The American Journal of Human Genetics, Volume *109*

# Supplemental information

# Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls

**Siddharth Avadhanam and Amy L. Williams**

# Supplemental Data for Simultaneous inference of parental admixture proportions and admixture times from unphased local ancestry calls

Siddharth Avadhanam[1] and Amy L. Williams[1,*]

[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA

| **Scenarios**<br>**Settings** | **Scenario (1)** | **Scenario (2)** | **Scenario (3)** |
|---|---|---|---|
| **Admixture time** | $t_A = t_B = t$; $t \in \{2, 9\}$ | $t_A > t_B$; $t_A = 9, t_B \in \{2, 8\}$ | $(t_A, t_B) \in \{(0, 0), (1, 1)\}$ |
| **Admixture proportions and number of individuals** | 22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0.25, 0.5), (0.4, 0.6), (0.25, 0.75)\}$ | 22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0.25, 0.5), (0.4, 0.6), (0.25, 0.75)\}$ | 22 individuals for each of $E[(p_A, p_B)] \in \{(0, 0), (1, 1)\}$;<br>44 individuals for $E[(p_A, p_B)] = (1, 0)$;<br>22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0, 0.5)\}$ |

Table S1: **Simulation scenarios.** Summary of all the admixture scenarios simulated in this paper. Columns are the scenario labels, the first row gives the admixture time settings, and the second row provides the number of individuals and admixture proportion setting simulated *for each* setting of $(t_A, t_B)$.

| **Model Setting**<br>**Inference Mode** | **Binomial** | **Binomial w/Error Model** | **HMM** | **HMM w/Error Model** | **Full** | **Full w/Error Model** |
|---|---|---|---|---|---|---|
| **GD** | 7.3s | – | 21.7s | 23.6s | 23.9s | 27.0s |
| **MCMC** | 5m 23s | – | 8h 36m 49s | 8hr 58m 31s | 8hr 5m 58s | 8hr 55m 6s |

Table S2: **PAPI runtimes for each model and inference mode.** Times are the per-individual average from analyzing simulated samples with $t_A = t_B = 5$, including 22 individuals for each of $E[(p_A, p_B)] \in \{(0.5, 0.5), (0.25, 0.5), (0.4, 0.6), (0.25, 0.75)\}$, for a total of 88 individuals per run. Table gives wall clock times from compute nodes with Intel Xeon E5 4620 processors utilizing 4 GB of RAM.

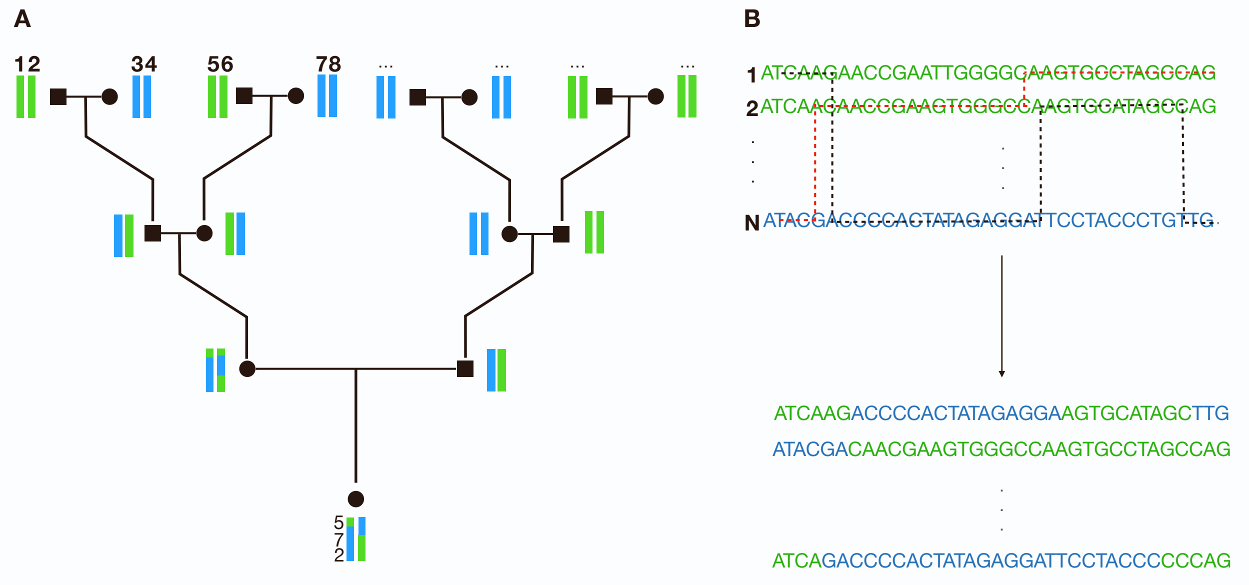


Figure S1: **Overview of simulation procedure.** (A) Given a pedigree topology (specified by $(t_A, t_B)$), Ped-sim generates a breakpoint file for the simulated samples containing recombination break points between founder haplotypes, each of which have a unique id number. Subsequently, we assign these haplotype ids to population labels according to the desired values of $E[(p_A, p_B)]$ and $(t_A, t_B)$. In the example depicted here, $(t_A, t_B) = (2, 1)$ and $E[(p_A, p_B)] = (0.5, 0.5)$, where parent $A$ is on the left and $B$ is on the right. (B) Given the adapted breakpoint file containing population labels, admix-simu generates offspring haplotypes by sampling non-overlapping paths (black and red dashed lines representing the sampling paths of two haplotypes) from the input unadmixed haplotypes.

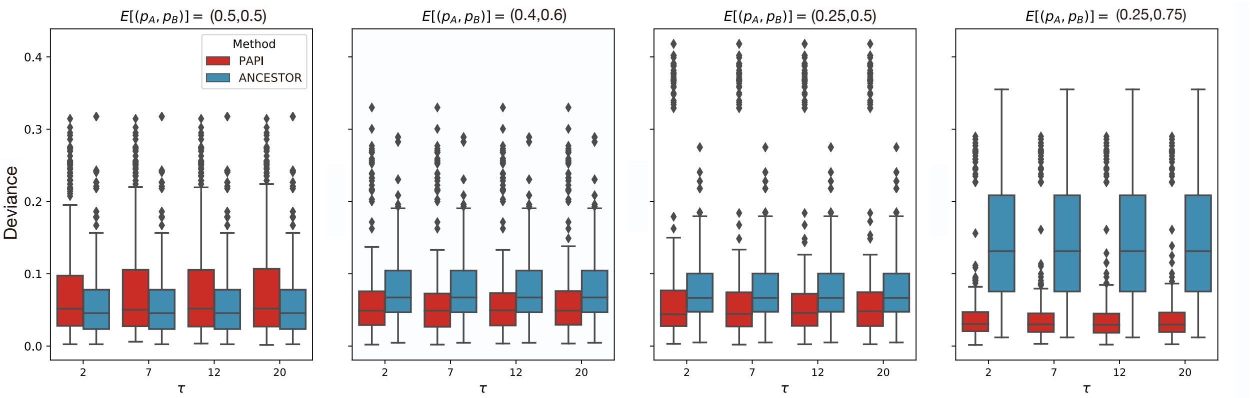


Figure S2: **Comparison of PAPI's admixture proportion estimates using the full model under different settings of $\tau$.** Each panel displays box plots of $d_{abs,p}$ from PAPI and ANCESTOR versus $\tau$ for one of the settings of $E[(p_A, p_B)]$ for all scenario (1) and (2) data. PAPI is strongly robust to variable specification of $\tau$. Note that ANCESTOR results are the same for each x-axis tick, as ANCESTOR does not utilize $\tau$.

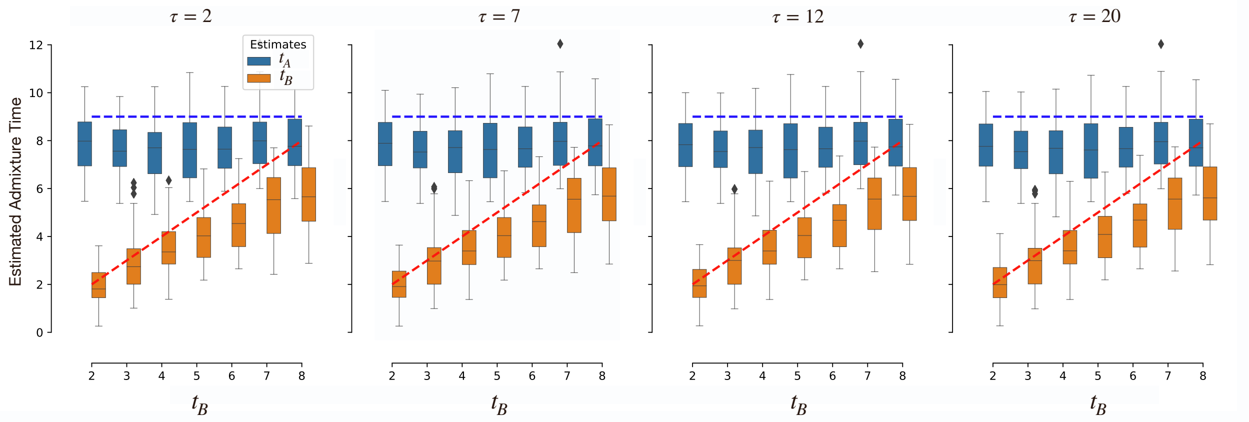


Figure S3: **Comparison of PAPI's admixture time estimates using the full model under different settings of $\tau$.** Each panel displays box plots of PAPI's time estimates versus $t_B$ for a different setting of $\tau$ for all scenario (1) datapoints. PAPI's results are largely unchanged regardless of the specification of $\tau$.

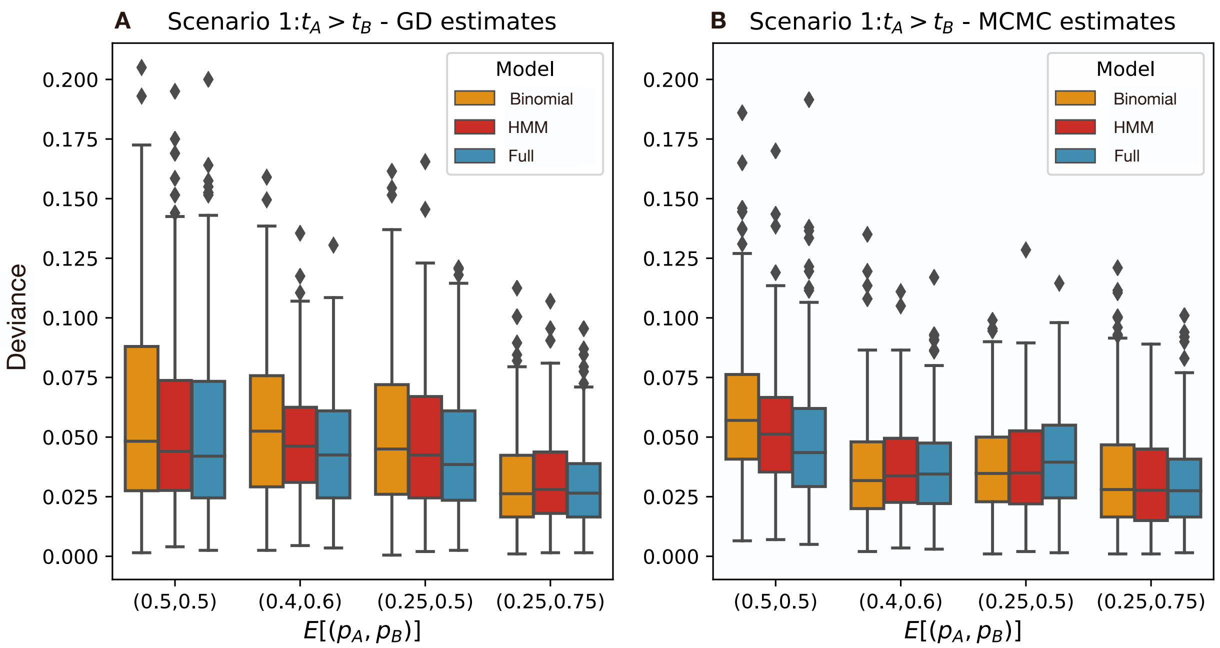


Figure S4: **Deviance statistics for inferred parent ancestry proportions** $(p_A, p_B)$ **for each of PAPI's component models.** Average absolute deviances ($d_{abs,p}$) on all scenario (1) data under the (A) GD and (B) MCMC inference modes.

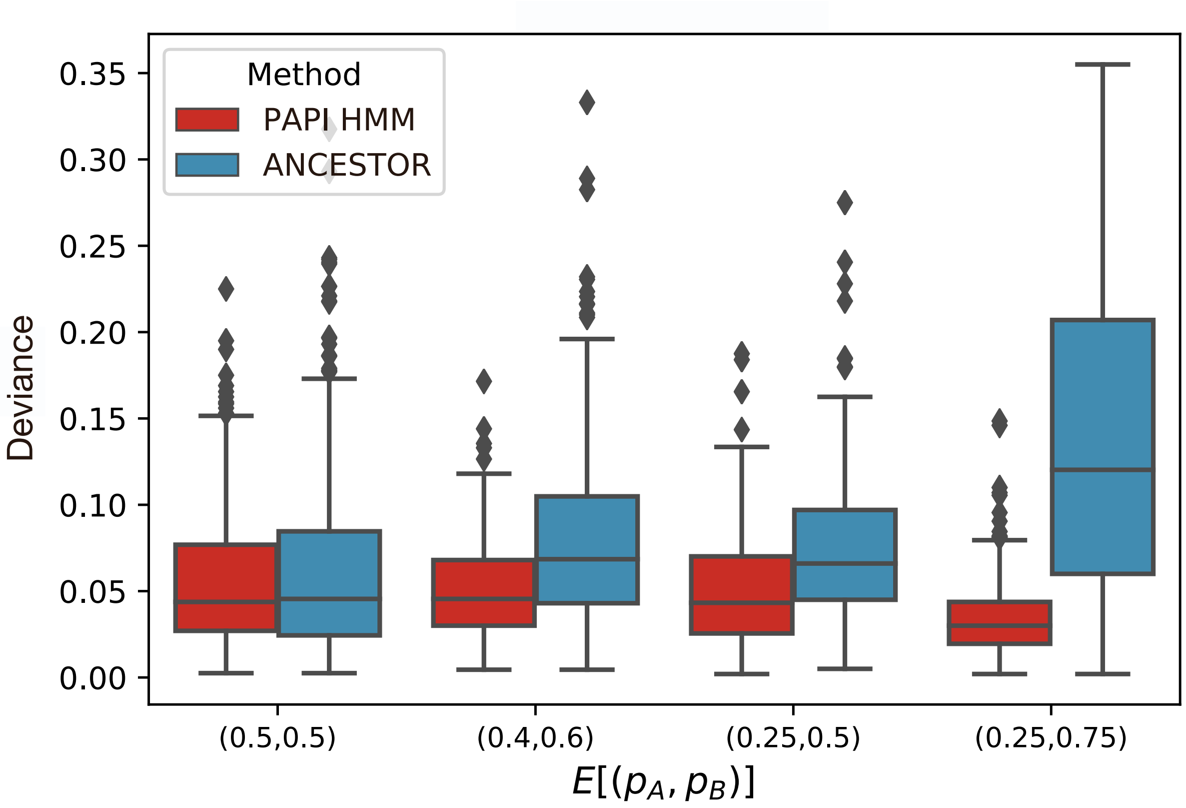


Figure S5: **Deviance statistics for inferred parent ancestry proportions** $(p_A, p_B)$ **from ANCESTOR and PAPI's HMM alone for all scenario (1) and (2) data points.**

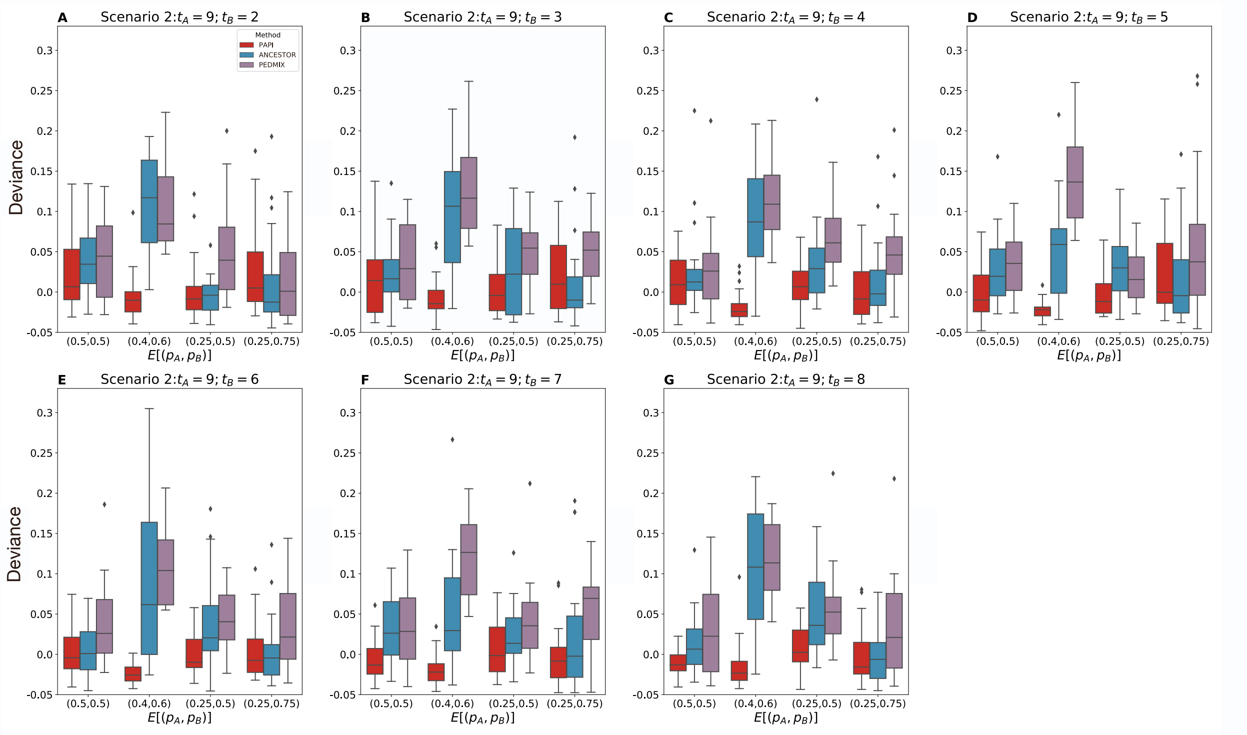


Figure S6: **Deviance statistics for inferred parent ancestry proportions** $(p_A, p_B)$ **from PAPI, ANCESTOR and PedMix for each scenario (1) parameter setting.** (A-G) Average absolute deviances $(d_{abs,p})$ for each value of $(t_A, t_B)$ as indicated in the panel title.

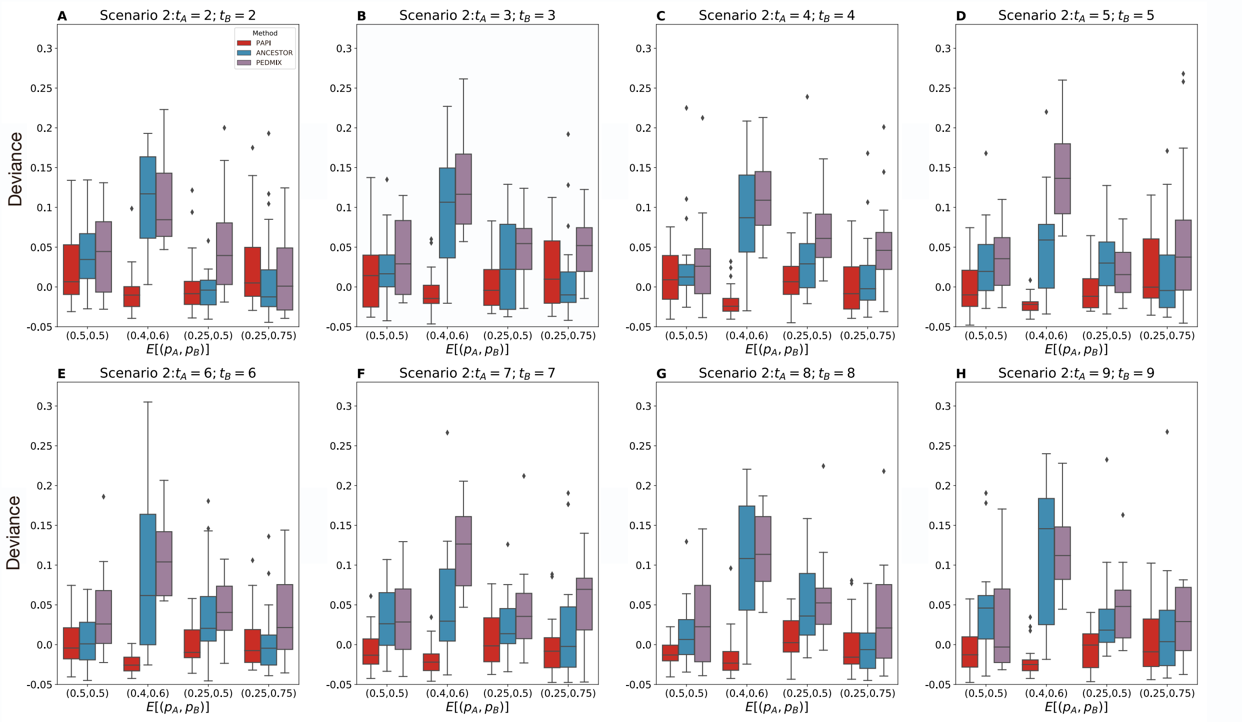


Figure S7: **Deviance statistics for inferred parent ancestry proportions** $(p_A, p_B)$ **from PAPI, ANCESTOR and PedMix for each scenario (2) parameter setting.** (A-H) Average absolute deviances $(d_{abs,p})$ for each value of $(t_A, t_B)$ as indicated in the panel title.

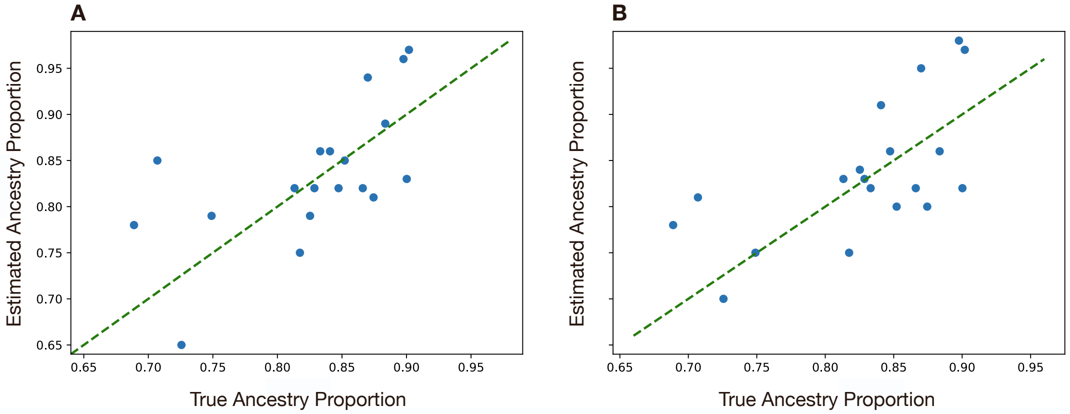


Figure S8: **Accuracy of admixture proportion estimates for PAPI's binomial model versus PAPI's HMM in the HapMap ASW trio children.** Scatter plots of estimated versus true parent ancestry proportions $p_j$ for (A) estimates from PAPI's binomial model (B) estimates from PAPI's HMM.

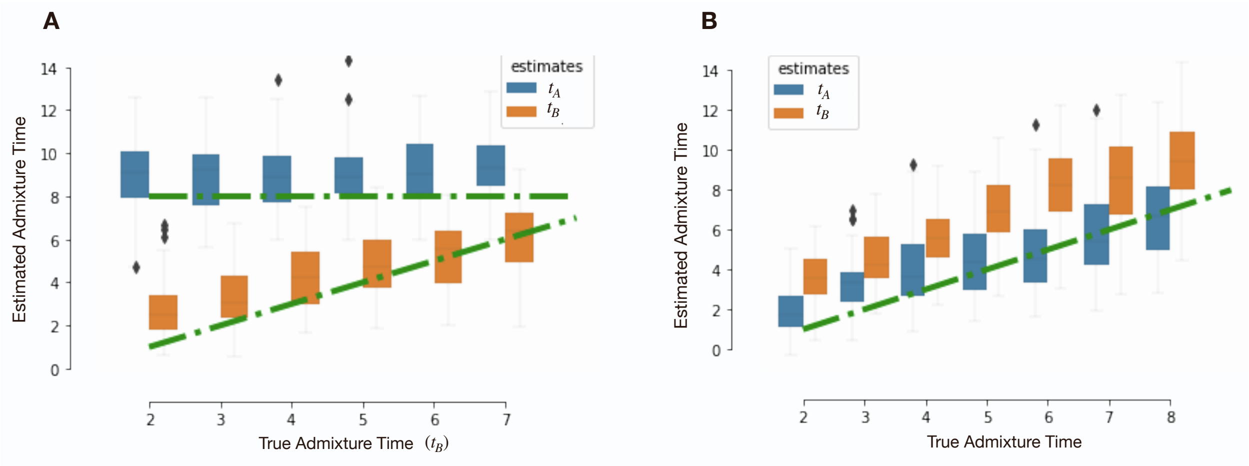
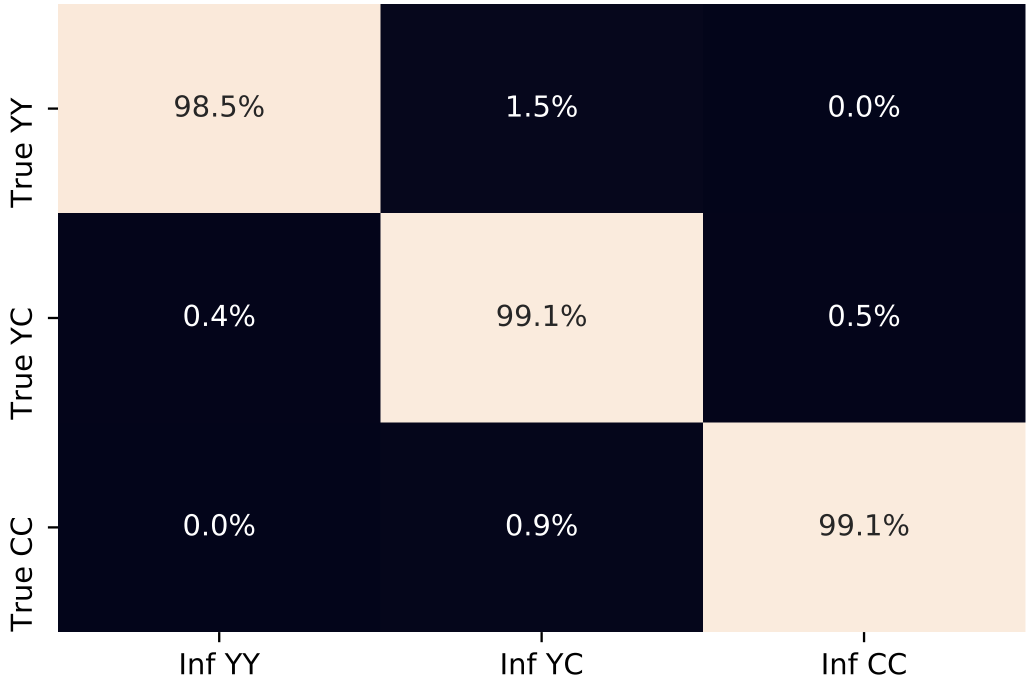


Figure S9: **PAPI's estimated time since admixture in simulated individuals using local ancestry tracts inferred by LAMP-LD.** Box plots of estimated times for (A) scenario (1) samples (where $t_A > t_B$) and (B) scenario (2) data (where $t_A = t_B$).



| | Inf YY | Inf YC | Inf CC |
|---|---|---|---|
| True YY | 98.5% | 1.5% | 0.0% |
| True YC | 0.4% | 99.1% | 0.5% |
| True CC | 0.0% | 0.9% | 99.1% |

Figure S10: **Per-marker confusion matrix of true diploid ancestries versus HAPMIX-inferred diploid ancestries.** HAPMIX shows a high true positive rate (> 98.5%). Matrix includes across all scenario (1) and (2) simulated data points. The rows are labeled true ancestry states, and the columns are inferred ancestry states, with Y and C representing HapMap YRI (African) and CEU (European) ancestry, respectively. For example, the cell at row 1, column 1 contains the percentage of true YY calls (diploid African) that were inferred to be YY.

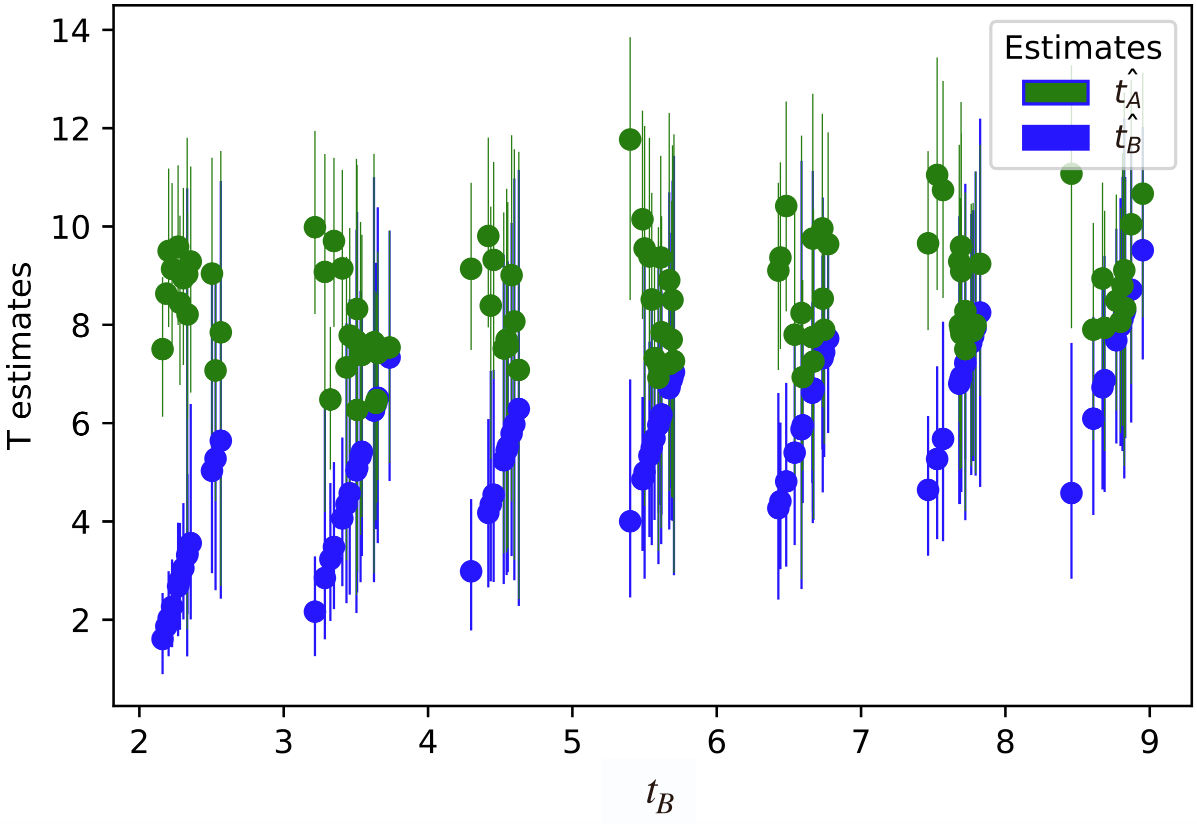


Figure S11: **Admixture time estimates with credible intervals for scenario (1) data points.** Plot shows point estimates along with their 90% credible intervals from PAPI's MCMC inference mode run under the full model. The x-axis coordinate for each point is shifted slightly from the truth to aid visualization: the estimate pairs $(\hat{t}_A, \hat{t}_B)$ are sorted by the value of $\hat{t}_B$.

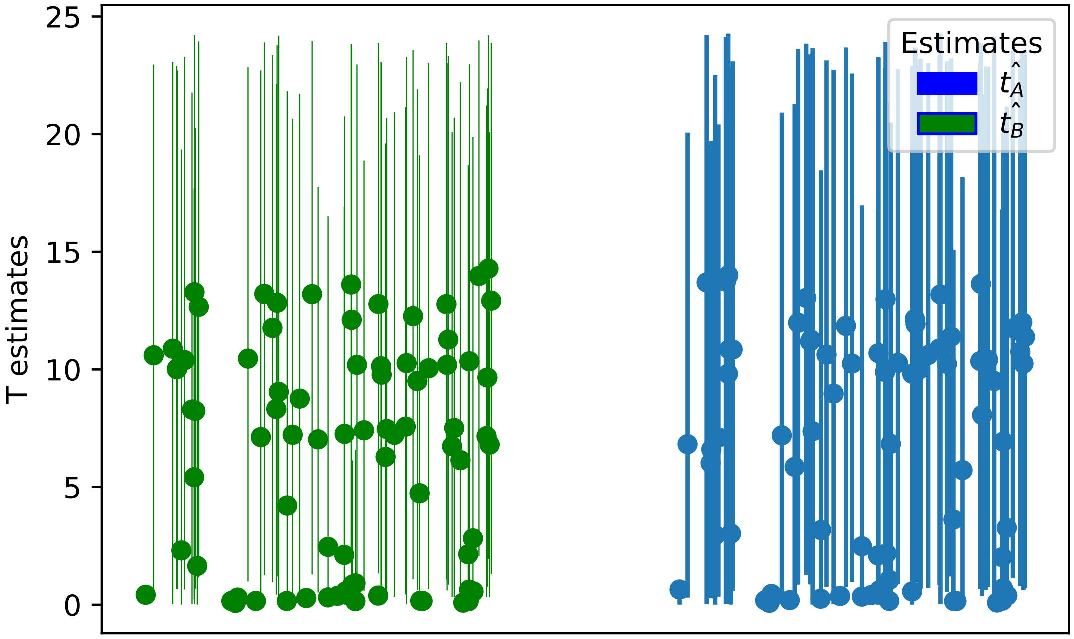


Figure S12: **Admixture time estimates with credible intervals for the subset of scenario (3) data points with $t_A = t_B = 0$.** Plot shows point estimates along with their 90% credible intervals from PAPI's MCMC inference mode run under the full model. Many estimates are biased and have very large 90% credible intervals because of a small number of erroneous tracts.

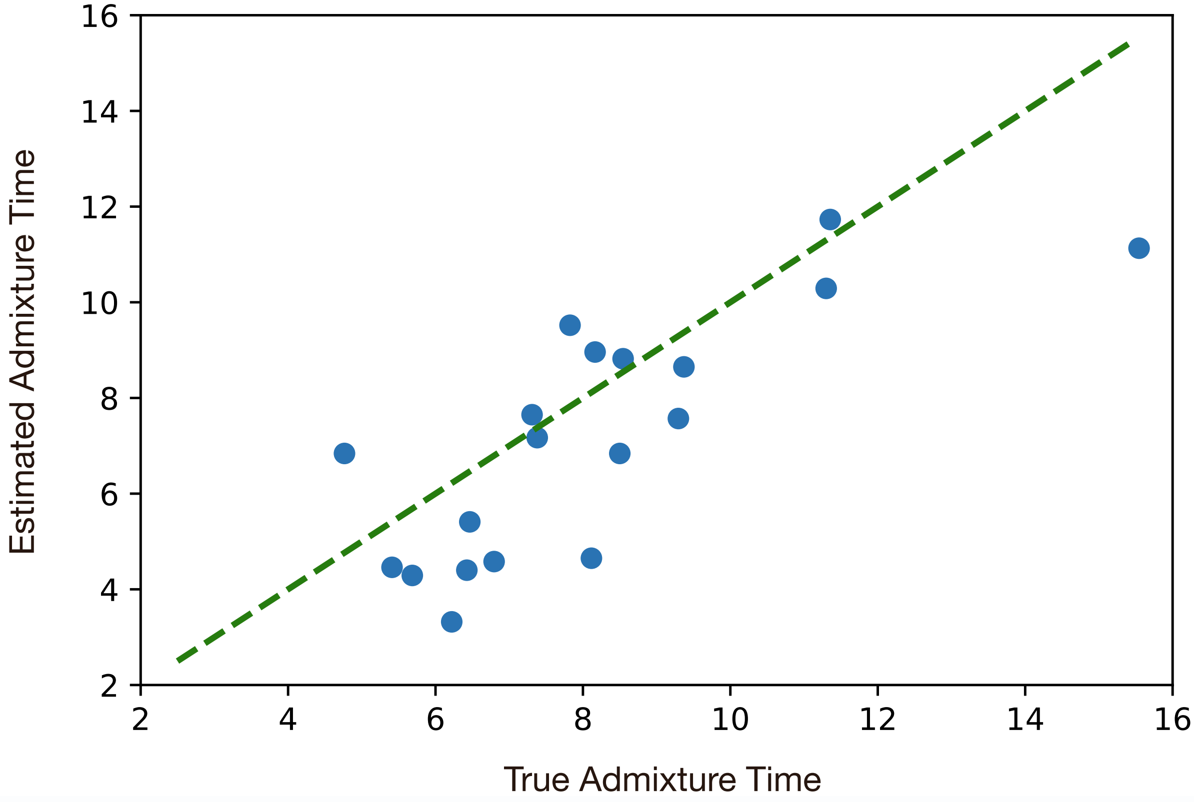


Figure S13: **Accuracy of admixture time estimates for PAPI's full model in the HapMap ASW trio children.** Scatter plots of estimated versus true parent admixture proportions $t_j$ for PAPI's full model.

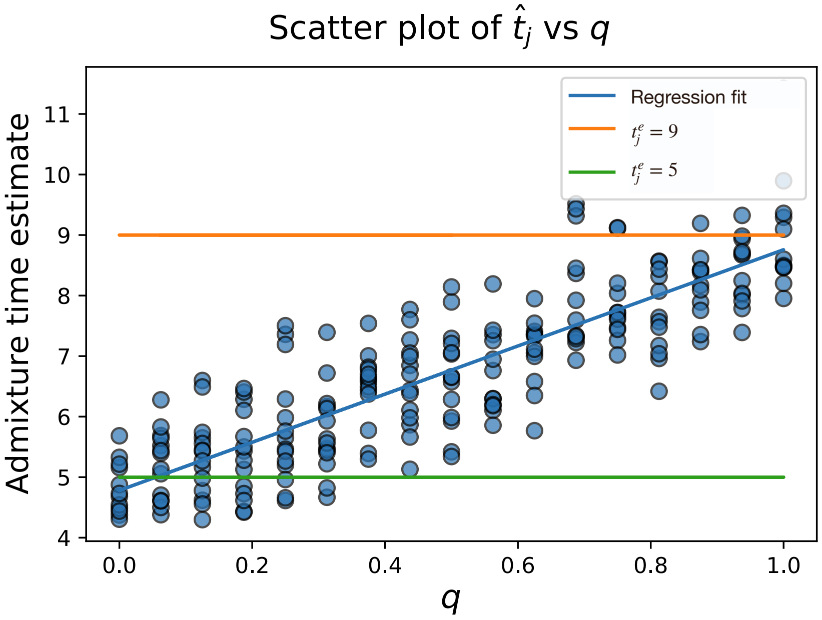


Figure S14: **PAPI's estimated time since admixture with two migrant pulses where $t_j^e = 9$ and $t_j^l = 5$.** Plot of estimated time since admixture versus $q$ (the proportion of unadmixed couples in the $t_j^e$ generation of different ancestries) when $E[p_j] = 0.5$. The blue line depicts the least-squares regressions on the plotted data points.

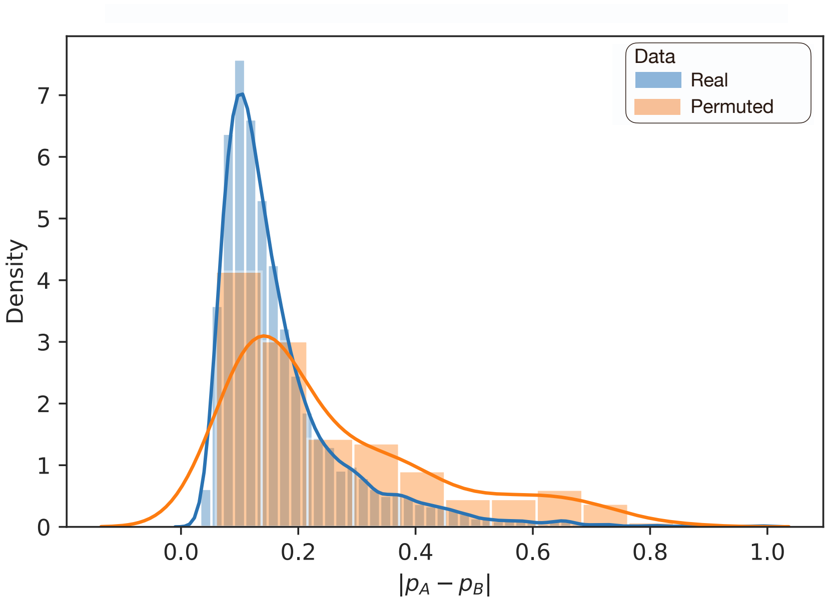


Figure S15: **Real versus simulated random mating distribution of $|p_A - p_B|$ in the PAGE dataset.** The orange distribution uses parent ancestry proportions randomly sampled from the inferred $\hat{p}_j$ values from the PAGE African Americans.