Final report

Female Reproduction PhenoBank: a database to facilitate the delivery of genomics technologies towards improving the fertility of Northern Australian cattle.

Project code:	L.GEN.1710
Prepared by:	Marina Fortes The University of Queensland
Date published:	December 2021
Research collaborators:	Ben Hayes (UQ-QAAFI); Toni Reverter, Laercio Porto Neto, and Sigrid Lehnert (CSIRO); Gehan Jayawardhana; Gretel Bailey Preston, and Tim Schatz (NTDITT); Cody Wilson, and Bryce Little (ABRI/CSIRO)

PUBLISHED BY
Meat & Livestock Australia Limited
PO Box 1961
NORTH SYDNEY NSW 2059

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Abstract

Female reproduction traits are heritable and can be improved with DNA-informed breeding decisions (i.e., genomic selection). The female reproduction PhenoBank project was undertaken to collect datasets that are informative for building a reference population, which will facilitate genomic selection in Northern Australian cattle. The datasets in this project were collected with support from beef producers that volunteered the performance records of their cows, together with a hair or tissue sample for DNA analyses. The used chip technology for genotyping DNA, followed by data analyses discovered associated DNA markers and computed the accuracy of genomic selection for reproduction traits. The traits used were relatively easy to measure, to facilitate future adoption. They were based on pregnancy outcomes of the first and second breeding seasons. Heifer pregnancy outcomes and their rebreeding capacity were heritable traits, within the expectations for female reproduction traits. Prediction accuracies for these traits demonstrates their potential in terms of using DNA data to improve reproductive efficiency. The PhenoBank project has shown that it is possible to work directly with producers to collect useful datasets for female reproduction traits, which can boost reference populations, increase prediction accuracies, and facilitate the adoption of genomic technologies. The performance records matched with genotypes formed a reference population with 10,507 cows. This reference population benefits the industry as it can be used to derive prediction equations for genomic estimated breeding values. The PhenoBank reference population may be integrated with other projects and platforms to aid genomic selection for reproduction traits. To facilitate data integration, management, storage and sharing, this project also developed a software solution: the PhenoBank database. The PhenoBank database can be accessed via a windows application or a web-based platform. Access is controlled through usernames and passwords that give permission to view and download data. The reference population and the database that can house phenotypes and genotypes are the main outputs of this project. These outputs can be used to develop a framework for engaging with producers that informs and facilitates the use of genomic technologies. Now that we have the infrastructure and capability, it is possible to add new data and new users so that PhenoBank grows to enhance its impact.

Executive summary

Background

Female reproduction performance is a major driver of on-farm profitability and currently has large potential for genetic improvement. Sub-optimal reproduction rates in Australia's tropical and subtropical beef herds are a complex and multifactorial problem and has a significant negative impact on the productivity of Australian beef. Traditional obstacles for genetic improvement through selective breeding for female reproductive traits in extensively managed northern herds include the difficulty of collecting accurate farm records and the low heritability of female reproduction traits.

Research has demonstrated the benefits gained in tropical beef cattle breeds by selecting for heritable, early-in-life female reproduction traits, such as age at puberty and post-partum anoestrus interval after first calving. Genomic selection for heritable traits offers a promising tool for achieving genetic gain but requires large reference populations for the identification and validation of predictive genetic markers. Such reference populations, with measured phenotypes, do not exist for reproductive traits in many tropical breeds relevant in northern Australia.

This project aimed to merge the data from existing research and industry beef herds with reproductive phenotype records into a single, easy to access platform — the Female Reproduction PhenoBank database — and to demonstrate its utility for achieving genetic progress. PhenoBank will serve as a large reference population to underpin the discovery and validation of predictive genetic markers for female reproduction traits. The resources created in this project benefit all producers through genetic improvement of beef cattle in the northern Australian herd.

Objectives

This project achieved all four of its stated objectives:

Objective 1: Create the Female Reproduction PhenoBank. Multiple datasets were integrated into one database with a standardised vocabulary to define female reproduction traits. Custom software was developed to integrate the multiple datasets of phenotypes and genotypes and provide easy, yet controlled access for end-users (i.e., PhenoBank collaborators) of the data. PhenoBank currently has integrated and matched female reproduction phenotypes and genotypes for 10,507 cows. This level of genome-wide data for tropical breeds is a world-class resource.

Objective 2: Identification of functional mutations for female reproduction traits. The PhenoBank database was used to identify mutations associated with female reproduction traits, and confirmed that female reproduction phenotypes are polygenic traits, with multiple associated mutations, all with relatively small effects individually.

Objective 3: Validation and Industry Herds. The mutations identified in Objective 2 were used to inform the subsequent genomic selection analyses and in this context, they were validated. This analysis adds evidence, or proof of concept, to the idea of using associated mutations to enhance the predictions in industry datasets. The measurable output of this work is estimated accuracies of genomic Estimated Breeding Values (gEBVs) across independent cattle populations.

Objective 4: Database protocol. The establishment of an appropriate infrastructure for PhenoBank data storage required standardised protocols for importing, manipulating, and exporting data, including meta-data to ensure source identification and appropriate use. A PhenoBank manual was developed.

Methodology

Collaboration between research partners was accomplished through a series of workshops to establish the terms of reference for data sharing, the scoping of database requirements and a standardised structure for data merging.

Genotypes from medium density TropBeef SNP chip (NEOGEN) were imputed to high density (BovineHD illumine chip with approx. 770 thousand DNA markers) and to whole genome sequence level (approx. 15 million DNA markers). Animals from existing research herds were included in PhenoBank if they had useful phenotype data and a retrievable DNA sample for genotyping, or genotype already available. Legacy data from the CRC for Beef Genetic Technologies (Beef CRC) was included in PhenoBank, with new phenotypes developed for those cows so that they could match the new industry datasets. Animals from existing and new industry herds were integrated into the PhenoBank database.

A standardized nomenclature for female reproduction phenotypes was described and used to facilitate the merge of datasets. New phenotypes for early in life reproduction traits were described and existing raw phenotype data for each animal were recoded to create the new phenotypes. The aim for the new phenotypes was to be relatively easy-to-measure so that we could use research data in conjunction with performance records volunteered by beef producers. Performance records were the outcomes of the first two breeding seasons, and they underpin the two focus traits for PhenoBank: PREG1 (pregnancy outcome of the first breeding season) and REB (a rebreeding score that considers the pregnancy outcome of both the first and the second breeding seasons). Producers who collaborated with PhenoBank needed only three pieces of information: pregnancy diagnosis for the first two seasons, a record of cow crop (year and season of birth), and a sample for DNA analyses. Keeping the data requests simple allowed more producers to participate.

Genome-wide association studies (GWAS) were conducted using a subset of PhenoBank data to identify single nucleotide polymorphisms (SNP) associated with the newly defined phenotypes. SNPs are the most common type of DNA markers. They are used frequently in GWAS and genomic selection alike. Genomic selection analyses of PhenoBank data explored two current methods (GBLUP and Bayes R) and future work could expand by investigating additional methodologies.

Results/key findings

PhenoBank database was constructed and populated with the reproduction phenotypes and genotypes for 10,507 cows from a combination of research and industry herds. Industry herds were contributed by producers who became PhenoBank collaborators. The software solution created for PhenoBank is a cloud-based system, which uses the CSIRO Livestock Information Platform (LIP) to store the data. Both windows and web-based front-end applications were developed for PhenoBank to access LIP. The applications were developed in consultation with all PhenoBank researchers, and the software improved in every interaction. User access is controlled by the administrator, currently this is ABRI in partnership with CSIRO. Each end user has a username and password, which they can use to upload and retrieve data safely. Upon upload, the user will assign privilege to others and therefore control who can access their data. As PhenoBank is currently populated with project data, all PhenoBank users are researchers that can download the entire data for research purposes. Genotypes stored in the PhenoBank database can be retrieve in a format that is easy to analyse (i.e., PLINK files). Genotypes and phenotypes on PhenoBank are linked through unique individual identifiers, created for each animal. The database can be used to search, merge, and download data for analyses. Now that we have the infrastructure and capability, it is possible to add new data and new users so that PhenoBank grows to enhance its impact.

Preliminary genomic analyses using HD data identified promising genetic markers. Genomic models for the two traits, PREG1 and REB, estimated heritabilities of 0.17 and 0.21 respectively. The heritabilities of the traits used were commensurate with those previously reported for similar female reproduction traits. These heritabilities confirm the possibility of targeting these industry-based traits for selection. Moreover, the accuracies of genomic selection improved when GWAS results were used to inform the predictions. Prediction accuracy estimates for these traits demonstrates the future potential for genomic technologies to improve reproductive efficiency and transform the productivity of Australia's northern beef herd.

Benefits to industry

The northern beef industry now has a solid platform to underpin substantial genetic gains in female reproduction traits. The database created and the data assembled can be made available to cattle geneticists in Australia to improve their predictions of genetic merit for female reproduction traits. PhenoBank resources can inform the use of genomic data for breeding tropically adapted cattle. The main audiences for this research are beef producers, especially breeding herds of tropically adapted cattle in Northern Australia, and geneticists seeking to provide information and service to these producers.

Future research and recommendations

This project had showcased how engagement with beef producers to collect on-farm measurements related to female reproduction performance can promote continuous growth of the reference population. It is foreseen that producer engagement through projects such as PhenoBank, which target easy-to-measure traits, creates a platform for ongoing communications between researchers and producers that may enhance adoption of genomic technologies. For example, producers benefited from DNA diagnostics, such as the Poll/Horn test, even before genomic analyses were carried.

Four key recommendations may be put forward: 1) to continue research in this area by using the full PhenoBank data, 2) to further develop the PhenoBank database so that it can store sequence level data (i.e. millions of DNA markers per animal), 3) to investigate if PhenoBank resources can be integrated with other projects to further enhance the reference population available for female reproduction traits, and 4) to provide the herd specific phenotypes and genotypes back to the producer who volunteered their data and samples. Some producers have already requested their cattle data because their breed association can use the genotypes generated for PhenoBank to provide gEBVs via BREEDPLAN. In short, we recommend that geneticists in Australia and the industry at large are encouraged to use the PhenoBank resources.

Table of contents

1 Background 8

2 Objectives 11

2.1 Objective 1: Female Reproduction PhenoBank Database 11

2.2 Objective 2: Identify functional mutations for female reproduction traits 11

2.3 Objective 3: Validate genomic markers in different herds 11

2.4 Objective 4: Design an appropriate database protocol 11

3 Methodology 12

3.1 Develop a working collaboration 12

3.1.1 Background Intellectual Property (IP) and permissions to access data 12

3.2 Scoping of database solutions for the PhenoBank 13

3.3 Genotypes 13

3.4 Imputing genotype data 14

3.5 Choosing animals for inclusion in PhenoBank 15

3.6 Redefining existing phenotypes 16

3.7 Genomic analyses 16

3.7.1 Genotypes 16

3.7.2 Principal component analysis and genomic relationship matrices 16

3.7.3 GWAS (Genome-wide association studies) 17

3.7.4 GBLUP genomic predictions 17

3.7.5 Bayes R model genomic predictions 17

3.7.6 Genomic prediction accuracy 18

4 Results 18

4.1 Research partner agreements 18

4.2 Data storage solutions 18

4.3 Dictionary of traits 19

4.4 Data entry into PhenoBank 22

4.5 Heritability of farm-based phenotypes 23

4.6 Genetic analyses 24

4.6.1 Genotype imputation 24

4.6.2 Identification of SNP associated with female reproduction traits 24

4.6.3 Genomic selection predictions and accuracies 26

5 Conclusion 30

5.1 Key findings 30

5.2 Benefits to industry 31

6 Future research and recommendations 32

6 References 33

7 Appendix 1: Summary and Manual for PhenoBank Database 34

1 Background

The ideal beef cow will calve before 2 years of age, maintain a calving interval of approximately one year, and successfully wean a calf each year (Engelken, 2008). She will also be tolerant of environmental stressors, diseases, and parasites, and require minimal assistance when calving or raising her young (Snelling et al., 2012). Cows in the northern beef herd in tropical and sub-tropical Australia generally do not meet this ideal. To tolerate the harsh environmental stressors, the beef breeding herd in Northern Australia is strongly influenced by *Bos indicus* genetics. Characteristically, *Bos indicus* cattle are older at puberty and have prolonged post-partum anoestrus: two key traits that lead to lower reproduction rates and a reduced rate of genetic gain for all traits of interest.

Sub-optimal reproduction rates have a negative impact on the productivity of Australian beef, especially in northern regions. This problem is complex and multifactorial. Heat stress, poor quality forage, infectious diseases, and parasites, phosphorous deficiency, and cattle genetics are all factors that influence reproduction rates (McGowan et al., 2014; Moore et al., 2021). Reproduction rates are calculated based on breeding outcomes divided by the number of cows exposed in each season. For example, the percentage of pregnant cows after a breeding season or the calculation of weaning rates (number of calves weaned per number of cows exposed) serve as indicators of herd reproductive performance. Heifer pregnancy rates are an early indicator of female reproduction performance.

Female reproduction performance is a major driver of on-farm profitability and currently has large potential for genetic improvement, as many northern properties have low reproductive rates (McGowan et al., 2014; Harburg et al., 2020). Recent modelling demonstrated that genetic selection targeting heifer pregnancy rates leads to economic gains in Northern Australian beef herds (Harburg et al., 2020). This is a recent study and heifer pregnancy rates are not yet used for selective breeding in Australia.

Female reproduction traits used for selective breeding are typically derived from farm records. In stud herds, pregnancy diagnostics, controlled mating seasons and recording of birth dates for every animal create a rich resource that is used in BREEDPLAN to calculate 'days to calve' estimated breeding values (EBV). Using the "days to calve" EBV, it is possible to drive genetic improvement for better reproduction rates (Graser et al., 2005; Moore et al., 2021). However, not all farmers are capable of mothering up and providing precise data for each calf. In addition, female reproduction traits are often low in heritability and/or are expensive to measure, which are hurdles for the adoption of genetic improvement through selection alone (Cammack et al., 2009). These hurdles make female reproduction traits an ideal target for genomic technologies and integrated approaches.

Research has demonstrated that substantial genetic gain leading to improved weaning rates in tropical beef cattle breeds can be achieved by focusing on early-in-life female reproduction traits (Johnston et al., 2010; Johnston et al., 2013). Age at puberty and post-partum anoestrous interval are heritable traits in tropical beef cattle (Johnston et al., 2009; Johnston et al., 2010), making genomic selection for these traits a promising strategy for improving reproduction rates in the tropically adapted cattle typically grown under extensive conditions in Northern Australia. In addition, the accuracy of genomic selection is often greater than the accuracy achieved with pedigree-based selection (Goddard and Hayes, 2007; Goddard et al., 2010; Goddard et al., 2011; Erbe et al., 2012; Bolormaa et al., 2013; Bolormaa et al., 2014). A requirement to establish a robust genomic selection program is to have a reference population, with phenotypes and genotypes, available. Lacking a reference population for female reproduction traits in relevant breeds is a major impediment for establishing a genomic selection program in Northern Australia.

Genetic markers, such as SNP, are valuable tools for genomic selection of cattle, especially if they are effective across breeds and particularly if the functional mutation underpinning the variation of reproduction traits is found (for a review see: Weller and Ron, 2011). Advances in genomic technologies now allow thousands of SNP markers to be assayed simultaneously in a single chip. Functional mutations in key genes can be used to increase the accuracy of genomic selection (Snelling et al., 2013), reduce the necessity of recalibration of prediction equations and facilitate the adoption of genomic tools across breeds. Combining genome-wide association studies (GWAS) and sequenced genomes has empowered the search for functional mutations associated with bull reproduction (completed MLA project B.NBP.0786). This work led to the discovery of the key gene *TEX11*, identified as harbouring a functional mutation associated with percentage of normal sperm and scrotal circumference in both Brahman and Tropical Composite bulls (Lyons et al., 2014). Similar work targeting female reproduction traits has merit. The PhenoBank project is making it possible.

Genomic selection accuracy is dependent on the identification of predictive genetic markers that have been extensively validated across a representative genetic pool of animals. Such validation requires large reference populations with both phenotypic performance records and DNA genotypes. Large reference populations exist for Holstein cattle and underpin the progress that the dairy industry achieved with genomic selection (Weller et al., 2017). A similar resource is not available across the many tropical beef breeds that are relevant in northern Australia and this deficit constrains the use of genomic selection to improve reproduction rates in northern herds.

While a large reference population does not currently exist for tropical beef breeds, there are many existing datasets from previous research projects — such as Beef CRCs, breed specific information nucleus (BINs), Repronomics, Northern Genomics, SmartState and others — with extensive phenotype records for female reproduction traits, and many with existing DNA genotypes. These datasets, if appropriately merged, could fast track the creation of a large reference population capable of underpinning the development of robust genomic Estimated Breeding Values (gEBVs), which would transform the rate of genetic gain in the northern beef herd (Meuwissen et al., 2013). Once established and validated, gEBVs are more likely to be adopted in Northern Australia, because they reduce the need for labour intensive data collection (e.g., birth date, and lifetime reproductive performance). In addition, if newly developed gEBVs consider phenotypes that are easier to measure, such as heifer pregnancy rates, more producers will be able to supply data and benefit from the selection program. Albeit less precise, heifer pregnancy is correlated with ‘days to calve’.

This project aimed to create a reference population for female reproduction traits in Brahman and composite breeds of cattle. This reference population is at the core of the Female Reproduction PhenoBank. A software tool for easy access to the assembled data was proposed because the construction of one unified database will greatly improve the capacity and outputs possible, when compared to working with any of the individual datasets. At completion, the project will have gathered reproduction phenotypes and genotypes for at least 10,000 cows, comprising existing and new datasets. By utilising existing and new datasets, the project links with research to reduce the establishment time and costs of assembling a reference population. PhenoBank aims to provide an easy-to-use data sharing platform, to transfer information to the beef industry.

This project also aimed to interrogate the compiled PhenoBank data to explore its utility for genomic selection of two early-in-life and easy to measure traits: heifer pregnancy outcomes after the first mating season (PREG1, success = 1, or failure = 0) and rebreeding score (REB, score from 1 to 4) defined by the outcomes of the first and second mating seasons. Investigating these traits provides highly valuable information to underpin the future commercial roll-out of cost-effective genomic technologies in Australia’s northern beef herd. To help drive genetic gain for female reproduction efficiency, it is important to know if traits such as PREG1 and REB can be easily volunteered by producers and could add value when integrated to more established platforms such as BREEDPLAN. As an alternative, genomic predictions for easy to measure traits might offer a new avenue for selecting cattle that might appeal to producers who did not access breeding programs before.

2 Objectives

The aims of the project were met through the following four key objectives:

1. Create of a Female Reproduction PhenoBank database
2. Identify functional mutations for female reproduction traits
3. Validate identified genomic markers in different herds to demonstrate feasibility of genetic improvement
4. Design an appropriate database protocol

2.1 Objective 1: Female Reproduction PhenoBank Database

Creation of a Female Reproduction PhenoBank database required the successful integration of multiple datasets in one database, with a standardised vocabulary to define female reproduction traits linked through genotypes. PhenoBank will require and easy-to use database platform that can store phenotypes, genotypes, and annotations, with secure data management, storage and sharing.

2.2 Objective 2: Identify functional mutations for female reproduction traits

Combined phenotypes, genotypes and sequencing resources enabled the application of genome-wide association studies (GWAS) in identification of functional mutations associated with female reproduction traits. GWAS will use high-density genotypes from single nucleotide polymorphism (SNP) chip assays, available after imputation.

2.3 Objective 3: Validate genomic markers in different herds

Validation herds and industry herds were used to apply the identified genomic selection tools for female reproduction traits and demonstrate the feasibility of genetic improvement for reproduction traits. The measurable output of this objective is the estimated accuracy of genomic Estimated Breeding Values (gEBVs) across populations.

2.4 Objective 4: Design an appropriate database protocol

Critical to PhenoBank’s utility is the establishment of a database infrastructure for genomic selection of female reproduction traits. The database required a standardised protocol for importing data into the PhenoBank, including linked meta-data so that phenotypes, genotypes, and data ownership can be easily identified and utilised appropriately.

3 Methodology

The Female Reproduction PhenoBank was proposed as a means to create and demonstrate the use of a large reference population, which is necessary for robust genomic selection. The creation of PhenoBank required the following steps:

1. Develop a working collaboration between the various research partners that were able to contribute female reproduction phenotype data.
2. Develop an appropriate database infrastructure and agreed structure for data storage, including a standardised dictionary of terms.
3. Develop a software tool for data management within the database.
4. Select an appropriate genotyping platform and establish boundaries for data quality.
5. Select herds for inclusion in PhenoBank.
6. Receive DNA samples and facilitate their genotyping.
7. Curate and clean the data prior to importing into PhenoBank.

Once established, the utility of PhenoBank was validated through genomic analyses. These analyses demonstrate the use of PhenoBank in different populations for identifying biologically significant genetic markers and estimating the accuracy of genomic predictions. PhenoBank is a multibreed project focused on tropically adapted breeds with varying degrees of *Bos indicus* content.

3.1 Develop a working collaboration

Two initial workshops were conducted with collaborating research partners: The University of Queensland (UQ, including QAAFI), The University of New England (UNE/AGBU), CSIRO Agriculture and Food, and The Northern Territory Department of Industry Tourism and Trade (NTDITT), formerly known as Department of Primary Industries and Resources (NTDPIR). Through these workshops the project activities and principals of the collaboration were agreed on. Terms of reference (TOR) were created in consultation with all participants and documented to underpin PhenoBank activities. Ongoing workshops were conducted throughout the project to ensure there was consensus among participants about the project's direction.

3.1.1 Background Intellectual Property (IP) and permissions to access data

MLA and the listed research organizations co-own the background IP on the Beef CRC and other previous MLA projects. Therefore, together, MLA, CSIRO, UQ, UNE and NTDITT have the freedom to operate and use this background IP for research purposes, including the PhenoBank activities.

NTDITT own the IP related to their phenotypes and DNA samples, which they have agreed to make available to the project.

Industry herds recruited to the project will sign a material transfer agreement created by UQ legal to address data transfer for research use. Industry participants will have access to their data in PhenoBank, including data analyses.

MLA is custodian of the project IP, including each dataset. During the project, all PhenoBank data is permitted to be shared among project participants for research purposes as detailed in the research contract signed by MLA and UQ. Upon project completion, MLA will control access to the PhenoBank database and own the PhenoBank project IP.

A head agreement contract was signed between UQ and MLA for the governance of PhenoBank. Subsequently, UQ signed sub-contracts with CSIRO, ABRI, and NTDITT. Dr Gehan Jayawardhana was sub-contracted as a consultant to facilitate engagement with NT producers. UNE/AGBU researchers were invited to participate of PhenoBank from the start. Their participation was part of the initial proposal put to MLA and they did contribute to many meetings and workshops during the initial years of the project. Unfortunately, a sub-contract between UQ and UNE/AGBU was not signed in a timely manner and the project had to proceed without their further contributions.

3.2 Scoping of database solutions for the PhenoBank

The pros and cons and relative costs of three possible options for the PhenoBank database were explored, with the understanding that the database platform would need to align with the requirements of a national research tool, as envisaged by MLA.

1. Build an interface that facilitates the link between existing databases and can retrieve data from various sources.
2. Build a new database dedicated to PhenoBank and then populate it with existing data and new data.
3. Enhance an existing database so that it can house existing and new data in a format that facilitates PhenoBank analyses.

Currently, multiple projects store their data as flat files in various computers under the care of the researchers who conducted the original projects, while some data is housed in more than one platform. The database solution was scoped through a series of workshops and put to tender with appropriate suppliers. Noteworthy, the ability to store and manage genotypes was considered an essential tool and a key enhancement compared to previous databases such as the Beef CRC legacy database (i.e., storage of phenotypes only).

Initial consultation and the requesting of quotes raised three potential providers of IT solutions for the PhenoBank project. They were: QFAB Bioinformatics (https://www.qfab.org/), Rezare Systems (https://www.rezare.co.nz/), and Agricultural Business Research Institute (ABRI). Quotes and solutions from these providers were compared in a workshop with all PhenoBank researchers. The proposal presented by ABRI was considered the most cost-effective and appropriate for PhenoBank needs. Hence, ABRI became a sub-contractor of the PhenoBank project.

3.3 Genotypes

The PhenoBank team decided to use NEOGEN's GGP TropBeef SNP chip as the preferred genotyping tool for the project. NEOGEN operates locally, which obviates the need for expensive transport, and the chip was designed for use with tropical breeds. The main competitor, Weatherbys in Ireland, uses a chip designed for taurine breeds and would incur transport risks and additional costs. Quotes from both providers were obtained at the beginning of the project and informed the genotyping decision.

Data reports from SNP genotyping service providers were read and combined using the SVS Golden Helix software. Quality control analyses were performed using standard thresholds for SNP quality: SNPs with a call rate lower than 0.90 and a minor allele frequency lower than 0.05 were discarded from future analyses. A sample call rate of 0.90 was used as the acceptable threshold for sample quality.

Where necessary, such as in situations where a new breed with a diverse genetic background is included in PhenoBank, the Illumina BovineHD chip (770K) should be used across a small number of representative sires to facilitate accurate imputation to higher density data. Similarly, sequence level data would be generated for a small sub-set of animals when necessary.

3.4 Imputing genotype data

All native platform genotypes (20K, 35K, 50K or 80K) were imputed to the Bovine high-density chip (BovineHD; 770K SNP) using the Beef CRC data as the reference population. The HD genotypes were then imputed up to whole genome sequence level using bi-allelic DNA marker data from the 1000 Bull Genome Project, run8 (http://www.1000bullgenomes.com/).

Genotype imputation is a multi-step process. The step-by-step process used was as follows:

1) The genotypes of samples with higher density (actual HD genotypes) had their genotypes phased using the software Eagle (Loh et al., 2016) to be used as "reference".
2) The genotypes of samples with lower density (Bovine SNP50 v1 or v2 or NEOGEN Tropical Chip v1 and v2) were also phased using Eagle, but this time imputation of missing genotypes was not performed (to be used as “target”).
3) Genotype imputation of lower density ("target") up to high-density ("reference") was performed using the software Minimac 3 (Das et al., 2016) for all autosomes and Minimac 4 for the X chromosome.
4) SNP genotypes for 668 animals were extracted from the 1000 Bull Genome project. Selected animals were representative of the relevant breeds for this project (composite breeds, tropically adapted cattle, and cattle with varying degrees of *Bos indicus* content). The data availability for relevant breeds is described in **Table 1**. The raw data was filtered, and only bi-allelic DNA markers were kept, which had at least four copies of the minor allele in this population. These whole-genome sequence samples were phased as per Step 1 and used as a reference for the last step.
5) Samples recently imputed up to HD were then phased using the whole-genome data.
6) Using the same procedure as Step 3, the samples were imputed from HD up to whole-genome sequence, and
7) SNP with imputation quality score (rsq) >0.8 were kept for future analyses.

Table 1: Number of animals per breed with actual HD genotypes or sequence data. These animals were used as the reference panel for genotype imputations.

Breeds	Reference HD 700K SNP	Reference Sequence data
Afrikander		5
Angus	195	50
Angus Red		30
Beefmaster		16
Belmont Tropical Composite	130	
Bonsmara	32	
Boran	24	21
Brahman	863	200
Brangus		5
Charolais		50
Composite	12	

Droughtmaster	345	37
Gir		7
Hereford		50
Limousin		50
Murray Grey		2
Nelore		12
Santa Gertrudis	467	28
Senepol		12
Shaiwal		2
Shorthorn		33
Tropical Composite	351	56
Tuli	33	2
Total	**2,452**	**668**

3.5 Choosing animals for inclusion in PhenoBank

Animals were chosen for inclusion in PhenoBank if they had useful phenotype data and a retrievable DNA sample for genotyping with the selected SNP platform.

PhenoBank included animals from previous research projects, the Northern Territory DITT herds and industry herds. Research data, and data collected by NTDITT staff are deemed more precise and enabled the development of more phenotypes. For example, the Beef CRC Legacy data has day of birth for all cows and regular ovarian scanning, which permits the annotation of puberty defined as the age at first *corpus luteum* (AGECL). Day of birth was also available for the NTDITT cattle and so we could calculate age at first calving (AFC), a phenotype that is a proxy for age at puberty (Moore et al., 2021). Not all producers are able to annotate day of birth and so alternative phenotypes were developed to enable the inclusion of additional industry herds into PhenoBank. Investigating the value of including animals with industry phenotypes, which are less precise, is an important research question. Animals with industry phenotypes were included to expand the genetic groups represented in the PhenoBank database. This was also an opportunity to expand our engagement with producers. To investigate the value of these farm-based phenotypes we redefined female reproduction traits using only basic information that was available for both research and industry herds. The basic information deemed necessary was the year/season of birth that is the cow crop and a pregnancy diagnosis after the first mating season. For example, animals born between November 2018 and February 2019 are all classified as the 2019 crop. Crop year and farm name are combined to create the contemporary groups. Pregnancy tests create records of success and failure that can potentially be used for selection (Moore et al., 2021). The PhenoBank objective was to test if these less precise phenotypes could be used in genomic predictions, to justify their inclusion in the PhenoBank and for future use in reference populations.

Producers submitted the results of pregnancy tests (success or failure) after the first and second mating seasons for all cows within a contemporary group (same crop year), along with a sample for DNA genotyping. Some herds could provide more information than others. We liaised with each producer individually to translate their farm records into PhenoBank phenotypes, according to the dictionary we created for the project, see below.

3.6 Redefining existing phenotypes

A standardised nomenclature and phenotype description was developed to encompass all the phenotypes from the various source projects that were combined in the PhenoBank database. New traits were described to facilitate the merge of research datasets with industry data. The new phenotypes were easier to measure so that the research data would be comparable with phenotypes measured on farm by producers. For example, raw data from the Beef CRC was recoded to create PREG1: a binary trait that represents pregnancy success (1) if the cow got pregnant or failure if not (0) in the first breeding season (among other traits). A dictionary of the nomenclature defined and used in PhenoBank is provided in section 4.3 of this report.

Preliminary analyses of PREG1 indicate that the trait is heritable and correlates with rebreeding in the second mating season (see the results section). After preliminary analyses, we deemed PREG1 and other farm-based traits useful and proceeded to engage with more producers to gather additional data.

3.7 Genomic analyses

The property (business name or herd name) and cow crop (year of birth) were used to create contemporary groups that were used as fixed effects for the genomic analyses. These fixed effects were combined to make a single fixed effect called *cohort* in the datasets.

Genomic analyses were carried in rounds that relate to data availability. The first preliminary analyses used only a sub-set pf Brahman cows. Subsequently, as data becomes available, all analyses were updated and multibreed analyses were carried.

3.7.1 Genotypes

The high-density genotypes of cows in the PhenoBank populations were passed through quality control filter (MAF > 0.05, Genotype Call Rate > 0.9) and SNP not fulfilling the quality control criteria were discarded. In all subsequent analyses, the filtered high-density genotypes used.

3.7.2 Principal component analysis and genomic relationship matrices

The genotypes that passed the quality control filter were used in a principal component analysis that was performed including all animals (Reich et al., 2008). The principal component analysis was done using SNP & Variation Suite v8.x (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com). The top two principal components were then used in GWAS and genomic selection analyses to account for population structure, more specifically to account for breed differences. The top principal components derived from genotypes are a useful covariant in genetic studies of *Bos indicus* influenced cattle, as established by previous work (Porto-Neto et al., 2014).

The genotypes that passed the quality control filter were used to construct separate genomic relationship matrices (GRMs) for each trait using GCTA software (Yang et al., 2011). SNPs in the X-chromosome were also used in the construction of GRM. The GRM accounts for the polygenic effect in our GWAS and genomic selection models. Further this use of GRMs, allows to include cattle of all breeds and without pedigree links in one analysis, deriving multi-breed gEBVs for PREG1 and REB.

3.7.3 GWAS (Genome-wide association studies)

A linear mixed model was used for GWAS of PREG1 and REB. Following is the description of the model used.

$$y = X\beta + Z\mu + S\alpha + \varepsilon$$

where **X** is vector of fixed effects, **β** is estimate of fixed effects. **Z** is an incidence matrix of random polygenic effects; **μ** is the estimate of random polygenic effects normally distributed as $\sim N(0, G\sigma_{\mu}^{2})$. **G** is the genomic relationship matrix (GRM) that was constructed using the SNP genotypes. **S** is the incidence matrix assigning the additive genotype score for each SNP; **α** is the SNP effect. Residual effect is represented as **Ɛ** distributed normally $\sim N(0, \sigma_{e}^{2})$. This model tested single-SNP single trait associations.

In addition to single trait GWAS, a meta-analysis was performed. The meta-analyses used estimated SNP effects for PREG1, REB and first conception score (FCS) to identify significant SNP for early-in-life female reproduction traits, as described in our published paper (Tahir et al., 2021).

3.7.4 GBLUP genomic predictions

All GBLUP analyses were performed using GCTA software (Yang et al., 2011). For the purpose of calculating gEBVs with GBLUP, high-density genotypes were deemed sufficient.

A five-fold cross validation approach was used for calculation of accuracy of genomic predictions. The datasets of both PREG1 and REB traits were randomised separately and divided into five equal subsets. The training and validation sets were produced in a way that four subsets were merged to make a training set and the fifth subset was declared as Validation set. Following this common cross validation strategy, five training and validation sets were produced for each trait.

The training sets for both traits were subjected to GBLUP analysis for the estimation of genomic breeding values (gEBVs), which were used to estimate GBLUP-SNP effects. The following linear mixed model was used to perform GBLUP:

$$y = X\beta + Z\mu + \varepsilon$$

where **X** is vector of fixed effects, **β** is estimate of fixed effects. **Z** is an incidence matrix of random polygenic effects; **μ** is the estimate of random polygenic effects normally distributed as $\sim N(0, G\sigma_{\mu}^{2})$. **G** is the genomic relationship matrix (GRM) that was constructed using the SNP genotypes. Residual is represented as **Ɛ** distributed normally $\sim N(0, \sigma_{e}^{2})$.

The SNP effects calculated from the training sets informed the prediction equations used to calculate the gEBVs of the respective validation sets. These analyses follow the principles first described by Meuwissen et al. (2001).

3.7.5 Bayes R model genomic predictions

The Bayes R model accounts for the biological fact that not all SNP are equally important for each of the target traits. Some SNP will carry more weight than others (van den Berg et al., 2017).

To save computational time for Bayes R analysis, the SNP genotypes were pruned for LD using PLINK software (Purcell et al., 2007) with the window size of 50 SNP, window siding of 5 SNP and variance inflation factor (VIF) threshold of 2. The following model was used for Bayes R analysis as described previously (van den Berg et al., 2017).

$$y = X\beta + Z\mu + Wv + \varepsilon$$

In the Bayes R model, SNP effects (v) were drawn from one of four normal distributions: N(0, 0 x σ_g^2), N(0, 0.0001 x σ_g^2), N(0, 0.001 x σ_g^2) and N(0, 0.01 x σ_g^2). The prior distribution of proportion of SNPs in each of these normal distributions was **P** ~ Dirichlet(α), α = [1, 1, 1, 1]. The number of iterations done in analysis were 50,000 with the first 35,000 iterations being burn-in.

3.7.6 Genomic prediction accuracy

The gEBV and phenotypes adjusted for fixed effects of validation sets were used to calculate the accuracy of genomic prediction for both GBLUP and BayesR approaches. The accuracy of genomic prediction was calculated in two ways as described in literature:

1. Accuracy = r is the correlation between gEBVs and phenotypes adjusted for fixed effects.
2. Accuracy = $r/\sqrt{h^2}$
   Where, the correlation between gEBVs and phenotypes adjusted for fixed effects are divided by the square root of $\boldsymbol{h^2}$ which is heritability of the trait.

4 Results

4.1 Research partner agreements

The collaborating research institutions agreed that the paradigm for collaboration was mutual, ongoing respect for the principal investigators, the background IPs, and the experimental design that underpinned the existing datasets. Therefore, all this information is considered important meta-data and is captured in the PhenoBank database.

The data inventory to accompany all datasets included the following:

- meta-data, with the principal investigator, the Intellectual Property agreements (IP) and the publications related to each dataset (already published or drafted/planned).
- clear definitions for each derived phenotype, to standardise the language used for female reproduction traits.
- information about fixed effects and contemporary groups that are relevant to each phenotype, within each dataset.

4.2 Data storage solutions

It was agreed that the PhenoBank data should be stored independently of the original datasets in a custom-built IT platform, providing controlled access to approved users. In this way, PhenoBank will enhance access to and use of existing datasets as well as offer a backup to protect the investment made in previous projects.

Data storage in PhenoBank was organised in three levels:

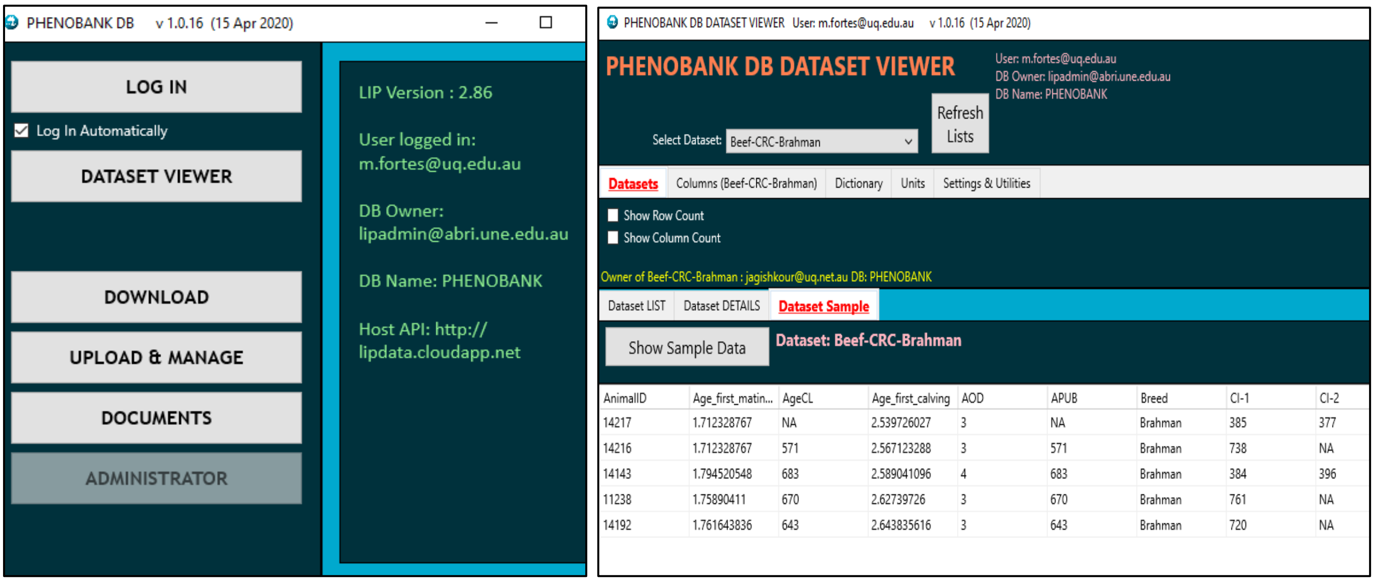
1. the source project meta-data, including
   - principal investigator.
   - the Intellectual Property agreements (IP).
   - publications related to each dataset (already published or drafted/planned).
   - descriptive level data that articulates research design.

2. the phenotype description within each source project, annotated with all fixed effects that will inform genetic models
3. animal level data – individual records for phenotypes and genotypes. Each animal entered into PhenoBank was referenced by an individual alpha numeric animal ID.

The Livestock Information Platform (LIP), co-created by ABRI and CSIRO, was chosen as a foundation for the PhenoBank database as it meets PhenoBank's requirements for secure access, data storage and retrieval, and served as a starting point to develop project specific functionality. The PhenoBank database tool allows users to upload data, merge data, allow and restrict access to data and download data for analysis. It is available for both Windows and web environments. A manual for the PhenoBank platform is provided in Appendix 1.

The PhenoBank platform was extensively beta tested with draft datasets before distribution among the PhenoBank collaborators. The current version of PhenoBank is V 1.0.28 and this has been made available to all users during the project.

Figure 1 demonstrates the PhenoBank database App log in page and the dataset viewer, a very useful feature of the platform. In the dataset viewer, anyone can see a sample of 10 records in a dataset that has been uploaded by a project participant. Allowing researchers to view a sample of each dataset will encourage collaboration, without disclosing the entire data. This is both time effective and it protects the data in the PhenoBank. The dataset viewer creates a vitrine for genomic selection in Australia and may be used to workshop the App and attract more industry partners.



☑ Log In Automatically

LIP Version : 2.86

User logged in: m.fortes@uq.edu.au

DB Owner: lipadmin@abri.une.edu.au

DB Name: PHENOBANK

Host API: http://lipdata.cloudapp.net

PHENOBANK DB DATASET VIEWER User: m.fortes@uq.edu.au v 1.0.16 (15 Apr 2020)

User: m.fortes@uq.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: Beef-CRC-Brahman

□ Show Row Count
□ Show Column Count

Owner of Beef-CRC-Brahman : jagishkour@uq.net.au DB: PHENOBANK

Dataset: Beef-CRC-Brahman

AnimalID	Age_first_matin...	AgeCL	Age_first_calving	AOD	APUB	Breed	CI-1	CI-2
14217	1.712328767	NA	2.539726027	3	NA	Brahman	385	377
14216	1.712328767	571	2.567123288	3	571	Brahman	738	NA
14143	1.794520548	683	2.589041096	4	683	Brahman	384	396
11238	1.75890411	670	2.62739726	3	670	Brahman	761	NA
14192	1.761643836	643	2.643835616	3	643	Brahman	720	NA

Figure 1: Login screen for the PhenoBank database and dataset viewer for the Windows version. The dataset viewer function provides the first ten individual cow records in each dataset, to showcase the data being collected in the PhenoBank.

4.3 Dictionary of traits

The language used for female reproduction traits was standardised, as detailed in **Table 2**. This dictionary was used to extract easy to measure traits from raw data, focussing on early in life traits measured over the first three breeding seasons to enable faster genetic gains for female reproduction. The genetic correlations between these newly defined traits and other important traits defined from previous research were verified in preliminary analyses and merit further investigations. The dictionary was constructed with input from all participating researchers and was shared with producers to facilitate conversations and help with collecting new phenotypes for the project. Having a dictionary brings clarity as to the meaning of each phenotype.

Descriptive statistics for all recorded traits in PhenoBank are detailed in

Table 3.

Table 2: Basic dictionary of female reproduction traits created for the PhenoBank project.

PhenoBank Trait	Long description
PREGY	Pregnant (= 1) or not (= 0) after the first mating opportunity, for heifers exposed to mating as yearling (between 10 and 18 months of age).
PREG1	Pregnant (= 1) or not (= 0) after the first mating opportunity, regardless of age. PREG1 and PREGY are the same for producers who practice yearling mating.
PREG2	Pregnant (= 1) or not (= 0) after the second mating opportunity.
PREG3	Pregnant (= 1) or not (= 0) after the third mating opportunity.
AGECL	Age at first *corpus luteum* detected with regular ultrasound scans after weaning. This trait is available in the Beef CRC Legacy data.
APUB	Age at puberty from the 600 days scan (Repronomics definition).
AFC	Age at first calving, age in days when the cow calved for the first time.
DTC	Days to calving is the time between first mating and calf being born (similar to BREEDPLAN definition).
ANLY	Anoestrus length after yearling mate is the time in days between calving and the first cycle post-partum (observation of heat or CL or calculated from foetal ageing records of the next pregnancy) after yearling mating (heifers exposed to mating between 10 and 18 months of age).
ANL1	Anoestrus length after first mate is the time in days between calving and the first cycle post-partum (observation of heat or CL or calculated from foetal ageing records of the next pregnancy) after first mating, regardless of age at first mating.
ANL2	Anoestrus length after second mate is the time in days between calving and the first cycle post-partum (observation of heat or CL or calculated from foetal ageing records of the next pregnancy) after second mating.
ANL3	Anoestrus length after third mate is the time in days between calving and the first cycle post-partum (observation of heat or CL or calculated from foetal ageing records of the next pregnancy) after third mating
REB	Rebreeding score (from 1 – 4) is scored as 1 when PREG1 and PREG2 are both 0, scored 2 when PREG1 is 0 and PREG2 is 1, scored 3 when PREG1 is 1 and PREG2 is 0, and scored 4 when PREG1 and PREG2 are both 1.
REPROSCORE	Reproductive score from 1 (not pregnant and dry) to 12 (pregnant and wet) as per previous publication (Reverter et al., 2016).
FCS	First Conception Score describes the age of first conception using approximate age data. Score 3 = cows that first conceived before 29 months of age; score 2 = cows that first conceived between 29 and 36 months of age; Score 1 = cows that conceived for the first time after 36 months of age.
RT_SCORE	Scan of the reproductive tract, score from 0 – 5 (infantile to mature).

Note: a more comprehensive dictionary is also available upon request, with definitions for all traits captured in the PhenoBank database, including traits defined by previous projects and definitions for fixed effects.

Table 3: Descriptive statistics for all recorded traits in PhenoBank (as of August 2021).

	N	MEAN	SD	MEDIAN	MIN	MAX	RANGE	SE
CROP	9193	2009	7.24	2009	1999	2019	NA	NA
AFC	4121	1106.92	202.03	1089.00	629.00	1890.00	1261.00	3.15
PREGY	3035	0.41	0.49	0.00	0.00	1.00	1.00	0.01
PREG1	6983	0.62	0.48	1.00	0.00	1.00	1.00	0.01
PREG2	5008	0.67	0.47	1.00	0.00	1.00	1.00	0.01
PREG3	2308	0.77	0.42	1.00	0.00	1.00	1.00	0.01
REB	3822	2.96	0.95	3.00	1.00	4.00	3.00	0.02
AGECL	1872	701.17	138.06	720.00	344.00	1211.00	867.00	3.19
DTC	2007	316.05	16.79	314.00	273.00	384.00	111.00	0.37
DTC1	1590	311.10	20.03	306.00	259.00	382.00	123.00	0.50
DTC2	1300	323.44	27.68	317.00	260.00	445.00	185.00	0.77
DTC3	1562	315.96	21.81	310.00	275.00	385.00	110.00	0.55
PPAI	1414	158.88	110.49	107.00	17.00	484.00	467.00	2.94
PW	3773	0.35	0.48	0.00	0.00	1.00	1.00	0.01
REPROSCORE	1410	7.20	3.13	9.00	1.00	11.50	10.50	0.08
PREG600D.WEEKS.	400	2.00	5.16	0.00	0.00	24.00	24.00	0.26
RT_SCORE	1026	2.85	1.53	3.00	0.00	5.00	5.00	0.05

Note: new data is still being curated and integrated into the PhenoBank database and so these numbers are not final.

4.4 Data entry into PhenoBank

Phenotypic data available from existing research datasets was manually re-coded and curated for the PhenoBank project using the dictionary of traits described above.

In addition, a number of industry herds have enrolled and supplied new phenotypes for PhenoBank, with the aim of expanding the reference population so it includes other genetic compositions. Newly supplied phenotypes are first evaluated against PhenoBank's criteria for goodness of fit before genotyping is performed.

At the completion of the project, PhenoBank contained sequence level data on 10,507 cows, all with early-in-life fertility phenotypes and genotypes available for use. A summary of the sources of data currently included in PhenoBank is provided in **Table 4**.

Table 4: Sources of data currently included in PhenoBank.

Datasets submitted to PhenoBank	Number of cows with phenotypes
Alpha Brangus (producer data)	458
Beef CRC Legacy	2,010
Beef CRC Rachel's Cows	2,339
Bimbadeen (producer data)	110
Coodardie (producer data)	147
Gyranda (producer data)	1,026
Iffley (CSIRO contribution)	550
Kamilaroi (CSIRO contribution)	1,410

Malilangwe (producer data)	71
Mt Eugene (producer data)	505
NT DITT commercial herd	515
NT DITT stud herd	718
REB Grazing (producer data)	279
Rocky Springs (producer data)	187
Roxborough (producer data)	182
Total	**10,507**

4.5 Heritability of farm-based phenotypes

Preliminary analyses were performed to explore the potential for using newly defined traits: PREG1, FCS and REB. The trait PREG1 is of particular interest to the project as it is typical of what can be measured on extensive properties, requiring only a pregnancy diagnosis along with knowledge of the season of birth (cow crop year) and the season of mating. Many properties do not know the exact day of birth for their heifers, but they do know the season (or crop) they belong to, thus forming a contemporary group for data analyses.

Heritabilities were estimated for PREG1, FCS, and REB in Brahman animals sourced from Beef CRC Brahman (n=962), NTDITT Brahman (n= 179 Braham within 965 cattle) and Kamilaroi (n=1410). heritabilities and the correlations between them are shown in **Table 5**. The accuracy of genotype imputation in these data was 0.95 for Beef CRC animals, 0.92 for Kamilaroi cattle and 0.93 for NTDITT herd. The analyses that used only the Brahman sub-set of PhenoBank data were published earlier this year (Tahir et al., 2021)

Table 5: Genetic and phenotypic parameters of three reproductive traits in Brahman. Estimated heritabilities are shown on the diagonal; genetic correlations are shown above the diagonal and phenotypic correlations are shown below the diagonal.

Traits	PREG1	FCS	REB
PREG1	**0.17 (0.03)**	0.839 (0.06)	0.799 (0.07)
FCS	0.86 (0.01)	**0.11 (0.03)**	0.756 (0.1)
REB	0.73 (0.02)	0.65 (0.02)	**0.28 (0.05)**

Correlations between these traits are high, as expected, because they are generated from the same field observations. Heritability estimates between 0.11 and 0.28 are encouraging for reproduction traits, and especially in the context of less precise farm-based phenotypes.

More recent analyses have used multibreed data, available in PhenoBank, and focused on the two traits that were promising from the above Brahman analyses, PREG1 and REB. The heritabilities estimated in the multibreed analyses are 0.17 for PREG1 and 0.21 for REB, when single step GBLUP analyses were used. More details about the updated analyses are presented in section 4.6.3 when we discuss the genomic selection results, see Table 9. In short, heritability estimates suggest that is possible to use pregnancy success and failure to select for female reproduction traits. Recording the full cohort of heifers and young cows, after the first and second mating seasons, is key for these traits.

4.6 Genetic analyses

Records for a total of 8,461 cows—phenotypes and genotypes imputed to HD level data were available in PhenoBank at the time of reporting. This sub-set of the total data was used for genomic analyses of reproductive traits measured early in life. After curating and cleaning, the merged dataset contained 8,371 cows. Contemporary groups with less than 3 females were removed from analyses.

4.6.1 Genotype imputation

In the HD version, all cows have over 720,000 SNP genotypes available after quality control. At sequence level, the number of SNPs per cow is approximately 15 million after quality control and removal of SNPs with low imputation accuracies. This level of genome-wide data for tropical breeds is a world-class resource that will help to develop and deliver genomic predictions for Northern Australia.

4.6.2 Identification of SNP associated with female reproduction traits

Identification of SNP associated with female reproduction traits is an ongoing exercise within the PhenoBank project. Initially, Braham data was used in a preliminary GWAS (Tahir et al., 2021). More recently we performed a second GWAS in a multibreed analyses. Below we report first the GWAS of the Brahman sub-set and after the multibreed analyses.

The initial study investigated the utility of three of the newly defined traits, PREG1 (pregnancy at first mating opportunity), FCS (first conception score), and REB (Rebreeding score) to estimate genomic merit. Approximately 2,400 Brahman cattle (from the Beef CRC, Kamilario and NT DITT herds) were used to perform GWAS and multi-trait meta-analysis to determine genomic regions associated with heifer fertility. Heritability estimates for the traits examined in this population were 0.17 (0.03) for PREG1, 0.11 (0.03) for FCS and 0.28 (0.05) for REB. The three traits were highly genetically correlated (0.75-0.83), for details consult the published paper (Tahir et al., 2021).

Meta-analysis using SNP effects estimated for each of the three traits and adjusted for standard identified 1359 significant SNPs (*P*-value < $9.9\text{x}10^{-6}$ at FDR <0.0001). Genomic regions of 0.5 Mb around each significant SNP from the meta-analysis were annotated to create a list of 2560 positional candidate genes. The most significant SNP was in vicinity of a genomic region on chromosome 8, encompassing the genes *SLC44A1*, *FSD1L*, *FKTN*, *TAL2* and *TMEM38B*. The genomic region in humans that contains homologs of these genes is associated with age at puberty in girls.

Significant SNPs pointed to additional fertility related genes, again within a 0.5 Mb region, including *ESR2*, *ITPR1*, *GNG2, RGS9BP, ANKRD27*, *TDRD12*, *GRM1, MTHFD1, PTGDR* and *NTNG1.* Functional pathway enrichment analysis resulted in many positional candidate genes relating to known fertility pathways, including GnRH signaling, estrogen signalling, progesterone mediated oocyte maturation, cAMP signalling, calcium signalling, glutamatergic signalling, focal adhesion, PI3K-AKT signalling and ovarian steroidogenesis pathway.

The comparison of results from this study with previous transcriptomics and proteomics studies on puberty in different populations of Brahman identified 392 genes in common, with the following genes also forming part of the above-mentioned pathways: *BRAF, GABRA2, GABR1B, GAD1, FSHR, CNGA3, PDE10A, SNAP25, ESR2, GRIA2, ORAI1, EGFR, CHRNA5, VDAC2, ACVR2B, ORAI3, CYP11A1, GRIN2A, ATP2B3, CAMK2A, PLA2G, CAMK2D and MAPK3.* The biological functions of the positional candidate genes and their annotation to known pathways allowed integration of these results into an overview of molecular mechanisms related to puberty in the hypothalamus-pituitary-ovarian axis. A reasonable number of genes, common between previous puberty studies and the present study, corroborate the proposed molecular mechanisms.

Multibreed GWAS of PREG1 and REB confirmed the polygenic nature of these traits. Considering multiple tests, no SNP was significant in this analysis (Figure 2). Further work is necessary to explore each breed as an individual dataset. Breed specific analyses and meta-analyses, like suggested above, may lead to the identification of additional SNP for female reproduction traits.

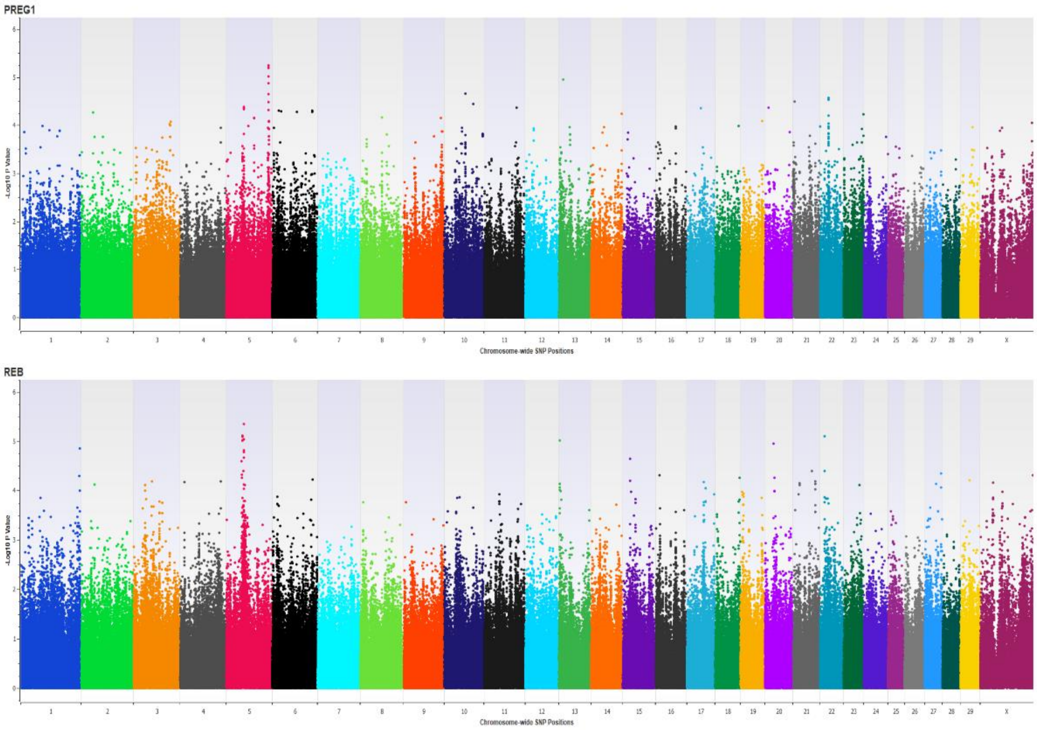


Figure 2: Manhattan plot from genome-wide association studies for reproduction traits measured early in life. Top: PREG1 (first pregnancy); Bottom: REB (rebreeding).

4.6.3 Genomic selection predictions and accuracies

We used the discoveries of our initial Brahman GWAS to select SNP for genomic selection analyses. SNP present in the vicinity of the puberty-related candidate genes were selected to create a set of biologically informed SNP, while the remainder of SNPs were considered biologically uninformed. Four SNP sets were used to estimate genomic breeding values (gEBVs) for the experimental population of Brahman cattle (described above) and a validation population of 1350 Tropical Composite cows from NT DITT and Beef CRC herds. The four SNP sets were: 1) meta-analyses SNP, multi-omics SNP, complete SNP and random SNP. Meta-analyses and multi-omics SNP sets are biologically informed, while complete and random were uninformed SNP sets, as detailed in our published work (Tahir et al., 2021). Records for the reproductive traits PREG1 and REB were available for both herds, but the FCS trait was only available for the Beef CRC herd.

Heritabilities and within-breed accuracies of genomic predictions for PREG1, FCS and REB are shown for the experimental Brahman population in **Error! Reference source not found.** and for the tropical composite validation population in **Error! Reference source not found.**. Although the accuracy of prediction was lower in the validation population, the accuracies of predictions improved in both populations when biologically informed SNP sets were used compared to uninformed sets, indicating that use of selected SNP can be beneficial for genomic predictions, even across breeds. The idea of biologically informed SNP sets is gaining momentum and recent literature is in accord with the results presented herein (Xiang et al., 2021). Future research should expand on these concepts of multi-omics data that can benefit genomic selection accuracies.

Table 6: Heritabilities and within breed accuracies of genomic prediction for the early reproductive traits, PREG1, FCS and REB in the experimental Brahman population (n=2,400).

SNP Sets for Genomic Prediction	h^2	SE	Correlation	Accuracy
PREG1				
Meta-analysis SNP Set	0.38	0.03	0.35	0.57
Multi-omics SNP Set	0.24	0.04	0.21	0.42
Complete SNP set	0.17	0.03	0.15	0.37
Random SNP Set	0.08	0.03	0.09	0.31
FCS				
Meta-analysis SNP Set	0.31	0.03	0.28	0.51
Multi-omics SNP Set	0.16	0.04	0.15	0.37
Complete SNP set	0.11	0.03	0.10	0.31
Random SNP Set	0.04	0.02	0.06	0.32
REB				
Meta-analysis SNP Set	0.47	0.04	0.38	0.55
Multi-omics SNP Set	0.35	0.05	0.27	0.50
Complete SNP set	0.28	0.05	0.22	0.42
Random SNP Set	0.18	0.05	0.17	0.42

Of note, the accuracies reported that used the meta-analyses SNP for the experimental Brahman population are artificially high because this is the same population used for the GWAS work. Therefore, we expect that the associated SNP will have a high predictive power in this population. To provide an independent validation, and a more realistic scenario, we worked with the Tropical Composite cattle. Future work using the full data now available in the PhenoBank database will further test the validity of these selected SNP.

Table 7: Heritabilities and within breed accuracies of genomic prediction for the early reproductive traits, PREG1, FCS and REB in the validation tropical composite population (n=1,350)

SNP Sets for Genomic Prediction	h^2	SE	Correlation	Accuracy
PREG1				
Meta-analysis SNP Set	0.14	0.04	0.18	0.49
Multi-omics SNP Set	0.18	0.05	0.20	0.46
Complete SNP set	0.16	0.05	0.17	0.44
Random SNP Set	0.13	0.05	0.16	0.43
FCS				
Meta-analysis SNP Set	0.09	0.04	0.12	0.42
Multi-omics SNP Set	0.12	0.05	0.13	0.36
Complete SNP set	0.08	0.05	0.10	0.37
Random SNP Set	0.08	0.05	0.10	0.34
REB				
Meta-analysis SNP Set	0.07	0.04	0.12	0.45
Multi-omics SNP Set	0.09	0.04	0.12	0.38
Complete SNP set	0.11	0.05	0.11	0.34
Random SNP Set	0.10	0.04	0.10	0.33

A larger subset of the PhenoBank curated data (8,371 cows) with genotypes imputed to HD and phenotypes for PREG1 and REB were used for this analysis: 5473 cows had useful phenotype and fixed effects records for the trait PREG1, and 3119 animals had useful phenotype and fixed effects records for the trait rebreed score (REB). These records belonged to different properties, which were Gyranda, Beef CRC, Kamilaroi, Mt Eugene, Rocky Springs and NTDPI studs. The dataset populations for both traits included animals from multiple breeds: Brahman, Tropical Composite, Santa Gertrudis, Angus black and Angus crossbred.

A combination of farm and year of birth was used as the cohort to account for contemporary group effect in the models. Any cohort that had three or less animals were removed from analyses. **Table 8** describes the distribution of animals in different cohorts for each trait.

Table 8: Distribution of animals for cohorts and phenotype score for the traits PREG1 and REB.

Cohort	PREG1 Score 0 (n)	PREG1 Score 1 (n)	PREG1 Total (n)	REB Total (n)	REB Score Average
2001DD				9	3.78
2002DD				10	3.40
2003DD	27	11	38	38	2.42
2004DD				9	3.44
2007DD	14	6	20	20	2.35
2008DD	34	22	56	56	2.57
2009DD	71	22	93	93	2.34
2010DD	39	43	82	81	2.84
2011DD	39	63	102	102	3.01
2012DD	16	54	70	70	3.29
2012KA	68	383	451		
2013DD	25	40	65	65	2.97
2013GY	286	14	300		
2013KA	27	437	464		
2014DD				59	3.24
2014GY	243	114	357		
2014KA	314	177	491	424	1.99
2015DD	30	32	62	62	2.79
2015GY	264	101	365		
2017ME	12	93	105	45	3.51
2018ME	110	7	117		
2019ME	148	34	182		
2019RS	17	30	47	29	3.76
PBE00	12	47	59	59	2.88
PBE01	41	165	206	206	3.30
PBE02	25	195	220	220	3.44
PBE03	28	130	158	158	3.28
PBP01	10	126	136	136	3.65
PBP02	27	216	243	243	3.29
PBP03	7	65	72	72	3.49
PSL01	77	86	163	163	2.66
PSL02	63	141	204	204	2.83
PSL03	5	31	36	36	3.28
PTR01	37	156	193	193	3.36
PTR02	67	190	257	257	3.31
Grand Total	**2183**	**3221**	**5414**	**3118**	

Principal component analyses (PCA) performed on this dataset are shown in **Figure 3**. The top two PCs explained 17% of the total variance for the population used in PREG1. Variance explained by top two PCs for the REB population was 14%. Therefore, fitting these components is useful for multibreed reference populations. This is the same population used for the multibreed GWAS for which the Manhattan plots are shown in **Figure 2**.

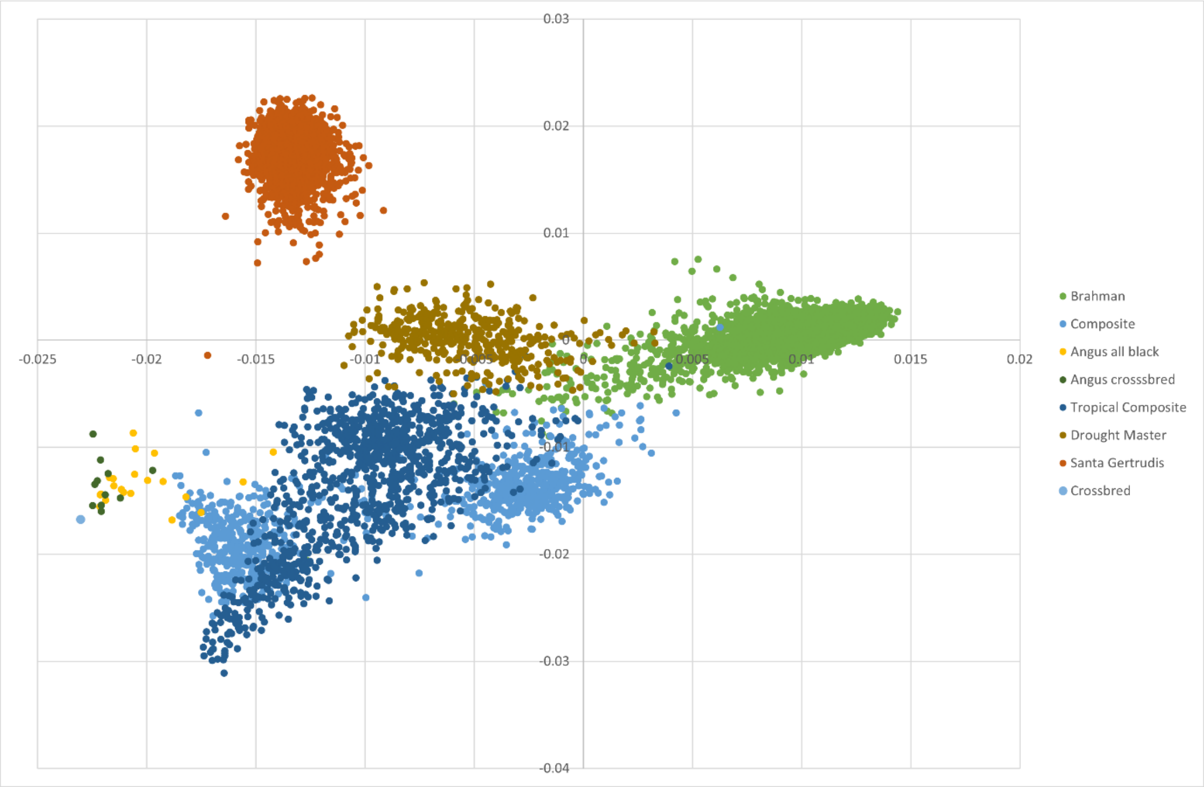


Figure 3: **Principal Component Analyses (PCA) of the merged dataset. High-density genotypes were used to visualise the genetic similarities and differences between animals in the dataset. The clustering of animals conforms to expectations considering the known breed of each animal. This illustrates the multi-breed population structure of the PhenoBank data.**

The accuracy of genomic predictions, calculated using gEBVs of validation datasets and phenotypes adjusted for fixed effects, are shown in **Table 9**.

Table 9: Genomic predictions for PREG1 and REB (Rebreed Score) using PhenoBank data. Cohort (farm and year of birth) was the fixed effect used to adjust the phenotype, and the first two principal components derived from all the genotypes were used as covariates in the model.

Trait	Method	h^2	S.E.	Correlation gEBVs x adj phenotypes	Accuracy (Correlation /$\sqrt{h^2}$)	SNPs	Total Population (n)	Training Population (n)	Validation Population (n)
PREG1	GBLUP	0.17	0.02	0.19	0.47	669,396	5414	4331	1083
REB	GBLUP	0.21	0.03	0.19	0.42	662,896	3118	2495	623
PREG1	Bayes R	0.18	0.02	0.18	0.43	669,396 96,992*	5414	4331	1083
REB	Bayes R	0.19		0.19	0.43	662,896 96,033*	3118	2495	623

* Number of SNP remaining after LD pruning.

5 Conclusion

Now that a reference population has been established with defined fertility traits, and a dedicated software tool has been created to access the data, there is an opportunity to further use the resources created in the PhenoBank project to drive on-farm practice change and genetic improvement in female reproduction. Reproductive performance of northern herds has traditionally been limited by difficulties in recording performance metrics and slow responses to selection related to low heritability of key production traits (Johnston 2013). The fertility traits proposed in PhenoBank have similarly low heritability, but they are easy to measure, and they align with the traditional recommendations about culling empty breeders. Producers who engage with PhenoBank will receive timely reports on DNA diagnostics as part of the PhenoBank Bundle, created with NEOGEN. The access to parentage verifications and DNA diagnostics incentivise producers to volunteer their farm records and submit samples for genotyping.

Therefore, as PhenoBank can be a platform to demonstrate genomic technologies for producers across northern Australia, there is a clear link between this project and the MLA-initiated northern breeding business (NB2) program of work. This demonstration could target those breeding herds captured in the extension activities of NB2, which could also benefit from DNA diagnostics. The extension of PhenoBank activities that potentially adds value to NB2 comprise customised workshops organized by PhenoBank researchers. The NB2 structure provides a system to engage with producers and can help to standardise phenotypic and genotypic data collection. As information is feedback to producers it provides practical decision tools, and opportunities to facilitate the adoption of genomic technologies.

The incorporation of superior genetics to deliver reproductive gains for northern herds is often disconnected from other core management options. These workshops could incorporate explanations about how to collect samples, what DNA diagnostics are available and how to use them. PhenoBank researchers could readily demonstrate how genomic selection can benefit northern breeding herds and support management decisions, creating an awareness to drive adoption and long-term practice change. The message to be reinforced is that the first two breeding cycles are key to female reproductive performance, plus the fact that genotyping is useful immediately, for diagnostics, and later on in the form of accurate gEBVs.

5.1 Key findings

- This project established a database of reliable measures of early-in-life female reproduction phenotypes. This database is a critical requirement for the discovery of DNA markers and the implementation of genomic technologies to drive genetic improvement for reproduction traits. Dataset creation was fast-tracked by mining existing research datasets to define new reproductive measures that are simple to measure on-farm; these were then merged with producer collected data to facilitate the rapid development of a large reference population for discovery and evaluation of gEBVs.
- The project also developed a software platform capable of merging data from different sources, storing and retrieving each animal's phenotype and SNP data. Data access is controlled through a simple user interface.
- Phenobank now offers high density genotype data on 10,507 cows with phenotype data for early-in-life female reproduction traits. This will be a valuable data storage hub for both

researchers and industry stakeholders working on future discovery projects and answering other genomic queries.

- The reference population available through PhenoBank is useful for the development of gEBVs as demonstrated in our analyses.
    - Two genome-wide association studies were performed to identify SNP markers associated with female reproduction traits in single breed and multibreed analyses.
    - SNP associated with the traits PREG1 and REB that were in the vicinity of the puberty-related candidate genes were selected to create a set of biologically informed SNP. Heritabilities and accuracies of genomic predictions were determined for both the experimental Brahman population and a validation population of tropical composite cattle. These analyses demonstrated that the accuracies of prediction improved when biologically informed SNP sets were used.
    - The newly defined traits, PREG1 and REB, are heritable if all females in the cohort are measured (i.e., by pregnancy diagnosis of all females within a crop year that were mated).
    - The use of selected associated SNP can be beneficial for genomic predictions of these traits, even across breeds.
- Discussions and collaborations developed during this project with industry stakeholders—both project participants and field day attendees—generated substantial producer interest in the use of genomic technologies for tropical beef cattle. An easy first step for producer adoption would be to raise awareness about parentage verification and simple diagnostics (such as Pompe's disease test and the Poll/Horn test) that can be bundled with SNP sets designed to assist with genomic selection for improved female reproduction traits.
    - A win-win within PhenoBank was the fact that researchers could use SNP chip data for genomic predictions while producers had rapid access to parentage verification and simple diagnostics that are available as soon as the animal is genotype. The research gEBVs are provided after data analyses.

5.2 Benefits to industry

It is well recognised that rapid genetic gains are possible when genomic tools are used to assist selection strategies. It is also well known that the development of powerful genomic tools to drive genetic improvement is underpinned by the existence of large reference populations, which traditionally are limited by the substantial time and funds required for establishment.

This project has demonstrated the utility of such a reference population for discovering simple and yet heritable female reproduction traits: ones that can be measured inexpensively on farm by producers themselves. The power of those traits to make genetic gains in female reproduction efficiency is tangible given the accuracies of our predictions. As Phenobank expands and researchers fully explore the data available, it is envisaged that this resource will have a significant positive influence on improving female reproductive performance in the northern Australian beef herd.

Importantly, this project has also provided a valuable template for the rapid and relatively inexpensive assembly of future reference populations that could underpin the discovery of additional genomic applications. The combination of data mining from existing research datasets, merging with contemporary data collected from industry herds, and managed through a flexible software platform might be the start of a digital infrastructure that has far-reaching implications for improved production in the red meat industry.

6 Future research and recommendations

The establishment and validation of PhenoBank and the producer interest generated so far attest to the success of this project. The PhenoBank database has the potential to kick-start the adoption of genomic technologies in the north and drive improvement in female reproduction efficiency in beef cattle. As northern Australian cow-calf operations are a significant component of the national red meat industry, improving its efficiency has positive flow-on effects for the entire sector.

Producers who have already collaborated with Phenobank, as well as research collaborators, can act as “adoption champions”, encouraging others to consider adopting this new framework for managing female fertility. Fully engaging producers will necessitate further extension activities, such as workshops, field days and Beef-up forum. These activities should showcase the project, its anticipated benefits, and the advantages gained from collecting simple measures, even if producers are unable to collect calf birth dates or mothering up data.

Phenobank is its infancy and its potential benefit for the northern beef herd is significant. To realise this benefit, the database should continue to expand in size with input from industry herds and be fully interrogated by researchers for SNP associated with female reproduction traits. Software upgrades that allow storage and usage of sequence level data will enhance discovery activities. There is also potential to explore genetic correlations between the traits in Phenobank and those measured in other projects, such as Repronomics and Smart futures. Potentially, these projects could become another source of data for Phenobank to ultimately generate more robust predictions. Likewise, PhenoBank data could be made available to producers, researchers and genetic improvement programs so that data collected with MLA funds benefits everyone in the industry. A few producers involved in this project have already requested that the genotypes on their cattle are to be made available to the breed association so that official gEBVs can be calculated. PhenoBank researchers are supportive of these requests, which are evidence of immediate impact of the project in benefit of producers.

Investment in the above-mentioned future activities should generate positive consequences. As more producers contribute data to Phenobank, and research identifies robust genomic markers, the accuracy of genomic predictions will continue to improve, and producers will experience the benefits in profitability. Projects such as PhenoBank, which engage directly with producers and genotyped industry herds should help embed and standardise the use of genomic technologies for beef cattle improvement.

6 References

Bolormaa, S.;Pryce, J. E.;Kemper, K.;Savin, K.;Hayes, B. J.;Barendse, W.;Zhang, Y.;Reich, C. M.;Mason, B. A.;Bunch, R. J.;Harrison, B. E.;Reverter, A.;Herd, R. M.;Tier, B.;Graser, H. U. and Goddard, M. E. 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in bos taurus, bos indicus, and composite beef cattle. J. Anim. Sci. 91:3088-3104.

Bolormaa, S.;Pryce, J. E.;Reverter, A.;Zhang, Y. D.;Barendse, W.;Kemper, K.;Tier, B.;Savin, K.;Hayes, B. and Goddard, M. E. 2014. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. Plos Genetics 10.

Cammack, K. M.;Thomas, M. G. and Enns, R. M. 2009. Review: Reproductive traits and their heritabilities in beef cattle. Prof. Anim. Sci. 25:517-528.

Engelken, T. J. 2008. Developing replacement beef heifers. Theriogenol. 70:569-572.

Erbe, M.;Hayes, B. J.;Matukumalli, L. K.;Goswami, S.;Bowman, P. J.;Reich, C. M.;Mason, B. A. and Goddard, M. E. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114-4129.

Goddard, M. E. and Hayes, B. J. 2007. Genomic selection. J. Anim. Breed. Genet. 124:323-330.

Goddard, M. E.;Hayes, B. J. and Meuwissen, T. H. E. 2010. Genomic selection in livestock populations. Genetics Research 92:413-421.

Goddard, M. E.;Hayes, B. J. and Meuwissen, T. H. E. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409-421.

Graser, H. U.;Tier, B.;Johnston, D. J. and Barwick, S. A. 2005. Genetic evaluation for the beef industry in australia. Australian Journal of Experimental Agriculture 45:913-921.

Harburg, S.;McLean, I. and Hayes, B. 2020. Economic consequences of selection for cow fertility in northern australian beef herds. Bush Agribusiness.

Johnston, D. J.;Barwick, S. A.;Corbet, N. J.;Fordyce, G.;Holroyd, R. G.;Williams, P. J. and Burrow, H. M. 2009. Genetics of heifer puberty in two tropical beef genotypes in northern australia and associations with heifer- and steer-production traits. Anim. Prod. Sci. 49:399-412.

Johnston, D. J.;Barwick, S. A.;Fordyce, G. and Holroyd, R. G. 2010. Understanding the genetics of lactation anoestrus in brahman beef cattle to enhance genetic evaluation of female reproductive traits 9th WCGALP, Leipzig, Germany.

Johnston, D. J.;Barwick, S. A.;Fordyce, G.;Holroyd, R. G.;Williams, P. J.;Corbet, N. J. and Grant, T. 2013. Genetics of early and lifetime annual reproductive performance in cows of two tropical beef genotypes in northern australia. Anim. Prod. Sci. 54:1-15.

Lyons, R. E.;Nguyen To, L.;Dierens, L.;Fortes, M. R. S.;Kelly, M.;McWilliam, S. S.;Li, Y.;Bunch, R. J.;Harrison, B. E.;Barendse, W.;Lehnert, S. A. and Moore, S. S. 2014. Evidence for positive selection of taurine genes within a qtl region on chromosome x associated with testicular size in australian brahman cattle. BMC Genetics 15:(10 January 2014)-(2010 January 2014).

McGowan, M.;Fordyce, G.;O'Rourke, P.;Barnes, T.;Morton, J.;Menzies, D.;Jephcott, S.;McCosker, K.;Smith, D.;Perkins, N.;Marquart, L.;Newsome, T. and Burns, B. 2014. B.Nbp.0382 northern australia beef fertility project: Cashcow, Meat and Livestock Australia.

Meuwissen, T.;Hayes, B. and Goddard, M. 2013. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci 1:221-237.

Meuwissen, T. H. E.;Hayes, B. J. and Goddard, M. E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Moore, K. L.;Wolcott, M. L. and Johnston, D. J. 2021. Genetic improvement of cow reproduction in northern australia beef cattle breeds. p 151-162.

Porto-Neto, L. R.;Reverter, A.;Prayaga, K. C.;Chan, E. K.;Johnston, D. J.;Hawken, R. J.;Fordyce, G.;Garcia, J. F.;Sonstegard, T. S.;Bolormaa, S.;Goddard, M. E.;Burrow, H. M.;Henshall, J. M.;Lehnert, S. A. and Barendse, W. 2014. The genetic architecture of climatic adaptation of tropical cattle. PLoS One 9:e113284.

Purcell, S.;Neale, B.;Todd-Brown, K.;Thomas, L.;Ferreira, M. A. R.;Bender, D.;Maller, J.;Sklar, P.;de Bakker, P. I. W.;Daly, M. J. and Sham, P. C. 2007. Plink: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81:559-575.

Reich, D.;Price, A. L. and Patterson, N. 2008. Principal component analysis of genetic data. Nat. Genet. 40:491-492.

Reverter, A.;Porto-Neto, L. R.;Fortes, M. R. S.;McCulloch, R.;Lyons, R. E.;Moore, S.;Nicol, D.;Henshall, J. and Lehnert, S. A. 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of brahman cows with unknown pedigree. J. Anim. Sci. 94:4096-4108.

Snelling, W. M.;Cushman, R. A.;Fortes, M. R. S.;Reverter, A.;Bennett, G. L.;Keele, J. W.;Kuehn, L. A.;McDaneld, T. G.;Thallman, R. M. and Thomas, M. G. 2012. Physiology and endocrinology symposium: How single nucleotide polymorphism chips will advance our knowledge of factors controlling puberty and aid in selecting replacement beef females. J. Anim. Sci. 90:1152-U1119.

Snelling, W. M.;Cushman, R. A.;Keele, J. W.;Maltecca, C.;Thomas, M. G.;Fortes, M. R. S. and Reverter, A. 2013. Breeding and genetics symposium: Networks and pathways to guide genomic selection. J. Anim. Sci. 91:537-552.

Tahir, M. S.;Porto-Neto, L. R.;Gondro, C.;Shittu, O. B.;Wockner, K.;Tan, A. W. L.;Smith, H. R.;Gouveia, G. C.;Kour, J. and Fortes, M. R. S. 2021. Meta-analysis of heifer traits identified reproductive pathways in bos indicus cattle. Genes (Basel) 12.

van den Berg, I.;Bowman, P. J.;MacLeod, I. M.;Hayes, B. J.;Wang, T.;Bolormaa, S. and Goddard, M. E. 2017. Multi-breed genomic prediction using bayes r with sequence data and dropping variants with a small effect. Genet. Sel. Evol. 49:70.

Weller, J. I.;Ezra, E. and Ron, M. 2017. Invited review: A perspective on the future of genomic selection in dairy cattle. J. Dairy Sci. 100:8633-8644.

Weller, J. I. and Ron, M. 2011. Invited review: Quantitative trait nucleotide determination in the era of genomic selection. J. Dairy Sci. 94:1082-1090.

Xiang, R.;MacLeod, I. M.;Daetwyler, H. D.;de Jong, G.;O'Connor, E.;Schrooten, C.;Chamberlain, A. J. and Goddard, M. E. 2021. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. Nat. Commun. 12:860.

Yang, J. A.;Lee, S. H.;Goddard, M. E. and Visscher, P. M. 2011. Gcta: A tool for genome-wide complex trait analysis. American Journal of Human Genetics 88:76-82.

7 Appendix 1: Summary and Manual for PhenoBank Database

PhenoBank Database Summary

A final summary note prepared by Bryce Little (CSIRO) and Cody Wilson (ABRI) – 23/11/2021

Developing the Phenobank database and user interfaces for both Windows and Web Browser environments (Collectively known as the PhenoBank Database) was particularly challenging given the constraints of budget and time available to us. The project would not have been possible to complete to the level achieved without the prior development of the Livestock Information Platform (LIP), an API accessible database developed jointly by CSIRO and ABRI.

The first requirement was to build a database structure and tools capable of storing and retrieving phenotype data. This was achieved in part due to the capabilities of LIP. However, new work was completed to enable the required merging of datasets.

The second requirement, and the most challenging nature of the build was the creation of a genotype database capable of efficiently storing and retrieving large quantities of SNP data. We concentrated upon the efficient storage of the SNP data itself, which was successfully achieved using prior capability within LIP, while keeping the storage of Genomic Map information as simple tables. The Map capability however needs more investment to enhance the speed of uploading genomic maps with large (one million plus) numbers of SNP. In short, further investment is required to manage sequence level data.

The greatest technical challenge was successfully met with the capability of merging multiple genomic datasets into single datasets using common SNP names. This was achieved and is built into the LIP system and available to the applications that access it.

The final challenge was to develop a comprehensive interface to the PhenoBank database. We elected to initially develop this via a Windows interface, due to prior work done within previous LIP based projects. The result is a comprehensive tool suitable for an expert user to upload, download and search phenotype and genotype data. “Administrators” who manage the PhenoBank Database access can also elect to add and remove user access.

The Windows interface was supplemented with a less comprehensive Web interface that includes basic functionality to download phenotype and genotype data. We recommend further investment in the Web Interface so that the capability can be enhanced to a level approaching that of the Windows interface.

PhenoBank Database Manual

Phenobank DB Manual

For the use of Dataset Viewers, Downloaders, Uploaders, Document Archivists and Administrators

Bryce Little, CSIRO & Cody Wilson, ABRI

19th Apr 2021

Version 1.0.25

Contents

Contents

1 What is Phenobank DB?

2 Setup Phenobank DB

3 Main Window

3.1 Log In 44

3.2 Dataset Viewer 44

3.3 Download 45

3.4 Upload & Manage 45

3.5 Documents 45

3.6 Administrator 45

4 Log In

4.1 Log in Window 46

4.2 Log in Field Definitions 46

4.3 LOG IN function 47

5 Dataset Viewer

5.1 View List of Datasets 49

5.2 Select a Dataset 49

5.3 View list of Column Names in a Dataset 50

5.4 View Dataset Details 51

5.5 View Dataset Sample 52

5.6 View Dictionary 53

5.6.1 What is the Dictionary? 53

5.7 View Units 54

5.7.1 What are Units? 54

6 Download

6.1 Selecting Datasets 56

6.2 Selecting Columns for merged Dataset 58

6.3 Downloading merged file 58
6.4 Settings 60
7 Search Download
7.1 Parameters 61
7.2 Result 64
8 Upload & Manage
8.1 Defining User Access 66
8.2 Uploading a Dataset 70
8.2.1 Uploading a NEW Dataset 70
8.2.2 Uploading and merging with an EXISTING dataset 73
9 Documents
9.1 Uploading Documents 77
9.2 Searching for Document 80
9.2.1 Selecting a document from list of documents 80
9.2.2 Searching for a range of documents 81
10 Administrator
10.1 Database Management 83
10.1.1 Setting Permissions 83
10.1.2 User Access 85
10.1.3 Master User 85
11 Genomic Manager
11.1 Accessing the Genomic Manager 87
11.2 The Genomic Catalog 88
11.2.1 Delete Genomic Catalog Entry 89
11.3 Genomic dataset relationships 90
11.4 Import & Upload 91
11.4.1 Import Genomic File 91
11.4.2 Importing “Simple” format files 92
11.4.3 Upload Genomic Data 92
11.4.4 Reset button 93

11.4.5 "Check is in catalog" button 93

11.5 Download & Export 94

11.5.1 Download Genomic Dataset 94

11.5.2 Export Genomic Dataset 94

11.5.3 One Step Download and Export 95

11.5.4 Parentage search 95

11.6 Link Genotype Sample ID to Animal ID 96

11.6.1 Select Genomic datasets from Catalog....................96

11.6.2 Import Animal ID file....................96

11.7 Link Phenotype Dataset to Genotype 96

11.8 Duplicate ID Test 97

12 Generating Phenobank ID

12.1 Accessing Phenobank ID Generator 98

12.2 Generating Phenobank ID's 98

13 Protocols

13.1 Phenobank upload protocol for phenotypes.................... 100

13.1.1 Phenotype source file 100

13.1.2 Uploading phenotype data 100

13.1.3 Generating a unique identifier 100

13.2 Phenobank upload protocol for genotypes.................... 100

13.2.1 Acceptable genotype source files....................100

13.2.2 Uploading Illumina files 100

13.2.3 Uploading PLINK files....................101

14 Format Descriptions

14.1 Simple Format.................... 102

14.1.1 Simple "geno" file type....................102

14.1.2 Simple "map" file type....................102

15 Phenobank on the Web

15.1 Accessing the Phenobank Web Page 103

15.2 The Phenobank Dashboard Page 104

15.3 Export Phenotype Data 104

15.3.1 Select Dataset for export 104

15.3.2 Select Columns for export 105

15.3.3 Select Key (ID) column & export 106

15.4 Export Genomic Data 107

15.4.1 Export Genomic Functions 107

15.4.2 Search and select a genomic dataset 107

15.4.3 Exporting a genomic dataset 108

1 What is Phenobank DB?

Phenobank DB is a uniquely designed tool that works with the cloud based “Livestock Information Platform” (LIP) that will help store and retrieve livestock phenotype and genotype data.

With the help of the Phenobank DB tool, users will be able to upload data, merge data, allow and restrict access to data and download data for analysis.

The Phenobank DB Tool is now available for both Windows and Web environments. The Windows App is a comprehensive tool for uploading data, downloading data and general database management.

The Web based Phenobank Application is designed to meet the needs of casual researchers who wish to download phenotype and genotype datasets.

2 Setup Phenobank DB

Phenobank can be installed on your PC using the Phenobank DB "Setup" application. This can be downloaded at the following web location:

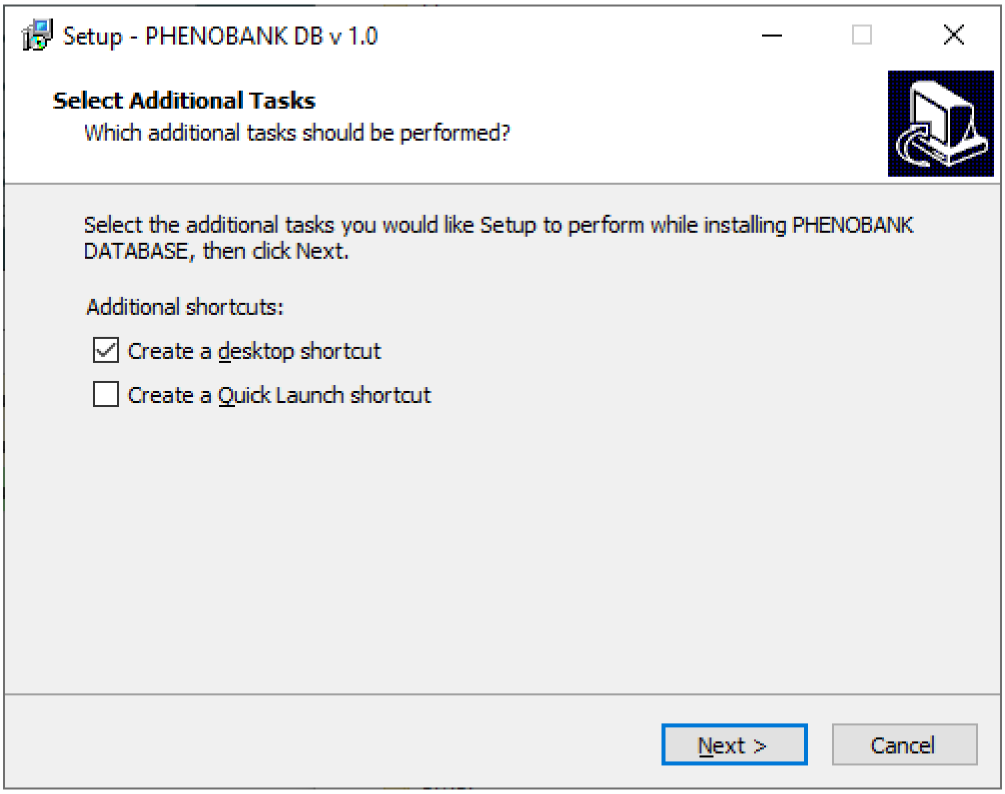
http://lipstorage.blob.core.windows.net/acc-lipadmin-abri-une-edu-au-phenobank/setup_PhenobankDB_15.exe

NOTE! You will need "Administrator" privileges on you Windows PC to be able to install Phenobank DB. If you are using a PC belonging to your employer, it is possible that you do not have Administrator privileges. In this case you will need to contact your IT department.

To install follow these steps:

1. Run this application to commence the installation.
2. You may be asked by Windows if you wish to allow the application to make changes to Windows. Answer 'yes' to this question.

You will be rewarded with initial installation window:



Setup - PHENOBANK DB v 1.0

Select Additional Tasks

Which additional tasks should be performed?

Select the additional tasks you would like Setup to perform while installing PHENOBANK DATABASE, then click Next.

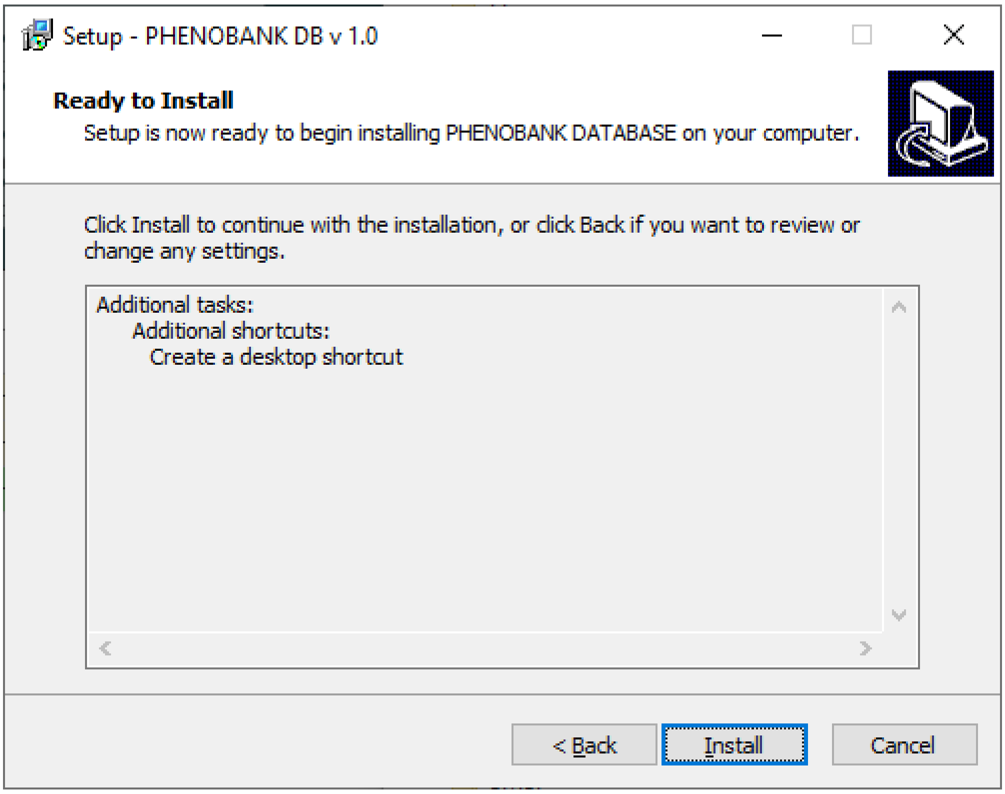
Additional shortcuts:

☑ Create a desktop shortcut

□ Create a Quick Launch shortcut

Choose the shortcut options, if you wish.

Click "Next" to get to the next step which will appear as follows:



Setup - PHENOBANK DB v 1.0

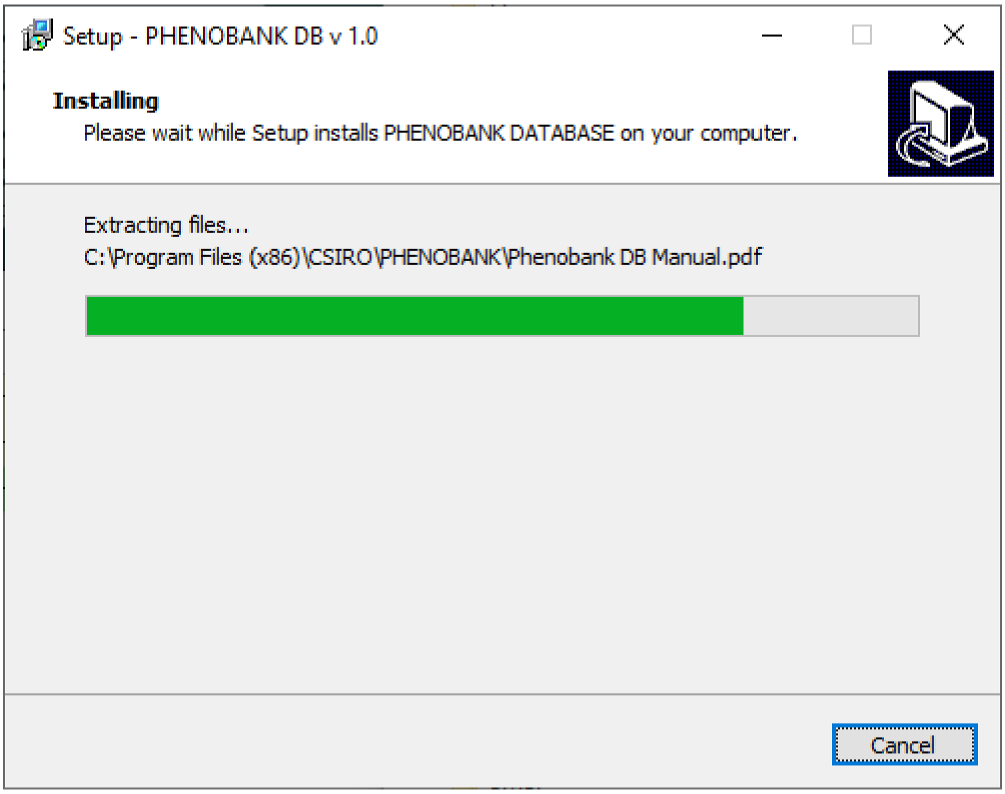
Ready to Install

Setup is now ready to begin installing PHENOBANK DATABASE on your computer.

Click Install to continue with the installation, or click Back if you want to review or change any settings.

Additional tasks:
Additional shortcuts:
Create a desktop shortcut

Click “Install” to complete the installation. You will see a window similar to the following:



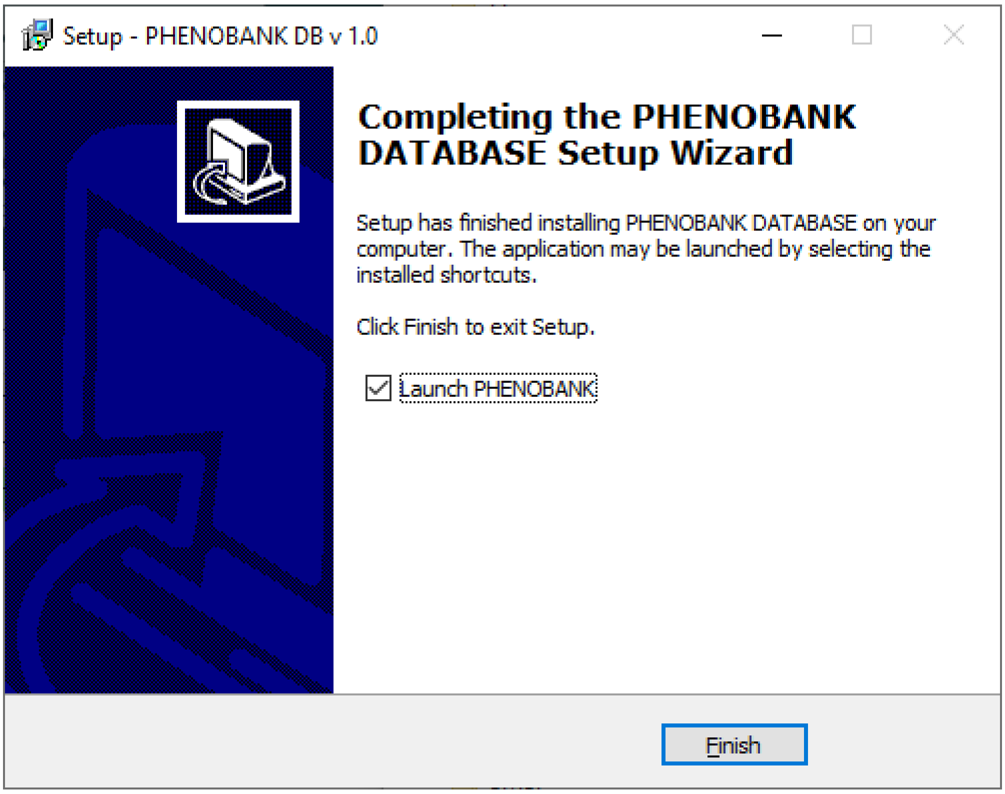
Setup - PHENOBANK DB v 1.0

Installing

Please wait while Setup installs PHENOBANK DATABASE on your computer.

Extracting files...
C:\Program Files (x86)\CSIRO\PHENOBANK\Phenobank DB Manual.pdf

When the installation has completed, you will be rewarded with the final screen as follows:



Completing the PHENOBANK DATABASE Setup Wizard

Setup has finished installing PHENOBANK DATABASE on your computer. The application may be launched by selecting the installed shortcuts.

Click Finish to exit Setup.

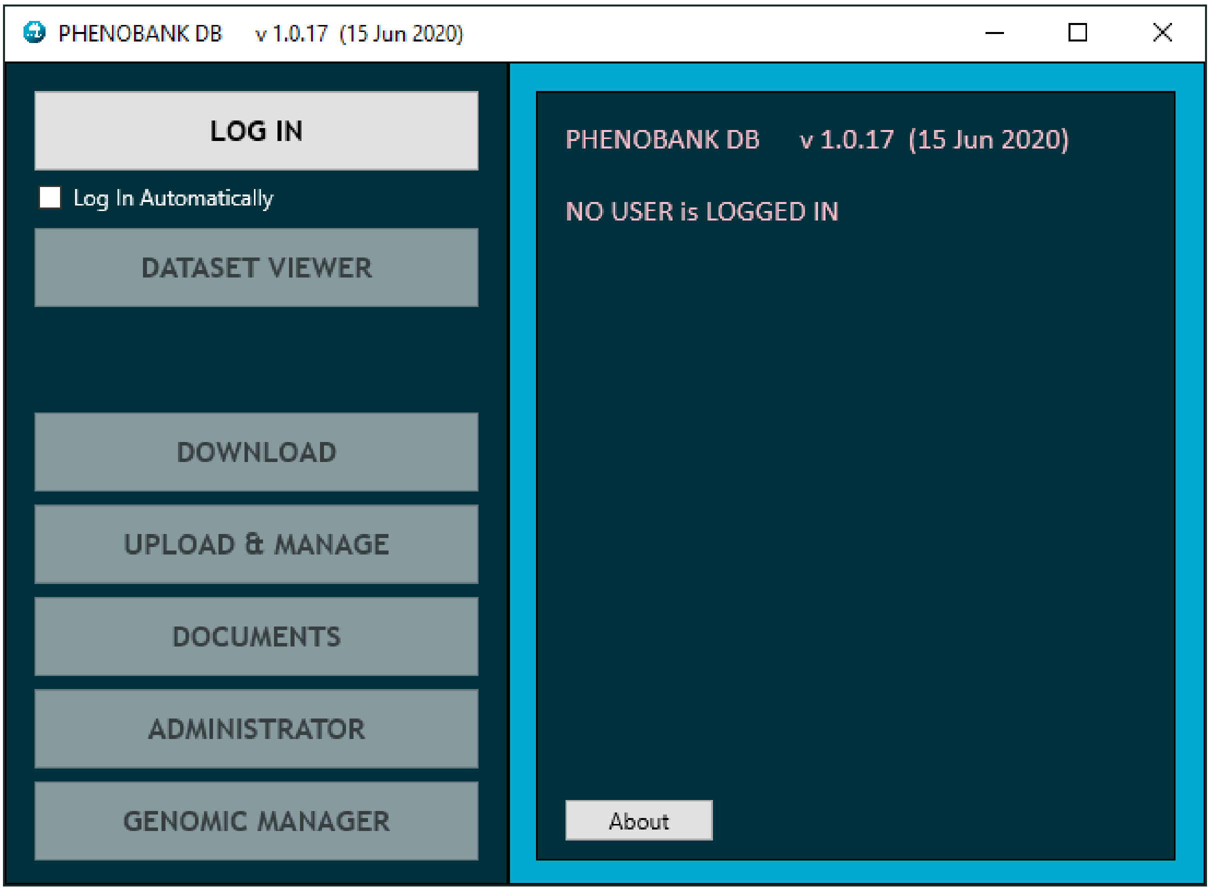
☑ Launch PHENOBANK

If you wish to run Phenobank DB straight away, tick “Launch Phenobank”.

3 Main Window

The Phenobank DB functions are accessible via the Main Window.

At time of publication, the main window appears as follows:



PHENOBANK DB   v 1.0.17 (15 Jun 2020)

NO USER is LOGGED IN

All the major functions in Phenobank DB are accessible via the Main Window, after the user has successfully logged in. Their purpose is described in the following sections.

3.1 Log In

All users must have an account before they can upload or download data. Accounts are created by users with “administrative” permissions. Typically, administrative user accounts are created by the LIP licensee (currently ABRI).   See the “Log In” chapter for more information.

3.2 Dataset Viewer

The Dataset Viewer has a simple interface that allows users to see a “snapshot” sample of data in each of the datasets available to them.

3.3 Download

The Download function has two types of Downloads. The “Merge Download” allows users to choose which columns in which datasets they wish to download. Where column data is associated with data in companion columns, these are automatically downloaded as well. Data from more than one dataset can downloaded, one after the other, into a single merged file. Columns with common names or a shared dictionary definition will appear in the same columns.

The “Search Download” function allows users to search SINGLE datasets for records that satisfy a range of values for up to three columns. Resulting datasets can be saved in several formats including CSV and TAB delimited.

3.4 Upload & Manage

Providers of data can upload the data. Providers can then annotate the data, describing each column, and associating columns to each other, or link a column to a dictionary definition. Dictionaries of column definitions are shared amongst one or more columns in one or more datasets. Data providers can also decide who has access to the datasets that they have provided, and what permission those users have, such as just reading the data, or adding to the data.

3.5 Documents

Users can upload and download “Documents” to share with either the public, or only those with access to the Phenobank Document Catalog. The documents can be ANY file such as Word doc’s, pdf’s, jpg’s, text or any data or binary file. Each document can be annotated with a document “Catalog” entry, to assist future users to locate.

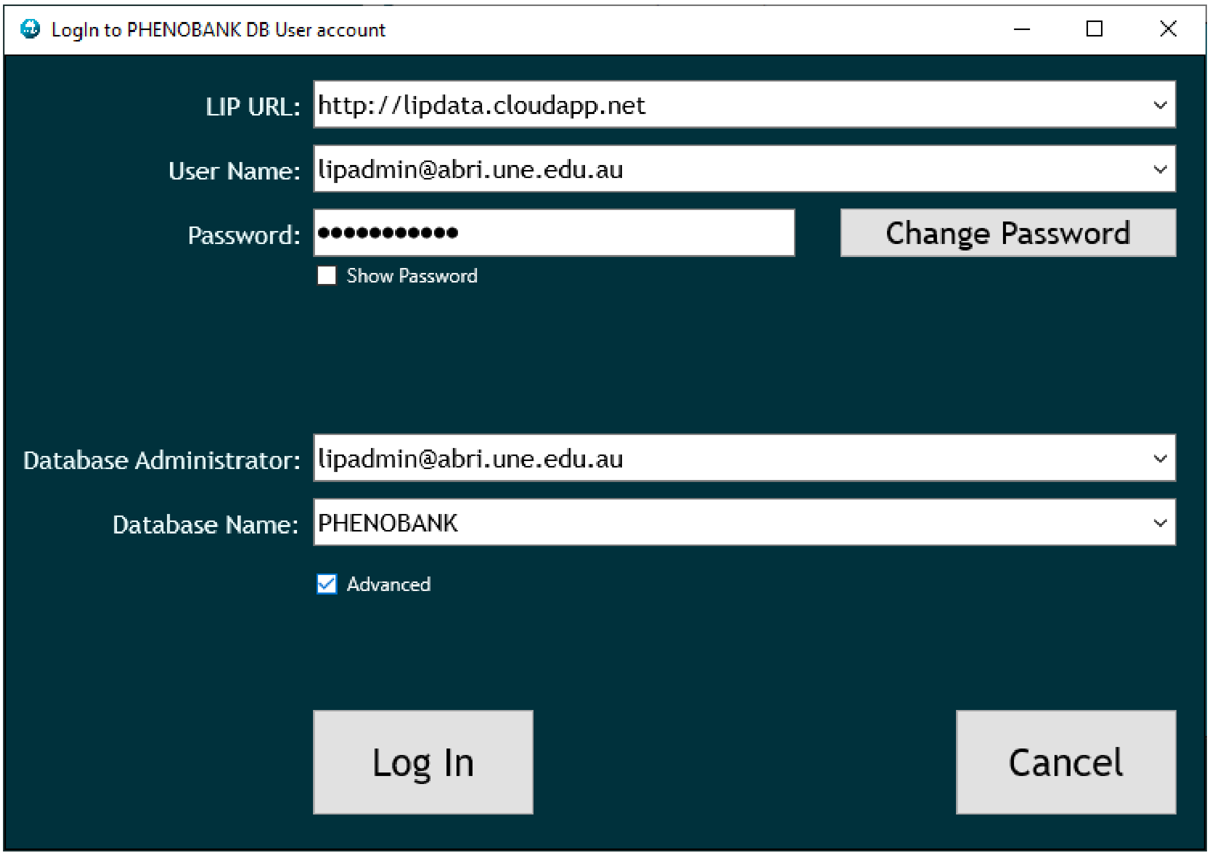
3.6 Administrator

The Administrator functions include many capabilities for managing all the datasets in the database.

4 Log In

4.1 Log in Window

Click the “Log In” button on the main screen. The following Windows will appear:



LogIn to PHENOBANK DB User account

LIP URL: http://lipdata.cloudapp.net

User Name: lipadmin@abri.une.edu.au

Password: ●●●●●●●●●●●●

□ Show Password

Database Administrator: lipadmin@abri.une.edu.au

Database Name: PHENOBANK

☑ Advanced

The following fields need to be populated. They are described as follows:

4.2 Log in Field Definitions

7.1.1.1 LIP URL:

This is the Web Address of the cloud database where the Phenobank DB is stored. Select the above from the drop-down menu.

7.1.1.2 Username:

This is YOUR username. The PHENOBANK DB project administrator will create your user account for you, or if you are a previous LIP user, your existing user name could be used.

7.1.1.3 Password:

This is the password assigned to you. Once you have logged in, you can select this Log In Window again and click the “Change Password” button to change your password.

7.1.1.4 Database Administrator:

This is the user account the contains the database where the Phenobank datasets are stored. This name will be supplied to you by the Phenobank DB administrators.

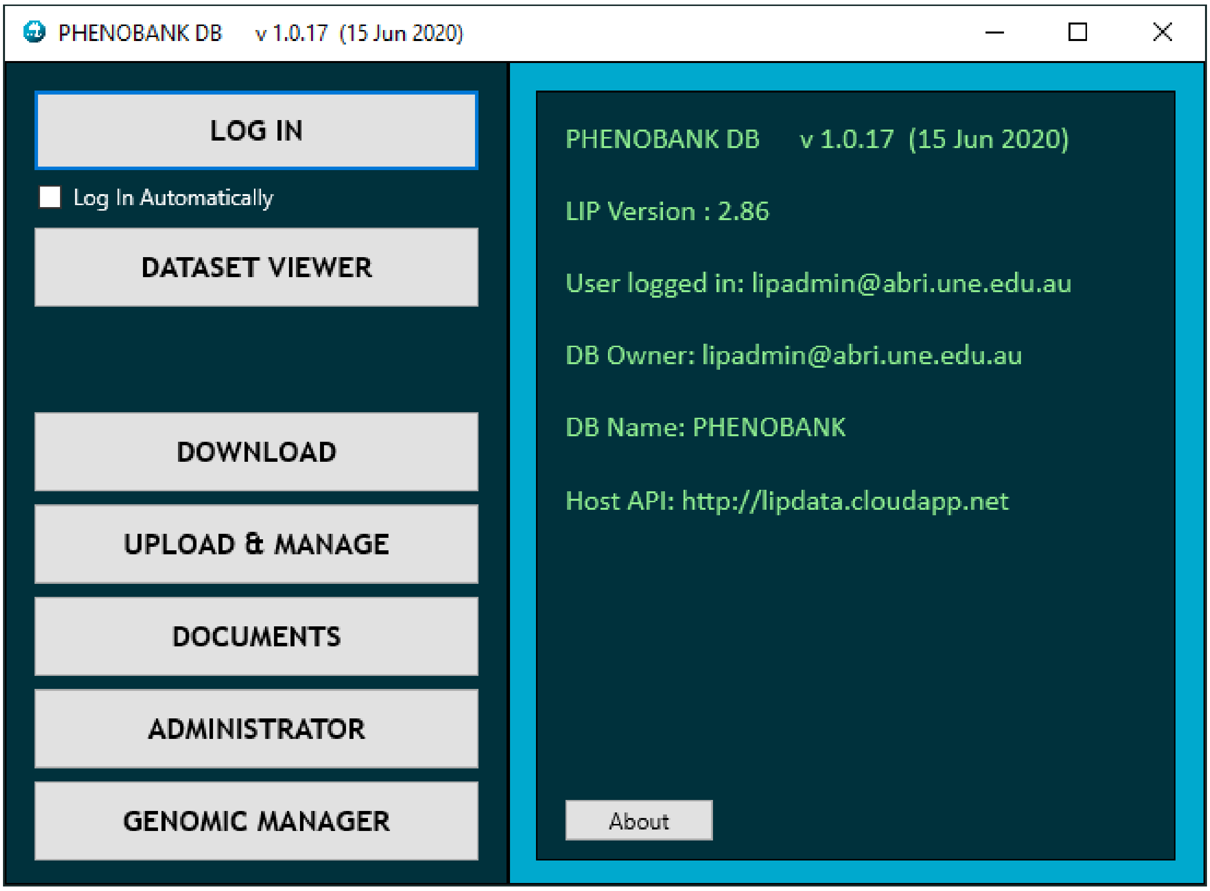
7.1.1.5 Database Name:

7.1.1.6 This is the name of the Database within the Database Owner's account where the Datasets will be stored.

NOTE that the values you enter in these fields will be remembered next time you Log In. For security purposes passwords are stored encrypted on your local personal computer, and accessible and decrypted only by this application.

4.3 LOG IN function

Finally, click the "Log in" button to gain access to the LIP system. You should be rewarded with an updated Main Window that will show you that you have successfully logged in with Log In details in green in the left-hand pane like this:



PHENOBANK DB v 1.0.17 (15 Jun 2020)

LIP Version : 2.86

User logged in: lipadmin@abri.une.edu.au

DB Owner: lipadmin@abri.une.edu.au

DB Name: PHENOBANK

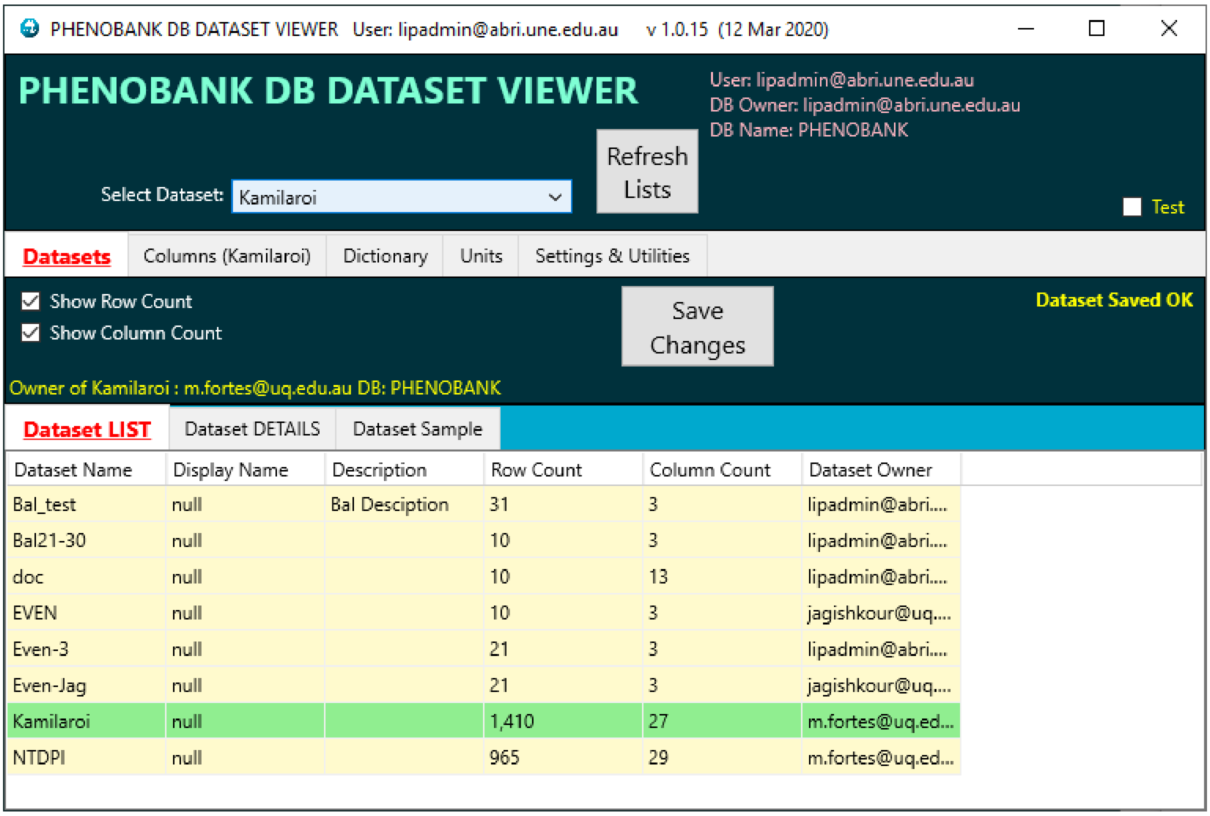
Host API: http://lipdata.cloudapp.net

7.1.1.7 User Tip:

Tick the “Log In Automatically” box, so that when you next run PHENOBANK DB, you will automatically be logged into the PHENOBANK DB.

5 Dataset Viewer

Click the “DATASET VIEWER” button on the Main Window to see the Dataset Viewer, like the following.



User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: Kamilaroi

☑ Show Row Count
☑ Show Column Count

Dataset Saved OK

Owner of Kamilaroi : m.fortes@uq.edu.au DB: PHENOBANK

Dataset Name	Display Name	Description	Row Count	Column Count	Dataset Owner
Bal_test	null	Bal Desciption	31	3	lipadmin@abri....
Bal21-30	null		10	3	lipadmin@abri....
doc	null		10	13	lipadmin@abri....
EVEN	null		10	3	jagishkour@uq....
Even-3	null		21	3	lipadmin@abri....
Even-Jag	null		21	3	jagishkour@uq....
Kamilaroi	null		1,410	27	m.fortes@uq.ed...
NTDPI	null		965	29	m.fortes@uq.ed...

5.1 View List of Datasets

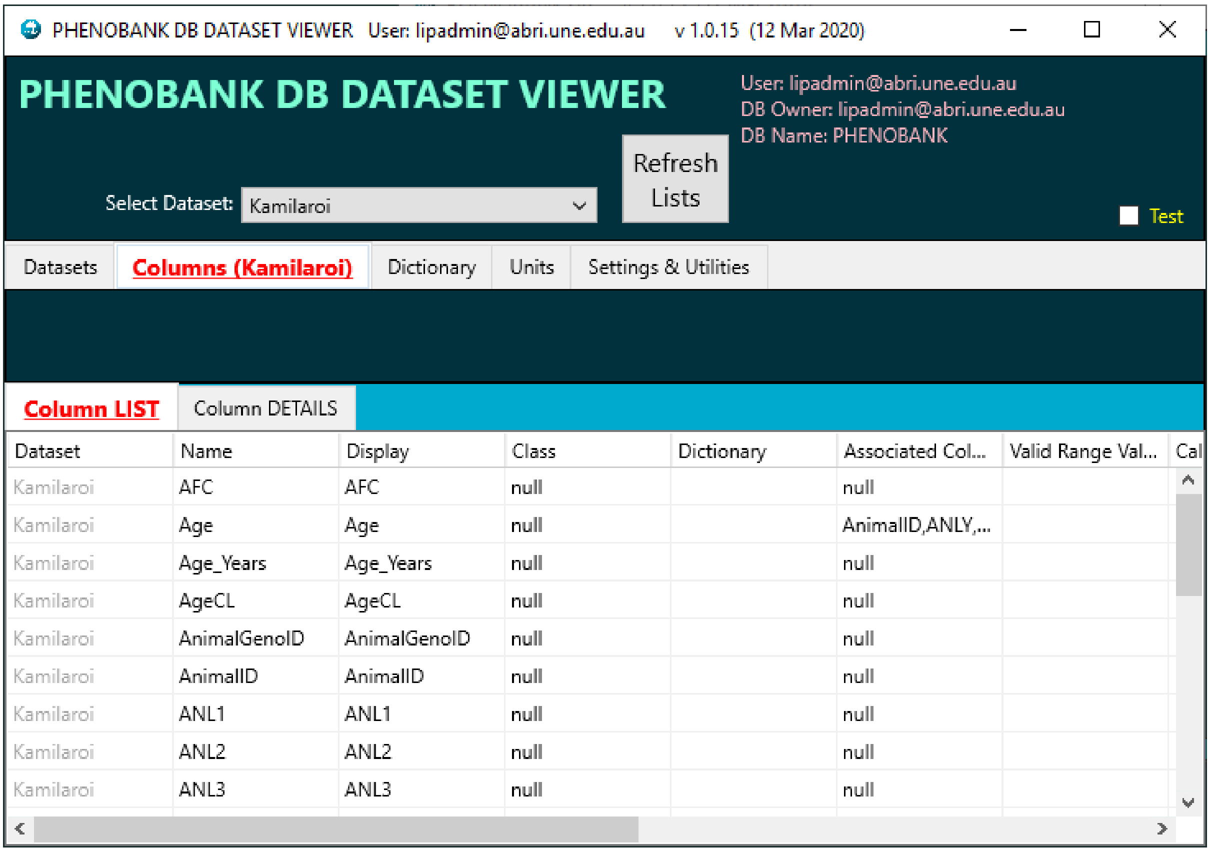
Select the “Datasets” tab and the “Dataset List” tab, as above.

5.2 Select a Dataset

To select a Dataset for further investigation, select from the Drop-down list labelled “Select Dataset” OR click on the row with the Dataset named in the Dataset List. When selected the chosen Dataset will be highlighted.

5.3 View list of Column Names in a Dataset

Click on the “Columns” tab to view the list of columns in the selected Dataset.

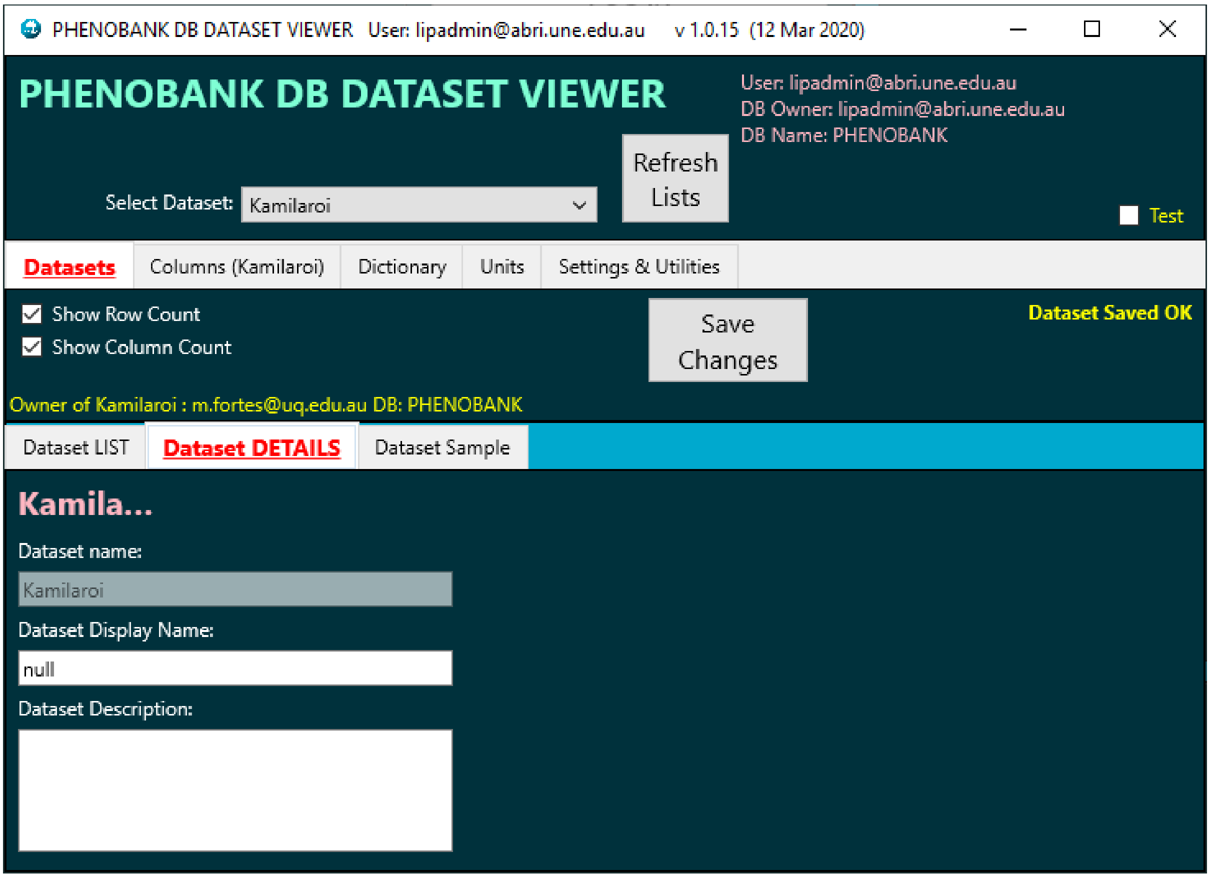

PHENOBANK DB DATASET VIEWER User: lipadmin@abri.une.edu.au v 1.0.15 (12 Mar 2020)
PHENOBANK DB DATASET VIEWER
User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK
Refresh Lists
Select Dataset: Kamilaroi
□ Test
Datasets
Columns (Kamilaroi)
Dictionary
Units
Settings & Utilities
Column LIST
Column DETAILS

Dataset	Name	Display	Class	Dictionary	Associated Col...	Valid Range Val...	Cal
Kamilaroi	AFC	AFC	null		null		
Kamilaroi	Age	Age	null		AnimalID,ANLY,...		
Kamilaroi	Age_Years	Age_Years	null		null		
Kamilaroi	AgeCL	AgeCL	null		null		
Kamilaroi	AnimalGenoID	AnimalGenoID	null		null		
Kamilaroi	AnimalID	AnimalID	null		null		
Kamilaroi	ANL1	ANL1	null		null		
Kamilaroi	ANL2	ANL2	null		null		
Kamilaroi	ANL3	ANL3	null		null		

5.4 View Dataset Details

Select the “Dataset DETAILS” tab to view the descriptive details of a dataset. In addition to the dataset name this includes:

- “Display name” to be used as an alternative name for Graphs or Reports.
- “Dataset description” to describe the dataset. This would be written by the original provider of the dataset.



PHENOBANK DB DATASET VIEWER User: lipadmin@abri.une.edu.au v 1.0.15 (12 Mar 2020)

PHENOBANK DB DATASET VIEWER

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: Kamilaroi

☑ Show Row Count
☑ Show Column Count

Dataset Saved OK

Owner of Kamilaroi : m.fortes@uq.edu.au DB: PHENOBANK

Kamila...

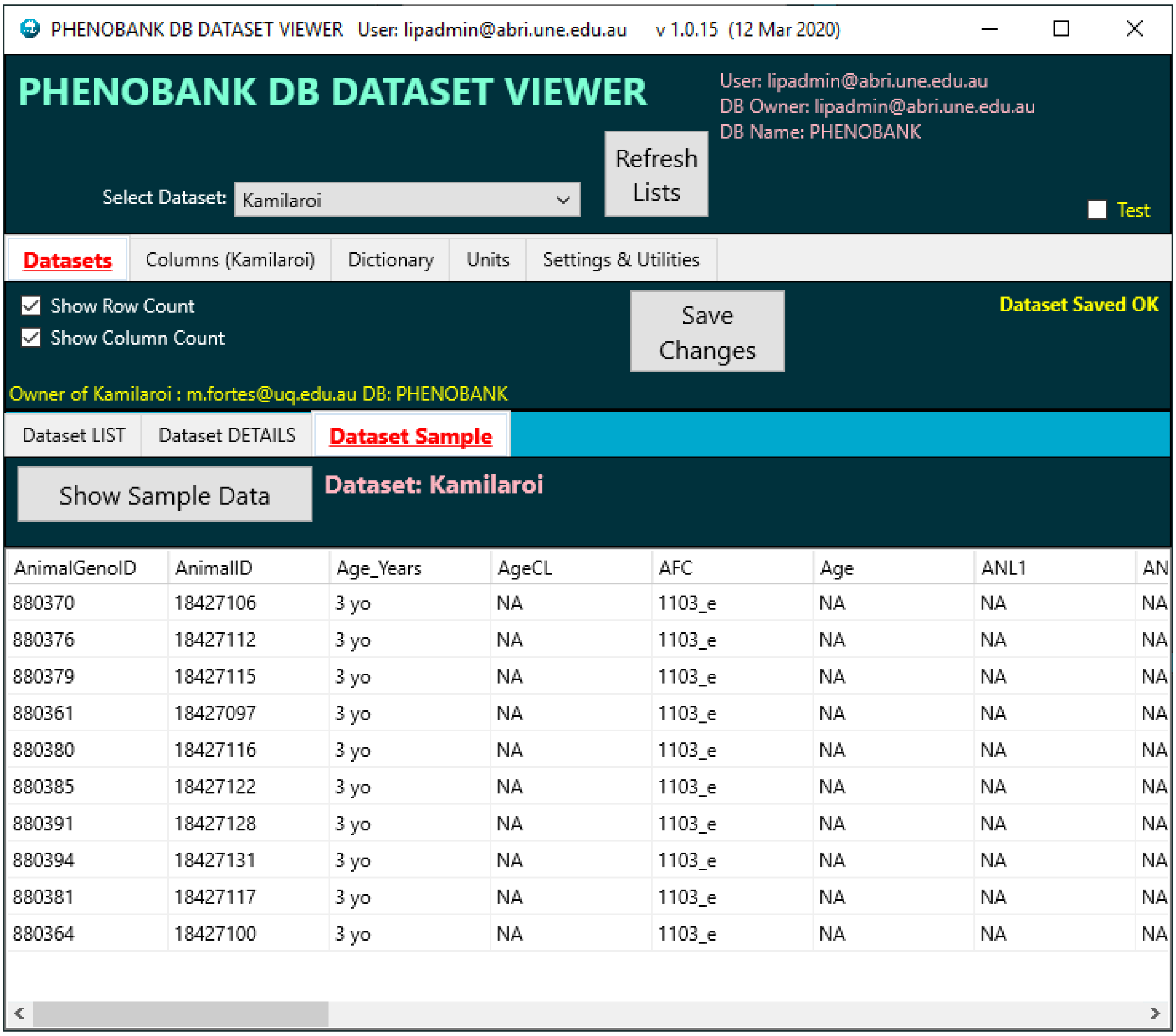
Dataset name: Kamilaroi

Dataset Display Name: null

Dataset Description:

5.5 View Dataset Sample

Select the “Dataset Sample” tab. If you have permission to see the whole Dataset, the first 10 lines of the currently selected dataset will be shown when you click the “Show Sample Data” button. If you do not have permission to see the whole dataset, only 5 lines will be shown.



PHENOBANK DB DATASET VIEWER   User: lipadmin@abri.une.edu.au   v 1.0.15 (12 Mar 2020)

PHENOBANK DB DATASET VIEWER

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: Kamilaroi

☑ Show Row Count
☑ Show Column Count

Dataset Saved OK

Owner of Kamilaroi : m.fortes@uq.edu.au DB: PHENOBANK

Dataset: Kamilaroi

AnimalGenoID	AnimalID	Age_Years	AgeCL	AFC	Age	ANL1	AN
880370	18427106	3 yo	NA	1103_e	NA	NA	NA
880376	18427112	3 yo	NA	1103_e	NA	NA	NA
880379	18427115	3 yo	NA	1103_e	NA	NA	NA
880361	18427097	3 yo	NA	1103_e	NA	NA	NA
880380	18427116	3 yo	NA	1103_e	NA	NA	NA
880385	18427122	3 yo	NA	1103_e	NA	NA	NA
880391	18427128	3 yo	NA	1103_e	NA	NA	NA
880394	18427131	3 yo	NA	1103_e	NA	NA	NA
880381	18427117	3 yo	NA	1103_e	NA	NA	NA
880364	18427100	3 yo	NA	1103_e	NA	NA	NA

5.6 View Dictionary

5.6.1 What is the Dictionary?

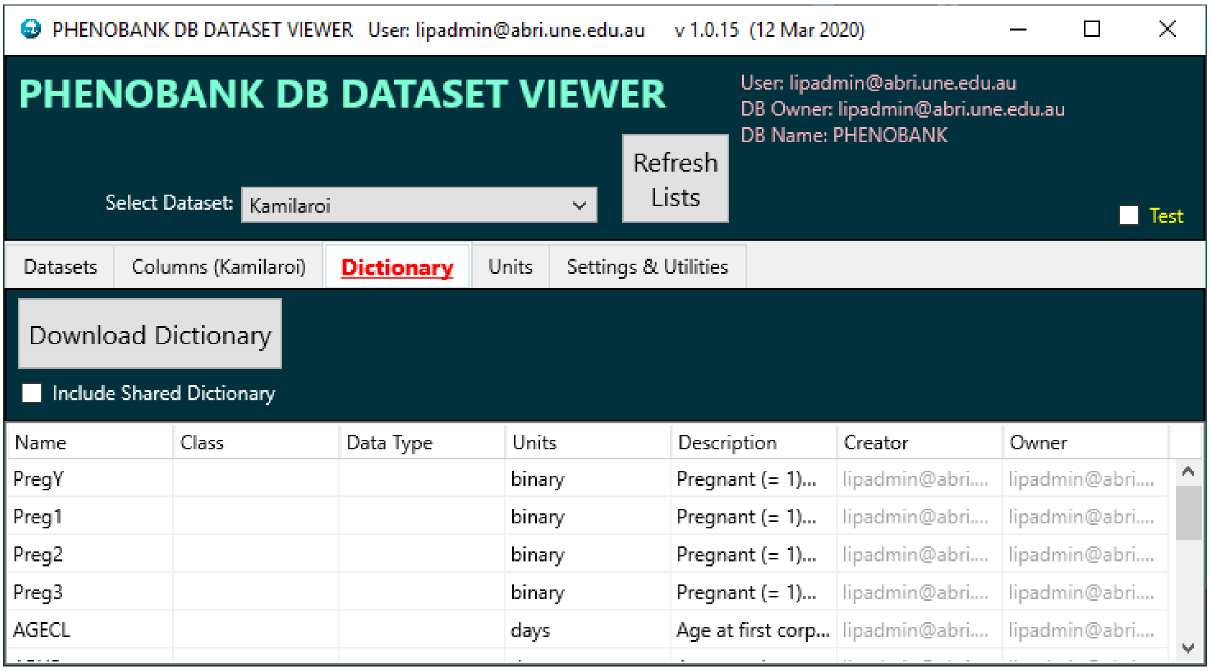
The Dictionary contains a list of Dictionary Definitions. Any column may be defined with a single Dictionary Definition. Through this capability, PHENOBANK DB is able to associate columns in one dataset with columns of another dataset.

A Dictionary definition is created by a user who can upload a dataset to a database.

The Dictionary is shared with all Datasets in all Databases within a user’s account.

Each Dictionary Definition can be defined by several parameters. These are:

- Name – The given name of the definition
- Class – A name given to group definitions into like classifications. Examples might be “Identifier” or “Phenotype”.
- Data Type – The kind of data being stored, such as “Date”, “Integer” or “Text”.
- Units – The name of the Units used for measuring data that is defined by this Dictionary definition. See the Units Table for the full description of Units.
- Nature – For future use
- Scope – For future use
- Description – A user defined description of the Dictionary Definition
- Creator – The user name of the Creator of this Definition.
- Owner – The owner of the record, who has permission to alter or delete the item.



PHENOBANK DB DATASET VIEWER User: lipadmin@abri.une.edu.au v 1.0.15 (12 Mar 2020)

PHENOBANK DB DATASET VIEWER

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: Kamilaroi

Name	Class	Data Type	Units	Description	Creator	Owner
PregY			binary	Pregnant (= 1)...	lipadmin@abri....	lipadmin@abri....
Preg1			binary	Pregnant (= 1)...	lipadmin@abri....	lipadmin@abri....
Preg2			binary	Pregnant (= 1)...	lipadmin@abri....	lipadmin@abri....
Preg3			binary	Pregnant (= 1)...	lipadmin@abri....	lipadmin@abri....
AGECL			days	Age at first corp...	lipadmin@abri....	lipadmin@abri....

5.7 View Units

5.7.1 What are Units?

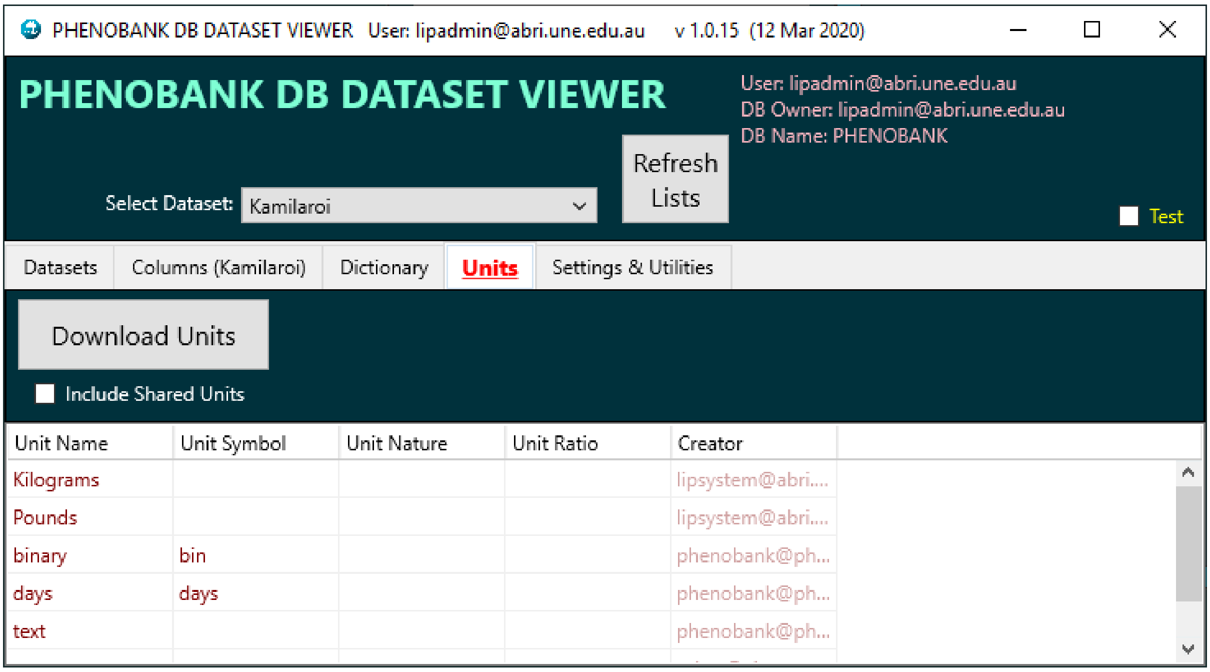
Units are the measurement unit types used for defining data values.

Lists of Units can be set up for all a user's account, but also a system wide standard list of units are available for all users to use.

Each Dictionary definition can optionally be attached to a Unit.

Each Unit has the following parameters:

- Unit name – A unique name for the unit
- Unit Symbol – A short name or symbol to represent the unit.
- Unit nature – Units the measure the same physical type, such as "mass". In this case the unit "kg" or "pound" or "g" could be given the same nature "mass".
- Unit Ratio – Units of the same nature can be converted to a common value by using the Unit Ratio. This is a number. For example, "g" might be given the value "1" and "kg" the value "1000" and "lb" the value "453.592".



PHENOBANK DB DATASET VIEWER   User: lipadmin@abri.une.edu.au   v 1.0.15 (12 Mar 2020)

PHENOBANK DB DATASET VIEWER

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: Kamilaroi

□ Test

□ Include Shared Units

Unit Name	Unit Symbol	Unit Nature	Unit Ratio	Creator
Kilograms				lipsystem@abri....
Pounds				lipsystem@abri....
binary	bin			phenobank@ph...
days	days			phenobank@ph...
text				phenobank@ph...

With this capability, future versions of Phenobank and other applications that use LIP will be able to translate data into common units for comparison or reporting purposes.

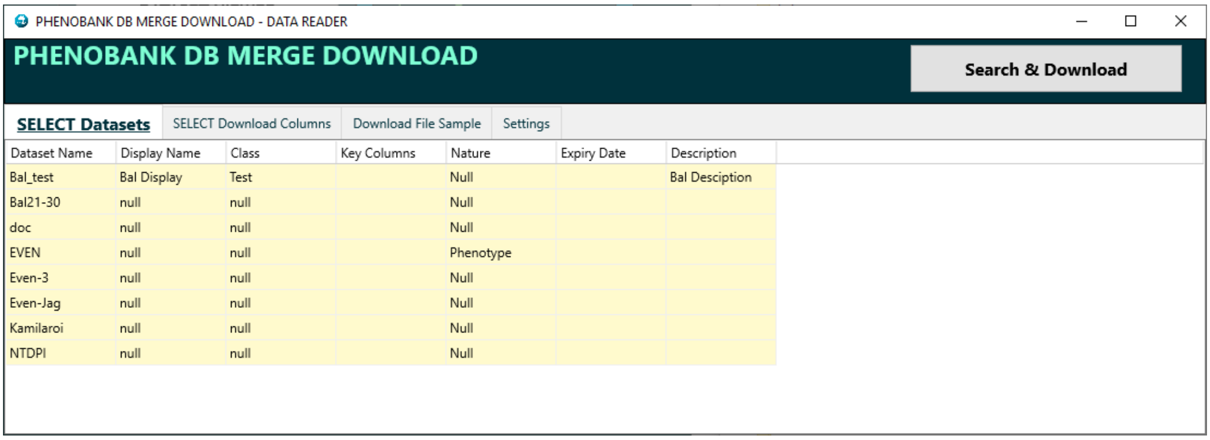
6 Download

The Download function allows access to two variations of Downloads:

- A “Merge Download” function allows multiple full datasets to be merged into one super-dataset with common columns.
- A “Search Download” function that allows more precise filtering for a set of records in a single dataset.

The “MERGE DOWNLOAD” Application allows users to select **one or more** datasets, merge them into a single dataset and download it.

Click the “DOWNLOAD” button on the Main Window to see the Dataset Viewer, like the following:

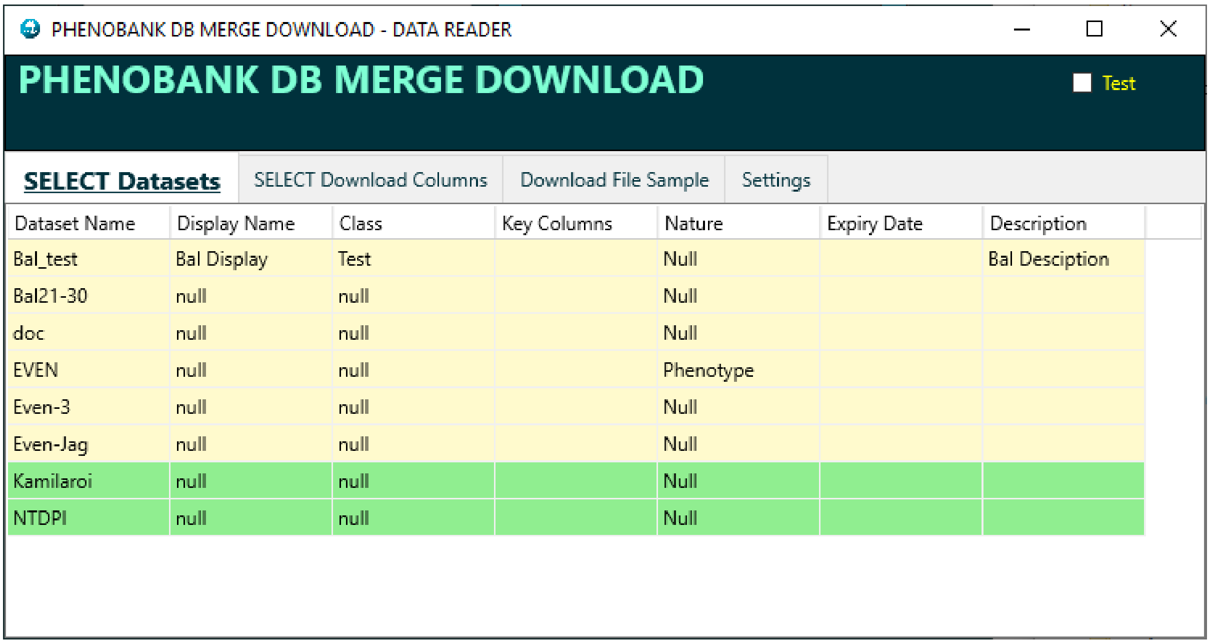


Dataset Name	Display Name	Class	Key Columns	Nature	Expiry Date	Description
Bal_test	Bal Display	Test		Null		Bal Desciption
Bal21-30	null	null		Null		
doc	null	null		Null		
EVEN	null	null		Phenotype		
Even-3	null	null		Null		
Even-Jag	null	null		Null		
Kamilaroi	null	null		Null		
NTDPI	null	null		Null		

6.1 Selecting Datasets

Click the “SELECT Datasets” tab.

Click on each Dataset that you wish to source data from. One or more may be selected. The selected datasets will be highlighted, as Dataset1, Dataset3 and Dataset4 are highlighted below:



Dataset Name	Display Name	Class	Key Columns	Nature	Expiry Date	Description
Bal_test	Bal Display	Test		Null		Bal Desciption
Bal21-30	null	null		Null		
doc	null	null		Null		
EVEN	null	null		Phenotype		
Even-3	null	null		Null		
Even-Jag	null	null		Null		
Kamilaroi	null	null		Null		
NTDPI	null	null		Null		

6.2 Selecting Columns for merged Dataset

Select the “SELECT Download Columns” tab.

Click on each column you wish to download.

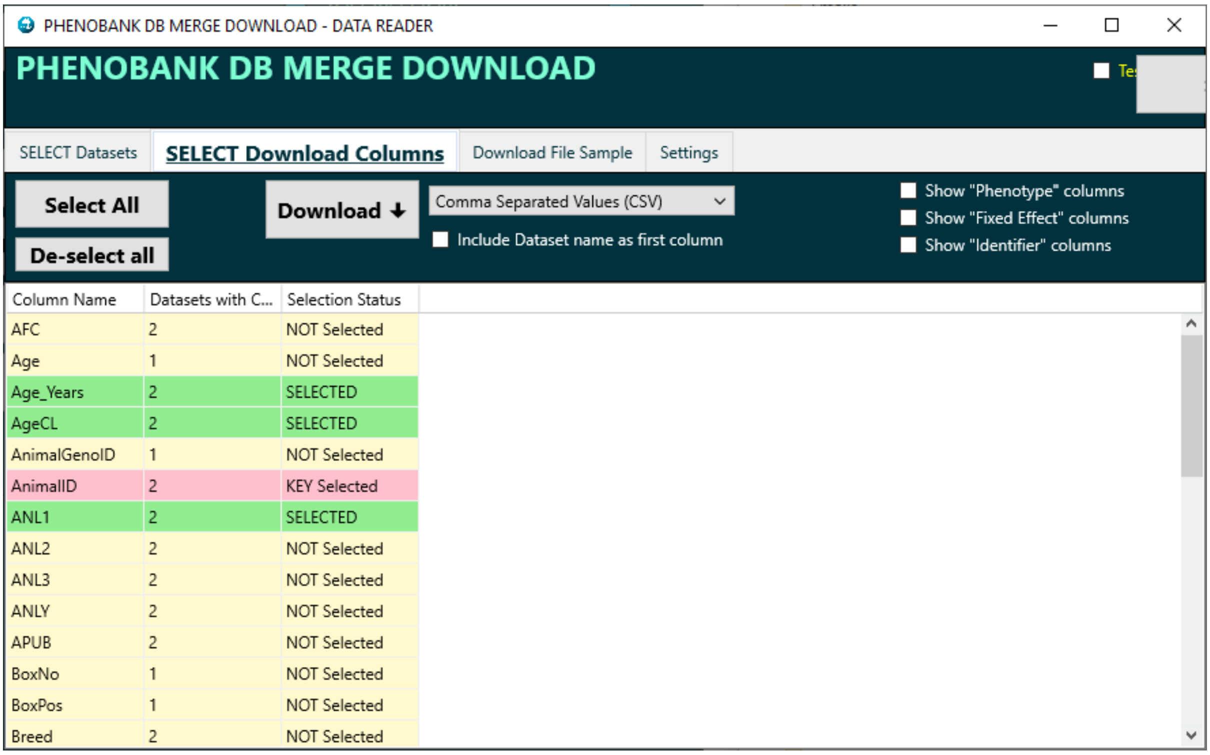
NOTE:

1. Where Columns are COMMON between datasets these columns will appear just once.
2. Where Columns have a DICTIONARY Definition, the *Definition name* will be SUBSTITUTED for the Column name.

Also supplied, for your information, is the number of Datasets each column appears in.

In this example four columns are selected. ONE of the selected columns must be clicked AGAIN to make it the “KEY”.   This is shown in PINK in the example below.

You may also RESTRICT the available columns shown in this list by ticking the boxes at the top right (Show “Phenotype” columns etc.)



Comma Separated Values (CSV)

☐ Include Dataset name as first column

☐ Show "Phenotype" columns

☐ Show "Fixed Effect" columns

☐ Show "Identifier" columns

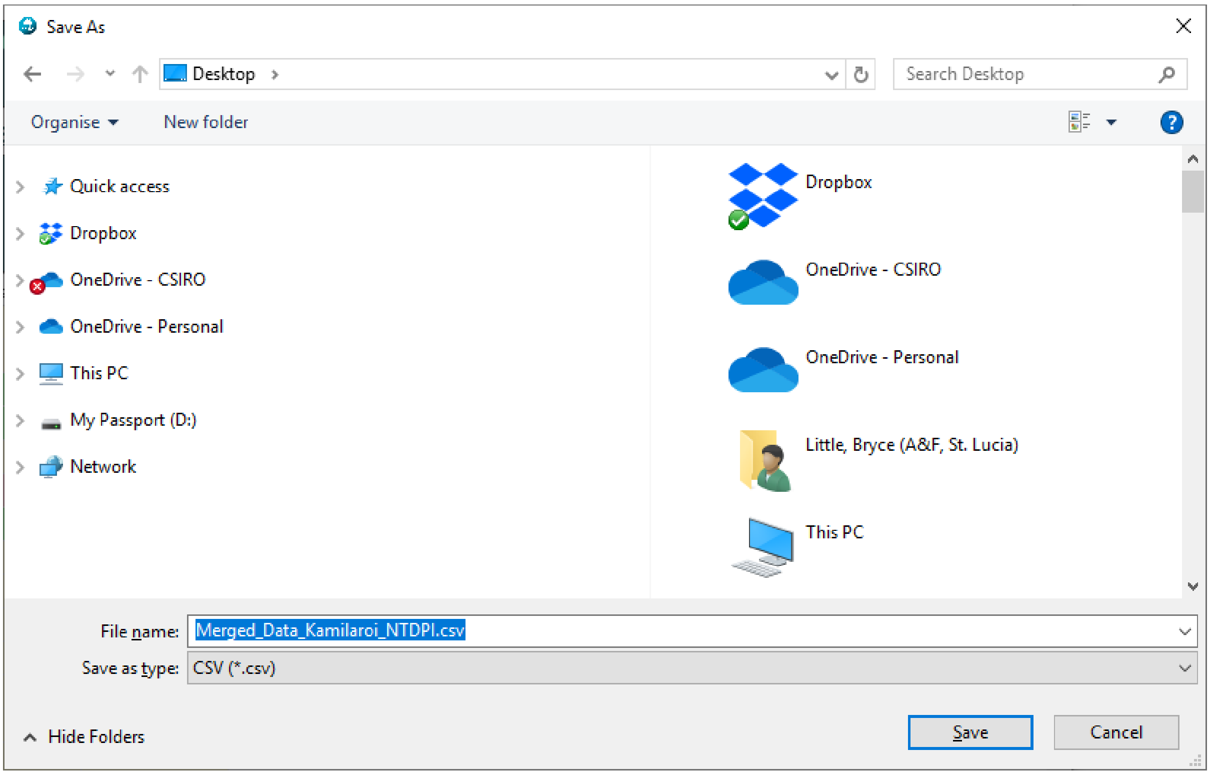
Column Name	Datasets with C...	Selection Status
AFC	2	NOT Selected
Age	1	NOT Selected
Age_Years	2	SELECTED
AgeCL	2	SELECTED
AnimalGenoID	1	NOT Selected
AnimalID	2	KEY Selected
ANL1	2	SELECTED
ANL2	2	NOT Selected
ANL3	2	NOT Selected
ANLY	2	NOT Selected
APUB	2	NOT Selected
BoxNo	1	NOT Selected
BoxPos	1	NOT Selected
Breed	2	NOT Selected

6.3 Downloading merged file

Users may download the file which will merge all of the datasets selected, by clicking the “Download” button in the “SELECT Download Columns” tab.

An additional column with the Dataset name can be downloaded by ticking the box labelled “Include Dataset name as first column”.

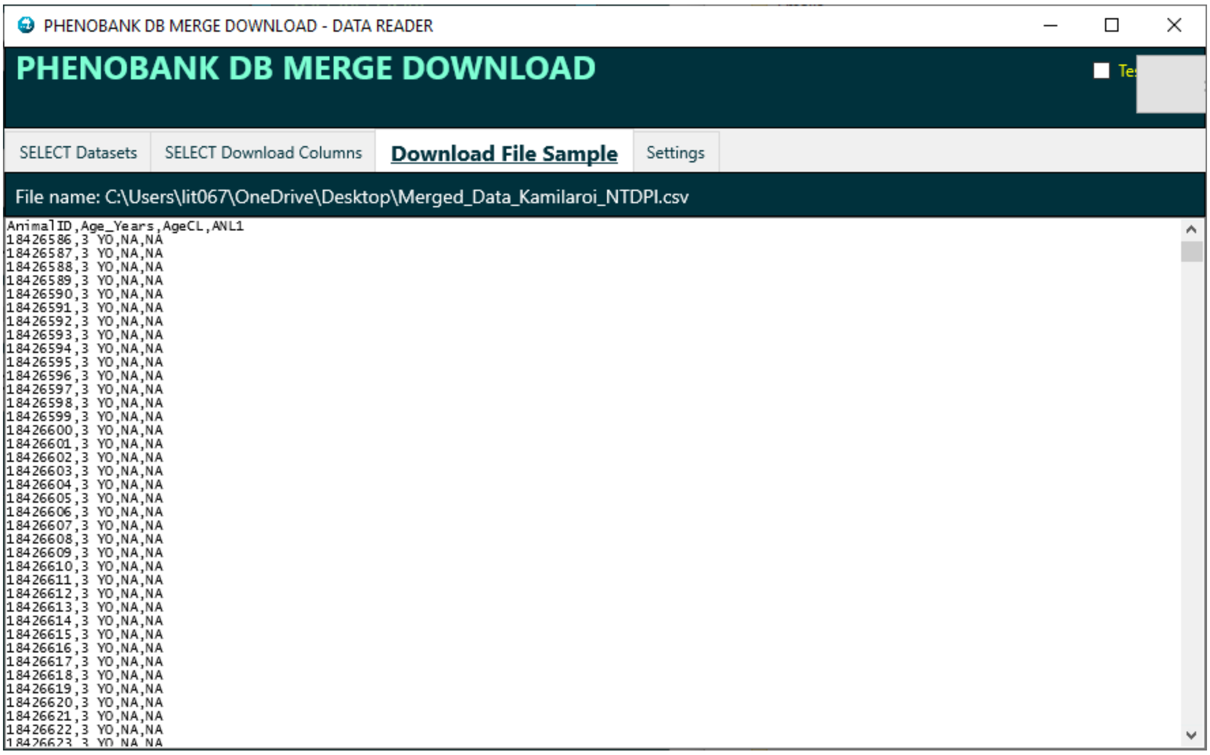
Click “Download” to download the data and create the merged file. A Windows box will appear requesting a location for the file and a file name, like the following:



File name: Merged_Data_Kamilaroi_NTDPI.csv

Save as type: CSV (*.csv)

Once downloaded, a sample of the file can be viewed at the “Download File Sample” tab. Note that data that is not available will be shown with the value “NA”.


PHENOBANK DB MERGE DOWNLOAD - DATA READER
PHENOBANK DB MERGE DOWNLOAD
Te
SELECT Datasets
SELECT Download Columns
Download File Sample
Settings

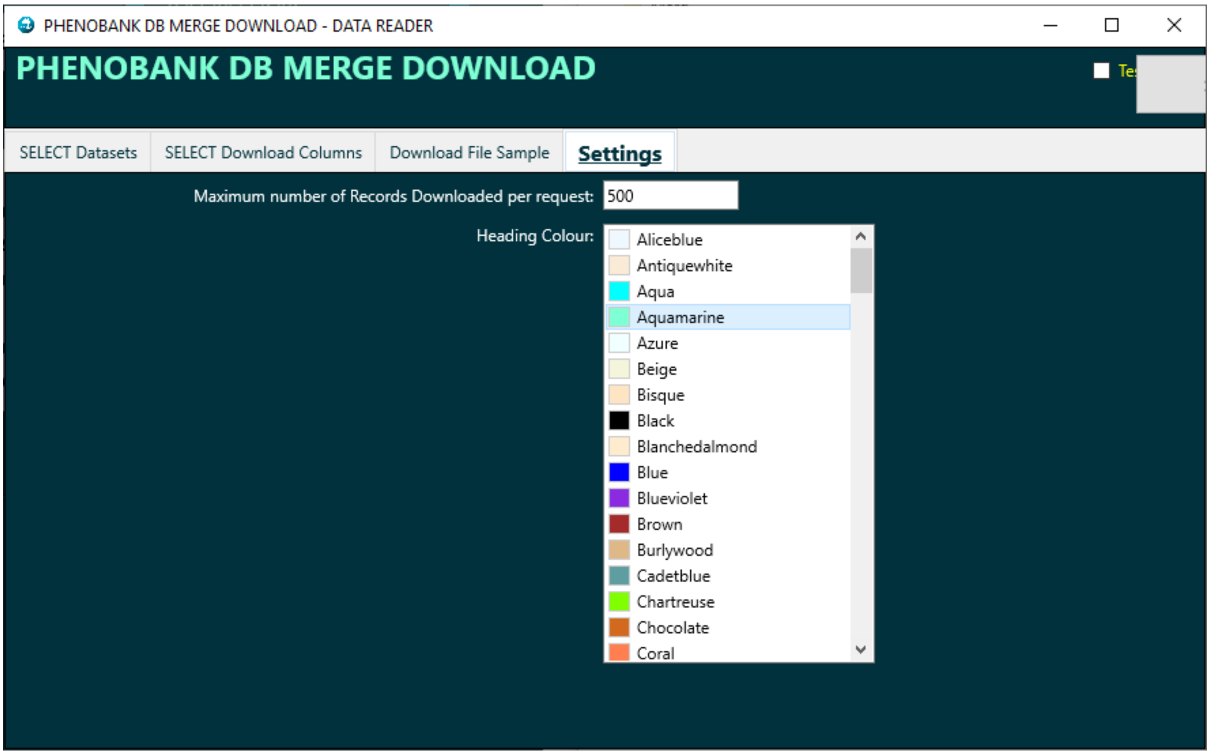
File name: C:\Users\lit067\OneDrive\Desktop\Merged_Data_Kamilaroi_NTDPI.csv

```
AnimalID,Age_Years,AgeCL,ANL1
18426586,3 YO,NA,NA
18426587,3 YO,NA,NA
18426588,3 YO,NA,NA
18426589,3 YO,NA,NA
18426590,3 YO,NA,NA
18426591,3 YO,NA,NA
18426592,3 YO,NA,NA
18426593,3 YO,NA,NA
18426594,3 YO,NA,NA
18426595,3 YO,NA,NA
18426596,3 YO,NA,NA
18426597,3 YO,NA,NA
18426598,3 YO,NA,NA
18426599,3 YO,NA,NA
18426600,3 YO,NA,NA
18426601,3 YO,NA,NA
18426602,3 YO,NA,NA
18426603,3 YO,NA,NA
18426604,3 YO,NA,NA
18426605,3 YO,NA,NA
18426606,3 YO,NA,NA
18426607,3 YO,NA,NA
18426608,3 YO,NA,NA
18426609,3 YO,NA,NA
18426610,3 YO,NA,NA
18426611,3 YO,NA,NA
18426612,3 YO,NA,NA
18426613,3 YO,NA,NA
18426614,3 YO,NA,NA
18426615,3 YO,NA,NA
18426616,3 YO,NA,NA
18426617,3 YO,NA,NA
18426618,3 YO,NA,NA
18426619,3 YO,NA,NA
18426620,3 YO,NA,NA
18426621,3 YO,NA,NA
18426622,3 YO,NA,NA
18426623,3 YO,NA,NA
```

6.4 Settings

Users may adjust the number of records downloaded in each “batch”. The default is 500, which we recommend is left unchanged.

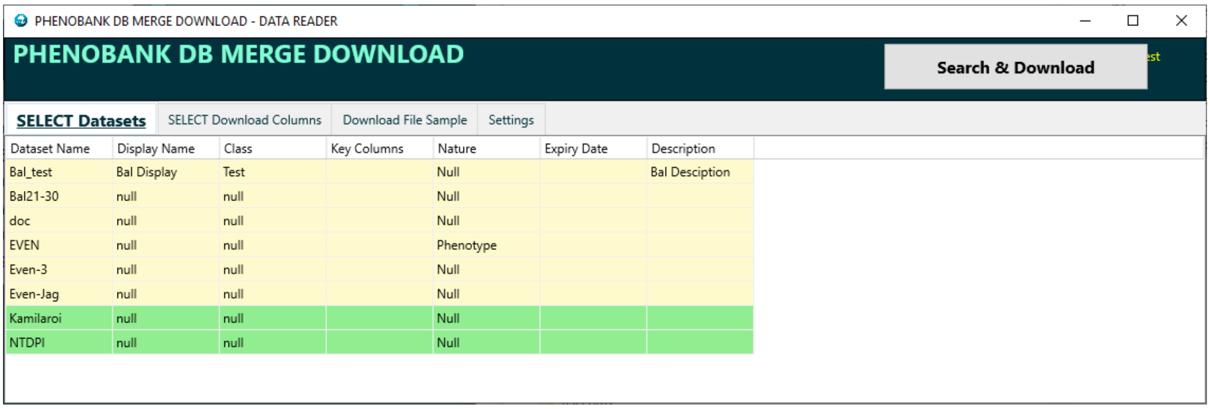
Users can also personalise their windows by selecting the colour of headings.



Maximum number of Records Downloaded per request: 500

7 Search Download

Users may access the Searh & Download function via the “Search & Download” button on the Download window.

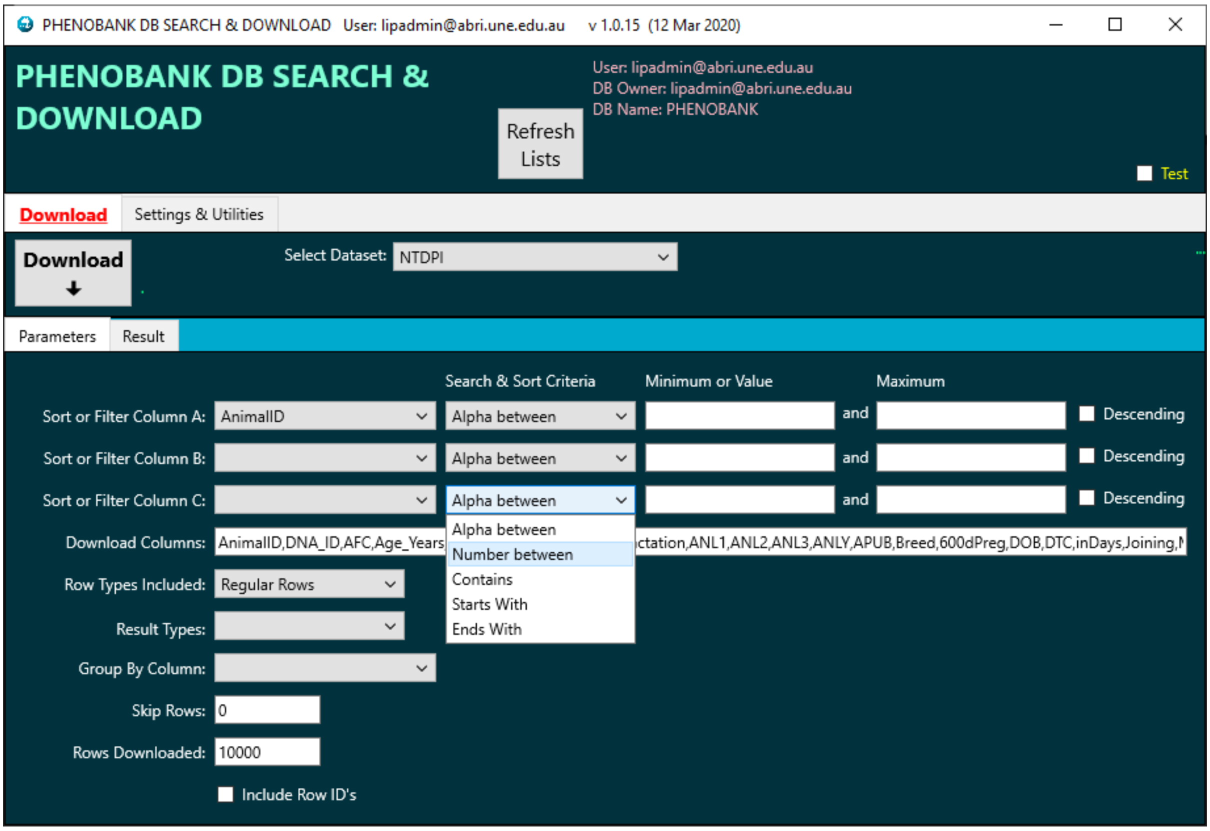


PHENOBANK DB MERGE DOWNLOAD - DATA READER

PHENOBANK DB MERGE DOWNLOAD

Dataset Name	Display Name	Class	Key Columns	Nature	Expiry Date	Description
Bal_test	Bal Display	Test		Null		Bal Desciption
Bal21-30	null	null		Null		
doc	null	null		Null		
EVEN	null	null		Phenotype		
Even-3	null	null		Null		
Even-Jag	null	null		Null		
Kamilaroi	null	null		Null		
NTDPI	null	null		Null		

The “SEARCH DOWNLOAD” function is used where a selected range of rows are required from a nominated **single** dataset. A window similar to the following will appear:



PHENOBANK DB SEARCH & DOWNLOAD User: lipadmin@abri.une.edu.au v 1.0.15 (12 Mar 2020)

PHENOBANK DB SEARCH & DOWNLOAD

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: NTDPI

Sort or Filter Column A: AnimalID | Alpha between | and | □ Descending

Sort or Filter Column B: | Alpha between | and | □ Descending

Sort or Filter Column C: | Alpha between | and | □ Descending

Download Columns: AnimalID,DNA_ID,AFC,Age_Years ctation,ANL1,ANL2,ANL3,ANLY,APUB,Breed,600dPreg,DOB,DTC,inDays,Joining,

Row Types Included: Regular Rows

Result Types:

Group By Column:

Skip Rows: 0

Rows Downloaded: 10000

□ Include Row ID's

7.1 Parameters

Several parameters can be selected to allow you to nominate to limit the range of data. These are available on the Parameters tab, and include the following:

7.1.1.8 Select Dataset

Click the drop down to select the dataset with the source of the information of interest

7.1.1.9 Filter Columns

Up to three columns can be selected for filtering data to rows of interest. For each column, a Minimum and Maximum can be given. The data is sorted prior to filtering based upon criteria as follows:

- Number between (e.g. 10 is greater than 2)
- List between (e.g. 2 is after 10, because 2 is after 1, “alphabetically”)
- Contains (e.g. “fred” will include “Frederick” and “Alfred”)
- Starts With
- Ends With

NOTE that there must be at least one Filter Column selected, even if there is no minimum and maximum range supplied.

Finally tick the “Descending” box if you wish the result to be sorted in reverse order.

7.1.1.10 Group By Columns

This option is relevant only if you select the “Summary” Result Type. In that case it will summarise data rows by each value of the field you enter here.

7.1.1.11 Download Columns

List the columns that you wish to download, delimited by commas. If this is left blank, all columns will be downloaded.

7.1.1.12 Include Row IDs (this may not be available to all users)

Phenobank DB’s underlying database technology creates a unique identifier for each row, called the RowID. Some users may wish to download this, and this is possible if this box is ticked.

7.1.1.13 Row Types Included (this may not be available to all users)

Rows are either “Regular” or “Deleted”. This drop-down option allows previously deleted rows to be retrieved.

7.1.1.14 Result Types

The results retrieved may be in several forms. These are:

- Data Only.
- Summary (mean, sum, count)
- Extended Summary (Summary + standard deviation)

7.1.1.15 Skip Rows

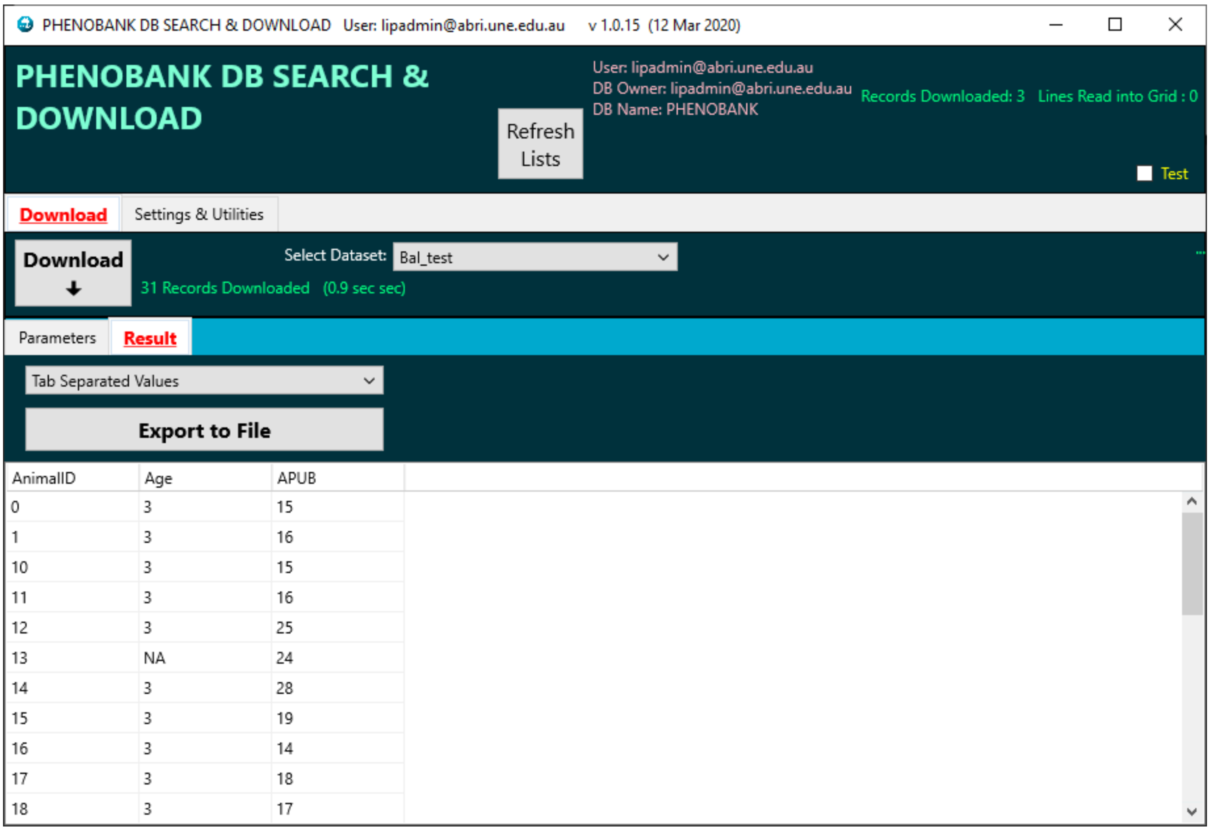
The user supplied number of rows are skipped before data from subsequent rows are output.

7.1.1.16 Rows Downloaded

This is the maximum number of Rows downloaded, rounded up to the nearest “page” (typically 500 rows, but may be adjusted by user on the settings page).

7.2 Result

Click the “Download” button to retrieve the data which, once downloaded, will appear in the table on the Result tab.



PHENOBANK DB SEARCH & DOWNLOAD User: lipadmin@abri.une.edu.au v 1.0.15 (12 Mar 2020)

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Records Downloaded: 3 Lines Read into Grid : 0

Select Dataset: Bal_test

31 Records Downloaded (0.9 sec sec)

Tab Separated Values

AnimalID	Age	APUB
0	3	15
1	3	16
10	3	15
11	3	16
12	3	25
13	NA	24
14	3	28
15	3	19
16	3	14
17	3	18
18	3	17

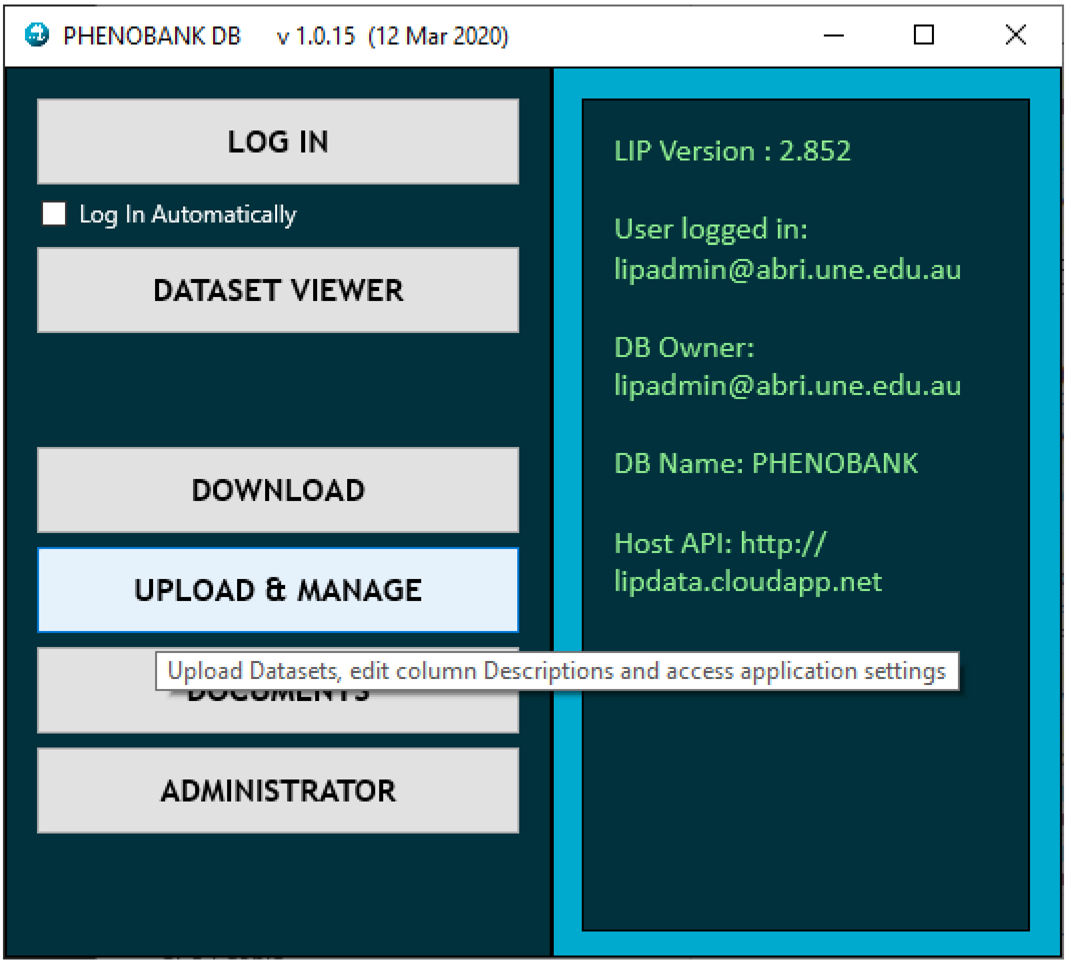
This table can be exported to a file by clicking the “Export to File” button. There are a choice of export formats available. These are:

- CSV (UTF-8, what used to be called ASCII)
- Tab delimited (UTF-8)
- Space separated values (UTF-8)
- Tab delimited, Unicode Text (UTF-16, including international character sets)

8 Upload & Manage

This function allows users to add new data to the Phenobank DB, as well as edit or correct data. Users who upload a dataset "own" the dataset, and therefore can control who has access to the dataset.

This ability can be accessed by clicking the DOCUMENTS button on the main window.



□ Log In Automatically

Upload Datasets, edit column Descriptions and access application settings

LIP Version : 2.852

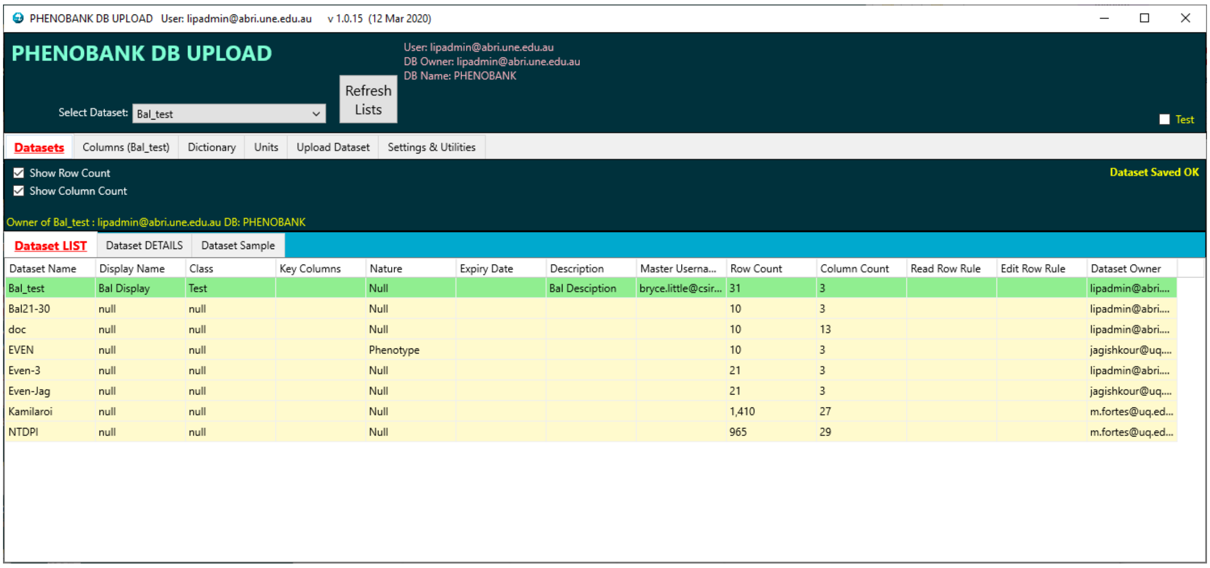
User logged in: lipadmin@abri.une.edu.au

DB Owner: lipadmin@abri.une.edu.au

DB Name: PHENOBANK

Host API: http:// lipdata.cloudapp.net

Upon clicking, you will be rewarded with the following:



User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: Bal_test

☐ Test

☑ Show Row Count
☑ Show Column Count

Owner of Bal_test : lipadmin@abri.une.edu.au DB: PHENOBANK

Dataset Name	Display Name	Class	Key Columns	Nature	Expiry Date	Description	Master Userna...	Row Count	Column Count	Read Row Rule	Edit Row Rule	Dataset Owner
Bal_test	Bal Display	Test		Null		Bal Desciption	bryce.little@csir...	31	3			lipadmin@abri....
Bal21-30	null	null		Null				10	3			lipadmin@abri....
doc	null	null		Null				10	13			lipadmin@abri....
EVEN	null	null		Phenotype				10	3			jagishkour@uq....
Even-3	null	null		Null				21	3			lipadmin@abri....
Even-Jag	null	null		Null				21	3			jagishkour@uq....
Kamilaroi	null	null		Null				1,410	27			m.fortes@uq.ed...
NTDPI	null	null		Null				965	29			m.fortes@uq.ed...

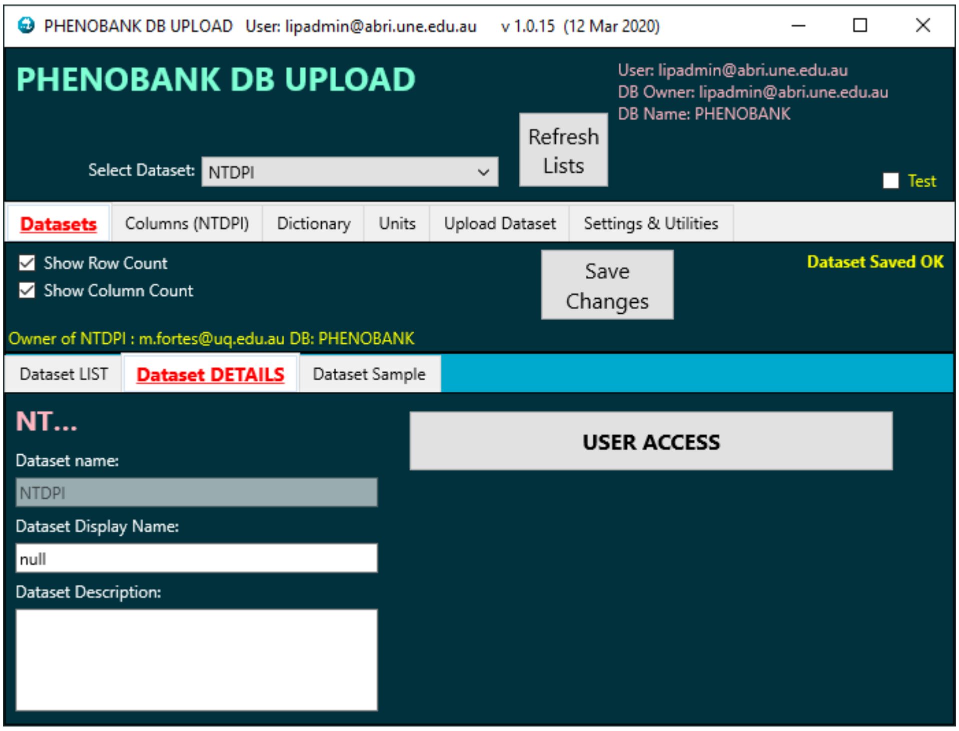
The “Upload” function uses the same functional template as the “Dataset Viewer” but with additional functionality.   This additional functionality is given in this chapter, however refer to chapter 5 for basic functionality.

8.1 Defining User Access

Access to individual users for EACH Dataset is available via the Dataset Details Tab. For any given dataset you may be able to determine which users have access to the Dataset. This is only possible if you are:

- The creator of the dataset or,
- The creator of the database that houses the dataset.

If this capability is available for the Dataset in question, you will see a “User Access” button on the Dataset Details tab, as here:


PHENOBANK DB UPLOAD User: lipadmin@abri.une.edu.au v 1.0.15 (12 Mar 2020)

PHENOBANK DB UPLOAD

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: NTDPI

□ Test

☑ Show Row Count
☑ Show Column Count

Dataset Saved OK

Owner of NTDPI : m.fortes@uq.edu.au DB: PHENOBANK

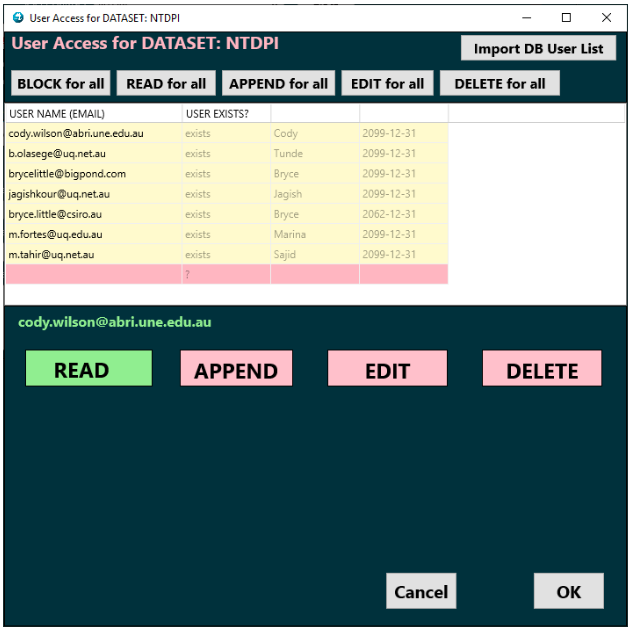
NT...

Dataset name: NTDPI

Dataset Display Name: null

Dataset Description:

You will be rewarded with a window that will be initially without any users listed if the dataset has been recently created, or it may have a list of users like this:

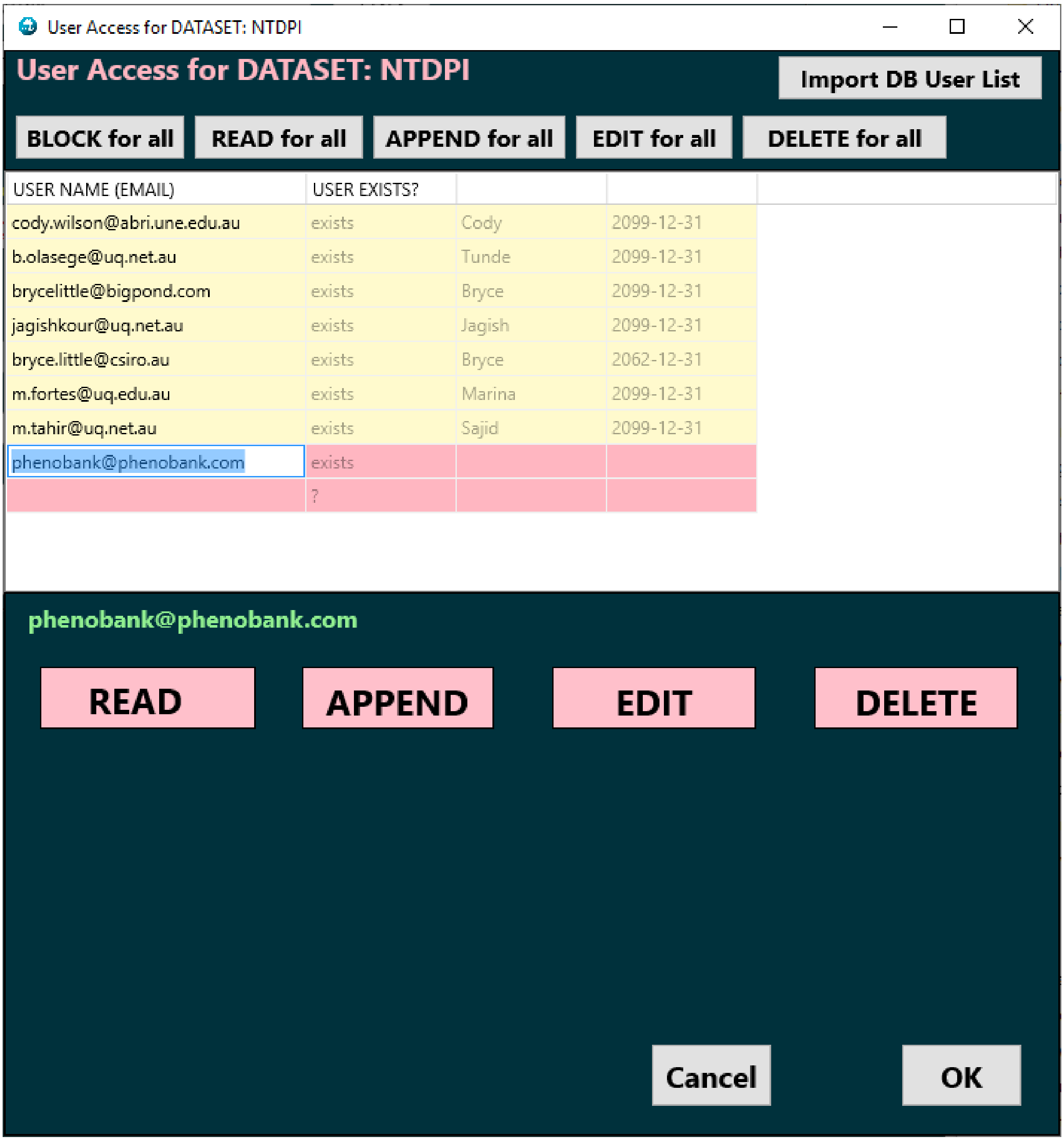

User Access for DATASET: NTDPI

User Access for DATASET: NTDPI

USER NAME (EMAIL)	USER EXISTS?		
cody.wilson@abri.une.edu.au	exists	Cody	2099-12-31
b.olasege@uq.net.au	exists	Tunde	2099-12-31
brycelittle@bigpond.com	exists	Bryce	2099-12-31
jagishkour@uq.net.au	exists	Jagish	2099-12-31
bryce.little@csiro.au	exists	Bryce	2062-12-31
m.fortes@uq.edu.au	exists	Marina	2099-12-31
m.tahir@uq.net.au	exists	Sajid	2099-12-31
	?		

cody.wilson@abri.une.edu.au

To allow a new user to access the dataset, enter the user name at the next blank fields in the “USER NAME” column and press “Enter”. If the user exists in LIP this will be noted in the “USER EXISTS?” column, as in this case:



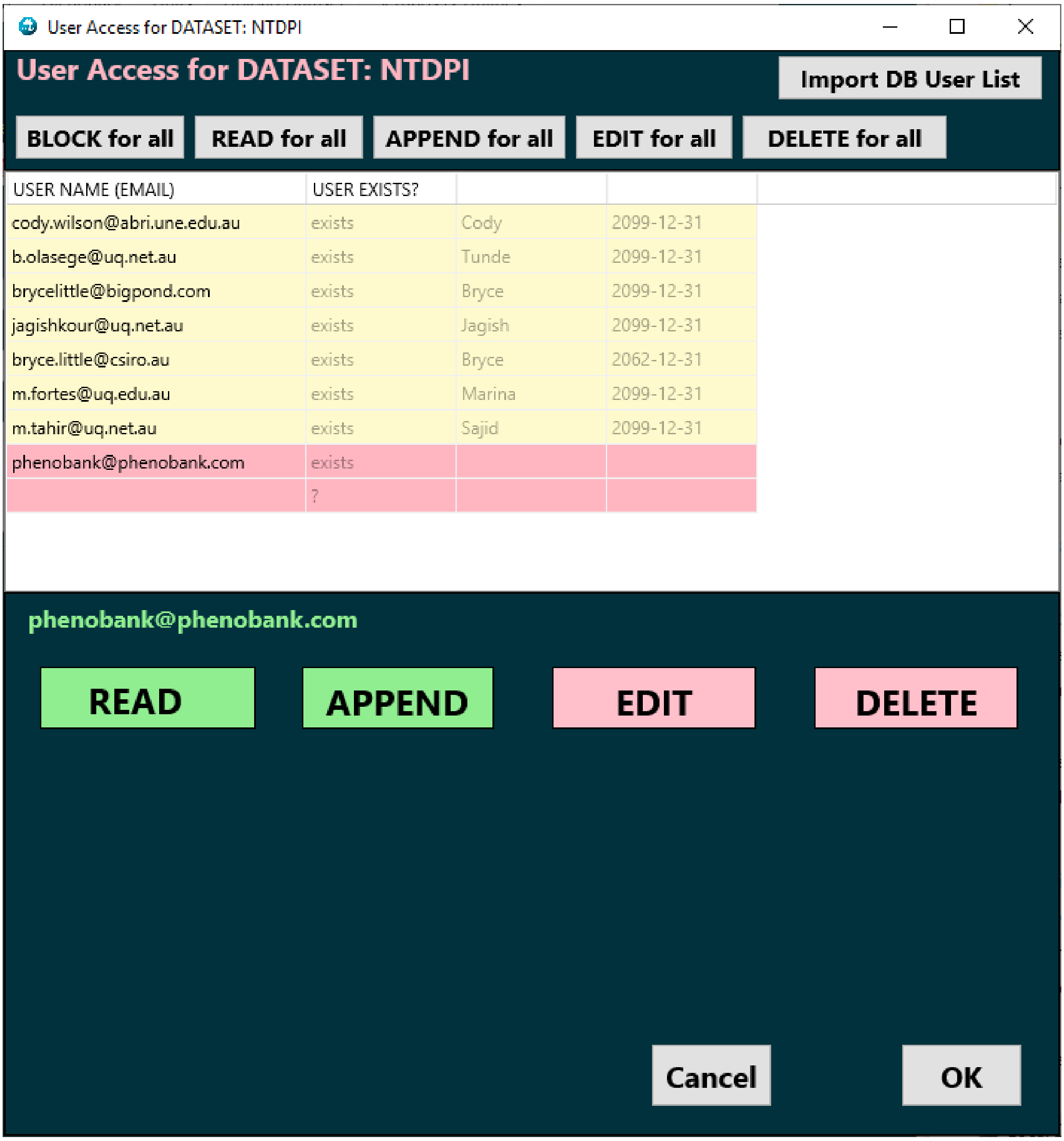
User Access for DATASET: NTDPI

USER NAME (EMAIL)	USER EXISTS?		
cody.wilson@abri.une.edu.au	exists	Cody	2099-12-31
b.olasege@uq.net.au	exists	Tunde	2099-12-31
brycelittle@bigpond.com	exists	Bryce	2099-12-31
jagishkour@uq.net.au	exists	Jagish	2099-12-31
bryce.little@csiro.au	exists	Bryce	2062-12-31
m.fortes@uq.edu.au	exists	Marina	2099-12-31
m.tahir@uq.net.au	exists	Sajid	2099-12-31
phenobank@phenobank.com	exists		
	?		

phenobank@phenobank.com

Next you may grant access by simply clicking on the rectangles labelled “READ”, “APPEND”, “EDIT” or “DELETE” depending upon the level of access and control that you wish to give to the user.

In the following example READ and APPEND permission are being granted. Click OK to Save.



USER NAME (EMAIL)	USER EXISTS?		
cody.wilson@abri.une.edu.au	exists	Cody	2099-12-31
b.olasege@uq.net.au	exists	Tunde	2099-12-31
brycelittle@bigpond.com	exists	Bryce	2099-12-31
jagishkour@uq.net.au	exists	Jagish	2099-12-31
bryce.little@csiro.au	exists	Bryce	2062-12-31
m.fortes@uq.edu.au	exists	Marina	2099-12-31
m.tahir@uq.net.au	exists	Sajid	2099-12-31
phenobank@phenobank.com	exists		
	?		

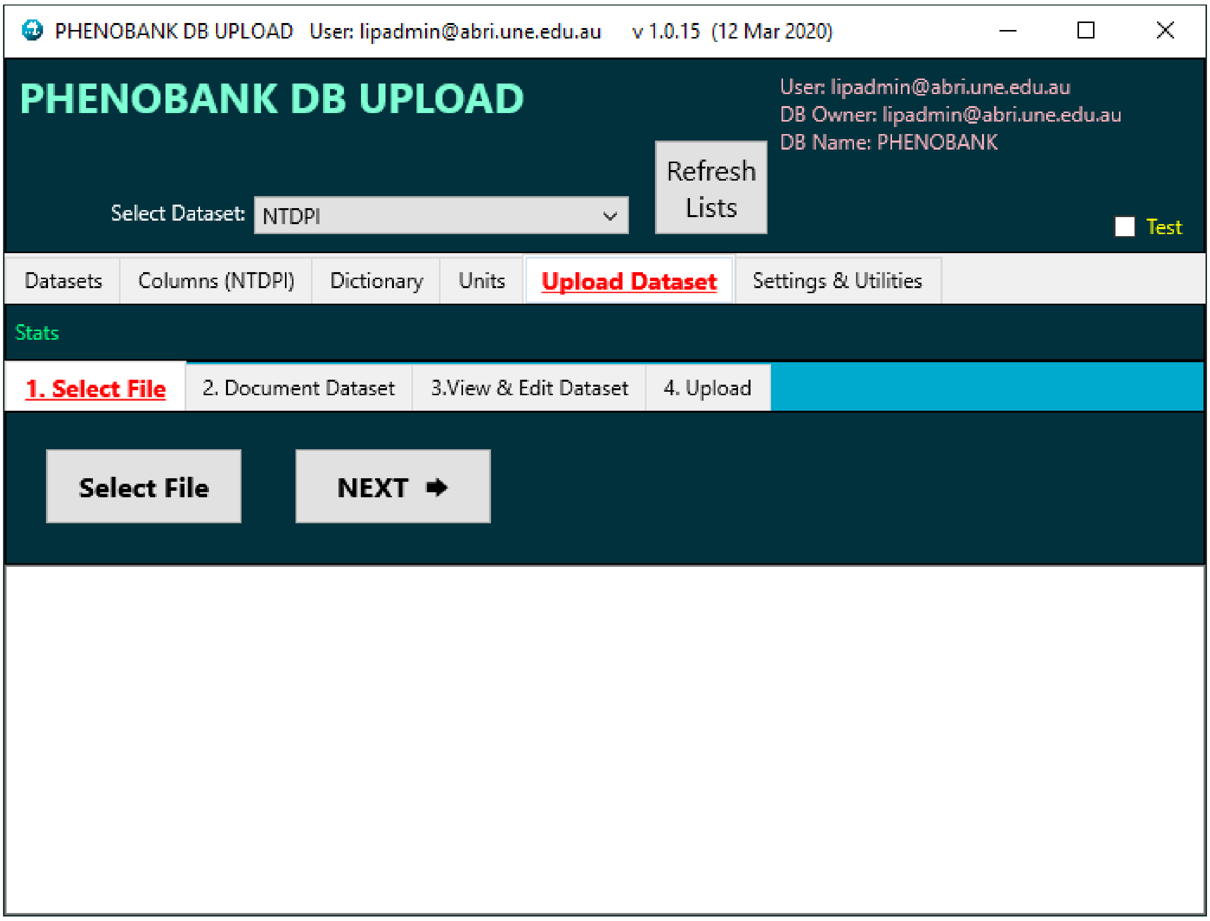
phenobank@phenobank.com

8.2 Uploading a Dataset

Datasets can either be:

- Uploaded and stored as NEW datasets or,
- Uploaded and merged with EXISTING datasets.

To undertake any of these functions select the “Upload Dataset” tab, like the following:



PHENOBANK DB UPLOAD User: lipadmin@abri.une.edu.au v 1.0.15 (12 Mar 2020)

PHENOBANK DB UPLOAD

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Dataset: NTDPI

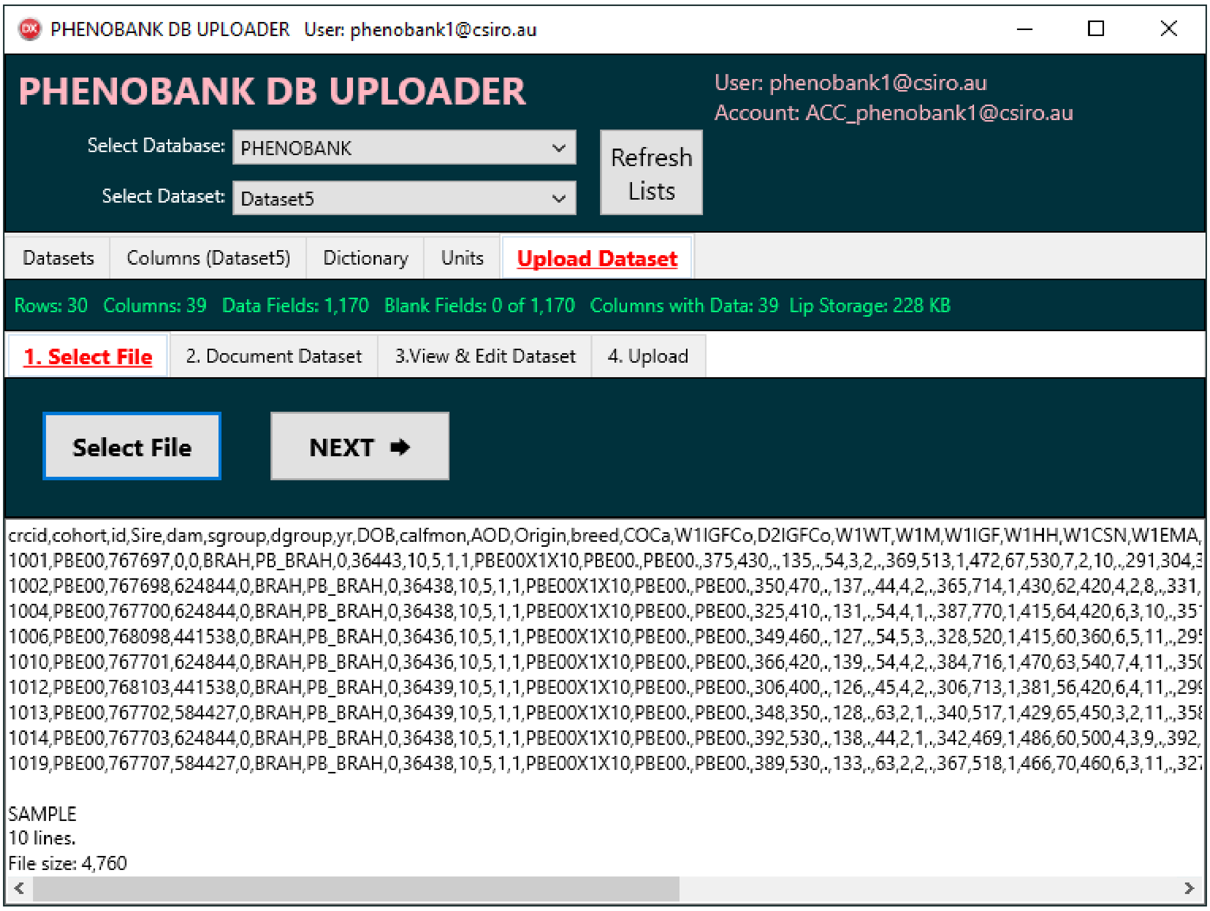
□ Test

8.2.1 Uploading a NEW Dataset

At time of publication the only types of files that can be uploaded with this Windows tool are CSV “Phenotype” files. “Genotype” files and various non-CSV files will be compatible with future versions of PHENOBANK DB.

Follow these steps:

- Click the “Select File” button on the “Select File” tab to locate the file on your computer that you wish to upload.



User: phenobank1@csiro.au
Account: ACC_phenobank1@csiro.au

Select Database: PHENOBANK
Select Dataset: Dataset5

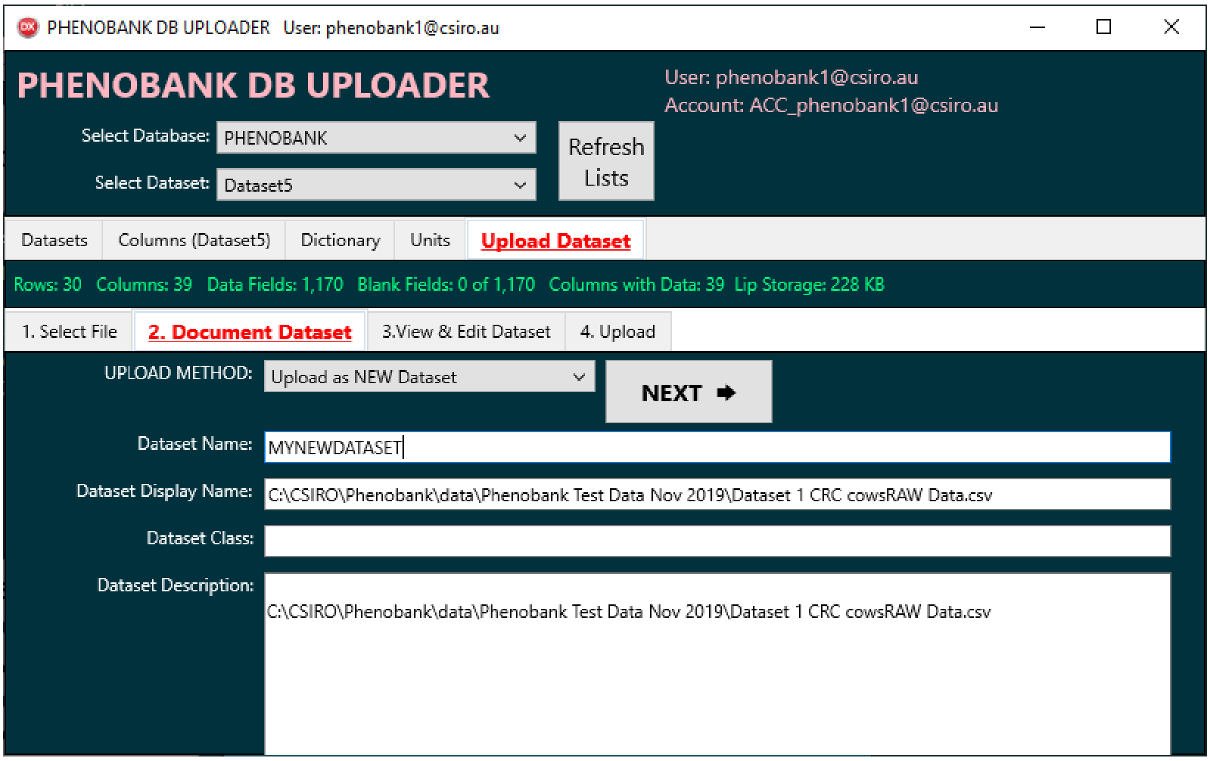
Rows: 30   Columns: 39   Data Fields: 1,170   Blank Fields: 0 of 1,170   Columns with Data: 39   Lip Storage: 228 KB

```
crcid,cohort,id,Sire,dam,sgroup,dgroup,yr,DOB,calfmon,AOD,Origin,breed,COCa,W1IGFCo,D2IGFCo,W1WT,W1M,W1IGF,W1HH,W1CSN,W1EMA,
1001,PBE00,767697,0,0,BRAH,PB_BRAH,0,36443,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,375,430,.,135,.,54,3,2,.,369,513,1,472,67,530,7,2,10,.,291,304,3
1002,PBE00,767698,624844,0,BRAH,PB_BRAH,0,36438,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,350,470,.,137,.,44,4,2,.,365,714,1,430,62,420,4,2,8,.,331,
1004,PBE00,767700,624844,0,BRAH,PB_BRAH,0,36438,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,325,410,.,131,.,54,4,1,.,387,770,1,415,64,420,6,3,10,.,35
1006,PBE00,768098,441538,0,BRAH,PB_BRAH,0,36436,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,349,460,.,127,.,54,5,3,.,328,520,1,415,60,360,6,5,11,.,29
1010,PBE00,767701,624844,0,BRAH,PB_BRAH,0,36436,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,366,420,.,139,.,54,4,2,.,384,716,1,470,63,540,7,4,11,.,35
1012,PBE00,768103,441538,0,BRAH,PB_BRAH,0,36439,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,306,400,.,126,.,45,4,2,.,306,713,1,381,56,420,6,4,11,.,29
1013,PBE00,767702,584427,0,BRAH,PB_BRAH,0,36439,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,348,350,.,128,.,63,2,1,.,340,517,1,429,65,450,3,2,11,.,35
1014,PBE00,767703,624844,0,BRAH,PB_BRAH,0,36438,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,392,530,.,138,.,44,2,1,.,342,469,1,486,60,500,4,3,9,.,392,
1019,PBE00,767707,584427,0,BRAH,PB_BRAH,0,36438,10,5,1,1,PBE00X1X10,PBE00.,PBE00.,389,530,.,133,.,63,2,2,.,367,518,1,466,70,460,6,3,11,.,32

SAMPLE
10 lines.
File size: 4,760
```

You will be rewarded with a sample 10 lines from the file in the bottom pane.

- Click "NEXT" enter the upload criteria:



User: phenobank1@csiro.au
Account: ACC_phenobank1@csiro.au

Select Database: PHENOBANK
Select Dataset: Dataset5

Rows: 30   Columns: 39   Data Fields: 1,170   Blank Fields: 0 of 1,170   Columns with Data: 39   Lip Storage: 228 KB

UPLOAD METHOD: Upload as NEW Dataset

Dataset Name: MYNEWDATASET

Dataset Display Name: C:\CSIRO\Phenobank\data\Phenobank Test Data Nov 2019\Dataset 1 CRC cowsRAW Data.csv

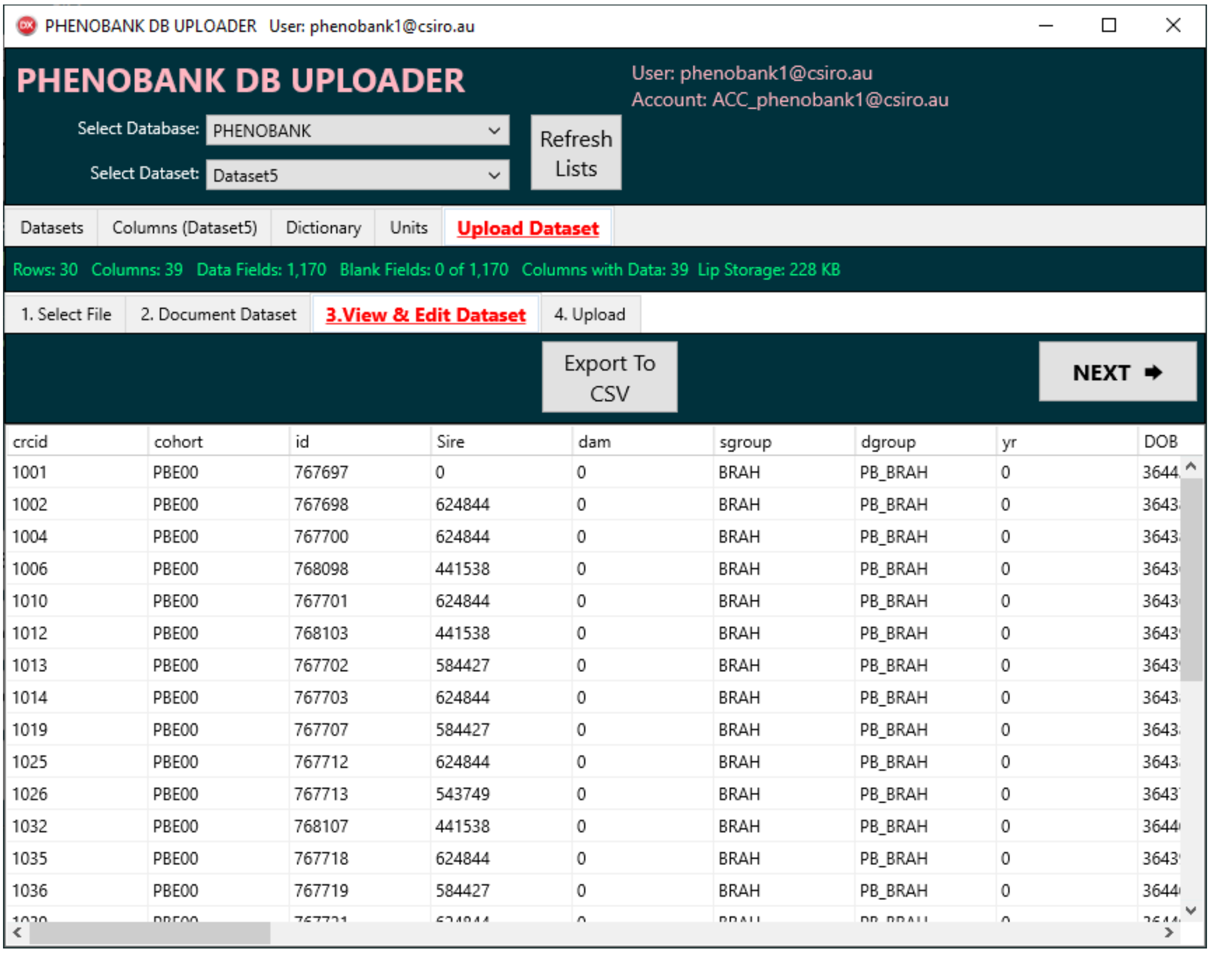
Dataset Class:

Dataset Description: C:\CSIRO\Phenobank\data\Phenobank Test Data Nov 2019\Dataset 1 CRC cowsRAW Data.csv

- Select "Upload as NEW Dataset" for the UPLOAD METHOD.

- Enter a Dataset Name if you wish to give it a different name to that of the orginal file name.
- Enter Display name, Class and Description as user wishes. (Class is a user defined field that can be used to group Datasets)
- Click “NEXT”

At this stage users can View (and Edit) the Dataset before the final step of Uploading to the Phenobank Database. When done, click “Next”



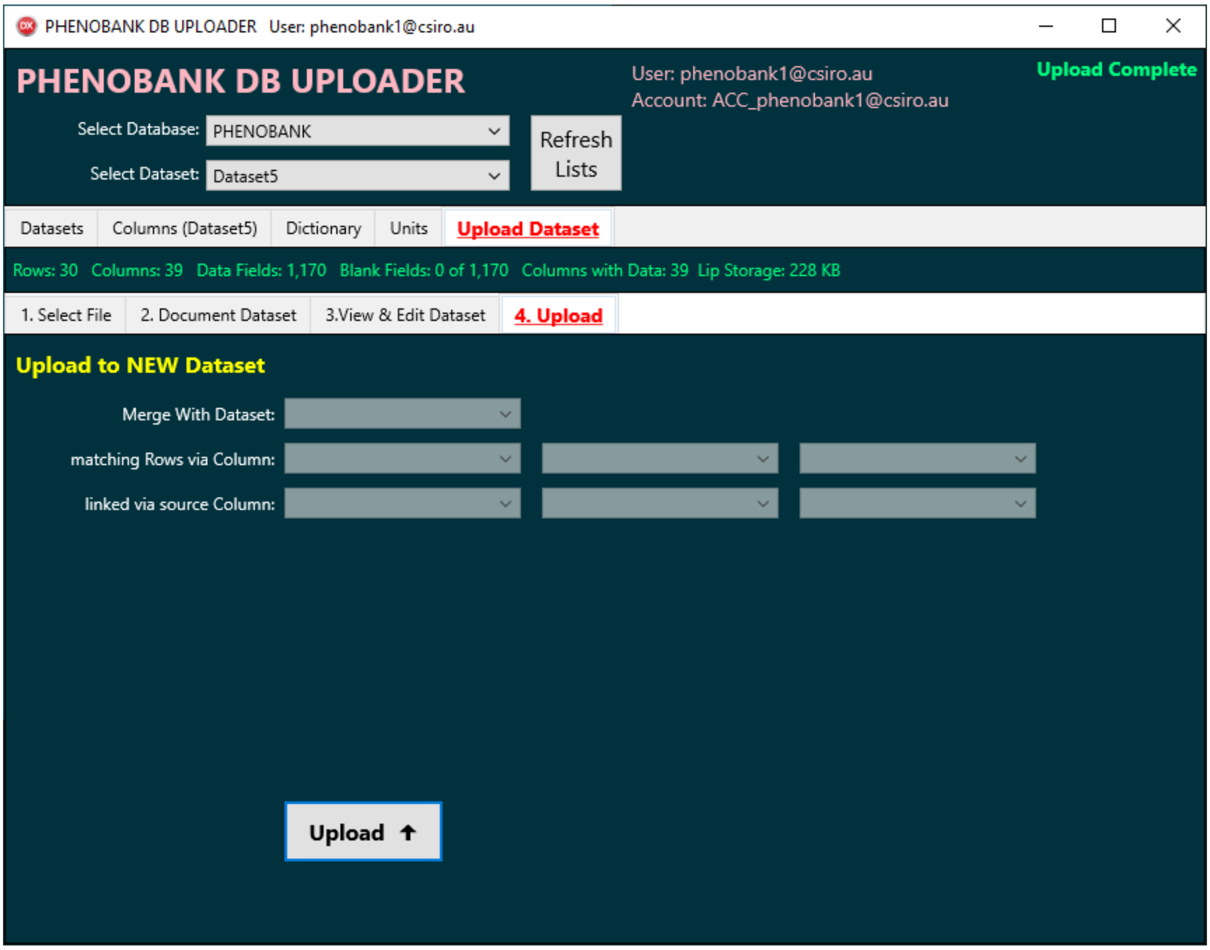
Select Database: PHENOBANK

Select Dataset: Dataset5

Rows: 30   Columns: 39   Data Fields: 1,170   Blank Fields: 0 of 1,170   Columns with Data: 39   Lip Storage: 228 KB

crcid	cohort	id	Sire	dam	sgroup	dgroup	yr	DOB
1001	PBE00	767697	0	0	BRAH	PB_BRAH	0	3644
1002	PBE00	767698	624844	0	BRAH	PB_BRAH	0	3643
1004	PBE00	767700	624844	0	BRAH	PB_BRAH	0	3643
1006	PBE00	768098	441538	0	BRAH	PB_BRAH	0	3643
1010	PBE00	767701	624844	0	BRAH	PB_BRAH	0	3643
1012	PBE00	768103	441538	0	BRAH	PB_BRAH	0	3643
1013	PBE00	767702	584427	0	BRAH	PB_BRAH	0	3643
1014	PBE00	767703	624844	0	BRAH	PB_BRAH	0	3643
1019	PBE00	767707	584427	0	BRAH	PB_BRAH	0	3643
1025	PBE00	767712	624844	0	BRAH	PB_BRAH	0	3643
1026	PBE00	767713	543749	0	BRAH	PB_BRAH	0	3643
1032	PBE00	768107	441538	0	BRAH	PB_BRAH	0	3644
1035	PBE00	767718	624844	0	BRAH	PB_BRAH	0	3643
1036	PBE00	767719	584427	0	BRAH	PB_BRAH	0	3644
[illegible]	PBE00	767721	624844	0	BRAH	PB_BRAH	0	3644

The final step is to simply click “Upload” to store in the Cloud. A message in the top right corner will show progress, and finish with an “Upload Complete” as in the following example:



Select Database: PHENOBANK

Select Dataset: Dataset5

Rows: 30   Columns: 39   Data Fields: 1,170   Blank Fields: 0 of 1,170   Columns with Data: 39   Lip Storage: 228 KB

Upload to NEW Dataset

Merge With Dataset:

matching Rows via Column:

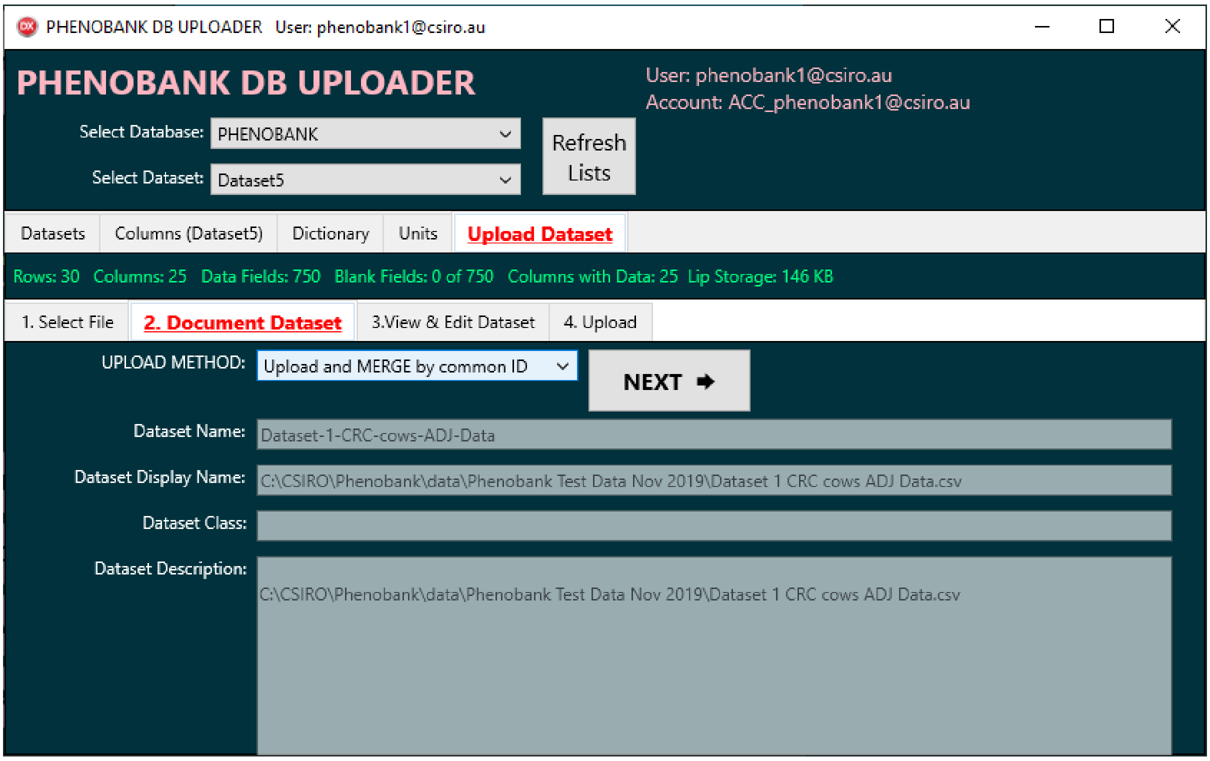
linked via source Column:

8.2.2 Uploading and merging with an EXISTING dataset

This capability allows users to add additional COLUMNS to a dataset using existing rows. For example, additional animal phenotypes might be added to existing animal rows.

Uploading and Merging a user's dataset with an existing Dataset follows a similar process to the New Dataset process described previously. The steps are as follows:

- Select the file as described in the previous section.
- At the "Document Dataset" step select the UPLOAD METHOD: "Upload and MERGE by common ID", as can be seen here:



User: phenobank1@csiro.au
Account: ACC_phenobank1@csiro.au

Select Database: PHENOBANK

Select Dataset: Dataset5

Rows: 30 Columns: 25 Data Fields: 750 Blank Fields: 0 of 750 Columns with Data: 25 Lip Storage: 146 KB

UPLOAD METHOD: Upload and MERGE by common ID

Dataset Name: Dataset-1-CRC-cows-ADJ-Data

Dataset Display Name: C:\CSIRO\Phenobank\data\Phenobank Test Data Nov 2019\Dataset 1 CRC cows ADJ Data.csv

Dataset Class:

Dataset Description: C:\CSIRO\Phenobank\data\Phenobank Test Data Nov 2019\Dataset 1 CRC cows ADJ Data.csv

Note that there is no need to enter a Dataset Name or Description since that has already been entered when the original Dataset was created.

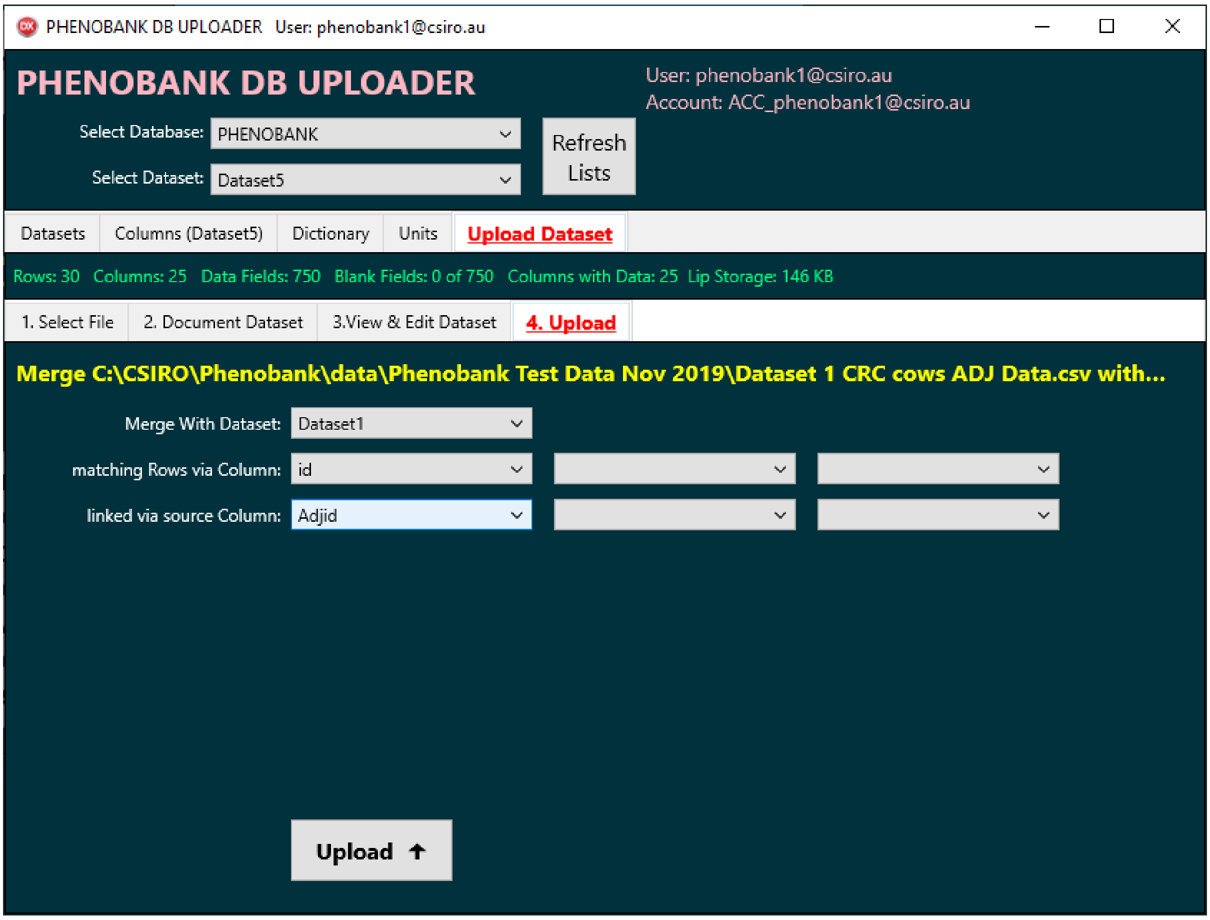
- Click NEXT to get to the Upload tab or click the Upload tab directly.
- At the UPLOAD Tab select the Merge Dataset and select the Columns that will be used to match the rows in the new Dataset with rows in the existing Dataset.

In the following example, the Uploaded Dataset will be merged with existing dataset named “Dataset1”. The column “id” in Dataset1 has been selected from the list to be the column for matching the rows from the new (or “Source”) dataset. The column in the Source dataset in this case has been selected as “Adjid”.

For each ROW if a match is SUCCESSFUL, the new column data will be added to the row, and columns with the same name will be REPLACED with data in the new source dataset.

For each ROW if a match is UNSUCCESSFUL, the ROW will be stored as an entirely NEW row in the existing dataset.

(NOTE that further options are to be added, as per feedback from users).



PHENOBANK DB UPLOADER User: phenobank1@csiro.au

PHENOBANK DB UPLOADER

User: phenobank1@csiro.au
Account: ACC_phenobank1@csiro.au

Select Database: PHENOBANK

Select Dataset: Dataset5

Rows: 30 Columns: 25 Data Fields: 750 Blank Fields: 0 of 750 Columns with Data: 25 Lip Storage: 146 KB

Merge C:\CSIRO\Phenobank\data\Phenobank Test Data Nov 2019\Dataset 1 CRC cows ADJ Data.csv with...

Merge With Dataset: Dataset1

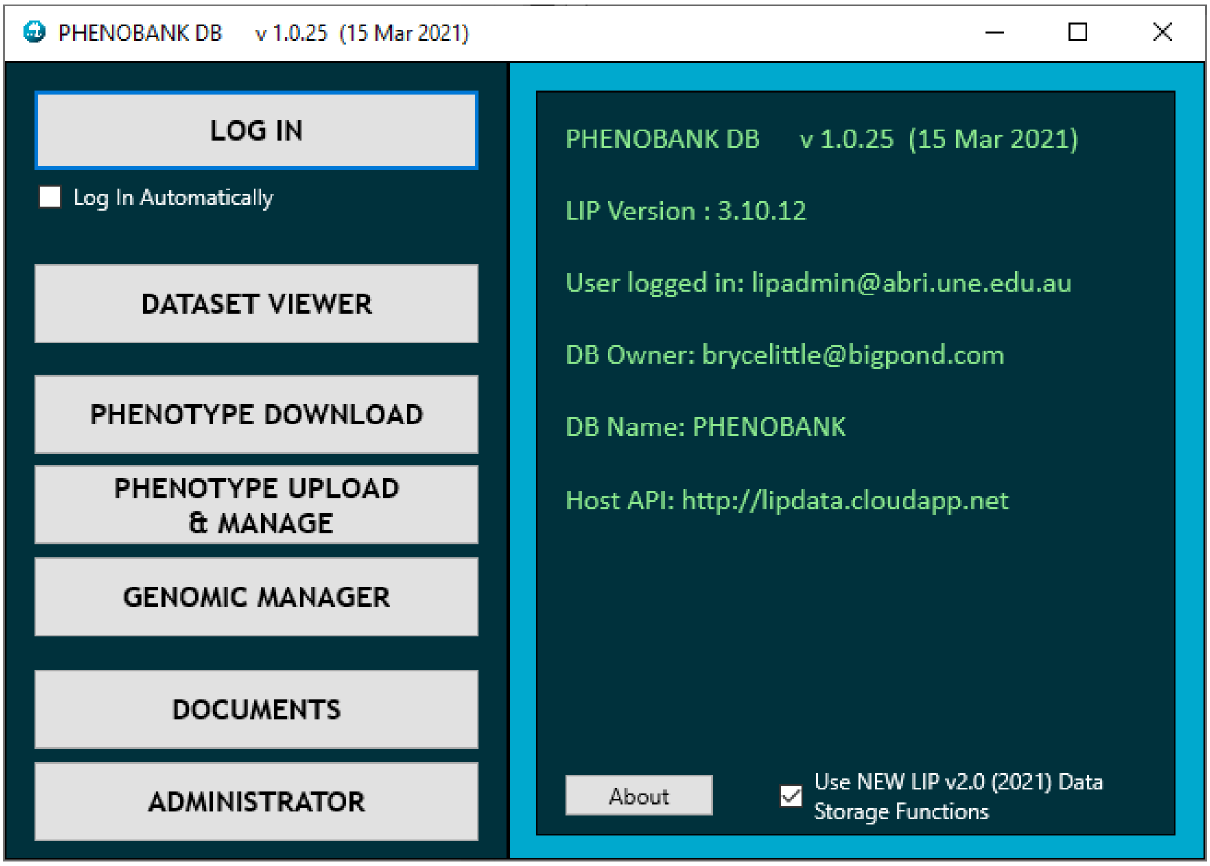
matching Rows via Column: id

linked via source Column: Adjid

9 Documents

9.1 Uploading Documents

Phenobank DB will allow you upload any file into the Phenobank DB document storage system. This ability can be accessed by clicking the DOCUMENTS button on the main window.



PHENOBANK DB   v 1.0.25 (15 Mar 2021)

LIP Version : 3.10.12

User logged in: lipadmin@abri.une.edu.au

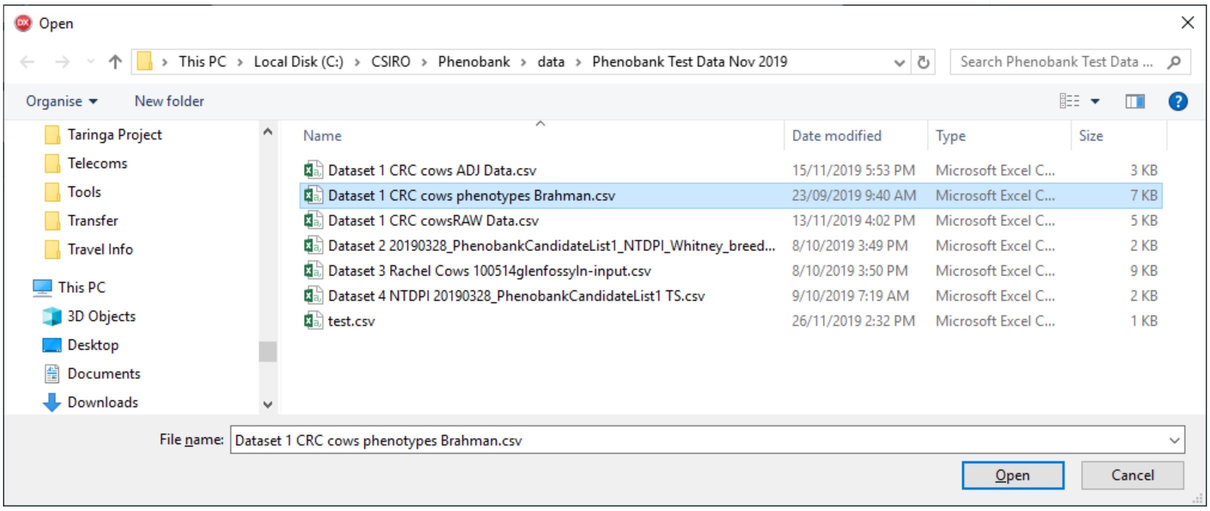
DB Owner: brycelittle@bigpond.com

DB Name: PHENOBANK

Host API: http://lipdata.cloudapp.net

☑ Use NEW LIP v2.0 (2021) Data Storage Functions

To select the file, click the button “Fetch Document, File or Picture”. A Windows File Explorer window will appear for you to select the file that you wish to upload, like the following:



Name	Date modified	Type	Size
Dataset 1 CRC cows ADJ Data.csv	15/11/2019 5:53 PM	Microsoft Excel C...	3 KB
Dataset 1 CRC cows phenotypes Brahman.csv	23/09/2019 9:40 AM	Microsoft Excel C...	7 KB
Dataset 1 CRC cowsRAW Data.csv	13/11/2019 4:02 PM	Microsoft Excel C...	5 KB
Dataset 2 20190328_PhenobankCandidateList1_NTDPI_Whitney_breed...	8/10/2019 3:49 PM	Microsoft Excel C...	2 KB
Dataset 3 Rachel Cows 100514glenfossyln-input.csv	8/10/2019 3:50 PM	Microsoft Excel C...	9 KB
Dataset 4 NTDPI 20190328_PhenobankCandidateList1 TS.csv	9/10/2019 7:19 AM	Microsoft Excel C...	2 KB
test.csv	26/11/2019 2:32 PM	Microsoft Excel C...	1 KB

File name: Dataset 1 CRC cows phenotypes Brahman.csv

Click on the file to be uploaded and click “Open”.

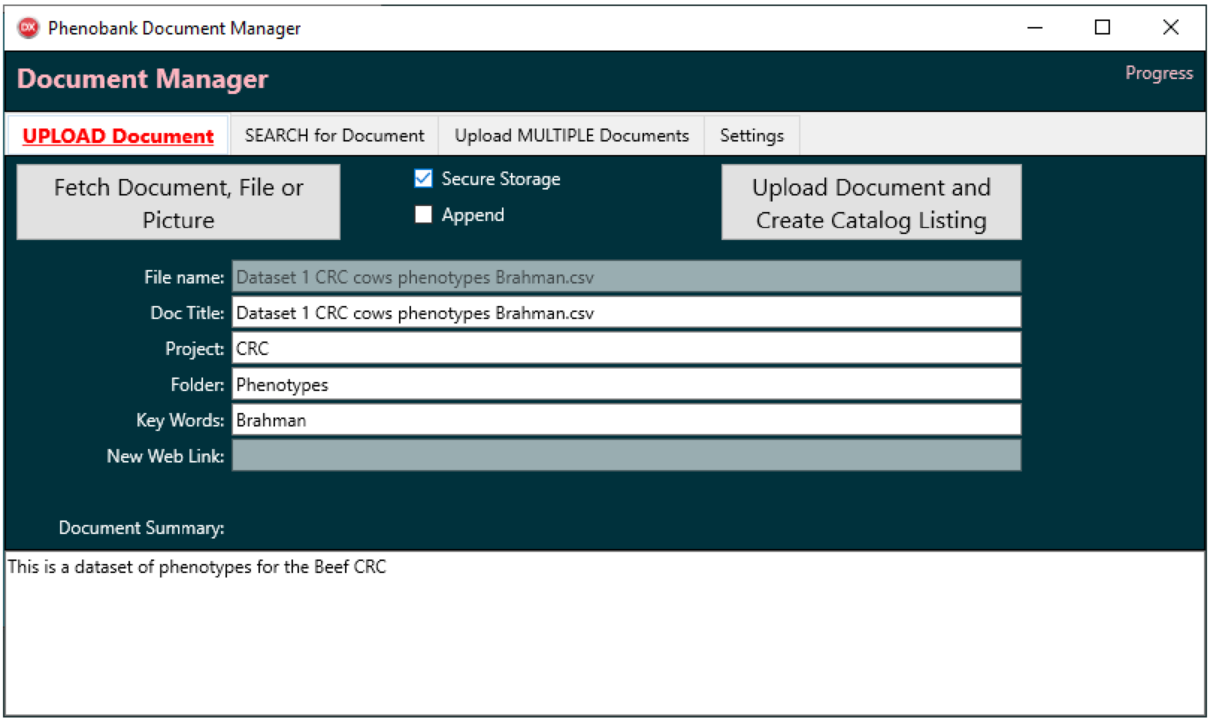
Now at the UPLOAD Document tab, enter additional information about the file including:

- title,
- project name,
- folder name
- key words and
- Summary

At this stage there are two options that you can tick. These are:

- “Secure Storage” (where the document is not available to the general public)
- “Append” (where the document is to be appended to the end of an existing document of the same File Name, Project Name and Folder Name.

The UPLOAD Document tab will appear like the following:


Phenobank Document Manager
Document Manager
Progress
UPLOAD Document
SEARCH for Document
Upload MULTIPLE Documents
Settings
Fetch Document, File or Picture
Upload Document and Create Catalog Listing

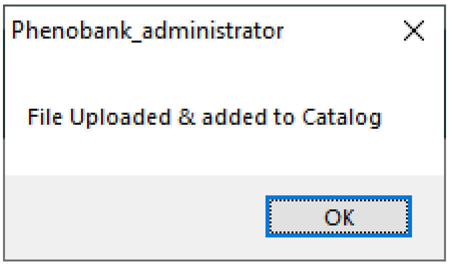
☑ Secure Storage
□ Append

File name: Dataset 1 CRC cows phenotypes Brahman.csv
Doc Title: Dataset 1 CRC cows phenotypes Brahman.csv
Project: CRC
Folder: Phenotypes
Key Words: Brahman
New Web Link:

Document Summary:
This is a dataset of phenotypes for the Beef CRC

To proceed with the Upload AND create the document catalog entry click “Upload Document etc.”

Wait for the following before proceeding:

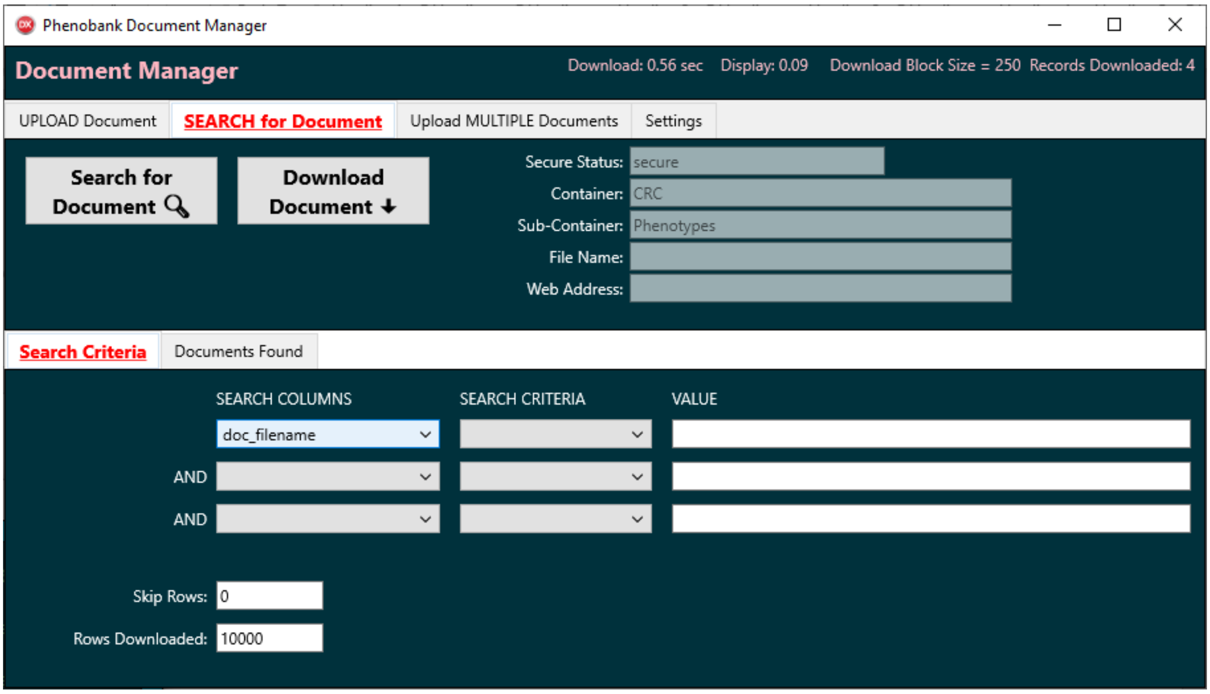


File Uploaded & added to Catalog

9.2 Searching for Document

9.2.1 Selecting a document from list of documents

Select the “SEARCH for Document” tab. To list ALL documents, do not enter any criteria other than the first search column, which can contain any of the drop-down column names, such as this:



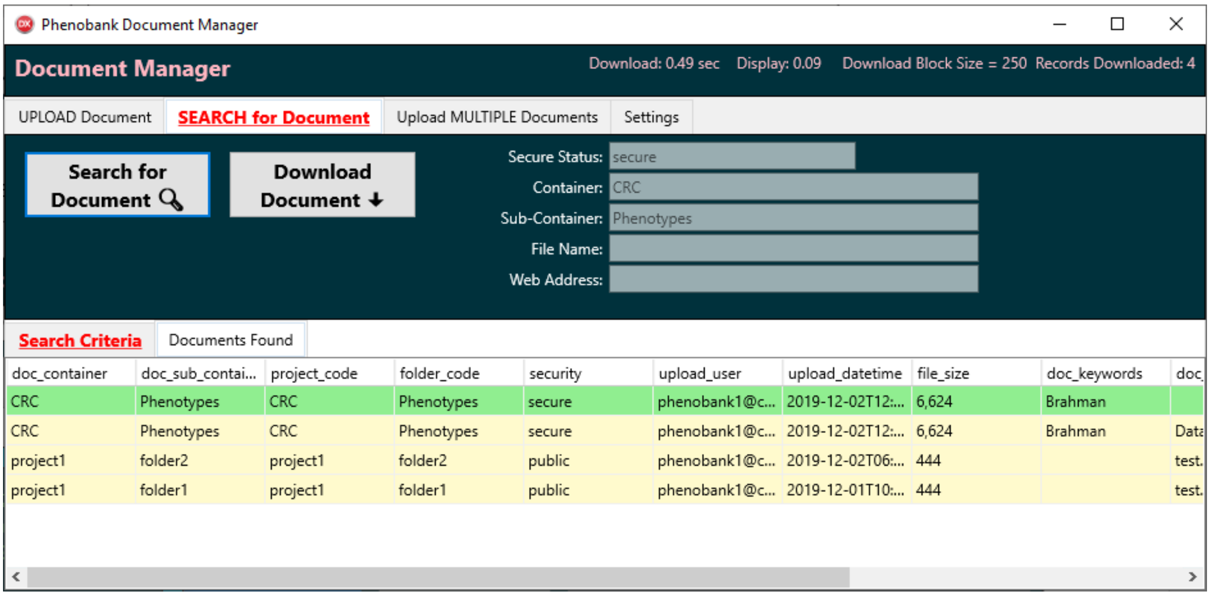
Download: 0.56 sec Display: 0.09 Download Block Size = 250 Records Downloaded: 4

Secure Status: secure
Container: CRC
Sub-Container: Phenotypes
File Name:
Web Address:

doc_filename
AND
AND

Skip Rows: 0
Rows Downloaded: 10000

Click “Download Document” and all Documents will be listed (only 4 in this example).



Download: 0.49 sec Display: 0.09 Download Block Size = 250 Records Downloaded: 4

Secure Status: secure
Container: CRC
Sub-Container: Phenotypes
File Name:
Web Address:

doc_container	doc_sub_contai...	project_code	folder_code	security	upload_user	upload_datetime	file_size	doc_keywords	doc_
CRC	Phenotypes	CRC	Phenotypes	secure	phenobank1@c...	2019-12-02T12:...	6,624	Brahman	
CRC	Phenotypes	CRC	Phenotypes	secure	phenobank1@c...	2019-12-02T12:...	6,624	Brahman	Data
project1	folder2	project1	folder2	public	phenobank1@c...	2019-12-02T06:...	444		test.
project1	folder1	project1	folder1	public	phenobank1@c...	2019-12-01T10:...	444		test.

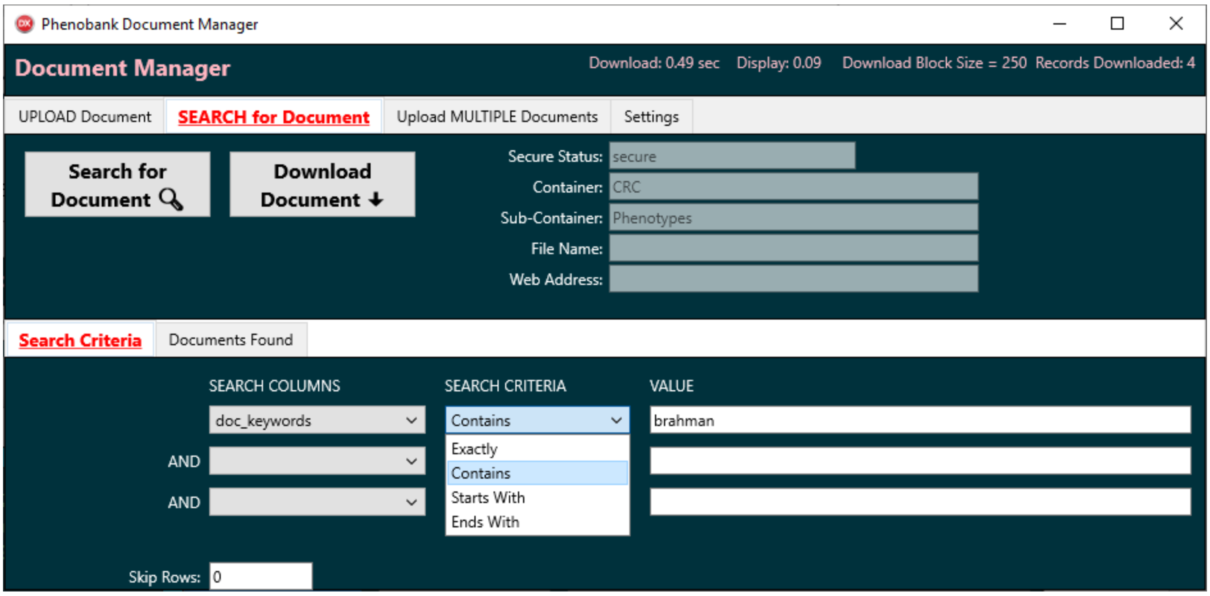
To select a document for Downloading, either:

- Click on the row representing the document of interest, and then click “Download Document” or,
- Double-click on the row

The document's web address will be passed to the default browser, initiating a download if the file is not viewable by the browser. The Web address will be a permanent address if the document is "public", otherwise for secure documents, the web address is a temporary, computer generated web address.

9.2.2 Searching for a range of documents

The range of documents can be reduced by entering restrictive criteria. In the below case, only those documents in the catalog with "brahman" in the "doc_keywords" column will be found.



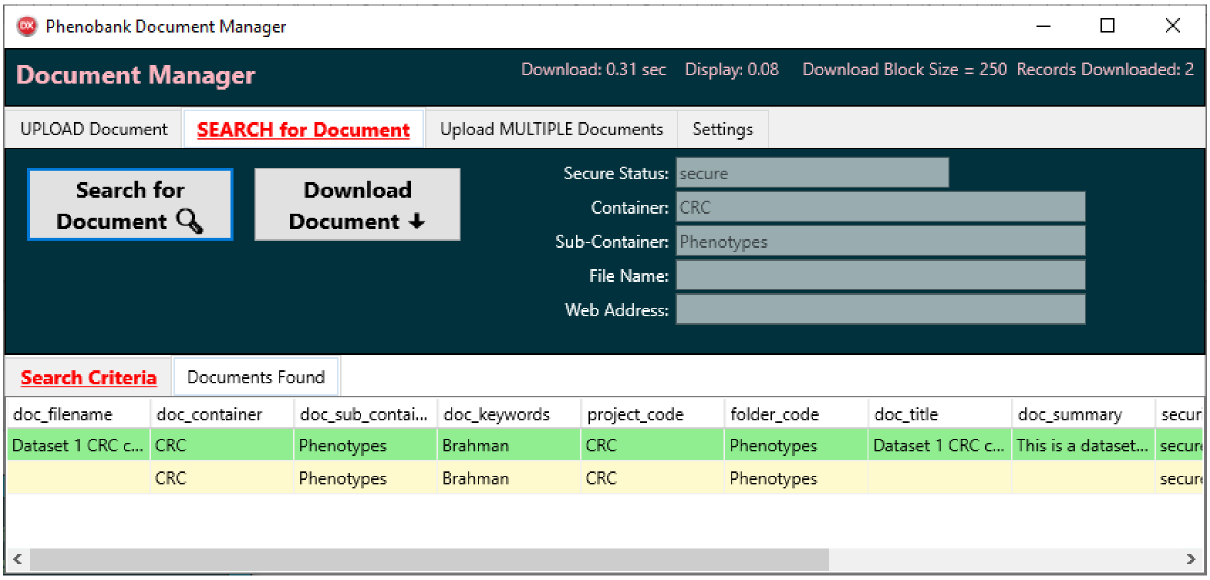
Download: 0.49 sec   Display: 0.09   Download Block Size = 250 Records Downloaded: 4

Secure Status: secure
Container: CRC
Sub-Container: Phenotypes
File Name:
Web Address:

	SEARCH COLUMNS	SEARCH CRITERIA	VALUE
	doc_keywords	Contains	brahman
AND		Exactly Contains Starts With Ends With	
AND			

Skip Rows: 0

Click "Search for Document" to reveal the reduced list. In this case, two of the documents satisfy the criteria:



Download: 0.31 sec   Display: 0.08   Download Block Size = 250 Records Downloaded: 2

Secure Status: secure
Container: CRC
Sub-Container: Phenotypes
File Name:
Web Address:

doc_filename	doc_container	doc_sub_contai...	doc_keywords	project_code	folder_code	doc_title	doc_summary	secur
Dataset 1 CRC c...	CRC	Phenotypes	Brahman	CRC	Phenotypes	Dataset 1 CRC c...	This is a dataset...	secur
	CRC	Phenotypes	Brahman	CRC	Phenotypes			secur

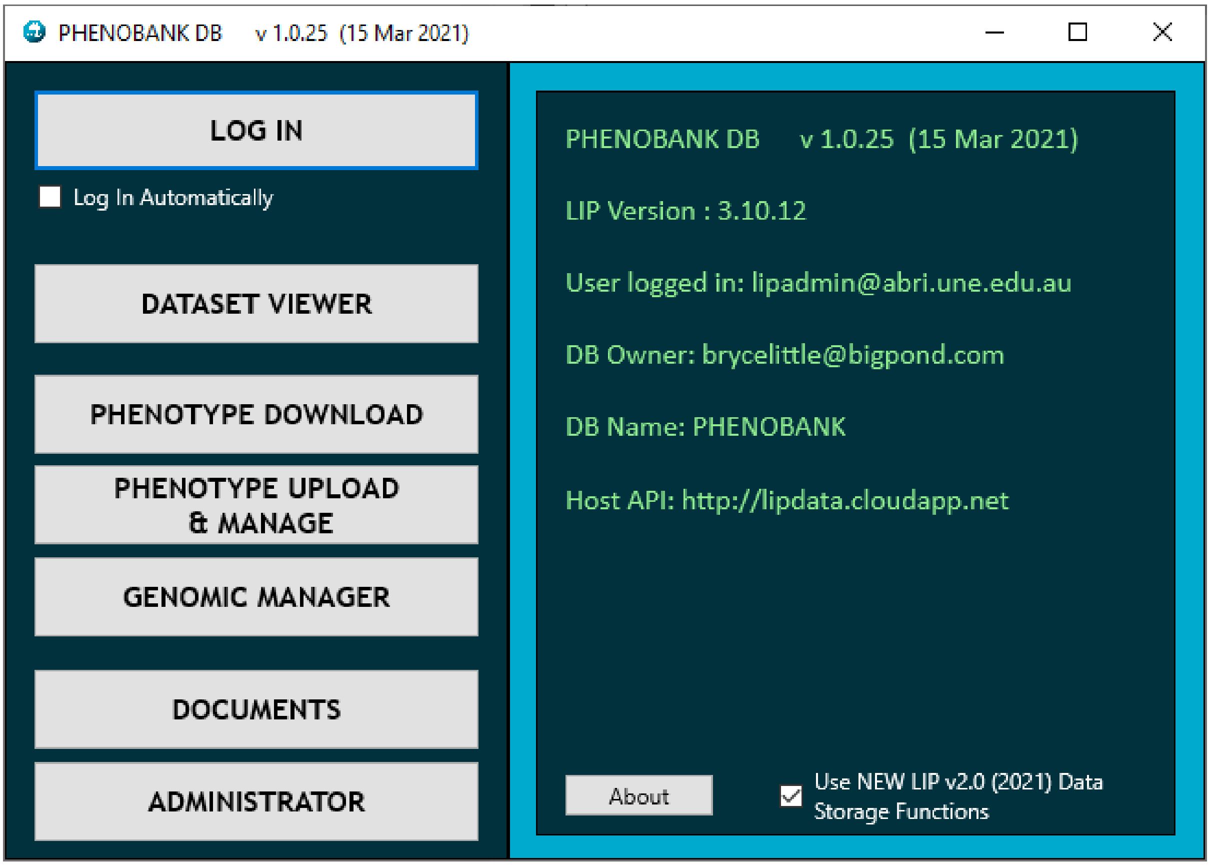
10 Administrator

The Administrator functions are accessible only if you the user logged in is the owner of the database. The administrator of the Phenobank DB is lipadmin@abri.une.edu.au. This user account will be managed a staff member of ABRI at the University of New England (currently Cody Wilson) and a staff member at CSIRO (currently Bryce Little).

The Administrator has overall access to all datasets within the PHENOBANK database whether uploaded by the administrator or uploaded by other users.

An important additional function available to the Administrator is to maintain a list of users who have permission to log into the database.

Click the “ADMINISTRATOR” button on the Main Window to gain access to the Administrator functions.



PHENOBANK DB v 1.0.25 (15 Mar 2021)

LIP Version : 3.10.12

User logged in: lipadmin@abri.une.edu.au

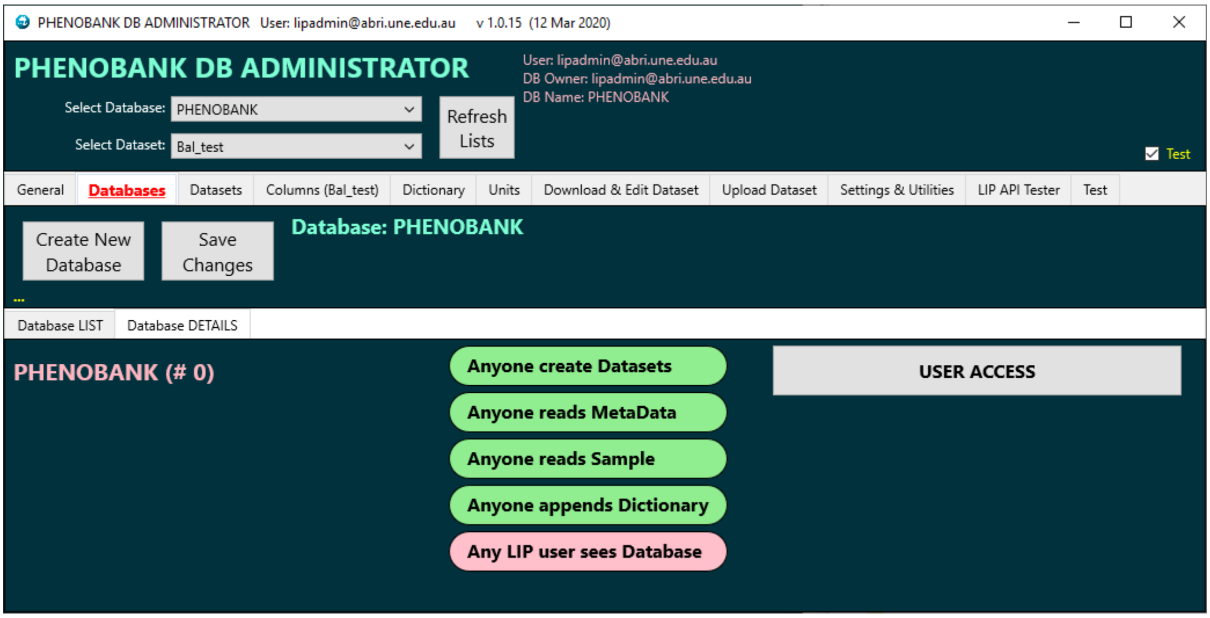
DB Owner: brycelittle@bigpond.com

DB Name: PHENOBANK

Host API: http://lipdata.cloudapp.net

☑ Use NEW LIP v2.0 (2021) Data Storage Functions

The Database Tab is accessible as below:



PHENOBANK DB ADMINISTRATOR User: lipadmin@abri.une.edu.au   v 1.0.15 (12 Mar 2020)

PHENOBANK DB ADMINISTRATOR

User: lipadmin@abri.une.edu.au
DB Owner: lipadmin@abri.une.edu.au
DB Name: PHENOBANK

Select Database: PHENOBANK

Select Dataset: Bal_test

☑ Test

Database: PHENOBANK

PHENOBANK (# 0)

10.1 Database Management

Select the “Databases” tab to gain access to the settings available for each Database. See the above Window image for an example. To edit the options available for each Database, either select from the drop down, or click on the row for the Database of interest.

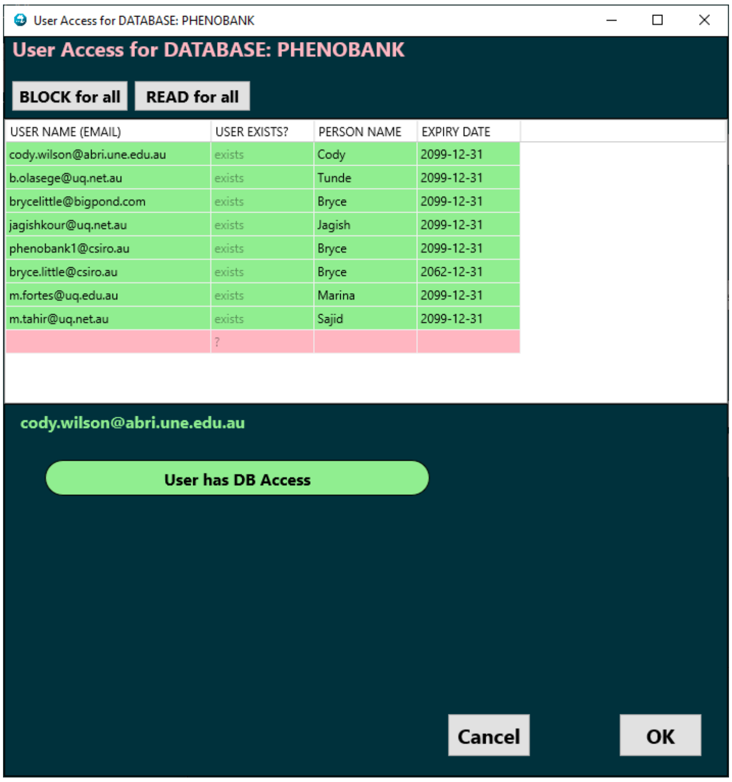
10.1.1 Setting Permissions

Three options are available for setting permission levels for all users. These can be changed by clicking on the relevant oval. Green for permission granted, pink for permission Not granted. The options available are, for each Database:

- Anyone can create Datasets in this Database
- Anyone can read the “MetaData” for Datasets in this Database. MetaData refers to the column names and other details about the Dataset, *other* than the data itself.
- Anyone can read a sample of Data (typically the first 10 lines) from any Dataset in this Database
- Anyone has permission to add new Dictionary Words to the common Phenobank Dictionary.
- Any user on the shared LIP platform is able to gain access to the Database, although individual Datasets within the database may not be accessible.

10.1.2 User Access

Click the User Access button to access a function to maintain the list of users who can access the database.


User Access for DATABASE: PHENOBANK
User Access for DATABASE: PHENOBANK
BLOCK for all
READ for all

USER NAME (EMAIL)	USER EXISTS?	PERSON NAME	EXPIRY DATE
cody.wilson@abri.une.edu.au	exists	Cody	2099-12-31
b.olasege@uq.net.au	exists	Tunde	2099-12-31
brycelittle@bigpond.com	exists	Bryce	2099-12-31
jagishkour@uq.net.au	exists	Jagish	2099-12-31
phenobank1@csiro.au	exists	Bryce	2099-12-31
bryce.little@csiro.au	exists	Bryce	2062-12-31
m.fortes@uq.edu.au	exists	Marina	2099-12-31
m.tahir@uq.net.au	exists	Sajid	2099-12-31
	?		

cody.wilson@abri.une.edu.au

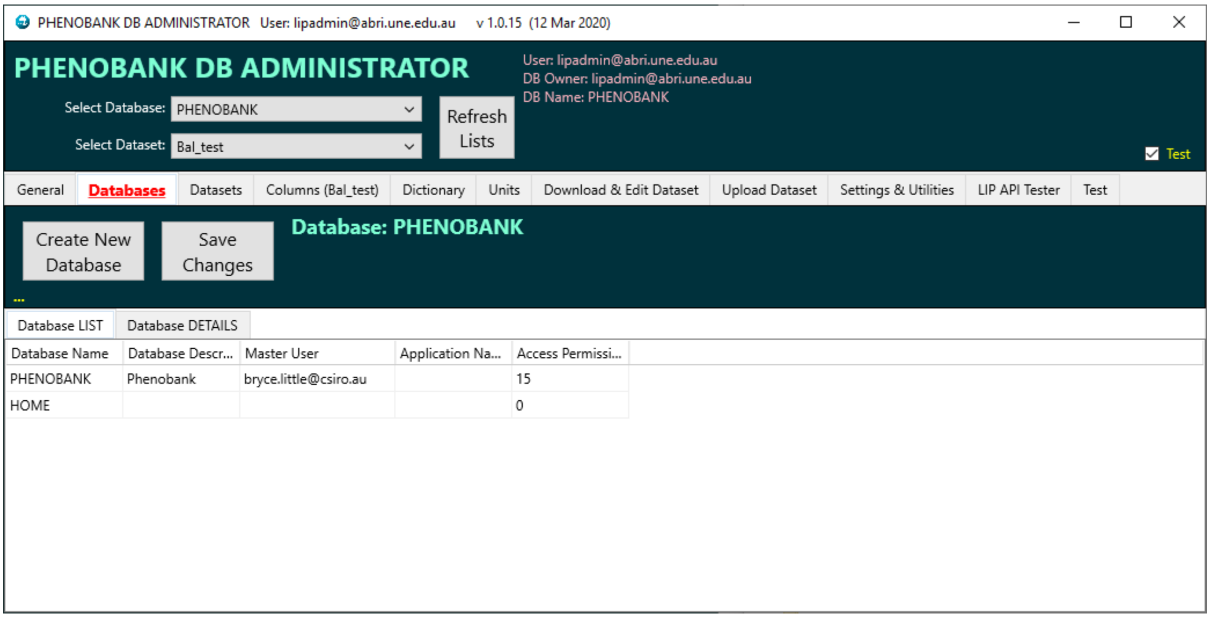
To Add a user, enter the new user name at the bottom of the “USER NAME” column and press Enter. If the users exists (Having previously been created by a LIP Administrator), the word “exists” will appear in the “USER EXISTS?” column.

To enable access or disable access, click the row with the user name of interest, and then click the round-ended shape. Green for access, pink for no access.

10.1.3 Master User

The “Master User” is a user, OTHER than the Database Owner who has been granted full permission to administer this database. This capability is important to allow a “back-up”, or deputy Administrator to be appointed so that management capability can continue if the primary Administrator is unavailable.

Select the Database Tab and the “Database List” Tab to be able to edit the master user.



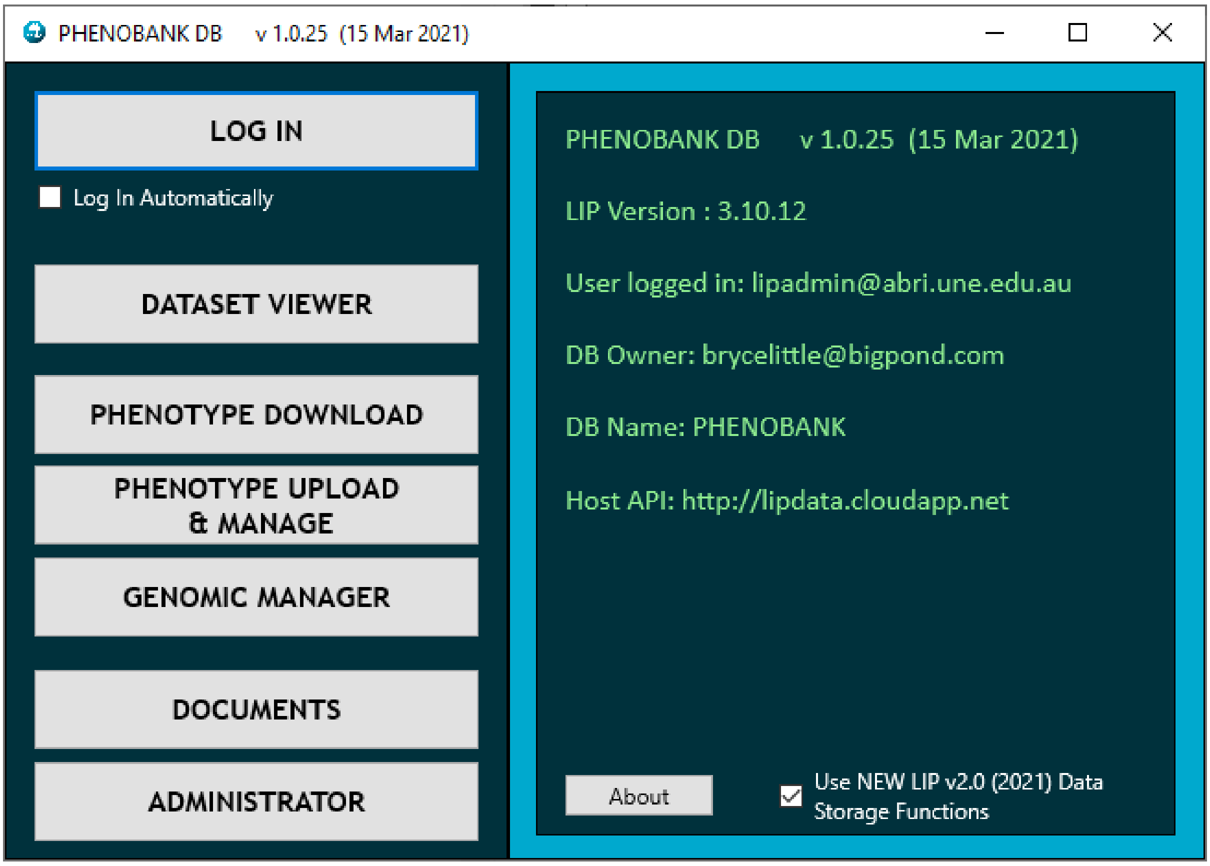
Database Name	Database Descr...	Master User	Application Na...	Access Permissi...
PHENOBANK	Phenobank	bryce.little@csiro.au		15
HOME				0

To enter the Master User, enter the person’s username in the “Master User” column, and the row of the Database of interest.

11 Genomic Manager

11.1 Accessing the Genomic Manager

At the main menu, select the “Genomic Manager” function.


PHENOBANK DB   v 1.0.25 (15 Mar 2021)
LOG IN
□ Log In Automatically
DATASET VIEWER
PHENOTYPE DOWNLOAD
PHENOTYPE UPLOAD & MANAGE
GENOMIC MANAGER
DOCUMENTS
ADMINISTRATOR
About
☑ Use NEW LIP v2.0 (2021) Data Storage Functions

PHENOBANK DB   v 1.0.25 (15 Mar 2021)

LIP Version : 3.10.12

User logged in: lipadmin@abri.une.edu.au

DB Owner: brycelittle@bigpond.com

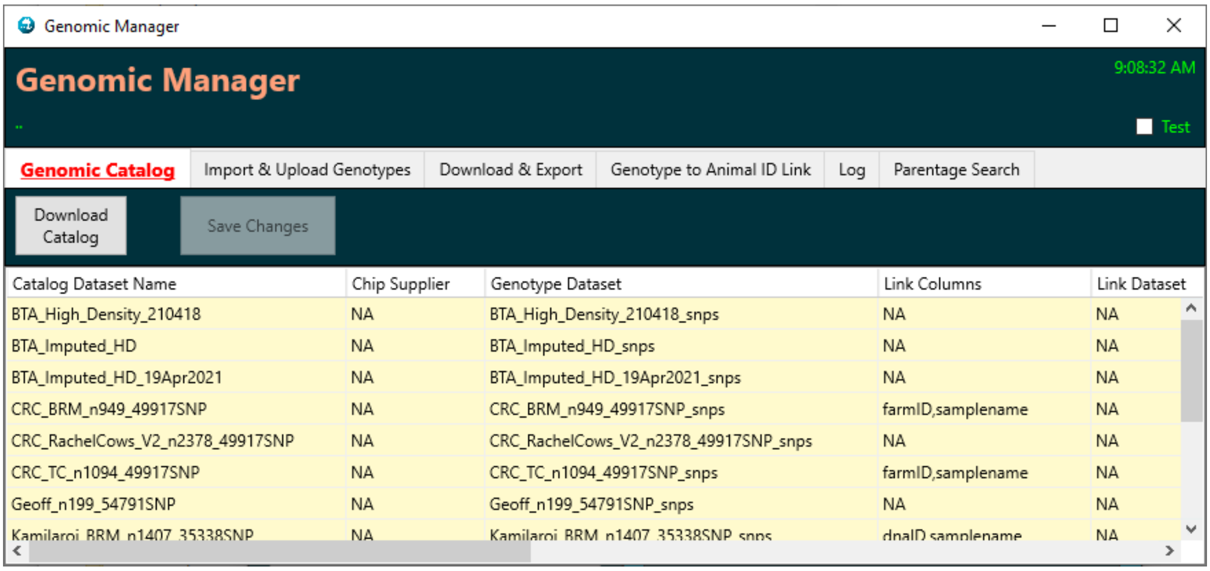
DB Name: PHENOBANK

Host API: http://lipdata.cloudapp.net

The Genomic Manager allows users to upload and download genotype data files in a variety of formats. The genotype (i.e. SNP) data is stored efficiently to minimise storage requirements and maximise speed of access.

11.2 The Genomic Catalog

The Genomic Manager maintains a Genomic “Catalog” that is a specialised dataset to store the links between the phenotype datasets and the genotype datasets. The Genomic Catalog is accessible via the Genomic Catalog Tab.



Catalog Dataset Name	Chip Supplier	Genotype Dataset	Link Columns	Link Dataset
BTA_High_Density_210418	NA	BTA_High_Density_210418_snps	NA	NA
BTA_Imputed_HD	NA	BTA_Imputed_HD_snps	NA	NA
BTA_Imputed_HD_19Apr2021	NA	BTA_Imputed_HD_19Apr2021_snps	NA	NA
CRC_BRM_n949_49917SNP	NA	CRC_BRM_n949_49917SNP_snps	farmID,samplename	NA
CRC_RachelCows_V2_n2378_49917SNP	NA	CRC_RachelCows_V2_n2378_49917SNP_snps	NA	NA
CRC_TC_n1094_49917SNP	NA	CRC_TC_n1094_49917SNP_snps	farmID,samplename	NA
Geoff_n199_54791SNP	NA	Geoff_n199_54791SNP_snps	NA	NA
Kamilaroi_BRM_n1407_35338SNP	NA	Kamilaroi_BRM_n1407_35338SNP_snps	dnaID,samplename	NA

Each line represents one Genomic Catalog record. Each record consists of:

1. Catalog Name
2. Chip Supplier (e.g. Illumina)
3. Genotype Dataset (name of Dataset containing SNP data)
4. Link Columns (name of common columns to link Genotype Sample with Phenotype Sample)
5. Link Dataset (Optional if the list of common values between Phenotype and Sample Datasets is housed in a separate dataset)
6. Map Dataset (the SNP definitions, chromosome and base pair positions)
7. Measure Dataset (for storing date specific measurements for animals listed in the Phenotype Dataset)
8. Phenotype Dataset (Animals and their Phenotypes that are descriptive rather than measurements)
9. Sample Dataset (The list of genotype sample records, identified by a unique “Samplename”)

A Catalog entry is automatically created via the process of Importing and Uploading a Genomic file.

11.2.1 Delete Genomic Catalog Entry

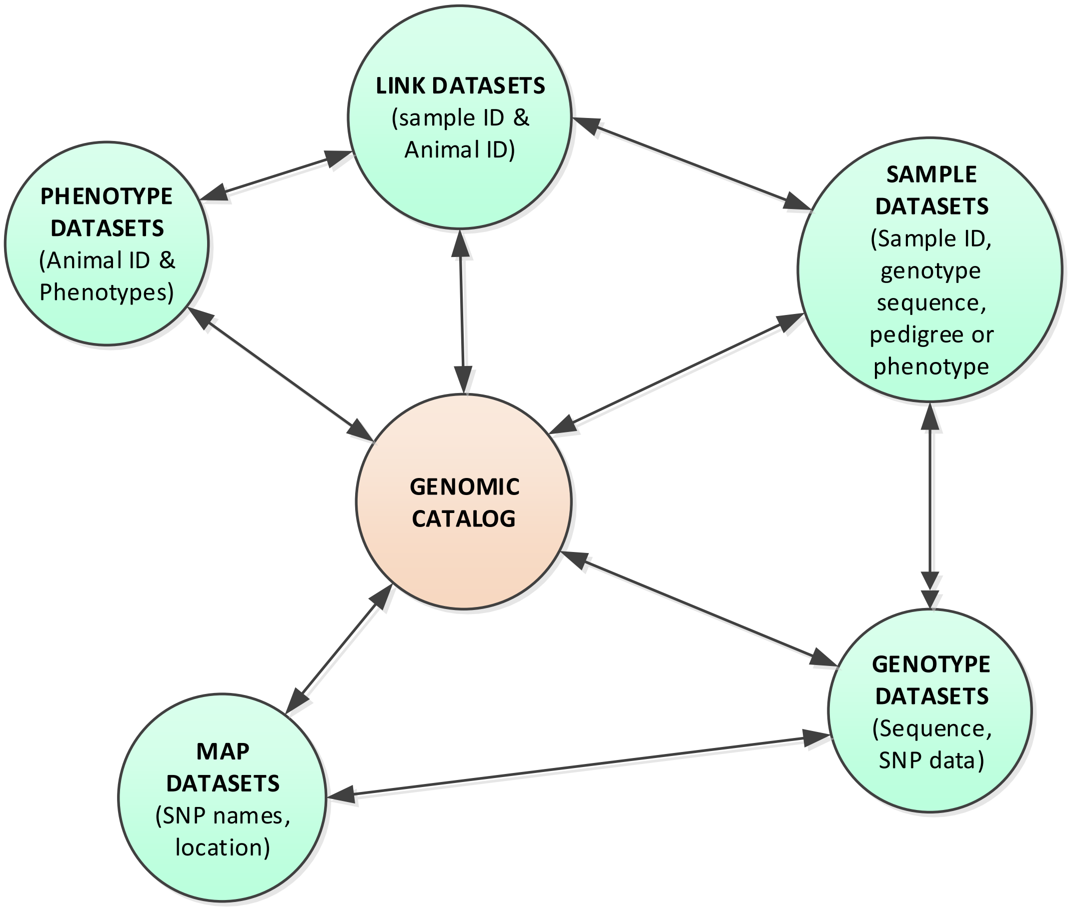
Users can delete the Catalog line, however the datasets given in that line will not be deleted. Datasets can be deleted via the Administrator functions. To delete a catalog entry follow these steps:

- Left click on Catalog row to select
- Right click to see the pop-up menu and select “Delete Catalog Item”
- Answer “Yes” to the question “Are you sure you wish to delete?”

11.3 Genomic dataset relationships

The following diagram shows the links between the various datasets that make up the complete set of Genotype and Phenotype data. Each Genomic Catalog record contains the names of:

- Genotype Dataset (containing SNP data with up to 26,000 SNPs per record)
- Sample Dataset (containing Sample ID and a Sequence number to link with the Genotype dataset.
- Link Dataset (containing Sample ID & Animal ID, and optionally Phenotype data)
- Phenotype Dataset (containing Animal ID & Phenoypes)
- Map Dataset (containing Snp Names, Snp Sequence number to link with the Genotype Dataset)



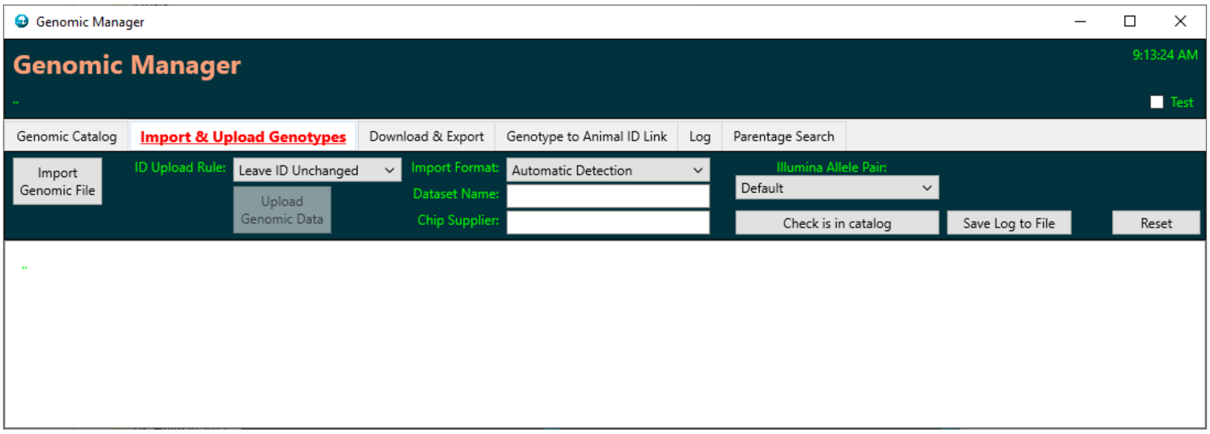
11.4 Import & Upload

Genomic Files can be imported and then uploaded to the LIP cloud server. Three formats are understood by the Phenobank DB. These are:

- PLINK (.ped and .map formats)
- VCF
- ILLUMINA Final Report
- “SIMPLE” (.geno and .map simple formats)

NOTE that SNP information is converted into the values 0 (Homozygous A,A), 1 (Heterozygous A,B), 2 (Homozygous B,B) and 3 (undefined). Quality, Allele order or CTAG information is not stored in the Phenobank database. If the user wishes to store the files as given, the Document function could be used for that purpose.

The Import & Upload Tab appears as follows:



ID Upload Rule: Leave ID Unchanged

Import Format: Automatic Detection

Dataset Name:

Chip Supplier:

Illumina Allele Pair: Default

There are two main steps: Importing and then Uploading.   An exception is if the Import Format selected is “SIMPLE”. In this case the uploading will occur automatically.

11.4.1 Import Genomic File

Click the “Import Genomic File” to select and import the file. Where there are multiple files, such as with PLINK, select one of the files and the system will automatically find the other file and import both, if they reside in the same folder. You may either select the import format manually or allow the system to automatically determine the format from the file content. If this is incorrectly given you will need to manually select the format.

This operation may take many minutes depending upon the size of the files. A typical import speed would be about 1 minute per GB for PLINK files.

When the operation is complete, the “Upload Genomic Data” button will be available.

11.4.2 Importing “Simple” format files

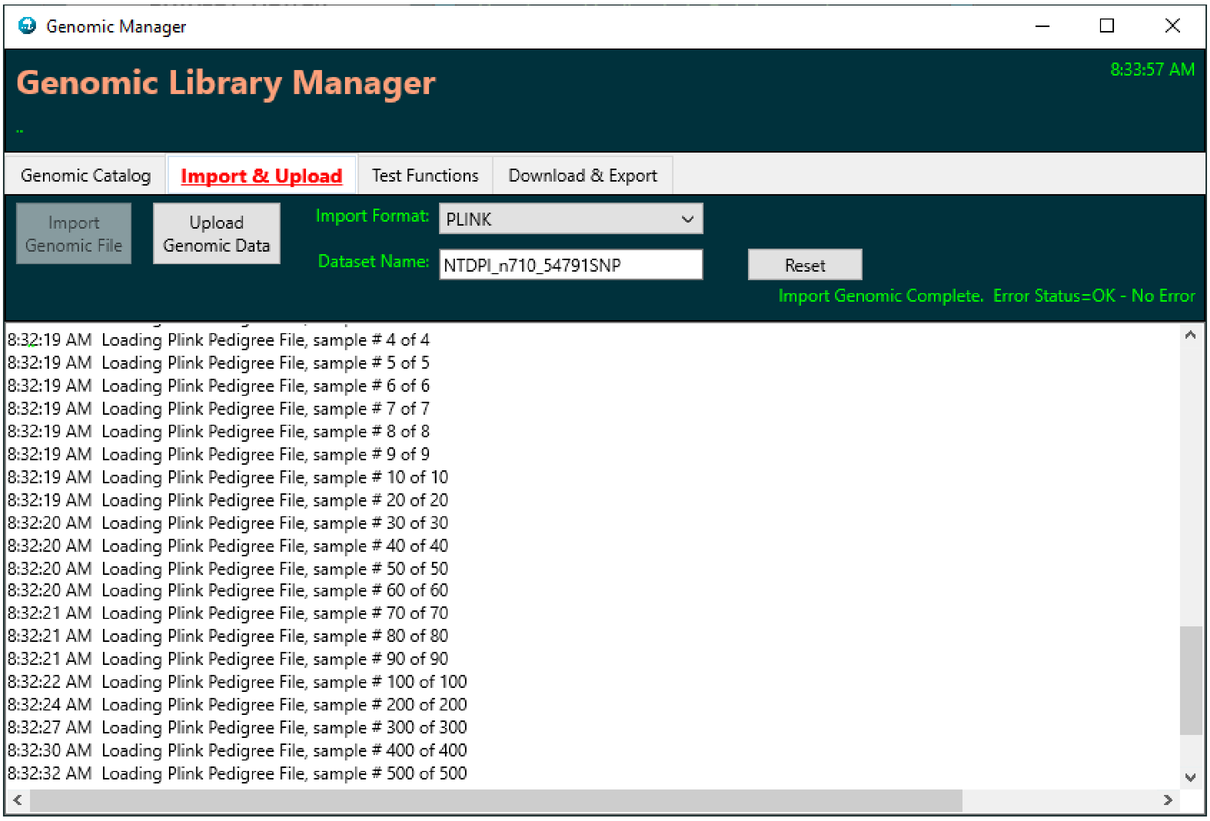
Importing Simple format files will be automatically followed by Uploading to the LIP database server. This is done to accommodate the uploading of massive Genotype datasets, split into multiple pairs of “.geno” and “.map” files. Each pair may represent the SNPs from one chromosome, but not necessarily so.

This method also will support average sized computers as less memory is required to temporarily hold the entire genotype dataset prior to uploading. See the separate chapter below for a detailed description of the format.

11.4.3 Upload Genomic Data

Prior to upload you will need to check the name of the Dataset. You can accept the one given by the program during the import, or change it now, prior to the upload. It is possible to change the name later, after upload, via the Administrator functions.

After Import and prior to the upload, the Tab may appear as follows.



Import Format: PLINK

Dataset Name: NTDPI_n710_54791SNP

Import Genomic Complete. Error Status=OK - No Error

```
8:32:19 AM  Loading Plink Pedigree File, sample # 4 of 4
8:32:19 AM  Loading Plink Pedigree File, sample # 5 of 5
8:32:19 AM  Loading Plink Pedigree File, sample # 6 of 6
8:32:19 AM  Loading Plink Pedigree File, sample # 7 of 7
8:32:19 AM  Loading Plink Pedigree File, sample # 8 of 8
8:32:19 AM  Loading Plink Pedigree File, sample # 9 of 9
8:32:19 AM  Loading Plink Pedigree File, sample # 10 of 10
8:32:19 AM  Loading Plink Pedigree File, sample # 20 of 20
8:32:20 AM  Loading Plink Pedigree File, sample # 30 of 30
8:32:20 AM  Loading Plink Pedigree File, sample # 40 of 40
8:32:20 AM  Loading Plink Pedigree File, sample # 50 of 50
8:32:20 AM  Loading Plink Pedigree File, sample # 60 of 60
8:32:21 AM  Loading Plink Pedigree File, sample # 70 of 70
8:32:21 AM  Loading Plink Pedigree File, sample # 80 of 80
8:32:21 AM  Loading Plink Pedigree File, sample # 90 of 90
8:32:22 AM  Loading Plink Pedigree File, sample # 100 of 100
8:32:24 AM  Loading Plink Pedigree File, sample # 200 of 200
8:32:27 AM  Loading Plink Pedigree File, sample # 300 of 300
8:32:30 AM  Loading Plink Pedigree File, sample # 400 of 400
8:32:32 AM  Loading Plink Pedigree File, sample # 500 of 500
```

Click “Upload Genomic Data” to proceed with the upload of the data to the cloud. This function will create the following in the cloud:

- A “MAP” dataset with a record for each SNP definition
- A “SAMPLE” dataset with a record for each animal Sample name (but not SNPs)

- A “SNP” dataset with one or more records for each animal Sample. Typically Phenobank will store approximately 26,000 SNPs per record.

Upload speed will be typically 80 SNP MAP descriptions per second. (That is around 10 minutes for a 50K Chip definition). For the SNP data proper, upload speed is typically 2 million SNPs per second.

11.4.4 Reset button

If the tab has a Genotype dataset in memory, and is ready for Uploading, the dataset will be cleared from memory and the tab “reset” to be ready to import a new genotype dataset.

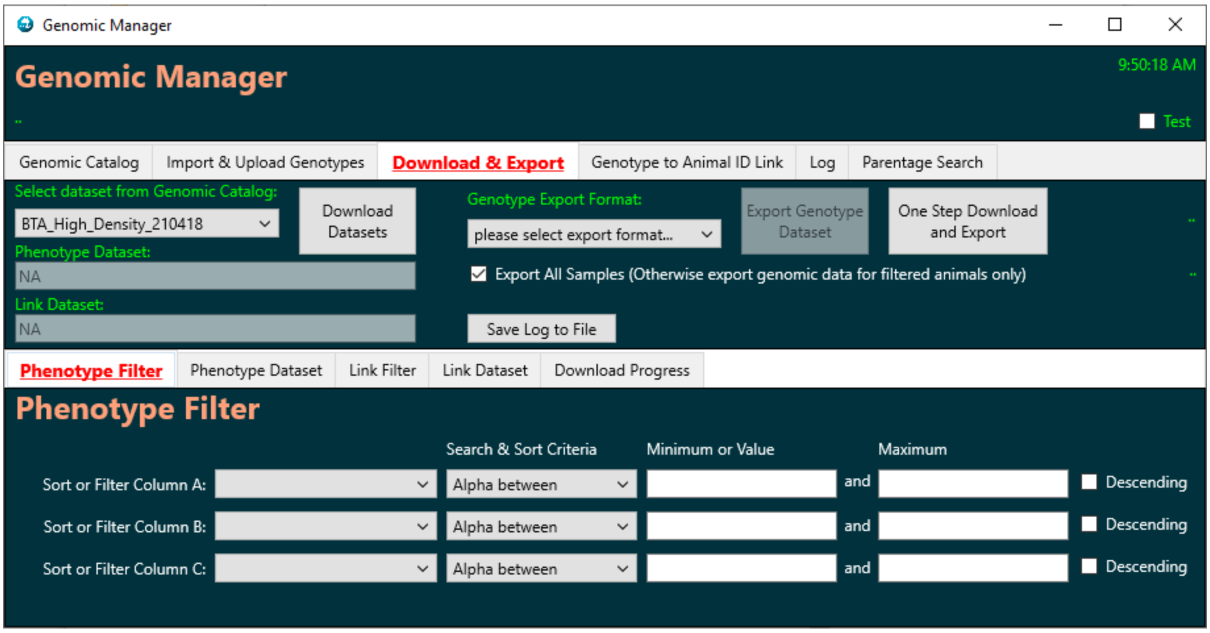
11.4.5 “Check is in catalog” button

A quick check to determine if the given Dataset Name is already in the Genomic Catalog.

ty

11.5 Download & Export

Select the Download and Export tab to retrieve Genomic Data from the Phenobank DB.



Select dataset from Genomic Catalog: BTA_High_Density_210418

Phenotype Dataset: NA

Link Dataset: NA

☑ Export All Samples (Otherwise export genomic data for filtered animals only)

11.5.1 Download Genomic Dataset

Select the Genomic Catalog name from the drop-down box. If a phenotype dataset has been linked to this catalog name, the user can select the range of animals of interest. For example, in the following, animals with a Rebreeding Score of 4 and a Year-of-Birth of 2010 will be retrieved when “Download Dataset” is selected.

This function will retrieve the list of animals that satisfy the users criteria and the Genotype Data associated with the Catalog name.

11.5.2 Export Genomic Dataset

To export the genomic data as either PLINK, VCF or Illumina format, follow these steps:

- First, you must download the genomic dataset as explained in section 11.5.1 above. This may take several minutes depending upon the size of the genomic dataset.
- Select the export format from the “Export format:” drop down list, such as PLINK, Illumina or VCF.
- Click the “Export Genomic Dataset” button, and when requested select the folder and file name to which the data is to be exported to.

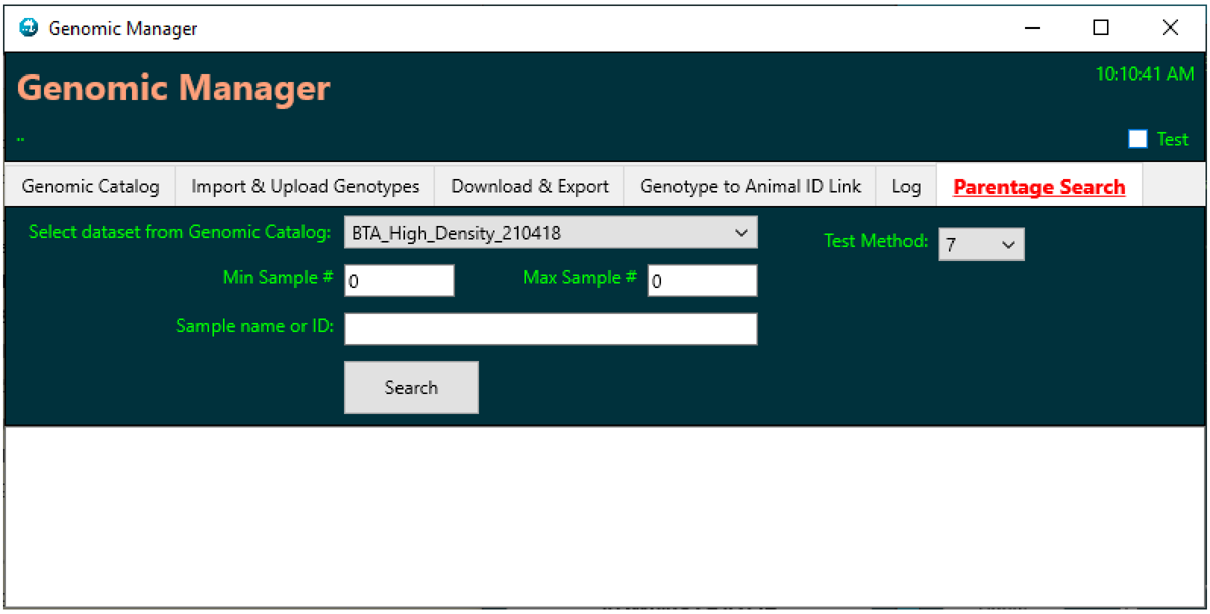
11.5.3 One Step Download and Export

Both steps to Download to memory and Export to a file can be undertaken with this button. As of Phenobank DB Build #25 (April 2021) you will still need enough RAM to store the entire genotype dataset. That is around 40 billion SNPS for a 16GByte computer.

11.5.4 Parentage search

Samples that are ancestors or descendants of the nominated sample are found and listed. These are based upon compatible alleles for each SNP.

Samples that are close matches and likely to be related are also listed.



Select dataset from Genomic Catalog: BTA_High_Density_210418

Test Method: 7

Min Sample # 0

Max Sample # 0

Sample name or ID:

11.6 Link Genotype Sample ID to Animal ID

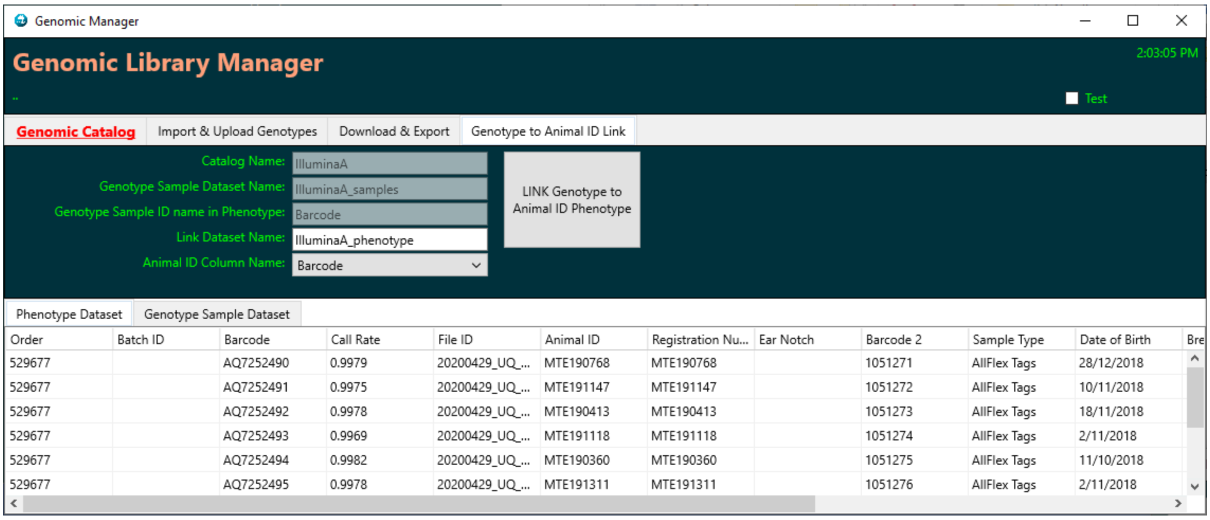
11.6.1 Select Genomic datasets from Catalog

- Open Genomic Manager, and select “Genomic Catalog” tab.

11.6.2 Import Animal ID file

- Left-Click on row with the Genomic datasets previously imported.
- RIGHT-Click to see pop-up menu and select the option: “Link & Import Animal ID File”.
- A Windows file search window will appear. Use it to select a CSV file with the Animal ID and the Genotype Sample ID as used to identify the Genotype samples. This file may or may not include additional columns such as phenotype data. Find file and click “Open”.
- You will be taken to the “Genotype to Animal ID Link” tab. Select the “Animal ID Column Name” from the drop-down list of possible contenders.
- Click the “Link Genotype to Animal ID Phenotype” button.

The link file will then be uploaded, and the catalog updated accordingly with the name of the Link dataset and the link fields.



Catalog Name: IlluminaA

Genotype Sample Dataset Name: IlluminaA_samples

Genotype Sample ID name in Phenotype: Barcode

Link Dataset Name: IlluminaA_phenotype

Animal ID Column Name: Barcode

Order	Batch ID	Barcode	Call Rate	File ID	Animal ID	Registration Nu...	Ear Notch	Barcode 2	Sample Type	Date of Birth	Bre
529677		AQ7252490	0.9979	20200429_UQ_...	MTE190768	MTE190768		1051271	AllFlex Tags	28/12/2018	
529677		AQ7252491	0.9975	20200429_UQ_...	MTE191147	MTE191147		1051272	AllFlex Tags	10/11/2018	
529677		AQ7252492	0.9978	20200429_UQ_...	MTE190413	MTE190413		1051273	AllFlex Tags	18/11/2018	
529677		AQ7252493	0.9969	20200429_UQ_...	MTE191118	MTE191118		1051274	AllFlex Tags	2/11/2018	
529677		AQ7252494	0.9982	20200429_UQ_...	MTE190360	MTE190360		1051275	AllFlex Tags	11/10/2018	
529677		AQ7252495	0.9978	20200429_UQ_...	MTE191311	MTE191311		1051276	AllFlex Tags	2/11/2018	

11.7 Link Phenotype Dataset to Genotype

In addition to the “Link Dataset”, another dataset can be linked to the genotype datasets. Typically, this would be an existing dataset that contains a column with the same Animal ID as that in the Link Dataset. To make the link:

- Select the “Genomic Catalog” tab in the Genomic Manager.

- Left click on the row with the genomic catalog record that you wish to link the phenotypes.
- Right click and select “Link Phenotype Dataset”.

A new window will appear listing the existing phenotype datasets. Now:

- Click on the dataset that you wish to link
- Click “OK”.

The selected dataset’s name will now be inserted into the genomic catalog record in the “Phenotype Dataset”.

If there is no field in the Phenotype Dataset that matches the fields given in the “Link Columns”, you will be given a warning message.

11.8 Duplicate ID Test

A test is available in the Genomic Manager to check that sample identifiers and animal identifiers are UNIQUE.

To access this function, follow these steps:

- Select the Genomic Catalog
- Select the row with the genomic and phenotype datasets that you wish to check.
- Right Click and select the “Unique ID Check” function.

The test will commence to ensure that the following tables in the row will be checked for duplicate ID’s:

- Phenotype Dataset
- Link Dataset
- Sample Dataset

The columns to be checked are listed in the “Link Columns” column in the genomic catalog.

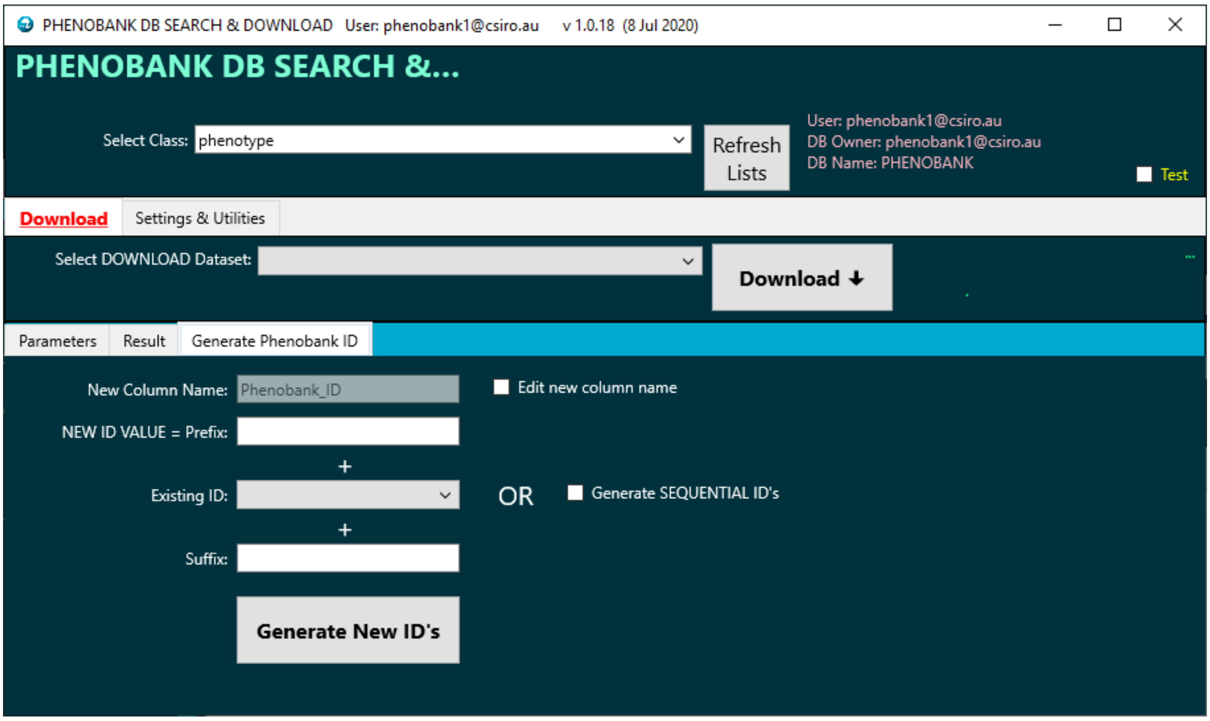
12 Generating Phenobank ID

12.1 Accessing Phenobank ID Generator

The “Generate Phenobank ID” function is available via the “Search & Download” function. To access:

- select “DOWNLOAD” at the main menu,
- click the “Search & Download” button to view a SEARCH & DOWNLOAD Window.
- Click the “Download Tab”
- Click the “Generate Phenobank ID” Tab

Your current window should appear like the following:



PHENOBANK DB SEARCH & DOWNLOAD User: phenobank1@csiro.au v 1.0.18 (8 Jul 2020)

PHENOBANK DB SEARCH &...

Select Class: phenotype

User: phenobank1@csiro.au
DB Owner: phenobank1@csiro.au
DB Name: PHENOBANK

Select DOWNLOAD Dataset:

New Column Name: Phenobank_ID

NEW ID VALUE = Prefix:

Existing ID:

Suffix:

12.2 Generating Phenobank ID’s

- Select, or type in, the Class of Datasets to help locate the Dataset that you wish to apply the ID generator. Typically, this will be “phenotype”.
- Select the DOWNLOAD Dataset that you wish to add the Phenobank ID.
- Enter the New Column Name if the default “Phenobank_ID” is not the name that you wish to call the new column. Tick the “Edit new column name” box to enter a new name.
- Enter a Prefix and Suffix for the new ID. The prefix will default to “PBID-SelectedDatasetName-“ as the Prefix. This can be edited prior to generating the new Phenobank ID’s.

- Select the column name with the existing ID from the drop-down box
- Click “Generate New ID’s” to create the new column and populate with the new ID’s.

13 Protocols

13.1 Phenobank upload protocol for phenotypes

13.1.1 Phenotype source file

The source file must be either a CSV (comma separated value) text file, or a Unicode text file with tab delimiters.

The first line is the header containing the column names.

The second and subsequent lines contain the data. Each line presents a single, unique animal and its phenotype values. These might include descriptive data such as colour or breed, as well as performance data such as average growth rate, or EBV.

Note that one of the columns should contain a unique animal identifier within the file.

13.1.2 Uploading phenotype data

Refer to section 8.2 for instructions to upload the file to Phenobank DB and create a dataset.

13.1.3 Generating a unique identifier

A unique "Phenobank" identifier will need to be created for each animal to ensure that the animal is uniquely identified when datasets are merged and downloaded for analysis. Refer to chapter 12 for instructions to add a unique identifier to a dataset.

13.2 Phenobank upload protocol for genotypes

13.2.1 Acceptable genotype source files

Users of Phenobank DB may provide either:

- Illumina
- PLINK

13.2.2 Uploading Illumina files

Three Illumina files can be uploaded:

- Final Report (with Genotype data)
- A Sample ID to Animal ID link file
- A Map file

All of these files should be supplied directly from the genotype testing company.

The Illumina Final Report is uploaded via the Genomic Manager upload described in section 11.4.

The Link file is uploaded via the Genomic Manager link function described in section 11.6.

The Map file is uploaded via the Upload dataset function described in section 8.2 AND then the user should link this to the genotype dataset described in section 11.7.

13.2.3 Uploading PLINK files

Two PLINK files are required to successfully upload genotype data. These are:

- The “ped” file, which contains both the Animal ID and the Genotype data
- The “map” file, which contains details of SNP name, location, chromosome etc.

These can be uploaded via the process described in section 11.4.

In addition, a third CSV “link” file may be required to link the genotype sample file with phenotypes. This file must contain both the Sample ID’s and the Animal ID’s. Refer to section 11.6 for instructions.

14 Format Descriptions

14.1 Simple Format

The Simple format has two file types, ".geno" and ".map".

14.1.1 Simple "geno" file type

There is no header. Each line represents one sample. The fields are:

1. Sample name followed by a delimiter, which can be a TAB or a SPACE.
2. A list of SNP values being either 0,1,2 or (3 or 5) representing AA, AB, BB or Undefined respectively. The list may be optionally delimited by spaces. The order of the SNPs being represented is defined by the corresponding "map" file with the otherwise identical file name.

14.1.2 Simple "map" file type

There is no header. Each line represents one SNP. The fields are tab or space delimited representing:

1. Chromosome
2. SNP name
3. Undefined at present
4. Base pair position

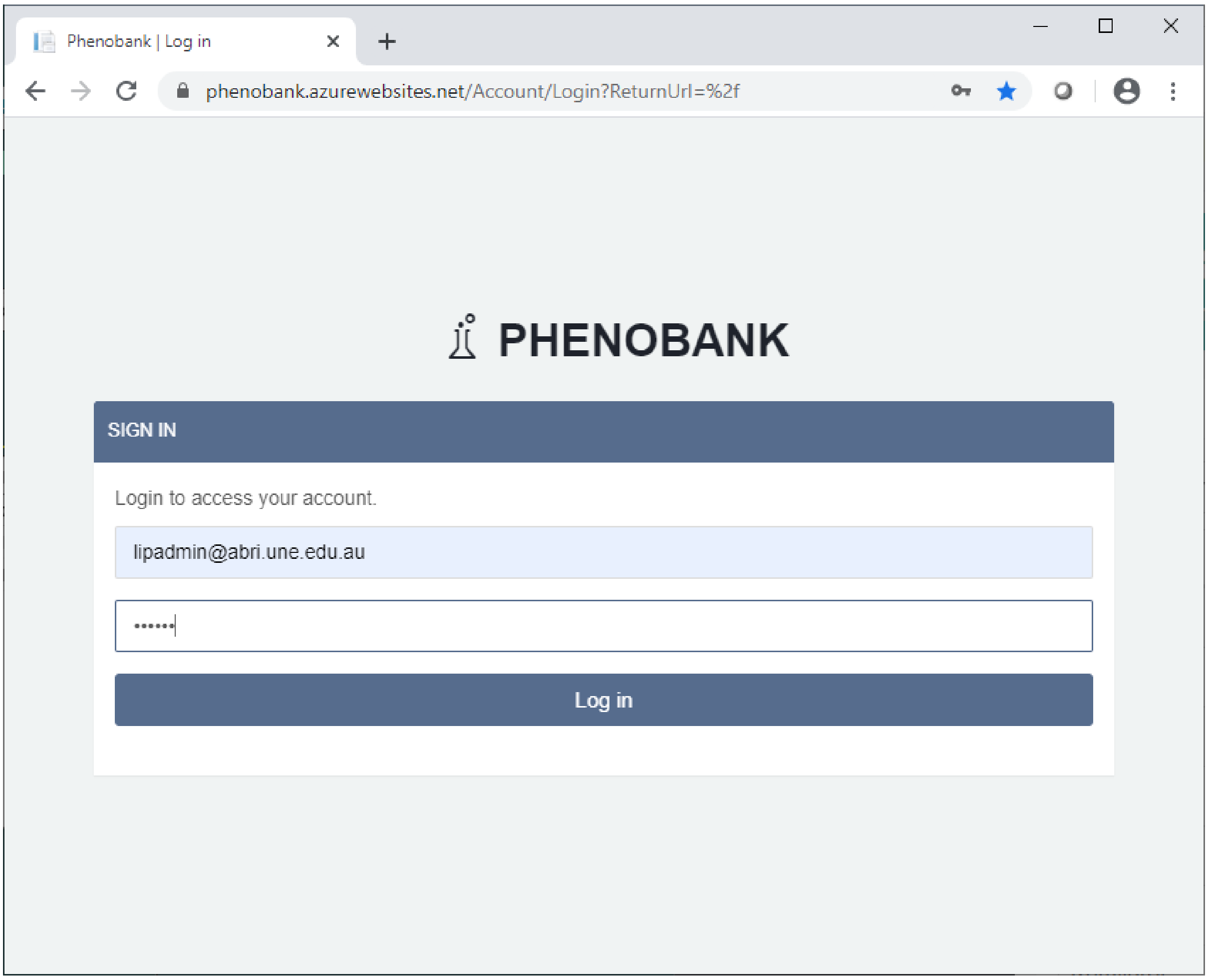
15 Phenobank on the Web

15.1 Accessing the Phenobank Web Page

The Phenobank DB Web Access Portal is accessible here:

phenobank.azurewebsites.net

The page will appear as below. Enter your username and password. New users will be supplied with a password by the Phenobank Administrator.



Login to access your account.

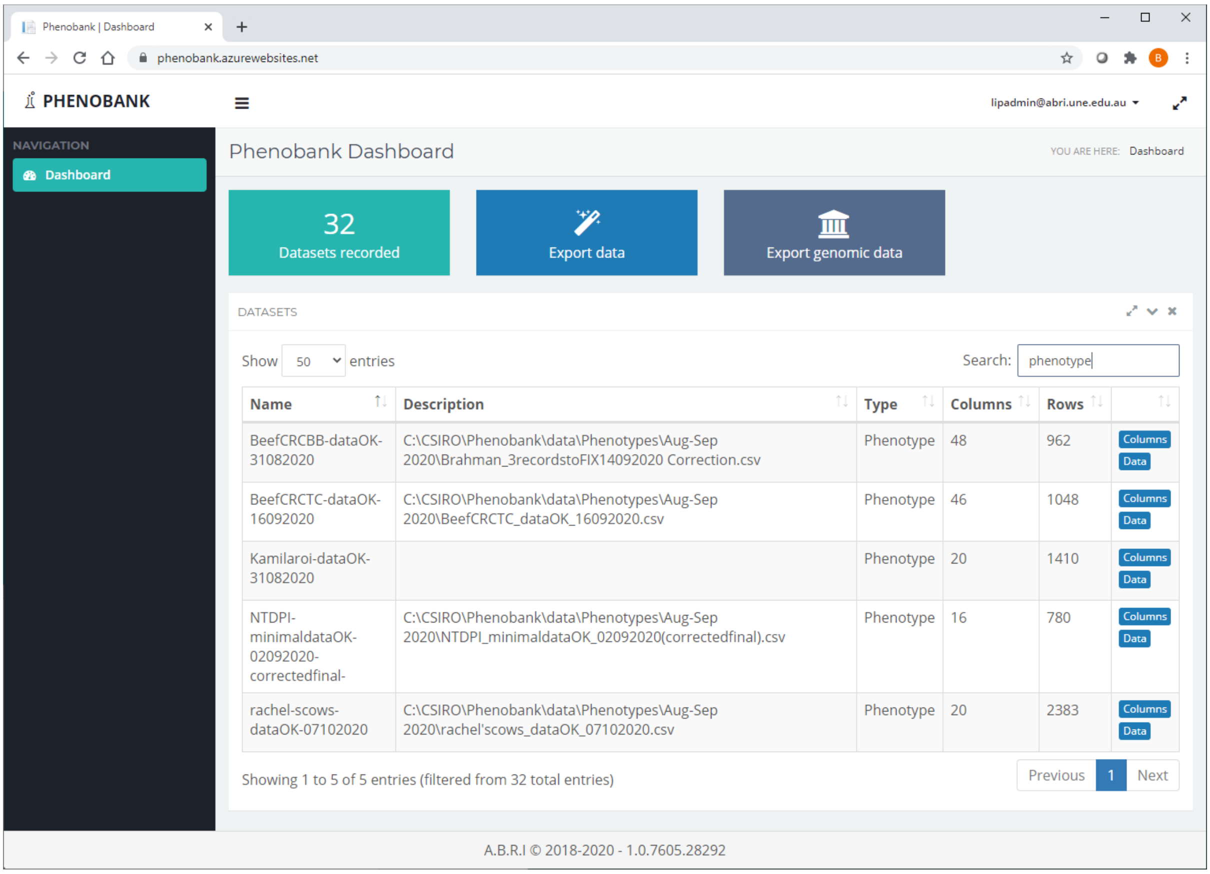
lipadmin@abri.une.edu.au

••••••

15.2The Phenobank Dashboard Page

The initial page gives users access to a list of all available datasets. Click “Columns” to access the column details and click “Data” to see a sample of the first 5 or 10 lines if users have permission to see.

To see only phenotype datasets, enter “phenotype” in the Search box at top right above the table.



Show 50 entries

Search: phenotype

Name	Description	Type	Columns	Rows	
BeefCRCBB-dataOK-31082020	C:\CSIRO\Phenobank\data\Phenotypes\Aug-Sep 2020\Brahman_3recordstoFIX14092020 Correction.csv	Phenotype	48	962	Columns Data
BeefCRCTC-dataOK-16092020	C:\CSIRO\Phenobank\data\Phenotypes\Aug-Sep 2020\BeefCRCTC_dataOK_16092020.csv	Phenotype	46	1048	Columns Data
Kamilaroi-dataOK-31082020		Phenotype	20	1410	Columns Data
NTDPI-minimaldataOK-02092020-correctedfinal-	C:\CSIRO\Phenobank\data\Phenotypes\Aug-Sep 2020\NTDPI_minimaldataOK_02092020(correctedfinal).csv	Phenotype	16	780	Columns Data
rachel-scows-dataOK-07102020	C:\CSIRO\Phenobank\data\Phenotypes\Aug-Sep 2020\rachel'scows_dataOK_07102020.csv	Phenotype	20	2383	Columns Data

Showing 1 to 5 of 5 entries (filtered from 32 total entries)

Previous 1 Next

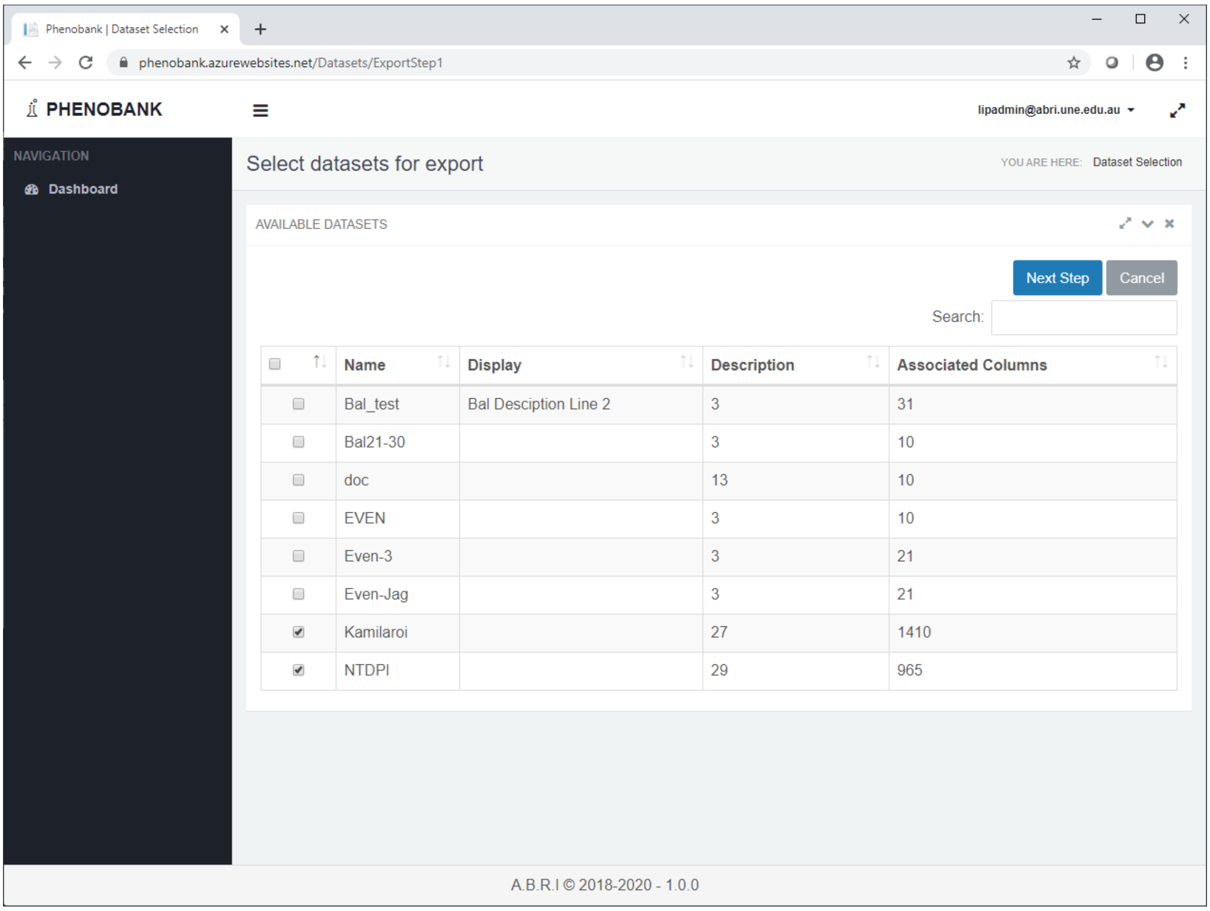
A.B.R.I © 2018-2020 - 1.0.7605.28292

15.3 Export Phenotype Data

Click the “Export Data” box to download a text file with data merged from one or more PHENOTYPE datasets.

15.3.1 Select Dataset for export

Tick the box at the left hand if you wish to include the dataset in the export file. Click the “Next Step” when ready.



Phenobank | Dataset Selection
phenobank.azurewebsites.net/Datasets/ExportStep1
PHENOBANK
lipadmin@abri.une.edu.au
NAVIGATION
Dashboard
YOU ARE HERE: Dataset Selection
AVAILABLE DATASETS
Next Step
Cancel
Search:
A.B.R.I © 2018-2020 - 1.0.0

Select datasets for export

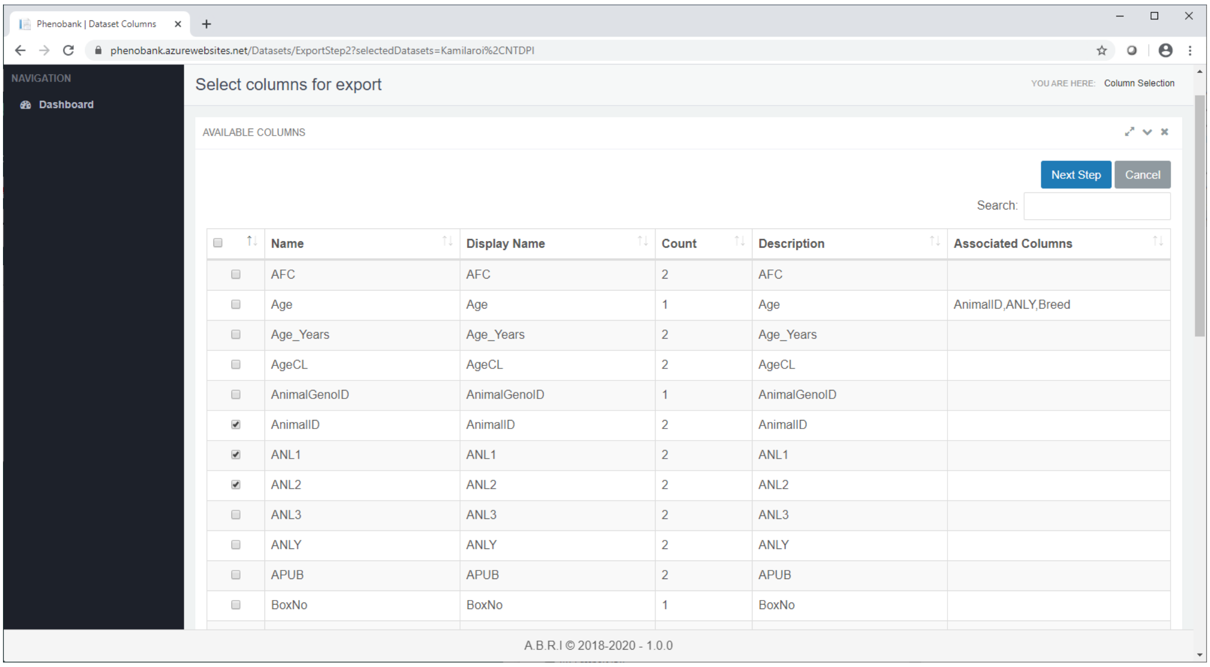
□	Name	Display	Description	Associated Columns
□	Bal_test	Bal Desciption Line 2	3	31
□	Bal21-30		3	10
□	doc		13	10
□	EVEN		3	10
□	Even-3		3	21
□	Even-Jag		3	21
☑	Kamilaroi		27	1410
☑	NTDPI		29	965

15.3.2 Select Columns for export

This Page allows the user to select which columns that need to be exported. As for datasets, the small square at the left hand of each row can be ticked to be selected. Alternatively, all columns can be selected in one step by ticking the square at the top left corner of the table.

The “Count” column informs the user how many of the datasets that the column is present.

Click the “Next Step” box to continue.


Phenobank | Dataset Columns
phenobank.azurewebsites.net/Datasets/ExportStep2?selectedDatasets=Kamilaroi%2CNTDPI
NAVIGATION
Dashboard
YOU ARE HERE: Column Selection
AVAILABLE COLUMNS
Next Step
Cancel
Search:

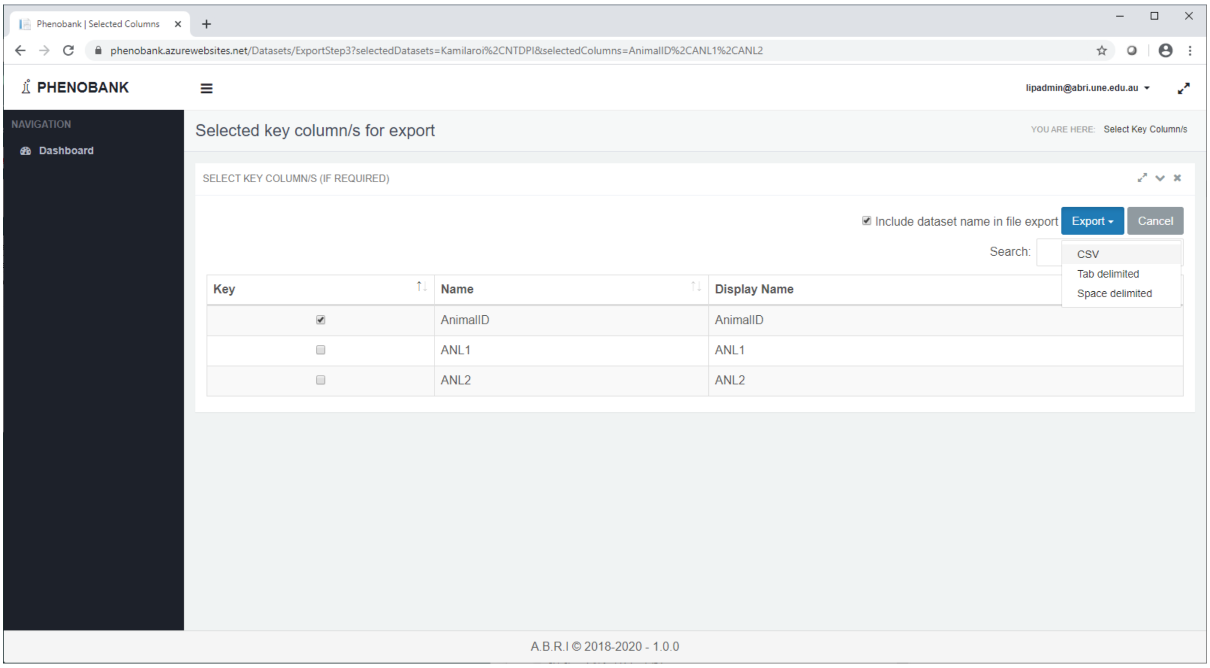
Select columns for export

□	Name	Display Name	Count	Description	Associated Columns
□	AFC	AFC	2	AFC	
□	Age	Age	1	Age	AnimalID,ANLY,Breed
□	Age_Years	Age_Years	2	Age_Years	
□	AgeCL	AgeCL	2	AgeCL	
□	AnimalGenoID	AnimalGenoID	1	AnimalGenoID	
☑	AnimalID	AnimalID	2	AnimalID	
☑	ANL1	ANL1	2	ANL1	
☑	ANL2	ANL2	2	ANL2	
□	ANL3	ANL3	2	ANL3	
□	ANLY	ANLY	2	ANLY	
□	APUB	APUB	2	APUB	
□	BoxNo	BoxNo	1	BoxNo	

A.B.R.I © 2018-2020 - 1.0.0

15.3.3 Select Key (ID) column & export

Finally, a key or Animal ID column will need to be selected prior to export. Tick the preferred key column. You may also tick to include the dataset name as the first column of the export file. A "drop down" list on the Export button will allow you to select the export file format. Once the format is selected the file will be created by the database. When ready an email will be sent to the user. The email will include a link enabling the user to download the file via a web browser.


Phenobank | Selected Columns
phenobank.azurewebsites.net/Datasets/ExportStep3?selectedDatasets=Kamilaroi%2CNTDPI&selectedColumns=AnimalID%2CANL1%2CANL2
PHENOBANK
lipadmin@abri.une.edu.au
NAVIGATION
Dashboard
YOU ARE HERE: Select Key Column/s
SELECT KEY COLUMN/S (IF REQUIRED)
Export
Cancel
Search:
CSV
Tab delimited
Space delimited

Selected key column/s for export

☑ Include dataset name in file export

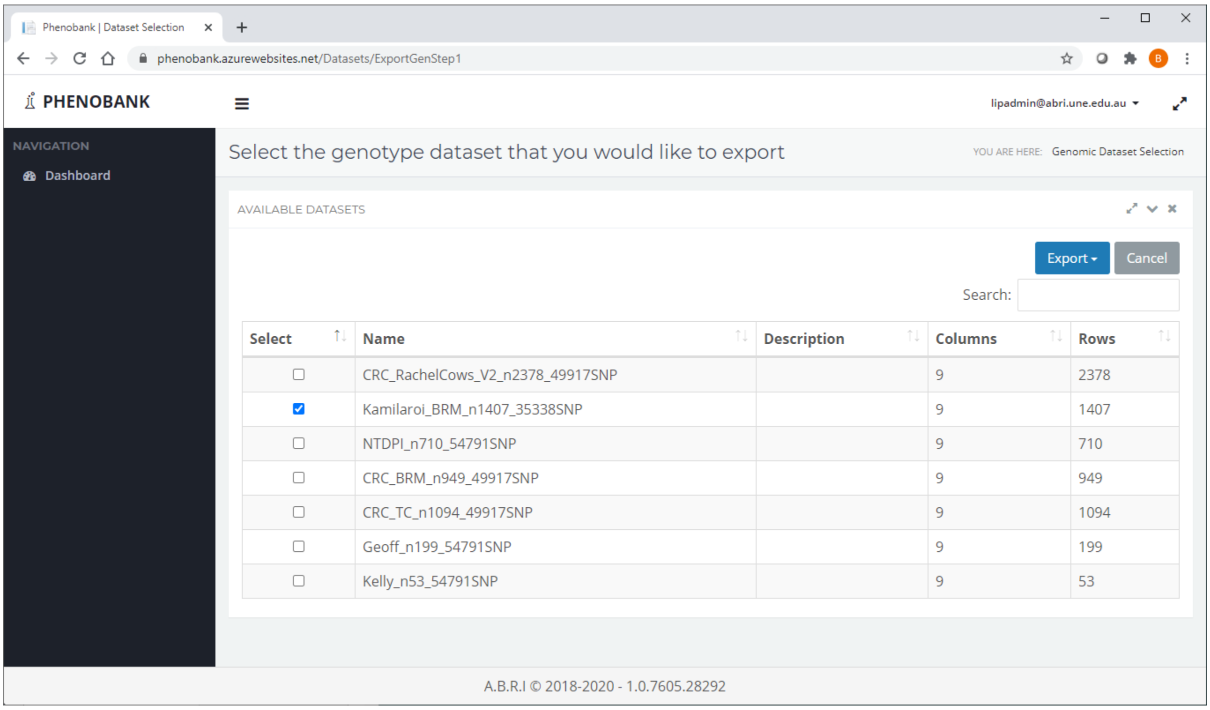
Key	Name	Display Name
☑	AnimalID	AnimalID
□	ANL1	ANL1
□	ANL2	ANL2

A.B.R.I © 2018-2020 - 1.0.0

15.4 Export Genomic Data

15.4.1 Export Genomic Functions

While at the Phenobank Dashboard, click the "Export genomic data" button. You will reach a page like the following:


Phenobank | Dataset Selection
phenobank.azurewebsites.net/Datasets/ExportGenStep1
PHENOBANK
lipadmin@abri.une.edu.au
NAVIGATION
Dashboard
YOU ARE HERE: Genomic Dataset Selection
AVAILABLE DATASETS
Export
Cancel
Search:

Select the genotype dataset that you would like to export

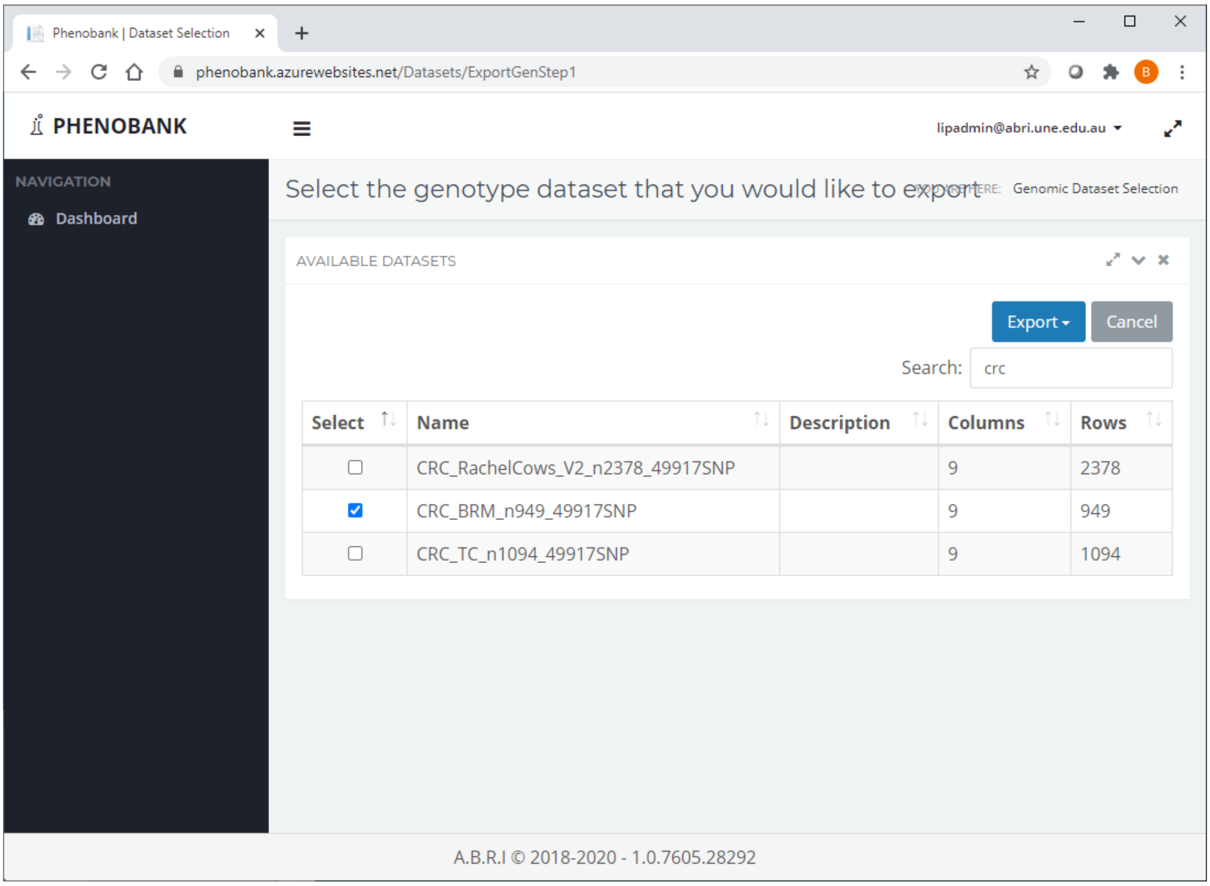
Select	Name	Description	Columns	Rows
□	CRC_RachelCows_V2_n2378_49917SNP		9	2378
☑	Kamilaroi_BRM_n1407_35338SNP		9	1407
□	NTDPI_n710_54791SNP		9	710
□	CRC_BRM_n949_49917SNP		9	949
□	CRC_TC_n1094_49917SNP		9	1094
□	Geoff_n199_54791SNP		9	199
□	Kelly_n53_54791SNP		9	53

A.B.R.I © 2018-2020 - 1.0.7605.28292

The names given will be those given for the genomic catalog item (GCI). A single GCI links together genotype, sample, phenotype and map datasets, along with the names of linkage columns.

15.4.2 Search and select a genomic dataset

The selection of datasets can be narrowed by entering text in the search box above the top right of the table. In the example below "CRC" has narrowed the search to three GCI's, one of which has been selected.



Phenobank | Dataset Selection
phenobank.azurewebsites.net/Datasets/ExportGenStep1
PHENOBANK
lipadmin@abri.une.edu.au
NAVIGATION
Dashboard

Select the genotype dataset that you would like to export

YOU ARE HERE: Genomic Dataset Selection
AVAILABLE DATASETS
Export
Cancel

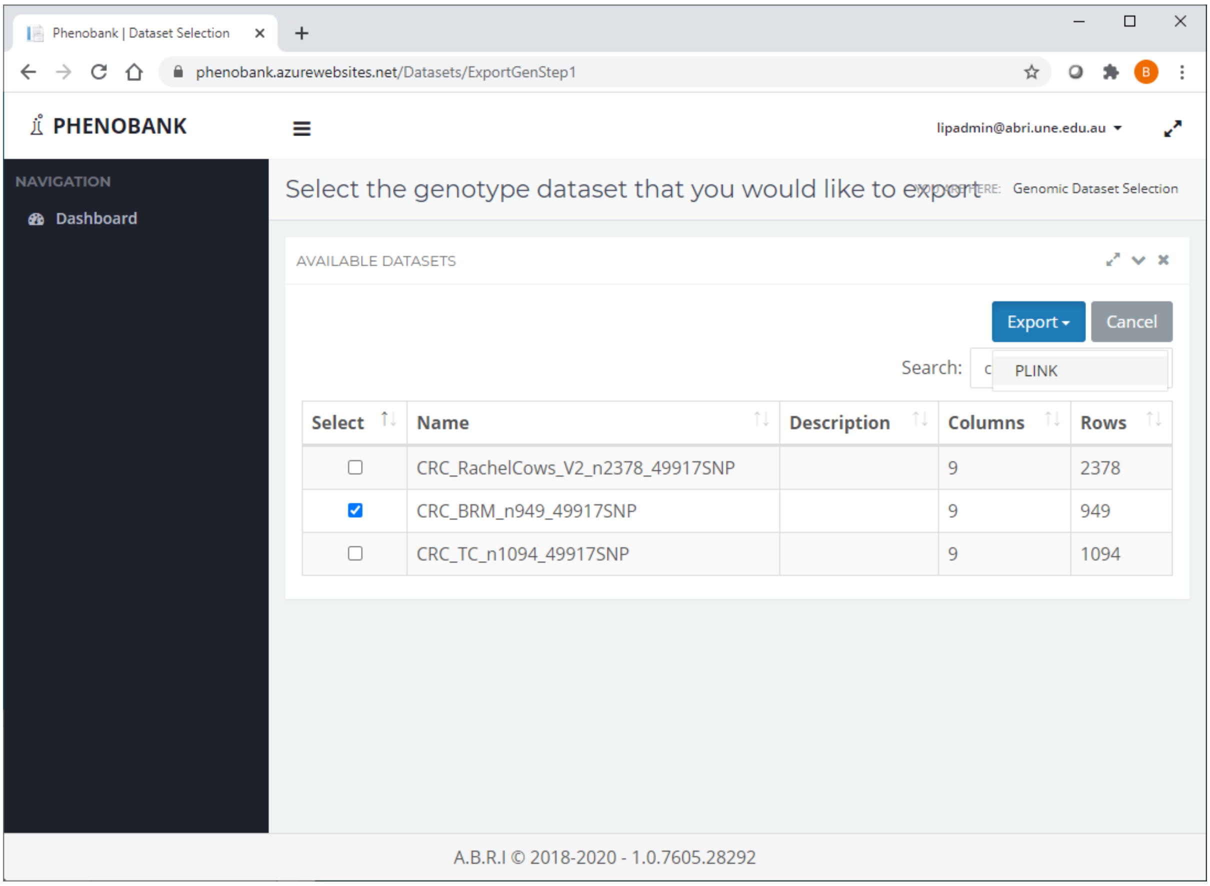
Search: crc

Select	Name	Description	Columns	Rows
☐	CRC_RachelCows_V2_n2378_49917SNP		9	2378
☑	CRC_BRM_n949_49917SNP		9	949
☐	CRC_TC_n1094_49917SNP		9	1094

A.B.R.I © 2018-2020 - 1.0.7605.28292

15.4.3 Exporting a genomic dataset

Click “Export” to see the drop down of available export file formats. In the following example, the “PLINK” format is available.


Phenobank | Dataset Selection
phenobank.azurewebsites.net/Datasets/ExportGenStep1
PHENOBANK
lipadmin@abri.une.edu.au
NAVIGATION
Dashboard

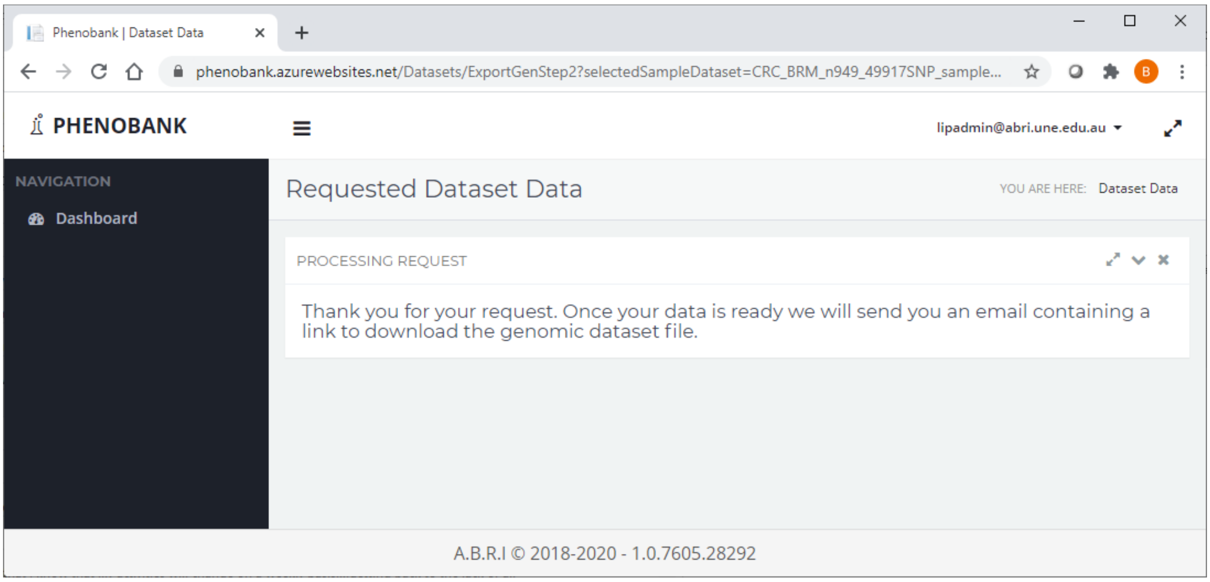
Select the genotype dataset that you would like to export

YOU ARE HERE: Genomic Dataset Selection
AVAILABLE DATASETS
Export
Cancel
Search:
PLINK

Select	Name	Description	Columns	Rows
☐	CRC_RachelCows_V2_n2378_49917SNP		9	2378
☑	CRC_BRM_n949_49917SNP		9	949
☐	CRC_TC_n1094_49917SNP		9	1094

A.B.R.I © 2018-2020 - 1.0.7605.28292

Click “PLINK” and the export files will be prepared. An email will be sent to you with links to a download location of the file, and you will be able to download the files (both the PLINK “ped” file and “map” file) via your browser. The time taken for your email to arrive will depend on a number of factors, however typically the files will be prepared in under 5 minutes for a 1,000 animal/50,000 SNP dataset.


Phenobank | Dataset Data
phenobank.azurewebsites.net/Datasets/ExportGenStep2?selectedSampleDataset=CRC_BRM_n949_49917SNP_sample...
PHENOBANK
lipadmin@abri.une.edu.au
NAVIGATION
Dashboard

Requested Dataset Data

YOU ARE HERE: Dataset Data
PROCESSING REQUEST

Thank you for your request. Once your data is ready we will send you an email containing a link to download the genomic dataset file.

A.B.R.I © 2018-2020 - 1.0.7605.28292

For further information

CSIRO Agriculture & Food
Bryce Little
+61 7 3214 2691
Bryce.little@csiro.au
csiro.au/agriculture

Agricultural Business Research Institute (ABRI)
Cody Wilson
+61 2 6773 3555
Cody.wilson@abri.une.edu.au
Abri.une.edu.au