CASSAVA DATABASE WORKSHOP
Chaired by Morag Ferguson; Meeting notes taken by Ismail Y. Rabbi

JT - Joe Tohme
MG – Melaku Gedil
GMcL – Graham McLaren
SP – Simon Prochnik
SR – Steve Rounsley
LM – Lukas Mueller
EP – Elizabeth Parkes
RB – Rebecca Bart
WW - Wenquan Wang
JY - Jun Yang
KK – Kathy Kahn
CH – Claire Hershey
PK – Peter Kulakow
WG – Wilhelm Gruissem
JD - Jorge
IN - Inosters Nzuki
MF – Morag Ferguson
HK – Heneriko Kulembeka
JL – Jim Lorenzen

Introduction by the Chair:
There are a number of database initiatives and enormous amounts of new data becoming available through next-generation sequencing (NGS). Some of these resources are documented by Ayling et al. (2012) (Information Resources for Cassava Research and Breeding. Tropical Plant Biology 5:140-151). In one way the cassava community is in a fortunate position that relatively little sequence information exists and there is an opportunity to get databases established, organized and integrated before an enormous amount of data is generated. These developments raise a number of questions relating to how the cassava community is going to store, integrate and share these data.

Purpose of the meeting:
1. Inform different users in the cassava community of the different database initiatives
2. Discuss options for database organization, integration and population

Presentations:
Below are the presenters and the key points of their presentations:

Graham McLaren - presentation
Presented on the Generation Challenge Program's (GCP) Integrated Breeding Platform (IBP)
Summarized information cycle for crop improvement into the following groups:

- Genetic resources; Environmental characterization; Crop information (pedigree, phenotype data)
- Genomics and genetics databases

These databases can be developed by different teams; the key element of cassava information system would be linking these databases together.
How do plant breeders want to access the information from the other domains and link them with the information that they collect themselves?
GCP develops tools to allow breeding programs to manage the logistics of their breeding projects and their day to day breeding information, one of which is having access to public crop information. There is a need for coordinators who will distill down/ filter the information that will be useful for breeders. The GCP is trying to establish a network of crop-lead centers to do this work.
The key to integrating crop information is trait ontologies for different crops.
GMcL stressed that there should be a commitment to adopt public ontologies and other standards out there (or develop new ones if not already available).

Genealogy management system:

- Can be used to calculate the coefficient of parentage
- Manages sample tracking

Phenotyping

- Most complex data (volume is rather small)
- Factor (variate that is measured) is annotated with the trait ontology, the scale or units that it is reported in, and the protocol by which it is measured. These components are managed by the community through ontologies and ensure cross-linkage of data from different studies.

Genotyping:
Currently generating a large volume of data; IBP not intensely focusing on the area of genomics data.

Q&A
SR: What is the capability of the GCP platform on handling the genomics data?
GMcL: Have the 'Genotyping Data Management System' (GDMS), a database that handles a small to medium scale volume; i.e. 10's or 20's of markers on 1000 lines.
No capability to go into genome-wide markers / selection.
KK: What part of the database system at GCP is functional already?
GMcL: All the geneology and phenotyping data management is functional now. Tools for management of genotyping data are not operational yet, but will be in the next 6 months.

Peter Kulakow – presentation
IITA in cooperation with CIAT and NRCRI are developing the ICASS database.
Cropontology.org has a list of 120 traits agreed upon by the aforementioned partners.
The cassava trait ontology originated with the publication of Clair Hershey's publication in 1983 with the IBPGR cassava descriptor list.
Implementing ICASS takes a lot of effort
Invested heavily in developing the crop ontology
These different databases should be able to communicate

Lukas Mueller - presentation
Presented on http://www.cassavabase.org
This database, currently unfunded, is hosted by Boyce Thompson Institute (BTI) located at Cornell University. This database has been purposefully developed to enable genomic selection in cassava. The genomic selection algorithm will be developed so that much of it will be implemented in the database.

Maps and markers module: a powerful map browsing and mapping tool. Genetic map developed by Dr Rabbi at IITA has also been loaded.

Phenotype module: can enter phenotypes, photographs, comments, literature references etc. The database currently holds phenotypic and pedigree data from the IITA breeding program (>1200 accessions, approx. 40K phenotypic measurements; Data not currently visible to outside users).
Other database functionalities include: Gene module, Pathway module, Genome and sequences model, and Breeder's toolbox.

Q&A
SR: is the GBS data handling in place?
LM: This is in place but is yet to be tested.
CH: who is sponsoring this and what is the long term sustainability?
LM: This is part of a consortium formed by the Cornell University within the frame-work of the Next Generation Cassava Breeding project.
PK: The database will be hosted at IITA, in order to give it a future beyond the project.
KK: There was an existing grant on wheat and maize to which additional funds were allocated to test the GS approach in cassava. The new project will give the database the opportunity to grow.

Simon Prochnik - presentation

- Presented on 'Phytozome', the database and portal hosting the genome database.
- There are 30 completed plant genomes, and more coming. The database is a place for biologists to browse genome data. JGI bioinformaticians analyze the data and present the results in a format that is useful for biologists.
- Everything on phytozome is fully accessible and downloadable (gene, gene families, genome sequences, etc).
- Run automated gene annotations.
- Search methods: text (name) or sequence (BLAST).
- Showed a range of useful features of phytozome.

Whats coming in cassava?
Version 5; dense genetic maps, RNSseq data, cDNA sequences from RIKEN; better assemblies (chromosome-like scaffolds)

Fully digestible data versions (i.e. SNP data) will be in cassavabase or the IBP platform.

Q&A
CH: Who is sponsoring this effort?
SP: Said he works for US DoE - JGI. He works on a large portfolio of crops, including citrus and algae, peach. He would like to have a more formal commitment in the grant being developed by Kathy. Whatever happens, he and JGI will ensure that cassava will have an updated genome.
JD: why use gbrowse?
SP: gbrowse is very easy to deploy and is less confusing compared with other browsers.

Wilhem Gruissem (WG) (ETH) - presentation
Presented the bioinformatics efforts and database developed at ETH
Genevestigator: a Gene expression database search engine. Cassava is not yet in this database because of lack of gene expression platform. They are ready to populate the database as soon as information is available from the cassava community. The appropriate threshold of putting data in database is at least 400 microarray experiments. The database comes with interesting tools to mine the database.

Protein database:
Started with *Arabidopsis*, and there are >16000 proteins available, and each has quantitative information that is linked with Genevestigator database.

His group has identified about 4000 cassava proteins and these proteins have been mapped to the genome.

They would like to develop a cassava proteomics data repository along the lines of the Arabidopsis database. ETH would like to hear from anyone working in this area for collaboration.

WG mentioned a European collaborative project on Arabidopsis leaf number 6 – to try to understand the developmental process, measuring everything from metabolites to epigenetic modifications, and the way in which this biochemical information was then linked to phenotypic information. All information was linked in a common database. He suggested that such a network could be established for cassava and urged people to work together.

-------------------END of the presentations---------------------------------

Open discussion session

1. What do the users require (diverse users, from researchers to breeders)

Simon Prochnik (BROWSERS):
A central browser platform will be the best option for genomic data integration and currently, there are two browsers (JGI and ETH).
GMcL: IBP is not involved in looking at sequence data.
SP suggested JGI can host the ETH's data as a way of collaboration. Alternatively, GFF files can be shared between JGI and ETH. Make sure that if one goes to genepage in Phytozome, there will be a link to ETH databases.

GMcL sought clarification on which users will use Gbrowse?
SP clarified that he does not think that what they do is going to apply directly to breeders.
WG pointed out that the breeders are also increasingly becoming interested in molecular information. Gave an example of how useful it is if one can link SNP to expression of gene.
JT: Suggested that the discussions should not be closed to just cassava as a plant but also pests and diseases (e.g. whiteflies).

WW: CAAS has a database with genome sequence and transcriptome sequence data.
WG: Suggested that the data be available under the Toronto Accord, which states that if one puts out data that have not been published, everyone can work with the data on condition that they contact the producer of the data, and the producer gets included in publications arising from their data.

MG: Pointed out that there are intermediate users who interface between genomics database and breeders.

PK: There is basically a need to know gene function, but a breeder needs to use the data quickly and easily without many layers of bioinformatics. The data has to be something that breeders can actually use.

JD: A major technical limitations is that it is not easy to see data from different accessions, and suggested that it would be useful to have functionality to, for instance, download SNPs between two accessions that one would like to use as parents of mapping populations.

EO: Asked how all the information available in the database would be useful to breeders. Also asked how if we could link to genomic info from other crops that may be useful in cassava?

IN: Asked about what efforts are in place to make people to use these databases.

SR: Gave a perspective on database use from a breeding company:
They have very distinct systems: (i) production breeding database (a LIMs system) providing a decision support tool for breeders; (ii) research database that is for the "random walk" scientist who is interested in exploring and making discoveries. Suggested that it is worth keeping those separate but with links rather than have everything in one database.
LM: Encouraged by the complementary nature of the databases discussed - 'Cassavabase' for marker, QTL data and genotyping data, IBP's 'Workbench' for germplasm, pedigree information, ontology, phenotypic data and breeders decision making tools, 'Genevestigator' for protein sequences, gene expression data. The challenge now is to interlink these databases, for example, Cassavabase may decide to not to have a genome browser but link to the JGI browser.

WG: Maintaining databases is expensive and the database that is not curated may quickly become obsolete. Suggested that a centralized information as much as possible to make it easy to finance the maintenance.

WW: Suggested integrate database to generate new information.

EP: Wondering how breeders have used the databases of Arabidopsis and rice that have been developed. Can learn from the users of the aforementioned databases?

SP: suggested that cassavabase.org could be the central portal solution; where one can go to other specialist database e.g. phytozome, proteomics database, or IBP to look for specific information.

RB: Suggested that there needs to be a volunteer taskforce made up of database expert, breeder, (functional) genomics to figure out how to integrate these tools.
SR: Suggested a "cassava breeders association"

HK: Breeders mostly deal with phenotypes and currently there is more genotypic information in the databases than phenotypes data. He is interested on how to make a link between genotype and phenotype.

GMcL: reiterated Steve's suggestion about the need to have a breeders' production database (routine breeding logistics, phenotypic data) and a research database (tools for genome, marker-trait associations). There is a gap in translating what is discovered in the research level and passing them on to breeders for use (phenotype to genotype is also key). Making the connections the two major classes of databases depend on (a) Trait ontologies and (b) Unique identification of germplasm.

EO: Asked SR for information on physical maps on cassava (e.g. where a QTL is located).
SR: Physical maps will become available in due course.

Communication from breeders is needed to guide the database programmers in order to develop a more intuitive and easy user interphase.

LM: said that having a large monolithic database is risk in terms of funding (gave TAIR, the Arabidopsis database as an example).

WG: funding – there has to be commitment from funding agencies that fund data generation and users (e.g. fee for accessing the data) in order to ensure long term sustainability of the databases.

JL: Uses TAIRs regularly. Emphasized the need for a champion to steer the integration and population of databases with existing data (specifically mentioned proteomics).

EO: GIS data information would be useful to integrate to the other databases in order to cross-reference with climatic data.

JY: Suggested user-user and user-developer communication tools; query history and Lukas answered that cassavabase has a forum for exchange of ideas, mailing list, comments.

LM: cassavabase will track queries using ‘Google analytics’ (gives stats on page visits etc).

SR: Internet connectivity should be considered when designing databases i.e. some the webpage should be designed with low-bandwidth connectivity in mind.

Working group volunteer: will include the main people from the different databases (at least one breeder and one functional genomics researcher). This will enable an understanding which is the needs. Lukas will initiate the working group, and we may build on ongoing groups such as cassava community of practice (COP).

Population of the databases (What are the issues and lessons from other crops? Are people willing to populate the databases, data curation).

WG: Anyone generating data should be willing to share the data. ETH can participate in curating and ensuring that the data is high quality if people are willing to contribute.

MG: Attaching contributor name is quite useful as it gives credit to the contributor, and gives accountability which helps to ensure quality.

EO: Noted that it would be helpful to have guidance on what acceptable quality of data is.

GMcL: Sees an opportunity for passive collection of data (via electronic field books) (even from wide array of environments) in an appropriately annotated way that is easy for breeders to use. There is a need to collect data from local database and populate central database, using simple tools. Graham also agreed that attribution of data to the data collector(s) is important from professional, courtesy and quality control point of views.

Getting data collected in the same way by different collectors needs to be emphasized and the use of IBP’s field book is a way of ensuring this.

Wrap-up:
It was noted that this was the first time that cassava users sat together to discuss the database issues.
A roadmap was defined by forming the working group chaired by Lukas.