Crop Information Engine and Research Assistant (CIERA) for managing genealogy, phenotypic and genotypic data for breeding programs

Shawn Yates[1], Martin Lagüe[2], Ron Knox[1], Richard Cuthbert[1], Fran Clarke[1,4], John Clarke[3], Yuefeng Ruan[1], Jatinder Sangha[1] and Vijai Bhadauria[1*]

[1]Swift Current Research and Development Centre, Agriculture and Agri-Food Canada, Swift Current, Canada

[2]Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, Canada

[3]Department of Plant Sciences, University of Saskatchewan, Saskatoon, Canada

[4]retired

*Correspondence: Vijai Bhadauria, E-mail: vijai.bhadauria@canada.ca

Abstract

Background: With the advent of next-generation marker platforms and phenomics in crop breeding programs, the volume of both the genotypic and phenotypic data produced has increased exponentially. Often the data remain underutilized if not properly collated, managed and accessed. Effective management of the data is paramount to making sound and timely decision on cross planning in order to accelerate genetic gain (ΔG) in crops for disease resistance, agronomic and end-use quality traits.

Results: To address the challenges in managing and efficient utilization of the sheer volume of data generated in a crop breeding program, we developed an electronic information system called the Crop Information Engine and Research Assistant (CIERA). The CIERA, written in Visual Basic, runs on the Microsoft Windows operating system and requires the .Net Framework 4.7 as well as the MySQL Community Server 5.7. The highly intuitive graphical user interface of CIERA includes user-friendly query tools to facilitate the collation of data across relevant phenotypic environments from its phenotypic data management database and can combine that information with the genealogy and genetic data from its genealogy management and genetic data management databases, respectively.

Conclusions: Using CIERA, breeders can build a comprehensive profile of germplasm, within a few minutes, to assist them in planning crosses for enhancing genetic gain by selecting superior lines for crosses.

Key words: Quantitative genetics, Wheat breeding and genetics, Pedigree database, Molecular markers, Biometrics

Background

The success in plant breeding has extensively relied on the availability and access to phenotypic and genotypic information on parental breeding material. However, crop breeding practices have changed dramatically over the past decade. The evolution of marker technology has rapidly increased breeders' understanding of the germplasm that they develop, and brought a new interest to historical phenotypic and genealogical information. With this advancement in marker technology, the amount of resulting genotypic data can be staggering and often difficult to manage, especially when dealing with next-generation DNA sequencing (NGS) data [1]. In addition, making sense of this genotypic data requires a strong understanding of the phenotype of the germplasm. Handling this massive amount of data is much easier with the use of an electronic information system [2]; although no single system will fulfill all the needs of its users.

In order to achieve trust and acceptance from breeders, a data management system must adapt to the existing practices of the breeding program, be intuitive, and not add significantly to the workload of technicians and researchers. The International Crop Information System (ICIS), was an open-source project led by researchers at the International Rice Research Institute (IRRI) [3], and was considered a good solution for addressing the information challenges faced by breeding programs. ICIS was employed by the wheat breeding programs at the Swift Current Research and Development Centre (SCRDC) from 2001-2015. During this period, the ICIS Genealogy Management System (GMS) and Data Management System (DMS) data models met most of the needs for handling SCRDC breeders' pedigree and phenotypic data [4].

Researchers found many applications of the ICIS data management capabilities during this time [5, 6]. While ICIS met many of the basic requirements, some of the components were limited or missing, especially in managing genotypic data.

After the ICIS project was discontinued in 2011, an evaluation of existing and up-and-coming breeding information systems was undertaken to search for a suitable replacement. The criteria was that a new system would have to meet the data management requirements of all SCRDC breeding programs, have an intuitive interface so that users could adopt it easily, be able to deploy to a local area network (LAN) for multi-user access, and be inexpensive to manage and operate.

Three alternative information systems were evaluated (Table 1): 1) The Integrated Breeding Platform - a Generation Challenge Programme/Bill and Melinda Gates Foundation initiative which uses the Breeding Management System (BMS) for data management purposes, as well as providing a suite of statistical and modelling tools, 2) Agrobase Generation II - a proprietary system from Agronomix Software Inc. which handles plant breeding data, some statistical analysis, reporting, and a host of other features related to a breeding project, and finally, 3) Katmandoo - a system from New South Wales Government Department of Primary Industries which handles phenotypic and genotypic datasets and has some capabilities to handle genealogy data.

Realizing that the existing ICIS implementation was still the best option for meeting many of the SCRDC breeding program requirements, we decided to develop an improved version, retaining the most used functionality of the predecessor ICIS, and adding new features to address the limitations. This also ensured that bugs and new features would be addressed immediately, and users would not have to wait for a 3rd

party to adopt these tasks. To develop this new system, we considered the experiences and user feedback of the users from over a decade. This helped to ensure that the new system would not only meet immediate needs of SCRDC breeding programs, but also be flexible enough to adapt to new breeding practices and evolving data collection technology.

Here we report the development, qualities and use of a new stand-alone electronic information system named the Crop Information Engine and Research Assistant (CIERA), useful for managing breeding data with any crop. This system, in addition to employing ICIS's approach to handle the genealogy and phenotypic information, has modules to manage genotypic data as well (Fig. 1). CIERA employs a data model that is flexible enough to evolve with the NGS data, besides handling the data generated from Single Sequence Repeat (SSR), Diversity Arrays Technology (DArT), Single-Nucleotide Polymorphism (SNP) and Kompetitive Allele Specific PCR (KASP) marker technology. This data is loaded through the Genetic Dataset Loader and queried through the Data Mining Tool modules.

The CIERA standalone version is installed on a client's machine, but can allow multiple users access to the database on a LAN at the same time. A CIERA web version is under development to port the system to a web interface and move the tools to a web server to improve speed and performance. CIERA, due to ease of use, advanced features of data handling and low cost maintenance, will be useful to the scientific community interested in bringing together all the current and historical phenotypic or genotypic information related to a crop for use in breeding programs.

Implementation

System Requirements

The CIERA standalone version currently runs on Microsoft Windows operating system, and requires the .Net Framework 4.7. Due to the potential size of datasets and output, it is recommended that, at minimum, Microsoft Excel 2010 be installed. MySQL Community Server 5.7 is currently the database backend for CIERA. An installer is available for the CIERA launcher and tools on SourceForge (https://sourceforge.net/projects/ciera/).

To create a modern and visually enriched interface, Microsoft Windows Presentation Foundation (WPF) (Microsoft 2017) was used for the graphical user interface (GUI) windows, while the rest was developed using Visual Basic (Microsoft 2017).

CIERA has been in use with the breeding programs at SCRDC since 2016, and most issues relate only to hardware limitations particularly on machines with limited RAM when datasets reach millions of rows. No issues with CIERA are reported on high-end machines having at least 16 GB of RAM.

Databases

In 2012, the International Maize and Wheat Improvement Center (CIMMYT) released a public wheat database to all ICIS/BMS users in the ICIS GMS schema format, containing pedigree information on over six million germplasm from international wheat breeding programs. The information in this database provides wheat breeders a valuable resource about the ancestry of many germplasm lines in the background of the

wheat breeding programs, so only minor tweaks were made to the ICIS GMS schema (Fig. 2) to avoid the daunting task of porting the records from this database to a completely new schema. Since users of the legacy ICIS genealogy tools found them quite intuitive, the decision was made to convert the 32-bit tools to a 64-bit platform, but keep the look and feel of their user interface.

The CIERA DMS is the database schema for storing raw and analyzed phenotypic data at SCRDC, and is based on the ICIS DMS example (Fig. 3). It has proven to accommodate any experimental design, variable or data type required for the breeding programs, and allows breeding programs to define their own variables and nomenclature. This flexibility allows CIERA to adapt to a breeding program, rather than having breeding programs learning new terminology which can lead to mistakes and be cumbersome for staff.

ICIS never had a fully functioning Genotypic Data Management System (GDMS) database schema or tools in production for handling genotypic data, whereas for CIERA, a database was worked out using a modified schema from the GDMS proposed by the Integrated Breeding Platform (IBP) [7]. New objects were added to this GDMS schema, to address missing requirements from the SSR, DArT, SNP and KASP marker datasets that were being used in the breeding programs at the SCRDC (Fig. 4). The schema was designed to handle marker and dataset metadata, alleles (both numeric and character), gene/trait-marker associations and identification of control markers in datasets, quantitative trait locus (QTL) analysis results, and genetic maps. The CIERA GDMS could handle all types of genotypic data by designating the data points as numeric or character, much more useful with NGS technology, generating millions of

data points, compared to the few dozen that an SSR dataset might contain. It will be important for researchers to make decisions on which information they want to store in the GDMS, as not all raw reads are useful and these sequences can quickly push the database to an enormous size.

CIERA Launcher

A CIERA Launcher interface was constructed to facilitate logging into the system and navigating the upload and query tools. The launcher's interface was developed in consultation with technicians and researchers from various breeding programs to ensure the graphical user interface (GUI) was intuitive.

Users create an account and are assigned an ID within the system that tracks the data they upload into the database and controls the tools and data they have access to, as assigned by the administrator. The menus within the launcher focus on regular tasks that a researcher would query for when looking at breeding program data (Fig. 5). By putting the tools in sub-menus that are related to the same task, the user can easily maneuver between tasks and not be overwhelmed by a long list of available tools.

Genealogy Tools

The tools used for loading and querying pedigree data in the stand-alone version of CIERA are based on the legacy ICIS tools: Genealogy Management System Search (GMS Search), and Set Generation (SetGen). The functionality of these tools is well

defined by Portugal et al. [3] and McLaren et al. [4]. By carefully examining the ancestry of potential parents, breeders can make more informed decisions on the sources of disease resistance genes and phenotypic characteristics. This leads to more efficient cross planning and better understanding of the germplasm within the breeding program.

Genetic Dataset Loader

The Genetic Dataset Loader was developed to facilitate the management of genotypic data. This tool allows users to load, edit and remove genetic data from CIERA.

The first step to loading any genotypic data is ensuring the marker metadata is in the system. The Genetic Dataset Loader handles the loading of marker metadata in two ways: by allowing the user to add information for individual markers, or the user can batch load information on multiple markers at once by filling the marker information into a Microsoft Excel template and loading it. Once in the GDMS, users can edit the marker metadata, link the markers to known phenotypic traits and genes, or remove the marker and all corresponding information (including data from datasets and maps) from the GDMS (Fig. 6).

There are also templates for loading SSR, DArT, KASP and SNP datasets, which can be downloaded from the CIERA SourceForge site. Smaller datasets can be loaded via Microsoft Excel templates, while larger datasets use tab-delimited text files. The Genetic Dataset Loader performs data validation checks prior to importing the dataset into the GDMS, such as making sure the markers and germplasm lines included in the dataset are loaded in CIERA, and that all required fields are filled in.

The Genetic Dataset Loader also manages the importing of genetic maps into the GDMS. The user fills a Microsoft Excel template with map metadata and map information. Upon importing, the tool will again ensure that all markers on the template are already loaded in CIERA, and then move the information to the tables in the database.

Finally, users can add Quantitative Loci Locus, Marker Controls and Marker-Trait Association information through the interface, simply by entering the required metadata on the input screen. This information is used in the genetic data queries to provide more detailed representation of the marker data and assist researchers in making more informed decisions on vast volumes of data.

Fieldbook Manager

The Fieldbook Manager was developed to facilitate the loading of phenotypic datasets into the DMS. A major problem with managing phenotypic data in any information system is ensuring that the variables are standardized and well-defined within the system. Prior to any loading of phenotypic data, a dictionary of all measured variables from a breeding program should be collected and added to CIERA. This dictionary needs to include the preferred acronym, property (trait) name, scale used, and the method of how the data is obtained, and this should be the starting point for breeding programs wishing to utilize any information system. To assist with this, the Ontology Manager was developed, which allows users to clearly define all aspects of a variable in CIERA.

To import a Microsoft Excel sheet of phenotypic data, the user maps each column heading acronym to a defined variable in the system, or adds a new one. This allows the user to keep acronyms in their spreadsheets that are familiar to them, but ensures at loading time that they are standardized in the DMS. After some validation checks to ensure the data in the worksheet does not have any obvious errors (characters in numeric fields, duplicate rows, etc.), the Fieldbook Manager imports the data to the DMS.

Users can also create their own field books with the Fieldbook Manager, based on a generic template that the SCRDC wheat breeding programs have been using for many years (Fig. 7). The user creates a germplasm list of cultivars,, loads a randomization for the test (in a tab-delimited text file, generated from other applications) and some study metadata, along with the variables they wish to include. The tool then generates a standardized Microsoft Excel field book, in plot order, that can be directly sent to collaborators at each breeding test location. The advantage to using these field books is that at loading time, the system will recognize the column headings and directly load the file and all study metadata, saving the time of having to map each column heading.

The Fieldbook Manager also handles deleting phenotypic studies, should there be any errors to loaded datasets.

Data Mining Tool

Querying and collation of the genealogy, phenotypic and genotypic data across environments and datasets is a major advantage to using an electronic information

system. While using the ICIS implementation, the Data Mining Tool was created at SCRDC using Visual Basic for querying the database. For CIERA, this tool was modified and expanded to incorporate more queries and performance. This tool now has numerous queries (Table 2) to allow researchers to combine phenotypic and/or genotypic data from across datasets and export it to a Microsoft Excel sheet, tab-delimited text file or Extensible Markup Language (XML) file (Fig. 8). The building of a query through a wizard-like approach allows the user to customize the output or filter results not relevant to their needs. This becomes essential when the output from queries is very large. Collation of pertinent phenotypic and genotypic data is paramount when planning crosses and making decisions on which test lines to advance in a breeding program. Making use of the query wizard in CIERA allows breeders to allocate more resources on the decision-making process and less time on the gathering, standardizing and formatting of datasets.

To assist researchers in evaluating the marker datasets with the visualization tool Flapjack [8], functionality was added to the output routines of the Data Mining Tool to export results from genotypic queries to Flapjack input files.

Network Deployment

At SCRDC, the CIERA Launcher and tools were installed on users' machines from a network share via ClickOnce deployment [9], which allows users for easy updating and addresses the problem of having different versions of the software on the computer, by alerting the users to an update upon launching CIERA. A stand-alone

installer was also created using the Visual Studio Setup and Deployment project add-in (Microsoft 2017), for users without access to a network, or may require off-line access with the understanding that manual updates will be required in that configuration.

CIERA Web

In addition to the standalone CIERA, a web-based version is under development. (Fig. S1). This project is using ASPX.net and C# as the programming languages (Microsoft 2017). CIERA web will provide many advantages. The main being the users' accessibility for CIERA implementation, making the CIERA platform-independent as it will run through a web browser, rather than on the user's machine. CIERA web also resolves the updating issue, because the tools for CIERA will run off the web server so users will always run the latest version. Another advantage is that, by having the heavy lifting of queries run on the server, users do not need to worry about having high-end workstations for large data sets. Simple devices could be used to access the server (laptops, tablets and even phones), and data could be queried and entered in real time wherever internet is accessible (i.e. greenhouse, laboratory, etc.).

Results and discussion

The concept for CIERA was to not only store breeding material data, but also to take the next step in assisting researchers to use the data with full potential. The Wheat GMS of CIERA at SCRDC currently stores pedigree and passport information on over six million wheat germplasm. These include crosses, registered varieties, breeder lines, and accessions. Breeders can use this information when planning crosses and can

assist in determining the relationships between lines. It also allows breeding programs to track their germplasm samples through CIERA, and link Germplasm IDs with specific bags or envelopes of seed, which is extremely important when tracking seed source of germplasm lines.

The CIERA DMS at SCRDC houses over nine hundred phenotypic studies. These studies include registration yield trials, population evaluations, disease nurseries and quality testing. Where available, both raw data and LS means are stored. The variables are standardized across years and studies, which allows for querying and analyzing the data across environments, and assists breeders in determining candidate germplasm lines to advance through their breeding programs.

As for genotypic data, the CIERA GDMS at SCRDC stores metadata on over forty thousand markers, which include SSR, KASP, SNP and DArT marker types. Close to one hundred genotypic datasets and sixteen genetic maps have also been loaded and are accessible to CIERA wheat users. This data is used for marker assisted selection purpose, but also in marker development and marker validation. As mentioned at the outset, genomics is becoming a major part of modern breeding programs, and NGS technology has been constantly producing massive amounts of genotypic data, which needs proper management and analysis. Deciding which data to store and what format to be used is a challenging task requiring constant evaluation.

CIERA allows the researchers to query the breeding data for genealogy, phenotyping and genotyping information together and simultaneously assemble phenotype and marker profiles of germplasm lines in the breeding program. The Data Mining Tool reduces the time spent on cumbersome efforts of scanning through large

spreadsheets filled with thousands of markers on hundreds of germplasm lines to a few minutes, thus saving the resources which would otherwise be required for manual query through datasets and online databases.

Other Systems

Selecting an information system for breeding program data management is purely a subjective choice; each organization and breeder faces different challenges and employs various approaches to plant breeding. Therefore, a breeder has to assess their program requirements, and choose a system that meets the majority of those needs. Three very capable systems were evaluated as a possible replacement to ICIS at SCRDC in 2015, with the understanding that no one system fits all requirements.

The Breeding Management System (BMS) (https://www.integratedbreeding.net/) is the data management component of the IBP. This system not only offers data management tools but also some modelling and statistical tools for breeders, currently not possible with ICIS or CIERA. The BMS is very good at creating phenotypic field books, managing germplasm lists, generating various experimental designs, and loading phenotypic datasets. It also comes with a public wheat database from CIMMYT, which in addition to the genealogy information, has phenotypic and genotypic data as well. However, the feedback from SCRDC users was that the BMS interface was difficult to use, the genealogy tools were not as intuitive as the legacy ICIS tools, and querying phenotypic datasets across studies was difficult, or did not work. Users found they could not search for both phenotypic and genotypic data at the same time, and there were several issues with loading genotypic data. The statistical and

modelling components of the BMS were not interesting to SCRDC breeders, as they use other software packages for these purposes. The IBP also requires an annual fee to use some portions of the BMS, with no guarantee that free modules would remain free in the future. Finally, there was a concern that SCRDC issues may not be addressed in a timely manner, as the IBP mandate is to support users in developing countries first.

Agrobase Generation II (https://www.agronomix.com), from Agronomix Software Inc., also offers similar data management features to ICIS and the BMS. It has a suite of tools for performing statistical analysis, tools for printing reports, labels and designing field plans. But Agrobase also requires an annual subscription, and there was concern about how it handles genotypic datasets, specifically NGS data. Being a proprietary system, it would not be possible to develop any new tools to address missing elements or incorporate any in-house tools, as the code is closed to the public. No public wheat database is distributed with Agrobase, so all the pedigree information from the wheat ICIS GMS databases would have to be imported manually through the interface. There was also concern that by using a proprietary system, features may change or be removed without consultation and that issues raised by SCRDC may not be addressed unless they were wide-spread.

Katmandoo (http://katmandoo.org.au/help/Index.htm), from NSW Government Department of Primary Industries, is an open source system that was originally designed to assist breeders in identifying notable germplasm lines through their phenotypic and genotypic data, and subsequently evolved into a breeding data management system. Feedback from SCRDC users regarding Katmandoo was that its

interface was not intuitive and required a learning curve to run. There were also concerns that it did not manage genealogy data as well as the ICIS GMS. As with Agrobase, no public wheat database comes with Katmandoo, so the information from the wheat ICIS GMS databases would have to be migrated to the new database or be imported through the interface. Also, it was felt that there were some limitations to being able to load all the different types of phenotypic and genotypic data used in the breeding programs at SCRDC, and that Katmandoo lacked querying capabilities for the data that the Data Mining Tool provided in ICIS.

Other Crops

In 2016, the Agriculture and Agri-Food Canada (AAFC) Fredericton Research and Development Centre (FRDC) showed interest in CIERA. Their centre focuses on potato breeding, which is a very different type of breeding program from the wheat breeding programs at SCRDC. Since ICIS data models were used for data management in international breeding programs for dozens of different crops in 2011, it was expected that CIERA would also inherit this ability. After some demonstrations and discussions, the breeders at the FRDC did not find any major barriers in having CIERA manage their potato breeding data and this led to collaboration between the Research Centres to share resources. Having a different crop, which utilizes a different type of breeding program, will verify the flexibility of CIERA and corrections will be made should any problem arise, which is one of the major advantages to open source, collaborative development.

Future developments

The focus going forward will be on completing the CIERA web to replace the standalone version of CIERA. With the large datasets being generated in modern breeding programs, it is necessary to reduce any speed issues while querying or loading data and 64-bit web architecture would provide increased efficiency in computer power [10].

In addition to the CIERA web development, another focus is to integrate many of the independent in-house tools that technicians at SCRDC use for seeding and harvest tasks, similar to the suite of tools in Agrobase. These tools will make use of the information stored in the database instead of exporting and importing input files. This will increase the efficiency of the breeding program activities and also reduce errors. Breeding programs also use many 3rd Party proprietary and open source software tools to perform various tasks, analysis, visualization, reports and printing labels. While it is not prudent to develop new tools to replace existing 3rd Party software that already meets the required needs of users, input files can be prepared if the software can make use of the data stored in CIERA. Manually creating these input files can be time-consuming, and there is always concern for errors during preparation.

Conclusions

CIERA was built upon the knowledge and experiences gained from collaborating for several years with an excellent development team at IRRI and the feedback from researchers, technicians and computer science staff from various organizations and breeding programs. The goal for CIERA was to create a system that facilitates the

researchers working with modern breeding program data, as this data has become too cumbersome to handle solely with spreadsheets. CIERA is highly efficient as it possess all the abilities for data management for standardization, storage, and computing power. The lessons learned during the collaboration between SCRDC and FRDC, each with very different breeding programs, will provide a path to expanding CIERA to other crops and research centres going forward. Future developments will focus on finding ways to improve the usage of the data and make it more informative for researchers.

Availability and requirements

Project name: Crop Information Engine and Research Assistant (CIERA)

Project home page: https://sourceforge.net/projects/ciera/

Operating System: Microsoft Windows

Programming Language: Visual Basic

Other requirements: MySQL .NET Connector, MySQL Community Server 5.7, Microsoft Excel 2010 (minimum)

License: GNU General Public License, Version 3.

Any restrictions to use by non-academics: No restrictions

List of abbreviations:

AAFC – Agriculture and Agri-Food Canada

BMS – Breeding Management System

CIMMYT – International Maize and Wheat Improvement Center

DArT – Diversity Arrays Technology

DMS – Data Management System (phenotypic)

FDRC – Fredericton Research and Development Centre

GDMS – Genotypic Data Management System

GMS - Genealogy Management System

GUI – Graphical User Interface

IBP – Integrated Breeding Platform

ICIS – International Crop Information System

IRRI – International Rice Research Institute

KASP – Kompetitive Allele Specific PCR

LAN – Local Area Network

NGS – Next-generation sequencing

QTL – Quantitative Trait Locus

RAM – Random Access Memory

SCRDC – Swift Current Research and Development Centre

SNP - Single-Nucleotide Polymorphism

SSR – Single Sequence Repeat

WPF – Microsoft Windows Presentation Framework

XML – Extensible Markup Language

Acknowledgements

The development of CIERA was funded by the Canadian National Wheat Cluster program. We would like to thank the technical and development assistance of Thomas MacGregor, Jonathon Florek, Jordan LeBlanc and Calvin Blacquier. We would also like to acknowledge the excellent work put in by the ICIS development team lead by Dr. Graham McLaren, Dr. Richard Bruskiewich, Arllet Portugal and Dr. Thomas Metz, formerly at the International Rice Research Institute (IRRI).

Funding

The development of CIERA was funded by the Canadian National Wheat Cluster program. The funders had no role in the software design or development, or the decision to submit the article for publication.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

- A demonstration MySQL database, filled with data from templates containing various phenotypic and genotypic data is available on the CIERA sourceforge.net repository, https://sourceforge.net/projects/ciera/files/Demonstration%20Files
- A demonstration tutorial is also available on the project wiki on sourceforge.net,. https://sourceforge.net/p/ciera/wiki/Demonstration

Competing financial interests

The authors declare that they have no competing interests

Author contributions

SY, VB and ML developed the CIERA; RK, FC, JC, RC and YR provided input, testing and feedback for CIERA; SY and VB wrote the manuscript; and JS provided suggestions on the manuscript.

References

1. Batley J, Edwards D. Genome sequence data: Management, storage, and visualization. BioTechniques 209; 46(5): 333-336.

2. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015; 13(7): e1002195.

3. Portugal A, Balachandra R, Metz T, Bruskiewich R, McLaren G. International crop information system for germplasm data management. Methods Mol Biol. 2007; 406: 459-72.

4. McLaren C, Bruskiewich R, Portugal A, Cosico A. The International Rice Information System. A Platform for Meta-Analysis of Rice Crop Data. Plant Physiol. 2005; 139: 637-642.

5. Clarke FR, Yates SM, Clarke JM, DePauw RM, McLaren CG. The International Crop Information System: a useful tool for wheat breeding systems. In: Pogna NE, Romano M, Pogna EA, Galterio G, editors. Paestum: Proceedings of the Tenth International Wheat Genetics Symposium; 2003. P. 684-686.

6. Clarke FR, Clarke JM, Pozniak CJ, Yates SM. Application of the International Crop Information System for retrieval and usage of historic and phenotypic data for use in breeding applications. Rome: International Symposium Genetics and Breeding of Durum Wheat; 2013.

7. Generation Challenge Programme. 2009. http://mbp.generationcp.org/confluence/display/MBP/Application+2.2.1+Tool+2.10+-+Database+++++2.10.3+IBDB+Genotyping+Data+Management+System+(GDMS)+Schema+v1.0. Accessed 07 August 2017.

8. Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas W, et al. Flapjack – graphical genotype visualization. Bioinformatics 2010; 26(24):3133-3134.

9. MacDonald M. ClickOnce Deployment. In: Pro WPF in C# 2010: Berkeley: Apress; 2010. p. 1079-1097.

10. Bourque, B: 32-bit vs. 64-bit – What's the difference, and what does it mean for your PC? http://www.digitaltrends.com/computing/32-bit-64-bit-operating-systems. Accessed 03 March 2017.

Table 1. Comparison of existing plant breeding data management systems.

Softwares	LAN Install	Genealogy Management	Phenotypic Management	Genotypic Management	Support Available	In-house Development Opportunity	Cost
CIERA	Yes	Yes	Yes	Yes	Yes	Yes	Open source (free)
ICIS	Yes	Yes	Yes	No	No	Yes	Open source (free)
BMS	Yes	Yes	Yes	Limited	Yes	Yes	Annual subscription for some components
Katmandoo	No	Limited	Yes	Limited	Limited	Yes	Open source (free)
Agrobase	Yes	Yes	Yes	Limited	Yes	No	Multi-user license

Table 2. Data Mining Tool queries, description of query function, output format, and application of the data output.

Query Name	Description	Output	Use Case
Search for data by lines	Enter germplasm identifier individually and retrieve all associated phenotypic and genotypic data; option to add check germplasm lines from the phenotypic datasets	Microsoft Excel Tab-delimited text file XML file	Cross planning; Evaluation of germplasm over environments and time
Search for Phenotypic data by study/dataset	Gather whole selected phenotypic datasets	Microsoft Excel Tab-delimited text file XML file	Comparing all germplasm in a particular study, nursery or environment; Evaluating genetic populations; Collating data for reports
Search for data on lines in an existing list	Get phenotypic and genotypic data on all germplasm from a list; Get all synonyms for germplasm; Get an overview of all studies the germplasm in the list appear in	Microsoft Excel Tab-delimited text file XML file	Cross planning; Working with international introduction nurseries; Reports of registered variety performance; Quick repeat of queries as new datasets are added
Search for genetic dataset or genetic map	Gather whole selected genotypic datasets and/or genetic maps	Microsoft Excel Tab-delimited text file XML file Flapjack .dat and .map files	Evaluating genetic populations; Marker-assisted selection to identify favorable germplasm for crossing
Search for information on marker	Search for marker metadata, genotypic data and genetic maps	Microsoft Excel	Follow how germplasm respond to a marker

Figure legends

Fig. 1. CIERA database system. The data models handle data from phenotypic and genotypic studies, as well as the genealogy of the germplasm.

Fig. 2. CIERA Genealogy Management System (GMS). All genealogy information on cultivars, passport information and germplasm lists are stored in the GMS.

Fig. 3. CIERA Data Management System (DMS). Handles the data and study metadata for all types of phenotypic studies, including raw data and summaries.

Fig. 4. CIERA Genetic Data Management System (GDMS). Schema stores the marker metadata, data from genetic datasets (SNP, DArT, SSR and KASP), QTL information and marker-trait associations.

Fig. 5. Graphical user interface of the CIERA launcher. The screenshot is of the main menu when starting CIERA.

Fig. 6. CIERA Genetic Dataset Loader marker information editor: This is a screenshot of the user interface to edit or add marker metadata in CIERA.

Fig. 7. CIERA Field Book cover page: This is an example of the fieldbook cover page that cooperators fill out with study information, such as planting/harvest dates, fertilizer added, and any notes regarding the growing season.

Fig. 8. CIERA output preview/export screen: This is a screenshot of the output preview screen, where a user can have a quick look at their query results and choose which format they wish to export to.

Supplementary information

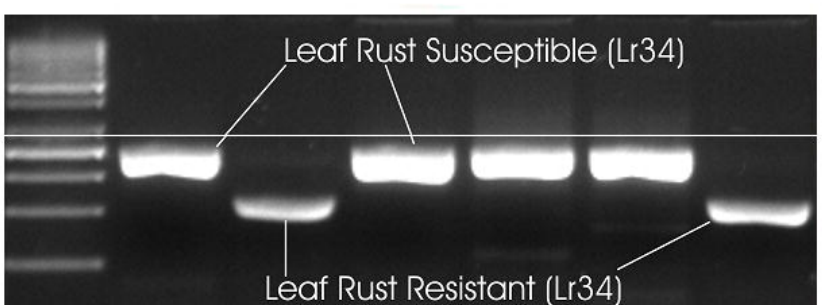
Fig. S1. Web version of CIERA: This is a screenshot of the CIERA web interface, which is still under construction.

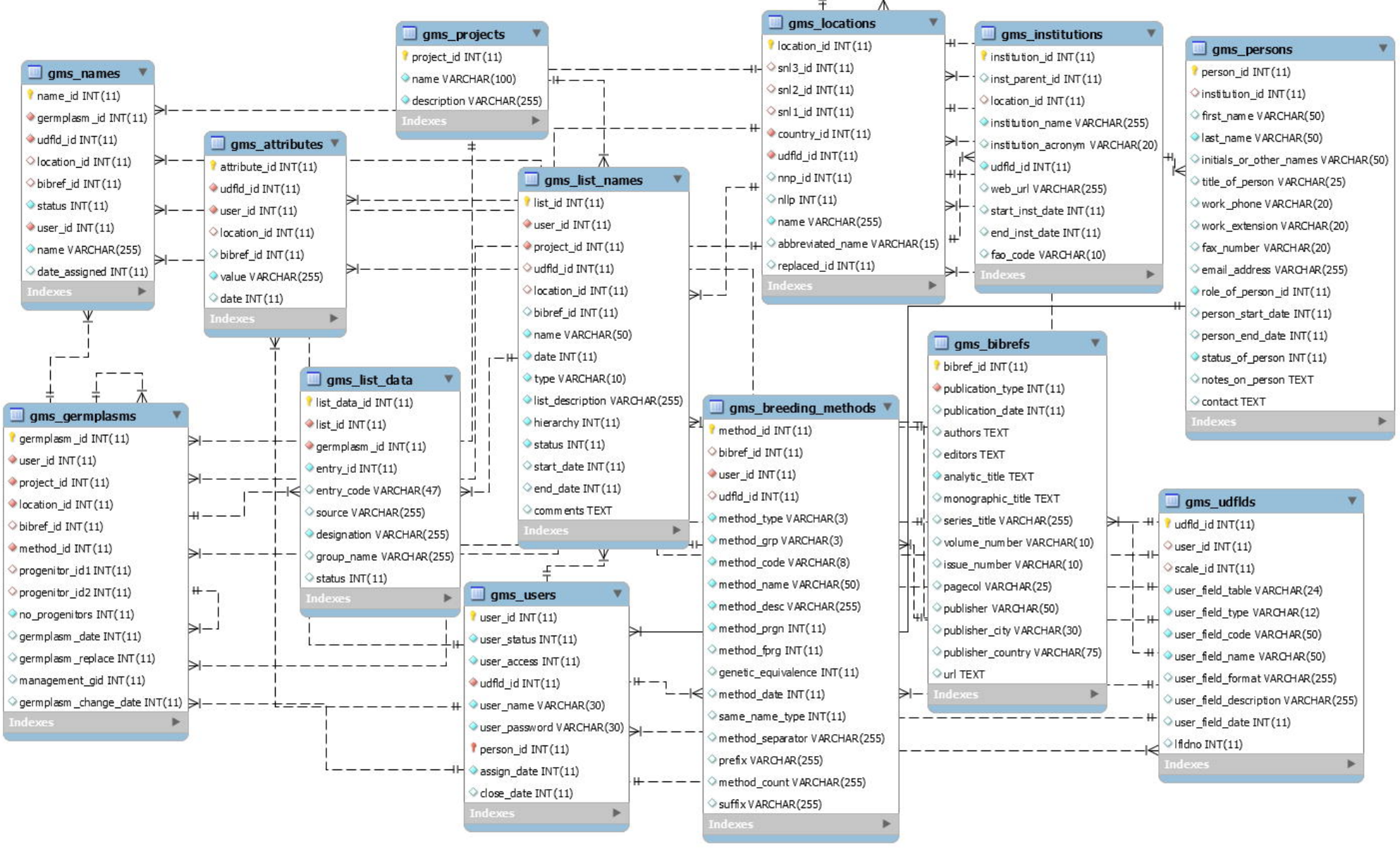
Genealogy Data

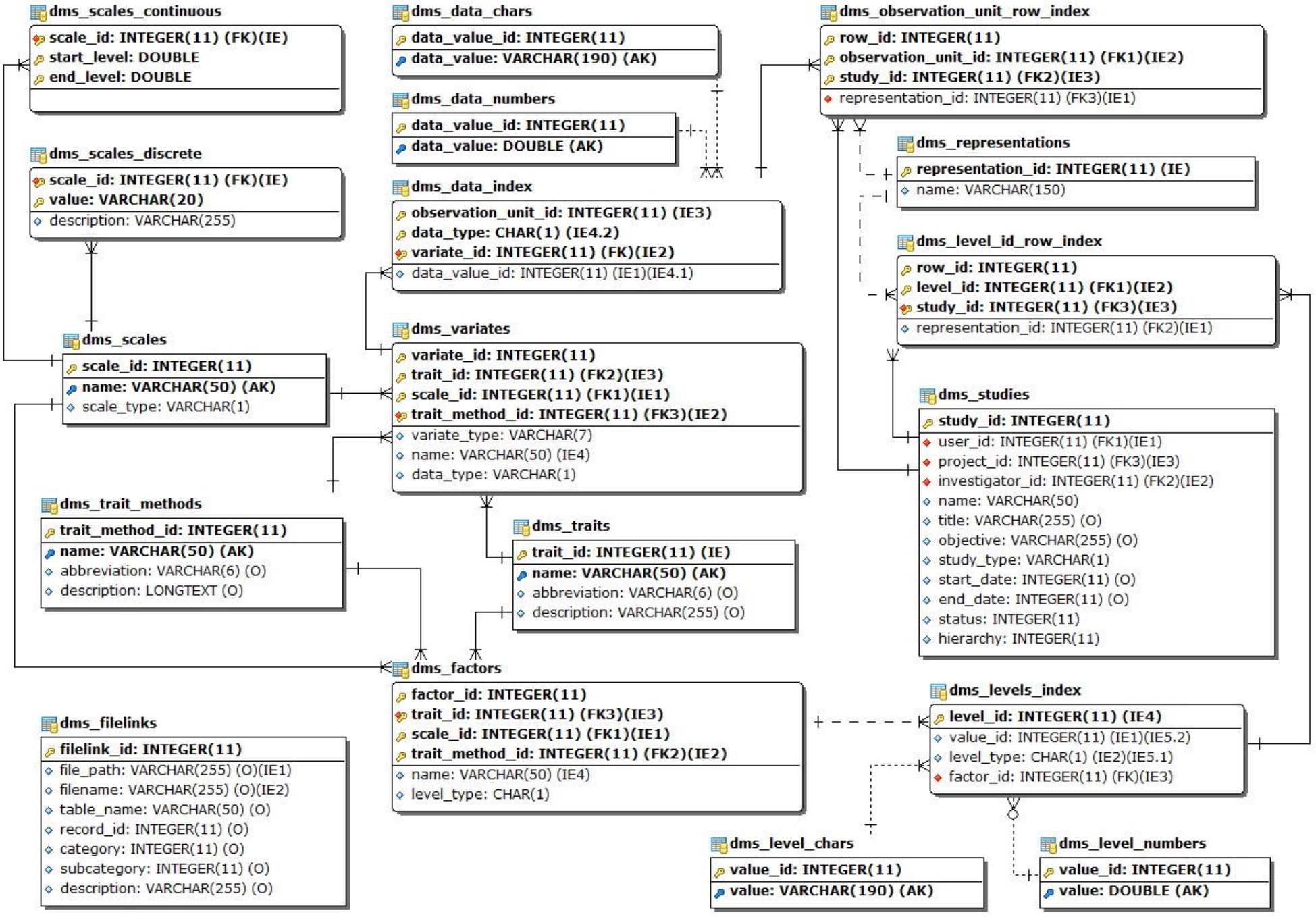
Agronomic Data

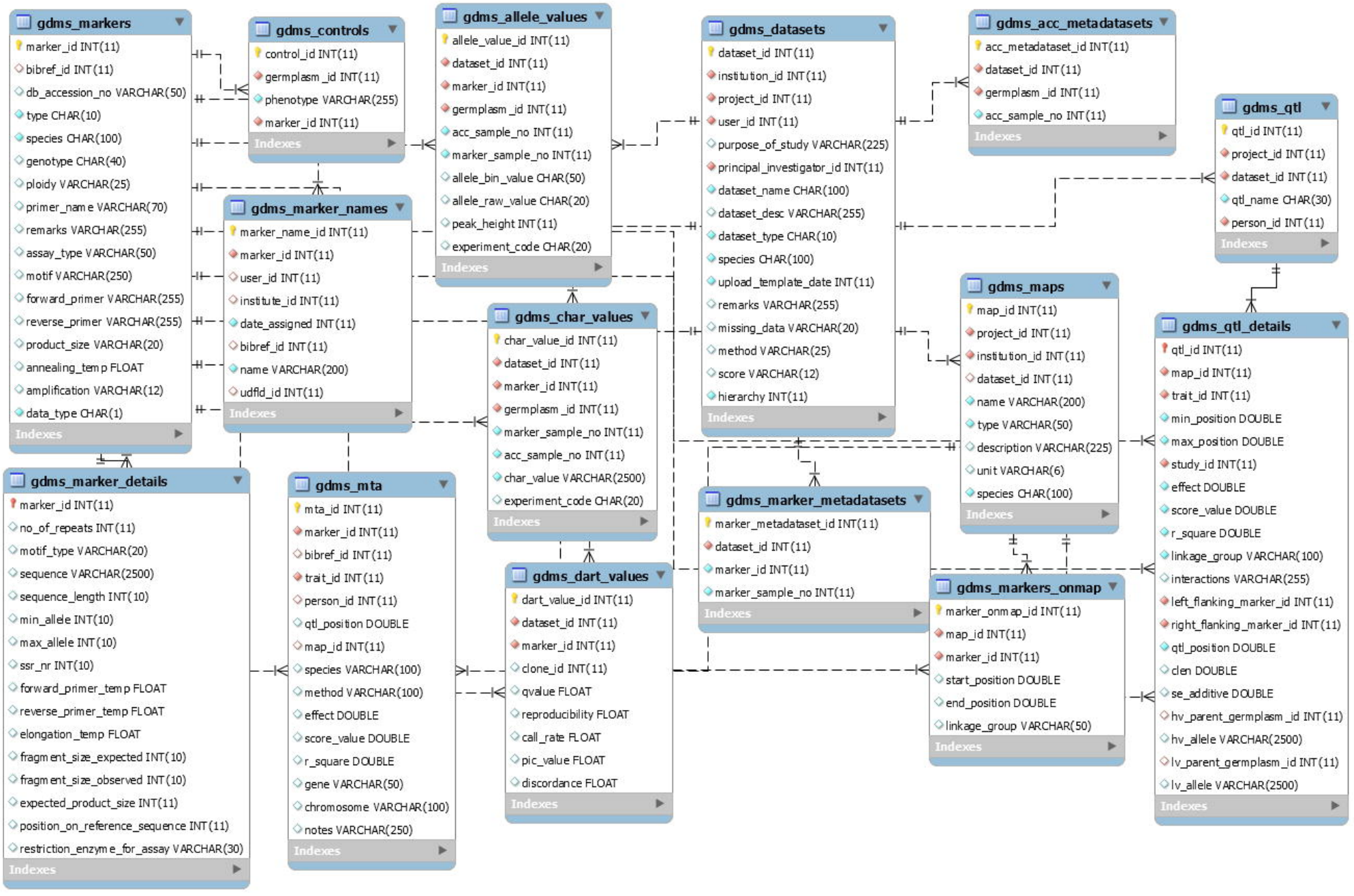
CIERA Database

Genotypic Data




gms_projects
project_id INT(11)
name VARCHAR(100)
description VARCHAR(255)
Indexes
gms_locations
location_id INT(11)
snl3_id INT(11)
snl2_id INT(11)
snl1_id INT(11)
country_id INT(11)
udfld_id INT(11)
nnp_id INT(11)
nllp INT(11)
name VARCHAR(255)
abbreviated_name VARCHAR(15)
replaced_id INT(11)
Indexes
gms_institutions
institution_id INT(11)
inst_parent_id INT(11)
location_id INT(11)
institution_name VARCHAR(255)
institution_acronym VARCHAR(20)
udfld_id INT(11)
web_url VARCHAR(255)
start_inst_date INT(11)
end_inst_date INT(11)
fao_code VARCHAR(10)
Indexes
gms_persons
person_id INT(11)
institution_id INT(11)
first_name VARCHAR(50)
last_name VARCHAR(50)
initials_or_other_names VARCHAR(50)
title_of_person VARCHAR(25)
work_phone VARCHAR(20)
work_extension VARCHAR(20)
fax_number VARCHAR(20)
email_address VARCHAR(255)
role_of_person_id INT(11)
person_start_date INT(11)
person_end_date INT(11)
status_of_person INT(11)
notes_on_person TEXT
contact TEXT
Indexes
gms_names
name_id INT(11)
germplasm_id INT(11)
udfld_id INT(11)
location_id INT(11)
bibref_id INT(11)
status INT(11)
user_id INT(11)
name VARCHAR(255)
date_assigned INT(11)
Indexes
gms_attributes
attribute_id INT(11)
udfld_id INT(11)
user_id INT(11)
location_id INT(11)
bibref_id INT(11)
value VARCHAR(255)
date INT(11)
Indexes
gms_list_names
list_id INT(11)
user_id INT(11)
project_id INT(11)
udfld_id INT(11)
location_id INT(11)
bibref_id INT(11)
name VARCHAR(50)
date INT(11)
type VARCHAR(10)
list_description VARCHAR(255)
hierarchy INT(11)
status INT(11)
start_date INT(11)
end_date INT(11)
comments TEXT
Indexes
gms_bibrefs
bibref_id INT(11)
publication_type INT(11)
publication_date INT(11)
authors TEXT
editors TEXT
analytic_title TEXT
monographic_title TEXT
series_title VARCHAR(255)
volume_number VARCHAR(10)
issue_number VARCHAR(10)
pagecol VARCHAR(25)
publisher VARCHAR(50)
publisher_city VARCHAR(30)
publisher_country VARCHAR(75)
url TEXT
Indexes
gms_germplasms
germplasm_id INT(11)
user_id INT(11)
project_id INT(11)
location_id INT(11)
bibref_id INT(11)
method_id INT(11)
progenitor_id1 INT(11)
progenitor_id2 INT(11)
no_progenitors INT(11)
germplasm_date INT(11)
germplasm_replace INT(11)
management_gid INT(11)
germplasm_change_date INT(11)
Indexes
gms_list_data
list_data_id INT(11)
list_id INT(11)
germplasm_id INT(11)
entry_id INT(11)
entry_code VARCHAR(47)
source VARCHAR(255)
designation VARCHAR(255)
group_name VARCHAR(255)
status INT(11)
Indexes
gms_users
user_id INT(11)
user_status INT(11)
user_access INT(11)
udfld_id INT(11)
user_name VARCHAR(30)
user_password VARCHAR(30)
person_id INT(11)
assign_date INT(11)
close_date INT(11)
Indexes
gms_breeding_methods
method_id INT(11)
bibref_id INT(11)
user_id INT(11)
udfld_id INT(11)
method_type VARCHAR(3)
method_grp VARCHAR(3)
method_code VARCHAR(8)
method_name VARCHAR(50)
method_desc VARCHAR(255)
method_prgn INT(11)
method_fprg INT(11)
genetic_equivalence INT(11)
method_date INT(11)
same_name_type INT(11)
method_separator VARCHAR(255)
prefix VARCHAR(255)
method_count VARCHAR(255)
suffix VARCHAR(255)
Indexes
gms_udflds
udfld_id INT(11)
user_id INT(11)
scale_id INT(11)
user_field_table VARCHAR(24)
user_field_type VARCHAR(12)
user_field_code VARCHAR(50)
user_field_name VARCHAR(50)
user_field_format VARCHAR(255)
user_field_description VARCHAR(255)
user_field_date INT(11)
lfldno INT(11)
Indexes


dms_scales_continuous
scale_id: INTEGER(11) (FK)(IE)
start_level: DOUBLE
end_level: DOUBLE
dms_data_chars
data_value_id: INTEGER(11)
data_value: VARCHAR(190) (AK)
dms_observation_unit_row_index
row_id: INTEGER(11)
observation_unit_id: INTEGER(11) (FK1)(IE2)
study_id: INTEGER(11) (FK2)(IE3)
representation_id: INTEGER(11) (FK3)(IE1)
dms_data_numbers
data_value_id: INTEGER(11)
data_value: DOUBLE (AK)
dms_representations
representation_id: INTEGER(11) (IE)
name: VARCHAR(150)
dms_scales_discrete
scale_id: INTEGER(11) (FK)(IE)
value: VARCHAR(20)
description: VARCHAR(255)
dms_data_index
observation_unit_id: INTEGER(11) (IE3)
data_type: CHAR(1) (IE4.2)
variate_id: INTEGER(11) (FK)(IE2)
data_value_id: INTEGER(11) (IE1)(IE4.1)
dms_level_id_row_index
row_id: INTEGER(11)
level_id: INTEGER(11) (FK1)(IE2)
study_id: INTEGER(11) (FK3)(IE3)
representation_id: INTEGER(11) (FK2)(IE1)
dms_scales
scale_id: INTEGER(11)
name: VARCHAR(50) (AK)
scale_type: VARCHAR(1)
dms_variates
variate_id: INTEGER(11)
trait_id: INTEGER(11) (FK2)(IE3)
scale_id: INTEGER(11) (FK1)(IE1)
trait_method_id: INTEGER(11) (FK3)(IE2)
variate_type: VARCHAR(7)
name: VARCHAR(50) (IE4)
data_type: VARCHAR(1)
dms_studies
study_id: INTEGER(11)
user_id: INTEGER(11) (FK1)(IE1)
project_id: INTEGER(11) (FK3)(IE3)
investigator_id: INTEGER(11) (FK2)(IE2)
name: VARCHAR(50)
title: VARCHAR(255) (O)
objective: VARCHAR(255) (O)
study_type: VARCHAR(1)
start_date: INTEGER(11) (O)
end_date: INTEGER(11) (O)
status: INTEGER(11)
hierarchy: INTEGER(11)
dms_trait_methods
trait_method_id: INTEGER(11)
name: VARCHAR(50) (AK)
abbreviation: VARCHAR(6) (O)
description: LONGTEXT (O)
dms_traits
trait_id: INTEGER(11) (IE)
name: VARCHAR(50) (AK)
abbreviation: VARCHAR(6) (O)
description: VARCHAR(255) (O)
dms_factors
factor_id: INTEGER(11)
trait_id: INTEGER(11) (FK3)(IE3)
scale_id: INTEGER(11) (FK1)(IE1)
trait_method_id: INTEGER(11) (FK2)(IE2)
name: VARCHAR(50) (IE4)
level_type: CHAR(1)
dms_levels_index
level_id: INTEGER(11) (IE4)
value_id: INTEGER(11) (IE1)(IE5.2)
level_type: CHAR(1) (IE2)(IE5.1)
factor_id: INTEGER(11) (FK)(IE3)
dms_filelinks
filelink_id: INTEGER(11)
file_path: VARCHAR(255) (O)(IE1)
filename: VARCHAR(255) (O)(IE2)
table_name: VARCHAR(50) (O)
record_id: INTEGER(11) (O)
category: INTEGER(11) (O)
subcategory: INTEGER(11) (O)
description: VARCHAR(255) (O)
dms_level_chars
value_id: INTEGER(11)
value: VARCHAR(190) (AK)
dms_level_numbers
value_id: INTEGER(11)
value: DOUBLE (AK)


gdms_markers
marker_id INT(11)
bibref_id INT(11)
db_accession_no VARCHAR(50)
type CHAR(10)
species CHAR(100)
genotype CHAR(40)
ploidy VARCHAR(25)
primer_name VARCHAR(70)
remarks VARCHAR(255)
assay_type VARCHAR(50)
motif VARCHAR(250)
forward_primer VARCHAR(255)
reverse_primer VARCHAR(255)
product_size VARCHAR(20)
annealing_temp FLOAT
amplification VARCHAR(12)
data_type CHAR(1)
Indexes
gdms_controls
control_id INT(11)
germplasm_id INT(11)
phenotype VARCHAR(255)
marker_id INT(11)
Indexes
gdms_allele_values
allele_value_id INT(11)
dataset_id INT(11)
marker_id INT(11)
germplasm_id INT(11)
acc_sample_no INT(11)
marker_sample_no INT(11)
allele_bin_value CHAR(50)
allele_raw_value CHAR(20)
peak_height INT(11)
experiment_code CHAR(20)
Indexes
gdms_datasets
dataset_id INT(11)
institution_id INT(11)
project_id INT(11)
user_id INT(11)
purpose_of_study VARCHAR(225)
principal_investigator_id INT(11)
dataset_name CHAR(100)
dataset_desc VARCHAR(255)
dataset_type CHAR(10)
species CHAR(100)
upload_template_date INT(11)
remarks VARCHAR(255)
missing_data VARCHAR(20)
method VARCHAR(25)
score VARCHAR(12)
hierarchy INT(11)
Indexes
gdms_acc_metadatasets
acc_metadataset_id INT(11)
dataset_id INT(11)
germplasm_id INT(11)
acc_sample_no INT(11)
Indexes
gdms_qtl
qtl_id INT(11)
project_id INT(11)
dataset_id INT(11)
qtl_name CHAR(30)
person_id INT(11)
Indexes
gdms_marker_names
marker_name_id INT(11)
marker_id INT(11)
user_id INT(11)
institute_id INT(11)
date_assigned INT(11)
bibref_id INT(11)
name VARCHAR(200)
udfld_id INT(11)
Indexes
gdms_char_values
char_value_id INT(11)
dataset_id INT(11)
marker_id INT(11)
germplasm_id INT(11)
marker_sample_no INT(11)
acc_sample_no INT(11)
char_value VARCHAR(2500)
experiment_code CHAR(20)
Indexes
gdms_maps
map_id INT(11)
project_id INT(11)
institution_id INT(11)
dataset_id INT(11)
name VARCHAR(200)
type VARCHAR(50)
description VARCHAR(225)
unit VARCHAR(6)
species CHAR(100)
Indexes
gdms_qtl_details
qtl_id INT(11)
map_id INT(11)
trait_id INT(11)
min_position DOUBLE
max_position DOUBLE
study_id INT(11)
effect DOUBLE
score_value DOUBLE
r_square DOUBLE
linkage_group VARCHAR(100)
interactions VARCHAR(255)
left_flanking_marker_id INT(11)
right_flanking_marker_id INT(11)
qtl_position DOUBLE
clen DOUBLE
se_additive DOUBLE
hv_parent_germplasm_id INT(11)
hv_allele VARCHAR(2500)
lv_parent_germplasm_id INT(11)
lv_allele VARCHAR(2500)
Indexes
gdms_marker_details
marker_id INT(11)
no_of_repeats INT(11)
motif_type VARCHAR(20)
sequence VARCHAR(2500)
sequence_length INT(10)
min_allele INT(10)
max_allele INT(10)
ssr_nr INT(10)
forward_primer_temp FLOAT
reverse_primer_temp FLOAT
elongation_temp FLOAT
fragment_size_expected INT(10)
fragment_size_observed INT(10)
expected_product_size INT(11)
position_on_reference_sequence INT(11)
restriction_enzyme_for_assay VARCHAR(30)
Indexes
gdms_mta
mta_id INT(11)
marker_id INT(11)
bibref_id INT(11)
trait_id INT(11)
person_id INT(11)
qtl_position DOUBLE
map_id INT(11)
species VARCHAR(100)
method VARCHAR(100)
effect DOUBLE
score_value DOUBLE
r_square DOUBLE
gene VARCHAR(50)
chromosome VARCHAR(100)
notes VARCHAR(250)
Indexes
gdms_dart_values
dart_value_id INT(11)
dataset_id INT(11)
marker_id INT(11)
clone_id INT(11)
qvalue FLOAT
reproducibility FLOAT
call_rate FLOAT
pic_value FLOAT
discordance FLOAT
Indexes
gdms_marker_metadatasets
marker_metadataset_id INT(11)
dataset_id INT(11)
marker_id INT(11)
marker_sample_no INT(11)
Indexes
gdms_markers_onmap
marker_onmap_id INT(11)
map_id INT(11)
marker_id INT(11)
start_position DOUBLE
end_position DOUBLE
linkage_group VARCHAR(50)
Indexes

Crop Information Engine and Research Assistant

Crop Information Engine and Research Assistant

Lookup Data:

Pedigree Information

Phenotypic and Genetic Data

Helpful Tools:

Tools to assist in seeding/harvest setup

Load/Edit Data:

Manage Data

Data Analysis:

Analysis of data

CIERA

Version 2.0.0.0

Crop Information Engine and Research Assistant

Enter the marker name:

Search (if looking for existing marker):

Marker Type:

Assay Type:

Ploidy:

Remarks:

Motif:

Forward Primer:

Reverse Primer:

Product Size:

Annealing Temp:

Amplification:

No. of Repeats:

Motif Type:

Sequence:

Sequence Length:

Min. Allele:

Max. Allele:

No. Identified SSRs:

Fwd Primer Temp:

Rvs. Primer Temp:

Elongation Temp:

Fragment Size Exp:

Fragment Size Obs:

Exp. Product Size:

Alias (",," between):

Allele Type:

Reference:

Species:

Genotype:

Accession ID:

Primer Name

Marker Information

Add New Marker

Update Marker Info

Delete Marker

Test Name: **Durum Coop**

Fill in the following spreadsheet and email with the agronomic data to: yuefend.ruan@agr.gc.ca

Start Date (yyyymmdd)	20170223	End Date (yyyymmdd)	20170223	Year	2016
Test Design	6*5 Lattice			Location	AAFC INDIAN HEAD RESEARCH FARM
Repetitions				Cooperator	
Replications	3			Email	
Blocks	6			Coordinator	Yuefeng Ruan (306) 770-4506
Treatments	5				Semi-arid Prairie Agricultural Research Centre
Entries	30				
Total Plots	90				

Seeding

	day	month	year		
SOWING DATE	13	04	2016	or Julian Date	

	sparse
	normal
	dense

Emergence

Relative to long-term norm: early / normal / late

DELAYED BY DRY SEED BED (Y/N)?

HARVEST

day	month	year		
27	09	2016	or Julian Date	

PLOT DIMENSIONS

	SOWN		HARVESTED	
number of rows			number of rows	
length of rows		m	length of rows	
space between rows		m		
cover crop between plots (Y/N)?				

plot area sown		m2
plot area harvested		m2
conversion factor		

YIELD GIVEN IN: g/plot ____ kg/plot ____ kg/ha ____

OTHER COMMENTS, PROBLEMS AND OBSERVATIONS ON PLANT STRESSES (EXCLUDING WEATHER)

WEATHER (GENERAL COMMENT, ESPECIALLY DEVIATIONS FROM NORMAL)

FERTILIZER APPLIED

Specify unit of fertilizer applied if different from kg/ha

Crop Information Engine and Research Assistant

Here is a preview of the data requested. Please choose an export option below.

Phenotypic Data

YEAR	STUDY	NURSERY	LOCATION ID	GID	DESIGNATION	ALTERNATE NAMES
2012	Pheno Demo Data	YIELD TRIAL	15988	5811041	5600HR	BW 238
2012	Pheno Demo Data	YIELD TRIAL	15988	600012181	5702PR	99S3008-05, HY977
2012	Pheno Demo Data	YIELD TRIAL	15988	261373	8021-V2	AUS-24546, 8021V2, SC 8021-V2, 8021-V-2, KENYA 321/
2012	Pheno Demo Data	YIELD TRIAL	15988	199089	AC BARRIE	BARRIE, PI-593658, 8405-BK3C, AC BARRIE, BW 661, C03
2012	Pheno Demo Data	YIELD TRIAL	15988	3825962	AC CADILLAC	CWI75580, L8900-AL2A10, BW 689, C0600C-5007, C110(
2012	Pheno Demo Data	YIELD TRIAL	15988	3819618	AC CRYSTAL	CWI75585, 8828-BG2, HY 417, C0500F-117, C0600C-512
2012	Pheno Demo Data	YIELD TRIAL	15988	199092	AC DOMAIN	BW35189, K-62919, BW 148, RL 4630, ###188386, DOM.

- [x] Microsoft Excel
- [ ] Comma delimited text file
- [ ] XML file

No genetic data chosen, so nothing to preview

- [ ] Microsoft Excel
- [ ] Comma delimited text file
- [ ] XML File
- [ ] FlapJack Input Files

Next