The Triticeae Toolbox: Combining Phenotype and Genotype Data to Advance Small-Grains Breeding

Victoria C. Blake, Clay Birkett, David E. Matthews, David L. Hane, Peter Bradbury, and Jean-Luc Jannink*

Abstract

The Triticeae Toolbox (http://triticeaetoolbox.org; T3) is the database schema enabling plant breeders and researchers to combine, visualize, and interrogate the wealth of phenotype and genotype data generated by the Triticeae Coordinated Agricultural Project (TCAP). T3 enables users to define specific data sets for download in formats compatible with the external tools TASSEL, Flapjack, and R; or to use by software residing on the T3 server for operations such as Genome Wide Association and Genomic Prediction. New T3 tools to assist plant breeders include a Selection Index Generator, analytical tools to compare phenotype trials using common or user-defined indices, and a histogram generator for nursery reports, with applications using the Android OS, and a Field Plot Layout Designer in development. Researchers using T3 will soon enjoy the ability to design training sets, define core germplasm sets, and perform multivariate analysis. An increased collaboration with GrainGenes and integration with the small grains reference sequence resources will place T3 in a pivotal role for on-the-fly data analysis, with instant access to the knowledge databases for wheat and barley. T3 software is available under the GNU General Public License and is freely downloadable.

The Triticeae Toolbox

A Brief History

The last generation of plant breeders have employed marker-assisted selection (Dubcovsky, 2004) wherein each line is evaluated with only a few genetic markers. New technologies have shifted the paradigm to an analysis of high-density (between a thousand and a million) genome-wide markers at equal or lower cost. In addition to high-throughput sequencing, automated phenotype measurement methods are rapidly progressing (Mon-tes et al., 2011; Furbank and Tester, 2011), though some types of automation remain specialized and expensive. The TCAP, funded by the United States Department of Agriculture's National Institute of Food and Agriculture (USDA-NIFA), aims to use these data-rich technologies to discover genes in wheat (Triticum aestivum L.) and barley (Hordeum vulgare L.) that will contribute to improved water and nitrogen use efficiency in future varieties. To this end, TCAP is capitalizing on current genotyping and sequencing technologies and applying new phenotyping strategies like canopy spectral reflectance (CSR) that indeed bring plant breeding into the realm of “Big Data” (Howe et al., 2008; Marx, 2013). To

Published in Plant Genome
© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
This is an open access article distributed under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Abbreviations: CSR, canopy spectral reflectance; GBS, genotype-by-sequencing; GRIN, Germplasm Resource Information Network; GWAS, genome-wide association studies; NSGC, National Small Grain Collections; SNP, single nucleotide polymorphism; T3, The Triticeae Toolbox; TCAP, Triticeae Coordinated Agricultural Project; THF, The Hordeum Toolbox.
handle this quantity of data within TCAP, increased data handling and improved tools for users and curators were clearly going to be needed.

While plant genomic databases like GrainGenes (Matthews et al., 2003; Carollo et al., 2005) and Maize-GDB (Lawrence et al., 2004; Sen et al., 2009) are rich in genetic information such as maps and sequences, they are primarily intended as knowledgebases that facilitate access to published results of genetic analyses but not to the original data and analyses themselves. Thus, T3 was adopted from The Hordeum Toolbox (THT; Blake et al., 2012) to integrate and provide access to phenotypic and genotypic data generated by TCAP.

The Triticeae Toolbox is actually several distinct databases (Table 1) holding data in separate production databases for barley, wheat, and oat, and results of U.S. Uniform Regional Nurseries for the U.S. Wheat and Barley Scab Initiative. Sandbox databases are available to data contributors to test the conformity of their data files before submitting them to the managing curator, who loads the data onto a production database. Sandboxes are mirrored at all times. A library of current upload template files is available, and individual template files are embedded assuring that T3 has the most up-to-date information available, and individual template files are embedded with further instructions to complete the data files. The project maintains Sandbox versions of the databases for participants to format and troubleshoot their data in preparation for submission to the managing curator.

### Table 1. Public databases using The Triticeae Toolbox (T3) schema.

<table>
<thead>
<tr>
<th>Database</th>
<th>URL</th>
<th>Description</th>
<th>Support†</th>
</tr>
</thead>
<tbody>
<tr>
<td>T3 Wheat</td>
<td>triticeaetoolbox.org/wheat</td>
<td>genotype/phenotype database and analysis tools for wheat (Triticum spp.)</td>
<td>TCAP</td>
</tr>
<tr>
<td>T3 Barley</td>
<td>triticeaetoolbox.org/barley</td>
<td>genotype/phenotype database and analysis tools for barley (Hordeum spp.)</td>
<td>TCAP</td>
</tr>
<tr>
<td>T3 Oat</td>
<td>triticeaetoolbox.org/oat</td>
<td>genotype/phenotype database and analysis tools for oat (Avena spp.)</td>
<td>NAMA, AAFC</td>
</tr>
<tr>
<td>Breeder’s Datafarm (wheat)</td>
<td>malt.pw.usda.gov/t3/bd/</td>
<td>genotype/phenotype database for US Regional Uniform Nurseries measuring traits related to Fusarium spp. infection in wheat</td>
<td>USWBSI</td>
</tr>
<tr>
<td>Breeder’s Datafarm (barley)</td>
<td>malt.pw.usda.gov/t3/bdb/</td>
<td>genotype/phenotype database for US Regional Uniform Nurseries measuring traits related to Fusarium spp. infection in barley</td>
<td>USWBSI</td>
</tr>
</tbody>
</table>

† AAFC, Agriculture and Agri-Food Canada; NAMA, North American Millers’ Association; TCAP, USDA-ARS Triticeae Coordinated Agriculture Project; USWBSI, U.S. Wheat and Barley Scab Initiative.

### Table 2. Database content in The Triticeae Toolbox (T3) Wheat and T3 Barley in May 2015.

<table>
<thead>
<tr>
<th>Database content†</th>
<th>T3 Wheat</th>
<th>T3 Barley</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phenotype trials</td>
<td>307</td>
<td>633</td>
</tr>
<tr>
<td>Genotype trials</td>
<td>50</td>
<td>111</td>
</tr>
<tr>
<td>TCAP data programs</td>
<td>19</td>
<td>27</td>
</tr>
<tr>
<td>Lines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line records</td>
<td>13,192</td>
<td>38,781</td>
</tr>
<tr>
<td>Breeding programs</td>
<td>49</td>
<td>24</td>
</tr>
<tr>
<td>Lines with genotyping data</td>
<td>9,351</td>
<td>15,967</td>
</tr>
<tr>
<td>Lines with phenotype data</td>
<td>8,655</td>
<td>9,413</td>
</tr>
<tr>
<td>Phenotype data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traits</td>
<td>144</td>
<td>111</td>
</tr>
<tr>
<td>Total phenotype data</td>
<td>407,781</td>
<td>423,282</td>
</tr>
<tr>
<td>Genotype data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-SNP markers</td>
<td>2,301</td>
<td>1,001</td>
</tr>
<tr>
<td>SNP markers</td>
<td>105,246</td>
<td>9,641</td>
</tr>
<tr>
<td>GBS markers</td>
<td>3547,662</td>
<td>55,782</td>
</tr>
<tr>
<td>Total genotype data</td>
<td>157,333,013</td>
<td>49,534,252</td>
</tr>
</tbody>
</table>

† GBS, genotype-by-sequencing; SNP, single nucleotide polymorphism; TCAP, Triticeae Coordinated Agricultural Project.

growing regions. T3 barley also holds data from Bolivia, Ethiopia, and Hungary. Traits in T3 are partitioned into categories that are customized to best reflect the types of data collected for each species. T3 Barley and T3 Wheat each have trait categories for agronomic, morphological, and disease-related traits. T3 Wheat reports quality traits in one category, and T3 Barley has the three additional categories: winter growth habit, malt quality, and food/feed quality.

A major strength of T3 is the ease with which data can be loaded. No programming or database skills are required to upload data, as T3 was developed to accept upload files that can easily be created by data contributors and loaded as Microsoft Excel or text files. These data are integrated immediately into the database, assuring that T3 has the most up-to-date information at all times. A library of current upload template files is available, and individual template files are embedded with further instructions to complete the data files. The project maintains Sandbox versions of the databases for participants to format and troubleshoot their data in preparation for submission to the managing curator.
The curator experience is interactive, and the T3 team has strived to deliver clear and explicit messages to aid troubleshooting upload files when errors occur.

**Hardware**

The server holding the production versions of T3 has 1.5 TB of disk space, 512 GB of RAM, and 32 CPU cores at 2.0 Ghz. The Sandbox versions, the development databases, and the Breeders Datafarms run on different machines (Table 1), assuring that most users can access the tools and resources on T3 without competing for processing time with the developers and curators. The T3 website and database use standard components: Linux, Apache HTTP Server (Apache Software Foundation), a MySQL (My Structured Query Language) relational database, and PHP (PHP hypertext preprocessing) programming language. Additional packages are used for analysis and visualization of data: R script language (Ihaka and Gentleman, 1996, R Development Core Team, 2008), X3DOM (Behr et al., 2009), and ViroBLAST (Deng et al., 2007).

**Database Initialization, Data Acquisition, and Maintenance**

The core design of T3 allows new projects to adapt easily, but several constraints require consensus by the project members for the resulting data to be useful. Members must agree on a controlled vocabulary, as well as scales and units for all phenotypic traits. The managing curator needs to upload a great deal of background information (i.e., institutions, project members, data types, and trait descriptions) before the flow of data can begin. Germplasm records must exist in T3 before data can be loaded. T3 requires the primary name to be one string, therefore some names will require underscores to connect words (ex. Chinese Spring becomes Chinese_Spring) while spaces can simply be removed between letters and numbers (ex. Sumai 3 becomes Sumai3).

Data is submitted to T3 by means of templates that are either comma- or tab-delimited text or Excel (Microsoft Corp., Redmond, WA) files, depending on the data type. The underlying principle in designing T3, preserved from that of THT, is to enable researchers to submit data using simple templates to organize data generated by their programs. A simple Web-based interface on T3 interacts with managing and user-curators concerning the format of their data (for example, using the most up-to-date template), or the integrity of the data (for example, verifying that measurements are within the range set for a particular trait). This enables one curator to manage data produced from a large number of sources. Curator tools built into T3 enable data managers to easily edit many of the fields of existing records, add new traits, and delete phenotype trials, enabling T3 to remain current.

Some tools in T3, such as 3D Cluster Lines by Genotype, and Genomic Association and Prediction tools, may require extensive calculations. Estimated execution time is calculated based on the number of lines and markers after filtering. When T3 estimates that the execution time will be over 2 min, the estimated time appears, and a link is provided to check if the job is finished. If the user is logged in to T3, then T3 will notify the user by email on completion. For shorter calculations, T3 shows a progress indicator until the calculation completes, then the page finishes loading.

**Genotype Data Storage and Retrieval**

To quickly retrieve large amounts of data from the database, the genotype information is stored into two-dimensional tables. This allows T3 to retrieve all the genotype information for specific lines or markers as one block of data. Records are also linked to experiment, program, and genotyping platforms. Because T3 serves to help a large group of PI’s collaborate, emphasis was placed on developing a simple interface to find datasets coming from different groups and then collate them for ease of downstream analysis. The menus designed for T3 reflect the workflow that we expect in such analyses, grouping together functions to select data, to analyze the resulting selections, or to download them for analysis on the client side. Our most innovative selection tool is the Wizard that allows the user to search and combine lines or phenotypes that share breeding program origins, were evaluated in the same locations, or for the same traits. Alternatively, starting from a previously selected set of lines, it is possible to identify trials where any or all lines were evaluated. These functions allow data from disparate trials to be quickly identified and brought together, ensuring that data management is not an impediment to collaboration. Complementing these search-and-combine functions, T3 enables users to define panels of germplasm lines for repeated use. For more complex analysis than allowed by internal T3 tools, data can be downloaded formatted specifically for tools like TASSEL (Bradbury et al., 2007), rrBLUP (Endelman, 2011), Flapjack (Milne et al., 2010), and synbreed (Wimmer et al., 2012).

**Genome-Wide Association Studies**

The most exciting new feature in T3 is the ability to perform genome-wide association studies (GWAS; Risch and Merikangas, 1996) in real time on selected data (Fig. 1). To access GWAS tools in T3, users click the Analyze menu bar, and the wizard will walk them through a series of steps beginning with assembling a dataset that has both phenotypes and markers for one trait. Upon returning to the GWAS interface through the Analyze tab, users will be prompted to select a genetic map and T3 will calculate the number of markers mapped for the line set in the queue per given map in T3, allowing for the best marker coverage in downstream applications.

With all necessary elements selected, T3 will now give users access to GWAS and genomic prediction tools (Hamblin et al., 2011). Both tools use the R package “rrBLUP” (Endelman, 2011). To execute a GWAS, the user specifies the number of principal component axes to be used as fixed effects and whether to use an
A. Genomic Association and Prediction

<table>
<thead>
<tr>
<th>Trait: breeders plump grain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
</tr>
<tr>
<td>Training</td>
</tr>
<tr>
<td>Validation</td>
</tr>
</tbody>
</table>

Warning: 79 lines removed from validation set because they are in training set.

Minimum MAF ≥ 5% Remove markers missing > 10% of data Remove lines missing > 10% of data

---

B. CAPWUE_2012_HighN_Bozeman breeders plump grain

C. Principal Components Analysis

---

Fig. 1. (continued on next page) Genomic Association and Prediction Suite in T3. (A–D) Prediction tools for plump grain using unique nurseries in Bozeman, MT, in 2008 and 2012. (E–G) GWAS tools using the plump grain trait in the CAPWUE_2012_HighN_Bozeman trial. (A) Interface showing selected training and validation phenotype trials, with information about overlapping germplasm lines and interactive text boxes to change Minimum minor allele frequency (MAF) and missing marker and line tolerance in the data. (B) Histogram comparing phenotype scores for “breeders plump grain” in the selected trials. (C) Population structure visualization plotting lines using the first two principal component eigenvectors. (D) Accuracy of the predicted values in the Validation Population. (E) Scree plot of the principal component eigenvalues. (F) Q-Q plot showing observed GWAS p-values against expected under the null hypothesis. (G) Manhattan plot of GWAS and top five markers scored in the analysis.
Fig. 1. Continued.
EMMA or EMMAX type of random effect control for population structure (Kang et al., 2010). The page then displays a histogram of the phenotypic observations, a scree-plot of the principal component eigenvectors, a Q-Q plot of observed versus expected–log$_{10}$ $p$-values, and a Manhattan plot (any unmapped markers in the dataset are binned in a chromosome 0). If multiple trials are selected, a fixed effect for trial is added to the model. If multiple traits are selected, only the first trait is analyzed.

In the example in Fig. 1, the trait “breeders plump grain” was used to determine if data generated in 2008 in a diverse North American collection of breeder’s lines could successfully predict the outcome of a similar set of lines grown in the same location in 2012. To begin with, the trait category “malt quality” was selected, then “breeders plump grain,” and the trial CAP3_2008_BozemanIrr, a 778 line irrigated nursery, was selected. The genetic map UC04162008 (Szuys et al., 2009) was selected to provide marker data, which had the best coverage for the selected lines with 3072 of 4608 markers mapped. Data was filtered using the default parameters (minimum minor allele frequency ≤ 5%; Remove markers missing > 10% of data; Remove lines missing > 10% of data) resulting in a training set with 2171 markers on 773 lines.

We show here a synthesis of results from both GWAS and genomic prediction tools. GWAS was implemented using three principal components as fixed effects to control for population structure. The mixed model used the EMMAX method. A separate validation set, a TCAP nursery CAPWUE_2012_HighN_Bozeman (256 lines with 74 removed because they were also in the training set), was chosen. Predictions based on model training from the 2008 data had a prediction accuracy of 0.41 on that validation set.

Rather than GWAS, the same inputs can be sent to a cross-validation analysis of genomic prediction using ridge regression, which is equivalent to a GBLUP analysis (Lorenz et al., 2011). In that case, five-fold cross-validation is performed on the data. The output is the histogram of phenotypes, a scatter-plot of the genotypes using the first two principal component eigenvectors, and a scatter-plot of the observed phenotype against its cross-validated prediction. Prediction accuracy is evaluated separately in each trial and observed versus prediction points in the scatterplot are given in a different color for each trial. Alternatively, the user may choose to select a second target dataset for prediction or cross-validation. In that case, the first dataset serves solely to train the prediction model. If the second dataset has the trait being modeled in the training set, T3 produces a scatter plot of observed against predicted phenotypes and calculates the correlation between the two in this validation set. Otherwise, T3 shows the same cross-validation plot described above. In all cases, a link is provided to a comma-separated values file with the lines, their cross-validated or direct prediction, and an indicator of whether the lines were in the training set or in the target set. When a second dataset is selected, the principal component scatterplot is also modified to show both training and target datasets.

While significant markers may commonly arise in a GWAS analysis when testing just one trial, when analyzing the genetic effect on a trait over many locations under varying conditions, more data is required. To test the hypothesis that more phenotype data in T3 for a given trait will increase the likelihood that a significant marker is discovered among several environments, GWAS was performed on data for the trait “breeders plump grain” that was available for all trials in THT in 2007 and 2008 (Waugh et al., 2009) and the data currently held in T3, which is nearly triple the germplasm tested. As shown in Fig. 2, data in combined locations in 2007 and 2008 was not sufficient to generate a significant marker, but by 2015 the increase in phenotype data in T3 enables the GWAS tools to identify two significant markers so T3 draws a horizontal LOD score line on the Manhattan plot to indicate significance.

**Genotype-By-Sequencing Big Data in T3**

Genotype-by-sequencing (GBS) is a low cost, high-throughput method to employ next-generation sequencing in genomes whose complexity has been reduced, usually by restriction enzymes (Elshire et al., 2011; Poland et al., 2012). The T3 project has strived to stay current with GBS data for wheat and barley, and a list of trials can be accessed via the “Genotype experiments” link on the left “Quick links” menu, or from the “Content status” page under the “About T3” menu. While there are synonymous markers discovered among the early wheat GBS trials (Table 3), until there are reference sequences available, the markers will remain distinct within the trials, and no synonyms will be created, thus demoting a marker to an alias rather than a primary marker name.

GBS markers in that are published without marker names are processed in the following manner:

1. The markers are filtered for duplicates and those whose reported single nucleotide polymorphism (SNP) position is beyond the length of the marker.
2. The A and B SNP alleles are alphabetized and placed in brackets within the marker sequence, keeping in mind that the published SNP position may begin with 0 rather than 1.
3. Markers are filtered again for duplicates that arise with the alphabetization.
4. Markers are then ordered alphabetically and named with a prefix to identify the experiment (e.g., gbsHWW, gbsCNL) and numbered sequentially.

**Selection Index Generator**

A selection index, first developed for animal breeders (Hazel, 1943), ranks breeding lines for multiple traits simultaneously, then computes a single number representing a combined score for all the traits. Plant breeders will find the selection index generator in T3 intuitive and flexible for helping to select germplasm to advance in programs and evaluate nurseries for traits of...
Though conceptually intuitive, complications arise because the numerical magnitude and the variance of some traits are greater than others, and because the desirable values of a trait may not always align to more-is-better rule (e.g., malt protein, heading date).

In the design of the T3 Selection Index generator, the user can choose to rescale the trait values either by standardizing to a distribution with a mean of zero and standard deviation of 1.0, or as percentage of the value for a reference line. A trait’s scale can be reversed and traits can be weighted unequally. Figure 3 provides an example where barley germplasm that was planted for 2 yr in Aberdeen, ID (irrigated with normal N), are indexed for the agronomic trait “heading date (Julian)” and the “Malting quality” traits “grain protein” and “plump grain” are selected (Fig. 3A). Since earliness to flower and low protein are desirable, the reverse ranging is selected (Fig. 3B). In this example, most of the top performing lines were from the North Dakota program, with one Montana line and the ND released variety Pinnacle (PI 643354; Fig. 3C).

### 3D Cluster by Genotype

T3 now provides three functions to cluster lines by genotype, a 2D method adopted from THT (Blake et al., 2012), and two new methods to generate a 3D representation of diversity by clustering. All functions require a selection of lines that have marker data. Marker data in columns are centered and missing marker scores imputed to zero (this is equivalent to mean imputation, Rutkoski et al., 2013). The first function uses the “partitioning around medoids” clustering algorithm (Maechler et al., 2013) applied to a Manhattan distance matrix calculated from the marker scores in R (R Development Core Team, 2008). The user specifies the number of clusters, $k$, the algorithm chooses $k$ breeding lines to be medoids, and then identifies which other lines to associate with each medoid such that the sum of distances between each line and its associated medoid is minimized. The second function is experimental and uses random projections of the marker data to create a similarity matrix. That matrix is then entered into the R “hclust” function for clustering. The 3D functions represent the clustering output by plotting a colored
point for each germplasm line using three-dimensional coordinates obtained from the first three eigenvectors of the singular value decomposition of the marker score matrix. The graphical display allows the user to rotate the three-dimensional array of colored points freely with the mouse. The graphical display software is the JavaScript package X3DOM (Behr et al., 2009).

Canopy Spectral Reflectance Plot-Level Data Management

Canopy Spectral Reflectance is a remote sensing technique to evaluate water and nitrogen use in crop plants (Tucker, 1979; Xue et al., 2004). The CSR data is an example of raw plot level phenotype measurements and are stored as discreet files. This is in contrast to the means or adjusted scores submitted with all other types of phenotype data that are integrated into the database on upload. A typical use of the CSR data is to calculate an index based on the measurements at specific frequencies. T3 provides tools for calculating 10 commonly used indices. The wavelength parameters and formula can be modified to create new indices. Built into the analysis tool is a plot of each CSR measurement that can be used to identify questionable data or for adjusting the wavelength parameters of the analysis. After calculating the CSR index, the results can be saved as a plot level trait. If the user is logged in, the line means can be calculated and saved in the database. These data can then be used as trial data for T3’s GWAS tools. The TCAP project has contributed data using the Jaz, USB2000 (Ocean Optics, Dunedin, FL), and CropScan (Cropscan Inc. Rochester MN) instruments. The metadata for the recording instrument and field layout is stored in the database.

Integration with the Android Field Book

The Android Field Book is “an open-source application for electronic data capture that runs on consumer-grade Android tablets” (Rife and Poland, 2014). If the user is working with a field trial for which T3 has a spatially explicit field layout (i.e., the row and column coordinates of plots are given), T3 can create and download a comma separated values (.csv) file compatible with the Field Book. Field data can be collected directly onto an Android tablet with the Field Book application, then uploaded directly to T3 without further formatting. T3 can obtain spatially explicit field layouts in two ways. Users can upload field layouts. Alternatively, T3 has an experimental design interface page that can design experiments and produce field plans. Designing experiments on T3 directly decreases risks of error from, for example, mismatches between files uploaded for the field layout and for the phenotypic measurements and eliminates the need to upload the design separately.
Discussion and Future Direction for the Toolbox

The completion the TCAP in 2016 (projected) will not spell the end of T3. It will persist. Other consortia have adapted the T3 database schema, including a North American Wheat and Barley Scab Initiative (Table 1).

The existing T3s for wheat and barley are in negotiation to merge with GrainGenes (http://wheat.pw.usda.gov; accessed 4 Dec. 2012; verified 4 Feb. 2016). GrainGenes is the knowledgebase for the Triticeae and Avena, and is currently overhauling the database and website with new funding to bring that resource current (personal communication, Ann Blechl, 2014). GrainGenes primarily contains genetic and mapping information from peer-reviewed studies, whereas T3 is a “diversity base” containing the raw information used in such studies, and is much better designed for storing and delivering phenotype data.

Analytical Tools in Development

**Training Population Design**

This upgrade will allow the user to specify a set of \( N_t \) lines that are candidates to be in a training population of size \( N_t \) \((N_t > N_c)\) and a set of lines that are the target of prediction. The analysis will then select the \( N_t \) such that the prediction accuracy on the target set is expected to be maximized.

**Definition of Core Germplasm Sets**

Similar to the above, the user will define a set of \( N_c \) lines that are candidates to be in a core set of size \( N_c \) \((N_c > N_t)\) thus determining if the variation is present in the core set, and if this set possesses the ability to predict the performance of other lines.

**Multivariate Analysis**

Initially, this will be limited to multivariate prediction methods, rather than GWAS. The user specifies a set of \( N_t \) training lines, a set of traits, and a set of lines that are the target of prediction. The analysis predicts values for all traits for the targets.

**Improved Handling of GBS Markers and Big Data**

T3 currently uploads and can retrieve and process GBS data. Nevertheless, with the size of the datasets, manipulations are slow and need to be improved.

**Experimental Designs with Field Maps**

T3 currently allows researchers to design trials with a modified augmented design that is spatially explicit. Such designs can be uploaded to the Android Field Book to facilitate data collection and can also be analyzed with spatial statistical models. We will expand the list of experimental designs to randomized complete block and augmented designs.

As reference genomes become available, T3 will incorporate these to anchor the current marker data for genotype by sequencing and high throughput SNP arrays, thus providing more confidence for GWAS and genomic prediction studies. A genomic browser for T3 is in development and should be available in late 2015.

Acknowledgments

This research was supported by the Agriculture and Food Research Initiative Competitive Grant No. 2011-68002-30029 from the USDA National Institute of Food and Agriculture. Maintenance and further development of T3 by GrainGenes is supported by USDA-ARS project 5325-21000-014-00, “An Integrated Database and Bioinformatics Resource for Small Grains.” 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


