Design and implementation of a global collaborative framework on Cacao Genetic Resources: incentives, constraints and institutional structures

Selim Louafi[1,2], Andrew Meter[1,2], Brigitte Laliberté[3], Viviana Medina[4]

1. Cirad, UMR AGAP, F-34398 Montpellier, France
2 AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
3. Bioversity International, Headquarter, Rome ,Italy
4. Bioversity International, Costa Rica office, Turrialba

Abstract

T7 - Marketing, Socio-Economics and Technology/ innovation adoption/transfer

Design and Implementation of a Global Collaborative Framework on Cacao Genetic Resources Research: incentives, constraints and institutional structures

Research in cacao genetics plays a crucial role for the sustainability of the cacao sector. Effective management and improvement of cacao genetic resources relies on the exchange of resources such as genetic material, data or knowledge between different countries and across continents. It often involves global collaboration among a range of diverse actors interested in cacao genetic resources but with different capacities, aspirations and motivations. The cacao genetic community has already engaged in large-scale research collaboration in the past especially through the international CFC/ICCO/Bioversity projects from 1998 to 2010 and a new collaborative initiative is currently being discussed, the Collaborative Framework for Cacao Evaluation (CFCE). This paper aims at understanding the opportunities and constraints for the formation process of collaborative inter-organisational initiative in cacao genetic research. It identifies the range of challenges to be addressed by the cacao community to make more informed choices about definition of common objectives, process and governance structure in establishing a collaborative initiative. This paper draws from an analysis of a survey conducted in April 2016 on a sample of 391 people involved in cacao genetic resources related activities, a bibliographic analysis as well as an in-depth evaluation and interviews carried out on the CFC/ICCO/Bioversity projects, drawing out the key lessons learnt and recommendations. Preliminary results show that existing barriers can potentially play against global collaboration and undermine a perceived sense of convergent interests. However, these constraints are more than counterbalanced by the existence of institutions that have the ability to support global collaboration and by pre-existing social relationships, including the CFC/ICCO/Bioversity project, that reflect a sense of strategic interdependency among potential participants. Therefore, the community's capacity to build on the awareness of the benefits of global collaboration and to agree on global objectives will depend on its ability to overcome tensions created by geographical distances, disciplinary divides or differences in capacity and to design a collaborative framework that will take advantage of existing converging forces while minimizing the effects of diverging forces.

Introduction

Research in cacao genetics plays a crucial role for the sustainability of the cacao sector. Effective management and improvement of cacao genetic resources relies on the exchange of resources such as genetic material, data or knowledge between different countries and across continents. It often involves global collaboration among a range of diverse actors with different capacities, aspirations and motivations. The cacao genetic community has already engaged in large-scale research collaboration in the past, especially through the international CFC/ICCO/Bioversity projects from 1998 to 2010 – "Cocoa Germplasm Utilization and Conservation: A Global Approach (1999-2004)" and "Cocoa Productivity and Quality Improvement: A Participatory Approach. (2004-2009)" (Eskes and Efron, 2006; Eskes, 2011). A new collaborative initiative, coordinated by Bioversity International, is being developed, the Collaborative Framework for Cacao Evaluation (CFCE).

This short policy paper reflects on the global collaborative context among the various actors of the community interested in cacao genetic resources. It draws from the findings of a Master Thesis (Meter, 2016) based on a bibliographic analysis and a survey that was conducted in April 2016 on a sample of 391 people involved in cacao genetic resources related activities. This paper is also based on the results of an in-depth evaluation of the CFC/ICCO/Bioversity projects (Medina *et al.*, 2017), drawing out the key lessons learnt and recommendations. This paper aims at guiding a collective reflection on the constraints and opportunities driving international collaboration in cacao genetic research and on an appropriate collaborative framework for a future initiative.

1 Collaboration formation: Initial conditions

Before engaging in a collaborative multi-stakeholders initiative, many factors can be considered as enabling and shaping the initiative's scope, goals and structure. Facilitating factors such as financial resources are determinant in the realization of the collaboration. Nonetheless, more structural factors related to the characteristics of the community are seen in the literature on collaboration in science as critical to understand the limits and opportunities for international collaboration. Initial positions of the community vis-à-vis the following factors set the agenda for partners in a collaborative multi-stakeholders initiative:

i. The clarity of potential benefits deriving from collaboration and the existence of a sense of convergent interests that may lead towards a desire to seek common ground (further discussed in Section 1.1)

ii. Proximities or distances between participants and strategic interdependency among actors (obstacles/opportunities, further discussed in Section 1.2)

iii. Preexistence of social relationships that provide for initial mutual understanding and trust, allowing the partners to start collaborating more rapidly and easily, and past experiences (further discussed in section 1.3)

The position of a community with regard to these initial conditions defines the community's "readiness" for collaboration which spans from rather spontaneous formation process to emergent or engineered process that require significant amount of managerial attention[1]. Collaboration also happens more easily when prior enabling conditions are met, less easily when they are not. In the process of developing a collaborative initiative, it is crucial for potential participants to collectively assess the constraints and opportunities for collaboration. In this section, we ask: Is there a common awareness of the potential benefits of global

[1] See Ring et al., 2005

collaboration within the cacao genetics research community? What are the constraints and opportunities for global collaboration in cacao genetic research?

1.1 Existence of a sense of convergent interests

Global issues related to cacao genetic resources hold an unusual position compared to other cultivated plants. Cocoa being mainly produced by small holder farmers, lack of economic incentives for the private sector limits its involvement in cacao breeding and conservation and use of cacao genetic resources. Public national research institutes and universities working on cacao supported by punctual private and public funds, along with a few private research structures, ultimately carry the burden of governing the cacao genetic resource public good – through conservation (see list of institutions holding accessions in Appendix 3) and/or use in genetic research and breeding.

This situation has spawned a common perception that separate efforts from different sectors (profit, non-profit, public) have failed or are likely to fail in addressing global challenges related to genetic resources. Efforts by one actor/sector alone, including the most powerful ones, cannot provide the full global public goods and services associated with cacao genetic resources. The existence of such perception is materialized by initiatives such as the Global Network on Cacao Genetic Resources (CacaoNet) or the development of a Global Strategy for the Conservation and Use of Cacao Genetic Resources (CacaoNet, 2012), which set priorities and emphasize the need for coordinated global efforts on the matter.

Key global challenges exist that can effectively incentivize collaboration. This has been observed through the CFC/ICCO/Bioversity projects, originally framed as an opportunity to stimulate under-funded cacao breeding programs worldwide while tackling the issue of cacao susceptibility to pests and diseases. In a Workshop organized on June 3rd 2016 by CacaoNet and the International Group for Genetic Improvement of Cocoa (INGENIC) concerning the development of the CFCE, selected topics such as the spread of pests and diseases and climate change adaptation have been recognized as globally critical and of current interest to all (national research organizations, donors and the industry).

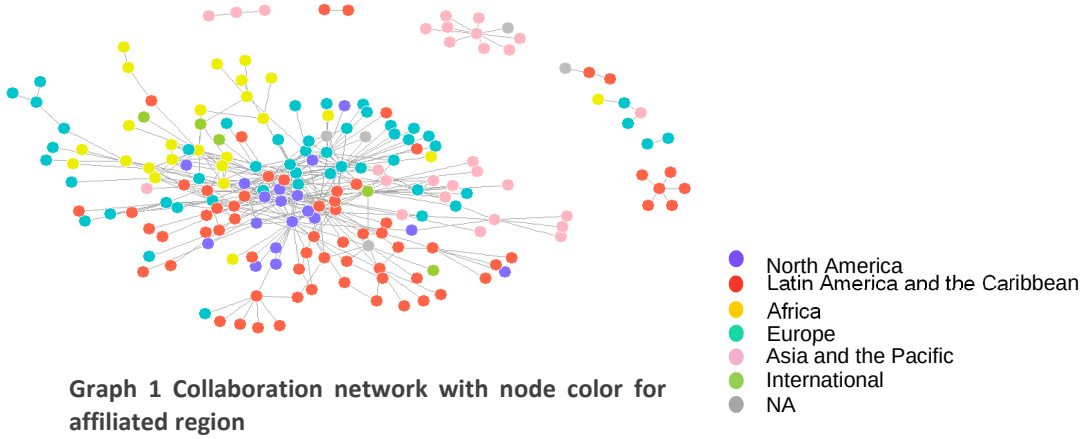
The cacao genetic research community's awareness of the benefits of global collaboration is decisive in the community's ability to find common ground. However, motivation and capacity to engage in a long term global collaborative initiative depends on a set of opportunities and constraints that can potentially play against or in favor of global collaboration. These opportunities and constraints will be further discussed as convergent and divergent forces in section 1.2 and as prior social relationships and past experiences in section 1.3.

1.2 Convergent and divergent forces

1.2.1 Convergent forces: structural and institutional opportunities for global collaboration

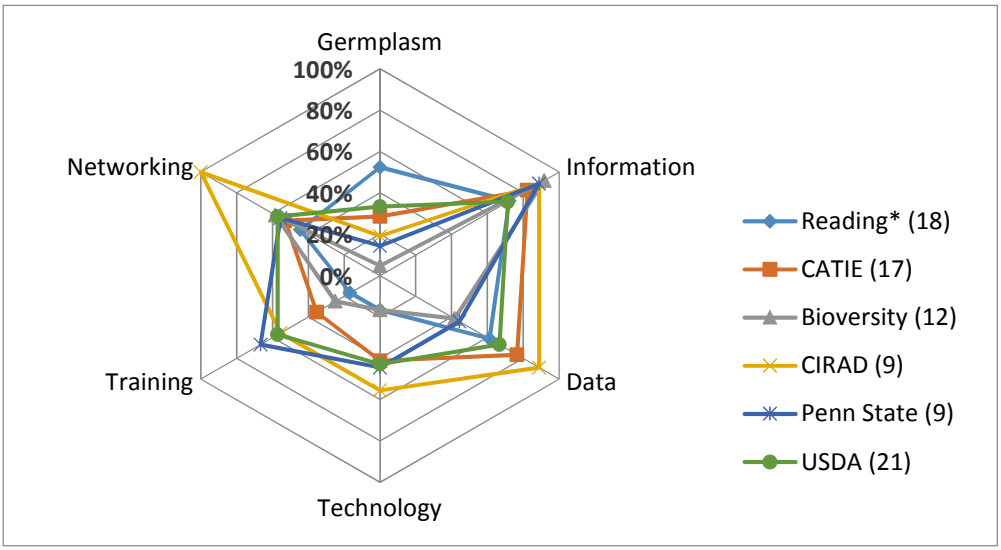
Six years after the end of this project, a social network analysis has been conducted based on reported ties between members of the cacao genetic resources community relating to the exchange of different types of resources[1]. Findings reveal the existence of a cohesive community that connects around central/key actors (see graph 1 below).

[1] This work was conducted between April and September 2016 through a Master's Thesis (Meter, 2016). Lists of collaborators provided by 84 respondents to the survey served as the basis for a social network analysis (network comprised of 196 individuals). See Appendix 1 for methodology and results



Graph 1 Collaboration network with node color for affiliated region

The centrality of institutions was measured based on the sum of the centrality of its members in the network – i.e. number of incoming ties from members of other institutions. Ties between individuals refer to the exchange of resources such as germplasm, data or information. Among the most central institutions are the United States Department of Agriculture (USDA); Mars; the Reading Hub including Reading University, the International Cocoa Quarantine Centre, Reading (ICQCR) and the International Cocoa Germplasm database (ICGD); the Cocoa Research Center (CRC); the Center for Tropical Agricultural Research and Education (CATIE); Bioversity International and the Centre for International Cooperation in Agricultural Research for Development (CIRAD). By analyzing resource-type subnetworks – networks formed by ties involving only one type of resource, such as germplasm or information – it also appears that these central institutions have different and complementary profiles as resource and service providers (see graph 2 below and graphs in Appendix 1.2, p. 15). Members of the community are also connected through



formal networks such as regional cocoa breeders' working groups, and at the international level the INCOCOA Groups[1] and CacaoNet, which are also important institutional resources.

Graph 2 Proportion of exchanges for each type of resources with individuals outside of the organization (total number of ties are mentioned next to the organizations)
***Includes individuals affiliated to the University of Reading, ICQCR and ICGD**

This underlines a key feature of the community: the presence of central actors and global networks, recognized by the international community as brokers for the exchange of specific sets of resources and services and hence facilitate or at least support collaboration. This is no minor observation especially when

[1] INGENIC, INAFORESTA, INCOPED and INCOSOM

considering the suitability of their attributes for international collaboration[1]. For instance, CATIE and CRC' international cacao *ex situ* collections are under the multilateral system of the International Treaty on Plant Genetic Resources for Food and Agriculture (IT-PGRFA), ICQCR offers an international platform for the safe exchange of cacao genetic material (a platform that is lacking in coffee for instance), and CIRAD stands out by its regular collaboration with a wide set of geographically dispersed actors[2].

The existence of central actors is an indicator of the gradual awareness of strategic interdependency within the community in the past 20 years. Given their role of boundary spanners, some of these institutions have the ability to act as a binding force and develop trust within the global community while simultaneously fulfilling their role in providing global public goods or services.

1.2.2 Divergent forces: potential obstacles to global collaboration

The varied set of actors involved in the conservation and use of cacao genetic resources (public research institutions, government bodies, private companies, non-profit) face locally embedded issues and have their own objectives, values and practices related to cacao genetic resources. Universal common ground is not innate to the cacao genetic resources community. In this context, some actors tend to share more common interests, some less, some may be more isolated, and some more connected. Overall, some clustering forces could potentially play against global collaboration as they divide the community into separate clusters.

Some of these are:

- **The regional and biological divides**, due to numerous factors such as regional specificity of pest and diseases, quality and flavour driving different markets, the existence of regional/national regulatory frameworks, or simply geographical distances
- **The disciplinary divide** and more particularly, the persistent division between conservation/diversity characterization, breeding and bio-informatics. They do not see genetic resources the same way as they all value different types of genetic material and information associated to them.
- **The market divide**, specialty/high premium versus the larger quantity markets and, more generally, the private company strategies with regard to genetic resources-related research and partnerships

Identifying these clustering forces and their effect on community structure is not straightforward as they tend to overlap. Results from the social network analysis did not allow for a clear identification of these divergent forces[3]. Yet, clustering processes such as regional or disciplinary groupings appear in the findings from the bibliometric analysis (See Appendix 2 for methodology and results). Generally speaking, high levels of convergent interests among members of each of the separate clusters may be an obstacle to the identification and achievement of more global common interests. In sum, the resulting coexistence of high proximity among some members and high distance across clusters can be a source of mistrust, an obstacle for smooth global resource exchanges, and therefore undermine existing incentives for global collaboration.

[1] This also explains their high centrality and exchanges with diverse actors worldwide.

[2] CIRAD cocoa researchers' tendency to publish with many scientists from different countries was clearly identified through our bibliographic analysis – see Appendix 2.2 p.18 (group K5).

[3] This also comes to show that no clustering effect appears strong enough to clearly structure the network. See end of Appendix 1.1

1.3 Antecedents of global collaboration

Through a succession of collaboration initiatives in the past 20 years, a set of relational ties between actors at the regional and global level have emerged. Some of these relationships are formalized by networks at regional and international levels (e.g. Regional Cocoa Breeders' Working Groups, INCOCOA groups, CacaoNet).

Reflecting on the process of their emergence underlines the importance of the CFC/ICCO/Bioversity projects (1998 – 2010), coordinated by Bertus Eskes from Cirad/Bioversity International. The two projects, aimed at providing new cacao varieties with improved yielding capacity, disease resistance and quality traits for increasing global cocoa outputs (Eskes and Efron, 2006; Eskes, 2011). Several activities, from multi-site trials to participatory activities involving producers, were carried out by participating national research institutions with the support of government bodies, the private sector and central actors.

The CFC/ICCO/Bioversity projects acted as strong catalysts for international collaboration, which was sorely needed at the time the projects started in 1998: cacao breeding during the 1990s suffered seriously from the low price of cocoa, links between national collections and the main international cacao collections were weak, links between breeding and conservation programs were generally weak or non-existent, and many breeders lacked adequate training and frequently operated under rather isolated conditions. The project raised awareness on the necessity for international collaboration within the community, and was an introduction to global collaboration for many participating research institutions. In addition to delivering concrete outputs (reinforcement and re-initiation of cacao breeding programs, validation and exchange of selected material, distribution of new material to farmers, generated and exchanged information, insights gained in resistance testing methodologies etc.), the projects also enabled participants to formalize their strengthened relational ties through the formation of networks such as the African Cocoa Breeders Working Group.

On the other hand, participants of the CFC/ICCO/Bioversity projects have also witnessed the limits of global collaboration and experienced various difficulties when collaborating and implementing trials. For instance, problems arose regarding the exchange of genetic material or the implementation of standardized working procedures among institutions. Having been introduced to the various limits and constraints of a global collaborative initiative, members of the community and especially participants of the CFC/ICCO/Bioversity projects might show reluctance, skepticism or at least sound criticism on certain aspects of an upcoming initiative. While this may complicate decision making processes, the experience gained by the community should also be regarded as valuable insight for better planning and for setting more attainable goals.

Current initiatives in cocoa research, some incorporating issues linked to genetic resources, are also proof of the ability for the actors of the broader cocoa research community to overcome obstacles to collaboration. Such initiatives may also present complementarities with a global collaborative initiative on cacao genetic research and can be seen as an opportunity for synergies.

The experience from the CFC/ICCO/Bioversity projects indicates that it is possible to design a global cooperative framework that can leverage on convergent forces and attenuate divergent forces while serving the interests of the widest range of participating actors. Drawing lessons from the CFC/ICCO/Bioversity projects can then help in the design of a future collaborative initiative – see the in depth review recently carried out (Medina *et al.*, 2017). Reflecting on the opportunities presented by ongoing research projects in cacao more generally will also help in identifying important gaps in global collaboration efforts linked to research on cacao genetic resources and therefore maximizes the value of an upcoming collaboration initiative. Links with these ongoing initiatives could also scale up the impacts of an upcoming initiative while limiting duplication of research efforts.

2 Process of Collaboration

Evidence from the literature on collaborative inter-organizational relations and lessons learned from the CFC/ICCO/Bioversity projects show that management challenges could vary greatly depending on the choices concerning:

- the scope, primary focus and goals of the initiative,
- the range of varied actors that will be ultimately gathered, and
- the mix of resources that will be primarily pooled and produced within the initiative.

There are no simple and uniform ways of addressing these challenges and hence it is not appropriate to develop specific recommendations. Rather, our approach is to direct members of the cacao genetic research community to consider the tensions that exist between alternatives in light of the current context.

2.1 Goal setting

Previous research on collaboration in science in health and plant genetics and genomics allows to identify three broad categories of goal orientations (Welch, Louafi, Fusi, 2016):

- **A research-oriented approach**, which might include different levels of research goal aggregation. Initiatives can provide technical support to already existing research projects (low level of goal integration) or can support the community in developing common practices and research methods across projects (medium level of goal integration); it can aggregate partners towards overarching, common research goals (high level of goal integration).
- **Community-building approach**, which give emphasis in generating continuous interactions among members to promote sharing and learning over time. It can materialize in different activities ranked in terms of resource intensiveness: exchanging information on existing projects (low level); brokering services and expertise (medium level), and providing capacity development (high level).
- **Service provision approach,** which can consist in a variety of products and services ranked in terms of resource intensiveness and level of commitment in the long term: providing tools and access to technology, (low level); adding the deployment of technical standards that allow interoperability across locations (medium level); offering and maintaining a platform for pooling resources (knowledge, data, germplasm..) (high level).

These goal orientations do not need to be mutually exclusive. While some initiatives may only focus on one of these goals, many integrate all three. Nevertheless, a relative and sometimes subtle focus on one to the expense of another greatly influences the structure as well as the output of an initiative.

The CFC/ICCO initiative primarily aimed at integrating and organizing scientific and technical efforts globally and at guiding partners towards overarching, common research goals. The CFCE's primary goal is to optimize the use of cacao genetic diversity in development of improved, diverse and locally-adapted varieties through international collaboration, bringing together players in public and private sectors. Discussions are now focused on which specific common research goals should be set as drivers for the initiation of this research collaboration. Pests and diseases being region specific appear to hold potential for division within the community – although the CFC/ICCO/Bioversity projects have successfully integrated this issue as a driving force for research collaboration. The urgent issue of climate change and particularly tolerance to drought, heat and high levels of CO2 appears to have to potential to federate all partners. However, initiatives that have broad global missions expose to trade-offs and challenges. For example, a broad research agenda is likely to include heterogenous partners with diverse perceptions about benefits and contributions which may require putting more efforts on community-building activities to increase goal consensus and resolve differences.

Given this primary orientation, what does it imply in terms of actors to be involved and resources to be exchanged and managed, or in terms of governance structures and mechanisms?

2.2 Set of actors to be involved

Large scale global collaborations often involve the aggregation of a set of actors who may differ on a wide range of aspects: academic disciplines, sector, culture and economic interest but also, in relation to these attributes, variations in endowment, political objectives, wealth and entitlement. This heterogeneity of actors is often important to initiate collective action but in the long run, it may deter participation and lead to some coordination problems, including poor compliance, consensus building, distributional conflicts, low trust and low provision and use of common resources.

Reducing the size of the group and/or the level of heterogeneity of interacting groups may seem the most obvious way to reduce collective action problem. However, this solution immediately raises legitimacy issues at the global level, especially in complex and politically charged environment where production of results depends upon the ability to aggregate a various set of (material and digital) inputs and the skills of a various actors (interdependence).

Hence, for an initiative to be successful, it is crucial to designing mechanisms to deal with heterogeneity. Such mechanisms could either consist in : i) structural solutions such as developing small-scale pilot projects that prove to be more manageable; or developing homogeneous sub-communities (by region, research topic or disciplines) or developing a phased approach where inclusion of more heterogeneous actors is undertaken after a consolidation phase of an initial more homogenous group; ii) motivational solutions that would focus on changing partners' perceptions of the social environment and hence their willingness to collaborate. The type(s) of heterogeneity that act as the strongest barriers to trust should be primarily targeted. For example, heterogeneity in capacity that are easily found in large scale international collaboration could lead to distributional conflicts related to input allocation and outputs and benefits redistribution. If left uncheck and not managed, such heterogeneities could generate dysfunctional and conflictual perceptions of equity among the various parties involved over time. In such context, attention need to be paid not only on increasing knowledge and resources through collaboration (expanding the pie) but also on how these knowledge and resources are actually accessed, used and valorized among members with differentiated capacities (sharing the pie).

2.3 Resource Mix

Collaboration in genetic research involves the pooling and management of multiple types of resources both serving as input for and produced through the research process. A first set of resources includes **genetic materials** such as seeds or other propagation materials, plant material or DNA and genomic or phenotypic **data associated to this material.** Other resources, perhaps too often overlooked, also play an important role in collaboration. These are **technical resources** (including equipment, software, human resources for assistance with access and use of existing data etc.); **organizational resources** (which facilitate interaction, collaboration, deliberation, or dissemination among individuals or groups); **institutional resources** (comprising data and material sharing standards such as the *technical guidelines for the safe movement of cocoa germplasm*, or assistance with the development and understanding of legal and regulatory issues); **knowledge resources** (including the knowledge outcomes of collaboration that are embedded in journal articles or research protocols and tacitly understood by scientists); and **social capital** (referring to the availability of relational resources such as access to new networks).

These resources have distinctive properties that entail very different management and regulatory challenges concerning their pooling, accessibility, use and sharing. In order to avoid long and complex adjustments of varied (proprietary) rules that may apply to these different resources, it is often considered in global collaboration context that it is easier to i) manage the type of resources separately (e.g. Material

Transfer Agreement (MTA) for the exchange of material); ii) privilege open systems as a way to facilitate exchange and sharing of resources. However, managing a single set of resource through open systems (e.g. open data) could become problematic in heterogeneous context, as effective use of open systems requires pre-existing infrastructures, knowledge and skills that are most likely to be found among the wealthier or higher capacity entities (e.g. research organizations, countries or stakeholders). A well-functioning open system does not resolve the issues posed by great differences in the capacity to valorize a particular resource. It is suggested that cooperative behaviors in relation to sharing of resources such as biological material and information are often made easier and sustained over time when embedded in broader collaborative research frameworks that recognizes and establishes linkages across multiple resources and activities. Such embeddedness offers opportunities to find cooperative equilibriums that single-resource transaction alone cannot easily achieve.

This requires deciding about the resource mix to be aggregated, produced, accessed and sustained over time and what pooling solution is most adapted to the capacity of the different actors involved. In particular, reflecting on the use and sharing of genetic material is of crucial importance. The value of exchanging genetic material for use in research (especially across continents) can be tremendous. Yet access and benefit regulations and risks of spreading diseases continue to be a challenge for such exchanges. The existence of the two international collections at CATIE and CRC and the ICQCR offers great perspectives for the introduction of new genetic diversity across continents in the framework of a global collaboration initiative – and was a cornerstone of the CFC/ICCO/Bioversity projects. Ten years later, could a similar framework for cacao genetic material exchange be sought for in an upcoming initiative?

3 Conclusion

In total, although the existence of divergent forces can potentially play against global collaboration and undermine perceived sense of convergent interests, this is balanced by the existence of strong converging forces, materialized by the existence of institutions and networks that reflect a sense of strategic interdependency among actors. Together, these institutions can provide the range of motivational, institutional and technical support needed for a global collaboration initiative. Furthermore, considering pre-existing social relationships among potential participants, the community appears in many aspects ready and well-conditioned for collaboration at a global level. In practice, it appears that an entity is still needed to trigger collaboration – consistent with an engineered process of collaboration formation (Ring et al. 2005). Nevertheless, the role of this triggering entity is less about starting from scratch and creating awareness on the benefits of collaboration than about helping overcome key challenges related to the scale at which collaboration needs to be established: heterogeneity of actors, pooling and enabling access to resources (especially germplasm), complexity of large scale coordination etc. The key players able to act as bridge builders or even triggering entities face the challenge of jointly designing formal arrangements that must overcome high distances between actors while taking advantage of their proximities and of existing cohesive institutional structures. They also need to come up with a governance process that will build trust and willingness of a wide range of varied actors to commit time and resources.

References

CacaoNet, 2012. A Global Strategy for the Conservation and Use of Cacao Genetic Resources, as the Foundation for a Sustainable Cocoa Economy (B. Laliberté, compiler). *Bioversity International*, Montpellier, France.

Eskes, A.B. and Y. Efron, editors. 2006. Global Approaches to Cocoa Germplasm Utilization and Conservation. Final report of the CFC/ICCO/IPGRI project on "Cocoa Germplasm Utilization and Conservation: a Global Approach" (1998-2004). CFC, Amsterdam, The Netherlands/ICCO, London, UK/IPGRI, Rome, Italy.

Eskes AB, editor. 2011. Collaborative and Participatory Approaches to Cocoa Variety Improvement. Final report of the CFC/ICCO/Bioversity project on "Cocoa Productivity and Quality Improvement: a Participatory Approach" (2004-2010). CFC, Amsterdam, The Netherlands/ICCO, London, UK/Bioversity International, Rome,Italy.

Medina, V., Meter, A., Demers, N. and Laliberte, B. 2017. *Review of the CFC/ICCO/Bioversity project on cacao (1998-2010)*. Costa Rica: Bioversity International, forthcoming.

Meter, A. 2016. *Network Collaboration in Science for the Global Genetic Resources Commons - A study of the global collaboration in the cacao genetic resources community.* Master's Thesis : EcoDEVA, Montpellier SupAgro, 56 p. [online]. Retrieved from https://web.supagro.inra.fr/pmb/opac_css/doc_num.php?explnum_id=3171 (accessed on 09/28/2017)

Ring, P.S., Doz, Y.L. & Olk, P.M., 2005. Managing formation processes in R&D Consortia. California Management Review, 47(4), p.137–156.

Welch, E., Louafi, S., Fusi, F., 2016, Institutional and Organizational Factors for Enabling Data Access, Exchange and Use Aims for DivSeek, ASU/Cirad.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n° 628785/FP7-PEOPLE-2013-IOF. The authors would also like to thank Mathieu Thomas for his support and inputs to this paper.

Appendix

Appendix 1 – Survey data and social network analysis

Appendix 1.1 – Methodology

A survey was carried out in May 2016 in order to gather information on existing collaboration patterns in the cacao resources community. This survey is the result of the collaboration between Selim Louafi, Mathieu Thomas and Andrew Metter from CIRAD, Brigitte Laliberté from Bioversity International and Michelle End from the Cocoa Research Association Ltd Uk (CRA Ltd). It was developed with the objective of gathering two types of data:

- information on the respondents and their collaboration ties, in order to map out the cacao genetic resource community
- respondents' feedback concerning the CFC/ICCO/Bioversity project

The sampling list has been established based on the assumption that individuals use and/or exchange cacao germplasms and/or associated resources. **The final e-mail list, counting 391 e-mail addresses**, was the result of a selected combination of e-mail lists provided by Brigitte Laliberte from Bioversity International (contact lists from CacaoNet) and from Michelle End (INCACAO group contact list). In parallel to these provided lists, names and e-mail addresses from diverse sources were gradually collected from the early stage of the internship[1]. The final sample list is thought to include the large majority of the cacao genetic resource community.

Two main types of results from this survey will be used for our analysis. The first was a list of collaborators given by a respondent. The person filling up the survey was asked the following:

> "Please list your most frequent collaborators on issues related to Cacao genetic resources, within or outside your organization, in the last two years."

One could cite up to 30 people or none at all. Once the respondent had answered this question, the following was asked:

> "Please list what resources you have exchanged with your most frequent collaborators on Cacao genetic resources in the past two years. Check all the resource categories that apply."

Respondents had the possibility to tick one or several of the options below:

- Sharing of **genetic material**
- Sharing of **advice, information**
- Sharing of **data, results**
- Sharing **equipment, technologies**
- **Training**, **mentorship**
- Access to **networks or projects**

Therefore, a first set of data is a list of relationship ties and their corresponding set of resource type(s) involved. Three types of individual will be listed: survey respondents having cited at least one collaborator, people mentioned and included in the e-mail list, and people mentioned and not included in the e-mail list. This raw data was cleaned-up through RStudio. By crossing information from several sources, the spellings of the names were harmonized and each individual was assigned an institution from a broader list of institutions linked to cacao. This cleaned-up version of the raw data was turned into an edge list, a matrix containing two vectors: one with the names of a respondent having cited a collaborator, the other with the corresponding collaborators – hereby referred to as the edge list. Using the *igraph* R package, the edge list was converted into a graph object used to plot a network and extract metrics for our analysis.

[1] All 75 contacts identified were referenced either in the list provided by Brigitte Laliberte or by Michelle End, which contributes to our trust in the completeness of our e-mail list.

Using raw data from the survey relative to the types of resources exchanged, each edge could be given a set of resource type attribute (1 or 0 for each resource type). These edge attributes allow us to create subgraphs. However, it should be noted that an error in the online survey led to the "Sharing genetic material" option not to appear to respondents. By the time this error was noticed, some respondents had cited collaborators and were not able to elect this option. Eventually, one third of the ties reported between respondents and collaborators are concerned. While this issue does not totally prevent the extraction of valuable information from the "germplasm collaboration network", related results must be interpreted with caution.

Attributes were also given to each node on the basis of a node's affiliated institution. An institution was assigned to each node by crossing information from several sources (survey response, contact list, internet, and journal articles). Basic information on each institution was then gathered and enriched the attributes associated to each node/individual: name; type (university or college, research center, private industry including trade associations, government organization or agency, non-profit); country; region (Europe, USA, South and Central America, Africa, Asia and the Pacific, International); if they hold accessions; etc.).

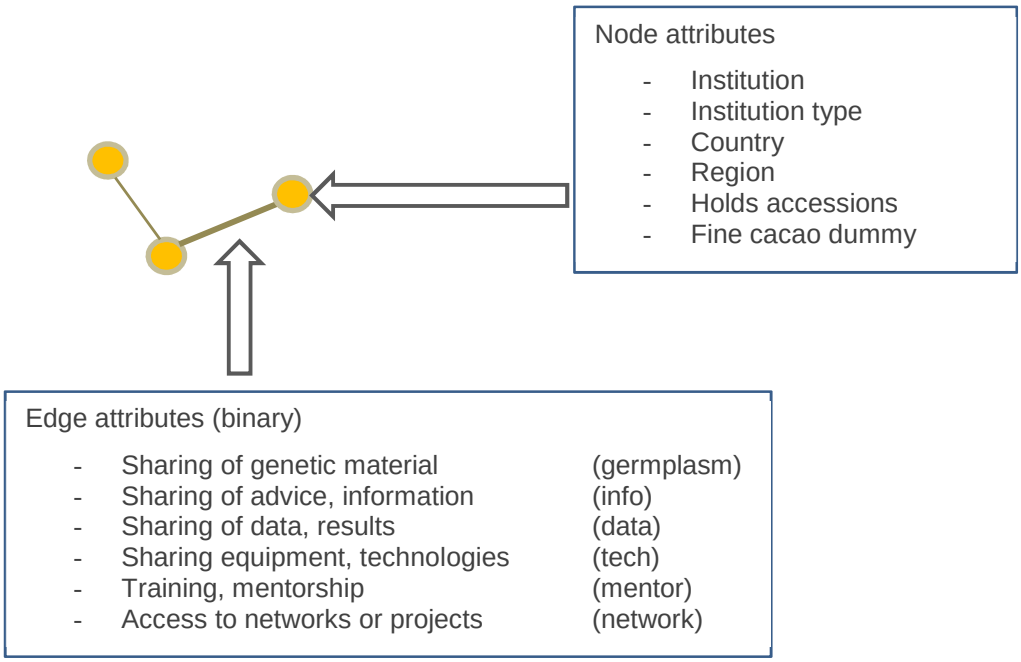


Illustration of attributes given to nodes and edges

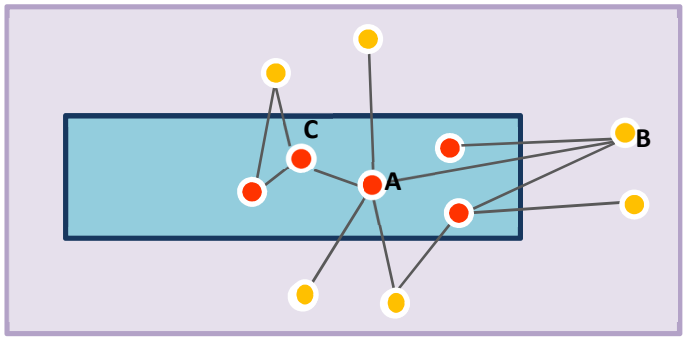
The figure above illustrates how attributes are given to our original graph network – which we will be referring to as the collaboration network. The edge attributes are used to create subgraphs, which are networks derived from our collaboration network on the basis of edges having a value of 1 for germplasm, info, data, tech, mentor or network – hereby referred to as resource type networks. Any measurement on the collaborative network can then be compared to results on these resource type networks. Degree centrality was used in this case.

Degree centrality is the number of edges one node has. In a more formal description, if we consider an adjacency matrix A with an entry (i, j) noted a_{ij}, the degree d_i of node i is then

$$d_i = \sum_j aij$$

While it is perhaps the simplest measurement of centrality, degree centrality is straightforward in identifying key actors within a network. In a directed graph, there are two types of degrees: in and out. Out degrees are edges that exit a node, while in-degrees correspond to receiving edges. If a network results from a survey, some nodes may have only been mentioned while others are respondents. In this case, a more pertinent measure may be in-degree. On the figure below, one can see that node A appears to have the highest degree: it is connected to five

other nodes. However, node B has the highest in-degree. When considering in-degrees, node C may also have a higher degree than A.



Effect of sampling on network structure

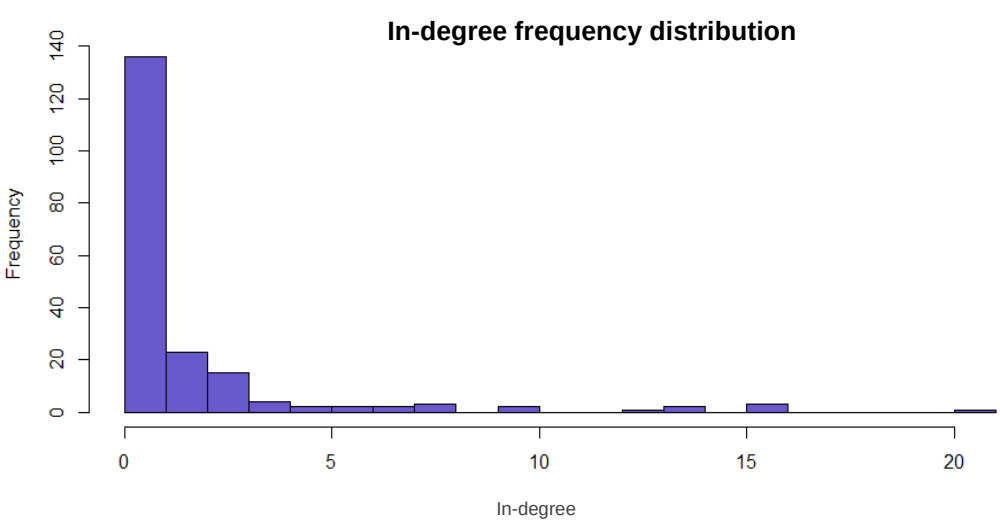
Therefore, node in-degrees D_i were measured, aggregated at the institution level and **edges connecting nodes from the same institution were discarded** so that each institution k would have their own centrality measure $\boldsymbol{D_k}$**.** This was done for all networks – collaboration and resource type networks.

Finally, the structure of the collaboration network as a whole was analyzed using a Stochastic Block Model. Essentially, the SBM algorithm divides individuals within an adjacency matrix in k groups by maximizing the probability of nodes from group k_i of actually being part of it while minimizing k. This allows for the identification of relations between identified groups and to understand what might be structuring the network. However, results were not conclusive as it appeared that information related to the way results were obtained seems to have been captured by the SBM. This also comes to show that no divergent force (such as regional divides) was stronger than the "sampling effect".

Appendix 1.2 – Results from the collaboration network analysis (survey data)

Out of **391 people included** in the e-mail list and having received the e-mail, **144 responded** (37% response rate). Of these 144 respondents, **84 listed at least one frequent collaborator**, which amounts to a **23% (84/391) response** rate for our collaboration list. The final collaboration list, regrouping respondents and their mentioned collaborators, includes **196 individuals**. Each individual was assigned an institution, to which were associated regions and types.

The in-degrees of each node are summed up in the following frequency distribution (Figure 13). The in-degree distribution follows a power law distribution of node's degree, which is a common feature of social network structure where many people tend to have few connections and a small number of nodes concentrate a high number of edges.



The table below sums up the results for the 10 highest ranked institutions based on their aggregated in-degree D_k within the general collaboration network (all resource types combined). We believe that their in-degree based on external ties (intra-organizational ties having been discarded) best captures their potential importance on the international collaboration arena.

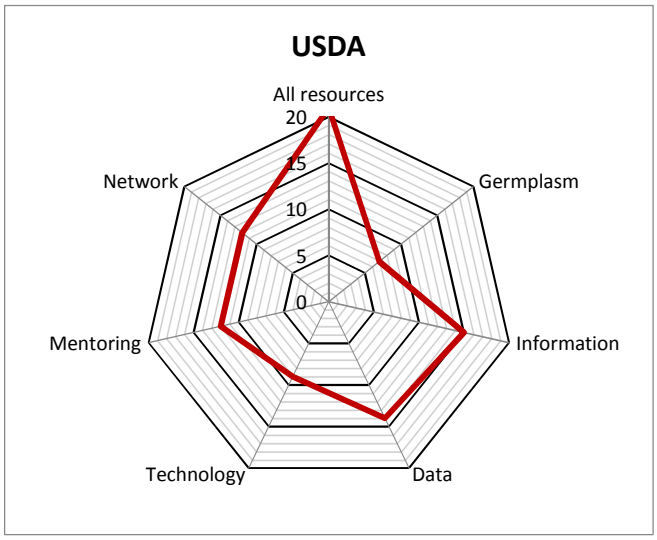
Results of centrality measures in the collaboration network (top 10 ranked institutions of 84)

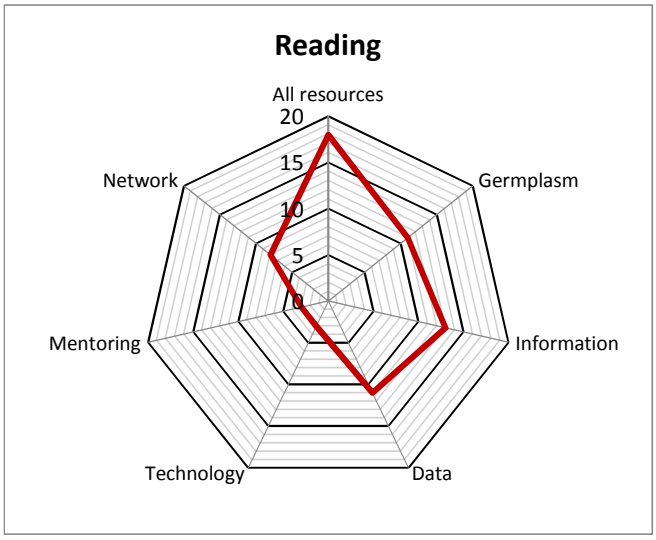
Institution	N° of Individuals	D_k
USDA	9	21
Reading*	6	18
Mars	5	17
CATIE	5	17
CRC	9	15
Bioversity	9	12
CIRAD	15	9
Penn State	2	9
CRA	1	9
CEPLAC	6	6

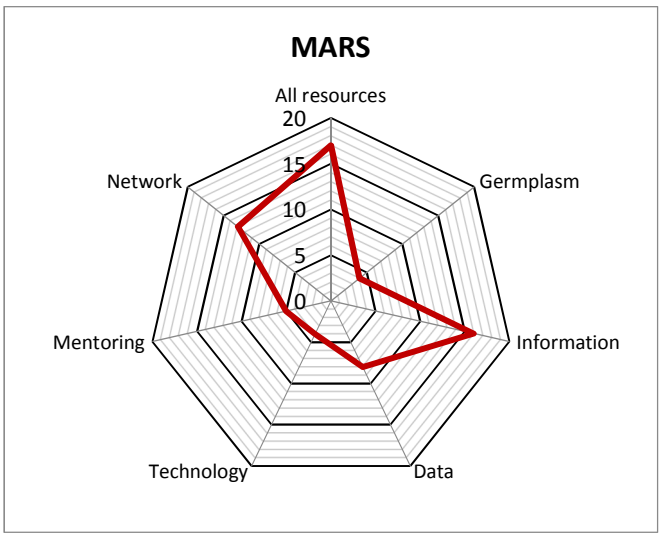
*Includes individuals affiliated to the University of Reading, ICQCR and ICGD

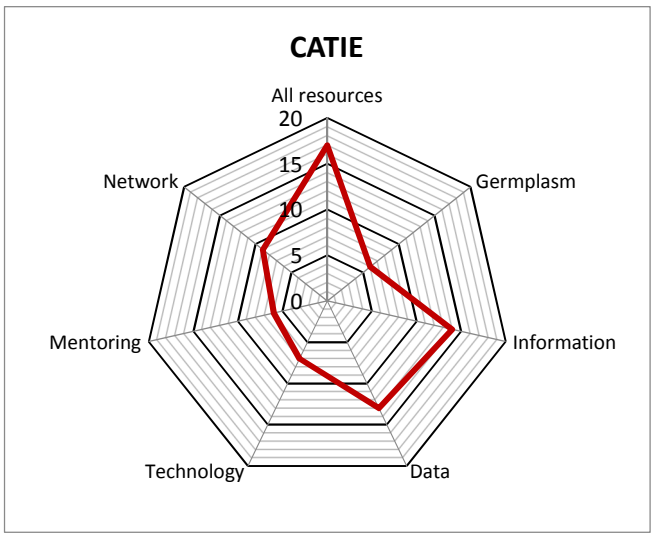
D_k= In-degree centrality without intra-organization collaboration

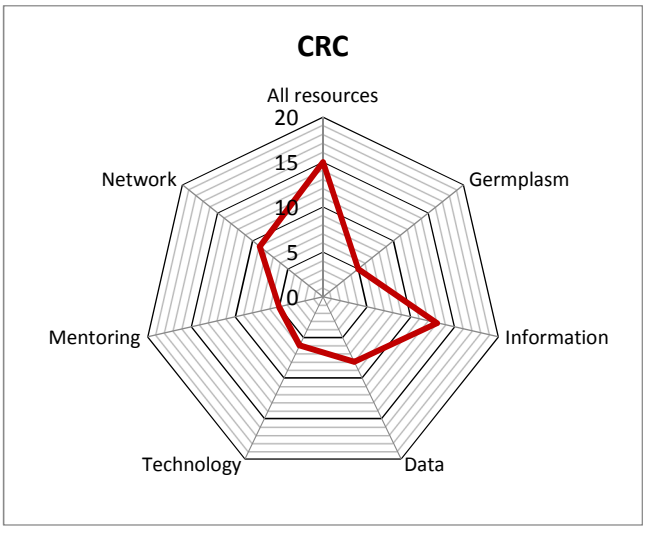
The following graphs show measures of aggregated in-degree (without intra-organization ties) for a selection of institutions in each resource-specific network. These graphs inform us on what type of resources are mostly solicited within an institution, showing its “resource provision” profile.

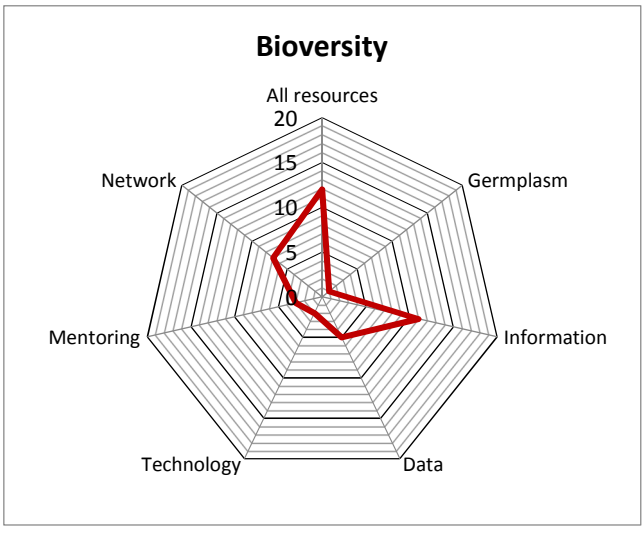

USDA
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network

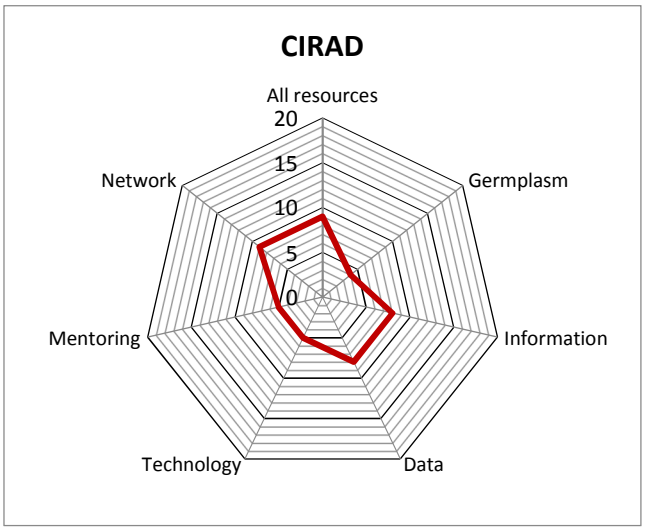

Reading
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network

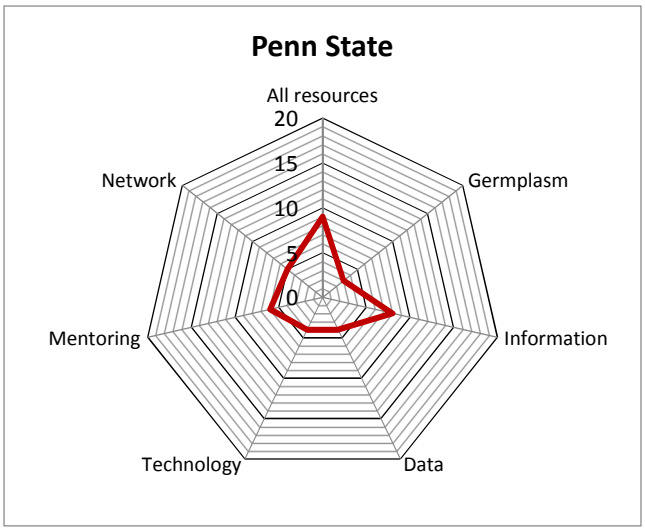

MARS
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network

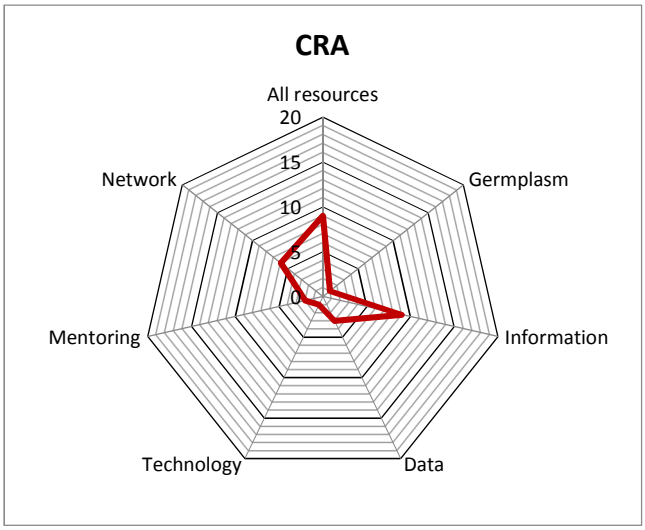

CATIE
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network

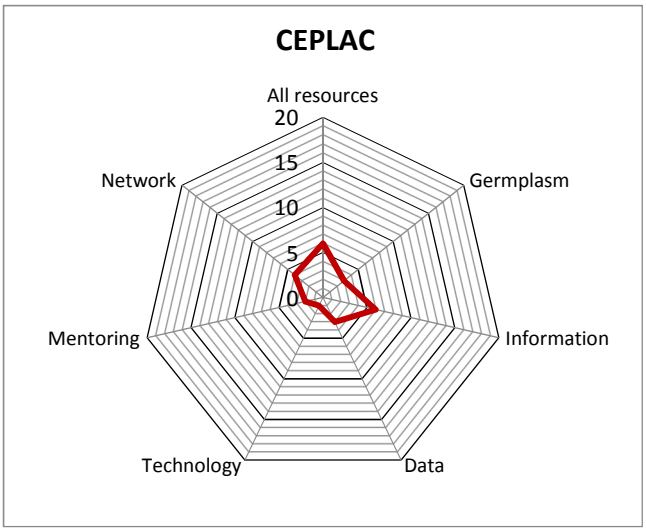

CRC
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network


Bioversity
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network


CIRAD
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network


Penn State
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network


CRA
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network


CEPLAC
All resources
20
15
10
5
0
Germplasm
Information
Data
Technology
Mentoring
Network

Appendix 2 – Bibliometric data and co-publishing network analysis

Appendix 2.1 – Methodology

Many actors involved in the use of cacao genetic material are researchers: geneticists, agronomists, plant pathologists etc. Collaboration on scientific studies creates ties between scientists, and when such studies lead to the publication of scientific articles these relationships can be materialized by co-publishing ties. Co-publication networks are often studied as proxies for collaboration in science (Wagner & Leydesdorff 2005). Therefore a bibliometric analysis using SNA tools was thought to provide interesting results for our study.

A literature research was conducted using different key words (cacao, genetic, genomic etc.) in "Web of Knowledge" and relevant scientific journals. The resulting 95 references were later found within a large reference database distributed at a Penn State symposium entitled "Frontiers in Science and Technology for Cacao Quality, Productivity, and Sustainability". This folder contains results of an online search done on May 5th 2016 of several databases for all literature with the words "Theobroma", or "cacao" or "cocoa" in the title including all years. The following databases were searched by Penn State genetic scientist Mark Guiltinan: Library of Congress, Natl Lib of Medicine, Pennsylvania State University, PubMed (NLM), Web of Science Core Collection. In total, 7422 references were obtained.

From this "cacao literature" data base, documents were selected based on the occurrence of the keywords "genetic" and/or "genomic", and on having been published after the year 2000. The year 2000 was an arbitrary choice, it seemed appropriate to capture long lasting co-publishing ties while not extending our analysis to "obsolete" relationships. After controlling for duplicates and non-relevant articles, the sample was reduced to 325 articles. The overall logic in the selection of scientific papers was that their title suggested authors may be part of the cacao genetic resource community by implying that the type of work conducted may have demanded cacao genetic material or related resources. A substantial manual post processing clean-up work was required for making these references network-analysis-ready – especially harmonizing author names. After reviewing the titles of all 325 articles, two types of research were considered to be too distant from the study of cacao genetics, and articles falling in these categories were discarded. These categories are the following:

1) Studies on the health effects of cocoa (as in cocoa powder and chocolate). Example:

   "Ibero-Baraibar, I. et al., 2016. *Cocoa extract intake for 4 weeks reduces postprandial systolic blood pressure response of obese subjects, even after following an energy-restricted diet*. Food Nutr Res, 60"

2) Studies on microbacteria involved in the fermentation of cacao beans and its processing to cocoa powder. Example:

   "Cleenwerck, I. et al., 2008. *Acetobacter fabarum sp nov., an acetic acid bacterium from a Ghanaian cocoa bean heap fermentation*. *International Journal of Systematic and Evolutionary Microbiology*, 58, p.2180–2185"

In total, 862 authors were referenced. Institutions were affiliated with as many authors as possible. Because this process was time consuming, only authors having published in at least 3 papers were taken into account. Out of all the authors referenced, 130 published in at least 3 papers. 100 of these 130 authors were eventually affiliated to an institution (in total, 100 out of 862). This is room for improvement in the processing of this data; however we believe that data on the most "influential" authors was a good start for this analysis. The set of references was cleaned-up through RStudio. The resulting adjacency matrix (authors in rows and columns) is the basis for our co-publication network. An R script was developed in order to convert this information into a bipartite matrix with authors in rows and articles in columns. This bipartite matrix was used as input for a probabilistic model called Latent Block Model (LBM) using blockmodel package in R studio (Leger 2016; Airoldi et al. 2009). In a co-publication two-mode network for instance, an LBM will identify k groups of authors and q groups of publications by maximizing the probability that authors from group k_i publish together in publications of group q_j.

For further analysis of the LBM results, betweeness centrality of authors was also taken into account. Betweeness measures how often a node falls along the shortest path between other nodes (Borgatti et al. 2013). A path is a sequence of connected nodes that never revisits a node. Betweenness centrality is measured by computing the proportion of all shortest paths from one node to another that pass through the focal node. This is done for every pair of nodes other than the focal node, and is summed into one value. It is defined as:

$$B_j = \sum_{i<k} \frac{g_{ijk}}{g_{ik}}$$

Where g_{ijk} is the number of shortest paths connecting i and k through j, and g_{ik} is the total number of shortest paths connecting i and k. High betweenness is evidence of a node's broker position: many nodes pass by him to reach other nodes (Borgatti et al. 2013)

Appendix 2.2 – Results of the co-publishing network analysis

As we explained, the LBM algorithm seeks at grouping authors and publication in K author clusters and Q by maximizing joint distribution probabilities and minimizing K and Q. Results of the LBM find 8 authors clusters (K) and 5 publication clusters (Q). A starting point is to look at this probability matrix: in rows, the 8 author clusters K, And in columns, the 5 publication clusters Q. The results can give clusters of authors that may be consistently publishing together, without having a clear type of publication associated to it, or clusters of authors that often publish together, but that have in common to publish in a very consistent type of publication, or both, and for some groups… no clear connection can be made. The results indicate that authors from group $K1$ for instance have a high probability (0,74) to publish together in publication-type $Q3$. In the case of author group $K7$, no clear association with a publication type appears, they tend to publish together in all types.

Probabilities for nodes from group K to co-author a publication from group Q

	$Q1$	$Q2$	$Q3$	$Q4$	$Q5$
$K1$	0,02	0,08	**0,74**	0,00	0,07
$K2$	**0,19**	0,00	0,01	0,00	0,02
$K3$	0,00	0,02	0,00	0,00	0,08
$K4$	0,00	0,00	0,00	0,01	0,00
$K5$	0,05	0,07	0,00	0,00	**0,79**
$K6$	0,00	0,00	0,08	0,00	0,00
$K7$	0,02	0,01	0,02	0,06	0,05
$K8$	0,00	**0,17**	0,01	0,01	0,03

To determine the "ground-truth" of these author clusters, we combined the information from this probability matrix with information on publications in Q clusters, our own knowledge on the community and help from an expert in cacao genetics at CIRAD. Our interpretation for these author clusters is the following:

Group K1: USDA and MARS

This group, composed of only 3 authors, obviously corresponds to USDA and Mars scientists that have published in many papers (most of which can be found in Q3). Results from ANOVA and Tukey tests show that the mean number of publications and betweenness centrality of this group is significantly different and higher (*p-value < 0.01)* than all other author clusters K (Appendix 2.2.1 and 2.2.2).

Group K2: BRAZIL Researchers/ Phyto-Pathology

Researchers in group K5 are all from Brazilian institutions, mainly CEPLAC and UESC. This "Brazil" team also tends to publish together in Q1 type publications. After taking a look at the type of research papers in Q1 publication

group, and consistent with an experts' opinion, this group is characterized by their work in phyto-pathodology in cacao.

Group K5: CIRAD

A small group of scientist from CIRAD, are associated with high probability with type Q5 publication, which are characterized by their large numbers of authors. An ANOVA and Tukey test show that the mean number of authors in publications from group Q5 is significantly different (*p-value < 0.01)* and higher than all other publication clusters Q (Appendix 2.2.3). This is consistent with the impression from a CIRAD scientist part of the group K5, who helped us analyze this data, that they tend to publish with a lot of people from different countries.

Group K7: Penn State and others

This small cluster of 7 authors is thought to correspond to a group of highly influential geneticists from Penn State and other highly influential scientists. They appear to be characterized by the publication of all types of papers and high betweeness centrality (second after group K1).

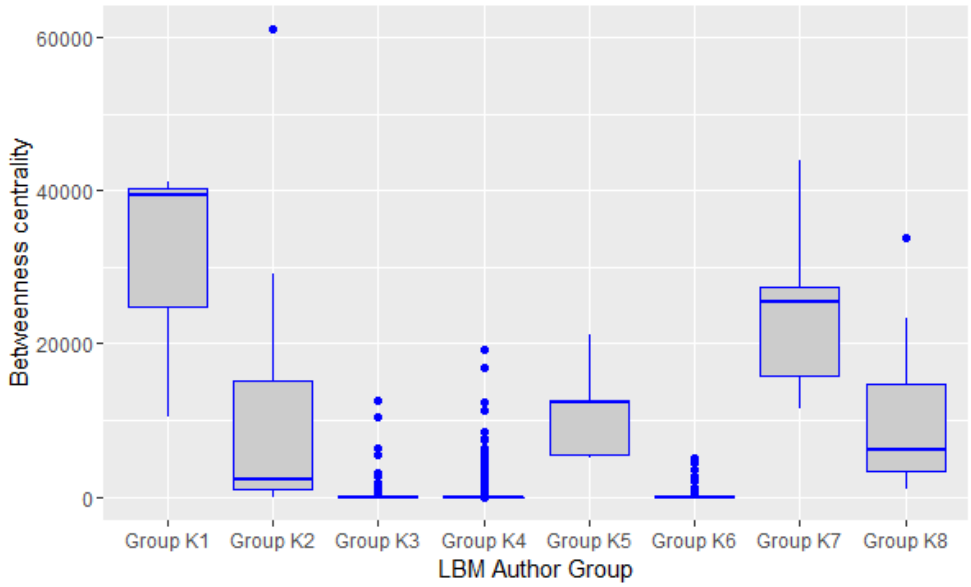
Group K8: Characterization cacao diversity

Group K8, with 15 people, seems characterized by the types of papers in which they publish. After taking a look at the type of research papers in Q2 publication group, and consistent with an experts' opinion, this group seems to stand out by their work in the characterization of cacao genetic diversity (genetic groups, diversity in collections).

Group K3, K4, and K6: ??

No clear results have come out from these groups. Small probabilities in the association matrix limit our interpretation. The group is also composed of authors with low betweenness centrality and having published in few papers (appendix 2.2.1 and 2.2.2). Lack of data on their affiliated institutions also limits our analysis. This is particularly true for group 4, which counts 672 authors out of 862 and seems to concentrate most authors having published only once.

Appendix 2.2.1 Results concerning betweenness centrality of authors in different groups K from LBM model



Results of ANOVA TEST: p-value < 0.05 At least one mean of betweenness centrality is different between K Groups

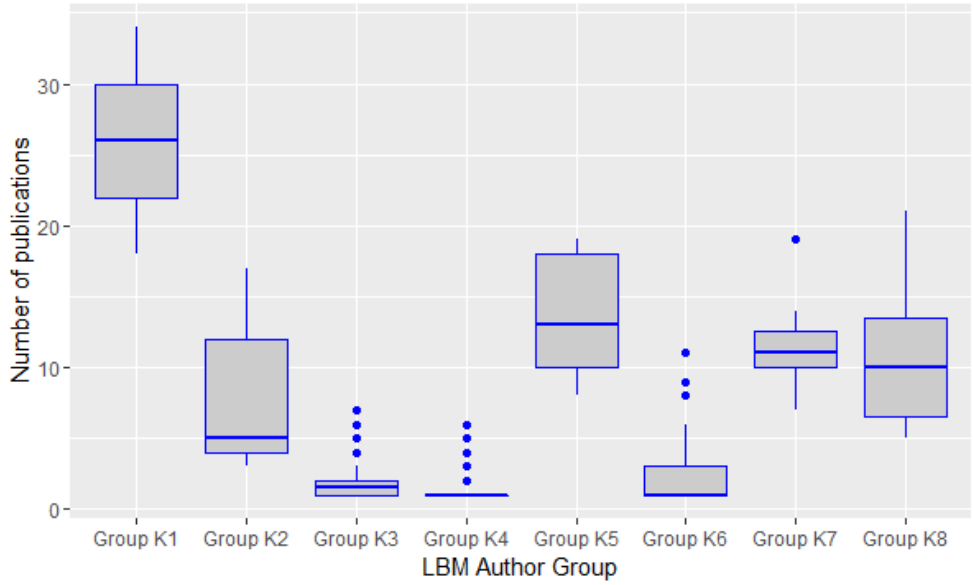
Df	Sum Sq	Mean Sq	F value	Pr(>F)

```
mbrshp        7 9.889e+09 1.413e+09   136.1 <2e-16 ***
Residuals   854 8.865e+09 1.038e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results of ANOVA Tuskey Test:

Groups,		Betweenness centrality mean
a	Group K1	30320
b	Group K7	23900
c	Group K5	11290
c	Group K8	10420
c	Group K2	10370
d	Group K3	659
d	Group K6	393
d	Group K4	367.3

Appendix 2.2.2 Results concerning number of publications from authors in different groups K of the LBM model



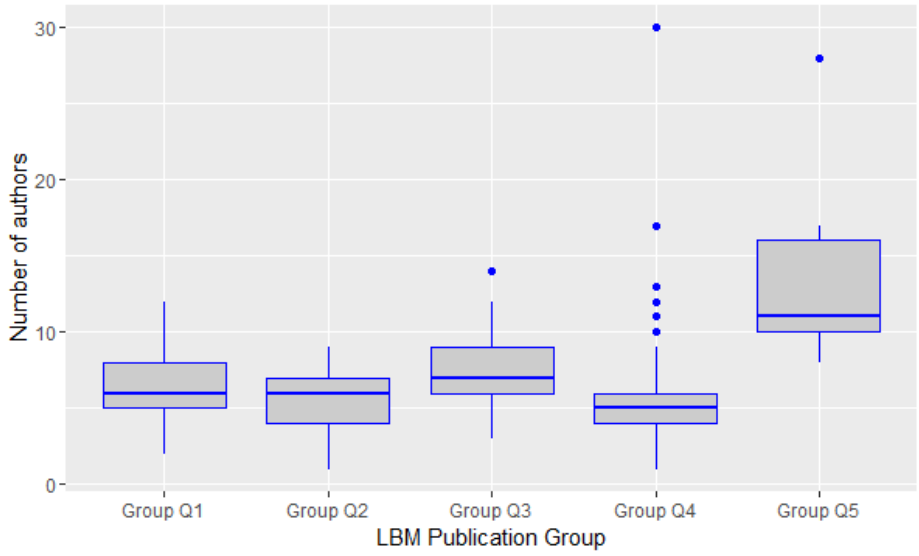
Results of ANOVA TEST: p-value < 0.05 At least one mean of number of publications is different between K Groups

```
             Df Sum Sq Mean Sq F value Pr(>F)
mbrshp        7   5044   720.6   339.4 <2e-16 ***
Residuals   854   1813     2.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results of ANOVA Tuskey Test:

Groups,		Number of publications means
a	Group K1	26
b	Group K5	13.6
bc	Group K7	11.71
c	Group K8	10.4
d	Group K2	8.118
e	Group K6	2.281
e	Group K3	1.919
f	Group K4	1.351

Appendix 2.2.3 Results concerning number of authors in publications in different publication groups Q from LBM modeling



Results of ANOVA TEST: p-value < 0.05 At least one mean of number of Authors is different between Q Groups

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mbrshp	4	718.8	179.70	19.21	**5.63e-14** ***
Residuals	280	2619.4	9.36		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Results of ANOVA Tuskey Test:

Groups,		N means
a	Group Q5	13.4
b	Group Q3	7.444
b	Group Q1	6.833
c	Group Q4	5.383
c	Group Q2	5.25

N = number of co-authors in a publication

Appendix 3 – List of institutions holding cacao germplasm accessions

From the 2012 Global Strategy, Table 2. Number of accessions in cacao *ex situ* collections (Source: Data from the CacaoNet surveys 2008-2012).

Country	Institute	Date of info	Foundation Year of the collection	No. of accessions - 2012
Benin	CRA-SB	March 2012	1986	15
Brazil	CEPEC-CEPLAC	June 2008	1967	1,302
Brazil	CEPLAC/SUEPA	May 2012	1965	2,504
Brazil	CEPLAC/SUERO	May 2012		773
Brazil	ICA	July 2011		130
Colombia	CORPOICA La Selva	FAO-VIEWS, 1998		745
Costa Rica	CATIE	February 2012	1944	1,146
Côte d'Ivoire	CNRA	August 2011	1973	1,605
Cuba	EIC-ECICC	June 2008	1982	127
Dominican Republic	IDIAF	July 2011	1974	115
Ecuador	INIAP	March 2012	1940	2,332
Fiji	Dobuilevu	SPC Dir. 2004*		115
France	CIRAD	February 2012	1985	138
French Guiana	CIRAD	February 2012	1980	508
Ghana	CRIG	August 2008	1943	1,366
Guyana	MHOCGA	July 2008	1920, 1950	65
Honduras	FHIA	March 2012	1987	31
India	CPCRI	July 2012	1970	291
Indonesia	Bah Lias	March 2012	1978	305
Indonesia	ICCRI	April 2012	1995	714
Malaysia	MCB	May 2011	1992	2,263
Nicaragua	UNAN	March 2012	2009	51
Nigeria	CRIN	August 2011	1948	1,100
Papua New Guinea	CCI	August 2011	1994	1,200
Peru	CEPICAFE	March 2012		30
Peru	ICT	July 2012	1999	607
Peru	UNSAAC	March 2012	2000	72
Peru	UNAS	February 2012	1987	422
Solomon Islands	Black Post Cocoa Unit	SPC Dir. 2004*		95
Thailand	CHRC	March 2012	1979	34
Togo	CRAF	August 2011	1968	217
Trinidad and Tobago	CRC/UWI	April 2012	1982	2,400
United Kingdom	ICQC,R	February 2012	1983	395
United States of America	USDA	August 2011	1930**	200
Vanuatu	VARTC	SPC Dir. 2004*		85
Venezuela	INIA	February 2012	1994	872
36 collections			**Total**	**24,370**

* Directory of Plant Genetic Resources Collections in the Pacific Island Countries and Territories – Secretariat of the Pacific Community (SPC), 2004.

** 1930s, re-established in 2000.