Sequence coverage required for accurate genotyping by sequencing in polyploid species

Lin Wang | Jixuan Yang | Hong Zhang | Qin Tao | Yuxin Zhang | Zhenyu Dang | Fengjun Zhang | Zewei Luo

Abstract

Polyploidy plays an important role in the evolution of eukaryotes, especially for flowering plants. Many of ecologically or agronomically important plant or crop species are polyploids, including sycamore maple (tetraploid), the world second and third largest food crops wheat (hexaploid) and potato (tetraploid) as well as economically important aquaculture animals such as Atlantic salmon and trout. The next generation sequencing data enables to allocate genotype at a sequence variant site, known as genotyping by sequencing (GBS). GBS has stimulated enormous interests in population based genomics studies in almost all diploid and many polyploid organisms. DNA sequence polymorphisms are codominant and thus fully informative about the underlying genotype at the polymorphic site, making GBS a straightforward task in diploids. However, sequence data may usually be uninformative in polyploid species, making GBS a far more challenging task in polyploids. This paper presents novel and rigorous statistical methods for predicting the number of sequence reads needed to ensure accurate GBS at a polymorphic site bared by the reads in polyploids and shows that a dozen of reads can ensure a probability of 95% to recover all constituent alleles of any tetraploid genotype but several hundreds of reads are needed to accurately uncover the genotype with probability confidence of 90%, subverting the proposition of GBS using low coverage sequence data in the literature. The theoretical prediction was tested by use of RAD-seq data from tetraploid potato cultivars. The paper provides polyploid experimentalists with theoretical guides and methods for designing and conducting their sequence-based studies.

KEYWORDS

genotyping by sequencing, polyploids, RAD-seq data, sequence coverage, Solanum tuberosum L
1 | INTRODUCTION

Advancement in next-generation sequencing technology (NGS) allows identification of DNA sequence variants at the genome-wide scale and at a very competitive cost in comparison to traditional DNA molecular marker development techniques such as restriction fragment length polymorphisms (RFLPs, Botstein et al., 1980), amplified fragment length polymorphisms (AFLPs, Vos et al., 1995) and Microsatellites (Jarne & Lagoda, 1996). The new sequencing technique has been fast evolved for the discovery, sequencing and genotyping of sequence variant markers at a genome-wide scale, at a population level and at a much competitive cost (Davey et al., 2011). The digital nature of sequence based marker data enables genotyping by sequencing (GBS) at a much competitive cost (Davey et al., 2011). The digital nature of sequence based marker data enables genotyping by sequencing (GBS) at the marker loci, which has motivated enormous interests in population and quantitative genetic analyses in diploid species (Nielsen et al., 2011; Narum et al., 2013; van de Geijn et al., 2015).

There may be three possible genotypes, that is, two homozygotes and one heterozygote, at any polymorphic locus in diploids. Diagnosing an individual genotype at any diploid sequence variant site may be done directly from surveying whether one or two alleles are present at the variant site. Thus, GBS in diploids is a straightforward task. However, it is much more challenging to carry out GBS in polyploid species because a much larger number of genotypes would be possible for a polyploid individual at a polymorphic locus with a given number of polymorphic alleles. Taking tetraploids as an example, there would be three possible heterozygote genotypes at a bi- or triallelic locus. Thus, sequence data from polyploids may be usually uninformative to recover the genotype at sequence variant sites with the certainty that in diploids as explained above.

Two fundamental questions are raised on designing GBS experiments and the following analysis in polyploids in order for accurate genotype prediction from the sequence data. First, for a polyploid individual, what number of sequence reads is needed to ensure the presence of the marker alleles at a sequence variant site with a given probability confidence? Second, what does sequence coverage at a polymorphic site mean? Two key questions is obviously critical for any genetic or genomic studies in polyploids. This study reports statistically rigorous methods to answer these two questions in tetraploid species as an example and the analyses presented here may be extended to any other polyploids without substantial technical obstacles. We explored the statistical prediction using the sequence data from restriction-site associated DNA sequencing (RAD-seq, Baird et al., 2008) experiment with tetraploid cultivars of potato (Solanum tuberosum L.) and discussed limitations and challenges of GBS in tetraploids in light of the theoretical prediction and complexities of sequence data in quantitative genetic analyses.

2 | MATERIALS AND METHODS

2.1 | Design parameters of GBS experiments in tetraploids

Although analysis presented below is generic for polyploids at any ploidy level, we focus here on tetraploids as an example. To design a GBS experiment in tetraploid species, one needs to address two key questions. First, what sequence coverage at a polymorphic site is needed to ensure all k constituent alleles of a tetraploid genotype at the site will be present in the sequence data for a given probability confidence P? Second, what sequence coverage is needed to ensure a tetraploid genotype to be accurately predicted with a given probability confidence interval P ± δ.

To answer these two questions, we assume that generation of sequence data from a tetraploid genotype can be assimilated as a sampling process of multinomial distribution. We consider a tetraploid individual genotype at a polymorphic site as \( g_i = (k_A, k_C, k_G, k_T) \), where \( k_x \) is the number of nucleotide \( x \) being the constituent alleles of a tetraploid genotype at the site. Data of the sequence reads covering the polymorphic site from a tetraploid individual is denoted by \( n = \{n_A, n_C, n_G, n_T\} \) with \( n_i \) being the number of sequence reads carrying nucleotide \( x = A, C, G \) or \( T \) at the polymorphic site. We assume that \( n = \{n_A, n_C, n_G, n_T\} \) follows a multinomial distribution as

\[
Pr(n_A, n_C, n_G, n_T \mid k_A, k_C, k_G, k_T) = \binom{n}{n_A, n_C, n_G, n_T} (k_A/4)^{n_A} (k_C/4)^{n_C} (k_G/4)^{n_G} (k_T/4)^{n_T} \tag{1}
\]

To formulate the questions aforementioned, we consider \( m \), the number of segregating alleles, possibly being equal to 2, 3 and 4, respectively. It needs to be noted that \( n \) is the number of sequence reads of a tetraploid individual.

For question 1, the smallest number of \( n \) sequence reads, which is required to all \( k \) constituent alleles of a tetraploid genotype at the site to be present in the sequence data for a given confidence \( P \), will satisfy

\[
\min_n \left\{ P \leq \sum_{n_i=1}^{n} Pr(n_A, n_C, n_G, n_T \mid g_i = (k_A, k_C, k_G, k_T)) \right\} \tag{2}
\]

Equation (2) is a generic form and needs to be specified for a given \( m \), the number of alleles at the polymorphic site.

When \( m = 2 \),

\[
\min_n \left\{ P \leq \sum_{n_i=1}^{n} Pr(n_A, n_T \mid g_i = (k_A, k_T)) \right\} = \sum_{n_i=1}^{n} \binom{n}{n_A} (k_A/4)^{n_A} (1-k_T/4)^{n_T-n_A} \tag{3}
\]
where \( X, Y = A, C, G \) or \( T( X, Y ) \), \( k_x = 1, 2 \) or 3.

When \( m = 3 \),

\[
\min_n P \leq \sum_{n=0}^{n} \Pr \left( \left( n_x, n_y, n_z | r_x, r_y, r_z \right) = \left( k_x, k_y, k_z \right) \right) = \sum_{n=0}^{n} \sum_{n=0}^{n} \sum_{n=0}^{n} \left( \begin{array}{l} n \n_x \n_y \n_z \end{array} \right) (k_x/4)^{n_x} (k_y/4)^{n_y} (k_z/4)^{n_z} \tag{4}\]

where \( X, Y, Z = A, C, G \) or \( T(X \neq Y = Z) \), \( k_x, k_y, k_z > 0 \) and \( k_x + k_y + k_z = 4 \).

When \( m = 4 \),

\[
\min_n P \leq \sum_{n=0}^{n} \Pr \left( \left( n_x, n_y, n_z, n_t | r_x, r_y, r_z, r_t \right) = \left( k_x, k_y, k_z, k_t \right) \right) = \sum_{n=0}^{n} \sum_{n=0}^{n} \sum_{n=0}^{n} \sum_{n=0}^{n} \left( \begin{array}{l} n \n_x \n_y \n_z \n_t \end{array} \right) (k_x/4)^{n_x} (k_y/4)^{n_y} (k_z/4)^{n_z} (k_t/4)^{n_t} \tag{5}\]

For question 2, the smallest number of \( n \) sequence reads satisfies

\[
\min_n \left\{ P - \delta \leq \hat{P} \leq P + \delta \right\}, \tag{6}\]

where \( \hat{P} = n_x/n \) as formulated in Equation (1), \( P \) and \( \delta \) are prior given. Equation (6) needs to be specified for a given \( m \), the number of alleles at the polymorphic site.

When \( m = 2, P = \frac{1}{2} \) (e.g., \( g_i = (A, A, A, G) \) or \( \frac{1}{2} \) (e.g., \( g_i = (A, A, G, G) \)). Let \( \hat{P} = n_x/n \) with \( X = A, C, G \) or \( T \) and \( x = n_x - nP \)/\( \sqrt{nP(1 - P)} \) ~ \text{N}(0, 1), i.e., the standard normal distribution. Logically,

\[
Pr \left( \hat{P} \in P \pm \delta \right) \Phi \left( \pm \delta \sqrt{nP(1 - P)} \right) = 1 - 2 \Phi \left( \delta' \right) \tag{7}\]

where \( \delta' = \delta \sqrt{nP(1 - P)} \) and \( \Phi ( \cdot ) \) (or \( \Phi^{-1} ( \cdot ) \)) is the accumulative (or inverse) probability function of the standard normal distribution. From Equation (7), one can have \( \delta \sqrt{nP(1 - P)} = \Phi^{-1} (1 - q/2) \) with \( q = Pr \left( \hat{P} \in P \pm \delta \right) \) and thus,

\[
n = P(1 - P) \Phi^{-1} \left( (1 - q)/2 \right)^2 / \delta^2 \tag{8}\]

When \( m = 3 \) and without loss of generality, let \( q = Pr \left( \hat{P}_i \leq \hat{P} \right) \), \( i = 1, 2, 3 \) where \( \hat{P}_1, \hat{P}_2, \hat{P}_3 = \left( \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right), \left( \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right) \text{ or } \left( \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right) \text{ and } \left( \hat{P}_1, \hat{P}_2, \hat{P}_3 \right) = \left( n_1, n_2, n_3 \right)/n \). For \( i = 1, 2 \), let \( x_i = n_i - nP \)/\( \sqrt{nP(1 - P)} \) which follows a standard normal distribution and \( Pr \left( \hat{P}_i \in P \pm \delta \right) = Pr \left( \left| x_i \right| \leq \delta \sqrt{nP(1 - P)} \right) \) is the accumulative (or inverse) probability function of the standard normal distribution and \( Pr \left( \hat{P}_i \in P \pm \delta \right) = Pr \left( \left| x_i \right| \leq \delta \sqrt{nP(1 - P)} \right) \) is the accumulative (or inverse) probability function of the standard normal distribution. Since \( n_3 = n - n_1 - n_2 \),

\[
x_3 = \left( n_3 - nP \right) / \sqrt{nP(1 - P)} = \frac{(n_1 - nP_1 + n_2 - nP_2)}{\sqrt{nP(1 - P)}} - x_2 = \frac{P_1(1 - P_1)}{P_3(1 - P_3)} - x_2 \frac{P_2(1 - P_2)}{P_3(1 - P_3)} = - (r_1x_1 + r_2x_2) \tag{9}\]

where \( r_i = \sqrt{P_1(1 - P)}/P_3(1 - P_3) \right) = \{ 1, 2 \} \text{ or } \right) = \{ 1, 2 \} \text{ and }

\[
q = Pr \left( \hat{P} \leq \hat{P} \leq P \pm \delta, i = 1, 2, 3 \right)
\]

\[
= Pr \left( x_i \in \pm \delta' \right) \sim \text{N} \left( \begin{array}{l} 0 \1 \rho \end{array} \right) \tag{10}\]

and approximate

\[
Pr \left( x_1 \in \pm \delta' \right) \sim \text{N} \left( \begin{array}{l} 0 \1 \rho \end{array} \right) \tag{11}\]

where \( x_i, x_2 \) is a sample realization of \( (x_1, x_2) \). Note that the right hand side of (11) depends on the sample size \( n \), which can be estimated by solving the equation

\[
\# \left\{ x_1 \in \pm \delta' \right\} \sim \text{N} \left( \begin{array}{l} 0 \1 \rho \right) \times 10^{-6} = q \tag{12}\]

When \( m = 4 \), the formulation is similar to the case of \( m = 3 \). Specifically, it is to solve \( n \) from

\[
q = Pr \left( \hat{P} \leq \hat{P} \leq P \pm \delta, i = 1, 2, 3 \right)
\]

\[
= Pr \left( x_1 \in \pm \delta' \right) \sim \text{N} \left( \begin{array}{l} 0 \1 \rho \right) \times 10^{-6} = q \tag{13}\]
where $P_i = \frac{1}{3}$, $\bar{x}_i$, $d_i$, and $r_i$ are similarly defined as the above.

\[
(x_1, x_2, x_3) \sim N\left(0, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \right)
\]

with $\rho_i = -P_i P_j / \sqrt{P_i (1 - P_i) P_j (1 - P_j)}$.

Equation (12) may be numerically solved for $n$ as programmed in R scripts or through a Monte Carlo approach described above.

2.2 | Collection and analysis of RAD-seq data from tetraploid potato cultivars

We conducted a breeding programme to generate an outbred segregation population from crossing two autotetraploid potato clones, 12-1-20 and 12-5-12, which are offspring from two autotetraploid potato cultivars, Atlantic and Longshu-3 (a Chinese variety cultivated in the northwest provinces of China). The outbred offspring population is composed of 304 individuals. The young leaf samples were collected from these offspring and their parental plants. Genomic DNA samples were extracted from the leaf samples using the Qiagen kit. The DNA samples were checked for quality and quantity by use of NanoDrop 2000c, then justified to a concentration of 100 ng/µl and stored at −80°C for further use. It is noted that each of the two parents were prepared in two biological replicate samples but no replication was made for the offspring individuals.

2.2.1 | Construction of RAD-seq libraries

The DNA samples made above were used to carry out restriction site associated DNA sequencing (RAD-seq) experiments. Protocol for constructing the RAD-seq library was modified from our previous work (Jiang et al., 2016), which confers the optimal target of minimizing the presence of undesirable DNA segments such as mitochondria DNA and DNA representing the RNA genes in the library and the optimal target of an even coverage of the DNA segments over the target genome. Each genomic DNA sample of 1 µg was firstly digested by two restriction enzymes (EcoRI-HF and Mspl, 5 U NEB). Adapters were added to each side of the enzyme digested segments by use of the T4 DNA ligase (200 U NEB). The Illumina adapters used in the experiment were in Table S1. These segments from 12 samples were mixed with an equal proportion. The pooled DNA sample was purified using the AxyPrep MAG PCR Clean-Up Kit (AxyGen) and then subjected to length selection for the DNA segments with 300–500 bp using the Pippin Prep (Sage Science). The DNA samples were further purified and amplified using the AxyPrep MAG PCR Clean-Up Kit (AxyGen) and then subjected to the second round of digestion by the restriction enzymes HhaI and HinfI (5 U, NEB). The enzyme-digested products were scanned for the DNA segments labelled with biotin using the Dynabeads MyOne Streptavidin T1 Streptavidin MagneSphere (Invitrogen). The selected DNA segments were used as templates for PCR amplification in a 50 µl reaction system, which were made up of 30–50 ng template, 1.2 µl (or 25 µM) forward and backward primers and 25 µl NEBNext High-Fidelity 2 × PCR Master Mix (M0541S, NEB). The PCR system was predegenerated at 98°C for 2 min and then subjected to 10 cycles of reaction, specifically degenerating at 98°C for 10 min, annealing at 56°C for 30 s and elongating at 72°C 1 min. Finally followed was elongating at 72°C for 5 min. The PCR products were purified using the Agencourt AMPure XP kit (Beckman Coulter, USA). The RAD-seq libraries so constructed were checked for quality by use of Agilent 2100 Analyser. The libraries were sequenced by Illumina Hiseq 3000 sequencer to generate a designed 4 MB sequence reads of 2 × 150 bp for each of the samples. Figure S1 summarizes a diagrammatic workflow of the optimizing RAD-seq library construction in the present study.

2.2.2 | Bioinformatic process of the sequence data

Sequence data from each of the pooled samples was separated into individual samples according to the corresponding barcode, and checked for quality using the trim_galore to trim the barcode and adaptor sequences as well as those nucleotides called with a Q value <30. The reads were aligned to the potato reference genome (DM v4.04 at http://solanaceae. plantbiology.msu.edu/pgsc_download.shtml) and only those reads, which were uniquely mapped with the mapping quality score ≥20, were to be included in further analyses. After the above steps of quality filtering, the sequence reads were subjected to screening for polymorphic sites embedded. A polymorphic site was declared if there were two or more different alleles (i.e., nucleotides) observed at the sequence site, that is, a SNP (single nucleotide polymorphism). When there were two or more polymorphic sites detected within a sequence read, the one with the highest Q value was chosen as the polymorphic site for the read and the others were ignored. For a demonstration purpose, only the sequence data from the two parental lines was used in the following analyses. The computational scripts developed here can be straightforwardly implemented to analyse any other data set.

3 | RESULTS

3.1 | Expected sequencing coverages required for detecting and genotyping at a SNP marker in tetraploids

Table 1 lists the number of sequence reads needed to ensure the presence of polymorphic alleles of a tetraploid genotype (m) in the sequence data for a given probability confidence (P) as well as the number of sequence reads required to ensure a tetraploid genotype (g) to be accurately predicted for a given probability confidence (P ± δ) with δ = 5%. It can be seen from the table that about one dozen of reads would be sufficient to ensure the presence of all constituent alleles of a tetraploid genotype in the sequence data with a probability of >95%. Obviously, a larger number of sequence reads are required to detect a larger number of constituent alleles of
a tetraploid genotype. For a given \( m \) (the number of constituent alleles of a tetraploid genotype), the evenness of the genotype alleles (e.g., AAGG) requires fewer sequence reads to detect their presence than the unevenness of the genotype alleles (e.g., AAAG or AGGG).

Although a moderate number of sequence reads are needed to detect the presence of alleles of a tetraploid genotype at a high statistical confidence (\( p > 95\% \)), the number of reads needed to accurately predict the tetraploid genotype will be few hundreds, depending on \( m \), the number of genotype constituent alleles. The larger is \( m \), the larger number of sequence reads is required to predict the genotype accurately. Again, the evenness of the genotype alleles requires a small number of sequence reads to predict the genotype than the unevenness of the alleles. It needs to be stressed that one may plug in any value of \( P \) and \( \delta \) in the program for numerical analyses.

### 3.2 Observed sequencing coverages required for detecting and genotyping at a SNP marker in tetraploid potato data

In the present study, we focus on the polymorphic marker data identified from the sequence data of the parental lines, each of which had two replicates of expected 4 M or a total 8 M sequence reads. Following the criteria described above for screening SNPs, we detected a total of 1,686,373 SNPs from the sequence data. Genome distribution of the SNPs is illustrated in Figure 1, showing a well even coverage of the SNPs across the potato reference genome, which is what we expected from the optimizing targets of the RAD-seq experiments and one of the important properties of the sequence based molecular marker data for any downstream population genetic analysis.

After filtering out those SNPs from the reads with coverage of less than 10, we obtained a total of 11,151 and 11,908 SNPs in the two parental plants, 12-1:20 and 12-5:12. Of the SNP markers, 11,116 (11,751), 33 (35) and two (122) have two, three and four segregating alleles respectively from plant 12-1:20 (or 12-5:12). This indicates that it is feasible to detect the presence of component alleles of a tetraploid genotype from the sequence data according to the theoretical prediction (Table 1). Because up to 95% of sequence variants show two alleles, we focused on the marker loci with two segregating alleles. There are three possible tetraploid genotypes, \( g_1, g_2 \) and \( g_3 \), at a two-allele SNP locus. For given \( n_1 \) and \( n_x \) observed numbers of alleles at the locus, we calculated

\[
\hat{p}_{k|n_1, n_2} = Pr [g_1|n_1, n_2] = Pr [n_1, n_2|g_1] / \sum_{g_j} Pr [n_1, n_2|g_j] \quad (13)
\]

and

\[
\hat{p}_{k|n_1, n_2} = Pr [g_2|n_1, n_2] = Pr [n_1, n_2|g_2] / \sum_{g_j} Pr [n_1, n_2|g_j] \quad (14)
\]

with

\[
Pr [n_1, n_2|g_k] = \binom{n}{n_1, n_2} \binom{k_1}{n_1} \binom{k_2}{n_2} \text{ and } Pr [n_1, n_2|g_k] = \binom{n}{n_1, n_2} \binom{n_1}{n_1} \binom{n_2}{n_2} \text{ respectively, where } g_k = (k_1, k_2) = (1, 3), (2, 2) \text{ or } (3, 1)
\]

Table 1 shows the number of sequence reads needed to detect the presence of polymorphic alleles in the sequence data (upper panel) and to predict genotype at the polymorphic site (lower panel).

| Number of alleles | \( m = 2 \) | \( m = 3 \) | \( m = 4 \) |
|-------------------|-------|-------|-------|
| Dosage            | 1:3   | 2:2   | 1:1:1 |
| \( p = 95\% \)    | 11    | 6     | 13    |
| \( p = 99\% \)    | 16    | 8     | 19    |
| \( p = 90\% \pm 0.05 \) | 271   | 203   | 352   |
| \( p = 95\% \pm 0.05 \) | 384   | 288   | 462   |

**Note:** \( M \), number of segregating alleles; Dosage, \( k_x: k_C: k_g: k_r \), indicates allele configuration of a tetraploid genotype; \( P (or P \pm \delta) \) is probability confidence of the prediction.

Figure 2 illustrates how \( \delta \) is related to coverage of sequence reads from which SNP genotypes are predicted, and shows a clear pattern that the certainty in prediction of tetraploid genotyping increases as the coverage increases. It is clear from this figure that there is still a large proportion of predicted genotypes deviating from the corresponding ones expected. Moreover, pattern of the frequencies is comparable between the two potato plants, suggesting repeatability of the relationship between the certainty of tetraploid genotype prediction and the coverage of sequence reads. To be more focusing, we calculated \( \rho = 0.05 \pm 0.05 \), the proportion of the SNP sites at which the genotype prediction lies within \( -0.05 \leq \delta \leq 0.05 \) and is presented in Table 2. It is clear that the proportion increases as the coverage of the sequence reads increases as predicted in Table 2. However, it is also clear that there is a substantially large proportion of genotypes predicted with a poor level of certainty even from sequence data with an exceptionally high level of coverage, and thus highlighting that it is clearly not a trivial task to recover a tetraploid genotype at a sequence variant site.

It should be noted that the two questions addressed in the present study will merge into one in diploids because appearance of two alleles at any sequence variant site of a diploid genome will suffice to infer an individual heterozygote genotype at the site. Assuming binomial distribution of sampling alleles from a diploid heterozygote genotype, one can simply the above equations to predict \( n \), the number of sequence reads that are needed to ensure the presence of two alleles of a heterozygote genotype in diploid with a prior given probability \( P \), from \( \log(1 - P)/\log(1/2) + 1 \), from which \( n \approx 5, 6 \) or 8 for \( p = 0.90, 95 \) or .99, respectively. This clearly indicates that sequencing coverage is a much less demanding for GBS in diploids than in polyploids. The formulation above has been programmed in R scripts which are presented as Text S1.
DISCUSSION

The next generation sequencing (NGS) technique enables generation of a large number of short sequence reads which may represent a reference genome with a predesigned coverage. Fast advancement in NGS makes it unprecedentedly economically doable to discover genome-wide sequence variants at a population scale. This has greatly facilitated sequence data driven population based genetic and genomic studies in ecology and evolutionary biology (Ellegren, 2014; Hill, 2012; Kim et al., 2016), and also motivated enormous methodological researches for modelling and analysing the NGS data (Li, 2011; Catchen et al., 2013; Begun et al., 2007). Digital nature of sequence reads data generated from NGS experiments allows genotyping at a polymorphic site embedded in any sequence reads, that is, genotyping by sequence (GBS). Codominance of polymorphic nucleotide alleles makes it feasible to predict individual genotype at any polymorphic site in diploids even with low coverage sequence reads (Wickland et al., 2017) because GBS in diploids merely needs to survey whether there is one (homozygote) or two alleles (heterozygote) present at the site. However, it is much more challenging to recover genotype of a polyploid even from codominant marker data including those identified from the sequence reads as detailed in Luo et al. (2000).

The challenge of GBS in polyploids may be attributed to several characteristic features of polyploid sequence data. First, the sequence data is highly relevant to but may be usually uninformative about allele dosage of a polyploid genotype at any sequence variant site (Margarido & Heckerman, 2015). Second, polyploid sequence data shows complex patterns of bio- and nonbiological variations (Dang et al., 2021; Gerard et al., 2015). Although enormous research efforts have been made to explore GBS in polyploid species (Griffin et al., 2011; Margarido & Heckerman, 2015; McKinney et al., 2018; Poland & Rife, 2012), little is yet known to what extent a polyploid genotype at a polymorphic site may be reliably recovered from sequence reads data. Third, there has been evidence showing there is allele loss at a substantial proportion of sequence variant sites in diploid and polyploid sequence data in the literature (Gautier et al., 2013; Shestak et al., 2021) and from our own unpublished tetraploid potato data. This adds an extra angle of complexity to GBS in polyploids.

The present study exploits two logically related but conceptually different questions, the numbers of sequence reads required to detect the presence of polymorphic alleles and to recover genotype at the polymorphic site in polyploids. These are highly relevant to how to effectively design a GBS experiment and how to reliably call for genotype from the sequence reads data in polyploids. Taking tetraploid as example, the paper provides generic statistical methods to predict the numbers of sequence reads to be required to ensure the presence or/and dosage of constituent alleles of a tetraploid genotype with a pregiven probability confidence. The formulation is made by assuming a multinominal distribution to model the sampling process of generating sequence reads from a given tetraploid genotype as in many polyploid sequence data modelling and analysis (Griffin et al., 2011; Margarido & Heckerman, 2015; McKinney et al., 2018; Poland & Rife, 2012).

Under the multinominal distribution assumption, we showed that a dozen of sequence reads are sufficient to ensure the presence of any tetraploid genotype alleles in the reads data with a probability confidence of 95%. However, it is much more demanding in term of the sequence coverage to accurately predict a tetraploid genotype at a polymorphic site embedded in sequence reads, specifically a few hundreds of the sequence reads are needed to reliably predict the genotype from the sequence data with a probability

FIGURE 1 Distribution of genomic DNA sequence variants identified from sequence reads data of the RAD-seq experiment with two tetraploid potato cultivars. The black bars indicate candidate chromosomal regions of the centromeres.
**FIGURE 2** Frequency of predicted tetraploid genotypes at two allele loci which are predicted with varying measures of accuracy $d$. The left (or right) panel of histograms are for potato plant 12-1-20 (or 12-5-12). From top to bottom of the histograms in each of the panels, coverage of SNP bearing sequence reads varies from 10+, 50–199, 200–999 and 1,000+.
Instead of sequencing the whole genome, restricted site association (RDA-seq) has been demonstrated to be effective in improving the efficiency of quantitative genetic analysis in tetraploids (Hackett et al., 2014). On the other hand, use of genotype information at molecular markers at a genome wide scale for many laboratory-based projects. On the other hand, it could be financially infeasible to sequence a polyploid population over the target genome. It must be pointed out that the number of sequence variant markers with sufficiently high coverage so that GBS could be accurately predicted.

It needs to be acknowledged that two parameters, $P$ and $\delta$, involved in the formulation presented in this study may take any values. Those considered here are for a numerical demonstration purpose and may be recognized to be of typical interests for GBS analyses in tetraploids. With the formulation presented and computational tools provided in the paper, experimentalists may explore any values of the parameters according to their needs. Specifically, choose of different values of these parameters has a direct impact on sensitivity and specificity of GBS experiments. With a more stringent (or larger) value of $P$ and/or a more stringent (or smaller) value of $\delta$, the number of sequence variant markers that pass the screening criteria would be smaller. There is a counterbalance between accuracy of GBS and the number of sequence variant sites scanned, and thus experimentalists need to explore various combinations of the parameters through the method provided by the present study so as to reach the one appropriate for their own research.

The present study has been focused on GBS prediction directly from sequence data of any tetraploid individual. It has been well documented that accuracy in predicting individual genotype at a polymorphic marker locus may be substantially improved from using the individual’s pedigree information in addition to its marker data as demonstrated in diploids (Chen et al., 2014) and in tetraploids (Luo et al., 2000). We will show how offspring sequence data may be implemented to improve the accuracy of prediction of their parental genotypes (Luo et al., unpublished data).

It has been well documented that newly formed polyploid genomes have been subject to the dynamic changes in both structure and function when compared to their diploid relatives and ancient counterparts (Solits & Solits, 1995; Comai et al., 2000). This may imply sequence coverage for accurate GBS may be more demanding in the newly formed polyploids than in ancient polyploids with more stable genomes. Tetraploid potato (Solanum tuberosum L.) has a domestication history of 8000 to 10,000 years (Spoon, 2005), suggesting a fairly genetically stable genome and thus that the sequence coverage estimated from the present study would be informative for accurate GBS in the tetraploid crop.

| TABLE 2 Proportion of predicted two-allele tetraploid genotypes with $-0.05 \leq \delta \leq 0.05$ from sequence data with varying coverages of sequence reads collected for the two parental potato cultivars (12-1-20 and 12-5-12) |
|---|---|---|---|---|---|
| c | 12-1-20 ($n = 11,116$) | | 12-5-12 ($n = 11,715$) | |
| 50 $\leq c \leq 200$ | 2429 | 1817 | 617 | 2950 | 2164 | 1009 |
| 200 $< c \leq 1000$ | 1817 | 617 | 2950 | 2164 | 1009 |
| $c > 1000$ | 50% | 25% | 1009 |
| $p_{-0.05 \leq \delta \leq 0.05}$ | 24% | 33% | 50% | 25% | 33% | 47% |

Note: $C$, number of sequence reads covered the polymorphic site; $n_c$, the number of polymorphic sites; $n_c$, number of sequence variant markers with sufficiently high coverage estimated from the present study would be informative for accurate GBS in the tetraploid crop.
Numerical analysis involves evaluation of bi- and/or multinomial probability distribution and may be recognized to be computationally challenging, particularly in the case of calculating multinomial probabilities. Method was proposed to approximate multinomial probability function in the classical literature in statistics, for example, Johnson (1960). By combining it with a more recently proposed method for approximating gamma function in Yang and Tian (2018), we tested and demonstrated the efficiency and accuracy of the method in statistically modelling sequence data from tetraploid potato (Dang et al., 2021). Additionally, numerical computation took only about 10 minutes on a Dell precision workstation computer to analyse the sequence data set from an individual sample.

ACKNOWLEDGEMENTS

The research is supported by research grants (31871240) from the National Natural Science Foundation of China, BBSRC (BB/N008952/1) and the Leverhulme Trust in the UK. We state that none of the authors have any competing interests. We thank three reviewers and the subject editor for their comments and suggestions which have helped improve clarity and presentation of an earlier version of the manuscript.

DATA AVAILABILITY STATEMENT

The sequence data analysed in the present study has been made available at http://figshare.com/articles/dataset/1-20-1_quality_inf_

REFERENCES

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One, 3(10), e3376.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y. P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., & Pachter, L. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biology, 5(11), e307. https://doi.org/10.1371/journal.pbio.0050310

Blischak, P. D., Kubatko, L. S., & Wolfe, A. D. (2018). SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics, 34, 407–415. https://doi.org/10.1093/bioinformatics/btx587

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics, 32, 314–331.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. Molecular Ecology, 22(11), 3124–3140. https://doi.org/10.1111/ mec.12354

Chen, N., Van Hout, C. V., Gottipati, S., & Clark, A. G. (2014). Using Mendelian inheritance to improve high-throughput SNP discovery. Genetics, 198, 847–857. https://doi.org/10.1534/genet ics.114.169052

Comai, L., Tyagi, A. P., Winter, K., Holmes-Davis, R., Reynolds, S. H., Stevens, Y., & Byers, B. (2000). Phenotypic instability and rapid gene silencing in newly formed Arabidopsis allotetraploids. The Plant Cell, 12, 1551–1567. https://doi.org/10.1105/tpc.12.9.1551

Dang, Z., Wang, L., Yang, J., Tao, Q., Zhang, F., Zhang, Y., & Luo, Z. W. (2021). Sampling variation of RAD-Seq data from diploid and tetraploid potato (Solanum tuberosum L.). Plants, 10(2), 319. https://doi.org/10.3390/plants10020319

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12, 499–510. https://doi.org/10.1038/nrg3012

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. Trends in Ecology and Evolution, 29(1), 51–63. https://doi.org/10.1016/j.tree.2013.09.008

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdélhué, C., Pudlo, P., Cornuet, J.-M., & Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology, 22, 3165–3178. https://doi.org/10.1111/mec.12089

Gerard, D., Ferrao, L. F. V., Garcia, A. A. F., & Stephens, M. (2015). Genotyping polyploids from messy sequencing data. Genetics, 210, 789–807. https://doi.org/10.1534/genetics.115.301468

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R. C., Antolin, R., & Hickey, J. M. (2017). Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. Crop Sciences, 57, 1404–1420. https://doi.org/10.2135/crops sci2016.08.0675

Griffin, P. C., Robin, C., & Hoffmann, A. A. (2011). A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poo grasses. BMC Biology, 9, 19. https://doi.org/10.1186/1741-7007-9-19

Hackett, C. A., Bradshaw, J. E., & Bryan, G. J. (2014). QTL mapping in autotetraploids using SNP dosage information. TAG. Theoretical and Applied Genetics, 127, 1885–1904. https://doi.org/10.1007/s00122-014-2347-2

Hill, W. G. (2012). Quantitative genetics in the genomics era. Current Genomics, 13(2), 196–206.

Jarne, P., & Lagoda, P. J. (1996). Microsatellites, from molecules to populations and back. Trends in Ecology & Evolution, 11, 424–429. https://doi.org/10.1016/S0169-5347(96)00494-5

Jiang, N., Zhang, F., Wu, J., Chen, Y., Hu, X., Fang, O. U., Leach, L. J., Wang, D. I., & Luo, Z. (2016). A highly robust and optimized sequence-based approach for genetic polymorphism discovery and genotyping in large plant populations. TAG. Theoretical and Applied Genetics, 129, 1739–1757. https://doi.org/10.1007/s00122-016-2732-9

Johnson, N. L. (1960). An approximation to multinomial distribution: some properties and applications. Biometrika, 47(1), 93–102.

Kim, G., Guo, H., Kong, R., Chandri, R., Shuang, L. S., & Paterson, A. H. (2016). Application of genotyping sequencing technology to a variety of crop breeding programs. Plant Science, 242, 14–22.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. Bioinformatics, 27, 2987–2993.

Luo, Z. W., Hackett, C. A., Bradshaw, J. E., McNicol, J. W., & Milbourne, D. (2000). Predicting parental genotypes and gene segregation for tetrasomic inheritance. Theoretical and Applied Genetics, 100, 1067–1073. https://doi.org/10.1007/s001220051388

Luo, Z. W., Wu, C.-I., & Kearsey, M. J. (2002). Precision and High-Resolution Mapping of quantitative trait loci by use of recurrent selection, backcross or intercross schemes. Genetics, 161, 915–929. https://doi.org/10.1093/genetics/161.2.915

Margarido, G. R. A., & Heckerman, D. (2015). ConPASE: Genome ensemble ploidy estimation from next-generation sequencing data. PLoS Computational Biology, 11(4), e1004229. https://doi.org/10.1371/journal.pcbi.1004229
McKinney, G. J., Waples, R. K., Pascal, C. E., Seeb, L. W., & Seeb, J. E. (2018). Resolving allele dosage in duplicated loci using genotyping-by-sequencing data: A path forward for population genetic analysis. *Molecular Ecology Resources, 18*(3), 570–579. https://doi.org/10.1111/1755-0998.12763

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology, 22*(11), 2841–2847. https://doi.org/10.1111/mec.12350

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics, 12*(6), 435–451. https://doi.org/10.1038/nrg2986

Poland, J. A., & Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome, 5*(3). https://doi.org/10.3835/plantgenome2012.05.0005

Shestak, A. G., Bukaeva, A. A., Saber, S., & Zaklyazminskaya, E. V. (2021). Allelic dropout Is a common phenomenon that reduces the diagnostic yield of PCR-based sequencing of targeted gene panels. *Frontiers in Genetics, 12*, 620337. https://doi.org/10.3389/fgene.2021.620337

Soltis, D. E., & Soltis, P. S. (1995). The dynamic nature of polyploid genome. *PNAS, 92*, 8089–8091.

Spooner, D. M., McLean, K., Ramsay, G., Waugh, R., & Bryan, G. J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. *PNAS, 102*(41), 14694–14699.

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods, 12*, 1061-1063. https://doi.org/10.1038/nmeth.3582

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. V. D., Hornes, M., Fritters, A., Pot, J., Paleman, J., Kuiper, M., & Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research, 23*, 4407–4414. https://doi.org/10.1093/nar/23.21.4407

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics, 18*, 586. https://doi.org/10.1186/s12859-017-2000-6

Yang, Z. H., & Tian, J. F. (2018). An accurate approximation formula for gamma function. *Journal of Inequalities and Applications, 56*. https://doi.org/10.1186/s13660-018-1646-6

Yu, X., Zhang, M., Yu, Z., Yang, D., Li, J., Wu, G., & Li, J. (2020). An SNP-based high-density genetic linkage map for tetraploid potato using specific length amplified fragment sequencing (SLAF-Seq) technology. *Agronomy, 10*(1), 114. https://doi.org/10.3390/agronomy10010114

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Wang, L., Yang, J., Zhang, H., Tao, Q., Zhang, Y., Dang, Z., Zhang, F., & Luo, Z. (2022). Sequence coverage required for accurate genotyping by sequencing in polyploid species. *Molecular Ecology Resources, 22*, 1417–1426. https://doi.org/10.1111/1755-0998.13558