SNP Consortium Pre-proposal

Joanne Labate, Angela Baldo and Larry Robertson USDA Geneva

Cultivated tomato (*Lycopersicon esculentum* var. *esculentum*) is known to have relatively low molecular genetic diversity. This limited genetic variation has restricted the use of molecular markers as tools for genetic studies or crop improvement. The USDA, ARS Plant Genetic Resources Unit (PGRU) in Geneva, NY currently conserves more than 5,000 accessions of cultivated tomato. We are developing DNA sequence-based molecular markers (Single Nucleotide Polymorphisms, known as SNPs) to characterize our collection. Polymorphic markers that we develop will also be useful to our stakeholders (e.g. tomato breeders) for saturating genetic maps and marker-assisted selection.

Tomato is ideal for pioneering SNP prediction and confirmation because of the wealth of publicly available DNA sequence data. Baldo has designed and implemented a high-throughput Expressed Sequence Tag (EST) analysis system which takes a NCBI Unigene set as input, and produces high-quality annotation and consensus sequences of subclusters, virtual mapping of consensus sequences by matching to known markers, SNP predictions, and SSR discovery. An additional module designs and selects optimal PCR primers flanking the regions of interest. Using the more than 150,000 EST and cDNA sequences from over 15 cultivars comprising the 3,000+ member Unigene set for *L. esculentum* in Genbank, 2,527 SNPs in 764 Unigenes were predicted. In 2004 we tested 85 independent amplicons from the 764 Unigenes for predicted SNPs by sequencing two or three cultivars per amplicon. We discovered 62 SNPs and 13 small insertion/deletion polymorphisms in 21 amplicons. For the 64 remaining amplicons, one primer pair did not amplify, thirty two showed no evidence of the predicted SNP, 20 appeared heterozygous or gave multiple PCR bands, and 11 gave insufficient data (poor quality sequence, fragment too large to sequence predicted SNP site, or data from only one cultivar). Based on the 53 amplicons that gave clear results thus far, this method discovered cultivated tomato SNPs with approximately 21-fold more efficiency compared to sequencing random genomic DNA (1 SNP per 300 nucleotides versus 1 SNP per 7 kb). We have submitted a manuscript for publication of the 21 polymorphic markers and are continuing development and testing of the remaining set of 64 of the 85 originally tested.

We propose to develop and test the remaining 679 Unigenes with predicted SNPs in a collaborative effort with private companies such as Campbell's, Western Seed, Syngenta/Roger's Seed, Rijk Zwaan, Zeraim Gedera, and other interested parties. Based on our results we anticipate successfully developing at least 25% of the 764 Unigenes (191 loci) into polymorphic markers by sequencing two to three tomato cultivars with predicted SNPs among them. This will entail - obtaining the minimal set of cultivars needed based on NCBI's original EST data, growing the cultivars in PGRU's greenhouse and isolating DNA from leaf tissue, PCR primer design and PCR amplification for the remaining 679 Unigenes for two or three cultivars each, DNA sequencing, and analyses of the sequence data. The anticipated cost of carrying out this project in-house at PGRU is detailed in the attached budget pages. Products will be robust PCR and sequencing primers for genomic DNA sequences, identification of polymorphic sites among the assayed cultivars for those sequences, and annotation of sequences including predicted proteins and matches to known tomato markers. All results will be published in peer-reviewed journals and all collaborators will have access to preliminary and pre-published data.

The proposed research is part of PGRU's "Conservation of Vegetable Crops" CRIS, Objective 2 "Enhance the effectiveness of germplasm maintenance through the application of genomic sequencing and molecular marker techniques". It will serve as a model for discovery of enhanced value within crop germplasm when faced by limited molecular genetic diversity, ensuring the future viability of U.S. farms and a nutritious food supply.

Budget

Item	**Year 1**	**Year 2**	**Total**
Materials			
PCR and DNA sequencing ($4.95 per sample)	$12,997	$12,997	$25,994
Primers	$11,270	$1,691	$12,961
Instrument maintenance	$2,192	$2,192	$4,383
Labor			
Technician (GS-4)			
salary	$24,698	$26,288	$50,986
benefits	$7,409	$7,886	$15,296
3 Supervisory scientists (5% each)	$16,778	$17,615	$34,393
Total materials and labor	$75,344	$68,669	$144,013
Other Direct Costs			
PGRU overhead IRC (0.1905)	$14,353	$13,081	$27,434
Total Direct Costs	$89,697	$81,750	$171,447
Indirect Costs			
ARS overhead (0.1111)	$9,965	$9,082	$19,048
Total Direct and Indirect Costs	**$99,662**	**$90,832**	**$190,495**

Sample sizes

genes	cultivars	sequencing reactions (forward and reverse)	cost/sequence	cost	number of additional reactions (25%) for redos	cost for redos	total sequencing cost	total number of sequencing reactions
700	3	2	$4.95	$20,795	1050	$5,199	$25,994	5250

no. of primer pairs	forward and reverse	size of primer	cost per base	cost	number of additional primers (15%) added for redesigns	cost of redesigned primers	total primer cost	total number of primers
700	2	23	$0.35	$11,270	210	$1,691	$12,961	1610

In-house lab costs are itemized as follows:

Estimated costs of PCR and sequencing per sample

	cost/unit	number expended/sample	cost/ sample
10ul unplugged pipette tips	0.0173	1	0.02
200ul unplugged pipette tips	0.014	2	0.03
10ul plugged pipette tips	0.049	7	0.34
200ul plugged pipette tips	0.063	2	0.13
large orifice 200ul unplugged pipette tips	0.026	2	0.05
large orifice 200ul plugged pipette tips	0.058	2	0.12
50ul plugged pipette tips	0.05	2	0.10
black 96-well clini plates	0.02	2	0.04
PCR plates	0.07	2	0.14
1.5ml tubes	0.03	1	0.03
Molecular Probes picogreen reagent (1 includes forward and reverse)	0.14	0.5	0.07
dNTPs	0.05	2.2	0.11
Promega GoTaq PCR enzyme	0.22	2.2	0.48
EdgeBio PCR clean-up (1 includes forward and reverse)	1.03	0.5	0.52
ABI BDT cycle sequencing enzyme	1.57	1	1.57
EdgeBio sequencing rxn. clean-up	0.52	1	0.52
ABI sequencing polymer POP6	0.26	1	0.26
ABI 3100 capillary array	0.43	1	0.20
		total	**$4.95**

ABI 3100 DNA sequencer maintenance

lower polymer block	$2,200
reserve polymer syringe	$258
array-fill syringe	$150
maintenance contract (10% over 2 years)	$1,775
total	**$4,383**

Budget narrative

Primers
Custom-synthesized oligos are purchased from Integrated DNA Technologies, Inc. PCR/sequencing oligos cost $0.35/base at 25 nmole scale. Generally 4 pmol primer is used per PCR or sequencing reaction. We estimate 1,610 primers will be sufficient to PCR and sequence 700 loci taking into account needs for occasionally redesigning primers and replenishing primer stocks (15% added). 23-mer x $0.35 x 1,610 = $12,961

Laboratory costs
We routinely test all newly synthesized PCR primers across the 2 or 3 genomic DNAs to be sequenced in small volume PCR reactions. After this initial optimization a larger volume (50 ul) PCR reaction is performed, PCR products are cleaned using an EdgeBioSystems Quickstep kit, DNA is quantified in a picogreen assay, ABI BDT ver. 3.1 cycle sequencing reactions are performed, samples are cleaned on an EdgeBioSystems DTR plate, dried down, resuspended in formamide, and run on the ABI 3100 capillary sequencer.

Itemized costs per sample are in the budget table. Plastic disposables include pipette tips, 96-well plates, and eppendorf tubes. Disposable capillary arrays for the ABI 3100 cost $695 each and last approximately 100 runs (1 run = 16 samples). Costs of reagents including dNTPs, GoTaq and BDT enzymes, picogreen, POP6 polymer, and EdgeBio kits are estimated on a per sample basis. Cost per sequence = $4.9513 x 5,250 sequences = $25,994

Instrument maintenance
The throughput of our ABI 3100 DNA sequencer is two 96-well plates in 30 hours. For 5,250 reactions this is equivalent to 8 months of operation @ 4 hours per day. We replace the lower polymer block, array-fill syringe, and reserve polymer syringe approximately every 6 months as routine maintenance. In 2004 the ABI service maintenance contract cost $8,750. We request $2,192 per year to defray instrument maintenance costs.

Labor
One full-time ARS Biological Science Technician (GS-4) is requested to carry out the majority of the labor for this project. Salary was estimated at entry level with 30% added for benefits. There will be three supervisory scientists whose costs are estimated at 5% of their salary and benefits each. Their respective roles can be described as follows - Vegetable Crops Curator Dr. Larry Robertson will be responsible for obtaining seed of required tomato cultivars, overseeing its planting and harvesting for tissue collection, and subsequent sample tracking and data curation. Larry has developed a database for PGRU for efficient tracking of samples from planting, through all laboratory assays, to final storage of molecular data. Bioinformaticist Dr. Angela Baldo will be responsible for all computational marker discovery: identification and downloading of public EST, cDNA, and genomic sequences, high-throughput in-house clustering, annotation, in silico mining for markers, and high-throughput primer design. Molecular Biologist Dr. Joanne Labate's primary responsibility on this project will be to train the technician in laboratory techniques and to oversee the collection and analyses of the molecular

data, ensure quality control of laboratory-generated data, and dissemination of high quality data to collaborators.

Overhead
Local overhead at PGRU is estimated as 16% of total direct costs. National overhead for ARS is estimated as 10% of total direct and indirect costs.

Facilities and equipment
Physical facilities include three greenhouses with a total of 7500 ft^2 for growing plants. The 1400 ft^2 laboratory is well-equipped to efficiently generate DNA sequence data. Major equipment includes an MJ Research Tetrad2 and two BioRad iCycler thermocyclers, one ABI 3100 and one ABI 310 Genetic Analyzer, a Tecan Genesis RSP 150 robotic liquid handling system, a plant tissue grinder (GenoGrinder), a refrigerated-centrifuge and a speed-vac that can hold 96-well plates, an Alpha Innotech FluorChem 8900 imaging and analysis system, and a 96-well plate reader for fluorescence and absorbance assays. We also have all necessary small equipment that we require such as microcentrifuges, incubators, shakers, autoclave, agarose gel rigs and power supplies, electronic multi-channel pipettes, -20°C and -80°C freezers, a freeze-drier, microcomputers, etc.
We have at our disposal a Dell PowerEdge 6600 server with four 2.4GHz Intel side-bus processors, 16GB of RAM, and 876GB of storage in a RAID configuration, with an additional hot spare. This server currently runs Linux kernel 2.4, and shoulders the majority of the Unit's computational analyses. We have two secondary servers: a Dell 530 Workstation with a i686 single 1.7GHz processor server with 512MB memory, and 80GB of storage; Dell PowerEdge 4100 with 200MHz processor with 256MB of memory and 22GB of storage, and a desktop workstation, all running Linux. Finally we have a Sony PCG-GRX700P notebook with i686 architecture, 2.20GHz, 512MB RAM, and a 60GB hard drive, running a notebook-optimized version of the same Linux distribution. All five Linux machines are configured such that they can read and write to the shared Novell fileserver (Dell PowerEdge 2650, single processor 1.8GHz, 1MB RAM, with 76GB of storage) and each other's hard drives. The rest of the computer network available for this project consists of approximately 14 Dell desktop PCs running Microsoft Windows connected by the Novell print and fileserver mentioned above.