# SUPPLEMENTAL TEXT

## INCORPORATING UNEVEN POOLING OF INDIVIDUALS PRODUCES MORE REALISTIC ESTIMATES OF TRUE ALLELE FREQUENCIES

Our ability to measure the accuracy of HAFs and raw AFs depends on our ability to determine the true contribution of each pooled individual. Since uneven pooling is a source of error known to affect pool-seq samples[15](https://paperpile.com/c/5dF0nm/QwCFZ), we estimated the relative contribution of DNA from each individual by calculating the average genome-wide allele frequency at sites private to each founder. While each founder could be detected in the pool, we found substantial variation in their relative representation ([Supp. Fig. 3](#_qvu4he5v6nu8)). ‘True’ frequencies for the experimental pooled sample were thus calculated by weighting founders known to contain the alternate allele by their estimated representation in the pool. We assessed whether these ‘true’ allele frequencies were better recapitulated by experimental reads than ‘true’ allele frequencies calculated without incorporating uneven pooling at all fully genotyped sites (both private and common). We found that the effective coverage using unevenly pooled weighted values (126x) was higher than the effective coverage assuming evenly pooled individuals (120x). We used these same estimates of uneven pooling to simulate reads in uneven proportions from different haplotypes for the synthetic sample as well.

## IMPUTING MISSING FOUNDER GENOTYPES INCREASES THE ACCURCY OF HAFS

While missing information can be accommodated by many haplotype inference tools (i.e. an N in place of a missing call), it is unclear how missing calls affect inference accuracy, and what the best practices should be when missing calls are present in the reference founder set.

We first examined whether haplotype frequencies estimated for founders with many missing calls or few missing calls systematically deviated from an expected haplotype frequency of 0.0101 (1/99). We found that across individual inference windows, there was a clear negative correlation between the number of missing calls per founder, and the haplotype frequencies estimated for that founder ([Supp. Fig. 1](#_jgy0outt7ap1)). To determine whether the observed skewed haplotype frequencies were directly associated with the presence of missing sites, we tested whether imputing genotype calls for missing sites would reduce bias in haplotype frequency assignment. While a number of sophisticated methods for imputing rare SNPs do exist [42–44](https://paperpile.com/c/5dF0nm/AVZlk%2ByO47Z%2Bf0SmJ), and may in some cases improve HAF accuracy, here we used a simple approach. To perform imputation, at each site we first calculated the allele frequency among called founder genotypes and used this value as a probability for assigning genotypes to missing calls. We found that imputation significantly reduced the skewed haplotype frequency distribution by 4-6-fold for all empirical coverages and window sizes tested. We expect that imputation with more advanced tools would achieve even better results.

We next examined how imputation of haplotype frequencies can impact the overall accuracy of HAFs. We also confirmed that haplotype inference using imputed calls produced more accurate HAFs than using a subset of sites with no missing calls. Thus, we include imputation as a key step in our analysis pipeline.