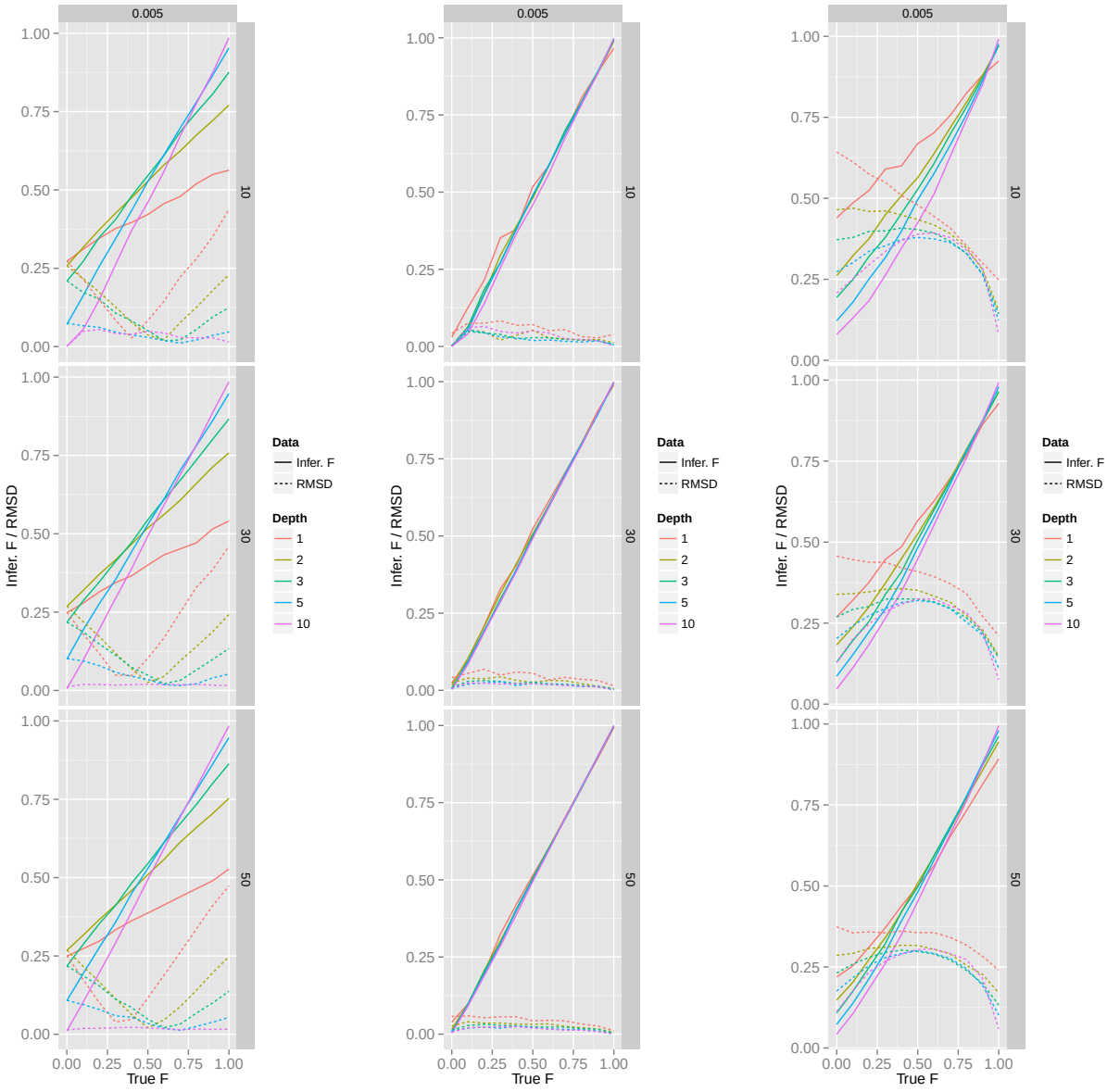
Suplementary Figure 2: Estimation of inbreeding coefficients per individual from called genotypes. Panels depict the performance of the EM algorithm for inferring F_{ind} from called genotypes over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Filled lines represent the inferred value for each simulated scenario, while dotted lines represent its RMSD.



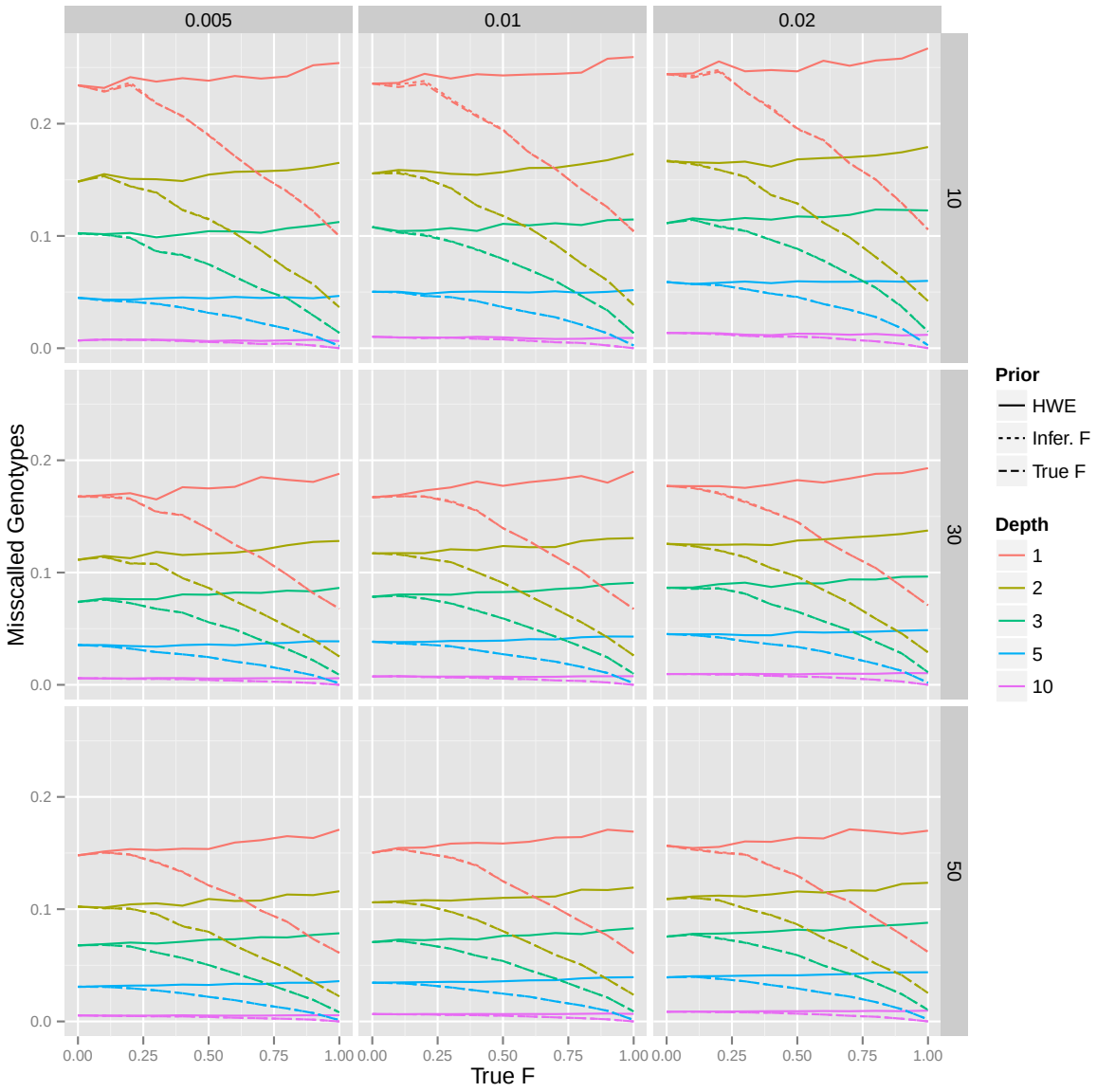
Suplementary Figure 3: Estimation of inbreeding coefficients per individual from genotype likelihoods. Panels depict the performance of the EM algorithm for inferring F_{ind} from genotype likelihoods over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Filled lines represent the inferred value for each simulated scenario, while dotted lines represent its RMSD.



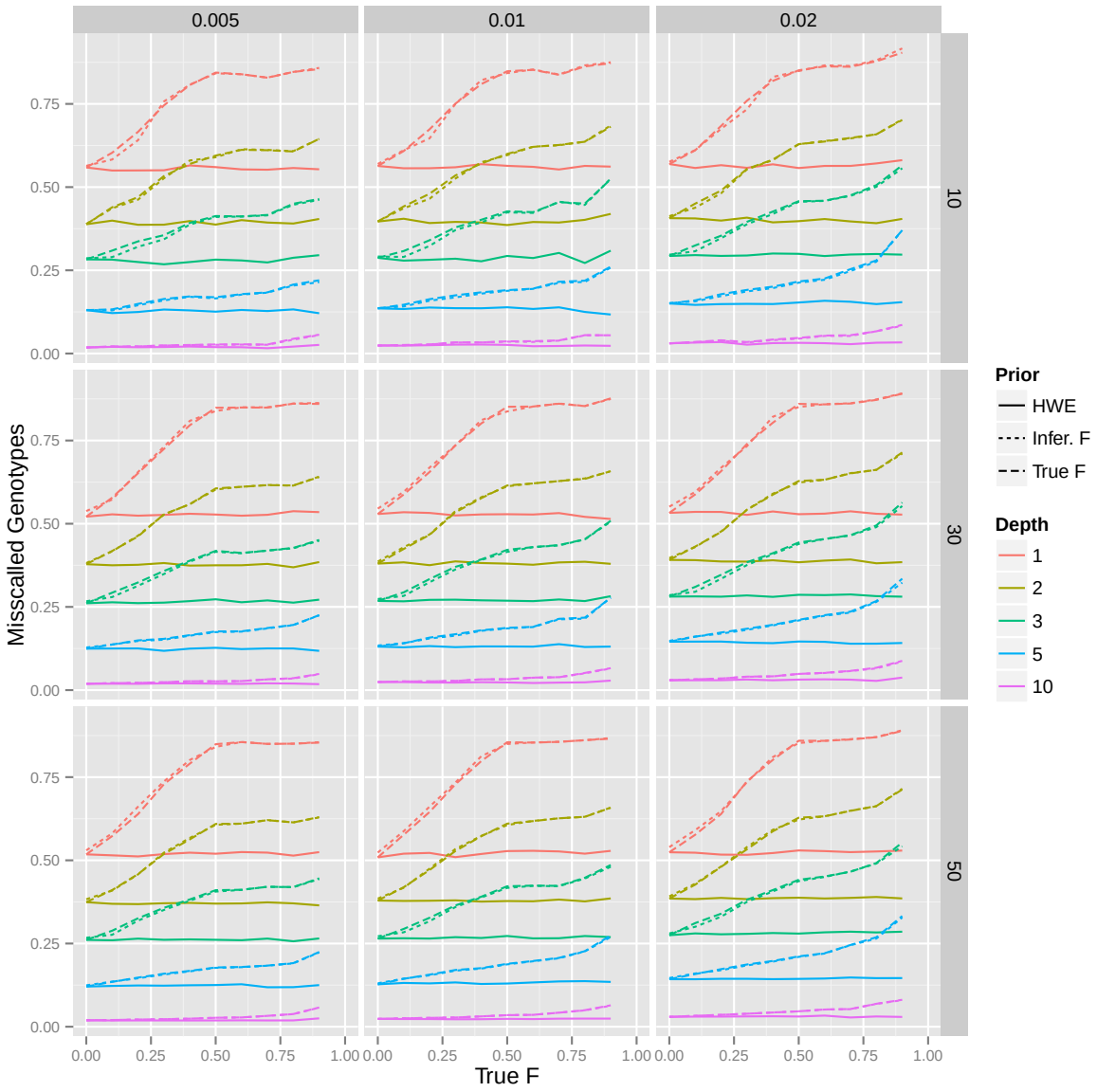
Suplementary Figure 4: Estimation of inbreeding coefficients per individual from genotype likelihoods but on 5′000 variable sites. Panels depict the performance of the EM algorithm for inferring F_{ind} from genotype likelihoods over 5′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Filled lines represent the inferred value for each simulated scenario, while dotted lines represent its RMSD.



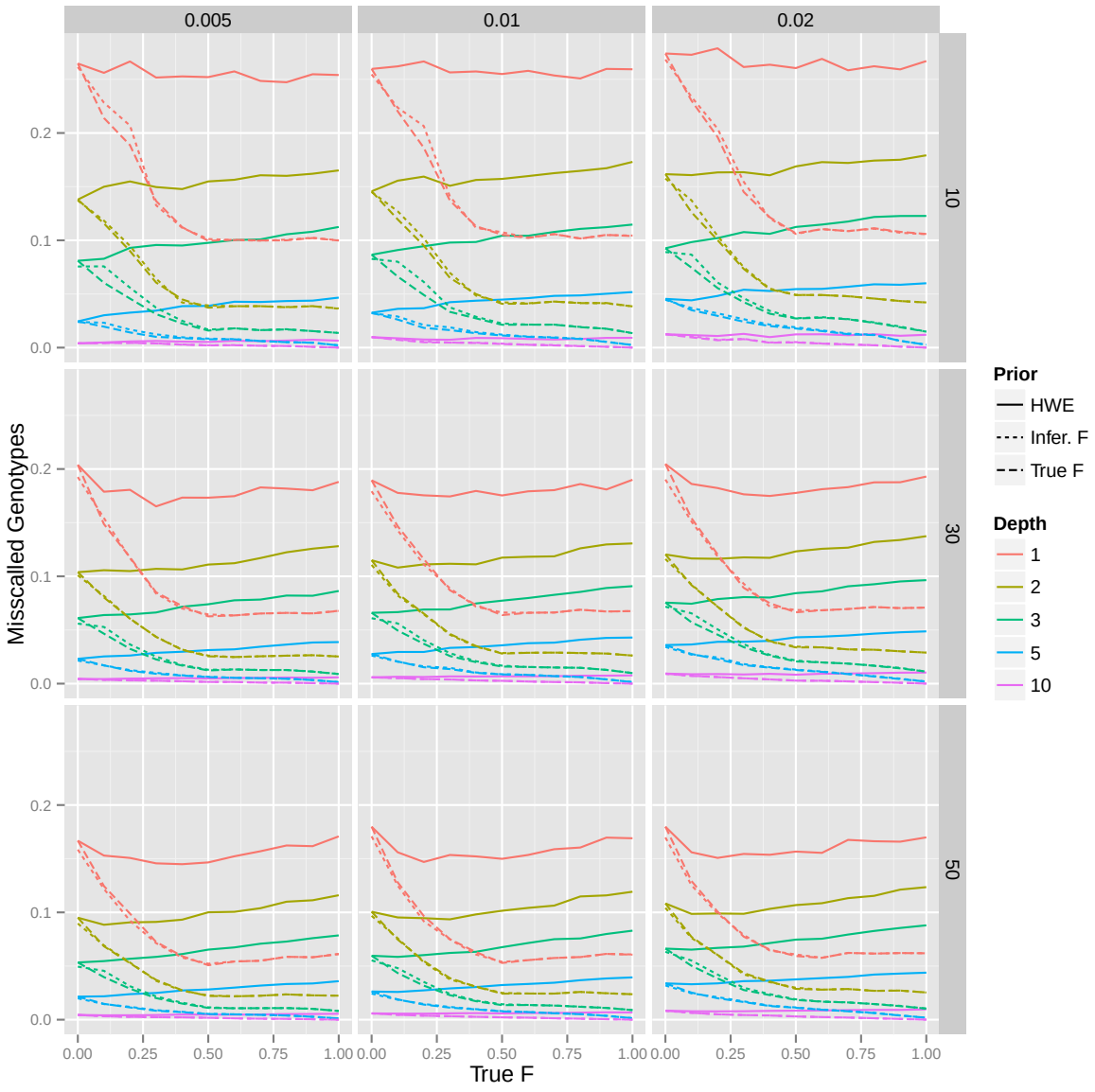
Suplementary Figure 5: Estimation of inbreeding coefficients over 1'000'000 sites. Panels depict the performance of our EM algorithms for inferring F_{ind} from called genotypes or using genotype likelihoods (left and center panels, respectively) and F_{site} (right panel) over 1′000′000 simulated sites, from which 1% are variable. For sake of clarity, only the 0.5% error rate was plotted. Rows represent different sample sizes (10, 30 and 50 individuals). Colored lines stand for different simulated sequencing coverages (see legend). Filled lines represent the inferred value for each simulated scenario, while dotted lines represent its RMSD.



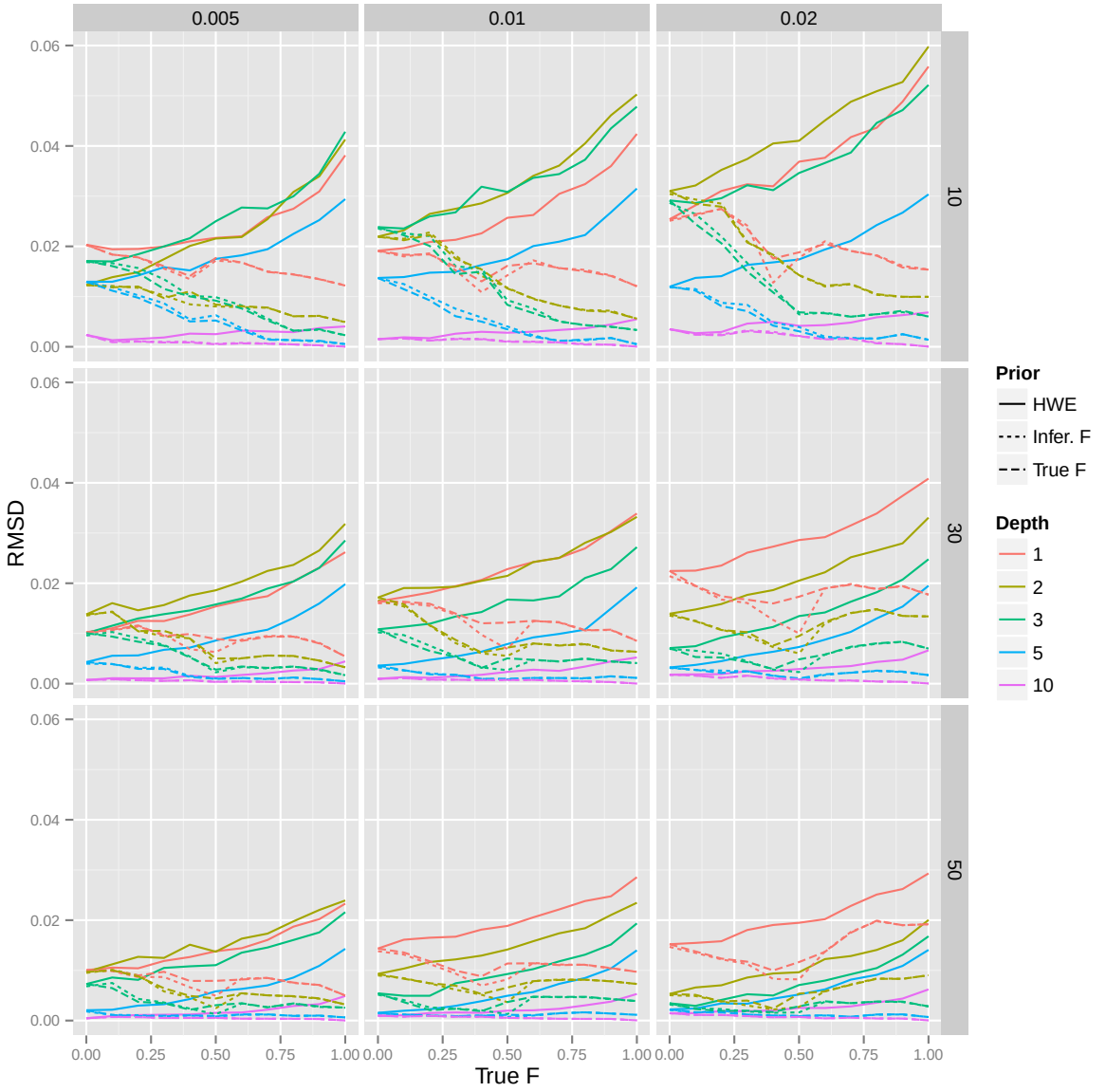
Suplementary Figure 6: Effect of the inbreeding coefficient on genotype calling. Panels depict the performance of genotype calling over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Line types represent the level of inbreeding used in the priors: $F = 0$ (HWE; filled), inferred value of F (dotted line) and true value of F used in the simulations (large dashes). Lines for the inferred and true values of F partially overlap.



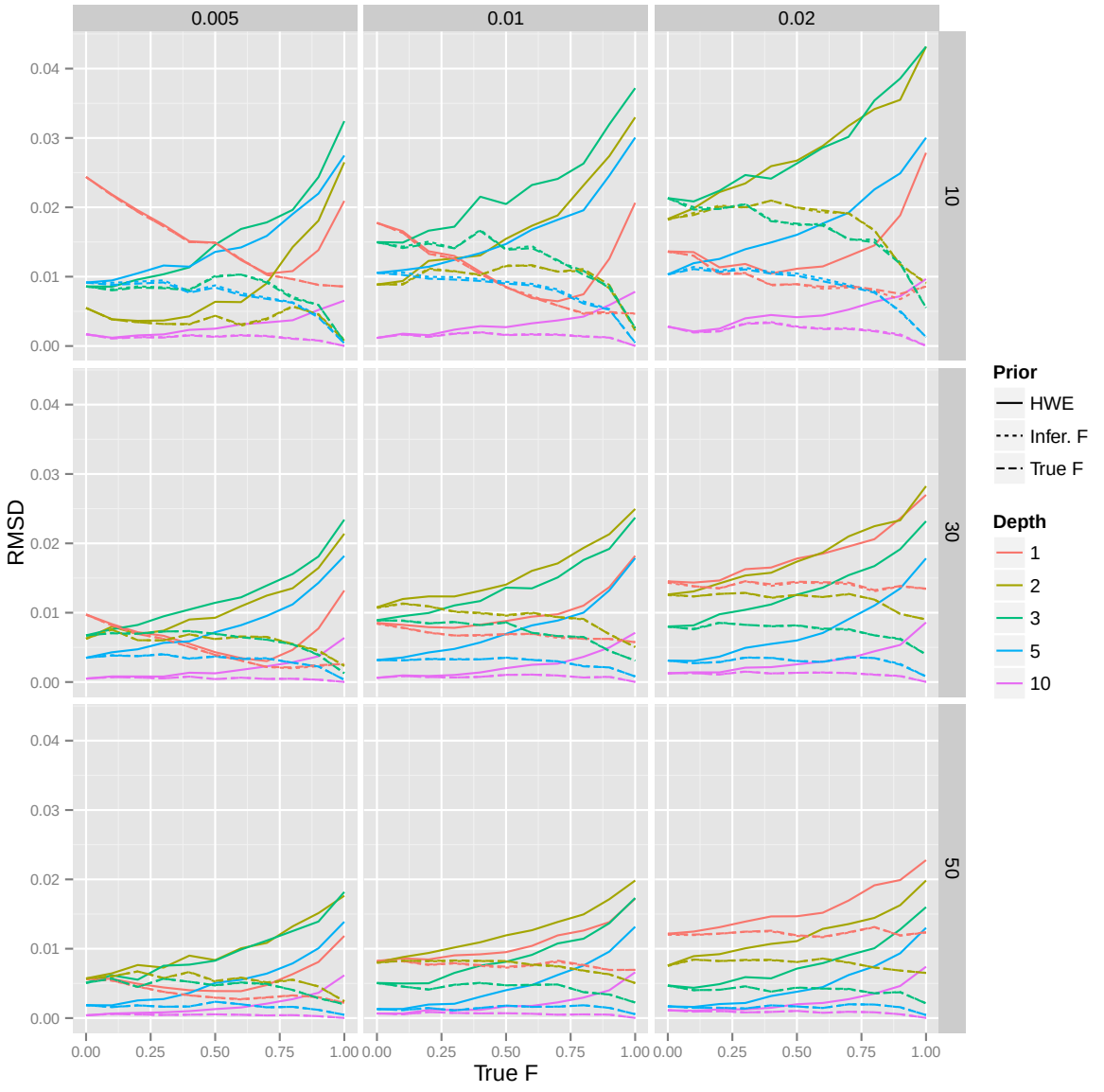
Suplementary Figure 7: Effect of the inbreeding coefficient on heterozygous genotype calling. Panels depict the performance of heterozygous genotype calling over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Line types represent the level of inbreeding used in the priors: $F = 0$ (HWE; filled), inferred value of F (dotted line) and true value of F used in the simulations (large dashes). Lines for the inferred and true values of F partially overlap.



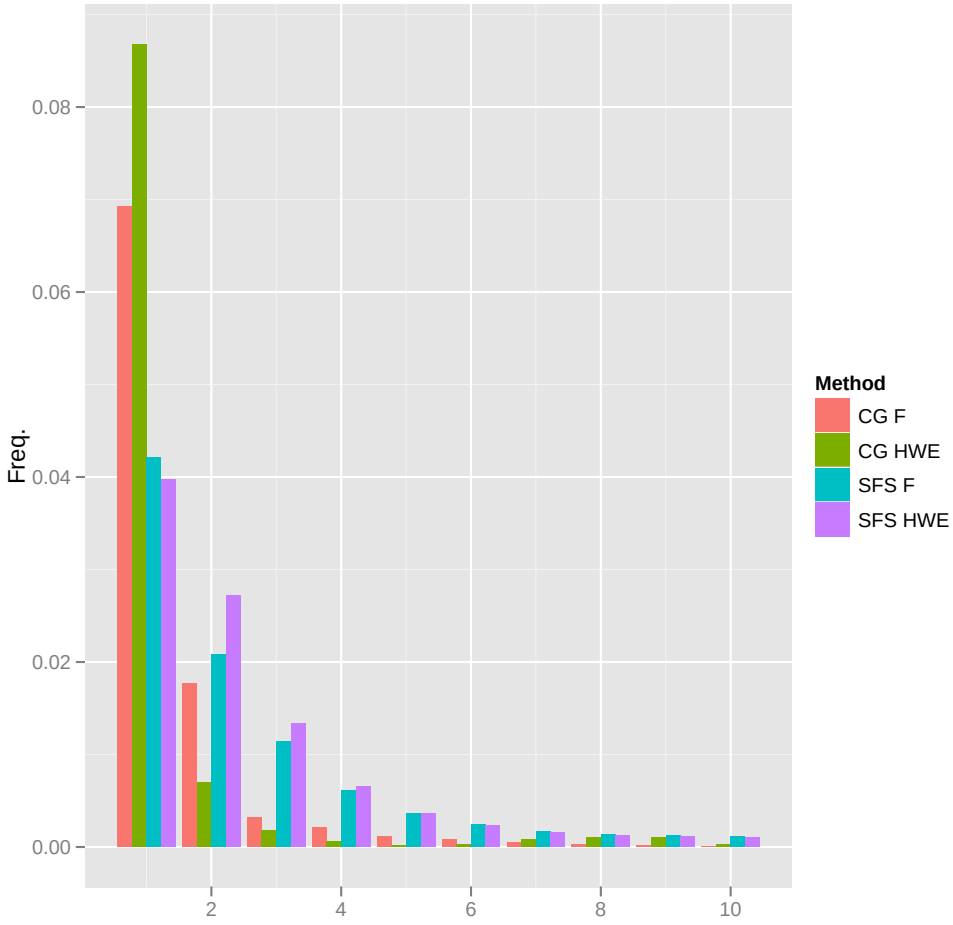
Suplementary Figure 8: Effect of the inbreeding coefficient on homozygous genotype calling. Panels depict the performance of homozygous genotype calling over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Line types represent the level of inbreeding used in the priors: $F = 0$ (HWE; filled), inferred value of F (dotted line) and true value of F used in the simulations (large dashes). Lines for the inferred and true values of F partially overlap.



Suplementary Figure 9: Effect of the inbreeding coefficient on SFS estimation from called genotypes. Panels depict the RMSD associated with SFS estimation from called genotypes over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Line types depict the level of inbreeding used in the priors: $F = 0$ (HWE; filled), inferred value of F (dotted line) and true value of F used in the simulations (large dashes). Lines for the inferred and true values of F partially overlap.



Suplementary Figure 10: Effect of the inbreeding coefficient on SFS estimation using Nielsen *et al.* (2012) method for SFS estimation taking inbreeding into account. Nielsen *et al.* (2012) method was used to calculate the allele frequency's posterior probabilities per site; these were summed over all sites to get an overall SFS. Panels depict the RMSD associated with SFS estimation over 10′000 simulated variable sites. Rows represent different sample sizes (10, 30 and 50 individuals) and columns different error rates (0.5%, 1% and 2%). Colored lines stand for different simulated sequencing coverages (see legend). Line types represent the level of inbreeding assumed in the priors: $F = 0$ (HWE; filled line), inferred value of F (large dashes) and true value of F used in the simulations (dotted line). Lines for the inferred and true values of F partially overlap.



Suplementary Figure 11: SFS from the low coverage *O. rufipogon* accessions using Nielsen *et al.* (2012) (SFS) and called genotypes (CG) methods, under two different priors of HWE and estimated inbreeding coefficient (F). For clarity sake, the 0-category was not included.

References

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PLoS One*, **7**(7), e37558.

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., Dong, Y., Gutenkunst, R. N., Fang, L., Huang, L., Li, J., He, W., Zhang, G., Zheng, X., Zhang, F., Li, Y., Yu, C., Kristiansen, K., Zhang, X., Wang, J., Wright, M., McCouch, S., Nielsen, R., Wang, J., and Wang, W. (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol*, **30**(1), 105–11.