Supplement to "Genotyping Polyploids from Messy Sequencing Data"

David Gerard[*,1], **Luis Felipe Ventorim Ferrão**[†], **Antonio Augusto Franco Garcia**[‡] **and Matthew Stephens**[§,**]

[*]Department of Mathematics and Statistics, American University, Washington, DC, 20016, USA, [†]Horticultural Sciences Department, University of Florida, Gainesville, FL, 32611, USA, [‡]Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, SP, 03178-200, Brazil, [§]Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA, [**]Department of Statistics, University of Chicago, Chicago, IL, 60637, USA

ABSTRACT This document contains optimization details, theorems and proofs relating to preferential pairing, and supplementary figures.

KEYWORDS GBS; RAD-Seq; sequencing; hierarchical modeling; read-mapping bias

EM Algorithm for an F1 Cross

1. Let (see (4))

$$a_{k_i\ell_1\ell_2} := Pr(p_i = k_i/K | \tilde{p}_1 = \ell_1/K, \tilde{p}_2 = \ell_2/K). \tag{S1}$$

2. E-Step: Set

$$\theta_i := \frac{\pi \sum_{k_i=0}^{K} \text{BB}(y_i|n_i, \xi(k_i/K, \epsilon, h), \tau) a_{k_i\ell_1\ell_2}}{\pi \sum_{k_i=0}^{K} \text{BB}(y_i|n_i, \xi(k_i/K, \epsilon, h), \tau) a_{k_i\ell_1\ell_2} + (1-\pi)\,\text{BB}(y_i|n_i, 1/2, 1/3)} \tag{S2}$$

$$\tilde{\theta}_j := \frac{\pi\,\text{BB}(\tilde{y}_j|\tilde{n}_j, \xi(\ell_j/K, \epsilon, h), \tau)}{\pi\,\text{BB}(\tilde{y}_j|\tilde{n}_j, \xi(\ell_j/K, \epsilon, h), \tau) + (1-\pi)\,\text{BB}(\tilde{y}_j|\tilde{n}_j, 1/2, 1/3)} \tag{S3}$$

3. M-Step: Set

$$\pi = \frac{1}{N+2}\left(\tilde{\theta}_1 + \tilde{\theta}_2 + \sum_{i=1}^{N} \theta_i\right) \tag{S4}$$

$$\begin{aligned}(\ell_1, \ell_2, \tau, h, \epsilon) = \underset{(\ell_1,\ell_2,\tau,h,\epsilon) \in 1{:}K \times 1{:}K \times [0,1] \times \mathbb{R}^+ \times [0,1]}{\arg\max} \Bigg[& \sum_{i=1}^{N} \theta_i \log \sum_{k_i=0}^{K} \text{BB}(y_i|n_i, \xi(k_i/K, \epsilon, h), \tau) a_{k_i\ell_1\ell_2} \\ & + \tilde{\theta}_1 \log \text{BB}(\tilde{y}_1|\tilde{n}_1, \xi(\ell_1/K, \epsilon, h), \tau) + \tilde{\theta}_2 \log \text{BB}(\tilde{y}_2|\tilde{n}_2, \xi(\ell_2/K, \epsilon, h), \tau) \\ & - \frac{1}{2\sigma_h^2}(\log(h) - \mu_h)^2 - \log(h) - \frac{1}{2\sigma_\epsilon^2}(\text{logit}(\epsilon) - \mu_\epsilon)^2 - \log(\epsilon) - \log(1-\epsilon)\Bigg],\end{aligned} \tag{S5}$$

where

$$\xi(p_i, \epsilon, h) = \frac{f(p_i, \epsilon)}{f(p_i, \epsilon) + h(1 - f(p_i, \epsilon))}, \tag{S6}$$

$$f(p_i, \epsilon) = (1-\epsilon)p_i + \epsilon(1-p_i). \tag{S7}$$

Theorems Relating to Preferential Pairing

See Section Accounting for Preferential Pairing for notational definitions.

Theorem S1. *The number of configurations* $\boldsymbol{m}$ *given a K-ploid parent has* ℓ *copies of* $\boldsymbol{A}$ *are*

$$\begin{cases} \frac{K}{2} - \left\lceil \frac{\ell}{2} \right\rceil + 1 & \text{if } 2\ell \geq K, \\ \left\lfloor \frac{\ell}{2} \right\rfloor + 1 & \text{if } 2\ell < K, \end{cases} \tag{S8}$$

where $\lceil \cdot \rceil$ *and* $\lfloor \cdot \rfloor$ *are the ceiling and floor functions, respectively.*

[1]Corresponding author: Department of Mathematics and Statistics, American University, 3501 Nebraska Ave NW, Don Myers Building, Washington, DC 20016-8050, U.S.A. E-mail: dgerard@american.edu

Proof. We begin by placing as many A's into an AA pair as possible. This is $\lfloor \ell/2 \rfloor$ AA pairs. In this configuration, there are $K/2 - \lceil \ell/2 \rceil$ aa pairs. We then navigate to a new configuration by subtracting 1 from m_1 and m_3 and adding 2 to m_2. We do this until either $m_1 = 0$ or $m_3 = 0$. Proceeding in this way, we reach all possible configurations. Counting the number of configurations we obtained, we get (S8). □

Summing over ℓ, we can also get a result for the total number of possible values of $\boldsymbol{m}$ not conditional on ℓ.

Theorem S2. *The total number of possible values of* $\boldsymbol{m}$ *is*

$$\frac{(K+4)(K+2)}{8} = \binom{\frac{K+4}{2}}{2}. \tag{S9}$$

Proof. The proof involves summing (S8) over ℓ under the two cases where $K/2$ is even and $K/2$ is odd.

We begin with the case where $K/2$ is even. Then the total number of elements is

$$2\sum_{\ell=0}^{\frac{K-2}{2}} (\lfloor \ell/2 \rfloor + 1) + \lfloor K/4 \rfloor + 1 \text{ (by symmetry)} \tag{S10}$$

$$= 4\sum_{i=1}^{K/4} i + K/4 + 1 \text{ (counting each element in summand twice)} \tag{S11}$$

$$= K(K+4)/8 + K/4 + 1 \text{ (equation for sum of } i) \tag{S12}$$

$$= \frac{(K+4)(K+2)}{8}. \tag{S13}$$

For the case where $K/2$ is odd, we have the total number of elements is

$$2\sum_{\ell=0}^{\frac{K-2}{2}} (\lfloor \ell/2 \rfloor + 1) + \lfloor K/4 \rfloor + 1 \text{ (by symmetry)} \tag{S14}$$

$$= 2\sum_{\ell=0}^{\frac{K}{2}} (\lfloor \ell/2 \rfloor + 1) - \lfloor K/4 \rfloor - 1 \tag{S15}$$

$$= 4\sum_{i=1}^{\frac{K+2}{4}} i - \frac{K-2}{4} - 1 \text{ (counting each element in summand twice)} \tag{S16}$$

$$= \frac{(K+2)(K+6)}{8} - \frac{K-2}{4} - 1 \text{ (equation for sum of } i) \tag{S17}$$

$$= \frac{(K+4)(K+2)}{8}. \tag{S18}$$

□

Theorem S3. *The total possible number of pairings where chromosomes are labeled but pairings are not labeled is*

$$\binom{K}{2}\binom{K-2}{2}\ldots\binom{2}{2}/\left(\frac{K}{2}\right)! \tag{S19}$$

Proof. The proof is very similar the proof for the derivation of combinations. First, choosing 2 chromosomes from K is $\binom{K}{2}$, then choosing 2 chromosomes from $K-2$ is $\binom{K-2}{2}$. We continue in this manner until we exhaust all of the chromosomes. We then need to divide by the number of orderings for iteratively choosing 2 chromosomes, which is just $(\frac{K}{2})!$. □

Theorem S4. *Let*

$$f(\ell, m) = \begin{cases} 1 & \text{if } m = 0, \\ \binom{\ell}{2}\binom{\ell-2}{2}\ldots\binom{\ell-2m+2}{2}/m! & \text{otherwise.} \end{cases} \tag{S20}$$

Then the number of pairings that result in a configuration $\boldsymbol{m}$ *where the chromosomes are labeled but the pairings are not labeled is*

$$f(K-\ell, m_1) f(\ell, m_3) m_2! \tag{S21}$$

Proof. First, we select the number of pairings resulting in aa and we obtain $f(K-\ell, m_1)$. The proof of this is very similar to the proof in Theorem S3 so we omit the details. Similarly, the number of pairings that result in AA is $f(\ell, m_3)$. The number of ways we can pair the leftovers into Aa is $m_2!$ because we first choose one of m_2 A's to pair with the first a, then $m_2 - 1$ of the remaining A's to pair with the second a, and continue in this way until all A's and a's are paired. We then simply multiply these counts together to obtain (S21). □

Theorem S5. *If there is no preferential pairing, then the probability of configuration* $\boldsymbol{m}$ *is*

$$\frac{f(K-\ell, m_1) f(\ell, m_3) m_2! (\frac{K}{2})!}{\binom{K}{2}\binom{K-2}{2}\cdots\binom{2}{2}}, \tag{S22}$$

where $f(\ell, m)$ *is defined in* (S20).

Proof. There are $\binom{K}{2}\binom{K-2}{2}\cdots\binom{2}{2}/(\frac{K}{2})!$ possible configurations where the chromosomes are labeled but the pairings are not labeled (Theorem S3). Of those, $f(K-\ell, m_1) f(\ell, m_3) m_2!$ contain configuration $\boldsymbol{m}$ (Theorem S4). If each configuration, where the chromosomes are labeled but the pairings are not, have an equal probability, then the probability of configuration $\boldsymbol{m}$ is just the ratio of these two quantities, (S22). □

Supplementary Figures

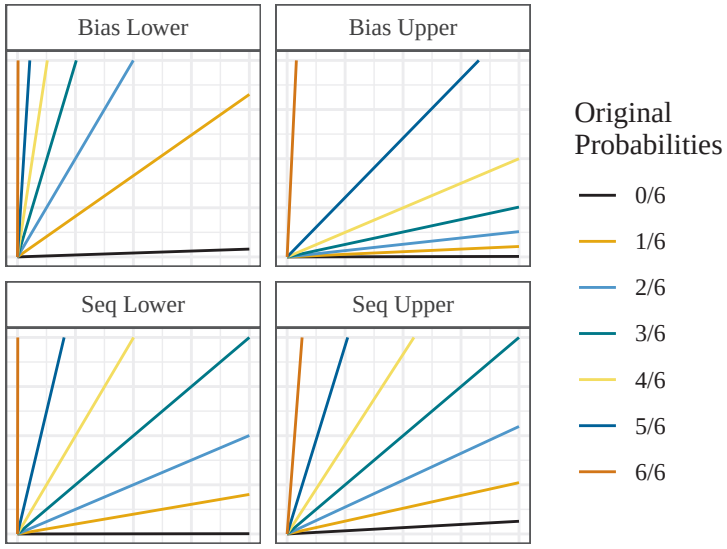

Figure S1 Considering an autohexaploid loci, -2 standard deviations of the bias parameter under our default prior (with a sequencing error rate of 0.01) (top left), +2 standard deviations of the bias parameter under our default prior (with a sequencing error rate of 0.01) (top right), -2 standard deviations of the sequencing error rate under our default prior (with a bias parameter of 1) (bottom left), +2 standard deviations of the sequencing error rate under our default prior (with a bias parameter of 1) (bottom right). The x-axis is the number of alternative reads and the y-axis is the number of reference reads.

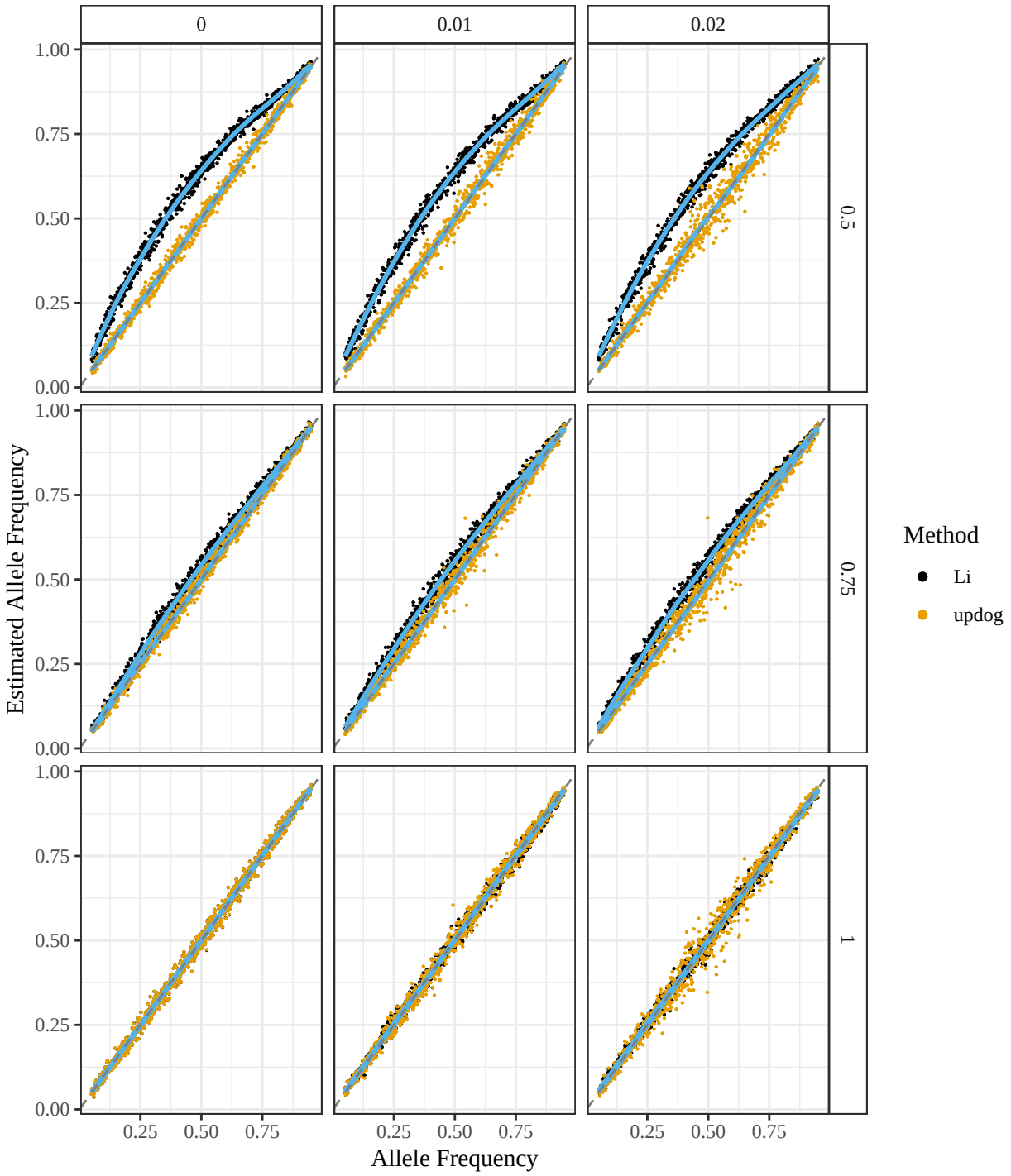

Figure S2 Estimated allele frequency (y-axis) for the `updog` method (black) and the method of Li (2011) (orange) versus true allele frequency (x-axis). An unbiased method would result in most points lying along the $y = x$ line. A smooth generalized additive model was fit to the results of both methods (blue lines). The column facets distinguish between different levels of the overdispersion parameter and the row facets distinguish between different levels of the bias parameter (with 1 indicating no bias).

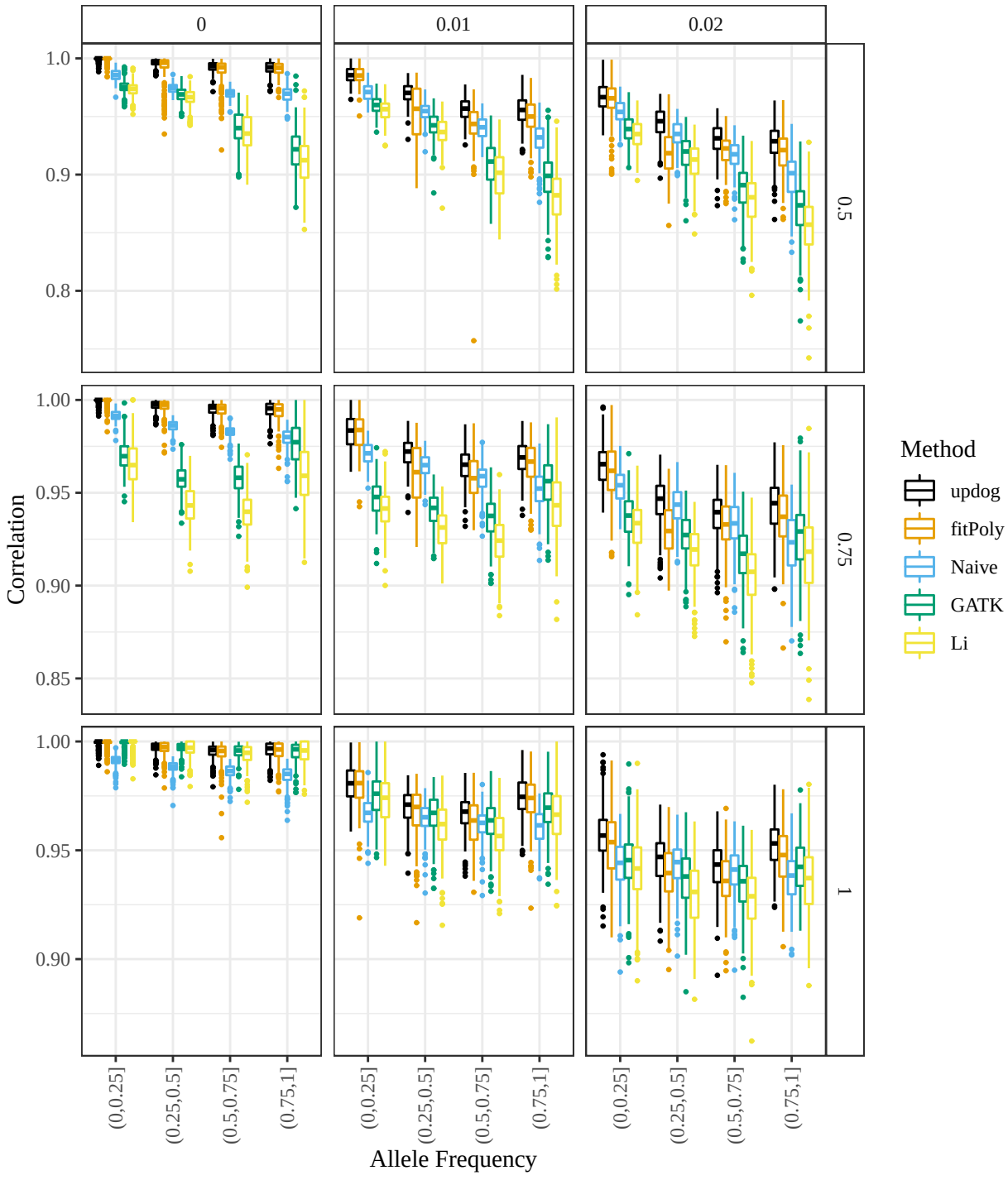

Figure S2 Boxplots of the correlation between the true and estimated genotypes (y-axis) stratified by the binned values of the major allele frequency (x-axis). The boxplots are color-coded by method: `updog` (black), `fitPoly` (orange), the naive method (blue), GATK (green), and the method of Li (2011) (yellow). The column facets distinguish between different levels of the overdispersion parameter and the row facets distinguish between different levels of the bias parameter (with 1 indicating no bias).

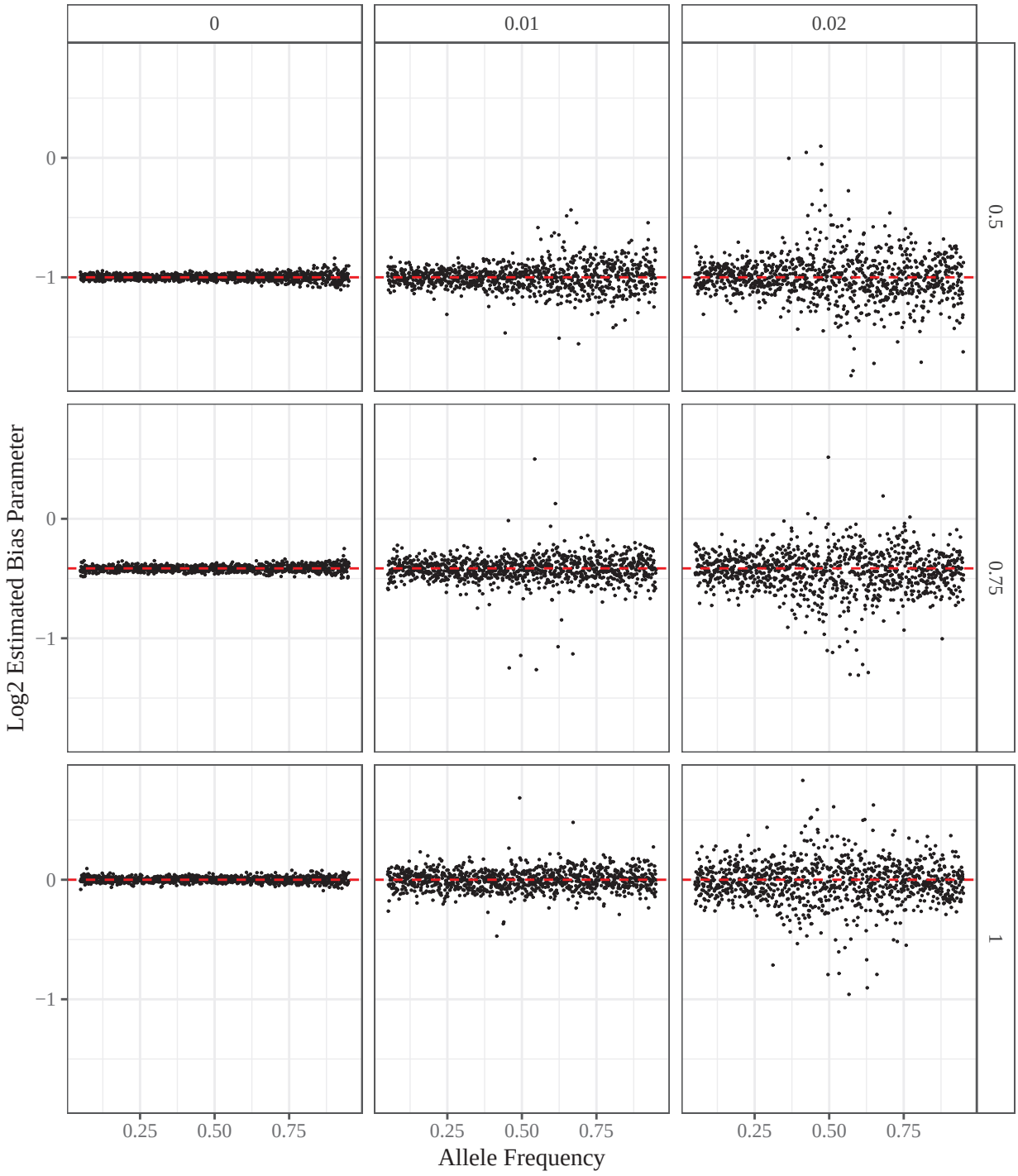

Figure S3 Allele frequency (x-axis) versus $\log_2$ of the estimated bias (y-axis). The row facets distinguish different levels of bias (with 1 indicating no bias) and the column facets distinguish different levels of overdispersion.

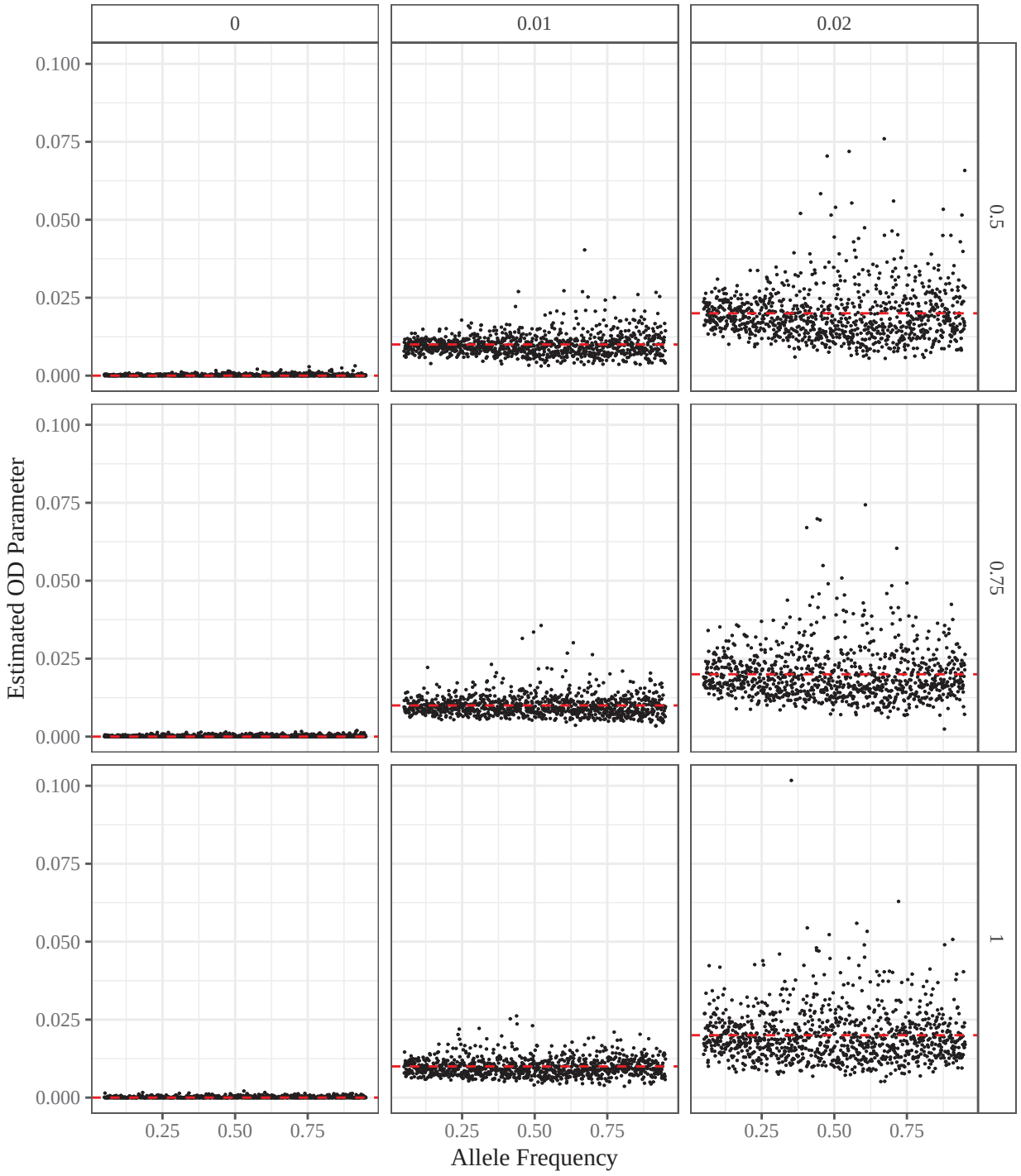

Figure S4 Allele frequency (*x*-axis) versus estimated overdispersion level (*y*-axis). The row facets distinguish different levels of bias (with 1 indicating no bias) and the column facets distinguish different levels of overdispersion.

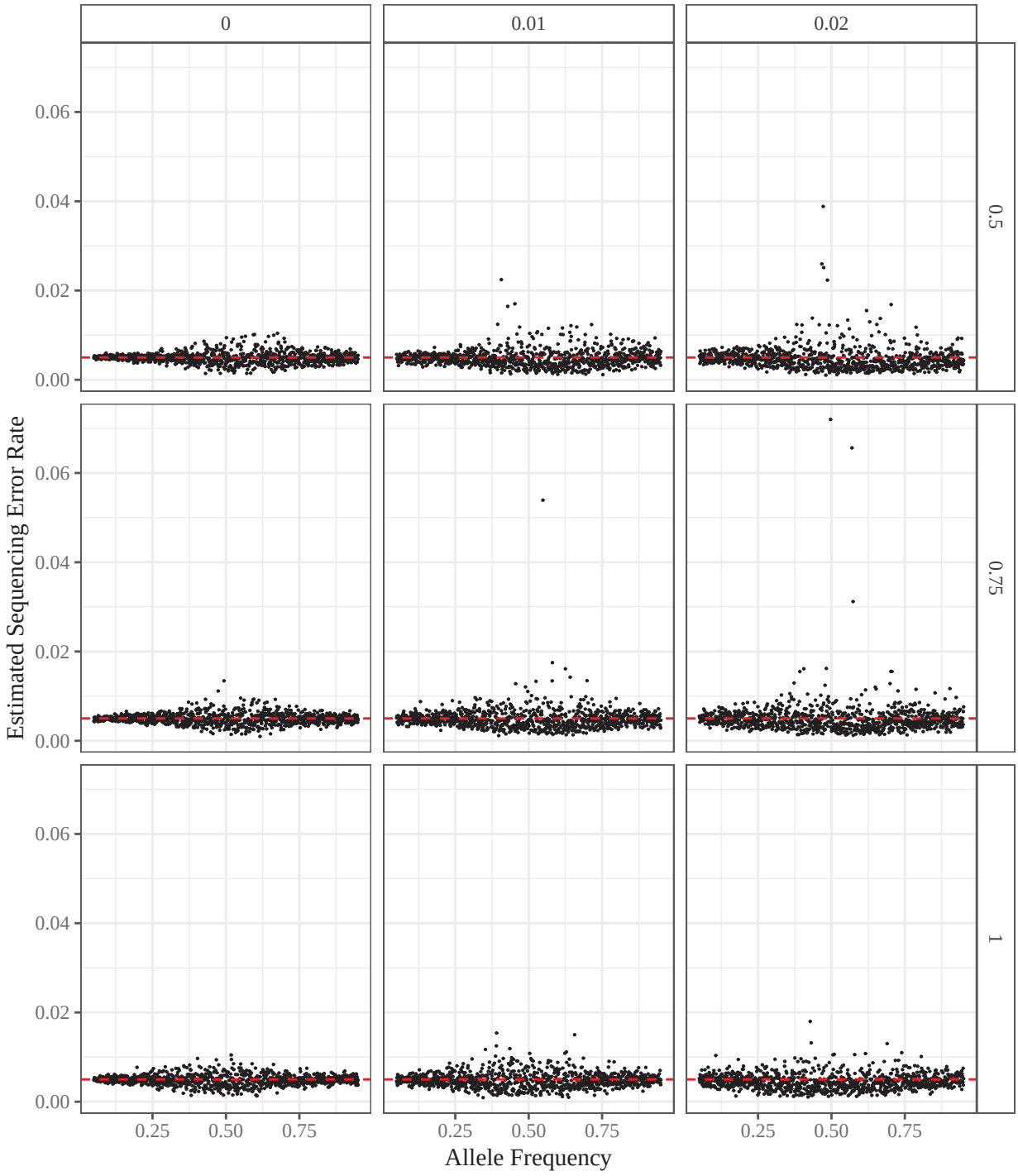

Figure S5 Allele frequency (*x*-axis) versus estimated sequencing error rate (*y*-axis). The row facets distinguish different levels of bias (with 1 indicating no bias) and the column facets distinguish different levels of overdispersion.

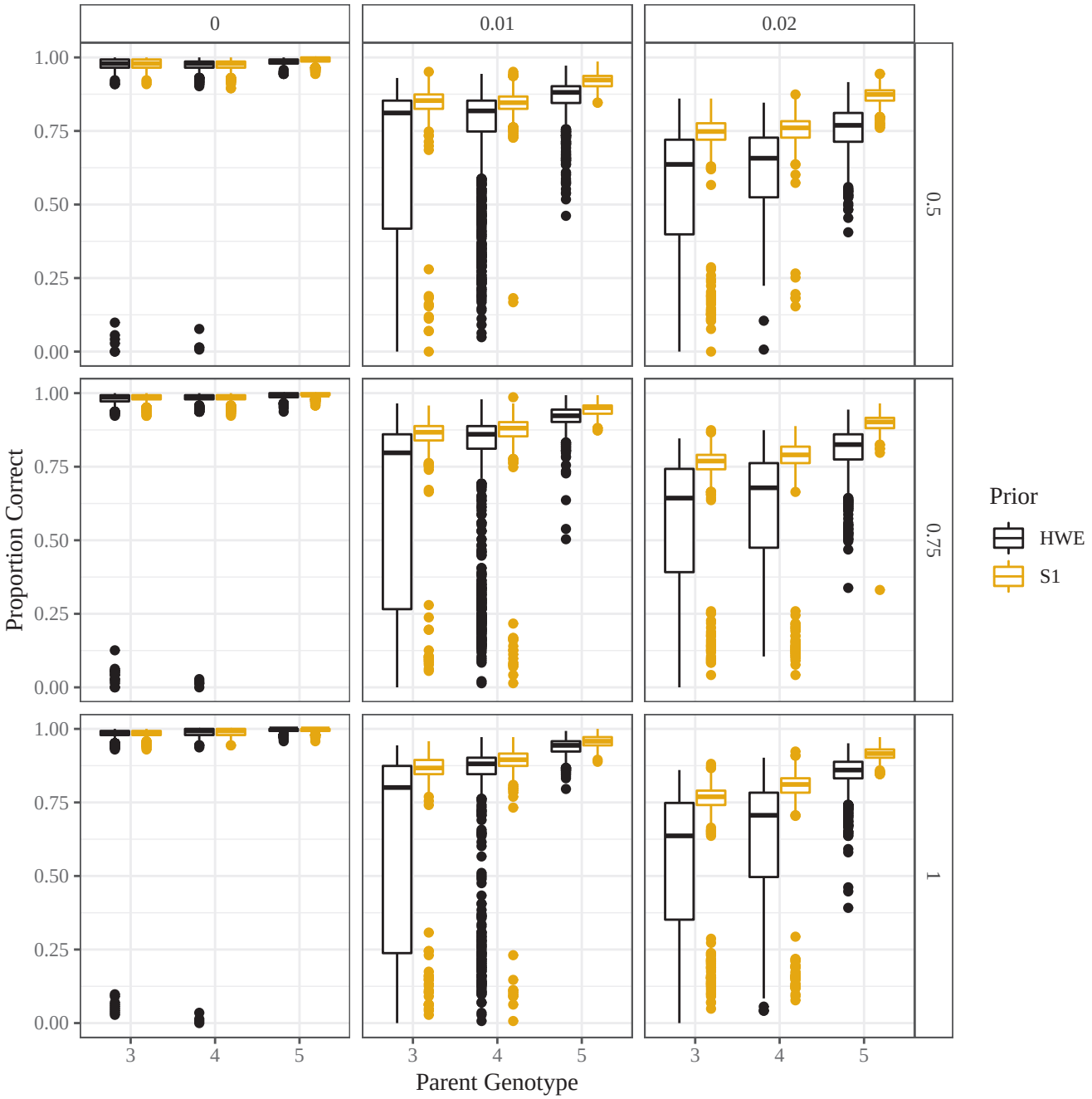

Figure S6 Box plots of the proportion of individuals genotyped correctly in an `updog` fit (y-axis) stratified by parent genotype (x-axis). Column facets distinguish different levels of overdispersion while row facets distinguish different levels of bias (with 1 indicating no bias). Orange box plots correctly assume the individuals are from an S1 population while black box plots incorrectly assume the population is in Hardy-Weinberg equilibrium.

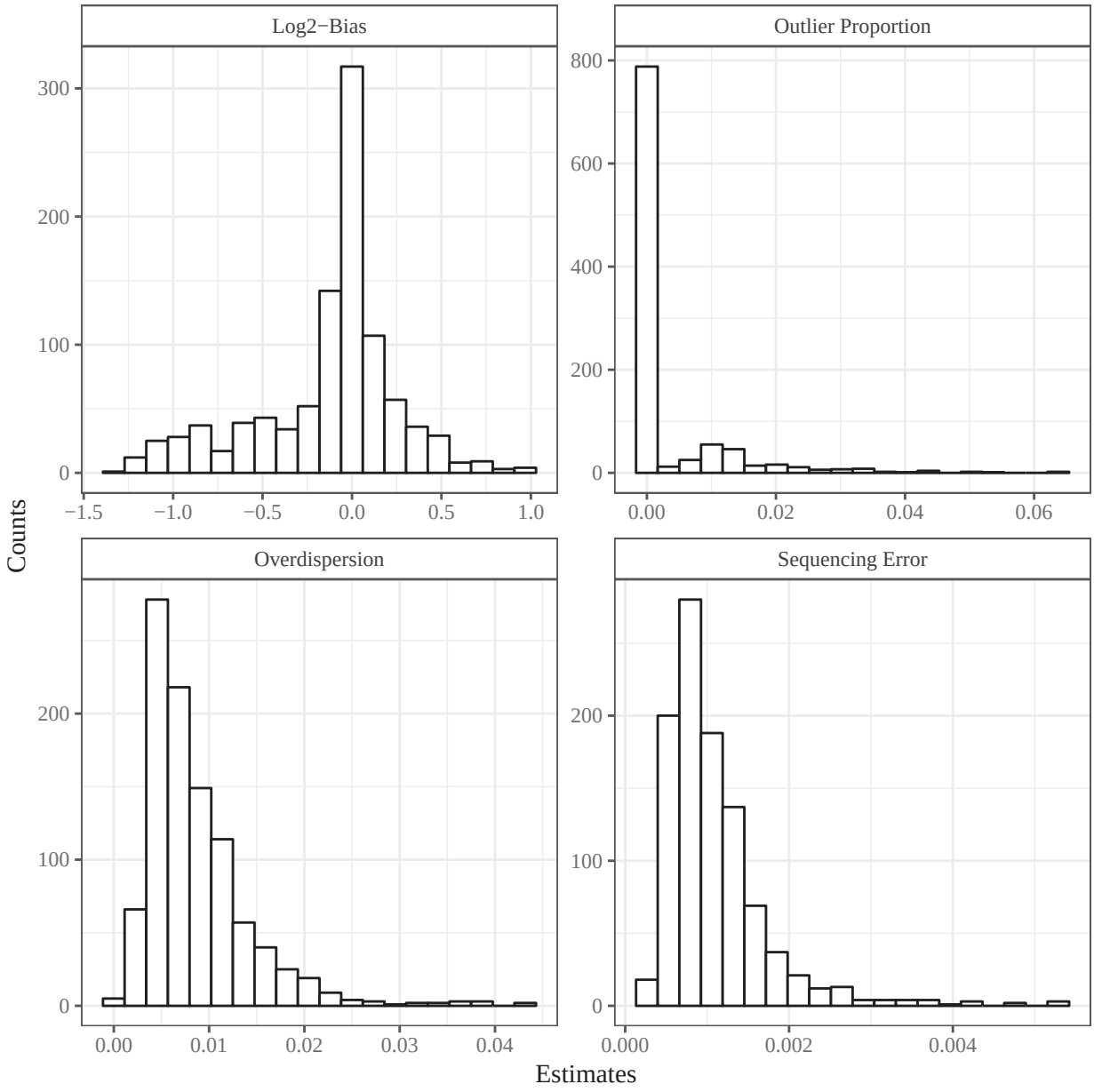

Figure S7 Histograms of `updog` estimates of parameters in 1000 SNPs from the data from Shirasawa *et al.* (2017).

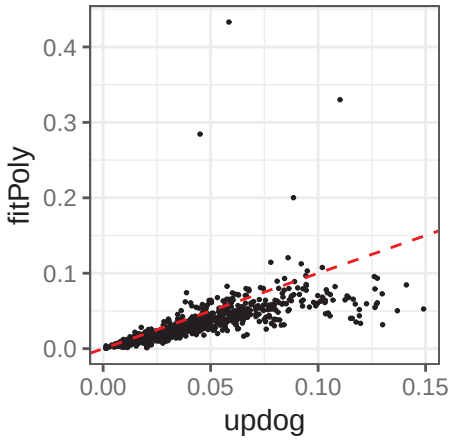

Figure S8 `Updog`'s estimated proportion of individuals genotyped incorrectly (x-axis) versus `fitPoly`'s estimated proportion of individuals genotyped incorrectly (y-axis) for the 1000 SNPs in the dataset from Shirasawa *et al.* (2017) with the largest read depth. The dashed line is the $y = x$ line. `Updog` has a larger estimate of incorrectly genotyped individuals for points below this line, and a smaller estimate for points above this line.

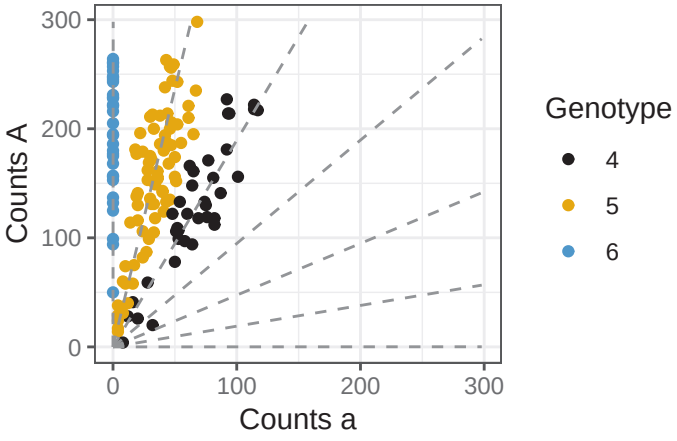

Figure S9 A genotype plot of the SNP from the dataset of Shirasawa *et al.* (2017) where `updog` estimated that 4.5% of the individuals were genotyped incorrectly and `fitPoly` estimated that 28.5% of the individuals were genotyped incorrectly. The figure is annotated by an `updog` fit.

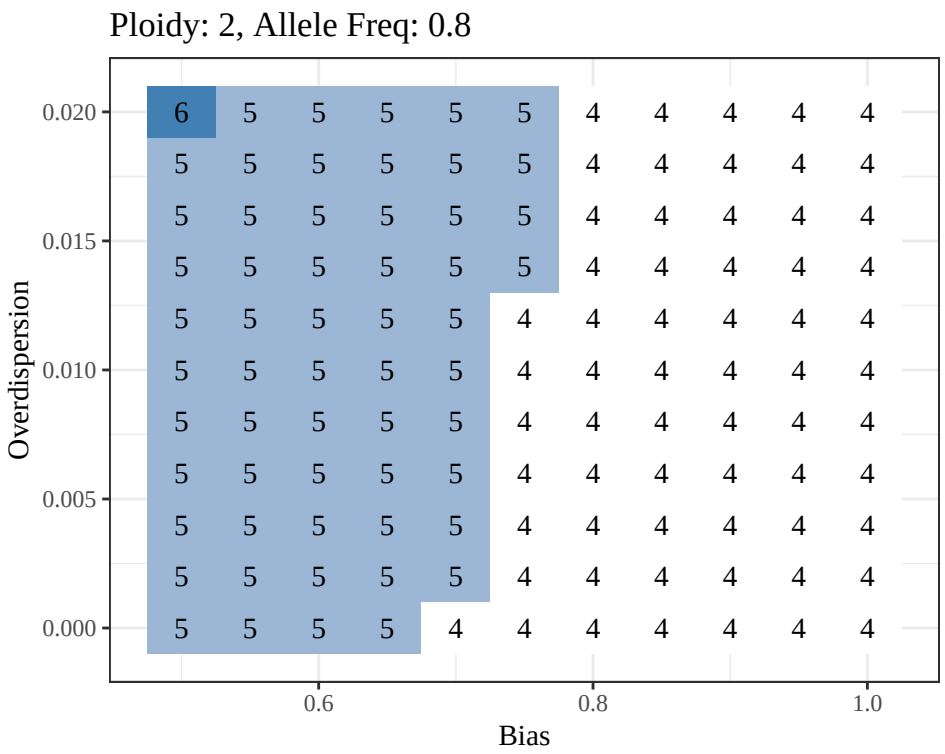

Figure S10 The minimum read depth required to obtain an oracle genotyping error rate less than 0.05 under different levels of overdispersion (*y*-axis) and allelic bias (*x*-axis) while the sequencing error rate is fixed at 0.001. This is for a diploid population under Hardy-Weinberg equilibrium with a major allele frequency of 0.8.

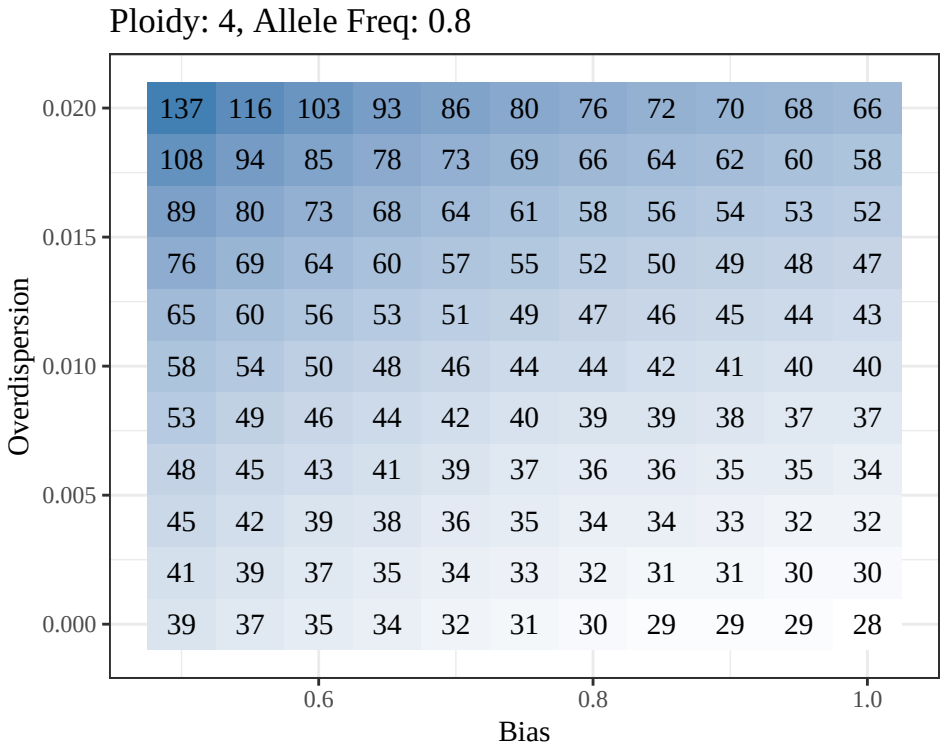

Figure S11 The minimum read depth required to obtain an oracle genotyping error rate less than 0.05 under different levels of overdispersion (*y*-axis) and allelic bias (*x*-axis) while the sequencing error rate is fixed at 0.001. This is for a tetraploid population under Hardy-Weinberg equilibrium with a major allele frequency of 0.8.

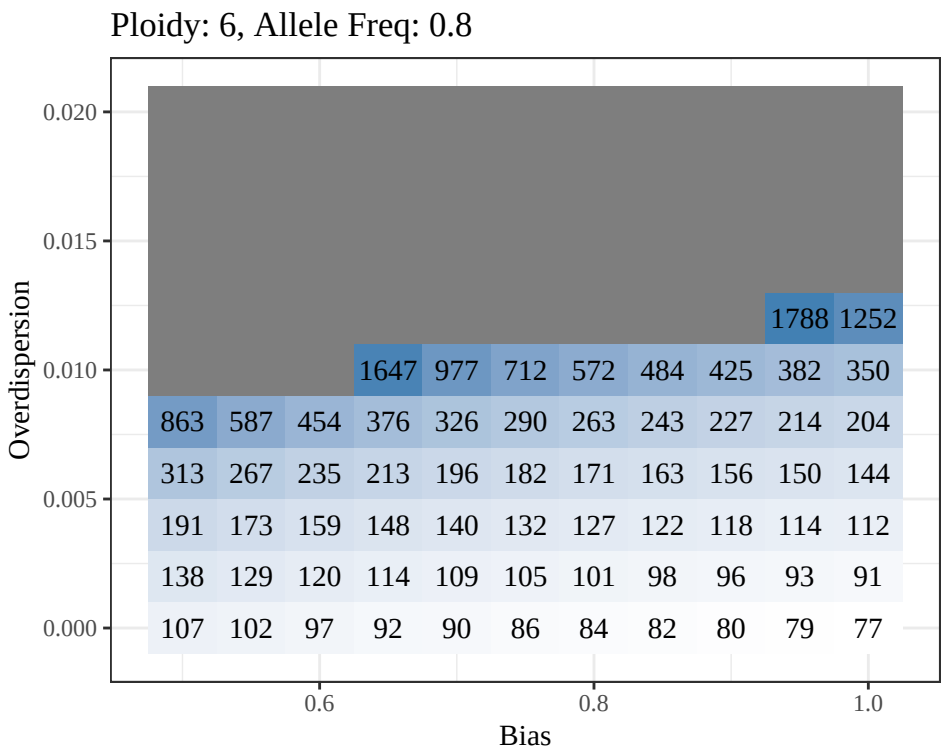

Figure S12 The minimum read depth required to obtain an oracle genotyping error rate less than 0.05 under different levels of overdispersion (y-axis) and allelic bias (x-axis) while the sequencing error rate is fixed at 0.001. This is for a hexaploid population under Hardy-Weinberg equilibrium with a major allele frequency of 0.8. Grey regions require a read depth greater than 3000.

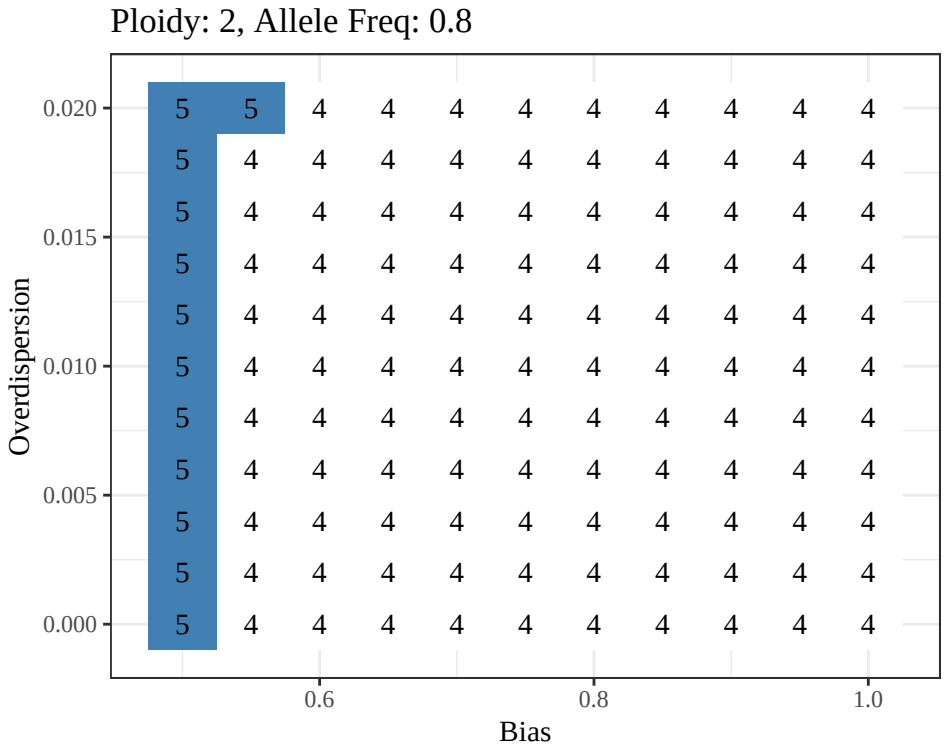

Figure S13 The minimum read depth required to obtain a correlation of over 0.9 between the true and oracly estimated genotypes under different levels of overdispersion (y-axis) and allelic bias (x-axis) while the sequencing error rate is fixed at 0.001. This is for a diploid population under Hardy-Weinberg equilibrium with a major allele frequency of 0.8.

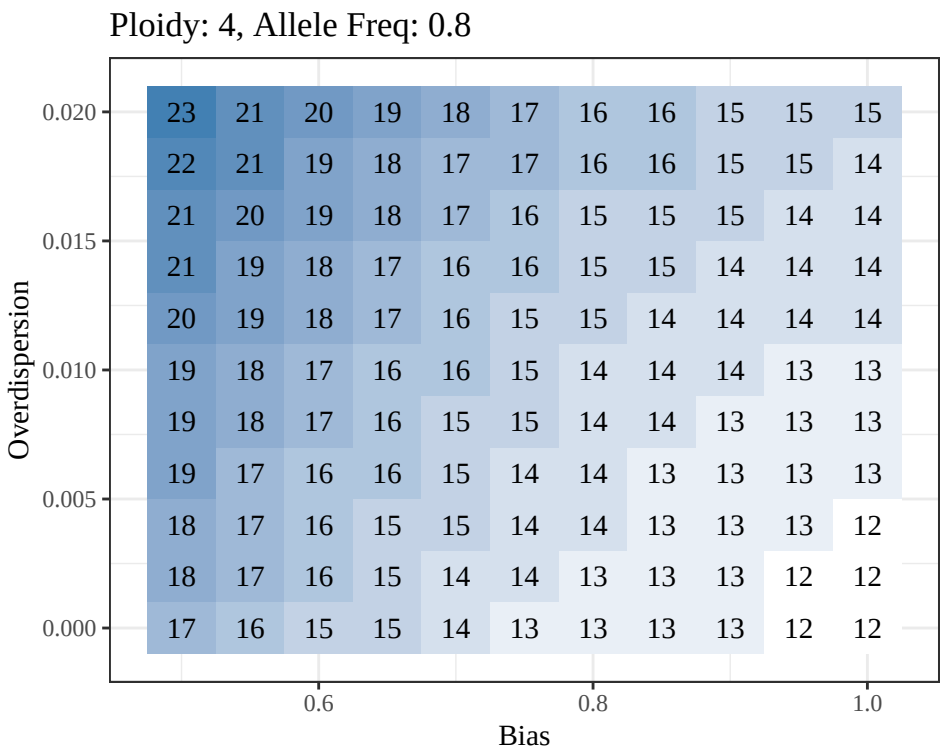

Figure S14 The minimum read depth required to obtain a correlation of over 0.9 between the true and oracly estimated genotypes under different levels of overdispersion (y-axis) and allelic bias (x-axis) while the sequencing error rate is fixed at 0.001. This is for a tetraploid population under Hardy-Weinberg equilibrium with a major allele frequency of 0.8.

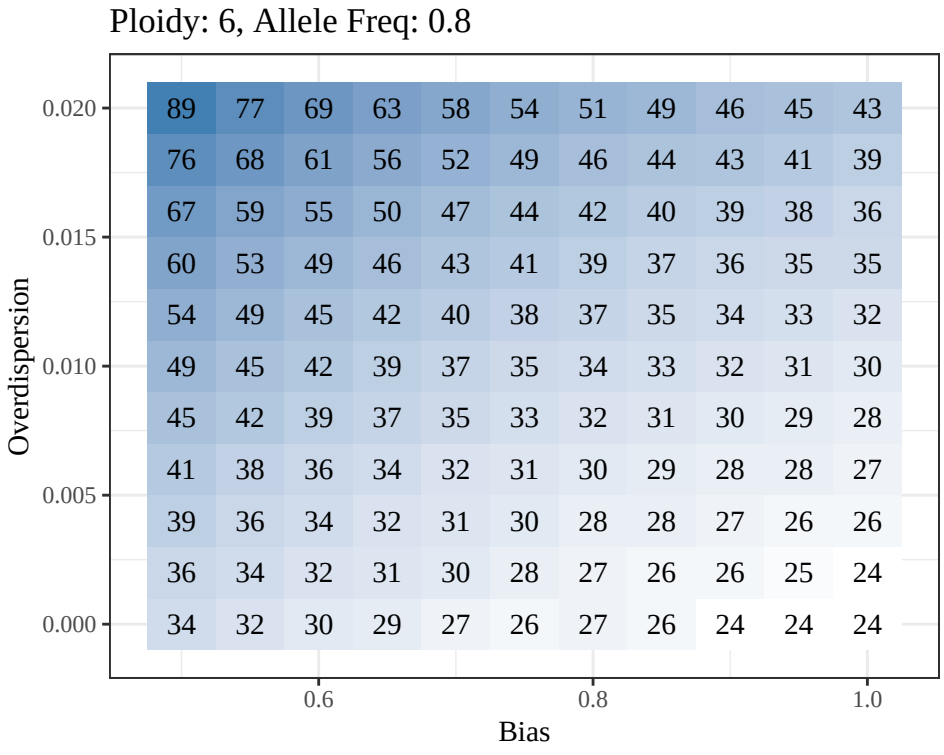

Figure S15 The minimum read depth required to obtain a correlation of over 0.9 between the true and oracly estimated genotypes under different levels of overdispersion (y-axis) and allelic bias (x-axis) while the sequencing error rate is fixed at 0.001. This is for a hexaploid population under Hardy-Weinberg equilibrium with a major allele frequency of 0.8.

Literature Cited

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**: 2987.

Shirasawa, K., M. Tanaka, Y. Takahata, D. Ma, Q. Cao, *et al.*, 2017 A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (*Ipomoea batatas*). Scientific Reports **7**.