


















ORIGINAL RESEARCH ARTICLE

Crop Breeding & Genetics

Building the Embrapa rice breeding dataset for efficient data reuse

Flavio Breseghello¹  | Raquel Neves de Mello¹  | Patrícia Valle Pinheiro¹  |
 Dino Magalhães Soares¹ | Sergio Lopes Júnior¹ | Paulo Hideo Nakano Rangel¹  |
 Elcio Perpétuo Guimarães¹ | Adriano Pereira de Castro¹  | José Manoel Colombari
 Filho¹  | Ariano Martins de Magalhães Júnior²  | Paulo Ricardo Reis Fagundes²  |
 Péricles de Carvalho Ferreira Neves¹  | Isabela Volpi Furtini¹  |
 Marley Marico Utumi³  | José Almeida Pereira⁴ | Antônio Carlos Centeno Cordeiro⁵  |
 Austrelino Silveira Filho⁶ | Guilherme Barbosa Abreu⁷  | Francisco Pereira de Moura
 Neto¹  | Julian Pietragalla⁸  | Mateo Vargas Hernández⁹  | José Crossa¹⁰ 

¹ Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, Brazil

² Embrapa Clima Temperado, Pelotas, RS, Brazil

³ Embrapa Rondônia, Vilhena, Rondonia, Brazil

⁴ Embrapa Meio-Norte, Teresina, PI, Brazil

⁵ Embrapa Roraima, Boa Vista, Roraima, Brazil

⁶ Embrapa Amazônia Oriental, Belém, PA, Brazil

⁷ Embrapa Cocais e Planícies Inundáveis, São Luís, MA, Brazil

⁸ Integrated Breeding Platform, Texcoco, Mexico, Mexico

⁹ Universidad Autónoma Chapingo, Texcoco, Mexico, Mexico

¹⁰ International Maize and Wheat Improvement Center, CIMMYT, Texcoco, Mexico, Mexico

Correspondence

Flavio Breseghello, Embrapa Arroz e Feijão,
 Rod. GO-462, km 12, Santo Antônio de
 Goiás, GO, Brazil. 75375-000.
 Email: flavio.breseghello@embrapa.br

Assigned to Associate Editor Stanley Omar
 Samonte.

Abstract

Embrapa has led breeding programs for irrigated and upland rice (*Oryza sativa* L.) since 1977, generating a large amount of pedigree and phenotypic data. However, there were no systematic standards for data recording nor long-term data preservation and reuse strategies. With the new aim of making data reuse practical, we recovered all data available and structured it into the Embrapa Rice Breeding Dataset (ERBD). In its current version, the ERBD includes 20,504 crosses involving 9,974 parents, the pedigrees of most of the 4,532 inbred lines that took part in advanced field trials, and phenotypic data from 2,711 field trials (1,118 irrigated, 1,593 upland trials), representing 226,458 field plots. Those trials were conducted over 38 years (1982–2019), in 247 locations, in latitudes ranging from 3°N to 33°S. Phenotypic traits included

Abbreviations: BMS, Breeding Management System; ERBD, Embrapa Rice Breeding Dataset; MET, multiple environment trials; VCU, Value for Cultivation and Use

© 2021 The Authors. Crop Science © 2021 Crop Science Society of America

grain yield, days to flowering, plant height, canopy lodging, and five important fungal diseases: leaf blast, panicle blast, brown spot, leaf scald, and grain discoloration. The total number of data points surpasses 1.27 million. Descriptive statistics were computed over the dataset, split by cropping systems (irrigated or upland). The mean heritability of grain yield was high for both systems, at around .7, whereas the mean coefficient of variation was 13.9% for irrigated trials and 18.7% for upland trials. The ERBD offers the possibility of conducting studies on different aspects of rice breeding and genetics, including genetic gain, G×E analysis, genome-wide association studies and genomic prediction.

1 | INTRODUCTION

The Brazilian Agricultural Research Corporation, Embrapa, started its rice (*Oryza sativa* L.) breeding program in 1975, gathering and evaluating germplasm of interest from domestic and international sources. Controlled crosses started in 1977 with the purpose of developing improved cultivars for all sites with relevant rice production in Brazil (Martínez et al., 2014). The program targets the two major rice systems in Brazil: irrigated lowland and rainfed upland rice. These two subprograms handle separate germplasm, except for common sources of disease resistance and some eventual cross-pool hybridizations.

Crosses are carried out using manual emasculation and pollination. Segregating families are advanced through a modified pedigree method, including the evaluation of yield of $F_{2:4}$ families. Fixed lines ($F_{5:7}$) are evaluated in preliminary yield trials in three to five locations. Selected lines advance to regional trials, and from those, top-performing materials are admitted in the multiple environment trials (MET) of value for cultivation and use (VCU).

Phenotypic data from MET, spanning a long period and representing a broad geographic region, constitute a valuable historical dataset. According to Mackay et al. (2011), “not to exploit such valuable historical datasets is wasteful”. Reanalyzing historical data can help increase the efficiency of the core program and to leverage genetic research connected to the germplasm of interest.

Historical datasets from plant breeding programs are naturally unbalanced since different materials are added and dropped every season, testing sites change, and plots or entire replicates are lost for different reasons. Unbalanced datasets pose problems for ordinary least squares models; however, the implementation of robust and efficient mixed model packages in free statistical programs like R (R Core Team, 2018) turned the task of computing variance components and best linear unbiased predictors (BLUPs) from this type of data more practical for breeding teams.

Some of the possible applications of historical datasets are (a) computing the genetic gain of the program for different traits and periods, subsidizing the ex-post analysis of different strategies used and projecting future breeding methods (Breseghello et al., 2011; Laidig et al., 2014; Streck et al., 2018); (b) modelling genotype × environment interactions, possibly including environmental covariates, improving the ability to predict germplasm performance (Heslot et al., 2014); (c) identifying mega-environments and optimizing resource allocation in MET (González-Barrios et al., 2019); (d) estimating the genetic value of genotypes in specific mega-environments (Piepho et al., 2008); (e) training models for genomic prediction based on subsets of the elite gene pool to reduce breeding cycle duration through genomic selection strategies (Gapare et al., 2018; Rutkoski et al., 2015); (f) guiding the search for useful germplasm from gene banks through genomic prediction (Jarquin et al., 2016); (g) performing genome-wide association analyses and studying quantitative trait loci (QTL) × environment interaction, using MET means as phenotypes (Migicovsky et al., 2016).

A dataset including thousands of genotypes gains additional value if their pedigrees are known. The numerator relationship matrix (usually denoted as A-matrix) derived from pedigrees can be used in mixed models to compute BLUPs more accurately (Piepho et al., 2008) and to increase the prediction accuracy of the genetic value of genotypes in the context of genomic selection (Pérez-Rodríguez et al., 2017). Additionally, knowledge of pedigrees helps the breeder make better decisions in all steps of the program, from planning crosses to releasing new cultivars for a target environment. Thus, connecting the genotypes in historical datasets through pedigrees expands the usefulness of the phenotypic data.

Having recognized the value of the historical data, the Embrapa rice breeding team undertook the task of joining all the historical records into a structured, computer-readable format which could be easily analyzed to extract useful information. In the past, the program did not use an information system or strict notation standards which resulted in

a myriad of variations in text and numeric records scattered into thousands of files. The objective of this article was to report the steps of the process of recovering and consolidating fragmented records into the Embrapa Rice Breeding Dataset (ERBD) and to describe the basic features of the ERBD as a resource for rice breeding and research.

2 | MATERIALS AND METHODS

2.1 | Collecting original data

The main sources of information used to develop the ERBD were (a) the book of crosses, a series of spreadsheets with cross records as free text with no consistent writing standards; (b) the book of inbred line coding, a series of spreadsheets connecting derivative names (which include cross codes) to inbred line codes; (c) a collection of computer files with data and metadata from individual field trials in .sas or .xls formats.

The first step for organizing the information was defining nomenclature standards for all the germplasm. Strict naming rules were defined for crosses, segregating families, inbred lines, and cultivars (Supplemental Table S1). Consolidating unique names for all genotypes was one of the most time-consuming tasks in the whole data recovery process. In the original records, the same genotype received several different names (“synonyms”) and several spellings for each name (with and/or without spaces, hyphens, underscores, special Portuguese characters, or abbreviations). Additionally, numerous typographical errors were found and corrected. The consolidated numbers of unique genotypes are presented in the Supplemental Table S2.

2.2 | Construction of the germplasm database

The book containing the records of more than 20,000 crosses made since the beginning of the breeding program was organized in a single spreadsheet, with the following columns: Cross Code, Female Parent, Male Parent, Year, Subprogram, Population, and Breeder. The first three columns were required to recover pedigrees, whereas the others were ancillary information. All crosses were coded in a biparental format (female/male). In double, triple or backcrosses, the code of the preceding cross was used as the parent name. All the parent names were revised for spelling consistency. In the case of synonyms, the most recent name was used according to the following hierarchy: cultivar name > line name > derivative name or introduction name.

The list of cross parents was extracted and imported into the germplasm database in Breeding Management System (BMS ,

Core Ideas

- Historical datasets are useful for genetic research and development of breeding tools.
- Phenotypic data from crop breeding programs connect elite genotypes with target environments.
- Converting available records in analyzable format is a worthwhile investment.

Integrated Breeding Platform, <https://integratedbreeding.net>). The chain of crosses (CNAX0001 to CNAX21572, from which 20,504 crosses are recorded) was imported into BMS. Whenever a parent was a cultivar or a coded line, it was necessary to use the information recorded in the book of inbred line codes to connect it to a previous cross. With this procedure, BMS computed pedigrees for all the germplasm, with a variable number of generations, depending on the data available.

2.3 | Construction of the phenotypic dataset

More than 3,000 computer files in text or spreadsheet format, corresponding to individual field trials, were recovered from the institutional mainframe, personal computers, and magnetic disks. Each file was read in SAS or Microsoft Excel and manipulated as necessary to comply with predefined standards. Trials with more than 25% of missing data for grain yield were discarded. Names and scales of variables were checked for coherence.

Metadata were collected from file headings and tabulated in a spreadsheet format where each line represented a trial. The design used, when not informed in the heading, was deducted from design factors. Experiments with only replicates or blocks were assigned a randomized complete block design, whereas experiments with replicates and blocks within replicates were assigned a lattice design.

No information about agronomic practices applied to trials was recovered. As a general rule, advanced yield trials were conducted under the usual crop management for the test location. Weeds and insect pests were controlled as needed, using chemical or mechanical methods, according to local availability. With few exceptions, no fungicides were used, letting the genotypes express their genetic resistance to the natural infection by common pathogens.

Revised files were converted into .csv files, with one row of column names, one row for each plot, and one column for each variable, without headings or captions, according to the “tidy data” concept (Wickham, 2014). Those files were read and stacked into a single dataset in the R environment (Supplemental Material).

2.4 | Dataset quality check

The grain yield of each individual trial was analyzed according to its experimental design, fitting a mixed model with genotypes as random factors. The broad sense heritability (H^2) and the coefficient of variation (CV) were computed for each trial using Meta-R (Alvarado et al., 2020). Only trials with $H^2 \geq .05$ and $CV \leq 50\%$ were kept in the dataset. The trial mean, CV and H^2 , along with the location and year, were used to identify and drop redundant trials in the dataset.

The assignment of metadata, like crop system, year, and trial type was done based on text headings of individual files. To check the accuracy of this classification, we used a cluster analysis based on the similarity of the set of genotypes tested in each trial and plotted the results in heat maps, allowing the visual identification of misclassified trials (Supplemental Materials). Several instances of wrong trial annotations were found and corrected. The summary of the classification of trials per year and location is presented in the Supplemental Table S3 and S4, respectively.

Extensive postcompilation revision was required. Extreme upper and lower values of grain yield, days to flowering, and plant height were inspected to detect obvious outliers. In case of detecting a single outlier in a trial, that data point was set as missing. When several values were out of the reasonable range for a trial, the variable was discarded. Grain yield was the only trait considered essential, such that, when this trait was not useful, the trial was discarded entirely. All the numbers presented in this paper refers to ERBD after quality check.

2.5 | Descriptive statistics and data storage

The R language was used to build and manipulate the phenotypic data in the ERBD dataset. Simple descriptive statistics were computed, separated by crop system, using the package ‘summarytools’. Central and dispersion parameters were computed for numeric traits (grain yield [GY], days to flowering [DTF], plant height [PHT]) and frequency distributions were computed for categorical traits scored on a scale from 1 to 9 (LOD, lodging; LBL, leaf blast; PBL, panicle blast; BSP, brown spot; LSC, leaf scald; GDS, grain discoloration). Pearson correlation coefficients between traits were computed using pairwise complete observations, separated by crop system, and plotted using the package ‘corrplot’. Variance components related to each factor were estimated for numeric traits with a joint analysis of VCU trials using the R package ‘lme4’ and considering all factors as random.

The full phenotypic dataset (named ERBD) was deposited in SIExp, Embrapa’s corporative information system for experimental data management. The germplasm database with genealogical relationships was built and is managed in BMS.

TABLE 1 Number of parents used in crosses and number of distinct crosses performed by each subprogram within the Embrapa rice breeding program from 1977 to July 2020

Subprogram	Number of parents	Number of crosses	Percentage of crosses
Upland	5,455	11,197	56.2%
Irrigated	3,617	6467	32.5%
Pre-breeding	324	973	4.9%
Special grain types	775	874	4.4%
Backcross	406	397	2.0%
Total	9,974	19,908	100%

3 | RESULTS

3.1 | From crosses to released cultivars

In the period from 1977 until July 2020, the Embrapa rice breeding program performed 20,504 crosses. From those, 596 were redundant or reciprocal crosses and 19,908 were distinct crosses (identical crosses performed in different subprograms were considered distinct), involving 9,974 parents (Table 1). Parents used in the irrigated and the upland subprograms were mostly different, although 325 parents participated in both subprograms. Out of the crosses performed, 56.2% were carried out for the upland rice program and 32.5% for the irrigated rice program. The remaining 11.3% of the crosses were done for prebreeding purposes to develop special grain types and for the introgression of traits through backcrosses.

Since the beginning of its breeding program, Embrapa released 45 improved cultivars of irrigated rice, 46 cultivars of upland rice, and five cultivars of special grain types. From those, 71 cultivars were derived from crosses conducted at Embrapa and 25 cultivars (13 irrigated, 11 upland, and 1 special) were derived from segregating families or inbred lines introduced from international institutions (21 from CIAT, 3 from CIRAD, and 1 from IRRI).

The BLUPs of cultivars for grain yield, days to flowering, and plant height were plotted against their year of release (Figure 1). The last cultivars from introduced elite germplasm were released in 2003 for upland and 2007 for irrigated rice. Herbicide-resistant cultivars entered the portfolio starting in 2009 for irrigated rice and 2015 for upland rice. The trend lines resulting from the linear regression of cultivar BLUPs on years indicated an upward trend for grain yield and a downward trend for days to flowering and plant height in the upland cultivars. This simple regression should be regarded as a first indication of genetic progress based on cultivars. A more detailed analysis of genetic gains, taking into consideration different cultivar types, target regions, and program phases is underway.

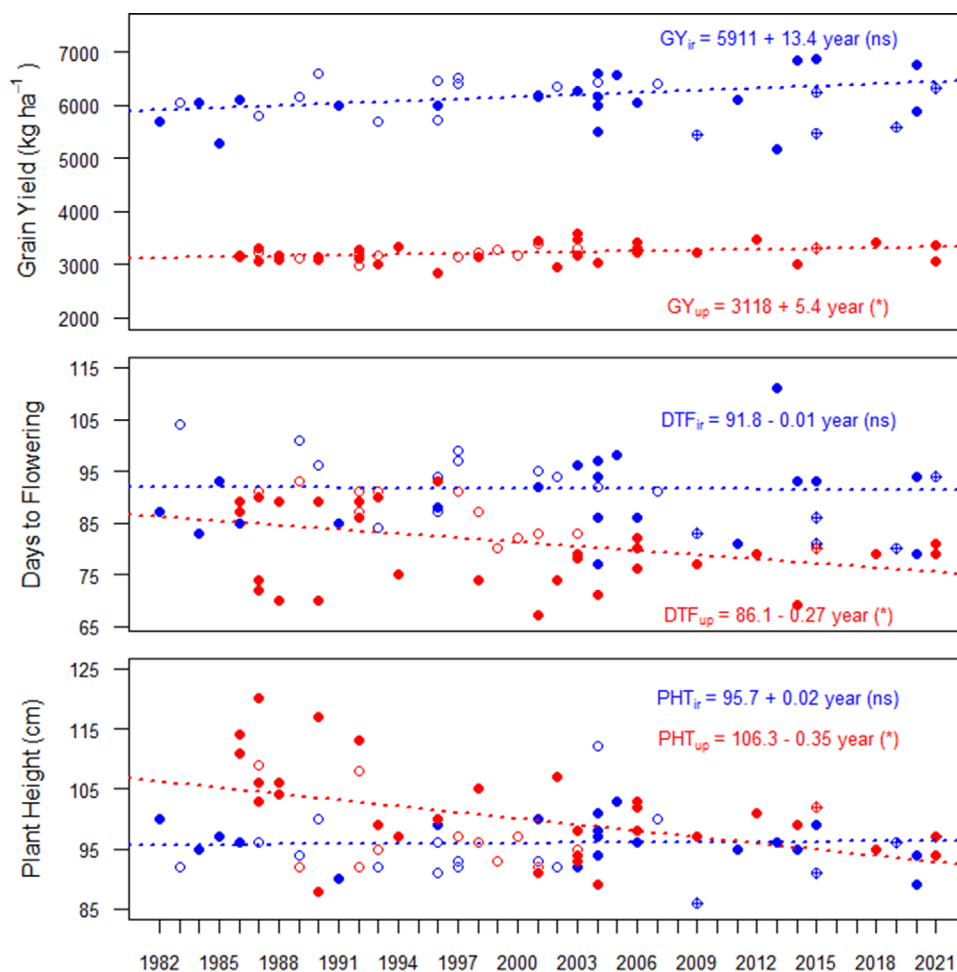


FIGURE 1 Cultivar best linear unbiased predictors (BLUPs) for grain yield (GY), days to flowering (DTF), and plant height (PHT) versus year of commercial release, from 1982 to 2021, for irrigated rice (blue) and upland rice (red). Open circles indicate cultivars released from external germplasm, filled circles indicate cultivars bred from crosses made in Brazil, and crossed circles indicate backcross-derived herbicide-resistant cultivars. Trend lines are based on linear regression of all cultivar BLUPs on year of release. * Significant at the .05 probability level; ns, nonsignificant

3.2 | Structure of ERBD

The ERBD includes the germplasm database with pedigrees, the metadata, and the phenotypic data from field trials. The metadata presented in Table 2 allow sorting trials and taking subsets of the dataset for specific meta-analyses. The primary metadata is the crop system, which splits the ERBD into two nearly independent datasets. Within those subsets, the metadata include trial type and design, year, and location, as well as some parameters of experimental quality for GY (mean, H^2 , and CV). The phenotypic dataset in the ERBD includes three experimental factors (REP, replication; BLO, block; GEN, genotype), three quantitative traits, and six traits recorded in a 1 to 9 rating scale (Table 3). The common variable TRIAL in Tables 2 and 3 allows metadata to be joined with data for analysis.

The dataset includes 2,711 field trials from 38 yr (1982–2019) of the irrigated rice program and 37 yr (1982–2018) of

the upland rice program (the COVID-19 lockdown impeded the harvest of the upland rice trials in the 2019–2020 season). From those trials, 1,118 were from the lowland irrigated breeding program and 1,593 from the upland rice program. Regarding trial type, the dataset includes 694 Regional trials and 2,017 VCU trials (Supplemental Table S3)

Grain yield was recorded in all trials, along with plant height and days to flowering in most cases. Plant lodging and diseases were evaluated according to their occurrence. The dataset contains 1.27 million data points (Table 4), from which 65.8% comes from the upland rice program. Those numbers reveal the historical importance of upland rice breeding in Brazil.

The geographic distribution of trials includes 247 locations (103 for irrigated rice and 185 for upland rice) in all five regions and 26 states of Brazil (Supplemental Table S4), spanning a latitude from 3° N (Boa Vista, RR) to 33° S (Santa Vitória do Palmar, RS). There was a larger number of upland

TABLE 2 Structure of the metadata related to field trials included in the Embrapa Rice Breeding Dataset

Nickname	Name	Details
TRIAL	Trial code	Unique string identifying the trial
SYST	Crop system	Indicates both the breeding subprogram and the trial environment. Levels: Irrigated or Upland
YEAR	Year of the trial	Year of trial preparation. e.g., 2005: season 2005/2006
DATE	Planting date	Day of planting dry seeds. Format DD/MM/YYYY
ST	State of Brazil	State of Brazil where the trial was conducted
LOCATION	Location of planting	Name of the municipality where the trial was conducted
LOC	Location of planting	Short tag indicating the municipality
TYPE	Trial type	Type of trial. ER, Regional Yield Trials; VCU, Value for Cultivation and Use (Advanced Yield Trials)
DESIGN	Experimental design	The statistical design of the trial. RCB, randomized complete block design; LAT, lattice design
MEAN	Grain yield mean	Trial grand mean of grain yield (kg ha ⁻¹)
H ²	Heritability	Broad-sense heritability of grain yield
CV	Coefficient of variation	Experimental coefficient of variation for grain yield (%)

TABLE 3 Structure of data related to field plots from trials in the Embrapa Rice Breeding Dataset, including sources of variation and phenotypic traits

Nickname	Name	Type	Details
TRIAL	Trial code	Link to metadata	Unique string identifying the trial
REP	Replicate number	Design Factor	Integer indicating the replicate within trial
BLO	Block number	Design Factor	Integer indicating the block within replicate (only in lattice design)
GEN	Genotype name	Experimental Factor	Identification of the germplasm (inbred line, landrace or cultivar)
GY	Grain yield	Numeric	Weight of paddy rice at 13% moisture (kg ha ⁻¹)
PHT	Plant height	Numeric	Height of the plant from the ground to the tip of the primary panicle, at pre-harvest stage, in cm
DTF	Days to flowering	Numeric	Number of days from planting dry seeds until 50% of the plants are flowered
LOD	Lodging	Scores 1 to 9 ^a	Level of lodging of the plot canopy, evaluated at pre-harvest stage
LBL	Leaf blast	Scores 1 to 9	Severity of the rice blast disease, caused by <i>Magnaporthe oryzae</i> , evaluated in leaves in the vegetative stage
PBL	Panicle blast	Scores 1 to 9	Severity of the rice blast disease, caused by <i>Magnaporthe oryzae</i> , evaluated in panicles in the pre-harvest stage
BSP	Brown spot	Scores 1 to 9	Severity of the disease caused by <i>Bipolaris oryzae</i> , evaluated on leaves in the preharvest stage
LSC	Leaf scald	Scores 1 to 9	Severity of the disease caused by <i>Monographella albescens</i> , evaluated in leaves in the preharvest stage
GDS	Grain discoloration	Scores 1 to 9	Severity of grain darkening or spots, caused by several fungi, evaluated on glumes in the preharvest stage

^aHigher scores indicate increasing levels of lodging or disease.

trials in the Central-West Region and of irrigated trials in the South Region which reflects the predominant rice cropping systems in those environments. In the other regions, there was a balance between the two cropping systems. The State of Goiás (GO) concentrates the largest number of trials because it harbors the base of the program conducted at Embrapa Rice and Beans.

The total number of genotypes evaluated in the period between 1982 and 2019 was 4,532 (Supplemental Table

S2), out of which 2,615 and 2,138 genotypes were evaluated under irrigated rice and upland rice conditions, respectively (221 genotypes participated in trials in both crop systems). In irrigated rice, 54.6% of the genotypes evaluated in VCU trials came from regional trials, whereas in upland rice, this proportion was 66.1%. This difference shows that the irrigated trials received more introductions from other breeding programs than the upland program.

TABLE 4 Number of trials, locations, years, and total data points for grain yield (GY), days to flowering (DTF), plant height (PHT), and scores for lodging (LOD), leaf blast (LBL), panicle blast (PBL), brown spot (BSP), leaf scald (LSC), and grain discoloration (GDS), from irrigated and upland rice trials in the Embrapa Rice Breeding Dataset, from 1982 to 2019

Traits	Irrigated				Upland			
	Trials	Locations	Years	Data points	Trials	Locations	Years	Data points
GY	1,118	103	38	91,127	1,593	185	37	133,045
DTF	924	90	38	75,612	1,346	163	37	113,358
PHT	991	92	38	80,111	1,505	172	37	125,870
LOD	443	61	38	35,973	965	141	37	81,715
LBL	210	31	36	17,267	713	100	37	58,575
PBL	361	36	36	30,174	914	119	36	75,956
BSP	434	41	36	36,222	969	134	37	85,064
LSC	415	44	36	35,167	947	126	36	83,660
GDS	400	45	33	33,320	931	133	36	80,023
Total				434,973				837,266
Grand total								1,272,239

3.3 | Results of the analysis of variance

The experimental quality of individual trials was assessed with an analysis of variance of grain yield and computing the heritability (H^2) and coefficient of variation (CV) for each trial. Trials with $H^2 < .05$ or $CV > 50\%$ were considered of poor quality and discarded. The mean CV was 13.9 and 18.7% for irrigated and upland rice trials, respectively, indicating a higher experimental quality for the irrigated conditions. Nevertheless, H^2 was high for both systems, at .71 and .69 for irrigated and upland, respectively. This result indicates that a broader genetic variability among upland rice genotypes may have compensated for a lower experimental precision, inherent to the rainfed environment.

The H^2 and CV were compared among the trials in the ERBD. The CV is easily computed from the analysis of variance by fixed models ($CV = [\sigma_e / \mu] 100$) and is influenced by the experimental error, trial management, and environmental quality. Heritability can be computed from mixed models ($H^2 = \sigma_g^2 / \sigma_p^2$), with genotypes as random effects, and has the advantage of considering the genetic variance in its formula and not being affected by the trial mean. The correlation between H^2 and CV was $-.46$ for irrigated trials and $-.36$ for upland trials, both highly significant. However, the scatterplot in Figure 2 shows a large spread of the data, indicating that trials with high CV can have high H^2 and vice versa. Therefore, the use of both statistics is advisable since they convey different and complementary information.

The overall importance of variance components related to sources of variation for grain yield, days to flowering, and plant height were estimated by the joint analysis of VCU trials, in random effect models (Table 5). The interaction genotype \times year was the most important factor for grain yield, both in irrigated and upland rice, which highlights the impor-

tance of multiyear field evaluation of elite materials for deciding on commercial release. The second most important factor for grain yield was location for irrigated rice, where trials were conducted in well-defined environments with irrigation infrastructure, such that location captures most of the variance between trials. In upland rice, trial was more relevant than location since this factor accounts for a wide variation in field management and the random variant of rainfall. Days to flowering and plant height were mostly accounted for by genotype, location, and the interaction genotype \times year. Effects of interactions involving trials were computationally challenging due to extreme lack of balance in the dataset and were not estimated. Replicate within trial was of lesser importance for all traits.

3.4 | Data distribution

The descriptive statistics for the quantitative traits in ERBD (Table 6) show that the mean grain yield is almost twice as high in irrigated rice than in upland rice. Additionally, the upland rice genotypes flowered 8.1 d earlier and were 6.7-cm taller than the irrigated rice genotypes, on average. The histograms (Figure 3) show that the three traits approximate a normal distribution. Nevertheless, grain yield in upland trials presents higher skewness and kurtosis than in irrigated rice, probably due to more challenging growth conditions, which compresses the mean downwards.

The categorical traits in ERBD consist of scores of canopy lodging, plus five common diseases in the rice crop in Brazil. Those traits were visually evaluated with a scale from 1 to 9, where 1 indicates the absence of lodging or disease symptoms and 9 indicates their maximum severity. There is a predominance of lower scores in the dataset, as expected for elite

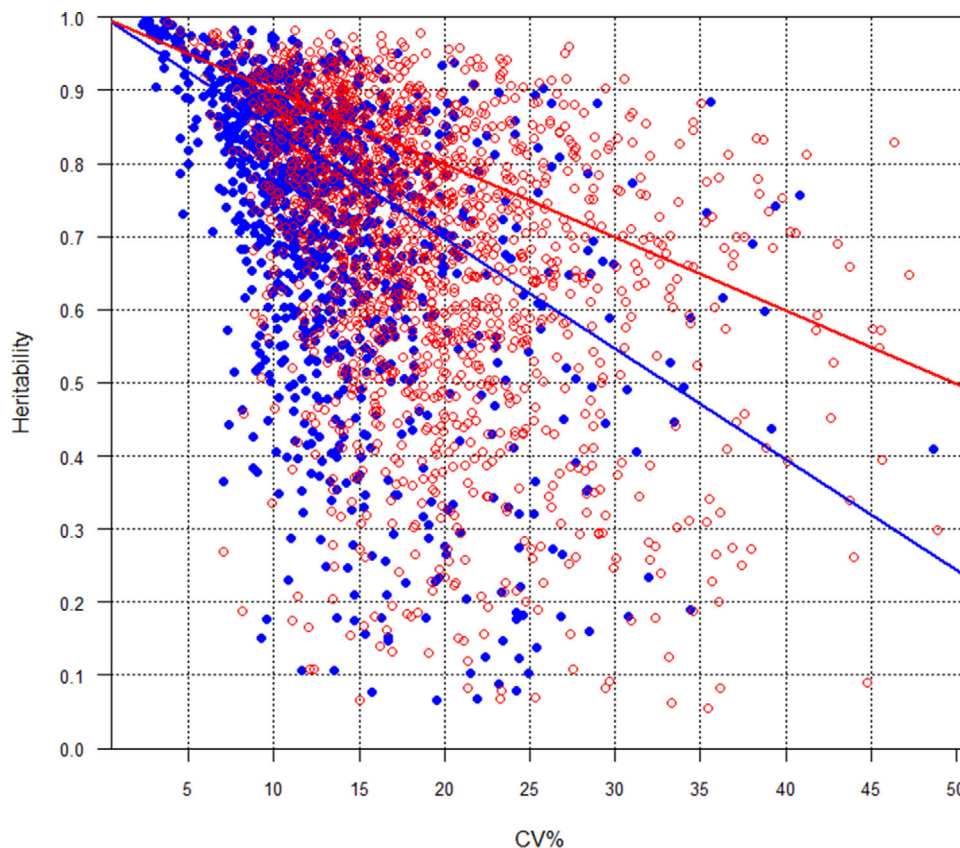


FIGURE 2 Scatterplot of heritability (H^2) versus coefficient of variation (CV) for rice trials conducted in irrigated (blue dots) and upland conditions (open red circles) in the Embrapa Rice Breeding Dataset. Straight lines indicate a regression of H^2 on CV, with the intercept set to (0, 1)

germplasm (Figure 4). The relative frequency of scores for each trait is very similar for irrigated and upland rice, showing that those stresses are equally important for both systems. Odd scores are more frequent than even scores, since odd scores are formally described in reference manuals for the evaluation of rice (International Rice Research Institute, 2002). Lodging scores are mostly low, which may be due to a combination of germplasm resistance and environments that are nonconducive to lodging.

Trait correlations were tested based on all pairwise complete observations in the irrigated and upland datasets, separately. All correlations were highly significant ($p < .01$), except in irrigated rice, leaf blast \times lodging ($p = .013$), and in upland rice, no correlation was found for brown spot versus days to flowering, plant height, or lodging, for leaf scald versus plant height, or for panicle blast versus lodging (Figure 5). Diseases were positively correlated among them and negatively correlated with grain yield, especially in irrigated trials. Days to flowering presented negative correlations with grain yield, favoring the selection of early flowering genotypes without penalty on harvest. Those correlations are favorable to breeding since they may lead to desirable correlated responses to selection. On the other hand, plant height is positively correlated with grain yield and lodging, espe-

cially in upland rice, which poses a challenge for breeding high-yielding, lodge-resistant cultivars.

4 | DISCUSSION

The recovery of historical data is a one-time task with long-term effects on the program. The cumbersome task of gathering scattered, unstructured, and nonstandardized data was the first step towards reusing the information generated and preserved by the Embrapa rice breeding program through its nearly four decades of operations. Once this task is completed, researchers can focus on clues for interesting scientific questions from this dataset, instead of spending most of their time collecting and cleaning data (Wickham, 2014). We expect this dataset to expand through time, adding more years, locations, genotypes, and traits. With the adoption of an information system for the whole program (e.g., BMS), this expansion should occur naturally.

The dataset is composed of two almost independent subsets, corresponding to the upland and irrigated rice breeding programs. The upland rice subset represents approximately two-thirds of the phenotypic data points (Table 4), which is a consequence of Embrapa's strong focus on this crop system

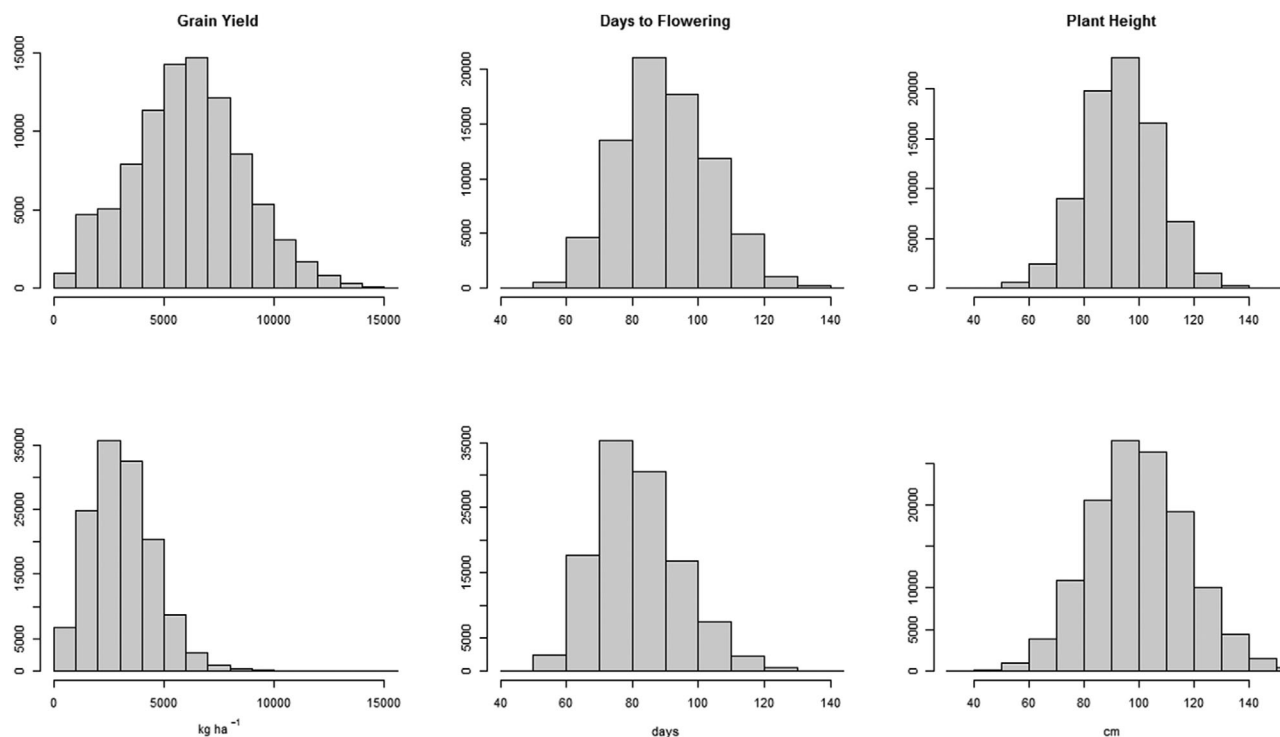


FIGURE 3 Histograms of grain yield (GY), days to flowering (DTF), and plant height (PHT) in the Embrapa Rice Breeding Dataset, for irrigated rice (upper row) and upland rice field trials (lower row), from 1982 to 2019

TABLE 5 Variance components of grain yield (GY), days to flowering (DTF), and plant height (PHT) related to main effects and two-way interactions between genotypes, years, and locations, plus trials and replicates within trial, estimated from random-model analyses of all irrigated and upland value for cultivation and use trials in the Embrapa Rice Breeding Dataset, from 1982 to 2019. Due to computational limitations, three-way interactions and interactions involving trials were not included in the model, adding to the residual variance

Factor	Irrigated			Upland		
	GY	DTF	PHT	GY	DTF	PHT
Genotype (G)	117,455	54.72	34.38	27,172	61.66	71.77
Year (Y)	529,945	2.04	16.71	277,005	3.76	8.70
Location (L)	1,577,102	63.60	46.97	103,218	54.75	72.47
G × Y	1,794,240	25.23	39.76	599,222	22.01	106.52
G × L	155,501	8.42	5.73	32,053	2.59	8.01
Y × L	347,701	14.93	11.06	161,891	12.12	20.88
Trial	1,044,298	28.49	30.01	479,499	9.06	41.69
Replicate	105,390	0.32	3.94	46,049	0.53	5.85
Residual	854,922	10.28	23.39	426,824	8.52	48.77

in the 20th Century, as reported by Martínez et al. (2014) and Breseghello et al. (2011). In the period since the year 2000 there is a near balance between the two systems (Supplemental Table S3).

An important feature of the ERBD is the partial replication of genotypes through years and locations which confers

TABLE 6 Descriptive statistics of the quantitative traits grain yield (GY), days to flowering (DTF), and plant height (PHT), in irrigated and upland field trials in the Embrapa Rice Breeding Dataset, from 1982 to 2019

Parameter	Irrigated			Upland		
	GY	DTF	PHT	GY	DTF	PHT
	kg ha ⁻¹ d		cm	kg ha ⁻¹ d		cm
No. Valid data	91,127	75,612	80,111	133,045	113,358	125,870
% Valid data	99.0	82.1	87.0	99.0	84.4	93.7
Mean	6,117.6	90.5	94.2	3,112.0	82.4	100.9
Std. Dev.	2,545.0	13.9	13.6	1,432.7	12.8	17.9
Skewness	0.188	0.225	0.054	0.620	0.561	0.191
Kurtosis	-0.058	-0.164	0.577	0.658	0.252	0.226
Minimum	252	42	32	109	42	30
Quartile 1	4,406	81	85	2,063	73	89
Median	6,095	90	94	2,996	81	100
Quartile 3	7,750	100	103	4,010	90	113
Maximum	16,774	151	210	14,142	142	221

connectivity of data within cropping systems. This property allows meta-analyses encompassing sections of environments or time intervals, or even the whole dataset, using appropriate mixed models (Arief et al., 2015). Aside from its internal connectivity of phenotypic data, the ERBD can be connected with other types of data, extending its applicability for novel statistical and genetics analyses.

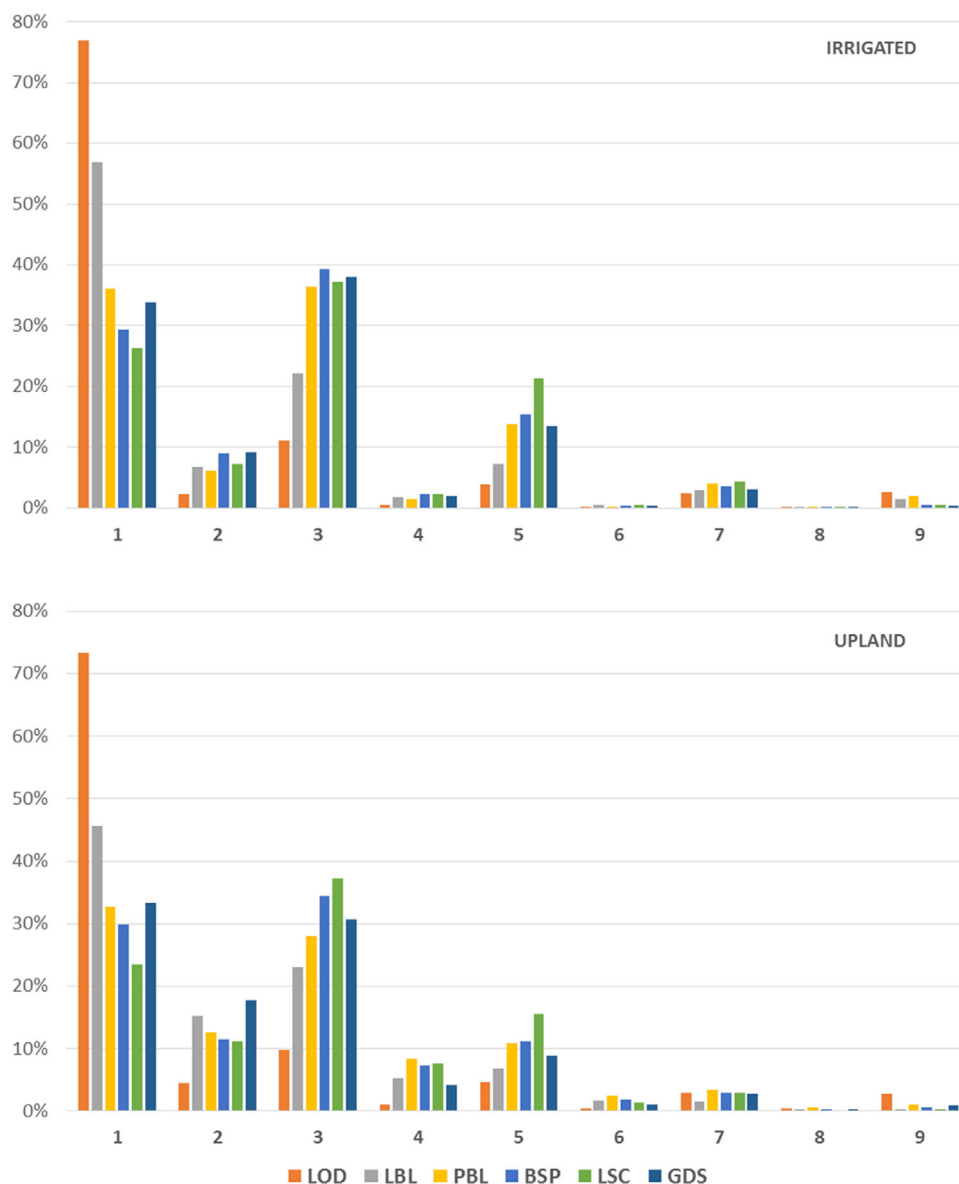


FIGURE 4 Relative frequency of scores of lodging (LOD) and diseases (LBL, leaf blast; PBL, panicle blast; BSP, brown spot; LSC, leaf scald; GDS, grain discoloration) in the Embrapa Rice Breeding Dataset, from irrigated and upland field trials, from 1982 to 2019

4.1 | Aggregating pedigree information

We connected all the information available in cross books and line coding books to build pedigrees as deep as possible, using the BMS software as the information platform. Although we made great progress since the free-text records, minor corrections are still being made. With this resource, we expect to be able to identify all the founding genotypes of the program and quantify its contribution to the current germplasm.

Once full pedigrees are available, a matrix of coefficients of parentage can be derived for any set of genotypes (Pérez-Rodríguez et al., 2017). Piepho et al. (2008) showed that using the A-matrix (derived from coefficients of parentage) in mixed models helps to estimate genotype breeding values through BLUPs with higher accuracy. The A-matrix also

increases the accuracy of genomic prediction based on models trained on samples of the historical germplasm (Pérez-Rodríguez et al., 2017).

The bridge built by pedigrees allows genotypes to borrow information from relatives (Dawson et al., 2013). In the context of the ERBD, this approach could provide phenotypic predictions for genotypes in locations where they have not been tested, thus saving resources and increasing the overall efficiency of the program.

4.2 | Studying genotype × environment interaction

The variance components in Table 5 show that year, location, and trial have a strong effect on rice grain yield. This

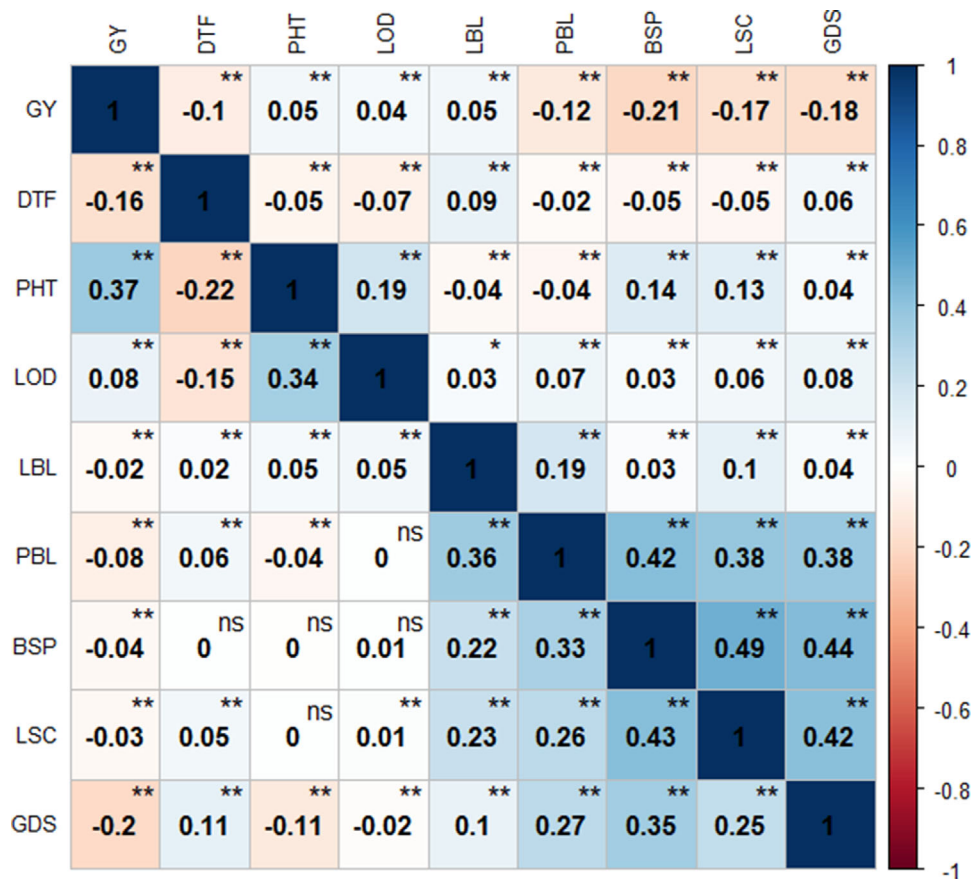


FIGURE 5 Pearson correlation coefficients between traits (GY, grain yield; DTF, days to flowering; PHT, plant height; LOD, lodging; LBL, leaf blast; PBL, panicle blast; BSP, brown spot; LSC, leaf scald; GDS, grain discoloration) in the Embrapa Rice Breeding Dataset, for irrigated rice trials (upper right) and upland rice trials (lower left), from 1982 to 2019. *Significant at the .05 probability level; **Significant at the .01 probability level; ns, nonsignificant

result highlights the need to consider genotype \times environment (G \times E) interaction when evaluating advanced breeding materials over large geographical areas, as in the case of Brazil. The ERBD offers opportunities for G \times E analyses between selection sites and rice production sites which can be highly informative for the improvement of resource allocation and genetic gain in target environments (González-Barrios et al., 2019).

Considering that a large proportion of the phenotypic variance is explained by the environment, it is worth exploring G \times E interaction in more detail by incorporating environmental covariates. The study of the reaction norm of germplasm to environmental conditions could lead to a deeper understanding of germplasm adaptability to normal or extreme weather conditions (Heslot et al., 2014; Morais Júnior et al., 2018).

Although the year of planting and days to flowering were recorded for most trials, exact planting dates (metadata “DATE” in Table 2) were not systematically recorded and are available for only 37% of the trials in ERBD. The lack of planting dates hinders the connection between phenotypic data and climatic data at a daily level. Nevertheless, it is still possible to

use environmental covariates at the season level, considering usual planting dates.

4.3 | Aggregating molecular data

The ERBD provides phenotypic data for relevant rice germplasm, connected by pedigrees, in target environments across Brazil. Adding molecular marker data to ERBD will open new possibilities for marker-aided breeding. The selection of priority germplasm for genotyping can be done based on the importance of the genotype within ERBD. Entries with more data, in general, should be prioritized for genotyping since they have phenotypic values estimated with more precision and sampled over a larger environmental variation. Priority materials should include founding genotypes, influential genotypes with a large number of progenies in the program, sources of disease resistance, and released cultivars.

Germplasm with genotypic and phenotypic data can be used as training populations for genomic prediction models to be applied on the current germplasm (Gapare et al., 2018).

Nevertheless, Rutkoski et al. (2015) highlight the importance of the training population being genetically close to the testing population. The parentage between the training and testing populations can be inferred from pedigrees.

The ERBD phenotypes can be coupled with high-density marker data for gene discovery through genome-wide association studies (GWAS; Ogonna et al., 2020). The GWAS panel can be selected based on ERBD data, aiming at maximizing the quality of phenotypic data and minimizing the within-panel population structure (Bressegello & Sorrells, 2006). Genes and alleles detected through GWAS in a panel extracted from ERBD are expected to be relevant for the current elite germplasm. Diagnostic markers for major genes, novel or not, could also be tracked in the breeding population all the way to released cultivars.

5 | CONCLUSION

The ERBD is a valuable dataset comprising pedigree information and phenotypic data over a wide geographical area and a long timeframe which offers many options to perform genetic research and increasing the efficiency of the breeding program. We hope this article could highlight the need to implement systematic data management in plant breeding and inspire other public breeding programs to undertake the task of data recovery and reuse. The ERBD can be shared with partners under specific agreements for scientific collaboration in studies of common interest.

ACKNOWLEDGMENTS

The authors acknowledge the work of all collaborators from partner institutions and the support staff that contributed to the conduction of the field trials and post-harvest processing. They also recognize the help of Dr. Fernando Toledo, from CIMMYT – BSU, with the R codes.

Most of the data in the ERBD was generated under the coordination of Dr. Orlando Peixoto de Moraes, who worked abreast with this team until the last days of his life.

AUTHOR CONTRIBUTION

Flavio Bressegello: Conceptualization; Data curation; Formal analysis; Writing-original draft. Raquel Neves de Mello: Data curation. Patrícia Valle Pinheiro: Data curation. Dino Magalhães Soares: Data curation. Sergio Lopes Júnior: Software. Paulo Hideo Nakano Rangel: Investigation; Resources. Elcio Perpétuo Guimarães: Investigation. Adriano Pereira de Castro: Investigation; Resources. José Manoel Colombari Filho: Investigation; Resources. Ariano Martins de Magalhães Júnior: Investigation. Paulo Ricardo Reis Fagundes: Investigation. Pércles de Carvalho Ferreira Neves: Investigation. Isabela Volpi Furtini: Investigation. Marley Marico Utumi: Investigation. José Almeida Pereira: Investigation. Antônio

Carlos Centeno Cordeiro: Investigation. Austrelino Silveira Filho: Investigation. Guilherme Barbosa Abreu: Investigation. Francisco Pereira de Moura Neto: Investigation. Julian Pietragalla: Software. Mateo Vargas Hernández: Data curation. José Crossa: Methodology

DATA AVAILABILITY STATEMENT

The ERBD data is available for collaborations between Embrapa and partners under agreement.

ORCID

Flavio Bressegello  <https://orcid.org/0000-0002-5476-7173>

Raquel Neves de Mello  <https://orcid.org/0000-0002-3647-4967>

Patrícia Valle Pinheiro  <https://orcid.org/0000-0003-0461-1821>


Paulo Hideo Nakano Rangel  <https://orcid.org/0000-0002-5741-3426>

Adriano Pereira de Castro  <https://orcid.org/0000-0002-6202-4767>

José Manoel Colombari Filho  <https://orcid.org/0000-0002-3143-8924>


Ariano Martins de Magalhães Júnior  <https://orcid.org/0000-0002-0756-4648>

Paulo Ricardo Reis Fagundes  <https://orcid.org/0000-0001-5294-2380>

Pércles de Carvalho Ferreira Neves  <https://orcid.org/0000-0002-4794-8953>

Isabela Volpi Furtini  <https://orcid.org/0000-0002-4932-9937>

Marley Marico Utumi  <https://orcid.org/0000-0002-4940-9363>

Antônio Carlos Centeno Cordeiro  <https://orcid.org/0000-0002-8197-2439>

Guilherme Barbosa Abreu  <https://orcid.org/0000-0003-3536-0527>

Francisco Pereira de Moura Neto  <https://orcid.org/0000-0002-1056-4438>

Julian Pietragalla  <https://orcid.org/0000-0001-9911-0960>

Mateo Vargas Hernández  <https://orcid.org/0000-0002-0735-3242>

José Crossa  <https://orcid.org/0000-0001-9429-5855>

REFERENCES

- Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., Pérez-Rodríguez, P., & Lopez-Cruz, M. A. (2020). META-R: A software to analyze data from multi-environment plant breeding trials. *The Crop Journal*, 8 (5), 745–756. <https://doi.org/10.1016/j.cj.2020.03.010>
- Arief, V. N., Delacy, I. H., Crossa, J., Payne, T., Singh, R., Braun, H. - J., Tian, T., Basford, K. E., & Dieters, M. J. (2015). Evaluating testing strategies for plant breeding field trials: Redesigning a

- CIMMYT international wheat nursery. *Crop Science*, 55(1), 164–177. <https://doi.org/10.2135/cropsci2014.06.0415>
- Breseghello, F., De Moraes, O. P., Pinheiro, P. V., Silva, A. C. S., Da Maia De Castro, E., Guimarães, É. P., De Castro, A. P., Pereira, J. A., De Matos Lopes, A., Utumi, M. M., & De Oliveira, J. P. (2011). Results of 25 years of upland rice breeding in Brazil. *Crop Science*, 51(3), 914–923. <https://doi.org/10.2135/cropsci2010.06.0325>
- Breseghello, F., & Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics*, 172(2), 1165–1177. <https://doi.org/10.1534/genetics.105.044586>
- Dawson, J. C., Endelman, J. B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., Manès, Y., Sorrells, M. E., & Jannink, J. L. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research*, 154, 12–22. <https://doi.org/10.1016/j.fcr.2013.07.020>
- Gapare, W., Liu, S., Conaty, W., Zhu, Q. H., Gillespie, V., Llewellyn, D., Stiller, W., & Wilson, I. (2018). Historical datasets support genomic selection models for the prediction of cotton fiber quality phenotypes across multiple environments. *G3: Genes, Genomes, Genetics*, 8, 1721–1732. <https://doi.org/10.1534/g3.118.200140>
- González-Barríos, P., Díaz-García, L., & Gutiérrez, L. (2019). Mega-environmental design: Using genotype × environment interaction to optimize resources for cultivar testing. *Crop Science*, 59(5). <https://doi.org/10.2135/cropsci2018.11.0692>
- Heslot, N., Akdemir, D., Sorrells, M. E., & Jannink, J. L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, 127(2), 463–480. <https://doi.org/10.1007/s00122-013-2231-5>
- International Rice Research Institute (2002). *Standard evaluation system for rice (SES)*. International Rice Research Institute.
- Jarquín, D., Specht, J., & Lorenz, A. (2016). Prospects of genomic prediction in the USDA soybean germplasm collection: Historical data creates robust models for enhancing selection of accessions. *G3: Genes, Genomes, Genetics*, 6(8), 2329–2341. <https://doi.org/10.1534/g3.116.031443>
- Laidig, F., Piepho, H. P., Drobek, T., & Meyer, U. (2014). Genetic and non-genetic long-term trends of 12 different crops in German official variety performance trials and on-farm yield trends. *Theoretical and Applied Genetics*, 127(12), 2599–2617. <https://doi.org/10.1007/s00122-014-2402-z>
- Mackay, I., Horwell, A., Garner, J., White, J., McKee, J., & Philpott, H. (2011). Reanalyses of the historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. *Theoretical and Applied Genetics*, 122(1), 225–238. <https://doi.org/10.1007/s00122-010-1438-y>
- Martínez, C. P., Torres, E. A., Chatel, M., Mosquera, G., Duitama, J., Ishitani, M., Selvaraj, M., Dedicova, B., Tohme, J., Grenier, C., Lorieux, M., Cruz, M., Berrió, L., Corredor, E., de San Martín, G. Z., Breseghello, F., Peixoto, O., Colombari Filho, J. M., de Castro, A. P., ... & Bruzzone, C. B. (2014). Rice breeding in Latin America. *Plant Breeding Reviews*, 38. <https://doi.org/10.1002/9781118916865.ch05>
- Migicovsky, Z., Gardner, K. M., Money, D., Sawler, J., Bloom, J. S., Moffett, P., Chao, C. T., Schwaninger, H., Fazio, G., Zhong, G. Y., & Myles, S. (2016). Genome to phenome mapping in apple using historical data. *The Plant Genome*, 9(2). <https://doi.org/10.3835/plantgenome2015.11.0113>
- Morais Júnior, O. P., Duarte, J. B., Breseghello, F., Coelho, A. S. G., Morais, O. P., & Magalhães Júnior, A. M. (2018). Single-step reaction norm models for genomic prediction in multi-environment recurrent selection trials. *Crop Science*, 58(2), 592–607. <https://doi.org/10.2135/cropsci2017.06.0366>
- Ogbonna, A. C., Braatz De Andrade, L. R., Rabbi, I. Y., Mueller, L. A., Jorge De Oliveira, E., & Bauchet, G. J. (2020). Large-scale GWAS using historical data identifies a conserved genetic architecture of cyanogenic glucosides content in cassava (*Manihot esculenta* Crantz.) root. *The Plant Journal*, 105(3), 754–770. <https://doi.org/10.1111/tj.15071>
- Pérez-Rodríguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., Legarra, A., Autrique, E., Campos, G. D. L., Burgueño, J., & Dreisigacker, S. (2017). Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. *The Plant Genome*, 10(2). <https://doi.org/10.3835/plantgenome2016.09.0089>
- Piepho, H. P., Möhring, J., Melchinger, A. E., & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1–2), 209–228. <https://doi.org/10.1007/s10681-007-9449-8>
- R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., & Sorrells, M. E. (2015). Efficient use of historical data for genomic selection: A case study of stem rust resistance in wheat. *The Plant Genome*, 8(1). <https://doi.org/10.3835/plantgenome2014.09.0046>
- Streck, E. A., De Magalhães, A. M., Aguiar, G. A., Henrique Facchinello, P. K., Reis Fagundes, P. R., Franco, D. F., Nardino, M., & De Oliveira, A. C. (2018). Genetic progress in 45 years of irrigated rice breeding in Southern Brazil. *Crop Science*, 58(3), 1094–1105. <https://doi.org/10.2135/cropsci2017.06.0383>
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1–23. <https://doi.org/10.18637/jss.v059.i10>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Breseghello F, de Mello R N, Pinheiro P V, Soares D M, Lopes Júnior S, Nakano Rangel P H, Guimarães E P, de Castro A P, Colombari Filho J M, de Magalhães Júnior A M, Fagundes P R R, de Carvalho Ferreira Neves P, Furtini I V, Utumi M M, Pereira J A, Cordeiro A C C, Filho A S, Abreu G B, de Moura Neto F P, ... Crossa J. Building the Embrapa rice breeding dataset for efficient data reuse. *Crop Science*. 2021; 1–13. <https://doi.org/10.1002/csc.2.20550>