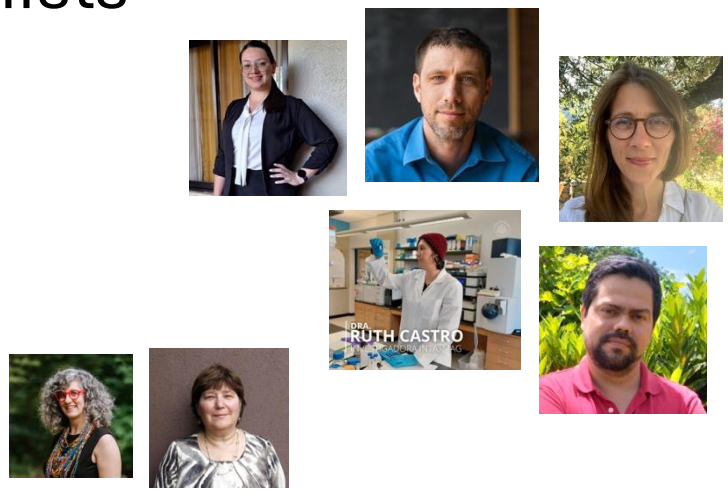
Varietal Knowledge and Genomic Uncertainty in Adoption Studies using DNA fingerprinting

Martina Occelli | mo386@cornell.edu | Cornell University | EQUAL Lab

This presentation

- Nuanced considerations while designing survey-fingerprinting protocols

- Work led by INTA (Costa Rica), with a team of geneticists, breeders and agricultural economists



Introduction

- Complementing observational studies, DNA fingerprinting has quickly emerged as the new protocol-to-use in adoption studies (Euler et al. 2022; Yigezu et al. 2019)
- There are detailed reviews of the DNA fingerprinting-based studies conducted since 2015, their findings and implications for the field of technology adoption (e.g., Euler et al. 2022; Stevenson et al. 2018)

- In this booming literature, few attention has been paid to domain-specific methodological choices (Poets et al. 2020; Gimode et al. 2025)

Introduction

- We ask: what happens to estimates of farmer- and fingerprinting-based varietal identification when genomic- and social science-specific methodological choices are made explicit?

- Which farmer should the team “go into the field with” for varietal identification? [social science-specific methodological choice]
- Is there a “confidence interval” attached to the DNA fingerprinting exercise? [genomic-specific methodological choice]

Why it matters?

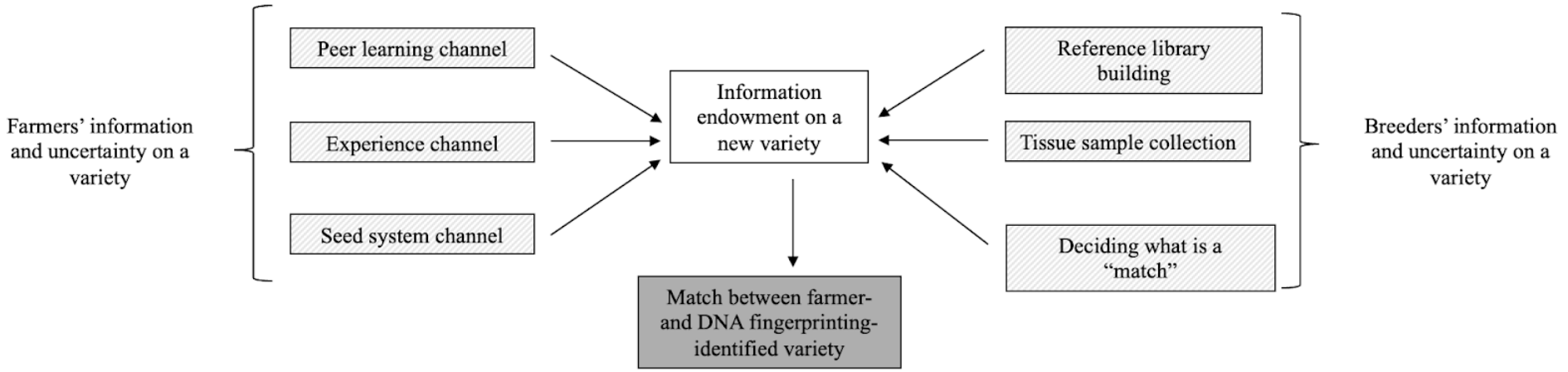
- Reversing the “burden of proof” (Waldman et al. 2025) and bringing at the forefront methodological choices to deliver better adoption data

- Targeting instruments, especially in cases when public breeding programs have limited resources

- Pushing the boundary between disciplines and knowledge paradigms

The study

- We hypothesize that varietal identification by farmers and scientists is guided by two different processes



The study

- We investigate if rates of varietal misidentification change, when we are more explicit on aspects of those processes
- We use the term match

- The sample: bean growing households in Costa Rica
  - Two-step sampling procedure (600+100)
    - HH type 1: one member engaged in bean cultivation
    - HH type 2: at least two members engaged in bean cultivation
  - 17% replacement
  - 670 households, 668 full dataset | 420 HH type 1 + 248 HH type 2

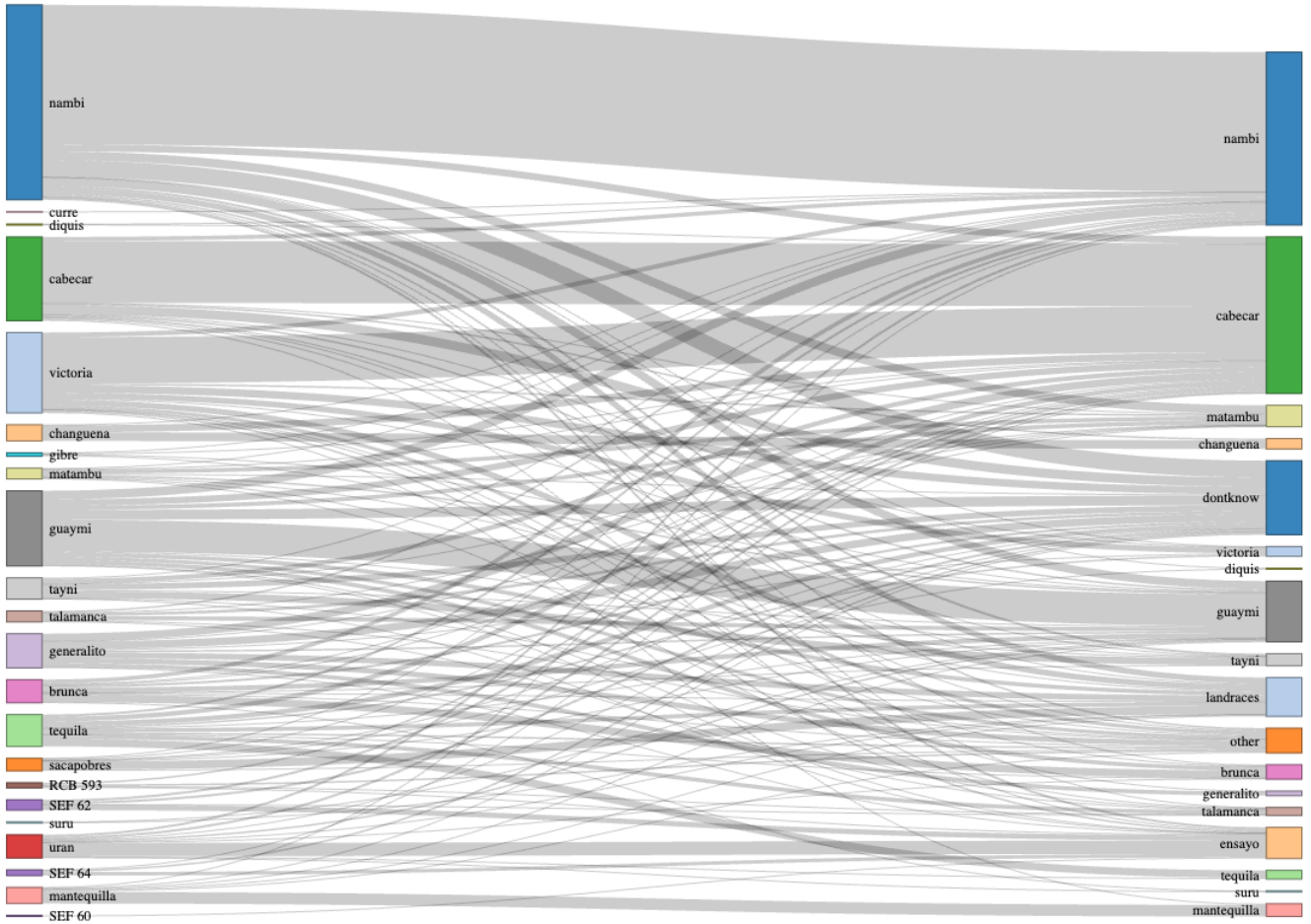
The protocols

- Step 1: standard vs respondent-intentional fingerprinting protocol
  - Information index

- Step 2: tissue collection and genomic analysis protocol
  - Up to 5 varieties, starting from the main plot
  - Random walk, 7 leaves (plastic tube with silica)
  - 1,221 plots in total (1,148 samples fingerprinted)
  - Reference library: 23 genotypes (14+6+3)
  - Genetic distance and similarity were used to determine DNA identity

Variables of interest

- Match



Sankey diagram capturing the relationship between common bean varieties identified through DNA fingerprinting (left) and comm on bean variety names given by farmers in the sample (with data cleaning) (right)

The bars indicate percentage of total varieties, while lines describe the relationship.

Variables of interest

- Match
- Information index
  - It doesn’t measure who is correct

Variables of interest

- Reference library variability
  - Genetic distance within varieties of our reference library (n=3 replicates per variety)
  - The standard deviation of the within-variety genetic distance in the reference library (standardized, mean 0 | variance 1)

- Fingerprinting Confidence Interval (within FCI)
  - From values of the pairwise similarity matrix within varieties of the genetic library, we compute the distribution for each variety (min, max and mean genetic distance)
  - We take the max as the cut-off threshold for each variety, below which we cannot declare that two samples represent a different clone
  - We then compare the genetic distance between the sample of the variety taken in the field and the first matching variety identified by fingerprinting, with the cut-off threshold for that variety in the reference library (within-variety genetic distance)
  - If the value of the genetic distance between the sample and the first fingerprinting match is WITHIN the cut-off threshold which makes within-variety replicates undistinguishable, DNA fingerprinting can identify that variety with a good degree of certainty.
  - To signal these cases, we construct a dummy variable named Within Fingerprinting Confidence Interval (or Within FCI), to which we assign value 1.

The analysis

- Covariate balancing

Variable	HH with one member engaged in beans	HH with at least two members engaged in beans	Difference p-value
Age	50 (13.01)	49 (13.76)	0.31
Gender (=1, women)	0.14 (0.34)	0.13 (0.34)	0.76
Education	6.15 (2.98)	5.87 (3.27)	0.15
Household size	3.88 (1.79)	4.49 (1.31)	<0.01***
Decision-maker on bean varieties to be planted (=1, respondent alone)	0.83 (0.36)	0.29 (0.45)	<0.01***
Varieties planted during this growing season (number, mean)	1.79 (0.89)	1.65 (0.76)	0.03**
Plots (number, mean)	2.79 (0.83)	2.67 (0.76)	0.03**
Bean intercropped (=1, yes)	0.09 (0.29)	0.06 (0.24)	0.06*
Bean sold (=1, yes)	0.94 (0.23)	0.87 (0.33)	<0.01***
Regular contacts with extension system (=1, yes)	0.12 (0.32)	0.11 (0.31)	0.61
N	420	248	

Significance level: p-value < .01 (***); < .05(**); < .10 (*). In parentheses, standard errors.

The analysis

- Covariate balancing

- We regress the information index on the *match* variable (first fingerprinting hit, >90% similarity)
- We vary the definition of a *match* and we regress on those definitions two proxies of fingerprinting variability and uncertainty
- We study determinants of higher varietal knowledge by respondents

Result (1) Knowledge driven gains in varietal identification

- High knowledge relate to an eight pp increase in match (19%)
- No difference in low knowledge among the two protocols
- Findings depend on the knowledge of the respondent, not simply by choosing someone else

Table 1 | Treatment effects on varietal identification

	Match (1)	Match (2)	Match (3)
Low Knowledge	0.00	0.04	0.04
	(0.04)	(0.04)	(0.04)
High Knowledge	0.08**	0.07**	0.07**
	(0.04)	(0.03)	(0.04)
Comparison mean value	0.41	0.41	0.41
Observations	1121	1121	1083
Variety-level controls	no	yes	yes
Socioeconomic controls	no	no	yes
Region Fixed Effects	yes	yes	yes

Notes: Table reports coefficient estimates from a linear probability model using as outcome variable match between the farmer and the first-match variety identified by the DNA protocol. Match is restricted to cases when the average genetic distance is lower than 10%. Outcome variable *match* is measured at the variety level. *High Knowledge* variable equals one when a farmer level ranks in the top 50% of the information index distribution. Variety-level controls include indicator variables identifying when a farmer did not know the name of the variety, varieties reported by farmers not included in the reference library, landraces, and varieties from breeding trials with no common names. Socioeconomic controls include education level in levels and indicators for single-member households, gender, participation in INDER's extension program, seed source, participation in Tricot trials and whether it is the first time planting the variety. Robust standard errors clustered at the farm level are reported in parenthesis.
Significance level: p-value <0.01 (***); < 0.05(**); < 0.10 (*).

Result (2) Varietal identification varies at the varying of DNA fingerprinting uncertainty

- Match rates increase from 41% to 58%
- High reference variability is linked to an eight pp decrease in the match rate
- Approx. a third of the predicted match rate (23 pp out of 0.56) can be attributed to cases with no uncertainty about the reference library's ability to identify varieties

Table 2 | Variation in Treatment Effects by Similarity and Reference Uncertainty

	First Match + Similarity >90%	First Match	Second Match	Third Match	Fourth Match
	(1)	(2)	(3)	(4)	(5)
Low Knowledge	0.03	0.03	0.02	0.04	0.01
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
High Knowledge	0.08**	0.09***	0.09**	0.07**	0.05
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Within FCI	0.23***	0.17***	0.09***	0.06**	0.04
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Reference variability	-0.08***	-0.08***	-0.09***	-0.08***	-0.08***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Comparison mean value	0.41	0.45	0.53	0.55	0.58
Mean genetic similarity	0.94	0.94	0.89	0.88	0.87
Observations	1083	1083	1083	1083	1083
Region Fixed Effects	yes	yes	yes	yes	yes

Notes: The table reports coefficient estimates from linear probability models where the outcome variable indicates a match between farmer-reported and DNA-identified bean varieties, based on varying levels of genetic similarity. Column 1 reports matches based on the closest genetic match, defined as the sampled variety with the lowest genetic distance to the reference library and restricted to cases with similarity above 90%. Column 2 reports matches based solely on the closest genetic match, regardless of similarity threshold. Columns 3 through 5 progressively expand the outcome to include the second, third, and fourth closest matches. All models include variety-level controls and region fixed effects. Robust standard errors, clustered at the farm level, are reported in parentheses.
Significance level: p-value <0.01 (***); < 0.05(**); < 0.10 (*).

Result (3) Channels relating to higher varietal knowledge

Table 3 | Determinants of varietal knowledge

	Knowledge Index (1)	Knowledge Index 1 km (2)	Knowledge Index 5 km (3)
Intentional Selection	0.28*	0.30*	0.28*
	(0.16)	(0.16)	(0.16)
Nearest-Neighbor Matches		0.05**	0.01*
		(0.02)	(0.01)
Tricot Trials	0.33**	0.40***	0.42***
	(0.15)	(0.15)	(0.15)
INDER Program	0.21	0.19	0.17
	(0.19)	(0.20)	(0.20)
Seed Source	-0.07	-0.03	-0.00
	(0.14)	(0.15)	(0.15)
Education	0.01	0.01	0.01
	(0.02)	(0.03)	(0.03)
Gender	0.10	0.06	0.06
	(0.18)	(0.19)	(0.19)
Constant	5.66***	5.52***	5.44***
	(0.27)	(0.28)	(0.31)
Observations	1121	1049	1049
Region Fixed Effects	yes	yes	yes

Notes: The table reports coefficient estimates from a linear regression model, where the outcome variable is represented by the score-based information index. Column 1 shows results for the information index overall, while Columns 2 and 3 account for estimates of the nearest-neighbor matches in a 1km and 5km radius, respectively. All models include variety-level controls and region fixed effects. Robust standard errors, clustered at the farm level, are reported in parentheses.

Significance level: p-value <0.01 (***); < 0.05(**); < 0.10 (*).

Discussion (still in progress)

- Results suggest that mismatch may arise when DNA fingerprinting protocols rely on unstructured respondent selection
  - Importance of knowledge-based targeting

- We find evidence that inherent uncertainty in genomic methods accounts for a share of the observed mismatch between survey- and fingerprinting-based estimates
  - Degrees of freedom available to researchers

Thank you!

Martina Occelli | mo386@cornell.edu | Cornell University | EQUAL Lab