File S1

Accuracy of genotype calling

We tested the accuracy of different methods for genotype calling on simulated data. Specifically, our goal was to quantify the overall genotyping error and the False Positive and False Negative rates in SNP calling, using different strategies to assign individual genotypes.

First, we called genotypes solely based on directly tabulating the occurrence of alternative bases among reads. Specifically, an individual was considered heterozygous if the minor allele was observed at least once among all reads for the individual (we label this procedure $GC1$). In a second scenario, to be heterozygous required that the minor allele was observed at least twice among all reads ($GC2$). These methods represent strategies for data analysis similar to the ones used on SNP genotype data and Sanger sequencing data where the genotypes for each individual are assumed to be be unambiguously determined.

Current NGS studies perform genotype calling on genotype likelihoods. We therefore computed genotype likelihoods for each individual at each site as described in Equation 22, and called the genotype with the highest likelihood (we label this procedure $GC3$). Bayesian methods assign individual genotypes from genotype likelihoods and a specific prior. We calculated genotype posterior probabilities as in Equation 9. The prior is calculated from the estimated per-site population allele frequencies (Kim et al., 2011). We assigned the genotypes with the highest posterior probability. We label this procedure GC for consistency with the main text.

We simulated sequencing data at different sequencing coverage as previously described (see Material and Methods). In particular, we simulated a total of 7M sites. In order to rule out the effect of different imputation strategies in case of missing data, we retained only sites where we had data for all individuals. Even if this is not a common practice, it allows us to directly compare different genotype calling procedures. Missing data in case of genotype calling from posterior probabilities is handled by the use of a prior estimated from the whole data (see Materials and Methods). For these reasons, the actual genotyping accuracy in case of genotype calling from counts of reads and genotype likelihoods will be lower that the values herein presented.

Results show that the lowest genotyping error is achieved when calling genotypes from genotype posterior probabilities at almost all simulated scenarios (Table S1). At low sequencing coverage, the lowest False Positive rate in SNP calling is obtained with $GC2$ although the rate steadily increases when more reads data is available (Table S2). GC provides the lowest False Negative rate in SNP calling at low coverage (Table S3). In general, calling genotypes from posterior probabilities provides the optimal balance between False Positive and False Negative rates in SNP calling. We should also notice that these results are conservative towards accuracy of GC because missing data, which are removed from these analyses, are likely to bias other genotype calling procedures at a larger extent.

Other methods to estimate F_{ST} without calling genotypes

We tested two additional methods to quantify population genetic differentiation without calling genotypes. One possible strategy for estimating F_{ST} is to calculate the posterior expectation of the sample allele frequencies, and then use these expectations to compute a method-of-moments estimator of F_{ST}. Recalling Materials and Methods, let $\pi^{(k)}_{(i,s)} = P(\hat{p}_{(i,s)} = k/(2n_i)|Y_{(i,s)})$ be the posterior probability that a site in population i has derived sample allele frequency $\hat{p}_{(i,s)} = k/(2n_i)$, in a sample of n_i diploid individuals, given the read data $Y_{(i,s)}$. Then the expected sample allele frequency, and its square value, conditional on the read data, at site s for population i is given by:

$$E[\hat{p}_{(i,s)}|Y_{(i,s)}] = \sum_{k=0}^{2n_i}(\frac{k}{2n_i})\pi^{(k)}_{(i,s)} \tag{1}$$

and

$$E[\hat{p}^2_{(i,s)}|Y_{(i,s)}] = \sum_{k=0}^{2n_i}(\frac{k}{2n_i})^2\pi^{(k)}_{(i,s)}. \tag{2}$$

Similarly, the expected square difference in the sample allele frequency between two distinct populations i and j is given by:

$$E[(\hat{p}_{(i,s)}-\hat{p}_{(j,s)})^2|Y_s] = E[\hat{p}^2_{(i,s)}+\hat{p}^2_{(j,s)}-2\hat{p}_{(i,s)}\hat{p}_{(j,s)}|Y_s] = E[\hat{p}^2_{(i,s)}|Y_{(i,s)}]+E[\hat{p}^2_{(j,s)}|Y_{(j,s)}]-2E[\hat{p}_{(i,s)}\times\hat{p}_{(j,s)}|Y_s] \tag{3}$$

where

$$E[\hat{p}_{(i,s)} \times \hat{p}_{(j,s)}|Y_s] = \sum_{k=0}^{2n_i}\sum_{z=0}^{2n_j}(\frac{k}{2n_i})(\frac{z}{2n_j})\pi^{(k,z)}_{(i,j,s)} \tag{4}$$

and $\pi^{(k,z)}_{(i,j,s)}$ is the joint posterior probability of sample allele frequencies $P(\hat{p}_{(i,s)} = k/(2n_i), \hat{p}_{(j,s)} = z/(2n_j)|Y_s)$. We substituted these expectations in the original F_{ST} formulation (Equations 1-4). We label this estimator $F_{ST.Ef2}$.

Alternatively, we simply computed an estimate of the sample allele frequency $\hat{p}_{(i,s)}$ at site s for population i as:

$$\hat{p}_{(i,s)} = \arg\max \pi_{(i,s)} \tag{5}$$

and substituted these values in the original F_{ST} formula (Equations 1-4). We label this estimator $F_{ST.Ef1}$. Table S4 summarizes all tested methods to estimate F_{ST} from NGS data used in this study.

Results from simulated data show that $F_{ST.Ef1}$ and $F_{ST.Ef2}$ have greater accuracy than methods based on genotype calling, but less that the new method based on the expectations of genetic variance components (Figure S2).

Supporting Tables

Table S1: **Genotype calling errors**. Genotype calling errors (in %) for different scenarios of sequencing depth and different genotype calling procedures. $GC1$ and $GC2$ assign a heterozygous state if at least 1 or 2 alternate alleles are observed, respectively. $GC3$ and GC assign genotypes according to the maximum genotype likelihood or genotype posterior probability, respectively. We retained only sites with no missing data.

Sequencing depth	Number of valid sites	$GC1$	$GC2$	$GC3$	GC
2X	2,148	2.53	1.72	2.53	1.77
6X	633,751	4.45	0.63	3.27	0.47
20X	700,007	13.36	0.47	0.074	0.0076

Table S2: **SNP calling false positive rates**. SNP calling false positive rates (in %) for different scenarios of sequencing depth and different genotype calling procedures. $GC1$ and $GC2$ assign a heterozygous state if at least 1 or 2 alternate alleles are observed, respectively. $GC3$ and GC assign genotypes according to the maximum genotype likelihood or genotype posterior probability, respectively. We retained only sites with no missing data.

Sequencing depth	Number of valid monomorphic sites	$GC1$	$GC2$	$GC3$	GC
2X	2,001	43.28	0.15	43.18	35.78
6X	588,989	83.28	1.75	72.48	11.18
20X	650,838	99.70	17.55	2.93	0.22

Table S3: **SNP calling false negative rates**. SNP calling false negative rates (in %) for different scenarios of sequencing depth and different genotype calling procedures. $GC1$ and $GC2$ assign a heterozygous state if at least 1 or 2 alternate alleles are observed, respectively. $GC3$ and GC assign genotypes according to the maximum genotype likelihood or genotype posterior probability, respectively. We retained only sites with no missing data.

Sequencing depth	Number of valid polymorphic sites	$GC1$	$GC2$	$GC3$	GC
2X	147	4.76	57.14	4.76	1.36
6X	44,762	0.26	5.87	0.39	1.58
20X	49,169	0.0041	0.016	0.018	0.042

Table S4: F_{ST} **estimators**. Names, brief descriptions, referring Equations and Figures for all different F_{ST} estimators tested in this study.

Name	Description	Equation(s)	Figure(s)
$\hat{F}_{ST.GC}$	from called genotypes	9	1-2, S1
$\hat{F}_{ST.Ef2}$	from expectation of sample allele frequency	28-31	S2
$\hat{F}_{ST.Ef1}$	from sample allele frequency calculated as the maximum posterior probability	32	S2
$\hat{F}_{ST.Ev}$	from expectation of genetic variance components	10-12	1-2, S1
$\hat{F}_{ST.ML.GC}$	ML estimator from called genotypes	9	2
$\hat{F}_{ST.ML}$	ML estimator without calling genotypes	17	2

Table S5: **Computational time for F_{ST} computation**. Computation time, in seconds, to compute F_{ST} for different number of simulated sites (S) and sample size (N) for each of the 2 populations, at 2X sequencing depth. 'Genotype p.p.' includes computing genotype posterior probabilities. 'Frequency p.p.' includes estimating the SFS and computation sample allele frequency posterior probabilities and it is required to compute $\hat{F}_{ST.Ev}$. $\hat{F}_{ST.Ev}$ also includes estimating the 2D-SFS. Calculations were run on a Unix desktop machine, Intel Core 2 Duo CPU E8600 @ 3.33GHz x 2. Maximum memory usage was $< 0.1G$.

S	N	Simulation	Genotype p.p.	Frequency p.p.	$\hat{F}_{ST.GC}$	$\hat{F}_{ST.Ef1}$	$\hat{F}_{ST.Ef2}$	$\hat{F}_{ST.Ev}$
10k	20	3	2	25	< 1	< 1	< 1	2
10k	40	6	3	116	< 1	< 1	< 1	10
50k	20	14	7	167	< 1	< 1	< 1	12
50k	40	29	14	357	< 1	< 1	< 1	47

Table S6: **Computational time for PCA computation**. Computation time, in seconds, to perform PCA for different number of simulated sites (S) and sample size (N) for each of the 3 populations, at 2X sequencing depth. 'Genotype p.p.' includes computing genotype posterior probabilities. 'Frequency p.p.' includes estimating the SFS and computation sample allele frequency posterior probabilities and it is required to perform PCA as in 'w/o GC (2)'. 'GC' refers to estimate C from called genotypes. 'w/o GC (1)' and 'w/o GC (2)' estimate C without calling genotypes. 'w/o GC (2)' also weights each site by its probability of being variable. Computations refer to estimation of the reduced matrix C as in Equation 18 and do not include the eigenvector decomposition. Calculations were run on a Unix desktop machine, Intel Core 2 Duo CPU E8600 @ 3.33GHz x 2. Maximum memory usage was $< 0.1G$.

S	N	Simulation	Genotype p.p.	Frequency p.p.	GC	w/o GC (1)	w/o GC (2)
10k	20	4	1	76	< 1	< 1	< 1
10k	40	8	3	143	2	3	2
50k	20	20	6	408	2	3	3
50k	40	41	12	561	11	17	18

Supporting Figures

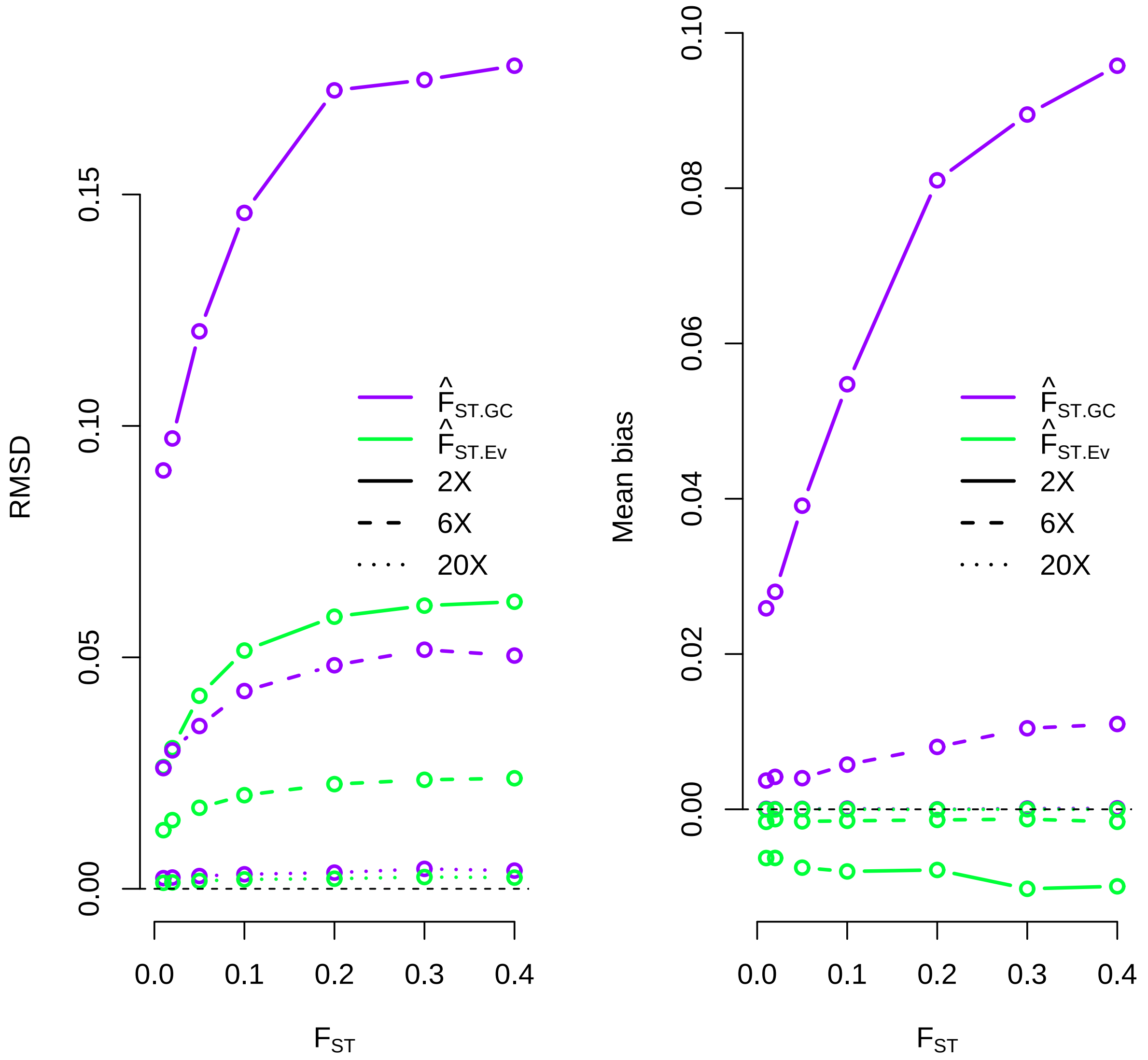

Figure S1: RMSD (left panel) and mean bias (right panel) for estimating F_{ST} under different sequencing coverage (2X, 6X and 20X). We compared the accuracy of the new method which does not rely on genotype calling ($\hat{F}_{ST.Ev}$), while also using the true 2D-SFS as a prior, and a method based on allele frequencies after calling genotypes ($\hat{F}_{ST.GC}$) (see Material and Methods). We simulated 20 individuals for each population and $10,000$ sites for each scenario.

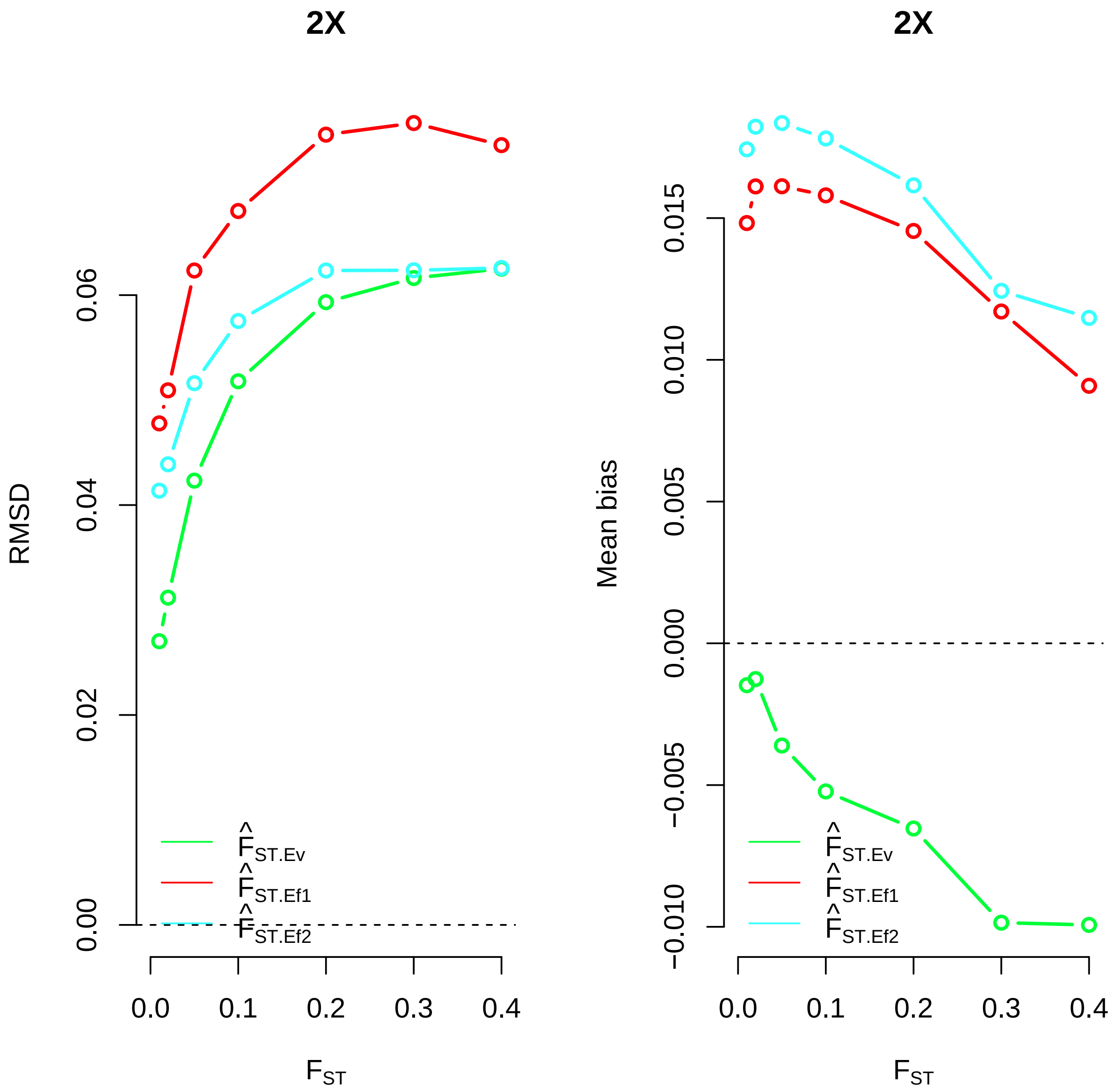

Figure S2: RMSD (left panel) and mean bias (right panel) for estimating F_{ST} at 2X sequencing coverage. We compared the accuracy of the new method which does not rely on genotype calling ($\hat{F}_{ST.Ev}$) and of two methods based on computing population allele frequency as the sample allele frequency with the highest posterior probability, $\hat{F}_{ST.Ef1}$, and as the expected allele frequency, $\hat{F}_{ST.Ef2}$ (see Material and Methods). We simulated 20 individuals for each population and $10,000$ sites for each scenario.

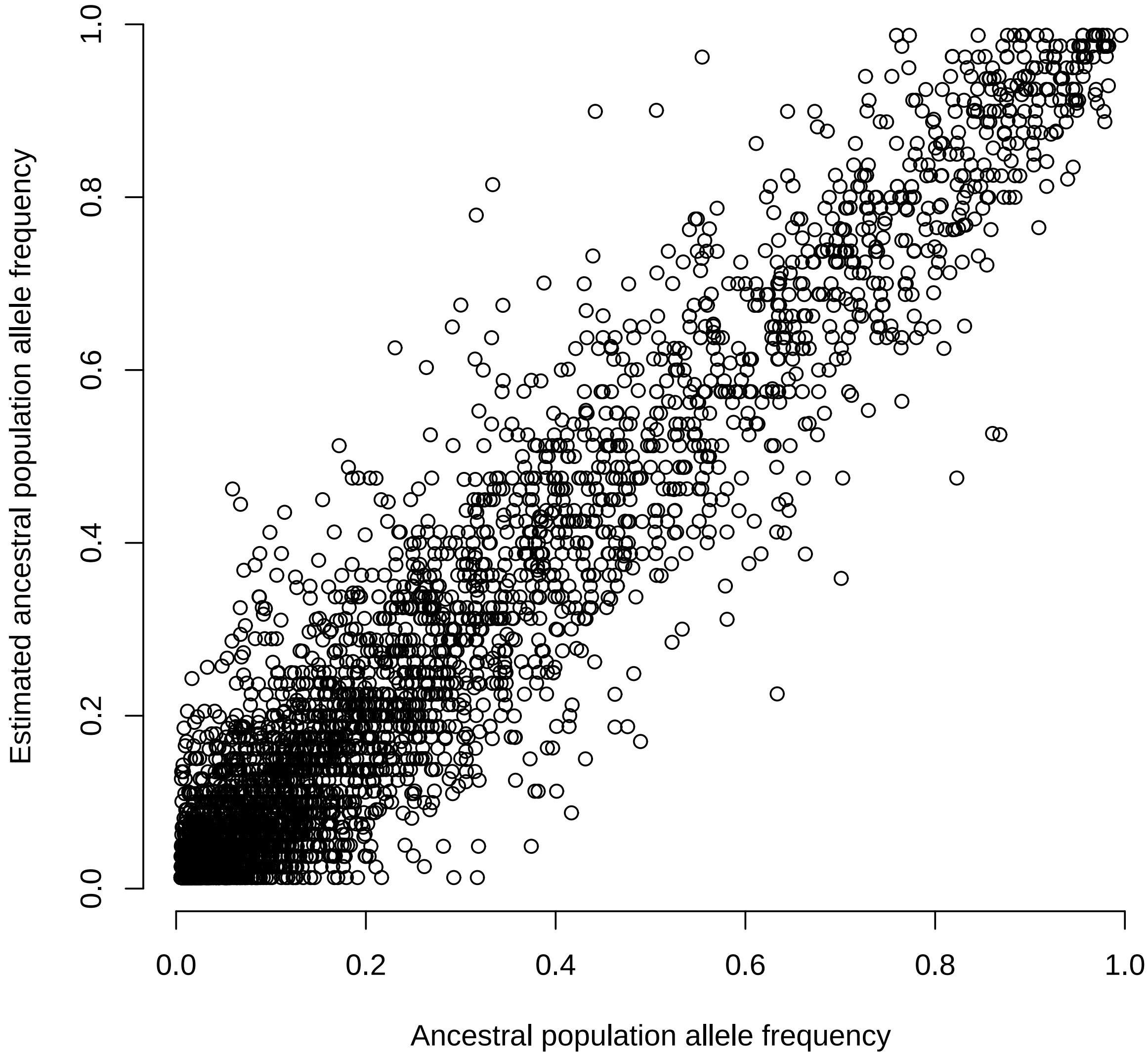

Figure S3: Ancestral population allele frequency estimated from a Maximum Likelihood procedure with unknown genotypes versus the true value used in the model. We simulated 20 individuals for each population and a total of $7,000$ sites, using data from Figure 2.

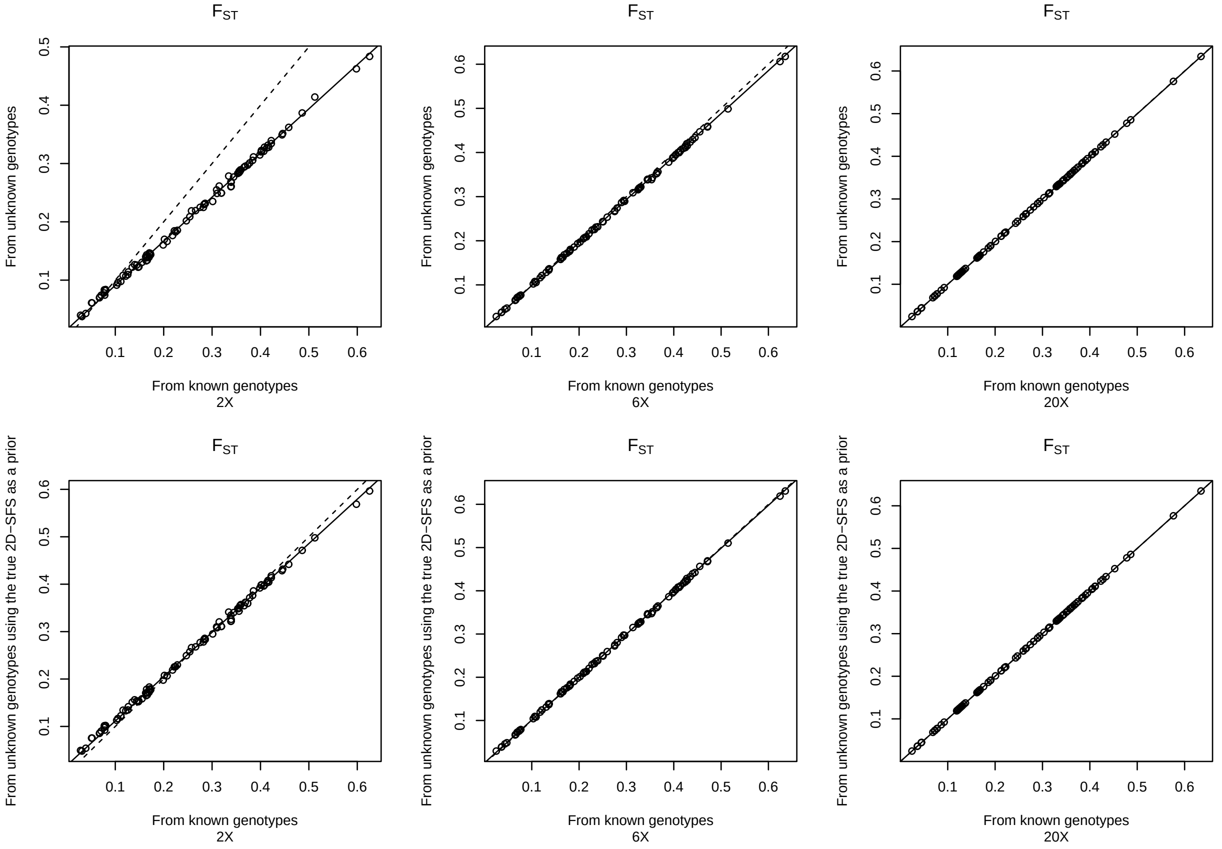

Figure S4: F_{ST} for 100 10kb regions where only 10% of the sites are variable in the population. F_{ST} is computed using the estimated global 2D-SFS (first row) or the true 2D-SFS as a prior (second row) (see Material and Methods). Dotted line represents the diagonal while the continuous line is the regressed line between true and estimated F_{ST}. We simulated a total of 1M sites at 2X, 6X and 20X sequencing coverage and 20 individuals for each population.

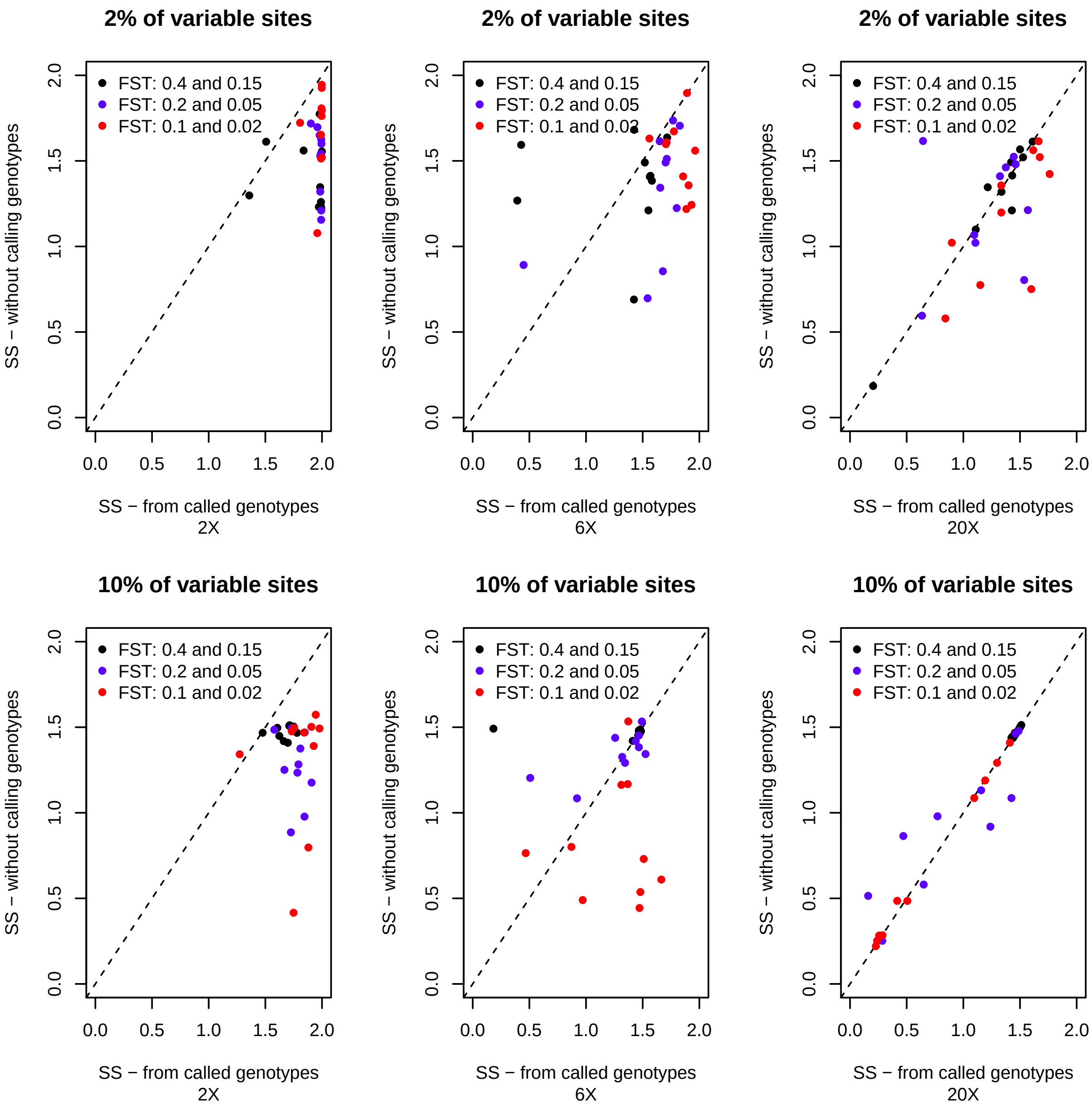

Figure S5: Sum-of-squares (SS) between PC1 and PC2 computed from called genotypes from genotype posterior probabilities (on x-axis) or with the new proposed method which does not rely on genotype calling (on y-axis). We simulated 3 populations of 20 individuals at 2X, 6X and 20X sequencing coverage. Populations are differentiated by F_{ST} of 0.4 - 0.15, 0.2 - 0.05 and 0.1 - 0.02. We simulated 10,000 sites with 2% and 10% of sites being variable in the population.

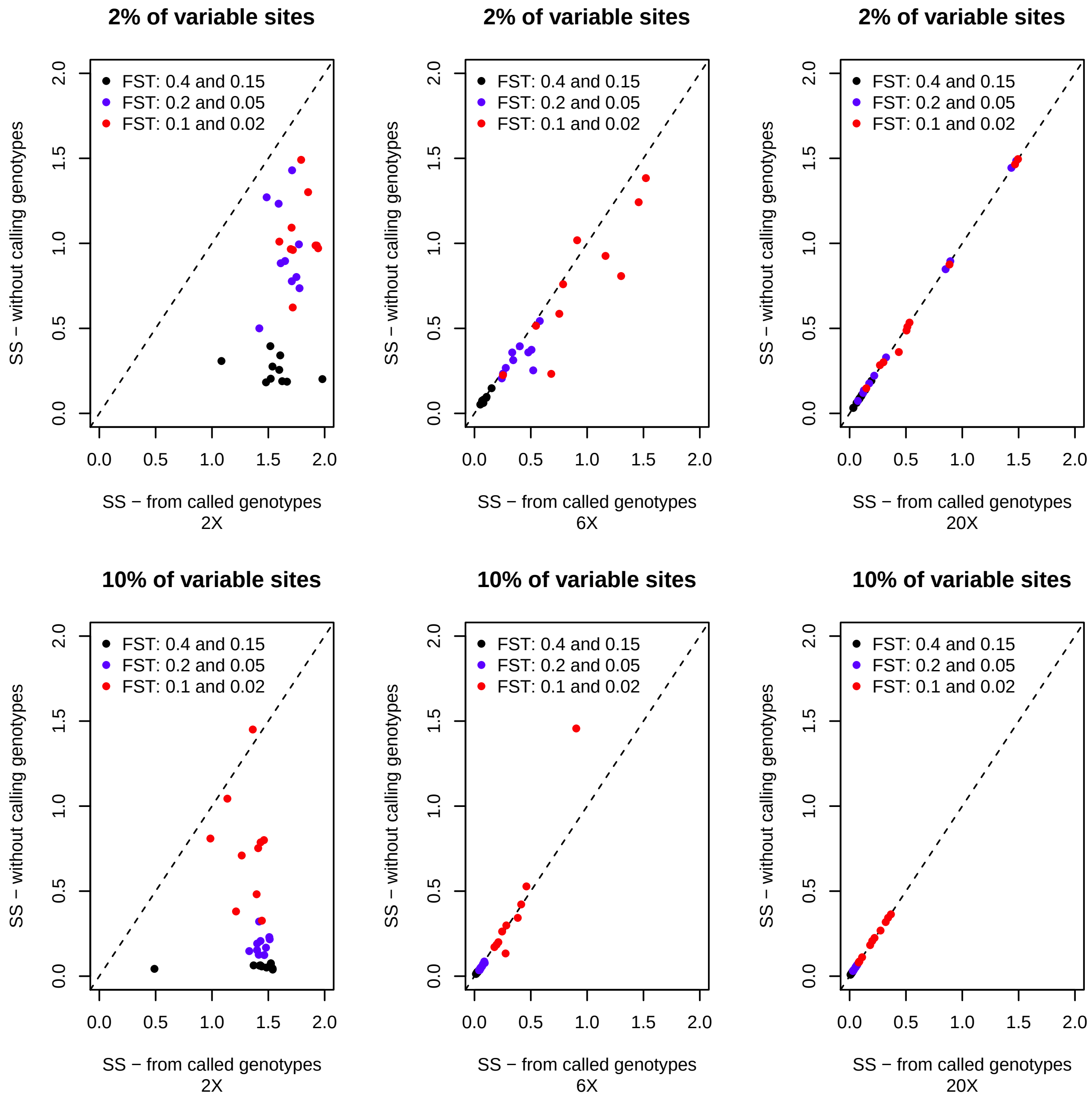

Figure S6: Sum-of-squares (SS) between PC1 and PC2 computed from called genotypes from genotype posterior probabilities (on x-axis) or with the new proposed method which does not rely on genotype calling (on y-axis). We did not normalize the standardized allele frequencies to have the same variance. We simulated 3 populations of 20 individuals at 2X, 6X and 20X sequencing coverage. Populations are differentiated by F_{ST} of 0.4 - 0.15, 0.2 - 0.05 and 0.1 - 0.02. We simulated $10,000$ sites with 2% and 10% of sites being variable in the population.

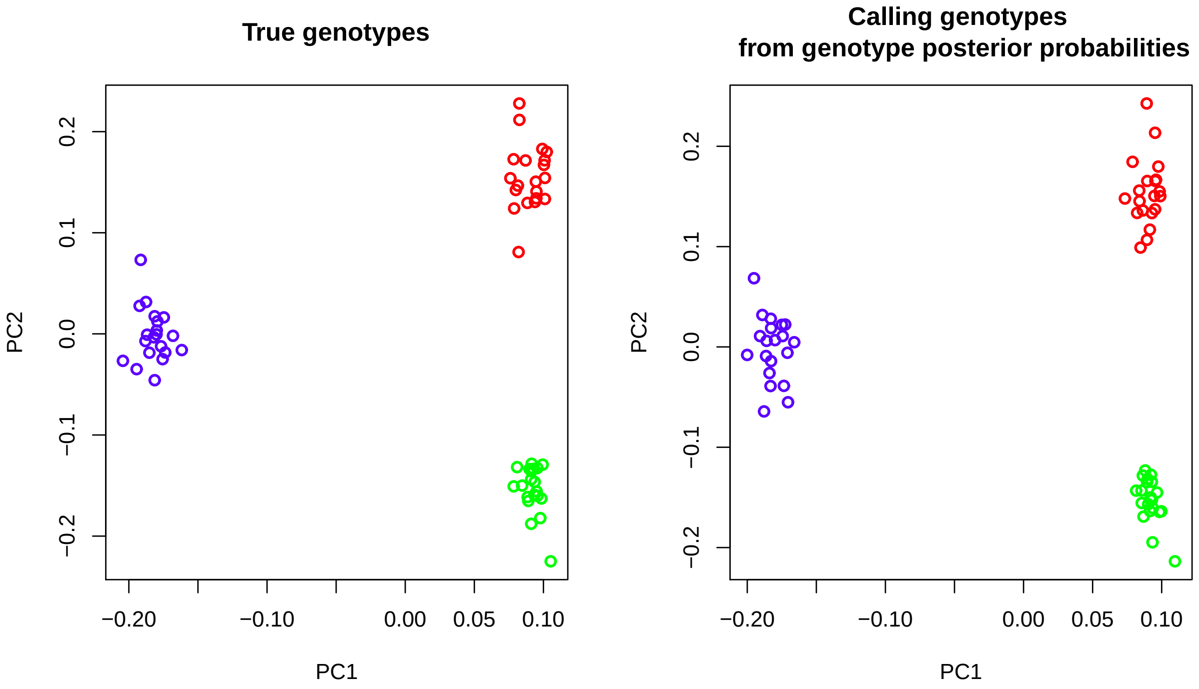

Figure S7: PCA plots from known genotypes and from called genotypes using genotype posterior probabilities. We simulated 3 populations of 20 individuals each at 20X sequencing coverage. Colors are coded according to each simulated population. Blue and green/red populations are differentiated by an F_{ST} of 0.4 while green and red populations are differentiated by an F_{ST} of 0.15. We simulated $10,000$ sites, all variable in the population.

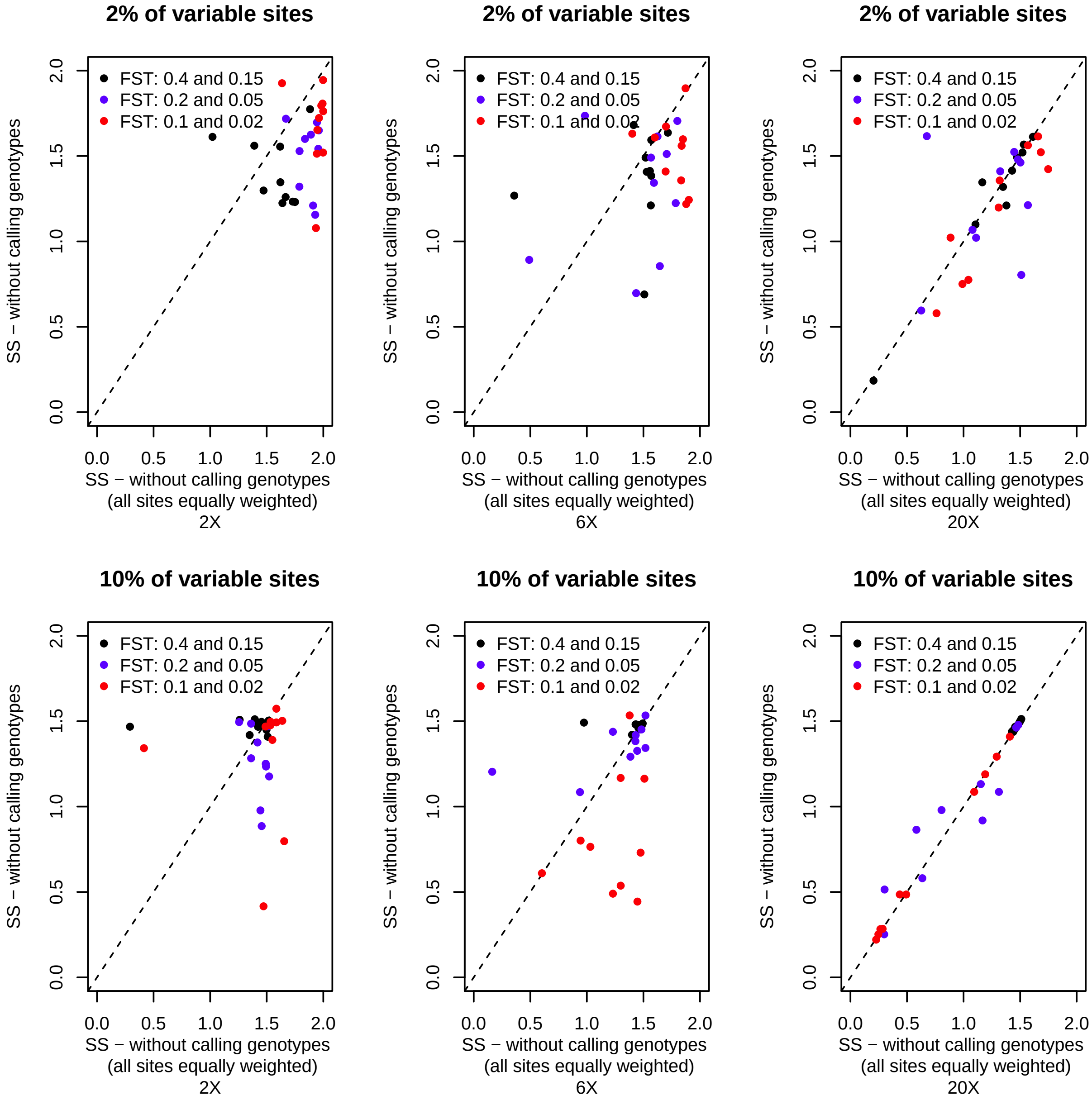

Figure S8: Sum-of-squares (SS) between PC1 and PC2 computed with the new proposed method, which does not rely on genotype calling, (on y-axis) or with the new method but without weighting each site for its probability to be variable (on x-axis). We simulated 3 populations of 20 individuals at 2X, 6X and 20X sequencing coverage. Populations are differentiated by F_{ST} of 0.4 - 0.15, 0.2 - 0.05 and 0.1 - 0.02. We simulated $10,000$ sites with 2% and 10% of sites being variable in the population.

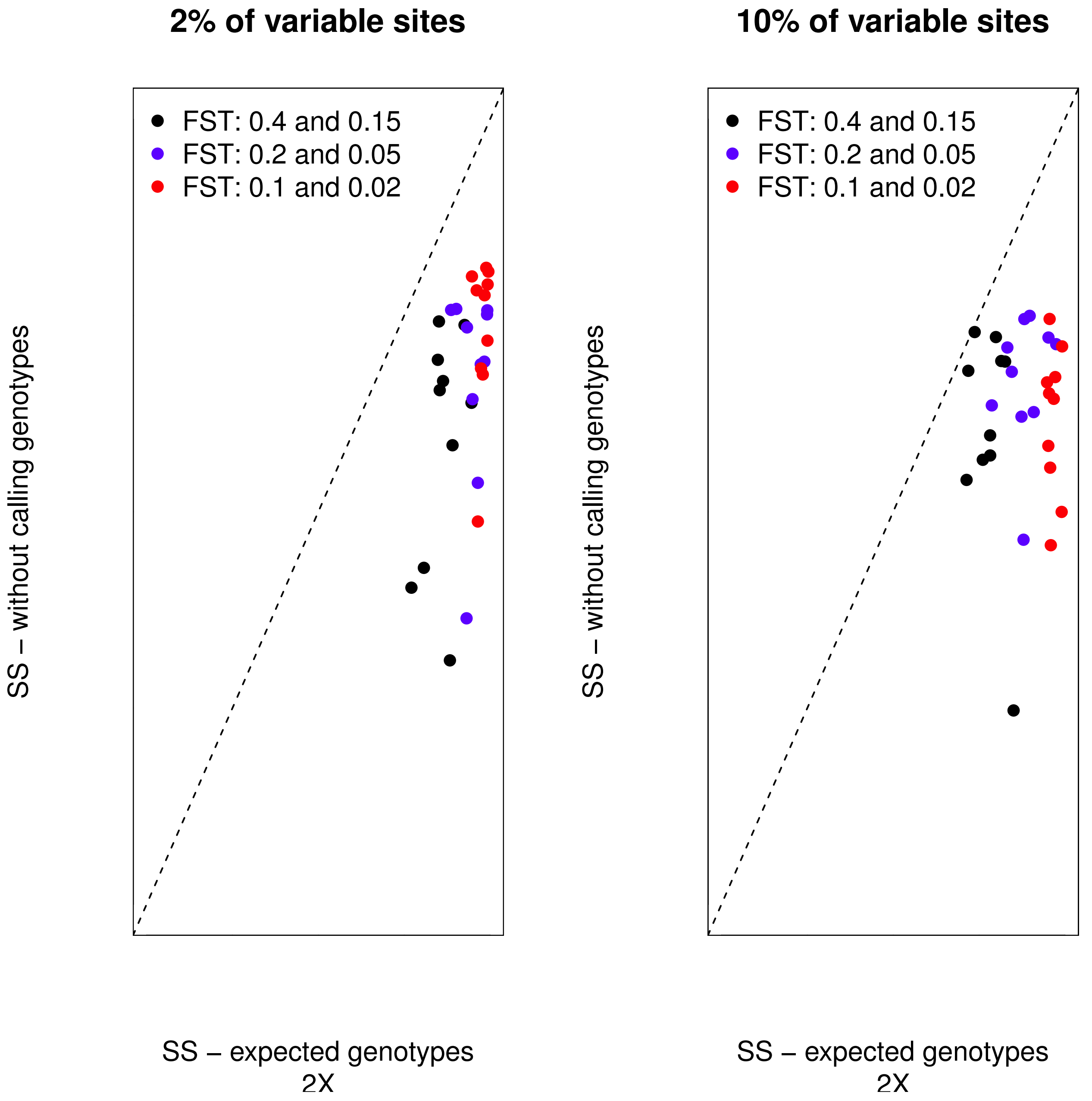

Figure S9: Sum-of-squares (SS) between PC1 and PC2 computed with the new proposed method, which does not rely on genotype calling, (on y-axis) or with a method based on computing the expectations of genotypes from genotype posterior probabilities (on x-axis). We simulated 3 populations of 20 individuals at 2X sequencing coverage. Populations are differentiated by F_{ST} of 0.4 - 0.15, 0.2 - 0.05 and 0.1 - 0.02. We simulated $10,000$ sites with 2% and 10% of sites being variable in the population.

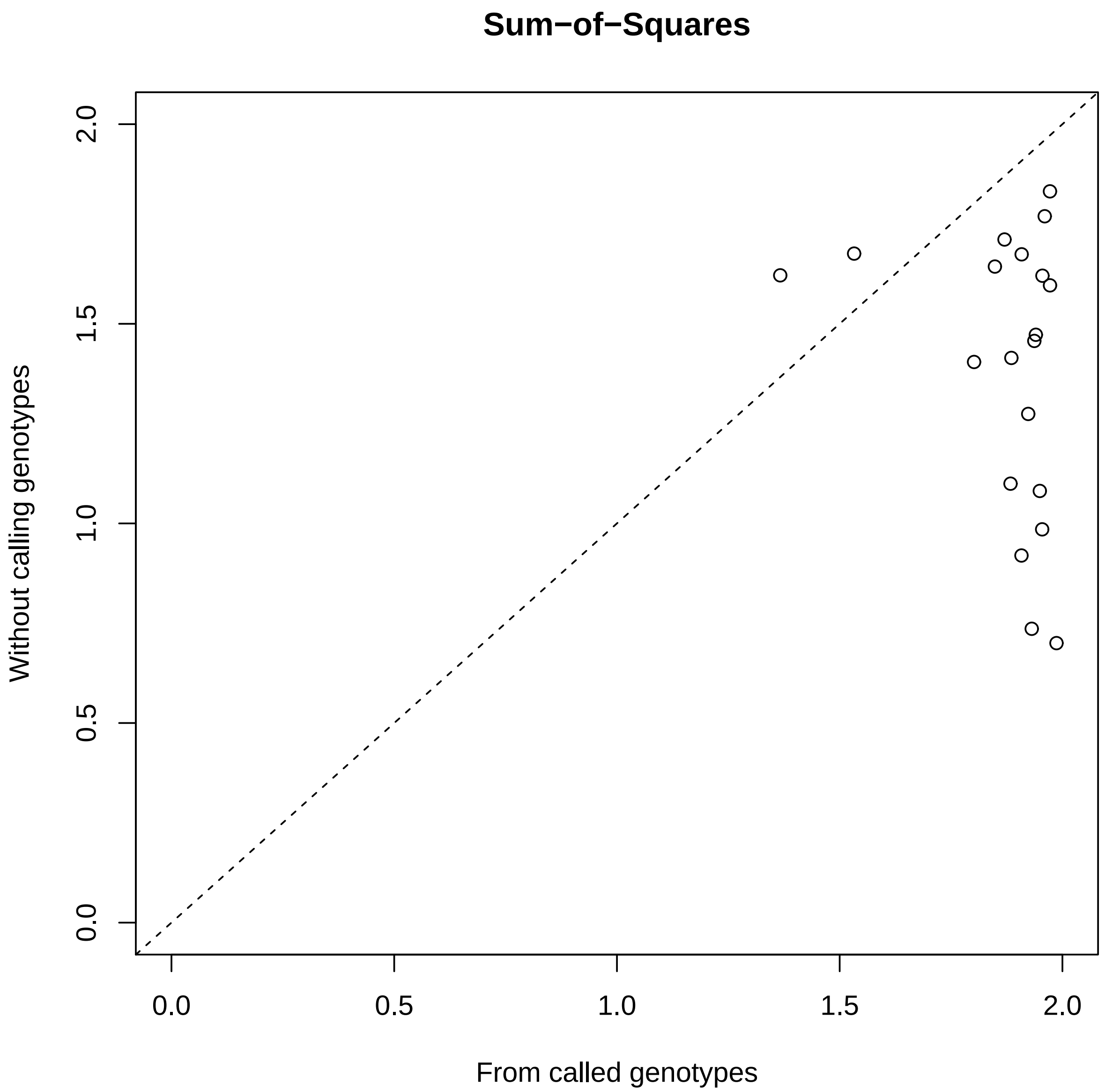

Figure S10: Sum-of-squares (SS) between PC1 and PC2 computed from called genotypes (on x-axis) or with the new proposed method which does not rely on genotype calling (on y-axis). We simulated 1 populations of 40 individuals: half of them were sequenced at 2X coverage and the other half were sequenced at 20X coverage. We simulated $10,000$ sites with 10% of sites being variable in the population.

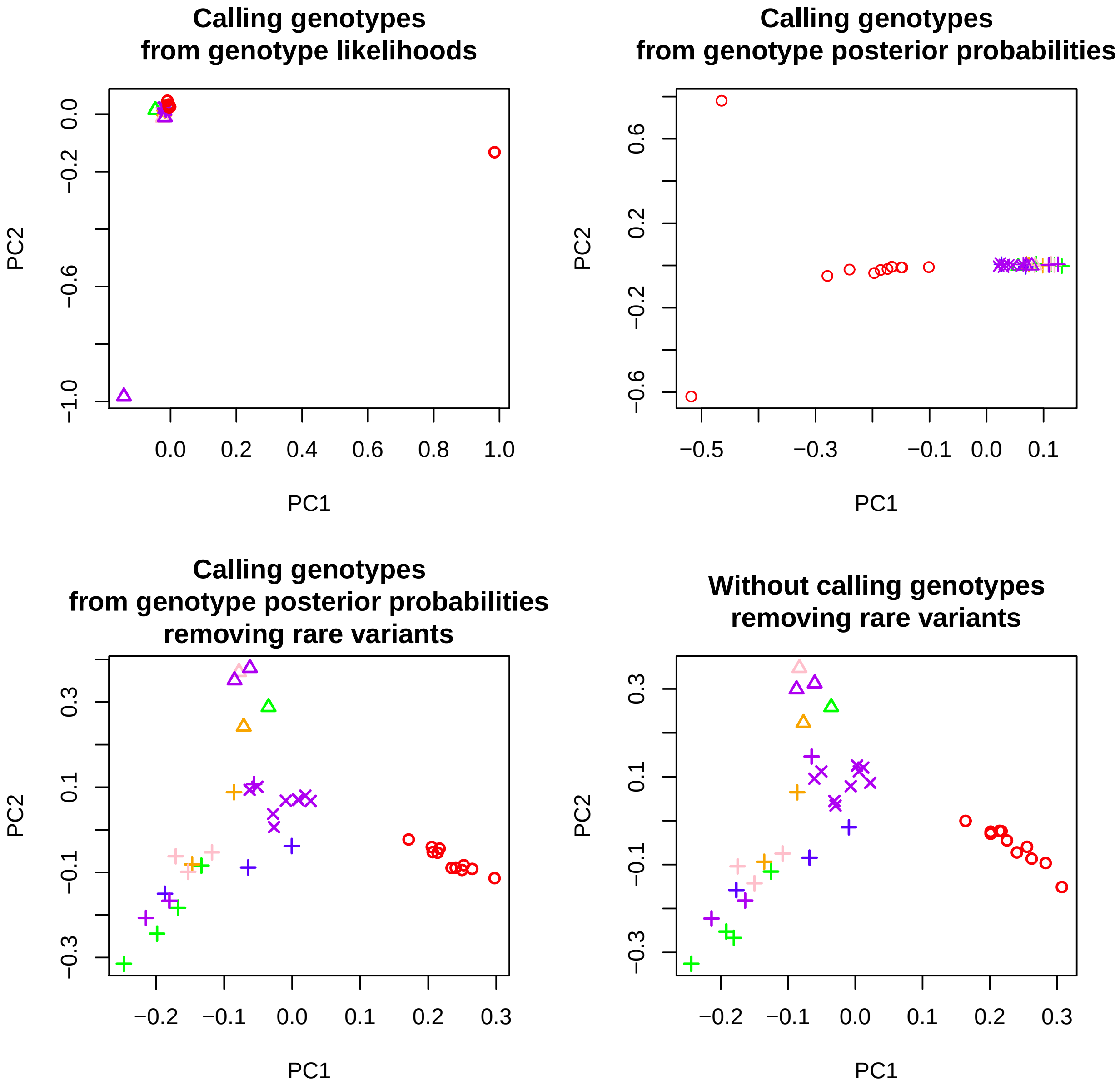

Figure S11: PCA plots for wild and domesticated samples using different strategies of calling genotypes and of filtering data. Legend is the same as Figure 4. Specifically, each lineage has a different shape pattern: hollow circles, wild lineage; hollow triangle: domesticated strain 1; plus sign, domesticated strain 2; multiplication sign, domesticated strain 3. Silkworm systems are colored-coded: green, Japanese; orange, tropical; blue, European; pink, mutant system; purple domesticated from China; red, wild from China.

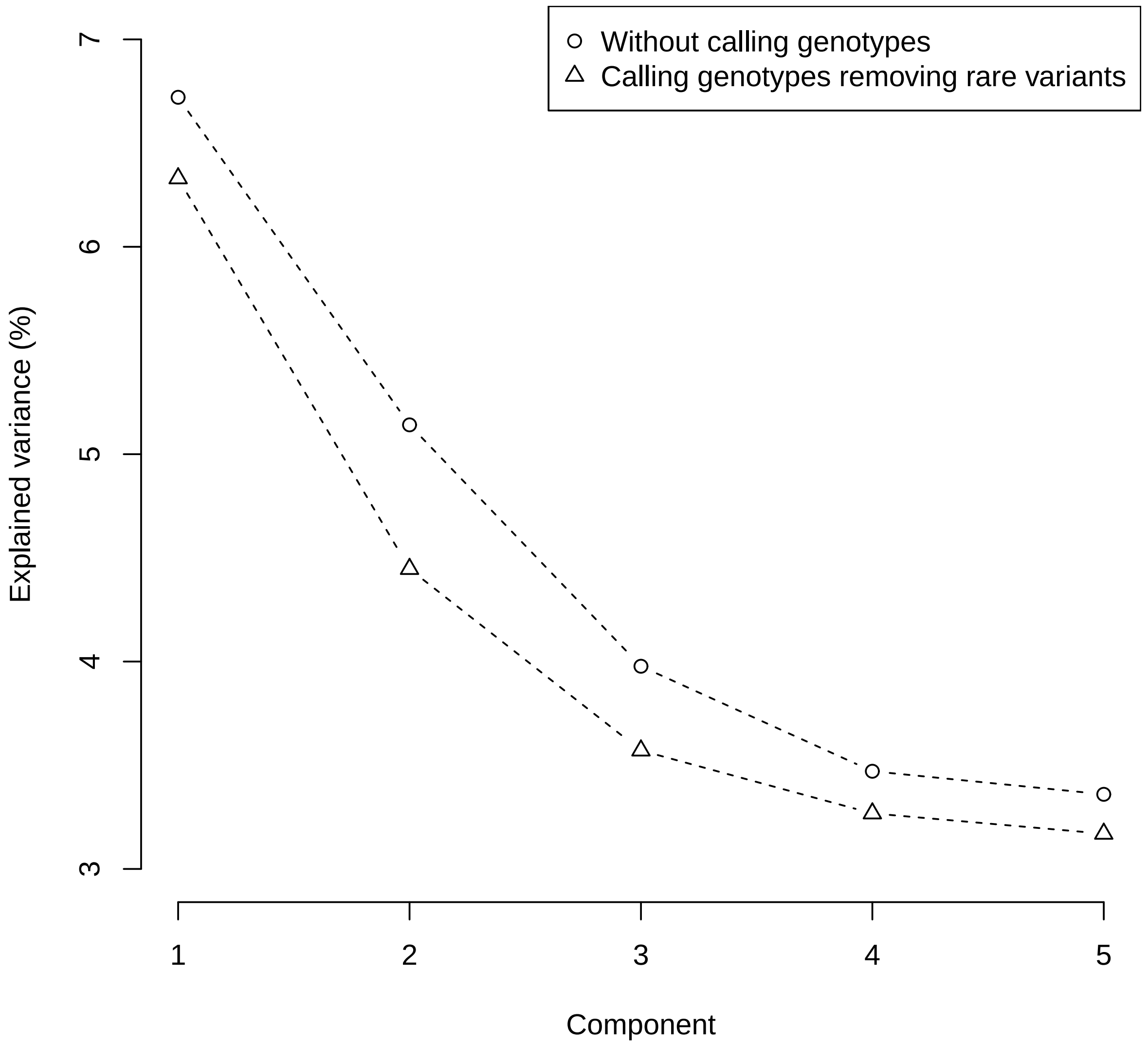

Figure S12: Percentage of explained variance from first components of PCA for wild and domesticated samples from called genotypes or without calling genotypes.

References

S. Y. Kim, K. E. Lohmueller, A. Albrechtsen, Y. Li, T. Korneliussen, G. Tian, N. Grarup, T. Jiang, G. Andersen, D. Witte, T. Jorgensen, T. Hansen, O. Pedersen, J. Wang, and R. Nielsen. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics*, 12:231, Jun 11 2011.