Photo: G. Smith/The Alliance

CGIAR | Standing Panel on Impact Assessment

DNA Fingerprinting and Bioinformatics

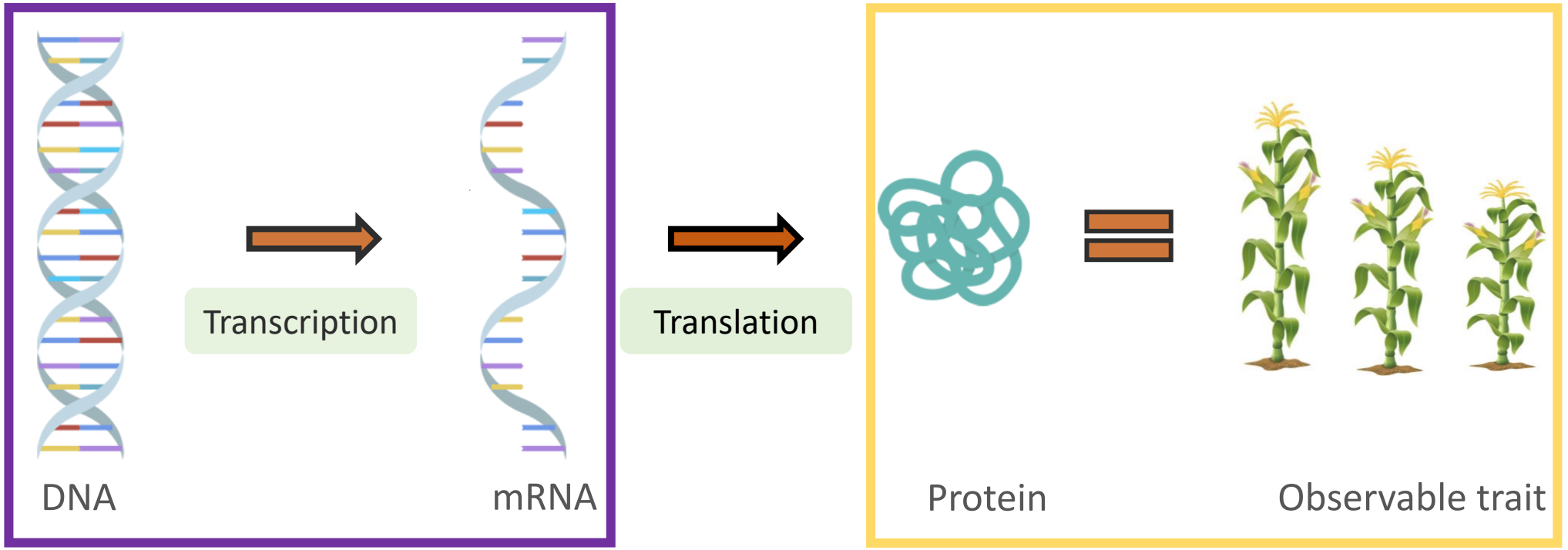
Davis Gimode
Bioinformatics Research Specialist, SPIA Country Studies

Webinar #4, 28th February 2025

Definitions

Central dogma of molecular biology

- Genetic information flows *typically* in one direction, from DNA, to RNA, to protein -Francis Crick



Genotype

Phenotype

- Fingerprinting: Distinguishing between individuals by the unique characteristics of their genotype

DNA Fingerprinting for adoption tracking

CGIAR Standing Panel on Impact Assessment

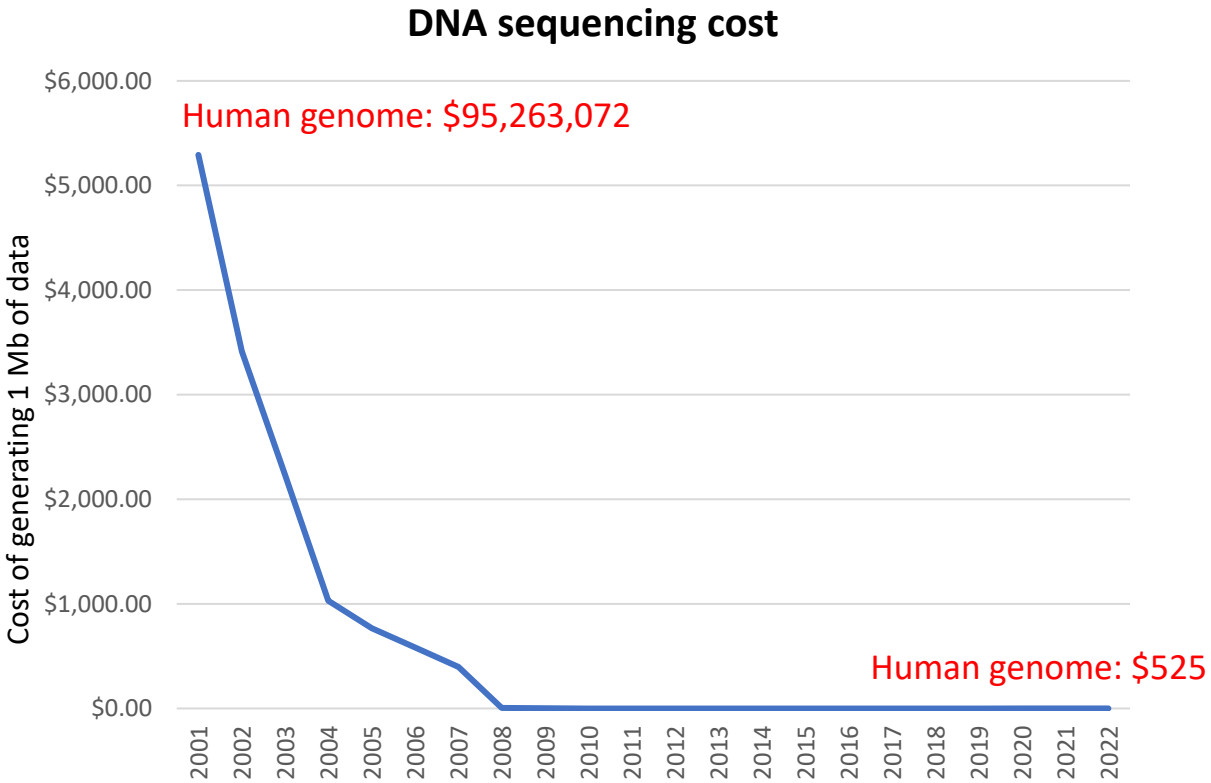
- Traditionally adoption tracking relies on farmers' self-reporting in household surveys
- Increasingly evident that farmer self-reporting may be unreliable
  - Inability of farmers to correctly identify varieties used
  - Inconsistency between farmer names and official records
  - Loss of genetic identity

    - Floro IV et al., 2018; Wineman et al., 2020
- Hence the necessity of using DNA fingerprinting to accurately identify varieties in adoption tracking studies

- Fingerprinting: Distinguishing between individuals by the unique characteristics of their genotype

DNA Fingerprinting for adoption tracking

- Incorporation of DNA fingerprinting to improve accuracy is becoming mainstream
- Motivated by drastic decrease in cost of generating genome scale data
- Evidenced by increasing numbers of adoption tracking literature that include DNA fingerprinting



- Fingerprinting: Distinguishing between individuals by the unique characteristics of their genotype

What it entails

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

- Fingerprinting: Distinguishing between individuals by the unique characteristics of their genotype

Reference library compilation

What is a variety?

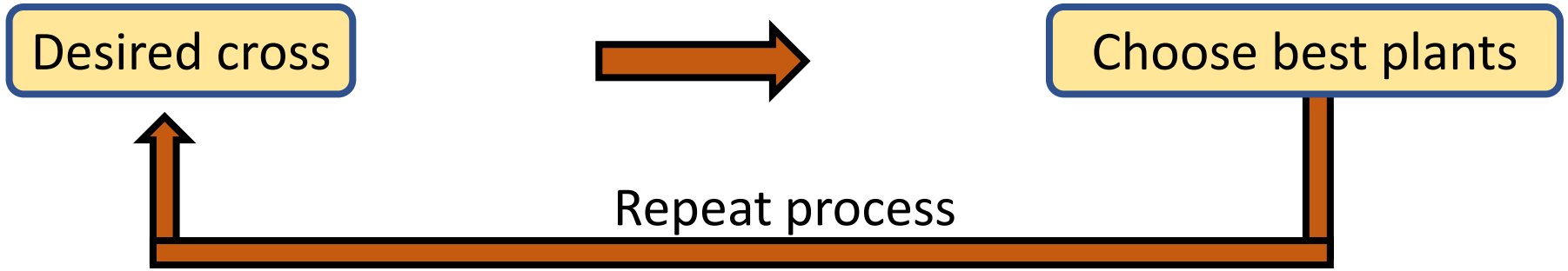
Modern variety

×

Landrace/Wild

=

Field evaluation



Reference library compilation

What is a variety?

Old variety

Improved variety

Reference library compilation

Important considerations

- This is the most important step since samples can not be identified without a good reference library
- The reference library should be:
  - Complete
    - The level of completeness is determined by the purpose of the study
  - Distinct
    - Individual varieties should be sufficiently differentiated from each other
  - Pure
    - Varieties should not be contaminated through outcrossing or mixing with other varieties

Improved variety

Reference library compilation

Obtaining reference material

- The best source for reference material is breeder seed
- This is the stock of genetically pure material that is maintained by a single institutional owner within a country
- Breeder seed and information on varieties can be obtained from:
  - International Agricultural Research Centers (IARCs)
  - National Agricultural Research Centers (NARCs)
  - Variety release committees
  - Seed multiplication agencies
  - Farmer's seed cooperatives
  - Private companies

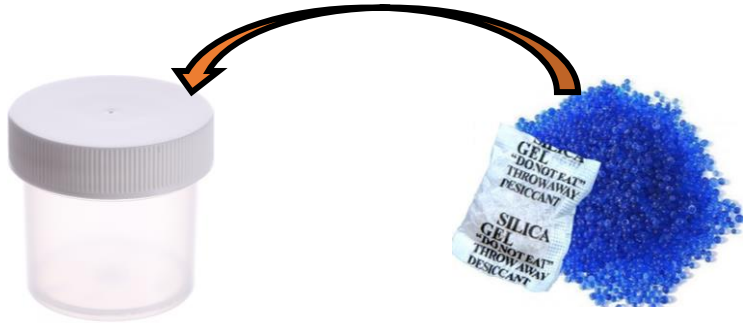
Improved variety

Sample collection

- Sample collection happens in the course of the Household survey
- Enumerators should be trained on the process before hand
- Enumerators should be familiar with field sampling manual

Some of the things to prepare in advance

Barcoded labels



Collection tubes

Silica gel

Device with barcode reader

Leaf punch

Alcohol wipes

Sample collection

Some key considerations

- Two main factors inform the nature of samples to be collected
  - Reproductive strategy of the crop
  - Expectation of homogeneity vs heterogeneity in the fields

Clonals

Cassava

Sweet potato

Inbreds

Groundnut

Bean

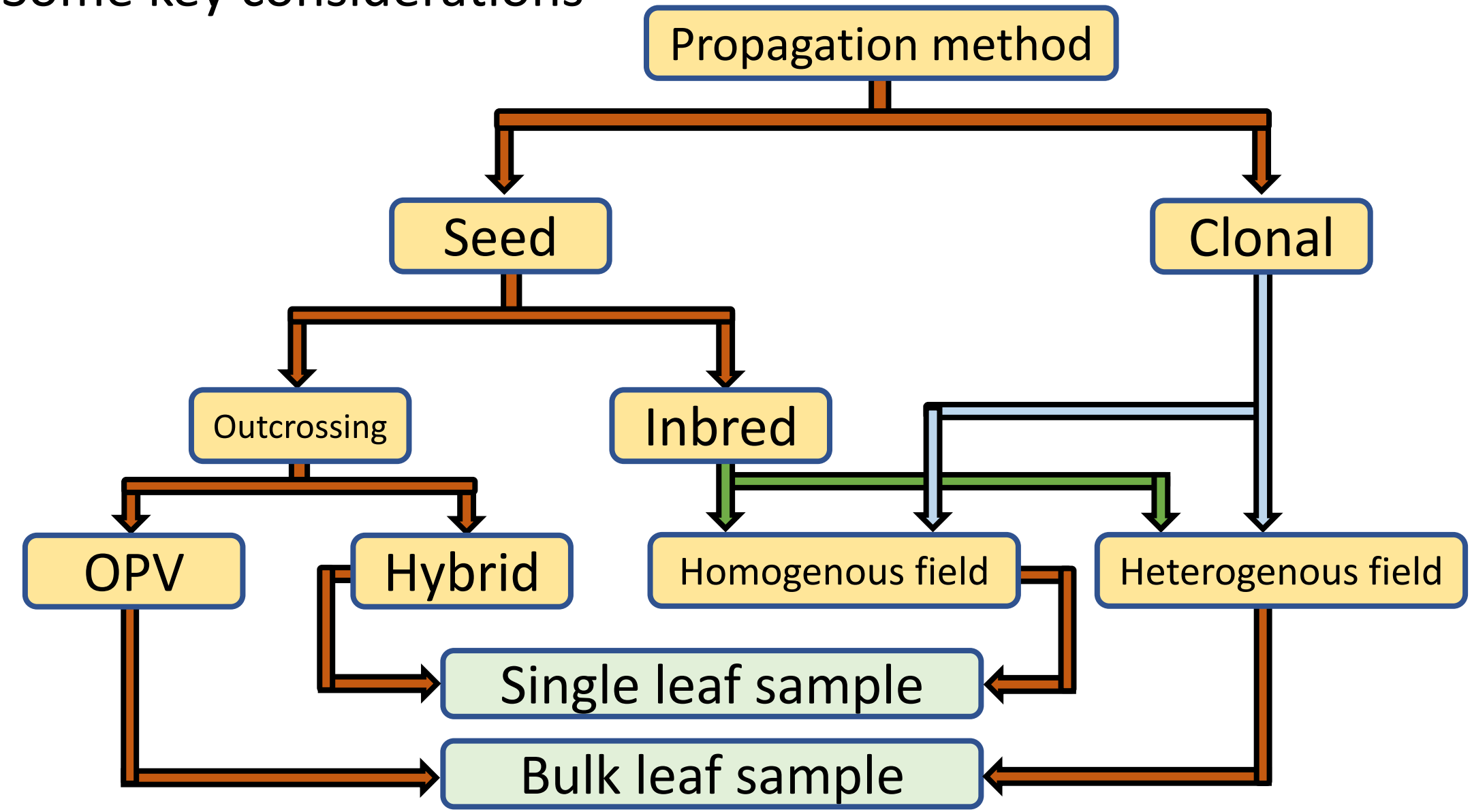
Outcrossing

Maize

- The clonals, Cassava and Sweet potato are outcrossing while potato is inbred
- However, they are cultivated asexually hence

Sample collection

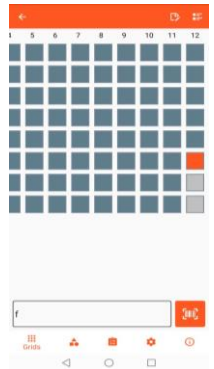
Some key considerations



Sample collection

Sample plating

- Sampling process is completed after samples are transferred to 96 well plates
- Liaise with a lab that can help with the process
- Use the coordinate app to guide in plating and to generate a sample tracking file



Coordinate app

Sample

96 Well plate

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

Sample tracking file

Genotyping

Samples shipped to GSP → DNA extraction → Library prep, sequencing & SNP calling

- It is best to bundle DNA extraction with genotyping
- GSP instructions on plastic ware and shipping should be strictly adhered to

Genotyping

Choosing a GSP

- Choice of GSP depends on
  - Nature of samples
  - Purpose of study
- Examples of GSPs
  - DArT
  - Agriplex
  - Intertek
  - Psomagen
  - CIAT etc

	Specific Identification	Varietal purity
Single leaf	➢MDP ➢DArT ➢Agriplex	➢LDP ➢Intertek
Bulk	➢MDP, GBS, DArTseq ➢DArT	➢LDP, MDP ➢Intertek ➢DArT

- The GSP will deliver genotyping data
- In some cases eg DArT & CIAT, preliminary analysis can be provided

Genotype data analysis

SNPs – Basis of fingerprinting

Improved variety

Genotyped region AATGCCTCGTAATCGCTCGTCC

Local variety

Genotyped region AATGCCTCGTACTCGCTCGTCC

Genetic region of interest AATGCCTCGTAATCGCTCGTCC (Improved variety)

Genetic region of interest AATGCCTCGTACTCGCTCGTCC (Local variety)

Single Nucleotide Polymorphism (SNP)

Genotype data analysis

SNP data formats

Sequencing

Pre-processing

SNP data

Data delivery

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

Most platforms: Score data, allelic

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

DArT output: Count data

Convert

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
SNP_001	1	1	1	1	1	0
SNP_002	1	1	1	0	1	1
SNP_003	0	0	0	0	0	0
SNP_004	0	0	1	0	0	0

Score data, numeric

Analysis pipelines

Simple comparison

- Simple visual comparison can work with
  - Few SNPs eg LDP
  - One or few references

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

Analysis pipelines

Distance methods

- Works by checking the similarity between the samples and references at each locus
- Examples of algorithms:
  - Identity by State (IBS)
  - Hamming
  - Nei's
- Typical Threshold of 5% genetic distance used to determine assignment

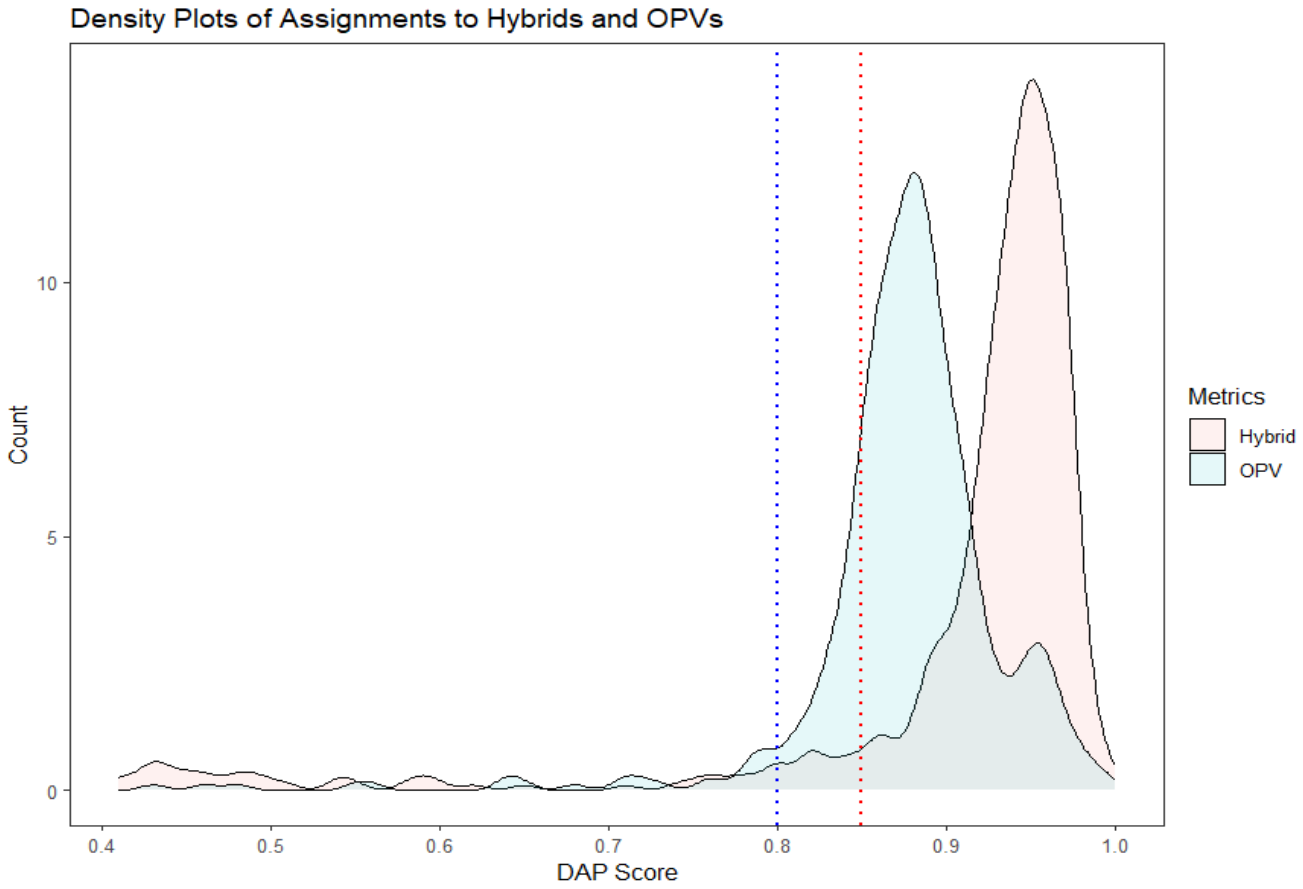
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Sample 1	0	0.002053	0.372163	0.37119	0.071311
Sample 2	0.002053	0	0.370782	0.370571	0.071575
Sample 3	0.372163	0.370782	0	0.00085	0.365269
Sample 4	0.37119	0.370571	0.00085	0	0.36394
Sample 5	0.071311	0.071575	0.365269	0.36394	0
Sample 6	0.071782	0.071987	0.365426	0.363788	0.001641

- Results in a *n* x *n* matrix showing relationships between all samples and all references

Analysis pipelines

Allele frequency methods

- These take advantage of high allele density count data
- Data from DArT used
- Compares allele proportions or frequencies between field samples and references to assign best matches
- Examples
  - Purity – DArT
  - DAP – DArT
  - Cluster – IMAGE



Example of score distributions following analysis by DAP algorithm

Conclusion

- DNA fingerprinting is a useful tool for varietal identification
- There are many steps involved and numerous considerations that need to be carefully evaluated
- We recommend starting the process as early as possible
  - Particularly reference library compilation
- We are available to offer guidance on
  - Conceptualization
  - Implementation
  - Identifying and negotiating with GSPs
  - Recommending analysis pipelines etc

Next steps

- Create a fingerprinting community of practice - Asana
  - Team leads
  - Point persons
  - Fill the questionnaire
- We can have more detailed discussions on specific activities
  - Reference library compilation
  - Sampling
  - Variety identification
- We'll avail a preliminary draft of a DNA fingerprinting guidance manual shortly