Primary Outcome 2*. Improved efficiency of yam breeding programs through use of faster and more precise tools and methods*

Deployment of innovative molecular-marker aided techniques requires fundamental resources such as whole genome sequences, genetic maps, and predicative markers associated with desirable traits. Early efforts to develop linkage maps and to identify QTLs used first generation molecular markers such as RAPD and AFLP (Mignouna et al., 2003; Mignouna et al., 2002a; Mignouna et al., 2002b) were developed but the utility in marker-aided breeding has been limited. More recently, EST sequences were developed from a bi-parental populations segregating for anthracnose disease, which led to the discovery of 1152 EST-SSRs (Narina et al. 2011).

However, utilization of the EST-SSR is at initial phase. The latest endeavor in developing yam genomic resources is being undertaken by a joint project by IITA-IBRC-JIRCAS aimed at developing whole genome sequencing of Guinea yam and their relatives using Illumina GAIIx and HiSEQ DNA sequencers. While the assembly of the draft genome is near completion, the daunting task of gene annotation and post-genome resource development is the next crucial step.

In parallel with the de novo genome sequencing, the collaborative project has resequenced 10 Guinea yam genotypes for discovery of SNPs. In addition, the de novo sequencing of the parental genotypes for anthracnose disease (*D. alata*) has been completed and the progenies are being genotyped using GBS in a collaborative USAID-Linkage project with USDA-ARS, Stoneville and CUGI, Clemson to develop additional genomic resources.

IITA yam breeding has been developing multi-traits bi parental populations of both water and white yam. Multiparent (half –sib) populations have also been developed for white yam. These populations combine yield, shape, multiple tubers, YMV resistance, tuber flesh, dry matter content, oxidation, cooking quality, tuber earliness, and storability, high zinc and iron content. For water yam, populations combine anthracnose and viruses' resistance, tuber quality after cooking, yield, shape and aerial bulbil production.

Generation of new multi-trait mapping populations of *D. alata* and *D. rotundata* will be carried out using bi-parents, open pollinated and sexed diallel crossing blocks. For any breeding program phenotyping techniques need to be accurate and standard uniformed techniques followed through all environments. F1 seed of new populations will be generated from parents selected. Applying the phenotyping techniques in three contrasting environments in Nigeria and representative yam production areas in the other countries of the project will generate phenotypic data to be used in association analysis. Seedlings obtained from F1 seed will be multiplied using the vine cutting technique two months after sowing, obtained vine cuttings will be transplanted in nursery bags and two month after transplanting, vine cutting will be harvested again to increase the number of plants at least five times per clone in one year. Tubers obtained from this system will be used for phenotyping through the selected environments. Seed set from parents will also be multiplied using the same technique and also planted in the same environments.

Based on the re-defined environments two contrasting environments for each trait will be selected for phenotyping.
Experimental design will
a) Develop experimental questions or hypothesis,
b) Define variables (treatment, checks and response),
c) Define experimental and sample size (experimental units, sampling unit),
d) Estimate sample size and error, and
e) Randomization and layout.

IITA has generated several mapping populations for targeted traits such as anthracnose disease, tuber quality, earliness and others. We propose to use the whole-genome re-sequencing (WGRS) technique in which only two bulks representing extreme phenotypes for the targeted trait will be sequenced. This will reduce the cost and enable rapid QTL mapping and gene isolation using methods such as QTL-seq (Takagi et al., 2013) and MutMap (Abe et al., 20112) that have been developed at IBRC, Japan. Both QTL-seq and MutMap depend on WGRS of bulk DNAs of progeny derived from bi-parental crosses. Using these WGS-based methods, > 20 novel genes have been isolated and > 10 QTLs identified in rice. With the QTLs and genes identified, additional DNA markers will be developed and used for marker-assisted selection (MAS) and for pyramiding of useful genes in elite cultivars.

The activities will include selection of those bi-parental populations for which good phenotyping data is available so that the progenies could be bulked into two extreme phenotypes. WGRS will then be applied on those two bulks for further QTL analysis and isolation of candidate gene (s).

We will use selected bi-parental populations that are segregating for traits of interest and apply QTL-seq (Takagi et al., 2013) to rapidly identify QTL by whole genome re-sequencing (WGRS) of two bulked DNAs. Briefly, in F1 population of e.g. 250 progeny, we score the phenotype and then select 20 individuals showing high trait value and additionally 20 individuals with low trait value, and prepare two bulked DNAs, respectively. These bulked DNAs are subjected to WGRS for ~30× genome coverage and aligned to the reference genome sequence of either of the parents. The alignment data is used to plot a graph (SNP-index graph) relating the SNP ratio of the sequence reads and to genomic positions. Genomic regions showing contrasting patterns of SNP-index graph between high and low bulks reveal the QTL positions governing traits of interest. This is an accurate and cost-effective method of QTL identification only made possible by WGRS. We may even be able to identify the genes responsible for the trait.

We will apply QTL-seq to existing IITA Guinea yam families derived from at least five bi-parental crosses for target traits and further extend the application to the families generated during the course of this project. This will result in development of DNA markers for use in breeding programs for marker assisted selection.

Although quality is an important selection criterion in breeding programs, the genetic basis of characters that determine the quality of tuber is not known. Several characters involved in the variability of quality are measured routinely in breeding programs (color and oxidation of the flesh, tuber shape). The acceptability of any newly developed variety depends also on its organoleptic properties. Organoleptic qualities depend on several physico-chemical characteristics, the most important are the starch content, dry matter and sugars (Martin, 1974 Lebot et al., 2006 Lebot et al., 2009). Martin (1974) observed that most appreciated varieties in Puerto Rico contains high levels of dry matter, and that it is associated with a fine structure of the flesh and dense feel. Lebot et al. (2006) found that the most appreciated varieties in Vanuatu are characterized with high dry matter and high starch content.

In West Africa, where the dominant food form of yam is pounded, *D. rotundata* is usually preferred to *D. alata* owing to its ease of dough formation when pounded. However, observations from IITA have shown that certain genotypes of *D. alata* have the ability to form good dough, comparable to or even better than that of some genotypes of *D. rotundata*. This has important implications for food security, especially as *D. alata* has a greater agronomic flexibility, considering the increasing challenge of deterioration in soil fertility (Egesi et al, 2003). The capacity of varieties to be pounded it is related to the starch content.

Approaches: QTL meta-analysis will be conducted in two mapping populations of 150 individual each, maintained at CIRAD by crossing contrasting diploid progenitors. Phenotyping of characters commonly evaluated in selection schemes (oxidation of the flesh, flesh color, tuber shape) as well as several physico-chemical characteristics (sugar content, starch content, dry matter, protein) to identify a maximum of genomic regions involved in different quality traits. The physico-chemical analyses will be carried out at the Laboratoire d'Analyses Agricultural Teyssier, Bordeaux, France, according to AFNOR (French Association, the Association of French standards) following EEC methods.

IITA has 2368 clones for which there are several years of historical phenotypic data generated in several environments and they still exist in the breeding collections. Working collections from national programs also have years of phenotypic data from a single site. Phenotypic data have been generated through preliminary clonal evaluation, yield trials, advanced yield trials, uniformity trials, and more recently from family replicated trials.
This data will be curated by assembling, checking and standardization.

The availability of clones with historical data is summarized in 2368 clones of *D. rotundata* from mapping populations, clones under resequencing, crossing blocks, advanced yield trials, family replicated trials and clones from the core collection. For *D. alata*, 1860 clones will be available from the same set of trials as described for *D. rotundata*. A total of 2293 clones from the four countries of the project will be incorporated in this process.

After assembling data from each source, it will be curated for quality and standardized using the Yam Crop Ontology generated within the GCP program for further distribution and sharing with different partners and analysis.

The potential for increasing yam productivity through use of molecular markers in breeding is well acknowledged, especially given the lengthy breeding cycle and challenges associated with field-based phenotyping of the crop. Additionally, more targeted use of the elite genotypes routinely used in breeding requires a better understanding of the genetic structure and fine-scale relationships. But progress in these two areas has been limited due to lack of adequate genomic resources for yam and it is for this reason that a complementary funding project was initiated by the Roots, Tubers and Bananas (RTB) CGIAR Research Program (CRP) to employ genotyping-by-sequencing (GBS) to generate SNP data for 488 yam breeding clones from IITA.

GBS uses a bioinformatic pipeline to call SNPs from next-generation sequencing of reduced-representation, bar-coded libraries. Because of the efficiency, high marker number, low cost, and lack of bias in GBS data, we chose GBS as our basic marker system. We are using a protocol developed in the Buckler lab at Cornell University (Elshire et al. 2011). For this project, we would genotype additional 1000 clones from the participating national programs. The GBS approach is expected to yield at least 20,000 SNPs markers.

Using *D. rotundata* genotypes that represent the genetic diversity of breeder's collection, we will carry out whole genome association study (WGAS) using whole genome sequencing (WGS). The research activities include

1) cultivation of 300 genotypes under different environmental conditions to obtain accurate phenotype data,
2) WGS-based WGAS analysis. The activity 2) comprises sequencing of each accession to depth of > 10× (~ 6Gb). The generated sequence reads are aligned to Guinea yam reference sequence, and the association between trait value and SNPs are assessed over the entire genome. In view of the estimated low levels of linkage disequilibrium (LD) of Guinea yams caused by obligate outcross associated with the dioecious nature of *Dioscorea*, GWAS requires extremely high density of DNA markers. We propose that WGS is the best choice to attain this objective. We will also be able to use WGS data to make contigs representing genomic regions specific to a given genotype, which would facilitate identification of QTL/gene from a wide range of genetic resources. Following WGAS of 300 genotypes, GBS data of additional lines could be used to impute their genome sequences.

The first three years will involve whole genome re-sequencing of 100 genotypes each to identify QTL (s) for target traits. In the fourth and fifth year, the markers for targeted traits will be identified and used in breeding programs for marker-assisted selection.

Genetic mapping of loci underlying important quality and agronomic traits has not been extensively carried out in yam making most selections to be carried out through phenotyping. We propose to carry out genome-wide association (GWAS) to rapidly identify markers linked to desired phenotypes in breeding collections.

GWAS is known to be an efficient way of determining genetic basis of complex traits. All that is required are quality phenotypic data for traits of interest and high density SNP data from a panel of genotypes. The phenotypic data will be from historical evaluations carried out in the IITA breeding program and augmented with additional phenotyping carried out in this project. The use of broader genetic variations available in entire breeding collections in association mapping enables many alleles and traits to be evaluated simultaneously.

The power of association mapping depends on the population size and density of markers across the genome. The latter is to ensure that there is at least a single SNP in strong linkage disequilibrium (LD) with the causative mutation for the target trait. The number of desired markers depends on the extent of LD in the breeding germplasm. If the LD decays very rapidly, much larger number of markers is required to saturate the genome. To achieve this density, IBRC and IITA propose to carry out whole genome re-sequencing of 300 yam genotypes that capture most of the haplotypes in the regional breeding collection and then use imputation approaches to interpolate the SNPs from WGS to the entire population that will be genotyped through GBS.

The ultimate objective of the genetic mapping efforts is to employ identified markers in selections and bypassing the need for phenotyping. All markers identified as linked to desired traits will be converted to breeder-friendly and readily accessible genotyping platforms like LGC Genomic's KASP assay. These will then be validated in yam germplasm, not included in the discovery panel, which have good phenotypic data.

Molecular breeding approaches will certainly result in the generation of large sets of genomic, genotyping, and phenotypic data which entails efficient data management system. An integrated database that combines storage of diverse data as well as bioinformatics and statistical analysis pipeline is imperative.

An open access yam database (YAMBASE) to host pedigree data, phenotypic data from field trials and laboratory assays, and molecular marker data has been created; this will integrate breeding decision making tools and genetic analysis pipelines. Yambase can be accessed here: https://yambase.org/

There will also be an electronic forum for breeders and scientists to exchange information, email lists, and Wikis, as well as links to twitter feeds and IRC channels (real-time Internet text messaging or synchronous conferencing) that will enable communication between project members. The software infrastructure will be based on software from SGN, which has developed genome and breeder specific databases for over ten years. All code is released as open source code on the web (http://github.com/solgenomics/).

The database will endeavor as much as possible not to duplicate other already available tools but interphase with such projects as the Integrated Breeding Platform (IBP) and the android-based Field Book app for field data collection. The latter is designed to efficiently gather data in the field using mobile devices. Standardized entry for phenotypic data will rely on available yam trait ontology. The West Africa Regional Hub of the Integrated Breeding Platform (IBP) is being set up at IITA headquarters through an agreement with CIMMYT (for the Generation Challenge Program).

Since genotype data will be applicable in all locations where a yam clone is tested, so data sharing through a secure and well curated database is necessary for international sharing of germplasm and breeding applications. The proposed database infrastructure will allow breeders to manage the anticipated genetic mapping studies in bi-parental populations, genome-wide association studies and future Genomic Selection (GS) process through the website by uploading data for their breeding populations, including phenotypic and genotypic datasets. Quality control of phenotypic data will be performed using simple statistical analyses for the detection of outliers.