

# Genotyping Polyploids from Messy Sequencing Data

David Gerard,<sup>\*,1</sup> Luis Felipe Ventorim Ferrão,<sup>†</sup> Antonio Augusto Franco Garcia,<sup>‡</sup>  
and Matthew Stephens<sup>§,\*\*</sup>

<sup>\*</sup>Department of Mathematics and Statistics, American University, Washington, DC 20016, <sup>†</sup>Horticultural Sciences Department, University of Florida, Gainesville, Florida 32611, <sup>‡</sup>Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, 13418-900, Brazil, and <sup>§</sup>Department of Human Genetics and <sup>\*\*</sup>Department of Statistics, University of Chicago, Illinois 60637

ORCID IDs: 0000-0001-9450-5023 (D.G.); 0000-0002-9655-4838 (L.F.V.F.); 0000-0003-0634-3277 (A.A.G.); 0000-0001-5397-9257 (M.S.)

**ABSTRACT** Detecting and quantifying the differences in individual genomes (*i.e.*, genotyping), plays a fundamental role in most modern bioinformatics pipelines. Many scientists now use reduced representation next-generation sequencing (NGS) approaches for genotyping. Genotyping diploid individuals using NGS is a well-studied field, and similar methods for polyploid individuals are just emerging. However, there are many aspects of NGS data, particularly in polyploids, that remain unexplored by most methods. Our contributions in this paper are fourfold: (i) We draw attention to, and then model, common aspects of NGS data: sequencing error, allelic bias, overdispersion, and outlying observations. (ii) Many datasets feature related individuals, and so we use the structure of Mendelian segregation to build an empirical Bayes approach for genotyping polyploid individuals. (iii) We develop novel models to account for preferential pairing of chromosomes, and harness these for genotyping. (iv) We derive oracle genotyping error rates that may be used for read depth suggestions. We assess the accuracy of our method in simulations, and apply it to a dataset of hexaploid sweet potato (*Ipomoea batatas*). An R package implementing our method is available at <https://cran.r-project.org/package=updog>.

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

**N**EW high-throughput genotyping methods (*e.g.*, Davey *et al.* 2011) allow scientists to pursue important genetic, ecological, and evolutionary questions for any organism, even those for which existing genomic resources are scarce (Chen *et al.* 2014). These methods combine high-throughput sequencing with preparation of a reduced representation library, to sequence a small subset of the entire genome across many individuals. This strategy allows both genome-wide single nucleotide polymorphism (SNP) discovery and SNP genotyping at a reduced cost compared to whole-genome sequencing (Chen *et al.* 2014; Kim *et al.* 2016). Specific examples of these methods include “restriction site-associated DNA sequencing” (RAD-seq) (Baird *et al.* 2008) and “genotyping-by-sequencing” (GBS) (Elshire *et al.* 2011). Both of

these approaches have been widely used in recent biological research, including in population-level analyses (Byrne *et al.* 2013; Schilling *et al.* 2014), quantitative trait loci mapping (Spindel *et al.* 2013), genomic prediction (Spindel *et al.* 2015), expression quantitative trait loci discovery (Liu *et al.* 2017), and genetic mapping studies (Shirasawa *et al.* 2017).

Statistical methods for SNP detection and genotype calling play a crucial role in these new genotyping technologies. And, indeed, considerable research has been performed to develop such methods (Nielsen *et al.* 2011). Much of this research has focused on methods for diploid organisms—those with two copies of their genomes. Here, we focus on developing methods for polyploid organisms—specifically for autopolyploids, which are organisms with more than two copies of their genome of the same type and origin, and present polysomic inheritance (Garcia *et al.* 2013). Autopolyploidy is a common feature in plants, including many important crops (*e.g.*, sugarcane, potato, several forage crops, and some ornamental flowers). More generally, polyploidy plays a key role in plant evolution (Otto and Whitton 2000; Soltis *et al.* 2014) and plant biodiversity (Soltis and Soltis 2000), and understanding polyploidy is important when performing genomic

Copyright © 2018 by the Genetics Society of America

doi: <https://doi.org/10.1534/genetics.118.301468>

Manuscript received August 3, 2018; accepted for publication August 21, 2018; published Early Online September 5, 2018.

Supplemental material available at Figshare: <https://doi.org/10.25386/genetics.7019456>.

<sup>1</sup>Corresponding author: Department of Mathematics and Statistics, American University, 3501 Nebraska Ave. NW, Don Myers Bldg., Washington, DC 20016-8050. E-mail: [dgerard@american.edu](mailto:dgerard@american.edu)

selection and predicting important agronomic traits (Udall and Wendel 2006). Consequently, there is strong interest in genotyping polyploid individuals, and, indeed, the last decade has seen considerable research into genotyping in both non-NGS (next-generation sequencing) data (Voorrips *et al.* 2011; Serang *et al.* 2012; Garcia *et al.* 2013; Bargary *et al.* 2014; Mollinari and Serang 2015; Schmitz Carley *et al.* 2017) and NGS data (McKenna *et al.* 2010; Li 2011; Garrison and Marth 2012; Blischak *et al.* 2016, 2018; Maruki and Lynch 2017; Clark *et al.* 2018).

Here, we will demonstrate that current analysis methods, though carefully thought out, can be improved in several ways. Current methods fail to account for the fact that NGS data are inherently messy. Generally, samples are genotyped at low coverage to reduce cost (Glaubitz *et al.* 2014; Blischak *et al.* 2018), increasing variability. Errors in sequencing from the NGS platforms abound (Li *et al.* 2011) (*Modeling sequencing error*). These are two well-known issues in NGS data. In this paper, we will further show that NGS data also face issues of systematic biases (e.g., resulting from the read-mapping step) (*Modeling allelic bias*), added variability beyond the effects of low-coverage (*Modeling overdispersion*), and the frequent occurrence of outlying observations (*Modeling outliers*). Our first contribution in this paper is highlighting these issues on real data and then developing a method to account for them.

Our second contribution is to consider information from Mendelian segregation in NGS genotyping methods. Many experimental designs in plant breeding are derived from progeny test data (Li *et al.* 2014; Tennessen *et al.* 2014; McCallum *et al.* 2016; Shirasawa *et al.* 2017). Such progenies often result from a biparental cross, including half and full-sib families, or from selfing. This naturally introduces a hierarchy that can be exploited to help genotype individuals with low coverage. Here, we implement this idea in the case of autopolyploids with polysomic inheritance and bivalent non-preferential pairing. Hierarchical modeling is a powerful statistical approach, and others have used its power in polyploid NGS SNP genotyping, not with Mendelian segregation, but in assuming Hardy-Weinberg equilibrium (HWE), or small deviations from HWE (Li 2011; Garrison and Marth 2012; Maruki and Lynch 2017; Blischak *et al.* 2018). Using Mendelian segregation for SNP genotyping has been used in non-NGS data (Serang *et al.* 2012; Schmitz Carley *et al.* 2017), and in diploid NGS data (Zhou and Whittemore 2012), but, to our knowledge, we are the first to implement this for polyploid NGS data—though others have used deviations from Mendelian segregation as a way to filter SNPs (Chen *et al.* 2014, e.g.) or for the related problem of haplotype assembly (Motazed *et al.* 2018).

Though wielding Mendelian segregation in species that exhibit bivalent nonpreferential pairing can be a powerful approach, in some species this assumption can be inaccurate. Chromosomal hom(o)eologues might exhibit partial or complete degrees of preferential pairing (Voorrips and Maliepaard 2012). Our next contribution is developing a model and inference procedures to account for arbitrary

levels of preferential pairing in bivalent pairing species. In particular, this approach allows us to estimate if a species contains completely homologous chromosomes (autopolyploids), completely homeologous chromosomes (allopolyploids), or some intermediate level of chromosomal pairing (Stift *et al.* 2008) (“segmental allopolyploids,” Stebbins 1947). We harness these estimates of preferential pairing to genotype (auto/allo)polyploids.

Our final contribution concerns the question raised by the development of our new model: “How many reads are needed to adequately genotype an individual?” This is a complicated by the fact that read depth calculations depend on many factors: the level of allelic bias, the sequencing error rate, the overdispersion, the distribution of genotypes in the sample, as well as the requirements of the intended downstream analyses. This last consideration is especially difficult to consider when coming up with general guidelines. For example, in association studies, merely a strong correlation with the true genotype is necessary (Pritchard and Przeworski 2001), and a smaller read depth might be adequate, whereas for mapping studies, a few errors can inflate genetic maps (Hackett and Broadfoot 2003), and so a larger read depth might be needed to improve accuracy. Rather than try to account for all of these contingencies, we took the approach of developing functions that a researcher may use to obtain lower bounds on the read depths required based on (i) the specifics of their data, and (ii) the needs of their study. This is in alignment with the strategy in elementary statistics of not providing general guidelines for the sample size of a *t*-test, but rather provide functions to determine sample size based on the variance, effect size, power, and significance level.

Our paper is organized as follows. We develop our method using a real dataset (Shirasawa *et al.* 2017) as motivation in *Methods*. During this development, we highlight several issues and solutions to genotyping NGS data, including overdispersion, allelic bias, and outlying observations. We then evaluate the performance of our method using Monte Carlo simulations (*Simulation comparisons to other methods* through *Simulations on preferential pairing*) and demonstrate its superior genotyping accuracy to competing methods in the presence of overdispersion and allelic bias. We then use our method on a real dataset of hexaploid sweet potato (*Ipomoea batatas*) in *Sweet potato*. Oracle genotyping rates are explored in *Using oracle results for sampling depth guidelines* and *Sampling depth suggestions*. We finish with a discussion and future directions (*Discussion*).

## Methods

We now describe models and methods for genotyping polyploid individuals. The models incorporate several features we have observed in real data. To help highlight these features, and for ease of explanation, we start with a simple model and gradually incorporate each additional feature. Notation is introduced gradually as required, and summarized for convenience in Table 1.

**Table 1 Summary of notation**

Term	Definition
$K$	Ploidy of the species.
$x_{ij}$	$j$ th read for $i$ th individual. Equal to 1 if a read is an A and 0 if a read is an a.
$y_i$	$= \sum_j x_{ij}$ . The number of A reads in individual $i$ .
$n_i$	Number of reads in individual $i$ .
$p_i$	Allele-dosage for individual $i$ , $p_i \in \{0/K, 1/K, \dots, K/K\}$ .
$\theta_i$	Probability sample $i$ is a nonoutlier.
$\tilde{y}_\ell$	Number of A reads in the $\ell$ th parent.
$\tilde{n}_\ell$	Number of reads in the $\ell$ th parent.
$\tilde{p}_\ell$	Allele-dosage for the $\ell$ th parent.
$\tilde{\theta}_\ell$	Probability the sample from parent $\ell$ is a nonoutlier.
$\epsilon$	Sequencing error rate.
$h$	Allelic bias parameter. $Pr(a \text{ observed after selected}) / Pr(A \text{ observed after selected})$ .
$\tau$	Overdispersion parameter for the read counts.
$\pi$	Proportion of samples that are nonoutliers.

To illustrate key features of our model, we give examples from a dataset on autohexaploid sweet potato samples (*I. batatas*) ( $2n = 6x = 90$ ) from a genetic mapping study by Shirasawa *et al.* (2017). These data consist of an S1 population of 142 individuals, genotyped using double-digest RAD-seq technology (Peterson *et al.* 2012). Here and throughout, “S1” refers to a population of individuals produced from the self-pollination of a single parent. We used the data resulting from the SNP selection and filtering procedures described in Shirasawa *et al.* (2017). These procedures included mapping reads onto a reference genome of the related *I. trifida* to identify putative SNPs, and then selecting high-confidence biallelic SNPs with coverage of  $\leq 10$  reads for each sample and with  $< 25\%$  of samples missing, yielding a total of 94,361 SNPs. Further details of the biological materials, assays, and data filtering may be found in Shirasawa *et al.* (2017).

For each SNP, we use A and a to denote the two alleles, with A being the reference allele (defined by the allele on the reference genome of the related *I. trifida*). For each individual the data at each SNP are summarized as the number of reads carrying the A allele and the number carrying the a allele. For a  $K$ -ploid individual there are  $K + 1$  possible genotypes, corresponding to  $0, 1, \dots, K$  copies of the A allele. We use  $p_i \in \{0/K, 1/K, \dots, K/K\}$  to denote the A allele dosage of individual  $i$  (so  $Kp_i$  is the number of copies of allele A). Genotyping an individual corresponds to estimating  $p_i$ .

Figure 1 illustrates the basic genotyping problem using data from a single well-behaved SNP. In this “genotype plot” each point is an individual, with the  $x$  and  $y$  axes showing the number of a and A reads, respectively. The lines in the plot indicate the expected values for possible genotype  $p_i \in \{0/K, 1/K, \dots, K/K\}$ , and are defined by

$$\frac{A}{A + a} = p_i, \quad (1)$$

where  $A$  is the count of A reads and  $a$  is the count of a reads. Genotyping each sample effectively corresponds to determining which line gave rise to the sample’s data. In this case, the

determination is fairly clear because the SNP is particularly well-behaved and most samples have good coverage. Later, we will see examples where the determination is harder.

### Naive model

A simple and natural model is that the reads at a given SNP are independent Bernoulli random variables:

$$x_{ij}|p_i \sim \text{Bernoulli}(p_i), \quad (2)$$

where  $x_{ij}$  is 1 if read  $j$  from individual  $i$  is an A allele, and is 0 if the read is an a allele. The total counts of allele A in individual  $i$  then follows a binomial distribution

$$y_i = \sum_{j=1}^{n_i} x_{ij} \sim \text{Binomial}(n_i, p_i). \quad (3)$$

If the individuals are siblings, then, by the rules of Mendelian segregation, the  $p_i$  values have distribution (Serang *et al.* 2012):

$$\sum_{i=0}^k HG\left(i, \frac{K}{2} \middle| \ell_1, K\right) HG\left(k - i, \frac{K}{2} \middle| \ell_2, K\right) \quad (4)$$

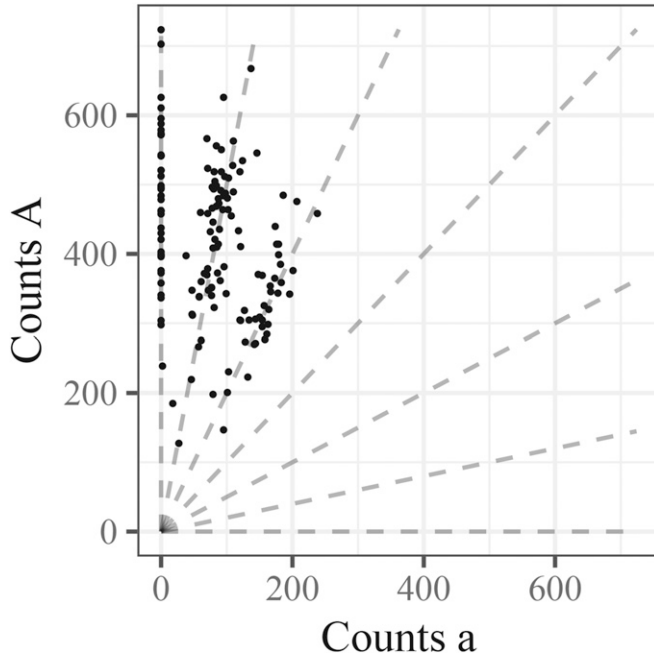
where  $\tilde{p}_j$  is the A allele dosage of parent  $j$ , and  $HG(a, b|c, d)$  is the hypergeometric probability mass function:

$$HG(a, b|c, d) = \frac{\binom{c}{a} \binom{d-c}{b-a}}{\binom{d}{b}}. \quad (5)$$

Equation 4 results from a convolution of two hypergeometric random variables. The distribution (4) effectively provides a prior distribution for  $p_i$ , given the parental genotypes. If the parental genotypes are known, then this prior is easily combined with the likelihood (3) to perform Bayesian inference for  $p_i$ . If the parental genotypes are not known, then it is straightforward to estimate the parental genotypes by maximum likelihood, marginalizing out the  $p_i$ , yielding an empirical Bayes procedure.

### Modeling sequencing error

Though model (2) is a reasonable starting point, it does not account for sequencing error. Even if sequencing errors rates are low (e.g., 0.5–1% (Li *et al.* 2011)), it is crucial to model them because a single error can otherwise dramatically impact genotype calls. In particular, if an individual truly has all reference alleles ( $p_i = 1$ ) then (in the model without errors) a single nonreference allele observed in error would yield a likelihood (and hence posterior probability) of 0 for the true genotype. Biologically, this means that a homozygous individual can be erroneously classified as heterozygous, which may impact downstream analyses (Bourke *et al.* 2018). We demonstrate this in Figure 2 where, in the right panel, we fit the model described in *Naive model* to a single SNP. We labeled six points on Figure 2 with triangles that we think



**Figure 1** Genotype plot of single well-behaved SNP in a hexaploid species. Each point is an individual with the number of alternative reads along the x-axis and the number of reference reads along the y-axis. The dashed lines are defined by  $y/(x+y) = p$  for  $p \in \{0/6, 1/6, \dots, 6/6\}$ , which correspond to genotypes  $\{aaaaaa, Aaaaaa, \dots, AAAAAA\}$

intuitively have a genotype of AAAAAA but were classified as having a genotype of AAAAAa due to the occurrence of one or two a reads.

To incorporate sequencing error, we replace (3) with

$$y_i \sim \text{Binomial}(n_i, f(p_i, \epsilon)), \quad (6)$$

where

$$f(p_i, \epsilon) = p_i(1 - \epsilon) + (1 - p_i)\epsilon, \quad (7)$$

and  $\epsilon$  denotes the sequencing error rate. Estimating this error rate in conjunction while fitting the model in *Naive model* results in intuitive genotyping for the SNP in Figure 2 (left panel). This approach is also used by Li (2011) (and seems more principled than the alternative in Li *et al.* (2014)); see also Maruki and Lynch (2017) for extensions to multi-allelic SNPs.

### Modeling allelic bias

We now address another common feature of these data: systematic bias toward one allele or the other. This is exemplified by the central panel of Figure 3. At first glance, it appears that all offspring have either five or six copies of the reference allele. However, this is unlikely to be the case: since this is an S1 population, if the parent had five copies of the reference allele, we would expect, under Mendelian segregation, the genotype proportions to be (0.25, 0.5, 0.25) for six, five, and four copies of the reference allele, respectively. Indeed, the proportion of individuals with >95% of their

read-counts being the reference allele is 0.197—relatively close to the 0.25 expected proportion for genotype AAAAAA (one-sided  $P$ -value = 0.085). That leaves the other points to represent a mixture of AAAAAa and AAAAAa genotypes. Thus, for this SNP, there appears to be bias toward observing an A read compared to an a read.

One possible source of this bias is the read mapping step (van de Geijn *et al.* 2015). For example, if one allele provides a better match to a different location on the genome than the true location, then this decreases its probability of being mapped correctly. van de Geijn *et al.* (2015) describe a clever and simple technique to adjust for allele-specific bias during the read-mapping step. However, we see three possible problems that may be encountered in using the approach of van de Geijn *et al.* (2015): (i) in some instances, a researcher may not have access to the raw data files to perform this procedure; (ii) the procedure requires access to a reference genome, which is unavailable for many organisms (Lu *et al.* 2013); (iii) there could plausibly be other sources of bias, requiring the development of a method agnostic to the source of bias.

To account for allelic bias, we model sequencing as a two stage procedure: first, reads are chosen to be sequenced (assumed independent of allele); and, second, chosen reads are either “observed” or “not observed” with probabilities that may depend on the allele they carry. Let  $x_{ij}$  denote the random variable that is 1 if the chosen read carries an A allele, and 0 otherwise; and let  $u_{ij}$  denote the random variable that is 1 if the chosen read is actually observed and 0 otherwise. To model the first stage we use

$$[x_{ij}|p_i, \epsilon] \sim \text{Bernoulli}(f(p_i, \epsilon)). \quad (8)$$

To model the second stage we assume

$$[u_{ij}|x_{ij}, c, d] \sim \begin{cases} \text{Bernoulli}(c) & \text{if } x_{ij} = 1 \\ \text{Bernoulli}(d) & \text{if } x_{ij} = 0. \end{cases} \quad (9)$$

Allelic bias occurs when  $c \neq d$ . Since we can only determine the alleles of the reads we observe, we are interested in the distribution of  $x_{ij}$  conditioned on  $u_{ij} = 1$ , which is given by Bayes rule:

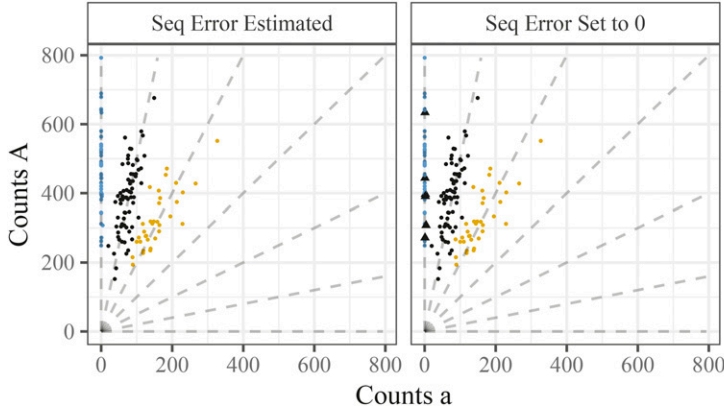
$$[x_{ij}|u_{ij} = 1, p_i, \epsilon, c, d] \sim \text{Bernoulli}(\xi(p_i, \epsilon, c, d)), \quad (10)$$

$$\xi(p_i, \epsilon, c, d) = \frac{cf(p_i, \epsilon)}{d(1 - f(p_i, \epsilon)) + cf(p_i, \epsilon)}. \quad (11)$$

Notice that  $\xi$  depends on  $c$  and  $d$  only through the ratio  $h := d/c$ . Specifically:

$$\xi(p_i, \epsilon, h) = \frac{f(p_i, \epsilon)}{h(1 - f(p_i, \epsilon)) + f(p_i, \epsilon)}. \quad (12)$$

We refer to  $h$  as the “bias parameter,” which represents the relative probability of a read carrying the two different alleles being observed after being chosen to be sequenced. For



**Figure 2** Two genotype plots demonstrating the need to model sequencing error. A SNP from hexaploid sweet potato genotyped in conjunction with estimating the sequencing error rate (left panel), or by setting the sequencing error rate to 0 (right panel). Triangles are points that we think look mis-classified.

example, a value of  $h = 1/2$  means that an A read is twice as probable to be correctly observed than an a read, while a value of  $h = 2$  means that an a read is twice as probable to be correctly observed than an A read.

Both the bias parameter  $h$  and the sequencing error rate  $\epsilon$  modify the expected allele proportions for each genotype. However, they do so in different ways: lower values of  $h$  push the means toward the upper left of the genotype plot, while higher values push the means toward the lower right. Higher values of  $\epsilon$  tend to squeeze the means toward each other. These different effects are illustrated in Figure 4.

### Modeling overdispersion

Overdispersion refers to additional variability than expected under a simple model. Overdispersion is a common feature in many datasets. In sequencing experiments, overdispersion could be introduced by variation in the processing of a sample, by variations in a sample's biochemistry, or by added measurement error introduced by the sequencing machine. These could all result in observed read-counts being more dispersed than expected under the simple binomial model (3). Figure 5 (which is an annotated version of the left panel in Figure 3) illustrates overdispersion in these data.

To model overdispersion, we replace the binomial model with a beta-binomial model (Skellam 1948). The beta-binomial model assumes that each individual draws their own individual-specific mean probability from a beta-distribution, then draws their counts conditional on this probability:

$$[q_i | p_i, \epsilon, h, \tau] \sim \text{Beta}\left(\xi(p_i, \epsilon, h) \frac{1 - \tau}{\tau}, [1 - \xi(p_i, \epsilon, h)] \frac{1 - \tau}{\tau}\right) \quad (13)$$

$$[y_i | n_i, q_i] \sim \text{Binomial}(n_i, q_i). \quad (14)$$

Here  $\xi(p_i, \epsilon, h)$  (12) is the mean of the underlying beta-distribution, and  $\tau \in [0, 1]$  is the overdispersion parameter, with values closer to 0 indicating less overdispersion and values closer to 1 indicating greater overdispersion. (The parameter  $\tau$  can also be interpreted as the “intra-class correlation” (Crowder 1979)). The  $q_i$ 's are dummy variables representing individual

specific probabilities, and are integrated out in the following analyses. We denote the marginal distribution of  $y_i$  as

$$[y_i | n_i, p_i, \epsilon, h, \tau] \sim \text{BB}(n_i, \xi(p_i, \epsilon, h), \tau). \quad (15)$$

### Modeling outliers

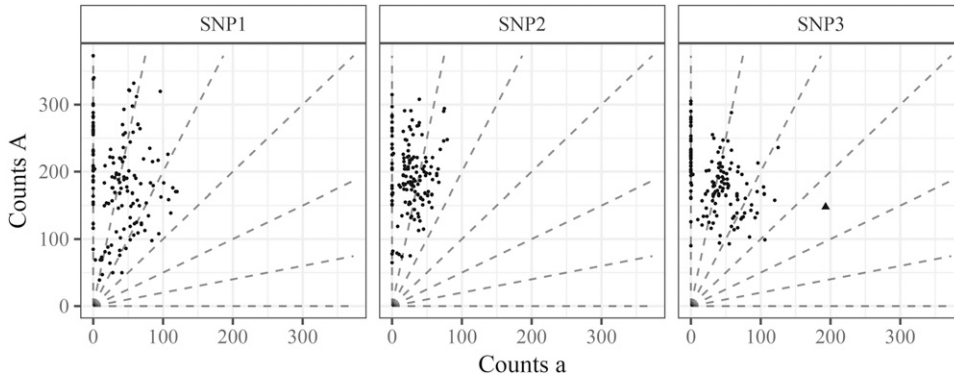
The right panel of Figure 3 illustrates another important feature of real data: outliers. Here, most of the points appear well-behaved, but one point (marked as a triangle) is far from the remaining points. Furthermore, taking account of the fact that these data came from an S1 population, this outlying point is inconsistent with the other points. This is because the other points strongly suggest that the parental genotype is AAAAAa, meaning that the possible genotypes are AAAAAA, AAAAAa, and AAAAAa, and the triangle lies far from the expectation for any of these genotypes. Indeed, under a fitted beta-binomial model, if the individual's genotype were AAAAAa, then the probability of seeing as few or fewer A counts in this individual as were actually observed is  $8.7 \times 10^{-6}$  (Bonferroni-corrected  $P$ -value of 0.0012). There are many possible sources of outliers like this, including individual-specific quirks in the amplification or read-mapping steps, and sample contamination or mislabeling.

There are several common strategies for dealing with outliers. These include using methods that are inherently robust to outliers (Huber 1964); identifying and removing outliers prior to analysis (Hadi and Simonoff 1993); or modeling the outliers directly (Aitkin and Wilson 1980). Here, we take this last approach, which has the advantage of providing measures of uncertainty for each point being an outlier.

Specifically, we model outliers using a mixture model:

$$[y_i | n_i, p_i, \epsilon, h, \tau, \pi] \sim \pi \text{BB}(n_i, \xi(p_i, \epsilon, h), \tau) + (1 - \pi) \text{BB}(n_i, 1/2, 1/3), \quad (16)$$

where  $\pi$  represents the proportion of points that are not outliers. Here, the second component of this mixture represents the outliers, and the parameters  $(1/2, 1/3)$  of this outlying distribution were chosen so that the underlying beta distribution is uniform on  $[0, 1]$ . (We also tried estimating the underlying beta parameters for the outlying distribution,



**Figure 3** Three genotype plots of SNPs demonstrating common features of the data from Shirasawa *et al.* (2017), considering autohexaploid sweet potato. Overdispersion (left), bias (middle), and outlying observations (right).

but we found that this can overfit the data in some instances; not shown.)

#### Prior on sequencing error rate and bias parameter

In some cases we found that maximum likelihood estimation of the bias parameter  $h$  and error rate  $\epsilon$  gave estimates that were unrealistic, and led to undesirable results. For example, we often observed this problem at SNPs where all individuals carry the same genotype (*i.e.*, monomorphic markers).

Figure 6A shows an example of this problem for simulated data from an autotetraploid species in which an AAAA parent is crossed with an aaaa parent, which results in all offspring having the same genotype AAaa. Using the model described up to now yields a maximum likelihood estimate of sequencing error rate  $\hat{\epsilon} = 0.37$ , which is unrealistically high. This further creates very poor genotype calls (Figure 6B).

To avoid this problem, we place priors on  $\epsilon$  and  $h$  to capture the fact that  $\epsilon$  will usually be small, and  $h$  will not deviate too far from 1. Specifically we use

$$\text{logit}(\epsilon) \sim N(\mu_\epsilon, \sigma_\epsilon^2), \text{ and} \quad (17)$$

$$\text{log}(h) \sim N(\mu_h, \sigma_h^2) \quad (18)$$

with software defaults  $\mu_\epsilon = -4.7$ ,  $\sigma_\epsilon^2 = 1$ ,  $\mu_h = 0$ , and  $\sigma_h^2 = 0.7^2$ . With these defaults, 95% of the prior mass is on  $\epsilon \in [0.0012, 0.061]$  and  $h \in [0.25, 4.1]$  (see Supplemental Material, Figure S1 for graphical depictions.). We chose these defaults based on empirical observations (Figure S8), inflating the variances to make the priors more diffuse (less informative) than what we observe in practice. Our prior on the sequencing error rate is also consistent with literature estimates from various platforms (Goodwin *et al.* 2016), though some sequencing technologies can have error rates as high as 15% and, if using such technologies, the prior should be adjusted accordingly.

Ideally, one would incorporate these prior distributions into a full Bayesian model, and integrate over the resulting posterior distribution on  $\epsilon$  and  $h$ . However, this would require nontrivial computation, and we take the simpler approach of simply multiplying the likelihood by the priors (17) and (18) and maximizing this product in place of the likelihood.

(Effectively this corresponds to optimizing a penalized likelihood.) See *Model* for details.

Using both these priors results in accurate genotypes for our simulated example (Figure 6C).

#### Incorporating parental reads

In genetic studies involving mating designs, researchers almost always have NGS data on parent(s) as well as offspring. Such data can be easily incorporated into our model.

Specifically, we model the number of A reads in parent  $\ell$  (denoted  $\tilde{y}_\ell$ ) given the total number of reads in parent  $\ell$  (denoted  $\tilde{n}_\ell$ ) and allelic dosage  $p_\ell$  by:

$$[\tilde{y}_\ell | \tilde{n}_\ell, \tilde{p}_\ell, \epsilon, h, \tau, \pi] \sim \pi \text{BB}(\tilde{n}_\ell, \xi(\tilde{p}_\ell, \epsilon, h), \tau) + (1 - \pi) \text{BB}(\tilde{n}_\ell, 1/2, 1/3). \quad (19)$$

We treat the parental read count data as independent of the offspring count data (given the underlying genotypes), so the model likelihood is the product of (19) and (6).

#### Model

We now summarize our model and fitting procedure.

Our model is:

$$[y_i | n_i, p_i = k_i/K, \epsilon, h, \tau, \pi] \sim \pi \text{BB}(n_i, \xi(k_i/K, \epsilon, h), \tau) + (1 - \pi) \text{BB}(n_i, 1/2, 1/3), \quad (20)$$

$$[\tilde{y}_j | \tilde{n}_j, \tilde{p}_j = \ell_j/K, \epsilon, h, \tau, \pi] \sim \pi \text{BB}(\tilde{n}_j, \xi(\ell_j/K, \epsilon, h), \tau) + (1 - \pi) \text{BB}(\tilde{n}_j, 1/2, 1/3) \quad (21)$$

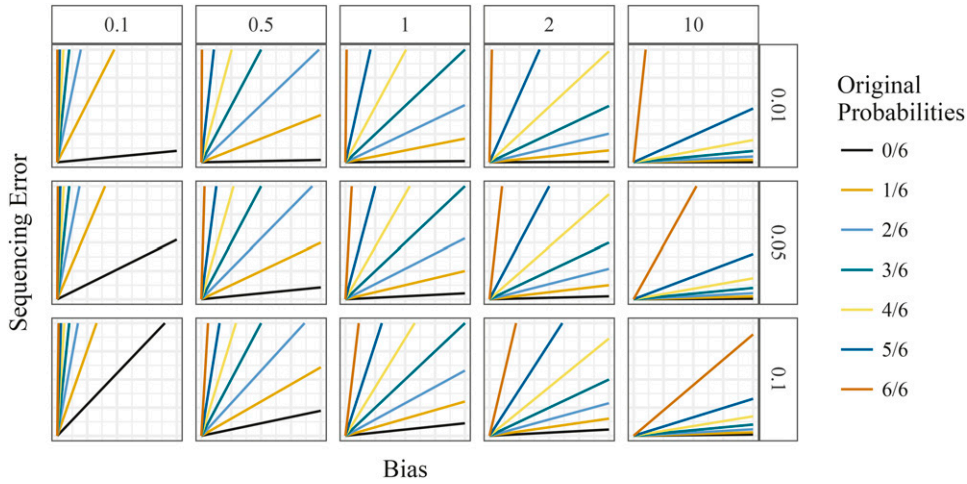
$$\sum_{j=0}^{k_i} \text{HG}\left(j, \frac{K}{2} \middle| \ell_1, K\right) \text{HG}\left(k_i - j, \frac{K}{2} \middle| \ell_2, K\right) \quad (22)$$

$$\text{logit}(\epsilon) \sim N(\mu_\epsilon, \sigma_\epsilon^2), \quad (23)$$

$$\text{log}(h) \sim N(\mu_h, \sigma_h^2). \quad (24)$$

To fit this model for offspring from two shared parents (an F1 cross), we first estimate  $\epsilon$ ,  $h$ ,  $\tau$ ,  $\pi$ ,  $\ell_1$ , and  $\ell_2$  via maximum likelihood (or, for  $h$ ,  $\epsilon$ , the posterior mode):





**Figure 4** Mean counts of the mapped reads under different levels of sequencing error rates (row facets) and different levels of allelic bias (column facets), considering an autohexaploid locus. The y-axis is the number of reference counts, and the x-axis is the number of alternative counts.

$$\begin{aligned}
 & \arg \max_{(\epsilon, \hat{h}, \tau, \pi, \ell_1, \ell_2) \in [0,1] \times \mathbb{R}^+ \times [0,1] \times [0,1] \times 0:K \times 0:K} p(h)p(\epsilon) \\
 & \times p(\tilde{y}_1 | \tilde{n}_1, \ell_1/K, \epsilon, h, \tau, \pi) p(\tilde{y}_2 | \tilde{n}_2, \ell_2/K, \epsilon, h, \tau, \pi) \\
 & \times \prod_i \sum_{k_i=0}^K p(y_i | n_i, k_i/K, \epsilon, h, \tau, \pi) p(k_i/K | \ell_1/K, \ell_2/K).
 \end{aligned} \quad (25)$$

We perform this maximization using an expectation maximization (EM) algorithm, presented in File S1. The M-step (S5) involves a quasi-Newton optimization for each possible combination of  $\ell_1$  and  $\ell_2$ , which we implement using the `optim` function in R (R Core Team 2017).

Given estimates  $\hat{\epsilon}, \hat{h}, \hat{\tau}, \hat{\pi}, \hat{\ell}_1$ , and  $\hat{\ell}_2$ , we use Bayes' theorem to obtain the posterior probability of the individuals' genotypes:

$$\begin{aligned}
 & Pr\left(p_i = \frac{k_i}{K} \middle| y_i, \hat{\epsilon}, \hat{h}, \hat{\tau}, \hat{\pi}, \hat{\ell}_1, \hat{\ell}_2\right) \\
 & = \frac{p\left(y_i \middle| n_i, \frac{k_i}{K}, \hat{\epsilon}, \hat{h}, \hat{\tau}, \hat{\pi}\right) Pr\left(p_i = \frac{k_i}{K} \middle| \tilde{p}_1 = \frac{\hat{\ell}_1}{K}, \tilde{p}_2 = \frac{\hat{\ell}_2}{K}\right)}{\sum_{k_i=0}^K p\left(y_i \middle| n_i, \frac{k_i}{K}, \hat{\epsilon}, \hat{h}, \hat{\tau}, \hat{\pi}\right) Pr\left(p_i = \frac{k_i}{K} \middle| \tilde{p}_1 = \frac{\hat{\ell}_1}{K}, \tilde{p}_2 = \frac{\hat{\ell}_2}{K}\right)}.
 \end{aligned} \quad (26)$$

From this we can obtain, for example, the posterior mean genotype or a posterior mode genotype for each individual. The  $\theta_i$  values from the E-step in (S2) in File S1 may be interpreted as the posterior probability that a point is a nonoutlier.

Modifying the EM algorithm in File S1 to deal with offspring from an S1 (instead of F1) cross is straightforward: simply constrain  $\ell_1 = \ell_2$  and remove  $p(\tilde{y}_2 | \tilde{n}_2, \ell_2/K, \epsilon, h, \tau, \pi)$  from (25).

This procedure is implemented in the R package `updog` (using parental data for offspring genotyping).

### Screening SNPs

No matter the modeling decisions made, there will likely be poorly behaved SNPs/individuals whose genotypes are

unreliable. Such poorly behaved SNPs/individuals may originate from sequencing artifacts, and it is important to consider identifying and removing them. The Bayesian paradigm we use naturally provides measures of genotyping quality, both at the individual and SNP levels.

Let  $q_i$  be the maximum posterior probability of individual  $i$ . That is,

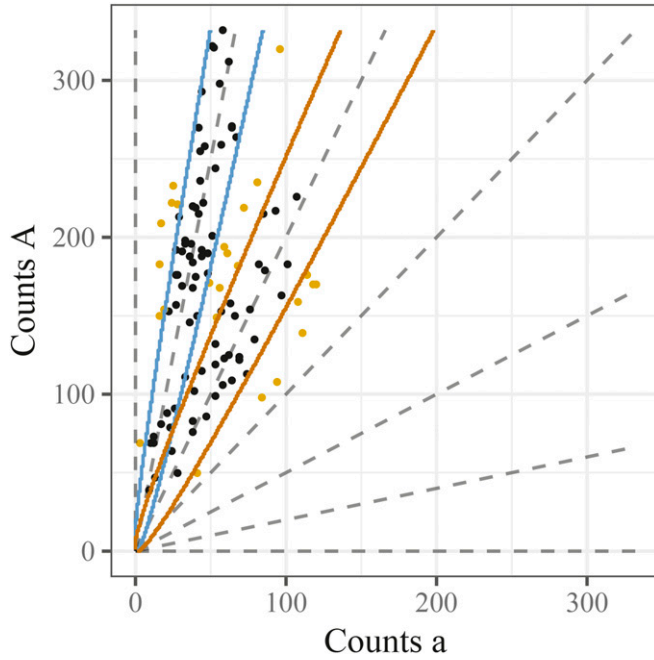
$$q_i := \max_k Pr(p_i = k_i/K | y_i, \hat{\epsilon}, \hat{h}, \hat{\tau}, \hat{\pi}, \hat{\ell}_1, \hat{\ell}_2), \quad (27)$$

where the right-hand side of (27) is defined in (26). Then, the posterior probability of an individual being genotyped incorrectly is  $1 - q_i$ , and individuals may be filtered based on this quantity. That is, if a researcher wants to have an error rate  $< 0.05$  then they would remove individuals with  $1 - q_i > 0.05$ . The overall error rate of a SNP is  $r := 1 - (1/n) \sum_i q_i$ , where  $n$  is the number of individuals in the sample. That is,  $r$  is the posterior proportion of individuals genotyped incorrectly at the SNP. Whole SNPs that have a large value of  $r$  may be discarded.

We evaluate the accuracy of  $r$  in *Simulation comparisons to other methods*.

### Accounting for preferential pairing

Up until now, we have used a prior (4) that assumes a particular form of meiotic segregation—autopolyploids with polysomic inheritance and bivalent nonpreferential pairing. If working with a species with different meiotic behavior—or a species where the meiotic behavior is unknown *a priori* (Bourke *et al.* 2017)—one should adjust this prior accordingly. In this section, we develop a model and inference procedures that account for one type of deviation from this ideal Mendelian segregation scenario: arbitrary levels of preferential pairing. Note, though, that there are other ways to deviate from Mendelian segregation that we do not cover in this section. For example, if multivalent pairing were allowed, then we would need to adjust our prior to account for the possibility of double reduction (Stift *et al.* 2010). However, multivalent pairing can occur relatively infrequently in many



**Figure 5** A genotype plot illustrating overdispersion compared with the simple binomial model. This figure shows the same SNP as the left panel of Figure 3 but with the points with  $>95\%$  A reads removed. Solid lines indicate the 0.025 and 0.975 quantiles for the Binomial distribution with probabilities  $4/6$  (red) and  $5/6$  (blue). Points that lie within these lines are colored black; outside are colored orange. Under the binomial model only, 5% of the points should be orange, but, in fact, a significantly higher proportion (23.6%;  $P$ -value  $1.2 \times 10^{-11}$ ) are orange.

stable autopolyploids (Bomblies *et al.* 2016), and, when it does occur, the levels of double reduction can be relatively low (Stift *et al.* 2008), and so the method below might act as a reasonable approximation in the presence of multivalent pairing. Another example of Mendelian violations would be distortion resulting from an allele being semilethal.

Preferential pairing describes the phenomenon where some homologous chromosomes are more likely to pair during meiosis. Preferential pairing affects the segregation patterns of haplotypes, and, hence, the distribution of genotypes in a sample. For a good overview of preferential pairing, see Voorrips and Maliepaard (2012).

To begin, we define a particular set of chromosome pairings to be a “pairing configuration.” Again, we suppose that all pairing during meiosis is bivalent. Since chromosomes are not labeled (we are only interested in the counts of reference and alternative alleles, and not the specific chromosome carrying these alleles), we may represent each configuration by a 3-tuple  $\mathbf{m} \in \mathbb{N}^3$ , where the first element contains the number of aa pairs, the second element contains the number of Aa pairs, and the third element contains the number of AA pairs. For example, for the two configurations of a hexaploid parent with three copies of A,

$$\text{Aa} \mid \text{aa} \mid \text{AA} \text{ and } \text{Aa} \mid \text{Aa} \mid \text{Aa},$$

we represent the first configuration by  $\mathbf{m} = (1, 1, 1)$  and the second by  $\mathbf{m} = (0, 3, 0)$ .

Given a configuration  $\mathbf{m}$ , a  $K$ -ploid parent will segregate  $K/2$  chromosomes to an offspring. The counts of reference alleles that segregate to the offspring, which we denote by  $z$ , will follow an “off-center” binomial distribution:

$$[z - m_3 \mid \mathbf{m}] \sim \text{Bin}(m_2, 1/2), \quad (28)$$

since, in aa pairs, no A alleles segregate; in AA pairs, only A alleles segregate; and in Aa pairs, an A allele segregates with probability  $1/2$ . However, the configuration that forms  $\mathbf{m}$  is a random variable. Suppose there are  $q(\ell)$  possible configurations [see Theorem S1 in File S1 for deriving  $q(\ell)$ ] given a parent has  $\ell$  copies of A. Also suppose configuration  $\mathbf{m}_i$  has probability  $\gamma_i$  of forming. Then the distribution of  $z$  is a mixture of (28), with density

$$p(z \mid \gamma, \ell) = \sum_{i=1}^{q(\ell)} \gamma_i \text{Bin}(z - m_{i3} \mid m_{i2}, 1/2). \quad (29)$$

The values of  $\gamma_i$  in (29) determine the degree and strength of preferential pairing. If there is no preferential pairing, then the segregation probabilities (29) will equal those from the hypergeometric distribution (5). We derive the values of  $\gamma_i$  that result in the hypergeometric distribution in Theorem S5 in File S1. Any preferential pairing will result in deviations away from the weights in Theorem S5.

If neither parent in an F1 population exhibits preferential pairing, then the genotype distribution will follow (4). However, deviations from (4) can occur if some bivalent pairings occur with a higher probability than random chance would allow (Theorem S5 in File S1). In which case, the offspring genotype distribution would be a convolution of two densities of family (29). Specifically, if parent 1 has genotype  $\ell_1$  and configuration probabilities  $\gamma_1$ , and parent 2 has genotype  $\ell_2$  and configuration probabilities  $\gamma_2$ , then the offspring genotype distribution is

$$\Pr(p_i = k/K \mid \ell_1, \gamma_1, \ell_2, \gamma_2) = \sum_{i=0}^K p(i \mid \gamma_1, \ell_1) p(k - i \mid \gamma_2, \ell_2). \quad (30)$$

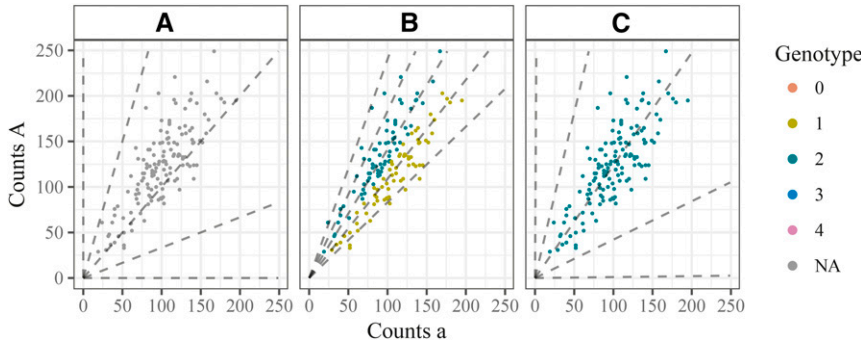
Allowing for arbitrary levels of preferential pairing is simply replacing prior (4) with (30). We have made this modification in updog, where we estimate both the parental genotypes ( $\ell_1$  and  $\ell_2$ ) and the degree and strength of the preferential pairing ( $\gamma_1$  and  $\gamma_2$ ) by maximum marginal likelihood.

This model for preferential pairing is similar to that of Stift *et al.* (2008), although there are major differences: (i) they do not use this model for genotyping, (ii) the chromosomes are not labeled in our setting and are in theirs, and (iii) we allow for arbitrary ploidy levels while they just allow for tetraploidy.

### Extension to population studies

We have focused here on data from an F1 (or S1) experimental design, using parental information to improve genotype calls.





**Figure 6** Three genotype plots illustrating the need to place priors on the bias and sequencing error rate parameters. Simulated autotetraploid NGS data where all individuals have the AAaa genotype. (A) Unlabeled NGS data, (B) labeled by a fit without any penalties on the bias and sequencing error rate, and (C) labeled by a fit with penalties on both the bias and the sequencing error rate.

However, a similar approach can also be applied to other samples (e.g., outbred populations) by replacing the prior on offspring genotypes (4) with another suitable prior. For example, previous studies have used a discrete uniform distribution (McKenna *et al.* 2010), a binomial distribution that results from assuming HWE (Li 2011; Garrison and Marth 2012), and a Balding-Nichols beta-binomial model on the genotypes (Balding and Nichols 1995, 1997) that assumes an individual contains the same overdispersion parameter across loci (Blischak *et al.* 2018). (All of these previous methods use models that are more limited than the one we present here: none of them account for allelic bias, outliers, or locus-specific overdispersion; and most implementations assume the sequencing error rate is known).

In our software, we have implemented both the uniform and binomial (HWE) priors on genotypes. The former is very straightforward. The latter involves modifying the EM algorithm in File S1 by replacing  $a_{k_i \ell_1 \ell_2}$  in (S1) with the binomial probability  $\Pr(k_i | n_i, \alpha)$ , where  $\alpha$  is the allele frequency of A, and then optimizing over  $(\alpha, \tau, h, \epsilon)$  in (S5).

Though the uniform distribution might seem like a reasonable “noninformative” prior distribution, in practice it can result in unintuitive genotyping estimates. This is because the maximum marginal likelihood approach effectively estimates the allelic bias  $h$  and sequencing error rate  $\epsilon$  in such a way that the estimated genotypes approximate the assumed prior distribution. As most populations do not exhibit a uniform genotype distribution, this can result in extreme estimates of the bias and sequencing error rates, resulting in poor genotype estimates. We thus strongly recommend against using the uniform prior in practice, and suggest researchers use more flexible prior distributions, such as the binomial prior.

### Using oracle results for sampling depth guidelines

In this section, we develop oracle rates for incorrectly genotyping individuals based on our new likelihood (20). These rates are “oracle” in that they are the lowest achievable by an ideal estimator of genotypes. Using these oracle rates, we provide a method that returns a lower bound on read depths required to get an error rate less than some threshold. This method also returns the correlation of the oracle estimator with the true genotype, which might be useful when choosing the read depths for association studies.

To begin, we suppose that the sequencing error rate  $\epsilon$ , bias  $h$ , overdispersion  $\tau$ , and genotype distribution  $\pi(k/K)$  are all known. That is, the probability of dosage  $p_i = k/K$  is  $\pi(k/K)$ . Developing read depth suggestions when these quantities are known will provide lower bounds on the read depths required when they are not known. However, standard maximum likelihood theory guarantees that the following approximations will be accurate for a large enough sample of individuals (even for low read depth).

Given that these quantities are known, we have a simplified model with a likelihood  $p(y_i | p_i = k/K)$  [defined in (20)] and a prior  $\pi(p_i)$ , where the only unknown quantity is the allele dosage  $p_i$ . Under 0/1 loss, the best estimator of  $p_i$  is the maximum a posteriori (MAP) estimator

$$\hat{p}_i = \arg \max_{p_i} p(y_i | p_i) \pi(p_i). \quad (31)$$

Notice that  $\hat{p}_i$  is only a function of  $y_i$  and not  $p_i$ . Thus, under this simplified model, one can derive the joint distribution of  $p_i$  and  $\hat{p}_i$  by summing over the possible counts  $y_i$ .

$$\Pr(p_i = k/K, \hat{p}_i = \ell/K) = \sum_{y_i: \hat{p}_i(y_i) = \ell/K} p(y_i | k/K) \pi(k/K). \quad (32)$$

Given this joint distribution (32), one can calculate the oracle misclassification error rate

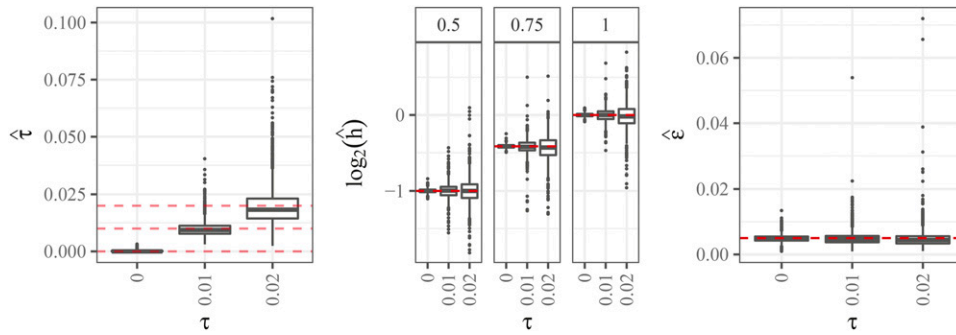
$$\Pr(p_i \neq \hat{p}_i) = 1 - \sum_{k=0}^K \Pr(p_i = k/K, \hat{p}_i = k/K). \quad (33)$$

One can also calculate the oracle misclassification error rate conditioned on the individual’s genotype

$$\Pr(p_i \neq \hat{p}_i | p_i = k/K) = 1 - \Pr(p_i = k/K, \hat{p}_i = k/K) / \pi(k/K). \quad (34)$$

Also possible is calculating the correlation between the MAP estimator and the true genotype. All of these calculations have been implemented in the updog R package.

For sample size calculations, a researcher could specify the largest levels of bias, overdispersion, and sequencing error that they expect in their data. They could then specify a few genotype distributions that they expect (which is a finite



**Figure 7** Left: Boxplots of estimated overdispersion parameter (y-axis) stratified by the value of overdispersion (x-axis). The horizontal lines are at 0, 0.01, and 0.02. Center: Boxplots of  $\log_2$ -transformed estimates of the bias parameter (y-axis) stratified by the value of overdispersion (x-axis). The column facets distinguish different levels of the true bias parameter (with 1 indicating no bias). The horizontal lines are at the true bias parameter. Right: Boxplots of estimated sequencing error rate (y-axis) stratified by the value of overdispersion (x-axis). The horizontal line is at the true sequencing error rate of 0.005.

number if the sample is from an F1 population), and then find the read depths required to obtain the oracle sequencing error rate (33) below some threshold that depends on their domain of application. We provide an example of this in *Sampling depth suggestions*.

In the machine learning community, Equation (33) is often called the Bayes error rate (Hastie *et al.* 2009, Section 2.4).

#### Data availability

The methods implemented in this paper are available in the updog R package available on the Comprehensive R Archive Network at <https://cran.r-project.org/package=updog>. Code and instructions to reproduce all of the results of this paper (doi: <https://doi.org/10.5281/zenodo.1203697>) are available at [https://github.com/dcgerard/reproduce\\_genotyping](https://github.com/dcgerard/reproduce_genotyping). Supplemental figures, and supplemental text containing optimization details, theorems, and proofs, have been deposited at Figshare. Supplemental material available at Figshare: <https://doi.org/10.25386/genetics.7019456>.

## Results

### Simulation comparisons to other methods

We ran simulations to evaluate updog's ability to estimate model parameters and genotype individuals in hexaploid species. Since most competing methods do not allow for an F1 population of individuals [although see Serang *et al.* (2012) implemented at <https://bitbucket.org/orserang/supermassa.git>], we compared the HWE version of updog (*Extension to population studies*), a GATK-like method (hereafter, "GATK") (McKenna *et al.* 2010) as implemented by updog (by using a discrete uniform prior on the genotypes, fixing  $h$  to 1 and  $\tau$  to 0, while specifying  $\epsilon$ ), fitPoly (Voorrips *et al.* 2011), and the method of Li (2011) as implemented by Blischak *et al.* (2018). Other methods are either specific to tetraploids, or only have implementations for tetraploids (Maruki and Lynch 2017; Schmitz Carley *et al.* 2017), so we did not explore their performances.

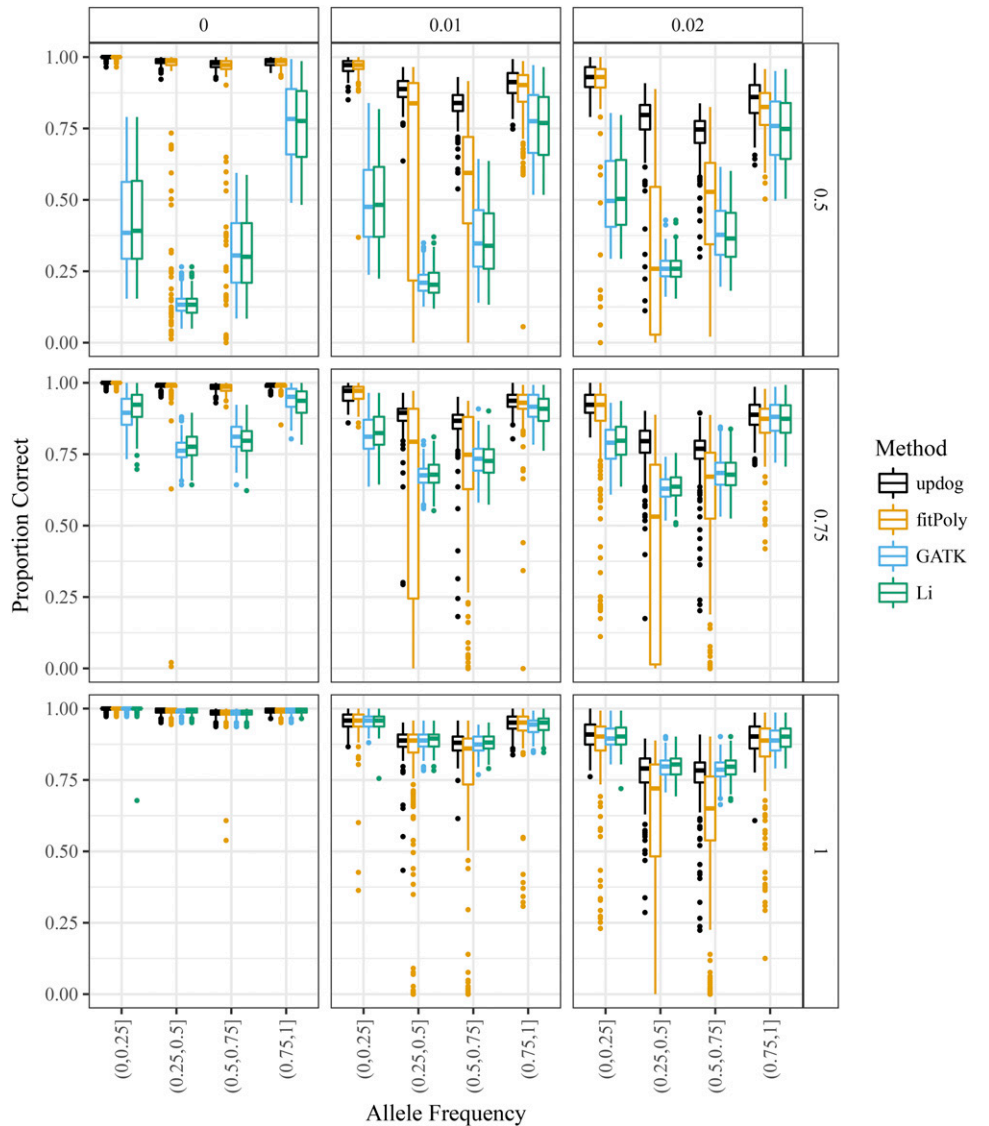
Specifically, we simulated 142 unrelated individuals [the number of individuals in the dataset from Shirasawa *et al.*

(2017)] under the updog model with sequencing error rate  $\epsilon = 0.005$ , overdispersion parameter  $\tau \in \{0, 0.01, 0.02\}$ , and bias parameter  $h \in \{0.5, 0.75, 1\}$  (with  $h = 1$  indicating no bias). These parameter values were motivated by features in real data (Figure S8). We did not allow for outliers, neither in the simulated data nor in the fit. For each combination of  $\tau$  and  $h$ , we simulated 1000 datasets. The  $n_i$  values were obtained from the 1000 SNPs in the Shirasawa *et al.* (2017) dataset with the largest read-counts. The distribution of the genotypes for each locus was distributed binomially using an allele frequency chosen from a uniform grid from 0.05 to 0.95.

Figure 7 contains boxplots for parameter estimates from the updog model. In general, the parameter estimates are highly accurate for small values of overdispersion and become less accurate, though still approximately unbiased, for larger values of overdispersion. The accuracy of the parameter estimates vary gradually for different levels of allele frequencies (Figures S4 and S5). We note that, in real data, only a small fraction of estimates of the overdispersion parameter are higher than 0.02 (Figure S8).

We then compared the genotyping accuracy of updog with the method from Li (2011), GATK (McKenna *et al.* 2010), and fitPoly (Voorrips *et al.* 2011). In Figure 8, we have boxplots of the proportion of samples genotyped correctly stratified by bins of the allele frequency, color coding by method. We draw the following conclusions:

1. All methods show similar performance when there is no bias ( $h = 1$ ) and no overdispersion ( $\tau = 0$ ). These are the (unstated) assumptions of GATK and the method of Li (2011).
2. Overdispersion generally makes estimating genotypes much more difficult, and accurate genotyping can only be guaranteed for small levels of overdispersion.
3. When there is any bias, updog has much superior performance to GATK and the method of Li (2011).
4. fitPoly performs well when there is no overdispersion, even for large levels of bias. However, in the presence of small amounts of overdispersion, fitPoly provides unstable genotype estimates.



**Figure 8** Boxplots of the proportion of individuals correctly genotyped (y-axis) for updog (black), fitPoly (orange), GATK (blue), and the method of Li (2011) (green) vs. the true allele frequency (x-axis). 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).

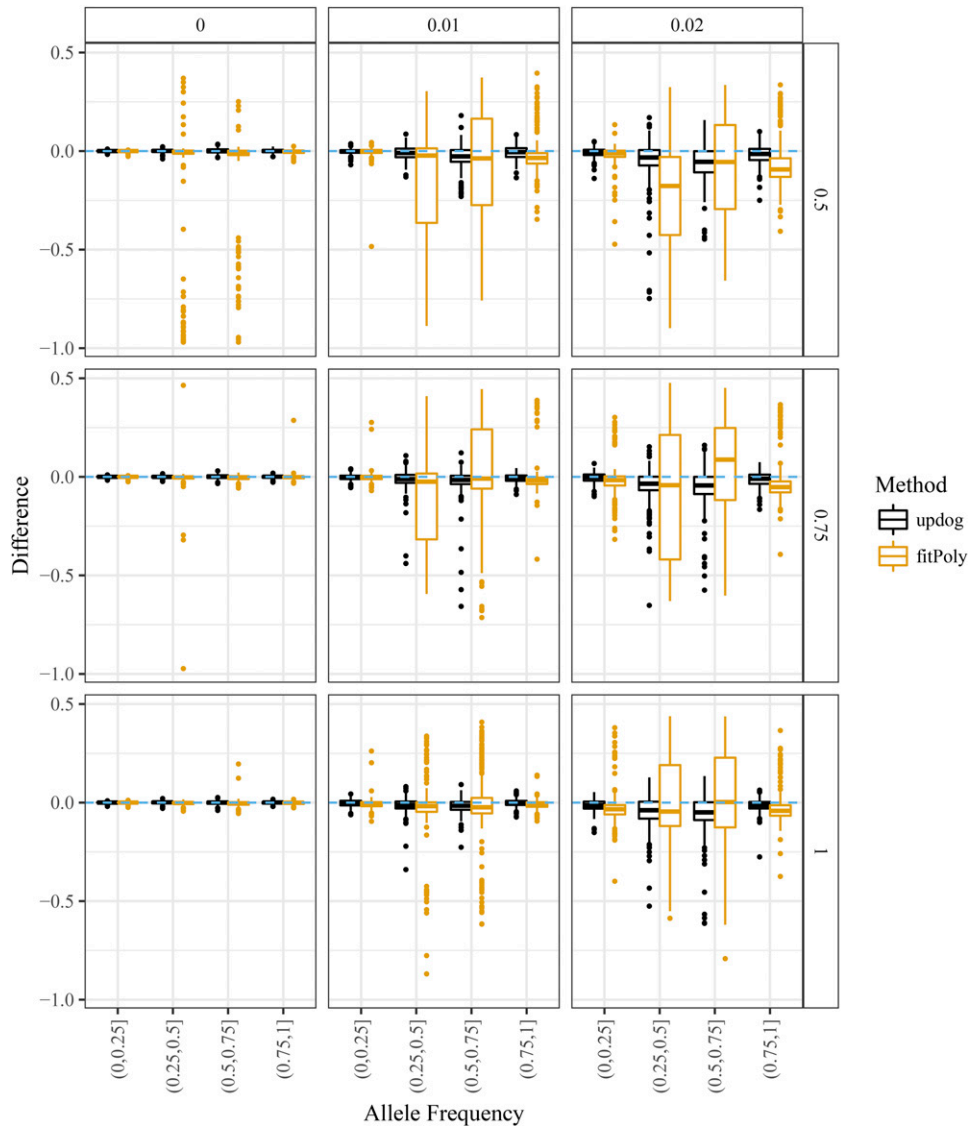
5. Even when there is no bias, updog performs as well as GATK and the method of Li (2011), except for some datasets where the allele frequency is close to 0.5 and the overdispersion is large. This is because GATK and the method of Li (2011) both effectively assume the bias parameter  $h$  is known at 1, while updog is using a degree of freedom to estimate the bias parameter. In the case where there is actually no bias, this can make updog's estimates somewhat more unstable.

Even though the genotyping results are not encouraging for large amounts of overdispersion, updog has some ability to estimate this level of overdispersion to provide information to researchers on the quality of a SNP (Figure 7).

We compared the allele frequency estimates from updog and the method of Li (2011). Here, the results are more encouraging (Figure S2). Even for large levels of overdispersion and bias, updog accurately estimates the allele frequency (though less accurately than with small overdispersion). As

expected, the method of Li (2011) is adversely affected by bias, which it does not model, and tends to overestimate the allele frequency in the direction of the bias.

Also encouraging are the resulting correlations between the true and estimated genotypes. Figure S3 contains boxplots of the correlation of the true genotypes with the posterior means of each method—except for the method of Li (2011), where we calculate the correlation with the posterior mode as the software of Blischak *et al.* (2018) does not return posterior means. We used the posterior means because the posterior mode minimizes 0/1 loss, and so would be suboptimal for other, continuous metrics. We also included the naive method of returning the proportion of reference reads as the estimated genotype (Grandke *et al.* 2016). From Figure S3, we see that updog outperforms all other methods in every scenario we explored. We see the largest gains over the naive method when there is no overdispersion and some bias, while the largest gains over fitPoly occur in high bias and high overdispersion scenarios. These results indicate that using



**Figure 9** Difference between the estimated and true proportion of individuals genotyped incorrectly (y-axis) stratified by bins of the allele frequency (x-axis). Values above zero (the dashed horizontal line) correspond to overestimating the proportion of individuals genotyped incorrectly. 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).

updog's posterior means could improve the results of association studies.

Finally, we compared the ability of updog and fitPoly to estimate the proportion of individuals genotyped incorrectly. Accurate estimates of this quantity can aid researchers when filtering for high-quality SNPs (*Screening SNPs*). The software implementation of updog returns this quantity by default, while we derived this quantity from fitPoly using the output of the posterior probabilities on the genotypes. We subtracted the true proportion of individuals genotyped incorrectly from the estimated proportion and provide boxplots of these quantities in Figure 9. Generally, updog provides more accurate estimates for this proportion than fitPoly, particularly for small levels of overdispersion and allele frequencies close to 0.5.

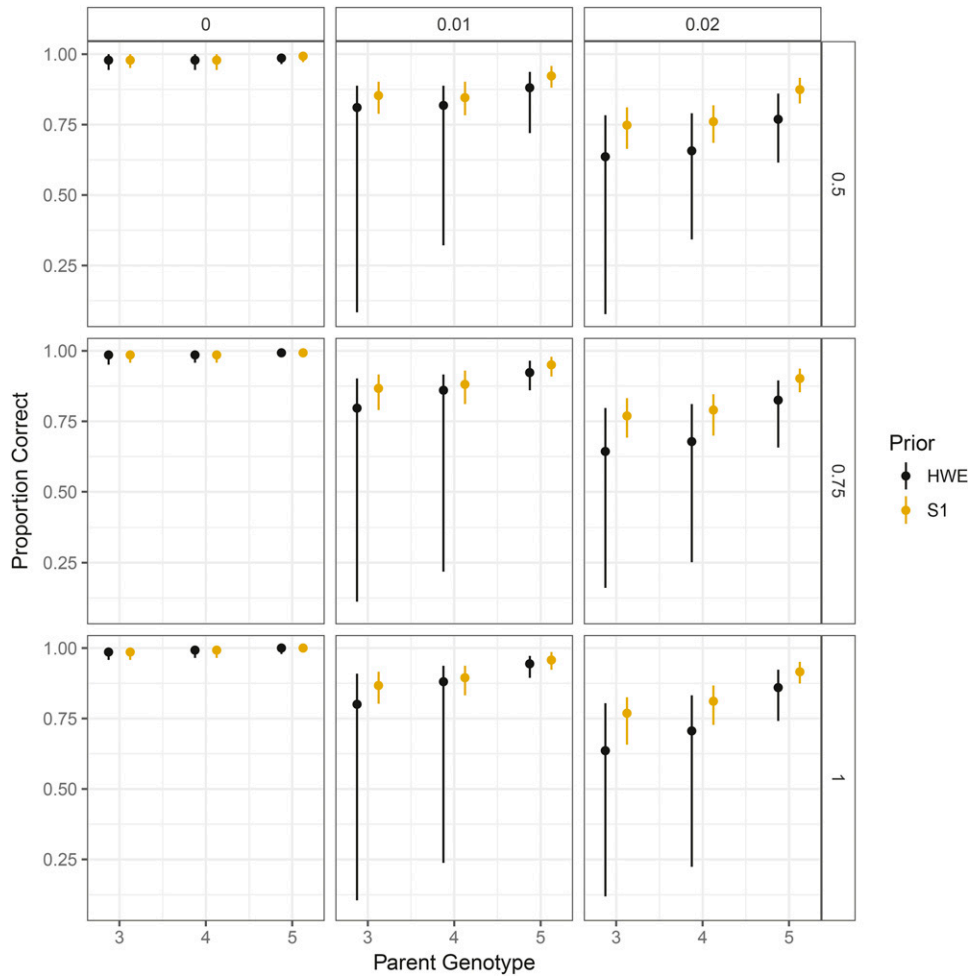
#### **Simulation comparing use of S1 prior and HWE prior**

We ran simulations to evaluate the gains in sharing information between siblings in an S1 population. We drew the

genotypes of an S1 population of 142 individuals where the parent contains either three, four, or five copies of the reference allele. We then simulated these individuals' read-counts under the updog model using the same parameter settings as in *Simulation comparisons to other methods*:  $\epsilon = 0.005$ ,  $\tau \in \{0, 0.01, 0.02\}$ , and  $h \in \{0.5, 0.75, 1\}$  (with  $h = 1$  indicating no bias). For each combination of parental genotype, overdispersion, and allelic bias, we simulated 1000 datasets. The  $n_i$  values were again obtained from the 1000 SNPs in the Shirasawa *et al.* (2017) dataset with the largest read-counts.

For each dataset, we fit updog using a prior that either assumes the individuals were from an S1 population (*Model*) or were in HWE (*Extension to population studies*). We plot summaries of the proportion of individuals genotyped correctly across datasets on Figure 10. In every combination of parameters, correctly assuming an S1 prior improves performance, particularly when there is a small amount of overdispersion ( $\tau = 0.01$ ). Boxplots of the proportion of individuals genotyped correctly may be found in Figure S7.





**Figure 10** The 0.05, 0.5, and 0.95 quantiles (across datasets) 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 intervals correctly assume the individuals are from an S1 population while black intervals incorrectly assume the population is in HWE.

### Simulations on preferential pairing

In *Accounting for preferential pairing* we developed a procedure to account for arbitrary levels of preferential pairing. In this section, we evaluate the accuracy of updog in the presence of preferential pairing.

We consider a tetraploid S1 population of individuals. In such a population, preferential pairing only affects the segregation probabilities when the parent has two copies of the reference allele (because this is the only scenario in a tetraploid population where there is more than one pairing configuration—see Theorem S1 in File S1). In this setting, the two possible pairing configurations are  $\mathbf{m} = (1, 0, 1)$  and  $\mathbf{m} = (0, 2, 0)$ . Under a nonpreferential pairing setting, the weights of each configuration would be  $\gamma_1 = 1/3$  for the first configuration and  $\gamma_2 = 2/3$  for the second (Theorem S5 in File S1).

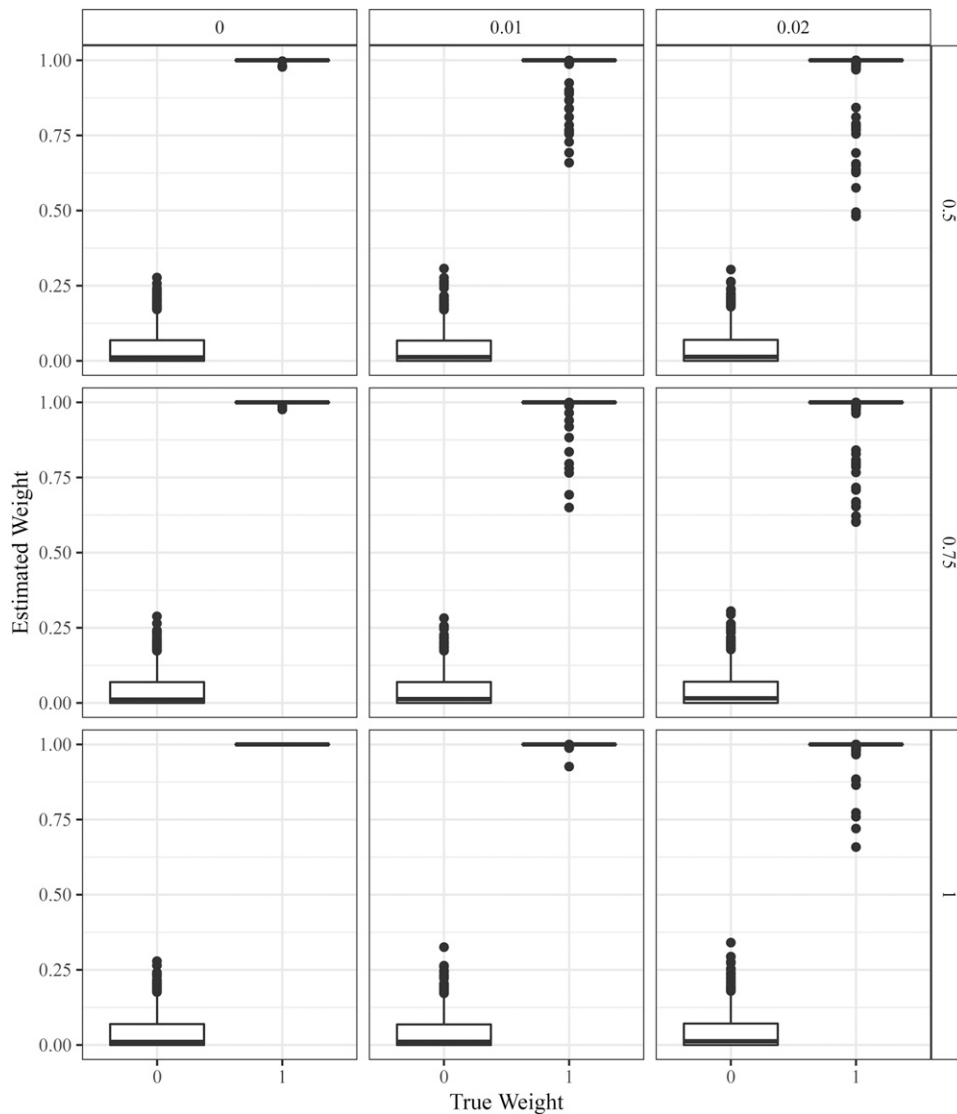
To explore the performance of updog in the case of the most extreme levels of preferential pairing possible in a tetraploid species, we simulated offspring genotypes by setting either  $\gamma_1 = 0$  or  $\gamma_1 = 1$ . We did this while setting  $\epsilon = 0.005$ ,  $\tau \in \{0, 0.01, 0.02\}$ , and  $h \in \{0.5, 0.75, 1\}$ . Figure 11 contains boxplots of the estimated levels of  $\gamma_1$  against the true values of  $\gamma_1$ . updog can estimate  $\gamma_1$  reasonably accurately,

though these estimates become less stable as the overdispersion increases.

We compared the preferential-pairing version of updog to fitPoly and the nonpreferential pairing version of updog. Figure 12 contains boxplots of the proportion of individuals genotyped correctly stratified by  $\gamma_1$ . The preferential pairing version of updog performs better than fitPoly and the nonpreferential pairing version of updog, particularly in the presence of large amounts of overdispersion, and when  $\gamma_1$  is large.

### Sweet potato

We fit the Balding-Nichols version of the method of Blischak *et al.* (2018), the method of Serang *et al.* (2012) (implemented in the SuperMASSA software at <http://statgen.esalq.usp.br/SuperMASSA/>), fitPoly (Voorrips *et al.* 2011), and updog to the SNPs presented in Figure 3. The Balding-Nichols model that Blischak *et al.* (2018) uses has parameters that are shared across loci to estimate the genotype distributions of the SNPs, so we actually fit the method of Blischak *et al.* (2018) on the 1000 SNPs with the most read-counts and just present the results for the three SNPs from Figure 3. As the method of Blischak *et al.* (2018) requires the sequencing



**Figure 11** Boxplots of the estimated mixing proportion of one of the two possible pairing configurations (y-axis) stratified by the actual mixing proportion (x-axis) in a tetraploid S1 population where the parent has two copies of the reference allele. The row facets index the allelic bias (where 1 indicates no bias), and the column facets index the overdispersion.

error rate to be known, we use the sequencing error rate estimates provided by updog.

FitPoly allows for F1, but not S1, populations, and seemingly only when parental counts are provided. To account for this, we copied the read-counts data from the parent and specified this duplicated point as a second parent. As recommended by the authors' documentation, we fit these data using the `try.HW = TRUE` setting.

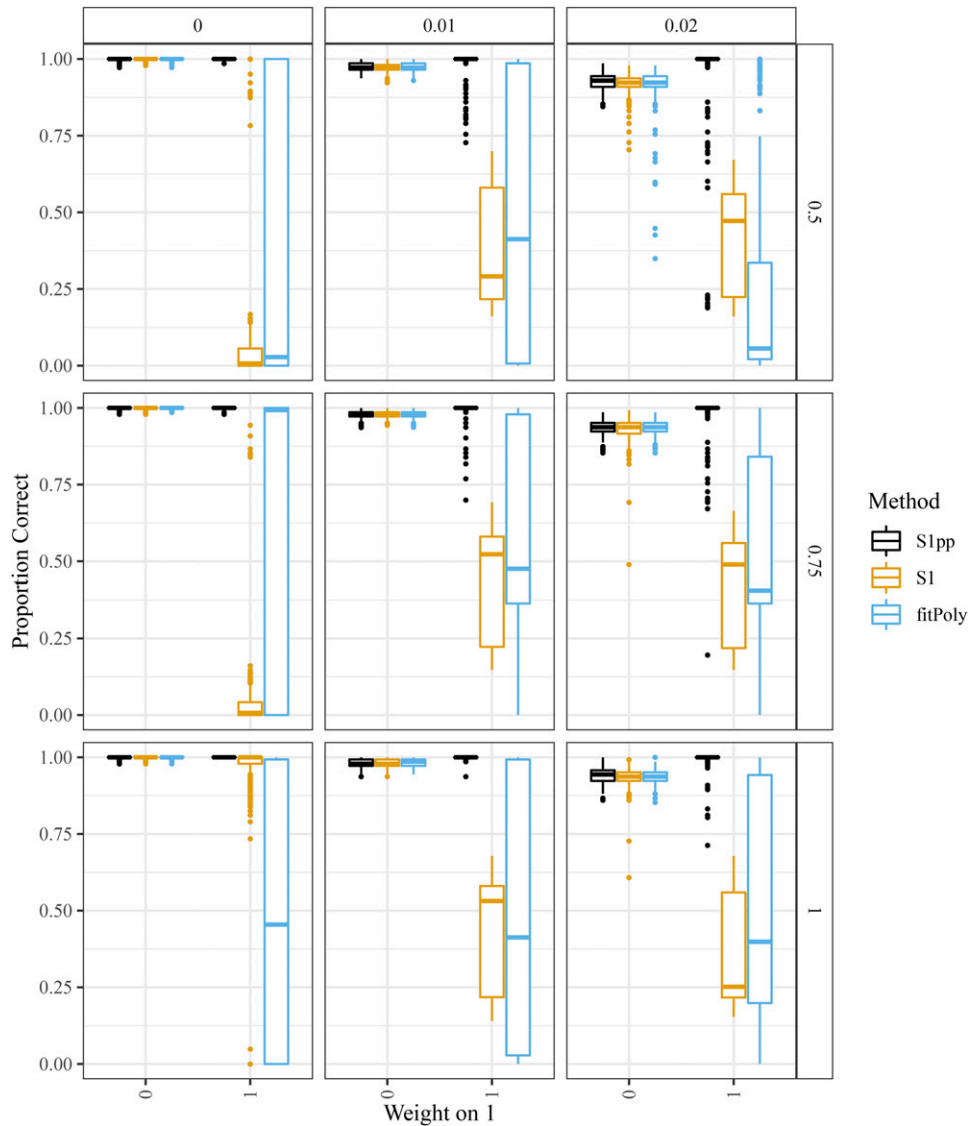
The fits for all four methods can be seen in Figure 13. Our conclusions are:

1. Serang *et al.* (2012) and Blischak *et al.* (2018) provide unintuitive results for SNP2. As we know this is an S1 population, the genotype distribution should be closer to a 1:2:1 ratio for genotypes AAAAAA, AAAAAa, and AAAAaa. Both updog and fitPoly correctly accounts for this. Updog does so by estimating an extreme bias while fitPoly does so by allowing the mean proportion of reference counts for each genotype to depend only linearly or quadratically on the dosage level.

2. The method of Serang *et al.* (2012) was designed for Gaussian, not count, data and, as such, provides some unintuitive genotyping. In particular, for SNPs 1 and 2 we see a few points on the left end of the AAAAAA genotypes that are coded AAAAAA when these could probably only result from an extreme sequencing error rate given that genotype.
3. The point in SNP3 that we have described as an outlier is removed by updog. The other methods are not able to cope with outliers.

Generally, fitPoly returns more similar genotyping estimates to updog on these data than other methods. However, their posterior measures of uncertainty differ, and sometimes substantially. Figure S9 plots the estimates for the proportion of individuals genotyped incorrectly by both updog and fitPoly. Updog usually returns larger levels of uncertainty, though there are some datasets where fitPoly estimates that a much larger proportion of individuals are genotyped incorrectly than updog. To us, these seem like





**Figure 12** Boxplots of the proportion of individuals genotyped correctly (y-axis) stratified by the weight on one of the two possible pairing configurations (x-axis) in a tetraploid S1 population where the parent has two copies of the reference allele. The row facets index the allelic bias (where 1 indicates no bias), and the column facets index the overdispersion. Boxplots are color-coded by method: an updog fit that accounts for preferential pairing (black), an updog fit that does not account for preferential pairing (orange), and fitPoly (blue).

anomalous results by fitPoly, as these SNPs seem mostly well-behaved (Figure S10).

### Computation time

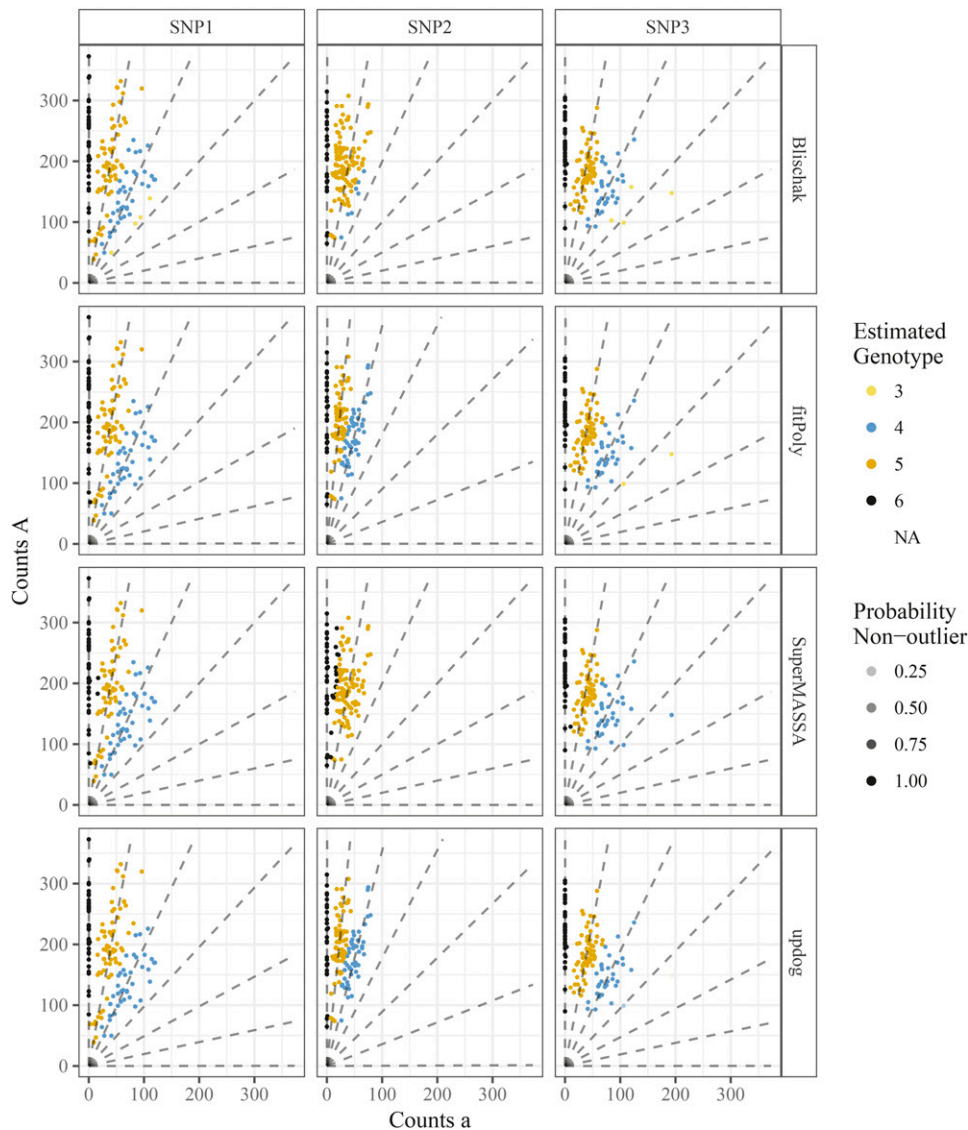
We measured the computation time required to fit updog to the 1000 SNPs with the highest read depth from the dataset of Shirasawa *et al.* (2017). These computations were run on a 4.0 GHz quad-core PC running Linux with 32 GB of memory. It took on average 3.3 sec per SNP to fit updog, with 95% of the runs taking between 1.81 and 6.56 sec. This is much larger than the total time of 121 sec it took for the software of Blischak *et al.* (2018) to fit their model on all 1000 SNPs. But it is almost exactly equal to the computation time of fitPoly, where 95% of the runs took between 1.77 and 6.55 sec. Also, since the SNPs in updog are fitted independently, this allows us to easily parallelize this computation over the SNPs.

### Sampling depth suggestions

In *Using oracle results for sampling depth guidelines*, we derived the oracle genotyping error rates given that the

sequencing error rate, the allelic bias, the overdispersion, and the genotype distributions are all known. In this section, we explore these functions in the particular case when a SNP is in HWE with an allele frequency of 0.8.

We found the minimum read depths needed to obtain an error rate (33)  $< 0.05$  under different levels of ploidy  $K \in \{2, 4, 6\}$ , bias  $h \in \{0.5, 0.55, 0.6, \dots, 1\}$ , and overdispersion  $\tau \in \{0, 0.002, 0.004, \dots, 0.02\}$  while fixing the sequencing error rate at  $\epsilon = 0.001$ . A heatmap of the minimum read depths required for the different ploidy levels may be found in Figures S11–S13. We draw the following conclusions: (i) For diploid individuals, the levels of bias and overdispersion matter relatively little in terms of genotyping accuracy, and accurate genotyping may be made with an extremely small read depth. (ii) Read depth requirements seem to grow exponentially with the ploidy. (iii) The levels of bias and overdispersion greatly impact the read depth requirements in higher ploidy species. For reasonable levels of bias and overdispersion (Figure S8), the read depth requirements can be



**Figure 13** Genotype plots as in Figure 3 but color-coded by the estimated genotypes from the method of Blischak *et al.* (2018), method of Serang *et al.* (2012), and fitPoly (Voorrips *et al.* 2011), or updog (row facets). The column facets distinguish between the different SNPs. Transparency is proportional to the probability of being an outlier (only for updog).

many times larger to obtain the same error rates as when there is no bias and no overdispersion.

Though the read depth requirements to obtain a small error rate seem depressing under reasonable levels of bias and overdispersion, the results are much more optimistic to obtain high correlation with the true genotypes. In Figures S14–S16, we calculated the minimum read depths required to obtain a correlation with the true genotypes of  $>0.9$  under the same parameter settings as Figures S11–S13. There, we see that the read depth requirements are much more reasonable, where, for hexaploid species, a read depth of 90 is adequate to obtain this correlation under large levels of bias and overdispersion. Correspondingly, for tetraploids a read depth of 25 seems adequate to obtain high correlation under large levels of bias and overdispersion.

Once again, we would like to stress that these are *lower bounds* on the read depth requirements. Actual read depths should be slightly larger than these, based on how many individuals are in the study.

## Discussion

We have developed an empirical Bayes genotyping procedure that takes into account common aspects of NGS data: sequencing error, allelic bias, overdispersion, and outlying points. We have shown that accounting for allelic bias is vital for accurate genotyping, and that that our posterior measures of uncertainty (automatically taking into account the estimated levels of overdispersion) are well calibrated and may be used as quality metrics for SNPs/individuals. We confirmed the validity of our method on simulated and real data.

We have focused on a dataset that has a relatively large read-coverage and contains a large amount of known structure (via Mendelian segregation). In many datasets, one would expect to have much lower coverage of SNPs and less structure (Blischak *et al.* 2018). For such data, we do not expect the problems of allelic bias, overdispersion, and outlying points to disappear. From our simulations, the most insidious of these issues to ignore is the allelic bias. If a

reference genome is available, then it might be possible to correct for the read-mapping bias by using the methods from van de Geijn *et al.* (2015). However, this might not be the total cause of allelic bias. And, without a reference genome, it is important to model this bias directly. We do not expect small coverage SNPs to contain enough information to accurately estimate the bias using our methods. For such SNPs, more work is needed. It might be possible to borrow strength between SNPs to develop accurate genotyping methods.

We have assumed that the ploidy is known and constant between individuals. However, some species (*e.g.*, sugarcane) can have different ploidies per individual (Garcia *et al.* 2013). If one has access to good cytological information on the ploidy of each individual, it would not be conceptually difficult to modify updog to allow for different (and known) ploidies of the individuals. However, estimating the ploidy might be more difficult, particularly in the presence of allelic bias. In the presence of such bias, one can imagine that it would be difficult to discern if a sample's location on a genotype plot was due to bias or due to a higher or lower ploidy level. More work would be needed to develop an approach that works with individuals having unknown ploidy levels. Serang *et al.* (2012) attempts to estimate the ploidy level in Gaussian data, but they do not jointly account for allelic bias, which we hypothesize would bias their genotyping results.

Garrison and Marth (2012) uses a multinomial likelihood to model multiallelic haplotypes. This allows for more complex genotyping beyond SNPs. The models presented here could be easily extended to a multinomial likelihood. For example, to model bias with  $k$  possible alleles, one could introduce  $k - 1$  bias parameters which measure the bias of each allele relative to a reference allele. One could use a Dirichlet-multinomial distribution to model overdispersion and a uniform-multinomial distribution (uniform on the standard  $k - 1$  simplex) to model outliers. Modeling allele-detection errors (corresponding to sequencing errors in our NGS setup) would be context-specific.

The methods developed here may also be useful for genotyping diploid individuals. We are not aware of diploid genotyping methods that account for allelic bias, overdispersion, and outlying points. However, considering our oracle results in *Sampling depth suggestions*, accounting for these features is probably more important for polyploid species, as determining allele dosage is more difficult than determining heterozygosity/homozygosity.

Genotyping is typically only one part of a large analysis pipeline. It is known that genotyping errors can inflate genetic maps (Hackett and Broadfoot 2003), but it remains to determine the impact of dosage estimates on other downstream analyses. In principle, one would want to integrate out uncertainty in estimated dosages, but, in diploid analyses, it is much more common to ignore uncertainty in genotypes and simply use the posterior mean genotype—for example, in GWAS analyses (Guan and Stephens 2008). It may be that similar ideas could work well for GWAS in polyploid species (see also Grandke *et al.* 2016).

## Acknowledgments

We sincerely thank the authors of Shirasawa *et al.* (2017) for providing their data and Paul Blischak for providing useful comments. D.G. and M.S. were supported by National Institutes of Health (NIH) grant HG002585 and by a grant from the Gordon and Betty Moore Foundation (Grant GBMF #4559). L.F.V.F. and A.A.F.G. were partially supported by grant 2014/20389-2, FAPESP/CAPES (São Paulo Research Foundation). A.A.F.G. was supported by a productivity scholarship from the National Council for Scientific and Technological Development (CNPq).

## Literature Cited

- Aitkin, M., and G. T. Wilson, 1980 Mixture models, outliers, and the EM algorithm. *Technometrics* 22: 325–331. <https://doi.org/10.1080/00401706.1980.10486163>
- Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver *et al.*, 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3: e3376. <https://doi.org/10.1371/journal.pone.0003376>
- Balding, D. J., and R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity, pp. 3–12 in *Human Identification: The Use of DNA Markers*, edited by B. S. Weir. Springer, Berlin. [https://doi.org/10.1007/978-0-306-46851-3\\_2](https://doi.org/10.1007/978-0-306-46851-3_2)
- Balding, D. J., and R. A. Nichols, 1997 Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity* 78: 583–589. <https://doi.org/10.1038/hdy.1997.97>
- Bargary, N., J. Hinde, and A. A. F. Garcia, 2014 Finite mixture model clustering of SNP data, pp. 139–157 in *Statistical Modeling in Biostatistics and Bioinformatics: Selected Papers*, edited by G. MacKenzie, and D. Peng. Springer International Publishing, New York. [https://doi.org/10.1007/978-3-319-04579-5\\_11](https://doi.org/10.1007/978-3-319-04579-5_11)
- Blischak, P. D., L. S. Kubatko, and A. D. Wolfe, 2016 Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Mol. Ecol. Resour.* 16: 742–754. <https://doi.org/10.1111/1755-0998.12493>
- Blischak, P. D., L. S. Kubatko, and A. D. Wolfe, 2018 SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* 34: 407–415. <https://doi.org/10.1093/bioinformatics/btx587>
- Bomblies, K., G. Jones, C. Franklin, D. Zickler, and N. Kleckner, 2016 The challenge of evolving stable polyploidy: could an increase in “crossover interference distance” play a central role? *Chromosoma* 125: 287–300. <https://doi.org/10.1007/s00412-015-0571-4>
- Bourke, P. M., P. Arens, R. E. Voorrips, G. D. Esselink, C. F. S. Koning-Boucoiran *et al.*, 2017 Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *Plant J.* 90: 330–343. <https://doi.org/10.1111/tpj.13496>
- Bourke, P. M., R. E. Voorrips, R. G. F. Visser, and C. Maliepaard, 2018 Tools for genetic studies in experimental populations of polyploids. *Front. Plant Sci.* 9: 513. <https://doi.org/10.3389/fpls.2018.00513>
- Byrne, S., A. Czaban, B. Studer, F. Panitz, C. Bendixen *et al.*, 2013 Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PLoS One* 8: e57438. <https://doi.org/10.1371/journal.pone.0057438>
- Chen, N., C. V. Van Hout, S. Gottipati, and A. G. Clark, 2014 Using Mendelian inheritance to improve high-throughput SNP discovery. *Genetics* 198: 847–857. <https://doi.org/10.1534/genetics.114.169052>



- Clark, L. V., A. E. Lipka, and E. J. Sacks, 2018 polyRAD: genotype calling with uncertainty from sequencing data in polyploids and diploids. *bioRxiv* 380899. <https://doi.org/10.1101/380899>.
- Crowder, M. J., 1979 Inference about the intraclass correlation coefficient in the beta-binomial ANOVA for proportions. *J. R. Stat. Soc. B* 41: 230–234.
- Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12: 499–510. <https://doi.org/10.1038/nrg3012>
- Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Garcia, A. A., M. Mollinari, T. G. Marconi, O. R. Serang, R. R. Silva *et al.*, 2013 SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Sci. Rep.* 3: 3399. <https://doi.org/10.1038/srep03399>
- Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907v2 [q-bio.GN]*.
- Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9: e90346. <https://doi.org/10.1371/journal.pone.0090346>
- Goodwin, S., J. D. McPherson, and W. R. McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17: 333–351. <https://doi.org/10.1038/nrg.2016.49>
- Grandke, F., P. Singh, H. C. M. Heuven, J. R. de Haan, and D. Metzler, 2016 Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: a comparative study in hexaploid chrysanthemum. *BMC Genomics* 17: 672. <https://doi.org/10.1186/s12864-016-2926-5>
- Guan, Y., and M. Stephens, 2008 Practical issues in imputation-based association mapping. *PLoS Genet.* 4: e1000279. <https://doi.org/10.1371/journal.pgen.1000279>
- Hackett, C., and L. Broadfoot, 2003 Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* 90: 33–38. <https://doi.org/10.1038/sj.hdy.6800173>
- Hadi, A. S., and J. S. Simonoff, 1993 Procedures for the identification of multiple outliers in linear models. *J. Am. Stat. Assoc.* 88: 1264–1272. <https://doi.org/10.1080/01621459.1993.10476407>
- Hastie, T., R. Tibshirani, and J. Friedman, 2009 *The Elements of Statistical Learning*. Springer-Verlag, New York. <https://doi.org/10.1007/978-0-387-84858-7>
- Huber, P. J., 1964 Robust estimation of a location parameter. *Ann. Math. Stat.* 35: 73–101. <https://doi.org/10.1214/aoms/1177703732>
- Kim, C., H. Guo, W. Kong, R. Chandnani, L.-S. Shuang *et al.*, 2016 Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci.* 242: 14–22. <https://doi.org/10.1016/j.plantsci.2015.04.016>
- Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
- Li, X., Y. Wei, A. Acharya, Q. Jiang, J. Kang *et al.*, 2014 A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. *G3 (Bethesda)* 4: 1971–1979. <https://doi.org/10.1534/g3.114.012245>
- Li, Y., C. Sidore, H. M. Kang, M. Boehnke, and G. R. Abecasis, 2011 Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 21: 940–951. <https://doi.org/10.1101/gr.117259.110>
- Liu, H., X. Luo, L. Niu, Y. Xiao, L. Chen *et al.*, 2017 Distant eQTLs and non-coding sequences play critical roles in regulating gene expression and quantitative trait variation in maize. *Mol. Plant* 10: 414–426. <https://doi.org/10.1016/j.molp.2016.06.016>
- Lu, F., A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney *et al.*, 2013 Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 9: e1003215. <https://doi.org/10.1371/journal.pgen.1003215>
- Maruki, T., and M. Lynch, 2017 Genotype calling from population-genomic sequencing data. *G3 (Bethesda)* 7: 1393–1404. <https://doi.org/10.1534/g3.117.039008>
- McCallum, S., J. Graham, L. Jorgensen, L. J. Rowland, N. V. Bassil *et al.*, 2016 Construction of a SNP and SSR linkage map in autotetraploid blueberry using genotyping by sequencing. *Mol. Breed.* 36: 41. <https://doi.org/10.1007/s11032-016-0443-5>
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Mollinari, M., and O. Serang, 2015 Quantitative SNP genotyping of polyploids with MassARRAY and other platforms, pp. 215–241 in *Plant Genotyping: Methods and Protocols*, edited by J. Batley. Springer, New York. [https://doi.org/10.1007/978-1-4939-1966-6\\_17](https://doi.org/10.1007/978-1-4939-1966-6_17)
- Motazed, E., D. de Ridder, R. Finkers, S. Baldwin, S. Thomson *et al.*, 2018 TriPoly: haplotype estimation for polyploids using sequencing data of related individuals. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bty442>
- Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12: 443–451. <https://doi.org/10.1038/nrg2986>
- Otto, S. P., and J. Whitton, 2000 Polyploid incidence and evolution. *Annu. Rev. Genet.* 34: 401–437. <https://doi.org/10.1146/annurev.genet.34.1.401>
- Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, 2012 Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7: e37135. <https://doi.org/10.1371/journal.pone.0037135>
- Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69: 1–14. <https://doi.org/10.1086/321275>
- R Core Team, 2017 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- Schilling, M. P., P. G. Wolf, A. M. Duffy, H. S. Rai, C. A. Rowe *et al.*, 2014 Genotyping-by-sequencing for populus population genomics: an assessment of genome sampling patterns and filtering approaches. *PLoS One* 9: e95292. <https://doi.org/10.1371/journal.pone.0095292>
- Schmitz Carley, C. A., J. J. Coombs, D. S. Douches, P. C. Bethke, J. P. Palta *et al.*, 2017 Automated tetraploid genotype calling by hierarchical clustering. *Theor. Appl. Genet.* 130: 717–726. <https://doi.org/10.1007/s00122-016-2845-5>
- Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS One* 7: e30906. <https://doi.org/10.1371/journal.pone.0030906>
- 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*). *Sci. Rep.* 7: 44207.
- Skellam, J. G., 1948 A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. R. Stat. Soc. B* 10: 257–261.
- Soltis, D. E., C. J. Visger, and P. S. Soltis, 2014 The polyploidy revolution then...and now: Stebbins revisited. *Am. J. Bot.* 101: 1057–1078. <https://doi.org/10.3732/ajb.1400178>
- Soltis, P. S., and D. E. Soltis, 2000 The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl. Acad.*

- Sci. USA 97: 7051–7057. <https://doi.org/10.1073/pnas.97.13.7051>
- Spindel, J., M. Wright, C. Chen, J. Cobb, J. Gage *et al.*, 2013 Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor. Appl. Genet.* 126: 2699–2716 [corrigenda: *Theor. Appl. Genet.* 129: 201–202 (2016)]. <https://doi.org/10.1007/s00122-013-2166-x>
- Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard *et al.*, 2015 Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11: 1–25. <https://doi.org/10.1371/journal.pgen.1004982>
- Stebbins, G. L., 1947 Types of polyploids: their classification and significance, pp. 403–429 in *Advances in Genetics*, Vol. 1, edited by M. Demerec. Academic Press, New York. [https://doi.org/10.1016/S0065-2660\(08\)60490-3](https://doi.org/10.1016/S0065-2660(08)60490-3)
- Stift, M., C. Berenos, P. Kuperus, and P. H. van Tienderen, 2008 Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *rorippa* (yellow cress) microsatellite data. *Genetics* 179: 2113–2123. <https://doi.org/10.1534/genetics.107.085027>
- Stift, M., R. Reeve, and P. H. Van Tienderen, 2010 Inheritance in tetraploid yeast revisited: segregation patterns and statistical power under different inheritance models. *J. Evol. Biol.* 23: 1570–1578. <https://doi.org/10.1111/j.1420-9101.2010.02012.x>
- Tennessen, J. A., R. Govindarajulu, T.-L. Ashman, and A. Liston, 2014 Evolutionary origins and dynamics of octoploid strawberry sub-genomes revealed by dense targeted capture linkage maps. *Genome Biol. Evol.* 6: 3295–3313. <https://doi.org/10.1093/gbe/evu261>
- Udall, J. A., and J. F. Wendel, 2006 Polyploidy and crop improvement. *Crop Sci.* 46: S3–S14. <https://doi.org/10.2135/cropsci2006.07.0489tpg>
- van de Geijn, B., G. McVicker, Y. Gilad, and J. K. Pritchard, 2015 WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* 12: 1061–1063. <https://doi.org/10.1038/nmeth.3582>
- Voorrips, R. E., and C. A. Maliepaard, 2012 The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics* 13: 248. <https://doi.org/10.1186/1471-2105-13-248>
- Voorrips, R. E., G. Gort, and B. Vosman, 2011 Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics* 12: 172. <https://doi.org/10.1186/1471-2105-12-172>
- Zhou, B., and A. S. Whittemore, 2012 Improving sequence-based genotype calls with linkage disequilibrium and pedigree information. *Ann. Appl. Stat.* 6: 457–475. <https://doi.org/10.1214/11-AOAS527>

Communicating editor: G. Churchill