24th CGIAR System Council Meeting
10–11 June 2026, Antalya, Türkiye

Additional Reference (no SC24 agenda item)	**Update from the Standing Panel on Impact Assessment (SPIA)**
Distribution notice:	☒ May be distributed without restriction ☐ Restricted to Members and Active Observers ☐ Restricted to Voting Members only
Issue date:	27 May 2026
Purpose:	☐ For consent agenda/information ☐ For discussion ☐ For discussion and decision ☒ Other — Additional Reference Material for the System Council
Requested action:	For information
Executive summary:	SPIA is providing five informational pre-reads to the 24th meeting of the System Council to update members on ongoing analytical, methodological, and data quality work. Materials include a country studies update covering survey implementation, DNA fingerprinting, remote sensing, and technical reporting progress, alongside visual stocktaking summaries from Côte d'Ivoire and Morocco. The package also includes a recently published *Scientific Reports* article on rice varietal adoption in Vietnam, a non-technical visual summary of findings, a practical DNA fingerprinting guidebook based on multi-country experience, and updated SPIA Data Documentation Standards to strengthen reproducibility, transparency, and scientific rigor across country studies.
Draft decision point:	N/A
Supporting materials:	• Annex with Attachment Abstracts • Attachments 1-5 detailed in Annex
Prepared by:	Standing Panel on Impact Assessment (SPIA) and its Professional Team

CGIAR
STANDING PANEL ON
IMPACT ASSESSMENT

Annex

As a part of the pre-reads package, SPIA is providing 5 pre-reads to 24th meeting of System Council. There is not a specific agenda item for SPIA. The information is supplied for Council's further information about SPIA's ongoing work.

Attachment 1. SPIA Country Studies Update for SIMEC & SC provides a country-wise update for each of the ongoing country studies. The document highlights, by country, planned survey rounds, crops selected for DNA fingerprinting, scope for remote sensing, and the progress demonstrated to SPIA during the technical reporting process. **Pre-reads 1.a and 1.b** show visualizations of the stocktaking results submitted by the Cote d'Ivoire and Morocco country teams respectively.

Attachment 2. Vietnam_Rice_ScientificReports_Publication_2026 is the manuscript that was recently published in Scientific Reports (of the Journal Nature). The study involved sequencing the DNA of samples from 766 rice-growing households across Vietnam (sampled through the Vietnam Household Living Standards Survey) and matching the genetics against a reference library of major varieties released in the country.

Attachment 3. Vietnam_Carousel summarizes the main findings from this paper in a visual format for non-technical audiences.

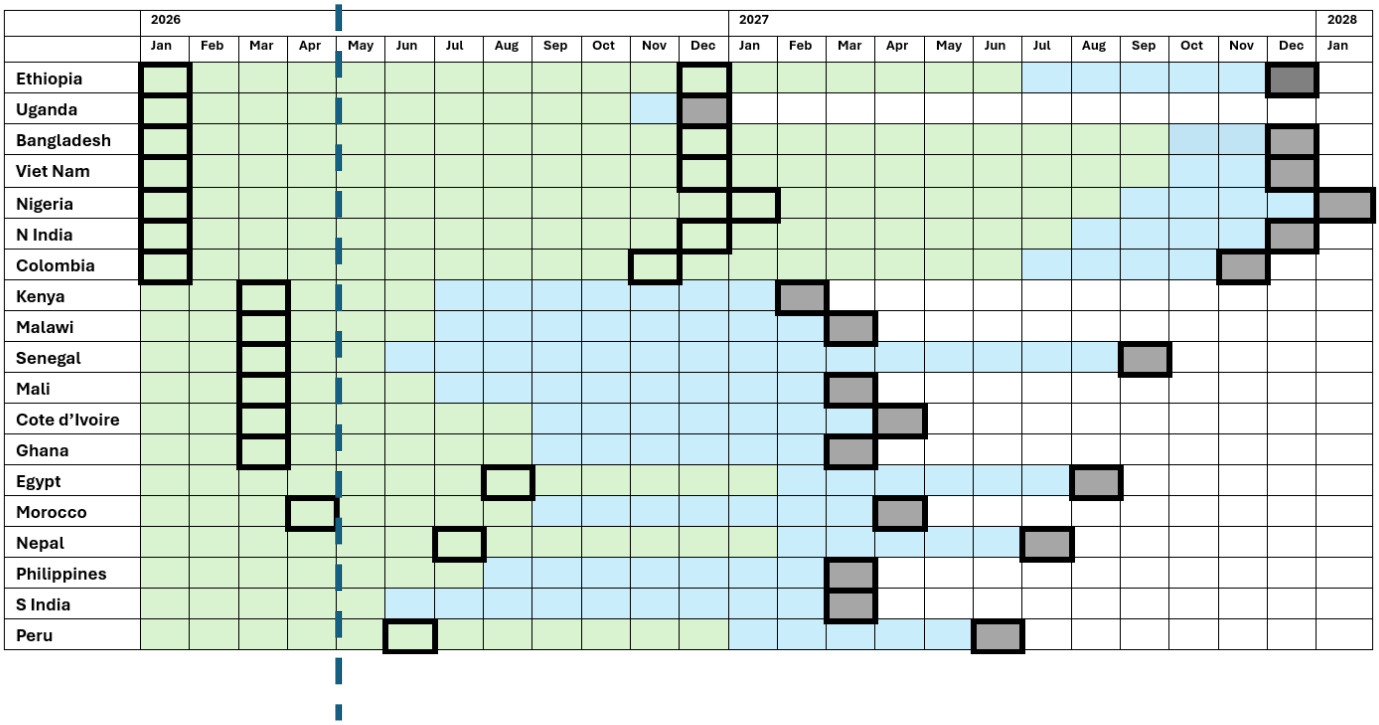
Attachment 4. DNA_Guidebook draws on prior SPIA experience in Ethiopia, Uganda, Vietnam and Bangladesh to offer a practical overview of DNA fingerprinting to researchers from various disciplines who wish to enhance the quality of crop varietal adoption data. It covers the necessary stages (from sample collection to data analysis), required skills (emphasizing collaboration among researchers with diverse expertise), and detailed cost considerations (covering survey-related, lab-related, and technical expertise costs). This publication is currently under peer review.

Attachment 5. SPIA_Data_Documentation_Standards sets out the minimum requirements for data documentation and reproducibility that all SPIA country studies must fulfil. It provides a shared framework for data documentation, articulates why these standards are indispensable to ensure scientific rigor, and provides guidance on best practice across the full range of data types.

SPIA Country Studies Update for SIMEC / System Council

May 19th 2026

Across the country study portfolio, the contracted teams have been making strong progress as revealed in this period of technical reporting. Progress reports were received between December 2025 and April 2026 as per the Gantt Chart below. In this update we pull out select details from each study that are novel or indicative of interesting analyses that are in the pipeline.



Ethiopia: PIs Tanguy Bernard (Bordeaux); Kaleab Baye (Addis Ababa University)

The project will soon launch the Ethiopian Atlas of Food Security and Nutrition, bringing together all major household surveys in Ethiopia, along with price data, administrative datasets, and satellite imagery, and harmonizing them across time and space using a 5×5 km grid. Using machine learning, geospatial, and Bayesian modeling techniques, the Atlas generates high-resolution estimates for 120 indicators related to Drivers, Food Supply, Food Environment, Individual Factors, Outcomes, and Cross-cutting issues. These estimates can be aggregated to any administrative level (Woreda, Zone, or Region) over a 13-year period.

Uganda: PIs Leah Bevis (Ohio State); Jeff Michler (Arizona)

In November, SPIA took the difficult decision to wind down the Uganda country study early, saving significant funds that are being reallocated to impact case studies across the country study portfolio. Despite this, work is ongoing to embed DNA fingerprinting of cassava varieties in a forthcoming survey round with Uganda Bureau of Statistics, as well as complementary analyses by the Leah Bevis and collaborators examining the role of extension in promoting disease-resistant cassava varieties. The SPIA

Uganda report (2025) highlighted how disease-resistant cassava has started to be adopted at scale in Uganda over the past decade.

Bangladesh: PIs Andrew Bell, Martina Occelli (Cornell), Saiful Islam (Bangladesh Agricultural University)

Plans are being put in place to implement a full fourth survey round of the Bangladesh Integrated Household Survey (BIHS), with a focus on rice and aquaculture innovations. The team will collect GPS coordinates of a sub-sample of rice plots to allow for remote sensing approaches to be used to train models for ongoing monitoring of change over time.

Viet Nam: PIs Matin Qaim, Tung Nguyen (U Bonn)

The forthcoming survey round will incorporate DNA fingerprinting for rice and cassava, along with a major focus on collecting ground truth data for training models for detecting Alternate Wetting and Drying (AWD) at scale using remote sensing. SPIA has long been concerned about how we measure AWD adoption as it is a complex innovation, but it is amenable to remote sensing and this survey represents the most comprehensive effort yet to collect adoption data in rich detail over space and time.

Nigeria: PI Yonas Alem (U Cape Town)

The team is collaborating with the Nigerian Bureau of Statistics (NBS) to embed an entire survey round in the sample frame used by the National Agricultural Sample Survey (NASS) in 2023. The survey will feature DNA fingerprinting of cassava and either cowpea or sorghum (choice being made in the coming weeks). The sample will allow the team to examine adoption in "urban" enumeration areas that prior survey data suggests have higher rates of adoption of innovations than seen in exclusively rural enumeration areas.

India: PIs Madhura Swaminathan (Foundation for Agrarian Studies / Indian Statistical Institute), Gaurav Datt (Monash U)

The team are designing a survey across several states in the north of the country, embedding data collection protocols designed to capture priority innovations identified through an outstanding stocktaking report. The project has a website that they update regularly with news on their activities. Q1 2026 the team carried out pilot data collection with four candidate survey companies, with a view to testing their capabilities for implementing the full survey round.

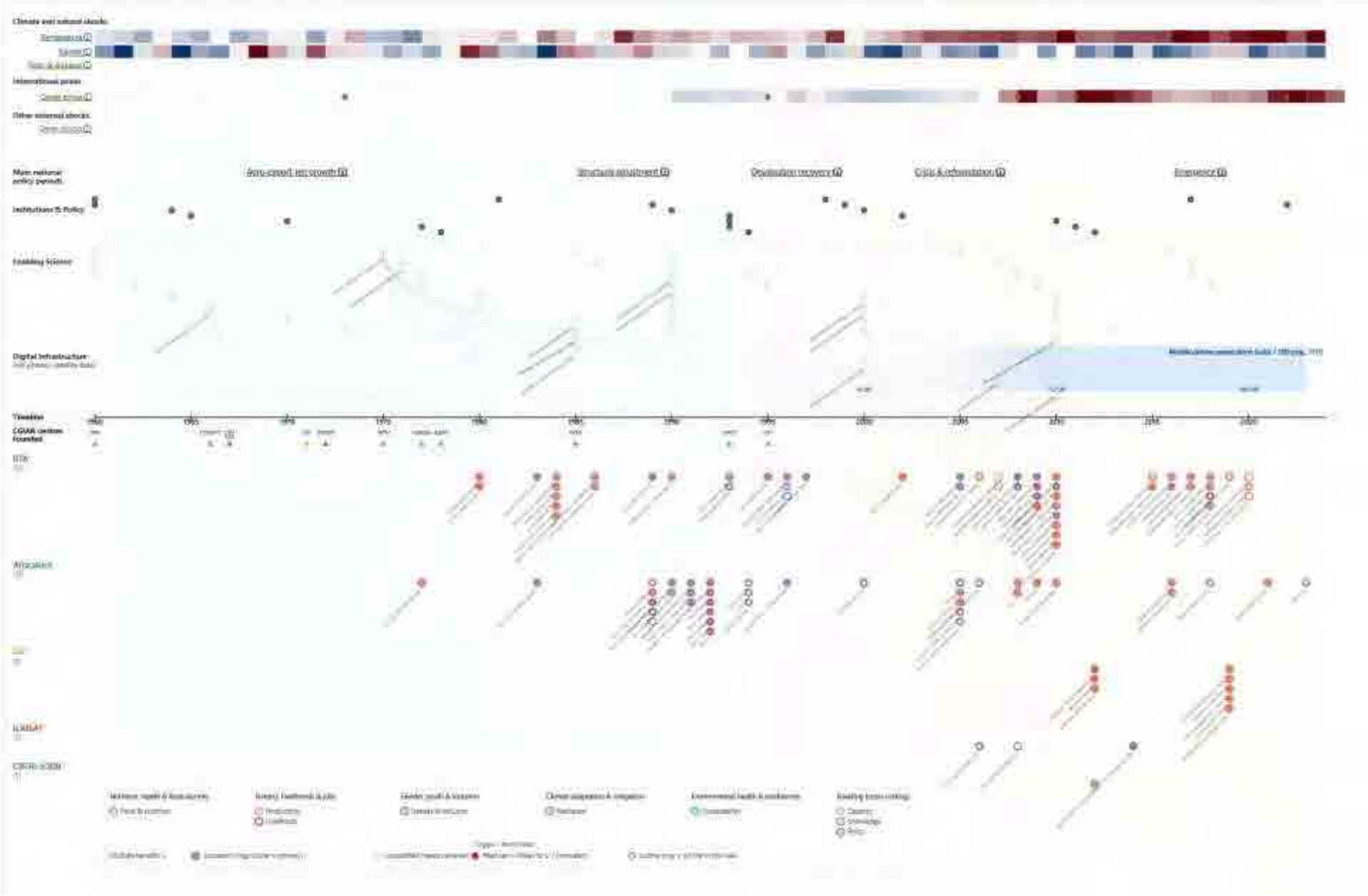
Colombia: PI Rachid Laajaj (U Los Andes)

The team have established a focus on improved forages (for DNA fingerprinting and understanding adoption at scale), rice varieties and management practices (the latter using remote sensing), and silvo-pastoral systems (using remote sensing). Q1 2026 the team carried out a pilot study to trial a survey data collection instrument and to practice leaf tissue sampling for rice.

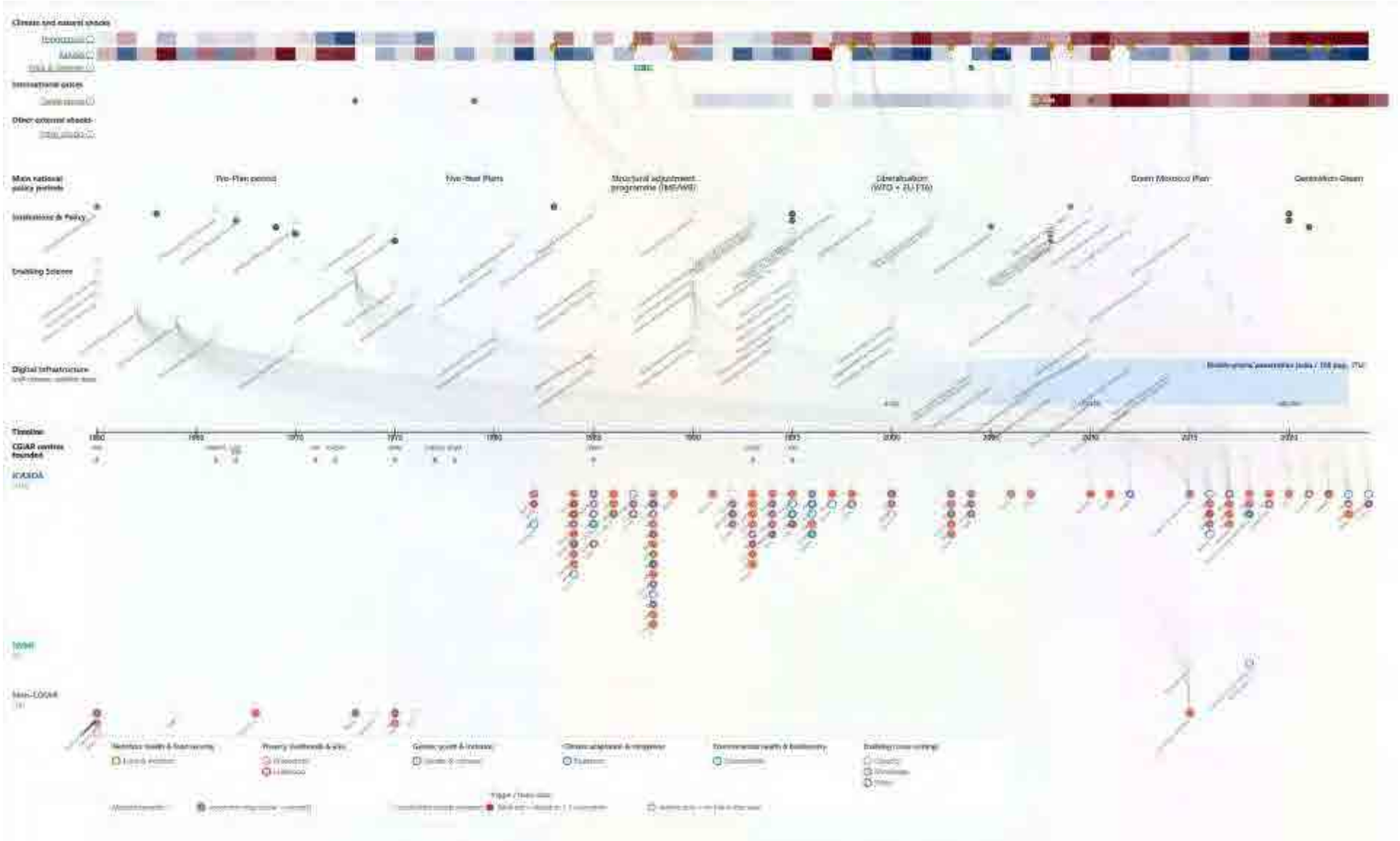
Kenya, Senegal, Malawi: Several potential impact cases are being scoped for potential top-up funding with monies saved from the early cessation of the Uganda study.

Cote d'Ivoire, Morocco: Both countries have developed outstanding visualizations of the stocktaking results, putting innovations in their historical context. Both are interactive html files and are attached to this report for download and exploration.

SPIA_1.a. CoteIvoire_Stocktake_Swimlane_30Apr2026



SPIA_1.b. Morocco_Stocktake_Swimlane_30Apr2026



Scientific Reports

https://doi.org/10.1038/s41598-026-44331-9

Article in Press

Farmers more likely to adopt rice varieties with higher density of quantitative trait loci (QTL) in Viet Nam

Received: 27 October 2025

Accepted: 11 March 2026

Published online: 08 April 2026

Cite this article as: Kosmowski F., Visaria S., Stevenson J. *et al.* Farmers more likely to adopt rice varieties with higher density of quantitative trait loci (QTL) in Viet Nam. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-44331-9

Frédéric Kosmowski, Sujata Visaria, James Stevenson, Davis Gimode & John Damien Platten

Farmers More Likely to Adopt Rice Varieties with Higher Density of Quantitative Trait Loci (QTL) in Viet Nam

Frédéric Kosmowski[a]*, Sujata Visaria[ac], James Stevenson[a], Davis Gimode[a], and John Damien Platten[b]

[a] CGIAR Standing Panel on Impact Assessment

[b] International Rice Research Institute

[c] City St. George's, University of London

* Email: f.kosmowski@cgiar.org

Abstract

This paper offers the first nationwide analysis of Vietnamese rice varieties by combining DNA-based identification with quantitative trait loci (QTLs) and household survey data. Using nationally representative data from 2022, we found that 51% of rice farmers grew improved varieties. These varieties contained significantly more beneficial QTLs associated with yield, grain quality, and resistance to biotic and abiotic stresses than genetically unidentified varieties. On average, the improved varieties cultivated had been released 14 years prior to 2022. Farmer's socioeconomic characteristics correlated with adoption patterns: belonging to an ethnic minority or residing in a government-classified poor commune significantly reduced the likelihood of growing an improved variety. Among adopters, varietal traits were further associated with specific adoption choices. Each additional trait-related QTL was associated with a 0.9 percentage point increase in a province's adoption rate. Traits conferring tolerance to abiotic stress were positively associated with adoption, suggesting farmers may prefer varieties that enhance resilience to environmental stressors.

Keywords: Adoption; QTL; Marker-assisted selection; Molecular breeding; Rice; Viet Nam;

Introduction

Rice has been cultivated in Viet Nam for over 4000 years. This staple crop underpins its economy and food security. Since the *Đổi Mới* reforms of 1986, a transition began that moved rice agriculture from subsistence farming to market-oriented production. Today Viet Nam is the world's second-largest rice exporter, and Vietnamese rice is exported to high-quality markets such as Japan, the European Union, and the United States. In 2023, Viet Nam exported 8.29 million tons of rice, 90% of which originated from the Mekong River Delta[1].

Since the early 1990s, molecular breeding, particularly the use of molecular markers, has played an increasingly important role in enhancing the traits in rice varieties that determine growth rates, yield and adaptability. Marker-Assisted Selection (MAS) and Marker-Assisted Backcrossing (MABC) allow breeders to identify and more precisely incorporate favorable genetic regions into elite cultivars[2]. Central to these approaches is the identification of Quantitative Trait Loci (QTLs)—specific genomic regions associated with variation in complex traits[3–5]. QTL analysis aims to determine the number, location, interactions, and effects of these loci[6].

Viet Nam has embraced these tools rapidly. Since 1984, 565 improved rice varieties have been certified nationally, of which 413 were released in the last two

decades. Modern breeding programs in Viet Nam increasingly employ QTL pyramiding, a strategy that combines multiple QTLs associated with desirable traits into a single genetic background to enhance performance across the dimensions of productivity, quality, and resilience[7]. Breeding in Viet Nam has targeted yield-related traits, such as plant vigor, tillering, panicle length, and grains per panicle[8], as well as shorter growth durations to facilitate multiple cropping cycles. In addition, there has been a focus on quality traits, such as low amylose content and aroma, in line with consumer preferences both domestically and in export markets[9]. There has also been an effort to breed varieties that are resistant to pests (e.g., brown planthoppers) and diseases (e.g., blast, bacterial leaf blight)[10,11], and those that can tolerate abiotic stresses such as salinity, submergence[12], and drought[13].

However, we have limited evidence about Vietnamese farmers' adoption of improved rice varieties and their genetic composition. A report by the Ministry of Agriculture and Rural Development (MARD)[14] identified 12 major rice varieties accounting for 47% of Viet Nam's rice-growing area, led by IR50404 (13%), OM5451 (7%), and OM4900 (5%). Similarly, a 2021 MARD survey found that five varieties were cultivated on over 100,000 hectares in Southern Viet Nam. The General Statistics Office (GSO) data also highlight OM5451, Bac Thom 7, and IR50404 as the dominant varieties (GSO, 2022). Through expert elicitation in the Mekong River Delta [15] estimated that 44% of households planted salt-tolerant varieties[15]. Rather than relying on farmers' reports of the variety they planted, which is known to be prone to substantial measurement error, this study uses DNA-based varietal identification of the crops that they grow [16,17]. The DNA data also provides insights into the genetic architecture, particularly the QTL content, of the rice varieties in farmers' fields.

We examine whether farmers are more likely to be growing rice varieties that contain more QTLs. There are good reasons why this may be the case. QTL pyramiding enables the integration of multiple beneficial traits into a single variety, so that varieties containing more QTLs likely have many of the traits that farmers value[18,19]. Multi-QTL varieties generally offer greater yield stability and profitability and may provide better tolerance to biotic and abiotic stresses. In addition, if farmers' produce meets the quality traits associated with many of these QTLs, they are more likely to have access to premium markets[20]. A QTL count is therefore a reliable proxy for breeding success, as it signifies the cumulative enhancement of a variety's adaptive and productive capabilities.

Yet there are also good reasons to doubt whether farmers would preferentially plant varieties that contain more QTLs. A growing literature has documented the fact that especially in low and middle income countries, farmers are often unaware of the (promised) traits of the varieties they are cultivating[17,21]. This could be due to a lack of information but it could also occur if the phenotypic traits fail to appear

in the plants—for instance, when specific genes are expressed only under particular environmental conditions. Varieties may also not be disseminated wide enough for farmers to have access to them.

This study addresses four research questions: (1) What is the current genetic composition of rice varieties cultivated by Vietnamese farmers?; (2) Which socioeconomic factors are associated with the adoption of trait-related Quantitative Trait Loci (QTLs)?; (3) Is the density of QTLs in rice varieties related to observed adoption patterns?; and (4) Which specific rice traits are associated with higher adoption levels?

By assessing farmers' varietal adoption using the underlying genetic information about their varieties, this study provides the first robust assessment of the reach of genetic improvements in Viet Nam's rice cultivation.

Materials and Methods

Survey design and rice sample collection

This study leverages data from the Viet Nam Household Living Standards Survey (VHLSS), conducted annually by Viet Nam's General Statistics Office. The VHLSS follows a stratified two-stage cluster sampling design and is nationally and regionally representative. In 2022, the survey covered approximately 46,995 households, including 13,650 rice-growing ones.

Although the VHLSS typically does not collect plot-level data or conduct direct field observations, through a collaboration with the CGIAR's Standing Panel on Impact Assessment, a new rice crop-sampling module was introduced in the 2022 survey. This module was administered to a random subsample of 832 rice-growing households, selected by odd-numbered GSO household IDs. Up to five households were selected per enumeration area.

In this exercise, trained enumerators identified plots that contained rice plants that were at least 20 days old. One plot was selected per household. One rice plant was selected from each of four randomly placed quadrants in this plot, and one leaf per plant was sampled using a hole punch. The samples from the four plants were then bulked, stored in silica gel. Each plot-level sample was barcoded, and samples were shipped to Hanoi.

Due to COVID-19-related delays in starting data collection, we were unable to spread the sample collection evenly across the year. The collected samples covered multiple cropping seasons: 73% from the winter–spring season, 11% from the summer–autumn season, and 16% from the autumn–winter season. The presented results represent household-level patterns of varietal use rather than production-weighted national averages.

In total, 832 samples were collected. Of these, 766 were successfully matched to household data. These samples were collected from 248 enumeration areas (EAs). The distribution across regions shows the Mekong River Delta most heavily sampled (n= 194 households), followed by the Red River Delta (n = 155), Northern Midlands & Mountain Areas (n = 178), North Central & Central Coastal Areas (n = 171), then smaller samples in the Central Highlands (n = 44) and Southeast (n = 24). This sample distribution appears aligned with the actual concentration of rice cultivation across Viet Nam.

Reference library creation and genotyping protocol

To accurately identify a crop variety using genomic techniques, one requires a robust reference library. This allows a genetic matching of the rice samples from the farm to known cultivars[22]. For this study, we assembled a comprehensive reference set consisting of 122 unique rice cultivars officially released in Viet Nam and 174 breeding lines from the International Rice Research Institute (IRRI) in the Philippines.

To build this library, the team contacted 17 public and private Vietnamese institutions engaged in rice breeding and seed release and requested specific breeder seed samples. A total of 147 requests were made, and 99 samples were obtained from six organizations. The remaining requests were declined, primarily because the institutions had discontinued seed production or because the variety was temporarily unavailable due to seasonality. For cultivars whose breeder seeds could not be obtained, commercial seeds and grains were purchased from seed retailers. This added 23 unique cultivars to the reference library. The IRRI elite lines were genotyped in a previous project and incorporated into the library. A full list of the reference materials is provided in Table S1.

Genotyping of the collected farm-based rice leaf samples was conducted by Agriplex Genomics using the PlexSeq platform, which supports high-level multiplexing for mid-density Single Nucleotide Polymorphism (SNP) analysis. This study employed Version 4 of the IRRI Rice Custom Amplicon (RICA) SNP panel, as described by [23], comprising 1,024 SNP markers. This included 797 SNPs from the Cornell 6 K Infinium Rice Chip[24], 205 trait-linked SNPs, and 22 quality control SNPs. This high throughput genotyping approach supports precise varietal matching and enables downstream analysis of genetic traits linked to rice performance and adaptation.

Variable measurements

Measurement of rice varietal adoption

A key part of this study was the estimation of genomic relationships between rice samples from farmers' fields and known rice varieties present in reference materials. In total, 1,017 single nucleotide polymorphisms (SNPs) were used to genotype the rice samples, along with 297 reference samples designated for use as references in matching.

The genotypes were first filtered using a minimum call rate threshold of 0.5 to eliminate those with excessive missing data. The SNPs were then filtered to retain only robust and informative loci. This entailed removing markers with a low minor allele frequency (MAF < 0.01), excluding those with call rates below 0.8, and eliminating monomorphic markers. Additional quality metrics, including polymorphism information content (PIC), observed heterozygosity, and inbreeding coefficient, were evaluated to help understand the genetic properties of the samples and the information content of the markers used in downstream analyses.

After quality control, 903 SNPs were retained, including 772 general SNP markers and 131 trait-specific markers. However, these marker types were not distinguished during varietal assignments. Among the field samples, 789 passed quality control and were retained for downstream analysis. Reference samples that displayed high heterozygosity levels and low inbreeding coefficients were retained in the reference library since they may represent the true genetic makeup of some varieties in circulation.

To assign field samples to known varieties, pairwise genetic comparisons were conducted between the field and reference genotypes. This involved calculating percentage similarity (to assess purity) and identity-by-state (IBS) genetic distance (rendered as 1 minus IBS, with distances closer to zero indicating closer relationship). IBS analysis was done using the R package SNPRelate [25]. A field genotype was considered to be related to the reference if the IBS distance was <0.05. In cases where a sample matched multiple references (common due to close genetic similarity), the "Top Reference" was selected based on the lowest IBS score and the highest purity score, and this reference ID was assigned to the sample.

In total, 390 of the 766 field samples were successfully matched with a reference variety. For classification purposes, non-assigned samples (likely landraces) were grouped by region—Northern Midlands and Mountainous Area (NMMA); North Central Coast and Central Highlands (NCCCA); Red River Delta (RRD); Central Highlands (CH); Southeast Region (Southeast); and Mekong River Delta (MRD)—on the basis that these landraces are likely to share similar agro-ecological contexts, which may result in some degree of clustering.

Measurement of rice QTLs

The presence of QTLs in the field samples was directly inferred from the allelic state of specific trait-linked SNP markers. The RICA V4 panel[23] was designed in collaboration with IRRI and includes 205 SNPs selected for their linkage to key major-effect QTL loci widely used in rice[26]. We assessed the presence of beneficial alleles based on these markers as diagnostic of trait-conferring QTLs in the background of improved varieties.

The QTLs we focused on in this study reflect the combined shifting priorities of rice breeding programs over the past several decades[8,27]. It relies on a standardized classification where the 'positive' [+] designation is assigned specifically to the allele or haplotype that confers the trait improvement desired by farmers. Trait-linked QTLs were grouped into categories based on the literature, as follows.

Yield: Eleven markers were used to assess traits related to grain number, panicle architecture, plant height, and flowering time. These included *Gn1a, Ehd1, RFT1, Hd1, GFR1, Hd2, Ghd7, NGR5, NAL1, Hd9/qDTY3.2* and *Hd3a*.

Quality: Key markers such as *SLG7, fgr, Alk, Chalk5, GS3, NAS3*, and *TGW6* play a role in determining grain quality by influencing its shape, starch composition, and nutritional content, thereby impacting consumer preference and market value.

Biotic stress resistance: These QTLs provide defense against major rice pathogens and pests, including rice blast (*Pita, qPi33*), bacterial blight (*Xa26* and *Xa4*), brown planthoppers (*Bph17* and *Bph32*), and tungro disease (*TSV1* and *STV11*).

Abiotic stress tolerance: These QTLs included drought tolerance (*qDTY1.1, DRO1, qDTY12.1, qDTY2.1, qDTY2.2, qDTY3.1, qDTY3.2 and qDTY4.1*), heat tolerance (*qHTSF4.1* and *TT1*), cold tolerance (*COLD1, qSCT1, qCTS10, and qPSST6*), salinity tolerance (*Saltol* and *qSIS1L*), submergence tolerance (conferred by *qSub1*, and *qAG3*, which specifically enables seeds to germinate and seedlings to establish under flooded, oxygen-deprived conditions).

To count the number of QTLs in each category that a household’s rice plants had, genomic data were aggregated across samples per variety collected on farmers’ fields. While trait-linked markers act as proxies for underlying functional genes, this approach has limitations when analyzing unassigned samples or landraces. The absence of specific marker alleles in landraces does not necessarily imply the absence of the associated traits. In landraces that have not undergone controlled breeding, natural recombination can disrupt marker–QTL associations, making direct comparisons with improved varieties more complex.

Measurement of household’s socio-economic characteristics

The information on socioeconomic characteristics of households was obtained from data collected by Viet Nam’s General Statistics Office (GSO) through its VHLSS questionnaire modules. These characteristics included whether the household was

headed by a female or an ethnic minority member, the age of the household head, and the highest education level completed by the household head. Additional indicators included annual household income (in million VND), the percentage of households located in poor communes, and the percentage of households with access to asphalt roads.

Survey weights

In VHLSS 2022, rice crop-sampling was conducted with 766 randomly chosen households, without regard to the population size of rice-growing households within each commune, leading to a sample that is not representative of the population of rice growers. To address this issue, household weights in the VHLSS were adjusted to account for the fact that the rice leaf samples were taken from randomly chosen households, with a limit of 5 households per enumeration area (See Text S1).

Empirical estimation

We aimed to understand whether farmers were more likely to adopt rice varieties that featured improved genetic traits, and if so, which traits were associated with adoption. To measure the genetic traits in the rice that farmers grew, we used the total number of trait-related QTLs (QTL count) from four trait categories: yield-related, grain quality, biotic stress resistance and abiotic stress tolerance. While our use of a count-based metric assumes that QTLs are equal despite differences in effect and environmental influence, in the absence of location-specific phenotype data, this serves as an indication of genetic enrichment as a result of allele pyramiding. This metric captures both QTL density and varietal improvements.

First, to examine whether the adoption of improved rice varieties is associated with households' socioeconomic characteristics, we estimated a probit regression model. The outcome variable included the seed status of the rice grown by the households (improved or not). Key independent variables included household-level characteristics.

Second, to test the hypothesis that varieties with a higher QTL count are more widely adopted by households, we estimated a series of linear regression models at the province-variety level. The dependent variable is the percentage of farming households in each province that adopted a given improved variety. Since adopters of improved rice varieties differ in both observable and unobservable characteristics from non-adopters, we restricted the following analysis to households that cultivated improved rice varieties. This restriction facilitates a more direct comparison with breeding outcomes, highlighting how modern genetic improvements relate to observed results.

In our baseline model, the only explanatory variable is the count of QTLs. We then sequentially add controls: regional fixed effects to account for agro-ecological and institutional differences, and next, province-level covariates. The province-level covariates—average years of education per household, average household income (in million VND), and a dummy for whether the main access road is asphalt—were aggregated from household data. All models were estimated using ordinary least squares (OLS) with the following specification:

$$\text{Adoption rate}_{vp} = \alpha + \beta \cdot \text{QTL Count}_{v} + \text{Region}_{p} + \gamma X_{p} + \epsilon_{vp}$$

where v indexes rice varieties and p indexes provinces. QTL Count_v represents the number of specific trait loci identified per variety, Region_p captures province-level fixed-effects, and X_p is a vector of province-level controls.

Third, to assess which specific traits are linked to higher levels of adoption, we adopted the previous approach but instead used the number of QTLs linked to specific trait categories— yield-related traits, grain quality, biotic stress resistance, and abiotic stress tolerance as independent variables. The empirical analysis was performed using R version 4.4.2. [28]

All methods were carried out in accordance with relevant guidelines and regulations. The study protocol was reviewed and approved by the Institutional Review Board of Hanoi University of Public Health (Decision No. 732/QD-DHYTCC). Informed consent was obtained from all participants prior to data collection.

Robustness checks

Robustness checks were conducted using two complementary approaches. First, we re-estimated the main specifications using only Winter-Spring season data, the largest sample in our dataset; the association of QTL count with adoption remained statistically significant (p = .1), with an effect size of similar magnitude (Table S5). This finding demonstrates that our main results are not driven by seasonality. Second, we used the average number of QTLs in the top three adopted varieties as an alternative dependent variable. The coefficient on QTL count remained positive and significant across all models (Table S6).

Results

Summary Statistics

We begin with summary statistics on rice varietal adoption in Viet Nam based on DNA fingerprinting data collected in tandem with VHLSS 2022. Figure 1 illustrates the share of different rice varieties grown by households in Viet Nam. The rice grown by nearly one-half (47%) of households could not be assigned to any of the references in the library. When visualized on a dendrogram (Figure S1), most of

these formed distinct separate clusters, suggesting that they are not necessarily noisy versions of elite varieties, but likely genetically distinct traditional landraces or local varieties. In contrast, assigned samples which represented 53% of the households clustered with their corresponding reference varieties. These showed considerable heterogeneity, corresponding to 28 distinct improved rice varieties, with no single variety holding more than 8% of the share. The most widely cultivated varieties were BT7 (7.71%), OM4900 (7.01%), and Dai Thom 8 (5.91%), and these were followed by TBR225, Thien Uu 8, and KD18. Other varieties, such as OM5451, N98, and IR50404, represented smaller proportions. Most samples belonged to the indica subspecies, with only 1.8% classified within the japonica group[1]. Collectively, these data revealed considerable diversity in the improved rice cultivars currently planted in Viet Nam, with no mega varieties. By 2022, the improved rice varieties present in farmers' fields were released, on average, 14 years ago. Additional details, including age, pedigree, and origin of the identified cultivars, are provided in Table S2.

[1] These samples exhibited at least 90% similarity to the japonica varieties Asiminori and Taichung65.

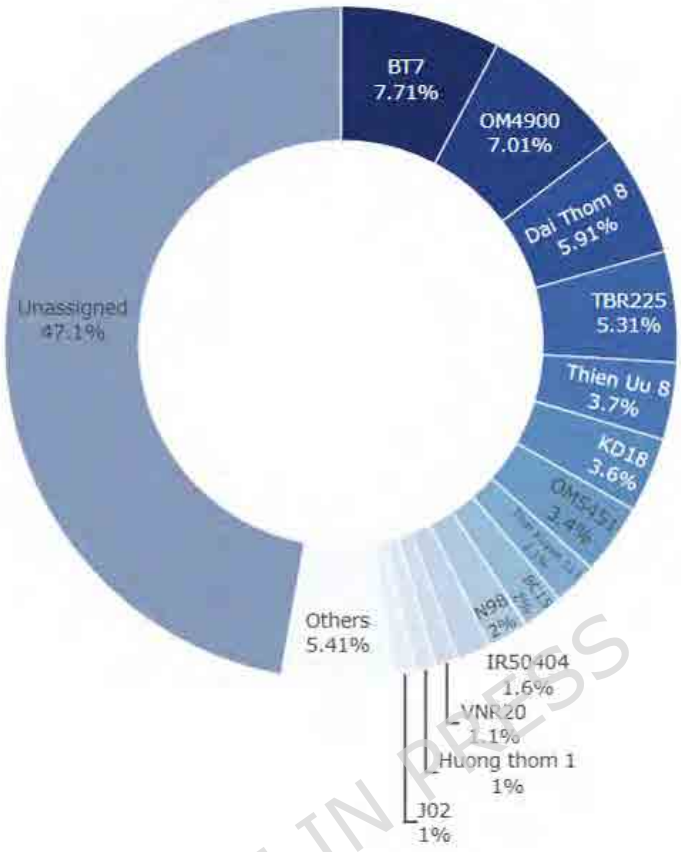


Figure 1. Distribution of Rice Varieties Grown by Farmers in Viet Nam in 2022. Each household in the sample is assigned a rice variety based on a DNA match between the rice sample collected from its plot and the reference library. The figure presents the percentage of households growing each rice variety identified. “Others” includes improved varieties found in less than 1% of households: HUONG UU 98, IR17494, N97, ST24, Q5, Jasmine 85, LH12, LTH31, RVT, VNR10, OM576, OM6162, OM7347, and SH14.

A key question is whether the varieties observed on farmers’ fields were developed using marker-assisted selection (MAS or MABC), which has been promoted in Viet Nam since the 2010s [8]. Although some of the most popular varieties, such as OM4900 and its derivative Dai Thom 8, are direct or indirect outcomes of MAS, most widely adopted varieties—including BT7, KD18, TBR225, Tien Uu 8, Thai Xuyen 111, OM5451, and BC15—were either developed through conventional breeding or derived from elite imported lines. These varieties generally originate from genetic backgrounds already rich in high-quality alleles, a fact supported by our QTL screening.

The spatial distribution of improved variety adoption revealed additional patterns. As shown in Figure 2, improved varieties were more likely to have been adopted in the delta regions, characterized by intensive rice cultivation, and in the southeastern region of the country. OM4900 and OM5451 were particularly

prevalent in the southern regions, whereas Dai Thom 8 and BC15 were the dominant varieties in the Red River Delta. Additionally, Figure S2b reveals spatial heterogeneity in the QTL density of rice varieties cultivated across Viet Nam. Provinces in the MRD and southeast regions tend to have rice varieties with higher mean numbers of QTLs. In contrast, several northern provinces and upland areas exhibited lower QTL count, consistent with their lower rates of adoption of improved varieties.

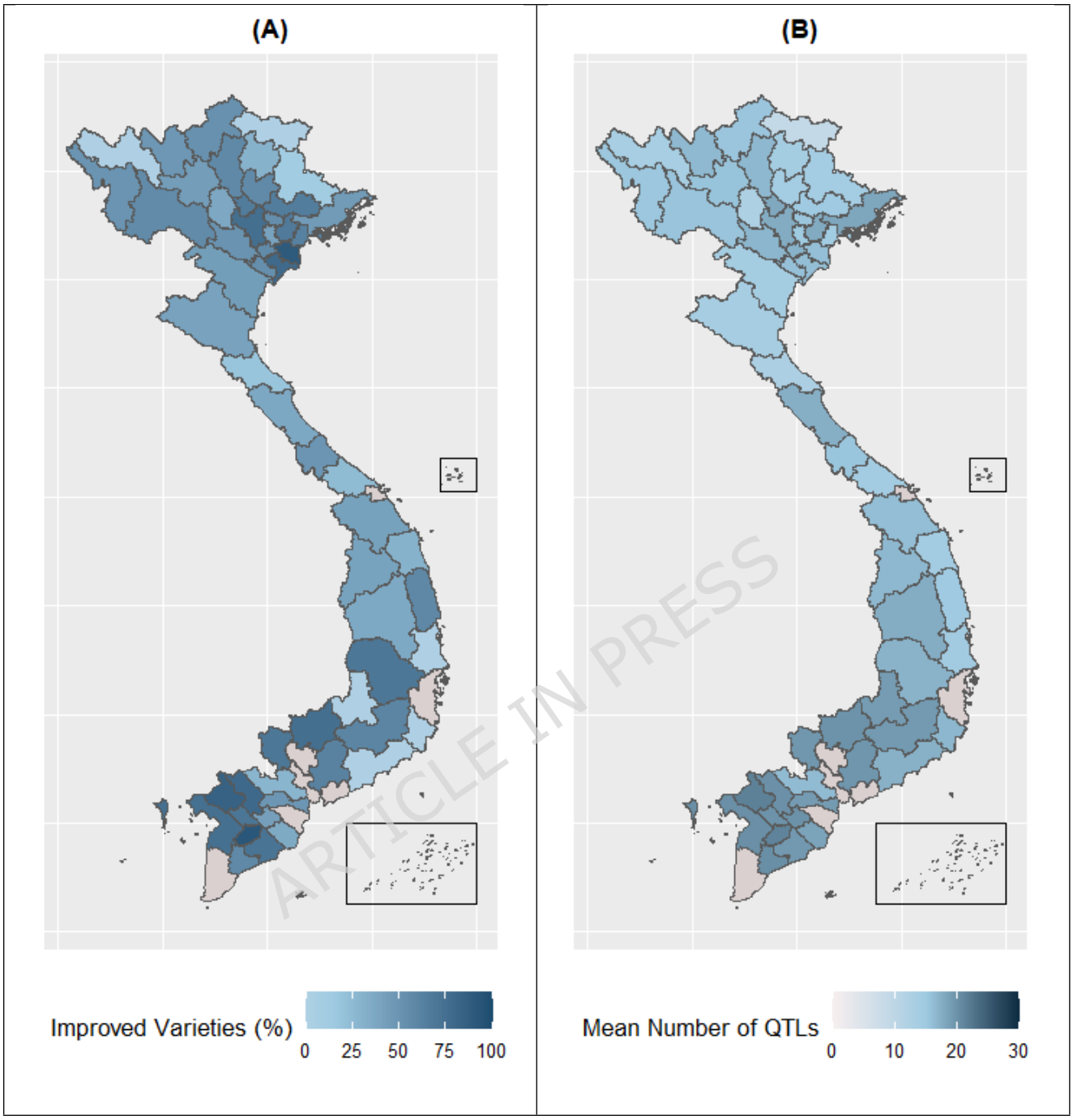


Figure 2. Spatial distribution of (A) improved rice varietal adoption and (B) the mean number of QTL count per rice variety in Viet Nam in 2022. Gray areas indicate provinces where no households grew rice at the time of rice sample collection. The map was generated using R version **4.2.3**. Source: VHLSS 2022.

At the national level, unassigned rice samples and those classified as improved varieties are present in roughly equal numbers, allowing for an effective comparison. In Table S3, we present the differences in QTLs between improved rice varieties and unassigned samples collected from farmers' fields. A t-test comparing the number of QTLs by seed status indicated that the average improved

rice variety that farmers grew had a higher number of QTLs (mean = 18.27, range = 9-20) than the unassigned samples (mean = 13.03, range = 10- 17), with the difference significantly different from zero (t = -27.41, df = 673.17, p < 0.0001). These descriptive patterns support the idea that improved rice varieties in Viet Nam possess more advantageous alleles, as expected from a successful breeding program.

Correlates of improved rice adoption

To understand the correlates of improved rice adoption, we analyzed household-level characteristics associated with adoption. Table 1 reports marginal effects from a multivariate probit regression.

Table 1. MVP estimates of the relationship between improved rice varietal adoption and a set of household's socio-economic characteristics.

	Marginal effects with 95% CI
Household head from an ethnic minority	-0.116 [-0.117, -0.115]
Household head is female	0.020 [0.019, 0.022]
Age of household head (years)	0.007 [0.007, 0.008]
Highest education level of household head	0.008 [0.007, 0.008]
Total agricultural land area (hectares)	0.062 [0.062, 0.062]
Commune classified as poor	-0.186 [-0.187, -0.184]
Main road surface is asphalt	-0.119 [-0.120, -0.118]
Observations	719
AIF	9,866,433

The marginal effects in **Error! Reference source not found.** indicate that adoption patterns differ across social groups, with all estimates significant at the p < 0.01 level. Belonging to an ethnic minority or residing in a classified poor commune substantially reduces the probability of adopting improved varieties. By contrast, higher education and age are positively associated. Although the pooled model suggests female-headed households have slightly higher adoption probabilities, an interaction analysis reveals this effect is spatially heterogeneous. The positive coefficient is primarily driven by the NCCCA and South East regions; however, the latter result warrants caution as it is based on a restricted subsample. Notably, the association of female headship correlates with lower adoption probabilities in the RRD and MRD regions.

Trait-related QTLs and Improved Rice Adoption

Given the correlation between farmer characteristics and adoption, we next examine whether adopters grew varieties with a broader suite of agronomic traits. To test this, we examine whether a higher density of trait-related QTLs in the rice variety grown is associated with an increased likelihood of adoption. We first analyzed the direct relationship between QTL count and varietal adoption at the national level (

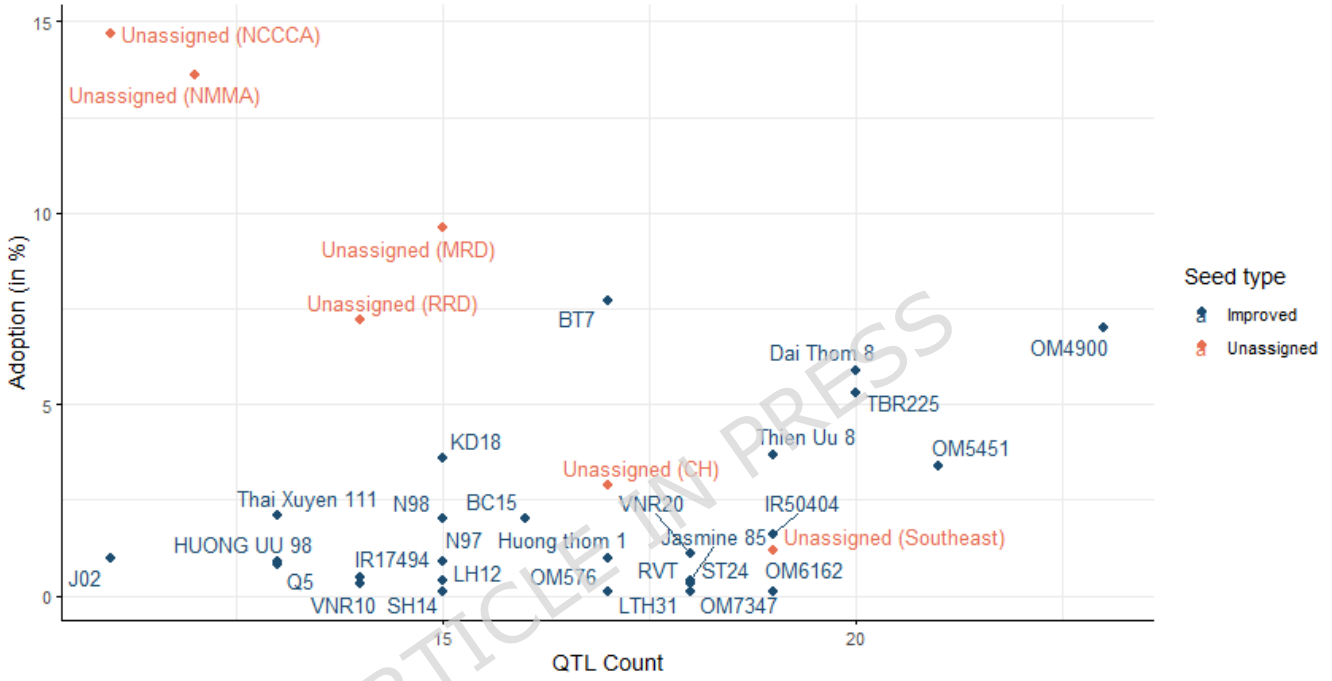


Figure 3). We then conduct regression analyses that progressively incorporate controls for potential confounding factors—first adding regional fixed effects and subsequently adjusting for province-level socioeconomic and infrastructure characteristics.

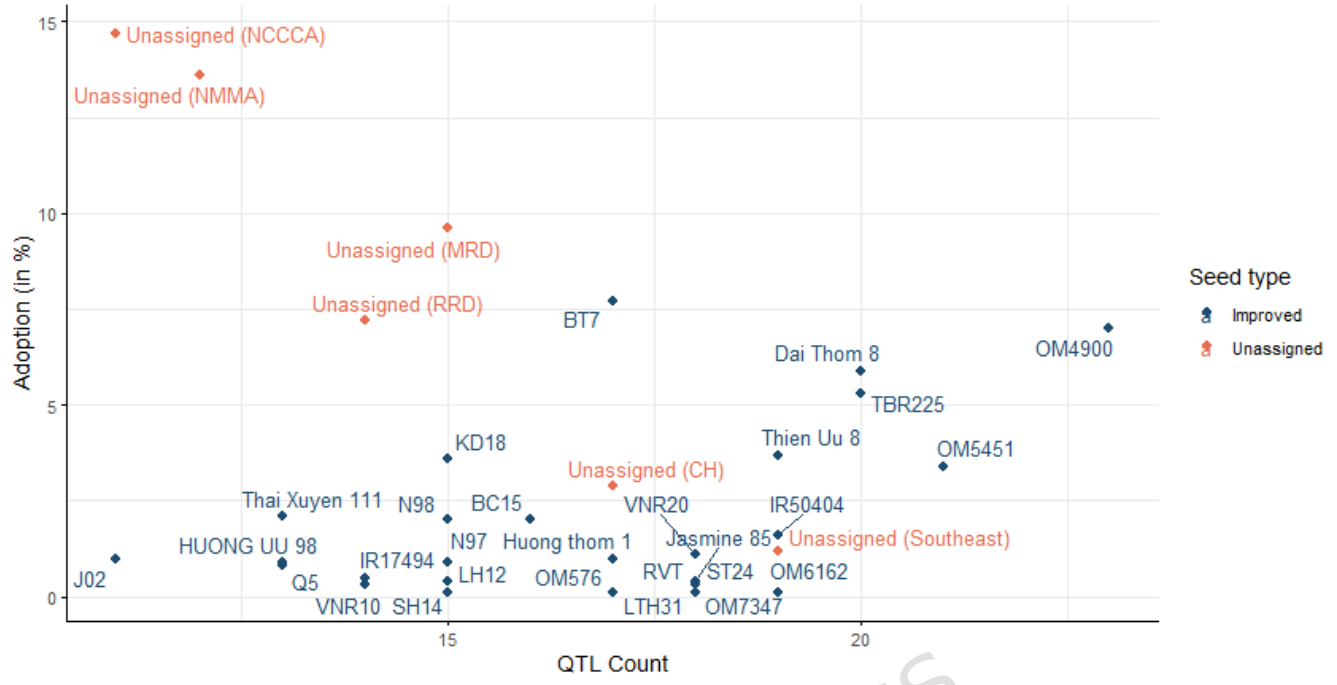


Figure 3. Patterns of Rice Adoption and Number of Trait-Related QTLs at the National Level in Viet Nam, 2022. The x-axis represents the number of QTLs, while the y-axis shows adoption rates at the national level. Unassigned samples were categorized according to their region of origin. NCCCA = North Central Coast and Central Highlands; NMMA = Northern Midlands and Mountainous Area; MRD = Mekong River Delta; RRD = Red River Delta; CH = Central Highlands; Southeast = Southeast Region.

ARTICLE IN PRESS

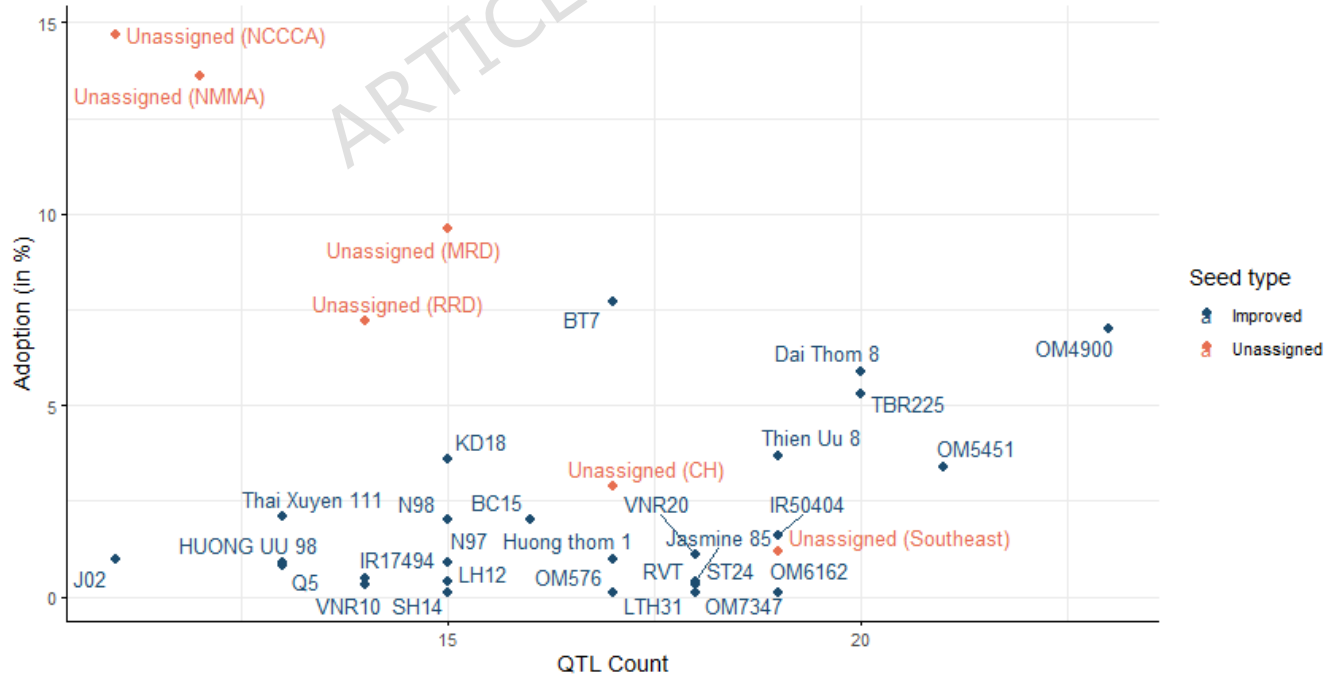


Figure 3 shows two clear patterns in the relationship between varietal adoption and the number of QTLs. In general, the unassigned samples tend to appear toward the left side of the x-axis, indicating that they possess fewer QTLs. Among unassigned samples, adoption rates are inversely correlated with number of QTLs. Despite this, they are found among a large proportion of the households surveyed in the Northern Central Coastal and North Midlands and Mountains regions (NCCCA and NMMA). These genetic entities possess a relatively small number yet diverse QTLs, such as *RFT1, Ehd1, Hd1*, and *Hd2* for flowering time and yield stability; *Chalk5, Alk, GS3, and DTH8* for grain quality; *Xa26* and *STV11* for biotic resistance; and *DRO1, qAG3*, and *HIS1* for abiotic stress tolerance. In contrast, improved varieties generally have a larger number of trait-related QTLs, reflecting greater genetic enrichment, but adoption rates tend to be lower, except for OM4900, BT7, and Dai Thom 8, which as noted earlier, are cultivated widely.

The relationship between province-level adoption rates and QTL count is further explored by aggregating the household-level data to the province level, controlling for regional fixed effects, and subsequently adjusting for province-level socioeconomic and infrastructure factors. The results from the most constrained model are shown in Figure 4. Overall, improved rice varieties with a higher number of QTLs display higher adoption rates at the province level. Each additional QTL in a variety was associated with a 0.9 percentage point increase in the provincial share of households adopting improved rice varieties. Across all three models, these associations are statistically significant (Table S4).

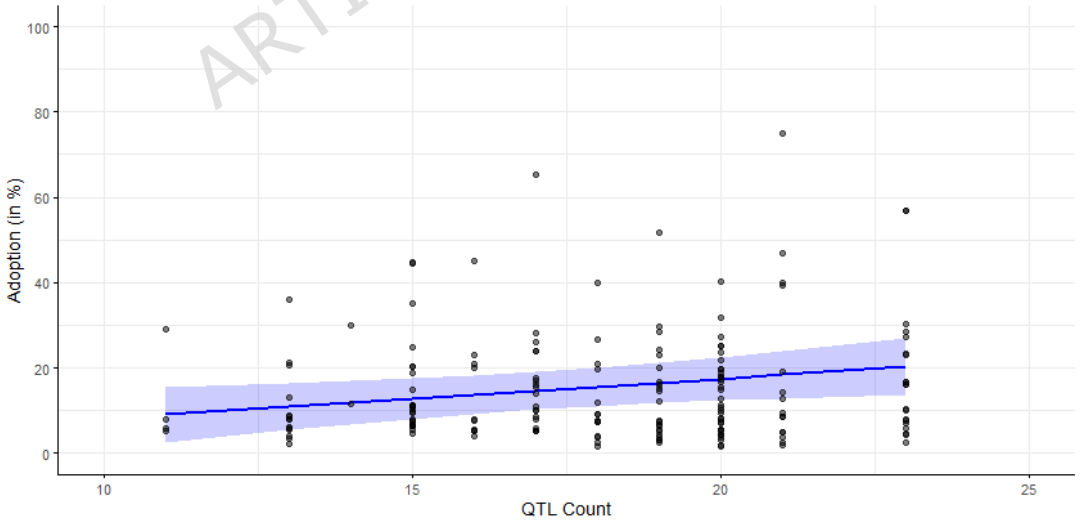


Figure 4. Marginal Effect of QTL count on Improved Varietal Adoption at the Province-Level. The adoption rate of an improved variety is measured as the percentage of farming households in each province that cultivate that variety. The linear regression model estimated the relationship between adoption rates and trait-related QTL count, controlling for

regional fixed effects and province-level covariates such as age of the variety, average education, household income, and infrastructure quality. The model was estimated using ordinary least squares (OLS) at the province level. $R^2 = 0.08$; $p < 0.05$.

Trait Preferences in Household Adoption

Farmers tend to grow improved rice varieties that exhibit a broader set of agronomic traits. Our final analysis examines whether Vietnamese farmers show preferences for specific traits. In Figure 5 we present the overlap in adoption percentages across specific trait categories in improved rice varieties. The results indicate that multi-trait adoption is primarily driven by yield, with joint adoption of yield and another trait category exceeding 60% in all cases. By contrast, quality traits and abiotic stress tolerance traits appear as complementary choices, often adopted alongside yield but less frequently on their own.

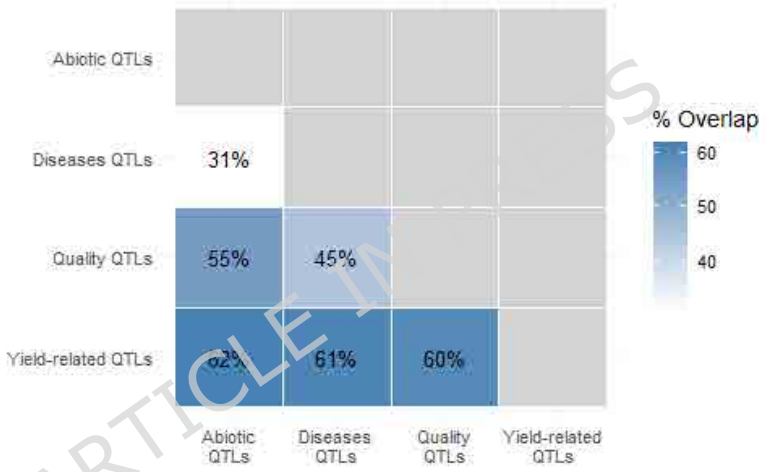


Figure 5. Percentage Overlap in Improved Rice Adoption among Four Key Rice Traits. For each trait, we used the median number of QTLs per variety as a threshold; only varieties with above-median QTL count for a given trait were considered in the corresponding column and cell of the cross-tab.

Building on these descriptive patterns, Table 2 reports results from a multivariate regression analysis assessing the association of different trait categories with improved rice adoption. This approach helps to isolate the relative importance of yield, grain quality, disease resistance, and abiotic stress tolerance traits in shaping varietal adoption, while controlling for household socioeconomic characteristics, infrastructure conditions, and regional fixed effects.

Table 2. Association Between QTL Categories and Adoption of Improved Rice Varieties at the Province Level (2022). Model 1 includes the total QTL Count as the primary independent variable. Model 2 incorporates regional fixed effects. Model 3 adds province-level covariates to the Model 2 specification. Standard errors are reported in parentheses below the

coefficients. NMMA = Northern Midlands and Mountainous Area; NCCCA = North Central Coast and Central Highlands; CH = Central Highlands; Southeast = Southeast Region. MRD = Mekong River Delta. Significance level: *p<0.1.

	% of Households Adopting		
Variable	**(1)**	**(2)**	**(3)**
QTL Count: Yield-related	0.999 (1.244)	1.327 (1.259)	1.352 (1.274)
QTL Count: Quality	-0.399 (1.853)	-0.528 (1.875)	-0.399 (2.078)
QTL Count: Disease	-0.301 (0.767)	0.118 (1.017)	-0.178 (1.066)
QTL Count: Abiotic	1.961* (1.155)	1.980* (1.175)	1.967* (1.183)
Region: NMMA		3.223 (2.731)	3.603 (3.113)
Region: NCCCA		0.456 (3.007)	1.685 (3.281)
Region: Central Highlands		4.258 (4.217)	5.268 (4.953)
Region: South East		7.177 (5.155)	4.81 (5.478)
Region: MRD		-1.691 (3.594)	-3.757 (4.153)
Age of Variety (in yrs)			0.089 (0.133)
Highest education level of household head			-1.064 (2.71)
Annual household income (million VND)			0.038 (0.028)
Main road surface is asphalt			-3.3 (3.85)
Constant	2.997 (7.979)	-0.816 (8.263)	-4.702 (10.465)
Observations	178	178	178
R-squared	0.022	0.060	0.076

Abiotic traits are the only category showing a consistent positive association with adoption across Models (1) and (2) (~1.96, p = .1).

The coefficients remain positive and significant in Model (3), after controlling for socioeconomic and infrastructure variables. This finding suggests that farmers may place greater value on traits that enhance resilience to environmental stress. However, the wider confidence interval for abiotic traits (-2.18; 2.49) likely reflects

environmental heterogeneity; these traits may be valued in stress-prone provinces but may be neutral in others. Notably, while the aggregate count of abiotic stress QTLs is a significant predictor, an analysis of individual QTLs (e.g., specific drought or salinity markers) yielded no statistically significant results (Table S7). In contrast, other breeding targets—including yield, grain quality, and disease resistance—showed no significant association with adoption in any model.

Discussion

[2,29]The observed effect sizes of QTL count in this study align with existing literature, suggesting that while individual loci may have small to moderate impacts on complex traits, their cumulative stacking enhances crop performance[30,31]. Our findings confirm that targeted introgression is successfully reaching farmers' fields in Viet Nam, as evidenced by the presence of improved varieties containing these beneficial genetic markers. While QTL effect sizes can theoretically be subject to heterogeneity, our study operates on the empirical observation that the loci explored are consistently desirable across diverse production contexts. In rice breeding, evidence for significant epistasis or environmental interactions that negate the primary effect of established genes is remarkably limited; generally, these genes maintain stable expression across various genetic backgrounds and ecological conditions [32,33].

When placed in a regional context, Viet Nam's adoption landscape appears more dynamic than that of its neighbors. Previous DNA-based studies in Bangladesh have reported lower rates of improved variety adoption; for instance, [34] found that only 21% of households grew varieties that matched breeders' seed reference samples during the rainfed (Aman) season. More recent data from [35] indicates that while adoption has risen to 39% in the Aman season and 59% in the irrigated Boro season, the average varietal age remains between 21 and 24 years. This is significantly older than the 14-year average observed in Viet Nam. Furthermore, a defining characteristic of the Bangladeshi rice sector is the persistence of mega-varieties such as BRRI dhan29. These varieties, released in the 1990s, continue to dominate the landscape decades later. In contrast, our study found no such dominance of legacy mega-varieties in Viet Nam, suggesting a more fluid varietal turnover.

These findings carry significant policy implications; reducing the average varietal age has become a strategic priority for major agricultural economies seeking to accelerate genetic gain in the field. The Government of India, for instance, has codified this through the National Food Security Mission, which restricts subsidies for cultivars exceeding 10 years of age to incentivize the adoption of modern, climate-resilient alternatives[36]. Ultimately, a seed system's ability to deliver up-to-

date cultivars—combined with the integration of genetic tracking—is essential to ensuring that farming communities benefit from the latest agricultural innovations.

Although improved varieties with higher QTL densities show higher overall adoption rates, their distribution remains uneven across regions and socioeconomic groups. Our findings reveal that ethnic minority households and smallholders in government-classified poor communes are significantly less likely to access or adopt these cultivars. Climate change exacerbates this inequity, as these populations often reside in areas equally exposed to adverse weather events, yet they lack the genetic tools to mitigate such risks. These results point to an adoption gap possibly caused by restricted seed access, credit constraints, or a potential mismatch between high-performing varieties and the specific needs of marginalized environments. Without targeted interventions, genetic innovations risk bypassing the most vulnerable farmers.

To bridge this gap, a shift toward decentralized, farmer-centric models may prove effective. Participatory Varietal Selection (PVS) has been proposed as a critical solution to ensure that QTL pyramiding aligns with the specific trait preferences demanded by resource-constrained producers[37,38]. By involving farmers in the selection process, breeding programs can validate varietal performance under the marginal conditions typical of poor communes. Furthermore, the establishment of Village-Based Seed Enterprises (VBSE) offers a sustainable mechanism for reaching remote areas where commercial seed companies find operation unprofitable. As highlighted by [39], VBSEs can be effective at introducing stress-tolerant varieties. This strategy not only eliminates the high transportation costs that render improved seeds prohibitively expensive but also creates local income-generating opportunities.

A central question in this study is whether farmers have begun to prioritize environmental resilience over traditional productivity. Our findings provide only suggestive evidence that farmers may favor resilience traits; and this preference appears secondary to entrenched market demands. In Viet Nam, and specifically the Mekong River Delta, yield and grain quality likely remain the well-established pillars of variety selection. Consumer preferences are heavily weighted toward specific sensory attributes—such as slender shape, soft texture, and aroma [20,40]. Consequently, while the literature indicates that farmers value traits like salinity tolerance, they often face a difficult quantity-quality trade-off. Many current stress-tolerant varieties command lower market prices because their grain quality fails to match these established consumer standards[15]. This economic barrier suggests that the weak evidence for resilience preference found in our study may reflect a market constraint rather than a lack of interest: farmers cannot afford to prioritize abiotic stress tolerance unless it is bundled with the high-quality characteristics the market demands.

This study has several limitations that should be considered when interpreting the findings. First, the sampling frame restricted our ability to fully investigate seasonal patterns. In Viet Nam, particularly the Mekong River Delta, regions may experience up to three cropping seasons annually, each with distinct requirements for growth duration and biotic stress resistance. Consequently, our results may not capture the full diversity of trait preferences across different times of the year.

Furthermore, there is a risk of misclassification bias that complicates comparisons with improved varieties. Although our reference library was designed to be comprehensive—encompassing all released cultivars of known importance—it remains possible that some field samples represent improved varieties not currently included in our database. This lack of genomic information could lead to some degree of misclassification. Another limitation is the high level of heterozygosity observed, affecting 30% of field samples, which may reflect the presence of hybrid varieties, accidental admixture within plots, or contamination by weedy rice. The potential impact of heterozygosity on downstream QTL profile analysis was mitigated by aggregating genomic data across field samples assigned to a particular variety. This way, any potential noise from individual plot level admixture was minimized ensuring that the QTL count reflected the representative genetic makeup of the varieties in circulation.

Finally, our results represent a blended reflection of farmer preferences and seed accessibility. Because farmer choices are constrained by the availability of seeds in local markets, it is difficult to isolate inherent trait preferences from the logistical realities of seed supply. Therefore, the observed variety distribution likely indicates what is available and chosen by farmers, rather than a clean expression of farmers' demand.

Future research should explore the dissemination pathways of the on-farm performance of marker-assisted varieties that have not yet gained widespread traction, such as AS996 (released 2002) or HDT8 (released 2011). Longitudinal studies tracking varietal adoption over multiple seasons and environments could clarify seasonal effects. This study is unique in providing a molecular-level perspective on genetic gains in rice, offering a model that can be replicated in other rice growing nations as well as to other crops using established molecular platforms. However, replicating the approach requires investment in construction of robust reference libraries of locally released varieties based on breeder seed for accurate varietal assignment. Also, QTL profiling should be done with care to ensure they reflect local realities such as quality traits desired by markets. Leveraging these genomic resources could provide broader insights into genetic improvements in smallholder farms.

Conclusion

This study offers novel insights into the genetic improvements achieved in farmers' fields by examining the presence of trait-related QTLs in rice varieties cultivated across Viet Nam. Our results validate that improved varieties carry significantly more QTLs related to yield, grain quality, and stress resistance than the unassigned samples, supporting the effectiveness and scaling of targeted introgression through breeding programs. However, adoption remains uneven; socioeconomic constraints—primarily ethnic background and geographic location—were negatively associated with improved varietal adoption. Among those who have adopted improved rice, a higher concentration of beneficial QTLs strongly correlates with variety adoption, with suggestive evidence that farmers may favor traits that bolster resilience against environmental stressors. These results illustrate the power of integrating DNA-based varietal identification with socioeconomic data, offering a sound methodology for monitoring the impact of agricultural research and ensuring the equitable distribution of genetic gains.

Data availability

All data and associated R scripts are stored in the OpenICPSR Repository: https://www.openicpsr.org/openicpsr/project/239565 under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

References

1. IPSARD. *Rice Annual Report*. https://thitruongnongsan.gov.vn/images/2013/DANGNHAP/VnSAT_Lua%20gao/2023/N%C4%83m/Bao%20cao%20thuong%20nien%20lua%20gao%202023%20-%20EN.pdf (2023).

2. Collard, B. C. Y. & Mackill, D. J. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**, 557–572 (2008).

3. Geldermann, H. Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. *Theoretical and Applied Genetics* **46**, 319–330 (1975).

4. Broman, K. W. & Sen, S. *A Guide to QTL Mapping with R/Qtl*. (Springer, New York, 2009).

5. Paterson, A. H. What has QTL mapping taught us about plant domestication? *New Phytologist* **154**, 591–608 (2002).

6. Miles, C. & Wayne, M. Quantitative trait locus (QTL) analysis. *Nature Education 1(1):208* (2010).

7. Muthu, V. *et al.* Pyramiding QTLs controlling tolerance against drought, salinity, and submergence in rice through marker assisted breeding. *PLoS One* **15**, (2020).

8. Khanh, T. D., Duong, V. X., Nguyen, P. C., Xuan, T. D. & Trung, N. T. Rice Breeding in Vietnam: Retrospects, Challenges and Prospects. *Agriculture* **11**, 1–21 (2021).

9. Pham, V. P., Huu, M. S. & Vo, C. T. Researching and selecting high quality rice varieties in Mekong Delta. *Can Tho Uni. J. Sci.* **15b**, 97–105 (2010).

10. Khoa, T. *et al.* Pyramiding the Candidate Genes of Rice Bacterial Leaf Blight Resistance xa5, Xa7 and xa13 into the Elite Rice Variety. *Journal of Scientific and Engineering Research* **4**, 92–98 (2017).

11. Nguyen Thi Pha & Nguyen Thi Lang. Marker Assisted Selection in Rice Breeding for Bacterial Leaf Blight. *Omonrice* **12**, 19–26 (2004).

12. Luu M.C. *et al.* Application of Marker Assisted Backcrossing to Introgress the Submergence Tolerance QTL Sub1 into the Vietnam Elite Rice Variety-AS996. *Am. J. Plant Sci.* **03**, 528–536 (2012).

13. Thị, V. *et al. Assessment of Genetic Diversity in Vietnamese Rice Using SSR Markers to Identify Cross Combinations for Developing Drought-Tolerant Cultivars*. (2013).

14. Khanh, T. D., Trung, K. H. & Hoi, P. X. Future perspectives and some approaches to develop rice in Vietnam. in *Biotechnology and Perspectives of Applications in Rice Breeding in Vietnam* (ed. Hanoi National University) 671–701 (Hanoi, 2018).

15. Paik, S. Y., Le, D. T. P., Nhu, L. T. & Mills, B. F. Salt-tolerant rice variety adoption in the Mekong River Delta: Farmer adaptation to sea-level rise. *PLoS One* **15**, 1–23 (2020).

16. Stevenson, J., Macours, K. & Gollin, D. The Rigor Revolution: New Standards of Evidence for Impact Assessment of International Agricultural Research. *Annu. Rev. Resour. Economics* **15**, (2023).

17. Stevenson, J., Gantier, M., Traxler, G. & Kosmowski, F. *The Challenge of Tracking the Reach of Post-Green Revolution Crop Breeding*. Preprint. DOI: 10.21203/rs.3.rs-3028333/v1 (2023).

18. Sandhu, N. *et al.* Positive interactions of major-effect QTLs with genetic background that enhances rice yield under drought. *Sci. Rep.* **8**, (2018).

19. Isnaini, I., Nugraha, Y., Baisakh, N. & Carsono, N. Toward Food Security in 2050: Gene Pyramiding for Climate-Smart Rice. *Sustainability* **15**, (2023).

20. Bairagi, S., Demont, M., Custodio, M. C. & Ynion, J. What drives consumer demand for rice fragrance? Evidence from South and Southeast Asia. *British Food Journal* **122**, 3473–3498 (2020).

21. Euler, M., Krishna, V. V, Jaleta, M. & Hodson, D. Because error has a price: A systematic review of the applications of DNA fingerprinting for crop varietal identification. *Outlook Agric.* **51**, 384–393 (2022).

22. Poets, A., Silverstein, K., Pardey, P., Hearne, S. & Stevenson, J. *DNA Fingerprinting for Crop Varietal Identification : Fit-for-Purpose Protocols and Their Cost and Analytical Implications*. (2020).

23. Arbelaez, J. D. *et al.* 1k-RiCA (1K-Rice Custom Amplicon) a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice. *Rice* **12**, (2019).

24. Thomson, M. J. *et al.* Large-scale deployment of a rice 6 K SNP array for genetics and breeding applications. *Rice* 10:40 (2017) doi:10.1186/s12284-017-0181-2.

25. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

26. CGIAR. Rice mid-density genotyping service. *https://excellenceinbreeding.org/toolbox/services/rice-mid-density-genotyping-service* (2025).

27. McCouch, S. R. Gene nomenclature system for rice. *Rice* **1**, 72–84 (2008).

28. R Core Team. R: A Language and Environment for Statistical Computing. Preprint at (2024).

29. Mackay, T. F. C. The Genetic Architecture of Quantitative Traits. *Annu. Rev. Genet.* **35**, 303–339 (2001).

30. Paterson, A. H. *et al.* Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**, 721–726 (1988).

31. Muthu, V. *et al.* Pyramiding QTLs controlling tolerance against drought, salinity, and submergence in rice through marker assisted breeding. *PLoS One* **15**, (2020).

32. Zuo, J. & Li, J. Molecular Genetic Dissection of Quantitative Trait Loci Regulating Rice Grain Size. *Annu. Rev. Genet.* **48**, 99–118 (2014).

33. Wing, R. A., Purugganan, M. D. & Zhang, Q. The rice genome revolution: from an ancient grain to Green Super Rice. *Nat. Rev. Genet.* **19**, 505–517 (2018).

34. Kretzschmar, T. *et al.* DNA fingerprinting at farm level maps rice biodiversity across Bangladesh and reveals regional varietal preferences. *Sci. Rep.* 1–12 (2018) doi:10.1038/s41598-018-33080-z.

35. Singla, S. *et al. SPIA Study 2025: Updating the Green Revolution*. https://iaes.cgiar.org/sites/default/files/pdf/SPIA_Bangladesh_Study_2025.pdf (2025).

36. NNFSN. *Operational Guidelines: National Food Security and Nutrition Mission*. (2024).

37. Sperling, L. *et al.* A framework for analyzing participatory plant breeding approaches and results. *Euphytica* (2001).

38. Nguyen, N. De & Kotaro Ohara. Participatory approaches for improving rice breeding in the Mekong Delta of Vietnam. in *Rice is life: scientific perspectives for the 21st century* (CABI, 2005).

39. MANZANILLA, D. O. *et al.* SOCIAL AND GENDER PERSPECTIVES IN RICE BREEDING FOR SUBMERGENCE TOLERANCE IN SOUTHEAST ASIA. *Exp. Agric.* **50**, 191–215 (2014).

40. Demont, M. & Rutsaert, P. Restructuring the Vietnamese rice sector: Towards increasing sustainability. *Sustainability (Switzerland)* **9**, (2017).

Acknowledgements

We gratefully acknowledge the support of the General Statistics Office (GSO) of Viet Nam and the Ministry of Agriculture and Rural Development (MARD) during the period of research and data collection. We note that, following government restructuring in 2025, the GSO has since become the National Statistics Office (NSO), under the Ministry of Finance, and MARD has been merged into the Ministry of Environment and Agriculture (MEA). We thank Bùi Chí Bửu for very helpful comments.

Author contributions

F.K., S.V., J.S., D.G. and J.P. contributed to the conception and design of the work and to the interpretation of data. F.K., D.G. and J.P. performed the analysis. F.K. drafted the manuscript. S.V., J.S., D.G. and J.P. reviewed and revised the manuscript. All authors approved the final version of the manuscript.

Funding information

The authors gratefully acknowledge funding from the CGIAR System Council for the Viet Nam Country Study.

Competing interests

The authors declare that they have no competing interests.

Improved rice varieties in Vietnam

A new study, published in *Nature Scientific Reports* by SPIA colleagues, sheds insight on which CGIAR innovations are actually reaching farmers.

CGIAR STANDING PANEL ON IMPACT ASSESSMENT

Rice is crucial to Vietnam's economy and food security.

Since the early 1990s, molecular breeding has played an increasingly important role in **enhancing the traits in rice varieties** that determine growth rates, yield, and adaptability. These include resistance to pests and disease, as well as tolerance to environmental stressors such as salinity, submergence, and drought.

But what varieties are farmers actually growing?

It’s a surprisingly difficult question to answer.

Variety names can get swapped, mixed, and renamed along the supply chain, making name-based identification unreliable. For CGIAR, which heavily invests in improved varieties, **this question is crucial.**

So the SPIA team turned to the rice itself.

They carried out the first nationwide analysis of Vietnamese rice varieties, sequencing the DNA of samples pulled from **766 rice-growing households** across Vietnam (sampled through the Vietnam Household Living Standards Survey) and matching the genetics against a reference library of major varieties released in the country.

What information emerged?

51% of the households grew improved rice varieties.

28 distinct varieties were identified, showing significant heterogeneity.

No variety held more than 8% of the share.

The most widely cultivated varieties were BT7, OM4900 and Dai Thom 8.

Unlike in neighboring Bangladesh, there are no “megavarieties” dominating Vietnam.

Varietal turnover is dynamic: the average varietal age is **14 years**, compared to 21-24 years in Bangladesh.

Access to improved varieties is uneven.

Ethnic minority households and smallholders in poor communies are significantly less likely to adopt.

Want to learn more?

Download the report to read more about what the team found.

CLICK THE LINK IN THE POST.

CGIAR
STANDING PANEL ON IMPACT ASSESSMENT

Identifying farmers' crop varieties using DNA fingerprinting: A guidebook

Davis Gimode, Frederic Kosmowski, James Stevenson

CGIAR Standing Panel on Impact Assessment

Version 2.0 for peer-review: April 21st 2026

Comments / feedback on this version please, to j.stevenson@cgiar.org

Contents

Introduction

This guidebook addresses the critical need for accurate crop varietal adoption data in developing countries to improve agricultural research and policies. Traditional farmer self-reporting methods have proven unreliable due to factors such as informal seed systems, visual similarities among improved varieties, and the use of inconsistent local names. To address these issues, this guidebook champions the integration of DNA fingerprinting into household surveys as a more objective and reliable approach to identifying crop varieties.

The guidebook provides a practical overview of DNA fingerprinting, including the necessary stages (from sample collection to data analysis), required skills (emphasizing collaboration among researchers with diverse expertise), and detailed cost considerations (covering survey-related, lab-related, and technical expertise costs). It stresses the importance of compiling a comprehensive and accurate reference library using breeders' seed whenever possible. Drawing on empirical studies from Ethiopia, Uganda, Viet Nam and Bangladesh across various crops (maize, beans, rice, groundnut, banana, sweet potato and cassava), the guidebook offers valuable insights for economists, social scientists, and agronomists seeking to enhance the quality of crop varietal adoption data. While acknowledging the complexities and costs associated with DNA fingerprinting, it emphasizes its potential to reduce measurement error and provide more accurate data for informed decision-making in agricultural development. The guide ultimately serves as a non-technical resource for researchers looking to design surveys and collaborate effectively to implement DNA fingerprinting methodologies.

We assume that the audience for this guidebook will be implementing household surveys using Computer-Assisted Personal Interviewing (CAPI) on tablets. We do not address the specifics of the programming of the broader questionnaire, but note that a CAPI survey opens the door to having a barcode scanner embedded in it that can be switched on at the appropriate time in the interview process, in order to capture a barcoded sample of plant tissue.

Some of the areas of covered by the guidebook have good practice that is well-established and uncontroversial. For those we make recommendations. Other aspects are still subject to ongoing methodological research to understand the potential implications of different sampling or bioinformatic approaches as DNA fingerprinting is not “one size fits all”. The specifics of the method should be tailored to the crop and context. We have tried to give a sense of the underlying biological or behavioral factors that could be in play to convey some of this complexity but a guidebook can only go so far. Do not hesitate to reach out to SPIA for advice through research design, fieldwork and data analysis.

Section 1. Understanding DNA fingerprinting

KEY POINTS

- DNA fingerprinting of crop varieties involves taking samples from the growing crop in farmers' fields
- There are five main stages in a DNA fingerprinting study and all need to be executed correctly to ensure high quality data on crop varietal adoption

DNA as Data

Cells of living organisms contain DNA (deoxyribonucleic acid) molecules which are condensed, organized and packaged into chromosomes. The basic building blocks of DNA are nucleotides or bases denoted by the letters "A", "T", "G" and "C"[1]. The bases are arranged in a sequence, linked by a sugar-phosphate backbone. Each base pair is around 0.34 nanometers[2] long (Annunziato, 2008). The wheat genome, for instance, is a book that is 16.5 billion letters long, but containing only four letters. The bases are grouped together to form genes, which are the informational basis of heredity. Genes consist of DNA segments that are transcribed and translated to amino acids; the chemical building blocks of proteins and that ultimately enable form and function in biology. For example, the sequence "CAT" encodes the amino acid Histidine; "ACT-CAT-GGT" encodes the amino acid sequence of Threonine-Histidine-Glycine. Thus, by this means, DNA instructs cells about which proteins to make.

When used for the purpose of identifying crop varieties, DNA fingerprinting uses markers called Single Nucleotide Polymorphisms (SNPs). SNPs are specific points in the genome where variations – polymorphisms – in the DNA sequence can pinpoint differences when comparing genomes from different individuals. SNPs may fall within coding sequences of genes, non-coding regions of genes, or between genes.

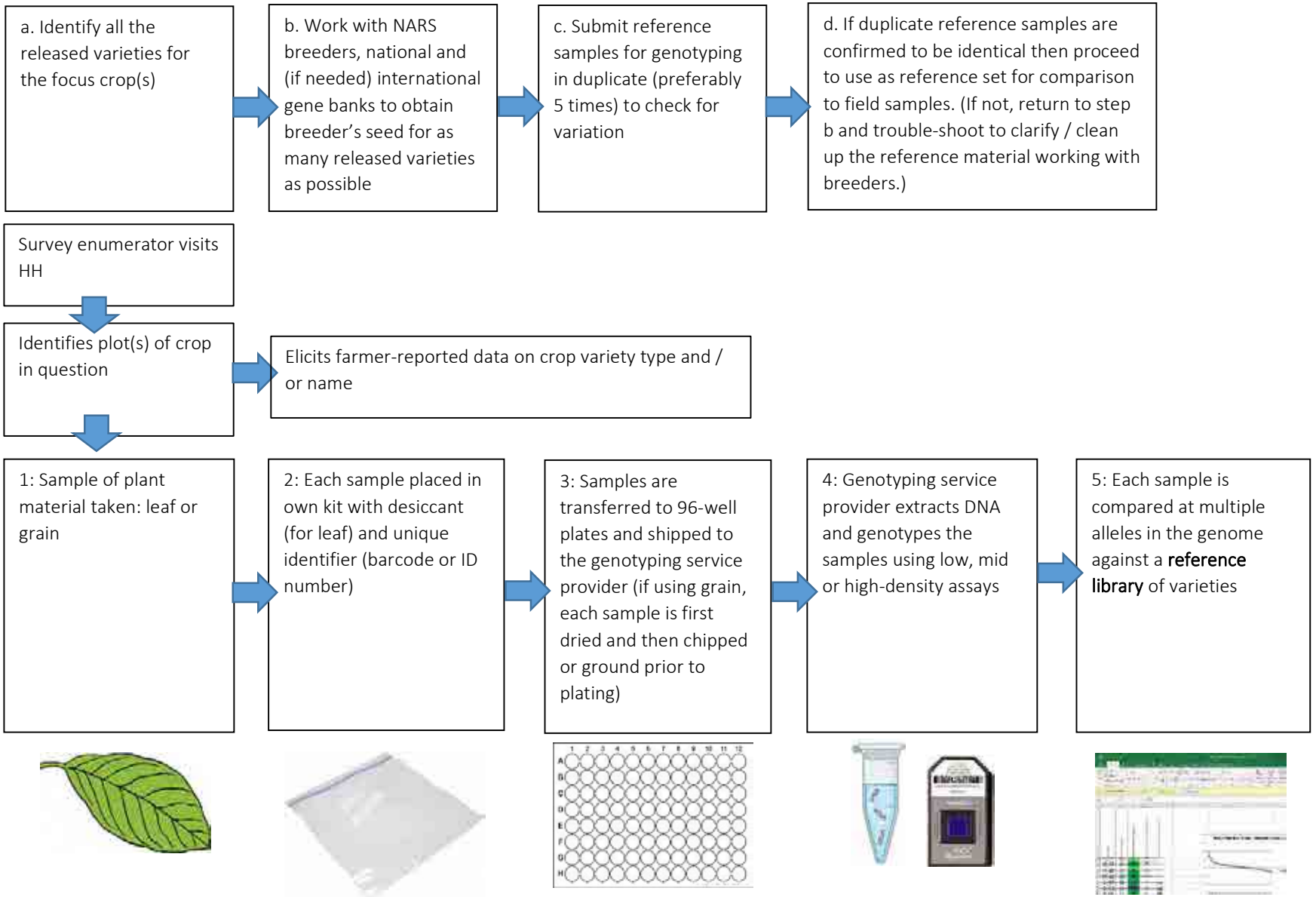
Modern plant breeding makes extensive use of genomics, allowing the identification of specific markers that are believed to favor the expression of specific traits. For example, marker-assisted selection helps breeders to identify plants that have targeted alleles for specific traits, and genome-wide association studies scan the entire genomes of plants to find variations with specific traits, thereby allowing for the discovery of quantitative trait loci (QTLs) for complex, multi-genic traits.

Key Steps in a DNA Fingerprinting Study

DNA fingerprinting for crop varietal identification is the process through which information from the genome of sampled plants is matched to a reference library of genetic profiles of plants of known identity (i.e. improved varieties). A match is made between each field sample and its closest reference sample, based on the degree of genetic similarity. If a match cannot be made, within a specified tolerance, then the field sample may be returned as "unassigned" or "mixed".

[1] DNA is composed of adenine, guanine, cytosine and thymine (A, G, C, T); RNA is composed of adenine, guanine, cytosine and uracil (A, G, C, U).
[2] Or 3.4e-8 centimeters

Figure 1. Parallel stages in DNA fingerprinting study of crop varietal identity: reference library construction (steps a – d) and field sampling (1 – 5)



Cost, Capacities and Technical Requirements

Consider three sources of additional costs to your household survey from integrating DNA fingerprinting: survey-related costs, lab-related costs, and technical expertise costs. We provide a detailed breakdown of three examples in but here discuss more generally about how to think about costs for DNA fingerprinting.

Survey-related costs

The cost of fielding a standard multi-module household survey can range between $50 to $200 USD per household. One of the first order sources of additional costs of integrating DNA fingerprinting is the need to visit plot(s) cultivated by the household. If the survey would not otherwise be collecting plot-level data, then this can have a significant impact on survey operations and costs. It is for this reason that we have targeted surveys that would otherwise be collecting plot-level data such as plot boundaries, and possibly even carrying out crop-cuts. Indeed, if crop-cutting is already planned, then the marginal survey-related costs of integrating DNA fingerprinting are negligible – a sub-sample of the dried and weighed grain sample can be diverted to a genotyping pipeline.

If, however, a separate sampling process is required then the enumeration team must visit the plot, follow a protocol for sampling randomly from the plant material within the plot, and administer to the logistics of handling the plant material in the appropriate way. It can take an additional 30 – 60 mins per plot, at least initially. The survey team may speed up over time as they become more familiar with the routine.

Depending on the protocol for sampling the plant material, you may need to account approx. 1-2 USD per sample for consumables used by the enumerators in the process of sample collection, such as: silica gel, screw-top pots, barcodes, gloves, ethanol-based wipes and leaf-punches.

Lab-related costs

There are two kind of lab-related costs – at a local lab, and at the destination genotyping service provider lab. At the local lab, you will need to pay for the time of technicians trained in the laborious task of **initial sample processing**. On a per-sample basis, this is not a major cost as many samples can be processed per hour. It helps to have the same people working across the whole process, and you should also budget to cover some days of time for a senior scientist at the institution who can help train and supervise their staff.

Extracting the DNA from the plant tissue can sometimes be carried out as part of the initial lab work in-country, but where possible it is better to bundle it with the contract with the genotyping service provider. The fewer steps that need to be followed, the lower the chance of tracking errors, degradation or contamination of samples. So long as the correct plasticware (i.e. 96-well plates) is used for submission of samples, service providers should have machinery for automating DNA extraction. DNA extraction per sample costs approx. $2 USD.

The major lab-related costs are for the **genotyping assay**. The costs vary according to the density (i.e. the number of SNPs used to look for polymorphisms that characterize specific varieties), between a few dollars per sample for low-density assays, around $10 USD for mid-density assays, and $25 – 30 or more for high-density assays. As the technology continues to develop and improve these costs per sample continue to fall.

Technical expertise costs

This is the hardest cost to approximate , and depends on the nature of the research team that can be assembled. Table 1 outlines the skills needed, some of which can be brought in through collaboration and the prospect of co-authorship of the resulting work. In some cases, you will need to hire consultants which can push up the cost considerably.

Table 1: Human resources considerations for a DNA fingerprinting study

Who	Role	Key skills
Genotyping service providers (GSPs)	• Private sector business or public / academic research laboratories, providing assays under contract • Can be bundled with DNA extraction from plant tissue at low marginal cost • Advice on tissue sampling protocol and compatible plasticware for submission of samples • May play role in supporting analysis and interpretation of output	Technical knowledge of genotyping assays Helps tremendously if they are good communicators to lay people
Local laboratory in country where survey is implemented	• Receiving samples from the field and preparing them for submission to genotyping laboratory • Plating of samples and may carry out DNA extraction if deemed necessary • Possible secondary, complementary analyses of the plant material	Diligence and attention to detail to avoid tracking errors
Specialists in the crop in question	• Insights about how best to compile the reference library • Advising on field sampling plans – in terms of plot-level protocol and sample design • Interpretation of output	Knowledge of the crop biology and common farmer practices in the survey context
Bioinformaticians (not always necessary)	• If service provider does not provide output in a way that can already be interpreted by the team, it may be necessary	Ability to make genotyping "calls" based on genetic distance matrices and other outputs from genotyping
Economist / social scientists	• Responsible for maintaining link between plant samples and corresponding survey data (i.e. barcoding and CAPI system) • Training and supervising enumerators • Analysis of output and report compilation	Ability to integrate the needs of the plant sampling protocol into the other logistical operations of the survey

For a real-world example of how fingerprinting was incorporated in an existing household survey, the Ethiopian Socioeconomic Survey (ESS 2018/19), we detail the logistical implications in Appendix A. We give a full breakdown of the marginal costs associate with incorporating DNA fingerprinting in the survey in Appendix B.

Section 2. Why Use DNA fingerprinting to Identify Crop Varieties?

KEY POINTS

- Estimating adoption of crop varieties using DNA fingerprinting is less prone to biases
- Integrating DNA fingerprinting within household surveys allows us to understand both what the farmer thinks they are growing and what they actually have in their plots – both are important for analysis

Why Better Data on Varietal Adoption Matters

The adoption of improved crop varieties is strongly associated with the process of modernizing the agricultural sector and, with it, the development of rural areas across the globe. Agricultural research to develop improved varieties, and government programs to promote the production and dissemination of improved seed, are central to the approach of many developing country governments, donor agencies, and multilateral organizations to meeting Sustainable Development Goal 2 ("End hunger, achieve food security and improved nutrition and promote sustainable agriculture").

Data at a national level on crop varietal adoption are scarce. The data that are available are typically based on highly imperfect methods. Some degree of measurement error is inevitable in all realms of applied economic analysis - everything is measured with error, and economists are accustomed to working with imperfect data. However, in the case of self-reported survey data from farmers on their adoption of improved crop varieties, there are particularly strong reasons for expecting a high level of measurement error. Consider the following.

First, the way in which farmers obtain seed (used hereon to refer to both seed and planting material for clonally propagated crops, unless specified) for many crops is **highly informal**, particularly in sub-Saharan Africa. The informality is such that to even conceptualize farmers as procuring a specific "variety" can often become somewhat dubious. When farmers do make commercial purchases of seed, there is heterogeneity across countries as to seed quality they can expect, including whether what they purchase is indeed the variety they think they are getting (see www.tasai.org for an overview).

Second, the improved varieties released in Asia in the Green Revolution of the 1970s had new traits, such as being "semi-dwarf", that were visibly distinct and different from the prior landraces – the plants put their energy into developing and filling their cereal grains, rather than growing long and lanky (and in doing so, often falling over with wind or rain). Later generations of improved varieties have often had **no obvious visual signature** by which they can be identified with any confidence. That's not to say that the improved varieties are not substantially different from one another in terms of economically important traits such drought tolerance, nutritional content, tolerance to specific pests or diseases. Rather, that these traits of more recent improved varieties are explicitly bred into a genetic background that is based on a pool of prior improved varieties, making distinguishing them that much harder.

Third, we know that crop varieties are given **local names** to stand in for their clunky official names. Local variety names are typically catchy and more salient to the population of farmers that invent them. However, in this process we lose the fidelity to the underlying genotype along the way, with the same local name being given to multiple different varieties, and vice versa – the same variety may have multiple different local names. There may also be sub-national variation in how these names map to genotypes.

Empirical evidence is at hand to support these concerns. Stevenson et al (2023) report on the findings from a series of studies to help us understand the extent of the problem, summarizing the results from 16 different empirical studies in which self-reported data were collected from farmers for the same plots as plant samples were taken for subsequent DNA fingerprinting. With some degree of variation across cases, on average farmers report the correct

"type" of variety only 71% of the time. Correct matches regarding the specific name the farmer reports are only 24% of observations.

Furthermore, these errors could be different for farmers with different levels of skill, knowledge, wealth etc. Such a scenario would make misclassification about variety type a potential source of endogeneity in causal settings using survey data to study the impact of adoption on yields, or other higher outcomes.

Box 1. A word of caution on methodological choices

In a study of bean-growing households in Costa Rica, Occelli et al. (2025) demonstrate that the accuracy of varietal adoption data depends significantly on both social and genomic methodological choices. By intentionally targeting respondents with high varietal knowledge rather than using standard random selection, the researchers achieved an eight-percentage-point increase in identification accuracy. The findings suggest that data quality is driven by the specific expertise of the person interviewed, rather than merely by selecting a different family member. Furthermore, the study introduces the concept of genomic uncertainty, revealing that variability within genetic reference libraries often causes mismatches previously attributed to farmer error. When accounting for these genomic confidence intervals, the match rate between farmer reports and DNA results rose from 41% to 58%. Ultimately, the research highlights that varietal identification is a process that requires knowledge-based respondents and transparent acknowledgment of the inherent uncertainties in DNA fingerprinting.

Given that both the theoretical reasons for concern and body of empirical evidence collected to date point in the same direction, it seems clear that taking samples of plant tissue from farmers' plots for laboratory analysis offers a more objective way of measuring adoption of crop varieties. The purpose of this guidebook is to introduce to economists, social scientists and agronomists how this can indeed be done. We draw on examples from SPIA's first phase of country studies in which we embedded DNA fingerprinting into nationally-representative surveys in Ethiopia, Uganda, Bangladesh and Vietnam.

Before we start, it seems appropriate to sound a note of caution. The methodology for successful implementation of DNA fingerprinting in household surveys is complex, will have implications for the logistics of survey operations, and is certainly more costly than asking a survey question. However, the numerous studies published in the past decade, reviewed in Stevenson et al (2023) have shown that it is possible.

Empirical Studies Used in this Guidebook

DNA fingerprinting for varietal identification (e.g. Morell et al, 1995) and testing of the composition of processed foods (e.g. Terzi et al, 2005) has seen commercially oriented applications in rich countries for many years. However, the past decade has seen the piloting and refinement of DNA fingerprinting for crop varietal identification for public research and statistics purposes in both Africa and Asia.

Table 2 provides summary details of such studies that the current authors have been involved with under the auspices of SPIA's country-level studies.

These crops offer a range of reproductive and propagation systems – cross-pollinated (maize), self-pollinated (rice, beans, groundnut), and clonally-propagated crops (cassava, sweet potato, banana) – and thus the opportunity for highlighting how the intricacies of the underlying genetics and seed systems of the crop have important implications for how fieldwork and analysis is organized in a DNA fingerprinting study. The four countries where these studies were carried out are of high strategic importance for international plant breeding efforts.

For each of these studies, multiple non-rival approaches were used to collect the same data about the same crop standing in the same field. First, a farmer provided self-reported data, and then a sample of plant tissue was taken from the plot in question for DNA fingerprinting, to compare the genetic profile of that sample to a reference set of samples of known identity. For some of these studies we leveraged a pre-existing arrangement in which a crop-cut of grain was routinely taken, primarily for the purposes of productivity estimation. In other cases we collected multiple samples within a plot, either using the farmers own information about varieties they considered to be present in the plot, or duplicates of single plant samples that were taken to represent the plot.

Table 2: Overview of source studies used in these guidelines

Crop	Country	Survey year	DNA fingerprinting approach	Sample type	Reference
Maize	Ethiopia	2018 2021	DArT	Crop-cut grain	Kosmowski et al, 2020; Alemu et al, 2024
Maize	Uganda	2022	DArTSeq	Crop-cut grain (2 per plot)	Ilukor et al, 2025
Beans	Uganda	2021	DArTag	Bulked leaf sample	Ilukor et al, 2025
Cassava	Uganda	2022	DArTSeqLD	Single plant leaf sample per "farmer-declared" variety	Ilukor et al, 2025
Banana	Uganda	2022	DArTSeqLD and Intertek KASP	Single plant leaf sample per "farmer-declared" variety	Ilukor et al, 2025
Sweetpotato	Uganda	2021	DArTSeqLD	Single plant leaf sample per "farmer-declared" variety	Ilukor et al, 2025
Groundnut	Uganda	2022	DArTag	Single plant leaf sample per "farmer-declared" variety	Ilukor et al, 2025
Rice	Vietnam	2022	Agriplex Rica 4	Single plant leaf sample (4 per plot)	Kosmowski et al, 2025
Cassava	Vietnam	2023	Intertek	Single plant leaf sample per plot	Kosmowski et al, 2025
Rice	Bangladesh	2024	Agriplex Rica 4	Single plant leaf sample (with sub-sample duplicated)	Singla et al, 2025
Beans	Costa Rica	2025		7 leaves per plot	Occelli et al. (2025)

Poets *et al* (2020) – a companion to this guidebook with an overlap in authorship – provides a more technical, genetics perspective to choosing specific genotyping approaches, as well as highlighting some of the analytical implications associated with those choices. Our unique contribution here is to provide a non-technical account of what we have learned about good practice on integrating DNA fingerprinting with household surveys. Our intended audience is economists and social scientists, with the goal of helping them understand the issues sufficiently to design surveys and find the appropriate collaborators to work with.

Section 3. Building a Reference library

KEY POINTS

- Learn as much as you can about the varietal releases, from different sources (including breeders)
- Investigate the existence of a prior validated reference library
- Genotype the reference library before starting fieldwork, if possible

The process of compiling a reference library – a comprehensive collection of reference materials used for matching with field samples – is critical. Building a reference library consists of collecting improved plant materials that will act as a benchmark to evaluate field samples' closeness to these true-to-type genotypes. Deciding what should go into the library, and identifying the appropriate source for each variety, is thus very important. Challenges in establishing a reference library are context- and crop-specific. A reference library should be as **exhaustive** as possible of all improved varieties released in the country, at least those released within the timescale you are interested in, and especially those known to have been bred through the research networks you are interested in (e.g. NARS and/or CGIAR).

The material collected for the references should be **correct**, meaning that the reference sample is a true representation of that genotype. Typically this means that it should be breeders' seed, obtained from the breeding program or from the genebank for the crop in question. Finally, the reference library should be **distinctive** – i.e. that reference varieties known to be genetically different (by definition, crop varieties must be distinctive) can be observed to be distinctive. Where they cannot be distinguished you may need to adjust the specific type of DNA fingerprinting approach (or "assay") to allow for a greater density of markers, thereby increasing the chances of finding two similar varieties to actually be distinctive.

Depending on the crop and context, it is worth enquiring about the existence of genotyped reference materials, provided that these are representative or can be complemented with new references. This consideration – the existence of references from prior studies using specific genotyping platforms – may therefore play a major role in the choice of genotyping lab to work with (see Section 7. Choosing and Using Genotyping Services). If in doubt, seek advice from SPIA.

What constitutes an improved variety?

The terms improved varieties and cultivars, used interchangeably in the literature, refer to plants that have been developed for cultivation following a process of selective breeding. **Improved varieties** are usually considered the opposite of **landraces** (also called local, or traditional) varieties, that have arisen through a combination of natural selection and farmer selection in a specific geography but for which no formal plant breeding was performed[4]. For some clonally propagated crops, there is an intermediate category of **selected landraces**. These are landraces that researchers identified as being popular with farmers that are then brought on-station to be characterized and "cleaned up". This might involve virus indexing to ensure propagation and multiplication only of virus-free planting material. While the process of selecting and cleaning landraces does not involve plant breeding, these may be considered improved varieties if the role of evaluating, cleaning and multiplying the planting material is thought to be a substantial contributor to promoting adoption. However, note that individual instances of farmers' adoption of selected landraces will be observationally equivalent to instances in which the farmer has just persisted with the original landrace.

[4] Farmers will likely have made their own selections over time by evaluating within their carry-over seed from one harvest to the next. This process can be very locally effective but is quite distinct from the process of formal plant breeding.

How are improved varieties released?

The most effective way to enumerate the universe of improved varieties is to follow the institutional framework established for seed approval and releases. Most countries now operate national varietal release *committees* that maintain a list of improved varieties that have been tested, evaluated for their superiority to landraces regarding specific performance criteria (typically yield, but also resistance to pests and diseases or abiotic stresses like drought, heat, salinity or flood). Varieties approved by these committees can be released and either commercialized by the private sector or multiplied by government and NGO actors for dissemination to farmers.

Each improved variety should have a unique name, a year of release as well as a list of plant descriptors and agro-ecological suitability indicators that justify its release. The Diffusion and Impact of Improved Varieties in Africa (DIIVA) project, completed in 2013, compiled the names and origins of new varieties released per country/crop combination from 1960 to the early 2000s[5]. In cases where improved materials released in the past cannot be identified or breeder seed samples cannot be obtained, the use of DNA fingerprinting should be reconsidered (unless the interest is solely in identifying recent releases). With an incomplete reference library, DNA fingerprinting arguably will not do any better than other methods as many samples will be unidentified. For crops with commercial seed varieties (particularly maize) on the market, researchers may choose to include commercial seed as references if breeders' seed for those varieties is unavailable[6].

To start, it is useful to identify the relevant institutions in the country working on the specific combination of crop(s) you want to collect data about. Such a list includes:

- International Agricultural Research Centers (IARC): collaborate with national research agencies on germplasm exchanges and breeding. This is the main channel by which germplasms from international collections can be integrated into national breeding programs.

- National Agricultural Research Systems (NARS): apply the specific breeding procedure and are responsible for selecting the best varieties. In some countries, NARS are also responsible for producing breeder seed, the first generation of true-to-type varieties that will then be multiplied at a larger scale.

- Variety release committees: ensure the uniqueness and distinctiveness of varieties and deliver the approval for distribution.

- Seed multiplication agencies: taking foundation seed and producing certified seed / planting material for dissemination / sale to farmers may be in hands of a dedicated government agency.

- Farmer's seed cooperatives: commercialize the dissemination of varieties, particularly for crops not covered by the private sector.

- Private seed companies: can commercialize their own improved varieties (ex: Monsanto, Pannar, Pioneer, Seed Co) or have license to multiply and sale varieties developed by NARS

An official list of improved varieties released may exist, typically at the Ministry of Agriculture. This list shows the varieties that have been released to date. If it is possible to get a copy, it is worth checking it with knowledgeable people to ensure it is accurate and up to date as it may not be well maintained. The number of entries in the varietal

[5] Data are available at *https://www.asti.cgiar.org/diiva*

[6] Using certified seed from commercial companies as references requires significant caution. Even in countries with highly regulated seed systems with a good degree of quality assurance / quality control (QA/QC) there is the potential for a commercially sourced certified seed sample to not faithfully represent the original breeders' seed owing to admixture, genetic drift or mistakes in seed production.

release list will vary by crop, depending on the level of plant breeding effort for the crop in question and the number of years since the establishment of the official system of tracking varietal releases.

Box 1. Accessing germplasm through international genebanks

Operating at a worldwide level, genebank platforms possess a collection of germplasms that contain best performing varieties as well as local landraces collected in specific locations. When a country has engaged in breeding efforts using foreign germplasm, the original line can often be traced back. We advise complementing samples collected within the country seed system with relevant genebank accessions where possible.

The CGIAR Genebank Platform encompasses 11 genebanks that conserve accessions of crops and trees on behalf of the global community under the International Treaty on Plant Genetic Resources for Food and Agriculture (CGIAR, 1994). Genebanks have a legal obligation to conserve and make available germplasms for research purposes. Improved and conserved seeds are thus freely available upon request to any research institution, worldwide. Appendix C provides details on how specific crop germplasms can be requested.

Choosing the right planting material for the reference library

The definitions in Table 3 below represent the levels in the seed multiplication pipeline that proceed following development of a variety. The stages and terminologies along the pipeline can differ across various national and international seed systems. For example, in India, an extra stage of nucleus seed precedes the breeder seed while in the United States an intermediate stage of registered seed preceding certified seed is common. The terms breeder seed and foundation seed can also be used interchangeably with pre-basic and basic seed respectively. Regardless of nomenclature, breeder seed is the optimal source material for compiling the genetic reference library. This represents the true-to-type genetically pure variety with sufficient fidelity to act as a standard of comparison for field samples.

Table 3: Classification of seed following variety development

Breeder Seed (or “pre-basic seed”)	**Foundation Seed** (or “basic seed”)	**Certified Seed**
This refers to seed produced under the direct selection of a breeder or breeding institution that develops a variety. Conditions for breeder seed production ensure sufficient controls that guarantee the genetic purity and varietal fidelity of resultant seed	This is seed that directly descends from breeder seed. It is produced under conditions designed to ensure maintenance of high genetic purity and identity	This seed descends from the foundation seed. It is produced under the supervision of a seed certification agency that ensures legally defined standards of purity are met

Reference library meta-data

It is important to keep track of where and when reference samples are obtained, and at which step of the chain. Establishing a database with this information is strongly recommended using the minimum set of meta-data about the reference samples, as outlined in Table 4. While much of this is self-explanatory, we expand on a few items here. The tissue type column refers to whether the reference sample comprises seed or leaf tissue. The replicate column has information on whether the sample is replicated more than once. In those cases, there should be multiple rows to allow for multiple barcodes (1 per physical reference sample). It may be useful to expand the number of fields to include information on traits that are of interest (as obtained either from the release list data or from breeders) and information on the role of international agricultural research in contributing germplasm for the variety in question.

Table 4. Reference library meta-data fields: Example

Crop	Variety name	Maintainer	Year of release	Provider	Tissue type	Single_Bulk	Replicate	Date collected
SORGHUM	76TI#23	Melkassa ARC	1979	Melkassa ARC	Seed	Bulk	Y	1/6/2016
SORGHUM	Abshir	Melkassa ARC	2000	Sirinka ARC	Seed	Bulk	Y	12/6/2016
SORGHUM	Birhan	Melkassa ARC	2002	Melkassa ARC	Seed	Bulk	Y	12/6/2016
SORGHUM	Dekeba	Melkassa ARC	2012	Sirinka ARC	Seed	Bulk	N	1/6/2016

Ensuring Exhaustiveness, Purity, and Distinctiveness

The meta-data will be of great help in demonstrating to users that the reference library is representative and exhaustive, but we also recommend assessing seed purity of the samples provided, and checking for distinctiveness of materials, ideally before field survey operations. We examine these issues below.

Ensuring exhaustiveness: which varieties to include?

The first step in building a relevant reference library consists of understanding the seed system and its actors. Who are the main actors carrying out plant breeding for the crop and country in question? Is there a private sector for breeding the crop? What role have international research centers played in providing exotic germplasm for local breeding? If a local research institute has undertaken collection, screening, evaluation and multiplication of local landrace materials, is this considered an improvement to be measured? The definition of what constitutes an improved variety is important and may vary in different contexts. The goal should be to have a reliable reference sample for each varietal release. Preferably, reference samples are replicated to build confidence in the uniformity of the reference material.

The extent of comprehensiveness in the reference set should be guided by the research question intended to be addressed. It may not be necessary to collect a complete set of references if the goal is to simply identify whether farmers are growing a certain set of improved varieties (e.g. those released after a specific date). If the goal is to identify the full range of varieties that farmers are growing, then as comprehensive a set should feature. If a variety is not in the reference library, it cannot be matched to the field samples to be identified.

For hybrid maize it is recommended to include the precursor breeder material used to construct the final varieties in the reference profile. These would include inbred lines and OPVs that are crossed to generate three-way hybrids and varietal hybrids respectively. While these precursors will not be used for variety identification of samples, they serve as important checks to confirm the integrity of the hybrid varieties.

Should local landraces be included in the reference library?

While the emphasis so far has been on ensuring exhaustiveness of improved varieties, it may be wise to also include local landraces for which material is available. Landrace references could be important in resolving ambiguity when samples fail to match known varieties – it is more satisfying to positively identify samples than leave them unidentified and having to assume that they are landraces. However, we advise caution as there is less control over local landraces than is the case for the maintenance of improved varieties. For example, in the case of the Uganda maize study, we initially included what the NARS breeding program had told us was a popular landrace OPV collected from farmers, only to find that it was genetically indistinguishable from the dominant released OPV (LONGE 5D). Therefore, we excluded this “landrace” and re-ran the analysis without it, as it was confounding the main thing we wanted to know, which was the adoption of the varieties originating from the breeding program.

In the case of a study of common bean in Uganda, an “exhaustive” set comprising more than 400 landraces was included among the references. The majority of these were un-curated collections that turned out not to be entirely

genetically distinct. Consequently, a lot of effort had to be expended on reducing redundancy analytically prior to their use for variety identification. On the other hand, for groundnut in Uganda, Red Beauty was a landrace accession whose identity had been carefully maintained over time. Hence, its inclusion in the reference library was critical for tracking its adoption.

Checking purity of reference materials

Since DNA fingerprinting assumes that the reference library is correct, it is necessary to confirm the purity of the seeds used for the reference. It is possible that genetic lines were not sufficiently separated during the breeding process – particularly for an outcrossing crop like maize – leading to varietal outcrossing with an undesirable pollen donor. Or, in the case of clonally propagated crops like cassava, it is possible that breeders' germplasm for some varieties may be admixed.

To confirm reference seed purity, we recommend planting and growing three to five seeds, then collecting leaf tissue from each plant and genotyping them. Individual genotyping provides the most detailed assessment of purity. Breeders' seed are expected to be pure, hence deviations from this expectation can be flagged and corrected at this first round of genotyping. However, information about this original discrepancy should be retained as it may help with interpretation of what we subsequently see in farmers' fields – if there are issues with the purity of breeders' seed, then these may well have been perpetuated through all subsequent steps in the seed system for that crop. We should know which varieties are affected by such potential impurities. Despite being more costly due to higher number of samples, this is the most rigorous method for establishing a reliable reference. Bulk sample genotyping can be used as an approach to reduce genotyping costs. In this case, tissue from multiple plants are physically pooled into a single sample for genotyping. With this approach genetic mixing (admixture) can be effectively detected, though the source of contamination cannot be determined. While less precise, it can be a more affordable way to flag a potentially contaminated seed source than processing multiple reference samples per variety.

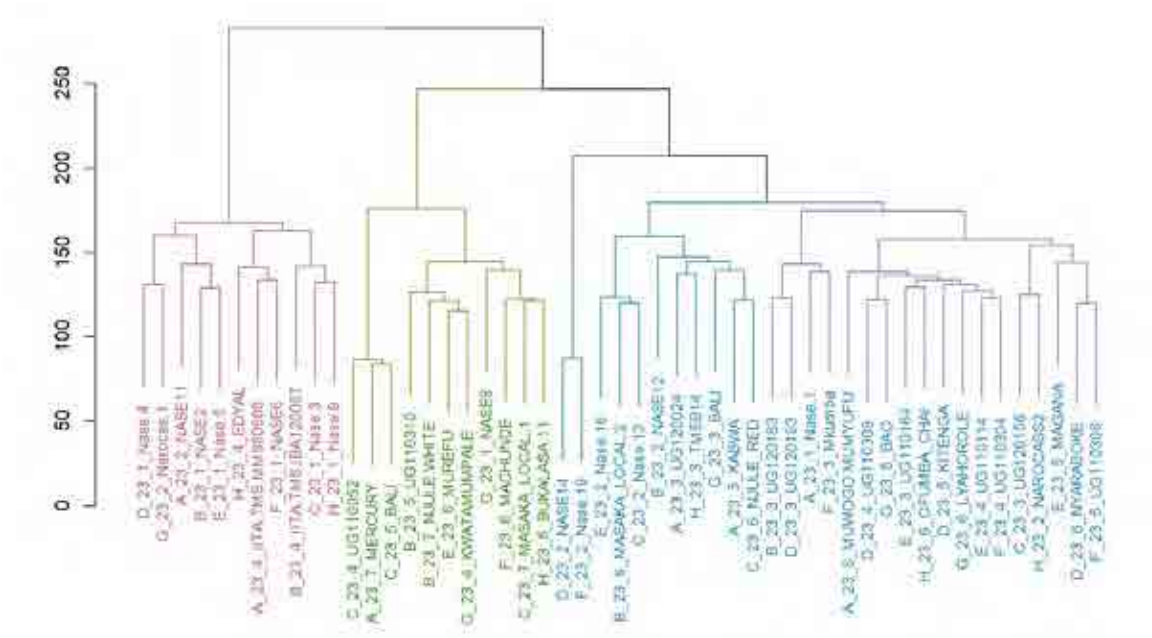
Assessing distinctiveness

To ensure reliable identification of field samples, there should be sufficient genetic distance, or discrimination between reference varieties. The risk of insufficient discrimination among varieties is real and can originate from shared pedigree of varieties, seed sample purity or the density of the assay used in genotyping (i.e. how many places we look for polymorphisms). To ensure sufficient genetic distance between reference samples, we recommend sending the reference library for genotyping before field data collection. If the similarity among varieties is due to the biology of the crop - as would be anticipated with clonally propagated crops[7] or a particular genotyping assay is found to be insufficiently discriminating, we can in some cases "dial up" the density of the assay (at greater cost per sample) rather than returning a lot of samples that cannot be meaningfully differentiated from one another. The genotyping should be done on the same platform as is to be used for genotyping the field samples. Importantly, consulting with the plant breeders about the pedigrees of the varieties and how they were developed can also help interpret the dendrogram of distinctiveness generated on the reference set. The examples of cassava and banana in Figure 2 illustrate how genotyping data can be used to assess the levels of variety distinctiveness in the reference library.

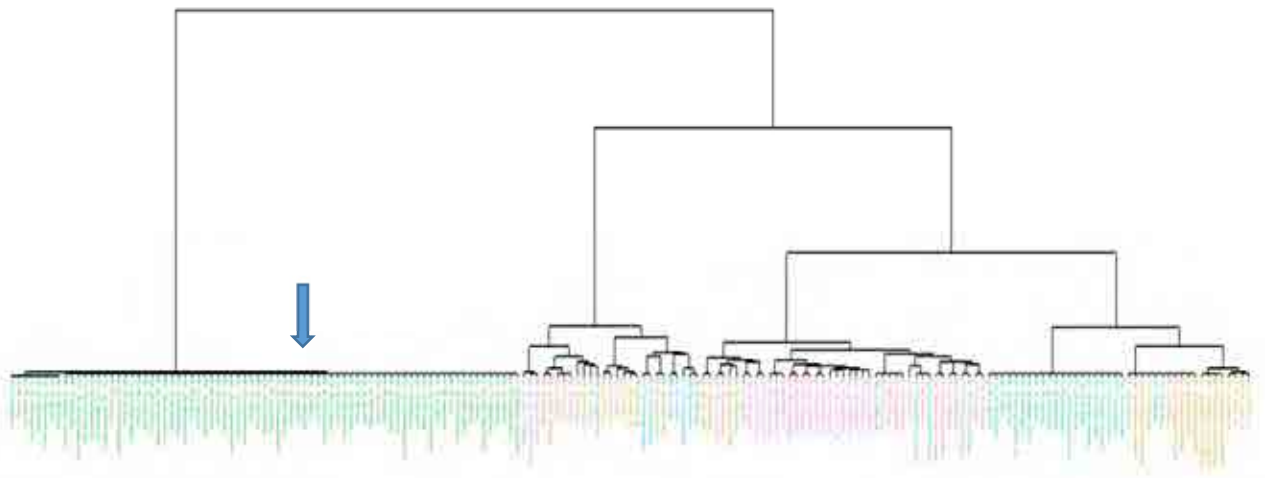
[7] Clonally propagated crops such as cassava, banana and sweet potato are either effectively sterile (as in triploid hybrid bananas) or can theoretically set seed and reproduce sexually but in practice this is a very rare event (sweet potato and cassava).

Figure 2. Dendrogram showing genetic clustering of reference varieties for a) cassava and b) banana in Uganda

(a) Cassava references from study in Uganda (from Ilukor et al, 2025). The clustering patterns show good distinctiveness between the varieties



(b) Banana references from study in Uganda (from Ilukor et al, 2025). The different landraces cluster together under EAHBs (left main branch) with no meaningful distinctiveness, but the improved varieties (right main branch) show clear genetic distinctiveness



Section 4. Sampling strategies

KEY POINTS

- The broader context of the survey will determine the opportunity you have for sampling – whether it should be grain from crop-cuts or leaf samples from a random walk within the plot.
- Leaf sampling protocols (single vs. bulk) must match the expected varietal heterogeneity and crop propagation method.

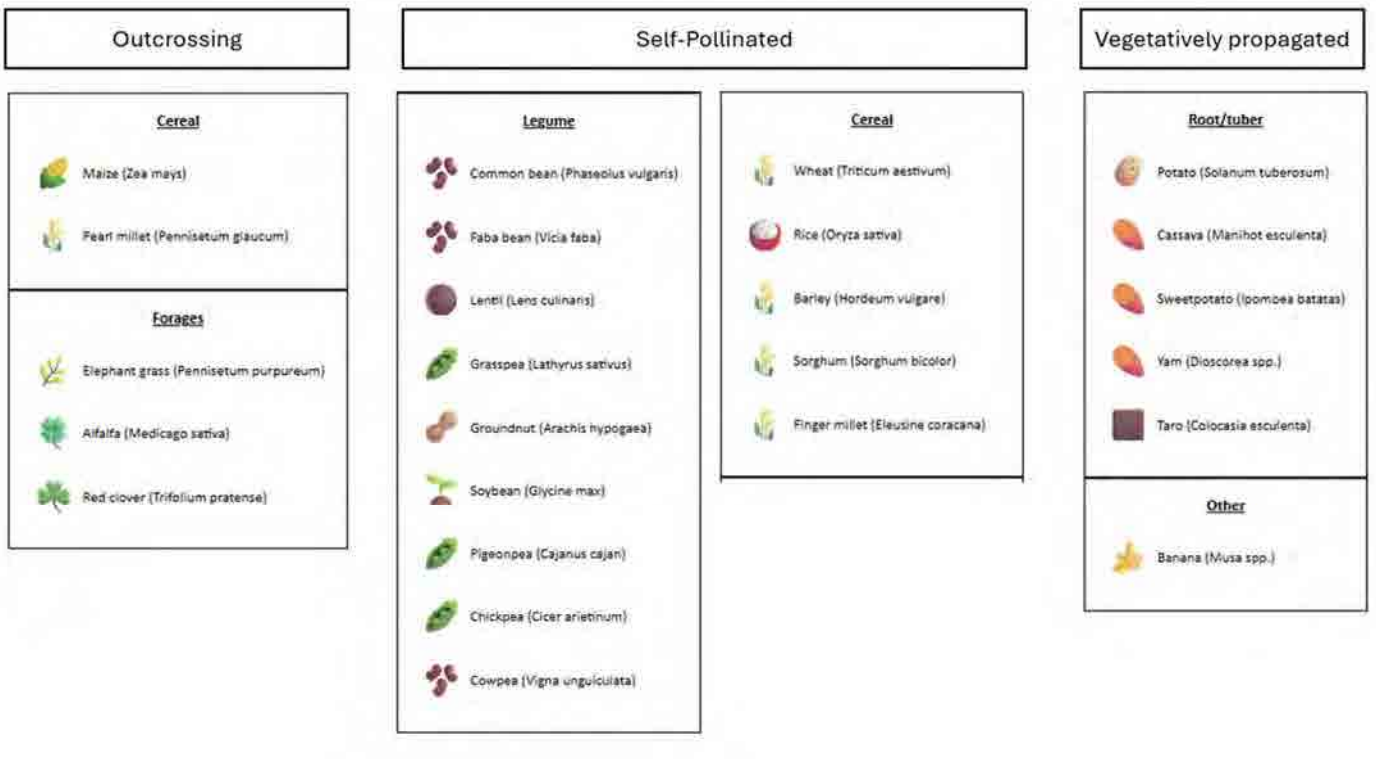
In this section, we shed light on key considerations for three levels where choices are made in determining what to sample: the ex-ante survey design, sampling plots and intra-plot sampling. These choices must be informed by the biology of the crop in question.

Crop-level sampling considerations

At the crop level, the optimal sampling strategy depends on the crop's reproductive biology—whether it is outcrossing, self-pollinated, or clonally propagated—and the expected genetic uniformity of the plot.

Plants have evolved different reproductive strategies. Most crops produce seeds as a result of either cross-fertilization (outcrossing) where pollen from one plant fertilizes another or self-pollination (also referred to as inbreeding or selfing) where the pollen fertilizes the ovules of the same plant. In the latter category, a degree of cross-fertilization can occur with rates varying according to specific crop genotype and environmental conditions. However, despite their potential for propagating using mixed mating strategies, they are generally bred and managed as inbreeding species. Seeds are genetically recombined offspring of the parent plants. Self-pollinated crops therefore generate much less genetic variation than cross-pollinated crops and tend toward uniformity until admixed (physically combined with a different seed source). Vegetative propagation is a reproduction strategy used by a distinct class of crops where new plants are produced from vegetative organs (e.g., tubers, suckers, cuttings) that produce genetically identical plants (clones) without sexual reproduction / pollination. In Figure 3 we provide information on a comprehensive list of crops under the mandate of CGIAR centers and how they map onto these categories.

Figure 3. Classification of 26 crop species according to their primary propagation method



Below, we review the various options available to survey practitioners. The distinction between self-pollinated and clonally propagated crops and cross-pollinated crops serves as a starting point for developing an effective crop-level sampling method. Ultimately, the final choice reflects a strategic trade-off between the costs and logistical challenges of sampling and the desired precision of varietal identification.

Self-pollinated and vegetatively propagated crops

For self-pollinated and vegetatively propagated crops, individual plants within a pure variety are expected to be genetically uniform. This makes **leaf sampling** a reasonable approach.

The key decision is whether to use leaf tissue from a single plant, taking multiple single plant samples, or bulking leaf tissue taken from several different plants into a single sample. Taking a single plant sample runs the risk of committing sampling error – you may make a draw of a non-representative individual plant from a heterogenous population. Taking multiple single plant samples is more expensive and time consuming but can build confidence in your varietal identity assignment if all individual samples are assigned to the same genetic identity. Bulking material from multiple plants to a single sample will reduce the genotyping costs (relative to multiple individual samples) and will positively identify genetically uniform plots but will be difficult to assign to a varietal identity if the plot is physically admixed. A single sample made up of 80% of tissue material from one variety and 20% from another will be difficult to analyze.

The choice depends on the expected uniformity of the plot and involves the tradeoff of cost and time against the risk of measurement error from unidentified variety mixtures. The following scenarios give guidance and recommendations to aid in selecting between the two choices.

Scenario A: Plot is expected to be homogenous

- Recommendation: Single tissue sampling.
- Procedure: Collect a leaf disc from one representative plant.
- Rationale: If the plot consists of a single uniform variety, a single plant is genetically representative of the whole plot. This is the most cost-effective method and leads to straightforward variety identification.

Scenario B: Plot is expected to be heterogeneous (admixed)

- Recommendation: Bulk tissue sampling. A single plant sample would be misleading. The choice of how to bulk depends on the study's objective.

Option 1: To simply confirm admixture

- Procedure: Collect a small bulk sample (e.g., leaves from 5 individual plants).
- Rationale: This is a cost-effective way to test for heterogeneity. The resulting genotype will show high levels of heterozygosity if the plot is mixed, serving as a clear indicator of admixture. The logistics are comparable to single-leaf collection. While this confirms a mixture, it may not be sufficient to identify the dominant variety.

Option 2: To identify the dominant variety

- Procedure: Collect a large bulk sample (e.g., leaves from 20-30 individual plants).
- Rationale: Increasing the number of plants in the bulk increases the probability that the allele frequencies of the most dominant variety will overshadow the "noise" from minor varieties, improving the chances of a successful ID. However, this carries significant logistical challenges. For instance, if 25 plants are selected, it would be necessary to prepare for five collection tubes per plot, each of which will accommodate five leaf discs. The samples would be plated separately and instructions given to the genotyping service provider to combine them after DNA extraction just prior to genotyping. This procedure still may not yield a clear ID if no single variety is dominant.

Option 3: To identify all component varieties

- Procedure: Collect and process multiple single tissue samples from different plants within the plot.
- Rationale: If it is critical to know the composition of an admixed plot, this is the only option. It provides the highest resolution but vastly increases the cost of genotyping and analysis.

Outcrossing crops

For outcrossing crops like maize, samples in the reference libraries should not be made up of individual plants. This is because their outcrossing nature results in genetically heterogeneous populations, where each individual plant is a new, unique genetic combination. Thus, the varieties are differentiated based on population-level characteristics rather than individual genotypes.

When dealing with maize, it is important to be aware of the two main variety types:

a. Open-Pollinated Varieties (OPVs): These are genetically diverse population of maize where individual plants are heterozygous and the overall population is heterogeneous. Consequently, a single plant is not representative of the variety as a whole.

b. **Hybrids**: These result from a cross between two or more distinct inbred parent lines, yielding plants that are individually heterozygous but forming homogenous populations. The level of uniformity at the population level depends on the number of parent lines used.
    - **Single-cross hybrids** (Inbred A x Inbred B) result in plants that are genetically identical to each other, resulting in maximum uniformity.
    - **Three-way hybrids** [(Inbred A x Inbred B) x Inbred C] are produced by crossing a single-cross hybrid (A x B) with a third inbred line (C). This introduces controlled levels of diversity resulting in plants that are highly heterozygous but not genetically identical to each other. The populations formed are slightly less uniform than a single-cross hybrid.

Box 2. Heterozygosity in plants

Heterozygosity refers to the genetic state where an individual has two different forms (alleles) of a gene. Using SNP markers, this can be observed as the presence of two different DNA bases at a particular SNP site, one inherited from each parent (e.g., an 'A' base and a 'G' base).

In inbreeding crops, the process of reproduction through self-fertilization systematically reduces heterozygosity by 50% with each generation, leading to the rapid development of highly homozygous or inbred lines. Consequently, the rate of heterozygosity in these crops is low. When present, this is often an indication of a recent and likely accidental cross-pollination event. In bulk samples, a heterozygous signal can also be a good indicator of admixture, suggesting the sample contains a mix of individuals with different DNA bases.

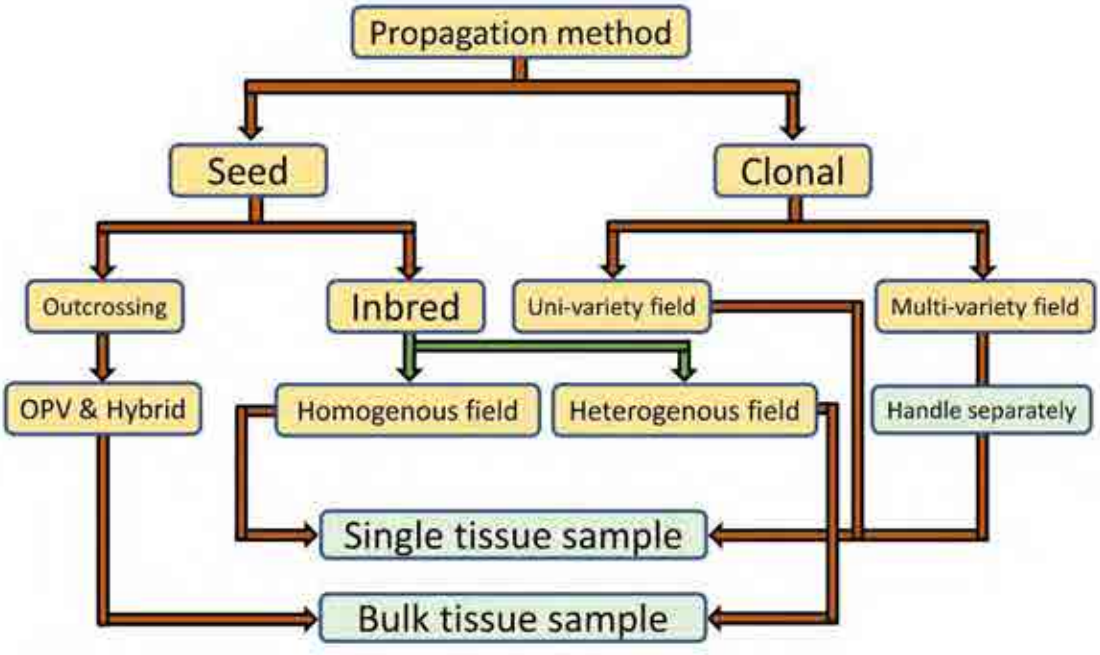
In outcrossing crops, pollen is exchanged between different plants, hence promoting the active combination of genetic material between plants, resulting in high levels of heterozygosity within the population. It is worth noting that commercial hybrid varieties of these crops are typically created by crossing inbred parent lines. While the final hybrid plants are highly heterozygous the precursor parent lines are, by design, highly homozygous.

For both OPVs and hybrids, the goal is to capture the representative genetic profile of the population in the plot. Therefore, **the samples should be constituted from bulk samples** representing the diversity within a variety. **We recommend sampling grain from crop cuts** or harvested ears rather than leaf tissue. While collecting leaf tissue ensures the sampled DNA comes directly from the plants grown in that specific season (avoiding the risk of pollination by distant males), it is logistically challenging. Bulking fresh leaves from many plants (e.g. 30 - 50) creates a sample with high moisture content, requiring immediate drying or careful storage in multiple tubes to prevent DNA degradation. This adds significant time, cost, and potential for error during fieldwork. Crop cuts should be laid several meters to the interior of the plot as this will result in a very low probability of the maize plants being fertilized by pollen blown in from male maize genotypes outside the plot. Fertilization decreases exponentially with distance from the source and lines of maize planted on the borders are most susceptible to pollination from outside (Bannert and Stamp, 2007).

Technical note for genotyping of bulk samples

When bulk samples are collected for maize, or for suspected admixed inbreds where the objective is to identify the dominant variety, the genotyping must be performed using a platform capable of delivering data that is amenable to quantitative allele frequency estimation (e.g., DArTag, DArTseq, GBS). This is essential for interpreting the composite DNA signal from a mixed sample.

Figure 4. Decision tree on whether to collect single leaf sample of bulk leaf sample



Survey Design Considerations

The design of the survey being implemented will shape how DNA fingerprinting can be integrated. In some cases, DNA fingerprinting may be a central focus, giving you full control over sampling. In others, it must fit within the constraints of an existing survey, which may already have established protocols.

Optimizing Survey Design for Effective DNA Fingerprinting Integration

Practical considerations in survey design and field constraints significantly influence how DNA sampling is implemented. Two illustrative cases—maize in Ethiopia and banana in Uganda—highlight different approaches and challenges in optimizing survey design for effective DNA fingerprinting integration.

Case 1: Maize in the Ethiopian Socioeconomic Survey (2018 and 2021)

In the Ethiopian Socioeconomic Survey (ESS) in 2018 and 2021, DNA fingerprinting was incorporated into an existing design where resident enumerators were already performing crop-cuts for yield estimation. We worked with the Ethiopian Central Statistics Agency (later the Ethiopian Statistical Service) to embed DNA fingerprinting into the Ethiopian Socioeconomic Survey, waves 4 (2018/19, reported in Kosmowski et al, 2020) and 5 (2021/22, reported in Alemu et al, 2024). The marginal costs of diverting a sub-sample of these crop cut grain from the yield measurement process into a pipeline for DNA fingerprinting was relatively low. It made sense for SPIA to leverage the crop-cuts not just for cost considerations but also because it meant that we would have high quality yield data and GPS-measured plot area for the same plot as we have the genetic identity data (see Jovanovic and Ricker-Gilbert, 2025 for an application of these data to the question of the impact of adoption of improved varieties on yields).

Case 2: Banana in Uganda (2022)

The sampling for six crops was integrated into the Uganda Household Integrated Survey 2021/22 (Ilukor et al, 2025). The fieldwork for the survey was organized over a 12-month implementation calendar in three visits to the households. Each visit had different crops featured for sampling. Banana was sampled during the second visit to the household, and the goal was to sample a single individual plant of each banana variety that the farmer reported cultivating. However, this proved unmanageable during the early fieldwork as a large share of farmers reported cultivating seven or more varieties. The sampling was taking too long, and enumerators were understandably objecting to the survey burden. Consequently, the protocol was revised to focus only on the three "most important" varieties as declared by the farmer. Interestingly, the genetic analysis later revealed that most of the farmer-reported phenotypic variation was not reflected in genetic differences. A large majority of the bananas sampled belonged to a single, genetically uniform cluster of East Africa Highland Banana landraces.

Timing of data collection

When integrating a new module into an existing survey, it is essential to understand the timing and logistical constraints.

- *Grain from Crop-Cuts*: Enumerators must arrive at harvest time – too early and farmers may resist, too late and the crop may be gone. Coordination with local contacts is essential and such a strategy will either rely on the survey already having resident enumerators that can easily adjust their visit for sampling, or work with a dedicated crop-cutting team that can be deployed to an area when harvests are imminent.
- *Leaf from Random Walks:* Green tissue for the focus crop must be available during the time of enumerator visits. Young, green leaves yield the best DNA for genotyping.

Case 3: Groundnut in Uganda (2022)

In this case, the survey team arrived late in the growing season to collect leaf samples from groundnut plots (Ilukor et al, 2025). By the time of sampling, many plots had already been harvested. As a result, the remaining plots available for sampling were not representative—they excluded early-maturing groundnut varieties and the plots of more organized farmers who had already completed their harvests. This timing issue led to a biased "selected" sample that does not accurately reflect the diversity of groundnut varieties or farmer practices in the area.

Case 4: Rice in Viet Nam (2022)

In 2022, a new crop-cut module for rice DNA fingerprinting was added to the Vietnam Household Living Standard Survey (Kosmowski et al, 2024). The aim was to collect all samples during March and April, covering the main winter-spring cropping season. However, COVID-19-related delays meant that only 62% of samples were collected during this period, with the rest gathered later. Ultimately, the final sample included rice from multiple cropping seasons: 73% from winter–spring, 11% from summer–autumn, and 16% from autumn–winter. This variation in timing and cropping seasons added complexity to the data collection and analysis.

Sampling plots

To ensure representativeness across diverse farming systems and agroecological zones, combine probability sampling at the cluster level with systematic random selection of households and plots within clusters.

For example, the Ethiopia Socioeconomic Survey (ESS) employs a clear sampling pipeline for selecting plots for crop cuts, using a two-stage stratified cluster sampling process within each sampled enumeration area (EA). In the first stage, EAs are selected using probability proportional to size (PPS), stratified by urban and rural areas. In the second

stage, within each rural EA, 10 agricultural households are randomly selected for sampling. Following household selection, an inventory of eligible crop plots for crop cutting is compiled. If more than five crop plots of a given crop exist within an EA, five plots are randomly selected for crop cutting. If there are five or fewer eligible plots, all plots are included. Mixed or intercropped plots are generally excluded if there are sufficient monocrop plots available; otherwise, they may be considered for sampling when fewer than five pure stand plots exist (World Bank, 2020).

Plot-level Considerations: On-Farm Realities

At the plot level, the primary consideration is anticipating the expected genetic variation within individual plots. This requires understanding both the local seed systems and farmer practices, as these factors influence the extent of variety mixing within plots. In many contexts, farmers do not use the concept of "varieties" to guide what they plant – it may be a foreign or unfamiliar concept to them. Rather, farmers may manage their plots based on identifying trusted sources of planting material (i.e. seed from specific people or institutions) or finding desired consumption traits from among informal sources (e.g. color, taste, cooking qualities) including the grain market. In general, if farmers struggle to confidently give the name of the variety or appear unsure of the concept of a variety, then we should plan for the possibility of finding high levels of physical admixture in their plots as it is likely that they are not managing for varietal fidelity. We further explore two sampling methods commonly used for this purpose: crop cuts and random walks.

Assessing Intra-Plot Varietal Heterogeneity

Before deciding on a sampling approach at the plot level, it is important to consider both the seed system and farmer behavior.

Seed Systems:

- Is there a formal, regulated seed system?
- What is the (approximate) share of farmers who buy new seed, exchange with others, or recycle their own?
- *Recommendations:* Review literature, use existing data, and conduct qualitative plotwork to understand seed sourcing and incentives.

Farmer Behavior:

- Do farmers mix varieties or seed sources within a plot?
- *Recommendations:* Use survey data, consult crop experts, and conduct qualitative fieldwork to understand planting decisions. If there is heterogeneity in different regions, be conservative and base your decision on the scenario presented by the most complex strategies farmers use.

Ultimately, determining the plot-level protocol relies on the way the crop is cultivated in the country of study.

Box 3. A proposed method for assessing intra-plot heterogeneity

To understand the overall rate of variety mixture in your study area without incurring the cost of intensive sampling in every plot, we recommend the following strategy:

In a randomly selected subset of 20-30% of all surveyed plots, collect two separate and independent samples.

- For a plot under a bulk sampling protocol (e.g., 5-leaf bulk), collect two separate 5-leaf bulks from different sections of the plot.
- For a plot under a single-leaf protocol, collect two single-leaf samples from different plants.
- For a maize plot, perform two separate crop-cuts and process the grain independently.

This approach of collecting two samples from a 20-30% subset of plots offers a crucial balance. It provides sufficient data to statistically estimate the prevalence and patterns of intra-plot heterogeneity across the broader population, allowing for more informed extrapolation of heterogeneity rates to the majority of plots where only one sample is taken. This ameliorates the logistical and cost burdens that would be incurred in intensive sampling of every single plot.

Sampling using Crop-Cuts

A crop-cut is a standardized agronomic technique used to estimate crop yield by harvesting a small, representative section of a plot. This typically involves marking out a defined area—such as a 4m x 4m square—and harvesting all the crop within that area. The harvested produce is processed, dried, and a subsample is taken for analysis. While the primary purpose of crop-cuts is yield measurement, these samples are also ideal for DNA fingerprinting because they are already dry, stable, and easy to store and transport. Crop-cut protocols are detailed in manuals such as those from the ESS and CGIAR (World Bank, 2020; CGIAR, 2023).

Measurement error regarding crop production can have significant implications for inference, potentially leading to spurious relationships and conflicting evidence (Abay et al., 2019; Lobell et al., 2020). Analytical objectives will dictate the required level of precision and representativeness. Higher levels of accuracy can be achieved by increasing the size of the crop cut, but this will also increase costs. Expectations of plot constitution should inform the implementation of the sampling by crop cut method.

Scenario A: Plot is expected to be homogeneous

This would be the case when a farmer is expected to have planted a single, known variety - especially a commercial hybrid and the plot appears visually uniform. In this case, a crop-cut from a standard 4 by 4 m quadrant is likely to accurately reflect the plot's genetic makeup. The decision of the quadrant size should also be taken using data on mean plot size (Kosmowski et al., 2021).

Recommended procedure:

1. Move at least 5 meters in from the plot edge to minimize pollen contamination from neighboring fields
2. Establish a single 4m x 4m crop-cut quadrant and ensure that this size is standardized for the entire survey
3. Harvest all maize cobs from the plants within the quadrant.
4. After shelling and drying the grain – following the main yield protocol, create a composite sample for DNA extraction by taking 50-100 grains sourced from at least 15-20 different cobs from the harvest

Scenario B: Plot is expected to be heterogeneous

This would be the case when a farmer is unsure of the variety planted, reports planting saved seed, planted grain from the market, or intentionally planted multiple varieties. The field may show high variability in plant height, color, or cob characteristics.

Scenario B1: Low expected heterogeneity

This might occur for instance, if a farmer planted a known OPV, but with some seed from a neighbor mixed in. The recommended procedure is to collect two standard crop-cuts as follows:

1. Randomly select two separate starting points within the plot avoiding edges
2. Establish two 4m x 4m crop-cuts
3. Harvest the cobs from each quadrant separately
4. Create a single composite sample for DNA extraction by taking 30-50 grains from 10-15 cobs from each quadrant's harvest, and combine them into one.

Scenario B2: Medium expected heterogeneity

This might occur if a farmer reports planting two different varieties in distinct sections of the same plot. The recommended procedure is to collect stratified crop-cuts as follows:

1. Ask the farmer to identify the different sections of the plot.
2. Establish one 4m x 4m crop-cut within the approximate center of each identified section.
3. Harvest, process, and sample each quadrant separately. Label the resulting grain samples using different barcodes, ensuring that all known sources of variation are captured as distinct samples.

Scenario B3: High expected heterogeneity

This might occur for instance, if a farmer planted recycled seed of unknown origin or grain from a market.

The recommended procedure is to collect multiple, smaller crop-cuts and process separately as follows:

1. Randomly establish 3 to 4 smaller crop-cuts (e.g., 2m x 2m) across the plot.
2. Harvest, process, and create a grain sample from each quadrant separately, labeling each with a unique barcode.
3. The decision on genotyping procedure can then be decided based on the aims of the study:

    - If the objective is to understand the plot's composition, each sample can be genotyped individually to provide higher resolution but at higher cost.
    - If the objective is to get an overall picture of the plot's genetic profile, combine the grain from all samples into a single bulk sample for genotyping. This will be more cost effective but will provide lower resolution.

However, as should be clear from the description, this is a very intrusive and disruptive data collection approach and may meet with significant resistance from farmers unless this is already part of an existing well-established survey approach used by the statistical agency.

Sampling using Random Walks

Random walk sampling is a practical method used to collect leaf samples within a plot by systematically moving through the plot in a randomized pattern to select individual plants. This approach helps capture spatial variation, hence avoiding selection bias. The standard method entails walking in a "W" or "Z" pattern, taking samples at regular intervals. This approach is recommended for self-pollinated and vegetatively propagated crops in plots with expected heterogeneity. For homogeneous plots—where a single variety is anticipated, a single plant is expected to be genetically representative of the whole plot. In such a case, random walking can be useful for collecting samples from multiple plants for confirming intra-plot homogeneity. In heterogeneous plots containing multiple varieties or seed sources, collection of a larger number of samples is necessary to adequately capture the full genetic diversity in line with the aims of the study; either, to confirm admixture, identify the dominant variety or identify all component varieties. The recommended procedure for doing a random walk is as follows:

1. Walk a "W" pattern across the plot.
2. Collect one leaf disc from the pre-determined number of plants based on the study objective at regular intervals along the walk.
3. Place all leaf discs into a single sample tube or in individual samples tubes in line with the study objectives.

If the plot is large or very highly admixed, stratified sampling can be incorporated into the strategy. In this case:

1. Mentally divide the plot into two halves.
2. Perform a "W" pattern walk in the first half, collecting the pre-determined number of samples into a single or individual tubes.
3. Perform a second "W" pattern walk and repeat the procedure in the other half of the plot. This will result in a higher number of samples per plot with the potential of providing the highest resolution data but at higher cost.

Section 5. Implications for Questionnaire Design

KEY POINTS

- Ensure your data collection instruments allow you to link barcoded physical samples to farmer-reported data about all relevant farmer-declared “varieties” by plot by household.
- Be mindful of the fact that many farmers will not be managing their planting materials as if they are distinct varieties. Indeed the concept of a variety may not be one they are familiar with or use in any meaningful way.

A well-structured questionnaire is a critical foundation for collecting reliable DNA fingerprinted data. The questionnaire design must ensure that plant samples can be accurately linked to household, plot, and variety-level information to enable meaningful analysis. This section provides examples and recommended module content to guide the integration of DNA fingerprinting into household surveys.

Structuring the Questionnaire for DNA Integration

Household Plot Roster

A plot roster provides a comprehensive list of all plots managed or cultivated by a household. This module facilitates the random selection of plots for sampling.

Table 2 illustrates a plot roster used to select cassava plots in the VHLSS 2023 survey.

Table 4. Example of a plot roster for selecting cassava plot in the VHLSS 2023

1. Is this household selected for cassava crop sampling?	Yes……..……. .	1 (>> Q2)
	No……..	>>> Next module

ENUMERATOR: PLEASE LIST ALL CASSAVA PLOTS CULTIVATED THIS SEASON

PLOT CODE	**2**	**3**	**4**	**5**
	Plot Name	What is the cultivated area of the plot, in sq. meters?	When was cassava planted on […] ?	*[Note: Automatic filling]* RANDOM SELECTION. REPORT THE PLOT ID ON NEXT SECTION. Randomly select a plot that was planted at least 1 month ago
	NAME	**AREA**	**MONTH/YEAR**	**NUMBER**
1				
2				
…				

Plot-level module

This module should be designed to gather all relevant information about the selected plot(s) for DNA fingerprinting. Key features include:

- Barcode Scan: The most critical element of this module is the barcode scan, which uniquely identifies the sample and connects it directly to the household survey records.
- GPS Measurement: Precise GPS coordinates of the plot are generally recorded to support spatial analysis. This variable can also be used to make sure the sample was randomly chosen within the plot, and that no bias exist towards plot selection (based on distance from household homestead)
- Seed Information: The module collects detailed information on the seed variety planted, sources and recycling practice. This is helpful in interpreting results from DNA fingerprinting (See Table 4).
- Enumerator Guidance: The module may serve as an operational guide for enumerators conducting complex crop-cut sampling or random walk procedures in the field. Random numbers can be generated directly within the data collection form, streamlining the process.

Table 5. Example of plot-level module for rice DNA fingerprinting in the VHLSS 2022 in Viet Nam

1	2	3	4	5	6	7	8	9	10

Household ID *[Note: Automatic filling]*	What is the ID code of the plot selected for crop sampling ? *[Note: Automatic filling]*	Plot Description [Enumerator: Use this name in the following questions] *[Note: Automatic filling]*	What is the cultivated area of the plot, in sq. meters?	How many varieties of rice were planted on this plot?	What is the name of the main variety planted on this plot? [Enumerator: continue with the main variety only]	What type of rice is the main variety planted on this plot? Traditional.....1 Improved.....2 Don't know.....3	Did you use certified seeds on this plot ? Yes…1 No …2	What is the source of the main seeds planted on this plot? Self-produced.....1 Farmer Group/Seed Club...2 Seed company....3 Research Institutes/Universities …4 Extension services ..5 Cooperative …6 Private stores/dealer …7 Other …8 If 1 >>> Q9 If not 1 >>> Next module	For how many seasons have you re-used the main variety planted on this plot? [Eumerator: If seeds are newly purchased, enter 0]
ID	**ID**		**AREA**	**NUMBER**	**NAME**	**CODE**	**CODE**	**CODE**	**NUMBER**

In the example outlined in Table 6, rice in Viet Nam is predominantly planted as a single variety. Farmers do use the concept of varieties to guide their planting decisions and the plots are relatively homogeneous. In these circumstances, we chose to ask a farmer about a main variety. From qualitative work prior to fieldwork, it is highly desirable to try and first understand this decision-making process for the crop and context in which you are sampling prior to confirming your sampling plans. There will be geographic variation and other sources of heterogeneity in farmer strategy. If in doubt, be conservative and assume that farmers don't have a lot of information about their planting material and its purity. In this scenario, we recommend three things. First, enumerators should be trained not to insist too strongly on soliciting a response to the number of varieties (ie. in table 6, Q5) and allow for "Don't know". Second, we should not list every possible variety name for Q6 in the CAPI and ask farmers to select from the list. Rather it is preferable to train the enumerators to record text of the response given by farmers. This text data will require cleaning in a later step to find "fuzzy matches" – where responses are very similar to the name of an improved variety and can be inferred to be equivalent – between what the farmer can tell us and the names of known varieties, but it is worth it to allow for genuine capture of what farmers know about their varieties and how they express it. Third, researchers could inquire about specific, observable traits the farmer believes the seed possesses. This variable can later be used to disentangle discrepancies between farmer self-identification and DNA fingerprinting results.

If farmers do state that they are cultivating multiple varieties in a given plot, you must decide whether to follow up on a) every "variety" the farmer considers they have; b) a sub-set of them (e.g. three "most important") or c) limit to what the farmer considers to be the "main variety". As you go from option a) to option c) there is a clear loss of information, but this is inversely proportional to the cost and time required for the data collection. The value of the additional information expected from sampling more farmer-declared varieties, and concomitant lower risk of committing false negative measurement error, must be traded off against these practical considerations. For highly mixed plots where farmers obtain planting material from multiple sources, we use a variety x plot module as outlined in the next section.

Variety x Plot module

In cases where you decide to sample from more than one farmer-stated "variety", it is worth collecting additional data for each variety. When combined with the results of the DNA fingerprinting, these additional data are

potentially of value to researchers, including closing the loop back to plant breeders. Do farmers perceive the traits that the variety was bred to have? To ensure there is no confusion about which variety is being referred to it is recommended that these additional details are collected as part of a plot visit for plant tissue sampling.

The example of beans in Uganda is outlined in Table 7 below. For each variety that the farmer says is cultivated in the plot, we collect additional information that can help us understand how the farmer perceives the benefits of the variety. In question K01_11 for example, we might include a long list of traits that relate to either the targets of breeding or are known consumer or producer preferences that shape decision-making about what to plant. These categories may all overlap in some fortunate cases. In other instances, there could be a trade-off between what is agronomically beneficial and what consumers demand. Examples here include yield, taste, nutritional qualities, cooking properties, disease and/or pest resistance, time to maturation, drought tolerance, storage quality, etc.

Table 6. Example of variety-plot module (beans in Uganda, collected as part of Ilukor et al, 2025)

K02_0	K02_1	K02_2	K02_3	K02_4	K02_5	K02_6	K02_7	K02_8	K01_11			K02_12			K03
What is the name of this [VARIETY]?	What % of the plot is planted to [VARIETY]?	What type of variety is [VARIETY]?	What growth habit does [VARIETY] have?	Is [VARIETY] biofortified (to be rich in iron & zinc)?	What type of bean seed did you obtain for planting [VARIETY]?	Did you obtain all new seed for this planting from outside your own farm?	For how many seasons have you re-used / recycled the seed?	From whom did you obtain the seed of [VARIETY] originally?	What do you like about [VARIETY]?			What do you dislike about [VARIETY]?			Can we please take a leaf sample?
RECORD THE NAME THEY TELL YOU AS ACCURATELY AS POSSIBLE		PLEASE READ THE OPTIONS Traditional= 1 Improved = 2 Don't know = 3	PLEASE READ THE OPTIONS Bush = 1 Climbing = 2 Semi-climbing = 3 Don't know = 4	Yes = 1 No = 2 Don't know = 3	PLEASE READ THE OPTIONS Certified = 1 Quality-declared = 2 Informal / Grain = 3 Don't know = 4	Yes = 1 >>Q8 No = 2 Don't know = 3	ENTER NUMBER IN SEASONS	NGO = 1 NAADS = 2 Gov'ment / OWC / MPS = 3 Farmer group = 4 Research center = 5 Local seed business = 6 Another farmer = 7 Grain market = 8 Agro-dealer = 9 Don't know = 10	LIST UP TO 3 REASONS, MOST IMPORTANT FIRST Refer to codebook for traits			LIST UP TO 3 REASONS, MOST IMPORTANT FIRST Refer to codebook for traits			PLEASE PICK LEAF SAMPLE FOR [VARIETY] AND CAPTURE BARCODE ON SAMPLE BOTTLE

Finally, it is useful to collect data on how confident farmers are in providing varietal data. From either of the modules outlined in tables 6 or 7, we could expand after asking the other questions to ask: "How confident are you?" with response codes on a 5-point Likert scale from very confident to not at all confident. These could refer to the module as a whole or could be targeted in some one or two important questions.

Section 6. Field-to-Lab Logistics: Getting the Process Right

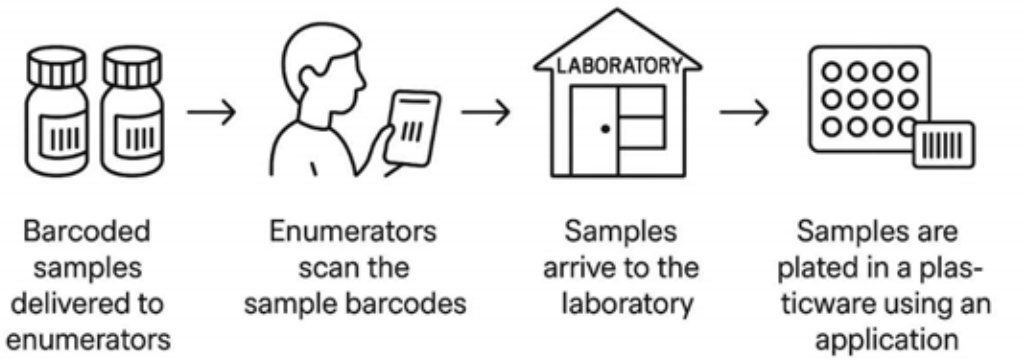
KEY POINTS

- Set up a barcoding system and make sure everyone knows how it works
- A tracking file takes up where a barcoding system leaves off – it is used to convey the barcode data associated with each sample in relation to the samples' location on the 96-well plates used by genotyping labs

Managing the Cold Chain: Keeping Samples Viable

The type of plant material collected has implications for field activities, material used, logistics and storage. Leaf samples are generally more challenging to collect than grains, due to the rapid deterioration of plant tissue. A barcoding system should be established along the chain, ensuring samples are tracked at each step. Appendix D provides an example protocol that can be adapted for field sample collection.

Figure 5. Sequential Workflow for Barcoded Samples



Barcoding

Establishing a reliable tracking system from the field to the lab is critical, and requires the use of barcodes or QR codes. Using barcodes allows tracking samples along the various step of the chain while also minimizing human errors involved in data entry. Several data collection applications (Survey Solution, ODK collect) offer barcodes question types, allowing the barcode ID to be attached to the questionnaire.

Before establishing a barcoding system, two important criteria to consider are barcode size and the number of duplicates needed per sample.

Barcode size: Along the tracking chain, barcodes may need to be attached to different supports (sampling kits or bags, laboratory tubes). It is thus important to avoid unfitted sizes for each use.

Technical replicates: Collected samples may be used for different purposes and/or sent to different locations for analysis. It is thus common to have an initial sample that is split into different samples. We recommend naming replicated barcodes with an underscore. For example, a replicate of sample 08733 should be named as 08733_1. In the case of duplicates (two samples of the same material that are independent), each sample should have its own unique barcode.

Box 4. Options for generating barcodes

Here, we rank available options from "higher cost, higher convenience" to "lower cost, lower convenience":

i) This service is offered by companies. The product can generally be delivered on-site, thus facilitating logistics. The cost can vary from $10 to $15 per 100 barcodes.

ii) A label printer can be purchased, along with its label management software. The portability of these devices allows in-field barcode generation. Zebra offers a range of relevant products (printers, ribbons, and labels). Zebra specific (Zebra Designer) barcode generating software for about $350 while a desktop barcode printer average price is from $300 to $600. The DYMO LabelWriter along with the DYMO Label software is a portable option.

iii) A Do-It-Yourself approach is possible, by purchasing blank custom printed labels (ex: AVERY L7651) and generating the barcodes with Microsoft Word[1]. Each page generally contains 50-75 printable stickers. Several label formats are incorporated into Microsoft Word, allowing the user to generate a document with the barcode IDs matching the format of the label. The list of barcode IDs is generated through Microsoft Excel. An office printer can then be used to print the barcodes.

Field survey conditions might lead to deterioration of the barcode reading. Stickers without protection may suffer from transport, rain, or dust. We thus recommend not to stick barcoded samples are given an additional layer of protection – some kind of simple plastic covering material depending on the nature of the samples.

It is also important to check the suitability of devices used for barcode reading. Some devices or applications may not be correctly configured to read barcodes or may have a deteriorated camera resolution. All steps of the field data collection protocol used should be piloted, including the successful deployment of barcoding.

Preservation (leaf/grain sampling)

Seed or grain samples are usually stable if they are sufficiently dry. They can yield DNA of sufficient quality provided they reach the extraction laboratory in a timely manner. If leaf samples are used, sufficient measures should be taken to rapidly dry the tissue to prevent growth of mold as well as degradation of DNA. A desiccant such as silica gel should be placed in the sample container prior to harvesting to rapidly absorb moisture from the leaf samples. Sufficiently dry samples can be handled and processed at ambient temperatures.

Cleaning material

For leaf material, samples should be collected by punching out small discs. After each punch, enumerators should clean the leaf punch with an ethanol-based wipe. To reinforce this procedure, a note or reminder page should be included in the digital survey form.

Sample processing

Seeds can be reduced to flour with a grinder and should be ground to a uniform "espresso coffee" consistency. Submitting material of heterogeneous particle size complicates DNA extraction and may affect results. . While only 2mg of tissue is necessary for extracting DNA, it is advisable to grind 20 to 50 gr of material to ensure maximum representativeness of the collected material. To avoid contamination, the grinders must be carefully cleaned after sample processing with a brush and ethanol.

Keeping remnant samples

It is recommended that sampling should be done to allow for maintenance of remnant samples. Thus, if the protocol requires submission of two leaf discs for genotyping, four discs can be harvested and placed in the sample collection tube. After plating, two leaves will remain in the tubes. These remnant samples should be safely stored at least until the genotyping results are received. The remnant samples are an important backup that can be handy in the event of loss or damage of samples during shipping or if a problem occurs during genotyping. Having backup tissue prevents the need to re-visit the field to collect tissue again.

Preparing the samples for genotyping

Plating

In the physical flow of samples from the field to the laboratory, the barcode conveys the sample ID. The enumerators will capture the barcode ID at the time of sampling, and those digital IDs stay with the remaining household, plot, and variety-level databases, ready to be merged with the eventual output from the DNA fingerprinting exercise – the varietal assignment calls.

The process of plating the submitted samples makes use of a **sample tracking file** which is used to describe the physical location of each sample within the 96-well plates that are the workhorses of the genotyping world. A sample tracking file contains the following fields (see table 8 below with an example for sorghum).

Table 8. Tracking file fields

Plate ID	Row	Column	Organism	Species	Genotype
1	A	1	Sorghum	Sorghum bicolor (L.) Moench	114608
1	B	1	Sorghum	Sorghum bicolor (L.) Moench	114615
1	C	1	Sorghum	Sorghum bicolor (L.) Moench	114622

Each plate contains 96 wells and is numbered from 8 rows (A to H) and 12 columns (1 to 12). Laboratories usually use wells G12 and H12 for controls and these should thus stay empty. Genotype corresponds to the sample ID from the barcode, regardless of whether it is part of the reference library or the field samples. In general, it is advisable to cluster reference library samples together on the same plates in case they require additional analyses prior to, or sometimes subsequent to, the field samples.

Box. 5: Plasticware – a cautionary tale

It is critical to ensure that the samples are plated in the exact plasticware required by the lab to avoid downstream difficulties. This entails getting confirmation from the point person at the genotyping lab about the currentness of their recommended catalogues for procurement. In the Bangladesh rice study (Singla et al, 2025), for example, an out-of-date catalogue for purchasing plasticware was shared by the service provider. Following the catalogue, we procured plate covers that failed to fit on the 96-well plates, necessitating workarounds for sealing the plates. Eventually, aluminum heat-sealing foils were procured but in the absence of proper plate sealing machinery, the local team in Bangladesh had to iron the seals on to the plates. Several plates were unavoidably damaged (melted at the top) in this process which initially led the genotyping service provider to reject the samples as the robotics in the lab could not penetrate fully into the wells. However, we were able to pay for the time of a technician at the genotyping lab to painstakingly remove excess melted plastic with a scalpel, thereby rectifying the heat damage and making the plates amenable for high-throughput processing.

The tracking file is sent to the laboratory or uploaded to their online portal. We recommend using the **Coordinate app** to generate the tracking file during the process of plating of the samples. Appendix E gives instructions of how to use the Coordinate app to track the plating process. This step is so important to make sure that samples are correctly positioned in the 96-well plates. From that point on, the barcodes can no longer help us identify the samples – the individual tubes in the 96-well plates are too small to have a barcode sticker attached to the, so we rely on the sample IDs from the barcodes being entered in the correct place in the tracking file. We rely on the app to scan the barcodes during plating to avoid transcription errors – just as we do when entering the sample IDs during fieldwork using the CAPI application barcode scanners.

DNA Extraction

DNA extraction should ideally be bundled with genotyping service at the same service provider. This can usually be added to the genotyping invoice at an additional cost of approx. $1 USD per sample and all handled with robotics in the lab. The fewer steps in which humans have to handle the materials the better.

Occasionally, there may be limitations on movement of plant material out of a country, necessitating extraction of DNA in-country prior to shipping for genotyping in a second country. This is a sensitive step that can potentially introduce errors if not handled correctly. In such a situation, we recommend partnering with a lab with good experience in extracting DNA from plant samples. Such a lab may be present in CGIAR centers or in the NARS. The lab will need to provide DNA that meets the requirements of the genotyping service provider in terms of quantity and quality.

Shipping DNA out of the country may also necessitate considerations such as government requirements like material transfer agreements or non-infectious certificate for export. Additionally, the integrity of the DNA during shipping may need to be considered, possibly necessitating freeze drying of the DNA or shipping in dry ice.

Section 7. Choosing and Using Genotyping Services

KEY POINTS

- The most effective strategy is to leverage an existing, validated SNP marker panel from research institutions and select a service provider with a technology platform that is compatible, cost-effective, and suited to your specific crop needs.
- Critically, all reference varieties and field-collected samples must be genotyped using the exact same marker panel and technology to ensure the resulting datasets are directly comparable for accurate identification
- If in doubt, ask for advice – these are highly technical decisions

Identifying Suitable SNP Markers

Selecting the right genotyping platform is a critical decision for ensuring the generation of reproducible, cost-effective, and analytically reliable datasets. For most major crops, the easiest way is to use a publicly validated and widely used marker panel. These are curated SNP sets that have been selected for their ability to differentiate varieties. Various research programs and institutions have developed such resources for applications like quality control (QC) studies, genetic diversity studies, marker-assisted selection (MAS), genome-wide association studies (GWAS), and genomic selection (GS).

Marker panels are usually categorized by their density - number of SNPs they contain. Low-density panels typically have < 1,000 SNPs and offer the advantage of lower per-sample costs and can comfortably distinguish between genetically distant varieties. Conversely, they lack the resolution to reliably separate very closely related varieties such as sister lines. Mid-density panels contain between 1,000 and 20,000 SNPs and in most cases have sufficient resolution for effective separation of even closely related varieties.

Information on availability of such panels can be obtained by:

- Consulting CGIAR centers and crop-specific research institutions (e.g., IRRI for rice, CIMMYT for maize, CIAT for cassava and IITA for banana).
- Exploring platforms like the Excellence in Breeding (EiB) Platform, which catalogs and provides access to validated genotyping resources.
- Contacting the original authors of major genetic studies for a target crop.

Worth noting, an established panel can often be ported to different commercial genotyping service providers. For instance, the RICA v4 rice panel used in the Vietnam and Bangladesh studies (Table 2) is available as a low-cost service through Agriplex Genomics, but can also be accessed as a DArTag panel at Diversity Arrays Technology.

Choosing a Service Provider

Once a marker panel is identified, you need a service provider with the right technology to process the samples. This will usually entail implementing a form of targeted genotyping approach such as KASP, DArTag or amplicon sequencing. These methods use molecular probes designed to detect a specific, pre-defined set of SNPs (based on the identified panel) accurately and cost-effectively. Providers like Intertek (KASP) and Agriplex (amplicon sequencing) specialize in these targeted assays.

A majority of genotyping in our previous studies (refer to Table 2) have relied on the DArT portfolio of sequencing technologies that includes:

- DArTseq: A high-density method for marker discovery and detailed genetic profiling.
- DArTseqLD: A low-density, cost-effective version for routine screening using a subset of informative markers discovered through DArTseq.
- DArTag: A targeted method, analogous to KASP, that uses DArT's sequencing technology to score a specific, pre-selected set of markers.

A key advantage of sequencing-based platforms like DArT is their suitability for analyzing bulk samples (essential for outcrossing crops like maize). By counting the sequence reads for each allele, the platform generates a proxy for allele frequency. This profile of a bulk field sample can then be matched to the profile of a bulked reference sample, enabling possible identification of heterogeneous varieties.

Options for Marker Discovery

In some cases, especially for orphan crops or regionally specific landraces, a suitable public marker panel may not exist. In other instances, such as the case for clonals with low diversity, existing panels might fail to differentiate varieties sufficiently. In this situation, a marker discovery phase may be necessary. However, this is a more intensive and expensive undertaking that is essentially a research project in itself. The typical approach involves:

1. Assembling a discovery panel consisting of the reference varieties and a representative selection of field-collected samples.
2. Sequencing this panel using a high-throughput platform that allows for *de novo* SNP discovery such as:
    - Genotyping-by-Sequencing (GBS) approaches (DArTseq, GBS): Here, small, representative fractions of the genome are sequenced enabling the simultaneous discovery and scoring of thousands of SNPs without prior knowledge. This is the most common and cost-effective method.
    - Whole-Genome Sequencing (WGS): This is the most comprehensive but also the most expensive option. WGS provides up to base-pair level resolution across the entire genome, offering the maximum possible number of markers. It is typically used for foundational genomic resource development rather than routine screening.
3. Bioinformatic Analysis: The sequence data is then analyzed to identify high-quality, informative SNPs that can reliably distinguish the varieties of interest. These SNPs form the new custom panel.

Regardless of the platform or panel selected, it is critical to ensure that reference varieties and field-collected samples are genotyped with the exact same marker panel and, ideally, on the same technology platform. Use of different genotyping strategies will result in generation of incomparable datasets.

Seek advice

SPIA can play an important role matching up the crop and country context to the specific genotyping service provider. Don’t hesitate to reach out for advice.

Section 8. Making Sense of the Data: Interpretation and Outputs

KEY POINTS

- Several different analytical approaches can be used in making varietal assignment "calls" for each sample
- While drawing from the same data sources (genotyped reference profiles and genotyped field samples) different analytical pipelines can yield different results
- If in doubt, seek advice from SPIA on how to analyze the genotyped data

Bioinformatics of DNA fingerprinting

At least two datasets are needed for routine varietal identification. The first dataset describes reference library samples at the "bin" level; the second is at the field samples level. In the case where all the genetic material in the reference library could be distinguished, each "bin" will simply represent a single unique reference sample. In some cases, some reference varieties may be genetically indistinguishable from each other – in these cases they are group together in a "bin" (See bin 1 in Table 9).

Table 9. Example of genotyping output describing the reference library bins used for varietal matching

Bin ID	References in Bin
1	Nase 19, NASE14
2	Nase 3
3	OFUMBA_CHAI

Several analytical pipelines are available for processing genotyping output, including genetic distance-based assignment, purity, Diversity Arrays Pipeline (DAP) and clustering. The choice of pipeline depends on whether the samples are made up of tissue from single plants or whether bulked samples were collected and also, in the case of high diversity outcrossing crops like maize. For the former, particularly in the case of inbreds and clonally propagated crops, genotype data that show the allele state at each locus (i.e score data) are sufficient. Most service providers give data in this manner. For the latter, it is best to use a service provider that gives the results in form of counts of alleles of each state at every locus. Such datasets are most conveniently obtained from DArT genotyping platforms.

Assignment by genetic distance

Distance based assignment uses score data and works by comparing the similarity between the samples and references at each locus. An example of a genetic distance method is identity by state (IBS). This involves determination of the probability that alleles at each locus are of the same state. Another example of a genetic distance method is the Hamming distance. This depends on the presence of minimal changes of alleles at loci being compared. Further, an approximation of purity between samples and references can be done by calculating pairwise comparisons between all samples and all references to determine the percentage similarity.

DAP, Purity and Clustering methods

These methods use the allele count data for both references and field samples to compare proportions of allele frequency similarity between the samples and each reference to determine the best matching reference. The final result is presented in the form of percentage similarity between the two.

In an ideal situation where a comprehensive set of distinct references is available and the samples are without impurities, all the methods perform comparably well. However, variation of performance and sensitivity can be observed in less ideal circumstances which likely represent a majority of cases for the CGIAR. With equal availability of all pipelines, it is recommended to use them all in initial exploration of datasets. Case by case evaluation can inform the optimal choice of pipeline when divergence of output is observed.

Box 6. Divergent results from different pipelines

Particularly in African smallholder settings, we see a preponderance of farmer plots that contain admixtures of genetically distinct plants rather than pure stands of a single variety. This necessitates collection of bulked tissue (e.g., multiple leaves or grains from a single plot) for genotyping, to capture a representative genetic profile.

While bulk sampling is a pragmatic solution to field heterogeneity, it shifts the analytical bottleneck from logistics to data interpretation. A bulk sample from an admixed field can produce a complex genetic signal broadly resulting in three analytical outcomes that pipeline must handle:

1. Clear Match: The sample is a pure or near-pure representation of a known reference variety.
2. Imperfect Match: The sample is a mixture, but with a predominant variety. It contains a strong signal for a known reference but is contaminated with other genetic material.
3. Unassignable: The sample is a complex mixture with no clear primary component, or it represents a variety completely absent from the reference panel.

While most pipelines agree on Scenario 1 and, to some extent Scenario 3, the handling of Scenario 2 is the primary point of divergence. This ambiguous middle ground is where the choice of algorithm most significantly influences the final adoption figures.

A **clustering pipeline** is well positioned studying a well-defined technology. A variety is a technology with a specific package of traits; yield potential, drought tolerance, etc. Thus, when a sample is 50% Variety X and 50% unknown landrace, it is no longer the technology the breeders released. Assigning the sample the label Variety X may be misleading because it will not perform as expected. In such situations, the low assignment rate from clustering is an accurate measurement of how few farmers are cultivating the technology in its pure, effective form. In the event of shocks (such as drought or disease), the measured output from the clustering pipeline, gives a reliable, estimate of the adoption of the intact technology package. Where clustering is open to critique is when we are interested in understanding instances of imperfect adoption.

Alternative best-match based methods extend assignment analysis further into scenario 2, accompanied by a risk of generating false positive results. These alternative methods recognize that adoption in a complex agricultural landscape is not a simple binary event. A farmer's field with a predominant improved variety remains a valid instance adoption worthy of being documented, albeit an imperfect one. It reflects the reality of informal seed systems and a reality that is useful to capture. Such an approach enables the quantification of imperfection, since the resultant scores give a probabilistic statement of evidence. That is, the evidence for a sample being an imperfect example of a particular variety is generated, but is not guaranteed to be correct.

A case can be made for the complementary use of both pipelines, providing for good analytical flexibility. For instance, modelling of impact can be based on mutual assignment by both clustering and best match pipelines. A conservative assignment using only adoption rates based on high confidence assignment using clustering can tell us how widely adopted varieties are in a form that are highly likely to have “varietal fidelity”. Yet at the margin there may be a large share of samples that are not assigned varietal identity using clustering but for which varietal assignments can be made based on a best-match pipeline with probability scores.

This larger pool of varietal assignments will represent a mix of false and true positive varietal assignments. The relative share of these false vs true positive assignments will determine the rigor of using best-match pipeline. However, we do not consider conservative assignment using clustering to be synonymous with rigor until such a time as we have evidence that clustering does not result in a lot of false negatives, thereby leaving many imperfect varietal samples unassigned that could and should be properly assigned with a predominant varietal identity using an alternative pipeline. We believe that experimental work, using blinded controlled ratios of admixture that simulate farmer planting behavior for bulked samples, is the best way to make further progress in this regard.

Having two rival adoption rates – one based on clustering, one based on best-match methods – being published for each DNA fingerprinting exercise allows for a more nuanced understanding of adoption levels, acknowledging the complexities of real-world agricultural systems. SPIA already makes use of lower and upper bound estimates for each country study, reflecting different degrees of certainty in the estimates of the reach of CGIAR. These bounded estimates can be quite far apart but help tell a nuanced story. One practical proposal for moving forward is that only varietal assignment rates using the clustering approach enter the lower bound estimates; additional assignments using best-match methods enter the upper bound for estimated reach of CGIAR.

Data output

An example of final assignment results is shown in Table 10. The samples are arranged in rows. The information on the first row should be read as: sample identified with barcode 600725 in well position A1 was matched to reference variety Nase 19. The IBS genetic distance between sample and reference is 1%, the genotyped alleles are 96% similar (Purity). The selected reference is in Bin 1 which contains the varieties Nase 19, NASE14. The varieties in the bin are highly similar to each other.

Table 10. Example of data output for varietal matching.

Barcode ID	Well	Top Reference	IBS	Purity	Bin ID	References in Bin
600725	A1	Nase 19	0.01	0.96	1	Nase 19, NASE14
600464	A2	Nase 3	0.01	0.97	2	Nase 3
603763	A3	OFUMBA_CHAI	0.02	0.95	3	OFUMBA_CHAI

As outlined in box 6 above, varietal assignment calls can vary over different analytical pipelines. SPIA can help with the process of evaluating the optimal choice of pipeline and its associated output. The final output most social scientists are interested in is a simple column of data with the top reference match for each sample, and the associated barcode ID for each sample. The barcode ID column can then be used to merge with the corresponding sample IDs in the household / plot / variety level dataset, and you are ready for analysis.

References

Abay, K.A., Abate, G.T., Barrett, C.B. and Bernard, T., 2019. Correlated non-classical measurement errors,'Second best'policy inference, and the inverse size-productivity relationship in agriculture. *Journal of Development Economics*, *139*, pp.171-184.

Alemu, S., Ambel, A., Khanal, A., Kosmowski, F., Stevenson, J., Taye, L., Tsegay, A. and Macours, K., 2024. SPIA Ethiopia Report 2024: Building Resilience to Shocks. *Rome: CGIAR Standing Panel on Impact Assessment (SPIA)*.

Annunziato, A. 2008. DNA Packaging: Nucleosomes and Chromatin. Nature Education 1(1):26

Bannert, M. and Stamp, P., 2007. Cross-pollination of maize at long distance. *European Journal of Agronomy*, *27*(1), pp.44-51.

CGIAR, 2023. Protocol for Crop Cut Surveys Yield Gap Decomposition Add-on. https://cgspace.cgiar.org/server/api/core/bitstreams/063e23dc-7580-456a-8538-7d8df30c060b/content

Ilukor, J., Letaa, E., Khanal, A., Barros, J., Taye, L., Gimode, D., Ponzini, G., Asea, G., Ssennono, V., Stevenson, J.R. and Lybbert, T., 2025. SPIA Uganda Report 2025: Agricultural Diversity Under Stress. *Rome: CGIAR Standing Panel on Impact Assessment.*

Jovanovic, N. and Ricker-Gilbert, J., 2025. Estimating the direct and indirect effects of improved seed adoption on yields: Evidence from DNA-fingerprinting, crop cuts, and self-reporting in Ethiopia. *Journal of Development Economics*, *174*, p.103466.

Kosmowski, F., Alemu, S., Mallia, P., Stevenson, J. and Macours, K., 2020. Shining a brighter light: Comprehensive evidence on adoption and diffusion of CGIAR-related innovations in Ethiopia. *Rome: CGIAR Standing Panel on Impact Assessment.*

Kosmowski, F., Chamberlin, J., Ayalew, H., Sida, T., Abay, K. and Craufurd, P., 2021. How accurate are yield estimates from crop cuts? Evidence from smallholder maize farms in Ethiopia. *Food Policy*, *102*, p.102122.

Kosmowski, F., Le, T.B., Chavez, S., Nguyen, T.H., Nguyen, P., Gimode, D., Biradavolu, M., Pelletier, J., Stevenson, J., Visaria, S. 2024. SPIA Viet Nam Report: Global Ambitions, Sustainable Pathways. *Rome: Standing Panel on Impact Assessment (SPIA).*

Lobell, D.B., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T. and Murray, S., 2020. Eyes in the sky, boots on the ground: Assessing satellite-and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics*, *102*(1), pp.202-219.

Morell, M.K., Peakall, R., Appels, R., Preston, L.R. and Lloyd, H.L., 1995. DNA profiling techniques for plant variety identification. *Australian Journal of Experimental Agriculture*, *35*(6), pp.807-819.

Occelli, M. et al. (2025). Varietal knowledge and genomic uncertainty in adoption studies using DNA fingerprinting [Presentation]. Cornell University; EQUAL Lab.Poets, A., Silverstein, K., Pardey, P., Hearne, S. and Stevenson, J., 2020. DNA fingerprinting for crop varietal identification: Fit-for-purpose protocols, their costs and analytical Implications. *International Maize and Wheat Improvement Center (CIMMYT), CGIAR Standing Panel on Impact Assessment (SPIA)*.

Singla, S., Islam, T., Hassan, F., Monteiro, I., Stevenson, J., Emerick, K. 2025. SPIA Bangladesh Study: Updating the Green Revolution. Rome: CGIAR Standing Panel on Impact Assessment.

Stevenson, J., Gantier, M., Traxler, G., Kosmowski, F., Macours, K. 2023. The Challenge of Tracking the Reach of Post-Green Revolution Crop Breeding, 07 June 2023, PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.3.rs-3028333/v1

Terzi, V., Morcia, C., Gorrini, A., Stanca, A.M., Shewry, P.R. and Faccioli, P., 2005. DNA-based methods for identification and quantification of small grain cereal mixtures and fingerprinting of varieties. *Journal of Cereal Science*, *41*(3), pp.213-220.

World Bank 2020. World Bank. Ethiopia Socioeconomic Survey (ESS) 2018-2019. Microdata Library. https://microdata.worldbank.org/index.php/catalog/3823

Appendices

Appendix A. Large-scale DNA fingerprinting: Example of barley, maize and sorghum in ESS 2018/19 in Ethiopia

The timeline of this 2-year exercise is indicated in Table 11 while major steps are detailed in the text below.

Table 11. Timeline of activities to collect and genotype crop samples in the ESS 2018/10. Fieldwork indicated in grey

Activity	2018												2019											
	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
Collection of the reference library	X	X	X	X																				
Samples logistics (3000 bags with barcodes)				X	X																			
ESS questionnaire design					X	X	X																	
Training of Trainers								X																
Training of enumerators / distribution of sampling material									X															
Crop cuts, sample collection and drying of									X	X	X	X												
Reception of seed samples													X	X	X									
Sample grinded into floor (four technicians)															X	X								
DNA extracted and shipment (two RAs using																		X	X	X				
Quality control after sample reception																						X		
DNA genotyped and results obtained																							X	X

The Ethiopian Socio-economic Survey

The Ethiopian Socio-economic Survey (ESS) is a nationally representative household panel survey conducted by the Ethiopian Central Statistical Agency (CSA) in collaboration with the World Bank LSMS-ISA team. In 2018/19, the fourth wave of the ESS was collected along with an additional crop sampling protocol, for DNA fingerprinting varietal identification.

Given the geography of improved seed distribution, crop samples of barley, maize, and sorghum were collected in the regions of Amhara, Dire Dawa, Harar, Oromia, SNNP, and Tigray. The collected DNA fingerprinting sample is representative at the household level across major growing areas.

Collection of the reference library

For maize, the reference library for Ethiopia was previously compiled under a CIMMYT/EIAR DNA fingerprinting research project funded by the Bill and Melinda Gates Foundation. As there were no readily available reference libraries for barley and sorghum, we compiled collections of breeders' seed from the Ethiopian Institute of Agricultural Research (EIAR) and its regional centers. The reference library comprised 41 of the 46 food barley varieties released and 17 of the 19 malt barley varieties released in Ethiopia since 1990. A total of 29 varieties were included in the reference library including all varieties that are still under production by EIAR (42 varieties have been released since 1990). Data on improved seeds released and included in the reference library are available in [2020 report].

Training of trainers and enumerators

The procedure for crop sampling and barcording was demonstrated to supervisors and enumerators in September 2019.

Samples logistics (3000 bags with barcodes)

A total of 3000 barcoded cotton sample bags were provided to enumerators and an additional barcode question was integrated into the crop-cut module. Cotton samples were manufactured in Addis, with a plastic pocket sewn on the cotton bags. Barcodes were generated with Microsoft Word, printed on blank custom printed labels, and added to sample bags.

Each Enumeration Area received a large plastic bag containing 60 sampling bags already barcoded. These large bags had been barcoded but the plastic did not survive field operations and this level of information was lost.

Crop cuts, sample collection, and drying of the samples

The ESS is among the rare agricultural surveys that implement crop-cuts for yields measurement, where resident enumerators are hired for survey work. In each Enumeration Area (EA), plots grown during the agricultural are listed for 21 temporary crops. Crop-cut plots are then randomly selected among ESS households in the EA, for up to ten plots per crop. Enumerators are trained on implementing a specific procedure where a 4 x 4 m quadrants is randomly laid over the plot. Once the plot is harvested, enumerators will weigh the total amount of crops harvested from the quadrant; while the dry weight will be taken after two weeks.

Reception of seed samples

While the framework for collecting dry samples was already in place, the protocol involved some logistical challenges regarding sample transportation. The operation relied on CSA branches and was in communication with CSA headquarters to insure the operation smoothness. The advancement of sample collection by enumerators was followed online once data from Survey solutions questionnaire was uploaded to servers. The collaboration with 20 CSA branches resulted in a total of n = 1147 samples obtained (n = 261 barley, n = 518 maize and n = 368 sorghum). Only a few samples (n = 12) went missing and could not be recovered.

Table 12. Summary of collected sample per CSA Branch

Region	CSA Branch	N Barley	N Maize	N Sorghum	N
Oromia	Adama	14	13	5	32
Oromia	Ambo	18	11	4	33
SNNP	Arba Minch	12	16	19	47
Oromia	Aseba Teferi		14	22	36
Benishangul Gumuz*	Asosa*		4		4
Amhara	Bahir Dar	17	48	6	71
Amhara	Desie	47	40	32	119
Amhara	Debre Berhan	13	1	3	17
Dire Dawa	Dire Dawa		2	31	33
Oromia	Goba		2	3	5
Amhara	Gondar	21	48	32	101
Harari	Harar	1	77	67	145
SNNP	Hawasa		7		7
SNNP	Hosaena	21	28	26	75
Oromia	Jimma	11	37	14	62
Tigray	Mekelle	66	62	46	174
SNNP	Mizan Teferi	6	16	6	28
Oromia	Negelle	4	27		31
Oromia	Nekemte	4	21	9	34
Tigray	Shire		35	34	69
SNNP	Sodo	6	9	9	24

* Benishangul Gumuz was not part of the regions targeted for DNA fingerprinting, but one EA located in Oromia was covered by this branch

Sample processing: grinding and DNA extraction

After sample processing (50 grams of flour), DNA was extracted in ILRI's laboratory in Addis Ababa. Two research assistants (MSc level) were hired for the task that was completed in three months using Qiagen Dneasy plant mini kits. The concentration of samples was adjusted before shipment. The results of extraction indicated a lack of quality DNA for barley samples only, and these samples were processed again.

The Ethiopian Institute of Bioversity Conservation was contacted to obtain all necessary authorizations for sample shipment. Tracking files were submitted online and samples were shipped to Australia for genotyping using DArT portfolio of sequencing technologies.

Appendix B. Cost estimates

Table 13. Costs associated with DNA fingerprinting of barley, maize, and sorghum in the ESS 2018/19 (in 2019 USD)

Item	Unit of measurement	Qty (no. of units)	Total Cost ($)	Unit cost ($)
Assumptions				
Total # of Enumeration Areas (EA)	/	197	/	
Maximum number of plant sample per EA	/	25	/	
ESS survey			$50,000	$43.55
Survey costs contribution	Invoice			
Survey supplies			$4,433	$3.86
Avery Dennison Mini Labels Laser L7651-25	Pack 1625	10		
Cotton sampling bags with sewed plastic cover	Bag	4925		
Plastic bags	Bag	197		
Sample processing			$911	$0.79
Supply costs				
Grinder	Item	4		
Tissue, alcohol, brushes	Item	5		
Staff costs				
Research Lab Technician	Person-month	4		
DNA extraction			$13,773	$12.00
Laboratory costs				
Space charge	Month	3		
Supply costs				
DNeasy Plant Mini Kit	Pack 250	6		
Microcentrifuge Tube 2.0 ml	Pack 500	6		
Micro Tube 1.5 ml	Pack 5,000	4		
Tip 50-1000µl	Pack 5,000	1		
Micro Tip, 0.1-10ul	Pack 1,000	3		
Thermo-Fast 96 PCR Plate Skirted	Pack 25	1		
Staff costs				
Research Associate I / JG 10	Person-month	6		
Genotyping			$78,623	$68.49
DaRT Barley Varietal ID and Purity (1.0)	Sample	261		
Sorghum DArTseq PH (1.0)	Sample	509		
Maize DArTseq PN (1.0)	Sample	378		
Total direct costs (USD)			$147,740	$128.69

Note: These estimates does not include cost associated with the reference library construction and project management staff. Exchange rate of August 2019 (ETB 1 = $ 0.034) was used. The total cost for the ESS survey was $1.8 m USD.

Appendix C. Seed accessions requests to CGIAR genebanks

Table 14. Overview of the CGIAR Genebank platforms where seed material can be requested

Genebank	Crop	Location	Accessions	Seed requests
Africa Rice	Rice	Bouake, Cote d'Ivoire	22,000	http://eservices.africarice.org/argis/
Alliance CIAT Bioversity	Banana	Leuven, Belgium	1,500	https://www.crop-diversity.org/mgis/mylist
Alliance CIAT Bioversity	Beans, cassava, forage grasses	Palmyra, Colombia		https://alliancebioversityciat.org/genebank-germplasm-requests
CIMMYT	Maize	Mexico	28,000	https://www.cimmyt.org/seed-request/#maize
CIMMYT	Wheat	Lima, Peru	150,000	https://www.cimmyt.org/seed-request/#wheat
CIP	Potato and sweet potato	Lima, Peru	17,898	http://genebank.cipotato.org/gringlobal/files/How%20to%20request%20material.pdf
ICARDA	Wheat, barley, chickpea, lentil, faba bean and grass pea	Rabat, Morocco and Beirut, Lebanon		https://icarda.org/research/genetic-resources
ICRAF	Forage and multipurposetrees	Nairobi, Kenya	190	http://www.worldagroforestry.org/products/grunew/index.php/seeds
ICRISAT	Sorghum, millet, pearl millet, chickpea, pigeon pea, and groundnut	Patancheru, India	123,000	http://genebank.icrisat.org/Common/Viewer?ctg=Section&ref=Sec7
IITA	Cowpea, cassava, banana, yam, soybean, groundnut, and maize	Ibadan, Nigeria	28,000 (including 15,122 cowpea)	http://gringlobal.iita.org/gringlobal/search.aspx
ILRI	Forages species	Addis Ababa, Ethiopia	19,000 accessions of over 1,000 species	https://genebank.ilri.org/gringlobal/search
IRRI	Rice	Los Baños, Philippines	132,000	https://www.irri.org/rice-seeds

Appendix D. An example of a field sample collection protocol

Below is an example of how a field sample collection protocol can be developed to serve as a checklist and guide for enumerators to collect material for genotyping during a household survey. The example of rice – an inbred, is used.

Requirements for purchase

- Sticky labels
- Sampling tubes/container (30ml) – According to household numbers + 20%
- Silica gel, a desiccant – 4g x number of pots
- Leaf punch (5cm diameter) – According to number of enumerators
- Alcohol based wipes
- Sample holding bags/boxes
- 96 well plates and seals

Pre-survey

Identify number of households

Print barcode stickers according to number of expected samples (e.g. number of households x 2.5 varieties on average)

Stick barcode on each 30 ml container

Pre-load 30 ml containers with dry silica gel (4g per container)

Make sure all enumerators have the appropriate barcode scanning app on their tablet

Make provisions for extra devices in case of malfunctions

Field survey checklist

Barcode-labelled 30 ml containers pre-loaded with silica gel

Tablets with CAPI application AND barcode scanning app installed

Leaf punch

Alcohol based wipes

Procedure

- For error tracking, a proportion (20%) of the plots will have two samples collected from them as technical replicates. The enumerators should have the same number/proportion of plots from which duplicate samples will be collected.
- Prior to sample collection, ensure that the silica gel granules in the tubes are optimally dry as indicated by the manufacturer stated color indication for dry silica gel.
- Identify the plot from which sample is to be collected

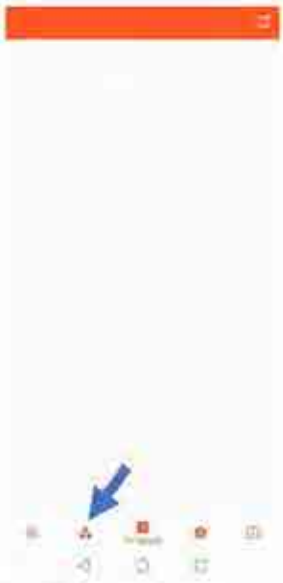
- For duplicate sample collection, Identify two healthy plants (with no obvious signs of disease on leaves). The plants should be away from the plot edges, preferably from the center of the plots.
- Excise a young growing leaf from the first plant, fold it 4 times to allow for collection of 4 leaf discs and punch
- Insert the leaf disc into the labelled collection container corresponding to the sample and seal tightly.
- Repeat the procedure for the second plant to result in collection of two samples per plot.
- Note that for duplicate sample collection, the leaf discs from each plant are stored in separate collection bottles.
- The procedure for sample collection for plots with no duplicate sampling is the same, except that only one plant will be identified and sampled, resulting in one sample per plot.
- Once the samples are harvested, they can be handled and shipped at ambient temperature.
- At the lab, received samples will be scanned to monitor any missing sample or tracking error that may have occurred. With the aid of the coordinate app, the leaf discs will be transferred to 96 well plates and a sample tracking file will be generated.

Appendix E. How to use Coordinate App for plating and creating tracking files

The coordinate app is useful for streamlining the plating of samples onto 96 well plates and it helps generate a tracking file that can accompany the samples. The app is compatible with Android phones and tablets and can be downloaded from the Play Store.

1. Download the app from the Play Store

2. Skip Storage location definer
   - Cancel loading available sample data



3. Click Templates

4. Click New Template

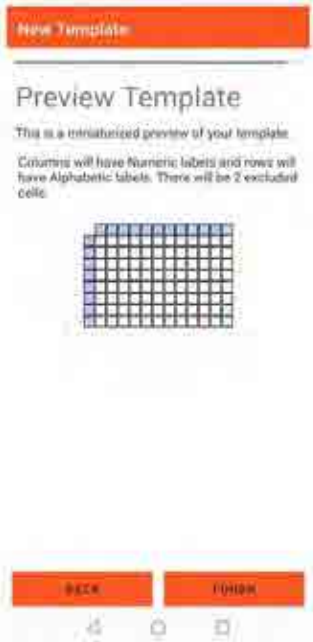
6. Fill Template name

7. Click Next

8. Select "Selection"

9. Select 7-12 and 8-12 to exclude

10. Keep these defaults



11. Confirm H12 and G12 are excluded

- Click Finish

12. Rice DNA template is created

- Select Projects

13. Name it Referencing Plating

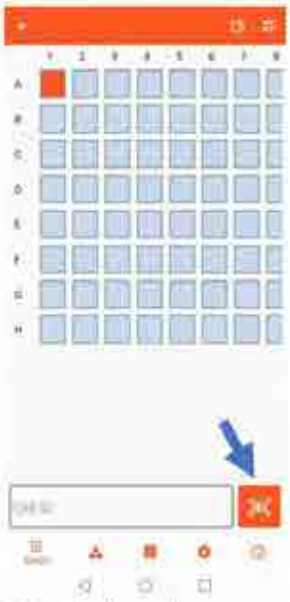
- Later, create a second project for Field Samples

14. Add Grid

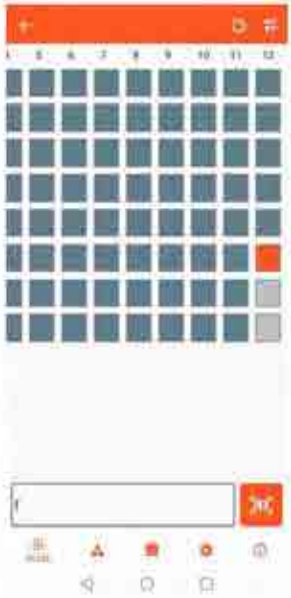
15. Select Rice DNA Template

16. Enter Plate ID for Identification

- Fill name



17. Use the barcode scanner to select sample names



18. The Grid fills for 94 samples as required and is ready for export

- Select back arrow to export

19. Filled plates ready to export

- Click save button to export
- Accept the file name
- It will export to "Exports" folder in internal storage

	A	B	C	D	E	F
1	Value	Column	Row	Identification	Person	Date
2	Sample_1	1	A	B_Os_1	Davis	6/9/2024
3	Sample_2	1	B	B_Os_1	Davis	6/9/2024
4	Sample_3	1	C	B_Os_1	Davis	6/9/2024
5	Sample_4	1	D	B_Os_1	Davis	6/9/2024
6	Sample_5	1	E	B_Os_1	Davis	6/9/2024
7	Sample_6	1	F	B_Os_1	Davis	6/9/2024
8	Sample_7	1	G	B_Os_1	Davis	6/9/2024
9	Sample_8	1	H	B_Os_1	Davis	6/9/2024
10	Sample_9	2	A	B_Os_1	Davis	6/9/2024
11	Sample_10	2	B	B_Os_1	Davis	6/9/2024
12	Sample_11	2	C	B_Os_1	Davis	6/9/2024
13	Sample_12	2	D	B_Os_1	Davis	6/9/2024
14	Sample_13	2	E	B_Os_1	Davis	6/9/2024
15	Sample_14	2	F	B_Os_1	Davis	6/9/2024
16	Sample_15	2	G	B_Os_1	Davis	6/9/2024
17	Sample_16	2	H	B_Os_1	Davis	6/9/2024

20. The app generates a CSV file with plate co-ordinates of every sample

- This file can be adapted modified to comply with the requirements of most genotyping service providers

Country Studies Data Documentation and Reproducibility Standard (Guide)

Table of Contents

1. Introduction and purpose

There is a great deal of responsibility associated with research that influences institutional decision-making, investments, or policy. Transparency, reproducibility, and replicability are foundational to SPIA's work. Policymakers and other research consumers who rely on SPIA's findings to inform decisions must be able to track, validate, and reproduce those findings independently. Achieving this requires that all analytical work be supported by clearly documented data sources, well-structured and annotated scripts, and results that are unambiguously reproducible by a third party. The practical foundation for regularly meeting that standard across all SPIA projects, teams, and outputs is provided in this guide. This guide helps build that research compendium well, so that producing the replication package at submission is straightforward.

This guide sets out the minimum requirements for data documentation and reproducibility that all country studies contributing to this programme must fulfil. It outlines the shared framework for data documentation and articulates why these standards are indispensable to the programme's scientific rigour. It also provides guidance on good practice across the full range of data types that country studies are likely to generate.

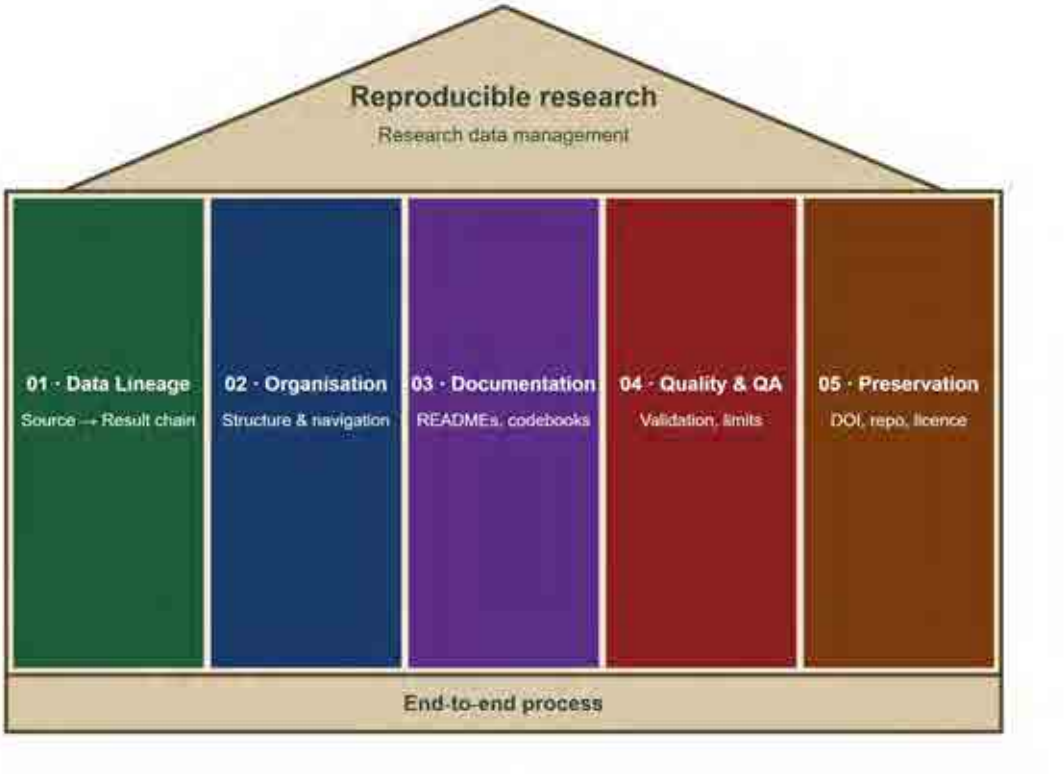
The programme is collectively building a body of evidence that spans countries, contexts, and years. The value of that evidence depends critically on how well any one team's data can be understood and verified by researchers who were not present when the data were collected. Replication packages reproducible only under specific conditions (for example, on a single operating system, or only with manual intervention or containing documentation gaps) make independent verification difficult. This guide addresses those specific points where replication can fail, as well as the broader documentation standard that SPIA expects of all submissions.

The guide is written with an awareness that teams operate in diverse national, institutional, and regulatory environments, and that many teams are simultaneously navigating their own Institutional Review Board (IRB) processes. Its aim is to provide a shared framework that helps teams align with one another and with prevailing standards in academic publishing, without imposing a parallel administrative layer on top of existing institutional requirements.

The guide has two interrelated purposes:

- Provenance: to ensure that the data generated including those underlying any published table or figure in a paper or report (hereon "exhibits") can be traced back to its original source through an unbroken, documented chain of transformation.
- Reproducibility: to ensure that all published exhibits can be independently reproduced by a third party executing the submitted code on the submitted data, on any standard computing platform.

2. Our Shared Approach



Country teams work within different institutional environments, use different software, and collect very different types of data. What is shared across the programme is a common framework; five areas that are relevant to every study. This framework draws on established data management principles, including the FAIR data principles (Findable, Accessible, Interoperable, Reusable), and the data sharing principles and practices.

The five pillars described below (and depicted in figure 1 above) define a minimum standard that all teams are expected to meet, irrespective of study design, data type, or publication venue. They are not presented as an exhaustive guide to research data management, but as the baseline requirements necessary for the programme's outputs to be credible, verifiable, and usable by future researchers. Each pillar is described briefly in Table 1 and in detail in the sections that follow.

The pillars are not independent. Good organisation makes documentation easier; thorough documentation enables meaningful quality validation; and all these pillars together make access and preservation actionable rather than aspirational. Teams that address all five pillars systematically will find that the work of assembling a replication package is substantially less burdensome than when it is approached retrospectively.

Table 1: The Five Pillars of Data Documentation

Pillar	Core Question	Description and rationale
1. Data Lineage and Provenance	*Where did the data come from?*	Traces data from its original source through every transformation to the final analytical dataset, with an unbroken, documented chain of transformations.

2. Organisation and Naming	*How is the package structured and named?*	Logical structure and consistent naming conventions across data, code, and documentation enable efficient navigation, collaboration, and future work; particularly in multi-year, multi-person projects and when data is shared across sites.
3. Documentation	*What does it mean?*	Makes variables, methods, and decisions interpretable to diverse audiences across time. Codebooks, data dictionaries, and READMEs are complete.
4. Quality & Validation	*How reliable is it?*	Quality control procedures, limitations, and validation checks are documented across both data and analytical processes. Establishes trust in data quality and enables appropriate interpretation of findings.
5. Access & Preservation	*Where will it live, who can access it, and under what condition for how long?*	Ensures data, code, and documentation remains findable and usable long-term, supporting future research and synthesis. Decisions about access, licencing, and preservations need to be made deliberately and early; especially where data is sensitive, jointly owned, or subject to legal restrictions.

3. Data lineage and provenance

3.1 Understanding data lineage

Data lineage is the documented journey of data, from its original source through every transformation to the final analytical dataset that underpins the reported results. The following core elements must be documented for all data types:

Table 2: Core Elements of Data Lineage Documentation

Original source	Organisation, platform, collection effort, or programme that produced the data. Include a URL or accession reference where one exists.
Acquisition method	How you obtained the data: direct collection, download (with date and URL), API call (with endpoint and parameters), data request or partnership agreement, purchase. Include any data use agreement reference.
Initial state	What processing had already been done before you received the data?
Transformations	Every cleaning, merging, recoding, filtering, and aggregation step you applied, in the order applied. These should be implemented in and traceable through numbered scripts.
Tools & Versions	Software, packages, libraries and version numbers used for processing and analysis steps.

Known limitations	Data quality issues, coverage gaps, reporting biases, or definitional changes relevant to interpretation. Stated per dataset, table, or figure, not as a blanket disclaimer.

3.2 Lineage by data type

3.2.1 Household and Farm Surveys

Survey data constitute the primary data type for most country studies. The key principle is that everything that affects how the data should be interpreted – sampling design, instruments, field decisions, cleaning choices – must be recorded.

Table 3: Documentation Requirements for Survey Data

Element	What to Document
Sampling design	Target population, sampling frame, sampling method, IRB/ethics clearance reference.
Survey design	Number of waves, timing, duration, mode of administration, language(s) used.
Data collection	Enumerator training procedures, supervision protocols, platform or tools used.
Quality assurance	Field validation procedures, back-checks, consistency checks.
Processing steps	Cleaning rules, outlier treatment, weighting procedures, any imputations.
Confidentiality	Whether individual, household, or institutional identities can be inferred from combinations of variables. Agreements governing publishable analyses should be documented.
Multi-wave studies	For panel data: the relationship to previous waves; a panel linkage file; comparability issues clearly explained; crosswalk files where variable definitions changed; attrition documented by wave with reasons.

3.2.2 Remote Sensing and Geospatial Data

This section covers the provenance documentation of both input and output data streams. Input data encompasses two distinct data streams that require separate provenance records: satellite and remote sensing data, and field-collected geospatial data, with field observations serving as training or validation inputs for satellite-derived products. Output data refers to the final product generated from these inputs – such as crop classification, AWD classification, yield estimation, or silvopastoral maps – which are intended for delivery and use by end users. Documentation applies equally to both data streams to ensure full traceability from raw input through to final products.

Reproducibility is arguably more fragile for remote sensing than for other data types: satellite products have different versions; cloud-cover thresholds applied interactively are invisible unless scripted; processing choices interact with scene selection in ways that cannot be reconstructed without explicit documentation. Where data were generated or exported from a cloud platform (e.g. Google Earth Engine) and no downloadable source file exists, the export script is the provenance record and must be archived in scripts/ and referenced by path.

Table 4: Documentation Requirements for Satellite and Remote Sensing Data

Element	What to Document
Source and product identity	Satellite or sensor name (Landsat, Sentinel, MODIS, etc.), specific product name and version number. For cloud platform exports (e.g. Google Earth Engine), no source file exists, archive the export script in scripts/ and record its path as the source record.
Acquisition parameters	Date range, cloud cover thresholds, spatial resolution, seasonal constraints, orbit/path selection, specific tiles and scene identifiers used.
Processing level received	What processing had been applied when you received it as delivered by the provider (e.g. Level 2 surface reflectance, atmospheric correction applied by the provider).
Processing steps	Preprocessing: Atmospheric correction method; cloud/shadow masking algorithm beyond provider delivery (with threshold value); band selection/spectral information, and index calculations (NDVI, EVI) with formula; temporal compositing method; spatial aggregation approach (include target resolution and resampling algorithm). Processing: Where classification/regression was applied: the algorithm, training sample source and total size, train/validation split, key hyperparameters, and accuracy metrics achieved. Post-processing: Accuracy assessment (including metrics used: confusion matrix, user's/producer's accuracy, F1-score); filtering or smoothing applied; spatial or temporal refinements; and any adjustments to the final outputs.
Coordinate Reference System (CRS)	CRS/projection used throughout; always report the EPSG code of the CRS in which outputs were saved.
Unique identifier	DOI (preferred) or UUID assigned to the dataset for citation, state the licence applied.
Output data (final product)	Format (GeoTIFF, CSV table, geojson, shapefile), spatial resolution, temporal resolution (e.g., monthly composite, annual) units (scaled -1 to 1), class definitions (if applicable), file structure. Include spatial metadata of the final products (ISO 19115 metadata standard preferred)
Limitations and uncertainty	Known limitations, sources of uncertainty, and assumptions affecting the data

Software/platform	Software/platform used (e.g., Google Earth Engine, Python, R libraries)
Data access (optional)	Storage location (repository, cloud bucket, path) and access conditions.

Field-collected geospatial data require a separate provenance record. These data are the spatial link between household or ecological observations and the landscape, collection conditions are unrepeatable, and the record must be specific enough. GPS coordinates recorded at household level must be de-identified before any public deposit.

Table 5: Documentation Requirements for Field-Collected Geospatial Data

Element	What to Document
Instrument	The GPS device make, model, and positional accuracy specification. For drone surveys: aircraft make and model, sensor type, and flight altitude.
Sampling design	The sampling frame, stratification approach, and randomisation procedure followed. Reference the design document where possible
Collection quality assurance	The positional accuracy threshold and repeat-measurement protocol applied in the field.
Coordinate reference system and transformations	The EPSG code of the CRS in which data were archived. Where reprojection was performed, state the source and target CRS by EPSG code.
Integration with RS products	Where field points were used as training or validation samples, or linked to raster products for analysis: the extraction method (pixel value, window size, statistic), buffer radius if used, and a reference to the RS product by path or DOI. The RS lineage record must also cross-reference this field dataset by the same identifier.
Unique identifier and access	A DOI (preferred) or UUID. Note whether household GPS coordinates were de-identified before public deposit, in accordance with ethics approvals.
Variable description	Variables collected, units, and coding scheme (e.g., class labels, survey responses).

3.2.3 DNA Fingerprinting / Genotyping Data

"Genotyping" covers a range of experimental and bioinformatic design choices that all have consequence for the data output and interpretation. All quality control filters must be implemented in reproducible scripts. The following fields are the technical metadata required for any genotyping dataset:

Table 6:Documentation Requirements for Genotyping Data

Element	What to Document
Platform and sequencing	Instrument model and manufacturer; sequencing chemistry or array platform; marker types; total marker count before QC; genome coverage.
Sample and extraction	Tissue type; DNA extraction protocol (DOI or internal SOP filename); sample storage conditions.
Reference genome and pipeline	Assembly name, version, and source; alignment software and version; variant calling software and version; imputation details if applied.
QC thresholds	Call rate, minor allele frequency (MAF), sample missingness, SNP/marker missingness, and any additional filters applied.
QC summary statistics	Samples and markers before and after QC; call rate and MAF distributions; exclusion counts per filter step.
File formats	Provide genotype data and metadata as received from the genotyping service provider in csv format.

Genotyping of plant material sits at the intersection of research ethics and binding international law. Two treaties are directly relevant to country teams collecting or transferring plant genetic material:

1. International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA)
2. Nagoya Protocol; Supplementary Agreement to the Convention on Biological Diversity

What this means for data use:

- ✓ Any plant material collected in, or transferred across the borders of, a signatory country is subject to one or both frameworks
- ✓ The legal basis for collection, transfer, and use must be documented before data are collected/generated
- ✓ Data derived from material obtained without the required agreements (e.g., SMTA, MTA, PIC, MAT) may be unpublishable
- ✓ If obligations are unclear for a specific collection context, teams should seek guidance from their institution's legal or technology-transfer office.

These fields in Table 7 below serve two purposes: they satisfy the documentation requirements of the ITPGRFA and Nagoya Protocol, and they make the material traceable and citable for future research. Every field should be completed at the point of collection/accession; gaps that emerge at the submission stage are difficult to fill retrospectively.

For plant genetic material, the following provenance fields must be recorded for each crop or species genotyped:

Table 7: Provenance Fields Required for Plant Genetic Material

Provenance Field	What to Record
Crop / species	Full scientific name (genus, species, authority). Record common name alongside.
Annex I status	Is this species listed in Annex I status of the ITPGRFA? (See fao.org/plant-treaty.)
Collection country	Country name and ISO 3166-1 alpha-3 code.
Collection location	GPS coordinates (decimal degrees) or administrative zone (region, district, kebele/ward equivalent). Both preferred.
Source of material	Genebank accession, farmer field, on-farm landrace, market sample, breeding station, or other. If genebank: record accession number and WIEWS code.
Material holder / provider	Full name and institutional affiliation. For farmer-managed material: community name and the position of the consenting individual
Collection date	ISO 8601 format (YYYY-MM-DD or YYYY-MM).
GLIS DOI	Any Global Information System (GLIS) Digital Object Identifier assigned to this accession. If none exists, record whether DOI assignment was requested.
Laboratory processing	Full name, country, and institutional type (national research institution, university, private company, international centre) of the laboratory conducting extraction and/or genotyping.
Cross-border transfer	Yes or No. If Yes: destination country, receiving institution, and reference to transfer agreement.
Data repository deposit	Repository name; accession number or DOI. If restricted: access conditions stated.

The structure below extends the folder's recommended layout for genotyping-specific documents:

```
documentation/
├── data_use_agreements/
│   ├── SMTA_[crop]_[institution]_[date].pdf      ← File if ITPGRFA materials obtained through multilateral system
│   ├── MTA_[country]_[institution]_[date].pdf    ← Non-Annex I species/ bilateral transfers outside the multilateral system
│   ├── MAT_[country]_[date].pdf      ← Accompanies PIC; specifies permitted uses and benefit-sharing conditions
│   └── PIC_[country]_[date].pdf      ← Prior Informed Consent if applicable (any Nagoya Protocol signatory country)
├── ethics/
```

```
|   ├── IRB_approval_[institution]_[date].pdf
|   ├── biosafety_approval_[country]_[date].pdf    ← Where required by national law
|   ├── export_permit_[country]_[date].pdf   ← Phytosanitary or ABS export permit (before cross-border transfer)
|   └── consent_forms/
|       ├── participant_consent_[country]_[language].pdf
|       └── community_consent_[country]_[date].pdf    ← For ABS material
└── compliance_log.xlsx                    ← Completed at programme level; referenced here
```

3.2.4 Qualitative Evidence

For interviews, focus groups, and observational data:

- Data collection method: Interview format, observation protocols, focus group structure, recording method used
- Participant selection: Sampling approach, recruitment method, selection criteria, sample size
- Analysis approach: Thematic analysis, grounded theory, narrative analysis, or other analytical method
- Anonymisation: Procedures for protecting participant identity in transcripts and quotes
- Triangulation: How findings from qualitative data were checked against other sources or methods. How contradictions or inconsistencies across sources were handled and resolved.
- Software used: List all software employed for data management, coding, or analysis. Include version numbers if relevant.

3.2.5 Process Tracing Data

For process tracing data analyses, the following elements should be documented:

Table 8: Provenance fields required for process tracing data

Element	What to Document
Evidence sources	Documents reviewed, stakeholders interviewed, events observed, archival records accessed — each logged with date, source type, and retrieval method.
Case selection	Logic for case selection, justification for its relevance,
Temporal sequencing	Event log documenting the order and timing of conditions, decisions, and outcomes relative to the hypothesized causal mechanism/chain.
Causal mechanism documentation	Hypothesized mechanism/pathway, expected evidence patterns, tests applied, rival explanations documented including the basis on which the primary account is preferred.
Evidence classification	How evidence was categorized, including confidence ratings, source reliability assessments, and potential bias of each piece of evidence.
Triangulation	How multiple evidence sources were integrated and cross-validated.
Iteration and updating	Whether and how prior evidence shaped subsequent evidence-seeking, including any revision to the hypothesis or evidence strategy mid-analysis.

Element	What to Document
Sufficiency of evidence	Statement of when evidence collection was considered adequate to support a defensible inferential claim, and the basis for that judgment.
Chain of custody	Consent basis, anonymisation decisions, handling of interview notes or recordings, and respondent review arrangements for qualitative data collected.

3.2.6 Administrative and External Data

For government records, programme data, or third-party datasets:

- Source agency: Government ministry, NGO, private provider, or another organisation
- Access method: How data was obtained (official data request, API with endpoint and date, partnership agreement, purchase, public portal download)
- Coverage: Time period, geographic scope, population covered, completeness assessment
- Known limitations: Data quality issues, missing records, reporting biases, definitional changes over time
- Use restrictions: Data sharing agreements (clearly regulate data publication), licensing constraints, citation requirements and any restrictions on what analyses or results may be published
- Integration procedures: How you matched or merged with primary data sources

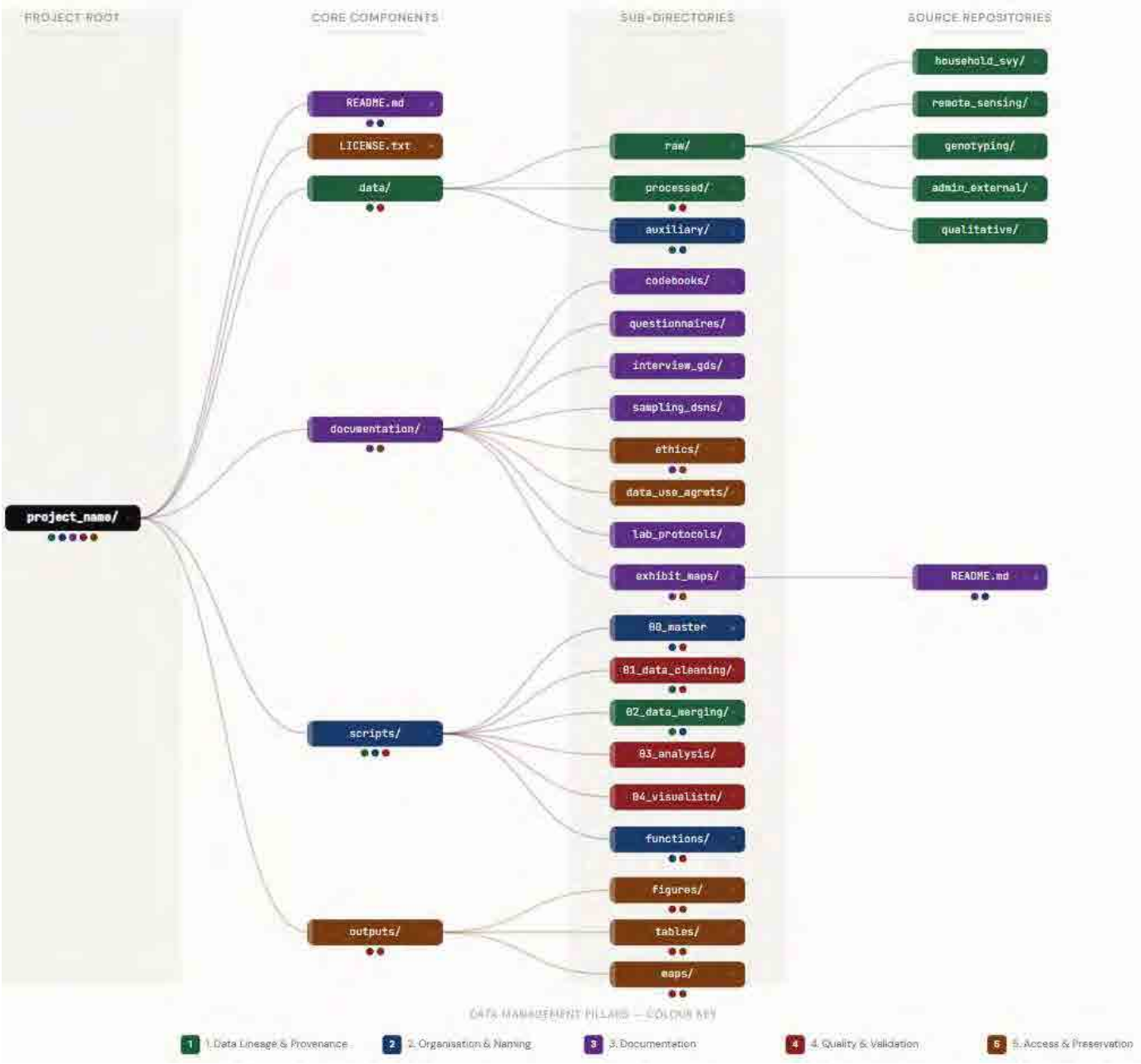
Having established the requirements for data lineage and provenance across all data types, the next pillar concerns how data, code, and documentation files ought to be organised and named for easy navigation.

4. Organisation and Naming

4.1 Folder structure

The structure below is a reference – a recommended structure – not a mandatory template. Adapt it to your study design, and document any departures in the README. What matters most is that the structure is logical, consistent, and self-explanatory. Where possible, scripts, documentation, and outputs may further be split by thematic area for easy navigation.

```
project_name/
├── README.md   ← Start here: overview, navigation, reproduction instructions, system requirements, data description, exhibit map
├── LICENSE.txt   ← Data and code licensing terms
├── data/          ← Provide the data in file formats that can be used independently of statistical package choice
│   ├── raw/   ← Original source files
│   │   ├── household_survey/   ← De-identified raw survey files by wave.
│   │   ├── remote_sensing/
│   │   ├── genotyping/
│   │   ├── admin_external/
│   │   └── qualitative/
│   ├── processed/   ← Project-generated. Produced by cleaning scripts.
│   └── auxiliary/
├── documentation/
│   ├── codebooks/
│   ├── questionnaires/   ← As used in the field, all waves.
│   ├── interview_guides/
│   ├── sampling_designs/
│   ├── ethics/   ← IRB approvals, consent forms.
│   ├── data_use_agreements/
│   └── lab_protocols/   ← For genotyping data
├── scripts/
│   ├── 00_master.do / 00_master.R / 00_master.py   ← Runs all sub-scripts in order
│   ├── 01_data_cleaning/
│   ├── 02_data_merging/
│   ├── 03_analysis/
│   ├── 04_visualization/
│   └── functions/   ← Reusable helper functions
└── outputs/
    ├── figures/
    ├── tables/
    └── maps/
```


PROJECT ROOT
CORE COMPONENTS
SUB-DIRECTORIES
SOURCE REPOSITORIES
project_name/
README.md
LICENSE.txt
data/
documentation/
scripts/
outputs/
raw/
processed/
auxiliary/
codebooks/
questionnaires/
interview_gds/
sampling_dsns/
ethics/
data_use_agrmts/
lab_protocols/
exhibit_maps/
00_master
01_data_cleaning/
02_data_merging/
03_analysis/
04_visualistn/
functions/
figures/
tables/
maps/
household_svy/
remote_sensing/
genotyping/
admin_external/
qualitative/
README.md
DATA MANAGEMENT PILLARS — COLOUR KEY
1 1. Data Lineage & Provenance
2 2. Organisation & Naming
3 3. Documentation
4 4. Quality & Validation
5 5. Access & Preservation

4.2 File naming

Apply the following conventions consistently across all files and folders:

- Use lowercase throughout
- Separate words with underscores; avoid spaces and special characters
- Use ISO 8601 date format (YYYYMMDD) in date-stamped files.Use a version suffix (v01, v02) for iterative outputs.

Sound organisation provides the framework upon which meaningful documentation can be built. The following section sets out what that documentation must contain.

5. Documentation

Documentation transforms data from raw files into interpretable research assets. Good documentation answers three questions: what is it? (content), why was it done this way? (decisions), and how can it be reproduced? (process).

Table 9: Required Documentation Elements

Element	What to Include	Priority
README	Project overview; data sources summary; folder navigation guide; reproduction instructions; system and computational requirements; list of all exhibits (tables and figures) in the paper; licence and citation information; data availability statement	Essential
Data dictionaries	Variable definitions, units, value ranges, missing data codes and any other relevant notes.	Essential
Codebooks	Coding schemes, variable labels, skip patterns for survey and administrative data. Weighting formulas. Crosswalk tables for multi-wave or multi-country variables.	Required
Exhibit map	Maps every table and figure in the submitted paper to the specific script(s) and output file(s) that produce it, including the appendix exhibits. Every exhibit must appear; every output file must correspond to an exhibit.	Required
Data collection tools	Questionnaires, interview guides, field sample collection protocol, observation protocols as used in the field.	Required
Decision log	Record of analytical choices made during cleaning and analysis, with justification. Especially important for choices that could reasonably have been made differently.	Recommended
Metadata files	Spatial, temporal, technical metadata for all data types (can reference applicable metadata standard).	Required

Ethical documentation	IRB approvals, consent procedures, data protection measures, data availability statement.	If applicable (Essential)
Environment documentation	Operating system (Name, version, and architecture), Software version, Package list with version numbers, estimated pipeline runtime.	Required
Scripts & Code	All cleaning, processing, analysis, and visualisation code with clear comments. A master script that installs packages, sets the random seed, defines folder paths, and calls every sub-script in sequence.	Required

5.1 The README File

The README is the landing page for any compendium. A researcher encountering your package for the first time should be able to answer the following questions from the README alone, without opening any other file:

Table 10: Required Content of the README File

Study/data description	What is contained in the study package, data description, what research paper it addresses, and the purpose. Two to three sentences.
Data availability statement	For each dataset: where it comes from, whether it is included in the package, and if not included, how a reviewer can access it and under what conditions.
How to run the code files	The single instruction a reviewer needs to reproduce all results (typically: change one path in 00_master and run it). Software and version requirements. Estimated runtime and computational resources required.
Folder and file guide	A brief description of what is in each folder and the purpose of each key file. The structure should be navigable from this section alone.
Exhibit map	A complete list of every table and figure in the paper and appendix, each linked to the script and output file that produces it. This can also be provided as part of the documentation if not included in the readme.
Known issues	Any exhibit that cannot be reproduced from the provided data, any script that requires manual intervention, and any result that is sensitive to software version or hardware.
Licence and citation	How the package should be cited, the DOI, and the terms under which the data and code may be reused.

The README should be in an open, platform-independent format (ASCII text, Markdown, or PDF).

5.2 Exhibit map

The exhibit map is one of the most useful documents a team can produce. It is a simple table linking every table and figure in the published report to the script and output file that produced it. Building this map before submission saves significant time and reveals gaps while they can still be addressed.

Table 11: Exhibit Map: Required Fields and Example

Exhibit in Paper	Output/Intermediary file	Producing Script	Notes

Table 1 (Baseline Characteristics)	outputs/tables/table_01.csv	scripts/03_analysis/03a_descriptives.do	Weighted using survey_weights.dta; weights constructed in 02_merging/02c_weights.do

5.3 Scripts and Code documentation

This is where reproducibility is made or broken. Every transformation applied to the data must be scripted, numbered, and explained.

Table 12: Script and Code Documentation Requirements

Element	Action at this stage
Numbered script sequence	Scripts must be numbered to reflect execution order (01_, 02_, etc.). A master script must call every sub-script in sequence from raw data to final outputs.
Script headers	Every script must include: author name, date last edited, purpose of the script, inputs (list all input data required upfront), and outputs.
No manual steps	All cleaning, recoding, outlier treatment, and QC decisions must be implemented in script. No manual edits to data files. No copy-paste into Excel. No interactive adjustments to charts.
Set random seeds	Every script that involves randomisation (sampling, simulation, imputation) must set the random seed explicitly at the top of the script.
Decision log	Every non-obvious analytical choice — an outlier threshold, a variable construction decision, a sample restriction — must be recorded in the decision log with a brief justification.
Intermediate outputs	Intermediate datasets should be saved at key stages of the pipeline to facilitate debugging and partial re-runs.
Output naming	Every output file must be named to match the exhibit it corresponds to in the paper (e.g., table_1_main.csv, fig_3_yield_density.png).

5.4 Ethical and Regulatory Documentation

Country teams are subject to different national legal and regulatory frameworks, and many are simultaneously navigating their own institutional IRB processes. This guide does not prescribe the IRB process itself, which is governed by teams' respective institutions and national laws. What it does require is that the outcomes of those processes are documented and included in the submission package.

At a minimum, each submission must include or reference the following, as applicable:

- IRB or ethics committee approval, including the reference number, issuing institution, and date.
- Participant consent forms as administered, in all languages used, for all waves of data collection.

- Any community-level consent documentation for studies involving traditional knowledge or community resources.
- Data use agreements (DUAs) or memoranda of understanding with data-providing institutions, with clear notation of any restrictions on publication or secondary use.
- Biosafety approvals and export permits, where required by national law (particularly relevant for genotyping studies).
- A data availability statement specifying the access conditions applicable to each dataset in the package.

Where ethics documents cannot be made publicly available due to confidentiality requirements, teams must provide metadata describing what documentation exists and whom to contact for access, and must note this in the README.

Thorough documentation of methods and decisions sets the foundation for the quality and validation procedures described in the following section.

6. Quality & Validation

Quality documentation is not a separate activity from cleaning and analysis; it is the record of how those activities were carried out. A brief data quality report submitted alongside the data package, tells a reviewer or a future user exactly what they are working with before they open a single file.

Table 13: Quality and Validation Documentation Requirements

Element	What to Document
QC Procedures	Validation checks applied during and after collection; range checks, consistency checks, duplicate detection. These should be implemented in reproducible scripts.
Outlier Treatment	Rules applied; thresholds used; how decisions were made (automated or manual); documented in the cleaning script with the rationale for any case-by-case decisions.
Missing Data	Proportion missing per variable; missing data mechanism assessed where possible; imputation approach if applied, with method and assumptions stated.
Limitations	Known quality issues, coverage gaps, reporting biases, or definitional changes; stated per data source.
Data Quality Report	Summary of QC outcomes: n records processed, n records excluded, flags raised, versions produced.
Version Control	Scripts versioned (Git strongly recommended); random seeds set and recorded; software versions documented.

With quality control procedures established and documented, the final pillar addresses how data are preserved and made accessible for future use.

7. Access & Preservation

Research data should be as open as possible and as closed as necessary, in accordance with FAIR data principles. Decisions about access, licensing, and long-term storage need to be made deliberately and early; they are much harder to revisit once a project has concluded, particularly where consent language, legal agreements, or sensitive data are involved.

These practices are consistent with the requirements of the major development economics and agricultural research journals' data-sharing mandates. If your institution or a target journal has its own reproducibility or data sharing requirements, those take precedence; the practices described here align with them.

Table 14: Access and Preservation Documentation Requirements

Element	What to Document
Storage Location	Repository, archive, or platform where data is deposited.
Access Rights	Who can access, under what conditions, and the specific licensing terms.
Unique Identifiers	DOI, UUID, or dataset ID for citation and discovery.
Retention Policy	How long data is kept, by whom, and the archival procedures.
Ethical Documentation	IRB approvals, consent forms, data use agreements archived alongside the data.

7.1 Sensitive and restricted data

Where data cannot be made publicly available, because of participant consent conditions, data use agreement terms, national data protection law, or the sensitivity of the subject matter, the following approach applies:

- Provide stub files: placeholder dataset files containing correct variable names, labels, and formats but no observation-level data. A reviewer can thereby confirm that the code would run on the correct data structure.
- Document the access procedure in the README: who holds the data, under what conditions access may be granted, and the contact point for access requests.
- Consider depositing in a restricted-access archive rather than withholding entirely.
- Synthetic data— data generated to preserve the distributional properties of the original dataset whilst containing no real observations — is increasingly accepted as a complement to restricted access. Document the method used to generate it.

- Metadata must be publicly accessible even where data is restricted or embargoed. A reviewer should be able to understand what the dataset is, how it was collected, and how to request access, without needing to access the data itself.

The five pillars described above converge in the replication package —the practical vehicle through which all requirements are fulfilled and submitted for review.

8. The replication package

This section does not repeat what earlier sections have covered, it assumes that the data are clean, the code is organised, the provenance chain is documented, and access decisions have been made. It addresses how to assemble those components into a coherent package drawn from different data types, analytical workflows, and compliance environments. A replication package is the complete, self-contained digital object that accompanies a submitted report or paper. It contains everything a reviewer needs to independently verify your results: the data, the code, the outputs, and the supporting documentation. The package is deposited in a trusted repository and assigned a persistent identifier.

Three organising principles should guide its construction:

- Data, methods, documentation, and outputs are clearly separated.
- The folder structure is self-explaining: an independent researcher should be able to navigate it using only the README.
- The computational environment is specified: a reviewer should be able to recreate the software conditions under which the analysis was run.

8.1 Replication Package components

Component	What to include	Cross-reference
Data	All datasets needed to reproduce reported exhibits. Stub files and access instructions for restricted data. Compliance documentation.	Data lineage and provenance; Quality & Validation; Access & Preservation
Code	All scripts from raw data to final outputs, numbered sequentially, called from a single master script.	Organisation and Naming; Scripts and Code documentation; The master script
Outputs	Every final output file, named to match its exhibit. No undocumented exhibits.	Exhibit map; File naming.
README	The entry point for the package. Orients a reviewer to the structure, data sources, software environment, and steps to reproduce results. Cross-references rather than duplicates other documentation.	The README File
Supporting documentation	Original data collection instruments (all languages, all waves); ethics clearance and consent forms or references to them; data	Ethical and Regulatory Documentation

Component	What to include	Cross-reference
	use agreements or references; a copy of or link to the associated paper.	

8.2 Computational Reproducibility

Computational reproducibility means that an independent researcher, given the data and scripts in the package, can execute the full pipeline and arrive at results numerically consistent with those reported in the paper. This is about recording the computation itself: the software, the execution, and the numerical outputs.

8.2.1 The master script

The master script is the single-entry point that calls every other script in the correct order. A reviewer should be able to:

- Change one global path (the top-level project directory) at the top of the master script.
- Run the master script from start to finish (include expected runtime).
- Find that all outputs in the outputs folder that match the exhibits in the submitted paper.

The master script should: document the software version and set random seed; install all required packages (or reference an environment file); define all folder-path globals; and call every sub-script in sequence from raw data to final outputs, with no manual steps in between.

If a master script is not provided, at minimum, provide an explicit numbered list of scripts so that the order is unambiguous, including inputs and outputs of each.

8.2.2 Random seeds

Any analysis involving stochastic processes— randomisation, bootstrapping, cross-validation, simulation— requires a random seed. Set and record all seeds explicitly at the top of the script and confirm that results are stable across repeated runs with the same seed. If results vary between runs with the same seed, investigate and resolve before submission.

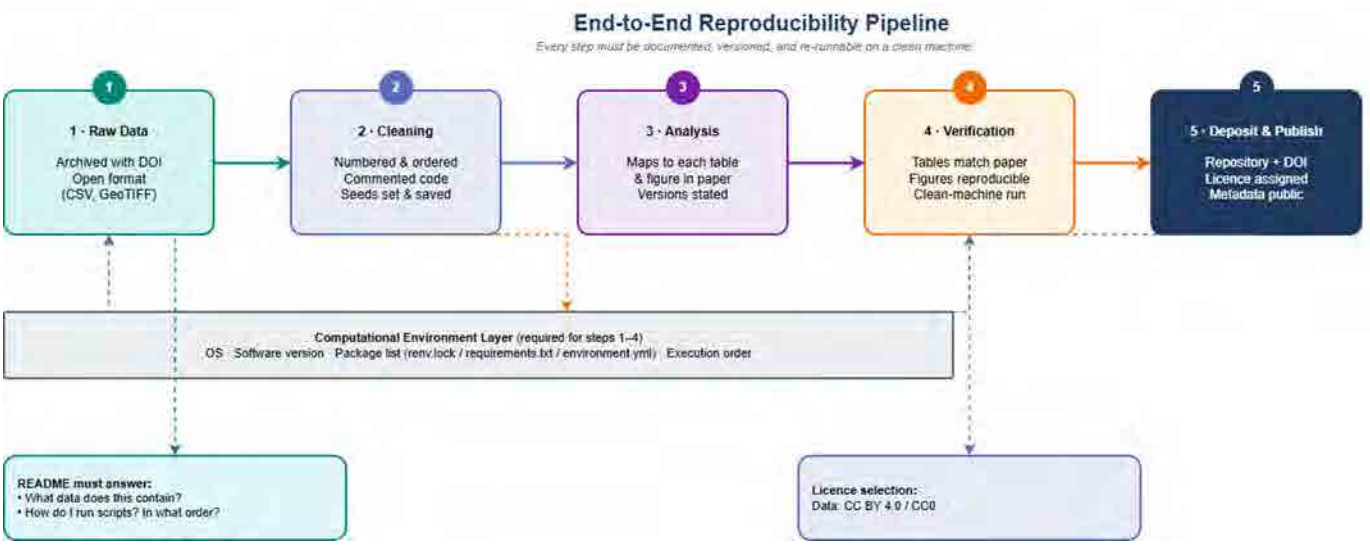
8.2.3 Capturing the computational environment

Document the software environment upon which your code depends: your operating system, the exact programming language version, all supporting packages and libraries with version numbers, and —where relevant — hardware specifications.

The following table describes the desired levels of environment capture,

Level	What to Capture
Minimum required	Operating system (name, version, architecture); software language and exact version number; complete package list with version numbers.
Recommended	All of the above, plus a machine-readable environment file that enables recreation of the software environment.

Teams should, where possible, test the full pipeline on at least one operating system other than the one used for development before submission.



The checklist that follows consolidates all requirements from the preceding sections into a practical pre-submission tool for country teams.

9. Pre-Submission Checklist

Use this checklist before finalising your submission: It maps the five pillars to concrete deliverables and is designed to be completed alongside your data package.

✓	Requirement
DATA LINEAGE AND PROVENANCE	
☐	A lineage record exists for every dataset: original source, acquisition method, initial state, transformations applied, and tools and versions used.
☐	Known limitations are documented per dataset, table, or figure
☐	For surveys: sampling design, instruments as administered, field quality assurance procedures, and enumerator identification scheme documented.
☐	Open file formats used for archival (CSV, GeoTIFF, plain text).
☐	For remote sensing: satellite product name and version, EPSG code, spatial and temporal resolution, full processing pipeline documented, data selection criteria, output data characteristics (accuracy assessment)
☐	For genotyping: platform metadata complete; all quality control filters scripted; material provenance records complete; GLIS DOIs checked and recorded.
☐	For qualitative data: instruments, participant selection, analytical approach, anonymisation procedures, and audit trail documented.
☐	For administrative data: source agency, access method, coverage, use restrictions, and data use agreement filed.
☐	Ethics and consent documentation included or referenced for all data types. For genetic material: SMTA, MTA, PIC, and MAT documents filed and referenced where applicable
ORGANISATION AND NAMING	
☐	Folder structure is logical, self-explanatory, and documented in README.
☐	File naming is consistent: lowercase, underscores, ISO date format where used.
☐	Exhibit map is complete: every paper exhibit linked to a script and output file.
☐	No output file exists without a corresponding paper exhibit or documented intermediate purpose.
DOCUMENTATION	
☐	README answers: what is this, how do I run it, in what order, what are the exhibits.
☐	README includes a data availability statement.
☐	Data dictionary or codebook provided for all key datasets.
☐	Data collection instruments included for all waves and all languages.
☐	Processing scripts are clearly commented; rationale for non-obvious decisions documented.
☐	Decision log provided for key analytical choices made during cleaning or analysis.

☐	Computational environment documented: OS, software version, package list, environment file.
☐	Estimated pipeline runtime stated.
☐	Known issues and limitations documented.
QUALITY AND VALIDATION	
☐	QC procedures documented; validation checks implemented in scripts.
☐	Outlier treatment rules documented with thresholds and rationale.
☐	Missing data proportions reported per variable; imputation approach stated.
☐	Known limitations documented per data source (not a blanket disclaimer).
☐	Data quality report produced and included in package.
☐	Random seeds set and recorded; software versions documented.
☐	For public-use files derived from sensitive data: statistical disclosure control (SDC) procedures applied, documented, and scripted.
☐	All key results reproducible by running provided scripts on provided data.
☐	Workflow clear: what runs in what order from raw data to final outputs.
ACCESS AND PRESERVATION	
☐	Metadata publicly accessible even if data itself is restricted or embargoed.
☐	Data licence assigned and stated in README and repository metadata.
☐	Access conditions and embargo end dates (if applicable) stated per data stream.
☐	For data that cannot be shared: stub files provided, access procedure documented, or synthetic data included.
☐	Ethical documentation archived: ethics clearance, consent forms, data protection measures.
☐	Applicable data protection law identified for each country of data collection.
☐	For cross-border transfers: legal basis documented (adequacy decision, SCCs, or equivalent).
☐	For genetic material: SMTA, MTA, PIC, MAT documents filed and referenced; GLIS DOIs recorded.
☐	Retention policy stated: how long data is kept, by whom, and archival procedure.
MULTI-WAVE AND PANEL STUDIES (where applicable)	
☐	Relationship to previous waves documented; panel linkage file included.
☐	Comparability issues clearly explained: instrument changes, sample frame revisions.
☐	Crosswalk files provided where variable definitions changed across waves.
☐	Attrition documented by wave with reasons.

This guide is a living document. We welcome insights, questions, and suggestions for improvement based on diverse experiences conducting country studies.