G3, 2026, **16(1)**, jkaf261
https://doi.org/10.1093/g3journal/jkaf261
Advance Access Publication Date: 8 November 2025
Investigation

Whole-genome resequencing of the wild barley diversity collection: a resource for identifying and exploiting genetic variation for cultivated barley improvement

Rebecca Spanner [1], Ahmad H. Sallam,[1,2] Yu Guo,[3] Murukarthick Jayakodi [3,4], Axel Himmelbach,[3] Anne Fiebig,[5] Jamie Simmons,[1] Gerit Bethke [6], Yoonjung Lee,[1] Luis Willian Pacheco Arge,[6] Yinjie Qiu [7], Ana Badea,[8] Michael Baum,[9] François Belzile [10,11,12], Roi Ben-David,[13] Robert Brueggeman [14], Austin Case,[15] Luigi Cattivelli,[16] Michael Davis,[17] Christoph Dockter [18], Jaroslav Doležel [19], Antonin Dreiseitl [20], Ryan Gavin,[1] Lior Glick [21], Stephan Greiner,[22] Ruth Hamilton,[23] Patrick M. Hayes,[24] Scott Heisel,[17] Cynthia Henson,[25] Benjamin Kilian,[26] Takao Komatsuda,[27] Chengdao Li [28,29], Cheng Liu [27], Ramamurthy Mahalingam,[25] Maren Maruschewski,[5] Oadi Matny [1], Andreas Maurer [30], Klaus F.X. Mayer,[31,32] Itay Mayrose [21], Matthew Moscou [33], Gary J. Muehlbauer,[6] Youko Oono [34], Frank Ordon,[35] Hakan Özkan,[36] Ales Pecinka [19], Dragan Perovic [35], Klaus Pillen [30], Mohammad Pourkheirandish [37], Joanne Russell,[23] Jan Šafář,[38] Silvio Salvi [39], Miguel Sanchez-Garcia,[9] Kazuhiro Sato [40,41,42], Thomas Schmutzer [30], Uwe Scholz [5], Jeness Scott,[1] Gurcharn Singh Brar,[43] Kevin P. Smith [6], Mark E. Sorrells,[44] Manuel Spannagl [31], Nils Stein [45,46], Alessandro Tondelli,[16] Roberto Tuberosa [39], James Tucker,[8] Thomas Turkington,[47] Jan Valkoun,[48] Ramesh Pal Singh Verma,[9] Marcus A. Vinje,[25] Maria von Korff [49,50], Jason G. Walling [25], Robbie Waugh [23,51,52], Roger P. Wise [53,54], Brande B.H. Wulff [55], Shengming Yang [56,57], Guoping Zhang [58], Peter L. Morrell [6], Martin Mascher [3,59,*], Brian J. Steffenson [1,*]

[1]Department of Plant Pathology, University of Minnesota Twin Cities, Saint Paul, MN 55108, United States
[2]Plant Physiology and Genetics Unit, U.S. Arid Land Agricultural Research Center, Maricopa, AZ 85138, United States
[3]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Saxony-Anhalt 06466, Germany
[4]Department of Soil and Crop Sciences, Texas A&M AgriLife Research, Dallas, TX 75252, United States
[5]Bioinformatics and Information Technology, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Saxony-Anhalt 06466, Germany
[6]Department of Agronomy and Plant Genetics, University of Minnesota Twin Cities, Saint Paul, MN 55108, United States
[7]Minnesota Supercomputing Institute, University of Minnesota Twin Cities, Minneapolis, MN 55455, United States
[8]Brandon Research and Development Centre, Agriculture and Agri-Food Canada, Brandon, Manitoba R7A 5Y3, Canada
[9]Biodiversity and Crop Improvement, International Center for Agricultural Research in the Dry Areas (ICARDA), Rabat BP6299, Morocco
[10]Département de Phytologie, Université Laval, Québec G1V 0A6, Canada
[11]Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec G1V 0A6, Canada
[12]Centre de Recherche Et d'innovation Sur Les Végétaux (CRIV), Université Laval, Québec G1V 0A6, Canada
[13]Department of Vegetables and Field Crops, ARO-Volcani Center, Rishon LeZion 7505101, Israel
[14]Department of Crop and Soil Sciences, Washington State University, Pullman, WA 99164-6420, United States
[15]Anheuser-Busch InBev, Fort Collins, CO 80524, United States
[16]Research Centre for Genomics and Bioinformatics, Council for Agricultural Research and Economics (CREA), Fiorenzuola d'Arda, Piacenza 29017, Italy
[17]American Malting Barley Association, Inc., Brookfield, WI 53005, United States
[18]Raw Materials, Carlsberg Research Laboratory, Copenhagen DK-1799, Denmark
[19]Centre of Plant Structural and Functional Genomics, Institute of Experimental Botany, Czech Academy of Sciences, Olomouc CZ-77900, Czech Republic
[20]Department of Integrated Plant Protection, Agrotest Fyto Ltd., Kroměříž CZ-767 01, Czech Republic
[21]School of Plant Sciences and Food Security, Tel Aviv University, Tel Aviv 69978, Israel
[22]Roche Information Solutions (RIS), Data, Analytics & Research (DA&R), Roche Diagnostics GmbH, Penzberg, Bavaria 82377, Germany
[23]Cell and Molecular Sciences, The James Hutton Institute, Dundee, Scotland DD2 5DA, UK
[24]Department of Crop and Soil Science, Oregon State University, Corvallis, OR 97331, United States
[25]Cereal Crops Research Unit, USDA-ARS, Madison, WI 53726, United States
[26]Global Crop Diversity Trust, Bonn, North Rhine-Westphalia 53113, Germany
[27]Shandong Academy of Agricultural Sciences (SAAS), Crop Research Institute, Jinan, Shandong 250100, China
[28]Western Crop Genetics Alliance, Food Futures Institute/School of Agriculture, Murdoch University, Murdock, WA 6150, Australia
[29]Department of Primary Industries and Regional Development, Government of Western Australia, Perth, WA 6151, Australia
[30]Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle, Saxony-Anhalt 06120, Germany
[31]Plant Genome and Systems Biology (PGSB), Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Neuherberg, Bavaria 85764, Germany
[32]School of Life Sciences, Technical University of Munich, Freising, Bavaria 85354, Germany
[33]Cereal Disease Laboratory, USDA-ARS, Saint Paul, MN 55108, United States
[34]Institute of Crop Science, National Agriculture and Food Research Organization (NARO), Tsukuba, Ibaraki 305-8602, Japan
[35]Institute for Resistance Research and Stress Tolerance, Julius Kühn Institute (JKI), Quedlinburg, Saxony-Anhalt 06484, Germany
[36]Department of Field Crops, Faculty of Agriculture, University of Çukurova, Sarıçam, Adana 1250, Turkey
[37]School of Agriculture, Food, and Ecosystem Sciences (SAFES), University of Melbourne, Melbourne, Victoria 3010, Australia
[38]Institute of Experimental Botany, Centre of Plant Structural and Functional Genomics, Olomouc 779 00, Czech Republic
[39]Department of Agricultural and Food Sciences, University of Bologna, Bologna 40127, Italy
[40]Institute of Plant Science and Resources, Okayama University, Kurashiki, Okayama 710-0046, Japan

Received on 21 March 2025; accepted on 01 October 2025
Published by Oxford University Press on behalf of The Genetics Society of America 2025.

[41]Department of Frontier Research and Development, Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan
[42]Faculty of Agriculture, Setsunan University, Hirakata, Osaka 573-0101, Japan
[43]Department of Agricultural, Food and Nutritional Sciences, University of Alberta, Edmonton, AB T6G 2P5, Canada
[44]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853-1902, United States
[45]Genebank, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Saxony-Anhalt 06466, Germany
[46]Crop Plant Genetics, Institute of Agricultural and Nutritional Sciences, Martin Luther University of Halle-Wittenberg, Halle (Saale), Saxony-Anhalt, 06120, Germany
[47]Agriculture and Agri-Food Canada, Lacombe Research and Development Centre, Lacombe, Alberta T4L 1W1, Canada
[48]Genetic Resources Unit, International Center for Agricultural Research in the Dry Areas (ICARDA), Praha 15000, Czech Republic
[49]Institute of Plant Genetics, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, North Rhine-Westphalia 40225, Germany
[50]Cluster of Excellence on Plant Sciences, Düsseldorf, North Rhine-Westphalia 40225, Germany
[51]Division of Plant Sciences, The University of Dundee, Dundee, Scotland DD2 5DA, UK
[52]The University of Adelaide, Adelaide, SA 5064, Australia
[53]Corn Insects and Crop Genetics Research Unit, USDA-ARS, Ames, IA 50011, United States
[54]Department of Plant Pathology, Entomology, and Microbiology, Iowa State University, Ames, IA 50011, United States
[55]Plant Science Program, Biological and Environmental Science and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[56]Cereal Crops Improvement Research Unit, USDA-ARS, Fargo, ND 58102, United States
[57]Department of Plant Sciences, North Dakota State University, Fargo, ND 58102, United States
[58]Department of Agronomy, Zhejiang University, Hangzhou, Zhejiang 310029, China
[59]Halle-Jena-Leipzig, German Centre for Integrative Biodiversity Research (iDiv), Leipzig, Saxony 04103, Germany

*Corresponding authors: Brian J. Steffenson, Department of Plant Pathology, University of Minnesota, Saint Paul, MN 55108, United States. Email: bsteffen@umn.edu; Martin Mascher, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Saxony-Anhalt 06466, Germany. Email: mascher@ipk-gatersleben.de

To exploit allelic variation in *Hordeum vulgare* subsp. *spontaneum*, the Wild Barley Diversity Collection was subjected to paired-end Illumina sequencing at ~9× depth and evaluated for several agronomic traits. We discovered 240.2 million single nucleotide polymorphisms (SNPs) after alignment to the Morex V3 assembly and 24.4 million short (1 to 50 bp) insertions and deletions. A genome-wide association study of lemma color identified one marker-trait association (MTA) on chromosome 1H close to *HvBlp*, the cloned gene controlling black lemma. Four MTAs were identified for seedling stem rust resistance, including 2 novel loci on chromosomes 1H and 6H and one co-locating to the complex RMRL1-RMRL2 locus on 5H. The whole-genome sequence data described herein will facilitate the identification and utilization of new alleles for barley improvement.

Keywords: *Hordeum vulgare* subsp. *spontaneum*; whole genome sequence data; genome-wide association study; agronomic traits

Introduction

Barley (*Hordeum vulgare* L. subsp. *vulgare*) was one of the first crops domesticated in the Near East ~10,000 years ago (Zohary et al. 2012) and is currently cultivated over 47 million hectares worldwide (FAO 2017). Its main uses include animal feed, malt for various alcoholic beverages, and human food. Through the domestication process and modern plant breeding, the genetic diversity of barley has been eroded (Russell et al. 2016; Milner et al. 2019; Cíváň et al. 2024), leaving the crop vulnerable to various biotic and abiotic threats and limiting further improvements for key traits. The primary gene pool of barley includes varieties, breeding lines, landraces, and wild barley (*H. vulgare* L. subsp. *spontaneum* C. Koch. Thell.), the latter of which can readily hybridize with the cultivated forms (Harlan and Zohary 1966; Liu et al. 2024). Studies aimed at identifying unexploited genes for use in barley breeding programs typically include panels more closely related to elite germplasm, thereby preserving the genetic linkages of favorable alleles for yield, quality, and agronomic traits. When a particular trait cannot be found in the cultivated forms of the primary gene pool, researchers often seek the desired alleles in the wild progenitor. To capture the allelic variation in wild barley, an ecogeographically diverse collection, known as the Wild Barley Diversity Collection (WBDC), was assembled (Steffenson et al. 2007). The WBDC comprises 318 accessions from across the range of *H. vulgare* subsp. *spontaneum* and has been evaluated for various agronomic, morphological, nutritional, and disease/pest resistance traits. These evaluations revealed a high level of variation for all the characterized traits, leading to subsequent genetic and genome-wide association studies (GWAS) based on various molecular marker technologies (Roy et al. 2010; Sallam et al. 2017; Mahalingam et al. 2020; Walling et al. 2022). Here, we describe the whole-genome resequencing of 281 WBDC accessions with ~9× coverage and demonstrate its utility for identifying both previously described and novel genes in *Hordeum vulgare* using an association genetic approach.

Methods

Wild barley germplasm

Collection site data for longitude and latitude, elevation, high and low temperature, rainfall, and soil type (Supplementary Table 1; Fig. 1a) were used to assemble the WBDC at the International Center for Agricultural Research in the Dry Areas (ICARDA) (Steffenson et al. 2007). The proportion of samples included was generally reflective of the density of populations in the Fertile Crescent, Central Asia, North Africa, and Caucasus regions. Of the 318 WBDC accessions selected initially, 37 were not included in resequencing due to failed genotyping or sequencing, duplication, or seed admixtures. The final sequenced panel comprises 281 accessions from 19 countries. Single plant selections were initially made from each accession and then selfed for 5 successive generations in the greenhouse before being used for DNA extraction and sequencing.

DNA extractions

The first and second leaves of each accession were harvested, flash-frozen in liquid nitrogen, and stored at −80°C until the DNA extractions were performed. For the extractions, tissue was first ground to a fine powder in liquid nitrogen using a mortar and pestle. Then, genomic DNA was extracted using a modified CTAB protocol (Yu et al. 2017). Agarose gel electrophoresis was

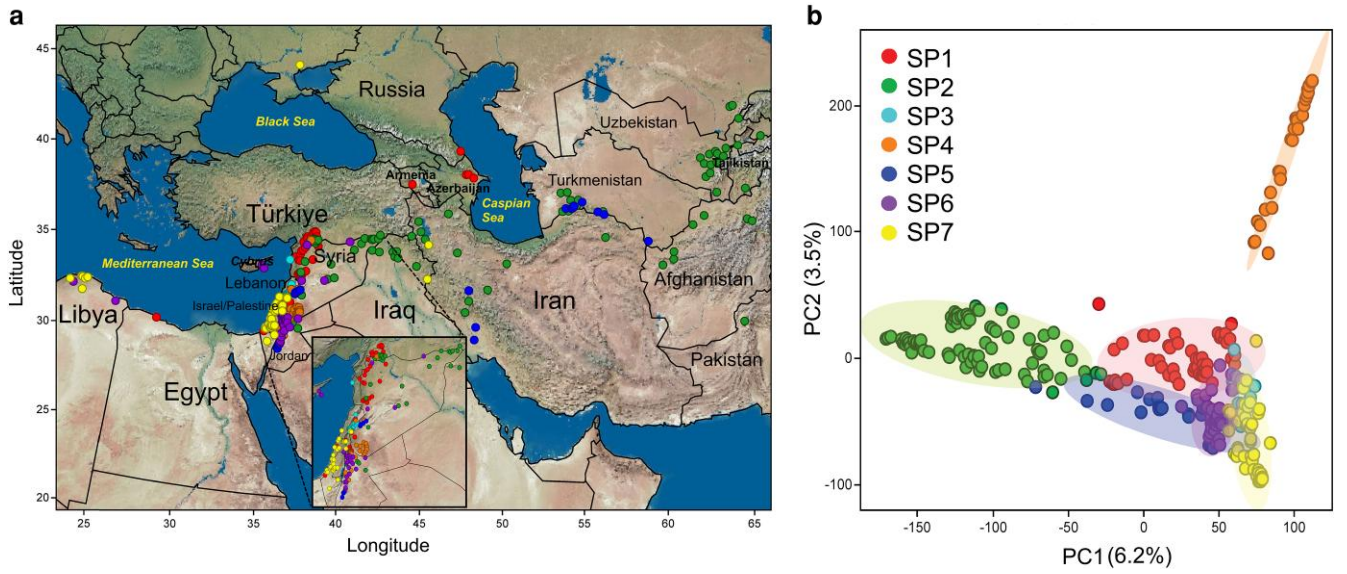

Fig. 1. a) Geographic distribution of 281 *Hordeum vulgare* subsp. *spontaneum* accessions of the Wild Barley Diversity Collection (WBDC) and b) Principal component analysis determined from ~1.3 million single nucleotide polymorphisms (SNPs). Accessions are color-coded by sub-population, defined by *k*-means clustering.

used to confirm that the genomic DNA was of high molecular weight (>10 kb). DNA quality was assessed using a NanoDrop spectrophotometer.

Library preparation and whole-genome sequencing (WGS)

WGS libraries were prepared using the "Illumina Nextera DNA Flex Library Preparation Kit" (workflow for 100 to 500 ng DNA input, 5 PCR-cycles for the addition of indexes) according to manufacturer's instructions (Illumina, Inc., San Diego, CA, USA). The final library pool was quantified by qPCR (Mascher et al. 2021). The pool was sequenced (XP workflow, paired-end, 2 × 151 cycles) using the Illumina NovaSeq6000 device and standard protocols from the manufacturer.

Variant calling

Quality assessment, read mapping, deduplication, and coverage estimation utilized scripts in the RepAdapt pipeline (https://github.com/RepAdapt/snp_calling_simple). This involved quality assessment and adapter trimming with FASTP (Chen et al. 2018), read mapping with BWA MEM (Li, 2013), read deduplication with Picard (Broad Institute 2019) and coverage estimation with samtools (Li et al. 2009; Danecek et al. 2021). Both SNP and indel variants were called using GATK version v4.1.2 (McKenna et al. 2010), with recommended GATK filtering as follows: SNP filtering "QD < 2.0 || FS > 60.0 || MQ < 45.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0 || DP > 4654.61"; and Indel Filtering "QD < 2.0 || FS > 200.0 || ReadPosRankSum < −20.0". Heterozygous and multiallelic sites were retained in the data set, and no frequency filter was applied to variants. The general feature format (GFF3) descriptions of both high and low confidence genes from the Morex V3 assembly (Mascher 2020) were used to create a BED file defining "gene space." Based on the descriptions in the GFF, gene space here comprises 5′ and 3′ UTRs, exons, and introns.

Variant statistics

Variant statistics, including sample size estimation and read depth per SNP, level of missingness, and related statistics, were calculated using the bcftools "+fill-tags" plugin (Danecek et al. 2021). Sample-level statistics were calculated with bcftools stats. The Variant Effect Predictor (VeP) (McLaren et al. 2016), along with GFF3 annotations for Morex V3, was used to annotate all variants. VeP results were used to parse variants by class for calculating the site frequency spectrum.

The folded site frequency spectrum (SFS) was estimated based on the minor allele frequency for biallelic SNPs genome-wide and for synonymous sites. The expectation for the SFS under a neutral coalescent history was generated using msprime (Baumdicker et al. 2022) based on nucleotide sequence diversity estimated as $\theta = 4N_e\mu = 0.008$ (Morrell et al. 2006; Schmid et al. 2018) and recombination rate of $\rho/\theta = 1.5$ (Morrell et al. 2006). We simulated 281 haploid samples with 1,000 replicate simulations with a locus length sufficient to generate ~100 SNPs per simulation or 100,000 variants that could be compared to the folded SFS for the empirical datasets.

Cluster analysis

The SNP dataset was filtered prior to analyses by setting heterozygote calls to missing and retaining biallelic sites with ≤10% missing data and ≥5% minor allele frequency. The SNP dataset was further pruned by discarding sites with $r^2 > 0.2$ in windows of 50 sites. Principal component analysis was performed in TASSEL v5.0 (Bradbury et al. 2007). *K*-means clustering was used to partition the wild barley panel into subpopulations (Supplementary Table 1). Based on our previous knowledge of the panel (Sallam et al. 2017), 7 subpopulations were assigned to the cluster analysis in JMP 17 (JMP Statistical Discovery LLC, Cary, NC, USA). JMP was used to plot the PCA results and create the map for the WBDC panel (Fig. 1).

Phenotyping

To demonstrate the utility of the WGS dataset for identifying trait-associated loci in wild barley, we selected 2 important traits for study: (1) lemma color and (2) stem rust resistance. Lemma color was assessed by taking digital images of mature seeds and then analyzing each color channel using the Fiji package

Table 1. Summary statistics for variants obtained after aligning whole-genome sequencing reads from 281 accessions of the Wild Barley Diversity Collection to the Morex V3 reference genome.

Data set	Variant #	Multiallelic	Ts/Tv	Proportion missing
SNPs	240,171,785	9,214,159	1.39	0.094 (+0.177) −0.018
Indels	24,387,195	6,493,811	1.07	0.148 (+0.217) −0.039
Biallelic	59,520,067	…	1.47	0.083 (+0.166) −0.014
SNPs—Gene space	6,385,855	166,712	1.90	0.068 (+0.158) −0.007

Biallelic sites include positions with unique mutations and no overlapping indels. Gene space includes coding regions and UTRs from Morex V3 gene annotation.

Table 2. Variant effect predictor (VeP) results for genome-wide SNPs and indels.

Variant	SNPs		Indels	
	Count	Proportion (%)	Count	Proportion (%)
Splice acceptor	2,007	0.0008	986	0.00404
Splice donor	1,747	0.0007	1,201	0.00492
Stop gained	18,571	0.0077	2,329	0.00955
Frameshift variant	…	…	74,148	0.30404
Stop lost	1,520	0.0006	190	0.00078
Start lost	1,507	0.0006	267	0.00109
Inframe insertion	…	…	15,642	0.06414
Inframe deletion	…	…	24,768	0.10156
Missense	807,753	0.3363	334	0.00137
Protein-altering variant	…	…	1,163	0.00477
Splice donor 5th base	4,472	0.0019	1,248	0.00512
Splice region	47,271	0.0197	7,983	0.03273
Splice donor region	12,910	0.0054	2,225	0.00912
Splice polypyrimidine tract	46,258	0.0193	11,488	0.04711
Start retained	…	…	9	0.00004
Stop retained	989	0.0004	98	0.0004
Synonymous	685,452	0.2854	71	0.00029
Coding sequence variant	…	…	781	0.0032
5′UTR	96,094	0.04	36,888	0.15126
3′UTR	229,979	0.0958	56,752	0.23271
Intron	2,757,606	1.1482	554,672	2.27444
Upstream gene	9,135,842	3.8039	1,560,586	6.3992
Downstream gene	7,049,038	2.935	1,194,674	4.89878
Intergenic	219,000,000	91.2983	20,838,692	85.44932

(Schindelin et al. 2012). The average color channel value (CCV) of 2 representative seeds of each accession for each color channel in the RGB color model was measured. Each RGB value was converted into a single 24-bit integer for GWAS analysis using the formula: $\text{Color} = (R \times 256^2) + (G \times 256) + B$. Stem rust assays on seedlings were performed with 2 races (MCCFC and QCCJB) of the wheat stem rust pathogen (*Puccinia graminis* f. sp. *tritici* (*Pgt*) and one isolate (92-MN-90) of the rye stem rust pathogen (*P. graminis* f. sp. *secalis*) (*Pgs*) as described in Sallam et al. (2017).

Genome-wide association mapping

To identify markers associated with the two traits, GWAS was conducted for 281 WBDC accessions using the following methods: (1) Mixed Linear Model (MLM) that accounts for population structure (Q) + kinship (K) (Yu et al. 2006), (2) Fixed and random model Circulating Probability Unification (FarmCPU) (Kusmec and Schnable 2018) that utilizes fixed and random effects iteratively to improve association power, and (3) a Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) (Huang et al. 2019) that utilizes Bayes and linkage disequilibrium to improve both association power and computation efficiency. The SNP dataset was filtered prior to GWAS by setting heterozygote calls to missing and retaining biallelic sites with ≤10% missing data and ≥5% minor allele frequency. The SNP dataset was further pruned by discarding sites with $r^2 > 0.2$ in windows of 50 sites. All association mapping methods were executed in the R package GAPIT v3.5 using ~1.3 million SNP markers (Wang and Zhang 2021). Marker trait associations (MTAs) identified using two or more methods or those detected with a single method but across 2 different datasets are presented. The Bonferroni test was performed to declare significant associations.

Results and discussion

Variant calling with GATK in the 281 sequenced WBDC accessions (~9× coverage, Supplementary Fig. 1) resulted in the identification of 240.2 million SNPs and 24.4 million indels (Table 1). In addition to these reported indels, there were a further 3.2 million sites where one of the variants at an indel site had a single base pair difference from the reference. These are among 6.5 million multiallelic variants, with 26.6% of indel sites called as multiallelic.

The vast majority of detected indels were one bp deletions; the second most abundant class was one bp insertions, with roughly half as many one bp insertions identified relative to deletions (Supplementary Fig. 2). The majority of variants identified were annotated as intergenic variants, including 219.2 million SNPs and 20.8 million indels (Table 2). Among coding SNPs, 53.3% (807,753) were missense changes, 45.2% (685,452) were synonymous changes, and 1.2% (18,571) were stop-gained. Among indels affecting coding regions, 61.9% (74,148) were frameshift variants, 20.7% (24,768) were inframe deletions, 13.1% (15,642) were inframe insertions, and 1.9% (2,329) were stop-gained.

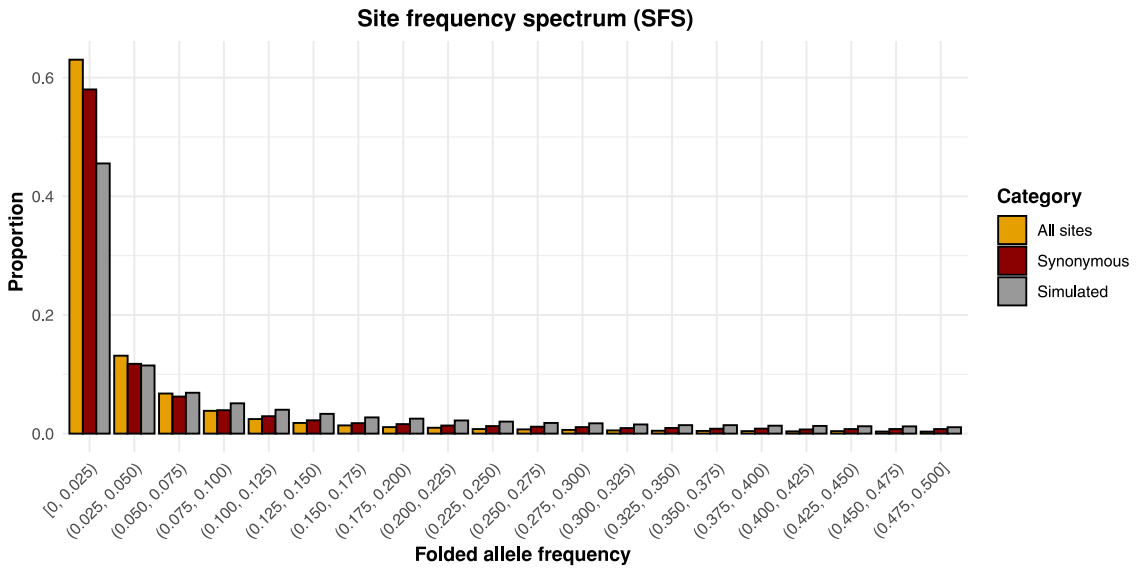

Fig. 2. Folded site frequency spectrum for genome-wide biallelic SNPs. The plot includes all biallelic sites, synonymous sites, and variants simulated under a neutral coalescent history.

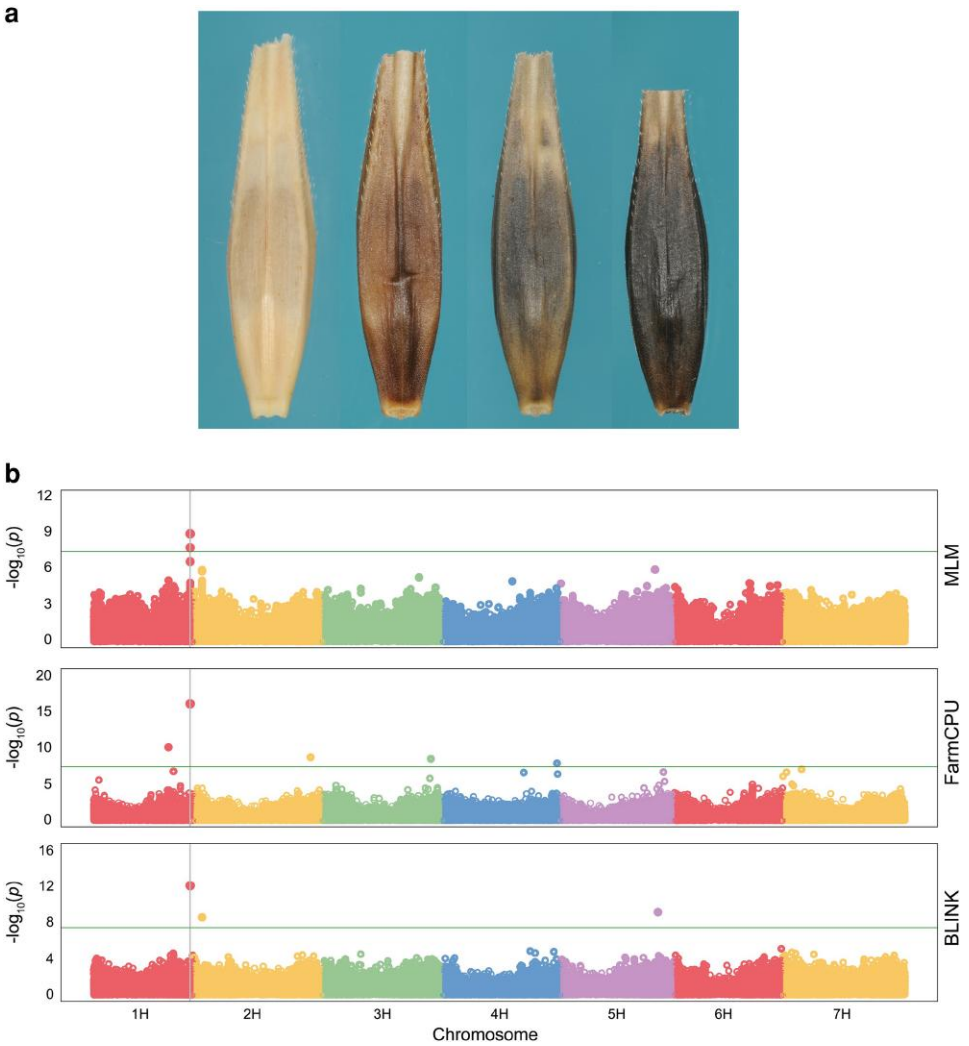

Fig. 3. a) Examples of different lemma colors in the wild barley diversity collection; from left to right: yellow (straw) from WBDC045, brown from WBDC204, diffuse black from WBDC014, and dark black from WBDC355. b) Manhattan plots displaying single-polymorphism (SNP) markers significantly associated with lemma color in the Wild Barley Diversity Collection. Three models were used in the analysis: (1) a Mixed Linear Model (MLM), 2) a Fixed and random model Circulating Probability Unification (FarmCPU), and 3) a Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK). Bonferroni significance threshold is shown with a horizontal solid green line.

We also partitioned the data set into variants found within gene space as defined by Morex V3 annotations. SNPs within genic regions showed much lower rates of multiallelic polymorphisms at 2.6% and a lower missingness rate at 0.068 (±0.158) with a median of 0.007, consistent with the relative ease of read alignment and variant calling within gene space (Table 1).

The transition to transversion ratio (Ts/Tv) can vary among classes of variants and is a potential metric of variant call quality. Values in partitions for the dataset ranged from 1.07 for indels to 1.90 for variants in gene space. These values align with prior reported values of 1.7 for Sanger sequencing in wild barley (Morrell et al. 2006) and Illumina exome capture sequencing from domesticated barley (Kono et al. 2016).

There were 59.5 million biallelic SNPs in the dataset that occurred outside of indels and thus were unique mutations. The folded site frequency spectrum in Fig. 2 includes all biallelic SNPs. We compared frequencies with expectations under a standard coalescent model of a panmictic population with constant population size. The SFS shows that a large proportion of variants reside in the rarest frequency class, here <2.5% frequency. While this was consistent with expectations under a standard coalescent model (Tajima 1989), rare variants at the whole-genome level were more abundant than expected based on neutral coalescent simulations (Fig. 2). The frequency spectrum for synonymous sites more closely resembles the expectation for neutral variants based on coalescent simulations, but again demonstrates more variants in the rarest frequency classes. This result likely reflects both the challenges of variant calling in a highly repetitive genome and an excess of rare variants at most wild barley loci, consistent with a recent population expansion in the species' coalescent history (Morrell et al. 2006).

Prior to GWAS, population structure was assessed by *k*-means clustering and principal component analyses (Fig. 1b). Consistent with previous results on the population structure of wild barley (Fang et al. 2014; Russell et al. 2016; Sallam et al. 2017), genetic relatedness mirrored geographic distance: the distribution of population centers roughly traced a path from the North African coast and the Southern Levant along the Fertile Crescent to Central Asia (Fig. 1a). A detailed analysis of population structure in wild barley and its relationship to domesticated accessions was undertaken by Guo et al. (2025) using the present data set.

Lemma color

Lemma color in the WBDC ranged from pale yellow (straw-colored) to brown and dark black based on visual inspection (Fig. 3a). Converted RGB values from digital images of pale yellow and dark black seed generally ranged from 10,750,000 to 13,550,000 and 4,350,000 to 6,570,000, respectively (Supplementary Table 1; Supplementary Fig. 3). Black lemma is a classic morphological trait in barley and is controlled by the *Blp* locus, which is composed of different alleles contributing to the intensity and distribution of color (Franckowiak and Lundqvist 1997). GWAS identified one association (WBDC_LC_1H_499.0) by a single SNP (S1H_499023721) on chromosome 1H using all 3 models (MLM, FarmCPU and BLINK) (Fig. 3b, Table 3). This SNP explained 17.5% of the phenotypic variation and lies in close proximity to *HvBlp*, the recently cloned gene controlling black lemma color positioned between 498.5 and 499.0 Mbp on 1H in the Morex V3 assembly (Li et al. 2024) (Table 3). Due to the complexity of the locus and a duplicated fragment of *HvBlp*, it is difficult to state with certainty the physical relationship of the identified SNP marker and this gene.

Table 3. Single-nucleotide polymorphism (SNP) markers significantly associated with lemma color and stem rust resistance in 281 *Hordeum vulgare* subsp. *spontaneum* accessions of the Wild Barley Diversity Collection.

Phenotype	Association designation[a]	Treatment or Trait[b]	SNP[c]	Chr[d]	Pos[e]	MAF[f]	GWAS detection method[g]	P-value range[h]		R^{2i}
Lemma color	WBDC_LC_1H_499.0	Lemma color	S1H_499023721	1H	499.0 Mbp	0.06	MLM/FarmCPU/BLINK	9.78×10^{-17}	1.36×10^{-9}	17.51%
Stem rust resistance	WBDC_SR_1H_11.7	MCCFC	S1H_11651434	1H	11.7 Mbp	0.07	MLM/FarmCPU	9.83×10^{-11}	1.81×10^{-10}	20.02%
		QCCJB	S1H_11651434	1H	11.7 Mbp	0.07	MLM/FarmCPU	2.17×10^{-14}	5.76×10^{-9}	14.86%
	WBDC_SR_1H_67.4-71.5	MCCFC	S1H_67388912, S1H_71536803	1H	67.4 to 71.5 Mbp	0.07	MLM MLM/FarmCPU/BLINK	1.50×10^{-24}	1.87×10^{-8}	22.46 to 31.03%
		QCCJB	S1H_67388912	1H	67.4 Mbp	0.07	BLINK	5.77×10^{-14}		18.71%
	WBDC_SR_5H_562.9	MCCFC	S5H_562922829	5H	562.9 Mbp	0.07	MLM/FarmCPU/BLINK	3.25×10^{-31}	3.49×10^{-12}	22.38%
		QCCJB	S5H_562922829	5H	562.9 Mbp	0.07	MLM/FarmCPU/BLINK	7.98×10^{-34}	3.77×10^{-13}	24.08%
		92-MN-90	S5H_562922829	5H	562.9 Mbp	0.07	MLM/FarmCPU/BLINK	2.89×10^{-42}	5.09×10^{-17}	32.28%
	WBDC_SR_6H_501.8	MCCFC	S6H_501789703	6H	501.8 Mbp	0.08	MLM	1.84×10^{-9}		15.80%
		QCCJB	S6H_501789703	6H	501.8 Mbp	0.08	MLM/FarmCPU	9.49×10^{-10}	2.60×10^{-8}	15.43%

[a] Association designation is based on the germplasm (WBDC), trait abbreviation (eg lemma color), chromosome location (1H), and physical position from the Morex (V3) genome assembly.
[b] Treatment or trait includes lemma color, reaction to races MCCFC and QCCJB of the wheat stem rust pathogen (*Puccinia graminis* f. sp. *tritici*), and reaction to isolate 92-MN-90 of the rye stem rust pathogen (*P. graminis* f. sp. *secalis*).
[c] SNP designation is based on the chromosome and physical position from the Morex (V3) genome assembly.
[d] Barley chromosome and arm designation: S = short or L = Long.
[e] Physical position based on the Morex (V3) genome assembly.
[f] Minimum allele frequency.
[g] Only associations detected with 2 or more methods or with a single method but in more than one dataset are shown. MLM denotes Mixed Linear Model (MLM); FarmCPU denotes fixed and random model Circulating Probability Unification; and BLINK denotes Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway.
[h] Range of *P*-values for the associations identified through different models.
[i] R^2 values for the SNP marker found significantly associated with the trait.

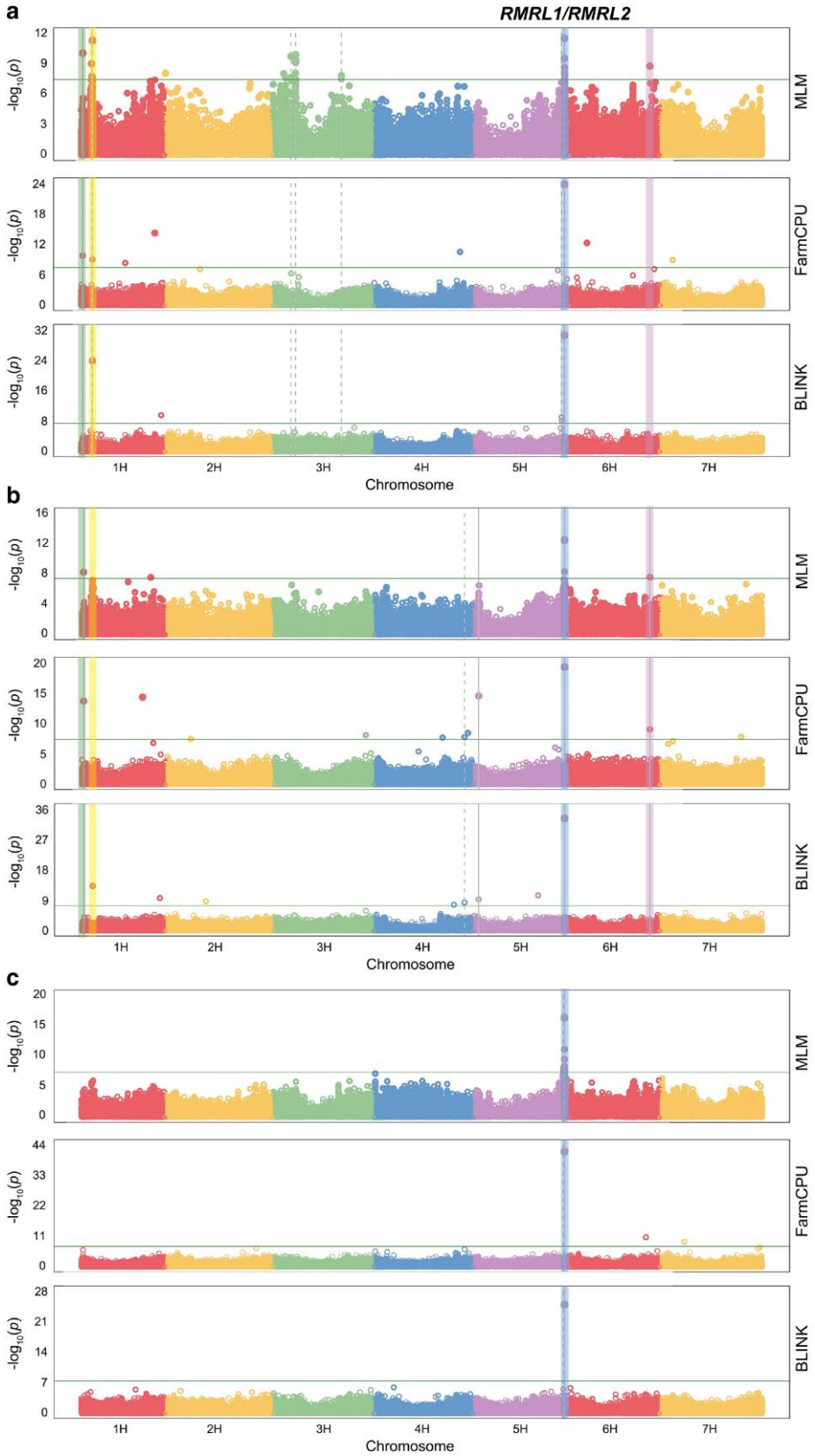

Fig. 4. Manhattan plots displaying single nucleotide polymorphism (SNP) markers significantly associated with resistance to the wheat stem rust (*Puccinia graminis* f. sp. *tritici*, *Pgt*) and rye stem rust (*P. graminis* f. sp. *secalis*, *Pgs*) pathogens: a) race *Pgt*-MCCFC, b) race *Pgt*-QCCJB, and (c) isolate *Pgs*-92-MN-90 in the Wild Barley Diversity Collection. Three models were used in the analysis: (i) a mixed linear model (MLM), (ii) a fixed and random model circulating probability unification (FarmCPU), and (iii) a Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK). The Bonferroni significance threshold is shown with a horizontal solid green line. The vertical blue, purple, yellow, and green lines show the significant associations consistently identified for resistance to 2 cultures of *P. graminis* with at least 1 or 2 models or to 1 culture with all 3 models. RMRL1/RMRL2 is a complex of several stem rust resistance genes.

Stem rust

Based on a coefficient of infection threshold of 2.7, only 15 (5.0%), 39 (14.0%), and 54 (19.0%) of the sequenced WBDC accessions were classified as resistant to *Pgt*-MCCFC, *Pgt*-QCCJB and *Pgs*-92-MN-90, respectively (Supplementary Table 1; Supplementary Fig. 4). Four MTAs (WBDC_SR_1H_11.7, WBDC_SR_1H_67.4–71.5, WBDC_SR_5H_562.9, and WBDC_SR_6H_501.8) were identified for stem rust resistance. WBDC_SR_1H_11.7 was novel and mapped to chromosome 1H in response to both *Pgt*-MCCFC and *Pgt*-QCCJB, explaining 20.0% and 14.9% of the variation, respectively (Fig. 4, Table 3). WBDC_SR_6H_501.8 was also novel and positioned on 6H in response to both *Pgt*-MCCFC and *Pgt*-QCCJB, explaining 15.4% to 15.8% of the variation (Table 3). WBDC_SR_1H_67.471.5 was mapped on chromosome 1H in response to races *Pgt*-MCCFC and *Pgt*-QCCJB (Table 3). The 2 different SNPs (S1H_67388912 and S1H_71536803) identified in the MTA were in moderate linkage disequilibrium ($r^2 = 0.565$). The position of this MTA is close to S1H_71499376, a genotyping-by-sequencing (GBS)-derived marker that was found significantly associated with resistance to both *Pgt*-MCCFC and *Pgt*-QCCJB based on 314 individuals (Sallam et al. 2017). This MTA explained 18.7% to 31.0% of the variation in this study (Table 3). WBDC_SR_5H_562.9 was identified on chromosome 5H (S5H_562922829) after challenge with all 3 *P. graminis* cultures using all 3 models and explained 22.4% to 32.3% of the variation (Table 3). It co-located to the position of the complex RMRL1-RMRL2 loci (Wang et al. 2013) from which several component resistance genes (eg *rpg4* and *Rpg5*) were cloned (Brueggeman et al. 2008; Arora et al. 2013).

Whole-genome sequencing data for diverse accessions of a crop and its wild relatives are essential for population genomic studies, the informed selection of genotypes for full genome sequence assembly (pangenomics), and the isolation of agronomically important genes. Our dataset complements similar short-read datasets for 1,315 domesticated barleys (Jayakodi et al. 2020, 2024) and 100 wild barleys from another collection (Jayakodi et al. 2020). Chromosome-scale genome assemblies of 9 WBDC accessions have been completed (Jayakodi et al. 2024), with more accessions to follow in the future. Applying GWAS to the WBDC, we demonstrated the utility of high-coverage sequence data for identifying novel genetic variation that may be useful in barley improvement. Additionally, we also validated major genes controlling key traits in barley, such as *Blp* for black lemma color and RMRL1/RMRL2 for stem rust resistance. Thus, this dataset may serve as a starting point for the identification of candidate genes underlying other important traits. In a companion paper, Guo et al. (2025) demonstrated the utility of WBDC sequence data in a population genomic study. They analyzed this dataset together with sequence data from other diverse wild and domesticated barley accessions to reconstruct the evolutionary history of wild barley and elucidate the origin of haplotypes in cultivated barley. The sequenced WBDC genomes will help connect target phenotypic traits to chromosome positions. Reference genome positions, as identified by HORVU I.D.s in the Morex V3 assembly (Mascher 2020), serve as anchors to protein–protein interactome hubs (Velásquez-Zapata et al. 2022) and the potential for engineering the molecular and cellular mechanisms by which key phenotypes are expressed.

Data availability

Seed of the complete WBDC ($N = 318$) can be obtained from the USDA-ARS National Small Grains Collection as accessions PI 681726 to PI 682043. Raw sequence data are deposited in the European Nucleotide Archive (ENA) under project ID PRJEB56087. The variant data are deposited in the GrainGenes database (Yao et al. 2022) at https://graingenes.org/snpversity/. Data are also available for download at FigShare: https://doi.org/10.6084/m9.figshare.30531372. SNP names from previous barley genotyping platforms (Close et al. 2009; Comadran et al. 2011 , 2012; Bayer et al. 2017) are added as annotations. Scripts used for variant calling, filtering, and other analyses can be found in GitHub repository: https://github.com/SteffensonLab/Barley_IPK_variant_calling. We used stem rust reaction type data from a previously published G3 paper (Sallam et al. 2017): https://doi.org/10.1534/g3.117.300222. These data are also included in Supplementary Table 1. Supplementary Figs. 14 and Supplementary Table 1 are available to download at G3 online.

Supplemental material available at G3 online.

Web resources

https://graingenes.org/snpversity/

Acknowledgments

A.H. acknowledges Ines Walde for expert technical assistance in the sequencing operation. B.J.S. thanks Harold Bockelman, Matthew Martin, and Tamas Szinyei for technical assistance in the handling of the wild barley materials. We acknowledge the University of Minnesota Supercomputing Institute (MSI) for providing computing resources.

Funding

The research of BJS was supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture (USDA), Hatch project (#MIN-22-085 “Exploiting Wild Relatives for Cultivated Wheat and Barley Improvement”); Lieberman-Okinow Endowment at the University of Minnesota; United States-Israel Binational Agriculture Research and Development Fund (BARD US-5089-18 “Cloning and comparative sequence analysis of powdery mildew and leaf rust resistance gene complements in wild barley”); and American Malting Barley Association. Barley genome research in the laboratory of MM, US, and NS is supported by a grant from the German Ministry of Research and Education (BMBF, SHAPE-P3, 031B0190). The wild barley research of MM is supported by a grant from the German Research Foundation (DFG, 460265804). AT was supported by European project RECOBAR (Recovering and Exploiting Old and New Barley Diversity for Future-Ready Agriculture). BK was supported by Biodiversity for Opportunities, Livelihoods and Development (BOLD) Project funded by the Government of Norway [grant number: QZA-20/0154]. BBHW was supported by funding from King Abdullah University of Science and Technology. The research of CD is supported by the Carlsberg Foundation (CF15-0236). JD, AP, and JŠ were supported from the project TowArds Next GENeration Crops, reg. no. CZ.02.01.01/00/22_008/0004581 of the European Regional Development Fund Programme Johannes Amos Comenius. For RPW, research was supported in part by the National Science Foundation—Plant Genome Research Program grant 13-39348, and USDA-Agricultural Research Service projects 3625-21000-067-00D and 5030-21220-068-000-D. For CH, RM, MAV, and JGW, research was supported by USDA-Agricultural Research Service projects 5090-21430-011-000D and

5090-43440-007-000D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA-Agricultural Research Service, or the National Science Foundation. USDA is an equal opportunity provider, employer and lender. MJM was supported by USDA-Agricultural Research Service CRIS #5062-21220-025-000D.

Conflicts of interest. None declared.

Author contributions

BJS and MM designed and led the study. AHS, GB, and JS grew and harvested the plant materials and extracted DNA. AH and AF performed the sequencing operations. RS, AHS, YL, MJ, MM, YQ, PLM, LWPA, and YG analyzed the data. BJS, RS, AHS, YL, PLM, MM, and YG drafted the manuscript and made the revisions. All authors reviewed and edited the manuscript.

Ethics declarations

Publisher's note Genetics Society of America remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Literature cited

Arora D, Gross T, Brueggeman R. 2013. Allele characterization of genes required for *rpg4*-mediated wheat stem rust resistance identifies *Rpg5* as the R gene. Phytopathology. 103:11531161. https://doi.org/10.1094/PHYTO-01-13-0030-R.

Baumdicker F *et al.* 2022. Efficient ancestry and mutation simulation with msprime 1.0. Genetics. 220:iyab229. https://doi.org/10.1093/genetics/iyab229.

Bayer MM *et al.* 2017. Development and evaluation of a barley 50k iSelect SNP array. Front Plant Sci. 8:1792. https://doi.org/10.3389/fpls.2017.01792.

Bradbury PJ *et al.* 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 23:2633–2635. https://doi.org/10.1093/bioinformatics/btm308.

Broad Institute. 2019. Picard. GItHub.

Brueggeman R *et al.* 2008. The stem rust resistance gene *Rpg5* encodes a protein with nucleotide-binding-site, leucine-rich, and protein kinase domains. Proc Natl Acad Sci U S A. 105:14970–14975. https://doi.org/10.1073/pnas.0807270105.

Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Civáň P *et al.* 2024. Genetic erosion in domesticated barley and a hypothesis of a North African centre of diversity. Ecol Evol. 14: e70068. https://doi.org/10.1002/ece3.70068.

Close TJ *et al.* 2009. Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics. 10:582. https://doi.org/10.1186/1471-2164-10-582.

Comadran J *et al.* 2011. Patterns of polymorphism and linkage disequilibrium in cultivated barley. Theor Appl Genet. 122: 523–531. https://doi.org/10.1007/s00122-010-1466-7.

Comadran J *et al.* 2012. Natural variation in a homolog of antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. Nat Genet. 44: 1388–1392. https://doi.org/10.1038/ng.2447.

Danecek P *et al.* 2021. Twelve years of SAMtools and BCFtools. GigaScience. 10:giab008. https://doi.org/10.1093/gigascience/giab008.

Fang Z *et al.* 2014. Two genomic regions contribute disproportionately to geographic differentiation in wild barley. G3 (Bethesda). 4: 1193–1203. https://doi.org/10.1534/g3.114.010561.

FAO. 2017. FAOSTAT crops. http://www.fao.org/faostat/en/#data/QC.

Franckowiak JD, Lundqvist U. 1997. New and revised names for barley genes. Barley Genetics Newsletter. 26:4–8. https://graingenes.org/ggpages/bgn/26/.

Guo Y *et al.* 2025. A haplotype-based evolutionary history of barley domestication. Nature. (accepted for publication, in press). https://doi.org/10.1038/s41586-025-09533-7.

Harlan JR, Zohary D. 1966. Distribution of wild wheats and barley. Science. 153:1074–1080. https://doi.org/10.1126/science.153.3740.1074.

Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. 2019. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience. 8: giy154. https://doi.org/10.1093/gigascience/giy154.

Jayakodi M *et al.* 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. Nature. 588:284–289. https://doi.org/10.1038/s41586-020-2947-8.

Jayakodi M *et al.* 2024. Structural variation in the pangenome of wild and domesticated barley. Nature. 636:654–662. https://doi.org/10.1038/s41586-024-08187-1.

Kono TJY *et al.* 2016. The role of deleterious substitutions in crop genomes. Mol Biol Evol. 33:2307–2317. https://doi.org/10.1093/molbev/msw102.

Kusmec A, Schnable PS. 2018. FarmCPUpp: efficient large-scale genomewide association studies. Plant Direct. 2:e00053. https://doi.org/10.1002/pld3.53.

Li B *et al.* 2024. Transcriptional convergence after repeated duplication of an amino acid transporter gene leads to the independent emergence of the black husk/pericarp trait in barley and rice. Plant Biotechnol J. 22:1282–1298. https://doi.org/10.1111/pbi.14264.

Li H *et al.* 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li H. 2013. [Heng Li—Compares BWA to other long read aligners like CUSHAW2] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [preprint]. arXiv [q-bio.GN]. https://doi.org/10.48550/arXiv.1303.3997

Liu C *et al.* 2024. Phenotypically wild barley shows evidence of introgression from cultivated barley [preprint]. bioRxiv 601622. https://doi.org/10.1101/2024.07.01.601622.

Mahalingam R, Sallam AH, Steffenson BJ, Fiedler JD, Walling JG. 2020. Genome-wide association analysis of natural variation in seed tocochromanols of barley. Plant Genome. 13:e20039. https://doi.org/10.1002/tpg2.20039.

Mascher M. 2020. Pseudomolecules and annotation of the third version of the reference genome sequence assembly of barley cv. Morex [Morex V3]. E!DAL—Plant Genomics and Phenomics Research Data Repository (PGP). https://doi.org/10.5447/ipk/2021/3.

Mascher M *et al.* 2021. Long-read sequence assembly: a technical evaluation in barley. Plant Cell. 33:1888–1906. https://doi.org/10.1093/plcell/koab077.

McKenna A *et al.* 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303. https://doi.org/10.1101/gr.107524.110.

McLaren W *et al.* 2016. The Ensembl variant effect predictor. Genome Biol. 17:122. https://doi.org/10.1186/s13059-016-0974-4.

Milner SG *et al.* 2019. Genebank genomics highlights the diversity of a global barley collection. Nat Genet. 51:319–326. https://doi.org/10.1038/s41588-018-0266-x.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2006. Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. Genetics. 173: 1705–1723. https://doi.org/10.1534/genetics.105.054502.

Roy JK *et al.* 2010. Association mapping of spot blotch resistance in wild barley. Mol Breed. 26:243–256. https://doi.org/10.1007/s11032-010-9402-8.

Russell J *et al.* 2016. Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. Nat Genet. 48:1024–1030. https://doi.org/10.1038/ng.3612.

Sallam AH *et al.* 2017. Genome-wide association mapping of stem rust resistance in *Hordeum vulgare* subsp. *spontaneum*. G3 (Bethesda). 7:3491–3507. https://doi.org/10.1534/g3.117.300222.

Schindelin J *et al.* 2012. Fiji: an open-source platform for biological-image analysis. Nat Methods. 9:676–682. https://doi.org/10.1038/nmeth.2019.

Schmid K, Kilian B, Russell J. 2018. Barley domestication, adaptation and population genomics. https://doi.org/10.1007/978-3-319-92528-8_17.

Steffenson BJ *et al.* 2007. A walk on the wild side: mining wild wheat and barley collections for rust resistance genes. Aust J Agric Res. 58:532–544. https://doi.org/10.1071/AR07123.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595. https://doi.org/10.1093/genetics/123.3.585.

Velásquez-Zapata V, Elmore JM, Fuerst G, Wise RP. 2022. An interolog-based barley interactome as an integration framework for immune signaling. Genetics. 221:iyac056. https://doi.org/10.1093/genetics/iyac056.

Walling JG *et al.* 2022. Quantitative trait loci impacting grain β-glucan content in wild barley (*Hordeum vulgare* ssp. *spontaneum*) reveals genes associated with cell wall modification and carbohydrate metabolism. Crop Sci. 62:1213–1227. https://doi.org/10.1002/csc2.20734.

Wang X *et al.* 2013. The *rpg4*-mediated resistance to wheat stem rust (*Puccinia graminis*) in barley (*Hordeum vulgare*) requires *Rpg5*, a second NBS-LRR gene, and an actin depolymerization factor. Mol Plant Microbe Interact. 26:407–418. https://doi.org/10.1094/MPMI-06-12-0146-R.

Wang J, Zhang Z. 2021. GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genomics Proteomics Bioinformatics. 19:629–640. https://doi.org/10.1016/j.gpb.2021.08.005.

Yao E *et al.* 2022. GrainGenes: a data-rich repository for small grains genetics and genomics. Database. 2022:baac034 https://doi.org/10.1093/database/baac034.

Yu J *et al.* 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38:203–208. https://doi.org/10.1038/ng1702.

Yu G, Hatta A, Periyannan S, Lagudah E, Wulff BBH. 2017. Isolation of wheat genomic DNA for gene mapping and cloning. In: Kang PB, editor. Methods in molecular biology, vol. 1659. Humana Press. p. 207–214. https://doi.org/10.1007/978-1-4939-7249-4_18.

Zohary D, Hopf M, Weiss E. 2012. Cereals. In: Domestication of plants in the old world: the origin and spread of domesticated plants in southwest Asia, Europe, and the Mediterranean basin. Oxford University Press. p. 52–58. https://doi.org/10.1093/acprof:osobl/9780199549061.001.0001.

Editor: T. Jamann