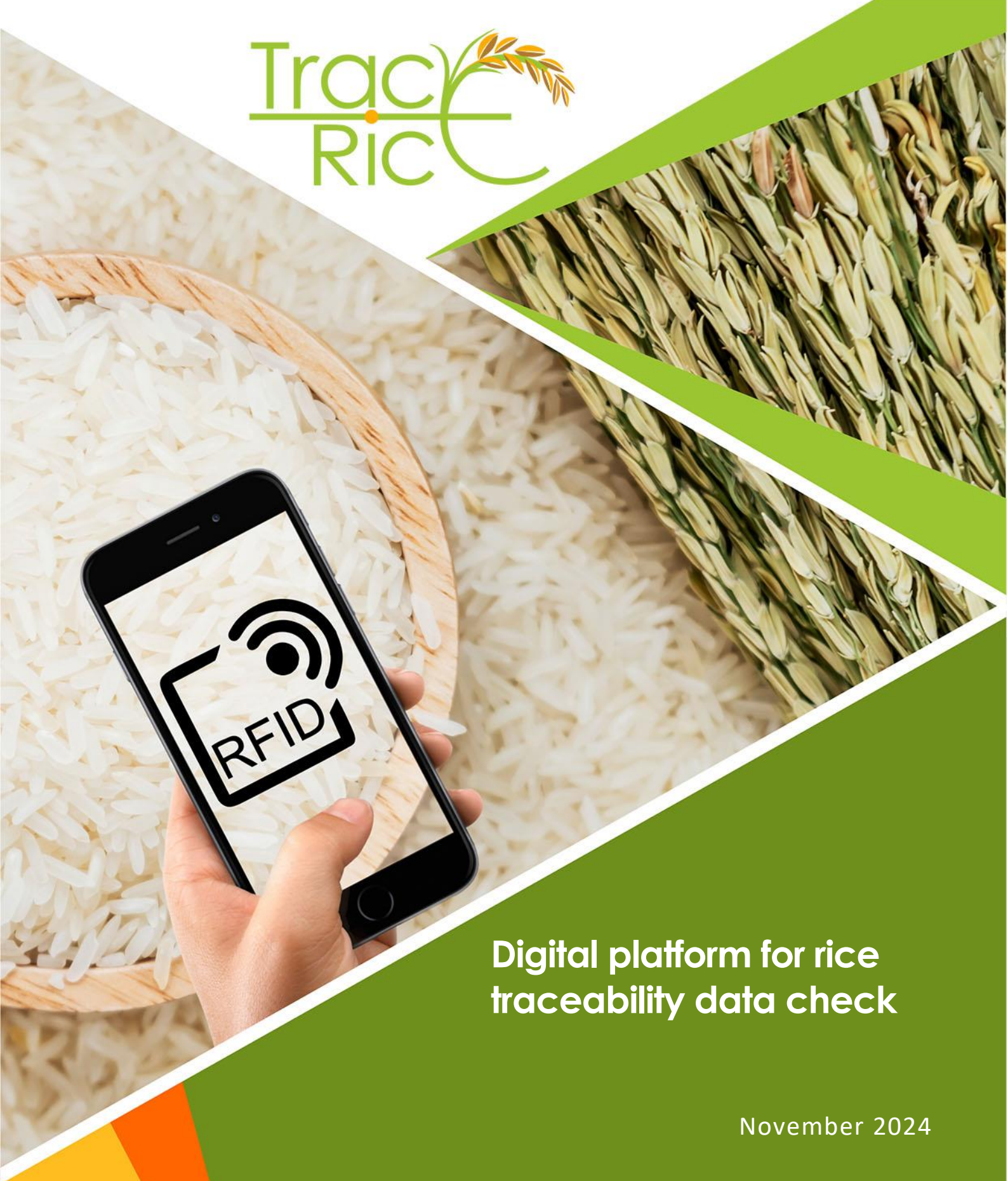

Digital platform for rice traceability data check

November 2024

TRACE-RICE with Grant nº 1934, is part of the PRIMA Programme supported under Horizon 2020, the European Union's Framework Programme for Research and Innovation

PRIMA
PARTNERSHIP FOR RESEARCH AND INNOVATION IN THE MEDITERRANEAN AREA

http://trace-rice.eu

TECHNICAL REFERENCES

Project Acronym	TRACE-RICE
Project Title	Tracing rice and valorizing side streams along Mediterranean blockchain
Project Coordinator	Carla Moita Brites carla.brites@iniav.pt
Project Duration	September 2020 – October 2024 (50 months)

Deliverable No.	1.7
Dissemination level*	CONFIDENTIAL
Work Package	1
Task	1.5
Lead beneficiary	INIAV instead of M. Dynamics
Contributing beneficiaries	UNL-ITQB
Due date of deliverable	31 October 2024
Actual submission date	17 December 2024

Written by: Pedro Barros, Hugo Rodrigues, Pedro Sampaio, Carla Brites

HISTORY OF CHANGES			
Date	Beneficiary	Version	Change
31/10/2024	UNL-ITQB	V1	Version sent to coordinator
18/12/2024	INIAV	V2	Final version approved by project coordinator

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The TRACE-RICE digital platform provides a centralized repository designed to integrate and manage genomic, phenotypic raw data, predictive models for rice quality traits using MATLAB and metadata associated with Mediterranean rice varieties. Hosted on the DMPortal via BioData.pt—the Portuguese node of ELIXIR (European life sciences infrastructure)—the platform complies with FAIR principles (Findable, Accessible, Interoperable, Reusable), ensuring transparency, free accessibility, and collaborative utility for research stakeholders.

Key Features:

1. Comprehensive genomic data repository covering 22 rice varieties, accessible via the European Nucleotide Archive (ENA).
2. Centralized dataset supporting quality control, research reproducibility, and variety authentication.
3. Advanced bioinformatics workflows for sequencing, mapping, and variant analysis, enabling the discovery of key polymorphisms.
4. Integration of phenotypic metadata, genetic insights and predictive models for rice quality traits into a unified framework, promoting traceability and innovation in rice value chain.

The TRACE-RICE repository provides not just data storage but a foundation for future research, helping to meet global demands for food security and sustainability.

1. Objectives of the Digital Platform

The TRACE-RICE digital platform, hosted on the DMPortal, the national Dataverse instance developed by Biodata.pt, serves as a long-term centralized repository for genomic, phenotypic, and associated metadata generated under the TRACE-RICE project. It is accessible at: TRACE-RICE Dataverse.

The platform aims to:

- **Integrate Data:** Enable streamlined integration for further analysis of genomic and phenotypic data for Mediterranean rice varieties.
- **Ensure FAIR Compliance:** Promote data accessibility and transparency by adhering to FAIR principles.
- **Support Collaboration:** Provide structured, standardized datasets to facilitate collaborative research.

2. Genomic Data Analysis Workflow and Integration Framework

2.1. Data Collection and Submission

- **Sequencing:** Whole-genome sequencing (WGS) was performed on 20 rice varieties using the Illumina NovaSeq 6000 platform (Macrogen, South Korea).
- **Data Deposition:** Sequencing data (FASTQ files) were deposited in the European Nucleotide Archive (ENA) under project accession code PRJEB64146.
- **Additional Data:** Genomic data for two additional varieties, *Bomba* and *Puntal*, were retrieved from ENA for further analysis.

2.2. Preprocessing

- **Quality Assessment:** Raw read quality was evaluated using FastQC (v0.11.9).
- **Quality Filtering:** Trimmomatic (v0.39) was used to trim low-quality regions in the *Bomba* and *Puntal* datasets.

2.3. Mapping

- Reads were aligned to the Nipponbare 1.0 reference genome (IRGSP-1.0, release 52) using BWA-MEM (v0.7.17).
- SAMtools (v1.7) was employed to compute coverage, sort and index data, and mark duplicates.

2.4. Variant Calling

- Short variant discovery was conducted using GATK4, following best practices for germline variant analysis.
    - Variants were called using GATK HaplotypeCaller and aggregated into a cohort VCF file.
- The cohort VCF file was deposited in the European Variation Archive (EVA) under project accession code PRJEB83571.

3. Genomic Data Upload Procedures

The TRACE-RICE dataset includes 22 gVCF files in compressed (.gz) format, adhering to the Variant Call Format (VCF) specifications.

VCF Structure

- **Metadata:** Lines prefixed with ## describe file structure.
- **Header:** A header line prefixed with # defines column content (**see Table 1**)
- **Variant Data:** Each line represents a genomic position and corresponding genotype information.

Data Summary

- The VCF files detail genomic regions differing from the reference genome (i.e., variants).
- The total dataset comprises 14,776 variants (~14.7 GB).
- Full specifications of the VCF file format can be found at VCF Specifications on GitHub.

Table 1: VCF Header Line Specification

Col	Field	Brief description
1	CHROM	Chromosome: the chromosome identifier/number from the reference genome.
2	POS	Position: the reference position (1st base = position 1) sorted numerically in increasing order for each reference CHROM.
3	ID	Identifier: Unique identifier when available.
4	REF	Reference base(s): the base(s) A,C,G,T or N present in this POS in the reference genome.
5	ALT	Alternate base(s): the alternate base(s) present in this POS for this Sample. <NON_REF> means that the base in this POS for this Sample differ from the REF.
6	QUAL	Quality: quality score for the assertion made in ALT [i.e. -10log prob(call in ALT is wrong)].
7	FILTER	Filter status: '.' or 'MISSING' for no filters applied; 'PASS' for if position passed all filters.
8	INFO	Additional information: additional information can be specified here. END: End reference position indicating the variant spans position POS-END on reference/contig CHROM.
9	FORMAT	Format: colon-separated keys specifying genotype fields.
10	Sample ID (e.g., Arborio)	Genotype fields for this specific Sample in this POS. GT: Genotype; DP: Read depth; GQ: Conditional genotype quality; MIN_DP: Minimum read depth; PL: Phred-scaled genotype likelihoods

This collection of files, in VCFv4.2 format, was generated from whole-genome sequencing data of 22 rice varieties using advanced bioinformatics tools, including GATK. In table 2 is the list of variation data for these 22 rice varieties.

Table 2: List of Variation Data for 22 rice varieties:

Rice variety	File name	File size (Mb)
Albatros	Albatros_g.vcf.gz	450
Arborio	Arborio_g.vcf.gz	466
Arelate	Arelate_g.vcf.gz	673
Ariete	Ariete_g.vcf.gz	433
Basmati-TypeIII	BasmatiTypeIII_g.vcf.gz	866
Bomba	Bomba_g.vcf.gz	980
CL-28	CL28_g.vcf.gz	471
Caravela	Caravela_g.vcf.gz	476
Carnaroli	Carnaroli_g.vcf.gz	631
Elettra	Elettra_g.vcf.gz	820
Gageron	Gageron_g.vcf.gz	806
Giza-177	Giza177_g.vcf.gz	650
Giza-181	Giza181_g.vcf.gz	780
JSendra	JSendra_g.vcf.gz	766
Lusitano	Lusitano_g.vcf.gz	716
Maçarico	Macarico_g.vcf.gz	1070
Manobi	Manobi_g.vcf.gz	315
Puntal	Puntal_g.vcf.gz	1110
Ronaldo	Ronaldo_g.vcf.gz	806
Super-Basmati	Super_Basmati_g.vcf.gz	599
Teti	Teti_g.vcf.gz	435
Ulisse	Ulisse_g.vcf.gz	457
-	-	TOTAL: 14,776 (14.7 Gb)

4. User Interface

4.1 Genomic Data

Public genomic data for 22 rice varieties, stored in ENA, can be accessed under project accession PRJEB64146. These files are listed in table 3 and have been publicly available since December 10, 2024.

Table 3: Genomic Data produced by TRACE-RICE for 20 rice varieties, stored in ENA Project Accession - PRJEB64146

Rice variety	ENA Accession code	File names
Albatros	ERR11777755	TRACE_RICE_ITQB_Albatros_1.fastq.gz TRACE_RICE_ITQB_Albatros_2.fastq.gz
Arborio	ERR11789060	TRACE_RICE_ITQB_Arborio_1.fastq.gz TRACE_RICE_ITQB_Arborio_2.fastq.gz
Arelate	ERR11777790	TRACE_RICE_ITQB_Arelate_1.fastq.gz TRACE_RICE_ITQB_Arelate_2.fastq.gz
Ariete	ERR11777858	TRACE_RICE_ITQB_Ariete_1.fastq.gz TRACE_RICE_ITQB_Ariete_2.fastq.gz
Basmati-TypeIII	ERR11777987	TRACE_RICE_ITQB_BasmatiTypeIII_1.fastq.gz TRACE_RICE_ITQB_BasmatiTypeIII_2.fastq.gz
CL-28	ERR11779317	TRACE_RICE_ITQB_CL28_1.fastq.gz TRACE_RICE_ITQB_CL28_2.fastq.gz
Caravela	ERR11778234	TRACE_RICE_ITQB_Caravela_1.fastq.gz TRACE_RICE_ITQB_Caravela_2.fastq.gz
Carnaroli	ERR11778992	TRACE_RICE_ITQB_Carnaroli_1.fastq.gz TRACE_RICE_ITQB_Carnaroli_2.fastq.gz
Elettra	ERR11779370	TRACE_RICE_ITQB_Elettra_1.fastq.gz TRACE_RICE_ITQB_Elettra_2.fastq.gz
Gageron	ERR11779532	TRACE_RICE_ITQB_Gageron_1.fastq.gz TRACE_RICE_ITQB_Gageron_2.fastq.gz
Giza-177	ERR11783480	TRACE_RICE_ITQB_Giza177_1.fastq.gz TRACE_RICE_ITQB_Giza177_2.fastq.gz
Giza-181	ERR11783914	TRACE_RICE_ITQB_Giza181_1.fastq.gz TRACE_RICE_ITQB_Giza181_2.fastq.gz
JSendra	ERR11784529	TRACE_RICE_ITQB_JSendra_1.fastq.gz TRACE_RICE_ITQB_JSendra_2.fastq.gz
Lusitano	ERR11784649	TRACE_RICE_ITQB_Lusitano_1.fastq.gz TRACE_RICE_ITQB_Lusitano_2.fastq.gz
Maçarico	ERR11786074	TRACE_RICE_ITQB_Macarico_1.fastq.gz TRACE_RICE_ITQB_Macarico_2.fastq.gz
Manobi	ERR11787405	TRACE_RICE_ITQB_Manobi_1.fastq.gz TRACE_RICE_ITQB_Manobi_2.fastq.gz
Ronaldo	ERR11787704	TRACE_RICE_ITQB_Ronaldo_1.fastq.gz TRACE_RICE_ITQB_Ronaldo_2.fastq.gz
Super-Basmati	ERR11788330	TRACE_RICE_ITQB_SuperBasmati_1.fastq.gz TRACE_RICE_ITQB_SuperBasmati_2.fastq.gz
Teti	ERR11789060	TRACE_RICE_ITQB_Teti_1.fastq.gz TRACE_RICE_ITQB_Teti_2.fastq.gz
Ulisse	ERR11792779	TRACE_RICE_ITQB_Ulisse_1.fastq.gz TRACE_RICE_ITQB_Ulisse_2.fastq.gz

Additionally, files from the *Bomba* and Puntal rice varieties (Table 4) were retrieved from genomic data produced by another research group and integrated into the TRACE-RICE analysis.

Table 4: Genomic Data of Bomba and Puntal retrieved from ENA Project Accession - PRJEB13328

Rice variety	ENA Accession codes	File names
Bomba	ERR1368635 ERR1368762 ERR1368812	ERR1368635_1.fastq.gz ERR1368635_2.fastq.gz ERR1368762_1.fastq.gz ERR1368762_2.fastq.gz ERR1368812_1.fastq.gz ERR1368812_2.fastq.gz
Puntal	ERR1388864 ERR1388865 ERR1388866	ERR1388864_1.fastq.gz ERR1388864_2.fastq.gz ERR1388865_1.fastq.gz ERR1388865_2.fastq.gz ERR1388866_1.fastq.gz ERR1388866_2.fastq.gz

- **Example of Access:**
  - View: ERR1368635
  - Download: ftp://ftp.sra.ebi.ac.uk/vol1/err/ERR136/005/ERR1368635

4.2 Genetic Variants Data

Joint variation data for the 22 varieties are stored in EVA and accessible under project accession PRJEB83571. This file is listed in Table 5 and has been publicly available since January 01, 2025.

Table 5: Genetic variant data produced by TRACE-RICE for 20 rice varieties, stored under the EVA Project Accession PRJEB83571.

Rice varieties	File name
All previously mentioned 22 rice varieties	TRACE-RICE_all_varieties.vcf.gz

5. Integration in DMPortal

The TRACE-RICE Dataverse hosted on the DMPortal acts as a centralized repository for genotypic, phenotypic, and physico-chemical data from 22 targeted rice varieties with the following key features and hosted datasets:

1. **Physico-Chemical Data**
   - **Source**: INIAV Dataset.
   - This dataset provides detailed measurements of the physico-chemical properties of rice, including grain hardness, moisture content, amylose levels, essential for evaluating functional quality.
2. **Metadata for 20 Rice Varieties**
   - Metadata of the genotypic and phenotypic experiments performed in TRACE-RICE
   - **Compliance**: adheres to the MIAPPE (Minimum Information About a Plant Phenotyping Experiment) standard.

- **Access**: Dataset Link

3. **Genetic Variants of Interest**
   - A curated dataset of genetic variants linked to important phenotypic traits.
   - **Access**: Dataset Link.
   - Enables researchers to explore genotype-to-phenotype relationships in rice.
4. **Cohort VCF File**
   - **Repository**: Available via the European Variation Archive (EVA), under project accessionPRJEB83571.
   - This file consolidates variant call data for the 22 varieties, facilitating genetic analyses.
5. **Predictive Models for Rice Quality Assessment**
   - Developed in MATLAB (R2023a) using advanced methodologies such as PLS, PLS-DA, SVM, Machine Learning, and ANN.
   - Provides tools for assessing rice quality traits and classifying rice types via Near-Infrared (NIR) spectroscopy.
   - Access: Dataset Link

The TRACE-RICE Dataverse seeks to bridge the gap between genomic insights and applied rice research by supporting breeding programs aimed at enhancing rice quality and sustainability, advancing quality control and traceability initiatives within the rice supply chain, and fostering collaborative research efforts across genomic and agronomic domains. All datasets are securely hosted and readily accessible to the TRACE-RICE research community and partners through the project BioData Portal.

6. Conclusion

The TRACE-RICE project has developed a robust framework for integrating genomic and phenotypic data, significantly enhancing rice variety authentication and traceability. This digital platform serves as an invaluable resource for the scientific community, fostering deeper exploration of rice genomics and promoting sustainable production practices. Future directions involve expanding the dataset to encompass additional rice varieties and diverse environmental conditions, as well as collaborating with stakeholders to implement effective traceability solutions across the agricultural rice value chain.

TRACE-RICE

TRACE-RICE Consortium

Instituto Nacional de Investigação Agrária e Veterinária, I.P.

iBET Instituto de Biologia Experimental e Tecnológica

ITQB Universidade Nova de Lisboa

Casa do Arroz

ernesto morgad desde 1920

Mater Dynamics

BGI www.bgi.pt

Grupo Desarrollo

iata Instituto de Agroquímica y Tecnología de Alimentos

Alexandria University