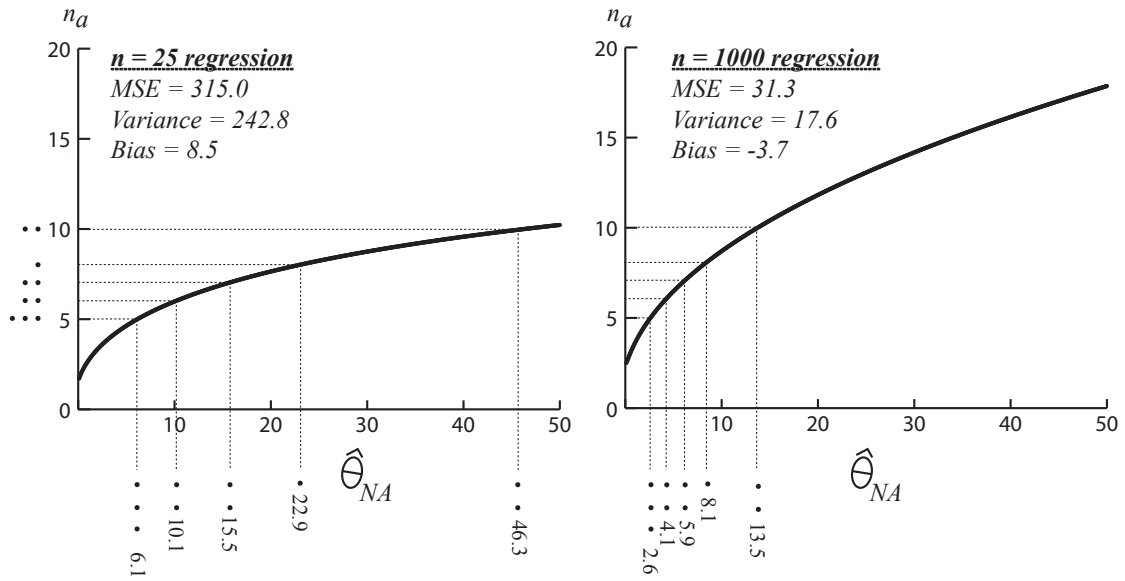
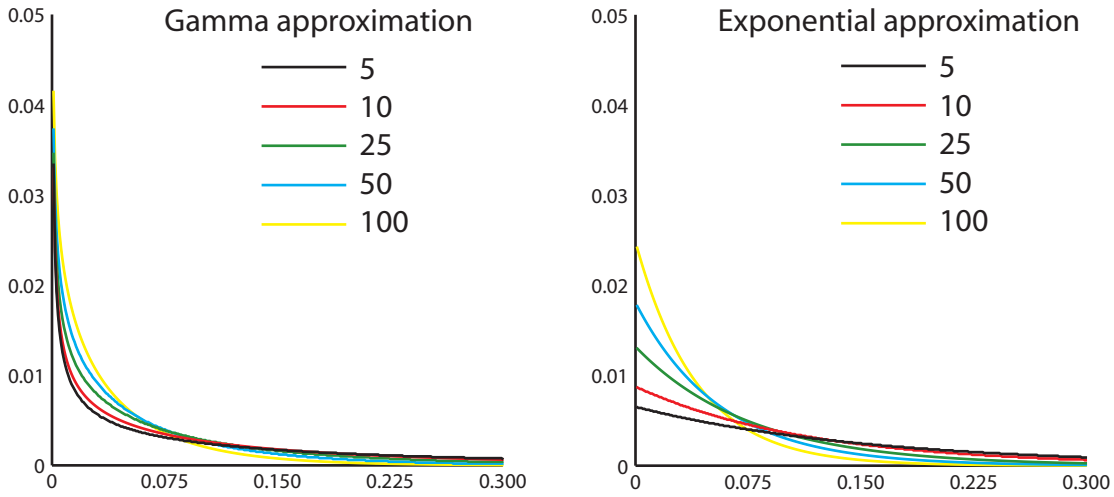
Supplementary Information for Haasl & Payseur.



SUPPLEMENTARY FIGURE 1. Using a version of equation (7) that is specific to the actual sample size can lead to large overestimates of θ. In this case, 10 data sets of $n = 25$ were simulated with true $\theta = 10$. This particular set of 10 data sets was chosen for its extremity. The n_a were: 5, 5, 5, 6, 6, 7, 7, 8, 10, 10. These values of n_a are mapped to estimates $\hat{\theta}_{NA}$ using the $n = 25$ regression (left panel) and $n = 1000$ regression (right panel). Despite the fact that the n_a vs θ curve in the left panel accurately corresponds to the average n_a for each θ when $n = 25$, its flat slope can lead to wild overestimates of θ when an unusually large value of n_a is obtained. Using the $n = 1000$ regression buffers against such overestimates, leading to an MSE that is 1/10th of that using the $n = 25$ regression.



SUPPLEMENTARY FIGURE 2. When estimating θ based on the vector of observed allele frequencies, the exponential approximation to the microsatellite frequency spectrum leads to more accurate estimates of θ than the gamma approximation. This is despite the fact that the gamma approximation describes empirical frequencies much better than the gamma approximation. Here, spectra for $\theta = 5, 10, 25, 50,$ and 100 are plotted on the interval (0, 0.3], with the gamma approximation on the left and the exponential approximation on the right. As θ changes, changes to the empirical approximation are most extreme along the very allele frequency intervals that change the most with θ. For example, exponential approximations for $\theta = 5$ and $\theta = 25$ are quite divergent for allele frequencies on the interval (0, 0.05], while the corresponding gamma approximations show much less divergence. To a lesser extent, divergence between the expoential approximations on the allele frequency interval (0.075, 0.3) are greater than those between the corresponding gamma approximations. As θ increases, allele frequencies < 0.05 (especially < 0.01) become much more common and allele frequencies > 0.1 become much more rare. The form of the exponential approximation is better suited to diagnose these differences.

SUPPLEMENTARY TABLE 1. **Intercepts and regression coefficients for equations (7) and (8)**.

	equation (7)			equation (8)
n	c_0	c_1	c_2	a_0
25	1.5384	-0.05558	-0.01983	0.57991
50	1.6831	-0.03792	-0.01655	0.53729
100	1.8157	-0.04686	-0.01017	0.49990
150	1.8816	-0.05574	-0.00673	0.46722
200	1.9210	-0.05735	-0.00557	0.45206
250	1.9486	-0.06295	-0.00387	0.43252
500	2.0357	-0.07910	-0.00007	0.39704
1000	2.0845	-0.08290	0.00110	0.35311

SUPPLEMENTARY TABLE 2. **Sample size and θ estimation**. MSE and bias (in parentheses) are shown. All statistics based on 150 independent data sets. n and θ in the two leftmost columns refer to the simulation parameters. The first row of the header refers to the version of equation (7) used to perform estimation. The results imply a complicated relationship between θ, n, and our methods of estimation. Generally, MSE and bias are reduced by using a version of equation (7) specific to a value of n that is greater than actual sample size.

		$\underline{n=25}$		$\underline{n=50}$		$\underline{n=100}$		$\underline{n=150}$		$\underline{n=250}$		$\underline{n=1000}$	
n	θ	$\hat{\theta}_{EFS}$	$\hat{\theta}_{NA}$	$\hat{\theta}_{EFS}$	$\hat{\theta}_{NA}$	$\hat{\theta}_{EFS}$	$\hat{\theta}_{NA}$	$\hat{\theta}_{EFS}$	$\hat{\theta}_{NA}$	$\hat{\theta}_{EFS}$	$\hat{\theta}_{NA}$	$\hat{\theta}_{EFS}$	$\hat{\theta}_{NA}$
25	10	165 (5.6)	148 (5.2)	118 (3.6)	96 (3.3)	25.8 (-2.2)	25.2 (-2.4)	25.2 (-3.2)	25.3 (-3.3)	26.2 (-3.6)	26.2 (-3.7)	31.9 (-4.7)	32.0 (-4.8)
	100	19920 (89)	15843 (73)	18664 (82)	14799 (66)	3753 (-59)	4004 (-61)	4280 (-6)	4530 (-66)	4710 (-67)	4946 (-70)	5536 (-74)	5338 (-75)
50	10	-	-	75.4 (2.5)	75.2 (2.7)	36.2 (0)	35.7 (0.12)	28.2 (-1.1)	27.8 (-1)	24.9 (-1.9)	24.2 (-1.8)	25.4 (-3.1)	24.6 (-3)
	100	-	-	14196 (55)	8334 (35)	2342 (-12)	1822 (-21)	2003 (-27)	1967 (-33)	2266 (-36)	2352 (-41)	2977 (-49)	3224 (-53)
100	10	-	-	-	-	38.5 (0.7)	40.3 (1.0)	37.4 (0.5)	38.9 (0.8)	32.7 (-0.1)	33.6 (0.21)	24.3 (-2)	24.3 (-1.8)
	100	-	-	-	-	3102 (5.8)	2343 (-0.8)	3252 (8.8)	2470 (2.1)	2282 (-6.2)	1831 (-12)	2032 (-25.6)	1928 (-30)

SUPPLEMENTARY TABLE 3. **Effect of subsampling on θ estimation**. All subsampling results are based on 10,000 independent estimates of θ. For ease of comparison, the results from non-subsampling estimates are repeated from Table 1. Boldface statistics indicate combinations of θ and n for which subsampling improved estimation in terms of MSE.

θ	n	$\hat{\theta}_{\bar{x}}$	$\hat{\theta}_{NA}$	$\hat{\theta}_{EFS}$
5	150 subsampling	**5.75 (-1.00)**	**6.77 (-0.85)**	**6.88 (-1.02)**
	150	6.66 (-0.36)	8.22 (-0.12)	7.89 (-0.35)
	250 subsampling	**5.64 (-0.93)**	**6.68 (-0.78)**	**6.79 (-0.96)**
	250	7.30 (-0.11)	9.29 (0.17)	8.65 (-0.15)
	500 subsampling	**5.52 (-0.78)**	**6.68 (-0.61)**	**6.72 (-0.80)**
	500	7.60 (0.06)	11.78 (0.71)	9.95 (0.27)
	1000 subsampling	**5.41 (-0.85)**	**6.45 (-0.68)**	**6.53 (-0.88)**
	1000	8.34 (0.30)	13.33 (0.99)	10.79 (0.48)
10	150 subsampling	**20.03 (-2.67)**	**22.14 (-2.02)**	**22.80 (-2.23)**
	150	20.84 (-1.46)	27.45 (-0.61)	26.22 (-0.98)
	250 subsampling	**19.67 (-1.94)**	**22.42 (-1.72)**	**23.01 (-1.94)**
	250	22.58 (-0.82)	23.43 (0.15)	28.68 (-0.27)
	500 subsampling	**19.10 (-2.26)**	**22.16 (-1.53)**	**22.66 (-1.78)**
	500	22.92 (-0.66)	37.99 (1.04)	33.26 (0.38)
	1000 subsampling	**18.75 (-2.14)**	**22.00 (-1.39)**	**22.47 (-1.64)**
	1000	24.72 (-0.14)	43.79 (1.70)	36.86 (0.90)
25	150 subsampling	130.81 (-6.48)	**119.96 (-6.48)**	**125.61 (-6.58)**
	150	114.45 (-6.08)	138.76 (-2.57)	138.17 (-3.07)
	250 subsampling	126.28 (-8.65)	**117.47 (-7.01)**	**123.37 (-6.15)**
	250	113.93 (-4.53)	161.29 (0.55)	152.06 (-1.34)
	500 subsampling	119.64 (-8.12)	**113.95 (-5.35)**	**120.03 (-5.49)**
	500	118.14 (-3.67)	200.40 (1.92)	188.00 (0.88)
	1000 subsampling	**118.21 (-7.92)**	**114.35 (-5.10)**	**120.74 (-5.24)**
	1000	122.70 (-2.57)	231.52 (3.42)	210.54 (2.12)
75	150 subsampling	1587.9 (-37.05)	1166.4 (-27.81)	1159.8 (-26.12)
	150	1144 (-26.32)	963.6 (-13.19)	1044 (-13.37)
	250 subsampling	1474.6 (-35.15)	1076.5 (-25.24)	**1084.9 (-23.39)**
	250	1010 (-20.38)	1061 (-5.35)	1178 (-6.08)
	500 subsampling	1386.6 (-33.42)	**1024.2 (-22.88)**	**1054.4 (-20.79)**
	500	1001 (-17.60)	1560 (3.69)	1730 (4.08)
	1000 subsampling	1350.4 (-32.77)	**999.1 (-22.00)**	**1037.5 (-19.85)**
	1000	987 (-13.56)	1917 (9.76)	2044 (9.32)