Mining the gene networks involved in water-stress response in bambara groundnut: A machine learning approach to translating traits in model species to minor crops (Code: Bam1-004).

Venkata Suresh Bonthala

Background:
Bambara groundnut (BG) is an underutilised legume crop and is cultivated as landraces. This crop shows good drought tolerance and has been proven to nutritionally valuable. But so far the crop is under-studied and there is insufficient knowledge available to develop high yield or biotic/abiotic stress resistant varieties.

The model species/crops, *Arabidopsis*, *Oryza*, *Medicago*, etc., have been well studied using various molecular technologies. So far a number of resources have been developed and huge amount of knowledge gained from various research experiments for these model species/crops. The resources and knowledge developed for these model species/crops could be used to help to translate the acquired knowledge into the minor crops to investigate various molecular mechanisms which are biologically significant.

Aim:
I would like to investigate the gene networks involved in response to water-stress in bambara groundnut by translating knowledge developed for the model species/crops like *Arabidopsis*, *Oryza*, *Medicago*, etc., using machine learning and/or Bioinformatics approaches (SVM, NN,Rule-based approaches, etc.).

Objectives:
1. Identification of candidate genes and construction of gene networks.
2. Identification of functional modules and functional annotation.
3. Sequence analysis, phylogenetic and evolutionary analysis.
4. Comparative genomics.
5. Validation of identified candidate genes.
6. Development of web-based resource based on this study.
7. Development of eQTLs (expression quantitative trait loci)

Approaches:

1. Identification of candidate genes and construction of gene networks:
It is possible to identify functionally similar genes of model species in BG based on the hypothesis that the functionally related genes tend to be transcriptionally coordinated i.e., co-expressed. To investigate such relationships, I would like to perform the following tasks: (1) identifying candidate probe sets (genes) which have same expression pattern in both model species (for example: *Medicago*/*Arabidopsis*) and in BG in same stress condition (i.e., translating knowledge from model species to BG). (2) Mapping of identified candidate probe sets to BG and model species genome/transcriptome to further confirm their candidature and to reduce the number, and (3) construction of co-expression network individually for both model species and BG using network-based approaches.

2. Identification of functional modules and functional annotation:
The physical manifestations of stress are usually organized as relatively separable modules of highly

interconnected genes in the co-expression networks and we need to partition them into biologically significant clusters. Graph clustering techniques are ideal for this purpose. For example: Markov Cluster (MCL) algorithm. Further, we will carry out functional annotation of partitioned clusters to identify sets of functionally related genes based on the high gene connectivity in expression.

3. Sequence analysis, phylogenetic and evolutionary analysis:
To further explore the identified candidate genes, I would like to analyze them with respect to their phylogenetic and evolutionary analysis and, basic sequence analyses.

4. Comparative genomics:
To further know whether the identified candidate genes have any orthologous relationship with closely related species, we will map the identified candidate genes on to the closely related genes.

5. Validation of identified candidate genes:
Gene/functional modules could be validated using quantitative RT-PCR to further strengthen the evidence that these are involved in the plant's molecular response to a particular stress.

6. Development of web-based resource based on this study:
Finally, I would like to develop web-resources to make sure the availability of the results of this study to the research community around the globe.

7. Development of eQTLs:
Based on the available expression data (either NGS based transcriptome or Xspecies microarray data), it is possible to develop/find the QTL which are responsible for survival of the plant under water stress.