A Pan-Genome Data Structure Induced by Pooled Sequencing Facilitates Variant Mining in Heterogeneous Germplasm

Patrick A. Reeves (pat.reeves@usda.gov)
USDA-ARS: Fort Collins, CO, US https://orcid.org/0000-0001-9991-1397
Christopher M. Richards
USDA ARS National Laboratory for Genetic Resources Preservation: Fort Collins, Colorado, US
https://orcid.org/0000-0002-9978-6079

Research Article

Keywords: bioinformatics, CRISPR, crop wild relatives, domestication, haplotype, phasing, sequence variant

Posted Date: January 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1296984/v1

Version of Record: A version of this preprint was published at Molecular Breeding on June 25th, 2022. See the published version at https://doi.org/10.1007/s11032-022-01308-6.

Abstract

Valuable genetic variation lies unused in gene banks due to the difficulty of exploiting heterogeneous germplasm accessions. Advances in molecular breeding, including transgenics and genome editing, present the opportunity to exploit hidden sequence variation directly. Here we describe the pangenome data structure induced by wholegenome sequencing of pooled individuals from wild populations of *Patellifolia spp.*, a source of disease resistance genes for the related crop species sugar beet (*Beta vulgaris*). We represent the pangenome of a heterogeneous population sample as a set of phased reads mapped to a reference genome assembly derived from the sequence pool or other sources. We show that this basic data structure can be queried by homology to identify short haplotypic variants present in the wild relative, at genes of agronomic interest in the crop. Further we demonstrate the possibility of cataloging short haplotypic variation in all *Patellifolia* genomic regions that have corresponding single copy orthologous regions in sugar beet. The data structure, termed a "phased read archive", can be produced, altered, and queried using standard tools to facilitate discovery of agronomicallyimportant sequence variation in heterogeneous germplasm.

Introduction

Germplasm collections are a source of novel genetic variants for breeding improved traits in cultivars of crop species. Within a breeding program, if donor parent germplasm from a gene bank accession is unimproved, many cycles of backcrossing to the elite parent are necessary to mitigate the inevitable introduction of undesirable traits genetically linked to the variant of interest. When donor parent germplasm is heterogeneous, isogenic lines may need to be produced prior to introgressive breeding to limit the scope of linkage drag and increase phenotypic uniformity. Recurrent backcrossing and production of isogenic lines is laborious and time-consuming, sometimes requiring many years (Rojas et al. 2009; Biancardi et al. 2010; McCouch et al. 2012). In some species the production of isogenic lines may be precluded by inbreeding depression (Li and Brummer 2009; Lindhout et al. 2011).

In breeding programs where the trait of interest is welldefined genetically, meaning most genes impacting the trait are known, genome editing or transgenic approaches can be used to target and modify gene expression patterns to produce desired traits with little need for subsequent "clean up" via recurrent backcrossing (Wolter et al. 2019). In these cases sequence data repositories, as opposed to living collections of germplasm, have the potential to facilitate targetgenespecific generation of genetic diversity in crops and other species (RodríguezLeal et al. 2017; Scheben and Edwards 2018; Belzile et al. 2020). This approach is distinct from conventional breeding wherein germplasm repositories are queried for desirable traits biologically, via controlled crosses *in vivo*, rather than informatically, by identifying useful homologous sequence variation or target sequences *in silico*.

Currently there is momentum to commence systematic, wholegenome, wholecollection genotyping of gene bank holdings (McCouch et al. 2020). Encouraging progress has been made in barley, rapeseed, rice, sorghum, and *Capsicum* (Wang et al. 2018; Hu et al. 2019; Milner et al. 2019; Wu et al. 2019; Tripodi et al. 2021). Crop varietal production often involves winnowing phenotypic variation by progressive inbreeding,

such that a new variety may be near isogenic, to satisfy the requirements that it be distinct, uniform, and stable, for intellectual property protections. Indeed, extensive varietal differentiation is testament to a crop species' capacity to be recombined to create a wide array of genetically homogeneous and phenotypically predictable lines. Accordingly, gene bank accessions from domesticated crops tend towards genetic homogeneity. Whole-collection genomewide genotyping efforts focused on crop varieties and other homogeneous accessions appropriately sample only one or a few individuals per accession because that is all that is needed.

Wild species retain a vast reserve of genetic variation that was left unsampled as emerging crops passed through the domestication bottleneck (Hawkes 1977; Doebley et al. 2006). In contrast to crop varieties, accessions from wild populations are often genetically heterogeneous. In addition to ubiquitous single nucleotide polymorphisms (SNPs), gene content also varies dramatically among individuals, with ~25% of genes belonging to the "dispensable" fraction of the species pangenome (Gao et al. 2019). Due to widespread structural variation, the noncoding dispensable fraction may be even higher (Hübner et al. 2019). In *Beta vulgaris*, for example, flow cytometric estimates of genome size varied from 633–875 MB, a 40% difference (McGrath et al. 2020; Castro et al. 2013). For heterogeneous accessions, genomewide data from a single individual represents an incomplete accounting of sequence variation and is a poor predictor of what a user can expect to receive when the accession is requested for use in a breeding program.

Gene banks are increasingly interested in cataloging the interesting, useful, and valuable genetic variants hidden in collections of wild germplasm (Tanksley and McCouch 1997; Gur and Zamir 2004; Hajjar and Hodgkin 2007; Mascher et al. 2019). One use of such a catalog is to enable molecular breeding initiatives to generate diversity in elite cultivars using transgenic or genome editing approaches, informed by the wide range of sequence variants present in unimproved germplasm. Variants present alternative, naturallyoccurring models for sitespecific base editing in the crop, or targetsequence information for "*de novo* domestication" of unruly wild relatives with otherwise desirable stress tolerance, disease resistance, morphological, or nutritional properties (Zsögön et al. 2018; Li et al. 2018; Lemmon et al. 2017). Additionally, for conventional breeding projects, the catalog could support prediction of the frequency of particular traits or trait values recoverable from accessions based on the frequency of haplotypic variants of defined function segregating therein. For traits that are wellunderstood genetically, this could enable initial selection of accessions for integration into a breeding program *bioinformatically*, instead of via extensive growouts and phenotyping or imprecise suggestions from passport data and surrogate predictors of genetic diversity (Reeves et al. 2020).

We here describe the pangenomelike data structure induced by the application of whole genome sequencing to pools of individual samples. We represent the pangenome as a queryable catalog of wholegenome multiallelic short haplotype ("microhaplotype" *sensu* Kidd et al. 2014) variation in six heterogeneous samples of *Patellifolia* spp., a wild relative of sugar beet (*Beta vulgaris*). Using this data structure, we calculate the proportion of expressed sugar beet genes for which orthologous variation can be found in distantlyrelated *Patellifolia*, which diverged from *Beta* ~25 Mya (Romeiras et al. 2016). We further consider the proportion of the entire sugar beet genome for which haplotype variants could be mined from

Patellifolia and we catalog those variants, using two sugar beet loci of agronomic importance as exemplars of the approach.

Materials And Methods

Leaf tissue from 17-25 individuals was collected from six wild populations of *Patellifolia* spp. in the Canary Islands and mainland Spain (Table 1; Frese et al. 2019). DNA was extracted using the DNeasy 96 Plant Kit (Qiagen GmbH, Hilden, Germany), concentrationnormalized to 20 ng/uL, and pooled for DNA sequencing. Sequencing libraries were generated from DNA pools using the KAPA HyperPrep Kit (Roche, Basel, Switzerland) with a PCRfree workflow and average insert size of 300 bp then sequenced on a NovaSeq instrument (Illumina Inc., San Diego, USA) producing 1.8E8–3.4E8 150 bp paired end reads per pool (Table 2).

Table 1
Patellifolia spp. sampling for pooled sequencing.

Species	Location	Latitude, Longitude	Individuals in pool
webbiana	Gran Canaria	28.172482, -15.419560	25
procumbens	Tenerife	28.553550, -16.348550	17
procumbens	El Hierro	27.747923, -18.098359	25
patellaris	Spain	37.557349, -1.168413	25
patellaris	Tenerife	28.376967, -16.799400	25
patellaris	Spain	37.504414, -1.425755	25

Table 2
DNA sequencing and pool genome assembly.

Species	Location	Sequence read pairs	Assembly size (Mbp)	Coverage (per individual)	Contigs	N50
webbiana	Gran Canaria	2.2E8	747	88x (3.5x)	258036	13789
procumbens	Tenerife	2.2E8	790	84x (4.9x)	285422	13366
procumbens	El Hierro	3.4E8	790	129x (5.2x)	271266	12459
patellaris	Spain	2.3E8	1114	62x (2.5x)	162082	17866
patellaris	Tenerife	1.8E8	1136	48x (1.9x)	202527	16679
patellaris	Spain	3.2E8	1090	88x (3.5x)	129336	20349

Details of bioinformatic procedures are at https://github.com/NCGRP/mb1suppl. We used MASURCA 3.2.4 (Zimin et al. 2013) to produce a crude genome assembly for each pool (hereafter, "pool assembly"). Trimmed read pairs (TRIMMOMATIC, Bolger et al. 2014) were mapped to the pool assembly using BWA-

MEM (Li 2013) to confirm proximity, filtered by quality using SAMTOOLS (Li et al. 2009), and duplicates removed with SAMBAMBA (Tarasov et al. 2015) before fusing into individual phased sequence reads, sometimes called merged reads or "FLASHed reads" (Bushnell et al. 2017; Sundaram et al. 2020). If read pairs were proximal and oriented correctly but were nonoverlapping then the intervening region was padded with a string of Ns of a length predicted from pool assembly contigs. This procedure was automated with our pipeline, HAPX (https://github.com/NCGRP/hapx), which uses a reference genome (here, the pool assembly) to filter by alignment quality, proximity, and orientation during read pair phasing. HAPX was set to produce phased reads only when both read pairs were mapped, properly oriented, within 1000 bp of one another, had a minimum mapping quality of 1, and belonged to the primary alignment, with no split reads allowed. The referencefree procedure used by FLASH and BBMERGE (Magoč and Salzberg 2011; Bushnell et al. 2017), which considers sequence overlap alone, is also suitable for phasing read pairs (Baetscher et al. 2018).

Binary alignment map files (BAM files) specifying the mapping between phased reads and pool assembly contigs were merged into one. The mapped phased reads contained therein plus associated pool assembly contigs were processed into a single sorted multiFASTA file. A BLAST database was constructed for the multiFASTA file and for the pool assembly alone using BLAST+ 2.5.0. We define a data structure, hereafter referred to as a "phased" or "flashed read archive" (FRA), that contains: 1) an indexed reference genome, 2) a BAM map between phased DNA sequences and the reference genome, 3) a sorted FASTA file containing phased DNA sequences and contigs from the BAM map, and 4) a BLAST database for the FASTA file and reference genome. This amalgamated data structure allows the use of standard software to query and retrieve short haplotypic variants from pooled sequence data by genome position (e.g. SAMTOOLS) or homology (BLAST).

We evaluated FRA quality by calculating average phased read length and coverage of the pool assembly. We explored the utility of *Patellifolia* FRAs as a source of variants for sugar beet genes by determining the proportion of the sugar beet transcriptome for which homologous sequence was present in *Patellifolia* pools. Phased read BLAST databases were queried with all 24255 primary transcripts in the sugar beet EL10 transcriptome (McGrath et al. 2020) with up to 100K matches returned per query. BLASTN results were filtered to exclude gene models that matched < 40 and > 1000 phased reads. These cutoff values were determined empirically to capture the linear portion of the sigmoid curve relating cumulative query frequency and log BLAST hit count (Supplemental Figure 1). Remaining matches were considered to represent the fraction of homologous genes, excluding highly repetitive, unmatched, and poorly represented genes. This cutoff resulted in discarding 1422 genes as repetitive (chloroplast, rDNA, and mitochondrial genes, plus gene models with repeated amino acid motifs), 3793 as unmatched, and 2130 genes as poorly represented in the *Patellifolia* pools. We defined orthologous genes operationally, using BLASTN, as EL10 transcript queries that matched only one contig in the pool assembly from diploid *P. procumbens* or *webbiana* (i.e. they were present as single copy genes in the assembly), and two or fewer contigs from tetraploid *P. patellaris*.

We determined the proportion of the *Patellifolia* pangenome represented in the pools that was homologous to the sugar beet genome. The nine chromosomal contigs from sugar beet genome assembly EL10 v1.0

were fragmented into sequential 1 Kbp sequences, each of which was then used as a BLASTN query against the FRA for the purposes of determining orthology, as was done with the transcriptome except that up to 10K matches were allowed to be returned with no subsequent filter on phased read depth per query applied, in order to retain information on short repetitive sequences.

To demonstrate the capacity of pooled sequencing data to facilitate variant mining, we characterized variant frequencies within *Patellifolia* pools at agronomicallyimportant cyst nematode resistance gene *Hs4* (Kumar et al. 2021) and the *Patellifolia* ortholog of pseudoresponse regulator *BvBTC1*, which determines annual versus biennial life cycle in sugar beet (Pin et al. 2012). Full length mRNA sequences were used to query FRA BLAST databases using BLASTN to identify *Patellifolia* contigs containing *Hs*4 and *BvBTC1*, along with the filtered phased reads mapped to those contigs, as contained in the FRA BAM file. For detailed analysis and visualization, a region containing ~12 Kbp and ~3 Kbp was arbitrarily defined for *BvBTC1* and *Hs4*, respectively, which encompassed complete coding sequence exons, introns, and some adjacent sequence. Short haplotypes were identified, and variant frequencies were estimated using HAPXM (https://github.com/NCGRP/hapxm) along a tiling path that maximized locus variation, accuracy, and length, in order to simplify presentation of results.

Results

Pool genome assemblies varied in size slightly, averaging ~775 Mbp in diploids *P. webbiana* and *procumbens*, ~1.1 Gbp in tetraploid *P. patellaris* (Table 2). Schlotterer et al. (2014) showed that pooled sequencing coverage in excess of 1x per individual produces allele frequency estimates that are equal to or more accurate than those computed from individual sequencing. Estimated coverage per individual in our pools varied from 1.9–5.2x. Pool assemblies were fragmented, containing on average 2.7E5 contigs with mean N50 ~13 Kbp for diploid pools, 1.6E5 contigs, mean N50 ~18 Kbp for tetraploids.

After read phasing and filtering, coverage per individual remained above 1x (1.4–3.5x). Average phased read length ranged from 251–275 bp, a substantial increase from the initial 150 bp reads (Table 3). Seventy percent of the 24255 primary transcripts in the EL10 transcriptome were found to have complementary phased reads in *Patellifolia*. Twenty five to thirty percent were single copy in *Patellifolia* pool assemblies. The latter genes are straightforward targets for variant mining using associated phased read data. Similarly, 72–80% of the sugar beet genome was found in the *Patellifolia* pools, 15–21% of it as single copy regions (Table 3). Thus, approximately one fifth of the sugar beet genome is directly accessible for improvement using sequence diversity mined from *Patellifolia*, without the complicating issue of paralogy, and about three quarters otherwise (Figure 1). Put differently, no homologous counterpart for about 25% of sugar beet genome EL10 was found in the *Patellifolia* pangenome using our procedure.

Table 3
Proportion of sugar beet transcriptome and genome with homologous sequence reads in wild relative *Patellifolia* spp.

Species	Location	Phased read coverage, per individual (count)	Average phased read length, bp	Proportion of EL10 transcriptome found in *Patellifolia* (transcript count)	Proportion of EL10 transcriptome single copy in *Patellifolia* (transcript count)
webbiana	Gran Canaria	2.4x (1.8E8)	251	0.71 (17179)	0.36 (8617)
procumbens	Tenerife	3.5x (1.7E8)	273	0.71 (17104)	0.32 (7797)
procumbens	El Hierro	3.5x (2.6E8)	268	0.70 (16909)	0.29 (6953)
patellaris	Spain	1.8x (1.8E8)	273	0.70 (16995)	0.28 (6721)
patellaris	Tenerife	1.4x (1.4E8)	273	0.69 (16784)	0.25 (6031)
patellaris	Spain	2.5x (2.4E8)	275	0.69 (16779)	0.30 (7289)

Table 3
, cont'd.

Species	Location	Proportion of EL10 genome found in *Patellifolia*	Proportion of EL10 genome single copy in *Patellifolia*
webbiana	Gran Canaria	0.72	0.17
procumbens	Tenerife	0.72	0.16
procumbens	El Hierro	0.80	0.15
patellaris	Spain	0.74	0.21
patellaris	Tenerife	0.72	0.21
patellaris	Spain	0.75	0.21

BLASTN query of FRA BLAST databases using *BvBTC1* and *Hs4* coding sequences yielded, in both cases, one matching contig in diploid and two matching contigs in tetraploid *Patellifolia* pool assemblies. In tetraploid *P. patellaris* pools, the matching homeolog was identified using indels shared with *P. procumbens* and *P. webbiana*. Depending on pool, between 1493 and 5384 phased reads were mapped to the ~12 Kbp genomic region containing *BvBTC1* (depth 34x–122x); 296–812 phased reads mapped to the ~3 Kbp region containing *Hs4* (24x–66x). The tiling paths across *BvBTC1* and *Hs4* contained 3220 and 754 short haplotype loci with a mean (± 1 SD) length of 2.99 ± 1.39 bp and 3.79 ± 2.58 bp, respectively. The number of variants per locus ranged from 1.65 ± 0.62 to 2.91 ± 0.94 for *BvBTC1*, and 1.01 ± 0.85 to 2.47 ± 0.89 for *Hs4*. These values were correlated with depth because no minor allele frequency cutoff was used except that singletons were disallowed; some low frequency variants attributable to sequencing error are therefore included in the estimates. 82% of *BvBTC1* short haplotype loci comprised indel variants only, 18%

contained at least some multi-nucleotide polymorphisms (MNPs). For *Hs4*, 80% of loci were indelonly, 20% contained MNPs. Among pools, the major variant frequency ranged from 0.87–0.91 for *BvBTC1* and 0.86–0.94 for *Hs4*. Major variant differences between pools across the genes are visualized in Figure 2. Perpool descriptive statistics are in Table 4.

Table 4

Short haplotype variation present in *Patellifolia* pools across *BvBTC1* and *Hs4*. [1]Multinucleotide polymorphism.

Species	Location	Phased reads, *BvBTC1* (depth)	Phased reads per locus (± 1 SD), *BvBTC1*	Variants per locus, *BvBTC1*	Major variant frequency, *BvBTC1*	Indelonly variants, *BvBTC1* (proportion)	MNPs[1], *BvBTC1* (proportion)
webbiana	Gran Canaria	3608 (77x)	49.8 ± 9.1	2.23 ± 0.72	0.91 ± 0.06	2764 (0.97)	93 (0.03)
procumbens	Tenerife	3442 (79x)	47.6 ± 8.4	2.37 ± 0.69	0.88 ± 0.07	2972 (0.97)	103 (0.03)
procumbens	El Hierro	5384 (122x)	70.9 ± 11.0	2.91 ± 0.94	0.87 ± 0.08	2977 (0.94)	198 (0.06)
patellaris	Spain	2071 (48x)	30.2 ± 6.5	2.00 ± 0.62	0.89 ± 0.07	2606 (0.99)	32 (0.01)
patellaris	Tenerife	1493 (34x)	19.9 ± 5.0	1.65 ± 0.62	0.91 ± 0.10	1837 (0.99)	15 (0.01)
patellaris	Spain	2534 (59x)	35.4 ± 6.6	2.07 ± 0.53	0.89 ± 0.06	2885 (0.99)	15 (0.01)

Table 4
, cont'd.

Species	Location	Phased reads, *Hs4* (depth)	Phased reads per locus (± 1 SD), *Hs4*	Variants per locus, *Hs4*	Major variant frequency, *Hs4*	Indelonly variants, *Hs4* (proportion)	MNPs[1], *BvBTC1* (proportion)
webbiana	Gran Canaria	568 (43x)	26.6 ± 6.6	1.62 ± 0.66	0.94 ± 0.08	367 (0.93)	27 (0.07)
procumbens	Tenerife	812 (66x)	35.3 ± 7.1	2.23 ± 0.75	0.87 ± 0.10	609 (0.91)	62 (0.09)
procumbens	El Hierro	786 (63x)	38.2 ± 8.5	2.47 ± 0.89	0.86 ± 0.11	600 (0.88)	79 (0.12)
patellaris	Spain	584 (47x)	28.7 ± 6.5	1.94 ± 0.63	0.90 ± 0.08	577 (0.98)	14 (0.02)
patellaris	Tenerife	296 (24x)	9.7 ± 7.7	1.01 ± 0.85	0.92 ± 0.09	469 (0.99)	5 (0.01)
patellaris	Spain	698 (57x)	30.8 ± 8.7	1.97 ± 0.66	0.90 ± 0.07	591 (0.98)	13 (0.02)

Supplemental Figure 1. Identification of sugar beet genes that are poorly represented or repetitive in the *Patellifolia* spp. poolseq pan-genome. The non-repetitive, homologous fraction of the *Patellifolia* spp. pan-genome (shaded box) was defined empirically as containing those EL10 transcripts represented by 40–1000 phased reads in BLAST searches.

Discussion

Biological exploration of germplasm through careful breeding and artificial selection has been used to improve crops since the dawn of agriculture. Digitization of germplasm collections so that they may also be explored using *information* is a longstanding objective of the gene banking enterprise (Volk et al. 2021). Increasingly standardized and interoperable data bases have facilitated query of collections' basic descriptive, or "passport", data (Wiese et al. 2020). Enhancing our ability to interrogate collections informatically, at the level of DNA sequence variation in addition to passport and phenotypic data, will accelerate agricultural progress (McCouch et al. 2020).

In this study we describe the cataloging of DNA sequence variants from heterogeneous collections of individuals to inform crop molecular breeding. We focus on crop wild relatives in the secondary and tertiary gene pools, where biological interrogation of variation using crosses is difficult due to reproductive barriers, and primary gene pool members where complex growth requirements or inbreeding depression prevents extensive development of inbred lines. We construct a pangenome data structure for this purpose that represents sequence diversity within the sample and which may be queried by homology or genomic position.

Pooled sequencing, the process of sequencing DNA from multiple individuals simultaneously, induces a pangenome data structure in its output, with sequence variation represented as independent reads derived from individuals in the pool. Reads may vary in length and configuration depending on sequencing technology, but, assuming PCRfree library construction, should represent actual sequence variants at the approximate frequency they occur in the pool of individuals. This form of pangenome representation differs from the common one (separately assembled genomes from multiple individuals) and lacks the power to evaluate properties such as adjacency and synteny or core versus disposable fractions (Bayer et al. 2020). However, it in principle can reveal the totality of sequence complexity in the sample, albeit as fragments instead of chromosome length pseudomolecules.

For cataloging sequence variation, the "poolseq pangenome" is a costeffective and appropriate alternative to sequencing individuals. In a typical outcrossed wild accession, all individuals are expected to be unique, so it is not possible to capture any particular sequenced genome again by repeated grow outs. Similarly, long haplotypes, those beyond the typical rate of linkage decay, may not be easily recovered. Because recombinatorial introgression of desirable variants via crossing is often not an option when working with secondary or tertiary gene pools, knowledge of the longdistance physical linkage relationships that emerge from sequencing individuals may not be relevant for crop improvement.

In cases where introgressive breeding is possible, the poolseq pangenome more accurately represents what is actually delivered when a heterogeneous accession is requested: the capacity to recombine into a target line some proportion of the sequence variation segregating in the accession. Thus a poolseq pangenome data structure encapsulates the constraints of reproductive biology inherent to the breeding process, as well as relating logically to most phenotypic characterization and evaluation data held in gene bank data bases, which is usually assessed at the level of populations not individuals (Galewski and McGrath 2020).

Production of a "phased read archive" (FRA) to enable query of a poolseq pangenome involves three main steps: read phasing, mapping of phased reads to a reference assembly, and production of BLAST data bases. Read phasing is essential for short read data because it increases the complexity of sequences exposed to homology query by increasing their length. The mapping step uses a reference genome assembly as a scaffold upon which to filter phased reads based on proximity, orientation, and alignment quality. The map enables query by position in the reference as opposed to query by homology. In general, direct homology query of phased reads is preferred because they represent real molecules present in the pool that come from a single individual, not bioinformaticallyderived contig sequences. Nevertheless, query

by reference position can be useful in certain situations, such as for polyploids or when duplicated genes or common sequence motifs would cause homology query to return excessive numbers of phased reads.

As proof of concept, we considered in detail two genes of agronomic importance in sugar beet. Haplotypic variation in the "bolting gene", *BvBTC1*, produces differences in bolting time, most coarsely at the level of assuming a biennial vs an annual life cycle, but with some additional variation within categories attributable to rare haplotypes and probably mediated by other genes in the flowering pathway (Pin et al. 2012; Höft et al 2018; Kuroda et al. 2019). Variation of this nature can affect decisions regarding planting time and may be relevant to improving earlybolting resistance.

The *Patellifolia* ortholog of *BvBTC1* was found by querying the FRA pool assembly. Examination of mapped phased reads revealed many differences in short haplotype major variants between pools across the genic region (Figure 2). Because major variant frequencies were generally high (Table 4), one could recover variants of interest with high probability by resampling the biological material from which the pooled sequencing data were drawn. Moreover, for most major variants, only two alleles were found (visualized in Figure 2 using red and green) which can be recovered with relative ease, by accessing only two accessions, at least for this set of six pools. Because allele frequencies for all short haplotype variants, minor or major, in the poolseq pangenome can be estimated, the minimum probability of recovering longer haplotypes from an accession can be approximated, as the joint probability of recovering physically adjacent short haplotypes. The probability of recovering sets of unlinked variants scattered across the genome can be similarly calculated. This allows one to set the scale of an allele mining experiment, in terms of the number of individuals required to ensure recovery of variants, at any position, or set of positions, in the genome.

The cyst nematode resistance gene *Hs4* has been transferred to some sugar beet germplasm via the translocation of large genomic segments from *Patellifolia* (Kumar et al. 2021). Via query of pool assemblies, we found that *Hs4* is single copy in *P. procumbens* and *P. webbiana*. The homeolog from tetraploid *P. patellaris* diverges less from the *P. procumbens/webbiana* ortholog than the analogous situation in *BvBTC1*, but orthologs are still identifiable by the presence of shared indels. Thus in the *Patellifolia/ Beta vulgaris* tertiary gene pool relationship, ploidy variation seems to be a surmountable problem. This may not always be the case in other extended gene pools due to the nature of initial polyploidy events ("allo" vs "autopolyploidy") and evolutionary processes affecting homeologous regions since then. Polyploids are an underutilized resource for exploring sequence variation due to technical difficulties encountered when applying many methods of polymorphism assessment. Polyploid sequence variation can be recovered readily from the FRA because phased reads represent actual sequence variation in the pool regardless of which genome they originate from.

The poolseq pangenome can be utilized to select germplasm from large gene bank collections. As with *BvBTC1*, major allele frequencies across the *Hs4* genic region were high and the number of variants tended towards two (Table 4). As an example, let's suppose one was especially interested in *Hs4* exon 5. We show that there are three major haplotypic variants spanning that ~182 bp region, composed of four short haplotype loci, among the six pools examined (Figure 2). To retrieve these three haplotypes with greatest

efficiency one should use accessions represented by the *P. webbiana* and two *P. patellaris* pools, because these have the highest major variant frequencies at the four short haplotype loci. In this way, the poolseq pangenome could be used to develop locus specific core collections.

Whole genome data sets are large and complex. They are often used collectively to produce a summary of population structure among accessions, which, in turn, is used to classify germplasm and assist in selection (MuñozAmatriaín et al. 2014; Milner et al. 2019). For heterogeneous accessions, whole genome data may also be profitably employed using the principle of *query*, to select accessions based on sequence variation at loci of interest. The nature of species is such that most variation is shared among populations, with the level of allelic diversity primarily dependent on mutation rate and population size (Kimura and Crow 1964). The distribution of allelic variation at loci under selection (e.g. agronomic loci) can deviate substantially from the predominantly neutral loci used for population structure analyses (Reeves et al. 2012). Poolseq pangenome data structures enable query and selection of accessions without explicit regard to population structure, accession provenance, or passport information, which may not be meaningful predictors of the occurrence of desirable sequence variation (Reeves and Richards 2018; Reeves et al. 2020).

Summary and visualization of whole genome data requires a reduction in complexity, and thus a reduction in accuracy when variation at specific sets of loci is desired. Major variant frequency variation displayed in Figure 2 is one such reduction; there are many other short haplotype variants that are not shown. However, a crop genome in its entirety can be mapped to orthologous sequence variants from its broader gene pool (Figure 1). All loci so mapped are accessible for improvement using sequence information from, in this case, a set of populations from the tertiary gene pool. To express this idea visually, for every vertical bar in Figure 1, a Figure 2 can be constructed. The resulting preprocessed data structure could be integrated into gene bank data bases to enable rapid query by homology or genome position for variant frequencies, opening up the possibility of selecting accessions based not only on passport data and population structure, but also by targeted query of sequence variation, at any locus. The FRA itself could be held in taxonspecific data bases, or with minor changes to the bioinformatic pipeline to retain quality scores, may qualify for submission to NCBI's Sequence Read Archive (SRA).

As crop improvement increasingly supplements conventional field breeding practices with *in vitro* techniques like transgenics and gene editing, the importance of accurate, comprehensive, sequencebased characterization of gene bank accessions grows. Knowledge of the full complement of sequences is important to ensure genome editing targets are present and to avoid offtarget effects (Danilevicz et al. 2020). Poolseq pangenomes allow collections to be characterized progressively, one accession at a time, which contrasts with population structurebased characterization, which is dependent on a particular referencebased variant calling pipeline. No reanalysis of existing data is required when poolseq pangenomic data are added for new accessions. Pooled sequencing data is therefore extensible at the level of accessions, but also at the level of the haplotype within accessions, because there is no conceptual barrier to adding singlemolecule long read data to existing phased short reads.

We have proposed one option for a queryready data structure that captures whole genome sequence variation for heterogeneous accessions, where representation by a single individual is inadequate. A data structure based on relatively unprocessed DNA sequence, closely representing the physical molecules from which it was constructed, is likely to be more "futureproof" than derived analytical products, and will provide novel opportunities for crop improvement as new analytical methods are developed.

Declarations

Acknowledgements

We thank Lothar Frese for collecting *Patellifolia* from the wild and sharing leaf tissue with us, Ann Fenwick for preparing the DNA pools, Mitch McGrath and Paul Galewski for coordinating DNA sequencing, and Kevin Dorn for comments on the manuscript. Mention of a trade name or proprietary product does not constitute endorsement by the USDA or a recommendation over other products that may be suitable. USDA is an equal opportunity provider and employer.

Funding: This work was supported by USDA-ARS National Program 301 Project 3012-21000-015-000-D.

Conflicts of interest/Competing interests: The authors have no conflicts of interest to declare that are relevant to the content of this article.

Availability of data and material: The data structures described here are available from the authors.

Code availability: https:/github.com/NCGRP. Repositories named in Materials and methods.

References

Baetscher DS, Clemento AJ, Ng TC, Anderson EC, Garza JC (2018) Microhaplotypes provide increased power from shortread DNA sequences for relationship inference. Mol Ecol Resour 18:296–305

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020) Plant pangenomes are the new reference. Nat Plants 6:914–920

Belzile F, Abed A, Torkamaneh D (2020) Time for a paradigm shift in the use of plant genetic resources. Genome 63:189–194

Biancardi E, McGrath JM, Panlla LW, Lewellen RT, Stevanato P (2010) Sugar Beet. In: Bradshaw JE (ed) Handbook of Plant Breeding 7: Root and Tuber Crops. Springer, Switzerland, pp 173–219

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Bushnell B, Rood J, Singer E (2017) BBMERGE—accurate paired shotgun read merging via overlap. PLoS ONE 12:e0185056. https://doi.org/10.1371/journal.pone.0185056

Castro S, Romeiras MM, Castro M, Duarte MC, Loureiro J (2013) Hidden diversity in wild *Beta* taxa from Portugal: Insights from genome size and ploidy level estimations using flow cytometry. Plant Sci 207:72–78

Danilevicz MF, Fernandez CGT, Marsh JI, Bayer PE, Edwards D (2020) Plant pangenomics: approaches, applications and advancements. Curr Opin Plant Biol 54:18–25

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127:1309–1321

Frese L, Nachtigall M, Iriondo JM, Teso MLR, Duarte MC, de Carvalho MÂAP (2019) Genetic diversity and differentiation in *Patellifolia* (Amaranthaceae) in the Macaronesian archipelagos and the Iberian Peninsula and implications for genetic conservation programmes. Genet Resour Crop Evol 66:225–241

Galewski P, McGrath JM (2020) Genetic diversity among cultivated beets (Beta vulgaris) assessed via populationbased whole genome sequences. BMC Genomics 21:189

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, BurzynskiChang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu Y, van der Knaap E, Huang S, Klee HJ, Giovannoni JJ, Fei Z (2019) The tomato pangenome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51:1044–1051

Gur A, Zamir D (2004) Unused natural variation can lift yield barriers in plant breeding. PLoS Biol 2:1610–1615

Hajjar R, Hodgkin T (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica 156:1–13

Hawkes JG (1977) The importance of wild germplasm in plant breeding. Euphytica 26:615–621

Höft N, Dally N, Hasler M, Jung C (2019) Haplotype variation of glowering time genes of sugar beet and its wild relatives and the impact on life cycle regimes. Front Plant Sci 8:2211. doi: 10.3389/fpls.2017.02211

Hu Z, Olatoye MO, Marla S, Morris GP (2019) An integrated genotyping by sequencing polymorphism map for over 10,000 sorghum genotypes. Plant Genome 12:180044

Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, Lee JS, Baute GJ, Owens GL, Grassa CJ, Ebert DP, Ostevik KL, Moyers BT, Yakimowski S, Masalia RR, Gao L, Ćalić I, Bowers JE, Kane NC, Swanevelder DZH, Kubach T, Muños S, Langlade NB, Burke JM, Rieseberg LH (2019) Sunflower pangenome analysis shows that hybridization altered gene content and disease resistance. Nat Plants 5:54–62

Kidd KK, Pastis AJ, Speed WC, Lagacé R, Chang J, Wootton S, Haigh E, Kidd JR (2014) Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci Int Genet 12:215–224

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Kumar A, Harloff HJ, Melzer S, Leineweber J, Defant B, Jung C (2021) A rhomboidlike protease gene from an interspecies translocation confers resistance to cyst nematodes. New Phytol. https://doi.org/10.1111/nph.17394

Kuroda Y, Takahashi H, Okazaki K, Taguchi K (2019) Molecular variation at BvBTC1 is associated with bolting tolerance in Japanese sugar beet. Euphytica 215:43. doi: 10.1007/s10681-019-2366-9

Lemmon ZH, Reem NT, Dalrymple J, Soyk S, Swartwood KE, RodríguezLeal D, Van Eck J, Lippman ZB (2018) Rapid improvement of domestication traits in an orphan crop by genome editing. Nat Plants 4:766–770

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Li T, Yang X, Yu Y, Si X, Zhai X, Zhang H, Dong W, Gao C, Xu C (2018) Domestication of wild tomato is accelerated by genome editing. 36:1160–1163

Li X, Brummer EC (2009) Inbreeding depression for fertility and biomass in advanced generations of inter and intrasubspecific hybrids of tetraploid alfalfa. Crop Sci 49:13–19

Lindhout P, Meijer D, Schotte T, Hutten RCB, Visser RGF, van Eck HJ (2011) Towards F_1 hybrid seed potato breeding. Potato Res 54:301–312

Mascher M, Schreiber M, Scholz U, Graner A, Reif JC, Stein N (2019) Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. Nat Genet 51:1076–1081

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963

McCouch SR, McNally KL, Wang W, Sackville Hamilton R (2012) Genomics of gene banks: a case study in rice. Am J Bot 99:407–423

McCouch S, Navabi K, Abberton M, Anglin NL, Barbieri RL, Baum M, Bett K, Booker H, Brown GL, Bryan GJ, Cattivelli L, Charest D, Eversole K, Freitas M, Ghamkhar K, Grattapaglia D, Henry R, Valadares Inglis MC, Islam T, Kehel Z, Kersey PJ, Kresovich S, Marden E, Mayes S, Ndjiondjop MN, Nguyen HT, Paiva S, Papa R, Phillips PWB, Rasheed A, Richards C, Rouard M, Amstalden Sampaio MJ, Scholz U, Shaw PD, Sherman B, Staton SE, Stein N, Svensson J, Tester M, Montenegro Valls JF, Varshney R, Visscher S, von Wettberg E, Waugh R, Wenzl PWB, Rieseberg LH (2020) Mobilizing crop biodiversity. Mol Plant 13:1341–1344

McGrath JM, Funk A, Galewski P, Ou S, Townsend B, Davenport K, Daligault H, Johnson S, Lee J, Hastie A, Darracq A, Willems G, Barnes S, Liachko I, Sullivan S, Koren S, Phillippy A, Wang J, Liu T, Pulman J, Childs K, Yocum A, Fermin D, Mutasa-Göttgens E, Stevanato P, Taguchi K, Dorn K (2020) A contiguous *de novo* genome assembly of sugar beet EL10 (*Beta vulgaris* L.) bioRxiv 2020.09.15.298315; doi: https://doi.org/10.1101/2020.09.15.298315

Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, Schüler D, Sharma R, Pasam RK, Rutten T, Guo G, Xu D, Zhang J, Herren G, Müller T, Krattinger SG, Keller B, Jiang Y, González MY, Zhao Y, Habekuß A, Färber S, Ordon F, Lange M, Börner A, Graner A, Reif JC, Scholz U, Mascher M, Stein N (2019) Genebank genomics highlights the diversity of a global barley collection. Nat Genet 2019:319–326

MuñozAmatriaín M, CuestaMarcos A, Endelman JB, Comadran J, Bonman JM, Bockelman HE, Chao S, Russel J, Waugh R, Hayes PM, Muehlbauer GS (2014) The USDA barley core collection: genetic diversity, population structure, and potential for genomewide association studies. PLoS ONE 9:e94688. doi:10.1371/journal.pone.0094688

Pin PA, Zhang W, Vogt SH, Dally N, Büttner B, SchulzeBuxloh G, Jelly NS, Chia TYP, MutasaGöttgens ES, Dohm JC, Himmelbauer H, Weisshaar B, Kraus J, Gielen JJL, Lommel M, Weyens G, Wahl B, Schechert A, Nilsson O, Jung C, Kraft T, Müller AE (2012) The role of a pseudoresponse regulator gene in life cycle adaptation and domestication of beet. Curr Biol 22:1095–1101

Reeves PA, Panella LW, Richards CM (2012) Retention of agronomically important variation in germplasm core collections: implications for allele mining. Theor Appl Genet 124:1155–1171

Reeves PA, Richards CM (2018) Biases induced by using geography and environment to guide ex situ conservation. Conserv Genet 19:1281–1293

Reeves PA, Tetreault HM, Richards CM (2020) Bioinformatic extraction of functional genetic diversity from heterogeneous germplasm collections for crop improvement. Agronomy 10:593. doi:10.3390/agronomy10040593

RodríguezLeal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB (2017) Engineering quantitative trait variation for crop improvement by genome editing. Cell 171:470–480

Rojas MC, Pérez JC, Ceballos H, Beina D, Morante N, Calle F (2009) Analysis of inbreeding depression in eight S_1 cassava families. Crop Sci 49:543–548

Romeiras MM, Vieira A, Silva DN, Moura M, SantosGuerra A, Batista D, Duarte MC, Paulo OS (2016) Evolutionary and biogeographic insights on the Macaronesian *BetaPatellifolia* species (Amaranthaceae) from a timescaled molecular phylogeny. PLoS ONE 11:e0152456. doi:10.1371/journal.pone.0152456

Scheben A, Edwards D (2018) Towards a more predictable plant breeding pipeline with CRISPR;Casinduced allelic series to optimize quantitative and qualitative traits. Curr Opin Plant Biol 45:218–225

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals—mining genomewide polymorphism data without big funding. Nat Genet 15:749–763

Sundaram AYM, Garseth ÃH, Maccari G, Grimholt U (2020) An Illumina approach to MHC typing of Atlantic salmon. Immunogenetics 72:89–100

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P (2015) Sambamba: fast processing of NGS alignment formats. Bioinformatics 31:2032–2034

Tripodi P, Rabanus-Wallace MT, Barchi L, Kale S, Esposito S, Acquadro A, Schafleitner R, van Zonneveld M, Prohens J, Diez MJ, Börner A, Salinier J, Caromel B, Bovy A, Boyaci F, Pasev G, Brandt R, Himmelbach A, Portis E, Finkers R, Lanteri S, Paran I, Lefebvre V, Giuliano G, Stein N (2021) Global range expansion history of pepper (*Capsicum* spp.) revealed by over 10,000 genebank accessions. PNAS 118:e2104315118

Volk GM, Byrne PF, Coyne CJ, FlintGarcia S, Reeves PA, Richards C (2021) Integrating genomic and phenomic approaches to support plant genetic resources conservation and use. Plants 10:2260. https://doi.org/10.3390/plants10112260

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann JC, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Weise S, Lohwasser U, Opermann M (2020) Document or lose it—on the importance of information management for genetic resources conservation in genebanks. Plants 9:1050

Wolter F, Schindele P, Puchta H (2019) Plant breeding at the speed of light: the power of CRISPR/Cas to generate directed genetic diversity at multiple sites. BMC Plant Biol 19:176

Wu D, Liang Z, Yan T, Xu Y, Xuan L, Tang J, Zhou G, Lohwasser U, Hua S, Wang H, Chen X, Wang Q, Zhu L, Maodzeka A, Hussain N, Li Z, Li X, Shamsi IH, Jilani G, Wu L, Zheng H, Zhang G, Chalhoub B, Shen L, Yu H, Jiang L (2019) Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. Mol Plant 12:30–43

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29:2669–2677

Zsögön A, Čermák T, Naves ER, Notini MM, Edel KH, Weinl S, Freschi L, Voytas DF, Kudla J, Peres LPP (2018) *De novo* domestication of wild tomato using genome editing. Nat Biotechnol 36:1211–1216

Figures

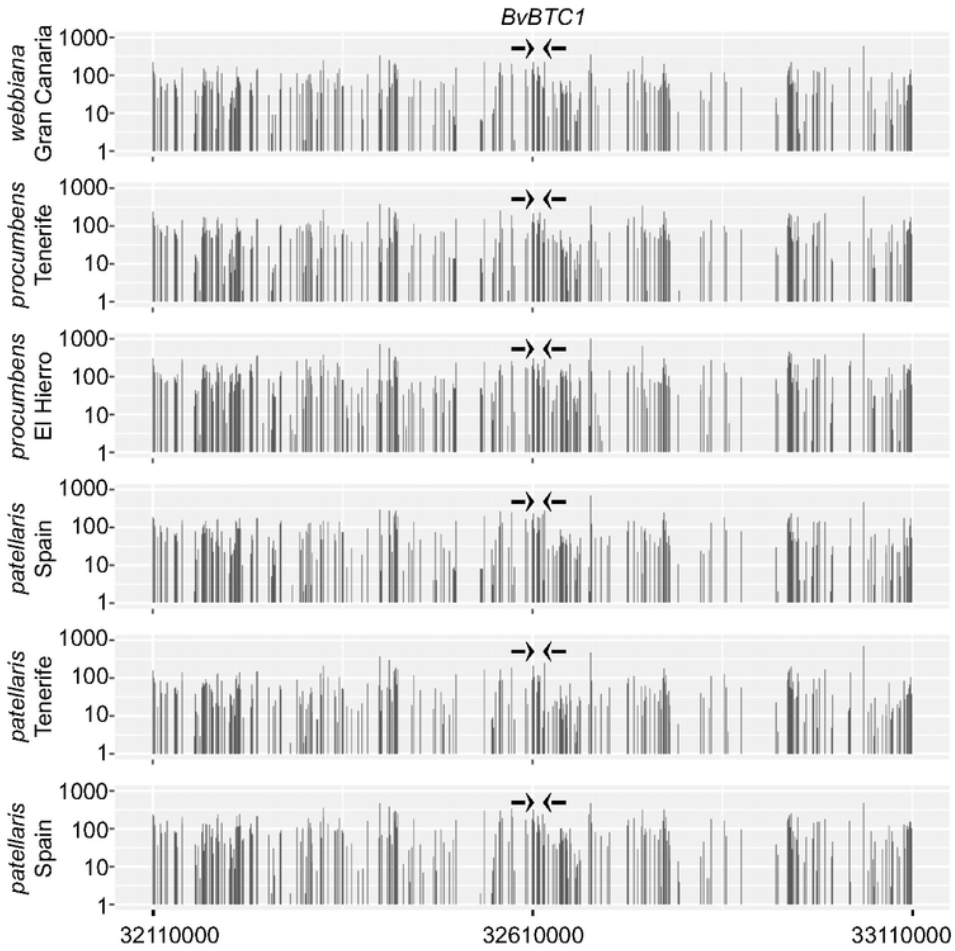

Figure 1.

Figure 1

Depth of *Patellifolia* phased reads at their orthologous map position in the sugar beet genome. A 1 Mbp portion of EL10 chromosome 2 containing the bolting gene *BvBTC1* is shown. Vertical bars are 1 Kbp in length and correspond to regions of the EL10 genome present as single copy contigs in *Patellifolia* pool genome assemblies. Height of bars indicates the number of phased reads from the *Patellifolia* pool that map to each 1 Kbp region. Approximately 1/5 of the sugar beet genome is present as single copy sequence in *Patellifolia*.

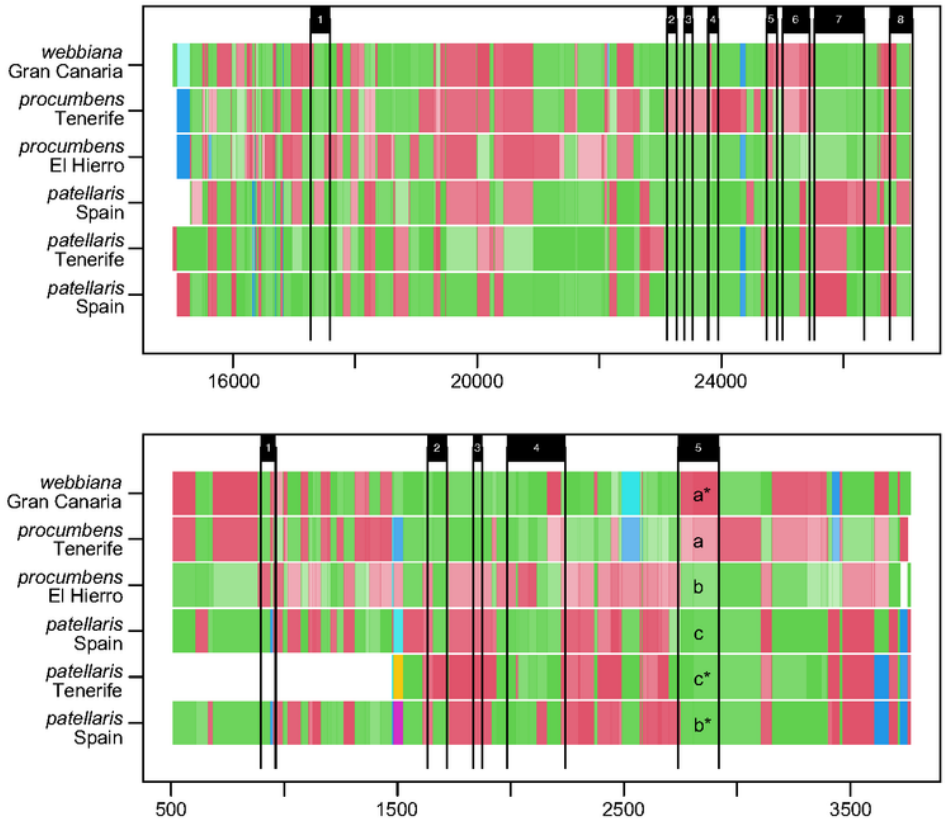

Figure 2.

Figure 2

Major variant frequencies at short haplotype loci across the *Patellifolia* ortholog of the sugar beet bolting gene *BvBTC1* (top) and the cyst nematode resistance gene *Hs4* (bottom). Exons labeled at top. Colors indicate different variants, shading within a color indicates variant frequency (lighter = lower). Only those loci where the major variant differed between the six pools are shown. A portion of *Hs4* containing exon 1 was absent from the Tenerife *P. patellaris* pool assembly. For each pool, the horizontal bar pattern visualizes a probability distribution expressing the likelihood of sampling the most common allelic variant, at each position in the gene. Three *Hs4* exon 5 major variants exist (a,b,c), and can be recovered most efficiently by sampling the *P. webbiana* and two *P. patellaris* pools (marked with an asterisk).

Supplementary Files

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

- SuppFigure1.png