



Genome diversity of the potato

Binqun Huang^{a,1}, David M. Spooner^{b,c,1}, and Qiqi Liang^{d,1}

In a recent issue of PNAS, Hardigan et al. (1) state that the genetic diversity of the potato is much greater than that of other major crops, based on 68.9 million SNPs identified from the resequencing of 67 accessions of wild and cultivated potatoes. We questioned this conclusion based on our own original analysis of wild and cultivated potato species (2) and estimates of genomic SNPs in other major crops by next-generation resequencing with a few to 15 million SNPs. Examples include soybean (3), pigeon pea (4), cotton (5), tomato (6), and potato (2) (Table 1). To explore this further, we reanalyzed the raw data from ref. 1 using standard, stricter methods to filter SNPs and then reanalyzed the data from both studies (1, 2) with similar subsets of cultivated and wild species, focusing on only diploid germplasm. Since greatly relaxed procedures for filtration of SNPs in ref. 1 were used, it is likely that false SNPs were identified. Results from our analysis led to a different conclusion than that of ref. 1, which provides higher estimates of diversity in wild and cultivated potatoes relative to our study (Table 2).

Hardigan et al. (1) identify 46,797,252 SNPs in 20 accessions of diploid wild species and 26,560,638 SNPs in 10 diploid landrace genotypes. We obtained many fewer, 10,473,482 SNPs from 20 diploid wild potatoes and 8,108,352 SNPs from 10 diploid landrace genotypes. We suggest that the large numbers of SNPs in ref. 1 resulted from relaxed filtration procedures, likely leading to an overestimate of diversity.

The hard filter in GATK (7) is employed to remove false SNPs. This filter removes (i) alignments with an MQ score <40, (ii) genotype quality (GQ) <20, (iii) low-quality sites (QUAL) <20, (iv) Fisher strand (FS) >60, and (v) QualByDepth (QD) <4. Hardigan et al. (1) used GATK to call SNPs but did not apply the filtering criteria QD and FS to the raw SNPs, thus retaining low-confidence SNPs. Finally, their MQ score was <20, not <40. In sum, their analysis likely retained many false SNPs, which led to overestimates of diversity. Our new reanalysis of Hardigan et al. (1) with stricter SNP retention criteria obtained 10,713,582 SNPs (GATK, hard filter) and 11,264,745 SNPs (SAMtools and GATK), leading to genetic diversity estimates in diploid wild and cultivated potatoes of 0.5334×10^{-3} and 0.3558×10^{-3} , respectively. This is consistent with estimates in our diploid wild and cultivated potato populations, $\pi_w = 0.4149 \times 10^{-3}$ and $\pi_c = 0.2696 \times 10^{-3}$ (Table 2). These results have broad implications for potato breeding. Diploid cultivated relatives have been used for base broadening in breeding programs (8, 9). High levels of genetic diversity in this germplasm would support this strategy but phenotypic variation does not necessarily require genetic diversity. Genetically similar individuals may be highly variable for traits of interest to breeders, such as tuber shape and color (10). Finally, wild germplasm is generally used by breeders for specific traits, so genome-wide diversity is not as important as diversity at selected loci.

^aDepartment of Plant Sciences, University of Oxford, Oxford OX1 3RB, United Kingdom; ^bVegetable Crops Research Unit, Agricultural Research Service, US Department of Agriculture, Madison, WI 53706; ^cDepartment of Horticulture, University of Wisconsin–Madison, Madison, WI 53706; and ^dNovogene Bioinformatics Institute, 100083 Beijing, China

Author contributions: B.H., D.M.S., and Q.L. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

Published under the [PNAS license](#).

¹To whom correspondence may be addressed. Email: binqun.huang@plants.ox.ac.uk, David.Spooner@ars.usda.gov, or liangqiqi@novogene.com.

Published online June 22, 2018.

Table 1. Genome sizes and numbers of SNPs in five major crop species

Species	Genome size (millions)	No. of accessions	Average coverage depth of genome	Filtered high-quality SNPs	Ref.
Soybean	950	302	>11	9,790,744	3
Pigeon pea	833	292	12	15,162,233	4
Cotton	2,500	352	6.9	7,497,568	5
Tomato	900	360	5.7	11,620,517	6
Potato	844	167	>12	6,487,006	2
Potato	844	67	11	68,914,903	1

Table 2. SNP variation in diploid cultivated and wild potatoes from Hardigan et al. (1) and data reanalyzed here from Li et al. (2)

Population composition	Population	SNP sites (different filtration methods)			Genetic diversity π (10^{-3})	
		Hardigan et al. (1)	GATK (hard filter) from Van der Auwera et al. (7)	GATK and SAMtools*	GATK and SAMtools [†]	Hardigan et al. (1)
Hardigan et al. (1)	Wild species, 2x (20) [‡]	46,797,252	2,049,897	7,202,681	0.5334	12.2
	Landrace, 2x (10)	26,560,638	1,761,567	5,516,552	0.3548	8.7
	Wild and landrace, 2x (30)	—	2,061,888	7,354,643	0.5577	—
	Full panel (67) [§]	68,914,903	10,713,582	11,264,754	1.2833	—
Similar accessions to Hardigan et al. (1)	Wild species, 2x (20) [¶]	—	1,924,256	10,473,482	0.4149	—
	Landrace, 2x (10) [¶]	—	1,641,373	8,108,352	0.2696	—
	Wild and landrace, 2x (30)	—	1,936,687	11,018,742	0.4409	—

*Common sites of SNPs based on both GATK and SAMtools were considered as the raw SNP candidates. High-quality SNPs were supported by at least five mapped reads, rms mapping quality ≥ 20 , phred-scaled genotype quality ≥ 5 , and less than 0.2 missing data.

[†]Based on high-quality SNPs using both GATK and SAMtools.

[‡]Number of genotypes in parentheses.

[§]Full panel, including diploid and tetraploid landraces, cultivars, wild species, and the outgroup.

[¶]Accessions of wild and cultivated potatoes have the same name, same ID, or similar geographic distribution as the accessions used in Hardigan et al. (1). Four of 20 wild species accessions and 6 of 20 landrace accessions are identical.

- 1 Hardigan MA, et al. (2017) Genome diversity of tuber-bearing *Solanum* uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc Natl Acad Sci USA* 114:E9999–E10008.
- 2 Li Y, et al. (2018) Genomic analyses yield markers for identifying agronomically important genes in potato. *Mol Plant* 11:473–484.
- 3 Zhou Z, et al. (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol* 33:408–414.
- 4 Varshney RK, et al. (2017) Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat Genet* 49:1082–1088.
- 5 Wang M, et al. (2017) Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat Genet* 49:579–587.
- 6 Lin T, et al. (2014) Genomic analyses provide insights into the history of tomato breeding. *Nat Genet* 46:1220–1226.
- 7 Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:1–33.
- 8 Bradshaw JE, et al. (2006) Genetic resources (including wild and cultivated *Solanum* species) and progress in their utilisation in potato breeding. *Potato Res* 49:49–65.
- 9 Plaisted RL, Hoopes RW (1989) The past record and future prospects for the use of exotic potato germplasm. *Am Potato J* 66:603–627.
- 10 Jansky SH, Dawson J, Spooner DM (2015) How do we address the disconnect between genetic and morphological diversity in germplasm collections? *Am J Bot* 102:1213–1215.