DNA fingerprinting for estimating varietal adoption: Summary of Case Studies
Cassava in Ghana

Partners: Mywish Maredia (Michigan State University), Byron Reyes (CIAT); Joe Manu (Ghana Crops Research Institute); Awere Dankyi (Agriculture Innovations Consult (AIC)); Peter Kulakow, Ismail Rabbi, Elizabeth Parkes, Tahirou Abdoulaye and Gezahegn Girma (IITA); Ramu Puna (Cornell)

Overview

- **Geographic scope:** The pilot study was implemented to be representative of three regions – Brong Ahafo, Ashanti and Eastern (these 3 regions account for 61% of cassava production).
- **Objectives:** The main objective of the study was to test different approaches of collecting variety-specific adoption data and to validate them against the benchmark of DNA-fingerprinting to determine which method is most accurate and cost-effective in measuring varietal adoption. Alternative methods tested:
  - V DNA fingerprinting (used as a benchmark to compare/validate other methods)
  - A Farmer elicitation (name and type of variety)
  - B Farmer elicitation based on series of photographs of plants at different growth stages and later identifying varieties based on morphological characteristics
  - C Trained enumerators/experts visiting the field and:
    1. Recording observations on varietal characteristics (phenotyping); and
    2. Identifying the variety based on observation (phenotyping)
  - D Taking photos of the plant in the field or seeds harvested by farmers for latter identification by experts (i.e., breeders, agronomists, etc.)
- **Current status:** Working paper in progress and a draft article will be prepared for publication.

Methodology

- **Sampling and data collection**

A total of 500 households across 100 villages (5 farmers in each village) were targeted for the survey using a multistage cluster sampling method (district, village and farmer samples selected randomly). Survey conducted in October-November 2013, and the team consisted of: Enumerators to complete the household modules; Cassava expert in-charge of completing the field survey module (method C and D); and DNA sampling expert for collecting, labeling and storing the plant tissue samples as per the protocol established.

- **Sampling for DNA fingerprinting**

For each HH, a cassava 'expert' visited one cassava field with the largest number of cassava varieties as declared by the farmer. For each variety as identified by the farmer, the 'expert' randomly selected one representative plant and collected leaf tissues from the youngest (apical) leaves. A versatile and economical sampling kit was developed. Leaf tissues were collected in a small screw-capped plastic jar with ~20 g dessicated silica gel. Samples were also collected from plants that had observed variations in morphological characteristics as assessed by the 'expert', but were identified by farmers as belonging to the same variety. A total of 917 samples were collected for genotyping from 495 cassava plots visited.

- **DNA Genotyping method, logistics and data analytics**

Getting high quality DNA for genotyping is a challenge, especially when samples are collected from farmers' fields, hundreds of miles from extraction labs. In the case of Ghana, samples from the farmers' fields were first brought to Kumasi, and then shipped to IITA (Ibadan) where a low-cost and high-throughput DNA extraction system was used to isolate DNA from > 1000 samples in less than one week. DNA was freeze-dried and shipped to Cornell for varietal identification using 'genotype by sequencing' (GBS) methodology.

A total of 64 accessions of released varieties (n=18) and popular landraces (n=46) were included in the reference library. Samples of these accessions along with the samples collected from farm surveys were all genotyped at 56,849 single nucleotide polymorphisms (SNP) loci. Genetically identical sets of clones were then identified by using

distance-based hierarchical clustering and model-based maximum likelihood admixture analysis (done by researchers at IITA).

Results and Insights

About 180 variety names were reported by farmers across 914 accessions collected from farmers' fields. These 914 field samples and 64 accessions for the reference library were classified into 11 unique varietal clusters and several hybrids or admixtures based on DNA fingerprinting. There are several interesting findings revealed by DNA fingerprinting of the farmer samples and library accessions that have implications on the breeding program, the seed system, and for impact assessment. **First**, some improved varieties were found to be genetically identical, and many fall under the same admixture group. This means, genetically identical (or closely similar) 'improved' cassava varieties have been released and are promoted in Ghana as distinct varieties. **Second**, except for 4 varieties, all other 14 released varieties were found to be hybrids or admixtures with different percentage of ancestry coming from the 11 unique varietal groups. This makes it challenging to precisely match the farmer samples with these 14 released varieties. **Third**, library accessions representing both 'released varieties' and 'landraces' fall under the same varietal cluster groups. This is the case with four unique varietal groups.

The third result especially poses a challenge for classifying farmer samples as improved vs. local varieties. The problem with this finding is that for farmer samples that fall in these four variety cluster groups, there is ambiguity as to whether they should be classified as 'improved/released' varieties or should they be classified as local/landrace varieties? This ambiguity makes it challenging to test the effectiveness of different methods, estimate varietal adoption, and to estimate the impact of breeding research. To circumvent this potential problem, effectiveness of different methods against the benchmark of DNA fingerprinting in this case study is done under two scenarios / assumptions: 1) Liberal scenario assumes that all the farmer samples that fall in a variety cluster in which there is at least one released variety are essentially improved varieties; and 2) Conservative scenario assumes the opposite. The only exception (under the conservative scenario) is Cluster group 4 (variety Afisiafi), which according to IITA cassava experts is unambiguously a cluster of improved varieties that previously did not exist.

The classification of 914 farmer samples as improved variety ranges from 31% under the liberal scenario to 4% under the conservative assumption. On aggregate level adoption of IV, methods based on farmer elicitation and field observations by an expert provide closest estimates under the conservative scenario, but with high error rate. No methods come closer to the 'truth' in adoption estimates under the liberal scenario. Identifying cassava varieties accurately by NAME in a setting where hundreds of variety names exist is a challenge across all the methods tested. Adoption estimates by the experts (based on photos) are substantially higher than other methods and has much higher type I error (false positives).

Key considerations, questions and challenges for scaling up

- Reference library – need to make sure materials used as references are unambiguous and represent their true identity. Need to first genotype the reference library of released varieties and accessions of landraces (obtained from a reliable source) to make sure the released varieties are genetically distinguishable from each other, and from the landraces. If the breeder seeds/materials of released varieties are not distinguishable from landraces obtained from a reliable source, then varietal identification will not be possible.
- Potential for scaling up as part of household surveys will depend on several factors:
    - Logistics of collecting, tracking, storing and transporting the samples from farmers' fields to a lab facility to get high quality DNA
    - Cost of DNA fingerprinting which includes—establishing the reference library, DNA extraction, and genotyping service. In this study the estimated cost per data point was ~$30
    - Capacity to do high volume DNA fingerprinting within the country or easy access to such capacity internationally (i.e.., no government restrictions on the shipment of plant tissues or DNA samples to other countries for analysis)

DNA fingerprinting for estimating varietal adoption: Summary of Case Studies
Case study on *cassava in Colombia & Vietnam*

Ricardo Labarta, Luis Augusto Becerra, Dung Phuong Le, Tatiana Ovalle, Stefan de Haan (CIAT), Mywish Maredia (Michigan State University), Vu Nguyen, Nguyen Trong Hien (Root Crop Center)

Overview

Geographic scope SIAC is funding the study in Vietnam, but this was built on a similar experience in Cauca, Colombia. In Colombia, the study targeted Cauca department where 30,000ha of cassava are grown for food consumption and for starch production. In Vietnam, we have a nationally representative study (500,000ha) covering all cassava main production areas that mainly target starch production.

Objectives In Colombia, the main objectives were 1) to develop the appropriate protocols (field sampling, planting material collection, labelling, DNA extraction), 2) test method of SNPs for DNA analysis developed by CIAT, and 3) to compare determinants of adoption using both farmers' self-identification and DNA fingerprinting identification. In Vietnam, we validated protocols tested in Colombia and added, a test for intra-plot diversity of varieties and the search for cost-effectiveness of DNA finger printing using different sampling strategies (number of households) at each community.
Current Status The Colombia study is completed and the first publications (methods & results) being finalized. In Vietnam, field level data collection is completed and being cured, All DNA extracted (3,500 samples) and being analyzed in CIAT

Methodology

In Colombia, due to the absence of official statistics, we reconstruct the data on cassava producing areas of the Cauca department. In our final sample, we visited farmers in 11 different municipalities (90% of cassava production in Cauca) and interview 305 cassava growers in 39 different communities. We collected plot and household level data for the full sample. We aimed to collect one sample for each household self-reported variety. However, many farmers had harvested the cassava or had young planting material by the time of the interview. We only collected 434 samples from 217 households.

Planting material collected followed a morphological identification and a careful labelling. We decided to transport samples collected to CIAT headquarters and plant them in greenhouses. After a month of emergence, DNA was extracted from each sample. This not only to facilitated the process but allow us to go back and re-extract the DNA when needed. We used Single Nucleotide Polymorphisms (SNPs) to construct improved genetic maps and look for trait associations, which was based on fluidigm genotyping.

In Vietnam, we replicated the protocols followed in Colombia. Some of the differences are: we had a nationally representative sample (power calculations performed) and selected 984 households in 82 farm communities across the country. We collected cassava-planting material from 834 households and 1,629 cassava stakes. We added a intra-plot diversity sampling and in a sub-sample of 100 farmers we collected randomly 15 cassava stakes from the same cassava plot. We used a similar approach to collect planting material and concentrate them in a central location (Root crop center, Hanoi). DNA was extracted in Vietnam and shipped to CIAT were DNA analysis is being done.

Results and Insights In Colombia we found the following results:

	Farmer Self-Identification		Identification though DNA fingerprinting	
	Improved	Landrace	Improved	Landrace
Number of Households	44	173	20	197
Percentage of Households	19.35%	80.65%	9.22%	90.78%
Percentage of Acreage	24.39%	75.61%	12.63%	87.37%

There are differences in determinants of adoption when considering farmers self-identification of improved varieties or when using DNA identification.

In Vietnam, we have preliminary estimated adoption of self-reported varieties:

Variety Name	English Translation	Total area(ha)	%	Number of HH	%
Improved			**95.2%**		**80.8%**
Cao San	High yielding	2,836.6	22.3%	314	33.1%
Tai Do	Red ear	1,780.3	14.0%	28	3.0%
Moi	New	1,106.4	8.7%	28	3.0%
Giong	Breeding	1,002.3	7.9%	29	3.1%
KM94	KM94	1,000.5	7.8%	88	9.3%
cut	Cut	449.2	3.5%	38	4.0%
Vedan	Vedan	423.8	3.3%	46	4.8%
Tay Ninh	Tay Ninh	271.4	2.1%	18	1.9%
Rau Muong	Spinach	242.0	1.9%	20	2.1%
Lai	Hybrid	97.4	0.8%	37	3.9%
Do	Red	62.7	0.5%	42	4.4%
Other improved (70)		2,841.3	22.4%	80	8.2%
Landraces			**4.8%**		**19.2%**
La Tre	Bamboo leaf	376.1	3.0%	86	9.1%
Xanh	Green	199.4	1.6%	55	5.8%
Other land races (5)		23.8	0.2%	41	4.3%

Insights of the process: For cassava, a vegetative propagated root crop, not using DNA fingerprinting led to an overestimation of adoption of cassava varieties**.** One challenge is availability of planting material or leave samples at household surveys time. Investing in enumerators training and labelling crucial, establishing several checks from the cassava field to the lab. Replanting seed/stakes has worked very well as backup strategy (possibility to correct an inadequate DNA extraction). Establishing the library for the analysis is a good investment for repeating DNA analysis (US$ 20-30,000 investment). The cost per sample analyzed could be between US$ 15-20. For us, this has been a process of developing capacity of national partners (DNA could be extracted in Vietnam and local scientists are trained to replicate the method). In the case of Vietnam, this opportunity has served to fingerprint the national cassava collection and eliminate duplicates in existing germplasm collection. With the knowledge accumulated, this method can be widely used and repeated at an affordable cost. The complementary morphological analysis performed in Colombia support the varietal identification done with DNA fingerprinting.

Adoption and impact of improved cassava varieties in Nigeria

Contact: Abdoulaye, Tahirou (IITA) <T.Abdoulaye@cgiar.org>

How the method was implemented

For the Nigerian case study, DNA based varietal identification has been conducted from cassava leaf samples obtained from 2500 households. Since most households grow more than one variety, the total samples collected from farmer's plot for DNA based varietal identification is about 7428. Herein, we summarise the procedures followed for field sample collection, establishment of tracking system from field to lab, implementing high throughput DNA extraction, genotyping, Bioinformatics and variety identification analysis.

1) Field sample collection, preservation and tracking

Leaf samples were collected and preserved in plastic tube containing silica gel for all the farmers identified varieties growing in each household and transported to bioscience laboratory at IITA for DNA extraction. Information including region ID, state ID, local government area ID, enumeration area ID, household ID, FieldID, PlotID and household head's name were captured on a booklet. In addition information on variety name, the GPS coordinates of the household where survey take place and the farmer's field were measured. In particular a detailed procedure (see Fig below) on how to collect samples was developed for training the enumerators.

As shown in the above figure, a standard tracking system is important particularly when dealing with a large sample size to reduce any possible introduction of human errors of sample mismatch and mix ups. A multiple layers of tracking system-using barcode label, self-adhesive stickers, booklet and tablet computer for capturing sample and sample-associated information were implemented. This process has improved the accuracy and reliability of the data. Duplicate barcode was prepared and pasted both on sample collection tube and booklet for each sample collected.

2) DNA extraction and Genotyping by Sequencing

DNA extraction was done from the above collected sample (a total of 7428 genotypes collected from 2500 households). In house modified protocol that enables to extract up to 10 plates of 96 samples each per day was implemented. All the extracted DNA samples were quantified using spectrophotometer and agarose gel electrophoresis for quality and quantity assessments. Furthermore, test digestion with restriction enzyme was performed for 10% of the samples extracted as suggested by Genetic Diversity Facility (GDF) at Cornell University for standard Genotyping by Sequencing (GBS) library preparation. DNA samples with high concentration were diluted to 1000ng/μl. All extracted samples that pass the minimum quantity requirement (300ng/μl) were shipped to GDF for genotyping by sequencing. For genotyping-by-sequencing library preparation, the ApekI restriction enzyme (recognition site: G|CWCG) that produces less variable distributions of read depth and therefore a larger number of scorable SNPs in cassava were used. Eighty 96-plex GBS libraries were constructed following the standard procedure and sequenced at the Institute of Genomic Diversity at Cornell University using the Illumina HiSeq2500.

3) Establishing a reference library:

IITA has a reference library with a total of 3891 diverse genotypes comprising a collection of known improved lines, IITA regional breeding program, IITA germplasm collection, wild species and CIAT collection. This reference library was used for the Nigerian case study.

4) Marker-based variety identification procedure

The following steps were implemented for varietal identification

1. Quality control of SNP data by removing those with > 30 % missing data and minor allele frequency of < 0.01%.
2. Calculate pairwise distance between all accessions,
3. Determine threshold for declaring two accessions as identical (using redundantly genotyped accessions)
4. Identify genetically identical sets of clones using distance-based hierarchical clustering,
5. Identify varieties by matching accessions from farmers to those in the reference library,
6. Determine ancestry estimates of hybrid clones using admixture analysis – especially if not direct match found in library.

Preliminary results

We then matched farmers own self-reported and DNA based varietal identification to identify the extent of adoption of improved cassava varieties. **What is an improved variety?** Classifying varieties into improved and land race is not straight forward even after DNA based varietal identification due to measurement and library matching issues. As a result, we developed two scenarios for defining improved varieties. In the first scenario, the distinction between improved and land race will be defined based on improvement status. As such, all varieties with genetic gain plus local varieties imported from abroad will be considered as improved. In the second scenario, we defined improved variety based on improvement status as scenario 1 but included land races that were purified and released. Based on the above definition, we estimated that 60% of farmers are using improved cassava varieties using self-reported data and 66% and 77% using scenario1 and 2 respectively.

DNA Gynotyping for Assessing Variety Area Estimates based on Farmer Identification: Case of Rice in Eastern India

by
P.C. Veettil , I. Gupta, T. Yamano, Z. Huelgas, G. Carino, T. Kretzschmar, and others

1. Introduction

To assess the area estimates based on farmers' identification of variety names in eastern India, we have collected 2,797 rice seed samples from 1,380 farmers in 2015. To verify the identity of the seed samples, we have also collected breeder seeds from seed companies and public institutes. The gynotyping of the farmer and breeder seed samples was conducted by using Illumina Infinium 6K SNP chips (Illumina Infinium 6K SNP - http://gsl.irri.org/services/infinium-6k). By using this approach, we effectively compare more than 4,000 SNP points across seed samples.

2. Data

The sampling of the 2015 survey used a simple self-weighting design across states. The total number of villages in each state was determined based on the total rice area in each state. A simple random sampling was used to select villages within each state by using the 2001 Census. In each of the selected villages, 10 households were randomly selected after listing households in the village. The total number of households interviewed was 6,740. From the sample households, 20 percent of them were selected for rice seed collection. In this paper, we use data from 2,797 seed samples collected from 1,380 households. The average number of samples collected per household was 1.8. More samples were collected from rice farmers in Odisha because they produce more varieties in kharif 2015.

3. DNA Fingerprinting - Method

The collected seeds were sent to a private company called SciGenom (http://www.scigenom.com/) located in Chennai, India, and DNA gynotyping was conducted by using Illumina Infinium 6K SNP chips (Illumina Infinium 6K SNP - http://gsl.irri.org/services/infinium-6k). From 6K data pointes, about 4K data points were selected for identifications. This suggests that 100% match indicates that only less than 20 SNP points are difference between two samples. It is rare but is still possible for two different samples to share more than 3,980 data points. It is possible for two closely related varieties, such as Swarna and Swarna-Sub1. Therefore, we also check for availability of SUB1 QTL markers to identify Sub1 varieties.

4. Results

4.1 Area estimated based on a pooled farmer surveys

Based on the farmer survey, we estimated areas under different rice varieties based on the farmer variety identification. The results indicate that the most popular variety in eastern India was Swarna (4.2 million ha – 29%), followed by Mahsuri (1.3 million ha, 8.9%), Pooja (1.0 million ha, 6.6%), and Lalat (0.9 million ha, 6.0%). The estimated area under Swarna-Sub1 was 0.4 million ha (2.7%). Note that, through the DNA gynotyping, we can only identify modern varieties that we have breeder seeds (we do not have reference data for traditional and hybrid rice varieties). According to the farmer identification, less than 73% of the total areas is under modern varieties.

4.2DNA Fingerprinting Results

Out of the 2,797 seed samples, we identified 650 samples (23.2%) with breeder seeds. We used 96% match as a cut-off point for all varieties. Because the results are still preliminary and can have commercial implications, we use pseudo names for variety names of breeder seeds, except for a few varieties such as Swarna and Sabhagidhan.

In Table 1, we identified 241 samples (8.6%) as Swarna, 92 samples (3.3%) as Variety 1, and 74 samples (2.6%) as Variety 2. These varieties are called by different names. Swarna, for example, called as Swarna (122 samples), Mahsuri (13), and others. Out of 92 samples of Variety 1, 41 samples were mistaken as MTU1010, but 10 samples were called Lalat, although Lalat is a common name used for Variety 2: 39 sample out of 74 sampled identified as Variety 2 was called Lalat. Well known variety names, such as Swarna, Lalat, and Puja, are used for different rice varieties.

Table 1. Major Varieties Identified by the 6K SNP analysis

Variety	Number of seed samples identified [A]	Variety names commonly used by farmers
	(A)	(C)
	Number (%)	
Swarna	241 (8.6)	Swarna (122), Mahsuri (13), Niranjan (4), Moti Gold (2), others
Variety 1	92 (3.3)	MTU1010 (41), Lalat (10), Swarna (4), Moti (1), others
Variety 2	74 (2.6)	Lalat (39), MTU1001 (4), Swarna (2), Puja (2), others
Variety 3	74 (2.6)	Lalat (48), Puja (2), MTU1001 (2), others
Variety 4	38 (1.4)	Moti (8), others
Variety 5	28 (1,0)	Swarna (16), Ranjit (2), Puja (1), others
Variety 6	22 (0.8)	MTU1010 (16), Lalat (2), others
Variety 7	16 (0.6)	Sarju 52 (14), others
Sabhagi Dhan	15 (0.5)	Sahbhagi Dhan (4), Swarna (2), others
…		
Not identified	2,147 (76.8)	
Total	2,797 (100)	

Note: [A] A cut-off point of 96% is used (accepted if 96% or higher).

By farmers, 382 seed samples were called Swarna. Among them, only 32% were identified correctly named as Swarna, and the rest were identified wrongly as Swarna (False-positive, Type I error). In addition, 119 samples were called otherwise and were identified as Swarna (False negative – Type II error). Other samples have been matched with breeder seeds, but we need to conduct additional analysis to cross-check if the names given by farmers were correct.

5. Caveats and plan for additional analyses

- Breeder seeds used for reference need to be examined. Which breeder seed should be treated as “true”?
- We are in the process of identifying factors associated with the correct identification of varieties. The factors examined include main seed source, household characteristics, and locations.

DNA Gynotyping for Assessing Variety Area Estimates based on Farmer Identification: Case of Rice in Bangladesh

by
T. Yamano, M.L. Malabayabas, M.A. Habib, S.K. Das, Z. Huelgas, G. Carino, T. Kretzschmar, and others (IRRI)

1. Introduction

Based on farmers' identification of variety names, areas under different rice varieties have been estimated by previous studies. Farmers' knowledge on rice variety names, however, is considered unreliable in developing countries where farmers obtain seeds from various sources. Only a small portion of farmers buy seeds in certified packages. Even certified seeds could be incorrectly labeled or adulterated. In Bangladesh, rice variety names are more confused than in other countries, because many Indian rice varieties have been brought to the country unofficially. This situation makes it difficult to track diffusion of new rice varieties, such as submergence-tolerant rice varieties. To mitigate flood damages on rice production, two submergence-tolerant rice varieties, called BR11-Sub1 and Swarna-Sub1, have been developed and distributed in Bangladesh since 2010.[1] The submergence-tolerant rice varieties have a single major quantitative trait locus (QTL) responsible for submergence tolerance, named Sub1 QTL, allowing rice varieties to withstand up to 14 days of complete submergence. To assess the area estimates based on farmers' identification of variety names, we collected 1,289 rice seed samples from 554 farmers in 2014 and 2015. The seed samples were collected from a pooled samples of 3,000 farmers who were interviewed either in 2014 or 2015. The gynotyping of the farmer and breeder seed samples was conducted by using Illumina Infinium 6K SNP chips (Illumina Infinium 6K SNP - http://gsl.irri.org/services/infinium-6k).

2. Data

For the surveys, we randomly selected 16 districts out of 57 districts in Bangladesh, after excluding remote districts where rice is not grown. In the 16 districts, we randomly selected 75 *thanas* out of 117 in the districts. Two villages were randomly selected from each sample thana, and ten farmers were randomly selected in each village (see Appendix Figure 1). One of the two villages in each thana was selected for seed selection, and randomly selected four out of 10 sample households were selected for seed collection. Finally, from each of the selected households, we listed all rice varieties that they produced in the last aman season, and collected aman rice seeds/grains up to four varieties.

3. DNA Fingerprinting - Method

The collected seeds were kept at the IRRI office in Dhaka and germinated at once in early 2016. When rice plants were grown at knee high, leaf samples were taken and small pieces were packed in plastic tubes with Silica gel to keep moisture low. Boxes of plastic tubes with leaf samples were sent to the Genotyping Services Laboratory (GSL) located at the International Rice Research Institute (IRRI) Head Quarters in the Philippines. Gynotyping was conducted by using Illumina Infinium 6K SNP chips. From 6K data pointes, about 4K data points were selected for identifications. This suggests that 100% match indicates that only less than 20 SNP points are difference between two samples. It is rare but still possible for two different samples to share more

[1] BR11-Sub1 is called BRRI dhan 52, and Swarna-Sub1 is called BRRI dhan51 in Bangladesh.

than 3980 data points, especially for closely related varieties, such as Swarna and Swarna-Sub1. Therefore, we also check for availability of SUB1 QTL markers to identify submergence-tolerant rice varieties with the Sub1 QTL.

4. Results

4.1 Area estimated based on a pooled farmer surveys

Based on the pooled farmer surveys of 3,000 farmers, we estimated areas under different rice varieties using the farmer variety identification. The results indicate that the most popular variety in Bangladesh is an Indian variety called Swarna (1 million ha – 21%), followed by BR11 (276'000ha, 5.9%), Binadhan 7 (240,400ha, 5.1%), and Sadamota 206,450ha, 4.4%). The estimated area under BR11-Sub1 (BRRI dhan52) was 51,750ha (1.1%), and that of Swarna-Sub1 (BRRI dhan51) was 37,750ha (0.8%). Thus, the submergence-tolerant rice varieties combined covered 89,500 ha (1.9%). Note that, through the DNA gynotyping, we can only identify modern varieties that we have breeder seeds (we do not have reference data for traditional and hybrid rice varieties). According to the farmer identification, less than 50% of the total areas is under modern varieties.

4.2DNA Fingerprinting Results

Out of the 1,289 seed samples, we identified 186 samples (14.4%) with breeder seeds. We used 100% match as a cut-off point for all varieties, except for Swarna. For Swarna, we used 96% as a cut-off point because we used 3K SNP data as reference for Swarna, and none of the seeds samples is expected to have 100% match with the 3K SNP data of Swarna. On other varieties, we found: BR11 (38 samples), BR11-Sub1 (23), BR23 (17), BR10 (13), and Swarna-Sub1 (5). On the submergence-tolerant rice varieties, we found that the share of the submergence-tolerant rice varieties is 2.2% of the total sample size, and the share is close to the area share estimate of 1.9%. Regarding the Sub1-QTL, it was found on all of the Swarna-Sub1 samples and 78.3% of the BR11-Sub1 samples. It is also found a few samples of BR11, BR10, and BR23. Notably, it is found on about 8% of un-identified samples.

Farmers' variety identification is found poor. Only 31% of seed samples named Swarna (with some variations in their names) were named correctly. The rest, 69%, were wrongly identified as Swarna (False-positive, Type I error). At the same time, 30 seed samples were named otherwise but identified as Swarna (False negative – Type II error). Regarding other varieties, 30% of BR11 seed samples were correctly identified, and so were 13.8% of BR10. None of the new varieties, such as BR11-Sub1, Swarna-Sub1, and Binadhan 7 were correctly identified. Further analysis is required, however, to confirm the preliminary findings provide in this study.

5. Caveats and plan for additional analyses

- Breeder seeds used for reference need to be examined. Which breeder seed should be treated as "true"?
- Rice seeds used by farmers might have gone through mutations or cross pollinations. How should we consider this?
- We merge the seed data with household data and find determinants of correct identification against the main seed source, household characteristics, locations, etc.

DNA fingerprinting for estimating varietal adoption: Summary of Case Studies
Title: Case study on Beans in Zambia

Mywish Maredia (Michigan State University), Byron Reyes, Enid Katungi, Clare Mukankusi, Bodo Raatz and Allan Male (CIAT); Petan Hamazakaza and Kennedy Mui Mui (Zambia Agricultural Research Institute)

1. Overview

- **Geographic scope:** The pilot study was implemented in the Muchinga and Northern provinces.
- **Objectives:** The main objective of the study was to test different approaches of collecting variety-specific adoption data and to validate them against the benchmark of DNA-fingerprinting to determine which method is most accurate and cost-effective in measuring varietal adoption.
- **Current status:** Working paper in progress and a draft article will be prepared for publication.

2. Methodology

Sampling frame, field sampling methods, sample size

The pilot study was designed to take advantage of an already planned bean varietal adoption and impact study by the Zambia Agricultural Research Institute with support from PABRA and CIAT. The Muchinga and Northern provinces were purposively selected because of their importance in bean production (about 70% of the area), and because most of the prior seed dissemination efforts were concentrated in this part of the country. A total of 7 districts were selected based on the importance of the bean crop: 4 districts in the Northern Province and 3 in the Muchinga Province, which together represent 59% of the total bean area in Zambia.

After the districts were selected, a two-stage cluster sample selection method was used. In the first stage, villages were randomly selected from each district according to the proportion of villages within the selected districts in each province. In the second stage, six households were systematically selected within each village. The sample size was determined based on the available budget. Thus, 41 and 26 villages were selected in the Northern and Muchinga provinces, respectively and 6 farmers per village were surveyed to get a total sample size of 402 farm households.

To select the households, a systematic random sampling procedure was followed where a village register list was obtained and served as the sampling frame and each household in this list was numbered sequentially. The first household was selected at random, and the remaining five households were chosen at a fixed interval x=N/6 (N=number of households on the list) until the target was reached.

DNA Genotyping method, logistics of sample collection, post-sample logistics, DNA analysis process

A total of 4 methods were evaluated against the benchmark of DNA fingerprinting (so 5 methods were implemented). For **method A**, farmers were asked to provide the name(s) (method A1) and type (improved vs. local, method A2) of varieties planted in the last completed season. **Method B** involved showing the farmer seed samples representing different varieties and asking him/her to identify the seed sample that matched the varieties grown on their farm. **Method C** consisted of taking photographs of a sample of harvested seeds that was requested from farmers and later using these pictures for varietal identification by a panel of experts. **Method D** consisted ~~for~~ on collecting seed samples of varieties grown by the farmer for later identification by breeders or other bean experts.

To implement the benchmark of DNA fingerprinting (**method V**), 10-15 seed grains from each variety were collected from the farmers during the interview. Small manila envelopes were used to collect each

sample and these were properly labeled (household ID, variety ID and variety name) and sealed for transport. Using paper envelopes is better than using plastic bags. The seeds were germinated by the ZARI bean breeder and with the help of a CIAT technician, leaf tissue samples from young germinated bean plants were collected in 96 well-plate leaf sampling kits and shipped to LGC Genomics lab in U.K. for genotyping. All the farmer samples were genotyped using 66 SNP markers (or assays) that were identified by the research collaborator (a bean genetics expert) from CIAT specifically for this study.

3. Results and Insights

Main results and implications on adoption and impact studies

- The DNA results suggest that 16% of the datapoints corresponded to improved varieties (IVs). All other methods evaluated showed that 4-71% of datapoints (or samples) were identified as IVs, which highlights the variability in estimates of adoption one can get depending on the method used.
- As expected, all methods had type I (i.e., a local variety incorrectly identified as IV) and type II (i.e., an IV incorrectly identified as local) errors. While for method B was more common to observe type I error, for all other methods (A1, A2, C, D) type II error was more common.
- When compared to the benchmark, all methods underestimated adoption. However, while identifying varieties by type was more accurate for methods A (i.e., A2) and B, there were no differences between identifying IVs either by name or type for methods C and D.
- The results suggest that method B (showing seed samples to farmers) was more accurate than all other methods to estimate adoption of IVs.

Insights on process, sampling, DNA genotyping

- Collecting seed samples from farmers had the limitation that some farmers did not have seed to share, as they had consumed/sold it all. Thus, the best time to implement this method would be prior to planting or soon after harvest, which is when farmers have the planting material available.
- Labeling samples along the process is extremely important. Problems can arise when many steps are involved, especially if the labels are changed during this process.
- Expertise in genetics may be needed to plan this type of studies and will be required to interpret DNA results.

Key considerations, questions and challenges: costs, partnerships, capacity, and scaling up potential

- Need to determine what type of material will be used to extract DNA for analysis (e.g., seeds, tissue) as this will have great implications during field activities and storage.
- Cost of DNA analysis was roughly US$34/sample (equivalent to US$0.38/data point) and is expected that this cost will decrease as technology becomes more widely used.
- Farmers may mix varieties after harvest, which complicates the DNA analysis and interpretation.
- Finding good partners is key. Having local capacity to extract DNA could simplify the process (and reduce cost and possibilities of error) as it is easier to mail abroad DNA samples than seeds.
- Some of the methods evaluated were not very accurate so if few methods must be selected for scaling up, showing seed samples to farmers (along with the traditional way of inquiring about the varieties planted) may be a best bet to scale up. With fewer methods, more resources can be used for (the additional samples that will require) DNA extraction.

DNA fingerprinting: Lessons from individual case-studies

NOT FOR CITATION – Contact James.stevenson@fao.org for questions

Maize in Uganda (SPIA / World Bank LSMS-ISA / Diversity Arrays / MSU / NaCCRI / LGC Genomics)

We have three separate field data collection exercises to report, and a further extension in progress in 2016

1 Leaf sampling 2014 – MSU / NaCCRI / LGC Genomics

As part of the planned DTMA (Drought Tolerant Maize in Africa) adoption survey by CIMMYT in three districts in Eastern Province of Uganda, MSU had designed and implemented modules and protocols to test the effectiveness of the following three household-based methods of tracking varietal adoption for maize.

A. Elicitation from farmers by asking him/her the names of varieties planted and some basic questions for each variety planted
B. Asking farmers to show the bag in which maize seeds were obtained and enumerator recording the name of the variety.
C. Enumerator recording observations on phenotypic characteristics by visiting the field. The analyst will later use this information to identify varieties based on the varietal characteristics data.

Field data were collected in June 2014 and leaf tissues from 416 maize fields across 34 villages were collected for DNA analysis. The National Crops Research and Resource Institute (NaCRRI) of NARO served as the 'technical' partner for DNA analysis through their ongoing project with the University of Ghana (under a Gates funded project).

About 50% of samples were lost due to mold development before they reached LGC Genomics. The remaining samples were put in production line for analysis by LGC Genomics in December and we were informed that as they began the process of DNA extraction, they found that almost all the remaining sample plates contained mold. The desiccants had been changed in order to stabilize the leaf material, but leaf material in the tubes had been compacted owing to enumerators not following the recommended number of leaf punches. Thus unfortunately, all the samples collected in June 2014 were lost.

2 Leaf sampling 2015 – SPIA / World Bank LSMS-ISA / NaCCRI / LGC Genomics

and

3 Grain sampling 2015 – SPIA / World Bank LSMS-ISA / Ugandan Bureau of Statistics / NaCCRI / Diversity Arrays

Due to the delays and difficulties experienced during this project, LGC offered a credit to carry out the leaf sample genotyping work for this project before June 2015 (for 34 sample plates x 146 assays) as they had not actually run any genotyping. The alternative was explored with SPIA to piggy back on a planned LSMS experiment on maize in Uganda in 2015, and management of the study was thereupon transferred over from MSU to SPIA in March 2015.

Since March 2015, under SPIA, the context for the study has now shifted to a large methods experiment run by the World Bank LSMS-ISA team and UBOS on estimating maize productivity – the Methodological Experiment on Measuring **Ma**ize **P**roductivity, Varieties, and **S**oil Fertility (**MAPS**). The following three methods for varietal identification were embedded in the design of the experiment:

A. Asking the farmer to identify the variety.
B. Asking the farmer to answer questions related to 15 phenotypic characteristics (using a visual aid), checked against sets of reference responses for each variety using alternative decision rules.
C. Focus group meeting with a number of experts.

These will be benchmarked against two DNA genotyping methods:

D. DNA fingerprinting using 140 SNP markers on samples from maize leaf tissue (using the credit with LGC Genomics under their contract with MSU/NaCCRI)
E. DNA fingerprinting using DArT method of genotyping on samples from maize grain.

Field work for the whole survey took place over three visits to a sample of 900 households (post-planting; crop-cutting; post-harvest) over the period April 2015 – August 2015 in 5 districts in Uganda. For budget reasons, DNA fingerprinting was possible only on a subset of 550 farms in two districts – Iganga and Mayuge. Enumerators from UBOS were recruited

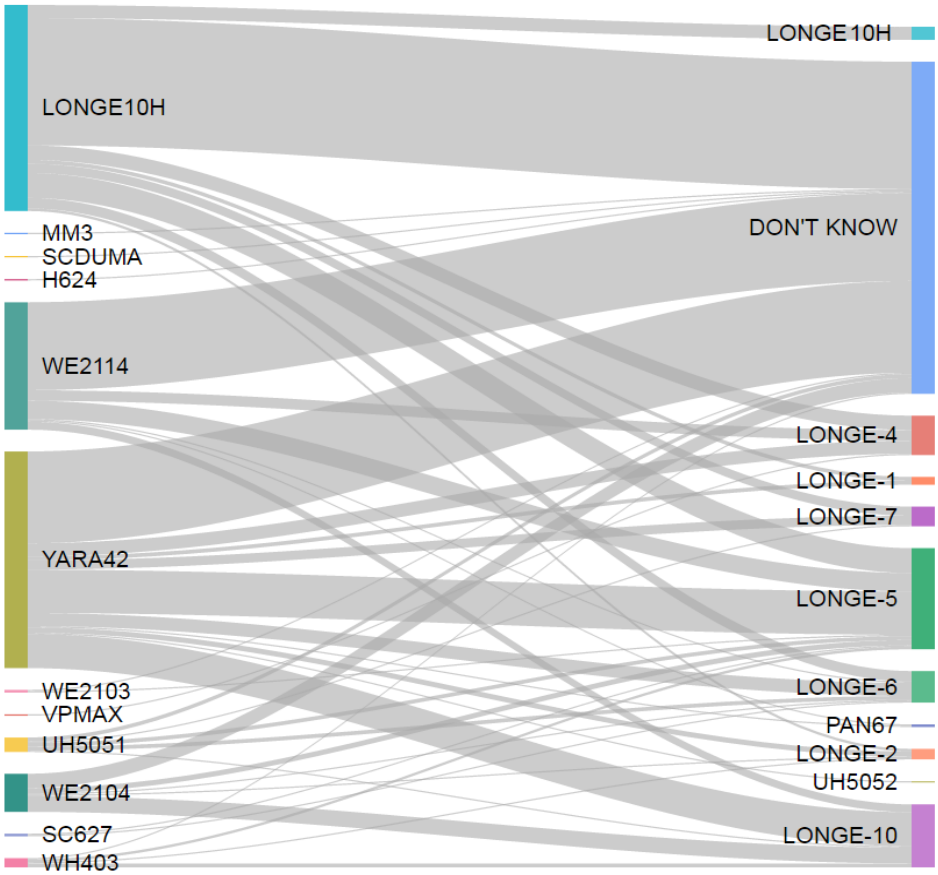
and trained intensively for one month, and survey data collection was facilitated by the use of networked tablets for real-time data management and processing. Leaf samples were collected at the post-planting visit in April and May 2015, from within the quadrant laid down by enumerators for subsequent crop-cutting, using leaf collection kits from LGC ltd. Four leaf punches from an individual maize plant were taken as a single sample, and this was repeated for 12 individual maize plants within the crop cut quadrant. Grain samples subsequently collected from these quadrants in the follow-up crop-cut visit in June and July 2015. The SNP-based genotyping data for the leaf samples was received from LGC in September 2015. Grain samples were processed (dried, ground to flour, labelled) by NACCRI in August and September 2015, and shipped to Diversity Arrays in Australia at the end of October 2015.

Both genotyping methods (D and E) have been successfully applied and the results are currently being compared by colleagues at Diversity Arrays and NaCCRI. Early results show that the SNP-based genotyping used on the leaf samples was insufficiently discriminating among improved varieties – some of the varieties in the reference library appear genetically identical when screen against the relatively low number of SNPs for maize (approx. 140), whereas when a highly quantitative assay used in DArt is applied (examining more than 10,000 alleles), then genetic distance between these two seemingly identical varieties is observed.

The table below shows summary results for **unique identification** for some of the methods applied (which is different from correct identification, which can only be established by method E, DArt). Before any consistency with the correct genotype can be established, these results show that unique identification is only possible only for 47% of farmer responses, and 13% of responses to questions in the morphological protocol. Farmers submit "don't know" responses to the open question of what the variety is that they are growing in 53% of cases. In theory, a set of 11 morphological questions is sufficient to uniquely discriminate among all varieties in the reference library. However, in practice, farmers are clearly unable to respond to the morphological questions with sufficient accuracy to allow for unique identification. Only 60 (13%) of the response sets from farmers interviewed correspond completely to a valid set of responses (as determined by characterization of the reference library). How many of these 60 response sets are consistent with the genotyped identification (as opposed to occurring by chance from the set of multiple choice questions) is currently being investigated.

	Farmer elicitation (A)	Morphological protocol (B)	Grain-based fingerprinting (E)
Don't know	253	417	0
Uniquely identified	224 (47%)	60 (13%)	477 (100%)
Not uniquely identified	0	477	477
Number of varieties	11	14	12

When we consider **correct identification**, only 2% of the sample of 477 farmers are able to correctly identify the variety. The correspondence between the varietal name provided by the farmer (method A) and the correct genotype (method E) is shown in the Sankey diagram below – correct genotypes are on the left, farmer responses on the right. Every farmer sampled (all 477) are growing an improved variety of maize – an astonishing finding. Furthermore, there is deep penetration of the commercial varieties from Western Seed company (WE varieties) and YARA seeds, whereas farmer's expectations are largely confined to the LONGE series released by the government NARO.



GENOTYPE of the primary constituent (Method E) **Farmer elicited variety name** (Method A)

However, while this suggests that improved materials are reaching farmers' fields some way or another, there is also a deeper, less clear picture that emerges when one considers the heterogeneity of the samples. There is considerable genetic heterogeneity within the samples used for the reference library. A good cut-off for acceptable heterogeneity in the reference material is 15%, whereas our mean reference library heterogeneity level is 33%, suggesting that genetic lines have not been well separated in the breeding process, or the seed sampled for the reference library was not pure breeders' seed. Second, there is a very low average purity level of the field samples. Only a minority of field samples are at 80% purity or above, with the average purity at 63%. This simply means that for the average plot, 37% of the genetic material sown is not from the primary improved variety planted in the plot, but is either a deliberate strategy on the part of the farmer to sow seed from more than one variety, or has been introduced in the sown seed for a single variety either deliberately (through counterfeiting) or through error in the seed supply chain.

The MAPS experiment also has first-rate data on agricultural productivity, soil quality, varietal identification, and household characteristics. We can estimate the same simple model for the determinants of productivity twice over with different data: once using the data that are typically collected in surveys, based on farmer testimony; and a second time using objective methods for varietal identification (DNA fingerprinting), yields (crop-cuts), soil quality (soil samples taken and analysed in laboratory). These results will help us understand more about the importance of data quality in context – does improved data quality substantially impact on our understanding of some fundamental issues in impact assessment?

4 Grain sampling 2016

Talip Kilic and John Ilukor are currently leading a second round of the MAPS experiment, which is currently in the field between June and September 2016. We will repeat the genotyping of the grain samples, and hope to dig deeper into the issue of low average purity in a single plot.

DNA fingerprinting for estimating varietal adoption: Summary of Case Studies

Title: *Case study on Adoption and Diffusion of Potato Variety Cooperation 88 in Yunnan (China)*

G. Hareau, W. Pradel, K. Xie, J. Qin, G. Forbes, D. Ellis, N. Barkley **(CIP)**
S. Myrick, J. Alwang, G. Norton, C. Larochelle **(Virginia Tech)**
Canhui Li **(Yunnan Normal University)**

1. Overview

A Province-level representative sample of 615 potato producing households of Yunnan (China) were surveyed between August – September 2015 to estimate adoption and diffusion of potato variety Cooperation 88 (C88). DNA fingerprinting was conducted to a sub-sample of 141 households who declared planting C88, to confirm genetic identity of putative C88 plants. The protocol initially developed by CIP for dried leaf samples was then extended to tuber samples because harvest season was underway when survey was conducted. DNA collection and analysis was conducted by a team of the Yunnan Normal University (YNU) based in Kunming. Out of the 141 samples, 137 (97%) were confirmed to be of true C88 type. CIP genebank leaders provided supervision of methods and confirmed interpretation of results.

2. Methodology

The procedure used by YNU to get DNA fingerprinting varietal confirmation of C88 included the following steps:

a. Declared C88 tuber/leaf sample collection from household survey
b. Visual identification of putative C-88 tuber samples
c. Identification of putative C-88 samples based on cytoplasm genome diversity
d. Identification of putative C-88 samples based on SSR marker analysis of nuclear genome.

Each leaf sample comprised 10-15 leaves randomly collected from at least five plants in the inspected field of farmers and dried with silica in a sealable plastic bag. Each tuber sample comprised 7-10 tubers of putative C-88 plants either randomly collected from harvests or supplied by farmers from recent harvest. Tubers were stored on tables in a dark room at room temperature in plastic bags.

- Based on visual observation, only one tuber sample of the C-88 collected from Zhanyi county of Qujing city was found to be mixed with another red-skin cultivar.
- According to the results of cytoplasmic type detection, one leaf sample collected from Ninglang county of Lijiang city had different cytoplasmic type (T/β type).
- Additionally, the SSR marker-based fingerprinting further clarified three samples showed different SSR genotypes at two loci (STM1049 and STM3032a) in comparison with the other samples and the reference C-88.
- Confirmed that over 97% (137/141) of the fresh samples (leaves and tubers) were C-88.

3. **Results and Insights**
   - **Main results and implications on adoption and impact studies**
   - **Insights on process, sampling, DNA genotyping**
   - **Key considerations, questions and challenges: costs, partnerships, capacity, and scaling up potential**

   - The DNA fingerprinting of samples of putative plants was effective in confirming genetic identity of C88 and gave additional confidence to the adoption estimates. In that respect, the procedure was worth doing.
   - Cost of DNA fingerprinting extraction and analysis was included in the total cost of the adoption survey (~USD100/survey) and it is not possible to determine the exact costs of the DNA fingerprinting procedure alone. Initial estimates of YNU put it at USD 50-70 per sample, but CIP scientists considered it to be too high. CIP genebank costs are around USD 10-20 per sample.
   - Coordination with local partners was challenging because of different reasons:
     - Initial protocol suggested by CIP genebank leaders was different from the suggested Chinese protocols. We could not get access to an English version of the complete Chinese protocol. At the end, they followed a mixed approach that was considered acceptable by CIP molecular biology scientists.
     - Logistics can be challenging. Coordination of HH surveys with leaf/tuber sample extraction was worth doing for cost purposes, but added additional constraints on the window of time in which the field work could be conducted. Several issues arise: larger team in the field and therefore higher supervision and transportation costs, and accessibility to HH can be compromised, amongst other. Trade-offs need to be careful assessed and detailed planning becomes more critical.
   - Matching samples with HH surveys continues to be a critical issue, not difficult to address but can easily fail in the field if supervision fails. New techniques such as barcoding of instruments can add precision and easily introduced with careful planning.
   - In the case of C88 study in Yunnan, results show that visual identification was enough to confirm more than 97% of the sampling. To confirm identity of one single variety in locations where diversity is not large, this might be in many cases a very effective and low cost method and enumerators can be trained to phenotype this particular variety. This could be the first option in these cases. Confirmation of all varieties growing in the field might need more trained personnel.
   - We did not conduct type II error confirmation in the field, although we believe that given the conditions under which C88 is grown in Yunnan this could be small. Could have been done with more time for planning and training of enumerators before going to the field.

Case study on sweet potato in Ethiopia

Frédéric Kosmowski[1], Abiyot Aragaw[2], Andrzej Kilian[3], Alemayehu Ambel[4], John Ilukor[1], Biratu Yigezu[5] and James Stevenson[1]

[1] CGIAR Standing Panel on Impact Assessment, Food and Agriculture Organization of the United Nations, Rome, Italy
[2] International Potato Center (CIP), Hawassa, Ethiopia
[3] Diversity Arrays Technology Pty. Ltd., Yarralumla, ACT, Australia
[4] The World Bank, Washington, DC, USA
[5] Central Statistical Agency of Ethiopia, Addis Ababa, Ethiopia

1. Overview

- ✓ The survey took place in January 2015 in the Wolayita zone (Southern Ethiopia).
- ✓ The objective was to assess the accuracy of three household-based methods for identifying sweet potato varieties using DNA fingerprinting as the benchmark. The methods used were:

  A - Elicitation from farmers with basic questions for the most widely planted variety;
  B - Farmer elicitation on five sweet potato phenotypic attributes by showing a visual-aid protocol
  C - Enumerator recording observations on five sweet potato phenotypic attributes using a visual-aid protocol and visiting the field
  D - DNA fingerprinting using GBS on samples from sweet potato leaves (as the benchmark).

- ✓ The manuscript is under review (PLoS One).

2. Methodology

- ✓ The survey was implemented in five different communities (kebelles) using snowball sampling. We used tablets equipped with the Open Data Kit application to complete the survey questionnaire. The survey focused on the farmer's most widely grown sweet potato variety.
- ✓ Leaf tissues from 259 fields were collected with a unique ID and conserved into a plastic bag.
- ✓ To establish the library, we included all CIP genebank accessions (1004 samples) as well as 19 improved materials collected from the agricultural research centers of Awassa, Adami Tulu and Baco. Six improved materials could not be included in the reference library because they were either not maintained anymore on research stations or were unlikely to be found in the variety collection area.

- ✓ DNA were extracted at ILRI (Addis Ababa) following the CetylTrimethylAmmonium Bromide (CTAB) method and shipped to Australia. From the 259 leaf tissue collected, a total of 231 samples were DNA fingerprinted.

- ✓ For genotyping by sequencing, a combination of a DArT complexity reduction methods and next generation sequencing platforms was used (Kilian *et al.*, 2012; Courtois *et al.*, 2013; Raman *et al.*, 2014; Cruz *et al.*, 2013).

3. Results and Insights

- ✓ Method D (DNA analysis) finds that 63% of the farmers are cultivating improved varieties and 37% are using local varieties;

- ✓ Method A produces almost identical results in the aggregate: across all responses 64% of the cultivars are self-identified as improved, 35% as local and 1% unknown.

- ✓ However, 32% of the actual (DNA-based) improved varieties were identified by farmers as local and 52% of farmers identified a variety as improved when in fact it was a local.

- ✓ Variety names (for both improved and local) given by farmers delivered inconsistent and uncertain varietal identities.

- ✓ Visual-aid protocols employed in methods B and C were more accurate than method A, but still far below the adoption estimates given by the DNA fingerprinting method.

- ✓ Estimating the adoption and impact of improved varieties of sweet potato in this area of Ethiopia with methods based on farmer self-reports is not reliable and indicates a need for wider use of DNA fingerprinting.

- ✓ Insights on process:
    - o Leaf samples were defrost during the trip back to Addis, leading to a lower quality of extracted DNA. Fortunately, these samples could be restored by DAT Australia.
    - o The extraction of DNA from tuber samples was successfully tested. Therefore, using tuber instead of leaves is recommended as an easier method for scaling-up.
    - o Some mistakes were done during the collection of reference materials at agricultural research centers (two highly different varieties where described as genetically similar, which is impossible).
    - o Laboratory performing genotyping should indicate precise instructions for DNA extraction. These guidelines, likely to be crop specific, may be developed at the CGIAR level.

Adoption of improved lentil varieties in Bangladesh: comparison between expert estimates, nationally representative farm household survey and DNA fingerprinting

Organization: ICARDA; Contact: Aden A Aw-Hassan

In general, food legumes get less attention in adoption and impact studies. But, food legumes are major sources of nutrition-protein in Bangladesh, (Datta et.al. 2013). They are also major source of dietary protein for the poor, who cannot afford animal protein and other nutritive and diverse foods (Gowda and Kaul, 1982). Bangladesh and other low income countries of the world with very high rate (26%) of child malnutrition where 41.3% of children are underweight (UNICEF, 2009 and IFPRI Hunger Index, 2012) need improvements in high nutritional intake particularly for children. Another important gap in adoption studies is the accuracy of variety identification- hence precision of adoption. DNA fingerprinting allows comparison between farmer-identified varieties and varietal records and can improve the accuracy of the adoption results.

The study objectives are to:
1. Estimate the extent of adoption of improved lentil varieties (in terms of both area and number of farmers) in Bangladesh using expert opinion elicitation, farm household survey and DNA fingerprinting.
2. Identify factors influencing the adoption of improved lentil varieties using novel econometric techniques.
3. An assessment of congruence/divergence between farmers' preferred attributes and breeders' breeding objectives will be conducted.
4. Measure the effects of lentil varietal adoption on household nutritional outcomes

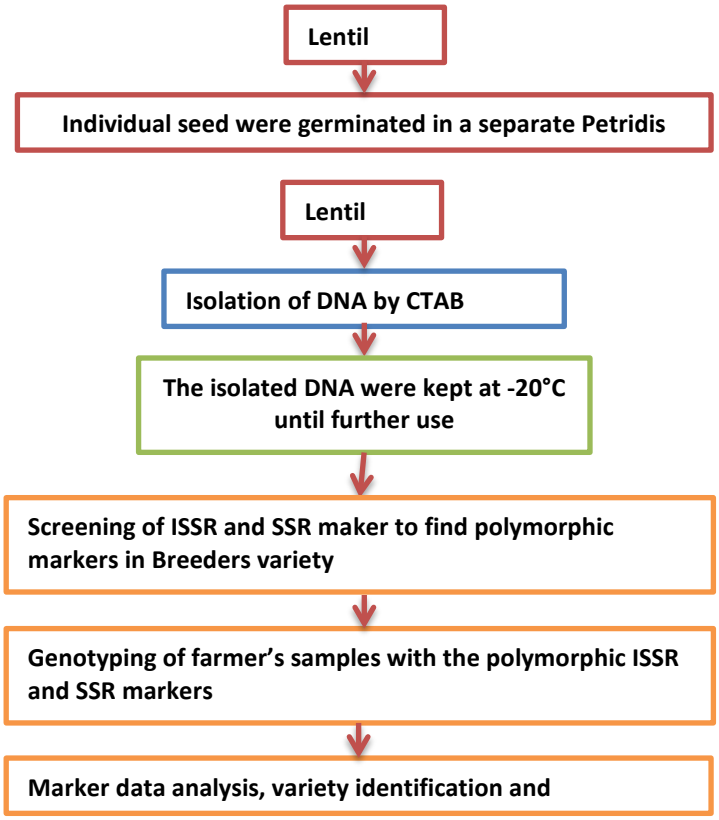
Technology
The technology under investigation is improved lentil varieties. Bangladesh Agricultural research and ICARDA recently released five lentil varieties: BARI M3, BARI M4 (ILL8006), BARI M6 (ILL10848), BARI M7 and BINA M4. These varieties have yield advantages of up to 30%-100% over the local checks. It appears that the higher productivity of lentil varieties has increased the area under lentil increased from about 1.2 million ha in 2007/08 to an average of 1.6 million ha in the period 2010 -2013 showing a 33%. Currently, the five improved varieties are estimated to cover 75% of total lentil area (1.2 million ha) in Bangladesh.

Sampling
Multistage sampling strategy is applied where the administrative units (districts and *mauzas*) is used as clusters. The sample size of 1000 households provided at least 95% confidence and 3% precision levels. The survey covered 10 districts in two agro-ecological zones namely active Ganges floodplain and high Ganges river floodplain.

DNA fingerprinting
We collected seed samples from 4 different sources. We first collected breeder seeds for each of the varieties released in the country were collected from Pulses Research Center (PRC), Bangladesh and Bangladesh Institute of Nuclear Agriculture (BINA) and serve as the reference samples. Secondly we collected two samples from each of the available released varieties from randomly selected seed bags in the Bangladesh Agricultural Development Corporation (BADC) seed storage facilities using grain probes. Thirdly we collected two samples from each of the varieties sold by a random sample of 2 dealers per sub district using grain probes. Fourthly, we collected a handful of (100-200 grains) of lentils from the particular bag or any other container where farmers stored the seed/grain from the particular plot and the seed/grain sample were placed into the plastic bag and sealed it using the zip locker. For each

sample, a unique identification code which is easy to understand and to match with survey questionnaire identification was assigned.

The flow chart of DNA fingerprinting of lentil varieties is illustrated below.



Partial comparison of DNA results of field data and breeders varieties

Sample ID	Identified variety by farmer	Matched with breeder's variety by DNA
Q0029	BARI-4	BARI-3
Q0360	BARI-6	Unmatched
Q0438	BARI-4	BARI-4
Q0456	BARI-3	BARI-4
Q0460	BARI-4	BARI-4
Q0789	BARI-3	BARI-3
Q0820	BARI-3	BARI-3
Q0848	Local	BARI-4
Q0007	BARI-7	BARI-7

The unmatched cases of farmers' identification and DNA identification are marked red. This is only a preliminary result and the full results are under preparation.

DNA fingerprinting for estimating varietal adoption: Summary of Case Studies
Wheat and Lentils in Bihar

Mywish Maredia, David DeYoung and Mukesh Ray (Michigan State University)
Lead on DNA analysis: Mahendar Thudi and Rajeev Varshney (ICRISAT)

Overview

- **Geographic scope:** Bihar state in India
- **Objectives:** The main objective of the study is to validate adoption estimates for the two crop-country-combinations (wheat and lentil in Bihar) obtained through the expert opinion elicitation method by conducting a representative farmer survey and DNA fingerprinting of seed samples collected during these surveys
- **Current status:** Field data collection has been completed. Genotyping of seed samples is currently ongoing at the ICRISAT lab. No DNA fingerprinting results have yet been received.

Methodology

- **Sampling and data collection**

A total of about 3400 households across 340 villages in 35 districts of Bihar were targeted for the survey using a multistage cluster sampling method. Village sample per district were selected randomly based on the wheat and lentil growing area in the district, and farmer samples were selected randomly within the village. Three districts (Kishanganj, Sivhar and Janui) were excluded as they each represent less than 1% of total wheat or lentil area planted in Bihar. Surveys were conducted in November 2015-January 2016, and data correspond to the Rabi planting season. Names of varieties planted in Rabi 2013 and 2014 were also collected to correspond with the year for which expert opinion elicitation method was used. The sample includes a total of 1,000 lentil farmers from across 30 districts and a total of 3,278 wheat farmers from all 35 districts. Survey was conducted using the SurveyBe CAPI program.

- **Sampling for DNA fingerprinting**

For DNA fingerprinting, about 25-50 seed grains for each wheat and lentil variety that farmers had planted in the Rabi season were collected during the interviews. Small plastic ziplock bags were used to collect each sample and these were properly labeled (household ID, variety ID and variety name) and sealed for transport. A total of 3162 packets of seed samples for wheat and 880 packets of seed samples for lentils were collected from the farmers. For about 10-15% of farmers, we were not able to collect any seed samples as all the seeds were planted by the time of the field survey.

Given the restriction of not able to ship any seed or DNA sample out of India, we had limited choice in finding a service provider for genotyping. After exploring the government laboratory and a private sector option, which both turned out to be non-viable (the government option due to lack of high throughput capacity and the private sector option due to high cost), we decided to use the biotechnology lab at ICRISAT for genotyping, based on cost and promised timeframe for completing the task. In the case of lentils all the farmer collected seeds were shipped to ICRISAT for genotyping in February 2016. The process involves first germinating the seeds, then extracting the DNA, and then doing the sequencing. Breeder seed samples for 16 released varieties of lentil were obtained from Bihar and included in the reference library.

The cost of genotyping (which includes DNA extraction) and data analytics is estimated to be $50/sample. Given this high cost, not all the wheat samples are subjected to DNA fingerprinting. About 50% of the wheat seed samples were purposively selected for DNA fingerprinting based on the following criteria: 1) samples from all the 335 villages are represented; 2) wheat samples of all the unique names were selected from each village. 3) if multiple samples of varieties with the same name were collected from a village, then only 1 or 2 samples of that variety name were randomly selected. Based on this rule, a total of about 1,686 samples were selected for wheat varietal identification

using GBS method. Breeder seeds of 64 released wheat varieties were obtained from different wheat breeding program with the help of CIMMYT and included in the reference library.

Interim results and Insights on varietal identification

Adoption of wheat and lentil varieties: Based on expert opinion, area planted to improved wheat varieties in Bihar is estimated to be 100%, with the top 15 varieties by area planted is as indicated in Table 1. Compared to this, area planted to improved varieties as reported by farmers is only 61%. In the case of lentil also there is a wide discrepancy in the estimates of improved varieties as a group reported by experts (42.7%) and as reported by farmers in the farm survey (9.4%). Comparison of results of expert opinion elicitation and farmer survey (table 1) indicate wide discrepancies in the estimates of varietal adoption by names as well.

Table 1. Adoption of top 15 wheat and lentil varieties by name: Comparison of expert elicitation and farmer elicitation method in Bihar, India

Expert opinion % of wheat area		Farmer survey % of wheat area	
PBW343	30.0%	UP262	24.9%
HD2733	21.8%	PBW343	17.9%
PBW502	14.3%	LOK-1	8.6%
PBW373	12.1%	PBW502	8.0%
HD2967	2.6%	PBW154	5.5%
UP262	2.1%	NL	5.2%
PBW550	1.9%	HUW234	4.9%
WH711	1.8%	Kedar	4.5%
HD2824	1.3%	HI1563	3.2%
HUW234	1.1%	HD2967	1.9%
HI1563	0.8%	HUW55	1.9%
K307	0.4%	PBW373	1.7%
PBW154	0.4%	Sher-e-Punjab	1.1%
GANGA HD2643	0.4%	HD2733	0.7%
LOK-1	0.1%	AINAL	0.7%

Expert opinion % of lentil area		Farmer survey % of lentil area	
Arun (PL 77-12)	16.2%	Titua	36.3%
Malika (K-75)	7.7%	Choti Masur	29.9%
HUL-57	6.6%	Dehati	12.7%
PL 639	3.8%	Arun (PL 77-12)	7.9%
NDL-1	2.8%	Meetha	4.5%
Pant L 406	1.6%	Pathal	2.3%
Other improved	1.5%	Pahalwan	1.7%
KLS-218	1.4%	Shipra	1.5%
IPL-406	0.3%	Desi	1.3%
PL 8	0.3%	Kanar	0.4%
IPL-81 (Noori)	0.2%	Savitri	0.3%
WBL-77	0.1%	Bada Masur	0.2%
PL 6	0.0%	Jaya	0.1%
PL 7	0.0%	Shipli	0.1%
Local varieties	57.3%	DKL-50	0.1%

Lessons and challenges for scaling up

- Reference library – It was very challenging to access breeders' seed from the Indian research institutes (ICAR) for all the released varieties to establish the reference library. Finally, with the help of CIMMYT scientists in India, seed samples of at least 64 (out of more than 90 varieties that appear on the Bihar list of wheat varieties) varieties were obtained through their network of wheat breeders. Similar challenges were also encountered for lentils, and finally we were able to access seeds of some released varieties through ICRISAT collaborators.
- Logistics of collecting, tracking, storing and transporting the samples from farmers' fields to a lab facility remains a challenging task for a survey of such a large magnitude. Matching samples with HH surveys continues to be a critical issue. This is not difficult to address but can easily fail in the field if supervision fails.
- The restrictions by Indian government on the shipment of seeds or DNA samples out of the country is a huge constraint in doing this type of studies in India. Limited option for doing high throughput genotyping within India has implications on the cost of DNA fingerprinting. Among all the case studies conducted by MSU (6 so far), the cost of DNA fingerprinting estimated for this study is on the higher end (~$50/sample).
- Potential ways to reduce the cost and to make the logistics more manageable for a large representative surveys would be to use DNA fingerprinting as a method of validation on a random sub-sample of households rather than all the households. This was essentially what was done for wheat in this study with only about 50% of collected samples were fingerprinted. Whether this cost-cutting strategy is effective and gives consistent results still needs to be seen.