PAISPO
Provision of adequate tree seed portfolios

World Agroforestry Centre

EFCC ETHIOPIA

Norway's International Climate and Forest Initiative (NICFI)

Consultancy on application of genetic markers in BSO/SSO design, establishment and management for selected priority species

November-December 2019

TECHNICAL

Consultancy on application of genetic markers in BSO/SSO design, establishment and management for selected priority species

November-December 2019

Consultant: Ole K Hansen

Provision of Adequate Tree Seed Portfolios (PATSPO) in Ethiopia, 2017-2020

Table of Contents

Objective..... 2

Framing and mission of this report..... 2

Overall assessment and recommendations..... 8

References..... 9

Objective

The objective of the consultancy is to support the PATSPO team in best possible preparation for-and establishment in the field of each of the BSO/SSO for 2020 and beyond. This by providing an assessment of if and how application of genomic tools can contribute to more efficient and economic breeding programmes for selected species; and provide recommendations for possible testing of methods proposed in 2020 and beyond.

Framing and mission of this report

This report treats specifically the use of DNA tools to reconstruct pedigrees in forest tree stands. Such stands can be established using seed from specific seed source(s). By making phenotypic measurements on the trees in the stands and combine it with the recovered pedigree information as input to a quantitative genetic analysis, so-called quasi-field trials (QFTs) can be established and used as a breeding approach (Hansen & McKinney 2010; Jing et al. 2018; Hansen et al. 2020) – see figure 1.

QFTs provides many advantages compared to traditionally established field trials. For example, tree breeders have often experienced to lose their field trials, before producing any results, because of the long period between establishment and measurement. Some misfortunes that can ruin field trials include fires, wind throw, ice storms, insect, animal and disease damage, conversion to urban areas and vandalism (Libby et al. 1969). Common in all such occurrences is that the costs of establishing the field trial have already been paid (e.g. keeping track of separate seed lots, tagging and registration of every individual/family and producing the field trial report), without any results being derived. One of the biggest advantages of the QFTs is their flexibility and non-dependence on continuity. Contrary to traditional field trials, where staff and institutional resources are needed to establish the trials, maintain and analyse them many years later, establishment and exploitation of QFTs can be completed simultaneously, thereby reducing the risk of wasting resources on a trial that is never measured. In some cases, the tedious tasks of keeping track of separate seed lots, tagging and registration is a barrier to get the trials established at all.

By the above it also follows, that although “genomic” is mentioned in the objective of the consultancy, this report is not dealing with an evaluation of the breeding approach “genomic selection” (Meuvissen et al. 2001; Isik 2014), where DNA information (genotypes) is used to predict the phenotype of trees and subsequently make direct selection of trees. Genomic selection has not yet been proven effective for breeding and improvement in any tree species. Therefore, although genomic selection is currently a topic of interest for tree breeders working with major commercial species including eucalypts and several conifers, it is not realistic to apply for the species, which are included in the PATSPO project.

If pedigree reconstruction is to play an important role in future operational breeding of forest trees, cost is a very important factor, in order to be able to work with large numbers of genotyped trees, and this is especially true in species of minor economic value. Central to this is a relatively cheap, robust and versatile genotyping method. Therefore, the choice of genotyping method is a key issue. Simple Sequence Repeats (SSRs), also known as microsatellites markers, have been the workhorse marker type for countless studies of parentage. However, for future breeding initiatives, the potential for automation and higher throughput in the genotyping of Single Nucleotide Polymorphisms (SNPs), compared to SSRs, seems to speak in favour of the former. Not at least because SNPs can benefit from the generic sequencing techniques, with ever decreasing per-base-costs.

With the above framing and prerequisites, the main tasks that should be accomplished in order to establish QFTs for any given species, boils down to (see also Figure 2):

1. Development of SNPs
2. DNA extraction
3. Genotyping method (in praxis – 10,000s of trees)
4. Pedigree reconstruction → MAP of Breeding Seedling Orchard (BSO) (PATSPO goal)

For the actual breeding to take place, additional steps are required, but those are similar to the conventional breeding approach, namely:

5. Quantitative analysis using pedigree information and phenotypic measurements

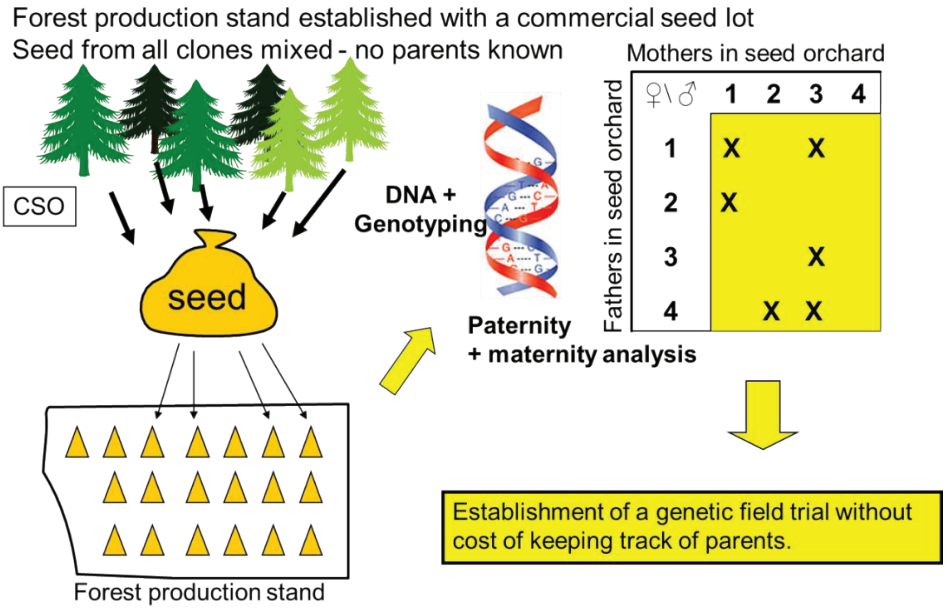

Figure 1. Illustration of the principle of quasi-field trials (QFTs) – here exemplified by a forest production stand established with seed from a clonal seed orchard (CSO). The DNA marker analysis reconstruct the pedigree, i.e. the information of parentage of the trees.

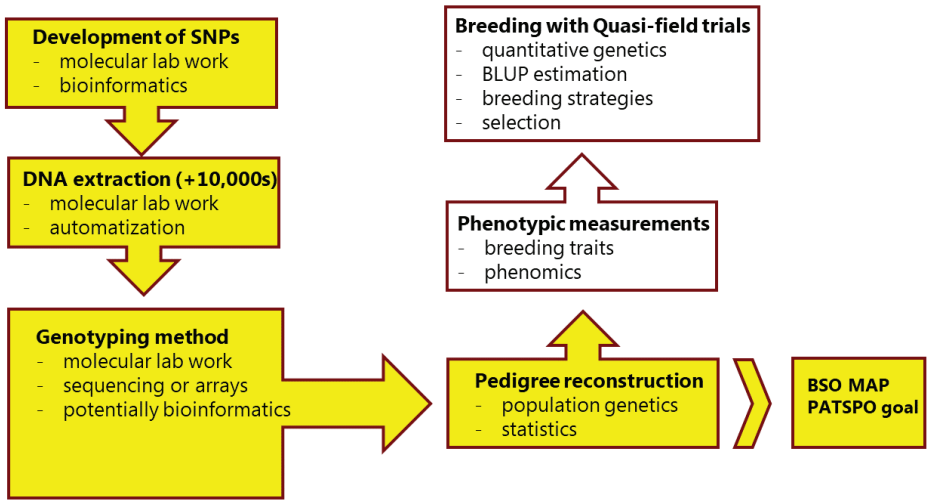

Figure 2. Flow chart for the process of using quasi-field trials (QFTs) for breeding. Yellow boxes are what this technical report focus on. The white parts (e.g. phenotypic measurements and quantitative genetic analysis) are the same processes, regardless whether DNA markers or traditional marking and book keeping of seed lots are applied.

In the following, the 4 point listed above and marked with yellow in Figure 2, are addressed one by one.

1) Development of SNPs

SNPs are single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater (Brookes 1999). Theoretically, four alleles can occur at each nucleotide position (because of the four existing nucleotide types), but in practice only two variants occur (Khlestkina & Salina 2006).

The expression "development of SNPs" actually means the identification of SNPs and surrounding DNA sequence, which are already there in a given organism. The surrounding DNA sequence is interesting because information of this is the prerequisite for obtaining the genotype of the SNP, at least for some genotyping methods. The easiness of finding the SNPs is always linked to the amount of genomic resources available for the species of interest. Therefore, if you have full genome sequence for a number of individuals of a species (e.g. 8+ individuals), you can just align (compare) some regions of the DNA sequence and look for positions where there is variation in the bases. I.e. where there is one base (A, C, G or T) in some individuals at this position in the genome and another of the four possible bases in other individuals. All this is possible *in silico* via bioinformatics. Until a few years ago, this approach was mainly reserved for model species like humans, mouse, Arabidopsis etc. However, during the last decade, a growing amount of genomic resources for many various species has emerged; this development has been caused by a steadily decrease in the cost of obtaining DNA sequence. The latter also means that to obtain sufficient genomic information to develop a smaller number of SNPs (e.g. 100-300) is now a rather affordable and manageable task, even for species with no prior genomic resources at all. For example, Delord et al. (2018) aimed to retrieve SNP markers for 40 diverse species including plants, invertebrates, fish, and mammals in a cost effective manner using RAD-sequencing. They developed a Python-based pipeline to isolate c. 100–500 high-quality SNP markers for each species that could be genotyped through classical PCR amplification

methods. A run of the pipeline applying stringent filtering parameters enabled the successful design of between 130 and 3,492 SNP markers for 30 of the 40 study species.

In this context, we have to put a comment on RAD-sequencing and it´s various forms like Genotype-By-Sequencing (GBS). The RAD-sequencing family of methods share the attributes of using restriction enzyme digests and barcoded adaptor ligation to guide high-throughput sequencing of subsets of genomes for many samples. Because DNA library preparation is simply based on restriction enzyme digest and subsequent adaptor ligation, RAD-sequencing methods can be implemented without prior genomic resources and can be used to rapidly and inexpensively generate data for large numbers of individuals (Parchman et al. 2018). In that way, it simultaneously allows SNP discovery and genotyping of the self-same SNPs, and should thereby be a good option for forest tree species. However, we do not think that it is a very robust and easily applicable method in practical forest breeding situations. One major drawback is the use of restriction enzymes, which makes certain demands on the quality and amount of DNA. The latter can be a challenge in many woody species, especially if working with DNA from non-juvenile tissue, like old leaves. Another is the high demands to bioinformatics skills. In conclusion, while we think that RAD-sequencing may be included in some approaches for developing SNPs, we do not think that it will be a viable and feasible option for the operational genotyping of the many individuals in a breeding programme – see later paragraph on genotyping.

If no current genomic resources exist for a given tree species, and one does not want to perform RAD-sequencing, another approach would be to do some low coverage sequencing of the species of interest and then align to a set of BUSCO sequences (**B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs - https://busco.ezlab.org/) from the nearest relatives that have genome sequences. Based on this alignment, targeted sequencing primers can be designed – and SNPs can be identified within the sequences. Another way could be just to design targeted sequencing primers based on a set of BUSCO sequences from all the close relatives, without the initial low coverage sequencing of the target species. A higher proportion of the designed primers would then not work in the target species, but the cost and work of the initial sequencing is avoided.

2) DNA extraction

Compared to animals, plants are in general difficult to isolate DNA from due to 1) the need to remove the cell walls of cellulose and 2) the existence of plant contaminants like polyphenols, polysaccharides, and tannins, which my all inhibit downstream applications (e.g. like genotyping).
For decades, numerous protocols to isolate quality DNA from a variety of plant species have been developed. Protocols, which can be used depending on the purpose of isolation. However, no published method has established itself as universally applicable to all plant varieties. In general, researchers need to modify a protocol or blend two or more different procedures to obtain DNA of desired quality for a particular requirement. This is due to the numerous factors that influence the methodology of genomic DNA isolation from plants, making it an art rather than a science (Varma et al. 2007).
This issue is especially problematic for the PATSPO project, which works with several dozens of different species, because there will most probably not be one single DNA extracting protocol which can cover all species.
The ideal DNA isolation protocol in relation to the goal of reconstructing pedigrees in the PATSPO BSOs should meet the following requirements:

1) Be relatively cheap per sample in chemicals and consumables, since there will be many thousands of trees to genotype.

2) Be amendable to high-throughput upscaling (e.g. 96 microtitter plate format), again because of the high number of trees to isolate from.
3) Deliver DNA of sufficient quality (purity and/or integrity) and amount in relation to the downstream application – here the chosen genotyping method.

How do we fulfill the above 3 points? In order to get a low price per sample in chemicals/consumables (point 1), it seems at first sight, as we need to use non-commercial protocols, published by the scientific community and using reagents made from scratch. No commercial companies sell ready-to-use kits at a low price. High-end kits from companies like Qiagen charge minimum around 3 US dollars per sample for their kits, and the lower-end kits are minimum around half that price (e.g. the E.Z.N.A.® Kits from Omega Bio-Tek). The per sample cost in chemicals/consumables for non-commercial protocols are often just a fraction of the cost of commercial kits – often around 10-15 % (e.g. Lamour & Finley 2006) but sometimes even lower (e.g. Spadoni et al. 2019).

If the strategy with non-commercial protocols is chosen, how do we get a high-throughput (point 2)? High-throughput is not exactly defined, but presently the DNA-lab at IGN, UCPH runs Qiagen 96 format plate columns, where the extraction of 192 samples takes around 4 hours. However, it also takes considerable time to prepare samples by putting the tissue into the 96 plates for initial homogenization. Though this first step does not need to be done by highly skilled lab-technicians, student workers or the like is fine. The IGN-lab does not use robots and the Qiagen 96 is a centrifuge- and column-based protocol. However, such protocols are not economically scalable or practical to use in a high-throughput manner; magnetic beads seem a much better option for this (Oberacker et al. 2019). However, developing, or at least adapting, homemade protocols using magnetic beads to suite ones needs is not an easy task and would require considerable intellectual input.

To which degree a high-throughput protocol can be applied is also related to point 3 – the quality (purity and/or integrity) and amount of the obtained DNA, which again is closely related to the choice of genotyping method (see later section). In general, the genotyping methods based on an initial PCR step are more robust to low DNA content and/or quality.

Another issue of importance in relation to use DNA based genotyping in large scale is the question about which tissue to use, and especially how to collect and store the tissue in order to preserve the DNA for later isolation. Two main approaches exist: a) harvested tissue are kept cool or frozen with liquid nitrogen, and thereafter kept at freezing temperatures. Or, b) samples are dried in one way or the other, and kept dry until DNA isolation. The latter option, e.g. a collection and storage method using a drying agent like silica-gel, seems the most evident choice for a project like PATSPO. Both because the freezer/refrigerator chain may be difficult to keep un-broken in all regions where samples should be collected in Ethiopia, but also because the cost of storage at room temperature, which is possible with dried samples, is much cheaper and less vulnerable than storage in freezers.

One solution may be to approach some specialized biotech companies who supply DNA isolation as a service – examples are LGC, Biosearch Technologies (https://www.biosearchtech.com/) or Omega Bioservices (http://omegabioservices.com/). Per sample prices are normally high for small and medium sized research projects with hundreds of samples. I inquired for a price from Omega Bioservices in relation to the potential for PATSPO, and they charge presently around 9-10 US $ per sample for such amounts (20.000+ samples). More interesting was the offer from LGC, Biosearch Technologies, where just pure DNA extraction (PCR quality) was around 1.5 Euro, and when including a special system for collecting and shipping the plant tissue, the cost is around two Euro for batch volumes of 10,000 samples.

3) **Genotyping method**

The genotyping method is the operational step where 10,000s of trees planted in the BSOs are to be genotyped for the identified SNPs in section 1. Evidently, this also requires a high-throughput approach.

SNP chips is the high-throughput choice for the big projects which seeks to develop genomic selection for all sorts of organisms (see "Framing of this report") and SNP chips currently constitutes the gold standard in relation to data reproducibility across samples, batches and laboratories. SNP chips are also simple and convenient to use (no need for sequence analysis capability) and are available from multiple commercial companies (Grattapaglia et al. 2018). However, that convenience comes with a cost, and a major drawback for SNP chips is the high capital cost of developing and testing them. This will not be realistic for a multitude of species with small/ limited economic interests, like most/all of the ones included in PATSPO. Furthermore, the genotyping cost of using such a chip would be quite high – maybe around 20-25 US $ per tree. This is maybe not expensive if calculated as the cost per marker, since SNP chips can genotype tens of thousands of SNPs in one go, but for pedigree reconstruction we only need in the order of 100-200 SNPs (Telfer et al. 2019).

The development of SNP-chips usually requires solid (=high coverage) whole genomic sequence of the target species. There are, however, other well-established possibilities for genotyping hundreds to thousands of markers across the genome without an available genome to start with; one such approach is DArTs (Sansaloni et al. 2011). However, two major obstacles for using this on a large scale in PATSPO would be that a) this is a proprietary technique offered by a commercial company in Australia (https://www.diversityarrays.com/), so the per sample cost could be prohibitive. And b), this approach involves the use of restriction enzymes, which makes certain demands on the quality and amount of DNA, which again may be hindering it, as it will not be possible or too costly to provide tens of thousands of DNA extractions of this standard.

Hansen et al. (2020) recently proposed and demonstrated how to use a multiplex PCR approach, coupled with high-throughput sequencing of the PCR amplicons, to genotype around 150 SNPs in *Abies procera*. Based on an initial PCR amplification of some targeted regions (identified through the development of SNP phase), this method is very robust in relation to both amount of and quality of the DNA to begin with. Specifically in relation to the latter - the relatively small amplicon sizes contribute to make this genotyping approach usable for degraded DNA.

Hansen et al. (2020) used a proprietary technique for the specific multiplex PCR amplification, namely MonsterPlex (https://floodlightgenomics.com/), but there are similar non-proprietary techniques published like e.g. Genotyping-in-Thousands by sequencing (GT-seq) (Campbell et al. 2015)

4) Pedigree reconstruction

When the reliable genotypes of all individuals are at hand (i.e. from both offspring and preferably from potential parents), the task of establishing the pedigree among the individuals is the final step. There are several analytical approaches to this parentage analysis – see Jones et al. (2010) for a review. The setup in the PATSPO situation is principally rather simple, as we have no overlapping generations in the offspring generations. In the ideal situation, we have genotypes for all markers in all offspring and available genotypes for all trees where seed was harvested. In that case, categorical allocation (also known as parentage assignment), using maximum likelihood calculations, is the evident choice. Categorical assignment approaches can handle scoring errors or mutations and can include methods for determining

confidence in parentage assignment (Jones et al. 2010). One of the most used softwares is the CERVUS software (Kalinowski et al. 2007; Marshall et al. 1998).
Several deviations from the ideal situations are likely to appear; e.g. that we do not have genotypes from all the seed trees. In cases where we have no genotypic information about the parents, sibship clustering may be the way to proceed (Huisman 2017).

Another plausible situation is that some of the target species in the PATSPO project are polyploids and not diploids; a very likely situation as polyploids are estimated to be 30–80% among plant species (Meyers & Levin 2006). In such cases, one can either chose to modify the population genetic method (e.g parentage analysis) developed for diploids to accommodate the polyploid inheritance model and the uncertainty of genotype data. Or to convert the polyploid genotypes to pseudodiploid genotypes such that many methods developed for diploids are applicable without modification (Wang & Scribner 2014).

Many freely available softwares are out there, including stand-alone programs as well as R-packages. To use them will require population genetics competences and, for some programs computer skills.

Overall assessment and recommendations

Overall, the assessment is that there are good opportunities and arguments to proceed with exploring the possibilities to apply genetic markers in BSO/SSO establishment and management for tree species in the PATSPO project. Instead of trying to implement it for all new species from a certain time point, it is evidently better to select a few test species and run those in parallel with the remaining species where normal procedure is continued.

The whole course of action, going from the development of the DNA markers (SNPs) to reconstruct the pedigree in the BSOs and later using this pedigree in the quantitative breeding and tree improvement, is a chain of interrelated events, where each step is relying on all the others. Few people in the world works on this approach, and several highly specialized techniques and competencies are required. However, all these techniques and competences are at hand, it is just the combination required that is rare. Combining those through supportive research and researcher education is therefore indeed possible, and may give PATSPO the opportunity to be an example to follow for other similar projects, and thereby expand their impact beyond the region where the project operates. The initiated pilot project on *Cordia africana*, outlined during the stay in Addis Ababa (see annex 1), is one concrete recommendation of exploring some of the practical challenges that such a new modus operandi will cause. The strategy for this pilot project is to use a commercially available genotyping technique, so the project can test some of the different aspects of the approach, but without the challenging first step of development of SNPs, which requires good bioinformatics skills, and may take relatively long time. Instead, the pilot project will give insight in other practicalities like how to physically collect and store samples, how and where should the DNA be isolated, how to analyse the SNP data coming from the commercial supplier, and how efficient are the pedigree analyses, with and without genotypes of the mother trees, etc. etc.

One of the biggest obstacles among the four steps appears to be able to get the DNA isolations made cost-effectively. The best solution may be to approach some specialized biotech companies who supply DNA isolation as a service – examples are LGC, Biosearch Technologies or Omega Bioservices (http://omegabioservices.com/). Per sample prices are normally high for small and medium sized research projects with hundreds of samples. However, a considerable volume discount may be possible, if PATSPO place an order of e.g. 60,000 DNA extractions, where dried plant tissue samples are delivered in 96-well plates.

References

Brookes AJ (1999) The essence of SNPs. Gene 234 (1999) 177–186.

Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Resources 15, 855–867 doi: 10.1111/1755-0998.12357

Delord C, Lassalle G, Oger A, et al. (2018) A cost-and-time effective procedure to develop SNP markers for multiple species: A support for community genetics. Methods Ecol Evol. 9:1959–1974. https://doi.org/10.1111/2041-210X.13034

El-Kassaby YA, Lstibůrek M (2009) Breeding without breeding. Genet. Res. 91: 111–120

Grattapaglia D, Silva-Junior OB, Resende RT, Cappa EP, Müller BSF, Tan B, Isik F, Ratcliffe B, El-Kassaby YA (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. Front. Plant Sci. 9:1693.doi: 10.3389/fpls.2018.01693

Hansen OK, McKinney LV (2010) Establishment of a quasi-field trial in *Abies nordmanniana*—test of a new approach to forest tree breeding. Tree Genet Genomes 6:345–355. https://doi.org/10.1007/s11295-009-0253-6

Hansen OK, Lamour K, Whetten R, Xu J, Cuenca A, Islam MS, Nielsen UB (2020) Ad hoc breeding of a genetically depauperate landrace of noble fir (Abies procera Rehder) using SNP genotyping via high-throughput targeted sequencing. Under review in Tree Genet Genomes

Huisman J (2017) Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. Molecular Ecology Resources 17, 1009–1024 doi: 10.1111/1755-0998.12665

Isik F (2014) Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forests 45:379–401. Doi: 10.1007/s11056-014-9422-z

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. Molecular Ecology Resources 10, 6–30. doi: 10.1111/j.1755-0998.2009.02778.x

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol. Ecol., **16**, 1099-1106

Khlestkina EK, Salina EA (2006) SNP Markers: Methods of Analysis, Ways of Development, and Comparison on an Example of Common Wheat. Russian Journal of Genetics, Vol. 42, No. 6, pp. 585–594

Lamour K, Finley L. A strategy for recovering high quality genomic DNA from a large number of Phytophthora isolates. Mycologia, 98(3), 2006, pp. 514–517.

Libby WJ, Stettler RF, Seitz FW (1969) Forest genetics and forest-tree breeding. Annu Rev Genet 3:469–494

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol., **7**, 639-655.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Meyers LA, Levin DA (2006) On the abundance of polyploids in flowering plants. Evolution 60: 1198-1206

Oberacker P, Stepper P, Bond DM, Höhn S, Focken J, Meyer V, et al. (2019) Bio-On-Magnetic-Beads (BOMB): Open platform for highthroughput nucleic acid extraction and manipulation. PLoS Biol 17(1): e3000107. https://doi.org/10.1371/journal.pbio.3000107

Parchman TL, Jahner JP, Uckele KA, Galland LM, Eckert AJ (2018) RADseq approaches and applications for forest tree genetics. Tree Genetics & Genomes (2018) 14: 39. https://doi.org/10.1007/s11295-018-1251-3

Sansaloni C, Petrol C, Laccoud D, Carling J, Deterinh F, Grattapaglia D, Kilian A (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. BMC Proceedings 2011 5(Suppl 7):P54.

Spadoni A, Sion S, Gadaleta S, Savoia MA, Piarulli L, Fanelli V, Di Rienzo V, Taranto F, Miazzi MM, Montemurro C, Sabetta W (2019) A Simple and Rapid Method for Genomic DNA Extraction and Microsatellite Analysis in Tree Plants. J. Agr. Sci. Tech. Vol. 21(5): 1215-1226

Telfer E, Graham N, Macdonald L, Li Y, Klápště J, Resende M, Jr, Neves LG, Dungey H, Wilcox P (2019) A high-density exome capture genotype-by-sequencing panel for forestry breeding in *Pinus radiata*. PLoS ONE 14(9): e0222640. https://doi.org/10.1371/journal.pone.0222640

Varma A, Padh H, Shrivastava N (2007) Plant genomic DNA isolation: An art or a science. Biotechnol. J. 2007, 2, 386–392. DOI 10.1002/biot.200600195

Wang J, Scribner KT (2014) Parentage and sibship inference from markers in polyploids. Molecular Ecology Resources 14, 541–553 doi: 10.1111/1755-0998.12210

Xu J, Nielsen UB, Hansen OK (2018a) Ad hoc breeding of *Abies bornmülleriana* for Christmas tree production using a combination of DNA markers and quantitative genetics—a case study. Tree Genet Genomes 14:5. doi: 10.1007/s11295-018-1276-7

Pilot study of pedigree reconstruction in a *Cordia Africana* BSO – the PATSPO project

During November-December 2019, a consultancy was made to support the PATSPO team in best possible preparation for-and establishment of each of the BSO/SSO for 2020 and beyond. This by providing an assessment of if and how application of genomic tools can contribute to more efficient and economic breeding programmes for selected species. Furthermore, it should provide recommendations for possible testing of methods proposed in 2020 and beyond.

Description of pilot project in 2020

In order to make a concrete initiative, besides theoretical considerations, it was suggested to do a pilot project for 2020 within the framework of the ongoing PATSPO project. In short, we suggested to test the overall quasi-field trial (QFT) methodology on the BSO with *Cordia Africana* on the campus of ILRI (F3). We will do that with the DarT-technology, which PSH already uses in the ICRAF lab in Nairobi. This technology is probably not the one, which could be used on a larger scale, but it could give data for a moderate number of individuals to demonstrate the possible feasibility of the approach. At the same time, we would get valuable information about the easiness/difficulty on how to obtain DNA from *C. africana*. It was furthermore decided to involve Kedra Mohammad (KM) as much as possible in the execution of the pilot project, as this process could be good experience and support for her potential future PhD project.

Specific plan

Target is to obtain SNP-genotypes for 188 trees of *C. africana*. The sampling of leaf samples and extraction of DNA will follow the plan outlined in the table below.

Target (# trees with good DNA)	Provenance	To be collected (# leaf samples)	# of DNA samples
4 families with 20 trees each	P31 - Adwa	4 X 30 = 120	4 X 20 = 80
4 families with 20 trees each	P34 - Harar	4 X 30 = 120	4 X 20 = 80
1 bulk collected provenance	P14 - Harar	1 X 30 = 30	1 x 20 = 20
8 mother trees to the 8 HS families*	P31 & P34	8 x 4 = 32*	8 x 1 = 8
	Total	302	188

* Maybe consider to sample leafs from all mother trees in P31 and P34 where trees have been planted out in BSOs – i.e. all around 20 mother trees at each of the two sites. Can be valuable later in a larger project/expansion.

Tasks, staff and budget

PSH and OKH will talk to the leadership of PATSPO about available funds to implement the pilot study. Below is a table of tasks, staff to do them and corresponding costs. At the time of writing (April 2 2020) the Corona virus situation has made it difficult to make very solid planning. Some task may be done by others than originally planned – e.g. the DNA extraction. Therefore, in the cost column it is often just a description of the associated cost and not a precise estimate.

Tasks	**Staff**	**Costs**	**Notes**
Collect leaf samples from mother trees in P31 & P34	KM, CATO	Transport within Ethiopia. Internal staff costs PATSPO. 1-2 weeks.	Consider to collect from all available mother trees
Mark up trees in BSO F3	KM, CATO	Internal staff costs PATSPO. Max. 2 days.	Do not choose families before verified that mother tree can be sampled + sufficient number of trees in the BSO for this family
Collect leaf samples from progeny in F3	KM, Technician from ICRAF Nairobi	Internal staff costs PATSPO. Max. 2 days.	Transport of technician. **QUESTION: Is technician from Kenya actually needed?**
Transport of leaf samples to ICRAF Nairobi	KM	Travel expenses KM.	Maybe this step will be skipped – if we choose to extract DNA at ILRI or to use a private company.
Extraction of DNA at ICRAF Nairobi	KM, Technician from ICRAF Nairobi	Depends on which solution is chosen.	Costs both to chemicals/extraction kits as well as to the staying of KM at ICRAF Nairobi.
Shipment samples and arrangement of DarT genotyping	PSH	Shipment of DNA on dry ice: maybe 500 US$ Genotyping by Australian company: 2-3,000 US$	
Filtering of raw genotyping data	PSH, KD, OKH	Internal staff costs PATSPO.	
Pedigree reconstruction	KM, OKH	Internal staff costs PATSPO.	
Report	KM, OKH, PSH	Internal staff costs PATSPO.	

During the consultancy visit in November 2019, CATN, KM, PSH and OKH visited the BSO (F3) with *Cordia Africana* on the campus of ILRI (F3).

Observations: The BSO is well established with a very high survival rate (>95%). Large efforts have been done to ensure this; e.g. manual watering and weed control but also ditch construction to prevent water logging during the rainy season. Due to the very good condition and survival rate, this BSO, located on the

Appendix 1 - Pilot study of pedigree reconstruction in a *Cordia Africana* BSO

ILRI campus, is an evident study object to do a pilot study – an activity that we subsequently suggested – se later for further description.

Whole-day trip to see the BSO (F4) with *Cordia Africana* in Suba Mountain.

Observations: the BSO F4 is also well established with a very high survival rate (>95%). The plants are generally smaller, but for harvest of tissue samples in relation to DNA studies, this is not important. If it shows up that sampling from F3 cannot take place in due time – e.g. no new shoots in February 2020, this BSO can be a very good substitute. The BSO is also of the closest to Addis Ababa – in case other new sub-projects for 2020 should include quantitative genetic analyses of BSOs.

Current ideas on the table is something about quantitative genetic analyses of the first BSOs - a sub-project which could also underpin the (hopefully) PhD scholarship from KU to KM. This will be discussed with Jon K Hansen (JKH) and Jing Xu when coming back to DK. Another subject was new methods to phenotyping trees in the BSOs – also to be discussed with JKH.

PATSPO
Provision of adequate tree seed portfolios

World Agroforestry Centre

PATSPO/ICRAF Office
c/o ILRI Campus, Gurd Shola, P.O. Box 5689,
Addis Ababa, Ethiopia

Phone: 251-116172000
ext. 2491
Email:K.Hadgu@cgiar.org

Website: https://www.worldagroforestry.org/project/provision-adequate-tree-seed-portfolio-ethiopia