American Journal of Potato Research (2021) 98:328–332
https://doi.org/10.1007/s12230-021-09846-z

Assessing SNP Heterozygosity in Potato (*Solanum)* Species— Bias Due to Missing and Non-allelic Genotypes

John Bamberg[1] · Alfonso del Rio[2] · Lisbeth Louderback[3] · Bruce Pavlik[4]

Accepted: 31 August 2021 / Published online: 18 September 2021
© The Potato Association of America 2021

Abstract

Potato has about 100 related wild *Solanum* species growing naturally in the Americas. The US Potato Genebank aims to keep samples useful for research and breeding to improve the crop, often in the form of botanical seed families. A key component of genebank efficiency is assessing diversity within and among populations, and DNA marker sequence diversity is a powerful proxy for trait diversity. We previously reported on three factors which can cause under-estimation of heterozygosity: ascertainment, allele frequency, and ploidy bias. We here report, using GBS data for four diploid potato species, that average percent of apparent heterozygosity increases as data is more complete—the maximum difference was 2% heterozygotes when only a few individuals are called, to 36% when nearly all individuals were called. However, there was evidence that estimates of average heterozygosity based only on loci for which every individual has data can also be biased upward. Implausibly high levels of heterozygosity suggest non-segregating non-homologous SNPs, which occurred as 5–9% of all loci with complete data. We propose that best estimates of average heterozygosity in unselected seedlings should be based on loci with data for all samples after eliminating those loci that appear to be artificially fixed as heterozygous, which reduces observed heterozygote frequency by 16–26%. On that basis, the wild species examined have similar heterozygosity to the cultivated *phureja*.

Resumen

La papa tiene alrededor de 100 especies silvestres relacionadas con *Solanum* que crecen naturalmente en las Américas. El banco de Germoplasma de Papa de los Estados Unidos tiene como objetivo mantener muestras útiles para la investigación y el mejoramiento para mejorar el cultivo, a menudo en forma de familias de semillas botánicas. Un componente clave de la eficiencia del banco de germoplasma es evaluar la diversidad dentro y entre las poblaciones, y la diversidad de secuencias de marcadores de ADN es un poderoso indicador

✉ John Bamberg
John.Bamberg@usda.gov
https://www.ars.usda.gov/midwest-area/madison-wi/vegetable-crops-research/people/john-bamberg/bamberg-lab/

Alfonso del Rio
adelrioc@wisc.edu

Lisbeth Louderback
llouderback@anthro.utah.edu

Bruce Pavlik
bruce.pavlik@redbutte.utah.edu

[1] USDA/Agricultural Research Service, US Potato Genebank, 4312 Hwy. 42, Sturgeon Bay, WI. 54235, USA

[2] Department of Horticulture, US Potato Genebank, University of Wisconsin, 1575 Linden Drive, Madison, WI, USA

[3] Anthropology Department, Natural History Museum of Utah, University of Utah, 301 Wakara Way, Salt Lake City, UT 84108, USA

[4] Department of Conservation, Red Butte Garden and Arboretum, University of Utah, 300 Wakara Way, Salt Lake City, UT 84108, USA

de la diversidad de caractéres. Previamente informamos sobre tres factores que pueden causar una subestimación de la heterocigosidad: comprobación, frecuencia de alelos y sesgo de ploidía. Aquí informamos, utilizando datos de GBS para cuatro especies de papa diploides, que el porcentaje promedio de heterocigosidad aparente aumenta a medida que los datos son más completos: la diferencia máxima fue del 2% de heterocigotos cuando solo se consaidera a unos pocos individuos, al 36% cuando se incluye a casi todos los individuos. Sin embargo, hubo evidencia de que las estimaciones de la heterocigosidad promedio basadas solo en loci para los cuales cada individuo tiene datos también pueden estar sesgadas hacia arriba. Inverosimilmente, Los niveles altos de heterocigosidad sugieren SNP no segregantes no homólogos, que ocurrieron como 5–9% de todos los loci con datos completos. Proponemos que las mejores estimaciones de la heterocigosidad promedio en plántulas no seleccionadas deben basarse en loci con datos para todas las muestras después de eliminar aquellos loci que parecen estar fijados artificialmente como heterocigotos, lo que reduce la frecuencia de heterocigotos observada en un 16–26%. Sobre esa base, las especies silvestres examinadas tienen una heterocigosidad similar a la de *phureja* cultivada.

Abbreviations

USPG US Potato Genebank
GRIN Germplasm Resources Information Network (https://npgsweb.ars-grin.gov/gringlobal/search)
GBS Genotyping By Sequencing

Introduction

Like a craftsman who wants to be prepared for any challenge by stocking a diversity of tools in his toolbox, genebanks want to stock the widest genetic diversity possible so they will have the tools needed for potato research and breeding, even when we don't know what challenges the crop will face in the future (Bamberg et al. 2018). The practical diversity in genebanks is phenotypic traits, but trait expression is often tedious, difficult and costly to measure, and results may not be consistent over different environments. Since we assume phenotypes are ultimately linked to genotypes, DNA variation has been a widely-used proxy for trait diversity. Although many alleles per locus may exist, most markers define two alleles per locus.

Using DNA variation to make inferences about germplasm heterogeneity depends on some assumptions that, when faulty, can lead to wrong conclusions. In a previous paper (Bamberg and del Rio 2020), we provided empirical evidence for large diversity in populations of wild diploid potato species, and showed simulations that mimic ascertainment, allele frequency, and ploidy bias account for an under-estimation of true genetic diversity. We here report the examination of GBS datasets with anomalies that suggest two more sources of bias in diploid potato germplasm: missing data and detection of presumed non-allelic SNP fragments.

SNPs can provide many thousands of putative marker loci—increasingly more when more loci with data (calls) missing for more samples are accepted (Torkamaneh and Belzile 2015). Filtering for quality is routinely done. A rare SNP, seen in just one plant, for example, is more likely to be a technical mistake. Data is often missing for many of the plants, and filtering out loci with no more than about 25% missing calls is standard (Qiao et al. 2020; Revord et al. 2020). Since datasets are intended to represent single two-allele loci, one also wishes to filter out loci that do not detect both alleles in replicated reads and multiple individual samples as expected in diploid heterozygotes (Melo et al. 2016).

Materials and Methods

Species germplasm and sampling Three diploid wild species, *Solanum boliviense* (blv), *S. jamesii* (jam), and *S. microdontum* (mcd), and the diploid cultivated species *S. phureja* (phu) were examined. Samples for mcd and phu were individual plants from multiple populations. For blv, samples were from individuals from a single population. Leaf samples of jam were collected from 337 plants from 12 sites in the wild as documented in Bamberg (2019). The number of families and individuals are given in Table 1. Specific identities of materials are provided in Supplemental Table 1. All were obtained from USPG, the US Potato Genebank. Complete details of stocks used for blv, mcd, phu can be obtained online by searching for the six-digit PI number in GRIN (https://npgsweb.ars-grin.gov/gringlobal/search).

GBS data generation methods– DNA extraction and GBS library construction Plant DNA for all the samples was extracted using QIACube HT (Qiagen, Maryland, USA), a Qiagen's automated extraction robot, which performs high throughput extractions in a 96-well plate format. DNA samples were used for genotyping by sequencing (GBS), using 48- and 96-plex plates. The first step was to generate a reduced representation of their genomes using restriction endonucleases. This process aimed to cleave genomic DNA into manageable fragments for PCR and sequencing. The potato species went through a selection and optimization of restriction enzymes to determine which enzymes were more effective for GBS. Restriction enzymes EcoT22I-AvaIII proved suitable for mcd, phu and blv samples, PstI-BfaI for jam. After enzymatic cleavage sticky ends of DNA were used as anchors for the ligation of barcoded adaptors. Barcoded adaptors were used (a) to identify each sample uniquely from the others in pooled DNA sequencing, (b) served as primer binding sites for PCR amplification of the DNA before sequencing, and (c) to adhere other sequencing instrument specific adaptors to. After barcode ligation the

Table 1 Stocks assessed by GBS, loci with completely called individuals, and assessment of loci with too many heterozygotes to be plausibly single-locus segregating SNPs at $p < 5\%$

Species[1]	Seedlings (individuals)	Families[2]	Total loci		100% called individuals loci		Statistically unlikely homologous SNP individual loci	Statistically unlikely homologous SNPs group of loci
blv	19	1	14 K	→	517 = 4%	→	54 = 10%	38 = 8%
mcd	100	50	53 K	→	1710 = 3%	→	201 = 12%	152 = 9%
phu	34	34	42 K	→	1077 = 3%	→	58 = 5%	52 = 5%

[1] *Solanum boliviense, S. microdontum, S. phureja. S. jamesii* samples were not known to be unselected seedlings so not included

[2] See Supplemental Table 1 for details

DNA was amplified via PCR, and then purified and quantified before sequencing. The sequencing was conducted following Single-Read sequencing (100 bp) using HiSeq 2500 Sequencer (Illumina, San Diego, USA) for mcd and blv samples, and NovaSeq 6000 Sequencer (Illumina, San Diego, USA) for jam samples. These sequencers are part of the sequencing facilities at the University of Minnesota Genomics Center (https://genomics.umn.edu/) and the University of Wisconsin-Madison Biotech Center (https://www.biotech.wisc.edu/), respectively.

GBS data generation methods– sequence analysis and identification of SNPs For the sequence analysis, low-quality reads, reads with uncalled bases, and reads with adapter sequences were removed using computational pipelines developed at the Bioinformatics Resource Center (BRC), part of the Advanced Genome Analysis Resource unit of the Biotechnology Center at the University of Wisconsin–Madison (https://www.biotech.wisc.edu/). The trimming software Skewer (Jiang et al. 2014) was used to pre-process raw fastq files. Skewer implements an efficient dynamic programming algorithm designed to remove adapters and/or primers from sequence reads. Reads that were trimmed too short and did not meet the adequate sequence length criteria were also discarded from downstream analysis. The goal of this process was to recover true target DNA sequences. The reads were then checked for quality using the FastQC tool and each sequence was identified using its unique barcode adapter with a barcode splitter tool program (https://sourceforge.net/projects/gbsbarcode/). The raw sequences were processed using Tassel 3.0 and Universal Network Enabled Analysis Kit (UNEAK) pipelines (Lu et al. 2013) for de novo SNP discovery. These bioinformatics computational steps were performed on the Unix platform "Zcluster" at the UW-Biotech Center. The raw reads were analyzed for quality and trimmed, then trimmed reads were de-multiplexed generating high-quality reads for each genotype. The DM Potato Reference Genome and a GBS-specific simulated reference generated from a group of high-quality reads were used for alignment of sequences. The reads were mapped to generate standard alignment files using SAMtools version 1.3.1 (Li et al. 2009). A SNP master matrix for each potato species group was created after the pipeline for SNP discovery and genotype calling. The pipeline for SNP discovery followed standard conditions of depth coverage (minDP ≥ 2), maximum mismatch for alignment (n = 3), Maximum Missing Data (MaxMD = 80%), and Minimum Minor Allele Frequency (MinMAF ≥ 0.05).

Analysis To assess the bias against apparent heterozygosity due to missing data, the number of missing calls and percent heterozygous calls across individuals was calculated for all loci and plotted against each other. To assess bias favoring apparent heterozygosity by non-allelic SNPs, loci with calls for all individuals were sorted by number of heterozygous individuals, highest to lowest. Then the probability of observing the degree of heterozygosity observed at the most heterozygous loci was calculated with Chi^2, with the expectation of 1/3 heterozygous loci if SNP allele frequencies were assumed to be random, with random independent allele segregation at a single locus. All calculations were done with Microsoft Excel®.

Results and Discussion

Missing calls are associated with lower apparent heterozygote frequency The four species covered datasets with a small to large number of individual plant samples, but in all cases, there was a strong positive correlation between missing calls and (apparent) missing detection of maximum heterozygosity (Fig. 1.). The impact of more calls is particularly strong in the part of the curve representing 75–100% called plants often recommended for GBS data analysis. Thus, to avoid underestimation of heterozygosity, one should consider only loci with nearly complete calls (for every individual). This heterozygosity bias based on missing calls could also be enhanced due to inherent differences in average missing calls among taxa. In previous studies, the average missing calls among four potato species were

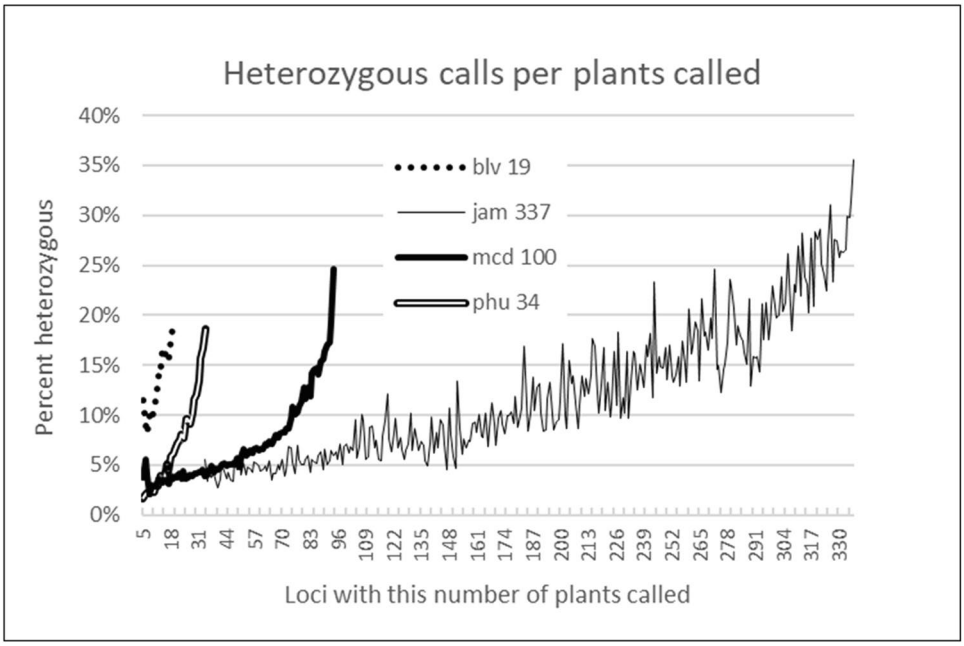

Fig. 1 The effect of missing data (less called plants) on apparent average heterozygosity for species *S. boliviense* (blv), *S. jamesii* (jam), *S. microdontum* (mcd) and *S. phureja* (phu). Number following species abbreviation in the legend is the total number of plants sampled

highly significantly different (Bamberg et al. 2015). In the current study, average missing calls were even different within a species, with the 12 populations of jam showing highly significantly different average missing calls (data not shown).

Excess heterozygotes indicate non-allelic fragments DNA marker data is organized as a table in which one dimension (e.g., the columns) represent the different DNA samples, and the values in a single row are assumed to represent genotypes of homologous segments that are potentially-segregating alleles at the same single locus. There are a few ways one can detect DNA marker genotypes that are not plausibly allelic. For example, with AFLPs, blanks represent the absence of a specific sequence and are contrasted with the dominant band that represents detecting that sequence. Apparent differences between two individuals are expected to be (1:0) for about the same number of loci as are (0:1). When the second individual appears to be very different because it has many loci with (0) genotype so most differences are (1:0), one concludes that this is better explained by a general failure of the technique to produce bands for the second individual.

Another implausible pattern is too much heterozygosity. Consider a locus for which every population examined has two SNP alleles. It is, of course, plausible to observe an AB heterozygote from each such population if the individuals genotyped had been deliberately *selected*– in nature, for example, as observed in Bamberg et al. (2009). But observing the AB genotype in every unselected seedling is a special case that we know is not usually plausible. Observing the AB genotype demonstrates that both alleles are indeed present in each population—each population appears to be polymorphic. But the maximum heterozygote frequency without selection is 50% and occurs when both A and B allele frequency are 50%. So, the chance of n segregating samples being all AB genotypes cannot be more than $(½)^n$. A corresponding implausibility cannot be calculated when every individual is homozygous unless there is an independent way of knowing that all of the populations from which the seedlings were drawn do possess the alternate allele.

The maximum average 50% heterozygosity of unselected random seedlings is too conservative for eliminating loci that are improbably heterozygous among all individuals, since empirical data suggests that average polymorphic loci tend to have more unbalanced allele frequencies (Bamberg and del Rio 2004; Bryan et al. 2017), so have a lower expected average heterozygosity than 50%. A reasonable compromise could be to propose *random* allele frequencies for polymorphic loci. The expected average heterozygote frequency for that is about 1/3, so we may set it as the expected proportion used to calculate the Chi2 probabilities for too-heterozygous loci.

For the three species blv, mcd, phu, we found an implausible number of uniformly heterozygous individuals in loci with complete calls (i.e., called for every individual). These could be a gauge of the number of apparent polymorphic SNP loci that are really not single-locus alleles. Table 1. shows that apparently polymorphic "loci" that are improbably allelic range from 5–9% of completely-called loci. The number of loci with implausibly frequent heterozygotes was

not presented for jam since those samples from the wild were not known to originate from unselected random seedlings. About 1% of the loci called for all 337 jam individuals were completely heterozygous which is far too many unless there was selection, non-allelic fragments, or redundant sampling of the same heterozygous clone. Considering only loci with all plants called and excluding loci with implausibly high heterozygote frequency, the average heterozygosity was blv = 14%, jam = 22%, mcd = 20%, phu = 16%.

Single plants underestimate heterozygosity The nineteen individuals of blv allowed assessment of SNP loci within a single population. A total of 517 SNP loci were called in all individuals—all loci were actually polymorphic, meaning they represented twice the potential genetic diversity of monomorphic loci. But the average observed heterozygote frequency in single plants was only 19%. If average allele frequencies were a balanced 50:50 we expect that a random single unselected seedling detects a heterozygote only half of the time. But in blv, the average allele frequencies must be closer to 10:90, making underestimation of heterozygotes about fivefold. We cannot precisely gauge within-population heterozygosity of a species based on a single-plant sample without knowing the allele frequencies.

Conclusions

The proportion of called SNP loci varies with the taxon assessed. And analysis of SNP data on individuals of four potato species showed that detection of heterozygosity increases with the proportion of called individuals for a given locus. On the other hand, many loci were heterozygous for too many individuals to be plausible under the assumption of allelic SNPs. Percent heterozygosity based on loci completely called, but not implausibly overwhelmingly heterozygous, may provide the best estimate of heterozygosity, and that estimate in the current study suggests diploid wild species are not less heterozygous than the diploid cultivated species examined. It was demonstrated that using a single plant to represent a population greatly underestimates heterozygosity.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12230-021-09846-z.

Acknowledgements Financial support for collection of *S. jamesii* samples was provided by the National Science Foundation, award no. BCS-1827414. The authors thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility for providing GBS facilities and services.

References

Bamberg, J.B. 2019. Southwest 2019 Potato (*Solanum*) Collecting Trip Report. Online at: https://npgsweb.ars-grin.gov/gringlobal/accessiondetail?id=2096809. Accessed 17 Aug 2021.

Bamberg, J.B., and A.H. del Rio. 2004. Genetic heterogeneity estimated by RAPD polymorphism of four tuber-bearing potato species differing by breeding system. *American Journal of Potato Research* 81: 377–383.

Bamberg, J.B., A.H. del Rio, J. Coombs, and D. Douches. 2015. Assessing SNPs versus RAPDs for predicting heterogeneity in wild potato species. *American Journal of Potato Research* 92: 276–283.

Bamberg, J.B., and A.H. del Rio. 2020. Assessing under-estimation of genetic diversity within wild potato (*Solanum*) species populations. *American Journal of Potato Research* 97: 547–553.

Bamberg, J.B., A.H. del Rio, and Rocio Moreyra. 2009. Genetic consequences of clonal versus seed sampling in model populations of two wild potato species indigenous to the USA. *American Journal of Potato Research* 86: 367–372.

Bamberg, J.B., A.H. del Rio, S.J. Jansky, and D. Ellis. 2018. Ensuring the genetic diversity of potatoes. In *Achieving sustainable cultivation of potatoes* No. 26, Vol.1, ed. Prof. Gefu Wang-Pruski, Chapter 3, pp 57–80. Cambridge, UK: Burleigh-Dodds Science Publishers.

Bryan, G.J., K. McLean, R. Waugh, and D.M. Spooner. 2017. Levels of intra-specific AFLP diversity in tuber-bearing potato species with different breeding systems and ploidy levels. *Frontiers in Genetics* 8: 119.

Jiang, H., R. Lei, S.W. Ding, and S. Zhu. 2014. Skewer: A fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15: 1–12. https://doi.org/10.1186/1471-2105-15-182.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, and N. Homer. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Lu, F., A.E. Lipka, J. Glaubitz, R. Elshire, J.H. Cherney, M.D. Casler, et al. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genetics* 9: e1003215. https://doi.org/10.1371/journal.pgen.1003215.

Melo, A.T.O., R. Bartaula, and I. Hale. 2016. GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics* 17: 29. https://doi.org/10.1186/s12859-016-0879-y.

Qiao, Y., F. Guo, N. Huo, L. Zhan, J. Sun, X. Zuo, Z. Guo, Y.Q. Gu, and Y. Liu. 2020. Genotyping-by-sequencing to determine the genetic structure of a Tibetan medicinal plant *Swertia mussotii* Franch. *Genetic Resources and Crop Evolution*. https://doi.org/10.1007/s10722-020-00993-6 (**Online 06 Aug 2020**).

Revord, R.S., S.T. Lovell, P. Brown, J. Capik, and T.J. Molnar. 2020. Using genotyping-by-sequencing derived SNPs to examine the genetic structure and identify a core set of *Corylus americana* germplasm. *Tree Genetics & Genomes* 16: 65. https://doi.org/10.1007/s11295-020-01462-y.

Torkamaneh, D., and F. Belzile. 2015. Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. *PLoS ONE* 10 (7): e0131533. https://doi.org/10.1371/journal.pone.0131533.

American Journal of Potato Research is a copyright of Springer, 2021. All Rights Reserved.