MEDICAGO TRUNCATULA GENOME INFORMATICS WORKSHOP EXECUTIVE SUMMARY

Medicago truncatula is a reference legume whose euchromatic gene-space is now being sequenced by an international consortium. In June 2005, the National Science Foundation (supplemented with a gift from Monsanto Company) sponsored a workshop in Asilomar, California to develop a "roadmap" for *Medicago* annotation and bioinformatics. A group of 36 international experts in plant genomics participated in the workshop, addressing the following questions:

- What does the community expect from the *Medicago* sequencing effort?
- Where does *Medicago* genome annotation stand today and what should come next?
- What *Medicago*-specific bioinformatic resources are needed in the context of broader informatics efforts already underway?

A clear message from the workshop was that *Medicago* is widely viewed as a key reference species, and therefore needs a comprehensive, high quality genome sequence. However, recent results indicate the gene-space of *Medicago* is approximately 25% larger than originally estimated. In the short-term, a genome sequence with physical contiguity and sequence gaps composed of "ordered and oriented" BACs will be useful. In the long run, a "gold standard" sequence of the entire euchromatin should remain the goal for *Medicago.*

In the area of annotation, the *Medicago* community is fortunate to have the "IMGAG Initiative," which already provides useful and reliable automated annotation accepted community-wide. Once the euchromatic sequence is complete, substantial genome curation will be needed. In the process, full-length cDNA sequencing will enhance the quality of the *Medicago* genome annotation. Moreover, the *Medicago* informatics community should play a larger role in the "Gene Ontology" and "Plant Ontology" consortia, as well as the Biological Database Curators Consortium.

Current *Medicago* genome sequence databases were designed primarily for the *Medicago* sequencing effort, but the broader community uses them as well. These database resources should be modified to simplify access to key datasets and interconnections with other species. In the medium term, many in the user community want a simple portal to explore the entire range of *Medicago* genome sequence and genomics data. Nonetheless, *Medicago* informatics needs to be embedded in cross-legume informatics, including the Legume Information System. In the realm of functional genomics, data types are highly diverse. To ensure that web-accessible data are truly comparable, the community needs to coordinate experimental conditions, methodologies, and data formats. Eventually, functional genomics should provide a basis for annotating biological functions to genes. Participants recognized that funding for *Medicago* informatics will be required to make these goals a reality,

MEDICAGO TRUNCATULA GENOME INFORMATICS WORKSHOP
ASILOMAR, CALIFORNIA
JUNE 9 AND 10, 2005

CONTEXT FOR THE WORKSHOP

The sequencing of *Medicago*'s gene-rich euchromatin is an important milestone in plant genomics. By the end of 2006, most of the *Medicago* euchromatic gene-space will be sequenced and researchers worldwide will be using the sequence to clone genes, predict protein function, dissect biological networks, and unravel genome evolution. To achieve its potential, the research community needs to develop a "roadmap" for *Medicago* annotation and informatics analysis. With this in mind, the National Science Foundation through award DBI 03-21460 sponsored a workshop on *Medicago* genomics and informatics at Asilomar, California on June 9 and 10, 2005. Funding for the workshop was supplemented by a generous gift from Monsanto Company.

A group of 36 international experts in plant genomics participated in the workshop, which began with overview presentations by Dr. Robin Buell of The Institute for Genomic Research (TIGR) and Dr. Sue Rhee of Carnegie Institute of Washington. Dr. Buell is a leader in the rice genome sequencing and annotation effort, while Dr. Rhee leads The Arabidopsis Information Resource (TAIR). Their talks reminded participants about the importance of using genome sequencing, annotation, and database development to build a stronger *Medicago* community. Participants were told they should identify the informatics needs of the U.S. and international *Medicago* research communities and work to extend the *Medicago* sequence to other species, especially crops. The speakers reminded participants that even as they focus on the needs of *Medicago* informatics, they should seek out alliances, such as the Gene Ontology and Biocurators Consortia. Finally, Dr. Buell and Dr. Rhee highlighted the many important decisions the *Medicago* genomics community will need to make – such as deciding when to produce new genome sequence releases, how to agree on stable identifiers, and how to share improvements and minimize duplication of effort.

The presentations were followed by a series of discussions that sought to frame the future of *Medicago* sequencing and informatics. The first discussion focused on the question: What does the community expect from the *Medicago* sequencing effort? The second discussion addressed the question: Where does *Medicago* genome annotation stand today and what should come next? Finally, participants envisioned the *Medicago*-specific bioinformatic resources that are needed in the context of broader informatics efforts already underway in other legumes and plant communities.

QUESTION 1. Should the community complete the genome sequence of *Medicago*, and if so, how?"

***Medicago* is widely viewed as a key reference with broad support for a comprehensive, high quality genome sequence.** Participants confirmed that a growing community of scientists relies on the *Medicago* genome sequence as a critical component in their research. Many scientists are planning future experiments with the expectation of a finished, high quality sequence for the entire *Medicago* gene-space. For example, the poplar sequencing community has already utilized the *Medicago* sequence to order and orient its whole genome shotgun assembly. In agriculturally important legumes, especially soybean, a comprehensive, high quality *Medicago* sequence will provide an essential foundation for successful genome assembly.

The gene-rich euchromatin of *Medicago* is larger than estimated at the start of the sequencing project. It is increasingly clear that the gene-rich portion of the *Medicago* genome is larger than originally estimated, though still quite modest in total size. Based on the proportion of all ESTs found within currently sequenced BACs, it appears that the *Medicago* gene-space lies between 250 and 300 Mbp. This is far less than the size of the entire genome (500-550 Mbp, depending on genotype), but still larger than the original 200 Mbp estimate for *Medicago*'s euchromatic gene-space.

In the short-term, a genome sequence with physical contiguity and sequence gaps composed of "ordered and oriented" BACs will be very useful to the research community. In an effort to balance the reality of a larger sequencing target with the community's need for as much high quality genome sequence as soon as possible — workshop participants considered strategies to optimize the efforts of the sequencing consortium over the next two years. A consensus developed around "physical contiguity with sequence gaps." Ideally, this would consist of BACs "finished" to phase 3, though in practice some BACs would be finished only to phase "2.55" in order to increase throughput and decrease cost per BAC. Phase 2.55 was defined as ordered and oriented BAC clones (phase 2) with at least 5 kbp of uninterrupted sequence at both ends and consisting of no more than five fragments. BAC sequences should be joined into pseudomolecules that run all the way from FISH-characterized telomeric markers to FISH-based markers defining boundaries with pericentromeric regions, characterized by sequence data indicating low gene densities and high repeat densities. Gaps should consist of scaffolds held together by paired BAC-ends, FPC contigs, or FISH-based estimates of physical gap sizes.

A "gold standard" sequence of the entire gene-rich euchromatin should remain the goal for *Medicago.* Participants agreed that in the short-run a *Medicago* genome sequence that reaches the benchmark of physical contiguity with some sequencing gaps would be extremely useful to the research community. Moreover, this goal is clearly achievable within a reasonable period

of time. Still, it misses the mark of a comprehensive, high quality sequence for the entire gene-rich euchromatin of *Medicago*. Participants reiterated their strong support for the additional effort required to reach the goal of a gold standard genome sequence as a foundation for positional cloning of biologically interesting genes, efficient sequencing in other legumes, and the broader goal of providing a third high quality plant genome sequence for comparative and evolutionary genomics.

QUESTION 2. What should be the future of *Medicago* genome annotation?

The IMGAG initiative is providing useful, reliable automated annotation that is accepted community-wide. Workshop participants were impressed by the success of the International *Medicago* Genome Annotation Group (IMGAG). This group is taking the lead in automated gene-calling, even while genome sequencing is still underway. IMGAG has developed community-wide standards and nomenclature, a definition for a reference whole genome data set, data exchange protocols, and corresponding pipelines and procedures. However, there are still no mechanisms (or funding) for expert annotation or curation and only limited efforts in the areas of ontology assignment or characterization of non-coding genome sequence.

Significant genome curation should be carried out once the sequence is complete. Participants agreed that once the *Medicago* euchromatin sequence is complete, substantial expert annotation and curation would be needed. The *Arabidopsis* experience (gene-by-gene annotation during the course of sequencing) demonstrated the value of waiting until sequence assemblies are stable and global views of gene families are available before intensive annotation/curation. Once stable genome builds for *Medicago* are available, a small number of informaticians need to take the lead in order to ensure standardized and consistent criteria. These informaticians need to work with communities of experts in specific gene families and functions, including experimentally-based annotation. Potentially, future annotation of *Medicago* could be performed in coordination with that of *Lotus* and other legume genome sequences. Eventually, *Medicago* annotation should include standardized annotation about non-coding regions. Identifying the resources to implement these important goals, while recognizing that *Medicago* informatics is unlikely to receive funding at the scale of earlier plant models like *Arabidopsis*, remains a major challenge to the *Medicago* community. Ideally, additional investment in *Medicago* annotation can take place in the same time frame as the genome sequenced is being completed.

An annotation jamboree can help to build community, but has limited value for large-scale, expert genome annotation. Workshop participants were told that jamborees are relatively inefficient in providing large-scale expert annotation for eukaryotic genomes, though they do have real value in building a sense of

"community." Significant lag-time is needed to bring participants up to speed and quality varies substantially among different groups at the same jamboree session. Nevertheless, members of the broader *Medicago* and legume communities would genuinely benefit from greater familiarity with the annotation process, especially if non-informaticists are expected to participate meaningfully in future distributed annotation activities.

Full-length cDNA sequencing will enhance the quality of *Medicago* genome annotation significantly. Beyond the genome sequence itself, there was widespread and substantial support for the sequencing of full-length *Medicago* cDNAs. High quality gene annotation critically depends on this type of sequence data, which is inadequately addressed with partial EST coverage. Participants noted that full-length cDNAs are the most informative and cost-effective resource for improving gene annotation, and should therefore be pursued as soon as possible.

The *Medicago* community should begin to play a role in the "Gene Ontology" and "Plant Ontology" consortia, as well as the Biological Database Curators Consortium and related initiatives. Because *Medicago* will be among the first generation of high quality plant genome sequences, participation in the Gene, Plant, and Trait Ontology efforts is important. A representative of the legume community already participates in the Trait Ontology effort. The *Medicago* community needs to decide soon on the best way(s) to contribute to the ontology effort, especially with regard to legume-specific attributes, while at the same time optimizing the utility of ontologies in the annotation of *Medicago* itself. Ideally, the International *Medicago truncatula* Steering Committee should identify a representative soon. The *Medicago* community will also benefit from participation in the Biological Database Curators Consortium once a long-term plan for genome curation is established.

QUESTION 3. Given multiple cross-legume resources — What *Medicago*-specific informatics resources are needed for the *Medicago* sequencing project? For the *Medicago* research community? For the legume community? For the broader plant research community?

Comparative Genomics Focus

Current resources to use the *Medicago* genome in comparative studies with other species seem scattered and difficult to navigate. Participants expressed frustration at the dispersed nature of current *Medicago* informatics resources and the difficulty in translating insights from the *Medicago* genome to other legume species. Participants were reminded that it is impractical to create comprehensive informatics tools for a still incomplete genome sequence. Connections that do exist tend to be sequence-based, so comparisons between *Medicago* and *Lotus* are often insightful, while sequence-based comparisons

between *Medicago* and soybean should improve as more soybean sequence becomes available. By contrast, comparisons to most other legume species will probably remain limited for some time to come.

Current *Medicago* genome sequence databases were designed primarily for *Medicago*, especially the sequencing effort. They should be modified to simplify community access to key datasets and connections with other species. Current *Medicago* genome databases (such as *Medicago*.org at Minnesota and comparable sites at TIGR and Oklahoma) were established primarily to serve the sequencing effort. These databases focus on *Medicago* with only limited connections to other species. Still, these databases could do a better job providing access to all sequenced BACs, snapshots of current sequence builds, and links to other databases. The existing *Medicago* databases should also work together to create an index that provides up-to-date information about existing *Medicago* informatics resources, including their features and interconnections. In the process, precise mechanisms and criteria by which specific datasets have been curated should be provided. Finally, there needs to be better interconnectivity among existing *Medicago* databases, as well as with the Legume Information System (LIS) with its distinctly comparative genomic focus.

Many in the community would like a single entry-point (portal) that enables efficient exploration of the entire range of *Medicago* genome sequence and related data. From the point of view of many *Medicago* researchers, a single *Medicago* website that efficiently integrates with other informatics sites is highly desirable. Even with the existence of cross-species sites like LIS, users of *Medicago* genome informatics generally wish to access *Medicago* information through a well-defined “portal” that logically moves among other databases. Participants frequently mentioned TAIR as a model for the type of *Medicago* portal they would like to see, though the level of funding support for *Medicago* is likely to be lower. Participants also recognized that it is crucial to establish a balance between *Medicago*-specific informatics, best addressed by *Medicago* resources, and the many other informatics questions better addressed by cross-species sites like LIS. Right from the start, *Medicago* and other legume databases should establish plans for interoperability. And while some redundancy among *Medicago* and other legume resources may be desirable, unproductive duplication must be minimized.

Informatics groups should make greater effort to inform the community about existing *Medicago* genome data resources. Participants noted that the broader research community seems to have limited awareness about the range of informatics sites and resources that already exist for *Medicago*. Current developers of *Medicago* websites and databases should do a better job communicating their resources to the *Medicago* and broader communities, potentially including a listserver and newsletter.

The community needs one or more *Medicago* stock centers that include a web-accessible database component. As *Medicago* becomes a more widely used system for biological research, a small number of stock centers need to be developed. One such center is already in development in France, though additional facilities will probably be required. Essential for the success of *Medicago* stock centers will be a web-interface in which different types of stocks can be maintained at different locations, but accessed through a common portal.

A task force to consider a portal for *Medicago* informatics should be established. While it is still some time before the euchromatic sequence *Medicago* reaches completion, a task force to consider the essential features of a *Medicago* informatics portal should be established soon. Ideally, this group should prepare a written report distributed back to workshop participants, the International *Medicago truncatula* Steering Committee, the broader research community, and relevant funding agencies.

Functional Genomics Focus

There is a broad range of data types that needs to be captured for successful *Medicago* functional genomics. This goes beyond transcript profiling to include proteomics and metabolomics, which are emerging as important research areas in *Medicago*. In contrast to genome sequence data, functional genomics data are diverse and heterogeneous. They range all the way from simple transcriptional profiling to RNAi and mutagenesis, from proteomics to metabolomics, all encompassing diverse technologies. Success of *Medicago* as a reference legume will require efficient capture, storage, and integration among these diverse data types. While a small number of archival databases have been established for transcriptional profiling data, the *Medicago* community itself needs to take the lead in integrating and analyzing most aspects of functional genomics results. Some tools will need to be developed specifically for *Medicago*, though many can be adapted from those created in other systems. Given the diverse nature of functional genomics data, analysis tools need to be highly modular, along the lines of Gbrowse and other GMOD software, working to avoid “one-off” designs whenever possible. Any tools that are developed need to be logically interconnected to the *Medicago* genome sequence as well as the informatics resources of other species.

The community should coordinate experimental conditions, methodologies, and data formats so web-accessible data are truly comparable. Comparisons between functional genomics experiments, including ecotypes, populations, experimental conditions, time points, and developmental stages, all require comparable data sets. Strategies to integrate meta-data, including experimental methodology and data analysis methods into *Medicago* functional genomic databases are essential. This will require proactive coordination among leaders in these areas, including the *Medicago* and broader

community of legume researchers. Therefore, a task force to establish standards for cross-referencing among the features of different array platforms and other functional genomics tools should soon be established. Moreover, participation in the Plant and Trait Ontology initiatives should be pursued, providing a structured environment to coordinate how information about different experimental conditions is described.

Eventually, functional genomics results should provide a basis for annotation of biological function to genes. The results of functional genomics will soon enable gene annotation that is far more informative than automated or even expert annotation. In the not-too-distant future, results of functional genomics should form the basis for gene annotation and gene ontology assignments. Planning for this stage of post-genomic community annotation should begin early and involve cooperation between the genome sequencing and functional genomics communities.

CONCLUDING REMARKS

Clearly, there is great enthusiasm for the *Medicago* genome sequencing effort and the research opportunities it provides, especially as a basis for successful genome sequencing in crop legumes. The *Medicago* sequence needs to be finished to high quality with progressively better annotation. This workshop sought to develop a roadmap for the future of *Medicago* sequencing, annotation, and informatics. Automated annotation, currently embodied in the IMGAG initiative, needs to be enhanced by full-length cDNAs and the finished genome sequence, eventually embracing community input. A thoughtful combination of *Medicago*-centered and legume-wide databases is essential, and these resources need to include both bioinformatics centers and “grass root” efforts. Success will depend on continued progress by the sequencing consortium, frequent communication between biologists and informaticians, active planning to integrate *Medicago* data into broader legume genomics, and interoperability among informatics initiatives.

To accomplish these goals, the momentum established at the Asilomar workshop needs to be maintained. The task forces envisioned at the workshop should be set up soon and coordinated with the International *Medicago truncatula* Steering Committee. Ideally these task forces will also coordinate closely with existing groups that focus on the genomics of other legumes, potentially joining forces to create a new body that helps to coordinate genomics research across all legumes. In the end, success will rely on a combination of engaged and creative researchers with adequate and timely support from funding agencies.

***Medicago* Genomics & Informatics Workshop**
Asilomar, California — June 9 – 10, 2005
Participants

BEAVIS Bill	National Center for Genome Resources
BUELL Robin	The Institute for Genomic Research
CANNON Steven	University of Minnesota
CHEUNG Foo	The Institute for Genomic Research
COOK Doug	University of California - Davis
DEBELLE Frederic	INRA / Toulouse, France
DENNY Roxanne	University of Minnesota
ELLIS Noel	John Innes Centre, U.K.
FRUGOLI Julia	Clemson University
GEPTS Paul	University of California - Davis
GOUZY Jerome	INRA / Toulouse, France
HARRISON Maria	Cornell University
HERNANDEZ Gina	CCG-UNAM, Cuernavaca, Mexico
JACKSON Scott	Purdue University
KNAPP Steve	University of Georgia
MAY Greg	Noble Foundation
MAYER Klaus	Munich Informatic Center (MIPS), Germany
MENDEZ Pedro	Virginia Tech / Virginia Bioinformatics Institute
MUELLER Lukas	Cornell University
OLDROYD Giles	John Innes Centre, U.K.
RETZEL Ernie	University of Minnesota
RHEE Sue	Carnegie Institute of Washington
ROE Bruce	University of Oklahoma
ROMBAUTS Stephane	Ghent University / VIB, Belgium
SAMAC Debby	University of Minnesota
SATO Shusei	Kazusa Center, Japan
SCHOOF Heiko	Max Planck / Cologne, Germany
SIEGFRIED Majesta	University of Oklahoma
SHOEMAKER Randy	ARS / Iowa State University
SINGH Karam	CSIRO / West Australia
STACEY Gary	University of Missouri
TOWN Chris	The Institute for Genomic Research
TUSKAN, Jerry	Oak Ridge National Laboratory
UDVARDI Michael	Max Planck / Golm, Germany
VANDENBOSCH Kate	University of Minnesota
WURTELE Eve	Iowa State University
YOUNG Nevin	University of Minnesota
ZORN Manfred	National Science Foundation / BIO-DBI (Observer)