Maize Informatics Research Coordination Network Workshop Outcomes

Twenty-eight researchers from U.S. academic institutions, USDA-ARS, related industries, and federal funding agencies met for a workshop in Madison, WI, on September 26-27, 2019. The goal of this workshop was to identify and articulate informatics issues relevant to the maize community and develop a vision to prioritize addressing these issues. There were two primary focus areas of this workshop. The first area was 'annotation and comparison of many genomes', and the second was 'collection, curation, and availability of phenomics data'. This white paper provides a summary of discussions of this two day workshop including the current community status, future challenges, and recommendations of needs that can be addressed through strategic funding decisions.

Workshop Attendees

Name	Affiliation	Email Address	Attendance
Carson Andorf	USDA-ARS	carson.andorf@usda.gov	In person
Ed Buckler	USDA-ARS/Cornell University	esb33@cornell.edu	In person
Ben Busby	National Center for Biotechnology Information	ben.busby@gmail.com	In person
Kapeel Chougule	Cold Spring Harbor Laboratory	kchougul@cshl.edu	In person
Jennifer Clarke	University of Nebraska	jclarke3@unl.edu	In person
Natalia de Leon	University of Wisconsin	ndeleongatti@wisc.edu	In person
Kevin Fengler	Corteva Agriscience	kevin.a.fengler@corteva.com	In person
Candice Hirsch	University of Minnesota	cnhirsch@umn.edu	In person
Matthew Hufford	Iowa State University	mhufford@iastate.edu	In person
David Jackson	Cold Spring Harbor Laboratory	jacksond@cshl.edu	In person
Shawn Kaeppler	University of Wisconsin-Madison	smkaeppl@wisc.edu	In person
Carolyn Lawrence-Dill	Iowa State University	triffid@iastate.edu	In person
Paula McSteen	University of Missouri	mcsteenp@missouri.edu	In person
Nirav Merchant	University of Arizona	nirav@email.arizona.edu	In person
Jack Okamuro	USDA-ARS	Jack.Okamuro@ars.usda.gov	In person
Jesse Poland	Kansas State University	jpoland@ksu.edu	In person
John Portwood	USDA-ARS	john.portwood@usda.gov	In person
Michael Schatz	John Hopkins University	mschatz@cs.jhu.edu	In person
James Schnable	University of Nebraska	schnable@unl.edu	In person
Nathan Springer	University of Minnesota	springer@umn.edu	In person
Ruth Wagner	Bayer Crop Science	ruth.wagner@bayer.com	In person
Doreen Ware	USDA-ARS/Cold Spring Harbor Laboratory	ware@cshl.edu	In person
Margaret Woodhouse	USDA-ARS	margaret.woodhouse@usda.gov	In person
R. Kelly Dawe	NSF	rdawe@nsf.gov	Virtual
Diane Okamuro	NSF	dokamuro@nsf.gov	Virtual
Gerald Schoenknecht	NSF	gschoenk@nsf.gov	Virtual
Anne Sylvester	NSF	asylvest@nsf.gov	Virtual
Clifford Weil	NSF	cweil@nsf.gov	Virtual

Workshop Speakers – Working Across Multiple Reference Assemblies:

Ben Busby	“Lessons learned from other communities working with multiple reference assemblies”
Michael Schatz	“Lessons learned from other communities working with multiple reference assemblies”
Kevin Fengler	“Tools and lessons learned from an industry perspective”
Nirav Merchant	“CyVerse infrastructure and interaction with other communities”

Breakout Group Discussion Questions – Working Across Multiple Reference Assemblies:

1) What are the most important tools to have for the future to work across multiple reference genome assemblies?
2) What mechanism(s) do we want to use to allow for community input on quality of gene models and to provide value added information about the models?
3) What is the process to allow for iterative improvement of gene model annotation? How does this process change in the context of many reference genome assemblies?
4) What are the needs of different members of the community and how do we balance providing the resources wanted or needed from different community members?

Workshop Speakers – Phenomics:

Jennifer Clarke	“Lessons learned from the North American Plant Phenotyping Network (NAPPN)”
Jesse Poland	“Lessons learned from the wheat phenomics community”

Breakout Group Discussion Questions – Phenomics:

1) What are the current biggest needs to bring the community together and make things more interoperable?
2) How do we implement these “biggest needs”?

Annotation and Comparison of Many Genomes

For many years it was believed that having access to a high-quality reference genome assembly would unlock the answers to countless unanswered biological questions. In 2009, the first maize reference genome assembly of inbred line B73 was published[1] and several improved assemblies have been subsequently released[2]. While having access to this iteratively improved reference genome assembly for B73 facilitated significant biological findings, there are also limitations to the biological scope of questions that can be addressed in the context of a single reference genome assembly for the species. Over the last five years more than 10 additional maize inbred lines have been assembled at various qualities and are publicly available (https://maizegdb.org/genome). Comparative analyses of these genome assemblies has shed light on the extensive variation that exists in haplotypes of shared regions of these genomes, as well as an understanding of the extent to which there are genomic regions that are present in only one or a subset of individuals in the species (i.e., dispensable genome sequence). The comparison of these genomes has also begun to reveal the complexities of comparing functional features among maize genomes that vary substantially in their sequence content.

It is anticipated that in the next five years there will be well over 100 reference quality genome assemblies for maize that will cover a large portion of the maize pan-genome sequence space. As genome assembly becomes more straightforward, our community limitation will shift from not knowing the sequence of an individual or that of the collective maize pan-genome, to a limited understanding of the functional elements within these genomes, and a limit in our ability to compare functional features among these genomes. Annotating a genome is a difficult computational challenge as we are often using incomplete or imperfect evidence, and working in a system in which “rules” of function are rarely absolute. This exacerbates the challenge of automating the processes of accurate annotation of features of a genome.

Access to a large number of reference genome assemblies is enabling and exciting for the community. However, it also magnifies existing challenges when generating a representation of a genome (assembly) and subsequently seeking to define the locations and functions of genes through the generation of gene models, definition of regulatory regions, and annotation of other relevant features of the genome. Finally, new challenges arise that do not exist within the context of a single reference assembly such as how to link features of the genome across assemblies. These opportunities and challenges were discussed, and a number of recommendations were put forward based on the points raised by the external speakers that presented at the workshop and by the workshop breakout discussion participants. These points included the following needs:

1. There is a need to determine who, what, when, where, why, and how gene model annotations will be initially generated and how subsequent versions will be handled. As sequencing technologies have advanced to generate longer sequence reads and algorithms have been developed to support assembly of these reads, the ability to create a reference quality genome assembly in version 1 is now possible. Consensus accuracy still remains challenging, and this can potentially impact subsequent annotation[3].

[1] Schnable et al., 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326:1112-1115.

[2] Jiao Y, et al., 2017. Improved maize reference genome with single molecule technologies. Nature. 546(7659):524-527.

[3] Watson and Ware, 2019. Errors in long-read assemblies can critically affect protein prediction. Nature Biotechnology. 37:124-126.

Still, genome assemblies that are generated in the future will likely be relatively static after the first version is released. In contrast, annotated features in the assemblies will require many iterations of improvement as we increase our knowledge of the structure and function of features of a genome through experimental evidence and data science approaches. This is further complicated by the fact that the function of a sequence may change in different cell types or developmental stages. The community holds a great deal of knowledge about which gene models are correct and which need to have modifications. A process and tools need to be developed such that this information is documented and iteratively used to improve gene model annotation. This will require determining how data is collected with defined community best practice standards, what information is to be collected (e.g., new whole-genome data sets such as CAGE data, ATAC-seq, information on specific genes that have been cloned, etc.), when will this iteration happen (e.g., annually, bi-annually, etc.), and how will nomenclature issues be resolved (e.g., how different does a gene model need to be before it gets a new name, etc.). It will also be important to consider how genomes generated by different research groups will be commonly annotated, and how annotation tools can be shared to ensure similar annotation quality to allow for comparisons of genome features. Another important question that will need to be addressed is how we will determine which genome assemblies are given a portion of the finite resources available for this activity of iterative improvement, and which will rely on porting information over from those assemblies that have resources to improve annotation.

2. There is a need to develop tools to link features across genome assemblies. Assuming that every functional feature of a given genome assembly can be accurately annotated, the next challenge is how to link these features across the assemblies. This linkage is required to address questions such as, I have my region of interest in genetic background X, is there genetic variation in genetic background Y and Z? Is this functional variation? This is particularly useful for those working on cloning and characterizing genes that may be part of larger structural variations between genomes. Ultimately, these links across genomes will likely be represented in a graph-based (or graph-related) format. However, due to the complexity of variation in maize and the current status of graph-based methods to represent variation, this is not possible at this time in maize. Engaging with the broader computer science community that has research interests in graphs early in the implementation of graph based methods will be critical. In the interim, there are other non-graph based approaches that are in development (e.g. gene beads) that can be used to link features across assemblies. It will also be important to generate useful visualization methods to show this variation in a comprehensible format such as gene trees, offset graphs, variation graphs, etc.

3. *There is a need to identify and support the different user groups that utilize maize genome assembly and annotation resources.* The individuals that utilize maize genome resources include maize focused and non-maize focused researchers, individuals with genomics/bioinformatics expertise, individuals with limited genomics expertise, as well as educators. The tools that are needed and mode of accessing these data are likely quite different across these user groups. For genomics researchers, a highly organized central place that hosts all data as flat files (i.e., FTP site, Cyverse, and eventually API-level access) that can be downloaded and parsed as needed is likely the ideal point of access. Detailed descriptions and metadata will be required for each available data set. For other members of the community, visual tools (i.e., genome browsers, gene tree browsers) and graphical user interfaces (GUIs) that facilitate access to the data will be necessary. It is important to recognize that maize is a model system with extensive resources that can be leveraged by other communities and it is important to create a system that is intuitive to those outside of the maize community.

These points of discussion resulted in some more immediate and some longer term goals. The more immediate goals are to continue to provide useful gene-centric resources through MaizeGDB (https://www.maizegdb.org) with improved accessibility to other communities, and to create a curated data repository with metadata to use for iterative annotations and cross-genome linking of features. The more long-term goal is to enable graph-based genome representations and comparisons to maximize the utility of having many reference quality genome assemblies within the species.

Collection, Curation, and Availability of Phenomics Data

One of the biggest challenges of the 2000's was to sequence and assemble a reference genome for maize. A major challenge of the current decade is to determine the function of all of the bases that exist in the pan-genome of a species, and how those functions change in response to environmental conditions. The ability to assign functions to elements of the genome (e.g., genes, transposable elements, regulatory regions, etc.) will require extensive phenotypic evaluation across scales from individual cells to whole plant and field based measurements of plants growing in different environments. It will also require expertise across disciplines from biologists to computer scientists and engineers.

Community resources that allow data to be leveraged across experiments and research groups will facilitate these efforts. The ability to make connections across experiments that utilize similar genotypes or environments will require data to be collected and curated in a standard way with accompanying metadata. Phenotyping is expensive and often labor intensive, and the generation of this data comes at a substantial cost. It is also important to recognize that we cannot replicate an environment after it happens, and thus the data that is collected can never be regenerated in future experiments. This places a much higher need on resources to store not only the phenotypic data itself, but also the metadata that describes when, where, and how the plants were grown and the data was collected. Development of data resources that are easily usable and well-curated will be critical to attract creative scientists from other disciplines to join our community to tackle fundamental problems in plant phenotyping.

The Genomes by Environment (GxE) Project within the Genomes To Fields Initiative (https://www.genomes2fields.org) is a useful case example from which to see how data collection, curation, and release can be done across groups and used to leverage larger biological findings than would be possible by any individual group[4]. The GxE Project has brought together 51 investigators from 21 institutions to collect the same 14 phenotypic traits from approximately 180,000 plots grown across 162 unique environments over the past six years. This process required standard operating procedures (SOPs) to be developed for phenotypic trait collection and standardization of required metadata to be collected, including GPS coordinates of field locations, management practices, soil composition, weather data, etc. There have been many successes in this project with regards to data standardization that have lead to the

[4] Gage, J., et al. 2017. The effect of artificial selection on phenotypic plasticity in maize. Nature Communications. 8(1):1348.

release of valuable data sets that contain consistently collected and curated data[5,6]. However, this project also exposed issues that will be critical to address in future large scale phenotyping efforts. These include how to share image data and work across scales as data collection moves into high throughput image based phenotyping, and the need for a database to store data from various imaging modalities with geospatial capabilities.

A number of recommendations were put forward that build from the experiences of the Genomes To Fields Project, points raised by external speakers that presented at the workshop, and the workshop breakout discussions. These included the following:

1. ***There is a need for an NCBI-like repository dedicated to housing phenotypic data and associated metadata with geospatial capabilities.*** Having this infrastructure in place will allow for efforts to create indexed and curated data sets that are useful in conducting meta-analyses as well as in providing data sets to developers interested in creating new innovative analysis methods and tools to be shared with the community. As with genomic data types, when a database exists it is then possible for funding agencies and peer-review journals to mandate data standards and data deposition. The current scale for such a database is in the terabyte range and it is expected in the next five years this will be in the petabyte range.

2. ***There is a need for standardization of data collection through standard operating procedures.*** Different research groups often utilize different methods for collecting the same traits (e.g. plant height from the base of the plant to the ligule of the flag leaf or to the base or tip of the tassel) and different traits that can be obtained from the same image type (e.g., meristem size and shape from scanning electron microscope images). This issue spans across scales from low throughput histological images to high throughput drone images. Development of community-approved standard operating procedures will expand the utility of current data collection efforts and maximize the biological knowledge that can be gained from the finite amount of resources that are available for these efforts.

3. ***There is a need for the development of analysis tools to extract trait data from images and conduct meta-analyses across the scales at which data is collected.*** The skills required to process image data at the scale that is available using UAVs, remote sensing technologies, and other high throughput data acquisition platforms are not necessarily in the repertoire of biological domain experts. Development of analysis tool kits for commonly performed image extraction methods would not only democratize the use of these technologies, but will also incentivize conforming to standards for data collection that are required to use these analysis workflows. This will likely require large exemplar training data sets (i.e., labeled images that can be used to train machine learning algorithms) to be used for development of analysis methodologies.

[5] AlKhalifah, N., et al. 2018. Maize Genomes to Fields: 2014 and 2015 Field Season Genotype, Phenotype, Environment, and Inbred Ear Image Datasets. BMC Research Notes. 11:452.

[6] McFarland, B., et al. 2019. Maize Genomes to Fields (G2F): 2014-2017 field season's genotype, phenotype, climatic, soil, and inbred ear images datasets. (*under review at BMC Research Notes*)