2017 International Symposium on Cocoa Research (ISCR), Lima, Peru, 13-17 November 2017

Adopting Reference Genotypes to Identify Off-types in Cacao Collections

C.J. Turnbull [1], A.J. Daymond [1], O. Gutierrez [2], P. Hadley [1], D. Livingstone [3], J.C. Motamayor [3], W. Phillips [4], P. Umaharan [5] and D. Zhang [6]

[1] School of Agriculture, Policy and Development, University of Reading, RG6 6AR, UK
[2] USDA-ARS, 13601 Old Cutler Road, Miami, FL 33158, USA
[3] Mars Inc., 13601 Old Cutler Road, Miami, FL 33158, USA
[4] Department of Agriculture and Agroforestry, CATIE, Turrialba, Costa Rica
[5] CRC, Sir Frank Stockdale Building, University of the West Indies, St. Augustine, Trinidad and Tobago
[6] USDA-ARS, 10300 Baltimore Avenue, Building 001 BARC-WEST, Beltsville, MD 20705, USA

Abstract

It is important for researchers, curators and breeders to have confidence in data being generated by germplasm evaluation and breeding programmes, and one of the key concerns is the true identity of the plants being used. Mislabelling of cocoa accessions is a significant problem, with estimates as high as 30%. Naming errors will have a large, detrimental impact on conservation, breeding and research, since mislabelled material is unlikely to exhibit the same genetic makeup or combination of traits as its true-to-type namesake. Farmers may also receive poorly performing material as a result of mislabelling. Data gathered from a mislabelled accession will result in misleading recommendations or conclusions, such as the genetic gain reported by breeding programmes.

Mislabelling events can also complicate the comparison of multiple data sets from different locations. It is particularly important to be able to link data to germplasm with confidence when combining the large trait evaluation and molecular analysis datasets necessary to identify and screen for key genes of interest.

The importance of correctly identifying material has been widely recognised for many years and several groups are currently generating genetic fingerprints using SNP markers. However, reliable identification of individuals will only be possible if a single genetic fingerprint is established as a reference for all others to be compared to and a core set of markers are consistently used. With this in mind, the Reference Genotype Working Group was formed in May 2016 with the aim of coordinating the verification of cacao germplasm.

This process has led to the development of an online tool to compare SNP profiles of individual trees and assign each a verification status; off-type, verified true-to-type or reference (original material if available). Although the tool can work with any number of SNP markers, the group have proposed a core set of widely-used markers to allow robust and consistent comparisons of profiles to be made across collections.

The verification status of accessions in a collection will be included in the International Cocoa Germplasm Database (ICGD) and made widely available to the cocoa community, with web pages created to highlight reference genotypes and compare other genetic fingerprints to these.

This paper describes the work of the Reference Genotype Working Group, initially focussing on the international collections and the International Cocoa Quarantine Centre (University of Reading), but with the intention to invite further collaboration from key partners and work towards the inclusion of other collections.

Introduction

Mislabelling in Cacao

Concerted efforts to identify germplasm through SNP and SSR fingerprinting and compare against referenced standards have confirmed high levels of off-types across different cocoa collections, relative to rates recorded in other species (Motilal *et al.*,2004; Motilal *et al.*, 2011; DuVal *et al.*, 2017). A figure of 30% is often quoted when discussing off-types in cocoa (from Schnell *et al.*, 2005), though off-types in hybrid trials have been estimated to be between 4.7% (DuVal *et al.*, 2017) to 54.5% (Padi *et al.*, 2015).

A recent study by DuVal *et al*. (2017) demonstrated that even <5% off-types present in a breeding programme altered selections by 48%, impacting heritability estimations for all of the traits analysed (including a 41% difference in estimated heritability for yield).

The frequent use of off-type parents in hybrid seed gardens of West Africa could be a major contributing factor to failures in meeting predicted productivity, with off-types reported to vary between 0 and 100% within a plot (Padi *et al*., 2015).

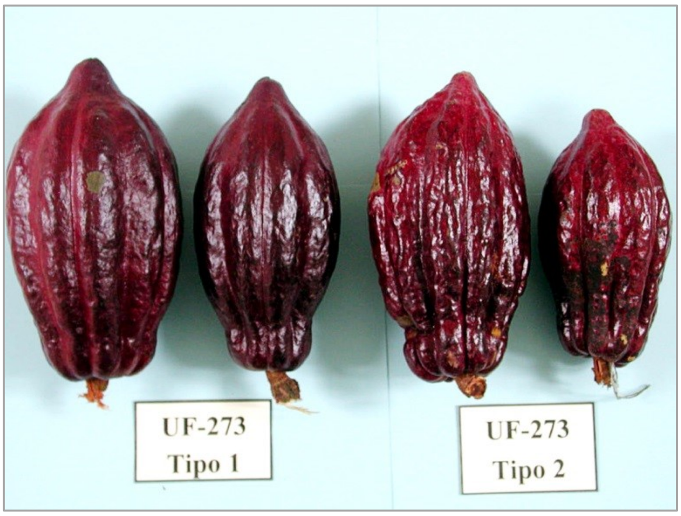


Figure 1. A photo showing the similarity between true-to-type UF-273 (Tipo 1) and an off-type (Tipo 2) at CATIE.

Identifying the existence of the off-type using genetic fingerprints made it possible to interpret previously confusing field trials data. As aresult of this work, UF-273 has become the most important and widespread source of resistance to Frosty Pod.

How can mislabelling occur?

There are a number of ways in which mislabelling can occur (Turnbull *et al*., 2004):

- Plants may lose their labels or the labels may become illegible;
- Plants may be moved before being properly labelled;
- Labels may get mixed up during vegetative propagation;
- Detached labels on the ground may be re-attached to the wrong plant;
- Dieback may result in labels being lost when a dead branch breaks;
- Chupons may grow from the rootstock and be confused with the scion;
- Established seedlings may be confused with the original tree;
- Some plants may be mislabelled in the greenhouse (human error);
- Introduction of synonymous germplasm (with different names) from abroad;
- Simple transcription errors can occur during plant propagation or label replacement.

Genotypes are usually difficult to distinguish on the basis of appearance, particularly during greenhouse propagation and initial planting when there are no flowers or pods. Consequently, identification relies heavily on the plant labels and field maps. However, maps become outdated when trees die or are replaced, or even when an old tree falls and a new "main trunk" becomes established in a new location.

Furthermore, hand annotation of maps is prone to misinterpretation and the problem can be confounded by high planting densities and irregular-shaped field plots with unclear boundaries; this can lead to ambiguities if individual trees are not labelled.

What should happen to mislabelled plants?

The curator of a germplasm collection makes the final decision as to whether to keep mislabelled material or not. Off-types that do not match a known clone should not automatically be removed since they may have valuable agronomic traits; for example, resistance to Frosty Pod has been found in two accessions in the cocoa collection at CATIE that are known to be off-types (Turnbull *et al*., 2004). However, any mislabelled material should be renamed, to avoid confusion with true-to-type material and prevent proliferation of the mislabelling event.

Renaming mislabelled clones

There is a need for a coordinated policy on the renaming of mislabelled accessions that can be followed by all curators of cocoa germplasm collections. Newly assigned names should be unique to the clone, they should have some meaning and should assist in documenting the origin of an off-type.

We propose the following procedure to avoid future confusion and provide mislabelling information for other users of the clone (Turnbull *et al*., 2004):

- The new name should be unique and already used within the collection, typically the local accession identifier (e.g. RUQ 1347).
- In addition, further information in brackets can be included with the name to indicate that the clone was originally misidentified:
    - The identifier MIS denotes that the clone has been mislabelled;
    - This is followed by the FAO code identifying the country and collection in which the mislabelled clone was found (e.g. the International Cocoa Quarantine Centre, Reading is 'GBR207');
    - The full name on the original label is retained at the end of the new clone name (e.g. CCN 51). This could be 'UNKNOWN' or left blank if, for example, the plant had lost its label;
- These parts would be separated by underscores (not previously used in clone names) for clear identification of the parts; e.g. RUQ 1347 (MIS_GBR207_CCN 51).

The new name (e.g. RUQ 1347) can be used on its own, since it is unique, but the inclusion of the additional information in written records (such as publications, labels, etc.) would help to highlight the mislabelling event. In particular, it records the original name that was given to the material, which may have been used when publishing data or distributing material. It is possible that some off-types may only be renamed temporarily, since a positive identification might be possible once a comprehensive database of DNA fingerprints is available.

The Reference Genotype Working Group

The Reference Genotype Working Group was formed in May 2016 during the 'Frontiers in Science and Technology for Cacao Quality, Productivity and Sustainability' symposium at Penn State University, USA (31st May to 2nd June, 2016). Following the symposium, a small group of researchers that were involved in the verification and/or curation of the international collections (in Trinidad and Costa Rica), as well as the collections at the University of Reading quarantine facility (UK) and the Mars Centre for Cocoa Science (Brazil), got together with the aim of coordinating the verification of cacao germplasm.

The objective is to bring together the existing SNP data being generated to compare genotypes using multilocus matching initially, then pedigree information and structure/population analysis. Where possible, a single reference genotype will be allocated for all other individuals to be compared to. Although ideally the most original example of a genotype, reference status will be assigned on a case by case basis, taking other factors in to account (such as access to material and data, and the quality and quantity of information available). Once a reference genotype has been designated, other material can be compared and assigned a verification status of 'true-to-type' or 'off-type'.

A similar approach was taken with SSR markers (Cryer *et al*., 2006), where allelic size standards and genotype standards were proposed, but the problems associated with genotyping errors and the standardisation of profile scoring between laboratories have meant that genotyping using SSRs has typically been restricted to specific projects and/or collections. The use of SNP-based multilocus fingerprints significantly improves the efficiency of genotype identification (Takrama *et al*., 2014), with little or no differences expected between profiles produced by research groups using the same markers. However, groups must still reference standard genotypes and use a core set of shared markers if SNPs are to provide reliable identification of individuals across collections.

Core Markers

The identification of panels of SNP markers which are suitable for identity and population ancestry analysis is needed to achieve comparative and transferable results among international collaborators (Motilal *et al*., 2017). A set of 96 SNP off-typing markers for cacao have been proposed by DuVal *et al.* (2017), which are included in the supplementary material for the article (Table S2), as well as another set proposed by Motilal *et al.* (2017). These will contribute to the final core panel of SNP markers to be put forward by the working group.

International Cocoa Germplasm Database (ICGD)

The group agreed that ICGD (www.icgd.reading.ac.uk) would maintain the SNP profiles available and develop the current webpages to enable reference genotypes to be highlighted and to allow comparison of other genetic fingerprints with these reference genotypes. Tools will also be developed to help identify the reference genotypes and assign a verification status to material. Accessions will be flagged as having been assessed, either as the reference, true-to-type or an off-type (when accurate identification has not yet been possible), so that they are not repeatedly compared as the process continues. The verification status will also be associated with accessions when displayed in ICGD.

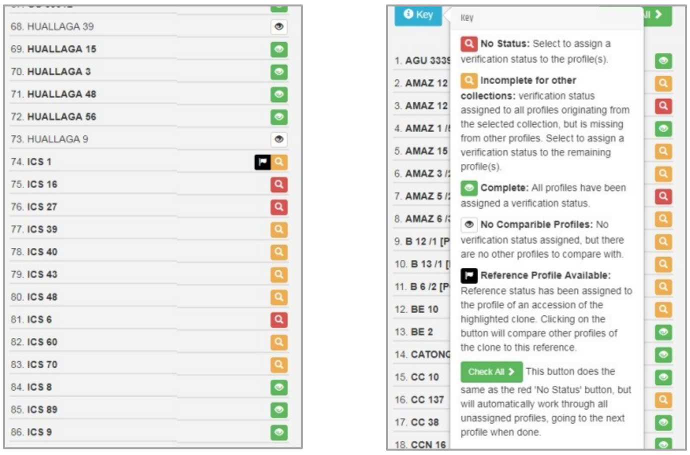
Online Tools

Although the SNP datasets are currently being finalised, several tools have already been created and are available to members of the working group (via a login page) through a new section of the ICGD website. These currently include 'Genotype Verification' using multilocus matching or population structure, a check for 'Contradictory Data', 'Collection Summaries', and the option to 'Match Off-types'), which are described in more details below. A tool that will allow users to input their own SNP profiles to be compared with reference genotypes is currently being developed.

Genotype Verification: SNP

This tool first presents the user with a drop-down list to select a collection in order to view a list of clones with SNP profiles available. Selecting the default option (a tick box) to 'Only show clones that require checking' will limit the results to accessions in the selected collection that have not yet been assigned a verification status and where SNP profiles are available from other locations for comparison. However, if not selected, all of the clones with SNP profiles in the collection are listed, with coloured icons highlighting the verification status:

- **Red 'Search' Icon** (magnifying glass) means no status has been assigned and clicking the button will continue to the next step in assigning a verification status to the profile(s).
- **Orange 'Search' Icon** means a verification status has been assigned to all profiles originating from the selected collection, but is missing from the profiles of other collections. Select to continue to the next step in assigning a verification status to the remaining profile(s).
- **Green 'View' Icon** (eye) means all profiles have been assigned a verification status. Follow the link to view the profiles.
- **White 'View' Icon** means no verification status has been assigned, but there are no other profiles to compare with. Follow the link to view the profile.
- **Black Flag Icon** next to one of the other 'View' or 'Search' icons means that reference status has been assigned to the profile of an accession of the highlighted clone. Clicking on the button will compare other profiles of the clone to this reference.



68. HUALLAGA 39
69. HUALLAGA 15
70. HUALLAGA 3
71. HUALLAGA 48
72. HUALLAGA 56
73. HUALLAGA 9
74. ICS 1
75. ICS 16
76. ICS 27
77. ICS 39
78. ICS 40
79. ICS 43
80. ICS 48
81. ICS 6
82. ICS 60
83. ICS 70
84. ICS 8
85. ICS 89
86. ICS 9

Key

No Status: Select to assign a verification status to the profile(s).

Incomplete for other collections: verification status assigned to all profiles originating from the selected collection, but is missing from other profiles. Select to assign a verification status to the remaining profile(s).

Complete: All profiles have been assigned a verification status.

No Comparible Profiles: No verification status assigned, but there are no other profiles to compare with.

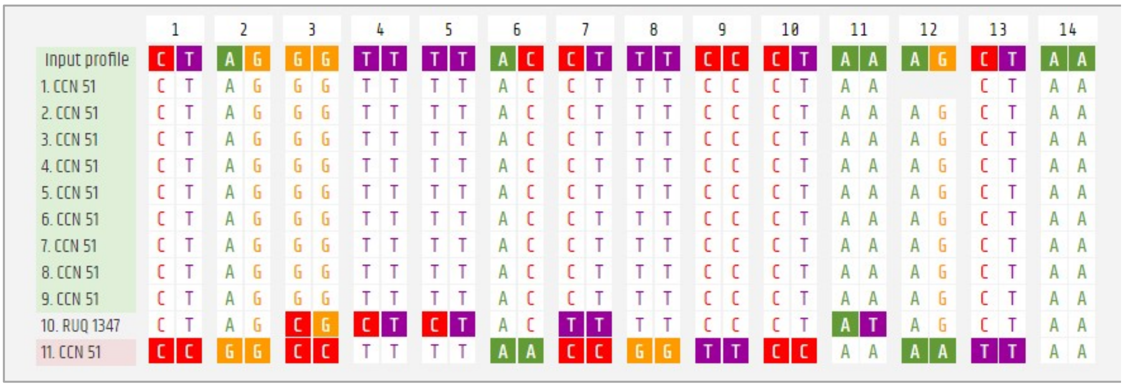
Reference Profile Available: Reference status has been assigned to the profile of an accession of the highlighted clone. Clicking on the button will compare other profiles of the clone to this reference.

Check All > This button does the same as the red 'No Status' button, but will automatically work through all unassigned profiles, going to the next profile when done.

Figure 2. Screenshots showing a clone list for a collection, with an example of each of the coloured icons indicating the verification status (*left image*). A key to these icons is available at the top of the page (*right image*).

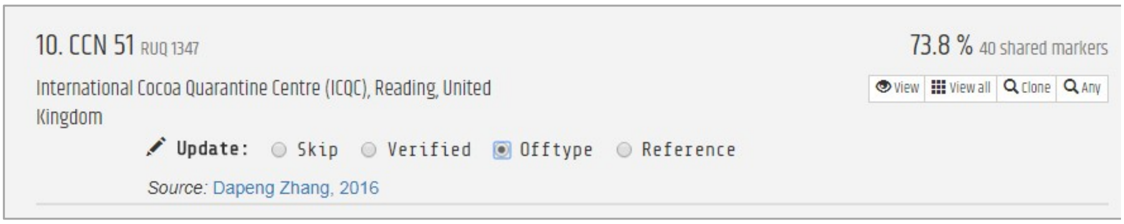
If a particular profile has been selected (e.g. the Reference Profile using the black flag icon) then that will be displayed first and all other available profiles lined-up below it. Each individual nucleotide making up the reference profile is displayed with a coloured background (A = green, C = red, G = orange and T = purple) and any differences in the profiles being compared are also highlighted with coloured backgrounds; the background remains white when matching. The percentage similarity, based on the number of exact matches of individual nucleotides and the number of shared markers, is displayed at the bottom of the page, along with the number of markers in common. If no particular profile has been selected, then all the profiles are aligned with coloured background, though no similarity scores. However, both views include the option to select another profile to compare others against and generate similarity scores.

If a profile has not been assigned a verification status, then the summary at the bottom of the page will include the option to update its status to verified (true-to-type), off-type or reference, though the reference option is only available if a reference accession has not already been designated. There is also an option to 'Skip' the profile if its verification status still requires further consideration.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Input profile	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
1. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A		C T	A A
2. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
3. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
4. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
5. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
6. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
7. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
8. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
9. CCN 51	C T	A G	G G	T T	T T	A C	C T	T T	C C	C T	A A	A G	C T	A A
10. RUQ 1347	C T	A G	C G	C T	C T	A C	T T	T T	C C	C T	A T	A G	C T	A A
11. CCN 51	C C	G G	C C	T T	T T	A A	C C	G G	T T	C C	A A	A A	T T	A A

Figure 3. Screenshots showing the comparison of CCN 51 profiles, with differences between the aligned profiles of the selected true-to-type and the two off-types highlighted.



10. CCN 51 RUQ 1347

73.8 % 40 shared markers

International Cocoa Quarantine Centre (ICQC), Reading, United Kingdom

Update: ○ Skip ○ Verified ● Offtype ○ Reference

Source: Dapeng Zhang, 2016

Figure 4. A section of the web page showing the similarity score and the number of shared markers for an off-type of CCN 51 from ICQC,R that is highlighted in Figure 3. In this example, the accession has yet to be designated an off-type, so the option to assign a verification status is available. **Note:** This accession is no longer called CCN 51 in the quarantine collection, it has been renamed 'RUQ 1347' (see example in the 'Renaming mislabelled clones' section above) and new accessions of CCN 51 have been imported.

Genotype Verification: Population

Some accessions have population assignment data available (based on the groups defined by Motamayor *et al*., 2008), which can be used as complimentary evidence to identify off-types. The user is first presented with a drop-down list to select a collection, with the option to 'Only show clones that require checking' (default) or all clones. Clones that have population data are listed, with icons to highlight their status:

- **White 'View Population' Icon** (tree) is available for all clones and is a link to the population data.
- **Red 'Off-type' Icon** ('X') means at least one accession has been designated an off-type.
- **Green 'Includes Expected' Icon** (tick) means at least one accession matches the expected population structure based on its origin or pedigree.

If an accession for which there is structure/population assignment data available has not been given a verification status, then the summary at the bottom of the page will include the option to update its status to 'Matches Expected' or 'Off-type', depending on whether or not the data agrees with the known origin or pedigree of the clone (the data is not specific enough to assign a reference or true-to-type status to material). There is also an option to 'Skip' the profile if its verification status still requires further consideration.

Contradictory Data

This tool lists any accessions that have been assigned a verification status based on both multilocus matching and structure/population analysis, but where the outcomes disagree. Although it is possible for an accession designated an off-type based on multilocus matching to be assigned the status 'Matching

Expected' using structure/population data, it should not be possible for the reverse to be true and such instances would need checking in more detail.

Collection Summaries

This option generates a simple report on the status of the collection selected by the user. It calculates the number of individual accessions, plus the clones they represent, that have been assigned a verification status (reference, true-to-type and off-type), as well as the number that remain unverified. A total for the number of accessions and clones is also reported, including a figure for the percentage of true-to-type (including reference) accessions and off-types in the collection.

Match Off-types

Although using SNP profiles to identify and rename off-types is not robust enough on its own, especially if only a small number of markers are shared between the off-type and any profiles that match it. However, the output of the tool can help to identify potential candidates and give an indication of the source of the error; for example, if an off-type of 'GU 144 /L' matches the reference profile for 'GU 114 /L', then it is likely the result a typing error. The tool lists all the clones across collections that include off-type accessions. If more than one accession has been classified as an off-type, a link to 'View' the profiles is displayed, from where selecting the red 'Off-type' button on any one of them will search for matching profiles. Where only one accession has been designated an off-type, a green 'Match' button is visible, which will initiate the search for matching profiles. Only profiles that are > 90% similar, with a minimum of 25 shared markers, are displayed. Any unmatched nucleotides are highlighted in the aligned profiles, with a similarity score and shared markers reported below.

Next Steps

Once the verification status of material in the international collections has been finalised and appropriate individuals have been designated as 'Reference' accessions, this information will be included in ICGD and made widely available to the cocoa community. This will include flagging the verification status in accession lists and SNP profiles generated in ICGD, as well as adding the option for the user to input a SNP profile to compare with that of the reference. The objective also includes collaboration with new partners to expand the system to cover other collections. A core set of widely-used markers will be proposed, based on those already being utilised by a number of groups.

Conclusions

Mislabelling is a significant problem in cocoa and can have a large, detrimental impact on conservation, breeding and research. By designating reference genotypes and core SNP markers, as well as providing access to the associated information and profile comparison tools, the Reference Genotype Working Group will help to make the validation of genotypes at all stages of a research, conservation or breeding programme a cost effective option, and one that should be widely adopted by the cocoa community.

References

Cryer, N.C., Fenn, M.G.E., Turnbull, C.J. and Wilkinson, M.J. (2006). Allelic size standards and reference genotypes to unify international cocoa (*Theobroma cacao* L.) microsatellite data. Genetic Resources and Crop Evolution 53:1643–1652. Doi: 10.1007/s10722-005-1286-9

DuVal, A., Gezan, S.A., Mustiga, G., Stack, C., Marelli, J-P., Chaparro, J., Livingstone, D. III., Royaert, S. and Motamayor, J.C. (2017). Genetic Parameters and the Impact of Off-Types for *Theobroma cacao* L. in a Breeding Program in Brazil. Front. Plant Sci. 8:2059. doi: 10.3389/fpls.2017.02059

Motamayor, J.C., Lachenaud, P., da Silva e Mota, J.W., Loor, R., Kuhn, D.N., Brown, J.S. & Schnell, R.J. (2008). Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (*Theobroma cacao* L). PLoS ONE 3(10): e3311. doi:10.1371/journal.pone.0003311

Motilal, L. A., Sounigo, O., Butler, D. R., & Mooleedhar, V. (2004). Misidentification within global cacao germplasm collections. In Vitro Culture, Transformation and Molecular Markets for Crop Improvement, 121–129.

Motilal, L. A., Zhang, D. P., Umaharan, P., Mischke, S., Pinney, S., & Meinhardt, L. W. (2011). Microsatellite fingerprinting in the International Cocoa Genebank, Trinidad: accession and plot homogeneity information for germplasm management. Plant Genetic Resources-Characterization and Utilization, 9(3), 430–438. https://doi.org/10.1017/S147926211100058x

Motilal, L., Mahabir, A., Gopaulchan, D., Sankar, A. and Umaharan, P. (2017). Identification of a Core SNP Panel for Cacao Identity and Population. Poster Presentation: International Symposium on Cocoa Research (ISCR), Lima, Peru, 13-17 November 2017.

Padi, F. K., Ofori, A., Takrama, J., Djan, E., Opoku, S. Y., Dadzie, A. M., Bhattacharjee, R., Motamayor, J.C. and Zhang, D. (2015). The impact of SNP fingerprinting and parentage analysis on the effectiveness of variety recommendations in cacao. Tree Genet. Genomes 11, 44. doi: 10.1007/s11295-015-0875-9

Schnell, R. J., Olano, C. T., Brown, J. S., Meerow, A. W., Cervantes-Martinez, C., Nagai, C. and Motamayor, J.C. (2005). Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. J. Am. Soc. Hortic. Sci. 130, 181–190.

Takrama, J., Kun, J., Meinhardt, L., Mischke, S., Opuku, S., Padi, F. K. and Zhang, D. (2014). Verification of genetic identity of introduced cacao germplasm in Ghana using single nucleotide polymorphism (SNP) markers. Afr. J. Biotechnol. 13, 2127–2136. doi: 10.5897/AJB2013.13331

Turnbull, C. J., Butler, D. R., Cryer, N. C., Zhang, D., Lanaud, C., Daymond, A. J., Ford, C.S., Wilkinson, M.J. and Hadley, P. (2004). Tackling mislabelling in cocoa germplasm collections. INGENIC Newsletter. 9, 8–11