The Effects of Gene Flow from Unsampled 'Ghost' Populations on the Estimation of Evolutionary History under the Isolation with Migration Model

Melissa Lynch*[1] and Arun Sethuraman†[1]

[1]Department of Biological Sciences, California State University San Marcos

August 12, 2019

Abstract

Unsampled or extinct 'ghost' populations leave signatures on the genomes of individuals from extant, sampled populations, especially if they have exchanged genes with them over evolutionary time. This gene flow from 'ghost' populations can introduce biases when estimating evolutionary history from genomic data, often leading to data misinterpretation and ambiguous results. To assess the extent of this bias, we perform an extensive simulation study under varying degrees of gene flow to and from 'ghost' populations under the Isolation with Migration (IM) model. Estimates of popular summary statistics like Watterson's θ, π, and F_{ST}, and evolutionary demographic history (estimated as effective population sizes, divergence times, and migration rates) using the `IMa2p` software clearly indicate that we a) under-estimate divergence times between sampled populations, (b) over-estimate effective population sizes of sampled populations, and (c) under-estimate migration rates between sampled populations, with increased gene flow from the unsampled 'ghost' population. Similarly, summary statistics like F_{ST} and π are also affected depending on the amount of gene flow from the unsampled 'ghost'.

*lynch026@cougars.csusm.edu
†asethuraman@csusm.edu

Introduction

Studies that apply population genetics methods to infer evolutionary history, including those in the fields of conservation, agriculture, evolution, and anthropology often begin with a sampling strategy. As a rule of thumb, we would expect that the more extensive the sampling of individuals across their geographical range (or according to the biological question at hand), the better the resolution of population genetic analyses. However, in any such study, genomic data collected harbors signatures of evolutionary processes of drift, selection, and gene flow involving both sampled and unsampled 'ghost' populations (Beerli, 2004). Such 'ghost' populations are ubiquitous, and have been increasingly discovered across numerous species complexes. For example, several studies of the evolutionary history of African Hunter-Gatherers (Lachance et al., 2012; Hey et al., 2018; Durvasula & Sankararaman, 2019) identify significant gene flow from an unsampled archaic population (diverged from modern humans around the same time as Neanderthals) into multiple modern Hunter-Gatherer lineages. Similar studies in black rats (Konečnỳ et al., 2013), Brachypodium sylvaticum bunchgrass (Rosenthal, Ramakrishnan, & Cruzan, 2008), bonobos (Kuhlwilm, Han, Sousa, Excoffier, & Marques-Bonet, 2019) and modern humans (summarized in (Nielsen et al., 2017)) also describe the occurrence of unsampled 'ghost' populations in their respective systems. The degree of such unsampled genomic variation in sampled genomic data is a result of either (a) incomplete population sampling on the part of the researcher, or (b) population extinction or decline, and (c) the amount of gene flow from such unsampled, or extinct populations over evolutionary time.

Importantly, not accounting for such unsampled 'ghost' population gene flow into extant, sampled populations can lead to erroneous estimates while using summary statistics, or phylogenetic/mutation model-based methods, and thus, erroneous conclusions about the evolutionary history of the species. Interestingly, the studies that account for unsampled 'ghosts' are still part of a minority in the landscape of population genomics research and publications.

For example, under an evolutionary scenario where two populations of a species don't directly exchange genes between each other, but exchange genes bidirectionally with an unsampled 'ghost', one would expect that measures of population differentiation (F_{ST}) between sampled extant populations would be very low, compared to what we would expect under a model where the gene flow from the 'ghost' is unaccounted for. Similarly, other summary statistics, such as the allele frequency distribution (Tajima's D), heterozygosity, nucleotide diversity (π), effective population size (N_e), and estimates of time to most recent common ancestor (T_{MRCA}) can all be biased depending on the degree of gene flow from unsampled 'ghost' populations.

Model-based methods to estimate population structure and gene flow, including Bayesian or maximum

likelihood approaches, as implemented in software such as `structure` (Pritchard, Stephens, & Donnelly, 2000), `ADMIXTURE` (Alexander, Novembre, & Lange, 2009), `MIGRATE-n` (Beerli & Felsenstein, 2001), and `IMa2p` (Sethuraman & Hey, 2016) are also not infallible to the effects of 'ghost' gene flow, in that identical estimates of phylogenetic and mutational history may be achieved through a number of unique demographic histories (Lawson, Van Dorp, & Falush, 2018). For example, Lawson et al., 2018 describe three scenarios of evolutionary history - one of recent admixture from four divergent populations, one of gene flow from a 'ghost' population, and one of recent bottlenecks in sampled, extant populations, which all estimate the same population structure and admixture proportions while using the programs `structure` or `ADMIXTURE`. Correspondingly, not accounting for 'ghost' gene flow has been known to bias estimates of migration rates (Hey et al., 2018), divergence times (Lachance et al., 2012) while utilizing other model-based estimators of evolutionary history.

Biases in estimates of effective population sizes and migration rates in the presence of the unsampled 'ghost' populations was previously investigated by Beerli (2004) and Slatkin (2004). Using an 'n'-island model (Slarkin, 1985), where each of 'n' populations of constant size can exchange genes at constant rates with the 'n-1' other populations, Beerli estimated the effect of the magnitude and direction of migration in the presence of a 'ghost' population. He simulated three populations of identical effective population size and per generation mutation rates under this 'n' island model, such that each set of populations either had high or low magnitude of unidirectional or bidirectional migration with the 'ghost' population (Figure 1). Beerli also estimated the effect of the number of 'ghost' populations by sampling two populations out of a larger, varied set of unsampled populations. Each of these datasets was then analyzed using the `MIGRATE-n` software. Beerli's study identified several effects of the 'ghost' population gene flow, including (1) a higher migration rate, both unidirectional and bidirectional to and from the 'ghost' population, led to an overestimation of migration rates between the sampled populations, (2) increasing the number of 'ghost' populations increased the bias in estimates of population sizes, but had little effect on the migration rate estimates, and (3) increasing the number of sampled loci did not improve or affect the estimation of migration rates.

Here we extend Beerli's study to a more complex model of evolutionary history, popularly termed the Isolation with Migration (IM) model (Hey & Nielsen, 2004, 2007; Nielsen & Wakeley, 2001). This class of models is widely used as a model of divergence with gene flow, where two sampled populations or subpopulations have diverged from the ancestor, and maintain gene flow via exchange of migrants since divergence. Numerous tools have been developed to estimate evolutionary history under the IM model, including the IM suite (`IMa2p`, and `IMa3` (Sethuraman & Hey, 2016; Hey et al., 2018), `MIST` (Chung & Hey, 2017), and `dadi` (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009). The IM suite of tools are

genealogy samplers that use a Metropolis Coupled Markov Chain Monte Carlo (MCMCMC) to explore the parameter space of evolutionary demographic parameters, propose updates to parameters and genealogies, and examine the posterior density distribution of the set of estimable parameters given genomic data from two or more sampled populations. These programs estimate divergence times, migration rates, and effective population sizes of all populations included in the model. Additionally, these programs also have the ability to allow for the presence of a single 'ghost' population, where a population is added to the model that is assumed to be the outgroup to all sampled populations (this aligns with Beerli's (2001) finding that the addition of more than one 'ghost' population doesn't significantly affect estimates of migration rates), and estimate parameters of the IM model under various phylogenetic models. This allows us to compare and contrast how accounting or not-accounting for the presence of a 'ghost' population could potentially bias estimates of divergence times, effective population sizes, and migration rates between sampled extant populations.

Our study utilizes an extensive set of simulations under the IM model, with (1) varying degrees of unidirectional and bidirectional gene flow from unsampled 'ghost' populations, (2) varying the number of sampled genomic loci, to quantify the biases in (a) popularly utilized summary statistics in population genomics, and (b) estimates of evolutionary history under the IM model.

Methods

Simulations

The `ms` software (Hudson, 2002) was used to generate genomic data under IM model 'versions' of the Beerli (Beerli, 2004) simulations (Figure 2). This software uses coalescent methods to generate genomic data from sampling haplotypes from populations evolving under a variety of models. Briefly, three populations were assumed in all models, A, and B (extant, sampled), and C (unsampled 'ghost'). All populations were assumed to have a constant effective population size (mutation rate per locus per generation scaled population size, $\theta = 4N_e\mu = 0.01$). Populations A and B were assumed to have diverged from their common ancestor (D) at t_{ABD}=0.5 (divergence time is scaled by mutation rate per locus per generation as t/μ). Population C diverged from D at a time t_{CD}=2. Ten individuals were sampled per population. There were five different migration scenarios simulated under this model. Under the first scenario, only populations A and B exchange genes since divergence from D at a rate of $4N_A m_{A\rightarrow B} = 4N_B m_{B\rightarrow A} = 1.0$, which is equivalent to one migrant individual every fourth generation. Neither A nor B exchange any immigrants with the 'ghost' population C. Under the second scenario, A and B exchange individuals at a rate of $4N_A m_{A\rightarrow B} = 4N_B m_{B\rightarrow A} = 1.0$ as in scenario 1, but individuals also emigrate out of the

'ghost' C into A and B at the same rate ($4N_C m_{C\to A} = 4N_C m_{C\to B} = 1.0$). Under the third scenario, the 'ghost' population C exchanges $4N_C m_{C\to A} = 4N_C m_{C\to B} = 10.0$ individuals every fourth generation, while A and B continue to exchange at the same low rate ($4N_C m_{C\to A} = 4N_C m_{C\to B} = 1.0$). Under the fourth scenario, all three populations exchange genes at the same low rate ($4N_e m = 1.0$). Under the last scenario, C exchanges genes bidirectionally at a high rate ($4N_C m_{C\to A} = 4N_A m_{A\to C} = 4N_C m_{C\to B} = 4N_B m_{B\to C} = 10.0$), while A and B exchange migrants at a low rate of $4N_A m_{A\to B} = 4N_B m_{B\to A} = 1.0$. Under all scenarios, ten individuals were sampled from each population, and separate data-sets were constructed with two or five genomic loci. Ten replicate datasets were simulated under each scenario, to construct confidence intervals around estimates.

Summary Statistics

The R PopGenome (Pfeifer, Wittelsbuerger, Ramos-Onsins, & Lercher, 2014) package was used to calculate measures of population differentiation between populations F_{ST}, Tajima's D, nucleotide diversity (π), effective population size (N_e) measured as the Watterson estimator of genetic diversity (θ), and the number of segregating sites (S).

Estimates of evolutionary history

Evolutionary history was then estimated using the `IMa2p` (Sethuraman and Hey, 2014) software under three separate models: (1) a two population model, wherein the simulated genomic data was down-sampled to only include populations A and B, (2) a three population model, wherein all three populations, A, B, and C were included in a population model with A and B sharing a more recent common ancestor, and C as the outgroup, and (3) a two population model (same as model 1), but with the addition of an outgroup 'ghost' population. Prior values for effective population sizes, divergence times, and migration rates were set using estimates of Watterson's θ, and setting the upper bound on the θ estimates to be set to five times the geometric mean of Watterson's θ across all loci, the upper bound on the divergence times were set to two times the geometric mean of Watterson's θ across all loci, and the upper bound on migration rates was set to 2 divided by the geometric mean of Watterson's θ across all loci, according to the recommendations of Hey (2011). All runs were performed using (?) chains distributed across 56 processors, discarding 1×10^6 MCMC iterations as burn-in, followed by sampling 100,000 genealogies. All chains were ensured to be mixing appropriately, swapping genealogies across chains, and that the chains had converged (adjudged by observing the autocorrelations between parameter estimates across iterations, and effective sample size values), prior to sampling genealogies. All sampled genealogies were then used in estimating marginal posterior densities of parameters, and the modes of these marginal distributions and 95% confidence intervals around the modes were computed, and compared to the 'true'

simulated parameters used in `ms` (Hudson, 2002).

Results

Effects on the estimation of divergence times

Divergence times (t, such that true divergence time in years $= \frac{t}{u}$ where u is the mutation rate per locus per generation) between sampled populations were consistently under-estimated with increased migration from the unsampled 'ghost' population, compared to the true value of $t_0 = 5$ across all our simulations. Nonetheless, the 95% confidence intervals of estimates encompassed the true simulated value across a majority of our simulations. However, sampling more genomic loci led to a reduction in the confidence intervals around divergence time estimates. The two population model with high bi-directional 'ghost' migration (Scenario E) was the only model to under-estimate divergence times without the confidence interval including the true value (simulated divergence time), while sampling either 2 or 5 loci. Estimates of divergence times of the common ancestor of the two sampled populations from the 'ghost' population were consistent with the true simulated divergence time of $t_1 = 20$, except for the 3 population model of scenario E. Inclusion of a 'ghost' population in model led to underestimation in no migration model (A), but led to seemingly better estimates with smaller confidence intervals in most other models.

Effects on the estimation of effective population sizes

While estimating scaled effective population sizes ($\theta = 4N_e u$, where N_e is the effective population size, and u is the mutation rate per locus per generation) of sampled populations, the inclusion of a 'ghost' population resulted in estimates similar to having sampled all three populations, while the two population models had wider confidence intervals across all estimates. Nonetheless, greater migration rates (under scenarios C and E) resulted in over-estimation of effective population sizes in both sampled populations.

The level of gene flow from the 'ghost' population greatly affects estimates of effective population size of the common ancestor of the two sampled populations. Estimates from the two population model have smaller confidence intervals when there is no or low migration with ghost (scenarios A, B), but in scenarios where migration is high (C and E), the effective population size is consistently overestimated to be $\approx 2 - 10$ times the 'true' value ($\theta = 10$).

When estimating the effective population sizes of the common ancestor of the two sampled populations

and the 'ghost' population, the two population model with inclusion of a 'ghost' performed as well as the three population model in scenarios of low or no migration (A and B). In scenarios of high migration however (C, E), the three population model obtained better estimates, while the model with the inclusion of a 'ghost' mostly underestimated θ_4. Increasing the number of loci did not seem to affect estimates of effective population sizes across all our scenarios.

Effects on the estimation of migration rates between sampled populations

Migration rates (scaled as $m = M/u$, where M is the population migration rate per locus per generation, and u is the mutation rate per generation per locus) between sampled populations were relatively accurately estimated between sampled populations under both 2 and 5 locus sampling schemes. The inclusion of a 'ghost' population did not change affect these estimates from two and three population models. Scenarios with highest true migration rates, regardless of direction (C and E) have consistently greater estimates of migration rates between sampled populations for all models. Estimates of migration rates under scenarios with lower true migration rates (B and D) are closer to the true migration rates, and the scenario with no migration to or from 'ghost' (A) has greatest under-estimate in all models. Migration rate estimates directly to and from the 'ghost' population were greatly under-estimated when 'true' migration was high, with estimates approximating ten times below the 'true' value. Additionally, increasing the number of sampled loci did not improve estimates of migration rates.

Effects on the estimation of summary statistics

Estimates of Tajima's D (Figure 2A) did not vary with increased migration among populations. The number of segregating sites (S), Watterson's θ and nucleotide diversity (pi) all show similar patterns across scenarios (Figure 2B, D, E), such that populations with increased migration from the 'ghost' show greater estimates of all diversity statistics. Estimates of population differentiation between populations (F_{ST}) showed different patterns depending on sampling strategy. A 3 population model reflected higher differentiation among populations where there was no migration between 'ghost', with differentiation among populations decreasing as migration rates increase (scenarios B-E) (Figure 2C). A 2 population model does not show any real differences in differentiation between sampled populations, with F_{ST} estimates higher in scenarios of unidirectional gene flow from 'ghost' (scenarios B, C) and lower F_{ST} in the scenario with highest bidirectional gene flow (scenario E) (Figure 2C). Increasing the number

of sampled loci did not change estimates of summary statistics, although the differences between high migration scenarios and low migration scenarios were minimized when sampling more loci.

Discussion

Estimates of evolutionary history and population genetics of species are ubiquitously affected by gene flow from unsampled 'ghost' populations. The degree to which differing rates of 'ghost' gene flow affect (a) popularly used summary statistics, and (b) model-based estimates of evolutionary history are scarcely quantified. Here we perform extensive simulations of genomic data under the Isolation with Migration (IM) model, under a variety of evolutionary scenarios (with differing degrees of gene flow from unsampled 'ghost' populations), to quantify these potential biases. Our simulations show that the degree of differentiation between sampled populations (measured as F_{ST} is always under-estimated with increased gene flow from unsampled 'ghosts' (Fig. 2C, 3C). Correspondingly, the degree of genomic diversity (measured as Watterson's θ, nucleotide diversity, and π) are always over-estimated with increased gene flow from unsampled 'ghosts'. These patterns can potentially be minimized by sampling more genomic loci, but regardless, not accounting for the presence of gene flow from an unsampled 'ghost' will always lead to erroneous conclusions about the degree of genomic diversity in sampled extant populations. For instance, consider a scenario where two species are geographically separated (and genomically disparate owing to allopatric speciation), and we sample individuals from these two species to estimate the degree of differentiation between them. In the event that there happens to be greater degree of gene flow from an unsampled 'ghost' population, the differentiation between the sampled extant populations would always be lower, which could lead the investigators to conclude that the two species are not genetically disjunct, and that they could be directly exchanging migrants between them. This would be an erroneous conclusion, since both the extant populations had in fact exchanged genes in the past with a now extinct unsampled 'ghost'.

Similarly, not accounting for an unsampled 'ghost' in a model-based estimation of evolutionary history can also lead to significant errors in our deductions of the true evolutionary history. Our simulations consistently show that we (a) under-estimate divergence times between sampled populations, (b) over-estimate effective population sizes of sampled populations, and (c) under-estimate migration rates between sampled populations, all with increased gene flow from the unsampled 'ghost' population. These findings are in line with those of (Beerli, 2004), showing that not accounting for an unsampled 'ghost' can skew estimates of evolutionary history. A few findings were perplexing - for instance, estimates of unidirectional migration rate from one sampled population into the unsampled 'ghost' were always over-estimated when

there was no gene flow into the 'ghost', while under-estimated when there was a larger degree of gene flow into the 'ghost' (e.g. Fig. 22,23). This could be an artifact of the fact that the divergence time between the common ancestor of the sampled populations, and the unsampled 'ghost' population is as yet too small, leading to erroneous conclusions of migration rates between recently diverged populations (also see (Hey, Chung, & Sethuraman, 2015)).

Increasing the number of sampled genomic loci generally improved the confidence intervals around all estimates - this could be an ideal strategy, especially in this age of next generation sequencing, and the ability to obtain long haplotypic segments at lower costs.

Estimates of effective population sizes were largely robust to the evolutionary model - for instance, in a scenario where there is no migration from an unsampled 'ghost' population, the effective population size of the common ancestor of the sampled populations is accurately estimated (Fig. 12,13).

Across all our simulations, including an unsampled 'ghost' population in the model while inferring evolutionary parameters improved all estimates. This should be an ideal strategy across all population genetics studies, where the true population tree or history is unknown. For instance, in a recent study, (Hey et al., 2018) showed while estimating the evolutionary history of African Hunter-Gatherers (Hadza, Sandawe), and Pastoralists/Agriculturalists (Yoruba, Baka), the inclusion of a 'ghost' population (a) led to the inference of significant unidirectional migration from the 'ghost' into Baka, Yoruba, and Sandawe, and the common ancestor of all four populations, while ignoring the 'ghost' did not estimate any migration between all sampled populations, and (b) accurate estimation of smaller effective population sizes of all sampled populations. Both these patterns are mimicked in our simulation study as well.

Interestingly, the same study (hey2018phylogeny) estimated the evolutionary history of chimpanzees using 200 genomic loci, and did not find significant evidence of contemporary gene flow to or from an unsampled 'ghost' population. However, another recent study by (Kuhlwilm et al., 2019) uses whole genome data from chimpanzees and bonobos to estimate significant unidirectional gene flow from an extinct 'ghost' lineage into modern bonobos. This is perhaps indicative of the importance of using a large number of genomic loci while inferring evolutionary history.

Our study re-iterates the key points made by (Beerli, 2004), in that while it is impossible to always sample all species, or large number of informative genomic loci, it is imperative to account for the presence of unsampled 'ghost' populations while estimating summary statistics or the evolutionary history of sampled species.

Acknowledgments

This work was supported by an NSF ABI Grant 1564659 to Arun Sethuraman and Jody Hey. This research includes calculations carried out on Temple University's HPC resources and thus was supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189.

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, *19*(9), 1655–1664.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Molecular Ecology*, *13*(4), 827–836.

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, *98*(8), 4563–4568.

Chung, Y., & Hey, J. (2017, 02). Bayesian Analysis of Evolutionary Divergence with Genomic Data under Diverse Demographic Models. *Molecular Biology and Evolution*, *34*(6), 1517-1528. doi: 10.1093/molbev/msx070

Durvasula, A., & Sankararaman, S. (2019). Recovering signals of ghost archaic introgression in african populations. *bioRxiv*, 285734.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics*, *5*(10), e1000695.

Hey, J., Chung, Y., & Sethuraman, A. (2015). On the occurrence of false positives in tests of migration under an isolation-with-migration model. *Molecular ecology*, *24*(20), 5078–5083.

Hey, J., Chung, Y., Sethuraman, A., Lachance, J., Tishkoff, S., Sousa, V. C., & Wang, Y. (2018). Phylogeny estimation by integration over isolation with migration models. *Molecular biology and evolution*, *35*(11), 2805–2818.

Hey, J., & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *Genetics*, *167*(2), 747–760.

Hey, J., & Nielsen, R. (2007). Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, *104*(8), 2785–2790.

Hudson, R. R. (2002). Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338.

Konečnỳ, A., Estoup, A., Duplantier, J.-M., Bryja, J., Bâ, K., Galan, M., ... Cosson, J.-F. (2013). Invasion genetics of the introduced black rat (rattus rattus) in senegal, west africa. *Molecular Ecology*, *22*(2), 286–300.

Kuhlwilm, M., Han, S., Sousa, V. C., Excoffier, L., & Marques-Bonet, T. (2019). Ancient admixture from an extinct ape lineage into bonobos. *Nature ecology & evolution*, *3*(6), 957.

Lachance, J., Vernot, B., Elbers, C. C., Ferwerda, B., Froment, A., Bodo, J.-M., ... others (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse african hunter-gatherers. *Cell*, *150*(3), 457–469.

Lawson, D. J., Van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret structure and admixture bar plots. *Nature Communications*, *9*(1), 3258.

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302.

Nielsen, R., & Wakeley, J. (2001). Distinguishing migration from isolation: a markov chain monte carlo approach. *Genetics*, *158*(2), 885–896.

Pfeifer, B., Wittelsbuerger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). Popgenome: An efficient swiss army knife for population genomic analyses in r. *Molecular Biology and Evolution*, *31*, 1929-1936. doi: 10.1093/molbev/msu136

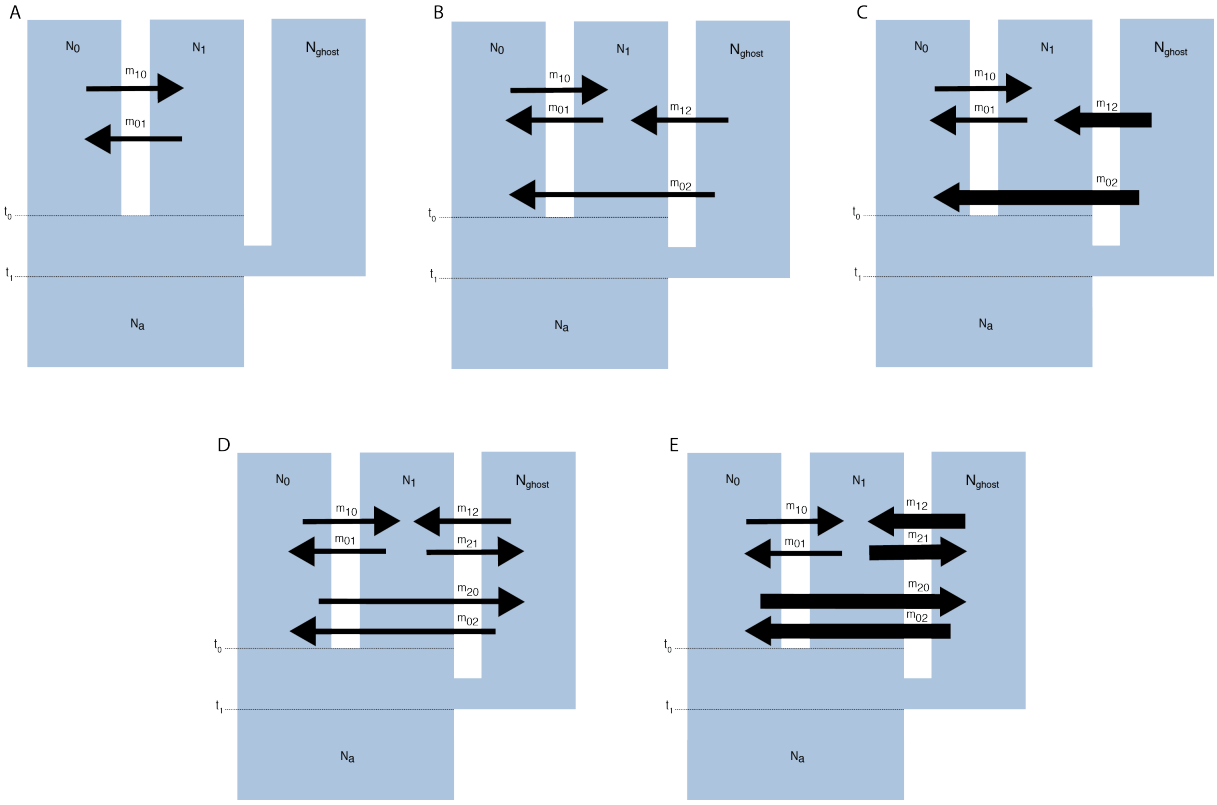
Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Rosenthal, D. M., Ramakrishnan, A. P., & Cruzan, M. B. (2008). Evidence for multiple sources of invasion and intraspecific hybridization in brachypodium sylvaticum (hudson) beauv. in north america. *Molecular Ecology*, *17*(21), 4657–4669.

Sethuraman, A., & Hey, J. (2016). Im a2p–parallel mcmc and inference of ancient demography under the isolation with migration (im) model. *Molecular ecology resources*, *16*(1), 206–215.

Slarkin, M. (1985). Gene flow in natural populations. *Annual review of ecology and systematics*, *16*(1), 393–430.

Figure 1: Isolation with migration models. N_0 and N_1 are sampled populations (which diverged at time t_0), N_{ghost} is the unsampled 'ghost' population (represented by θ_2 in figures), and N_a (represented by θ_3 in figures) represents the common ancestor between N_0, N_1 and 'ghost' (which diverged at time t_1). The arrows indicate either high (bold arrow) or low (small arrow) level of gene flow, as well as direction.



Data Accessibility

All simulation scripts, `IMa2p` datasets will be made available on the Author's GitHub page.

Author Contributions

ML conceptualized the study, performed all the simulations, ran all the IM analyses, and AS and ML wrote the paper.

Tables and Figures

Figure 2: Summary Statistic Estimates for 2 loci

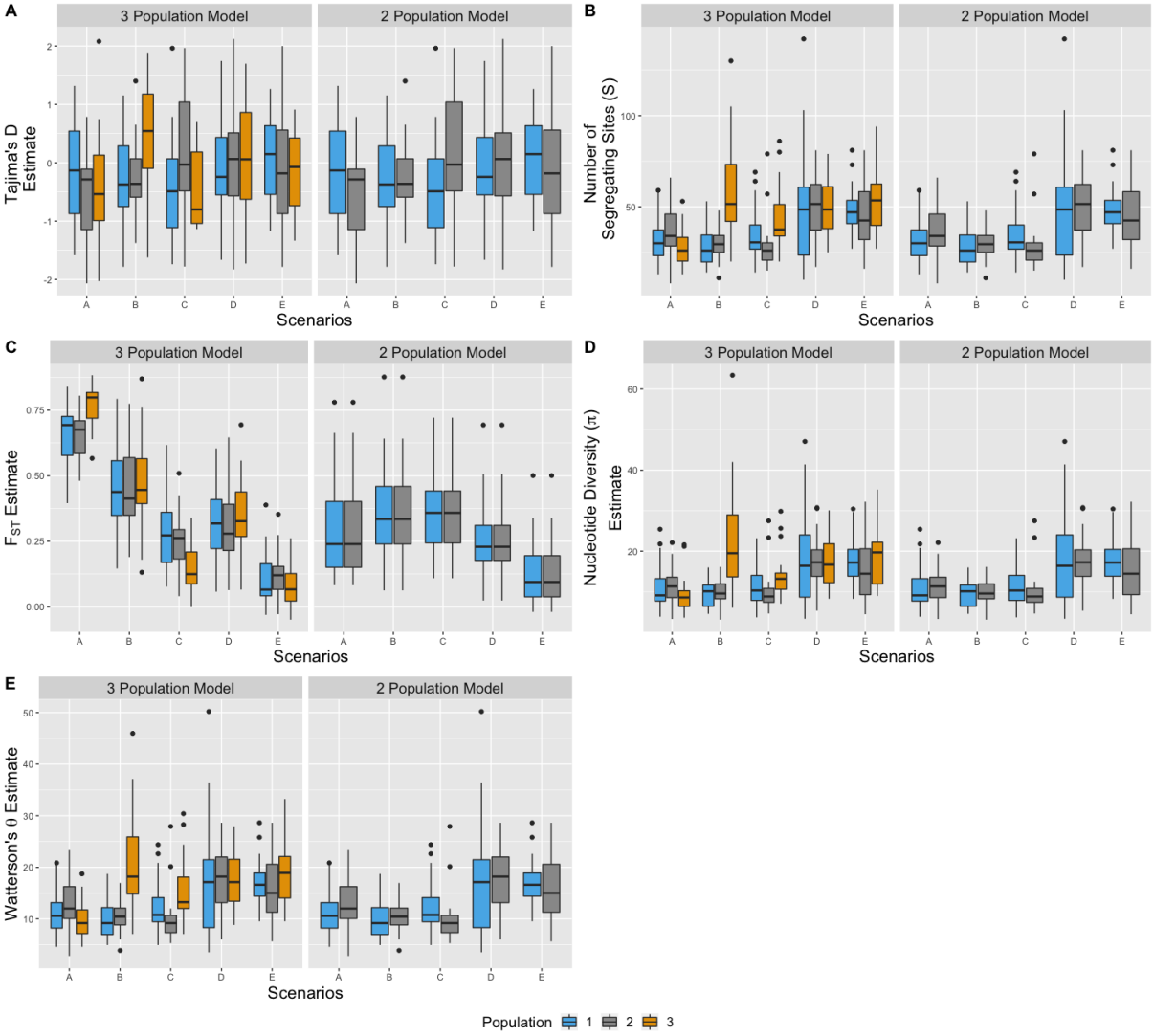


Figure 3: Summary Statistic Estimates for 5 loci

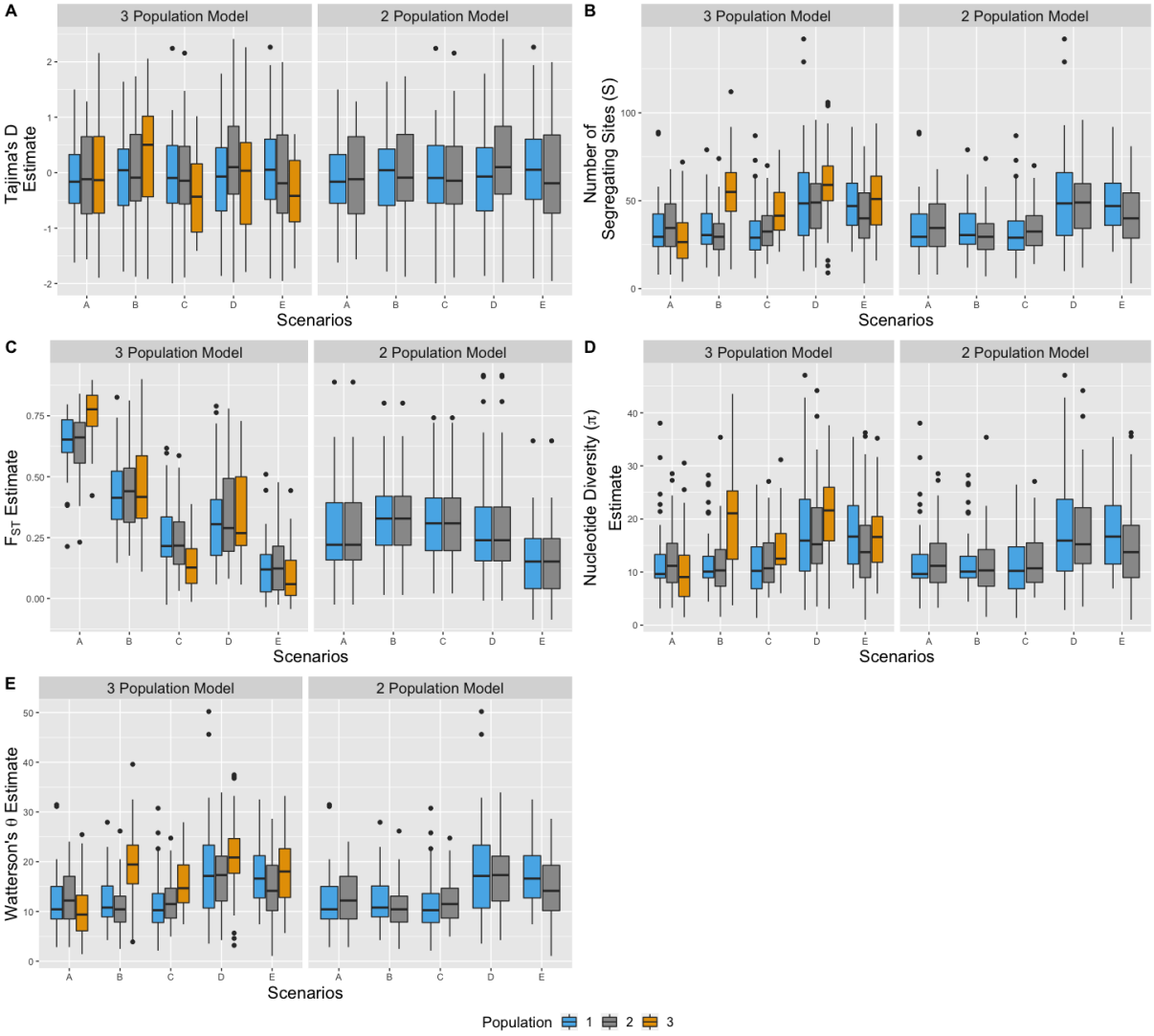


Figure 4: Divergence time (t_0) estimates between sampled populations, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

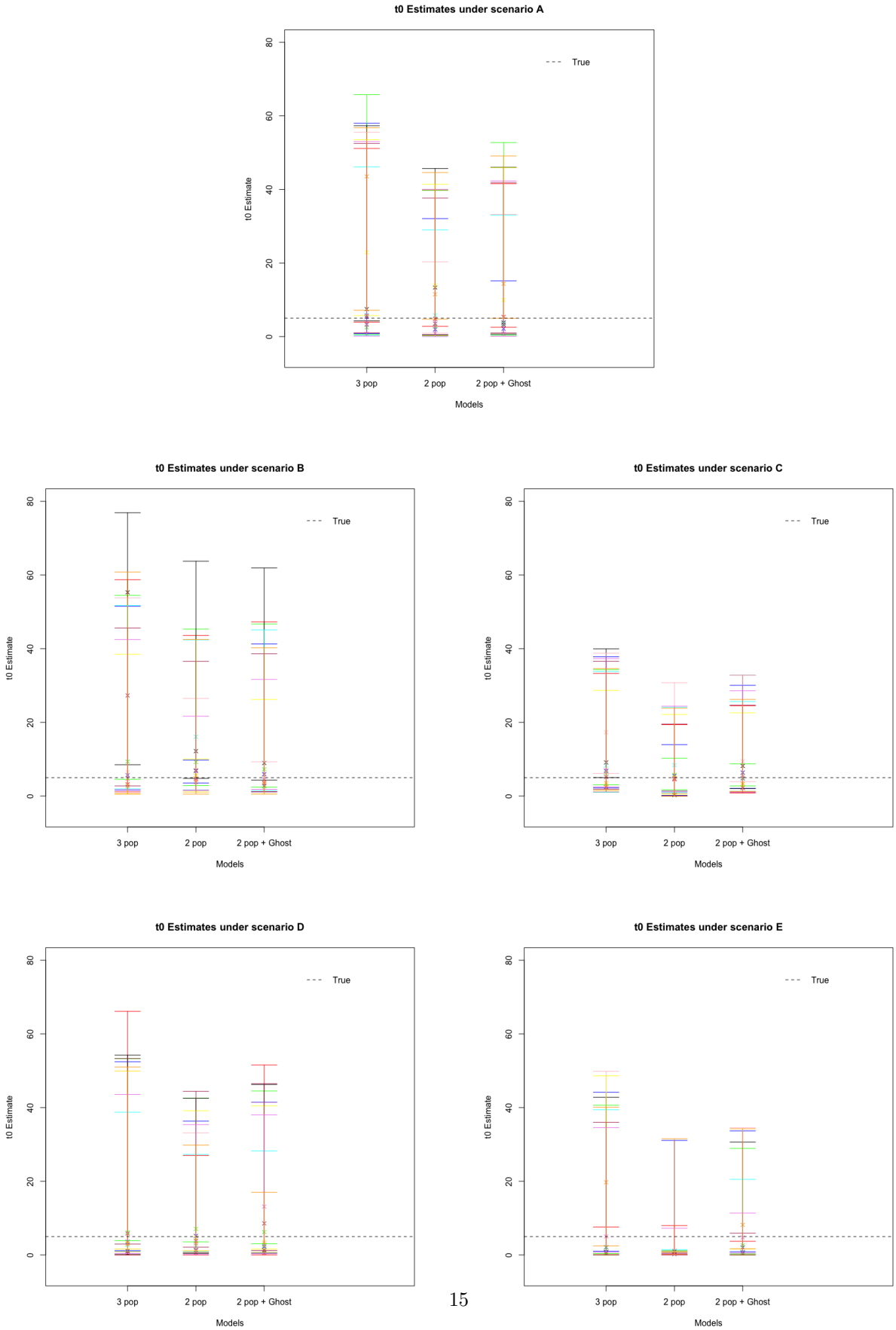


Figure 5: Divergence time (t_0) estimates between sampled populations, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

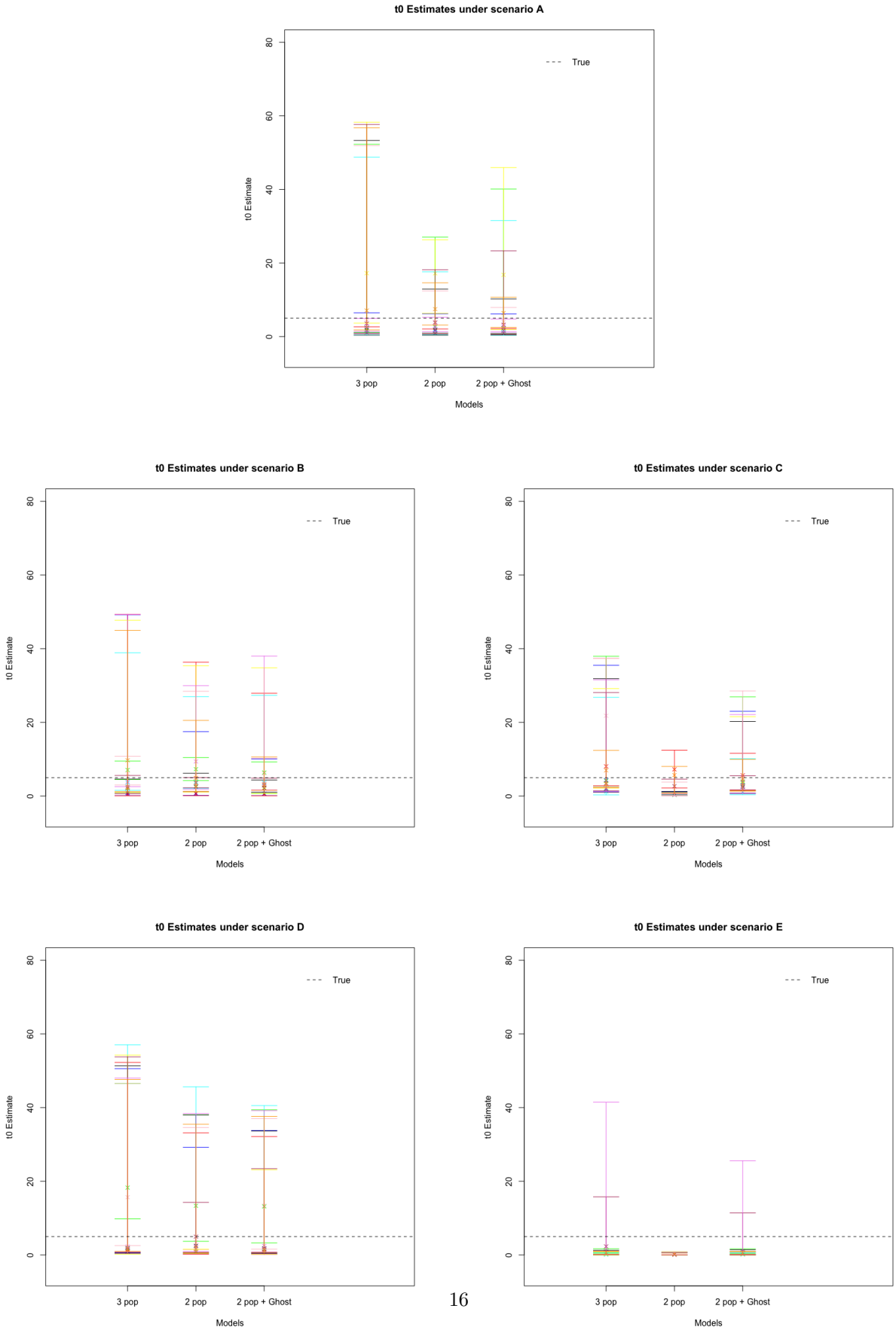


Figure 6: Divergence time (t_1) estimates between the common ancestor of sampled populations, and the ghost outgroup, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

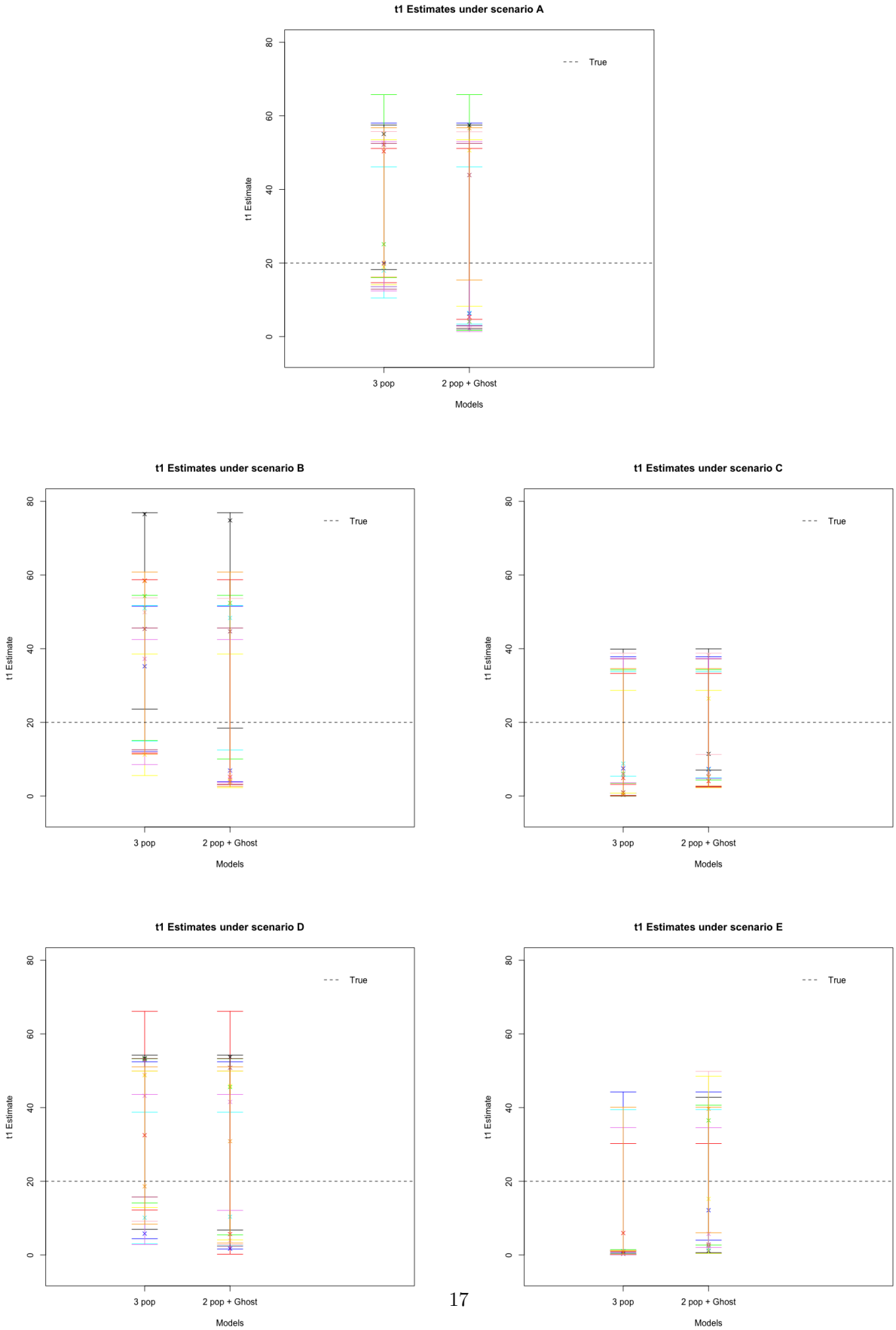


Figure 7: Divergence time (t_1) estimates between the common ancestor of sampled populations, and the ghost outgroup, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

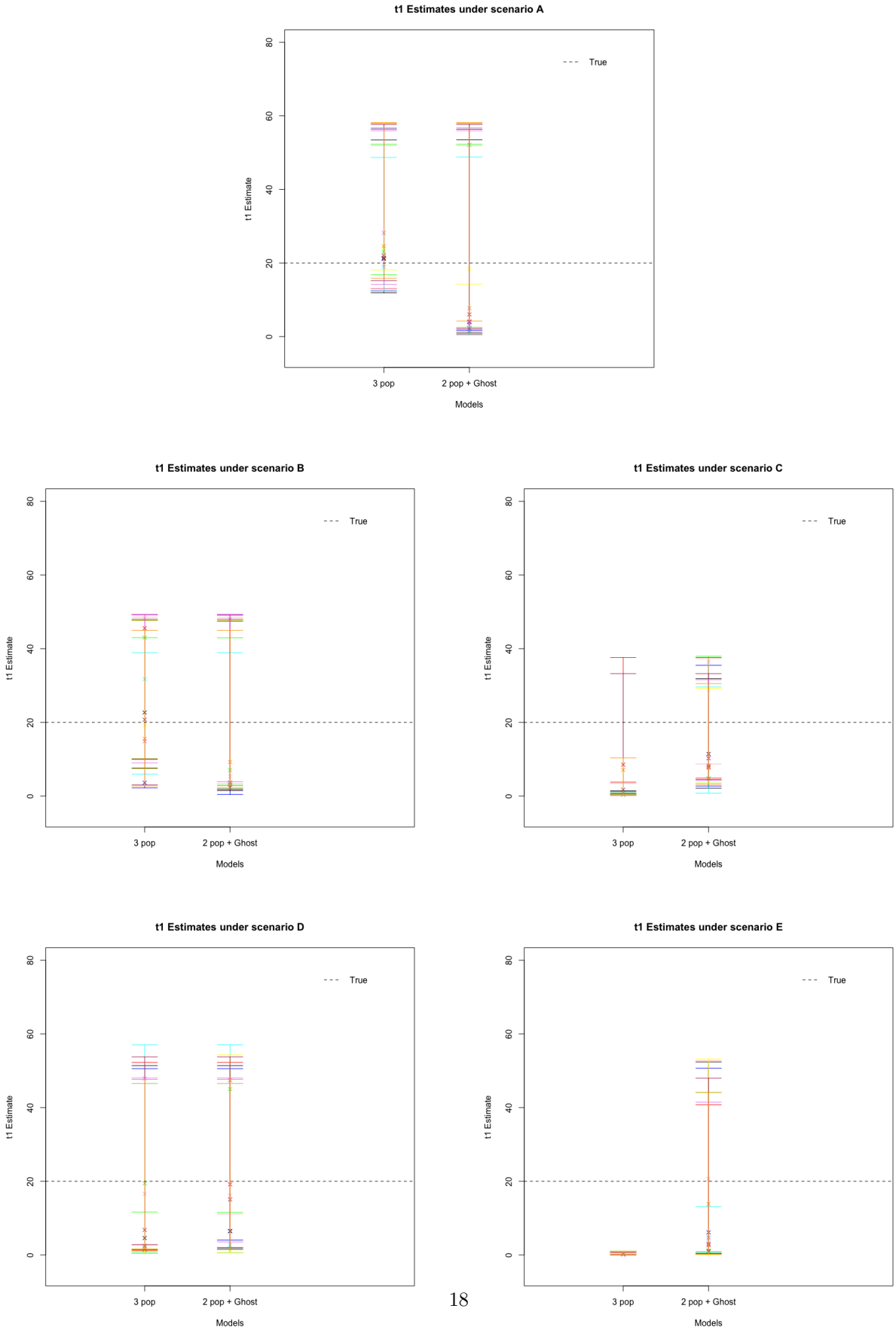


Figure 8: Scaled effective population size (θ_0) estimate of the first sampled populations, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

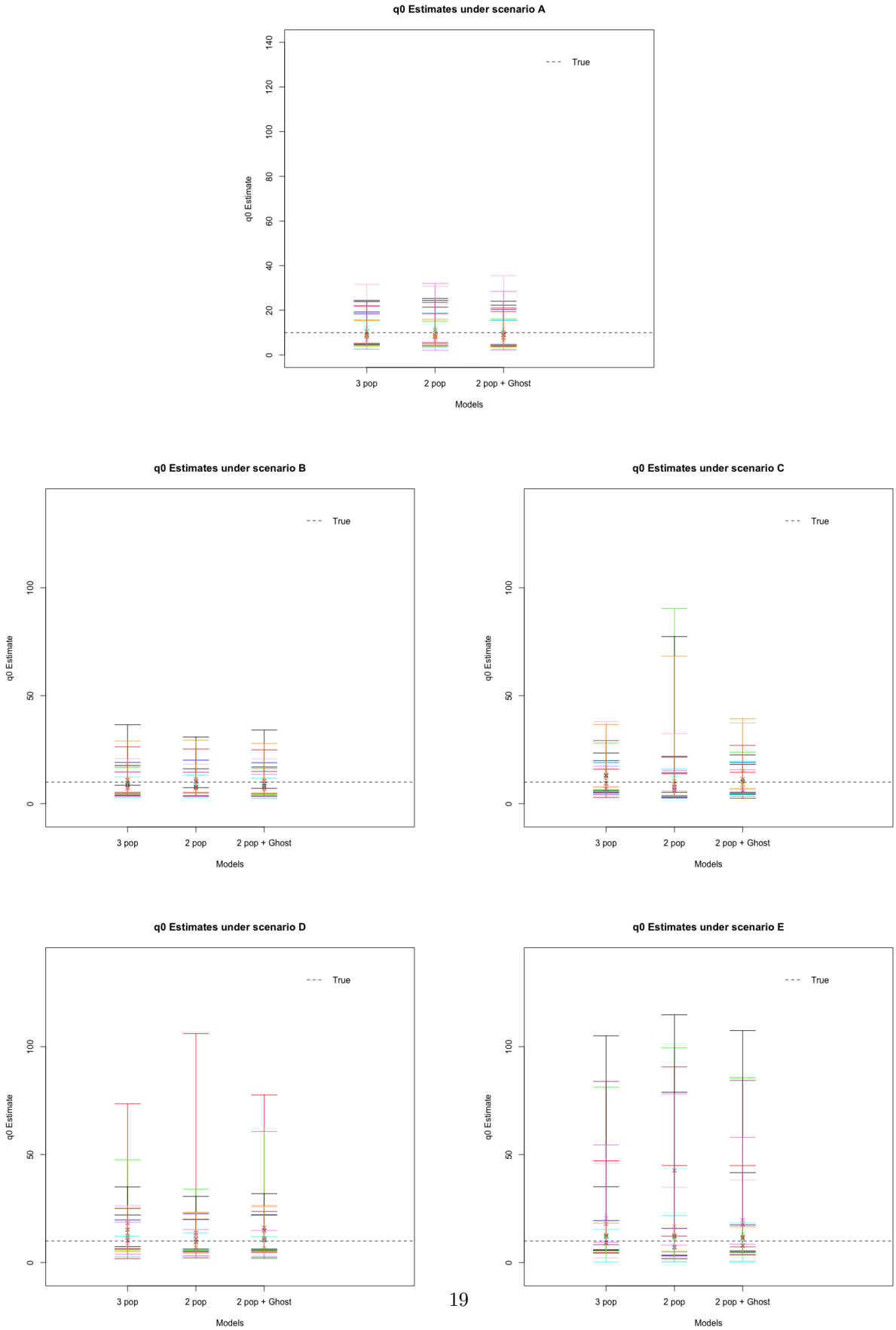


Figure 9: Scaled effective population size (θ_0) estimate of the first sampled population, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

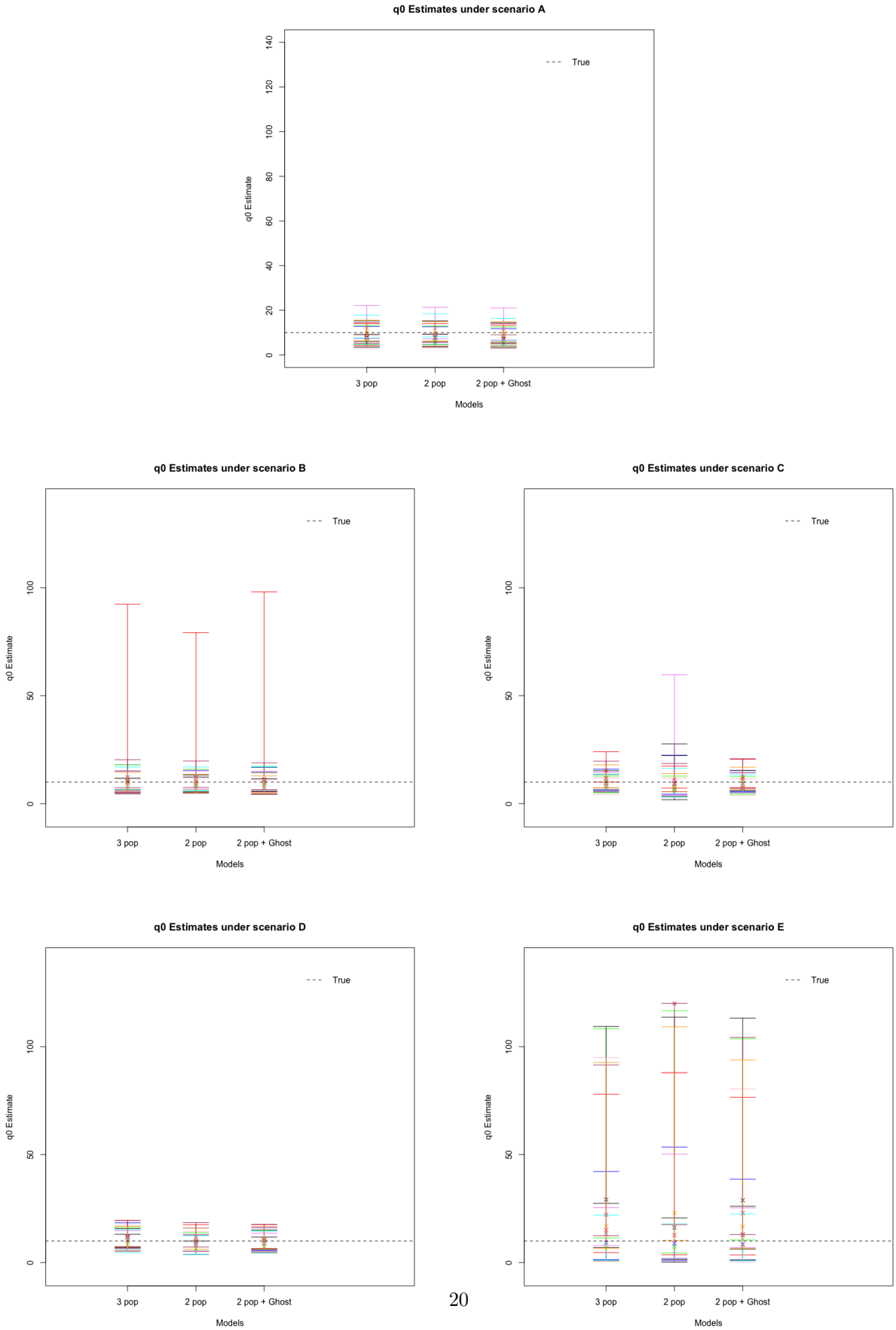


Figure 10: Scaled effective population size (θ_1) estimate of the second sampled population, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

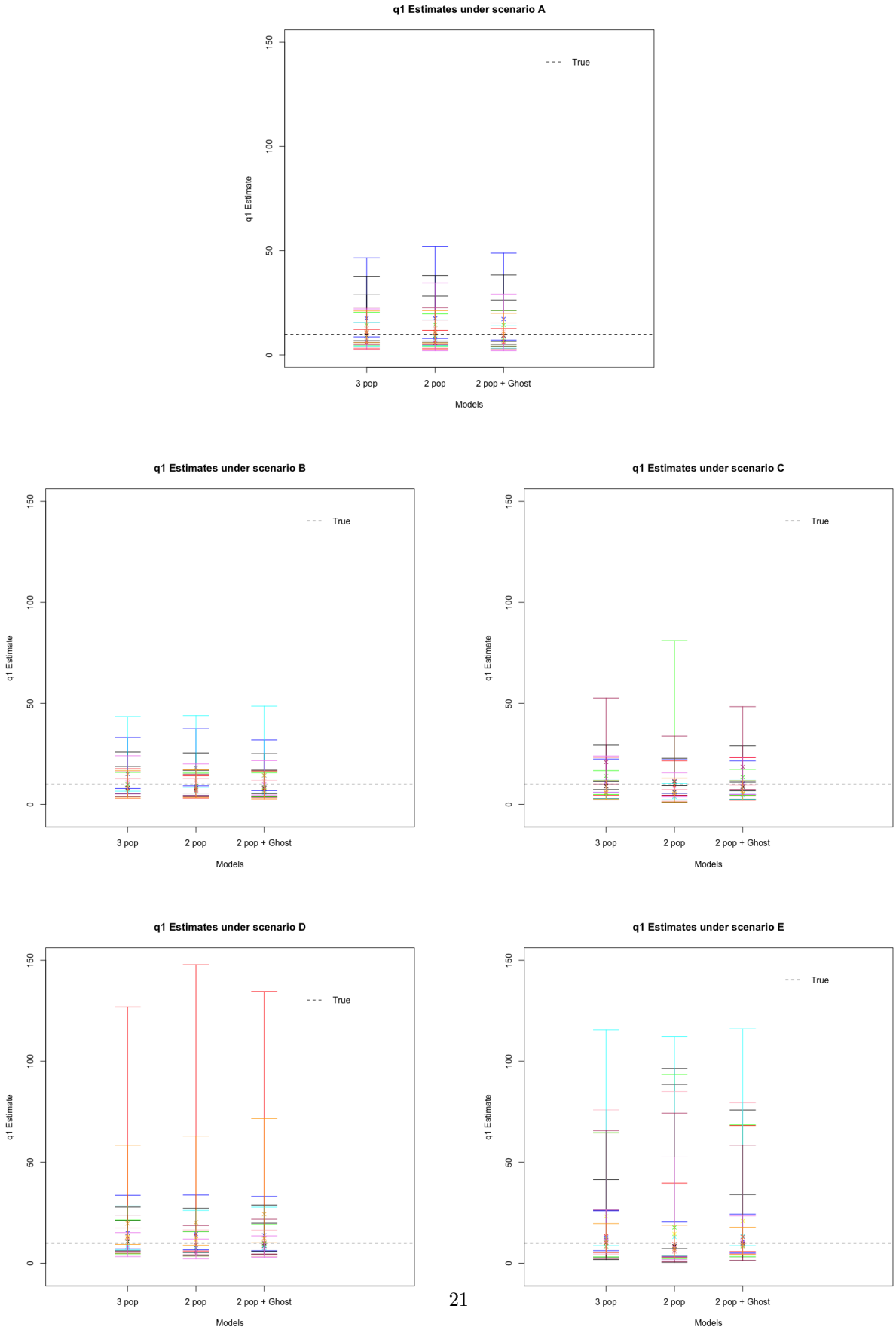


Figure 11: Scaled effective population size (θ_1) estimate of the second sampled population, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

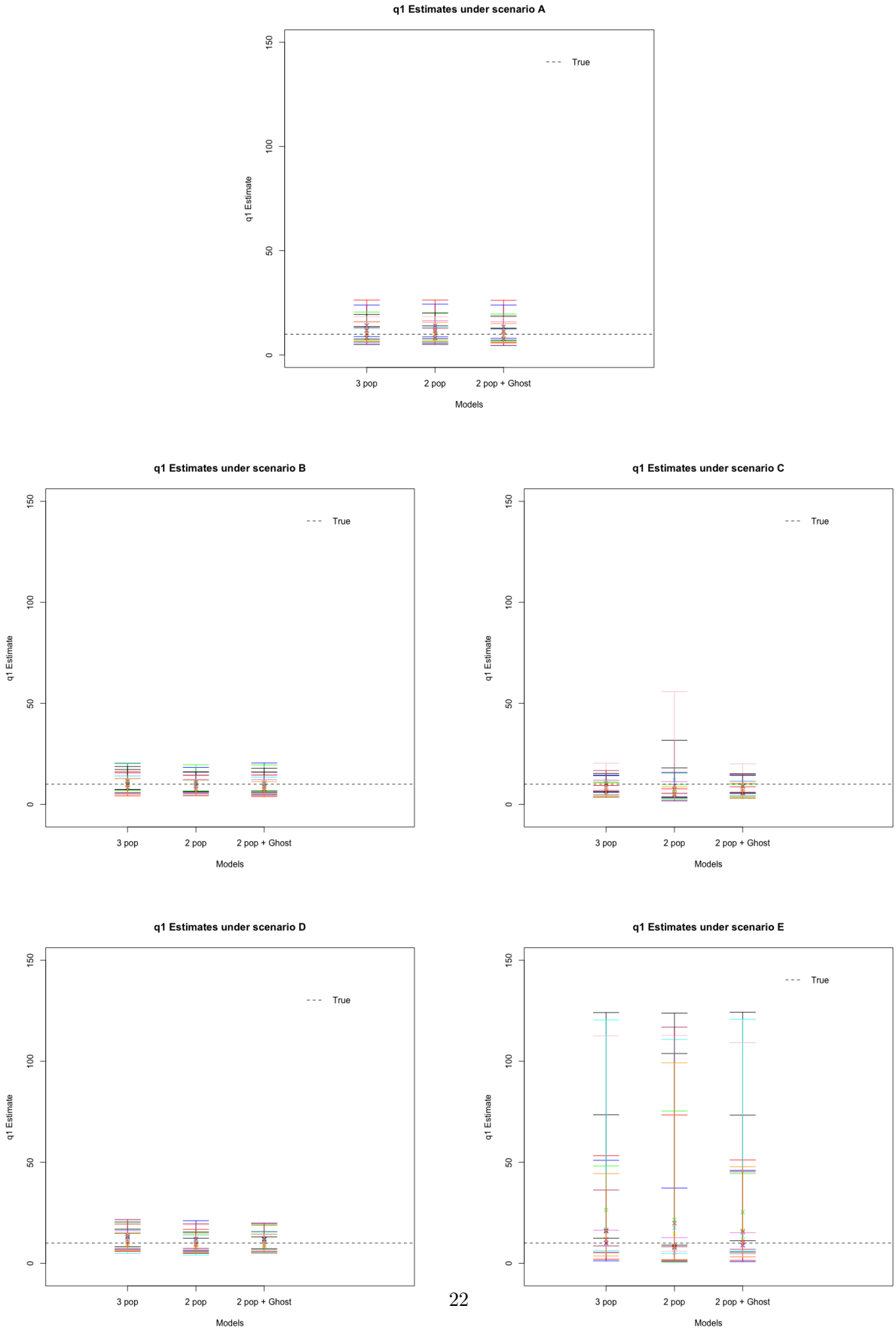


Figure 12: Scaled effective population size (θ_2) estimate of the common ancestor of the two sampled populations, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

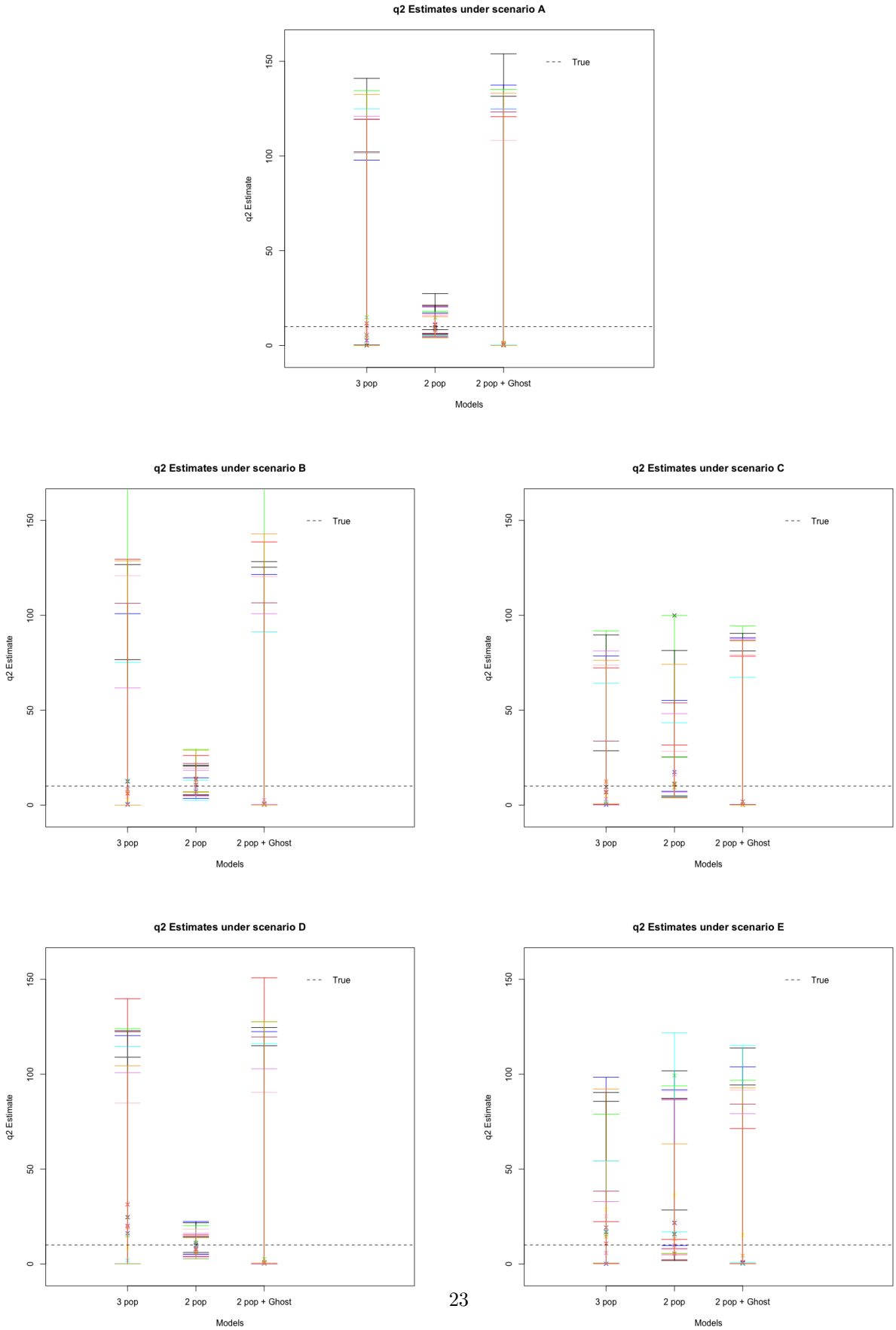


Figure 13: Scaled effective population size (θ_2) estimate of the common ancestor of the two sampled populations, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

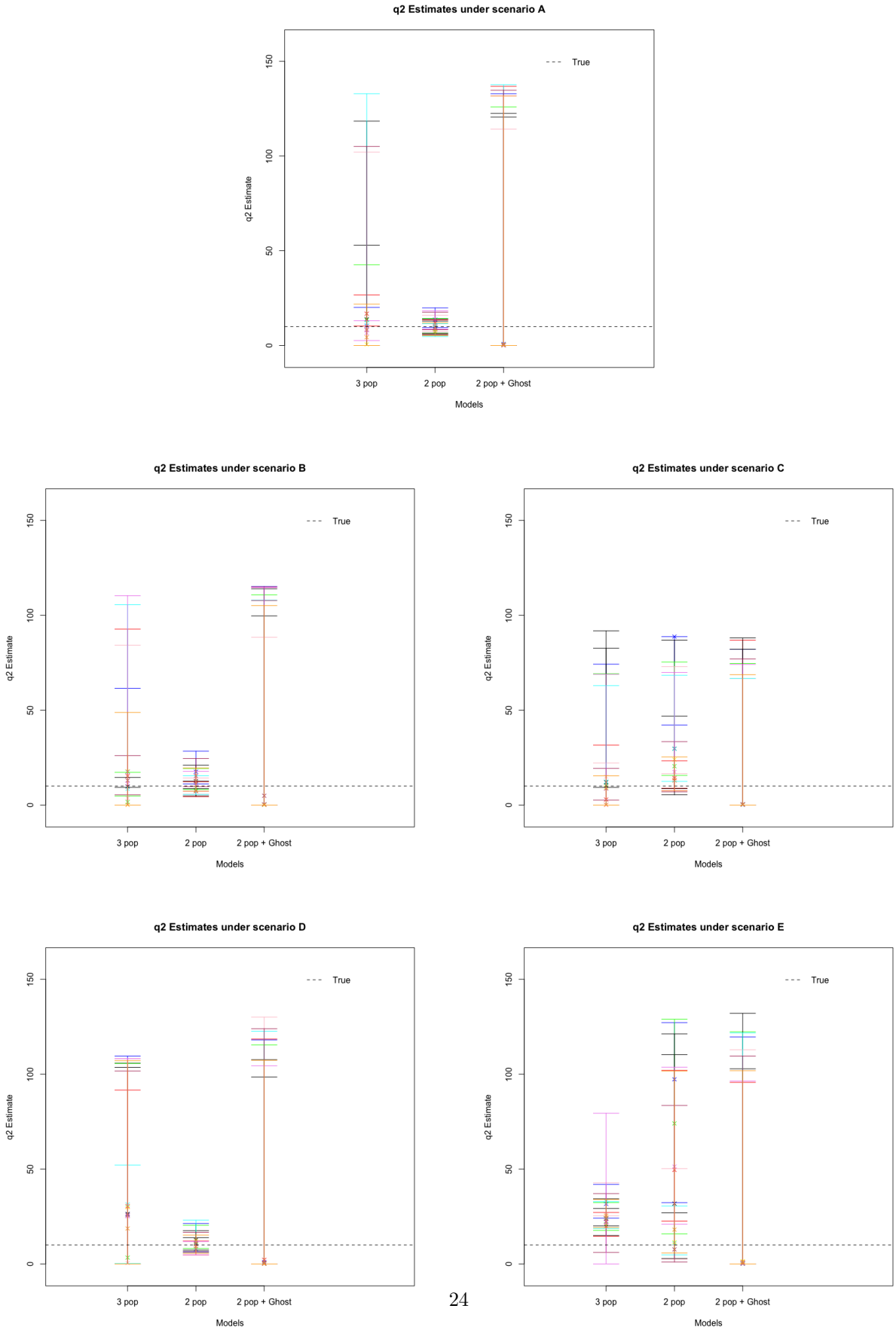


Figure 14: Scaled effective population size (θ_3) estimate of the ghost outgroup, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

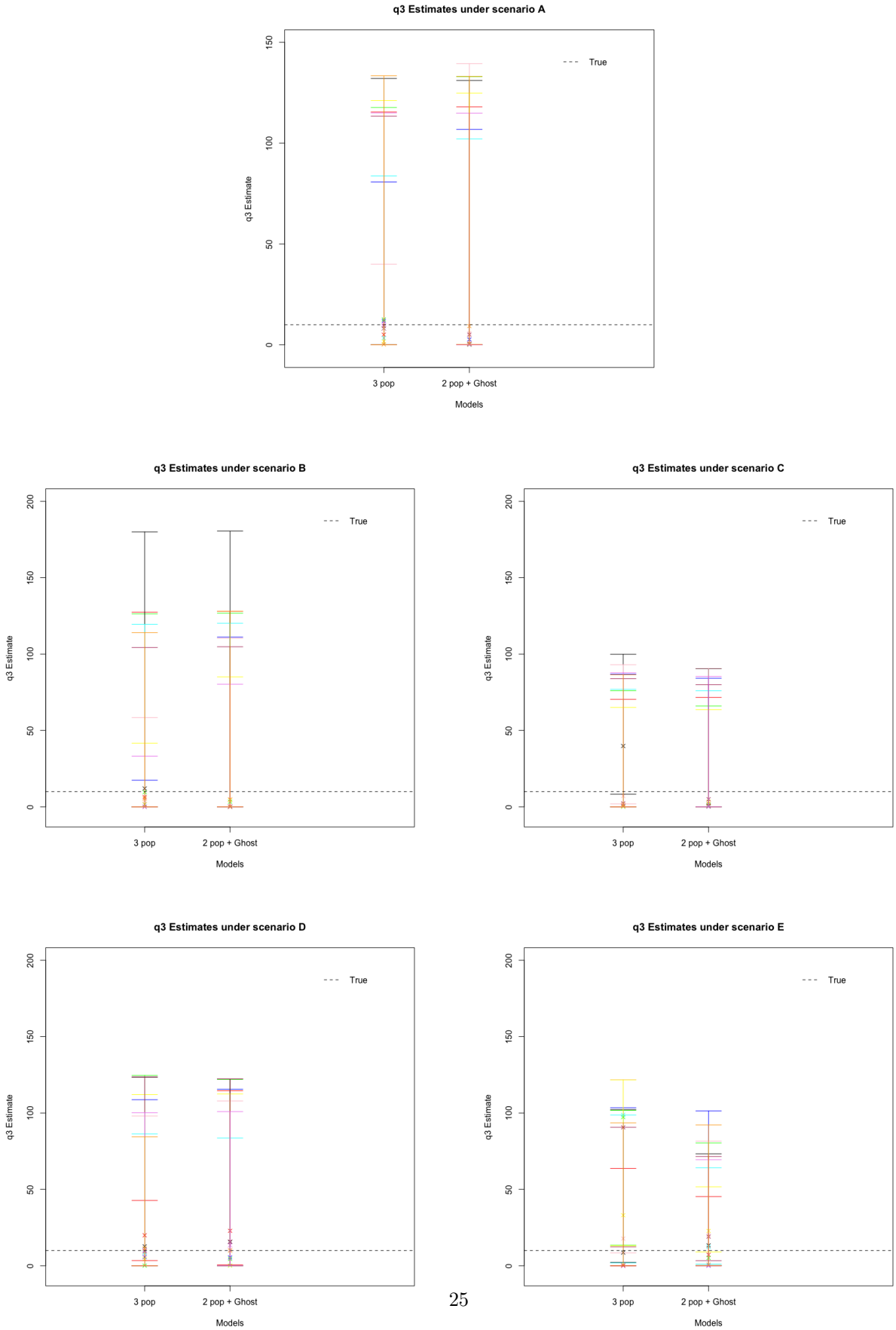


Figure 15: Scaled effective population size (θ_3) estimate of the ghost outgroup, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

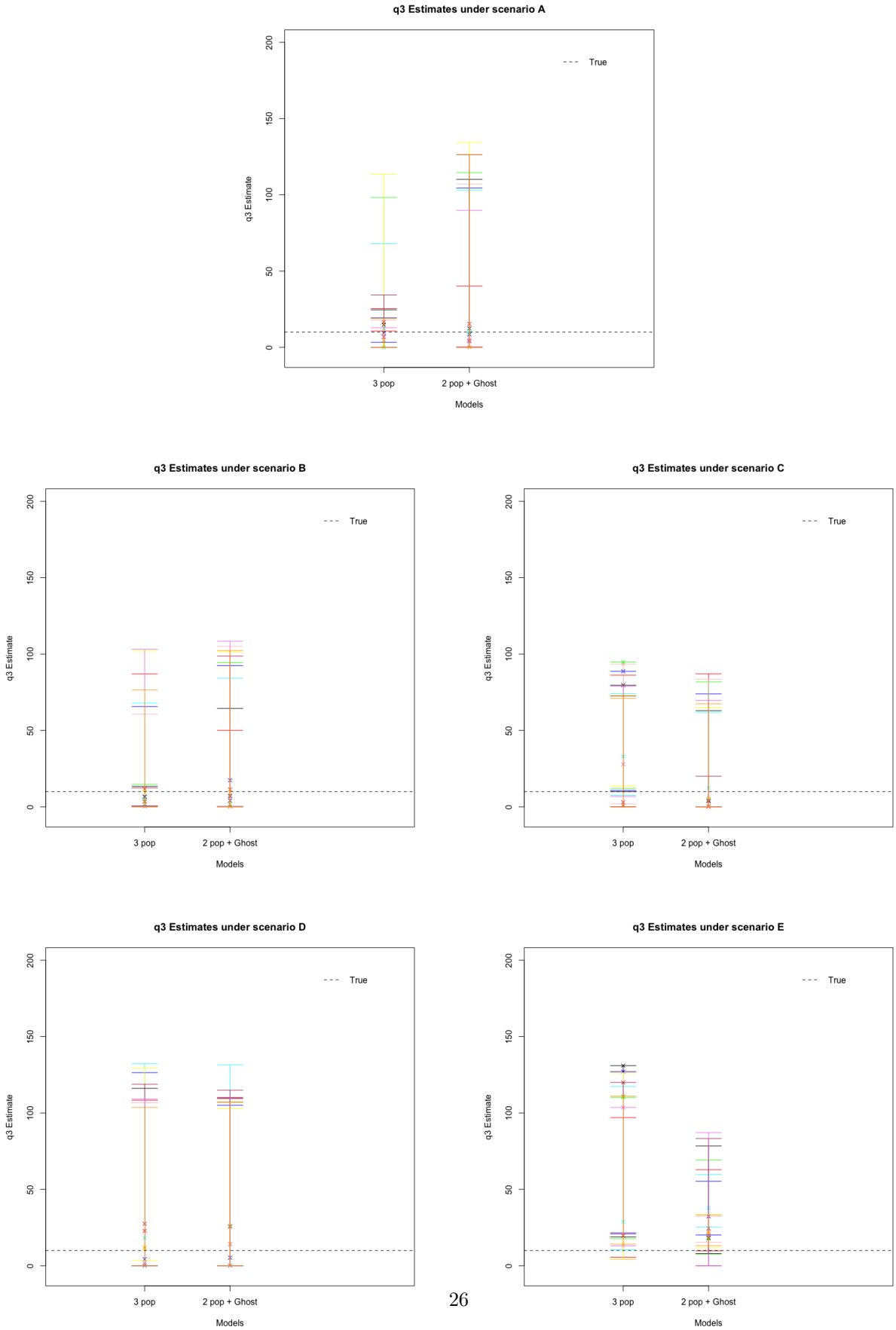


Figure 16: Scaled effective population size (θ_4) estimate of the common ancestor of the two sampled populations, and the ghost outgroup, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

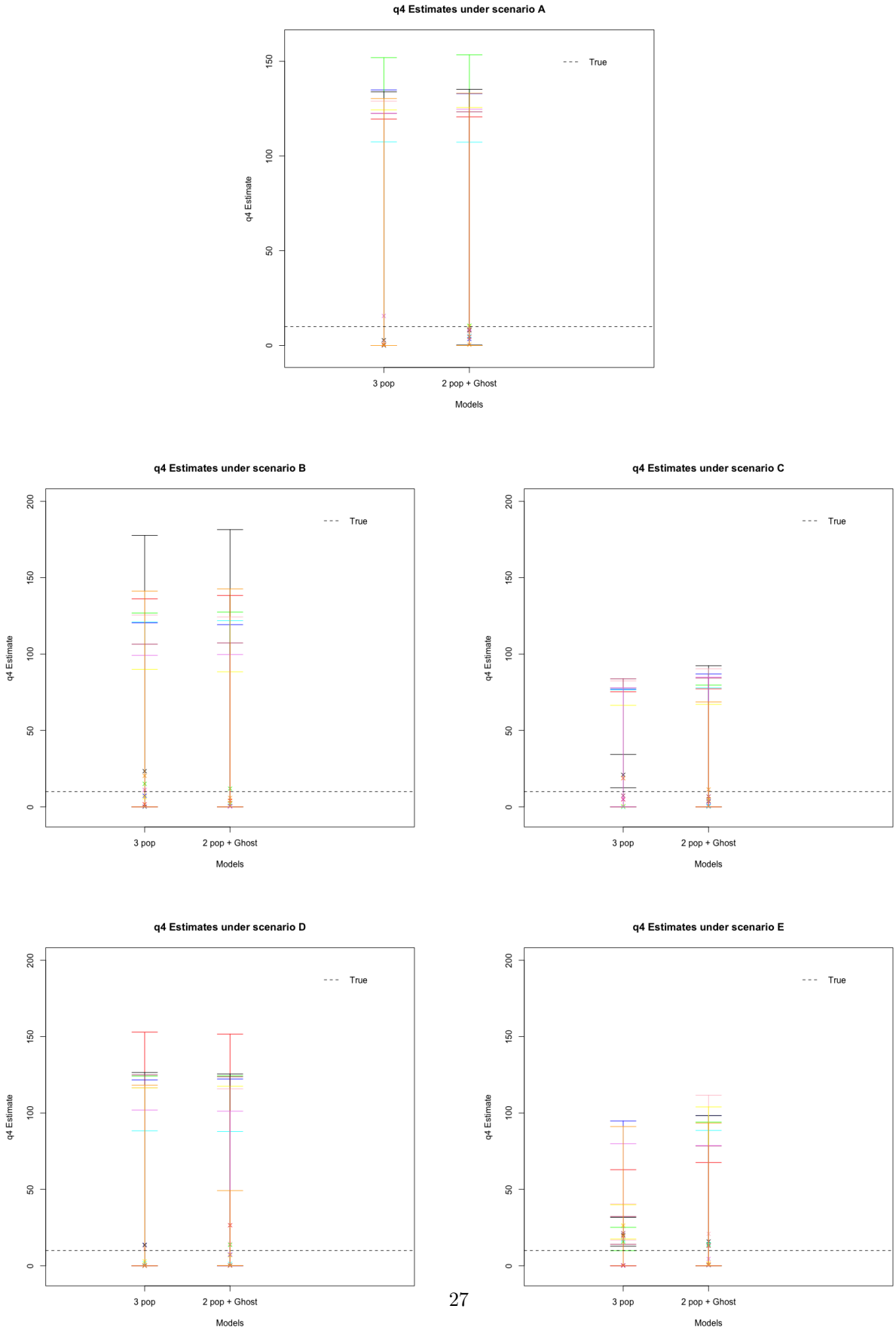


Figure 17: Scaled effective population size (θ_4) estimate of the common ancestor of the two sampled populations, and the ghost outgroup, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

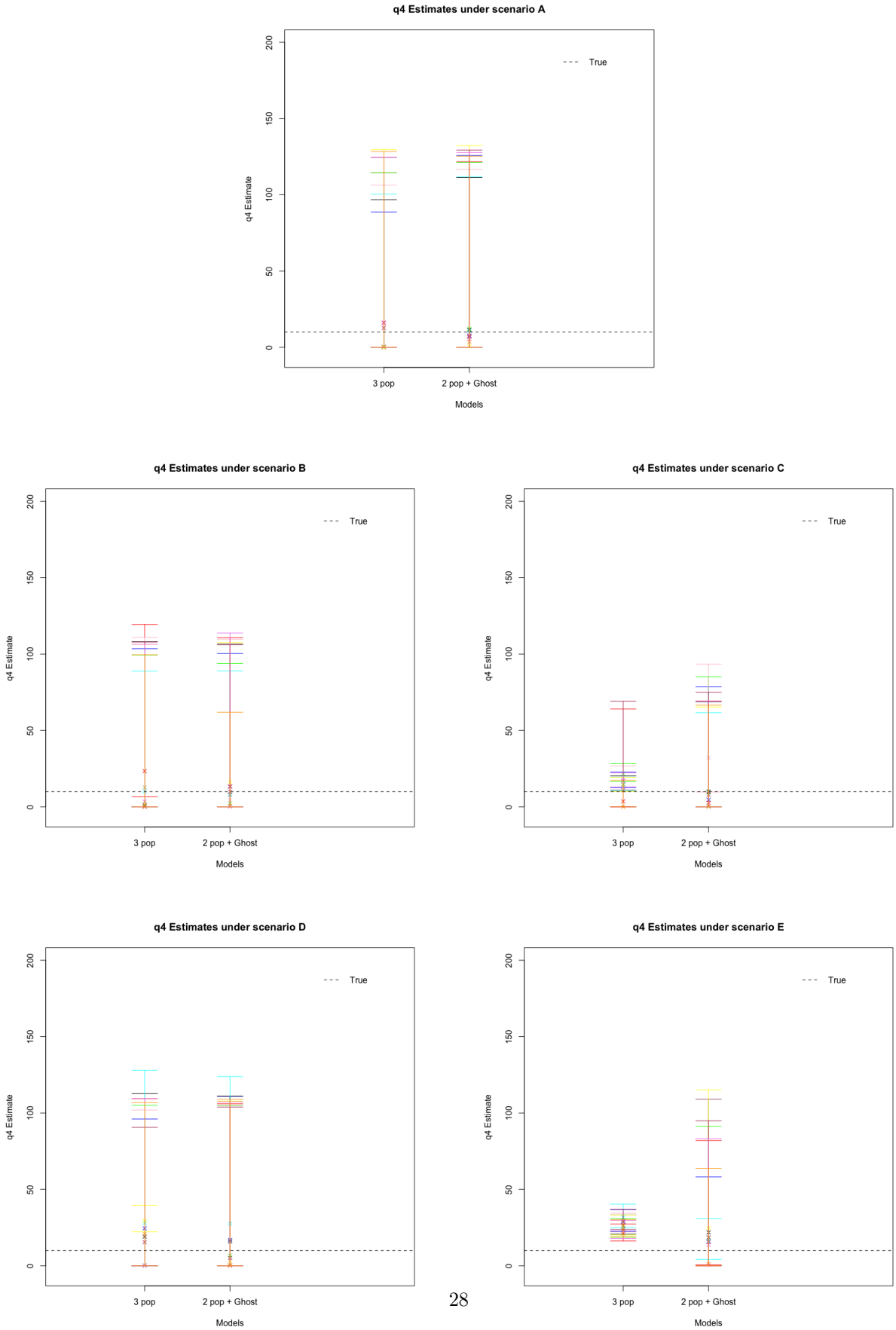


Figure 18: Scaled migration rate (m_{10}) estimate between the two sampled populations, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

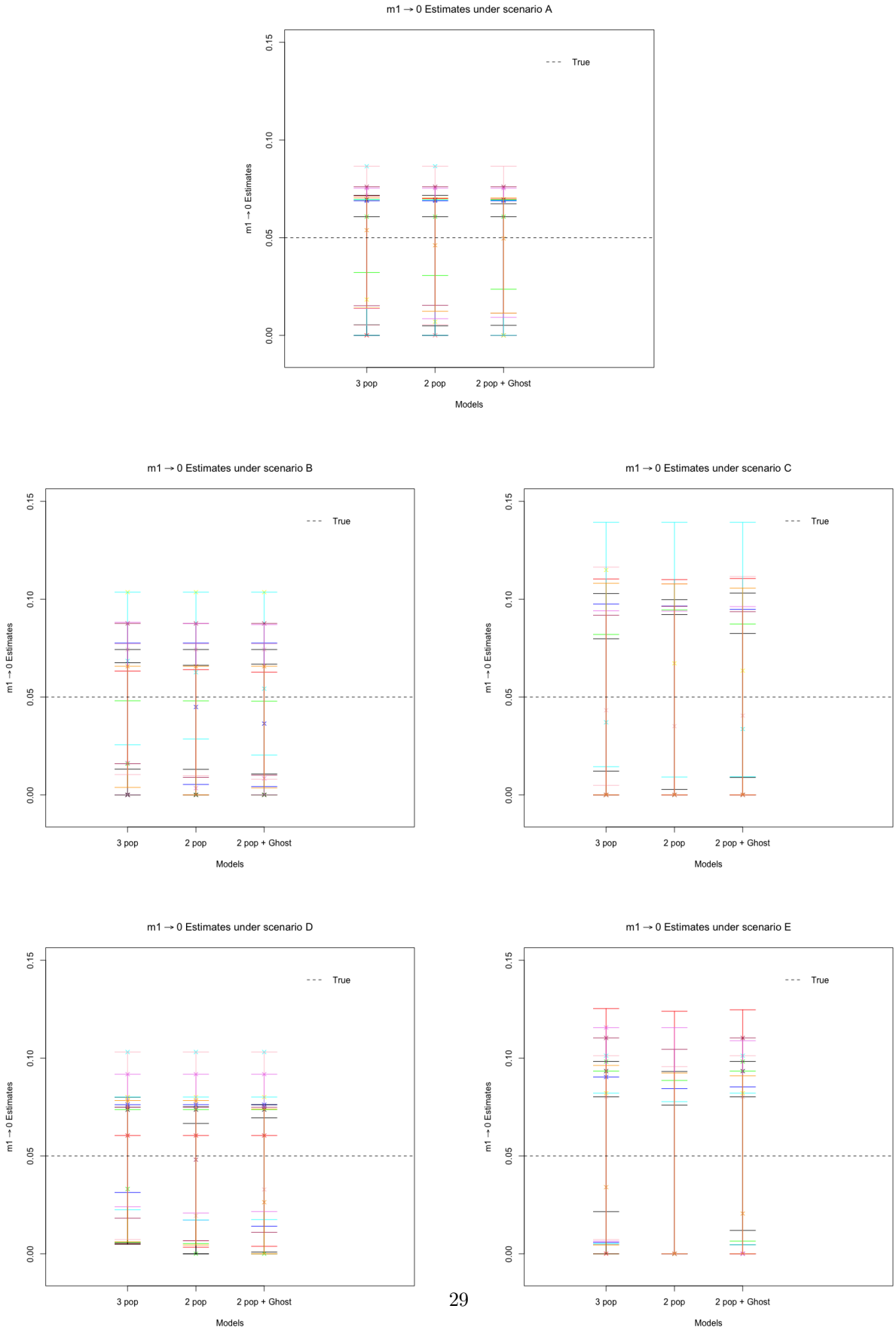


Figure 19: Scaled migration rate (m_{10}) estimate between the two sampled populations, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

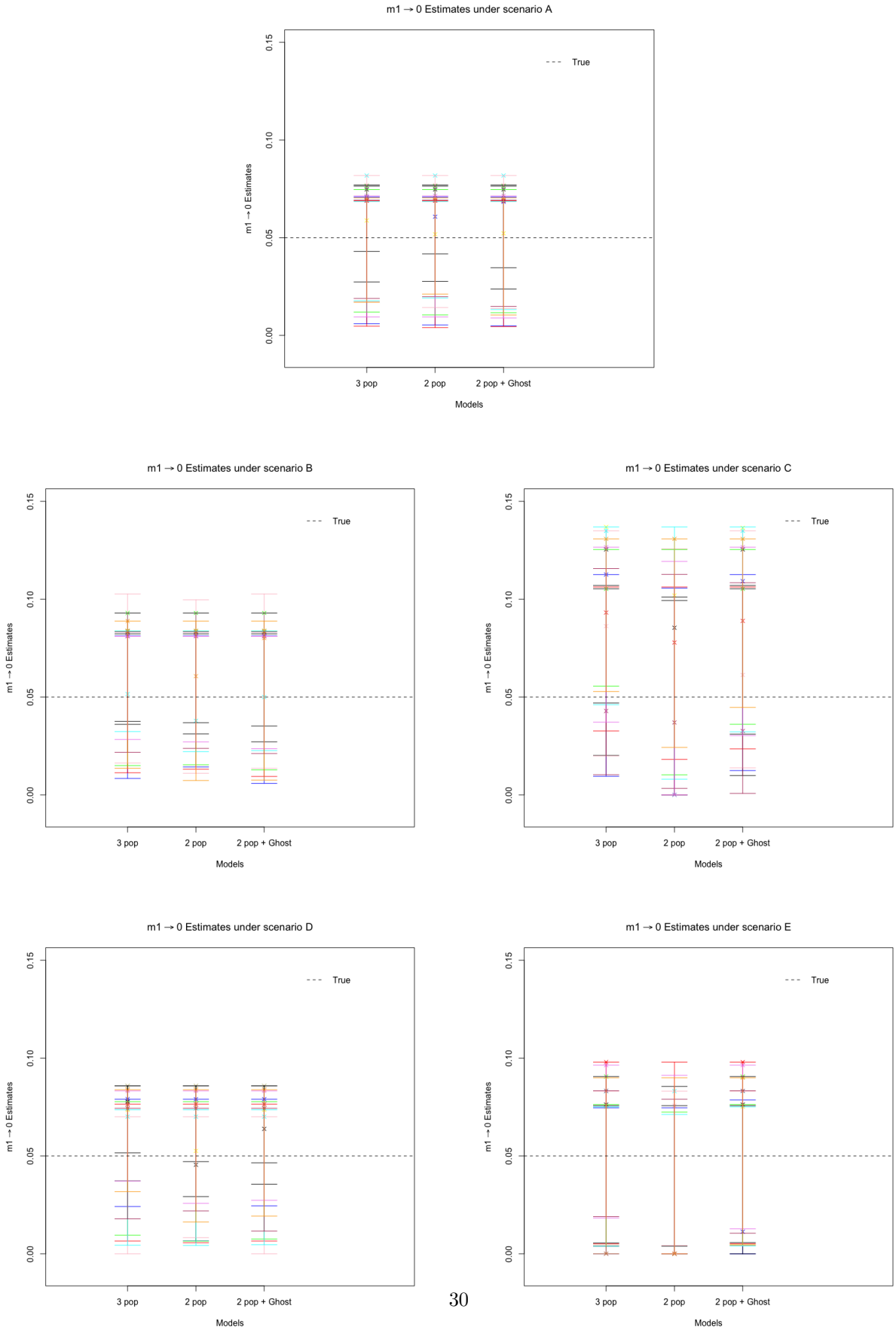


Figure 20: Scaled migration rate (m_{01}) estimate between the two sampled populations, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

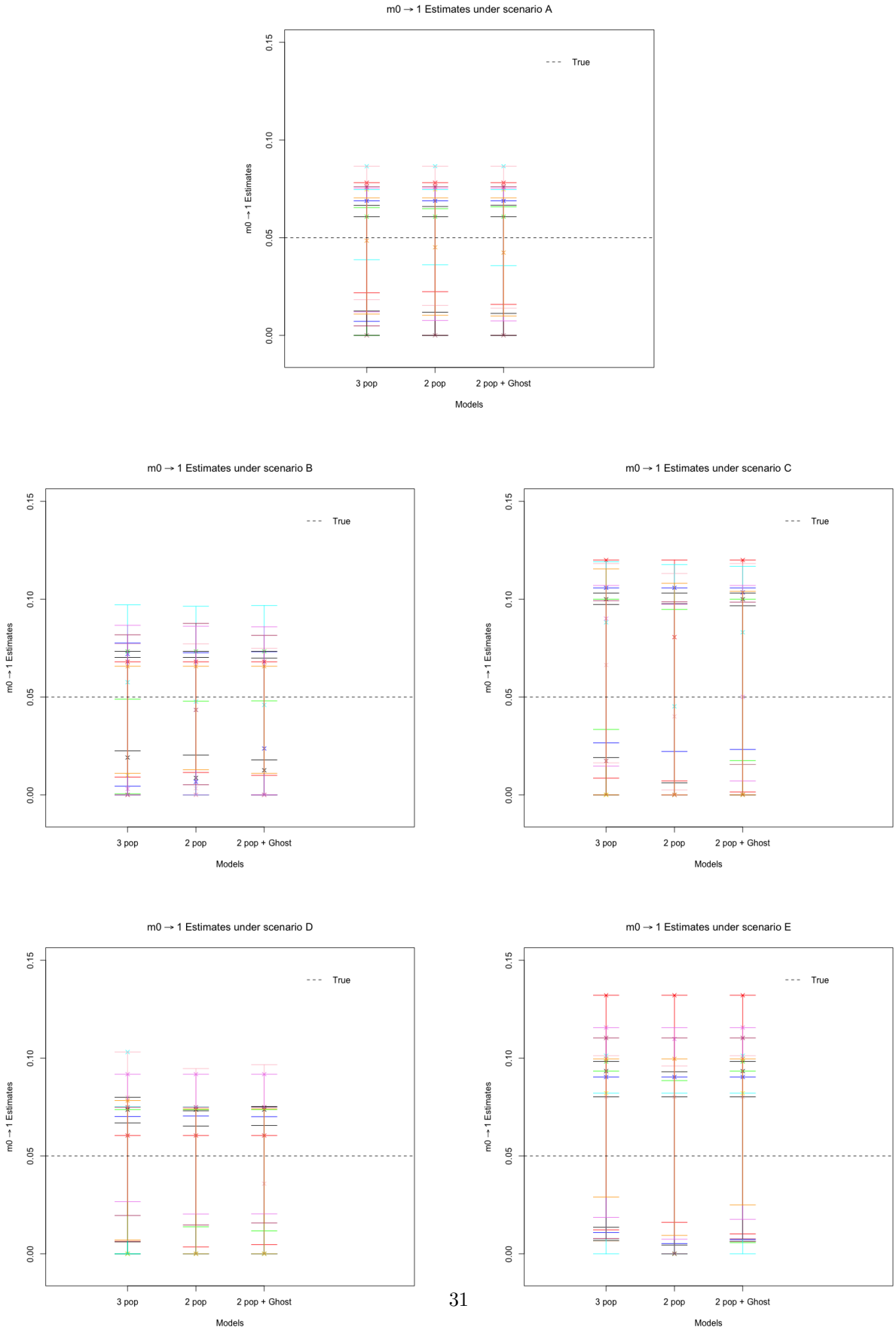


Figure 21: Scaled migration rate (m_{01}) estimate between the two sampled populations, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, 2-population model, and a 2-population model with a ghost outgroup.

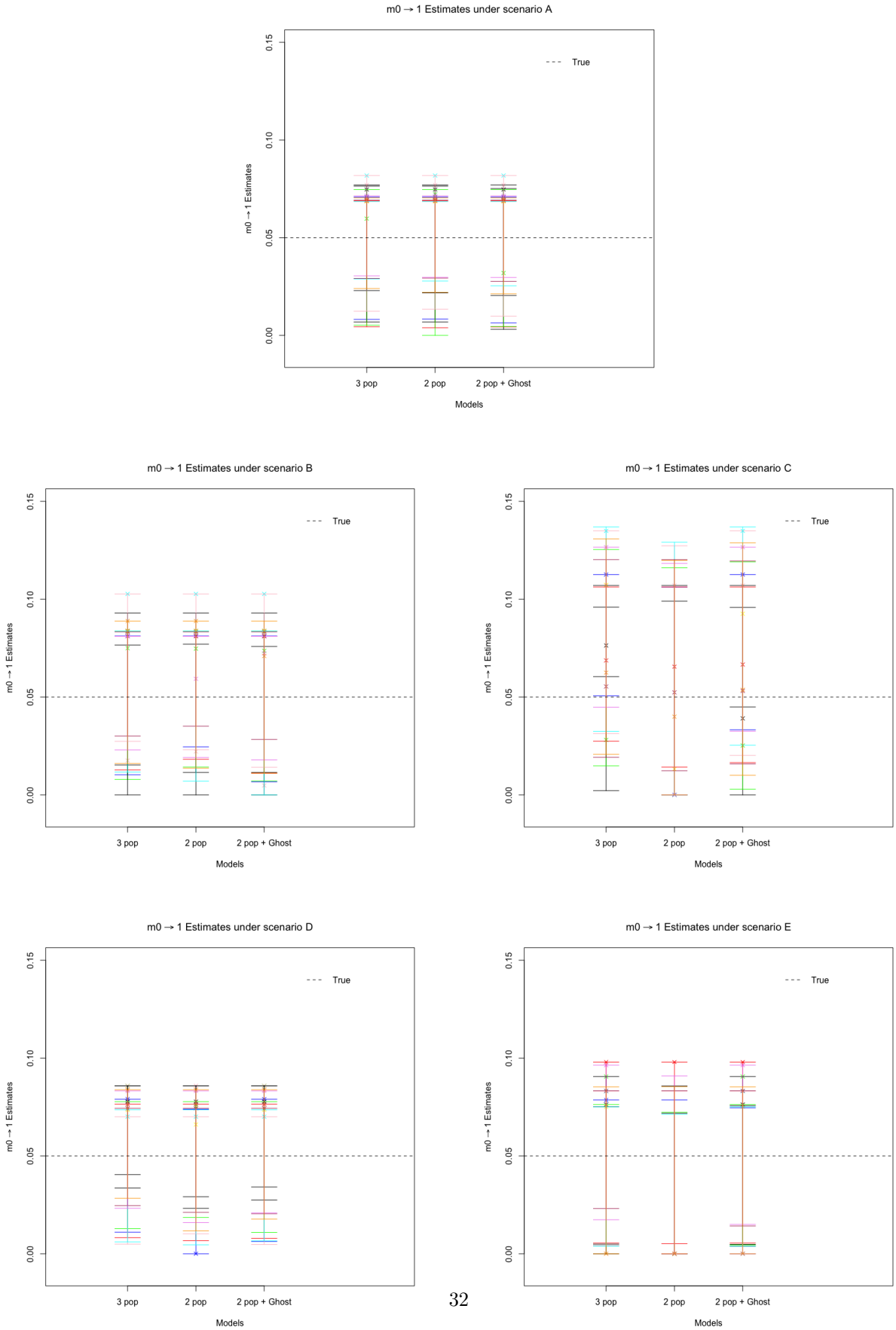


Figure 22: Scaled migration rate (m_{20}) estimate between one sampled population and the ghost, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

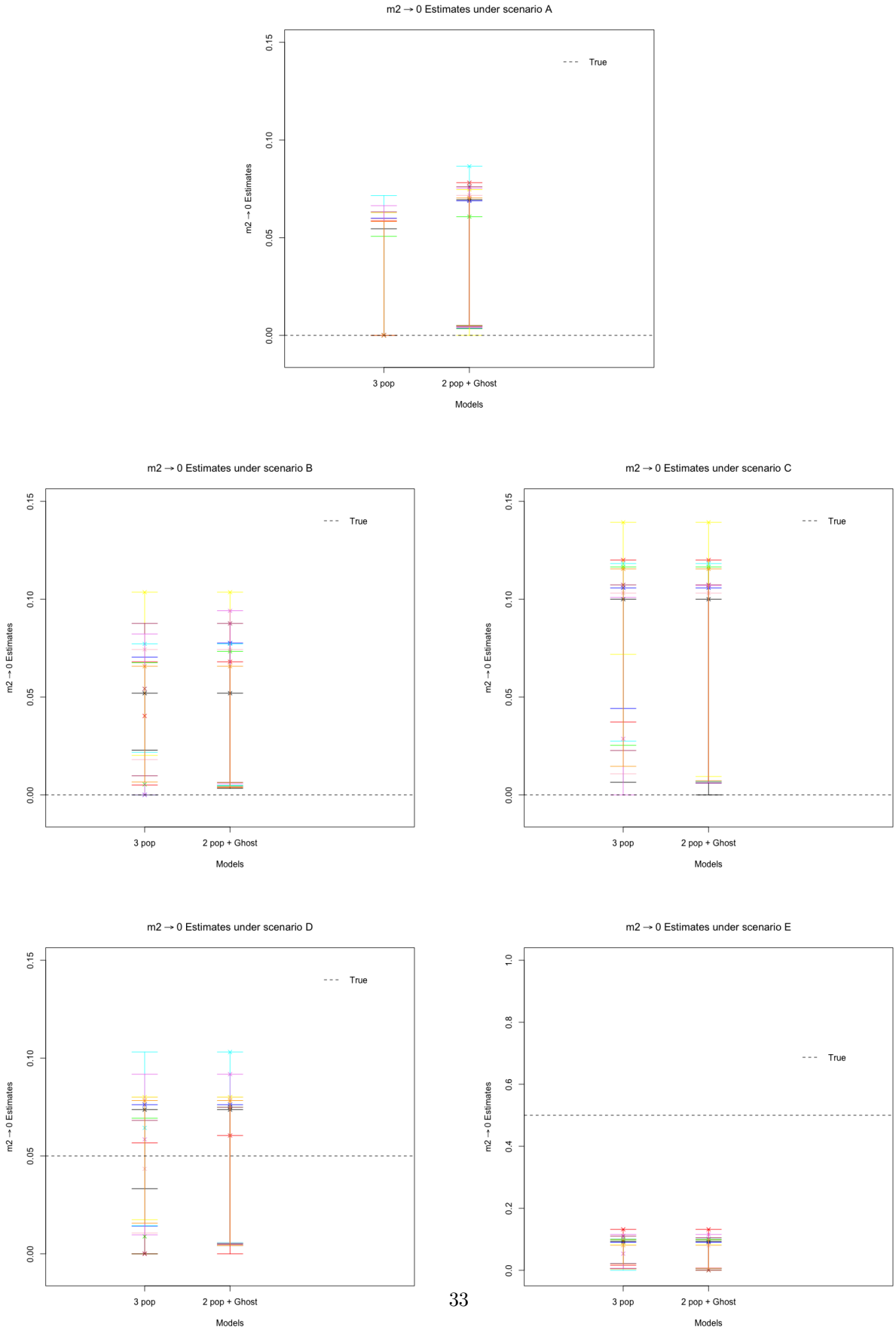


Figure 23: Scaled migration rate (m_{20}) estimate between one sampled population and the ghost, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

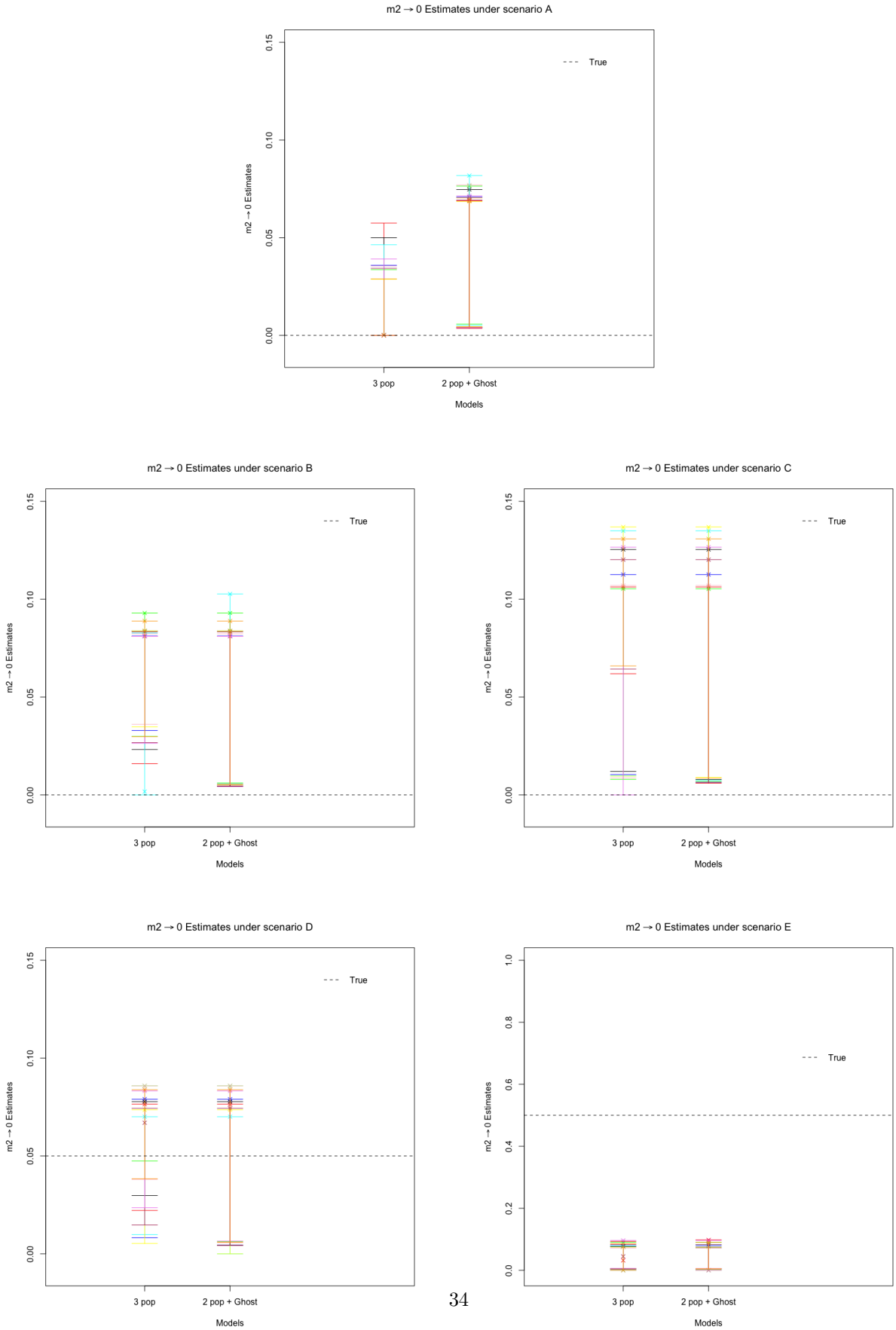


Figure 24: Scaled migration rate (m_{02}) estimate between one sampled population and the ghost, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

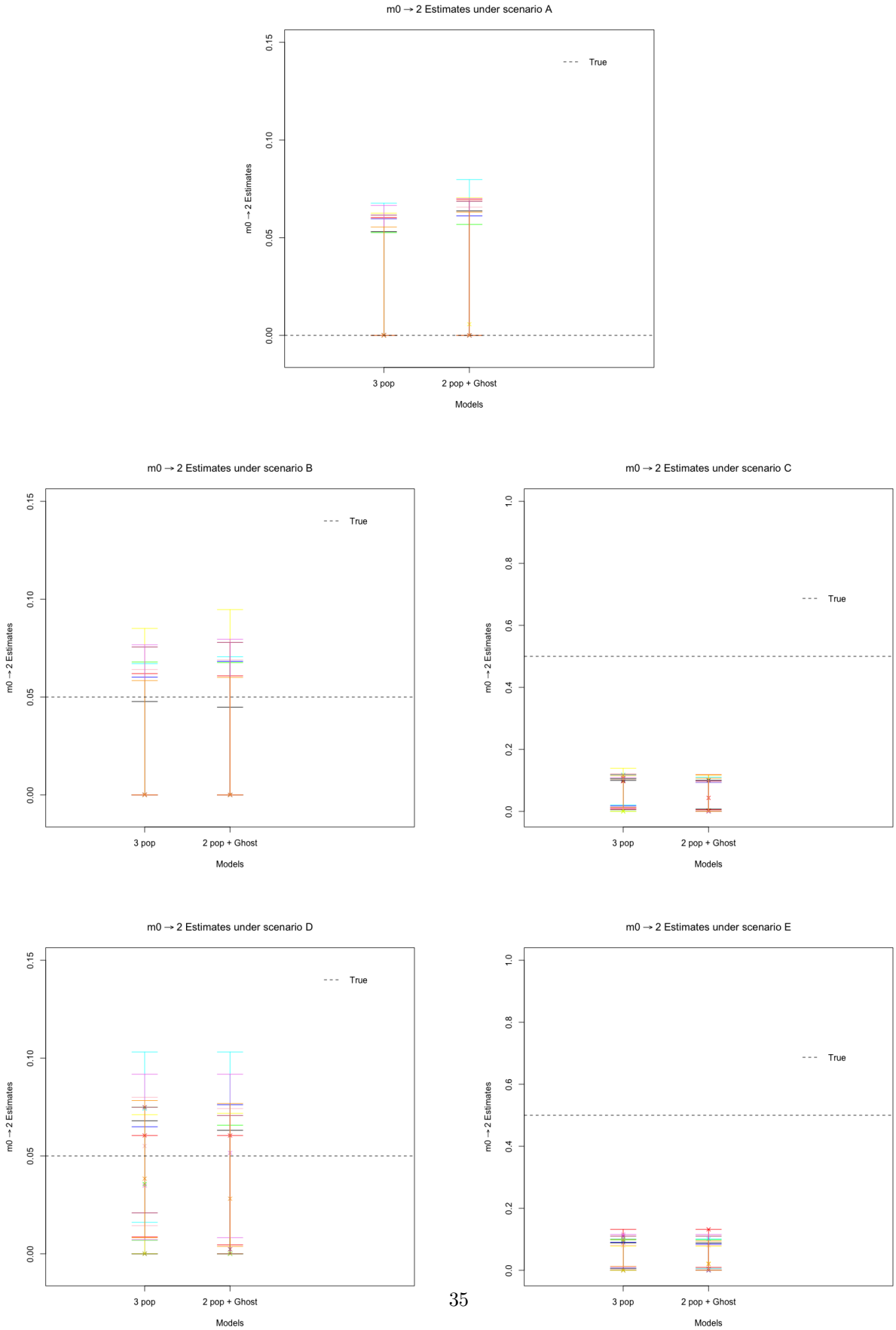


Figure 25: Scaled migration rate (m_{02}) estimate between one sampled population and the ghost, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

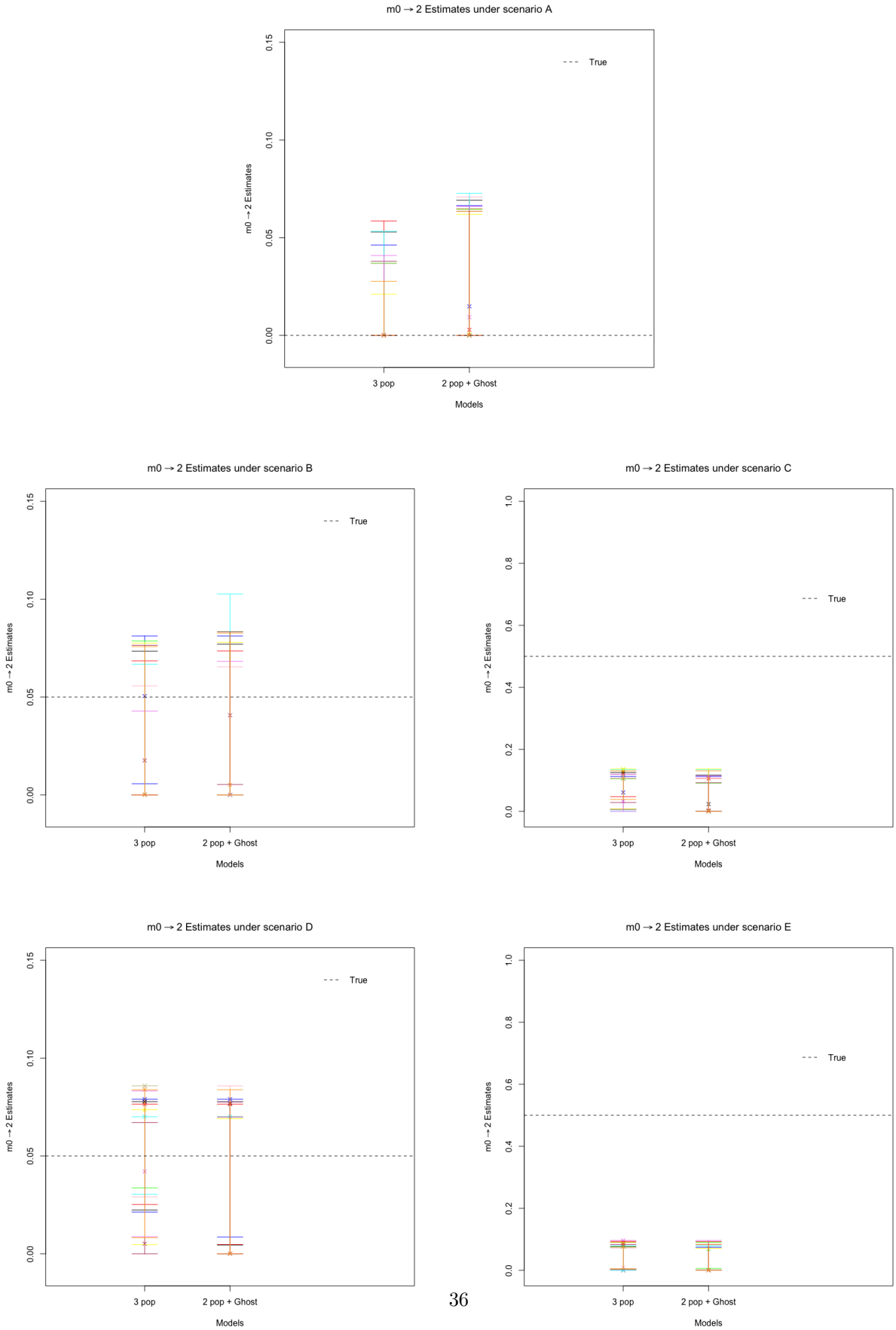


Figure 26: Scaled migration rate (m_{21}) estimate between one sampled population and the ghost, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

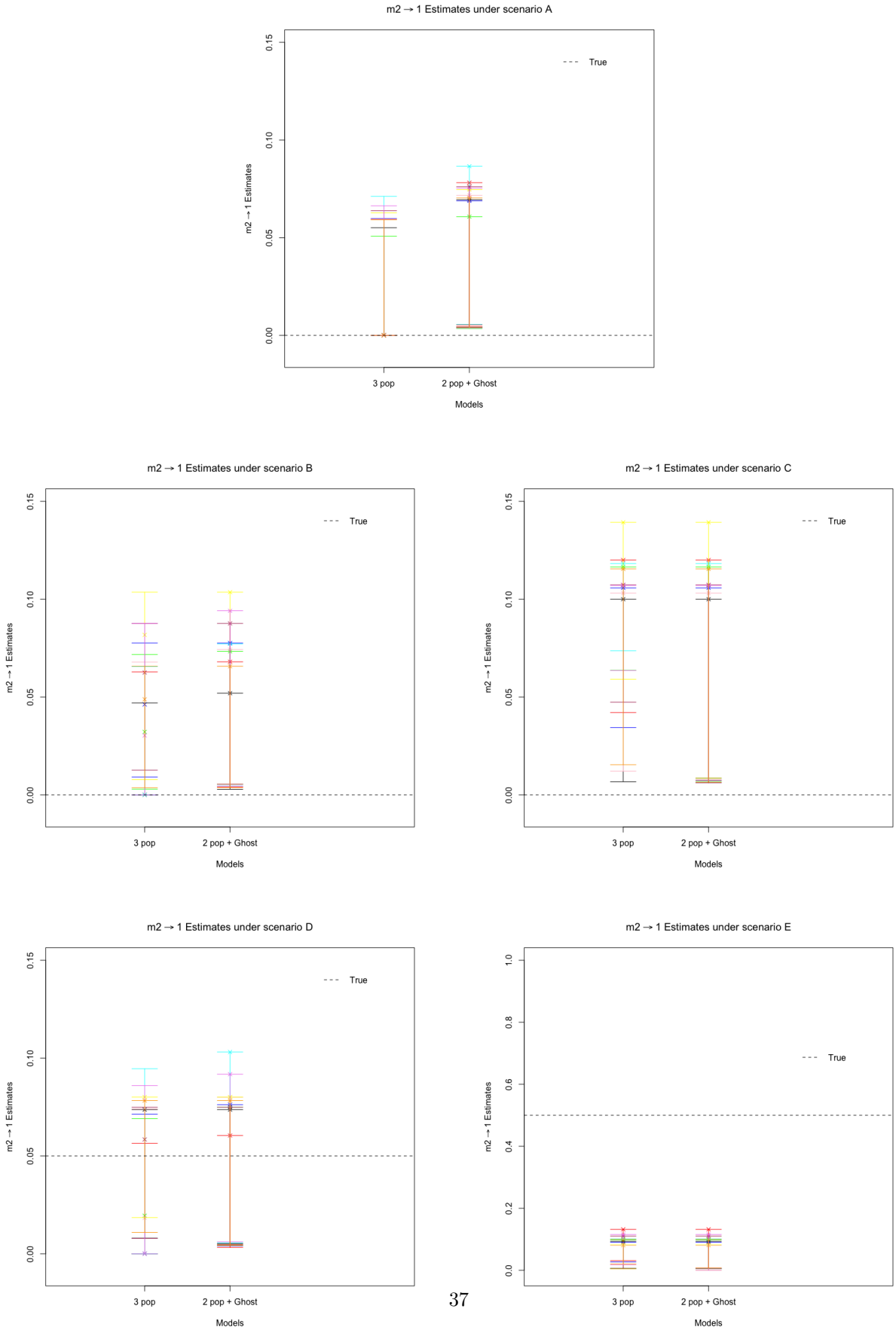


Figure 27: Scaled migration rate (m_{21}) estimate between one sampled population and the ghost, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

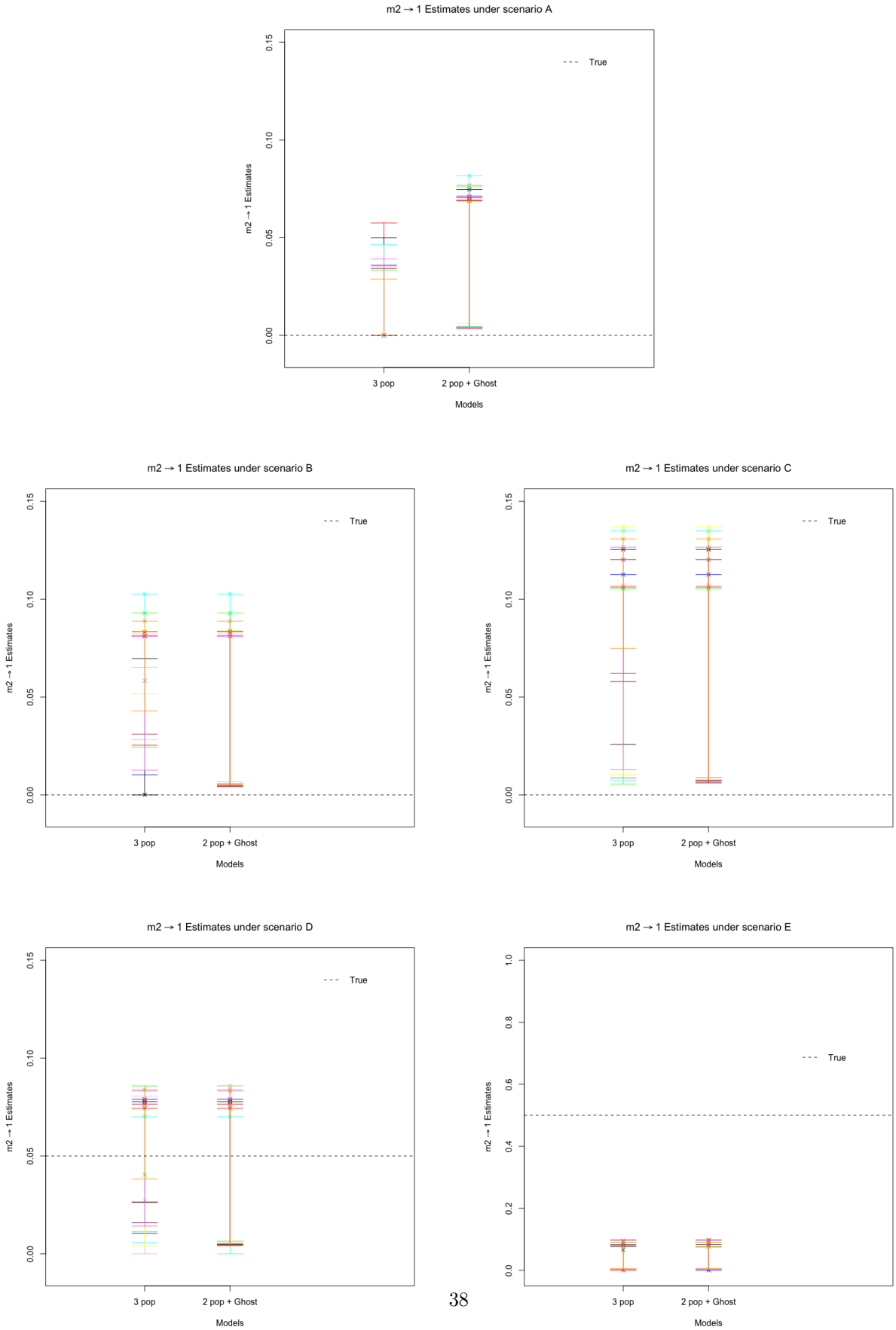


Figure 28: Scaled migration rate (m_{12}) estimate between one sampled population and the ghost, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

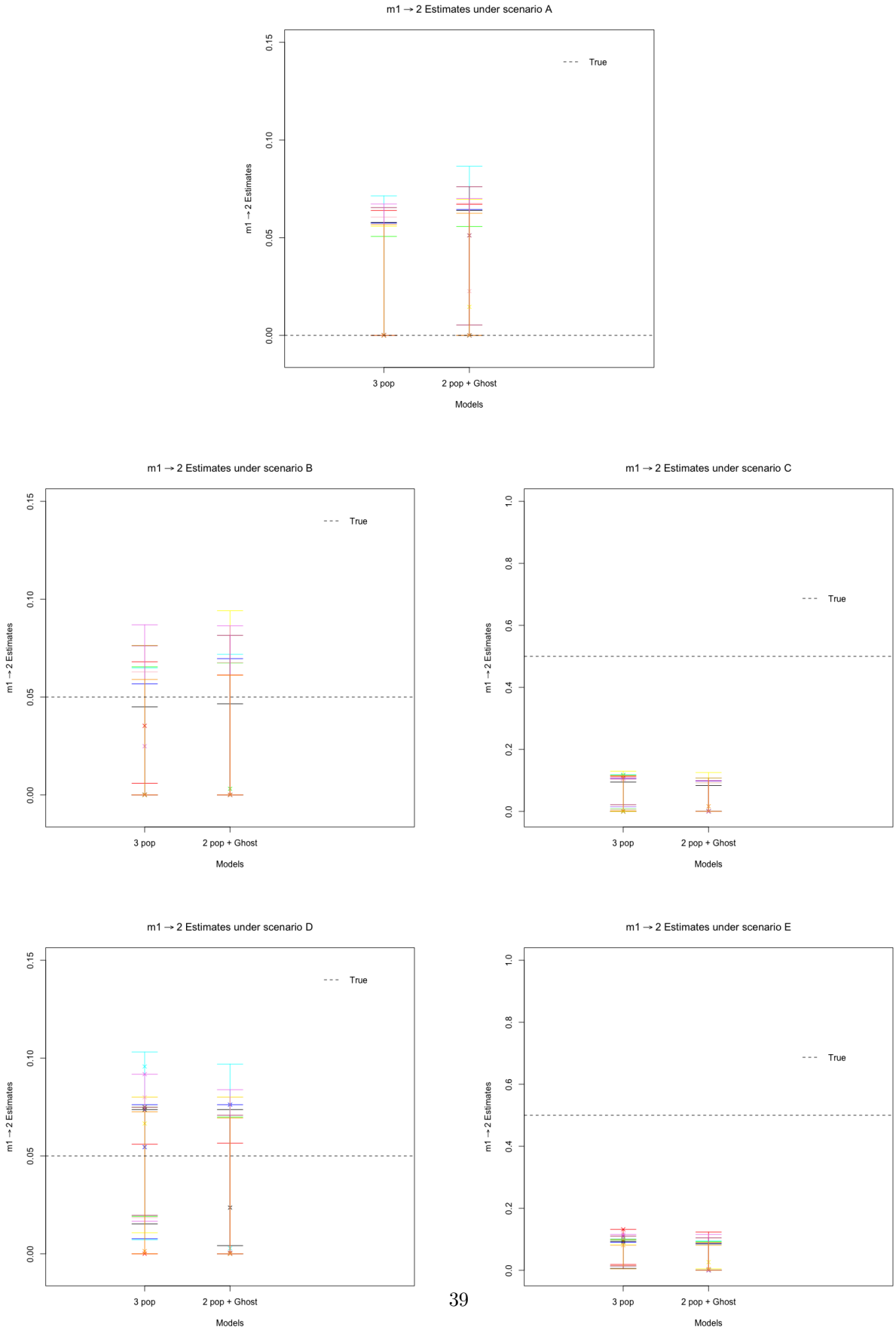


Figure 29: Scaled migration rate (m_{12}) estimate between one sampled population and the ghost, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

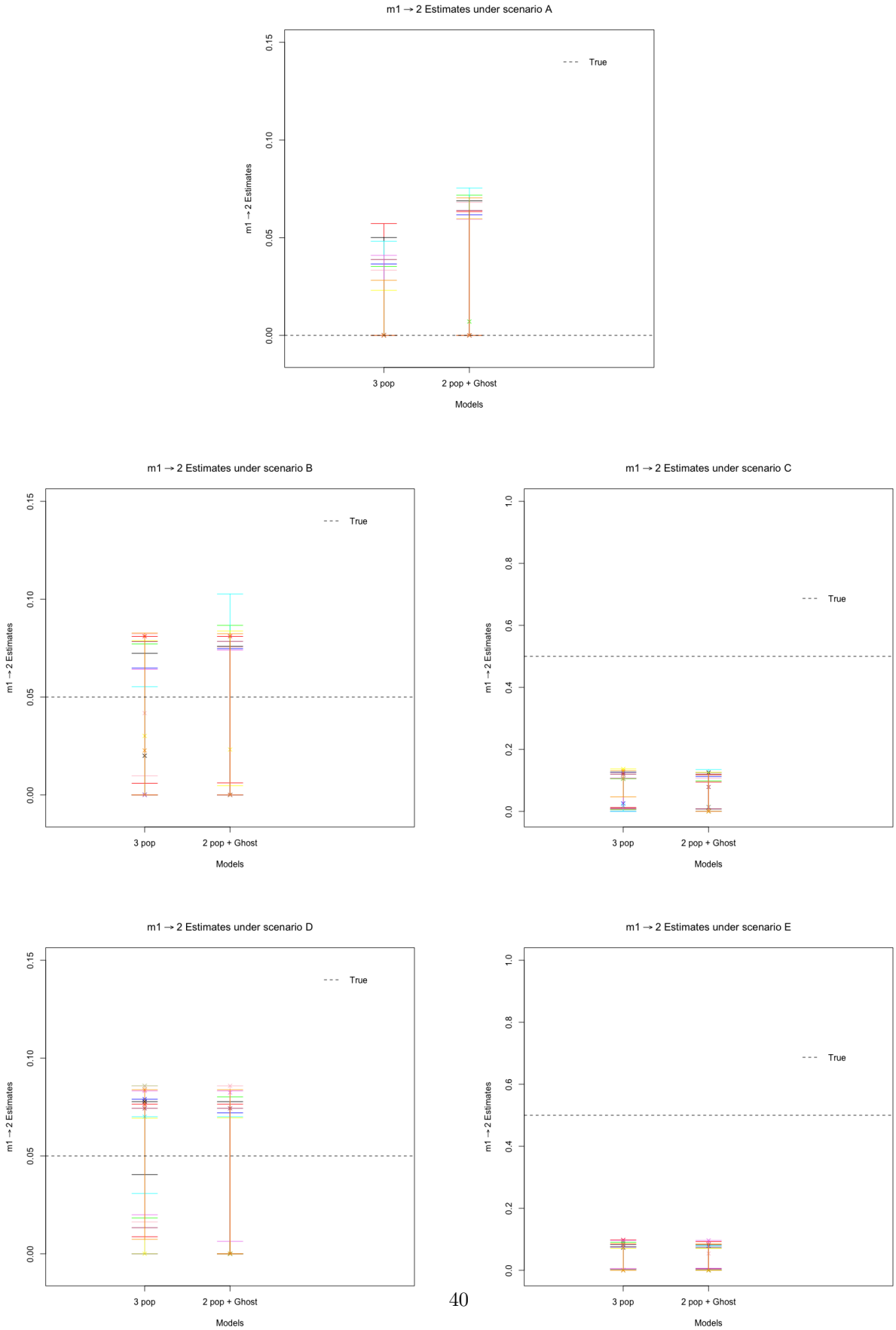


Figure 30: Scaled migration rate (m_{23}) estimate between the common ancestor of the two sampled populations and the ghost, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

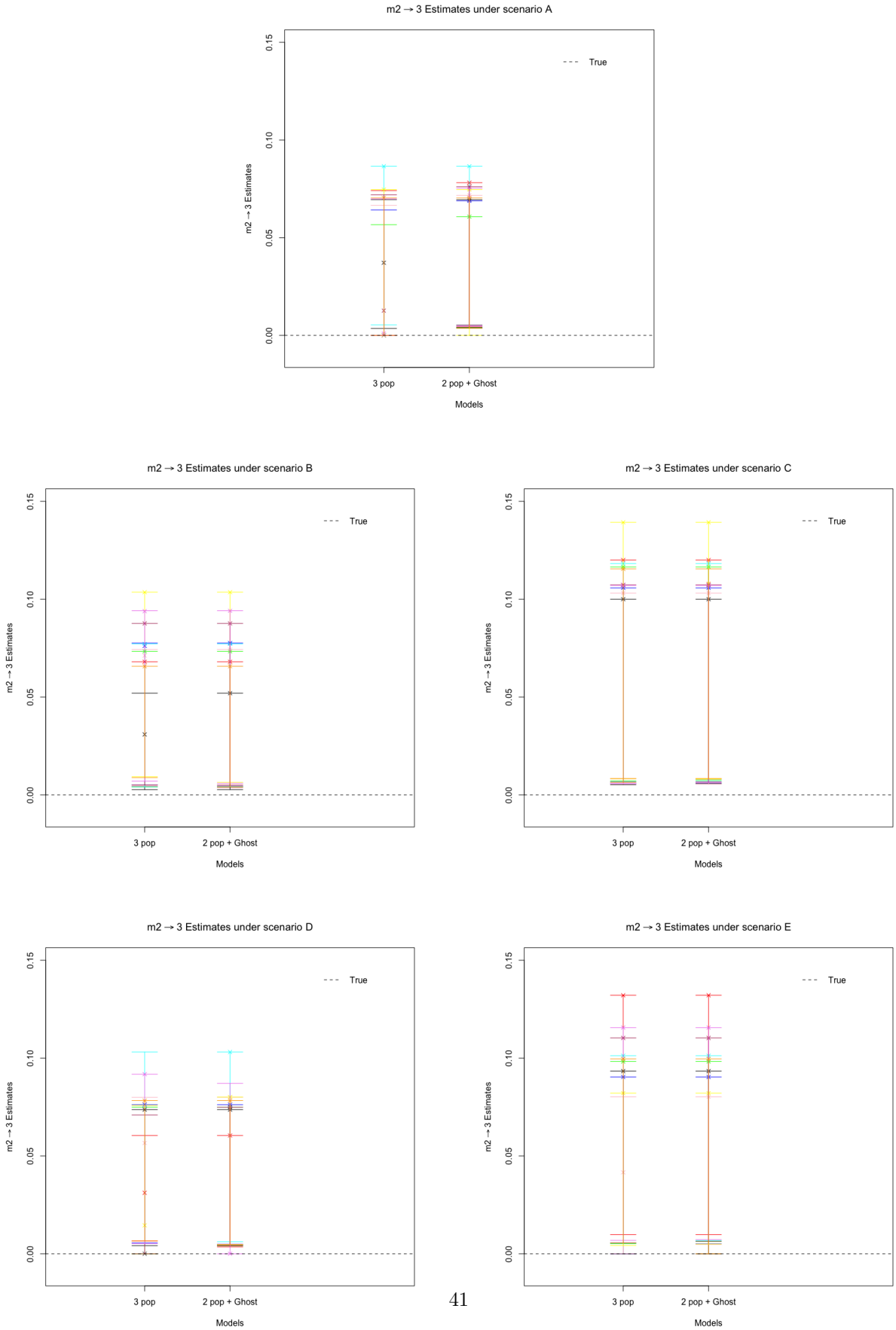


Figure 31: Scaled migration rate (m_{23}) estimate between the common ancestor of the two sampled populations and the ghost, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

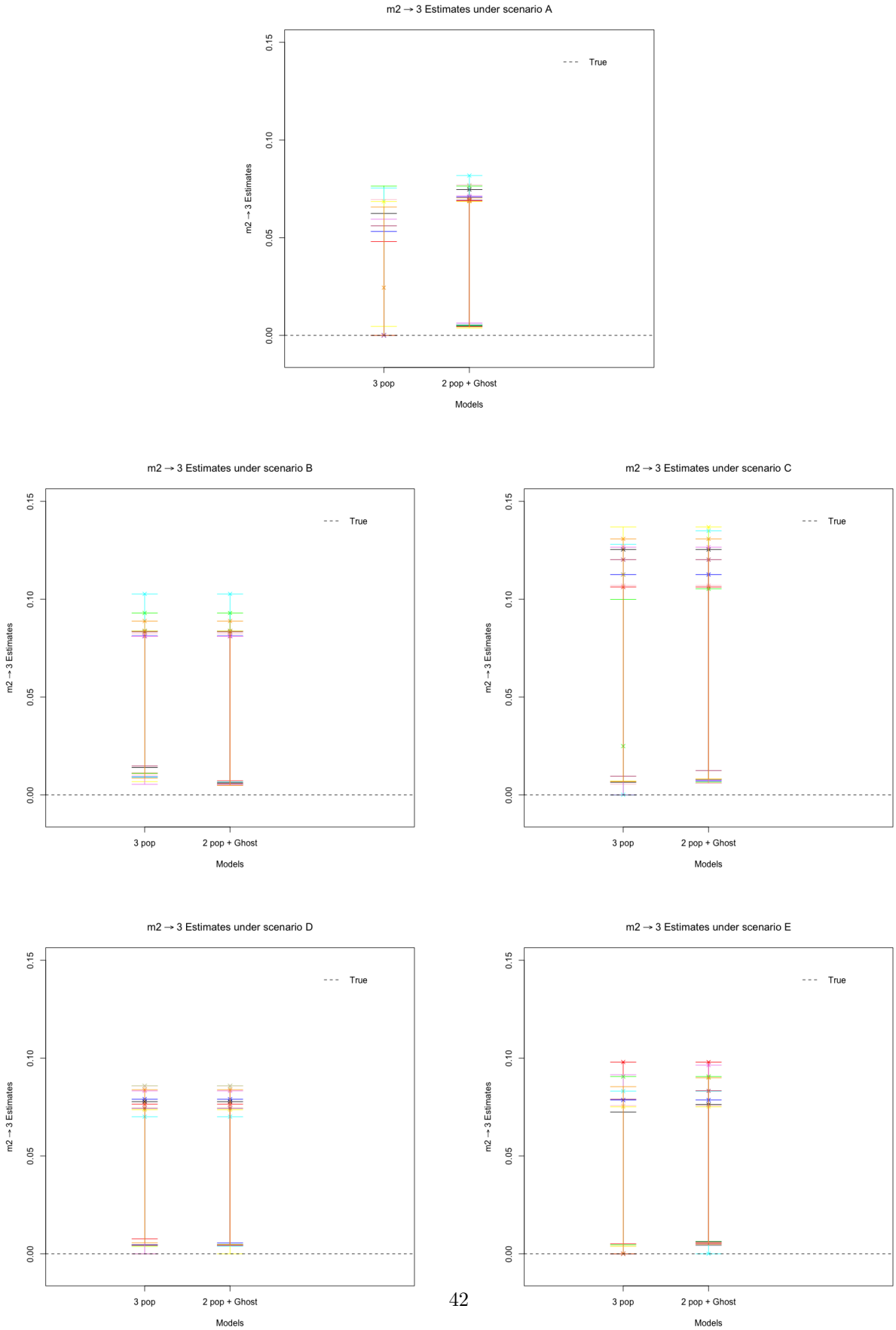


Figure 32: Scaled migration rate (m_{32}) estimate between the common ancestor of the two sampled populations and the ghost, using 2 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.

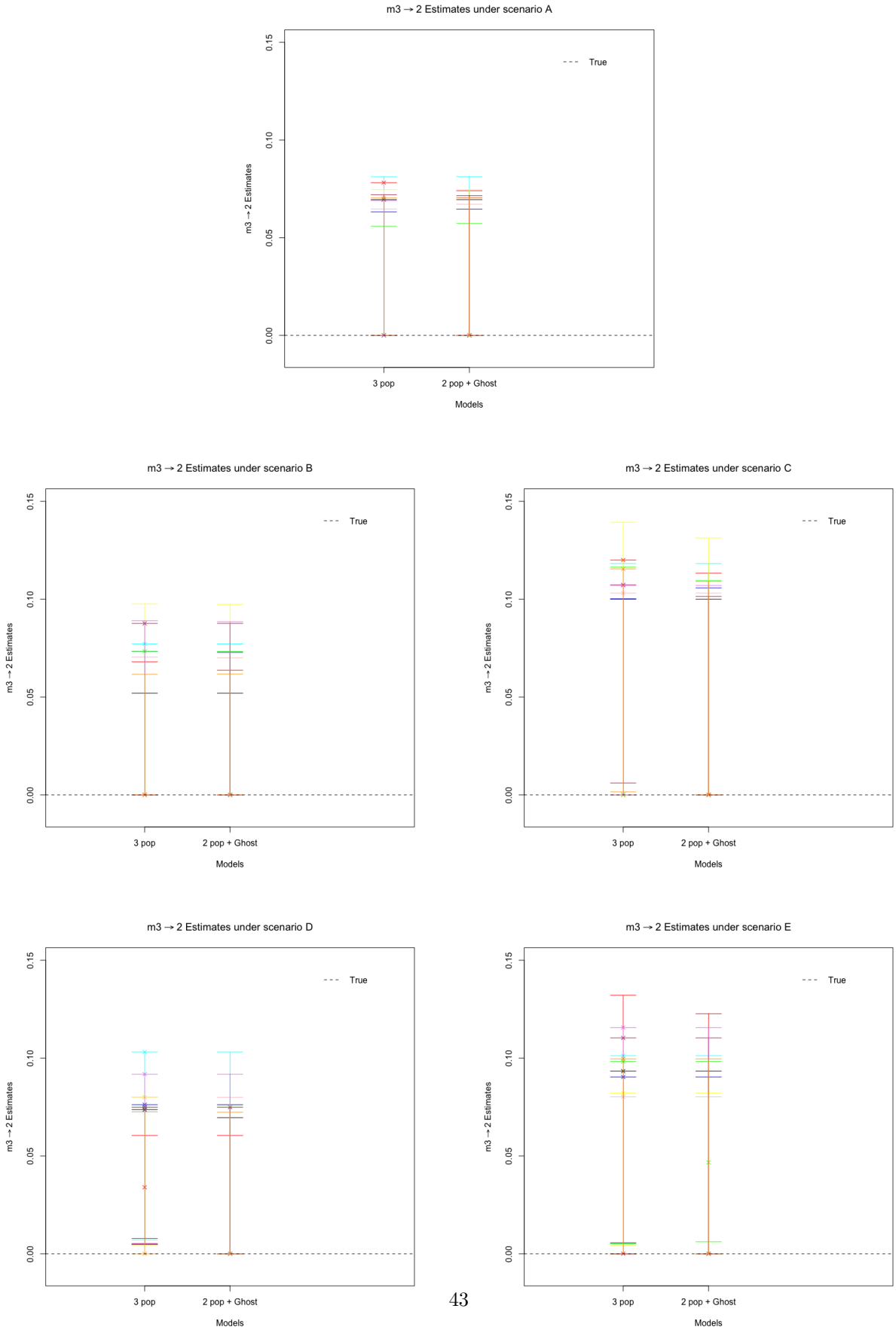
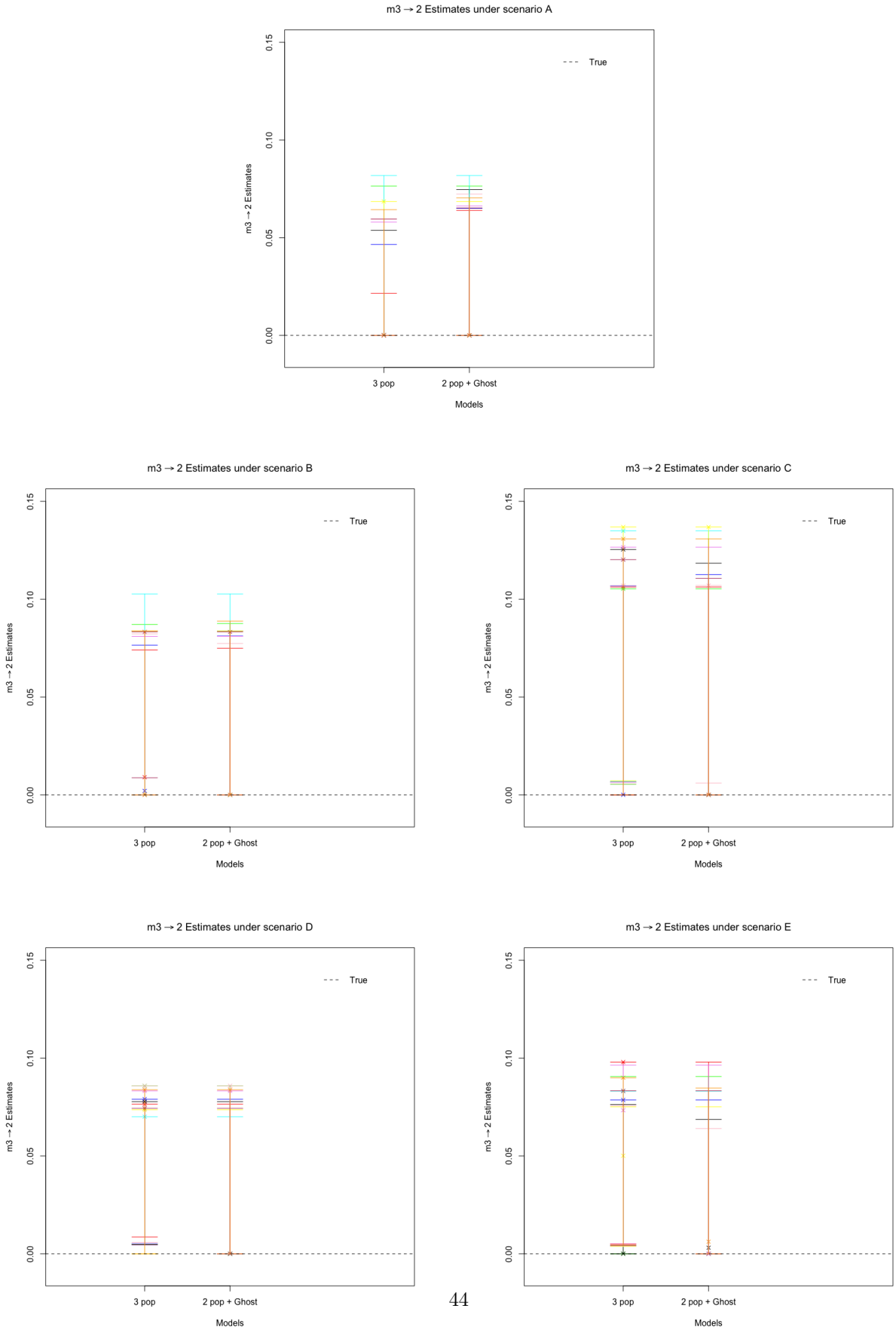



Figure 33: Scaled migration rate (m_{32}) estimate between the common ancestor of the two sampled populations and the ghost, using 5 genomic loci under scenarios A-E, estimated using a 3-population model, and a 2-population model with a ghost outgroup.