

Project Award Number: 2008-55300-04757

Progress Report: Project Director: Douches

Translating Solanaceae Sequence Diversity and Trait Variation into Applied Outcomes through Integrative Research, Education, and Extension

Since the SolCAP project started in October 2008, we have made significant progress relative to 2008-09 research, education, and extension goals. At the January 2009 Plant and Animal Genome (PAG) meeting, we held our first advisory committee meeting and received feedback from the Solanaceae Coordinated Agricultural Project (SolCAP) advisory group that was used to further refine the deliverables for the project 2009-12 proposal (see report at project website: URL 18) . Distance learning infrastructure has been established with Cornell Transnational Learning (*Obj. 1*). We led an eXtension.org workshop for plant translational genomics CAPs at the 2009 PAG. This workshop created buy-in from other plant translational genomics CAPs to establish a web 2.0 workspace for the eXtension community of practice (URL 11). We reached a consensus to name, create, and initiate a Plant Breeding and Genomics Community of Practice (PBGCoP) and are in the process of submitting an application to eXtension to formalize the CoP. Several other plant translational genomics CAPs have already initiated their own workgroups within the PBG workspace (*Obj. 2*). Breeder workshops are set for this year in conjunction with the Tomato Breeders Roundtable and Potato Association of America. A survey has been initiated to assess the baseline knowledge and expectations of breeders for markers. SolCAP co-PDs have made over 10 presentations to audiences in the US, China, and South America on the objectives and goals of our project. Such venues include the 2009 PAG, North Central and Western Regional Potato Breeding and Genetics Technical Committee meetings, Michigan Potato Industry Commission winter potato conference, Mid-Atlantic Fruit and Vegetable Conference (potato session), Potato Summit in South Carolina, Michigan High School Science Teachers Conference, California Seed Association, and Great Lakes Vegetable Working Group meeting (*Obj. 2*). Discovery of new polymorphisms for elite tomato and potato breeding populations is based on both *de-novo* sequencing and bioinformatic investigation of existing data sets. Normalized libraries for sequencing have been constructed from multiple tissues including leaves, tubers, flowers and callus for potato. For tomato, they are being constructed from three stages of fruit development, callus, leaves, roots, and flowers. To maximize SNP discovery, we have optimized bioinformatic pipelines to identify putative SNPs (eSNPs) within existing tomato and potato Expressed Sequence Tag (EST) data sets and emerging genome sequence. Pipelines have been developed to assemble Sanger-derived ESTs on a single accession basis and then identify high confidence eSNPs between genomic and EST contigs as well as between EST contigs of different accessions. We are also developing bioinformatic pipelines to identify eSNPs between genomic or Sanger-derived EST contigs and short read transcript data such as those being generated in this project. To link these bioinformatic predictions with experimental data, we are validating these predicted SNPs to benchmark our pipeline (*Obj. 4*). Illumina genotyping chips are planned for 1536 and 7600 loci for tomato and potato, respectively, beginning in 2009 (*Obj. 5*). For the tomato and potato field trials, we finalized the germplasm and intellectual property agreements for the crop panels, are developing the cooperators guides, and have begun planting these trials. The germplasm being used for the potato and tomato genome sequencing projects is integrated into the potato/tomato panels so that the SNP genotyping will be linked with the sequencing efforts (*Obj. 3*). Flexible funds from year 1 are being used to collect centralized sugar, chip color and vitamin C data in a 4x mapping population as well producing seed tubers to increase the size of the 4x mapping

population (*Obj. 6*). Relative to our education and extension goals, we established a project web site (URL 18) which is a centralized resource for SolCAP participants. The site includes project news and personnel, databases, tools, resources, project objectives and downloads. We have published three newsletters that reach nearly 400 people in the Solanaceae community (*Obj. 7*).

PROJECT DESCRIPTION

INTRODUCTION

SolCAP focuses on the two most important vegetable crops in the Solanaceae: potato and tomato. Our vision is to move translational genomics beyond commodity boundaries toward an emphasis on taxonomic groups and DNA sequence homology so that we can leverage knowledge and resources across species. Sequencing efforts in Solanaceae have led to extensive expressed EST resources and genome sequence is emerging for both potato and tomato with preliminary assemblies of both of these genomes anticipated in the near future (URL 21, 13, 7). Ultimately, understanding variation at the DNA sequence level is useful in crop improvement only to the extent that it helps us understand and/or predict phenotypic variation for agriculturally important traits. *The primary research objective of this proposal* is to provide the infrastructure to link allelic variation in genes to valuable traits in cultivated germplasm of potato and tomato. Focusing on elite breeding material will increase the probability that these solanaceous crops will benefit from genotype-based selection. *The extension and education components* will integrate training in genomics and plant breeding with curriculum aimed at students coming into the field and existing breeders seeking to make better use of sequence data in the context of crop improvement. To foster interaction across plant translational genomics CAPs we have invited other projects into our eXtension community of practice, created a Plant Breeding and Genomics workspace to leverage Web 2.0 interactive functions (URL 11), and we are developing content for publication to eXtension.org. Developing curriculum for new and practicing breeders will facilitate the use of genomic resources and databases to solve problems in the Solanaceae. We are requesting funding for the final three years of our initial four-year Solanaceae CAP awarded in 2008.

SPECIFIC OBJECTIVES AND DELIVERABLES:

- Obj. 1** *Create an education program to train graduate students in genome-based breeding.* We will educate future plant breeders in the theory and practice of translational genomics. This objective will be achieved by developing and delivering curricula through a team-taught distance education course across four universities.
- Obj. 2** *Amplify outreach efforts by developing an eXtension Plant Breeding and Genomics Community of Practice (PBGCoP) to develop continuing education material aimed at practicing plant breeders, their staff, and seed industry professionals.* SolCAP is creating a web community (PBGCoP) and is providing leadership to foster cross-disciplinary and cross-commodity networking and collaboration to generate outreach materials for publication on eXtension.org. Yearly workshops will also be tailored to plant breeders and industry professionals in the theory and practice of translational genomics.
- Obj. 3** *Collect standardized phenotypic data across multiple environments for tomato and potato.* A standardized database of phenotypes for key traits across core collections of each crop will be linked to the genotypic data, accessible through the Solanaceae Genome Network website (SGN; URL 7).

- Obj. 4** *Develop extensive sequence data of expressed genes, and identify Single Nucleotide Polymorphism (SNP) markers distributed across the genome and associated with specific candidate genes for sugar, carbohydrate, and vitamin biosynthetic pathways.* Over 2 Gb of sequence is being generated from normalized cDNA libraries for each of three potato and six tomato genotypes for SNP discovery.
- Obj. 5** *Establish centralized facilities for genotyping a core set of SNP markers in standard germplasm panels in tomato and potato.* We will develop a database of integrated and mapped markers and genotypes for at least 480 accessions for each crop. We anticipate >85% success for 7,600 potato SNPs and 1,536 tomato SNPs.
- Obj. 6** *Address regional, individual program and emerging needs within the Solanaceae community through a small grants program.* To increase breeder engagement with genomics, we will provide opportunities to genotype up to 16 populations and validate marker linkages to major QTL.
- Obj. 7** *Create integrated, breeder-focused resources for genotypic and phenotypic analysis by leveraging existing databases and resources at SGN.* We will create integrated genomic and phenotypic databases, the breeder's toolbox, that serves the entire Solanaceae breeding and genetics community through SGN.

BACKGROUND

The Solanaceae is a diverse family containing economically important species, including two vegetable crops that collectively provide substantial amounts of starch, sugars, vitamins, and antioxidants to our diet. Potato (*Solanum tuberosum* subsp. *tuberosum*) and tomato (*S. lycopersicum*) annually account for \$6.3 billion in farm value and are grown on 1.5 M acres in the US (URL 9). Through SolCAP, we can exploit homology within genomes and positively impact applied breeding and crop quality across traditional commodity boundaries.

Tomato. Tomato is a diploid ($2n = 24$) and has provided a reference genetic map against which both potato and other Solanaceae maps have been aligned (21, 52, 53, 63, 64). Extensive breeding efforts exist in the private sector, yet eight University programs continue to develop relevant commercial varieties and germplasm. Private efforts span large and small vegetable companies with comprehensive programs across market niches. In the public sector, fresh-market programs in New York, Pennsylvania, North Carolina, Oregon, and Florida span trait development, inbred parent, and hybrid development. Efforts in Ohio, New York, Pennsylvania, and California focus on trait discovery, pre-breeding, parent development, and hybrid development for the processing market.

Potato. The most widely cultivated potato species, *S. tuberosum* subsp. *tuberosum*, is an autotetraploid ($2n=4x=48$) and is the world's most important dicot food crop ranking fourth in overall production, after rice, wheat, and maize. Unlike many other crops, over 90% of potato breeding is still conducted in the US public sector. There are 11 public sector and one private (Frito-Lay) breeding programs, along with four USDA/ARS geneticists focused on germplasm enhancement, taxonomy, and curation of germplasm. The public programs have a strong history of partnering with regional industry to develop and release varieties for the processing (60% market share) and fresh markets. Breeding activities are directed towards improved germplasm with biotic resistances combined with improved quality and nutritional components. Breeders in the US also have a history of accessing germplasm from the National Plant Germplasm System (NPGS, NRSP-6) for these traits.

The most critical need in the Solanaceae breeding and genetics community is for a sufficient number of markers that are polymorphic in elite germplasm. Currently, the

predominant use of marker-assisted selection (MAS) in the Solanaceae is to select for desirable wild species alleles during backcrossing. Recently, more sophisticated MAS applications in tomato have been used to reduce introgression size (7) and to select for coupling phase recombinants for disease resistance (66). In potato, mapping studies (at the 2x and occasionally 4x levels) have been conducted since the late 1980's, but MAS is not yet practiced in varietal breeding in the US. The traits of most interest to growers and processors remain genetically uncharacterized. As a result, only a few markers linked to simply inherited traits are being used in commercial programs. The potential for MAS to analyze whole genomes to improve estimation of genetic variance, recover elite backgrounds, and select superior varieties has not been fully realized in Solanaceae crops (27, 30).

Solanaceae resources are significant. Leveraging genomic resources across species is the basis of SolCAP. Over the past six years, numerous sequence and functional genomics resources have been generated for the Solanaceae. The SOL Genomics Network (SGN; URL 17) is a community driven, Solanaceae-specific database integrating map, EST as well as genome sequence, genotyping, locus annotation, and phenotypic information. Large collections of ESTs are available for tomato (259,421 ESTs; URL 4) including an on-going tomato full-length cDNA sequencing effort in Japan (URL 8). Despite these resources, there remains a need to develop translational tools. Only 106 validated and mapped SNPs have been developed by mining EST resources (29, 65). One reason for the disparity between sequence resources and validated markers for breeding populations is that current resources lack both breadth of varieties and focus on relevant populations. The potential of a broadly focused translational effort is demonstrated by the recent release of 1487 new SNPs representing 302 loci in tomato with 579 of the new SNPs representing 162 polymorphic loci in breeding germplasm (56). Additional resources are emerging from the tomato genome sequencing initiative which is focused on sequencing euchromatic regions of Heinz 1706 using a bacterial artificial chromosome (BAC) - by-BAC approach combined with next generation sequencing. To date, 1092 BAC sequences have been generated by this initiative (URL 15). Several TILLING populations are in development for tomato functional genomics and discovery of novel alleles including a newly funded USDA project (URL 2).

For potato, 236,547 ESTs (16, 44) are available (URL 4). These ESTs were primarily generated from three cultivars: "Kennebec", "Shepody", and "Bintje". In addition, an international consortium is sequencing the entire potato genome of the heterozygous diploid line RH89-039-16 (60) and the *S. tuberosum* group *phureja* doubled monoploid clone, DM1-3 516R44. We expect SNP discovery to be straightforward because of the high level of sequence diversity within potato. Previous deep-sequencing has revealed multiple haplotypes within potato genes. For example, 3-5 common haplotypes were reported at the disease resistance locus *Stve1* (47), the dihydroflavonol 4-reductase locus (12, 13), and acid invertase *invGE* (32).

Comparative maps exist between tomato and potato, pepper and eggplant (34, 51, 60, 63, 64) which will allow these resources to be exploited across species and guide marker discovery with application to breeding programs and genetic studies. Buell previously identified simple sequence repeats from 12 Solanaceae species including tomato and potato and designed primers for use by the community (URL 22). Buell's data mining, performed under the SolCAP pre-award year, has also identified putative SNPs (eSNPs) in the current potato EST collection; from the Sanger-derived ESTs, 9,620 high confidence candidate SNPs (> 4x coverage of alternate allele, no other SNP within 25 bp) could be identified in 2,332 contigs (URL 23). For the Sanger-derived tomato ESTs, 1,849 high confidence SNP positions in 903 contigs were identified. We

benchmarked our computational pipeline by determining how many of the known, validated SNPs from tomato (Francis, D. pers. comm.) were present in our eSNP output. Using a set of 132 tomato sequences that were present within 77 contigs, we were able to identify 32 contigs with 92 eSNPs in our high confidence tomato SNP data set, suggesting a sensitivity of 42% (32/77). Lack of detection of all validated SNPs can be attributed to our high threshold for calling a SNP; the validated SNP occurs within an intron rather than the coding sequence; and/or the lack of the alternate allele in the public EST data set as not all validated SNPs were designed using only publicly available EST data.

While this type of data-mining is productive in potato, unfortunately, the majority of putative polymorphisms from tomato are of limited use as a number of the ESTs in the public tomato EST collection are derived from Micro-Tom, a tomato with seven wild-species in its pedigree. Thus, few data-mined tomato SNPs are useful within breeding populations. To identify polymorphisms among elite tomato germplasm, we developed a bioinformatic pipeline to identify SNPs between genomic sequence and assembled ESTs. Using the tomato TA496 ESTs in Genbank (108,685 ESTs), we cleaned and then clustered and assembled these sequences into 12,349 contigs. We then aligned the TA496 EST contigs to the emerging tomato genome sequence derived from Heinz 1706. Thirty-one percent (3,831) of the TA496 contigs aligned with the Heinz 1706 genomic sequence. Within these contigs, 86 high confidence SNPs were predicted within 32 contigs; an eSNP discovery rate of < 1% with 2-3 SNPs per contig.

SolCAP Community. We have created an open community consisting of applied and basic researchers working on crops in the Solanaceae. Our stakeholder involvement was extensive, including ten organizational meetings over the last three years. We engaged international scientific efforts such as EUSol (URL 5; *see letter of support*) and the international genome sequencing projects for tomato and potato. Publicized open workshops were held in 2005 in July at the ASHS meeting, in Las Vegas and in November at UC-Davis. The latter included 70 participants from the vegetable seed industry, public and private breeders, processors, NPGS staff, and grower groups. Four meetings were held in conjunction with the PAG conference (2006-2009). In addition, a SolCAP translational genomics section was held in conjunction with the international Solanaceae Genome Workshop in 2007. Commodity-specific input was provided through the tomato breeder's roundtable (2006, 2007) and the Potato Association of America (2005-2007) meetings. Summaries from major meetings were posted on the web and published (58) to facilitate community discussion. Priorities were systematically developed for traits and tools that would benefit more than one commodity in the Solanaceae (57). Needs assessment placed priority on candidate gene approaches that would facilitate future breeding efforts for these traits. Crop-specific traits will be best addressed indirectly, in part through genotyping a diverse array of germplasm and small grants. Further refinement of SolCAP objectives and deliverables were made through scientific and stakeholder advisory board recommendations and also through the eXtension.org workshop held at the January 2009 PAG.

Traits. Community discussions lead to a consensus that consumer and processor quality traits would provide strong and mutually beneficial cross-commodity interaction (*see letters of support*). Sugar, carbohydrate (CHO)-related phenotypes, and vitamin content represent high impact traits common to both tomato and potato. ***CHO and sugar metabolic pathways are well defined.*** At the genetic and biochemical levels, the pathways of potato and tomato carbohydrate synthesis and breakdown are moderately well characterized (5, 22, 49). Numerous QTL for starch and sugar content have been identified in potato and tomato (5, 15, 37, 45), and the map

locations of many QTL overlap with known carbohydrate metabolism genes (6, 37). Many of these have been shown to co-segregate in tomato (17, 18, 32) and potato with allelic and expression variation (3, 31, 52). These studies validate an integrated candidate gene and whole genome approach for trait dissection.

RATIONALE and SIGNIFICANCE

SolCAP will allow the US Solanaceae community to pursue marker-based breeding by delivering integrated marker tools and breeder-friendly databases. SolCAP leverages operational and scientific strengths of the participating scientists and their institutions across Solanaceae crops. Potato and tomato are in need of tools for genotyping in elite germplasm. There are practical limitations to the homology-based approach in that SNP markers must be specific to a base change in the context of a gene. Thus, SNPs will be unique to species and even crosses within species. However, there is immense potential to leverage information as well as database structure and content within the two crops to achieve a level of efficiency when characterizing and/or manipulating traits shared across species. Our approach will focus on comparative analysis of variation in specific CHO, sugar, and vitamin metabolic pathways and will emphasize deep resources within each crop as well as cross-species comparisons (URL 1). Working together allows us to maximize our collective impact by cost-effectively sharing many resources including SNP genotyping, bioinformatics, databases, education, and extension. *These improved resources and tools are being tied to directed education and extension programs for current and future breeders.* We will help coordinate education of industry and stakeholders across all plant translational genomics CAPs through a web-based CoP called the Plant Breeding and Genomics Community (PBGCoP). We envision SolCAP providing the tools and training for breeders to link traits with markers, so that Solanaceae crop varieties can be developed more rapidly, precisely, and efficiently.

RESEARCH METHODS

Research, education, and extension logic models are described in the logic model PDF.

Obj. 1. Create an education program to train graduate students in genome-based breeding. (De Jong, Francis, Douches, Van Deynze, Liedl)

Graduate students. Our first goal is to train graduate students who can actively conduct genomics-assisted breeding. We are currently developing a team-taught distance education course. The course will consist of modular, synchronized video, voice and powerpoint/chalkboard presentations, and carefully designed problem sets to teach and assess progress. The course will be made available without charge through the Cornell Transnational Learning Website (URL 3). CTL has extensive experience in developing and distributing plant breeding-related courses for a wide audience through streaming video. The course will be taught once a year and students will interact live with instructors on a weekly basis using text and voice features of Adobe Acrobat Connect. Graduate students at instructor institutions will be permitted to take the course for credit; others students seeking completion certificates will be offered extension or continuing education credits. Learning objectives for the course will be developed through extensive consultation between Douches, Buell, Van Deynze, Francis, and De Jong, with the latter taking the overall lead. Desired learning outcomes for students will include ability to: 1) describe current marker technologies; 2) describe appropriate marker use for indirect selection in breeding; 3) develop markers linked to any trait of interest; 4) explain limitations of markers; 5) navigate genome databases, including Solanaceae, to address breeding problems; 6) utilize analysis pipelines related to the collection and analysis of both phenotypic data and marker-trait

associations using SolCAP resources; 7) master key breeding/genetics concepts, such as the relationship between estimates of heritability and estimates of the strength of a QTL, and comparing trait-based to marker-based selection. Course assessment and objectives will be modified with input from the Cornell Center for Teaching Excellence, feedback from students, and technological and conceptual changes in the discipline. Suitable members of Cornell plant breeding graduates will be recruited to help develop the curriculum and in the process, internalize many principles of translational genomics themselves.

Specific outcomes and Timetable: We will educate graduate students in the theory and practice of translational genomics for vegetable crops. This objective will be achieved by developing and delivering curricula through a team-taught distance education course across four universities. ***Year 1:*** Graduate student curricula will be taught as a live pilot course at Cornell University. Course material will be added to the SolCAP website. ***Years 2 and 3:*** A course will be added to the plant breeding curriculum at the participating institutions as a live course. Course material will be updated in subsequent years, integrating data and resources developed by SolCAP. In the third year, the course will be team-taught as a distance learning course.

Obj. 2. Amplify outreach efforts by developing an eXtension Plant Breeding Community of Practice (CoP) to develop continuing education material aimed at practicing plant breeders, their staff and seed industry professionals. (Stone, Francis, Van Deynze, Liedl and De Jong)

SolCAP is creating an eXtension Community of Practice (CoP) that fosters cross-disciplinary and cross-commodity outreach (Plant Breeding and Genomics CoP, PBGCoP). We are using web 2.0 resources to exchange research- and experience-based objective information as well as communicate across disciplines and commodities beyond SolCAP. The CoP is collaboratively generating outreach materials that will be published on the public eXtension.org site. SolCAP's target audience, the **Community of Interest (CoI)**, is plant breeders, their staff, and allied professionals and our goal is to help this community turn research into practice. Currently, within the Solanaceae community, many breeding programs lack knowledge and access to technology, and lack experience with genomics tools and analysis techniques to facilitate applied outcomes. The CoP will address this need. We are working in a Drupal powered **collaborative workspace** (PBGWorks; URL 11) that includes functionalities to foster long distance networking and collaboration. The workspace supports private/public small group spaces, collaborative authorship, image sharing, forums/discussions and blogging, web-conferencing, and tools for workflow management including peer review. The site also includes data portability; content can be electronically 'fed' for publication on other sites, including www.eXtension.org.

Leadership and support for conceptual/scientific aspects eXtension content development will come from Co-PD David Francis and the executive committee (EC). Francis will supervise a dedicated coordinator (program manager) to manage content development workflow, including peer review and evaluation by the project assessor. Co-PD Stone will supervise technical support of the Drupal workspace through Oregon State University's Academic Publishing Technology Group (APT Group). John McQueen, the PBGWorks technical coordinator, is also the technical coordinator for the eOrganic workspace. Dr. Debby Lewis, Ohio State University, will facilitate development of the CoP in years 1 and 2. Dr. Lewis is an expert in Extension program development and evaluation. She and Dr. Namuth will represent eXtension on the stakeholder group to guide content development. To develop the vision and functionality of the CoP, SolCAP organized a telephone conference (Sept. 2008) with USDA and eXtension in order to clarify goals and then organized a workshop with other plant translational genomics CAPs at the

January 2009 PAG Conference where we invited these CAPs to participate in the PBGCoP. Groups planning new plant translational genomics CAPs are using the workspace as well. With input from these other groups, the CoP structure was modified. The CoP application is in process and a mechanism by which new projects can voluntarily join the CoP was established (URL 11).

Direct interaction with SolCAP's CoI: SolCAP will develop workshops to be held in conjunction with national and regional breeder meetings. Pre- and post-workshop surveys will be conducted to assess baseline and new knowledge, desired knowledge and effective curricula delivery formats. During these workshops we will build on curriculum developed through PBGWorks and refine it for publication to eXtension.org. By targeting breeders through workshops we can increase both their experience and access to newly developed tools. Learning objectives (developed from stakeholder meetings) include: understanding marker detection platforms; developing skills to interpret data (mapping, analyzing population structure, QTL detection, and statistical analysis); and database usage and mining. Discussion of population structures, sizes and understanding the issues and opportunities associated with MAS will be incorporated. Hands-on sessions using tools from SGN will be an integral part of the training. Workshops will be managed by Van Deynze, De Jong, Liedl (WVSU) and Sue DiTomaso and tied to annual conferences (such as Potato Association of America, regional breeders meetings, Tomato Breeder's Roundtable) and UC-Davis to maximize the opportunity for the target audience to participate. The UC-Davis courses are independent of the more general SBC courses currently offered (URL 28). Week-long internships to provide practical experience in using markers and analyzing data will be offered for breeders and their support staff in SolCAP labs in the final two years of the project. Lesson materials developed specifically to accomplish workshop and internship learning objectives will be made available through eXtension.

Specific outcomes and Timetable: Our CoI will gain knowledge and ability to use genomic information in the context of applied plant breeding programs. The CoI will also have improved access to genomic data through the breeder's toolbox (see Obj. 7). At the same time, genome scientists will gain a better understanding of breeding through engagement. Assessment will be measured through surveys conducted by Dr. Coe. **Year 1:** Additional CAPs will be added to the CoP. SolCAP eXtension content will include 12 featured articles, 100 FAQs, and 6 resource tutorials. These will be integrated through learning lessons into a single "start-to-finish" example of marker-assisted selection (MAS) focused on a high-heritability trait. We will seek "eXtension certification" for the CoP. Three workshops will be conducted for plant breeders and staff. **Year 2:** Additional public content will be launched, including 18 featured articles. PBG content on eXtension.org will include contributions from other CAPs as well as news and interactive elements. Curricula will be expanded through learning lessons and database tutorials that encompass more complicated examples of MAS for quantitative trait loci. SolCAP will provide week-long breeder training in SolCAP labs. Two workshops will be conducted for plant breeders and staff. **Year 3:** PBG content published on eXtension.org will include 24 featured articles, and content will be evaluated for its utility to agricultural professionals. The PBGWorks collaborative workspace will be evaluated annually for efficacy in fostering cross-disciplinary and cross-commodity networking and collaboration, as well as interaction amongst breeders, geneticists, and agricultural professionals. Workshops (3) are planned for breeders and staff.

Obj. 3. Collect standardized phenotypic data across multiple environments for tomato and potato. (Francis, Douches and De Jong)

Germplasm panels form the core of SolCAP, linking all objectives. Each panel was assembled with extensive community input, including basic researchers, applied breeders, USDA Crop Germplasm Committees, and key germplasm repositories.

Tomato germplasm was assembled to facilitate the applied objectives of SolCAP while providing a basis for historical comparisons and assessment of the role of introgression in contributing to modern varieties. A comparison of patterns of polymorphism in lineages defined by market niche or regional adaptation will lead to insight concerning selection during domestication and breeding (1) and allow breeders to access variation found within elite material. In assembling this collection, we are drawing from community core collections for *S. lycopersicum* var. *Cerasiformae* (*Slc*), and other members of *Solanum* sect. *Lycopersicon* (Tanksley, pers. comm.; (URL 25)). We include landraces (LR) representing geographic range, diversity of fruit shapes, and genetic variation (26, 43, 62, Labate, pers. comm.). Heirloom (H) varieties have been selected such that the major fruit morphologies are represented as well as selections for flavor. The elite germplasm representing freshmarket (FM), 144 accessions, and processing (P), 140 accessions, forms the most important resource in the collection. Varieties were chosen to represent major environmental adaptations in the US (humid and arid) and public breeding programs. Public germplasm includes parents for multiple mapping populations, inbred lines from commercial companies, and inbred lines generated by self-pollinating commercial hybrids. Genetic diversity was also considered in assembling the collection of modern varieties (40, URL 26). Accessions selected for these studies will complement efforts of the EUSol project (*see letter from D. Zamir*).

Potato germplasm is focused on specific market classes (russet frozen processing, and round white chip processing) and will include a broad genetic base. We incorporated pedigree information, which has been well documented for North American cultivars since the mid-1800s. A subset of the panel will comprise lines with resistance to sugar accumulation at low temperatures. All active US breeding programs are contributing varieties and advanced lines of value to the public breeding community (*see letters of support*) and a small selection of international germplasm is included for comparative analyses. The collection includes the leading US varieties, clones of historical significance and introgressions from other *Solanum* species, and exhibits substantial variation in processing ability, starch content, appearance, and resistance to major diseases. A core set of *Solanum* species and accessions (provided by D. S. Spooner, USDA/ARS *Solanum* taxonomist) that have been introgressed into tetraploid germplasm are included in the panel to provide a taxonomic perspective. Genetic stocks that will link the SNP development work to the potato genome sequencing projects are included. A tetraploid mapping population from a russet cross (primary market type for French fry processing), segregating for vitamin C and reducing sugar accumulation, will also be evaluated, as will a small diploid reference population to genetically map SNPs.

Traits. The phenotyping for agronomic and quality traits will be conducted by cooperating breeders. Germplasm coordinators and cooperators have developed a “cooperators guide” to standardize data collection for both field performance and fruit/tuber quality. The guiding principles will be to collect standardized objective data, to use consistent language, and order traits into ontologies where appropriate. Each cooperator will also evaluate germplasm for traits specific to their breeding program (e.g. resistance and quality traits). Buell is assisting in phenotypic database design, descriptor language and rating scales. Final formatted data will be

posted on SolCAP and SGN. For tomato UPOV/IPGR descriptors will be collected to describe plant habit, inflorescence structure, leaf shape. Fruit size and shape data will be collected from scanned images, as will color (L^* , a^* , b^* , Hue, Chroma) and color uniformity data, using TomatoAnalyzer software (4, 11, 23). Standard measurements of total soluble solids, pH, and citric acid will also be collected. These phenotype ontologies have all ready been integrated into SGN (URL 16). For potato standard measures will include glucose, sucrose, chip (or fry) color and specific gravity (an indirect measure of starch content). Other measurements will include vitamin C, total yield, and vine maturity along with Plant Variety Protection (PVP) descriptors such as flower color, vine type, tuber skin color, tuber flesh color, tuber pigment pattern, tuber shape, skin brightness, and tuber eye depth. Sugar and vitamin HPLC measurements will be performed at the University of Wisc., Madison. Individual sites will score opportunistic disease and quality data such as frequency of heat sprouts, internal heat necrosis, reaction to common scab, etc.

Field Evaluation of Germplasm Panels in each crop will occur over at least two years and two environments per year with a randomized complete block design. Wild germplasm will not be included in the field trials although cooperators have evaluated sub-populations for specific traits (e.g. fruit quality assessment performed at Oregon State University for *Slc* accessions). Tomato evaluations will be performed at Campbell's R & D (Processing germplasm) with assistance from Chetelat and Van Deynze; at UC-Davis (fresh market collection); at OSU (Francis; entire collection) and UFL (Scott; FM, H, LR). De Jong and Novy (USDA/ARS, ID) will provide clean seed of the clonal potato material and tetraploid mapping population, respectively, for the replicated field tests. Thill (UMN) and Yencho (NCSU) will provide additional sites for mapping population evaluations, while Jansky (USDA/ARS, WI) and Vales (Oregon State) will evaluate clonal material. Haynes (USDA/ARS, MD) will provide experimental design, data management and analysis of the phenotypic data for potato.

Pre-analysis and quality control QC of phenotypic data. The evaluation of phenotypic data will be conducted by Francis for tomato and Douches and Haynes for potato. This QC will include an analysis of data distribution and appropriate transformations or normalization. Inconsistency among replicates, spatial variation and other anomalies within trials will be accounted for to the degree possible, through increased replication of check lines. The use of lsmeans and Best Linear Unbiased Predictions (BLUPs) will be introduced as appropriate. As many of these analyses are routine within the scope of our breeding programs, existing expertise and analysis pipelines are in place. QC of phenotypic data will also include the assessment of heritability and of G x E and environmental variation.

Specific Outcomes and Timetable: phenotypic data, described above, for core germplasm sets in tomato and potato will be integrated into the SGN database. The germplasm panels will serve as a long-term breeding community resource. Breeders will be able to access data from these panels to assess parents of populations already developed for marker polymorphism as well as to facilitate the design of future crosses by allowing identification of parents with desirable combinations of marker and trait variation. **Year 1:** Potato seed multiplication; potato and tomato field evaluations by breeder-cooperators. Standardized phenotypic data collected by all cooperators, quality control and analysis. **Year 2:** Seed multiplication; field evaluations by breeder-cooperators. Phenotypic data uploaded to the SolCAP website and made available to SGN. **Year 3:** Final phenotypic data uploaded to the SolCAP website and made available to SGN.

Obj. 4. Develop extensive sequence data of expressed genes, and identify Single Nucleotide Polymorphisms (SNPs) markers distributed across the genome and associated with specific candidate genes for sugar, carbohydrate and vitamin biosynthetic pathways. (Francis, Van Deynze, Buell, De Jong)

Generation of sequence data. This objective is being achieved by sequencing normalized cDNA populations from six tomato and three potato cultivars, breeding lines, and accessions using the Illumina Genome Analyzer (GA). The sequence data will be combined with existing sequence resources to computationally identify polymorphisms. This approach is cost efficient, amenable to bioinformatic analyses and provides unbiased discovery of genes and markers for any trait.

Sequencing in tomato will focus on one processing tomato variety which is genetically distinct from existing EST and genomic sequence available for processing lines E6203 (TA496) and H1706; three genetically distant fresh market varieties (from FL, NC, and CA), a land-race variety, and a *S. pimpinellifolium* accession. The purpose of this last entry will be to gain insight into the number of novel alleles that we might discover among closely related wild relatives. We believe that this approach will maximize SNPs that are useful across populations. The three potato varieties selected (Atlantic, Snowden and Premier Russet) represent current fresh market and processing types.

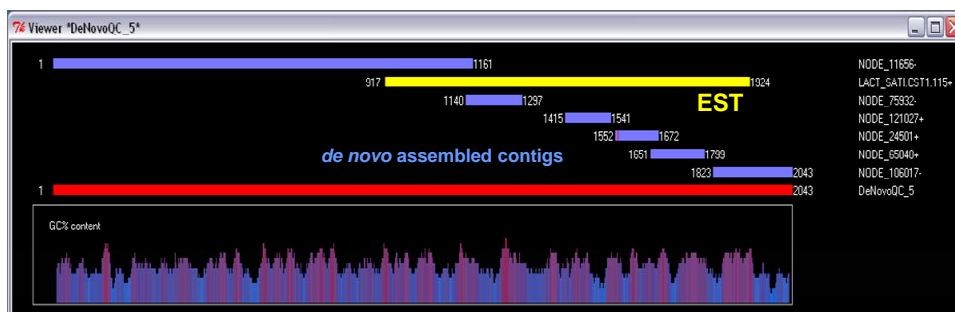


Figure 1. Example of assembly of 25 million reads 24-60 bp, resulting in 70,000 lettuce contigs of 120 bp or longer. The blue lines indicate GA reads that cover and extend existing Sanger unigenes. (A. Kozik and R. Michelmore, UC-Davis).

We are using Evrogen-based technology (URL 6) in which re-association kinetics and double-stranded nuclease (DSN) is used to isolate full-length sequences to normalize the cDNA population across tissues. At UC-Davis, GA throughput approaches 7-15 Gb per run in single or paired-read format, with the vast majority of sequences between 35-80 bp (56). Our aim is to sample polymorphisms in as many genes as possible with 2 Gb coverage in paired-end reads (2 x 80 bp) for each variety. This level of coverage in normalized cDNA libraries has yielded 28-40 fold redundancy in nine cotton (tetraploid) genomes (36, Van Deynze unpublished) giving very high confidence in SNP calls, even in polyploids. Using established pipelines in the Buell lab, sequences will be trimmed and mapped to reference transcript and genomic assemblies using assembly and variant detection software including MAQ, Velvet, and Bowtie/TopHat (33, 69, URL 27). Sequences will be mapped to the emerging genome sequence for both potato and tomato. We will also utilize Sanger-derived EST assemblies as a reference sequence and construct accession-specific EST assemblies for potato and tomato to more readily identify SNPs. Paired-end reads will aid in *de novo* assemblies and contig resolution, especially in autotetraploid potato. *De novo* assemblies have shown to add to both gene discovery and SNP discovery over simple mapping to known unigenes (Fig. 1)

Identifying candidate genes. The primary focus of our efforts will be genes involved in the metabolism, catabolism, and transport of simple sugars, carbohydrates and vitamins. For CHO in potato and tomato, we will focus on all genes and QTL related to sugar and starch content previously identified and mapped (5, 6, 14, 37, 45), as well as genes closely associated with starch synthesis, starch degradation, and hexogenesis (URL 1). For tomato, 63 sequences related to CHO have been previously mapped in wide crosses (5) as well as 56 sequences associated with metabolic and transcription profiles (2). Gene orthologs will be mined from these and related species such as Arabidopsis which has biochemical mappings available through AraCyc (URL 20). We anticipate that our candidate gene search will expand to include genes involved in transport and regulation. We will utilize existing bioinformatic pipelines in the Buell Lab at MSU that utilize assignment of Gene Ontologies and KEGG pathways to identify genes of interest for CHO, sugar, and vitamin (including vitamin c and pro-vitamin A) production and/or regulation. For genes that are part of multi-gene families, priority will be given to members expressed in potato tubers or tomato fruit. Other candidate genes are being solicited from the community using a web-based survey (<http://www.solcap.msu.edu/candidates.shtml>).

Expected results. We estimate that 25,000-35,000 genes (50-75% of tomato unigenes totaling 25-35 Mb) (55) with an average size of 1 kb (based on contig length in assembled tomato ESTs) will be present in our cDNA populations. SNPs occur one out of 7,500-8,500 bp in exons of cultivated tomato (29, 39, 65), but are not distributed evenly across genes (65). Taking into account tomato SNP rate and distribution, we anticipate identifying 2,900 to 4,600 new SNPs in any single pair-wise comparison between varieties. Existing SNP data suggest that as many as 75% will be redundant in subsequent comparisons. Based on relative SNP rates in exons reported for potato (1/175 bp) (19, URL 14), we anticipate detecting 142,000 to 200,000 SNPs in potato based on a single pair-wise comparison.

All sequences will be deposited in the NCBI Short Read Archive (URL 10) and made available through the project website. All high confidence SNPs will be made available through the project website and SGN. The paired-end read approach will help reduce mis-assemblies due to paralogs. We expect to characterize allelic variation within multigene families by analysis of haplotypes afforded by in-depth analysis of *de novo* sequence data relative to established unigenes and contigs.

Timeline and Specific Outcomes: Normalized cDNA libraries and 2 Gb sequence for each of 3 potato and 6 tomato genotypes will be completed in 2009; SNPs discovery will be completed in 2009. After filtering SNPs based on sequence quality and context we expect the minimum number of high quality SNPs to design Illumina assays for 1,536 and 7,600 for tomato and potato, respectively.

Objective 5. Establish centralized facilities for genotyping a core set of SNP markers in standard germplasm panels in tomato and potato. (Van Deynze and Francis)

We will utilize the tomato and potato SNP data generated under Objective 4 and SNPs mined from public sources (29, 56, 54, 59, 65; SCRI) to design our arrays. We will genotype the core collections of 480 accessions for each crop using the 1,536 probe Illumina design in tomato and the Illumina Infinium 7,600 probe design for potato. An additional 672 genotypes will be run for potato under the small grants program. Genotype data quality assessment, collation, and release are discussed in detail in Objective 7.

Polymorphism rates are sufficiently high for potato (URL 14) that we will be maximizing coverage of candidate genes and including multiple probes per gene so that haplotype information can be deduced from SNP genotyping. We propose that up to 5 SNPs be associated

with each candidate gene in potato. We envision genotyping over 150 candidate genes from the CHO, sugar and vitamin biosynthetic pathways along with 250 other genes suggested by the research community. Even with a large number of candidate genes, we anticipate that well over 75% of the SNPs will come from randomly selected genes, thus ensuring genome wide coverage. In tomato, our expectations are that fewer candidate genes will be polymorphic (28, 56, 65). Buell has been identifying candidate tomato SNPs (eSNPs) by analyzing existing genome and EST sequence data. Validation studies are being conducted with a sample of eSNPs to determine the criteria for successful calling prior to construction of the 1,536 array.

For potato, SNPs will initially be mapped in an existing diploid reference population (n=94). At the tetraploid level, power to map traits will be considerably enhanced if SNP data allow us to deduce allele dosage and haplotypes with accuracy. Both the Illumina and Luminex platforms will allow us to deduce haplotype and, in at least some instances, dosage, as is achieved with TaqMan (12). Preliminary data derived from sequencing sucrose phosphate synthase, granule bound starch synthase, and four ESTs from 10 potato cultivars demonstrates the feasibility of representing haplotypes by a subset of SNPs. For example, 3/10 SNPs represented five haplotypes in a 260 bp EST (De Jong, unpublished).

For tomato, determining the genetic map locations of SNPs will be addressed primarily through BLAST to emerging genome sequence because of the low SNP polymorphism rates. Mapping of SNPs in tomato will be costly if we rely on bi-parental crosses, as only a fraction of markers will segregate in any given cross. Genotyping and mapping community populations will be most cost-efficient with a two-tiered approach. Fixed-array technology (Illumina Golden Gate or Infinium, UC-Davis) will be used to genotype the core collections. Liquid bead assays (Luminex or Illumina BeadXpress (OSU)) will provide flexibility and cost effectiveness for individual populations where only a small proportion of the fixed-array SNPs are polymorphic. To assess the feasibility of positioning SNPs using a BLAST search, we assessed the ability to map 325 loci using available sequence as of 3/1/2009 (estimated 39% of euchromatic sequence). We were able to find high quality hits to 33% (>95% homology over >90% of clone length) of the loci, with 15% matching mapped BACs. We estimate that 80% of the SNPs will be mapped relative to euchromatic BACs sequenced through the current tomato genome effort. It is highly probable that we will be able to map over 1000 SNPs. A recently funded joint Dutch/Indian effort will be generating next-generation shotgun sequence for the cultivar M82 and the wild tomato accession, LA716 (*S. pennellii*). As these resources become available, a higher percentage of SNPs will be localized to genome sequence.

Application of Genotyping. As population structure and linkage disequilibrium (LD) is not yet defined in either potato or tomato, it is not clear what the optimal population is for association mapping. Our genotyping will represent almost all US breeding programs, and we recognize that the utility of association analysis for any one breeding program might be limited. A direct outcome of our germplasm survey will be to visualize haplotype blocks within and among subpopulations, estimates of LD decay and a description of population structure. Estimates of population subdivision and kinship can be incorporated into mixed-model analysis of variance to test for associations between markers and traits (9, 24, 42, 67). We are experienced at estimating population structure using pedigree data to predict identity by descent (IBD) (e.g., 8), Bayesian models of Hardy-Weinberg equilibrium as implemented in STRUCTURE (42), and principal components analysis (41).

Sequence data across several hundred genes in ~100 cultivated tomato accessions (35) suggest that recombination is low and LD decay occurs over 5-10 cM rather than Kb (Francis,

unpublished). These preliminary data from tomato also suggest population structure between lineages representing different market classes and within market classes due to environmental adaptation. Gametic phase disequilibrium also occurs within tomato breeding programs and market classes, and is almost always associated with loci that have been fixed within individual breeding programs. These data also suggest that tomato breeding programs have very small effective population sizes. We recognize limitations of association mapping given these constraints, and a probable outcome for SolCAP genotyping will be to focus on parents selected as diverse founders for nested association mapping strategies (68). It is also likely that introgression approaches will continue, and the data generated under Objective 5 will provide a means to ensure that this approach is systematic. Thus, benefits beyond QTL discovery through association mapping including benefits to parent selection, population development, and allele mining, are an expected outcome of the proposed research.

Potato, being asexually propagated, has undergone few meioses since introduction in the mid-1850s. The decay of LD in modern *S. tuberosum* is in the 3-8 cM range (10). If this is the case for our population, then a small number of markers may be sufficient for genome-wide association studies, although resolving power will be limited. Consistent with this, Simko et al. (46) reported that while LD decays rapidly over distances of a Kb, it does not decay to background over the length of a BAC clone. A mixed-model approach, taking into account pedigree information, and using resistance gene analogs as markers, has been successfully used for association mapping of late blight resistance loci in potato (35). There is little evidence of population structure, and SNPs are shared across market classes (10). Finding potato SNPs will not be difficult. The most important goal, assuming that population size is not a restriction, is having sufficient marker density so that we can elucidate valuable QTL.

For both potato and tomato we expect to detect QTL that contribute 5-7% of the proportional variance when traits are independent of market classes ($n = 288$, $P < 0.05$). We do not expect to detect QTL within a market-class unless they contribute 10-15% of the variance. Raising the probability threshold to $P < 0.001$ will require that a QTL detected across cultivated genotypes account for ~10% of the variance, while a QTL detected within market-class must account for 15-20% of the variance. We believe that an important outcome of the SolCAP genotyping will not only be in the potential to discover marker-trait associations, but also the potential to stress approaches to breeding (breeding systems) that include estimation of breeding value for individual lines, for populations, and for pedigrees. Such approaches are not widely used or even discussed in vegetable crop breeding (25).

Timetable and Specific Outcomes: With an anticipated SNP validation and assay success rate of $\geq 85\%$, we will deliver 627,000 data points in tomato and 7.4M data points in potato. Of the 1,536 SNPs assayed in tomato and 7,600 SNPs in potato, with up to 25% be from candidate genes. The position of about 1000 SNPs will be determined for tomato relative to emerging sequence and 4,000 for potato in the 2x mapping population. QTL analysis in the 4x potato mapping population should allow discovery of loci affecting $>10\%$ of the variation for a trait. Analysis of the core collections will define population structure, define LD decay, identify cryptic introgressions and describe patterns of variation and allelic diversity within and among market classes. We will complete genotyping of the tomato and potato collections in **year 1** and add genotypes from specific mapping populations to both crops in **years 2 and 3** through the small grants program.

Objective 6. Address regional, individual program and emerging needs within the Solanaceae community through a small grants program. (Executive Committee)

To increase breeder engagement with genomics, we will provide opportunities to genotype up to 20 populations and validate marker linkages to major QTL. This program provides a mechanism to reallocate resources and implement community-based strategies for validation and MAS. Small grant funds will be used primarily to genotype populations provided by breeders, thus leveraging SolCAP-developed SNPs to map traits (e.g., agronomic, disease resistance, insect resistance and quality) not addressed in our core germplasm panels. When requested, small grant funds will also be used to facilitate marker conversion in order to empower individual programs to pursue specific traits or loci and/or validate candidate QTL identified in our core germplasm panel. Many in the solanaceous community have begun to develop populations for SolCAP to genotype; where we know of such populations, we will genotype the parents at the same time we genotype the germplasm panels. We have set aside funds specifically to provide genotyping service in years 1, 2 and 3. Note that *due to the nature of the genotyping platforms, the funds for genotyping populations from the community are built into the UC-Davis and Ohio State University budgets in year 1*. Based upon current genotyping costs, we anticipate being able to genotype up to 16 potato and tomato populations through the small grants program. Populations submitted by the community following a call for proposals will be awarded on a competitive basis (*see Management Plan*). Priority will be given to populations with existing replicated phenotypic data and completed statistical analyses. During the application review process we will consider population size, distribution of the phenotypic data, replication, and coefficient of variance for traits. Data from initial genotyping under Objective 3 will be used to select the appropriate platform for mapping any population. Polymorphism rate and number of individuals in a population will be taken into account to choose platforms for SNP mapping.

Fragment length polymorphisms (FLPs) continue to be the primary assays employed in small breeding labs. Although single nucleotide extension and ligation assays have become the method of choice for genotyping SNPs, we (Francis and Van Deynze, unpubl.) enjoy a high frequency of success in converting SNPs into markers that are easily assayed on agarose gel systems. To ensure that SolCAP SNP marker assays are relevant for small breeding labs, we will work with individual programs to convert *trait-linked* SNP markers into “breeder-friendly” markers (CAPs or SNAPs; e.g. 20, 38, 61) using flexible funds under the small grants program.

Specific Outcomes and Timetable: An integrated Solanaceae genetic map with ~1000 SNPs in tomato and ~4000 in potato will be incorporated into SGN. Markers for CHO, vitamin and many other traits will be made available for individual breeding programs to use. ***Year 1:*** Small grants will be awarded to genotype 3-5 potato populations. Data will begin to be incorporated into SGN. ***Years 2 and 3:*** Small grant funding for SNP genotyping will continue with up to 8 mapping populations for potato and 8 mapping populations for tomato; validation studies for CHO and vitamin QTLs will be initiated; marker conversion projects will be initiated and completed, QTL validation or other future national needs (as prioritized by the small grants funding) will be addressed.

Obj. 7. Create integrated, breeder-focused resources for genotypic and phenotypic analysis by leveraging existing databases and resources at SGN and MSU. (Buell, Mueller, Francis)

SolCAP will leverage existing bioinformatics resources from the Buell lab (URL 19) and SGN (URL 17) to provide interpretation and dissemination of the project data. The division of effort between these two bioinformatics resources will allow the expertise of each group to be optimized and provide the community with integrated interfaces for data-mining. Hyperlinks between the two resources will ensure broad dissemination of the project deliverables. Douches

and Buell will continue to manage the SolCAP website, while Mueller will manage the SGN database. A SolCAP Workspace has been created in the PBGCoP to develop content for publishing to eXtension.org. Buell will facilitate project communication by interlinking project, database, bioinformatics, education and extension computational resources. The Buell lab will be the primary interface for those who generate raw data in the SolCAP project while SGN will be the primary interface for the public and breeders to access processed data. This will allow for rapid, custom interfaces to be developed for collecting, QC and dissemination of data to multiple entities, to avoid unnecessary duplication of effort.

Project Web Site. The Buell and Douches labs have constructed the project-specific website (URL 18) which currently includes project information pages and publicly available potato and tomato SSR and SNP markers. All SolCAP data will be centralized and available on the project website for bulk download. Fixed formats will be defined for all data types and controlled vocabularies will be used to obtain the phenotype and genotype data from the data providers. Specifically, the project website will contain the following: 1) Project description, 2) Links to Outreach/Education material, 3) Sequence data generated in this project, 4) SNP query tools, 5) Descriptive pages for genotype, phenotype, mapping data generated by SolCAP participants with linkages to SGN pages for integrated queries, 6) Simplistic query/viewing pages of SolCAP data with hyperlinkages to SGN, dbEST, etc. and 7) Bulk download of genotype, phenotype, and mapping data.

Sequence data. The Buell lab will deposit all of the sequence generated in this project to the NCBI Short Read Archive and will make the assembled sequences available through search and download tools on the project website. Using multiple bioinformatics analyses, the Buell group will mine the sequences generated in this project, along with all publicly available potato and tomato genome and transcriptome data, to identify genome wide putative SNPs in each crop. The polymorphism data will be available through a bioinformatics infrastructure in place at the Buell lab at MSU to identify, store, and display polymorphisms among and within genomes. As the tomato and potato genome sequences become available, we will use these as additional reference sequence for our sequence and polymorphism data and display this on a Genome Browser analogous to what is provided on the MSU Rice Genome Resource Project (URL 12). Query tools will be developed to identify SNPs based on breeding lines, annotation to genes, chromosome regions, etc.

Genotype Data. The genotype data generated through this project will be quality checked by the SolCAP participants, collated by the Buell lab and made available through simple web interfaces (with descriptive text and web based tables) on the project website that are hyperlinked to SGN, and through bulk downloads. Accessing data through both MSU and SGN will allow centralized preprocessing and QC of data and dissemination. We will be able to efficiently process entire data sets for use by a greater research community through SGN. Data release on the web site and deposit to SGN will occur immediately after quality control tests.

Phenotype data. Buell and Francis will work with Mueller to link breeder “character descriptors” and “evaluation descriptors”, as defined by UPOV and International Plant Genetic Resources Institute (IPGR), with the standardized phenotype ontologies and controlled vocabularies implemented in SGN. Phenotype data will be collected from the SolCAP participants. The data will then be reviewed by the Germplasm Coordinators (Douches, Francis) for fitting within the biological range of variation, proper use of rating scales and formatting. The Buell lab will perform consistency tests on the data (format, nomenclature). The phenotype data will be made available as web pages linked to SGN and for download on the SolCAP

website with simple descriptive text of the experimental metadata. Phenotype data will be deposited in bulk to SGN and made publicly available immediately following QC.

SGN website. Managed by Mueller, SGN (URL 17) will provide the public interface for the genotype, phenotype, sequence and mapping data in a breeder and biological context, continuously updating the links of these data to other data types in the SGN database, and providing links to the SolCAP website. The interface will be accessible through a breeder-specific “portal”. SGN will receive all data from the SolCAP project from the Buell lab in a bulk format. SGN currently has the ability to store, query and visualize phenotypic and genotypic information associated with plant accessions and to calculate QTLs on the fly; data types include images, ontology annotations, mapping information, individual marker scores, trait data (both qualitative and quantitative), literature, allele, polymorphism and other information. Plant accessions are grouped into populations, which can represent mapping populations, mutagenized plants of the same background, introgression lines, or accessions that have been grouped together for some other reason. Accession trait data can be compared to population mean, minima and maxima, and genotypic data is visualized in a map-viewer; all trait and genotype data can be downloaded in bulk and searched through query tools.

Improve access to information through the “Breeder’s Toolbox”: An explicit outreach objective of SolCAP is to improve access to information housed in SGN. Curriculum developed through PBGWorks and workshops are expected to facilitate this goal. At the same time using CoI input to improve access to data housed in SGN through modification of the SGN “Breeder’s Toolbox” will assure that the bridge between developer and user communities is built from both directions. The Breeder’s Toolbox is being developed as a portal into phenotypic, genotypic and sequence data developed by the project. SGN search functionalities are currently genomics and ontology oriented. There is a need to bridge phenotype ontologies developed by the genomics community with terminology used by breeders. To accomplish this we will implement more intuitive searches, including drop down menus with terms and word completion scripts, as well as organizational structures based on market classes and niches that may not necessarily track to ontologies. Several of these approaches have already been implemented into databases that are being ported to SGN (e.g., Tomatomap.net (URL 26)) and Tomato fruit morphology database (URL 24). Success of this objective will be evaluated based on tracking use of SGN resources through the portal, prior to and following specific updates. In the SGN user database, we will add specific fields that will let users self-identify as breeders, allowing us to better track the fraction of users that are breeders. In addition, we will create an SGN breeder specific mailing list, which will allow SGN to announce new breeder specific functionalities and data, and will also let subscribers post messages. Number of subscribers and number of posts to the list will be an additional metric to measure breeder participation.

Specific Outcomes and Timetable: Breeder-friendly interactive interfaces to query and display data from trait to gene including germplasm, phenotypic, genotypic, genomic data with links to protocols and references; marker analysis/assay/conversion resources will be developed; Breeder adoption of genomic tools and markers in their breeding programs will be measured by Dr. Coe. **Year 1:** Data will begin to be uploaded to the MSU SolCAP and SGN will initiate SolCAP data integration. **Years 2 and 3:** Year 2 efforts will continue with refinement based upon assessments of Dr. Coe.

SolCAP Evaluation (MSU IRB X09-177)

The long-term research goals of SolCAP are beyond the timeframe of the specific objectives of the proposed research and their measurement would require evaluation activities

beyond the funding period. For example, new varieties take time to reach market and gain share (48). Intermediate goals, such as increased awareness and use of MAS for breeding programs (48), increased educational opportunities and engagement of students in the further development of marker technologies, and increased productivity through improved technologies (and access to technology) are more easily measured during the life of the project. Our external evaluation will therefore assess awareness and adoption of genetic marker technology, capacity building, and increased skills of participants. Project evaluation and mentoring will be provided by the stakeholder and scientific advisory committees.

External evaluation of SolCAP education and extension objectives will be provided by Dr. Michael Coe. Dr. Coe has extensive experience in evaluating educational projects funded by the US Department of Education, the National Science Foundation, and numerous state and local educational agencies, many with a focus on educational uses of technology. He is currently serving as external evaluator for the Conifer Translational Genomics Network.

Surveys, interviews, document review, and participant observation at key meetings and events will be used to assess baseline status and progress on the education and extension objectives. The breeder workshops will be initiated at the Tomato Breeder Roundtable and the Potato Association of America meetings using funding from the USDA/NRI. Through input from the executive committee, surveys have been developed that target the tomato and potato communities. These surveys will be used to target workshop content. In addition to survey results, regular informal communication and annual formal reports will be used to provide feedback to assist in project management and continuous improvement, as well as to document cumulative progress and impact through years 1-3 of this USDA/AFRI project.

Subject-specific surveys evaluating knowledge, skills, attitudes, and use of genomic tools and markers will be conducted before and after key education and extension activities. The effectiveness of education and extension resources will also be evaluated using interviews with select users, surveys, and assessment tools to be integrated in the online modules to gather information on frequency and ease of use, quality and format of materials, and size and type of audience. Workshop participants will be asked questions about their use of markers and genomic tools in breeding, how often they access SolCAP developed tools and other similar resources, and their use of SolCAP developed education materials within their programs.

In addition to the education and evaluation activities included in Objectives 1 and 2, the external evaluator work with project leaders and staff to assess the usefulness and impact of the public web portal operated by SGN. Surveys and interviews with users and a review by a panel of web design specialists will provide feedback for continuous improvement and will be documented along with quantitative measures of the use of the portal.

Activities such as the creation and use of curriculum materials, databases, and other resources will also be documented through review of the materials themselves, evidence of their use in course syllabi, participant counts, publications, citations, and website use statistics. Additional surveys and interviews with users of these resources will be conducted.