University of Naples Federico II
Department of Agriculture

Agrobiology e Agrochemstry Ph.D.
XXVII Cycle (2012-2015)

THE GENOME SEQUENCE OF *SOLANUM COMMERSONII* DUN., A VALUABLE SOURCE OF RESISTANT GENES FOR POTATO BREEDING

Ph.D. Dissertation
by Dr. Felice Contaldi

TUTOR:
Prof. Domenico Carputo

Index

1.1 The potato, its economic importance and breeding targets

The cultivated potato, *Solanum tuberosum* L., originated in the Andean region of South America. Since its introduction in Europe in 1570, it has been spread globally and nowadays it is the fourth most important crop in the world, following maize, rice and wheat (www. FAO.com). Potato is a rich source of energy, with a starch content that accounts for 80% of the tuber dry weight and with a high content of quality protein and vitamin C (Scott et al., 2000). In addition, it yields on average more food, energy and protein per unit of land than cereals (Horton, 1988). Potato is a crop of temperate climate and it is sensitive to frost. The young plants grow well at 24 °C. Late growth is favoured at 18 °C. Tuber production is maximum at 20 °C and decreases with rise in temperature. At about 30 °C tuber production is heavily compromised.

Due to its agronomical plasticity, the potato is grown in more countries and agro-ecological zones than any other crop. Based on a projection of human population growth rate and some economical parameters such as increase in household income and a more intensive participation of women in the labour force, it is estimated that potato production worldwide will increase to 403.5 millions metric tons per year by the year 2020 (Scott et al., 2000). In developing countries the production of potato will increase from 94.3 to 194.0 millions metric tons per year. China, India and regions including South Asia, West Asia/North Africa and Sub-Saharan Africa are expected to have the highest increase in potato production (Scott et al., 2000).

Today, considering the EU situation, Italy ranks ninth in terms of area invested and sixth in terms of production. After tomato, in Italy potato is the most widespread crop, with a production of about 1.3 million tons and an area down by about 72,000 hectares (data FAOSTA 2013). Here, due to the favourable climatic conditions, potato production is almost uninterrupted throughout the course of the year. In recent years, potato cultivation has been concentrated in southern Italy, where it has 62% of the total number of companies and 70% of the cultivated area. The Italian potato production is divided into three harvests per year, with cultivation from November to May (early potato), or from April to September (common potato), or from August to December (late potato). The main national product is the common potato, widespread in northern regions, where the climatic conditions favour high yields per hectare. Campania, Puglia, Sicily and Sardinia show lower yields per hectare compared to the North. However, in these regions there is 90% of Italian production of the most profitable early potato. The national potato trade have weakness at various steps of the production and distribution system. One of the most serious problem is related to the extreme fragmentation of its production, both in terms of production companies and in terms of cultivated varieties; these, in fact, are usually selected in Northern Europe, where the soil and the climate conditions are different from the Italian one.

The development of new cultivars is one of the top priorities for cultivation worldwide. However, potato breeding is a hard grind. Accelerating the process, for example, by applying progeny selection to identify superior crosses, probably by massively expanding the nuclear seed of a promising variety in the early stages of evaluation (e.g. through state-of-the-art minituber production) is fraught with the risk of serious financial exposure. To reiterate the point, breeding successful potato varieties is a big challenge. Using a gross average, each sector can reasonably expect one new variety enter commercialization every decade. In potato breeding, yield potential, tuber internal and external quality traits, range of adaptation, and disease and pest resistances are the most important criteria for selection. The range of adaptation (or degree of stability in performance over a wide range of environments) is particularly difficult to breed for as the physiology of potato shows a very strong influence of genotype x environment interactions. A better understanding of the physiological and genetic regulation of agronomically important traits such as resistance to stresses will facilitate the breeding of varieties able to perform well under specific sets of conditions. This is particularly true for environmental circumstances that are considered unfavourable for potato, but where the urgent need for a highly productive crop with high nutritional value makes potato cultivation desirable.

1.2. Classical potato breeding

The genetic characteristics of the potato make the genetic studies on the heritability of characters useful for the development of new varieties difficult and complicate. *S. tuberosum* is a tetraploid tetrasomic in which, at a given locus (e.g. A), five genotypes are possible: AAAA (quadruplex), AAAa (triplex), AAaa (duplex), Aaaa (simplex), aaaa (nulliplex). Traditional approaches of genetic improvement of potato are focused on selection of clones coming from crosses between tetraploid clones (Carputo and Frusciante, 2011). Sexual interspecific hybridization, which is based on the crossing between cultivated potato and wild species, is the best strategy that allows to access the rich genetic pool of wild species. It is often accomplished through the manipulation of ploidy levels (genome engineering) to overcome sexual barriers or differences between parental chromosomal makeup (Carputo and Barone, 2005). It is based on the possibility to reduce ploidy levels through haploidization or increase them through sexual or somatic polyploidization.

While the cultivated varieties are tetraploid, most wild species are diploid (2n =2x=24). 4x x 2x crosses between tetraploid varieties and diploid species are not possible due to a strong triploid block. Changes of the ploidy level of parents allow the crossing with wild species. As shown in Figure 1.5 Panel A, haploid plants (2n = 2x = 24) are obtained from tetraploid varieties. Haploids can be easily crossed with many diploid species. If diploid hybrids produce 2n gametes, they may be crossed with tetraploid varieties in order to obtain tetraploid progeny. One additional problem that potato breeders have is that not all diploid potato species can cross with *S. tuberosum* haploids as previously described. In potato, the degree to which two species are interfertile must be tested experimentally but can be predicted to a great extent from their ploidy and Endosperm Balance Number (EBN). These are of interest for taxonomic and phylogenetic considerations, but are also of practical importance for breeders. The EBN system forms a strong isolating mechanism in sect. *Petota*. The EBN hypothesis was first published by Johnston et al. (1980) to explain success or failure of intra- and interspecific crosses, due to the functioning or breakdown of the endosperm after fertilization. The EBNs are hypothetical genetic factors independent of ploidy and empirically determined relative to other EBNs. They are based on crossability with standard EBN test crossers or other species of known EBN, and are published with the actual ploidy of the species. In potato, these are 2x (1EBN), 2x (2EBN), 4x (2EBN), 4x (4EBN), and 6x (4EBN). Differences in EBN are important barriers in nature, but ploidy manipulations to bridge EBN barriers, and general lack of strong crossability barriers within species sharing EBN, have allowed relatively easy access of these wild species by breeding programmes.

Due to the presence of EBN barriers the strategy described in Figure 1.5 A is not always possible. To overcome these incompatibilities between wild diploid 2x (1EBN) species and haploid of the cultivated varieties [2x (2EBN)] approaches are possible based on the production of bridge F1 triploids. Fig 1.5 B shows the breeding scheme used to introduce *S. comme*rsonii (2X, 1EBN) into the cultivated gene pool (Carputo et al., 1997).

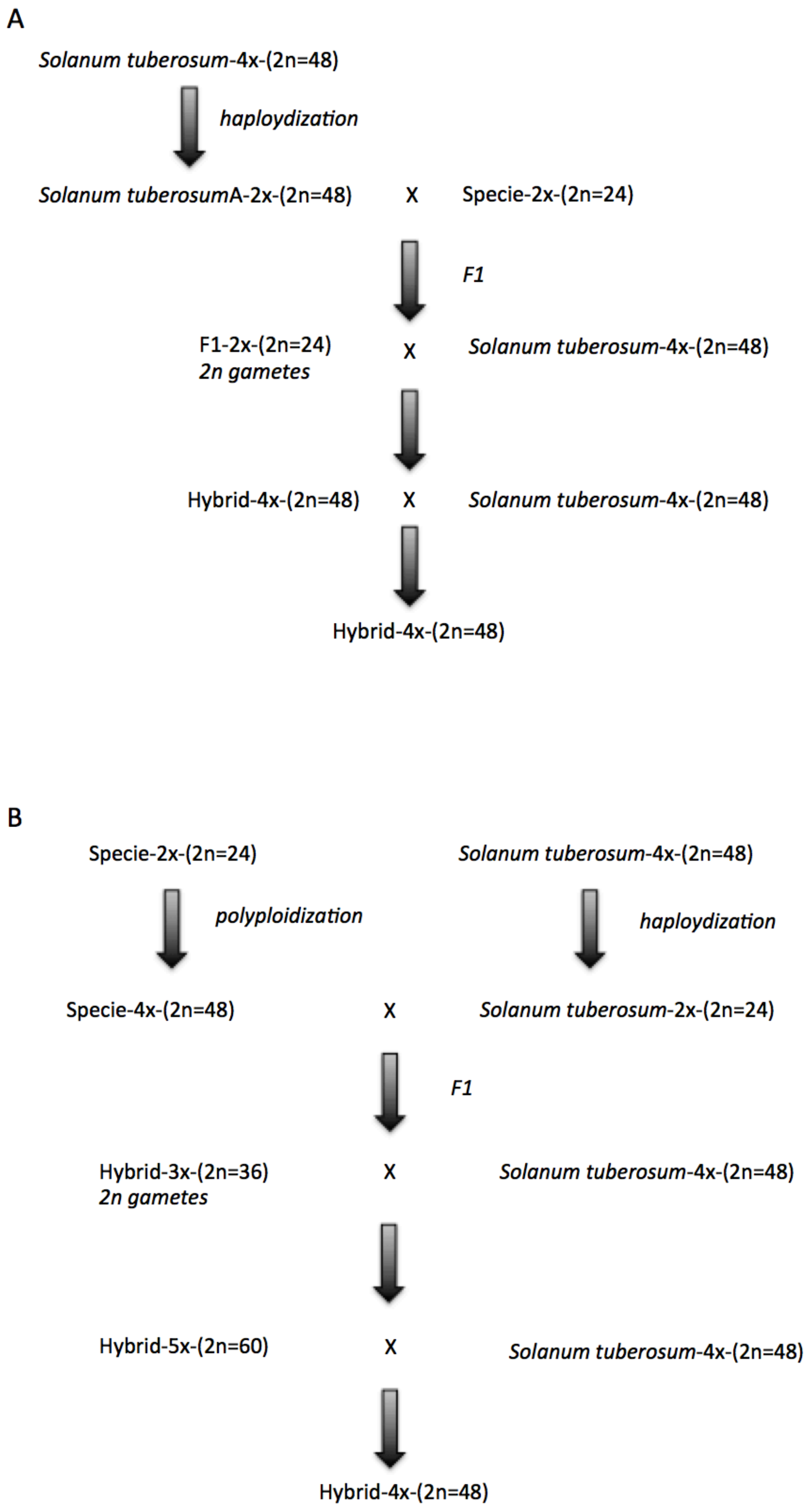

A
Solanum tuberosum-4x-(2n=48)
haploydization
Solanum tuberosumA-2x-(2n=48)
X
Specie-2x-(2n=24)
F1
F1-2x-(2n=24)
2n gametes
X
Solanum tuberosum-4x-(2n=48)
Hybrid-4x-(2n=48)
X
Solanum tuberosum-4x-(2n=48)
Hybrid-4x-(2n=48)
B
Specie-2x-(2n=24)
Solanum tuberosum-4x-(2n=48)
polyploidization
haploydization
Specie-4x-(2n=48)
X
Solanum tuberosum-2x-(2n=24)
F1
Hybrid-3x-(2n=36)
2n gametes
X
Solanum tuberosum-4x-(2n=48)
Hybrid-5x-(2n=60)
X
Solanum tuberosum-4x-(2n=48)
Hybrid-4x-(2n=48)

Figure 1.5: Breeding schemes to introgres exotic germoplasm, based on the use of diploid species compatible with *S. tuberosum* haploid (panel A) and bridge crosses (3x and 5x) (panel B).

The bridge F1 triploids can be produced by crosses between wild species (2X, 1EBN) and the *S. tuberosum* haploids (2X, 2EBN), if the wild parent produces 2n gametes or, if it is impossible, through crosses between tetraploid forms of the wild species (4X, 2EBN) and *S. tuberosum* haploids (2X, 2EBN). In this case the wild parent is exposed to a process of somatic polyploidization to double the chromosomes and make it compatible with haploids. If triploids produce 2n gametes they can be backcrosses with tetraploid varieties, producing pentaploid hybrids. With the pentaploid hybrids it is relatively easy to proceed with backcrosses, as pentaploids are easily cross-referenced with *S. tuberosum*.

As already pointed out, potato breeding is not an easy task due to several factors, including the tetraploid level of the cultivated potato, sexual barriers to interspecif crosses, ploidy levels and self-incompatibility at the diploid levels. However, in the last few years genetics, genomics and breeding have emerged as three overlapping and complimentary disciplines for comprehensive and fine-scale analysis of plant genomes and their precise and rapid improvement. While genetics and plant breeding have contributed enormously towards several new concepts and strategies for elucidation of plant genes and genomes as well as development of a huge number of crop varieties with desirable traits, genomics has depicted the chemical nature of genes, gene products and genomes and also provided additional resources for crop improvement. In today's world, teaching, research, funding, regulation and utilization of plant genetics, genomics and breeding essentially require a thorough out understanding of their components including classical, biochemical, cytological and molecular genetics; and traditional, molecular, transgenic and genomics-assisted breeding.

1.3 The potato wild germplasm

The cultivated potato is an unusual crop that has extremely large secondary and tertiary gene pools consisting of related wild species that are tuber-bearing, albeit with small inedible tubers. The taxonomy of the cultivated potato and its wild relatives has been the subject of study for many years. Less than 200 species and many intraspecific taxa have been described. These taxa have been classified in series, with different authors recognizing different numbers of series, often with different circumscriptions. Two authorities essays (Correll, 1962; Hawkes, 1990) recognized 26 and 21 series, respectively. Hawkes (1989) suggested a division of the series into two super series, *Stellata* and *Rotata*, emphasizing the outline of the corolla as a major distinctive character (Table 1). Some of the series contain only one or just a few species, indicating that their relationship to the other species is not clear. Series such as *Piurana* and, especially, *Tuberosa*, are large groups of species that may not be closely related to each other. Many wild potato species look similar to, and are easily confused with, cultivated potatoes. Figure 1.2 shows a variety of phenotypes, ranging from *Solanum hougasii*, typical for a 'cultivated-like' form, to *S. agrimonifolium*, with somewhat parallel lanceolate leaflets seen in many members of series *Conicibaccata*; to *S. morelliforme*, with simple leaves and an epiphytic habit to *S. infundibuliforme,* with a diminutive stature and linear leaflets. Flower shapes range from star-shaped (stellate), typical of many Mexican diploid species and some other species from South America, to highly wheel shaped (rotate), and with intermediate shapes referred to as pentagonal or rotate pentagonal (Spooner and Van den Berg, 2001) (Figure 1.2). Tubers grow on stolons (underground stems) that can attain a length of a meter or more. Tubers vary greatly in size from a few millimetres (e.g., *S. clarum, S. morelliforme*), to that of some of the cultivated species (e.g., *S. burtonii* or *S. candolleanum*). Their form varies from globose to tubular (straight to curved), with many intermediate shapes (Figure 1.3). Most species have tubers at the end of stolons, as in the cultivated species, but most species in ser. *Piurana* have tubers arranged along the stolons like beads on a string (Figure 1.3 F) (Salas et al., 2001). Flower colours range from white, to cream white, to various shades of pink, purple and blue. Hijmans and Spooner (2001) documented the geographic distribution of wild potato species, with the majority occurring in Argentina, Bolivia, Mexico and Peru, many with only restricted distribution areas (Figure 1.4). Potato species largely differ in ploidy level, with 12 being the basic (x) chromosome number. The ploidy level of 14 of the 196 wild potatoes is not known, leaving 182 with known ploidy. Of these, 139 are diploid, and six of these diploids have additional triploid populations with 36 chromosomes (3x). Seven species are exclusively triploid, 22 exclusively tetraploid (48 chromosomes, 4x), one exclusively pentaploid (60 chromosomes, 5x), and 12 exclusively hexaploid (72 chromosomes, 6x). Three species have populations with more than one even ploidy level (*S. acaule* 4x, 6x; *S. leptophyes* 2x, 4x; *S. oplocense* 2x, 4x, 6x). The triploid and pentaploid populations are generally highly sterile. They are less likely to be discovered as most germplasm collecting expeditions collect seed rather than tubers because tubers contain less genetic diversity and more diseases. It is likely, therefore, that the number of species with additional triploid or pentaploid populations is greater than currently known.

Table 1 Potato series according to Hawkes (1990)

Subsection *Estolonifera*	
	Series *Etuberosa*
	Series *Juglandifolia*
Subsection *Potatoe*	
Superseries *Stellata*	
	Series *Morelliforme*
	Series *Bulbocastana*
	Series *Pinnatisecta*
	Series *Polyadenia*
	Series *Commersoniana*
	Series *Circaeifolia*
	Series *Lignicaulia*
	Series *Olmosiana*
	Series *Yungasensa*
Superseries *Rotata*	
	Series *Megistacroloba*
	Series *Cuneoalata*
	Series *Conicibaccata*
	Series *Piurana*
	Series *Ingifolia*
	Series *Maglia*
	Series *Tuberosa*
	Series *Acaulia*
	Series *Longipedicellata*
	Series *Demissa*

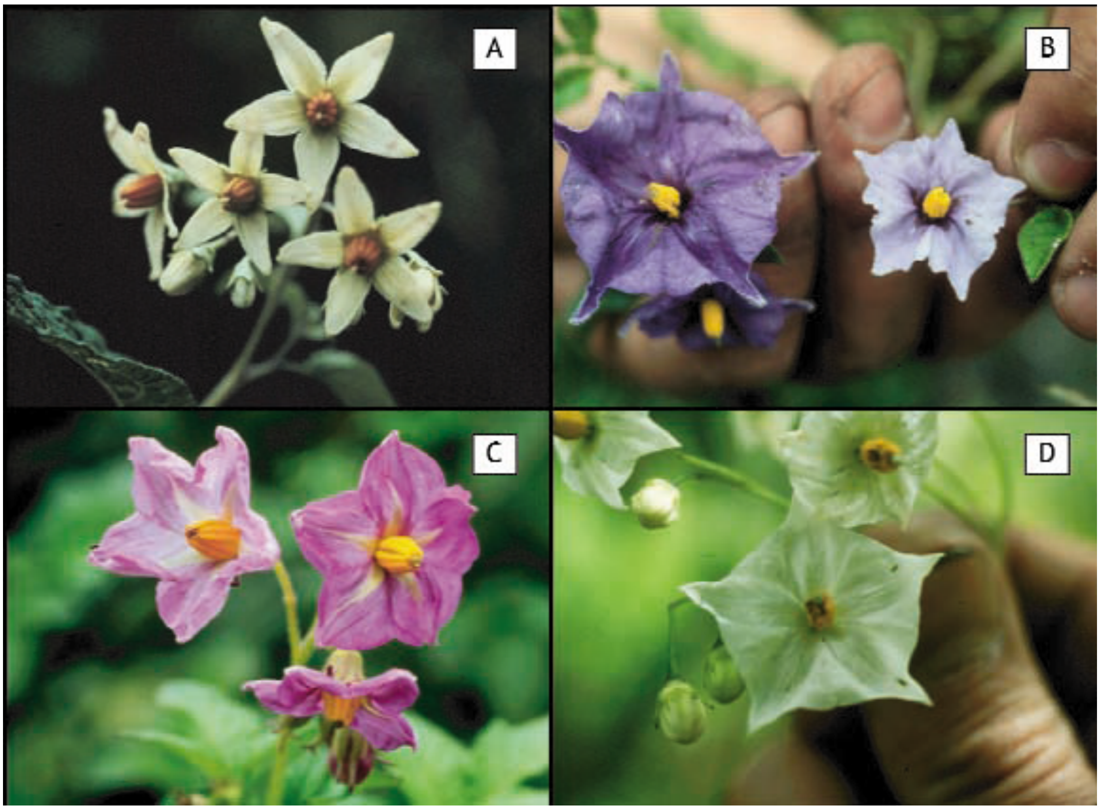


Figure 1.1 Wild potato plant forms. A: *Solanum hougasii*; B: *S. agrimonifolium*; C*: S.morelliforme*; D: *S. infundibuliforme* (photos by David Spooner).

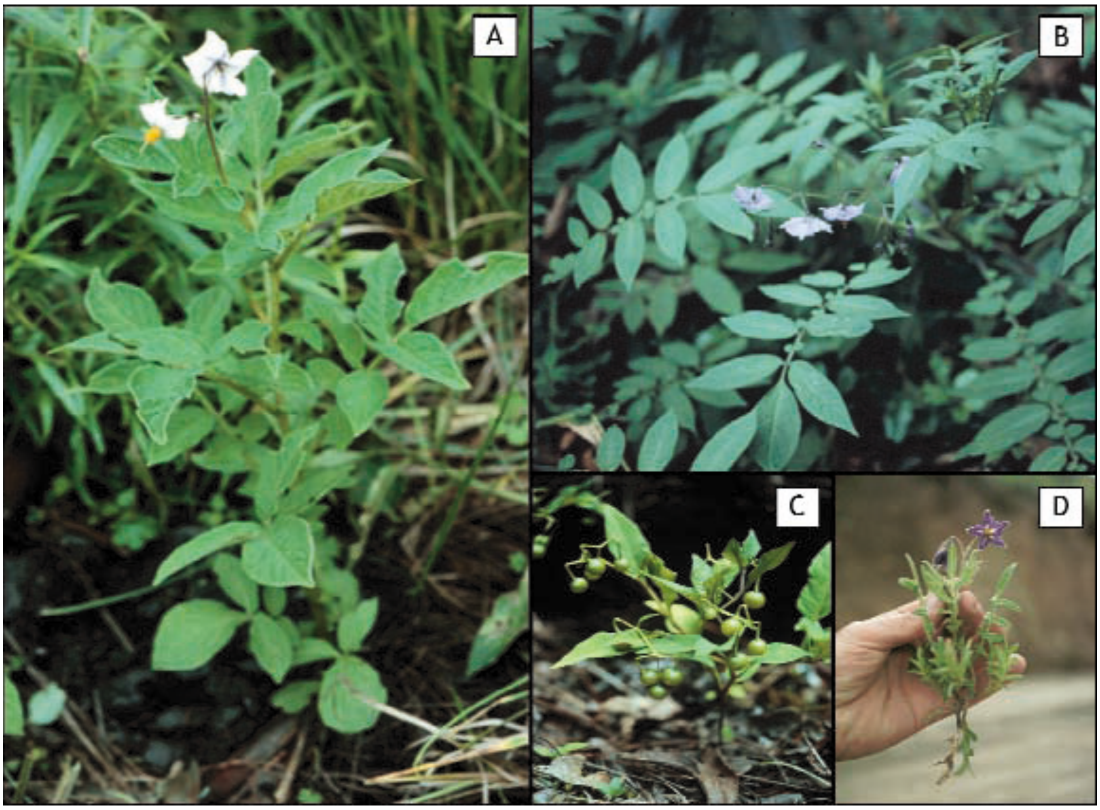


Figure 1.2 Wild and cultivated potato flowers. A: *Solanum bulbocastanum*; B: *S. paucijugum*; C: *S. tuberosum* (cultivated species); D: *S. colombianum* (photos by David Spooner).

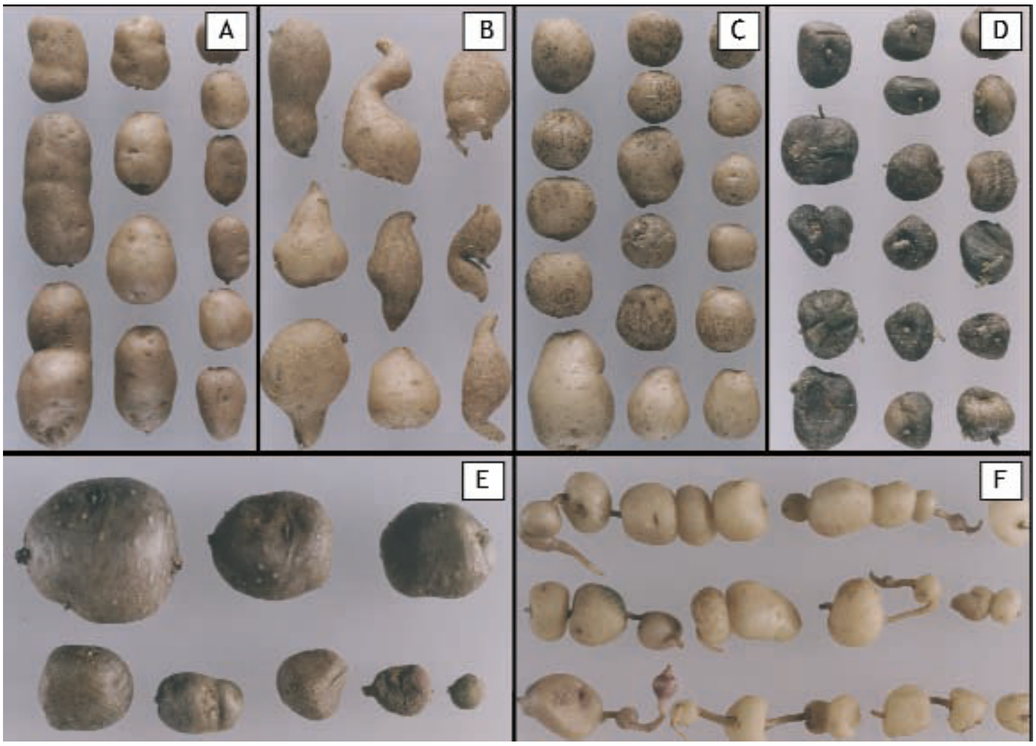


Figure 1.3 Wild potato tubers. A: *Solanum sparsipilum*; B: *S. polyadenium*; C: *S. acaule*; D: *S. chiquidenum*; E: *S. commersonii*; F: *S. piurae*. Scale across photos is comparable; the size of the largest tuber on panel E is approximately 3 cm (photos by Candelaria Atalaya).

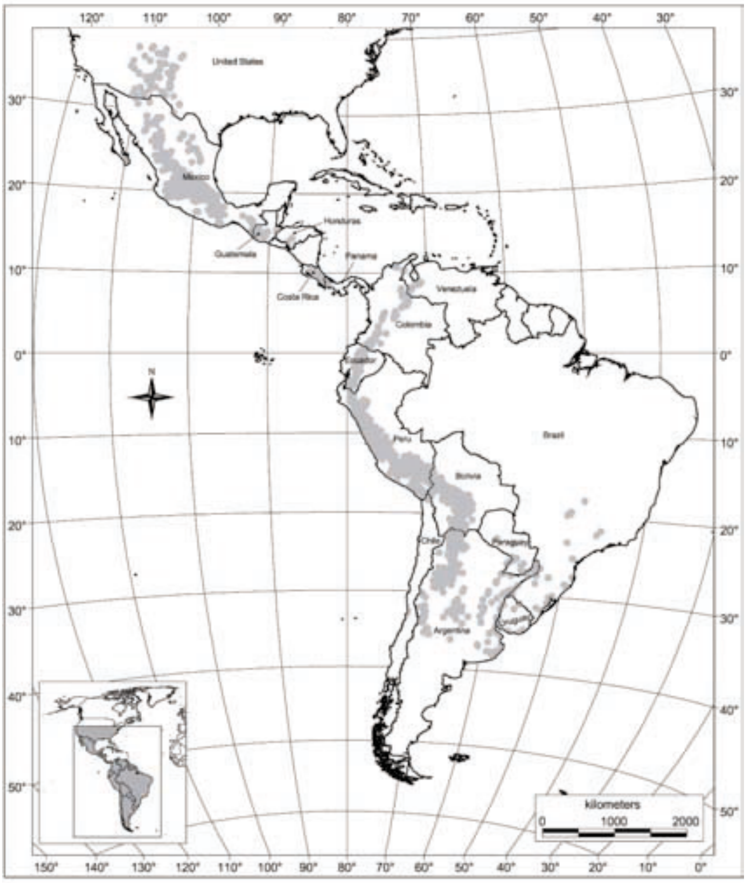


Figure 1.4 Area of known distribution of wild potatoes (grey shade).

The cultivated potato*,* *S. tuberosum*, is classified in series *Tuberosa*, a rather large and variable group without clear diagnostic characters. The link between wild and cultivated potatoes, the direct ancestors of the crop, must be looked for in the so-called *brevicaule* complex, a group of morphologically variable, diploid species within series *Tuberosa*. Within this complex, about 20 species have been distinguished, but Ugent (1970) suggested that these could be drastically reduced to one species (*Solanum brevicaule*) and Van den Berg et al. (1998) confirmed that conclusion. Morphologically, many of the wild species in the *brevicaule* complex are similar to some of the cultivated potatoes, the main differences being found in leaf dissection, in corolla colour and obviously in the tuber. The origin of the cultivated potatoes has been described as the result of successive hybridizations between diploid members of the *brevicaule* complex, accompanied by chromosome doubling leading to the tetraploid forms. The crop itself has been classified into seven cultivated species (*Solanum ajanhuiri, Solanum chaucha, Solanum curtilobum, Solanum juzepczukii, Solanum phureja, Solanum stenotomum* and *S. tuberosum* with two subspecies, *tuberosum* and *andigena*), showing several ploidy levels from the diploid to the hexaploid.

The discussion about the taxonomic status of cultivated plant material (Hetterscheid and Brandenburg, 1995) suggests that the taxon ‘species’ (with its connotation of a product resulting from evolutionary processes) is not suitable for the classification of cultivated plants as the influence of humans seriously disturbs the patterns of variation used to classify species. Rather, cultivated material should be treated as artificial entities such as landraces or cultivars and classified into cultivar-groups as advocated in the International Code of Nomenclature of Cultivated Plants (ICNCP, 2004). This was anticipated by Dodds (1962), who suggested the informal groups *Stenotomum, Phureja, Chaucha, Andigena* and *Tuberosum* within the species *S. tuberosum* to accommodate the cultivated potatoes. Huaman and Spooner (2002) suggested a similar solution with eight groups (*Ajanhuiri, Juzepczukii, Curtilobum, Chilotanum, Andigenum, Chaucha, Phureja* and *Stenotomum*). The crop ‘potato’, making up the total of these groups, can still be assigned to the ‘species’ *S. tuberosum*, if so desired. This species name should then be considered as a cultigen (a species consisting of cultivated plants only, and as such without wild representatives, without a natural geographic distribution area and without a natural population structure).

The potato has the most extensive and rich wild germplasm of the plant kingdom. The potato germplasm banks, including such the International Potato Centre (CIP, Lima, Peru) or Potato Introduction Station (Sturgeon Bay, Wisconsin - USA) (Huaman, 1997), as well as banks germplasm available online (http://www.potatogenebank.org) were built thanks to germplasm collections from Central and South America. Wild species of *Solanum* are useful genes lacking in cultivated varieties and represent a reservoir of variability from which to find the characters needed for the constitution of new varieties. Among the most important traits are the high content of dry matter, the low content of reducing sugars, resistance to biotic stresses (mushrooms, insects, nematodes, bacteria and viruses) and abiotic (especially cold and drought), and many other related to quality and productivity. Sources of resistance to the *Leptinotarsa decemlineata* were found in *S. pinnatisectum* and *S. tarijense*. Resistance to *Verticillium spp*. and *Clavibacter spp*. has been reported in *S. lesteri, S. polyadenium* and *S. jemesii.* Particularly interesting are the wild species *S. bulbocastanum*, *S.* and *S. polyadenium pinnatisectum*, resistant to the fungus plant pathogen *Phytophthora infestans*, the main adversity of the potato in terms of economic losses. This fungus has spread to all areas of the world where is cultivated potato. In the past it was responsible for real famines caused by the destruction of entire crops (Ireland 1845-46). The knowledge of the life cycle, the main sources of inoculum and climatic conditions favourable to its development can, with appropriate agronomic techniques, reduce the damage of any attack. However, the use of fungicides inevitably causes an increase in costs of production and especially damage to the environment. To this we must add extraordinary plasticity of genetic oomycetes and their adaptability and variability seemingly limitless.

Potato species are also a source of allelic diversity necessary to obtain heterosis for the production of tubers and other features subject to polygenic control (Carputo and Barone, 2005). The wild germplasm is the richest source of variability, but not the only one. Important genetic resources are also represented by varieties, landraces, hybrid euploid and aneuploid and other materials developed during the activity of breeding, part of the primary "*gene pool*" . Increasingly expanded genepools are becoming available for crop improvement. There are now a variety of types of genes from diverse sources that might be used to enhance disease resistance. These include cloned resistance genes, genes that are involved in the induction of the resistance response, resistance-response genes, genes with antimicrobial activity from non-host sources and novel genes that have been generated *in vitro*. Among potato species, a very interesting one is diploid *S. commersonii.* It is a 2x (1EBN) species possessing resistance to biotic and abiotic stresses. Next paragraph gives a detailed overview on this valuable source of germplasm for potato breeding.

1.4 *Solanum commersonii*, a source of useful genes

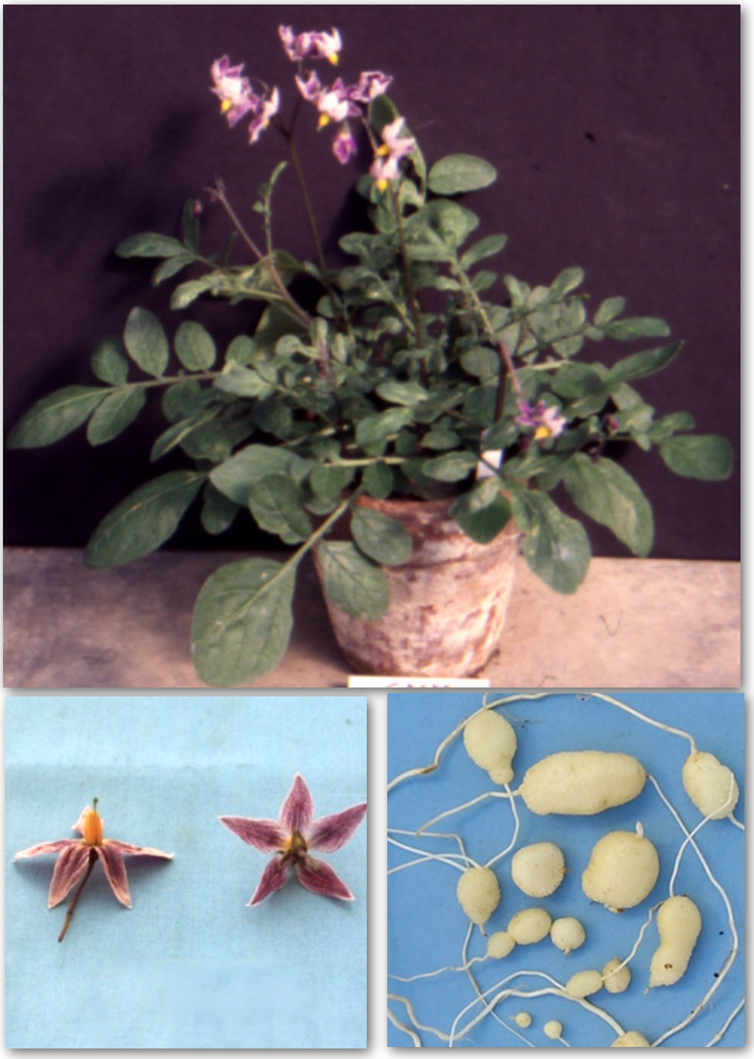


Figure 1.6 Wild potato *Solanum commersonii.*

S. commersonii Dun. is a tuber-bearing wild potato species native to Central and South America (Figure 1.6). The French taxonomist Michel-Felix Dunal named this species in honor of Philibert Commerson (1727-73), who collected the type specimen (No. 47) in 1767 at Montevideo, Uruguay. This was probably the first wild potato to be collected on a scientific expedition (Hawkes, 1990). Analyses of chloroplast genome restriction sites and nitrate reductase gene sequence confirmed that *S. commersonii* is phylogenetically distinct from the cultivated potato (Spooner et al., 2005; Rodriguez et al., 2009; Rodríguez and Spooner, 2009). Consistently, *S. commersonii* and *S. tuberosum* are sexually incompatible (Jackson and Hanneman, 1999) and have been assigned different endosperm balance numbers (EBNs) (Johnston et al., 1980), with *S. commersonii* and *S.*

tuberosum reported as 1EBN and 4EBN, respectively. Despite being genetically isolated from the cultivated potato, *S. commersonii* has garnered significant research interest. It possesses several resistance traits not found in cultivated potato including resistance to root knot nematode, soft rot and blackleg, Verticillium wilt, Potato Virus X (PVX), Tobacco Etch Virus (TEV)*,* common scab, and late blight (Hanneman and Bamberg, 1986; Hawkes, 1990; Micheletto et al., 2000). Particularly attractive is its freezing tolerance and capacity to cold acclimate (i.e., able to increase cold tolerance after exposure to low, non-freezing temperatures). In contrast, the cultivated potato is classified as sensitive to low temperatures and unable to cold acclimate (Palta and Simon, 1993). Both frost resistance and cold acclimation capacity are important breeding traits, since temperatures below 0°C are a major cause of yield losses in several production regions.

From the literature it is known that plants use a variety of adaptive mechanisms to take on environmental stress caused by cold temperature. Seed dormancy is a physiological condition that delays seed germination until the embryo has gone through an after-ripening period, during which certain biochemical and enzymatic processes occur for the seed to attain full maturity. When plants are exposed to gradually decreasing temperatures below a certain threshold, they acclimatize (low-temperature acclimation) to the stress, a process called *cold hardening*. In spite of various adaptations to cold, plants may be injured through exposure to cold temperatures in a variety of ways, depending on the temperature range. One type of injury, called chilling injury, occurs at drop below the freezing point of water. Sometimes, ice crystals form in the protoplasm of cells, resulting in cell death and possibly plant death. Plants may be classified into three groups according to their tolerance to low temperatures. Frost-sensitive plants are intolerant of ice in their tissues and are hence sensitive to chilling injury. The plant (e.g., beans, corn, tomato) can be killed when temperatures fall just below 0°C. Frost-resistant plants can tolerate some ice in their cells and can survive cold temperatures of up to − 40°C. Cold-hardy plants are predominantly temperate woody species. They can survive temperatures of up to − 196°C. The capacity of a genotype to tolerate low temperatures has been extensively studied. Whereas it is agreed that low-temperature tolerance is a complex trait (quantitative), researchers are not unanimous on the mode of gene action governing the expression of the trait. Genes that condition varying levels of low-temperature tolerance occur within and among species. This genetic variability has been exploited to a degree in cultivar development within production regions. A large amount of low-temperature tolerance research has been conducted in wheat. Low-temperature tolerance in cereals depends on a highly integrated system of structural, regulatory, and developmental genes. Several vernalization genes have been identified (e.g., *vrn1, vrn4*). The *vrn1* is homeoallelic to the locus *Sh2* in barley and *Sp1* in rye. These two genes have been linked to genetic differences in low-temperature tolerance. Winter cereals also produce several proteins in response to low-temperature stress, for example the dehydrin families of genes (*dh5*, *Wcs120*). Whit in *Solanum* wild species, *S. commersonii* is studied since many years and attempts have been made to elucidate the mechanisms of freezing tolerance and identify biochemical traits that are accompanied by increased freezing tolerance during cold acclimation. The first study go back to 1980 in which Chen and Li compared the biochemical changes involved in the cold acclimation process in the leaves of freezing tolerant *S. acaule* and *S. commersonii* with those in freezing sensitive *S. tuberosum*. Their research showed that there are similar increases in sugar and starch content during cold acclimation, althrough *S. tuberosum* failed to acclimate but the net synthesis of soluble proteins and the level of total lipid and phospholipids were higher in freezing tolerant wild species. Chen et al. (1992), have also observed changes in endogenous ABA levels to increase transiently in *S. commersonii* after four days of cold acclimation and also detected two separate peaks of free ABA on the second and sixth days of cold acclimation. Evidence for the role of ABA as a signalling molecule was revealed when exogenously applied ABA was observed to improve freezing tolerance at room temperature in *S. commersonii.* The lipid composition of freezing tolerant *S. commersonii* and freezing sensitive *S. tuberosum* was also studied. The comparison revealed that in the tolerant species several changes in plasma membrane lipid composition occurred which could not be found in *S. tuberosum* following cold acclimation (Palta et al., 1993). *S. commersonii* and in *S. sogarandinum*were the most interesting wild potato species in which the effect of acclimation on gene expression has been studied. Three osmotin-like cDNAs (pA13, pA35, pA81), which were not only low temperature inducible, but were also induced by ABA and pathogens in *S. commersonii*. The induction of osmotin-like protein mRNA was relatively slow at low temperature and it did not result in an elevated protein level. Instead, osmotin-like protein accumulated after pathogen (*Phytophthora infestans*) infection (Zhu et al., 1995).

In *S. commersonii* freezing tolerance is defined in genetic terms as a complex quantitative trait controlled by several, as yet unknown, combinations of genes and gene families (Stone et al., 1993). Studies on the molecular mechanisms behind improved freezing tolerance in cold acclimated plants have revealed information on cold-induced genes from over 50 different plant species. The genetic complexity of the trait makes genetic engineering challenging. The modern methods of molecular plant breeding have indicated that the physiological and biochemical bases of plant freezing tolerance need to be studied more in order to engineer the trait or to find reliable selection markers. Progenies that have been derived from interspecific hybrids between freezing tolerant and sensitive plant species can provide valuable information on the mechanisms underlying freezing tolerance and acclimation capacity and the genetic basis of freezing tolerance in plants.

In spite of the lack of information on target traits and its sexual isolation, breeders have successfully introgressed genes from *S. commersonii* into cultivated potato (Cardi et al., 1993; Bamberg et al., 1994; Carputo et al., 1997). The success of this approach is widely documented in the literature. For example Carputo et al. (2009) have transferred the resistance to *Ralstonia solanacearum*, cold and soft rot in sexual pentaploid hybrids used bridge ploidies between *S. commersonii* and *S. tuberosum*. Carputo et al. (2000) have also transferred the resistance to *Erwinia carotovora* from *S. tarijense* to *S. tuberosum* through crosses 2x × 4x. Despite such efforts, very little progress has been made in the release of new varieties. This is at least partially due to the lack of genomics resources available for *S. commersonii* and other potato wild relatives. The genome sequence of the cultivated potato was published in 2011 (Potato Genome Sequencing Consortium et al., 2011), rushing in a new era of potato functional and comparative genomics. By contrast, no report is available on the genome sequence of a wild potato species. Comparative sequencing of cultivated potato and wild relatives will increase the use of wild species for crop breeding and improvement (Stupar, 2010).

1.5 Biotechnological advances: applications and opportunities for potato breeding

The immense amount of diversity within cultivated potato and its relatives provides both opportunities and challenges. The demands of industry and the public for rapid introduction of new and better varieties contrasts with the long-term efforts that are needed to incorporate beneficial diversity not present in the adapted germplasm. In these years, the application of genomics technology within potato germplasm evaluation and breeding programs is reviewed and new opportunities identified. DNA markers and sequencing information have been critical in taxonomic studies within wild and cultivated potato species. Markers have helped to elucidate the phylogenetic relationships within the primary and secondary gene pools of cultivated potato and determine the most likely progenitors of cultivated *S. tuberosum*. Various strategies for incorporating markers into potato breeding programs have been discussed. The greatest impact of molecular breeding in potato will likely be for pre-breeding and parental development activities. The typical model of genetic modification of existing cultivars may not be sufficient as a mean to make continual breeding progress to address the varied requirements of industry and consumers. Cisgenic technology has the potential to revolutionize how breeders access genes for pre-breeding and parent development. Enhanced knowledge of the genetic factors controlling traits important to industry as well as those traits important for breeding success will make the coming years exciting for potato improvement. Molecular breeding provides the means to transfer this knowledge to the development of new cultivars for a variety of uses.

Moreover, considerable scientific and commercial progress has been made since these early reports, and a large number of single-gene and multiple-gene traits have been engineered into a wide range of potato and related germplasm (Millam, 2005). One example is *Amflora*, a genetically modified potato cultivar, approved for industrial application in the European Union market which is used for industrial production of starch, a biopolymer of two monomers: amylose and amylopectin. The latter is the important monomer for industrial purposes. Potatoes typically produce starch that is 20% amylose and 80% amylopectin, but thanks to genetic modification, *Amflora* potatoes produce only amylopectin. In addition to the commercial applications of GM technology, which are increasing in both total area and the number of countries growing GM crops, gene transfer methods can be used for a wide range of fundamental studies, contributing to a better understanding of the mechanisms of plant: pathogen interactions and biosynthetic pathways in plants. Examples of *in vitro* manipulation technology in potato vary from the low-input and widely adopted (and non-GM) technology of micropropagation to the complex manipulation of multi-gene biosynthetic pathways through transgenic intervention. Potato is considered, due to its high *in vitro* regeneration capacity, a model species for methods such as somatic hybridization and *Agrobacterium* -mediated transformation. Indeed, historically, potato was one of the first crop plants to be successfully transformed. In addition, the related technology of micropropagation is an important component of many breeding schemes, and anther culture methodology has been adapted recently to enable the production of haploid lines from an increasing number of genotypes. Furthermore, recent advances in somatic embryogenesis in this species have increased the scope for the rapid propagation of novel materials (Sharma and Millam, 2004).

Additional new disciplines, known as –omics sciences, have been developed that are having a great applicability to potato breeding. Among them transcriptomics is the study of the transcriptome, which is defined as the complete set of transcripts or a specific subset of transcripts (e.g. messenger RNA molecules) produced from the genome in one cell or population of cells of a given organism at any one time. Recent studies of plant development and environmental stress responses have converged on the roles of RNA and its metabolism as primary regulators of gene action (Hollick, 2008). Most translational regulation factors such as proteases and kinases are also products of transcripts. Since *S. tuberosum* is autotetraploid, the transcriptome and transcriptional regulation are expected to be highly complex. Transcriptomics approaches are not only powerful but also necessary for understanding many aspects of the cell biology responsible for the performance and quality of this tetraploid crop. Considerable progress has been made in generating functional potato genomics resources, understanding the nature and regulation of the potato transcriptome, and applying potato transcriptomics resources for crop improvement (Rensink and Buell 2005; Bryan and Hein 2008; Li et al., 2008). Several different technologies that have been or are likely to be used to study the potato transcriptome, including sequence tagging methodologies, various types of microarrays, PCR-based technologies, and various emerging sequencing and gene expression nanotechnologies. Additionally, numerous online resources have emerged, supporting transcriptomics research efforts in potato.

Microarray technology is used to examine simultaneous gene expression profiles of different cells and tissues. Since it has been employed in this research, a few highlight on this technique are provided. For potato transcriptome analysis, cDNA microarrays and oligo-based microarrays have been developed. Additionally, oligo-based tomato microarrays have been successfully used to study the potato transcriptome. cDNA microarrays are constructed using probes consisting of cDNA library clone inserts. Transcriptome analysis in potato using cDNA microarrays has been employed in the study of tuber development (Kloosterman et al., 2005; Van Dijk et al., 2009), light-regulated transcription (Rutitzky et al., 2009), autopolyploidization (Stupar et al., 2007), and response to cold, heat, salt (Kim et al., 2003; Rensink et al., 2005; Oufir et al., 2008), and drought (Schafleitner et al., 2007). During cold exposure (up to 50% cells injured), transcriptome analysis revealed a differential response between the wild *Solanum phureja* and cultivated *S. tuberosum* plants (Oufir et al., 2008). Gene expression profiling of potato seedling responses to cold (4°C), heat (35°C), and salt (100 mM NaCl) stresses identified a total of 3,314 genes that had significant up- or down-regulation in response to at least one stress condition. cDNA microarrays have also been employed to study the potato transcriptome in response to *Phythopthora infestans* (Restrepo et al., 2005; Tian et al., 2006; Wang et al., 2008), revealing a role for metabolic and signaling (e.g., salicylic acid, jasmonic acid, and ethylene) pathways. A plastidic carbonic anhydrase gene was found to have a very different expression pattern in compatible vs. incompatible interactions in potato plants during infection and, interestingly, silencing this gene increases *Nicotiana benthamiana* susceptibility to *P. infestans* (Restrepo et al., 2005).

In the last few years, thanks to sequencing technologies and bioinformatics advances, a somehow altered view of the genetic structure of organisms has emerged. Consequently, a changed perspective of plant breeding is growing, that is reaching an unprecedented precision. DNA sequencing is a fast-moving science with technologies and platforms being updated at breathtaking speed. The hallmark of next generation sequencing (NGS) has been a massive increase in throughput and a decrease in price compared with previous technologies. The first next-generation DNA sequencing machine was introduced to the market by 454 Life Sciences (Basel, Switzerland) in 2005. The technology is based on a large-scale parallel pyrosequencing system, which relies on fixing nebulized and adapter-ligated DNA fragments to small DNA-capture beads in a water-in-oil emulsion. The Illumina's (CA, USA) Genome Analyzer was released in 2007 and marked a true revolution for genome sequencing in which short reads became significant to genomic applications. The technology is based on reversible dye terminators. DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed. Life Technologies' (CA, USA) SOLiDTM technology employs sequencing by ligation. In this technology, a pool of all possible oligonucleotides of a fixed length is labelled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal that is informative of the nucleotide at that position. The so-called 'third-generation' technologies directly sequence individual DNA molecules rather than relying on any amplification prior to sequencing. The recently released PacBio system can produce 35–45 Mb of data per cell with an average read length of 1,500 bp. The Ion Torrent Personal Genome Machine (PGM) is another third-generation platform that uses standard sequencing chemistry, but with a novel, semiconductor-based detection system. This technology already claims read lengths of approximately 200 bp with high accuracy, and the latest PGM 318 chip can produce 1.0 Gb of data in a 2-h run. When the implications of NGS technology became apparent, several assemblers were designed to deal with the new problems, i.e., assembly of short NGS reads in order to reconstruct the main longer sequences. Assembly process can be done either having a reference genome available (mapping) or without having a reference genome available (de Novo assembly).

Knowing about the order (sequence) of nucleotides in DNA, the molecule in which the genetic information of all organisms is stored, has revolutionized biology and resulted in our better understanding of life's secrets (BBSRC Review of Next- Generation Sequencing—final version). The first two DNA sequencing techniques, which are known as first-generation DNA sequencers, historically were developed by Fredrick Sanger (1977, University of Cambridge) and Allan Maxam and Walter Gilbert (1976–1977, Harvard University), independently. Sanger's method, which earned him a Nobel Prize in Chemistry in 1980, became popular, and in fact was the sole method for DNA sequencing for three decades, as a result of its lesser technical complexity and lesser amount of toxic chemicals used.

In the Sanger sequencing method, which is also known as ''chain termination'' or the ''dideoxy method,'' modified nucleotides (fluorescently labeled dideoxynucleotides) are used in the reaction in addition to normal nucleotides; this method was gradually improved and became automated (the first automatic sequencing machine, AB370, was introduced in 1987 by Applied Biosystems), and therefore has been the method of choice for large-scale sequencing projects, e.g., whole-genome sequencing for various species, for about 30 years. In classical Sanger sequencing technology, which is sequencing by the synthesis method, the sequencing reaction is performed in the presence of the singlestranded DNA template, DNA primers, DNA polymerase, four normal DNA nucleotides, and four fluorescently labeled modified nucleotides (ddATP, ddCTP, ddGTP and ddTTP). The DNA template is initially divided into four separate sequencing reactions containing primers, polymerase and normal nucleotides. In each reaction in the presence of a small amount of one of four modified nucleotides (which lack the 3'-OH group required for the extension), which randomly incorporates into the growing strands, terminates DNA elongation and results in DNA fragments with various lengths. The obtained DNA fragments are then separated by size through high-resolution polyacrylamide gel electrophoresis (capillary electrophoresis) with each of four reactions run in one of four individual lanes (lanes A, C, G and T). DNA bands that correspond to DNA fragments with differing lengths are then visualized, using UV light or X-ray autoradiography, and the order of nucleotides can be determined according to the relative positions of DNA bands among four different lanes (Figure 1.7).

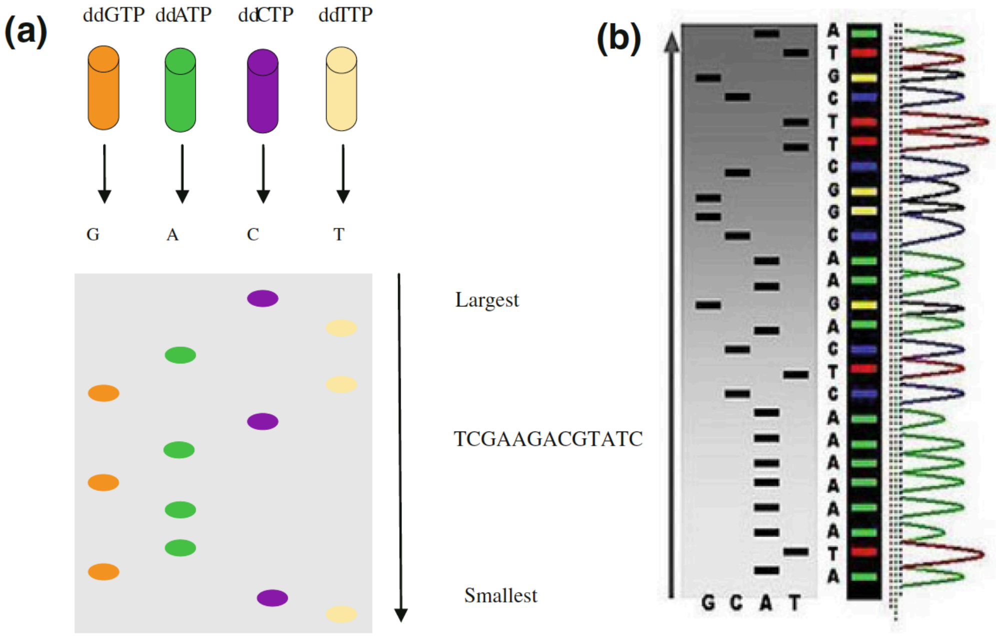


Fig. 1.7 Sanger sequencing procedure.

A. Four distinct reactions are taking place in the presence of all required materials for DNA synthesis. Besides in each separate reaction, a distinct type of fluorescently labeled dideoxy nucleotides is added which after completion DNA synthesis cycles, results in the DNA strands each of which terminated in specific dideoxy nucleotide present on that reaction.
B. After reaction completion, the content of four separate reactions is electrophoresed using high-resolution polyacrylamide gel (www.Wikipedia.org)

The Maxam-Gilbert technique relies on the cleaving of nucleotides by chemicals and is more efficient with small nucleotide polymers (Figure 1.8). Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A + G, C, C + T). Due to the advancements in chain termination methodology, the Maxam-Gilbert method has become redundant. It became obsolete due to its less ergonomical feasibility, and it is also considered unsafe because of the extensive use of toxic chemicals. As a result of using less toxic chemicals and lower amounts of radioactivity than the Maxam and Gilbert method, and because of its comparative ease, the Sanger method was soon automated and was the method used in the first generation of DNA sequencers.

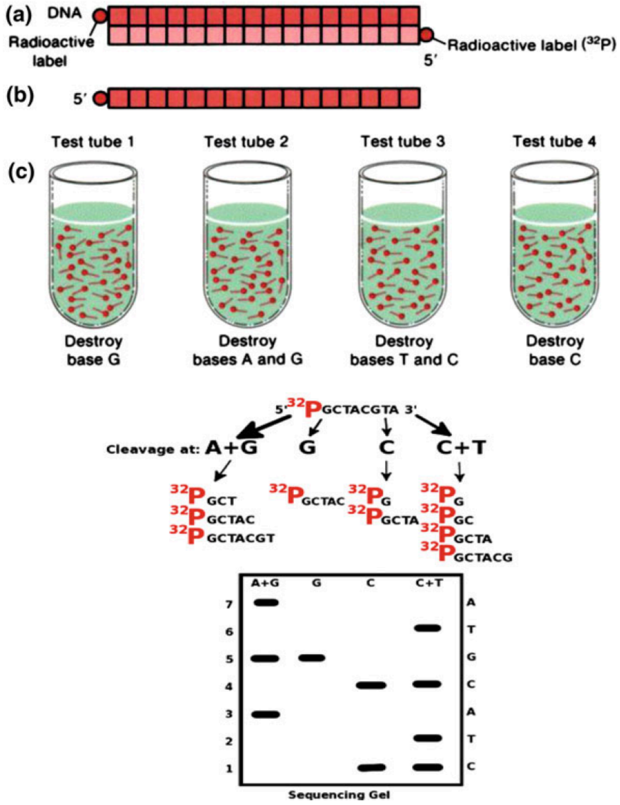


Fig. 1.8 Maxam-Gilbert chemical degradation sequencing technique.

A. Double-stranded DNA is labeled at 50 ends.
B. B Single-stranded DNA fragment is produced. C DNA fragments are distributed in four parallel test tubes. Each test tube is subjected to a specific base-degrading chemical. The content of each tube will be electrophoresed in the next step for fragment size separation

Next Generation Sequencing (NGS): The revolution

Although in the past few years the genomes of several species, as well as humans, were sequenced using the automated Sanger method, the above-mentioned limitations of this method indicated a need to develop new and improved sequencing technologies to sequence the large number of human genomes and to find answers to biological problems of interest that could not be addressed before (Scheibye-Alsing, 2009; Li, 2008). For example, advances in sequencing technology would help in the development infields such as comparative genomics, which involves comparing the genome of distinct organisms to learn about their molecular programs, biomedical research, through which so many problems concerning the genetic basis of susceptibility to diseases, multi-factorial diseases, and cancer therapy can be investigated. The detection of different genomic and epigenomics alterations, such as single nucleotide mutations, small insertions and deletions, chromosomal rearrangements, copy number variations, and DNA methylation can be facilitated using advanced sequencing technologies (Li, 2008).

Subsequently, several research centers initiated the designing of new sequencing technologies not needing gels, which would allow sequencing large numbers of samples in parallel (Butler, J., 2008). These technologies are known as Next-generation DNA sequencing (NGS) methods in which the bacterial cloning steps have been removed (in comparison with the Sanger method). Three major NGS methods that are routinely used in many laboratories today include:

1. The Roche/454 FLX (http://www.454.com)
2. The Illumina/Solexa Genome Analyzer (http://www.illumina.com)
3. The Applied Biosystems SOLiDTM System (http://marketing.appliedbio appliedbiosystems. com)

Additional massively parallel technologies that have been introduced more recently include the Polonator (Dover/Harvard), the HeliScope Single Molecule Sequencer technology (Helicos; Cambridge, MA, USA) (Ariyaratne, 2011 ; Huang, 1999), and the Ion Semiconductor (Torrent Ion Sequencing). The single molecule real time (SMRT) [Pacific Biosciences] and Nanopore Sequencing (Batzoglou, 2002) are another two newly introduced technologies that are based on the sequencing of single molecules.

There are two major types of sequencing projects in terms of application. In *de novo* sequencing, the genome of an organism is sequenced for the first time. In resequencing projects, the whole genome of an organism or parts of it is sequenced while the reference sequenced genome for the species of that organism is already available. Maximum sequencing efficiency is achieved as a consequence of both depth (coverage) and uniform read distribution (breadth). Sequencing depth or coverage concerns the average number of times each base in the genome is sequenced. For example, to sequence a 3 Gb human genome with 10x coverage, 30 Gb of sequenced data is needed. Sequencing breadth refers to the percentage of the genome that is covered by sequenced reads.

1.6 Bioinformatics in NGS

As the various sequencing technologies follow different paths, everyone provides as output sequences. Substrings of this sequence, of variable length from a few tens up to several hundred bases, are combined together, usually in a FASTQ file. This is a type of file used in biology to store genetic sequences and their quality scores, namely the score that the algorithm and links were the analysis assigns to the string, and which is then used to choose the best match against the reference genome. At this point, the bioinformatics analysis is divided into three steps: alignment, which search correspondences between the reads and the reference genome, variant calling, that attempts to separate from references due to genetic errors by instrument errors made in analysis, filtering and annotation (Figure 1.9), which attempt to align reads to the reference genome.

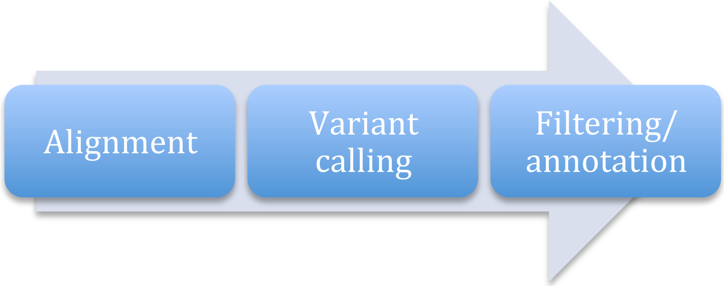


Figure 1.9: Basic steps in the analysis of the output produced by bioinformatics technology of next generation sequencing

Alignment

Alignment is the process by which you map reads to a genome reference. It is a complex task, since the software must compare each reads with each position of the reference DNA. It is a passage computationally demanding, and expensive in terms of time. The SAM (Sequence Alignment Map) and BAM (Binary Alignment Map) are reference standards for saving data obtained by the alignment for the new generation of technology. There are many commercially available softwares, free or paid, to perform this task. Most of these uses a method based on indices, which are faster than in the search for alignment positions without gap in the reference genome. Other algorithms instead allow the search for alignments with gaps. The various approaches to the problem involve the use of hash tables (e.g., MAQ, ELAND), based algorithms Burrows-Wheeler Transform (e.g., BWA, Bowtie, SOAP2), an algorithm compression reversible exchange the order of the characters, without changing the value, genome-based hash (e.g., Novoalign, SOAP), or an approach spaced-seeds (e.g., shrimp), a variant of the pattern matching which allows presence of a certain number of errors in the positions of the string xed. The algorithms can provide the best result based on heuristics, other instead they can provide all the results that give feedback. They differ also between algorithms that take into account the gap (e.g., BWA, Bowtie2) and algorithms that however does not take into account (e.g., MAQ, Bowtie).

Variant Calling

After the alignment of reads, the DNA analysis can be compared to the genome reference, thus having the ability to identify changes. These variations can be due to disease, or can be only noise, without any harmful effect. The complexity of this topic is to identify variants true, like mutations in the genome, from variations due to sequencing errors. The continuous development and improvement of new generation technologies hopefully will bring advantages in this area, improving input data that the variants analysis receives. The difficulty in the analysis of variants is mainly caused by the presence of indels, like phenomena of insertion or deletion. These are in fact the cause of most of the false-positives detected by these algorithms that number increases if not executed an array, which considers the gap. Another cause of the phenomenon are the mistakes made in the preparation of sequences, due to problems in the analysis PCR. The solution to this problem is to increase the sensitivity of the instruments in the analysis, and improve the software used in the alignment, as well as having a database for comparison of considerable size, so as to have a set for comparison wider.

Filtering and annotation

With the above steps generates a list of thousands of potential differences between the genome under study and reference are generated. The next step is therefore to determine which of these variations, not due to sequencing errors, contribute to the problem under study, thereby reducing the amount of material to be analysed. The method involves altering the results obtained by removing variations that follow models of other phenomena, and note, looking information changes, identifying those, which resonate with the process in question. Filtering can be performed with the comparison of reads against a tree genetic reference, so try all those elements for which you know the function, or, in the case of a plant disease, for example, analysing the sequenced genome of diseased tissue and healthy tissue of the individual, excluding those variations present in both tissues. In addition to filtering, annotation provides another tool to select and restrict our sample, carrying out research on past studies, or applying models of the functional effects predicted. The effectiveness and low cost of these instruments is leading to the discovery of a number of genetic variations.

1.8 Objectives of thesis

The large biodiversity available in the germplasm of wild potato species and the information produced from the Next Generation Sequencing technology will definetively to improve the selection processes but also will contribute to the development of new genetic variability. The potato is the fourth most important crop for humans, but it is a classic example of species with low genetic variability. In fact the cultivated varieties derived from a small number of genotypes. Many studies are oriented towards an efficient utilization of genetic resources, in order to broaden the genetic base of the cultivated potato. Despite the many characters of interest attributed to wild species, ecotypes and local varieties it is not always possible to use the classical breeding methods, which require long times for the selection and laborious methods for their use. The new information based on the structural and functional genomics that can be produced by the new sequencing strategy can be useful for a more adequate and efficient exploitation of this heritage, which is often underutilized. This PhD thesis work has been developed within the research programs come out at the Department of Agriculture in Portici aimed at protecting, enhancing, study and employment of wild potato germplasm, with particular attention to the development of new genomic tools applicable to the potato genetic improvement. Part of this work was funded by the Italian Ministry of University and Research (MiUR)- PON02 R&C 2007–2013 PON02_00395_3215002 GenHORT (D.D. n. 813/Ric.) and by *Sicurezza, Sostenibilità e Competitività delle Produzioni Agroalimentari della Campania (CARINA)*.

Specifically, the objectives of the work carried out can be summarized as follows:

1. Development of *de novo* sequence of *S.commersonii* and gene annotation. Towards this goal, a high-throughput genome and transcriptome analysis was carried out using an Illumina technology and different bioinformatics tools. We have studied the genome composition in terms of number of annotated genes, level of heterozygosity and number of repetitive sequences.
2. Study the effects of duplication events during the evolution of *S. commersonii*. We have extrapolated the orthology and paralogy information to assess evolutionary relationships between *S. commersonii* and other sequenced plant genomes.
3. Cold responsive gene analysis and Pathogen-Receptor Genes annotation. A new bioinformatic tool was experimented for the annotation of the previous reported classes of genes. Moreover, transcriptome analysis was carried out using a custom microarray able to detect all the expressed genes in two different experimental conditions.

2. Materials and Methods

2.1 Genetic background of sequenced material

We sequenced the genome of clone cmm1t from *S. commersonii* (accession PI243503), derived from a single seed. This clone has been widely characterized and used at the University of Naples as source of resistance genes to biotic and abiotic stress (Carputo et al., 1997; 2007; 2009; 2013). To produce plant material for this study, one-month old plants were transferred from *in vitro* cultures into styrofoam trays filled with sterile soil and acclimated to *ex vitro* conditions in a growth chamber at 18-20°C (day/night). After two weeks, they were transferred to 5-cm-diameter plastic pots and grown in a temperature-controlled (20–24°C) greenhouse. DNA from leaves was purified using DNeasy Plant Maxi Kit (Qiagen, Valencia, USA) according to manufacturer's instructions.

2.2 Library construction, sequencing, and quality control

A total amount of 2.5 µg of genomic DNA was sonicated with a Covaris S2 instrument (Covaris, inc., Woburn, MA) to obtain fragments ranging from 200bp to 1000bp in length. Preparation of *S. commersonii* DNA libraries was carried out using the TruSeq DNA Sample Prep Kit v2 (Illumina, San Diego, CA) accordingly to manufacturer's instructions. Libraries were size selected at 400bp, 550bp and 700bp on 1.5% agarose gel cassettes using a Pippin Prep instrument (Sage Science, Beverly, MA). Preparation of *S. commersonii* cDNA libraries was carried out starting from 2.5 µg of total RNA extracted from leaf tissue grown under the conditions specified above. The RNA was purified using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA) accordingly to manufacturer's instructions. Mate-pair libraries of 3Kb, 5Kb and 10Kb target insert sizes were constructed by Fasteris SA (Geneva, Switzerland) using an in-house modified Roche MP protocol. Quality control of libraries was performed using High Sensitivity DNA Kit (Agilent, Wokingham, UK). Libraries were quantified using qPCR with a KAPA Library Quantification kit (KapaBiosystems, USA). cDNA libraries were then pooled and sequenced using Illumina HiSeq 1000 with TruSeq SBS Kit v3-HS and TruSeq PE Cluster Kit v3-cBot-HS kits (lllumina, USA) generating 100-bp paired-end sequences. To estimate sequencing depth, we counted the copy number of a certain K-mer (e.g., 17-mer) present in sequence reads, and plotted the distribution of copy numbers. The peak value of the frequency curve represents the overall sequencing depth. We used the algorithm: $(N \times (L - K + 1) - B)/D = G$, where N is the total sequence read number, L is the average length of sequence reads and K is K-mer length, defined as 17 bp here. To minimize the influence of sequencing error, K-mers with low frequency (<4) are discarded. B is the total number of low frequency 17-mer. G denotes the genome size, and D is the overall depth estimated from K-mer distribution. Genome size was calculated using the total length of sequence reads divided by estimated sequencing depth.

2.3 Read filtering

To obtain a set of high-quality reads, all sequenced reads from *S. commersonii* were preprocessed by first discarding reads with more than 10% of undetermined bases or with more than 50 bases of qualities lower than 7; duplicated reads were discarded as well. All reads were preprocessed to clip sequencing adapters with scythe (https://github.com/vsbuffalo/scythe). After clipping, the 3' ends of reads were quality trimmed with a quality threshold of 20 over a window of 10 bases with sickle (https://github.com/najoshi/sickle). Mate-pair reads were further filtered with Deloxer (Van Nieuwerburgh et al., 2012) (http://genomes.sdsc.edu/downloads/deloxer/) to identify and discard unpaired and paired-end reads.

2.4 Genome size estimation

We estimated the genome size of *S. commersonii* using the using flow cytometry. *S. commersonii* and *Glycine max* nuclei were isolated, propidium iodide-stained and analyzed simultaneously (Doležel et al., 1998). Soybean (*G. max* 'Polanka', 2C= 2.50 pg DNA) served as internal reference standard. The absolute DNA amount of *S. commersonii* was calculated on the values of G1 peak means as follow: (G1 peak means *S. commersonii*/G1 peak means of *G. max*) × G. max DNA content.

2.5 Genome assembly and SNP calling

High quality reads from the paired-end libraries were assembled into contigs using SOAPdenovo v2.04 (Luo et al., 2012), a tool belonging to the SOAP (Short Oligonucleotide Analysis Package) suite (http://soap.genomics.org.cn/), with multiple k-mers between 79 and 99; paired-end and mate-pair libraries were used for scaffolding by increasing library size. Gaps that resulted from the scaffolding were closed using GapCloser v1.12 (a SOAP suite tool) and all sequences shorter than 1000 bases were discarded from the final assembly.

The gene space of the assembled genome was assessed by aligning Core Eukaryotic Genes (CEGs) (Parra et al., 2009) over the assembly using BLAST (Altschul et al., 1990) with a 65% identity threshold. High quality paired-end reads were aligned to the assembled genome using SOAPaligner v2.21 (a SOAP suite tool) with standard parameters but "-r" option set to "0" in order to avoid reporting of any repeat hits. To detect SNPs, SOAPsnp v1.03 (a SOAP suite tool) was used with standard parameters but "-u" and "-n" options enabled to give a better accuracy to heterozygous

SNP detection. Further filtering conditions were set as sequencing depth of the site of more than 10 and less than 300, quality scores of the consensus genotype of more than 20, and mapped best and second-best bases supported by at least four unique reads. Finally, sites with best base calling read count less than four times second-best base calling reads count were identified as heterozygous sites. Only heterozygous sites were considered effective.

2.6 Genome annotation

To investigate the nature of repetitive DNA in *S. commersonii*, we annotated repeat clusters using similarity to known repetitive DNA, using a RepBase library (Jurka et al., 2005) and RepeatMasker (RepeatMasker Open-3.0. URL http://www.repeatmasker.org). The RepeatMasker (http://www.repeatmasker.org) suite (Smit et al., 1996-2004)[i] was used with the public Solanaceae libraries to mask the assembled genome, using default parameters. The assembled masked genome of *S. commersonii* was annotated using the program MAKER pipeline (Holt and Yandell 2011). As evidence of gene annotations we used (i) protein alignments to a set of 162,435 available proteins derived from the species *A. thaliana* (35,386 proteins, TAIR10), *S. tuberosum* (56,218 proteins, PGSC v. 3.4), *S. lycopersicum* (34,727 proteins, ITAG 2.3) and downloaded from the Swiss-Prot Plants protein database (36,104 proteins, 13/04/2013) and a protein identity threshold of 40%; (ii) nucleotide alignments to 548,500 EST sequences coming from *S. commersonii* (67 sequences, NCBI, 17/04/2013), *S. tuberosum* (250,127 sequences, NCBI, 17/04/2013) and *S. lycopersicum* (298,306 sequences, NCBI, 17/04/2013); (iii) nucleotide alignments to 117,816 contigs de novo assembled from the high-quality RNA-seq reads of *S. commersonii* using Trinity release 2013/02/25 (Grabherr et al., 2011) with default parameters but a minimum contig length of 300 bp and at least two independent reads covering each contig; and (iv) predictions from SNAP (Korf 2004) and Augustus (Stanke and Waack 2003), all trained with first iteration of MAKER with (i), (ii) and (iii) evidence and standard parameters, and predictions from GeneMark (Ter-Hovhannisyan et al., 2008), trained using randomly selected scaffolds covering about 40 Mbps, in accordance with author's instructions. MAKER was run twice consecutive, each with (i), (ii) and (iii) external evidence and (iv) predictions; the gene models from the second and last iteration with Annotation Edit Distance (AED) (Yandell and Ence 2012) higher than 0.5 were discarded from the final annotation. Predicted open reading frames (ORFs) from the annotation were aligned against the NR database (06/2012 release) with BLAST (BlastP, E-value < 10-5) and functionally annotated by automatic annotations performed with Blast2GO (Conesa and Götz 2008) .

2.7 Comparative genome analyses

The OrthoMCL pipeline (Li et al., 2003) was used to identify and estimate the number of paralogous and orthologous gene clusters between *S. commersonii*, *S. tuberosum* and *A. thaliana*. Standard settings (BLASTP, E-value < 10-5) were used to compute the all-against-all similarities. Syntenic blocks (≥5 genes per block) between *S. commersonii* and *S. tuberosum* were identified using MCScanX (Wang et al., 2013) based on the orthologous and co-orthologous gene pairs found by OrthoMCL pipeline.

2.8 Long non-coding RNA and miRNA annotation

Raw reads coming from RNA-seq performed in root, tuber, leaf and flower samples were checked for quality with FastQC v0.10.1 (www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimming and removal of adapters were performed with AdapterRemoval 1.5.2 (Lindgreen et al. BMC Research Notes 2012) and FASTX Toolkit 0.0.13.2 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). High quality reads (minimum length 35bp, minimum base quality 25) were then mapped against the *S. commersonii* genome sequence with TopHat v2.0.11 (Kim et al., 2013), providing the reference gene annotation file with known transcripts. Duplicated reads were then removed from the mapped files (bam files) with Picard Tools 1.110 (http://picard.sourceforge.net) and the resulting files were used to annotate new transcripts with Cufflinks v2.2.0 (Trapnell et al., 2010). The resulting assembly files were merged and compared to the reference with Cuffmerge v1.0.0 (Trapnell et al., 2010). A new annotation file was created by removing the isoforms contained in other isoforms and those belonging to the class "s" as reported by Cuffmerge. Long non-coding RNAs (lncRNAs) were identified using the approach described by Boerner and McGinnis (2012, PLOS One, 10.1371/journal.pone.0043047). In order to distinguish truly lncRNA from precursors of other ncRNA, the set of lncRNAs was first analyzed with cmscan (e-value 0.01) from Infernal 1.1 (Nawrocki and Eddy 2013) against the database of covariate models of Rfam 11.0. Non-coding transcripts were blasted as well against a database of plant mature miRNA sequences from miRBase (http://www.mirbase.org/) to identify homologous miRNAs. Start and end positions of the match between the query and the hit were retrieved to check whether those positions correspond to miRNAs using MIReNA (Mathelier and Carbone 2010). The transcripts annotated as rRNA, tRNA, miRNA, or other ncRNA by cmscan and those validated positively by MIReNA were excluded. The remaining transcripts were analyzed with MIReNA without providing any position in order to identify novel putative pre-miRNAs. The remaining transcripts were considered lncRNAs. Cufflinks v2.2.0 (Trapnell et al., 2010) was then used to obtain RPKM expression levels for each annotated transcript. miRNA target prediction was performed by using psRNATarget (Dai and Zhao 2011) with default settings.

2.9 MATRIX RELOADED: a new approach to gene prediction

The high efficiency prediction system MATRIX was developed to analyze large data sets in order to identify sequences with a putative resistance function. A first necessary step in the creation of this new system of analysis, is the subdivision of *Solanaceae* family resistance genes by phylogeny. The genes protein sequences have been collected in one multi fasta file and aligned through MUSCLE program (Edgar, 2004). Then, alignments were manually cleaned up and used as a map for the creation of the modules HMM (Knudsen et al., 2003). The group of genes with the high homology regions were aligned using *perl* written programs. Through *ad hoc* scripts were extrapolated highly conserved regions of the individual R groups of proteins and the HMM profiles were created using the *hmmbuilt* of the package HMMER version 2.3.2 (Zhang et al., 2003) and calibrated by *hmmcalibrate*. The constructed profiles were then aligned with unknown proteins using *hmmalign*, in order to identify putative resistance genes. To automate the alignment process of the HMM profiles a *perl*, written program was used to assolve the following goals: read the protein sequences in sequential multi fasta files; align one at a time all HMM profiles against the proteins; produce an homology value between the single profile and the protein, using a BLOSUM62 substitution based matrix. The results of analysis is a matrix of numbers in which homology of all the proteins is analysed against all the HMM profiles. Despite of the software complex performances we were able to build up an optimized system, greatly reducing the computational time and, at the same time, process a whole genome (consisting of about 30,000 proteins) in about 3-5 minutes.

2.10 Identification of resistance genes

Identification of abiotic resistance genes in *S. commersonii*, *S. tuberosum*, *S. lycopersicum* and *A. thaliana* was based on (i) a manually edited list of GO terms found by Blast2GO related to "response stress" and "defense"; (ii) comparison to *A. thaliana* sequences from database STIFDB V2.0 (Stress Responsive Transcription Factor Database) (Naika et al., 2013) based on BLAST (BlastP), with filtering conditions set as E-value < 10-5, query length equal or more than 90% of hit length and equal or less than 110% of hit length, alignment length equal or more than 90% of query length, and, retention of only the best hit, if any. To detect abiotic resistance genes, proteins of *S. commersonii*, *S. tuberosum*, *S. lycopersicum* were subjected to both criteria (i) and (ii). By contrast, proteins of *A. thaliana* were subjected only to criterion (i) and were integrated with *A. thaliana* sequences from database STIFDB V2.0. Identification of biotic resistance genes (R-genes) in *S. commersonii* was based on (i) a manually edited list of GO terms found by Blast2GO related to "response stress" and "defense"; (ii) comparison to sequences from database PRGDB (Plant Resistance Gene Database) (Sanseverino et al., 2010) based on BLAST (BlastP), with filtering conditions set as E-value < 10-5, query length equal or more than 90% of hit length and equal or less than 110% of hit length, alignment length equal or more than 90% of query length, and, retention of only the best hitm, if any; and (iii) screening of conserved motif structure for NBS (PFAM: PF00931), KIN (PFAM: PF00069), TIR (PFAM: PF01582) LRR (PFAM: PF00560) and CC domains with HMMER V.3 (http://hmmer.janelia.org/software) and PAIRCOIL2 (McDonnell et al. 2006) to specifically detect CC domains.

2.11 R-Genes analysis

Matrix-R is a custom-made pipeline written in perl to automatically retrieve, annotate and classify plant resistance genes. Translated proteins sequences of 91 cloned R-genes falling in the major four R-classes (http://prgdb.crg.eu/) were used as the starting point of the pipeline. Protein sequences belonging to single R-classes were aligned using MUSCLE 3.6 (Edgar, 2004) with manual editing. The resulting alignments for each group were used as a base for the creation of an aligned subset of conserved regions and a set of hidden Markov models (HMMs) using the HMMER v3 package (http://hmmer.janelia.org/). A total of 60 HMM were build, 15 for the CNL class, 24 for the TNL class, 8 for the RLK class and 13 for the RLP class. The input dataset of Matrix-R is a plant proteome. For each protein, Matrix-R calculates: 1) the coils potential, with COILS (Lupas et al. 1991;) to detect CC domains; 2) the putative Transmembrane domains with TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) to infer the putative protein localization; 3) the matching score with the HMM modules previously created. According to user defined thresholds, all proteins with a significant match with the HMM modules were stored and then assigned to R-classes on the basis of the type of HMM, the presence/absence of coils and/or transmembrane domains. Matrix-R was used to screen the proteomes of *S. commersonii* and *S. tuberosum* (37.662 and 39.031 proteins, respectively). Annotated genes (39.031) from the PGSC whole genome annotation of DM assembly were used (PGSC_DM_v3_superscaffolds.fasta.zip; http://potatogenomics.plantbiology.msu.edu/index.html). The set of predicted proteins identified via HMM profiling was further analyzed using INTERPROSCAN software version 5.0 (http://www.ebi.ac.uk/Tools/pfa/iprscan5/) to verify the presence of conserved domains and motifs characteristic of R-proteins (Nucleotide Binding Sites, NBS; Leucine Rich Repeats, LRR; Toll-Interleukin receptor, TIR; KINASE; SERINE/ THREONINE).

2.12 Cold resistance gene analysis

To annotate putative cold resistance genes in *S. commersonii*, a set of reference proteins were selected from *A. thaliana*. In detail, 58 proteins annotated with the Gene Ontology term *cold acclimation* (CA), 28 proteins annotated as *cellular response to cold* (CRTC), and 619 proteins as *response to cold* (RC) were selected. INTERPROSCAN was used to identify the domains of the proteins included in each gene family (Table 2.1).

Table 2.1 Number of non-redundant protein families annotated with the Gene Ontology term cold acclimation (CA). as cellular response to cold (CRTC). and proteins as response to cold (RTC) and related number of proteins in A, thaliana and S, tuberosum

GO category	Protein families	*A. thaliana* proteins	*S. tuberosum* proteins
CA	17	177	239
RTC	146	1429	2833
CRTC	10	96	208
Total	173	1702	3280

The proteins showing the same domain composition were grouped and aligned using MUSCLE 3.6 (Edgar, 2004) thus generating a consensus sequence. For each protein group, the generated consensus sequence was used to interrogate the proteome of *A. thaliana* (*ftp://ftp.arabidopsis.org/home/tair/Proteins/TAIR10_protein_lists/TAIR10_pep_20101214*) with a BlastP (E-value ≤ 10-3), in order to identify all the proteins with the same domain composition. The same analysis was carried out in *S. tuberosum* (*http://potato.plantbiology.msu.edu/data/PGSC_DM_v3.4_pep.fasta.zip*). For each protein family, the identified *A. thaliana* and the *S. tuberosum* sequences were aligned and the resulting alignments were used as a base for the creation of hidden Markov models (HMMs) using MATRIX RELOADED (Supplemental dataset 7 online: *https://drive.google.com/open?id=0BzNl4eO_bJumMDVNVXhUS1FVNEU&authuser=0*). The obtained HMM modules were used to identify putative cold responsive proteins in *S. tuberosum* and *S. commersonii.* Several filtering steps were then performed to remove false positives. First, protein blast were performed against the proteins used to create the HMM modules, with filtering conditions set as e-value ≤ 10-5 and the alignment length being at least 90% of the query length. Secondly, a promoter analysis was performed to identify genes having putative promoter binding sites for transcription factors related to response to cold, as reported by *Kyonoshin et al*. In detail, using the (*PGSC_DM_V403_genes.gff.zip*) gene annotation of *S. tuberosum* and of *S. commersonii*, 1000 bp upstream of each gene were extracted and analyzed for the presence of cold responsive transcription factor binding motifs. To produce the final dataset of proteins putatively involved in response to cold, only the proteins encoded by genes with cold responsive motifs were kept and when two or more predicted proteins resulted from alternative splicing events, only the longest protein was retained. All proteins were then classified according to the type of HMM and the presence/absence of the previously identified domains.

2.13 Transcriptional analysis

Twelve clonally propagated plants from cmm1t (PI243503) were cultured in a growth chamber under cool white fluorescent lamps (350-400 mmol $m^{-2}s^{-1}$) at 24°C and than exposed at -2°C for 6 hours to test their resistance to low temperature in non acclimated (NAC) conditions. To evaluate the resistance following acclimation (AC test) six plants were first transferred from a 24°C growth chamber to a cold room (4°C) under cool white fluorescent lamps (100 mmol $m^{-2}s^{-1}$) for two weeks and then exposed to -2°C for 6 hours. For each test, RNA was isolated from a 100 mg of leaf tissue pooled from five different plants. Pooled tissue was homogenized (TissueLyzer by Qiagen) using a TRIZOL reagent (Life Technologies) and RNA was extracted following TRIZOL Life Technologies protocols. The concentration and purity of extracted RNAs were estimated using the NanoDrop spectrophotometer (Thermo Fisher Scientific). The quality and the integrity of RNA were checked after electrophoresis 1 ìg of RNA samples on 1% agarose gel stained with SYBR® Safe (Life Technologies). The synthesized and labeled antisense-RNA (aRNA) was performed using the Kreatech's kit RNA ampULSe: Amplification and Labeling Kit for CombiMatrix (Kreatech Biotechnology, Amsterdam, The Netherlands) arrays with Cy5 dye. The purified, labeled aRNA was quantified by spectrophotometer and 4 ìg was hybridized to the Combimatrix array according to the manufacturer's directions. Pre-hybridization, hybridization, washing and imaging were performed according to the manufacturer's protocols (http://www.combimatrix.com/support_docs.htm). Imaging of array slides was performed using a GenePix® 4400A Microarray Scanner controlled by the GENEPIX PRO V.7 software (Molecular Devices) at 5ìm resolution. The GENEPIX PRO v.7 software was also used for densitometry analysis and raw data extraction. Probe signals higher than negative control values plus twice the standard deviation were considered as 'present'.

The analysis was performed on a Combimatrix *S. tuberosum* chip produced by the Plant Functional Genomics Center at the University of Verona. The chip contained 27.234 non-redundant probes in triplicate, composed of 35-40-mer oligos. Probes were designed on tentative consensus sequences (TCs; 23.453 probes) and singletons with a 3' poly(A) tail (46 probes) derived from the SolEST database (D'Agostino et al., 2009) using Oligoarray 2.1 (Rouillard et al., 2003). Oligo probes were designed to identify the 3'-UTR region of genes. Results from BLASTx comparisons against the UniPortKB/Swiss-Prot database were exploited to determine the correct open reading frame and to define forward/reverse TC orientation. 13.207 TC sequences had a forward orientation, while 2.027 had a reverse orientation. In the case of 9.000 TC sequences no BLAST hits were found and it was not possible to assess where their 3'-UTR region was located for these sequences. As a consequence, we filtered out 6.000 TCs generated by assembling the largest number of ESTs and considered both the orientations for probe design. Nine bacterial oligonucleotide sequences provided by CombiMatrix, 40 probes designed on seven Ambion spikes and 11 additional negative probes based on *Bacillus anthracis*, *Haemophilus ducreyi* and *Alteromonas phage* sequences were used as negative controls. Three to four replicates of each probe were distributed randomly across the array. Three technical and three biological replicates were used for each hybridization experiment.

Data analysis was performed by using the R package limma (Smyth, 2005). The median of the signal was used for the analysis. Flagged spots with a value of -50 were downweighted. Hierarchical clustering based on the euclidean distance between the samples checked the good quality of the replicates. Samples not clustering with their corresponding replicates were discarded. Maximum likelihood normexp was used for background correction, and the arrays were normalized by quantile normalization. Fitting a linear model including the correlation between replicated probes followed by bayesian test performed identification of differentially expressed probes. Raw p-values were adjusted for multiple corrections with Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Adjusted p-values <=0.05 were considered statistically significant. The TC sequences used to design the Combimatrix probes were blasted (BLASTn, e-value < 0.01) against the transcriptome of *S. commersonii* in order to determine the match between probes and annotated loci.

To validate the microarray data, we performed real time PCR analysis for three cold-regulated genes from our list. These included COR413 (SOLTUB01G046490), Histone demethylase (SOTUB05G023460.1.1), and Histone-lysine N-methyltransferase (SOTUB10G019470.1.1). The qPCR results showed that the three genes are all cold regulated, and their expression kinetics is very similar to those obtained from microarray analysis. These results support the validity of the *S. commersonii* cold-regulated transcriptome from the PotatArray analysis.

2.14 Phylome reconstruction

Proteins encoded in 12 fully-sequenced plant genomes, including the wild-potato transcriptome, were downloaded from various sources (Table 2.1).

Table 2.1. Detected one-to-one orthologs between a given species and *S. commersonii*

Species Name	one-to-one orthologs
S. tuberosum	17.297
S. lycopersicum	16.821
M. guttatus	7.058
B. vulgaris	6.799
C. melo	6.684
A. thaliana	5,.862
G. max	1.667
T. aestivum	1.160
Z. mays	3.913
B. distachyon	4.968
O. sativa subsp. *japonica*	4,492

The final database used for the phylome reconstruction contained 37,477 unique protein sequences for *S. commersonii*. The resulting phylome comprises 35,182 gene trees, representing 93.88% of the predicted proteins. To perform the phylome reconstruction, a Smith-Waterman (Smith and Waterman, 1981) search was used to retrieve homologs using an e-value cut-off of 1e-5 and considering only sequences that aligned with a continous region representing more than 50% of the query sequence. Then, selected homologous sequences were aligned using three different programs: MUSCLE v3.8 (Edgar, 2004), MAFFT v6.712b (Katoh and Toh, 2008), and Kalign v2.04 (Lassmann et al., 2009). Alignments were performed in both forward and reverse directions (i.e using the Head or Tail approach (Landan and Graur, 2007)), and the six resulting alignments were combined using M-Coffee (Wallace et al., 2006). The resulting combined alignment was subsequently trimmed with trimAl v1.4 (Capella-Gutiérrez et al., 2009), using a consistency score cut-off of 0.1667 and a gap score cut-off of 0.1 to remove poorly aligned regions. Phylogenetic trees based on a Maximum Likelihood (ML) approach were inferred from these alignments. ML trees were reconstructed using the best fitting evolutionary model. The selection of the evolutionary model best fitting each protein family was performed as follows: a phylogenetic tree was reconstructed using a Neighbour Joining (NJ) approach as implemented in BioNJ (Gascuel, 1997);

the likelihood of this topology was computed, allowing branch-length optimization, using nine different models (JTT, WAG, MtREV, VT, LG, Blosum62, Dayhoff, DCMut and CpREV), as implemented in PhyML v3 (Guindon et al., 2010); the two evolutionary models best fitting the data were determined by comparing the likelihood of the used models according to the AIC criterion (Akaike, 1974). Then, ML trees derived using these two models and the default tree topology search method NNI (Nearest Neighbor Interchange) were compared and the tree with the best likelihood was used for further analyses. A similar approach based on NJ topologies to select the best fitting model for a subsequent ML analysis has been shown previously to be highly accurate. Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML. In all cases, a discrete gamma-distribution with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data.

2.15 Phylogeny-based prediction of orthology and paralogy

Orthology and paralogy relationships among *S. commersonii* genes and those encoded by the other genomes included in the phylome were inferred using a phylogenetic approach (Gabaldón, 2008). In brief, a species-overlap algorithm, as implemented in ETE v2 (Huerta-Cepas et al., 2010), was used to label each node in the phylogenetic tree as duplication or speciation depending on whether or not the descendant partitions have at least one common species (i.e. using a Species Overlap Score of 0). The resulting orthology and paralogy predictions can be accessed through phylomeDB.org (Huerta-Cepas et al., 2014). These predictions have been used in subsequent analyses such as orthology-based functional annotation, identification of gene expansions, or duplication dating.

2.16 Species tree reconstruction and shared genomic content

A phylogeny for the species included in the phylome was inferred using two complementary approaches, which rendered identical topology results. First, a super tree was inferred from all the trees in the phylome (34,633 separate single gene trees) using a Gene Tree Parsimony approach as implemented in the dup-tree algorithm (Wehe et al., 2008). This procedure finds the species topology, which minimizes the number of total duplications implied by a collection of gene family trees, i.e. the phylome. Secondly, 454 gene families with a clear, phylogeny-based, one-to-one orthology present in at least 11 of the 12 species included in the analyses were used to perform a multi-gene phylogenetic analysis. Protein sequence alignments were performed as described above and then concatenated into a single alignment. Species relationships were inferred from this alignment using a ML approach as implemented in PhyML (Guindon et al., 2010), using JTT as the evolutionary model, since in 357 out of 454 gene families this model was the best fit. The tree topology search method was set to SPR (Subtree Pruning and Regrafting). Branch supports were computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution. A comparative analysis, in terms of homology relationships, was performed among the 12 species included in the phylome. To carry out such analysis, a Blast search of all species against all species was performed to retrieve homologous sequences with a cut-off e-value of 1e-5 and coverage of 50%, e.g. aligned region length divided by the total query length. Then, results showing different patterns of homology (i.e. genes present in all species or specifically in Asterids), were computed and plotted in the species tree inferred.

2.17 Phylostratigraphic dating of duplication events

The phylome was scenned to detect and date duplication events, using a previously described algorithm (Huerta-Cepas and Gabaldón, 2011). We focused on events assigned to three different relative evolutionary periods: 1) *S. commersonii* specific, 2) Potato ancestor, 3) *Solanum* ancestor, and 4) Basal to Asterids. Individual trees were scanned and all duplication events that involved the seed and others wild potato proteins were dated. *S. commersonii* proteins duplicated at different ages were analyzed, looking for any functional enrichment. Enrichment analyses for overrepresented GO terms for the dated duplicated proteins compared to the whole set of annotated proteins were performed using FatiGO as implemented on the Babelomics webserver (Medina et al., 2010). A Fisher exact test looking for overrepresented terms in specific sets of proteins against the whole annotated genome was used with a e-value cut-off of 0.01. Then, GO terms redundancy was reduced using the REViGO webserver (Supek et al., 2011), setting a similarity threshold of 0.5, using as quality score the ratio of log odds values, and SimRel as the semantic similarity algorithm.

Focus was given on lineage-specific genome expansions encompassing 5,119 genes (~13.6%). Impacted genes were grouped into clusters with at least 50% of overlap of shared genes. 2,937 (~57.40%) of these genes were assigned to a unique cluster. Only clusters comprising 10 or more proteins were considered in this analysis. Clusters comprising expanded protein families were analyzed, looking for any statistically significant functional enrichment. Functional enrichment is provided for the 10 largest clusters displaying statistically significant enriched terms. Enrichment analyses of overrepresented GO terms for these expanded families compared with the annotate *S. commersonii* genes were performed using FatiGo as implemented on the Babelomics webserver (Medina et al., 2010) using the Fisher exact test for genome comparison and an e-value cut-off of 0.01. Then, GO term redundancy was reduced using the REViGO webserver (Supek et al., 2011) setting a similarity threshold of 0.5, using the ratio of odd log as a metric and SimRel as a semantic similarity algorithm.

2.18 Functional annotation

Protein coding genes predicted in *S. commersonii* genome were functionally annotated using two complementary approaches, one based on protein signatures and the other based on orthology relationships. In the first approach, each protein was inspected, looking for different signatures such as families, regions, domains, repeats and binding sites using InterProScan over a set of more than 10 different databases. Using this approximation, 91,566 gene ontology (GO) terms were assigned to 21,352 proteins. In the case of phylogeny-based analyses, 12,435 one-to-one orthology relationships among *S. commersonii* genes and genes from species used in the phylome with some GO annotation were found. Using these predictions 59,820 GO terms were transferred from the *S. commersonii* counterparts to corresponding proteins in other species. After filtering redundant annotations, 39,574 GO terms were assigned to *S. commersonii* proteins using this approximation. Figure 2.1 shows the overlap between the two approaches. *S. commersonii* proteins duplicated at different ages were analyzed, looking for any functional enrichment.

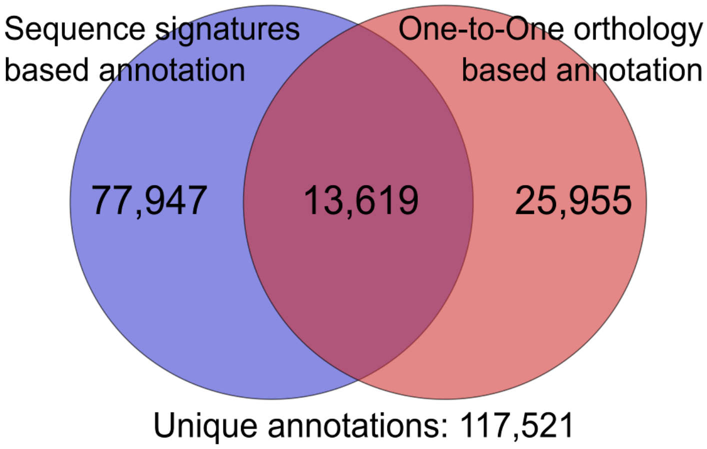


Figure 2.1. Comparison of two strategies for functional annotation of *S. commersonii* proteins.

Enrichment analyses for overrepresented GO terms for the dated duplicated proteins compared to the whole set of annotated proteins were performed using FatiGO as implemented on the Babelomics webserver (Medina et al., 2010). A Fisher exact test looking for overrepresented terms in specific sets of proteins against the whole annotated genome was used with a e-value cut-off of 0.01. Then, GO term redundancy was reduced using the REViGO webserver (Supek et al., 2011), setting a similarity threshold of 0.5, using as a quality score the ratio of log odds values, and SimRel as a semantic similarity algorithm.

2.19 Divergence time analysis

We computed transversion rates at fourfold degenerate sites (4DTv) as a conservative genetic distance to estimate recent major evolutionary events. To assess divergence between species, individual gene trees in the *S. commersonii* phylome were scanned to detect one-to-one orthologs between *S. commersonii* and *S. tuberosum*, and between *S. commersonii* and *S. lycopersicum*. To estimate age of duplication waves, we used paralogous gene pairs assigned to the three relevant evolutionary time points: 1) *S. commersonii* specific, 2) Potato ancestor, and 3) *Solanum* ancestor. Aligned gene pair sequences were extracted from the consensus untrimmed alignments in PhylomeDB and back-translated into their corresponding Coding DNA sequences (CDS) using trimAl v1.3 (Wallace et al., 2006) 4DTv values were computed for gene pairs with at least 10 degenerated sites. Raw 4DTv values were corrected for potential multiple transversions at the same sites using the formula 4DTv_corrected = -1/2 x log(1 - 2 x 4DTv_uncorrected) after (Tang et al., 2008). After discarding cases for which a corrected 4DTv values could not be generated or with few fourfold degenerate sites, 17,539, 21,928 and 13,442 pairs of paralogs assigned to the three mentioned relative time points as well as 15,739 and 16,354 pairs of one-to-one orthologs with *S. tuberosum* and *S. lycopersicum*, respectively, were analyzed and the corrected 4DTv values plotted.

3. RESULTS

Genome sequencing and assembling

To obtain a whole-genome shotgun (WGS) assembly of *S. commersonii* clone cmm1t, we produced size-selected sequencing pair-end and mate-pair libraries based on six insert sizes ranging from 400 bp to 10 kb. A total of 145.93 Gb of sequence reads were produced. After filtering low-quality sequences, the remaining 88 Gb were assembled into 278,460 contigs with an N50 contig length of 6,506 bp (Table 3.1).

Table 3.1. Metrics of *S. commersonii* genome assembly

Genome Assembly Statistics	
N50 index (contigs), number	27,829
N50 length (contigs), bp	6,506
Contig (>100 bp), number	278,460
Large (>500 bp) contig, number	226,195
Longest contig (bp)	170,543
Average contig length (bp)	2,932
N50 index (scaffolds), number	4,833
N50 length (scaffolds), bp	44,298
Longest scaffold (bp)	458,668
Average scaffold length (bp)	13,543

All contigs were further assembled into 64,665 scaffolds (>1 kb), of which 4,833 containing 50% of the assembly were 44.3 kb or larger (N50 = 44,298 kb, Table 3.1 and Table 3.2).

Table 3.2. Summary of the *S. commersonii* genome assembly

	Conting		**Scaffold**	
	Size (bp)	Number	Size (bp)	Number
N90	1,178	146,855	5,763	26,615
N80	2,108	94,918	12,735	15,653
N70	3,258	63,880	21,439	10,432
N60	4,628	42,804	31,743	7,132
N50	6,506	27,829	44,298	4,833
Longest	170,543	-	458,668	-
Total number (>100 bp)	-	278,460		-
Total number (>500 bp)	-	226,195		-
Total number (> 1kb)	-	-		64,655
Total number (> 2kb)	-	-		-

With interactive mapping approach using the potato genome as a reference, the scaffolds were anchored onto each chromosome, resulting in 12 pseudomolecules representing the *S. commersonii* scaffolds linked and ordered according the homology information with *S. tuberosum* (Figure 3.1).

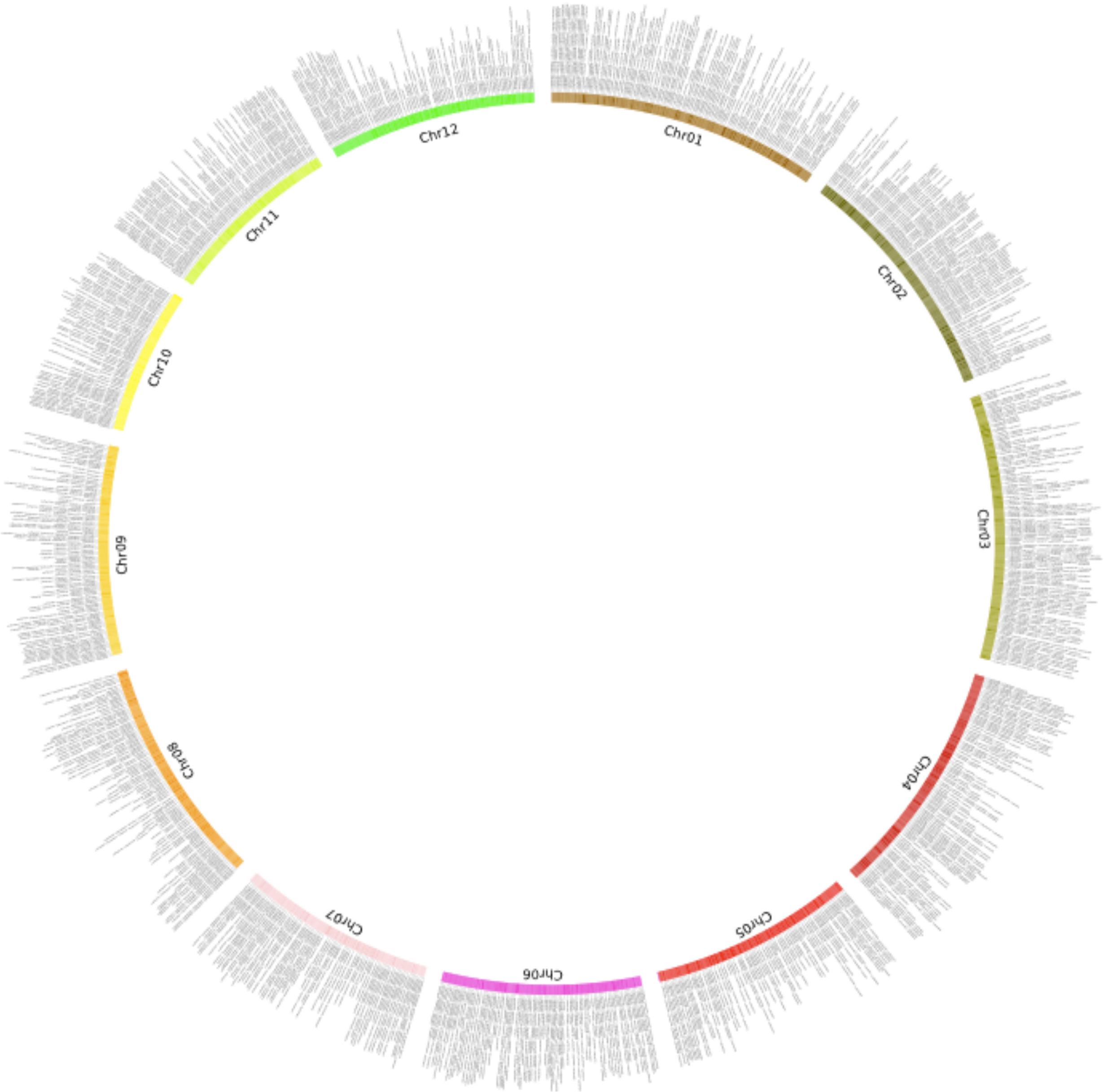


Figure 3.1. Ideograms of the 12 pseudochromosomes of *S. commersonii* (in Mb scales).

The *S. commersonii* genome size was evaluated at ~830 Mb by flow cytometry, consistent with genome size estimation (838 Mb) by the 23-nucleotide depth distributionIn Figure 3.2 the histogram of relative DNA content was obtained after flow cytometric analysis of propidium iodide-stained nuclei of *S. commersonii* and *Glycine max*, which were isolated, stained and analysed simultaneously. Soybean (*Glycine max* 'Polanka', 2C= 2.50 pg DNA) served as internal reference standard. The absolute DNA amount of *S. commersonii* was calculated based on the values of G_1 peak means as follow: (G_1 peak means *S. commersonii*/ G_1 peak means of *G. max*) × *G. max* DNA content. Genome size of the *S. commersonii* was estimated to be 830 Mb.

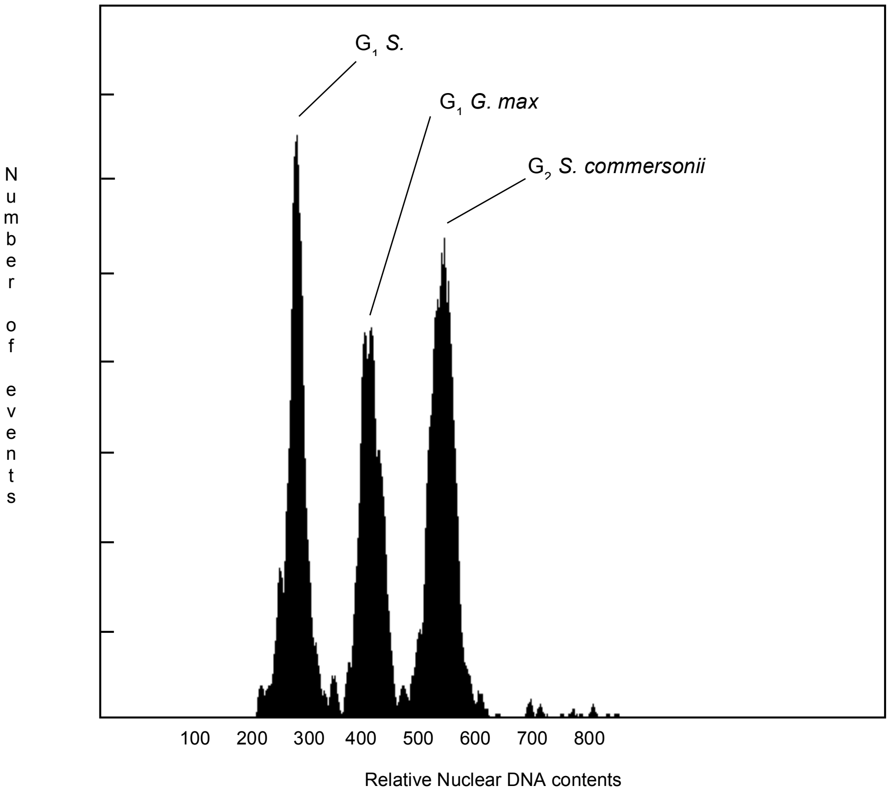


Figure 3.2

Estimation of absolute nuclear DNA amount (genome size) in *S. commersonii*.

The volume of K-mers is plotted against the frequency at which they occur. The left-hand, truncated, peak at low frequency and high volume represents K-mers containing essentially random sequencing errors, while the right-hand distribution represents proper (putatively error-free) data. The total K- mer number is 54.703.986.536, and the volume peak is 64. The genome size can be estimated as (total K-mer number)/(the volume peak), which is 838 Mb. (Figure 3.3). The sum of the Illumina sequences obtained represented ~105x coverage (filtered reads) of the *S. commersonii* nuclear genome. Gaps within scaffolds ranged in length from 1 to 8,369 bp, with a median length of 213 bp (Figure 3.4). The GC content within *S. commersonii* coding DNA sequence was 34.5% Table 3.3). To assess the proportion of the gene space captured in this draft genome assembly, we aligned 248 sequences from the non-redundant core eukaryotic genes (CEGs) to the genome assembly. In total, 243 (98%) CEGs homologs were found in the *S. commersonii* genome, suggesting that the assembly captured a large majority of the gene space (Figure 3.5).

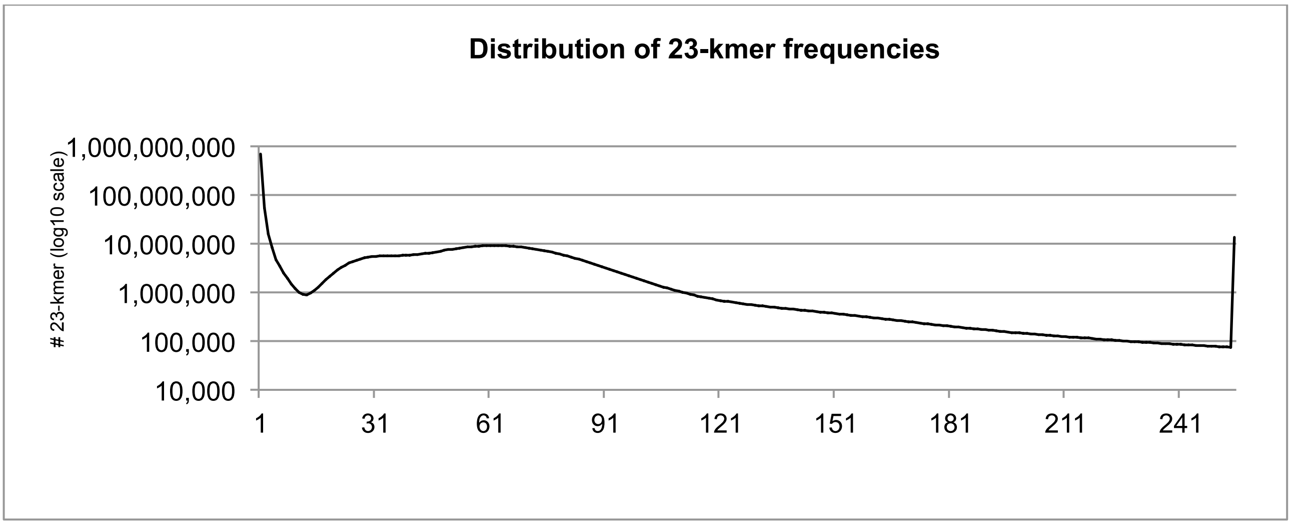


Figure 3.3 Distribution of Illumina 23 k-mer frequency for *S. commersonii*.

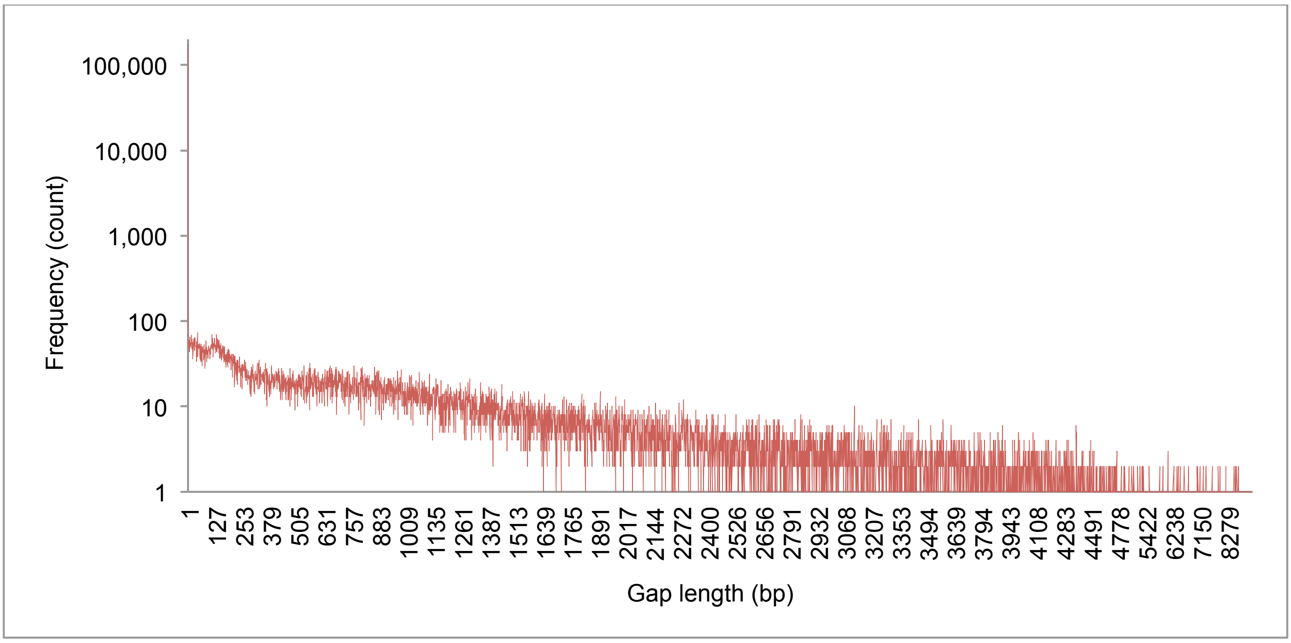


Figure 3.4. Distribution of gap length within the scaffold assembly of *S. commersonii*

Table 3.3 CG content in *S. commersonii* genome

Feature	# A	# C	# G	# T	# N	%GC content
Total	267,803,084	141,392,099	140,663,862	266,680,757	45,924,484	34.54%
Intergenic	212,916,516	109,799,013	109,139,327	211,976,646	40,869,266	34.01%
Genic	54,886,568	31,593,086	31,524,535	54,704,111	5,055,218	36.55%
Intronic	36,299,327	18,750,225	18,721,983	36,136,268	5,050,255	34.09%
Exonic	18,587,241	12,842,861	12,802,552	18,567,843	4,963	40.84%

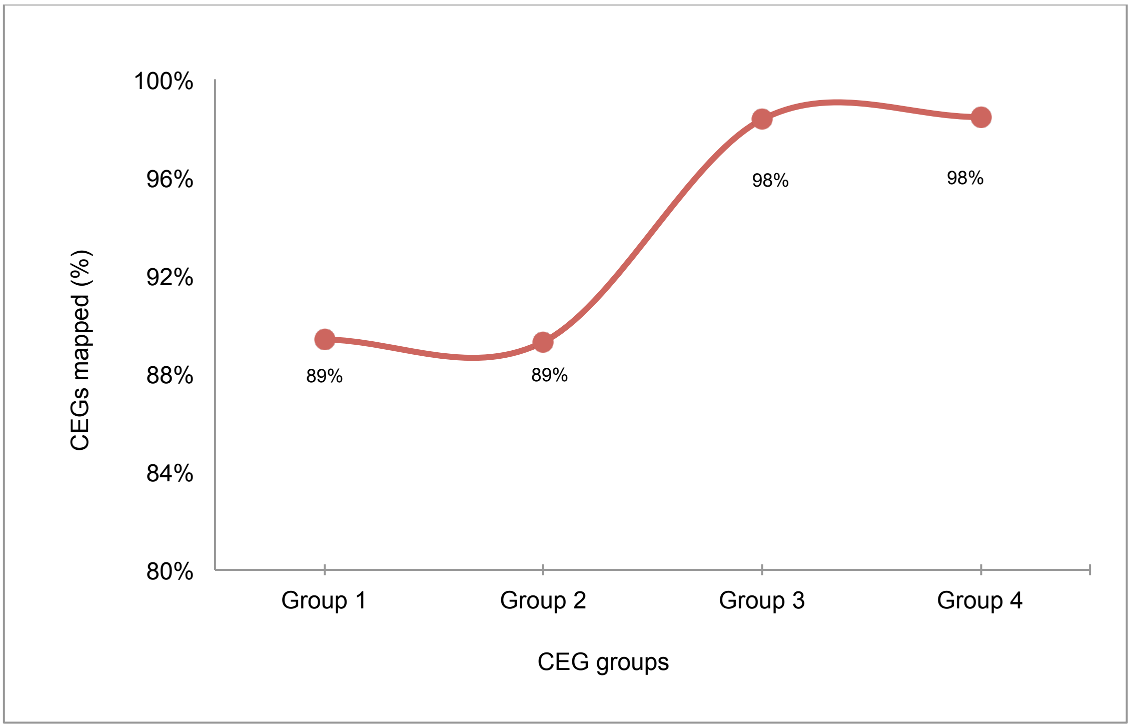


Figure 3.5

Percentage of Core Eukaryotic Genes (CEGs) mapping on the *S. commersonii* draft genome. Group 1 represents the least conserved genes while Group4 the most conserved. Overall, 233 out of the 248 CEGs were detected (94%).

Genomic variations

Compared with the cultivated potato, the *S. commersonii* genome showed a lower level of heterozygosity (Hirsch et al., 2013). A total of 9,894,571 reliable single-nucleotide polymorphisms (SNPs) were identified among 662,040,919 reliable genome bases (Table 3.4), yielding a SNP frequency of 1.49%.

We evaluated the structural and functional impact of SNPs. Ninety-two per cent of all SNPs had a distance of less than 50 bp to its nearest neighboring SNP. Overall 12,412 genes encompassed SNPs, of which 11,608 had a SNP rate smaller than 1% (Figure 3.6; Figure 3.7).

Table 3.4 Heterozygosity in *S. commersonii* genome

Features	Bases affected	Length	Frequency
Genome*	9,894,571	662,040,919	1.4946%
Gene	261,398	149,307,299	0.1751%
Intron	159,793	99,092,644	0.1613%
Exons	141,821	50,152,571	0.2828%
3' UTR	14,216	4,660,982	0.3050%
5' UTR	10,594	4,070,026	0.2603%
UTR	24,810	8,731,008	0.2842%
CDS	117,011	41,421,563	0.2825%

* only reliable position are considered, not the whole genome

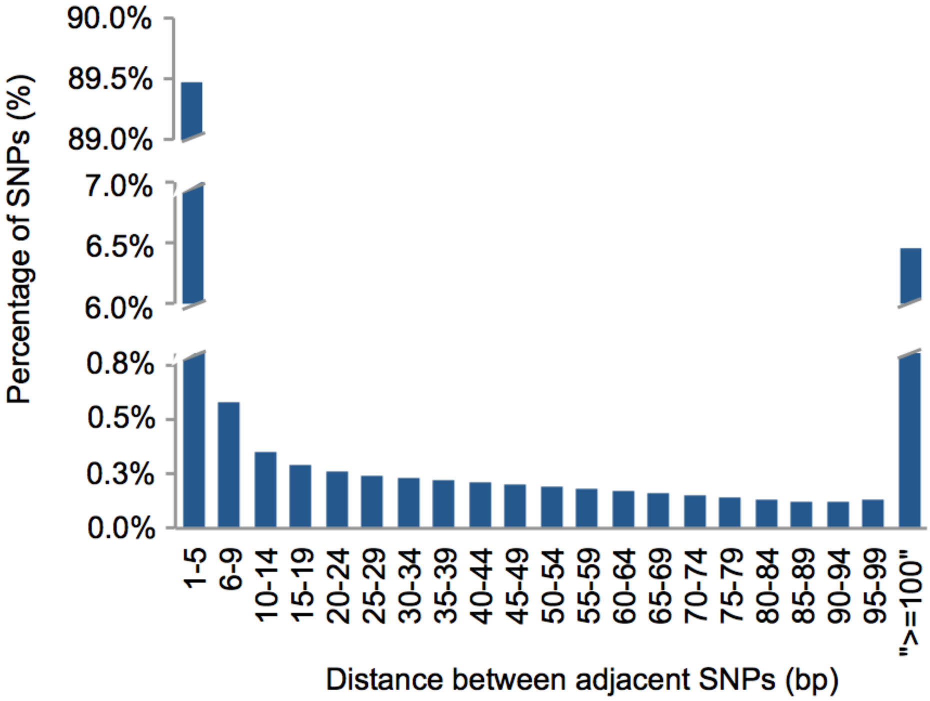


Figure 3.6 SNP spacing in the *S.commersonii* genome. The distribuition of distance between SNPs.

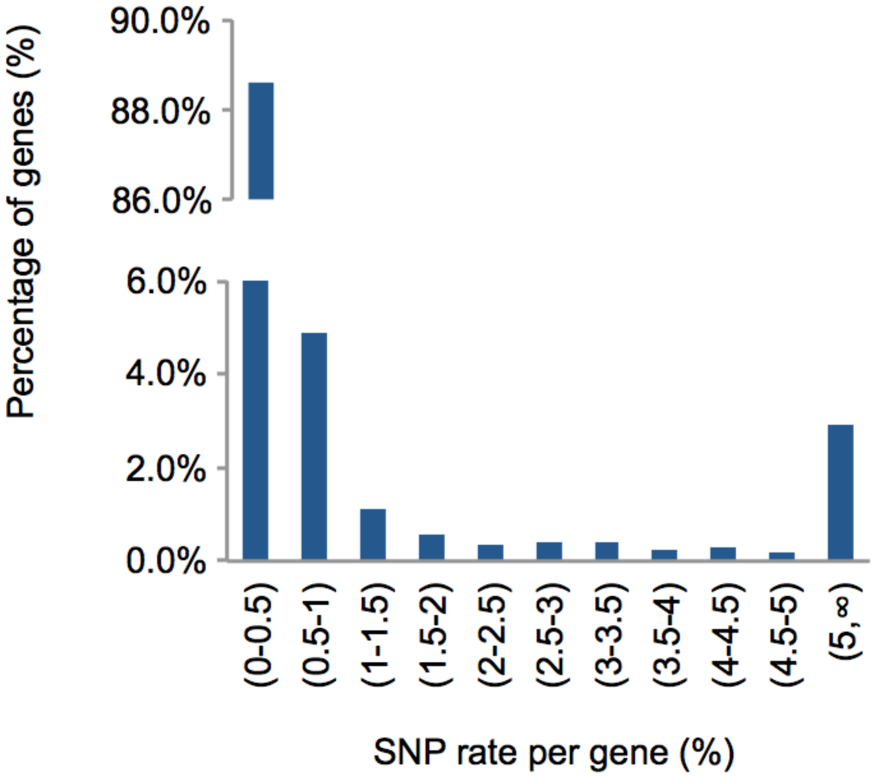


Figure 3.7 SNP spacing in the *S.commersonii* genome. SNP frequency per gene

With regard to functional annotation, most of the identified SNPs (84%) were located in intergenic regions (Table 3.5). The 12,412 SNP-containing genes encompass overrepresentation of some major functional categories, including macromolecule metabolic processes, response to stimuli, carbohydrate derivative binding, localization, and ion binding (Supplemental dataset 1 online: *https://drive.google.com/open?id=0BzNl4eO_bJumYkt1eE9fWURLOWs&authuser=0*).

Table 3.5. Annotation of SNPs detected in *S. commersonii*

SNP Effect*	Count	Genes Affected, number	Percentage, %
Intergenic	8,340,599	-	84.29
Intragenic	70,012	-	0.71
Upstream	294,797	-	2.98
Downstream	375,589	-	3.8
Intron	281,752	19,142	2.85
UTR_5_prime	18,865	2,199	0.19
UTR_3_prime	24,747	2,856	0.25
Splice site acceptor	3,017	1,710	0.03
Splice site donor	3,027	1,697	0.03
Start lost	1,687	1,037	0.02
Non synonymous start	459	462	0
Stop lost	1,546	914	0.02
Stop gained	25,404	4,127	0.26
Non synonymous coding	330,095	16,571	3.34
Codon change	1,017	269	0.01
Synonymous start	2	2	0
Synonymous stop	298	289	0
Synonymous coding	106,405	13,196	1.08
Not processed**	15,253	-	0.15
Total SNP	9,894,571		

*only the most deleterious effect for each SNP is considered, thus every SNP is counted one time

** variants that software (SnpEff) cannot classify due to java errors

The genome size difference between *S. commersonii* (830 Mb) and *S. tuberosum* (838 Mb) was mainly due to differences in intergenic sequence length (Figure 3.8).

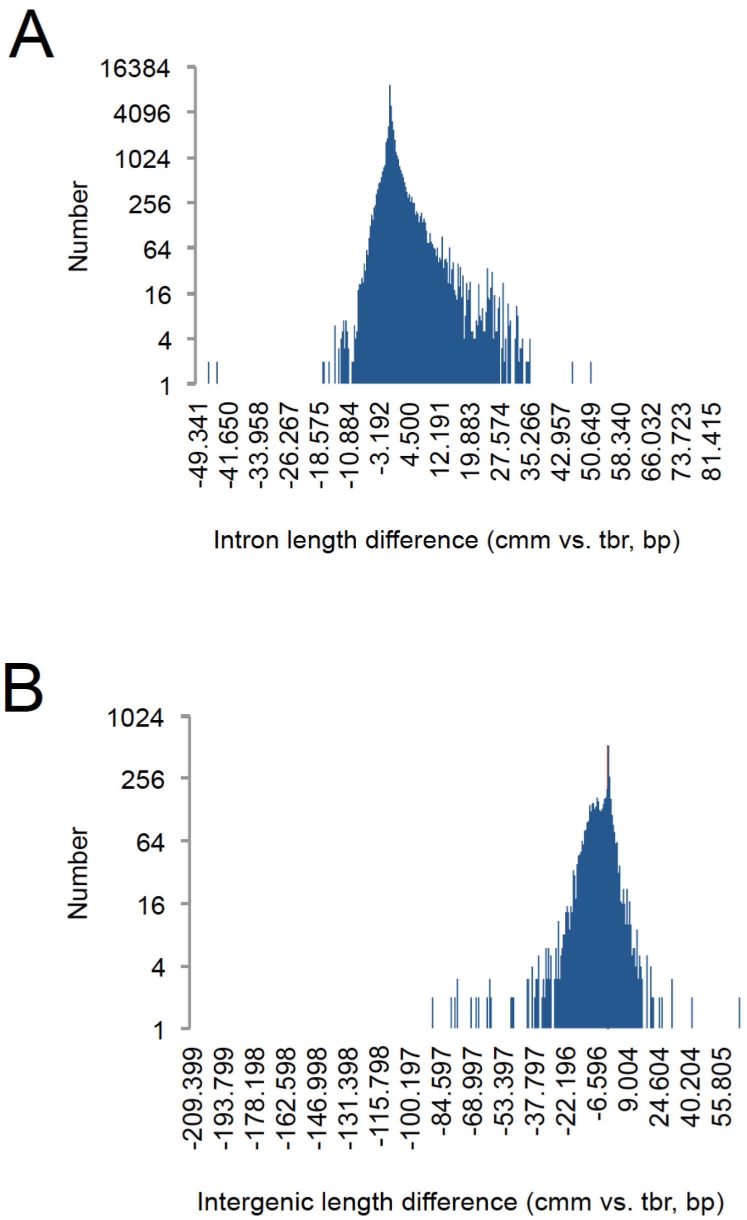


Figure 3.8 Influence of introns and intergenic regions to genome size variation.

A. Differences in intron size of orthologous genes between *S. commersonii* (cmm) and *S. tuberosum* (tbr)

B. Differences in the size of orthologous intergenic regions between *cmm* and *tbr*

The results of microsynteny analyses revealed greater frequency of SNPs in and insertion-deletion events (indels) spanning intergenic regions, consistent with this observation. Roughly 383 Mb of repetitive sequences were identified, accounting for 44.5% of the current assembly of the *S. commersonii* genome. Compared to potato (Potato Genome Sequencing Consortium 2011), *S. commersonii* has a lower amount of repetitive DNA (44.5% vs. 55%), which might explain its smaller genome size and predict different genome dynamics in these two species since their separation from a common ancestor (Figure 3.9). The repetitive fraction of the *S. commersonii* genome assembly is dominated by Long Terminal Repeat (LTR)-retrotranspons (~34%) with lower levels of several other repeat types (Figure 3.10).

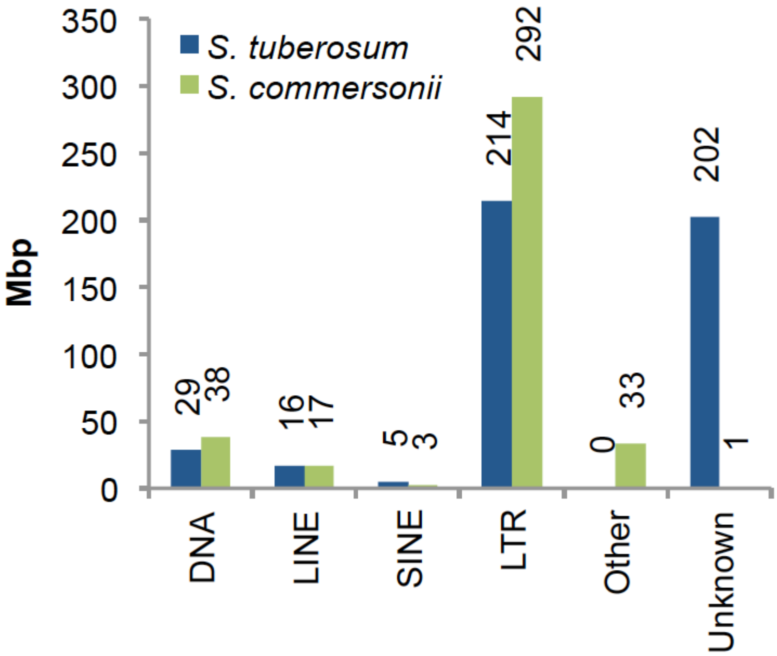


Figure 3.9. Repetitive sequence annotation in the draft genome of *S. commersonii*. . Length of transposable elements in *S. commersonii* and *S. tuberosum*

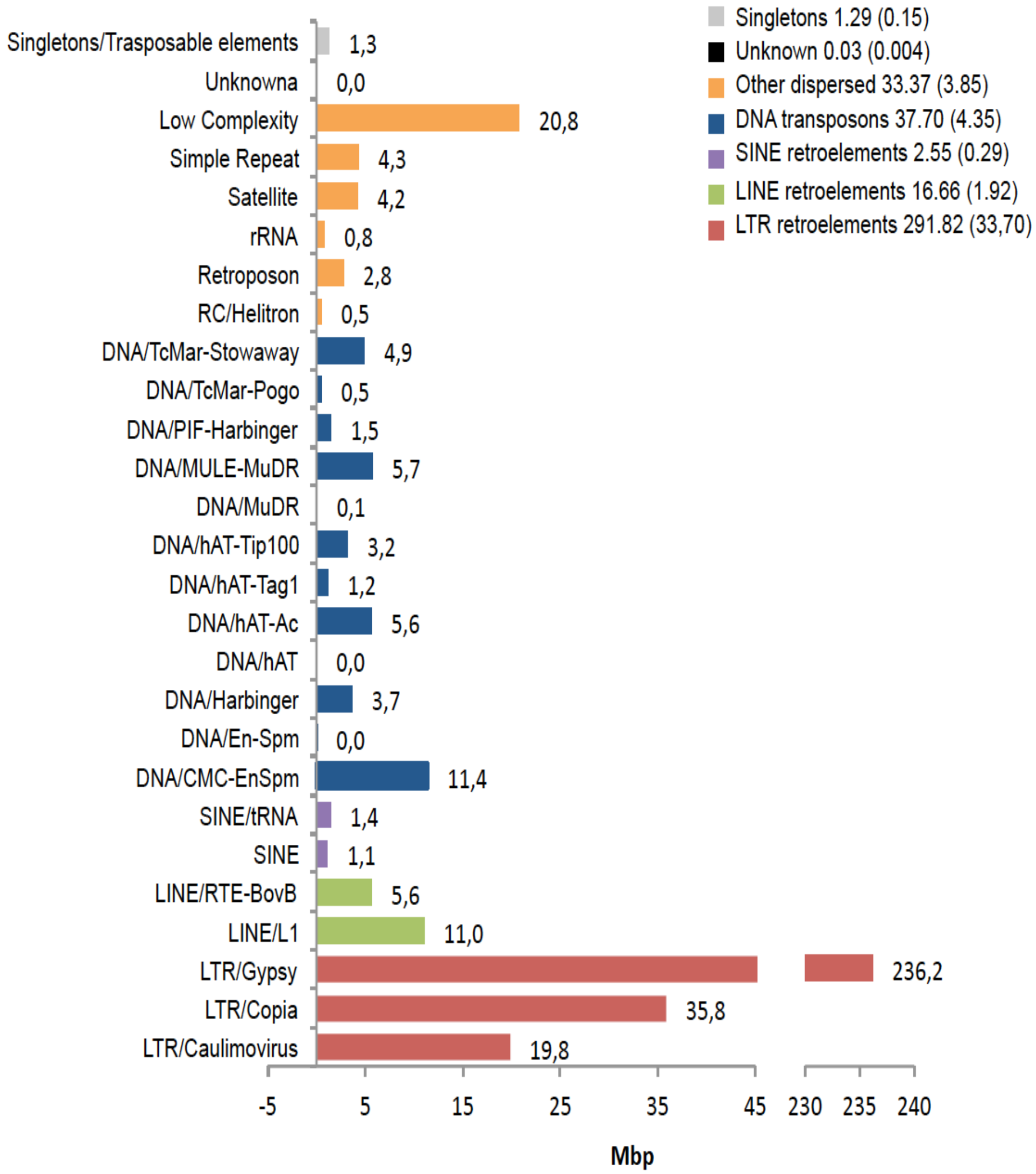


Figure 3.10 Classification of repetitive sequences in *S. commersonii*

Gene annotation

Gene prediction was performed by combining results obtained from *ab initio* prediction, homology searches and experimental support (cmm EST). The *de novo* assembled transcriptome encompassed ~96% of all predicted *S. commersonii* genes (Table 3.6).

Table 3.6. *De novo* assembled transcript

Assembled sequences. number	117,816
Maximum length. bp	53,539
Average length. bp	1,369.13
Minimum length. bp	301
Median	1,026
N50	1,887
no mapping against assembly	113,559
% mapping against assembly	96.39%

Fewer genes (37,662) (AED≤0.5) were predicted in *S. commersonii* than in potato (~39,000), but the wild species had more than tomato (34,727). Of predicted *S. commersonii* genes, 30,477 predicted protein-coding genes had significant Blast similarity to protein-coding genes from other organisms in the non-redundant (nr) NCBI database. Near 20,500 genes were assigned to Gene Ontology (GO) terms and more than 4,900 proteins were annotated with a four-digit EC number. This implied that more than 24% of the predicted proteome of *S. commersonii* has enzymatic function. A large number of transcripts (20,994) with no apparent coding capacity was predicted in *S. commersonii*. These noncoding RNAs (ncRNAs) comprised a diverse group of transcripts, including 22 transfer RNAs (tRNA), 40 ribosomal RNAs (rRNA), 18,882 long non-coding RNAs (lncRNA), and 1,703 putative microRNAs (miRNA) precursors (Table 3.7). Among these latter, 360 were predicted to fold in a secondary structure leading to the typical miRNA/miRNA* double stranded RNA duplexes.(Table 3.9). In addition, 47 out of those 360 transcripts showed similarity to known mature miRNAs. A key study for understanding the biological functions of 1,703 predicted miRNAs was made through the identification of 4,437 target sites. According to GO term classification, 22% (976) of these are involved in cold response categories and 10 are potential regulators of transcripts annotated as responsive to cold (Table 3.8)

Table 3.7. Micro RNA statistics

Predicted miRNA precursors	1703
Prediction of mature miRNAs	1515
Putative target transcripts	4437
Average Nr of targets per miRNA	12
Minimum Nr of targers per miRNA	1
Maximum Nr of targers per miRNA	64
Average Nr of miRNA per target	2.2
Minimum Nr of miRNA per target	1
Maximum Nr of miRNA per target	70
Nr of putative target loci in	
Cold Acclimation-Like	277
Cellular Response to cold-Like	45
Response to Cold-Like	654

Table 3.8 Transcripts annotated as responsive to cold stress and of their potential miRNA regulators

Transcript_target	Annotation	miRNA_precursor	miRBase hit	RFAM hit
augustus_masked_scaffold2559_abinit_gene_0_8	avr9 cf-9 rapidly elicited protein 275	TCONS_00020996	stu-miR6023	
augustus_masked_scaffold27265_abinit_gene_0_2	cf-9 precursor	TCONS_00020996	stu-miR6023	
augustus_masked_scaffold31010_abinit_gene_0_1	rna recognition motif-containing protein	TCONS_00067712	stu-miR7988	
augustus_masked_scaffold370_abinit_gene_0_0	wd40 yvtn repeat and bromo-wdr9-i-like domain-containing protein	TCONS_00060297	stu-miR7988	
augustus_masked_scaffold40820_abinit_gene_0_3	receptor-like protein 12-like	TCONS_00020996	stu-miR6023	
augustus_masked_scaffold4372_abinit_gene_0_0	avr9 cf-9 rapidly elicited protein 275	TCONS_00020996	stu-miR6023	
augustus_masked_scaffold5238_abinit_gene_0_0	cf-9 precursor	TCONS_00020996	stu-miR6023	
augustus_masked_scaffold7053_abinit_gene_0_1	phosphoglycerate mutase	TCONS_00029235	peu-miR2916	
augustus_masked_scaffold712_abinit_gene_0_0	avr9 cf-9 rapidly elicited protein 275	TCONS_00020996	stu-miR6023	
genemark_scaffold21357_abinit_gene_0_4	peru 2	TCONS_00020996	stu-miR6023	
genemark_scaffold363_abinit_gene_0_19	g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like	TCONS_00031602	stu-miR8025-5p	mir-399
genemark_scaffold363_abinit_gene_0_19	g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like	TCONS_00031603	stu-miR8025-3p	
maker_scaffold10612_augustus_gene_0_22	probable lrr receptor-like serine threonine-	TCONS_00020996	stu-miR6023	

	protein kinase at5g10290-like			
maker_scaffold10960_snap_gene_0_61	leucine-rich repeat protein kinase-like protein	TCONS_00060297	stu-miR7988	
maker_scaffold15760_snap_gene_0_35	peru 1	TCONS_00020996	stu-miR6023	
maker_scaffold1691_snap_gene_1_59	protein kinase chloroplast	TCONS_00046799	stu-miR7981-3p	
maker_scaffold1754_snap_gene_0_10	peru 1	TCONS_00020996	stu-miR6023	
maker_scaffold17583_augustus_gene_0_17	g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like	TCONS_00031602	stu-miR8025-5p	mir-399
maker_scaffold17583_augustus_gene_0_17	g-type lectin s-receptor-like serine threonine-protein kinase rlk1-like	TCONS_00031603	stu-miR8025-3p	
maker_scaffold20925_augustus_gene_0_30	arginine serine-rich-splicing factor rsp40-like	TCONS_00046799	stu-miR7981-3p	
maker_scaffold20968_augustus_gene_0_48	vacuolar cation proton exchanger 5-like	TCONS_00005906	stu-miR7997c	
maker_scaffold23900_augustus_gene_0_18	peru 1	TCONS_00020996	stu-miR6023	
maker_scaffold24560_snap_gene_0_68	pentatricopeptide repeat-containing protein	TCONS_00060297	stu-miR7988	
maker_scaffold2531_augustus_gene_0_75	receptor-like protein kinase	TCONS_00020996	stu-miR6023	
maker_scaffold27257_snap_gene_0_14	peru 1	TCONS_00020996	stu-miR6023	
maker_scaffold27265_augustus_gene_0_25	peru 1	TCONS_00020996	stu-miR6023	
maker_scaffold32581_snap_gene_0_11	peru 2	TCONS_00020996	stu-miR6023	
maker_scaffold4372_snap_gene_0_34	lrr receptor-like serine threonine-protein kinase gso2-like	TCONS_00020996	stu-miR6023	
maker_scaffold7225_snap_gene_0_60	cationic peroxidase isozyme 40k precursor	TCONS_00031603	stu-miR8025-3p	
maker_scaffold8156_snap_gene_1_55	receptor-like protein kinase	TCONS_00020996	stu-miR6023	
maker_scaffold8450_snap_gene_0_34	receptor-like protein kinase	TCONS_00020996	stu-miR6023	
maker_scaffold8450_snap_gene_0_34	receptor-like protein kinase	TCONS_00029235	peu-miR2916	
snap_masked_scaffold15959_abinit_gene_0_11	protein	TCONS_00068368	sly-miR1918	
snap_masked_scaffold16580_abinit_gene_0_9	transcriptional adapter ada2-like	TCONS_00049373	stu-miR7998	
snap_masked_scaffold6229_abinit_gene_0_56	catalase	TCONS_00029235	peu-miR2916	
snap_masked_scaffold6491_abinit_gene_0_41	udp-d-glucuronate 4-epimerase 2	TCONS_00064460	mtr-miR5298d	

Table 3.9 Putative miRNA precursor showing miRNA/miRNA* duplexes and similarity to know miRNAs. Similarity with miRNAs was checked by blasting against MiRBase and with RFAM.

Transcript_id	miRBase hit	RFAM hit
TCONS_00001190	mtr-miR319a-5p	
TCONS_00002050	ahy-miR3508	
TCONS_00002051	ahy-miR3508	
TCONS_00005906	stu-miR7997c	
TCONS_00012360	stu-miR7985	
TCONS_00019794	stu-miR7998	
TCONS_00020996	stu-miR6023	
TCONS_00022572	gma-miR1520o	
TCONS_00024816	pab-miR3698	
TCONS_00025232	bdi-miR5164	
TCONS_00029235	peu-miR2916	
TCONS_00031426	osa-miR5837.1	
TCONS_00031603	stu-miR8025-3p	
TCONS_00043860	ppt-miR1033e	
TCONS_00045720	stu-miR7998	
TCONS_00047702	stu-miR7998	
TCONS_00049373	stu-miR7998	
TCONS_00053681	gma-miR4995	
TCONS_00055885	ptc-miR169af	
TCONS_00058398	mtr-miR2670g	
TCONS_00060297	stu-miR7988	
TCONS_00064460	mtr-miR5298d	
TCONS_00067712	stu-miR7988	
TCONS_00068368	sly-miR1918	
TCONS_00075645	osa-miR1863a	
TCONS_00020719	stu-miR7986	
TCONS_00031602	stu-miR8025-5p	mir-399
TCONS_00038446	gma-miR4995	
TCONS_00046799	stu-miR7981-3p	
TCONS_00076957	stu-miR8006-5p	mir-166
TCONS_00058937		mir-598
TCONS_00028908		mir-308
TCONS_00033773		MIR1023
TCONS_00034001		MIR396
TCONS_00036819		mir-785
TCONS_00050504		MIR821
TCONS_00032245		lin-4
TCONS_00017293		mir-198
TCONS_00059748		mir-62
TCONS_00006315		mir-156
TCONS_00063573		MIR477
TCONS_00055774		mir-48
TCONS_00062719		MIR821
TCONS_00005261		MIR1122
TCONS_00059744		MIR807
TCONS_00059483		mir-598
TCONS_00006798		MIR820

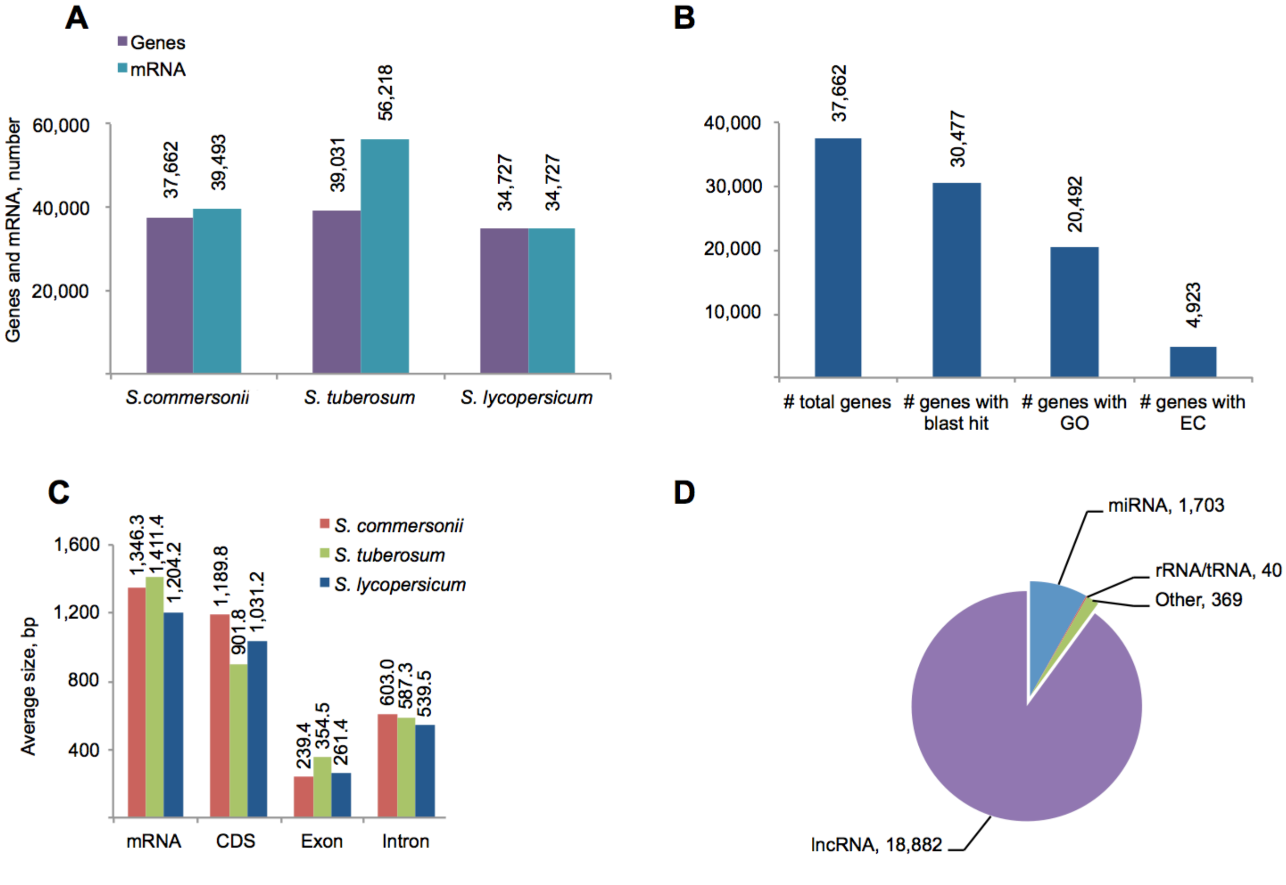


Figure 3.9. Functional annotation of *S. commersonii* transcriptome.

A. Comparison of gene (AED≤0.5) and mRNA numbers in *S. commersonii*, *S. tuberosum* and *S. lycopersium*.
B. Number of predicted protein-encoding genes with significant BLAST similarity, with GO annotation and with a 4-digit EC number.
C. mRNA, CDS, exon and intron average size in *S. commersonii*. The mean number of exons and intron per gene are reported as well.
D. Non-coding RNA gene classes in *S. commersonii*, including transfer RNA (tRNA), ribosomal RNA (rRNA), microRNA (miRNA). Small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) were included in "other" category.

Phylogenetic analysis and genome evolution

To gain insight into the evolution of the *S. commersonii* genome, we compared its virtual proteome with predicted proteins of 11 other fully-sequenced plant genomes (Table 3.10), including *S. tuberosum* and *S. lycopersicum* (Figure 3.10).

Table 3.10. Overview of the species used for the comparative genomics analyses,

Species Name	**Genes**	**Unique, longest transcripts**	**Source**	**As in**
Solanum commersonii	37.662	37.477	Genome Project	04/2014
Solanum tuberosum	39.021	38.781	Ensembl Plants - Release 22	04/2014
Solanum lycopersicum	34.727	34.635	International Tomato Annotation Group	02/2012
Mimulus guttatus	28.140	27.980	Phytozome 10 by JGI	04/2014
Beta vulgaris	27.421	27.363	CRG	11/2012
Cucumis melo	27.427	27.376	melonomics,upv,es	04/2011
Arabidopsis thaliana	27.416	27.233	Ensembl Plants - Release 17	04/2013
Glycine max	54.174	53.821	Ensembl Plants - Release 17	04/2013
Triticum aestivum	98.779	94.236	Ensembl Plants - Release 22	04/2014
Zea mays	39.475	38.773	Ensembl Plants - Release 22	04/2014
Brachypodium distachyon	26.552	26.470	Ensembl Plants - Release 22	04/2014
Oryza sativa subsp, japonica	35.679	35.445	Ensembl Plants - Release 22	04/2014

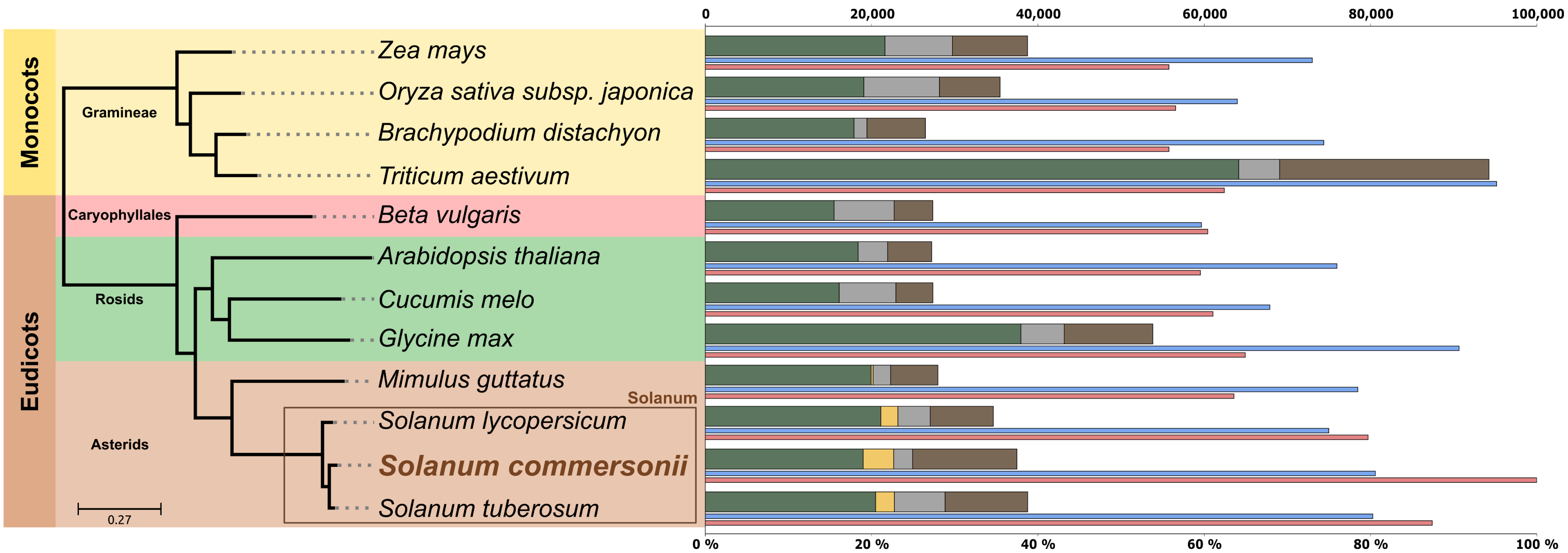


Figura 3.10 Comparative genomics of 12 fully sequenced plant species where phylogeny is based on maximum-likelihood analysis of a concatenated alignment of 454 widespread single-copy proteins. Different background colors indicate taxonomic groupings within the species used to make the tree. Inset square highlights the species belonging to the genus *Solanum*. Bars represent the total number of genes for each species (scale on the top).

1. Bars are divided to indicate different types of homology relationships.

- Dark green: widespread genes that are found in at least 11 of the 12 species.
- Yellow: widespread but asterids-specific genes that are found in at least 3 of the 4 rosids species.
- Gray: Species-specific genes with no (detectable) homologs in other species.
- Brown: genes without a clear homology pattern.

2. The thin purple line under each bar represents the percentage of genes with at least one paralog in a given species.

3. The thin dark blue line represents the percentage of wild potato genes that have homologs in a given species.

The resulting 35,182 phylogenetic trees, available through PhylomeDB (Huerta-Cepas et al., 2014), were scanned to predict phylogeny-based orthology and paralogy relationships (Gabaldón, 2008), to detect and date duplication events (Huerta-Cepas and Gabaldón, 2011), and to transfer annotations to *S. commersonii* genes from their functionally characterized one-to-one orthologs (Huerta-Cepas and Gabaldón, 2011). Roughly 17,300 (44%) and 16,821 (42%) *S. commersonii* genes showed one-to-one orthology with genes from *S. tuberosum* and *S. lycopersicum*, respectively, but only 7,058 (18%) with genes from the more distantly related asterid *Mimulus guttatus* (Table 3.10). Out of 35,182 phylogenies obtained, 9,445 (24%), 7,316 (21%), and 14,061 (40%) showed at least one duplication event at the *S. commersonii*, potato ancestor, and *Solanum* ancestor nodes, respectively, compared to only 1,814 trees (5%) showing a duplication at the base of Asterids (Table 3.11). The overall average number of duplications per branch (duplication density) was 0.66, 0.93, and 0.94 for *S. commersonii,* potato-ancestor and *Solanum* ancestor, respectively, whereas we found a low rate of 0.066 for the common ancestor of Asterids was found (Table 3.12).

Table 3.11. Detected one-to-one orthologs between a given species and *S. commersonii*

Species Name	**one-to-one orthologs**
S. tuberosum	17.297
S. lycopersicum	16.821
M. guttatus	7.058
B. vulgaris	6.799
C. melo	6.684
A. thaliana	5,.862
G. max	1.667
T. aestivum	1.160
Z. mays	3.913
B. distachyon	4.968
O. sativa subsp. *japonica*	4,492

Table 3.12. Statistics about the number of duplication events detected in single gene trees according to their relative ages

Age	**Events**	**Trees with events (all trees: 35,182)**	**Ratio (events / all trees)**
1:*S.commersonii* specific	23,133	9,445	0.6575
2: Potato Ancestor	32,680	7,316	0.9289
3: Solanum Ancestor	33,185	14,61	0.9432
4: Basal to Asterids	2,331	1,814	0.0663

To gain further insight into the divergence between *S. commersonii* and the domesticated potato, we measured transversions at four-fold degenerate sites (4DTv) for orthologous gene pairs between *S. commersonii* and either the domesticated potato or tomato (Figure 4), and between paralogous gene pairs diverged from duplications at each of the three relevant duplication periods investigated. Substitution rates at orthologous sites between *S. commersonii* and tomato peaked at 0.225, whereas those between *S. commersonii* and domesticated potato at 0.077. Assuming a divergence time between tomato and potato of 7.3 Mya (Potato Genome Sequencing Consortium 2011), and a constant mutation rate between the three lineages, this renders an estimate of ~2.3 Mya for the separation of domesticated potato and *S. commersonii* lineages. The analysis of the paralogous pairs revealed at all three relative ages showed a similar pattern, with the two most prominent peaks largely preceding the divergence of *S. commersonii* and tomato. Paralogous genes mapped to the *S. commersonii*-specific duplication (age 1) did show an additional, younger peak at 4DTv values. We assessed the genomic organization of these recent duplicates and found that most of them were present in tandem (314) or at least closely associated in the same contig (141). Finally, we assessed functional enrichment among gene families duplicated at each of these three periods (Supplemental dataset 2 online: https://drive.google.com/open?id=0BzNl4eO_bJumTnRZZ1hIbEUtUm8&authuser=0). Response to salt stress and water transport were terms found to be enriched exclusively among *S. commersonii* specific duplications. The ancestral potato duplication was enriched in terms related to cadmium, metal ion binding, or synthesis of terpenes, whereas terms related to nitrogen starvation, response to ethylene, response to gamma radiation, and maltose metabolism were enriched in the duplications preceding the common *Solanum* ancestor. All three duplication periods shared enrichments in defence response and growth. Transposon related terms were enriched in the largest expanded families in *S. commersonii*, indicating active expansion of transposons.

Pathogen-receptor genes annotation

A catalogue of 942 and 1,406 non-redundant pathogen recognition proteins was created from the *S. commersonii* and *S. tuberosum* proteomes, respectively. We classified the corresponding genes into various structural categories based on the arrangement of encoded domains (Table 3.13). In *S. commersonii*, 286 coiled-coil (CC)-nucleotide-binding site (NBS)-leucine-rich repeat (LRR) (CNL), 71 NBS, 143 Toll/interleukin-1 receptor (TIR)-NBS, and 37 TIR genes were found. More than 250 receptor like kinases (RLK) and 280 receptor like proteins (RLP) were also recorded. In comparison, in *S. tuberosum,* 506 CNL, 199 NBS, 199 TIR-NBS, 36 TIR, 313 RLK, and 237 RLP genes were identified. The *S. tuberosum* genome also encodes 14 TIR-LRR genes. Previously, using similar approaches, the pathogen recognition proteins from tomato were catalogued (Andolfo et al., 2014a; 2014b). While the *S. tuberosum* genome encoded nearly twice as many CNL genes as *S. commersonii* (506 vs. 286, respectively), the tomato roughly encoded half of *S. commersonii* CNL genes (81 vs. 186, respectively). In contrast, *S. commeronii* and *S. tuberosum* possess a larger complement of TIR-NBS-LRR (TNL) and RLP proteins that does tomato. These findings suggest that the pathogen receptor gene repertoire in each *Solanum* species is uniquely shaped based on pathogen pressures and life history. Syntenic relationships between pathogen receptor genes in *S. commersonii* and *S. tuberosum* were further explored by comparative analysis of three loci involved in *Phytophthora infestans* resistance, *Rpi-blb2* (van der Vossen et al., 2005), *Tm-2* (Lanfermeijer et al., 2003) and *R1* (Ballvora et al., 2002). All are members of the CNL superfamily and all exist within clusters of related gene copies. In *S. tuberosum, Rpi-blb2* is part of a 15-gene cluster (van der Vossen et al., 2005), but in *S. commersonii,* only four corresponding gene copies were present. Similarly, in *S. tuberosum,* the *Tm-2* cluster comprises four gene copies (Lanfermeijer et al., 2003) while in *S. commersonii,* only two genes were annotated. For *R1,* there was a clear variation between the species in terms of the physical cluster size and number of genes included. The *S. tuberosum R1* cluster was longer than that of *S. commersonii* (300 kb versus 37 kb, respectively) and comprised more gene copies. The *S. commersonii* genome encodes three *R1* genes with clear orthologous relationship to genes in *S. tuberosum.* Specifically, *S. commersonii ScR1B-23-like*, *ScR1C-3-like* and *ScR1A-4-like*, correspond to *S. tuberosum StR1B-23, StR1C-3* and *StR1A-4*. These three *ScR1* orthologous pairs exhibited an average nucleotide identity of 93%, but were arranged in reverse order in the two species. Furthermore, additional, unrelated genes found in the *StR1B-23* to *StR1C-3* interval and in the *StR1C-3* and *StR1A-4* interval were completely absent in the *S. commersonii R1* gene cluster. Comparison of *S. commersonii* and *S. tuberosum* orthologous *R1* gene copies revealed further substantial structural variation in some cases. In particular, while the

CDS length of *R1B-23* and *R1A-4* were similar for *S. tuberosum* and *S. commersonii* genes, *StR1C-3* was about three times shorter (4,207 bp) than *ScR1C-3* (12,386 bp). Differences in the number of exons and introns between *R1* orthologs were also found (Figure 3.11).

Table 3.13. Numbers of *S. commersonii* and *S. tuberosum* genes encoding proteins with domains similar to those found in plant pathogen receptor proteins

Family ID[1]	Domain	*S. tuberosum,* number	*S. commersonii,* number
Canonical cytoplasmic R-genes			
CNL or NL	CC-NBS-LRR	194	186
TNL	TIR-NBS-LRR	46	36
Single domains or incomplete structures			
NL	NBS-LRR	165	98
N	NBS	199	71
T	TIR	36	10
L	LRR	199	144
TN	TIR-NBS	14	12
TL	TIR-LRR	2	1
Canonical transmembrane domains			
RLK	Receptor like kinase	313	252
RLP	Receptor like protein	237	180
New combinations of domains or new structures			
MN	Metallophos-NBS	1	0
GN	Glutaredoix-NBS	1	0
RPW8N	RPW8-NBS	1	0
RPW8NL	RPW8-NBS-LRR	1	0
TPP2	TIR-PP2	4	1
ANL	Aldolase-NBS-LRR	1	0
RLP-M	RLP-Malectin	4	0
RLP-U	RLP-Ubiquitin	4	0
RLK-UPP	RLK-UPP	1	1
CelNL	Cellulose_syntase-NBS-LRR	0	1
PN	Peroxidase-NBS	0	2
PhN	Phage_GPO-NBS	0	1
HN	Homoserin-NBS	0	4
AL	aldolase-LRR	0	1
HL	hidrolase-LRR	0	1
LL	Lipase-LRR	0	1
PL	Peptidase-LRR	0	1
YL	YDG-LRR	0	1
ML	Malectin-LRR	0	6
RLK-L	RLK-Lipase	0	1
RLK-M	RLK-Malectin	0	6
RLK-P	RLK-PPR	0	1
SN	SBF-NBS	0	1
PT	PthA_Avr-TIR	0	1

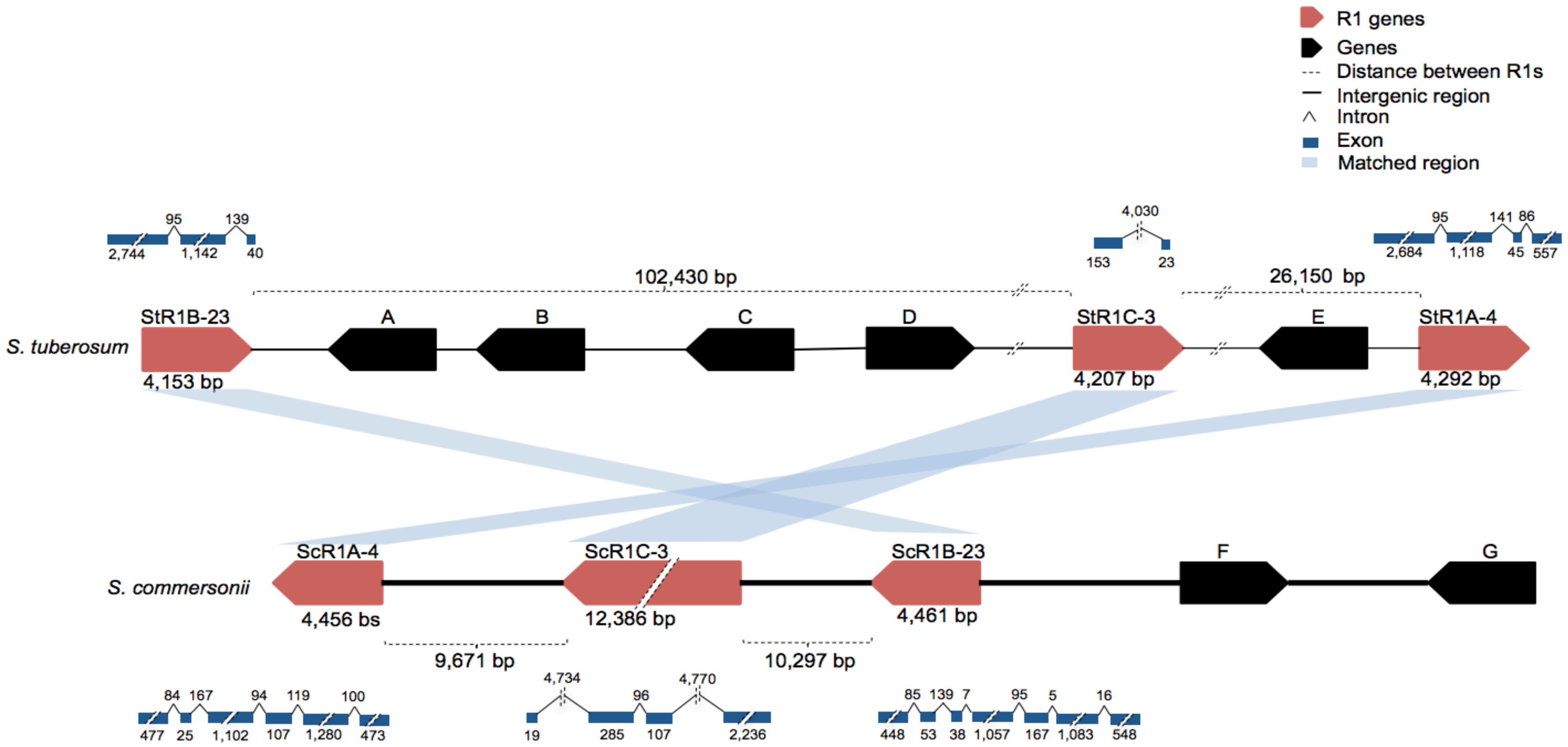
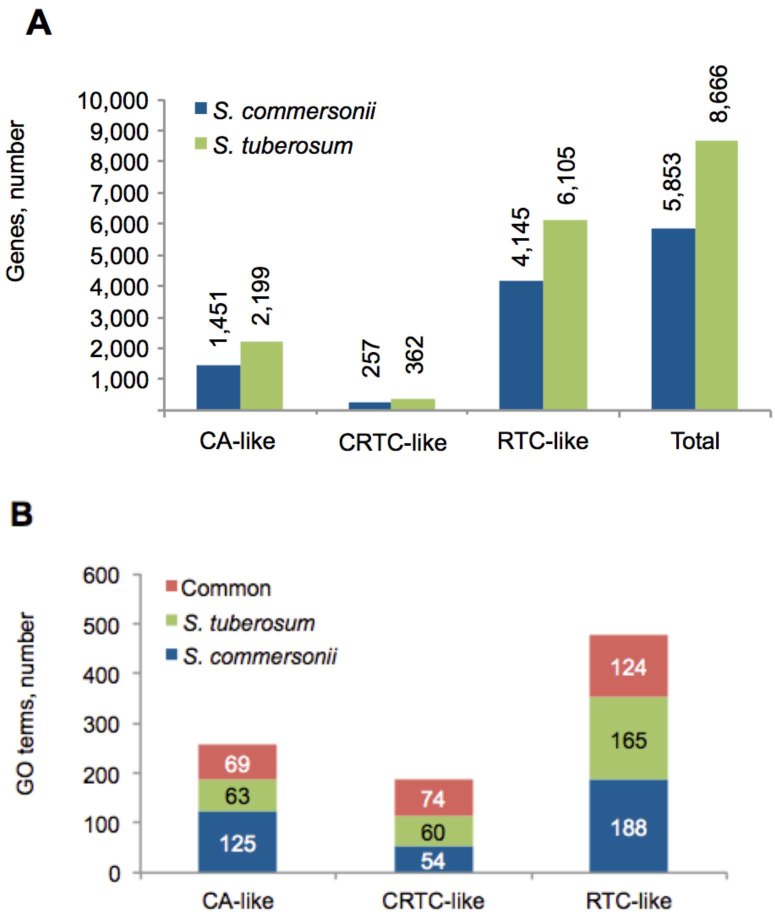


Figure 3.11 *R1* cluster in *S. commersonii* and *S. tuberosum.*
R1-gene homologues and genes are indicated in red and black filled oriented boxes, respectively. Numbers below the *R1* homologue boxes indicate their length (bp). For each R1 homologue intron-exon structure is shown. Intergenic regions are drawn as thicker solid lines, whereas thick dashed lines indicate distance between *R1*-gene homologues. Blue-shaded areas between *S. tuberosum* and *S. commersonii* genotypes designate homology among *R1* sequences. Figure not drawn to scale.

Cold-responsive genes analysis

A total of 5,853 and 8,666 predicted protein sequences similar to *Arabidopsis* proteins annotated as responsive to cold was identified in *S. commersonii* and *S. tuberosum*, respectively. In *S. commersonii*, 1,451 proteins were homologous to *Arabidopsis* sequences annotated with the GO term *cold acclimation* (hereinafter CA-like), 257 with the term *cellular response to cold* (hereinafter CRC-like) and 4,145 with the term *response to cold* (hereinafter RC-like). In *S. tuberosum,* 2,199 were in the CA-like group*,* 362 in the CRC-like group and 6,105 in the RC-like group. Enriched GO term categories were found in both species (Figure 3.12 A; Supplemental datasets 3 online: *https://drive.google.com/open?id=0BzNl4eO_bJumblVtRW1fQ3NLTTA&authuser=0* and 4 online: *https://drive.google.com/open?id=0BzNl4eO_bJumZm1sQ0FMc2tzNUE&authuser=0*). Roughly 2,860 genes were enriched in *S. commersonii* (707, 85 and 2,072 belonging to the CA-like, CRC-like and RC-like groups, respectively). By contrast, only 532, 181 and 1,539 genes were enriched in *S. tuberosum* (Figures 3.12 B-C). GO annotation also revealed that 34 CA-like, CRC-like, and RC-like categories encompassed a large number (1,546) of cold responsive genes harboring a SNP (12.4% of total) (Figure 3.12 D; Supplemental dataset 5 online: *https://drive.google.com/open?id=0BzNl4eO_bJumZm1sQ0FMc2tzNUE&authuser=0).*



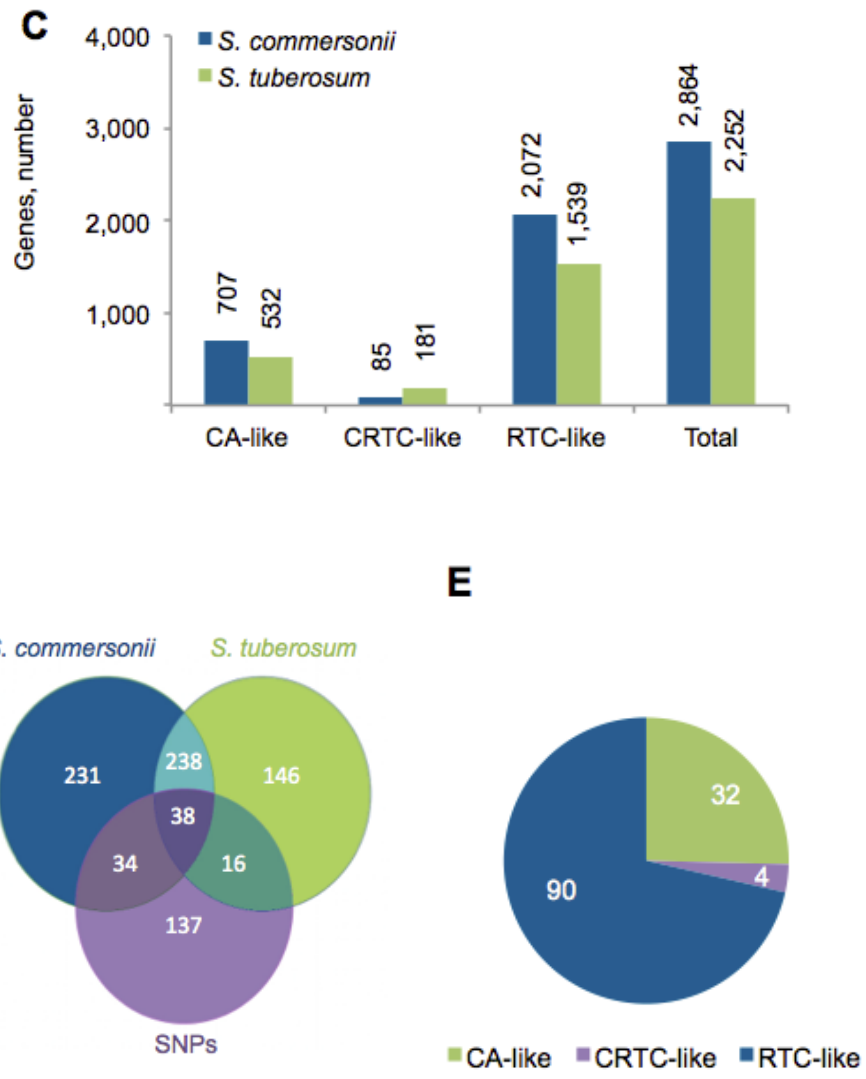


Figure 3.12. Cold responsive genes annotation analysis.

To annotate putative cold resistance genes, a set of reference proteins was selected from *Arabidopsis thaliana*. CA: Cold Acclimation; CRTC: Cellular Response To Cold; RTC: Response To Cold.

A. Number of genes having putative binding sites for transcription factors related to responsive to cold.
B. Results of enrichment GO analysis.
C. Number of genes with unique GO term in *S. commersonii* and *S. tuberosum.*
D. Cold-responsive GO Terms significantly enriched (FDR < 0.05) in genes containing SNPs both in *S. commersonii* and *S. tuberosum.*
E. Number of unique genes involved in tolerance to cold in *S. commersonii*.

In addition, of 126 unique annotated *S. commersonii* genes involved in response to cold, 32 belonged to CA-like, four to CRTC-like, and 90 to RTC-like GO terms (Figure 3.12 E). To identify genes involved in freezing and cold-acclimation responses, the transcript expression profile of AC and NAC plants were both compared to that of non frost-stressed, grown at 24°C. We identified 855 differentially expressed genes: 720 in AC conditions and 784 in NAC conditions. Venn diagram analysis indicated that 71 genes were differently expressed (mostly upregulated) exclusively under AC condition, whereas 135 only in NAC conditions. Roughly 649 genes were found responsive to both conditions (Supplemental dataset 6 online: *https://drive.google.com/open?id=0BzNl4eO_bJumWnhkUnc5WFYxYTg&authuser=0*). Different functional categories appeared to be enriched in either AC or NAC conditions (Table 3.14). Among the NAC differentially expressed genes, the most significantly enriched groups were those involved in cytoplasmic part (GO:0044444), organelle metabolisms (GO:0044422 and GO:0043226) and in phytosteroid and brassinosteroid metabolic processes (GO:0016128 and GO:0016131, respectively). As far as AC differently expressed genes are concerned, the most represented GO terms were in response to metal and cadmium ions (GO:0010038 and GO:0046686) symplast (GO:0055044) and vacuolar part (GO:0044437). Protein kinases and phosphatases with altered expression under NAC and AC conditions were among the most differentially expressed groups (Supplemental dataset 6 online: *https://drive.google.com/open?id=0BzNl4eO_bJumWnhkUnc5WFYxYTg&authuser=0*). In particular, 76 receptor-like kinase (RLK) were either up or downregulated under both conditions, whereas six RLKs exhibited contrasting kinetics: two RLKs (At1g11050-like; At4g35230-like) were upregulated under NAC, but repressed under AC and four (a putative *hormone-sensitive lipase* 1 - HSL1-like, At5g15080-like, At5g35380-like and a putative *recombinant brassinosteroid insensitive* 1, BSK1) were upregulated under AC and downregulated under NAC. In addition, proteins involved in the cold response machinery, such as antioxidant cascades, secondary metabolism, cell wall polysaccharide remodeling, starch metabolism, and protein folding (heat shock protein 70, HSP70) were found. Out of 855 cold-DEG, 56 (6.5%) were annotated as transcription factors (TFs) with known DNA binding domains (Figure 3.13).

Table 3.14. Functional enrichment analysis results after removing redundancy for the 10 biggest clusters of specifically expanded clusters of proteins in *S. commersonii* with statistically significant enriched functional terms

Cluster	Size	Ontology	Go Term	Go Term Name
cluster 4369	191	Biological Process	GO:0006278	RNA-dependent DNA replication
cluster 4369	191	Molecular Function	GO:0003676	nucleic acid binding
cluster 4369	191	Molecular Function	GO:0003723	RNA binding
cluster 4369	191	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
cluster 4369	191	Molecular Function	GO:0004523	ribonuclease H activity
cluster 4368	158	Biological Process	GO:0006278	RNA-dependent DNA replication
cluster 4368	158	Molecular Function	GO:0003723	RNA binding
cluster 4368	158	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
cluster 4368	158	Molecular Function	GO:0016787	hydrolase activity
cluster 4364	138	Molecular Function	GO:0003676	nucleic acid binding
cluster 4364	138	Molecular Function	GO:0004523	ribonuclease H activity
cluster 4363	121	Biological Process	GO:0006278	RNA-dependent DNA replication
cluster 4363	121	Molecular Function	GO:0003676	nucleic acid binding
cluster 4363	121	Molecular Function	GO:0003723	RNA binding
cluster 4363	121	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
cluster 4363	121	Molecular Function	GO:0008270	zinc ion binding
cluster 4362	119	Molecular Function	GO:0004386	helicase activity
cluster 4362	119	Molecular Function	GO:0005524	ATP binding
cluster 4360	98	Biological Process	GO:0006278	RNA-dependent DNA replication
cluster 4360	98	Molecular Function	GO:0003676	nucleic acid binding
cluster 4360	98	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
cluster 4359	67	Molecular Function	GO:0008270	zinc ion binding
cluster 4355	60	Molecular Function	GO:0003676	nucleic acid binding
cluster 4354	57	Biological Process	GO:0006278	RNA-dependent DNA replication
cluster 4354	57	Molecular Function	GO:0003723	RNA binding
cluster 4354	57	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
cluster 4354	57	Molecular Function	GO:0004523	ribonuclease H activity
cluster 4350	52	Biological Process	GO:0051252	regulation of RNA metabolic process
cluster 4350	52	Molecular Function	GO:0003676	nucleic acid binding
cluster 4350	52	Molecular Function	GO:0004523	ribonuclease H activity

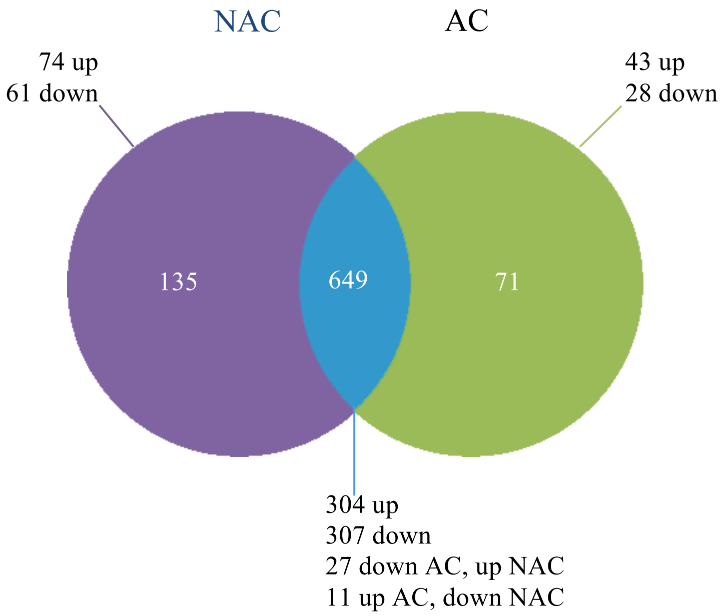
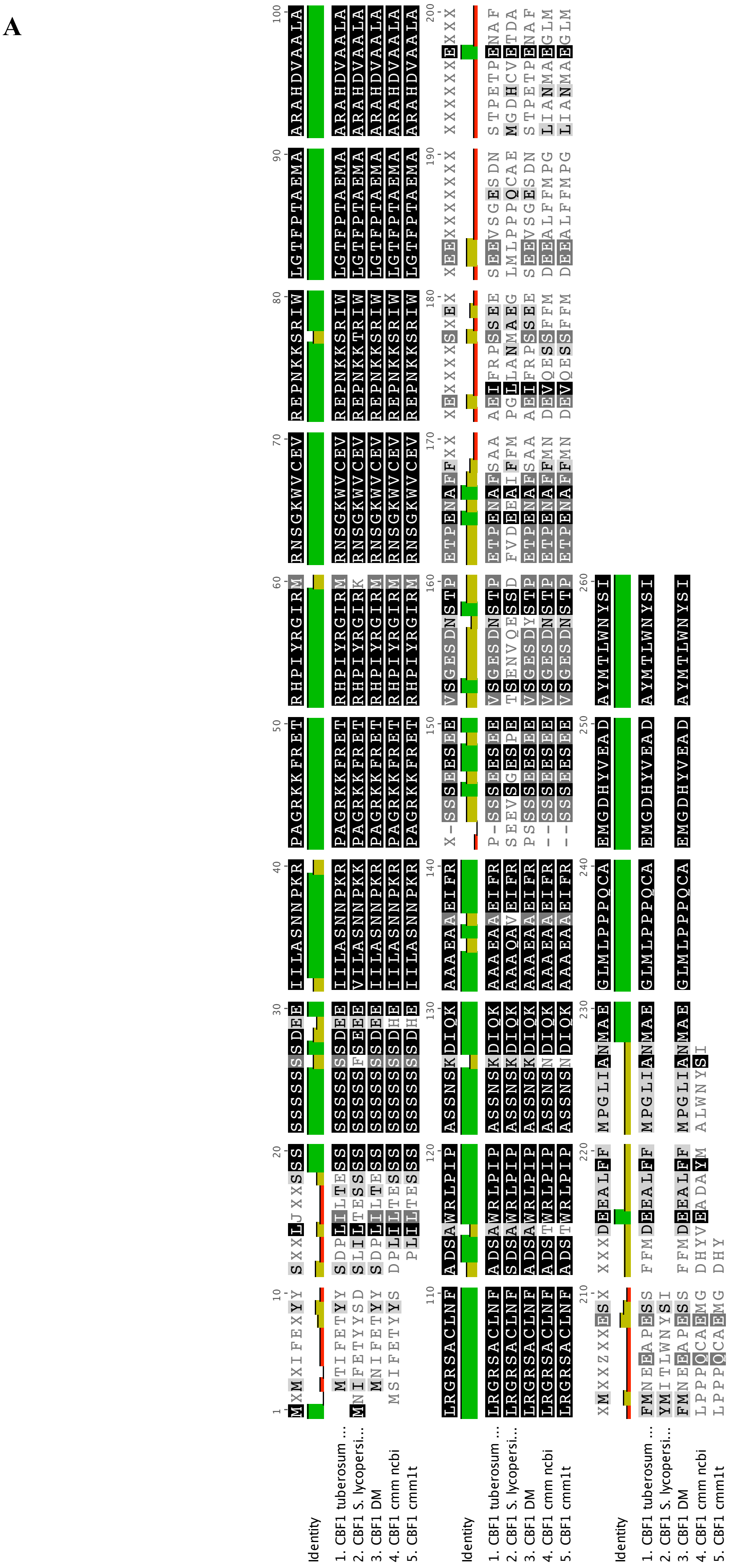
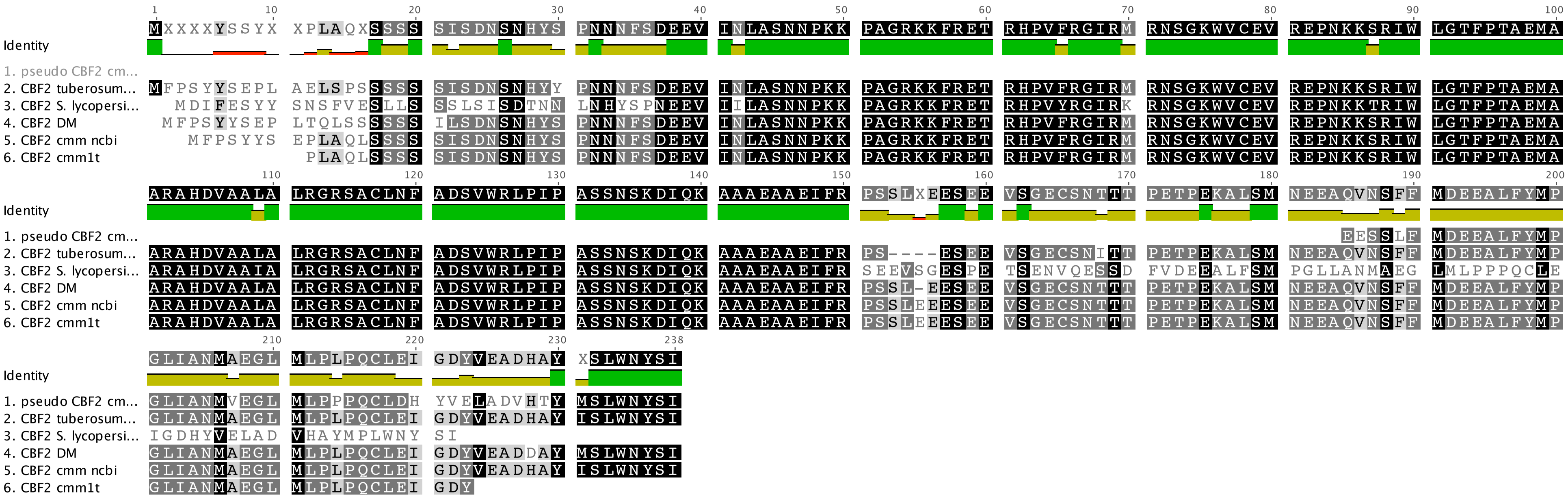


Figure 3.13. Common and differentially expressed genes between AC and NAC conditions.

Thirty-eight TFs were differentially expressed under both AC and NAC conditions, 15 only under NAC and three exclusively under AC. The apetala 2/ethylene-response element binding factor (AP2/ERF) domain was the most represented TF family, accounting for 17 differently expressed TFs. A set of 19 genes encoding cold-sensing and signalling proteins was specifically analyzed under both NAC and AC conditions (Figure 3.15). The cold acclimation pathway is initiated when plants sense low temperatures through membrane rigidification that is followed by a surge of Ca^{2+} into the cytosol. Plants possess groups of Ca^{2+} sensors, including CDPKs (Ca^{2+}-dependent protein kinases), CBL (calcineurin B-like protein), and CIPKs (CBL interacting protein kinase). In *S. commersonii* the expression of CDPK7, -17, and -19, as well as CIPK1, -3, and -23 were profiled both under AC and NAC conditions. They were all activated under NAC and AC conditions, except for CIPK1, which was suppressed under NAC, and CDPK17 and -19, whose expression of which was not impacted by acclimation. Inside the nucleus, the low-temperature signal transduction pathway triggers the expression of C-repeat binding factor (CBF) genes, and their upstream regulators, namely ICE1 (inducer of CBF expression), a positive regulator of CBF3, HOS1 (high expression of osmotically sensitive), a negative regulator of ICE1, and SIZ1, a SUMO E3 ligase, which mediates sumoylation (SUMO conjugation) of ICE1. In *S. commersonii*, we found *ICE1* transcription suppressed under both conditions tested, whereas *HOS1* and SIZ1 expression was consistently downregulated under both NAC and AC. In light of their prominent role in plant cold acclimation, we examined the *S. commersonii* CBF (ScCBF) structural organization and surveyed their expression patterns under AC and NAC conditions. In total, we found four ScCBFs (ScCBF1, -2, -3, and -4) and pseudogenes of CBF2 and CBF3. ScCBFs were collinear with *S. tuberosum* CBFs (StCBF), although StCBF5 was missing in *S. commersonii*. Structural variant analysis revealed the presence of few insertions and/or deletions (indels) within the coding sequences analyzed. By contrast, a substantially low conservation level of the CBF2 and CBF4 upstream regions was observed (Figure 3.17). The ScCBF2 pseudogene possessed only portions of a coding sequence with numerous nonsense codons in all reading frames (Figure 3.14 A) (Pennycooke et al., 2008). The duplicated ScCBF3 possessed the amino acid block ASP-ALA-SER-TRY-ARG (hereafter DASWR) positioned immediately downstream from the 60-amino-acid AP2/ERF DNA-binding domain in the CBF protein. However, it lacked the PKKPAGR sequence positioned upstream of the domain (Figure 3.14 B). Phylogenetic analysis showed that ScCBF2 and ScCBF3 grouped with StCBF sequences (Figure 3.16), supporting robust orthology relationships. By contrast, ScCBF1 was independent of the *Solanum* CBF1 clade, probably due to poorly conserved sequences flanking the AP2/ERF domain (Figure 3.14 C). We wondered whether the ScCBF transcripts accumulated differently under NAC and AC conditions. Interestingly, we found that all ScCBFs were highly responsive, regardless of whether the plant experienced acclimation, with ScCBF1 and ScCBF3 being the most active (Figure 3.15). CBFs can activate expression of a battery of downstream target genes, also called CBF regulon genes, by binding CRT/DRE elements in target promoter regions. Among these, cold responsive (COR) genes act in concert to enhance freezing tolerance. In *S. commersonii*, COR genes responded differently to the cold stimulus depending on whether the plants were first acclimated. In particular, COR47 and COR78 were upregulated under both AC and NAC conditions. In contrast, COR15a and COR413 were upregulated under AC but downregulated under NAC. In *Arabidopsis*, different components of the Histone acetyltransferases (HATs) complex were described to interact with CBF1 *in vitro* and are needed for CBF1 function (Stockinger et al., 2001). Therefore we checked the expression of *ScADA2b* and *ScGCN5*, two components of HATs complex. Our data showed that both *ScAda2b* and *ScGCN5* were transcribed whether or not acclimation occurred.


A
Identity
1. CBF1 tuberosum ...
2. CBF1 S. lycopersi...
3. CBF1 DM
4. CBF1 cmm ncbi
5. CBF1 cmm1t
1
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260

B



1 10 20 30 40 50 60 70 80 90 100
MXXXXYSSYX XPLAQXSSSS SISDNSNHYS PNNNFSDEEV INLASNNPKK PAGRKKFRET RHPVFRGIRM RNSGKWVCEV REPNKKSRIW LGTFPTAEMA
Identity
1. pseudo CBF2 cm...
2. CBF2 tuberosum... MFPSYYSEPL AELSPSSSSS SISDNSNHYY PNNNFSDEEV INLASNNPKK PAGRKKFRET RHPVFRGIRM RNSGKWVCEV REPNKKSRIW LGTFPTAEMA
3. CBF2 S. lycopersi... MDIFESYY SNSFVESLLS SSLSISDTNN LNHYSPNEEV IILASNNPKK PAGRKKFRET RHPVYRGIRK RNSGKWVCEV REPNKKTRIW LGTFPTAEMA
4. CBF2 DM MFPSYYSEP LTQLSSSSSS ILSDNSNHYS PNNNFSDEEV INLASNNPKK PAGRKKFRET RHPVFRGIRM RNSGKWVCEV REPNKKSRIW LGTFPTAEMA
5. CBF2 cmm ncbi MFPSYYS EPLAQLSSSS SISDNSNHYS PNNNFSDEEV INLASNNPKK PAGRKKFRET RHPVFRGIRM RNSGKWVCEV REPNKKSRIW LGTFPTAEMA
6. CBF2 cmm1t PLAQLSSSS SISDNSNHYS PNNNFSDEEV INLASNNPKK PAGRKKFRET RHPVFRGIRM RNSGKWVCEV REPNKKSRIW LGTFPTAEMA
110 120 130 140 150 160 170 180 190 200
ARAHDVAALA LRGRSACLNF ADSVWRLPIP ASSNSKDIQK AAAEAAEIFR PSSLXEESEE VSGECSNTTT PETPEKALSM NEEAQVNSFF MDEEALFYMP
Identity
1. pseudo CBF2 cm... EESSLF MDEEALFYMP
2. CBF2 tuberosum... ARAHDVAALA LRGRSACLNF ADSVWRLPIP ASSNSKDIQK AAAEAAEIFR PS----ESEE VSGECSNITT PETPEKALSM NEEAQVNSFF MDEEALFYMP
3. CBF2 S. lycopersi... ARAHDVAAIA LRGRSACLNF ADSVWRLPIP ASSNSKDIQK AAAEAAEIFR SEEVSGESPE TSENVQESSD FVDEEALFSM PGLLANMAEG LMLPPPQCLE
4. CBF2 DM ARAHDVAALA LRGRSACLNF ADSVWRLPIP ASSNSKDIQK AAAEAAEIFR PSSL-EESEE VSGECSNTTT PETPEKALSM NEEAQVNSFF MDEEALFYMP
5. CBF2 cmm ncbi ARAHDVAALA LRGRSACLNF ADSVWRLPIP ASSNSKDIQK AAAEAAEIFR PSSLEEESEE VSGECSNTTT PETPEKALSM NEEAQVNSFF MDEEALFYMP
6. CBF2 cmm1t ARAHDVAALA LRGRSACLNF ADSVWRLPIP ASSNSKDIQK AAAEAAEIFR PSSLEEESEE VSGECSNTTT PETPEKALSM NEEAQVNSFF MDEEALFYMP
210 220 230 238
GLIANMAEGL MLPLPQCLEI GDYVEADHAY XSLWNYSI
Identity
1. pseudo CBF2 cm... GLIANMVEGL MLPPPQCLDH YVELADVHTY MSLWNYSI
2. CBF2 tuberosum... GLIANMAEGL MLPLPQCLEI GDYVEADHAY ISLWNYSI
3. CBF2 S. lycopersi... IGDHYVELAD VHAYMPLWNY SI
4. CBF2 DM GLIANMAEGL MLPLPQCLEI GDYVEADDAY MSLWNYSI
5. CBF2 cmm ncbi GLIANMAEGL MLPLPQCLEI GDYVEADHAY ISLWNYSI
6. CBF2 cmm1t GLIANMAEGL MLPLPQCLEI GDY

C

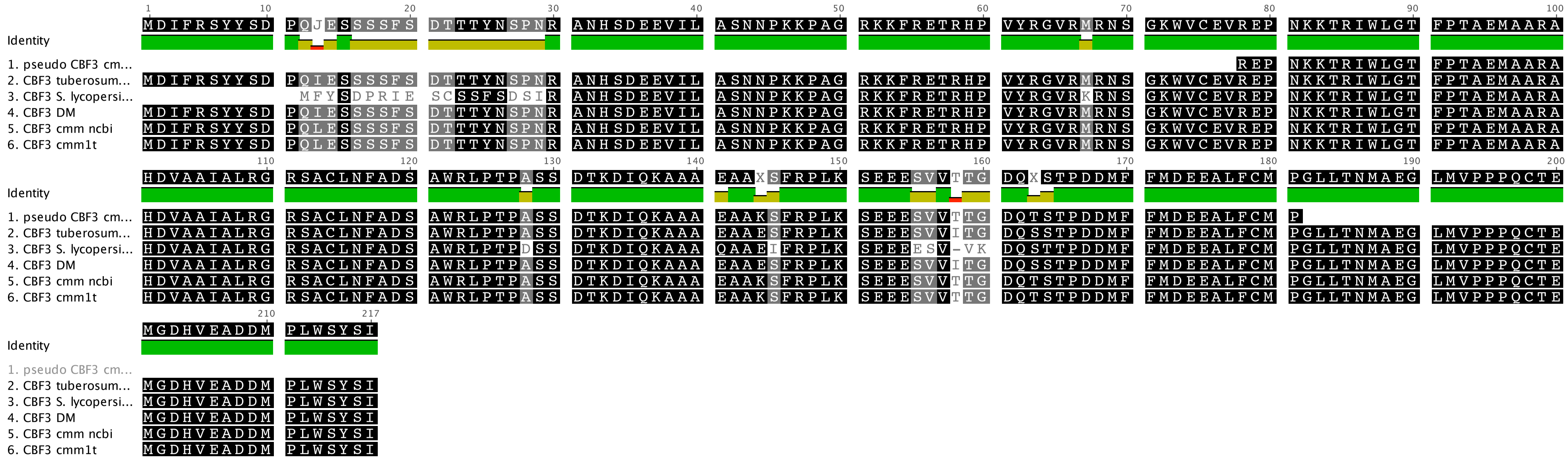

1
10
20
30
40
50
60
70
80
90
100
MDIFRSYYSD PQJESSSSFS DTTTYNSPNR ANHSDEEVIL ASNNPKKPAG RKKFRETRHP VYRGVRMRNS GKWVCEVREP NKKTRIWLGT FPTAEMAARA
Identity
1. pseudo CBF3 cm... REP NKKTRIWLGT FPTAEMAARA
2. CBF3 tuberosum... MDIFRSYYSD PQIESSSSFS DTTTYNSPNR ANHSDEEVIL ASNNPKKPAG RKKFRETRHP VYRGVRMRNS GKWVCEVREP NKKTRIWLGT FPTAEMAARA
3. CBF3 S. lycopersi... MFYSDPRIE SCSSFSDSIR ANHSDEEVIL ASNNPKKPAG RKKFRETRHP VYRGVRKRNS GKWVCEVREP NKKTRIWLGT FPTAEMAARA
4. CBF3 DM MDIFRSYYSD PQIESSSSFS DTTTYNSPNR ANHSDEEVIL ASNNPKKPAG RKKFRETRHP VYRGVRMRNS GKWVCEVREP NKKTRIWLGT FPTAEMAARA
5. CBF3 cmm ncbi MDIFRSYYSD PQLESSSSFS DTTTYNSPNR ANHSDEEVIL ASNNPKKPAG RKKFRETRHP VYRGVRMRNS GKWVCEVREP NKKTRIWLGT FPTAEMAARA
6. CBF3 cmm1t MDIFRSYYSD PQLESSSSFS DTTTYNSPNR ANHSDEEVIL ASNNPKKPAG RKKFRETRHP VYRGVRMRNS GKWVCEVREP NKKTRIWLGT FPTAEMAARA
110
120
130
140
150
160
170
180
190
200
HDVAAIALRG RSACLNFADS AWRLPTPASS DTKDIQKAAA EAAXSFRPLK SEEESVVTTG DQXSTPDDMF FMDEEALFCM PGLLTNMAEG LMVPPPQCTE
Identity
1. pseudo CBF3 cm... HDVAAIALRG RSACLNFADS AWRLPTPASS DTKDIQKAAA EAAKSFRPLK SEEESVVTTG DQTSTPDDMF FMDEEALFCM P
2. CBF3 tuberosum... HDVAAIALRG RSACLNFADS AWRLPTPASS DTKDIQKAAA EAAESFRPLK SEEESVVITG DQSSTPDDMF FMDEEALFCM PGLLTNMAEG LMVPPPQCTE
3. CBF3 S. lycopersi... HDVAAIALRG RSACLNFADS AWRLPTPDSS DTKDIQKAAA QAAEIFRPLK SEEEESV-VK DQSTTPDDMF FMDEEALFCM PGLLTNMAEG LMVPPPQCTE
4. CBF3 DM HDVAAIALRG RSACLNFADS AWRLPTPASS DTKDIQKAAA EAAESFRPLK SEEESVVITG DQSSTPDDMF FMDEEALFCM PGLLTNMAEG LMVPPPQCTE
5. CBF3 cmm ncbi HDVAAIALRG RSACLNFADS AWRLPTPASS DTKDIQKAAA EAAKSFRPLK SEEESVVTTG DQTSTPDDMF FMDEEALFCM PGLLTNMAEG LMVPPPQCTE
6. CBF3 cmm1t HDVAAIALRG RSACLNFADS AWRLPTPASS DTKDIQKAAA EAAKSFRPLK SEEESVVTTG DQTSTPDDMF FMDEEALFCM PGLLTNMAEG LMVPPPQCTE
210
217
MGDHVEADDM PLWSYSI
Identity
1. pseudo CBF3 cm...
2. CBF3 tuberosum... MGDHVEADDM PLWSYSI
3. CBF3 S. lycopersi... MGDHVEADDM PLWSYSI
4. CBF3 DM MGDHVEADDM PLWSYSI
5. CBF3 cmm ncbi MGDHVEADDM PLWSYSI
6. CBF3 cmm1t MGDHVEADDM PLWSYSI

Figure 3.14. Comparison between CBF2 (A), CBF3 (B) and CBF1 (C) protein sequences of *S. commersonii* (clone cmm1t) and the orthologous sequences of *S. commersonii* (NCBI, Pennycook et al. 2009), *S. tuberosum* DM1-3 516 R44, *S. tuberosum* cv. Umatilla and *S. lycopersicum*. For *ScCBF3 and ScCBF2* the corresponding pseudogenes were reported in A and B, respectively.

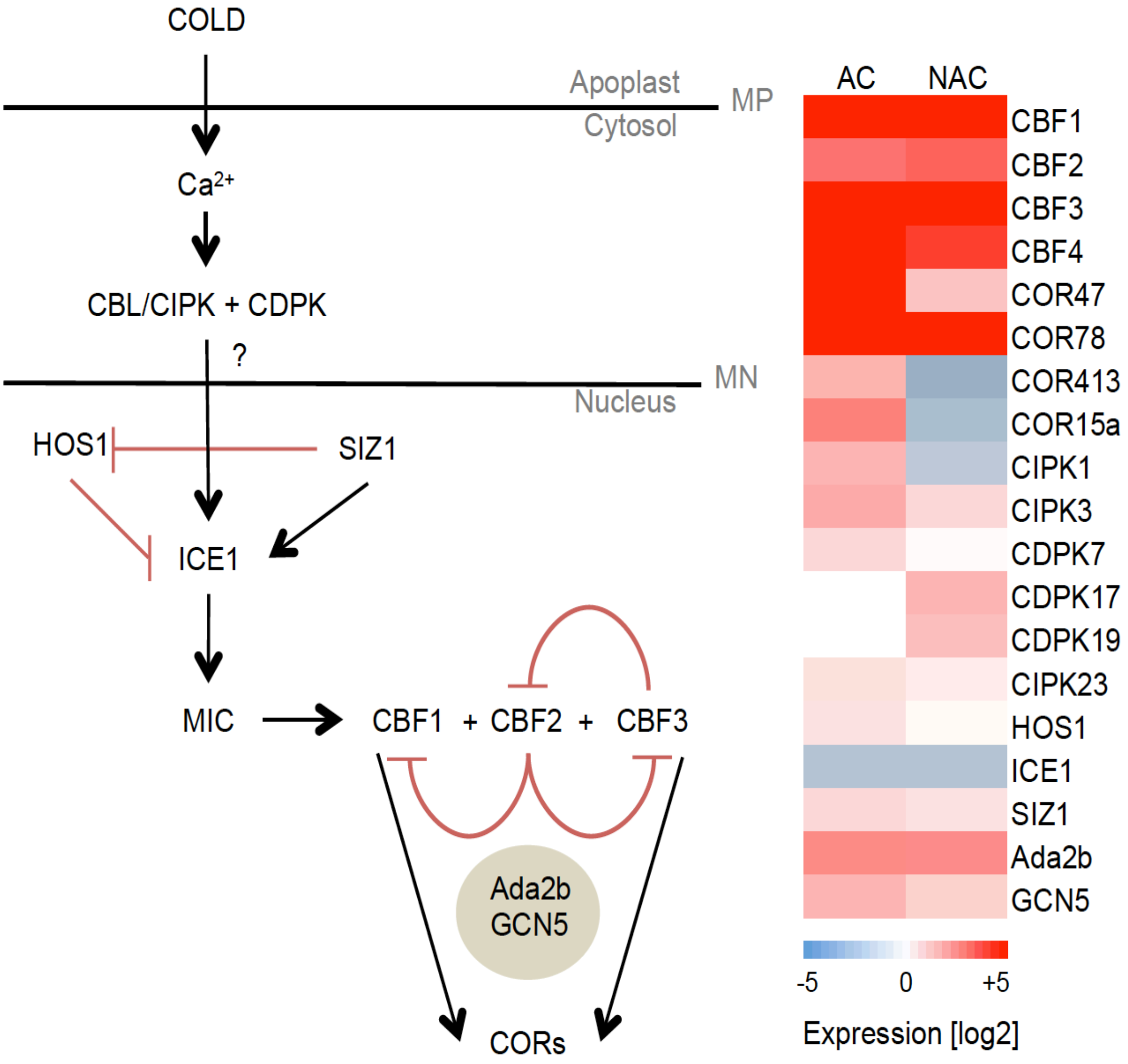


Figure 3.15. Cold-sensing and signaling pathway and gene expression heat map. Expression levels are indicated by shades of blue (down-regulation) and red (upregulation), where white indicates no differences between control and stressed plants (AC or NAC).

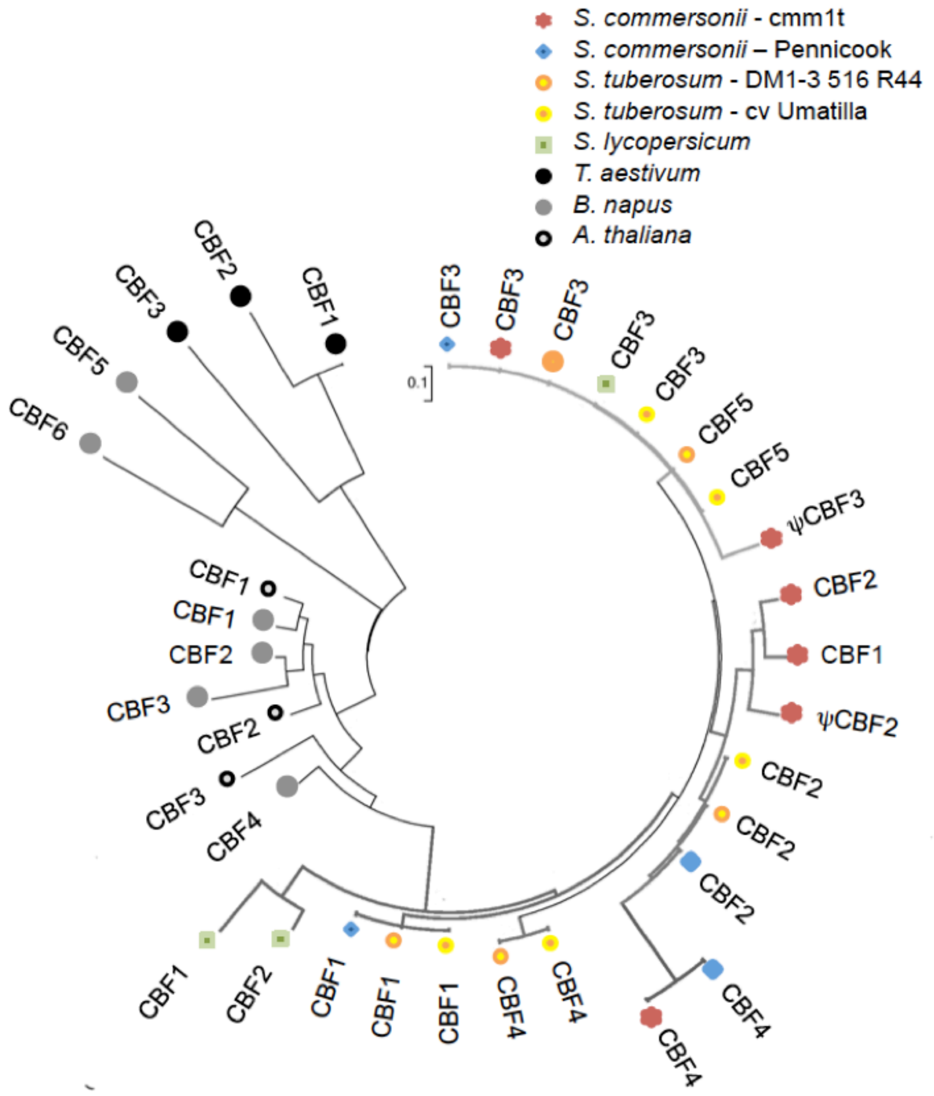


Figure 3.16 Structural organization of the StCBF and ScCBF regions.

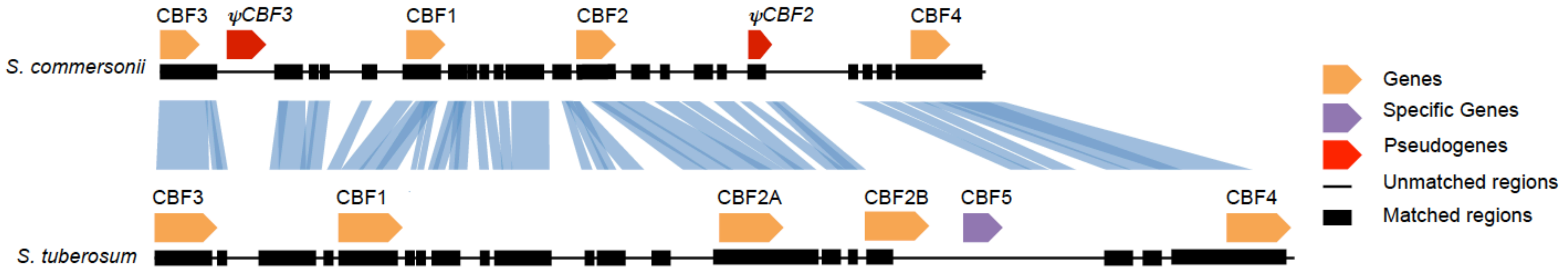


Figure 3.17. Similarity tree of the Solanum CBFs in relation to the *Arabidopsis, Brassica* and *Triticum* polypeptides having the CBF signature sequences. The bar indicates branch length scale

4. Discussion

Genome sequencing, assembly and gene annotation

To understand plant heritable traits, it is important to understand its genome. If a plant has a very complex genome it may be a challenge to correlate genome variation with important agronomic traits. Genomes have been sequenced for a number of different species including rice, sorghum, potato, tomato and maize (Matsumoto et al., 2005; Paterson et al., 2009; Potato Genome Sequencing Consortium, 2011; Tomato Genome Consortium, 2012; Schnable et al., 2009). The availability of genome sequences for these species enhanced the ability to understand quantitative trait loci (QTL), genes associated with domestication and drought tolerance in rice (Li et al., 2006; Degenkolbe et al., 2009), shoot fly resistance in sorghum (Satish et al., 2009), as well as biotic and abiotic stress resistance in maize, tomato and potato (Chung et al., 2010; Andolfo et al., 2012). In each of these cases, the genome sequence provided the foundation to better understand important agronomic traits and assist crop improvement. For this purpose, there are different genetic approaches that can be used and facilitate the conservation and utilization of genetic diversity and accelerate plant breeding. The terms "next generation plant breeding" is increasingly becoming popular in crop breeding programmes, conferences, scientific fora and social media. Being a frontier area of crop science and business, it is gaining considerable interest among scientific community and policy makers and funds flow from entrepreneurs and research funding agencies. Plant breeding is a continuous attempt to alter the genetic architecture of crop plants for efficient utilization in food, fodder, fiber, fuel or other end uses. Although the scientific concepts in plant breeding originated about 100 years ago, domestication and selection of desirable plants from prehistoric periods have contributed tremendously to ensure food security to the human being (Gepts, 2004). In the last few years, supported crop improvement programmes for major crops have started reaping benefits from cutting edge technologies of biological sciences, with examples in molecular markers and transgenic crop development which, in combination with conventional phenotype based selection, defines the current generation plant breeding practices. Recent genome sequencing efforts in other species such as apple, strawberry, cocoa are promising similar usefulness to researchers seeking to improve varieties (Velasco et al., 2010; Argout et al., 2011; Shulaev et al., 2011). In the potato, as in all species of agricultural interest, the tuber and plant characteristics are closely related to genome and this one can be improved by identifying appropriate breeding strategies aimed to obtain new varieties with gene constitutions able to ensure the desired performance. The potato commonly grown in Europe, originated from a small number of genotypes and it is a classic example of cultivated species that lacks genetic variability. Therefore its wild relatives have long been used in breeding programs as sources of valuable characters to improve and enrich the genetic base (Carputo et al., 2002). The introduction of useful genes from the wild gene pool to the cultivated potato is often hindered by the presence of reproductive isolation mechanism. The majority of wild potato species, in fact, is sexually isolated from the common potato not only for the different ploidy but also because there are sexual post-zygotic barriers that determine the hybrid endosperm degeneration (Camadro et al., 2004). These barriers can be overcome thanks to ploidy manipulation, bridge crosses, hormone treatment, recovery of embryos and somatic hybridation by protoplasm fusion. Despite all the interesting traits of wild potato species and the available of methods to access to this genetic resources, wild potato is still underutilized (Pavek and Corsini, 2001). Only a small part of the wild germplasm is used in the transfer of resistance characters; example are the resistance to Colorado potato beetle transferred from *S. berthaultii*, or the resistance to potato cyst nematode (*Globodera pallida*) transferred from *S. vernei* and *S. sparsipilium* (Ortiz et al., 1997). It follows that there is a high disparity between the large number of wild species that show interesting characters (about 200) and their realistic use in potato breeding. In the development of new verities breeders often do not have enough time or resource to transfer useful genes from wild species to cultivated potato, as this process involves numerous cycles of crossing and selection to eliminate unknown characters (Pavek and Corsini, 2001). The main problem is that the use of wild species in breeding program often involves the introduction, into the new variety, not only of useful genes but also deleterious genes and negative effects characteristic of linkage drag. This is true also for *S. commersonii* the species used in this study. It is a diploid species with endosperm balance number (EBN) = 1 and so it is sexually isolated from diploid 2EBN species and both tetraploid (2n = 4x = 48, 4EBN) and haploid (2n = 2x = 24, 2EBN) *S. tuberosum*. To overcome these incompatibility barriers in Portici breeding approaches based on bridge ploidies have been used (Carputo et al., 1997). The ploidy, and consequently the EBN, of *S. commersonii* has been doubled through in vitro regeneration of leaf explants and 4x (2EBN) genotypes obtained have been crossed with *S. tuberosum* (2ENB). The triploid hybrids produced 2n eggs and have been used with success in 3x x 4x that have produced pentaploid further used in backcrosses. The selection of the hybrids obtained with the classic methods is very difficult. To solve this problem, one drive is the use of genomics methods to reveal the genetic blueprints for different plant species as well as resolving genome differences in hybrids generated through the approach aforementioned. Genome technology has advanced substantially since publication, due to its suitability as a model species for plant research and its small genome, of the first plant genome sequence of *Arabidopsis thaliana*, in 2000 (Arabidopsis Genome Initiative, 2000). Plant genomics researchers have readily embraced new algorithms, technologies and approaches to generate genome, transcriptome and epigenome datasets for model and crop species that have permitted deep inferences into plant biology. Challenges in sequencing any genome include ploidy, heterozygosity and paralogy, all which are amplified in plant genomes compared to animal genomes due to the large genome sizes, high repetitive sequence content, and rampant whole- or segmental genome duplication. Traditional genome sequencing approaches are increasingly giving way to *de novo* assembly of NGS data. This new approach sacrifices assembled sequence quality for speed and greatly reduced costs. The ability to generate *de novo* genome assemblies provides an alternative approach to bypass these complex genomes and access the gene space of these recalcitrant species. In this work we *de novo* sequenced the genome of the stress-tolerant *S. commersonii*, as an integral step towards deciphering the genetic bases of traits that can be improved with this wild germplasm donor. The resulting *S. commersonii* genome assembly is comparable in length to the reference *S. tuberosum* genome, but divergence between the two sequences is demonstrated by the presence of SNPs and indels impacting target intergenic regions. The genome dimension is the first interesting data that it is possible to compare with the information of other plant genome sequencing. In the last years, more then fifty pant genomes were studied than span several orders of magnitude from the carnivorous corkscrew plant (*Genlisea aurea*) at 63 megabases (Mb) to the rare Japanese *Paris japonica* at 148,000 Mb (Bennett and Leitch, 2011). The smallest published genome is the carnivorous bladderwort (*Utricularia gibba*) at 82 Mb, while the largest, the Norway Spruce (*Picea abies*), stands by itself at 19,600 Mb. The genome estimating size of *S. commersonii* placed it exactly between the two other most significant *Solanum* species sequenced: potato and tomato.

These rearranged sites, together with genome-wide analyses of SNPs and indels, will shed light on selection processes shaping intergenic spaces and will likely facilitate the identification of polymorphic markers. In addition, the distribution of SNPs across *S. commersonii* genes indicated high variability in genes related to specific biological processes such as macromolecule metabolic processes, response to stimuli, carbohydrate derivative binding, localization, and ion binding. The enrichment of SNPs within particular functional classes of genes may reflect the environment in which *S. commersonii* evolved, and this is particularly interesting because genome-wide structural and gene content variations may drive the phenotypic diversity that characterize different species (McHale et al., 2012). Our data highlighted a striking difference between *S. commersonii* heterozygosity (1.5%) and that of the common potato (53-59%, (Hirsch et al., 2013)). The high level of potato heterozygosity reflects progress made in the past ~150 years of concerted potato breeding to maximize heterosis (Jansky, 2006). Since the magnitude of the difference in heterozygosity in *S. commersonii* relative to *S. tuberosum* was considerable, fundamental questions arise concerning the genetic constitution of wild relatives of potato with regard to gene flow and population structure (Hirsch et al., 2013). From a practical perspective, it seems that the use of species-wide diversity rather than individual accessions would be much more desirable to broaden the narrow genetic base of and increase allelic diversity in cultivated potato.

In my study, was also found that differences exist between *S. commersonii* and other solanaceous species in terms of repetitive sequences. Transposable elements (TEs) are major components of all plant genomes studied, shaping genome structure and organization. In a comprehensive review of the first 50 sequenced plant genomes, Michael and Jackson, (2013) reported that genome repetitive content ranged from 3% (*Utricularia gibba*) to 85% (*Zea mays*). Plant genomes are very often composed of a large part with TEs (Bennetzen 2000), which contain protein-coding sequences that are often annotated as genes. One example is reported by Bennetzen et al, (2004) that estimated in rice that only 40,000 of the more than 55,000 annotated genes are effectively genes and the more than 10,000 are TEs-usually low copy. TEs include various families that move via copy-and- paste (class I) and cut-and-paste (class II) mechanisms. TEs can strong influence the real size of a genome and the best example occurred in a relative of rice *(Oryza australiensis*) with a genome nearly two-fold larger than rice (Piegu et al., 2006). TEs strongly affect expression levels and transcript splicing and consequently may impact plant phenotypes. Compared to potato (55%) and tomato (63%), *S. commersonii* showed a lower amount of repetitive DNA (~383 Mb, accounting for 44.5% of the current assembly). As in other *Solanaceae* species, there were many more Ty3-*gypsy* type than Ty1-*copia* type LTR-RTs identified in *S. commersonii*, suggesting that the former elements have been somewhat more successful in colonizing and persisting in *Solanaceae* genomes. Moreover, the ratio of Ty3-*gypsy*:Ty1-*copia* might also be driven by variation in the efficacy of illegitimate recombination and/or unequal homologous recombination in removing LTR-RTs from the genomes, as reported in Arabidopsis, maize, barley and rice (Bennetzen, 2007). LTR-RTs play a substantial role in genome size variation, and the lower frequency of TEs in *S. commersonii* may contribute to its smaller genome size as well as underline the occurrence of different evolutionary dynamics in individual *Solanaceae* species genomes since their separation from a common ancestor. Vitte and Bennetzen, (2006), suggested that the proportion of TEs in different genomes might be influenced by destabilization of epigenetic regulation. Nevertheless we must consider that the TEs biology is a very interesting area of research but is very complicated and relies on relatively complete genomes so that TEs are captured in sequence contigs and can be accurately annotated. Schemes for classification of TEs have been agreed on (Wicker et al., 2007), but annotation of non-LTR TEs is complicated by the lack of structural clues that allows routine *ab initio* prediction (El Baidouri and Panaud, 2013). Another complication is that in genomes produced by short read DNA sequencing technology, TEs are often missed in the assembly due to their repetitive nature. So only thanks to the excellent quality sequencing of *S. commersonii* it was possible to annotate with precision the per cent and to classify the different family of TEs.

In this study we targeted different tissues to best represent the *S. commersonii* transcript repertoire. Annotation of plant genomes is difficult especially as the definition of what constitutes a gene continues to evolve. We must remember that a large parts of the genome are 'expressed' in that RNAs are formed, but do not correspond to traditional genes in that they are not translated to a protein. However, most annotated plant genomes showed between 22,285 (*Selaginella moellendorffii*) and 94,000 (*Triticum aestivum*) genes and in the *de novo* assembly of *S. commersonii* transcriptome from leaf, flower, stolon and tuber tissues allowed identification of 37,662 genes. Even though the number of genes found in *S. commersonii* was similar to that of *S. tuberosum*, the number of transcripts differed greatly between the two species. This might highlight the presence of more prominent alternative splicing activities in potato than in *S. commersonii*. This is consistent with observations by the Potato Genome Consortium (Potato Genome Sequencing Consortium, 2011) that ~25% of potato genes encoded two or more isoforms, indicative of more functional variation than is represented by the gene set alone. It is possible to hypothize that during the evolution polyploidization events, natural but also artificial selection including domestication, breeding and cultivation have probably influenced gene evolution. Even in comparison to the number of genes and transcripts annotated between *S. commersonii* and *S. tuberosum* it was necessary to take into account the different evolutionary processes that these two species have followed over the years. Differences between genomes most likely lies in the tools used for annotation and how relaxed the annotators were in calling genes as well as lineage-specific genes and gene family expansions. Genomes produced by next generation sequencing typically have smaller contig and scaffold sizes that complicate annotation as genes may not exist on single contigs but may be broken across contigs, thus inflating the number of annotated genes (e.g., pigeon pea, Varshney et al., 2012). So it is important to start from an excellent genome assembly quality to perform the gene annotation. One measure of genome assembly quality is the contiguity or the length of contigs and scaffolds at which 50% of the assembly can be found; this is commonly referred to as N50. *Sorghum*, *Brachypodium distachyon*, soybean, and foxtail millet have the top four scaffold contiguities with 62.4, 59.3, 47.8, and 47.3 Mb respectively and all four were sequenced using Sanger as part of the JGI pipeline. However, the genome with the ninth largest scaffold N50 is the tomato genome at 16 Mb, which was predominantly assembled using 454. Each scaffold is comprised of thousands of contigs and contig length generally drives the completeness and quality of the gene predictions. The *S. commersonii* genome was assembled using Illumia technology and has produced N50 length (scaffold) of 44 Mb. This is a quality result that has subsequently allowed to produce a genome annotation of high value.

We also identified ~21,000 *S. commersonii* ncRNAs. Emerging evidence has revealed that ncRNAs are major products of the plant transcriptome (Rymarquis et al., 2008). They may have significant regulatory importance, especially during stress situations (Matsui et al., 2013). In this study, the perfect or near-perfect match to target-sites allowed effective prediction of the target sequences by computation (Rhoades et al., 2002) and revealed that more than 20% of ncRNAs targeted cold responsive genes. This would strengthen the claim that ncRNAs play a role in adaptation to stresses and tie their importance specifically to the unique *S. commersonii* phenotype. Manipulation of miRNA/siRNA-guided gene regulation may enable engineering of plants for improved stress-resistance (Sunkar et al., 2006; Katiyar-Agarwal et al., 2007). Therefore, detailed analyses on miRNA-guided stress responsive gene regulation in *S. commersonii* may lead to new insights for an efficient exploitation of this germplasm.

Phylogenomic Analysis Across Plant Species

To assess evolutionary relationships between *S. commersonii* and other sequenced plant genomes, we undertook a comprehensive phylogenomic approach. This involved reconstruction of the complete collection of evolutionary histories of all *S. commersonii* protein-coding genes across a phylogeny of 12 sequenced plants (i.e., the phylome). The usefulness of this approach in the annotation of newly sequenced genomes has been demonstrated in other plants (Garcia-Mas et al., 2012; Dohm et al., 2014) and animals (Garcia-Mas, 2013). In total, 17,297 (44%) *S. commersonii* genes showed a one-to-one orthologous relationship with *S. tuberosum* genes and 16,821 (42%) with *S. lycopersicum* genes, but only 7,058 (18%) with genes from the more distantly related Asterid *Mimulus guttatus.* This scenario likely results from the past genome duplication shared by the three *Solanum* species (Tomato Genome Consortium, 2012), followed by differential loss of paralogous genes in each of the species. The overall average number of duplications per node (duplication density) was 0.66, 0.93, and 0.94 for *S. commersonii,* potato-ancestor and *Solanum* ancestor, respectively, whereas we found a low rate of 0.066 for the common ancestor of Asterids. Collectively, these numbers suggest multiple rounds of duplications, at least at the node separating the *Solanum* ancestor from other Asterid taxa, represented in this analysis by *M. guttatus*. These two rounds of ancestral duplications were also previously suggested by comparison of the domesticated potato and tomato genomes (Potato Genome Sequencing Consortium et al., 2011; Tomato Genome Consortium, 2012). Our study showed that these ancestral duplications were also shared with the wild species lineage represented by *S. commersonii*, as would be predicted by commonly accepted *Solanum* phylogeny. The analysis of paralogous gene pairs revealed that all three relative ages show a similar pattern the two most prominent peaks largely preceding the divergence of *S. commersonii* and tomato. These results, in combination with the topological dating of duplications, strongly suggest that these major duplications predate the divergence of the *Solanum* species, and that most paralogous pairs dated as potato-specific or *S. commersonii*-specific result from differential retention of duplicated pairs in each of the investigated lineages. We assessed the genomic organization of these recent duplicates and found that most were present in tandem (314) or were closely associated along the same contig (141). Thus, results obtained are not compatible with a recent, specific genome duplication in *S commersonii* but rather with differential retention of paralogues from ancient duplications and additional lineage-specific segmental duplications that blurred the syntenic and one-to-one correspondence between *S. commersonii* and *S. tuberosum*. This in turn may underlie the sexual incompatibility between these two related species.

Pathogen-Receptor Genes

During the stage in Spain at Sequentia Biotech, a strong effort was dedicated to the development of a new pipeline for the research of pathogen receptor genes and the identification of new genes involved in freezing tolerance. As already reported, *S. commersoni* is the potato species with highest resistance to low temperatures but it also has other interesting traits such as resistance against Bacterial wilt, Black leg, Early blight, Ring rot, Verticillium, PVY*, Meloidogyne spp.* and *Phytophthora.* The plants, using accurate and sophisticated defence systems to interact with the surrounding reality, survive and evolve. Fundamentally the interaction plant-pathogen, and so the susceptibility and resistance of any species, is due to a complex molecular machine that can recognize pathogens, activate and defend, at multiple levels (physical, chemical, molecular), the attacked organism. (Jones and Dangl, 2006). This short description of the plant immune system shows how it is deeply interconnected with other units of plant and how the study of this topic is complicated by the presence of multiple factors and the complexity of their interactions. Among different ways in which plants exert this function, there is one in which the main factors are the sentinel so called *R-genes,* or resistance genes (Ellis et al., 2000) These genes produce proteins that can recognize pathogen specific proteins and transduce a signal that activates the cellular defence responses (Means et al., 2000). The literature reports that the *R-gene* family is divided into 4 classes, each with specific characteristics that make proteins discriminable from each other. The protein structures of the four classes are composed of different domains, but all have the same function: to recognize a molecule and transduce a signal. Recognition is given in all classes from a specific domain: the LRR. This domain, with functions of binding and recognition of molecules (DeYoung and Innes, 2006), is ubiquitous in all the resistance genes and can be connected to different transducers, which complete their function. The transducers of R-proteins that are know can be characterize in:

A: very complex, as the case of the class TNL, constitute from TIR domains (Means et al., 2000) and NBS (McHale et al., 2006) that transduce the signal and bind ATP;

B: medium complex, as the class that contains the ATP-binding CNL (NBS domain) has a supercoiled region, or like the class RLK that has a kinase domain linked to the LRR domain (Morillo and Tax, 2006), with the function of signal transducer;

C: low complex, such as class RLP where signal transduction occurs directly from the LRR domain closely related to a domain serine - threonine kinase;

Despite the fact the *R-genes* can be easily classified in the respective classes, their structure is to such an extent that a single amino acid change can completely change the recognition and the protein function (Anderson et al., 1997). The idea to develop in *silico* systems for the detection of *R-genes*, instead of using of classical gene characterization approach through molecular biology, was born observing the developments of the systems of genomic analysis in recent years. The development of new prediction systems is designed to create a bridge between the information stored in the public data base and the genetics plant world (Solovyev and Salamov, 1997). MATRIX RELOADED is a prediction program, developed during my stage at Sequentia Biotech, that was used for the annotation of *R-genes* in *S. commersonii* and *S. tuberosum.*

MATRIX RELOADED is a *de novo* prediction program based on algorithms and softwares already known, but that proposes a new system for the prediction of protein structures. Like many protein analysis software (PFAM uses the same operating system, but in a more generic way) it is based on the use of HMM profiles (Del Sal et al., 1989). Using the PFAM's HMM profiles (for the LRR domains, NBS and TIR) for search *R-genes* in known data set, we managed to get only 25% of the actual number of *R genes* present in that particular set. So in this project we built a very stringent HMM profile started from sequences of the *R-gene* family. The proteins analysed through MATRIX RELOADED represent the amino acid sequences capable of taking specific conformations and carry out specific functions. Analysing large data sets with this system it is possible to predict all the similar proteins to the *R-genes*, but also makes it easy to see which proteins are associated with new proteins whose function is already known. MATRIX RELOADED has not been built as an aseptic bioinformatics tool, but as a genetic tool that uses computer science to give a biological sense to the results that it produces. For the first time, this new tool was used to check the proteome of a wild potato species and the candidate *S. commersonii R-genes* were catalogued and compared to the *R-gene* complement comprising the cultivated potato and tomato genomes. Our data revealed that *S. commersonii* encodes fewer *R-gene* candidates than *S. tuberosum,* but more than tomato.

Polyploidization, genome size variation, natural selection, artificial selection including domestication, breeding and cultivation, and gene family interactions have probably influenced pathogen recognition gene evolution in *Solanum* (Andolfo et al., 2013). Differences in copy number of *R-gene* family are important source of genetic variation likely to play a role in phenotypic diversity and adaptation in different species (Peele et al., 2014), but also genome expansion or contraction favour diversification. Our analysis pinpointed that different R locus arrangements occurred after the separation from a common *Solanum* ancestor. Indeed, the size of R1 locus could increase/decrease of 10 times among genomes tested. Previous comparative analysis of R1 locus revealed highly conserved collinear flanking regions, showing high variability and tandem duplicated genes, namely R1 homologs and F-box containing genes (Ballvora et al., 2007). Solanum *R-genes* architecture seems to be shaped by the interplay of large-scale gene organization that determine the conservation of loci order and an extensive local genome rearrangements mediated by tandem duplication, transposon microRNA and other shuffling elements that determine distinct loci arrangements (Zhang et al., 2014). The *Solanum* local adaptive divergence events caused by specific needs can underlie substantial phenotypic resistance diversity and should be further investigated.

Nonacclimated and Cold Acclimated Gene Expression and Regulation

The most interesting resistance traits of *S. commersonii* that are being studied since many years at the University of Naples are the resistance to low temperature and the capacity to cold acclimate. Freezing temperatures adversely affect plant productivity in many parts of the world. Traditional plant breeding methods for improvement of freezing resistance in crop plants by using field selection (frost or winter survival) have achieved only limited success. Much of the failure to achieve greater success has been attributed to lack of genetic diversity, lack of effective selection criteria, and limited or inconsistent information on genetic control of freezing resistance. In the last years different comparative studies were carried out between freezing-tolerant, cold-acclimating *S. commersonii* and a freezing-sensitive, non acclimating *S. tuberosum*. Since no sequencing data on this wild potato genome was well known, most studies were focused on specific metabolic pathways that in literature were described as involved in freezing tolerance. One example of this comparative study is that reported by Palta et al., (1993). The author determined differentiation of plasma membrane lipid changes associated with increased freezing tolerance following acclimation. The lipid changes analysed included a decrease in palmitic acid, an increase in unsaturated to saturated fatty acid ratio, an increase in free sterols, an increase in sitosterol, and a slight decrease in cerebrosides. These changes were either absent or opposite in the NA species, suggesting an association of these lipid changes with cold acclimation. Nevertheless, the molecular basis of non-acclimated tolerance is poorly understood although it has been reported that it may be genetically determined by loci independent of acclimated tolerance in potato (Stone et al., 1993), willow (Tsarouhas et al., 2004), and oilseed rape (Teutonico et al., 1995). Thanks to the annotation of *S. commersonii* genes and the plasticity of Matrix-R it was possible, for the first time, to identify a complete set of putative cold resistance genes in this species and at a later stage to analyse the cold-responsive transcriptome of *S. commersonii*. Overall, the whole-genome expression data highlighted an extensive reorganization of the transcriptome under cold stress, with enhanced expression of genes affecting ROS scavenging enzymes (e.g. superoxide dismutase, SOD; catalase, CAT; ascorbate peroxidase, APX), those involved in cell repair (such as heat-shock proteins, HSPs and dehydrins, DHNs), and those encoding proteins that may function as osmoprotectans. Among the latter, we found a significant up regulation of *S. commersonii* galactinol synthase (ScGOLS1) under both NAC and AC conditions. Previously, over-expression of MfGolS1 (from *Medicago falcata*) or BhGolS1 (*Boea hygrometrica*) promoted the biosynthesis of increased amounts of raffinose family oligosaccharides (RFOs), such as galactinol, raffinose and stachyose and resulted in elevated tolerance to low temperatures in transgenic tobacco plants (Zhuo et al., 2013; Wang et al., 2009). We hypothesize that high expression of ScGOLS1 in conjunction with the increased activity of the afore mentioned cold-associated and -inducible proteins might contribute to *S. commersonii* frost tolerance. One notable observation was that several genes were responsive to cold relative to control conditions, but with contrasting kinetics, under AC vs. NAC. For instance, brassinosteroid-signaling kinase 1 (BSK1) was activated under AC and suppressed under NAC. Conversely, one MYB and one bHLH TF were cold induced under NAC and repressed under AC. As MYB and bHLH proteins often interact with each other to control transcription (Stracke et al., 2001; Heim et al., 2003; Ramsay and Glover, 2005), this differential expression of MYB and bHLH TF suggests that the regulation of some cold-responsive genes may be achieved by modulating the ratio of these partners. TFs were mostly upregulated under both conditions, as was observed in *Arabidopsis* (Lee et al., 2005). This is consistent with overall up regulation, rather than repression, of gene transcription following cold stress. Specifically, we identified 25 TFs correlating positively with acclimated and non-acclimated tolerance, and only 11 that showed negative correlations. Among the negatively correlating TFs was a Cys-2/His-2-type (C2H2) zinc-finger protein (Sakamoto et al., 2004). C2H2 zinc-finger-type TFs have been found to work downstream of DREB1/CBF and to be responsible for stress tolerance in plants (Sakamoto et al., 2004).

The comparison of cold-responsive gene expression profiles between AC and NAC stressed plants highlighted remarkable features of cold-responsive plant genes known to be critical in cold sensing and signaling pathways. Two calcium-dependent protein kinases, *CDPK17* and *CDPK 19 (CPK8)* were differentially expressed. Interestingly, *CDPK19* was up regulated only under NAC conditions, whereas *CDPK17* expression required acclimation. Neither gene has been previously implicated in cold stress response. Thus, the induction of *CDPK19 (CPK8)* and *CDPK17* in *S. commersonii* suggests possible independent roles in response to freezing and cold acclimation, respectively. Also the structural organization and transcriptional activity of the ScCBFs (C-repeat binding factors) revealed intriguing features. Our cross-species comparisons indicated that the CBFs underwent to rapid expansion via duplication processes. In *S. commersonii* we found two pseduogenes, ye found two pseduogenes, dent roles in response to freezing and cold acclimation, respectively. Also the structural organization and transcriptional activity of the ScCBFs (C-repeat binding factors) revealed intriguing features. Our cross-species comparison potato and *S. commersonii* from their most recent common ancestor. In particular, the paucity of sequence change indicates that, after the divergence of the two species lineages, there were likely strong constraints on CBF3 that conserved protein sequence. A different situation was found for the *ScCBF2* pseudogene, which shares only 80% of identity with the functional *ScCBF2* gene. This suggests that the gene duplication occurred prior to divergence of the *S. tuberosum* and *S. commersonii* lineages from their most recent common ancestor, with the duplicated copy subsequently undergoing rearrangements as observed also in other duplicated genes (Ohno, 1970; Lynch and Force, 2000). Phylogenetic analysis highlighted a common origin of CBFs in *Solanum* species with respect to other plants from temperate regions that can cold acclimate, such as *Arabidopsis*, wheat and *Brassica napus* (Jaglo et al., 2001). This suggested a homogenization mechanisms occurring in *Solanum,* as previously reported (Pennycooke et al., 2008). Despite observed orthology for most of the CBF1 gene family, ScCBF1 clustered apart the other CBF1 sequences analysed. This might be the result of strong selection pressure towards functional diversification of ScCBF1. Taken together, our data are consistent with a hypothesis of rapid evolution of CBFs within the genus *Solanum* (Carvallo et al., 2011). We hypothesize that a duplication event occurred after the *S. tuberosum-S. commersonii* divergence and may have led to a different functionalization of the ScCBF3 pseudogene, resulting in enhanced cold response capability in *S. commersonii*. To more deeply investigate the role of CBFs in *S. commersonii*, transcript levels were monitored both under AC and NAC. Our data showed that all ScCBFs were up regulated under all tested conditions relative to controls. This is in contrast with previous reports that *CBF1*, but not related CBFs, were responsive to low temperatures in both *S. commersonii* and *S. tuberosum* (Pennycooke et al., 2008; Carvallo et al., 2011). Our observations parallel patterns observed in tomato species. In cold-sensitive cultivated tomato, only *CBF1* was unregulated in response to cold treatments, whereas in the cold-tolerant wild tomato species *S. peruvianum,* all three CBF genes were cold responsive (Mboup et al., 2012). High expression of *S. commersonii* CBF genes and genes regulated by CBF proteins (e.g., COR genes) may be directly responsible for enhanced cold tolerance and acclimation ability in this species.

5. Conclusions

The potato ranks fourth among the most important crops in the world after wheat, corn and rice. The annual production is estimated about 300 million tons. It is a tetraploid species (2n = 4x = 48), propagated through clones and has a very low genetic variability. However among the wild potato species a high genetic diversity is available. It has been estimated that only 10% of these species has been used to develop new improved potato verities (Gavrilenko, 2011). Therefore, there is a large gap between the high biodiversity available in the about 200 wild potato species and their effective use breeding (Pavek and Corsini, 2001). Many wild species particularly resistant to pathogens or abiotic stress are also sexually isolated from the tetraploid cultivated potato. Therefore, there are many difficulties in their introduction and use in classical breeding programs, which often require long and laborious methods for obtaining hybrids and for the selection of materials. To broaden the genetic base and to transfer interesting genes from wild potato species of *Solanum* in the scientific community it is increasingly felt the need to develop new tools based on the structural and functional genomics, allowing a more efficient exploitation of the wide genetic background of wild potatoes and therefore, to increase the current strategies of potato breeding. The experimental work carried out in this PhD reports the first genome sequence for a wild relative of the cultivated potato. We believe that this work will contribute to extend the limited information available on the genomic structure of wild potato species from a practical perspective, the genome sequence of *S. commersonii* and all the data obtained will provide the genomic tools for an efficient exploitation of this species and its noteworthy genes. We think that four main evidences emerged from this research.

- The resulting *S. commersonii* genome assembly is comparable in length to the reference *S. tuberosum* genome. The new gene annotation show that the two compared species have a similar number of genes but divergence between the two sequences is demonstrated by the presence of SNPs and indels impacting target intergenic regions. Moreover, interesting is that the number of transcripts annotated differed greatly between the two species. This might highlight the presence of more prominent alternative splicing activities in potato than in *S. commersonii*. Our data also highlighted a striking difference between *S. commersonii* heterozygosity and that of the common potato. The high level of potato heterozygosity reflects progress made in the past ~150 years of concerted potato breeding to maximize heterosis.
- We have used the orthology and paralogy proteins relationships for the reconstruction of the complete collection of evolutionary histories of all *S. commersonii* protein-coding genes

across a phylogeny of 12 sequenced plants. Our results show that during its evolution, *S. commersonii* has undergone three different ancient duplications events and that most paralogous pairs dated as potato-specific or *S. commersoni*-specific result from differential retention of duplicated pairs. Thus, results obtained are not compatible with a recent, specific genome duplication in *S commersonii* but to an additional lineage-specific segmental duplications though which the wild specie has preserved a set of paralogues genes, loosed by the cultivated potato during it evolution. This in turn may underlie the sexual incompatibility between *S .tuberosum* and *S. commersonii.*

- Our results showed that the number of *R-genes* annotated in *S. commersonii* is lower than that of potato but higher than that of tomato. Also the analysis of a specific locus involved in resistance to *Phytophthora infestans* suggested that it has suffered a substantially rearrangement compared with the same locus of the cultivated potato. This marks the fact that polyploidization, genome size variation, natural selection, domestication, breeding and cultivation, and gene family interactions have probably influenced pathogen recognition gene evolution in *Solanum.*
- The last consideration is based on the identification of new cold regulated genes. It is the first time that a complete catalogue of these genes for this species is produced. The information we generated provides a foundation for further experiments to explore the network of gene regulation required for cold tolerance and acclimation and to determine the function of cold-responsive genes through molecular and cellular approaches. Future challenges include translation of this new knowledge into advances in potato. A GBS map will be developed and molecular markers linked to genes on interest will be identified. This will allow a combined breeding approach based on molecular and bioinformatics tools, and in the knowledge of classical genetics and breeding principles.

6. References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**: 716–723.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.

Anderson, P.A., Lawrence, G.J., Morrish, B.C., Ayliffe, M.A., Finnegan, E.J. and Ellis, J.G. (1997) Inactivation of the flax rust resistance gene M associated with loss of a repeated unit within the leucine-- ‐ rich repeat coding region. Plant cell. **9**: 641 ‐ 651.

Andolfo, G., Aversano, R., Frusciante, L., Ercolano, M.R., and Sanseverino, W. (2013). Genome-wide identification and analysis of candidate genes for disease resistance in tomato. Mol Breeding: 1–7.

Argout, X. , J. Salse , J.-M. Aury , M. J. Guiltinan , G. Droc , J. Gouzy , M. Allegre , et al . (2011). The genome of the obroma cacao . Nat Genet. **43** : 101 – 108 .

Ariyaratne, P. N., & Sung, W. K. (2011). PE-Assembler: De novo assembler using short paired-end reads. Bioinformatics. **27**: 167–174.

Bamberg, J.B., Martin, M.W., and Schartner, J.J. (1994). Elite selections of tuber-bearing Solanum species germplasm (Inter-Regional Potato Introduction Station,NRSP–6: Sturgeon Bay, Wisconsin).

Ballvora, A., Jöcker A, Viehöver P, Ishihara H, Paal J, Meksem K, Bruggmann R, Schoof H, Weisshaar B, Gebhardt C. (2007) Comparative sequence analysis of Solanum and Arabidopsis in a hot spot for pathogen resistance on potato chromosome V reveals a patchwork of conserved and rapidly evolving genome segments. BMC Genomics. **8**:112.

Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, JP., Lander, E.S. (2002). ARACHNE: A whole-genome shotgun assembler. Genome Res. **12**: 177–189.

Bennett, M.D., and Leitch, I.J. (2011). Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. Ann. Bot. **107**: 467–590.

Bennetzen, J.L., Coleman, C., Liu, J. Ma, R., Ramakrishna ,W. (2004). Consistent over-estimation of gene number in complex plant genomes. Curr. Opin. Plant Biol. **7**: 732–736.

Bennetzen, J.L. (2007). Patterns in grass genome evolution. Curr. Opin. Plant Biol.

Boerner, S. and McGinnis, K.M. (2012). Computational identification and functional predictions of long noncoding RNA in Zea mays. PLoS ONE. **7**: e43047.

Bradeen, J.M. and Haynes, K.G. (2011). Introduction to potato. In Genetics, genomics and breeding of potato, J.M. Bradeen and C. Kole, eds (CRC Press/Science Publishers: Enfield, NH), pp. 1–19.

Bradshaw, J.E. and Ramsay, G. (2005). Utilisation of the Commonwealth Potato Collection in potato breeding . Euphytica. **146**: 9-19.

Bryan, G.J., Hein, I. (2008). Genomic resources and tools for gene function analysis in potato. Int J Plant Genom. **2008**:216513.

Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., Jaffe, D.B. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res. **18**: 810–820.

Camadro E., Carputo D., Peloquin S.J. (2004): Substitutes for genome differentiation in tuberbearing Solanum: interspecific pollen-pistil incompatibility, nuclear cytoplasmic male sterility, and endosperm. Theoretical and Applied Genetics. **109**:1369-1376.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. **25**: 2.

Cardi, T., D'Ambrosio, E., Consoli, D., Puite, K.J., and Ramulu, K.S. (1993). Production of somatic hybrids between frost-tolerant Solanum commersonii and S. tuberosum: characterization of hybrid plants. TAG Theor Appl Genet. **87**: 193-200.

Carputo, D. et al. (2009). Resistance to Ralstonia solanacearum of Sexual Hybrids Between Solanum commersonii and S. tuberosum. Am. J. Pot Res. **86**: 196–202.

Carputo, D., Alioto, D., Aversano, R., Garramone, R., Miraglia, V., Villano, C., and Frusciante, L. (2013). Genetic diversity among potato species as revealed by phenotypic resistances and SSR markers. Plant Genetic Resources: Characterization and Utilization **11**: 131–139.

Carputo, D. and Barone, A. (2005): Ploidy manipulation in potato through sexual hybridization. Annals of Applied Biology, **146**: 71-79.

Carputo, D., Basile, B., Cardi, T., Frusciante, L. (2000). Erwinia resistance in backcross progenies of Solanum tuberosum x Solanum tarijense and Solanum tuberosum(+)Solanum commersonii hybrids. Potato R. **43**:135-142.

Carputo, D., Aversano, R., Barone, A., Di Matteo, A., Iorizzo, M., Sigillo, L., Zoina, A., Frusciante, L. (2009): Resistance to Ralstonia solanacearum of Sexual Hybrids Between Solanum commersonii and S. tuberosum. American Journal on Potato Research; 86:196-202.

Carputo, D., Barone, A., Cardi, T., Sebastiano, A., Frusciante, L., and Peloquin, S.J. (1997). Endosperm balance number manipulation for direct in vivo germplasm introgression to potato from a sexually isolated relative (Solanum commersonii Dun.). Proc. Natl. Acad. Sci. U.S.A.. **94**: 12013–12017.

Carputo, D., Castaldi, L., Caruso, I., Aversano, R., Monti, L., and Frusciante, L. (2007). Resistance to frost and tuber soft rot in near-pentaploid Solanum tuberosum- S-commersonii hybrids. Breeding Science **57**: 145–151.

Carvallo, M.A., Pino, M.-T., Jeknic, Z., Zou, C., Doherty, C.J., Shiu, S.-H., Chen, T.H.H., and Thomashow, M.F. (2011). A comparison of the low temperature transcriptomes and CBF regulons of three plant species that differ in freezing tolerance: Solanum commersonii, Solanum tuberosum, and Arabidopsis thaliana. J. Exp. Bot. **62**: 3807–3819.

Cheng, J., M.G., Bolyard, R.C., Saxena, and Sticklen, M.B.. (1992). Plant Sci. 81**,** 83.

Chung , C.-L. , T. Jamann , J. Longfellow , and Nelson, R.. (2010). Characterization and fine-mapping of a resistance locus for northern leaf blight in maize bin 8.06. Theor Appl Genet. **121**: 205 – 227.

Conesa, A. and Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. **2008**: 619832.

Correll, D.S. (1962). The Potato and Its Wild Relatives. Contributions from the Texas Research Foundation, Botanical Studies.

Dai, X. and Zhao, P.X. (2011). psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res. **39**: W155–9.

De Young, B.J. and Innes, R.W. (2006) Plant NBS ‐ LRR proteins in pathogen sensing Nat Immunol. **7**: 1243-9.

Degenkolbe, T., P. Do, E. Zuther, D. Repsilber, D. Walther, D. Hincha, and K. K Öhl . (2009) . Expression profiling of rice cultivars differing in their tolerance to long-term drought stress. Plant Mol Biol. **69**: 133 – 153 .

Del Sal, G., Manfioletti, G. and Schneider, C. (1989) The CTAB ‐ DNA precipitation method: a common mini ‐ scale preparation of template DNA from phagemids, phages or plasmids suitable for sequencing. BioTechniques. **7**: 514 ‐ 520.

Dodds, K.S. (1962). The Potato and Its Wild Relatives. Contributions from the Texas Research Foundation. Botanical Studies. **4**: 517.

Dohm, J.C., Dohm, J.C., Minoche, A.E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., Rupp, O., Sörensen, TR., Stracke, R., Reinhardt, R., Goesmann, A., Kraft T., Schulz, B., Stadler, P.F., Schmidt, T., Gabaldón, T., Lehrach, H., Weisshaar B., Himmelbauer, H. (2014). The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature. **505**: 546–549.

Doležel, J., Greilhuber, J., Lucretti, S., and Meister, A. (1998). Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison. Annals of Botany. **82** : 17-26.

Ducreux L.J.M., W.L. Morris, P.E. Hedley, T. Shepherd, H.V. Davies, S. Millam and M.A. Taylor. (2005). J. Exp. Bot. 56, 81.

Eberlein, C. V., M. J. Guttieri and J. Steffen-Campbell. (1998). Bromoxynil Resistance in Transgenic Potato Clones Expressing the Bxn Gene. Weed Sci. **46**: 150-157.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. **5**: 113.

El Baidouri, M., and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol. Evol. **5**: 954–965.

Ellis, J., Dodds, P. and Pryor, T. (2000) The generation of plant disease resistance gene specificities. Trends Plant Sci, **5**: 373-379.

Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics? Genome Biol. **9**: 235.

Garcia-Mas, J. et al. (2012). The genome of melon (Cucumis melo L.). PNAS. **109**: 11872–11877.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**: 685–695.

Gavrilenko (2011): Application of molecular cytogenetics in fundamental and applied research of potato. In Genetics, genomics and breeding of potato. Edited by Bradeen J.M. and Kole C. CRC Press New York. pp. 184-206.

Gepts, P. (2004). Crop domestication as a long term selection experiment. Plant Breed. Rev. **24**: 1–44.

Grabherr, M.G. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**: 644–652.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**: 307–321.

Hanneman, R.E. and Bamberg, J.B. (1986). Inventory of tuber-bearing Solanum species 1st ed. (Potato Introduction Station: Sturgeon Bay, Wisconsin).

Hawkes, J.G. (1989). Nomenclatural and taxonomic notes on the infrageneric taxa of the tuber-bearing Solanums (Solanaceae). Taxon **38**: pp. 489-492.

Hawkes, J.G. (1990). The potato: evolution, biodiversity and genetic resources. (Belhaven Press: Washington, USA).

Heim, M.A., Jakoby, M., Werber, M., Martin, C., Weisshaar, B., and Bailey, P.C. (2003). The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol. Biol. Evol. **20**: 735–747.

Hetterscheid, W.L.A. and. Brandenburg, W.A. (1995), Taxon 44, 161.

Hirsch, C.N., Hirsch, C.D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., De Jong, W., Veilleux, R.E., Jansky, S.H., Bethke, P., Douches, D.S., and Buell, C.R. (2013). Retrospective view of North American potato (Solanum tuberosum L.) breeding in the 20th and 21st centuries. G3 (Bethesda) **3**: 1003–1013.

Hollick, J.B. (2008). Sensing the epigenome. Trends Plant Sci **13**: 398–404.

Holt, C. and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12**: 491.

Huamán, Z., Golmirzaie, A., and Amoros, W. (1997). The potato. In: D. Fuccillo, L. Sears, and P. Stapleton (eds.). Biodiversity in trust: conservation and use of plant genetic resources in CGIAR Centres. Cambridge University Press, Cambridge, UK. pp. 21–28.

Huamán, Z., Hoekstra, R., and Bamberg, J.B. (2000). The Inter-genebank potato database and the dimensions of available wild potato germplasm. Am. J. Pot Res. **77**: 353–362.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Research. 9: 868–877.

Huerta-Cepas, J. and Gabaldón, T. (2011). Assigning duplication events to relative temporal scales in genome-wide studies. Bioinformatics **27**: 38–45.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., and Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. **42**: D897–902.

Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python Environment for Tree Exploration. BMC Bioinformatics **11**: 24.

Huaman, Z. and Spooner, D.M. (2002). Reclassification of landrace populations of cultivated potatoes (Solanum sect. Petota). Am J Bot. **89**: 947-65.

Jackson, S.A. and Hanneman, R.E. (1999). Crossability between cultivated and wild tuber-and non-tuber-bearing Solanum. Euphytica. **109**: 51-67.

Jaglo, K.R., Kleff, S., Amundsen, K.L., Zhang, X., Haake, V., Zhang, J.Z., Deits, T., and Thomashow, M.F. (2001). Components of the Arabidopsis C-Repeat/Dehydration-Responsive Element Binding Factor Cold-Response Pathway Are Conserved inBrassica napus and Other Plant Species. Plant physiology. **127**: 7.

Jansky, S.H. (2006). Overcoming hybridization barriers in potato. Plant breeding. **125**: 1-12.

Johnston, S.A., Nijs, den, T.P.M., Peloquin, S.J., and Hanneman, R.E., Jr (1980). The significance of genic balance to endosperm development in interspecific crosses. TAG Theoretical and Applied Genetics. **57**: 5–9.

Jones, J.D. and Dangl, J.L. (2006) The plant immune system. Nature. **444**: 323-329.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**: 462–467.

Katiyar-Agarwal, S., Gao, S., Vivian-Smith, A., and Jin, H. (2007). A novel class of bacteria-induced small RNAs in Arabidopsis. Genes Dev. **21**: 3123–3134.

Katoh, K. and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics. **9**: 286–298.

Kawchuk, L.M., Martin, R.R. and McPherson, J.. (1991). Mol. Plant Microbe Inter. 4, 247.

Kim, D.Y., Lee, H.E., Yi, K.W., Han, S.E., Kwon, H.B., Go, S.J., Byun, M.O. (2003). Expression pattern of potato (Solanum tuberosum) genes under cold stress by using cDNA microarray. Kor J Genet. **25**: 345–352.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14**: R36.

Kloosterman, B., Vorst, O., Hall, R.D., Visser, R.G.F., Bachem, C.W. (2005). Tuber on a chip: Differential gene expression during potato tuber development. Plant Biotechnol J. **3**: 505–519.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics. **5**: 59.

Landan, G. and Graur, D. (2007). Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. Mol. Biol. Evol. **24**: 1380–1383.

Lassmann, T., Frings, O., and Sonnhammer, E.L.L. (2009). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res. **37**: 858–865.

Lee, B.-H., Henderson, D.A., and Zhu, J.-K. (2005). The Arabidopsis cold-responsive transcriptome and its regulation by ICE1. Plant Cell. **17**: 3155–3175.

Li, R., Li, Y., Kristiansen, K., Wang, J. (2008). SOAP: Short oligonucleotide alignment program. Bioinformatics. **24**: 713–714.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**: 2178–2189.

Li, X.Q. (2008). Molecular characterization and biotechnological improvement of the processing quality of potatoes. Can J Plant Sci. **88:** 639–648.

Li, C., Zhou, A., and Sang, T.. (2006). Genetic analysis of rice domestication syndrome with the wild annual species, *Oryza nivara*. New Phytol. **170** : 185 – 194 .

Luo, R. et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. **1**: 18.

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science **252**: 1162–1164.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics **154**: 459–473.

Lyapkova N.S., Loskutova, N.A., Maisuryan, A.N., Mazin, V.V., Korableva, N.P., Platonova, T.A., Ladyzhenskaya, E.P. and Evsyunina, A.S.. (2001). Appl. Biochem. Microbiol. 37, 301.

Malarkey, T. (2003). Human health concerns with GM crops. Mutat. Res. **544**: 217-21.

Maruyama, K. et al. (2012). Identification of cis-acting promoter elements in cold- and dehydration-induced transcriptional pathways in Arabidopsis, rice, and soybean. DNA Res. **19**: 37–49.

Maruyama, K., Sakuma, Y., Kasuga, M., Ito, Y., Seki, M., Goda, H., Shimada, Y., Yoshida, S., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2004). Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. Plant J. **38**: 982–993.

Mathelier, A. and Carbone, A. (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. Bioinformatics. **26**: 2226–2234.

Matsui, A., Nguyen, A.H., Nakaminami, K., and Seki, M. (2013). Arabidopsis non-coding RNA regulation in abiotic stress responses. Int J Mol Sci. **14**: 22642–22654.

Matsumoto, T., J. Z. Wu, H. Kanamori, Y. Katayose, M. Fujisawa, N. Namiki, H. Mizuno et al. (2005). The map-based sequence of the rice genome. Nature. **436**: 793 – 800 .

Mboup, M., Fischer, I., Lainer, H., and Stephan, W. (2012). Trans-species polymorphism and allele-specific expression in the CBF gene family of wild tomatoes. Mol. Biol. Evol. **29**: 3641–3652.

McDonnell, A.V., Jiang, T., Keating, A.E., and Berger, B. (2006). Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics. **22**: 356–358.

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A., and Stupar, R.M. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant physiol. **159**: 1295–1308.

McHale, L., Tan, X., Koehl, P. and Michelmore, R.W. (2006) Plant NBS‐LRR proteins: adaptable guards. Genome biol. **7**: 212.

Means, T.K., Golenbock, D.T. and Fenton, M.J. (2000) The biology of Toll‐like rece*pto*rs. Cytokine growth factor rev. **11**: 219‐232.

Medina, I. et al. (2010). Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. **38**: W210–3.

Michael, T.P. and Jackson, S. (2013). The First 50 Plant Genomes. The Plant Genome. **6**: 7.

Micheletto, S., Boland, R., and Huarte, M. (2000). Argentinian wild diploid Solanum species as sources of quantitative late blight resistance - Springer. Theoretical and Applied Genetics.

Millam, S., (2005). In: I. Curtis (ed.), Transgenic Crops of the World – Essential Protocols, pp. 257–270. Kluwer Academic, Dordrecht.

Morillo, S.A. and Tax, F.E. (2006) Functional analysis of rece*pto*r‐like kinases in monocots and dicots. Curr Opin Plant Biol. **9**: 460‐469.

Naika, M., Shameer, K., Mathew, O.K., Gowda, R., and Sowdhamini, R. (2013). STIFDB2: an updated version of plant stress-responsive transcription factor database with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in Arabidopsis and rice. Plant Cell Physiol. **54**: e8.

Nawrocki, E.P. and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. **29**: 2933–2935.

Ohno, S. (1970). Evolution by gene duplication (Springer-Verlag: New York).

Ortiz, R., Franco, J., Iwanaga, M. (1997): Transfer of resistance to potato cyst nematode (Globodera pallida) into cultivated potato *Solanum tuberosum* through first division restitution 2n pollen. Euphytica. **96**: 339-344.

Oufir, M., Legay, S., Nicot, N., Van Moer, K., Hoffmann, L., Renaut, J., Hausman, J.F., Evers, D. (2008). Gene expression in potato during cold exposure: Changes in carbohydrate and polyamine metabolisms. Plant Sci. **175**: 839–852.

Palta, J.P. and Simon, G. (1993). Breeding potential for improvement of freezing stress resistance: genetic separation of freezing tolerance, freezing avoidance, and capacity to cold acclimate. In Advances in Plant Cold Hardiness, P.H. Li, ed (CRC Press: Boca Raton, Florida, USA), pp. 299–310.

Park, Y. and Cheong, H. (2002). Protein Express Purif. **25**: 160.

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. Nucleic Acids Res. **37**: 289–297.

Paterson, A. H., Bowers J. E., Bruggmann R., Dubchak, I., Grimwood, J., Gundlach H., Haberer G., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457**: 551 – 556.

Pavek, J.J. and Corsini, D.L. (2001). Utilization of potato genetic resources in variety development - Springer. Am. J. Pot Res. **16**: 839–852.

Peele, H.M., Guan, N., Fogelqvist, J., Dixelius, C. (2014). Loss and retention of resistance genes in five species of the Brassicaceae family. BMC Plant Biol. **14**: 298.

Pennycooke, J.C., Cheng, H., Roberts, S.M., Yang, Q., Rhee, S.Y., and Stockinger, E.J. (2008). The low temperature-responsive, Solanum CBF1 genes maintain high identity in their upstream regions in a genomic environment undergoing gene duplications, deletions, and rearrangements. Plant Mol. Biol. **67**: 483–497.

Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S.A.,. Wing, R.A., Panaud, O. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition- driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genet. Res. **16**: 1262–1269.

Potato Genome Sequencing Consortium et al. (2011). Genome sequence and analysis of the tuber crop potato. Nature. **475**: 189–195.

Ramsay, N.A. and Glover, B.J. (2005). MYB-bHLH-WD40 protein complex and the evolution of cellular diversity. Trends Plant Sci. **10**: 63–70.

Rensink WA, Buell CR. (2005). Microarray expression profiling resources for plant genomics. Trends Plant Sci. **10**: 603–609.

Restrepo, S., Myers, K.L., Del Pozo, O., Martin, G.B., Hart, A.L., Buell, C.R., Fry, W.E., Smart, C.D. (2005). Gene profiling of a compatible interaction between Phytophthora infestans and Solanum tuberosum suggests a role for carbonic anhydrase. Mol Plant-Microbe Interact. **18**: 913–922.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. Cell. **110**: 513–520.

Rodriguez, F., Wu, F., Ané, C., Tanksley, S., and Spooner, D.M. (2009). Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? BMC Evol. Biol. **9**: 191.

Rodríguez, F. and Spooner, D.M. (2009). Nitrate Reductase Phylogeny of Potato (Solatium sect. Petota) Genomes with Emphasis on the Origins of the Polyploid Species. Systematic botany. Systematic Botany. **34**: 207-219

Rouillard, J.-M., Zuker, M., and Gulari, E. (2003). OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. **31**: 3057–3062.

Rymarquis, L.A., Kastenmayer, J.P., Hüttenhofer, A.G., and Green, P.J. (2008). Diamonds in the rough: mRNA-like non-coding RNAs. Trends Plant Sci. **13**: 329–334.

Rutitzky, M., Ghiglione, H.O., Curá, J.A., Casal, J.J., Yanovsky, M.J. (2009). Comparative genomic analysis of light-regulated transcripts in the Solanaceae. BMC Genom **10**: 60.

Sakamoto, H., Maruyama, K., Sakuma, Y., Meshi, T., Iwabuchi, M., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2004). Arabidopsis Cys2/His2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. Plant physiology **136**: 2734–2746.

Salas, A.R., D.M. Spooner, Z. Huamán, R. Vinci Torres Maita, R. Hoekstra, K. Schüler, and R.J. Hijmans. (2001). Taxonomy and new genetic resources of wild potato collections in Central Peru (Departments of Ancash, Huancavelica, Junin, La Libertad, Lima) in 1999, and new data on collections in southern Peru in 1999. American Journal of Potato Research **78**:197–207.

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., Frusciante, L., and Ercolano, M.R. (2010). PRGdb: a bioinformatics platform for plant resistance gene analysis. Nucleic Acids Res. **38**: D814–21.

Satish, K., Srinivas G., Madhusudhana, R., Padmaja, P., Nagaraja Reddy, R., Murali Mohan, S., Seetharama, N. (2009). Identification of quantitative trait loci for resistance to shoot fly in sorghum . [Sorghum bicolor ; (L.) Moench] Theor Appl Genet. **119**: 1425 – 1439

Schafleitner, R., Gutierrez Rosales, R.O., Gaudin, A., Alvarado Aliaga, C.A., Martinez, G.N., TincopaMarca, L.R., Bolivar, L.A., Delgado, F.M., Simon, R., Bonierbale, M. (2007). Capturing candidatedrought tolerance traits in two native Andean potato clones by transcription profiling of field grown plants under water stress. Plant Physiol Biochem. **45**: 673–690.

Scheibye-Alsing, K., Hoffmann. S., Frankel, A., Jensen, P., Stadler, P.F., Mang, Y., Tommerup, N., Gilchrist, M.J., Nygård, A.B., Cirera, S., Jørgensen, C.B., Fredholm, M.,Gorodkin, J. (2009). Sequence assembly. Computational Biology and Chemistry. **33**: 121–136.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., et al . (2009**)**. The B73 maize genome: Complexity, diversity, and dynamics. Science. **326**: 1112 - 1115.

Scott, G.J., Rosegrant M.W.vand Ringler, C. (2000). Roots and tubers for the 21st century – trends, projections, and policy options. International Food Policy Research Institute. http://www.ifpri.org.

Sharma, S.K. and Millam, S. (2004). Somatic embryogenesis in Solanum tuberosum L.: a histological examination of key developmental stages. Plant Cell Rep. **23**: 115-9.

Shulaev, V., Sargent, D.J., Crowhurst, R.N., Mockler, T.C., Folkerts, O., Delcher, A.L., Jaiswal, P., et al. (2011). The genome of woodland strawberry (Fragaria vesca). Nature Genet. **43**: 109 – 116.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. J. Mol. Biol. **147**: 195–197.

Snow A.A.(2002). Nat. Biotechnol. **20**: 542.

Solovyev, V. and Salamov, A. (1997) The Gene ‐ Finder computer tools for analysis of human and model organisms genome sequences. Proceedings - International Conference on Intelligent Systems for Molecular Biology ; ISMB. **5**: 294 ‐ 302.

Spooner, D.M. and R.J. Hijmans. (2001). Potato systematics and germplasm collecting, 1989–2000. American Journal of Potato Research. **78**: 237–268.

Spooner, D.M. and R.G. Van den Berg. (2001). Quantitative assessment of corolla shape variation in Mexican Solanum sect. Petota. In: R.G. Van den Berg, G. Barendse, G.W. van der Weerden, and C. Mariani (eds.), Solanaceae V: Progress in taxonomy and utilization. Nijmegen University Press, Nijmegen, Netherlands. pp. 61–72.

Spooner, D.M., McLean, K., Ramsay, G., Waugh, R., and Bryan, G.J. (2005). A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. Proc. Natl. Acad. Sci. U.S.A. **102**: 14694–14699.

Stanke, M. and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. **19 Suppl 2**: ii215–25.

Stevenson, W.R. (2001). Compendium of Potato Diseases (Amer Phytopathological Society).

Stockinger, E.J., Mao, Y., Regier, M.K., Triezenberg, S.J., and Thomashow, M.F. (2001). Transcriptional adaptor and histone acetyltransferase proteins in Arabidopsis and their interactions with CBF1, a transcriptional activator involved in cold-regulated gene expression. Nucleic Acids Res. **29**: 1524–1533.

Stone, J.M., Palta, J.P., Bamberg, J.B., Weiss, L.S., and Harbage, J.F. (1993). Inheritance of freezing resistance in tuber-bearing Solanum species: evidence for independent genetic control of nonacclimated freezing tolerance and cold acclimation capacity. Proc. Natl. Acad. Sci. U.S.A. **90**: 7869–7873.

Stracke, R., Werber, M., and Weisshaar, B. (2001). The R2R3-MYB gene family in Arabidopsis thaliana. Curr. Opin. Plant Biol. **4**: 447–456.

Stupar, R.M. (2010). Into the wild: The soybean genome meets its undomesticated relative. PNAS. **107**: 21947–21948.

Stupar, R.M., Bhaskar, P.B., Yandell, B.S., Rensink, W.A., Hart, A.L., Ouyang, S., Veilleux, R.E., Busse, J.S., Erhardt, R.J., Buell, C.R., Jiang, J. (2007). Phenotypic and transcriptomic changes associated with potato autopolyploidization. Genetics. **176**: 2055–2067.

Sunkar, R., Kapoor, A., and Zhu, J.-K. (2006). Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. Plant Cell. **18**: 2051–2065.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE. **6**: e21800.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res. **18**: 1944–1954.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. **18**: 1979–1990.

Teutonico, R.A., Yandell, B., Satagopan, J.M., Ferreira, M.E., Palta, J.P., and Osborn, T.C. (1995). Genetic analysis and mapping of genes controlling freezing tolerance in oil seed Brassica. Mol Breeding. **1**: 329–339.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature. **408**: 796-815

Tian, Z.D., Liu, J., Wang, B.L., Xie, C.H. (2006). Screening and expression analysis of Phytophthora infestans induced genes in potato leaves with horizontal resistance. Plant Cell Rep **25**:1094–1103.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**: 635–641.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**: 511–515.

Tsarouhas, V., Gullberg, U., and Lagercrantz, U. (2004). Mapping of quantitative trait loci (QTLs) affecting autumn freezing resistance and phenology in Salix. Theor. Appl. Genet. **108**: 1335–1342.

Ugent, D. (1970). The potato. Science. **170**: 1161–1166.

Van den Berg R.G., Miller, J.T., Ugarte, M.L., Kardolus, J.P., J. Villand, J. Nienhuis and Spooner, D.M., (1998). Am. J. Bot. **85**: 92.

Van Dijk, J.P., Cankar, K., Scheffer, S.J., Beenen, H.G., Shepherd, L.V.T., Stewart, D., Davies, H.V., Wilkockson, S.J., Leifert, C., Gruden, K., Kok, E.J. (2009). Transcriptome analysis of potato tubers-effects of different agricultural practices. J Agri Food Chem. **57**: 1612–1623.

Van Nieuwerburgh, F., Thompson, R.C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhanian, P., and Head, S.R. (2012). Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Res. **40**: e24.

Vardy K.A., M.J. Emes and M.M. Burrell. (2002). Funct. Plant Biol. **29**: 975.

Varshney, R.K., Chen, W., Li, Y., Bharti, A.K., Saxena, R.K., Schlueter, J.A., Donoghue, M.T.A., Azam, S., Fan, G., Whaley, A.M., Farmer, A., Tuetja, R., Penmetsa, R.V., Wu, W., Upadhyaya, H., Yang, S.-P., Shah, T., Saxena, K.B., Ward, E., Michael, T., McCombie, W. R., Yang, B., Jones, J.D.G., Spillane, C., Cook, D.R., May, G.D., Xu, X., Jackson, S.A. (2012). Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat Biotechnol. **30**: 83-9.

Velasco, R., A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana, et al . (2010). The genome of the domesticated apple (Malus ×domestica Borkh.). Nature Genet. **42**: 833 – 839.

Vitte, C. and Bennetzen, J.L. (2006). Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad. Sci. U.S.A. **103**: 17638–17643.

Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. **34**: 1692–1699.

Wang, B.L., Liu, J., Tian, Z.D., Song, B.T., Xie, C.H. (2008). cDNA microarray analysis of metabolismrelated genes in inoculated potato leaves expressing moderate quantitative resistance to Phytophthora infestans. J Hort Sci Biotechnol. **83**: 419–426.

Wang, Y., Li, J., and Paterson, A.H. (2013). MCScanX-transposed: detecting transposed gene duplications based on multiple colinearity scans. Bioinformatics. **29**: 1458–1460.

Wang, Z., Zhu, Y., Wang, L., Liu, X., Liu, Y., Phillips, J., and Deng, X. (2009). A WRKY transcription factor participates in dehydration tolerance in Boea hygrometrica by binding to the W-box elements of the galactinol synthase (BhGolS1) promoter. Planta. **230**: 1155–1166.

Wang-Pruski, G. and Schofield, A. (2012). Potato: Improving Crop Productivity and Abiotic Stress Tolerance. In Improving Crop Resistance to Abiotic Stress, N. Tuteja, S.S. Gill, F.A. Tiburcio, and R. Tuteja, eds (WILEY-VCH Verlag: Weinheim, Germany), pp. 1121–1153.

Wehe, A., Bansal, M.S., Burleigh, J.G., and Eulenstein, O. (2008). DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. **24**: 1540–1541.

Wicker, T., Sabot, F., Hua-Van, A.,. Bennetzen, J.L, Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**: 973–982.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. **13**: 329–342.

Zhang, C., Liu, L., Wang, X., Vossen, J., Li, G., Li, T., Zheng, Z., Gao, J., Guo, Y., Visser, R.G., Li, J., Bai, Y., Du, Y. (2014). The Ph-3 gene from Solanum pimpinellifolium encodes CC-NBS-LRR protein conferring resistance toPhytophthora infestans. Theor Appl Genet. **127(6)**:1353-64

Zhuo, C., Wang, T., Lu, S., Zhao, Y., Li, X., and Guo, Z. (2013). A cold responsive galactinol synthase gene from Medicago falcata (MfGolS1) is induced by myo-inositol and confers multiple tolerances to abiotic stresses. Physiologia Plantarum. **149**: 67–78.