Photo: G. Smith/The Alliance

CGIAR | Standing Panel on Impact Assessment

Guidance on Reference Compilation, Field Sampling & Survey Integration

James Stevenson,
SPIA Sr. Research Fellow

Frederic Kosmowski,
SPIA Sr. Scientist

Davis Gimode,
SPIA Bioinformatics Research Specialist

15th October 2025

Outline

- Introduction: **James**
  - One size does not fit all
  - Behavior, biology, context
  - Introducing our guidebook

- Technical details of guidebook: **Davis**
  - Re-cap of previous webinar
  - Understanding what a modern variety is
  - A good reference library
  - Field sampling considerations

- Linking samples to surveys: **Frederic**
  - Implications for questionnaire design
  - Sample tracking – barcodes and tracking files

DNA fingerprinting guidance

CGIAR Standing Panel on Impact Assessment

- One size does not fit all
- Our Goal: To understand the contribution that the CGIAR is currently making to the production of a specific crop in a particular country
- The **crop** and **country context** matters for how we should sample
- We should be mindful of two sources of measurement error
  - **False positives**: don't want to falsely ascribe "adoption" of a variety
  - **False negatives**: don't want to miss adoption

Behavior

- Do farmers plant varieties for this crop?

Kurt Waldman et al 2025 Environ. Res. Lett. 20 081002

"The distinction between understanding discrepancies as measurement error versus a conceptual mismatch is fundamental to how we conduct research with smallholder farmers. Skepticism of farmer self-reported data naturally leads to a preference for more objective measures."

- What can we learn prior to launching fieldwork that can inform us on how farmers make decisions about planting material?

- A "variety" is our concept. Farmers may seek out and plant specific varieties. But they may just plant what they can get hold of.

Biology

- Plants of different species reproduce in different ways
- The biology of each crop significantly shapes the correct way to take samples

Context

- Who are the players in the seed system for the crop?
- Private sector seed companies: how well-regulated is the sector?
- Research centers: how active is the NARS in pushing out varieties for which there is no functioning private sector?
- Farmer groups / NGOs: where do they go to get the planting material they distribute?

→ Anticipating the expected genetic make-up of the population of the plants in farmers' plots

Guidebook

- “Soft launch” with you all today
- You have until November 15th to provide feedback! PLEASE!
- We’ll make changes
- December: External peer review
- December: Infographics and comms
- Publication in collaboration with the World Bank / FAO / IFAD “50 x 2030” Initiative in 2026

Brief re-cap of previous webinar

Standing Panel on Impact Assessment

- The four main steps of DNA fingerprinting for adoption tracking
  - Compiling a reference library
  - Collecting samples from the field
  - Genotyping samples and references
  - Analysis: Assigning variety Ids to samples

Understanding varieties

What is a variety?

Landrace/Wild

- Yield
- A/biotic stress tolerance

Modern variety

- Yield
- A/biotic stress tolerance

Field evaluation

Desired cross

Choose best plants

Repeat process

Understanding varieties

What is a variety?

Old variety

Improved variety

Compiling the varieties: start point

Seed classification after variety development

Breeder Seed

- Seed produced under direct selection of a breeder
- Purest representation of the variety

Foundation Seed

- Directly descends from breeder seed
- Production conditions guarantee genetic purity

Certified Seed

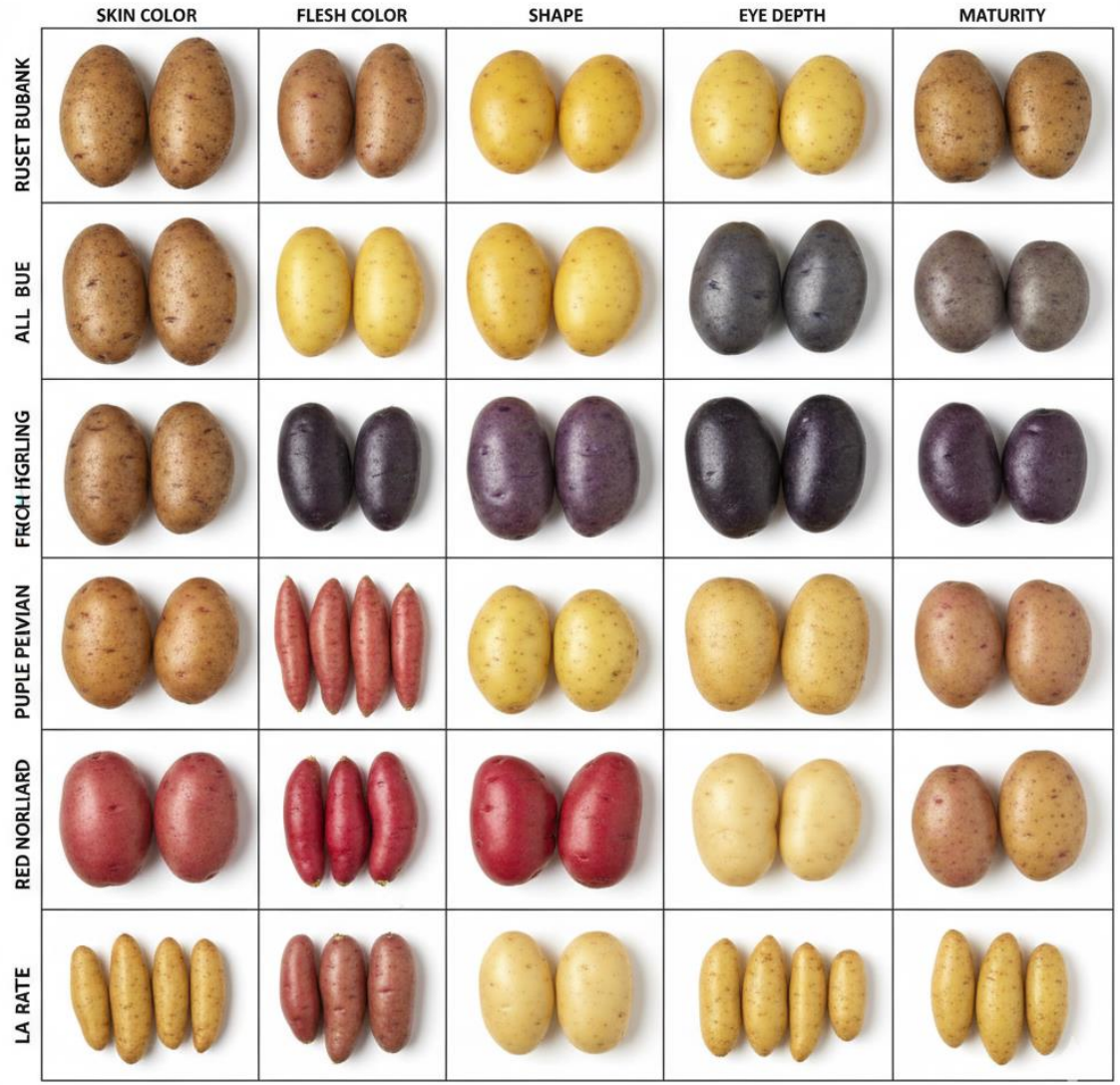
- Descends from foundation seed
- Produced under supervision of a certification agency to guarantee purity

Breeder seed is the best source for compiling the genetic reference library

Compiling the varieties

Attributes of a good reference library

- The reference library should be:
  - Pure
    - Varieties should not be contaminated through outcrossing or mixing with other varieties
  - Exhaustive
    - The level of completeness is determined by the purpose of the study
  - Distinct
    - Individual varieties should be sufficiently differentiated from each other



Verifying purity of references

CGIAR Standing Panel on Impact Assessment

Necessary because:

- Unintended crossing can contaminate a genetic line
- Accidental seed mixtures can occur

- **Procedure:** Plant 3-5 seeds from the reference sample, grow them, and collect leaf tissue from each plant individually for genotyping
- **Advantage:** This is the most detailed and reliable method. It allows for a precise assessment of purity and helps identify any deviations in the breeder's seed.
- **Consideration:** While more costly, this method is crucial because impurities found in breeder seed can be passed down through the entire seed system.

Exhaustiveness

Guided by research question

Partial reference set

- Suitable if the goal is to identify a specific subset of improved varieties
  - Recent releases
  - Those with particular traits
- Less resource-intensive

Comprehensive reference set

All grown varieties

- Essential if the goal is to identify the full range of varieties farmers are growing
- Any variety not in the library cannot be identified.

Exhaustiveness

Should you include landraces?

- **Potential benefit:** Allows for full positive identification when field samples don't match known improved varieties – resolving ambiguities
- **Considerations & Cautions:**
  - **Lack of control:** Less control over genetic purity and maintenance compared to improved varieties
  - **Genetic indistinguishability:** Landraces may be genetically indistinguishable from released varieties, leading to confounding results
  - **Redundancy:** Large collections of un-curated landraces can lead to significant redundancy and increased analytical effort
  - **Careful curation:** Only include landraces whose identity has been carefully maintained and verified
- **Recommendation:** Proceed with caution and verification when including landraces.

Distinctiveness

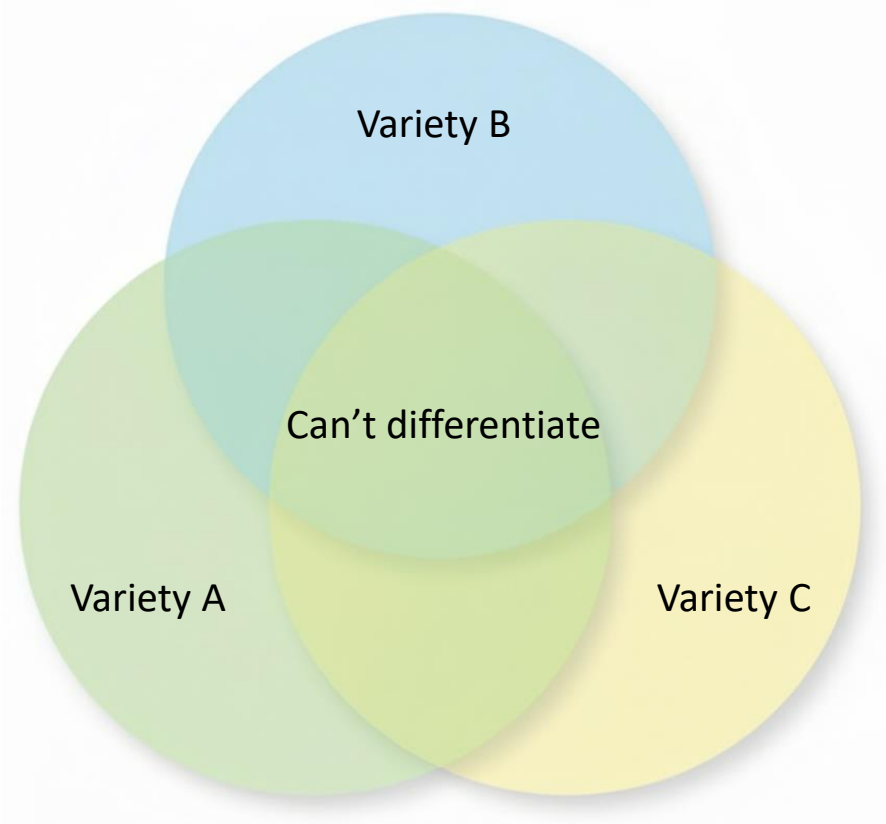
We need sufficient genetic distance between each sample

Causes of low distinctiveness may include

- **Shared pedigree:** Varieties may be closely related (e.g., a variety and its improved version)
- **Sample impurity:** Contamination or mislabeling of the reference seed
- **Low assay density:** The genotyping method doesn't use enough markers (SNPs) to find the subtle differences between varieties

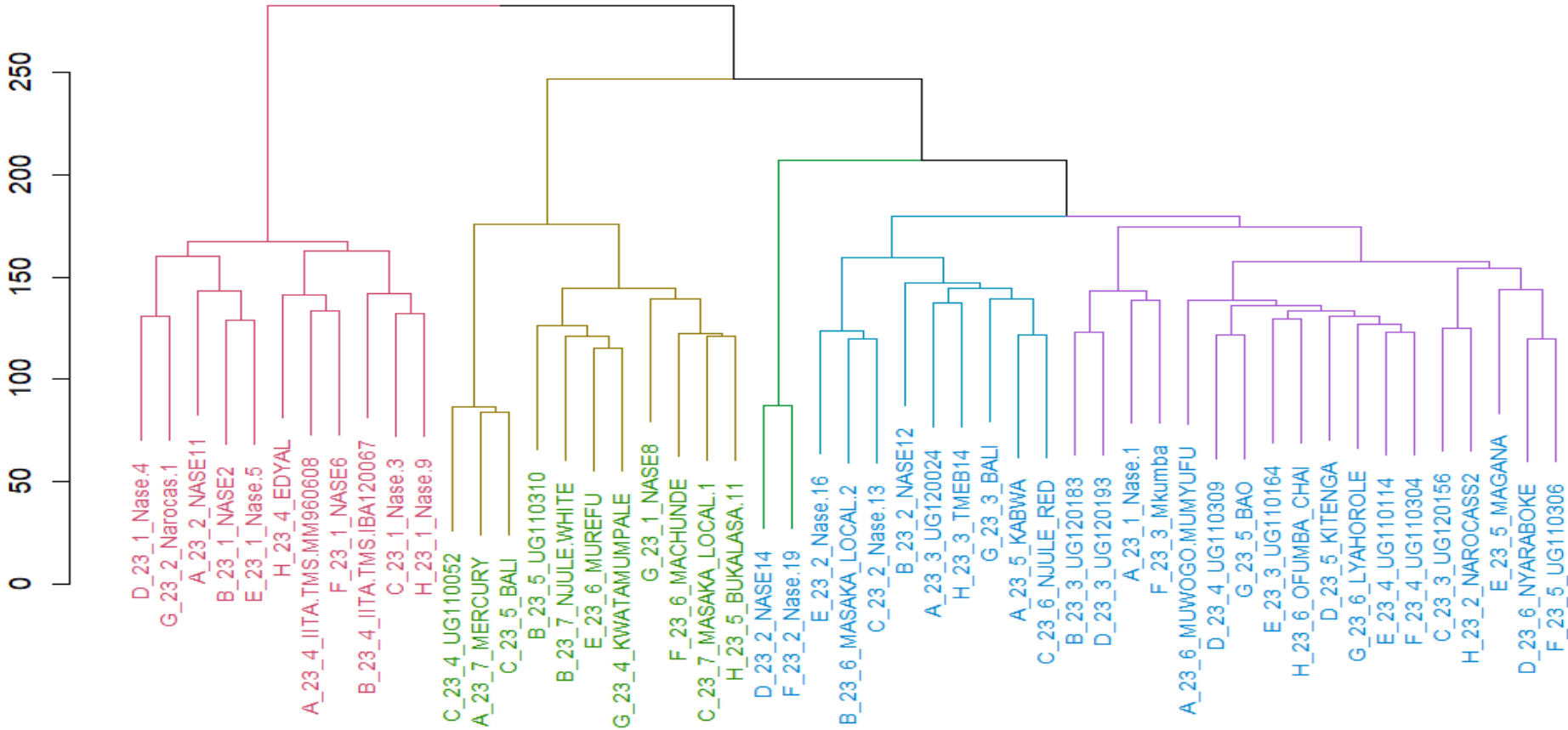
Recommendation: Genotype before survey

- Genotype the entire reference library before collecting field samples
- Analyze the data to create a genetic clustering dendrogram to visualize the distance between varieties
- If distinctiveness is too low, you can "dial up" the marker density (at a higher cost) or consult breeders to understand the genetic relationships. This avoids collecting field data that cannot be matched.



Distinctiveness

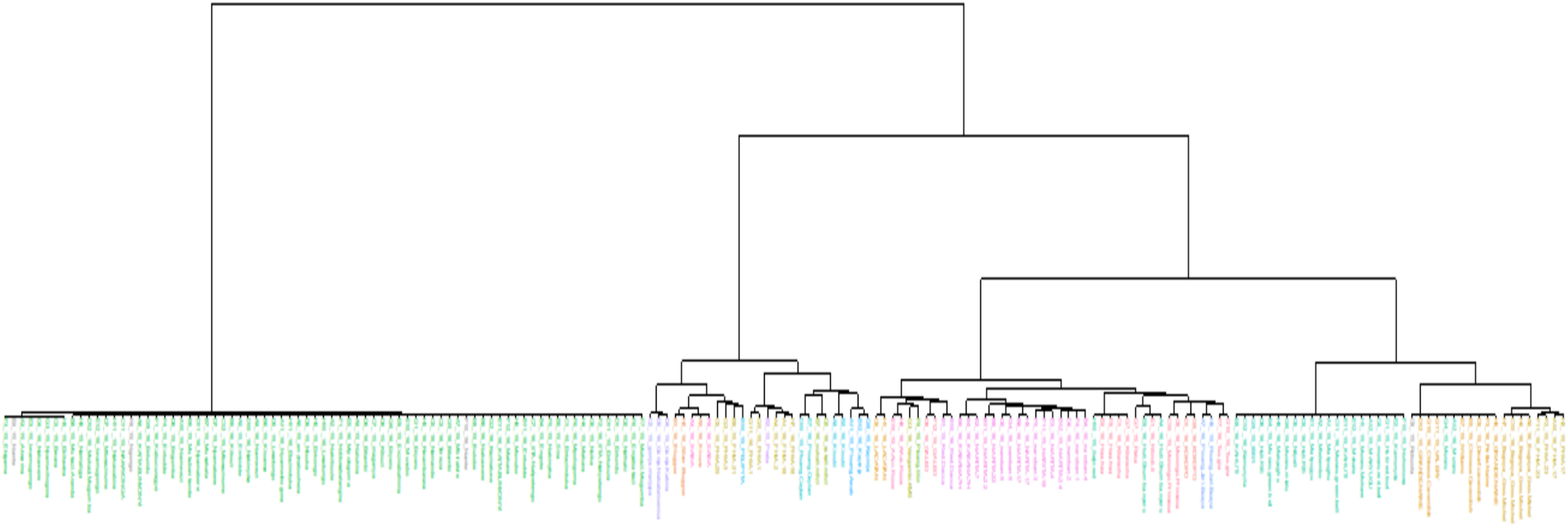
Cassava example



Good branching on the dendrogram indicates ability to distinguish between varieties

Distinctiveness

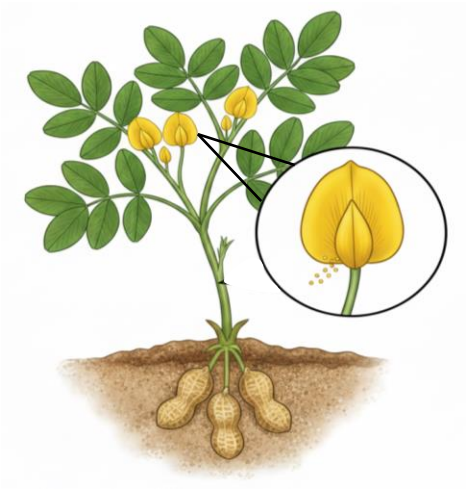
Banana example

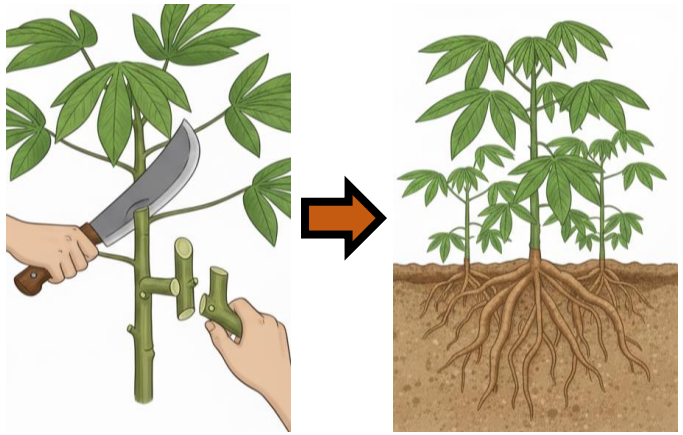


Contrast between indistinct varieties (Matooke landrace) and more distinct improved varieties

Field sampling considerations

Crop reproductive biology is key





Outcrossing

- Pollen from one plant fertilizes a different plant
- High genetic variation. Each seed is a new, unique combination of its parents
  - Maize
  - Pearl millet
  - Alfalfa

Self-pollinated

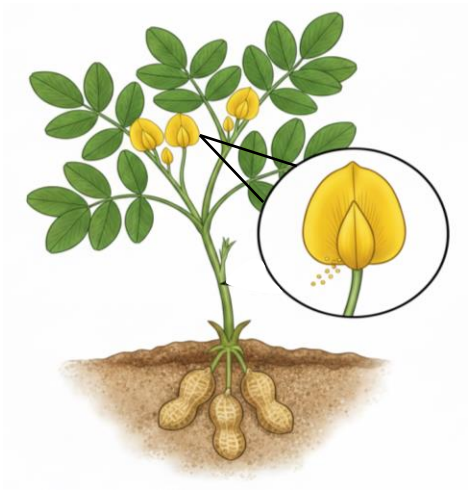
- The plant's pollen fertilizes its own ovules
- Very low genetic variation, leading to highly uniform varieties
  - Rice
  - Wheat
  - Beans

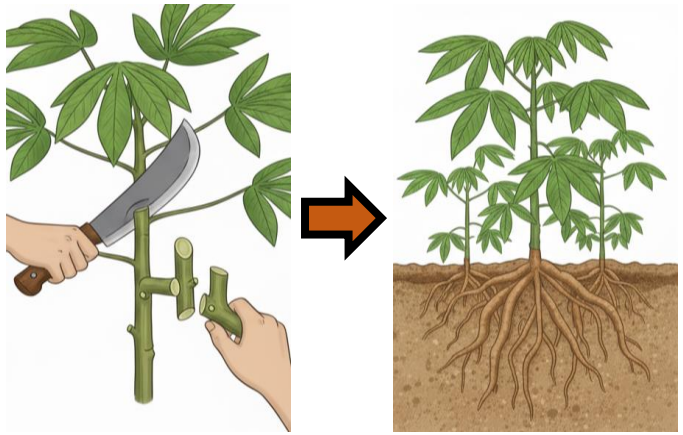
Clonally propagated

- New plants grow from vegetative parts e.g., tubers & cuttings
- The new plants are genetically identical clones of the parent
  - Cassava
  - Potato
  - Banana

Field sampling considerations

Special notes





Heterozygosity: Genetic state where an individual has two different forms (alleles) of a gene

Outcrossing

- Shuffles genes every generation
- Results in individual plants with high levels of heterozygosity
- Populations are genetically diverse, hence heterogeneous

Self-pollinated

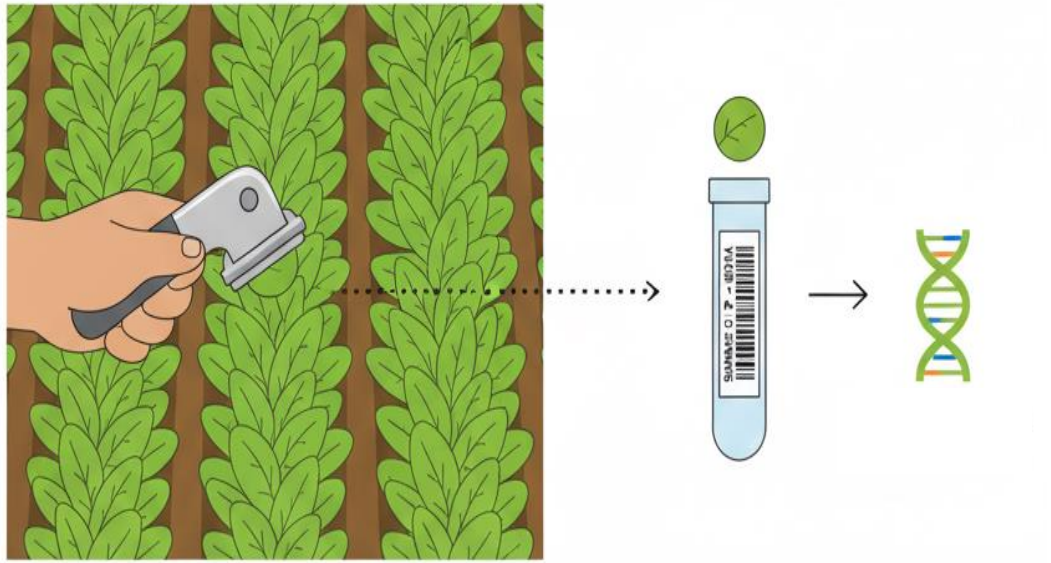
- Systematically reduces heterozygosity by 50% with each generation
- Results in homozygous plants and homogenous populations
- High heterozygosity means:
  - A recent - likely accidental, cross-pollination event occurred or
  - Presence of physical mixture

Clonally propagated

- Levels depend on underlying reproductive biology of the crop
- Clonal propagation freezes the genetic profile of the parent
- Seed germination in fields alters the original variety profile

Sampling self-pollinated and clonal crops

Homogenous plots expected

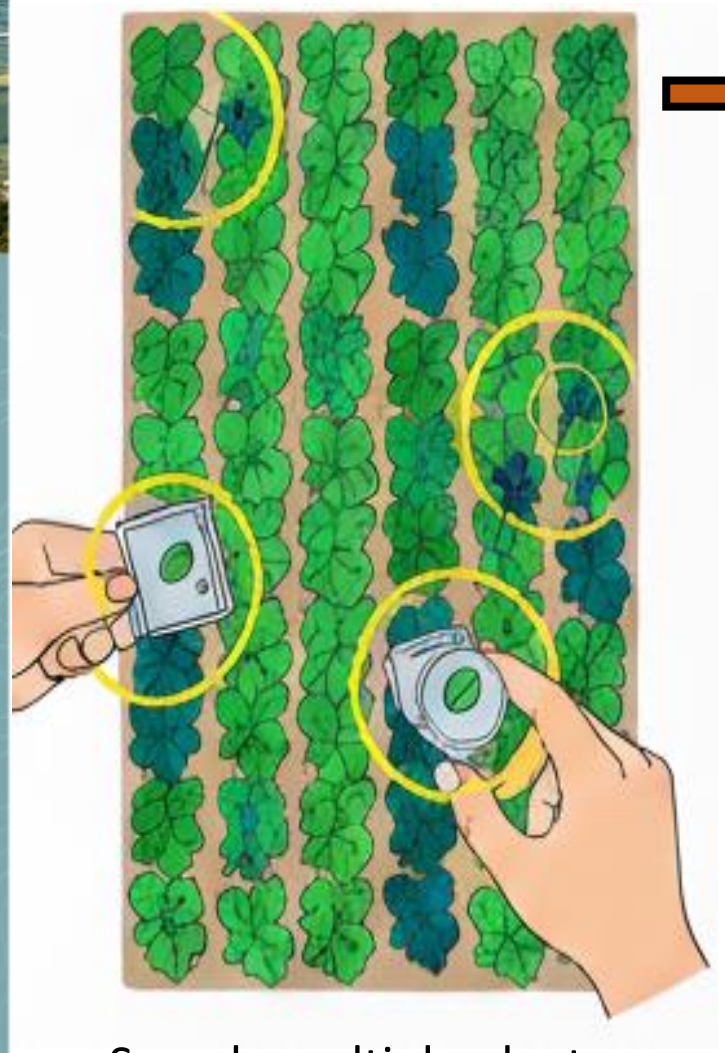


Plots consists of a single, uniform variety

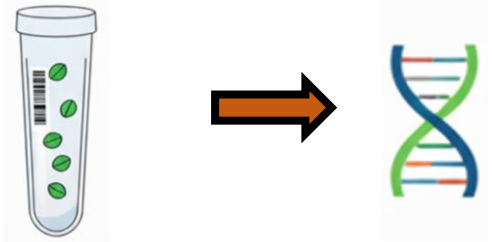
- **Recommendation:** Single Tissue Sampling
- **Procedure:** Collect a leaf disc from one representative plant
- **Rationale:**
  - If the plot is truly uniform, a single plant accurately represents the entire plot's genetic profile
  - This is the most cost-effective and straightforward method
  - Leads to unambiguous variety identification

Sampling self-pollinated and clonal crops

Heterogenous plots expected



Sample multiple plants



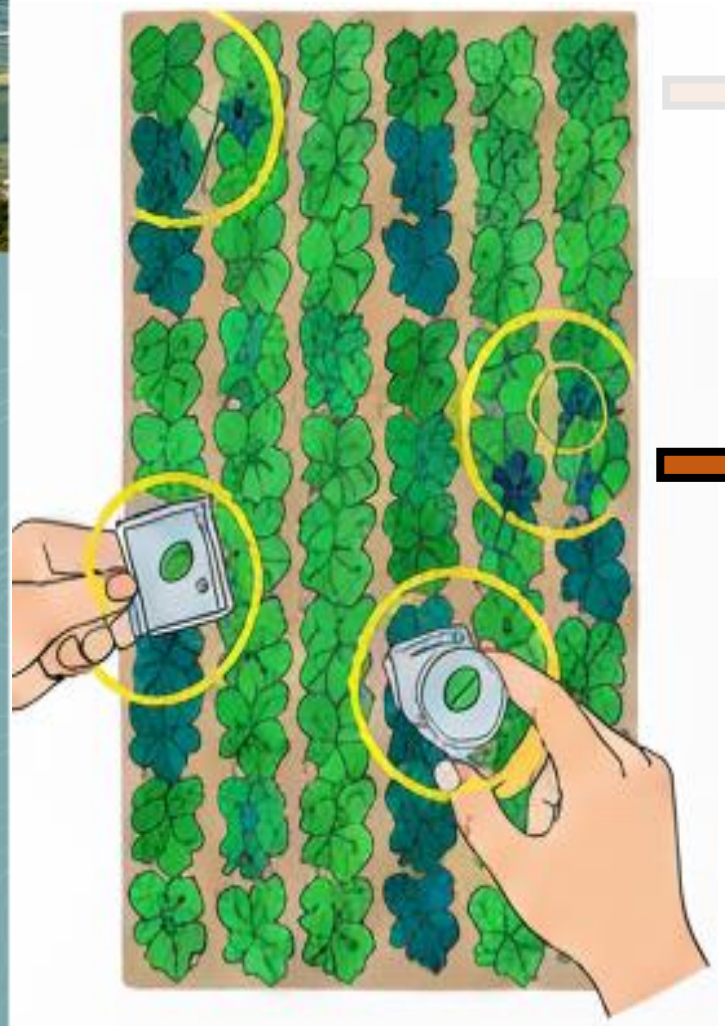
Confirming admixture

Objective: Admixture confirmation

- **Procedure:** Collect a small bulk sample
- **Rationale:** Cost-effectively tests for heterogeneity
  - Will show high heterozygosity if plot is mixed
  - It cannot identify the dominant variety

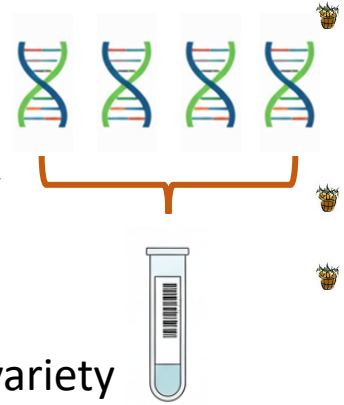
Sampling self-pollinated and clonal crops

Heterogenous plots expected



Sample multiple plants

Identify dominant variety

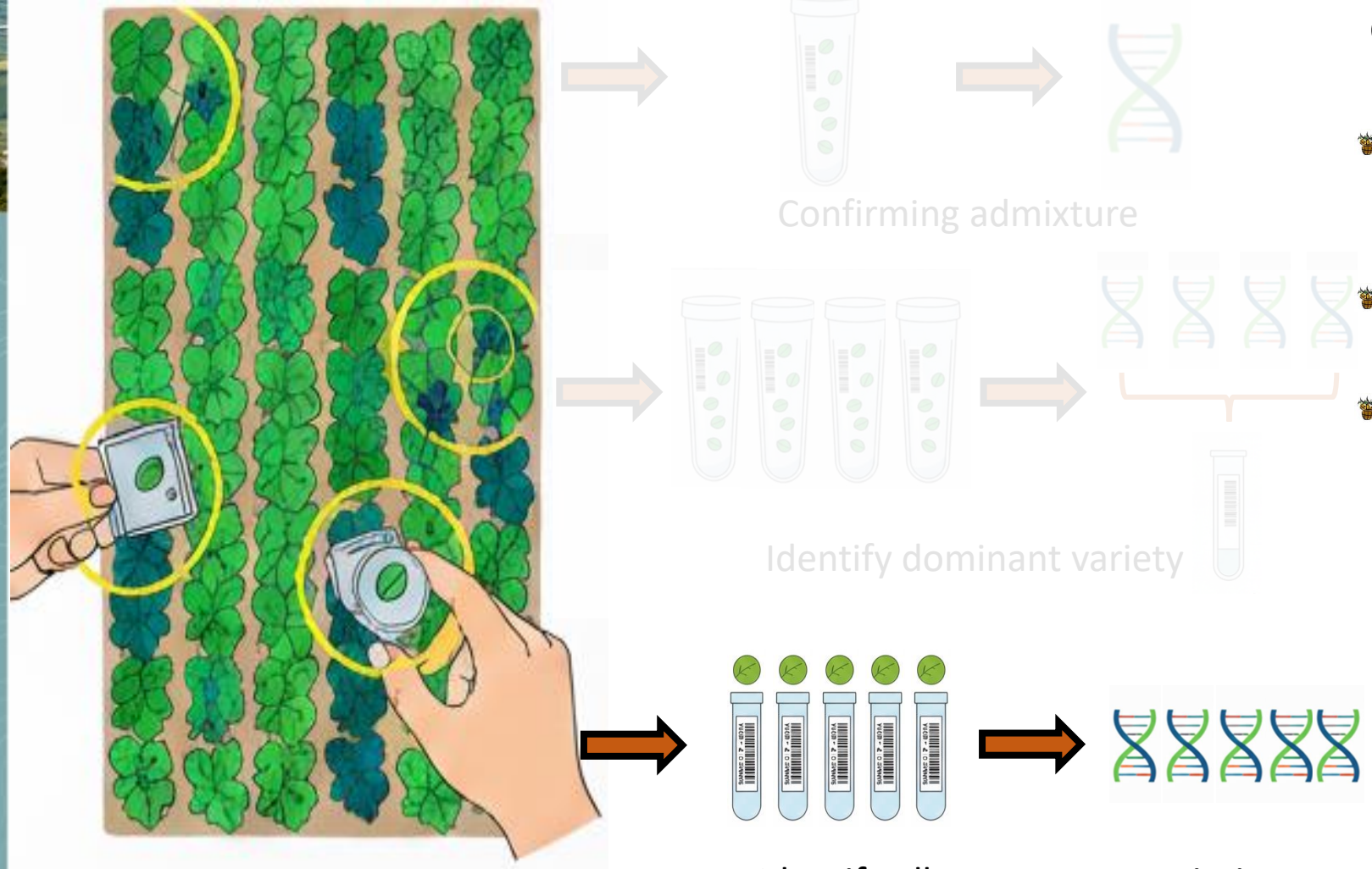


Objective: Identify dominant variety

- **Procedure:** Collect a large bulk sample
- **Rationale:** High sample numbers increases chance detecting dominant variety signal
- Logistically challenging
- May not yield a clear ID if no single variety is dominant

Sampling self-pollinated and clonal crops

Heterogenous plots expected



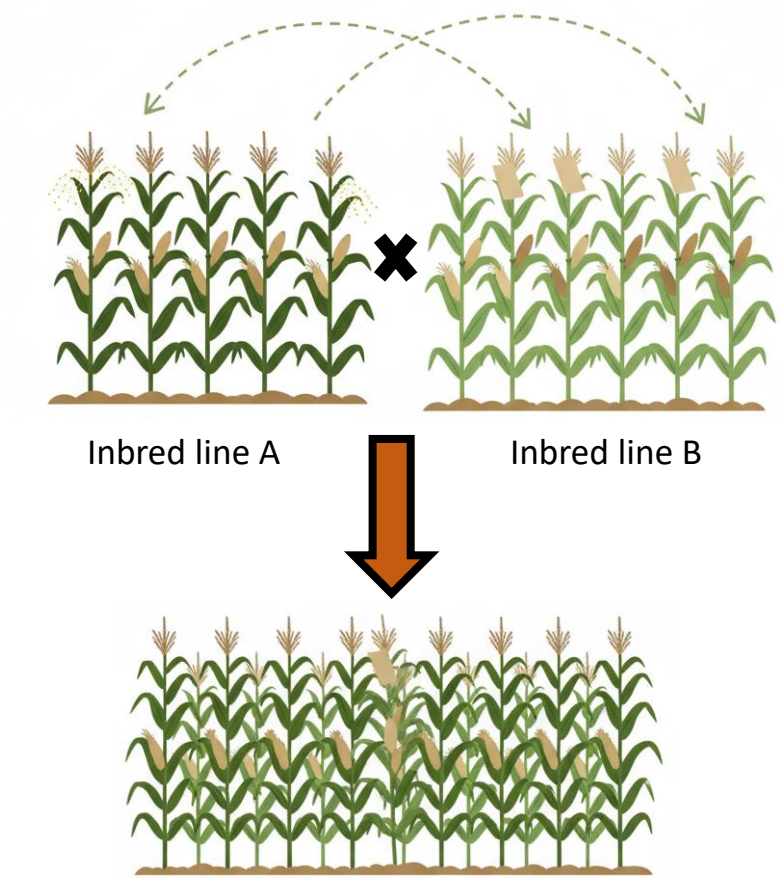
Objective: Identify all component varieties

- **Procedure:** Collect and process multiple single tissue samples from different plants within the plot
- **Rationale:** Only way to know exact composition of an admixed plot
- Provides high resolution but at high cost

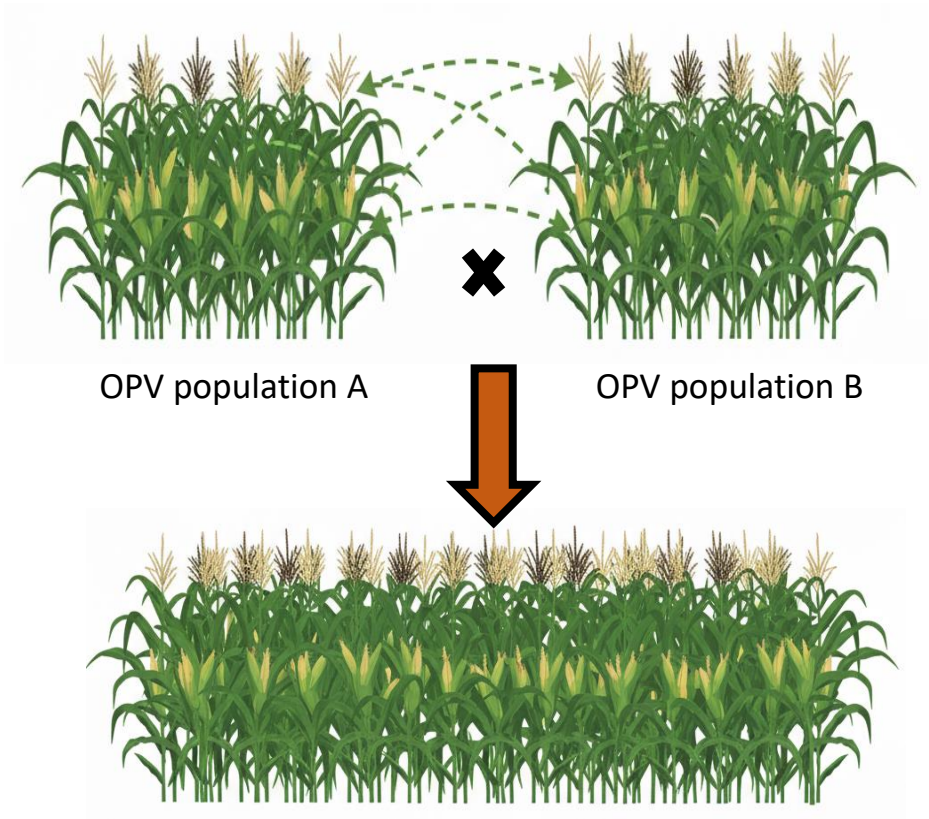
Sampling cross-pollinated crops: maize

Hybrids vs OPVs

Hybrid maize



OPV maize



In both cases, sampling using crop-cuts is recommended

Sampling by crop-cuts: maize

In general

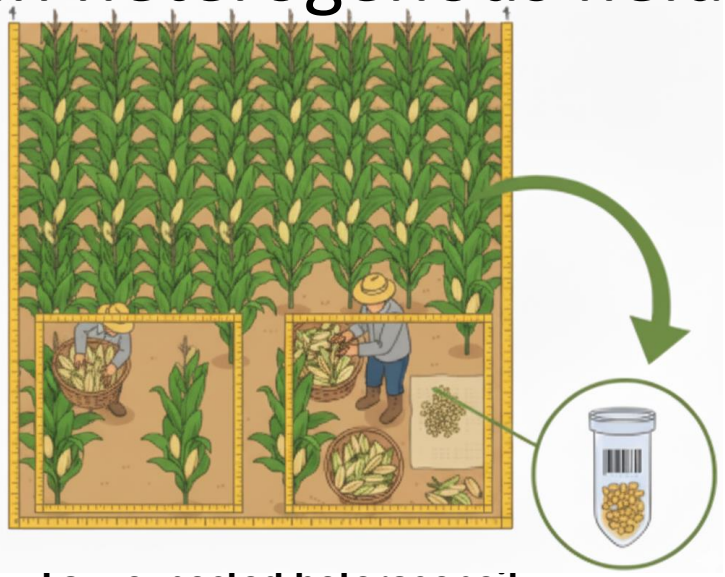
Yield estimation sampling tool used for collecting samples for DNA extraction

Procedure

- Move at least 5 meters in from the plot edge to minimize pollen contamination from neighboring fields
- Establish a single 4m x 4m crop-cut quadrant
- Harvest all maize cobs from the plants within the quadrant
- After shelling and drying the grain, create a composite sample for DNA extraction by taking 50-100 grains sourced from at least 15-20 different cobs from the harvest
- This works well for homogenous plots eg with a commercial hybrid and uniform visual appearance

Sampling by crop-cuts: maize

In heterogenous fields

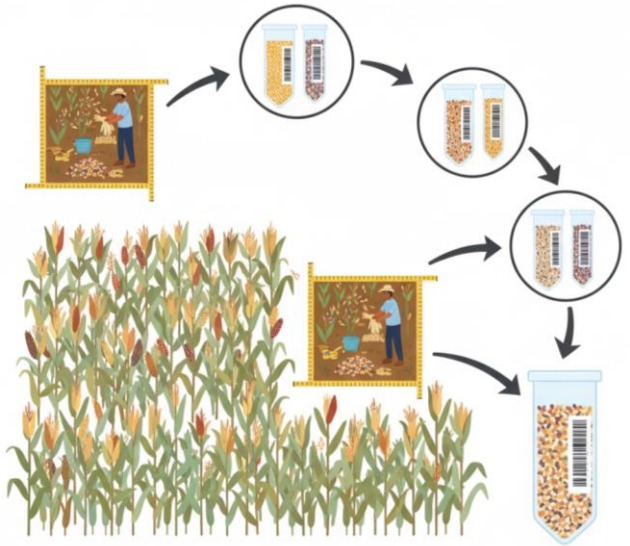


Low expected heterogeneity

- E.g. OPV field with a chance of mixture from neighbor
- Collect 2 standard crop-cuts
  - Establish two 4m x 4m crop-cuts
  - Harvest the cobs from each quadrant separately
  - Combine to create a single composite sample for DNA extraction

Moderate expected heterogeneity

- E.g. if farmer plants two varieties in distinct sections of the same plot
- Collect stratified crop-cuts
  - Ask farmer to identify the different sections of the plot
  - Establish a 4m x 4m crop-cut for each section
  - Harvest, label and process the grain separately

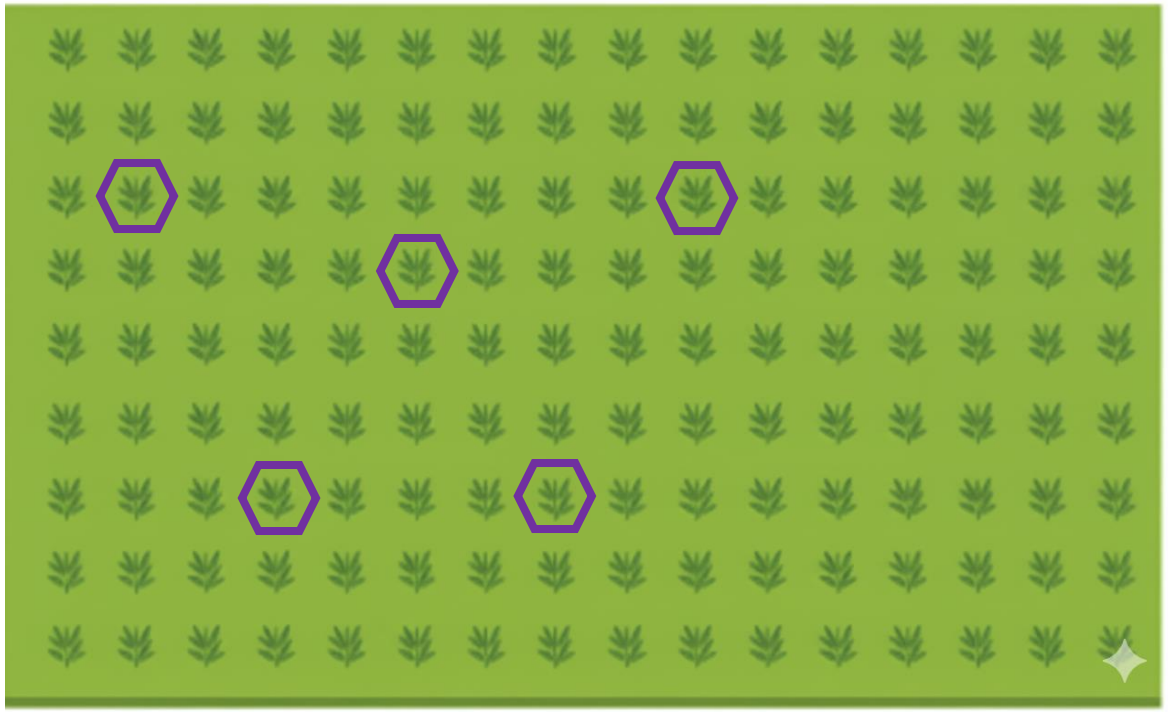


High expected heterogeneity

- E.g. if farmer plants recycled seed of unknown origin or grain from market
- Collect multiple smaller crop-cuts
  - Randomly establish 3 to 4 smaller crop-cuts (e.g., 2m x 2m) across the plot
  - Harvest, label and process the grain separately
  - Genotype separately or pool samples depending on study

Leaf sampling using random walks

Self-pollinated and clonal crops



Systematic random selection of plants in a plot

- Homogenous plots: Useful for sampling to confirm intra-plot homogeneity
- Heterogenous plots: Adequate capturing of full genetic diversity in line with study objectives

Procedure

- Walk a "W" pattern across the plot
- Collect samples from the pre-determined number of plants based on the study objective at regular intervals along the walk
- Place 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, stratify the plot by dividing it into two halves
- Perform the "W" pattern walk in both halves

Genotype data for bulk samples

SNP data formats

Score data

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
SNP_001	GG	GG	GG	GG	GC	AA
SNP_002	CC	CC	CC	CC	CG	GG
SNP_003	TT	TT	TT	TT	TA	CC
SNP_004	CC	CC	CC	CC	CG	AT

Most genotyping service providers supply score data

Counts data

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
SNP_001	163	191	243	246	235
SNP_001	104	10	81	0	8
SNP_002	45	16	60	24	63
SNP_002	8	13	9	30	0
SNP_003	89	39	124	42	60
SNP_003	177	126	70	99	123
SNP_004	10	5	30	4	16
SNP_004	59	40	31	64	60

Easiest to get counts data from Diversity Arrays

- Counts data is critical when bulk samples are collected for maize, or for suspected admixed inbreds where the objective is to identify the dominant variety
- Counts data enable quantitative allele frequency estimation for interpreting the composite DNA signal from a mixed sample.

Summary and key takeaways

Standing Panel on Impact Assessment
CGIAR

For the reference library...

- Be exhaustive: Start with official release lists from NARS and CGIAR
- Purity is key: Source breeder seed as the gold standard
- Verify distinctiveness: Genotype the library first to ensure markers can tell varieties apart

For field sampling...

- Know your crop: Understand the biology of your crop of focus
- Tailor strategy to the plot: Understand how expectations of homogeneity and heterogeneity impact sample collections for outcrossing, inbred and clonal crops

Structuring the Questionnaire

CGIAR Standing Panel on Impact Assessment

How many (farmer's declared) varieties to be sampled?

→ Be mindful of the fact that many farmers will not be managing their planting materials as if they are distinct varieties

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")
c) Limit to what the farmer considers to be the "main variety".

→ Loss of information, but this is inversely proportional to the cost and time required for the data collection.

Be conservative and assume that farmers don't have a lot of information about their planting material and its purity.

Structuring the Questionnaire

Standing Panel on Impact Assessment
CGIAR

Objective: accurately link the varietal identification data generated to household, plot, and variety-level information

Plot roster and module

Comprehensive list of all plots managed or cultivated by a household
Allows for random selection of plots for sampling

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				

Structuring the Questionnaire

CGIAR Standing Panel on Impact Assessment

Objective: accurately link the varietal identification data generated to household, plot, and variety-level information

Plot module: common questions

- **How many varieties of rice were planted on this plot?**

→ Enumerators should be trained not to insist too strongly on soliciting a response to the number of varieties in Q5 and allow for “Don’t know”.

- **What is the name of the <u>main variety</u> planted on this plot?**

→ We advise against listing every variety name and asking farmers to select from the list, but rather train the enumerators to record text of the response given by farmers.

- **What type of rice is the <u>main variety</u> planted on this plot?**
- **Did you use certified seeds on this plot ?**
- **What is the source of the main seeds planted on this plot?**
- **For how many seasons have you re-used the main variety planted on this plot?**

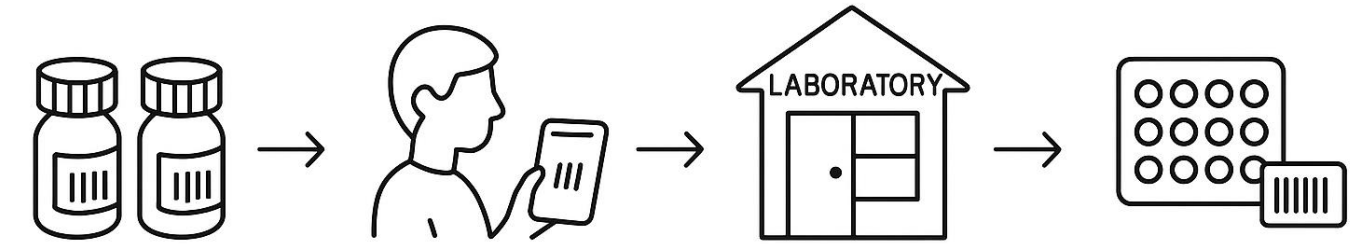
Structuring the Questionnaire

Standing Panel on Impact Assessment
CGIAR

Variety * Plot module [VHLSS 2022]

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

From Field to the Lab



Barcoded samples delivered to enumerators

Enumerators scan the sample barcodes

Samples arrive to the laboratory

Samples are plated in a plas-ticware using an application

HH survey data

Barcode ID

HH survey

Genetic data

Barcode ID

Results

From Field to the Lab

CGIAR Standing Panel on Impact Assessment

Barcoding

Allows tracking samples along the various step of the chain while also minimizing human errors involved in data entry

Barcode size:
Must fit to the support

Fieldwork conditions:
- Stickers without protection may suffer from transport, rain, or dust
- check the suitability of devices used for barcode reading

Technical replicates:
- Collected samples may be split for different purposes
- We recommend naming replicated barcodes with an underscore. For example, a replicate of sample 08733 should be named as 08733_1.

From Field to the Lab

CGIAR Standing Panel on Impact Assessment

Tracking File example

Only when submitting plated samples to the lab

PlateID	Row	Column	Barcode IDs
1	A	1	Q5-1476
1	B	1	Q5-1262
1	C	1	Q5-1443
1	D	1	Q5-1463
1	E	1	Q5-1066
1	F	1	Q5-1083
1	G	1	Q5-1467
1	H	1	Q5-1163
1	A	2	Q5-1235
1	B	2	Q5-1039
1	C	2	Q5-1173
1	D	2	Q5-1131
1	E	2	Q5-1345
1	F	2	Q5-1319
1	G	2	Q5-1357
1	H	2	Q5-1150
1	A	3	Q5-1478
1	B	3	Q5-1245
1	C	3	Q5-1105
1	D	3	Q5-1392

CGIAR Standing Panel on Impact Assessment

Questions