Micro-haplotyping in polyploids using massively parallel amplicon sequencing

Geert van Geest (✉ geert.vangeest@gmail.com)

Wageningen Universiteit en Research <https://orcid.org/0000-0002-1561-078X>

Johan Willemsen

Fresh Forward

Krishna Reddy Gujjula

Thermo Fisher Scientific

Prasad Siddavatam

Thermo Fisher Scientific

Christopher Roy Willis

Thermo Fisher Scientific

Ravindra Ramadhar

Thermo Fisher Scientific

Roeland E Voorrips

Wageningen Universiteit en Research

Paul Arens

Wageningen Universiteit en Research

Research

Keywords: Haplotyping, polyploids, potato, parallel amplicon sequencing

Posted Date: June 17th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-35215/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Haplotyping in polyploids is a very relevant but challenging task. Most methods based on short-read technologies require haplotype assembly, which is error-prone and computationally intensive. However, most (auto)polyploids have high SNP densities. Therefore, haplotyping could already be done in regions of a size that can be entirely covered by reads generated by current short-read sequencing techniques. Here, we used a method that enables massively parallel amplicon sequencing in order to generate such micro-haplotypes in autotetraploid potato and autohexaploid chrysanthemum.

Results

For potato, we generated 79 million reads to sequence 412 regions in 96 samples. The average region size was 154 base-pairs. On average, we found 2.4 haplotypes per locus per individual. For chrysanthemum, we generated a similar amount of reads to sequence 940 regions in 92 samples. The average region size here was 171 base-pairs. On average, 2.7 haplotypes per locus per individual were found. Concordance with dosage data based on SNP-arrays was up to 96.8% in potato and 94.2% in chrysanthemum.

Conclusions

Interpretation of genotyping data generated with next-generation sequencing can be challenging in polyploids, because estimation of allele dosages is often a requisite. Taking advantage of high SNP densities in many autopolyploids we show that massively parallel amplicon sequencing can generate high-quality data with high information content. This has a broad range of applications, including linkage mapping, diversity studies, and genomic selection.

Background

Innovation in DNA sequencing techniques have had a large influence on our understanding of genetics. Due to such innovations, it has become affordable to genotype a large number of samples for a large number of variants for many different species, including polyploids. However, genotyping polyploids is more challenging compared to diploids. This is because one individual can carry more than two alleles at a locus and these alleles can be present in different dosages. The information about allele dosage is important for genetic studies, because a wide range of allelic combinations can occur at a locus that can result in different phenotypes [1–4]

Similar to diploids, SNP genotyping has become state-of-art in polyploids. High-throughput genotyping often is performed by the use of conventional SNP arrays [5–10]. By making use of allelic dosage

estimation in single SNPs, significant progress has been made in GWAS [11, 12] and in linkage and QTL mapping [1, 13–15]. However, a limitation of such an approach is that a single bi-allelic SNP can only separate two alleles, whereas most polyploid crops are outbreeding and can carry multiple alleles at the same locus. Therefore, assays that tag multiple alleles enable better capture of the genetic diversity of a locus.

The identification of haplotypes enables the characterization of multiple alleles within a locus. To obtain haplotypes, sequencing data is often phased using haplotype assembly [16–18]. In diploids, this is quite straightforward because one haplotype defines the phasing of the other. However, in polyploids, this is much more complex. For example, defining one haplotype in a tetraploid leaves three other haplotypes to phase. This makes haplotype assembly much more computationally intensive.

Direct haplotyping within sequencing reads alleviates the difficulties associated with haplotype assembly. The length of the regions needed for direct haplotyping depends on the SNP density in the regions of interest. In general, SNP densities are high in polyploids. Potato for example has an estimated population-wide SNP density in genic regions of 1 in 24 base pairs, whereas each individual potato variety contains a single variant each 42 bp [19]. This means that direct multi-SNP haplotyping can be done already in relatively short regions of 150 to 200 base pairs, which is in the range of most short-read sequencing technologies. This concept is referred to as micro-haplotyping [20–22], and has great potential in polyploid organisms due to their high SNP densities.

With common PCR, short regions can be easily amplified in a sample and the resulting amplicons can be sequenced. With massively parallel PCR this can be done for a high number of regions in a single reaction. The AgriSeq technology [23] is developed to automate this for a large number of regions in a large number of genotypes, enabling high-throughput micro-haplotyping. In this study we evaluate this methodology for its suitability for haplotyping in polyploids. We applied this technology in tetraploid potato (*Solanum tuberosum*) and hexaploid chrysanthemum (*Chrysanthemum x morifolium*) to generate microhaplotypes. We used loci and individuals that previously have been genotyped with SNP arrays [1, 24], enabling the comparison of SNP dosage estimation from both methods.

Results

In order to investigate the potential of massively parallel amplicon sequencing, we used AgriSeq to sequence amplicons in 96 samples for potato and 92 samples for chrysanthemum. Samples from both species including the parents, members of an F1 population, and a small set of cultivars (Additional file 1 & 2).

Sequencing results

For potato, 79 million reads were generated with a mean read length of 152 bp which approached the mean target region size of 154 bp (Table 1). With these reads, we identified 2467 SNPs in the target

Loading [MathJax]/jax/output/CommonHTML/jax.js y spanned all identified SNPs within a region and were

considered useful for direct micro-haplotyping. These reads mapped to 409 out of the 412 (99.3%) regions selected on the potato genome [25]. This resulted in an average read depth of 1764 per sample on the target regions. The average number of SNPs on the target regions was 6.0, resulting in an average SNP density of 1 in 26 base pairs.

For chrysanthemum, 79 million reads were generated with a mean read length of 176 bp, which exceeded the mean target region size of 171 bp (Table 1). With these reads, we identified 6600 SNPs in the target regions. In total, 70.7 million (89.4%) entirely spanned all called SNPs within a target region. These reads mapped to 931 out of the 940 (99.0%) target regions selected on the *C. seticuspe* reference genome [26]. This resulted in an average read depth per sample on the target regions of 822. The average number of SNPs in the target regions was 7.1, resulting in an average SNP density of 1 in 24 base pairs.

Table 1
Summary of sequencing characteristics and data quality

	Potato	Chrysanthemum
Number of reads	79×10^6	79×10^6
Reads mapping to target*	71.2×10^6	70.7×10^6
Number of target amplicons	412	940
Average region length	154	171
Genotyped target amplicons	409	931
Average read depth	1764	822
Average number of SNPs per locus	6.0 (1/26)	7.1 (1/24)
Average haplotypes per locus per individual	2.4	2.7
*Reads that mapped and fully spanned the SNPs in the target regions.		

Haplotyping

Aligned sequence reads were filtered for read alignments that spanned all the SNPs within a target region. Therefore, we could directly construct and count haplotypes from the sequence alignment files. The haplotype distributions per individual per locus had maxima at the expected allele frequencies of 0, 1/4, 1/2 and 3/4 and 1 for potato and 0, 1/6, 1/3, 1/2, 2/3, 5/6 and 1 for chrysanthemum (Fig. 1). Dosages of haplotypes were based on haplotype counts using a maximum likelihood method. With this method, any regions and haplotypes that were unlikely to fit the expected dosage distributions within an individual were filtered out. Since most haplotype frequencies approached the expected distributions, estimating haplotype dosage was straightforward in most regions (Fig. 2; Fig. 3). Haplotype identification and dosage scoring resulted in 5.4 and 7.1 haplotypes per target region on average for potato and

chrysanthemum, respectively. On average, each individual carried 2.4 haplotypes per locus for potato and 2.7 haplotypes per locus for chrysanthemum.

Comparison to array data

Because genomic regions were selected based on regions around previously characterized SNP markers, the SNP dosages of these particular SNPs could be compared between the two genotyping methods. For potato we could compare 346 SNPs. Not all used individuals were previously genotyped with the array resulting in 63 common individuals. For chrysanthemum, we could compare 82 genotypes for 869 SNPs. Concordance was calculated as the number of equal non-missing dosage calls between the two methods by dividing the total number of non-missing calls. In general, concordance was over 90% (Table 3) for both crops. If the dosage did not match with the array, it differed usually by single dosage (Fig. 4).

The used datasets have high average coverages per locus (1764 for potato and 822 for chrysanthemum), and this might not be cost-effective in future applications. In order to investigate whether lower coverages would result in similar output, we sub-sampled the datasets to have an average coverage per locus of 100, 200 and 400 (Table 3). This did not decrease concordance with the SNP array. For potato, lower read depths actually increased the concordance slightly, as low-frequency spurious haplotypes were not detected anymore.

Table 3

Concordance between SNP dosage estimated using AgriSeq and genotyping array. Concordance is defined as the percentage of equal dosages as part of the number of non-missing dosage estimations.

	Potato		Chrysanthemum	
	All	Heterozygote only**	All	Heterozygote only**
All data	95.0%	95.6%	92.2%	90.4%
Filtered on likelihood ratio*	96.0%	96.8%	94.2%	93.2%
Mean coverage 400	95.0%	95.7%	92.2%	90.3%
Mean coverage 200	95.2%	95.8%	92.2%	90.3%
Mean coverage 100	95.8%	96.1%	92.0%	90.3%
*Data filtered with a threshold of likelihood ratio of multi-allelic dosage estimation. This threshold was chosen based on filtering resulting in < 10% missing values. The threshold was greater than 80 for potato and 10 for chrysanthemum.				
** Concordance of calls that were called as heterozygous by the AgriSeq method.				

Discussion

In this study, we used massively parallel amplicon sequencing in order to identify and genotype micro-haplotypes. In regions around previously characterized SNPs, micro-haplotypes can be used as multi-allelic

genotypic markers. The use of multi-allelic markers instead of bi-allelic markers can significantly increase the amount of genetic information in diploids [20, 27], and for polyploids in particular [28]. The reason is that polyploid species can carry multiple different alleles at the same locus, which are difficult to capture by bi-allelic markers. Multi-allelic markers can, for example, improve the construction of genetic linkage maps or detection of alleles associated with traits of interest using genome-wide association studies (GWAS). In this study, we show how a simple and straightforward short-read-sequencing-based technology can provide multi-allelic markers based on micro-haplotypes.

We employed amplicon-based sequencing (Agrisecq) in two polyploid crops: tetraploid potato (*Solanum tuberosum*) and hexaploid chrysanthemum (*Chrysanthemum x morifolium*). The design of the PCR target regions involved a selection from high-quality SNPs that had previously been used to generate genetic maps of both potato [24] and chrysanthemum [1] by SNP arrays. In general, the conversion rates between SNPs used in these arrays and the design considered here are high, despite the fact that the sequences used for marker design are different between the SNP arrays and amplicon sequencing. In a SNP array, a sequence directly flanking the SNP is used. For amplicon sequencing, primer annealing sites at both sides of the target amplicon are used. These locate at larger distances from the original focal SNP. Nevertheless, conversion rates between SNP locus and primer design were over 99%.

Haplotypes were directly identified from sequence reads that aligned to the reference. This was straightforward because all used reads entirely spanned the SNPs within a target region. Therefore, there was no need for haplotype assembly of overlapping reads, which is often needed in longer genomic regions [16–18]. This strongly reduced the sensitivity to errors and computational time.

After identification, we could estimate the allelic composition and dosage within an individual based on the haplotype counts. The individual SNP dosages that were estimated based on the haplotype dosages as estimated from the maximum likelihood approach showed a high concordance with previous results. In both crops, concordance of the dosages estimated from amplicon sequencing with the SNP array data was high: up to 96.8% for potato and 94.2% for chrysanthemum. The used maximum likelihood estimation method enabled us not only to estimate the most likely allele dosage, but also a confidence measure of the estimation in the form of a log likelihood. Filtering based on the log likelihood slightly increased the concordance with genotyping arrays, suggesting that low log likelihood estimates can be used to filter out erroneous dosage calls. In addition, in the cases in which there was a discordance in allelic dosages, in most cases this was a difference of one, which will have lower impact on downstream analyses compared to larger dosage differences. The relatively marginal discordance between the array data and current sequencing data therefore shows that the error rates of haplotype dosage estimations was low and are at least comparable to state-of-art SNP genotyping methods in polyploids.

When comparing the two crops, it appears that likelihood ratios and concordance with the SNP array were lower in chrysanthemum compared to potato. One important cause is the ploidy level. The expected read frequencies per dosage are closer together for hexaploid chrysanthemum compared to tetraploid potato. A small allele bias would therefore more likely result in errors in haplotype dosage estimation in a

hexaploid compared to a tetraploid. Therefore, while employing techniques as AgriSeq for genotyping, it should be considered that uncertainty caused by allelic bias or low read depth is higher in higher ploidy levels.

In order to make an application for AgriSeq more cost effective, one could consider to use lower average read depth compared to this study (1764 for potato and 822 for chrysanthemum). For the described dataset, lowering the average read-depth by sub-sampling did not decrease the concordance with the array, and in some cases it even slightly increased the number of high-quality regions because some low-frequency spurious haplotypes did not occur. Even coverage distributions are key for the success of using a lower number of sequence-reads per sample.

We think that the Agriseq method can be applied for analyses in polyploids that require high-quality marker information, like linkage mapping. It could be particularly interesting for linkage map integration, since this can be a serious challenge in polyploids [13]. A substantial proportion of the developed markers are multi-allelic, so they anchor loci relative to each other between homologous linkage groups. The accuracy of integration of linkage groups will therefore most likely strongly increase. Likewise these panels of reasonable high numbers of amplicons could potentially be used in applications such as genomic prediction, genetic diversity analyses and genotype validation of DNA samples.

Conclusions

Methods for genotyping by sequencing have had and still have a large impact on genetic research. This is because they have enabled the generation of large amounts of data at low costs per data point. A challenge is coping with the balance between costs and data quality. This is certainly the case in polyploids, in which for most applications, dosage scoring is a requisite. In this paper we have described a genotyping by sequencing method that can deliver high-quality data with high information content at relatively low costs. The method is particularly useful for polyploids, because haplotyping is straightforward in small regions with high SNP density. Future applications of the method are broad, ranging from diversity studies to linkage mapping and genomic selection.

Materials And Methods

Plant Material

For chrysanthemum and potato, we genotyped 91 and 96 samples, respectively. For potato, these included 75 members of the mapping population resulting from the cross between 'Altus' and 'Colomba' [24], three replicates of both parents and 15 cultivars (Additional file 1). For chrysanthemum, these included 74 members of the population resulting from the cross between DB36451 and DB39287 [1, 10], two replicates of the maternal parent, four replicates of the paternal parent and twelve cultivars (Additional file 2).

For chrysanthemum, 1152 SNPs were selected that properly mapped according to Van Geest et al. (2017a). The regions around selected SNPs were converted into regions with an average length of 171 base pairs. Primer design was based on the *C. seticuspe* genome sequence [26]. In total 940 SNP regions passed quality control. For potato 569 SNPs were selected that were originally part of the SOL-STW 20 k SNP array [6]. Like with chrysanthemum these markers were converted into amplicon regions of 154 bp. In total 412 regions passed quality control. Primer design was based on the potato DM genome sequence V4.03 [25].

Sequencing

Samples were prepared for sequencing using the AgriSeq™ HTS Library Kit (A34143-Life Technologies). In short, DNA concentrations were normalized to 3.3 ng/μL for a total of 10 ng DNA per 10 μL reaction. Normalized DNA was combined with the Ion AgriSeq primer panel and AgriSeq amplification master mix. For amplification of genomic targets, the following thermocycling programs were used; 99°C for 2 min, then 15 cycles of 99°C for 15 s and 60°C for 4 min for chrysanthemum and 99°C for 2 min, then 16 cycles of 99°C for 15 s and 60°C for 4 min for potato. Amplicons were then prepared for ligation with pre-ligation enzyme digestion at 50°C for 10 min, 55°C for 10 min, and 60°C for 20 min. IonCode™ Barcode Adapters 1-384 Kit (A29751-Life Technologies) were ligated to the digested products with barcoding enzyme and buffer. Labeled amplicons were then pooled, cleaned up, amplified and normalized. Following library preparation, libraries were loaded onto an Ion 540™ sequencing Chip Kit (A27765) via the Ion 540™ Kit-Chef (A30011-Life Technologies) and Ion Chef. Sequencing was then performed on the Ion S5 system (Thermo Fisher, Inc. Waltham, MA).

Bioinformatics and data analysis

SNPs were called using freebayes [29] and filtered using vcfilter for SNPs with a quality greater than 100 and at least one heterozygous individual. Sequence alignment files were filtered for reads entirely covering the first and last SNP in the target regions using bedtools [30]. The variant call format file resulting from the SNP call was used to identify SNP alleles in each read using pysam (<https://github.com/pysam-developers/pysam>).

In order to filter for spurious haplotypes that resulted from off target amplification, we filtered for haplotypes with a high frequency of occurrence in the individuals (> 80%) but low average frequency within individuals (less than 1/8 for potato and less than 1/12 for chrysanthemum). Filtering steps and data visualisation were done using R [31].

Likelihood-based test for genotyping

In this study we estimated haplotype dosages by taking into account all possible haplotype conformations of the detected haplotypes within a genotype. Haplotype dosages were estimated for each individual in each region using a likelihood-based method in line with well-established techniques (Li 2011; Gerard et al. 2018). The possible dosage of each allele was calculated based on ploidy level ($k = 4$,

$k = 6$) and the number of haplotypes found in a single individual. If we consider a region with e.g. three haplotypes (A, B, C) we easily inferred the dosages of homozygous classes to be AAAA, BBBB or CCCC.

Exclusion of erroneous haplotypes

Before genotyping each haplotype was tested if it was likely to originate from a sequencing error. This avoids situations where only a minor proportion of the reads leads to the incorrect genotype assignment. These reads are likely to represent sequencing error or contamination from other samples.

With the assumption of independence between sequencing read observations we can estimate the probability of observing k erroneous reads at n depth, given a sequencing error rate ε using a binomial model.

$$P(X = k) = \binom{n}{k} (1 - \varepsilon)^k (\varepsilon)^{n-k}$$

Where k represents the read counts of the most frequent haplotype in an individual, and n represents the total haplotype count/read depth at that position. Hence, we can estimate the probability of observing k reads given the total read depth n .

For instance, with sequence error rate at 2%, two errors at 100 depth occur in 32% of the cases. Likewise 10 errors at read depth of 1000 occur in 98% of the cases. This allows to separate the erroneous haplotypes from real haplotypes at a userspecified probability threshold. In this study we used a threshold of 0.9.

Genotype likelihoods

Subsequently, we calculated genotype likelihoods. In case of heterozygosity we have multiple genotype classes: In this example we have three genotype possibilities (G_i) in a tetraploid: AABC, ABBC and ABCC, with corresponding expected frequencies per haplotype (Table 1). As the individual haplotype counts are known, we can simply estimate the posterior probability of each genotypes by Bayes rule as:

$$Pr(R) = \frac{Pr(G_i) Pr(R|G_i)}{Pr(R)}$$

From this follows that the likelihood of observing a genotype is equal to:

$$Pr(R) = \frac{Pr(G_i) Pr(R|G_i)}{\sum_{n=1}^N Pr(R|G_n)}$$

Where R is a vector of read counts of all haplotypes. For bi-allelic variants we assume read counts to be Loading [MathJax]/jax/output/CommonHTML/jax.js

allelic variants to be to be multinomially distributed. Hence we

can compute the conditional probabilities of each $Pr(R)$ by calculating the multinomial probability of each genotype.

For the homozygous classes (in example AAAA we estimate the probability of observing a homozygous classes with given a sequencing error rate of ε :

$$Pr(R) = \binom{n}{A} (1 - \varepsilon)^A * (\varepsilon)^\alpha$$

where A represents the count of the major allele, and α represents the counts of the minor allele. This avoids that genotypes in which only a few alternative reads are present are seen as heterozygous. In any case the solution with highest likelihood is selected. As confidence measurement we use the log likelihood ratio between the best and second best solution.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets on potato generated and/or analysed during the current study are available in the Sequence Read Archive repository under accession number PRJNA635347, and can be accessed through:
<https://www.ncbi.nlm.nih.gov/sra/PRJNA635347>

The datasets on chrysanthemum used and/or analysed during the current study are available from the corresponding author on reasonable request.

Python scripts written for haplotype calling and dosage estimation can be found at
<https://github.com/GeertvanGeest/dirHap>.

- Project name: dirHap
- Project home page: <https://github.com/GeertvanGeest/dirHap>.
- Archived version: <https://github.com/GeertvanGeest/dirHap>.
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: Python 3.7 or higher

Loading [MathJax]/jax/output/CommonHTML/jax.js

- License: e.g. GNU GPL

Competing interests

The authors declare that they have no competing interests

Funding

Funding for this research was provided through the TKI project “Novel genetic and genomic tools for polyploid crops”, project number KV 1605-020.

Authors' contributions

GG executed the analysis, wrote the manuscript, and selected the target regions for chrysanthemum. JW selected the target regions from potato and developed the MLE method for dosage scoring. KRG, PS, CW and RR designed the panels and generated sequencing data. RR, RV and PA participated in coordination. All authors participated in drafting the manuscript.

Acknowledgements

Not applicable

References

1. Van Geest G, Bourke PM, Voorrips RE, Marasek-Ciolakowska A, Liao Y, Post A, et al. An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. *Theor Appl Genet.* 2017;130:2527–41.
2. Klaassen MT, Bourke PM, Maliepaard C, Trindade LM. Multi-allelic QTL analysis of protein content in a bi-parental population of cultivated tetraploid potato. *Euphytica Springer Netherlands.* 2019;215:1–18.
3. Kloosterman B, Abelenda JA, del Mar Carretero Gomez M, Oortwijn M, de Boer JM, Kowitwanich K. Naturally occurring allele diversity allows potato cultivation in northern latitudes. *Nature.* 2013;495:246–50.
4. Manrique-Carpintero NC, Coombs JJ, Pham GM, Laimbeer FPE, Braz GT, Jiang J, et al. Genome reduction in tetraploid potato reveals genetic load, haplotype variation, and loci associated with agronomic traits. *Front Plant Sci.* 2018;9:1–15.
5. Koning-Boucoiran CFS, Esselink GD, Vukosavljev M, van 't Westende WPC, Gitonga VW, Krens FA, et al. Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed

- genes and to develop the WagRhSNP 68 k Axiom SNP array for rose (*Rosa* L.). *Front. Plant Sci.* 2015;6:249.
6. Vos PG, Uitdewilligen JGAML, Voorrips RE, Visser RGF, van Eck HJ. Development and analysis of a 20K SNP array for potato (*Solanum tuberosum*): an insight into the breeding history. *Theor. Appl. Genet.* Springer Berlin Heidelberg; 2015;128:2387–401.
  7. You Q, Yang X, Peng Z, Islam MS, Sood S, Luo Z, et al. Development of an Axiom Sugarcane100K SNP array for high-resolution genetic map construction and QTL identification. *Theor. Appl. Genet.* Springer Berlin Heidelberg; 2019.
  8. Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, et al. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol J.* 2014;12:787–96.
  9. Bassil NV, Davis TM, Zhang H, Ficklin S, Mittmann M, Webster T, et al. Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria × ananassa*. *BMC Genom.* 2015;16:1–30.
  10. Van Geest G, Voorrips RE, Esselink D, Post A, Visser RG, Arens P. Conclusive evidence for hexasomic inheritance in chrysanthemum based on analysis of a 183 k SNP array. *BMC Genomics BMC Genomics.* 2017;18:585.
  11. Rosyara UR, De Jong WS, Douches DS, Endelman JB. Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *Plant Genome.* 2016;9:1–10.
  12. Ferrão LFV, Benevenuto J, Oliveira I, de B, Cellon, Olmstead C, Kirst J. M, et al. Insights Into the Genetic Basis of Blueberry Fruit-Related Traits Using Diploid and Polyploid Models in a GWAS Context. *Front. Ecol. Evol.* 2018;6.
  13. Bourke PM, Voorrips RE, Kranenburg T, Jansen J, Visser RG, Maliepaard C. Integrating haplotype-specific linkage maps in autotetraploid potato using SNP markers. *Theor Appl Genet.* 2016;129:2211–26.
  14. Hackett C, Boskamp B, Vogogias A, Preedy K, Milne I. TetraploidSNPMap: Software suite for linkage analysis and QTL mapping for autotetraploid populations using SNP dosage data. *Biomathematics & Statistics Scotland*; 2016.
  15. 15.
  16. Motazed E, Maliepaard C, Finkers R, Visser R, De Ridder D. Family-based haplotype estimation and allele dosage correction for polyploids using short sequence reads. *Front Genet.* 2019;10:1–12.
  17. Aguiar D, Istrail S. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics.* 2013;29:352–60.
  18. Siragusa E, Haiminen N, Finkers R, Visser R, Parida L. Haplotype assembly of autotetraploid potato using integer linear programming. *Bioinformatics.* 2019;35:3279–86.
  19. Uitdewilligen JGAML, Wolters AMA, D’hoop BB, Borm TJA, Visser RGF, van Eck HJ. A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato.

20. Kidd KK, Pakstis AJ, Speed WC, Lagacé R, Chang J, Wootton S, et al. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Sci Int Genet*. 2014;12:215–24.
21. Oldoni F, Kidd KK, Podini D. Microhaplotypes in forensic genetics. *Forensic Sci Int Genet Elsevier*. 2019;38:54–69.
22. Turchi C, Pesaresi M, Tagliabracci A. A microhaplotypes panel for forensic genetics using massive parallel sequencing. *Forensic Sci Int Genet Suppl Ser Elsevier*. 2017;6:e117–8.
23. Willis RC, Swimley M, Siddavatam P, Conrad R, Scientific TF. Modular automation solution for genotyping by sequencing for animal breeding. *Proc. World Congr. Genet. Appl. to Livest. Prod.* 2016.
24. Bourke PM, Voorrips RE, Visser RGF, Maliepaard C. The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*. 2015;201:853–63.
25. The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature*. 2011;475:189–95.
26. Hirakawa H, Sumitomo K, Hisamatsu T, Nagano S, Shirasawa K, Higuchi Y, et al. De novo whole-genome assembly in *Chrysanthemum seticuspe*, a model species of *Chrysanthemums*, and its application to genetic and gene discovery analysis. *DNA Res*. 2019;26:195–203.
27. Maliepaard C, Alston FH, Van Arkel G, Brown LM, Chevreau E, Dunemann F, et al. Aligning male and female linkage maps of apple (*Malus pumila* Mill.) using multi-allelic markers. *Theor Appl Genet*. 1998;97:60–73.
28. van Dijk T, Noordijk Y, Dubos T, Bink MC, Meulenbroek BJ, Visser RG, et al. Microsatellite allele dose and configuration establishment (MADCE): an integrated approach for genetic studies in allopolyploids. *BMC Plant Biol BioMed Central Ltd*. 2012;12:25.
29. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv*. 2012.
30. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
31. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019.

Figures

Figure 1

Distribution of haplotype frequencies over all regions and samples of tetraploid potato (A) and hexaploid chrysanthemum (B). Red vertical lines mark the expected frequencies according to allelic dosage.

Figure 2

Example of haplotyping in a single target region in potato with the original distribution (A) and estimated discrete dosage distribution (B). Each color represents a different haplotype. Horizontal lines correspond the expected frequencies at $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{3}{4}$. Each column represents a sample of which P1 and P2 correspond to the replicates of the parents, AxC to the F1 population, and PANEL to the cultivar panel.

Figure 3

Example of haplotyping in chrysanthemum. See caption of Figure 2 for explanation.

Figure 4

Concordance between SNP dosage estimated using a SNP array and using AgriSeq in potato (A) and chrysanthemum (B). The numbers on the sides depict SNP dosages as estimated by the SNP array and AgriSeq. Non-analyzed dosages (due to too unexpected signal ratios or low read coverage) are depicted with NA.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile2.xlsx](#)
- [Additionalfile1.xlsx](#)