ESTs and candidate gene approaches in the Compositae Genome Project

Richard Michelmore[1], Alex Kozik[1], María José Truco[1], Marta Matviencho[2], Oswaldo Ochoa[1], Mireille van Damme[1], Dean Lavelle[1], Hong Lin[2], Barnaly Pande[1], Leah McHale[1], Padma Sudarshana[1], Jason Argyris[1], Paula Ellison[1], Kent Bradford[1], Louise Jackson[1] and Rick Kesseli[3]

[1]*Department of Vegetable Crops, University of California, Davis, CA 95616.*
[2]*Celera Agen, 1756 Picasso Avenue, Davis, CA 95616.*
[3]*Biology Department, University of Massachusetts, Boston MA. 02125.*

Abstract. The Compositae Genome Project (CGP) is in its third year. The initial phase has focused on sequencing of expressed sequence tag (ESTs). Over 19,000 unigenes of lettuce have been identified, probably representing at least a third of all genes expressed in lettuce. The current focus is on mapping candidate genes to agriculturally important phenotypes.

Keywords: *Lactuca sativa*, genomics, expressed sequence tag, synteny, candidate gene.

Introduction

The Compositae Genome Project (CGP) is a collaboration between the laboratories of Richard Michelmore, Kent Bradford and Louise Jackson (all at the University of California, Davis), Steven Knapp (Oregon State University, Corvallis), Loren Rieseberg (Indiana University, Bloomington), and Rick Kesseli (University of Massachusetts, Boston). The different responsibilities are shown in Figure 1. The CGP builds on long-standing interactions among these labs and has been running formally since 1999.

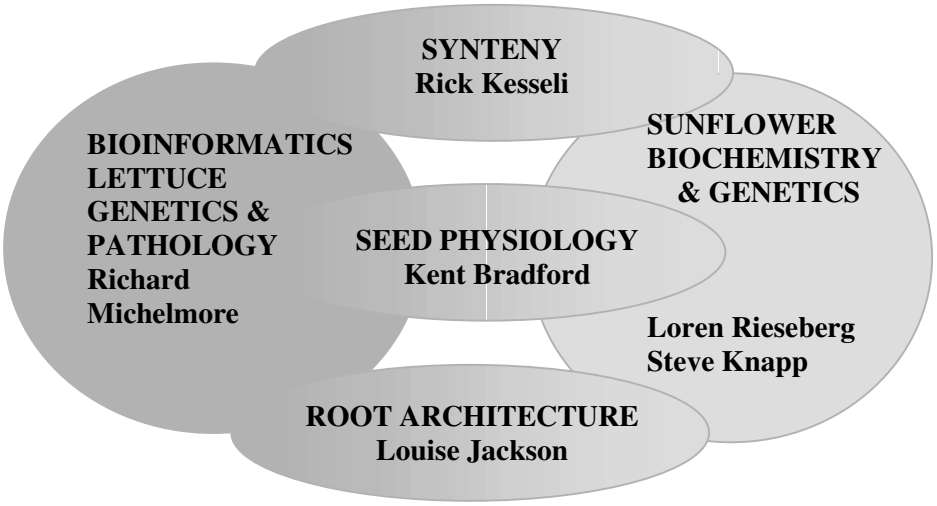

Figure 1. Areas of interest and responsibility in the Compositae Genome Project.

The goals of the CGP are:

- To develop comprehensive gene catalogs for economically important genera of the Compositae, particularly lettuce and sunflower.

- To develop detailed genetic maps integrating phenotypic data for agriculturally important traits with candidate gene sequences.
- Determine the extent of synteny between lettuce and sunflower and to *Arabidopsis* as well as to other plant species such as tomato.
- To enhance the introgression of agriculturally useful alleles from wild species.
- To establish tools and resources for lettuce and sunflower that will be the basis of genomic investigations in these genera and in the Compositae more generally.
- To understand genome evolution and phenotypic diversification in the Compositae.

The article reviews the progress with respect to lettuce. Parallel progress has been made with sunflower but is not the subject of the current paper. More details can be found at http://compgenomics.ucdavis.edu.

Background

The Compositae (Asteraceae) is one of the largest and most diverse families of flowering plants, comprising one-tenth of all known Angiosperm species. It is characterized by the compound inflorescence that has the appearance of a single "composite" flower. The Compositae is divided into two major subfamilies and one minor subfamily with 1,100 to 2,000 genera and over 20,000 species (Cronquist, 1977; Jansen *et al.,* 1991). The family has undergone extensive diversification producing a cosmopolitan array of taxa. Compositae are found in diverse habitats; anaerophytic, xerophytic, and halophytic specialists thrive in some of the more inhospitable habitats (vertisols, deserts, and salt marshes). The size and adaptive success of the Compositae have stimulated considerable research into its systematics and evolution. However, molecular characterization has lagged behind other families (Kesseli and Michelmore, 1997).

Multiple species have been domesticated within the Compositae including over 40 economically important species (Kesseli and Michelmore, 1997). These include food (lettuce, chicory, Jerusalem artichoke), oil (sunflower, safflower), medicinal (*Echinacea*, chamomile) and many ornamental (chrysanthemum, dahlia, zinnia, marigold) crops. High quality edible oils are low in saturated and high in mono- and di-unsaturated fatty acids. The Compositae are renowned for their variety of novel secondary chemicals including several novel industrial fatty acids (Caligari & Hind, 1996). The family is also a rich source of powerful insecticides and industrial chemicals, e.g., pyrethrum (*Chrysanthemum*) and rubber (guayule) (Heywood *et al*., 1977). Despite this interesting diversity, individual members of the Compositae have not been extensively studied genetically and until recently there had been minimal investment in developing genomics resources. Lettuce and sunflower are representatives of each of the two major subfamilies and are the best genetically characterized members of the Compositae.

Genetic Analysis

In collaboration with others, we have developed several genetic maps (Landry *et al*., 1987; Kesseli *et al*., 1994; unpublished data). One inter-specific cross *L. sativa* cv. Salinas x *L. serriola* UC96US23 is now the core mapping population for lettuce. In collaboration with Keygene and others, this population has so far been analyzed for 750 AFLP markers, 80 SSR loci, 51 ESTs (see below) as well as markers for the disease resistance clusters. We have completed the development of 115 F_8 recombinant inbred lines (RILs) from this population and an additional 185 families are at F_6. We now have an integrated map of over 1,400 markers and 9 linkage groups (R. Michelmore *et al*., unpublished). The core mapping population has been adopted by the European ANGEL Project (www.plant.wageningen-

ur.nl/projects/angel/) and the RILs have been distributed to several groups for additional mapping of markers and phenotypic traits.

Numerous horticultural and morphological traits have been and are being analyzed genetically. We have conducted quantitative trait locus (QTL) analyses of many traits including bolting, root architecture, and the ability to extract water from different levels of the soil profile (Johnson *et al*., 2000 & unpublished). We identified several major loci for bolting, including one with an allele for slow bolting from the wild parent as well as a potential QTL for tipburn sensitivity. We have identified and mapped many genes for resistance to several diseases (e.g. Kesseli *et al*., 1993, 1994; Maisonneuve *et al*., 1994; Robbins *et al.*, 1994; Witsenboer *et al*., 1995).

We are characterizing several genes of horticultural importance at the molecular level. Our emphasis has been on disease resistance genes, particularly the major cluster that contains *Dm3* and confers resistance to downy mildew. This has involved combinations of map-based cloning, mutagenesis, and candidate gene approaches (Meyers *et al*., 1998a, b; Shen *et al*., 1998, 2002). *Dm3* homologs comprise a large family of NBS-LRR (nucleotide binding site-leucine rich repeat) encoding genes. As part of the *Dm3* cloning strategy, we utilized PCR with degenerate oligonucleotide primers designed from resistance genes cloned from other species (Shen *et al*., 1998 & unpublished). At least 22 distinct families of RGC sequences have so far been identified. Several families had greater similarity to resistance genes from other species than to the other lettuce RGC sequences indicating that the different RGC families originated early in angiosperm evolution. Two families mapped to clusters of known resistance genes.

The Current CGP

Establishment of an extensive EST database for lettuce.
EST libraries were made using a modified SMART^{TM} (Clontech) approach. cDNAs were made from ten pools of RNA from different tissues/developmental stages/environmental conditions of each of *L. sativa* cv. Salinas x *L. serriola* UC96US23, the two genotypes that had been used as parents for the core mapping population. For each genotype, each cDNA was made using oligonucleotide primers that incorporated unique 5' and 3' sequence tags so that the source of each sequence could be subsequently identified. The cDNAs were pooled and then size-fractionated into four size classes. Each size class was directionally cloned into a medium-copy vector and transformed separately to reduce size bias.

Table 1. Summary of ESTs developed for lettuce.

Total number of reads	76,000
Total number of good reads	68,197
Average read length (nt phred 20)	527
Number ESTs in contigs	56,853
Total number of contigs	8,179
Number of singletons	11,344
Number of unigenes	19,523

Over 68,000 ESTs were generated from *L. sativa* cv. Salinas and *L. serriola* (UC96US23). Over 68,000 lettuce ESTs from the multiple libraries were assembled using CAP3 (Huang & Madan, 1999) into ~19,000 lettuce unigenes and organized in the CGP mySQL database.

Custom PHP and Python scripts were developed to manipulate the data and view the assemblies. Our utilization of diverse sources of mRNA and size fractionation strategy resulted in a very efficient gene discovery (Table 1). This number of unigenes probably represents at least a third of all genes expressed in lettuce.

Comprehensive EST data are displayed at the CGP web site. This includes the raw chromatograms, details on BLAST searches, SNP and indel polymorphisms, etc. In addition to releasing the sequences to GenBank, they were incorporated by TIGR into their latest gene indices (www.tigr.org/tdb/tgi/) and by MIPS into their SPUTNIK EST database (mips.gsf.de/proj/sputnik/lactuca/).

Development of high-density genetic maps based on transcribed sequences.
We are now focused on mapping numerous ESTs with an emphasis on candidate genes for agricultural traits (Table 2). Sequences derived from each parent were compared using Python custom scripts to identify putative indels and single nucleotide polymorphism markers (SNPs). We have so far identified ~150 indels, 14,000 SSRs and 1,500 SNPs for lettuce. Wet lab experiments have confirmed the predicted indel polymorphisms in ~90% cases and SNP polymorphisms in ~70% cases. So far 51 ESTs have been mapped in lettuce using several technologies. Indel polymorphisms have been analyzed using agarose and acrylamide electrophoresis. Lettuce SNPs are being analyzed using temperature gradient capillary electrophoresis (TGCE; REVEAL™, http://www.spectrumedix.com/Reveal.htm).

In order to maximize mapping efficiency, we utilized MapPop (Vision *et al*., 2000) and our software Genoplayer (http://compgenomics.ucdavis.edu/genoplayer.htm) to identify a subset of the most informative individuals so that marker and phenotypic analyses have only to be run on 46 individuals with little loss of genetic information. We currently have a pipeline to map ~ 20 ESTs/week including SNP discovery and validation.

Table 2. Summary of Candidate Clones for Traits in Lettuce.

	# candidates[a]	# polymorphic[a]	# so far mapped
Disease resistance[R1]	278	70	33
Developmental	49	24	1
Abiotic stress[R2]	49	30	0
Physiological	199	92	1
Other	104	102	16
All ESTs	679	370	51

[a] putative from database or experimentally confirmed, analysis on going.

***Studies of synteny between lettuce and sunflower and to* Arabidopsis.**
This phase of the project has only been initiated recently, as it required the availability of the EST sequences. To study macrosynteny, we have identified a COS (conserved ortholog sequence) set using a new clustering program, Graph9, and visualized using GenomePixelizer (Kozik et al., 2002; www.atgc.org/GenomePixelizer). About 1,200 lettuce putative COS markers have been identified (cgpdb.ucdavis.edu/database/est_vs_ath/arabidopsis_cos_map.html); subsets of these are currently being mapped. In addition, we are also utilizing the candidate genes as part of the synteny analysis whenever possible.

We have also sequenced the majority of our 158 previously mapped RFLP markers. Of these, ~80% had significant sequence similarity to orthologous or paralogous loci in *Arabidopsis*.

To study microsynteny, lettuce and sunflower EST assemblies were BLASTed against the *Arabidopsis* genome then the hits displayed as a linear graphical representation along the *Arabidopsis* genome. Each element represents a predicted ORF ordered according to position on the chromosome and there are links to each EST in the CGP database (cgpdb.ucdavis.edu/database/est_vs_ath/tigr_vs_let_and_sun.html). This has provided further candidate clusters of sequences that we are now mapping.

Databases

We have developed a series of public databases to facilitate ready access to the data. Lettcv (compositdb.ucdavis.edu/database/lettcv2/display) contains information on over 4,500 cultivars. The Compositae Genome Project site (compgenomics.ucdavis.edu) provides access to extensive EST information. Compositdb (compositdb.ucdavis.edu) contains the genetic map information and images for markers as well as descriptions of RFLP probes and PCR-based markers.

Future Directions of the CGP

Subject to funding, the next phase of the CGP will have the following objectives:

Extension of the EST database.
This will involve sequencing more ESTs from four species each of *Lactuca* and *Helianthus* as well as single genotypes of safflower and chicory with the primary goal of comprehensive gene discovery and identification of allelic diversity.

Expansion of high-density genetic maps based on transcribed sequences.
Over 2000 ESTs will be added to the consensus maps. This will establish genetic correlations between candidate genes and QTLs.

Analyses of synteny between lettuce, sunflower, tomato and Arabidopsis.
We will continue the analysis of macrosynteny and microsynteny based on genetic analysis as well as sequencing of BACs from syntenic regions in lettuce and sunflower.

Phenotypic and molecular analysis of natural variation in six genera.
This will examine variation for a spectrum of morphological and physiological traits at the phenotypic level and assess natural sequence variation for ~100 genes in 182 accessions distributed across the five and seven species of *Lactuca* and *Helianthus* respectively, as well as four wild and cultivated representatives of safflower, chicory, Echinacea and chrysanthemum. This will establish patterns of selection and indicate potential correlations between sequence and phenotype as well as the relative contributions of recombination and mutation to allelic diversity.

Acknowledgements

We thank the USDA Initiative for Future Agricultural Food Systems program for financial support.

References

Caligari, P. and Hind, D (eds.). 1996. Compositae: Biology and Utilization. The Royal Botanical Gardens, Kew. Whitstable Litho Printers. UK. 689 pp.

Cronquist, A. 1977. The Compositae revisited. Brittonia 29:137-153.

Heywood V.H., Harbourne J.B., Turner B.L. 1977. An overture to the Compositae. *In*: Heywood VH, Harbourne JB, Turner BL Eds. *The Biology and Chemistry of the Compositae.* Acad. Press. pp1-20.

Huang, X. and Madan, A. 1999. CAP3: A DNA Sequence Assembly Program. Genome Res. 9: 868-877.

Jansen R.K., Michaels H.J., Palmer J.D. 1991. Phylogeny and character evolution in the Asteraceae based on chloroplast DNA restriction site mapping. Syst. Bot. 16:98-115.

Johnson, W.C., Jackson, L.E., Ochoa, O., Peleman, J., van Wijk, R., St. Clair, D.A., Michelmore, R.W. 2000. A shallow-rooted crop and its wild progenitor differ at loci determining root architecture and deep soil water exploitation. Theor.Appl. Genet.101:1066-1073.

Kesseli, R.V. and Michelmore, R.W. 1997. The Compositae: systematically fascinating but specifically neglected. *In*: *Genome Mapping in Plants*. (A.H. Paterson) ed. R.G. Landes Co. Georgetown, TX. pp179 - 191.

Kesseli, R.V., Ochoa, O., Michelmore, R.W. 1991. Variation at RFLP loci in *Lactuca* spp. and origin of cultivated lettuce. Genome 34:430-436.

Kesseli, R.V., Paran, I., Michelmore, R.W. 1994. Analysis of a detailed genetic linkage map of *Lactuca sativa* (lettuce) constructed from RFLP and RAPD markers. Genetics 136:1435-1446.

Kesseli, R.V., Witsenboer, H., Stanghellini, M., Vandermark, G., Michelmore, R.W. 1993. Recessive resistance to *Plasmopara lactucae-radicis* maps by bulked segregant analysis to a cluster of dominant disease resistance genes in lettuce. Molec. Pl. Microbe Interact. 6:722-728.

Kozik. A., Kochetkova, E., Michelmore, R.W. 2002. GenomePixelizer-a visualization program for comparative genomics within and between species. Bioinformatics 18:335-336

Landry, B.S., Kesseli, R.V., Farrara, B., Michelmore, R.W. (1987. A genetic map of lettuce (*Lactuca sativa* L.) with restriction fragment length polymorphism, isozyme, disease resistance, and morphological markers. Genetics 116:331-337.

Maisonneuve, B., Bellec, Y., Anderson, P., Michelmore, R.W. 1994. Rapid mapping of two genes for resistance to downy mildew from *Lactuca serriola* to existing clusters of resistance genes. Theor. Appl. Genet. 89:96-104.

Meyers, B.C., Chin, D.B., Shen, K.A., Sivaramakrishnan, S., Lavelle, D.O., Zhang, Z., Michelmore, R.W. 1998a. The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. Plant Cell 10:1817-1832.

Meyers, B.C., Shen, K.A., Rohani, P., Gaut, B.S., Michelmore, R.W. 1998b. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell 10:1833-1846.

Robbins, M.A., Witsenboer, H. Michelmore, R.W., Laliberte, J.F., Fortin, M. 1994. Genetic mapping of turnip mosiac virus resistance in *Lactuca sativa*. Theor. Appl. Genet. 89:583-589.

Shen, K.A., Meyers, B.C., Islam-Faridi, M.N. Chin, D.B., Stelly, D.M., Michelmore, R.W. 1998. Resistance gene candidates identified using PCR with degenerate primers map to resistance genes clusters in lettuce. Mol. Plant-Microbe Interact. 11:815-823.

Shen, K.A., Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Lavelle, D.O., Wroblewski, T., Meyers, B.C., Michelmore, R.W. 2002. *Dm3* is one member of a large constitutively expressed family of nucleotide binding site-leucine-rich repeat encoding genes. Mol. Plant Microbe Interact. 15:251-261.

Vision, T.J., Brown, D.G., Shmoys, D.B., Durrett, R.T., Tanksley, S.D. 2000. Selective Mapping: A strategy for optimizing the construction of high-density linkage maps. Genetics 155:407-420.

Witsenboer, H., Kesseli, R.V., Fortin, M., Stanghellini, M., Michelmore, R.W. 1995. Sources and genetic structure of a cluster of genes for resistance to three pathogens in lettuce. Theor. Appl. Genet. 91:178-188.

[R1]The same situation with Leah. See next comment
[R2]I understood Padma did these and she was left out in the poster. I see Barnaly in the co-author list, I think you need to revise the list.