Theor Appl Genet (2017) 130:717–726
DOI 10.1007/s00122-016-2845-5

ORIGINAL ARTICLE

Automated tetraploid genotype calling by hierarchical clustering

Cari A. Schmitz Carley[1] · Joseph J. Coombs[2] · David S. Douches[2] · Paul C. Bethke[1,3] · Jiwan P. Palta[1] · Richard G. Novy[4] · Jeffrey B. Endelman[1]

Received: 19 August 2016 / Accepted: 22 December 2016 / Published online: 9 January 2017

Abstract

Key message **New software to make tetraploid genotype calls from SNP array data was developed, which uses hierarchical clustering and multiple F1 populations to calibrate the relationship between signal intensity and allele dosage.**

Abstract SNP arrays are transforming breeding and genetics research for autotetraploids. To fully utilize these arrays, the relationship between signal intensity and allele dosage must be calibrated for each marker. We developed an improved computational method to automate this process, which is provided as the R package ClusterCall. In the training phase of the algorithm, hierarchical clustering within an F1 population is used to group samples with similar intensity values, and allele dosages are assigned to clusters based on expected segregation ratios. In the prediction phase, multiple F1 populations and the prediction set are clustered together, and the genotype for each cluster is the mode of the training set samples. A concordance metric, defined as the proportion of training set samples equal to the mode, can be used to eliminate unreliable markers and compare different algorithms. Across three potato families genotyped with an 8K SNP array, ClusterCall scored 5729 markers with at least 0.95 concordance (94.6% of its total), compared to 5325 with the software fitTetra (82.5% of its total). The three families were used to predict genotypes for 5218 SNPs in the SolCAP diversity panel, compared with 3521 SNPs in a previous study in which genotypes were called manually. One of the additional markers produced a significant association for vine maturity near a well-known causal locus on chromosome 5. In conclusion, when multiple F1 populations are available, ClusterCall is an efficient method for accurate, autotetraploid genotype calling that enables the use of SNP data for research and plant breeding.

Introduction

Cytological and genomic studies reveal that whole genome duplications have occurred many times in the evolution of plants, such that nearly all plants can be considered "paleopolyploids" (Stebbins 1950; Leitch and Bennet 1997; Renny-Byfield and Wendel 2014). When more than two homologous chromosomes are present—a condition known as autopolyploidy—multiple pairs of bivalents, and sometimes higher order multivalent structures, form in meiosis (Mather 1936). This leads to unique segregation patterns in autopolyploids compared to diploids. In multivalents, recombination can occur between homologs that migrate to the same pole in meiosis I, also known as the reductional division. It is then possible for parts of sister chromatids

Communicated by Christine A. Hackett.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-016-2845-5) contains supplementary material, which is available to authorized users.

✉ Jeffrey B. Endelman
endelman@wisc.edu

[1] Department of Horticulture, University of Wisconsin, Madison, WI 53706, USA

[2] Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA

[3] USDA Agricultural Research Service, Madison, WI 53706, USA

[4] USDA–ARS Small Grains and Potato Germplasm Research Unit, Aberdeen, ID 83210, USA

to migrate to the same pole in meiosis II—a phenomenon called "double reduction" (Mather 1936).

Despite the ubiquity of ancestral autopolyploidy, there are fewer species in which the condition persists, presumably because of the higher frequency of complications during mitosis and meiosis (Comai 2005). The most common form of autopolyploidy in cultivated species is tetraploidy, which includes potato (Bourke et al. 2015), highbush blueberry (Krebs and Hancock 1989), leek (Jones et al. 1996) and alfalfa (Quiros 1982). In some species, such as peanut (Leal-Bertioli et al. 2015), only certain chromosomes have sufficient homology to exhibit tetrasomic inheritance—a condition known as segmental allotetraploidy.

Single nucleotide polymorphism (SNP) arrays are having a significant impact on the pace and scope of genetics and breeding research in tetraploids. Examples include the 8303 SNP array for potato (Hamilton et al. 2011; Felcher et al. 2012), the 9K array for alfalfa (Li et al. 2014a), and the 68K array for rose (Koning-Boucoiran et al. 2015). These arrays produce a fluorescent signal that varies with the dosage of the two alleles for each marker, which can be visualized in a two-dimensional plot (Fig. 1). When a marker works well, the homozygous samples should cluster near the axes, while heterozygous samples appear at intermediate positions. For diploids, one expects a single heterozygous cluster (AB) midway between the two axes; for tetraploids there can be up to three heterozygous clusters (AAAB, AABB, ABBB). The Cartesian (x,y) representation in Fig. 1 can be converted to polar coordinates (r,θ), where r is the magnitude of the signal and theta (θ) is the angle from the x-axis. Following the convention used in the Illumina GenomeStudio software, theta values have been normalized by $\pi/2$ radians to lie on the [0, 1] interval.

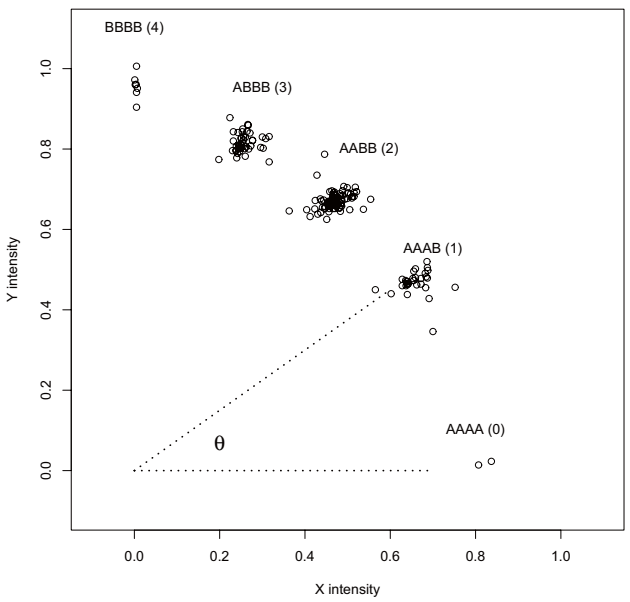

Fig. 1 Illustrating the relationship between relative fluorescence intensity from the A allele probe (X), the B allele probe (Y), and theta for a bi-allelic SNP. Marker genotypes are shown in both alphabetic (*AAAA* nulliplex; *AAAB* simplex; *AABB* duplex; *ABBB* triplex; *BBBB* quadruplex) and numeric (*0–4, in parentheses*) nomenclature. Theta is normalized by $\pi/2$ radians to lie on the [0, 1] interval

As illustrated in Fig. 1, theta values convey information about allele dosage, and both manual and automated approaches have been used to convert theta values into tetraploid genotypes. For the 8303 potato SNP array, Hirsch et al. (2013) visually inspected the theta values of a diversity panel combined with diploid and tetraploid mapping populations to devise a set of marker-specific theta "boundaries", which delineate the minimum and maximum theta value for each allele dosage. Although visual inspection is accurate, the method is time-consuming and impractical for arrays with tens or hundreds of thousands of SNPs.

Voorrips et al. (2011) developed an automated approach based on fitting a mixture of normal distributions, which is available as the R package fitTetra. Both Hackett et al. (2013) and Bourke et al. (2015) employed normal mixture models to assign genotypes in a single F1 population of potato (the latter used fitTetra while the former used an independent implementation), and Vos et al. (2015) used fitTetra for a panel of 537 potato lines. Bourke et al. (2015) reported that of the 15,137 SNPs called by fitTetra, 1370 were subsequently removed due to incompatible parent-offspring genotypes or poor fit to the expected segregation ratio; whether these problems were present with the allele intensity values or introduced during genotype calling was not determined. Vos et al. (2015) reported that fitTetra scored 15,271 markers and rejected 2716. Upon visual inspection of the rejected markers, Vos et al. (2015) found that 843 could be scored manually. Of the 15,271 fitted markers, Vos et al. (2015) rejected 378 based on the analysis of segregation ratios in an F1 mapping population. When an additional 1832 markers with more than 5% missing data were visually inspected, it was discovered that for 626 markers fitTetra "produced false negative genotype calls based on correct marker signal intensities" (Vos et al. 2015).

Our objective was to develop an improved, automated method for converting theta values into tetraploid genotypes. The pipeline was implemented as the R package ClusterCall, which is available at http://potatobreeding.cals.wisc.edu/software under the GNU General Public License. A detailed tutorial, or vignette, can be downloaded with the software.

Materials and methods

ClusterCall algorithm

The ClusterCall algorithm is applied independently to each marker and can be divided into training and prediction phases. In the training phase (function *CC.bipop*), hierarchical clustering within an F1 population is used to group samples with similar theta values, and then allele dosages are assigned to clusters based on expected segregation ratios (Fig. S1, Supplementary Material 1). In the prediction phase (function *CC.anypop*), multiple F1 populations and the prediction set are clustered together, and the mode allele dosage of the training set samples is used to predict all samples in the cluster.

In the first step, theta values within an F1 population are hierarchically clustered (R function *hclust*, method = "complete"; R Development Core Team 2015) based on Euclidean distance. The number of clusters is varied from $k = 2$ to 5, and only if the partitioning satisfies three criteria (range, separation, and compactness) is it passed to the genotype assignment module: (1) the range for each cluster is less than the parameter *max.range* (default = 0.25), (2) the separation distance between the medians for adjacent clusters exceeds the parameter *min.sep* (default = 0.05), and (3) the fraction of samples with cluster membership probability less than *min.posterior* (default = 0.7) is no more than *max.missing* (default = 0.05).

The *min.posterior* parameter is used to identify samples that do not cluster tightly and assign them no call (NA). The probability $p(x|\theta)$ of a sample with theta value θ belonging to cluster x is calculated using Bayes theorem, $p(x|\theta) = p(\theta|x)/\sum_z p(\theta|z)$, where the sum is taken over all clusters and uniform priors are assumed. After numerical experiments with several distributions for $p(\theta|x)$, we settled on *t*-distributions, for which the location, scale, and degrees-of-freedom parameters are fitted for each cluster independently by maximum likelihood, using the *fitdistr* function in R package MASS (Venables and Ripley 2002). The *t*-distribution fit sometimes fails for clusters near the boundaries of the [0, 1] interval. This error is caught by ClusterCall and leads to a fit of the heavy-tailed log-normal distribution.

At this point in the workflow, ClusterCall checks whether the marker is segregating or should be called monomorphic in the progeny. This decision is based on whether the proportion of total samples in any one cluster exceeds the parameter *mono.thresh* (default = 0.95). If this threshold is met, there are three possible outcomes: (1) if the range of all samples $\leq$ *max.range* and the median $\leq$*hom.theta* (a parameter for homozygous cluster proximity to the [0, 1] boundaries; default = 0.15), the parents and progeny are assigned genotype 0; (2) if the range of all samples $\leq$ *max.range* and the median $\geq 1 -$ *hom.theta*, the parents and progeny are assigned genotype 4; (3) if $\theta \leq$ *hom.theta* for one parent and $\theta \geq 1 -$ *hom.theta* for the other parent, the parents are assigned genotypes 0 and 4, respectively, and the progeny are assigned a genotype of 2.

For segregating markers, the clustering solutions that meet the range, separation, and compactness criteria are passed to the genotype assignment module. Initially, the k clusters are ordered from lowest ($\tilde{\theta}_1$) to highest ($\tilde{\theta}_k$) median theta value (the tilde denotes the median). When $k = 5$, genotypes 0–4 are assigned to the clusters, provided the lowest and highest clusters satisfy the *hom.theta* threshold: $\tilde{\theta}_1 \leq$ *hom.theta* and $\tilde{\theta}_k \geq 1 -$ *hom.theta*. For $k < 5$, any solutions with $\tilde{\theta}_1 \leq$ *hom.theta* and $\tilde{\theta}_k \geq 1 -$ *hom.theta* are discarded, as assigning genotypes 0 and 4 to the low and high clusters would create a skip in the genotype sequence, which is not expected in an F1 population. When $k = 4$, genotypes 0–3 are assigned if $\tilde{\theta}_1 \leq$ *hom.theta*, while if $\tilde{\theta}_k \geq 1 -$ *hom.theta*, genotypes 1–4 are assigned. For $k = 3$, the genotype assignment can be 0–2, 1–3, or 2–4, depending on the values of $\tilde{\theta}_1$ and $\tilde{\theta}_k$. For $k = 2$, the genotype assignment is 0–1 or 3–4, depending on the values of $\tilde{\theta}_1$ and $\tilde{\theta}_k$.

The goodness-of-fit for each genotype assignment is calculated using Pearson's chi-squared test (*chisq.test* function in *R*), which relies on the normalized sum of squared deviations between the expected and observed counts for each genotype class. The expected segregation ratio is based on the random bivalents model (aka random chromosome segregation, Gallais 2003) and depends on the genotypes of the parents, which, therefore, must be included in the population. Samples with genotypes that are not possible under the random bivalents models (e.g., due to double reduction) are treated as missing for the chi-squared test. The use of the random bivalents model to score goodness-of-fit does not prevent the software from making accurate genotype calls for double reduction products, as they occur at low frequency. If either parent fails the *min.posterior* threshold at a particular marker and, therefore, receives no call, the marker is not scored. If one or more samples were assigned genotypes incompatible with the parental genotypes at a marker (allowing for double reduction), that clustering solution is rejected. For example, if one parent has genotype 0, the maximum genotype for the progeny is 2.

The optimal number of clusters for a marker is selected based on maximizing the *p*-value from the chi-squared test. However, to prevent the spurious loss of small clusters with double reduction products due to sampling variation, a penalty is applied for reducing the number of clusters. To accept a solution with fewer clusters, the $\log_{10}p$ value must increase by at least the parameter *chi2p.step* (default = 1). The importance of this parameter is illustrated in the "Results".

To predict genotypes for an arbitrary population, ClusterCall uses the genotype assignments from two or more F1 populations as a training set. Theta values from the training and prediction sets are clustered together as one population, using the Euclidean distance metric. The genotype assignment for each cluster is the most prevalent genotype, or mode, of the training set samples in that cluster. If multiple clusters are assigned the same genotype, the clustering solution is rejected. Clusters without at least one called training set sample are assigned no call (NA), except when there are four other clusters with assigned genotypes. In this case, if the cluster without a genotype has median theta value less (greater) than *hom.theta* ($1 - hom.theta$), it is assigned a genotype of 0 (4).

Beginning with $k = 5$ clusters, the software checks that the solution meets the range, separation, and compactness criteria described previously. If not, ClusterCall reduces k by 1 and checks again, continuing in this way until a clustering solution is found that meets the criteria. Our analysis is limited to markers segregating in at least one F1 population, so $k = 2$ is the minimum number of clusters. No further reductions in k are attempted after a valid clustering solution is found to prevent the spurious loss of prediction set clusters with genotype values not present in the training set. If the proportion of samples with no call in the prediction set is less than max.missing (default = 0.05), the function *CC.anypop* returns the genotype calls and a concordance statistic, defined as the proportion of the training set for which the genotype call in the F1 population matches the genotype call from the multi-family clustering.

Application to potato

Three F1 populations and a diversity panel of tetraploid potatoes were used in this study. All samples were genotyped with the Infinium 8303 SNP array for potato (Hamilton et al. 2011; Felcher et al. 2012), and theta values were exported from the Illumina GenomeStudio software for each family and the diversity panel as separate projects. Using *hclust* in R, hierarchical clustering of samples based on the Euclidean distance (calculated from theta values) was used to identify spurious progeny for removal. Replicate samples were combined by taking the median value, resulting in $n = 160$, 191, and 162 progeny for the families of Atlantic × Superior (AxS, Zorrilla et al. 2014; Supplementary Materials 2 and 3), Wauseon × Lenape (WxL, Supplementary Materials 4 and 5), and Rio Grande × Premier (RGxP, Douches et al. 2014; Supplementary Materials 6 and 7), respectively. The diversity panel consisted of 187 tetraploid individuals (Hirsch et al. 2013; Supplementary Material 8) with phenotype data from Rosyara et al. (2016) for the following traits: vine maturity at 95 and 120 days, total yield, chip color, tuber width, length, shape, eye depth, fructose, sucrose and malic acid.

ClusterCall results are based on the default parameter values in version 1.3 of the package, which were chosen to maximize the number of markers with concordance ≥0.95 across the three potato families.

ClusterCall was compared against R package fitTetra (Voorrips et al. 2011; version 1.0) for the F1 populations. The *saveMarkerModels* function was used, which attempts to fit eight different normal mixture models, using different constraints on the means and mixing proportions of the distributions, and then selects the model with the lowest Bayesian Information Criterion (BIC). Initially we used the signal ratio, $X_{raw}/(X_{raw} + Y_{raw})$, as suggested in the fitTetra manual, but fitTetra produced more high-concordance markers when theta values were used. (Concordance can be calculated for any genotype calling method, based on clustering independent results for multiple families.) Three fitTetra parameters were adjusted to increase the number of markers with concordance ≥0.95 across the potato families: (1) *peak.threshold*, a maximum threshold of scored samples in a single cluster, (2) *p.threshold*, a minimum *p*-value threshold for assigning a sample to a cluster, and (3) *sd.threshold*, a maximum threshold on the standard deviation of sample signal ratios within clusters. *The peak.threshold* parameter was set at 1 to allow fitTetra to call monomorphic markers. The *p.threshold* parameter was varied between 0.7 and 0.99, and results are shown for a value of 0.85. Values in the range of 0.025–0.3 were tested for *sd.threshold*, and results are shown using the default value of 0.1.

Linkage analysis of the ClusterCall genotypes was used to further validate its performance in the F1 populations. Following the method of Luo et al. (2001), an R script was written to automate calculation of the coefficients of the log likelihood (LL) for the recombination fraction (r) between pairs of markers, for every possible combination of parental phases. Maximum likelihood estimates for r (denoted r_{ML}) were obtained by numerical maximization in the interval [0, 0.5], using R function *optimize*, and the phase configuration with the largest LL evaluated at r_{ML} was used to estimate the LOD score for linkage as $LOD = LL\{r = r_{ML}\} - LL\{r = 0.5\}$. Hierarchical clustering (*hclust*, method "single") of the LOD scores was used to split the polymorphic markers into linkage groups for each family.

For the diversity panel, ClusterCall was compared against the custom cluster file of Hirsch et al. (2013), which was developed by visual inspection of each marker. ClusterCall genotypes for the diversity panel were used for QTL mapping with R package GWASpoly (Rosyara et al. 2016) and compared against the GWAS results with the original genotype calls. Four marker-effect models were tested (additive, simplex-dominant, duplex-dominant, and

general), using a genome-wide significance level of 0.05 for each model based on 1000 random permutations.

Results

ClusterCall algorithm

In the training phase of ClusterCall, optimal genotype-to-cluster assignments are made for each marker and each F1 family independently (workflow in Fig. S1, Supplementary Material 1). For segregating markers there can be from two to five clusters, all of which are considered (dendrogram cuts in Fig. S2, Supplementary Material 1). Solutions are eliminated using criteria such as the maximum range for each cluster, the minimum separation between clusters, and the compactness of the clusters (see "Materials and methods").

Figure 2 illustrates this process using a "band" diagram for solcap_snp_c2_49209 in the RGxP family, in which the vertical axis is theta and the horizontal axis is an arbitrary sample index. The symbol for each sample is the genotype call (integer between 0 and 4). For this marker, only the three- and four-cluster solutions passed the range, separation, and compactness criteria. After assigning genotype values to each cluster (see "Materials and methods"), the goodness-of-fit *p*-value to the expected segregation ratio was calculated for each solution.

During the development of ClusterCall, we initially tried to select the optimal number of clusters based on maximizing the *p*-value, but Fig. 2 illustrates the problem with this approach. The only difference between the four- and three-cluster solutions is whether the sample with $\theta = 0.36$ is grouped with the cluster assigned genotype 2 (Fig. 2b) or forms its own cluster assigned genotype 1 (Fig. 2a). Visually, the four-cluster solution seems more likely to be correct, which is confirmed when this family is aligned with the other two families and the SolCAP diversity panel (Fig. S3, Supplementary Material 1). In Fig. 2, the parents of the family are circled in blue on the far left. With a genotype assignment of 3 for both parents, the expected segregation ratio under the random bivalents model is 1:2:1 for genotypes {2,3,4}. Since a genotype of 1 is not possible under the random bivalents model, we conclude that the $\theta = 0.36$ sample was the result of double reduction. Due to random segregation, the goodness-of-fit p-value is slightly higher when the $\theta = 0.36$ sample is assigned genotype 2 ($p = 0.61$) than when it is assigned genotype 1 ($p = 0.58$).

To prevent the loss of small-membership clusters containing double reduction products, ClusterCall considers *k*-cluster solutions from $k = 5$ down to 2, accepting a solution with fewer clusters as superior only when the $\log_{10}$p increases by an amount controlled by the parameter *chi2p.step*. After exploring a range of values for this parameter, a value of 1 was chosen as the ClusterCall default and used in this study. In the case of solcap_snp_c2_49209, the $\log_{10}p$ value only increased by 0.02 when the number of clusters decreased from 4 to 3, and the four-cluster solution was selected as optimal.

Performance measures for ClusterCall

Since the true genotype for each sample is unknown, ClusterCall was designed to quantify the consistency of the genotype calls made independently for several F1 populations. After clustering the theta values for multiple families together as one set, concordance is defined as the proportion of samples with genotype equal to the mode for each cluster. To illustrate, Fig. 3a shows the band diagram for solcap_snp_c1_10494 for three potato F1 populations. For this marker the genotype clusters line up across the families, and concordance is 1. Markers with low concordance can be visually inspected to determine if a satisfying genotype assignment is possible (Fig. S4a, Supplementary

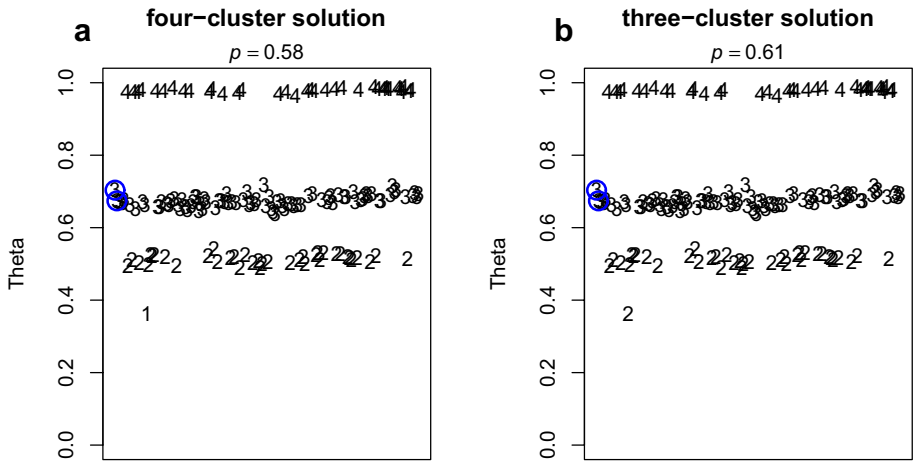

Fig. 2 Selecting the optimal clustering for solcap_snp_c2_49209 in the Rio Grande × Premier family. **a** With four clusters, the *p*-value is 0.58 for the chi-squared test to the expected segregation ratio. **b** With three clusters, the double reduction product at $\theta = 0.36$ is grouped with the duplex (2) cluster, but the *p*-value increases to 0.61. *Plotting symbols* are the assigned allele dosage (0 nulliplex to 4 quadruplex), and parental samples appear on the far *left*, with *blue circles*

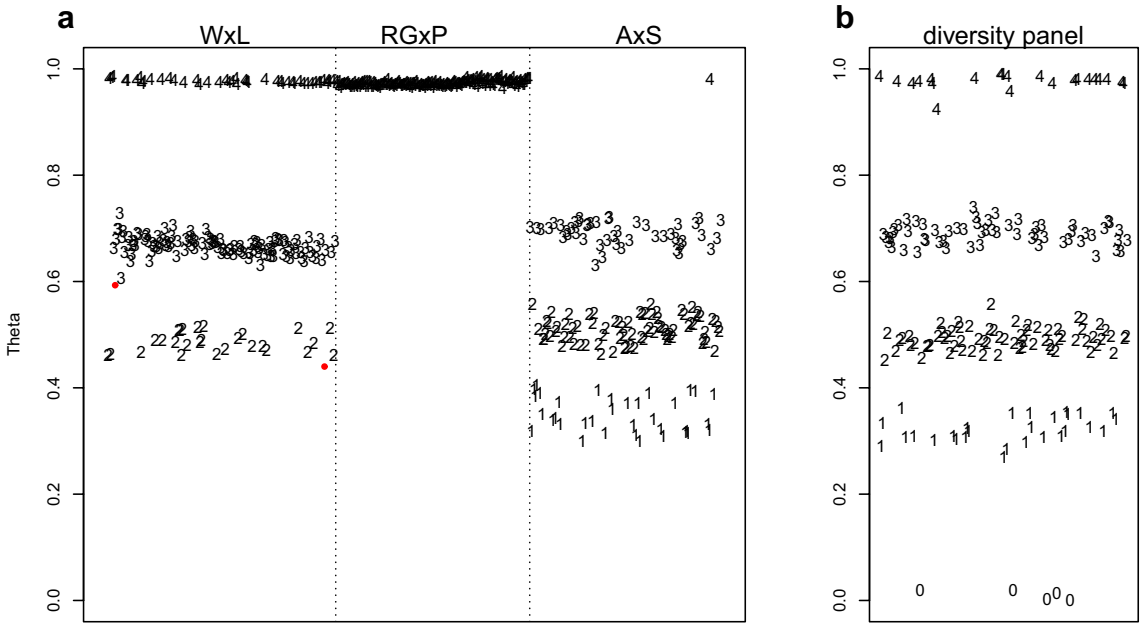

Fig. 3 Visualizing the concordance of ClusterCall genotypes at solcap_snp_c1_10494 for **a** the Wauseon x Lenape (WxL), Rio Grande × Premier (RGxP), and Atlantic × Superior (AxS) families (separated by *dashed vertical lines*), and **b** the SolCAP diversity panel, using the three families as a training set. *Plotting symbols* are the assigned allele dosage (0 nulliplex to 4 quadruplex), and *red points* indicate samples with no genotype call (NA)

Material 1) or if the marker appears to be unreliable (Fig. S4b, Supplementary Material 1).

Figure 4 compares the concordance distribution for markers segregating in at least one F1 population, when scored by ClusterCall vs. the software package fitTetra. With ClusterCall, 4217 markers had perfect concordance across the families, compared to 3478 with fitTetra. By relaxing the threshold to 0.95 concordance, the marker count for ClusterCall increased to 5729 compared to 5325 with fitTetra. Whereas 95% of the markers scored by ClusterCall had ≥0.95 concordance, only 83% of the fitTetra markers met this threshold. Of the 5434 markers scored by both software packages, 4611 SNPs had ≥0.95 concordance with both, 564 had ≥0.95 concordance with ClusterCall but not fitTetra, and 116 had the reverse pattern.

To further test the quality of the ClusterCall genotypes, linkage analysis was performed within each F1 population and then compared across the families. There were 5532 markers with ≥0.95 concordance and unique positions in version 4.03 of the DM reference genome (Potato Genome Sequencing Consortium 2011; Sharma et al. 2013), and 5503 were mapped to a chromosome in at least one family (Table S1, Supplementary Material 9). Of the 4976 markers mapped in at least two families, there were (1) 5 markers with conflicting chromosome assignments across families, (2) 57 previously unanchored markers, (3) 4905 markers for which the linkage analysis agreed with the reference genome, and (4) 14 markers with consistent results across

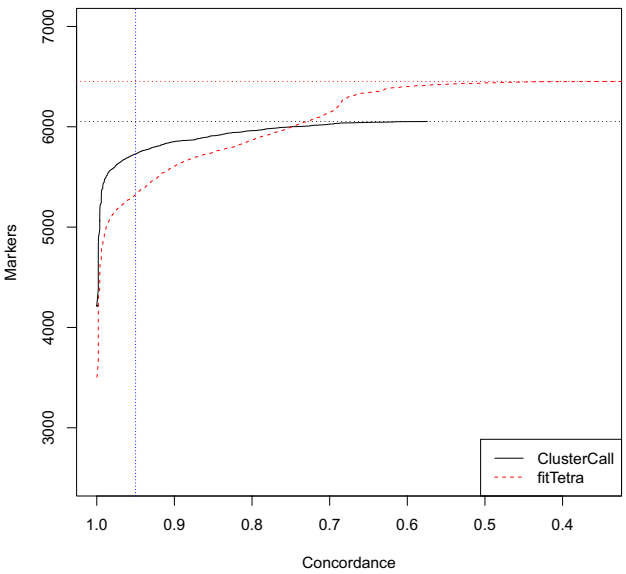

Fig. 4 Distribution of the genotype concordance across three potato F1 populations, using R packages ClusterCall and fitTetra. Concordance is defined as the proportion of samples within a cluster with genotype equal to the cluster mode. *Horizontal dotted lines* indicate the total marker count for each method. The *vertical dotted blue li*ne indicates a 0.95 concordance threshold

the families but a conflict with the reference genome, most of which lie on superscaffolds for which conflicts have been reported previously (Table 1).

Table 1 Conflicts between linkage groups and DMv4.03 pseudomolecules for families Rio Grande × Premier (RG×P), Atlantic × Superior (A×S), and Wauseon × Lenape (W×L)

SNP	RGxP	AxS	WxL	DMv4.03	Pos (bp)	Super scaffold	Literature support
solcap_snp_c1_1770	9	9	9	11	20,596,535	PGSC0003DMB000000211	
solcap_snp_c2_16746	4	4	4	3	41,496,608	PGSC0003DMB000000414	Hackett et al. (2013)
solcap_snp_c2_21309	NA	9	9	10	42,008,529	PGSC0003DMB000000220	Hackett et al. (2013)
solcap_snp_c2_37342	NA	11	11	6	55,960,752	PGSC0003DMB000000575	Hackett et al. (2013)
solcap_snp_c2_4393	9	9	9	7	18,715,095	PGSC0003DMB000000224	Felcher et al. (2012)
solcap_snp_c2_4567	NA	9	9	7	12,696,253	PGSC0003DMB000000348	Bourke et al. (2015), Hackett et al. (2013) and Felcher et al. (2012)
solcap_snp_c2_49280	NA	11	11	5	15,735,037	PGSC0003DMB000000317	Hackett et al. (2013)
solcap_snp_c2_49282	11	11	11	5	15,735,180	PGSC0003DMB000000317	Felcher et al. (2012)
solcap_snp_c2_49847	7	7	7	4	13,556,151	PGSC0003DMB000000067	Endelman and Jansky (2016)
solcap_snp_c2_49848	7	7	NA	4	13,555,806	PGSC0003DMB000000067	Endelman and Jansky (2016)
solcap_snp_c2_49850	7	7	7	4	13,555,777	PGSC0003DMB000000067	
solcap_snp_c2_49851	7	7	7	4	13,555,746	PGSC0003DMB000000067	
solcap_snp_c2_52829	NA	3	3	10	51,201,738	PGSC0003DMB000000106	Felcher et al. (2012)
solcap_snp_c2_54581	8	8	8	1	24,708,276	PGSC0003DMB000000195	Bourke et al. (2015), Endelman and Jansky (2016)

Literature support lists publications in which the marker was reported on the same linkage group as determined in this study

Genotype calls for arbitrary populations

To predict genotypes for an arbitrary population, such as a diversity panel or breeding population, ClusterCall uses the genotype assignments from two or more F1 populations as a training set. The prediction is made by hierarchical clustering of all the samples (training and prediction) together, and the genotype call for the cluster is the mode of the training set samples. Figure 3 shows the result for solcap_snp_c1_10494, illustrating how the clusters in the SolCAP diversity panel (Fig. 3b, $n = 187$) align with the corresponding clusters in the F1 populations (Fig. 3a).

If a cluster does not contain any training set samples, which will happen when the prediction set contains allele dosages not present in the training set, the samples in that cluster are not called (Fig. S5, Supplementary Material 1). An exception to this rule is when four other clusters are defined and the remaining cluster can be assigned the 0 or 4 genotype (see "Materials and methods"). Figure 3b illustrates this feature, as genotype 0 was assigned to the lowest cluster even though it contained no samples from the training set. To prevent the spurious lumping together of prediction set clusters that do not contain training set samples, ClusterCall uses the largest k (for clustering in k groups) consistent with the cluster range, separation, and compactness criteria mentioned in the previous section.

Using the AxS, WxL, and RGxP families as training data, genotypes were predicted for 5218 markers in the SolCAP diversity panel with ≥0.95 concordance and ≤5% missing data. From this set, 1858 markers were called with ≤5% missing data using the visually determined cluster boundaries from Hirsch et al. (2013). The genotype calls by the two methods were at least 95% identical at all but 9 markers, demonstrating the high quality of the automated calls made by ClusterCall.

As an additional test, a genome-wide association study (GWAS) was performed with the ClusterCall genotypes (Table S2, Supplementary Material 1) and compared against the results of Rosyara et al. (2016), who used the genotype calls from Hirsch et al. (2013). Both studies detected co-localized QTL for tuber shape and eye depth on chromosome 10 (van Eck et al. 1994; Li et al. 2005), but only with the ClusterCall genotypes was the well-known maturity QTL on chromosome 5 detected (Fig. S6, Supplementary Material 1; Bradshaw et al. 2008; Uitdewilligen et al. 2013). The most significant SNP for vine maturity at 120 days was solcap_snp_c1_14802 ($-\log_{10}p = 6.1$), which at 5,051,766 bp is close to the position (4,538,880...4,541,736) of the *StCDF1* gene (PGSC0003DMG400018408) implicated as the causal locus by Kloosterman et al. (2013). The solcap_snp_c1_14802 association was not detected by Rosyara et al. (2016) because the SNP was not called by Hirsch et al. (2013).

Discussion

For both technical and computational reasons, previous research has often failed to take advantage of allele dosage information in polyploid studies. Wu et al. (1992) proposed that polyploid linkage maps could be created using only

Table 2 Number of allele dosage levels in the training set vs. in the SolCAP diversity panel (prediction set), for markers with at least 0.95 concordance

Number of dosage levels in the training set	Number of dosage levels in the prediction set					
	0	1	2	3	4	5
2	83	1	284			
3	216	0	148	1180		
4	124	0	1	130	861	637
5	80	0	0	17	296	1671

The count in each cell of the two-way table is the number of markers

markers that are simplex in one parent and nulliplex in the other. With random chromosome segregation, these "single-dose" markers segregate 1:1:0:0:0, and the maximum likelihood estimator of recombination frequency between markers in coupling phase is independent of ploidy level, which means diploid mapping software can be used (Li et al. 1998, 2014b; Brouwer and Osborn 1999). In population genetic studies, all markers are typically used but the three heterozygous genotypes may not be differentiated to facilitate analysis with diploid software (Simko et al. 2006; Li et al. 2014a).

Because of the ability to score allele dosage with SNP arrays, new software that utilizes this information is being developed. Within the past few years, tools for tetraploid linkage mapping (Hackett et al. 2013), haplotype reconstruction (Zheng et al. 2016), and GWAS (Rosyara et al. 2016) have been published. These analysis tools depend on accurate allele dosage information, for which ClusterCall represents a significant advance.

A key feature of ClusterCall is the use of F1 populations to calibrate the theta-to-genotype relationship for each marker. The idea of using F1 populations to calibrate tetraploid genotype calls has been described before, notably by Hirsch et al. (2013) and Vos et al. (2015) in potato. Compared with these earlier studies, ClusterCall automates and streamlines this process and, by using the concordance metric, provides a quantitative basis for eliminating unreliable markers. The high quality of the ClusterCall output does not come at the expense of speed. To assign genotypes for the Atlantic × Superior family, ClusterCall required 9.5 min on a 3.2 GHz Intel Core i5 processor, compared with 7.3 h for fitTetra.

Whereas fitTetra requires no training population, the number and quality of the genotype calls made by ClusterCall depend on the number and size of the F1 populations. Larger populations improve the reliability of using the chi-squared segregation test to infer genotypes, and more populations increase the probability of observing all five dosages (0–4) in the training set. To minimize ascertainment bias, the parents of the F1 populations should be representative of the germplasm in the prediction set. A two-way table of the number of dosage levels in the training vs. prediction sets can be used to assess how many markers are potentially missed due to ascertainment bias (Table 2). The first column of Table 2 shows there were several hundred markers with high concordance across the training set but no call in the diversity panel. A hybrid approach can be used to score these markers, by combining training set calls from ClusterCall with prediction set calls from fitTetra (or some other method). Markers could be filtered for concordance across the training and prediction sets and then visually inspected using a theta band diagram.

ClusterCall was written exclusively for autotetraploids. Although in principle our pipeline of hierarchical clustering and model selection based on expected segregation ratios could be adapted for higher ploidy levels, this may be impractical even for autohexaploids. With tetraploids, a certain amount of spread for each genotype cluster can be tolerated because only five clusters need to be accommodated in the [0, 1] interval, but more stringent tolerances on the clusters would be needed to fit up to seven clusters for a hexaploid. Furthermore, it will prove difficult to use the goodness-of-fit to segregation ratios to determine the number of true clusters and genotype-to-cluster assignment, as a small shift in the allele dosage of the parents produces similar ratios that only large populations could differentiate. For ploidy levels greater than four, mixed ploidy populations, and segmental allopolyploids, we recommend the normal mixture model or probabilistic graphical model of Serang et al. (2012).

The focus of this study was the 8303 SNP array for potato, which is transforming potato breeding and genetics research (Douches et al. 2014). Vos et al. (2015) designed a 20K array for potato, but it is no longer available. In 2015, a 12K array was developed for North American potato by combining reliable markers from the original 8303 SNP array with additional SNPs from Hamilton et al. (2011). This array has been used to genotype hundreds of elite lines, but thus far we have been limited to using the SNPs derived from the 8303 array. Recently, several F1 populations were genotyped with the 12K array for the purpose of QTL mapping. With ClusterCall, we can now use these populations to realize the full potential of the array for GWAS and genome-wide prediction.

Author contribution statement Designed the research: JBE. Contributed marker data: JPP, JJC, DSD, PCB, RGN. Developed the software, analyzed the data, and prepared the manuscript: CSC, JBE.

Acknowledgements Financial support was provided by the National Institute of Food and Agriculture, U.S. Department of Agriculture, Award Number 2014-67013-22418 and Hatch Project Number 1002731.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Bourke PM, Voorrips RE, Visser RGF, Maliepaard C (2015) The double reduction landscape in tetraploid potato as revealed by a high-density linkage map. Genetics 201:853–863. doi:10.1534/genetics.115.181008

Bradshaw JE, Hackett CA, Pande B, Waugh R, Bryan GJ (2008) QTL mapping of yield, agronomic and quality traits in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*). Theor Appl Genet 116:193–211

Brouwer DJ, Osborn TC (1999) A molecular marker linkage map of tetraploid alfalfa (*Medicago sativa* L.). Theor Appl Genet 99:1194–1200. doi:10.1007/s001220051324

Comai L (2005) The advantages and disadvantages of being polyploid. Nat Rev Genet 6:836–846. doi:10.1038/nrg1711

Douches D, Hirsch CN, Manrique-Carpintero NC, Massa AN, Coombs J, Hardigan M, Bisognin D, De Jong W, Buell CR (2014) The contribution of the Solanaceae coordinated agricultural project to potato breeding. Potato Res 57(3–4):215–224. doi:10.1007/s11540-014-9267-z

Endelman J, Jansky S (2016) Genetic mapping with an inbred line derived F2 population in potato. Theor Appl Genet 129:935–943. doi:10.1007/s00122-016-2673-7

Felcher KJ, Coombs JJ, Massa AN, Hansey CN, Hamilton JP, Veilleux RE, Buell CB, Douches DS (2012) Integration of two diploid potato linkage maps with the potato genome sequence. PLoS One 7(4):e36347. doi:10.1371/journal.pone.0036347

Gallais A (2003) Quantitative genetics and breeding methods in autopolyploid plants. INRA, Paris

Hackett CA, McLean K, Bryan GJ (2013) Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. PLoS One 8(5):e63939. doi:10.1371/journal.pone.0063939

Hamilton JP, Hansey CN, Whitty BR, Stoffel K, Massa AN, Van Deynze A, De Jong WS, Douches DS, Buell CR (2011) Single nucleotide polymorphism discovery in elite North American potato germplasm. BMC Genom 12:302. doi:10.1186/1471-2164-12-302

Hirsch CN, Hirsch CD, Felcher K, Coombs J, Zarka D, Van Deynze A, De Jong W, Veilleux RE, Jansky S, Bethke P, Douches DS, Buell CR (2013) Retrospective view of North American potato (*Solanum tuberosum L.*) breeding in the 20th and 21st centuries. G3 3:1003–1013. doi:10.1534/g3.113.005595

Jones GH, Khazanehdari KA, Ford-Lloyd BV (1996) Meiosis in the leek (*Allium porrum* L.) revisited. II. Metaphase I observations. Heredity 76:186–191

Kloosterman B, Abelenda JA, Carretero Gomez MM, Oortwijn M, de Boer JM, Kowitwanich K, Horvath BM, van Eck HJ, Smaczniak C, Prat S, Visser RGF, Bachem CWB (2013) Naturally occurring allele diversity allows potato cultivation in northern latitudes. Nature 495:246–250. doi:10.1038/nature11912

Koning-Boucoiran CFS, Esselink GD, Vukosavljev M, van't Westende WPC, Gitonga VW, Krens FA, Voorrips RE, van de Weg WE, Schulz D, Debener T, Maliepaard C, Arens P, Smulders MJM (2015) Using RNA-seq to assemble a rose transcriptom with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rose* L.). Front Plant Sci 6:249. doi:10.3389/fpls.2015.00249

Krebs SL, Hancock JF (1989) Tetrasomic inheritance of isoenzyme markers in the highbush blueberry, *Vaccinium corymbosum* L. Heredity 63:11–18. doi:10.1038/hdy.1989.70

Leal-Bertioli S, Shirasawa K, Abernathy B, Moretzsohn M, Chavarro C, Clevenger J, Ozias-Akins P, Jackson S, Bertioli D (2015) Tetrasomic recombination is surprisingly frequent in allotetraploid Arachis. Genetics 199:1093–1105. doi:10.1534/genetics.115.174607

Leitch IJ, Bennett MD (1997) Polyploidy in angiosperms. Trends Plant Sci 2:470–476. doi:10.1016/S1360-1385(97)01154-0

Li X, van Eck HJ, Rouppe van der Voort JNAM, Huigen DJ, Stam P, Jacobsen E (1998) Autotetraploids and genetic mapping using common AFLP markers: the R2 allele conferring resistance to *Phytophthora intestans* mapped on potato chromosome 4. Theor Appl Genet 96:1121–1128. doi:10.1007/s001220050847

Li X, De Jong H, De Jong DM, De Jong WS (2005) Inheritance and genetic mapping of tuber eye depth in cultivated diploid potatoes. Theor Appl Genet 110:1068–1073. doi:10.1007/s00122-005-1927-6

Li X, Han Y, Wei Y, Acharya A, Farmer AD, Ho J, Monteros MJ, Brummer EC (2014a) Development of an alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. Plos One. doi:10.1371/journal.pone.0084329

Li X, Wei Y, Acharya A, Jiang Q, Kang J, Brummer EC (2014b) 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 4:1971–1979. doi:10.1534/g3.114.012245

Luo ZW, Hackett CA, Bradshaw JE, McNicol JW, Milbourne D (2001) Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics 157:1369–1385

Mather K (1936) Segregation and linkage in autotetraploids. J Genet 32(2):287–314. doi:10.1007/BF02982683

Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195. doi:10.1038/nature10158

Quiros CF (1982) Tetrasomic segregation for multiple alleles in alfalfa. Genetics 101:117–127

R Development Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Renny-Byfield S, Wendel FJ (2014) Doubling down on genomes: polyploidy and crop plants. Am J Bot 101:1711–1725. doi:10.3732/ajb.1400119

Rosyara UR, De Jong WS, Douches DS, Endelman JB (2016) Software for genome-wide association studies in autopolyploids and its application to potato. Plant Genome 9. doi:10.3835/plantgenome2015.08.0073

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS One 7:e30906

Sharma SK, Bolser D, de Boer J, Sønderkær M, Amoros W, Carboni MF, D'Ambrosio JM, de la Cruz G, Di Genova A, Douches DS, Eguiluz M, Guo X, Guzman F, Hackett CA, Hamilton JP, Li G,

Li Y, Lozano R, Maass A, Marshall D, Martinez D, McLean K, Mejía N, Milne L, Munive S, Nagy I, Ponce O, Ramirez M, Simon R, Thomson SJ, Torres Y, Waugh R, Zhang Z, Huang S, Visser RGF, Bachem CWB, Sagredo B, Feingold SE, Orjeda G, Veilleux RE, Bonierbale M, Jacobs JME, Milbourne D, Martin DMA, Bryan GJ (2013) Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. G3 3:2031–2047. doi:10.1534/g3.113.007153

Simko I, Haynes KG, Jones RW (2006) Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173:2237–2245. doi:10.1534/genetics.106.060905

Stebbins GL (1950) Variation and evolution in plants. Columbia University Press, New York

Uitdewilligen JGAML, Wolters AMA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One 8:e62355

van Eck HJ, Jacobs JM, Stam P, Ton J, Stiekema WJ, Jacobsen E (1994) Multiple alleles for tuber shape in diploid potato detected by qualitative and quantitative genetic analysis using RFLPs. Genetics 137:303–309

Venables WN, Ripley BD (2002) Modern Applied Statistics with S. 4th edition. Springer, New York.

Voorrips RE, Gort G, Vosman B (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinform 12:172. doi:10.1186/1471-2105-12-172

Vos PG, Uitdewilligen JGAML, Voorrips RE, Visser RGF, van Eck HJ (2015) Development and analysis of a 20 K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. Theor Appl Genet 128:2387–2401. doi:10.1007/s00122-015-2593-y

Wu KK, Burnquist W, Sorrells ME, Tew TL, Moore PH, Tanksley SD (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor Appl Genet 83:294–300. doi:10.1007/BF00224274

Zheng C, Voorrips RE, Jansen J, Hackett CA, Ho J, Bink MCAM (2016) Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. Genetics 203:119–131. doi:10.1534/genetics.115.185579

Zorrilla C, Navarro F, Vega S, Bamberg J, Palta J (2014) Identification and selection for tuber calcium, internal quality and pitted scab in segregating 'Atlantic' × 'Superior' reciprocal tetraploid populations. Am J Potato Res 91:673–687. doi:10.1007/s12230-014-9399-3