Received: 30 April 2020 | Accepted: 16 June 2020

DOI: 10.1002/csc2.20248

Crop Science

SPECIAL ISSUE ARTICLE

Adapting Agriculture to Climate Change: A Walk on the Wild Side

From bits to bites: Advancement of the Germinate platform to support prebreeding informatics for crop wild relatives

Sebastian Raubach[1] | Benjamin Kilian[2] | Kate Dreher[3] | Ahmed Amri[19] | Filippo M. Bassi[19] | Ousmane Boukar[12] | Douglas Cook[8] | Alan Cruickshank[15] | Christian Fatokun[11] | Noureddine El Haddad[19] | Alan Humphries[10] | David Jordan[15] | Zakaria Kehel[19] | Shiv Kumar[19] | Sandy Jan Labarosa[17] | Loi Huu Nguyen[16] | Emma Mace[15] | Susan McCouch[14] | Ken McNally[13] | David F. Marshall[7] | Erick Owuor Mikwa[9] | Iain Milne[1] | Damaris Achieng Odeny[9] | Mariola Plazas[4] | Jaime Prohens[4] | Loren H. Rieseberg[5] | Roland Schafleitner[18] | Shivali Sharma[6] | Gordon Stephen[1] | Huynh Quang Tin[16] | Abou Togola[11] | Emily Warschefsky[5] | Peter Werner[2] | All our CWR Pre-Breeding Partners and Collaborators | Paul D. Shaw[1]

[1] Department of Information and Computational Sciences, The James Hutton Institute, Errol Road, Invergowrie, Dundee, Scotland DD2 5DA

[2] Global Crop Diversity Trust, Platz der Vereinten Nationen 7, Bonn 53113, Germany

[3] International Maize and Wheat Improvement Center (CIMMYT), Edo. De Mexico, Texcoco, Mexico CP56237

[4] Universitat Politècnica de València, Valencia 46022, Spain

[5] Department of Botany, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

[6] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Andhra Pradesh 502324, India

[7] SRUC, Peter Wilson Building, West Mains Road, Edinburgh, Scotland EH9 3JG

[8] Department of Plant Pathology, College of Agricultural and Environmental Sciences, University of California–Davis, One Shields Avenue, Davis, CA 95616, USA

[9] International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)–Eastern and Southern Africa, PO Box 39063 - 00623, Nairobi, Kenya

[10] South Australian Research and Development Institute (SARDI), Waite Institute, Urrbrae 5064, Australia

[11] International Institute of Tropical Agriculture (IITA), Kano, Nigeria

[12] International Institute of Tropical Agriculture (IITA), HQ, Ibadan, Nigeria

[13] International Rice Research Institute (IRRI), Pili Drive, Los Baños, Laguna 4031, Philippines

[14] Plant Breeding & Genetics, Cornell University, Ithaca, NY 14850, USA

[15] University of Queensland and Department of Agriculture and Fisheries, Warwick, Qld 4370, Australia

[16] Mekong Delta Development Research Institute, Can Tho University, Can Tho, Vietnam

Abbreviations: API, application programming interface; BCNAM, backcross nested association mapping; BrAPI, plant breeding application programming interface; CSSL, chromosome segment substitution line; CWR, crop wild relative; DAS, days after sowing; DOI, data object identifier; FAO, Food and Agriculture Organization; GBS, genotyping-by-sequencing; HDF5, Hierarchical Data Format version 5; IL, introgression line; IRRI, International Rice Research Institute; MCPD, Multi-crop Passport Descriptors; ORSC, *Oryza rufipogon* Griff. species complex; SNP, single nucleotide polymorphism.

[17] The University of Bonn/Crop Trust, Bonn, Germany

[18] World Vegetable Center, Shanhua, Taiwan

[19] International Center For Agricultural Research In The Dry Areas (ICARDA), Rabat, Morocco

Correspondence
Paul D. Shaw, Department of Information and Computational Sciences, The James Hutton Institute, Errol Road, Invergowrie, Dundee, DD2 5DA, Scotland.
Email: paul.shaw@hutton.ac.uk

Assigned to Associate Editor Luigi Guarino.

Abstract

Management and distribution of experimental data from prebreeding projects is important to ensure uptake of germplasm into breeding and research programs. Being able to access and share this data in standard formats is essential. The adoption of a common informatics platform for crops that may have limited resources brings economies of scale, allowing common informatics components to be used across multiple species. The close integration of such a platform with commonly used breeding software, visualization, and analysis tools reduces the barrier for entry to researchers and provides a common framework to facilitate collaborations and data sharing. This work presents significant updates to the Germinate platform and highlights its value in distributing prebreeding data for 14 crops as part of the project 'Adapting Agriculture to Climate Change: Collecting, Protecting and Preparing Crop Wild Relatives' (hereafter Crop Trust Crop Wild Relatives project) led by the Crop Trust (https://www.cwrdiversity.org). The addition of data on these species compliments data already publicly available in Germinate. We present a suite of updated Germinate features using examples from these crop species and their wild relatives. The use of Germinate within the Crop Trust Crop Wild Relatives project demonstrates the usefulness of the system and the benefits a shared informatics platform provides. These data resources provide a foundation on which breeding and research communities can develop additional online resources for their crops, harness new data as it becomes available, and benefit collectively from future developments of the Germinate platform.

1 | INTRODUCTION

Harnessing the variation contained within cultivated crops, landraces, and wild relatives is a strategy that can be used to improve the resilience and sustainability of cropping systems to mitigate the effects of climate change (Dempewolf et al., 2017; Jarvis, Lane, & Hijmans, 2008; Nelson et al., 2009; Redden et al., 2015). High-throughput phenotyping (Fahlgren, Gehan, & Baxter, 2015) and genotyping (Onda & Mochida, 2016) technologies are increasingly used to evaluate plant genetic resources and breeding materials. The resulting large quantities of data are derived from a diversity of technologies (array-based genotyping, whole-genome sequencing, remote sensing, image files, etc.) with correspondingly diverse data formats and user needs. Unless forethought is given to storage, processing, and presentation, such data may languish (Marx, 2013; Nandyala & Kim, 2016). Careful planning is required to ensure that appropriate information systems are in place to manage, integrate, and store generated data. Information systems that provide user-friendly tools and that allow data to be efficiently stored, queried, and visualized are therefore essential.

The recording of experimental data from large, multiactor projects presents unique challenges that we have tried to address with our platform. These problems are not new and are common to every body of work that we have seen. The challenges include, but are not limited to, the following:

a. Data formats: It is usual for each group to have their own specific way to store and record data; however, this can be problematic at a number of levels. Firstly, sharing data between groups requires significant investment in data wrangling in order to standardize formats. This requires experienced bioinformaticians or data scientists with the necessary data handling skills. Secondly, data must be collected using standardized protocols so that data generated by different researchers can be integrated. For example, phenotypic data must be scored in a consistent and reproducible way, based on established phenotyping procedures. Initiatives to develop crop and trait ontologies (Shrestha et al., 2010, 2012) provide a standard platform for data collection but researchers must agree to follow the prescribed protocols and use the ontologies to facilitate downstream data integration.

b. Data sharing: The use of email for distribution and sharing of experimental data amongst project partners is a common practice but introduces problems that make it difficult to know which version of a data set is the most up to date and how to reconcile multiple updates to the same file by different people, particularly if the data set was sent to multiple partners simultaneously. Solutions that can help address this problem include the use of cloud-based online drive storage such as Google Drive (https://drive.google.com) or Office 365 (https://onedrive.live.com). These platforms provide storage and tools that allow users to work on the same documents and files in real time, with all changes and edits tracked to individual users. These services do, however, usually rely on maintaining an active internet connection, something which may not always be possible. Along with the clear advantages to sharing data quickly, these solutions also have the knock-on effect of ensuring that data owners are quick to update problems and undertake good data handling practice because of the visibility of their work amongst collaborators. Those working with the data directly and those developing bioinformatics solutions around these data have immediate access to the most up-to-date versions available.

c. Data versioning: When data sets are generated, it is common for updates to be made correcting typographical errors, incorrect formatting, and misuse of identifier names. This presents complications over time and so efficient ways of versioning data sets is crucial. Much like with data sharing, the use of cloud-based solutions can help address these issues. Other options include the use of version control software such as Github, which can be used to keep track of all changes made to a document over time. It is important that the number of copies of the same data set that is being worked on is kept to a minimum.

d. Historical data: Historical data are common and cause problems for uploading of data into databases. Ensuring that historical data are accurate is problematic, especially when the original curator or generator of the data is no longer available to answer questions or, indeed, when researchers no longer remember what was done. For this reason, care should be taken with historical data and a timely investment to make sense of older data before knowledge is lost is crucial.

Core Ideas

- Common data platform for pre-breeding data.
- Making data from genetic resources collections available.
- Information visualization.
- Crop wild relatives pre breeding data.

Information systems need to be flexible to adapt to new data types and new uses of existing data types (Germeier & Unger, 2019). This is especially important in academic projects that tend to be under-resourced in terms of informatics capability. There are rarely enough human or computational resources available to develop fully functional systems and few have long-term sustainability plans. The problem that emerges with stagnant information systems is that development is no longer active and it is not clear whether to invest significant resources to update and upgrade those systems or simply to start again using current information platforms and adapt the core functionality of older systems. A relatively small investment to improve or expand the functionality of current systems may lead to large gains in user satisfaction with increasing applicability to multi-investigator research projects. Buy-in from user groups is critical to the vitality and continued development of information systems, and active collaboration between researchers in the field and bioinformatics groups developing software should be encouraged. It is often the case that organizations creating information platforms do not fully use them in-house. This is a problem that can be addressed through codevelopment strategies and augmented by training and promotion of the benefits of the platforms wherever possible.

The use of standard platforms, such as Germinate, present a number of distinct advantages to those who want to make their data available. These include (a) adoption of an existing system is often easier than developing a new system; (b) Germinate provides tools that allow users to customize their interaction with the platform, allowing Germinate pages to be branded for specific projects and for users to turn off the interface with information pages that are not relevant to their work; (c) translations and internationalization are provided out-of-the-box, meaning that Germinate's interface can be quickly made available in multiple languages; (d) links are provided to other information resources, that is, to distributed data, additional information on projects, and to national and international germplasm collections through tools such as Genesys (https://www.genesys-pgr.org), EURISCO (Weise, Oppermann, Maggioni, Van Hintum, & Knupffer, 2017; https://eurisco.ipk-gatersleben.de), and Grin Global (Postman et al., 2010; https://www.grin-global.org): (e) standardized templates are provided for uploading of data into the system, and these can be shared with project partners; (f) contribution to the development of a platform benefits not only your own work but also the community; and (g) support is provided to help smaller groups host their data and address questions about their informatics requirements.

Research groups should embrace state-of-the-art informatics technology wherever possible. The use of mobile-based tools can also help in reducing the number of data errors introduced into experimental data sets using solutions that already exist (Rife & Poland, 2014; https://ics.hutton.ac.uk/get-germinate-scan). While there is greater resistance to the adoption of such technologies than would be expected, once adopted they offer significant advantages over manual recording of data using pen and paper. Such tools can allow the removal of ambiguity, enforce the use of ontologies, improve the speed of data collection, perform data sanity checks at the point of collection, and should be regarded as the benchmark for reliable data collection for input into information systems moving forwards.

One of the fundamental issues in developing new information systems is how to confront the challenge of integrating new kinds of data emerging from rapidly evolving technologies. There are significant problems in ensuring that systems and data schemas are flexible enough to allow incorporation of new technologies but not so generic they lose focus on their primary research objectives. The current shift from desktop and laptop computers to mobile devices means that users are increasingly reliant on being able to browse data on smaller screened devices. Ensuring that new tools and utilities, as much as possible, can be viewed on mobile devices is important for their uptake and use in the research community. To address some of these issues in Germinate, we have adopted the strategy that basic background information on germplasm will all be accessible from mobile devices using state-of-the-art responsive interfaces, whereas analyses requiring more complex data integration and visualization, while compatible in most cases, are recommended for desktop platforms.

There are three major data management components required for standardized management of genetic resource collections and their associated data (Shaw et al., 2017). The first of these are systems for germplasm management, enabling genebank or collection managers to manage information about the availability of resources in the collections for which they are responsible (Postman et al., 2010). Secondly, systems that allow the collation and integration of data across species and germplasm management systems, enabling researchers and breeders to identify and access all suitable germplasm for their requirements (Weise et al., 2017). The final type of system provides tools to allow the integration of experimental data, namely phenotypic, genotypic and environmental data, and allows a user to export relevant data based on filtering of passport and other germplasm categorization data (Blake et al., 2012; Shaw et al., 2017).

Here we present a significant redesign of the Germinate 3 platform. We describe the technical architecture of the system and the main standards that it implements then show how through user testing using domain experts we have been able to identify challenges users have had with the system and provided technical solutions and tweaks to our user interfaces to help mitigate them. We describe the development of new tools to help scientists navigate the large data sets that typify many modern breeding and prebreeding programs and to help them better classify, explore, and interact with the underlying germplasm. Finally, we conclude by describing these tools in the context of storing data obtained from 14 different species of significant agricultural importance to large parts of the world from the Crop Trust Crop Wild Relatives project whereby Germinate has become the primary information platform for distribution of the projects prebreeding data.

From this point forward we will refer to new developments as ‘Germinate’ and older versions suffixed with their version number for clarity. This change in branding is a reflection that as we move forwards, we are working to ensure that all Germinate instances will be compatible with one another and that users will be upgraded to the latest versions by default. We want to ensure that updates are as smooth as possible and that all Germinate users can benefit from future developments across the community quickly.

2 | MATERIALS AND METHODS

Germinate 3 (Shaw et al., 2017) is an open source plant resources platform that stores experimental data and provides a web-based visual query interface for plant genetic resources collections. The main purpose of Germinate is to store and make analyzed, clean data available to interested parties. Germinate has seen significant advances in its development from a platform capable of routinely handling tens of markers and hundreds of germplasm entries (Lee et al., 2005) to hundreds of thousands of markers and germplasm entries (Shaw et al., 2017) to the current version that is suitable for the storage of many millions of data points (markers, phenotype, and germplasm entries). Germinate continues to see long-term development, support, and stability through core support from the James Hutton Institute as well as funding from a number of organizations and national and international funding bodies. The initial implementation supported work on potato (*Solanum* spp.), barley (*Hordeum vulgare* L.), wheat (*Triticum* spp.), and maize (*Zea mays* L.), and today Germinate hosts data from 20 diverse species, including 17 publicly available collections as shown in Table 1. Data has been incorporated from several large international projects including Seeds of Discovery (https://seedsofdiscovery.org) and the Crop Trust Crop Wild Relatives project (https://www.cwrdiversity.org). In addition, it is used to support both commercial and academic projects, ranging from sweet cherry [*Prunus avium* (L.) L.] and soft fruits (*Rubus* spp., *Vaccinium* spp., and *Ribes* spp.) to characterization data for germplasm collections of national importance such as the Commonwealth Potato Collection (Bradshaw & Ramsay, 2005; Hawkes, 1951). The geographical distribution of collection sites and testing locations for which Germinate holds data is shown in Figure 1.

2.1 | Germinate architecture

The underlying Germinate database has been implemented using the commonly used MySQL (www.mysl.com) relational database management system and is compatible with versions from MySQL v5.7.22 or later. While we cannot guarantee complete compatibility, it also can be run without issues using MariaDB (https://mariadb.org). Additional database management systems introduce complexity but we are committed to supporting the most commonly used, freely available system MySQL/MariaDB. Germinate is composed of 70 tables and uses 31 views and six stored procedures to reduce the complexity of commonly used queries. The increase in complexity has allowed us to introduce new features and accommodate new data types, for instance chemical compound data, data licenses, asynchronous data import and export, and more detailed pedigree storage functionality.

The previous Germinate 3 was developed using GWT (http://www.gwtproject.org) web technologies, which were state of the art at the time. Recent advancements and the quick uptake and development of modern JavaScript libraries has meant that there are now technological solutions better suited to Germinate's development. After research and testing of the mainstream JavaScript libraries such as React (https://reactjs.org/), Vue.js (https://vuejs.org), and AngularJS (https://angular.io/), we decided that Vue.js offered us the best mix of functionality and features that were required to meet Germinate's web interface requirements. Based on the advantages offered by Vue.js, we adopted this technology for all Germinate development from this point forward. Vue.js is both state of the art and a commonly used library and has already provided us with four main advantages over the previous GWT-based Germinate 3 platform. These advantages include (a) enhanced stability of the system, (b) improved ease of maintenance, (c) faster development cycles, and (d) a significant reduction in the barrier to entry for new developers looking to contribute to the project—something we encourage. Bringing new users and developers into the Germinate ecosystem is important to ensure that Germinate becomes a community platform with buy-in from a diversity of users, primarily groups who currently have restricted bioinformatics support or no data distribution solutions in place. The main problems that needed to be addressed with the use of the older GWT technology were the significant increase in the complexity of the application when even minor updates or bug fixes were performed. The amount of time taken to recompile and deploy the application after updates was also an issue and this procedure was required even when making minor edits to user interface translations. The use of Vue.js has removed these problems. To offer flexibility, we have implemented all communication between the Germinate web interface and database backend through the new Germinate application programming interface (API). This API (Figure 2) offers greater flexibility for developers to tailor solutions that use desktop and mobile based interfaces and provide a more accessible interface for third-party tools searching and retrieving data held in Germinate. The development of new informatics components (including data upload, query, and visualization interfaces), maintenance of current features to reduce complexity and technical debt, and the ease and speed with which updates can be deployed have all been significantly simplified with the new Germinate platform.

One important design consideration was ensuring that Germinate is and will continue to be free to use, free to develop, and have no tie-ins restricting its use within academic or industrial settings. Only unrestrictive, open-source libraries have been used in the development of the platform in an effort to ensure future sustainability.

Germinate maintains all its original features (Shaw et al., 2017) including its full compatibility with the Multi-Crop Passport Descriptors (MCPD V2.1) presented by the United Nations Food and Agriculture Organization (FAO) (Alercia, Diulgheroff, & Mackay, 2015) but has evolved to include compatibility with a number of other defined standards including the Dublin Core Metadata standards to describe digital resources, data object identifiers (DOI) through the Global Information System (https://ssl.fao.org/glis) being developed by the FAO and the BrAPI (Selby

TABLE 1 Current publicly available Germinate databases showing publicly available databases along with a description of the volumes of data that each contains at the time of writing. In addition to these databases, similar resources for lentil and durum wheat will be available toward the end of 2020. Those databases marked with an asterisk (*) are discussed in this work and used in the examples presented for the Germinate user interface

Species	Germplasm	Taxa	Markers	Locations	Genotypic data sets	Phenotypic data sets	Traits
Alfalfa*	278	15	–	156	–	–	–
Barley*	1,803	12	577,011	1,505	1	–	–
Chickpea*	22,789	7	–	860	–	5	72
Cowpea*	12,895	1	–	6	–	17	51
Demo (JHI)[a]	2,024	2	5,000	509	2	2	51
Eggplant*	2,467	16	1,508	34	5	30	251
Finger millet*	681	8	27,312	4	2	4	28
Grasspea*	5,551	46	40202	1,335	1	8	64
Maize (CIMMYT)	61,084	24	1,633,755	10,251	3	49	110
Pigeon pea*	461	5	3,687	8	2	20	9
Pearl millet*	221	2	–	7	–	14	43
Potato (CPC; JHI)[a]	461	121	–	578	–	–	–
Rice*	3,304	4	–	3	–	1	14
Sorghum*	1,236	0	17,235	0	1	–	–
Sunflower*	884	23	3,891	177	1	2	6
Wheat (CIMMYT)	298,224	69	1,114,298	14,246	19	32	679

[a] JHI, James Hutton Institute; CPC, Commonwealth Potato Collection.

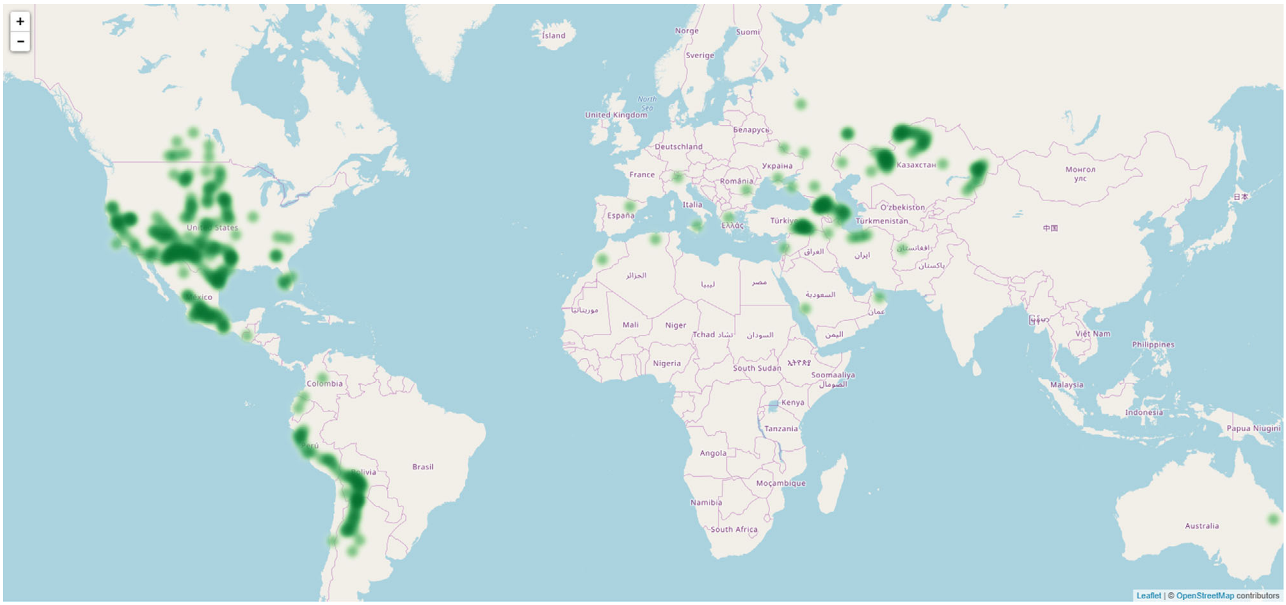


FIGURE 1 The current Germinate crop wild relative (CWR) database germplasm collecting sites and experimental trial sites for alfalfa, barley, chickpea, cowpea, eggplant, finger millet, pigeon pea, Commonwealth Potato Collection, rice, and sunflower. The darker areas show higher density of samples from that location. Overview statistics like this can be generated to show the distribution of data on a geographical map context and give an indication of the diversity of data that Germinate currently holds

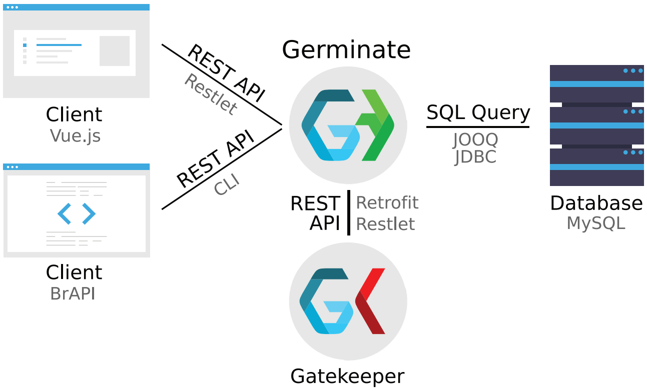


FIGURE 2 The current Germinate architecture. The Germinate core application consists of the Germinate user interface, optional user authentication using Germinate Gatekeeper, and the underlying database which is currently implemented in MySQL 5.7. The current Germinate platform supports a hybrid approach whereby public and private data can exist in the same database, a major advancement over the previous platform where a database was either fully public or private. This hybrid approach allows us not only to retain confidential data sets for authenticated users but allows users to upload new data sets and only make them public when they are ready to do so

et al., 2019). The result of these developments has been to open up Germinate's internal data schema to new tools and resources via a representational state transfer interface.

Germinate has advanced its capabilities as a platform when linking to external visual analytics tools such as Helium (Shaw, Graham, Kennedy, Milne, & Marshall, 2014) and Flapjack (Milne et al., 2010). We have implemented a Germinate BrAPI interface for Flapjack, which ensures users do not need to export data to an intermediary file for analysis.

2.2 | Germinate 3 user testing

User testing allows developers to engage in structured interaction with end users. It can be helpful in identifying issues users have with software and targets where improvements would increase software acceptance or utility. User testing has been shown to be a critical component in the development of software (Sedlmair, Meyer, Munzner, 2012; Munzer, 2009; Lam, Bertini, Isenberg, Plaisant, & Carpendale, 2011) but is sadly overlooked in many bioinformatics tools (Shaw et al., 2014). A subjective user evaluation was performed on the Germinate 3 platform (v3.6) to determine both user acceptance and to identify areas where improvements could be made. It also established a benchmark against which future Germinate developments could be compared. We used this evaluation to assess if users were able to perform basic data operations using the Germinate web-based interface and the resulting test data was used to undertake targeted development of the Germinate interface. The testing consisted of a Google Forms based online test that was undertaken by 17 domain experts. The process included answering questions using the Germinate user interface and comment-based feedback. The feedback requested was about how intuitive the experts found the interface to be and how they thought it could be improved compared with their current way of working.

2.3 | User testing methodology

The online test consisted of a prescreening questionnaire, user tasks, and a follow-up feedback section. The prescreening allowed us to establish the experience users had in their fields. This was important, as we wanted to ensure that testing was carried out with domain experts. The user tasks were developed based on previous work (Shaw et al., 2014). These tasks asked users to undertake standard procedures that allowed them to explore the data sets used during this process. User task questions were marked as either correct or incorrect. The follow up feedback section was divided into two sections: the first of which used attitude-scale questions about user views on the Germinate interface and then subjective open-ended follow up questions to obtain additional information about the user perception of Germinate outside the scope of the user tasks. Users were asked to provide comment-based feedback on how they interacted, what problems they encountered, and finally what they thought was good and bad about the interface. This feedback allowed us to tweak and fine-tune the Germinate interface to better meet the needs of our users and their research requirements. During the user testing process, notes were taken based on observations made about the users' interaction with Germinate. The user testing process took 40 min per user and took the form of a one-on-one session. Testing was undertaken in line with current European General Data Protection Regulation requirements.

2.4 | User testing results

2.4.1 | General background profiling

The 17 domain expert users that undertook this study were broken down into five classifications: six identified as geneticists, two plant breeders, two statisticians, one germplasm manager, and six scientists ranging from lab technicians to department heads. Out of the users, 76.5% were educated to the PhD level. Six users had worked in their area for more than 25 yr, two between 20 and 25 yr, three between 15 and 20 yr, and one between 10 and 15 yr. Five users had fewer than 10 yr experience. With regard to data use, 82.4% of users use genetic resources data in the normal course of their work, 10 users interacted with this sort of data at least on a weekly basis, four on a daily basis, and three on a yearly basis. Fifteen of the 17 users thought there was a current problem in the way genetic resources data are made available to researchers and that improvements were required. Through verbal feedback, it was clear that most users currently interact and manage their data using Microsoft Excel spreadsheets.

2.4.2 | User tasks

Users were asked to respond to 13 questions using the Germinate user interface. These questions were assigned categories and marked as correct or incorrect. The question categories were as follows: unexplained concepts, simple searching, data set statistics, pedigree data, table filtering, advanced searching, genetic map data, export formats, climate data, climate metadata, geographic map data, germplasm collecting sites, and trials data. These categories covered the most commonly used functionality of the user interface.

2.5 | Follow up feedback (attitudinal and open-ended)

2.5.1 | Features that users liked

1. The accession pages were clearly laid out and easy to understand.
2. The common layout on all pages meant finding navigation and options was easy.
3. Location of search bars was convenient, meaning it was easy to search for data.
4. The front page containing additional information on the project was useful and put the data into context.

2.5.2 | Features that users disliked or found confusing

1. It was difficult to use the 'like' search functionality and expected to use '*' for wildcards instead of the standard '%' used in SQL.
2. People found it difficult to use the geographic map features.
3. Some of the page headings such as 'Trial Overview' should be renamed to 'Summary Statistics' to ensure clarity.
4. Locations of some of the data types were not where people expected them to be; for example, data under environment should be under 'data' instead.
5. The number of items in data tables was not obvious.

6. Genetic map identifiers were confusing.
7. Data sets looked like links but are just hover over items to get additional information.
8. Filtering for tables was not obvious.
9. Terminology and naming may not apply to all domains or even across species.
10. It was difficult to see where you are on the accessions page and how far through it you had scrolled.

2.5.3 | New features that users would like to see after using the system

1. Download genotypic data as a matrix
2. Explore data across multiple locations
3. Additional statistical information available on data sets
4. User-provided list for data extraction
5. Exporting genotypic and phenotypic data together
6. Searching across additional fields
7. Images showing phenotypes and states
8. Conditions on data access and use
9. Additional features for climate data

2.5.4 | General comments on using the interface

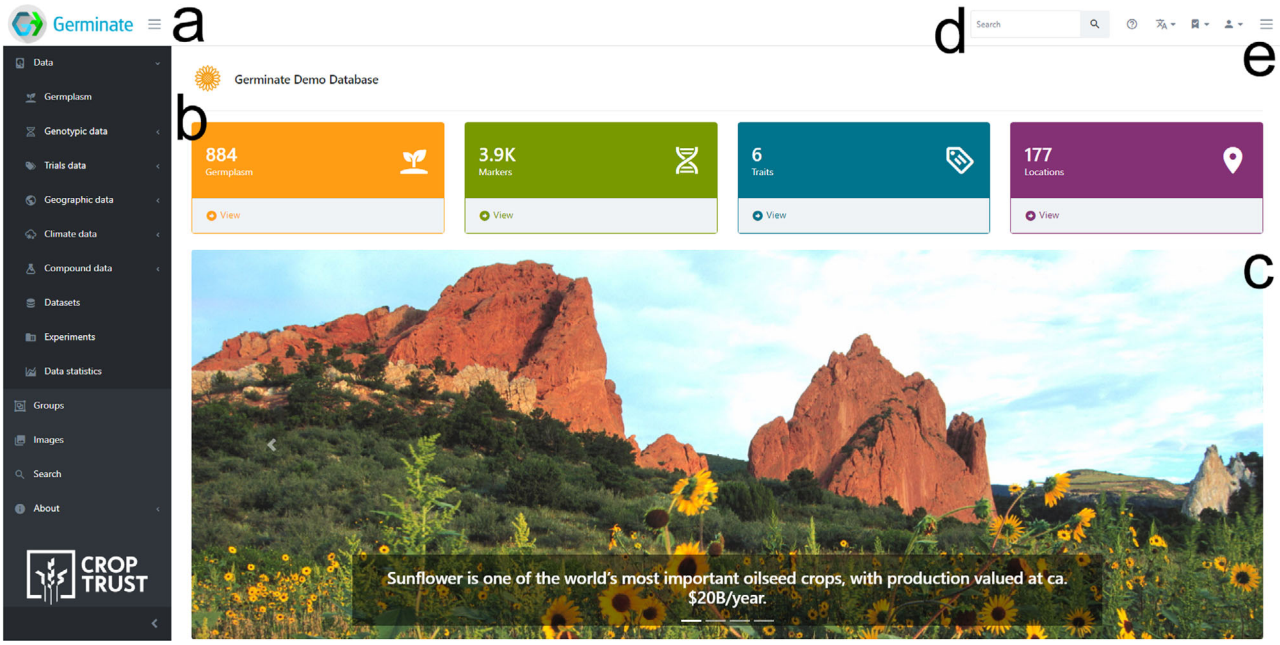
1. Awareness sessions detailing features would be useful for users
2. Graphic representations to show collection breakdown were well received
3. Test the system with data that the user is more familiar with
4. Add links to other databases and information resources
5. Provide alternative coloring options for tables of data and not just red to green gradients
6. Training and familiarization materials for users would be useful
7. More explicit links between Germinate and other systems such as the International Wheat Information System (CIMMYT), Genesys, and EURISCO
8. How are accessions with the same name dealt with?
9. How are errors in data resolved?

2.6 | Updates and new features added to Germinate based on user feedback

The user testing allowed us to prioritize which features to add and what tweaks to make to the Germinate 3.6 interface for the release of the new Germinate platform described here. Taking into account the information we gained through user testing, as well as feedback from domain experts, the following improvements and new features were developed. Figure 3 shows the redesigned user interface of Germinate. While the overall structure remains similar to Germinate 3, additional features have been added for convenience. These include a collapsible sidebar to give more room to the main page as well as a slide-in menu showing progress of asynchronous data import and export jobs.

2.7 | Asynchronous data download

A feature that Germinate 3 lacked was the ability to efficiently work with large data sets, perform complex queries, then finally download the resulting data. This worked well with data sets consisting of thousands of markers and genotypes but not so well with data sets that include millions of genetic markers and hundreds of thousands of plant lines, including large international prebreeding and germplasm categorization projects like the Seeds of Discovery (https://seedsofdiscovery.org/) and the Crop Trust Crop Wild Relatives project. Accessing these large genotypic data sets in Germinate 3 was problematic and slow. Germinate 3 used a naive method of choosing data sets, groups of markers and germplasm, then exporting the intersection of data. On smaller data sets (up to tens of thousands of markers) this operation was quick, with export times up to 20 s (query including data export and transfer time). With larger data sets containing hundreds of thousands of markers and thousands of lines, these queries were taking minutes to perform and presented significant challenges for web-based interfaces. To improve the user experience when performing queries on large data sets, we implemented asynchronous data export, allowing users to choose desired parameters; then after clicking export, they can perform other tasks using the system while their data is being exported. Once the data is ready for download, which may take a few minutes on larger data sets, the user is alerted from within the Germinate interface. This process allows users to set up complex or large exports, perform other tasks, and return when the data is ready for download. Data is held for a system-defined period of time depending on the set up of the platform. Data can only be downloaded by the user if the following conditions are met: the data is (a) fully public or (b) the user is authenticated with Germinate gatekeeper. Germinate gatekeeper controls access to specific databases and data sets based on trusted, authenticated users. This set up provides flexibility where some data sets can be fully public, while others require user authentication to access.



Sunflower is one of the world's most important oilseed crops, with production valued at ca. $20B/year.

FIGURE 3 Germinate interface redesign. This example shows the Germinate sunflower database (https://ics.hutton.ac.uk/cwr/sunflower) but applies to all Germinate installations. (a) The main navigation has been moved to the left-hand side to allow easier exploration. It can be completely collapsed to give room to the main display area. The home page provides an overview of (b) data statistics as well as information about the project supported by this Germinate database, news about the project, and the (c) development of Germinate and related projects. (d) The top menu allows direct access to the global search functionality as well as useful dropdown boxes for supported languages and marked items. (e) There is also another menu that slides in from the right showing the status of asynchronous data export jobs

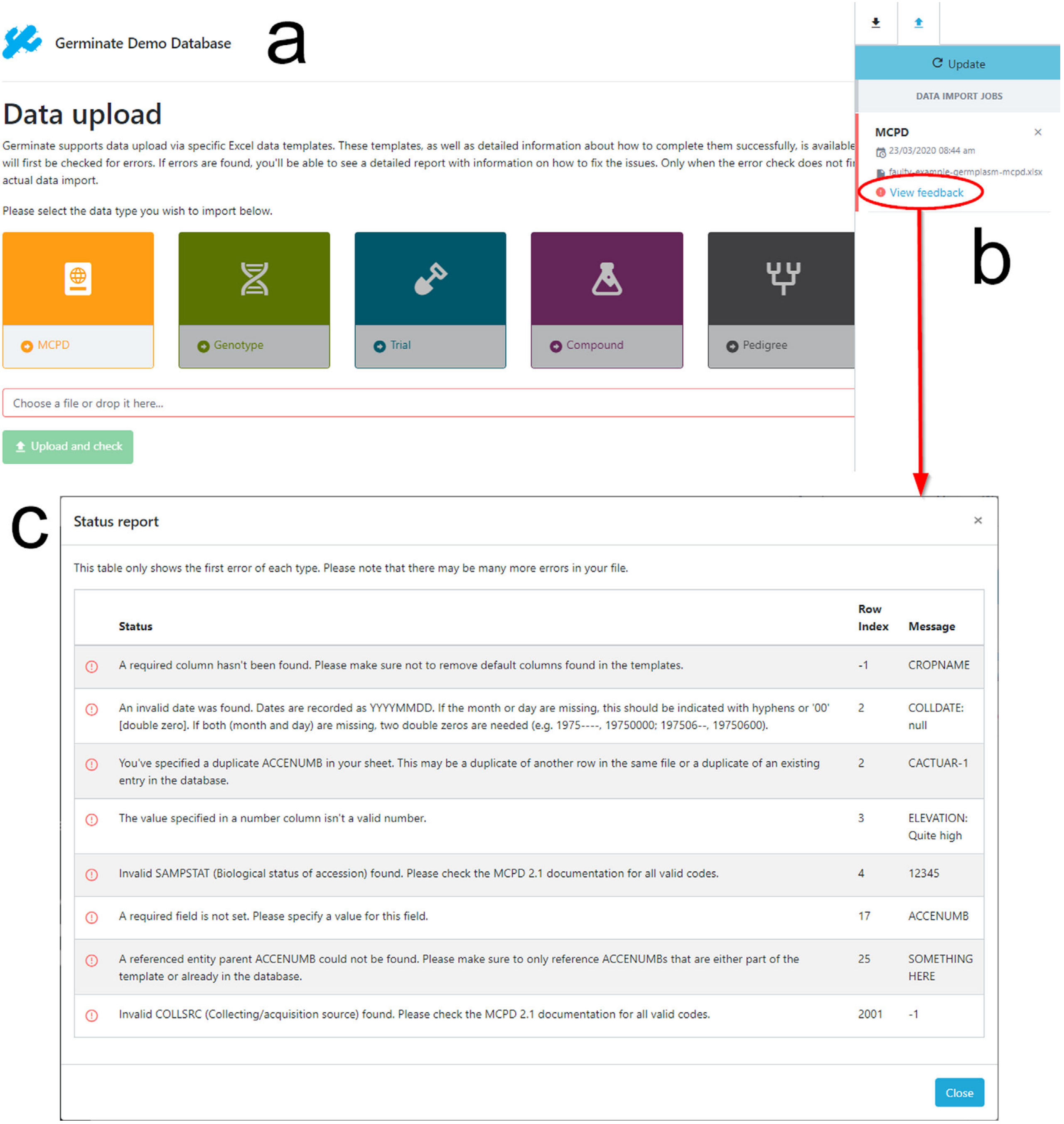
2.8 | Data upload and verification

Working with a diverse set of species and datatypes is challenging. One of these challenges is that not all groups have the same approaches to data standards and handling procedures. In order to standardize this, and to reduce the inevitable and expected data handling burden, a series of standard data templates were created that encapsulated the main data types required for categorization of genetic resources collections including genotypic, phenotypic, chemical compound, pedigree, groups, and passport data. These templates are based on the FAO MCPD standards (Alercia et al., 2015). A standard set of templates can now be provided to project partners along with examples on how these should be completed (https://github.com/germinateplatform/germinate-data-templates). To complement these templates, a series of data upload tools were developed that provide a web-based system for uploading data into Germinate using the standard data templates. These data verification tools do fundamental data sanity checking as well as more complex comparisons against information already stored in Germinate. This ensures that crucial primary identifiers for germplasm are defined in the database. Where these requirements are not met, the users are shown a detailed report listing all issues that need to be addressed before the data can be imported (Figure 4).

The ability to check data import templates against a target Germinate database and only commit data when it has met strict integrity checks is particularly useful when dealing with multiactor projects. Often collaborators and data providers are located worldwide, and data curators may not be familiar with the crop whose data they are curating. As is the case with download features, the data upload is also asynchronous.

2.9 | Hierarchical Data Format storage of genotypic data

The genotypic data export has seen multiple iterations of improved implementations to tackle the challenge of large data sets. The initial Germinate 2 version held the genotypic data in MySQL and exported it to a tab-delimited format on request. This worked well for small data sets but struggled quickly as data volumes increased. As the target data format is a tab-delimited data matrix, the updated method for genotypic data export in Germinate 3 made use of this and stored the data as a plain-text file. This allowed Germinate to scale to hundreds of thousands of markers by hundreds of thousands of genotypes. Germinate would extract only the requested genotypes and markers and write them to a file that could then be downloaded



Germinate Demo Database

Data upload

Germinate supports data upload via specific Excel data templates. These templates, as well as detailed information about how to complete them successfully, is available
will first be checked for errors. If errors are found, you'll be able to see a detailed report with information on how to fix the issues. Only when the error check does not fi
actual data import.

Please select the data type you wish to import below.

Status report

This table only shows the first error of each type. Please note that there may be many more errors in your file.

Status	Row Index	Message
A required column hasn't been found. Please make sure not to remove default columns found in the templates.	-1	CROPNAME
An invalid date was found. Dates are recorded as YYYYMMDD. If the month or day are missing, this should be indicated with hyphens or '00' [double zero]. If both (month and day) are missing, two double zeros are needed (e.g. 1975----, 19750000; 197506--, 19750600).	2	COLLDATE: null
You've specified a duplicate ACCENUMB in your sheet. This may be a duplicate of another row in the same file or a duplicate of an existing entry in the database.	2	CACTUAR-1
The value specified in a number column isn't a valid number.	3	ELEVATION: Quite high
Invalid SAMPSTAT (Biological status of accession) found. Please check the MCPD 2.1 documentation for all valid codes.	4	12345
A required field is not set. Please specify a value for this field.	17	ACCENUMB
A referenced entity parent ACCENUMB could not be found. Please make sure to only reference ACCENUMBs that are either part of the template or already in the database.	25	SOMETHING HERE
Invalid COLLSRC (Collecting/acquisition source) found. Please check the MCPD 2.1 documentation for all valid codes.	2001	-1

FIGURE 4 Data upload and verification. (a) Users with the role of data curator can upload data using the Germinate data templates. (b) These templates are asynchronously checked for validity before any changes are made to the database. (c) The user is then shown a detailed report of any potential issues with the data that require fixing before the template can be accepted and imported into Germinate. If a template file does not contain any errors, a final confirmation is required from the user before the data is imported

by users. While this approach significantly outperformed the initial Germinate 2 implementation, while also reducing disk space requirements, performance improvements would need to be made to scale to millions of markers by hundreds of thousands of genotypes. Germinate has now implemented a new storage approach using the Hierarchical Data Format version 5 (HDF5; https://www.hdfgroup.org). Hierarchical Data Format version 5 is a file format designed to handle large quantities of data while still providing fast query speeds by using efficient and optimized B-Tree indexing (Bayer & McCreight, 1972). The current version of Germinate also encodes the data to further decrease file sizes.

The new Germinate HDF5 approach outperforms the two older implementations in all cases. The runtime improvement factor ranges from 4 to 20 compared with

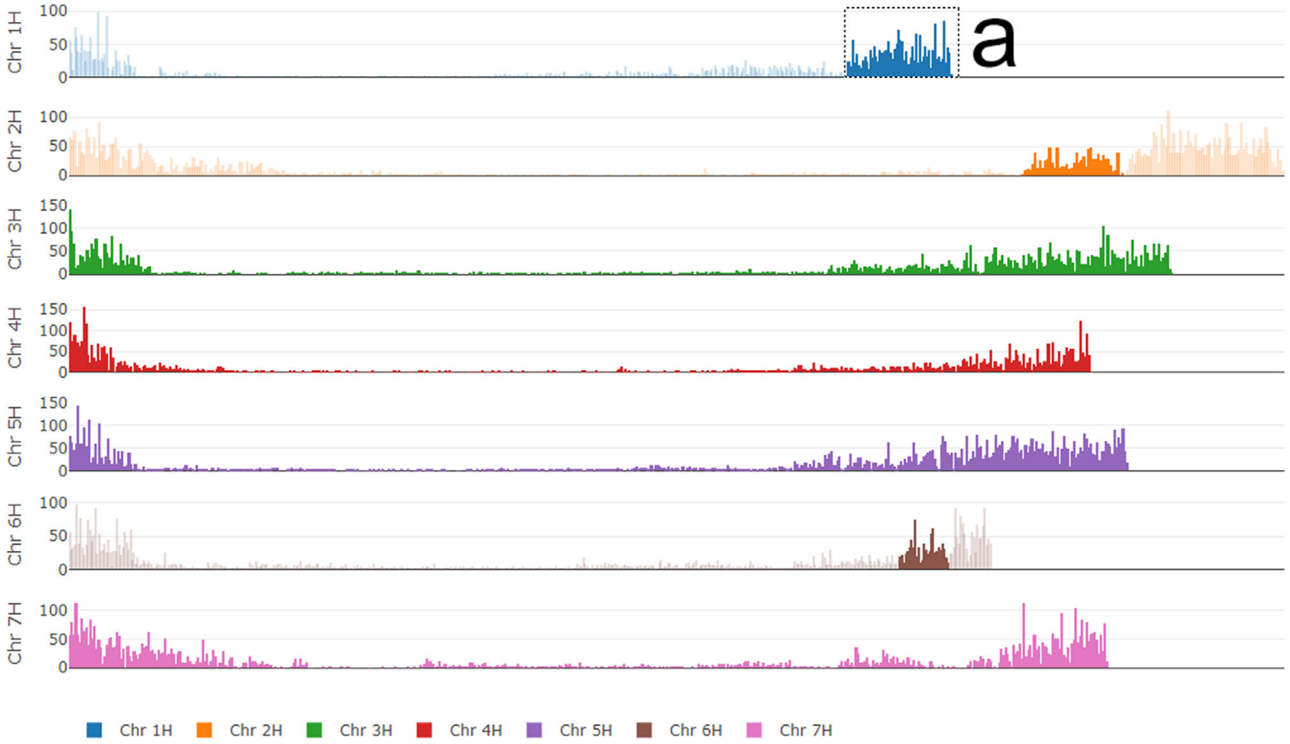


FIGURE 5 Germinate's map visualization and marker selection using barley (*Hordeum vulgare*) 50K SNP markers (Bayer et al., 2017). Germinate plots the marker distribution per linkage group or chromosome to highlight areas of high and low marker density. (a) Users can select areas of interest, which then enables the download of the specified regions or the addition of the markers in those areas to the marked item list. All the charts in Germinate are generated in real time so updates to the database are immediately propagated to the user interface

the flat-file approach and 6–100 compared with the export from the database. This represents a significant improvement when delivering data through the Germinte web platform. The HDF5 format can cope with large numbers of markers very efficiently, whereas with a larger number of lines the runtime is slightly longer. The HDF5 files are approximately 4.5× smaller when dealing with large numbers of markers but up to 2.2× larger for large numbers of lines compared with the plain-text file alternative used in Germinate 3. In most real-world situations, the numbers of markers will significantly outnumber germplasm entries and therefore the slight performance hit for increased line numbers would be uncommon.

While there are global initiatives to develop storage solutions for high-density genotypic data, such as GOBii (http://www.gobii.org) and Gigwa (Sempéré et al., 2019), these tools were either unsuitable or did not currently offer the required maturity of functionality and compatibility with Germinate that was required. While some platforms store binary data in two orientations (markers × genotypes and genotypes × markers), these platforms offer quick access to data but only if the user is subsetting by either axis. Germinate allows the subsetting of data for both axes so a hybrid approach is not appropriate nor does it offer significant benefits. It is our aim to migrate to a standard genotype storage platform such as GOBii should its demonstrated capabilities develop sufficiently to offer us tangible advantages over our current genotypic data storage implementation.

2.10 | Genetic and physical map export and visualization features

The exporting of genotypic data held within Germinate is handled through a standard interface. This interface allows users to export all data for a given map (where a map is a list of markers and their positions, either physical or genetic) by selecting regions across multiple chromosomes visually (Figure 5) and lastly by selection from a set of predefined options. These options include selecting all markers that are bound by user-defined flanking markers, selecting all markers that are bound by flanking markers and include a region of wobble to either side, and lastly by selecting all markers within a defined distance to either side of a specified marker. These more advanced marker selection mechanisms coupled with the close linking of Germinate with the graphical genotyping application Flapjack (Milne et al., 2010) means that selecting single nucleotide polymorphisms (SNPs) around a locus of interest, exporting the data, quickly visualizing allele calls, and returning to Germinate for additional information is seamless and straightforward.

2.11 | Data licenses

Germinate provides tools to allow users to apply licensing or usage restrictions at the data set level. The top level of data compartmentalization in Germinate is an experiment that contains one or more data sets that can be of varying types. In this way, an experiment can contain data sets that have different licensing terms. While this can be restrictive, for example where a user needs to agree to specific license terms in order to gain access to the data, discussions with users has shown that this functionality is seen as an important step—not in restricting access to data but highlighting the origin of data sets where they have been published and how the data providers would like their use to be promoted. Individual data sets can have a specific license associated with them. When a user agrees to the license, they will be granted access without being prompted to accept from that point forward. If Germinate is operating in open-access mode, details will need to be provided each session in order for a user to download data. If a user selects multiple data sets with multiple data access licenses, the user will need to accept each license before being granted access. Data can only be accessed if a license has been accepted. Access to data sets can also be controlled in a more restrictive way where users are required to be authenticated through the Germinate gatekeeper system in order to gain access to data. This provides users with the ability to have protected data sets that may be close to release but require additional data quality checks, ensuring that collaborators have early access to these data sets through the platform.

2.12 | Dublin Core and other metadata standards

The Dublin Core (https://www.dublincore.org) is a set of vocabulary terms used to describe digital resources. It standardizes the way in which data resources are defined by specifying a set of 15 terms ranging from the title and type of a resource to publisher and rights information. Germinate uses the Dublin Core to provide additional information about data sets, where each data set is described by an appropriate subset of Dublin Core terms. This information is made available through the Germinate web interface and can be downloaded alongside the data contained in a data set. Additional information on plant lines can also be defined as attributes on the main Germinate germplasm page. This allows for information that does not comply with MCPD or Germinate core terms to be added and be searchable from the Germinate interface.

2.13 | Germinate BrAPI implementation

A BrAPI implementation has been developed for Germinate, allowing database instances to communicate with other BrAPI-compliant databases and client applications. The BrAPI is a community-driven effort to design a representational state transfer standard for plant phenotype and genotype databases to communicate with each other and additional visualization and analysis tools. The BrAPI is compatible with the key data standards Germinate implements such as the MCPD.

The BrAPI 2.0 standard, released March 2020 (https://github.com/plantbreeding/API/releases/tag/V2.0), splits the API into four modules: core, phenotyping, genotyping, and germplasm. The BrAPI-compliant services are permitted to implement as much, or as little, of the specification as is suitable for their use cases, and Germinate's implementation has focused on the new Genotyping module that has been modeled on the Global Alliance for Genomics and Health Variants API (https://ga4gh-schemas.readthedocs.io/en/latest/api/variants.html). This part of the BrAPI specification is used by Flapjack to pull genotypic data from Germinate without using the Germinate user interface, therefore reducing the need to switch back and forth between applications. Germinate's BrAPI service is currently under development and will be expanded to implement the API calls from the other three BrAPI modules to enable data exchange with a wider range of plant breeding data clients as BrAPI matures and develops.

2.14 | Trials and phenotypic data visualization

To meet the demands of increased data volumes, Germinate's visualization tools for trials data have been significantly improved. To gain a better understanding of the performance of materials that have been evaluated in trials, users can select one or more data sets to explore from the trials browser page. A series of graphs can be created showing relationships between different traits using the interactive scatterplot matrix tool (Figure 6). Although this tool was present in the previous Germinate 3, new features have been added to improve the display and performance and to enable users to better connect the data points present in the chart to other data available within the platform. These new features include the ability to interact with charts using lasso or single-point selection

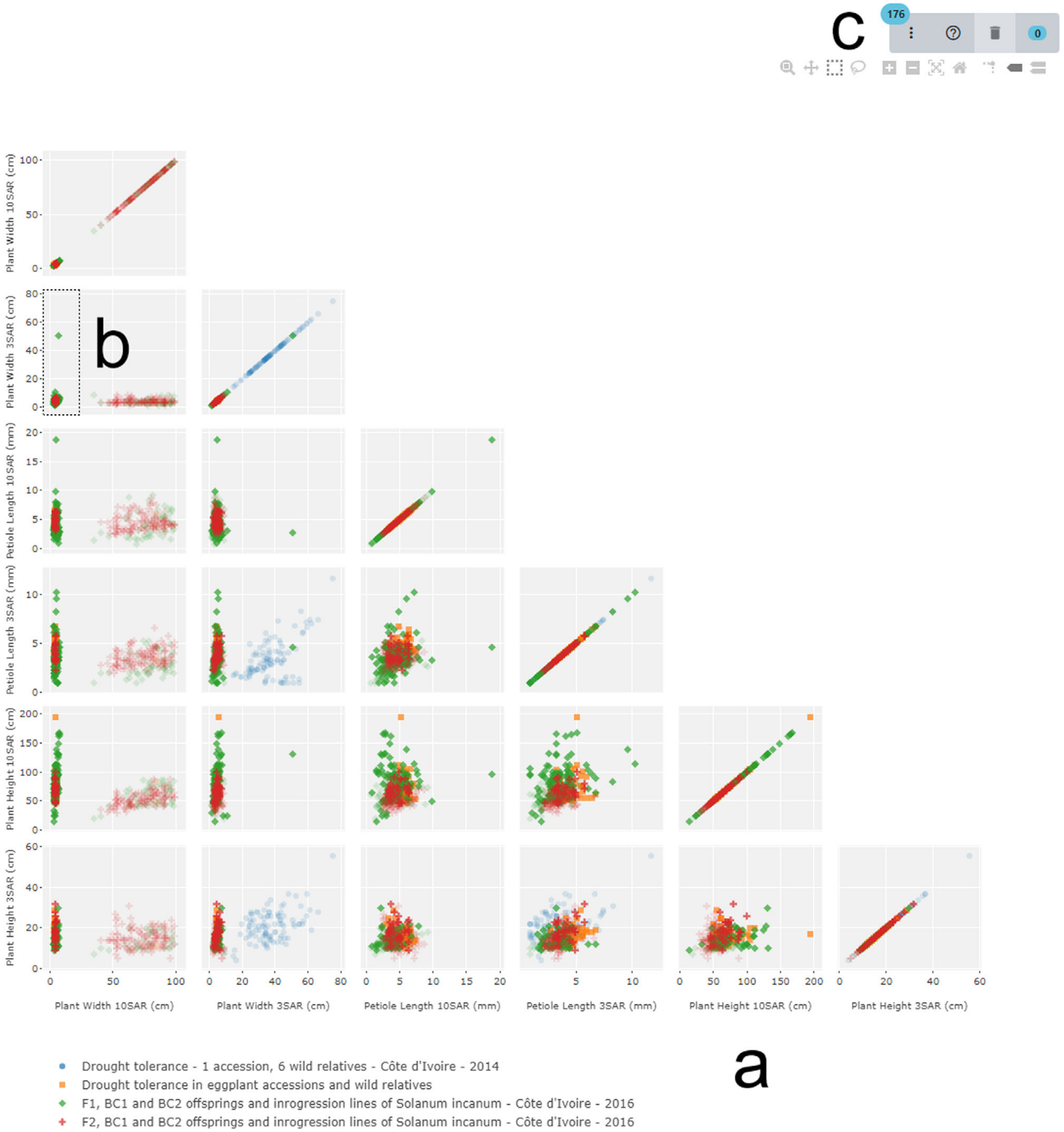


FIGURE 6 Trials data visualization in eggplant. Germinate offers various ways of exploring and visualizing data. This chart is a matrix of scatter plots where each individual cell represents a scatter plot between two traits. Trait names and measurement units are shown along the axes. (a) The coloring is based on the source data set, but the chart can also be colored by germplasm group, treatment or trial site. This kind of chart accentuates correlations between traits but also between data sets or selected germplasm groups. Most notably, in this example, there is a significant difference in plant width 10SAR (weeks post planting) (first column) between the four data sets. (b) Selections can be made in the chart using rectangular or lasso selection modes. (c) Germplasm selected in this way is highlighted in all subplots and can easily be added or removed from the marked item list. In this example, a selection of 176 eggplant genotypes has been made (data points within the selection rectangle). The same kind of chart is available for chemical compound and climate data cross data set comparisons can be made where trait names are identical. We hope to develop this further to include phenotype equivalence based on ontologies

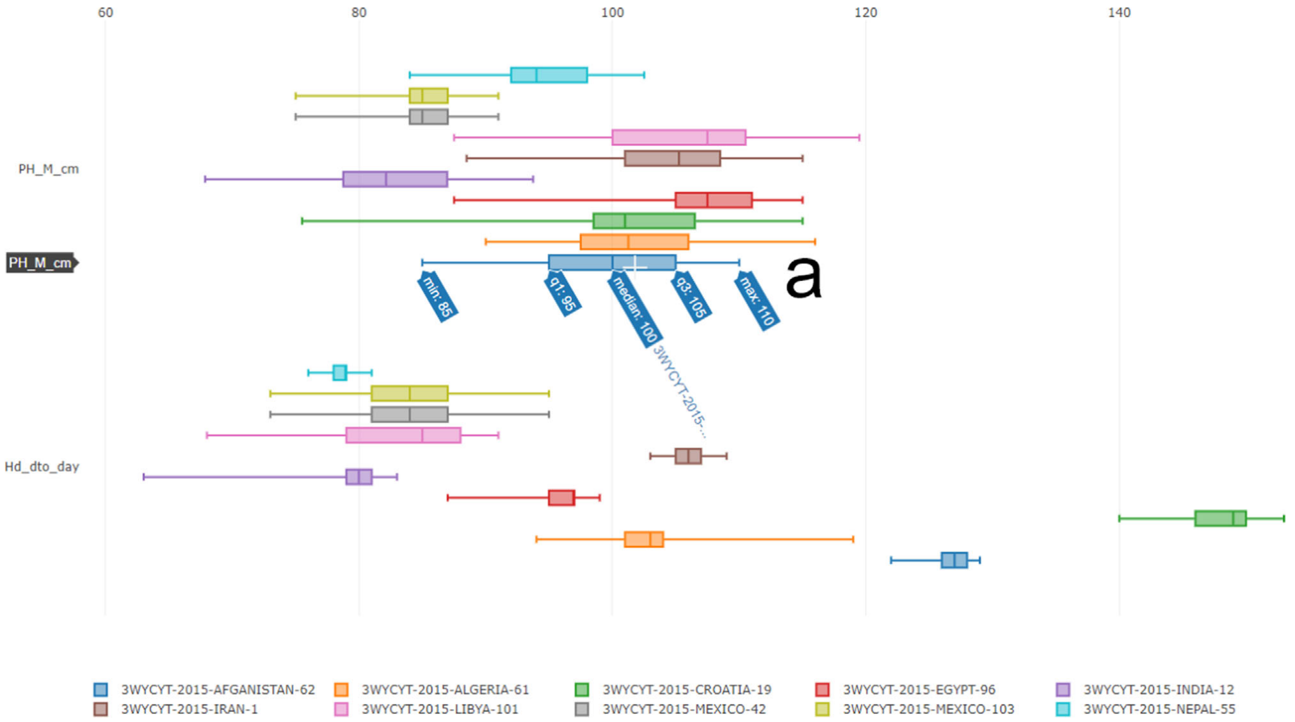


FIGURE 7 Interactive trait box plots. Box plots can be used to compare essential statistics between different data sources. Germinate can generate box plots for a user-defined selection of traits across either all the data or selected germplasm. (a) The chart shows the minimum, maximum, as well as all quartiles per data set and trait. The *y*-axis is broken down by selected traits. Within each band, there is a box plot for each data set, which allows comparison of different sets of data. In this example using CIMMYT's Germinate wheat database, there are two traits: plant height (PH) and heading date (Hd). Within each trait block, the individual box plots represent data sets in which this trait has been scored. Data obtained from Global Wheat Program; IWIN Collaborators; Reynolds, Matthew; Payne, Thomas. 3rd Wheat Yield Collaboration Yield Trial: 3WYCYT, Cycle 2015. CIMMYT Germinate Wheat. 2020. http://germinate.cimmyt.org/wheat

mechanisms to offer interactive ways to create groups of related germplasm. When lines are selected from within charts, related data is highlighted within linked plots and the lines can then be easily added to what is called 'lists'. These lists can then be converted into permanent groups that can be shared with other users or used to temporarily export data from any of the Germinate export pages and tools available into external analysis tools such as Flapjack.

A new dedicated traits page has been added to Germinate to offer a central location containing detailed information about a selected trait. This includes exemplar images of phenotypes, a trait data table showing all data points for this specific trait, a table showing all data sets this trait has been scored in, and a box plot visualization of the trait data per data set (Figure 7).

2.15 | Marked lists and groups

The groups functionality of the Germinate 3 platform has been improved and developed to allow the storage of two distinct group types for germplasm, location, and marker data: persistent groupings and temporary markings. Persistent groups are stored in the Germinate database, can have access permissions applied to them, and therefore can be shared with other users. Persistent groups can only be created by authenticated users or pregenerated by database administrators. The temporary marked items are user-specific and are stored by the user's browser. While the groups system was already part of Germinate 3, marked item lists have received extensive development and can now be used in all places that were previously reserved for groups. Groups and lists can be used to export selected data or for chart coloring to highlight and compare data of interest. While exploring data using either data tables or charts, germplasm, locations, and markers can dynamically be added to and removed from groups and lists. Figure 8b shows selection boxes for germplasm and marker groups during the genotypic data export as well as germplasm for phenotypic data export. The items below the horizontal line are persistent groups, while the items above the line represents marked items on the user-specific list. Selections can be made in these selection boxes and only the items within the selected groups or lists are exported.

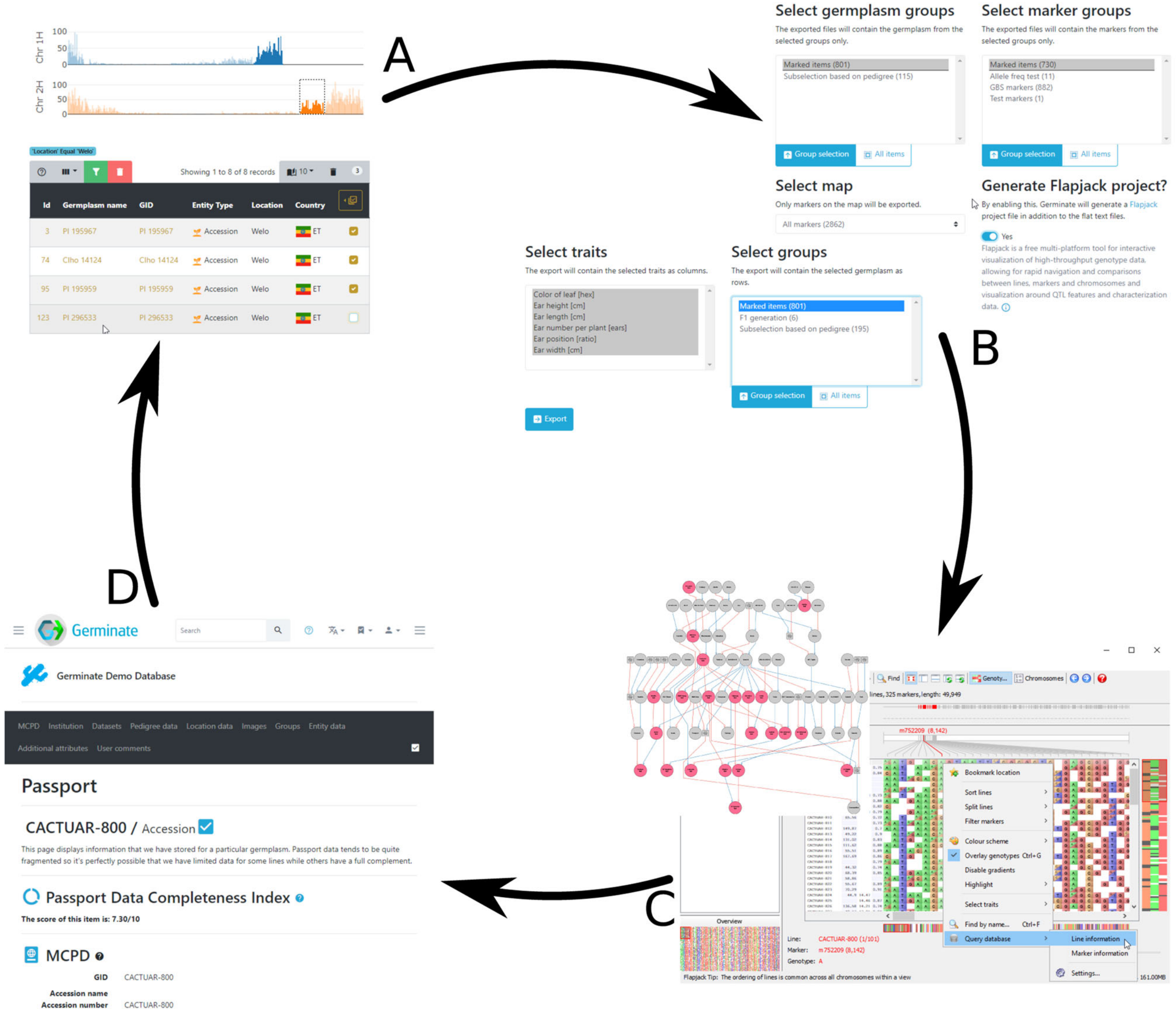

A
B
C
D
Chr 1H
Chr 2H
100
50
0
'Location' Equal 'Welo'
Showing 1 to 8 of 8 records
Id
Germplasm name
GID
Entity Type
Location
Country
3 PI 195967 PI 195967 Accession Welo ET
74 Clho 14124 Clho 14124 Accession Welo ET
95 PI 195959 PI 195959 Accession Welo ET
123 PI 296533 PI 296533 Accession Welo ET
Select germplasm groups
The exported files will contain the germplasm from the selected groups only.
Marked items (801)
Subselection based on pedigree (115)
Group selection
All items
Select marker groups
The exported files will contain the markers from the selected groups only.
Marked items (730)
Allele freq test (11)
GBS markers (882)
Test markers (1)
Select map
Only markers on the map will be exported.
All markers (2862)
Generate Flapjack project?
By enabling this, Germinate will generate a Flapjack project file in addition to the flat text files.
Yes
Flapjack is a free multi-platform tool for interactive visualization of high-throughput genotype data, allowing for rapid navigation and comparisons between lines, markers and chromosomes and visualization around QTL features and characterization data.
Select traits
The export will contain the selected traits as columns.
Color of leaf [hex]
Ear height [cm]
Ear length [cm]
Ear number per plant [ears]
Ear position [ratio]
Ear width [cm]
Select groups
The export will contain the selected germplasm as rows.
Marked items (801)
F1 generation (6)
Subselection based on pedigree (195)
Export
Bookmark location
Sort lines
Split lines
Filter markers
Colour scheme
Overlay genotypes Ctrl+G
Disable gradients
Highlight
Select traits
Find by name... Ctrl+F
Query database
Line information
Marker information
Settings...
Line: CACTUAR-800 (1/101)
Marker: m752209 (8,142)
Genotype: A
Flapjack Tip: The ordering of lines is common across all chromosomes within a view
Germinate
Search
Germinate Demo Database
MCPD Institution Datasets Pedigree data Location data Images Groups Entity data
Additional attributes User comments
Passport
CACTUAR-800 / Accession
This page displays information that we have stored for a particular germplasm. Passport data tends to be quite fragmented so it's perfectly possible that we have limited data for some lines while others have a full complement.
Passport Data Completeness Index
The score of this item is: 7.30/10
MCPD
GID CACTUAR-800
Accession name
Accession number CACTUAR-800

FIGURE 8 Data export selection based on groups from the Germinate demo database. Every data export page and every chart by default exports all available data within the selected data sets. This can be restricted in both dimensions, for example, germplasm and markers for genotypic data, to reduce the exported data volume to only the data that is of interest to the user. (a) Specific germplasm, markers, and locations can be added to marked item lists using visualizations and table filtering. (b) A combination of groups or marked items for either dimension can be selected resulting in an exported file that contains just elements from these selected groups. If a user, for example, has selected a few hundred germplasm lines and a few thousand markers, these selections can be used to export genotypic data against them. It is no longer required to make the marked item selection persistent by creating a group before using the export functionalities, which is a big step forward in terms of usability. In this example, genotypic and phenotypic data are exported and loaded into Flapjack. (c) Users can then jump back to Germinate to get additional information about germplasm and markers. (d) User selections for marked item lists can then further be adjusted to refine the selection and export data of the same type or across types

2.16 | Docker

To facilitate access to Germinate by a wide audience, we have created a docker-based instance of the platform (https://www.docker.com). This allows both prospective users and those who want to run Germinate from within a containerized environment to quickly deploy a working Germinate system to test with their own data and removes the need for manual compilation of the source code as was required for Germinate 3. It reduces the barrier to entry for the system and allows a clean install of Germinate to be quickly deployed to local or cloud-based hosting. Detailed instructions on how to use the docker image are available on GitHub (https://germinateplatform.github.io/germinate-server/setup.html#docker).

2.17 | Germinate deployment to support the Crop Trust Crop Wild Relatives project

Germinate has been used to store experimental prebreeding data for 14 of the crops that form part of the Crop Trust Crop Wild Relatives project (Table 1). Examples of their use of the Germinate platform in the distribution of prebreeding data are described here. Each of these projects is different in its aims and the data it generates but draws upon the common development goals of Germinate in order to ensure data is made publicly available to the community in a standard way. A full list of the species available under this project is available from https://ics.hutton.ac.uk/get-germinate. New data sets will be added to these resources as they become available.

2.18 | Germinate alfalfa

The crop wild relative (CWR), drought-tolerant alfalfa (*Medicago sativa* L.) project has developed a cohort of wild accessions and prebreeding lines with adaptation to different types of drought stress that include tolerance to winter freezing. The Germinate database, which can be accessed from https://ics.hutton.ac.uk/cwr/alfalfa, contains preliminary passport characterization and evaluation data (flower color, habit, pod coiling, height, spring forage yield, and seed size) for 434 CWR accessions. From this material, 47 new prebreeding lines have been developed targeting adaptation to warm- and cold-temperate environments. A full description of their parents and taxonomy is provided within Germinate. The prebreeding lines include 13 new hybrids that were developed between *M. sativa*, *M. arborea* L., and *M. truncatula* Gaertn. The Germinate database contains two generations of phenotypic evaluation of the *M. arborea* × *M. sativa* hybrid cultivar Alborea-101 (project identifier CTA018), which was donated to the project by Edwin Bingham (Bingham, Armour, Irwin, Jayaraman, & Ané, 2009). A statistical summary of this data shows that 25% of CTA018 plants had 20% greater forage yield scores than the local nondormant ‘SARDI 10 Series 2′. This population is also very diverse for seed size, plant height, and flower color.

The CWR alfalfa cohort has been evaluated in Inner Mongolia, China, northern and southern Kazakhstan, southeastern Australia, and central and Patagonian Chile. Data presented through Germinate includes measurements of establishment and persistence over time (survival), forage yield, and plant height (used to assess fall dormancy; Teuber et al., 2004). A basic statistical assessment of the data is presented using spatial analysis (restricted maximum likelihood; Genstat 20) for each individual trial site. A highlight of the results has been the excellent performance of the Chinese cultivar Zhongcao No.3 at Hohhot and Siziwang in China and Kokshetau in northern Kazakhstan. This variety has introgressions of CWR *M. sativa* subsp. *falcata* (L.) Arcang. (details of which can be traced back from the passport data held in Germinate) in its breeding and will contribute strongly to the future release of a new cultivar for central and northern Kazakhstan. This variety is also part of a seed-sharing scheme developed to improve seed distribution and adoption of alfalfa by smallholder farmers in Inner Mongolia. Alfalfa seed is also being distributed to smallholder farmers in central Chile, this time with the Australian cultivar SARDI Grazer, which was developed for tolerance to persistent grazing in low-rainfall environments.

The Alborea-101 hybrids have excellent potential to increase forage yields in both of these countries as demonstrated by experiments comparing the performance of CWR lines in both rainfed and semi-irrigated conditions in Adelaide, Australia and Cauquenes, Chile. Data from the Cauquenes experiment in Germinate includes physiological measurements of normalized difference vegetation index (Greenseeker), leaf area index (Ceptometer), and gas exchange (Li-Cor).

The Germinate alfalfa database also captures photographs of plants and people from the alfalfa Crop Trust project. It is hoped that the images will provide an insight into the environments assessed and the smallholder cooperating farmers that we hope to assist in this project. The data described here, which is held in Germinate, can be freely accessed (https://ics.hutton.ac.uk/cwr/alfalfa).

2.19 | Germinate barley

Barley is a multipurpose crop used for livestock feed, human food, and beverages and agronomically tends to show good resilience to adverse climate change. All over central and western Asia and the northern Africa region, barley is considered by most farmers to be a risk aversion crop needing limited inputs and being able to grow under harsher conditions than wheat. Breeding productive varieties resistant to major diseases, tolerant to drought and heat, and having the required quality attributes is a very appropriate strategy for enhancing barley production and use. Among CGIAR centers, ICARDA has the global mandate for the improvement of barley and holds in trust one of the most important and unique collections of *Hordeum* genetic resources. The collection has 32,783 accessions including 2,384 accessions of wild *Hordeum* species.

The project aims to strengthen prebreeding efforts by mobilizing novel diversity from a set of wild species accessions using the primary gene pool (*H. vulgare* L. subsp. *spontaneum* (K. Koch) Thell.) and the secondary gene pool

(*H. bulbosum* L.) to improve the performance and quality of cultivated barley. Resistance to major diseases, tolerance to drought and heat, and improvement of beta glucan content are the key traits targeted in interspecific crosses and selection in subsequent generations. A set of 117 accessions of *H. spontaneum* have been characterized and used as parental germplasm in crosses over the past four seasons. Crosses involving *H. bulbosum* in collaboration with our Institut national de la recherche agronomique–Morocco partner were also successful and evaluated together with introgression lines derived from crosses of cultivated barley with *H. bulbosum* (supplied by The Nordic Genetic Resource Centre). The first batch of advanced lines are ready for more evaluation and distribution to breeding programs at and outside ICARDA.

The phenotype data for the source accessions, passport data, and the derived prebreeding lines is stored in Germinate. Furthermore, the project helped with the genotyping of a large number of accessions of barley including CWRs. A total of 1,880 landraces and 411 accessions of wild *Hordeum* accessions were genotyped by the project partner The Leibniz Institute of Plant Genetics and Crop Plant Research–Germany using the genotyping-by-sequencing (GBS) approach. Germinate will enable cross-referencing to the accessions used in the prebreeding work.

2.20 | Germinate chickpea

The chickpea (*Cicer arietinum* L.) instance of Germinate is based on the de novo collection of 371 wild *Cicer* accessions as seed and 839 accessions as DNA. Single seed descent immortalized the living resource, which was deposited to genebanks of the multilateral system at the Aegean Agricultural Research Institute (Turkey), ICARDA, and the Australian National Genebank. Metadata in Germinate chickpea provides the respective genebank accessions numbers, along with a range of relevant data on genomics, traits, and habitats.

The wild chickpea Germinate data includes DNA-based collections because this data type offers the opportunity to understand genetic diversity and its ecological drivers without the burden of curation. Importantly, knowledge of genomic and ecological diversity can facilitate construction of suitably diverse living collections for trait analysis and breeding, and it can guide exploration of new collection sites to fill gaps in the living collection. The living collection of chickpea CWRs is dominated by the crop's immediate wild progenitor *Cicer reticulatum* Ladiz. (∼10,000 yr ago diverged) and the closest outgroup species *Cicer echinospermum* P.H. Davis (∼110,000 yr ago diverged). Species from the tertiary gene pool, *C. bijugum* Rech. f. and *C. pinnatifidum* Jaub. & Spach, represent minor fractions and are primarily available as DNA.

The collection encompasses the majority of the wild species' known and accessible native range in southeastern Turkey. As described in Von Wettberg et al. (2018), 985 accessions were genotyped using a GBS approach, revealing 12 genetic populations among *C. recitulatum* and *C. echinospermum* accessions, all of which are represented in Germinate chickpea. Although admixture was detected between wild species, gene flow was not detected between wild and cultivated genomes. During collection, all wild accessions were georeferenced. The ecological breadth of the collection was documented by collection of extensive plant–microsite soil chemistry and microclimate data in addition to the culture-independent sequencing of >800 wild plant microbiome samples (partially described in Greenlon et al., 2019). *Cicer recitulatum* and *C. echinospermum* are distinguished by their origin soil types (sandstone or limestone vs. basaltic, respectively). Taken together, these data feature the ecological and biological context of wild collection sites.

Genomic, population, genetic, and environment data were used to select 26 accessions of *C. reticulatum* and *C. echinospermum* that span the diversity of the collection. Analysis of whole-genome sequences of these wild accessions, along with a global collection 36 modern elite chickpea accessions, revealed a 97.4% reduction in genetic diversity (Von Wettberg et al., 2018). This analysis documents the vastness of diversity captured in the wild relative collection. Genomic data are linked through the Germinate database to BioProject PRJNA353637 at the National Center for Biotechnology Information.

These same 26 wild accessions were crossed into seven cultivated accessions that represent the primary cultivated environments of the crop. Approximately 15,000 independent recombinant inbred lineages were generated by multiple partners, increasing the diversity of the cultivated gene pool by ∼44-fold. Geminate contains data on recombinant inbred lines developed from four of these cultivated accessions crossed into these 26 wild parents. Genotype data on ∼2,300 F_2 lineages from a single cultivated parent provide opportunities for trait genetic analysis (Shin et al., 2019). Among 9,700 F_4 recombinant inbred lineages are 3,700 F_4–derived sibling pairs. These sibling pairs comprise a collection of semi-isogenic lineages in which ∼80% of the genome is fixed and similar between paired lineages and ∼20% is fixed and different. The structure of this material will Mendelize many polygenic traits between pairs and facilitate trait analyses by reducing heterogeneity of genome backgrounds.

Germinate chickpea provides access to phenotyping data for traits that are high priorities for crop improvement, including flowering time and days to maturity, seed traits, plant architecture, seed shattering, biomass, and seed yield. Not currently incorporated into Germinate chickpea are data on resistance to *Fusarium* wilt, *Ascochyta* blight, and pod borer; nitrogen fixation effectiveness; heat and drought tolerance; and aluminum tolerance.

2.21 | Germinate cowpea

Cowpea [*Vigna unguiculata* (L.) Walp.] is primarily a self-pollinated crop, and its genetic base is reported as narrow. In their quest to develop improved, highly productive uniform varieties, breeders use elite lines as parents in their crossing programs, thereby inadvertently contributing to the reduction of genetic variation in cowpea. As in the case of other crops, cowpea wild relatives offer important genetic diversity that can widen the genetic base of the crop. Through this project, whose data is available from Germinate (https://ics.hutton.ac.uk/cwr/cowpea), sources of tolerance or resistance to drought, heat, and aphid were identified and are being exploited in the genetic improvement of cowpea. Elite lines from the International Institute of Tropical Agriculture (Nigeria) and three National Agricultural Research System breeding programs (INERA of Burkina Faso, INRAN of Niger, and NACGRAB of Nigeria) were crossed to the wild relatives identified as sources of drought and heat tolerance and aphid resistance. Several hundreds of lines derived from 74 crosses constitute the foundations of strong prebreeding programs in these institutes to widen the genetic diversity of cowpea. These lines are sources of novel traits for cowpea genetic improvement under the challenging conditions of climate change. With support from the Crop Trust, International Institute of Tropical Agriculture has worked to develop cowpea informatics resources using Germinate as the data-sharing platform. This will ensure the long-term unrestricted access to data and provides a valuable resource for the cowpea community.

2.22 | Germinate durum wheat

Durum wheat [*Triticum turgidum* L. subsp. *Durum* (Desf.) van Slageren] is a traditional crop of the Mediterranean region used around the world for making pasta, couscous, and bulgur. Because durum wheat is often appreciated for its protein content, and it is deemed as more rustic than common wheat, it is often sown late in the season and in marginal lands. That exposes it to severe terminal droughts and heat stresses. ICARDA's genebank holds over 43,000 wheat accessions (>20,000 of durum wheat) including one of the largest global collections of its wild relatives. The partnership of genebank and breeders has resulted in strong utilization of these resources, with today over 12 varieties released around the world derived directly from CWRs.

The project Dissemination of Interspecific ICARDA Varieties and Elites through Participatory Research (DIIVA-PR; https://mel.cgiar.org/projects/741) aims at evaluating and further developing several CWR-derived elite lines for their response to several biotic and abiotic stresses. In addition, a critical component is the inclusion of over 20 on-farm tests with the aim of capturing their suitability for commercialization and appreciation or criticism by their final users. The multilocations testing of 24 CWR-derived elite lines of durum wheat exhibited >20% yield advantage over the national commercial check, better quality characteristics and disease resistance, and high nutritional value. Additionally, strong farmer appreciation was achieved for the new CWR-derived lines.

All raw data including numerous agronomic parameters (plant height, yield, and yield components), phenological traits (flowering time, physiological maturity), disease resistance, and nutritional quality for on-station trials across Morocco and Lebanon, as well on-farm trials, will soon be available from the Germinate durum wheat database (https://ics.hutton.ac.uk/cwr/wheat). This will provide free access and a significant resource for the scientific community in order to select superior germplasm well-adapted to severe climatic conditions.

2.23 | Germinate eggplant

Eggplant (*Solanum melongena* L.), also known as aubergine or brinjal, is a crop species of global importance and an important source of antioxidants, vitamins, and minerals (Raigón, Prohens, Muñoz-Falcón, & Nuez, 2008; Gramazio et al., 2014). Eggplant arose in Africa and was dispersed throughout the Middle East to Asia (Weese & Bohs, 2010; Knapp, Aubriot, & Prohens, 2019) and has several centers of domestication across Asia (Cericola et al., 2013, Meyer, Karol, Little, Nee, & Litt, 2012). Crop wild relatives of eggplant have been investigated as a source of genes for yield increase, fruit quality, disease resistance, or nutritional content. The Crop Trust Crop Wild Relatives projects have generated data on cultivated eggplant, 15 CWR species, and on interspecific hybrid backcross populations between *S. melongena* and seven wild species including *S. insanum* L., *S. incanum* L., *S. dasyphyllum* Schumach., *S. anguivi* Lam., *S. lichtensteinii* Willd., *S. torvum* Sw., and *S. lidii* Sunding.

The eggplant Germinate database (https://ics.hutton.ac.uk/cwr/eggplant) contains data sets from two project phases. From Phase 1, it contains 35 data sets (30 of phenotyping trials and five of genotypic data), encompassing 921 accessions of cultivated eggplant and CWRs, interspecific hybrids, and advanced backcross materials including introgression lines (ILs). In total, data from 18 accessions from 12 eggplant CWR species from the primary, secondary, and tertiary gene pools and derived materials after hybridization and backcrossing with seven accessions of eggplant are included. Over 33,500 data points are included in the database. Phase 2 has produced 24 data sets, including data from 17 phenotypic trials and seven genotype data sets.

Germinate is the repository for phenotyping trial data corresponding to field, greenhouse, and climatic chamber trials for characterization of morphological and agronomic traits, cross-compatibility and pollen fertility, seed yield and germination, tolerance to drought, including both morphological and biochemical data, resistance or tolerance to bacterial wilt, *Sclerotium rolfsii,* nematodes, spider mites and whitefly, as well as fruit composition. These diverse data reveal that eggplant is cross-compatible with a broad range of wild species and heterosis for vigor in many interspecific hybrids (Kaushik, Prohens, Vilanova, Gramazio, & Plazas, 2016). Characterization and evaluation data for tolerance to drought using different methodologies, including open-field trials under natural and artificial drought conditions and controlled experiments in greenhouse and climatic chambers, have been produced that are able to be stored in Germinate using the phenotype fields. Basic analysis is available from the tools in the application showing, for example in the case of fruit composition data, that eggplant CWRs generally had higher contents of chlorogenic acid, a phenolic acid with beneficial properties for human health, than cultivated eggplant while interspecific hybrids were intermediate or similar in contents to the wild parents (Kaushik et al., 2017).

Genotyping data, including high-throughput SNP genotyping data of cultivated and CWR, of early backcross materials of *S. incanum* and *S. insanum* are also available for storage in Germinate and will be available for marker-assisted selection in the future development of ILs as well as of already obtained introgression lines (Gramazio et al., 2017). Genotypic data of 45 ILs with *S. incanum* covering 71.7% of the genome of *S. incanum* in addition to further four BC2 or BC3 introgression lines of *S. incanum*, *S. dasyphyllum*, and *S. elaeagnifolium* are available in the Germinate database. These genotyping data have been useful to establish relationships between eggplant and its CWRs, for reducing the number of generations needed to develop ILs, and for fine genetic characterization of ILs, facilitating the detection of quantitative trait loci and the future development of sub–ILs. In this respect, markers for traits of interest in eggplant, such as prickliness, have been located through the genotyping of segregating backcross generations and ILs.

Overall, the data included so far in the eggplant Germinate database represent a dramatic increase on the information available on phenotypic and genotypic characterization of eggplant relatives, interspecific hybrids with eggplant, backcross generations, and ILs that is of great relevance for eggplant breeders. Further data from recent and ongoing experiments will complete the broad array of data available through the Germinate platform.

2.24 | Germinate finger millet

The data for the CWR finger millet (*Eleusine* spp.) project implemented in Kenya by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)–Eastern and Southern Africa in collaboration with Maseno University and Kenya Agricultural and Livestock Research Institute (KALRO)–Kisii has been deposited into Germinate and is freely accessible (https://ics.hutton.ac.uk/cwr/fingermillet). In this project we generated field and screenhouse data from both wild and cultivated finger millet for climate change related traits in order to establish the potential of using wild relatives as novel sources of traits of interest. The three main traits of interest were response of finger millet to *Striga hermonthica*, blast disease (*Magnaporthe grisea*), and drought. *Striga,* also called witch weed, is a parasitic weed to many cereals including maize, sorghum [*Sorghum bicolor* (L.) Moench], and rice (*Oryza sativa* L.). The mechanism of response to *Striga* remains unknown in finger millet, where research began only recently. In the current study, we generated three seasons' data from two hotspot *Striga* locations—Alupe and Kibos in western Kenya—using both wild and cultivated germplasm. The data was collected on the number of germinating *Striga* per plot every 2 wk until maturity. To avoid building the *Striga* seed bank in the fields and increasing the number of weedy wild finger millet in the environment the experiments were done in pots and each plant bagged at the flowering stage.

Data on response to blast disease was collected both under screenhouse and field conditions at an interval of 2 wk from 5 wk after germination until maturity and has been uploaded into the Germinate finger millet database. Since blast is most severe at maturity, we also ranked the genotypes depending on their blast scores at maturity. The drought experiment was undertaken in Kiboko, Kenya. The drought trials were given supplementary irrigation until 50% of the plots attained 50% flowering then the water was withdrawn. Data on agronomic traits (plant aspect, plant height, yield, and yield-associated traits) and traits associated with drought tolerance and escape (number of green leaves at physiological maturity and days to 50% flowering) were also taken and uploaded into the database.

The Germinate finger millet database will be expanded during 2020–2022 with the inclusion of new data generated under the Templeton Charity Foundation Inc. grant 'Safeguarding crop diversity for food security: Prebreeding complemented with Innovative Finance,' which will also allow us to develop a finger millet community and central hub for all finger millet news and research.

2.25 | Germinate grasspea

Grasspea (*Lathyrus sativus* L.) is a climate-resilient, nutrient-rich legume crop grown mainly in fragile agroecosystems in dry areas. Despite many desirable features, this crop is underutilized because of the presence of a plant toxin called ODAP (β-N-oxalyl-l-α, β diamino propionic acid) in its seeds and susceptibility to the parasitic weed bean broomrape (*Orobanche crenata* Forssk.). Within the cultivated gene pool, ODAP-free and *Orobanche*-resistant germplasm is not reported. The ICARDA genebank holds 4,450 accessions of grasspea including 1,555 accessions representing 45 wild species. In the CWR prebreeding project on grasspea, we generated field and screenhouse data on wild and cultivated species for *Orobanche* tolerance, phenological traits, morphological and yield traits, drought tolerance, and ODAP and micronutrient contents in seeds over a period of 3 yr (2017–2019) and have uploaded these to the Germinate grasspea database (https://ics.hutton.ac.uk/cwr/grasspea). The concentration of ODAP was determined spectrophotometrically using an ortho-phthalaldehyde fluorescent dye. Inductively coupled plasma emission spectrometry was used for the analysis of different minerals (selenium, iron, zinc, calcium, magnesium, potassium, and copper). Screening against *Orobanche* was taken up in a highly infested field designated as a sick plot at Marchouch, Morocco. Evaluation of 515 accessions of 17 *Lathyrus* species revealed significant variation for ODAP content ranging from 0.024 to 0.456% with an overall mean of 0.129. On an average, *Lathyrus annuus* L., *L. cicero* L., and *L. gorgoni* Parl. showed lower ODAP content than the cultivated species and offer a good source for introgression in the cultivated germplasm. Results of 285 accessions representing 14 species showed good sources of resistance to *Orobanche* in CWRs mainly in *L. ochrus* (L.) DC., and *L. cicera*. Observations on traits associated with drought tolerance and escape (days to 50% flowering, root length and shoot length, number of pods at physiological maturity, biological yield, and seed yield) were taken and loaded into the database. All of the data generated under the CWR grasspea (*Lathyrus* spp.) project by the International Center for Agricultural Research in the Dry Areas (ICARDA) has been uploaded into Germinate and is freely accessible (https://ics.hutton.ac.uk/cwr/grasspea). It contains eight data sets, encompassing 515 accessions of 17 *Lathyrus* species with 15,752 data points. The grasspea Germinate portal, which we are currently developing as part of the Templeton Charity Foundation Inc. grant 'Safeguarding crop diversity for food security: Prebreeding complemented with Innovative Finance' project will be an important resource for breeders to select promising donors for introgression of desirable traits into cultivated gene pool by expanding on the current Germinate database to include new data sets and fostering a community of grasspea research worldwide.

2.26 | Germinate lentil

Lentil (*Lens* spp.) is an important cool-season food legume that is well-adapted to low- and mid-temperature environments. ICARDA's genebank has over 14,000 accessions of lentil, mostly composed of landraces and CWRs. These resources have been used to create new promising lines with good productivity and high tolerance to the various biotic and abiotic stress in the Dissemination of Interspecific ICARDA Varieties and Elites through Participatory Research project (https://mel.cgiar.org/projects/741), which is led by ICARDA in partnership with INRA Morocco, LARI Lebanon, ISRA Senegal, and EIAR Ethiopia. A set of 24 CWR-derived lines from *Lens orientalis* (Boiss.) Hand.-Mazz. [syn. *L. culinaris* Medik. subsp. *orientalis* (Boiss.) Ponert], with Bakria as a control, were screened in Morocco and Lebanon for important morphological and phenological traits including growth vigor, time to flowering, grain yield, 100-seed weight, selenium, zinc, and iron. All these data from this trial will be available from the Germinate lentil platform (http://ics.hutton.ac.uk/cwr/lentil). The superiority of the CWR-derived lines was proved with 75% yield advantage over the best national commercial check. The best performing CWR-derived lines from the Marchouch station were promoted for multilocation testing in 2020 in Morocco, Senegal, Ethiopia, and Lebanon including 20 on-farm tests conducted across Morocco. These new data will also be distributed through the Germinate lentil database to help drive insights and knowledge and provide freely available resources for the lentil community.

2.27 | Germinate pearl millet

The CWR prebreeding project on pearl millet [*Cenchrus americanus* (L.) Morrone, syn. *Pennisetum glaucum* (L.) R. Br.] focused on improving terminal drought, flowering-stage heat, and blast resistance in cultivated pearl millet. Three populations derived from wild *C. violaceus* (Lam.)

Morrone [syn. *P. violaceum* (Lam.) Rich.] as the donor and cultivated pearl millet genotypes as recipients were evaluated for flowering-stage heat stress during the 2018 summer season at two locations in Gujarat state and one location in Uttar Pradesh state in India. Two sets of these breeding materials were planted at an interval of ∼ 10 d at each site to coincide heat stress with the flowering stage of these entries. Data were recorded in days to boot leaf stage, days to 50% flowering, and percentage seed set for each genotype in each planting and has been uploaded into the Germinate pearl millet database (https://ics.hutton.ac.uk/cwr/pearlmillet).

Blast (caused by *Pyricularia grisea*; teleomorph: *Magnaporthe grisea*) screening of wild *C. violaceus* accessions and four prebreeding populations consisting of 221 lines was carried out under controlled environmental conditions at ICRISAT, Patancheru, India. The prebreeding populations were evaluated for their response to five diverse pathotype-isolates, Pg 45, Pg 138, Pg 186, Pg 204, and Pg 232 of the blast pathogen. In each of the four prebreeding populations, data were recorded on blast severity for each pathotype in S0, S1, and S2 generations and is available to download.

Three prebreeding populations consisting of 105 IL were screened for response to striga in Niamey, Niger. Data were recorded on number of hills per plot after emergence, days to 50% flowering, plant height, panicle length, number of tillers per pot, number of tillers per plant, number of Striga emerged at 60 d after sowing (DAS), striga vigor score at 60 DAS, number of Striga emerged at 85 DAS, striga vigor score at 85 DAS, and total striga per entry at harvesting. Results of the preliminary screening experiment showed that 30 lines were highly resistant (total striga per entry < 1.0) to Striga. Data is available to download (http://ics.hutton.ac.uk/cwr/pearlmillet). These resistant lines will be rescreened to confirm the results and shared with pearl millet breeding programs in Africa.

2.28 | Germinate pigeonpea

Pigeonpea (*Cajanus cajan* L. Huth) has a narrow genetic base. Under the CWR pigeonpea prebreeding project, novel variability was generated by using wild *Cajanus* species from secondary [*C. acutifolious* (F. Muell.) Maesen, *C. cajanifolius* (Haines) Maesen, *C. scarabaeoides* (L.) Thouars] and tertiary [*C. platycarpus* (Benth.) Maesen] gene pools for use in pigeonpea breeding programs. These populations were screened for important biotic (phytophthora blight, sterility mosaic disease, fusarium wilt, and pod borer) and abiotic (salinity) stresses. Selected introgression lines having good agronomic performance as well as salinity and phytophthora blight tolerant lines were evaluated for yield-related traits across locations in India and Myanmar. Based on the high-yield potential, promising prebreeding lines were nominated to the initial varietal trials of All India Coordinated Research Project on pigeonpea. In 2019–2020, the prebreeding line ICPIL 17116 was nominated under the middle–early duration group (151–165 d to maturity) whereas the two high-yielding lines ICPL 15062 and ICPL 15072 were nominated under the medium group (166–185 d to maturity) for multilocation evaluation at the national level. Further, six high-yielding prebreeding lines including, ICPIL 17116, ICPL 15062, and ICPL 15072, were shared with farmers across locations in India for conducting farmers' participatory varietal selection trials during the 2019–2020 rainy season. Two advanced backcross populations derived from *C. acutifolius* and *C. cajanifolius* were used for mapping quantitative trait loci for yield-related traits. Data for these trials is available for download and exploration (https://ics.hutton.ac.uk/cwr/pigeonpea) and additional data sets will also be made available soon.

Pod borer (*Helicoverpa armigera* Hübner) is an important damaging pest in pigeonpea. New prebreeding populations were developed using wild *Cajanus* species. A simple and complex cross-approach was used for introgressing components of pod borer tolerance from wild species into the common cultivated genetic background of popular pigeonpea varieties Asha (syn. ICPL 87119) and Maruti (syn. ICP 8863). These populations were evaluated under unsprayed field conditions during the 2018–2019 crop season followed by rescreening of pod borer tolerant lines during the 2019–2020 crop season. Data recorded on pod damage, as well as leaf and pod bioassay and biochemical characterization of pod borer tolerant lines, will soon be available through our Germinate pigeonpea platform.

2.29 | Germinate rice

2.29.1 | International Rice Research Institute and Cornell University

The CWR rice (*Oryza* spp.) project was implemented at the International Rice Research Institute (IRRI) in the Philippines and at Cornell University (Ithaca, NY, USA). The project had three main goals: (a) to examine genotypic diversity in a collection of 286 accessions representing the *Oryza rufipogon* Griff. species complex (ORSC), the immediate wild ancestor of *O. sativa*; (b) to create four populations of interspecific ILs by crossing geographically and genetically diverse *O. rufipogon* donors into an elite, high-yielding *O. sativa* variety; and (c) to examine the genotypic and phenotypic diversity of the prebreeding populations as the basis for future variety development. Data

FIGURE 9 Example of crop wild relative (CWR)-derived rice trials at International Rice Research Institute, Philippines

from the project has been deposited into the Germinate rice database (https://ics.hutton.ac.uk/cwr/rice).

We initially assembled a collection of *O. rufipogon* accessions from the IRRI genebank to represent the geographical distribution of the ORSC. Accessions were purified via single seed descent for two generations in the greenhouse at Cornell, and DNA from individual plants was used for genotyping. Genotyping data, generated via GBS and targeted chloroplast sequencing, was used for phylogenetic analysis as described by Kim et al. (2016). Based on this initial study, we identified diverse clusters of ORSC accessions and selected four genetically and geographically diverse wild accessions for use as donors. The wild donors were crossed to a common recurrent parent, the elite indica line NSiC Rc 222 (also referred to as IRRI 154), and subsequently backcrossed to generate prebreeding populations. These four populations of wild ILs were genotyped in the BC3F2 generation using GBS and phenotyped in the BC3F3 generation for 14 traits under both drought and well-watered conditions during the 2016 dry season (DS2016) at IRRI (Figure 9). Finally, an additional population of wild chromosome segment substitution lines (CSSLs) (Arbelaez et al., 2015) were phenotyped for the same 14 traits under drought and well-watered conditions during the same DS2016 trial at IRRI. The CSSL population, which had been developed previously through a collaboration between The International Center for Tropical Agriculture (CIAT) and Cornell, was of particular interest for this study because it represented a cross between one of the same *O. rufipogon* donors as the IL population but in the genetic background of a drought-tolerant, upland tropical Japonica variety from Brazil. We were interested to determine whether any of the same introgressions conferred an advantage under drought in both the Indica and the tropical Japonica backgrounds.

The genotypic and phenotypic data sets outlined above are available from Germinate, and seeds from the collection of ORSC, the ILs, and the CSSLs are available through IRRI and CIAT. A selection of wild ILs developed on this project has also been shared with researchers at the Mekong Delta Development Research Institute in Dong Thap province, Vietnam, and is currently being evaluated for diverse traits related to climate resilience, disease and insect resistance, productivity, and grain quality by a participatory network of farmers' groups from the Mekong Delta, known as Seed Clubs (https://www.cwrdiversity.org/rice_in_vietnam). Initial results indicate that some of the ILs developed on this project show enhanced adaptation to drought, acid sulphate soils, and resistance to blast disease and brown planthopper.

2.29.2 | Vietnam

Two hundred BC3F3B prebreeding lines from the CWR rice prebreeding project were initially evaluated on-farm in cooperation with 13 farmer seed clubs in eight provinces of the Mekong Delta region in Vietnam. All the data from these farm-scale participatory evaluation trials will be made available through the Germinate rice platform (https://ics.hutton.ac.uk/cwr/rice).

A total of 1027 BC3F4 lines derived from the 200 prebreeding lines from IRRI (which originated from the interspecific cross between NSiC Rc222 and several CWR accessions developed through the collaboration of IRRI and Cornell University under the pilot CWR-prebreeding project from 2011–2016) were selected in 2019 for further evaluation. These lines were selected on the basis of their phenology, agronomic and yield traits, and resistance to brown plant hopper and rice blast disease (data for pest and disease resistance are yet to be shared). In total, in excess of 16 traits are recorded in the Germinate database.

Furthermore, 50 stable lines were screened for salinity tolerance at seedling stage in a hydroponic system using Yoshida nutrient solution with different salt concentration (4, 6, and 8%) under laboratory conditions at Can Tho University. Phenotypic responses such as salt injury symptoms, plant height, root length, shoot dry weight, and root dry weight were recorded and analyzed. The responses of the screened lines were compared with the responses of three check varieties: Pokkali (international resistant check), Doc Phung (national resistant check), and IR28 (susceptible check).

2.30 | Germinate sorghum

Sorghum is an important source of food and fodder in water-limited agricultural production systems in the tropics and subtropics. Crop simulation modeling predicts that climate change is likely to increase incidence of abiotic stress, destabilizing sorghum production (Burke, Lobell, & Guarino, 2009, Lobell et al., 2015). To protect and enhance the productivity and sustainability of the sorghum grain crop, breeders will need access to novel genetic variation in adaptive traits.

Knowing the value of backcross nested association mapping (BCNAM) populations for exploiting the genetic diversity present in sorghum CWRs, the sorghum CWR prebreeding project developed a BCNAM panel of 1224 lines using nine exotic parents from the *S. bicolor* (L.) Moench subsp. *verticilliflorum* (Steud.) de Wet ex Wiersema & J. Dahlb., *S. bicolor* (L.) Moench nothosubsp. *Drummondii* (Steud.) de Wet ex Davidse, and *S. margaritiferum* Stapf [syn. *S. bicolor* (L.) Moench subsp. *bicolor*] taxa crossed to two adapted elite lines (Macia and QL39). Whole genome SNP data have been generated for 1,219 individuals from the panel using the DArTseq GBS platform. As each population was developed, selection was practiced to ensure the resulting lines were nonshattering and photoperiod insensitive to make the more amenable to evaluation for adaptive traits without the confounding effects of these phenotypes. In contrast to the BCNAM populations previously developed by our group, stabilizing selection for height and maturity was not practiced to the degree used in previously developed BCNAM populations developed for cultivated sorghum (Jordan, Mace, Cruickshank, Hunt, & Henzell, 2011). A wide range of plant types have been conserved, from fine-stemmed 3-dwarf types to robust 1- and 2-dwarf types with thick stems.

In the 2017–2018 summer these 1,224 lines were grown for seed increase at Warwick, Queensland, Australia and two partially replicated preliminary characterization experiments were grown at Gatton (humid subtropical, 27°33′ S, 152°20′ E) and Emerald (semiarid tropical, 23°32′ S, 148°11′ E). Plant height of the main stem and a rust resistance rating were recorded at both the Gatton and Emerald sites. Because of constraints, days to flower were recorded at Emerald only. Data were analyzed with ASREML-R (Butler, Cullis, Gilmour, & Gogel, 2009) to generate best linear unbiased predictions for these traits. Trait variances for each pedigree and heritabilities for each trait were estimated.

Genotypic and phenotypic data has been loaded into Germinate (https://ics.hutton.ac.uk/cwr/sorghum). Seed of >95% of the lines is already available from the Australian Grains Genebank with the remainder to be logged soon.

2.31 | Germinate sunflower

The CWR prebreeding project on sunflower (*Helianthus annuus* L.) has had two primary goals: to create prebred lines that contain introgressions from wild *Helianthus* species and to identify which of these lines show resistance to abiotic and biotic stressors and could therefore be useful for sunflower breeding programs. Data from the entirety of this project has been made available through the Germinate sunflower portal (https://ics.hutton.ac.uk/cwr/sunflower).

In the first stage of this project, a diversity panel of 169 wild *Helianthus* accessions was examined. Twenty-eight individuals representing 10 species were selected as wild donors for the prebred lines, which were created by crossing wild donors with the elite sunflower cultivar HA89 (see Warschefsky and Baute et al., unpublished data, 2020). Data, already available on Germinate, includes high-throughput SNP genotypes, collection localities (latitude, longitude) for the wild diversity panel of 169 *Helianthus* accessions (including the 28 wild donors), and pedigree information for each of the 426 prebred lines. In the future, SNP genotype data for 1,428 individuals representing the 426 prebred is available from the Germinate sunflower database (https://ics.hutton.ac.uk/cwr/sunflower).

In the second stage of the CWR sunflower project, the prebred lines have been screened for resistance to important abiotic (drought, heat, low nutrient), and biotic (stem canker [*Diaporthe helianthi*], Verticillium wilt [*Verticillium dahliae*], and downy mildew [*Plasmopara halstedii*]) stressors in field trials. Additionally, trade-offs between stress resistance and performance have been examined as well as oil content and composition. Data from these trials, including various measures of performance, such as flowering time, head diameter, reproductive biomass, and 1000-seed weight, along with oil content and composition, have been shared publicly via our Germinate sunflower portal. Furthermore, photographs from which some data (e.g. specific leaf area) were generated has been shared as will carbon/nitrogen isotopic ratios for low-nutrient trials. With the combination of genotypic and phenotypic information, the Germinate sunflower portal will be an important resource for breeders and growers to select promising lines for their individual growing conditions and challenges.

3 | DISCUSSION

The new Germinate platform offers major improvements to the functionality, speed, and features of its predecessor, Germinate 3. These advancements in functionality bring a step change to the capabilities of the platform both in terms of the technologies used and the volumes of data that can be routinely handled. These updated features were developed as a result of both user testing and requirements of large international prebreeding projects that now use the platform as their primary data distribution mechanism. Undertaking user evaluations has tangible benefits in quantifying user experience and providing direction on what features can be improved to meet user needs. This hybrid approach has allowed us to target the features that users most want and then verify their effectiveness in a wide variety of species. We have demonstrated Germinate's utility through both user testing and the deployment of prebreeding data for 13 species from the Crop Trust Crop Wild Relatives project and CIMMYT's maize and wheat collections along with implementations for National Collections such as the Commonwealth Potato Collection. Germinate is a flexible application appropriate for a wide range of situations from stand-alone private access projects through to provision of the multiuser, multicrop platform increasing in demand to facilitate the handling and interaction of prebreeding material related to the accessions available in international genebanks. In its current form, it now provides tools that have been successfully deployed across 19 different species and includes new functionality for data visualization, data export, and data upload as well as better integration with germplasm catalogs, including Genesys and Eurisco, and compatibility with developing standards such as the use of DOIs. A single log-on can be used across Germinate databases, which is a useful feature for groups working on multiple species and projects.

While there are other tools that provide similar functionality to Germinate, to our knowledge, none offer the depth of data types and visualization tools and connectivity that Germinate provides (Table 2).

The development of common templates for data have helped reduce errors generated at the data import stage such as problems with naming of germplasm, duplication, typographical errors, and inconsistencies in trait names, which are common problems for similar databases. The structural changes introduced since Germinate 3 enable more continuous maintenance and upgrading under the new Germinate brand. This will facilitate further development and refinement of all aspects of the software, which will be freely available to users as automatic updates. Notable tasks that have been identified from our user survey include improving the data upload tools particularly for new users who have minimal bioinformatics and data handling skills. We will also integrate the Germinate platform with mobile applications to enhance its utility, from data collection to data distribution, for a variety of applications and use cases.

There is a need to improve the infrastructure supporting the reuse of scholarly data. This includes establishing principles for the long-term management and stewardship of data, for example, the FAIR (findable, accessible, interoperable, reusable) principles (Wilkinson et al,

TABLE 2 Comparison between tools offering similar functionality to Germinate. These include: Tripal (https://www.drupal.org/project/tripal), Intermine (http://intermine.org), T3 (https://triticeaetoolbox.org), Sol Genomics Network (https://solgenomics.net), and MaizeGDB (https://www.maizegdb.org). Asterisk (*) denotes that while the feature is available, it may require additional components or work to achieve. As is the nature with informatics projects features can change quickly so checking with the project websites for up to date information is essential

Core features	Germinate	Tripal	T3	Sol Genomics Network	MaizeGDB
Docker	Yes	Yes	No	No	No
Multi-crop Passport Descriptors	Yes	Yes*	No	No	No
Pedigree data	Yes	No	Yes	Yes	No
Genomic data	No	Yes	Yes	Yes	Yes
Genotypic data	Yes	Yes*	Yes	Yes	Yes
Phenotypic data	Yes	Yes*	Yes	Yes	Yes
Field-trial data	Yes	No	Yes	Yes	No
Image data	Yes	Yes	No	Yes	No
Data set-level access control	Yes	Yes*	No	No	No
Custom data set licensing	Yes	Yes*	No	No	No
User annotations	Yes	Yes	No	No	Yes
BrAPI[a] compatibility	Yes	Yes*	Yes	Yes*	No

[a] Plant breeding application programming interface.

2016), and clarifying the roles and responsibilities of the parties responsible for ensuring compliance with those principles. The Germinate infrastructure has a clear role to play by offering a data management tool that can be used in a number of crop germplasm contexts; however, it is important to realize that technology alone is not sufficient. There is a human obligation in both funding agencies and research organizations to ensure that data from crop science projects are not only suitably annotated, analyzed, and accessible but that the data and information generated by these projects are available and continue to be accessible for as long as they remain relevant. This may require a different approach from both funding agencies and research organizations to ensure that projects are funded adequately and that any project generating data includes a plan for responsible, long-term stewardship of those data in the context of a well-structured data repository.

An evolving list of updates to the Germinate platform can be found at the Germinate GitHub pages (https://github.com/germinateplatform) along with links and documentation to download and install Germinate, tutorial videos describing how to use the system, and example data sets that allow users to explore all the features offered by the platform.

The efficient management and distribution of experimental data from prebreeding projects is important in ensuring uptake of enhanced germplasm into breeding and research programs.

4 | AVAILABILITY

Germinate is open source and freely available from our Github page (https://github.com/germinateplatform). Additional information about the platform is also available (http://ics.hutton.ac.uk/get-germinate). We are particularly keen to hear from users who want to contribute to enhancing or improving the platform. Please contact us at germinate@hutton.ac.uk with your feedback, questions, or to discuss suggestions for new features. Germinate now holds data for a large number of species and we are actively trying to foster community interactions to demonstrate the benefits of using a common platform across all species. We are happy to help new users make experimental data available to the public via Germinate, with particular attention to smaller groups who may not have access to dedicated bioinformatics teams. Our online video-based tutorials provide step-by-step walk-throughs of Germinate installation, docker deployment, and basic usage and can be found as links from our main Germinate webpage (http://ics.hutton.ac.uk/get-germinate).

Germinate does not store tracking cookies apart from those required by Google Analytics.

- The pedigree visualization tool Helium is free to use for both commercial and noncommercial purposes and can be downloaded from https://ics.hutton.ac.uk/helium.
- The visualization tools Flapjack and CurlyWhirly are open source and can be downloaded from https://ics.hutton.ac.uk/flapjack and https://ics.hutton.ac.uk/curlywhirly.
- Germinate CWR alfalfa can be accessed from https://ics.hutton.ac.uk/cwr/alfalfa
- Germinate CWR barley can be accessed from https://ics.hutton.ac.uk/cwr/barley
- Germinate CWR chickpea can be accessed from https://ics.hutton.ac.uk/cwr/chickpea
- Germinate CWR cowpea can be accessed from https://ics.hutton.ac.uk/cwr/cowpea
- Germinate CWR durum wheat can be accessed from https://ics.hutton.ac.uk/cwr/wheat
- Germinate CPC can be accessed from https://ics.hutton.ac.uk/germinate-cpc
- Germinate CWR eggplant can be accessed from https://ics.hutton.ac.uk/cwr/eggplant
- Germinate CWR finger millet can be accessed from https://ics.hutton.ac.uk/cwr/fingermillet
- Germinate CWR grasspea can be accessed from https://ics.hutton.ac.uk/cwr/grasspea
- Germinate CWR lentil can be accessed from https://ics.hutton.ac.uk/cwr/lentil
- Germinate CIMMYT maize can be accessed from http://germinate.cimmyt.org/maize/ and requires registration for access to data.
- Germinate CWR pearl millet can be accessed from http://ics.hutton.ac.uk/cwr/pearlmillet
- Germinate CWR pigeonpea can be accessed from http://ics.hutton.ac.uk/cwr/pigeonpea
- Germinate CWR rice can be accessed from http://ics.hutton.ac.uk/cwr/rice
- Germinate CWR sorghum can be accessed from https://ics.hutton.ac.uk/cwr/sorghum
- Germinate CWR sunflower can be accessed from https://ics.hutton.ac.uk/cwr/sunflower
- Germinate CIMMYT wheat can be accessed from http://germinate.cimmyt.org/wheat/ and requires registration for access to data.
- Additional Germinate databases can be found at https://ics.hutton.ac.uk/get-germinate
- Germinate Scan can be downloaded through the Google Play Store https://play.google.com/store/apps/details?id=uk.ac.hutton.android.germinatescan

ACKNOWLEDGMENTS

This work was undertaken as part of the initiative 'Adapting Agriculture to Climate Change: Collecting, Protecting and Preparing Crop Wild Relatives,' which is supported by the Government of Norway. The project is managed by the Global Crop Diversity Trust with the Millennium Seed Bank of the Royal Botanic Gardens, Kew, and implemented in partnership with national and international genebanks and plant breeding institutes around the world. For further information, see the project website: http://www.cwrdiversity.org. The authors also gratefully acknowledge funding from the Scottish Government's Rural and Environmental Science and Analytical Services (RESAS) division, Scottish Environment Food and Agriculture Research Institutes (SEFARI), the International Barley Hub (IBH) and the Templeton World Charity Foundation, Inc. Grant ID TWCF0400 'Safeguarding crop diversity for food security: Prebreeding complemented with Innovative Finance.' Additional funding for Germinate components was provided by Seeds of Discovery (SeeD) Initiative component MasAgro Biodiversidad, funded through the Sustainable Modernization of Traditional Agriculture (MasAgro) Project by the Secretaría de Agricultura y Desarrollo Rural (SADER) of the government of Mexico as well as the MAIZE and WHEAT CGIAR Research Programs and the International Wheat Yield Partnership (IWYP). We would also like to thank colleagues at the James Hutton Institute, in particular, members of the Information and Computational Sciences and Cell and Molecular Sciences Departments in Invergowrie for their help and support throughout this work. Additionally, we would like to thank all of those who have been generous enough with their time, enthusiasm, and feedback over the development of Germinate and their help in carrying out user testing on the platform without whom projects such as these would not succeed.

AUTHOR CONTRIBUTIONS

P. D. Shaw cowrote this manuscript and leads the Germinate project. S. Raubach cowrote this manuscript and is the primary Germinate platform developer. B. Kilian contributed toward this manuscript and has provided feedback on Germinate. K. Dreher has contributed to this manuscript and is responsible for deployment of Germinate for the CIMMYT maize and wheat collections and has provided technical feedback on features. D. F. Marshall contributed toward the writing. G. Stephen and I. Milne developed components for asynchronous data transfer, BrAPI, and interaction with other visualization tools and provide sys-admin support. M. Plazas, R. Schafleitner and J. Prohens are responsible for the eggplant data and have contributed toward writing. L. Reiseberg and E. Warschefsky are responsible for the sunflower data and have contributed toward writing. D. Cook is responsible for the chickpea data and has contributed toward writing. D. Odeny and E. Mikwa are responsible for the finger millet data and have contributed toward writing. A. Humphries is responsible for the alfalfa data and has contributed toward writing. O. Boukar, A. Togola, and C. Fatokun provided the cowpea data and contributed toward the writing. S. Sharma provided the pigeonpea and pearl millet data and has contributed toward the writing of this manuscript. S. McCouch, K. McNaly, N. Loi, S. Labarosa, and H. Tin provided the rice data and contributed toward the writing of this manuscript. A. Amri, Z. Kehel, F.M Bassi, N. El Haddad, and S. Kumar provided the grasspea and barley data and contributed toward the writing of this manuscript. A. Cruickshank, D. Jordan, and E. Mace provided sorghum data. P. Werner undertook an evaluation of the platform, provided feedback, and contributed toward the writing of this manuscript.

ORCID

Sebastian Raubach https://orcid.org/0000-0001-5659-247X
Kate Dreher https://orcid.org/0000-0003-4652-4398
Filippo M. Bassi https://orcid.org/0000-0002-1164-5598
Alan Cruickshank https://orcid.org/0000-0002-7982-1746
Alan Humphries https://orcid.org/0000-0002-8700-4340
David F. Marshall https://orcid.org/0000-0001-9309-2570
Iain Milne https://orcid.org/0000-0002-4126-0859
Damaris Achieng Odeny https://orcid.org/0000-0002-3629-3752
Jaime Prohens https://orcid.org/0000-0003-1181-9065
Shivali Sharma https://orcid.org/0000-0001-5314-484X
Abou Togola https://orcid.org/0000-0001-6155-8292
Paul D. Shaw https://orcid.org/0000-0002-0202-1150

REFERENCES

Alercia, A., Diulgheroff, S., & Mackay, M. (2015). FAO/Bioversity Multi-Crop Passport Descriptors V.2.1 [MCPD V.2.1]. Retrieved from http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/

Arbelaez, J. D., Moreno, L. T., Singh, N., Tung, C. W., Maron, L. G., Ospina, Y., … McCouch, S. R. (2015). Development and GBS-genotyping of introgression lines (ILs) using two wild species of rice, *O. meridionalis* and *O. rufipogon*, in a common recurrent parent, *O. sativa* cv. Curinga. *Molecular Breeding*, *35*, 81. https://doi.org/10.1007/s11032-015-0276-7

Bayer, M., Ganal, M., Hedley, P. E., Macaulay, M., Plieske, J., Ramsay, L., … Waugh, R. (2017). Development and evaluation of a barley 50k iSelect SNP array. *Frontiers in Plant Science*, *8*, 1792. https://doi.org/10.3389/fpls.2017.01792

Bayer, R., & McCreight, E. M. (1972). Organization and maintenance of large ordered indexes. *Acta Informatica*, *1*, 173–189. https://doi.org/10.1007/BF00288683

Bingham, E., Armour, D., Irwin, J., Jayaraman, D., & Ané, J. M. (2009). *Report on progress hybridizing herbaceous* Medicago sativa *and woody* M. arborea. *Medicago Genetic Reports*, *9*. Retrieved from http://www.medicago-reports.org/volumes09.php

Blake, V. C., Kling, J. G., Hayes, P. M., Jannink, J. L., Jillella, S. R., Lee, J., … Dickerson, J. A. (2012). The Hordeum toolbox: The barley coordinated agricultural project genotype and phenotype resource. *The Plant Genome*, *5*, 81–91. https://doi.org/10.3835/plantgenome2012.03.0002

Bradshaw, J. E., & Ramsay, G. (2005). Utilisation of the Commonwealth Potato Collection in potato breeding. *Euphytica*, *146*, 9–19. https://doi.org/10.1007/s10681-005-3881-4

Burke, M. B., Lobell, D. B., & Guarino, L. (2009). Shifts in African crop climates by 2050, and the implications for crop improvement and genetic resources conservation. *Global Environmental Change*, *19*, 317–325. https://doi.org/10.1016/j.gloenvcha.2009.04.003

Butler, D. G., Cullis, B., Gilmour, A. R., & Gogel, B. (2009). *ASReml-R reference manual, Release 3. Technical Report*. Brisbane, Australia: Department of Primary Industries and Fisheries.

Cericola, F., Portis, E., Toppino, L., Barchi, L., Acciarri, N., Ciriaci, T., … Lanteri, S. (2013). The population structure and diversity of eggplant from Asia and the Mediterranean Basin. *PLoS ONE*, *8*, e0073702. https://doi.org/10.1371/journal.pone.0073702

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., & Guarino, L. (2017). Past and future use of wild relatives in crop breeding. *Crop Science*, *57*, 1070–1082. https://doi.org/10.2135/cropsci2016.10.0885

Fahlgren, N., Gehan, M. A., & Baxter, I. (2015). Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology*, *24*, 93–99. https://doi.org/10.1016/j.pbi.2015.02.006

Germeier, C. U., & Unger, S. (2019). Modeling crop genetic resources phenotyping information systems. *Frontiers in Plant Science*, *10*, 728. https://doi.org/10.3389/fpls.2019.00728

Gramazio, P., Prohens, J., Plazas, M., Andújar, I., Herraiz, F. J., Castillo, E., … Vilanova, S. (2014). Location of chlorogenic acid biosynthesis pathway and polyphenol oxidase genes in a new interspecific anchored linkage map of eggplant. *BMC Plant Biology*, *14*, 350. https://doi.org/10.1186/s12870-014-0350-z

Gramazio, P., Prohens, J., Plazas, M., Mangino, G., Herraiz, F. J., & Vilanova, S. (2017). Development and genetic characterization of advanced backcross materials and an introgression line population of *Solanum incanum* in a *S. melongena* background. *Frontiers in Plant Science*, *8*, 1477. https://doi.org/10.3389/fpls.2017.01477

Greenlon, A., Chang, P. L., Damtew, Z. M., Muleta, A., Carrasquilla-Garcia, N., Kim, D., … Cook, D. R. (2019). Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. *Proceedings of the National Academy of Sciences*, *116*, 15200–15209. https://doi.org/10.1073/pnas.1900056116

Hawkes, J. G. (1951). The Commonwealth potato collection. *American Potato Journal*, *28*, 465–471. https://doi.org/10.1007/BF02854979

Jarvis, A., Lane, A., & Hijmans, R. J. (2008). The effect of climate change on crop wild relatives. *Agriculture, Ecosystems and Environment*, *126*, 13–23. https://doi.org/10.1016/j.agee.2008.01.013

Jordan, D. R., Mace, E. S., Cruickshank, A. W., Hunt, C. H., & Henzell, R. G. (2011). Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. *Crop Science*, *51*, 1444–1457. https://doi.org/10.2135/cropsci2010.06.0326

Kaushik, P., Gramazio, P., Vilanova, S., Raigón, M. D., Prohens, J., & Plazas, M. (2017). Phenolics content, fruit flesh colour and browning in cultivated eggplant, wild relatives and interspecific hybrids and implications for fruit quality breeding. *Food Research International*, *102*, 392–401. https://doi.org/10.1016/j.foodres.2017.09.028

Kaushik, P., Prohens, J., Vilanova, S., Gramazio, P., & Plazas, M. (2016). Phenotyping of eggplant wild relatives and interspecific hybrids with conventional and phenomics descriptors provides insight for their potential utilization in breeding. *Frontiers in Plant Science*, *7*, 677. https://doi.org/10.3389/fpls.2016.00677

Kim, H., Jung, J., Singh, N., Greenberg, A., Doyle, J., Tyagi, W., … McCouch, S. (2016). Population dynamics among six major groups of the *Oryza rufipogon* species complex, wild relative of cultivated Asian rice. *Rice*, *9*, 56. https://doi.org/10.1186/s12284-016-0119-0

Knapp, S., Aubriot, X., & Prohens, J. (2019). Eggplant (*Solanum melongena* L.): Taxonomy and relationships. In M. Chapman (Ed.) *The eggplant genome. Compendium of plant genomes* (pp. 11–22). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-99208-2_2

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2011). Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, *18*, 1520–1536.

Lee, J. M., Davenport, G. F., Marshall, D., Ellis, T. H. N., Ambrose, M. J., Dicks, J., … Flavell, A. J. (2005). GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. *Plant Physiology*, *139*, 619–631. https://doi.org/10.1104/pp.105.065201

Lobell, D. B., Hammer, G. L., Chenu, K., Zheng, B., McLean, G., & Chapman, S. C. (2015). The shifting influence of drought and heat stress for crops in northeast Australia. *Global Change Biology*, *21*, 4115–4127. https://doi.org/10.1111/gcb.13022.

Marx, V. (2013). The big challenges of big data. *Nature*, *498*, 255–260. https://doi.org/10.1038/498255a

Meyer, R. S., Karol, K. G., Little, D. P., Nee, M. H., & Litt, A. (2012). Phylogeographic relationships among Asian eggplants and new perspectives on eggplant domestication. *Molecular Phylogenetics and Evolution*, *63*, 685–701. https://doi.org/10.1016/j.ympev.2012.02.006

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W. T. B., … Marshall, D. (2010). Flapjack–Graphical genotype visualization. *Bioinformatics*, *26*, 3133–3134. https://doi.org/10.1093/bioinformatics/btq580

Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, *15*, 921–928.

Nandyala, C. S., & Kim, H. K. (2016). Big and meta data management for U-Agriculture mobile services. *International Journal of Software Engineering and Its Applications*, *10*, 257–270. https://doi.org/10.14257/ijseia.2016.10.2.21

Nelson, G. C., Rosegrant, M. W., Koo, J., Robertson, R., Sulser, T., Zhu, T., … Lee, D. (2009). *Climate change: Impact on agriculture and costs of adaptation*. Washington, DC: International Food Policy Research Institute. http://doi.org/10.2499/0896295354

Onda, Y., & Mochida, K. (2016). Exploring genetic diversity in plants using high-throughput sequencing techniques. *Current Genomics*, *17*, 358–367. https://doi.org/10.2174/1389202917666160331202742

Postman, J., Hummer, K., Bretting, P., Kinard, G., Bohning, M., Emberland, G., … Guarino, L. (2010). GRIN-Global: An international project to develop a global plant genebank information management system. *Acta Horticulturae*, *859*, 49–56. https://doi.org/10.17660/actahortic.2010.859.4

Raigón, M. D., Prohens, J., Muñoz-Falcón, J. E., & Nuez, F. (2008). Comparison of eggplant landraces and commercial varieties for fruit content of phenolics, minerals, dry matter and protein. *Journal of Food Composition and Analysis*, *21*, 370–376. https://doi.org/10.1016/j.jfca.2008.03.006

Redden, R. J., Yadav, S. S., Maxted, N., Dulloo, M. E., Guarino, L., & Smith, P. (2015). *Crop wild relatives and climate change*. Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118854396

Rife, T. W., & Poland, J. A. (2014). Field book: An open-source application for field data collection on android. *Crop Science*, *54*, 1624–1627. https://doi.org/10.2135/cropsci2013.08.0579

Sedlmair, M., Meyer, M., & Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, *18*, 2431–2440.

Selby, P., Abbeloos, R., Backlund, J. E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O. E., … Wren, J. (2019). BrAPI—An application programming interface for plant breeding applications. *Bioinformatics*, *35*, 4147–4155. https://doi.org/10.1093/bioinformatics/btz190

Sempéré, G., Pétel, A., Rouard, M., Frouin, J., Hueber, Y., De Bellis, F., & Larmande, P. (2019). Gigwa v2—Extended and improved genotype investigator. *GigaScience*, *8*, giz051. https://doi.org/10.1093/gigascience/giz051

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

Shaw, P. D., Raubach, S., Hearne, S. J., Dreher, K., Bryan, G., Mckenzie, G., … Marshall, D. F. (2017). Germinate 3: Development of a common platform to support the distribution of experimental data on crop wild relatives. *Crop Science*, *57*, 1259–1273. https://doi.org/10.2135/cropsci2016.09.0814

Shin, M. G., Bulyntsev, S. V., Chang, P. L., Korbu, L. B., Carrasquila-Garcia, N., Vishnyakova, M. A., … Nuzhdin, S. V. (2019). Multitrait analysis of domestication genes in *Cicer arietinum–Cicer reticulatum* hybrids with a multidimensional approach: Modeling wide crosses for crop improvement. *Plant Science*, *285*, 122–131. https://doi.org/10.1016/j.plantsci.2019.04.018

Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., Hancock, D., … McLaren, G. (2010). Multifunctional crop trait ontology for breeders' data: Field book, annotation, data discovery and semantic enrichment of the literature. *AoB PLANTS*, *2010*, plq008, https://doi.org/10.1093/aobpla/plq008

Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., & Arnaud, E. (2012). Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop ontology developed by the crop communities of practice. *Frontiers in Physiology*, *3*, 326. https://doi.org/10.3389/fphys.2012.00326

Teuber, L. R., Taggard, K. L., Gibbs, L. K., McCaslin, M. H., Peterson, M. A., & Barnes, D. K. (2004). Fall dormancy, standard tests to characterize alfalfa cultivars. In *Standard tests to characterize alfalfa cultivars, Third Edition (Amended 2004)*. Retrieved from https://www.naaic.org/stdtests/Dormancy2.html

Von Wettberg, E. J. B., Chang, P. L., Başdemir, F., Carrasquila-Garcia, N., Korbu, L. B., Moenga, S. M., … Cook, D. R. (2018). Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation. *Nature Communications*, *9*, 649. https://doi.org/10.1038/s41467-018-02867-z

Weese, T. L., & Bohs, L. (2010). Eggplant origins: Out of Africa, into the Orient. *Taxon*, *59*, 49–56. https://doi.org/10.1002/tax.591006

Weise, S., Oppermann, M., Maggioni, L., Van Hintum, T., & Knupffer, H. (2017). EURISCO: The European search catalogue for plant genetic resources. *Nucleic Acids Research*, *45*, D1003–D1008. https://doi.org/10.1093/nar/gkw755

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). Comment: The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

How to cite this article: Raubach S, Kilian B, Dreher K, et al. From bits to bites: Advancement of the Germinate platform to support genetic resources collections and prebreeding informatics for crop wild relatives. *Crop Science*. 2020;1–29. https://doi.org/10.1002/csc2.20248