**PeanutBase Report – December, 2017**

Contents

Advisory committee 1

Summary and Statement of Purpose 1

Reccomendations 2

Emerging Needs (Marker-Assisted Selection) 2

Literature Curation 3

Incorporation of Large Datasets 4

Meeting Breeder Needs 5

Funding 6

# **Advisory committee**

* Mary Schaeffer (schaefferm@missouri.edu) – MaizeGDB emeritus, USDA and University of Missouri
* Kelly Chamberlin (Kelly.Chamberlin@ARS.USDA.GOV) - USDA, Wheat, Peanut, and Other Field Crops Research, Stillwater, OK
* Josh Clevenger (jclev@uga.edu) - Mars Corporation
* Shyam Tallury (Shyam.Tallury@ARS.USDA.GOV) - USDA, Plant Genetic Resources Conservation Unit, Griffin, GA.
* David Bertioli (djbertioli@gmail.com) - University of Georgia
* Ray Schnell (Ray.Schnell@effem.com) - Mars Corporation
* Manish Pandey (m.pandey@cgiar.org) – ICRISAT, India

Ex officio:

* Andrew Farmer (adf@ncgr.org)- LIS and NCGR
* Bob Parker (bparker@nationalpeanutboard.org) - National Peanut Board
* Steve Brown (sbrown@peanutsusa.com) - Peanut Foundation

# Summary and Statement of Purpose

PeanutBase belongs to the peanut community. Its fundamental objective is to support the development of improved peanut varieties by providing modern science-based resources and increased understanding of peanut biology. PeanutBase was launched in summer of 2013 to provide public access to the data produced by the 5-year International Peanut Genome Initiative. The 6 components of the initiative are 1) genome assemblies of the 2 diploid ancestors of cultivated peanut, and cultivated peanut; 2) development of high-density map and marker sets; 3) gene expression atlas; 4) exploration of current genome sequencing and assembly methods and technology; 5) phenotyping of important collections (US mini-core); 6) bioinformatics and breeders toolbox (PeanutBase) to permit access to and use of data produced by the previous 4 objectives.

The 5-year plan is now drawing to a close and the genome assemblies, map and marker datasets, gene expression data and analyses, and phenotype data provides a strong basis for peanut breeding and molecular research. The purpose of this advisory group is to help plan the next phase of PeanutBase, including identifying and prioritizing key areas of breeder and researcher support, planning for varying levels of continued support and contingency planning if one or more collaborations is reduced or lost. To this aim, the committee has prepared a report to address the main questions posed for the direction of PeanutBase moving forward.

**Emerging Needs**

 As peanut continues its journey into molecular genomics, it is incumbent for the community to be able to use this new information efficiently. The ultimate goal is important for us, which is efficient genetic improvement of elite cultivars to meet the emerging needs of the global community. To that end, information is our most powerful tool. PeanutBase currently does an excellent job collecting, curating, and presenting data. The committee has the following recommendations for improvement in order of importance

**• Link marker information with germplasm**

The most important information a breeder needs for utilization of marker-assisted selection can be broken into three components: How to use the marker (technology), in what populations is the marker useful, and what is the percentage of broken linkage associated with the marker. It is the recommendation of the committee that marker information be linked to relevant germplasm in an obvious way. Further, markers should be placed in levels of information and confidence. For example, currently there are two traits with markers where there are functional markers or near functional markers. These are high oleic acid content and nematode resistance. These markers would be placed in tier 1 based on the amount of work that has shown they will select for the trait. On a marker page for the nematode marker, there can be links to or just lists of germplasm where if in the pedigree of a breeder’s population, they can use this marker with confidence to select for their trait of interest. For nematode resistance the germplasm would include NemaTam, COAN, Tifguard, TifNV hiOl, Georgia-14N, etc. The marker page should also include links to or summaries of relevant published work for the marker. Valuable information includes potential trait effect, primer information, marker type, etc. This creates a “one stop shop” for a breeder to gain all of the information they need to make decisions on utilizing marker-assisted selection for their desired trait. Another example is linked markers for rust and late leaf spot resistance for transferring resistance QTLs from donor ‘GPBD 4’ to three elite popular varieties of India namely ICGV 91114, TAG 24 and JL 24. These markers have been shown clearly to select for strong resistance. Level 2 tiers would include markers that have been developed and validated by one program, but not across programs. For example, in the USDA breeding program in conjunction with Peggy Ozias-Akins program, markers were identified and developed to select for leaf spot resistance in the background of GP-NC WS 16. The markers were validate in a blind selection study, but have not been used in other programs. Level 3 tiers would include published work on QTLs that have not been experimentally validated outside of the original study identifying them.

**• Core and Mini Core collection pages**

With the emergence of high throughput genotyping data technologies for peanut, there is suddenly a influx of marker data on key core collections. For example, the US mini core, comprised of 112 accessions has recently been assayed using the *Axiom Arachis2* SNP array and has available 31,210 polymorphic and curated markers. In addition, exome capture sequencing data generated through Peanut Foundation funding, was analyzed by validated pipelines to discover over 1.5 million polymorphisms. There is new interest in genotyping larger collections, such as the US core collection and even the entire USDA germplasm collection. Further, important collections from India and China have published trait and genotyping data. Currently, it is not easy to access lists of accessions in each of the collections or trait data for those accessions.

 PeanutBase can lead the utilization of these resources and facilitate the new data being generated. The committee recommends a core collection section which includes for each section the PI numbers associated with each accession, any genotype data available for each accession, and any phenotype data available. For genotype and phenotype data, bulk downloads should be available that allow data to be retrieved in table format for any collection or any group of accession IDs which can be submitted as a list. Currently this can be done on GRIN, but it is rather obtuse and overly complicated. It can be greatly streamlined for ease of access.

 **• Standardization of phenotypic data**

 As peanut moves closer to utilizing genomic selection, the committee sees PeanutBase as a “pseudo seed company” in respect to compiling data on many traits across many environments on common genotypes that can be leveraged for genomic selection. This is a very large task of course, but as a first step, the peanut community needs to work towards standardizations for measuring certain phenotypes or implementation of ontologies so that phenotypes can be grouped and used as best linear unbiased predictions (BLUP) which can be used to normalize different data types. PeanutBase may also provide updated relevant information of most efficient genomic selection models and may plan to further customize these models for achieving better results for peanut research community.

**• Links to other international initiates related to genetic gains**

 Modernization of breeding programs is key for succeeding in coming days and several of the programs found this mandatory for achieving higher efficiency in their programs. Several international initiates are running and much of the information and tools developed by these initiatives may be very much application to peanut community with no or little customization. For example, Integrated Breeding Platform (IBP, <http://www.integratedbreeding.net>), Global Open Breeding Informatics Initiative (GOBII) (http://cbsugobii05.tc.cornell.edu/wordpress/index.php/about/) and Excellence in Breeding (EIB) Platform (http://excellenceinbreeding.org/).

**Literature Curation**

 PeanutBase can play a role in facilitating the translation of published work into the most informative and useful information for the peanut research community. For example, a QTL study may identify a major QTL, develop a tightly linked marker for the trait, and validate that marker as being useful for marker-assisted selection. The most important information for breeders and other researchers is where is the marker physically located in the genome, what kind of marker is it, what are the primer sequences for the marker, what is an estimate for linkage with the trait, what is the relevant genetic background, and what is the estimate trait effect for the marker (if qualitative trait). This information, in published manuscript form, is not always readily available in its simplest form. These studies can be disseminated and published in a clear form on PeanutBase to facilitate the use of the information.

 Because of considerations with the usage of time and effort, there is the question on what literature to focus on. It is the recommendation of the committee that literature curation focus top priority on studies that can provide new tools for marker-assisted selection. To prioritize literature that falls within this category, work can be grouped into levels in a similar fashion as described in the Emerging needs section.

• Highest priority should be given to literature representing tools for marker-assisted selection where any developed markers have been shown to be effective across experiments, programs, and publications. These markers represent high confidence for effectiveness across environments. In addition, any literature representing new tools for molecular breeding (bioinformatics, DNA extraction methods, marker protocols, etc) should be given highest priority for dissemination.

• Second priority should be given to literature representing tools for marker-assisted selection where any developed markers were validated in more than one experiment within a program.

• Third priority should be given to literature representing tools for marker-assisted selection where any markers are not validated outside of the experiment that identified them. For example, any QTLs identified from mapping experiments without additional validation.

**Incorporation of Large Datasets**

 The more information is available to solve a problem, the more efficiently the problem can be solved. In the context of genomics-assisted breeding, the more environments that alleles and genotypes are tested in, the greater the prediction accuracy for the true allele affects. To this end, large datasets generated by different programs across the world would need to be incorporated in one place and formatted in a consistent manner. The committee has the following recommendations for how to proceed locating and incorporating large datasets.

**• Establish guidelines for quality control/selection of highest quality/most informative datasets**

There is a danger in being too open in that poor data may drown out good data, making use of that data inefficient and leading to decreased usage by the community. However, it is also not the place of PeanutBase or a small group of individuals to judge the merits of every program’s scientific output. There is still a middle ground that will ensure the best possible incorporation of informative datasets that will enhance peanut research moving forward. The committee recommends the following guidelines:

 **• Establish a tiered priority-based system for data sets**

Priority can be given to datasets that possess certain traits. A tiered system can be put in place, with levels of datasets based on specific attributes.

 **• Tier 1 datasets**

Highest priority datasets will be comprised of published, peer reviewed work in international journals with complete description of materials and methods.

 **• Tier 2 datasets**

Datasets from published, peer reviewed work in international journals with incomplete descriptions of materials and methods where there is missing information will be prioritized after tier 1 datasets. Authors of the work can be contacted for missing information, but as this will take more time, these datasets can be prioritized second.

**• Tier 3 datasets**

The third priority datasets will be comprised of high quality datasets with appropriate controls and replications that for some reason will not be published and are informative for others. Program that generated the data must provide complete material and methods for the data.

**Meeting Breeder Needs**

The following survey was sent to breeders in the U.S. as well as abroad. Unfortunately, only four responses were received.

**PeanutBase Breeder Survey**

**1. Do you use PeanutBase in your program? If so, what features of the database do you find most helpful? If no, please elaborate on features that could be changed or added to make PeanutBase useful to you.**

**2. What can the database administrator(s) do to improve your experience when using PeanutBase?**

**3. Would you be interested in participating in a meeting workshop and/or online training module to learn how to use PeanutBase?**

Of the four that responded, two breeders use PeanutBase as a tool to search for sequences in the genome, look for markers associated with specific traits, and/or place markers on a specific chromosome or existing map. One other breeder indicated that PeanutBase was used extensively with their partner who has a molecular lab upon whom they reply for all molecular work associated with the breeding program. The remaining breeder runs a classical breeding program. All breeders that replied were favorable towards and would attend workshops held aimed at improving their familiarity with the data base and/or simply their use. All agreed that linking known markers for specific traits to germplasm known to possess those traits and sequences would be helpful to breeders searching for breeding material, especially in the case of disease resistance (this could be added to the MAS sections).

Of the breeders who responded that they personally use PeanutBase, one has a background in molecular genomics and leads a “hybrid” program that produces cultivars and does basic research as well. The other leads a breeding program that uses marker-assisted selection. This breeder mainly uses PeanutBase for locating primers for SSR markers that have been identified as selecting for a trait of interest.

The committee recommends that hands-on training sessions at workshops or online webinars would help realize the full potential of PeanutBase. Questions and feedback received during such sessions could serve to inform usability analyses and to help focus efforts on content development such as interactive tours. Another low-cost model for improving use was a mailing list where users can post questions and suggest ideas for improvements.

We believe that engagement and tiered, breeder-focused efficient information retrieval provide a basis for wider use of PeanutBase, especially when the number markers linked to traits increases.

**Funding**

 The leadership of Peanutbase outlined four options for the future based on funding possibilities; (1) Static, (2) Minimal, (3) Current level of development, and (4) current development plus expansion to meet emerging needs. For the peanut community to keep up the momentum established over the last 6 years, the first two options should not be considered. The committee recommends that the best option is the fourth option, which will require an estimated support of 2 ¼ FTE in funding. As is evident in this report’s coverage of the emerging needs, the community needs the expertise of PeanutBase for the utilization of these data and resources. However, the current state of peanut funding is undergoing a transition, in which certain obligations of investment from industry partners has been met and currently not renewed. In this new landscape, new funding opportunities need to be explored to meet funding needs. The committee recommends the following options for funding PeanutBase going in to the future.

**• Principle Investigators budget PeanutBase support in grant applications**

 The committee feels the best option for future funding is for the community to begin writing in PeanutBase support into their grant applications. Dissemination of data is a big part of the grant application process and deposition into public databases, although free, is time consuming and complicated. Further, utilization of data is higher when it is easily accessible. The committee realizes that this process will be difficult to sell to already limited budgets and may result in a few programs providing the majority of funds. The committee recommends the following process for implementation.

 **• Provide a framework for costs associated with different projects**

If researchers are provided a “menu” in which to write in support in their budgets, they will be more likely to participate. Further, it will ensure that every researcher pays the same amount for the same scale of work. For example, the leadership at PeanutBase can estimate how much support would be needed to deposit and curate a large phenotypic dataset, or genotype data, or sequencing data. This information could be made easily available on the website so that as PIs write grants they can easily find this information.

 **• Only accept data from projects that provided PeanutBase support**

For the life of the Peanut Genome Project, PeanutBase has been supported by true community support from industry investment. While it is important to maintain an open community, the reality is that new avenues of funding need to be found. Requiring including PeanutBase support into project budgets for deposition of data has the potential to quickly transition from top-down support to bottom-up support.

 **• Actively promote cost framework and highlight programs that contribute support**

Awareness in the community is the best way to change the norm. Actively promoting support frameworks and highlighting programs that provide support can lead to others wanting to “join the club”. Highlighting programs can include a featured lab page as is done at the Sol Genomics Network (<https://solgenomics.net/>). When datasets are used for demonstration in workshops, help tutorials, and presentations, highlight the program that contributed the data. Include in all presentations and workshops a slide or two explaining the need to for including PeanutBase support in grant budgets and show examples of budgeting frameworks.