A pan-genome data structure induced by pooled sequencing facilitates variant mining in heterogeneous germplasm

Patrick A. Reeves · Christopher M. Richards

Received: 25 January 2022 / Accepted: 8 June 2022 / Published online: 25 June 2022

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 pan-genome data structure induced by whole-genome 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 pan-genome as a map of reads from pooled sequencing of a heterogeneous population sample to a reference genome, plus a BLAST data base of the mapped reads. We show that this basic data structure can be queried by reference genome position or homology to identify sequence variants present in the wild relative, at genes of agronomic interest in the crop, a process known as allele or variant mining. Further we demonstrate the possibility of cataloging variants in all *Patellifolia* genomic regions that have corresponding single copy orthologous regions in sugar beet. The data structure, termed a "pooled read archive," can be produced, altered, and queried using standard tools to facilitate discovery of agronomically-important sequence variation.

Keywords Bioinformatics · Crop wild relatives · Domestication · Genome editing · Haplotype · Phasing · Sequence variant

Introduction

Germplasm collections are a source of novel genetic variants for crop improvement. 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).

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s11032-022-01308-6.

P. A. Reeves (✉) · C. M. Richards
Agricultural Research Service, United States Department of Agriculture, National Laboratory for Genetic Resources Preservation, 1111 South Mason Street, Fort Collins, CO 80521, USA
e-mail: pat.reeves@usda.gov

In conventional breeding, germplasm collections are explored for desirable traits biologically, via controlled crosses in vivo. For some crops and some traits, it may be more efficient to explore germplasm collections informatically, to identify useful sequence variation in silico. In breeding programs where the trait of interest is well-defined genetically, meaning most genes impacting it 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 target-gene-specific generation of genetic diversity (Rodríguez-Leal et al. 2017; Scheben and Edwards 2018; Belzile et al. 2020).

Currently there is momentum to commence systematic, whole-genome, whole-collection genotyping of gene bank holdings (McCouch et al. 2020). 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 a 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 genome-wide 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' pan-genome (Gao et al. 2019). Due to widespread structural variation, the non-coding dispensable fraction may be even higher (Hübner et al. 2019). In *Beta vulgaris*, for example, flow cytometric estimates of genome size varied from 633 to 875 MB, a 40% difference (Castro et al. 2013; McGrath et al. 2020). For heterogeneous accessions, genome-wide 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 involved in describing 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). A catalog of the wide range of sequence variants present in unimproved germplasm could support molecular breeding initiatives to generate diversity in elite cultivars using transgenic or genome editing approaches. Variants present alternative, naturally-occurring models for site-specific base editing in a crop, or target-sequence 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. 2018). Additionally, for conventional breeding projects, a catalog could support prediction of the frequency of particular traits or trait values in accessions. For traits that are well-understood genetically, population phenotypic variation can be predicted from variant frequencies at controlling genes. Prediction of phenotypic variation based on measurement of segregating genotypic variation could enable initial selection of accessions for integration into a conventional breeding program bioinformatically, instead of via extensive grow-outs and phenotyping or imprecise suggestions from passport data and surrogate predictors of genetic diversity (Reeves et al. 2020).

In this study, we demonstrate one approach for cataloging DNA sequence variants from heterogeneous collections of individuals, to inform crop molecular breeding. Our motivation is to improve access to potentially valuable variation in crop wild relatives in the secondary and tertiary gene pools, where biological interrogation of variation using crosses is difficult due to reproductive barriers, as well as primary gene pool members where complex growth requirements or inbreeding depression prevents extensive development of inbred lines. We describe the pan-genome-like properties of data acquired by whole genome sequencing of pools of individuals ("pooled sequencing" or "pool-seq"). We develop a data structure that supports query of this "pool-seq pan-genome" by sequence homology or genome position. Through query, we calculate the proportion of genes expressed in sugar beet (*Beta vulgaris*) for which orthologous variation can be found in *Patellifolia* spp. (Thulin et al. 2010), a distant wild relative that diverged from *Beta* ~ 25 Mya (Romeiras et al. 2016). We likewise calculate the proportion of the entire sugar beet genome with orthologous sequence in *Patellifolia*—the fraction for which variant mining is uncomplicated by differences in gene content. We catalog whole-genome multi-allelic short haplotype (SH) variation (akin to "microhaplotype" sensu Kidd et al. 2014 and Baetscher et al. 2018) for six heterogeneous samples of *Patellifolia* spp. Using the catalog, we describe orthologous sequence variants for two sugar beet loci of agronomic importance found in the wild relative.

Materials and methods

Pooled DNA sequencing

Leaf tissue from 17 to 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), concentration-normalized to 20 ng/µL, and pooled for DNA sequencing. Sequencing libraries were generated from DNA pools using the KAPA HyperPrep Kit (Roche, Basel, Switzerland) with a PCR-free 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).

Data structure construction

Detailed bioinformatic procedures including all software settings are at https://github.com/NCGRP/mb1suppl. MASURCA 3.2.4 (Zimin et al. 2013) was used to produce a genome assembly for each pool (hereafter, "pool assembly"). Trimmed read pairs

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 A	37.557349, – 1.168413	25
patellaris	Tenerife	28.376967, – 16.799400	25
patellaris	Spain B	37.504414, – 1.425755	25

Table 2 DNA sequencing and pool genome assembly

Species	Location	NCBI SRA (raw reads)	Read pairs	Assembly size (Mbp)	Coverage (per individual)	Contigs	N50
webbiana	Gran Canaria	SRX6944498	2.2E8	747	88x (3.5x)	258,036	13,789
procumbens	Tenerife	SRX6944497	2.2E8	790	84x (4.9x)	285,422	13,366
procumbens	El Hierro	SRX6944496	3.4E8	790	129x (5.2x)	271,266	12,459
patellaris	Spain A	SRX6944495	2.3E8	1114	62x (2.5x)	162,082	17,866
patellaris	Tenerife	SRX6944494	1.8E8	1136	48x (1.9x)	202,527	16,679
patellaris	Spain B	SRX6944492	3.2E8	1090	88x (3.5x)	129,336	20,349

(TRIMMOMATIC 0.33, Bolger et al. 2014) were mapped to the pool assembly using BWA-MEM 0.7.17 (Li 2013) to confirm proximity, filtered by quality using SAMTOOLS 1.8 (Li et al. 2009), and duplicates removed with SAMBAMBA 0.7.0 (Tarasov et al. 2015) before combining into "phased reads," sometimes called merged reads or "FLASHed reads" (Bushnell et al. 2017; Sundaram et al. 2020). We use the term phased read to refer to a single sequence derived from two or more sequences known to originate from the same physical DNA molecule. In the case of Illumina paired end sequencing, read phasing is a trivial operation of combining the two opposing reads of a read pair into a single sequence, since they derive from opposite ends of the same molecule.

Reads from each *Patellifolia* population were filtered by alignment quality, proximity, and orientation, using the population's pool assembly as a reference genome. Reads were combined into a single sequence (creating a phased read) when both reads of a pair 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. If read pairs met these criteria but were non-overlapping then the intervening region was padded with a string of Ns of a length predicted from the pool assembly contigs to which they mapped. Read pairs that could not be combined into a single sequence by this procedure (unphased reads) were retained because they contain much additional information, albeit within shorter sequences. The reference-free 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).

For each *Patellifolia* population, a binary alignment map file (BAM file) specifying the map of processed read pairs (including phased and unphased reads, hereafter "reads") onto pool assembly contigs was produced. Mapped reads and their associated pool assembly contigs were processed into a single sorted multi-FASTA file. A BLAST nucleotide database was constructed for the multi-FASTA file and for the pool assembly alone using BLAST + 2.5.0. We define a data structure, hereafter referred to as a "pooled read archive" (PRA), that contains (1) an indexed reference genome, (2) a BAM map between pool-seq reads and the reference genome, (3) a sorted FASTA file containing pool-seq reads 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 sequence variation from pooled sequence data by genome position (e.g., SAMTOOLS) or homology (BLAST).

Data structure evaluation

We evaluated PRA quality by calculating average read length and coverage of the pool assembly. We explored the utility of *Patellifolia* PRAs as a source of variants for sugar beet genes by determining the proportion of the sugar beet transcriptome with homologous sequence in *Patellifolia* pools. PRA BLAST databases were queried with all 24,255 primary transcripts in the sugar beet EL10_1.0 transcriptome (McGrath et al. 2020, https://phytozome-next.jgi.doe.gov/info/Bvulgaris_EL10_1_0) with up to 100 K matches returned per query. BLASTN results were filtered to exclude gene models that matched < 40 and > 1000 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 Fig. 1). Remaining matches were considered to represent the set of homologous genes, excluding highly repetitive (> 1000 hits), unmatched (0 hits), and poorly represented (1–39 hits) genes. We defined orthologous genes operationally, using BLASTN, as EL10_1.0 transcript queries from the homologous set 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* pan-genome represented in the pool-seq data that was homologous to the sugar beet genome. The nine chromosome scaffolds from sugar beet genome assembly EL10_1.0 (NCBI GCA_002917755.1) were fragmented into sequential 1 Kbp sequences, each of which was then used as a BLASTN query against the PRA for the purposes of determining orthology, as was done with the transcriptome except that up to 10 K matches were allowed to be returned with no subsequent filter on read depth per query applied, in order to retain information on repetitive sequences.

Variant mining

To demonstrate the capacity of pooled sequencing data to facilitate variant mining, we characterized variant frequencies within *Patellifolia* pools at agronomically-important cyst nematode resistance gene *Hs4* (Kumar et al. 2021) and the *Patellifolia* ortholog of pseudo-response regulator *BvBTC1*, which determines annual versus biennial life cycle in sugar beet (Pin et al. 2012). Full length mRNA sequences were used to query PRA BLAST databases using BLASTN to identify *Patellifolia* contigs containing *Hs*4 and *BvBTC1*, along with the reads mapped to those contigs, as contained in the PRA BAM file. For detailed analysis and visualization, a gene 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 haplotype loci, defined here as short genomic regions containing one or more SNPs segregating as haplotypes (based on Baetscher et al. 2018), were identified, and major variant frequencies were estimated across each gene region along a tiling path that maximized locus variation, accuracy, and length, in order to simplify presentation of results (details in Supplementary Information).

Results

Pooled DNA sequencing and data structure evaluation

Pool genome assemblies varied in size, averaging ~ 775 Mbp in diploids *P. webbiana* and *procumbens*, ~ 1.1 Gbp in tetraploid *P. patellaris* (Table 2). Estimated coverage per individual in the raw data pools varied from 1.9 to 5.2x. Pooled sequencing coverage > 1 × per individual produces allele frequency estimates that are equal to or more accurate than those computed from sequencing individuals (Schlötterer et al. 2014). Pool assemblies were fragmented, containing on average 271,575 contigs with mean N50 ~ 13 Kbp for diploid pools, 164,648 contigs with mean N50 ~ 18 Kbp for tetraploids (Table 2). After filtering during production of the PRA, coverage per individual remained above 1x (1.4–3.5x). Average processed read length ranged from 251 to 275 bp, a substantial increase from the initial 150 bp reads (Table 3). Complete PRA data structures occupied 222 ± 49 GB (± 1 SD) of disk space on average; compressed raw reads occupied 44 ± 16 GB. PRA data structures used in this study are available upon request from the authors.

Of the 24,255 primary transcripts in the EL10_1.0 transcriptome, 1422 were determined to be highly repetitive (chloroplast, rDNA, and mitochondrial genes, plus gene models with repeated amino acid motifs), 3793 unmatched, and 2130 poorly

Table 3 Evaluation of PRA data structures including proportion of sugar beet transcriptome and genome with homologous sequence reads in wild relative *Patellifolia* spp

Species	Location	Read coverage, per individual (processed read count)	Average read length, bp	Proportion of EL10_1.0 transcriptome found in *Patellifolia* (transcript count)	Proportion of EL10_1.0 transcriptome single copy in *Patellifolia* (transcript count)	Proportion of EL10_1.0 genome found in *Patellifolia*	Proportion of EL10_1.0 genome single copy in *Patellifolia*
webbiana	Gran Canaria	2.4x (1.8E8)	251	0.71 (17,179)	0.36 (8617)	0.72	0.17
procumbens	Tenerife	3.5x (1.7E8)	273	0.71 (17,104)	0.32 (7797)	0.72	0.16
procumbens	El Hierro	3.5x (2.6E8)	268	0.70 (16,909)	0.29 (6953)	0.80	0.15
patellaris	Spain A	1.8x (1.8E8)	273	0.70 (16,995)	0.28 (6721)	0.74	0.21
patellaris	Tenerife	1.4x (1.4E8)	273	0.69 (16,784)	0.25 (6031)	0.72	0.21
patellaris	Spain B	2.5x (2.4E8)	275	0.69 (16,779)	0.30 (7289)	0.75	0.21

represented in *Patellifolia* pools (Supplemental Fig. 1). Thus, 70% of the sugar beet transcriptome had homologous sequence regions in *Patellifolia*. Twenty-five to thirty percent was single copy, here considered orthologous, and therefore a straightforward target for variant mining. Similarly, homologous sequence for 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 accessible for improvement using sequence diversity mined from *Patellifolia*, without the complicating issue of paralogy (Fig. 1). Put differently, no homologous counterpart for about 25% of sugar beet genome EL10_1.0 was found in the *Patellifolia* pan-genome using our procedure.

Variant mining

BLASTN query of PRA BLAST databases using *BvBTC1* and *Hs4* coding sequences yielded 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 reads were mapped to the ~ 12 Kbp genomic region containing *BvBTC1* (depth 34x–122x); 296–812 reads mapped to the ~ 3 Kbp region containing *Hs4* (24x–66x). The tiling paths across *BvBTC1* and *Hs4* contained 3220 and 754 SH loci with a mean length of 2.99 ± 1.39 bp and 3.79 ± 2.58 bp, respectively. The number of variants per SH 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. Eighty-two percent of *BvBTC1* SH loci comprised indel variants only, 18% contained single-nucleotide or multi-nucleotide polymorphisms (MNPs). For *Hs4*, 80% of loci were indel-only;

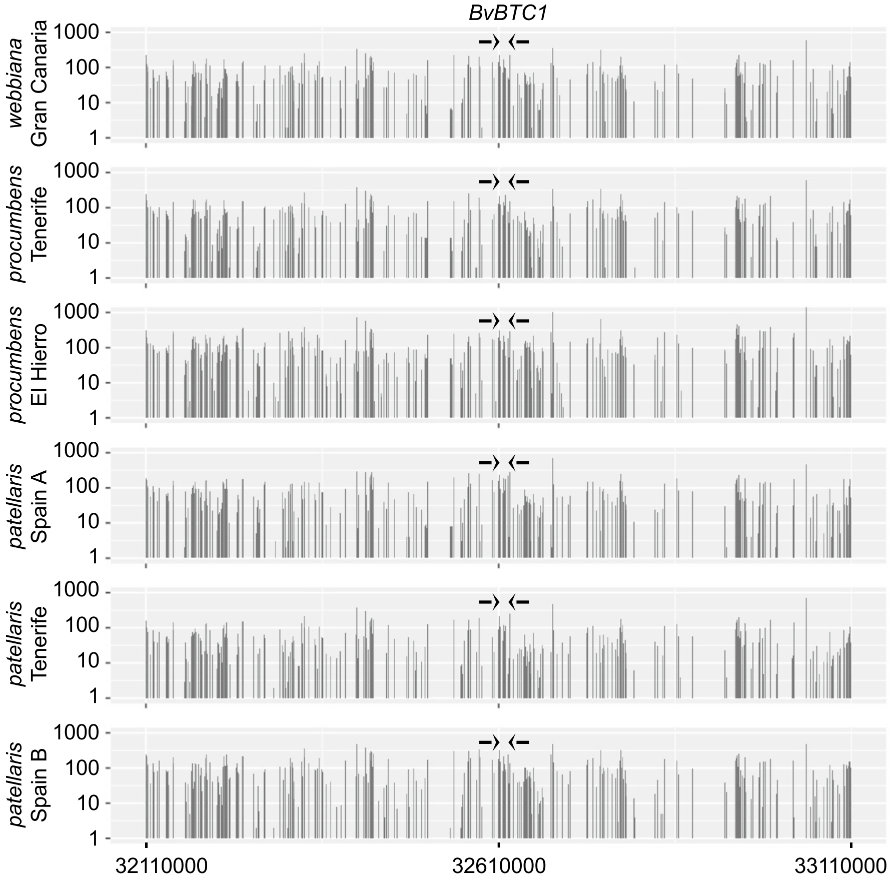

Fig. 1 Depth of *Patellifolia* reads at their orthologous map position in the sugar beet genome. A 1 Mbp span of sugar beet EL10_1.0 chromosome 2 is shown. This region contains the bolting gene *BvBTC1*, the boundaries of which are marked by opposing arrows in each plot. Vertical bars each represent 1 Kbp of sequence along the 1 Mbp span of chromosome 2. Height of bars indicates the number of reads that map to the 1 Kbp region. Portions of the 1 Mbp span with no bars plotted represent parts of sugar beet EL10_1.0 chromosome 2 with no orthologous counterpart in *Patellifolia*. Approximately 1/5 of the sugar beet genome can be found as single copy sequence in *Patellifolia*, including the region containing the bolting gene *BvBTC1*

20% contained SNPs or MNPs. Among pools, the major variant frequency ranged from 0.87 to 0.91 for *BvBTC1* and 0.86 to 0.94 for *Hs4*. Major variant differences between pools across the genes are visualized in Figs. 2 and 3. Per-pool descriptive statistics are in Table 4.

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 (Weise 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).

The pool-seq pan-genome

Pooled sequencing, the process of sequencing DNA from multiple individuals simultaneously, induces a pan-genome-like data structure in its output, with sequence variation captured as independent reads derived from different individuals in the

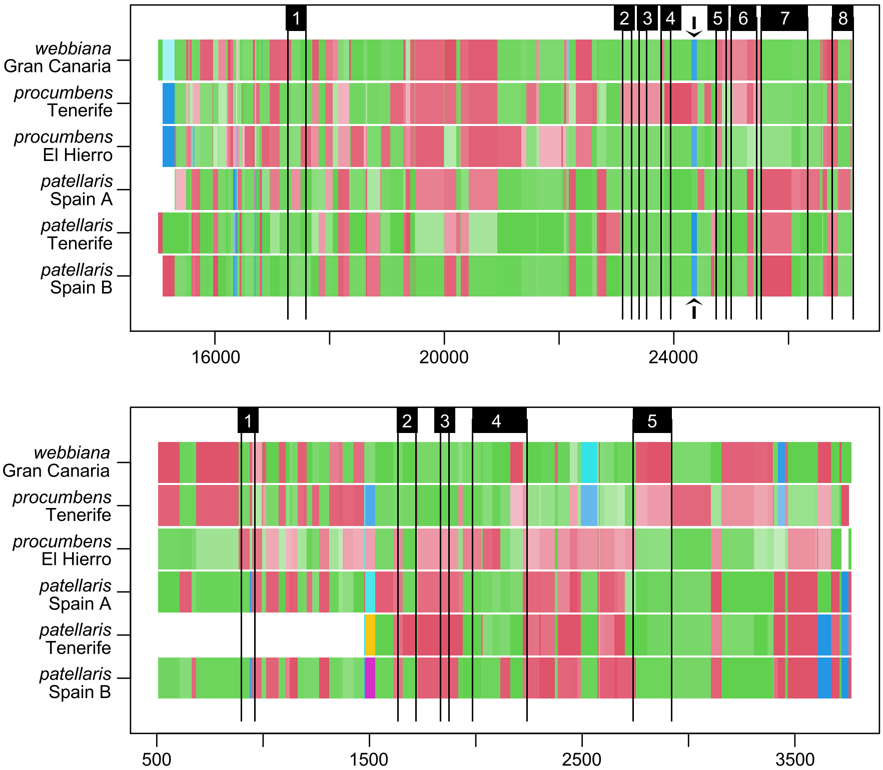

Fig. 2 Major variant frequencies at short haplotype (SH) loci across the *Patellifolia* ortholog of the sugar beet bolting gene *BvBTC1* (top) and the cyst nematode resistance gene *Hs4* (bottom). Exons labeled at top. Colored bars indicate different variants within an SH locus; shading within a color indicates variant frequency (lighter = lower). Only loci where the major variant differed between the six pools are shown. For better visualization, bars have been widened to cover intervening regions where the major variant was identical between pools (see Fig. 3). A single locus with three major variants (red and green found in one pool each, blue in four pools) in intron 4 of *BvBTC1* is marked using arrows to show how SH loci appear as bars of the same width, stacked vertically among the pools. Within a pool, the horizontal assemblage of bars shows which variant is the major one, and its probability of being sampled, by virtue of shading, relative to the other pools with the same major variant. A portion of *Hs4* was missing from the Tenerife *P. patellaris* pool assembly

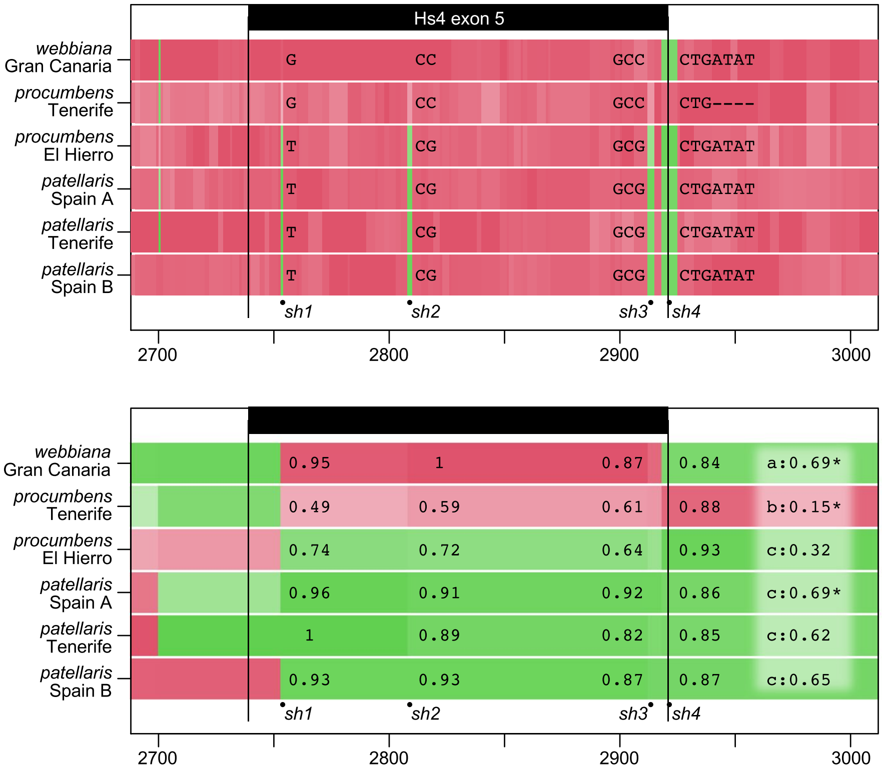

Fig. 3 Major variant frequencies at short haplotype (SH) loci across exon 5 of *Hs4* in *Patellifolia* pools. Colors indicate different variants; shading indicates frequency. Top panel shows major variants, width of bars is proportional to number of bp in the SH locus, and faint white lines mark locus boundaries. Most major variants at SH loci within exon 5 are identical among pools. Four SH loci where the major variant differed between the six pools were found, labeled *sh1–4* with positions marked using a bullet at bottom of panel. In the top panel, variant sequences are shown to the right of the colored bars for *sh1*, *sh2*, and *sh4*, and to the left for *sh3*. To aid visualization, in the bottom panel, colored bars for *sh1–4* have been widened to the right of the locus start point, until they contact the next SH locus with a major variant difference between pools, as in Fig. 2. Major variant frequency is indicated. The three *Hs4* exon 5 major variant haplotypes (a, b, c) and their minimum probability of being drawn from the pool is shown at right, calculated as the joint probability of drawing the four contributing SH locus major variants simultaneously. *Hs4* exon 5 haplotypes can be recovered most efficiently by sampling the pools marked with an asterisk. The joint probability of drawing haplotype b is low, recommending increased sampling intensity of the *procumbens*-Tenerife population in a variant mining experiment intent on recovering major variant haplotypes of *Hs4* exon 5

pool. Reads may vary in length and configuration depending on sequencing technology, but, assuming PCR-free library construction, should represent actual sequence variants at the approximate frequency they occur in the pool of individuals (Lynch et al. 2014; Schlötterer et al. 2014). This form of pan-genome 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 a pan-genome, albeit as short physical fragments instead of analytically derived contigs, scaffolds, or chromosome length pseudomolecules.

For cataloging sequence variation, the “pool-seq pan-genome” is a cost-effective and appropriate alternative to sequencing many individuals separately. In outcrossed wild populations, all individuals are expected to have distinct genome sequences due to recombination of standing variation during sexual reproduction history. Individual multi-locus genotypes and long DNA sequence

Table 4 Short haplotype (SH) variation present in *Patellifolia* pools across *BvBTC1* and *Hs4*

Species	Location	Reads mapped to *BvBTC1* (depth)	Reads per SH locus (±1 SD), *BvBTC1*	Variants per SH locus, *BvBTC1*	Major variant frequency, *BvBTC1*	Indel-only variants, *BvBTC1* (proportion)	S/MNPs[1], *BvBTC1* (proportion)	Reads mapped to *Hs4* (depth)	Reads per SH locus (±1 SD), *Hs4*	Variants per SH locus, *Hs4*	Major variant frequency, *Hs4*	Indel-only variants, *Hs4* (proportion)	S/MNPs[1], *Hs4* (proportion)
webbiana	Gran Canaria	3608 (77x)	49.8 ± 9.1	2.23 ± 0.72	0.91 ± 0.06	2764 (0.97)	93 (0.03)	568 (43x)	26.6 ± 6.6	1.62 ± 0.66	0.94 ± 0.08	367 (0.93)	27 (0.07)
procumbens	Tenerife	3442 (79x)	47.6 ± 8.4	2.37 ± 0.69	0.88 ± 0.07	2972 (0.97)	103 (0.03)	812 (66x)	35.3 ± 7.1	2.23 ± 0.75	0.87 ± 0.10	609 (0.91)	62 (0.09)
procumbens	El Hierro	5384 (122x)	70.9 ± 11.0	2.91 ± 0.94	0.87 ± 0.08	2977 (0.94)	198 (0.06)	786 (63x)	38.2 ± 8.5	2.47 ± 0.89	0.86 ± 0.11	600 (0.88)	79 (0.12)
patellaris	Spain A	2071 (48x)	30.2 ± 6.5	2.00 ± 0.62	0.89 ± 0.07	2606 (0.99)	32 (0.01)	584 (47x)	28.7 ± 6.5	1.94 ± 0.63	0.90 ± 0.08	577 (0.98)	14 (0.02)
patellaris	Tenerife	1493 (34x)	19.9 ± 5.0	1.65 ± 0.62	0.91 ± 0.10	1837 (0.99)	15 (0.01)	296 (24x)	9.7 ± 7.7	1.01 ± 0.85	0.92 ± 0.09	469 (0.99)	5 (0.01)
patellaris	Spain B	2534 (59x)	35.4 ± 6.6	2.07 ± 0.53	0.89 ± 0.06	2885 (0.99)	15 (0.01)	698 (57x)	30.8 ± 8.7	1.97 ± 0.66	0.90 ± 0.07	591 (0.98)	13 (0.02)

[1] Single- or multi-nucleotide polymorphisms

haplotypes (those beyond typical linkage disequilibrium decay distances) may not be easily recovered from the population more than once because they are broken up by recombination. A catalog of SH loci across the genome can represent what variants are present, and at what frequency they will be encountered, when sampling a population. A similar catalog of SH loci could be constructed by sequencing individuals separately, but only with added effort and expense relative to pooled sequencing.

Recombinatorial introgression of desirable variants via crossing is often not an option when working with secondary or tertiary gene pools. But in cases where introgressive breeding is possible, the pool-seq pan-genome expresses what is delivered when a heterogeneous gene bank accession is requested: the capacity to recombine into a target line some proportion of the sequence variation segregating in the accession. In other words, the data are an approximate representation of the gamete pool. Thus, a pool-seq pan-genome 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 PRA to enable query of a pool-seq pan-genome involves three main steps: read phasing, mapping of reads to a reference assembly, and production of BLAST data bases. Read phasing is useful 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 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 reads is preferred because they represent physical molecules present in the pool that come from a single individual, with no reference to bioinformatically imputed 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 reads.

Pooled sequencing data as a gene bank breeding resource

Application: variant mining

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). Sequence variants “mined” from *BvBTC1* might be useful for improving early-bolting resistance, affecting decisions regarding planting time.

The *Patellifolia* ortholog of *BvBTC1* was found by querying the PRA pool assembly. Examination of mapped reads revealed many differences in SH locus major variants between pools across the genic region (Fig. 2). Because major variant frequencies were generally high (Table 4), one could recover variants of interest with high probability by re-sampling the biological material from which the pooled sequencing data were drawn. Moreover, for most major variants in this set of six pools, only two alleles were found (visualized in Fig. 2 using red and green) which could be recovered with relative ease, by accessing only two populations.

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 hold for all extended gene pools—depending on initial polyploidy events ("allo" vs "autopolyploidy") and evolutionary processes affecting homeologous regions since then, it may not always be possible to distinguish homeologs.

Polyploids are an underutilized resource for mining sequence variants due to technical difficulties encountered in polymorphism assessment. Since phasing beyond the limits of a single read pair is not possible, within the pooled sequencing context a tetraploid individual is no different from two diploids—regardless of parental genome, sequence reads can be recovered from the PRA. But, accurately mapping reads to the correct parental genome will not be possible unless homeologs can be distinguished, making variant frequency estimation impossible in some cases.

Application: germplasm selection

The pool-seq pan-genome 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 us suppose one was especially interested in *Hs4* exon 5. We show that there are three major haplotypic variants covering that ~ 182 bp region, composed of four SH loci, among the six pools examined (Fig. 3). To retrieve these three haplotypes with greatest efficiency, one should use populations represented by the *P. webbiana*, *procumbens*-Tenerife, and *patellaris*-Spain A pools, because these have the highest major variant frequencies at the four SH loci.

The principle illustrated in Fig. 3 can be extended. Because the frequency of all SH locus variants, minor or major, in the pool-seq pan-genome can be estimated, the probability of recovering sets of unlinked variants scattered across the genome can be approximated as their joint probability, equal to the product of the individual variant frequencies under the assumption of independent assortment. The minimum probability of recovering long DNA sequence haplotypes from a population can be similarly calculated (it is a minimum probability because physically adjacent variants will usually violate independent assortment). This allows one to select germplasm accessions with the highest probability of delivering haplotypes of interest, and to set the scale of the experiment, in terms of the number of individuals needed to recover desired variants at any position, or set of positions, in the genome.

Whole genome data sets are large and complex. They are often used collectively to describe population structure among accessions or classify germplasm by genetic distance, which, in turn, is used to select accessions from a gene bank (Muñoz-Amatriaín 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 (Reed and Frankham 2001; Reeves et al. 2012). Pool-seq pan-genome 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).

Future opportunities

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 Fig. 2 is one such reduction; there are many other SH locus variants that are not shown. However, a crop genome in its entirety can be mapped to orthologous sequence variants from its broader gene pool (Fig. 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 bar in Fig. 1, a Fig. 2 can be constructed (see https://github.com/NCGRP/mb1suppl for visualization). The resulting pre-processed data 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.

As crop improvement increasingly supplements conventional field breeding practices with in vitro techniques like transgenics and genome editing, the importance of accurate, comprehensive, sequence-based 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 off-target effects (Danilevicz et al. 2020). Pool-seq pan-genomes allow collections to be characterized progressively, one accession at a time; no reanalysis of existing data is required to add data for new accessions. This contrasts with population structure-based characterization, which requires reapplication of the variant calling pipeline and reanalysis with each new sample. 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 single-molecule long read data to existing short reads. This could extend the length of SH loci recovered.

Conclusion

We have proposed one option for a query-ready data structure that captures whole genome sequence variation for heterogeneous populations, 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 "future-proof" than derived analytical products, and will provide novel opportunities for crop improvement as new analytical methods are developed.

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. The fair, thoughtful, and constructive criticism of two anonymous reviewers for *Molecular Breeding* greatly improved 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, employer, and lender.

Author contribution Conceptualization, P.A.R. and C.M.R.; data curation, P.A.R.; formal analysis, P.A.R.; investigation, P.A.R.; methodology, P.A.R and C.M.R; project administration, C.M.R.; resources, C.M.R.; software, P.A.R.; supervision, C.M.R.; visualization, P.A.R.; writing—original draft preparation, P.A.R.; writing—review and editing, P.A.R and C.M.R.

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

Data availability Raw sequence data are available in the NCBI Sequence Read Archive. Derived data structures are available from the authors.

Code availability https:/github.com/NCGRP/mb1suppl.

Declarations

Conflict of interests The authors declare no competing interests.

References

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

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020) Plant pan-genomes 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, Panella 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 population-based whole genome sequences. BMC Genomics 21:189

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang 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 pan-genome 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 (2018) Haplotype variation of flowering time genes of sugar beet and its wild relatives and the impact on life cycle regimes. Front Plant Sci 8:2211. https://doi.org/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 pan-genome 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 H-J, Melzer S, Leineweber J, Defant B, Jung C (2021) A rhomboid-like 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. https://doi.org/10.1007/s10681-019-2366-9

Lemmon ZH, Reem NT, Dalrymple J, Soyk S, Swartwood KE, Rodríguez-Leal 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 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

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

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

Lynch M, Bost D, Wilson S, Maruki T, Harrison S (2014) Population-genetic inference from pooled-sequencing data. Genome Biol Evol 6:1210–1218

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

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

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; 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

Muñoz-Amatriaín M, Cuesta-Marcos 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 genome-wide association studies. PLoS ONE 9:e94688. https://doi.org/10.1371/journal.pone.0094688

Pin PA, Zhang W, Vogt SH, Dally N, Büttner B, Schulze-Buxloh G, Jelly NS, Chia TYP, Mutasa-Göttgens 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 pseudo-response regulator gene in life cycle adaptation and domestication of beet. Curr Biol 22:1095–1101

Reed DH, Frankham R (2001) How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. Evolution 55:1095–1103

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

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

Rodríguez-Leal 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, Santos-Guerra A, Batista D, Duarte MC, Paulo OS (2016) Evolutionary and biogeographic insights on the Macaronesian *Beta-Patellifolia* species (Amaranthaceae) from a time-scaled molecular phylogeny. PLoS ONE 11:e0152456. https://doi.org/10.1371/journal.pone.0152456

Scheben A, Edwards D (2018) Towards a more predictable plant breeding pipeline with CRISPR/Cas-induced 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 genome-wide polymorphism data without big funding. Nat Genet 15:749–763

Sundaram AYM, Garseth ÅH, 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

Thulin M, Rydberg A, Theide J (2010) Identity of *Tetragonia pentandra* and taxonomy and distribution of *Patellifolia* (Chenopodiaceae). Willdenowia 40:5–11

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, Flint-Garcia 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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.