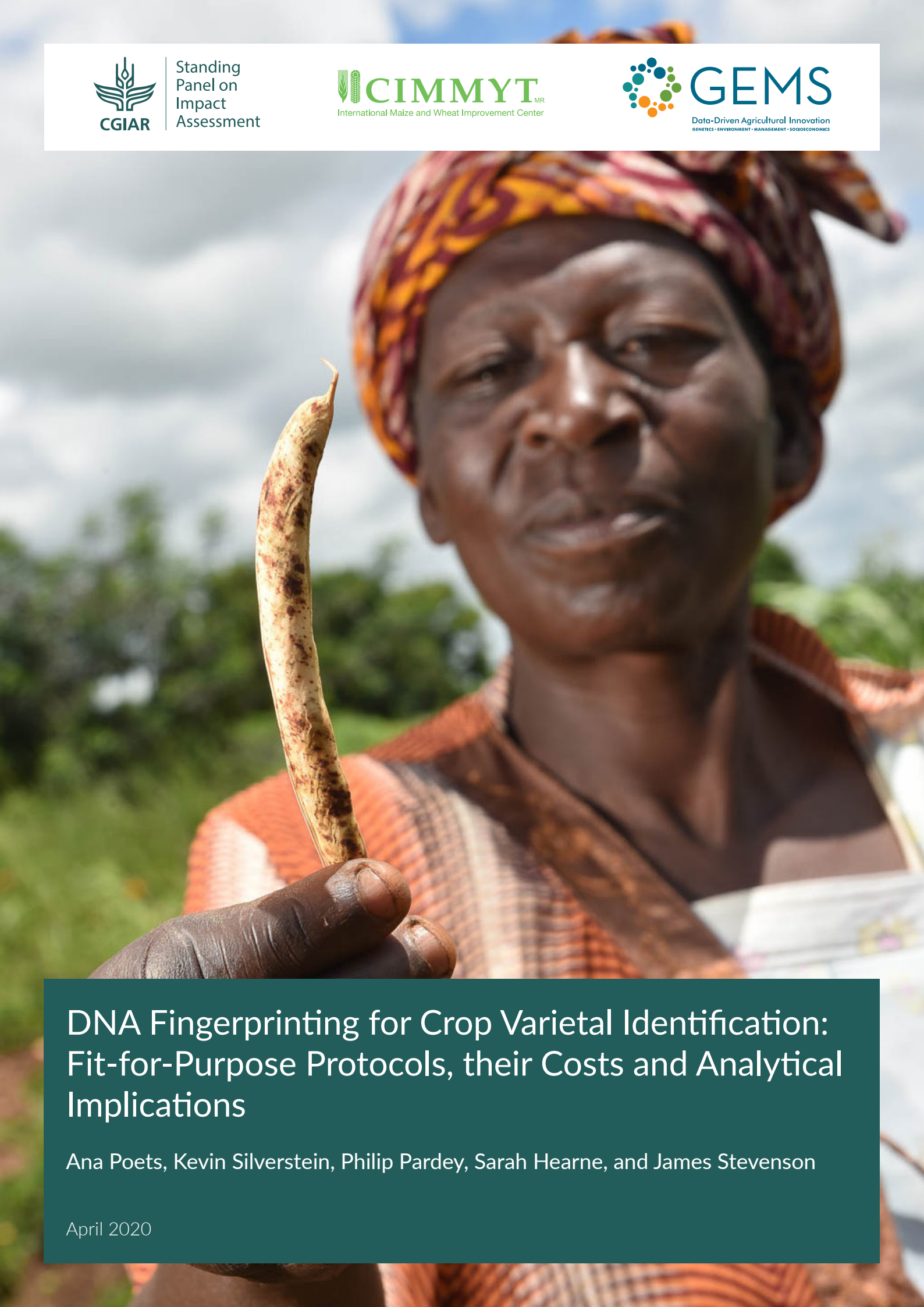
CGIAR Standing Panel on Impact Assessment

CIMMYT International Maize and Wheat Improvement Center

GEMS Data-Driven Agricultural Innovation
GENETICS · ENVIRONMENT · MANAGEMENT · SOCIOECONOMICS



DNA Fingerprinting for Crop Varietal Identification: Fit-for-Purpose Protocols, their Costs and Analytical Implications

Ana Poets, Kevin Silverstein, Philip Pardey, Sarah Hearne, and James Stevenson

April 2020

This document was prepared with support from the Standing Panel on Impact Assessment (SPIA) of CGIAR and the GEMS agroinformatics initiative at the University of Minnesota (UMN). Poets was an agroinformatics analyst with GEMS and is presently genetics project lead at Syngenta, Slater, IA; Silverstein is scientific lead at the Minnesota Supercomputing Institute and operations lead of GEMS; Pardey is director of global research strategy for UMN's College of Food, Agricultural and Natural Resource Sciences (CFANS) and co-director of GEMS; Hearne is a Principal Scientist at the International Maize and Wheat Improvement Center (CIMMYT); and Stevenson is a senior research fellow at the CGIAR Standing Panel on Impact Assessment. This paper solely reflects the opinions and findings of the authors.

The content of this report benefited greatly from discussions with participants at the meeting "Scaling Best Practice on Integrating DNA Fingerprinting of Crops into Large-Scale Household Surveys", convened by the CGIAR Standing Panel on Impact Assessment (SPIA) and the University of Minnesota and hosted by the Bill and Melinda Gates Foundation in Seattle, WA, on January 18-19, 2018. We are particularly grateful for detailed comments on prior drafts from Andrzej Kilian. We also thank Maxwell Mkondiwa, along with participants in the Seattle workshop—especially Ismail Rabbi (IITA), Marianne Banziger (CIMMYT), Augusto Becerra (CIAT), and Mywish Maredia (Michigan State University)—for valuable input in the preparation of this report, and Heidi Fritschel for outstanding editorial revisions to the manuscript.

Cover image: ©2016CIAT/NeilPalmer
Design and layout: Macaroni Bros

Ana Poets, Kevin Silverstein,
Philip Pardey, Sarah Hearne,
and James Stevenson

DNA Fingerprinting for Crop Varietal Identification: Fit-for-Purpose Protocols, their Costs and Analytical Implications

April 2020

TABLE OF CONTENTS

FIGURES

TABLES

BOXES

FOREWORD

In an effort to understand the diffusion of improved agricultural technologies in the developing world, researchers have long sought to measure farmers' adoption of improved crop varieties. A number of different approaches have been used: eliciting the opinions of informed experts, collecting self-reported data from farmers as part of household surveys, or imputing varietal areas from data on seed sales. Little information has been available, however, about the validity of these approaches, which undoubtedly lend themselves to various biases and multiple sources of measurement error.

In the past 15 years, as a result of several technological breakthroughs in the laboratory, the cost of genotyping has fallen significantly. It is now possible to mainstream the use of DNA fingerprinting for estimating varietal adoption. In experimental studies, data on varietal adoption can be collected using all three methods (expert opinion, farmer self-reports, and DNA fingerprinting of tissue collected from farmers' fields), allowing us to use the latter as an objective benchmark against which the earlier methods can be judged.

In most experimental studies that have used this benchmarking, we are finding significant differences between the estimates from DNA fingerprinting and those established using earlier methods. In some cases the prior estimates underestimated the true extent of adoption, but in many cases older methods overestimated adoption by farmers. However, we are only scratching the surface of the insights we stand to gain from scaling up DNA fingerprinting. The method can be applied to a host of second-order questions, such as varietal turnover, the age of varieties in farmers' fields, and the efficacy of the seed system in providing high-quality seed to farmers or of the extension system in promoting new varieties. To inform this process of scaling up, the CGIAR Standing Panel on Impact Assessment (SPIA) commissioned this report.

Researchers who want to use DNA fingerprinting to analyze the adoption of improved crop varieties in farmers' fields face multiple methodological options. They must make careful decisions to match protocols for sampling and analysis to their specific analytical needs. This document synthesies what we have learned about the state of the art regarding that process. This is a field that is rapidly shifting— the technology is changing, and with it the questions we can ask, thus this document should be seen as a set of best practices as of today.

The interdisciplinary team of authors assembled for this study (from genetics, data science, and economics) used evidence from multiple empirical studies, supplemented with their own research and consultations with experts. Much of the evidence was generated by studies carried out in the context of the five-year SPIA program "Strengthening Impact Assessment in the CGIAR" (SIAC), which ran from 2013 to 2017. Other fingerprinting studies were run independently by individual CGIAR centers, and these too played a significant role in informing the material presented in this report.

The Bill and Melinda Gates Foundation, and program officers Greg Traxler, Marianna Kim, and Richard Caldwell in particular, helped convene discussions of the methodological issues related to DNA fingerprinting, and the foundation provided significant grant funding to SIAC and other fingerprinting studies. Indeed, the foundation offices in Seattle hosted two methodological workshops on DNA fingerprinting—in August 2014 and again in January 2018—and many of the perspectives presented in this document were first debated by participants in those two events. We thank them all.

SPIA is grateful for the work that the author team carried out in producing this report. They have gone well beyond the original vision for this document, taking the initiative to get updated cost estimates and methodological details from alternative providers of genotyping services, as well as exhaustively reviewing the scientific literature. Given the speed at which the technology and commercial landscape for genotyping services is changing, we will likely need to revisit this document in a few short years to update it. The case for doing so will be all the stronger if it is widely used in the interim. We hope that CGIAR researchers and the broader agricultural research community will find this document useful and aspire to contribute the empirical evidence to inform the next edition.

Douglas Gollin
Professor of Development Economics
University of Oxford
SPIA Chair (2012-17) and SPIA Member (2017-19)

GLOSSARY

Allele	One of two or more alternative forms of a particular gene (usually a DNA sequence representation) that arise by natural or human-induced mutation and are found at the same place on a chromosome. Alleles can confer functional changes in a plant phenotype (e.g., plant height) or have a neutral effect. Different alleles can be used to distinguish different varieties of a crop
Allele frequency	Measure of the relative frequency of a particular allele at a specific location in the genome in a set of samples. Usually it is reported as a proportion or a percentage of representation of the allele in the germplasm panel evaluated
Allele frequency profile	Distribution of allele frequencies at a given set of genetic markers in a population or sample
Ascertainment bias	Systematic distortion in measuring the true profile of polymorphisms or frequency of a specific allele resulting from the way in which the data were collected or processed
Bulking	Combination of plant material from numerous individual plants to form a representative sample of a farmer's field or plot
DArT	Diversity Arrays Technology which comprises diverse proprietary protocols for optimized targeted genotyping methods
Discovery panel	The set of samples used to identify an initial set of single nucleotide polymorphisms (SNPs) in a population of interest
DNA fingerprinting	Process of using DNA information to characterize the genetic material planted in farmers' fields
DNA barcode	Short DNA sequence fragment physically attached to sample DNA that uniquely identifies it as a particular sample even when mixed with other sample DNA
GBS	Genotyping by sequencing
Haplotype	Chromosomal segments that are inherited together from a single parent
Heterozygosity	The state of having multiple versions of a variant (allele) at the same genetic position
Heterogeneity	The state of having multiple varieties of a particular crop growing in the same field or plot. Alternatively, the state of a bulk sample with two or more forms of alleles at one position in the genome
Homozygosity	The state of having only one variant (allele) at a defined genetic position.
Nucleotide	A compound of nucleoside linked to a phosphate group. Five nucleotides (the base pairs adenine, cytosine, guanine, and thymine/uracil) form the basic structural unit of DNA
Polymerase chain reaction	A technique used in molecular biology to generate thousands to millions of copies of a particular segment of DNA

Read depth coverage	The number of unique DNA sequence reads that align to a given position in the reference genome. Deep sequencing refers to the general concept of aiming for a large number of unique reads for each region of a sequence
Restriction enzyme	Enzyme produced primarily by bacteria. This enzyme has the property of cleaving DNA molecules at or near a specific sequence of bases
SNP	Single nucleotide polymorphism is a variation of a single nucleotide (base pair) in the genome that occurs at a specific position in the genome; also referred to as a variant
TAS	Targeted amplicon-based sequencing method
WGS	Whole genome sequencing method

HIGHLIGHTS

Traditionally, varietal identification relied heavily on morphological differences that farmers and experts could use to classify a sample as a specific variety. This method raised a number of challenges. Among certain varieties, there are few morphological differences. Where visible defining traits exist, their expression may be conditioned on crop management or environmental factors. Non-experts typically lack clear knowledge of these differences or experience of all the varieties available. DNA fingerprinting, therefore, is regarded as a more objective and, ostensibly, less error-prone method of identifying plant varieties than traditional methods.

DNA fingerprinting is the process of using fundamental genome coding, rather than morphological characteristics, to identify a variety. DNA is extracted from a field sample and compared with a reference library—that is, a set of genetic profiles from known improved and unimproved varieties. The sample is then classified by the closest match based on its genetic similarity to varieties in the reference library within a defined tolerance.

There are biological, technical, and practical trade-offs associated with genotyping different crops; therefore a one-size-fits-all approach to DNA fingerprinting to assess varietal use in farmers' fields is neither cost-effective nor practical. In this report we describe a range of alternative approaches, and their respective trade-offs, for varietal identification in a specific crop given practical constraints on external conditions.

The DNA fingerprinting strategies considered in this report involve protocols designed to (1) differentiate between improved and unimproved varieties, or (2) identify the specific varieties included in samples taken from farmers' fields. Evaluating the uptake of new varieties using DNA fingerprinting methods is a multidisciplinary undertaking that involves significant technical choices concerning sampling and sequencing strategies. These choices have direct implications for what can and cannot be assessed regarding varietal use in farmers' fields, and the accuracy of those assessments. The technical details of these sampling (plus sample handling) and sequencing procedures are laid out in this report in a rigorous, but hopefully accessible, fashion, along with the cost and analytical implications of choosing among these alternative procedures. A companion handbook tailored to social scientists wanting to incorporate these methods in field surveys is under preparation.

Sample Size

Based on theoretical and empirical evidence, the recommended sample size is 30 or more individual plants from each field for cross-pollinated crops (such as maize) and self-fertilized crops (such as wheat, barley, oats, and rice) for which an allele frequency profile (from a bulked collection) will be used as a reference library. There is not strong empirical evidence on the appropriate sample size for crops with reference libraries formed using individual (not bulked) plants. Nonetheless, the limited evidence available suggests that a similar sample size is sufficient for most improved varieties.

Heterogeneity in Fields versus Heterozygosity in Plants

Heterogeneity in the field refers to the presence of multiple varieties within a field. The varieties might be separated by plot or intermingled in the same plot (farmers might or might not be aware of the latter case). Heterozygosity is the presence of different versions of an allele (called a variant) within a single individual plant. Heterozygotes are common in hybrids and open-pollinated crops, existing naturally in a single released variety. Additionally, in some cases heterozygotes can result from a farmer's more recent intentional cross between varieties or from pollen contamination within the field from one variety to another.

In an ideal world, to retain the ability to distinguish between heterogeneous fields and heterozygous and homozygous plant varieties, the plants or seeds gathered from each field should be sampled and stored individually rather than being bulked. Practical fac-

tors, however, including logistical and cost considerations, may render this approach infeasible. In such cases it can be more appropriate to sample bulked plants and interpret the results with an appreciation of the analytic limitations of using bulked samples.

Control Samples

It is important to deploy control mechanisms during the fingerprinting process that can (1) reveal some of the problems that may arise during this process, and (2) enable researchers to correct or account for these problems during data analysis. Carefully tracking genetic material through the analytical process is essential to preventing the introduction of errors.

Genotyping Technologies

In this report we present seven genotyping technologies that vary in cost at the time of writing from $7 to $2,500 per sample. The cost varies as a function of the precision of each genotype or genetic profile developed and the amount of the genome covered. These two variables influence both the sensitivity of the data to differentiate varieties and the long-term analytical options. However, this influence is dependent on the species, with predominantly homozygous crops such as wheat requiring relatively few markers to distinguish among different varieties compared with heterozygous maize. Homozygous material produces more precise fingerprints even if the genome is comparatively large.

While several genotyping technologies are currently available, at the time of writing we recommend target amplicon-based sequencing and DArT optimized targeted genotyping technologies for routine analysis of variety identification.

We also recommend using a sequence-and-discover approach for the development of the reference library. Currently there are five technologies of this kind: whole genome sequencing, exome capture, genotyping by sequencing (GBS, Keygene), tunable GBS (tGBS, Data2Bio, Ott *et al.* 2017) and optimized targeted genotyping provided by Diversity Arrays Technologies (DArTseq, DArTcap, DArTag, DArTmp) (DArT 2019); these are presented in order from the technology that yields the most data points to the one that yields the fewest, though all technologies listed yield a robust number of data points.

Each genotyping technology detects different portions of the genome; some of the genetic variants can be captured by one or more platforms, whereas others are private to a particular platform. Thus it is a challenge to know whether the same variant has been found by two different platforms unless both refer to a unique identifier (ID). One possible way to assign an ID is to identify the position of such a variant in a genome reference (available for most major crops) and assign that position as an ID. However, since genome references undergo continual change, it is easier to create a cross-reference table of positions in different genome reference versions, and use these position cross-references to merge data sets developed at different times and places and by different technologies. This practice prevents the need to map each genetic variant ever detected by a particular sequencing technology to a new genome reference every time multiple data sets need to be merged.

Reference Library Composition

The ideal reference library for varietal identification should contain all the possible varieties from private and public breeding efforts, including improved and landrace varieties likely to be grown by farmers in the sampled area. It should not be limited by assumptions about what farmers might be growing. Given that the genetic profiles of varieties are determined by the seed or clones chosen to represent each variety, it is necessary to confirm the purity of the seeds or clones used to develop the reference library.

Developing and maintaining representative reference libraries are costly and time-consuming. Libraries are also not one-off ventures; they need to be maintained, updated over time, and made accessible in ways that respect any relevant intellectual property.

Field Sample Storage

Saving field samples—along with their associated data and metadata—at least for a period of time, serves to "future proof" past work and enable data interoperability with studies yet to be taken using new DNA fingerprinting strategies yet to be developed or fully deployed. The saving and reusing of samples must be conducted in alignment with relevant national and international laws and regulations.

1 INTRODUCTION

In a world where suitable new agricultural land is ever scarcer and cropping systems are increasingly challenged by changing climate regimes, limited water, and other natural factors of production, ensuring food security and access to healthy, nutritious, and safe foods remains a major objective for private and public breeding programs. The World Bank, the Bill and Melinda Gates Foundation, the U.S. Agency for International Development (USAID), and other institutions have long supported the efforts of national and international (e.g., CGIAR) research programs to develop and deploy improved varieties that are better suited to the challenging conditions facing crop farmers throughout the developing world. The success of these breeding programs and broader partnerships can be assessed based on the extent to which improved crop varieties are disseminated and adopted.

Much of the existing evidence on the adoption of improved crop varieties is derived from seed sales data or obtained from socioeconomic household surveys and research impact assessment studies that rely on self-reported data from farmers (Mkondiwa *et al.*, 2020). Estimates of the extent of the uptake of improved varieties in terms of area, sales, or production have also relied heavily on expert opinion, often obtained from breeders and extension services (Rabbi *et al.* 2015). Several factors make such methods error-prone. Estimates based on expert or farmer recall data can be affected by inconsistencies in the names assigned to varieties (e.g., released name versus local naming variants), farmers' inability to identify varieties by name (including farmers' lack of understanding of the morphological differences between improved modern varieties and unimproved traditional varieties), or the inability to differentiate between particular hybrids or varietal types, as noted by Floro *et al.* (2017) and Maredia *et al.* (2016). Morphology-based variety assessment can also be influenced by environmental conditions (which can affect the expression of various morphological details), the plant's developmental stage when the variety is assessed, and the limited number of morphological characteristics that actually differentiate closely related varieties, as shown by Kosmowski *et al.* (2016).

DNA fingerprinting is deemed a more objective and, ostensibly, less error-prone approach to identifying varieties than traditional methods, especially when there are few morphological differences among certain varieties. In this approach, genetic material is extracted from a field sample and compared with a reference library—that is, a set of genetic profiles from known improved and unimproved varieties. The sample is then classified as a particular variety based on its genetic similarity to or difference from the varieties in the reference library within a defined tolerance.

The use of genetic information to distinguish varieties is not new in the field of plant breeding (e.g., Chakravarthi and Naravaneni 2006; Warburton *et al.* 2010; Zhu *et al.* 2011). DNA fingerprinting has been used to assess varietal purity (e.g., Smith and Register 1998; Habernicht and Blake 1999) and to enforce intellectual property rights (Bhat 2008). In the past few years, a handful of social science studies have used DNA fingerprinting in studies of varietal adoption (e.g., Wossen *et al.* 2019). Using DNA fingerprinting data as a benchmark, a number of recent studies have assessed the accuracy of traditional varietal identification methods; they have revealed

a high rate of discrepancy (20–30 percent) when self-reported data from farmers were used to differentiate among varieties, and a similar range of discrepancy when phenotypic traits were used as the differentiating factor (e.g., Kosmowski *et al.* 2016; Maredia *et al.* 2016; Rabbi *et al.* 2015). Table 1 lists previous DNA fingerprinting studies and summarizes pertinent technical details for each study.

Table 1: Summary of studies using DNA fingerprinting for varietal identification

Source	Kosmowski *et al.* 2016	Maredia *et al.* 2016; Rabbi *et al.* 2015	Maredia *et al.* 2016; Rabbi *et al.* 2015	Floro *et al.* 2017
Year	2016	2015–2016	2015–2016	2017
Country	Ethiopia	Ghana	Zambia	Colombia
Organism	Sweet potato (*Ipomoea batatas*)	Cassava (*Manihot esculenta* Cranz)	Bean (*Phaseolus vulgaris*)	Cassava (*Manihot esculenta* Cranz)
Ploidy number	Hexaploid (2N=6x=90)	Diploid (2N=2x =36)	Diploid (2N=2x=22)	Diploid (2N=2x =36)
Genomesize	2.37 Gbp	772 Mbp	587 Mbp	772 Mbp
Propagation/ mating system	Sprout or vine cuttings	Clonally/ outcrossing	Seed/selfing	Clonally/ outcrossing
Tissue collected	Leaf (individual)	Apical leaf (individual)	Seed (individual)	Stems (individual)
Tissue for DNA extraction	Leaf	Leaf	Young leaf	Stems
Sampling strategy	Snowball	1 sample/ variety	10–15 seeds/ variety	1 stem/ plant
Total sample size	231	914	855	436
Genotyping technology	DArTseq	GBS	KASP (SNP)	SNPY-chip (yuca chip)
No. markers per sample	Not reported	56,849	66	93
Reference library composition	1,004 samples (CIP genebank + 19 improved)	18 released varieties + 46 landraces	11 released varieties + 2 landraces + 25 farmer-collected samples + 698 East/ Southern Africa	150 LAC landraces
Markers/ reference sample	Not reported	56,849	776	93

Source	Ilukor *et al.* n.d.	Kilic (n.d.)	Wossen *et al.* 2018	Yirga *et al.* 2016	Yirga *et al.* 2016
Year	2017		2018	2016	2016
Country	Malawi	Uganda	Nigeria	Ethiopia	Ethiopia
Organism	Cassava (*Manihot esculenta* Cranz)	Maize (*Zea mays*)	Cassava (*Manihot esculenta* Cranz)	Maize (*Zea mays*)	Wheat (*Triticum aestivum*)
Ploidy number	Diploid (2N=2x=36)	Diploid (2N=10x=20)	Diploid (2N=2x=36)	Diploid (2N=10x=20)	Hexaploid (2N=6x=42)
Genomesize	772 Mbp	2.5 Gbp	772 Mbp	2.5 Gbp	17 Gbp
Propagation/ mating system	Clonally/ outcrossing	Seed/ outcrossing	Clonally/ outcrossing	Seed/ outcrossing	Seed/selfing
Tissue collected	Leaf (individual)	Seed (bulk)	Leaf (individual)	Seed (bulk)	Seed (bulk)
Tissue for DNA extraction	Leaf	Bulked seed/ crop cut	Leaf	Bulked seed/ crop cut	Bulked seed/ crop cut
Sampling strategy	3 newly expanded leaf samples	5 2×2 crop cuts/farm	1 leaf/ variety/ plot	200 grains/ crop cut/ farm	200 grains/ crop cut/ farm
Total sample size	1,174	510	7,565	472	393
Genotyping technology	DArTseq	DArTseq	GBS	DArTseq	DArTseq
No. markers per sample	Not reported	Not reported	52,899	Not reported	Not reported
Reference library composition		38	3,891 improved varieties	39	75
Markers/ reference sample	Not reported	Not reported	52,899	Not reported	Not reported

Source: Compiled by authors.

Note: LAC = Latin America and the Caribbean.

DNA fingerprinting can be carried out using a number of methods. These methods vary in cost and precision, raising a set of practical questions about the appropriateness of different fingerprinting methods for different research questions. To the best of our knowledge, no published study has yet formally examined both the accuracy and the relative costs of alternative DNA fingerprinting methods for identifying varieties grown in farmers' fields, though a number of groups are working toward this goal.[1]

In practical terms, the accuracy of these methods depends on the whole chain of events from farm field to interpretation of lab results. The logistics involved in these studies are critical at each stage, including the collection of samples; the tracking, processing, and analyzing of DNA; the interpretation of results within the context of the specific research questions and the crop's biology; and the understanding of the seed system. A host of potential problems may arise, such as poor field sampling protocols, mislabe-

[1] For an example of the application of DNA fingerprinting for crop quality assurance, see Meibusch (2013).

ling of samples, contamination of samples, mixing of samples during various stages of the fingerprinting process, degradation of samples (due to improper storage), low-resolution genotyping, the use of DNA reference libraries that underrepresent or misrepresent the varieties (or combinations thereof) that have been sampled in farmers' fields, and inappropriate interpretation of data based on a poor understanding of species biology or the relevant seed system.

There are several DNA fingerprinting approaches, and their relative benefits and costs vary considerably. Nonetheless, because identification of varieties depends on genetic information within the sample rather than on human opinion, DNA fingerprinting should in principle be a more reliable means of identifying varieties than any approach based on data reported by farmers or experts. In particular, DNA fingerprinting should work better at discriminating between the real distribution of unimproved and improved varieties and among morphologically similar improved varieties (assuming a high-quality reference library). Furthermore, as stated earlier, depending on the genotyping platform used, it is possible to elucidate the level of purity of a sample, and the sources of mixture if the sample fails to perfectly match a particular variety in the reference library (e.g., Poets *et al.* 2015; Rabbi *et al.* 2015). These additional characteristics, when combined with the right sampling design, can provide important data on the state of the seed system for a specific crop in a particular country and serve as an invaluable resource when evaluating the causes of discrepancies between perceived and actual variety distribution and adoption.

Reliability (or accuracy) and reproducibility are important features to keep in mind when choosing a fingerprinting strategy. The stringency applied to these factors needs to reflect both the objectives of the work to be conducted (e.g., the goal may be to determine whether farmers are growing improved varieties or to identify what specific varieties farmers are growing) and the resource envelope available for the study (Boxes 1 and 2).[2] The goal of this manual is to offer practical guidance on the technical and practical cost-benefit trade-offs involved in choosing among DNA fingerprinting strategies for varietal identification. The fingerprinting strategies considered in this report involve protocols designed to (1) differentiate between improved and unimproved varieties, (2) identify the specific varieties represented by samples taken from farmers' fields, or (3) assess the purity of varieties sampled in farmers' fields. Although (1) and (2) may seem superficially similar, they actually differ in important ways, and the choice of objective has clear implications for subsequent choices about research design, and we hope to make the implications of these design choices clear to the non-specialist.

BOX 1. COMMON QUESTIONS ABOUT DNA FINGERPRINTING

1. Can DNA fingerprinting discriminate among closely related varieties?
2. Can identity be preserved for hundreds of samples through field collection, DNA extraction, and lab analysis?
3. Can sample contamination and biological degradation be controlled?
4. Are costs per sample low enough to show promise for widespread use in monitoring diffusion?
5. Do diffusion estimates differ from estimates based on surveys of farmers and experts?

Source: Traxler ***et al.*** (2015).

[2] Crop breeders typically structure their breeding trials to reveal the biologically optimal yield of a particular variety, which is rarely the economically optimal yield given the cost of inputs relative to the value of the resulting crop. Similarly, the scientific instinct when genotyping material is to strive for the "best practice" regardless of the cost-benefit trade-offs involved. Here we seek to reveal the nature of the technical versus economic trade-offs involved in practical fingerprinting settings.

BOX 2: TRADE-OFFS—A DNA FINGERPRINTING DECISION TREE

The extent to which information can be derived from the investment made in a DNA fingerprinting project depends on decisions made throughout the entire fingerprinting process, from field sampling to genotyping in the lab. The approach that will generate the largest amount of information involves sampling tissue from plants standing in the field and then handling each sample from each plant individually, from the collection site to the data analysis. Other approaches, however, might be preferred based on budget, specific characteristics of a crop, reference library composition, and specific questions to be addressed. Some of the key decisions are whether to (1) sample seeds or leaves, (2) sample in bulk or individual plants, and (3) genotype using a pooling (or multiplexed) method or maintaining the individuality of a sample.

SAMPLING SEEDS OR LEAVES

For self-fertilizing crops, which have a low risk of pollen contamination, sampling seed instead of leaf tissue is reasonable from a theoretical standpoint and likely preferred in practice. Preserving seeds is easier than preserving leaf tissue. The higher moisture levels in leaves can lead to mold proliferation and tissue loss, whereas seeds can be dried if necessary before storage and shipping to the lab. For cross-pollinated crops, in contrast, sampling seed is in theory problematic. The seed could represent the genetic material of multiple plants (e.g., the ovule from one plant and pollen from another), and drift will cause it to move further and further from the true genotype of the variety with each generation of cultivation away from the original hybrid or variety. When multiple varieties are detected in an individual seed from an outcrossing crop, it could result from one of two processes: (1) farmers' cultivation of plants that stem from crosses or drift and selections they or others made in prior seasons, or (2) pollen contamination from other varieties in the same or another field in the current season. Genotyping has no way of determining which process led to that signal. Despite these theoretical concerns, in practice the outcrossing risk has proved to be minimal (Kilian 2019; Hearne 2019), even in crops such as maize, where the odds of cross-contamination are comparatively high. Still, the impact of genetic drift from farmers' practice of saving seed warrants further attention and clear documentation of the impact on varietal identification.

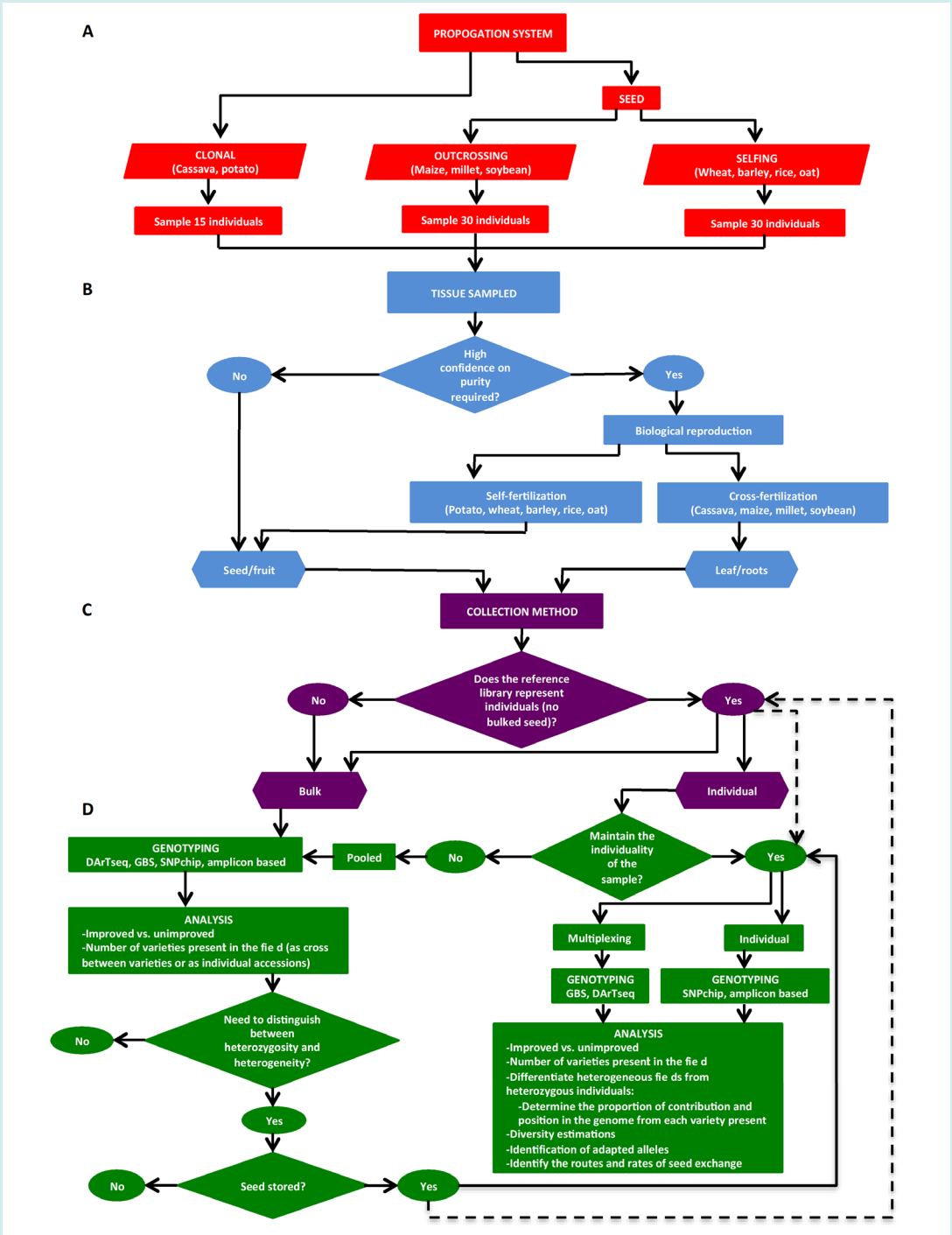
SAMPLING IN BULK OR INDIVIDUAL PLANTS

If the reference library to be used comes from bulked seed (or leaf tissue) and the DNA fingerprints for each variety in the reference library are allele frequency profiles, then it is necessary to obtain allele frequency profiles from farmers' fields, which can be obtain from a bulked sample. In such cases, no more information is obtainable by maintaining the individuality of each plant. There are times when the only material that can feasibly be collected is a bulked sample representing sampled plots (distinct from individual plants within the same field). Care should be taken to ensure that each individual sampled is equally represented in a bulk (e.g., use the same area of leaf tissue or the same number of seeds per individual). This bulked material can be used to identify the major varieties planted in the field and determine whether they are improved or unimproved. Without further study, however, it may not be possible to differentiate between heterogeneneity in the field (multiple varieties planted in the same plot) and heterozygosity in a sample (presence of different versions of an allele—called a variant—co-located within a single individual) (Box 4).

An alternative to this bulking method is to sample and store each individual plant's tissue separately and then either use part of the sample from each plant to bulk them before DNA extraction or pool them before genotyping. This approach will make it easier to assess heterogeneity versus heterozygosity and will enhance the differentiation of unimproved from improved varieties and potentially the number of

distinct varieties resolved. It comes, however, at a significant cost—currently some 20–30 times that of a bulk sample, depending on the numbers of individuals evaluated. Another potential advantage of individual-based assessment is the potential to return to the seed samples collected from each individual and later assess sibling seed used in the original genotyping. Seed from bulked samples can also be stored, but its potential for use in further, more in-depth analyses is limited, because additional resolution would require individually genotyping individual seeds from the bulk. This may be relevant if heterogeneity was detected in some samples, because it would allow in-depth analysis to be conducted on only those samples, saving some resources through a two-step process.

Figure B1: DNA Fingerprinting Decision Tree



REDUCE GENOTYPING COSTS BY MULTIPLEXING

In contrast to the sampling-in-bulk approach, the multiplexing method happens after the DNA is extracted from samples but just before genotyping. DNA from each sample is labeled with a unique DNA barcode that is used to distinguish the data from each sample that makes up the multiplex. The number of samples to multiplex is dependent on the number of barcodes available and the depth of sequencing desired. The more samples put into the multiplex, the lower the cost of genotyping for any one sample.

There are several key technical design choices to consider when deploying DNA fingerprinting at a large scale for the purposes of varietal identification. These choices can be clustered into three groups: (1) sample collection and handling, (2) genotyping technology, and (3) reference library. We outline the key steps involved—from field to data—in fingerprinting crop varieties and identify the options and trade-offs involved in each of these steps. In doing so we are conscious of how the design choices for DNA fingerprinting varieties will vary by country and crop context. At one extreme are commercial, monoculture production systems with efficient farm-to-lab infrastructure in place. At the other extreme are fairly common polyculture systems that contain multiple varieties (including landraces) and have less than ideal infrastructural and technical support. Approaches that make technical and practical sense in one setting may be inappropriate in another setting, and changing technological frontiers may also alter the balance over time.

Our review also covers six currently mainstream genotyping technologies that can be used to create DNA reference libraries and to identify varieties in farmer field samples. For each of these we consider their advantages and disadvantages. In addition, we discuss and evaluate protocols for the creation of a DNA reference library, with an eye to the purity and the representativeness of the library.[3]

[3] Sampling off-farm from within the seed supply chain can also be valuable. Current experience suggests that this is, in comparison with farmer surveys and collection, a low-cost activity with a potentially large payoff in terms of developing an understanding of the variety dynamics within the seed system.

2 SAMPLE COLLECTION STRATEGIES

During sample collection, household survey enumerators walk through farmers' fields and cut or harvest, store, and label plant tissue. One of the most challenging aspects of sample collection is the tracking and preservation of the samples through all the steps from field collection to genotyping. Without a robust chain of custody from farmer to data point, studies can be rendered uninterpretable. Recent DNA fingerprinting studies in Africa (Uganda, Malawi, Ghana) by Ilukor *et al.* (n.d.), Kilic (n.d.), and Rabbi *et al.* (2015) show the process involved from sample collection to data analysis. Each step of sample manipulation and shipment poses a risk for sample swaps, contamination, and degradation of plant tissue or DNA. Here we present a sample collection and handling protocol that seeks to minimize human errors to the extent possible, while also taking into account the practical, analytical, and cost implications of alternative options in the fingerprinting process (Figure B1 in Box 2).

2.1 Field Sample Size and Composition

The number of individual plants to sample from each farmer's plot depends principally on the propagation and biological reproductive (mating) system of the crop being studied (part A of Figure B1 in Box 2, and Box 3) and the genotyping protocol dictated by the composition of the reference library (e.g., bulked versus individual seed or leaf tissue).[4]

BOX 3: COMPLEXITY OF DNA FINGERPRINTING VIS-À-VIS PLANT CHARACTERISTICS

PLANT REPRODUCTIVE STRATEGIES

Although humans can propagate crops clonally or from seed, this might not be the biological reproductive strategy used by the plant species in a field setting. Plants reproduce using one of two main reproductive strategies: cross-fertilization (otherwise known as outcrossing) and self-fertilization (known as inbreeding or selfing). In cross-fertilization an ovule is fertilized by pollen from another plant, whereas in self-fertilized plants both the ovule and the pollen belong to the same plant. This has relevance for DNA fingerprinting because a plant can be fertilized by another individual within the field or from another field (regardless of how humans propagate the crop, e.g., clonally). This cross-pollination has the potential to produce seed that represents the genetic makeup of two varieties instead of the individual variety planted by the farmer.

[4] An important preparatory step in this fieldwork is the identification of the number, type, and location of the farms to sample. This choice in turn depends on the questions to be addressed by the fingerprinting survey (which may in fact be done as an adjunct to a survey intended for other purposes). If the intent is to survey "representative" farms, with the notion that the measured varietal use on sampled farm data will be scaled to provide broader (spatial) indicators of varietal use, then a host of (spatially explicit) sampling decisions are required, similar to those used in collecting other types of household farm data (see, for example, Pardey *et al.* 2020).

Therefore, the plant's inherent reproductive strategy can, in theory, affect the accuracy of genotyping when seed is sampled instead of leaf tissue. In species such as maize, where cross-pollination is favored, there is a tendency for pollen to come from nearby plants at a much higher rate than from distal plants. Additionally, since seeds are composed of mostly maternal tissue (Radchuk ***et al.*** 2011) and nutrient storage tissues that have a 2:1 ratio of genetic material from the mother compared with the father (Johnston ***et al.*** 1980; Yan ***et al.*** 2014; Costa ***et al.*** 2014), the contamination of the DNA signal from cross-fertilized pollen is lower than might be expected, a fact borne out by in-field testing (Kilian 2019; Hearne 2019).

PLANT PROPAGATION SYSTEMS

Crops can be propagated clonally or by seed. Producing seed for seed-propagated crops requires a generation of fertilization, which can occur through self-fertilization or cross-fertilization. Cross-fertilized crops represent the genetic content of two or more distinct and different parents, whereas self-fertilized crops maintain the integrity of one common parent with minimal changes, which are mainly due to natural processes.

Some crops, such as cassava, banana, and potato, are reproduced primarily through clonal propagation. This means that a part of the plant is used to generate a new individual with the parent plant's exact genetic content, carrying all its genetic characteristics. Thus tissue (other than fruit in most cases) collected from these types of crops should represent the material planted by the farmer. However, seed or fruit from a clonally propagated plant could result from a cross with pollen from another individual (if the organism is an outcrosser) or with pollen from the same individual (if it is selfer). Cassava, for example, is clonally propagated but produces seed through cross-fertilization. In addition, some species, including cassava, have strong self-incompatability. Therefore its seed will represent a cross-pollination event with pollen from another plant, likely a variety different from the one planted. For this reason, a leaf sample will more accurately represent the genetic material the farmer planted, though it may be more costly to obtain in practice.

Other clonally propagated crops, such as banana, produce fruit without the need of a fertilized ovule. For such crops, the fruit and any other plant tissue will represent the genotype of the plant planted by the farmer. Sampling the fruit, however, carries logistical challenges related to both its volume and its risk of tissue degradation.

Based on reproductive biology norms, clonally propagated crops such as cassava and potato could be represented by one sample per area (i.e., plot or farm field), if farmers were planting a single variety using uniform planting material. However, experts participating in the January 2018 Seattle fingerprinting workshop with experience in cassava breeding and varietal adoption studies noted that in practice many farmers plant multiple varieties within a sampling area, which gives rise to a high degree of varietal heterogeneity within each individual farmer plot. To ensure that this potential heterogeneity is reflected in efforts to assess the varietal diversity in the fields of farmers growing cassava, we recommend one of two strategies. If there are strong reasons for differentiating the planted cassava stands within a plot (for example, if the farmer differentiates and gives several different local names for the planting material used) then this information should be reflected in the sampling approach, with each "variety" being sampled separately. Absent this kind of differentiation, a random sample of 15 individual plants is recommended, as used by Le *et al.* (2017).[5]

[5] Girma *et al.* (2017, p. 6) recommends collecting "two newly expanded apical leaf tissues of approximately 6 cm from a single stem." However, this recommendation assumes complete homogeneity within each sampled field.

For crops with reference library data generated through fingerprinting of bulked material that is a composite of multiple individuals per sample, varietal identification relies on the accurate estimation of allele frequencies in the targeted sample. In population genetics studies, a sample of 20–30 randomly selected individuals is deemed sufficient to capture most of the allelic variation within a population (see, for example, Watterson 1975). Theoretical work by Fung and Keenan (2014) on the estimation of confidence intervals for population allele frequencies concluded that the sample size should be more than 30 individuals. However, empirical studies using highly heterogeneous maize landraces determined that sampling 20–30 individuals per population provided saturation of differentiation using sequence based genotyping methods (Hearne 2019). For establishing varietal identification, the optimal number in the sample would likely vary according to the specific question being addressed or the sensitivity required (Kilian 2019). Based on these findings, we recommend sampling 30 or more individuals from each field for cross-pollinated crops (such as maize) and self-fertilized crops (such as wheat, barley, oats, and rice) for which an allele frequency profile will be used as a reference.

In all cases, irrespective of the sample size, individual plants should be collected throughout the plot or field to ensure that the sampling is representative of the material growing in each plot or field. Plants from the peripheries of each plot or field should be avoided if collecting grain, as these are more prone to contamination from pollen from plants in nearby fields, especially in open-pollinated species (see below for more discussion of this matter).

Plant Tissue Collected

In theory, researchers have a variety of options for tissues to sample, depending on the crop propagation system (clonal or seed) and breeding system (self-pollinated or cross-pollinated) (part A of Figure B1 in Box 2). Because it is impractical to collect leaves without plant-to-plant contamination (Kilian 2019),[6] the most common practice is to collect other plant tissue: grain for seed crops and tubers for relevant clonally propagated crops (part B of Figure B1 in Box 2). One option is to sample one cob for each individual maize plant; one spike for each wheat, barley, and oat plant; one pod (averaging 2.5 seeds per pod) for each soybean plant; and one tuber for each clonally propagated tuber producing plant (with a caveat for cross-pollinated crops planted in heterogeneous plots, as discussed below and in Box 4). In cases where suitable, timely processing is available and there are large volume transport and chain-of-custody solutions, root and fruit collection can be considered for clonally propagated plants like cassava and musa species. In cases where these are not available, processes for leaf collection should be carefully considered and the cost-benefit of sampling each tissue type evaluated. Sampling just one spike, pod, tuber, or cob per plant typically provides enough tissue for subsequent DNA extraction.[7] The goal is to collect enough tissue to represent each individual plant to be sampled within the farm plot in a balanced manner (i.e., in a sample collected from 30 plants, one would not want to have only 1 seed from each of 28 plants and 40 seeds from the remaining 2 individuals) and to ensure that sufficient DNA can be extracted for genotyping.

[6] It is possible to sample young leaves from farmers' plots in a way that avoids the risk to varietal identification that could arise from out-crossing, but doing so requires robust training of those collecting samples and adequate infrastructure for collecting and maintaining un-degraded leaf tissue and applying robust chain-of-custody procedures. Practically speaking, however, robust training and adequate infrastructure are often lacking, especially in rural areas. Most surveys therefore collect other tissues that are more amenable to in-field processing. See Box 2.

[7] Sampling a complete tuber for every plant may be unwieldy for large studies, and so one option is to sample tuber cores. For many species, however, tuber cores degrade quickly.

Collection Method

Irrespective of the propagation system being sampled—i.e., clonal or sexual via seed—the collection method that preserves the most analytical options (but is also by far the most costly) is to sample individual plants and avoid bulking them during sample collection (part C of Figure B1 in Box 2, and Box 4). This procedure is especially important if the goal is (1) to identify the number or proportion of unique intercropped varieties planted in a particular field in a particular season, versus (2) where either a) the seed planted by the farmer derives from multiple varieties resulting from naturally occurring crosses within the current season, or b) the planted seed stems from crosses that farmers may have intentionally made over prior seasons. This approach serves as a means of differentiating between heterogeneity in the field (stemming from farmers' mixing of different types of seed, perhaps from multiple sources, in the process of planting) and heterozygosity in the individual plant (which is expected to be low in self-pollinated species and higher in cross-pollinated hybrids) (Box 4).

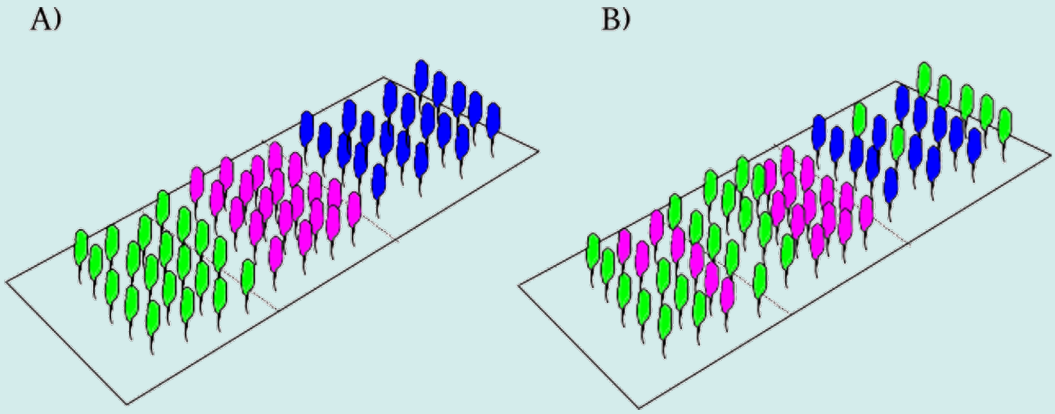
BOX 4: HETEROGENEITY IN FIELD VERSUS HETEROZYGOSITY IN PLANT

Heterogeneity in the field refers to the presence of multiple varieties within a field (Figure B2). Heterozygosity is the presence of different versions of an allele (called a variant) co-located within a single individual.

Crop breeders typically try to minimize the amount of heterozygosity in inbred species before a line is released as a variety. In the case of hybrid crops and some clonally propagated crops like cassava, the performance of the variety—that is, its hybrid vigor—is positively correlated with heterozygosity (hybrid maize for example is produced by crossing two or more highly differentiated lines). In each generation, half of the genetic material from a maternal plant is combined with half of the genetic material of a paternal plant through pollination, which results in seed formation (except in banana). The maternal and paternal contributions can come from the same plant (self-fertilized crops) or from two different plants (cross-fertilized crops). DNA fingerprinting analysis can help determine the level of heterogeneity in a field or heterozygosity in a sample. Before the analysis, however, it is possible that a field is heterogeneous even if a farmer believes it is homogeneous. Since it cannot be assumed that a field is homogeneous, it is important to consider the possibility of heterogeneity and its implications in DNA fingerprinting analysis when collecting different plant tissues.

In cross-fertilized crops there is a risk that pollen from one variety (in the same field or, less likely, from another nearby field) will fertilize a flower from a different variety. At the molecular level, the seed coming from such a cross will match two varieties in the reference library with a closer similarity to the maternal variety because seed contains predominantly two copies of the maternal genome and one of the paternal genome. The leaf tissue of the plant from which the seed was sampled will match only one variety (specifically, the variety planted by the farmer). Meanwhile another leaf from another sample from the same field might match yet another variety in the reference library. From evidence gathered from leaf tissue under this scenario, it could be concluded that multiple varieties are planted in that field—in other words, it is a heterogeneous field composed of two or more varieties. Fingerprinting based on seed could support two explanations: either (1) the farmer has a heterogeneous field of plants representing crosses of two or more varieties in previous seasons, or (2) there are two or more varieties planted distally to each other that had, at some unknown rate, exchanged pollen in the current season to produce the seed sampled. Although in both scenarios two or more varieties are detected, the cause and consequences of the two scenarios are very different.

Figure B2: Homogeneous versus heterogeneous plots



Notes: Panel (A) shows three plots within a farmer's field; each plot contains one variety (green, pink, and blue). Panel (B) shows three heterogeneous plots; each plot contains more than one variety so even samples taken from the middle of each plot might be contaminated by pollen from adjacent individuals.

Bulking plant samples collected from within a given field is recommended if the purpose is to assess the cultivation or distribution of improved versus unimproved varieties (i.e., within-farm plot variability is not a primary focus of the assessment). If the varieties grown in farmers' fields were generated in seed improvement programs that do not have formal and planned crossing (e.g., breeding of maize open pollinated varieties, where varietal selections are typically made by crossing populations of maize plants rather than crossing specific maize inbreds), allele frequencies are measured from bulked samples and matched directly to the allele frequencies recorded for bulked reference library material.

It is also possible that some of the multiple varieties found in a particular farmer's field come from crosses that the farmer made—purposefully or inadvertently—in prior seasons. If samples are collected and sequenced individually, population genetic approaches can potentially be used to identify the varieties used in such crosses. If samples are collected in bulk, however, the DNA signal implicating multiple entries in a reference library could be coming either from a heterogeneous field in which the farmer planted multiple varieties (while preserving their characteristics intact) or from individual plants that are the product of crosses made by the farmer in prior seasons (i.e., heterozygous samples). To retain the ability to distinguish between multiple varieties in one field and multiple varieties in one plant, the plants gathered from each field should be sampled and stored individually rather than bulked. Logistical or cost considerations, however, may make sampling in bulk the only practical option. In such cases, bulked tissue samples can be collected, but with the understanding of the analytic limitations of bulked samples, where heterozygosity in a line can be masked with heterogeneity of multiple varieties in the field (parts A, C, and D of Figure B1 in Box 2, and section 3 below).

2.2 Handling Samples

Among the most critical steps for the success of DNA fingerprinting are preserving the identity of the collected samples and maintaining the viability of the sampled material for DNA extraction and genotyping. The process thus requires clearly articulated sample handling protocols, a barcoding system (or equivalent) that is deployed from field to genotyping center, suitable methods of sample preservation, and good practices and protocols for genotyping that reduce the risk of sample contamination or degrada-

tion. Barcoding software is available for researchers to print their own sets of stickers (one for each step at which the sample will be transformed). Alternatively, rolls of barcode stickers can be ordered for delivery.

Sampling Leaf Tissue

To avoid DNA contamination when sampling leaf tissue from individual plants, measures should be taken to limit excessive cross-sample contamination. These measures include cleaning hole punches, scissors, or other cutting tools used to take samples. Cleaning is not done to forensic standards; rather, the aim is to remove excess large-volume contaminants like sap and leaf tissue. A protocol developed by LGC for leaf sampling in the field proposes washing the cutting tool between samples by placing the end of the instrument into a container with clean water and dipping it 5 to 10 times (LGC Genomics 2017). After washing, the instrument should be shaken until it is completely dry before it is reused. Alternatively, wiping the edge of the cutting instrument with a clean, damp tissue is an effective sampling method used extensively in other industries where leaf sampling is prevalent.

Samples can be collected and placed into suitable volume plastic tubes or directly into the wells of a 96-well plate (the standard size used for DNA extraction; Box 5). Barcoded labels should be printed and attached to each tube and plate ahead of time. All else being equal, we recommend the use of 96-well plates (rather than individual plastic tubes), as this reduces the number of steps from test tubes to well plate for DNA extraction and thus lowers the risk that samples will be mislabeled. Each sample should be recorded as it is taken. The risk of human error in data entry is minimized if computational tools (such as the Survey Solutions CAPI application used by Ilukor *et al.* n.d.) rather than manual methods are used to record each barcode and all the metadata related to each sample.

Sampling Seed / Grain

Seeds have some important advantages over many other tissues (including leaf) as a source of DNA for genetic identification testing because they have evolved to be much more robust for handling and have increased shelf life. Many tissues have a window in the development of the plant when the sampling provides suitable material for DNA extraction. Because seeds are much more “stable,” they can be effectively collected by enumerators with limited training or technical skill and are easily stored in dry conditions for a period of time after harvest. In some countries such as Ethiopia, grain samples are collected from farms in large-scale surveys ("crop cuts") at the time of harvest, as part of the data collection system underlying the country's official agricultural statistics, and these samples can then be used for genetic identification without a dedicated sampling of leaf or other tissue.

The optimal size of the seed sample depends on a number of factors:

1. Accuracy and precision required from the assay: A larger sample has a smaller sampling error and therefore a more accurate purity estimate and a more reliable genetic identification.

2. Size of the seed: It is much easier to take and transport a sample of small seeds, and to crush a sample of small seeds, than it is to process large seeds.

3. Logistics of sample processing and equipment available to pulverize seed samples.

4. Composition and genetics of material: A smaller number of seeds is sufficient for highly inbred material, especially in countries with a well-developed seed system where hybrids are used. Where seed systems are less developed, a larger number of seeds must be sampled from each enumeration unit.

As mentioned, in most cases the sample used for extraction should include at least 30 seeds.

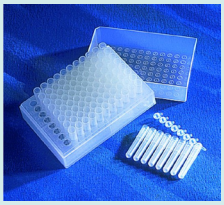
DNA Isolation

For all tissues, DNA extracted from samples needs to be of high quality or purity. The tolerance of different genotyping systems for contaminants varies; nonetheless, an un-degraded high-molecular-weight DNA sample without enzyme inhibitors is desired. The extraction method must deliver DNA at a reasonable concentration (preferably above 10 ng/µl) at a reasonable cost, given the available capacity (human, equipment, chemical handling, and waste disposal) at extraction centers. In the case of leaf samples of some species, seed samples, and tuber samples, the DNA extraction method needs to effectively eliminate polysaccharides from the extract; this is achieved by using appropriate lysis buffers and extraction conditions as well as, when needed, additional cleanup procedures. It is also important to obtain reasonably similar concentrations of DNA across the tested samples. In cases where DNA concentration varies significantly (e.g., over 10-fold) across samples, the variation can have a significant negative impact on genotyping data quality. It is therefore important to perform extractions in a highly standardized manner (preferably using robotics or at least microplate-based systems). For smaller projects using manual extraction, it is important to obtain proper quantitation of DNA concentration using gel eletrophoresis and adjusting the concentration. For all projects, proper DNA quality control testing for the presence of both inhibitors and DNases is an important prerequisite to securing good genotypic data and therefore good classification of test samples.

BOX 5: SAMPLE STORAGE TECHNOLOGIES

One option for storing samples in the field is to place them directly in 96-well plates or tubes. The options for preserving leaf tissue are to (1) use silica gel or another small packaged desiccant in each tube in a 96-well plate and cover each sample with a cap, or (2) place the sample in the tube, cover it with a permeable membrane, and place a molecular sieve desiccant package on top of the samples to absorb the moisture.

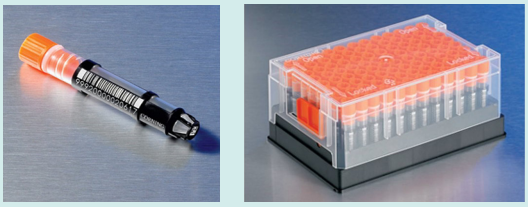
Figure B3: 96-well plate cluster tubes are composed of 8 (1.2ml) polypropylene tubes in strips that can be arranged in a 96-well rack



Note: Each strip can be covered with 8-cap strips or permeable membranes. The arrangement of a 96-well plate is the standard arrangement for DNA extraction and genotyping machines. Each axis in the rack holding the tubes tracks the position of each sample in the well plate.

Source:www.sigmaaldrich.com/catalog/product/sigma/cls4401?lang=en&region=US..

Figure B4: Micro tubes are individual tubes that can be individually barcoded for sample identification.



Note: Each tube can be independently covered with a cap. The picture shows an example of a micro tube that includes 1D and 2D barcoding. The tube shown (left) is from Corning™ and can be arranged in a 96-sample rack (right).

Source: www.fishersci.com/shop/products/corning-barcoded-storage-tubes-1d-2d-barcoded-9/p-4099031.

Figure B5: Traditional 2.0 ml micro tube with sealing film"

Note: The film allows gas exchange while keeping contents protected. The film could be used in the barcoded tubes (shown in Figure 5.2) or in each of the cluster tubes in a 96-well plate cluster (shown in Figure 5.1). In any case where this film is used, a molecular sieve desiccant package could be placed on top of the vials as they are added to the collection (preferably arranged in 96 samples), attached, and sealed with tape (Figure 5.2).

Source: https://www.usascientific.com/9126-1000breathe-easyandtradetubemembranes.aspx.

Figure B6: An example of a molecular sieve desiccant package.

Source: https://midsouthpackaging.com/desiccants/.

Control Samples

Physically moving material from sample collection through to genotyping involves several steps that require human or mechanical manipulation. Each step has the potential to introduce errors into the collection (e.g., sample mix-up, contamination, or loss) that will reduce the accuracy of DNA fingerprinting. Just as crop breeders plant control or check varieties in their field trials, researchers must deploy control mechanisms during the fingerprinting process that can (1) reveal some of the problems that may arise during this process and (2) enable them to correct or account for these problems during data analysis.

The most straightforward approach to tracking potential errors is to include technical replicates at both the sampling and the DNA extraction stages. During field sampling, a defined subset of farmers' plots are chosen at random to be sampled in duplicate, and these should ultimately show up as identical in terms of purity and identity at the lab stage. Details of which plots are duplicates should remain blind to the analyst. If duplicated plots are not self-evident at the analysis stage, a mix-up has occurred. In the same manner, a defined subset of samples can be assigned randomly or in a structured manner to duplicate DNA isolation. More elaborate control schemes could be prescribed that would allow surveyors to determine whether genotyping inconsistencies are errors due to DNA contamination as a result of inappropriate cleanup of the tools used for tissue collection or due to DNA contamination that occurred later, during DNA extraction or genotyping. In most field conditions, however, it is not practical to employ these schemes.

Controlling Genotyping Errors

Although most genotyping technologies intrinsically have low error rates, human errors made in handling material in the field, in the lab, and during the genotyping procedure itself are a constant risk. Some of the most common causes of genotyping errors are plate rotation and DNA contamination. Placing control samples in opposite corners of a 96-well plate can help to correct the directionality of the plate. DNA contamination from other well cells in a run can be assessed by the level of purity in the control samples placed randomly in the well plate and through analysis of duplicated samples (see points above).

3 GENOTYPING TECHNOLOGIES

In the process of identifying varieties through DNA fingerprinting, genotyping is the last step before data analysis (part D of Figure B1 in Box 2). Several genotyping technologies are currently used in crop development, and several emerging technologies are coming into practical play. These technologies vary in the amount of genome covered, from the capture of a set of fixed variants (e.g., SNP arrays or amplicon-based methods), to the capture of large portions of the genome (e.g., exome capture, DArTseq, tGBS, and GBS), to representation of the whole genome (e.g., whole genome sequencing, WGS) (Table 2 and Annex 1).

Table 2: Cost and performance aspects of genotyping technologies

Technology	No. of variants assessed/ sample	No. of samples assessed	Cost per sample ($)	Detection of novel variants	Processing required to obtain sample by variant matrix	Reproducibility
SNP arrays	6K–80K	48–1,152	50–100	No	High, but provided	>99.9%
Amplicon-based	100–100K	24– 1,000	40–80	Only in targeted regions	High, but provided	Depends on depth
Genotyping-by-sequencing	10K–1,000K	≤100K	30–90	High	High, but provided[a]	Depends on depth
DArT-seq	4K–30K	>40K	7–35	Moderated	High, but provided	99.7%
Exome capture	200K–10M	1,000–3,000	700–1,500	Only in targeted regions	High	>99%
Whole genome sequencing	>1M	<1,000	700–2,500	High	High	Depends on the region

[a] Licensed use of genotyping-by-sequencing technologies for plants is available at only a few facilities worldwide, where processing of raw sequence reads is done for the customer.

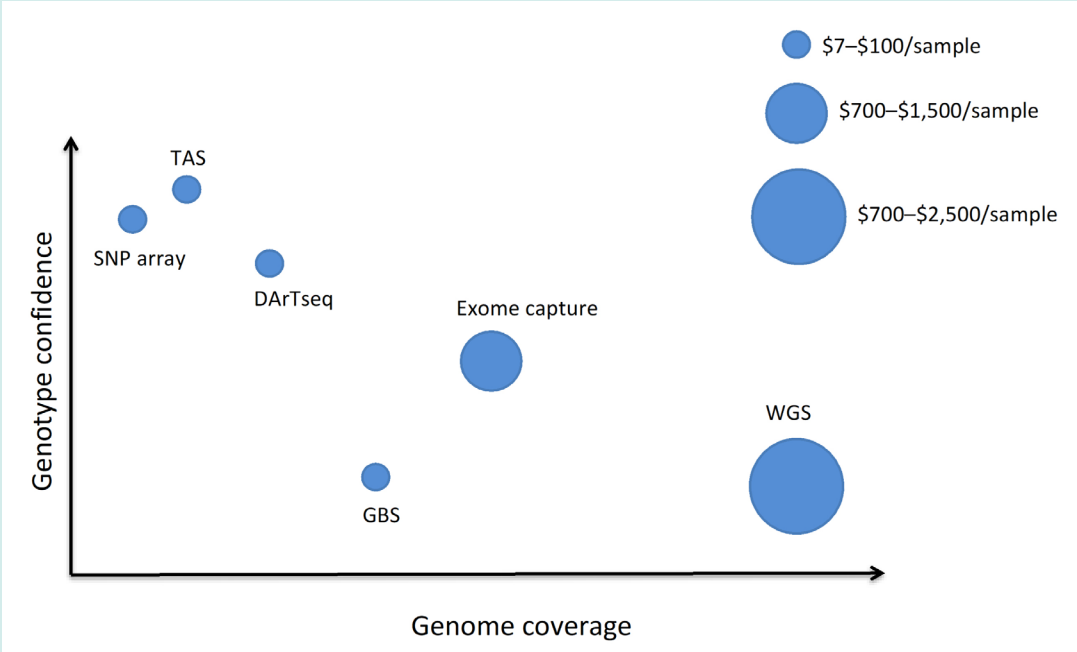
Source: Authors' compilation based on quotes from sales representatives.
Notes: Number of variants assessed indicates the number of markers that are identified with each technology. Number of samples indicates the approximate size limits for a project. Cost per sample depends on organism characteristics (genome size, ploidy, and complexity); cost estimates were made in late 2017. Detection of novel variants indicates the level at which variants can be observed for the first time from the data.

These applications also vary in cost from $7 to $2,500 per sample (late 2017 prices), with costs changing as a function of the precision of each genotype and the amount of genome covered (Box 6). Here we briefly review the technologies currently being used, noting that the information provided about them, including details about costs and reproducibility, are subject to change over time. We endeavor to highlight factors that might favor one technology over another in the decision process. Among these are crop-specific characteristics that can influence the genotyping strategy to use. These are ploidy level and genome size, which are described in more detail in Box 7.

BOX 6: GENOTYPING—COSTS AND CONFIDENCE

Figure B6 shows the difference in cost for six genotyping technologies depending on the relative proportion of the genome that each technology covers and the confidence in the genotypes obtained. Although the portion of the genome covered varies greatly among platforms, the genotyping confidence can be comparable between SNP arrays and sequence-and-discover approaches if a high read depth (>20X) is obtained; increases in the number of reads can significantly increase genotyping confidence at each site and also the cost per sample. The minimum number of variants assessed in this estimation is 100 SNPs.

Figure B7: Cost, genome coverage, and genotype confidence for six genotyping technologies



Source: Cost estimates are compiled by authors from quotes from LGC and UMGC (see also text Table 2).
Note: Costs were obtained in late 2017.

BOX 7: FACTORS AFFECTING THE NUMBER OF VARYING MARKERS REQUIRED FOR DNA FINGERPRINTING

The cost of genotyping depends on the number of variants to be measured and the accuracy required for each measurement. Several genotyping technologies are available for DNA fingerprinting (see Annex 1) with costs (using late 2017 prices) that range from $7 to $2,500 per sample (see Table 2). The number of markers required to distinguish one variety from another depends on the diversity of the crop and the size of its genome.

PLOIDY LEVEL

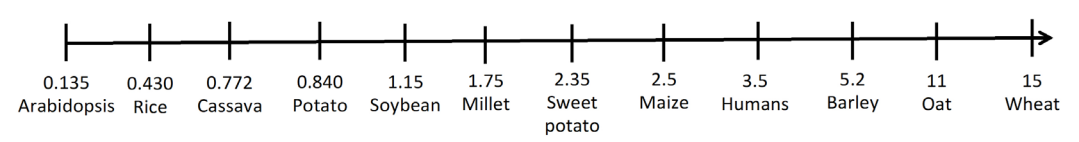
One of the many factors that can affect the amount of diversity in a crop is the ploidy level (other factors include rates of mutation, recombination, and migration). Ploidy level refers to the number of chromosome sets an organism has. Humans, for example, have two homologous (similar) sets of chromosomes (referred to as diploid), with 23 individual chromosomes in each set. Among the major crops produced in the world, the number of chromosome sets can range from two (e.g., barley, beans, cassava, maize, rice, millet, soybeans, and sweet potatoes) to four (e.g., cultivated potatoes, alfalfa) to six (e.g., bread wheat and oats) and even higher (e.g., strawberries). Plants with more than two sets of chromosomes are classified as polyploid. There are two types of polyploid: auto-polyploid and allo-polyploid, which differ in the origin of the chromosome sets. If the sets come from the same plant they are classified as auto-polyploid, and if they come from different plants they are classified as allo-polyploid.

Organisms with higher levels of genetic diversity require a larger number of genetic markers (SNPs) to capture the diversity and to determine DNA patterns that can be used as fingerprints. Higher ploidy levels bring together ancestrally related genes into the same genome, making it likely that the genome will have multiple copies of a gene. Detecting and discriminating between multiple copies of a gene versus multiple mutations of the same gene can be challenging, but it can be achieved with highly sensitive genotyping platforms such as SNP arrays or other Amplicon-based genotyping, and with the use of deep coverage sequencing for re-sequencing platforms (GBS, DArTseq, exome capture, and whole genome sequencing).

GENOME SIZE

Genome size refers to the number of nucleotides (base pairs) found across all chromosomes in an organism. Among the major crops genome size ranges from 0.43 Gb (giga basepairs) in rice to 15 Gb in wheat (Figure B8). Larger genomes require a larger number of markers to be gentoyped to get a good representation.

Figure B8: Genome size for selected organisms (Gb).



Source: Compiled by authors.

RECOMBINATION RATE AND DIVERSITY

Recombination typically occurs during meiosis (the formation of gametes). During recombination, DNA segments from one chromosome in a homologous set are swapped with the same segments of DNA from another chromosome in the homologous set. Recombination is a process that creates genetic diversity as different forms of genes are swapped, creating new and different gametes. The rate of recombination differs along the physical genome and also between species. In plants, the larger the genome, the lower the recombination rate (Stapley ***et al.*** 2017). The higher the recombination rate, the more markers are needed for a given genome size. The inherent diversity of a crop is a critical consideration—not just at a species level but in the context of the breadth and possible number of varieties in a market. Crops with high diversity require more markers than those with low diversity. It is important to appreciate the recombination rate and the underlying level of genetic diversity present in a species when considering how many markers are required in assessments.

3.1 Reference Library Approaches

One of the requirements for varietal identification is the ability to sample genetic variants or variant combinations (haplotypes) that clearly distinguish multiple varieties in the reference library. We recommend using a sequence-and-discover approach for the development of the reference library given the wide range of gene pools used in the construction of varieties in target geographies (see reference library creation section below). Currently there are four technologies of this kind: whole genome sequencing, exome capture, genotyping-by-sequencing variations, and Diversity Arrays Technologies (DArT) platform technologies.

Whole Genome Sequencing (WGS)

Whole genome sequencing (WGS) is the process of creating information on a near-complete DNA sequence of an organism. Even though it would be useful for future projects to collect this level of data for every sample, the data collected are not essential for varietal identification, the current cost per sample is high, and processing the output requires highly trained personnel.

Exome Capture (EC)

Exome capture is a type of reduced-representation sequencing, wherein only a portion of the genome is reproducibly captured and sequenced. In exome capture, a large expressed genomic space (including all the protein coding regions) is represented by genomic sequences from that space that are physically affixed to an array. DNA fragments from each sample that bind to these sequence "baits" are preferentially collected and sequenced. Genetic variants (SNPs) are later discovered from these sequence data. This technology requires a large investment to develop an unbiased array containing the targeted genomic space, the cost per sample is high, and it requires highly trained personnel to process the data coming from the sequencing center. The genomic space surveyed with this array depends on which segments were used as baits (likely resulting in ascertainment bias), which could, to a small extent, limit the detection of genetic differences between closely related varieties. Based on the cost and sophisticated processing capabilities required, we do not recommend this approach for the routine development of a reference library.

Genotype by Sequencing (GBS)

Genotyping by sequencing (GBS) (Elshire *et al.* 2011; He *et al.* 2014) and tGBS are other reduced-representation methods to discover variants from fragments of sequenced DNA. GBS uses restriction

enzymes that cut up the genome in repetitive regions and washes those regions away. Then, only the fragments lacking repetitive sequences are sequenced. These sequences can be aligned to each other de novo or to an existing genome reference sequence, and variants are identified within the pileup of reads at each position in the reference sequence. It produces highly reproducible results that are cheaper than either of the methods discussed above. Some ascertainment bias—i.e., over- or underrepresentation of one variant relative to another—is inherent in the data since the regions in the genome that the restriction enzymes cut may have high mutation rates and therefore be unrecognized by the restriction enzyme. Despite this caveat, this method is capable of identifying a large number of variants along the genome, and, if done with deep read coverage (>20 reads covering the same genomic space), it is also capable of identifying rare variants that can more easily distinguish between closely related varieties.

The quality at which each variant is captured intrinsically varies from sample to sample, such that variants may be observed with high quality (based on read depth and certainty of each base call in the alignment) in one sample but discarded from another sample owing to poor sequencing. This feature of the process results in a high level of missing data across samples. But given the large volume of data captured for each sample, there can still be many sites where high-quality variant calls are made for all the samples in the set—but the precise sites cannot be preselected. GBS technology is protected by patents owned by KeyGene N.V., and UMGC, BGI, and LGC (with global representation) have licenses from KeyGene to use this technology. tGBS is more similar to DArTseq technology, detailed in the following section. tGBS technology is owned by Data2bio.

DArTseq

DArTseq is yet another reduced-representation method designed to capture most of the functional part of the genome at an inexpensive cost per sample and with a high rate of reproducibility. DArTseq makes use of restriction enzymes to reduce the genome complexity by removing the repetitive portion of the genome. The remaining pieces contain predominantly active genes. These pieces can then be sequenced at a determined depth and aligned to each other or to a genome reference sequence. The advantage of this technology over other enzyme-based approaches like standard GBS is that DArT reduces the genome to the portion of the genome that can be sequenced in most of the samples. This lowers the number of missing data points across samples, which can be an advantage when creating a catalog of DNA fingerprints for inclusion in a reference library. The systematic nature of the procedure used makes it extremely reproducible through multiple runs, and data processing is aided and standardized by proprietary software. This technology is one of the lowest-cost reduced-representation genotyping technologies (<$35 per sample). DArTseq is a proprietary technology available at Diversity Arrays Technologies, Australia.

Either GBS-based systems or DArTseq are recommended for the identification of genetic variants that are required to create a reference library.

3.2 Genotyping for Routine Varietal Identification

After a reference library has been constructed and genotyped, a catalog of DNA fingerprints for each entry (or an allele frequency profile for bulked samples) can be determined for each variety in the library. The individuals collected from the field need to be genotyped with a technology that captures at least a fraction of the genetic variants included in the reference library, and this fraction must be robust enough to contain the unique DNA identifiers that distinguish one variety from another in the reference library. Here we present three of the technologies that can be used to identify just a subset of variants for discriminating among varieties.

SNP Arrays

SNP arrays detect genetic variation within populations by assaying single nucleotide polymorphism (SNP, a single site in the DNA). Although this technology is fast, accurate, easy, and cheap to use, it has two major drawbacks when applied to DNA finger-

printing: (1) it relies on variants found at medium to high frequency, so it might not contain rare variants (or variant arrangements) that are unique to a variety, and (2) variants on the SNP array are restricted to the genetic variants observed in the set of samples used to identify SNPs in the first place (known as a discovery panel). If the panel was shallow, the discovered SNPs may underrepresent the overall genetic diversity that actually exists among the varieties available for that crop.

Expanding the varietal diversity beyond the discovery panel greatly improves the ability of the array to identify varieties in the field. Therefore, unless the variants included on the SNP array are able to differentiate the samples in the reference library (and assuming that the reference library is complete and not missing data on important unimproved materials), this approach would not be able to discriminate among varieties taken from the field. If an existing array cannot discriminate between varieties in the reference library, developing a new array is costly and time consuming. There are several core processing centers worldwide that offer this service once the array exists, but given the drawbacks of this approach we do not recommend it for routine tests.

Targeted Amplicon-Based Sequencing (TAS)

Targeted amplicon-based sequencing (TAS) technology (e.g., the NuGen Allegro platform) has much of the convenience of SNP arrays but is more flexible for expanding the panel of variants found in ever-larger varietal collections. This technology uses restriction enzymes to genotype known variants identified in the reference library, but unlike SNP arrays, which need to be rebuilt entirely to accommodate new SNPs coming from new varieties, TAS technologies can be easily expanded. The cost per sample is presently $50–100, processing is fast and easy, and the data are highly reproducible. Any variants included in TAS can correspond directly to variants identified in other sequencing-based platforms. For instance, TAS technology uses 500 base pairs of DNA preceding the SNP, giving a designer ample ability to target precise locations in the genome to match sites from other platforms. One of the advantages of this approach is that variants selected for surveillance in a sample set—although fewer than those obtained by resequencing technologies like GBS and the DArT portfolio of sequencing technologies (DArT 2019)—can come directly from the variants that most distinguish one variety from other varieties in the reference library.

Other Marker-Based Technologies

DArT has developed a series of targeted-genotyping techniques[8] for routine genotyping as well as variety identification at a reduced cost compared with more comprehensive assays like DArTseq and GBS. These methods are cost-effective for larger assay volumes (>1,000 samples) and at the time of writing cost $5-10 per sample depending on the number of SNPs in the panel, which usually ranges from a few hundred to a few thousand markers. DArT platforms are able to capture both rare and common variants with high confidence owing to the high read depths supporting them, with a typical range from 100 to 300X. While the DArTcap method is restricted to the selection of markers derived from the DArTseq assay, the other two methods (DArTmp and DArTag) can use SNPs from any marker discovery platform. When derived from a bulk sample, the “counts” of sequence variants/alleles are used as a proxy for allele frequencies and combine into a profile that can be matched to a bulk reference variety profile generated using the same method. This is the recommended approach to use for routine DNA fingerprinting analysis when the reference library is composed of bulked samples that can be differentiated by their allelic profiles and when the SNP markers with good discrimination power for a target set of reference varieties are already in hand.

[8] These methods are described in more detail at https://www.diversityarrays.com/technology-and-resources/targeted-genotyping.

3.3 Multiplexing and Pooling

Multiplexing and pooling are two practices that can be used to lower genotyping cost once the DNA extraction has been completed for each individual sample (part D of Figure B1 in Box 2).[9] With multiplexing, each segment of the DNA from an individual plant is labeled with a unique sequence of DNA barcodes[10]; fragments from different samples are then combined and sequenced (Smith *et al.* 2010). With pooling, DNA from different samples is combined, sometimes in different combinations (combinatorial pooling) without being labeled, and then passed through one sequencing run. The maximum number of samples that can be included in one sequencing run depends on the sequencer throughput (number of reads), desired coverage (read depth), and the size of the genomic target (or number of targeted amplicons). The following equation determines how many samples to pool or multiplex for each organism:

samples to pool or multiplex = sequencer throughput/(coverage per sample × size of genomic targets).

Multiplexing and pooling methods are used during the creation of a DNA fragment library in the early steps for sequencing-based genotyping technologies (GBS, DArT, TAS, exome capture, and WGS). Each approach serves a different objective. Multiplexing using DNA barcodes enables tracking of the origin of each sequenced read, which facilitates the identification of all the variants found in each sample (e.g., Elshire *et al.* 2011). Pooling is used when the discovery of new variants across a population (rather than distinguishing variants among specific samples) is the goal (e.g., Douzery *et al.* 2013). While combinatory pooling can be used to genotype different samples, the designs and analysis procedures used are complex, with a clearer application in rare variant screening across samples (Cao and Sun 2016). The clear identification of the genetic makeup of an individual sample is vital to the success of the DNA fingerprinting approach, especially when closely related varieties are being analyzed or when the sample consists of a mixture of multiple varieties. Therefore multiplexing is the recommended choice for studies of varietal identification, especially when the objective is to identify the specific varieties included in samples taken from farmers' fields.

[9] This is in contrast to the bulking procedure mentioned above, where the sampled plants from each field or plot are combined before DNA extraction.

[10] Note that this use of the term "DNA barcode" is metaphorical and distinct from the discussion of barcode stickers for logistical tracking of samples from section 2.2.

4 REFERENCE LIBRARY: CREATION AND MAINTENANCE

The effectiveness of any method of varietal detection depends critically on the quality and representativeness of the reference library used. A lack of variety representation, misidentified varieties, or the inclusion of contaminated samples in the reference library may result in an erroneous assignment of a sample to a variety and/or a failure to achieve a varietal match between the sampled DNA and the reference material. We therefore emphasize the fundamental importance of investing sufficient resources (including time) in the creation of representative and pure reference libraries before investing in other efforts to survey varietal adoption.

4.1 Representative Library

The ideal reference library for varietal identification should contain all the possible varieties (private and public) likely to be grown by farmers in the sampled area and not be limited by assumptions about what farmers might be growing. This is especially pertinent for production systems where landrace cultivation, informal farmer-to-farmer seed dissemination, and seed recycling are commonplace. The need for a representative reference library poses a somewhat limiting chicken-and-egg problem concerning the use of DNA fingerprinting methods in areas subject to complex seed use systems. The primary question addressed by genomic data—e.g., distinguishing improved from unimproved varieties versus identifying what specific varieties are present—needs to be considered. Notably, if the primary objective of crop varietal fingerprinting is to identify what is grown in farmers' fields, a reference library of limited varietal representation will inevitably fall short in providing a detailed answer, irrespective of the genotyping platform or sampling approach used. Information should be sought from locally knowledgeable breeders in the national agricultural research systems, CGIAR, seed companies, processors, extension agents, and past surveys to determine the varietal entries deemed appropriate for the reference library. A good standard would be to include all improved varieties thought to have been disseminated in the relevant geography, plus all landrace varieties that have been collected from the same, the latter often being available from genebanks. The more comprehensive this resource is, the better any interpretation of the data will be.

4.2 Purity of a Variety in the Reference Library

DNA fingerprinting assumes not only that the reference library is complete (i.e., represents all the possible varieties available), but also that each genotype truly reflects one variety. In the case of inbred species or clones, this means that the seed or clone is not segregating morphologically and genetically (i.e., it is pure). In the case of a hybrid crop, it means that there is either no segregation or the segregation follows expected patterns (i.e., complex hybrids and OPVs). DNA fingerprinting also assumes that the samples are representative of the variety and that there is no genetic difference between seed or clone sources. It has been observed, however, that there can be significant variation in the uniformity of some reference samples. Haun *et al.* (2011) reported variation among individuals within some cultivars of soybean, and others have reported significant deviations from uniformity (impurities) in source breeder seed (Kilian 2019).

To confirm the purity of the seeds used as the reference variety, one option is to grow three to five seeds from the seed package belonging to each variety, then collect leaf tissue from each plant and genotype them individually or via the multiplex method. If the genotypes obtained reveal that the material in the seed package is heterogeneous, meaning the identity between pairs of samples from the same envelope is less than 95% identical (setting the threshold at the expected residual heterozygosity corresponding to the degree of inbreeding of the variety), discard the package and identify other seed sources deemed representative of that variety (and repeat this validation process). This procedure should not be applied to open-pollinated varieties or complex hybrids (e.g., three-way and double crosses) and crops that undergo bulk crossing such as maize, because the diversity present even in the reference materials would prevent convergence of the validation process.

4.3 Genotyping the Reference Library

The reference library should be genotyped with a preferred sequence-and-discover technology (e.g., GBS, tGBS, DArTseq, or WGS). Samples in the library could be multiplexed, but they should never be pooled or bulked unless bulked seed was used to create a variety (as is the case for maize). The average read depth per sample should be sufficiently high to result in accurate variant calls, especially for low-frequency variants that could be confounded with sample/reference heterogeneity and genotyping errors. While it is tempting to prescribe a certain minimum and average read depth, it is impossible to suggest one that fits all circumstances and species. In studies of maize landraces, a minimum of five to six reads is currently used in ongoing analyses with no significant deviation in data interpretation when higher baselines are used (Hearne 2019). The factors that must be considered are material heterogeneity (more heterogeneity requires more depth), the level of differentiation among references (tightly clustered references will require higher depth to distinguish), and the required level of accuracy and precision for any purity estimates.

Given how fast genotyping technologies are developing, for longer-term tracking of varietal change, it is imperative that fingerprinting evidence be developed in ways that aid the matching of variants identified using different (and continually evolving) genotyping technologies. When possible, a DNA bank of varieties should be maintained, enabling evaluation on future platforms, should that be desired. All sequences generated should be aligned to a genome reference if and when available (i.e., a genotype from the whole genome of a representative accession of a crop). Genome references are indexed (such that each site has a unique ID), which means that any variants discovered can be assigned a unique ID. Each genotyping technology detects different sets of SNPs, and some SNPs are detected by multiple technologies whereas others are detected in specific platforms. It is thus a challenge to know whether the same SNP has been found by two different platforms unless both have been properly cross-referenced to the reference genome.

Although this approach might aid in the use of multiple sources of data, it is limited to the genome reference used. The degree to which a variety's genome can be represented by one individual (referred to as the reference) depends on the degree of genetic diversity and the within-variety variation. A study of maize (Tenaillon *et al.* 2001) showed that in a comparison of two maize lines, maize has large nucleotide diversity (0.96 percent). This is close to the diversity between humans and chimpanzees (1.23 percent according to Mikkelsen *et al.* 2005). Moreover, even though crops like soybean have low nucleotide diversity (e.g., just 0.082 percent diversity according to a study by Zhu *et al.* 2003, they can have genome rearrangements and other structurally variable genetic features within one variety (i.e., heterogeneity). These challenges result in reference genomes that do not represent the entirety of species diversity and in some cases reference genomes that do not fully represent the genetic variation in other individuals from the same variety (Haun *et al.* 2011). Preserving the option of undertaking analyses beyond the initial fingerprinting exercise requires storing the plant material and DNA used for the creation of the reference library, the DNA from the material sampled from farmers'

fields, and the genetic data and associated metadata obtained from the initial fingerprinting exercise. With changing techniques and analytical platforms, the data being generated needs to be produced with an eye to the possibility that future researchers may be using different methods. In the Annex 2 we describe some of the more in-depth and potentially highly informative analyses beyond simply varietal identification made possible by tapping stored raw genetic data and the associated plant material.

5 DATA AND SAMPLE MANAGEMENT STRATEGIES

Tracking varietal change over time requires a data and sample management strategy as an integral part of the data generation procedures described above. It is costly and time consuming to develop and maintain representative reference libraries. Libraries are not one-off ventures; they need to be updated over time and made accessible in ways that respect any relevant intellectual property. This is especially important given that increasing shares (and in some countries, large shares) of the germplasm planted by farmers consists of proprietary varieties involving both public and private intellectual property. Genotyping such varieties for inclusion in reference libraries may contravene the rights conferred through issued plant patents or seed label restrictions, and so attention to managing these rights must also be an integral part of developing a reference library for any DNA fingerprinting purposes. The same data-sharing and intellectual property issues pertain to the varieties being collected and sequenced to monitor the farm-level use of these varieties.

Data and sample management strategies are also an imperative in modern science; most funding agencies and professional publications now insist on some form of data access and reproducibility standards. Equally important, as we describe in the Annex 2, a host of other important issues related to gene deployment strategies, beyond the uptake of new genetic material, could benefit from access to data accumulated over multiple fingerprinting exercises and spanning multiple time periods, agroecologies, and crop management scenarios. Such issues include the yield performance of these genes, pest resistance, and climate change analytics. Given the substantive sunk costs involved in collecting and sequencing material for varietal tracking purposes, the marginal costs of then storing this plant material, at least for some period of time, along with the associated data and metadata, are likely to realize large additional payoffs in terms of scientific understanding and the practical aspects of crop productivity.

Fortunately, solutions to these data and metadata creation and management problems related to data privacy or proprietary issues have been and are being developed. For example, the GEMS agroinformatics platform designed and developed by the University of Minnesota and its partners is structured to explicitly deal with these tricky data-stewarding and data-sharing problems. GEMS adheres to FAIR(ER) protocols (making data Findable, Accessible, Interoperable and Reusable, plus the even stricter standards of Ethical and Reproducible).[11] It also makes DNA fingerprinting data interoperable with others types of data (i.e., other genomic, environmental, management, and socioeconomic data that are made interoperable at varying spatial and temporal scales) and facilitates data analytics (including varietal impact assessments) along the entire varietal development, deployment, and stewardship chain. Another system is the Germinate database system developed by the James Hutton Institute. Germinate provides a FAIR-compliant platform for the storage, visualization, and integration of genomic, phenotypic, and spatial data such as climate data.

Another concern about the stewardship of the seed

[11] See Wilkinson *et al.* (2016) for a full description of the FAIR guiding principles for scientific data management and stewardship. For more details on GEMS, see www.agroinformatics.org.

(and other plant tissue) samples collected from farmers' fields, as well as the plant material used to develop the reference libraries, relates to changes in the technology of sequencing. As the discipline shifts toward portable, real-time single molecule technologies such as Oxford Nanopore (see Annex 1) that can perform sequencing in the field, new trade-offs arise. On the one hand, potential sources of error in the chain of custody to sequencing labs will be eliminated. On the other hand, enumerators may be tempted to discard samples for convenience, thus eliminating future analytical options that are possible when source material is conserved. Saving field samples and DNA from these samples—along with their associated data and metadata, at least for a period of time—serves to "future proof" past work and enable data interoperability with studies yet to be undertaken using new DNA fingerprinting strategies yet to be developed or fully deployed. Storing the extracted DNA is the best option for conserving the data for posterity (and potential reanalysis) as it is more stable than plant material.

6 CONCLUSION

Varietal identification using DNA fingerprinting is by no means free of challenges. There are numerous steps in the process from sample collection to genotyping that can substantively affect the results and reduce the accuracy of the varietal identification. In this report we outlined the most critical steps in deploying DNA fingerprinting methods to identify the diversity of crop varieties grown in farmers' fields. Rather than simply identify present "best scientific practice," we opted to identify the cost-benefit trade-offs involved in implementing fingerprinting processes from field to lab. We also offered a range of technical and practical options based on diverse scenarios that reflect differences in farmer germplasm use practices, the cost implications and logistical details of sampling and genotyping plant material, the errors associated with using this procedure for varietal identification, and the specific questions a DNA fingerprinting exercise is designed to address.

Given these varying factors, and the technical and practical trade-offs they imply, a one-size-fits-all approach to DNA fingerprinting to assess varietal use in farmers' fields is neither cost-effective nor practical. Different approaches will be preferable in different circumstances, and we sought to inform those choices in this report. That said, a number of technical matters undercut the value of DNA fingerprinting efforts, irrespective of the method of choice, and we highlighted those aspects in detail. Finally, the challenges in distinguishing varieties properly are likely to increase as seed systems and breeding technologies evolve and the introduction of closely related but distinct varieties onto the market becomes more commonplace (e.g., gene-edited varieties that differ from an existing variety by only one base).

REFERENCES

Barnaud, A., Trigueros, G., McKey, D., and Joly, H.I. (2008). High outcrossing rates in fields with mixed sorghum landraces: How are landraces maintained? *Heredity 101*(5): 445–452. https://doi.org/10.1038/hdy.2008.77

Bhat, K.V. (2008). DNA fingerprinting and cultivar identification. PhD thesis, University of the Free State, Bloemfontein, South Africa, 101–109. http://www.iasri.res.in/design/ebook/EBADAT/6-Other Useful Techniques/9-DNA Fingerprinting-IASRI-KVB.pdf.

Bøhn, B.M., Espinoza, L.C., Banda, A.E., De La Fuente Martínez, J.M., Tiznado, J.A.G., García, J.G., … García, F.Z. (2015). Pollen-mediated gene flow in maize: Implications for isolation requirements and coexistence in Mexico, the center of origin of maize. *PLoS ONE 10*(7): 1–12. https://doi.org/10.1371/journal.pone.0131549.

Buś, M.M., and Allen M. (2014). Collecting and preserving biological samples from challenging environments for DNA analysis. *Biopreservation and Biobanking 12*(1): 17–22

Cao, C. Chang, and Sun, X. (2016). Combinatorial pooled sequencing: Experiment design and decoding. *Quantitative Biology 4*(1): 36–46. https://doi.org/10.1007/s40484-016-0064-3

Chakravarthi, B.K., and Naravaneni, R. (2006). SSR marker based DNA fingerprinting and diversity study in rice (Oryza sativa. L). *African Journal of Biotechnology 5*(9): 684–688.

Costa, L.M., Marshall, E., Tesfaye, M., Silverstein, K.A., Mori, M., Umetsu, Y., Otterbach, S.L., Papareddy, R., Dickinson, H.G., Boutiller, K., VandenBosch, K.A., Ohki, S., and Gutierrez-Marcos, J.F. (2014) Central cell-derived peptides regulate early embryo patterning in flowering plants. *Science 244*(6180): 168–172. https://doi.org/10.1126/science.1243005.

DArT. (2019). Targeted Genotyping. Canberra: Diversity Arrays Technology. https://www.diversityarrays.com/technology-and-resources/targeted-genotyping/.

Day-Williams, A.G., McLay, K., Drury, E., Edkins, S., Coffey, A.J., Palotie, A., and Zeggini, E. (2011). An evaluation of different target enrichment methods in pooled sequencing designs for complex disease association studies. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0026279.

Douzery, E.J.P., Gissi, C., and Mastrototaro, F. (2013). Deep sequencing of mixed total DNA without barcodes mitochondrial genomes. *Genome Biology and Evolution 5*(6): 1185–1199. https://doi.org/10.1093/gbe/evt081.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5): 1–12. https://doi.org/10.1371/journal.pone.0019379.

Floro, V.O., Labarta, R.A., Becerra López-Lavalle, L.A., Martinez, J.M., and Ovalle, T.M. (2017). Household determinants of the adoption of improved cassava varieties using DNA fingerprinting to identify varieties in farmer fields: A case study in Colombia. *Journal of Agricultural Economics 69*(2): 518–536. https://doi.org/10.1111/1477-9552.12247.

Fung, T., and Keenan, K. (2014). Confidence intervals for population allele frequencies: The general case of sampling from a finite diploid population of any size. *PLoS ONE*, *9*(1), e85925. https://doi.org/10.1371/journal.pone.0085925.

Guo, Y., Cai, Q., Li, C., Li, J., Li, C.I., Courtney, R., … and Long, J. (2013). An evaluation of allele frequency estimation accuracy using pooled sequencing data. *International Journal of Computational Biology and Drug Design 6*(4): 279. https://doi.org/10.1504/IJCBDD.2013.056709.

Habernicht, D.K., and Blake, T.K. (1999). Application of PCR to detect varietal purity in barley malt. *Journal of the American Society of Brewing Chemists 57*(2): 64–71. https://doi.org/10.1094/ASBCJ-57-0064.

Haun, W.J., Hyten, D.L., Xu, W.W., Gerhardt, D.J., Albert, T.J., Richmond, T., … Stupar, R.M. (2011). The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiology 155*(2): 645–655. https://doi.org/10.1104/pp.110.166736.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science 5*: 1–8. https://doi.org/10.3389/fpls.2014.00484.

Hearne, S. (2019). Personal communication. El Batan, Mexico: CIMMYT.

Girma, G., Rabbi, I., Olanrewaju, A., Alaba, O., Kulakow, P., Alabi, T., Ayedun, B., Abdoulaye, T., Wossen, T., Alena, A., and Manyong, V. (2017). *A manual for large-scale sample collection, preservation, tracking, DNA extraction, and variety identification analysis*. Ibadan, Nigeria: International Institute of Tropical Agriculture (IITA). pp. i-32. ISBN: 978-978-8444-83-1. https://hdl.handle.net/10568/80560).

Ilukor, J., Kilic, T., and Moylan, H. (n.d.). Should DNA fingerprinting be part of adoption and impact studies of crop technologies? Results from the Cassava Variety Identification Experiment in Malawi, 1–30. World Bank Living Standards Measurement Study, unpublished manuscript.

Johnston, S.A., den Nils, T.P., Peloquin, S.J. and Hanneman, R.E., Jr. (1980). The significance of genic balance to endosperm development in interspecific crosses. *Theoretical and Applied Genetics 57*(1): 5–9. doi: 10.1007/BF00276002.

Kilian, A. (2019). Personal communication. Canberra: Diversity Arrays Technology.

Kilic, T. (n.d.). Blowing in the wind: The quest for accurate crop variety identification in field research, with an application to maize in Uganda. World Bank, Washington, DC. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=CSAE2018&paper_id=697.

Kosmowski, F., Aragaw, A., Kilian, A., Ambel, A., Ilukor, J., Yigezu, B., and Stevenson, J. (2016). *Varietal identification in household surveys results from an experiment using DNA fingerprinting of sweet potato leaves in southern Ethiopia*. Policy Research Working Paper 7812. Washington, DC: World Bank.

Lawson, D., Hellenthal, G.S.M., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics 8*(1): e1002453. https://doi.org/10.1371/journal.pgen.1002453.

Le, D. P., Labarta, R.A. and Maredia, M.K. (2017). Analysis of cassava varietal adoption in Vietnam using DNA fingerprinting approach. Cali, Colombia: International Center for Tropical Agriculture (CIAT); East Lansing: Michigan State University. http://veam.org/wp-content/uploads/2017/12/100.-Phuong-Dung-Le.pdf.

LGC Genomics. (2017). Plant Sample Collection Kit. www.lgcgroup.com/LGCGroup/media/PDFs/services/Genotyping/Plant-leaf-kit.pdf.

Maredia, M.K., Reyes, B.A., Manu-Aduening, J., Dankyi, A., Hamazakaza, P., Muimui, K., and Raatz, B. (2016). Testing alternative methods of varietal identification using DNA fingerprinting: Results of pilot studies in Ghana and Zambia. MSU International Development Working Paper No. 149. East Lansing: Michigan State University.

Meibusch, P. (2013). New quality assurance for wheat and barley. *GroundCover 105* (July–August). https://grdc.com.au/resources-and-publications/groundcover/ground-cover-issue-105-july-august-2013/new-quality-assurance-for-wheat-and-barley.

Mikkelsen, T.S., Hillier, L.W., Eichler, E.E., Zody, M.C., Jaffe, D.B., Yang, S.P., … Waterston, R.H. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature 437*(7055): 69–87. https://doi.org/10.1038/nature04072.

Mkondiwa, M., Pardey, P.G., and Chan-Kang, C. (2020). A meta-review of the conventional crop varietal use evidence for sub-Saharan Africa, 1988–2016. St Paul: International Science and Technology Practice and Policy (InSTePP) Center, University of Minnesota (in preparation).

Ott, A., Liu, S., Schnable, J., Yeh, C.-T., Wang, K.-S., and Schnable, P.S. (2017). tGBS® genotyping-by-sequencing enables reliable genotyping of heterozygous loci. *Nucleic Acids Research 45*(21): e178. https://doi.org/10.1093/nar/gkx853.

Pardey, P., Joglekar, A., Chan-Kang, C., Liebenberg, F., Luby, I., Senay, S., Azzarri, C. and Mkondiwa, M. (2020). What do we know about (procured) input use in African agriculture?: Supporting information. InSTePP Working Paper. St. Paul: International Science and Technology Practice and Policy center.

Patterson, N., and Gabriel, S. (2009). Combinatorics and next-generation sequencing. *Nature Biotechnology 27*(9): 826–827. https://doi.org/10.1038/nbt0909-826.

Poets, A.M., Fang, Z., Clegg, M.T., and Morrell, P.L. (2015). Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome Biology 16*(1): 1–11. https://doi.org/10.1186/s13059-015-0712-3.

Rabbi, I.Y., Kulakow, P.A., Manu-Aduening, J.A., Dankyi, A.A., Asibuo, J.Y., Parkes, E.Y., and Maredia, M.K. (2015). Tracking crop varieties using genotyping-by-sequencing markers: A case study using cassava (*Manihot esculenta* Crantz). *BMC Genetics 16*(1): 115. https://doi.org/10.1186/s12863-015-0273-1

Radchuk, V., Weier, D., Radchuk, R, Weschke, W., and Weber, H. (2011). Development of maternal seed tissue in barley is mediated by regulated cell expansion and cell disintegration and coordinated with endosperm growth. *Journal of Experimental Botany 62*(3): 1217–1227. https://doi.org/10.1093/jxb/erq348.

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., and Fischer, M.C. (2013). Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS ONE 8*(11): e80422. https://doi.org/10.1371/journal.pone.0080422.

Smith, A.M., Heisler, L.E., St.Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., and Nislow, C. (2010). Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples. *Nucleic Acids Research 38*(13): 1–7. https://doi.org/10.1093/nar/gkq368.

Smith, J.S.C., and Register, J.C., III. (1998). Genetic purity and testing technologies for seed quality: A company perspective. *Seed Science Research 8*(2): 285–294. doi:10.1017/S0960258500004189.

Stapley, J., Feulner, P.G.D., Johnston, S.E., Santure, A.W., and Smadja, C.M. (2017). Variation in recombination frequency and distribution across eukaryotes: Patterns and processes. *Philosophical Transaction of the Royal Society* B, 372 (1736): 20160455.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F., and Gaut, B.S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). *Proceedings of the National Academy of Sciences 98*(16): 9161–9166. https://doi.org/10.1073/pnas.151244298.

Traxler, G., Tizale, C.Y., Kim, M., and Alemu, D. (2015). Using DNA fingerprinting to estimate the diffusion of improved crop varieties in Ethiopia. Presentation at the 29th International Conference of Agricultural Economists, Milan, Italy, August 9–14, 2015. http://ru-ftf.rutgers.edu/Outputs%20for%20webpage/Traxler%20ICAE%20Milan%202015.pdf.

Warburton, M.L., Setimela, P., Franco, J., Cordova, H., Pixley, K., Bänziger, M., … Macrobert, J. (2010). Toward a cost-effective fingerprinting methodology to distinguish maize open-pollinated varieties. *Crop Science 50*(2): 467–477. doi: 10.2135/cropsci2009.02.0089.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology 7*(2): 256–276.

Wilkinson. M.D., Dumontier, M……Mons, B. (2016). Comment: The FAIR guiding principles for scientific data management and stewardship. *Nature: Scientific Data* 3, article 160018. DOI: 10.1038/sdata.2016.18.

Wossen, T., Abdoulaye, T., Alene, A., Nguimkeu, P., Feleke, S., Rabbi, I.Y., … Manyong, V. (2019). Estimating the productivity impacts of technology adoption in the presence of misclassification. *American Journal of Agricultural Economics 101*(1): 1–16.

Yan, D., Duermeyer, L, Leoveanu, C., and Nambara, E. (2014). The functions of the endosperm during seed germination. *Plant and Cell Physiology 55*(9): 1521–1533. https://doi.org/10.1093/pcp/pcu089.

Yirga, C., Alemu, D., Oruko, L., Negisho, K., and Traxler, G. (2016). T*racking the Diffusion of Crop Varieties Using DNA Fingerprinting Crop Varieties*. Technical Report. Addis Ababa: Ethiopian Institute of Agricultural Research. https://www.researchgate.net/publication/303913826_Tracking_the_Diffusion_of_Crop_Varieties_Using_DNA_Fingerprinting.

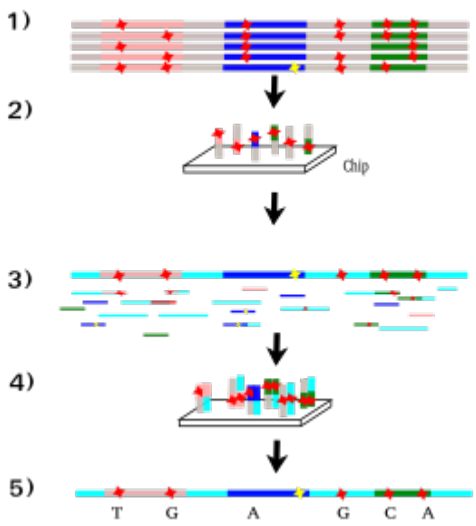
Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., … Cregan, P.B. (2003). Single-nucleotide polymorphisms in soybean. *Genetics 163*(3): 1123–1134. https://doi.org/10.1534/genetics.105.043877.

Zhu, Y., Hu, J., Han, R., Wang, Y., and Zhu, S. (2011). Fingerprinting and identification of closely related wheat (Triticum aestivum L.) cultivars using ISSR and fluorescence-labeled TP-M13-SSR markers. *Australian Journal of Crop Science 5*(7): 846–850.

ANNEX 1: A PRIMER ON DNA SEQUENCING TECHNOLOGIES

There are a number of technologies that can be used for genotyping. Here we describe five of the reduced-representation technologies (i.e., SNP arrays and GBS) as well as whole genome sequencing methods.

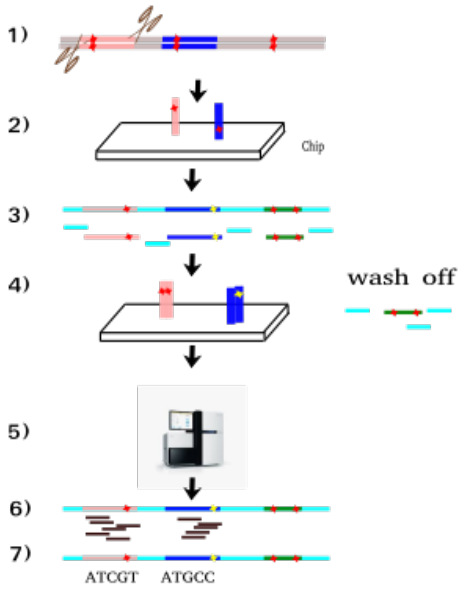
SNP Arrays



SNP arrays involve two major activities: development of the array (steps 1 and 2) and genotyping of the samples (steps 3 to 5). The development of a SNP array requires a "discovery panel" of known genotypes that are deemed representative of the diversity expected in the varieties planted in farmers' fields (step 1). In panel A, the gray (or cyan, in steps 3 to 5) bars represent the non-coding parts of a small segment of a genome. Pink, blue, and green bars are coding portions of a gene that translate to proteins. Genetic variants (SNPs) are represented as red stars (common among the panel) and yellow stars (observed in only a couple of individuals and thus constituting a rare variant). For common variants in the panel, small fragments (probes) of DNA containing 60–125 base pairs on each side of the variant are affixed to the array (step 2).

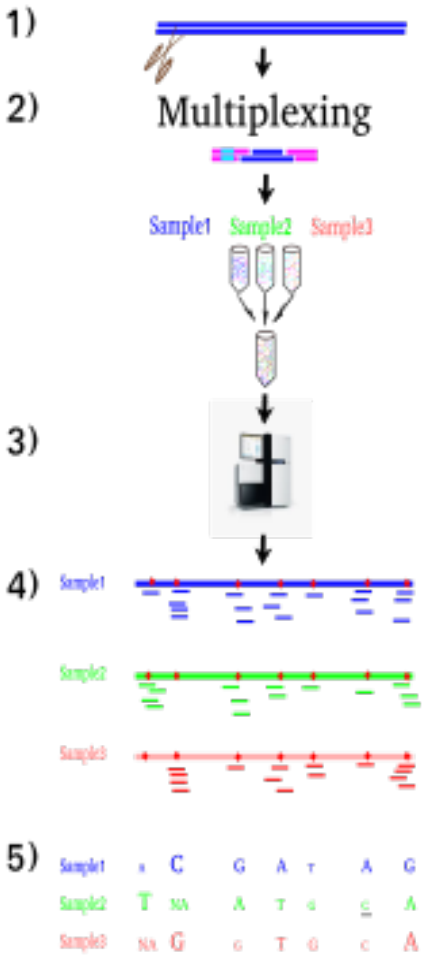
The array (or SNP chip), which is a DNA-coated glass surface (slide or bead), will only be useful for genotyping the variants that are included on the chip, which are restricted to the diversity in the discovery panel (thus raising the possibility of ascertainment bias if genetic variants in the material found in the field are not part of those in the discovery panel). Arrays can be made from low hundreds to millions of variants. Once the array has been constructed, it can be used to genotype new individuals (e.g., samples from a farmer's field). The genotyping procedure requires fragmenting the DNA of a sample (step 3). Then, the DNA fragments that match the features in the array are detected with a system based on fluorescent dye signals; this system records the matched segment and interprets which allele of a variant the sample carries (step 4). SNPs are a type of variant that have two possible alleles (nucleotides). Note that the rare variant (yellow star) in the discovery panel was not included in the array. Therefore the same variant (yellow star) in the assayed sample (in step 3) was not scanned for (in step 4), and hence not detected (absent red star in step 5).

Exome Capture



Exome capture is a technique for sequencing most of the protein-coding portion (exome) of genes in a genome. Instead of using a discovery panel as in SNP arrays, an array with exomes is built from exomes detected in a single individual (using specific restriction enzymes) (step 1). The exome fragments or probes (which can be short, tiled DNA sequences matching the genome or thousands of base pairs long) are affixed to an array (step 2) or used in solution. DNA from a sample is fragmented (step 3), and genomic DNA fragments are hybridized to the probes; fragments that do not match with any probe are washed away (step 4). (Note that if the sample being assayed contains novel structural regions that were absent from the reference genome used to create the exome capture array, those regions will be washed away—and hence lost to the assay.) The next step is to sequence the hybridized exomes using a high-throughput sequencer machine (step 5), which generates millions of genetic sequences that individually are referred to as reads. Computational tools are used to determine the exome to which each of those sequences corresponds (step 6) and to identify genetic variants (step 7). In contrast to SNP arrays, the exome capture approach does not require knowing where variants occur; it discovers variants from the generated data, thereby lowering the effects of ascertainment bias inherent in SNPs arrays. However, bias might be introduced by mutation at enzyme cut sites in the reference accession used to create the array, which would prevent the enzyme from recognizing a portion of the coding material in the genome. The methods sketched below reduce the effects of ascertainment bias even further.

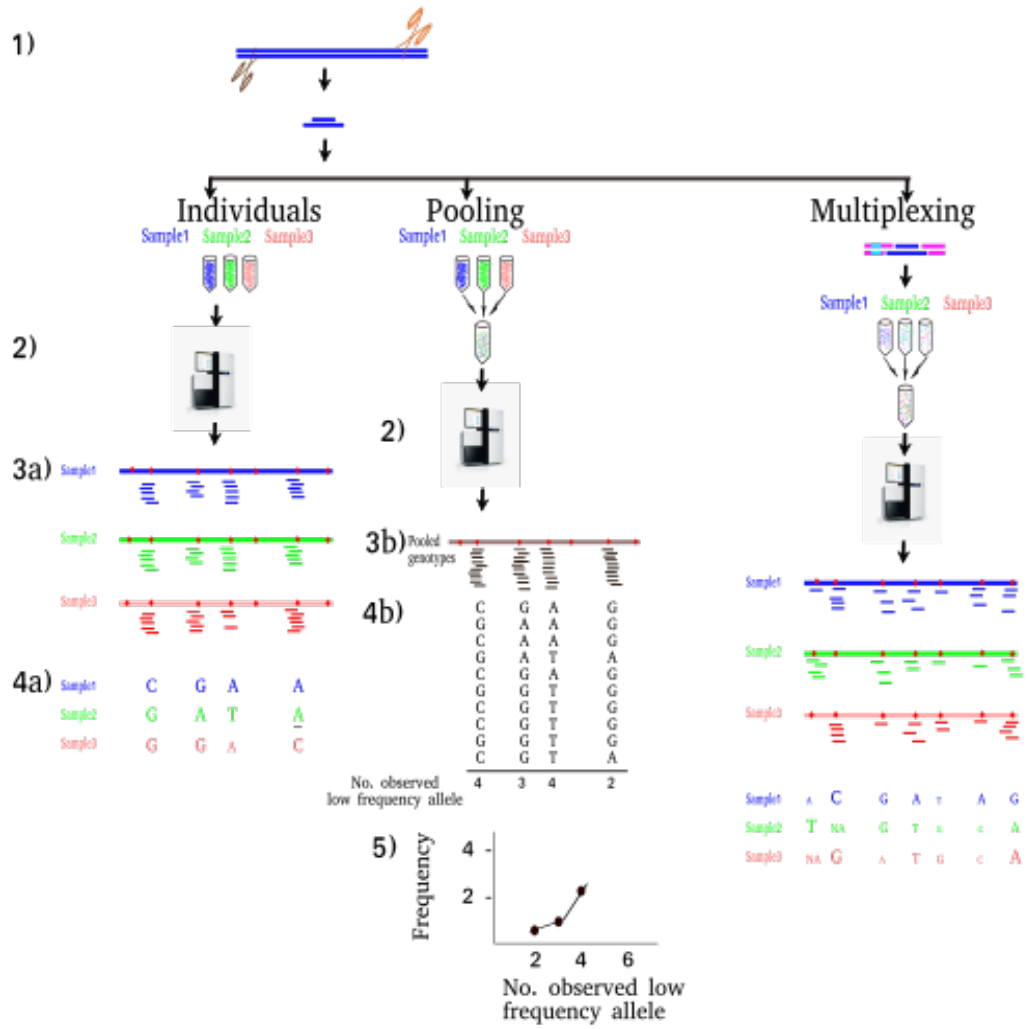
Genotype by Sequencing (GBS)



Genotyping by sequencing (GBS) uses restriction enzymes to reduce genome complexity (step 1) and makes it possible to genotype a larger portion of the genome in multiple samples (step 2) at a cost comparable to SNP arrays. Each sample is digested using one or two restriction enzymes that recognize specific genetic patterns that are known to be present throughout the genome. To be able to preserve the identity of each sample but still process multiple samples at once (multiplexing), samples are labeled with unique DNA barcodes (step 2, marked with a small cyan bar in the figure) and ligases (purple). Fragments from each sample are later amplified using a PCR (polymerase chain reaction). Similar amounts of DNA per sample are combined (step 2) and sequenced using high-throughput sequencing machines (step 3).

Using computational tools and the DNA barcodes, fragments can then be separated by sample and organized to identify their position in the genome (step 4). After the fragments have been ordered, variants can be identified. The accuracy of each variant call that is identified depends largely on the number of sequences stacked up over the same aligned position that provided the information (read depth) and the quality of each sequenced read. Therefore, the accuracy of the identified variants varies (higher accuracy is represented with a larger font size in the pictogram for variants with deep coverage) (step 5). By the nature of this method, some restriction sites might be absent or mutated and unrecognizable by the restriction enzymes, or fragments might fail to be amplified during PCR, resulting in unrepresented genomic segments and leading to missing data points across samples, which is a characteristic of GBS data.

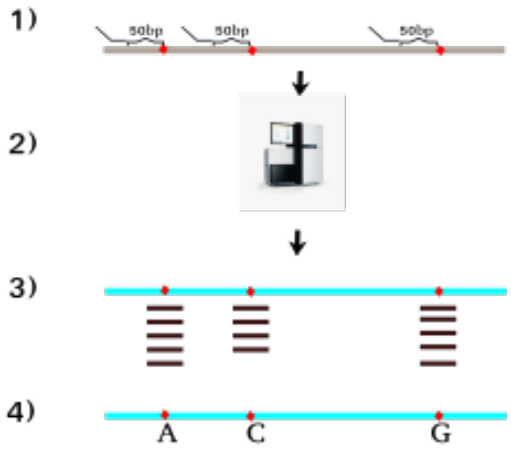
DArTseq



DArTseq is a type of GBS technology that, compared with GBS, uses an additional set of restriction enzymes that cleave off repetitive DNA. The remaining fragmented DNA per sample is sequenced from either a mixture of equal amounts of DNA from each sample (pooling method or multiplexing) or separated per each individual (step 2). The double reduction of genome complexity results in fewer fragments that are easier to sequence, allowing more resources to be invested in higher read depth per site (step 3a), which in turn increases the accuracy of variant identification (step 4b). In the pictorial example the last variant in sample 2 (step 4a) has been assigned to A (marked with an underscore) with high confidence owing to a deeper sequencing; the same variant was assigned incorrectly to C using GBS with a single read at the site.

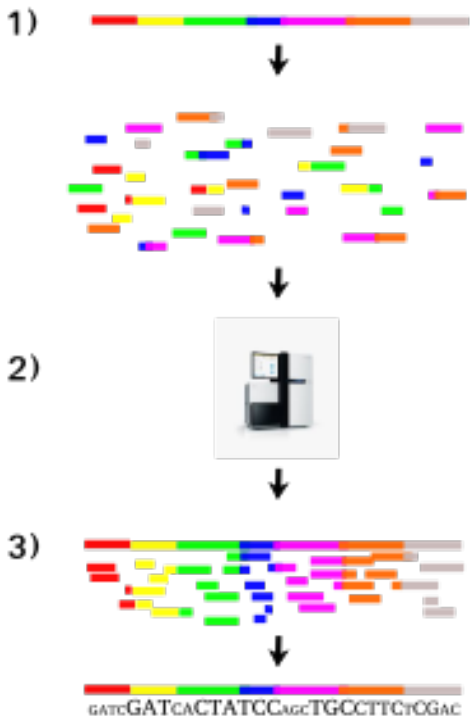
The double reduction of genome complexity also reduces the genome to the portion that is more conserved across samples, resulting in fewer missing data points across the set (step 4a). When the pooling method is used, sequenced fragments are often not sorted by the sample they came from; rather, with really deep coverage the frequency of each allele (A, C, T, or G) can be estimated (step 4b) and a profile of allele frequencies can be made (step 5). The profile is then compared with similar profiles in the reference library.

Targeted amplicon sequencing (TAS)



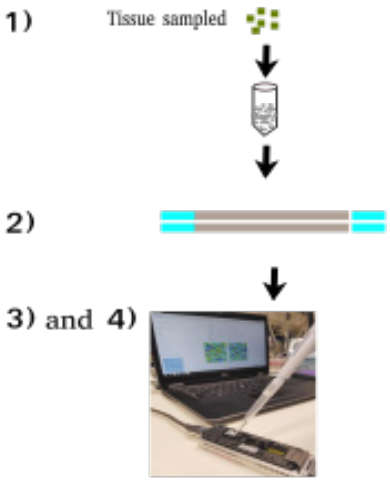
Targeted amplicon sequencing (TAS) is a technique focused on pieces of DNA amplified in PCR reactions of specific genes. This method combines the benefits of SNP arrays (i.e., restricting efforts to targeted genomic regions) with the advantages of high-throughput sequencing machines (i.e., high volumes of data at reduced cost). Oligonucleotide probes are designed to target and capture regions of interest (step 1). Probes are designed based on preconceived knowledge of the genetic sequence preceding the targeted region; this information can be obtained from other sequencing efforts such as WGS, GBS, DArTseq, or exome capture. The regions captured by the probes are then sequenced at high coverage (step 2). For each sample subjected to this approach, a series of sequenced fragments will be obtained (step 3). The fragments are then aligned to match the targeted regions, and variants are identified (step 4). The advantages of this approach are the high coverage, reduced sequencing cost and time, and the ability to genotype even difficult-to-sequence areas.

Whole genome sequencing (WGS)



Whole genome sequencing (WGS) entails sequencing not just parts but all of the chromosomal content of an organism. First, the entire DNA is fragmented (step 1). Then, using a high-throughput sequencing approach, all the fragments are sequenced (step 2). There are enough reads to cover the entire genome, but some regions may have or typically have low coverage and other regions may have or typically have high coverage, resulting in low- and high-accuracy base calls, respectively (step 3).

Oxford nanopore



Oxford nanopore is another WGS approach. It involves extracting DNA (represented in gray) from the tissue sample (step 1) and preparing a sequence library by attaching sequencing adapters (cyan) using the Oxford nanopore's easy-to-use kit (step 2). The prepared DNA is then loaded onto a MinION flow cell (step 3), which is directly connected to and powered by a computer via USB. The MinION contains a flow cell that has sensors to detect the nanopore signal as the molecule of DNA is analyzed, producing sequencing data. The nanopore signal is immediately processed by MinKNOW, which is the instrument's control software that provides real-time feedback on the process and controls parameters in the workflow. Analysis can start immediately, or data can be stored (step 4) to be analyzed later.

This technology is distinguished from most WGS technologies in that sequencing can be done in the field and individual sequence reads are extremely long (tens of thousands to hundreds of thousands of base pairs for conventional sequencing). Base pair accuracy has been significantly lower than traditional sequencing but has recently improved significantly owing to applications of artificial intelligence to the signal.

ANNEX 2: STORING DATA, DNA, OR PLANT TISSUE

In some cases, researchers may wish to reevaluate existing samples. The future repurposing of DNA fingerprinting data requires access to the raw genotyping data collected, a detailed description of how those data were processed before analysis, and the metadata of each sample in the collection. Here we present six case scenarios when stored raw genotypes, DNA, or plant tissue are required for additional analyses.

Stored Genotype Data

1. New genotyping efforts to update the reference library identified two or more seed sources (each from a homogeneous seed in an envelope) matching one named variety. This is a potential outcome when several seeds are selected from an early stage of a developing line; each seed may carry high levels of heterogeneity that will segregate and eventually fix at different loci. Thus, each seed has the potential to develop a population with different genotypes. This was observed for Williams 82, which was the accession used for the development of a reference genome in soybean (Haun *et al.* 2011). This phenomenon was discovered only after a large volume of genetic variants, including genome rearrangements and other structural variation (e.g., presence or absence of genomic segments), were obtained for each of the seeds derived from the same cross that gave rise to Williams 82. The storage of genotyping data for samples used for the reference library for DNA fingerprinting studies can aid the discovery of varieties maintained and used for crop improvement at different geographic locations or institutions under the assumption of a homozygous seed source. Note that this is a different issue from having heterogeneous seed within an envelope of seeds; we suggest testing those seeds in the field before a variety is added to the reference library. In the case discussed here, each institution might hold a different version of the same variety in an envelope of homozygous seeds; for the purposes of DNA fingerprinting, all versions of the variety should be added to the reference library.

2. The reference genome used to standardize variant names across platforms is not the same as the new reference available. In this case, raw genetic sequence data can be aligned to the new reference, and a new ID can be assigned to each variant with a direct mapping created to corresponding variants from other platforms. The probes used for SNP arrays and TAS can be matched to the new reference genome to assign new IDs.

3. The reference genome used to assemble and discover variants does not represent the variation in the species; therefore important genetic variation might be missing.
   The ideal reference genome will represent the genomic content of the species it represents. The example of Williams 82 illustrates not only the problem of heterozygosity in a sample but also the risky assumption that one sample can represent an entire species. In the case of soybean and Williams 82, the variation was not only at the nucleotide level, where one might expect significant variation within the species (not the case for soybean), but also in the structure of the genome itself. The reference version of Williams 82

demonstrated significant segmental variation in one version of Williams 82 relative to the other versions. This might not alter the classification of improved and unimproved varieties or the determination of the number of improved varieties, but it might hide information about specific adaptive genetic structures in a given variety. Therefore, one might be interested in using the raw sequence data to assemble it de novo without using a reference genome to identify genetic variants absent from the genome reference.

Stored Plant Tissue

1. First genotyping analysis was done using the pooling method and major varieties were identified, and now we want to discriminate between heterozygosity in a sample and heterogeneity in the field.

2. Using the pooling method (i.e., combining DNA) or bulking from a sample collected as individuals while storing tissue from each individual can reduce the cost of genotyping while allowing the possibility of further analysis. A new analysis to discriminate between a heterozygous sample and heterogeneous fields requires identifying the genetic composition of each sample collected. Therefore, if plant tissue was stored for individual samples, it could be genotyped again by either multiplexing or genotyping each accession individually.

3. A small set of variants were genotyped (e.g., using SNP arrays or TAS), varieties were discriminated, and heterozygosity was found. Now there is interest in determining the proportion and genetic location of parental sources on heterozygous samples.

   Although it may be possible to determine which two or more samples gave rise to a heterozygous individual, the value of this information is limited. It is more important to know what proportion of each variety is present in the sample, because this information will give an idea of the varieties that farmers are using to perform crosses (even if they are unconscious choices). The genetic location of each variety used in a cross can indicate the importance of that segment in the cross (if also observed in other samples). The resolution to which one can determine where each variety contributed depends on the number of variants used, which is in turn determined by the genome size and the diversity of the crop. Highly diverse crops and crops with larger genomes will require more markers. Studies to determine the origin of each portion of the genome are referred to as "chromosomal painting" (e.g., Poets *et al.* 2015; Lawson *et al.* 2012). Having a large number of variants can also help determine the timing of the events. Genetic transmission theory states that variants that segregate together (haplotype) have been inherited together. Moreover, as the generations pass those haplotypes are broken by a process at the chromosomal level called recombination. Therefore long co-segregating variants are indicators of recent events while tracks that are short suggest that the event occurred long ago. Population genetic studies using haplotype data and in the absence of phenotypic data can target potential sources of selection or putative candidate genes for adaptation to environments where farmers are cultivating. This can be a good source of information for plant breeders. Having said that, to archive the level of resolution optimal for this sort of studies, it might be necessary to re-genotype some of the tissue collected with a different platform.

4. Farmers' plants might not be in the reference library or might be a known variety but carry important adaptive variation that can be used for genetic improvements.

   Although landrace collections for the major world crops are available, it is likely that there are new landraces in farmers' fields that do not match any of the entries in the reference library. It would be valuable to retain this genetic material, cognizant of any prevailing biodiversity protocols or regulations.

Stored DNA

DNA storage provides a way to uplift and repurpose older survey efforts as new genotyping technologies become available. In 15 years' time, for example, surveys will inevitably use a much more advanced system of genotyping, generating more data with high accuracy. To empirically compare survey results it will be desirable to genotype old materials on this new platform. DNA is the best resource to store and reuse for these purposes. Unlike plant tissue, DNA takes little space and can be stored in appropriate conditions for decades with no appreciable deterioration in quality. Plant tissue requires much larger storage infrastructure, and even with good storage the quality of DNA present in tissue degrades over time. DNA is thus a useful long-term resource for future analysis.

Standing
Panel on
Impact
Assessment

CGIAR Advisory Services (CAS) Secretariat
Via dei Tre Denari, 472/a, Maccarese (Fiumicino), Italy
tel: (39-06) 61181 - email: cas@cgiar.org
https://cas.cgiar.org/