# Germinate 3: Development of a Common Platform to Support the Distribution of Experimental Data on Crop Wild Relatives

Paul D. Shaw,★ Sebastian Raubach, Sarah J. Hearne, Kate Dreher, Glenn Bryan, Gaynor McKenzie, Iain Milne, Gordon Stephen, and David F. Marshall

## ABSTRACT

Conservation and exploitation of crop wild relative species is an important component in ensuring food security and improving current agricultural output. By identifying agriculturally important characteristics that express favorable response to both biotic and abiotic stress currently unused by breeders, the incorporation of this new genetic material into genetic background stocks may help mitigate problems imposed by climate change, land degradation, and population pressure. This is particularly important in countries that will be more severely affected by the threat of reduced yields. The ability to effectively manage genetic resources collections and integrate unique and diverse data types is crucial in exploring, understanding, and exploiting the diversity contained within genebanks. Providing a common interface through which experimental and background data can be disseminated to both researchers and breeders will bring focus and facilitate community building into research communities. We have taken wild barley (*Hordeum* spp.) and potato (*Solanum* spp.) collections along with wheat (*Triticum* spp.) and maize (*Zea mays* subsp. *mays*) and their wild relatives and incorporated this data into web-based information resources built using the Germinate platform (https://ics.hutton.ac.uk/get-germinate, accessed 4 Apr. 2017). We have tailored these to better meet the demands of researchers by developing both new data visualization tools and integration with current software such as Helium, Flapjack, and CurlyWhirly (https://ics.hutton.ac.uk/software, accessed 4 Apr. 2017) and presented the data in a common platform. While the underlying species differ, the approach taken ensures that tools are compatible across all database instances. We will describe these database instances and show that Germinate offers a common platform that will aid in the exploration and wider use of these species.

P.D. Shaw, S. Raubach, I. Milne, G. Stephen, and D.F. Marshall, Information and Computational Sciences, The James Hutton Institute, Errol Road, Invergowrie, Dundee, DD2 5DA, Scotland; G. Bryan and G. McKenzie, Cell and Molecular Sciences, James Hutton Institute, Errol Road, Invergowrie, Dundee, DD2 5DA, Scotland; K. Dreher and S. Hearne, International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Edo. De Mexico, Mexico CP 56237. Received 28 Sept. 2016. Accepted 13 Mar. 2017. ★Corresponding author (paul.shaw@hutton.ac.uk). Assigned to Associate Editor Hem Bhandari.

**Abbreviations:** CIMMYT, Centro Internacional de Mejoramiento de Maíz y Trigo; CPC, Commonwealth Potato Collection; CWR, crop wild relative; DOI, data object identifier; GWAS, genome-wide association scan; GWT, GWT Web Toolkit; ITPGRFA, International Treaty on Plant Genetic Resources for Food and Agriculture; MCPD, Multi-Crop Passport Descriptors; PCA, principal component analysis; PCO, principal coordinate analysis; SeeD, Seeds of Discovery; SNP, single nucleotide polymorphism.

PLANT GENETIC RESOURCES of modern crop varieties, landraces, and wild relatives represent the gene pool on which our ability to sustainably respond to pressures from land degradation, population pressure (Godfray et al., 2010), and climate change (Nelson et al., 2009) depends (McCouch et al., 2013). Our ability to efficiently utilize such resources has been significantly enhanced by the modern development of high-throughput technologies for sequencing, genotyping (Onda and Mochida, 2016), and more recently phenotyping (Fahlgren et. al., 2015) plant germplasm. Depending on genome size and breeding system, technologies such as genotyping arrays, reduced representation sequencing, or even whole genome sequencing may be routinely deployed to characterize many thousands of plant lines in germplasm collections or breeding populations. Increasingly, with the dramatic

changes in the relative costs of sequencing, phenotyping on a comparable scale has looked to be the rate limiting component of germplasm characterization. However, with more recent developments in high-throughput phenotyping (based on imaging across a range of spectra and improved image analysis technologies), the numerical scale at which germplasm can be phenotypically characterized is increasing rapidly. These high-throughput technologies will revolutionize germplasm exploitation and utilization. In effect they have the potential to provide what is, in essence, a molecular and phenotypic index of collections. This will enable genebank managers to more efficiently manage their collections by structuring them to account for duplicates, or helping them to identify gaps that can serve to target additional collection priorities. From a genebank user perspective it will considerably improve the development of criteria that can more readily target the selection of requests for lines and accessions more likely to exhibit the sought-after properties. This has the joint benefit of reducing the scale of prebreeding task faced by the plant breeder and by targeting demand more precisely, thus reducing the administration and maintenance burden faced by genebank managers.

Access to, and efficient utilization of, the resulting data from these genotyping and phenotyping technologies will require the development of suitable data management systems able to integrate the metadata that describes the origin and the biological nature of the plant germplasm together with the associated characterization information. The realization of the impact and challenges that will come from these developments has recently led to the formation of the DivSeek (www.divseek.org, accessed 4 Apr. 2017) initiative, which brings together many key organizations and funding bodies to openly share approaches and solutions for mutual benefit.

In the international plant genetic resources community, considerable progress has been made in defining the metadata standards or "passport" information that are necessary to understand the biological nature of a plant line or accession, its historical maintenance in collections, and its origin either from other collections or from field sampling. The Multi-Crop Passport Descriptors (MCPD) (Alercia et. al.,2015) provide a well-established international standard for genebank accession metadata and can also provide a useful comparable standard for breeding lines. In the context of the international legal framework for access to and exchange of plant germplasm encapsulated in the Convention for Biological Diversity (https://www.cbd.int/convention/, accessed 4 Apr. 2017), the associated Nagoya Protocol on Access and Benefit Sharing (https://www.cbd.int/abs/), and the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA, http://www.fao.org/plant-treaty, accessed 4 Apr. 2017), a key element of germplasm access will be

the assignment and utilization of appropriate persistent unique identifiers (PUIds) for germplasm and for data. The ITPGRFA already supports the utilization of, and provides data object identifiers (DOIs) within, its domain of responsibility (http://www.fao.org/plant-treaty/areas-of-work/global-information-system/doi/, accessed 4 Apr. 2017), as does the Global Biodiversity Information Facility (http://www.gbif.org/, accessed 4 Apr. 2017). It is likely that such identifiers will become a requirement for germplasm exchange or for reference to an accession or associated data set for publication.

There are three major data management components required for systematic management of genetic resource collections and their associated data. The first of these is a system for stock management that enables genebank managers to control information on the collections they are responsible for, including MCPD data, regeneration and replenishment of seed stocks, and germplasm requests. Though many genebanks have developed their own bespoke systems, a more generic system, GRIN-Global (Postman et al., 2010), is being jointly developed by the USDA Agricultural Research Service (ARS), Bioversity International, and the Global Crop Diversity Trust. GRIN-Global is now being deployed in many international crop genebanks. The second major component is a system that can integrate information about accessible accessions across collections to enable plant breeders and scientists to explore the full range of internationally available germplasm for their target species. Genesys (https://www.genesys-pgr.org, accessed 4 Apr. 2017), developed with the support of the Global Crop Diversity Trust, fulfills this role by integrating passport data and, where available, evaluating information from an extensive range of national and international plant germplasm collections.

The final component required for efficient access to germplasm information are systems capable of integrating both genotypic and phenotypic data together with associated passport data and providing suitable query and browsing tools to support the facile exploration of the increasingly large and complex germplasm characterization data sets that high-throughput technologies are now generating. We have developed Germinate 3 to meet this demand.

On the basis of our experience in working with data sets from both plant germplasm collections and breeding lines, we have developed Germinate 3, which is a data warehouse tailored to the domain of plant germplasm collections together with associated characterization data. Germinate 3 provides the data infrastructure to store validated molecular and phenotypic characterization data together with either associated accession passport data or comparable breeding line information. In addition, it provides a versatile query visualization interface to identify groups of lines or accessions that meet selected passport, molecular, or phenotypic criteria and to download the
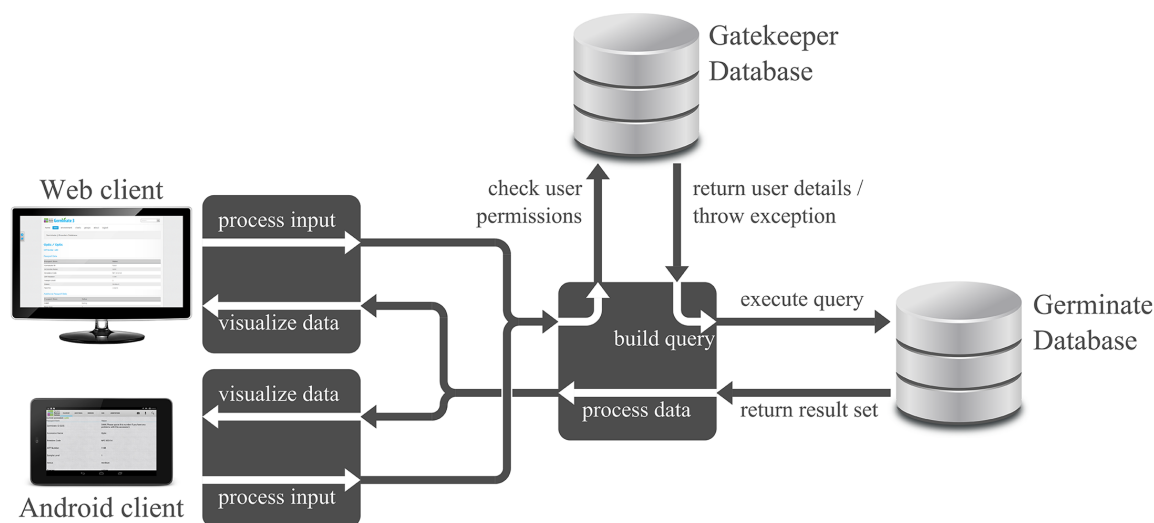
Fig. 1. Germinate architecture: Clients request data via a web interface or Android client that is processed by Germinate in two ways: (1) the data is freely made available and no authentication takes places and (2) the user's credentials are checked against the Germinate Gatekeeper database to provide authentication. Germinate can run in either of these modes. Germinate Gatekeeper means that a single user can be given varying permissions across a number of Germinate installations. Gatekeeper provides user tools to allow the automated recovery and setting of passwords as well as informing administrators of new account requests.

associated data in a variety of formats. It also integrates with the James Hutton Institute's powerful range of germplasm data visualization tools: Helium (Shaw et al., 2014), Flapjack (Milne et al., 2010), and CurlyWhirly (https://ics.hutton.ac.uk/software, accessed 4 Apr. 2017).

## MATERIALS AND METHODS
### The Germinate 3 Platform

Germinate is an open source plant database infrastructure and application programming platform on which complex data from genetic resources collections can be stored, queried, and visualized using common, reusable, programming components. Germinate utilizes modern web and database standards to provide a standard architecture and high performance web-based user interface and analytics functionality across a wide variety of data types. In addition it provides native coupling with a variety of external analysis tools. Germinate is fully compliant with the Multi-Crop Passport Descriptors (MCPD V.2.1) developed by the Food and Agriculture Organization (FAO) (Alercia et. al., 2015). The single point of access for all background data on genetic resources collections that Germinate implements provides a valuable digital curation resource as well as offering essential continuity on how new data is recorded and archived.

Germinate is a generic platform for the storage and dissemination of multiple data types associated with genetic resource collections. Examples of these include passport, phenotypic, field trial, pedigree, genetic, climatic, and geographic location data as well as user-submitted annotations. We have developed Germinate to be as flexible as possible and can add support for additional data types as users' needs dictate. Germinate not only acts as storage for experimental data but offers a user-friendly interface into the data and acts as a backend for visual analytics tools such as Helium for pedigree visualization, Flapjack for graphical genotyping, and CurlyWhirly for the simple display of Cartesian coordinate data such as principal coordinate analysis (PCO) and principal component analysis (PCA). We have prioritized the development of tools to allow the export of data in formats suitable for analysis in statistical analysis tools such as R (https://www.r-project.org/, accessed 4 Apr. 2017).

Germinate 3 builds on the existing Germinate platform, improving its performance while also adding new functionality and tools. A simpler installation process and architectural changes now allow deployment of the system within both server and desktop environments; previously only a Linux based operating system could be used. The prerequisites for installation are Java (https://java.com, accessed 4 Apr. 2017), Apache Tomcat (http://tomcat.apache.org/, accessed 4 Apr. 2017), and MySQL (http://www.mysql.com/, accessed 4 Apr. 2017), allowing Germinate to be installed natively on Linux, Windows, and macOS systems. Germinate moves away from Perl (https://www.perl.org, accessed 4 Apr. 2017) as its primary development language and has been completely rewritten using the Java-based GWT Web Toolkit (GWT) (http://www.gwtproject.org, accessed 4 Apr. 2017), which allows us to build feature-rich and interactive websites that not only support all modern web browsers but also enables Germinate to be easily customized for different regions and deployed in multiple languages. This is especially important where collections are collected and curated by diverse international groups. By default we offer Germinate translations in English and German as well as localizations offered by the GWT framework. To develop new, and maintain our currently supported, language translations, we utilize the Crowdin translation service (https://www.crowdin.com/, accessed 4 Apr. 2017). Examples of new tools that are offered include: an interactive web-based user interface built using current web standards; advanced data query and export tools; full internationalization (via resource qualifiers), localization, and user access control using the Germinate Gatekeeper database and user interface application (Fig. 1).

The Germinate application is distributed as a common web application archive (WAR) file that includes all required

resources for the Germinate web interface installation and can be easily deployed on Apache Tomcat. The database schema is distributed as a structured query language (SQL) file that can be loaded using MySQL.

Germinate has been developed using MySQL (current requirements are MySQL v5.7) and is, with some exceptions, compatible with MariaDB (https://mariadb.org/, accessed 4 Apr. 2017), which can be used as a drop-in alternative to MySQL. Germinate should also be compatible with a number of other relational database management systems with minor changes.

Germinate has been developed as an extensible platform, with an aim of offering an off-the-shelf solution for the storage of genetic resources data. Where the necessary skills exist, increased functionality may be added to the platform as required. Development of Germinate components are made immediately available to the community, and we have also developed tools that aid in the automatic deployment of new Germinate software and database updates as they become available as well as aid in the upload of data to the system.

## Database Schema

The principal goal of Germinate was to adopt a pragmatic approach to developing a data warehouse system able to handle the large and varied data types that are common in genetic resources collections. We have simplified the original Germinate schema (Lee et al., 2005), modifying and dividing into specific subschema, meaning new subschema can more easily be developed to add additional functionality to the platform. In its entirety, Germinate consists of 62 tables, although a proportion of these are fixed in nature and hold static or infrequently changing, information such as default measurement units and standard controlled vocabulary from MCPD terms. The main subschema include Germinate Core, which holds germplasm lists, MCPD information, associated identification data, and holding institutes; marker and genotype schema for background information on genetic markers and associated genotypic data; phenotype and field trials schema for information and scores relating to phenotypic datasets and field trial experiments; pedigree schema for pedigree data if available; locations schema used to store georeferenced collecting site details; groups schema, which is used to store user-defined groups; datasets and experiments schema, which is used to assign phenotypic and genotypic data to specific datasets and experiments; climate data schema, which is used to store climatic data relating to collecting sites held within Germinate.

For all intents and purposes, these MCPD terms should be classified as the *lowest common denominator* when it comes to passport data held in any germplasm database. Germinate is fully MCPD v2.1 compliant but also includes the ability to hold any other data types used in the categorization of germplasm collections.

## Additional Data Types

Not every type of data can be accounted for at the outset of any work. Because of this, it is important that Germinate offer flexibility to store additional data types. We have developed an attribute storage model, fully searchable via our search interface, that allows user-defined additional data to be associated with a germplasm entry. Attributes, including additional identifiers that are used to identify the accession, barcode identifiers, or locations where the accession is stored in a collection, can all be stored in this way. Multiple attributes, which can be changed over time, can be added to Germinate as projects develop. This provides flexibility to information storage.

While the attribute storage model is incredibly flexible, it must be emphasized that it should only be used for attributes that are low in volume or are added to a collection infrequently. For more formal storage of additional data types, it would be better to include them in the core Germinate germplasm table as new fields. If something is used frequently, it is better to include a formal definition rather than rely on loosely coupled data tables such as attributes. This is one of the areas in which Germinate has evolved from its original implementation where Lee stated preference to attribute type access over formal definition (Lee et. al., 2005).

## User Interface

In addition to the underlying database, the Germinate Platform also includes an out-of-the-box, web-based user interface designed to current web standards (HTTP, CSS [cascading style sheets], SVG [scalable vector graphics], JavaScript) and uses CSS-responsive layouts to ensure that our user interface works across a wide spectrum of desktop, tablet, and mobile devices. The interface also offers a standard set of tools for accessing and exporting data held in a Germinate database. The interface is broken down into a number of specific areas, described in Fig. 2.

### Germinate Group Concept

One of Germinate's key features is its ability to define what we have called *groups*. A group is defined as a user-defined list of accessions, collecting sites, or markers, and these groups can overlap and intersect in a number of ways, giving great flexibility to the system. Germinate's groups-based approach and implementation means that any number of groups can be defined against a given experimental dataset. A typical use case would be that a user would want to generate a group of lines that exhibit a particular characteristic. The user can search for these within the Germinate interface on the basis of a character or collection site attributes and then add them to the Germinate "shopping cart." The user can then create a custom user-defined group based on these accessions that can be used to export phenotypic or genotypic data from the database (Fig. 3). The groups are *permanent* and are held in Germinate until such time as the original user deletes the group. If a Germinate user is creating the group, there is the option when the group is created to make the group public; this means that *all Germinate users* will have access to the group and not just the user who created it.

### Geographic Searches

Germinate uses a number of geographic-based visualization tools that use the Google Maps API (application programming interface). Although we are also working on integration of Germinate with ArcGIS from ESRI (http://www.esri.com/, accessed 4 Apr. 2017), the use of Google components allows us to distribute Germinate without relying on commercial libraries for mapping visualizations. The use of ArcGIS may, however, be included as an optional component in future Germinate releases.

There are a number of tools that we provide to allow users to search on the basis of geographic location. First, Germinate allows the selection of germplasm and ordering of collecting sites on the

Fig. 2. The Germinate user interface: An example of the standard Germinate web interface that is implemented here for our Germinate Wild Barley database (https://ics.hutton.ac.uk/germinate-wild-barley). (a) Menu options are shown here on every page. The menus are dynamically generated on the basis of the content in your particular Germinate instance and can be customized but are standard across your site. (b) Search facilities allow searching across all data held within Germinate. When the search feature is used, the returned data is broken down into categories, then displayed to the user to make finding specific information easy. (c) Where translations are available, these are switchable by selecting options. By default Germinate will switch to a language based on the locale of your machine, but it can be manually changed here. (d) The main information pane contains the data that is returned from Germinate on the basis of your menu options or queries. Check boxes can be seen that allow you to add data to the "shopping cart" feature of Germinate to create custom germplasm, marker, and location groupings. Germinate can be customized for each individual instance using built-in admin tools that allow the branding of each site to suit the needs of the project. (e) Export options are always displayed just below the main data pane. If different data export formats are supported, these options will appear here.
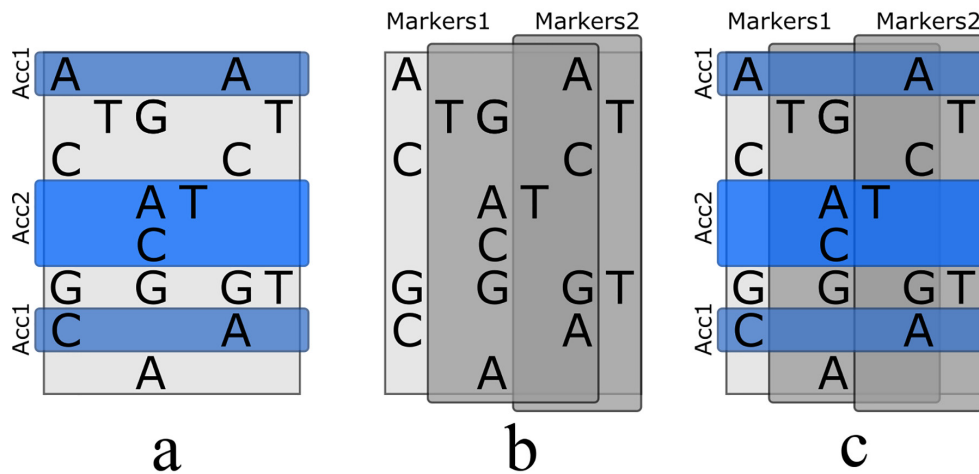
Fig. 3. Germinate groups feature: Germinate allows (a) accessions, (b) markers, and germplasm collecting sites (not shown here) to be arranged into groups. Groups allow the export of data for predefined or user-defined collections. Groups can be combined. In this example we have four groups: two accession groups and two marker groups. The example dataset shown here is for an allele call matrix, germplasm accessions on the horizontal axis (rows) and markers on the vertical (columns). (a) Accession groups Acc1 and Acc2 allow the selection of different set of germplasm. These groups can be combined by selecting the required groups when exporting data. (b) Marker groups Markers1 and Markers2 provide the same functionality for germplasm. Marker and accession groups can be combined (c) to allow the export of specific data intersections from larger datasets.

basis of distance from a geographic point. Second, Germinate allows users to select geographic regions, by means of drawing a polygon that bounds an area then selecting all accessions that are assigned to collecting sites within this region (Fig. 4).

We also provide a number of geographic visualizations to show collecting sites and site densities across the globe using Google Maps (Fig. 5) and export options to allow export into Google Earth (https://www.google.co.uk/earth/, accessed 4 Apr. 2017) format.

### Overview Statistics for Germinate Installations

Germinate has been implemented across a diverse range of species including both wild, landrace, and cultivated germplasm that includes wheat, maize, barley, potato, ryegrass (Lolium spp.) (Kopecky et al., 2009), blackcurrant (R*ibes* spp.), raspberry (*Rubus* spp.), pea (*Pisum* spp.) (Jing et al., 2010), sunflower (*Helianthus annulus* L.), and rice (*Oryza sativa* L.). Each of these databases differ in size from small project databases such as Germinate Grasses (which holds *Lolium* and *Festuca* accessions) to large scale genebank collections of significant global importance, including maize and wheat collections from the International Maize and Wheat Improvement Center (CIMMYT) in Mexico. The data volumes also vary between each of these instances, with the larger Germinate instances holding tens of thousands of accessions and a million markers. Detailed overview statistics for each of the current Germinate instances can be seen in Table 1.

Accessions held within Germinate can also be grouped on the basis of collection site. This is shown in Fig. 6.

Germinate has been implemented for a number of crop species. While most of the examples presented here are freely and openly accessible, Germinate Maize and Germinate Wheat, which are part of the Seeds of Discovery initiative, operate under a proactive intellectual property framework that requires users to register and agree to license conditions to access the information they contain.

### Examples of Implementations in Crop Wild Relative Species—Potato and Barley

The Commonwealth Potato Collection (CPC) is the United Kingdom's genebank of landrace and wild potatoes (Hawkes, 1951). The collection is one of a network of international potato genebanks and is held in trust for humanity by the James Hutton Institute in accordance with the International Treaty for Plant Genetic Resources for Food and Agriculture.

The CPC is comprised of 1518 accessions representing 80 different species, of which 821 accessions are wild and 697 are of primitive cultivated potato species. Each accession can be traced back, through collection records, to berries or tubers collected from 1938 to the late 1950s. These collections were performed in North, Central, and South America and are composed of wild accessions or accessions obtained from growers at local markets (Fig. 7).

Collections like the CPC are incredibly important, containing the basic resource for the improvement and adaptation to the world's fourth most important food crop (Bradshaw and Ramsay, 2005). The species contained within the collection come from a diverse range of environments, from the damp cloud forests of Bolivia to the arid scrubland of Mexico and, as such, are a valuable resource for novel traits not contained in currently cultivated material. The genetic material contained in this collection has made a major contribution to the potato crop worldwide and continues to be exploited for novel genetic content.

The International Treaty on Plant Genetic Resources for Food and Agriculture and the Convention on Biological Diversity place requirements on nations for the protection and sustainable use of the biodiversity in their care and for appropriate intellectual property control of such material according to the requirements of the International Treaty. The Germinate database system facilitates the management and utilization of the collection by allowing interested parties to interrogate the collection for suitable accessions and request access to these resources in accordance with the multilateral system.

Fig. 4. Geographic searching with Germinate: Germinate offers a number of tools to allow searching of collecting sites that are within defined geographic bounds. (a) The selection of collecting sites by means of drawing a polygon around a region of interest can be easily performed. The numbers are clusters of collecting sites within that area. (b) The number of collecting sites found and, subsequently, the accessions that are assigned to those collecting sites are also shown. (c) Additional search options are, where available, also presented. In this example we are selecting wild barley accessions from the Fertile Crescent region.

The Germinate Wild Barley database holds a set of wild and landrace germplasm that was assembled by three barley research groups: the James Hutton Institute, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), and the University of Minnesota. The collection represents accessions with a wide geographical distribution. The main aim of the study was to explore how barley adapts to the different environments across the wide geographical range of its distribution. All the accessions, which are georeferenced and available from the IPK genebank (http://www.ipk-gatersleben.de/en/genebank, accessed 4 Apr. 2017) have been exome captured and sequenced, resulting in 1.6 million single nucleotide polymorphisms (SNPs)
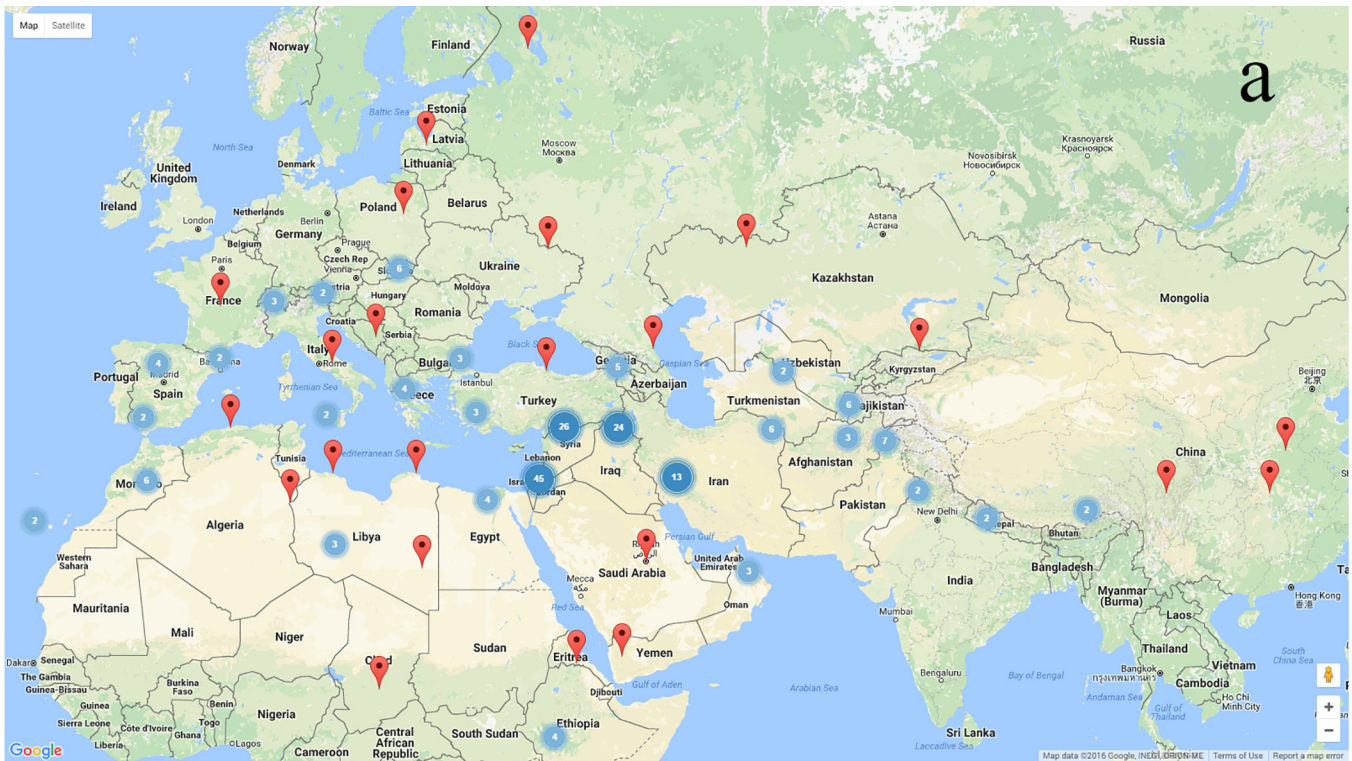
Fig. 5. (continued on next page) Collecting site densities: Wild barley accessions from the Germinate Wild Barley database. Germinate offers multiple ways to show collecting site densities. (a) Collecting sites displayed using marker clustering, (b) collecting sites displayed using a heat map representation, and (c) using individual markers for each site. Each has its advantages depending on what the user is trying to achieve. The Fertile Crescent is particularly clear when looking at the heat-mapped representation (b).

and over 140,000 indels (Mascher et al., 2013; Russell et al., 2016). Using this data, we observed geographically structured patterns of diversity and a high proportion of rare alleles for future exploitation and identified associations between environmental variables and SNPs.

## Examples of Implementations in Crop Wild Relative Species—Maize and Wheat

The germplasm bank at CIMMYT houses the world's largest collections of maize and wheat, with over 175,000 accessions
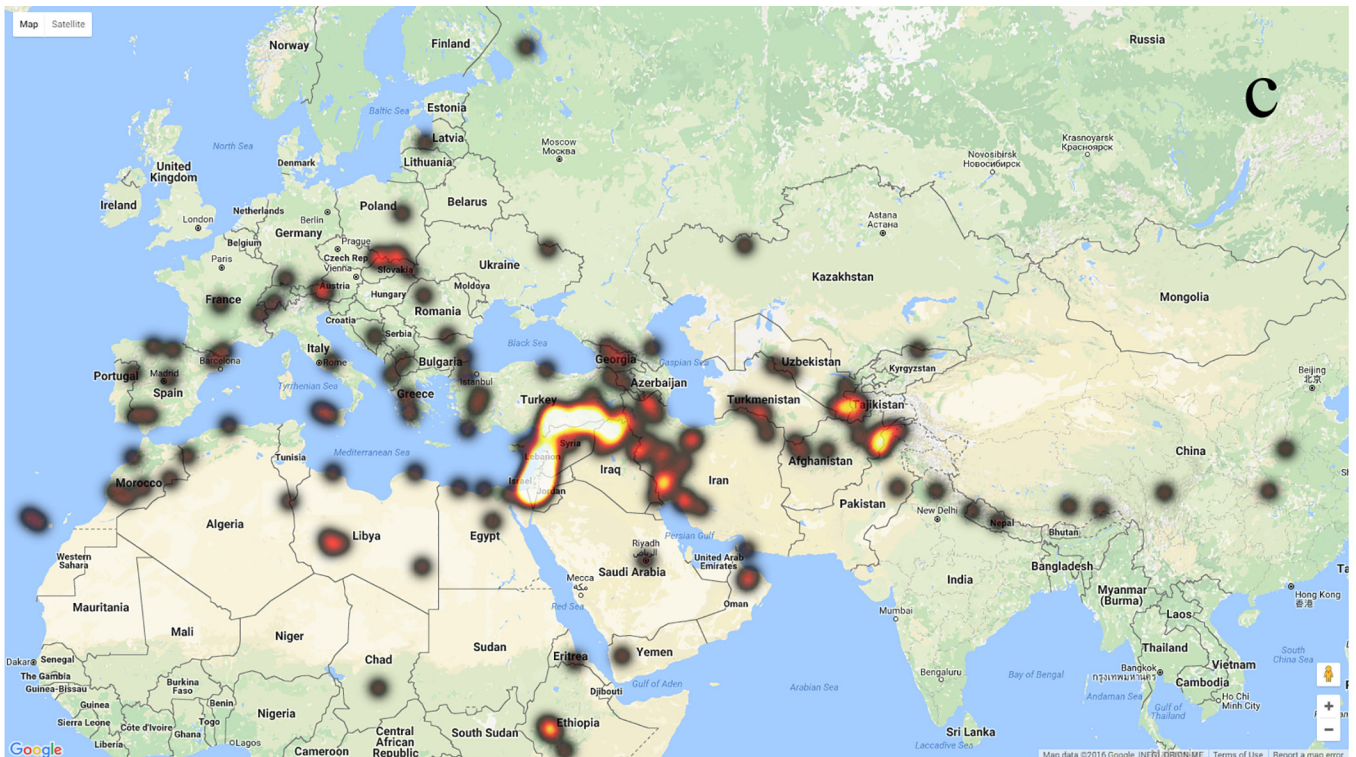
Fig. 5. Continued.

Table 1. Germinate installations overview statistics showing volumes of data being held. These results show the diversity of scale of a number of Germinate installations that cover crop wild relative species.

| Germinate instance | Accessions or germplasm instances | Markers | Collecting sites | Genotypes | Phenotypic values |
|---|---|---|---|---|---|
| Germinate Maize | 33,114 | 1,004,972 | 10,238 | 477,582 | 647,669 |
| Germinate Wheat | 85,519 | 150,456 | 1979 | 250,036,560 | 804,519 |
| Germinate Wild Barley | 267 | 557,344 | 221 | 127,074,432 | 10,261 |
| Germinate Potato | 1499 | 0 | 457 | 0 | 3299 |
| Germinate Demo | 2000 | 5000 | 500 | 8,472,803 | 59,400 |
| Germinate Grasses | 167 | 3884 | 28 | 644,744 | 0 |
| Germinate Pea | 3693 | 1469 | 213 | 244,496 | 0 |

stored in Mexico. Plus, many of the materials are backed up in the Svalbard Global Seed Vault in Norway. These seeds, and all information about them, are international public goods. Every year, researchers worldwide receive several tons of these materials for use in breeding and discovery.

The Seeds of Discovery (SeeD) initiative is a multipartner, collaborative effort led by CIMMYT that aims to unlock and utilize the novel genetic diversity held within wheat and maize germplasm banks to accelerate the development of maize and wheat varieties to meet the demands of a growing population and climate change. Seeds of Discovery has undertaken to evaluate the International Maize and Wheat Germplasm Bank collection managed by CIMMYT, along with materials managed by partner institutions' germplasm banks. This evaluation comprises the genotypic characterization of the collection, coupled with the targeted phenotypic evaluation of selected accessions, for a variety of priority traits. Management of this data, and other data pertinent to the accessions such as passport data and geographic information system data derived from analysis of collection locations, is a key challenge. Facilitation

of interlinked query and visualization of the various data types is key to enabling knowledge discovery not only by initiative partners but also in the broader wheat, maize, and larger scientific communities. CIMMYT decided to adopt Germinate as a flexible web-based resource to facilitate this sharing and value addition of SeeD data. Germinate has been deployed in two instances (http://germinate.seedsofdiscovery.org, accessed 4 Apr. 2017) to create a central component of what are termed the wheat and maize "molecular atlases." In this effort, the crop wild relatives (CWR) housed in CIMMYT's Maize and Wheat Germplasm Bank are included, along with other accessions comprising landraces and specific elite materials. GRIN-Global (http://www.ars-grin.gov/npgs/gringlobal/sb/home.html, accessed 4 Apr. 2017), available from the USDA, the Crop Trust, and Bioversity International, is used by the Germplasm Bank to manage the inventory of these accessions and to provide some basic information about the taxonomy, collection site, and MCPD for these genetic materials (Maize: http://mgb.cimmyt.org/gringlobal/search.aspx, accessed 4 Apr. 2017; Wheat: http://wgb.cimmyt.org/gringlobal/search.aspx, accessed 4 Apr. 2017). To begin to connect information on wild relatives of maize and
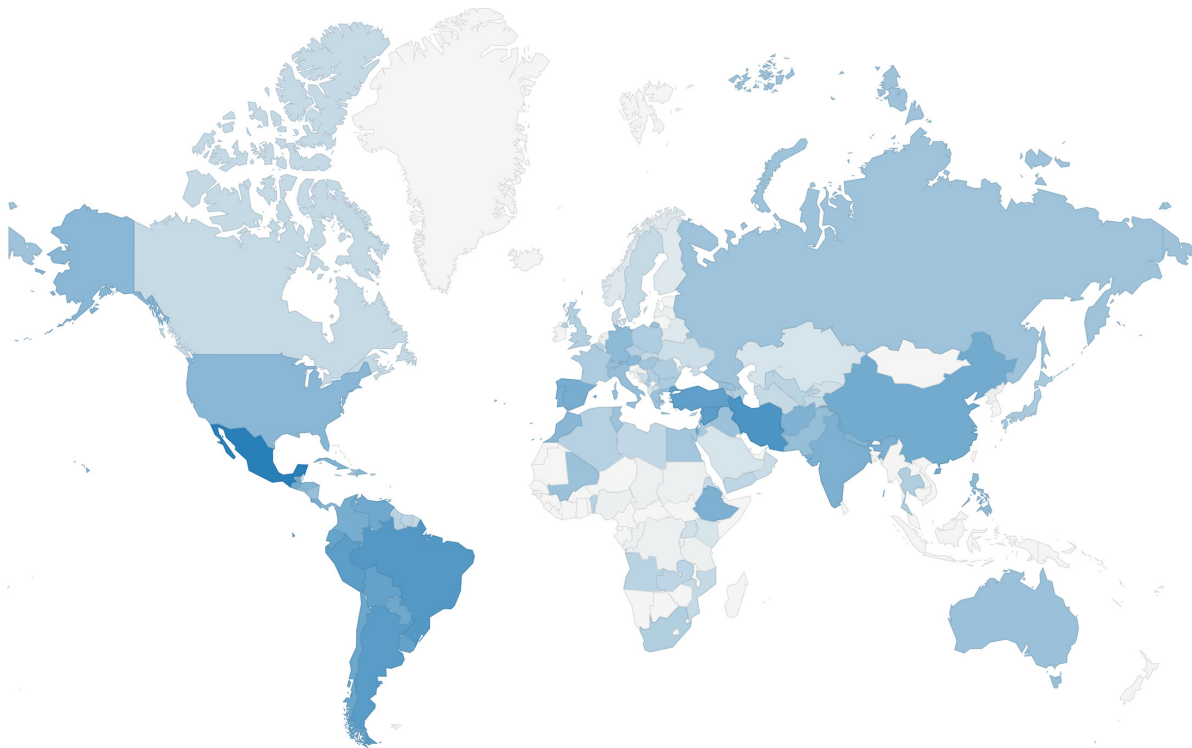
Fig. 6. Worldwide distribution of accessions held across all Germinate installations: Darker saturation of color represents a higher number of accessions from that country. The density of accessions from Mexico and South America is greater owing to the volume of CIMMYT's wheat and maize accessions in the databases.



Fig. 7. Germinate Commonwealth Potato Collection (CPC) database. The Germinate CPC database holds collecting information on 1518 wild potato accessions as well as disease-evaluation data for a number of important potato pathogens.

a



b

Collectingsite map

Fig. 8. (continued on next page) Visualizing climatic data for specific germplasm groups. Germinate allows the graphical representation of climatic data if available. (a) Line chart representation of precipitation in millimeters across all collecting sites within the Germinate Maize database. Min., max., and average values are represented by the varying series. Hovering over a point on the max. and min. series displays the site responsible for that value. (b) A clustered marker view along with overlays exported from ArcGIS are shown using Google Maps. (c) The data is shown in numerical form and color coded on the basis of the value from black (through red and yellow) to white (low to high). The data in can be exported to text or Excel format where the color coding from Germinate is preserved.

wheat to an array of different datasets generated under SeeD, germplasm background information from GRIN-Global has been loaded into Germinate. At this point in time, information is held in Germinate and not pulled automatically from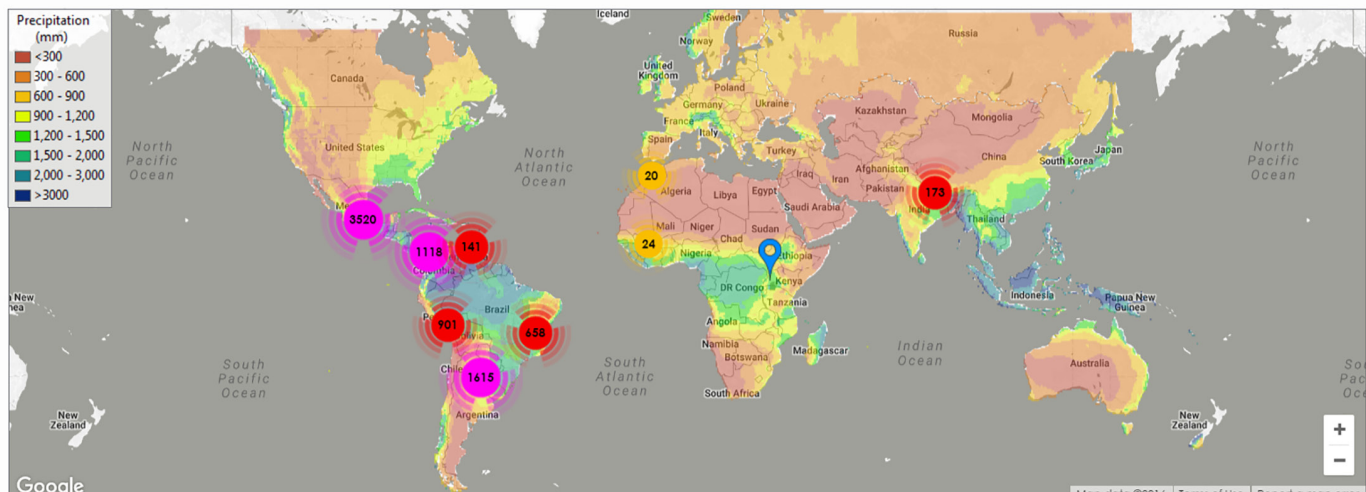 GRIN-Global, but we aim to develop tools to make the transfer of data from GRIN-Global to Germinate as germplasm is made publicly available. In Germinate Wheat, over 40 different species in the *Triticum* and *Aegilops* genera are present in the database, whereas there are over 20 species and subspecies in the *Zea* and *Tripsacum* genera in the Maize database.

## Finding Germplasm from a Wheat Species of Interest

Germinate Wheat supports the searching, retrieval, and analysis of CWR data held in the CIMMYT genebank. The first step in this process may be to search for accessions that belong to specific taxonomic groups. Germinate offers several pathways for identifying and creating custom groups of germplasm based on taxonomic information. First, in the accessions page, users have the option to directly filter for specific genera or species by typing free text values into the dialogue box at the top of the table. Boolean operators can be applied, and partial text matches are supported.

Users can then select all or some of the filtered items and download the available information about them from the accessions page; meanwhile, the selected items also appear in a "shopping cart" that can be used to store lists of accessions and then used to generate Germinate groups from this germplasm for later use. For example, a researcher who uses a filter to generate a list of 38 *Aegilops tauschii* accessions and selects them can create and save a group that can be used to retrieve species-specific phenotypic and genotypic data from the database for this

Color-coded detailed data

C

| Collsite | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 KILOMETROS AL OESTE DE MAYARI | 69.46 | 43.36 | 55.69 | 78.57 | 165.41 | 127.98 | 64.01 | 77.70 | 111.79 | 148.69 | 173.41 | 82.42 |
| 15 KILOMETROS AL OESTE DE MAYARI | 69.46 | 43.36 | 55.69 | 78.57 | 165.41 | 127.98 | 64.01 | 77.70 | 111.79 | 148.69 | 173.41 | 82.42 |
| 15 MI. E OF FREDERIKSTED | 160.13 | 105.21 | 72.58 | 141.90 | 197.03 | 193.29 | 213.37 | 208.62 | 281.65 | 224.96 | 207.77 | 175.27 |
| 15 MIN. WEST OF PORT AU PRINCE | 19.98 | 35.71 | 62.07 | 140.60 | 280.00 | 197.59 | 166.88 | 186.32 | 215.19 | 195.86 | 81.77 | 22.28 |
| 20 KILOMETROS AL OESTE DE HOLGUIN | 31.28 | 23.25 | 40.55 | 67.32 | 199.53 | 170.38 | 101.03 | 105.33 | 136.47 | 133.94 | 77.82 | 28.94 |
| 21 KILOMETROS ESTE DE ELIAS PINA | 31.69 | 32.40 | 35.29 | 94.12 | 204.85 | 141.48 | 133.82 | 138.84 | 170.27 | 179.66 | 90.85 | 33.82 |
| 24 DE MAYO | 38.06 | 67.12 | 85.14 | 93.04 | 57.10 | 31.63 | 16.27 | 18.71 | 31.66 | 50.62 | 49.13 | 51.03 |
| 25 DE MAYO, MACHAGAY | 120.43 | 114.02 | 144.50 | 116.19 | 54.69 | 29.57 | 23.76 | 19.85 | 43.59 | 101.36 | 120.87 | 130.31 |
| 27 DE ABRIL SANTA CRUZ | 0.00 | 0.00 | 2.42 | 27.22 | 241.36 | 274.55 | 176.63 | 234.57 | 343.76 | 392.73 | 129.37 | 9.83 |
| 27 DE ABRIL SANTA CRUZ | 0.00 | 0.00 | 2.42 | 27.22 | 241.36 | 274.55 | 176.63 | 234.57 | 343.76 | 392.73 | 129.37 | 9.83 |
| 28 DE AGOSTO | 9.55 | 7.02 | 6.38 | 6.24 | 16.98 | 38.85 | 54.13 | 53.73 | 54.89 | 29.85 | 9.51 | 11.51 |
| 3 KILOMETROS AL NOROESTE DE HOLGUIN | 31.28 | 23.25 | 40.55 | 67.32 | 199.53 | 170.38 | 101.03 | 105.33 | 136.47 | 133.94 | 77.82 | 28.94 |
| 4 KILOMETROS AL OESTE DE MAYARI | 69.46 | 43.36 | 55.69 | 78.57 | 165.41 | 127.98 | 64.01 | 77.70 | 111.79 | 148.69 | 173.41 | 82.42 |
| 40 KILOMETROS AL OESTE DE HOLGUIN | 29.11 | 31.84 | 35.28 | 48.80 | 192.50 | 191.46 | 128.46 | 127.30 | 168.75 | 133.94 | 51.88 | 17.37 |
| 50 KILOMETROS AL OESTE DE HOLGUIN | 29.11 | 31.84 | 35.28 | 48.80 | 192.50 | 191.46 | 128.46 | 127.30 | 168.75 | 133.94 | 51.88 | 17.37 |
| 56 KM. S.W. OF JEREMIE | 67.65 | 67.22 | 70.60 | 102.00 | 229.46 | 162.48 | 138.19 | 170.36 | 191.08 | 255.98 | 165.41 | 79.14 |
| 6 E ENTRONQUE GUAYACANES | 21.36 | 20.64 | 33.18 | 52.83 | 172.87 | 207.85 | 152.96 | 159.01 | 194.74 | 153.14 | 48.83 | 14.18 |
| 6 KILOMETROS NORTE DE ACALA | 17.50 | 12.42 | 21.39 | 54.28 | 143.22 | 319.29 | 252.33 | 244.79 | 325.65 | 229.04 | 51.84 | 42.85 |
| 60 KM. S.W. OF JEREMIE | 67.65 | 67.22 | 70.60 | 102.00 | 229.46 | 162.48 | 138.19 | 170.36 | 191.08 | 255.98 | 165.41 | 79.14 |

Low (0)                                                                                                                      High (511.53)

Fig. 8. Continued.

species. We aim to further develop the Germinate groups and "shopping cart" functionality to allow the ordering of germplasm based on custom and predefined user groups.

Groups of germplasm can also be constructed de novo from the Germinate groups page in a manner that permits the layering of several search criteria.

Once a group has been created, it appears in the dropdown menus for generating specific reports of germplasm, genotypic, and phenotypic data. Though few phenotypic data have been loaded for the CWR to date, a number of rye varieties have been included in the analyses. A user can create a group of the 274 rye accessions in the database and then select one or more phenotypic datasets from which they wish to retrieve data. Users can visualize or download data solely for the selected group of rye accessions that have been included in these selected datasets. The same species–based data selection method can be applied to the phenotype chart and matrix graphs as well.

## Wheat Genotypic Data

Over 40,000 wheat accessions, including CWRs, have been genotyped using the DArTSeq methodology (Cruz et al., 2013), which can generate over 100,000 SNP calls per DNA sample. These data are currently being prepared for release through Germinate. One of the benefits of sharing the data through Germinate is that it can accommodate the mapping of a single marker to multiple maps. Therefore, a given marker could be assigned to a physical or genetic position on a wheat D chromosome in a bread wheat reference map and could be independently mapped to a different position on an orthologous chromosome using the reference sequence available for the *Aegilops tauschii* genome, one of the wild progenitors of hexaploid wheat. When downloading genotypic data for *Aegilops* accessions, users could then elect to retrieve marker coordinates on the basis of the *A. tauschii* map or any other map that has been loaded to Germinate.

## Maize Landrace Collecting Site Information

As part of the work of the maize group within SeeD, georeferenced location data available for maize accessions were curated and long-term climate metrics for the collection sites were derived using WorldClim (Hijmans et al., 2005) and other relevant climatic datasets.

This climatic data can be viewed within Germinate, including cloud cover, aridity index, precipitation, and mean, minimum, and maximum monthly temperature metrics. The data, which can be viewed using Germinate groups where required, provides a graphical representation of variation of the data, including identification of those sites corresponding with the minimum and maximum variables and geographical representation of the data, displaying global variation for the trait of interest with collection sites of members of the group. Finally, color-coded tabular display of the data provides site-specific annual variation of the characteristic of interest (Fig. 8).

The ability to search for accessions within a particular geographical area is an important and widely used feature of Germinate. We have used the Germinate polygon search functionality when identifying accessions of interest from specific geographic regions such as those that come from areas where problematic pathogens are endemic.

## Maize Landrace Additional Characters

In addition to typical passport data, Germinate offers the flexibility to assign additional *user-defined* attributes to germplasm held within Germinate. In the case of maize, a number of landraces have been used in controlled crosses as part of a large genome-wide association scan (GWAS) experiment. Information relevant to this experiment, such as DNA sample identifiers, test germplasm identifiers, and corresponding testcross germplasm identifiers, have all been stored using these user-submitted descriptors, which can then be easily exported along with any other relevant attributes from Germinate.

## Maize Genotypic Diversity within a Collection

Genotypic characterization of the maize accessions offers a particular challenge given the allogamous breeding nature of the species involved. The genotyping approaches adopted within SeeD enable the derivation of allele frequency estimates for each of the loci evaluated within the accessions evaluated. In cases where the CWR accessions are not pure-breeding inbreds, it may also be useful to genotype for allele frequencies within the accession rather than to just indicate which alleles are present. Germinate caters to this need by offering specific tools to store and extract allele frequency data. To visualize a continuous variable in a graphical genotype viewer such as Flapjack, the frequency data must first be binned into discrete categories. Germinate handles this binning and, subsequently, the export of marker frequency data, along with relevant genetic or physical maps, in such a way as to allow users to define three possible binning strategies and vary the number of categories to view within Flapjack. This automation of binning reduces user error and better facilitates interpretation of the data prior to further analysis. In addition, users have the option of applying a default "Missing Data Filter" prior to generating the bins and can receive a report of the loci removed by this filter.

## Germinate Maize and Wheat User Access Control

While our data is free to access, the user access control features of Germinate (using Germinate Gatekeeper) gives CIMMYT the ability to ensure that registrants sign a license agreement that informs users of their rights and obligations concerning the access and use of data. For CWRs and other materials held in trust in CIMMYT's genebank, license agreements can help to ensure compliance with relevant treaties, protocols, and donor requirements. At CIMMYT, Germinate Gatekeeper handles user registration and manages user access to each crop-level database. However, in cases where finer grain access control is needed for sensitive or prepublication data, specific users can be granted permission to see and obtain data from specific datasets.

## Germinate Maize and Wheat Links to and from External Resources

Germinate houses a number of different useful data types at CIMMYT. However, in some cases, it is advantageous to store datasets externally or link to external datasets or data resources important to a project. Germinate has implemented the ability to connect to external datasets and data resources. This feature has been used to link to datasets directly managed by the SeeD initiative, and we envisage that this could be further expanded on by our other projects to direct users to data or additional information housed in other resources. For example, when the underlying FASTQ sequence data for a CWR obtained during DArTSeq analyses are loaded to the European Nucleotide Archive (http://www.ebi.ac.uk/ena, accessed 4 Apr. 2017), a link to this dataset will be made available from Germinate. Future plans also include a direct link from each Germinate accession page to its corresponding accession page in GRIN-Global to facilitate seed requests. This is achieved by passing the accession name to a Germinate page, via a uniform resource locator (URL) parameter, which returns all data for the specified accession. External resources can also link back to Germinate, and one of CIMMYT's priorities is to provide a reciprocal link from accession pages in GRIN-Global to the relevant Germinate accession page so that we can see exactly which datasets a particular germplasm accession has data for.

## Germinate Integration with Visualization and Data Collection Tools

By default, Germinate has data export tools that are suitable for a number of external visualization tools ranging from pedigree visualization to graphical genotype visualization. These export tools increase the utility of Germinate, allowing complex analysis to be performed in dedicated external applications while ensuring the easy-to-export data in these formats if required.

## Germinate Integration with Information Visualization and Collection Tools

Helium is a desktop pedigree visualization platform developed to aid in the comprehension of complex plant pedigrees (Shaw et. al., 2014). Germinate is fully Helium compatible and allows the export of pedigree and trait data from Germinate into ".helium" files that can be opened and viewed in Helium. Helium uses a simple import format that builds up pedigree graphs from a series of parent-child relationships. Phenotypic data is stored as a simple data matrix. Trait data types can be defined, which affects the way in which Helium handles color coding, depending on if the character is qualitative or quantitative.

Flapjack (Milne et al., 2010) is a visualization tool for graphical genotype visualization. It provides the framework to allow the navigation and comparisons between lines and their underlying genotypic data. Flapjack works with any marker technology, although by far the most commonly seen in our Germinate instances are SNP and genotype-by-sequencing (GBS) marker technologies. Data is represented as a data matrix, with markers along the $x$ axis and germplasm on the $y$ axis. Row intersections are color coded to represent the allelic state at that specific line by marker combination. Flapjack allows for comparisons between lines by performing similarity calculations and clustering and facilitates the integration of both phenotypic data and quantitative trait locus data where this is available. Germinate exports data into native Flapjack project file format.

If data is exported into Flapjack, it is then possible to select a marker or plant line from within Flapjack and link directly back to Germinate to retrieve additional information where available. This is an automatic process when exporting data from Germinate into Flapjack format and works by including the necessary information to identify the data source into the Flapjack project file.

CurlyWhirly provides a means of representing multidimensional information such as PCA and PCO analysis data. Where a Germinate instance has results from these types of experiments tools to export this data into CurlyWhirly format are available. If data has been exported in CurlyWhirly format from Germinate, users can select a data point and get additional information on the plant line from Germinate. In addition, user-selected groups from CurlyWhirly can be automatically linked to Germinate, which will generate a user-defined group

based on the selected germplasm. This means that users can select germplasm using the CurlyWhirly interface on the basis of specific requirements and automatically generate Germinate groups that can be used in subsequent data export.

Germinate Scan (https://ics.hutton.ac.uk/germinatescan, accessed 4 Apr. 2017) facilitates the collection of phenotypic data in the field or glasshouse and presents the collected data in a format suitable for upload into Germinate. We have developed a simple Android-based mobile application that interfaces with external Bluetooth barcode scanners (or uses the internal camera functionality of Android devices) to increase the speed and reliability that qualitative characters on plant lines can be collected in the field. Traits and their associated phenotypes are defined on data collection sheets. Data collection is performed by scanning the plant's barcode, then the trait, then one of the defined values on the data collection sheet, then proceeding to the next plant or trait. Each reading is tagged with the date and time of collection, and where global positioning system (GPS) data is available to the mobile device, this is also recorded for each reading.

Germinate Scan is being extensively used in the maintenance and running of the CPC by the James Hutton Institute by allowing phenotypic data and images on CWR species to be quickly and efficiently recorded each year.

## DISCUSSION

Germinate offers a platform for the distribution of genetic resources data across an increasing number of species and their wild relatives. The use of a common platform allows us to maintain consistency across installations and deploy new features across all installations both quickly and efficiently. We have shown that Germinate has to be deployed successfully across a number of species and data types and offers a functional online database to allow users to make their germplasm collections, and associated data, publicly available.

As Germinate continues to develop as a platform, we are looking toward increasing its functionality. Our aim is to develop improved Germinate components used to store complex pedigree data, DOI codes for both germplasm and datasets, and ontology features to allow us to browse phenotypic data and define available phenotypes in ontological terms. As with our other features, we will make these as generic as possible without compromising data integrity and performance of the platform.

## Availability

The Germinate platform is open source and freely available from https://ics.hutton.ac.uk/germinate (accessed 14 Apr. 2017). We welcome both input from Germinate users on how we can improve the platform and contributions to features. Our aim is to increase community involvement in the project, with a goal of providing tools to allow genetic resources projects to easily and freely make their experimental data available to the public, especially groups who may not have access to dedicated bioinformatics teams.

We also provide detailed installation instructions and tutorial videos that can be used to aid in the installation of Germinate https://ics.hutton.ac.uk/germinate/ and from https://www.youtube.com/channel/UC34WPRCVzrtN-I7U9RhIPGIA (accessed 14 Apr. 2017).

A standard installation in a Docker image for groups that wish to evaluate Germinate and test before committing to use our platform is available along with installation instructions from https://ics.hutton.ac.uk/germinate/germinate-docker-image (accessed 14 Apr. 2017).

For further information on Germinate please visit https://ics.hutton.ac.uk/germinate. We can also be contacted via the email address germinate@hutton.ac.uk.

The pedigree visualization tool Helium is free to use and can be downloaded from https://ics.hutton.ac.uk/helium (accessed 14 Apr. 2017).

The visualization tools Flapjack, CurlyWhirly, and Strudel are open source and can be downloaded from https://ics.hutton.ac.uk (accessed 14 Apr. 2017).

Germinate Pea can be freely accessed from http://bioinf.hutton.ac.uk/germinate_pea (accessed 14 Apr. 2017).

Germinate Wild Barley can be freely accessed from http://ics.hutton.ac.uk/germinate-wild-barley (accessed 14 Apr. 2017).

Germinate CPC can be freely accessed from https://ics.hutton.ac.uk/germinate-cpc (accessed 14 Apr. 2017).

Germinate Maize (SeeD) requires registration and can be accessed from http://germinate.seedsofdiscovery.org/maize (accessed 14 Apr. 2017).

Germinate Wheat (SeeD) requires registration and can be accessed from http://germinate.seedsofdiscovery.org/wheat (accessed 14 Apr. 2017).

Germinate Scan can be downloaded through the Google Play Store (https://play.google.com/store/apps/details?id=uk.ac.hutton.android.germinatescan, accessed 14 Apr. 2017).

# References

Alercia, A., S. Diulgheroff, and M. Mackay. 2015. FAO/Bioversity Multi-Crop Passport Descriptors V.2.1 [MCPD V.2.1]. http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/, accessed 4 Apr. 2017.

Bradshaw, John E., and Gavin Ramsay. 2005. Utilisation of the Commonwealth Potato Collection in potato breeding. Euphytica 146(1–2):9–19. doi:10.1007/s10681-005-3881-4

Cruz, M.V., A. Kilian, and D.A. Dierig. 2013. Development of DArT marker platforms and genetic diversity assessment of the U.S. collection of the new oilseed crop *Lesquerella* and related species. PloS One 8(5):E64062. doi:10.1371/journal.pone.0064062

Fahlgren, N., M.A. Gehan, and I. Baxter. 2015. Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. Current Opinion in Plant Biology 24(April):93–99.

Godfray, H., J. Charles, J.R. Beddington, I.R. Crute, L. Haddad, D. Lawrence, J.F. Muir, J. Pretty, S. Robinson, S.M. Thomas, and C. Toulmin. 2010. Food security: The challenge of feeding 9 billion people. Science (New York) 327(5967):812–18. http://science.sciencemag.org/content/327/5967/812.abstract, accessed 4 Apr. 2017.

Hawkes, J. G. 1951. The Commonwealth Potato Collection. American Potato Journal 28(1):465–71. doi:10.1007/BF02854979

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25(15):1965–1978. doi:10.1002/joc.1276

Jing, R., A. Vershinin, J. Grzebyta, P. Shaw, P. Smýkal, D. Marshall, M.J. Ambrose, T.H.N. Ellis, and A.J. Flavell. 2010. The genetic diversity and evolution of field pea (*Pisum*) studied by high throughput retrotransposon based insertion polymorphism (RBIP) marker analysis. BMC Evol. Biol. 10(January):44. doi:10.1186/1471-2148-10-44

Kopecky, D., J. Bartos, A. Lukaszewski, J. Baird, V. Cernoch, R. Kolliker, O.A. Rognli, et al. 2009. Development and mapping of DArT markers within the *Festuca-Lolium* complex. BMC Genomics 10(1):473. doi:10.1186/1471-2164-10-473

Lee, J.M., G.F. Davenport, D. Marshall, T.H.N. Ellis, M.J. Ambrose, J. Dicks, T.J.L. 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(2):619–631. doi:10.1104/pp.105.065201

Mascher, M., T.A. Richmond, D.J. Gerhardt, A. Himmelbach, L. Clissold, D. Sampath, S. Ayling, et al. 2013. Barley whole exome capture: A tool for genomic research in the genus *Hordeum* and beyond. Plant J. 76(3):494–505. doi:10.1111/tpj.12294

McCouch, S., G.J. Baute, J. Bradeen, P. Bramel, P.K. Bretting, E. Buckler, J.M. Burke, et al. 2013. Agriculture: Feeding the future. Nature 499:23–24. doi:10.1038/499023a

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(24):3133–3134. doi:10.1093/bioinformatics/btq580

Nelson, G.C., M. Rosegrant, J. Koo, R. Robertson, T. Sulser, T. Zhu, S. Msangi, et al. 2009. Climate change: Impact on agriculture and costs of adaptation. International Food Policy Research Institute. 30 p. doi:10.2499/089629535

Onda, Y., and K. Mochida. 2016. Exploring genetic diversity in plants using high-throughput sequencing techniques. Current Genomics 17:356–65. http://www.eurekaselect.com/140192, accessed 4 Apr. 2017.

Postman, J., K. Hummer, P. Bretting, G. Kinard, M. Bohning, G. Emberland, Q. Sinnott, et al. 2010. GRIN-Global: An international project to develop a global plant genebank information management system. Acta Hortic. 859:49–56. doi:10.17660/ActaHortic.2010.859.4

Russell, J., M. Mascher, I.K. Dawson, S. Kyriakidis, C. Calixto, F. Freund, M. Bayer et al. 2016. Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation. Nature Genetics 48(9):1024–30. doi:10.1038/ng.3612

Shaw, P., M. Graham, J. Kennedy, I. Milne, and D. Marshall. 2014. Helium: Visualization of large scale plant pedigrees. BMC Bioinformatics 15(1):259. doi:10.1186/1471-2105-15-259