Translational genomics of Vegetable Crops
Las Vegas, NV
July 21, 2005
David Francis and Allen Van Deynze

At the recent ASHS meetings in Las Vegas, a workshop "Translational Genomics of Vegetable Crops" sponsored by the Vegetable Breeding Working Group was held. The workshop title borrowed the word "Translational", from medical research in which it is understood to refer to the use of basic knowledge for applied outcomes, as in the "the process of translating discoveries in the laboratory into clinical interventions" (Minna and Gazdar, 1996). In applied plant science, "translational genomics" implies the adaptation of information derived from genome technologies for crop improvement. The purpose of the ASHS workshop was to raise awareness of translational research in the vegetable crops through an overview of current genome projects in the *Solanaceae* and *Compositae* and translational research for marker development, germplasm curation, and breeding. Following formal presentations a group discussion was held to initiate organizational efforts that may boost translational research in vegetable crops. In this letter, we summarize the workshop and principal issues that emerged from the meeting discussion.

Several themes emerged during the course of the workshop that are worth stating early in this summary letter. First, "translational" research that makes use of genome sequencing information requires that we think about agricultural research from the point of view of taxonomic groups and DNA sequence homology rather than traditional commodity boundaries. For example ornamental crops that are taxonomically related to tomato and potato may benefit from sequencing efforts in those crops if resources for translational research are mobilized. To maximize the use of resources, a research community must be willing to work beyond traditional commodity divisions. Despite the potential of translational research to benefit multiple species, most efforts in plant translational research focus on a single commodity indicating that established boundaries still limit thinking. Second, access to resources for translational research remains a primary limitation in applying genome sequence data to crop improvement in vegetable crops. The lack of resources for phenotypic characterization precludes the exploration of new populations, the collection of meaningful phenotypic data, and use of marker-assisted selection despite the abundance of DNA sequencing and genotyping facilities. Third, there remains a paucity of markers that can be applied to most breeding populations. Although the role of protein and DNA-based molecular markers has long been established for selection and introgression, academic research has often focused on wide crosses and thus the available markers are tailored for this use. For most horticultural crops, including those with well-developed sequence resources, there remains an insufficient number of polymorphic markers for application to intraspecific crop improvement efforts. Finally, in organizing and planning for large community-based efforts in translational genomics there must be a balance between achieving general goals and allowing sufficient resources to accomplish specific goals. For example an effort coordinated around a trait-based theme such as improving nutritional value would need to remain flexible enough to accommodate nutritional traits specific to individual crops. An effort that aimed to develop DNA-based markers that serve the need of multiple commodities must also meet the needs of individual market niches and breeding

programs. Having stated these general themes, it will be worth reviewing the formal talks.

The workshop began with an overview of two plant genome sequencing projects. The *Solanaceae* (http://www.sgn.cornell.edu/) and *Compositae* (http://cpgdb.ucdavis.edu/) projects are both providing a wealth of information that may be of practical use, and they provide two visions of project organization. In his overview of the *Solanaceae* project, Dr. Giovannoni (USDA-ARS, Ithaca, NY) stressed the development of an international consortium with the goal of sequencing the gene-rich regions of the genome of a single tomato variety. Over eleven countries are participating in this project with participants focusing on individual chromosomes or organelle genomes. A primary goal of the tomato sequencing effort is to establish an information network to address significant questions in plant biology such as "how can a common set of genes and proteins give rise to a wide range of morphologically and ecologically distinct organisms?" and "how can a deeper understanding of the genetic basis of plant diversity be harnessed to better meet the needs of society in an environmentally-friendly and sustainable manner?" (http://www.sgn.cornell.edu/help/about/us_tomato_sequencing.html). Sequencing efforts in tomato have all ready yielded results of translational value as data deposited into publicly accessible databases has been used to discover molecular markers of use for genetics and breeding research (Yang et al., 2004; Frary et al., 2005).

In her overview of the *Compositae* project, Leah McHale (U.C. Davis, Davis, CA) described the aim of deep sequencing expressed sequences from a number of species in the family. The project initially focused on cultivated lettuce (*Lactuca sativa* L. cv. Salinas), wild lettuce (*Lactuca serriola* L.) and Sunflower. Sunflower efforts included *Helianthus annuus* lines RHA801 and RHA280, *Helianthus paradoxus*, and *Helianthus argophyllus*. In the *Compositae* genome initiative, expressed sequence tags (ESTs) are being developed for the parents of key mapping populations for lettuce and sunflower such that the basic research has generated translational information from the outset. By including both wild and domesticated forms of lettuce and sunflower, the project aims to gain insight into the process of domestication. The project has been expanded to encompass six domestication events and ten weed species including safflower, *Echinacea*, chicory, and marigold. Species were chosen for their commercial interest as crops and economic impact as invasive plants. Despite the contrasting organization of the two projects, both are yielding results that can facilitate translational research. It is also clear that planning from the outset for translational outcomes can facilitate simultaneous discovery and application.

A primary goal of many translational research projects is the development and application of DNA-based molecular markers for use in trait discovery, genetic mapping, and selection. Allen Van Deynze (U.C. davis, Davis, CA) described several approaches that use existing sequence resources to discover or develop markers. Bioinformatic approaches that can be used to identify single nucleotide polymorphisms (SNPs) in existing sequence databases were described (e.g. Yang et al., 2004). An approach that combines bioinformatics and *de novo* sequencing to identify polymorphisms in introns was also described. By using the sequence of the *Arabidopsis* genome, it is possible to predict intron position in other plants with close to 95% accuracy. Primers can then be designed to span introns, and a comparison of sequenced PCR products can be used to

discover polymorphisms. This so-called "intron scanning" approach has been applied to tomato "conserved orthologous set" (COS) genes for the development of COS intron (COSi) markers. Rates of polymorphism have been found to be five-fold higher in the non-coding introns relative to the coding exons. Finally, advances in technologies for identifying genetic polymorphisms rapidly and accurately have dramatically accelerated the discovery of molecular markers. The detection of genetic differences using oligonucleotide arrays was first described for yeast (Winzeler et al., 1998). This methodology differs from the "traditional" techniques for marker discovery in that the approach is reversed and potential markers are affixed to a single high-density microarray. "Microarray" refers to an experimental approach in which biological materials (molecules) are deposited onto a solid support at high density in a known configuration. This method allows simultaneous analysis of many single feature polymorphisms (SFPs) in DNA or RNA using a single hybridization experiment. Dr. Van Deynze described how an oligonucleotide chip based on the AFFYMETRIX platform is being designed specifically for SFP discovery in lettuce, with oligonucleotides designed along a tilling path with a two-base shift. It is estimated that such a chip will be able to detect 5000 polymorphisms in lettuce breeding germplasm. Among the methods for rapid identification and genotyping, the detection of genetic differences with specifically designed DNA chip arrays is perhaps the most promising (Huber et al., 2002; Stears et al., 2003; Wong et al., 2004).

The application of DNA sequence data to germplasm curation was described by Joanne Labate (USDA-ARS, Geneva, NY). Two examples were provided, one based on a candidate gene approach in which genes affecting *Arabidopsis* were used to investigate genetic diversity in cauliflower and broccoli. The *ap1-1/cal-1* mutant of *Arabidopsis thaliana* results in a proliferating and arrested inflorescence similar in phenotype to cauliflower and broccoli leading to speculation that similar genes from *B. oleracea* may be responsible for this characteristic trait (King, 2003). Although genotypes at the *BoCAL-a* locus were not predictive of heading phenotype in *Brassica oleracea*, genetic variation was detected for this locus in the germplasm collection (Labate et al., 2003). The use of SNPs to genotype the USDA tomato collection was described with an emphasis on *S. lycopersicum* (formerly *Lycopersicon esculentum*) collections from a wide geographic range including land races and wild progenitors (*S. lycopersicum* var. *cereasiformae*). The polymorphic information content (PIC) of SNPs in tomato is approximately 1 in 7.7 to 8.5 Kb, and is therefore low (Nesbitt and Tanksley, 2002; Yang et al., 2004; Labate et al. 2005. Molecular Breeding, *in press*) The low polymorphism results from evolutionary processes such as genetic bottlenecks and intense selection that resulting from domestication events. Low levels of polymorphism create a challenge for characterizing germplasm collections. Despite a low PIC, SNPs provide an abundant class of DNA polymorphisms in plant and animal genomes. As molecular markers, SNPs can be used to saturate linkage maps in gene rich regions, can be tied directly to functional polymorphisms, and can be associated via linkage disequilibrium to quantitative and qualitative traits (Labate et al. 2005 Molecular Breeding, *in press*). As markers, SNPs also have advantages over simple sequence repeat markers (which can also be mined from sequence data), including the reduced occurrence of similarities that are not homologous (homoplasy). A divergence in the needs of

breeders and geneticists seeking to map traits or affect selection, is that genotyping for germplasm characterization should not focus on polymorphic markers alone.

In the final presentation, Shanna Moore (Cornell University, Ithaca, NY) presented two examples of how pepper breeding in the program of Molly Jahn has benefited from sequence data in related Solanaceae crop species. Both projects, emphasized a candidate gene approach in which genes likely to be involved in pungency or virus resistance were identified. In the case of pungency, the gene responsible can only be found in pepper species, and thus was isolated by identifying candidates among published sequences. One gene, AT3, showed significant similarity to acyltransferases, mapped to the Pun1 locus (required for the presence of capsaicin), and contained a deletion spanning the promoter and first exon of the predicted coding region in every non-pungent accession tested (Stewart et al., 2005). Potyvirus infection in pepper is caused by several destructive viruses. The translation initiation factor eIF4E had been identified in a number of host species as a recessive resistance factor. Genetic mapping demonstrated an association of the candidate gene, eIF4E, with the pvr1 locus in Capsicum. Further sequencing demonstrated that multiple mutations of the eIF4E homolog are associated with an allelic series that result in a range of potyvirus resistance from broad to narrow spectrum (Kang et al., 2005). Molecular markers based on pun1 and prv1 allow breeders to select directly for traits of interest in large populations and at early growth stages, thus saving time, space, and money.

Following the formal presentations, a discussion was held to develop a strategy to build communities of researchers to facilitate translational genomics of vegetable crops. Thirty-six public and private scientists, seed company representatives and commodity group representatives working in *Solanaceae*, *Compositae*, *Curcurbitae*, and *Leguminosae* and representing both vegetable and ornamental crops participated in the discussion. The USDA/NRI Coordinated Agricultural Project (CAP) program offers a template for organization of researchers for application of genomic resources in order to maximize the benefit of translational research. The applied plant genomics CAPs were initiated to bring together scientists and stakeholders with a shared vision and plan to facilitate translation of basic discoveries and technology. The goal is to create an inclusive community consisting of applied and basic, private and public researchers combined with participation of commodity groups, growers, and end users. Translational research that leverages genome sequencing information requires that we think from the point of view of DNA sequence homology rather than conventional commodity or department boundaries. To maximize the use of resources, a research community must be willing to work beyond traditional divisions. It is unclear, however, where the new divisions should be established. For example, the Asterid clade would include *Solanaceae, Rubiaceae* (coffee) and *Compositeae*. History, to date, has not supported such broad based efforts. The first CAP was funded for rice in 2004. In CAP planning efforts to date, organizational efforts have not been able to transcend traditional divisions, perhaps due to resource limitations or due to unique needs for each commodity. Planning efforts are underway for CAPs that focus on Wheat, Soy, Barley, and Cotton with several planning efforts described on the worldwide web. Thus a major hurdle in developing an organizational structure that spans taxonomic groups will be the development of resources that serve a general need while providing capital to address individual needs.

During the workshop we recommended a family-based CAP over a commodity-based CAP for vegetable crops. Species such as tomato all ready span several commodities with distinct germplasm pools for processing, field fresh market, and greenhouse crops. The tomato breeding and genetics community is accustomed to working across commodity lines. Likewise, the potato breeding and genetics community is well organized and accustomed to working across processing and fresh market commodity lines. A discussion of the potential for a *Solanaceae* CAP (SolCAP) was held during the annual potato industry meetings in Calgary. Strong support to develop a SolCAP having the potato research community as a significant partner exists. Thus expanding cooperation within traditional crops to encompass the *Solanaceae* will provide a model for family-based translational genomics. Despite this goal, balancing the development of general resources with the needs of specific crops and programs will remain an issue prior to implementation. The successful funding of a CAP in the future should not be considered the sole benefit of organizing. Any organizational structure that improves access to technology, resources, and methods for applied plant research will benefit public and private sector research, growers and end-users.

During the general discussion, each member of the group had an opportunity to provide input on the kind of tools that are needed for translational genomics and the best traits to target. By far the most common suggestion was the need for flexible tools that comprehensively sample variation in breeding populations. The limitation of markers for use in breeding populations stems from a long emphasis on the development of tools for wide crosses that has left a gap in tools for application. Discussion also covered the need to develop a core set of markers, which may include COS markers as anchors across species, and the need to develop common panels of germplasm for screening new markers across species. Genome projects that focus on generating basic information often fail to incorporate a breadth of economically relevant germplasm which limits translational research. It was noted that marker data should be curated and available in a common format, including databases such as the SNP database (dbSNP) which is now incorporated into the National Center for Biotechnology Information (NCBI) Entrez system. In order to facilitate large-scale archiving of information, a community will need to adopt common table and data formats.

Several participants noted that access to technology due to cost was a limitation to exploiting existing genomic resources. It is worth noting that the perception of financial limitations to technology access is not limited to agricultural research as translational research in the medical field "is hindered by insufficient targeted resources, a shortage of qualified investigators, [and] an academic culture that hinders collaboration between clinical and laboratory-based investigators" (Pober et al., 2001). Thus primary goals of community building efforts should be to focus on bridging the collaboration gap between applied and basic research, developing marker resources for applied research and building an infrastructure that increases access to translational tools.

There is a further need to identify resources for population development and characterization for relevant traits. Discussion regarding the prioritization of traits focused on the possibility of considering specific pathways or organs across species with an emphasis on nutritional targets. This later suggestion mirrors the potato community's suggestion that priorities should target health and environment issues rather than direct production constraints. Prioritizing traits will require further discussion as the general

needs of a community and individual needs of a breeding program may not always overlap. Follow up meetings for a *Solanaceae* CAP (SolCAP) are scheduled for November 15, in Davis CA, January at the Plant and Animal Genome Conference in San Diego, CA, and July 2006 at the Third *Solanaceae* Genome Workshop in Madison Wisconsin.

Recommendations for organizing the *Solanaceae* community follow:

- Seek partners from other commodity groups in the *Solanaceae* and organize around taxonomic groups and DNA sequence homology rather than traditional commodity boundaries.
- Reduce duplication, both by dividing the workload and improving information exchange, in order to help leverage scarce resources and build community resources.
- Develop flexible tools that comprehensively sample variation in breeding populations including a core set of markers for use as anchors across species.
- Develop common panels of germplasm for screening new markers across and within species.
- Create bioinformatic platforms that allow access, updating, and sharing of data and information among all researchers in the community.
- Curate marker data in a common format so that database tables can be shared and expanded.
- Adopt trait-ontology approaches for the collection of phenotypic data in standardized formats and promote the development of phenotypic databases.

In conclusion, the application of data from genome sequencing projects to the improvement of vegetable crops will benefit from community organization that spans commodity groups. A major goal of organization efforts to facilitate translational genomics and applied plant biology should be to reduce redundancy and improve access to technology. Efforts that reduce duplication, either by dividing the workload or improving information exchange, will help leverage scarce resources to accomplish more. A major research effort that helps discover polymorphic markers across species and within relevant germplasm pools appears to be emerging as a primary goal. Cost may be lowered and access to technology may be improved if the community can develop both a plan and infrastructure to share common reagents such as primers, DNA for a common panel of varieties, and other genotyping reagents. Information sharing that involves the collection of data in a common format and the development of tools that increase accessibility and ease of viewing will further strengthen research efforts and reduce duplication.

Acknowledgements:

We would like to thank the Organizing Chair, Dr. Mikel R. Stevens (Brigham Young University) and workshop presenters Dr. Jim Giovannoni (USDA/ARS and Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY), Leah McHale (U.C. Davis Genome Center, Davis, CA), Dr. Allen Van Deynze (Seed Biotechnology Center, Plant Reproductive Biology, UC Davis), Joanne Labate (Plant Genetic Resources Unit, USDA, Cornell University, Geneva, NY), and Dr. Shanna Moore (Plant Breeding,

Cornell University, Ithaca, NY) for their efforts and contributions. Dr. Allen van Deynze (U.C. Davis) and Dr. Ryan Hayes (USDA/ARS Salinas, CA) served as recorders and provided notes. Dr. David Francis organized the workshop and served as moderator. Additional ideas and discussion points were contributed by Dr. David Douches (Michigan State University, East Lansing, MI), Dr. Walter De Jong (Cornell University, Ithaca, NY), Dr. Steven Tanksley (Cornell University, Ithaca, NY), Dr. Molly Jahn (Cornell University, Ithaca, NY), Dr. Martha Mutschler (Cornell University, Ithaca, NY), Dr. John Stommel (USDA/ARS, Belltsville Maryland), and Dr. Dina St. Clair (University of California, Davis). We offer a special thanks to those workshop participants that contributed to the discussion.

References:

Huber M, Mundlein A, Dornstauder E, Schneeberger C, Tempfer CB, Mueller MW, Schmidt WM. 2002. Accessing single nucleotide polymorphisms in genomic DNA by direct multiplex polymerase chain reaction amplification on oligonucleotide microarrays. Anal Biochem. 303(1):25-33.

Frary, A., Y. Xu, J. Liu, S. Mitchell, E. Tedeshi, and S. Tanksley. 2005. Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor. Appl. Genet. 111: 291-312.

Kang, B. C., I. Yeam, J. D. Frantz, J. F. Murphy, and M. M. Jahn. 2005. The pvr1 locus in Capsicum encodes a translation initiation factor eIF4E that interacts with Tobacco etch virus VPg. Plant J. 42:392-405.

King, G. J. 2003. Using Molecular Allelic Variation to Undestand Domestication Processes and Conserve Diversity in Brassica Crops. In: International Symposium on Sustainable Use of Plant Biodiversity to Promote New Opportunities for Horticultural Production Development, E. Düzyaman and Y. Tüzel, editors. Acta Hort. (ISHS) 598:181-186

Labate, J., L. Robertson, and T. Bjorkman. 2003. Genotypes at the *BoCAL-a* locus in B. *oleracea* do not predict broccoli, caulifolower, and purple cauliflower phenotype. HortScience 38:736.

Minna, J.D. and A.F. Gazdar. 1996. Translational research comes of age. Nature Med. 2:974-975

Nesbitt, T.C. and S.D. Tanksley (2002). Comparative sequencing in the genus Lycopersicon: implications for the evolution of fruit size in the domestication of cultivated tomatoes. Genetics 162:365-379.

Pober, J.S., C.S. Neuhauser, and J.M. Pober. 2001. Obstacles facing translational research in academic medical centers. The FASEB Journal. 15:2303-2313.

Stears RL, Martinsky T, Schena M. 2003. Trends in microarray analysis. Nat Med. 9:140-5.

Stewart, C. B. C. Kang, K. Liu, M. Mazourek, S. L. Moore, E. Y. Yoo, B. D. Kim, I. Paran, and M. Jahn. 2005. The Pun1 gene for pungency in pepper encodes a putative acyltransferase. Plant J. 42: 675-688

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, Davis RW. 1998. Direct allelic variation scanning of the yeast genome. Science. 281(5380):1194-7.

Wong KK, Tsang YT, Shen J, Cheng RS, Chang YM, Man TK, Lau CC. 2004. Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. Nucleic Acids Res. 32(9):e69.

Yang, W., X. Bai, E. Kabelka, C. Eaton, S. Kamoun, E. van der Knaap, E., and D. Francis. 2004. Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. Molecular Breeding 14:21-34.