FILE S1

Corrections for pooled sampling

The nature of our experimental design creates additional noise that we must correct for in our population genetic estimates. In particular, the pooled DNA sequencing design of this manuscript creates a second level of binomial sampling, beyond what is associated with the "normal" population genetic survey. Throughout we assume that a sample of size n chromosomes is taken from nature and pooled for sequencing. This sequencing is performed to depth m which may be variable among loci/sites. Conditioning on a population frequency of the A_1 allele p, the probability of sampling i out of n A_1 alleles in our initial sample is simply binomial with parameters n and p. Thus the expected value of the sample frequency $E(i/n) = p$ and the second moment is $E((i/n)^2) = \frac{p(1-p)}{n} + p^2$.

We will first derive similar results for the pooled sampling design, and then move on to estimation of population genetic statistics. The probability of sampling k A_1 alleles out of m in our pooled sequences conditional upon having sampled i out of n in our initial sampling is again binomial, this time with parameters m and i/n. Thus the probability of sequencing k out of m reads of the A_1 allele conditional upon the population allele frequency is

$$Prob(X = k|m, n, p) = \sum_{i=0}^{n} \binom{m}{k} (i/n)^k (1 - i/n)^{m-k} \binom{n}{i} p^i (1-p)^{n-i}. \quad (1)$$

The expected value of the frequency of the allele in our sequenced sample, k/m, can then be easily found through the use of conditional expectations

$$\begin{aligned} E\{k/m\} &= E\{E\{k/m|i/n\}\} \\ &= \sum_{i=0}^{n} E\{k/m|i/n\} \times Prob(i) \\ &= \sum_{i=0}^{n} \frac{m(i/n)}{m} \binom{n}{i} p^i (1-p)^{n-i} = p \end{aligned} \quad (2)$$

We can find the second moment through similar means

$$\begin{aligned} E\{(k/m)^2\} &= E\{E\{(k/m)^2|i/n\}\} \\ &= \sum_{i=0}^{n} E\{(k/m)^2|i/n\} \times Prob(i) \\ &= \sum_{i=0}^{n} \left\{ \frac{(i/n)(1-i/n)}{m} + (i/n)^2 \right\} \times Prob(i) \\ &= \sum_{i=0}^{n} \frac{(i/n)(1-i/n)}{m} \times Prob(i) + \sum_{i=0}^{n} (i/n)^2 \times Prob(i) \end{aligned} \quad (3)$$

There are two terms in equation 3. This second term is immediately recognizable as the second moment that we examined above (i.e. $\frac{p(1-p)}{n} + p^2$). The first term after a bit of algebra turns into

$$\frac{1}{m}\{p(1-p) - \frac{p(1-p)}{n})\}$$

putting it all together, the expectation of the second moment, conditional upon the population allele frequency is

$$E\{(k/m)^2\} = p(1-p)/m - p(1-p)/mn + p(1-p)/n$$

With that result in hand we are now ready to write down the expectation of heterozygosity ($H = 2p(1-p)$) given our population allele frequency

$$\begin{aligned} E\{H\} &= E\{2p(1-p)\} = 2(E\{p\} - E\{p^2\}) \\ &= 2\Big(p - p(1-p)/m + p(1-p)/mn - p(1-p)/n\Big) \\ &= 2p(1-p)((n-1)/n)((m-1)/m) \end{aligned} \quad (4)$$

This leads to a simple bias correction on our estimates of heterozygosity which is $n/(n-1) \times m/(m-1)$. Figure S7 shows coalescent simulation results, where we generated samples from the standard coalescent model, and then resampled chromosomes with replacement to a coverage depth m. We then applied both the "double" correction derived here and the standard single correction. As can be seen in that figure, we do indeed have an unbiased estimator of heterozygosity if we correct for both the original size of our pooled sample and the coverage.

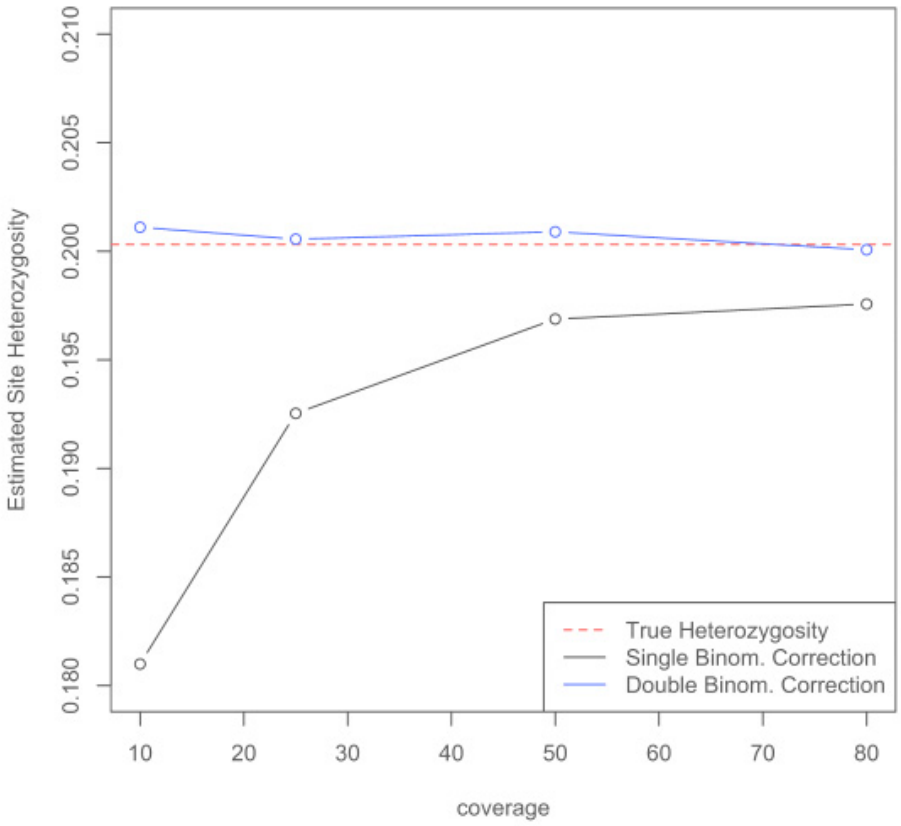

FIGURE S5.—Simulation results showing the corrected heterozygosity (eqn 4) is effective across a range of coverages used in this manuscript

Estimating θ

Of interest to us is coming up with an unbiased estimator of $\theta = 4Nu$ in the face of our pooled sampling strategy. Recently (Futschik and Schlotterer 2010) have done quite a bit of work on this problem, and they were able to come up with corrected estimators for $\theta_p i$ and θ_w. Here we derive ostensibly similar results through different means, and arrive at a generalized correction for pooled sampling which allows for construction of arbitrary estimators of θ as linear combinations of the site frequency spectrum using the system of (Achaz 2008).

We start by generalizing equation 1 of the supplement across the sample site frequency spectrum (SFS) expected under the standard neutral model. The probability of observing an allele segregating at frequency i out of n in a standard sample is $Prob(i|n) = 1/ia_n$, where $a_n = \sum_{i=1}^{n-1} 1/i$ (Ewens 2004). Thus the probability of observing an allele at frequency k out of m reads in our pooled sequence sample unconditional on the population

frequency is

$$\begin{aligned} Prob(k|m,n) &= \sum_{i=1}^{n-1} Prob(K|m,n,i)Prob(i) \\ &= \sum_{i=1}^{n-1} \binom{m}{k} (i/n)^k (1-i/n)^{m-k} (1/ia_n) \end{aligned} \tag{5}$$

This expression allows us to write down the expected number of sites segregating at frequency k out of m, call it Y_k as a function of the mutation rate θ. Conditioning on the expected number of segregating sites S in our initial sample of size n allows us to write down the expected values of the Y_ks as

$$\begin{aligned} E\{Y_k\} &= E\{S\} \times Prob(k|m,n) \\ &= \theta a_n \sum_{i=1}^{n-1} \binom{m}{k} (i/n)^k (1-i/n)^{m-k} (1/ia_n) \end{aligned} \tag{6}$$

To check the accuracy of this expression we performed coalescent simulations with a specified sample size n and mutation rate θ. The initial site frequency spectrum was recorded and then transformed to mimic our pooled sequencing by sampling alleles at each segregating site with replacement. This yields a transformed SFS Y – see figure S6.

Given the accuracy of our correction for the SFS we move on to derive a bias corrected estimation scheme for θ. In particular we note that rearrangement of equation 6 suggests a moment estimator of the type derived in Fu (1995),

$$\hat{\theta} = \frac{Y_i}{a_n} \frac{1}{Prob(k|m,n)}$$

Achaz (2009) noted that linear combinations of the SFS can be used to derive new estimators of θ given some arbitrary weighting scheme. In this context we can write down the bias corrected version of Achaz's generic estimator as

$$\hat{\theta}_\omega = \frac{1}{a_n \sum_k \omega_k} \sum_{k=1}^{m-1} \omega_k Y_k \frac{1}{Prob(k|m,n)}. \tag{7}$$

In this manuscript we focus attention on Tajima's nucleotide diversity ($\hat{\theta}_\pi$) and Fay and Wu's ($\hat{\theta}_H$) (Tajima 1983; Fay and Wu 2000). To show the potential generality of our correction scheme we present coalescent simulation results as before where we have estimated θ using six different weighting schemes: $\omega_{\pi,i}$, $\omega_{H,i}$, and $\omega_{W,i}$, each with and without use of singletons

$$\begin{aligned}\omega_{\pi,i} &= n - i \\ \omega_{H,i} &= i \\ \omega_{W,i} &= i^{-1}\end{aligned}$$

In the case where singletons are ignored each of the $\omega_1 = 0$. As can be seen in Figure S7 our bias corrected estimates are quite accurate, thus the framework we have introduced here should be general.

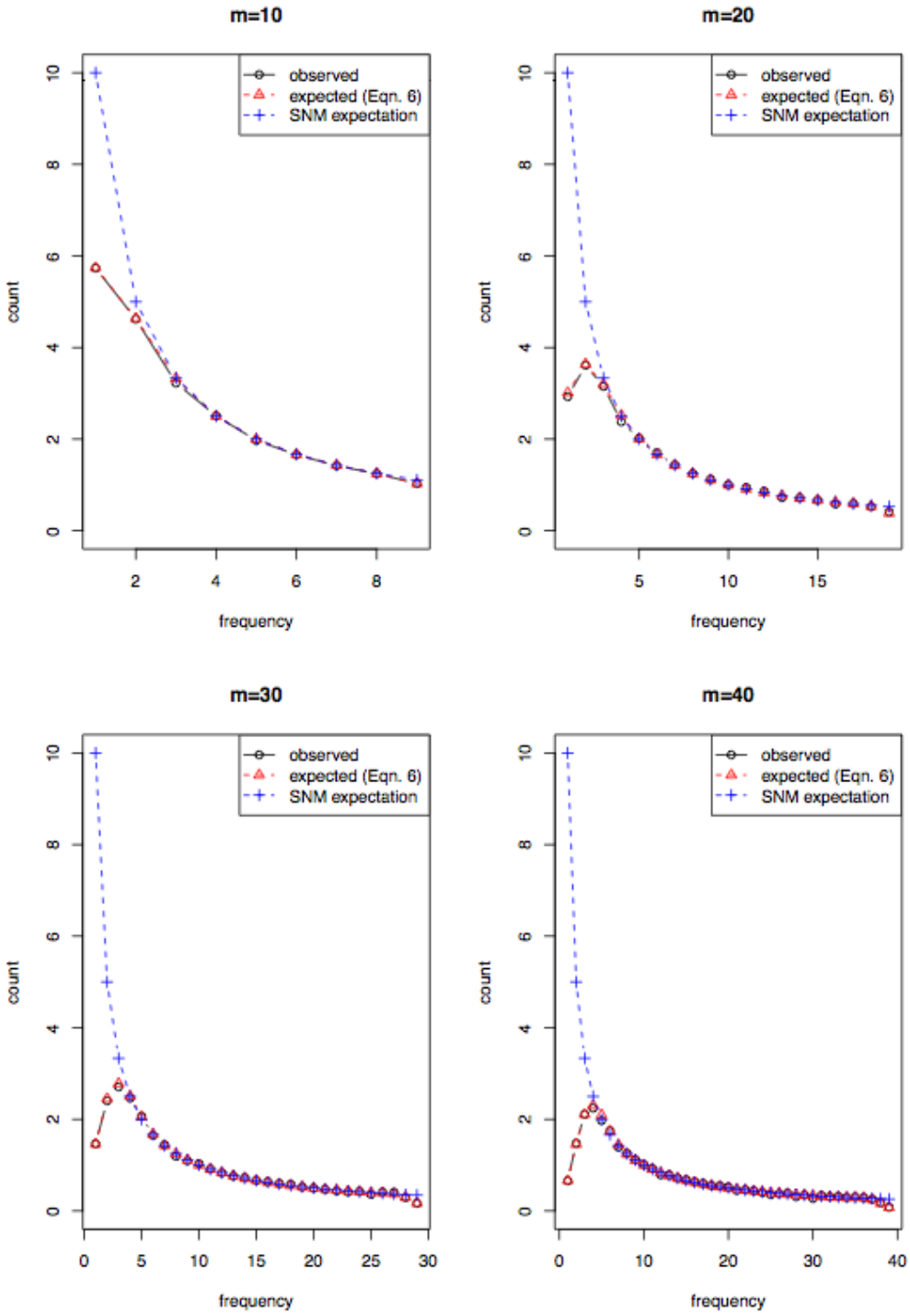

FIGURE S6.—Simulation results showing the correspondence between the observed and expected site frequency spectrum as m the sequencing depth changes. 1000 coalescent simulations were run with n=10 and θ=10. The expected values in red are derived from equation 6. Shown for comparison in blue are the expectations under the standard neutral model.

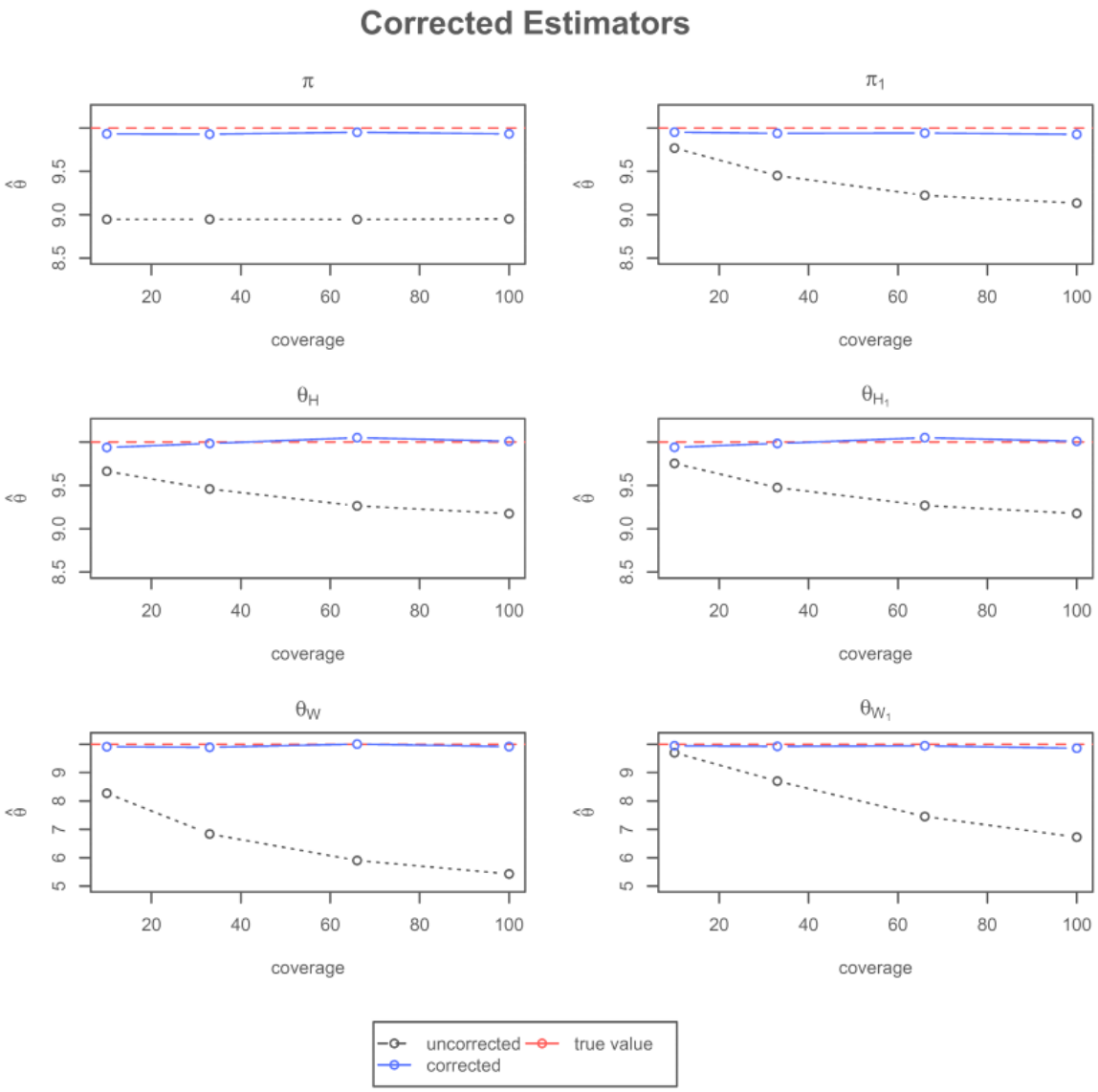

FIGURE S7.—Simulation results showing the performance of our bias corrected estimators of θ. 1000 coalescent simulations were run with $n = 40$ and $\theta = 10$. Uncorrected estimates are shown in black.

REFERENCES

ACHAZ, G., 2008, Jul) Testing for neutrality in samples with sequencing errors. Genetics *179* (3): 1409–1424.

ACHAZ, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics *183* (1): 249–58.

EWENS, W. J., 2004 *Mathematical population genetics* (2nd ed ed.), Volume v. 27. New York: Springer.

FAY, J. C. and C. I. WU, 2000, Jul) Hitchhiking under positive Darwinian selection. Genetics *155* (3): 1405–13.

FU, Y. X., 1995 Statistical Properties of Segregating Sites. Theoretical Population Biology **48**: 172–197.

FUTSCHIK, A. and C. SCHLOTTERER, 2010 Massively Parallel Sequencing of Pooled DNA Samples–The Next Generation of Molecular Markers. Genetics.

TAJIMA, F., 1983, Oct) Evolutionary relationship of DNA sequences in finite populations. Genetics *105* (2): 437–60.