Effect of error and missing data on population structure inference using microsatellite data

Patrick A. Reeves[1,4], Cheryl L. Bowker[2], Christa E. Fettig[2], Luke R. Tembrock[3], and Christopher M. Richards[1]

[1]United States Department of Agriculture, Agricultural Research Service, National Laboratory for Genetic Resources Preservation, 1111 South Mason Street, Fort Collins, CO, 80521, U.S.A.

[2]Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO, 80523

[3] Department of Biology, Colorado State University, Fort Collins, CO, 80523

[4]Corresponding author (pat.reeves@ars.usda.gov; phone: 970-492-7611; fax: 970-492-7605).

Keywords: Bayesian inference, coalescent simulation, Hardy-Weinberg equilibrium, multilocus genotype, neighbor-joining, principal coordinate analysis

Running title: Effect of error and missing data.

ABSTRACT

Missing data and genotyping errors are common in microsatellite data sets. We used simulated data to quantify the effect of these data aberrations on the accuracy of population structure inference. Data sets with complex, randomly-generated, population histories were simulated under the coalescent. Models describing the characteristic patterns of missing data and genotyping error in real microsatellite data sets were used to modify the simulated data sets. Accuracy of ordination, tree-based, and model-based methods of inference was evaluated before and after data set modifications. The ability to recover correct population clusters decreased as missing data increased. The rate of decrease was similar among analytical procedures, thus no single analytical approach was preferable. For every 1% of a data matrix that contained missing genotypes, 2–4% fewer correct clusters were found. For every 1% of a matrix that contained erroneous genotypes, 1–2% fewer correct clusters were found using ordination and tree-based methods. Model-based procedures that minimize the deviation from Hardy-Weinberg equilibrium in order to assign individuals to clusters performed better as genotyping error increased. We attribute this surprising result to the inbreeding-like nature of microsatellite genotyping error, wherein heterozygous genotypes are mischaracterized as homozygous. We show that genotyping error elevates estimates of the level of genetic admixture. Overall, missing data negatively impact population structure inference more than typical genotyping errors.

INTRODUCTION

Short, repetitive regions of the genome, known as microsatellite DNA, simple sequence repeats (SSRs), or short tandem repeats (STRs), are commonly used in molecular population genetic studies (Sunnucks 2000; Guichoux *et al.* 2011). SSR loci exhibit a unique mutational mechanism, slipped-strand mispairing, which causes the duplication or deletion of repeat units, resulting in sequence length variation among alleles (Levinson & Gutman 1987). SSR mutation rates vary widely depending on organism, repeat length, and repeat number, but are generally 1E3–1E4 times higher than a typical nucleotide substitution rate of 1E-8 per generation (Dallas 1992; Chakraborty *et al.* 1997; Vigouroux *et al.* 2002), thus SSR regions provide highly polymorphic markers, useful for distinguishing individuals, reconstructing population history, and estimating demographic parameters.

While single nucleotide polymorphisms (SNPs) may gradually supplant SSRs for certain population genetic applications (Brumfield *et al.* 2003), SSRs remain popular. Over 3500 papers that utilized SSRs were published in 2009 (Guichoux *et al.* 2011), and in 2013–2015, ~33% of articles in the journal *Molecular Ecology* included SSRs as a primary data source. SSR development and typing costs have dropped due to next generation sequencing (Gardner *et al.* 2011) and improved multiplexing protocols (Butler 2005; Holleley and Geerts 2009). We expect SSRs to remain in use due to low cost and their ability to outperform SNPs with fewer loci for individual identification (Seddon *et al.* 2005); parentage and sibship analysis (Glaubitz *et al.* 2003; Wang & Santure 2009); and population structure inference (Liu *et al.* 2005; Glover *et al.* 2010).

Most SSR data sets contain missing data and erroneous genotypes. Missing data are entered into an SSR data matrix when a particular sample does not produce an interpretable pattern of DNA fragments after PCR amplification. PCR failure is usually caused by poor quality template DNA or improper PCR conditions, including mispriming due to mutations at primer binding sites (Guichoux *et al.* 2011). The frequency of missing data is elevated across loci for samples with poor quality DNA. Suboptimal PCR conditions increase missing data across samples for specific loci. Consequently, missing genotypes typically occur non-randomly in data matrices, clumped in rows or columns. Missing data can be minimized by re-extracting DNA and repeating amplifications for problematic individuals, redesigning troublesome primers and optimizing PCR conditions, or excluding individuals and loci with high failure rates from the matrix.

SSR genotyping errors arise from three main sources: the occurrence of null alleles at a locus, preferential amplification of small DNA targets during PCR, and stuttered visualization of amplification products (DeWoody *et al.* 2006; Guichoux *et al.* 2011). “Null alleles” are genotypic variants that fail to amplify under the conditions specified for the locus. They often occur due to mutations at primer binding sites that inhibit amplification but may also arise from poor template quality. Because nothing is amplified, null alleles, by definition, cannot be observed and are not scored. Consequently, in diploid organisms, the genotype entered into the data matrix becomes—erroneously—homozygous for the other, visible allele, or, if both alleles are nulls, missing data. Several analytical approaches have been devised to detect null alleles and estimate their frequency (Chakraborty *et al.* 1992; Raymond & Rousset 1995; Van Oosterhout *et al.* 2004; Kalinowski *et al.* 2007) although only about 40% of studies use them (Guichoux *et al.* 2011). In most cases, eliminating null alleles requires primer redesign outside of highly mutable regions (Dakin & Avise 2004; Chapuis & Estoup 2007).

Preferential amplification of short DNA sequences causes “large allele dropout” error. All else equal, short sequences are more efficiently amplified than long. In a heterozygote with differently-sized alleles this bias may prevent the signal for the large allele from rising above the detection threshold, with the consequence that a heterozygous genotype will be erroneously scored as homozygous for the smaller allele (Wattier *et al.* 1998; Björklund 2005). Large allele dropout can be mitigated in some cases by excluding loci where amplicons exceed 200 bp (Sefc *et al.* 2003).

Slipped-strand mispairing during PCR results in the production of shadow peaks around the amplified allele (Murray *et al.* 1993), a phenomenon termed “stutter”. Stutter peaks are usually smaller than the target, and deviate in size by multiples of the repeat unit length, with progressively decreasing signal (Shinde *et al.* 2003). As with other types of error, stutter causes heterozygotes to be scored as homozygous, but always for the larger of the two alleles, and only when the alleles differ in size by a single repeat unit. Stutter can be reduced by avoiding SSRs with dinucleotide repeats (Chambers & MacAvoy 2000), decreasing denaturation temperature (Olejniczak & Krzyzosiak 2006), and using highly processive polymerases (Davidson *et al.* 2003).

The final assembly of an SSR data set requires considerable care. Filling all cells in a data matrix is time consuming; poor-performing loci must be optimized and recalcitrant individuals must be extracted and genotyped repeatedly. Consequently, the typical course for dealing with missing data is to eliminate problematic individuals or loci. This can produce biased sampling because the frequency of missing data may be similar in related populations (Amos 2006). Some authors have recommended that error rates be reported and efforts made to assess the reliability of conclusions given uncertainty in the genotypes (Bonin *et al.* 2004; Broquet & Petit 2004; Hoffman & Amos 2005). This practice is increasing—error rate estimates can be found in about a quarter of published papers (Guichoux *et al.* 2011)—but it requires sample replication and extensive post-genotyping data analysis, increasing costs. If missing data and genotyping error were to impact the accuracy of population structure inference in only minor ways these steps might be avoided.

The effect of missing data and error on linkage mapping (Hackett & Broadfoot 2003), parentage analysis (Dakin & Avise 2004; Hoffman & Amos 2005; Kalinowski *et al.* 2007), and the estimation of population genetic parameters (Chapuis & Estoup 2007; Hall *et al.* 2012; Peel *et al.* 2013) has been studied in detail, but there have been few investigations of the effect on population structure inference (Pompanon *et al.* 2005; Carlsson 2008; Chapuis *et al.* 2008). In this study we seek to quantify the extent to which missing data or genotyping error impact population structure inference, to assist researchers in developing strategies to produce SSR data sets that maximize accuracy while minimizing costs.

MATERIALS AND METHODS

Highly polymorphic, neutral marker data were simulated using a coalescent model and the software MSMS (Ewing & Hermisson 2010). The simulation approach is detailed in Reeves *et al.* (2012). Briefly, 1E4 data sets were generated, each containing 500 diploid individuals equally distributed among 50 populations, and 50 unlinked loci. An asymmetric island migration model was created by randomly assigning migration rates between populations. The population scaled mutation rate (θ) was varied from 0–0.5. The underlying mutation model of MSMS is an infinite sites model. Unique binary strings output by MSMS were converted into uniquely named alleles following Huelsenbeck and Andolfatto (2007), rendering the infinite sites model as an infinite allele model. We used an infinite allele model instead of a stepwise mutation model as an expedient, and because it is not clear the extent to which the stepwise mutation model fits real SSR data (Gaggiotti *et al.* 1999). It is not unusual for SSR data sets to better fit an infinite allele model than stepwise mutation (e.g. Estoup *et al.* 1995; O'Connell *et al.* 1997). The primary defining feature of real SSR data is a limited number of alleles per locus (Paetkau *et* *al.* 1997). Therefore, data sets containing levels of polymorphism comparable to typical SSR loci (on average < 30 alleles per locus, following Kalinowski 2002) were subsampled from the original 1E4 data sets. A total of 1367 were found and used for further analysis.

Simulated data sets were altered to include missing or erroneous genotypes using models describing the distribution of these data aberrations in typical SSR data sets. In the missing data model, the percent of the matrix containing missing data was varied between zero and 25. A "clumping" parameter was used to bias placement of missing genotypes towards certain loci or individuals. The clumping parameter was varied among data sets from zero, which caused a uniform distribution of missing data, to ten, which elevated the probability ten-fold that the next missing genotype would occur in a row or column already containing missing data. Missing genotypes were substituted for known genotypes one by one, with the probability of conversion for each row and column adjusted after each substitution using the clumping parameter, until the specified percentage of missing data was reached.

To create data sets that varied in their propensity to contain null alleles, a data-set-specific maximum null allele frequency parameter (ν_d) was defined, and selected at random from 0 to 20 percent, following Dakin and Avise (2004). Within each data set, the locus-specific null allele frequency parameter (ν_l) was chosen at random from 0 to ν_d for each locus. The number of null alleles per locus was then defined as ν_l multiplied by the number of distinct alleles at the locus, with the alleles that were to act as nulls chosen randomly. Alleles defined as nulls were treated as unknown, with the consequence that heterozygous genotypes became homozygous for the non-null allele, and homozygous null genotypes became missing data, in the modified data set.

To simulate large allele dropout, it was necessary to assign a probability of dropout for each allele that was proportional to its size. A locus-specific maximum probability of dropout (δ_l) was chosen at random for each locus from 0 to δ_d, the data-set-specific dropout probability. The ceiling on δ_d was set to 0.5 based on empirical studies (Taberlet *et al.* 1996; Gagneux *et al.* 1997; Buchan *et al.* 2005). We assumed a curvilinear function relating dropout probability to allele size, with the largest allele at each locus having a dropout probability of δ_l. Coalescent simulations only provide information on allelic state, not allele size, so relative allele sizes (σ_a) were assigned to all alleles for each locus by randomly sampling an exponential distribution with rate parameter λ, varied by locus from 0 to 10. The probability of retention (i.e. the probability that an allele does not drop out) was then computed for each allele as one minus the cumulative distribution function of the exponential ($\mathrm{CDF} = 1 - e^{-\lambda\sigma_a}$) rescaled to have a maximum value of δ_l. In this way, data sets exhibited varying levels of overall "dropout proneness" governed by the parameter δ_d. Within data sets, the largest alleles at a locus always had the highest dropout probability, but some loci were characterized by dropout probabilities that declined uniformly with consecutively smaller allele size (when $\lambda \rightarrow 0$), while others approximated a threshold effect where alleles above a particular size were highly dropout prone (when $\lambda \rightarrow 10$). If a uniform random number from 0 to 1 exceeded the probability of retention, the allele was made to drop out, and the data set was modified accordingly. Thus, the effect of simulated large allele dropout was to convert heterozygous genotypes to small allele homozygotes with probability $\delta_l(1 - e^{-\lambda\sigma_a})$.

To model stutter error, we identified all heterozygotes having consecutively-sized alleles using the arbitrary sizes from the large allele dropout model. These genotypes are called "adjacent-allele heterozygotes" (Hoffman & Amos 2005). Stutter error could only affect these genotypes, but was not assumed to affect all of them. A model was developed to assign a "probability of stutter error occurrence" to each adjacent-allele heterozygous genotype, thus permitting us to vary the "stutter error proneness" between data sets and among loci. A data-set-specific average probability of stutter error at adjacent-allele heterozygotes ($\overline{\tau_d}$) was randomly chosen from zero to one. $\widetilde{\tau_d}$, the standard deviation of $\overline{\tau_d}$, was randomly set from zero to one. The locus-specific probability of stutter error, τ_l, was sampled from the normal distribution defined by $\overline{\tau_d}$and $\widetilde{\tau_d}$. The conversion of adjacent-allele heterozygotes into large-allele homozygotes then occurred with probability τ_l, the locus-specific conversion probability.

Genotypes in the unmodified matrices were altered in the order that errors would arise during the genotyping process. Null allele errors, which are caused by mispriming, were added first, followed by large allele dropout errors, caused by poor amplification, then by stutter errors, caused by slippage during amplification but attributable primarily to poor scoring procedures. Realized error rates were then recalculated for each error type, for each data set.

Three categorically-distinct analytical approaches for inferring population structure were used. First, we applied a class of Bayesian Markov chain Monte Carlo (MCMC) methods introduced by Pritchard *et al.* (2000) in the software STRUCTURE. To avoid *ad hoc* model selection procedures for determining the number of populations (K) (Evanno *et al.* 2005), we used INSTRUCT v1.0 (Gao *et al.* 2007) and STRUCTURAMA v1.0 (Huelsenbeck and Andolfatto 2007) instead of STRUCTURE. For INSTRUCT analyses, we used no-admixture (mode 0) and admixture (mode 1) models, as well as a model that estimates individual inbreeding coefficients simultaneously with individual assignment (mode 5). Modes 0 and 1 of INSTRUCT are comparable to no-admixture and admixture models of STRUCTURE. A single

Markov chain was run for 1E5 generations and sampled every 25, with the initial 12500 generations discarded. The deviance information criterion (DIC) was used to determine the best value of K between 1 and 50 (Gao *et al.* 2011). A discrete assignment was created by assigning individuals to clusters based on the highest assignment probability in the Q-matrix. STRUCTURAMA estimates K alongside individual assignment. Chains were run as for INSTRUCT, the prior on number of populations was set to two, and no admixture was allowed.

Second, we used a tree-based approach. Neighbor-joining (NJ) trees were constructed using NTSYS (Rohlf 2008). Inter-individual distances were computed using Lynch's (1990) band sharing coefficient. Populations were counted as correctly inferred when they existed as a monophyletic group in the resulting tree. Third, an ordination method was applied. We used PCOMC, a procedure that couples ordination with cluster analysis to simultaneously determine population number and membership (Reeves & Richards 2009). Principal coordinate analysis was performed using NTSYS. Distance matrices, computed as for NJ, were double-centered prior to the calculation. Principal coordinate values were weighted according to their contribution to the total variance, then subject to the density clustering algorithm PROC MODECLUS (SAS Institute, Cary, NC).

Correct and incorrect clusters resulting from application of each analytical method to each data set were counted. Correct clusters were those that contained all 10 individuals that belonged to a single population as specified in the coalescent simulation model, and no others. Incorrect clusters contained some, but not all, members of a population in the model, or individuals from more than one population. The "performance ratio" was defined as the number of correct clusters resulting from analysis of a modified data set divided by the number of correct clusters in the unmodified data set from which it was derived. A performance ratio < 1 indicates that data modification reduces accuracy, while a performance ratio > 1 indicates improved accuracy. We used the partition distance of Gusfield (2002) as a second, less strict measure of accuracy for methods that produce partitions (INSTRUCT and STRUCTURAMA). The partition distance (PD) has been used previously for this purpose, and is useful because it can quantify partially correct matches between clusters, unlike the performance ratio (Huelsenbeck and Andolfatto 2007; Choi and Hey 2011). It is defined as the number of elements that must be moved between clusters to make one partition identical to another. We normalized the PD to the range 0–1 by dividing by the maximum possible PD (Charon *et al.* 2006), calling the result PD_n. We define the "partition distance ratio" as $\frac{1-PD_n^{modified}}{1-PD_n^{unmodified}}$, where numerator and denominator are formulated as similarities to simplify comparison with the performance ratio. $PD_n^{modified}$ is the normalized partition distance between the partition resulting from analysis of the modified data and the 50 cluster partition defined in the coalescent simulation model (likewise for $PD_n^{unmodified}$). The performance ratio and the partition distance ratio are non-identical, but correlated, measures of the accuracy of population structure inference (Supplementary Table 1). We preferentially report the performance ratio, because its interpretation is intuitive, and it is applicable to all methods.

The false discovery rate (FDR) was calculated as the number of incorrect clusters divided by the total number of clusters returned. A ratio, $\frac{FDR^{modified}}{FDR^{unmodified}}$, was calculated to express the difference in probability of recovering incorrect clusters between modified and unmodified data sets. Multiple regression and likelihood analysis were used to examine the effect of model factors on the accuracy of inference.

RESULTS

Data set validation. The 1367 simulated data sets contained an average of 14.48 ± 8.46 (1 sd) alleles per locus. Population structures ranged from virtually panmictic to highly subdivided (Figure 1). Levels of population subdivision, as quantified by Hedrick's (2005) G'$_{st}$, varied from 0.006 to 0.999, with an excess of low G'$_{st}$ values. The degree to which data sets were modified by the missing data model was not significantly related to the level of population differentiation ($r = 0.03$, $p = 0.32$); application of the error models resulted in a slight, but significant, negative correlation between G'$_{st}$ and the proportion of genotypes modified ($r = -0.14$, $p < 0.0001$ (Supplementary Figure 1).

Performance. The analytical methods differed in accuracy when unmodified data sets were analyzed (Figure 2). For a given level of genetic subdivision, correct clusters were found in NJ trees at a much higher frequency than in INSTRUCT, STRUCTURAMA or PCOMC analyses. This is not surprising because the criterion for identifying correct clusters for NJ was less strict than for other methods—a correct inference occurred when a node existed that defined a population correctly in the tree, but there was no mechanism to determine which nodes defined populations. PCOMC had the lowest rate of correct cluster recovery. STRUCTURAMA recovered more correct clusters and had a lower false discovery rate than INSTRUCT for high levels of genetic subdivision, while INSTRUCT was more accurate at lower levels, regardless of mode.

To avoid undefined values when calculating the performance ratio, data sets were excluded when zero correct clusters were inferred with unmodified data. Likewise, for the partition distance ratio, data sets were excluded when $PD_n^{unmodified} = 1$, i.e. when the observed partition was maximally distant from the simulated partition. Because the coalescent model was complex and population subdivision was often low, a substantial number of data sets were excluded using this restriction. The total number of useful data sets ranged from 63 to 473, depending on analytical method, when measured using the performance ratio (Table 1), and 115 to 464 for the partition distance ratio (Supplementary Table 1). Population subdivision in the excluded data sets was low (G'$_{st}$ ~5-fold lower than retained data sets), approaching or exceeding the limits of resolution of the methods applied.

Performance decreased in a roughly linear manner as missing data increased (Figure 3a–f). The slope of the regression line was significantly different from zero (at $\alpha = 0.01$, Holm-Bonferroni corrected, here and throughout) for all methods except 'INSTRUCT inbreeding' (mode 5) and PCOMC (Table 1). R^2_{adj} values for significant regressions ranged from 0.08 to 0.27 and the slope of performance loss ranged from $m = -1.8 - -3.5$ between analytical approaches (Table 1). The results using the partition distance ratio were similar. A significant negative correlation was found for 'INSTRUCT no admixture' (mode 0) and 'INSTRUCT admixture' (mode 1); the correlation was not significant for 'INSTRUCT inbreeding' or STRUCTURAMA (Supplementary Table 1).

Taking into account 95% confidence intervals on the slopes, a data matrix with 5% missing data is predicted to result in recovery of at least 72–81% of the correct clusters that could be recovered with a complete data matrix. Based on our data, 95% of recoverable clusters should be found when data matrices contain, for INSTRUCT, 2.5–2.7% missing data, or, for the other methods, 1.5–1.6% missing data. The FDR ratio increased significantly ($\alpha = 0.05$) with missing data for all methods except 'INSTRUCT inbreeding' and PCOMC, indicating that, in addition to fewer correct clusters, users should expect more erroneous clusters as missing data increase.

The clumping parameter had a statistically significant effect on performance for NJ only (Table 2). Performance of NJ was higher when missing values were more clumped ($\beta = 0.02$). Using ratios of the Akaike weights, the linear model for percent missing data alone had a much higher probability (10^{25}-fold) than the model for clumping parameter for NJ. Regardless of method, less than 4% of the variance in performance was attributable to clumping—a slight effect—thus the clumping parameter, despite adding realism to the simulation, was largely dispensable.

The effect of genotyping error differed between categories of analytical method. For distance methods NJ and PCOMC, performance declined as erroneous data increased (Figure 3k,l). The slopes of the regressions were significantly different from zero and about half the magnitude found for missing data (Table 1). A matrix with 5% erroneous data should result in recovery of 85% (NJ) or 82% (PCOMC) of the clusters that would be found if the data set contained no error. Ninety five percent of recoverable clusters were found when 2.9% (NJ) or 3.6% (PCOMC) of the data matrix was erroneous. For NJ, large allele dropout and stutter had the greatest effect on deteriorating performance (Table 2).

In contrast, accuracy of model based methods improved as erroneous data increased. When using the performance ratio, the slope of the regression was positive for all methods (0.41–1.36) (Figures 3g-j, Table 1). The relationship was statistically significant for STRUCTURAMA. When using the partition distance ratio, a significant positive correlation was found for all methods (Figure 4, Supplementary Table 1). Large allele dropout was the most important model effect for INSTRUCT and STRUCTURAMA, explaining 2–10% of the variation in performance (combined model, 5–16%), and holding 18–78% of the linear models'

Akaike weight (model probability). The FDR ratio did not change significantly with increasing error for any method.

DISCUSSION

Genomewide genotyping approaches are gaining popularity for studies of population structure but SSRs continue to be used due to low cost and high power of inference. Although SSR data sets are prone to containing missing data and erroneous genotypes, few studies have examined their impact on analyses of population structure (Pompanon *et al.* 2005). We explored the effect of missing data and genotyping errors on population structure inference using model-based Bayesian MCMC procedures (INSTRUCT, STRUCTURAMA), a tree-based method (NJ), and an ordination approach (PCOMC). Our goal is to provide users of SSRs with insight for how much missing data and error might be tolerated in order to achieve a desired level of accuracy.

In order to make general recommendations it was necessary to explore a diverse set of population structures. This was accomplished using coalescent modeling with key parameters—mutation, migration rates, migration directionality—set stochastically, but within plausible ranges (Reeves *et al.* 2012). The resultant data sets exhibited a large range of complex population structures with widely varying levels of subdivision (Figure 1, Supplementary Figure 1). Nevertheless, the extent to which simulated data can ever accurately represent nature is debatable. We attempted to produce a realistic subsample from the universe of plausible population structures. Other models describing population divergence exist, most notably the isolation with migration model (Nielsen & Wakeley 2001). Our conclusions should be interpreted as limited to modified island models similar to those simulated, and should not be expected to precisely predict outcomes for any single data set.

Key properties peculiar to SSR data sets were modeled, including the clumped nature of missing data, and the three best understood sources of error: null alleles, large allele dropouts, and stutter artifacts. We did not include an assessment of human error, which contributes substantially (Bonin *et al.* 2004), but is not easily modeled. While most real data sets will be affected by both missing data and error, we elected to study these factors independently because the course of action for mitigation seems distinct.

Impact of missing data. For data sets modified with missing data three basic observations were made. First, the degradation of performance as missing data increased was roughly linear (Figure 3a–f). Second, the rate at which performance declined was similar among methods (Table 1). Third, the degree to which missing data were clumped by locus or DNA sample was unimportant (Table 2)—contiguous blocks of missing genotypes did not affect performance much more than an equal number of randomly-distributed missing genotypes. These observations lead to a simple prediction: for every 1% of a data matrix that contains missing data, researchers should expect to recover 2-4% fewer correct clusters than would be inferred with a complete data set.

Impact of error on distance methods. Error caused a decline in performance for the tree-based NJ method and for the ordination approach PCOMC. Researchers should expect a ~2% decrease in the number of correct clusters recovered for every 1% of the data matrix impacted by error when using these methods.

Although NJ and PCOMC are quite distinct mathematically, the similarity in response between the methods may arise from their use of the same distance metric. Distance was calculated as $1 - S_{xy}$, where S_{xy} is the ratio of alleles common to two individuals, over the total, or $\frac{|A \cap B|}{|A \cup B|}$, i.e. the Jaccard index (here, A and B represent the sets of unique alleles from each individual). The mechanism by which performance decreases is loss of precision in the genetic distance estimate. Conversion of heterozygotes to homozygotes decreases the number of alleles, lowering $|A \cup B|$. The maximum number of distinct values the index can assume, $|A \cup B| + 1$, also decreases, imposing a limit on the assembly of individuals into different clusters. Missing data had a stronger effect than error (Figure 3e,f,k,l). Conversion of a heterozygous locus to missing data may reduce $|A \cup B|$ by two, whereas with error $|A \cup B|$ can only decrease by one, thus the erosion of resolving power is generally more rapid with missing data.

Impact of error on model-based methods. We observed a positive relationship between the amount of error introduced into a data set and the accuracy of inference when using model-based methods (Figures 3-4, Table 1, Supplementary Table 1). Our data suggest that a matrix containing 25% erroneous data could cause the recovery of 30–70% more correct clusters than an error-free matrix. Hereafter, we attempt to explain this unexpected result.

When a population consists of several subpopulations, a "heterozygote deficit" occurs, where the observed heterozygosity of the population analyzed as a whole is lower than predicted under Hardy-Weinberg equilibrium. This "Wahlund effect" (Wahlund 1928) forms the theoretical justification, and source of signal, for a class of model-based clustering methods that includes STRUCTURE, STRUCTURAMA, INSTRUCT, and others (e.g. Dawson and Belkhir 2001; Corander *et al.* 2003), which minimize the deviation from Hardy-Weinberg expectations by partitioning individuals into distinct subpopulations. Unfortunately, the signal upon which these methods rely is not uniquely generated by population subdivision. Heterozygote deficit also occurs as a consequence of selfing or inbreeding (Gao *et al.* 2007). Likewise, because SSR genotyping error typically involves the conversion of heterozygotes to homozygotes, it too induces a heterozygote deficit. In our simulation, a decline in observed heterozygosity with increasing error inflated the inbreeding coefficient $F = 1 - \frac{H_o}{H_e}$. As error increased, the heterozygote deficit $D = \frac{H_o - H_e}{H_e}$ likewise increased (since D is the additive inverse of F). Thus error produces an inbreeding-like signature in data sets, which magnifies the Wahlund effect and, as a byproduct, inflates the level of genetic subdivision, causing increased signal (Supplementary Figure 2).

If this sort of signal amplification is important we would expect model factors ΔF (the change in inbreeding between unmodified and modified data due to error), and/or ΔG’st (the change in genetic subdivision) to drive the positive relationship between error and accuracy. Some evidence supports this. When accuracy was measured using the partition distance ratio, ΔG’st was the second-most important model effect for all INSTRUCT modes, explaining 3-10% of the variation (Supplementary Table 3). For ‘INSTRUCT no admixture’, ΔF was the most important. ΔG’st and ΔF were also important model effects for STRUCTURAMA, explaining 5-15% of the variation in the partition distance ratio. However, for STRUCTURAMA, the overwhelming weight of evidence was that ΔK, the difference in the number of populations inferred, was the primary driver of the response. Thus the improvement in accuracy observed for STRUCTURAMA seems to have an additional underlying cause.

STRUCTURAMA uses a Dirichlet process prior to simultaneously infer K and individual assignment. A “Dirichlet process” describes a distribution of probability distributions and is a mathematical construct commonly used as a prior in MCMC procedures to categorize items into groups when the number of groups is unknown. During progression of the Markov chain in

STRUCTURAMA, a critical decision is made for each individual at each step: whether the individual belongs to an existing cluster or should be assigned to a new cluster. This decision affects both assignment and the inference of K. The algorithm proceeds roughly as follows. First, an individual *i* is selected at random from the existing partition and removed. Individual *i* is then re-assigned to whichever of the K clusters it fits best, or to a new cluster, all by itself. The probability that the individual is re-assigned to an existing cluster, *k*, is dependent on the number of individuals, η, in that cluster (large clusters are more attractive) and the marginal posterior probability of drawing *i*'s genotype from cluster *k*. In contrast, the probability of assigning *i* to a new cluster depends on the concentration parameter, α, of the Dirichlet process (the higher α is, the lower the probability that two randomly drawn individuals belong to the same cluster) and the probability of drawing *i*'s genotype from the prior distribution of allele frequencies, where all alleles are equiprobable.

In our unmodified data sets, the deviation from Hardy-Weinberg equilibrium is due to subpopulation structure plus some minor, random effects of the coalescent. However, when error is introduced, part of the deviation is then derived from the inbreeding-like effect of SSR error. With error, the relative probability of assigning an individual to an existing cluster versus a new cluster decreases—data with heterozygote deficit, by definition, do not fit nicely into a Hardy-Weinberg scenario—and, accordingly, it becomes more likely that an individual will be assigned to a new cluster, thereby increasing K. Consistent with this expectation, K increased significantly with error (Figure 5a). When more clusters are available in which to distribute individuals, more correct inferences can be made (equivalently, the number of correct inferences possible is suppressed for low levels of error). Consequently, performance improves with increasing error primarily because K increases. While convenient, the STRUCTURAMA approach for defining K may be incautious. Miller and Harrison (2014) argued that Dirichlet process models should not be used to infer the number of components.

INSTRUCT uses the Deviance Information Criterion (DIC) to select the best K value from a user-defined range. In the present case, a model with K=50 will usually have a higher likelihood than one with, say, K=4, but that does not necessarily mean it is better, it may merely be overfit. The DIC, like other model selection statistics, deals with this by imposing a penalty on the likelihood for adding parameters. The DIC value is, essentially, the mean likelihood across MCMC draws at stationarity, penalized for increasing K. The K value that minimizes the DIC across the range is deemed optimal. This method for choosing K appears largely insensitive to the level of error, and in this sense seems preferable to the Dirichlet process used by STRUCTURAMA. There is currently no consensus on the best procedure to estimate the number of groups in a finite mixture. Until the basic mathematical issues are resolved, the biological problem of estimating K for population structure inference will remain more craft than science.

Impact of error on the estimation of admixture proportions. Using an admixture model and STRUCTURE, Gao *et al.* (2007) showed that inbreeding causes two spurious results: exaggerated levels of admixture and elevated likelihoods for higher K values. Falush *et al.* (2003, p. 1572) predicted this: "…situations that might cause additional populations to be inferred by STRUCTURE…include a significant frequency of inbreeding, cryptic relatedness within the sample, or the presence of null alleles." The implication is that heterozygote-deficit-inducing factors other than population subdivision might promote the inference of more clusters, or more admixture among clusters, than is actually the case. For 'INSTRUCT admixture', the relative level of inferred admixture increased with error, and likelihoods were elevated for higher K values (Figure 5b,d). This did not, however, translate into the inference of higher K values under the DIC ($p > 0.19$).

To reduce artifacts from conflation of signal types, Gao *et al.* (2007) relaxed the assumption of Hardy-Weinberg equilibrium to estimate an inbreeding parameter alongside subpopulation membership. Relative to 'INSTRUCT admixture,' the relationship between error and admixture for 'INSTRUCT inbreeding' remained positive, but that between K and the model likelihood was reversed: *Ln*L decreased as K increased (Figure 5c,d). Thus the inbreeding model of Gao *et al.* (2007) compensates in some ways for genotyping error, but does not resolve the issue of admixture overestimation. The effects of error might be mitigated by explicitly modeling it as an additional parameter that alters Hardy-Weinberg expectations at each locus independently, distinct from the inbreeding coefficient estimated by INSTRUCT, which alters expectations uniformly across loci.

Whether over-estimation of admixture proportions due to genotyping error is a problem in real data sets is unknown, but the potential exists. False confidence in our understanding of the partitioning of genetic variation in nature notwithstanding, artifactually-elevated admixture estimates could have important economic implications, in endangered species management or genomewide association analysis, for example. Newer forms of data, like SNPs, are not immune to the problem. High throughput genotyping and next generation sequencing approaches vary in their capacity to accurately identify heterozygotes (Harismendy *et al.* 2009; Skotte *et al.* 2013), sometimes defaulting to the homozygous condition, akin to SSR error.

Recommendations. Missing data exert a greater negative effect on the inference of population structure than typical SSR genotyping errors. As missing data increase, the number of correct clusters recovered decreases and the number of incorrect clusters increases. In order to efficiently produce SSR data sets for inferring population structure we recommend limiting the percent of a matrix that contains missing data to ~2%, unless a greater amount can be justified based on the particular hypotheses under examination. This will allow researchers to retain most of the resolving power of their data while not incurring the extra costs associated with completing a data matrix. For analyses reliant on simple distance metrics, we recommend limiting the error rate to ~4% of scored genotypes. Model-based population structure inference methods handle genotyping error well. We recommend the use of admixture models to infer cluster membership, but caution that admixture estimates may be artificially elevated.

ACKNOWLEDGEMENTS

This study was conceived during a graduate seminar at Colorado State University organized by Mark P. Simmons, who we thank for critical reading of the manuscript. This research used resources from the Open Science Grid (supported by the National Science Foundation and the U.S. Department of Energy's Office of Science), the iPlant Collaborative (National Science Foundation grant #DBI-0735191. URL: www.iplantcollaborative.org), and the HTCondor group in the Department of Computer Sciences at the University of Wisconsin-Madison. We thank Nirav Merchant, Suchandra Thapa, Greg Thain, John Fonner and Nicole Hopkins for helping us find spare compute cycles and optimize HTC strategies at iPlant, the OSG, and UW.

REFERENCES

Amos W (2006) The hidden value of missing genotypes. Molecular Biology and Evolution, 23, 1995–1996.

Björklund M (2005) A method for adjusting allele frequencies in the case of microsatellite allele drop-out. Molecular Ecology Notes, 5, 676–679.

Bonin A, Bellemain E, Bronken Eidesen P et al. (2004) How to track and assess genotyping errors in population genetics studies. Molecular Ecology, 13, 3261–3273.

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. Molecular Ecology, 13, 3601–3608.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology & Evolution, 18, 249–256.

Buchan JC, Archie EA, Van Horn RC, Moss CJ, Alberts SC (2005) Locus effects and sources of error in noninvasive genotyping. Molecular Ecology Notes, 5, 680–683.

Butler JM (2005) Constructing STR multiplex assays. In: Forensic DNA Typing Protocols , pp. 53–65. Springer.

Carlsson J (2008) Effects of microsatellite null alleles on assignment testing. Journal of Heredity, 99, 616–623.

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proceedings of the National Academy of Sciences, 94, 1041–1046.

Chambers GK, MacAvoy ES (2000) Microsatellites: consensus and controversy. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, 126, 455–476.

Chapuis M-P, Estoup A (2006) Microsatellite null alleles and estimation of population differentiation. Molecular Biology and Evolution, 24, 621–631.

Chapuis M-P, Lecoq M, Michalakis Y, Loiseau A, Sword GA, Piry S, Estoup A (2008) Do outbreaks affect genetic population structure? A worldwide survey in Locusta migratoria, a pest plagued by microsatellite null alleles. Molecular Ecology, 17, 3640–3653.

Charon I, Denœud L, Guénoche A, Hudry O (2006) Maximum transfer distance between partitions. Journal of Classification, 23, 103–121.

Choi SC, Hey J (2011) Joint inference of population assignment and demographic history. Genetics, 189, 561–577.

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics, 163, 367–374.

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. Heredity, 93, 504–509.

Dallas JF (1992) Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. Mammalian Genome, 3, 452–456.

Davidson JF, Fox R, Harris DD, Lyons-Abbott S, Loeb LA (2003) Insertion of the T3 DNA polymerase thioredoxin binding domain enhances the processivity and fidelity of Taq DNA polymerase. Nucleic Acids Research, 31, 4702–4709.

Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genetical Research, 78, 59–77.

DeWoody J, Nason JD, Hipkins VD (2006) Mitigating scoring errors in microsatellite data from wild populations. Molecular Ecology Notes, 6, 951–957.

Estoup A, Garnery L, Solignac M, Cornuet J-M (1995) Microsatellite variation in honey bee (Apis mellifera L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. Genetics, 140, 679–695).

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 14, 2611–2620.

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics, 26, 2064–2065.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164, 1567–1587.

Gaggiotti OE, Lange O, Rassmann K, Gliddon C (1999) A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. Molecular Ecoloty, 8, 1513–1520.

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology, 6, 861–868.

Gao H, Bryc K, Bustamante CD (2011) On Identifying the Optimal Number of Population Clusters via the Deviance Information Criterion. PLoS ONE, 6, e21014.

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics, 176, 1635–1651.

Gardner MG, Fitch AJ, Bertozzi T, Lowe AJ (2011) Rise of the machines - recommendations for ecologists when using next generation sequencing for microsatellite development. Molecular Ecology Resources, 11, 1093–1101.

Glaubitz JC, Rhodes OE, DeWoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Molecular Ecology, 12, 1039–1047.

Glover K, Hansen M, Lien S et al. (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. Bmc Genetics, 11, 2.

Guichoux E, Lagache L, Wagner S et al. (2011) Current trends in microsatellite genotyping. Molecular Ecology Resources, 11, 591–611.

Gusfield D (2002) Partition-distance: a problem and class of perfect graphs arising in clustering. Information Processing Letters, 82, 159–164.

Hackett CA, Broadfoot LB (2003) Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity, 90, 33–38.

Hall N, Mercer L, Phillips D, Shaw J, Anderson AD (2012) Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies. Genetics Research, 94, 151–161.

Harismendy O, Ng PC, Strausberg RL et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol, 10, R32.

Hedrick PW (2005) A standardized genetic differentiation measure. Evolution, 59, 1633–1638.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. Molecular Ecology, 14, 599–612.

Holleley C, Geerts P (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. BioTechniques, 46, 511–517.

Huelsenbeck JP, Andolfatto P (2007) Inference of population structure under a Dirichlet process model. Genetics, 175, 1787–1802.

Kalinowski ST (2002) How many alleles per locus should be used to estimate genetic distances? Heredity, 88, 62–65.

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Molecular Ecology, 16, 1099–1106.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Molecular biology and evolution, 4, 203–221.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genetics, 6, S26.

Lynch M (1990) The similarity index and DNA fingerprinting. Molecular Biology and Evolution, 7, 478–484.

Miller JW, Harrison MT (2014) Inconsistency of Pitman-Yor process mixtures for the number of components. Journal of Machine Learning Research, 15, 3333–3370.

Murray V, Monchawin C, England PR (1993) The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. Nucleic Acids Research, 21, 2395–2398.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics, 158, 885–896.

O'Connell M, Danzmann RG, Cornuet J-M, Wright J, Ferguson MM (1997) Differentiation of rainbow trout (Oncorhynchus mykiss) populations in Lake Ontario and the evaluation of the stepwise mutation and infinite allele mutation models using microsatellite variability. Canadian Journal of Fisheries and Aquatic Sciences, 54, 1391–1399.

Olejniczak M, Krzyzosiak WJ (2006) Genotyping of simple sequence repeat factors implicated in shadow band generation revisited. Electrophoresis, 27, 3724–3734.

Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C (1997) An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics, 143, 1943–1957.

Peel D, Waples RS, Macbeth GM, Do C, Ovenden JR (2013) Accounting for missing data in the estimation of contemporary genetic effective population size (N_e). Molecular Ecology Resources, 13, 243–253.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics, 6, 847–846.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics, 155, 945–959.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Journal of Heredity, 86, 248–249.

Reeves PA, Panella LW, Richards CM (2012) Retention of agronomically important variation in germplasm core collections: implications for allele mining. Theoretical and Applied Genetics, 124, 1155–1171.

Reeves PA, Richards CM (2009) Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. PLoS One, 4.

Rohlf FJ (2008) NTSYSpc: Numerical Taxonomy System, ver. 2.11x. Exeter, Setauket, NY.

Seddon JM, Parker HG, Ostrander EA, Ellegren H (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. Molecular Ecology, 14, 503–511.

Sefc KM, Payne RB, Sorenson MD (2003) Microsatellite amplification from museum feather samples: effects of fragment size and template concentration on genotyping errors. The Auk, 120, 982.

Shinde D, Lai Y, Sun F, Arnheim N (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Research, 31, 974–980.

Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. Genetics, 195, 693–702.

Sunnucks P (2000) Efficient genetic markers for population biology. Trends in Ecology & Evolution, 15, 199–203.

Taberlet P, Griffin S, Goossens B et al. (1996) Reliable genotyping of samples with very low DNA quantities using PCR. Nucleic Acids Research, 24, 3189–3194.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes, 4, 535–538.

Vigouroux Y, Jaqueth JS, Matsuoka Y et al. (2002) Rate and pattern of mutation at microsatellite loci in maize. Molecular Biology and Evolution, 19, 1251–1260.

Wahlund S (1928) Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas, 11, 65–106.

Wang J, Santure AW (2009) Parentage and sibship inference from multilocus genotype data under polygamy. Genetics, 181, 1579–1594.

Wattier R, Engel CR, Saumitou-Laprade P, Valero M (1998) Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in Gracilaria gracilis (Rhodophyta). Molecular Ecology, 7, 1569–1573.

DATA ACCESSIBILITY

Computer programs used to simulate missing data and genotyping error are available from the authors.

AUTHOR CONTRIBUTIONS

Designed research, all authors. Ran analyses, all authors. Analyzed results, PAR. Wrote computer programs, PAR. Wrote the manuscript, all authors.

FIGURE LEGENDS

Figure 1. Some complex population structures generated by coalescent simulation, visualized using principal coordinate analysis. Top row, high population subdivision ($G'_{st} = 0.99$); middle row, intermediate ($G'_{st} = 0.5$); bottom row, low ($G'_{st} = 0.01$).

Figure 2. Accuracy of population structure inference for unmodified data sets. X-axis measures population subdivision, Y-axis shows the proportion of correctly inferred populations out of 50 possible (black), and the false discovery rate (grey). a) INSTRUCT no admixture; b) INSTRUCT admixture; c) INSTRUCT inbreeding; d) STRUCTURAMA; e) Neighbor-joining; f) PCOMC. False discovery rate not calculable for neighbor-joining.

Figure 3. Effect of missing data and error on clustering accuracy using the performance ratio. a,g) INSTRUCT no admixture; b,h) INSTRUCT admixture; c,i) INSTRUCT inbreeding; d,j) STRUCTURAMA; e,k) Neighbor-joining; f,l) PCOMC. For visual clarity, points are mean values binned with an interval width of 0.01. Error bars indicate SEM. Statistics were calculated without binning. Dotted curves mark the 95% confidence interval on the slope of the regression. Grey shaded area is the 95% confidence interval on the slope of the regression for the false discovery rate ratio. False discovery rate not calculable for neighbor-joining. The range of values in f) is truncated because principal coordinate analysis can accept limited missing data.

Figure 4. Effect of error on clustering accuracy using the partition distance ratio. a) INSTRUCT no admixture; b) INSTRUCT admixture; c) INSTRUCT inbreeding; d) STRUCTURAMA. Dotted curves mark the 95% confidence interval on the slope of the regression. For clarity, a total of 8 outlying points have been omitted from the plots, but were included in the regression.

Figure 5. a) As genotyping error increases the optimal value for K is elevated under a Dirichlet process prior using STRUCTURAMA. b) The level of inferred admixture, as measured by the Shannon index, increases with error for ‘INSTRUCT admixture’. c) The upward bias in the admixture estimate is not corrected by using ‘INSTRUCT inbreeding’. d) Error causes the likelihood to be elevated for higher values of K, plateauing after K > 10. This effect is reversed by using a model that compensates for inbreeding. Black line, ‘INSTRUCT admixture’; grey line, ‘INSTRUCT inbreeding’.

TABLES

Table 1. Regression analysis of clustering accuracy, measured with the performance ratio.

Data set	Method	n_{obs}	m^a	b^b	R^2_{adj}	p^c
Missing						
	INSTRUCT no admixture	188	-1.96	1.02	0.11	2.0E-06***
	INSTRUCT admixture	133	-1.97	0.95	0.08	6.6E-04**
	INSTRUCT inbreeding	129	-1.84	0.87	0.04	1.8E-02
	STRUCTURAMA	106	-3.46	0.96	0.27	6.4E-09***
	Neighbor-joining	473	-3.13	0.84	0.24	1.1E-30***
	PCOMC	63	-3.19	0.84	8.7E-03	0.22
Error						
	INSTRUCT no admixture	190	0.41	0.90	0.01	8.7E-02
	INSTRUCT admixture	134	1.19	0.84	0.02	3.8E-02
	INSTRUCT inbreeding	130	1.36	0.91	9.5E-03	0.14
	STRUCTURAMA	128	1.05	0.79	0.11	8.7E-05***
	Neighbor-joining	473	-1.86	0.96	0.13	1.5E-16***
	PCOMC	185	-1.67	0.95	0.04	6.2E-03*

[a] Slope with Y-intercept fixed at 1.

[b] Y intercept for natural regression.

[c] *significant at Holm-Bonferroni corrected α=0.05; ** α=0.01; *** α=0.001.

Table 2. Multiple regression analysis and linear model likelihoods.

Data set	Method	Model	R^2_{adj}	β	SE	p[a]	K	AIC	Δ_i[b]	w_i[c]
Missing										
	INSTRUCT no admixture									
		% missing { $\beta_0+\beta_1(x_1)$ }	0.11	-1.99	0.41	3.2E-6***	2	-301.76	1.01	0.38
		Clumping parameter { $\beta_0+\beta_2(x_2)$ }	0.02	0.02	0.01	0.09	2	-282.69	20.08	2.7E-05
		Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)$ }	0.12			2.9E-6***	3	-302.77	0	0.62
	INSTRUCT admixture									
		% missing { $\beta_0+\beta_1(x_1)$ }	0.08	-1.51	0.48	1.8E-3*	2	-216.92	2.24	0.24
		Clumping parameter { $\beta_0+\beta_2(x_2)$ }	0.04	0.03	0.01	0.04	2	-211.19	7.97	0.01
		Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)$ }	0.1			3.9E-4**	3	-219.16	0	0.74
	INSTRUCT inbreeding									
		% missing { $\beta_0+\beta_1(x_1)$ }	0.04	-0.99	0.46	0.03	2	-226.42	2.24	0.20
		Clumping parameter { $\beta_0+\beta_2(x_2)$ }	0.03	0.02	0.01	0.04	2	-226.05	2.61	0.17
		Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)$ }	0.06			7.6E-3	3	-228.66	0	0.63
	STRUCTURAMA									
		% missing { $\beta_0+\beta_1(x_1)$ }	0.27	-3.30	0.51	2.6E-9***	2	-184.27	2.80	0.20
		Clumping parameter { $\beta_0+\beta_2(x_2)$ }	0.02	-0.03	0.01	0.03	2	-152.4	34.67	2.4E-08
		Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)$ }	0.30			5.0E-9***	3	-187.07	0	0.80
	Neighbor-joining									
		% missing { $\beta_0+\beta_1(x_1)$ }	0.24	-2.2	0.18	1.4E-30***	2	-1066.74	14.94	0.00
		Clumping parameter { $\beta_0+\beta_2(x_2)$ }	0.04	0.02	4.7E-3	4.1E-5***	2	-950.623	131.06	3.5E-29
		Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)$ }	0.27			3.7E-33***	3	-1081.68	0	1.00
	PCOMC									
		% missing { $\beta_0+\beta_1(x_1)$ }	8.7E-3	-1.77	1.26	0.17	2	-98.52	0	0.50
		Clumping parameter { $\beta_0+\beta_2(x_2)$ }	-0.01	-0.01	0.02	0.43	2	-97.15	1.37	0.25
		Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)$ }	2.8E-3			0.34	3	-97.19	1.33	0.25
Error										
	INSTRUCT no admixture									
		Null percent { $\beta_0+\beta_1(x_1)$ }	-4.5E-3	0.14	0.86	0.87	2	-208.04	5.57	0.05
		Drop out percent { $\beta_0+\beta_2(x_2)$ }	0.02	2.07	0.88	0.02	2	-213.61	0	0.78
		Stutter percent { $\beta_0+\beta_3(x_3)$ }	-3.4E-3	-1.26	2.51	0.61	2	-208.25	5.36	0.05
		Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)$ }	0.16			0.12	4	-209.92	3.69	0.12

INSTRUCT admixture									
	Null percent { $\beta_0+\beta_1(x_1)$ }	-2.3E-3	0.99	1.49	0.51	2	-51.1	6.06	0.03
	Drop out percent { $\beta_0+\beta_2(x_2)$ }	0.04	3.90	1.44	7.6E-3	2	-57.16	0	0.59
	Stutter percent { $\beta_0+\beta_3(x_3)$ }	5.3E-3	-6.45	4.49	0.15	2	-52.12	5.04	0.05
	Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)$ }	0.05			0.03	4	-55.98	1.18	0.33
INSTRUCT inbreeding									
	Null percent { $\beta_0+\beta_1(x_1)$ }	-7.6E-3	-0.71	1.82	0.70	2	0.24	11.09	3.1E-03
	Drop out percent { $\beta_0+\beta_2(x_2)$ }	0.05	5.35	1.68	1.8E-3*	2	-7.91	2.94	0.18
	Stutter percent { $\beta_0+\beta_3(x_3)$ }	0.03	-15.29	5.82	9.7E-3	2	-4.54	6.31	0.03
	Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)$ }	0.09			2.1E-3*	4	-10.85	0	0.78
STRUCTURAMA									
	Null percent { $\beta_0+\beta_1(x_1)$ }	1.0E-3	1.26	0.83	0.13	2	-200.45	15.10	3.5E-04
	Drop out percent { $\beta_0+\beta_2(x_2)$ }	0.10	3.10	0.80	1.8E-4**	2	-214.15	1.40	0.33
	Stutter percent { $\beta_0+\beta_3(x_3)$ }	0.02	5.12	2.60	0.05	2	-202.97	12.58	1.2E-03
	Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)$ }	0.13			2.0E-4**	4	-215.55	0	0.67
Neighbor-joining									
	Null percent { $\beta_0+\beta_1(x_1)$ }	0.03	-1.31	0.29	1.2E-5***	2	-951.524	76.95	2.0E-17
	Drop out percent { $\beta_0+\beta_2(x_2)$ }	0.08	-1.67	0.26	5.7E-10***	2	-975.431	53.04	3.0E-12
	Stutter percent { $\beta_0+\beta_3(x_3)$ }	0.07	-6.69	0.99	4.2E-11***	2	-971.313	57.16	3.9E-13
	Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)$ }	0.18			1.2E-20***	4	-1028.47	0	1.00
PCOMC									
	Null percent { $\beta_0+\beta_1(x_1)$ }	2.0E-3	-1.19	0.74	0.11	2	-274.075	5.05	0.05
	Drop out percent { $\beta_0+\beta_2(x_2)$ }	0.02	-1.49	0.66	0.03	2	-277.417	1.71	0.26
	Stutter percent { $\beta_0+\beta_3(x_3)$ }	8.5E-3	-5.15	2.45	0.04	2	-275.274	3.85	0.09
	Combined { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)$ }	0.04			0.02	4	-279.128	0	0.61

[a] *Significant at Holm-Bonferroni corrected $\alpha=0.05$; ** $\alpha=0.01$; *** $\alpha=0.001$.

[b] Rescaled Akaike information criterion (AIC).

[c] Model probability.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Properties of simulated data sets. a) Frequency of data sets with differing levels of population subdivision. b) The amount of missing data introduced into data sets was unrelated to level of population subdivision. c) The frequency of erroneous genotypes was negatively correlated with population subdivision.

Supplementary Figure 2. The inbreeding-like effect of SSR genotyping error. a) The level of apparent inbreeding increases with genotyping error. b, c) The increase in F is driven by a decrease in observed heterozygosity rather than an increase in expected heterozygosity, which decreased significantly, but slightly. d) Error increases the apparent level of population subdivision, producing a tendency for more accurate assignment of genotypes to clusters as the amount of error increases.

Supplementary Table 1. Spearman's rank order correlation of clustering accuracy, measured with the partition distance ratio.

Data set	Method	n_{obs}[a]	ρ[a,b]	p[a,c]	n_{obs}[d]	ρ[d,b]	p[d,c]
Missing							
	INSTRUCT no admixture	437	-0.22	2.3E-06***	188	0.68	1.4E-26***
	INSTRUCT admixture	356	-0.23	1.1E-05***	133	0.63	2.6E-16***
	INSTRUCT inbreeding	357	-0.11	0.03	125	0.45	1.4E-07***
	STRUCTURAMA	115	0.12	0.19	100	0.05	0.61
Error							
	INSTRUCT no admixture	464	0.30	2.4E-11***	190	0.51	1.0E-13***
	INSTRUCT admixture	393	0.49	4.5E-25***	134	0.41	7.0E-07***
	INSTRUCT inbreeding	382	0.53	6.2E-29***	130	0.46	2.7E-08***
	STRUCTURAMA	141	0.18	0.03	115	0.57	4.4E-11***

[a] Spearman's rank order correlation between percent data set modification and the partition distance ratio metric.

[b] Spearman's coefficient, rho.

[c] *Significant at Holm-Bonferroni corrected α=0.05; ** α=0.01; *** α=0.001.

[d] Spearman's rank order correlation between the performance ratio and the partition distance ratio.

Supplementary Table 2. Multiple regression and likelihood analysis of six competing model effects explaining accuracy of model-based clustering methods subject to erroneous data, measured with the performance ratio.

Method	Model	R^2_{adj}	β	SE	p[c]	K	AIC	Δ_i[d]	w_i[e]
STRUCTURAMA									
	% error { $\beta_0+\beta_1(x_1)$ }	0.11	3.16	0.95	1.2E-03*	2	-215.01	32.01	4.4E-08
	Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	0.03	0.09	0.07	0.20	2	-204.35	42.67	2.1E-10
	Δ K { $\beta_0+\beta_3(x_3)$ }	0.24	0.10	0.02	2.6E-08***	2	-235.96	11.06	1.6E-03
	Δ G'st { $\beta_0+\beta_4(x_4)$ }	-7.8E-03	-0.86	4.14	0.84	2	-199.33	47.69	1.7E-11
	Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	1.7E-03	-3.48	1.70	0.04	2	-200.54	46.48	3.2E-11
	Δ F { $\beta_0+\beta_6(x_6)$ }	2.9E-03	-1.16	0.72	0.11	2	-200.7	46.32	3.5E-11
	Best 2 { $\beta_0+\beta_3(x_3)+\beta_5(x_5)$ }	0.28			6.4E-10***	3	-240.68	6.34	1.7E-02
	Best 3 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_5(x_5)$ }	0.30			2.0E-10***	4	-244.76	2.26	0.13
	Best 4 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)$ }	0.32			1.4E-10***	5	-247.02	0	0.40
	Best 5 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)$ }	0.32			2.8E-10***	6	-246.68	0.34	0.33
	All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.32			1.1E-09***	7	-244.73	2.29	0.13
INSTRUCT, no admixture									
	% error { $\beta_0+\beta_1(x_1)$ }	0.01	2.62	1.12	0.02	2	-210.86	12.65	5.5E-04
	Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	0.04	-1.2E-03	0.09	0.99	2	-216.59	6.92	9.6E-03
	Δ K { $\beta_0+\beta_3(x_3)$ }	-3.6E-03	1.0E-03	0.00	0.64	2	-208.21	15.30	1.4E-04
	Δ G'st { $\beta_0+\beta_4(x_4)$ }	0.03	-2.47	2.00	0.22	2	-214.90	8.61	4.1E-03
	Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	0.02	-5.02	3.62	0.17	2	-212.23	11.28	1.1E-03
	Δ F { $\beta_0+\beta_6(x_6)$ }	0.02	-1.03	0.78	0.19	2	-212.64	10.87	1.3E-03
	Best 2 { $\beta_0+\beta_1(x_1)+\beta_4(x_4)$ }	0.06			6.2E-04*	3	-220.91	2.60	0.08
	Best 3 { $\beta_0+\beta_1(x_1)+\beta_4(x_4)+\beta_6(x_6)$ }	0.08			2.9E-04*	4	-223.11	0.40	0.25
	Best 4 { $\beta_0+\beta_1(x_1)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.09			2.9E-04*	5	-223.51	0	0.30

Best 5 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.09			4.0E-04*	6	-223.08	0.44	0.25
All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.09			8.7E-04*	7	-221.30	2.21	0.10
INSTRUCT, admixture								
% error { $\beta_0+\beta_1(x_1)$ }	0.02	5.18	2.71	0.06	2	-54.77	0.22	0.19
Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	0.02	0.04	0.16	0.78	2	-53.84	1.15	0.12
Δ K { $\beta_0+\beta_3(x_3)$ }	-7.6E-03	-1.9E-04	7.7E-03	0.98	2	-50.41	4.59	0.02
Δ G'st { $\beta_0+\beta_4(x_4)$ }	-3.0E-03	8.87	7.02	0.21	2	-51.02	3.97	0.03
Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	-4.2E-03	-6.48	4.27	0.13	2	-50.85	4.14	0.03
Δ F { $\beta_0+\beta_6(x_6)$ }	-4.6E-03	-3.32	1.82	0.07	2	-50.80	4.19	0.03
Best 2 { $\beta_0+\beta_1(x_1)+\beta_6(x_6)$ }	0.03			0.04	3	-54.99	0	0.21
Best 3 { $\beta_0+\beta_1(x_1)+\beta_5(x_5)+\beta_6(x_6)$ }	0.04			0.05	4	-54.43	0.56	0.16
Best 4 { $\beta_0+\beta_1(x_1)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.04			0.05	5	-54.28	0.71	0.15
Best 5 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.04			0.09	6	-52.36	2.63	0.06
All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.03			0.14	7	-50.36	4.63	0.02
INSTRUCT, inbreeding								
% error { $\beta_0+\beta_1(x_1)$ }	9.5E-03	3.44	3.58	0.34	2	-1.99	0.44	0.16
Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	9.2E-03	-0.05	0.21	0.80	2	-1.95	0.48	0.15
Δ K { $\beta_0+\beta_3(x_3)$ }	-3.0E-03	0.01	0.01	0.33	2	-0.37	2.06	0.07
Δ G'st { $\beta_0+\beta_4(x_4)$ }	-7.3E-03	2.68	8.98	0.77	2	0.19	2.62	0.05
Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	2.5E-03	-8.53	5.50	0.12	2	-1.07	1.36	0.10
Δ F { $\beta_0+\beta_6(x_6)$ }	-5.1E-03	-1.58	2.41	0.51	2	-0.10	2.33	0.06
Best 2 { $\beta_0+\beta_1(x_1)+\beta_5(x_5)$ }	0.02			0.10	3	-2.43	0	0.19
Best 3 { $\beta_0+\beta_1(x_1)+\beta_5(x_5)+\beta_6(x_6)$ }	0.02			0.14	4	-1.38	1.05	0.11
Best 4 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_5(x_5)+\beta_6(x_6)$ }	0.02			0.17	5	-0.34	2.09	0.07
Best 5 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.01			0.27	6	1.59	4.02	0.03
All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	4.2E-03			0.37	7	3.52	5.95	9.9E-03

[a] Mean number of alleles per locus

[b] Mean private allele frequency

[c] *significant at Holm-Bonferroni corrected $\alpha=0.05$; ** $\alpha=0.01$; *** $\alpha=0.001$.

[d] Rescaled AIC.

[e] Model probability.

Supplementary Table 3. Multiple regression and likelihood analysis of six competing model effects explaining accuracy of model-based clustering methods subject to erroneous data, measured with the partition distance ratio metric.

Method	Model	R^2_{adj}	β	SE	p[c]	K	AIC	Δ_i[d]	w_i[e]
STRUCTURAMA									
	% error { $\beta_0+\beta_1(x_1)$ }	0.04	-0.78	0.72	0.28	2	-245.03	83.80	2.4E-19
	Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	6.8E-03	-0.01	0.05	0.84	2	-238.05	90.78	7.3E-21
	Δ K { $\beta_0+\beta_3(x_3)$ }	0.43	0.12	0.01	1.2E-15***	2	-318.23	10.60	1.9E-03
	Δ G'st { $\beta_0+\beta_4(x_4)$ }	0.05	-1.94	3.15	0.54	2	-246.64	82.19	5.4E-19
	Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	0.01	-0.13	1.37	0.92	2	-240.54	88.29	2.6E-20
	Δ F { $\beta_0+\beta_6(x_6)$ }	0.15	1.51	0.56	7.6E-03	2	-261.90	66.93	1.1E-15
	Best 2 { $\beta_0+\beta_3(x_3)+\beta_6(x_6)$ }	0.48			1.9E-20***	3	-328.83	0	0.38
	Best 3 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_6(x_6)$ }	0.48			6.8E-20***	4	-328.75	0.07	0.37
	Best 4 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_4(x_4)+\beta_6(x_6)$ }	0.48			4.0E-19***	5	-327.20	1.63	0.17
	Best 5 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_6(x_6)$ }	0.47			2.4E-18***	6	-325.24	3.59	0.06
	All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.47			1.3E-17***	7	-323.25	5.58	0.02
INSTRUCT, no admixture									
	% error { $\beta_0+\beta_1(x_1)$ }	0.06	1.32	0.52	0.01	2	-597.36	35.52	8.0E-09
	Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	-1.5E-03	0.10	0.05	0.07	2	-567.72	65.17	2.9E-15
	Δ K { $\beta_0+\beta_3(x_3)$ }	-9.0E-04	-1.3E-03	1.4E-03	0.36	2	-567.98	64.91	3.3E-15
	Δ G'st { $\beta_0+\beta_4(x_4)$ }	0.10	2.70	1.45	0.06	2	-616.39	16.50	1.1E-04
	Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	7.7E-03	-1.01	1.21	0.40	2	-571.96	60.93	2.4E-14

Δ F { $\beta_0+\beta_6(x_6)$ }	0.11	0.67	0.24	0.00	2	-621.13	11.76	1.2E-03
Best 2 { $\beta_0+\beta_3(x_3)+\beta_4(x_4)$ }	0.13			1.8E-14***	3	-629.15	3.73	6.4E-02
Best 3 { $\beta_0+\beta_3(x_3)+\beta_4(x_4)+\beta_6(x_6)$ }	0.13			1.8E-14***	4	-630.95	1.93	0.16
Best 4 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_4(x_4)+\beta_6(x_6)$ }	0.14			1.5E-14***	5	-632.89	0.00	0.41
Best 5 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_6(x_6)$ }	0.14			4.5E-14***	6	-631.80	1.084	0.24
All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.14			1.4E-13***	7	-630.51	2.38	0.13
INSTRUCT, admixture								
% error { $\beta_0+\beta_1(x_1)$ }	0.12	5.79	1.04	4.7E-08***	2	-45.33	6.06	1.5E-02
Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	8.0E-03	0.24	0.11	0.03	2	0.16	51.54	2.0E-12
Δ K { $\beta_0+\beta_3(x_3)$ }	-1.5E-03	-3.3E-04	5.0E-03	0.95	2	3.85	55.24	3.1E-13
Δ G'st { $\beta_0+\beta_4(x_4)$ }	0.05	6.24	2.93	0.03	2	-16.00	35.39	6.4E-09
Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	-6.0E-04	-3.42	2.35	0.15	2	3.50	54.88	3.7E-13
Δ F { $\beta_0+\beta_6(x_6)$ }	4.7E-03	-1.21	0.49	0.01	2	1.43	52.81	1.1E-12
Best 2 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)$ }	0.13			1.8E-12***	3	-48.28	3.11	0.07
Best 3 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)$ }	0.13			1.3E-12***	4	-50.62	0.7663	0.21
Best 4 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)$ }	0.14			1.9E-12***	5	-51.22	0.16	0.29
Best 5 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)$ }	0.14			3.0E-12***	6	-51.39	0.00	0.31
All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.14			1.2E-11***	7	-49.39	2.00	0.11
INSTRUCT, inbreeding								
% error { $\beta_0+\beta_1(x_1)$ }	0.12	6.89	1.06	3.2E-10***	2	-19.01	15.98	1.8E-04
Δ APL[a] { $\beta_0+\beta_2(x_2)$ }	4.6E-03	0.34	0.11	2.3E-03*	2	26.42	61.40	2.5E-14
Δ K { $\beta_0+\beta_3(x_3)$ }	-2.3E-03	-1.9E-04	5.5E-03	0.97	2	29.02	64.01	6.7E-15
Δ G'st { $\beta_0+\beta_4(x_4)$ }	0.03	6.35	3.01	0.04	2	15.03	50.02	7.3E-12
Δ p(1)[b] { $\beta_0+\beta_5(x_5)$ }	-2.5E-03	-5.82	2.41	0.02	2	29.08	64.07	6.5E-15
Δ F { $\beta_0+\beta_6(x_6)$ }	-5.0E-04	-1.61	0.52	2.2E-03*	2	28.32	63.31	9.5E-15
Best 2 { $\beta_0+\beta_3(x_3)+\beta_6(x_6)$ }	0.13			1.4E-12***	3	-23.86	11.13	0.00

Best 3 { $\beta_0+\beta_1(x_1)+\beta_3(x_3)+\beta_6(x_6)$ }	0.15	6.1E-14***	4	-32.04	2.95	0.12
Best 4 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_6(x_6)$ }	0.15	1.0E-13***	5	-32.47	2.52	0.15
Best 5 { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_5(x_5)+\beta_6(x_6)$ }	0.16	5.6E-14***	6	-34.99	0	0.53
All { $\beta_0+\beta_1(x_1)+\beta_2(x_2)+\beta_3(x_3)+\beta_4(x_4)+\beta_5(x_5)+\beta_6(x_6)$ }	0.16	2.3E-13***	7	-32.99	2.00	0.20

[a]Mean number of alleles per locus

[b]Mean private allele frequency

[c]*significant at Holm-Bonferroni corrected α=0.05; ** α=0.01; *** α=0.001.

[d] Rescaled AIC.

[e] Model probability.

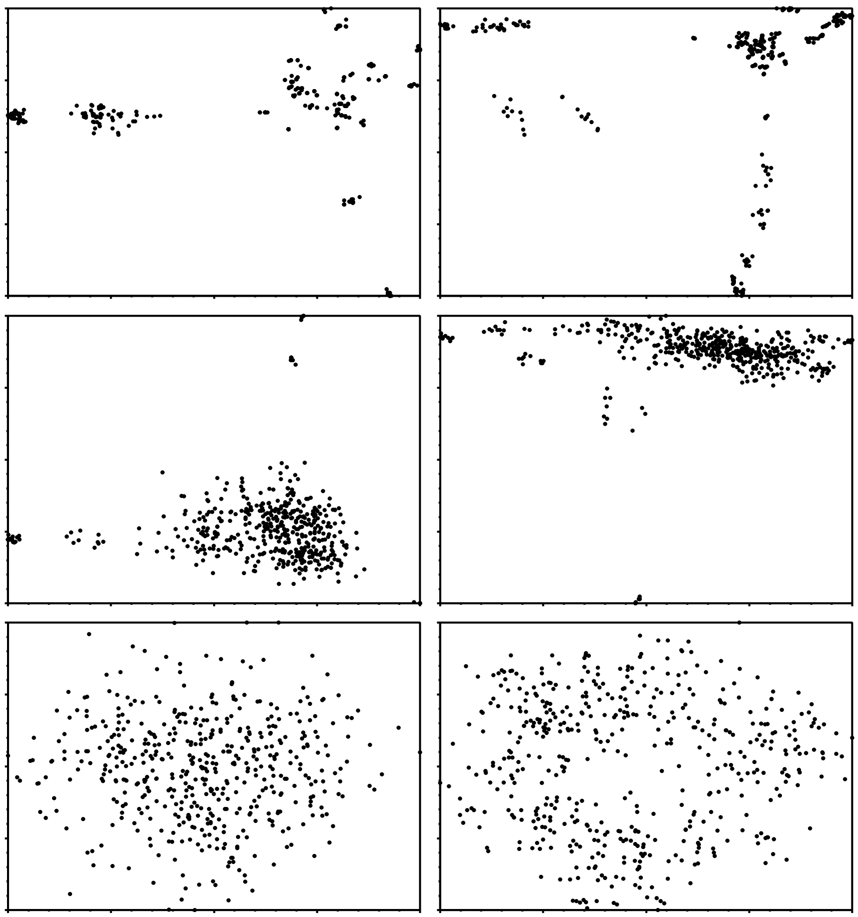

Figure 1.

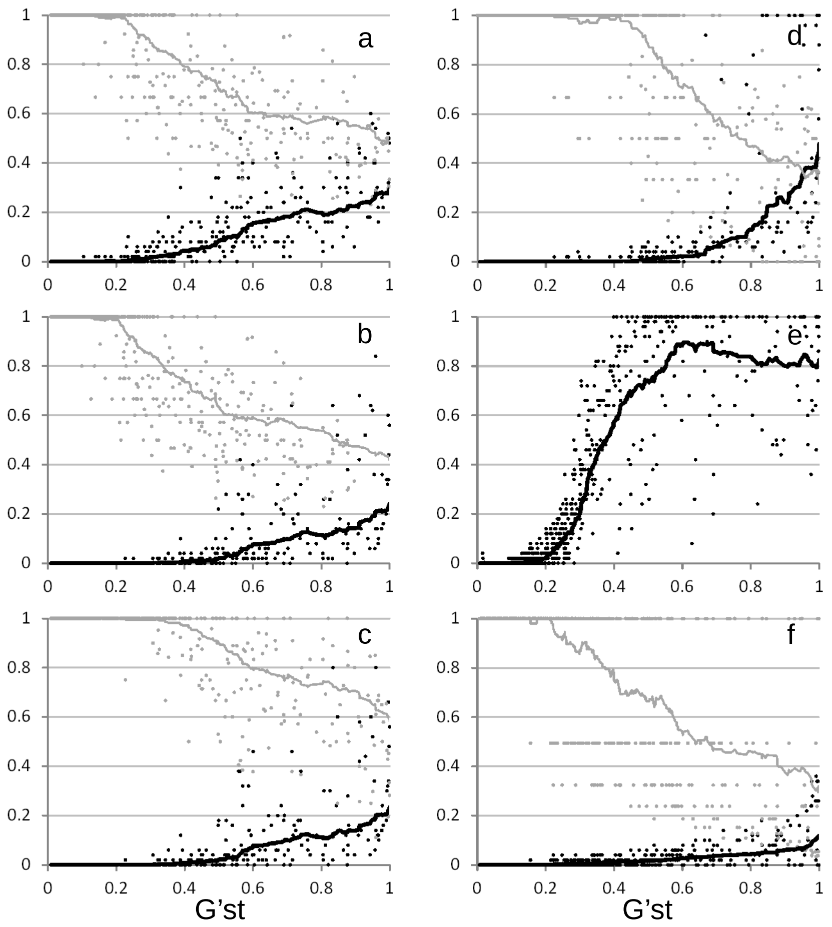

Figure 2.

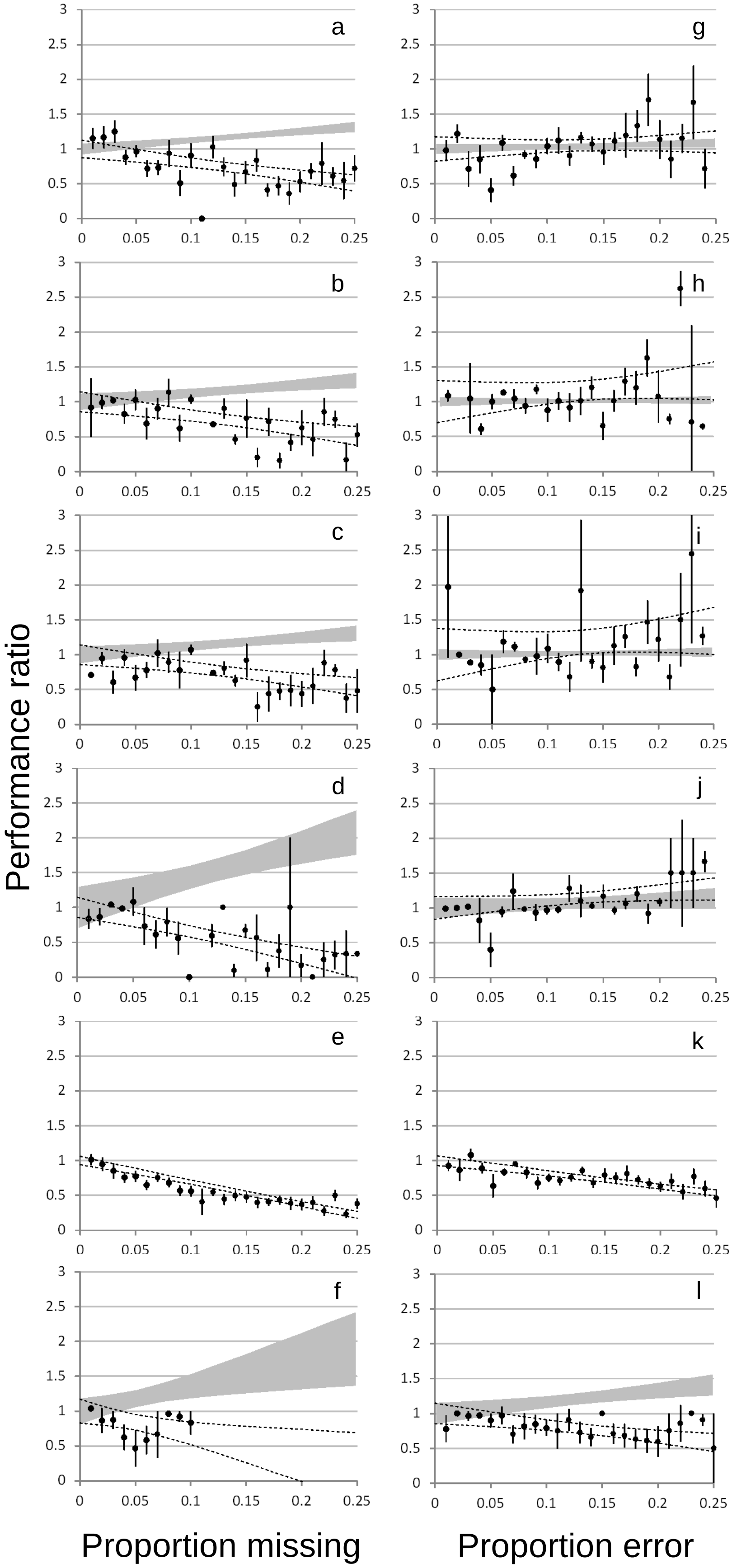
Performance ratio
a
b
c
d
e
f
g
h
i
j
k
l
0
0.5
1
1.5
2
2.5
3
0
0.05
0.1
0.15
0.2
0.25
Proportion missing
Proportion error

Figure 3.

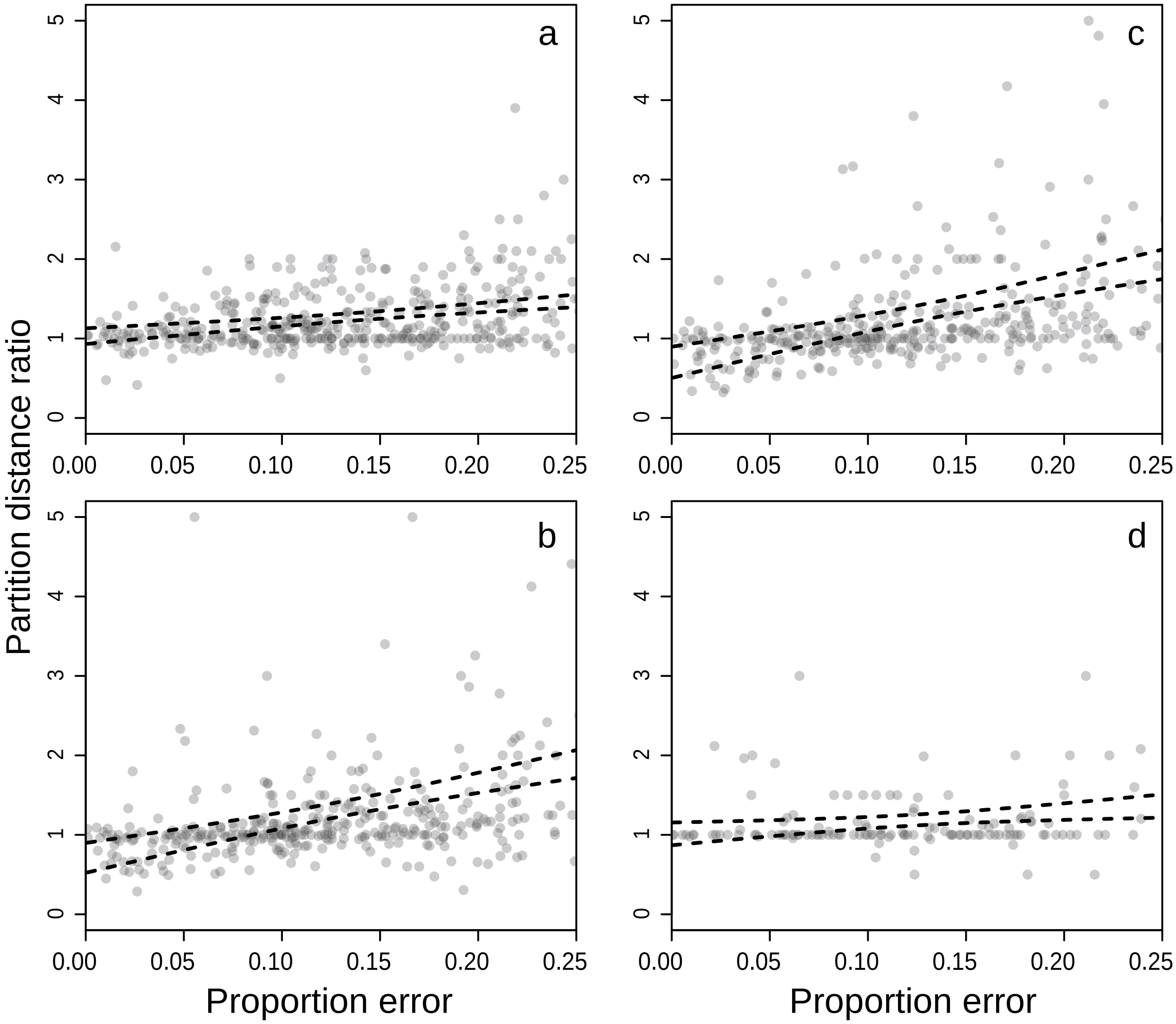
a
b
c
d
Partition distance ratio
0
1
2
3
4
5
0.00
0.05
0.10
0.15
0.20
0.25
Proportion error
Proportion error

Figure 4.

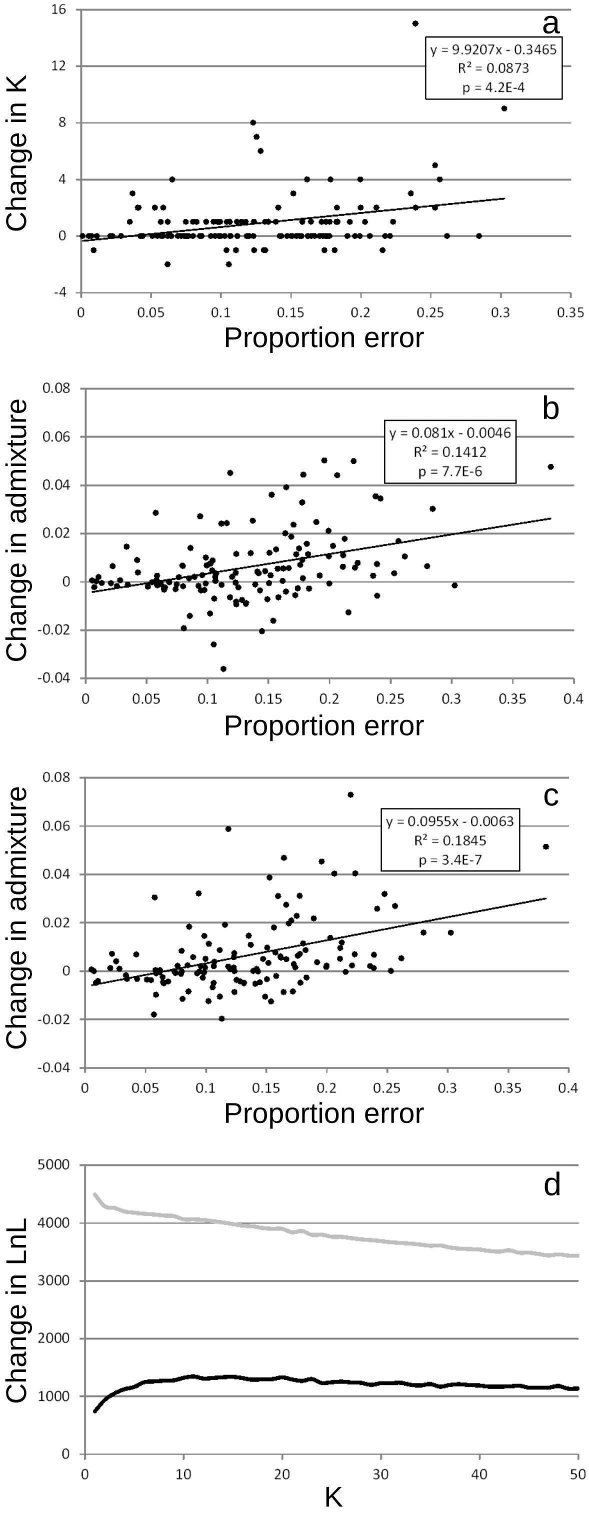
a
y = 9.9207x - 0.3465
R² = 0.0873
p = 4.2E-4
Change in K
16
12
8
4
0
-4
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Proportion error
b
y = 0.081x - 0.0046
R² = 0.1412
p = 7.7E-6
Change in admixture
0.08
0.06
0.04
0.02
0
-0.02
-0.04
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Proportion error
c
y = 0.0955x - 0.0063
R² = 0.1845
p = 3.4E-7
Change in admixture
0.08
0.06
0.04
0.02
0
-0.02
-0.04
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
Proportion error
d
Change in LnL
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
K

Figure 5.