**August 20, 2004**

**PREFACE**

This document is the third version of the SOL whitepaper and describes the ten-year goals of the International Solanaceae Genome Initiative. Since the last version, the following changes have been made: The whitepaper has been split up into different chapters that can be downloaded separately: the “vision” chapter, the “country writeups”, and the “standards” chapter. All parts will be updated regularly.

Happy SOL

# *The International Solanaceae Genome Project (SOL): Systems Approach to Diversity and Adaptation*

**Summary**

## The Questions*:*

Modern biology is expanding our view on life from the reductionist approach - analyzing individual components of biological systems, to the holistic view - integrating entire genetic programes and the complex events they dictate. Information generated by large-scale genomic sequencing has led to a major revolution in biological sciences through the revelation of all the genes required to encode major life forms. One of the biggest surprises has been that organisms that are evolutionary and morphologically distinct, share a very similar gene/protein content and even conserved linkage groups (e.g. human and mouse). Over the coming 10 years the International Solanaceae Genome Project (SOL) will integrate diverse disciplines and research groups from around the world to create a coordinated network of knowledge about the Solanaceae family aimed at answering two of the most important questions about life and agriculture:

***- How can a common set of genes/proteins give rise to such a wide range of morphologically and ecologically distinct organisms that occupy our planet?***

The corollary question of agricultural importance is:

***- How can a deeper understanding of genetic basis of plant diversity be harnessed to better meet the needs of society in an environmentally-friendly and sustainable manner?***

## The Family*:*

The family Solanaceae is ideally suited to address both of these questions. This taxon includes more than 3000 species many of which evolved in the Andean/Amazonian regions of South America in habitats that vary dramatically and include rain forests that receive more than 3 meters of rainfall annually, deserts with virtually no rainfall and high mountains with regular snowfall and sub freezing temperatures. The center of diversity of the Solanaceae is near the equator and thus species were undisturbed by the ice ages and have had time to accumulate adaptive genetic variation for extreme ecological niches.

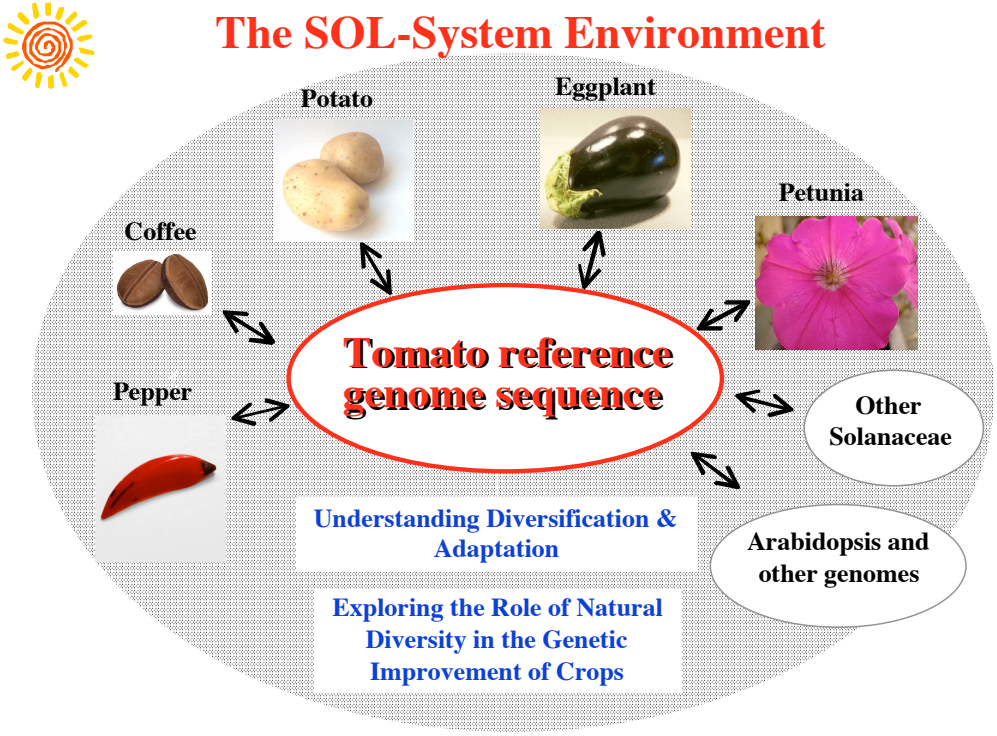
The Solanaceae is the third most important plant taxa economically and the most valuable in terms of vegetable crops. It encompasses the most variable of crop species in terms of their agricultural utility, as it includes the tuber-bearing potato (a food staple over much of the world), a number of fruit-bearing vegetables (e.g. tomato, eggplant, peppers, husk tomato), ornamental flowers (petunias, *Nicotiana*), edible leaves (*Solanum aethiopicum*, *S. macrocarpon*), and medicinals (e.g. *Datura, Capsicum*). Seeds can also be included in this list if we include the closely allied species coffee. Fruit and tubers are major contributors of vitamins, fiber, carbohydrates, and phyto-nutrient compounds in our diet. The nutritional importance of fruit and vegetables is reflected in current USDA recommendations of five or more servings of fruit or vegetables a day for a healthy diet. The World Health Organization and the United Nations Food and Agriculture Organization (FAO) recently launched an effort to enhance fruit and vegetable

consumption worldwide as low consumption is considered one of the top ten contributing factors to human mortality. The Solanaceae are unique in that multiple crop species in this family are major contributors to fruit and vegetable consumption and thus to our quality of life.

Solanaceae crops have been subjected to intensive human selection, allowing their use as models to study the evolutionary interface between plants and people. The unique and ancient mode of Solanaceae evolution, coupled with an exceptionally high level of conservation of genome organization at the macro and micro levels, makes the family a unique subject to explore the basis of phenotypic diversity and adaptation to natural and agricultural environments.

## The Concept*:*

The long-term goal of the SOL program is to create a network of map based resources and information to address key questions in plant adaptation and diversification. This will be done using the tools and philosophy of systems biology which is a multidisciplinary approach to tackle the complex interactions that occur at all levels of biological organization and their functional relationship to the organism as a whole. Moreover, from these studies we wish to provide a new outlook to how we value and utilize natural variation to impact the health and well being of humans in a more environmentally friendly and sustainable manner. Our international effort will not only impact Solanaceae biology but will also set the road map for implementing rational




strategies for improvement of other crop species that are important to human nutrition.

## Background on the Inception of SOL*:*

On November 3, 2003 researchers, from more than 10 countries, representing academic and government research labs, industry and extension/outreach specialists with interest in the Solanaceae met for a full day in Washington DC to kick off the 10 year initiative entitled “The International Solanaceae Genome Project (SOL)”. The forum united around a common set of

tools, populations and concepts with a firm commitment to work together to elevate our level of understanding of the network of interactions that lead to population diversity and adaptation. The agreed upon course of action for the first stage of this initiative was: 1) to obtain high quality sequence of the tomato genome as a reference for solanaceous plants as well as plants from other related taxa , 2) to display all data generated from around the world via a single virtual entry point for Solanaceae genomics, and 3) to establish a Steering Committee that will facilitate and coordinate research and funding for projects under the virtual umbrella of SOL.

## DIMENSIONS IN DIVERSITY AND ADAPTATION

### How can one genome code for diverse adaptive outcomes?

A basic descriptor in comparative genomics is the degree of homology between genome sequences and phenotypes of evolutionary divergent species. For example, comparisons between sequenced genomes of worms, yeast and flies revealed that many of the proteins encoded by these eukaryotic genomes are similar. These conserved features, which are shared by many of the living organisms, reflect the unity of life. At the same time, the diversity of characteristics among the earth's millions of species is staggering. Understanding the genetic basis of traits that distinguish closely related taxonomic groups is one of the great challenges in biology. Some of the differences between taxa reflect neutral ticks of the molecular evolutionary clock while others are associated with features of adaptation that enhance survival and reproduction in unique ecosystems. Natural selection is the ultimate determinant of adaptation where over evolutionary times, mutations, recombination and the element of drift resulted in descendent species that are dramatically different from their last common ancestor.

With the sequencing of the human genome, the mouse and the chimpanzee, the quest to discover the genetic basis of phenotypes that distinguish us is gaining momentum. Virtually all the protein-coding genes in human align with homologues in mice and the two genomes show a very high level of gene order conservation (synteny). Despite nearly 100 Myr of divergence, mouse and humans share a nearly gene-for-gene match between their genomes, raising the question of how a common set of proteins could lead to such dramatically different outcomes as a mouse and a human. Comparisons of the two genomes have begun to identify regulatory genomic regions, which substantially outnumber the
proteins repertoire, with potential functional roles in controlling gene expression and chromatin organization. This question of "which sequences make us human" must wait until more mammalian genomes, particularly apes, are fully sequenced to reveal new leads to explore the complexity of phenotypes.

It is now becoming clear that most of the traits involved in adaptation and diversification are polygenic and affect continuous or quantitative adaptive phenotypes. Moreover, most of these genetic changes are not "loss of function" mutations of the type induced in the laboratory. Rather, they are genetic variants that change the function of the proteins for which they code, or perhaps more often, the temporal and special expression of those genes. Understanding the nature of the genetic changes underlying adaptation and diversification, is a prerequisite to understand the basis of life and evolution. This will also enable us to fully appreciate and utilize the natural variation around us to better adapt plants to the need of humans in an environmentally, sustainable manner. Due to a number of inherent attributes, plants are ideal model for resolving the genetic basis of quantitative traits related to adaptation and diversification (the cloning of quantitative trait loci (QTL) was achieved in plants first). Some of these features include short generation time, large families and tolerance to inbreeding. Furthermore, the ability to generate

segregating populations from divergent plant species that are adapted to different growth conditions facilitates the mapping of numerous QTL that affect fitness. As genome sequences and phenotypes become available from a range of related plant species, we will be able to better understand how a common set of genetic building-blocks regulate the diverse outcomes that affect adaptation.

## The Solanaceae family is an ideal model to explore the basis of diversity and adaptation

From the rich diversity of ~300,000 higher plant species on our planet, the genomes of a single dicot (Arabidopsis) and a single monocot (rice) have been sequenced. Solanaceae represent a unique portion of the family tree of plants and its sequencing and exploration will enable comparative analysis for the discovery of distinct and common aspects of plant evolution. The family Solanaceae is anchored in a section of plants' evolutionary tree that is distant from both Arabidopsis and rice (Figure 1). Contained within these clades (Asterids I, II) are not only the Solanaceous crops, but also a number of other major crop plants such as coffee, lettuce, safflower and sunflower. The Solanaceae family includes more than 3000 species where the genus *Solanum* is the largest one with approximately 1500 species. Extensive current and ongoing knowledge exists about systematics of the family, including a recent generic conspectus and up-to-date family-wide molecular phylogenies (ref).

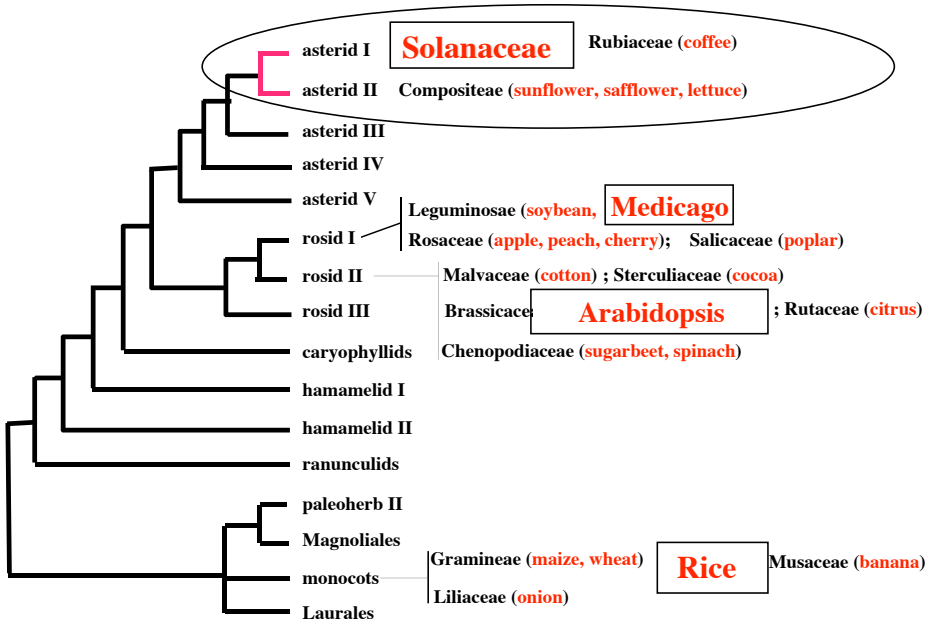



***Figure 1****: Solanaceae represent a unique portion of the family tree of flowering plants and its sequencing will enable comparative analysis for the discovery of distinct and common aspects of plant evolution.*

The Solanaceae is the third most valuable crop family exceeded only by the grasses (e.g. rice, maize, wheat) and legumes (e.g. soybean), and the most valuable in terms of vegetable crops. It also encompasses the most variable of crop species in terms of their agricultural utility, as it

includes the tuber-bearing potato (a food staple over much of the world), a number of fruit-bearing vegetables (e.g. tomato, eggplant, peppers, husk tomato), ornamental flowers (petunias, *Nicotiana*), edible leaves (*Solanum aethiopicum*, *S. macrocarpon*), and medicinals (e.g. *Datura, Capsicum*). Seeds can also be included in this list if we include the closely allied species coffee. Multiple important species in our family are major contributors to fruit and vegetable consumption and thus human health.

Solanaceae species thrive in some of the most diverse natural habitats that include rain forests that receive more than 3 meters of rainfall annually, deserts that receive virtually no rainfall and in which plants survive entirely on moisture from fog, to the high elevation Andean mountains where UV radiation is high and temperature plummet to sub freezing on a regular basis (Figure 2). Solanaceae species range in habit from tall forest trees and woody lianas to tiny annual herbs. Being near the equator and thus undisturbed by the ice ages, the Solanaceae have had time to adapt to diverse niches. Yet, despite this high level of phenotypic variation and ecological adaptation, the Solanaceae share very similar genomes and gene repertoire.


***Figure 2:*** *Solanaceae species evolved and are adapted to some of the most diverse and extreme habitats on earth.*

## The Solanaceae genome is uniquely conserved

In plants, the use of comparative genetic molecular mapping has revealed a high level of conservation of gene content and order within the grasses, crucifers, legumes and Solanaceae species. Sequencing of Arabidopsis and rice has shown that more than 80% of the genes that have been annotated in Arabidopsis were also found in rice; however, nearly 50% of the predicted rice genes do not have a match in Arabidopsis thus providing a basis for the specificities of the two organisms. Numerous episodes of polyploidy within both the grasses and Brassiceae have led to segmental duplications, selective gene losses and significant genome reshuffling. As a result, species in the grasses and crucifers are characterized by different chromosome numbers coupled with extensive loss of microsynteny between the paralogous segments of Brassica chromosomes, and between those and their Arabidopsis homoeologs. The Solanaceae family is unique in that there have been no large-scale duplication events (e.g. polyploidy) early in the radiation of this family. The polyploidy events (e.g. tetraploid potatoes and tetraploid tobacco) are all recent events and diploid forms of both of these species are still in existence. As a result, microsynteny conservation amongst the genomes of tomato, potato, pepper and eggplant is very high (Fig X

Technical Document ). This allows to predict regions between genomes that are identical by descent and to study the evolution of sequence and function of orthologous genes – a key to understanding diversification and adaptation. The highly conserved genome organization, both at the macro- and microsyntenic levels, allows extending the information basis beyond the individual species thus creating a common map-based framework of knowledge. Hence, the first goal of SOL is to determine, with great precision, the nucleotide sequence of the tomato genome and link it to the Solanaceae map.

The tomato map-based genome will provide a reference to interpret the sequence organization of other Solanaceae crops as the basis of understanding how plants diversify and adapt to new and adverse environments. Tomato was selected as a reference since it provides the smallest diploid genome (950 Mb) for which homozygous inbreds are available, as well as an advanced BAC based physical map to start the sequencing. Tomato is also the most intensively researched Solanaceae genome with simple diploid genetics, short generation time, routine transformation technology, and availability of rich genetic and genomic resource. The tomato genome encodes approx. 35,000 genes, which are largely sequestered in contiguous euchromatic regions corresponding to less than a 25% of the total DNA in the tomato nucleus (220~250 Mb of gene rich regions). Presently the *Solanaceae Genome Network* (SGN; http://www.sgn.cornell.edu/) hosts multiple information from diverse sources around the world in a (MySQL) relational database. SGN currently contains approximately 200,000 gene/EST sequences from tomato, potato, eggplant, pepper and petunia. As part of the SOL project, SGN will merge or integrate with other related plant genome databases to provide a virtual workbench to explore phenotypic diversity in the highly conserved genomes of the Solanaceae.

## Networks in physiology and biochemistry

The underlying genetic diversity in the Solanaceae is arrayed on a broad canvas of phenotypic variation, where the richer the genetic pool, the more diverse the resulting cellular processes and organismal complexity. Accordingly, genome-analysis projects that are undertaken within a framework of genetically diverse germplasm will, by definition, result in a far more profound and comprehensive understanding of biochemistry and physiology. Members of the Solanaceae collectively comprise precisely the germplasm diversity, and crucially the phenotypic diversity, that will prove essential for pushing forward the frontiers of plant biochemistry and physiology.

The plant kingdom is estimated to produce approximately 200,000 metabolites, many of which play specific roles in allowing adaptation to specific ecological niches. The spectrum of Solanaceous species collectively occupy almost every conceivable niche, presenting a remarkable opportunity to access the breadth of phenotypic variation and underlying biochemical diversity. In addition, many Solanaceae crops have been selected for specific characteristics related to biochemical composition and physiological traits. These include composition of sugars, organic acids, volatiles, a structurally diverse array of secondary metabolites and tolerance of environmental stresses. As a result of this biological specialization, a formidable knowledge base of plant biochemistry and physiology has been accumulated, in which members of the Solanaceae have been used as pioneering species in discovery-based research across the metabolic map. Examples include cell wall and storage polysaccharide synthesis and metabolism, volatile production, vitamin biosynthesis, biosynthesis and action of the hormones ethylene and brassinosteroids and biosynthesis of flavonoids and carotenoids. Fruits of tomato and peppers accumulate high amounts of carotenoid pigments. Carotenoids with beta-ring, known as provitamin A, are indispensable in the human diet because they are the only source of vitamin A. Epidemiological studies indicate carotenoids to be preventive agents against specific diseases such as prostate cancer (lycopene) and age-related macular degeneration (lutein/zeaxanthin). Carotenoids show protective activity *in vitro* and *in vivo* against a variety of degenerative diseases, possibly through their activity as antioxidants.

Solanaceous species have been an excellent source of all major classes of secondary metabolites, including alkaloids, terpenoids, flavonoids, amino acid and fatty acid derivatives, and sugars. For example, atropine, found in Datura (devil's apple) was used for many years as a shamanistic tool and now has an important role in modern medicine. Another example is the capsaicinoids compounds, which are synthesized in the placental septum of the pod or fruit in pungent *Capsicum* species. These compounds, which contain moieties derived from both the phenylproanoid and amino acid/fatty acid pathways, underlie the familiar burning sensation of hot peppers, and their human uses range from traditional painkilling to chemo-preventative agents in cancer treatment. While some progress has been made in elucidating a few of the secondary metabolites biochemical pathways, there are many other types of metabolites in the rich biodiversity of the Solanaceae for which little information is available regarding their synthesis and physiological function.


***Figure 3****: The genomic tools resulting from SOL will facilitate comparative biology of common phenotypes such as fruit development and carotenoid content.*

Another attribute of Solanaceae species that makes this taxa an attractive target of study of secondary metabolites is the presnce of specialized structures, easily isolated, that are a major site of production, storage and secretion of certain such compounds. These are the glandular trichomes found on the surface of most aerial organs and consist of long stems with a two- or four-celled gland at the tip (Figure 4). In tobacco, these glands produce and secrete the diterpenes cembratriene-diol (CBT-diol), which plays a major role in the interaction between the plant and colonizing aphids. Similar anatomical structures in *Lycopersicon* are the site of biosynthesis of monoterpenes and sesquiterpens, acylsugars, and methylketones. Procedures for separating these glands from the leaves and then isolating chemicals, proteins or mRNAs from them have been developed, and so large chemical and molecular databases (e.g., EST) have been obtained for glands from several Solanaceae species. Combined with genomic sequence information, these tools to study the biology of a single type of cell are extremely powerful and will guarantee quick

progress in elucidating the biochemical pathways leading to the synthesis of the myriad secondary compounds.

The scientific literature extends far beyond this shortlist, which could be readily extended to include the many physiological processes for which members of the Solanaceae represent model species, including source-sink relations, temperature, drought - and salinity stress tolerance. In addition Solanaceae species are in many cases experimentally far more tractable and attractive as models for biochemical and physiological research than smaller organisms such as Arabidopsis. It is important to recognize that members of the Solanaceae represent model species for large sectors of plant biochemistry and physiology and will continue to occupy this central position in plant science.

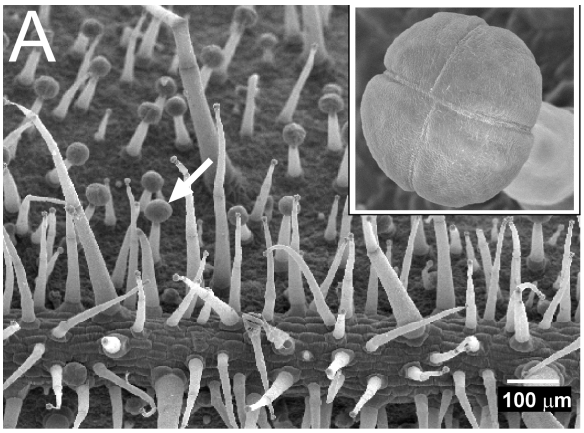



***Figure 4****: Scanning electron micrograph of the lower surface of the leaf of L. hirsutum. Type VI glands (four-celled; identified by white arrow, and in inset) predominate.*

It is clear that biological hypotheses will continue to be developed and tested in the Solanaceae. The genetic, phenotypic, biochemical and physiological diversity, coupled with the wealth of literature and expert knowledge base, when supported by a genome sequence, make this taxonomic group a clear front-runner when electing a model group of organisms in which to develop the paradigm of plant systems biology.

## Fruit and tuber biology provide a key to understand agricultural yield

In higher plants, sugars are produced photosynthetically, primarily in the leaves where carbon is fixed. These sugars are transported to other plant organs that are involved in active growth and development (such as roots, tubers, flowers, seed and fruits) where they are metabolized or stored. Because of the basic importance of the source-sink balance to human food and nutrition, the partitioning of assimilates is of key concern to basic and applied plant biology. Understanding of the genetic mechanisms regulating source-sink relationships is a prerequisite for optimal modulation of yield through crop design and breeding. The Solanaceae are a unique model to explore the basis of yield: in tomato, peppers and eggplant fruits represent the major sink tissue while potato tubers are the underground stems that are uniquely modified as starch-storing organs. The physiological and developmental processes that regulate sink strength are expected to be largely independent of the anatomical identity of the organ. Comparison of the phylogenetically close potato with the fruit bearing Solanaceae is an active area of research aimed at revealing molecular modulators and communication networks that connect source with sink.

By anatomical definition the fruit is a mature ovary and therefore typically includes carpel tissues in part or in whole. Evolutionary pressures have resulted in a variety of developmental manifestations of fruit tissues, resulting in structures that range in design and function from hardened fruit capsules or pods that forcefully expel seed at maturation, to forms optimized for seed movement by wind, water, animal fur or gravity, and finally those implementing developmental programs yielding succulent and flavorful tissues for organisms that consume and disperse the associated seed. In recent years, particular emphasis has been placed on tomato as an especially tractable system for molecular genetic analysis of fleshy fruit development and ripening. Fleshy fruits such as tomato undergo a ripening process in which the biochemistry, physiology and structure of the organ are developmentally altered to influence appearance, texture, flavor and aroma in ways designed to attract seed dispersing organisms. The importance of tomato as an agricultural commodity has resulted in decades of public and private breeding efforts, which have yielded numerous genomic resources including mapping populations, mapped DNA markers, extensive EST collections and publicly available microarrays. The sequence conservation of Solanaceae genes facilitated the use of the tomato microarrays to explore gene expression in fruits of other species (Figure 5) thus revealing a high commonality in fruit development among the family members.

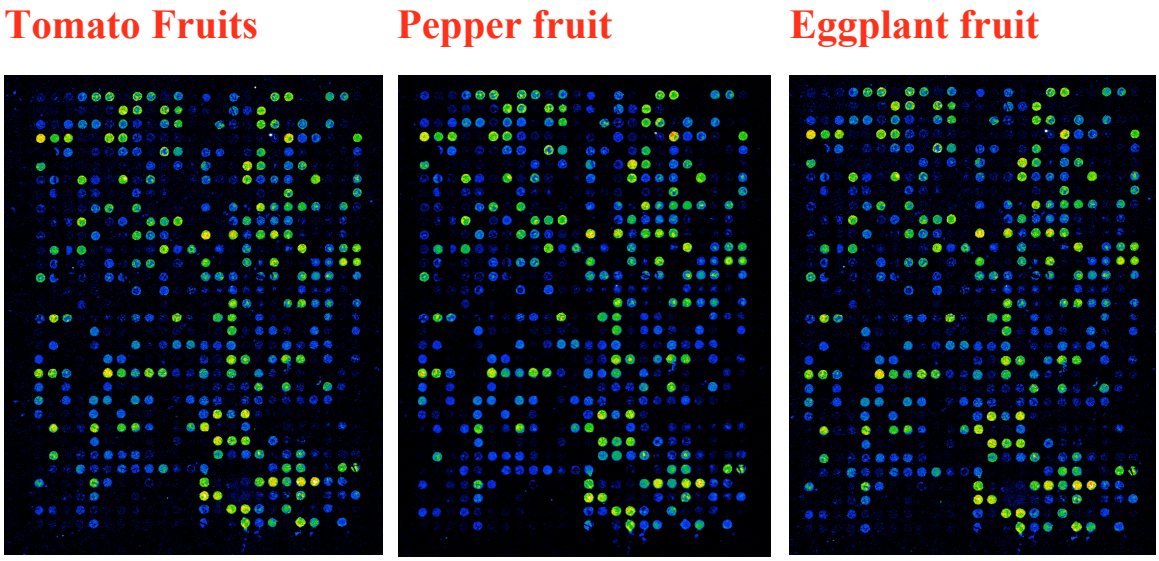



***Figure 5****: The high degree of sequence similarity among Solanaceae facilitates cross utilization of microarrays resources and tools - thus demonstrating that a single genome sequence will have a broad impact across the family.*

Tuber formation is the most critical physiological function involved in potato production. It involves a number of biological processes at the stolon tip, such as carbon partitioning, starch biosynthesis, signal transduction, and meristem determination. It is hypothesized that each physiological stage in tuber development is controlled by specific gene interactions. A number of genes has been found to be associated with early tuber formation including tubulins, S-adenosylmethionine decarboxylase, MADS box genes, acyl carrier protein thioesterase, and lipoxygenases. Due to the complexity of the tuber development process, it is difficult to fully understand the molecular mechanism of tuberization through the examination of individual genes.

A firm understanding of the molecular mechanisms that regulate cell growth during tuber development are critical to address the improvement of numerous tuber traits including starch and protein content, internal quality, tuber shape, fast bulking, and early maturity. Given the high degree of similarity of genomes among the Solanaceae it will be especially interesting to assess how these genes have changed through evolution and selection to bring about the developmental and anatomical diversity displayed between members of this family. Through comparative genomics approaches, it should be possible to determine the evolutionary changes that enable plants like potato to produce tubers.

## Diversity in Solanaceae defense responses

The Solanaceae species provide many unique advantages for studies of plant defense responses. First, they are hosts to many well-characterized and economically important pathogens and insects (viruses, bacteria, fungi, nematodes, chewing/sucking insects). Second, many of the Solanaceae (i.e., tobacco) have large leaves that are especially amenable to pathogen/insect assays. Large leaves are also important for the isolation of proteins that are present in low abundance in plant tissues. Third, Solanaceae species, unlike Arabidopsis, possess multicellular trichomes that play an important role in insect defense. Fourth, several of the Solanaceae species permit high efficiency of Agrobacterium-mediated transient expression that expedites characterization of disease resistance genes and defense-signaling genes (i.e. this approach has been especially useful in tobacco and tomato). Fifth, the family encompasses extensive (natural) genetic diversity that has been instrumental in the identification of host resistance genes and other defense factors. Sixth, over 100 disease resistance (*R*) genes are known in various Solanaceous species and many have been cloned (see below). Finally, virus-induced gene silencing (VIGS) has, to date, been found to be most efficient in the Solanaceae (*Nicotiana benthamiana*, tomato, tobacco).

Tomato was the first plant from which a "gene-for-gene" class of *R* gene was cloned and more than 12 R genes now have been isolated from tomato. These include genes conferring resistance to fungi, nematodes, aphids, bacteria and viruses. Additional R genes, some of which have been shown to function when transferred into tomato, have been isolated from related species including pepper, potato, and tobacco. Remarkably, four *R* genes from tomato are unique (i.e. *Pto, Ve, Asc, Cf* genes). No similar *R* genes have been identified yet from Arabidopsis, rice or any other plant species

In addition to yielding many *R* genes, tomato and related species have been used as a model system for other important advances in host defenses. Some of the major accomplishments include: 1) Discovery of systemin (present only in Solanaceous species). Systemin, which plays a key role in defense against herbivorous insects, was the first peptide hormone discovered in plants; 2) Isolation of the first mitogen-activated protein kinases (WIPK and SIPK); 3) Isolation of the first nitric oxide synthase gene from plants (iNOS); 4) Isolation of the first salicylic acid-binding proteins (SABPs) from plants; and 5) Providing some of the first insights into the mechanism of viral cross-protection and the development of virus-induced gene silencing (VIGS).

The many cloned *R* genes from tomato and other Solanaceae species and the abundance of information and resources related to diverse plant defense responses provides an unparalleled foundation for using the tomato genome sequence to increase our understanding of plant disease resistance and susceptibility. Finally, it is important to note that potato, pepper, and tobacco are also susceptible to many of the same pathogens as tomato (e.g. *Phytophthora* spp., *Fusarium* spp.). The availability of genome sequences for a range of Solanaceae plants will broadly benefit

the study of plant defense responses and their role in adaptation to natural and agricultural environments.

### What is the role of natural diversity in the genetic improvement of plants?

On the agricultural side, Solanaceae provide an opportunity to explore natural diversity as a sustainable resource to enrich the genetic basis of cultivated plants with novel genes that increase productivity. This approach is both a complement and an alternative to the GMO strategy for improving the quality and quantity of food output.

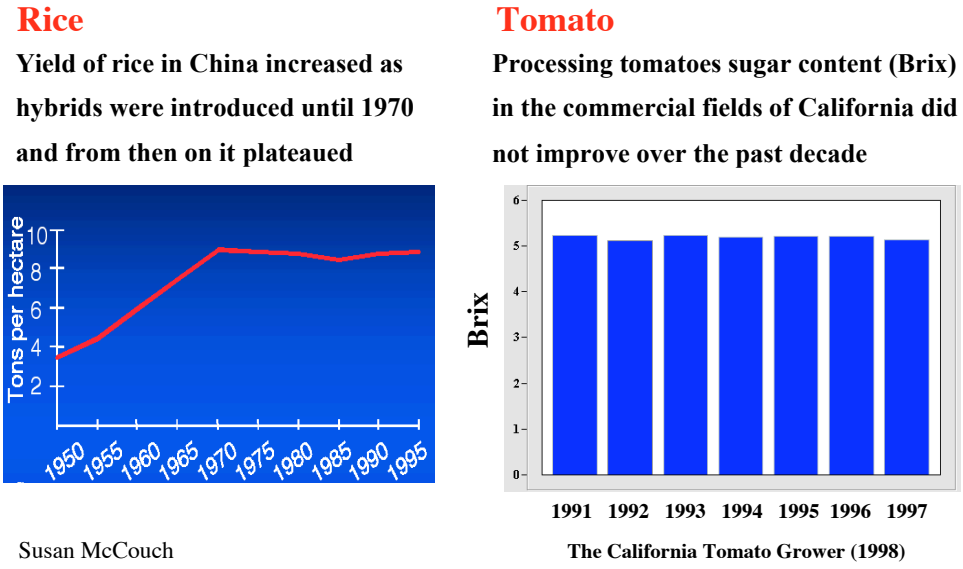



***Figure 6**: Depletion of the genetic variation for yield-associated traits in elite germplasm is the major factor for the slow progress in improving yield potentials.*

Plant evolution under domestication has lead to increased productivity, but at the same time has narrowed the genetic basis of crop species (Figure 6). The challenges facing modern plant breeders are to develop higher yielding, nutritious and environmentally friendly varieties that improve our quality of life by not harnessing additional natural habitats to agricultural production. Genetic variation is the engine that propels breeding to meet future challenges. The observation that wild genetic resources can contribute to crop improvement, combined with the alarming rate at which locally adapted landraces are being lost and at which natural habitats are being damaged, has led to the establishment of large germplasm collections. These seed banks initiate collection missions, maintain and characterize accessions, and make them available to the breeding community. The task facing us is to devise the tools and concepts that would allow us to rapidly utilize the genetic potential that exists in the rich genetic diversity of wild species.

To enhance the rate of progress of such introgression breeding, the Solanaceae community pioneered 15 years ago the development of permanent genetic resources, which comprise of marker-defined genomic regions taken from wild species and introgressed onto the background of elite crop lines. These introgression populations (lines with individual introgressed chromosomes or inbred backcross populations with a few small segments of wild species DNA in an otherwise

isogenic background), which are now available for a number of tomato and pepper species, serve as unique powerful reagents for the discovery and characterization of genes that underlie traits of agricultural value (Figure 7). The populations are being phenotyped by numerous labs for a range of yield-associated traits, including biotic and abiotic stresses, as well as for metabolic profiles of hundreds of distinct compounds. A user-friendly bioinformatic management system has been established such that this QTL data is available to the community over the Internet. The challenge facing SOL for the coming years is to develop methodologies that will enable genomic information to be associated with phenotypes of interest for crop improvement. The framework for this data organization is the highly conserved genetic map of the Solanaceae that will allow us to extend the information basis beyond the individual species. These exotic IL populations make a wide array of previously unexplored genetic variation rapidly available to plant breeders and geneticists. Either in combination with GM technology or without it, exotic genetic libraries represent a dynamic new resource base that can substantially enrich traditional crop improvement programs for many years to come. The vision uniting the SOL is to take advantage of the distinctive exotic breeding populations as a springboard to unite systems biology concepts with the practical discipline of plant breeding. This effort will not only impact Solanaceae biology but will also set the road map for implementing rational strategies for improvement of other crop plants that are important to human nutrition.

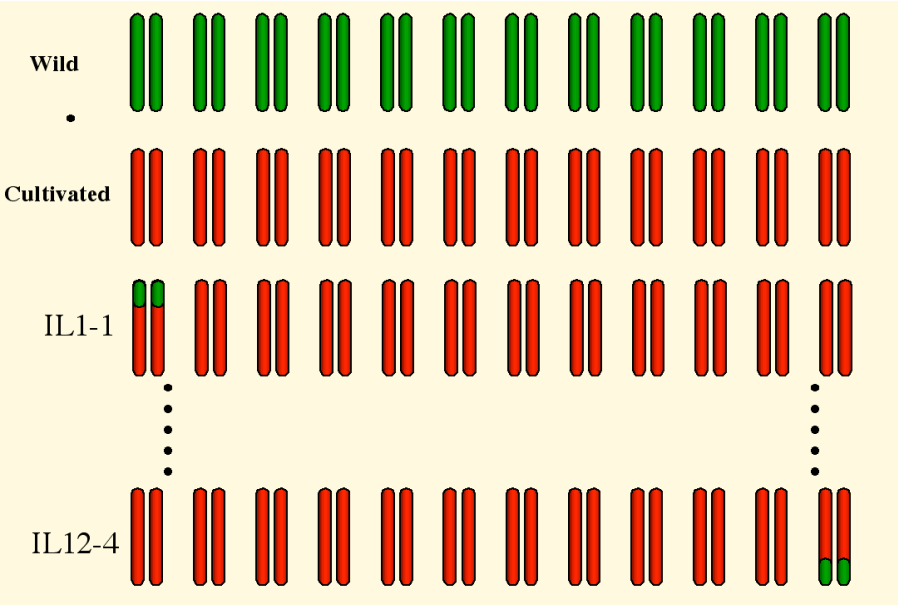



***Figure** 7: The construction of a set of interspecific introgression resources in elite genetic backgrounds will make a range of biodiversity available for breeders. This will allow scientists to unravel the basis of hidden wild genes for productivity, adaptation and other phenotypes that affect human well-being.*

## How can a system-level-approach in the Solanaceae help in resolving some of life's complexity?

The genomics era has largely matured around sequencing and studying the function of genes within a single organism (e.g. Drosophila, yeast, Arabidopsis). The tools for study most often involve under-expression (e.g. gene knockouts) or over-expression to deduce the function of genes. More and more, the approach taken is to study the response of the entire set of genes/proteins to perturbations in the expression of single genes. In reality, evolution of form and adaptation occurs not through the radical loss of gene function, but through the modification of gene function through changes in protein structure/activity and quantitative, temporal and special changes in gene expression. Moreover, the genetic variation that fuels these adaptive changes has passed through thousands if not millions of years of selection – and cannot be readily replicated in the laboratory. Through segregation and recombination that reshuffles the genomes of diverse donor parents in genetic crosses, the Solanaceae community has generated populations of offspring with many combinations of allelic variants, many of which date back thousands of years and honed through selective pressures. Such populations facilitate the study of traits that are determined by many genes, which almost always interact with each other and with the environment. Systems biology requires quantitative data on appropriate germplasm resources that are of high quality and all-inclusive and taken simultaneously at different stages of development from defined cell lineages. Looking ahead, genomics, quantitative genetics and computing sciences will be integrated in a comprehensive strategy of designing, modeling and analyzing complex biological data.

Systems biology is an approach to tackle the complex interactions that occur at all levels of biological organization and their functional relationship to the organism as a whole. This new branch of biology, though yet somewhat fuzzy, is gaining in popularity in recent years, as modern genomic technologies are capable of generating comprehensive data sets on a wide spectrum of attributes of the organism. The high throughput technologies describe components of living systems that include whole genome DNA sequence, RNA transcription and processing, protein synthesis, post translational modifications, the formation of protein complexes, metabolic networks as well as simple and complex morphological phenotypes. Progress in future biological research will depend on our ability to find ways to start tying together the independent components into higher order complexity with multiple dimensions. It is also becoming apparent that multidisciplinary research efforts, involving the increased input of chemistry, physics, statistics, mathematics and computing sciences, is crucial for the success of a multifactorial approach. The key question is how to obtain the greatest knowledge about complex biological systems through clever experimental designs, models and methods of analysis.

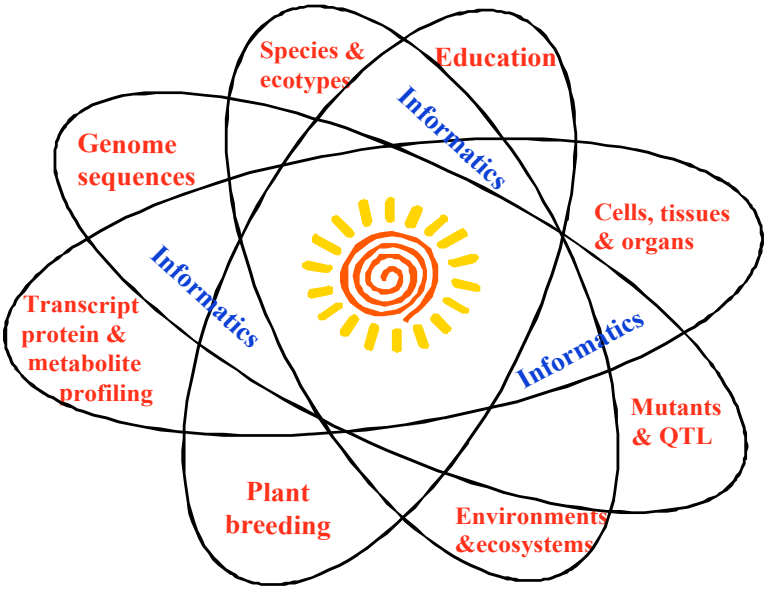



***Figure 8****: The systems approach aims to tackle the complex interactions that occur at all levels of biological organization and their functional relationship to the organism as a whole.*

## How can bioinformatics evolve to accommodate systems biology on the scale of the SOL project?

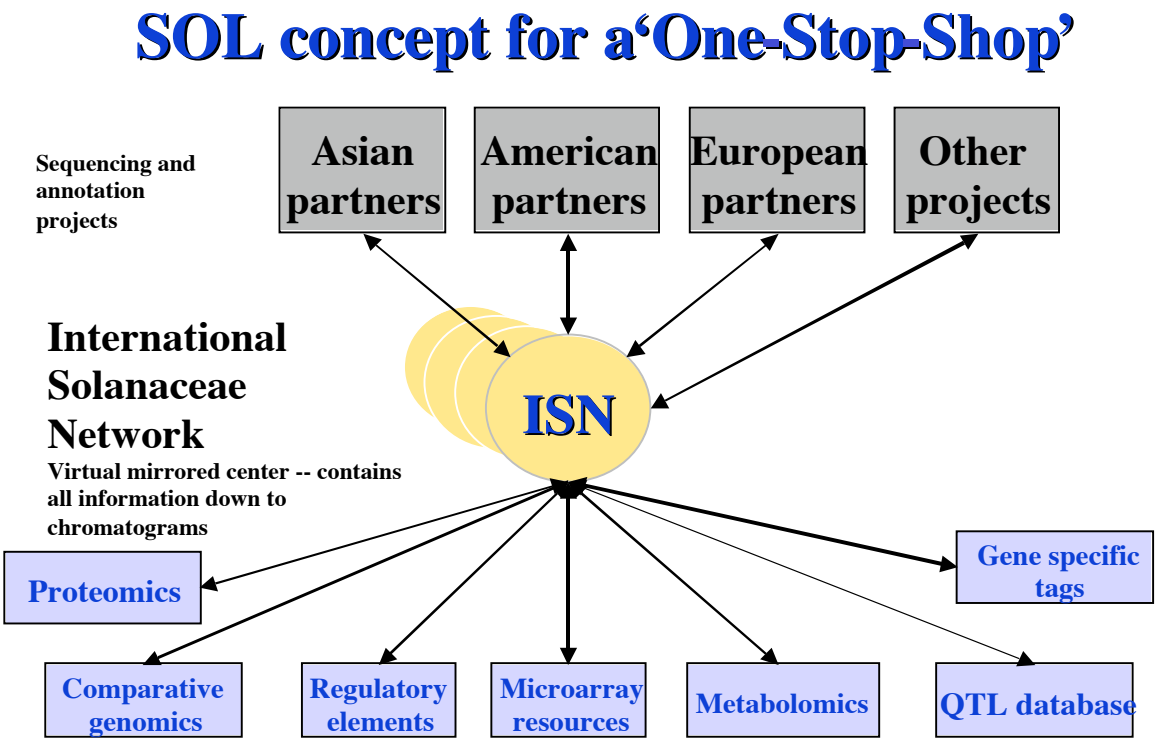
Presently, bioinformatics has most of its efforts focused on storage and retrieval of information in a format usable by as many scientists as possible. In the future, bioinformatics must evolve into a network of information that will become the driving force for creative ideas. In this new paradigm, the bioinformatics network (and its associated tools), will lead investigators into new hypotheses that can be tested *in silico* or through predictive laboratory experimentation. The proposed 10 years international SOL initiative will turn this vision into reality by making sure that the independent research conducted in the participating laboratories will focus on defined set of similarly grown genotypes, tissues and cells in a manner that would facilitate multifactorial analysis of combined datasets.

There is strong and broad agreement in SOL to develop a single virtual entry point (One-stop-shop) for Solanaceae genome sequence and related genomics and systems information. The systems biology framework that SOL has adopted poses new challenges to bioinformatics and in particular to data management. Bioinformatics is the discipline that analyses large-scale genome data and makes predictions on entities such as gene structure, gene function, protein structure, expression levels, and phenotypes and it also deals with data management and visualization issues. In the past few years, bioinformatics has been very successful in solving very specific problems such as gene finding and protein domain identification. Systems biology builds on these successes as the discipline that analyzes the networks that can be 'overlaid' onto these and other types of data, such as metabolic networks, gene expression networks, regulatory pathways, developmental pathways and even networks of interactions among organisms - and makes inferences about how these different networks interact and predictions about outcomes when these networks are disturbed. Thus, Systems biology allows a real understanding of what makes a living system tick. Obviously, the description of these extensive networks is an information-intensive task. A prerequisite to work intelligently with such data is a universal way to describe these networks. Second, we need these data accessible in one location (not necessarily physical) to be able to integrate and analyze them. Therefore, the ultimate aim of the bioinformatics strategy is to provide scientists with a knowledge environment that will enable them to generate new hypotheses in an exploratory fashion. The Solanaceae bioinformatics effort will work with the major model organism databases to make a more unified and advanced querying of information possible.

The importance of the bioinformatics part of this project to the overall success of the project cannot be overestimated. Bioinformatics is really the glue that holds a project of this magnitude together. Our view is therefore that such an effort should be closely coordinated, more so than in other genomics projects in the past. It is particularly important to establish, early on, standards and coordinates to ensure efficient use of resources, generation of consistent, high quality data, and to prevent major duplications of efforts. The collaborative spirit in the Solanaceae community, that was evident at the recent Solanaceae meeting in Washington, will also be reflected in the bioinformatics strategy. At the meeting, the representatives of a number of bioinformatics centers have agreed that the most desirable deliverable of such a project is a unified data-center on the web that holds all data and connects to other model organism databases. Although many projects have advocated an open source approach in the past, in practice, few projects really achieve a true collaborative environment. The prerequisites for a collaborative environment are common technical standards and a basic infrastructure to share codes and data. The centerpiece of the collaboration will be a repository where all centers share their pipelines, programs and data to the extent possible in common standard formats. This common repository will be the basis for database implementations at the participating

bioinformatics centers that mirror the complete information available. The standards will be established by a bioinformatics committee that is already working on standards for the first phase of the Solanaceae project - the sequencing of the tomato genome.

***Figure 9***: *The concept for organization of the international SOL bioinformatics network*



# *SOL Goals and Objectives*

Over the coming decade the International Solanaceae Genome Project (SOL) will create a coordinated network of knowledge about the Solanaceae family aimed at answering two of the most important questions about life and agriculture:

***How can a common set of genes/proteins give rise to such a wide range of morphologically and ecologically distinct organisms that occupy our planet?***

The corollary question of agricultural importance is:

***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?***

**Key components and milestones for the SOL project:**

1) Sequence the reference tomato genome on a BAC by BAC basis (see the technical sequencing paper) tying this information together in a common framework with Arabidopsis and rice.

2) Develop deep EST databases from various Solanaceae tissues and shotgun genomic sequencing of other Solanaceae with data integration. Align EST/shotgun sequence from other Solanaceae species against the contiguous tomato genome sequence so as to provide the basis for identifying sets of orthologs across as many species as possible. In addition, this process will allow the identification of both conserved genes (and sequence motifs – e.g. promoters) as well as genes/regions of the genome that are evolving rapidly under positive selection and may hence be related to species diversification

3) Complete a set of high resolution comparative genetic maps for solanaceous species and related taxa (e.g. coffee, based on the analysis of Conserved Orthologous Set (COS) markers).

4) Construct a set of interspecific introgression resources (e.g. introgression lines, backcross inbred lines etc.) for all Solanaceae crop species in order to provide the genetic material from which genes/QTL underlying species divergence, evolution of biochemistry, environmental adaptation, evolution of mating systems etc. and the development of associated phenotypic databases can be studied.

5) Establish saturation mutagenesis genetic resources, methods for high throughput identification of mutants associated with a specific sequence and comparative biology of developmental pathways between taxonomic groups.

6) Construct a comprehensive phylogenetic and geographical distribution information network that can be used for both researchers and educators – including online information

about natural distribution, habitats, species images, germplasm banks, botanical gardens and herbarium collections.

7) Application, on a basis of a broad phylogenetic sampling, of transcription, proteomic & metabolic profiling to begin understanding the range of evolution of plant chemicals and to provide a base line for follow-up studies in chemical ecology.

8) Application of non-destructive 'Real Time' physiology and phenotyping platforms including imaging techniques to dynamically monitor changes in cells and organs throughout development.

9) Improve the efficiency of plant breeding based on the use of wild species variation, marker assisted selection, and mutagenesis. This component of the SOL project will be conducted in collaboration with industries that rely on Solanaceae species in a manner that would allow dissemination of the results to the entire community.

10) Develop an education package, based on the Solanaceae, with which the public and students (of all ages) can become more aware of the natural diversity of plant, the process of domestication, and the role of these in the manner in which we utilize and sustain agriculture.

11) Develop an international bioinformatics platform, which will not only allow storage and retrieval of all information from the SOL project, but will simultaneously develop and apply new algorithmic tools which will promote and facilitate a systems approach to the study of this group of plants.

## Organization and Coordination of the SOL project on an International Level:

The Steering Committee of SOL will function as a virtual 'umbrella organization' for the project by coordinating and facilitating research along the vision proposed in this white paper. The members of the committee (24 – one for each tomato chromosome) will include representatives of the participating countries as well as scientists that are expert in specific fields (such as: systematics, bioinformatics, profiling technologies, education and theory and practice of systems biology). Committee membership will rotate as the project moves forward in time and knowledge **- from sequence to systems -** and the Committee will meet once a year in the International SOL Genomic Symposium (see Technical Sequencing Document for further details of organization). SOL will encourage and actively seek additional scientists, countries and funding agencies to participate in this expedition into higher order organization of biological information. Members of the scientific community, from around the world, who would like to submit local or international collaborative research grants under the umbrella of SOL would receive a warm letter of support from SOL. Grants that will be endorsed by SOL will include a section detailing how the generated data will be integrated with SOL bioinformatics system.