

# The Hordeum Toolbox: The Barley Coordinated Agricultural Project Genotype and Phenotype Resource

Victoria C. Blake,\* Jennifer G. Kling, Patrick M. Hayes, Jean-Luc Jannink, Suman R. Jillella, John Lee, David E. Matthews, Shiaoan Chao, Timothy J. Close, Gary J. Muehlbauer, Kevin P. Smith, Roger P. Wise, and Julie A. Dickerson

## Abstract

The use of DNA markers in public sector plant breeding is now the norm. Such markers are common across breeding programs and this commonality enables and enhances collaboration. Therefore, large collaborative research projects that measure several phenotypes across multiple environments coupled with the expanding amount of genotype data attainable with current marker technologies are on the rise and these projects demand efficient data delivery. However, development of computational tools for advanced data integration, visualization, and analysis is still a bottleneck, even though these resources have the greatest potential impact for users who are extracting and developing hypothesis-based solutions. The Hordeum Toolbox (THT) was developed as a data resource for the Barley Coordinated Agricultural Project (CAP) with the novel capability of constructing user-defined downloadable sets of phenotype and/or genotype data for downstream analysis. Internal tools in THT enable users to create clusters of a selected group of lines based on genotype data, parse pedigrees, and select germplasm based on haplotype, phenotype, and agronomic properties. The Hordeum Toolbox can be adapted to breeding programs or collaborations to assist researchers in germplasm selection, genotype data visualization, and the integration of complex data sets for statistical analysis.

**T**RADITIONALLY, plant breeders have collected phenotype data from breeding populations and used it to select for superior genotypes. Data access was limited to individual programs via spreadsheets or in-house databases. This approach has been successful in developing novel germplasm and varieties. However, with the exception of the few lines being grown in regional nurseries, the only scientists that had access to these extensive datasets were those that were intimately associated with the programs that generated the data. Therefore, there was little understanding of the relationship of germplasm between programs, and the ability to share germplasm between programs in an intelligent manner was restricted. Web-accessible databases that contain data on all germplasm within a breeding program provide the opportunity

V.C. Blake and J. Lee, USDA-ARS-WRRC, 800 Buchanan St., Albany, CA 94710; J.G. Kling and P.M. Hayes, Dep. of Crop and Soil Science, Oregon State Univ., Corvallis, OR 97331; J.-L. Jannink, USDA-REE-ARS-NAA, Robert W. Holley Center, Tower Road, Ithaca, NY 14853; S.R. Jillella, 2624 Howe Hall, Virtual Reality Applications Center, Iowa State Univ., Ames, IA 50010; D.E. Matthews, USDA-ARS, Dep. of Plant Breeding and Genetics, Cornell Univ., Ithaca, NY 14853; S. Chao, USDA-ARS-BRL, 1605 Albrecht Blvd. N, Fargo, ND 58102; T.J. Close, Dep. of Botany and Plant Sciences, Univ. of California, Riverside, CA 92521; G.J. Muehlbauer and K.P. Smith, Dep. of Agronomy and Plant Genetics, Univ. of Minnesota, St. Paul, MN 55108; R.P. Wise, Crop and Insects Genetics, Genomics, and Informatics Research Unit, USDA-ARS, Dep. of Plant Pathology and Microbiology, 411 Bessey Hall, Iowa State Univ., Ames, IA 50011; J.A. Dickerson, 3123 Coover Hall, Electrical and Computer Engineering Dep., Iowa State Univ., Ames, IA 50010. Received 19 Mar. 2012. \*Corresponding author (victoria.blake@ars.usda.gov).

Published in *The Plant Genome* 5:81–91.  
doi: 10.3835/plantgenome2012.03.0002  
© Crop Science Society of America  
5585 Guilford Rd., Madison, WI 53711 USA  
An open-access publication

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

**Abbreviations:** BOPA, barley oligo pool assay; CAP, Coordinated Agricultural Project; GWAS, genome-wide association studies; MAS, marker-assisted selection; OWB, Oregon Wolfe Barley; POPA, pilot oligo pool assay; SNP, single nucleotide polymorphism; SQL, structured query language; T3, The Triticeae Toolbox; T-CAP, Triticeae Coordinated Agricultural Project; TASSEL, trait analysis by association, evolution, and linkage; THT, The Hordeum Toolbox.

for collaboration and for sharing data and germplasm among breeding programs. Therefore, in the future, centralized databases will likely be the centerpiece for modern plant breeding programs.

Modern plant breeders use genetic marker data to predict phenotypes of novel germplasm, thus increasing the speed and efficiency of breeding while simultaneously reducing cost. These predictions are generated through the use of linked genetic markers to select individuals carrying the favorable alleles (marker-assisted selection [MAS]; e.g., Castro et al., 2003). Alternatively, for more complex traits, models are used to estimate the effects of a large number of markers distributed throughout the genome, which are subsequently used to predict the phenotypes of breeding lines, also known as genomic or genome-wide selection (e.g., Heffner et al., 2009). Both of these methods rely heavily on managing and analyzing large sets of trait and marker data. The rapid advancements in genotyping technology have radically changed the context in which MAS and genomic selection are performed. Therefore, new integrated marker and trait databases are needed to support the complex analyses essential to modern breeding efforts.

Several features of the current genetic marker technologies and the opportunities that they create are driving the demand for integrated marker and trait databases. (i) The size of data sets required for mapping and modeling marker effects is rapidly increasing. Traditionally, biparental mapping studies in barley (*Hordeum vulgare* L.) involved population sizes of several hundred individuals (e.g., Vales et al., 2005) and maps constructed from around 100 to 200 markers. Recently, approaches to genome-wide association studies (GWAS) used populations of thousands of individuals and over a million markers (e.g., Tian et al., 2011). (ii) Panels of genotypes used for mapping have gone from relatively static biparental mapping populations to ad hoc collections of breeding lines and genotypes that can be manipulated and increased in size by combining multiple data sets—this requires dynamic access to marker and trait data. Multiple users can both generate and analyze these data sets. (iii) Data analysis pathways have become more complex and time dependent to fit within critical breeding steps, resulting in the need for rapid access to appropriate data sets and analysis tools. Methods of analyses for association mapping and genomic selection are evolving and it is critical to have easy access to large data sets to reanalyze data using new approaches. (iv) To connect the results of marker–trait analyses to other plant genomic research it is necessary to link to other plant genomic resources. Taken together, integrated marker and trait databases are a necessary enabling tool to coordinate efforts between breeding programs and to fully capitalize on available genomic tools.

The Hordeum Toolbox (THT) is the barley genotype and phenotype database central to the USDA-National Institute of Food and Agriculture-funded Barley Coordinated Agricultural Project (CAP) (Waugh et al.,

2009). Derived from the Germinate database ([http://bioinf.scri.ac.uk/public/?page\\_id=159](http://bioinf.scri.ac.uk/public/?page_id=159) [accessed 19 May 2006]; Lee et al., 2005), THT was the first of its kind to integrate state-of-the-art genomics and a multi-institutional collaboration to measure agronomic, morphologic, quality, and disease parameters on large diverse populations as well as of breeder's lines and industry standard varieties over several years. User-defined datasets are delivered as files ready to load onto TASSEL (trait analysis by association, evolution, and linkage) (Bradbury et al., 2007) for association mapping or Flapjack (Milne et al., 2010) for graphical genotyping. Novel tools such as “Cluster by Genotype” and the “Haplotype Viewer” make THT a vital resource for the future of barley genome research. Interconnecting links to plant genomic resources such as PLEXdb (Plant Expression Database) (<http://plexdb.org> [accessed 19 May 2006]; Wise et al., 2007; Dash et al., 2012), HarVEST (<http://harvest.ucr.edu> [accessed 19 May 2006]; Close et al., 2007), Gramene (<http://gramene.org> [accessed 19 May 2006]; Jaiswal et al., 2006) and GrainGenes (<http://wheat.pw.usda.gov/GG2/index.shtml> [accessed 19 May 2006]; Matthews et al., 2003, Carollo et al., 2005) facilitate access to related contig alignments, oligo probe information, and a variety of gene function annotation data from the NCBI, PlantGDB (Plant Genome DataBase), TAIR (The Arabidopsis Information Resource), or rice (*Oryza sativa* L.) genome databases.

In this paper we describe the key elements of THT database developed as part of the Barley CAP (Fig. 1), including data curation and upload, main functions, analytical tools developed for THT, and a case study for downloading data. The Hordeum Toolbox is freely available at GitHub (<https://github.com/Dave-Matthews/The-Triticeae-Toolbox> [accessed 17 Nov. 2010]) and can be adapted to other crop breeding programs.

## MATERIALS AND Methods

### Database Adaptation: Germinate

The THT database is adapted from the schema for the Germinate database ([http://bioinf.scri.ac.uk/public/?page\\_id=159](http://bioinf.scri.ac.uk/public/?page_id=159); Lee et al., 2005), a generic plant data management system implemented in the public domain with the MySQL (My structured query language) relational database and freely available under the terms of the GNU public license (The GNU Operating System, <http://www.gnu.org> [accessed 19 May 19 2006]). A novel feature of Germinate, unlike any other plant data management systems at the time, was the combination of phenotypic and molecular data derived from germplasm panels, allowing queries among multiple independent data sets containing a common set of lines. Changes were made to the schema in the development of THT to generalize the structure, to deal with synonyms for line accession names and single nucleotide polymorphism (SNP) marker names, and to manipulate experimental metadata.

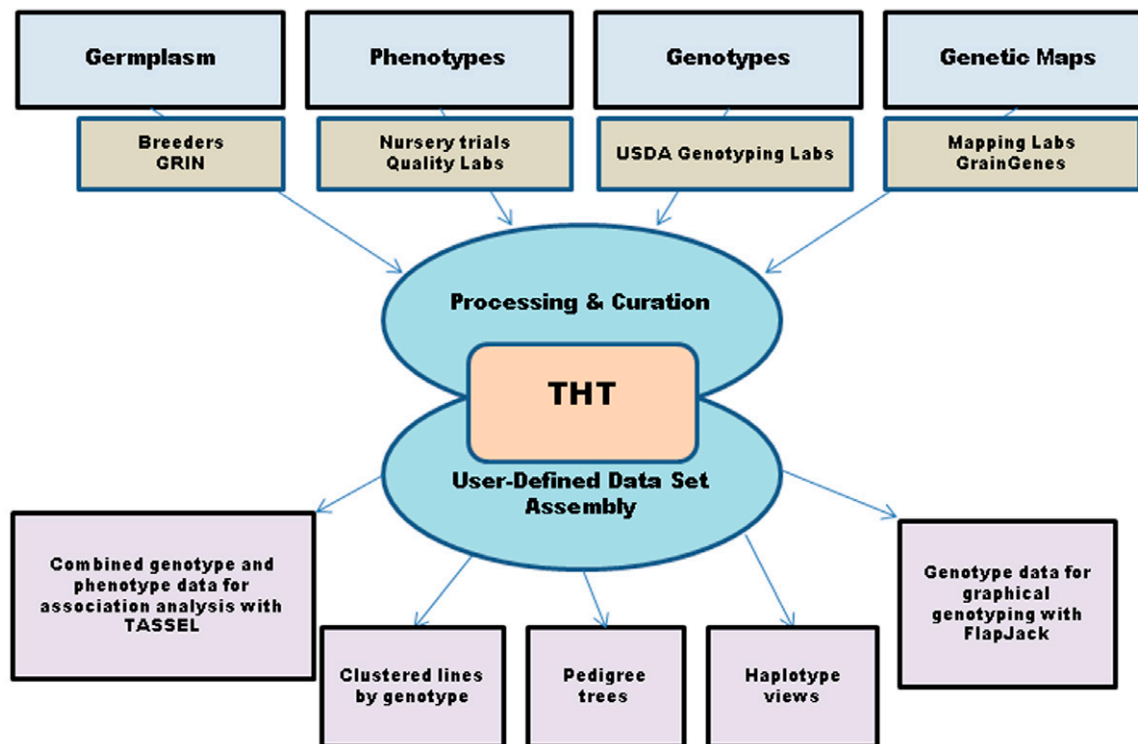


Figure 1. The Hordeum Toolbox (THT) input and output model.

## Data Acquisition

The Hordeum Toolbox combines curated genotype and phenotype data for lines selected by the 10 participating U.S. breeding programs in the Barley CAP. Each program chose 96 elite breeding lines ( $F_4$  or more advanced) each year for 4 yr: in total 3781 lines providing data. These lines were selected to be representative of the germplasm in the breeding program at the time they were submitted. Of the 10 programs, two are winter (Oregon and Virginia) and eight are spring programs; two programs breed exclusively six-row barley (Minnesota and the six-row program in North Dakota) while the others are primarily two-row or a combination (Table 1). Further detail about the participating programs can be found by visiting the THT homepage at <http://hordeumtoolbox.org> [accessed 16 June 2008], clicking on the “About THT” tab, and selecting “CAP Data Programs.”

All 10 programs planted their 96 chosen entries in replicated yield trials at locations suited to their breeding objectives, and agronomic data (yield, plant height, and heading date) were collected at all sites. Each year, all CAP germplasm lines or subsets of lines (spring, winter, or malting) were evaluated in collaborative trials for 58 other traits (Supplemental File S1). These additional 58 traits fell under the categories of agronomic, malting quality, disease, food quality, and winter growth habit traits (Table 2). In addition, data for other traits were collected when the opportunity arose. These ad hoc traits included incidental occurrence of scald or spot blotch diseases in the field. Trait definitions and measurement

units defined by the Plant Ontology (Avraham et al., 2008) were used when possible to promote more robust Triticeae ontologies and broaden the utility of THT.

All breeder’s germplasm lines as well as 94 “core” lines that included released cultivars, mapping, and key breeding parents from participating programs (Cuesta-Marcos et al., 2010; Comadran et al., 2011) were genotyped with the Illumina Golden Gate assay (Fan et al., 2003) using two 1536-SNP arrays, barley oligo pool assay (BOPA) 1 and BOPA2 (Close et al., 2009). Data generated in the development of the BOPA SNP arrays from the pilot oligo pool assays (POPAs) (POPA1, POPA2, and POPA3) are also on THT (see Supplemental File S2 for marker sequence, nomenclature, and map position). The POPAs were used to genotype the Steptoe  $\times$  Morex (Kleinbofs et al., 1993), Morex  $\times$  Barke (Close et al., 2009), and Oregon Wolfe Barley (OWB) biparental (Costa et al., 2001) doubled haploid mapping populations. These data, along with the BOPA1 panel used to genotype the Haruna Nijo  $\times$  OHU602 population (Sato et al., 2009), yielded a consensus map containing 2943 SNP loci covering a genetic distance of 1099 cM (Close et al., 2009), which is available at THT along with the Steptoe  $\times$  Morex, OWB\_2383, and Morex  $\times$  Barke genetic maps. When the allele data were entered into THT, the line names and SNP markers were verified to match entries in the database. Summary data such as the number and percentages of each allele in an experiment are computed by THT and can be used in the user’s decision on what data to include in the user-defined dataset.

**Table 1. United States breeding programs participating in the USDA Barley Coordinated Agricultural Project and phenotypic trials.**

Breeding program	Breeding objective		Lines tested	Phenotypic trials <sup>1</sup>				
	Growth habit	Row type		Agronomic	Malting quality	Disease	Food quality and miscellaneous	Winter adaptation
Busch Agricultural Resources, Inc.	Spring	2 and 6	379	87	69	68	3	
Montana State Univ.	Spring	2	384	7	9	13	3	
North Dakota State Univ. (2-row)	Spring	2	384	36	19	29	3	
North Dakota State Univ. (6-row)	Spring	6	384	38	33	29	3	
Oregon State Univ.	Winter	2 and 6	379	13	18	11	1	3
Univ. of Idaho	Spring	2 and 6	382	15	21	13	3	
Univ. of Minnesota	Spring	6	384	13	21	29	3	
Utah State Univ.	Spring	2 and 6	383	7	13	12	3	
Virginia Tech Univ.	Winter	2 and 6	339	22	19	22	3	2
Washington State Univ.	Spring	2 and 6	383	24	30	13	3	
Totals			3781	262	252	239	28	5

<sup>1</sup>Trials may have measured traits in more than one category.

**Table 2. Yearly summary of the number of experiments and collected data points for the 61 measured traits in the USDA Barley Coordinated Agricultural Project reported in The Hordeum Toolbox.**

	Experiments				Data points				Germplasm group <sup>†</sup>
	2006	2007	2008	2009	2006	2007	2008	2009	
<b><i>Agronomic Traits</i></b>									
Grain yield, kg ha <sup>-1</sup>	76	70	31	25	5004	2487	1329	1414	All
Head drop, 0–9	1	0	0	0	737	0	0	0	Spring
Head shattering, %	1	0	0	0	425	0	0	0	Spring
Heading date, days	68	67	32	25	5806	3258	1719	1414	All
Lodging, %	48	21	22	14	1841	549	810	666	All
Plant height, cm	72	67	31	25	5249	3314	1329	1414	All
Stem length, cm	26	19	16	8	1104	932	710	420	All
Straw breakage, %	0	2	0	0	0	55	0	0	Spring
Test weight, g L <sup>-1</sup>	31	22	15	7	2931	997	600	310	All
<b><i>Malting Quality Traits</i></b>									
Alpha amylase, 20°DU <sup>‡</sup>	46	58	16	16	1942	1901	544	580	Malt
Barley color, °ASBC <sup>§</sup>	33	20	16	16	1585	1038	544	580	Malt
Beta-glucan, %	3	3	2	2	956	960	1035	864	All
Beta-glucanase activity, U kg <sup>-1</sup> malt	12	6	0	0	491	277	0	0	Malt
Beta-glucanase thermostability, U kg <sup>-1</sup> malt	12	6	0	0	491	277	0	0	Malt
Breeders grain protein, %	0	7	4	1	0	320	205	100	Ad hoc
Breeders plump grain, % by weight on 2.4 mm (6/64") sieve	0	7	4	7	0	320	205	310	Ad hoc
Diastatic power, °ASBC	46	58	16	16	1942	1901	544	580	Malt
Grain protein, %	55	66	20	16	4000	2107	684	580	Malt
Kernel weight, mg	33	20	16	16	1585	1038	544	580	Malt
Lipoxygenase activity, U g <sup>-1</sup> malt	0	1	1	0	0	287	288	0	Malt
Malt beta-glucan, mg kg <sup>-1</sup>	46	58	16	16	1942	1900	544	580	Malt
Malt extract, %	46	58	16	16	1940	1901	544	580	Malt
Malt protein, %	0	38	0	0	0	863	0	0	Malt
Nondormant seeds, %	1	1	1	0	813	746	718	0	Spring
Plump grain, % by weight on 2.4 mm (6/64") sieve	53	59	22	12	4213	2001	782	440	All
Residual beta-glucanase, %	12	6	0	0	492	277	0	0	Malt
Soluble protein:total protein, %	46	58	16	16	1940	1901	544	580	Malt

(cont'd)

## Data Processing

Barley CAP collaborators conducted their own statistical analysis and reported the means for each genotype and summary statistics for the trials. Germplasm lines are assigned a unique identifier (UID) by THT, as are trial codes, traits, or any other field that will be routinely

accessed. The trial code provides a direct link to the experimental annotations and allows variables to be loaded separately from the same trial. For example, agronomic traits could be submitted directly after harvest, with malting quality traits reported at a later time. Comparisons across experiments are facilitated by the use of standard check cultivars and summary statistics for each

**Table 2. Continued.**

	Experiments				Data points				Germplasm group <sup>1</sup>
	2006	2007	2008	2009	2006	2007	2008	2009	
Wort color, °ASBC	33	20	16	16	1523	951	531	577	Malt
Wort protein, %	46	58	16	16	1940	1901	543	580	Malt
<b><i>Disease Traits</i></b>									
Barley yellow dwarf rating, 0–8	1	0	0	0	880	0	0	0	All
Common root rot severity, %	1	1	0	0	386	390	0	0	Spring
Deoxynivalenol, mg kg <sup>-1</sup>	3	7	11	8	1161	716	1382	1124	Spring (Midwest)
Fusarium head blight (FHB) incidence, %	4	0	0	0	382	0	0	0	Spring
FHB reaction type, 0–9	1	2	0	0	96	55	0	0	Spring
FHB severity, %	3	6	11	8	1156	1159	1381	1137	Spring (Midwest)
Leaf rust, 0–9	6	6	0	0	136	152	0	0	Winter
Leaf rust seedling, 0–4, with qualifiers	1	0	0	0	951	0	0	0	All
Net blotch, 1–10	11	9	0	0	1072	882	0	0	All
Net blotch net form reaction type, 0–9	12	4	0	0	202	110	0	0	Spring
Powdery mildew, 0–4	1	0	0	0	913	0	0	0	All
Scald, 0–8	3	0	0	0	90	0	0	0	Ad hoc
Scald reaction type, 0–9	18	7	0	0	540	133	0	0	Ad hoc
Septoria seedling infection response, 0–5	1	2	1	0	958	939	863	0	All
Septoria speckled leaf blotch, 0–9	0	2	0	0	0	55	0	0	Ad hoc
Spot blotch infection coefficient, %	1	1	1	0	673	771	768	0	Spring
Spot blotch reaction type, 0–9	5	0	0	0	23	0	0	0	Ad hoc
Spot blotch seedling infection response, 1–9	1	2	1	0	957	939	863	0	All
Spot blotch severity, %	4	1	1	0	763	771	768	0	Spring
Stripe rust severity, %	4	0	0	0	195	0	0	0	Winter
<b><i>Food Quality and Miscellaneous Traits</i></b>									
Amylose content, % dry weight basis	0	2	1	0	0	938	864	0	All
Grain hardness, SKCS <sup>‡</sup>	0	2	1	1	0	938	864	859	All
Grain width, mm	0	2	1	1	0	938	864	862	All
Grain weight, mg	0	2	1	1	0	937	864	862	All
Hull proportion, %	0	2	1	1	0	938	803	764	All
Kernels per spike	2	0	0	0	769	0	0	0	Spring
Phenolic compound content, %	0	2	1	0	0	938	864	0	All
Polyphenol oxidase activity, abs <sup>#</sup>	0	2	1	1	0	938	864	862	All
<b><i>Winter Adaptation Traits</i></b>									
Fall planting heading date, days	2	2	0	0	194	155	0	0	Winter
Spring planting heading date, days	2	2	0	0	194	155	0	0	Winter
Vernalization score, days	2	2	0	0	194	155	0	0	Winter
Winter hardiness, % survival	1	0	0	0	98	0	0	0	Winter

<sup>1</sup>All, winter, or spring subsets of germplasm were evaluated for different traits. The Ad hoc category includes traits that were observed in experiments not necessarily designed for that trait.

<sup>‡</sup>20°DU, dextrinizing units at 20°C.

<sup>§</sup>ASBC, American Society of Brewing Chemists standard.

<sup>‡</sup>SKCS, single kernel characterization system.

<sup>#</sup>abs, absorbance at 480 nm.

	Markers				Trait	Expts	Markers				
	1	2	3	4			1	2	3	4	
05WA-357.1	AA	BB	AA	BB	14.9	2	2	11_11223			
S47/E//S47-43	AA	BB	BB	BB	14.7	3	3	11_20502			
ND20798-12	AA	BB	AA	AA	14.5	5	4	11_21067			
2ND26384	AA	BB	AA	AA	14.4	1					
05WA-315.47	AA	BB	BB	BB	14.3	2					
S47/E//K51-29	AA	BB	BB	BB	14.3	2					
ND24335	AA	BB	AA	BB	14.3	2					
VA05B-93	AA	BB	AA	BB	14.2	1					
05WA-360.14	AA	BB	BB	BB	14.2	2					
S47/E//K47/E-4\	AA	BB	BB	BB	14.0	4					
ND24379	AA	BB	AA	BB	14.0	5					
05WA-316.32	AA	BB	AA	BB	14.0	2					
MT050081	AA	BB	AA	BB	14.0	7					
Z031P134Q	AA	BB	BB	BB	14.0	3					

Trait
grain protein

Experiments
ORELT_2006_Corvallis
ORELT_2006_Pendleton
ORELT_2006_Parma
ORELT_2006_Pullman
ORELT_2006_Filer
OSUCAPYT_2006_Aberdeen07
OSUCAPCB_2006_Corvallis07

Figure 2. The “Select Lines by Haplotype” analytical tool identifies germplasm lines that carry desired alleles. Here, alleles for two markers were designated (AA for 12\_30969 and BB for 11\_11223) and “any” alleles of two others (11\_20502 and 11\_22067) were allowed. Experimental means from the selected phenotypic trait “grain protein” are shown to the right of the allele chart with number and name of experiments analyzing this trait on the selected lines also reported.

trial, including the trial mean and number of replications that were measured for each trait. For replicated traits, the standard error of a mean and the probability value for the *F* test for genotypes from the analysis of variance or mixed model analysis were reported. The range of each variable was determined from the raw data sets. For genotypes that had values that were four or more standard deviations from the mean, results were compared across replications. Potential outliers were identified and inquiries were sent to the collaborator who submitted the data for consideration. The Hordeum Toolbox stores the original data sets (raw data) without modification for archive purposes.

### Data Access Levels

The Hordeum Toolbox defines four levels of user access, controlling what data are visible, what data can be added or edited, and what menus are available to manipulate the database. During the data acquisition period of the barley CAP, the public could view most of the data immediately, with the exception of data flagged for a brief delay in general release. Registered CAP participants had permission to view all data when logged onto THT. The curator’s access level allowed participants to upload data files and edit existing data on a record-by-record basis, using the data input functions in the Curation Menu. Curation status also allows users to access the Database Menu, which contains tools to generate reports on database content, review the schema, export data, and clean up temporary files. The administrator can grant curator-level access and perform database administration tasks using the Database menu.

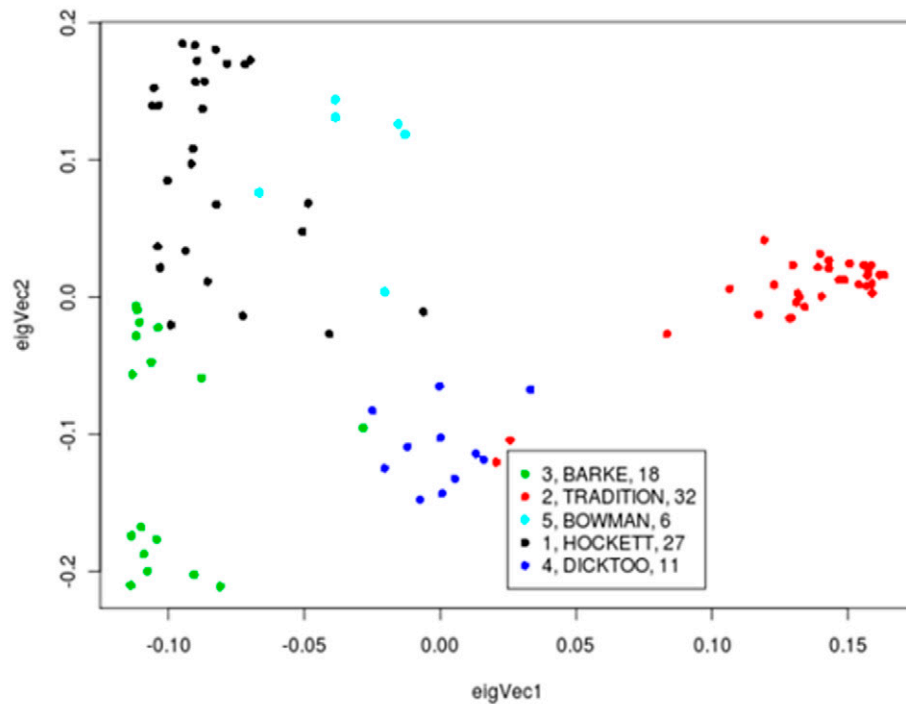
### Data Curation and Upload

In the development of THT, scripts for data curation were written for direct internet upload. The upload tools provide another layer of checking for trial code tracking, data ranges, line names, line aliases, CAP data programs, marker names, marker synonyms, and trait definitions. The curator can also interact directly with THT to add or modify traits and experiment annotations in real time. Most data load within a few seconds. Errors in the content or structure of the data generate specific messages usually enabling the curator to locate and correct errors in the data quickly. For very large datasets, which can take hours to load, processing is “off-line” with the success or failure of the data load reported via e-mail to the curator.

### Analytical Tool Development

The data are stored in THT in a MySQL database with the web interface generated using a combination of JavaScript and hypertext preprocessor (PHP) scripts affording a wealth of possibilities for querying and data delivery. For example, association analysis may indicate that a small number of markers are associated with favorable values of a trait or traits. The “Select Lines by Haplotypes” SQL query identifies all germplasm lines that carry the desired alleles of any combination of markers (Fig. 2). In addition, the phenotype value for a trait of interest can be displayed for each line.

The “Cluster by Genotype” analytical tool (Fig. 3) was designed to perform a clustering of a selected set of barley germplasm lines based on their alleles for all markers in the database, resulting in a two-dimensional projection, color coded by cluster. One or



Cluster	Labeled lines	Lines
1	HOCKETT	27
2	TRADITION	32
3	BARKE	18
4	DICKTOO	11
5	BOWMAN	6
<b>Total:</b>		<b>94</b>

Cluster	Lines
1	2B96-5038, 2B98-5312, AC_METCALFE, ARAPILES, B1202, B1215, BCD12, BCD47, CDC_COPELAND, CDC_KENDALL, CDC_STRATUS, CIHO4196, CONRAD, CRAFT, CREST, FARMINGTON, FLAGSHIP, FRANKLIN, GARNET, HARRINGTON, HARUNA_NIJO, HOCKETT, KLAGES, MERIT, NEWDALE, SUBLETTE, TR306
2	6B00-1526, 6B02-3394, 6B94-7378, 6B94-8253, 6B97-2245, 88AB536, 88AB536-B, B1602, B1614, CDC_SISLER, DRUMMOND, EXCEL, FEG55-14, FEG59-09, FEG66-08, FEG90-31, FOSTER, LACEY, LARKER, LEGACY, M122, M123, MNBRITE, MOREX, ND20508, NDB112, OWB_RECESSIVE, ROBUST, STANDER, STELLAR, TRADITION, WASHFORD
3	BARKE, BARONESSE, BISON1H, BISON1H+4H, BISON1H+4H+5H, BISON1H+5H, BISON4H, BISON4H+5H, BISON5H, BISON7H, C-14, CANELA, COLLINS, ESLICK, GERALDINE, HAYS, PASADENA, SCARLETT
4	BELFORD, DICKTOO, HOODY, KOLD, KOMPOLTI, OWB_DOMINANT, SHENMAI_3, STEPTOE, STRIDER, SUSSEX, WA1614-95
5	BOWMAN, CONLON, HAXBY, ND21863, ORCA, RAWSON

Select the clusters you want to use.

 1  
 2  
 3  
 4  
 5

Select

Figure 3. The “Cluster Lines by Genotype” analytical tool clusters a selected set of barley germplasm based on their alleles for all markers in The Hordeum Toolbox. Here the 96 members of the USDA Barley Coordinated Agricultural Project Core Mapping set are shown partitioned into five clusters, each of which can be selected independently for further analysis.

more clusters can be selected for further examination such as reclustering or export to TASSEL (Bradbury et al., 2007). For the analysis, Illumina A and B allele calls are replaced by the numerical values 0 and 1, and then missing data points for a marker are imputed as the mean marker score. For clustering, the first two principal component analysis eigenvectors are extracted from the line by marker data matrix. Each barley line is plotted as a point in a scatter graph with eigenvector scores for the  $x$  and  $y$  coordinates. In parallel, the lines are clustered in two to eight clusters using the partitioning around

medoids method (Kaufman and Rousseeuw, 1990). This method was chosen because it is fast and robust. The points in the graph are colored according to the cluster to which the corresponding line belongs. Names of known lines can be given to the analysis to construct a legend allowing clusters to be associated with those lines.

## Results and Discussion

After 4 yr of serving as a repository for the Barley CAP project, The Hordeum Toolbox database now contains 4216 germplasm line records, 3781 from the breeding

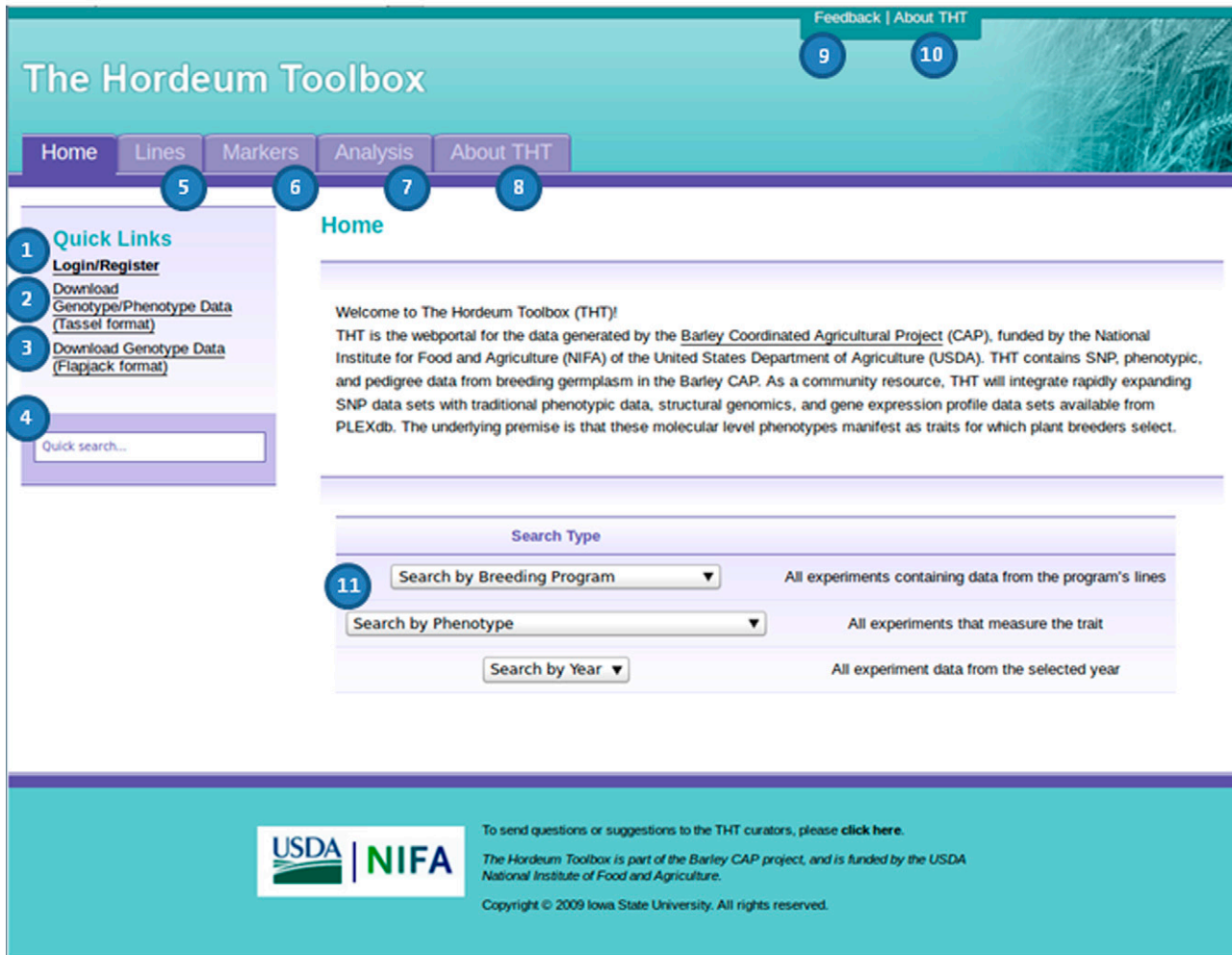


Figure 4. A screenshot of The Hordeum Toolbox (THT) homepage at <http://hordeumtoolbox.org> with key features and menus annotated. 1) Registered User and Curator Login. 2) TASSEL (trait analysis by association, evolution, and linkage) (Bradbury et al., 2007) download gateway for association mapping. 3) Flapjack (Milne et al., 2010) download gateway for visual genotyping. 4) Quick Search, a flexible search tool for all THT data. 5) Links to interactive menus to select germplasm lines by properties (end use, growth habit, row type, etc.) and/or by phenotype within a defined range. Includes access to the pedigree search page. 6) Marker selection interface by text input or mapping position, genetic map information and single nucleotide polymorphism allele sequences. 7) Links to the novel THT tools: Cluster Lines by Genotype, View Haplotypes, and Parse Purdy Pedigrees. 8) Descriptions of the Coordinated Agricultural Project (CAP) data programs and links to contributed data. 9) Suggestions e-mailed directly to the curator group. 10) Data access policy, software availability, development team, and collaborator credits. 11) Dropdown menus to navigate to data by breeding program, measured phenotype, and experiment year.

programs and 435 from mapping populations. Of these lines, 4209 have genotype data and 3701 have phenotype data. The Hordeum Toolbox contains sequence information for 5609 molecular markers, 4595 with genotyping data for a total of 14,114,103 genotype data points. Phenotype data from 417 separate experimental trials measuring any of 61 individual traits have provided 159,799 data points (Tables 1 and 2).

The design of THT has enabled researchers to have simple and rapid access to data. Pathways to data by phenotype, breeding program, or year, links to germplasm line and marker selection by user-defined criteria, analytical tools, and quick links to complex searches are all easily accessible from the THT homepage (<http://hordeumtoolbox.org>; Fig. 4). For simple searches, individual

lines, markers, and experiments can be accessed via the quick search text box on the homepage. The strength of THT, however, is the ability to build user-defined sets of data. The four basic data types are germplasm lines, genetic markers, phenotype or genotype experimental trials, and measured traits. The Hordeum Toolbox users can build unique datasets for download containing any or all of the available data for a given category.

### Germplasm Selection

Germplasm lines can be selected directly from a pasted list or by interactive menus to select lines by property or by their phenotypic values. Information for each line, available on a summary page, include properties (see below) synonyms, links to the Germplasm Resources Information Network



(GRIN) (Wiersema, 1995) when appropriate, pedigree (if available), and links to genotype and phenotype data.

### Select Lines by Properties

This interface uses seven categorical variables stored for each line, including the line name or synonym, breeding program, year included in the CAP, and primary end use (malt, food, forage, feed, and/or genetic stock) as well as the genetic properties: growth habit (spring or winter), inflorescence row type (2-row vs. 6-row), and hulled vs. hullless. For this search, any combination of values for any of the variables can be selected. The resulting set of lines can then be selected by the user and is stored in a buffer by THT. Further line selection by properties or phenotypes will add to, replace, or be selected from this original set, depending on the user's selection once additional searches are performed.

### Select Lines by Phenotype

This tool allows the user to select sets and subsets of lines based on quantitative traits and allows the selection on a range of values for a trait such as grain protein or heading date. For example, a user can select all lines for all 4 yr with a desirable grain protein value (e.g., 10–11%). They can then search for all lines in all years with a low leaf rust disease incidence (e.g., 0–4). Once those lines are queried, selecting the “Intersect (AND)” radio button on the results page will produce a set of lines fulfilling both parameters.

### Marker Selection

An individual marker record in THT contains the SNP sequence, nucleotides represented by A or B alleles for BOPA1 and BOPA2 data (Close et al., 2009), and synonyms and annotations to outside databases such as GrainGenes (<http://wheat.pw.usda.gov/GG2/index.shtml>) and HarVEST (<http://harvest.ucr.edu>). Lists of markers or their synonyms selected for further analysis can be manually entered into a “Search by Name” text box in the marker selection menu page. Marker sets can also be created by selecting a chromosome from one of the genetic maps in THT, such as the Steptoe × Morex map. When selecting a chromosome, the user can narrow the range of markers to a specific region on the map if desired. Before download, markers can be further narrowed on the basis of minor allele frequency to eliminate those with limited polymorphism and the amount of missing data.

### Cluster Lines by Genotype

Evaluation of relatedness or similarity among breeding lines is increasingly done on the basis of genotypic rather than pedigree data. The “Cluster by Genotype” analysis strives to be a visually rich implementation of this evaluation (Fig. 3). Two clear cases for its use are (i) the user has a panel of lines that will serve as a basis for association analysis. To get a feel for the extent of structure in the lines, the user uses Cluster by Genotype to visualize that structure; and (ii) the user has selected lines from different breeding programs knowing that they ultimately want to select a panel of lines from across the

different programs that are genetically the most similar, irrespective of their origin. The result is a visually informative presentation of differences and groupings among the currently selected lines. The user may take the next step of choosing one or more clusters and restricting the selected set of lines to the lines in those clusters.

### Haplotype Viewer

The haplotype viewer allows the THT user to identify barley lines that carry a specific combination of alleles at a limited number of markers, assuming that known major-effect QTL underlie those markers. Although the number of markers is not restricted by THT, the database performance is impacted if many markers are selected at a time. Once the user has selected markers in the THT buffer the “View Haplotypes” button links to a panel of dropdown menus for each marker to specify the desired allele state. At that stage, the user can also select phenotype trials, and the mean of the results for each given line will be displayed next to the allele state (Fig. 2).

### User-Defined Datasets for TASSEL and Flapjack

In the last few years, a wealth of genomic data coupled with fast and cheap computational power have enabled plant scientists to begin to look at the entire genome when designing breeding projects and searching for genes of interest. Software packages to assist in this effort are in the public domain and THT has been working with their developers to deliver user-defined data sets to use with these new tools. Specifically, THT can produce downloadable text files that can be directly loaded into TASSEL (Bradbury et al., 2007) for association analysis or loaded into Flapjack (Milne et al., 2010) for graphical genotyping.

The flexibility in designing the files to load into TASSEL (Bradbury et al., 2007) or Flapjack (Milne et al., 2010) reveal the strength and versatility of THT, and the simplicity of the downloaded files enables users of other analysis packages (e.g., JMP Genomics 6.0 (SAS Institute, 2010)) to use the data without extensive manipulation. The ability to create specific subsets of the germplasm lines is particularly important because population structure due to relatedness by descent can bias the statistical association between traits and markers as well as lead to spurious associations (Thornsberry et al., 2001). The gateways to data selection are under the “Quick Links” menu on the THT homepage (see Supplemental File S3 for a step-by-step creation of a TASSEL-ready file and subsequent analysis of results in THT).

### Pedigree

Pedigree information is stored as a text string using standard Purdy notation (Purdy et al., 1968). In the first phase of the Barley CAP, pedigrees were manually converted into a tabulated format that could be readily exported and used to calculate co-ancestries among lines. Data input for each cross included the names of both parents, the genetic contributions expected, and the level of inbreeding of each parent. The genetic contribution was assumed to be 0.5 but could

be modified to indicate additional generations of backcrossing. To avoid pitfalls, extensive discussions with breeders were necessary, since pedigree records are often inconsistent or incomplete (e.g., intermediate parents in pedigrees are often unknown and levels of inbreeding of those parents were seldom reported). Another complication is that barley varieties used in crosses differ considerably in their level of genetic uniformity. Some may have come from a single spike while others are essentially bulks of partially inbred populations. Although pedigree information is useful, the ability to collect accurate and extensive pedigrees was limiting. Therefore, we chose to discontinue collecting pedigree information and rely on the genetic data.

## Outlook

The THT bioinformatics tool was developed to help plant breeders more easily access and use our crop genetic resources. One of the challenges with this approach has been the tendency for collaborators to use long-established methods for data collection rather than standardized database protocols that would permit analyses across diverse breeding programs. Since its inception, THT has provided data for several barley genomics studies including the analysis of population structure in barley breeding populations using multilocus SNP data (Hamblin et al., 2010) and the assessment of population size and unbalanced data sets on GWAS (Wang et al., 2012). The success and ease of data handling in THT prompted the adoption of the database to the next generation of CAP projects, the Triticeae CAP (T-CAP) (<http://www.triticeaecap.org/> [accessed 1 Feb. 2011]), creating The Triticeae Toolbox (T3). The data in the THT database were imported into the barley T-CAP database (T3 Barley) and is available at <http://triticeaetoolbox.org/barley> (accessed 1 Feb. 2011) while a parallel database for wheat (*Triticum aestivum* L.) (T3 Wheat) at <http://triticeaetoolbox.org/wheat> (accessed 1 Feb. 2011) was built on the THT model with slight modifications, primarily in the properties of germplasm lines and measured phenotypic traits.

The flexible structure of THT makes it easy to add new species, traits, and data types into the database. The value of THT and its progeny databases, such as T3 Barley and T3 Wheat, will be continuously enhanced by adding links to rapidly growing statistical tools written in R (R Development Core Team, 2006) and by the addition of enhanced graphical data representation and summarization tools now being written, promising the utility of The Hordeum Toolbox for years to come.

## Supplemental Information Available

Supplemental material is available at <http://www.crops.org/publications/tpg>.

Supplemental File S1. Collaborative trials and large datasets in The Hordeum Toolbox (THT).

Supplemental File S2. Barley oligo pool assay (BOPA) marker nomenclature, map position, and sequence.

Supplemental File S3. Case study with The Hordeum Toolbox (THT) screenshots to guide the user through an exercise to select barley germplasm.

## Acknowledgments

The authors thank the outstanding contributions by Iowa State University undergraduate students who helped create the THT interface software: Gavin Monroe, Ethan Wilder, and Yong Huang. This research was supported by USDA-CSREES-NRI Grant No. 2006-55606-16722 and USDA-NIFA Grant No. 2009-85606-05701, "Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics, and Breeding for Gene Discovery and Barley Improvement." Maintenance and further development of THT by GrainGenes is supported by USDA-ARS project 5325-21000-014-00, "An Integrated Database and Bioinformatics Resource for Small Grains." This publication was supported by funds received from the "Triticeae Coordinated Agricultural Project (TCAP)" USDA-NIFA Grant no. 2011-68002-30029. 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 U.S. Department of Agriculture. Upon request, all novel materials described in this publication will be made available in a timely manner for noncommercial research purposes, subject to the requisite permission from any third-party owners of all or parts of the material. Obtaining any permissions will be the responsibility of the requestor.

## References

- Avraham, S., C.-W. Tung, K. Ilic, P. Jaiswal, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, F. Zapata, and D. Ware. 2008. The plant ontology database: A community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.* 36(S1):D449–D454. doi:10.1093/nar/gkm908
- Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635. doi:10.1093/bioinformatics/btm308
- Carollo, V., D.E. Matthews, G.R. Lazo, T.K. Blake, D.D. Hummel, N. Lui, D.L. Hane, and O.D. Anderson. 2005. GrainGenes 2.0. An improved resource for the small-grains community. *Plant Physiol.* 139:643–651. doi:10.1104/pp.105.064485
- Castro, A., F. Capetini, A. Corey, T. Filichia, P.M. Hayes, A. Kleinhofs, D. Kudrna, K. Richardson, S. Sandoval-Islas, C. Rossi, and H. Vivar. 2003. Mapping and pyramiding of qualitative and quantitative resistance to stripe rust in barley. *Theor. Appl. Genet.* 107:922–930. doi:10.1007/s00122-003-1329-6
- Close, T.J., P.R. Bhat, S. Lonardi, Y. Wu, N. Rostoks, L. Ramsay, A. Druka, N. Stein, J.T. Svensson, S. Wanamaker, S. Bozdag, M.L. Roose, M.J. Moscou, S. Chao, R.K. Varshney, P. Szucs, K. Sato, P.M. Hayes, D.E. Matthews, A. Kleinhofs, G.J. Muehlbauer, J. DeYoung, D.F. Marshall, K. Madishetty, R.D. Fenton, P. Condamine, A. Graner, and R. Waugh. 2009. Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* 10:582. doi:10.1186/1471-2164-10-582
- Close, T.J., S. Wanamaker, M.L. Roose, and M. Lyon. 2007. HarvEST: An EST database and viewing software. In: D. Edwards, editor, *Plant bioinformatics: Methods in molecular biology*. Humana Press, New York, NY, p. 161–178.
- Comadran, J., L. Ramsay, K. MacKenzie, P.M. Hayes, T.J. Close, G.J. Muehlbauer, N. Stein, and R. Waugh. 2011. Patterns of polymorphism and linkage disequilibrium in cultivated barley. *Theor. Appl. Genet.* 122:523–531. doi:10.1007/s00122-010-1466-7
- Costa, J.M., A. Corey, P.M. Hayes, C. Jobet, A. Kleinhofs, A. Kopsich-Obusch, S.F. Kramer, D. Kudrna, M. Li, O. Riera-Lizarazu, K. Sato, P. Szűcs, T. Toojinda, M.I. Vales, and R.I. Wolfe. 2001. Molecular mapping of the Oregon Wolfe barleys: A phenotypically polymorphic doubled-haploid population. *Theor. Appl. Genet.* 103:415–424. doi:10.1007/s001220100622
- Cuesta-Marcos, A., P. Szucs, T.J. Close, T. Filichkin, G.J. Muehlbauer, K.P. Smith, and P.M. Hayes. 2010. Genome-wide SNPs and re-sequencing of growth habit and inflorescence genes in barley: Implications for association mapping in germplasm arrays varying in size and structure. *BMC Genomics* 11:707. doi:10.1186/1471-2164-11-707

- Dash, S., J. Van Hemert, L. Hong, R.P. Wise, and J.A. Dickerson. 2012. PLEXdb: Gene expression resources for plants and plant pathogens. *Nucleic Acids Res.* 40(D1):D1194–D1201. doi:10.1093/nar/gkr938
- Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, and M.S. Chee. 2003. Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.* 68:69–78. doi:10.1101/sqb.2003.68.69
- Hamblin, M.T., T.J. Close, P.R. Bhat, S. Chao, J.G. Kling, K.J. Abraham, T. Blake, W.S. Brooks, B. Cooper, C.A. Griffey, P.M. Hayes, D.J. Hole, R.D. Horsley, D.E. Obert, K.P. Smith, S.E. Ullrich, G.J. Muehlbauer, and J.-L. Jannink. 2010. Population structure and linkage disequilibrium in U.S. barley germplasm: Implications for association mapping. *Crop Sci.* 50:556–566. doi:10.2135/cropsci2009.04.0198
- Heffner, E.L., M. Sorrells, and J.-L. Jannink. 2009. Genomic selection for crop improvement. *Crop Sci.* 49:1–12. doi:10.2135/cropsci2008.08.0512
- Jaiswal, P., J. Ni, I. Yap, D. Ware, W. Spooner, K. Youens-Clark, L. Ren, C. Liang, W. Zhao, K. Ratnapu, B. Faga, P. Canaran, M. Fogleman, C. Hebbard, S. Avraham, S. Schmidt, T.M. Casstevens, E.S. Buckler, L. Stein, and S. McCouch. 2006. Gramene: A bird's eye view of cereal genomes. *Nucleic Acids Res.* 34(S1):D717–D723. doi:10.1093/nar/gkj154
- Kaufman, L., and P.J. Rousseeuw. 1990. Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, Hoboken, NJ.
- Kleinhofs, A., A. Kilian, M.A. Saghai Maroof, R.M. Biyashev, P. Hayes, F.Q. Chen, N. Lapitan, A. Fenwick, T.K. Blake, V. Kanazin, E. Ananiev, L. Dahleen, D. Kudrna, J. Bollinger, S.J. Knapp, B. Liu, M. Sorrells, M. Heun, J.D. Franckowiak, D. Hoffman, R. Skadsen, and B.J. Steffenson. 1993. A molecular, isozyme and morphological map of the barley (*Hordeum vulgare*) genome. *Theor. Appl. Genet.* 86:705–712. doi:10.1007/BF00222660
- Lee, J.M., G.F. Davenport, D. Marshall, T.H. Ellis, M.J. Ambrose, J. Dicks, T.J. van Hintum, and A.J. Flavell. 2005. GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. *Plant Physiol.* 139:619–631. doi:10.1104/pp.105.065201
- Matthews, D.E., V.L. Carollo, G.R. Lazo, and O.D. Anderson. 2003. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.* 31:183–186. doi:10.1093/nar/gkg058
- Milne, I., P. Shaw, G. Stephen, M. Bayer, L. Cardle, W.T.B. Thomas, A.J. Flavell, and D. Marshall. 2010. Flapjack – Graphical genotype visualization. *Bioinformatics* 26:3133–3134. doi:10.1093/bioinformatics/btq580
- Purdy, L.H., W.Q. Loegering, C.F. Konzak, and C.J. Peterson. 1968. A proposed standard method for illustrating pedigrees of small grain varieties. *Crop Sci.* 8:405–406. doi:10.2135/cropsci1968.0011183X00800040002x
- R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org (accessed 1 Feb. 2011).
- SAS Institute. 2012. JMP Genomics. Release 6.0. SAS Inst., Cary, NC.
- Sato, K., N. Nankaku, and K. Takeda. 2009. A high density transcript linkage map of barley derived from a single population. *Heredity* 103:110–117. doi:10.1038/hdy.2009.57
- Thornberry, J.M., M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. *Nat. Genet.* 28:286–289. doi:10.1038/90135
- Tian, F., P.J. Bradbury, P.J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T.R. Rocheford, M.D. McMullen, J.B. Holland, and E.S. Buckler. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43:159–162. doi:10.1038/ng.746
- Vales, M.I., C.C. Schön, F. Capetini, X.M. Chen, A.E. Corey, D.E. Mather, C.C. Mundt, K.L. Richardson, J.S. Sandoval-Islas, H.F. Utz, and P.M. Hayes. 2005. Effect of population size on the estimation of QTL: A test using resistance to barley stripe rust. *Theor. Appl. Genet.* 111:1260–1270. doi:10.1007/s00122-005-0043-y
- Wang, H., K.P. Smith, E. Combs, T. Blake, R.D. Horsley, and G.J. Muehlbauer. 2012. Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theor. Appl. Genet.* 124:111–124. doi:10.1007/s00122-011-1691-8
- Waugh, R., J.-L. Jannink, G.J. Muehlbauer, and L. Ramsey. 2009. The emergence of whole genome association scans in barley. *Curr. Opin. Plant Biol.* 12:218–222. doi:10.1016/j.pbi.2008.12.007
- Wiersema, J.H. 1995. Taxonomic information on cultivated plants in the USDA/ARS germplasm resources information network (GRIN). *Acta Hort.* 413:109–116.
- Wise, R.P., R.A. Caldo, L. Hong, L. Shen, E. Cannon, and J.A. Dickerson. 2007. BarleyBase/PLEXdb. A unified expression profiling database for plants and plant pathogens. *Methods Mol. Biol.* 406:347–363. doi:10.1007/978-1-59745-535-0\_17