POPcorn – A Project Portal for Corn
Project Summary

SENIOR PERSONNEL
Carolyn J. Lawrence, PI, Iowa State University and USDA-ARS, Ames, IA, USA
Taner Z. Sen, co-PI, Iowa State University and USDA-ARS, Ames, IA, USA

SCIENTIFIC OBJECTIVES AND APPROACHES
Over the course of the past 7 years, the NSF, USDA, and DOE jointly supported and coordinated the development of resources to paint a picture of the maize genome's content and how it is organized. These resources enabled an informed approach to be developed for sequencing the genome, a project that is currently underway. Due to the creation of many online resources to enable access to these projects' data, a number of secondary challenges have arisen. Most importantly, maize researchers cannot easily leverage these data because the online locations of the generated resources are not easy to locate and the sequence-indexed resources generated by the individual projects must be searched independently. In addition, it is often the case that when a project's funding period ends, the generated data are lost because they are not moved to long-term repositories: these once-funded project sites degrade over time and sometimes disappear entirely.

This proposal offers a plan to meet these challenges in collaboration with the community of maize researchers by launching POPcorn (**P**r**O**ject **P**ortal for **corn**), a needs-driven resource and data pipeline. POPcorn will make available (1) a centralized Web-accessible resource to search and browse ongoing maize genomics projects, (2) a single, stand-alone tool that makes use of Web services and minimal data warehousing to enable researchers to carry out sequence searches at one location that return matches for all participating projects' related resources, (3) a set of tools that enable collaborators to migrate their data to MaizeGDB, the long-term model organism database for maize genetic and genomic information, at their projects' conclusion, and (4) generalized, freely available code that other research communities could use to meet similar needs.

INTELLECTUAL MERIT
By developing novel Web interfaces and data exchange protocols and collaborating with researchers to make their data available, a single point of access to ongoing research projects will be created, thus fulfilling a specific unmet need for maize research. By creating tools to upload data into MaizeGDB at the collaborating projects' close, the full life cycle of research projects will be supported and generated data will be preserved.

EXPECTED BROADER IMPACTS OF THE PROPOSED PROJECT
POPcorn will provide an essential public resource that will serve as a bridge for data storage and accessibility from maize project databases to long-term warehousing solutions. The underlying premise is that the development of the project portal, upload tools for maize data, and tutorials will serve as an outreach activity directly, support basic research, and accelerate the acquisition and utilization of new knowledge that translates directly into this important production crop's agronomic improvement. Thus, POPcorn will aid in the identification of the molecular-level phenotypes manifesting as traits that plant breeders select for and will lead to improvements in food, fuel, and nutrition. The development of the collaborative database management solutions for POPcorn represents a first effort to coordinate access to research outcomes within a research community, and developed solutions will be generalized for use by other research groups. In addition, by actively participating in the NSF's Plant Genome Research Outreach to American Indian Undergraduates program at Iowa State University, POPcorn personnel will contribute to mentoring underrepresented students who otherwise do not have the opportunity to participate in research at their home institutions. Thus, these activities will promote research, education, and dissemination of maize data to a broad audience, while developing a new generation of scientists.

POPcorn – A Project Portal for Corn
Project Description

I. INTRODUCTION

Although many resources have been developed to move plant genomics (and, more specifically, maize genomics) forward, a question remains: How accessible are the various resources that have been funded separately? It is often the case that researchers learn about a new data source relevant to their research (at a meeting or by word of mouth) only to become frustrated when they are unable to locate the project's online resources. Even more exasperating is the repetitious situation that arises when researchers know how to locate pertinent project sites, but find that the tools to access the outcomes of each project are stand-alone and specific to each individual project website. That is, the sites cannot be searched simultaneously. This is a noted problem for the maize research community (see the Allerton Report at http://www.maizegdb.org/AllertonReport.doc) because many projects create sequence-indexed resources (e.g., microarray probes, insertional mutants, etc.) that are available only at the projects' stand-alone online websites. In addition, when funding for a project ends, it is often the case that the data generated are not moved to long-term repositories for preservation. The once-funded project sites degrade over time, and sometimes disappear entirely. When the data disappear, generated resources are effectively lost.

The overall objective of this proposal is to develop unified public resources that facilitate access to the outcomes of maize genetics and genomics research projects and ensure their sustainability by migrating them to MaizeGDB, the maize Model Organism Database (MOD) [Lawrence *et al.* 2004, Lawrence *et al.* 2005, Lawrence *et al.* 2007]. POPcorn (the **P**r**O**ject **P**ortal for **corn**) will serve as a centralized Web-accessible resource for gaining access to outcomes of maize research projects. By working in tandem with MaizeGDB, the long-term repository for maize genetic and genomic information, a pipeline will be created whereby ongoing projects' data are accessible via POPcorn and their outcomes flow into MaizeGDB at the projects' conclusion. The underlying premise is that by creating a single point of access, researchers' valuable time will be saved and resources they might not otherwise know of will be accessible for inclusion in their analyses. In addition, by bringing collaborators' project data into MaizeGDB at the end of their funding periods, these valuable data will be preserved long-term.

SPECIFIC PROJECT OBJECTIVES

Objective 1: **Enable maize researchers to easily locate and leverage community databases that contain large-scale datasets by creating a single, unified portal: POPcorn.** The inability to locate project data efficiently impedes the utilization of large-scale datasets created to serve as community resources. To meet this need, access to distributed maize data will be centralized

Objective 2: **Enhance maize research by allowing scientists to carry out sequence searches through POPcorn, which will encompass all collaborating projects' sequence-indexed resources.** Focusing on a sequence-centric approach, integration of various distributed databases will be realized through the creation of the online portal, which will enable sequence searches across many databases to be initiated at one location. Collated results from the collaborating repositories will be returned to POPcorn, displaying sequences and related information listed side-by-side.

Objective 3: **Preserve maize projects' data long-term at their conclusion by transferring raw data and associated annotations to MaizeGDB.** Most (if not all) collaborating projects have only short-term funding. Migration of these projects' data and annotation to MaizeGDB will ensure long-term data preservation.

Objective 4: **Make the portal and all projects' data interoperable and accessible for third-party use.**

a) Because the maize research community is unlikely to be the only group in need of solutions to support centralized access to project data, the code to be developed will be generalized and made freely available for download and customization by interested groups.

b) All collaborating projects' data that are manipulated for inclusion in the POPcorn resource will be assigned semantic tags, making them accessible for use via Web services. Because the data could be of use in third-party pipelines, they will be made freely available for general use.

II. BACKGROUND

The creation and maintenance of bioinformatics tools and resources only has value to the extent that they are used, and integrating resources to enable ease of use must be actively pursued [Cannata *et al.* 2005]. **What would it take to integrate independent resources into a single point of access during the projects' active phase? How can we ensure that valuable resources developed are not lost after their term of funding ends?**

To understand how these problems can be best addressed, it is useful to consider different classes of data repositories as knuckles and nodes [Stein 2003] and to describe how these knuckles and nodes can interact. For this proposal, the short-lived project databases and associated websites are nodes whereas a long-term portal or repository that can overlay and organize the nodes' data into one resource serves as the knuckle. Integration of project data into the knuckle may be accomplished by way of at least three approaches: link integration, view integration, and data warehousing [Stein 2003]. Link integration can be at the level of incorporation of context-sensitive Web addresses into the knuckle's Web interface, or it can be implemented in a more sophisticated manner if the nodes are equipped (with respect to both technology and human resources) to allow the use of the Web services (published interfaces to a type of data or computation; [Stein 2002]). View integration requires the utilization of cross-database query languages at all participating sites, and has not really caught on in the research community. Data warehousing brings all data under one roof (such as the MOD) but is not an ideal answer to the problem during a project's active period. For example, while integrating all maize project data into MaizeGDB while a node is active and funded might seem to be a good solution, it is not reasonable given that active projects continually add new data to their databases. Migration of project data to the research community's MOD only makes sense once the project is completed. Otherwise updates to the MOD require the time of curators, which is the most limited resource for most (if not all) MODs. In addition, integration of the data into an overarching, single database oftentimes fails to preserve a real asset to the individual research projects: personnel who have interests and expertise that add value to the data [Stein 2003]. However, limited data warehousing that would enable improved link integration (described in Section IV.1.B) is expected to prove useful during a project's active phase. For long-term data accessibility after a node is no longer active, migration of project data to the knuckle is the solution.

Creation of a pipeline to support accessibility to ongoing maize projects will improve access to projects already funded by the NSF. Maize has long been the number one production crop in the US, but in 2001 it also became number one in the world. The diversity represented by maize is unparalleled in both a phenotypic and molecular sense, and it provides a unique vehicle to understand the basis of genetic diversity while exploring questions related to evolution, domestication, trait expression, functional allelic diversity, and the processes that shape such events and their outcomes. Thus far, the NSF Plant Genome Research Program alone has committed an estimated $93,058,482 (not including the maize genome sequencing project which is a NSF/DOE/USDA Joint Program estimated to cost $29,450,001) to the creation of maize-specific genomics resource development over the past 5 funding periods (see http://www.nsf.gov/bio/pubs/awards/pgr.htm years 2002-2006). To build upon these investments, to support basic research, to accelerate the acquisition and utilization of new knowledge, and to leverage existing resources for maize it is proposed here that a pipeline to support data accessibility from the funded project sites (nodes) be developed. This pipeline would serve the projects throughout their funding period by making all data accessible via an integrated portal (POPcorn; a short-term knuckle), and would make available mechanisms for project personnel to upload their valuable data to MaizeGDB (a long-term knuckle) toward the end of their funded period. The pipeline would benefit researchers by helping them to locate resources to aid in their own research and would ensure that valuable project data would be preserved long-term at MaizeGDB. **As the MOD for maize, MaizeGDB is funded by continuous congressional appropriations through the USDA-Agricultural Research Service (USDA-ARS). For MaizeGDB to serve as the ultimate repository for both project data and the long-term home for the POPcorn**

resource ensures that the outcomes of the project are preserved and that the technical solutions derived will be kept up-to-date. It is also proposed here that the POPcorn code be generalized as a software solution that any research community group or MOD personnel could download (from SourceForge; http://sourceforge.net), customize, and populate to fulfill their own specific project integration needs at little to no cost. Note that this project is complementary and synergistic with the existing maize resources and that its creation would be neither redundant nor duplicative.

III. RELEVANCE AND JUSTIFICATION – UNIFYING RESOURCES FOR MAIZE RESEARCH

III.1. The maize community needs a single portal to genetics and genomics data.

Leaders of the maize genetics community met to discuss the strengths, challenges, and initiatives that will define the future of maize research in March of 2007, immediately prior to the Annual Maize Genetics Conference, at the Allerton park and retreat center in Monticello, Illinois. Their report (accessible at http://www.maizegdb.org/AllertonReport.doc) is the community's assessment of the key biological issues that define research goals and the resources needed to help to achieve those goals. **Centralized informatics resources are a need repeatedly noted in the Allerton Report as being essential to realizing researchers' goals.** The following is a list of excerpts from that document:

1. "Centralized databases with increased funding are needed now."
2. "Indexed reverse genetic resources need to be finalized and will accelerate many areas of research. Current mutagenesis libraries should be indexed with new technologies."
3. "Databases and stock center capacity will be enhanced, coordinated and supported"
4. "…a fully annotated, accessible, and centralized sequence database will be essential because all additional resources depend upon robust integration of sequence information. The sequence project site must be transitioned into a community-based permanent platform that will have robust long-term support. The community expressed the desire that MaizeGDB should become the centralized sequence resource soon after the genome is complete (2009-2010)."
5. "…coordination of data deposition is essential to all advances in maize research"
6. "Long-term support for MaizeGDB from USDA-ARS was recognized; however, new and creative funding mechanisms are required now to provide sufficient resources to exploit a fully sequenced genome."
7. "To improve access to maize sequence data, resources that integrate various gene models and annotation sets must be made available to MaizeGDB. Complex datasets from federally funded projects should be deposited into MaizeGDB. However, collaborations should be established between MaizeGDB and the researchers who develop these complex datasets to insure efficient and cost-effective data flow directly into MaizeGDB."
8. "A sequence-indexed collection of mutations is essential for researchers to exploit the genome sequence fully. It was noted that multiple mutagens are necessary to insure broad coverage of the genome and generate a range of allelic lesions. These would include transposon insertions, small deletions, and point mutations. It is imperative that these lines be accessible through a community web browser to facilitate dissemination of the resource. Training in the use of the resource should be an essential and embedded component of dissemination. This collection should be searchable by BLAST, browsable, and linked to readily available seed stocks."
9. "A unified phenotyping effort proposed here will require new scales of coordination within the community, will require continued advances in cyberinfrastructure, and further development of centralized databases for analysis, synthesis and dissemination of phenomics data."

POPcorn will largely meet each of the needs outlined above, is the logical expansion of two ongoing projects directed/co-directed by the PI and co-PI on this proposal (see III.2 and III.3, below), and is of great interest to various projects for which the PI serves as the informatics contact (see III.4, below) as well as to the maize research community at large.

III.2. PlantGDB's Plant Genome Research Outreach Portal (PGROP) provides a portal to outreach resources.
PI C. Lawrence serves as co-PI for **PlantGDB**, a plant sequence-centric database that provides gene structure annotation in emerging and assembled genomes, which is fundamental to comparative, functional, and translational genomics. The PlantGDB resource (http://www.plantgdb.org [Dong *et al.* 2005]) serves as a portal to plant genomic sequence data and provides sequences and maize subproject datasets (e.g., the sequence indexed maize *Ac/Ds* and Uniform*Mu* datasets) to MaizeGDB. Sequence data (raw DNA sequences, contigs, associated gene ontology annotations, and sequence-indexed subproject data) are updated to MaizeGDB monthly. One major outreach component of PlantGDB is PGROP [Baran *et al.* 2004], the Plant Genome *Outreach* Portal, which will serve as the code base for POPcorn.

III.3. MaizeGDB is the long-term central repository for maize genetic and genomic data.
PI C. Lawrence is the Director for the maize (corn) community's Model Organism Database (MOD) **MaizeGDB**, which has long-term funding provided by the USDA-ARS. Co-PI T. Sen serves as the Computational Biologist for MaizeGDB. Available at MaizeGDB are diverse data that support maize research including maps, gene product information, loci and their various alleles, phenotypes (both naturally occurring and as a result of directed mutagenesis), stocks, sequences, molecular markers, references, and contact information for maize researchers worldwide. Also available through MaizeGDB are various community support service bulletin boards including the Editorial Board's list of high-impact papers, information about the Annual Maize Genetics Conference, and the Jobs board where employment opportunities are posted. MaizeGDB is freely available and can be accessed online at http://www.maizegdb.org. Described below are major collaborations between the MaizeGDB group and various ongoing NSF-funded projects to illustrate that the PI and co-PI are poised to integrate various database resources for maize research (see also attached letters of collaboration) and to take in collaborating project data to MaizeGDB at those projects' close.

III.4. The NSF encourages funded maize projects to deposit data into MaizeGDB for long-term preservation. Projects of interest for this proposal and for which the PI currently serves as the informatics contact or coordinates data migration into MaizeGDB include:

1. *Cytogenetic Map of Maize* NSF DBI-0321639, $1,019,029 (estimated), 9/1/03 – 8/31/07. H. Bass, PI. No funds to C. Lawrence.
2. *Cyberinfrastructure for (Comparative) Plant Genome Research Through PlantGDB* NSF DBI-0606909, $3,469,595 (estimated), 7/15/06 – 7/30/10. V. Brendel, PI, and C. Lawrence, co-PI. $224,964 (estimated) to C. Lawrence for outreach.
3. *Maize TILLING Project: Reverse Genetics of Maize Point Mutations* NSF DBI-0604765, $1,617,351 (estimated), 8/15/06 – 7/31/08. C. Weil, PI. No funds to C. Lawrence.
4. *Functional Genomics of Maize Chromatin* NSF DBI-0421619, $5,528,960 (estimated), 9/1/04 – 8/31/09. C. Cone, PI. No funds to C. Lawrence.
5. *Construction Of Comprehensive Sequence Indexed Transposon Resources For Maize* NSF DBI-0703273, $3,703,500 (estimated), 9/1/07 – 8/31/11. D. McCarty, PI. $73,885 to C. Lawrence.
6. *Cell Fate Acquisition in Maize* NSF DBI-0701880, Funding pending, 7/1/07 – 6/30/12. V. Walbot, PI. No funds to C. Lawrence.
7. *Functional Genomics of Maize Reproduction* NSF DBI-0701731, Funding pending, 7/01/07 – 6/30/12 Matt Evans, PI. No funds to C. Lawrence.

IV. PROJECT DESCRIPTION

In summary, this proposal outlines three resources to be created to support maize genetic and genomic research: (1) a portal to be created to enable access to active maize research project websites, (2) integration of sequence similarity search tools into the portal to create a single-point-of-access into all collaborators' sequence-indexed online resources, and (3) data upload tools which would enable researchers to migrate their data into MaizeGDB at the project's close.

This will be accomplished by adapting existing PGROP code for maize project representation, loading project-specific information into the created resource (called POPcorn), putting the services in place to enable centralized sequence searches, and the design and creation of data upload tools for MaizeGDB. Table 1 outlines a timeline and deliverables for the project.

Table 1. Milestones, deliverables, and timeline

Milestones and *Deliverables*	**Months**			
	1	7	13	19
Acquire the full PGROP code base from collaborator V. Brendel	X			
Contact maize researchers (via email and trips to project sites) to to assemble project descriptions, initiate discussions on Web services requirements, and define upload tool requirements,	XXX			
Release the public POPcorn website for maize with capability to browse to described projects (version 1.0)	XXX			
Design and create upload tools	XXXXXXXX			
Acquire the maize sequence set and configure the stand-alone POPcorn BLAST service	XXX			
Contact and visit project sites to set up Web services and other data exchange protocols			XX	
Release POPcorn version 2.0 with BLAST services in place and begin work toward generalizing the software			XX	
Generalize the POPcorn code for stand-alone, customized use by other research communities			XXXXXX	
Deposit developed code into SourceForge			XX	X
Migrate POPcorn to the generalized code				XXX
Add MaizeGDB upload tools to the POPcorn resource				X
Release POPcorn version 3.0 running on the generalized code base				X
Prepare and *submit for publication a description of the generalized software and POPcorn implementation for maize*				X
Migrate curation of POPcorn to MaizeGDB staff				XXX

IV.1 *Objective 1:* Enable maize researchers to easily locate and leverage community databases that contain large-scale datasets by creating a single, unified portal: POPcorn.

IV.1.A. Adapt the existing PGROP code base ([Baran *et al.* 2004]. PGROP (http://www.plantgdb.org/PGROP/pgrop.php) is a functional database and website combination made up of 31 MySQL tables and 87 PHP files. It was created as the outreach component to PlantGDB with the mission to provide a centralized access point for locating Plant Genome Research Outreach activities, programs, and resources. The site seeks to be an actively maintained portal that serves the needs of a wide-ranging audience including high-school students, teachers, undergraduates, university faculty, and even the public at large. It has integrated search and browse mechanisms, and enables the classification of records based upon record characteristics including target audience, research organism, and data type. It also makes available information on how to locate information on opportunities including internships, grants and fellowships, and educational enrichment. Collaborator V. Brendel is the PI for PlantGDB and PGROP, and has committed to making PGROP's source available to POPcorn project personnel for use as the code base for POPcorn and as the basis for the creation of a generalized tool for use by other research groups (see attached letter of collaboration).

To utilize the existing PGROP code for POPcorn, specific content within the source code must be changed. For instance, in the PGROP interface's left margin (Figure 1; p.7), the links and the content for "resources for: High School Students" could be replaced in the interface's source code with content "sequencing projects: B73 Genome" (Figure 2; p. 8) for POPcorn. In addition, information regarding which database connections to make, the color scheme, and other specifications would be altered.

It is anticipated that it will be more advantageous to adapt the PGROP code for POPcorn by customizing it for maize research and to secondarily generalize the code for other research groups' use instead of carrying out these processes in reverse order (i.e., generalizing the code first) for two reasons: (1) to enable the POPcorn Programmer to become familiar with the PGROP code by adapting it to maize the first time through and (2) to get a usable product out to the community of maize researchers more quickly.

IV.1.B. Acquire and add project-specific content. To include maize project-specific information in POPcorn, maize research project personnel must contribute project descriptions and the links to enable access to the project sites through POPcorn. Maize collaborators (working at the 'nodes'), including but not limited to those researchers who provided a letter of collaboration (see Table 2), will be contacted at conferences as well as via email and telephone to invite their participation and to ask them to provide a description of their project's online-accessible resources. Many projects already have committed to providing this sort of content (see attached letters of collaboration). This will enable maize project-specific data to be located and browsed through in the same way that, e.g., bioinformatics projects can be accessed through PGROP (see Figure 1). Once the PGROP code has been adapted for POPcorn and the database has been populated with project-specific information, those descriptions and links to the primary repositories will be added to the database and can serve as planning documents in developing the MaizeGDB data upload tools (see IV.3, Objective 3).

Table 2. Collaborators' anticipated roles.

Collaborators (see letters)	**Role and project website**
Alice Barkan	Providing maize orthologs as they are linked to rice including annotations of domain organization, gene models, phylogenetic trees, and intracellular targeting predictions. POGs (http://pogs.uoregon.edu)
Hank Bass	Delivering sequence information for the maize core bin markers and cytological probes to enable integration with various map types. CYTOMAIZE (http://www.cytomaize.org/)
Volker Brendel	Making available the PGROP software code base and lending project guidance. PGROP (http://www.plantgdb.org/PGROP/pgrop.php)
Tom Brutnell	Allowing BLAST searches of the f*Ds* collection. *Ds* insertion sites are mapped to BAC and GSS assemblies as well as a browsable list of putative gene insertion sites using the MAGI assemblies and their gene annotations. *Ds* insertions (http://www.plantgdb.org/prj/AcDsTagging/)
Karen Cone	Identifying maize chromatin genes including RNA-based silencing genes and making accessible new chromatin mutants for maize (RNAi lines and recessive mutations). The Maize Chromatin Project (http://www.chromdb.org/)
Matt Evans	Delivering data on the effect of gametophyte mutants' lethality based on the number of genes expressed in gametophytes. The Maize Gametophyte Project (new)
Damian Gessler	Creating semantic Web tools and protocols. The Virtual Plant Information Network (VPIN; http://vpin.ncgr.org/)
Jeff Glaubitz	Creating mechanisms of access to diversity data stored at Panzea. (http://www.panzea.org/)
Sarah Hake	Collating inflorescence information in ear and tassel cDNA sequences and mutant phenotypes. (http://gremlin1.gdcb.iastate.edu/MIP/EMSphenotypeDB/)
Don McCarty	Constructing sequence-indexed transposon resources using the Uniform*Mu* population. Uniform*Mu* Sequence Indexed Insertions (new)
Jo Messing	Accessing sequence-based maize storage protein data. Storage Proteins in Maize (http://pgir.rutgers.edu/)
Steve Moose	Discovering genes associated with nitrogen responses in maize through a functional genomics program. NitroGenes (http://nitrogenes.cropsci.uiuc.edu)
John Quackenbush	Integrating data from international Maize EST sequencing and gene research projects. Dana Farber's Gene Indices (ZmGI; http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=maize)

Pat Schnable	Creating centralized accessibility to the Maize Assembled Gene Islands assemblies and the Maize Genetic Mapping Project outcomes. (MAGI; http://magi.plantgenomics.iastate.edu/ and MaGMaP; http://maize-mapping.plantgenomics.iastate.edu/)
Virginia Walbot	Making available transcriptome and protein profiling data plus mutant and cytological analyses of hundreds of male-sterile lines to understand anther development. Maize Anther Development (new)
Doreen Ware	Moderating interactions with and connections to both Gramene and the Maize Genome Sequencing Consortium's project staffs. Gramene (http://www.gramene.org/) and the Maize Genome Browser (http://www.MaizeSequence.org)
Cliff Weil	Delivering Maize TILLING Project data as well as providing guidance on how to handle presentation of sequences and phenotypes associated with the various mutation projects. The Maize TILLING Project (http://genome.purdue.edu/maizetilling/)
Roger Wise	Moderating interactions with and connections to PLEXdb for maize microarray probe set sequence information integration. PLEXdb (http://plexdb.org/)
Yeisoo Yu	Providing acces to sequences of 30,000 FLcDNA clones from two cDNA libraries of varied tissues and stress treatments. The Maize Full-length cDNA Project (http://www.maizecdna.org/)

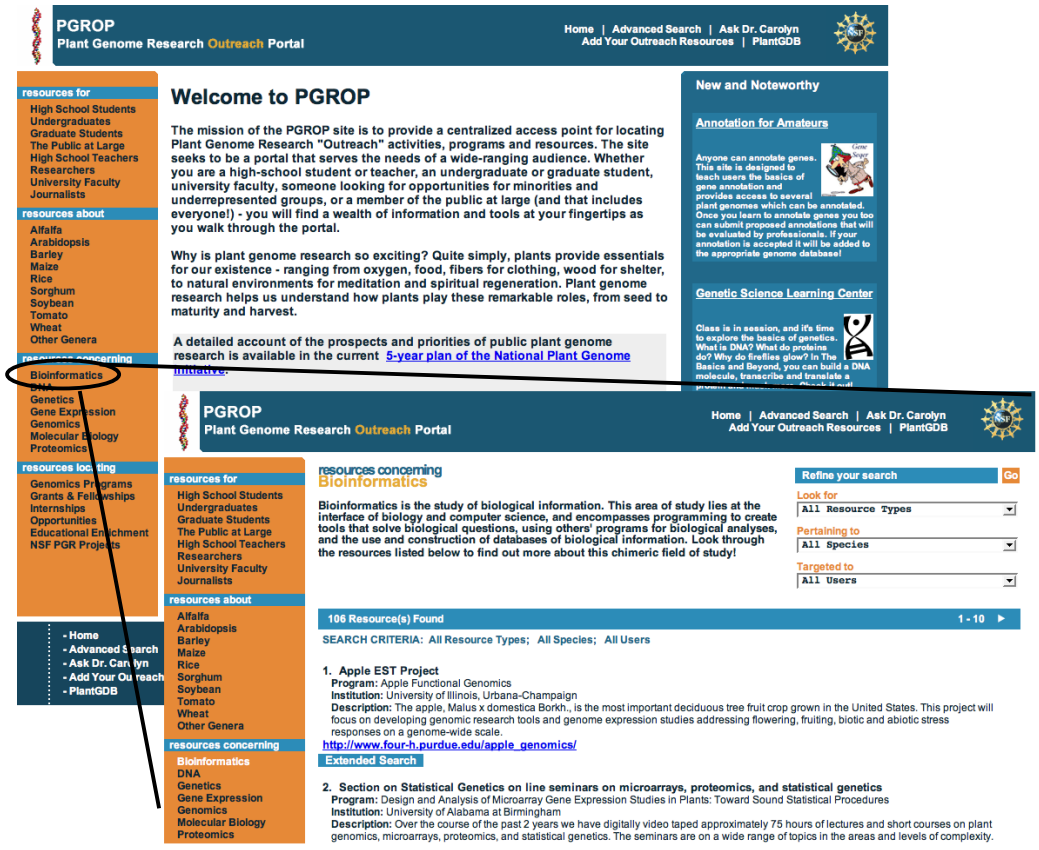


Figure 1 (left). Entry page and bioinformatics resource browser for PGROP. When researchers navigate to PGROP, they find the project icon on the upper left, a list of available resource types in the left margin, and tools at both the top right and bottom left. Clicking on the link to "resources concerning: Bioinformatics" leads the user to a page of bioinformatics resources made up of individual project descriptions and links to access the described resources off-site.

Figure 2 (below). Entry page mock-up for the POPcorn site. When researchers navigate to POPcorn, they will find the project icon in the upper left, a list of project types in the left margin, and tools at both the top right and bottom left. Note that for POPcorn v. 1.0 the BLAST service link in the upper right will not be present. Instead, a note that the BLAST service is anticipated for release on a particular date will be placed in that position. The main content of the page will describe what the POPcorn site seeks to provide for researchers, and the right margin will show links to high-profile and well-described projects accessible through the site.

The Programmer will carry out entry of project-specific data into the POPcorn database. Although there are integrated project description tools available in the PGROP package that will be adapted for use in POPcorn, the first release will need to be populated by project personnel to enable a fully-functional initial release. This will constitute POPcorn version 1.0.

As researchers begin to utilize POPcorn, it is anticipated that they will see that entering and updating the information directly is to their benefit. Once POPcorn has become a useful tool to the community, direct entry of project data by collaborating projects' personnel will be encouraged.

It should be noted that all project data listed at POPcorn will be made accessible via both search and browse functions, and will be available for download in various formats (e.g., HTML, MAGE/ML, XML, spreadsheet compatible, etc.) to enable researchers to manipulate the information easily.

IV.2. ***Objective 2:*** **Enhance maize research by allowing scientists to carry out sequence searches through POPcorn, which will encompass all collaborating projects' sequence-indexed resources.**

IV.2.A. Add sequence search capabilities to POPcorn. Drawing on our experience with PGROP, MaizeGDB, and PlantGDB, we will seek to provide a consistent Web interface for maize genomic data that enables access to diverse types of data from at least two perspectives: (1) by searching or browsing project types (e.g., projects generating insertional mutants, microarray, etc.; described in Objective 1) and (2) by submitting a query sequence for similarity analyses. This is especially important given that the maize research community has generated so many sequence-tagged resources that are currently available only via stand-alone repositories. Being able to search these sequences via POPcorn rather than directing researchers to GenBank makes sense because GenBank does not allow the connection of their sequence resources to the materials and reagents that give the sequences added value. By making the projects' sequences able to be searched from a single portal where the materials and other resources generated by the projects are listed alongside the sequence matches, the full utility of the generated resources will be accessible to researchers, thus making the projects' outcomes more easily accessible and useful to the community.

IV.2.B. Set up a BLAST server for POPcorn. Simultaneous with the Programmer's creation of POPcorn v. 1.0 (Objective 1), the Solution/Application Architect will be working with collaborator D. Gessler's VPIN staff (see letter) to select programming languages and softwares to support the creation of a centralized point of access for sequence-based searches.

The sequence set for BLAST [Altschul *et al.* 1997] analyses will consist of all PlantGDB-processed *Zea* subspecies' sequences present in GenBank, and the BLAST server setup will be handled in collaboration with MaizeGDB personnel. In order to set up the BLAST service, a dataset of sequences must be acquired. Through a customized sequence delivery pipeline, PlantGDB provides a sequence set for maize (consisting of all public maize sequences including EST, cDNA, GSS, STS, HTC, and genomic DNA sequences from GenBank as well as the Uniprot protein sequence set) to personnel at MaizeGDB for inclusion in the MaizeGDB's monthly update [Lawrence *et al.* 2005, Dong *et al.* 2005]. This same sequence set will serve as the basis for POPcorn's BLAST service. To set up the BLAST server itself, three major components must be installed and configured: the Web forms used input sequences and parameters, the actual BLAST program itself, and the tools and interfaces required to deliver and display search results [Dong & Brendel 2005]. In addition to the database of sequences and these three software components, POPcorn will require functioning Web services and (for those projects that cannot support Web services) files of GenBank identifiers utilized by the participating project groups. Data exchange with the participating project sites will be coordinated by the Programmer who will develop documentation on how to participate and will also travel to collaborating project site locations to train personnel on how to manage this method of access to their data. Once the POPcorn BLAST server is up and running, a link for accessing the service will appear in the upper right corner of the POPcorn website as shown in Figure 2. This release will constitute POPcorn v. 2.0 and will be announced via an email to maize cooperators through the MaizeGDB mailing list and will be noted in the MaizeGDB news column on the right margin of the MaizeGDB home page. For a diagram of how the system will work, see Figure 3 (next page).

To enable access to the nodes' data, Web services will be implemented in collaboration with D. Gessler's VPIN staff (see attached letter). To enable nodes that do not have the human or technical resources to utilize Web services to participate, a method for POPcorn to warehouse datasets of GenBank identifiers utilized by the nodes will be created. The resulting BLAST service with access to nodes' related information via link integration will be released to the public. This will constitute POPcorn v. 2.0.

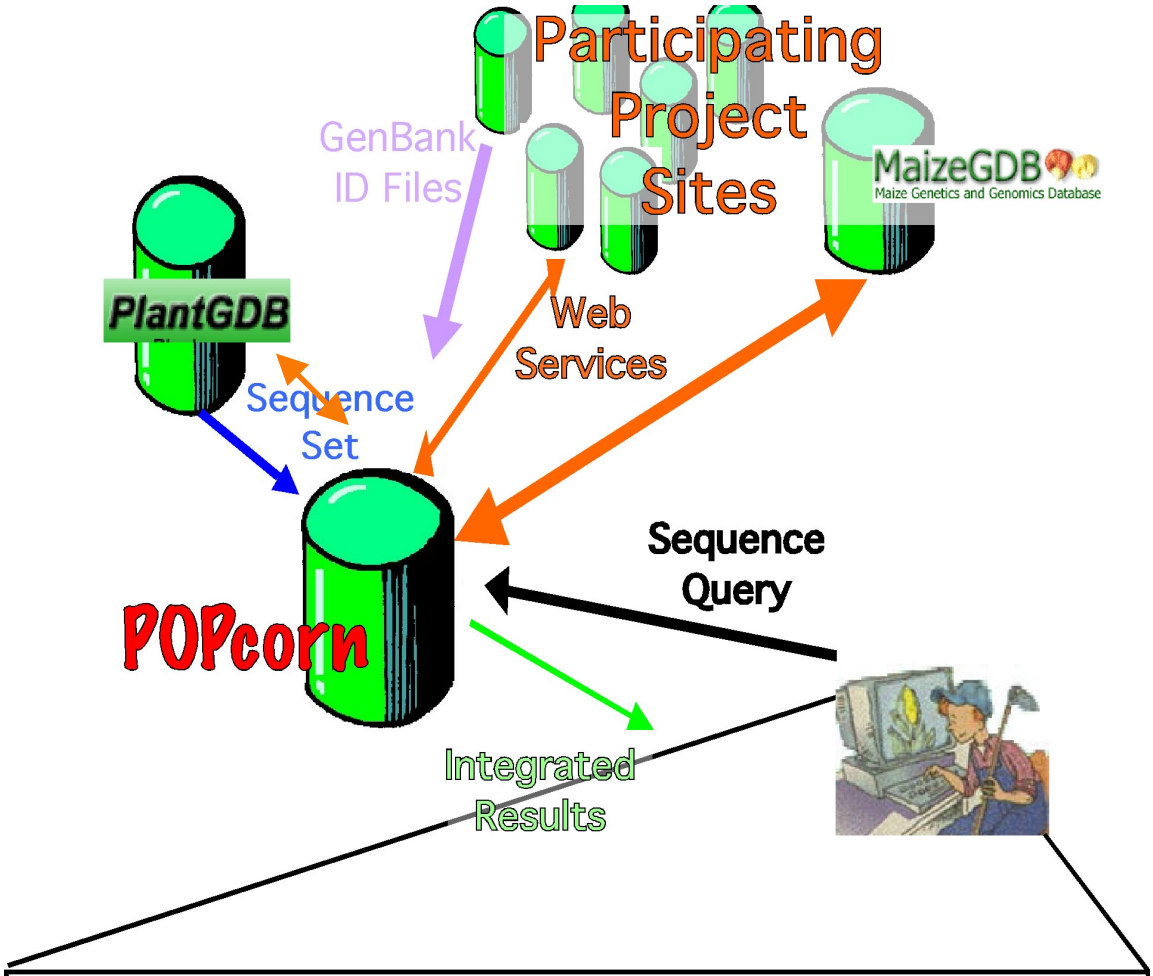
Shown in Figure 3 is a method of BLAST service management wherein the Web services component is limited to the exchange of GenBank identifiers. This would enable a single BLAST to be carried out at POPcorn, and participating project sites would not be required to maintain a separate, Web accessible BLAST service. In practice, there exist several research groups that would likely prefer to have Web services deliver the query sequence and have the BLAST search carried out on their own projects' machines using generated project data. Results from the project site's BLAST service would then be delivered back to POPcorn for integration into the results page. This sort of method of data handling would make it possible for POPcorn to deliver, e.g., projects' contig sequences that may not be present in GenBank. Both types of Web services would be enabled by POPcorn.

IV.3. *Objective 3:* Preserve maize projects' data long-term at their conclusion by transferring raw data and associated annotations to MaizeGDB.

Oftentimes resources are created with no clear plan for their continued support. **To ensure that the POPcorn resource for maize does not become stale, maintenance of the site's content and interface will be continued after the project's funding ends by developers and curators at MaizeGDB.** This is made possible because the needs to be met by POPcorn are in line with the mission of MaizeGDB: to support the bioinformatic needs of maize researchers. It also should be noted that maintenance of the POPcorn site also could serve as a queue for keeping track of which data sets are available and for keeping in contact with maize project personnel. Close cooperation with project sites through POPcorn will enable MaizeGDB personnel to be in contact with and aware of the maize projects' timelines, and will help MaizeGDB personnel to do a better job of acquiring data in the long run.

The MaizeGDB data upload tools will be built upon the Web services to be created for POPcorn as well as the existing Community Curation Suite of tools which currently enable researchers to upload their data into MaizeGDB directly [reviewed in Lawrence *et al*. 2005]. The existing Community Curation Suite of tools enables researchers to enter data piecemeal (i.e., one data point at a time) whereas the tools to be created in collaboration with POPcorn will enable bulk data upload from maize projects. Exact specifications for the tools functionalities will be developed in collaboration with maize researchers (see letters), and the Solution/Application Architect will define which existing languages and technologies will be utlilized as their foundation. Once a plan for the tools' creation has been laid out, the Programmer will set to work to create them.

Figure 3 (next page). Diagram showing how the BLAST service will work. Sequence files will be delivered to POPcorn from PlantGDB (blue arrow) as will GenBank Identifier files (lavender) for projects that cannot support Web services. When a researcher submits a query via the POPcorn BLAST (black arrow), the sequence set residing on the POPcorn server will be searched and a results file will be generated. From the results file list of hits, a query will be sent to the Web services participants (orange bi-directional arrows) to find out which databases have information associated with the results file's sequence identifiers. POPcorn will collate the Web services responses and will integrate those results with information from other projects' stored GenBank identifier files. The results file and all GenBank identifiers will be processed and delivered (green arrow) back to the user's local machine for Web display (lower panel). Listed below the identified sequences are links to results sorted by project (left) or data type (right).



Distribution of 2 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Color key for alignment scores

<40 | 40-50 | 50-80 | 80-200 | >=200

Query

0 350 700 1050 1400 1750

Distance tree of results NEW

Sequences producing significant alignments:		Score (Bits)	E Value	
gi|85540272|gb|BT024021.1|	Zea mays clone EL01N0561C08 mRNA sequ	2883	0.0	U
gi|21207833|gb|AY104755.1|	Zea mays PCO064769 mRNA sequence	2313	0.0	U

List Results by Project:

- MAGI
- *Ac/Ds*
- Uniform*Mu*
- MaizeSequence.org
- MaizeGDB

List Results by Data Type:

- Microarray Probes
- Stocks
- Loci
- Transposon Insertions
- Diversity

IV.4. *Objective 4:* Make the portal and all projects' data interoperable and accessible for third-party use.

IV.4.A. Generalize generated data-handling code developed for POPcorn for other research communities that need a unifying portal. The POPcorn Programmer will begin work toward generalizing the data handling solution by generating a modular software package that could be downloaded, customized, and populated. Particular emphasis will be placed on enabling providers to skin, swap components, and extend the software's functionality to underscore its intrinsic versatility. Once that software has been developed, its usability will be confirmed by migrating POPcorn v. 2.0 to the generalized software. Tutorials and instructional documentation on how to implement the system and manage the community participants will be created for distribution to groups interested in utilizing the software for their own needs. This version of POPcorn (v. 3.0) will be released, the generalized solution's code will be deposited into SourceForge (http://sourceforge.net) for public access, and the availability of the software will be announced by way of publication in the journal *Bioinformatics* as an Applications Note or via some other appropriate venue.

IV.4.B. Make available all collaborating projects' data via Web services for use by way of third-party tools and data analysis pipelines. Because the participating projects' data will be made accessible for POPcorn, they also will be available for use by third parties (e.g., D. Gessler's VPIN). This is a serendipitous by-product of the processes required to enable their access via POPcorn.

V. Anticipated Challenges and Vision for Project Evolution

V.1. Potential obstacles.

PI C. Lawrence and Co-PI T. Sen have extensive experience with maize genetic and genomic information and its representation. Most of the approaches POPcorn will use are well established, and no serious problems are anticipated for achieving Objectives 1, 2, and 3. One challenging aspect of this proposal might be in Objective 4 – the generalization and semantic tagging of the participating projects' code, which will require excellent coordination among the technical personnel of the participating projects, the Solution/Application Architect, and the POPcorn Programmer. To minimize the chances that this could impede progress, ample funds are allocated for travel for both the POPcorn Programmer and the Architect at various stages during the project to enable on-site, one-on-one consultation. Weekly meetings among project personnel also will be conducted.

Another potential obstacle is a lack of interest from the maize community to populate or use POPcorn for their research. However, as it is shown throughout the proposal and by way of letters of collaboration, the maize community is excited about the POPcorn project and notes the benefits of having the POPcorn project be ancillary to MaizeGDB. For the POPcorn project, we expect the maize community to show its unparalleled enthusiasm as it has with MaizeGDB. Because all data and annotations will be migrated to MaizeGDB when POPcorn's NSF funding ends, not only will the sustainability of the data, annotations, and accompanying semantic tagging technologies be ensured, but also the maize community will find that MaizeGDB itself has developed enhanced data representations and content by its connections to POPcorn. Therefore, it is expected that the addition of the POPcorn portal and the new data and annotations from the participating projects will increase MaizeGDB's overall utility and traffic.

V.2. Proposed technologies for use to support this project.

Web service implementations involve providers offering data or algorithms that conform to a common set of rules that allow clients to use the services. For the purposes of this proposal, POPcorn will serve as a client and the collaborating project resources will serve as providers. Many implementing Web services use SOAP (Simple Object Access Protocol; http://www.w3.org/2000/xp/Group) for sending XML (eXtensible Markup Language; http://www.w3.org/XML/) messages. This lets clients invoke a service as if it were a remote procedure call. To engage a service, a WSDL (Web Services Description Language; http://www.w3.org/TR/wsdl) document is provided to describe the details. While traditional Web services alone could be utilized for our efforts, it is unfortunate that Web services alone do not provide sufficient standards for inferring meaning using domain-specific annotations. We plan to work with collaborator D. Gessler to utilize semantic Web services, that is, Web services that assess

and discern appropriate data services available on the Web using semantics (i.e., the meanings and relationships of entities to one another) to better define the data to be exchanged by maize data providers and the POPcorn client. Use of Gessler's Simple Semantic Web Architecture and Protocol (SSWAP) combined with OWL (Web Ontology Language; http://www.w3.org/2004/OWL/) should enable POPcorn and its providers to implement data exchange rather quickly, and could serve as the basis for POPcorn's technological underpinnings.

V.3. Products and deliverables are the goal.
This proposal describes the current need for maize geneticists, which the creation of the described product (POPcorn) would meet. Web services and (more specifically) the semantic Web services technologies described above could be used to meet these needs, but if better technologies are identified or become available, the best solution to solve the problems will be utilized. The purpose for hiring a qualified Solution/Application Architect is to ensure that all possible technologies that could address this need are properly evaluated and considered. 'Spiral design' tactics wherein rapidly iterating design and adaptation cycles occur will be utilized to better enable agile development and adherence to emerging standards. The product, a package which consists of (1) a centralized resource to locate maize project sites, (2) sequence-based access to collaborating projects' resources, and (3) tools to enable deposition of collaborators' project data into MaizeGDB at their close, is the goal. The proposed technologies to be used to create these resources will be evaluated and re-evaluated over the course of the project, but the deliverables will go unchanged.

VI. BROADER IMPACTS

VI.1. POPcorn outreach to maize researchers.
One clear form of outreach to the community is the creation of tutorials that will demonstrate to researchers how best to utilize POPcorn. At the other end of the spectrum are the personnel working at participating project sites (the nodes). **Tutorials on how to exchange data (for those who do not have Web services capabilities) and on how to link to POPcorn (for those who do have Web services) will be created and made available. Visits to participating project sites during year two also will help to educate researchers on how best to link up with POPcorn.**

VI.2. Utility of the package for other research communities.
As has been mentioned throughout this proposal, the ability to locate the many maize research projects' outcomes has become nearly impossible given the number and breadth of projects that create large-scale datasets. It is anticipated that maize represents an 'indicator species' for problems to come in other systems, and that this phenomenon is emerging or already exists for other biological research communities. **The development of the collaborative database management solutions described herein represents a first effort at coordinating access to research outcomes within a community, and is expected to become a useful model for other groups to follow.** By making freely available the (1) generalized software, (2) semantic tagging technologies, and (3) detailed descriptions of how to manage the cooperative environment, POPcorn will serve as a basis for coordinating research groups in general in addition to coordinating data access for the maize community.

VI.3. Outreach to Native Americans.
The funds currently received by PI C. Lawrence from the PlantGDB collaboration support outreach to Native Americans. In an effort to increase their representation in the research community, an eight-week summer program to mentor Native American undergraduates in plant genomics research was begun on the Iowa State University campus. The first summer session occurred June 3 through July 29 of 2006 and was supported by supplemental funds to V. Brendel's "PlantGDB – Plant Genome Database and Analysis Tools" grant. Participating students studied *Zea mays* and other plants of importance to many Native Tribes. Students worked with USDA-ARS North Central Regional Plant Introduction scientists to carry out field-work and to collect and preserve plant material and learned to use molecular markers in the lab to characterize the Southwest maize collection, a group of accessions collected from and contributed by Native Americans to the National Plant Germplasm System. Data generated, as well as information describing cultural practices involving plants, were stored and made accessible online by the students who were taught data storage methods and how to code HTML at the command line. An advisory council made up of

Tribal Elders was involved in the program and traveled to Ames to work with the students. In this manner, the Elders' hopes for the students were conveyed alongside the outcomes anticipated by the researcher mentors. The website created by the students to document their summer experience can be viewed online at http://www.lawrencelab.org/Outreach/2006/home.html. It is anticipated that this outreach program will continue for four more years. Continued work on this front is supported by the "Cyberinfrastructure for (Comparative) Plant Genome Research through PlantGDB" grant. During the years during which the POPcorn project will take place, **POPcorn project personnel will contribute to the already successful outreach to Native Americans program by serving as mentors for one to two students who will learn to store biological information and to create Web interfaces to the outreach project's data and documentation.**

VI.4. Interagency Working Group on Plant Genomes.
Because the PI and co-PI are employees of the USDA-ARS in addition to having faculty appointments at Iowa State University, it should be noted that **the POPcorn project would constitute collaboration between the USDA-ARS and the NSF, which would be of great interest to and would further the goals of the National Plant Genome Initiative's Interagency Working Group on Plant Genomes (IWG), which includes representatives from the NSF, USDA, DOE among others.** The IWG coordinates and provides oversight for the federal investment in plant genome research through NPGI and provides oversight for the NPGI to garner interagency support and cooperation for plant genome sciences (http://www.nsf.gov/bio/pubs/reports/npgi2006/). Most significant among NPGI's accomplishments in 2005 was the initiation of the maize genome sequencing project which was jointly funded by the NSF, USDA, and DOE. Although sequencing the maize genome was a primary goal of the NPGI for some time, this joint interagency effort to develop a high-resolution picture of the maize genome became technically feasible only recently. It is anticipated that the maize genome sequence should be completed by late 2008. **Creating a better informatic infrastructure to support coordination of maize research as it pertains to understanding of the maize genome will advance our understanding of maize biology, which will in turn allow us to develop new and improved varieties as well as new uses for maize.**

VII. PERSONNEL AND MANAGEMENT PLAN

PI C. Lawrence will plan; hire and manage the project team; initiate collaboration with outside groups (including both MaizeGDB and other participating project groups); and author and release manuscripts and project announcements. **Co-PI T. Sen** will ensure the scientific relevance, usability, and data integrity of POPcorn; provide direct supervision and guidance for project personnel; and communicate with researchers in the participating projects to facilitate the integration of their data into POPcorn.

The **Solution/Application Architect** will travel to meetings and project sites to communicate with the participating projects' personnel, analyze collaborators' software architectures and database systems, and use his/her technical expertise to propose solutions to guide the POPcorn Programmer in the creation of Web services solutions for POPcorn's sequence analysis capabilities as well as in the creation of data upload tools for MaizeGDB.

The **POPcorn Programmer, in coordination with the Solution/Application Architect,** will adapt PGROP for maize, interact with participating websites' technical personnel directly by traveling to meetings and project sites, and define and implement data handling protocols for both POPcorn and the data upload tools for MaizeGDB.

Two undergraduates (Computer Science majors) will assist the Programmer by preparing data files, carrying out some software development, and beta testing developed software.

POPcorn – A Project Portal for Corn
References Cited

Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W, and Lipman, DJ (1997) Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs *Nucleic Acids Res.*, **25,** 3389-3402. http://nar.oxfordjournals.org/cgi/content/full/25/17/3389

Baran, SB, Lawrence, CJ, and Brendel, V (2004) Plant Genome Research Outreach Portal - A Gateway to Plant Genome Research 'Outreach' Programs *Plant Physiol*, **134,** 889. http://www.plantphysiol.org/cgi/content/full/134/3/889

Cannata, N, Merelli, E, and Altman, RB (2005) Time to Organize the Bioinformatics Resourceome *PLoS Comp Biol*, **1,** e76. http://compbiol.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pcbi.0010076

Dong, Q and Brendel, V (2005) Computational Identification of Related Proteins: BLAST, PSI-BLAST, and Other Tools *In* Walker, J. (ed.), *The Proteomics Protocols Handbook*. Humana Press, USA, pp. 555-570.

Dong, Q, Lawrence, CJ, Schlueter, SD, Wilkerson, MD, Kurtz, S, Lushbough, C, and Brendel, V (2005) Comparative Plant Genomics Resources at PlantGDB *Plant Physiol*, **139,** 610-618. http://www.plantphysiol.org/cgi/content/full/139/2/610

Lawrence, CJ, Dong, Q, Polacco, ML, Seigfried, TE, and Brendel, V (2004) MaizeGDB, the Community Database for Maize Genetics and Genomics *Nucleic Acids Res*, **32,** D393-D397. http://nar.oxfordjournals.org/cgi/content/full/32/suppl_1/D393

Lawrence, CJ, Schaeffer, ML, Seigfried, TE, Campbell, DA, and Harper, LC (2007) MaizeGDB's New Data Types, Resources, and Activities *Nucleic Acids Res*, **35**, D895-D900. http://nar.oxfordjournals.org/cgi/content/full/35/suppl_1/D895

Lawrence, CJ, Seigfried, TE, and Brendel, V (2005) The Maize Genetics and Genomics Database. The Community Resource for Access to Diverse Maize Data *Plant Physiol*, **139,** 610-618. http://www.plantphysiol.org/cgi/content/full/138/1/55

Stein, LD (2002) Creating a Bioinformatics Nation *Nature*, **417,** 119-120. http://www.nature.com/nature/journal/v417/n6885/full/417119a.html

Stein, LD (2003) Integrating Biological Databases *Nature Rev Genet*, **4,** 337-345. http://www.nature.com/nrg/journal/v4/n5/abs/nrg1065_fs.html